Literature Data Mining and Enrichment Analysis Reveal A Genetic Network of 423 Genes for Renal Cancer

Peng Zhou[1], Yuping Wang[2], Hongbao Cao[3], Lydia C Manor[4*]

[1] Department of BME, Tianjin University, Tianjin 300072, China; [2] Department of Biomedical Engineering, Tulane University, New Orleans, LA, USA; [3] Elsevier Inc., 5635 Fishers Ln, Rockville, MD 20852; [4] American Informatics Consultant LLC, Rockville, MD, 20852.

***Correspondence:** Lydia C Manor, Sr. Bioinformatics Scientist, American Informatics Consultant LLC, Rockville, MD, 20852, USA. Email: l.manor@gousinfo.com

ABSTRACT

Background: Renal cancer (RC) is a type of cancer that starts in the cells of the kidneys. Around 208,500 new cases of renal cancer worldwide are diagnosed yearly, accounting for just under 2% of all cancers. People who have a family history of RC have an increased risk of developing the disease. Recent years, an increased number of researches have been reported hundreds of genes related to the development of the disease. However, no systemic study has summarized these findings and has provided an objective view of these genes reportedly associated with RC.

Methods: We conducted a literature data mining (LDM) of over 1,100 articles covering publications from 1988 to April 2016, where 423 genes were reported to be associated with RC. We then performed a gene set enrichment analysis (GSEA) and a sub-network enrichment analysis (SNEA) to study the functional profile and pathogenic significance of these genes with RC. Lastly, we performed a network connectivity analysis (NCA) to study the associations between the reported genes. Literature and enrichment metrics analyses were used to discover genes with specific significance to the disease.

Results: 329/423 genes enriched 100 pathways ($p<1.2e-10$), demonstrating multiple associations with RC. Ten genes (IL6, VEGFA, HIF1A, EGFR, PTEN, TP53, FGF2, CTNNB1, HMOX1, and BRCA1) were identified as the top genes associated with leukemia in terms of both functional diversity and replication frequency. Additionally, three novel genes, CD274, NOTCH1, and CREB1, were found to play roles within many significant RC related pathways, suggesting that they were worthy of further study. Moreover, SNEA and NCA results indicated that many of these genes work as a functional network that plays roles in the pathogenesis of other RC related disorders.

Conclusion: Our results suggest that the genetic causes of RC were linked to a genetic network composed of a large group of genes. The gene lists, together with the literature and enrichment metrics provided in this study, can serve as a groundwork for further biological/genetic studies in the field.

Key Words: Renal Cancer; Literature Data Mining; Gene Set

ISSN:2397-9119

OPEN ACCESS

DOI: 10.20900/mo.20160010

Received: March 6, 2016

Accepted: May 10, 2016

Published: June 25, 2016

website:http://mo.qingres.com

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

Enrichment Analysis; Sub-network Enrichment Analysis; Network Connectivity Analysis

INTRODUCTION

Renal cancer (RC), also known as kidney cancer, is a type of cancer that starts in the cells in the kidney. Renal cell carcinoma (RCC) and transitional cell carcinoma (TCC) are the two most common types of RC, and their names reflect the type of cells from which the cancer develops. The lifetime risk of RC is approximately 1.6 percent for both men and women. (1) The number of new cases of kidney and renal pelvis cancer was 0.016%, and about 3.9 out of 100,000 men and women die of RC every year.(1) For cancers that were confined to the kidney, the five-year survival rate was 92%, if the cancer has spread to the surrounding lymph nodes, the survival rate was 65%, and if it has metastasized, the survival rate was 12%. (1) The highest rates were recorded in North America and the lowest rates in Asia and Africa. (2)

The prevalence of known risk factors includes cigarette smoking, obesity, regular use of NSAIDs and hypertension. (3) Genetic variations and their interactions with environmental exposures are believed to influence RC risk, but studies based on candidate gene approaches have not produced conclusive results. (4) Recent years have seen an increased number of articles reporting hundreds of genes/proteins which are related to RC, many of which were suggested as potential biomarkers for the disease, such as VEGFA, IL6 and MIR34A（5-7). Additionally, some genes (e.g., IL2) have been studied in clinical trials. (8)

Moreover, articles have reported genetic changes and quantitative changes of genes in the case of RC (9,10). Both increased and decreased gene expression levels/activities were observed. (10.11) To note, many genes were reported to influence the pathogenic development of RC with an unknown mechanism. (12)

Alternatively, some studies did suggest that a functional mechanism of a mutation can cause RC. Through exploring the effects of calcineurin inhibitors (CNI) on the expression and function of CXCR3 splice variants, Datta et al. found that CNI may mediate the progression of human RC by down regulating CXCR3-B and by promoting proliferative signals, likely through CXCR3-A. (13)

Nevertheless, no systematic analysis has evaluated the quality and strength of these reported genes as a functional network/group in order to study the underlying biological processes of RC. In this study, instead of focusing on a specific gene, we attempt to provide an encompassing view of the genetic-map through a comprehensive literature data mining (LDM), together with a gene set enrichment analysis (GSEA), as well as a sub-network enrichment analysis (SNEA) to study the underlying functional profile of the genes identified (14). We hypothesize that the majority, if not all, of the previously reported genes play roles in the development of RC, and that the major pathways/gene sets enriched by these genes are the candidate pathways through which those genes influenced the pathogenesis of the disease.

METHODS AND MATERIALS

The study is laid out as follows: 1) LDM to discover gene-MDD relations; 2) Enrichment analysis on the identified genes to study their pathogenic significance with RC. 3) Literature and enrichment metrics analysis to identify genes with specific significance. 4) NCA to test the functional association between these reported genes.

