Innovative microRNA-lncRNA-mRNA co-expression analysis to understand the pathogenesis and progression of diabetic kidney disease

Lihua Zhang
Department of Geriatric Endocrinology
First Affiliated Hospital of Kunming Medical University
Kunming, Yunnan 650031, China

Rong Li
Department of Nephrology
First People’s Hospital of Yunnan Province
Kunming, Yunnan 650032, China

Qiuping Yang
Department of Geriatric Endocrinology
First Affiliated Hospital of Kunming Medical University
Kunming, Yunnan 650031, China

Yanan Wu
Department of Geriatric Endocrinology
First Affiliated Hospital of Kunming Medical University
Kunming, Yunnan 650031, China

Jingshan Huang†
Computer Science Department, School of Computing
University of South Alabama
Mobile, Alabama 36688, U.S.A.

Bin Wu‡
Department of Endocrinology and Metabolism
First Affiliated Hospital of Kunming Medical University
Kunming, Yunnan 650031, China

Abstract—Diabetic kidney disease (DKD) is a serious disease that presents a major health problem worldwide. There is a desperate need to explore novel biomarkers to further facilitate the early diagnosis and effective treatment in DKD patients so that to prevent them to develop end-stage renal disease (ESRD). However, most of regulation mechanisms at genetic level in DKD still remain unclear. In this work-in-progress paper, we describe our innovative methodologies that integrate biological, statistics, and computational approaches to investigate important roles performed by regulations among microRNAs (miRs), long non-coding RNAs (lncRNAs), and messenger RNAs (mRNAs) in DKD. We conducted a series of experiments and identified a list of miRs and lncRNAs as potential novel biomarkers, along with the set of target genes regulated by discovered miRs. Our initial analysis results are promising in better understanding regulation mechanisms of miRs and lncRNAs on the pathogenesis and progression of DKD.

Keywords—microRNA, lncRNA, diabetes, nephropathy, diabetic kidney disease, co-expression analysis, biomarker.

I. INTRODUCTION

The International Diabetes Federation (IDF) report indicated that in 2011 there were 366 million diabetic patients worldwide — this number was estimated to increase to 552 million in 2030. In addition, the prevalence of diabetes has significantly increased in both developed and developing countries [1] [2]. For example, diabetes patients in China were approximately 113.9 million in 2013 [3]. Moreover, research has demonstrated that diabetes is the most common disease causing chronic kidney disease (CKD), and diabetic patients have a 2.6-fold risk to develop CKD compared to non-diabetic patients. In fact, diabetic kidney disease (DKD) is one of the most important microvascular complications of diabetic patients [4]. In most cases, there are no specific clinical manifestations during the early stage of DKD. More seriously, there are currently no effective methods to prevent DKD patients to develop kidney failure, which is also known as end-stage renal disease (ESRD), and which is one important cause of death in patients with diabetes.

The gold diagnostic criterion of various kidney diseases is to perform pathological biopsy on renal tissues, which unfortunately has high risk of patient injury. Quantitative test of urinary micro-protein is one popular non-invasive diagnostic indicator. But there are quite some limitations in this method due to various factors such as: the complexity of the protein, the variation caused by post-translational modifications, the stability of the specimen, and so forth. Especially, for patients at the early stage of nephropathy, the false negative rate of urinary micro-protein test is particularly high. Therefore, there is an urgent need to explore novel biomarkers that are both sensitive and specific to guide early diagnosis of DKD in an effective manner; it is then possible to give proper early intervention and treatment to diabetic patients to prevent the progression from DKD to ESRD.

Towards this end, we will need to significantly enhance our understanding of the genetic, genomic, and epigenetic foundation underlying DKD disease process. Thus, it is necessary to develop more advanced methodologies that are capable of integrating biological, statistic, and computational approaches in a seamless manner; only this way will it be possible for us to explore a more accurate representation of biological processes that regulate DKD and the progression from DKD to ESRD. That said, we report in this work-in-progress paper our on-
going efforts in effectively combining biological experiments and software analysis (based upon statistics, domain ontologies, and semantic technologies), so that to better investigate important roles performed by regulations among microRNAs (miRs), long non-coding RNAs (lncRNAs), and messenger RNAs (mRNAs) in DKD.

