Emerson et al. Journal of Clinical Bioinformatics (2015) 5:1
DOI 10.1186/s13336-015-0016-6

JCBi JOURNAL OF CLINICAL BIOINFORMATICS

RESEARCH Open Access

K-core decomposition of a protein domain co-occurrence network reveals lower cancer mutation rates for interior cores

Arnold I Emerson[1,2], Simeon Andrews[1,2], Ikhlak Ahmed[1,2], Thasni KA Azis[1,2] and Joel A Malek[1,2*]

Abstract

Background: Network biology currently focuses primarily on metabolic pathways, gene regulatory, and protein-protein interaction networks. While these approaches have yielded critical information, alternative methods to network analysis will offer new perspectives on biological information. A little explored area is the interactions between domains that can be captured using domain co-occurrence networks (DCN). A DCN can be used to study the function and interaction of proteins by representing protein domains and their co-existence in genes and by mapping cancer mutations to the individual protein domains to identify signals.

Results: The domain co-occurrence network was constructed for the human proteome based on PFAM domains in proteins. Highly connected domains in the central cores were identified using the k-core decomposition technique. Here we show that these domains were found to be more evolutionarily conserved than the peripheral domains. The somatic mutations for ovarian, breast and prostate cancer diseases were obtained from the TCGA database. We mapped the somatic mutations to the individual protein domains and the local false discovery rate was used to identify significantly mutated domains in each cancer type. Significantly mutated domains were found to be enriched in cancer disease pathways. However, we found that the inner cores of the DCN did not contain any of the significantly mutated domains. We observed that the inner core protein domains are highly conserved and these domains co-exist in large numbers with other protein domains.

Conclusion: Mutations and domain co-occurrence networks provide a framework for understanding hierarchal designs in protein function from a network perspective. This study provides evidence that a majority of protein domains in the inner core of the DCN have a lower mutation frequency and that protein domains present in the peripheral regions of the k-core contribute more heavily to the disease. These findings may contribute further to drug development.

Keywords: Domain co-occurrence network, K-core decomposition, Somatic mutations, Cancer, Cancer mutations, TCGA

Background

Domains are distinct functional or structural units in a protein. Most domains correspond to tertiary structure elements, and are able to fold independently. All protein domains exhibit evolutionary conservation and many either perform specific functions or contribute in a specific way to the structure of their proteins. Domains may exist in a variety of biological contexts, wherein similar domains can be found in proteins with different functions. Many proteins are composed of one or more domains that can fold independently into a stable core structure [1-3].

Many complex systems have been analyzed as networks by representing the system as nodes and interactions between them as edges. Studies on complex networks including the network of co-authorships, sexual contacts and the world-wide-web (WWW) reveal that their structure and growth is governed by a set of generic organizing principles [4,5]. Network biology is emerging as a new field in biology due to the increasing availability of genome-scale data sets of molecular interactions. These data are a result of new high-throughput technologies yielding information on protein interactions, regulatory

* Correspondence: jom2042@qatar-med.cornell.edu
[1]Department of Genetic Medicine, Weill Cornell Medical College, New York, NY, USA
[2]Genomic Core, Weill Cornell Medical College in Qatar, Qatar Foundation, Doha 24144, Qatar

networks and the metabolome. Biological systems like gene interaction networks, protein and metabolite networks have been found to exhibit a scale-free property [6-13].

The development of high-throughput, whole-exome/genome DNA sequencing has made it possible to evaluate normal and tumor tissue samples in a single study. These studies have revealed the connection between somatic mutations and cancer susceptibility, initiation and development [14]. A central goal of cancer genome analysis is the identification of cancer genes that, by definition, carry driver mutations. A key challenge will therefore be to distinguish driver from passenger mutations. Most studies thus far have attempted to identify driver mutations using gene-centric approaches [15-20]. Unfortunately, this method is limited to a small subset of genes and also leads to mischaracterized mutations [21]. The gene-based approach usually fails to reflect the position of mutation or the functional context the position of mutation provides in protein level. But a protein domain network enables the identification of mutations that are rare at the gene level, but that occur frequently within the specified domain. These highly mutated domains potentially reveal disruptions of protein function necessary for cancer development.

Several studies have been conducted on protein domain co-occurrence networks (DCN). These studies represent domains as nodes and their co-occurrence in a protein are denoted as edges. The networks have also been shown to possess a scale-free property [22-24]. Increasing complexity of the organisms were observed from bacteria to eukaryotes due to the links involved in the cell-cell interaction domains, signal transduction and cell differentiation domains. Studies on DCN have examined the network property [22], evolutionary traces among the species [25], architectural design of protein domain networks [26] and mapping somatic mutations to protein domains in colon cancer [27]. More recently, a disease-drug-phenotype matrix was also analyzed using protein domain networks [28]. However, each of these studies have focused either on domain co-occurrence networks or on a specific feature of the DCN and therefore, do not provide a generalized view of mutations in the domain co-occurrence network. In this study, we investigated the protein domain co-occurrence network in the context of various cancers and their mutations. We specifically focused on the highly connected protein domains of the DCN core by using k-core decomposition techniques.

The definition of k-core was first introduced by Seidman [29] to characterize the cohesive regions of graphs. Batagelj et al., developed an efficient algorithm to find the k-core decomposition of a graph [30]. K-core technique has been used in many areas including the alternative method for community detection algorithm [31] and for the identification of dense components in most of the complex networks [32-34].

K-core decomposition is a network analysis approach that helps in understanding interesting structural properties that are not otherwise captured by many other network topological parameters. The basic principle behind the k-core is decomposition to identify particular subsets of the network called k-cores. Each k-core is obtained by a recursive pruning method [29,35,36]. This decomposition method allows the study of the hierarchical properties of large complex networks by focusing on the network centrality and connectedness properties of nodes. The central cores of this analysis have more strongly connected vertices with large number of possible distinct paths between them. This helps in obtaining robust routing properties.

