Molecular Taxonomy of Urinary Diseases through Unsupervised Data-Driven Classification Based on 7000-plex Urinary Proteome

Vinaika Maruvada, Shri Chada, Yewei Ma, Kamala Vanarsa, Ramesh Saxena, Chandra Mohan
Department of Biomedical Engineering, University of Houston, Houston, TX

Background

Introduction:
Lupus Nephritis (LN) is a severe manifestation of Systemic Lupus Erythematosus (SLE) that affects approximately 40% of patients diagnosed with SLE. The heterogeneity of LN presents a significant clinical challenge, leading to variable treatment responses and patient outcomes. Current classifications, based on renal biopsy histology, are subjective and limited in their ability to capture the full molecular complexity of the disease. Addressing this heterogeneity through objective molecular stratification is crucial for advancing Personalized Medicine.

Overall Objective:
To establish a novel, data-driven molecular taxonomy of urinary diseases, including Lupus Nephritis, at the proteome level using comprehensive targeted proteomic screening and unsupervised machine learning.

Specific Aim:
- To identify inherent proteomic endotypes (patient groupings) and defining protein modules (co-expressed protein groups) from 7000-plex urine proteomic data derived from a diverse patient cohort.
- To functionally characterize these protein modules and correlate them with clinical and demographic variables.

Conclusion

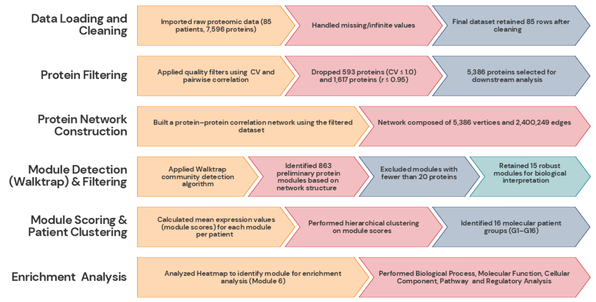
- This study demonstrates that unsupervised proteomic analysis of urine can successfully stratify patients into distinct molecular endotypes (Groups), revealing disease-specific signatures that are independent of their initial clinical diagnosis.
- In this work, Modules are defined as clusters of highly correlated proteins that represent a single functional unit. Groups (G1-G16) are defined as clusters of patients who share similar expression profiles across these modules.
- Hierarchical clustering successfully identified several clinically relevant endotypes. The largest patient clusters were G3 (n=21), G5 (n=20), G6 (n=11), and G4 (n=9).
- The clustering effectively isolated specific disease states:
  - Group G5 was strongly enriched for Bladder Cancer (BC) and Urological Control (UC) patients, who were molecularly indistinguishable and defined by a "low-expression" signature across most modules.
  - Group G4 was predominantly composed of Diabetic Chronic Kidney Disease (DCKD) patients (67%), identifying a specific molecular profile for this condition.
  - In contrast, Lupus Nephritis (LN) patients were highly heterogeneous and were distributed across 11 different endotypes (e.g., G2, G8-G11, G13-G15), suggesting multiple distinct molecular pathologies underlie the same clinical disease. The large, mixed groups G3 and G6 also contained a mix of LN, controls, and other disease states.
  - Analysis of Module 6, the most statistically significant module for separating the original clinical groups, revealed an enrichment for proteins involved in ciliary biology, centrosome organization, and intraflagellar transport. This finding suggests a novel link between ciliary dysfunction and the pathology of Lupus Nephritis.

Future Work

- Correlate module scores and patient endotypes (Groups) with key clinical and histological data (e.g., eGFR, SLEDAI, AI/CI scores) to establish their utility as biomarkers for disease severity or progression.
- Validate the upregulation of key Module 6 (ciliary) proteins in lupus nephritis (LN) patient kidney biopsies to confirm the link to ciliary dysfunction at the tissue level.
- Assess the robustness of the 15 modules and 16 endotypes using alternative algorithms