1. Literature data mining and article selection criterion

In this study, we performed a LDM for all articles available in the Pathway Studio database (www.pathwaystudio.com) until Apr. 2016, which coverd over 40 million scientific articles, seeking the ones that reported gene-RC relations. The LDM was conducted by employing the finely-tuned Natural Language Processing (NLP) system of the Pathway Studio software, which has the capability of identifying and extracting relationship data from scientific literature. Only the publications containing a biological gene-RC interaction defined by ResNet Exchange (RNEF) data format were included (http://www.gousinfo.com/).

2. Literature metrics analysis

For our literature metrics analysis, we propose two scores for each gene-disease relationship.

We define the reference number underlying a gene-disease relationship as the gene's reference score (RScore) in Eq. (1).

$$RScore = n \quad (1)$$

where n is the total number of references supporting a gene-disease relation.

We define the earliest publication age of a gene-disease relationship as the gene's age score (AScore) as Eq. (2).

$$AScore=\max_{(1\le i\le n)} ArticlePubAge_i \quad (2)$$

where n is the total number of references supporting a gene-disease relation, and

$$ArticlePubAge=Current\ date - Publication\ date + 1 \quad (3)$$

3. Enrichment metric analysis

Suppose a disease is associated with ith genetic pathways. We then define the gene-wise enrichment score (EScore) for the gene within a gene set as Eq. (4).

$$EScore_k=\sum_{(i=1)}^{m} (-\log_{10} pValue_i) / \max_{(1< i< n)} (-\log_{10} pValue_i) \quad (4)$$

Where $pValue_i$ is the enrichment score of the ith pathway with the gene set; $m \in R$ is the number of pathways including the kth gene; we define m as the PScore for the gene:

$$PScore_k = The\ number\ of\ pathways\ from\ R\ including\ the\ k th\ gene \quad (5)$$

We note here that PScore presents how many of the disease related pathways are associated with the genes, and EScore shows the significance of these pathways.

4. Enrichment analysis

To better understand the underlying functional profile and the pathogenic significance of the reported genes, we performed a GSEA and a sub-network enrichment analysis (SNEA) on 3 groups: 1) The whole gene list (423 genes); 2) 2-subgroups selected using the highest quality matrix scores. In addition, we conducted a network connectivity analysis (NCA) using the Pathway Studio network building module.

RESULTS

1 Summary of LDM results

In this study, we conducted a LDM of 1,100 articles reporting 423 genes associated with RC. According to the reported category of gene-RC relations, the articles can be generally clustered into 7 different classes: 1) biomarkers (4.91%), 2) cell Expression (1.64%), 3) clinical trials (1.82%), 4) genetic changes (42.64%), 5) quantitative changes (18.36%), 6) regulation (29.55%), and 7) state changes (1.09%).

For the 423 genes, 9.93% genes presented biomarker relationships to the disease, 3.55% with cell expression, 2.60% with clinical trials, 33.10% with genetic changes, 35.70% with quantitative changes, 39.24% with regulation, and 2.60% with state changes. To note, for a cadidate gene, there may be more than one articles reporting different relations with RC, and therefore one gene may have multiple relations. Specifically, a percentage (79.20 %) of the genes presented 1 type of relationship to the disease, whilst 20.80 % genes have been reported to have multiple relationships with the disease: 16.08 % genes have 2 types of relationships with the disease, 3.55 % with 3, and 1.18 % with 4, as shown in Fig. 1.

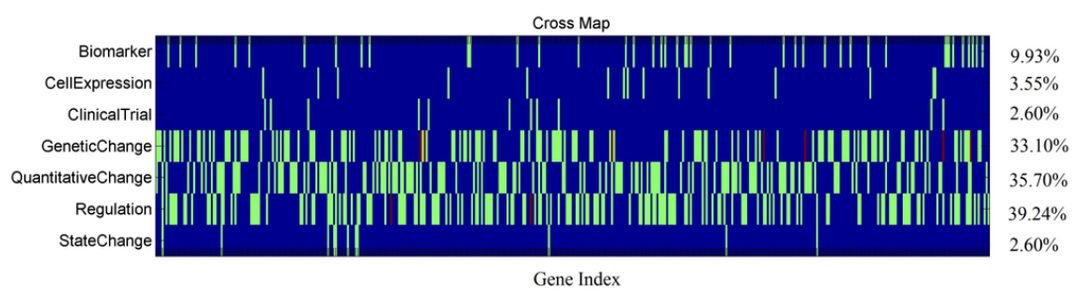
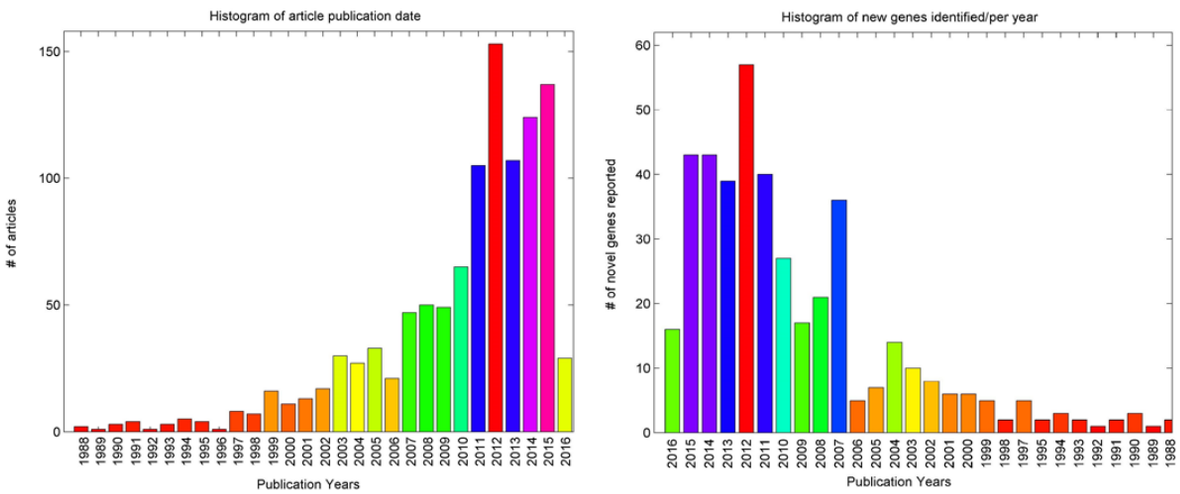