The rest of this paper is organized as follows. Section II summarizes state-of-the-art research in DKD and semantic technologies; Section III describes our methodologies; Section IV reports our current findings along with discussion; and finally, Section V concludes with important future work.

II. RELATED WORK

A. Related work in mechanisms performed by miRs and lncRNAs in DKD

Although far from being completely elucidated, the genetic regulation mechanisms performed by various miRs and lncRNAs in DKD have attracted a lot research work.

According to the experimental results in vitro animal models reported in [5] [6], miR-192 expression increases in DKD kidney tissue, and inhibition of miR-192 expression (or knockout of miR192) can reduce the discharge proteinuria and renal fibrosis in mice with type 1 diabetes. On the other hand, the expression of the same miR is reduced in DKD patients. In fact, miR-192 expression was shown to be negatively correlated with renal tubular interstitial fibrosis and renal function decline. Zhong et al. [7] demonstrated that high glucose station can stimulate TGF-β1/Smad3 signal pathways of kidney cells and thus cause higher miR21 expression. Lacking of Smad3 can prevent cells from up-regulating miR-21 in response to TGF-β1/Smad3 signal pathways, which promotes renal fibrosis. As a result, inhibition of miR-21 may be a therapeutic approach to suppress renal fibrosis. Another study [8] showed that miR21 expression was down-regulated in mice of type 2 diabetes model’s kidney tissue. When excessive miR-21 expression was inhibited, the occurrence of apoptosis can be found in animal model. Interestingly, the role of miR-29 in DKD has conflicting conclusions [9]: the expression of miR-29 can be found in animal model. Interestingly, the role of miR-29 in DKD has conflicting conclusions [9]: the expression of miR-29 in diabetes patients can be observed increased or decreased.

Although a small number of lncRNAs have been functionally characterized [10] [11], it remains questionable whether the majority is biologically meaningful or merely transcriptional noise. The few lncRNAs that have been characterized to date exhibit adverse range of functions and expression in specific cell types and/or localization to specific subcellular compartments. In recent years a number of studies [12–14] have found that quite some lncRNAs are associated with the onset of diabetes. In particular, the islet tissue specific anti-sense lncRNAs were reported to be associated with the pathogenesis of diabetes such as neonatal diabetes. In addition, MALAT1, ANRIL/CDKN2BAS, HI-LNC25, and KCNQ1OT1 were reported to closely associated with type 2 diabetes susceptibility genes. Research [15] has also shown that, whereas many lncRNAs are interacted with protein functions in chromatin remodeling and gene transcription, some lncRNAs function by regulating miR functions. While lncRNAs can be used as miR precursors, they can also be combined with miRs and participate in miR regulation network, thus affecting more functional gene expression.

B. Related work in semantic technologies

In biomedical investigation, when we need to integrate a large number of data sources that have heterogeneous semantics, semantic technologies, which are based on domain ontologies, can render great assistance.

Bio-ontologies have been widely utilized nowadays, such as: Gene Ontology (GO) [16], the most successful and widely used bio-ontology with three independent sub-ontologies (biological processes, molecular functions, and cellular components); Non-Coding RNA Ontology (NCRO) [17] [18], an Open Biological and Biomedical Ontologies (OBO) [19] candidate reference ontology in non-coding RNA (ncRNA) domain; and Ontology for MicroRNA Target (OMIT) [20] [21], an application ontology to provide the miR community with common data elements and data exchange standards.

Part of our proposed methodologies in this paper are closely related to semantic search [22], which usually requires the utilization of structured knowledge to model/interpret search queries, by using formal logic for example. One popular idea in numerous semantic search systems ([23–27] for example) is to expand the query keywords utilizing synonyms and other relations not originally part of the query. A second way to implement semantic search is to translate the original keyword-based search into some formal semantic queries through the adoption of domain ontologies.

III. METHODOLOGIES

The overview of our methodologies are exhibited in Figure 1, consisting of the following five steps.