Materials and methods

Construction of protein domain co-occurrence networks (DCN)

The DCN for *Homo sapiens* was constructed using the Ensembl database (version 72) that provides a comprehensive source of stable automatic annotation of individual genomes [37].

To ensure correct coordinates in our version domains present in any protein sequence of the human proteome were determined using the program PfamScan [38]. Domain hits with an e-value < =0.01 were considered for constructing a DCN. Each domain was represented as a node and if every two domains co-exist in one protein then they were connected by edges as shown in Figure 1 [25]. Figure 2 illustrates the largest sub-graph for *H. sapiens* consisting of 1929 nodes and 5171 edges as visualized by Cytoscape [39].

K-core decomposition of DCN

To identify the core-periphery organization of the domain co-occurrence network, we subjected it to core decomposition. The cores of different orders of a network can be obtained by iteratively removing all nodes which have less than k connections with other protein domains ($k = 1, 2, \ldots$). This is done by first identifying all nodes whose degree (i.e., number of connections) is less than k. After removing these, the network is reanalyzed to determine if the removal of these nodes has resulted in other nodes (which originally had degree > k) having now less than k connections. If such nodes are identified, then they are removed, and the process is continued, until no more nodes can be removed. The resulting sub-network is called the k-core of the network.

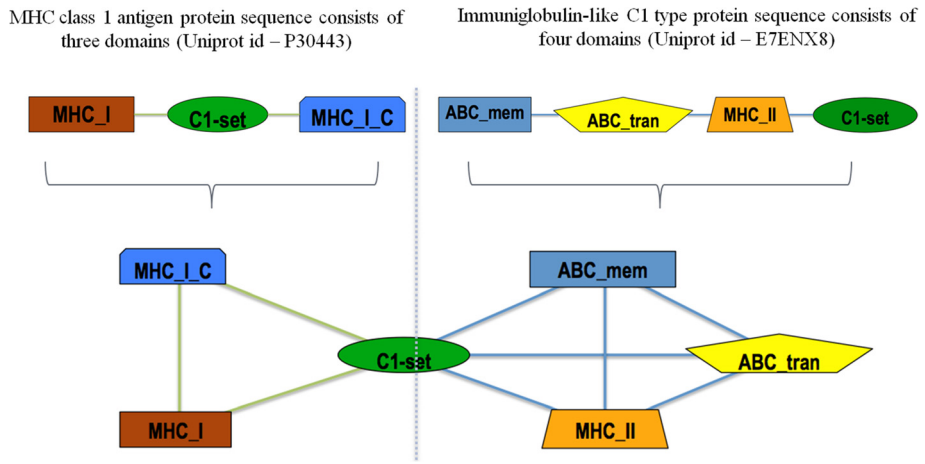

Figure 1 A small DCN consisting of two *H. sapiens* proteins. The two proteins share the same domain C1-set. An edge is drawn between domains co-occurring in the same protein and a fully connected sub-graph (clique) in the DCN corresponds to a protein.

Randomization of k-core

To determine the statistical significance of the properties calculated for members of an empirically determined *k*-core, we compared them with the mean and variance of the corresponding values obtained for a randomized ensemble. Each randomized *k*-core in the ensemble is obtained by random selection without replacement of N_k domains from the DCN, where N_k is the size of the empirically determined *k*-core. The randomized ensemble for every DCN considered was generated by constructing 100 such randomized *k*-cores.

Evolutionary conservation of protein domains

The evolutionary conserved protein domains were identified using the database PANDIT*plus* [40]. It consists of a database of PFAM alignment phylogenetic trees for known protein domains and their families. This database was constructed using a relational database which comprises of information regarding the functional categories, metabolic pathways, protein–protein interactions, disease associations, gene expressions, three-dimensional structures, as well as estimates from an evolutionary analyses of selective pressures.

Cancer mutation dataset

Somatic mutation data for ovarian, breast and prostate cancer were obtained from TCGA data portal (http://tcga-data.nci.nih.gov/tcga/) using mutation files from the hgsc.bcm.edu_COAD.IlluminaGA_DNASeq.1 and hgsc.bcm.edu_COAD.SOLiD_DNASeq.1 directories downloaded on March 30th, 2013. The silent and RNA mutations were filtered out from the data set as they were assumed unlikely to affect the cancer development. Somatic mutation counts for ovarian cancer were found to be 20,878. For breast and prostate cancer the values were found to be 35,558 and 23,349.

Mapping cancer SNPs to individual protein domains

Before mapping mutations to the individual protein domains, the protein domain positions need to be converted into their chromosome positions. Mutations obtained from TCGA data portal were reported with genomic locations while predicted PFAM domains documented in peptide coordinates. A Perl program was written using the ensemble Perl API module for converting the protein domain positions into chromosome positions. The Pfam domains from *Homo sapiens* were successfully mapped with the chromosome positions as shown in the Additional file 1: Table S1. For ovarian cancer sample set, almost one third of the mutations (30%) occurred inside annotated protein domain regions. Similarly 47.5% and 49.3% of all mutations in breast and prostate cancer sample sets were observed to have occurred inside the protein domain space (Table 1).

Procedure for normalizing the domain mutation frequency

To determine the domains that are frequently mutated in the human genome, we first obtained the count of mutations that fell within each domain. Since larger domains are generally expected to accumulate more mutations than the shorter domains, we normalized the domain mutation counts with domain length. This was done by dividing domain mutations counts by the cumulative length of the domain in the genome. That is, the summed length of all occurrences of the domain in the genome was used as total length. The normalized score for all the three cancer types are shown in Additional file 2: Table S2.

