

## Two-layer modular analysis of gene and protein networks in breast cancer

Srivastava *et al.*

RESEARCH ARTICLE

Open Access

# Two-layer modular analysis of gene and protein networks in breast cancer

Alok Srivastava<sup>1\*</sup>, Suraj Kumar<sup>1,2</sup> and Ramakrishna Ramaswamy<sup>3,4\*</sup>

## Abstract

**Background:** Genomic, proteomic and high-throughput gene expression data, when integrated, can be used to map the interaction networks between genes and proteins. Different approaches have been used to analyze these networks, especially in cancer, where mutations in biologically related genes that encode mutually interacting proteins are believed to be involved. This system of integrated networks as a whole exhibits emergent biological properties that are not obvious at the individual network level. We analyze the system in terms of modules, namely a set of densely interconnected nodes that can be further divided into submodules that are expected to participate in multiple biological activities in coordinated manner.

**Results:** In the present work we construct two layers of the breast cancer network: the gene layer, where the correlation network of breast cancer genes is analyzed to identify gene modules, and the protein layer, where each gene module is extended to map out the network of expressed proteins and their interactions in order to identify submodules. Each module and its associated submodules are analyzed to test the robustness of their topological distribution. The constituent biological phenomena are explored through the use of the Gene Ontology. We thus construct a “network of networks”, and demonstrate that both the gene and protein interaction networks are modular in nature. By focusing on the ontological classification, we are able to determine the entire GO profiles that are distributed at different levels of hierarchy. Within each submodule most of the proteins are biologically correlated, and participate in groups of distinct biological activities.

**Conclusions:** The present approach is an effective method for discovering coherent gene modules and protein submodules. We show that this also provides a means of determining biological pathways (both novel and as well those that have been reported previously) that are related, in the present instance, to breast cancer. Similar strategies are likely to be useful in the analysis of other diseases as well.

**Keywords:** Protein-protein interaction, Gene expression, Networks, Modules, Gene ontology

## Background

The current paradigm in a *systems* approach to biological phenomena is that of networks and the interactions among them. Advances in genomic, proteomic and high-throughput gene expression data, when integrated, can be used to map the interaction networks between genes and proteins, as well as their association with specific biological activities. It has also become increasingly clear that an integrated analysis of these extensive components is crucial, especially in the case of cancers [1].

Mutation in biologically correlated genes affects the translation of key proteins that do not function in isolation: distinct biological activities are the result of the coordinated action of multiple proteins [2] and a reduction in the synthesis of one protein can directly affect various specialized biological actions.

In the case of breast cancer for example, the interaction network of 6004 proteins is, in different combinations, associated with 5732 biological processes (BP), 1930 molecular functions (MF) and 879 cellular components (CC) as specified in the Gene Ontology Annotation (GOA) [3] database. On this scale it is difficult to interpret the organization principle of such networks that may be

\* Correspondence: foralok@gmail.com; r.ramaswamy@gmail.com

<sup>1</sup>C R RAO Advanced Institute of Mathematics, Statistics and Computer Science, University of Hyderabad Campus, Hyderabad 500046, India

<sup>3</sup>University of Hyderabad, Hyderabad 500046, India

Full list of author information is available at the end of the article

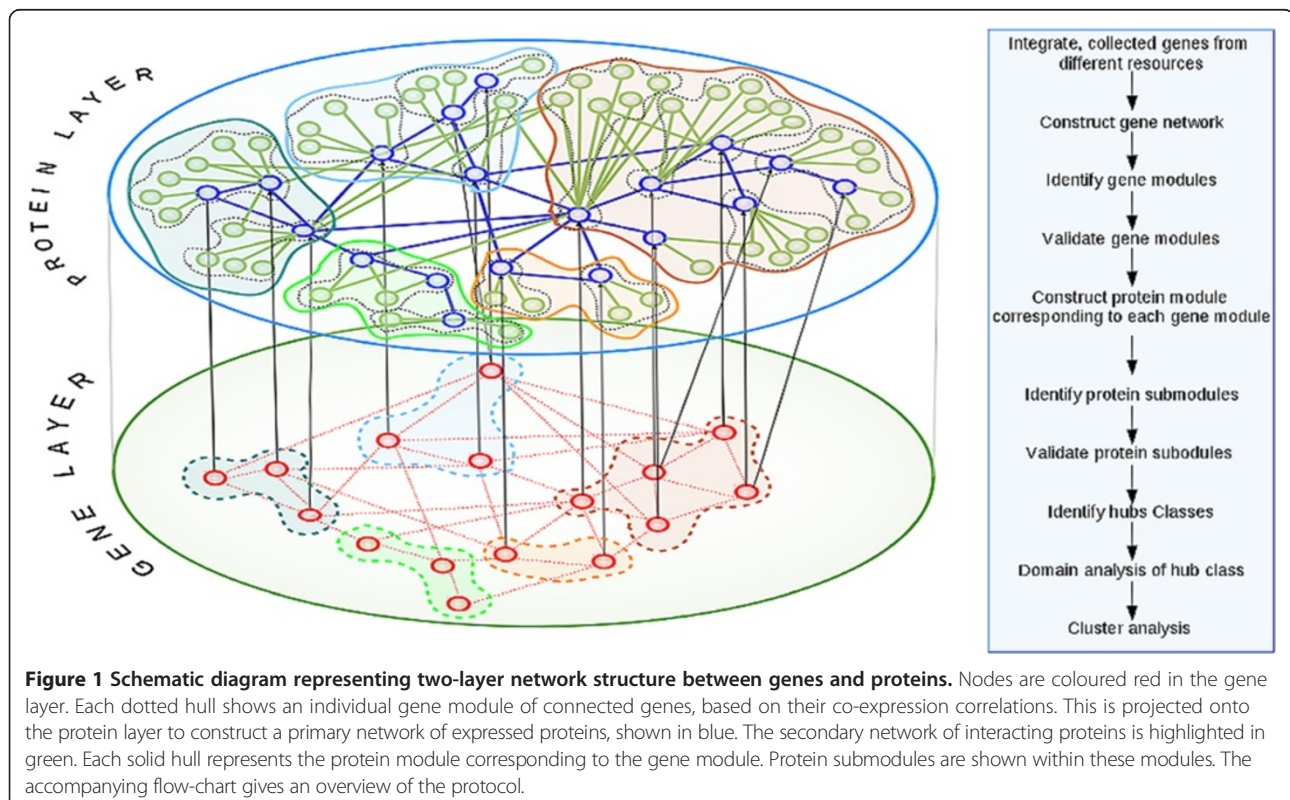
composed of thousands of structural subunits. The more highly connected subunits participate in multiple biological activities [4].

The alternative *bottom-up* approach, namely breaking up the complex network into several interacting sub-networks, can be more helpful. These sub-networks help to analyse the activity at various levels of specificity, especially in case of complex diseases where the main interest is to elucidate the coordination principle that controls the progression of the disease.

The modular nature of a wide variety of complex networks has been investigated in detail in recent years, ranging from social networks [5], cellular phone networks [6], collaboration networks [7], citation networks [8], gene co-occurrence network [9], protein-protein interaction (PPI) networks [10], and metabolic networks [10]. A module can be defined as a subset of the nodes such that nodes within the module are densely connected while being sparsely connected with nodes in other modules [11]. Modules are the building blocks of higher-level functional organization, and can exhibit hierarchical properties. In particular, modules can be recursively divided into smaller submodules; such submodules are potentially a rich source of information on biological networks [12-15]. Nodes within a submodule are more likely to have closely related biological properties, and thus separating a network into modules and submodules can make it possible to understand the more

specific domain of activities in which they participate, either singly or in a coordinated manner. (By domain we mean here the specializations defined by Gene Ontology (GO) [16], namely BP, MF or CC.)

In the present paper we propose a general framework for the analysis of breast cancer data. By combining information on correlated genes and knowledge of their associated proteins, it is possible to construct a model two-layer network that can be explored to uncover the underlying biological phenomena in cancers. Analyzing the network as a whole gives insight into all annotated biological activities and specific sub-activities within the cell. On the *gene layer* we construct a correlation network from the expression data available for disease specific genes using Graphical Gaussian Modeling (GGM) [17]. The corresponding protein layer is constructed from the known set of proteins expressed by each of the genes to make the primary network, and a secondary network that shows their corresponding interactions is constructed from available protein interaction data. This two layer analysis is schematically shown in Figure 1. On each layer, modules are identified through a fast greedy optimization algorithm [18] applied recursively. We also investigate the robustness of the various modules by examining the effect of node removal. Each protein submodule is further divided into an inter-modular bridge class and an intra-modular group class, so as to identify the domains with the largest number of participating



proteins. The overall aim of the present procedure is in effect to construct a network of networks and to examine the connections between the different domain groups within the protein submodules.

There are several methods that can be used to detect modularity, and a number of algorithms have been described [10,18-24]. In our study we focus on the gene module and protein submodule detection using a fast greedy modularity optimization technique [18] that is efficient in the analysis of large networks. The detailed topology of the submodules is interesting: the group of proteins within a submodule that interact physically with other proteins tend to form two sorts of hubs. The so called 'party hubs' interact with most (or all) of their partners simultaneously, while 'date hubs' are those that interact with different partners at different times [25]. Party hubs are intra-modular and the constituent proteins are believed to perform a specific biological activity within the submodule. In contrast, the activity of date hubs is intermodular and link different biological domains [26]. Proteins belonging to each hub within a submodule can be further explored from an ontological perspective: a group of proteins can be located in one or more cellular component, be active in one or more biological processes, and perform one or more molecular functions, but since domain activity is shared by several (and overlapping) groups, classification using GO terms can become very complex. This can be reduced to some extent by the implementation of a majority rule or by GO homogeneity (GO-H) analysis [27]. In the present case we use this latter reduction to identify GO terms associated with the largest fraction of proteins within a given hub class (i.e. party or date).

