SCIENTIFIC REPORTS

OPEN

Correspondence and requests for materials should be addressed to L.X. (lyxu@stu.edu.cn) or E.L. (nmli@stu.edu. cn)

* These authors contributed equally to this work.

Network based analyses of gene expression profile of *LCN2* overexpression in esophageal squamous cell carcinoma

Bingli Wu[1]*, Chunquan Li[1,2]*, Zepeng Du[3], Qianlan Yao[4], Jianyi Wu[1], Li Feng[4], Pixian Zhang[1], Shang Li[2], Liyan Xu[5] & Enmin Li[1]

[1]Department of Biochemistry and Molecular Biology, Shantou University Medical College, Shantou515041, China, [2]College of Medical Informatics, Daqing Campus, Harbin Medical University, Daqing163319, China, [3]Department of Pathology, Shantou Central Hospital, Shantou515041, China, [4]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin150081, China, [5]Institute of Oncologic Pathology, Shantou University Medical College, Shantou515041, China.

LCN2 (lipocalin 2) is a member of the lipocalin family of proteins that transport small, hydrophobic ligands. *LCN2* is elevated in various cancers including esophageal squamous cell carcinoma (ESCC). In this study, *LCN2* was overexpressed in the EC109 ESCC cell line and we applied integrated analyses of the gene expression data to identify protein-protein interactions (PPI) network to enhance our understanding of the role of *LCN2* in ESCC. Through further mining of PPI sub-networks, hundreds of differentially expressed genes (DEGs) were identified to interact with thousands of other proteins. Subcellular localization analyses found the DEGs and their directly or indirectly interacting proteins distributed in multiple layers, which was applied to analyze the possible paths between two DEGs. Gene Ontology annotation generated a functional annotation map and found hundreds of significant terms, especially those associated with the known and potential roles of LCN2 protein. The algorithm of Random Walk with Restart was applied to prioritize the DEGs and identified several cancer-related DEGs ranked closest to LCN2 protein. These analyses based on PPI network have greatly expanded our understanding of the mRNA expression profile of *LCN2* overexpresssion for future examination of the roles and mechanisms of *LCN2*.

*L*CN2 (lipocalin 2), also known as oncogene 24p3, uterocalin, siderocalin or neutrophil gelatinase associated lipocalin (NGAL), is a 24 kDa secreted glycoprotein and a member of the lipocalin family of proteins that transports small, hydrophobic ligands[1]. LCN2 protein is secreted into the extracellular environment and forms a heterodimer with matrix metalloproteinase-9 (MMP-9) through disulfide bonds, modulating the stability rather than the enzymatic activity of MMP-9[2]. By sequestering iron-laden siderophores, LCN2 deprives bacteria of a vital nutrient and thus inhibits their growth, suggesting its bacteriostatic effect or protection against bacterial infection[3]. Its small size, secreted nature and relative stability have led to it being investigated as a diagnostic and prognostic biomarker in many acute diseases, especially in acute kidney injury[4].

Dysregulated of *LCN2* has been observed in several benign and malignant diseases, including breast, colorectal, pancreatic, ovarian, gastric, thyroid, ovarian, and bladder, as well as kidney cancers[5]. Elevated LCN2 participates in various functions in malignant cells, even sometimes the conclusions were controversial. LCN2 inhibits apoptosis in thyroid cancer and decreases invasion and angiogenesis in pancreatic cancer, but increases proliferation and metastasis in breast and colon cancer[6]. Our previous studies have demonstrated that LCN2 is elevated in esophageal squamous cell carcinoma (ESCC), and its upregulation significantly correlates with cell differentiation and tumor invasion and could served as an independent prognostic factor[7,8]. To better understand the biological role of LCN2 in ESCC, we overexpressed *LCN2* in the EC109 ESCC cell line. Subsequently, Agilent whole genome oligo microarray (Agilent Technologies, USA) was applied for mRNA expression profile and hundreds of differentially expressed genes (DEGs) were obtained from *LCN2* overexpressed cell comparing with its control (data prepared in other manuscript).

Network-based analyses of protein-protein interactions (PPI) utilize known associations among the protein molecules to globally describe the interactions of these associations in context of of biochemistry, signal trans-

duction and biomolecular networks. Virtually all proteins perform specific functions through the interactions with other proteins in specific biological contexts[9]. In the recent years, the integrated analysis of large-scale gene expression data with PPI networks has received considerable attention[10,11]. Knowledge of the PPI network provides a number of applications, such as prediction of proteins interaction and protein function, and identification of functional protein modules, disease candidate genes identification, and drug targets identification[12,13].

To acquire a more global biological context for mRNA expression profiles, analyses should exceed the merely listing of affected genes and extend our knowledge to explain the enhanced biological phenotype resulting from the cascades of spatial or temporal interactions of target genes with other proteins. In this study, we analyzed the mRNA expression profile of LCN2 overexpression in ESCC using system biology method based on the knowledge of PPI network.

Results

PPI sub-network of DEGs derived from LCN2 overexpression. More than 200 DEGs, including 167 upregulated genes and 96 downregulated genes, were obtained, using a 2-fold threshold, from the mRNA expression profile following LCN2 overexpression. In order to gain insight into how the DEGs affected cellular biological activity, a full screen of their interactions with other proteins would provide important clues of their functions. The combination of PPI datasets from both acknowledged HPRD and BioGRID databases provides credible original data for subsequent analyses. Three kinds of PPI sub-networks were generated by mapping the downregulated, upregulated and total DEGs to the parental PPI network, respectively. Fifty-five downregulated DEGs had literature on interacting proteins, which formed a PPI sub-network with their first neighboring proteins that contains 834 nodes and 7005 edges (Supplementary Figure S1). On the other hand, eighty-two upregulated proteins had reported interacting proteins and formed a PPI sub-network with their first neighboring proteins containing 1813 nodes and 23380 edges (Supplementary Figure S2). The total DEG PPI sub-network was composed of 2458 nodes and 33671 edges, including 135 DEGs (Figure 1A). These three sub-networks indicated that the overexpression of LCN2 greatly disturbs the PPI network in ESCC as hundreds of DEGs interacted with thousands of other proteins to enlarge the biological consequences of its overexpression.

To focus on LCN2 protein, a PPI sub-network based on the axis of LCN2 → interacting proteins → DEGs → interacting proteins was also built to detect the relationship between LCN2 and the nearest DEG proteins. This central LCN2 sub-network contained 121 nodes and 132 edges, including 8 DEGs, the downregulated TGFB1, COL4A3, COL4A4, SDC2 and DCN, the upregulated LCN2, AREG and A2M. Currently, only four LCN2-interacting proteins (MMP2, MMP9, HGF and LRP2) have been reported and collected by HPRD and BioGRID, and their expression levels did not significantly change in our mRNA profile of LCN2 overexpression in ESCC (Figure 1B). However, three of the LCN2-interacting proteins interacted with LCN2 overexpression-related DEGs, such as MMP9 interaction with the downregulated TGFB1, COL4A3, COL4A4 and the upregulated A2M, HGF interaction with downregulated SDC2, and MMP2 interaction with downregulated DCN.

To detect whether there are internal interactions between DEGs, the DEG-DEG interactions were acquired. This sub-network contained 18 nodes (10 downregulations and 8 upregulations) and 17 edges, including a small module composed of 11 DEGs, a four-DEG interactions and two two-DEG interactions (Figure 1C).

