# Functional analysis of the nasopharyngeal carcinoma primary tumor-associated gene interaction network

FENGWEI AN[1*], ZHIQIANG ZHANG[2*] and MING XIA[3]

[1]Department of Otorhinolaryngology, Jinan Military General Hospital, Jinan, Shandong 250031;
[2]Department of Gastroenterology and Hepatology, People's Hospital of Huangdao, Qingdao, Shandong 266400;
[3]Department of Otorhinolaryngology, The Second Hospital of Shandong University, Jinan, Shandong 250031, P.R. China

Received October 16, 2014; Accepted June 22, 2015

DOI: 10.3892/mmr.2015.4090

**Abstract.** The aim of the present study was to investigate the molecular mechanism of nasopharyngeal carcinoma (NPC) primary tumor development through the identification of key genes using bioinformatics approaches. Using the GSE53819 microarray dataset, acquired from the Gene Expression Omnibus database, differentially expressed genes (DEGs) were screened out between NPC primary tumor and control samples, followed by hierarchical clustering analysis. The Search Tool for the Retrieval of Interacting Genes database was utilized to build a protein-protein interaction network to identify key node proteins. In total, 1,067 DEGs, including 326 upregulated genes and 741 downregulated genes, were identified between the NPC and control samples. The results of the hierarchical clustering analysis demonstrated that 95% of the DEGs were sample-specific. Furthermore, PDZ binding kinase (PBK), centromere protein F (CENPF), actin-binding protein anillin (ANLN), exonuclease 1 (EXO1) and chromosome 15 open reading frame 42 (C15ORF42) were included in the obtained network module, which was closely associated with the cell cycle and nucleic acid metabolic process GO functions. The results of the present study revealed that EXO1, CENPF, ANLN, PBK and C15ORF42 may be involved in the mechanism of NPC via modulating the cell cycle and nucleic acid metabolic processes, and may serve as molecular biomarkers for the diagnosis of this disease.

*Correspondence to:* Dr Ming Xia, Department of Otorhinolaryngology, The Second Hospital of Shandong University, 247 Beiyuan Avenue, Jinan, Shandong 250031, P.R. China
E-mail: mingxiamxer@163.com

*Contributed equally

*Key words:* nasopharyngeal carcinoma, protein-protein interaction network, exonuclease 1, centromere protein F, gene interaction network

## Introduction

The primary tumor or nasopharyngeal carcinoma (NPC) is a complicated malignant disease, originating from the epithelial cells located in the nasopharynx. There is markedly higher incidence of NPC in East Asia and Africa, compared with other regions of the world (1). The disease is attributed to multiple causative factors. One of the key risk factors identified is the Epstein-Barr (EB) viral infection (2,3). In addition, environmental effects and hereditary susceptibility contribute to the disease (4). The poor outcome of NPC treatment is attributed to the deficiency of effective therapeutic approaches and medicines, the complex structure of the nasopharynx, nonspecific clinical features, the difficulty of early diagnosis and variations in tumor histological types and differentiation (5,6). Therefore, there is an urgent requirement to identify specific molecular biomarkers for the early diagnosis of NPC.

It has been previously reported in Central and Southern China, that the miRNA-146a gene polymorphism is associated with the incidence of NPC (7). Additionally, EB virus-encoded microRNA has been reported to have an active role in NPC via modulating E-cadherin (8). It has been established that biological activities are performed by numerous interactions among proteins, DNA, RNA and other small molecules (9). Biological functions are achieved by a complex interaction network constructed by several functional units (10). Therefore, bioinformatics approaches have been widely used to investigate the associations among biological molecules, thus elucidating the complex mechanisms of disease (11). In addition, increasing studies have revealed that the roles of node proteins in the biological network topology are closely associated with their importance in cellular function, and networks with distinct topological features exhibit varying degrees of robustness in response to external environmental effects and internal conflicts (12,13). Consequently, the aims of topology-based investigations of biological networks are to investigate the association of critical nodes in the network, thus assisting in the understanding of the interactive topology and complex functions in cells. This provides valuable information for the diagnosis and treatment of disease, and designing novel drugs (14).