A domain cluster is a group of three or more GO terms connected together by a *parent-child* relationship, namely these form a directed acyclic graph (DAG) [16]. GO terms up to the 2nd level is general while terms at the lower levels are specific, and we explored the biological domain cluster of similar GO terms through OntoVisT [28], a general purpose tool developed earlier for interactive visualization and navigation of any ontology.

We present our findings using an exhaustive list of breast cancer genes, proteins and their PPI network from the Human Protein Reference Database (HPRD) [29] in combination with the comprehensive and well-established microarray datasets from breast cancer patients [30,31].

## Methods

### Materials: breast cancer data

We have integrated nine cancer resources in order to obtain a comprehensive list of breast cancer marker genes. The different resources focus on different aspects of biology:

- NCG (Network of Cancer Genes) has data on gene mutations [32],
- TGDBs (Tumor Gene Family of Databases) lists the target genes implicated in cancer-causing mutations [33],
- CGW (Cancer Genetics Web) details primary mutations that cause cancer, as well as secondary genetic abnormalities caused by cancer [34],
- CGC (Cancer Gene Census) catalogues all genes whose mutations have been implicated in cancer [35],
- KEGG (Kyoto Encyclopedia of Genes and Genomes) integrates current knowledge on molecular interaction networks [36],
- BCGD (Breast Cancer Gene Database) collects molecular genetic data related to genes involved in breast cancer [37],
- CGAP (Cancer Genome Anatomy Project) lists gene expression profiles of cancer cells [38],
- OMIM (Online Mendelian Inheritance in Man) is a compendium of information on genetic disorders and genes [39], and
- GAD (Genetic Association Database) contains genetic association study data reported in the literature [40].

From the above integrated sources we identified 975 breast cancer genes of which 956 are unique (since some genes have more than one identifier [41]). This is a more exhaustive list than what has been hitherto available, and is summarized in Table 1. Details of the complete marker gene list are available in Additional file 1: Table S1.

### Gene expression microarray data

We use the comprehensive and well curated microarray data sets first studied by West et al. [30] and Györfy et al. [31], using UniProt [41] in order to convert synonymous gene symbols to their current approved labels. These datasets are termed Set 1 and Set 2 respectively

**Table 1 Breast cancer gene information**

Data source	BC genes	Processed genes	Version	Reference
KEGG	9	9	Jan, 2010 (V-53)	WGS
CGC	19	20	Apr, 2011	LC
BCGD	62	62	1999	LC
TGDB	67	67	1999	LC
CGAP	69	69	Jan, 2010	GE
CGW	82	83	Apr, 2003	LC
NCG	140	144	Jun, 2011 (V-2.1)	HTMS
GAD	695	690	Jun, 2011	LC
Total	975	956		

WGS, Whole genome sequencing; LC, Literature curation; GE, Gene Expression; HTMS, High throughput mutational screening; V, Version;

throughout this paper. Samples with more than 20% missing values were discarded, and the standard  $K$  nearest neighbor (KNN) imputation technique [42] with  $k = 10$  was used to estimate the missing values in the other samples. Imputed data was then averaged over all replicates to obtain a processed data matrix; the first dataset thus contains 49 breast tumor samples separated into two classes: 25 positive samples for estrogen receptor (ER+) and 24 negative (ER-) with expression levels measured for 5728 genes. The second dataset has 1809 breast tumor samples separated into three classes: 295 HER2+ samples, 1285 ER+/HER2- samples, and 229 ER-/HER2- samples with expression levels measured for 12496 genes. The normalized expression for the ESR1 gene (Affymetrix ID 205225) at above 500 is considered as ER+, while ERBB2 gene (Affymetrix ID 216836\_s) above 4800 is regarded as HER2+ [31].

#### **Gene-protein association and protein-protein interaction data**

In order to construct the primary network of expressed proteins we map genes to the UniProt (July 2011) dataset [41] that lists 41149 proteins expressed by 19511 genes. The primary network is further extended to form a secondary network of interacting proteins using HPRD (Release 9, Aug 2011) [29]. This database has two classes of interaction data: binary (if two proteins interact directly) and complex (when several proteins form a complex). Since protein complexes also constitute functional groups and take part in activation or inhibition [43] we also included these complexes in our study with the assumption that all proteins in a complex interact with each other. There are a total of 91029 binary interactions between 13494 proteins and 1521 protein complexes that are associated with 3652 proteins. On combining both, we obtain 13691 proteins, with a total of 110613 interactions.

#### **Gene ontology annotation data**

GO data was used to explore the biological activity of a group of proteins, by analyzing the molecular functions (MFs) which they perform, the biological processes (BPs) in which they participate and their cellular components. The ontology consists of over thirty thousand terms distributed across 12 different levels of hierarchy, starting from generic terms at the highest level, to more specific terms at the leaf nodes. To incorporate the biological domain knowledge to the entire groups of proteins, we used GOA data (July 2011) [3] and found that the 40422 proteins are mapped to 7488 GO:BP terms, 3221 GO:MF terms, and 1053 GO:CC terms respectively.

#### **Methods**

##### **Gene network using graphical Gaussian modeling**

Microarray data for the integrated set of breast cancer genes from different resources were used to construct the gene-gene interaction network based on graphical Gaussian modeling (GGM) [17]. The false discovery rate (FDR) criterion is then chosen to filter out the least significant genes, thus providing a computational criterion that permits us to determine which edge is to be included in the network. To construct the gene network based on the GGM algorithm, we used the Gene Net algorithm that is included in the R package [44].

##### **Modularity optimization using fast greedy technique**

The gene network is further analyzed to identify modules using the CNM fast greedy optimization technique [18]; this identifies a subset of nodes that are densely connected within modules while being sparsely connected to nodes in other modules. We use the igraph implementation contained in the R package [45] to identify gene modules and protein submodules.

##### **Leave-one-out cross-validation (LOO-CV) and majority-voting technique**

To test the robustness of the modular structure obtained above, we implemented the standard (but computationally expensive) LOO-CV technique by removing a node and its corresponding interactions from the data to predict the changed module distribution. The process is repeated iteratively for each node in the complete network, and majority voting is then used to identify the most robust modules.

##### **Protein network construction**

Each gene module is projected onto the protein layer in order to construct the primary network of expressed proteins and further extended to include the secondary network of interacting proteins. Combining both primary and secondary network components gives the two-layer network shown in Figure 1. Each protein module is then further analyzed to identify the submodules, using the CNM algorithm along with LOO cross validation.

Each protein submodule obtained above is divided into bridge (B) and group (G) hub classes, and further analyzed using GO homogeneity (GO-H) to identify the domain groups of those proteins that have a majority participation in specific GO categories.

##### **GO analysis**

For a group of proteins contained in a hub class of submodule  $i$ , GO Homogeneity (GO-H), is defined as a

GO term that has the maximum fraction of proteins, among all the mapped GO terms.

$$GO-H_i = \max_j [n_j^i/n_i] = \max_j [GO h_j^i]$$

where  $n_i$  is the number of proteins in the group  $i$  that have any GO annotations, and  $n_j^i$  the number of proteins that have a specific GO term  $j$ .  $GO h_j^i$  represents fraction of protein contained in group  $j$ . After assigning  $n_j^i$  proteins to a GO term  $j$ ,  $n_i - n_j^i$  proteins still remain and these may participate in other groups. This is done iteratively for the complete list as shown in Figure 2 so as to assign GO terms as well as group labels to all the remaining proteins.

Unannotated proteins introduce errors when included in GO-H analysis, and are therefore removed from the list. In order to identify the domain cluster represented by a group of GO terms we use OntoVisT [28] which as mentioned above, is a general purpose ontological visualization tool.

## Results

As indicated in the Materials section we combined 9 different cancer resources to get 956 breast cancer marker genes, of which 610 and 840 genes were found to be present in Sets 1 and 2 respectively. These are then used to construct the gene layer networks for detailed analysis.