Network topological properties. Dependent on its distinguishing topological characteristics, the real biological networks (e.g. the PPI network) are significantly different from random networks. The power law of node degree distribution is one of most important criteria[14,15]. The distributions of node degree approximately followed power law distributions, with an $R^2 = 0.844$, 0.814 and 0.866 for the downregulated, upregulated and total DEGs sub-networks, respectively (Figure 2). This suggests that the three PPI sub-networks were scale-free, which is one of most important characteristics of true complex biological networks[16]. These results also indicate that a few protein nodes act as hubs with a large number of links to other protein nodes. Other topological parameters of these sub-networks, such as clustering coefficient, network centralization and network density are shown in Table 1. Several special network elements, including closeness centrality, topological coefficients, neighborhood connectivity distribution and average clustering coefficient distribution are indicated in Supplementary Fig. S3 with their definitions were described in Supplementary Text S1.

Subcellular localization of proteins in the PPI sub-networks. The appropriate subcellular localization and their translocations of proteins are crucial because they provide the physiological context for their function, such as complex formation, signal transduction, and protein modification. With Cerebral plugin, nodes were redistributed according to their intracellular localization without changing their connecting neighbors. The total DEG sub-network was divided into 9 layers in this study with their percentage as follows: Secreted (6.8%), Membrane (11.2%), Cytoskeleton (4%), Cytoplasm (33.1%), Secreted/Nucleus (1.4%), Cytoskeleton/Nucleus (0.9%), Cytoplasm/Nucleus (14.7%), Nucleus (20.1%) and Unknown (7.7%) (the proteins without subcellular location annotation) (Figure 3A). The subcellular locations of proteins in the total DEG PPI sub-network range from extracellular to intracellular and even nucleus. We also found at least 12 DEGs are able to translocate from cytoplasm to nucleus (Supplementary Table S1).

The subcellular location of LCN2 is variable, depending on its cellular functions. LCN2 is able to be secreted to the extracellular space, forming a complex with MMP9 by disulfide bond linkage, protecting MMP9 from proteolytic degradation to enhance tumoral invasiveness and diffusion[2]. The other principal characteristic of LCN2 is to capture iron-containing siderophores and transport them to the cell interior after interacting with specific membrane receptors (24p3R, megalin or NGALR), increasing cytoplasmic mineral levels and triggering the iron-dependent reactions[17–19]. The currently annotated LCN2-interacting proteins are mostly located in the extracellular space (e.g. MMP2, MMP9 and HGF), or in the membrane (LRP2). In addition to interacting with MMP2 and MMP9 extracellularly, LCN2 could also interact with its own receptor LRP2. Our previous study has identified a novel splicing variant of the LCN2 receptor in ESCC, and both LCN2 and its receptor are overexpressed in ESCC[19]. To detect whether there were any possibilities for LCN2 transform information into the nucleus, we also distributed the proteins of LCN2-central PPI sub-network according to their subcellular localizations. As shown in Fig. 3B, a dozens of LCN2 neighboring proteins, especially LRP2-interacting proteins, such as MAPK8IP1, HDAC7 and ANAPC10, were located in the nucleus or could translocate into the nucleus.

To further illustrate the strength of this kind analysis, we applied the shortest path algorithm to find the possible shortest path from LCN2 to FOXP1, and identify the linking proteins between LCN2 and FOXP1. We found 28 shortest paths from LCN2 to FOXP1 with all the path lengths equaling 4 (Table 2). In Table 2, we prioritized the list of paths first by the normalized intensity of LCN2 directly-interacting genes, followed by normalized intensities of subsequent genes participating sequentially down the signal cascade. For example, the four LCN2 interacting proteins were ranged by the order of MMP9, MMP2, HGF and LRP2 according their normalized intensity. Subsequently, the MMP9 interacting proteins were also ranged by
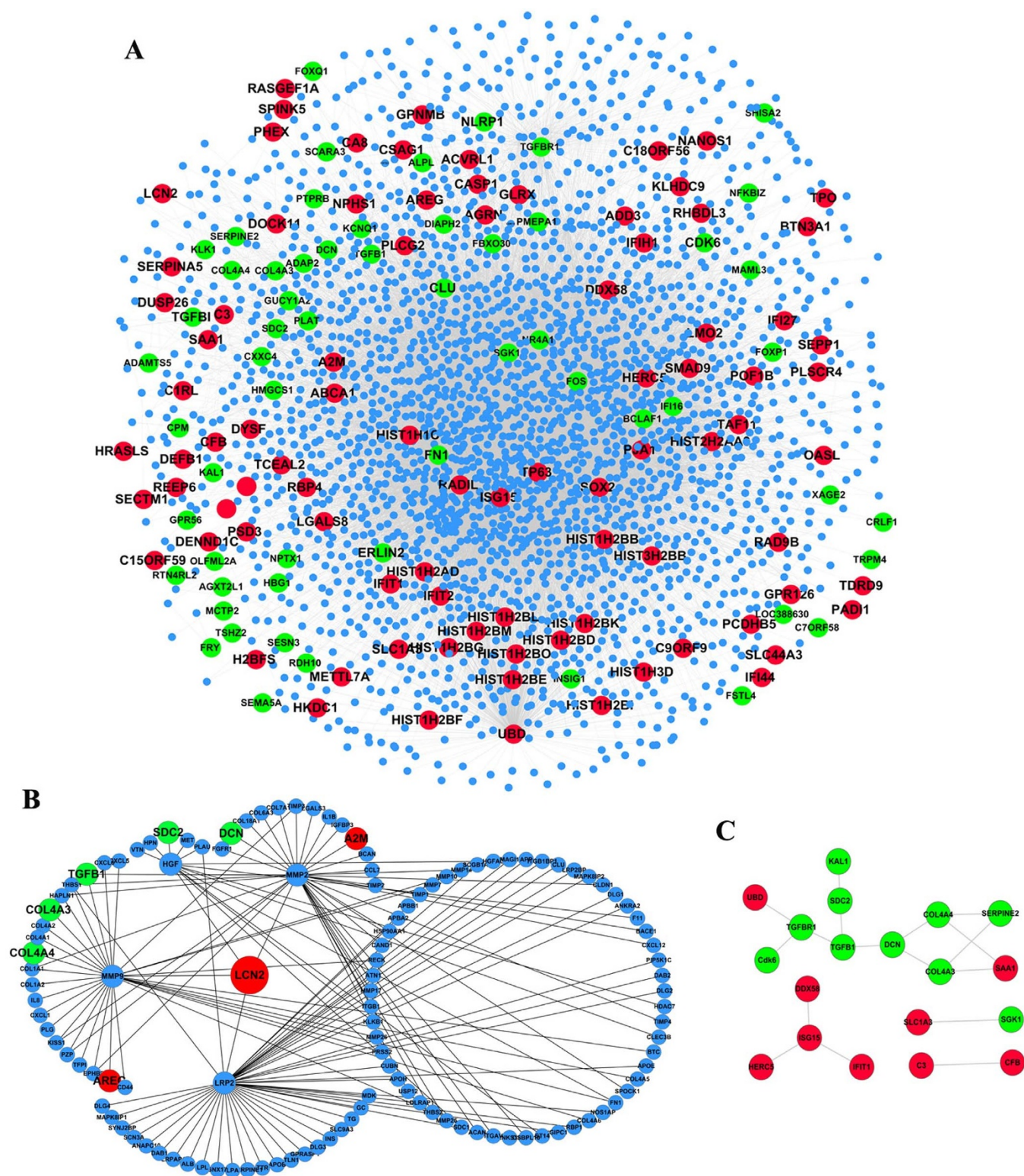
Figure 1 | PPI sub-network generation by mapping DEGs to the HPRD&BioGRID parental PPI network. (A) PPI sub-networks of total DEGs. (B) LCN2-central PPI sub-network. (C) Internal interactions of DEGs. Different colors of nodes indicate the types of proteins represented. Green and red nodes represent proteins encoded by down- and up-regulated genes, respectively. Blue nodes represent interacting proteins which were not significantly differentially expressed. The arrangement of nodes was applied to the "Spring Embedded" layout in Cytoscape.

the order of their normalized intensity (Supplementary Figure S4). We also distributed these proteins members in the paths according to their subcellular localizations. Most of these paths obey the principle of from extracellular to cytoplasm till nucleus (Figure 3C).