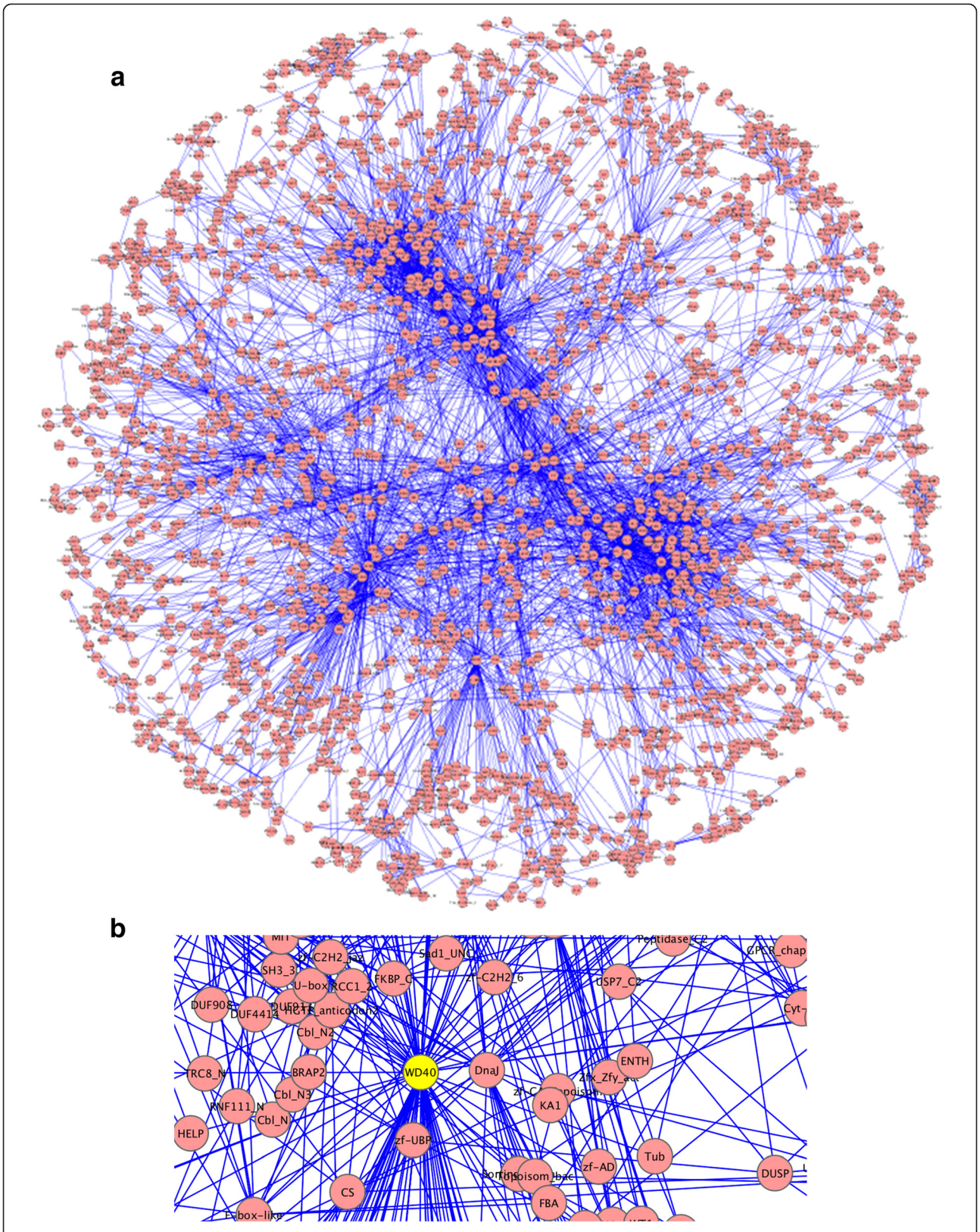

Figure 2 Domain co-occurrence network of *H. sapiens*. **(a)** The largest DCN sub-graph or the main graph of *H. -sapiens*, which consists of 1929 nodes and 5171 edges. **(b)** The enlarged partial view of *H. sapiens* main DCN, in which the domain WD40 (beta-transducin repeat) is represented as a hub.

Table 1 Percentage of mapped mutations in three forms of cancer

S. No	Cancer type	No. of Mutations	No. of mapped mutations	Percentage
1	Ovarian	20,878	5,842	30%
2	Breast	35,558	16,887	47.5%
3	Prostate	23,349	11,502	49.3%

Calculation of significantly mutated domains

Domain mutation counts were normalized with the cumulative length of the domain in the genome. We then obtained the relative frequencies from the normalized values and these frequencies were used as success probability (p). This probability (p) was normalized using the signal to noise ratio of the Bernoulli distribution, which resulted in a normalized score, z and is given by

$$Z = p/sqrt(p(1-p))$$

The **"locfdr"** package [41] from R was used to estimate the null distribution and these statistics were used to identify domains with a local false discovery rate < 0.1 [42]. The local false discovery rate values for each domain in all the three cancer types are shown in Additional file 3: Table S3a (sheet1), Table S3b (sheet2) and Table S3c (sheet3).

Results

Domains in the inner cores are more conserved than those at the periphery

The domain co-occurrence network of *Homo sapiens* was constructed and its statistical properties were determined. From the degree distribution plot (Additional file 4: Figure S1a), the DCN was found to have scale-free behavior. Additional file 4: Figure S1b shows the shortest path length distribution exhibiting a small-world phenomenon. The average clustering coefficient distributions and the node degrees are found to have an inherent hierarchical modularity (Additional file 4: Figure S1c). We applied the k-core decomposition algorithm to the *Homo sapiens* DCN [29]. The cores were found to have 10 nested k-cores, where k values ranged from 1 to 10. The property of k-core decomposition is that as the core increases the number of nodes in each core decreases. This property was observed in the *Homo sapiens* DCN.

To differentiate the protein domains in each core, we first identified the conserved domains in each core using PANDIT server. To verify whether the frequency of conserved domains in the inner cores is statistically significant, the empirical values were compared against the randomized cores. The percentage of conserved domains increased with increasing core order in contrast to the random cores that did not show significant deviations with the core order (Figure 3). The deviation of the empirical data with core order is much greater than the error bars obtained from the random ensemble. This suggests that the highly conserved nature of the inner core domains is significant in the empirical DCN.