### Gene layer construction and analysis of gene modules

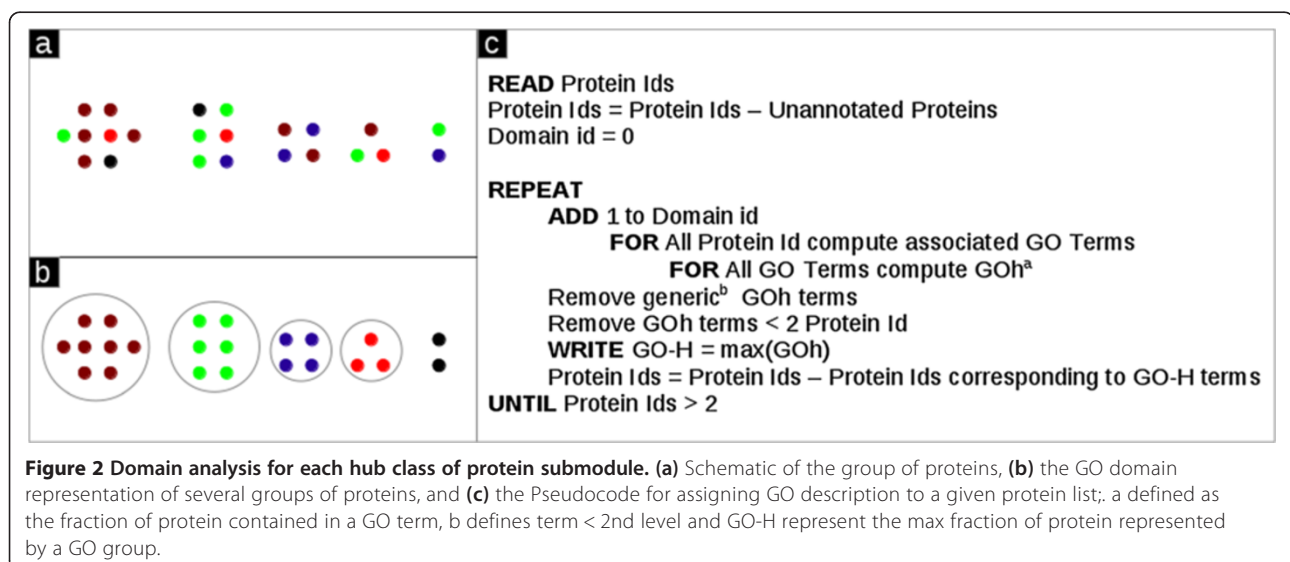
We construct a partial correlation based gene network using the GGM technique. For each edge the partial correlation (pcor), p-value (pval), FDR (qval) and corresponding posterior probabilities (prob) are calculated, and the FDR criterion is employed to assess its

significance. Table 2 summarizes the information regarding the gene layer network. The sparse nature of these networks with a fairly low number of genes results in a low modularity score  $Q_{max}$  for the optimal partitions.

For Set 1 a total of 4619 edges are found with a significance level of 30% FDR among the 513 genes (complete data given in Additional file 2: Table S2). The data was further stratified into ER+ and ER- subtypes, and the same analysis was carried out (see Additional file 2: Tables S2). For Set 2, there were a total of 3488 edges at the same significance among the 796 genes (see Additional file 3: Table S3), and the samples in Set 2 were stratified into the three subtypes HER2+, ER+/HER2- and ER-HER2- as discussed in the Materials and Methods section. The analysis for the stratified samples gives the result presented in Table 2.

The networks and modules that result from the present analysis are consistent across stratification with some variation. Consider Set 1. When all the samples are taken together, we find 7 modules with a modularity score of 0.25. Most of the genes are included in both the ER+ and ER- sample networks, and the results for the subtypes share a high similarity (86% and 74%, respectively) with the results for the total sample set. The number of modules for ER- is slightly larger and may indicate an increased heterogeneity of the subtype, but the modules themselves are quite similar.

In Set 2, stratification again gives results that are fairly consistent, with the number of genes being nearly constant across the subtypes. The number of modules varies although the similarity measure again gives a high level of concordance between the modules obtained by considering the entire sample set and the subtypes (87%, 83% and 88% respectively).



**Table 2 Summary of gene layer network for both datasets of breast cancer genes**

	Set 1: West et al. [30]			Set 2: Györfy et al. [31]			
	All samples	ER+	ER-	ALL samples	HER2+	ER+/HER2	ER-/HER2-
Number of samples	49	25	24	1809	295	1285	229
Number of genes	610	610	610	796	796	796	796
$\lambda$	0.80	0.81	0.83	0.03	0.14	0.04	0.18
qval	0.30	0.30	0.30	0.30	0.30	0.30	0.30
Genes	513	464	459	796	791	789	793
Number of interactions	4619	5718	1960	3488	6086	2745	7546
$\alpha$	1.23	1.37	1.38	1.35	1.25	1.40	1.21
Qmax	0.25	0.19	0.47	0.35	0.23	0.39	0.23
Number of modules	7	7	10	10	7	11	6
% accuracy, LOO-CV	78	71	60	57	61	76	72
Similarity measure		86	74		87	83	88

$\lambda$ , optimal shrinkage intensity; qval, cut off for the partial correlation (the threshold is adjusted so as to ensure that all well-studied breast cancer marker genes are included in the networks);  $\alpha$ , power-law exponent; Qmax, maximum modularity score; LOO-CV, Leave-one-out cross-validation.

Cross validation results (LOO-CV followed by majority voting) indicate that overall, 78% of the modular structure remains intact for Set 1 (71% for the ER+ subtype and 60% for ER-) while only 57% is preserved for Set 2 (and correspondingly 61% for HER2+, 76% for ER+/HER2-, and 72% for ER-/HER2-). See Table 2. Some mismatch might be expected due to the removal of hubs or some other important nodes from the network, but on the whole, these results reliably show that the breast cancer gene network is modular in nature.

#### Protein layer construction and analysis of protein submodules

Modules that are identified in the gene layer for entire sample set are mapped to their expressed proteins. This gives the primary protein network, and the secondary network is then constructed by examining all proteins that interact with the primary network. Both networks when combined give the protein module that corresponds to each gene module. Together this forms the protein layer, as explained in Figure 1. We remove loops and multiple edges prior to application of the CNM algorithm to obtain submodules in the protein layer.

Our results for Set 1 are summarized in Table 3. There are 7 protein modules, each corresponding to a gene module and these are numbered by rank. The degree distribution of the networks in each protein module follows a power-law, with exponents between 1.5 and 1.8. The protein modules are much larger than the corresponding gene modules and these also have a high interaction density (namely the number of interactions per module).  $Q_{max}$  varies from 0.632 for module 3 which has a large number of proteins and correspondingly the most interactions, to 0.798 for module 6, a sparse graph

of relatively fewer proteins.  $Q_{max}$  is inversely correlated to the number of protein-protein interactions. Module 7 has only 7 genes associated with a sparse graph of 113 interactions among 114 proteins. Redoing our analysis using LOO-CV suggests that more than 86% of the sub-modular structure is robust. The smallest modules, 6 and 7 give the same submodule structure regardless of which nodes are left out. These modules are very sparse and do not contain any hubs. In the larger modules where there are hubs, removal of such a node can give very different submodules although the average  $Q_{max}$  value does not change significantly, indicating that modularity is a robust property.

Set 2 when similarly analysed (see Additional file 4: Table S4) gives 10 protein modules with similar power-law distribution in degree, with exponent in the same range, 1.58-1.75. The  $Q_{max}$  value shows larger variation, from 0.62 to 0.87, and the number of submodules vary from 10 to 40 with considerable inter-module linkage. In other aspects such as modularity robustness, this dataset is similar to Set 1.

#### The biological properties of submodules

Proteins in each submodule can be divided into two classes: group proteins (G) that form so-called party hubs that interact primarily with other proteins within the same submodule, and bridge proteins (B) which are the “date” hubs that interact with proteins in two or more submodules. Their associated domain properties are obtained by first mapping them to the GOA [3] followed by GO-H analysis. Generic GO terms are discarded up to the second level in the hierarchy, as also terms with fewer than three proteins. Each domain group of a submodule class is identified by a group

**Table 3 Summary of topological and biological analysis of gene module and their corresponding protein modules for Set 1**

Module Id	1	2	3	4	5	6	7	#U Sum
<b>Topological analysis</b>								
Genes	142	142	125	40	35	22	7	513
Expressed proteins	640	486	439	159	152	79	10	1958
PPI	7465	6576	8163	2106	1317	885	126	24752
Processed protein in PPI	3284	3206	3187	1338	797	575	114	6519
Processed PPI	6951	6095	7712	1985	1250	794	113	23173
$\alpha$	1.606	1.632	1.576	1.701	1.658	1.698	1.777	
Qmax	0.637	0.662	0.632	0.714	0.792	0.798	0.793	
Modules (Isolated modules)	309 (7)	259 (4)	28 (11)	26 (10)	23 (8)	17 (7)	6 (4)	155 (51)
%age accuracy, LOO-CV	86.510	87.410	95.199	99.028	99.624	100	100	
Avg. Qmax, LOO-CV	0.639	0.664	0.634	0.714	0.797	0.797	0.791	
Module range, LOO-CV	24-33	20-28	24-34	25-32	21-25	16-18	5-6	
Avg. Modules, LOO-CV	28	26	28	26	22	16	5	
<b>Biological analysis</b>								
GO:BP, Groups	137	135	111	71	45	34	6	539
GO:BP, Terms	96	117	96	67	47	37	8	205
GO:BP, Referenced	42	54	46	32	25	19	8	85
GO:MF, Groups	131	124	115	69	44	35	7	525
GO:MF, Terms	79	72	72	51	33	30	10	145
GO:MF, Referenced	55	40	43	34	24	23	8	91
GO:CC, Groups	112	95	92	55	42	32	6	434
GO:CC, Terms	15	14	14	11	10	13	6	23
GO:CC, Referenced	12	11	12	9	8	9	6	16

Topological analysis represents the analysis of network topology, while biological analysis represents the domain analysis of individual modules; Genes, Number of genes; Expressed Proteins, Proteins expressed by genes in module; PPI, Number of interaction between proteins; Processed protein in PPI, Number of proteins after removing loops and multiple edges; Processed PPI, Number of interaction in processed PPI;  $\alpha$ , Power law exponent; Qmax, maximum modularity score; LOO-CV, Leave-one-out cross-validation; Avg. modules, Mean number of modules in LOO-CV; GO:BP, GO:MF, GO:CC represent the GO biological process, GO molecular function, GO cellular component respectively, while Groups represents the domain groups annotated with each module, Terms represent the unique number of GO term; and Referenced represents GO terms previously reported in literature, #U Sum, Unique sum.

