

# Identification of Candidate Genes Related to Renal Carcinoma using Protein Interactions and Structures

Lei Chen, Yuhang Zhang, Jian Zhang, Tao Huang, Yang Shu, and Yu-Dong Cai

**Abstract**—Renal carcinoma is a type of cancer that starts in the kidneys. It has been reported that the incidence rate of this type of cancer is increasing, inducing that effective treatments are urgently needed. However, it is very difficult to design effective treatments based on current limited comprehension of this disease. Extracting all related genes of this disease may provide help for good understanding of its mechanism. This study strengthened a recent reported computational method to identify new candidate genes related to renal carcinoma. The new candidate genes were found by applying a shortest path algorithm in a weighted network, constructed by protein interactions, to search shortest paths connecting any two known genes related to renal carcinoma. The further candidate genes were selected by a randomization test and measuring the associations between candidate genes and known genes based on their interactions and structures information. Finally, some obtained new genes were discussed, indicating that they are closely related to renal carcinoma.

**Index Terms**—Renal carcinoma, disease gene, protein interaction, BLAST

## I. INTRODUCTION

Renal carcinoma (RC) is a common group of kidney cancer that originates from the epithelium of the renal parenchyma (renal cell carcinoma) or the renal pelvis (renal

This work was supported in part by the National Basic Research Program of China (2011CB510101, 2011CB510102), the National Natural Science Foundation of China (61202021, 31371335), the Innovation Program of Shanghai Municipal Education Commission (12YZ120, 12ZZ087), and the Shanghai Educational Development Foundation (12CG55).

Lei Chen is with the College of Life Science, Shanghai University, Shanghai 200444, China and the College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China (e-mail: chen\_lei1@163.com).

Yuhang Zhang is with the Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China (e-mail: zhangyh825@163.com).

Jian Zhang is with the Department of Ophthalmology, Shanghai First People's Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai 200080, China and Shanghai Key Laboratory of Ocular Fundus Diseases, Shanghai First People's Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai 200080, China (e-mail: natalieeilatan@126.com)

Tao Huang is with the Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China (e-mail: tohuangtao@126.com).

Yang Shu is with the the Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China (e-mail: shuyang1986@gmail.com).

Yu-Dong Cai is with the College of Life Science, Shanghai University, Shanghai 200444, China (corresponding author to provide phone: 86-21-66136132; fax: 86-21-66136109; e-mail: cai\_yud@126.com).

pelvis carcinoma). Of all renal malignancies subtypes, renal cell carcinoma (RCC) accounts for nearly 90% of the neoplasms arising from the kidney [1]. RC has been reported to be related with several environmental and genetic factors, such as various age standardized incidences [2], difference in incidence rates in both sexes (male higher than female) (refer to [www.cancerresearchuk.org](http://www.cancerresearchuk.org)), smoking [3], potential occupational carcinogens [4], excess body weight [5] and, most significant, genetic changes [6].

With the development of sequencing technology, a surprising number of genetic mutations have been revealed to affect the initiation and progression of RC. Such mutated genes can be classified into several groups and have different occurrence frequency in different kinds of RC. The first gene identified for RCC is von Hippel-Lindau (*VHL*) tumor-suppressor gene [7], which has been detected in most cases of sporadic RCC. What's more, chromatin remodeling genes: *PBRM1*, *BAP1*, and *SETD2* have been proved to modify the SWI/SNF complex and may further regulate tumorigenesis and metastasis in RC [8]. Another group of genes: *KDM5a*, *ARID1A*, *UTX* are identified as histone modifiers which may also contain specific mutations [9]. Genetic changes associated with mTOR pathway signaling such as mutations in *PIK3CA*, *PTEN*, and *mTOR* probably contribute to the progression of RC. There still remain a couple of genes (such as *FH*, *MET*, *BHD*, *FL*, *CN*, etc) containing more gene mutations identified in different types of RC in spite of their ambiguous functional mechanism [10]-[12].

Genetic changes underlying RC influence crucial functions of certain key passways in RC. Essentially, genetic changes of specific factors involved in such passways may induce certain dysfunction and further affect the normal cellular metabolism through protein-protein interaction which is regarded as the foundation of cell signaling and biological function. What's more, not only the protein interacting with the known oncogenic mutated proteins but certain homologous protein sharing similar sequence with these mutated proteins may act as key proteins in tumor initiation and progression as well.