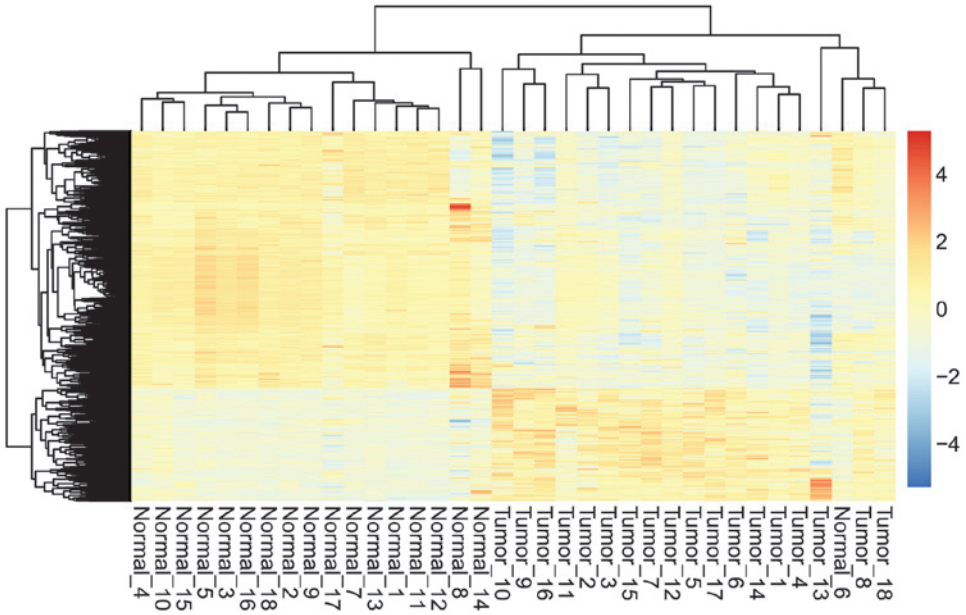


Figure 1. Heatmap from hierarchical clustering analysis. Changes in color between blue and orange indicate the progression of expression values of the differentially expressed genes between downregulation and upregulation, respectively. X-axis, sample name; Y-axis, fold change of the expression values of differentially expressed genes.

The present study aimed to investigate the molecular mechanism underlying NPC, by screening for the differentially expressed genes (DEGs) between NPC primary tumor and control samples, followed by hierarchical clustering analysis. The subsequent construction of a protein-protein interaction (PPI) network aimed to select hub proteins and perform network module analysis. The present study contributed to an enhanced understanding of the molecular mechanism of NPC and provided a basis for treating the disease.

## Materials and methods

*Microarray data preprocessing and DEG screening*. The GSE53819 microarray dataset was downloaded from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/), which is the largest open database of gene expression data (15). The data set used in the present study consisted of 18 samples of NPC primary tumor tissue and 18 control samples of normal nasopharyngeal tissue, based on the GPL6480 Agilent-014850 Whole Human Genome Microarray 4x44 K G4112F platform (Agilent Technologies, Inc., Santa Clara, CA, USA).

According to the platform, all probe numbers in the microarray data were mapped to their corresponding gene names. Regarding the genes corresponding to several probes, the average expression values of these probes were calculated to determine the expression value of the gene. Subsequently, the skewed distribution of data was converted into a normal distribution using a log 2 transformation, followed by normalization using the Median method (16). The Linear Models for Microarray Analysis package (http://www.bioconductor.org/packages/release/bioc/html/limma.html) (17) in R language was used to screen for the DEGs between the NPC and control tissue samples. Multiple testing correction (18) was also performed using the Benjamini-Hochberg method (19). |Log fold change|>1 and false discovery rate <0.05 were set as the strict cutoffs for DEG identification.