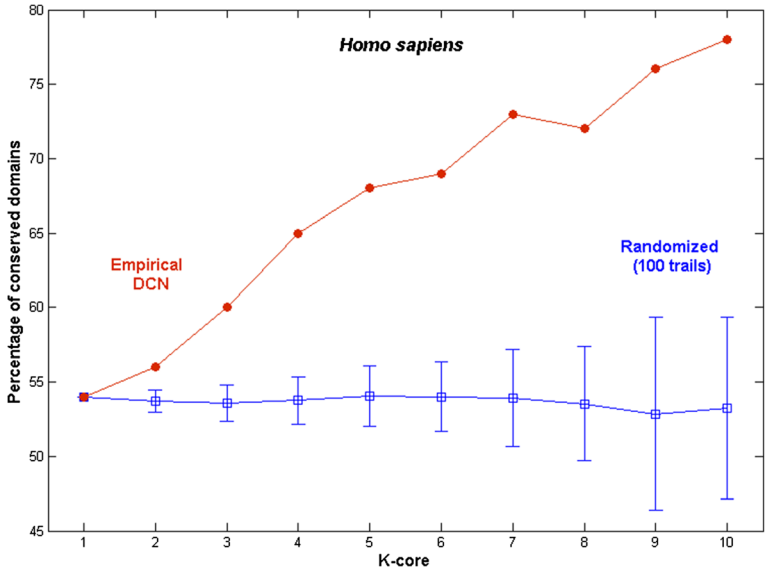

Figure 3 The percentage of conserved domains within each core order.

Domains in inner cores have fewer mutations than those at the periphery

Cancer mutations obtained from TCGA data portal were mapped to the individual protein domains. The normalized mutation score for each protein domain was also calculated (for details see Methods). To study the mutations in domain co-occurrence network, the normalized mutation scores were assigned to all the nodes (i.e., protein domain) in the network. In order to understand the nature of mutation in the *Homo sapiens* DCN we subjected it to core decomposition. We found that the normalized mutations per domain in each core gradually decreased with the core order. This observation occurred for all the three cancer types. To verify whether the findings are statistically significant, we compared the empirical DCN with its random counterparts. The pattern observed in the *Homo sapiens* DCN was not found in 100 random DCNs (i.e. $p < 0.01$).

As shown in Figure 4, the number of normalized mutations per domain corresponding to the random cores does not show any significant deviation with core order, unlike the case for the empirical DCN. The deviation of the empirical data with core order is much less than the error bars obtained from the random model. This suggests that the less mutated nature of the inner core domains is significant. A similar study shows that the domains in the inner cores of *S.cerevisiae*, *C.elegans, D. melanogaster, M.musculus and H. sapiens* have been preserved during evolution. This high conservation of inner core domains across species development may explain why they are also less mutated in comparison to the peripheral protein domains in cancer [25].

From the randomized simulations, we observed that the inner core domains had lesser rates of mutation compared to the peripheral cores. To verify if, the inner core significantly differed from the other cores (we wanted to investigate the extent to which this aspect was true and to also identify outliers), every domain's normalized mutation rates were plotted against the k-core values for the three types of cancers as shown in Additional file 5: Figure S2 (2a-ovarian cancer, 2b-breast cancer & 2c-prostate cancer). Results suggested that the normalized mutation rates gradually declined with the core order and the correlation values (R^2) between them were also found to be positive. Interestingly, the outlier of the inner core is found to be more significant as it comprises of lower mutation rates.

Identification of significant mutated domains

Significantly mutated domains in all of the three cancer types were identified using the local false discovery rate. On comparing the significantly mutated domains among the cancer types, we found several domains common between ovarian, breast and prostate cancer (Figure 5). The significantly different domains in all three cancers are tabulated in the Additional file 6: Table S4. Interestingly, we found that in all of the three cancer sets the P53 domain scores the highest number of mutations. Among all the cancer domains 11 pfam domains were commonly mutated (Table 2). To determine the functions overrepresented in our sets of significant mutated domains, we obtained the annotations for all the domains using DAVID [43,44]. A list of KEGG pathways and gene annotation terms from the enrichment analysis of significant domains for ovarian, breast and prostate cancer can be found in the Additional file 7: Table S5. A subsequent enrichment analysis of KEGG pathways annotated for significant PFAM domains in prostate cancer revealed pathways related to Toll-like receptor signaling, small cell lung cancer, complement and coagulation cascades, etc. A similar analysis of GO terms annotated for prostate cancer revealed an overrepresentation of GO terms related to death, development process, and cellular component organization among others. The complete set of KEGG pathway and GO term annotation for ovarian and breast cancer is tabulated in Additional file 7: Table S5a, b.

As discussed in the previous section, core domains were found to be less mutated in comparison to the peripheral domains. We also investigated the presence of significantly mutated domains in each core of the domain co-occurrence network. We calculated the percentages of significant domains in each core of the DCN as shown in Figure 6. Interestingly, we found that inner cores 8, 9 and 10 did not contain any significantly mutated domains. This indicates clearly that the inner core domains have been highly conserved through evolution and also less mutated in cancer. From the study done by Benjamin A. et. al., it was found that highly connected nodes in the domain interaction network had domains which were conserved and also involved in important biological roles within a cell [45].