Functional annotation map of the PPI sub-network. Cellular activities, likely cancer-related, should be influenced by the DEGs through their interactions in the PPI network. To identify potential cellular activities related to LCN2 activity, we analyzed over-
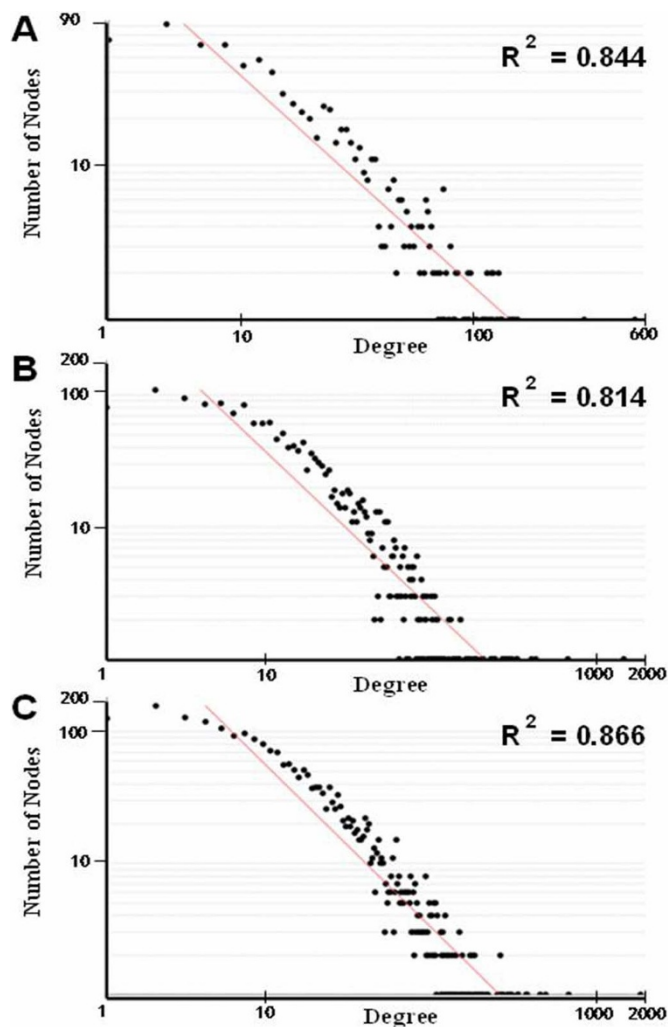
Figure 2 | Power law distribution of node degree. (A) Degree distribution of the downregulated DEG PPI sub-network. (B) Degree distribution of the upregulated DEG PPI sub-network. (C) Degree distribution of the total DEG PPI sub-network. The graph displays a decreasing trend of degree distribution with an increase in number of links displaying scale-free topology.

represented GO "Biological Process" terms of the total DEG PPI sub-network were analyzed. A functional annotation map containing 451 GO terms was generated in which proteins were ended up in nodes according to their enriched GO terms, with the edges connecting the GO terms indicative of proteins share the same enriched GO terms (Figure 4). To our great interest, several GO terms were potentially related to LCN2 functions. For example, a group of immunity-related terms were found, such as "regulation of immune respond", "activation of immune respond", "innate immune respond" and "deference

respond", etc. On the other hand, the proteins in the total PPI sub-network significantly involved the signal transduction. Many terms of different signal pathways were clustered, for example, "regulation of transforming growth factor beta receptor signaling pathway", "regulation of Wnt receptor signaling pathway", "immune response-regulating cell surface receptor signaling pathway". Another large GO term group was comprised of cell cycle-related GO terms, such as "G1 phase of mitotic cell cycle", "G1/S transition of mitotic cell cycle", "G2/M transition of mitotic cell cycle", "M

Table 1 | Topological parameters of three DEG PPI sub-networks

PPI sub-network	$y = \beta x^a$	Number of nodes	Number of edges	R^2	Correlation	Clustering coefficient[a]	Network centralization[b]	Network density[c]	Network diameter[d]
Downregulated DEGs	$y = 226.33x^{-1.132}$	834	7005	0.844	0.782	0.365	0.621	0.020	6
Upregulated DEGs	$y = 515.01x^{-1.172}$	1813	23380	0.814	0.662	0.331	0.803	0.014	6
Total DEGs	$y = 849.12x^{-1.218}$	2458	33671	0.866	0.696	0.306	0.739	0.011	6

[a]Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.
[b]Network centralization measures the degree of the effect when removing some central nodes in the whole network.
[c]Network density describes the portion of the potential connections in a network that are actual connections.
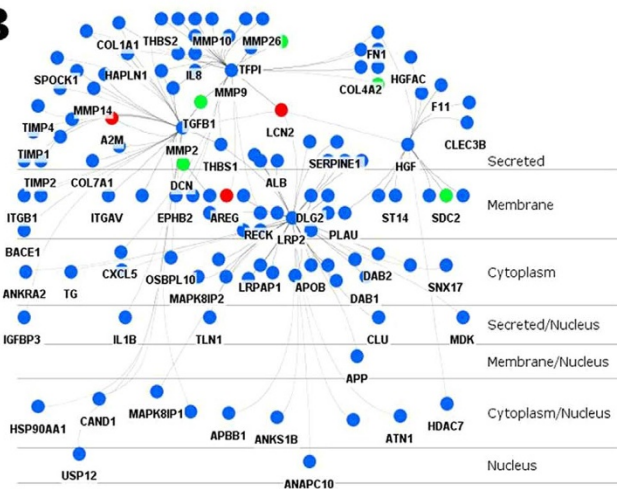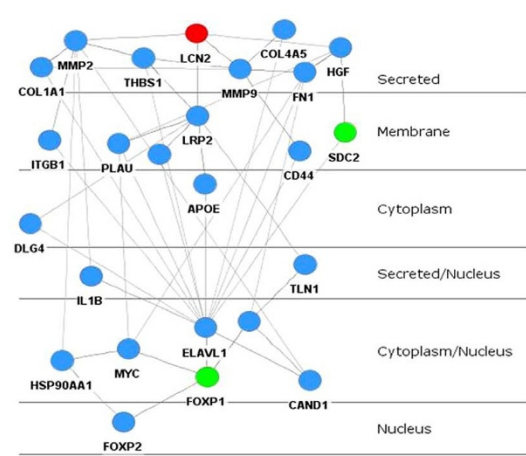[d]Network diameter representative of the linear size of a network.
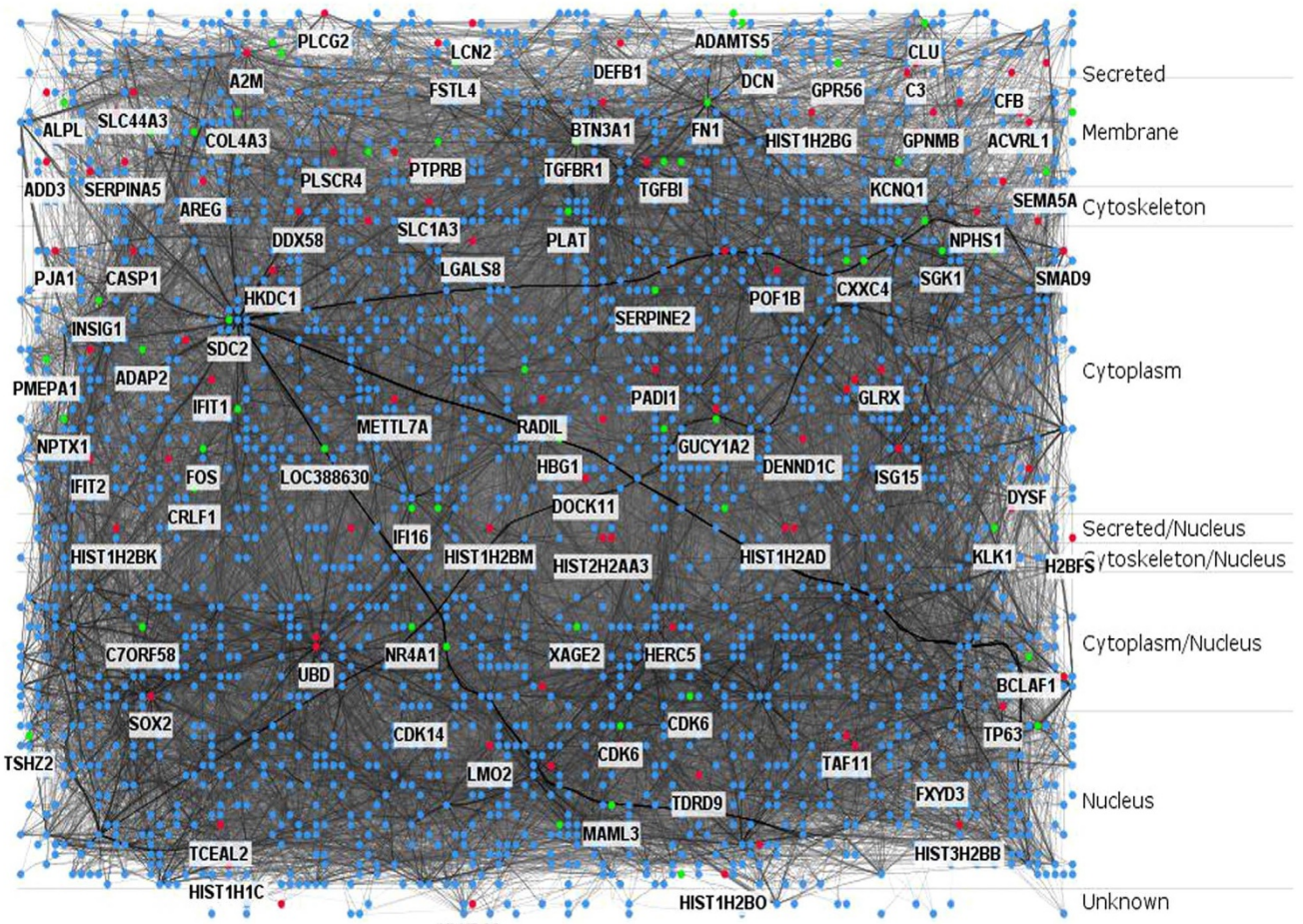
Figure 3 | Subcellular layers illustrating the PPI sub-network. (A) The total DEG PPI network. (B) LCN2-central PPI sub-network. (C) 28 possible paths from LCN2 to FOXP1.

