Capstone Project Team 1 "*SOX9-dependent fibrosis drives renal function in nephronophthisis*"

Leticia Castillón, Mai Soliman, Milda Sakalauskaite

1. Project Background

Nephronophthisis (NPHP) is a genetic disorder characterized by cystic kidney disease and progressive fibrosis, frequently linked to defects in primary cilia. It affects approximately 1 in 50,000 children and young adults, and is often linked to syndromic variants of Polycystic Kidney Disease, such as Joubert, Bardet–Biedl, and Meckel–Gruber syndromes. (Gupta et al., 2021). While multiple genes have been associated with NPHP, the molecular mechanisms driving renal fibrosis remain poorly understood. *Fbxw7* is an E3 ubiquitin ligase known to regulate the degradation of various transcription factors, including *c-Myc* and *Sox9*, which are implicated in cell proliferation and differentiation (Shimizu et al., 2018). Previous studies suggested that loss of *Fbxw7* may disrupt protein homeostasis and contribute to tissue remodelling, but its role in kidney-specific fibrotic processes was not fully elucidated (Petsouki et al., 2021). This study aimed to investigate the functional consequences of *Fbxw7* deletion in renal tissue, particularly in the context of ciliary maintenance, inflammation, and fibrotic progression. Using a mouse model with targeted deletion of *Fbxw7* in renal epithelial cells, and *in vitro* functional genetics experiments combined with quantitative mass spectrometry, the researchers sought to uncover downstream effectors and signalling pathways that mediate fibrosis, with the goal of identifying new therapeutic targets for NPHP and related ciliopathies.

2. Summary of the Original Study

This report details the processing and analysis of proteomics data from the PRIDE dataset PXD061542, which is part of the project *"SOX9-dependent fibrosis drives renal function in nephronophthisis"* (Patel et al., 2025). The study was published in EMBO in April 2025.

The study investigates the role of *Fbxw7* in the development of renal fibrosis associated with NPHP. Using *Fbxw7*-null mIMCD-3 cell lines, the researchers conducted quantitative bottom-up proteomic analysis using data-independent acquisition (DIA), where they processed the total protein from each sample. Briefly, the total protein suspension was reduced, alkylated and purified, digested with trypsin and separated by reverse phase. Eluted peptides were ionised by electrospray followed by mass spectrometric analysis in an Orbitrap Exploris 480 mass spectrometer. The resulting data were deposited in the PRIDE repository, a component of the ProteomeXchange consortium that facilitates the sharing of proteomics data. To uncover downstream molecular changes. Mass spectrometry was used to generate high-resolution proteomic data, resulting in the identification of 893,271 peptide-spectrum matches (PSMs), 38,737 unique peptides, and 4,923 unique proteins. The researchers do not report this numbers back in the paper, but provide a .mztab file in PRIDE from where the numbers have been extracted.

	Original publication
Peptide-spectrum matches (PSMs)	893,271
Unique peptides	38,737
Unique proteins	4,923

Table 1: Overview of Identified Spectra, Peptides, and Proteins.

One of the major findings was the downregulation of TMEM237, a protein critical for ciliary function, upon *Fbxw7* deletion (Fig.2A-C). This downregulation was associated with ciliary structural defects (Fig.2D-F). TMEM237 is a known component of the NPHP module, a group of proteins genetically associated with nephronophthisis (Gana et al., 2022). Its downregulation provides a mechanistic link between *Fbxw7* loss and ciliary dysfunction, establishing that the fibrotic phenotype observed in these mice may arise not only from transcriptional reprogramming (e.g., SOX9 upregulation) but also from direct structural defects in cilia due to impaired expression of transition zone components.