Discussion

On analyzing genomes studies have shown that more than 70% of eukaryotic proteins comprised of multiple domains. Domain-domain interactions are now becoming an upcoming trend of interest across numerous studies [46-51]. Studies on protein-protein and domain-domain interaction networks using graph models have revealed that domain levels are the most important aspects of evolutionary selection. In addition to this, protein structural domains seem to have been the most distinct and significant biological entities for interaction, function and evolution [47]. Modeling of domain interaction networks have identified that domains are often involved in the propagation of signal transduction and helps determine the recognition specificity of each domain

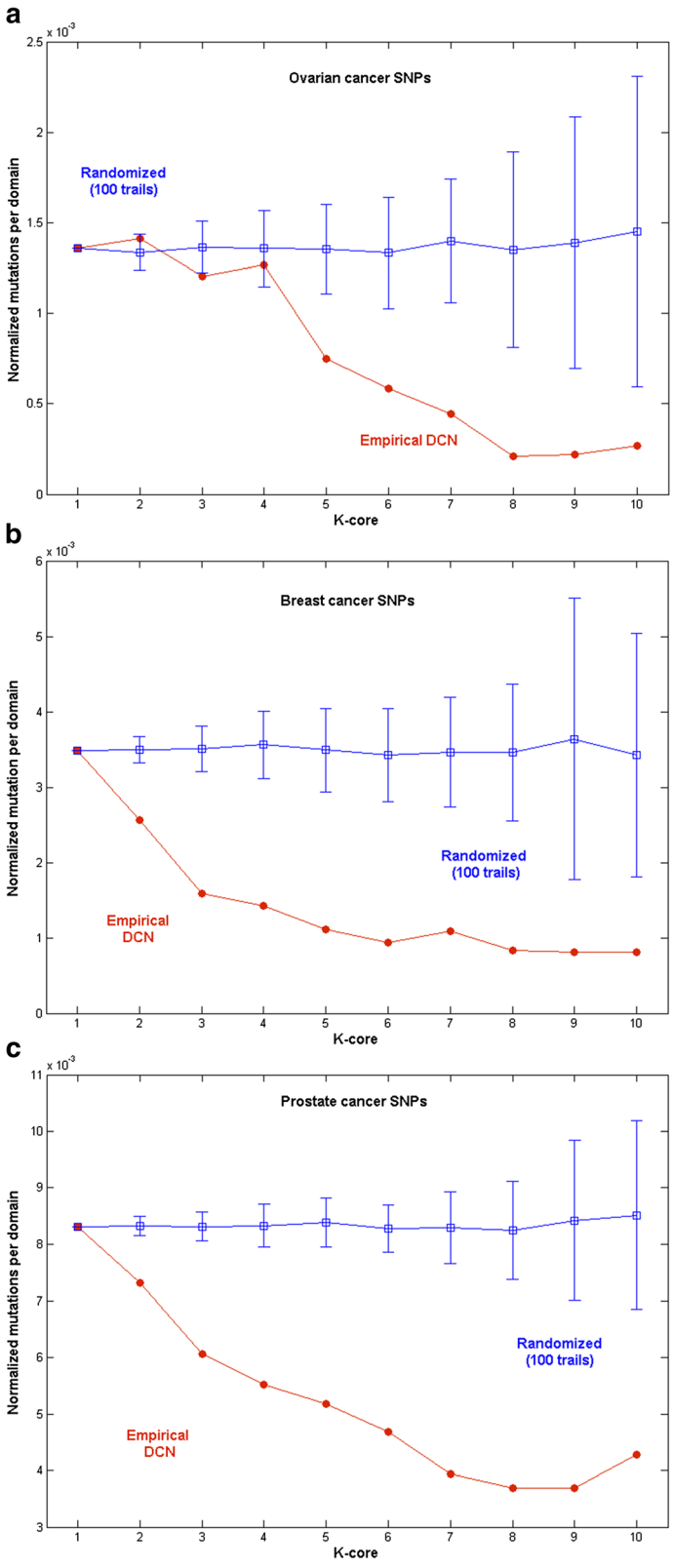

Figure 4 Variation of normalized mutation values with the core order a) Ovarian cancer b) Breast cancer and c) Prostate cancer.

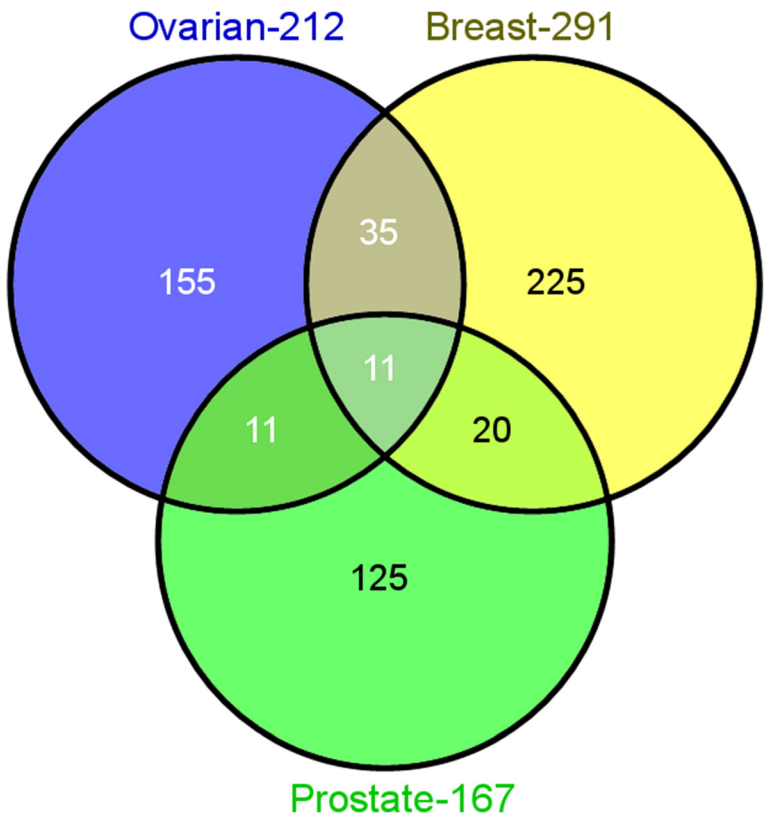

Figure 5 Overlap between significantly mutated domains of ovarian, breast and prostate cancer.

family member. This becomes an essential step toward a functional description of the global interactome [48].

