Ontology-based Modeling, Representation, and Analysis of Biomarkers in Healthy and Disease Kidney Tissue

Yingtong Liu[1,ξ], Wenjun Ju[1], Becky Steck[1], Sanjay Jain[2], Matthias Kretzler[1], and Yongqun He[1], for the Kidney Precision Medicine Project

[1] *University of Michigan Medical School, Ann Arbor, MI 48109, USA*
[2] *Washington University School of Medicine, St. Louis, MO 63110, USA*
[ξ] *Corresponding author*

Abstract

Biomarker reflects an underlying biological state or identity. Extensive kidney studies have resulted in the discovery of many kidney cell and disease biomarkers. Our manual literature annotations have identified 150 cell-specific markers for 73 kidney cell types and 38 diabetic kidney disease (DKD)-related biomarkers. To systematically study these biomarkers, we first surveyed and ontologically defined the term biomarker and different types of biomarkers. The Kidney Tissue Atlas Ontology (KTAO) has been further used as a platform to model and represent these kidney biomarkers by including the biomarker gene name, cell type, disease, and axioms linking the biomarkers and other terms. Gene Ontology (GO) analysis revealed 9 shared enriched GO terms in both biomarker sets. A DL-query was performed to demonstrate the advantages of ontology-based modeling and analysis of kidney biomarkers.

Keywords

Kidney biomarker, KTAO, DKD, ontology

1. Introduction

Molecular biomarkers, which are molecules indicating biological identities or states, are crucial in further understanding of kidney diseases and support the precision kidney medicine. Extensive research has found various kidney biomarkers associated with various kidney states, either healthy or diseased. With various kidney biomarkers identified, it is critical to systematically annotate, standardize, represent, integrate, and analyze these biomarkers and their associated features and conditions.

Ontology is an ideal tool for such study. Basically, an ontology is a human- and computer-interpretable representation of the entity types, entity properties, and their interrelationships that exist in a particular domain [2]. Ontologies have emerged to become an important platform for systematical data and knowledge representation, integration, sharing, and computer-assisted reasoning and analysis.

The Kidney Tissue Atlas Ontology (KTAO) is a community-based open-source biomedical ontology that systematically represents entities associated with kidney disease, kidney structures, cells, genes etc. [1]. KTAO is primarily developed by the NIH-supported Kidney Precision Medicine Project (KPMP, http://kpmp.org), an NIH-funded multi-year project aimed to understand and find ways to treat the chronic kidney disease (CKD) and acute kidney injury (AKI). Particularly, diabetic kidney disease (DKD), a subtype of CKD, occurs in people with diabetes.

In this study, we systematically annotated around 180 of biomarkers, mainly gene markers, from the literature, the HuBMAP and KPMP consortia. These biomarkers are associated with kidney healthy cells or diseased kidney. We have used KTAO to ontologically classify these biomarkers, their features, and the relations among these markers and their association with various kidney cell types, biological processes, and kidney diseases.

International Conference on Biomedical Ontologies 2021, September 16–18, 2021, Bozen-Bolzano, Italy
EMAIL: yingtliu@umich.edu (A. 1); wenjunj@med.umich.edu (A. 2); roesch@med.umich.edu (A. 3); sanjayjain@wustl.edu (A. 4); kretzler@med.umich.edu (A. 5); yongqunh@med.umich.edu (A. 6)
© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3073/paper8.pdf

2. Methods

2.1. Data Collection and Annotations

Two types of kidney biomarkers, normal kidney cell lineage biomarkers and diabetic kidney disease (DKD) were compiled from different resources. The normal cell type lineage biomarkers were obtained from the literature including our recent KPMP publication [2] and The Human BioMolecular Atlas Program (HuBMAP) (https://commonfund.nih.gov/hubmap) reference ASCT+B data [3]. A number of biomarkers were derived from single cell sequencing studies based on statistical cut off's and in some cases automated tools such as those in Seurat package (https://satijalab.org/seurat/) to label the cell types and the most significant markers [4, 5]. DKD protein biomarkers were collected mainly from a review article [6]. The selection criterion for a non-invasive DKD biomarker is its being reported in at least one non-high throughput research article using human patients' samples. Our study focused on the collection and analysis of DKD protein biomarkers. All the results are based on rigorous experiment design and statistical analysis [2]. The manual curation confirms and picks the ones seen in multiple technologies and across species and validations where possible.