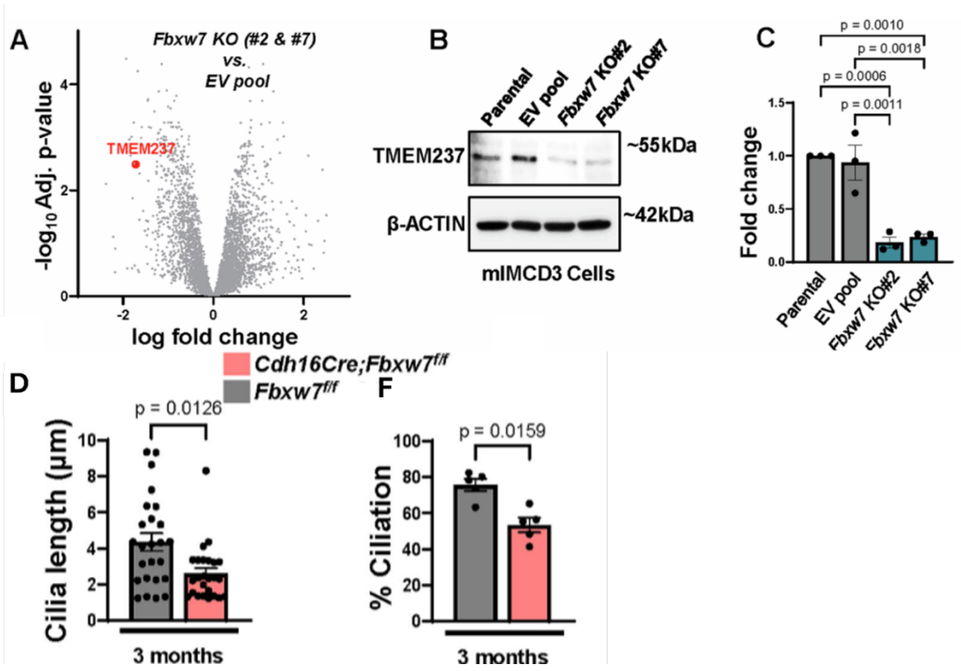

*Figure 1: **A**) Volcano plot showing differentially expressed proteins from Fbxw7-null cells (Fbxw7 KO#2 and Fbxw7 KO#7) versus empty vector (EV) pool mIMCD3 cells. The red dot represents the TMEM237 protein that is significantly downregulated in Fbxw7-null cells compared to EV pool mIMCD3 cells. n = 3 independent experiments. The data was normalized using cyclic loess, and statistical analysis was performed using linear models for microarray data (limma) with empirical Bayes (eBayes) smoothing to the standard errors. Proteins with an FDR-adjusted P value <0.05 and a fold change >2 were considered significant. (**B**) Validation of quantitative proteomics using immunoblotting and (**C**) quantification of TMEM237 from n = 3 independent experiments. Statistical analysis was performed using one-way ANOVA followed by Šídák's multiple comparisons test and is presented as the mean ± SEM. (**D**, **F**) Each data point represents the mean ciliary length per field of view from >20 images or percent ciliation in cystic cells from n ≥ 3 animals. Statistical analysis was performed using the Mann–Whitney test and is presented as the mean ± SEM.*

The study also investigated the expression levels of SOX9 and C-Myc. Even though the mass-spectrometry data did not highlight them as targets of *Fbxw7* deletion *in vitro*, the role of SOX9 and c-Myc in renal fibrosis has been documented before. The protein expression levels of these transcription factors were assessed through histological analysis and Western Blot. These revealed upregulation of transcription factors SOX9 and c-Myc, which is consistent with their known roles as FBXW7 degradation targets (Suryo Rahmanto et al., 2016). While both were elevated at the protein level, functional experiments revealed clear differences in their contributions to disease. c-Myc upregulation had little impact on fibrosis severity or inflammation (Fig.3A), suggesting it was not a key driver in this context. In contrast, SOX9 played a central role in mediating the fibrotic phenotype. Its expression correlated with fibrotic regions, and deleting SOX9 in Fbxw7-deficient mice significantly reduced fibrosis, inflammation, and tissue damage. The study also linked SOX9 to upregulation of WNT4 (Aggarwal et al., 2024), a profibrotic signalling molecule, indicating that SOX9 may activate a downstream transcriptional program promoting fibrosis (Fig.3B-G). These findings demonstrate that although both factors accumulate when FBXW7 is lost, only SOX9 is essential for disease progression. Additionally, the study identified WNT4 signalling dysregulation, likely acting downstream of SOX9, linking transcriptional reprogramming to fibrosis. These findings highlight a SOX9-dependent fibrotic program triggered by Fbxw7 deletion, with implications for future therapeutic targeting.

