




Modular organization of gene–tumor association network allows identification of key molecular players in cancer

THARMARAJ JESAN^{1,2,3} and SITABHRA SINHA^{1,2*} 

¹The Institute of Mathematical Sciences, CIT Campus, Taramani, Chennai 600 113, India

²Homi Bhabha National Institute, Anushaktinagar, Mumbai 400 094, India

³Health Physics Division, Bhabha Atomic Research Centre, Kalpakkam 603 102, India

*Corresponding author (Email, sitabhra@imsc.res.in)

MS received 18 October 2021; accepted 31 May 2022

The role played by the topological structure of biological networks in their dynamics and function is receiving increasing attention over the last decade as large-throughput experiments have provided large volumes of highly resolved data on the interactions between the components of such networks. This has provided new perspectives on systems diseases: for example, there has been a gradual shift in cancer research away from the study of individual molecules and of single gene mutations to the emerging consensus that it is a complex disease involving large-scale disruptions in the intracellular signaling network. One of the drawbacks of a systems- or network-based approach is the large number of cellular agents whose interactions need to be investigated. We tried to solve this problem by taking a mesoscopic view of the cancer diseases–genes network, whose modular organization we studied after projecting it onto two networks, one comprising only disease types and the other consisting of only genes related to one or more categories of cancer. Using community partitioning, we identified several modules in these networks. Projecting cancer gene clusters onto an abstract ‘modular space’ allows us to infer the relations between different tumor types. By classifying the functional role of particular genes in terms of their inter- and intra-modular connectivity, we identified a number of genes that play the key role of ‘connector hubs’ in the network. Using data from the human protein–protein interaction network we showed that genes that are ‘connector hubs’ or ‘global hubs’ are, in fact, much more likely to be related to cancer than other genes. More important from a therapeutic point of view, we showed that the connector hubs in the cancer gene network are involved in a significantly larger number of human signaling pathways associated with cancer than other types of cancer genes. Furthermore, the types of cancer linked to connector hub genes have significantly reduced survival rates compared with other types of cancer, thereby enhancing their importance in the search for potential therapeutic targets.

Keywords. Cancer; community structure; gene–tumor association; mesoscopic organization; network modules

1. Introduction

Cancer is the collective name of a group of diseases characterized by unconstrained cell growth which has significant mortality and high public health cost. In

developed countries where life expectancy has increased and ‘diseases of poverty’ such as tuberculosis have largely been controlled, cancer is one of the leading causes of death (WHO 2021). Despite years of sustained research efforts, cancer is still untamed (for a highly acclaimed popular account of the medical campaign to cure cancer, see Mukherjee 2010). This is at least in part because cancer is a ‘systems disease’

This article is part of the Topical Collection: Emergent dynamics of biological networks.

Supplementary Information: The online version contains supplementary material available at <https://doi.org/10.1007/s12038-022-00292-5>.

(Hornberg *et al.* 2006), i.e., it cannot be completely understood without considering the network of interactions between a number of different elements (Kreeger and Lauffenburger 2010). It is therefore unlikely that it can be cured by treating a single cause.

The difficulty of investigating cancer as a network disease is in the large number of elements that are involved and the myriad ways in which they interact. A possible approach to this complex disease is to compartmentalize the entire system of interactions into sub-networks that are easier to analyze by exploiting the modular nature of biological networks (Hartwell *et al.* 1999; Cloutier and Wang 2011). By focusing on the modules of networks related to cancer and the interactions between them, it is possible to use a mesoscopic approach for understanding the systems-level aspect of cancer without getting mired in the complexity of the large number of molecules and interactions involved.

In this article we have reconstructed a cancer gene network and a cancer category/tumor type network using a comprehensive database relating different categories of cancers and types of tumors with mutations of specific genes. This was done by taking projections from the bipartite network that connect the nodes representing genes with the nodes representing the types of tumors in which mutations of those genes have been implicated. Using community detection algorithms, we identified several modules in these networks. By classifying genes in terms of their connectivity to members of their own module and to members of different modules, we identified their functional importance in the cancer network (see supplementary figure 1 for a graphic summary of the methodology employed in this study). We show that genes playing the role of connector hubs and global hubs, i.e., having high connectivity with other modules in addition to genes belonging to their own module, have a disproportionately high representation in the human signaling pathways related to cancer. Therefore, these genes can be identified as potential targets for therapeutic intervention. The importance of connector and global hubs is further underlined by observing that nodes having these roles in the protein–protein interaction network have an extremely high probability of being related to cancer compared with the corresponding probability of a randomly chosen protein. Finally, we showed that genes that are connector hubs are associated with diseases that have a much lower survival rate than others, pointing to the critical positions they occupy in the cancer network.

2. Materials and methods

2.1 Connectivity data

For constructing the networks analyzed here we have used information (supplementary table 1) on the association of 927 cancer-related genes (supplementary table 2) with 35 different cancer categories (supplementary table 3) and 135 tumor types (supplementary table 4), obtained from the F-Census or Functional Census database of human cancer genes (Gong *et al.* 2010). These are derived from high-throughput mutational screens of cancer genomes collected from various sources including the Cancer Gene Census (CGC) (Futreal *et al.* 2006) and Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.* 2005). After pre-processing the associations to remove ambiguous entries, we obtained a list of 4066 links between tumor types and genes (supplementary table 5), which were used to construct (via projection, see Results) the network of cancer genes.

For constructing the protein–protein interaction network we used the data for interactions between different proteins obtained from the Human Protein Reference Database (HPRD) (available at <http://www.hprd.org/>). The data set we considered (supplementary table 6) comprised 9645 proteins whose interconnections have been identified using yeast two-hybrid analysis and *in vitro* or *in vivo* methods (Prasad *et al.* 2009).

2.2 Network randomization

Ensembles of randomized versions of the empirical networks were constructed via the procedure of degree-preserved randomization of links (Milo *et al.* 2002) by exchanging the terminal nodes of 10^6 randomly chosen pairs of links in the network. When applied to the weighted network of cancer-related genes obtained by projecting the bipartite network of genes and tumor types (TT-GWN, see Results), the empirical link weights associated with the connections were preserved. An alternative procedure was also carried out, in which the links in the empirical bipartite network itself were subjected to 10^6 random swaps. The link weights in the resulting randomized network are different from those in the empirical network. However, comparison of the empirical network with these two ensembles of randomized networks yields qualitatively similar differences.

2.3 Community detection

Modular organization of connections is one of the most prominent mesoscopic structural properties observed in many biological, social, and technological networks (Porter *et al.* 2009). Modules (also referred to as communities) are subnetworks characterized by relatively high connection density within their constituent nodes with comparatively sparser connections between nodes that are members of different modules (Newman 2004). Possibly the most widely used theoretical framework for identifying the modules of a network involves using different methods to maximize a quantitative measure of modularity, Q , defined for a given modular partitioning of the network as

$$Q = (1/L) \sum_{ij} B_{ij} \delta(c_i, c_j),$$

where $B_{ij} = A_{ij} - (k_i^{\text{in}} k_j^{\text{out}}/L)$ are the constituent elements of the modularity matrix \mathbf{B} (Clauset *et al.* 2004; Newman and Girvan 2004; Newman 2006). The connection topology is described by the adjacency matrix \mathbf{A} whose elements $A_{ij} = 1$ if a (directed) link connects node j to node i , and $A_{ij} = 0$, otherwise. The in-degree (number of incoming connections) and out-degree (number of outgoing connections) of a node i are represented by $k_i^{\text{in}} = \sum_j A_{ij}$ and $k_j^{\text{out}} = \sum_i A_{ji}$, respectively. The total number of connections in the network is indicated by $L (= \sum_i k_i^{\text{in}} = \sum_i k_j^{\text{out}})$. The Kronecker delta function $\delta(c_i, c_j) = 1 (=0)$ if the communities c_i, c_j to which the nodes i, j belong, respectively, are the same (are different).

2.3.1 Spectral method for optimal partitioning of a network into modules: One of the methods in wide use for achieving optimal partitioning of a network into modules by maximizing Q is the spectral method (Newman 2006). The network is first bisected by obtaining the eigenvector for the largest positive eigenvalue of the symmetrized modularity matrix $\mathbf{B} + \mathbf{B}^T$ and assigning each node to one of two communities depending on the sign of the corresponding eigenvector component. This partitioning is subsequently improved by exchanging the members of the two communities so as to achieve the highest value of Q . By carrying out the process recursively on each of the resulting communities until Q can no longer be increased, the method converges to an optimal partitioning of the network.