*Hierarchical clustering analysis*. Two-way hierarchical clustering analysis was performed for the identified DEGs using the pheatmap package in R language (http://cran.fhcrc.org/web/packages/pheatmap/index.html) (20). The clustering analysis grouped together genes with similarities in expression patterns, evaluating whether these DEGs were sample-specific. The clustering result of the DEGs enabled assessment of the sample type. The result was displayed as a heatmap.

*PPI network construction and hub protein analysis*. It has been established that the majority of biological networks are scale-free networks, in which only a minority of nodes possess a large number of links, while the majority of nodes have few links (21). Nodes which are connected to most of the proteins are defined as hub proteins and are the key in the network. To identify the hub proteins in the present study, the Search Tool for the Retrieval of Interacting Genes (22) online database (http://string-db.org/) was used to construct a PPI network using the proteins encoded by the DEGs. The path lengths of the nodes in the network were calculated to determine that the constructed network was scale-free. Subsequently, the degrees of the nodes corresponding to the links of the node protein were calculated, in order to screen for the hub proteins with the highest degrees.

*Network modules analysis*. Single proteins usually function via interactions with other proteins, rather than acting alone (23). Given that proteins in the same module are likely to perform similar functions, network modules with a degree ≥2 and K-core ≥2 were obtained using the Mcode plugin (24) from Cytoscape (www.cytoscape.org/) (25), which is software for network visualization and analysis. Gene ontology (GO) (26)

