Identifying Genetic Susceptibility Factors for Posterior Urethral Valves (PUV)

Yask Gupta[1], Tze Y Lim[1], Jonathan Diah[1], Gina Jin[1], Juntao Ke[1], Martino Jeremiah[1], Cathy L Mendelsohn[2], David B Goldstein[3], Simone Sanna-Cherchi[1]

[1]Division of Nephrology, Department of Medicine, Columbia University, New York, New York. [2]Department of Urology, Columbia University, New York, NY. [3]Institute for Genomic Medicine, Columbia University, New York, New York.

Background

Posterior urethral valves (PUV) indicate the presence of congenital obstruction at the level of the prostatic urethra (hence present only in males), and it is the most common congenital cause of lower urinary tract obstruction in males. Although affecting only 1 in 4,000 to 7,500 infants, PUV is responsible for up to 17% of pediatric end-stage kidney failure despite early surgical intervention. The role of genetic factors underlying PUV is unclear: the increased prevalence in patients with Down syndrome suggests that gene dosage might be a critical factor in the etiology of PUV, while the familial occurrence in rare cases suggests that single-gene defects can underlie PUV. Nevertheless, no mutations associated to isolated PUV have been identified so far, providing a unique opportunity for gene discovery.

Cohort Selection & Methods

- We performed a whole genome and exome sequencing study in 73 PUV trios (219 samples) and 17,865 controls. We then conducted two main gene discovery analyses: a) a case control analysis for enrichment of rare deleterious variants (exome-wide collapsing analysis); and b) a trios based analysis testing global and per-gene enrichment of ultra-rare de novo mutations.
- **Exome-wide collapsing analysis** (a.k.a. burden test): we used the 73 PUV index cases and compared them to 3,987 genetically-matched male controls (from 17,865 controls. We performed Principal Component Analysis (PCA) for dimensionality reduction on a set of predefined variants to capture population structure (Fig. 1a). Using the first 6 principal components (PCs) as input, we applied the Louvain method of community detection for identifying clusters within the data that reflect the ethnicity of the samples (Fig. 1b). We then selected high quality rare qualifying variants (QV) to perform collapsing analysis within each cluster. Afterwards, we extracted the number of cases/controls and use the Cochran-Mantel-Haenszel (CMH) test to assess the association between case/control status and QV status while controlling for cluster membership.

PCA and Louvain Clustering of 73 cases and 17,865 controls

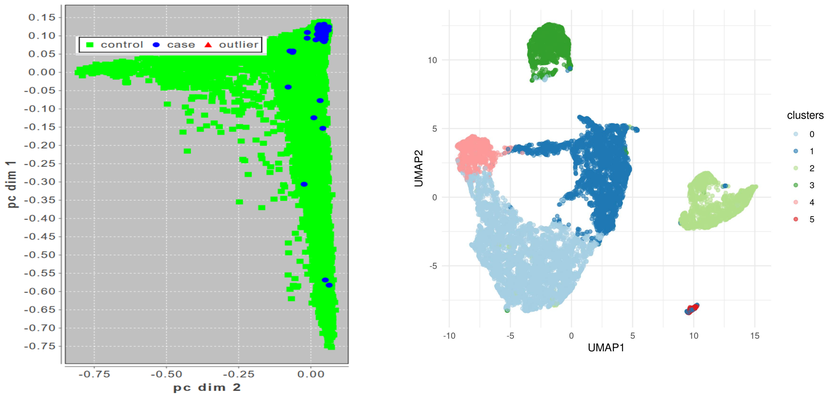

Fig. 1. a) PCA showing distribution of cases and controls across all ancestries. **b)** UMAP clustering of our PUV cohort showing improved resolution of population ancestries

- **Trios analysis.** Here we used WGS for all 219 individuals from the 73 PUV trios. Raw FastQ files were aligned to human genome (hg19) using Dragen software and processed for *de novo* variant identification using IGM ATAV workflow. After annotation, variants were filtered out if they were mapped to duplicated, low complexity and decoy regions of human genome, or when they were observed 3 times or more in our PUV cohort, or have allele frequency more than 0.001% in 1000 genome, EXAC or in Gnomad databases. We chose variants with highest impact as provided by ENSEMBL VEP to select de novo variants per gene per proband for downstream analysis. These variants were further annotated if they mapped in constrained coding regions CCR regions. Gene level annotation for sHET values (heterozygous loss of gene function) were performed by comparing *de novo* containing genes in our cohort to previously established genes sHET values in cohort of 60,706 individuals. Finally, we used denovoanalyzer and denovoWest for downstream analyses for genome-wide and per-gene burden of *de novo* variants in PUV trios.

