



The Network Organization of Cancer-associated Protein Complexes in Human Tissues

Jing Zhao¹, Sang Hoon Lee^{2,3}, Mikael Huss^{4,5} & Petter Holme^{2,6}

¹Department of Mathematics, Logistical Engineering University, Chongqing, China, ²IceLab, Department of Physics, Umeå University, Umeå, Sweden, ³Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford, United Kingdom, ⁴Science for Life Laboratory Stockholm, Solna, Sweden, ⁵Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden, ⁶Department of Energy Science, Sungkyunkwan University, Suwon, Korea.

SUBJECT AREAS:

PROTEOME
INFORMATICS

COMPUTATIONAL SCIENCE

ONCOGENESIS

COMPLEX NETWORKS

Received
24 October 2012

Accepted
7 March 2013

Published
9 April 2013

Correspondence and
requests for materials
should be addressed to

J.Z. (zhaojanne@
gmail.com) or P.H.
(petter.holme@
physics.umu.se)

Differential gene expression profiles for detecting disease genes have been studied intensively in systems biology. However, it is known that various biological functions achieved by proteins follow from the ability of the protein to form complexes by physically binding to each other. In other words, the functional units are often protein complexes rather than individual proteins. Thus, we seek to replace the perspective of disease-related genes by disease-related complexes, exemplifying with data on 39 human solid tissue cancers and their original normal tissues. To obtain the differential abundance levels of protein complexes, we apply an optimization algorithm to genome-wide differential expression data. From the differential abundance of complexes, we extract tissue- and cancer-selective complexes, and investigate their relevance to cancer. The method is supported by a clustering tendency of bipartite cancer-complex relationships, as well as a more concrete and realistic approach to disease-related proteomics.

Genome sequencing can, at least in an idealized world, list the repertoire of what a cell could possibly do; expression profiling, on the other hand, reflects what the cell actually is doing. Selective or differential gene expression profiles in specific cells, therefore, add valuable contextual information. It is quite natural to connect the differential gene expression profiles to disease states, whether they are genetic diseases or not. An overwhelming number of studies in this vein have been published: e.g., Refs. 1–6 to name just a few. Essentially all of these approaches make the assumption that genes are the units of biological functionality.

Even if the assumption cannot be denied, it has recently been pointed out that the relationships among proteins, not just properties of individual proteins, are essential ingredients in characterizing the entity of biological functions. The relationships can be binary protein-protein interactions (PPIs)^{7–10} or formation of stable structural and functional units called protein complexes^{11–15}. Proteins tend to function as members of complexes, and dysfunctions of different proteins in the same complex generally lead to similar disorders. Research has been conducted trying to identify disease-associated protein-protein interactions, signaling pathways and protein complexes by the integrated computational analysis of heterogeneous data sources^{16–22}.

Human diseases usually occur in one or more specific tissues and organs, while different types of organs and tissues make use of selective sets of expressed genes, protein-protein interactions and protein complexes²³. Genes predominantly expressed in one or a few biologically similar tissue types are defined as tissue-selective genes²⁴. Similarly, protein complexes showing significantly higher abundance levels in one or limited tissues are considered as tissue-selective complexes. Tissue-selective genes and complexes could be disease markers and potential drug targets. Although many approaches have been developed to identify tissue-selective genes and their relationships to diseases^{24–29}, the identification of tissue- and disease-selective complexes is still in its infancy due to the lack of adequate coverage on experimental proteomic data, so that gene expression levels have been used instead of protein abundance^{20,30,31}.

In this paper, by using the optimization algorithm for estimating differential abundance levels of protein complexes introduced in Ref. 15, we attempt to define the human tissue- and cancer-selective protein complexes. More specifically, we use the recently released E-MTAB-62 gene expression profile dataset³² and focus on 39 solid tissue cancers and 25 different normal tissues from some of which the cancers are originated (Table 1). From the abundance profiles of complexes, we classify the complexes associated with cancers and tissues into four different categories called Patterns 1–4, where the complexes over-expressed in cancers but under-expressed in originated



Table 1 | List of solid cancers and their originated normal tissues. Cancers were selected from the file “E-MTAB-62.sdrf.txt” whose columns “Characteristics [4 meta-groups]” and “Characteristics [Blood/NonBlood meta-groups]” are “neoplasm” and “non blood”, respectively. Cancer name and its originated normal tissue are taken from “Characteristics [DiseaseState]” and “Characteristics [OrganismPart]” of the file, respectively

Cancers	Originated normal tissue
Liposarcoma, Myxoid liposarcoma	adipose tissue
Bladder cancer	bladder
Chondroblastoma, Chordoma, Ewings sarcoma, Osteosarcoma, Spindle cell tumor	bone
Brain tumor, Ganglioneuroblastoma, Ganglioneuroma, Glioblastoma, Malignant peripheral nerve sheath tumor, Neuroblastoma, Neurofibroma, Schwannoma	brain
Chondromyxoid fibroma, Chondrosarcoma, Dedifferentiated chondrosarcoma, Fibromatosis, Monophasic synovial sarcoma, Sarcoma	connective tissue
Esophageal adenocarcinoma	esophagus
Oral squamous cell carcinoma	hypopharynx
Kidney carcinoma, Renal cell carcinoma	kidney
Hepatocellular carcinoma	liver
Lung cancer	lung
Uterine tumor	myometrium
Head and neck squamous cell carcinoma	hypopharynx
Ovarian tumor	ovary
Prostate cancer	prostate
Acute quadriplegic myopathy	skeletal muscle
Kaposi sarcoma	skin
Alveolar rhabdomyo sarcoma, Embryonal rhabdomyo sarcoma, Leiomyosarcoma	smooth muscle
Germ cell tumor	testis
Thyroid adenocarcinoma	thyroid

normal tissues are considered as most relevant and analyzed in terms of the bipartite relation between cancers and complexes. Finally, we show that the correlation structures of different cancers and tissues are preserved in our complex-based study, in comparison to the results from individual gene expression levels.

Results

Differentially expressed protein complexes in normal tissues.

First, we present our results of the differentially expressed protein complexes in normal tissues. For each of 25 solid tissues under study, using the average abundance levels over all the other tissues as the control set, we extracted over (under)-expressed complexes with a change more than a factor two, or less than a factor 1/2 (Table S1 and S2). A total of 106 and 209 distinct protein complexes were found over- and under-expressed in normal tissues, respectively. See Table S3 for the number of complexes differentially expressed in each tissue. The distributions of the number of different tissues in which complexes are over- or under-expressed are shown in Fig. 1. It can be seen that most complexes are over- and under-expressed only in a small number of tissues, suggesting that a large fraction of complexes

predicted by our method exhibits a high extent of tissue selectivity. Note that the tissues are (of course) not completely independent from one another, which may be responsible for some multiple numbers of tissues in which complexes are differentially expressed.

In the CORUM (Comprehensive Resource of Mammalian protein complexes³³) database, which we use for our complex list, functions of protein complexes are annotated by the Functional Catalogue (FunCat) scheme, whose hierarchical structure allows browsing for protein complexes with particular cellular functions or localizations^{33,34}. However, among all the 2837 mammalian protein complexes in the CORUM database, only 148 have information concerning specific animal tissue of the complex. Because of this lack of tissue-specific annotation, only 5 of the 106 over-expressed complexes predicted by our method have tissue annotation. As shown in Table 2, among the 5 complexes, 4 complexes are consistent with the annotation, suggesting the validity of our result. For instance, “thymus” (our predicted tissue) and “bone marrow” (CORUM) are compatible, as both of those are hot spots of T cell production and maturation³⁵. They are both considered (the only) “primary lymphoid organs”³⁵.