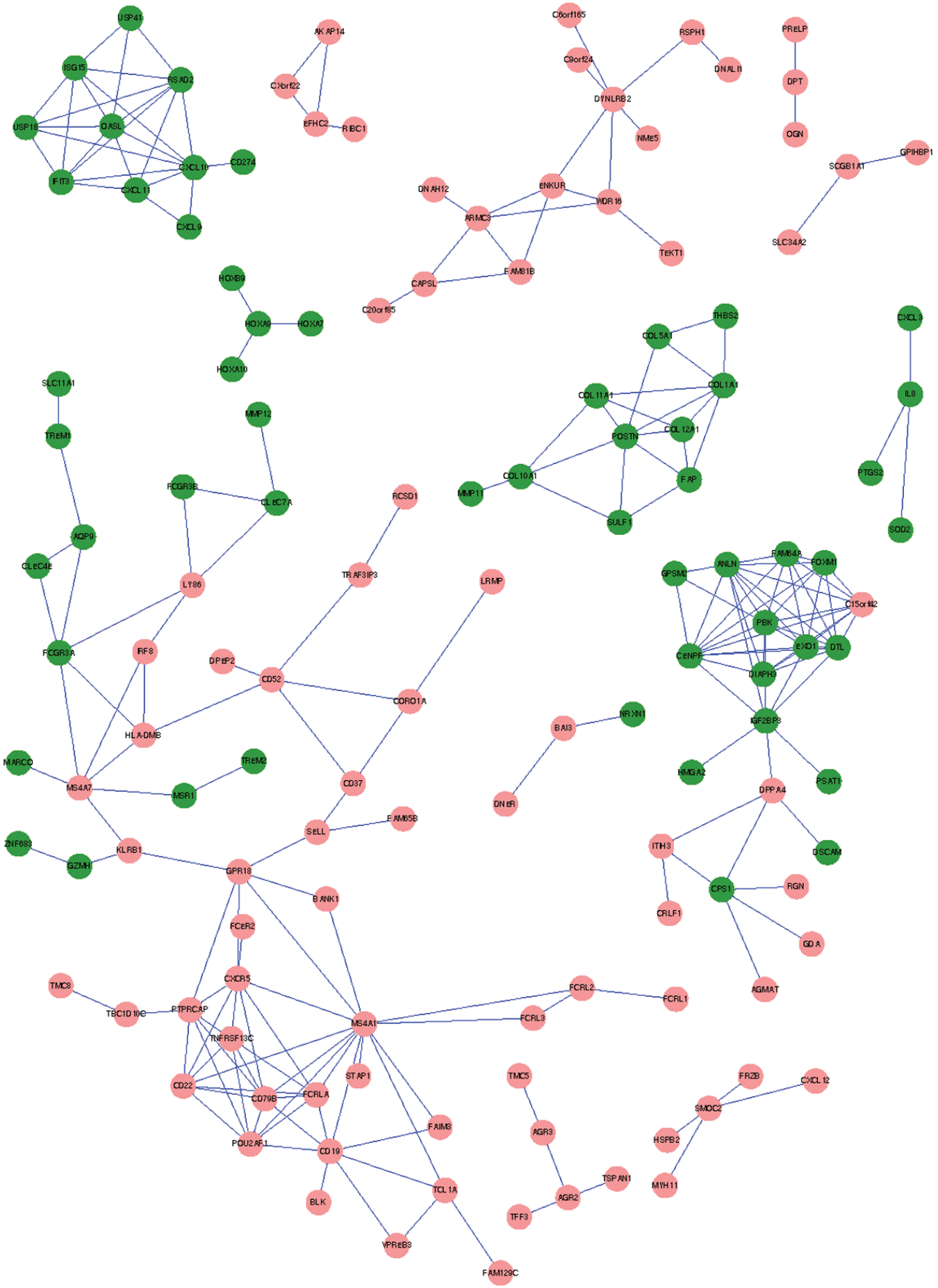


Figure 2. Protein-protein interaction network. Green nodes represent downregulated DEGs; pink nodes represent upregulated DEGs. Blue lines indicate the interaction between two proteins. DEGs, differentially expressed genes.

Table I. Top 10 node genes sorted in descending order of degree.

| Gene | Path length | Degree |
|---|---|---|
| MS4A1 | 3.16 | 13 |
| PBK | 2.05 | 10 |
| CENPF | 2.05 | 10 |
| ANLN | 2.05 | 10 |
| DTL | 2.10 | 9 |
| EXO1 | 2.10 | 9 |
| CD79B | 3.82 | 8 |
| C15ORF42 | 2.65 | 8 |
| IGF2BP3 | 1.80 | 8 |

MS4A1, membrane-spanning 4-domains, subfamily A, member 1; PBK, PBZ binding kinase; CENPF, centromere protein F; ANLN, anillin; DTL, denticleless protein homolog; EXO1, exonuclease 1; C15ORF42, chromosome 15 open reading frame 42; insulin-like growth factor 2 mRNA-binding protein 3.

Table II. GO functional enrichment analysis of network modules.

| GO ID | P-value | Adjusted P-value | Description |
|---|---|---|---|
| 7049 | $7.30\times10^{-7}$ | $2.38\times10^{-4}$ | Cell cycle |
| 51726 | $1.49\times10^{-6}$ | $2.43\times10^{-4}$ | Regulation of cell cycle |
| 90304 | $4.73\times10^{-4}$ | $6.75\times10^{-3}$ | Nucleic acid metabolic process |
| 6139 | $1.22\times10^{-3}$ | $1.17\times10^{-2}$ | Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process |
| 34641 | $2.46\times10^{-3}$ | $1.82\times10^{-2}$ | Cellular nitrogen compound metabolic process |
| 6807 | $3.16\times10^{-3}$ | $2.15\times10^{-2}$ | Nitrogen compound metabolic process |
| 44260 | $3.79\times10^{-3}$ | $2.32\times10^{-2}$ | Cellular macromolecule metabolic process |
| 16043 | $6.37\times10^{-3}$ | $2.91\times10^{-2}$ | Cellular component organization |
| 43170 | $7.92\times10^{-3}$ | $3.31\times10^{-2}$ | Macromolecule metabolic process |

GO, gene ontology.

functional enrichment analysis was performed for these obtained modules using the Bingo plugin (27) of Cytoscape. An adjusted P-value <0.05 was set as the threshold.

## Results

*DEG screening and hierarchical clustering analysis*. A total of 1,067 DEGs were screened between the NPC and control samples, including 326 upregulated genes and 741 down-regulated genes. The heatmap demonstrated that 95% of the DEGs were sample-specific (Fig. 1).