2.3.2 Robustness of partitioning: To ensure that the optimal partitioning of a network is not sensitively dependent on the specific method used to maximize Q , a stochastic simulated annealing approach (Good *et al.* 2010) was employed to obtain an ensemble of 350 optimal partitions. If there is lack of similarity between these different partitions (obtained from distinct realizations of the annealing algorithm), this will suggest a high level of degeneracy (and hence, ambiguity) in the identification of the modular composition of the network. Each simulated annealing realization first partitions the network into arbitrary communities and then iteratively changes it by any of the following possible moves (chosen with equal probability) at each step: (i) transfers a randomly chosen node to another module, possibly one that is newly created, (ii) joins a pair of modules that are randomly chosen, and (iii) bisects a module (that is randomly selected) so as to minimize the number of connections between these two parts. At each step, the new partition that results from the move selected is always accepted if it has a higher Q than the immediately preceding partition. If, on the other hand, it has a lower Q , it is accepted with a Boltzmann-like probability factor $\exp(-|\Delta Q|/T)$, where ΔQ is the change in Q and T is equivalent to temperature. The annealing involves decreasing T over successive iterations according to a pre-specified cooling schedule. The process halts when the number of unsuccessful attempts at changing the partitioning exceeds a threshold value. We observe that the Q values follow a bimodal distribution with one peak at 0 and the other with a peak close to the value of $Q \approx 0.42$ obtained using the Infomap method (see below). Note that we performed 100 different realizations of degree-preserved randomization on the gene–disease association bipartite network and obtained corresponding ‘randomized surrogate’ gene networks as a control that yielded mean Q of 0.315 ± 0.059 , the values for individual realizations ranging between a maximum Q of 0.371 and a minimum Q of 0.107. This suggests that the value of Q obtained for TT-GWN is extremely unlikely to have been obtained by chance in the absence of any modular organization. We focused on the 147 partitions whose Q belonged to the higher peak of the distribution and verified that the module membership of the large majority of nodes remains invariant across all these partitions, which underscores the robustness of the modular decomposition.

2.3.3 Consensus clustering: To quantitatively express the consistency between different modular partitionings obtained using simulated annealing, we computed a consensus matrix \mathbf{P} (Lancichinetti and Fortunato 2012). The matrix elements P_{ij} denote the fraction of realizations (i.e., of the different partitionings obtained that have Q values belonging to the higher peak of the bimodal distribution as described above) in which a pair of nodes i, j occurs in the same module. When the two nodes always occur in the same module, $P_{ij} = 1$, and $P_{ij} = 0$ when they never appear in the same module. In either case, this suggests a high degree of consistency, so that if \mathbf{P} comprises only 0s and 1s, it suggests the maximum degree of consistency across partitioning realizations.

2.3.4 Infomap method: While the bulk of the methods currently used to identify modular organization of a network involves some variant of maximizing the modularity measure Q , one of the exceptions is the Infomap method. Based on its performance in several benchmark tests, the Infomap method has emerged as one of the most efficient algorithms for partitioning a network into communities (Lancichinetti and Fortunato 2009; Fortunato 2010). The basic principle of the Infomap algorithm is that optimal compression of network topology uses the regularities in network structure, in particular, the occurrence of modules (Rosvall and Bergstrom 2008, 2010). An implementation of the code is available at <https://www.mapequation.org/infomap/>.

2.4 Modular spectra

We analyzed the relation between different tumor types (and cancer categories) by using a decomposition in terms of the overlap of their associated genes with the different modules of the gene network. Let the set of all genes be optimally partitioned into M modules. We then define an overlap matrix \mathbf{O} , whose rows correspond to the different groups of genes associated with specific tumor types or cancer categories, and the columns correspond to the different modules of the cancer gene network. An element of this overlap matrix O_{ij} is the number of genes in group i that are from the module j . Thus, the decomposition of the i -th group in the abstract M -dimensional basis space formed by the modules is $\{O_{i1}/N_i, O_{i2}/N_i, \dots, O_{iM}/N_i\}$, where $N_i = \sum_{k=1, \dots, M} O_{ik}$ is the total number of genes in the i -th group.

The distance between two groups i and j in this ‘modular’ space is defined as

$$d_{i,j}^{\text{modular}} = \sqrt{\sum_k \left[\frac{O_{ik}}{N_i} - \frac{O_{jk}}{N_j} \right]^2}$$

This measure can be used as a metric for closeness or proximity between different tumor types or cancer categories. For visualization of the relation between different groups, a dendrogram was constructed, where the ordinate represents the closeness between a pair of groups.

2.5 Determining the intra- and inter-modular role of a gene

The role played by each gene in terms of its connectivity within its own module and in the entire network is determined by two properties (Guimera *et al.* 2007): (i) the relative within module degree, z , and (ii) the participation coefficient, P . The within-community degree z -score measures how well connected a node i is to other nodes in the community to which it belongs, i.e., it distinguishes nodes that are hubs of their communities from those that are non-hubs. It is defined as

$$z_i = \frac{k_{c_i}^i - \langle k_{c_i}^j \rangle_{j \in c_i}}{\sqrt{\langle (k_{c_i}^j)^2 \rangle_{j \in c_i} - \langle k_{c_i}^j \rangle_{j \in c_i}^2}}$$

where $k_{c_i}^i$ is the number of links of node i to other nodes in its community c and $\langle \dots \rangle_{j \in c}$ is taken over all nodes in module c .

The nodes are also distinguished based on their connectivity profile over the entire network, in particular, their connections to nodes in other communities. Two nodes with same within module degree z -score can play different roles if one of them has significantly higher inter-modular connections compared with the other. This is measured by the participation coefficient P_i of node i , defined as

$$P_i = 1 - \sum_{c=1}^m \left(\frac{k_c^i}{k_i} \right)^2$$

where M is the total number of communities, k_c^i is the number of links from node i to other nodes in its community c and $k_i = \sum_c k_c^i$ is the total degree of node i . The participation coefficient of a node is close to 1 if its links are uniformly distributed among all the

communities. On the other hand, it is 0 if all links of a node are with members of its own community.

3. Results

3.1 Modular structure of the network of cancer-related genes

We first constructed a bipartite network consisting of two types of nodes, viz., cancer-related genes (G) and tumor types (TT) [alternatively, we also use cancer categories (CC)]. Nodes of different types were connected based on the association between genes and tumor types (or cancer categories) related to them according to the information obtained from the F-Census database (Gong *et al.* 2010). From this bipartite network, we produced a tumor type–cancer gene network (TT-GN) and tumor-type network (TTN)

by using the method of projections (figure 1a). According to this technique, from a bipartite network consisting of two categories of nodes, Type I and Type II, respectively, one can construct two networks, one comprising only Type I nodes (obtained by connecting any pair of Type I nodes that share as a common neighbor a Type II node in the bipartite network) and the other comprising only Type II nodes (connecting pairs of Type II nodes that share a common Type I node as neighbor) (Goh *et al.* 2007). In the *tumor type–cancer gene network* (TT-GN), the nodes represent different cancer-related genes. Two genes are connected to each other if they have at least one tumor-type that they are associated with in common. In the *tumor type–cancer gene weighted network* (TT-GWN), the links are weighted in proportion to the number of common tumor types associated with any pair of connected genes. In the *tumor-type network* (TTN), the nodes represent tumor types and two tumor types are