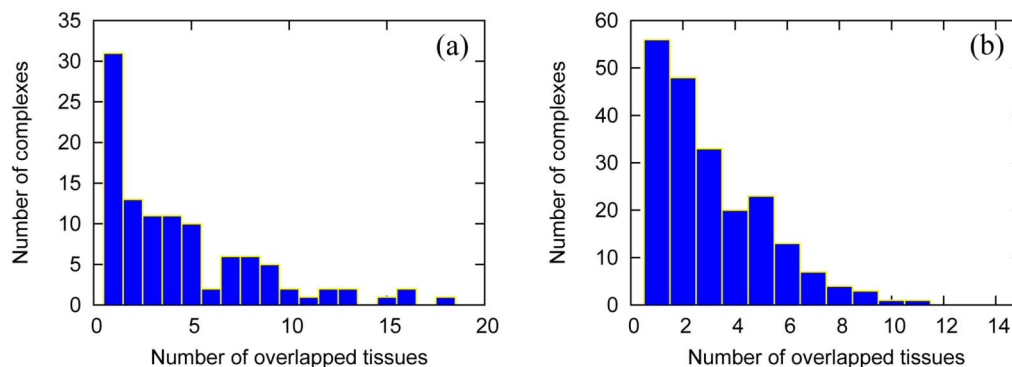


Figure 1 | Distributions of number of overlapped tissues for over-expressed (a) and under-expressed (b) complexes, in normal tissues. For each over- or under-expressed complex in normal tissues, we count the number of tissues where it is over- or under-expressed and define the number as the number of overlapped tissues.



Table 2 | Comparison of our results with tissue information of complexes in CORUM. Boldface marks consistent results

complex name	tissue information in CORUM	over-expressed predicted tissue
KCNQ1 macromolecular complex	muscle and heart muscle	adipose tissue bone brain heart liver smooth muscle testis
RICH1-PAR3-aPKC polarity complex	epithelium	adipose tissue hypopharynx lymph node skeletal muscle skin
SMAD3-SMAD4-FOXO3-FOXG1 complex	epithelium	connective tissue eye skin thymus thyroid ovary tonsil
PKC-alpha-PLD1-PLC-gamma-2 signaling complex, lacritin stimulated	epithelium	
YY1-Notch1 complex	bone marrow	thymus

Differentially expressed protein complexes in solid cancers. As in the normal tissue case, for each of 39 solid tissue cancers, using the abundance levels in the originated normal tissue as the control set, we extract over-(under)-expressed complexes with more (less) than 2-fold (1/2-fold) changes, respectively (Tables S4 and S5). A total of 283 and 294 distinct complexes were identified over- and under-expressed in the cancers, respectively. We call these complexes *cancer-associated complexes*. Again, from the distributions of the number of different cancers in which complexes are over- or under-expressed, shown in Fig. 2, we can observe the high degree of cancer selectivity of the complexes. The fact that several cancers are derived from the same normal tissues seems to be responsible for the larger number of overlapped cancers compared to the number of overlapped normal tissues in Fig. 1, and in fact, such cancer-cancer correlations will be presented later.

The most fundamental assumption of our approach is to treat the complexes as a functional unit, instead of individual component proteins. In other words, differential abundance profiles for complexes are more relevant than the ones for individual genes, since each gene may play different functional roles in different complexes, resulting in the situation that expression levels over different contexts are effectively “averaged out.” In Table 3, we compare over-expressed protein complexes of brain tumor with their up-regulated component genes which were shown associated with nerve system cancers in

GeneCards³⁶. We use the t-test to test if a gene is differentially expressed in the brain tumor and control samples. For such a large number of genes being simultaneously tested, the FDR³⁷ corrected p-values are used for screening differentially expressed genes. We consider genes with at least 2-fold change of log ratio for average expression level and FDR at most 0.05 as up-regulated in brain tumor. It can be seen that in complexes identified over-expressed in brain tumor by our algorithm, only a small fraction of component genes associated with nerve system cancers was up-regulated. Such a large difference is strong evidence supporting the fundamental assumption of complexes’ relevance to biological functions and dysfunctions compared to individual genes.

Considering that the database E-MTAB-62 we used is an integration of data generated in different laboratories, we conducted a within-laboratory comparison on over-expressed complexes in brain tumor to see to what extent our result is replicated across studies. The samples of brain tumor and normal brain tissue came from 2 and 6 different laboratories, respectively. By combining brain tumor samples from one lab with normal brain samples from another lab, we got 12 different sample sets. We ran our algorithm on each sample set and identified complexes over-expressed in brain tumor. As shown in Figure 3, most complexes identified by our algorithm are also identified by at least half of the sample sets. Then we ran our algorithm on each of the brain tumor and normal brain tissue samples,

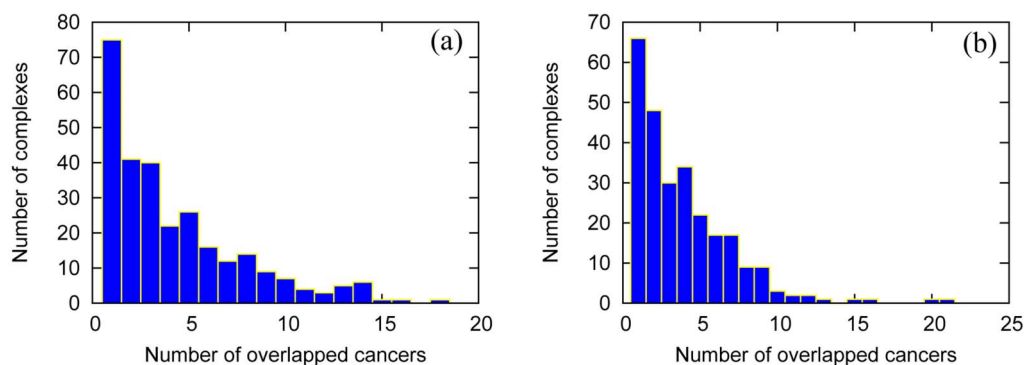


Figure 2 | Distributions of number of overlapped cancers for over-expressed (a) and under-expressed (b) complexes, in cancers. For each over- or under-expressed complex in cancers, we count the number of tissues where it is over- or under-expressed and define the number as the number of overlapped cancers.



Table 3 | Comparison of protein complexes over-expressed in brain tumor and their up-regulated component genes

Complex name	Also identified by GSEA	Number of genes in complex	Percentage of up-regulated genes in complex	Number of genes associated with nerve system cancers	Percentage of up-regulated genes in subset of genes associated with nerve system cancers
SMN complex, U7 snRNA specific	Y	5	100%	0	-
CDC2-CCNA2 complex	Y	2	50%	2	50%
VEGF transcriptional complex	Y	5	40%	1	100%
Cell cycle kinase complex CDK5	Y	5	40%	5	40%
Anti-HDAC2 complex	Y	17	17.65%	7	42.86%
Emerin complex 52	Y	23	17.39%	8	25%
RC complex during S-phase of cell cycle	Y	13	7.69%	7	0
WINAC complex	Y	14	7.14%	6	16.67%
EIF3 complex (EIF3B, EIF3G, EIF3I)	Y	3	0	0	-
CAV1-VDAC1-ESR1 complex	N	3	33.33%	3	33.33%
SMURF2-SMAD3-SnoN complex, TGF(beta)-dependent	N	3	33.33%	2	0
VHL-TBP1-HIF1A complex	N	3	33.33%	2	50%
RAF1-RAS complex, EGF induced	N	4	25%	4	25%
P2X7 receptor signalling complex	N	11	9.09%	5	20%
RNA polymerase II complex, chromatin structure modifying	N	18	0	5	0
MRN-TRRAP complex (MRE11A-RAD50-NBN-TRRAP complex)	N	4	0	2	0
PLC-gamma-1-SLP-76-SOS1-LAT complex	N	4	0	2	0
PlexinA1-NRP1-SEMA3A complex	N	3	0	2	0
SMARCA2/BRM-BAF57-MECP2 complex	N	3	0	2	0
TRAP complex	N	15	0	2	0
APP-TIMM23 complex	N	2	0	1	0
BCL6-HDAC7 complex	N	2	0	1	0
DNA polymerase alpha-primase complex	N	4	0	1	0
MCM8-ORC2-CDC6 complex	N	2	0	1	0
RICH1-PAR3-aPKC polarity complex	N	3	0	1	0
APLG1-Rababin5 complex	N	3	0	0	0
BLM-TOP3A complex	N	2	0	0	-
CTF18-CTF8-DCC1-RFC3 complex	N	2	0	0	-
FEN1-9-1-1 complex	N	4	0	0	-
Kinase-scaffold-phosphatase complex, PKA-AKAP79-CaN	N	3	0	0	-
PPP4C-PPP4R2-Gemin3-Gemin4 complex	N	3	0	0	-
Retrotranslocation complex	N	2	0	0	-
RFC2-Rlalpha complex	N	2	0	0	-
TRAP-SMCC mediator complex	N	7	0	0	-

respectively. By t-test and multiple testing corrections on the resulting complex abundance matrix of large samples, we identify complexes statistically over-expressed in brain tumor with $FDR < 0.05$. A total of 29 complexes identified over-expressed by this sample replication method are also identified by our method which used the average of samples as input (See Figure 3). These comparisons suggest the robustness of our algorithm on different data resources.