2.2. KTAO Biomarker Modeling, Representation and DL Query

Gene and protein markers are modeled. Genes and proteins are presented using the Ontology of Genes and Genomes (OGG) and Protein Ontology (PR). IDs were retrieved by using Ontobee (http://ontobee.org). Biomarker-related relations were defined to generate new axioms that semantically link gene/protein biomarkers and cell types (or diseases). Such axioms were built up using Ontorat [7]. The Description Logic (DL) queries were performed using DL Query plugin of Protégé 5.0 (beta 15, https://protege.stanford.edu/). After reasoning of the KTAO with the ELK reasoned (https://www.cs.ox.ac.uk/isg/tools/ELK/), DL queries were performed on biomarkers.

2.3. GSEA Analysis on Gene Markers and Incorporation to KTAO

For functional analysis of our collected gene markers, a gene set enrichment analysis (GSEA) was performed using the DAVID Bioinformatics Resources (https://david.ncifcrf.gov/tools.jsp). The default DAVID background was applied for analysis of enriched biological processes, cellular components and molecular functions as defined by the Gene Ontology (GO). The p-value ≤ 0.05 after FDR adjustment was used as the cutoff. These GO terms were retrieved by using Ontobee. The axiom relations between gene markers and GO biological processes and molecular functions are defined using the Relation Ontology (RO) [8] terms '*participates in*' and '*has participant*', and the axiom relation between genes and GO cellular components are defined using '*has part*' and '*part of*'. These corresponding axioms in KTAO were established using Ontorat [7].

3. Results

3.1. Kidney Biomarker Collection and Representation

Our kidney biomarker collection focused on the branches: kidney cell lineage biomarkers and DKD biomarkers. Our study found a total of 150 genetic biomarkers of 72 kidney cell lineage types using the method defined in the Methods section. Meanwhile, we collected a total of 38 DKD protein biomarkers. For each of these biomarkers, we have included at least one peer-reviewed publication. Note that our non-exhaustive collections are likely incomplete and have not included those markers identified from high throughput analyses [4, 9]. This focus on this study is primarily on the establishment of ontological modeling and knowledge representation of these biomarkers.

3.2. Ontological Definitions and Representation of Kidney Biomarkers

There have been many definitions of biomarkers. The NIH Biomarkers and Surrogate Endpoint Working Group defined a biomarker as "A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention" [10]. However, the term "characteristic" is vague and difficult to define ontologically. Mayeux defined biomarker as a type of "alterations in the constituents of tissues or body fluids" [11]. Biomarkers should be assessable or measurable. In a 2015 paper [12], Ceuters and Smith proposed three disjoint types of biomarkers: material biomarker, quality biomarker, and process biomarker. As the bearers of various dispositions, material entities can be measured. However, processes and qualities do not have assessable dispositions based on the Basic Formal Ontology (BFO). Therefore, it appears difficult to assess biomarkers as processes and qualities using the BFO framework. It also appears that the 2015 proposal of the three-type biomarker classification has not been adopted in the ontology community.

In our study, we define biomarker as a material entity only, and we do not consider quality or process as a biomarker. Meanwhile, the biomarker material entity has measurable qualities and processes that can be used as the indicator of some biological state. The biomarker material entity has measurable dispositions for a specific state/identity, which is the basis of the material entity being the biomarker. Specifically, we have defined the term 'biomarker' in the Ontology of Precision Medicine and Investigation (OPMI) [1] as:

biomarker = def. A material entity that has a measurable quality or process profile(s), which can be used as an indicator of an underlying biological state or identity.