By constructing and analyzing domain co-occurrence networks we gain new and fundamental insights into the qualitative arrangement and evolutionary utilization of the proteome. Domain databases like Pfam and Interdom provide comprehensive domain information but mapping cancer SNPs to the individual domains may help identify cancer targeted protein domains rather than just the proteins. Domains with high relative rates of mutation in three hormonal cancer types were identified along with their common domains. Recent studies from Liu et al, 2014 revealed that the PDZ and LIM protein domain promotes breast cancer cell migration, invasion and metastasis [52]. These two Pfam domains were also listed among the significantly mutated domains of the breast cancer.

Table 2 Significantly mutated domains found in all three cancers studied

S. No	Pfam id	Pfam domains
1	PF00870	P53
2	PF12129	Phtf-FEM1B_bdg
3	PF01192	RNA_pol_Rpb6
4	PF02020	W2
5	PF09801	SYS1
6	PF07941	K_channel_TID
7	PF00594	Gla
8	PF13096	CENP-P
9	PF01250	Ribosomal_S6
10	PF11629	Mst1_SARAH
11	PF05111	Amelin

Chi-squared goodness of fit test was employed to validate whether the observed and expected mutations are statistically significant. The expected mutations were calculated from the *Homo sapiens* DCN and the observed mutations included all the three cancer SNPs. The expected and observed mutations in different k-cores were calculated and plotted as shown in Figure 7. The observed mutations (red curve) were found to be lower than the expected mutations (blue curve) as the core order increased. This clearly suggests that the protein domain in the innermost core is less likely to get mutated as it was connected to many other protein domains and also corresponding to the set of domains with highest coreness. The p-value was found to be less than 0.05 suggesting that the observed mutation counts are not sampled from populations with the expected frequencies.

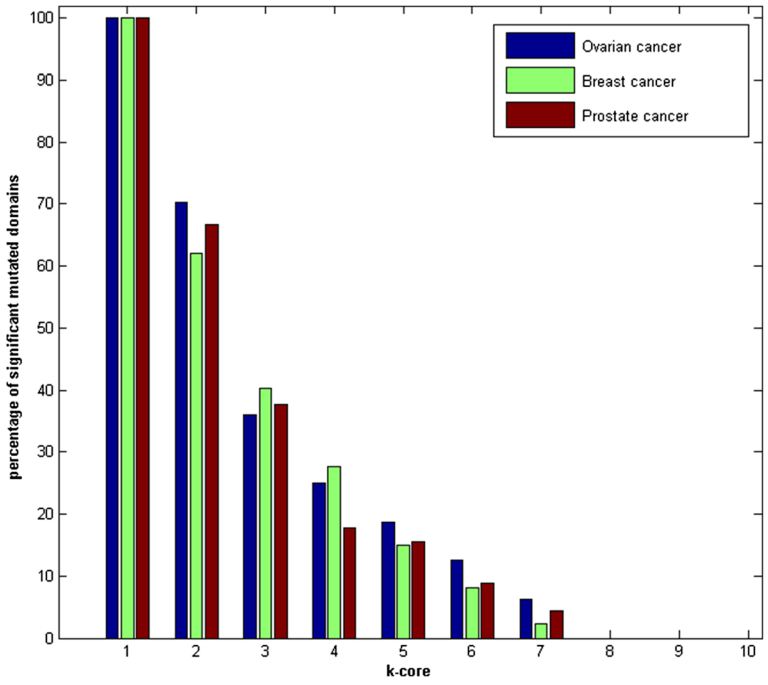

Figure 6 The percentage of significantly mutated domains in k-core decomposition.

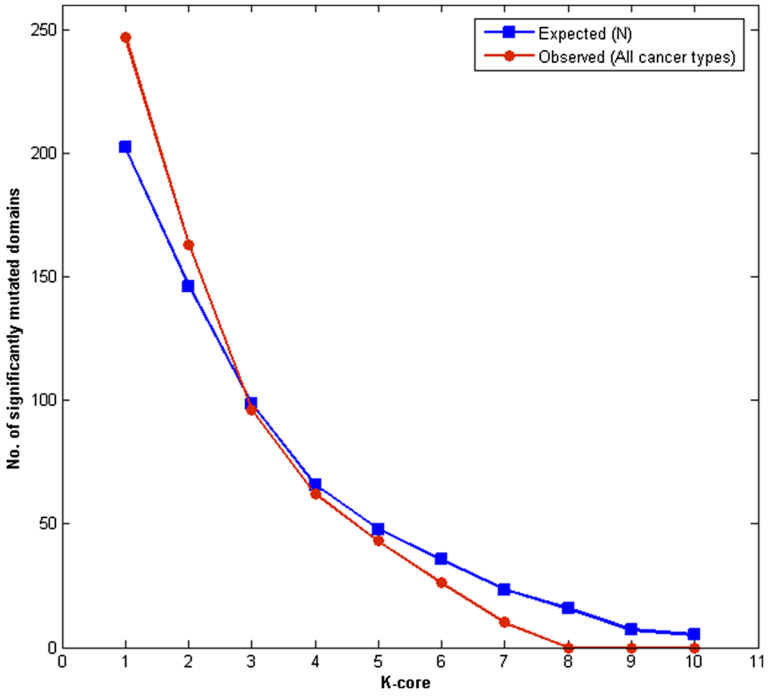

Figure 7 The expected versus observed significantly mutated domains in the core order.

Conclusion

Analyzing mutations in domain co-occurrence network helps in identifying crucial protein domains that aid in the progression of cancer disease. Highly connected protein domains are found to be evolutionarily conserved in the domain co-occurrence network. This implies that protein domains in the inner core are more conserved than the domains in the peripheral region. Significantly mutated protein domains which were identified further contributed to determining the disease target protein domains in all the three cancer diseases. Comparing the mutational landscape of somatic mutations in the protein domain co-occurrence network with the random counterparts, our findings revealed that there is a statistically significant difference between them.