We also compare our algorithm to a gene set testing approach, the Gene Set Enrichment Analysis (GSEA)³⁸. Using the CORUM complexes as gene sets, we conducted GSEA analysis on expression data of brain tumor and normal brain tissue. This method identifies 227 complexes that were significantly enriched in brain tumor tissue ($FDR < 25\%$). As shown in Figure 3 and Table 3, 9 of the 34 complexes over-expressed in brain tumor identified by our method are also identified by GSEA. From Table 3 we see that relatively more up-regulated genes appeared in the overlapped complexes, which is the principle of identifying enriched gene sets by GSEA. Complexes identified over-expressed only by our algorithm include genes reported associated with nerve system cancers, suggesting they may related with brain tumor. However, these complexes are not detected by GSEA because few genes were up-regulated. This comparison suggests that our algorithm, which considers stoichiometry of complexes from global point of view, could add some new information in complex prediction.

From Figure 3 we can see that several complexes, such as Anti-HDAC2 complex, SMN complex, EIF3 complex, CDC2-CCNA2

complex, are well identified over-expressed in brain tumor by all the four methods, suggesting strong expression signals of these complexes in brain tumor. Complexes such as CDC2-CCNA2 complex, Anti-HDAC2 complex and WINAC complex are more obviously associated with brain tumor due to their high fraction of component proteins related with nerve system cancer (see Table 3). However, from GeneCards and GoPubmed database, all the five component proteins of SMN complex (small nuclear ribonucleoprotein B, D, E, F, G) are not associated with nerve system cancer although they are highly associated with neurologic manifestations and neurodegenerative diseases. Our computations found this complex and its five component proteins are significantly over-expressed in brain tumor, indicating its relationship with brain tumor. More research deserves to be undertaken to validate such results.

Expression patterns of cancer-associated complexes in normal tissues. For complexes differentially expressed in a cancer, we compare their abundance levels in the cancer tissue with those in the originated normal tissue, and in the other normal tissues. Specifically, we mapped the differentially expressed complexes in each cancer to each normal tissue and classified differential expressions of these complexes according to the following four patterns:

Pattern 1: over-expressed in the cancer tissue but under-expressed in the normal tissue

Pattern 2: over-expressed in the cancer tissue as well as in the normal tissue

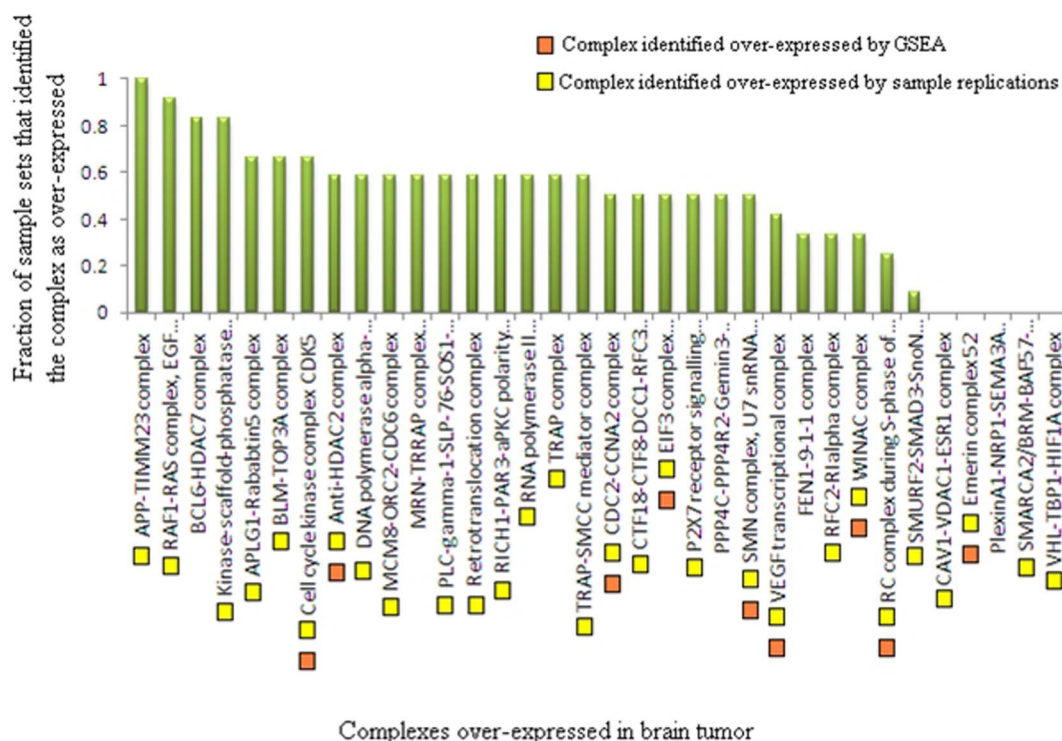


Figure 3 | Cross-validation of complexes over-expressed in brain tumor identified by our method by within-laboratory comparison, sample replication method and GSEA.

Pattern 3: under-expressed in the cancer tissue but over-expressed in the normal tissue

Pattern 4: under-expressed in the cancer tissue as well as in the normal tissue

For each cancer, we count the number of complexes in each tissue of different Patterns (see Table S6). Then for each cancer, we list the number of complexes in the tissue from which it originated, along with the largest number of complexes among the other tissue other than its originated tissue, classified as the different Patterns (see

Table S7). Figure 4 shows the distribution of the four differential expression patterns of cancer-associated complexes in their originated normal tissues. It can be seen that the dominant expression patterns are Patterns 1 (57.2%) and 3 (27.1%), whereas Patterns 2 and 4 complexes in originated normal tissues (1.15% and 3.87%) are minorities. In Table S7, we list the comparison of the four patterns in cancers' originated normal tissues with those in the other normal tissue with the maximum number of cancer-associated complexes. Table S7 shows that, compared with those in the other normal tissues, Pattern 1 complexes in originated normal tissues are much more numerous (57.2% vs. 22.6%); Pattern 2 and 4 complexes in originated normal tissues are much fewer (1.15% vs. 17.5% for Pattern 2; and 3.87% vs. 22.94% for Pattern 4); Pattern 3 complexes has no significant difference (27.1% vs. 26.9%). Moreover, by the t-test, the expressions of Pattern 3 complexes in originated normal tissues have no significant difference from those in other normal tissues; whereas the expressions of Pattern 1, 2 and 4 complexes are significantly different from those in the other normal tissues, respectively.