Fig. 1 Gene-wise relation type distribution of 423 genes

We presented the publication date distribution of these 1,100 articles in Fig. 2 (a), where we showed that this study covers literature data from the past 28 years (1988, 2016), with novel genes reported in each year (Fig. 2 (b)). To note, these articles have an average publication age of only 5.8 years, indicating that most of these articles were published in recent years. In addition, recent years, an increased number of publications have been seen, especially after 2010, with discoveries of more novel genes (Fig.2 (b)). Moreover, our analysis showed that the publication date distributions of the articles underlying each of the 423 genes were similar to that presented in Fig. 2.



(a) (b)

Fig. 2 Histogram of the publications reporting gene-disease relationships between RC and 423 genes. (a) presents the histogram of article publication date; (b) is the histogram of the number of novelty genes identified in each year

2. Marker Ranking

Using the 2 literature metric scores, we identified genes which were reported with supports from large numbers of articles, such as FH (72 articles), VHL (42 articles), and IL2 (39 articles). Some genes have been recently reported (roughly in the past year) such as FOXO4, HIST1H2APS4, and INPP5K.

Among these 423 genes, 16 were reported in 2016 and were listed in Table 1.The full results were provided in Supplementary Material 1. For comparison, Table 1 also lists the top 16 genes with the highest RScore (in descending order).

Table 1:Top 16 genes with reported associations with RC ranked by different scores

Genes with AScore=1	FOXO4; HIST1H2APS4; INPP5K;KLK3; MIR1236; MIR148B; MIR200A; MIR206; MIR22; MIR362; NOTCH1; SDPR; TICAM1; TRPC4; TRPM2; TRPM8
Genes By RScore	FH; VHL; IL2; MET; PTEN; FLCN; TSC2; MTOR; EGFR; TP53; HIF1A; VEGFA; WT1; EPAS1; PBRM1; SETD2

3. Enrichment Analysis

In this section, we present the GSEA and SNEA results for 3 different groups: All 423 genes, and both gene groups in Table 1.

3.1 Enrichment Analysis on all 423 genes

The full list of 100 pathways/gene sets enriched with 1.2e-10 (with 329/423 genes) is provided in Supplementary Material 2, where 20 pathways were enriched with p-values<1e-20 (with 272/423 genes) as listed in Table 2.

Among these 100 pathways/gene sets that were enriched, we identified 17 pathways/gene sets that were related to cell growth and proliferation (with 183/423 genes), 7 to cell apoptosis(148/423 genes), 4 to transcription factors (110/423 genes), 2 to protein phosphorylation (44/423 genes) and 1 to protein kinase (31/423 genes). In Table 2, the Jaccard similarity () is a statistic used for comparing the similarity and diversity of sample sets, and is defined by Eq. (6).

$$J_s(A,B) = \frac{A \cap B}{A \cup B} \qquad (6)$$

Where , A and B are two sample sets.

Table 2: Molecular function pathways/groups enriched by 423 genes reported

Pathway/gene set name	Hit type	GO ID	# of Entities	Overlap	p-value	Jaccard similarity
Response to drug	biological_process	0017035	509	72	6.79E-44	0.08
Positive regulation of cell proliferation	biological_process	0008284	568	69	1.18E-37	0.08
Negative regulation of cell proliferation	biological_process	0008285	471	63	7.62E-37	0.08
Response to hypoxia	biological_process	0001666	259	49	5.42E-36	0.08
Negative regulation of apoptotic process	biological_process	0006916	650	68	5.35E-33	0.07
Response to organic cyclic compound	biological_process	0014070	253	42	1.60E-28	0.07
Aging	biological_process	0016280	254	42	1.89E-28	0.07
Angiogenesis	biological_process	0001525	256	40	3.71E-26	0.06
Positive regulation of apoptotic process	biological_process	0043065	393	47	2.24E-25	0.06
Positive regulation of transcription from RNA polymerase II promoter	biological_process	0010552	1041	74	4.06E-25	0.05
Positive regulation of protein phosphorylation	biological_process	0001934	168	33	4.83E-25	0.06
Response to estradiol	biological_process	0032355	175	33	1.92E-24	0.06
Positive regulation of cell migration	biological_process	0030335	178	32	4.64E-23	0.06
Cell surface	cellular_component	0009929	645	55	4.71E-23	0.05
Response to organic substance	biological_process	0010033	153	30	7.73E-23	0.06
Positive regulation of gene expression	biological_process	0010628	293	38	6.36E-22	0.06
Cellular response to mechanical stimulus	biological_process	0071260	101	25	6.94E-22	0.05
Positive regulation of transcription, DNA-templated	biological_process	0045941	623	53	1.69E-21	0.05
Cellular response to organic cyclic compound	biological_process	0071407	94	24	2.09E-21	0.05
Tumor Suppressors	Pathway Studio Ontology	Pathway Studio Ontology	111	19	9.97E-21	0.04