It has been proved that interactive proteins or proteins with similar structures always share similar functions [13]-[15]. Thus, we employed the information of protein interactions and sequence similarity to find potential tumor associated genes participated in the initiation and progression of RC. The protein interactions were retrieved from STRING (Search Tool for the Retrieval of Interacting Genes/Proteins)

[16], a database of known and predicted protein interactions including direct (physical) and indirect (functional) associations, in which most known and predicted associations are scored and integrated. It provides not only easy and intuitive interfaces for searching and browsing the protein interaction data but also background algorithms to quantify statistic co-occurrence in database so as to calculate max link scores [17]. While BLAST (basic local alignment search tool) [13] is designed to find regions of similarity in long DNA, RNA or protein sequence. The proposed method using the aforementioned two types of protein information to find new genes of RC based on known RC-related genes retrieved from Uniprot [18], TSGene database [19] and NCI database [20]. Firstly, the protein interactions were used to construct a weighted network. Secondly, a shortest path algorithm was applied in this network to extract candidate genes. Finally, the candidate genes were filtered by a randomization test, protein interactions and protein sequence similarities. In the end of this study, we discussed some obtained new genes and found that they are related to RC.

## II. MATERIALS AND METHODS

### A. Genes Related to Renal Cancer

The known RC-related human genes were obtained from the following three databases: (1) Uniprot (<http://www.uniprot.org/>) [18]; (2) TSGene database (<http://bioinfo.mc.vanderbilt.edu/TSGene/search.cgi>) [19]; (3) NCI database (<https://gforge.nci.nih.gov>, released April, 2009) [20]. Specifically, 613 genes were retrieved from Uniprot by inputting "human renal cancer reviewed" as the keywords; 21 genes were chosen in the catalogue of renal cancer from TSGene database; 104 genes that are recorded as the renal cancer/ renal carcinoma related genes were collected from NCI database. After combining the aforementioned 738 genes, 708 human genes were finally obtained, which are listed in Online Supporting Information S1.

### B. Searching Method

To search new candidate genes related to RC, we referred to the method reported in a recent published study [21] and generalized it. According to the method in [21], we downloaded the file (protein.links.v9.1.txt.gz) containing great number of protein-protein interactions from STRING (<http://string.embl.de/>) [16] and extracted 1,640,707 protein-protein interactions of human, involving 18,600 ensembl IDs. These interactions are derived from genomic context, high-throughput experiments, (conserved) co-expression or previous knowledge (refer to <http://string.embl.de/>), thus they included direct and indirect associations of proteins. Furthermore, each of the 1,640,707 protein-protein interactions is labeled a score to measure the strength of the interaction. Its range is between 150 and 999. For each interaction between proteins  $p_1$  and  $p_2$ , its score was denoted by  $S(p_1, p_2)$ . All information of 1,640,707 protein-protein interactions was used to construct a weighted network by taking proteins occurring in 1,640,707 protein-protein interactions as nodes and connecting two nodes if the corresponding proteins can comprise a

protein-protein interaction. Furthermore, to reflect the strength of each interaction, each corresponding edge was assigned a weight defined by  $w(n_1, n_2)=1000- S(p_1, p_2)$ , where  $p_1$  and  $p_2$  were corresponding proteins of  $n_1$  and  $n_2$ , respectively.

After constructing the weighted network, the shortest path algorithm was applied in this network to search all shortest paths connecting any pair of 708 human genes. Based on these paths, we extracted all their inner nodes and the corresponding proteins were deemed to have special associations with RC and termed as candidate genes. In addition, we also counted the number of paths containing a certain candidate gene as an inner node and termed this value as betweenness.

### C. Screening Method

As described in Section II.A, some candidate genes can be identified. However, false positives were inevitable. The screening method is necessary to control them, which included two stages: (I) A randomization test used to filter universal genes; (II) selection of some candidate genes with core relationship with known RC-related genes.