phase of mitotic cell cycle", "M/G1 transition of mitotic cell cycle", suggesting LCN2 regulates the cell cycle. Two terms directly reflect the reported functions of LCN2 were also found, there were "cellular response to molecule of bacterial origin" and "extracellular matrix organization". The significant GO terms of interest were shown in Supplementary Table S2.

DEG prioritization. Since the overexpression of *LCN2* caused the expression change of hundreds of genes, it is interesting to detect how the DEGs were ranked by their importants when considering their relationship with LCN2. In this study, the RWR algorithm was used to analyze the closeness of proteins to LCN2 in the total DEG PPI network. Raw probability scores ranged from 0.705 to 7.96 e^{-9}. Since

Table 2	Possible shortest paths from LCN2 to FOXP1	
No.	The protein members of the path
1	LCN2 → MMP9 → CD44 → ELAVL1 → FOXP1
2	LCN2 → MMP9 → COL4A5 → ELAVL1 → FOXP1
3	LCN2 → MMP9 → THBS1 → ELAVL1 → FOXP1
4	LCN2 → MMP9 → FN1 → MYC → FOXP1
5	LCN2 → MMP9 → FN1 → SUMO2 → FOXP1
6	LCN2 → MMP9 → FN1 → ELAVL1 → FOXP1
7	LCN2 → MMP9 → COL1A1 → ELAVL1 → FOXP1
8	LCN2 → HGF → PLAU → MYC → FOXP1
9	LCN2 → HGF → PLAU → ELAVL1 → FOXP1
10	LCN2 → HGF → SDC2 → ELAVL1 → FOXP1
11	LCN2 → HGF → FN1 → MYC → FOXP1
12	LCN2 → HGF → FN1 → SUMO2 → FOXP1
13	LCN2 → HGF → FN1 → ELAVL1 → FOXP1
14	LCN2 → MMP2 → HSP90AA1 → MYC → FOXP1
15	LCN2 → MMP2 → HSP90AA1 → FOXP2 → FOXP1
16	LCN2 → MMP2 → ITGB1 → ELAVL1 → FOXP1
17	LCN2 → MMP2 → CAND1 → SUMO2 → FOXP1
18	LCN2 → MMP2 → CAND1 → ELAVL1 → FOXP1
19	LCN2 → MMP2 → THBS1 → ELAVL1 → FOXP1
20	LCN2 → MMP2 → IL1B → ELAVL1 → FOXP1
21	LCN2 → MMP2 → COL1A1 → ELAVL1 → FOXP1
22	LCN2 → LRP2 → DLG3 → ELAVL1 → FOXP1
23	LCN2 → LRP2 → PLAU → MYC → FOXP1
24	LCN2 → LRP2 → PLAU → ELAVL1 → FOXP1
25	LCN2 → LRP2 → TLN1 → SUMO2 → FOXP1
26	LCN2 → LRP2 → DLG4 → ELAVL1 → FOXP1
27	LCN2 → LRP2 → THBS1 → ELAVL1 → FOXP1
28	LCN2 → LRP2 → APOE → ELAVL1 → FOXP1

the scores of many nodes were very close, the scores were log10-transformed and to range from −1.27 to −8.10 (the more negative the score, the less significant.). The log-transformed score was regarded as the node attribute and displayed by the Cytoscape. The closer the protein to LCN2, the larger the node size (Figure 5A). The nodes of LCN2 interacting proteins (MMP2, MMP9, HGF and LRP2) were the biggest nodes, which was consistent with the idea of the algorithm of RWR. The DEGs alone are displayed in Fig. 5B for greater clarity in distinguishing differences. (Figure 5B). To better illustrate their closeness to LCN2, the DEGs were classified into different layers according to their range of score, e.g. only the seed node LCN2 was classified as the A layer, DEGs with a log-transformed of score −2.0 ~ −2.99 were classified as the B layer, and DEGs with a log-transformed score of −3.0 ~ −3.99 were classified as the C layer. The more negative the score, the further the node from LCN2. Based on Fig. 5B, these DEGs were rearranged into different layers also by the Celebral plugin (Figure 5C). As shown in Fig. 5C, downregulated SDC2, TGFB1 and DCN, upregulated A2M were ranked in the first closest class of DEGs to LCN2, while other DEGs such as AREG, PLAT were ranked in the second class, and so on. These result provided the prioritizations of DEGs when considering their relationship with LCN2.