- Step (1). Among plasma samples from DKD patients vs. controls, we will use the microarray method to simultaneously detect different expression levels of both miRs (using Agilent chips) and lncRNAs (using Affymetrix HTA2.0 chips). The results will be the first-round screened miRs and lncRNAs. We will then utilize GeneSpring package (version 13.1 by Agilent Technologies) [28] to perform the statistic analysis: mean ± standard deviation for normal distribution; single-factor analysis of variance for comparison between groups; and least significant difference (LSD) test for comparison between a pair.

- Step (2). We will performe miR-lncRNA-mRNA co-expression network analysis based on Pearson correlation coefficient calculation. A total of 300 differentially expressed molecules will be returned, that is, top 100 miRs, top 100 lncRNAs, and top 100 mRNAs (in terms of their connection degrees) will be returned. Next, we will utilize the igraph software package [29] to construct a network connection diagram. Following this, we can obtain a set of second-round screened miRs and lncRNAs.

- Step (3). OmniSearch software tool [30–32], which is based upon domain ontologies and semantic technologies, will be used to retrieve a set of target genes (i.e., target mRNAs) for each miR in the second-round screened list from Step (2). Along with target mRNAs, information on GO annotations and PubMed publications will also be obtained for each and every
TABLE I. CLINICAL BACKGROUND OF PATIENTS AND CONTROLS IN OUR EXPERIMENTS

<table>
<thead>
<tr>
<th></th>
<th>DKD 1</th>
<th>DKD 2</th>
<th>DKD 3</th>
<th>Control 1</th>
<th>Control 2</th>
<th>Control 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>male</td>
<td>male</td>
<td>female</td>
<td>male</td>
<td>male</td>
<td>female</td>
</tr>
<tr>
<td>Age (y)</td>
<td>45</td>
<td>51</td>
<td>46</td>
<td>43</td>
<td>45</td>
<td>57</td>
</tr>
<tr>
<td>Diabetes Mellitus (DM) Duration (y)</td>
<td>3</td>
<td>10</td>
<td>7</td>
<td>NGT</td>
<td>NGT</td>
<td>NGT</td>
</tr>
<tr>
<td>Body Mass Index (BMI) (kg/m²)</td>
<td>22.94</td>
<td>24.11</td>
<td>26.84</td>
<td>21.48</td>
<td>20.53</td>
<td>19.81</td>
</tr>
<tr>
<td>Glycosylated Hemoglobin (HbA1c) (%)</td>
<td>8.1</td>
<td>7.8</td>
<td>6.7</td>
<td>5.3</td>
<td>5.4</td>
<td>5.6</td>
</tr>
<tr>
<td>Creatinine (umol/l)</td>
<td>578.0</td>
<td>220.0</td>
<td>206.0</td>
<td>89.0</td>
<td>78.9</td>
<td>70.8</td>
</tr>
<tr>
<td>Urinary Albumin/Creatinine Ratio (ACR) (mg/g)</td>
<td>5630</td>
<td>600</td>
<td>11354</td>
<td>20</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>Urine Protein</td>
<td>+++++</td>
<td>++</td>
<td>+++++</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serum Cystatin C (mg/l)</td>
<td>4.98</td>
<td>2.89</td>
<td>2.14</td>
<td>0.56</td>
<td>0.76</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Fig. 1. Overview of our combined approach to investigate miR-lncRNA-mRNA co-expression in DKD.

Fig. 2. Affymetrix HTA2.0 outcome: differentially expressed lncRNAs of DKD patients compared to controls.

Fig. 3. Agilent miR microarray outcomes: differentially expressed miRs of DKD patients compared to controls.
TABLE II. THE SECOND-ROUND SELECTION OF miRs THROUGH CO-EXPRESSION ANALYSIS