From these observations, we can conclude that solid cancers tend to over-express complexes that are under-expressed in the normal tissues of the cancers' origin (Pattern 1). In other words, *complexes that are not supposed to be expressed in a specific tissue but are over-expressed in this tissue* can be related to cancers. Furthermore, solid cancers could over-express (or under-express) part of complexes that are over-expressed (or under-expressed) in normal tissues other than the cancer's tissue of origin (Patterns 2 and 4). These patterns could complement earlier findings on single gene expression pattern in cancers. For example, it was reported that genes over-expressed in human leukemias were rarely over-expressed in hematopoietic tissues³⁹. Generally, cancers over-express only a fairly small part of genes that are selectively expressed in their originated tissues²⁵. On the other hand, under-expressed complexes in cancers do not have statistically significant tendency to be over-expressed in the originated normal tissues (Pattern 3), which can be interpreted to mean

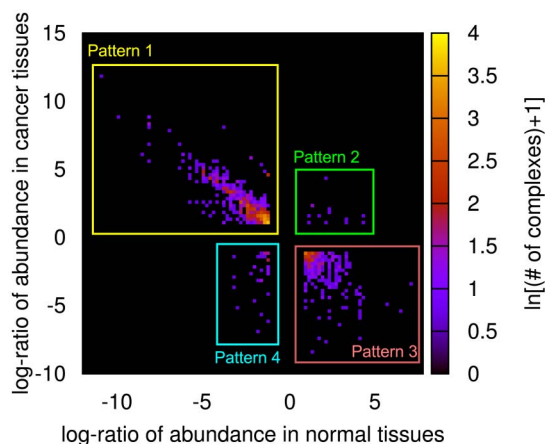


Figure 4 | Differentially expressed complexes in cancers and originated normal tissues. Log-ratio of abundance in cancers (vertical axis) are defined with respect to the originated normal tissues, and that in normal tissues (horizontal axis) are defined with respect to all the other normal tissues. The log-ratio values in the “normal” range (−1, 1) are excluded for both cancers and normal tissues. Four different patterns are noted according to their differential abundance levels in cancers and their originated tissues.

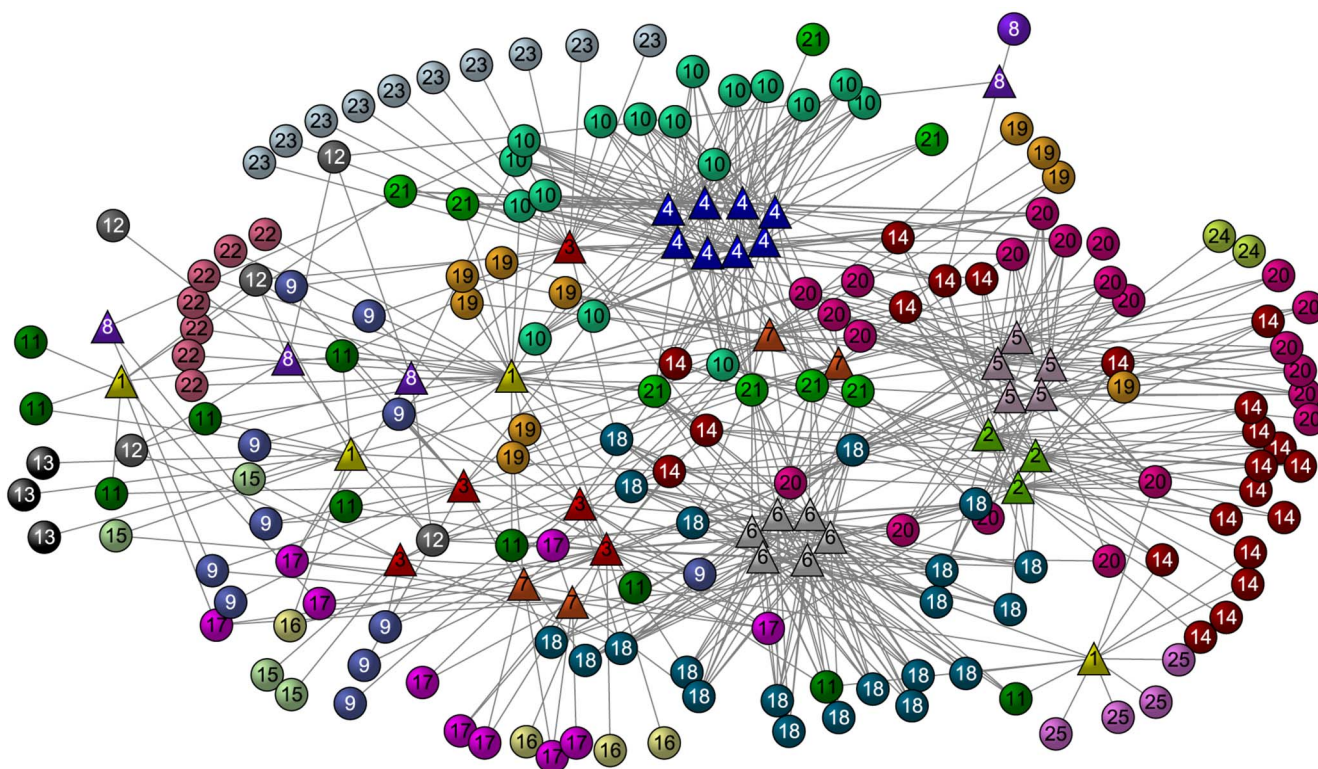


Figure 5 | Bipartite network of cancers and protein complexes of Pattern 1. Triangles (circles) represent cancers (complexes), respectively. The numbers (and corresponding colors) on vertices show the clustering structure defined with the Jaccard similarity index (see the text).

that the lack of necessary complexes does not tend to cause cancers, in contrast to the existence of unnecessary complexes in Pattern 1.

It is known that one form of cancer can affect many tissues, not only the tissue from which it originated. The expression patterns of cancer-associated complexes may indicate the cancer-tissue relations. One interesting way to verify the cancer-tissue relations from an external source is to use the Web search engine⁴⁰. Our basic assumption is that the more Web pages Google finds from the search query with [cancer name][tissue name], the more probably the tissue is related to the cancer. We measure cancer-tissue “Google correlation” (“Google page” column in Table S6). For a specific cancer A, most Google correlation values for [cancer A][originated tissue of cancer A] pair are ranked on the top among all the [cancer A][tissue name]. More precisely, 14 of the 39 cancers have the largest number of Google correlation value with their originated tissues. This result validates our assumption. In addition, from Table S6, for each cancer, we calculated the Pearson correlation coefficient between columns ‘Google pages’ and column ‘Patterns 1–4,’ as shown in Table S8.

The statistical significance test suggested that cancer-associated complexes are expressed according to Patterns 1, 2 or 4. Thus, we took the maximum values of Pearson correlation coefficient for Patterns 1, 2, and 4, and show them in the last column of Table S8. Most (about 3/4) of the Pearson correlation coefficients in the last column are positive, suggesting a positive correlation between cancer-tissues relations from Google correlation and those from the number of cancer-associated complexes with differential abundance levels.

Bipartite complex-cancer relations and common complexes associated with the same cluster of cancers. The previous subsection suggests that most cancer-associated complexes are Pattern 1 complexes in the originated normal tissues, i.e., over-expressed in the cancer tissue but under-expressed in the originated normal tissue. Thus we focus on these Pattern 1 complexes, and investigate the bipartite network between cancers and Pattern 1 complexes in

cancer tissues. We constructed a bipartite network between cancers and Pattern 1 complexes, in which a cancer node is connected to a complex node if and only if this complex is a Pattern 1 complex of this cancer. In the bipartite network, we measured the topological similarity of the vertices according to the following Jaccard similarity index:

$$J(u, v) = \frac{|N_u \cap N_v|}{|N_u \cup N_v|},$$

where N_u is the set of neighbors of node u . Then Ward’s clustering, a hierarchically agglomerative clustering method, was used to cluster the nodes in the network⁴¹. The hierarchical clustering starts off with each node being its own cluster and the distance between nodes u and v is defined as $d(u, v) = 1 - J(u, v)$. At each step, pair of clusters (u, v) with the smallest distance $d(u, v)$ is selected to be merged as a single cluster and distance measures between clusters are updated as the weighted sum of distances according to the Lance-Williams algorithm⁴², and the process is repeated until all nodes have been combined into one cluster, represented as a dendrogram with a hierarchical structure. In our case, $d(u, v) = 2$ is used as the threshold for cutting the hierarchical tree to yield the clustering structure. Figure 5 shows that some cancers are clustered because of their common over-expressed complexes, and also some complexes are clustered together.