The functional annotations obtained for all the significantly mutated domains were seen to be involved in all the three cancer diseases. Polymorphisms in inflammation-related genes, including those in the Toll-like receptor (TLR) signaling pathway, are hypothesized to be involved in prostate carcinogenesis [53]. This Toll-like receptor signaling pathway was enriched as one of the top ranked KEGG pathway in our results. Similarly, the ribosome pathway is also enriched as one of the top ranked KEGG pathway. This ribosome pathway are activated in aggressive human breast cancer cells [54] and comparison with other pathways showed that the ribosome pathway genes were up-regulated in ZR-75-1 (Human breast carcinoma cell line) [55].

A recent study done by Pasta A et al [56], showed that overexpressed genes in cancer stem cells (CSC) from patients with epithelial ovarian cancer are associated with glucose uptake, oxidative phosphorylation (OXPHOS) and fatty acid beta-oxidation. These overexpressed genes are consistent with a metabolic profile dominated by OXPHOS pathway [56]. In our results, 'Complement and coagulation cascades' was the most frequently perturbed pathway, as it was dysregulated in the ovarian cancer [57] and these two pathways are also found to be enriched in the significantly mutated domains. This clearly suggests that the statistically significant domains occur more commonly in cancer diseases. Further studies are however recommended to investigate the functional and structural constraints for the protein domain that evolves to be an inner core rather than outer core domain of the DCN.

Additional files

Additional file 1: Table S1. Chromosome start and end positions for each predicted domain from homo sapiens proteome.

Additional file 2: Table S2a (sheet1). Mutation count and their normalized score for each domain in Ovarian cancer. **Table S2b (sheet2).** Mutation count and their normalized score for each domain in breast cancer. **Table S2c (sheet3).** Mutation count and their normalized score for each domain in prostate cancer.

Additional file 3: Table S3a (sheet1). Local false discovery rate values for each domain in ovarian cancer. **Table S3b (sheet2).** Local false discovery rate values for each domain in breast cancer. **Table S3c (sheet3).** Local false discovery rate values for each domain in prostate cancer.

Additional file 4: Figure S1a. Degree distribution plot for Homo sapien's DCN. **Figure S1b.** Shortest path length distribution plot for Homo sapien's DCN. **Figure S1c.** Average clustering co-efficient distribution plot for Homo sapien's DCN.

Additional file 5: Figure S2a. Correlation between each domain's normalized mutation rate and k-core values in ovarian cancer. **Figure S2b.** Correlation between each domain's normalized mutation rate and k-core values in breast cancer. **Figure S2c.** Correlation between each domain's normalized mutation rate and k-core values in prostate cancer.

Additional file 6: Table S4a (sheet1). Significantly mutated domains in ovarian cancer using local false discovery rate values. **Table S4b (sheet2).** Significantly mutated domains in breast cancer using local false discovery rate values. **Table S4c (sheet3).** Significantly mutated domains in prostate cancer using local false discovery rate values.

Additional file 7: Table S5a (sheet1). KEGG pathways and gene ontology annotations enriched in significantly mutated domains for the ovarian cancer using DVAID tool. **Table S5b (sheet2).** KEGG pathways and gene ontology annotations enriched in significantly mutated domains for the breast cancer using DVAID tool. **Table S5c (sheet3).** KEGG pathways and gene ontology annotations enriched in significantly mutated domains for the prostate cancer using DVAID tool.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AIE contributed to study design, study execution, data analysis/interpretation, and manuscript preparation. JAM contributed to study design, study execution, data analysis/interpretation, and manuscript preparation. SA contributed to study execution and manuscript preparation. IA contributed to study execution and manuscript preparation. TKA contributed to study execution and manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful to Qatar Foundation Biomedical Research Program funding to Weill Cornell Medical College-Qatar (WCMCQ) for supporting this research.

Received: 18 November 2014 Accepted: 18 February 2015
Published online: 03 March 2015

References

1. Cheng J. DOMAC: an accurate, hybrid protein domain prediction server. Nucl Acids Res. 2007;35:w354–6.
2. Cheng J, Sweredoski M, Baldi P. Data mining and knowledge discovery. DOMpro. 2006;13:1–10.
3. Jaenicke R. Folding and association of proteins. Prog Biophys Mol Biol. 1987;49:117–237.
4. Albert R, Barabasi AL. Statistical mechanics of complex networks. Rev Mod Phys. 2002;74(1):47–97.
5. Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509–12.
6. Wagner A, Fell DA. The small world inside large metabolic networks. Proc Roy Soc London Series B. 2001;268(1478):1803–10.
7. Jeong H, Tombor B, Albert R, Oltvai Z, Barabasi AL. The large-scale organization of metabolic networks. Nature. 2000;407(6804):651–4.
8. Fell D, Wagner A. The small world of metabolism. Nature Biotech. 2000;18(11):1121–2.
9. Wuchty S. Small-worlds in RNA. Nucl Acids Res. 2003;31(3):1108–17.
10. Ravasz E, Somera A, Mongru D, Oltvai Z, Barabai A. Hierarchical organization of modularity in metabolic networks. Science. 2002;297(5586):551–1555.