<table>
<thead>
<tr>
<th>Selected miR</th>
<th>Co-expression Degree</th>
<th>Selected miR</th>
<th>Co-expression Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-miR-106b-5p</td>
<td>86</td>
<td>hsa-miR-5195-3p</td>
<td>71</td>
</tr>
<tr>
<td>hsa-miR-227a-3p</td>
<td>86</td>
<td>hsa-miR-6831-5p</td>
<td>74</td>
</tr>
<tr>
<td>hsa-miR-103a-3p</td>
<td>83</td>
<td>hsa-miR-7106-5p</td>
<td>71</td>
</tr>
<tr>
<td>hsa-miR-27a-3p</td>
<td>83</td>
<td>hsa-miR-4253</td>
<td>70</td>
</tr>
<tr>
<td>hsa-miR-223-3p</td>
<td>83</td>
<td>hsa-miR-4484</td>
<td>70</td>
</tr>
<tr>
<td>hsa-miR-227a-3p</td>
<td>83</td>
<td>hsa-miR-659-3p</td>
<td>70</td>
</tr>
<tr>
<td>hsa-miR-29a-3p</td>
<td>83</td>
<td>hsa-miR-6802-5p</td>
<td>70</td>
</tr>
<tr>
<td>hsa-miR-29c-3p</td>
<td>83</td>
<td>hsa-miR-4496</td>
<td>70</td>
</tr>
<tr>
<td>hsa-miR-425-5p</td>
<td>83</td>
<td>hsa-miR-190-3p</td>
<td>68</td>
</tr>
<tr>
<td>hsa-miR-997-5p</td>
<td>83</td>
<td>hsa-miR-6805-5p</td>
<td>67</td>
</tr>
<tr>
<td>hsa-miR-1249-5p</td>
<td>83</td>
<td>hsa-miR-150-5p</td>
<td>66</td>
</tr>
</tbody>
</table>

miR. Such information is highly valuable and can provide critical clues to further facilitate filtering out more likely miRs as candidate biomarkers.

- Step (4). We will perform real-time quantitative PCR (qPCR) to detect plasma expression levels of miRs and lncRNAs returned from Step (2); we will then conduct receiver operating characteristic (ROC) curve analysis on miR and lncRNA expression, followed by the calculation of the area under the curve (AUC) in the combined ROC of miR and lncRNA. Youden index will be calculated to determine the best boundary values, based on which we will generate a final list of candidate miRs and lncRNAs.

- Step (5). Biological experiments (dual luciferase reporter assay system and western blot) will be conducted to verify biological functions of those miRs and lncRNAs discovered in Step (4). Those biologically validated miRs and/or lncRNAs will be treated as novel biomarkers in DKD.

IV. RESULTS AND DISCUSSION

A. Subject selection

All subjects (informed consent signed) in our experiments were recruited from First People’s Hospital of Yunnan Province. We selected three DKD patients from the Department of Endocrinology and three controls (healthy persons) from the Physical Examination Center. Our selection criteria on DKD patients were: (1) urinary albumin excretion rate > 30 mg/d; (2) urinary albumin/creatinine ratio (ACR) > 30 mg/g; (3) without hypertension, other causes of renal insufficiency, severe hepatic insufficiency, or blood system diseases. Background information as well as routine lab results on all subjects are exhibited in Table I. Note that diabetes mellitus (DM) duration was recorded as either the number of years of DM duration or normal glucose tolerance (NGT).

B. Results from Step (1): microarray test to obtain the first-round screened miRs and lncRNAs

Figure 2 is the result on Affymetrix HTA2.0 chips. We used T test to calculate both the significant difference (P value) and standardized signal multiple fold change value. If the fold...
change was greater than or equal to 1.5 and the P value was less than or equal to 0.05, the corresponding lncRNA was selected. We identified a total of 127 lncRNAs, among which 101 were up-regulated and 26 were down-regulated. In Figure 2, x axis corresponds to \( \log_2 \) (fold change), y axis corresponds to \(-\log_{10} \) (P value), and blue and red dots represent differentially expressed lncRNAs.

Figure 3 is the result on Agilent miR microarray chips. Similarly, we used T test to calculate P value and fold change. If the fold change was greater than or equal to 2.0 and the P value was less than or equal to 0.05, the corresponding miR was selected. We identified a total of 88 miRs, among which 47 were up-regulated and 41 were down-regulated. Similarly to Figure 2, x axis in Figure 3 corresponds to \( \log_2 \) (fold change), y axis corresponds to \(-\log_{10} \) (P value), and blue and red dots represent differentially expressed miRs.