We classify the 39 cancers under study into six categories according to Medical Subject Headings (MeSH⁴³) annotation of their originated tissue categories: nerve tissue neoplasm, connective and soft tissue neoplasm, head and neck neoplasm, urogenital tissue neoplasm, digestive system neoplasm, and respiratory tract neoplasm. Biologically, cancers originated from same tissue should be correlated to some extent. In Table 4, we list the cluster indexes of the cancers in Figure 5 and their originated tissues. It can be seen that cancers originated from the same tissue category are clustered together. Figure 5 shows that cancers in the clusters 4, 5, 6 tend to link with complexes in clusters 10, 20 and 18 respectively, suggesting



Table 4 | Cancers classified by categories of their originated tissues and topology of the cancer-complex association network in Fig. 4

cluster index	cancer	originated tissue
1	acute quadriplegic myopathy	connective and soft tissue
	thyroid adenocarcinoma	head and neck
	germ cell tumor	urogenital tissue
2	uterine tumor	urogenital tissue
	alveolar rhabdomyo sarcoma	connective and soft tissue
	embryonal rhabdomyo sarcoma	connective and soft tissue
3	leiomyosarcoma	connective and soft tissue
	Hepatocellular carcinoma	digestive system
	esophageal adenocarcinoma	head and neck
4	lung cancer	respiratory tract
	bladder cancer	urogenital tissue
	ovarian tumor	urogenital tissue
	brain tumor	nerve tissue
	ganglioneuroblastoma	nerve tissue
	ganglioneuroma	nerve tissue
	glioblastoma	nerve tissue
5	malignant peripheral nerve sheath tumor	nerve tissue
	neuroblastoma	nerve tissue
	neurofibroma	nerve tissue
	schwannoma	nerve tissue
	chondroblastoma	connective and soft tissue
	chordoma	connective and soft tissue
	ewings sarcoma	connective and soft tissue
6	osteosarcoma	connective and soft tissue
	spindle cell tumor	connective and soft tissue
	chondromyxoid fibroma	connective and soft tissue
	chondrosarcoma	connective and soft tissue
	dedifferentiated chondrosarcoma	connective and soft tissue
	fibromatosis	connective and soft tissue
7	monophasic synovial sarcoma	connective and soft tissue
	sarcoma	connective and soft tissue
	liposarcoma	connective and soft tissue
	myxoidliposarcoma	connective and soft tissue
8	head and neck squamous cell carcinoma	head and neck
	oral squamous cell carcinoma	head and neck
	Kaposi sarcoma	connective and soft tissue
	kidney carcinoma	urogenital tissue
	renal cell carcinoma	urogenital tissue
	prostate cancer	urogenital tissue

the association of these complex groups with nerve tissue cancers (cluster 4) and connective tissue cancers (cluster 5 and 6) respectively. To verify the correlation of the complexes in cluster 10 with nerve tissue cancers (cluster 4), we searched PubMed⁴⁴ with complex names or gene names of the complex component proteins (in January of 2012) and listed the results in Table 5. A total of 13 of the 17 complexes show rank 1 association with cancers compared with all diseases, implying the important functions of these complexes in the occurrence or development of cancers. The associations of most complexes with nerve system diseases and nerve system cancers rank on the top of “All of Diseases” (more than 20 disease items) and “Neoplasms by Site” (more than 10 cancer tissue items) lists, respectively, demonstrating a high degree of correlation of complexes in cluster 10 with nerve system cancers. Moreover, proteins in some complexes such as cell cycle kinase complex CDK5, SMARCA2/BRM-BAF57-MECP2 complex and SMARCA2/BRM-BAF57-MECP2 complex have been extensively reported to be associated with eye cancer retinoblastoma, specifically implying the functions of these complexes in nerve systems cancers. In addition, 5 complexes in Table 5, CDC2-CCNA2 complex, Cell cycle kinase complex CDK5, Anti-HDAC2 complex, Emerin complex 52 and WINAC complex, are also identified over-expressed in brain tumor by GSEA (see Table 3), which cross-validates the correlation of these complexes with nerve system cancer. Similarly, the associations of complexes in cluster 20 with connective tissue cancers were shown in Table S9.

Cancer-cancer correlations deduced from gene expression and complex abundance profiles. From our results, we see that many complexes predicted by our algorithm are important biological modules involved in the occurrence and development of solid cancers, and these modules suggest correlations of cancers to some extent. To verify if the predicted complexes could reflect the relationships between different cancers as the original gene expression data do, we hierarchically clustered the gene expression profile and complex abundance profile of all cancers and normal tissues under study, respectively. Similarity between groups is defined as the mean Pearson correlation coefficient between the sample profiles (hierarchical clustering trees in Figs. S1 and S2). Three large tissue categories include more cancers—soft tissue, nerve tissue and urogenital tissue are clustered together in both cases; i.e. both clustering results show the correlations of cancers and normal tissues of similar tissue categories.

Similarly, according to the relative gene expression level and complex abundance of the cancers against their originated normal tissues by log-ratio values, we hierarchically clustered the cancers, respectively (Figs. 6 and 7). Figure 7 shows the heatmap of hierarchical clustering of the 39 cancers compared to each other, according to relative complex abundance of cancer against its originated normal tissue. Similar to the heatmap in Fig. 6, the clusters of cancers in Fig. 7 are mostly consistent with their tissue categories. We partitioned the cancers into 4 clusters according to the hierarchical trees in Figs. 6



Table 5 | GoPubMed search results for the associations of complexes in cluster 10 with nerve tissue cancer. (Complexes with higher specificity are shown in the boldface.) Disease hits: number of PubMed papers indicating the association of searched item with diseases; neoplasms hits/rank: number of PubMed papers indicating the association of searched item with cancers and the rank of paper numbers in “All of Diseases” item of GoPubMed results. Association with nerve tissue cancer: number of PubMed papers indicating the association of searched item with nerve system diseases (box in the first row) and nerve system cancers (box in the second row) and the rank of paper numbers in “All of Diseases” and “Neoplasms by Site,” respectively

complex name	item searched	disease hits	neoplasms hits/rank	association with Nerve tissue cancer	
				disease	hits/rank
Cell cycle kinase complex CDK5	CCND1	10111	8862/1	Retinoblastoma	983/3
RICH1-PAR3-aPKC polarity complex	PARD3	49	20/1	Nervous system neoplasms	204/9
Emerin complex 52	Emerin	305	25/9	Nervous system diseases	12/2
BCL6-HDAC7 complex	BCL6	758	684/1	Nervous system neoplasms	1/4
Anti-HDAC2 complex	HDACs	1047	641/1	Nervous system diseases	250/2
RNA polymerase II complex, chro structure modifying	RNA polymerase II complex	1529	568/2	Nervous system neoplasms	1/7
SMARCA2/BRM-BAF57-MECP2 complex	SMARCA2	112	81/1	Nervous system diseases	30/13
CDC2-CCNA2 complex	CDC2	3777	2581/1	Nervous sys neoplasms	15/5
	CCNA2	270	207/1	Nervous system diseases	126/5
CAV1-VDAC1-ESR1 complex	VDAC1	99	45/1	Nervous sys neoplasms	7/9
	CAV1	880	382/1	Nervous system diseases	193/7
TGF-beta receptor II-TGF-beta3 complex	TGFB3	874	259/1	Nervous sys neoplasms	13/7
Retrotranslocation complex	GEMIN4	26	9/2	Retinoblastoma	13/2
TRAP complex	SYVN1	36	8/5	Eye neoplasm	13/1
RAF1-RAS complex, EGF induced	Mediator complex	2518	805/1	Eye diseases	464/4
	RAF1	271	174/1	Eye neoplasm	444/1
	Ras	31223	22723	Eye diseases	25/4
APLG1-Rababtin5 complex	Rab effec protein	141	38/1	Eye neoplasm	22/4
WINAC complex	SMARCA2	112	81/1	Nervous sys diseases	31/2
	SMARCA4	219	155/1	Neuroblastoma	5/2
	SMARCB1	291	282/1	Nervous sys diseases	143/3
SNARE complex (STX11, VAMP2, SNAP23)	VAMP2	204	29/4	Nervous system neoplasms	12/7
MCM8-ORC2-CDC6 complex	CDC6	208	104/1	Nervous system diseases	89/11
				Nervous system neoplasms	9/8
				Nervous sys diseases	15/1
				Nervous sys diseases	9/5
				Nervous system diseases	397/4
				Nervous system neoplasms	20/6
				Nervous system diseases	44/6
				Nervous system neoplasms	8/5
				Nervous system diseases	2166/12
				Nervous system neoplasms	881/8
				Nervous sys diseases	19/5
				Nervous system diseases	11/8
				Retinoblastoma	13/1
				Eye diseases	20/7
				Retinoblastoma	17/2
				Nervous system diseases	121/2
				Nervous system neoplasms	116/1
				Nervous system diseases	41/3
				Eye diseases	13/8
				Retinoblastoma	12/2

