Randomness and preserved patterns in cancer network

Aparna Rai[1], A. Vipin Menon[2] & Sarika Jalan[1,2]

[1]Centre for Bio-Science and Bio-Medical Engineering, Indian Institute of Technology Indore, M-Block, IET-DAVV Campus, Khandwa Road, Indore 452017, India, [2]Complex Systems Lab, Discipline of Physics, Indian Institute of Technology Indore, M-Block, IET-DAVV Campus, Khandwa Road, Indore 452017, India.

Breast cancer has been reported to account for the maximum cases among all female cancers till date. In order to gain a deeper insight into the complexities of the disease, we analyze the breast cancer network and its normal counterpart at the proteomic level. While the short range correlations in the eigenvalues exhibiting universality provide an evidence towards the importance of random connections in the underlying networks, the long range correlations along with the localization properties reveal insightful structural patterns involving functionally important proteins. The analysis provides a benchmark for designing drugs which can target a subgraph instead of individual proteins.

One of the major goals of the post-genomic era is to understand the role of proteomics and genomics in human health and diseases[1]. The world health organization has estimated that about 11 percent of the total cancers accounts for breast cancer and a drop of about one-third of the cancer deaths could be attained if detected and treated early[2]. Molecular studies and healthcare research has shown that the detection of more and more genes after BRCA1 and BRCA2 implicative in breast cancer has rendered research, diagnosis and treatment strategies more ambiguous as the dangers being posed by those genes are yet uncertain[3]. Complexity as well as variations at every stage of the cancer render designing drug targets very difficult[4,5]. The ample availability of data in functional genomic and proteomic information and the development of high-throughput data-collection techniques has resulted from basic gene-based traditional molecular biology to a systems approach of network biology[6,7]. In this approach, biological processes are considered as complex networks of interactions between numerous components of the cell rather than as independent interactions involving only a few molecules[8–10]. Previous attempts to understand various diseases under network biology approach reveals that various types of cancers are interlinked to each other through some pathways which are altered in different diseases[11]. Further analysis of the centrosome dysfunction under the network theory framework demonstrates importance of hub proteins as well as those connected with them[12,13]. This paper, in order to achieve a deeper understanding of the complexity of breast cancer, its interacting patterns, role and importance of interaction patterns for the disease, analyzes protein protein interaction (PPI) network using a novel mathematical tool random matrix theory (RMT). This technique is known to develop fifty years back to explain interactions of complex nucleons[14], has recently exhibited its remarkable success in understanding complex systems arising from diverse fields ranging from quantum chaos to galaxy[15]. While structural parameters namely degree, clustering coefficient (CC), degree clustering correlation and diameter demonstrate similarities in both the normal and disease networks, the structural patterns such as cliques combined with proteins revealed through the spectral analysis indicate changes in both the networks, which might be important behind transformation of a cell from the normal to the disease state. The present work not only straightens the importance of structural patterns in the disease, further demonstrating the success of the network framework, but also for the first time analyzes a disease using the RMT techniques. This combined framework helps in detecting proteins, beyond their structural significance in the underlying network, which are crucial for the disease. A detailed analysis of the top contributing nodes (TCNs) in the localized eigenvectors reveal their importance in the occurrence of the disease state.

Results and discussion

Structural properties of cancer networks. The structural parameters of the largest connected cluster of the network using dataset described in the materials and methods section, are summarized in the Table 1.

The degree distribution $\rho(k)$ of both the networks follow the power law (Figure 1) indicating the presence of very high degree nodes. These nodes are known to keep the network robust against random external perturba-

Table 1 | Network parameters for both the normal and disease networks. The total number of proteins N collected using various databases (described in the Method section), number of proteins in the largest connected cluster N_{LCC} and connections NC_{LCC}, the average degree $\langle k \rangle$, average clustering coefficient $\langle CC \rangle$, the number of nodes having $CC = 1$ in the whole network (N_{CC}) and the average IPR of both the networks

Network	N	N_{LCC}	NC_{LCC}	$\langle k \rangle$	D	$\langle CC \rangle$	N_{CC}	$\langle IPR \rangle$
Normal	2464	2441	15118	12	11	0.28	153	0.004
Disease	2096	2046	14150	14	10	0.29	107	0.005