Disscussion

Esophageal cancer is the sixth most common fatal human cancer in the world, and the histological type of squamous cell carcinoma is one of the most common cancers in the Chinese population[20,21]. Accumulated researches have illustrated that an integrative analysis of gene expression and PPI networks can provide deep insights into the molecular mechanisms of diseases, or the specific genes involved[22,23]. In this study, we applied a system approach by linking public PPI data with DEGs of *LCN2* overexpression to provide unique insights into the mechanisms of LCN2 from the network aspect. The three sub-networks for downregulated, upregulated and total DEGs were composed thousands of protein nodes, indicating

LCN2 influences other proteins directly or indirectly, and its over-expression disturbs the PPI network to alter cell function in ESCC. Second, this analysis provided a full screen of LCN2 directly-interacting proteins and their neighbor proteins, and this method is more effectively than merely literatures research and manually curation one by one. To our surprise, all four LCN2 interacting proteins (LRP2, MMP2, MMP9 and HGF) have been found over-expressed in ESCC[8,24,25]. Moreover, some of neighboring DEGs were also reported aberrant expression in ESCC. The upregulated DEG of A2M, the downregulated DEGs of DCN and TGFB1 are found enhanced in ESCC[25–27]. Our previous study showed found SDC2 mRNA down-regulation in ESCC is related to a poor prognosis[28]. These evidences suggested that our PPI sub-network could discover the links between LCN2 and other ESCC related genes (proteins). The topologies of the these three sub-networks showed that they are scale-free biological networks rather than a random networks, with their node degree distributions following a power law, one of most important network characters. This indicates that the overexpression of *LCN2* has truly disturbs the of PPI network in ESCC.

Since LCN2 can distribute both extracellularly and intracellularly and its overexpression causes broad changes in gene expression pro-files, it is interesting to understand how LCN2 signals are transduced from the cell exterior or within the cytoplasm to the nucleus. Subcellular localization offers important clues for proteins to reveal their participating pathways that regulate cellular activities at the subcellular level. Studies of cellular signal transduction processes indicate that classical signaling pathways are integrated parts of lar-ger molecular interaction networks[29]. We assumed that the signaling is transduced by sequential PPIs, since the composition and bio-logical role of proteins vary with subcellular localization. For example, proteins located in the plasma membrane are primarily involved in cell adhesion, cytoskeleton and cell signaling, whereas in the nucleus, proteins are mainly involved in transcription and ribosomal assembly. In this study, subcellular localization informa-tion was incorporated into total DEG PPI sub-network, generating biologically intuitive pathway-like layouts of a network. That many of the interacting proteins of LCN2 receptor LRP2 are able to trans-location into nucleus provides evidence for such a pathway. For example, MAPK8IP1 (mitogen-activated protein kinase 8 interact-ing protein 1), also named JNK-interacting protein-1 (JIP1), is a scaffolding protein that enhances JNK signaling by placing JNK and upstream kinases in proximity, which is critical in oncogenic transformation involving gene expression, cell survival, growth, dif-ferentiation and death[30,31]. In a like manner, overexpression of *LCN2* might influence the PPI network directly or indirectly, affecting the signaling of extracellular-membrane-cytoskeleton/cytoplasm-nucleus cascades to cause the altered expressions of DEGs and con-sequent alterations in cell proliferation, cell morphology, invasion and metastasis.

We assumed the elevated LCN2 protein would cause a wide range of mRNA expression profile alternation through the cascade of PPI activities, and the transcription factors or transcriptional regulators in the PPI sub-network play critical roles in this expression alterna-tion. So we were interested in the transcription factors or transcrip-tional regulators in our PPI sub-network. FOXP1 is a member of the FOX family of transcription factors which has a broad range of functions. FOXP1 overexpression is associated with poor prognosis in diffuse large B-cell lymphoma, gastric MALT lymphoma and hepatocellular carcinoma but with good prognosis in breast can-cer[32,33]. Tang *et al.* found 1473 potential target genes of FOXP1 using genome-wide expression microarrays and ChIP-seq in Huntington's disease[34]. Among these potential target genes list, we also found 6 downregulated DEGs of our LCN2 overexpression microarray result (COL4A4, EGR1, FOS, PGCP, PMP22, TGFBI). These suggested that the mRNA expression profile alternation following LCN2 over-expression were through some critical transcription factors. The
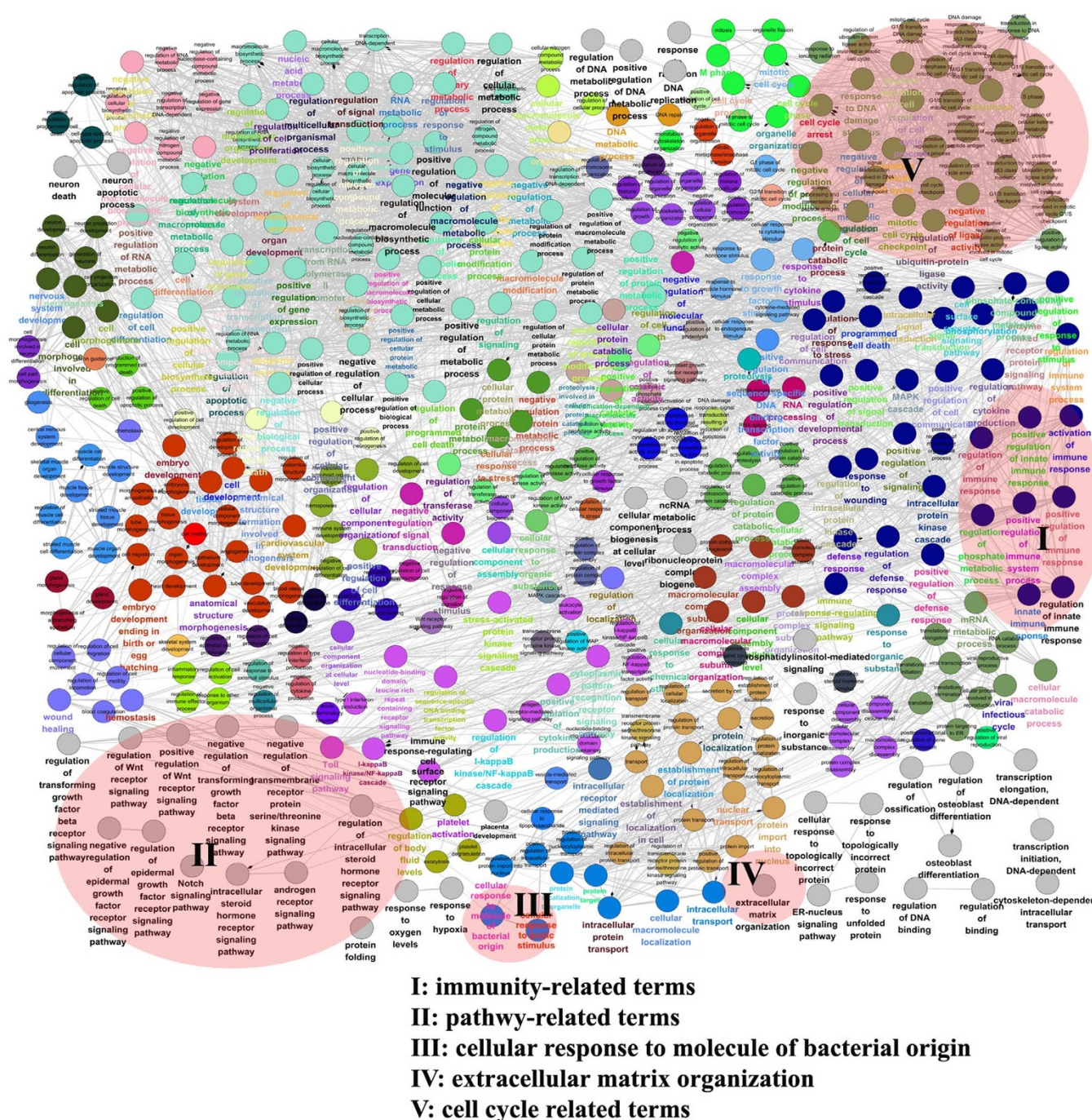
Figure 4 | Functional map of the total DEG PPI sub-network. Functionally grouped network with terms as nodes linked based on their kappa score level (≥0.3). Functionally related groups partially overlap. The similar GO terms were labeled in the same color. The interested GO term group related or potentially related to LCN2 function was indicated by a Roman numeral.