*Analysis of hub proteins in the PPI network*. In the PPI network (Fig. 2), 239 pairs of interactions among proteins were identified, in which 168 DEGs were involved. As shown in Fig. 3, the path lengths of the nodes in the network varied, ranging between one and nine, with the highest frequency at two, revealing that the network was scale-free. The degrees of the nodes are shown in Fig. 4. The top 10 node genes were sorted by degree in descending order (Table I). Among these 10 genes, membrane-spanning 4-domains, subfamily A, member 1 had the highest degree, with a degree of 13.

*Network module analysis*. As shown in Fig. 5, a network module including six genes exhibiting high degrees was obtained. The six genes involved were PDZ binding kinase (PBK), centromere protein F (CENPF), anillin (ANLN), denticleless protein homolog (DTL), exonuclease 1 (EXO1) and chromosome 15 open reading frame 42 (C15ORF42). The results of the GO functional analysis revealed that the network module was closely associated with the cell cycle and nucleic acid metabolic process (Table II), which were enriched in five of the genes exhibiting high degrees: EXO1, CENPF, ANLN, åPBK and C15ORF42. Of these five genes, PBK exhibited the highest degree (10).

## Discussion

NPC is an endemic malignant tumor in Southern China. The present study identified 1,067 DEGs between NPC and control

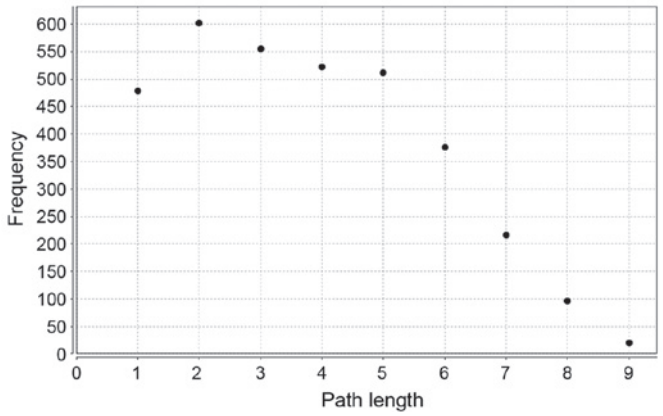


Figure 3. Analysis of the path lengths of the nodes in the protein-protein interaction network.

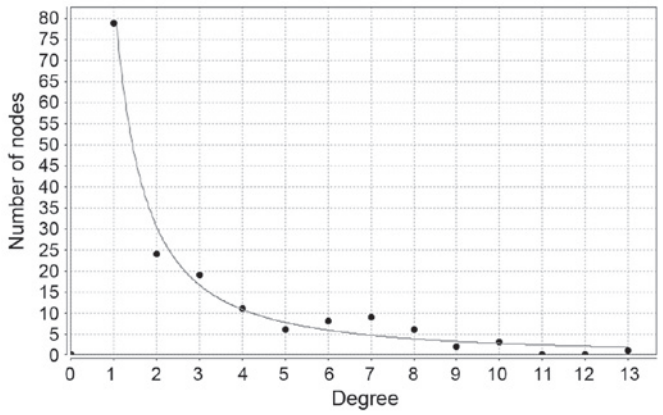


Figure 4. Analysis of the degrees of the nodes in the protein-protein interaction network.