A biomarker can be classified into different subtypes. Based on the clinical purpose of the biomarkers, there are diagnostic, predictive, prognostic, pharmacodynamic biomarkers, etc. [13]. Based on the types of the material entities, there are gene, protein, RNA, and metabolite biomarkers. Based on the restricted expression in cell types and tissues, there are cell type- and tissue-specific biomarkers. Furthermore, there are various morphological biomarkers.

As an example, decreased renal tubular cell expression and reduced urinary excretion of epidermal growth factor (EGF) has been observed in many human kidney diseases including acute kidney injury and CKD, and the EGF can serve as a biomarker for progression of these kidney diseases [14, 15]. In our definition, the urinary human EGF protein (hEGF) is a prognostic biomarker for kidney disease, and it can be used by measuring its concentration in urinary excretion. This case can be defined ontologically using the following ontological axiom:

'urine hEGF': has_role some ('prognostic biomarker role' and (realized_in some 'DKD process'))

Alternatively, we can make an equivalent axiom with a shortcut relation:

'urine hEGF': 'has prognostic biomarker role in' some 'DKD process'

The ontological hierarchical structure of various relevant biomarkers is shown in Figure 1A. The high level 'biomarker' term and several of its subclasses (including disease biomarker, immune biomarker and cell biomarker) were imported from the OPMI. We further defined many kidney specific biomarkers under 'kidney biomarker' in KTAO. Under 'kidney disease biomarker' are 'AKI biomarker' and 'CKD biomarker', which includes DKD biomarkers. Under 'kidney cell lineage biomarker', there are substructures of the kidney, specific cell types and different types of cell biomarkers. For example, kidney podocyte biomarker is a biomarker to the kidney podocyte cell lineage (Figure 1A). After ontology inferencing using a semantic reasoner, 4 biomarkers, NPHS1, NPHS2, PODXL and PTPRQ, were inferred to be the kidney podocyte biomarkers (Figure 1B).

Biomarker-related relations (or called object properties) were generated for new ontology axiom generations. For example, the relation '*is gene marker of cell*' semantically links a gene marker and a cell type, and the relation '*protein biomarker of disease*' links a protein biomarker with a disease. Existing relations in the RO [8] were also used. For example, all enriched GO terms associated with gene makers are connected by '*participates in*' defined in RO. For instance, NPHS1 is a kidney cell marker for podocytes (glomerular visceral epithelial cell) and a kidney disease protein marker for diabetic nephropathy. This marker is also a component of several GO-defined cellular components such as extracellular exome (Figure 1C). Each cellular component GO terms related to gene markers are also well-defined and has axiom '*has part*' to trace back to gene markers (Figure 1D). All the relations

between biomarkers with disease, with cell type and with GO terms are incorporated into KTAO. In total, there are 37 new relations between biomarkers and DKD, 107 between biomarkers with cell types, and over hundreds of new relations between biomarkers and GO terms.