tions, as well as have been found to be functionally important in many pathways[16]. The degree and CC of the normal and the disease networks are negatively correlated (Figure 1), as found in the case of other biological systems[17]. As mentioned in the Table 1, the average CC of both the networks is high as exhibited by most of the biological systems investigated under network theory framework indicating presence of functional modules[10]. What follows that the disease and normal networks exhibit overall similar statistics for widely investigated structural properties but the crucial differences between them which are of potential importance, are revealed through the analysis of cliques structures and spectra. As reflected in Table 1, the disease network bears less number of nodes with $CC = 1$ than the normal one. The value of CC for a node being one reflects the formation of complete subgraph or clique comprising of that node. The higher value of average CC implicates the presence of high number of clique structure in a network[18]. Further, cliques are known as building blocks of a network for making the underlying system more robust[19] and stable[20]. Additionally nodes forming cliques structure are known to be preserved during evolution[21]. What follows that the disease network having less cliques of order three as well as less number of nodes with CC being one as compared to the normal one indicate that there is a demolition of building blocks in the disease state, which may be leading to a more *unstable* underlying
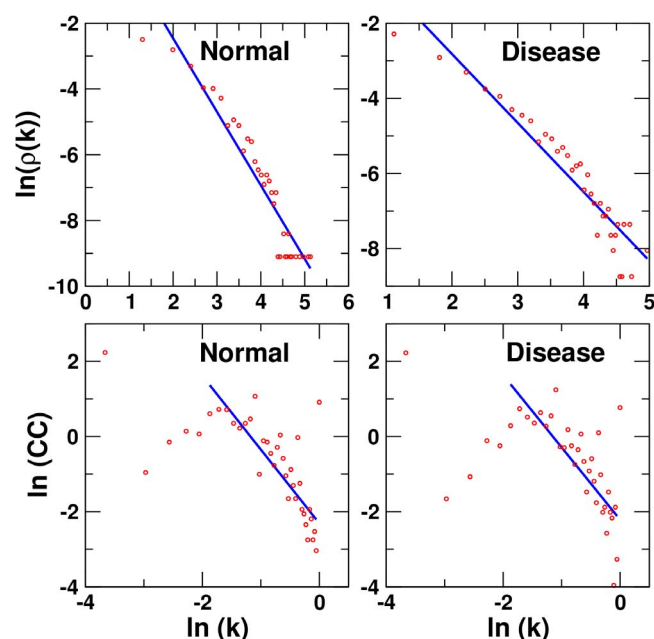
system, and might be one of the reasons behind occurrence of the disease.

The importance of clique structures would become more clear after we explain spectral properties and local patterns of top contributing nodes (TCNs) appearing in the spectral analysis performed under the powerful framework of RMT. This analysis not only reveals functionally important proteins but also helps in uncovering importance of structural pattern in the disease network. In the following, we provide the results pertaining to the global spectral properties, eigenvalue fluctuations and properties of nodes appearing in the localized eigenvectors. These are the most frequently used techniques in RMT for analysis of spectral properties in order to achieve a comprehensive understanding of the underlying complex system.

Universality and the deviation from the same. The eigenvalue statistics reflects typical triangular shape with the tail of the distribution (Figure 2) relating with the exponent of the power law of degree distribution as observed for many other biological and real world networks[22,23]. Both the disease and normal networks have about 30% eigenstates with zero eigenvalues. This high degeneracy at zero is not surprising as many of the biological networks have been shown to yield very high degeneracy at zero[22].

Further, as depicted in Figure 3, both the disease and normal networks follow GOE statistics of random matrix theory at the consecutive eigenvalues captured through the distribution of their ratio (Eq. 4), which reflects that both the networks have a *minimal amount of randomness*[22]. Randomness in a network might be arising due to some nonsense mutations[24] occurring in the underlying system. We remark that in dynamical systems, randomness may be related with the unpredictable nature of time evolution (for example: chaotic systems)[25], whereas for networks, randomness is referred to as random connections between nodes[26], which for biological systems might have evinced in the course of evolution randomly and not because of any particular functional importance of that connection. For instance, emergence of the modular structure in networks, which are known to be motivated by their specific functional role in the evolution[27] might be linked with random connections perhaps resulting from mutations[24]. Further, randomness in the interactions has been known to be important for functioning of the underlying system. For instance, information processing in the brain is considered to arise because of many random long-range connections among different modules[28] making the underlying system robust[19]. (Supplementary material contains details about RMT technique). The universal Gaussian orthogonal ensemble (GOE) statistics displayed by the disease network on one hand, indicates the *robustness* of the cell even in the disease state, which may be considered crucial for maintenance and housekeeping processes of cancerous cell and on other hand establishes that the breast cell can be modeled using GOE of RMT and we can apply all the techniques developed under
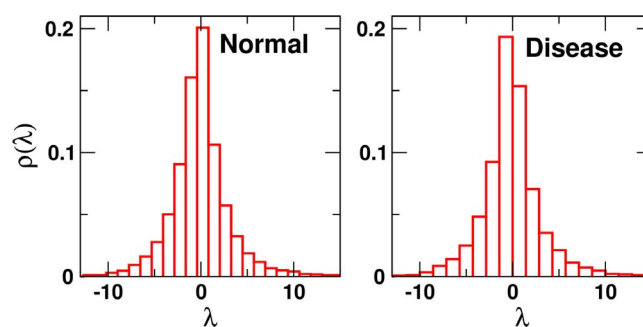

Figure 1 | Degree distribution and degree-CC correlations for the normal and disease networks. Left panel of the normal network show that the degree distribution follows power law and the degree clustering coefficient correlation shows are negatively correlated. The right panel gives us the same results for the disease networks.

Figure 2 | Eigenvalue distribution of both normal and disease networks. The plots depict triangular distribution for both the networks with a high degeneracy at zero.
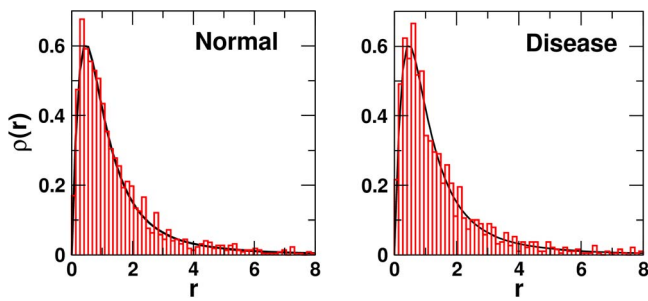
Figure 3 | Spacing distribution. The ratio of eigenvalues spacing follow GOE statistics for both the networks. The bars represent data points and solid line represents Eq.4 with parameters of GOE statistics.

the well established framework of the RMT to understand the breast cancer.