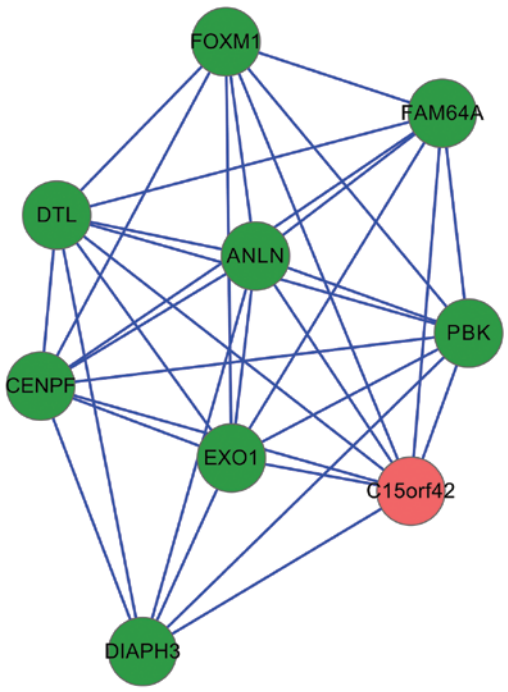


Figure 5. Network module obtained from the protein-protein interaction network. Green nodes represent downregulated DEGs; red nodes represent upregulated DEGs. Blue lines indicate the interaction between two proteins. DEGs, differentially expressed genes.

samples. These DEGs were revealed to be sample-specific by hierarchical clustering analysis. The constructed PPI network was confirmed to be scale-free and its hub proteins were analyzed. The results of network module analysis demonstrated that the obtained network module was associated with the cell cycle and nucleic acid metabolic process, in which EXO1, CENPF, ANLN, PBK and C15ORF42 were enriched DEGs with high degrees. These DEGs were downregulated, with the exception of C15ORF42. In agreement with the results of the present study, increasing studies have reported that cell cycle function is closely associated with the initiation and progression of NPC (28,29).

EXO1, one of the five DEGs identified, is an enzyme encoded by the EXO1 gene, which is involved in DNA repair and homologous recombination (30). It has been reported that genes associated with DNA repair are involved in the molecular mechanism underlying NPC (31). Similarly, the present study found that EXO1, the critical node protein in PPI network, was associated with the nucleic acid metabolic process, suggesting that EXO1 may be critically involved in the mechanism of NPC via regulating the nucleic acid metabolic process.

CENPF, a member of the centromere protein family, is involved in the formation of the nuclear matrix during the $G_2$ phase of the cell cycle and is involved in mitosis (32). Significant upregulation of CENPF has been previously reported in NPC cells, relative to normal nasopharyngeal cells, thus CENPF may be a molecular biomarker for the progression of NPC (33). Centromere protein H is also considered as a prognostic marker for the progression of NPC (34). In agreement with these reports, the present study demonstrated that CENPF was a critical node protein, exhibiting a high degree in the network module, indicating its importance in the mechanism of NPC.

Anillin, encoded by the ANLN gene, is a scaffolding actin-binding protein, which is involved in cytokinesis via connecting RhoA, actin and myosin (35). It has been demonstrated that anillin is upregulated in lung carcinogenesis, which may serve as a prognostic indicator for this disease (36). By contrast, the results of the present study suggested an undetermined role of anillin, which was downregulated in NPC. In addition, lymphokine-activated killer T-cell-originated protein kinase (TOPK), which is encoded by the PBK gene, is a serine/threonine kinase associated with mitogen-activated protein kinase kinase (37). TOPK has been previously identified to promote proliferation of breast tumor cells via p38 mitogen activated protein kinase activity (38). In addition, high expression levels of TOPK have been observed in melanoma cells (39). However, PBK was downregulated in the present study. These conflicting results may be a result of discrepancies between the experimental models and the samples.

The function of the C15ORF42 gene remains to be fully elucidated. It has been reported that another member of the same family, C16ORF13, is overexpressed in gastric cancer tissues, although the precise function of C15ORF42 remains elusive. In the present study, C15ORF42 was identified to be an upregulated node protein with a high degree in NPC, indicating a potentially critical role of C15ORF42 in the tumorigenesis of NPC for the first time, to the best of our knowledge.

In conclusion, the present study identified critical node proteins exhibiting close interactions with other proteins in the network module, including EXO1, CENPF, ANLN, PBK and

C15ORF42. These proteins may be involved in the tumorigenesis of NPC via modulating the cell cycle and nucleic acid metabolic process, and may be used as molecular biomarkers for the early diagnosis of NPC. The results of the present study assist in further understanding of the tumorigenesis of NPC, and provide potential targets for developing effective therapeutic treatment strategies for this disease.