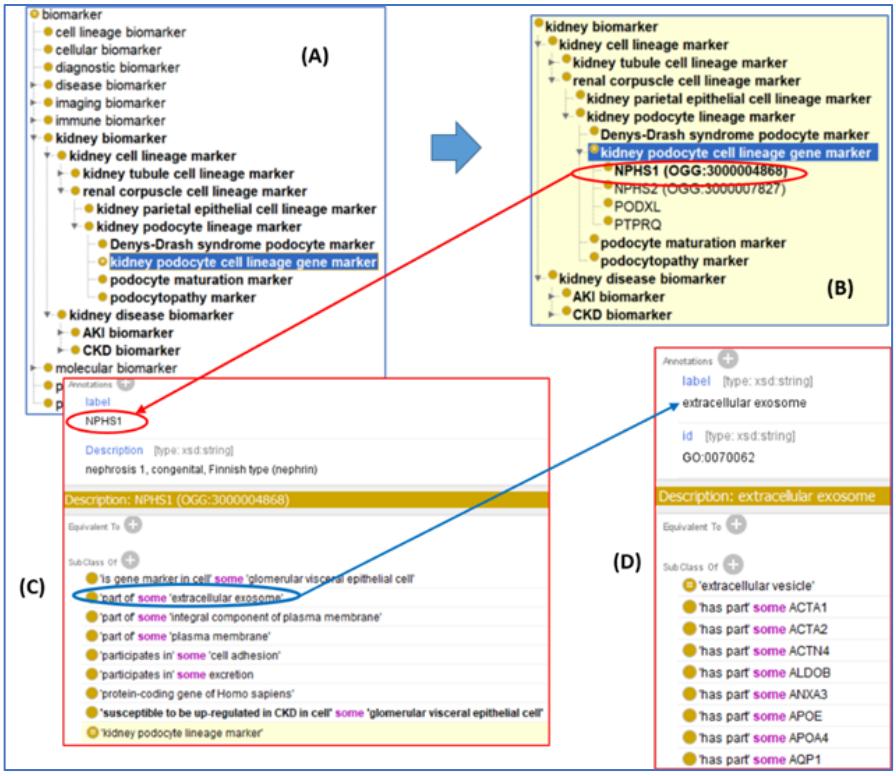

Figure 1. Ontological hierarchy and representation of biomarker axioms in KTAO. A) Hierarchical structure of different kinds of biomarkers in KTAO. B) After reasoning, the markers that belong to this classification of biomarkers. C) Example of biomarker representation. D) Example of associated GO terms representation.

3.3. KTAO-based Analysis and Query of Kidney Biomarkers

Based on our DAVID GO term enrichment analysis, 61 GO terms were enriched in kidney cell biomarkers, and 41 GO terms were enriched for DKD protein markers. Nine GO terms, including 5 cellular component (CC) terms and 4 biological process (BP) terms, were found to be shared between these cell biomarkers and DKD biomarkers (Table 1). Our analysis identified large numbers of gene markers located in extracellular space, extracellular region, extracellular exosome, and cell surface (Table 1), indicating that the gene markers for both cell types and diabetic kidney diseases are more likely to involve in extracellular and cell surface structures.

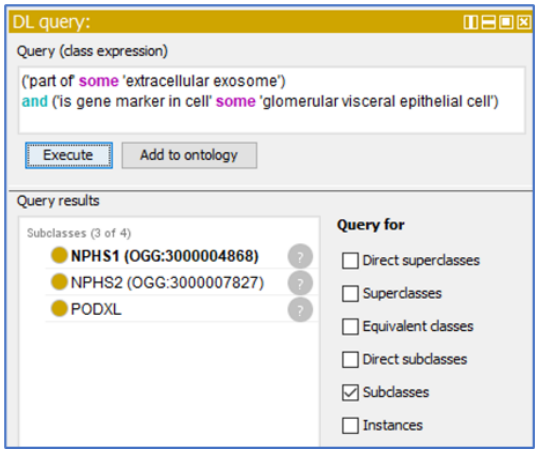

Figure 2. Example of DL-query. This query identifies 3 biomarkers that are podocyte biomarkers, and also part of extracellular exosome. The DL-query was performed in Protégé-OWL editor. ELK reasoned was used to perform the reasoning before the query was conducted.

The knowledge represented in KTAO can be interpretable by computers and reasoned with methods such as DL-query or SPARQL query. For example, we performed a DL query on KTAO to identify those biomarkers that are the biomarkers for podocytes (i.e., glomerular visceral epithelial cells), and meanwhile they are located in the extracellular exosome (Figure 2). Three axioms were queried together in this query example. Genes NPHS1, NPHS2, and PODXL were identified (Figure 2).