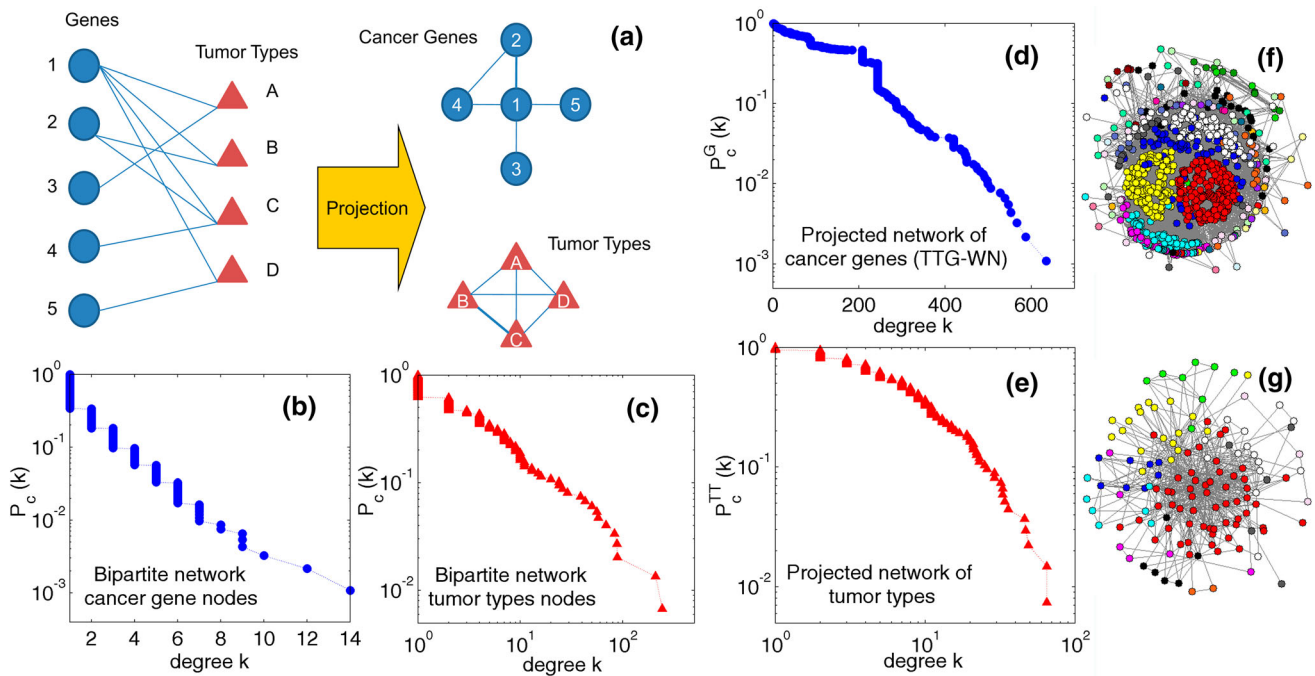


Figure 1. Networks of cancer genes and tumor types. (a) Schematic diagram showing the bipartite network comprising genes (represented by circles) and tumor types (triangles). A gene is connected to a tumor type if mutations in the gene result in a tumor of that specific type. The *tumor type–cancer gene network* (TT-GN) is obtained from a projection of the bipartite network, where two genes are connected if there is a tumor type that can be related to mutations in either of the two genes. In the *tumor type–cancer gene weighted network* (TT-GWN), a link between two cancer-related genes (e.g., 1 and 2) in the projected network is weighted by the number of tumor types (B, C, etc.) to which both nodes (viz., 1 and 2) are connected in the bipartite network. The other possible projection yields the tumor-type network (TTN), where two tumor types are connected if either can result from mutations in the same gene. Each link can be weighted in proportion to the number of common genes with which the two tumor types are associated. (b–d) The cumulative degree distribution $P_c(k)$, i.e., the probability that a node will have k or more links is shown for the (b) genes and (c) tumor types of the bipartite network, as well as for the two networks obtained by projection, viz., (d) the network of cancer genes and (e) the network of tumor types. (f–g) Pictorial representation of (f) the cancer gene network comprising 910 nodes and (g) the network of 135 tumor types.

connected if there is at least one common element in the set of genes that each is related to. The weight associated with a link is proportional to the number of genes that appear in common for both tumor types. The cumulative degree distribution $P_c(k)$ for cancer genes in the bipartite network shows a rapidly decreasing exponential nature (figure 1b), while that of the nodes corresponding to tumor types decays more slowly, resembling a power law (as indicated by the approximately linear nature in double logarithmic scale; figure 1c). However, the projected networks of cancer genes and tumor types both have rapidly decaying tails in the cumulative degree distribution (figure 1d–e). The representation of the two projected networks (figure 1f–g) appears to suggest that they both have a

densely connected core surrounded by a periphery of sparsely connected sets of nodes.

We analyzed the mesoscopic organization of TT-GWN by first identifying its modular arrangement using the Infomap method (see Methods). Figure 2a shows the clustering of the network into 25 communities (with the corresponding value of $0.42 \approx Q$) using this method, suggesting that the network has a strong modular organization. The modules are of heterogeneous sizes, the largest having 246 genes and the smallest has only 1 (we explicitly verified that non-inclusion of this single-gene module in further analysis did not affect our results). We also carried out a modular decomposition of the largest connected component of the human protein–protein interaction network

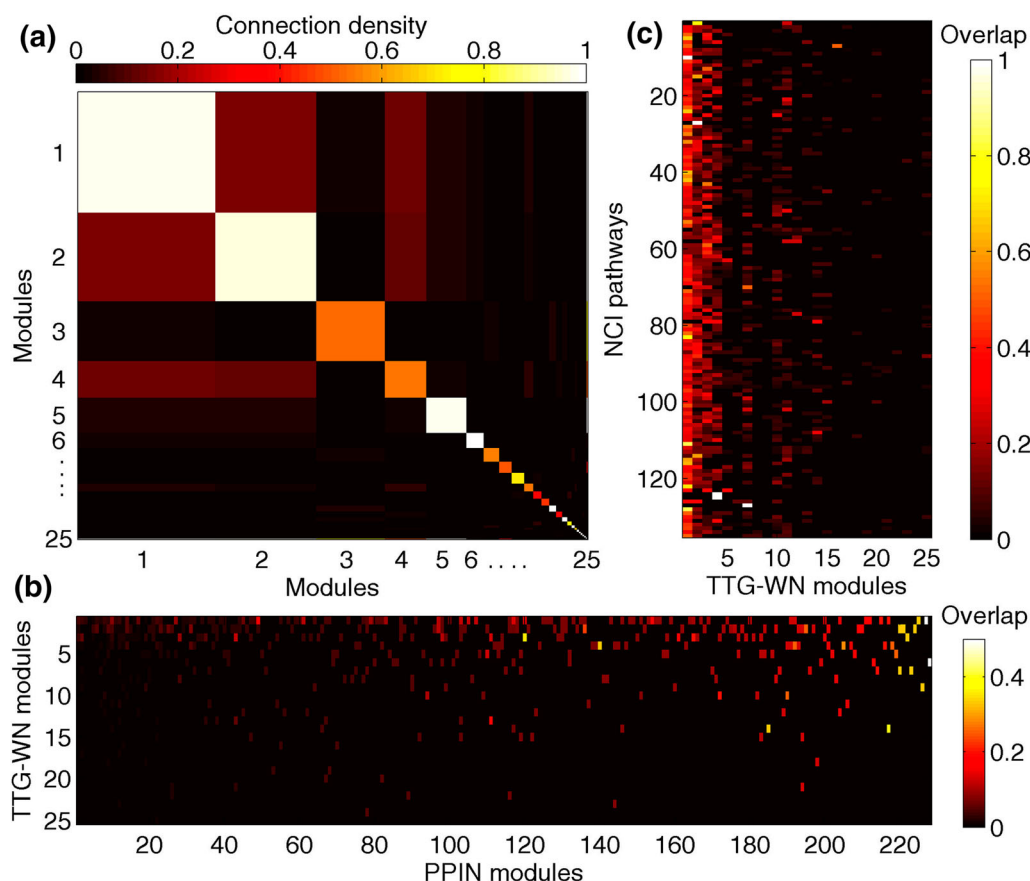


Figure 2. Modular interconnectivity in the tumor type–gene network. **(a)** Matrix representing the average connection density between genes occurring within modules and those in different modules of the TT-GWN, Note that the genes within a module are not only much more densely interconnected compared to the overall connectivity of the network, but modules 1,2,5 and 6 show almost complete intra-connectivity of the genes belonging to them. **(b)** The overlap between the modules of TT-GWN and those of the human PPIN with the modules arranged according to their decreasing size. Several of the smaller PPIN modules have a high degree of overlap with the larger modules of the TT-GWN, implying that some of the latter modules contain groups of genes encoding mutually interacting proteins. **(c)** The overlap between modules of TT-GWN and genes present in different human signaling pathways related to cancer obtained from the National Cancer Institute (NCI) database.