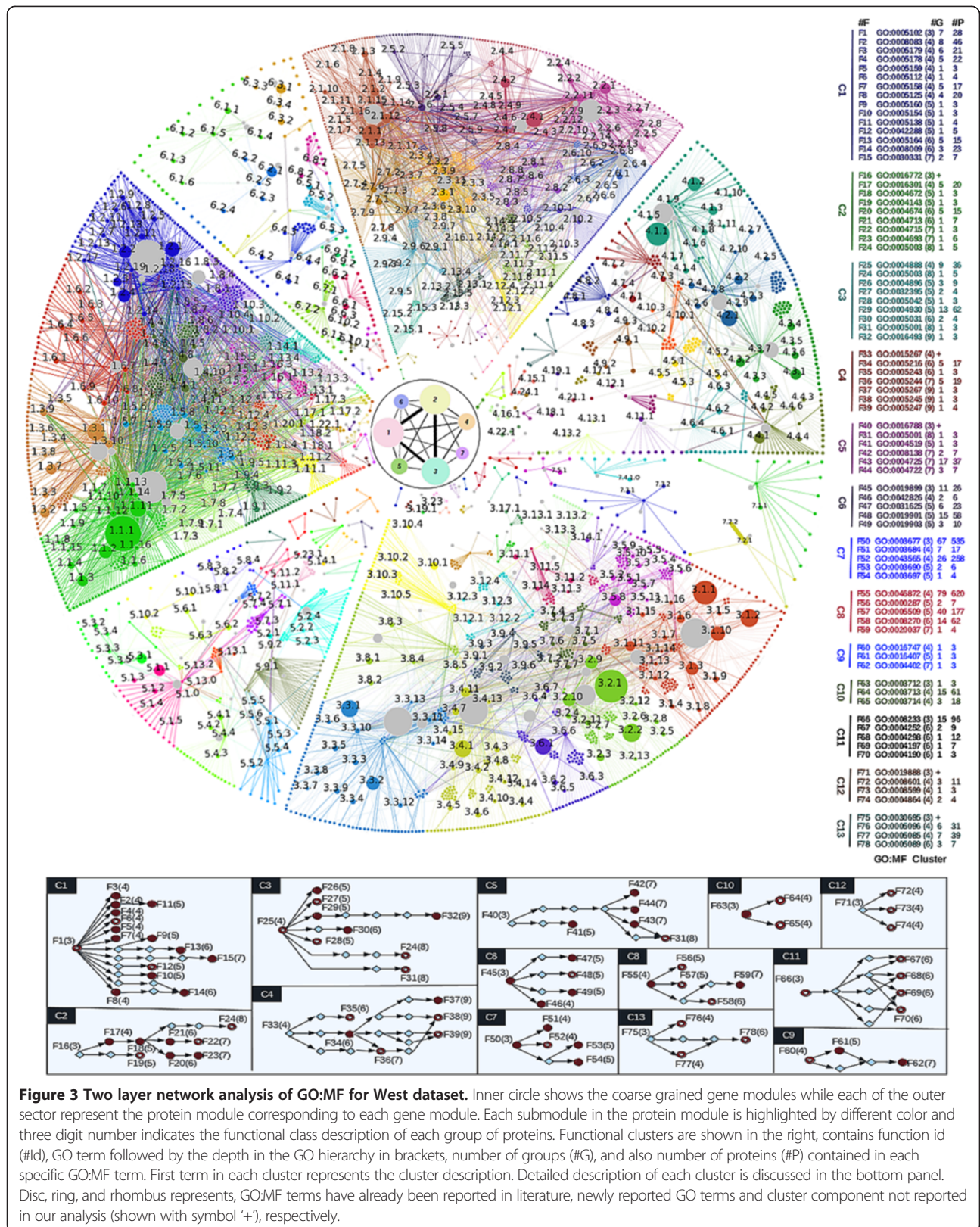
identifier (ID) which is a set of three integers written *a.b.c*, with *a* representing the module ID, *b* the submodule ID and *c* the domain ID assigned by our algorithm. The pseudocode is given above (see Figure 2). Each group ID is also associated with a group label G, B, OG or OB representing group, bridge, overlapping groups or overlapping bridges, based on their topological class category. To identify clusters represented by a given set of GO terms we use OntoVisT [28]; PUBMED is also queried in order to determine whether the GO term in the submodule has been previously reported as occurring in any breast cancer related study.

Both datasets are studied in order to infer the biological activity for different hub classes of each submodule. The three GO categories BP, MF, and CC are separately analyzed. The detailed analysis of the main cluster of each GO category in Set 1 is presented here, and results for the other clusters of both sets are provided in the supplementary information.

### Analysis of molecular functions

The two layer network deduced from the functional analysis of breast cancer genes found in Set 1 is shown in Figure 3. The inner circle depicts a coarse-grained image of the 7 modules in the gene layer. Details of the gene module and the protein submodules have been discussed above, and these are summarized in Table 3. The outer circle represents the protein layer, each sector representing the corresponding protein modules. Using HPRD [29] we have identified 23173 distinct protein-protein (PP) interactions in the network of 6519 proteins that are divided into 7 modules and 155 distinct submodules of which 51 are isolated. The interaction network in each sector shows the submodular network of the functional groups; these are highlighted in different colors. Each submodule is divided into group and bridge categories, based on their hub property, and these are then further analyzed for their functional associations using GO-H scores. Generic GO terms up to the second level





**Figure 3 Two layer network analysis of GO:MF for West dataset.** Inner circle shows the coarse grained gene modules while each of the outer sector represent the protein module corresponding to each gene module. Each submodule in the protein module is highlighted by different color and three digit number indicates the functional class description of each group of proteins. Functional clusters are shown in the right, contains function id (#f), GO term followed by the depth in the GO hierarchy in brackets, number of groups (#G), and also number of proteins (#P) contained in each specific GO:MF term. First term in each cluster represents the cluster description. Detailed description of each cluster is discussed in the bottom panel. Disc, ring, and rhombus represents, GO:MF terms have already been reported in literature, newly reported GO terms and cluster component not reported in our analysis (shown with symbol '+'), respectively.

in the hierarchy are discarded, as also those terms corresponding to fewer than three proteins. This results in 525 different functional groups that participate in 145 GO:MF terms (see Table 3). Each functional group category is marked with three digit identifier as discussed, and the details are included in Additional file 5: Table S5.

As can be seen in Figure 3, namely the cluster diagram, there are 73 GO:MF terms from 13 different functional clusters. Among these, the term “Receptor binding” gives the largest cluster. There are 15 terms in panel C1 and several of these have been reported in earlier breast cancer studies. For illustrative purposes, some of these are discussed briefly below.

For instance, thioredoxin, a redox protein with growth factor activity (F2<sup>++</sup>) increases cell proliferation of breast cancer cells [46]. Many cell line models have been used to identify genetic elements that mediate the progression of breast cancer, both hormone dependent as well as hormone independent metastatic growth [47]. Integrin binding (F4) plays a crucial role in breast cancer tumor growth and metastasis [48]. EGFR binding (F10) is mediated through the binding of a mitogenic peptide epidermal growth factor (EGF) to a surface membrane receptor, EGFR of breast cancer cells [49]. Down-regulation of Interleukin-6 and its receptor (F11) results in growth inhibition of MCF-7 breast cancer cells [50], while high affinity of Insulin receptor binding (F7) has been observed in the same cells [51]. TGFβR binding (F9) regulates insulin-like growth factor binding protein (F5) (IGFBP)-3 production, which is a major antiproliferative factor and a key element for TGFβ-induced growth inhibition in breast cancer cells [52]. The breast cancer suppressor gene tyrosine kinase (PTKs) (F13) is involved in TNF-activated receptor activity by interacting with ‘TNF receptor-associated factor interacting protein’ (TRIP) in breast epithelial cells [53]. ER binding (F15) profiles are used to predict breast cancer outcome [54]. In some cases, breast cancer survivals shows elevation in the serum marker associated with proinflammatory cytokine activity (F8) [55], while more specific studies suggests that estrogen exposure decreases chemokine activity (F14), increasing the chance of developing breast cancer [56].

The present protocol also suggests that MHC class I protein binding as well as notch binding, both of which have hitherto not been studied in the context of breast cancer as more specific functional subgroups of receptor binding which are deserving of more study since both terms may provide insight into receptor binding activity during breast cancer metastasis.

Our analysis also throws up new functional clusters that should be further investigated. In C11, peptidase activity (more specifically, serine, threonine, cysteine and aspartic-type endopeptidase activity) is indicated as

being associated with breast cancer. In panel C12, the focus is on protein phosphatase regulator activity (specifically protein phosphatase type 1 and 2A regulator activity) and protein phosphatase inhibitor activity. In panel C13, GTPase regulator activity instances are clubbed together, including guanyl-nucleotide exchange and Rho guanyl-nucleotide exchange factor activity. These might provide potential candidates for target discovery and therapeutics. Other functional clusters and terms included therein can be explored for relevance, as well as for detailed insight of specific molecular functions; see Additional file 5: Table S5.