and 7, respectively. Then we applied overlap score to quantify the similarity between the two partitions of cancers respectively generated from gene expression and protein complex profiles^{45,46} and got the value of overlap score as 0.72. We then generated 200 pairs of random clusters of the cancers, in which the cluster sizes are the same as in the real data. The average overlap score of the random ensemble was calculated as 0.24, while the z-score⁴⁶ for the overlap score of the two real partitions was 8.15, suggesting a fairly high extent of overlap between the two partitions of cancers with statistical significance. These results suggest that our predictions of complexes extract cancer modules from the expression data *while not changing the inherent correlations of the data*. Therefore, we can see that they reflect the intrinsic relationships among different cancers.

Discussion

Studies on the differential gene expression levels have added significant values to the genome-wide analyses having focused on genome

sequencing, due to their condition-dependent dynamic nature. In other words, they indicate how the biological functions are phenomenologically realized for given “blueprints” of genome sequences and different environments. Our method can successfully identify cancer-associated complexes. We believe that it, from the assumption that protein complexes are real biological functional units, leads us to one step closer to biological reality.

Our optimization procedure is based on linear programming (polynomial in computational time), implying that our method is feasible for future, larger studies. The method, as we apply it in this paper, rests on the assumption that expression levels are strongly correlated to protein abundance. Although signals from Affymetrix arrays used in our data sets can differ from the absolute protein abundance, considering the dataset’s broad coverage in terms of both cancers and various tissues, this study provides a novel approach that can be adopted by other researchers who are possibly in possession of better datasets currently or in the future, we believe. Moreover, the

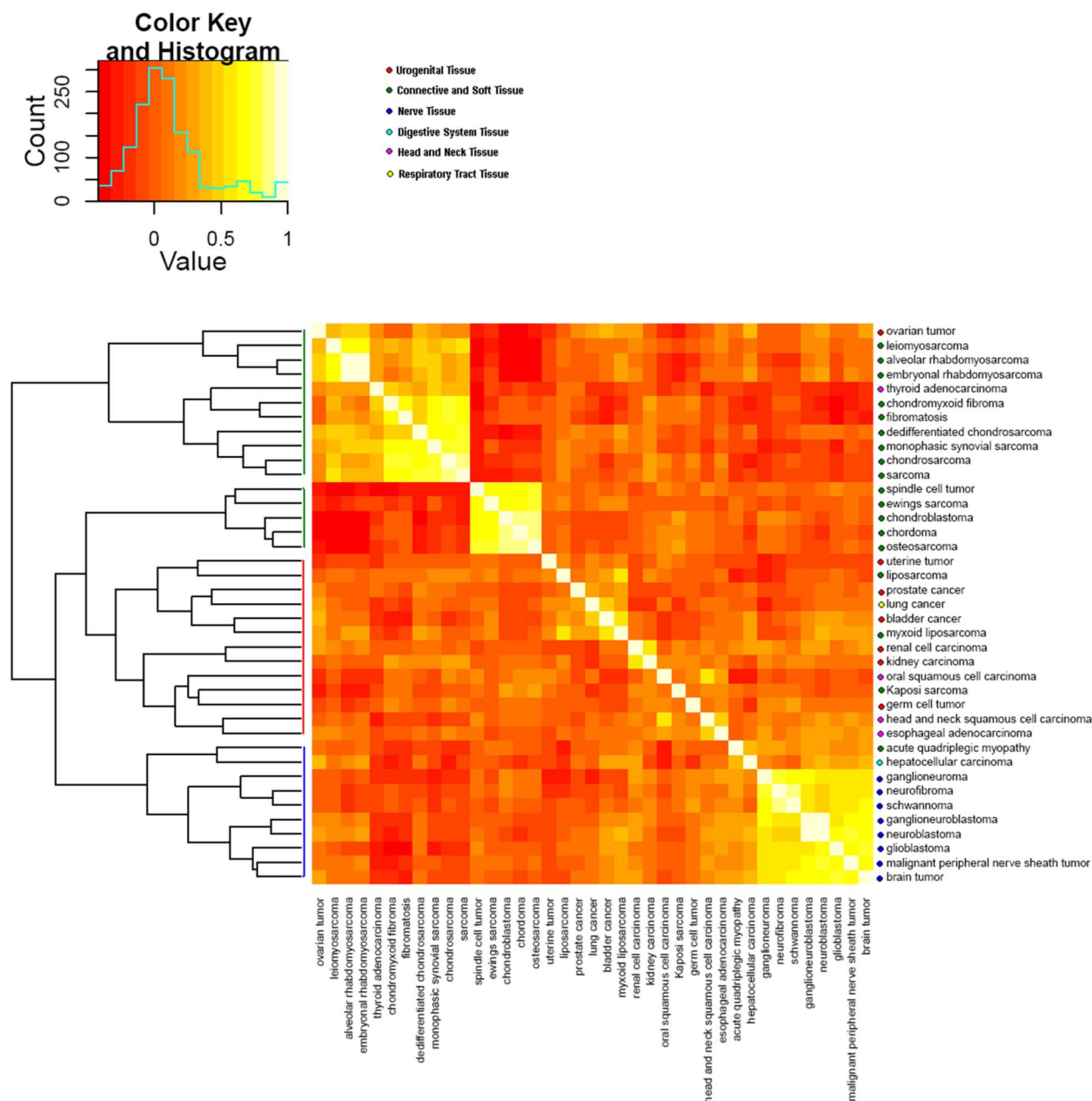


Figure 6 | Heat map and hierarchical clustering of 39 cancers. Similarity between cancers is defined as the Pearson correlation coefficient between the log-ratio expression profiles of genes that cancers contain.

advantage of protein-complex-based approaches, other than the identification of cancer-specific complexes, could be investigated further in the future.

Methods

Gene expression dataset. For gene expression data, we use recently released E-MTAB-62 in the Array Express repository³². It is an integration of 206 different experiments and 5372 samples generated in 163 different laboratories, including 369 different cell and normal tissue types, diseases, and cell lines. The most important aspect of this dataset is that all the data are from the same platform, pass data quality checks and get normalized so that we can compare the expression levels *across different cancers/tissues*. CEL files of samples that did not pass quality checks were removed. The retaining 5372 CEL files were normalized by Robust Multi-array Average (RMA) method, i.e., the raw intensity values were background corrected, log₂ transformed and then quantile normalized⁴⁷. In this work, we studied 39 solid tissue cancers (708 samples) and 25 normal solid tissue types (440 samples), in which 18 normal tissue types were where these cancers originated and thus were used as control sets (see Table 1).

Protein complex dataset. For the list of human protein complexes, we use the Comprehensive Resource of Mammalian protein complexes (CORUM) database³³,

where 1343 complexes and 2315 component proteins (the expression profiles of 2064 of these 2315 proteins are listed in E-MTAB-62 data) are listed in total as a core data. Among the core data, 1338 complexes, at least one of component proteins of which is assigned with the expression profile in E-MTAB-62 data, are used in our analysis.