It is known that some genes always have special associations with other genes, which induces that they are always selected by the method described in Section II.B even if we randomly selected some genes as RC-related genes. However, these genes have few associations with RC. In view of this, we randomly produced 500 gene sets whose sizes were same as that of the set consisting of RC-related genes. Then, for each gene set, the shortest path algorithm was used to find shortest paths connecting any pair of genes in the set and the betweenness of each candidate gene was counted based on these paths. Finally, a measure, named permutation FDR, was calculated for each candidate gene, which was defined as "the number of gene sets on which the betweenness was larger than its actual betweenness"/500. It is obvious that low permutation FDR indicates the candidate gene is a universal gene with small probability and is an actual gene related to RC with high likelihood. Thus, we set 0.05 as a threshold to further select candidate genes.

The second step of the screening method is to identify genes from the remaining genes, which have core associations with known RC-related genes, thereby being actual RC-related genes with high possibility. It has been reported in some studies that two proteins that can comprise an interaction always share similar functions [14], [15]. In view of this, the interacting genes of RC-related genes have higher likelihood to be RC-related genes than others. Therefore, for each remaining candidate gene  $g$ , we calculated the following maximum interaction score of  $g$ :

$$\text{score} - \text{interaction}(g) = \max \{S(g, g') : g' \text{ is a known RC-related gene}\} \quad (1)$$

On the other hand, proteins with similar structures always have similar functions [13]. Thus, we employed the basic local alignment search tool (BLAST) [13], which can search local similarities between the protein sequences. For each remaining candidate gene  $g$ , maximum alignment score of  $g$  was calculated as follows:

$$\text{score-sequence}(g) = \max \{S_b(g, g') : g' \text{ is a known RC-related gene}\} \quad (2)$$

where  $S_b(g, g')$  is the alignment score between  $g$  and  $g'$  obtained by BLAST. The remaining candidate genes were further selected by setting 900 as the threshold for maximum interaction score and setting 90 as the threshold for maximum alignment score.

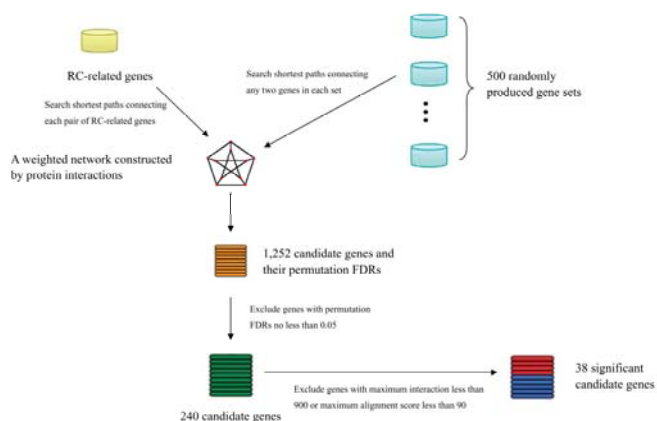


Fig. 1. The workflow of the procedures for identification of new candidate RC-related genes.

### III. RESULTS AND DISCUSSION

The procedures for the identification of new RC-related genes integrated the methods in Section II.B and II.C, their workflow is illustrated in **Fig. 1**.

#### A. Results of the Searching Method

According to the method mentioned in Section II.B, all shortest paths connecting any pair of known RC-related genes were searched in the constructed weighted network. Then, the nodes occurring in any of these paths as inner nodes were extracted and the corresponding proteins, which were not RC-related genes, were selected as candidate genes. The obtained 1,252 candidate genes are listed in Online Supporting Information S2, in which the betweenness of these genes are also provided.

#### B. Results of the Screening Method

The screening method included two steps as mentioned in Section II.C. Firstly, a randomization test was executed by calculating the permutation FDR for each candidate gene, which is listed in Online Supporting Information S2. After setting 0.05 as a threshold, we obtained 240 candidate genes with permutation FDRs smaller than 0.05, which are listed in Online Supporting Information S3. Then, the maximum interaction score and maximum alignment score for each of 240 candidate genes were calculated, which are also provided in Online Supporting Information S3. By setting 900 as the threshold for maximum interaction score and 90 as the threshold for maximum alignment score, 38 candidate genes (see the first 38 genes in Online Supporting Information S3) were finally obtained. These genes were deemed to be closely related to RC and termed as significant candidate genes for RC.