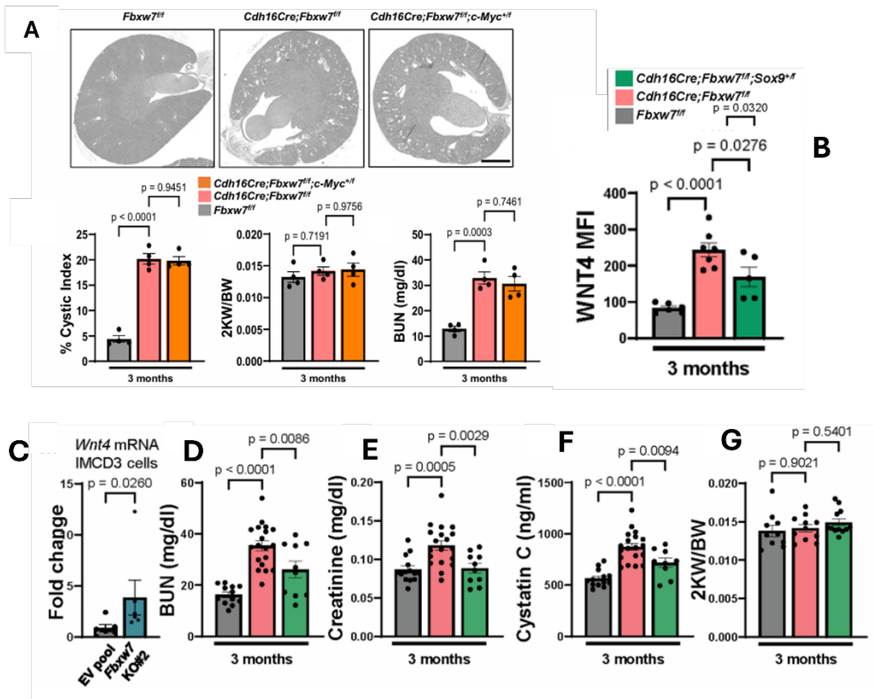

Figure 2: Effect of c-Myc and SOX9 on disease progression.

*(A) Representative images of whole kidney section scan, each data point represents one animal. Statistical analysis was performed using one-way ANOVA followed by Šídák's multiple comparisons test and is presented as the mean ± SEM. Scale bar: 400 µm. (**B**) an average of WNT4 (n ≥ 5) MFI per animal. Statistical analysis was performed using one-way ANOVA followed by Šídák's multiple comparisons test and is presented as the mean ± SEM. (**C**) qPCR of Wnt4 mRNA from EV pool and Fbxw7-null (Fbxw7 KO#2) mIMCD3 cells from n = 6 experiments. (**D**) Serum BUN (n ≥ 10), (**E**) Creatinine (n ≥ 10), (**F**) Cytastin C (n ≥ 9), and (**G**) 2KW/BW (n ≥ 10) from 3-month-old Fbxw7$^{f/f}$, Cdh16Cre;Fbxw7$^{f/f}$, and Cdh16Cre;Fbxw7$^{f/f}$;Sox9$^{+/f}$ mice. Each data point represents one animal. Statistical analysis was performed using one-way ANOVA followed by Šídák's multiple comparisons test and is presented as the mean ± SEM. Source data are available online for this figure*

3. Team Findings and Comparative Analysis

The authors did not provide .mzID files that we could directly use to assess the number of peptide-spectral matches (PSMs), nor the number of uniquely identified peptides and proteins. We decided to aim for a direct comparison with the publication results by using the mzML files containing the raw data and attempting to follow the methodology described in the paper. The first step was to use the mzML files for peptide-spectra matching and identification.