Results

Exome based collapsing (a.k.a. burden) analysis identifies three potential novel candidate genes for PUV

- Our clustering and harmonization methods allowed precise case-control matching and mitigation of sequencing platform heterogeneity as demonstrated by lambda inflation factor <1.1 in all models and similar distribution of synonymous variants in cases and controls (**Fig. 2**).
- Using exome-based collapsing methods, though we did not observe any study-wise significant association ($P<1x10^{-7}$; after correction for multiple models tested; **Fig 3.**). We detected one near-genome-wide significant signal *PLCD1* ($P=2.01x10^{-6}$; OR=162) and two suggestive for *DNAH11* ($P=2.89x10^{-5}$; OR=11) and *TLK2* ($P=5.99x10^{-5}$; OR=108) as shown in T**ab 1**.
- We next cross-referenced these finding with a mouse single-cell RNA-sequencing study on urethra, publicly available from Dr. Strand lab, and found that mouse orthologues Plcd1 was expressed in urethral luminal epithelial and Tlk2 in all urethral epithelial cells.

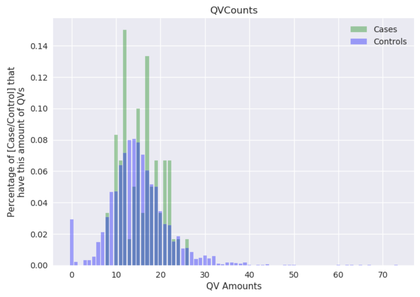

Fig. 2 Histogram plot showing pre-individual number of synonymous variants in the largest cluster. The nearly exact distribution of variants in cases (green) and controls (blue) indicate good matching for downstream analyses.

Class	Obs	Exp	Enrichment	P
Syn	13	13.7	0.95	0.6
Mis	32	30.9	1.04	0.4
Lof	17	4.3	3.97	2.8×10^{-6}
Prot	49	35.1	1.39	0.015
All	62	48.9	1.27	0.04

Tab. 2 Summary results form the exome-wide burden of de novo variants using denovoanalyzer. We observed a significant enrichment of de novo loss-of-function mutations, while no difference in de novo synonymous variants in our cases as compared to expected, suggesting a role of de novo LOF in the pathogenesis of PUV.

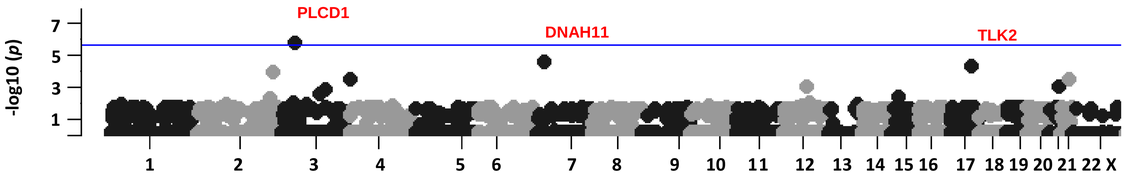

Fig. 3 Manhattan plot of the CMH results from the ultra-rare model in the collapsing analysis. Each dot corresponds to a gene, reported in their genetic coordinates along the 23 chromosomes (x-axis). The blue line indicates the standard exome-wide significance threshold of 2.5×10^{-6}.