I: immunity-related terms
II: pathwy-related terms
III: cellular response to molecule of bacterial origin
IV: extracellular matrix organization
V: cell cycle related terms

other reason is that the expression level of FOXP1 was also changed, which might be regulated by other transcription factors. The alternation of FOXP1 expression might also change the expression level of its target genes. Thus the transcription regulational cascade signals were formed and genome-wide expression was changed. So we take FOXP1 for exam to find the possible shortest path from LCN2 to transcription factor illustrating how LCN2 affect mRNA expression profile alternation. In total, 28 shortest paths between LCN2 and FOXP1 were found. We noticed that ELAVL1 (ELAV like RNA binding protein 1, also called HuR) is most frequent protein (17/28) in the 28 possible paths to reach FOXP1. Overexpression of ELAVL1 is also found in ESCC, which is associated with positive

lymph node metastasis, deep tumor invasion, high tumor stage, and poor survival[35]. According to their subcellular localizations, most of these paths follow a pathway starting from the extracellular space to the cytoplasm to the nucleus. Moreover, many DEGs are able transloated from cytoplasm to nucleus. With a number of proteins are capable of translocation into nucleus, it can be argued that the overexpression of *LCN2* should greatly impact on the ESCC gene expression profile.

The total DEG PPI sub-network, when annotated by GO also in the format of a network, show that the PPI sub-network disturbed by the overexpresion of *LCN2* involves various biological entities, closely related to the known functions of *LCN2*. Of interest, this func-
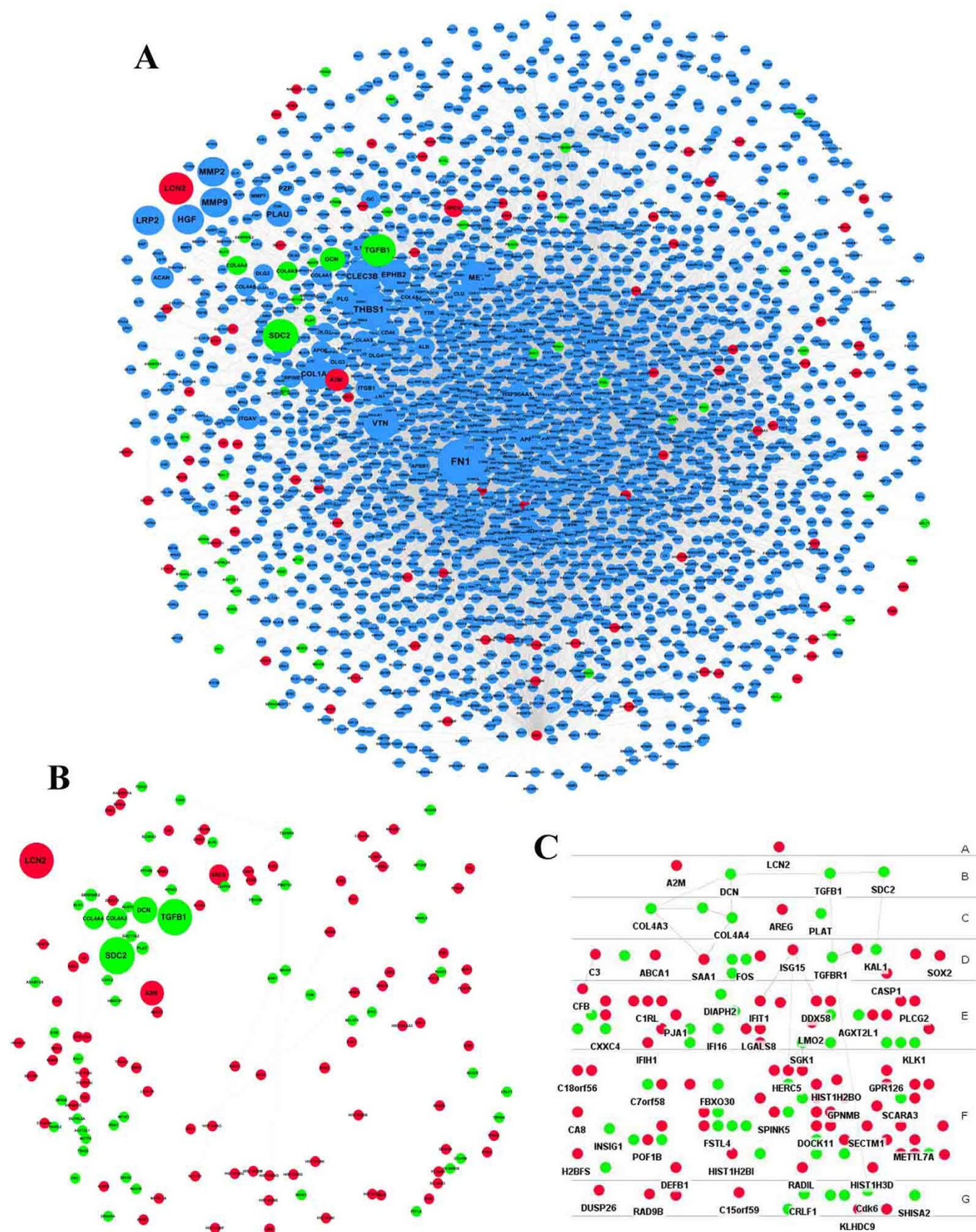
Figure 5 | Priorization analyses of DEGs in the total DEG PPI sub-network. (A) Random Walk with Restart algorithm was used to score all proteins in the PPI network for their network proximity to the seed node of LCN2. The node size in the PPI sub-network is designed in a gradient according to their scores. (B) The DEGs were extracted from (A) to better show their size. (C) The DEGs were re-arranged according to their closeness to LCN2 protein. The more negative the log10-transformed score, the further the node from LCN2. DEGs were classified into seven layers (from A to G, the Y axis) according to their range of scores as described in the Result section.

tional annotation map revealed many immunity-related GO terms, such as "regulation of immune respond", "activation of immune respond", "innate immune respond" and "deference respond", suggesting a role for LCN2 in the immune response. Direct evidence for an involvement of LCN2 in the immune response has been reported. Secreted LCN2 is involved in the the innate immune response to limit bacterial growth by sequestering the iron-laden siderophore[36]. That no iron metabolism related GO terms were found in this analysis, could be due to the possibility that there are no siderophores secreted by bacteria in the cell culture media for LCN2 to transport iron. Flo et al. reported that Lcn2$^{-/-}$ mice exhibit apparently normal iron metabolism. However, Lcn2$^{-/-}$ mice fail to mount efficient innate immune responses against bacterial infection[36]. Though we do not find significant GO terms associated with "cancer" or "tumor", the functional annotation map contained two large group terms of signaling pathways and the cell cycle, which is potentially related to the initiation or development of carcinoma. We assume LCN2 is not a proto-oncogene, but its biological influence in ESCC is multi-faceted, since so many signaling and cell cycle regulatory pathways are involved following its overexpression.

How to choose the DEGs for the subsequent functional experiments is still a huge challenge for the researchers after microarray analysis is completed. The RWR algorithm was applied to prioritize DEGs by ranking their closeness to LCN2. Many cancer-related genes were found closest to LCN2. For example, interaction of A2M (alpha2-macroglobulin) with low-density lipoprotein receptor-related protein-1 (LRP1) is associated with an inhibition of tumor cell proliferation, migration, invasion, spheroid formation, and anchorage-independent growth through inhibition of beta-catenin signaling in astrocytoma cells[37]. DCN (also called decorin) is known to interfere with cellular events of tumorigenesis mainly by blocking various receptor tyrosine kinases (RTK) such as the EGFR, Met, IGF-IR, PDGFR and VEGFR2. Genetic ablation of DCN leads to enhanced liver tumor incidence by providing an environment devoid of this potent pan-RTK inhibitor[38]. It has been suggested that frequent overexpression of TGFB1 promotes the progression of esophageal precancerous lesions via the proliferation of epithelial cells and angiogenesis, through the upregulation of vascular endothelial growth factor (VEGF) expression[25]. These results prioritize other DEGs, for examination of a relationship with LCN2, and provide important clues for experimental evaluation of the DEGs.