TABLE I  
SOME SIGNIFICANT CANDIDATE GENES OBTAINED BY OUR METHOD

Ensembl ID	Gene symbol	Betweenness	Permutation FDR	Maximum interaction score (Most related known gene)	Maximum alignment score (Most related known gene)
ENSP00000350896	EPHB4	681	0	905 (PIK3CA)	1074 (EPHB2)
ENSP00000367830	PRKCZ	1693	0.008	993 (AKT1)	884 (PRKC1)
ENSP00000379866	COL4A4	1361	0.002	999 (COL4A3)	357 (COL4A6)
ENSP00000331902	COL4A5	685	0.002	999 (COL4A6)	439 (COL4A1)
ENSP00000219070	MMP2	1363	0.028	995 (COL18A1)	190 (MMP7)
ENSP00000261799	PDGFRB	10634	0.032	998 (NCK1)	330 (FLT1)
ENSP00000365312	ABI1	748	0	925 (EGFR)	195 (ABI3)
ENSP00000315768	STAT2	681	0.008	998 (TYK2)	505 (STAT1)
ENSP00000287934	FZD1	2031	0.004	993 (WNT7B)	501 (FZD5)

#### C. Analysis of Significant Candidate Genes

By our method, we obtained 38 significant candidate genes that may involve in RC. The obtained genes have interactions and, simultaneously, share sequence homology with RC-related genes. These genes, despite lack of direct experimental evidence, may correlate with tumorigenesis potentially. Some of them, listed in **Table I**, were discussed as below.

Among these genes, specific kinases account for a large proportion. EPHB4 (see **Table I**, row 2), a member of the Eph family of receptor tyrosine kinases, has been proved to have both tumor-suppressing and tumor-promoting activities in breast cancer [22]. Together with its preferred ligand ephrin-B2 (EFNB2) which has been proved to be involved in RC, it regulates the Abl/Crk pathway and shows a tumor-suppressing phenotype [23]. Consistence with our calculation result, EPHB4/EPNB2 may constitute a complex which interacts with R-Ras and may further participate in the phosphorylation of AKT [24]. Such evidence above hints that EPHB4 may play a similar role in the initiation and progression of carcinoma, especially RC. Apart from EPHB4, another protein kinase, PRKCZ (see **Table I**, row 3) was also predicted to play a significant role in RC. It has been reported to be a calcium-and diacylglycerol-independent serine /threonine-protein kinase that functions in phosphatidylinositol 3-kinase (PI3K) pathway and

mitogen-activated protein (MAP) kinase cascade [25]. Further, PRKCZ is quite crucial in human prostate cancer which has been regarded as a specific marker [26].

As kidney tissue specific genes, we also identified COL4A4 and COL4A5 (see **Table I**, rows 4-5) as potential genes. These two genes encode the major structural component of glomerular basement membranes. COL4A4/COL4A5 may have indirect interaction with MMP2 which was also screened as a candidate gene (see **Table I**, row 6) [27]. MMP2 is an ubiquitous metalloproteinase that involved in diverse functions and has been proved to be relevant with tumor progression [28]. As the upstream of Ras and MMP2, candidate gene PDGFRB (see **Table I**, row 7) has been proved to be located on the cell membrane and participate in the MAPK signaling pathway which is also activated in tumor cells [29]. As a key transducer of signals from Ras to Rac, selected gene ABI1 (see **Table I**, row 8), may interact with phosphatidylinositol-3-kinase to further show tumorigenesis characteristics [30].

There are also quite a lot of crucial genes in specific pathways that involve in RCs. STAT2 (see **Table I**, row 9), as key component of JAK-STAT passway which has only been reported to be associated with inflammation and few tumors has been predicted to be a potential tumor-associated gene. Although there are few direct proofs that STAT2 participates in tumor initiation and progression, it has been reported to contribute to inflammation and, therefore, may perform as an assistive mutation that helps to remodel tumor microenvironment [22].

What's more, Wnt passway has been widely investigated and found to contribute to cell differentiation and tumor initiation [31]. Based on our algorithm, we screened out a specific receptor for Wnt proteins, FZD1, as a candidate gene (see **Table I**, row 10). FZD1 has been reported to activate disheveled proteins, inhibit GSK-3 kinase and, most significantly, activate Wnt target genes [32]-[34]. Furthermore, WNT2 is also selected through this algorithm, which is a proper ligand for members of the frizzled family of several transmembrane receptors. It has been reported to be associated with several kinds of tumors [35], [36]. Analysis above confirms that our predicted genes may actually play a specific role in RC.