After the short range correlations, which is analyzed through probability of ratio of consecutive level, the second most insightful step in the RMT is the analysis of long range correlations in eigenvalue using spectral rigidity test, which is generally done using Δ_3 statistics given by Eq. 5. This test reveals that both the networks follow RMT prediction of GOE statistics till a particular value of L (Eq.6) and deviates thereafter. Interestingly, the value of L for the disease network is less as compared to the normal one (Figure 4), suggesting that the normal network is more random than the disease network[26] or the disease network is more ordered than the normal one. This interpretation combined with the observation that the disease state has a less number of connections (pathways expressed) than that of the normal one, implicates that there are some interactions getting hampered or silenced during the course of mutation leading to the disease[11]. What follows that, these hampered pathways should be corresponding to or treated as *random pathways* and as *randomness* is one of the essential ingredient for the robustness of a cell[26], lack of *sufficient randomness* might be leading to the disease.

While universal part following RMT prediction reflects the importance of *random connections* in the biological networks, we will witness in the following that the non-universal part of the spectra deviating from RMT provides direct clue about the set of nodes (proteins) relevant for the occurrence of the disease state. This is achieved by analyzing localization properties of eigenstates which provides a quantitative picture of non universal part of the spectra.

Important proteins through eigenvector localization. Based on the IPR values calculated using Eq.7, the eigenstates can be divided into two components, one which follows RMT predictions of Porter-Thomas distribution[29], and another one which deviates from this universality and show localization (Figure 5). The average IPR calculated using Eq.8 comes out to be more for the disease network than for the normal one, which is in the direct relation with the
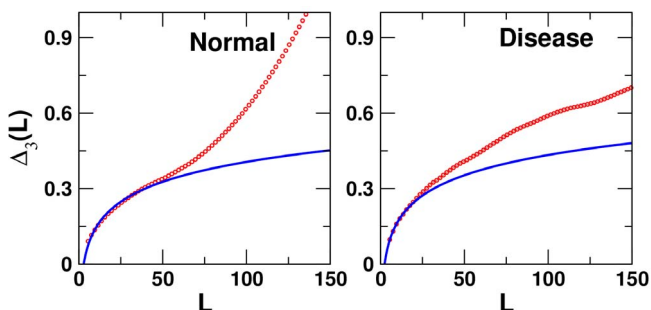
behavior of Δ_3 statistics demonstrating that the normal network is more random than the disease network. The part of the unfolded spectra following universality corresponds to random interactions in the underlying system[30], whereas non-universal part can be exploited to get the system dependent information. The non universal part annotating the importance of the localized eigenvectors reveals important proteins as explained below. The disease network yields 34 TCNs corresponding to the top 5 most localized eigenvectors extracted by taking the threshold as $1/IPR$[31], of which 18 appears to be unique in the disease and 14 are common to both the disease and the normal networks. In the following section, we discuss briefly the functions of these proteins selected through the localization property of eigenvectors (Table 2).

Functional properties of disease proteins. The most important outcome of the functional analysis of proteins corresponding to the TCNs is that all of them are involved in important pathways leading to breast cancer. The first localized eigenvector has five contributing nodes, of which MTHFD1L is responsible for synthesis of purines in mitochondria but its expression is up-regulated in breast cancer leading to proliferation, invasiveness and anti-apoptotic activity[32]. The protein corresponding to the next TCN is ALDH1L1 which in the normal cell controls the cell mobility, but in the disease state is silenced thereby making way for uncontrolled proliferation of cells[33]. The next two proteins CACNB1 and CACNB2 help in calcium transport in the normal state but are down-regulated in the breast cancer cell, thereby affecting the calcium metabolism which causes poor signaling of messages in the cell[34]. Mutation in the last TCN KCNB2, results in concomitant cell proliferation[35].

The second localized eigenvector consists of six TCNs, among which KCNB2 and CACNB2 have been discussed above. The third protein KCNH4 in a normal cell is responsible for the potassium transport but undergoes splicing and is silenced in the disease state, thereby hampering early event in apoptosis leading to prolonged survival of cells[36]. The next protein TEX1, under normal condition, plays a major role in transcription regulation, but in the breast cancer cells, is over expressed which causes down-regulation of transcription[37]. The next TCN, KIAA1, is a binding protein to RNA but in cancer state undergoes mutation and refrains from its function of binding to RNA resulting in hampering of proper translation in cancer[38]. The next protein THOC1 is a part of TREX complex which is responsible for regulating transcription in a normal cell but in breast cancer cell, it delays transcriptional expression leading to irregular central dogma which makes the cell unorganized[37].

There are seven TCNs in third most localized eigenvector. The first two proteins, KLK5 and KLK10, are from Kellikein gene family which are generally involved in serine-type peptidase activities. In
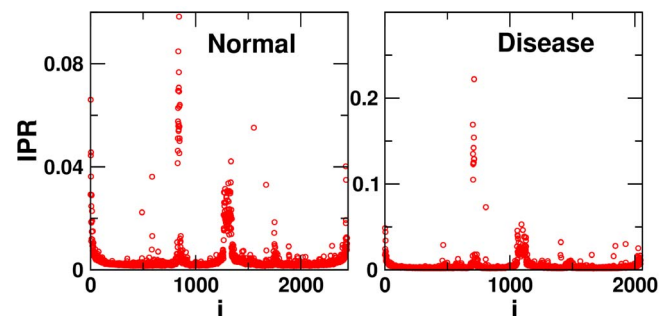

Figure 4 | Long-range Correlations (Δ_3 statistics) for the normal and the disease network. Circles denote data points for the normal and the disease networks whereas the solid line is the Δ_3 statistics for the GOE.