Conclusions

In summary, the analyses based on PPI network have greatly expand our understanding of the mRNA expression profile of LCN2 overexpression, as well as the potential biological roles of LCN2. Our study also provides a work flow to analyze expression data generated from high-throughput experiments.

Methods

The differentially expressed genes. LCN2 was overexpressed by transfection of the pcDNA3.0 plasmid, encoding LCN2, in the EC109 ESCC cell line. A control cell line was generated by transfection with an empty plasmid. The stably transfected cell clones were selected by Medium 199 (Invitrogen, USA) containing G418 (400 μg/ml) (Invitrogen, USA). Overexpression of LCN2 protein was confirmed by western blot analysis. The total RNA of LCN2 overexpressing cell and its control were extracted using TRIzol (Invitrogen, USA), respectively. Total RNA was amplified and labeled using the Agilent Quick Amp labeling kit by Cy3 or Cy5 and dye swapping. The labeled RNA was hybridized with Agilent whole Human genome oligo microarray (Agilent Technologies, USA) according its manual. After hybridization and washing, the processed slides were scanned with an Agilent DNA microarray scanner (part number G2505B) using settings recommended by Agilent Technologies. The raw data was treated by LOWESS (locally weighted scatterplot smoothing) normalization and log transformation. The expression data is in the GEO database (http://www.ncbi.nlm.nih.gov/geo/) under accession number of GSE57630. The differentially expressed genes (DEGs) were defined using a 2-fold threshold.

PPI sub-network construction. The newest versions of human protein-protein interaction datasets were available from both HPRD (http://www.hprd.org/) (Release 9) and BioGRID (http://thebiogrid.org/) (Release 3.2.107). These interactions were

derived from literatures of both low through-put and high through put experimentally validation. These two datasets have been widely applied in disease researches combined with human PPI network[39,40]. BioGRID also contains interactions from other species. In this study, the union interactions of Homo sapiens species from these two datasets were integrated manually, with each pair of interacting proteins in two lists of an Excel file. The redundancy from these two datasets was removed by the autofilter of Excel. The curated PPI data containing 18595 unique proteins and 174552 interactions were used as the parental PPI network. Cytoscape software was applied for visualization and analysis of PPI networks, which provides various plugins for different analyses[41]. PPI networks are illustrated as graphs in Cytoscape with the nodes representing the proteins and the edges representing their interactions. The different node attribution files and visual style files were imported into Cytoscape for better illustration in the context of biological networks.

We constructed five PPI sub-networks by mapping the DEGs to the HPRD&BioGRID parent PPI network by the following steps. First, the HPRD&BioGRID parent PPI network was imported in to Cytoscape. The DEGs (gene symbols) were listed in a text file (downregulated DEGs, upregulated DEGs and total DEGs, respectively) and mapped to the parental PPI network by the menu of "Select → Nodes → From ID List File". To confine the interactions only to those close to the DEGs and gain maximal significance, only first level interactions between DEGs and their neighbor were detected. We used Cytoscape menus of "Select → Nodes → First Neighbors of Selected Nodes" and "New → Network → From Selected Nodes, All Edges" to extracted the sub-network. Second, LCN2 was used as the query node and extracted interactions for the axis of LCN2 → neighbor proteins → DEGs → neighbor proteins by twice repeating the "First Neighbors of Selected Nodes", constructing the LCN2-central PPI sub-network. Third, a sub-network was generated by "New → Network → From Selected Nodes, All Edges" after total DEGs were mapped to the parental PPI network to detect the internal interactions between DEGs. Duplicated edges, single nodes and self-interactions of these sub-networks were regarded as redundant data and removed to avoid miscalculations of topological parameters of the PPI sub-network.

Network topological parameter analyses. The topological parameters of networks were analyzed by NetworkAnalyzer. By computing a comprehensive set of topological parameters, such as network diameter, density, centralization, heterogeneity, and clustering coefficient, neighborhood connectivity, average clustering coefficients and the distribution of node degrees, NetworkAnalyzer provides insights into the organization and structure of complex networks[42]. The degree of a node was the number of its directly connecting neighbours in the network. In this study, the power law of distribution of node degrees, one of most important network topological characteristics, was analyzed as we performed previously[43]. Briefly, the edges in all networks were treated as undirected. Distribution of node degree $P(k)$ is defined as the number of nodes with a degree k for $k = 0, 1, 2, \ldots$. The pattern of their dependencies can be visualized by fitting a line on the node degree distribution data. NetworkAnalyzer calculates the positive coordinate value for fitting the line where the power law curve of the form $y = \beta x^a$. R^2 value is a statistical measure of the linearity of the curve fit and used to quantify the fit to the power line. When the fit is good, the R^2 value is very close to 1. Moreover, other network parameters reflecting network properties were also analyzed and displayed.

Subcellular layers of the PPI sub-network. The subcellular localization information of each protein in the total DEG PPI sub-network was extracted by a custom R program from the newest Gene Ontology annotation file of Homo sapiens GO Annotations (released on 4/15/2014) at http://www.geneontology.org/GO.downloads.annotations.shtml. If one of the proteins was annotated with multiple localizations, especially for the proteins localizing in the nucleus (e.g. cytoplasm and nucleus), these localizations were integrated (cytoplasm/nucleus). The subcellular localization information was imported into Cytoscape as a node attribute. Cerebral software (http://www.pathogenomics.ca/cerebral/) was applied to re-distribute the nodes according to subcellular localization without changing their interactions, which provides a pathway-like diagram[44]. The igraph R program was applied to find the shortest path between LCN2 and FOXP1 (forkhead box P1) in the total DEG PPI sub-network. The shortest path algorithm is able to find the shortest connection between two nodes in the graph[45]. The protein members of these paths were also displayed according to their subcellular localization. These shortest paths were prioritized according to the normalized intensity of genes in their order with the signaling cascade.

Functional annotation map generation. We integrated Gene Ontology (GO) annotation into the total DEG PPI sub-networks by mining for enriched GO "Biological Process" terms of proteins using the ClueGO plugin, which allows the decoding and visualization of functionally grouped GO terms in the form of networks. ClueGO is a user friendly plugin to analyze interrelations of terms and functional groups in biological networks[46]. Only GO terms with a P-value < 0.001 were considered significant. A kappa score was calculated reflecting the relationships between the terms based on the similarity of their associated genes, which was set to 0.3 as the threshold in this study.

Random walk with restart to prioritize DEGs. A random walk on a graph is defined as an iterative walker's transition from a specific node to a random neighbor starting at a given source node (e.g. "protein A"). In this study, the algorithm of Random Walk

with Restart (RWR) we applied in which allow the restart of the walk in every time step at node "protein A" with probability r. The equation for the random walk with restart is defined as:

$$p^{t+1} = (1-r)\,Wp^{t} + rp^{0}$$

where r is the restart probability, W is the column-normalized adjacency matrix of the network graph, and p^{t} is a vector of size equal to the number of nodes in the graph where the i-th element holds the probability of being at node i at time step t. The initial probability vector p^{0} was constructed such that equal probabilities were assigned to the nodes representing members of the disease, with the sum of the probabilities equal to 1. In this study, RWR was carried out by a customized R program in the total DEG PPI sub-network with LCN2 protein set as the seed node. The probabilities of DEGs were regarded as node attributes and displayed by Cytoscape.