Table 1 Common Gene Ontology terms for gene or protein markers of kidney cell types and DKD

GO ID	Term Label	Type	# cell marker	# disease marker	# of common genes
GO:0005576	extracellular region	CC	34	29	2
GO:0005615	extracellular space	CC	37	27	2
GO:0070062	extracellular exosome	CC	56	18	3
GO:0006955	immune response	BP	17	8	1
GO:0005578	proteinaceous extracellular matrix	CC	9	6	1
GO:0006954	inflammatory response	BP	13	7	0
GO:0001666	response to hypoxia	BP	8	7	1
GO:0071356	cellular response to tumor necrosis factor	BP	6	5	0
GO:0009986	cell surface	CC	21	7	1

4. Discussion

In this paper, we systematically collected and annotated various biomarkers for regular kidney cells as well as for DKD, and ontologically represented these kidney biomarkers in KTAO. Such ontological representation provides standardized computer-interpretable knowledge representation and supports automated semantic queries and data analyses of kidney biomarkers. And the ontological representation can serve as a knowledge base and be used to compare and analyze the results from high throughput omics studies, supporting kidney diseases diagnosis, mechanisms study and rational treatment design. In addition, the KTAO representation can be incorporated into our KPMP web application development for more advanced browsing, reasoning, and data analysis.

The systematical and logical KTAO representation of the biomarker types, biomarker locations, and their involving biological processes supports integrative analysis of the biomarkers. We can systematically find those biomarker-associated cellular components or biological processes and inspect potential mechanisms of actions of these biomarkers in normal kidney functions or disease formation. For example, NPHS1 is a gene marker for DKD, and it is a kidney podocyte-specific marker, and it can

be found in extracellular exosomes. Detection of NPHS1 in podocytes may be a potential sign of DKD, and we can also infer that NPHS1 play important role in cell junctions. Some of the disease markers may also be potential drug targets for DKD treatment. To facilitate such analysis, we may develop new models and representation of the drug-biomarker associations in KTAO.

Admittedly, this work has limitation and further study is going on. Due to the time restriction, only a small number of papers were reviewed and used. Further literature mining is necessary to identify and annotate more kidney biomarkers. Our future work includes the expansion of the kidney biomarker presentation to include more inclusive biomarkers for DKD, and additionally for other types of CKD and AKI. We will also include other types of biomarkers such as RNA and metabolite markers. We welcome researchers interested in the topic to participate in the community-based KTAO development and its applications.

5. Conclusion

Kidney biomarkers are critical to the fundamental study of kidney functions and disease development. In this study, we started with 150 cell gene markers and 38 DKD protein markers, analyzed and incorporated them into the KTAO. Nine enriched Gene Ontology terms were identified in these two groups of biomarkers. These DKD gene markers and their associations with different kidney cell types, cellular components and biological processes were systematically represented in KTAO. A DL-query demonstrated the KTAO support for computer-assisted reasoning and query. Overall, our ontological knowledge representation facilitates systematic kidney biomarker standardization, integration, and analysis.

6. Acknowledgements

This project is supported by the following KPMP grants from the NIDDK: U2C DK114886, UH3DK114861, UH3DK114866, UH3DK114870, UH3DK114908, UH3DK114915, UH3DK114926, UH3DK114907, UH3DK114920, UH3DK114923, UH3DK114933, and UH3DK114937, and HuBMAP consortium grant U54HL145608 funded by the NIH. This work is also partially supported by George M. O'Brien Michigan Kidney Translational Core Center, funded by NIH/NIDDK grant 2P30-DK-081943, and The Nephrotic Syndrome Rare Disease Clinical Research Network (NEPTUNE). NEPTUNE is part of the Rare Diseases Clinical Research Network (RDCRN), which is funded by the National Institutes of Health (NIH) and led by the National Center for Advancing Translational Sciences (NCATS) through its Office of Rare Diseases Research (ORDR). NEPTUNE is funded under grant number U54DK083912 as a collaboration between NCATS and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Additional funding and/or programmatic support for this project has also been provided by the University of Michigan, NephCure Kidney International and the Halpin Foundation. All RDCRN consortia are supported by the network's Data Management and Coordinating Center (DMCC) (U2CTR002818). Funding support for the DMCC is provided by NCATS and the National Institute of Neurological Disorders and Stroke (NINDS).