Gene	Model	OR (95%CI)	P-value	QV Cases	QV Controls
PLCD1	Dominant Ultra-Rare	162 (12.9 - 8033.9)	2.01×10^{-6}	3/72	1/3893
DNAH11	Dominant Ultra-Rare	11 (2.6 - 34.2)	2.89×10^{-5}	4/71	20/3967
TLK2	Dominant Ultra-Rare	108 (5.6 – 6072.9)	5.99×10^{-5}	2/73	1/3896

Tab. 1 Suggestive candidate genes from the exome-wide collapsing analysis. OR=odds ratio; QV=qualifying variant.

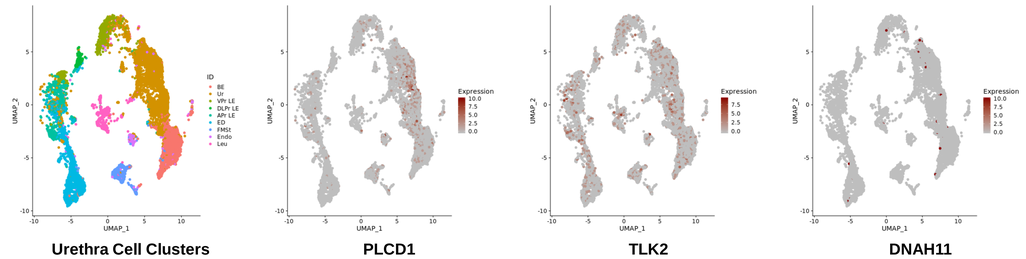

Fig. 4 The single cell RNA-seq (Strand Lab[6]) shows expression of *Plcd1*, *Tlk2* and *Dnah11* in mouse urethra cells. The umap plot shows data from single cell RNA-seq from adult mouse (10-12 weeks). The clusters are identified as BE, basal epithelia; Ur, urethral luminal epithelial; VPr LE, ventral prostate luminal epithelia; DLPr LE, dorsolateral prostate luminal epithelia; APr LE, anterior prostate luminal epithelia; ED, ejaculatory duct; FMSt, Fibromuscular stroma; Endo, Endothelia; Leu, leukocytes.

Trios Based analysis

- In the trios analysis we first explored the exome-wide burden of different classes of genetic variations in our cases, as compared to an expected null distribution based on XXX trios, using the software denovoanalyzer. We observed a ~4-fold increase in burden of *de novo* LOF ($P=2.8 \times 10^{-6}$), while no excess was found for either synonymous or missense variants (P>0.1).
- We next used simulation-based statistical tests (as implemented in deNovoWest software) to explore the per-gene excess of de novo variants. While we did not observe excess of synonymous DNMs for any gene as a proof of concept, we did detect suggestive excess burden ($P<1x10^{-5}$) for damaging de novo variants for the following genes: *FOXP2*, *MSLN*, *COL23A1*, *DRD4*, *SKA3*, *KRT77*.

Conclusions, Limitations, and Future Directions

- Here we provide the initial results and the optimization of the workflow for gene identification in PUV using trios-based and case-control analyses from whole-genome sequencing data.
- By analyzing the coding regions only of the genome (exomes), we report on several suggestive candidate genes for PUV that will need to be validate by genetic and functional studies.
- We have now expanded our whole genome sequencing study to over 150 PUV trios (more than double than the reported here) and acquired WGS data from >700 unaffected trios from the Simons collection.
- Comprehensive studies of both the coding and non coding genome in this extended cohort will maximize our power for gene identification for PUV.

References

1. Sanna-Cherchi *et al.* J Clin Invest. 2018, 128(1):4-15.
2. Verbitsky *et al.* Nat Genet. 2019, 51:117–127.
3. Sanna-Cherchi *et al.*, Am J Hum Genet. 2017, 101(5): 789–802
4. Povysil G. *et al.* Nat Rev Genet 2019, Dec;20(12):747-759
5. Ren Z. *et al.* BMC Bioinformatics 2021, Mar 23;22(1):149
6. Joseph D.B. et al. Prostate 2020, Aug;80(11):872-884.
7. Kaplanis J. et al. Nature. 2020 Oct 1; 586(7831): 757–762.

Funding:
NIDDK P20 DK116191,
RO1 DK103184, and R01 DK115574