Figure 5 | Eigenvector Localization both normal and disease network. They clearly reflect three regions (i) degenerate part in the middle, (ii) a large non-degenerate part which follow GOE statistics of RMT and non-degenerate part at both the end and near to the zero eigenvalues which deviate from RMT.

Table 2 | Eigenvector localization properties. Top most localized eigenvectors (E^k) for the disease dataset, their top contributing proteins and network parameters namely degree and clustering coefficient. The betweenness centrality of all the TCNs in the disease network is zero

E^k	TCN	k	CC
714	MTHFD1L, ALDH1L1, CACNB1, CACNB2, KCNB2	2, 2, 3, 3, 5	1, 1, 0.66, 0.66, 1
704	KCNB2, KCNH4, TEX1, CACNB2, KIAA1, THOC1	5, 5, 6, 3, 3, 3	1, 1, 1, 1, 1, 1
712	KLK5, KLK10, CNTN4, ASTN2, BRMS1L, ARID4B, KCNQ5	3, 3, 2, 2, 3, 3, 5	1, 1, 1, 1, 1, 1, 1
710	THOC1, BRMS1L, KCND3, KCNB2, KCNQ5, ARID4B, TEX1, KCNH4	6, 3, 5, 5, 5, 6, 5 3	1, 1, 1, 1, 1, 1, 1, 1
705	KCND3, FCRL3, FCRL5, TEX1, THOC6, KCNH4, KCNQ5, MTFHD1L	5, 2, 2, 6, 6, 5, 5, 2	1, 1, 1, 1, 1, 1, 1, 1

the breast cancer they are down-regulated leading to the suppression of tumorigenesis[39]. The next protein CNTN4 is implicated in nervous system development. As of now its exact function in breast cancer is not known, but its first interacting neighbor is BRCA1 which is potent candidate in breast cancer cells[40]. Similarly, the next contributing node, ASTN2, under normal condition controls neural migration but in the breast cancer is found to have undergone chromosomal rearrangement with PTPRG which is an important gene for the recognition of the cancerous state in the cell[41]. The next protein, BRMS1L, is found to reduce expression of mRNA in breast cancer[42]. Another protein, ARID4B, functions in diverse cellular processes including proliferation, differentiation, apoptosis, oncogenesis, and cell fate determination. In the disease state, it is found to cause irregular cell formation and proliferation[43]. KCNQ5 is another TCN which is a family member of KCN which affects the cell proliferation of the breast cell[35].

The fourth localized eigenvector has eight TCNs of which KCND3 elevates the influx of potassium ions in breast cancer cells[44]. The others appear in the three most localized eigenvector and have been already discussed. Among the top contribution nodes of the fifth localized eigenvector, except THOC6, FCRL3 and FCRL5, all other have been discussed above in different localized vector. These three proteins have been found in both normal and breast cell. THOC6, in normal cell accounts for negative regulation of apoptosis, whereas in the breast cancer cell have been found to be silenced[37]. Next two proteins FCRL3 and FCRL5 appear as a part of FCRL complex which under normal circumstances act as an adapter of protein as well as development in immunity. In breast cancer, they have been found to be over expressed in immune cells namely WBC, thereby making it more robust against treatments[45].

What follows that all the proteins corresponding to the TCNs, except few, in the five most localized eigenvectors have a major contribution in promoting breast cancer. Moreover, fourteen common proteins are important for the normal and the disease both but leading to a very different behavior of cell in the two states. Whereas for normal cell, these common proteins are involved in major functioning of the cell (Supplementary), in the disease cell they all are found to be abnormally expressed or mutated leading to the disease state.

Preserved structures in localized nodes. The TCNs, in addition to the functional importance pertaining to the occurrence of the disease state revealed, exhibits interesting structural properties. This is more remarkable in the light that all of these TCNs lie in the low degree regime in the networks. Moreover, their betweenness centrality also are zero further ruling out any trivial structural significance of these nodes. But importance of these nodes based on the analysis of their interactions reveals the existence of preserved local structural patterns. Most strikingly, all of them follow phenomenon of gene duplication[46], as depicted in the Figure 6, which shows TCNs being involved in the pair formation in which first node in each pair has exactly the same neighbors as of the second node. Most remarkably, there are 20 duplicates (proteins having the same number of neighbors and having more than one connection) in the whole

network of which 18 are found in the TCNs of the most localized eigenvectors.