EncyclopeDIA is a library search engine for peptide identification that counts with several algorithms for DIA data analysis (Searle et al., 2018). Briefly, these kinds of tools search the spectra of peaks against a FASTA database containing protein sequences to match peaks and peptides and/or proteins. EncyclopeDIA allows for the use of various workflows. In this project, we generated a chromatogram library (Searle et al., 2018) (.elib format) from the raw .mzML files available; together with a Prosit-generated library for *Mus musculus* that is used as scaffold for the analysis, and the corresponding protein FASTA file (both of which were downloaded from Scaffold Proteome Software). The software first creates the chromatogram library, which is then used for peptide and protein identification.

At this point, our workflow already differs from that of the authors: they used

an empirically corrected chromatogram library as input to Encyclopedia. To generate this, they collect 6

GFP-DIA acquisitions of the sample pool – using the same gradient that they use for the single-injection DIA acquisition they performed for each sample, as described in (Searle et al., 2020). However, they do not provide the GFP-DIA acquisition files so that an empirically corrected library can be generated. In our case, we simply searched each .mzML file against the predicted library. The software outputs several graphs that can be used to assess the quality of the analysis. An example from one of the samples is provided in

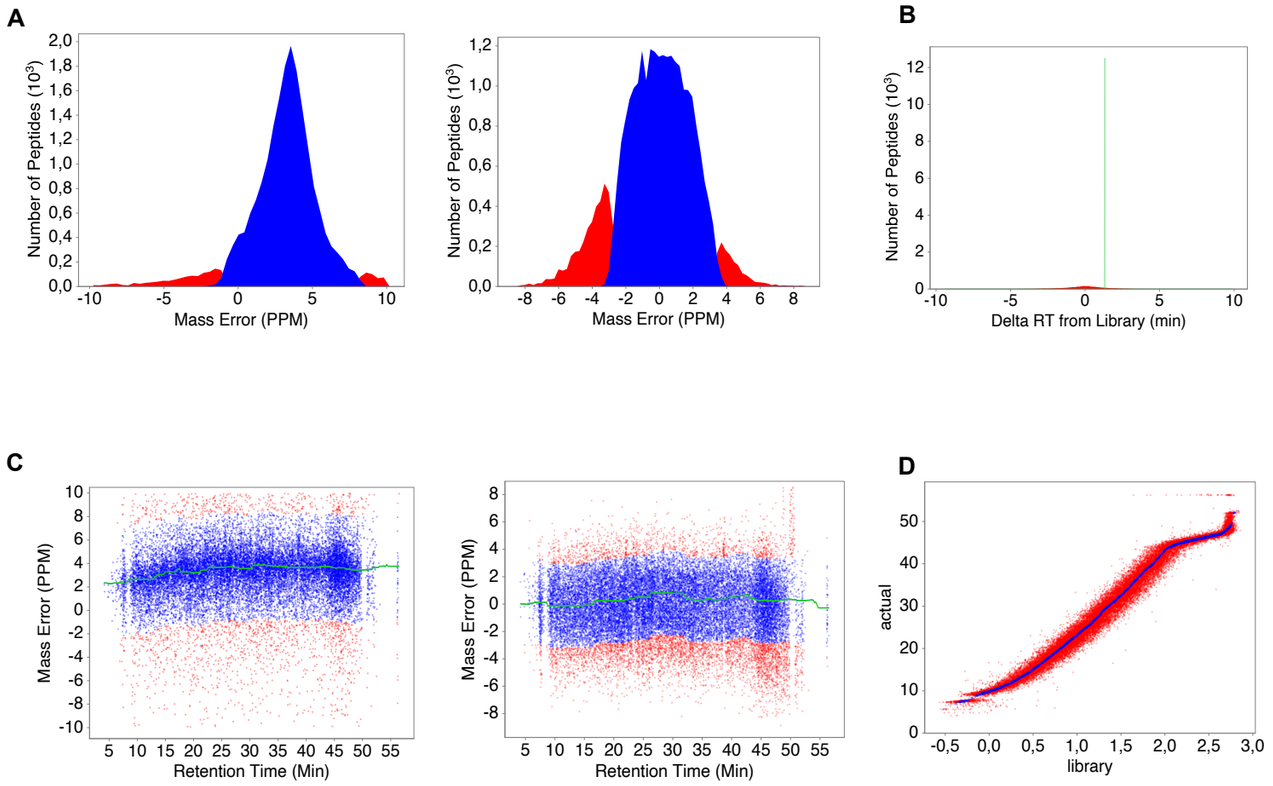

Figure 3. (A) MS1 (left) and MS2 (right) Mass error (PPM): mass accuracy of the fragment ions matched to the spectral library. In the x-axis, the mass error in parts per million. (B) Delta RT from library. The histogram shows the difference in retention time (RT) between the observed peptides in the sample and the expected RTs from the spectral library. The x-axis represents the D(RT) in minutes, while the y-axis shows the frequency. (C) Mass error plot for MS1(left) and MS2 (right), showing retention time in the x-axis and the mass error as parts per million (PPM) in the y-axis. The plot shows how much the observed mass of a peptide deviated from its theoretical mass. (D) Retention time (RT) fit plot. The X-axis shows the predicted retention time expected by the spectral library. The y-axis shows the actual retention time from the instrument.