11. Holme P, Huss M, Jeong H. Subnetwork hierarchies in biochemical pathways. Bioinformatics. 2003;19:532–8.
12. Barabasi A, Oltvai Z. Network biology: understanding the cell's functional organization. Nature Rev Gen. 2004;5(2):101–13.
13. Rives A, Galitski T. Modular organization of cellular networks. Proc Nail Acad Sci. 2003;100(3):1128–33.
14. National Cancer Institute. The Cancer Genome Atlas Homepage. http://cancergenome.nih.gov (30 March 2013, date last accessed).
15. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. In Science. 2006;314:268–74.
16. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. In Science. 2007;318:1108–13.
17. Ding L, Getz G, Wheeler D, Mardis E, McLellan M, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature Protoc. 2008;455:1069–75.
18. Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JC, et al. The genetic landscape of the childhood cancer medulloblastoma. Science. 2011;331:435–9.
19. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature. 2011;470:214–20.
20. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science. 2008;321:1801–6.
21. Zhong Q, Simonis N, Li QR, Charloteaux B, Heuze F, Klitgord N, et al. Edgetic perturbation models of human inherited disorders. Mol Syst Biol. 2009;5(1):321.
22. Wuchty S. Scale-free behavior in protein domain networks. Mol Biol Evol. 2001;18:1694–702.
23. Apic G, Gough J, Teichmann S. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol. 2001;310:311–25.
24. Wuchty S. Proteomics. Interaction and Domain Networks of Yeast. 2002;2:1715–23.
25. Wuchty S, Almaas E. Evolutionary cores of domain co-occurrence networks. BMC Evol Biol. 2005;5(1):24.
26. Hsu CH, Chen CK, Hwang MJ. The architectural design of networks of protein domain architectures. Biol Lett. 2013;9:20130268.
27. Nehrt NL, Peterson TA, Park D, Kann MG. Domain landscapes of somatic mutations in cancer. BMC Genomics. 2012;13:S9.
28. Fang H, Gough J. A disease-drug-phenotype matrix inferred by walking on a functional domain network, Mol Bio Syst. 2013;9:1686–96.
29. Seidman SB. Network structure and minimum degree. Soc Networks. 1983;5:269–87.
30. Batagelj V, Zaversnik M. An O (m) Algorithm for Cores Decomposition of Networks. 2003. *arXiv preprint cs/0310049* 2003.
31. Giatsidis C, Thilikos DM, Vazirgiannis M. D-cores: measuring collaboration of directed graphs based on degeneracy IEEE. In: Data Mining (ICDM), IEEE 11th International Conference on. 2011. p. 210–10.
32. Andersen R, Chellapilla K. Finding dense subgraphs with size bounds. In: Algorithms and Models for the Web-Graph. Berlin Heidelberg: Springer; 2009. p. 25–37.
33. Balasundaram B, Butenko S, Hicks IV. Clique relaxations in social network analysis: the maximum k-plex problem. Oper Res. 2011;59:133–42.
34. Kortsarz G, Peleg D. Generating sparse 2-spanners. J Algorithms. 1994;17:222–36.
35. Bollobas B, Thomason A. Random graphs of small order. North-Holland Mathematics Studies. 1985;118:47–97.
36. Batagelj V, Zaversnik M. Generalized Cores. 2002. *arXiv preprint cs/0202039* 2002.
37. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. Nucleic acids research. 2012;41:D48–55. gks1236.
38. Bateman A, Coin L, Durbin R, Finn R, Hollich V. The Pfam protein families database. Nucleic Acids Res. 2004;32:276–80.
39. Shannon P, Markiel A, Ozier O, Baliga N, Wang J. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
40. Dimitrieva S, Anisimova M. PANDITplus: toward better integration of evolutionary view on molecular sequences with supplementary bioinformatics resources. Trends Evol Biol. 2009;2:e1.
41. Efron B, Turnbull BB, Narasimhan B. Computation of Local False Discovery Rates, R package version 1.1-7. Vienna, Austria: R Foundation for Statistical Computing; 2011.
42. Efron B, Ibshirani R. Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol. 2002;23:70–86.
43. Huang D, Sherman B, Lempicki R. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4:44–57.
44. Huang D, Sherman B, Lempicki R. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.
45. Shoemaker B, Panchenko A, Bryant SH. Finding biologically relevant protein domain interactions: conserved binding mode analysis. Protein Sci. 2006;15:352–61.
46. Deng M, Mehta S, Sun F, Chen T. Inferring domain-domain interactions from protein-protein interactions. Genome Res. 2002;12:1540–8.
47. Moon HS, Bhak J, Lee KH, Lee D. Architecture of basic building blocks in protein and domain structural interaction networks. Bioinformatics. 2005;21:1479–86.
48. Santonico E, Castagnoli L, Cesareni G. Methods to reveal domain networks. Drug Discov Today. 2005;10:1111–7.
49. Emig D, Cline MS, Lengauer T, Albrecht M. Integrating expression data with domain interaction networks. Bioinformatics. 2008;24:2546–8.
50. Prieto C, De Las Rivas J. Structural domain-domain interactions: assessment and comparison with protein-protein interaction data to improve the interactome. Proteins. 2010;78:109–17.
51. Stein A, Ceol A, Aloy P. 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2011;39:D718–23.
52. Liu Z, Zhan Y, Tu Y, Chen K, Liu Z, Wu C. PDZ and LIM domain protein 1 (PDLIM1)/CLP36 promotes breast cancer cell migration, invasion and metastasis through interaction with α-actinin. Oncogene. 2014, doi:10.1038/onc.2014.64
53. Stark JR, Wiklund F, Gronberg H, Schumacher F, Sinnott JA, Stampfer MJ, et al. Toll-like receptor signaling pathway variants and prostate cancer mortality. Cancer Epidemiol Biomarkers Prev. 2009;18:1859–63.
54. Belin S, Beghin A, Solano-Gonzalez E, Bezin L, Brunet-Manquat S, Textoris J, et al. Dysregulation of ribosome biogenesis and translational capacity is associated with tumor progression of human breast cancer cells. PLoS One. 2009;4:e7147.
55. Mandal S, Davie JR. An integrated analysis of genes and pathways exhibiting metabolic differences between estrogen receptor positive breast cancer cells. BMC Cancer. 2007;7:181.
56. Pasta A, Bellio C, Pilotto G, Ciminale V, Silic-Benussi M, Guzzo G, et al. Cancer stem cells from epithelial ovarian cancer patients privilege oxidative phosphorylation, and resist glucose deprivation. Oncotarget. 2014;5:4305.
57. Lin P, Huang Z. Correlation analysis connects cancer subtypes. PLoS One. 2013;8:e69747.