Further insight to this local structure is surfaced when we analyze the interaction patterns of these proteins in the normal breast network. What comes out from the analysis of the local interaction patterns of TCNs in the normal to those of disease is that there is either addition(s) of new interaction in the disease state in order to build clique of order three, or preservation of clique structure from normal to the disease or removal of interactions while in the normal keeping the clique intact. For example, TEX1 and THOC6, retain the same clique structures in both the networks (Figure 6). There are other proteins, for example BRMS1L and ARID4B which shed off some connections from the normal network (Figure 6(right)) while keeping clique structure intact in the disease state (Figure 6(left)). Further, nodes KLK5 and KCND3 form new connection in the disease state yielding the clique structures (Figure 6(left)). The fact that in spite of less connections of the proteins in the disease state as compared to the normal as well as reduction of cliques of order three from the normal to the disease, cliques in TCNs are preserved. This may be one of the reason of the poor performance of drugs targeting these proteins, as cliques are known to be the building blocks of a system making the underlying system robust against external perturbations. While the less number of nodes with $CC = 1$ as well as less number of clique of order three in the overall disease network from the normal one implicate that there is a destruction of building block, the conservation or addition of cliques of order three in TCNs (whose function importance for the occurrence of disease has already been emphasized) reflects that mutation in the disease cell makes the proteins to form a stable structure.

Conclusions

We construct and investigate the breast cancer network under the RMT framework. The analysis reveals that the TCNs of the most localized eigenvectors, despite of lying at low degree regime and having zero betweenness centrality, exhibit structural significance. All of them form pairs possessing common neighbors. Most remarkably, in the disease network there are 20 duplicate proteins with connections more than one, and out of which 18 appear in the localized eigenvector. This interesting revelation turns out to be more intriguing in the light of the functional analysis of proteins corresponding to the TCNs which clearly confirms their functional importance for the occurrence of the disease state.

Furthermore, clique structure and duplication of genes, which have been emphasized important for robustness as well as the evolution of a system, are found to be crucial for the disease state. Most striking revelation is that while there is an overall reduction in cliques in the disease network from the normal one reflecting reduction in the building blocks for the disease state, conservation or formation of cliques involving important disease proteins reflects that robustness of the overall system is decreased in the disease but the interactions of the important proteins involved in promoting the disease are preserved and might be one of the reasons behind making those pathways involved with the important proteins highly resistant to various treatments.
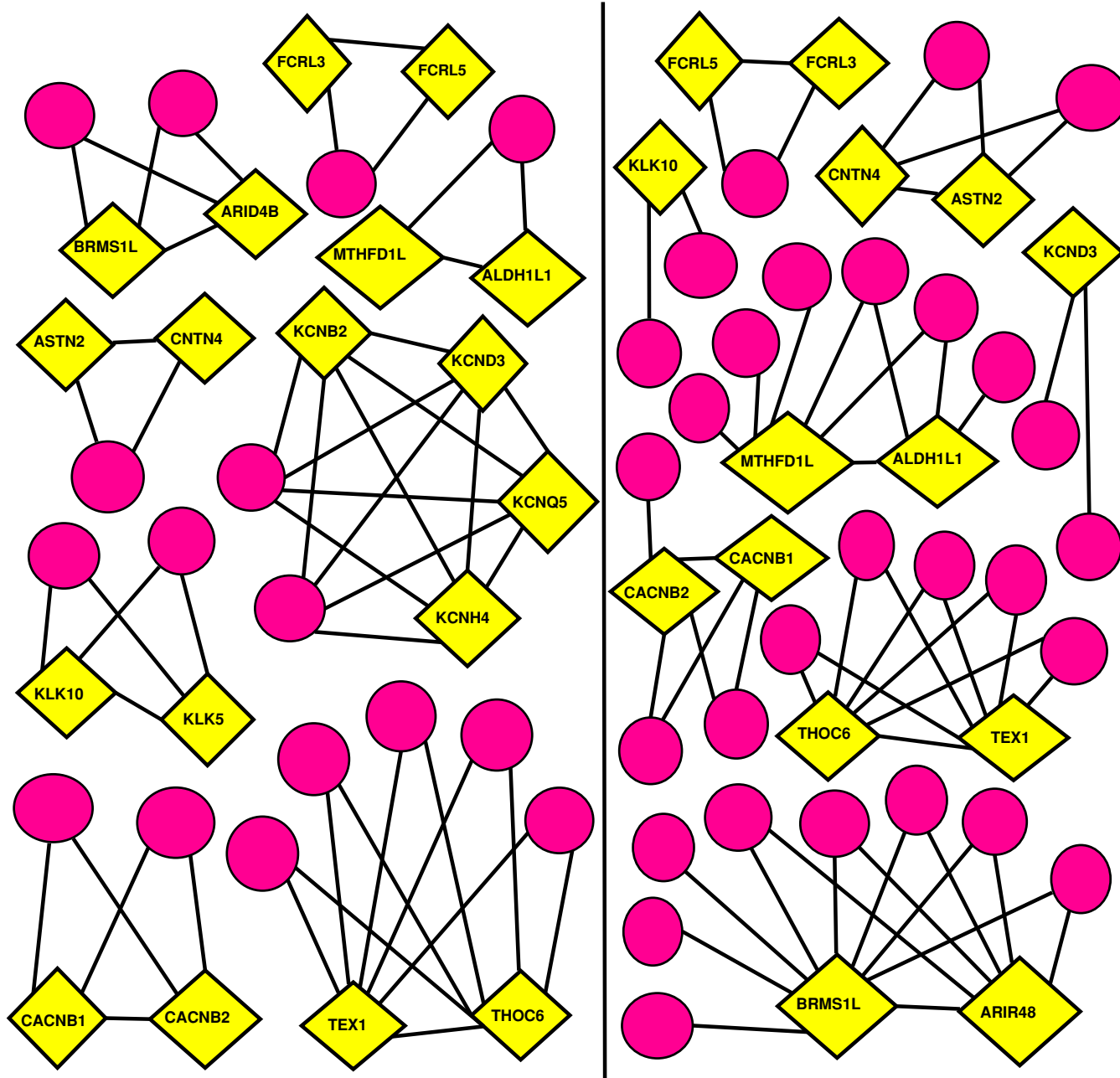
Figure 6 | Local structure of top contributing nodes. (Left panel) The local structure of all TCNs in the disease network. (Right panel) The local structure for the same proteins in the normal network. Yellow represents TCNs and pink represent their first neighbor.