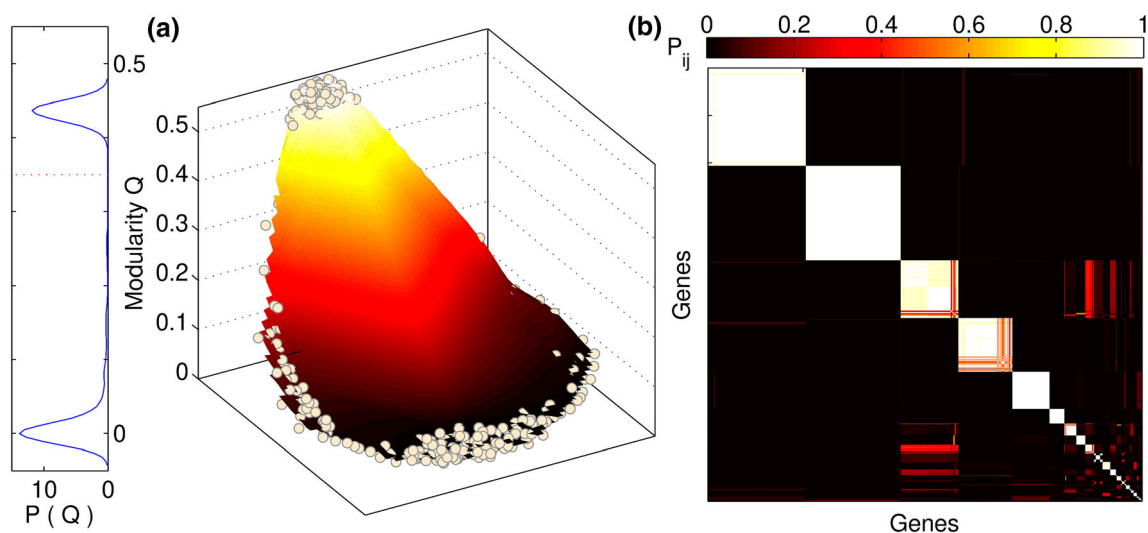


Figure 3. Robustness of the modular decomposition of the tumor type–gene network. **(a)** Modularity of the TT-GWN shown as a function reconstructed from 350 partitionings (circles) obtained through a simulated annealing method for determining communities (Good *et al.* 2010). The axes on the horizontal plane orthogonal to the vertical axis representing modularity Q correspond to embedding dimensions. These are complex functions of the partition space making their scale irrelevant. Positions of the circles on the horizontal plane are obtained by Curvilinear Component Analysis such that the distance between each pair indicates the extent of dissimilarity between the corresponding partitionings of the network into communities. The panel on the left shows the distribution of Q values in the ensemble, suggesting a strongly bimodal nature. The 147 partitionings that compose the peak at high Q (occurring above $Q = 0.35$, indicated by a broken line in the distribution) have been used to quantitatively establish the robustness of the modular identities of the different genes using the consensus matrix shown in **(b)**. The matrix shows the fraction P_{ij} of partitionings in which the pair of genes i, j ($= 1, \dots, 910$) occur in the same module. Most of the modules are seen to be almost completely consistent across all 147 partitionings as indicated by the diagonal blocks having almost all elements equal to 1.

(PPIN) comprising 9270 proteins using the Infomap method, which yielded 542 modules. Figure 2b shows that several of the smaller PPIN modules have large overlap with the larger TT-GWN modules, implying that the latter modules contain genes that code for mutually interacting proteins. The overlap between modules of TT-GWN and the genes present in the National Cancer Institute (NCI) pathway interaction database (figure 2c) indicates that many of the genes in the larger modules belong to different human signaling pathways related to cancer. In order to test the significance of the mesoscopic organization of the empirical network revealed by the modular decomposition, we performed the same analysis on surrogate ensembles of degree-preserved randomized networks (see Methods). The randomization of the TT-GWN results in a homogeneous network that does not have any apparent modular organization, suggesting that the observed mesoscopic structure of the cancer-related gene network is highly significant.

3.1.1 Establishing the robustness of modular decomposition: Given that modular decomposition of networks can be done in a number of different ways

(Schaub *et al.* 2017), it is important to establish that the modules reported here are not sensitively dependent on the method of community partitioning employed and are instead a fundamental attribute of the network connection topology. To this end, we carried out partitioning of the network using two very different methods: (i) a spectral method based on maximization of Q , a quantitative measure of modularity, and (ii) the Infomap method which is based upon optimally compressing information about dynamic processes on the network (see Methods for details). Supplementary figure 2 shows an alluvial diagram comparing the two modular decompositions, explicitly showing the extent of overlap in modular memberships in the two cases. While the Infomap method does appear to generate a larger number of modules, most of these modules are very small. Indeed, several of these smaller modules can be seen to be finer subdivisions of the modules obtained with the spectral method (and which are much fewer in number). This relatively high degree of overlap between the partitionings that have been generated by very different module decomposition techniques based on distinct

theoretical principles suggests that the modular nature of the network reported here is a fundamental property of the network of cancer genes.

As further verification of the robustness of the modular organization uncovered here, we created an ensemble of 350 realizations of the partitioning of the network using a stochastic simulated annealing algorithm (see Methods). By comparing the extent of overlap in the module compositions for these different realizations, we can determine the robustness of the communities in the network. Figure 3a shows the modularity Q values for these different realizations, using a representation such that the symbols (circles) representing partitionings that are similar in nature occur in adjacent positions in the 2D plane that is orthogonal to the vertical axis that represents Q . The coordinates of the symbols in the 2D plane are obtained by Curvilinear Component Analysis (CCA) (Lee and Verleysen 2007). As can be seen from the distribution of Q values obtained from the partitionings, it has a bimodal nature with a clear distinction between values of Q close to zero (and hence, failed attempts at optimal modular partitioning) and those that are relatively high. In fact, the latter realizations are clustered around the value of Q obtained from the Infomap map (≈ 0.42). Taken in conjunction with figure 3b showing the consensus matrix, which establishes that a pair of nodes occur either almost always or almost never in the same module across the different realizations, this result emphasizes the consistency (and hence, robustness) of the modular mesoscopic organization that we have described here.

3.2 Modules, cancer categories, and gene ontology

In order to discern the functional significance of the modular organization, we analyzed the composition of the different modules in terms of gene ontology. In figure 4, the modules are represented as circles connected by lines whose thickness is related to the total number of connections between genes belonging in one module with genes in the other module. We observed that most of the important cancer categories dominate a particular module. For example, breast cancer-related genes comprise about half of the members of module 1; genes related to cancers of the large intestine are responsible for more than half the genes belonging to module 2; cancers of the pancreas and central nervous system dominate modules 5 and 6, respectively, etc. More importantly, a major fraction of genes in eight of

the modules are related to haematopoietic and lymphoid tissue cancers, making this category the most prolific in terms of dominating the mesoscopic organization of the cancer gene network, even though fewer genes (237) are associated with it than breast cancer (244 genes).

Supplementary figures 3 and 4 show the modular dominance of other classes of ontology domains, viz., cellular components and biological processes, in the different communities of the TT-GWN. Apart from these, we also considered the molecular functions. However, unlike in the case of cancer categories, none of the modules of the TT-GWN can be considered to be related to a specific cellular component or biological processes or molecular function. These appear to have almost similar distribution in the different modules, e.g., the genes belonging to cellular locations corresponding to the cytoplasm, nucleus, and plasma membrane dominate most of the modules, while in the case of biological processes, genes responsible for cell communication, signal transduction, or regulation of nucleobase metabolism contribute the majority of elements in most modules. This result can be understood in light of the fact that cancer is a complex group of multi-factorial diseases involving several genes in multiple cellular locations and responsible for different biological processes and molecular functions.

3.3 Closeness between different cancer categories and tumor types

The relation between different cancer categories or tumor types can potentially be understood in terms of the degree of overlap of the genes associated with the different modules of TT-GWN or PPIN. To this end, we clustered the cancer categories or tumor types in terms of the similarity in their modular spectra (see Methods). The relations between the different cancer categories (CC) are represented in terms of a dendrogram shown in figure 5a, where the CC gene classes are projected on the space of modules of TT-GWN. Closely connected cancer categories that are related through environmental factors, viz., oral cancers, cancer of upper aerodigestive tract, liver cancer, and urinary tract cancer, have been highlighted. By performing a similar decomposition of the different tumor types in modular space we observed a closeness between breast and ovarian tumors, which are related by hormones, hereditary links, and clinical treatment (figure 5b).