Besides the 2 neuronal system related pathways listed in Table 2, there were 15 pathways/gene sets related to cell growth and proliferation (P-value: [1.7e-020, 9.5e-011]): positive regulation of smooth muscle cell

proliferation (GO: 0048661, p-value=1.7e-020, overlap: 21); regulation of cell proliferation (GO: 0042127, p-value=5.1e-018, overlap: 31); epidermal growth factor receptor signaling pathway (GO: 0007173, p-value=2.9e-017, overlap: 28); negative regulation of cell growth (GO: 0030308, p-value=6.1e-017, overlap: 24); growth factor activity (GO: 0008083, p-value=1.8e-016, overlap: 26); fibroblast growth factor receptor signaling pathway (GO:0008543, p-value=2.3e-016, overlap: 25); positive regulation of epithelial cell proliferation (GO: 0050679, p-value=1.4e-014, overlap: 17); vascular endothelial growth factor receptor signaling pathway (GO: 0048010, p-value=7.4e-014, overlap: 19); cellular response to transforming growth factor beta stimulus (GO: 0071560, p-value=1.8e-013, overlap: 15); cell proliferation (GO: 0008283, p-value=1.1e-012, overlap: 33); positive regulation of endothelial cell proliferation (GO: 0001938, p-value=2.3e-012, overlap: 15); positive regulation of fibroblast proliferation (GO: 0048146, p-value=2.7e-011, overlap: 13); negative regulation of smooth muscle cell proliferation (GO: 0048662, p-value=5e-011, overlap: 11); positive regulation of T cell proliferation (GO: 0042102, p-value=5.3e-011, overlap: 13); transforming growth factor beta receptor signaling pathway (GO: 0007179, p-value=9.5e-011, overlap: 18)

There were 5 additional pathways/gene sets related to cell apoptosis (P-value: [1.1e-019,1.2e-010] and 2 additional pathways/gene sets related to transcription factors: regulation of apoptotic process (GO: 0042981, p-value=1.1e-019, overlap: 35); apoptotic process (GO: 0008632, p-value=1.2e-019, overlap: 57); activation of cysteine-type endopeptidase activity involved in apoptotic process (GO: 0006919, p-value=1e-012, overlap: 17); negative regulation of neuron apoptotic process (GO: 0043524, p-value=7e-011, overlap: 19); apoptotic signaling pathway (GO: 0097190, p-value=1.2e-010, overlap: 17); negative regulation of transcription from the RNA polymerase II promoters (GO: 0000122, p-value=2e-012, overlap: 46); regulation of transcription from the RNA polymerase II promoters in response to hypoxia (GO: 0061418, p-value=1.8e-011, overlap: 10).

Furthermore, the results presented 1 extra pathway related to protein phosphorylation (P-value: [2.2e-014]) and 1 pathway was related to protein kinase (P-value: [7.2e-011]): positive regulation of peptidyl-serine phosphorylation (GO: 0033138, p-value=2.2e-014, overlap: 17) and protein kinase binding (GO: 0019901, p-value=7.2e-011, overlap: 31).

Besides GSEA, we also performed a SNEA using Pathway Studio with the purpose of identifying the pathogenic significance of the reported genes with other disorders that were potentially related to RC. We provide the full list of results in Supplementary Material 3. In Table 3, we present the disease related sub-networks enriched with a p-value<4.24E-167.

Table 3: Sub-networks enriched by the 423 genes reported

Gene Set Seed	Total # of Neighbors	Overlap	p-value	Jaccard similarity
Breast Cancer	3146	308	7.46E-187	0.09
Cancer of Stomach	1833	256	1.46E-183	0.13
Neoplasm Metastasis	1843	256	6.11E-183	0.13
Carcinoma, Hepatocellular	2417	279	6.83E-182	0.11
Lung Cancer	1723	249	5.84E-181	0.13
Colorectal Cancer	2291	270	2.02E-176	0.11
Ovary Cancer	1402	225	4.66E-170	0.14
Clear Cell Renal Cell Carcinoma	471	159	9.55E-170	0.22
Cancer of Pancreas	1159	211	1.74E-169	0.16
Prostate Cancer	1954	249	4.50E-167	0.12

From Table 3, we see that many of these reported RC related genes were also identified in other cancer diseases, with a large percentage of overlap (Jaccard similarity>=0.10).

3.2 Enrichment Analysis on the top 16 genes with highest scores

We compare here the top 16 genes listed in Table 1 in terms of GSEA and SNEA results. The top 10 pathways/sub-networks for the AScore group and the RScore group (Table 4 and Table 5) are presented here. The full report is in Supplementary Material 2 and 3.

Using the same enrichment p-value threshold (p<6E-004), we identified 23 pathways/gene sets that were enriched with the 16 genes with top AScores, while the number for the RScore group is 119. The full lists of

these pathways/gene sets are provided in Supplementary Material 2. Table 4 presents the top 10 pathways enriched with the 16 genes from AScore and RScore groups, respectively.

Table 4: Pathways/groups enriched by 16 genes with the highest AScore and RScore

	Pathway/gene set Name	GO ID	p-value
The first 10 pathways/ gene sets enriched by top 16 genes with highest AScores	store-operated Ca2+ channel	Pathway Studio Ontology	7.60E-08
	TC 1.A.4.5	Pathway Studio Ontology	2.13E-06
	Calcium channel activity	0005262;	8.33E-06
	Non-voltage Ca++ import proteins	Pathway Studio Ontology	9.10E-06
	Calcium ion transmembrane transport	0070588;	2.32E-05
	Calcium ion transport	0006816;	3.03E-05
	Ion channel activity	0005216;	5.89E-05
	Oligodendrocyte differentiation	0048709;	9.11E-05
	Ion transmembrane transport	0034220;	2.08E-04
	Negative regulation of angiogenesis	0016525;	4.00E-04
The first 10 pathways/ gene sets enriched by top 16 genes with highest RScore	Positive regulation of transcription from RNA polymerase II promoter	0010552;	1.48E-11
	Tumor Suppressors	Pathway Studio Ontology	1.18E-10
	Cellular response to hypoxia	0071456;	4.01E-10
	Regulation of thymocyte apoptotic process	0070243;	1.45E-09
	Regulation of transcription from RNA Polymerase II promoter in response to hypoxia	0061418;	3.91E-09
	Negative regulation of apoptotic process	0006916;	8.39E-09
	Negative regulation of cell proliferation	0008285;	2.63E-08
	Negative regulation of cell size	0045792;	5.98E-08
	Lactation	0007595;	1.00E-07
	Positive regulation of protein phosphorylation	0001934;	1.26E-07

From Table 4, we see that the genes with the top AScores and those with the top RScores are enriching different groups of pathways with different p-values (AScore group: 7.60E-08~4.00E-04; RScore group:1.48E-11~1.26E-07), indicating that the newly reported genes were functionally different than from those most frequently reported.