Estimation of abundance levels of complexes based on optimization. The detailed background and procedure of our optimization algorithm is described in Ref. 15. Assume that the copy number of protein i ($i = 1, \dots, N$; N is the number of proteins) and the number of complex j ($j = 1, \dots, M$; where M is the number of complexes) are given by P_i and c_j , respectively. Also, suppose that we denote the number of protein i in the complex j as S_{ij} , where $S_{ij} = 0$ if the complex j does not include the protein i as its component. In the ideal situation where all the proteins in a cell are of the exact amount to be used in forming a complex, the variable sets $\{P_i\}$ and $\{c_j\}$ satisfy

$$P_i = \sum_{j=1}^M S_{ij} c_j \quad (1)$$

The question is how to determine $\{c_j\}$ (variables) with known values of $\{P_i\}$ and $\{S_{ij}\}$ (constants). However, since the number of proteins N is usually larger than the number of complexes M , the set of linear equations above is *over-determined*, so in general it is not possible to satisfy all the equations in Eq. (1). In reality, therefore, the

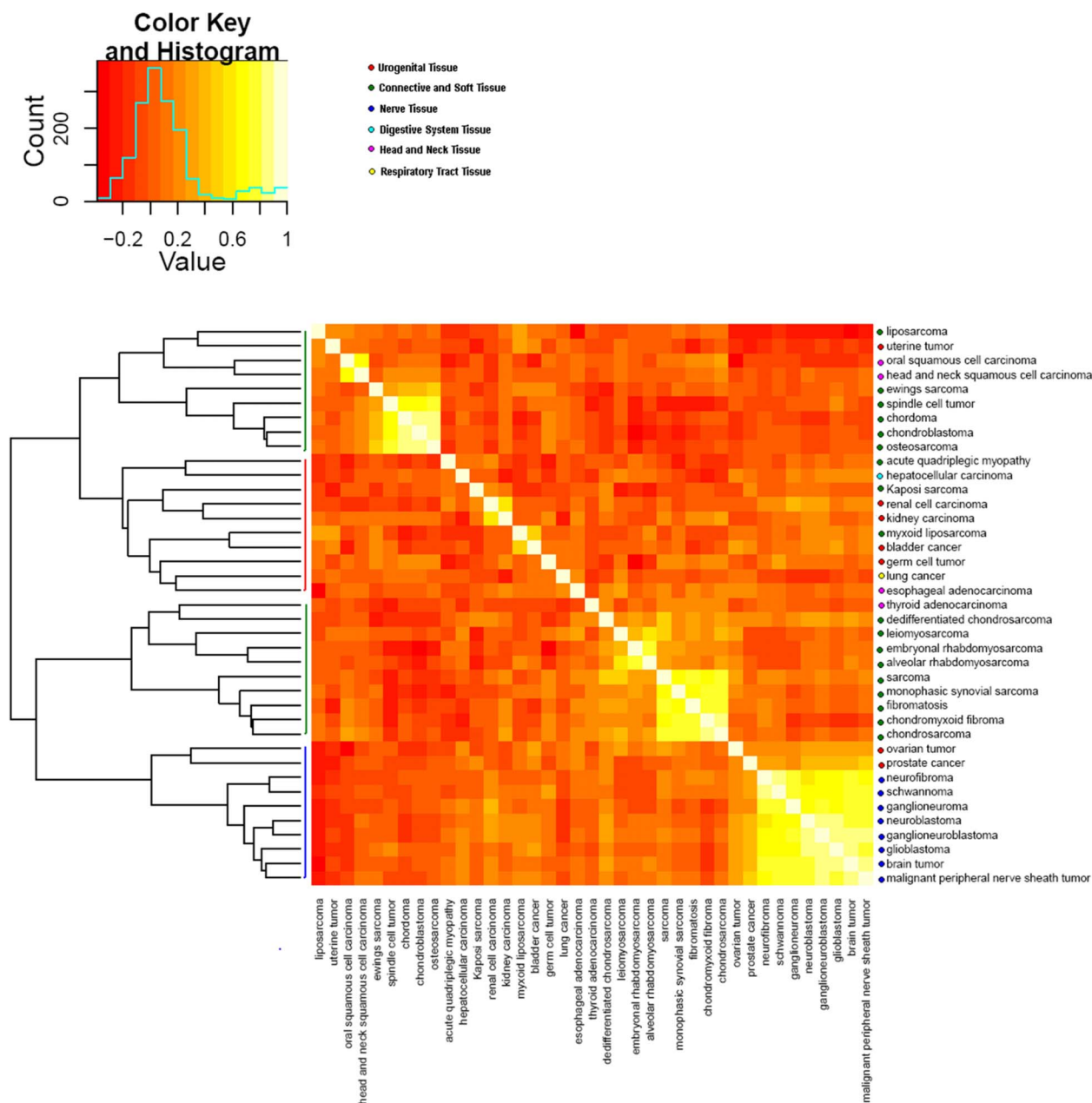


Figure 7 | Heat map and hierarchical clustering of 39 cancers. Similarity between cancers is defined as the Pearson correlation coefficient between the log-ratio abundance profiles of complexes that cancers contain.

number of proteins in a cell should be greater than or equal to that necessary to form complexes, i.e., $p_i \geq \sum_{j=1}^M S_{ij}c_j$, which is the basic constraint of our optimization scheme. Instead of finding an exact solution satisfying Eq. (1), we try to minimize the deviation from the ideal situation in Eq. (1), given by the object function

$$DA = \sum_{i=1}^N \left[1 - \sum_{j=1}^M (S_{ij}c_j) / p_i \right] \quad (2)$$

where the summation is only for indices i where $P_i > 0$. Now, for the given values of P_i and $\{S_{ij}\}$, our basic strategy is to determine c_j values that minimize DA in Eq. (2), and this problem is numerically solved by the linear programming (LP) technique. Moreover, after the determination of c_j values, if some values of P_i are unknown, we can assign those values of P_i using Eq. (1) for the ideal situation. This optimization is based on an assumption that organisms have been evolved in a way that increases efficiency by reducing wasted resources.

In this work, the average expression level of gene encoding protein i is used as the P_i -value³³ and the composition matrix S_{ij} is approximated by the binary value ($= 1$ if protein i is included in complex j , 0 otherwise). Ideally, it is more realistic to estimate protein complex levels from protein abundance as mRNA expression level cannot

completely represent the true protein abundance. However, although several large proteomics data sets are available^{48,49}, currently there are no equally rich genome-wide protein abundance data sets for tumor versus normal tissue samples. Several studies have found mRNA and protein expression levels to be well correlated^{50,51}. It is reported that approximately 40% of the variation in mammalian protein abundance is explained by mRNA levels⁵¹. It is known that signals from Affymetrix arrays used in our data sets can differ from the absolute protein abundance. However, our method does not, strictly speaking, need to use absolute abundances - it is sufficient that the relative abundances are accurately measured, since all the objective functions and constraints in our linear programming (LP) optimization are strictly linear by definition. Therefore, the direct usage of gene expression levels as protein abundance is not free from errors, but it could yield reasonable results.

Identification of differentially expressed complexes in cancers and normal tissues.

For each cancer or tissue case, individual genes' expression profiles are averaged over different samples in the E-MTAB-62 dataset, and the set is used as the input data of $\{p_i\}$ set. Our optimization procedure minimizing Eq. (2) will yield the $\{c_j\}$ set, i.e., complexes' abundance levels for the cancer or tissue. Then the abundance levels of all complexes in different cancers are compared with the abundance levels in the corresponding normal tissue in which these cancers originated; while the abundance



levels of all complexes in different normal tissues are compared with the average abundance levels in all the other normal tissues. Over-expression(under-expression) of a complex is defined as at least 2-fold (at most 1/2-fold) change of abundance level.