### Analysis of biological processes

GO:BP analysis for Set 1 suggests that 73.5% of all proteins in this study organize themselves in 539 different process groups that participate in 205 GO:BP terms of which only 70 have hitherto been reported in the literature related to breast cancer studies (see Table 3). Details are included in Additional file 6: Table S6.

The present protocol thus extends earlier work in suggesting novel susceptible or candidate gene targets. A representative case-study for the third module of Set 1 is shown in Figure 4. This contains the largest number of interacting proteins, namely 7712 interactions among 3187 proteins (see Table 3). Of these proteins, nearly 75% form 111 process groups that participate in 96 GO:BP terms and give 10 process clusters.

Regulation of cellular process is the largest cluster among these contains 22 predicted terms, eleven of which are known [57-67] to be involved in breast cancer and is shown in Figure 4 panel C1. These are enumerated below.

- 1) Signal transduction (P2).
- 2) Cell adhesion (P3).
- 3) Cell proliferation (P4).
- 4) Apoptosis (P6).
- 5) MAPK signaling pathway via the G-protein-coupled receptor (GPCR) (P7).
- 6) Notch signaling pathway (P8).
- 7) Estrogen induced cell proliferation (P5).
- 8) Cell death (P22).
- 9) Inhibition of epidermal growth factor receptor (EGFR) (P15) transcription.
- 10) Anti-apoptosis (P14).
- 11) Retinoblastoma cell apoptosis (P19).

In addition to these terms, cluster analysis also suggests eleven additional specific novel process terms that have not been explored so far in case of breast cancer. It includes small GTPase mediated signal transduction (P9), transmembrane receptor protein tyrosine kinase signaling pathway (P11), phosphatidylinositol-mediated

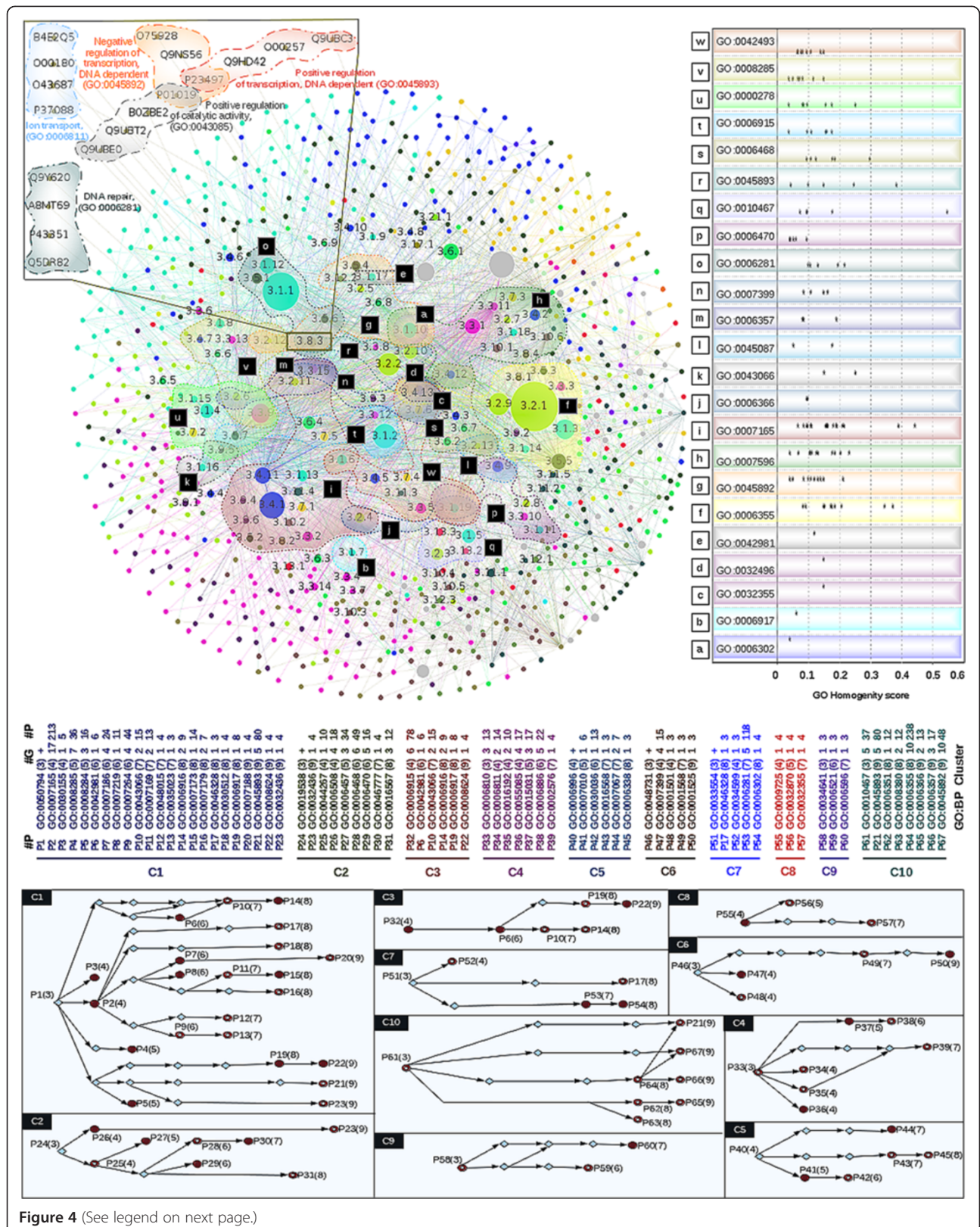


Figure 4 (See legend on next page.)

(See figure on previous page.)

**Figure 4 Network of GO:BP groups of submodular network within 3rd module of West dataset.** Each submodule is highlighted by different colors and three digit number represents the BP description of each group of proteins. Process clusters are shown in the bottom, contains process id (#d), GO terms followed by depth of the GO hierarchy in parenthesis, number of groups (#G), and number of proteins (#P). 23 GO terms common in first six module, and also those represented by gene contained in more than 2 databases, represented by dotted hull. Detail of each GO:BP term is shown in right panel. Inset represents the zoom in of a process group 3.8.3 made up of 3 overlapping and 2 independent group of proteins. Disc, ring, and rhombus represents, GO:BP terms have already been reported in literature, newly reported GO terms and cluster component not reported in our analysis (shown with symbol '+'), respectively.

signaling (P12), regulation of r-protein signal transduction (P13), transforming growth factor beta receptor signaling pathway (P16), regulation of JNK cascade (P17), negative regulation of transforming growth factor beta receptor signaling pathway (P18), G-protein signaling, coupled to cAMP nucleotide second messenger (P20), negative regulation of apoptotic process (P10), Positive regulation of transcription, DNA dependent (P21) and positive regulation of proteasomal ubiquitin-dependent protein catabolic process (P23). All these processes should be explored further to get additional insight into the regulation of cellular process in breast cancer.

A new candidate that has become evident through the present analysis is the Abraxas gene [68] that has two synonymous terms, FAM175A and CCDC98. Abraxas, which has recently been reported in breast cancer owing to its association with BRCA1 BRCT (BRCA1 C-terminal) repeats motif [69], links BRCA1 to a protein complex dedicated to ubiquitin chain recognition and hydrolysis at DNA double strand breaks, and is thus involved in BRCA1-dependent DNA damage response [70,71]. Abraxas and other member of this protein complex are required for the DNA damage checkpoints and cellular resistance to ionizing radiation (IR) in breast cancer [68]. Though CCDC98 is present in the GAD database, the synonymous term FAM175A is filtered out during gene network construction. The gene BRCA1 expresses two proteins P38398 and Q6IN79 in the primary network that interact with themselves as well as with four other proteins Q6UWZ7, P46736, Q96SD1 and Q9BX63 that also participate in double-strand break repair; these are in the secondary network. Q6UWZ7 is expressed by FAM175A: the two-layer protocol thus throws up a potential candidate gene that could be used as a novel target in the treatment of breast cancer. This also gives the motivation to explore other candidate genes included through the secondary network construction.