All in all, the obtained candidate genes either have been proved to be associated with tumor or contribute to regulation of crucial pathways that involve in cell proliferation and adhesion, which may be potential oncogene or tumor suppressor gene. Although the function of several genes have been confirmed to contribute to RC, such as *VHL*, *PBRM1*, *BAP1*, *SETD2* and so on, the diverse nature of RC implies there may still be potential tumor associated genes remained to be found [37]. Such genes have to be identified to clearly demonstrate the heterogeneity of mutations in different subtypes of RC. In conclusion, based on confirmed RC-related genes, the proposed algorithm has been proved to be effective to identify candidate tumor associated genes in RC.

#### IV. CONCLUSIONS

This contribution strengthened an existing computational method and applied this method to identify new candidate genes related to renal carcinoma. Compared to the original method, the new method added an additional screening procedure using the information of protein interactions and structures. The analysis of the final obtained genes indicates that this method is effective for identification of new candidate genes for renal carcinoma. However, the new findings in this study should be validated by solid experiments. It is hopeful that this method may give new lights to study renal carcinoma and other diseases.

#### REFERENCES

- [1] B. Curti, *et al.*, *Renal cell carcinoma*, 2014.
- [2] E. E. Calle and R. Kaaks, "Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms," *Nature Reviews Cancer*, vol. 4, pp. 579-591, 2004.
- [3] W. A. Chiu, *et al.*, "Key scientific issues in the health risk assessment of trichloroethylene," *Environmental health perspectives*, pp. 1445-1449, 2006.
- [4] J. Ferlay, *et al.*, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," *Int J Cancer*, vol. 127, pp. 2893-917, Dec 15 2010.
- [5] A. Franceschini, *et al.*, "STRING v9. 1: protein-protein interaction networks, with increased coverage and integration," *Nucleic acids research*, vol. 41, pp. D808-D815, 2013.
- [6] D. Cohen and M. Zhou, "Molecular genetics of familial renal cell carcinoma syndromes," *Clinics in laboratory medicine*, vol. 25, pp. 259-277, 2005.
- [7] A. A. Hakimi, *et al.*, "Adverse outcomes in clear cell renal cell carcinoma with mutations of 3p21 epigenetic regulators BAP1 and SETD2: a report by MSKCC and the KIRC TCGA research network," *Clinical Cancer Research*, vol. 19, pp. 3259-3267, 2013.
- [8] J. D. Hunt, *et al.*, "Renal cell carcinoma in relation to cigarette smoking: Meta-analysis of 24 studies," *International journal of cancer*, vol. 114, pp. 101-108, 2005.
- [9] E. Jonasch, *et al.*, "State of the science: an update on renal cell carcinoma," *Molecular Cancer Research*, vol. 10, pp. 859-880, 2012.
- [10] W. G. Kaelin Jr, "Von hippel-lindau disease," *Annu. Rev. Pathol. Mech. Dis.*, vol. 2, pp. 145-173, 2007.
- [11] L. Gossage, *et al.*, "Clinical and pathological impact of VHL, PBRM1, BAP1, SETD2, KDM6A, and JARID1c in clear cell renal cell carcinoma," *Genes, Chromosomes and Cancer*, vol. 53, pp. 38-51, 2014.
- [12] P. Kapur, *et al.*, "Effects on survival of BAP1 and PBRM1 mutations in sporadic clear-cell renal-cell carcinoma: a retrospective analysis with independent validation," *The lancet oncology*, vol. 14, pp. 159-167, 2013.
- [13] S. F. Altschul, *et al.*, "Basic local alignment search tool," *J Mol Biol*, vol. 215, pp. 403-10, Oct 5 1990.
- [14] L. L. Hu, *et al.*, "Predicting Protein Phenotypes Based on Protein-Protein Interaction Network," *PLoS ONE*, vol. 6, p. e17668, 2011.
- [15] L. L. Hu, *et al.*, "Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties," *PLoS ONE*, vol. 6, p. e14556, 2011.
- [16] L. J. Jensen, *et al.*, "STRING 8-a global view on proteins and their functional interactions in 630 organisms," *Nucleic acids research*, vol. 37, pp. D412-416, 2009.
- [17] I. Varela, *et al.*, "Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma," *Nature*, vol. 469, pp. 539-542, 2011.
- [18] U. Consortium, "Update on activities at the Universal Protein Resource (UniProt) in 2013," *Nucleic acids research*, vol. 41, pp. D43-D47, 2013.
- [19] M. Zhao, *et al.*, "TSGene: a web resource for tumor suppressor genes," *Nucleic acids research*, vol. 41, pp. D970-D976, 2013.
- [20] S. McNeil, *et al.*, "National Cancer Institute," *Imaging*, 2013.
- [21] Y. Jiang, *et al.*, "Identifying Gastric Cancer Related Genes Using the Shortest Path Algorithm and Protein-Protein Interaction Network," *BioMed Research International*, vol. 2014, p. 371397, 2014.