The Δ_3 statistics demonstrates less *randomness* in the disease state than the normal one, which might be arising due to the removal of random connections in the disease state, and not because of a probable enhancement in the modular structure as number of nodes having $CC = 1$ is indeed less in case of the disease state than the normal one. This in-turn depicts that randomness leads to the robustness of the system where, normal breast network is found to be more robust than the disease state.

Detection of important proteins involved in breast cancer using RMT platform provides a time efficient and cost effective approach for those diseases which lack in-depth information about important genes. Revelation of clique structure, being formed or preserved by these proteins, provides a further bench mark for designing drugs which can target a sub-graph instead of the individual protein. This analysis presented here can be extended further to study other diseases like other types of cancers, diabetes, hypertension etc to predict

various structural and functional aspect of biology[47], which in addition may help to compose novel drug targets and to introduce the concept of single medicine for multiple diseases[48].

Methods

Data assimilation and network construction. According to the network theory, there are two basic components of network namely nodes and edges. Here we study the PPI network where nodes are the proteins and edges denote the interactions. After diligent and enormous efforts, we collect the protein interaction data from various literature and bioinformatic sources. To keep the authenticity of the data we only take the proteins into account which are reviewed and cited. We use various different bioinformatic databases namely Gen bank from NCBI and UNIPROT[49,50], constituting data available from other resources like European Bioinformatic Institute, the Swiss Institute of Bioinformatics, and the Protein Information Resource. To add more information we also take the most widely studied normal and breast cancer cell lines whose protein expression data is known. There are numerous cell lines available but a very few have been exploited for their maximum proteomic insight. Here, we are use the data of HMEC cell line for normal[51,52] and MFC-7 for the

breast cancer network[53]. The collection and discussion to select the datasets and authenticating this data for RMT analysis is an extensive job and we generate this data after thorough literature search and validation in about 400 hrs of rigorous study. After collecting the proteins for both the datasets, their interacting partners are downloaded from STRING database[54]. The dataset contains 2464 nodes and 15131 connections for normal network followed by 2096 nodes and 14183 links for the breast cancer network. The networks turned to have one big cluster and several small clusters. We then investigate the structural properties of the network.

Structural measures. Several statistical measures are proposed to understand specific features of the network[55]. First we define the interaction matrix or the adjacency matrix of the network as follows:

$$A_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The most basic structural parameter of a network is the degree of a node (k_i), which is defined as the number of neighbors of the node has ($k_i = \sum_j A_{ij}$). Degree distribution $\rho(k)$, revealing the fraction of vertices with the degree k, is known as the fingerprint of the network. Another important parameter is the clustering coefficient (CC) of the network. Clustering is defined as the ratio of the number of connections a particular node is having by the possible number of connections the particular node can have. These are also known as cliques. Clustering coefficient of a network can be written as

$$CC = \frac{1}{n} \sum_{i=1}^{n} C_i \quad (2)$$

They are complete sub graphs in the network which are known to be the conserved part of the network[21]. The average clustering coefficient of the network characterizes the overall tendency of nodes to form cluster or groups[55]. Further, the betweenness centrality of a node i is defined as the fraction of shortest paths between node pairs that pass through the said node of interest[56]

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}, \quad (3)$$

where n_{st}^i is the number of paths from s to t that passes through i and g_{st} is the total number of paths from s to t in the network. Another parameter is the diameter of the network which measures the longest of the shortest path between the two nodes.

Spectral techniques. The random matrix analysis of the eigenvalue spectra considers (1) global properties such as spectral distribution of eigenvalues $\rho(\lambda)$, and (2) local properties such as eigenvalue fluctuations around $\rho(\lambda)$. We denote the eigenvalues of a network by $\lambda_i = 1, \dots, N$ and $\lambda_1 > \lambda_2 > \lambda_3 > \dots > \lambda_N$. The nearest neighbor spacing distribution (NNSD) has been known to be one of the most powerful technique in RMT and we analyze them by calculating the distribution of ratio of the consecutive eigenvalues which is represented as[57]

$$P_W(r) = \frac{1}{Z_\beta} \frac{(r+r^2)^\beta}{(1+r+r^2)^{(1+3/2\beta)}} \quad (4)$$

The benefit of analyzing ratio of nearest spacings over much used NNSD[58] is that this method does not require unfolding of the eigenvalues. Further, the NNSD accounts only for the short range correlations in the eigenvalues. We probe for the long range correlations in eigenvalues using $\Delta_3(L)$ statistics which measures the least-square deviation of the spectral staircase function representing average integrated eigenvalue density $N(\bar{\lambda})$ from the best fitted straight line for a finite interval of length L of the spectrum. In order to get the universal properties of eigenvalues through h Δ_3 statistics, it is customary in RMT to unfold eigenvalues by a transformation $\bar{\lambda}_i = \bar{N}(\lambda_i)$, where $\bar{N}$ is the average integrated eigenvalue density. Since we do not have any analytical form for N, we numerically unfold the spectrum by polynomial curve fitting. After unfolding, average spacings are unity, independent of the system, which in the absence of any analytical form of polynomial fitting of the eigenvalues works with the approximate numerical fitting[58] and is given by

$$\Delta_3(L; x) = \frac{1}{L} \min_{a,b} \int_x^{x+L} \left[N(\bar{\lambda}) - a\bar{\lambda} - b \right]^2 d\bar{\lambda} \quad (5)$$

where a and b are regression coefficients obtained after least square fit[58]. Average over several choices of x gives the spectral rigidity, the $\Delta_3(L)$. In case of GOE statistics, the $\Delta_3(L)$ depends logarithmically on L, i.e.