Overlap score. We use overlap score to measure the overlap extent of cancer clusters respectively generated from gene expression and protein complex profiles^{45,46}. Consider two different categories A and B (for example, two partitions of cancers got by different clustering methods) and assume each cancer is associated with a subset (cluster) of the partitions of A and B. Let $\phi_A(i)$ and $\phi_B(j)$ denote the fraction of cancers in cluster $i \in A$ and $j \in B$ ($i = 1, 2, \dots, m$; $j = 1, 2, \dots, n$), respectively. Let $\phi_{AB}(i, j)$ denote the joint frequency of i and j , i.e., the fraction of cancers that are partitioned in both cluster $i \in A$ and $j \in B$. In a random distribution of clusters the expectation value of $\phi_{AB}(i, j)$ is $\phi_A(i)\phi_B(j)$. If the clusters of differ partitions are overlapping, some $\phi_{AB}(i, j)$, the ones that overlap, will be larger than $\phi_A(i)\phi_B(j)$, while for the others, $\phi_{AB}(i, j)$ will be smaller than $\phi_A(i)\phi_B(j)$. Thus, the overlapping of clusters in partitions A and B can be quantitatively measured by:

$$\mu_{AB} = \frac{\sum_{i=1}^m \sum_{j=1}^n |\phi_{AB}(i, j) - \phi_A(i)\phi_B(j)|}{\sum_{i=1}^m \sum_{j=1}^n \phi_{AB}(i, j)} \quad (3)$$

Since the value of μ is affected by finite sizes, it is hard to judge if a μ -value indicates a good or bad overlap. Therefore, we normalize the μ -value against those of the perfect overlaps and define overlap score of partitions A and B as follows:

$$v_{AB} = \frac{\mu_{AB}}{\max(\mu_{AA}, \mu_{BB})} \quad (4)$$

The value of v is between 0 and 1, and it is 1 for perfect matches.

- Mohammadi, A., Sarrae, M. & Salehi, M. Identification of disease-causing genes using microarray data mining and Gene Ontology. *BMC Medical Genomics* **4**, 12 (2011).
- Ruan, X., Wang, J., Li, H., Perozzi, R. E. & Perozzi, E. F. The use of logic relationships to model colon cancer gene expression networks with mRNA microarray data. *Journal of Biomedical Informatics* **41**, 530–543 (2008).
- Nitsch, D. *et al.* Network Analysis of Differential Expression for the Identification of Disease-Causing Genes. *PLoS ONE* **4**, e5526 (2009).
- Zhao, J., Yang, T.-H., Huang, Y. & Holme, P. Ranking Candidate Disease Genes from Gene Expression and Protein Interaction: A Katz-Centrality Based Approach. *PLoS ONE* **6**, e24306 (2011).
- Wu, X., Jiang, R., Zhang, M. Q. & Li, S. Network-based global inference of human disease genes. *Molecular Systems Biology* **4**, 189 (2008).
- Yao, X., Hao, H., Li, Y. & Li, S. Modularity-based credible prediction of disease genes and detection of disease subtypes on the phenotype-gene heterogeneous network. *BMC Systems Biology* **5**, 79 (2011).
- Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
- Ito, T. *et al.* Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A* **97**, 1143–7 (2000).
- Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences* **98**, 4569–4574 (2001).
- Yu, H. *et al.* High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science* **322**, 104–110 (2008).
- Puig, O. *et al.* The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification. *Methods* **24**, 218–229 (2001).
- Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotech* **17**, 1030–1032 (1999).
- Gavin, A.-C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
- Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
- Lee, S. H., Kim, P.-J. & Jeong, H. Global organization of protein complexome in the yeast *Saccharomyces cerevisiae*. *BMC Syst Biol* **5**, 126 (2011).
- Parsons, D. W. *et al.* An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science* **321**, 1807–1812 (2008).
- Jones, S. *et al.* Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science* **321**, 1801–1806 (2008).
- Hwang, S. *et al.* A protein interaction network associated with asthma. *Journal of Theoretical Biology* **252**, 722–731 (2008).
- Qiu, Y.-Q., Zhang, S., Zhang, X.-S. & Chen, L. Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics* **11**, 26 (2010).
- Suthram, S. *et al.* Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets. *PLoS Computational Biology* **6**, e1000662 (2010).
- Zhao, J., Chen, J., Yang, T.-H. & Holme, P. Insights into the pathogenesis of axial spondyloarthritis from network and pathway analysis. *BMC Systems Biology* **6**, S4 (2012).
- Liu, K.-Q., Liu, Z.-P., Hao, J.-K., Chen, L. & Zhao, X.-M. Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics* **13**, 126 (2012).
- Bossi, A. & Lehner, B. Tissue specificity and the human protein interaction network. *Mol Syst Biol* **5**, 260 (2009).
- Liang, S., Li, Y., Be, X., Howes, S. & Liu, W. Detecting and profiling tissue-selective genes. *Physiol Genomics* **26**, 158–162 (2006).
- Axelens, J., Lotem, J., Sachs, L. & Domany, E. Genes overexpressed in different human solid cancer exhibit different tissue-specific expression profiles. *Proc Natl Acad Sci U S A* **104**, 13122–13127 (2007).
- Wang, L., Srivastava, A. & Schwartz, C. Microarray data integration for genome-wide analysis of human tissue-selective gene expression. *BMC Genomics* **11**, S15 (2010).
- Chang, C. *et al.* Identification of Human Housekeeping Genes and Tissue-Selective Genes by Microarray Meta-Analysis. *PLoS ONE* **6**, e22859 (2011).
- Liu, X., Yu, X., Zack, D., Zhu, H. & Qian, J. TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics* **9**, 271 (2008).
- Wang, L., Srivastava, A. & Schwartz, C. Microarray data integration for genome-wide analysis of human tissue-selective gene expression. *BMC Genomics* **11**, S15 (2010).
- Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotech* **25**, 309–316 (2007).
- Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci U S A* **105**, 20870–20875 (2008).
- Lukk, M. *et al.* A global map of human gene expression. *Nat Biotechnol* **28**, 322–324 (2010).
- Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes 2009. *Nucleic Acids Research* **38**, D497–D501 (2010).
- Ruepp, A. *et al.* The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research* **32**, 5539–5545 (2004).
- Parham, P. *The Immune System* (Garland Science, New York, 2009).
- Safra, M. *et al.* GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* **18**, 1542–1543 (2002).
- Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).
- Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–15550 (2005).
- Lotem, J., Netanel, D., Domany, E. & Sachs, L. Human cancers overexpress genes that are specific to a variety of normal human tissues. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 18556–18561 (2005).
- Lee, S. H., Kim, P.-J., Ahn, Y.-Y. & Jeong, H. Googling Social Interactions: Web Search Engine Based Social Network Construction. *PLoS ONE* **5**, e11233 (2010).
- Ward, J. Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* **58**, 236–244 (1963).
- Lance, G. & Williams, W. A general theory of classificatory sorting strategies. *Comput J* **9**, 373–380 (1967).
- Rogers, F. Medical subject headings. *Bull Med Libr Assoc Bull Med Libr Assoc* **51**, 114–116 (1963).
- Doms, A. & Schroeder, M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research* **33**, W783–W786 (2005).
- Holme, P. Model validation of simple-graph representation of metabolism. *J. Roy. Soc. Interface* **40**, 1027–1034 (2009).
- Zhao, J. *et al.* Reconstruction and Analysis of Human Liver-Specific Metabolic Network Based on CNHLPP Data. *Journal of Proteome Research* **9**, 1648–1658 (2010).
- Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Bioinformatics* **4**, 249–264 (2003).
- Ponten, F. *et al.* A global view of protein expression in human cells, tissues, and organs. *Molecular Systems Biology* **5**, 337 (2009).
- Huttlin, E. L. *et al.* A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. *Cell* **143**, 1174–1189 (2010).
- Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotech* **25**, 117–124 (2007).
- Schwanhauss, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

Acknowledgments

This work was supported financially by National Natural Science Foundation of China, grants no. 10971227 and 81260672 (J.Z.); the Swedish Research Council (S.H.L. and P.H.) and the World Class University program through National Research Foundation of Korea funded by Ministry of Education, Science and Technology R31–2008–10029 (P.H.). The



authors greatly thank Pan-Jun Kim for providing the LP optimization code and Daniel Kim for the Web search data from Google Search API.

Author contributions

J.Z., S.H.L., M.H. and P.H. conceived the study; S.H.L., J.Z. and M.H. developed the methods and analyzed the data; J.Z., S.H.L. and P.H. wrote the paper. All authors read an approved the final version of the manuscript.

Additional information

Supplementary information accompanies this paper at <http://www.nature.com/scientificreports>

Competing financial interests: The authors declare no competing financial interests.

License: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>

How to cite this article: Zhao, J., Lee, S.H., Huss, M. & Holme, P. The Network Organization of Cancer-associated Protein Complexes in Human Tissues. *Sci. Rep.* **3**, 1583; DOI:10.1038/srep01583 (2013).