## References


1. Chang ET and Adami HO: The enigmatic epidemiology of nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev 15: 1765-1777, 2006.
2. Wong AM, Kong KL, Tsang JW, Kwong DL and Guan XY: Profiling of Epstein, Barr virus, encoded microRNAs in nasopharyngeal carcinoma reveals potential biomarkers and oncomirs. Cancer 118: 698-710, 2012.
3. Dawson CW, Port RJ and Young LS: The role of the EBV-encoded latent membrane proteins LMP1 and LMP2 in the pathogenesis of nasopharyngeal carcinoma (NPC). Semin Cancer Biol 22: 144-153, 2012.
4. Fachiroh J, Sangrajrang S, Johansson M, Renard H, Gaborieau V, Chabrier A, Chindavijak S, Brennan P and McKay JD: Tobacco consumption and genetic susceptibility to nasopharyngeal carcinoma (NPC) in Thailand. Cancer Causes Control 23: 1995-2002, 2012.
5. Wei WI and Sham JS: Nasopharyngeal carcinoma. Lancet 365: 2041-2054, 2005.
6. Siddique MA, Sabur MA, Kundu SC, Mostafa MG, Khan JA, Ahmed S, Karim MA and Hanif MA: Difficulty in diagnosis of nasopharyngeal carcinoma. Mymensingh Med J 21: 158-161, 2012.
7. Huang GL, Chen ML, Li YZ, Lu Y, Pu XX, He YX, Tang SY, Chen H, Ding C and He Z: Association of miR-146a gene polymorphism with risk of nasopharyngeal carcinoma in the central-southern Chinese population. J Hum Genet 59: 141-144, 2014.
8. Hsu CY, Yi YH, Chang KP, Chang YS, Chen SJ and Chen HC: The Epstein-Barr virus-encoded microRNA MiR-BART9 promotes tumor metastasis by targeting E-cadherin in nasopharyngeal carcinoma. Plos Pathog 10: e1003974, 2014.
9. Segal E, Wang H and Koller D: Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics 19: i264-i271, 2003.
10. Ideker T, Ozier O, Schwikowski B and Siegel AF: Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18 (Suppl 1): S233-S240, 2002.
11. Moore JH, Asselbergs FW and Williams SM: Bioinformatics challenges for genome-wide association studies. Bioinformatics 26: 445-455, 2010.
12. Albert R, Jeong H and Barabási AL: Error and attack tolerance of complex networks. Nature 406: 378-382, 2000.
13. Jeong H, Mason SP, Barabási AL and Oltvai ZN: Lethality and centrality in protein networks. Nature 411: 41-42, 2001.
14. Csermely P, Agoston V and Pongor S: The efficiency of multi-target drugs: The network approach might help drug design. Trends Pharmacol Sci 26: 178-182, 2005.
15. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M and Edgar R: NCBI GEO: Mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 35 (Database Issue): D760-D765, 2007.
16. Fujita A, Sato JR, Rodrigues Lde O, Ferreira CE and Sogayar MC: Evaluating different methods of microarray data normalization. BMC Bioinformatics 7: 469, 2006.
17. Smyth GK: Limma: Linear models for microarray data. In: Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Gentleman R, Carey V, Dudoit S, Irizarry R and Huber W (eds). Springer, New York, pp397-420, 2005.
18. Noble WS: How does multiple testing correction work? Nat Biotechnol 27: 1135-1137, 2009.
19. He L and Sarkar SK: On improving some adaptive BH procedures controlling the FDR under dependence. Electron J Statist 7: 2683-2701, 2013.
20. Szekely GJ and Rizzo ML: Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. J Classif 22: 151-183, 2005.
21. Albert R: Scale-free networks in cell biology. J Cell Sci 118: 4947-4957, 2005.
22. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, *et al*: STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41 (Database Issue): D808-D815, 2013.