Figure 3. We could not compare these QC plots with the ones from the original publication since they were not provided by the authors.

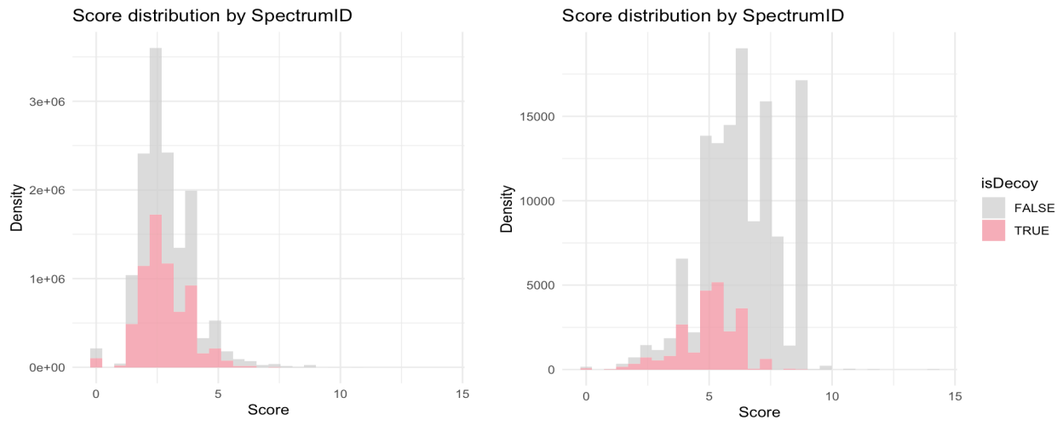

Figure 4. HyperScore distribution before (left) and after (right) rank filtering.

PSM object creation

EncyclopeDIA provides the identified peptides and proteins per spectra as a .features.txt file for each run. The feature file contains 40 variables (i.e. columns), among which are the sequence for the identified peptide for each spectrum number, the accession number(s) for the proteins that the peptide is matched, and other parameters including confidence scores for the matching.

We combined the features.txt file from each sample to create a combined features dataset, and we selected the variables that are of use for creating the PSM object creation. Because EncyclopeDIA does not rank the peptide-spectrum matches, we create a manual rank by grouping the spectrum IDs and ordering them according to the HyperScore variable.

Hyperscore is calculated by EncyclopeDIA and higher values indicate higher confidence in the peptide-spectrum match. The highest Hyperscore per spectrum would therefore correspond to rank = 1. We filtered the dataset to keep all those spectra with rank = 1 (Figure 4). We did not filter for FDR < 0.01 because EncyclopeDIA only includes those peptides identified at 1.0% FDR in their features.txt output - this can be checked in the EncyclopeDIA log files that are generated whith each run.

Identification summary

Once we filtered and selected for the variables that are useful to create the PSM object, we created it using PSM function (more details can be found in the script). Table 2 contains a summary of the PSM object.

Total PSMs	126,573
Target PSMs	103,811
Decoy PSMs	22,762
Unique peptides	54,114
Unique proteins	13,068

Table 2. PSM object summary.