1. Flower, D. R. The lipocalin protein family: structure and function. *Biochem J* **318**, 1–14 (1996).
2. Yan, L., Borregaard, N., Kjeldsen, L. & Moses, M. A. The high molecular weight urinary matrix metalloproteinase (MMP) activity is a complex of gelatinase B/MMP-9 and neutrophil gelatinase-associated lipocalin (NGAL). Modulation of MMP-9 activity by NGAL. *J Biol Chem* **276**, 37258–65 (2001).
3. Yang, J. *et al.* An iron delivery pathway mediated by a lipocalin. *Mol Cell* **10**, 1045–56 (2002).
4. Paragas, N. *et al.* NGAL-Siderocalin in kidney disease. *Biochim Biophys Acta* **1823**, 1451–8 (2012).
5. Bolignano, D1. *et al.* Neutrophil gelatinase-associated lipocalin (NGAL) in human neoplasias: a new protein enters the scene. *Cancer Lett* **288**, 10–6 (2010).
6. Chakraborty, S., Kaur, S., Guha, S. & Batra, S. K. The multifaceted roles of neutrophil gelatinase associated lipocalin (NGAL) in inflammation and cancer. *Biochim Biophys Acta* **1826**, 129–69 (2012).
7. Zhang, H. *et al.* Up regulation of neutrophil gelatinase associated lipocalin in oesophageal squamous cell carcinoma: significant correlation with cell differentiation and tumour invasion. *J Clin Pathol* **60**, 555–561 (2007).
8. Du, Z. P. *et al.* Neutrophil gelatinase-associated lipocalin and its receptor: independent prognostic factors of oesophageal squamous cell carcinoma. *J Clin Pathol* **64**, 69–74 (2011).
9. Wu, J. *et al.* Integrated network analysis platform for protein-protein interactions. *Nat Methods* **6**, 75–7 (2009).
10. Li, M., Wu, X., Wang, J. & Pan, Y. Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. *BMC Bioinformatics* **13**, 109 (2012).
11. Bapat, S. A. *et al.* Gene expression: protein interaction systems network modeling identifies transformation-associated molecules and pathways in ovarian cancer. *Cancer Res* **70**, 4809–19 (2010).
12. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol Syst Biol* **3**, 88 (2007).
13. Li, Y. & Li, J. Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. *BMC Genomics* **13 Suppl 7**, S27 (2012).
14. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–3 (2002).
15. Zhu, X., Gerstein, M. & Snyder, M. Getting connected: analysis and principles of biological networks. *Genes Dev* **21**, 1010–24 (2007).
16. Barabási, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–13 (2004).
17. Goetz, D. H. *et al.* Ligand preference inferred from the structure of neutrophil gelatinase associated lipocalin. *Biochemistry* **39**, 1935–41 (2000).
18. Hvidberg, V. *et al.* The endocytic receptor megalin binds the iron transporting neutrophil-gelatinase-associated lipocalin with high affinity and mediates its cellular uptake. *FEBS Lett* **579**, 773–7 (2005).
19. Fang, W. K. *et al.* A novel alternative spliced variant of neutrophil gelatinase-associated lipocalin receptor in oesophageal carcinoma cells. *Biochem J* **403**, 297–303 (2007).
20. Ferlay, J. *et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer* **127**, 2893–2917 (2010).
21. Yang, L. *et al.* Estimates of cancer incidence in China for 2000 and projections for 2005. *Cancer Epidemiology Biomarkers and Prevention* **14**, 243–250 (2005).
22. Zhu, W., Yang, L. & Du, Z. Layered functional network analysis of gene expression in human heart failure. *PLoS One* **4**, e6288 (2009).
23. Chen, J., Aronow, B. J. & Jegga, A. G. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* **10**, 73 (2009).
24. Wang, F. *et al.* HDAC inhibitor trichostatin A suppresses esophageal squamous cell carcinoma metastasis through HADC2 reduced MMP-2/9. *Clin Invest Med* **36**, E87–94 (2013).
25. Xu, Z. *et al.* TGFβ1 and HGF protein secretion by esophageal squamous epithelial cells and stromal fibroblasts in oesophageal carcinogenesis. *Oncol Lett* **6**, 401–406 (2013).
26. Saito, T. *et al.* Factors influencing the acute phase protein levels in patients with esophageal cancer. *Jpn J Surg* **21**, 402–11 (1991).
27. Wu, I. C. *et al.* Plasma decorin predicts the presence of esophageal squamous cell carcinoma. *Int J Cancer* **127**, 2138–46 (2010).
28. Huang, X. *et al.* Prognostic significance of altered expression of SDC2 and CYR61 in esophageal squamous cell carcinoma. *Oncol Rep* **21**, 1123–9 (2009).
29. Friedman, A. & Perrimon, N. Genetic screening for signal transduction in the era of network biology. *Cell* **128**, 225–231 (2007).
30. Stebbins, J. L. *et al.* Identification of a new JNK inhibitor targeting the JNK-JIP interaction site. *Proc Natl Acad Sci U S A* **105**, 16809–13 (2008).
31. Cui, J., Zhang, M., Zhang, Y. Q. & Xu, Z. H. JNK pathway: diseases and therapeutic potential. *Acta Pharmacol Sin* **28**, 601–8 (2007).
32. Koon, H. B., Ippolito, G. C., Banham, A. H. & Tucker, P. W. FOXP1: a potential therapeutic target in cancer. *Expert Opin Ther Targets* **11**, 955–65 (2007).
33. Katoh, M. *et al.* Cancer genetics and genomics of human FOX family genes. *Cancer Lett* **328**, 198–206 (2013).
34. Tang, B. *et al.* Forkhead box protein p1 is a transcriptional repressor of immune signaling in the CNS: implications for transcriptional dysregulation in Huntington disease. *Hum Mol Genet* **21**, 3097–111 (2012).
35. Zhang, C. *et al.* Cytoplasmic expression of the ELAV-like protein HuR as a potential prognostic marker in esophageal squamous cell carcinoma. *Tumour Biol* **35**, 73–80 (2014).
36. Flo, T. H. *et al.* Lipocalin 2 mediates an innate immune response to bacterial infection by sequestering iron. *Nature* **432**, 917–921 (2004).
37. Lindner, I. *et al.* Alpha2-macroglobulin inhibits the malignant properties of astrocytoma cells by impeding beta-catenin signaling. *Cancer Res* **70**, 277–87 (2010).
38. Horváth, Z. *et al.* Decorin deficiency promotes hepatic carcinogenesis. *Matrix Biol* **35**, 194–205 (2014).
39. Goel, R., Muthusamy, B., Pandey, A. & Prasad, T. S. Human protein reference database and human proteinpedia as discovery resources for molecular biotechnology. *Mol Biotechnol* **48**, 87–95 (2011).
40. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2013 update. *Nucleic Acids Res* **41(Database issue)**, D816–23 (2013).
41. Smoot, M. E. *et al.* Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–2 (2011).
42. Assenov, Y. *et al.* Computing topological parameters of biological networks. *Bioinformatics* **24**, 282–4 (2008).
43. Wu, B. *et al.* Dissection of miRNA-miRNA Interaction in Esophageal Squamous Cell Carcinoma. *PLoS One* **8**, e73191 (2013).
44. Barsky, A., Gardy, J. L., Hancock, R. E. & Munzner, T. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics* **23**, 1040–2 (2007).
45. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Systems* **5**, 1695 (2006).
46. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–3 (2009).

Acknowledgments

Author contributions

L.X. and E.L. participated in study design and coordination, material support for obtained funding, and supervised study. B.W. and C.L. performed most of the analyses and drafted the manuscript. Z.D., Q.Y., J.W., L.F., P.Z., S.L. carried out the R programming and helped write the manuscript. All authors reviewed and approved the final manuscript.

Additional information