- [22] N. K. Noren and E. B. Pasquale, "Paradoxes of the EphB4 receptor in cancer," *Cancer research*, vol. 67, pp. 3994-3997, 2007.
- [23] A. Bates, *et al.*, "Stromal matrix metalloproteinase 2 regulates collagen expression and promotes the outgrowth of experimental metastases," *J Pathol*, 2014.
- [24] C. G. A. R. Network, "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, vol. 499, pp. 43-49, 2013.
- [25] A. Chetcuti, *et al.*, "Expression profiling reveals MSX1 and EphB2 expression correlates with the invasion capacity of Wilms tumors," *Pediatric blood & cancer*, vol. 57, pp. 950-957, 2011.
- [26] S. Fransson, *et al.*, "Stage-dependent expression of PI3K/Akt- pathway genes in neuroblastoma," *Int J Oncol*, vol. 42, pp. 609-616, 2013.
- [27] X. Fu and Y. Feng, "QKI-5 suppresses cyclin D1 expression and proliferation of oral squamous cell carcinoma cells via MAPK signalling pathway," *International journal of oral and maxillofacial surgery*, 2014.
- [28] Y.-D. Gong, *et al.*, "A novel 3-arylethynyl-substituted pyrido [2, 3,-b] pyrazine derivatives and pharmacophore model as Wnt2/ $\beta$ -catenin pathway inhibitors in non-small-cell lung cancer cell lines," *Bioorganic & medicinal chemistry*, vol. 19, pp. 5639-5647, 2011.
- [29] Y. Li, *et al.*, "A Wnt/ $\beta$ -catenin negative feedback loop represses TLR-triggered inflammatory responses in alveolar epithelial cells," *Molecular immunology*, vol. 59, pp. 128-135, 2014.
- [30] K. Steinestel, *et al.*, "Expression of Abelson interactor 1 (Abl) correlates with inflammation, KRAS mutation and adenomatous change during colonic carcinogenesis," *PLoS ONE*, vol. 7, p. e40671, 2012.
- [31] I. Oinuma, *et al.*, "R-Ras controls axon specification upstream of glycogen synthase kinase-3 $\beta$  through integrin-linked kinase," *Journal of Biological Chemistry*, vol. 282, pp. 303-318, 2007.
- [32] J. Pollheimer, *et al.*, "Expression pattern of collagen XVIII and its cleavage product, the angiogenesis inhibitor endostatin, at the fetal-maternal interface," *Placenta*, vol. 25, pp. 770-779, 2004.
- [33] Y. W. Qiang, *et al.*, "Characterization of Wnt/ $\beta$ -catenin signalling in osteoclasts in multiple myeloma," *British journal of haematology*, vol. 148, pp. 726-738, 2010.
- [34] K. Saito-Diaz, *et al.*, "The way Wnt works: components and mechanism," *Growth Factors*, vol. 31, pp. 1-31, 2013.
- [35] Y. Sun, *et al.*, "Interplay between interferon-mediated innate immunity and porcine reproductive and respiratory syndrome virus," *Viruses*, vol. 4, pp. 424-446, 2012.
- [36] S. Yao, *et al.*, "Splice variant PRKC- $\zeta$ -PrC is a novel biomarker of human prostate cancer," *Br J Cancer*, vol. 107, pp. 388-399, 2012.
- [37] Y. Shibing, *et al.*, "AP2 suppresses osteoblast differentiation and mineralization through down-regulation of Frizzled-1," *Biochemical Journal*, vol. 465, pp. 395-404, 2014.