$$\Delta_3(L) \sim \frac{1}{\pi^2} \ln L \quad (6)$$

Further, we use the inverse participation ratio (IPR) to analyze localization properties of the eigenvectors[59]. For E_l^k denoting lth component of kth eigenvector E^k, the IPR of an eigenvector can be defined as

$$I^k = \frac{\sum_{l=1}^{N} [E_l^k]^4}{\left(\sum_{l=1}^{N} [E_l^k]^2 \right)^2} \quad (7)$$

which shows two limiting values : (i) a vector with identical components $E_l^k \equiv 1/\sqrt{N}$ has $I^k = 1/N$, whereas (ii) a vector, with one component $E_1^k = 1$ and the remainders zero, has $I^k = 1$. Thus, the IPR quantifies the reciprocal of the number of eigenvector components that contribute significantly. We further calculate the average IPR in order to measure an overall localization of the network calculated as[60],

$$\langle IPR \rangle = \frac{\sum_{l=1}^{N} [I^k]}{N} \quad (8)$$

Note that IPR defined as above separates out the TCNs by keeping the threshold as 1/IPR.

1. Venter, J. C. et al. The Sequence of the Human Genome. *Science* **16**, 1304–1351 (2001).
2. Ferlay, J., Parkin, D. M. & Fouchera, E. S. Estimates of cancer incidence and mortality in Europe in 2008. *EJC* **46**, 765–781 (2010).
3. Kiberstis, P. & Roberts, L. A Race Still Unfinished. *Science* **343**, 1451–1470 (2014).
4. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
5. Petricoin, E. F. et al. Mapping Molecular Networks Using Proteomics: A Vision for Patient-Tailored Combination Therapy. *J. Clin. Oncol.* **23**, 3614–3621 (2005).
6. Barabasi, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
7. Hartwell, L. H. et al. From molecular to modular cell biology. *Nature* **402**, 47–52 (1999).
8. Kitano, H. Systems Biology: A Brief Overview. *Science* **295**, 1662–1664 (2002).
9. Zhu, X., Gerstein, M. & Snyder, M. Getting connected: analysis and principles of biological networks. *Genes & Dev.* **21**, 1010–1024 (2007).
10. Barabasi, A. L. & Otavi, Z. N. Network biology: understanding the cells functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
11. Goh, K-II. et al. The human disease network. *PNAS* **104**, 8685–8690 (2008).
12. Pujana, M. A. et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genet.* **39**, 1338–1349 (2007).
13. Chuang, H. Y. et al. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007).
14. Wigner, E. P. Characteristic vectors of bordered matrices with infinite dimentions. *Ann. Math.* **62**, 548–564 (1955).
15. Papenbrock, T. & Weidenmüller, H. A. Random matrices and chaos in nuclear spectra. *Rev. Mod. Phys.* **79**, 997–1013 (2007).
16. Chung, F., Linyuan, L. & Van, V. Spectra of random graphs with given expected degrees. *PNAS* **100**, 6313–6318 (2003).
17. Friedel, C. C. & Zimmer, R. Influence of degree correlations on network structure and stability in protein-protein interaction networks. *BMC Bioinformatics* **8**, 297 (2007).
18. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
19. Alon, U. *An introduction to Systems Biology: Design principles of Biological Circuits* (Chapman and Hall/CRC, London, 2006).
20. Dwivedi, S. K. & Jalan, S. Emergence of clustering: Role of inhibition. *arXiv*, 1405.2413 (2014).
21. Lotem, E. Y. et al. Network motifs in integrated cellular networks of transcriptionregulation and proteinprotein interaction. *PNAS* **101**, 5934–5939 (2004).
22. Bandyopadhyay, J. N. & Jalan, S. Universality in complex networks: Random matrix analysis. *Phys. Rev. E* **76**, 026109 (2007).
23. Aguiar, M. A. M. de. & Bar-Yam, Y. Spectral analysis and the dynamic response of complex networks. *Phys.Rev.E* **71**, 016106 (2005).
24. Clancy, S. Genetic mutation. *Nat. Educat.* **1**, 187 (2008).
25. Atay, F. M., Jalan, S. & Jost, J. Randomness, chaos, and structure. *Complexity* **15**, 29–35 (2009).
26. Jalan, S. & Bandyopadhyay, J. N. Randomness of random networks: A random matrix analysis. *EPL* **87**, 48010 (2009).
27. Ravsaz, E. Hierarchical Organization of Modularity in Metabolic Networks. *et. al.*, *Science* **297**, 1551–1555 (2002).
28. Cohen, J. D. & Tong, F. The Face of Controversy. *Science* **293**, 2405–2407 (2001).
29. Zyczkowski, K. [Quantum mechanical eigenvector statistics of chaotic systems]. [153–168] *Quantum Chaos* [Cerderia, H. A., Ramaswami, R., Gutzwiller, M. C. & Casati, G. (ed.)] (World Scientific, Singapore, 1991).
30. Jalan, S. et al. Spectral analysis of gene co-expression network of Zebrafish. *EPL* **99**, 48004 (2012).
31. Plerou, V. et al Random matrix approach to cross correlations in financial data. *Phys. Rev. E* **65**, 066126 (2002).
32. Minguzzi, S. et al. An NTD-Associated Polymorphism in the 3 UTR of MTHFD1L can Affect Disease Risk by Altering miRNA Binding. *Hum. Mutat.* **35**, 96–104 (2014).
33. Ginestier, C. et al. ALDH1 Is a Marker of Normal and Malignant Human Mammary Stem Cells and a Predictor of Poor Clinical Outcome. *Cell Stem Cell* **1**, 555–567 (2007).
34. Wuchty, S., Arjona, D. & Bauer, P. O. Important miRs of Pathways in Different Tumor Types. *PLoS Comput. Biol.* **9**, e1002883 (2013).