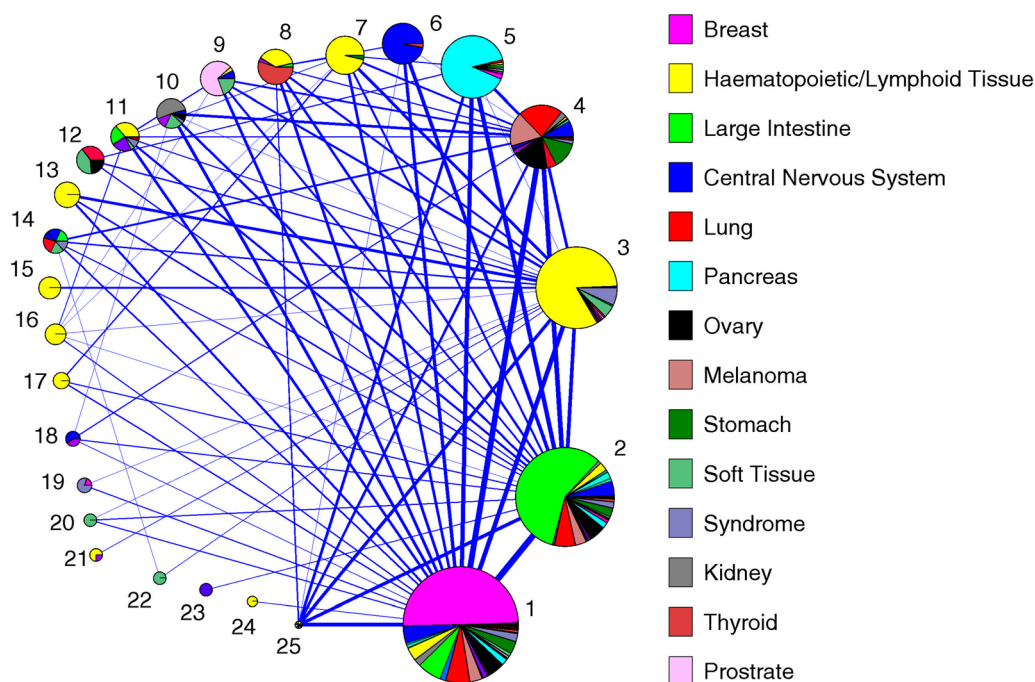


Figure 4. The composition of the modules of the TT-GWN in terms of the different cancer categories associated with the genes comprising each module. Each circle represents a module of TT-GWN with its size being proportional to the number of genes in that module. The fraction of genes in each module that are associated with particular cancer categories is shown in terms of the pie chart inscribed within each circle representing a specific module. For example, the 246 genes belonging to module 1 have in all 489 associations with different cancer categories. As 243 of these genes are linked to breast cancer, the association of this module with the category of breast cancer is represented by the fraction $243/489$, i.e., approximately 50% of the area of the circle representing the module. The thickness of the line connecting a pair of modules is related to the total number of connections that exist between the genes of the two modules.

3.4 Functional roles of cancer genes

We investigated the importance of individual genes linked to cancer in terms of their connectivity. This is revealed by a comparison between the localization of their connections within their own community and their global connectivity profile over the entire network. In order to do this, we focused on (i) the within-module degree z -score of a node within its module, which indicates the importance of a node (in terms of how many connections it has) within its own module, and (ii) its participation coefficient, P , which measures how dispersed the connections of a node are among the different modules (see Methods). A node having low within-module degree is called a *non-hub* ($z < 1$), which can be further classified according to their fraction of connections with other modules. Following Guimera *et al.* (2007), these were classified as (R1) *ultra-peripheral nodes* ($P \leq 0.05$), having connections only within their module; (R2) *peripheral nodes* ($0.05 < P \leq 0.62$), which have a majority of their links within their module; (R3) *non-hub connectors* ($0.62 < P \leq 0.8$), with many links connecting nodes outside

their modules; and (R4) *kinless nodes* ($P > 0.8$), which form links uniformly across the network. Hubs, i.e., nodes having relatively large number of connections within their module ($z \geq 1$), were also divided according to their participation coefficient into (R5) *provincial hubs* ($P \leq 0.62$), with most connections within their module; (R6) *connector hubs* ($0.62 < P \leq 0.8$), with a significant fraction of links distributed among many modules; and (R7) *global hubs* ($P > 0.8$), which connect homogeneously to all modules. The thresholds of P for classifying the nodes into different categories were chosen based on the criteria described in Guimera and Amaral (2005), viz., if $P < 0.05$, a node has almost all its connected neighbors in the same module, and if $P < 0.62$, more than 60% of its neighbors are in the module, while if $P > 0.8$, then it has fewer than 35% of its neighbors in the same module. This classification allowed us to distinguish nodes according to their different roles as brought out by their intra-modular and inter-modular connectivity patterns.

We used this classification method on the nodes of the TT-GWN in order to identify the genes that play a vital role in cancer through coordinating the behavior

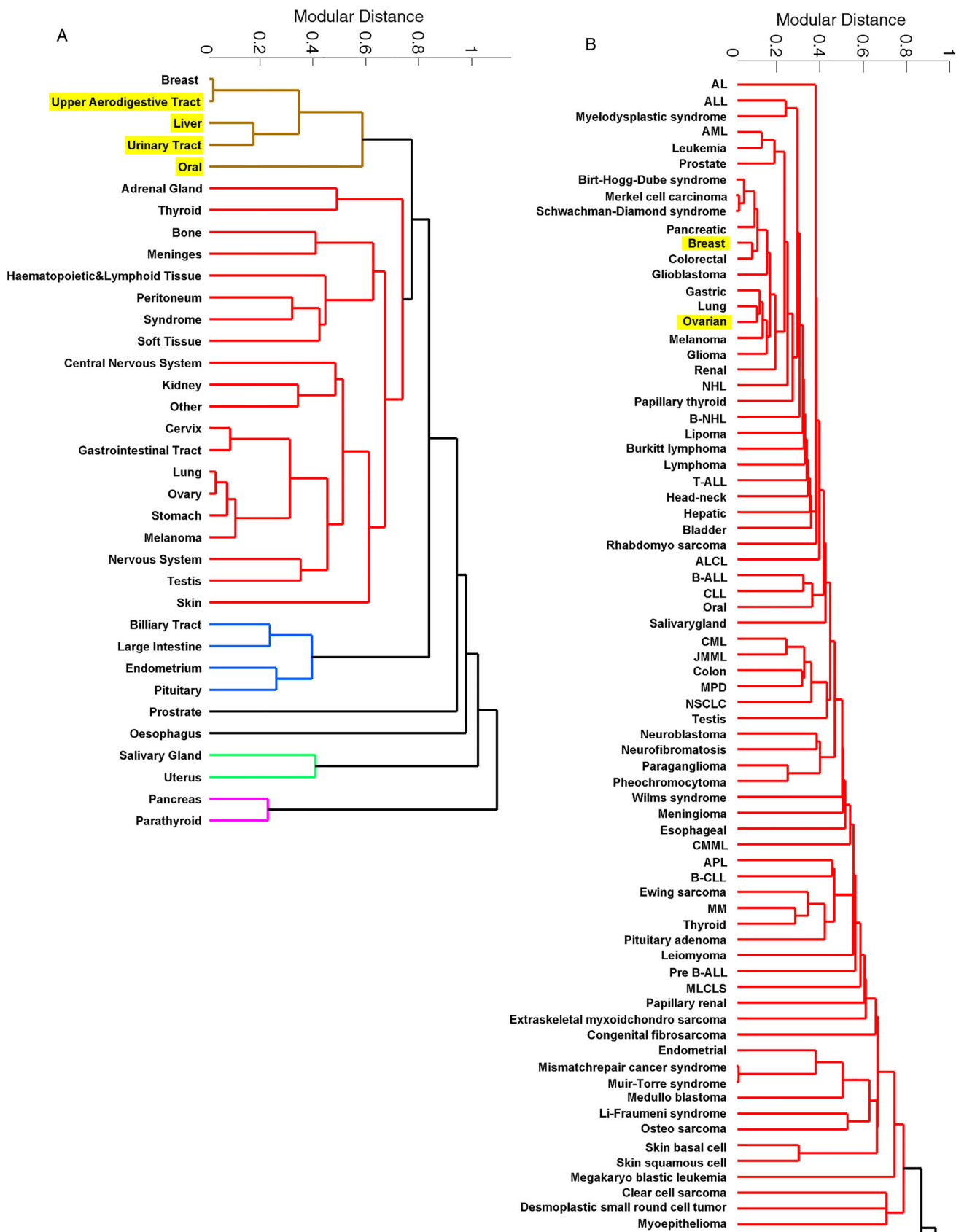


Figure 5. Relation between different cancer categories, and between different tumor types, based on modular spectra. **(a)** Dendrogram of cancer categories obtained by projecting CC gene classes over the space of modules of TT-GWN. Closely connected cancer categories related via environmental factors are highlighted. **(b)** Dendrogram of tumor types obtained by projecting TT gene classes over the space of PPIN modules (only a section of the entire tree is shown). The closeness of breast and ovarian tumor types, which are related by hormones, hereditary linkages, and clinical treatments, is indicated in the figure.