C. Results from Step (2): miR-lncRNA-mRNA co-expression network analysis to obtain the second-round screened candidate miRs and lncRNAs

Figure 4 demonstrates a miR-lncRNA-mRNA co-expression network obtained in our experiments. Based on the analysis on this network, we further reduced the number of candidate biomarkers during our second-round screening procedure — miRs were reduced from 88 to 32, and lncRNAs were reduced from 127 to 68. Information on these miRs and lncRNAs is detailed in Tables II and III, respectively.

D. Results from Step (3): search for miR target mRNAs through OmniSearch user interface

The OmniSearch software tool [30–32] was built upon semantic technologies (including semantic data annotation, semantic data integration, and semantic search) that are based on two domain ontologies, Ontology for MicroRNA Target (OMIT) [20] [21] and Non-Coding RNA Ontology (NCRO) [17] [18]. OmniSearch was developed to handle the significant challenge of miR-related data integration and semantic search query. Three miR target prediction databases (miRDB, TargetScan, and miRanda) and one miR validated target database (miRTarBase) have been integrated in OmniSearch. This software tool provides with users a “one-stop” visit that enables a convenient, side-by-side comparison among prediction results from numerous miR target prediction databases, as well as providing access to other valuable, relevant data sources such as GO annotations and PubMed publications. Altogether, information from OmniSearch has provided important clues, during Step (4), to enhance our ability to filter out likely miR candidates in an effective and efficient manner.

E. Discussion

Our initial experimental results reported in this section were based on a small sample size (3 patients vs. 3 healthy persons as controls), and we have focused on the first three steps in the proposed methodologies (out of five steps as detailed in Section III). As a proof of concept, we have successfully demonstrated the innovation and feasibility of our methods: we were able to identify significantly differentially expressed miRs and lncRNAs between DKD patients and healthy persons; next, using miR-lncRNA-mRNA co-expression network analysis, we successfully reduced the number of candidate miRs and lncRNAs by 64% and 46%, respectively; moreover, we retrieved adequate amount of relevant information on candidate miRs including their computationally putative target mRNAs, GO annotations, and related PubMed publications — during Step (4) in our methods, all such information can provide clues to help further reduce the total number of putative biomarkers, which will result in more precise and efficient discovery of novel biomarkers in DKD disease process.

Besides completing Steps (4) and (5) on subjects currently recruited in this research, we also plan to significantly increase the sample size by recruiting a total of 126 type 2 diabetes patients (ages ranging from 30 to 79, in-patients or out-patients from First Affiliated Hospital of Kunming Medical University and First People’s Hospital of Yunnan Province), along with a total of 40 healthy persons. In addition, it is expected that upon completion of Step (4), the total number of candidate miRs and lncRNAs will be further reduced by at least 80%, resulting in around three miRs and five lncRNAs as final putative biomarkers to be biologically verified in Step (5).

V. Conclusions

DKD is a disease causing major health problems worldwide. Not only there are no specific clinical manifestations during the early stage of DKD, but also more seriously, currently there exist no effective methods to prevent DKD patients to develop ESRD. It is necessary to discover novel biomarkers to facilitate the early diagnosis and effective treatment in DKD patients so that to prevent them to develop ESRD. Towards this end, we need to better understand various regulation mechanisms at genetic level in DKD. In this work-in-progress paper, innovative methodologies are presented to integrate biological, statistics, and computational approaches.
to investigate miR-lncRNA-mRNA regulations in DKD. Our initial experimental results have successfully identified a set of candidate miRs and lncRNAs that may perform important roles in DKD disease process, thus facilitating our understanding of genetic regulation mechanisms on the pathogenesis and progression of DKD.

Following the promising proof of concept exhibited in this paper, we will continue to complete the last two steps in our proposed methodologies on our currently recruited subjects. We also plan to significantly increase our sample size by including a much larger number of DKD patients in the future.

ACKNOWLEDGMENT

Research in this paper was partially supported by: Yunnan Health and Family Planning Commission (2016NS019); National Natural Science Foundation of China (Regional Science Fund Project 81560126); Yunnan Applied Basic Research Projects (2016FB127); Chengdu Medical College (CYZ15-08); and National Cancer Institute at the U.S. National Institutes of Health (U01CA180982). The views contained in this paper are solely the responsibility of the authors and do not represent the official views, either expressed or implied, of these funding agencies or the China and U.S. Governments.

REFERENCES