Methods

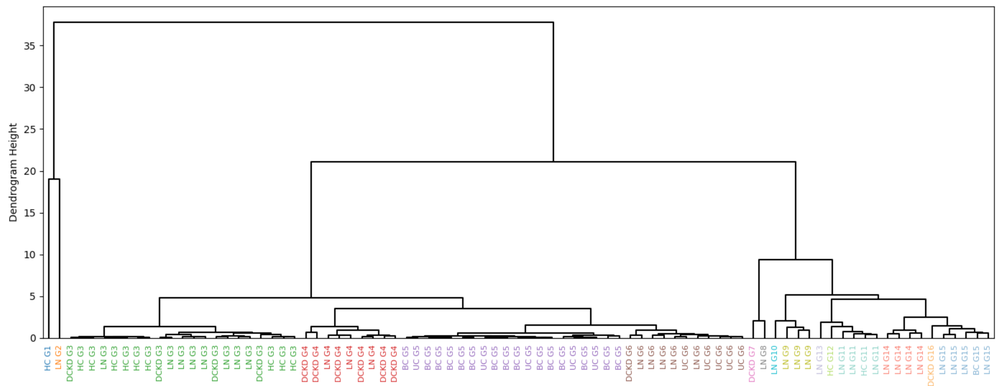
Figure 1: Urine Proteomic Analysis and Patient Stratification Workflow



This diagram illustrates the computational workflow used in this study. Urine samples from 85 individuals including healthy controls (HC), lupus nephritis (LN), diabetic chronic kidney disease (DCKD), urology clinic controls (UC, presenting with hematuria), and bladder cancer (BC) were profiled using a 7000-plex aptamer-based proteomic platform. After preprocessing, 5,386 proteins were used to construct a protein-protein correlation network. The Walktrap community detection algorithm identified 15 robust protein modules. Patient-specific module activity scores were then calculated. Hierarchical clustering using Ward's method was applied to these module profiles to stratify patients into molecular groups, or endotypes (see Figure 2), independent of clinical diagnosis. To understand the biological function of key modules (e.g., Module 6), a multi-level enrichment analysis was performed (see Figure 4). Protein-protein interaction (PPI) networks were also constructed using Cytoscape (see Figure 5) to identify core protein hubs and potential regulatory molecules.

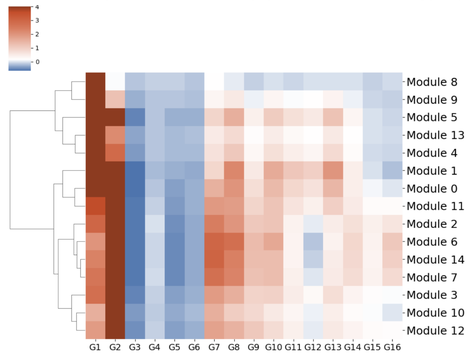
Results

Figure 2: Hierarchical Clustering of Urine Proteomic Profiles Identifies 16 Distinct Patient Endotypes (Groups G1-G16)



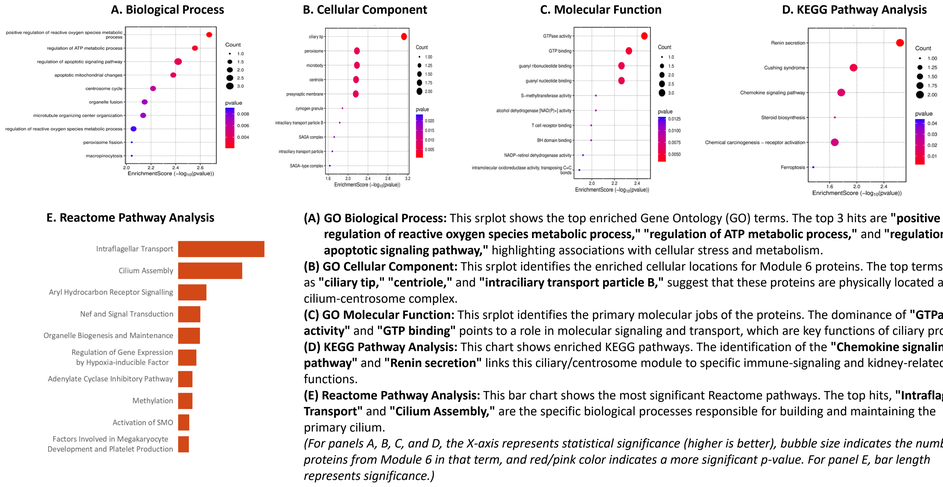
This figure displays the agglomerative hierarchical clustering dendrogram for all 85 patients, generated using the expression scores of the 15 protein modules.
X-axis (Patient & Group Labels): Each leaf on the dendrogram represents a single patient, labeled by their original clinical diagnosis (e.g., HC, BC, ALN). The colored bar below the dendrogram indicates the new molecular endotype (G1-G16) assigned to each patient.
Y-axis (Dendrogram Height): Represents the dissimilarity between clusters. Taller branches indicate that the joined clusters or patients are more different from each other.