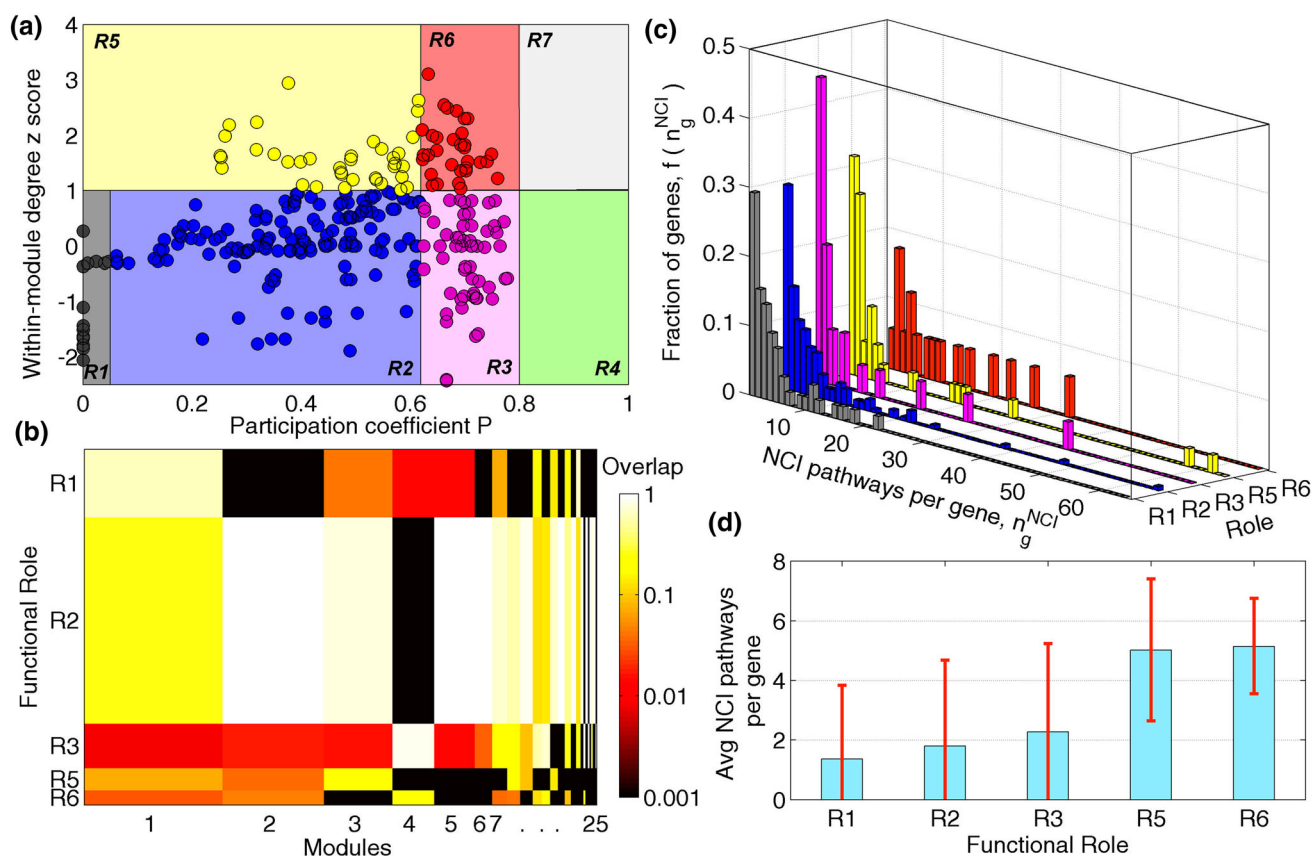


Figure 6. Classification of genes in terms of their functional role according to intra- and inter-modular connectivity in TT-GWN. **(a)** The within-module degree z-score of each gene in TT-GWN is shown against the corresponding participation coefficient P . The within-module degree measures the connectivity of a node to other nodes within its own module, while the participation coefficient measures its connectivity with nodes in the entire network. Nodes in different regions in the P - z space are categorized as R1: ultraperipheral nodes, i.e., nodes with all their links within the module; R2: peripheral nodes, i.e., nodes with most of their links within their module; R3: non-hub connector nodes, i.e., nodes with many links to other modules; R4: non-hub kinless nodes, i.e., nodes with links homogeneously distributed among all modules; R5: provincial hubs, i.e., hub nodes with the vast majority of links within their module; R6: connector hubs, i.e., hubs with many links to most of the other modules; and R7: global hubs, i.e., hubs with links homogeneously distributed among all modules. **(b)** matrix representing the overlap between modules of TT-GWN and the functional roles of their constituent elements (modules are arranged in terms of decreasing size). **(c)** The fraction of genes with a particular functional role associated with a specified number of human signaling pathways related to cancer (NCI database). **(d)** The mean number of signaling pathways in the NCI database that a gene with a specific functional role is associated with.

of the network either locally within their community or globally over the entire system (figure 6a). Our analysis revealed that while the network does not have any global hubs (R7), there are several connector hubs, e.g., *MAPK14*, *TP53*, *BCL10*, etc. (table 1; supplementary

table 7 contains the entire list of genes classified according to their functional role) that can be potential targets for therapeutic intervention through pharmaceutical drugs. Figure 6b shows the overlap between the modules and the functional role of the genes

Table 1. Identities of genes that are connector hubs (R6) in TT-GWN, and connector and global hubs (R7) in the human PPIN

Connector Hubs (R6) of TT-GWN			
<i>APC</i>	<i>INSRR</i>	<i>RPS6KA2</i>	<i>MELK</i>
<i>FAS</i>	<i>MARK1</i>	<i>MAP2K4</i>	<i>KIF1B</i>
<i>ATM</i>	<i>NRAS</i>	<i>TFE3</i>	<i>RAD54B</i>
<i>BRAF</i>	<i>ROR1</i>	<i>TGFBR2</i>	<i>TRIM33</i>
<i>MAPK14</i>	<i>PCMI</i>	<i>TP53</i>	<i>TEX14</i>
<i>EPHB1</i>	<i>PDGFRA</i>	<i>TTN</i>	<i>WNK4</i>
<i>FRAP1</i>	<i>PRCC</i>	<i>TRRAP</i>	<i>ALPK2</i>
<i>FYN</i>	<i>PTCH1</i>	<i>BCL10</i>	<i>NEK10</i>
<i>IGH@</i>	<i>ROS1</i>	<i>AATK</i>	<i>NEK8</i>
Connector Hubs (R6) of PPIN			
<i>CREBBP</i>	<i>GNAI1</i>	<i>SKP1</i>	<i>CCDC85B</i>
<i>DLG1</i>	<i>JUN</i>	<i>SRC</i>	<i>C1orf103</i>
<i>DLG4</i>	<i>SMAD1</i>	<i>TGFBR1</i>	<i>KRTAP4-12</i>
<i>ESR1</i>	<i>MDFI</i>	<i>TRIP13</i>	<i>SFRS12</i>
<i>FNI</i>	<i>PCNA</i>	<i>SLC9A3R1</i>	
<i>FYN</i>	<i>SHBG</i>	<i>SETDB1</i>	
Global Hubs (R7) of PPIN			
<i>ACTA1</i>	<i>EWSR1</i>	<i>PRKACA</i>	<i>YWHAB</i>
<i>ACTB</i>	<i>HDAC1</i>	<i>PRKCA</i>	<i>YWHAG</i>
<i>AR</i>	<i>HRAS</i>	<i>RAC1</i>	<i>GFI1B</i>
<i>CDC42</i>	<i>SMAD2</i>	<i>RBI</i>	<i>NDRG1</i>
<i>MAPK14</i>	<i>SMAD4</i>	<i>ATXN1</i>	<i>PRPF40A</i>
<i>CTNNB1</i>	<i>SMAD9</i>	<i>STX1A</i>	<i>ATF7IP</i>
<i>ATN1</i>	<i>MAGEA11</i>	<i>TP53</i>	<i>UBQLN4</i>
<i>EP300</i>	<i>PPP1CA</i>	<i>TRAF2</i>	<i>SUMO4</i>

belonging to them. The overlap is measured in terms of the fraction of genes in a specific module that has a particular functional role.