7. References

[1] E. Ong, L. L. Wang, J. Schaub, J. F. O'Toole, B. Steck, A. Z. Rosenberg, et al., Modelling kidney disease using ontology: insights from the Kidney Precision Medicine Project, Nat Rev Nephrol, Sep 16 2020.

[2] T. M. El-Achkar, M. T. Eadon, R. Menon, B. B. Lake, T. K. Sigdel, T. Alexandrov, et al., A multimodal and integrated approach to interrogate human kidney biopsies with rigor and reproducibility: guidelines from the Kidney Precision Medicine Project, Physiol Genomics, vol. 53, pp. 1-11, Jan 1 2021.

[3] K. Borner, S. A. Teichmann, E. M. Quardokus, J. Gee, K. Browne, D. Osumi-Sutherland, et al., Anatomical Structures, Cell Types, and Biomarkers Tables Plus 3D Reference Organs in Support of a Human Reference Atlas, bioRxiv, 2021.

[4] B. B. Lake, S. Chen, M. Hoshi, N. Plongthongkum, D. Salamon, A. Knoten, et al., A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys, Nat Commun, vol. 10, p. 2832, Jun 27 2019.
[5] B. B. Lake, R. Menon, S. Winfree, Q. Hu, R. M. Ferreira, K. Kalhor, et al., An atlas of healthy and injured cell states and niches in the human kidney, bioRxiv, 2021.
[6] F. C. Brosius, W. Ju, The Promise of Systems Biology for Diabetic Kidney Disease, Adv Chronic Kidney Dis, vol. 25, pp. 202-213, Mar 2018.
[7] Z. Xiang, J. Zheng, Y. Lin, Y. He, Ontorat: Automatic generation of new ontology terms, annotations, and axioms based on ontology design patterns, Journal of Biomedical Semantics, vol. 6, p. 4 (10 pages), July 24-27 2015.
[8] B. Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, et al., Relations in biomedical ontologies, Genome Biol, vol. 6, p. R46, 2005.
[9] S. Mulder, H. Hamidi, M. Kretzler, W. Ju, An integrative systems biology approach for precision medicine in diabetic kidney disease, Diabetes Obes Metab, vol. 20 Suppl 3, pp. 6-13, Oct 2018.
[10] Biomarkers Definitions Working Group, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework, Clin Pharmacol Ther, vol. 69, pp. 89-95, Mar 2001.
[11] R. Mayeux, Biomarkers: potential uses and limitations, NeuroRx, vol. 1, pp. 182-8, Apr 2004.
[12] W. Ceusters, B. Smith, Biomarkers in the ontology for general medical science, in Studies in Health Technology and Informatics 2015, pp. 155-159.
[13] L. M. Millner and L. N. Strotman, The Future of Precision Medicine in Oncology, Clin Lab Med, vol. 36, pp. 557-73, Sep 2016.
[14] D. S. Gipson, H. Trachtman, A. Waldo, K. L. Gibson, S. Eddy, K. M. Dell, et al., Urinary Epidermal Growth Factor as a Marker of Disease Progression in Children With Nephrotic Syndrome, Kidney Int Rep, vol. 5, pp. 414-425, Apr 2020.
[15] Y. Isaka, Epidermal growth factor as a prognostic biomarker in chronic kidney diseases, Ann Transl Med, vol. 4, p. S62, Oct 2016.