Additionally, we observed that 5 out of the 10 pathways/gene sets enriched by the RScore group (Table 4) are presented in Table 2, which lists the top 20 pathways/gene set enriched with 272 /423 genes, and the number for AScore group was 0.

For the SNEA analysis, we performed an enrichment analysis against disease sub-networks. The full list of results is provided in Supplementary Material 3. Table 5 presents the top 10 disease-related sub-networks enriched by the top 16 genes from AScore group and RScore group, respectively.

Table 5: SNEA results by 16 genes with the highest AScore and RScore

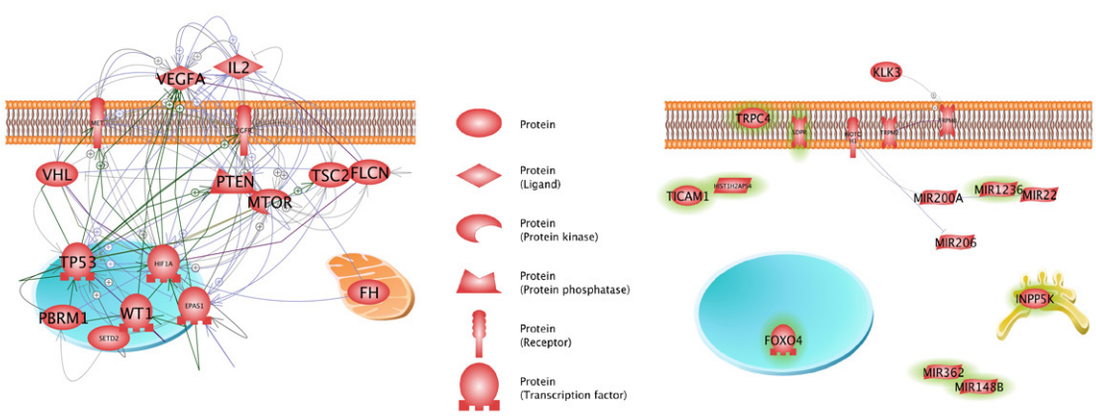
	Gene Set Seed	Overlap	p-value	Jaccard similarity

The first 10 pathways/gene sets enriched by top 16 genes with highest AScores	Diabetes Mellitus	11	2.03E-08	0
	Cancer of Head and Neck	5	1.62E-06	0.01
	monocrotaline-induced pulmonary hypertension	3	2.99E-06	0.05
	Pulmonary Disease, Chronic Obstructive	5	4.21E-05	0.01
	Lung Cancer	7	5.53E-05	0
	Carcinoma, Endometrioid	3	5.85E-05	0.02
	Adenocarcinoma, Clear Cell	3	6.00E-05	0.02
	visceral pain	2	6.81E-05	0.06
	Pulpitis	2	6.81E-05	0.06
	Cancer of Stomach	7	8.21E-05	0
The first 10 pathways/gene sets enriched by top 16 genes with highest RScore	von Hippel-Lindau Disease	11	2.14E-28	0.24
	Papillary Renal Cell Carcinoma	12	2.89E-28	0.16
	kidney cyst	12	1.33E-27	0.14
	Clear Cell Renal Cell Carcinoma	15	2.40E-25	0.03
	kidney metastasis	11	4.50E-25	0.14
	thyroid medullary carcinoma	10	3.02E-20	0.08
	Adenocarcinoma, Clear Cell	10	3.93E-20	0.08
	cancer family syndromes	9	4.83E-20	0.12
	Neuroendocrine Tumors	10	3.29E-18	0.05
	Li-Fraumeni Syndrome	8	1.02E-17	0.12

From Table 5, we see that both groups enriched other RC health-related sub-networks. However, the enrichment p-values of the RScore group are much more significant than those of the AScore group, with higher Jaccard similarities.

4. Connectivity Analysis

In addition to GSEA and SNEA, we performed a NCA on the top 16 genes with the highest RScores and AScores (from Table 1) to generate gene-gene interaction networks. Results for the RScore group showed 104 connections among 16/16 genes, with more than 300 literature support. In contrast, genes within the AScore group demonstrated only 6 relations among 7/16 genes, as shown in Fig. 3 (b), with 9 genes showing no direct relation with other genes in the group (Fig. 3 (b); highlighted in Green). This observation is consistent with the GSEA and SNEA, suggesting that genes with the smallest AScores are not as functionally close to each other as those from the RScore group.