Next, we investigated the significance of the functional role of a gene determined from its intra- and inter-modular connectivity by looking at its association with the probability that the gene is connected to one or more human signaling pathways related to cancer. For this purpose, we used the Pathway Interaction Database (PID, available via the NDEX database, <http://www.ndexbio.org/>) that was established as a collaboration between the U.S. National Cancer Institute (NCI) and Nature Publishing Group (Schaefer *et al.* 2009). This is a highly structured collection of 137 curated and peer-reviewed human signaling pathways assembled from 9248 known human biomolecular interactions and key cellular processes. Figure 6c shows the fraction of genes with a particular functional role associated with a specified number of human signaling pathways related to cancer. The distribution clearly shows that there are many more pathways associated with connector hubs (R6) than with genes having other functional roles. Further, genes that are hub nodes (R5 and R6) have a much higher number of signaling pathways associated with them, on average (figure 6d). Thus, this supports our conclusion that connector hub genes can be potential therapeutic targets.

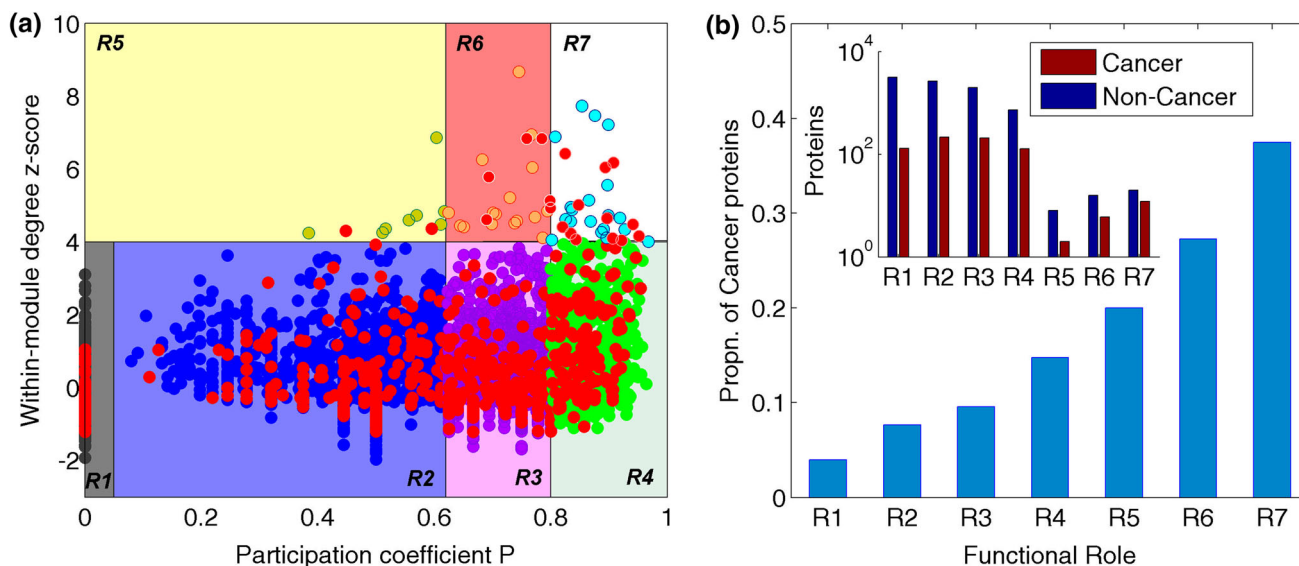


Figure 7. The role of individual proteins according to their intra- and inter-modular connectivity in PPIN. (a) The within-module degree z-score of each protein in the PPIN is shown against the corresponding participation coefficient P . The red filled circles represent the cancer proteins. The probability that a global hub (R7) or connector hub (R6) protein has a link with cancer is extremely high (0.38 and 0.27, respectively) compared with the corresponding average probability for any node in the PPIN ($=0.07$). (b) The proportion of cancer proteins (and the total number of cancer and non-cancer proteins, inset) in the population of proteins with each functional role.

Table 2. Five-year survival rates (5YSR) for different tumor types obtained from SEER program database

Tumor Type	5YSR	Tumor Type	5YSR	Tumor Type	5YSR
Acute leukemia	5.8	Esophageal	13.6	Oligodendroglioma	68.2
Anaplastic large-cell lymphoma	53.9	Ewing's sarcoma	48.4	Oral	59.4
Acute lymphocytic leukemia	62.2	Extra skeletal myxoidchondrosarcoma	91	Osteosarcoma	59.2
Acute myelogenous leukemia	16.5	Gastrointestinal	27.5	Ovarian	53.8
Acute promyelocytic leukemia	60	Glioblastoma	2.9	Pancreatic	4.8
Adrenal	38.7	Glioma	45.2	Papillary thyroid	98.7
Adreno cortical	41.2	Head-neck	57.1	Paraganglioma	65.1
B-cell non-Hodgkin's lymphoma	50.4	Hepatic	8	Parathyroid	93.1
Basal cell carcinoma	99.4	Hyper parathyroidism-jaw tumor syndrome	93.1	Pheochromocytoma	60.3
Bladder	81.9	Leiomyomata	51.9	Pilocytic astrocytoma	35.8
Brain	23.6	Leukemia	55	Pituitary adenoma	63.8
Breast	87.1	Lipoma	82.8	Prostate	97.6
Burkitt's lymphoma	45.4	Lymphocytic leukemia	79.5	Renal	60.2
Chronic lymphatic leukemia	74.9	Lymphoma	70.6	Retinoblastoma	93.5
Chronic myelomonocytic leukemia	37.7	Multiple myeloma	29.4	Rhabdomyosarcoma	64
CNS	69.5	Myelo proliferative disorder	31.7	Salivary gland	73.9
Cervical	71.5	Medullary thyroid	82.1	Schwannomatosis	99
Cholangiocarcinoma	4.5	Medulloblastoma	66.4	Sézary syndrome	88.4
Chondrosarcoma	81.6	Melanoma	90.2	Stomach	21
Clear cell sarcoma	83.4	Meningioma	60	T-cell acute lymphoblastic leukemia	24.3
Colon	64	Merkel cell carcinoma	62.8	Testis	96
Colorectal	62.6	Mesothelioma	8.2	Thyroid	96
Diffuse large B-cell lymphoma	50.4	Non-Hodgkin's lymphoma	60	Wilms' syndrome	78.1
Dermatofibrosarcoma protuberans	99.9	Non-small cell lung cancer	12.1		
Endometrial	74.6	Nasopharyngeal	56.6		

3.5 Functional roles of proteins in PPIN

The basis of all biological functions in the cell in health and disease are the interactions between proteins. Therefore, to further support our hypothesis regarding the importance of network elements R6 and R7 having functional roles, we also analyzed the largest connected component of the PPIN comprising 9270 proteins. Classification of proteins into different functional roles in terms of their intra- and inter-modular connectivity shows a preponderance of cancer genes among the connector hubs and global hubs (figure 7a). Compared with the probability of a randomly chosen element of the PPIN being related to cancer (0.07), the probability that a global hub (R7) or connector hub (R6) is related to cancer is seen to be

extremely high (0.38 and 0.27, respectively) (figure 7b). Our mesoscopic structural study of the PPIN revealed several global hubs of which 12 are known to be cancer genes. The 20 other genes which were also identified as being global hubs in our analysis may have previously unsuspected roles in the genesis and treatment of cancer (table 1).