The present study also finds a novel cluster, one that has not so far been studied in breast cancer and that appears to be worthy of further experimental investigation. This cluster includes the following specific biological processes: positive regulation of transcription, DNA-dependent (P21), transcription, DNA-dependent (P62), regulation of transcription, DNA-dependent (P64), transcription from RNA polymerase II promoter (P65),

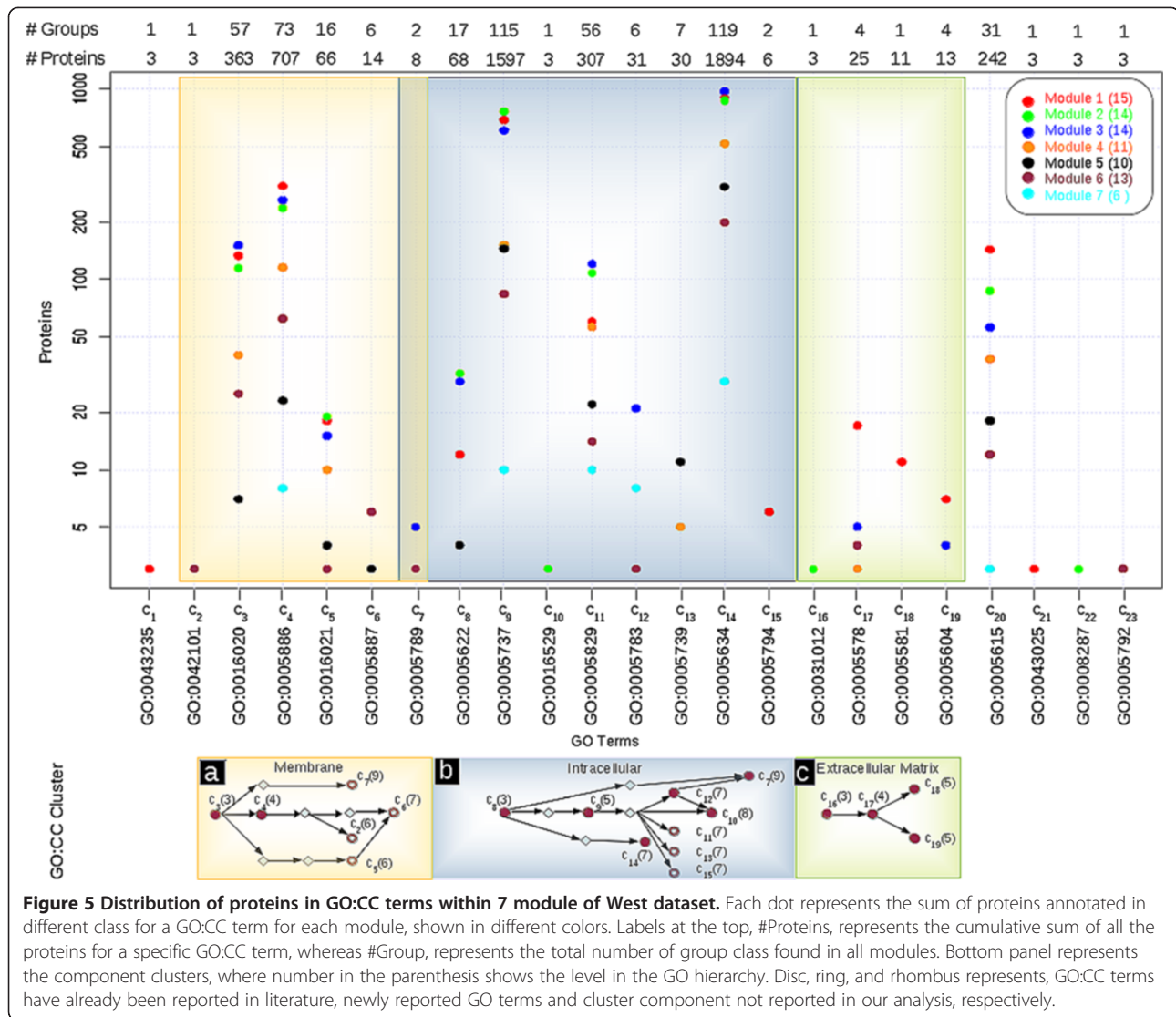
regulation of transcription from RNA polymerase II promoter (P66), negative regulation of transcription, DNA-dependent (P67) as a part of gene expression (P61) (Figure 4, panel C10).

#### Analysis of cellular components

The distributions of proteins at various cellular locations in all seven modules of Set 1 are shown in Figure 5. About 75% of the proteins organize themselves to form 434 different component submodules that are distributed across 23 different sub-cellular locations as mentioned in Table 3; the complete details are in Additional file 7: Table S7.

The largest cluster corresponds to the intracellular component (C8) (Figure 5, panel b) which appears to be the most favorable site for breast cancer progression. The BRCA1 gene that is involved in DNA repair is mainly active in the intracellular subcomponent, in the cytoplasm (C9) of breast ductal epithelial cells. Any dysfunction in BRCA1 correlates with greater risk of breast cancer symptoms [72]. Another gene BIK, an apoptotic inducer in breast cancer cell, also known as BCL2 interacting killer performs its activity primarily at another intracellular sub-location, the endoplasmic reticulum (C12) [73]. Thapsigargin (TG), a highly specific inhibitor of the endoplasmic reticulum (C12) Ca<sup>2+</sup> -ATPase pump and sarcoplasmic reticulum (C10), induced the apoptosis in the breast cancer cells as a subsequent to the secondary rise in [Ca<sup>2+</sup>] [74]. While the release of cytochrome-C from mitochondria (C8) served as an early signalling for the intrinsic pathway of apoptosis in breast cancer MCF-7c3 cells [75]. Caveolin-1 (CAV1), a highly conserved membrane-associated protein, is localized in the various subcellular locations including endoplasmic reticulum membrane (C7). Phosphorylation of CAV1 on Tyr-14 regulates paclitaxel-mediated apoptosis in MCF-7 breast cancer [76]. Also, various nuclear changes, such as enlargement, shrinkage, necrosis, vacuolation and pynknotic nuclei (C14) have already been reported in the breast cancer metastasis [77].

Cluster analysis also indicates more specific intracellular locations in the cytosol, mitochondrion, Golgi apparatus, sarcoplasmic reticulum and endoplasmic reticulum membrane as sites wherein breast cancer activity occurs. A recent study reports Caspase-2 releases from nucleus



**Figure 5 Distribution of proteins in GO:CC terms within 7 module of West dataset.** Each dot represents the sum of proteins annotated in different class for a GO:CC term for each module, shown in different colors. Labels at the top, #Proteins, represents the cumulative sum of all the proteins for a specific GO:CC term, whereas #Group, represents the total number of group class found in all modules. Bottom panel represents the component clusters, where number in the parenthesis shows the level in the GO hierarchy. Disc, ring, and rhombus represents, GO:CC terms have already been reported in literature, newly reported GO terms and cluster component not reported in our analysis, respectively.

to cytosol, which is partially required for the apoptosis induction by taxanes in breast cancer cells [78]. Similarly, other cellular component clusters can be explored for the detailed analysis, as discussed in Additional file 7: Table S7.

Our analysis of Set 2 for all three GO categories is included in Additional file 8: Table S8, Additional file 9: Table S9 and Additional file 10: Table S10. Results for both the datasets are similar, in as much as the majority of the biological activities in BP, MF, and CC are identical in both sets. This analysis thus makes it possible to

- i validate domains with relevant activity in breast cancer (as reported in literature).
- ii identify new domain terms to examine details of specific biological activity, and
- iii determine important domain clusters as the site of major biological activity.

## Discussion

In the past few years, the integrated analysis of gene expression data in context of PPI has received considerable attention. One objective has been to derive biologically interesting subnetworks or modules of interpretable size from large scale PPI data [79]. Subnetwork detection, in particular uncovering modules in the context of biology (similar to community detection in social networks) has received considerable attention [9,10]. In social networks such modules are shown to have a hierarchical structure with submodules being embedded in modules [80-82]. We have explored analogous organization principles in breast cancer data to construct the two-layer network of breast cancer genes and the associated proteins, using gene expression and PPI dataset.

There are, however, caveats. Noise in the available microarray or annotation data as well as incomplete annotations can make the network reconstruction ambiguous,

and the complexity of the underlying regulatory mechanism, in particular the combinatorics involved makes it difficult to draw direct inferences. We have considered all currently available annotations for the given set of proteins, but since many proteins are as yet unannotated, this will affect the fraction of proteins in the largest domains and thus can alter the GO-H score for any selected set of proteins. The protocol can be modified to include a predictive aspect by using the principle of guilt-by association [83] in order to discover the possible function of a protein of unknown function: if a protein has many neighbors with a particular GO classification, it is likely to belong to that particular class as well [84,85].

## Conclusion

In summary, we have constructed the two-layer “network of networks” of breast cancer genes and associated proteins. The breast cancer bilayer network has a hierarchical modular structure, with sub-modules inside modules. This structure is robust: the general topology is unaltered even when some specific proteins are removed (as seen when the LOO-CV protocol is implemented). Alternate methods that test the stability of a control network in comparison to networks formed from a random collection of genes or proteins are available, but these are not applicable here since there is extensive work on breast cancer genes and networks. Furthermore, subgroup analysis for small subgroups is unlikely to be effective [86] and randomization changes the overall topology of the control network and results in essentially no overlap with the primary network; this also reduces the probability of getting a similar prognosis among randomized groups.

The property of submodularity has advantages. Sub-modules are smaller units of biological organization that allow network abstraction at specific levels and thus helps determine domain activity through a small number of interacting submodular units rather than through the hundreds or thousands of proteins in a complex network. The different hub proteins within the submodules perform specific domain activities in a coordinated manner. We performed a complete ontological analysis to determine the GO profile that is involved in breast cancer. Some of these protein groups are involved in multiple biological activities and also form clusters of specific domain activities associated with various GO terms. This division into smaller groups helps to identify the set of proteins that participate in specific domain activities, and can be used to examine the organization and coordination of complex domain activities. We have also shown that specific submodules contain more conserved domain groups and these are assumed to be preserved during evolution [87].

The two-layer protocol allows not only the catalogue of the (well-studied) domain activities in breast cancer but also permits the prediction of more specific domain activities at a lower level of the ontological hierarchy. The analysis predicts new domain clusters which has not been studied in the context of breast cancer, and identifies Q6UWZ7, a protein expressed by a newly reported breast cancer marker gene *Abraxas* through the secondary network. In addition to identifying novel domain activities, this protocol can thus also point to new biomarkers.