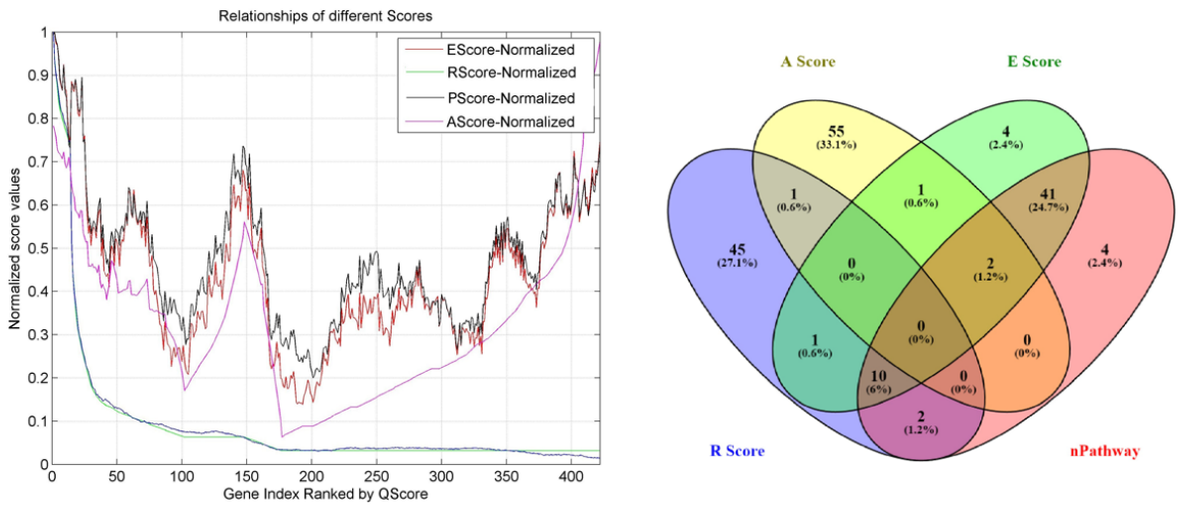
(a) By RScore group (b) By AScore group

Fig. 3 Connectivity networks built by 16 genes from different groups. The networks were generated using Pathway Studio; The un-related genes are highlighted in blue.

5. EScore Analysis

Through GSEA, we also generated two biological metrics, EScore and PScore, for each gene. The value of a PScore represents how many RC associated pathways involve the gene, and EScore shows the significance of those pathways.

To compare the EScore and PScore with the two literature metrics, we conducted a correlation analysis using the averaged metric values of all the 423 genes at a group level, shown in Fig.4 (a). We used a group size of 14 genes: we first sorted the 423 genes by RScore and averaged each type of metric values using a moving window of length 14. Results showed that the average scores were strongly correlated, especially for the top ones ranked by different socres, as shown in Fig. 4 (a) and Table 6. The group-wise PScore and EScore, were extremely correlated (p≈1).



(a) (b)

Fig.4 Comparison of different metrics ranking the 423 genes. (a) A Venn diagram of the top 59 genes selected by different metrics; (b) Comparison of average metrics values with a gene set size of 14.

Table 6: Pearson correlation coefficients between different metrics

	RScore	EScore	PScore	AScore
RScore	1.00	0.62	0.63	0.47
EScore	0.62	1.00	0.99	0.86
PScore	0.63	0.99	1.00	0.83
AScore	0.47	0.86	0.83	1.00

In addition to the group-wise correlation analysis, we also performed a cross-analysis of the top 59 genes selected using different scores (corresponding to the number of genes reported within the past two years (2015~Apr. 2016)), and present a Venn diagram in Fig.4 (a) (Oliveros, 2007-2015).

There was a strong overlap between PScore and EScore group (53/59). These 53 genes related to the most pathways that were significantly enriched. Additionally, we noted that the AScore group presented an overlap of one gene with RScore group (CD274: 5 references), and an overlap of 2 genes with both the EScore and PScore groups (NOTCH1: 1 reference, CREB1: 1 reference). These novel genes were reported within the last 2 years and demonstrated a relatively high frequency of replication or multiple functional associations with the disease, (PScore:15.00 ±1.41 pathways), suggesting that they are worthy of further study.

On the other hand, 10 genes were identified to overlap the EScore, PScore and RScore groups, including IL6, VEGFA, HIF1A, EGFR, PTEN, TP53, FGF2, CTNNB1, HMOX1, and BRCA1, with RScore=11.20 ± 6.88 references, PScore: 27.70 ± 8.39 pathways. Additionally, there were 41 genes observed in both the PScore group and the EScore group, but not in the RScore group, including: TGFB1, TNF, PDGFB, BCL2, PRRCA,

FAS, AKT1, PTK2, TGFBR2, CAV1, BMP2, IGF1, CDKN1B, KDR, MYC, HRAS, SERPINE1, MMP9, CCL2, CDKN1A, AGT, STAT1, IGF2, SFRP1, EPO, CDKN2A, IL4, FGF1, MDM2, HSPD1, GSK3B, NOD2, IFNG, MMP2, COL1A1, CASP1, AGER, TIMP1, IL18, CXCL12, and RPS27A. These genes play roles within many significant pathways of the disease (21.80 ±7.09 pathways). The RC-gene relationships involving these genes were old (ASocre: 10.59 ±6.78 years) and were not frequently replicated (1.49 ± 0.75 references).

DISCUSSION

We performed a LDM on 1,100 articles (from 1988 to April 2016) reporting 423 genes that were associated with RC. We provide in Supplementary Materials 1 the full gene list together with the literature and enrichment metrics scores. In addition, results from GSEA and SNEA support the literature that most of these genes may play roles in the pathogenesis of RC. Furthermore, NCA showed that many of these genes were functionally linked to one another.

As an automatic data mining approach, the NLP technique is effective and efficient in dealing with large amounts of literature data for LDM. However, the automatic LDM method may produce some false positives. Therefore, the results of this study were intended to provide a map for the current field of genetic study of RC and laid the groundwork for further biological/genetic studies in the area. For this purpose, we provided in Supplementary material 1 detailed information of all the 1,100 articles studied for further investigation, including the sentences where a specific relationship is located.