35. Choi, M. *et al.* K+ Channel Mutations in Adrenal Aldosterone-Producing Adenomas and Hereditary Hypertension. *Science* **331**, 768–772 (2011).

36. Moore, M. J. & Silver, P. A. Global analysis of mRNA splicing. *RNA* **14**, 197–203 (2008).

37. Chi, B. *et al.* Aly and THO are required for assembly of the human TREX complex and association of TREX components with the spliced mRNA. *Nucleic Acids Res.* **41**, 1294–1306 (2013).

38. King, M. C. & Welch, P. L. A novel functional screen for new breast cancer genes. *Gynecol. Oncol.* **98**, 68 (2005).

39. Yousef, G. M. *et al.* Kallikrein gene downregulation in breast cancer. *Br. J. Cancer* **90**, 167–172 (2004).

40. McClellan, J. & King, M. C. Genetic Heterogeneity in Human Disease. *Cell* **141**, 210–217 (2010).

41. Hampton, O. A. *et al.* A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res.* **19**, 167–177 (2009).

42. Stark, A. M. *et al.* Reduced metastasis-suppressor gene mRNA-expression in breast cancer brain metastases. *J Cancer Res Clin Oncol.* **131**, 191–198 (2005).

43. Wu, M. Y., Eldin, K. W. & Beaudet, A. L. Identification of Chromatin Remodeling Genes Arid4a and Arid4b as Leukemia Suppressor Genes. *JNCI* **100**, 1247–1259 (2008).

44. Ko, J. H. *et al.* Expression profiling of ion channel genes predicts clinical outcome in breast cancer. *Mol. Cancer* **12**, 106 (2013).

45. Inozume, T. *et al.* Novel melanoma antigen, FCRL/FREB, identified by cDNA profile comparison using DNA chip are immunogenic in multiple melanoma patients. *Int. J. Cancer* **114**, 283–290 (2005).

46. Bergthorsson, U., Andersson, D. I. & Roth, J. R. Ohno's dilemma: Evolution of new genes under continuous selection. *PNAS* **104**, 17004–17009 (2007).

47. Li, Y. & Agarwal, P. A Pathway-Based View of Human Diseases and Disease Relationships. *PLoS ONE* **4**, e4346 (2009).

48. Yildirim, M. A. *et al.* Drugtarget network. *Nat. Biotechnol.* **25**, 1119–1126 (2007).

49. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **40**, D48–D53 (2012) (http://www.ncbi.nlm.nih.gov/Date of access: 15/02/2014).

50. Bairoch, A. *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2005) (http://www.uniprot.org/Date of access: 17/02/2014).

51. Page, M. J. *et al.* Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties. *PNAS* **96**, 12589–12594 (1999).

52. Patwardhan, A. J. *et al.* Comparison of Normal and Breast Cancer Cell Lines Using Proteome, Genome, and Interactome Data. *J. Proteome Res.* **4**, 1952–1960 (2005).

53. Kulasingam, V. & Diamandis, E. P. Proteomics Analysis of Conditioned Media from Three Breast Cancer Cell Lines. *Mol. Cel. Proteomics* **6**, 1997–2011 (2007).

54. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011) (http://string-db.org/ Date of access: 09/03/2014)

55. Boccaletti, S. *et al.* Complex networks: Structure and dynamics. *Phys. Rep.* **424**, 175–308 (2006) and references therein.

56. Newman, M. E. J. The Structure and Function of Complex Networks. *SIAM Review.* **45**, 167–256 (2003).

57. Atas, Y. Y. *et al.* Distribution of the Ratio of Consecutive Level Spacings in Random Matrix Ensembles. *Phys Rev. Lett.* **110**, 084101 (2013).

58. Mehta, M. L. *Random Matrices* 1st edition (Elsevier, San Diego, 2004).

59. Haake, F. & Życzkowski, K. Random-matrix theory and eigenmodes of dynamical systems. *Phys. Rev. A* **42**, 1012–1016 (1990).

60. Jalan, S., Zhu, G. & Li, B. Spectral properties of directed random networks with modular structure. *Phys. Rev. E* **84**, 046107 (2011).

Acknowledgments

Author contributions

S.J. conceived and supervised the project. A.R. and A.V.M. collected the data. S.J. and A.R. performed the analysis and wrote the manuscript.

Additional information