Given the ever-increasing amount of interaction data available, we expect that the two-layer approach described herein will prove useful in ongoing efforts to explore the protein interaction universe and understand how domain building blocks are assembled together to perform or alter normal activity, not only in breast cancer, but in other complex diseases as well.

## Additional files

**Additional file 1: Table S1.** Gene extracted from different data sources.

**Additional file 2: Table S2.** Gene network analysis for Set 1.

**Additional file 3: Table S3.** Gene network analysis for Set 2.

**Additional file 4: Table S4.** Summary of topological and biological analysis of gene module and their corresponding protein module for Set 2.

**Additional file 5: Table S5.** Molecular function analysis of modular network of Set 1.

**Additional file 6: Table S6.** Biological process analysis of modular network of Set 1.

**Additional file 7: Table S7.** Cellular component analysis of modular network of Set 1.

**Additional file 8: Table S8.** Molecular function analysis of modular network of Set 2.

**Additional file 9: Table S9.** Biological process analysis of modular network of Set 2.

**Additional file 10: Table S10.** Cellular component analysis of modular network of Set 2.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

Conceived and design the protocol: RR, AS. Generate the results: AS, SK. Analyze the results: RR, AS, SK. Wrote the manuscript: RR, AS. Read and Approved: RR, AS, SK. All authors read and approved the final manuscript.

## Acknowledgments

AS and SK would like to acknowledge Department of Science and Technology, Centre for Mathematical Sciences for their financial support (DST-CMS project Lr.No.SR/S4/MS:516/07, Dt.21-04-2008).

## Author details

<sup>1</sup>C R RAO Advanced Institute of Mathematics, Statistics and Computer Science, University of Hyderabad Campus, Hyderabad 500046, India. <sup>2</sup>Indian Institute of Technology Bhubaneswar, Bhubaneswar 751013, India. <sup>3</sup>University of Hyderabad, Hyderabad 500046, India. <sup>4</sup>School of Integrative and Computational Sciences, Jawaharlal Nehru University, New Delhi 110 067, India.

Received: 3 March 2014 Accepted: 26 June 2014  
Published: 5 July 2014

## References

1. Ditttrich M, Klau G, Rosenwald A, Dandekar T, Müller T: **Identifying functional modules in protein-protein interaction networks: an integrated exact approach.** *Bioinformatics* 2008, **24**:223–231.
2. Ideker T, Sharan R: **Protein networks in disease.** *Genome Res* 2008, **18**:644–652.
3. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database. sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Res* 2004, **32**:D262–D266.
4. Palla G, Derényi I, Farkas I, Vicsek T: **Uncovering the overlapping community structure of complex networks in nature and society.** *Nature* 2005, **435**:814–818.
5. Traud AL, Kelsic ED, Mucha PJ, Porter MA: **Community structure in online collegiate social networks.** In *Proceedings of the 2009 Pittsburgh, PA: APS March Meeting*; 2009:16–20.
6. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E: **Fast unfolding of communities in large networks.** *J Stat Mech* 2008, **10**:P100–108.
7. Newman MEJ: **The structure of scientific collaboration networks.** *Proc Natl Acad Sci* 2001, **98**(2):404–9.
8. Rosvall M, Bergstrom CT: **Maps of random walks on complex networks reveal community structure.** *Proc Natl Acad Sci* 2008, **105**:1118–1123.
9. Wilkinson DM, Huberman BA: **A method for finding communities of related genes.** *Proc Natl Acad Sci* 2004, **101**:5241–5248.
10. Santo Fortunato: **Community detection in graphs.** *Phys Rep* 2010, **486**:75–174.
11. Girvan M, Newman MEJ: **Community structure in social and biological networks.** *Proc Natl Acad Sci* 2002, **99**(12):7821–6.
12. Yook SH, Oltvai ZN, Barabási AL: **Functional and topological characterization of protein interaction networks.** *Proteomics* 2004, **4**(4):928–942.
13. Rives AW, Galitski T: **Modular organization of cellular networks.** *Proc Natl Acad Sci* 2003, **100**(3):1128–1133.
14. Pieroni E, De la fuente van Bentem S, Mancosu G, Capobianco E, Hirt H, de la Fuente A: **Protein networking. insights into global functional organization of proteomes.** *Proteomics* 2008, **8**(4):799–816.
15. Pereira-Leal JB, Enright AJ, Ouzounis CA: **Detection of functional modules from protein interaction networks.** *Proteins* 2004, **54**:49–57.
16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology. tool for the unification of biology. The Gene Ontology Consortium.** *Nat Gene* 2000, **25**(1).
17. Schäfer J, Strimmer K: **A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.** *Stat Appl Genet Mol Biol* 2005, **4**(1):32.
18. Clauset A, Newman MEJ, Moore C: **Finding community structure in very large networks.** *Phys Rev E* 2004, **70**:066111.
19. Newman MEJ, Girvan M: **Finding and evaluating community structure in networks.** *Phys Rev E* 2004, **69**:026113.
20. Raghavan UN, Albert R, Kumara S: **Near linear time algorithm to detect community structures in large-scale networks.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2007, **76**:036106.
21. Newman MEJ: **Finding community structure in networks using the eigenvectors of matrices.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2006, **74**:036104.
22. Reichardt J, Bornholdt S: **Statistical Mechanics of Community Detection.** *Phys Rev E* 2006, **74**:016110.
23. Latapy M, Pons P: **Computing communities in large networks using random walks.** *Proc Lect Notes Comp Sci* 2005, **3733**:285–293.
24. Porter MA, Onnella JP, Mucha PJ: **Communities in networks.** *Not Am Math Soc* 2009, **56**(9):1082–1097.
25. Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network.** *Nature* 2004, **430**:88–93.
26. Fraser HB: **Modularity and evolutionary constraint on proteins.** *Nat Genet* 2005, **37**(4):351–2.
27. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M: **The human disease network.** *Proc Natl Acad Sci* 2007, **104**:8685–8690.
28. Srivastava AK, Sahni NS: **OntoVisT. A general purpose Ontological Visualization Tool.** *Bioinformatics* 2011, **6**(7):288–290.
29. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S, Rashmi BP, Shanker K, Padma N, Niranjana V, Harsha HC, Talreja N, Vrushabhendra BM, Ramya MA, Yatish AJ, Joy M, Shivashankar HN, Kavitha MP, Menezes M, Choudhury DR, Ghosh N, Saravana R, Chandran S, Mohan S, Jonnalagadda CK, Prasad CK, et al: **Human protein reference database as a discovery resource for proteomics.** *Nucleic Acids Res* 2004, **32**(Database issue):D497–501.
30. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles.** *Proc Natl Acad Sci* 2001, **98**:11462–11467.
31. Gyorffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, Szallasi Z: **An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients.** *Breast Cancer Res Treat* 2010, **123**:725–731.
32. Syed AS, D'Antonio M, Ciccarelli FD: **Network of Cancer Genes. a web resource to analyze duplicability, orthology and network properties of cancer genes.** *Nucleic Acids Res* 2010, **38**(Database issue):D670–5.
33. **Tumor Gene Family of Databases.** <http://www.tumor-gene.org/tgdf.html>.
34. **Cancer GeneticsWeb.** [<http://www.cancerindex.org/geneweb/>]
35. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4**(3):177–83.
36. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H: **KEGG Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**(1):29–34.
37. Baasiri RA, Glasser SR, Steffen DL, Wheeler DA: **The breast cancer gene database a collaborative information resource.** *Oncogene* 1999, **18**(56):7958–65.
38. Strausberg RL: **The Cancer Genome Anatomy Project. new resources for reading the molecular signatures of cancer.** *J Pathol* 2001, **195**(1):31–40.
39. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33**(Database issue):D514–7.
40. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database.** *Nat Genet* 2004, **36**:431–432.
41. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS: **UniProt. The Universal Protein knowledgebase.** *Nucleic Acids Res* 2004, **32**:D115–D119.
42. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altmann RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17**:520–525.
43. Ahn J, Yoon Y, Park C, Shin E, Park S: **Integrative gene network construction for predicting a set of complementary prostate cancer genes.** *Bioinformatics* 2011, **27**:1846–1853.
44. Ledoit O, Wolf M: **Improved estimation of the covariance matrix of stock returns with an application to portfolio selection.** *Finance: J Empir*; 2003:10.603–621.
45. Csardi G, Nepusz T: **The igraph Software Package for Complex Network Research.** *Inter J Complex Syst* 2006, **1965**:1–9.
46. Gallegos A, Gasdaska JR, Taylor CW, Paine-Murrieta GD, Goodman D, Gasdaska PY, Berggren M, Briehl MM, Powis G: **ransfection with human thioredoxin increases cell proliferation and a dominant-negative mutant thioredoxin reverses the transformed phenotype of human breast cancer cells.** *Cancer Res* 1996, **56**:5765–5770.
47. Nakshatri H, Bhat-Nakshatri P, Martin DA, Goulet RJJ, Sledge GWJ: **Constitutive activation of NF-kappaB during progression of breast cancer to hormone-independent growth.** *Mol Cell Biol* 2007, **17**:3629–3639.
48. Khalili P, Arakelian A, Chen G, Plunkett ML, Beck I, Parry GC, Donate F, Shaw DE, Mazar AP, Rabbani SA: **A non-RGD-based integrin binding peptide (ATN-161) blocks breast cancer growth and metastasis in vivo.** *Mol Cancer Ther* 2006, **5**:2271–2280.
49. Depenbrock H, Shirvani A, Rastetter J, Hanauske AR: **Effects of vinorelbine on epidermal growth factor-receptor binding of human breast cancer cell lines in vitro.** *Springer* 1995, **3**:187–193.
50. Jiang XP, Yang DC, Elliott RL, Head JF: **Down-regulation of expression of interleukin-6 and its receptor results in growth inhibition of MCF-7 breast cancer cells.** *Anticancer Res* 2011, **31**(9):2899–906.