Although our analysis did not specifically focus on individual gene, we noticed that the 423 genes identified were not equal in terms of publication frequency (RScore), or their novelties (AScore), or their functional diversity (EScore). Using the proposed quality metric scores, one is able to rank the genes according to different needs/significance and picked the top ones to further analysis (Table 1). For example, the top 5 genes by AScore, namely FOXO4, HIST1H2APS4, INPP5K, KLK3, and MIR1236 were recently reported. Alternatively, FH, VHL, IL2, MET and PTEN are the top 5 genes that were most often replicated in studies (with highest RScores), suggesting that they are common variables with RC.

Additionally, we noted that for the top 100 pathways enriched with 329/423 genes (Supplementary Material 2), some genes were duplicated in multiple significantly enriched pathways to present high EScore, such as TGFB1 (46/100 pathways), IL6 (38/100 pathways), TNF (37/100 pathways), VEGFA (39/100 pathways) and PDGFB (36/100 pathways). These genes played multiple roles within different genetic pathways associated with RC, indicating their biological significance with the disease.

To note, 10 genes were identified to overlap the EScore, PScore and RScore groups. These genes were frequently replicated (11.20±6.88 references) in previous studies showing association with RC, and play roles within multiple (27.70±8.39) significant pathways associated with RC. Our results indicated that these genes are highly likely to posses pathogenic significance to RC.

We also identified 3 novel genes (NOTCH1, CD274, and CREB1) in both of the EScore and PScore groups, which were reported last 2 years with a few references. However, they play roles within multiple significant pathways implicated with RC, warranting further study. For example, NOTCH1 was recently reported in 2016 with only one reference. However, this gene is involved in many pathways previously implicated with RC or other cancers, such as: positive regulation of cell proliferation (0008284), negative regulation of cell proliferation (0008285), angiogenesis (0001525), positive regulation of apoptotic process (0043065), positive regulation of cell migration (0030335), regulation of cell proliferation (0042127), positive regulation of epithelial cell proliferation (0050679), negative regulation of canonical Wnt signaling pathway (0090090), and organ regeneration (0031100). (15-17)

Furthermore, 41 genes were observed in both the PScore group and the EScore group, but not in the RScore group. Although the RC-gene relationships involving these genes were old (ASocre: 10.59 ± 6.78 years) and were not frequently replicated (1.49 ± 0.75 references), our results suggest that they may be worthy of further study.

In addition, we observed that most genes identified by this LDM were included in the pathways previously implicated with RC, including 17 cell growth and proliferation-related pathways, 2 protein phosphorylation-related pathways, 4 pathways/gene sets were related to transcription factors, 7 cell apoptosis-related pathways and 1 protein kinase related pathways [18-22]. We hypothesize that the majority of these reported genes, especially the ones that were identified from significantly enriched pathways, should be functionally linked with RC. Although there may be false positives from separate studies into the publications, it is less likely that a large group of genes were falsely perturbed. (14)

When the members of a gene set exhibit strong cross-correlation, GSEA can boost the signal-to-noise ratio and make it possible to detect modest changes in individual genes (14). The NCA analysis showed that many of the frequently reported genes relating to RC were functionally associated with one another (Fig. 3), which is supported by hundreds of scientific reports. Furthermore, we noticed that 329/423 of the genes were included in the top 100 enriched pathways (p-value<1.2e-10), and 272/423 genes are in the top 20 pathways listed in Table 2 (p-value<1e-20). If we define that two genes were functionally related to each other as their co-existence within same genetic pathway, then we saw that around 77.8% of the 423 genes are functionally related. The results indicate that these functionally linked genes likely presented their relationships as true discoveries rather than noise (false positives).

In addition to GSEA, we performed a SNEA, which was implemented in Pathway Studio using master casual networks generated from more than 6.5 million relationships derived from more than 4 million full text articles and 25 million PubMed abstracts. The ability of the Pathway Studio automated NLP technology to quickly update the terminologies and linguistic rules used by the NLP systems ensures that new terms can be captured soon after they entering regular use in the literature. Updating takes place on a weekly basis. This extensive database of interaction data provided high levels of confidence when interpreting experimentally-derived genetic data against the background of previously published results (Pathway Studio Web Help). SNEA results demonstrated that many of the 423 genes (>90%) that are also identified as causal genes for other health disorders (Breast Cancer, Stomach Cancer, Lung Cancer, et al) are strongly associated with RC. (23-25)

This study, however, has several limitations that should be considered in future work. The literature data of the 1,100 articles studied were extracted from the Pathway Studio database. Although the Pathway Studio database covers over 40 million articles, it is still possible that some articles studying gene-RC associations were beyond the scope of coverage. Additionally, the metrics scores, RScore, AScore, EScore, and PScore were proposed as significance measures of the literature reported gene-disease relations. Although they are related, they are not the direct biological significance measures of the genes to the disease. Experiments should be done in the future to test the networks and these metrics.

CONCLUSION

Results from this up-to-date LDM reveal that the 423 genes identified multiple types of associations with RC, and provided a map that provided an overview for the current genetic study of RC. The literature and enrichment metrics discovered top genes with specific significance. In addition, NCA and enrichment analysis results suggested that these genes play significant roles as a network in the pathogenesis of RC, as well as in the pathogenesis of many other RC-related disorders. Our results suggest that these genes may operate as a functional genetic network which influence the development of the disease.

We conclude that RC is a complex disease with genetic causes that were linked to a network composed of a large group of genes. LDM, together with GSEA, SNEA, and NCA, can serve as an effective approach in finding these potential target genes. This study provides an landscape map with metrics for the current field of genetic research into RC, and can be used as a groundwork for further biological/genetic studies in the area.

Declaration of interests

The authors declare no conflict of interests.