23. Przulj N, Wigle NA and Jurisica I: Functional topology in a network of protein interactions. Bioinformatics 20: 340-348, 2004.
24. Bader GD and Hogue CW: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2, 2003.
25. Smoot ME, Ono K, Ruscheinski J, Wang PL and Ideker T: Cytoscape 2.8: New features for data integration and network visualization. Bioinformatics 27: 431-432, 2011.
26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29, 2000.
27. Maere S, Heymans K and Kuiper M: BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21: 3448-3449, 2005.
28. Zhu XF, Liu ZC, Xie BF, Li ZM, Feng GK, Yang D and Zeng YX: EGFR tyrosine kinase inhibitor AG1478 inhibits cell proliferation and arrests cell cycle in nasopharyngeal carcinoma cells. Cancer Lett 169: 27-32, 2001.
29. Guo L, Tang M, Yang L, Xiao L, Bode AM, Li L, Dong Z and Cao Y: Epstein-Barr virus oncoprotein LMP1 mediates survivin upregulation by p53 contributing to G1/S cell cycle progression in nasopharyngeal carcinoma. Int J Mol Med 29: 574, 2012.
30. Desai A, Qing Y and Gerson SL: Exonuclease 1 is a critical mediator of survival during DNA double strand break repair in nonquiescent hematopoietic stem and progenitor cells. Stem Cells 32: 582-593, 2014.
31. Zhou X, Tian D, Wang S, Ruan Y, Qiu B, Zhang L and Lu B: Expressions of genes related to genome stability and DNA repair in nasopharyngeal carcinoma clustering families. Chin Ger J Clin Oncol 8: 713-718, 2009.
32. Liao H, Winkfein R, Mack G, Rattner JB and Yen TJ: CENP-F is a protein of the nuclear matrix that assembles onto kinetochores at late G2 and is rapidly degraded after mitosis. J Cell Biol 130: 507-518, 1995.
33. Cao JY, Liu L, Chen SP, Zhang X, Mi YJ, Liu ZG, Li MZ, Zhang H, Qian CN, Shao JY, *et al*: Prognostic significance and therapeutic implications of centromere protein F expression in human nasopharyngeal carcinoma. Mol Cancer 9: 237, 2010.
34. Liao WT, Song LB, Zhang HZ, Zhang X, Zhang L, Liu WL, Feng Y, Guo BH, Mai HQ, Cao SM, *et al*: Centromere protein H is a novel prognostic marker for nasopharyngeal carcinoma progression and overall patient survival. Clin Cancer Res 13: 508-514, 2007.
35. Piekny AJ and Glotzer M: Anillin is a scaffold protein that links RhoA, actin and myosin during cytokinesis. Curr Biol 18: 30-36, 2008.
36. Suzuki C, Daigo Y, Ishikawa N, Kato T, Hayama S, Ito T, Tsuchiya E and Nakamura Y: ANLN plays a critical role in human lung carcinogenesis through the activation of RHOA and by involvement in the phosphoinositide 3-kinase/AKT pathway. Cancer Res 65: 11314-11325, 2005.
37. Abe Y, Matsumoto S, Kito K and Ueda N: Cloning and expression of a novel MAPKK-like protein kinase, lymphokine-activated killer T-cell-originated protein kinase, specifically expressed in the testis and activated lymphoid cells. J Biol Chem 275: 21525-21531, 2000.
38. Ayllon V and O'connor R: PBK/TOPK promotes tumour cell proliferation through p38 MAPK activity and regulation of the DNA damage response. Oncogene 26: 3451-3461, 2007.
39. Zykova TA, Zhu F, Lu C, Higgins L, Tatsumi Y, Abe Y, Bode AM and Dong Z: Lymphokine-activated killer T-cell-originated protein kinase phosphorylation of histone H2AX prevents arsenite-induced apoptosis in RPMI7951 melanoma cells. Clin Cancer Res 12: 6884-6893, 2006.