3.6 Relating functional role of cancer gene and patient survivability

It is well known that survival probability of a cancer patient depends on the tumor type or cancer category. For instance, the 5-year survival rate for breast or prostate cancer patients is significantly higher than

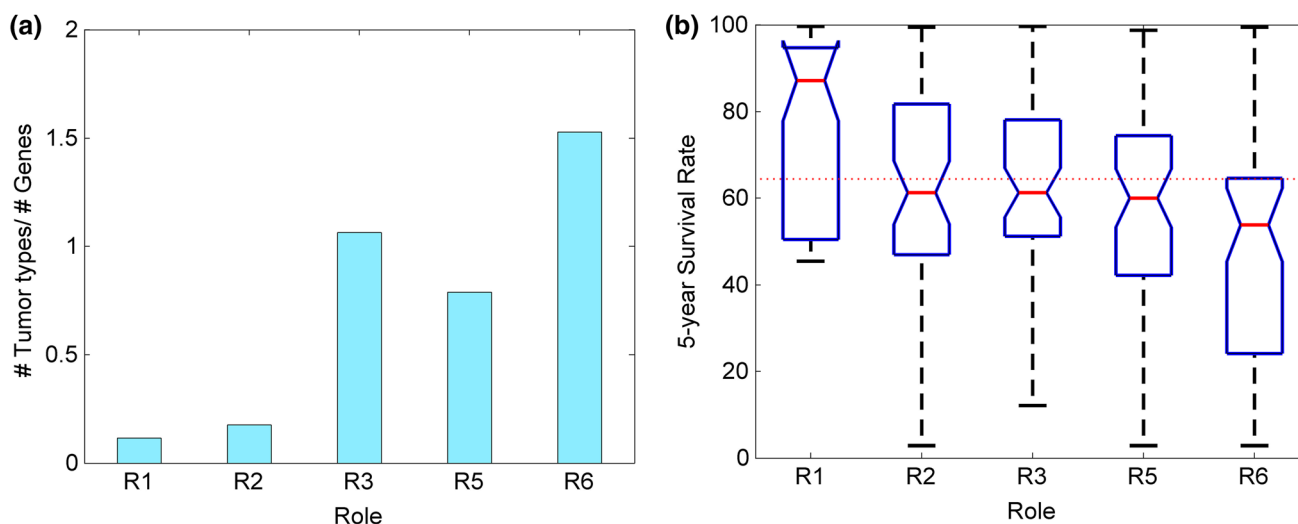


Figure 8. Distribution of cancer survival rates associated with genes having specific functional roles in TT-GWN. **(a)** The ratio of the number of tumor types to genes for each functional role category R1, R2, R3, R5 and R6 of genes in TT-GWN. **(b)** Box plot showing the mean 5-year survival rates for different tumor types corresponding to genes having different roles. The broken line represents the mean 5-year survival rate for all cancers. The data is for US population of cancer patients obtained from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute. The median survival rates become progressively low from R1 to R6 signifying that genes that are connector hubs (R6) are associated with tumors that have lower survival rates.

patients diagnosed with brain or lung cancer. Therefore, we have also investigated the relation between tumor types associated with genes that have specific functional roles (R1–R6) and the 5-year survivability rates for patients having these types of tumors. For this purpose, we used 5-year survival statistics from the Surveillance Epidemiology and End Results (SEER) Program database (available at <https://seer.cancer.gov/>), compiled by the National Cancer Institute as a service to researchers and physicians (Mariotto *et al.* 2014). The survival rates for 73 tumor types available from the SEER data (table 2) was compared with the classification into the 135 tumor types that we used for constructing the TT-GWN.

Figure 8a shows the ratio of the number of tumor types to genes for each functional role category of genes in the TT-GWN. This shows that connector hub genes are associated on average with a much larger number of tumor types than genes having other functional roles. Figure 8b shows that the 5-year survival rates for tumor types associated with connector hub genes are lower than those associated with genes having other functional roles. For comparison, the average 5-year survival rate for all tumor types (broken line) is shown. Thus, genes which act as connector hubs in the TT-GWN are associated with tumors having higher mortality and should be preferentially targeted for therapeutic intervention.

4. Discussion

Despite being one of the leading causes of death in the developed world, cancer is yet to be tamed owing to the complex, heterogeneous nature of the disease. Despite the understanding that cancer is a systems-level disease and cannot be treated by targeting a single factor, the large number of elements involved and the dense set of interactions between them have prevented a major breakthrough in this area. In this article we have adopted a mesoscopic approach by identifying structural modules in the network of cancer-related genes. This has helped us in identifying several genes that have the important functional role of connecting members in their own module with members of other modules. Thus, these genes help coordinate the behavior of the entire network in health and disease, and play a vital role in the origin and treatment of cancer. We validated our hypothesis by showing that tumors associated with these genes were involved in many human signaling pathways related to cancer. More importantly, we showed that patients suffering from tumors involving these genes had a much lower survival rate than those suffering from other types of tumors. The integrated knowledge of cancer networks gained by assembling and evaluating the functional roles of the different genes and proteins associated with many tumor types and cancer categories may provide

new insights for understanding the interconnectedness of key players in the genesis and treatment of the disease. This may have implications for enhancing the efficacy of multiple drug action and proper drug administration, as well as in the discovery of novel drug targets.

Acknowledgements

The work was partly supported by the ISc Complex Systems Project (XII Plan) funded by the Department of Atomic Energy, Government of India. We would like to thank Soumya Easwaran and Anand Pathak for their assistance in preparing figures, and Indrani Bose, Shaon Chakrabarti, Arjun Krishnan, Ramakrishna Ramaswamy, M S Santhanam and Somdatta Sinha for helpful discussions.

References

- Clauset A, Newman ME and Moore C 2004 Finding community structure in very large networks. *Phys. Rev. E* **70** 066111
- Cloutier M and Wang E 2011 Dynamic modeling and analysis of cancer cellular network motifs. *Integr. Biol.* **3** 724–732
- Fortunato S 2010 Community detection in graphs. *Phys. Rep.* **486** 75–174
- Futreal AP, Coin L, Marshall M, *et al.* 2006 A census of human cancer genes. *Nat. Rev. Cancer* **4** 177–183
- Goh KI, Cusick ME, Valle D, *et al.* 2007 The human disease network. *Proc. Natl. Acad. Sci. USA* **104** 8685–8690
- Gong X, Wu R, Zhang Y, *et al.* 2010 Extracting consistent knowledge from highly inconsistent cancer gene data sources. *BMC Bioinform.* **11** 76
- Good BH, De Montjoye YA and Clauset A 2010 Performance of modularity maximization in practical contexts. *Phys. Rev. E* **81** 046106
- Guimera R and Amaral LAN 2005 Cartography of complex networks: modules and universal roles. *J. Stat. Mech.* <https://doi.org/10.1088/1742-5468/2005/02/P02001>
- Guimera R, Pardo MS and Amaral LAN 2007 Classes of complex networks defined by role-to-role connectivity profiles. *Nat. Phys.* **3** 63–69
- Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA 2005 Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33** D514–D517
- Hartwell LH, Hopfield JJ, Leibler S and Murray AW 1999 From molecular to modular cell biology. *Nature* **402** C47–C52
- Hornberg JJ, Bruggeman FJ, Westerhoff HV and Lankelma J 2006 Cancer: a systems biology disease. *Biosystems* **83** 81–90
- Kreeger PK and Lauffenburger DA 2010 Cancer systems biology: a network modeling perspective. *Carcinogenesis* **31** 2–8
- Lancichinetti A and Fortunato S 2009 Community detection algorithms: a comparative analysis. *Phys. Rev. E* **80** 056117
- Lancichinetti A and Fortunato S 2012 Consensus clustering in complex networks. *Sci. Rep.* **2** 336
- Lee JA and Verleysen M 2007 *Nonlinear dimensionality reduction* (New York: Springer)
- Mariotto AB, Noone AM, Howlander N, *et al.* 2014 Cancer survival: an overview of measures, uses, and interpretation. *J. Nat. Cancer Inst. Monogr.* **2014** 145–186
- Milo R, Shen-Orr S, Itzkovitz S, *et al.* 2002 Network motifs: simple building blocks of complex networks. *Science* **298** 824–827
- Mukherjee S 2010 *The emperor of all maladies: A biography of cancer* (New York: Scribner)
- Newman MEJ 2004 Detecting community structure in networks. *Eur. Phys. J. B* **38** 321–330
- Newman MEJ 2006 Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582
- Newman MEJ and Girvan M 2004 Finding and evaluating community structure in networks. *Phys. Rev. E* **69** 026113
- Porter MA, Onnela JP and Mucha PJ 2009 Communities in Networks. *Notices AMS* **56** 1082–1097
- Prasad TSK, Goel R, Kandasamy K, *et al.* 2009 Human Protein Reference Database - 2009 update. *Nucleic Acids Res.* **37** D767–D772
- Rosvall M and Bergstrom CT 2008 Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA* **105** 1118–1123
- Rosvall M and Bergstrom CT 2010 Mapping change in large networks. *PLoS One* **5** e8694
- Schaefer CF, Anthony K, Krupa S, *et al.* 2009 PID: the pathway interaction database. *Nucleic Acids Res.* **37** D674–D679
- Schaub MT, Delvenne JC, Rosvall M and Lambiotte R 2017 The many facets of community detection in complex networks. *Appl. Netw. Sci.* **2** 4
- WHO 2021 Cancer fact sheet <https://www.who.int/en/news-room/fact-sheets/detail/cancer>

Corresponding editor: MOHIT KUMAR JOLLY