SUPPLEMENTARY

1. Marker summary at http://www.qingres.com/Upload/Excel/Supplementary 1_marker summary.xlsx
2. Pathways Enriched by different groups at http://www.qingres.com/Upload/Excel/Supplementary 2 _ Pathways Enriched by different groups.xlsx
3. SubNetworks Enriched by different groups at http://www.qingres.com/Upload/Excel/Supplementary 3_ subNetworks Enriched by different groups.xlsx

REFERENCES

1. National Cancer Institute. 2016. SEER Stat Fact Sheets: Kidney and Renal Pelvis Cancer. http://seer.

cancer.gov/statfacts/html/kidrp.html.

2. Wong-Ho Chow, Linda M. Dong and Susan S. Devesa. Epidemiology and risk factors for renal cancer. Nat Rev Urol. 2010; 7(5): 245–257.
3. Sudarshan S, Linehan WM. Genetic basis of cancer of the kidney. Semin Oncol. 2006; 33(5):544-51.
4. Wong-Ho Chow, Susan S. Devesa. Contemporary epidemiology of renal cell cancer. The Cancer J. 2008; 14:288–301.
5. Escudier B, Eisen T, Stadler W, Szczylik C, Oudard S, Staehler M, et al. Sorafenib for treatment of renal cell carcinoma: Final efficacy and safety results of the phase III treatment approaches in renal cancer global evaluation trial. Journal of Clinical Oncology. 2009; 27(20):3312-3318.
6. Vanden Berghe W, Vermeulen L, De Wilde G, De Bosscher K, Boone E, Haegeman G. Signal transduction by tumor necrosis factor and gene regulation of the inflammatory cytokine interleukin-6. Biochemical Pharmacology. 2000; 60(8):1185-1195.
7. Larkin S, Kyprianou N. Molecular signatures in urologic tumors. International Journal of Molecular Sciences. 2013; 14(9):18421-18436.
8. Fenton RG, Steis RG, Madara K, Zea AH, Ochoa AC, Janik JE, et al. A phase I randomized study of subcutaneous adjuvant IL-2 in combination with an autologous tumor vaccine in patients with advanced renal cell carcinoma. J Immunother Emphasis Tumor Immunol. 1996; 19(5):364-74.
9. Elbelt U, Trovato A, Kloth M, Gentz E, Finke R, Spranger J, et al. Molecular and clinical evidence for an ARMC5 tumor syndrome: Concurrent inactivating germline and somatic mutations are associated with both primary macronodular adrenal hyperplasia and meningioma. Journal of Clinical Endocrinology and Metabolism. 2015; 100(1):E119-E128.
10. Niers T, Richel D, Meijers J, Schlingemann R. Vascular endothelial growth factor in the circulation in cancer patients may not be a relevant biomarker. PLoS ONE. 2011; 6(5): e19873.
11. Hwang J, Uchio E, Linehan W, Walther M. Hereditary renal cancer. Urologic Clinics of North America. 2003; 30(4):831-842.
12. Li L, Shen C, Nakamura E, Ando K, Signoretti S, Beroukhim R, et al. SQSTM1 Is a Pathogenic Target of 5q Copy Number Gains in Kidney Cancer. Cancer Cell. 2013; 24(6):738-750.
13. Datta D, Contreras AG, Grimm M, Waaga-Gasser AM, Briscoe DM, Pal S. Calcineurin Inhibitors Modulate CXCR3 Splice Variant Expression and Mediate Renal Cancer Progression. J Am Soc Nephrol. 2008; 19(12): 2437–2446.
14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102(43):15545-50.
15. Evan GI, Vousden KH. Proliferation, cell cycle and apoptosis in cancer. Nature. 2001; 411(6835): 342-8.
16. Klaus A, Birchmeier W. Wnt signalling and its impact on development and cancer. 2008; 8(5):387-98.
17. Tataria M, Perryman SV, Sylvester KG. Stem cells: tissue regeneration and cancer. Semin Pediatr Surg. 2006; 15(4):284-92.
18. Zhang HJ, Tao J, Sheng L, Hu X, Rong RM, Xu M, et al. Twist2 promotes renal cancer cell proliferation and invasion by regulating ITGA6 and CD44 expression in the ECM-receptor interaction pathway. Onco Targets Ther. 2016; 9:1801-12.
19. Li S, Kong Y, Si L, Chi Z, Cui C, Sheng X, et al. Phosphorylation of mTOR and S6RP predicts the efficacy of everolimus in patients with metastatic renal cell carcinoma. BMC Cancer. 2014; 14:376.
20. Samarghandian S, Afshari JT, Davoodi S. Honey induces apoptosis in renal cell carcinoma. Pharmacogn Mag. 2011; 7(25): 46–52.
21. Schödel J, Grampp S, Maher ER, Moch H, Ratcliffe PJ, Russo P, et al. Hypoxia, Hypoxia-inducible Transcription Factors, and Renal Cancer. Eur Urol. 2016; 69(4):646-57.
22. Bracarda S, Caserta C, Sordini L, Rossi M, Hamzay A, Crinò L. Protein kinase inhibitors in the treatment of renal cell carcinoma: sorafenib. Ann Oncol. 2007; 18 Suppl 6:vi22-5.
23. Van Wynsberge LK, Vierling P, Lampel A. 2004. Breast cancer metastatic to a renal cell carcinoma. Aktuelle Urol.35(6):505-7.

24. Pollheimer MJ, Hinterleitner TA, Pollheimer VS, Schlemmer A, Langner C. Renal cell carcinoma metastatic to the stomach: single-centre experience and literature review. BJU Int. 2008; 102(3):315-9.

25. Merimsky O, Gez E, Weitzen R, Nehushtan H, Rubinov R, Hayat H, et al. Targeting pulmonary metastases of renal cell carcinoma by inhalation of interleukin-2. Ann Oncol. 2004; 15(4):610-2.