Figure 3: Module Expression Heatmap per Patient Endotype (G1-G16)



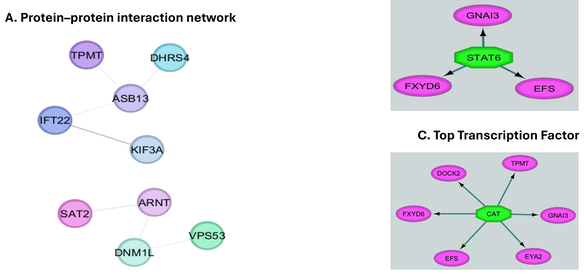
This heatmap visualizes the molecular signature of the 16 patient endotypes identified in **Figure 2**. The color of each cell represents the average module expression (z-score) for all patients within a given endotype.
Color Scale (Average Module Score): Represents the average expression of each module for all patients within a molecular group. Red indicates high average expression (upregulated), blue indicates low average expression (downregulated), and white represents a neutral baseline (z-score of 0).
X-axis (G1-G16): The 16 distinct molecular patient subgroups (endotypes) identified by the clustering.
Y-axis: The 15 robust protein modules.
Left Dendrogram: Shows clustering of modules that have similar expression patterns across the 16 patient groups.

Figure 4: Multi-Panel Functional Enrichment Analysis of Module 6



(A) GO Biological Process: This srplot shows the top enriched Gene Ontology (GO) terms. The top 3 hits are **"positive regulation of reactive oxygen species metabolic process," "regulation of ATP metabolic process,"** and **"regulation of apoptotic signaling pathway,"** highlighting associations with cellular stress and metabolism.
(B) GO Cellular Component: This srplot identifies the enriched cellular locations for Module 6 proteins. The top terms, such as **"ciliary tip," "centriole,"** and **"intraciliary transport particle B,"** suggest that these proteins are physically located at the cilium-centrosome complex.
(C) GO Molecular Function: This srplot identifies the primary molecular jobs of the proteins. The dominance of **"GTPase activity"** and **"GTP binding"** points to a role in molecular signaling and transport, which are key functions of ciliary proteins.
(D) KEGG Pathway Analysis: This chart shows enriched KEGG pathways. The identification of the **"Chemokine signaling pathway"** and **"Renin secretion"** links this ciliary/centrosome module to specific immune-signaling and kidney-related functions.
(E) Reactome Pathway Analysis: This bar chart shows the most significant Reactome pathways. The top hits, **"Intraflagellar Transport"** and **"Cilium Assembly,"** are the specific biological processes responsible for building and maintaining the primary cilium.
(For panels A, B, C, and D, the X-axis represents statistical significance (higher is better), bubble size indicates the number of proteins from Module 6 in that term, and red/pink color indicates a more significant p-value. For panel E, bar length represents significance.)

Figure 5: Module 6 Protein-Protein Interaction (PPI) and Regulatory Network



This figure visualizes the known connections between the 36 proteins in Module 6 using Cytoscape.
A) (Protein–protein interaction (PPI) network: Constructed from the 36 proteins that belonged to module 6, using the Cytoscape stringApp plug-in. Edge darkness corresponds to the strength of the correlation between protein nodes.
(B–C) Network visualization: Generated using Cytoscape. The top signaling molecule was **STAT6**, regulating 3 proteins, and the top transcription factor identified was **CAT**, regulating 6 proteins, from the total input set of 36 proteins.

References

1. Banchereau R, Hong S, Cantarel B, et al. Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients [published correction appears in Cell. 2016 Jun 2;165(6):1548-1550]. Cell. 2016;165(3):551-565. doi:10.1016/j.cell.2016.03.008
2. George BJ, Reif DM, Gallagher JE, Williams-DeVane CR, Heidenfelder BL, Hudgens EE, Jones W, Neas L, Hubal EA, Edwards SW. Data-driven asthma endotypes defined from blood biomarker and gene expression data. PLoS One. 2015 Feb 2;10(2):e0117445. doi: 10.1371/journal.pone.0117445. PMID: 25643280; PMCID: PMC4314082.
3. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun 9, 1090 (2018). https://doi.org/10.1038/s41467-018-03424-4

Acknowledgements

This research was made possible by generous funding and support from the National Institutes of Health (NIH) through grants #AR074096 and R01DK134055. We also gratefully acknowledge the support of the Lupus Research Alliance. For questions or correspondence, please contact: Chandra Mohan: cmohan@Central.UH.EDU | Vinaika Maruvada: vmaruvad@CougarNet.UH.EDU | Shri Chada: shrila777@gmail.com