51. Milazzo G, Yip CC, Maddux BA, Vigneri R, Goldfine ID: **High-affinity insulin binding to an atypical insulin-like growth factor-I receptor in human breast cancer cells.** *J Clin Invest* 1992, **89**(3):899–908.
52. Oh Y, Muller HL, Ng L, Rosenfeld RG: **Transforming growth factor- $\beta$ -induced cell growth inhibition in human breast cancer cells is mediated through insulin-like growth factor-binding protein-3 action.** *J Biol Chem* 1995, **270**:13589–13592.
53. Zhou Q, Geahlen RL: **The protein-tyrosine kinase Syk interacts with TRAF-interacting protein TRIP in breast epithelial cells.** *Oncogene* 2009, **28**:1348–1356.
54. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR, Ali S, Chin SF, Palmieri C, Caldas C, Carroll JS: **Differential oestrogen receptor binding is associated with clinical outcome in breast cancer.** *Nature* 2012, **481**:389–393.
55. Bower JE, Ganz PA, Aziz N, Fahey JL: **Fatigue and Proinflammatory Cytokine Activity in Breast Cancer Survivors.** *Psychosomatic Med* 2002, **64**:604–611.
56. Janis K, Hoeltke J, Nazareth M, Fanti P, Poppenberg K, Aronica SM: **Estrogen decreases expression of chemokine receptors, and suppresses chemokine bioactivity in murine monocytes.** *Am J Reprod Immunol* 2004, **51**:22–31.
57. Strobl JS, Wonderlin WF, Flynn DC: **Mitogenic signal transduction in human breast cancer cells.** *Gen Pharmacol* 1995, **26**:1643–1649.
58. Oka H, Shiozaki H, Kobayashi K, Inoue M, Tahara H, Kobayashi T, Takatsuka Y, Matsuyoshi N, Hirano S, Takeichi M, Mori T: **Expression of E-cadherin cell adhesion molecules in human breast cancer tissues and its relationship to metastasis.** *Cancer Res* 1993, **53**:1696–1701.
59. Yang L, Wei L, Zhao W, Wang X, Zheng G, Zheng M, Song X, Zuo W: **Down-regulation of osteopontin expression by RNA interference affects cell proliferation and chemotherapy sensitivity of breast cancer MDA-MB-231 cells.** *Mol Med Report* 2012, **5**(2):373–6.
60. Formby B, Wiley TS: **Progesterone inhibits growth and induces apoptosis in breast cancer cells. inverse effects on Bcl-2 and p53.** *Ann Clin Lab Sci* 1998, **28**:360–369.
61. Filardo EJ: **Epidermal growth factor receptor (EGFR) transactivation by estrogen via the G-protein-coupled receptor, GPR30. a novel signaling pathway with potential significance for breast cancer.** *J Steroid Biochem Mol Biol* 2002, **80**:231–238.
62. Al-Hussaini H, Subramanyam D, Reedijk M, Sridhar SS: **Notch signaling pathway as a therapeutic target in breast cancer.** *Mol Cancer Ther* 2011, **10**(1):9–15.
63. Tan H, Zhong Y, Pan Z: **Autocrine regulation of cell proliferation by estrogen receptor-alpha in estrogen receptor-alpha-positive breast cancer cell lines.** *BMC Cancer* 2009, **9**:31–42.
64. Perks CM, Newcomb PV, Norman MR, Holly JM: **Effect of insulin-like growth factor binding protein-1 on integrin signaling and the induction of apoptosis in human breast cancer cells.** *J Mol Endocrinol* 1999, **22**:141–150.
65. Jin W, Chen BB, Li JY, Zhu H, Huang M, Gu SM, Wang QQ, Chen JY, Yu S, Wu J, Shao ZM: **TIEG1 inhibits breast cancer invasion and metastasis by inhibition of epidermal growth factor receptor (EGFR) transcription and the EGFR signaling pathway.** *Mol Cell Biol* 2012, **32**(1):50–63.
66. Razandi M, Pedram A, Levin ER: **Plasma membrane estrogen receptors signal to antiapoptosis in breast cancer.** *Mol Endocrinol* 2000, **14**(9):1434–47.
67. Hsiao WT, Tsai MD, Jow GM, Tien LT, Lee YJ: **Involvement of Smac, p53, and caspase pathways in induction of apoptosis by gossypol in human retinoblastoma cells.** *Mol Vis* 2012, **18**:2033–2042.
68. Solyom S, Aressy B, Pylkäs K, Patterson-Fortin J, Hartikainen JM, Kallioniemi A, Kauppila S, Nikkilä J, Kosma VM, Mannermaa A, Greenberg RA, Winqvist R: **Breast Cancer-Associated Abraxas Mutation Disrupts Nuclear Localization and DNA Damage Response Functions.** *Sci Transl Med* 2012, **4**(122):122ra23.
69. Wang B, Matsuoka S, Ballif BA, Zhang D, Smogorzewska A: **Abraxas and RAP80 form a BRCA1 protein complex required for the DNA damage response.** *Science* 2007, **316**:1194–1198.
70. Sobhian B, Shao G, Lilli DR, Culhane AC, Moreau LA: **RAP80 targets BRCA1 to specific ubiquitin structures at DNA damage sites.** *Science* 2007, **316**:1198–1202.
71. Shao G, Patterson-Fortin J, Messick TE, Feng D, Shanbhag N: **MERIT40 controls BRCA1–Rap80 complex integrity and recruitment to DNA double-strand breaks.** *Genes Dev* 2009, **23**(6):740–754.
72. Chen Y, Chen C-F, Riley DJ, Allred DC, Chen P-L, Von Hoff D, Osborne CK, Lee WH: **Aberrant Subcellular Localization of BRCA1 in Breast Cancer.** *Science* 1995, **270**:789–791.
73. Fu Y, Li J, Lee AS: **GRP78/BIP Inhibits Endoplasmic Reticulum BIK and Protects Human Breast Cancer Cells against Estrogen Starvation-Induced Apoptosis.** *Cancer Res* 2007, **67**(8):3734–40.
74. Jackisch C, Hahm HA, Tombal B, McCloskey D, Butash K, Davidson NE, Denmeade SR: **Delayed micromolar elevation in intracellular calcium precedes induction of apoptosis in thapsigargin-treated breast cancer cells.** *Clin Cancer Res* 2000, **6**(7):2844–50.
75. Chiu SM, Xue LY, Usuda J, Azizuddin K, Oleinick NL: **Bax is essential for mitochondrion-mediated apoptosis but not for cell death caused by photodynamic therapy.** *Br J Cancer* 2003, **89**(8):1590–7.
76. Shajahan AN, Wang A, Decker M, Minshall RD, Liu MC, Clarke R: **Caveolin-1 tyrosine phosphorylation enhances paclitaxel-mediated cytotoxicity.** *J Biol Chem* 2007, **282**(8):5934–43.
77. Sethi D, Sen R, Parshad S, Khetarpal S, Garg M, Sen J: **Histopathologic changes following neoadjuvant chemotherapy in various malignancies.** *Int J App Basic Med Res* 2012, **2**:111–116.
78. Jelínek M, Balu Íková K, Kopperová D, McOva-Furstova V, Ramek J, Fidlerova J, Zanardi I, Ojima I, Kova J: **Caspase-2 is involved in cell death induction by taxanes in breast cancer cells.** *Cancer Cell Int* 2013, **13**(1):42.
79. Wagner A, Wright J: **Alternative routes and mutational robustness in complex regulatory networks.** *Biosystems* 2007, **88**(1–2):163–72.
80. Arenas A, D’az-Guilera A, Perez-Vicente CJ: **Synchronization reveals topological scales in complex networks.** *Phys Rev Lett* 2006, **96**(11):114102.
81. Lancichinetti A, Fortunato S, Kertesz J: **Detecting the overlapping and hierarchical community structure of complex networks.** *New J Phys* 2009, **11**:033015.
82. Shuqin Z: **Hierarchical Modular Structure in Gene Coexpression Networks.** China: IEEE 6th International Conference on Systems Biology (ISB); 2012:118–124.
83. Oliver S: **Guilt-by-association goes global.** *Nature* 2000, **403**:601–603.
84. Schwikowski B, Uetz P, Fields S: **A network of protein protein interactions in yeast.** *Nat Biotechnol* 2000, **18**:1257–1261.
85. Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T: **Assessment of prediction accuracy of protein function from protein–protein interaction data.** *Yeast* 2001, **18**:523–531.
86. Kaiser LD: **Stratification of randomization is not required for a pre-specified subgroup analysis.** *Pharm Stat* 2013, **12**(1):43–7.
87. Vespignani A: **Evolution thinks modular.** *Nat Genet* 2003, **35**:118–119.

doi:10.1186/1752-0509-8-81

Cite this article as: Srivastava et al.: Two-layer modular analysis of gene and protein networks in breast cancer. *BMC Systems Biology* 2014 **8**:81.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