Using an adjacency matrix, we investigated the number of proteins identified by a single peptide, the number of razor peptides, and the number of protein groups.

Peptide and protein quantification

The workflow used to generate the QFeatures object is detailed in Figure 5

Figure 5. QFeature object creation.

Briefly, we run EncyclopeDIA one more time, on the folder where the first analysis results are, and with a flag asking to create an integrated analysis. This creates an output.elib file that contains an integration of all the other results. We queried the .elib file using SQL and exported the peptide quantification table and the peptide to protein table. These were used to create the QFeatuers object.

Once we created the object, we transformed the intensities with value 0 to NA, since this means that the peptides were not really detected. Then, we remove those peptides containing over 20% of missing data, and we imputed the remaining missing data using the knn algorithm.

On the imputed data, we performed a peptide-based normalisation. We decided to use this because peptide-based normalisation better accounts for technical variation (since the peptide intensities is what we are measuring. First, we log2-transform the data, and we explore a couple of normalisation methods (Figure 6). The plots look the same, but in the end we decided for the center.median since summarizing using the median is more resistant to outliers than using mean. Then, we aggregated to proteins using the robustSummary method from QProteins.

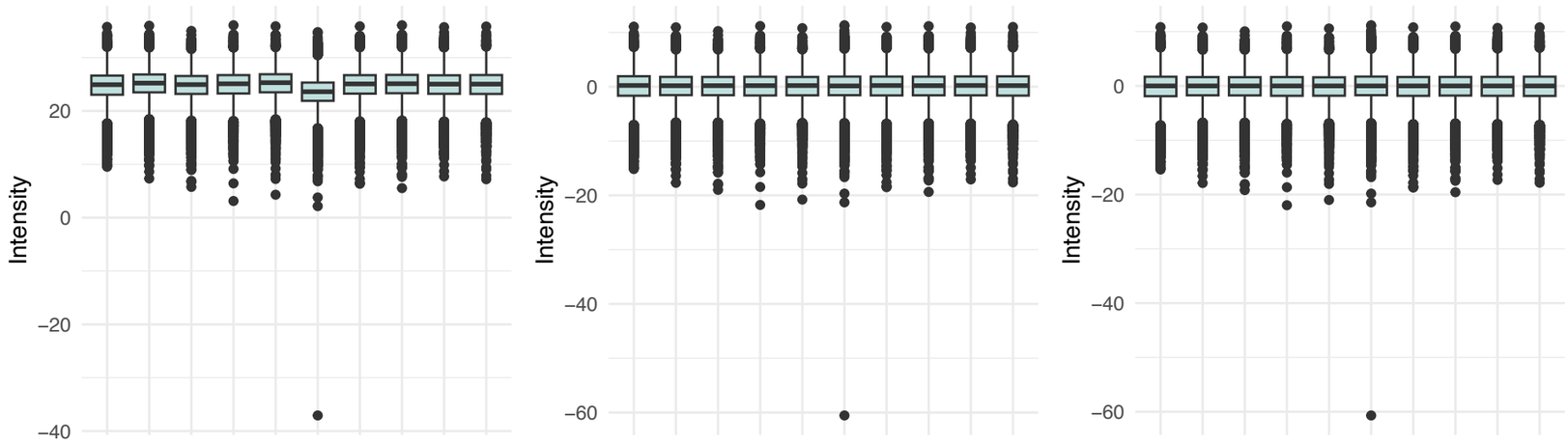

Figure 6. Normalisation on the log2-transformed data (before any other normalisation, on the left) using the center.mean method (center) or the center.median method (right) from QFeatures.

It is worth mentioning here that this and the following steps constitute a significant deviation from the described protocol in the original paper. The authors first assess the protein-exclusive intensity values using ProteiNorm and performed normalisation on the protein level using cyclic loess, followed by using limma with eBayes for differential expression analysis smoothing to the standard errors.

Visualising the data using Principal Components Analysis (PCA) (Figure 7), we can see that there is still some variation within samples that is not accounted for by the condition, but in general there is a good separation between WT and KO samples. Considering the variation that is inherit to biological experiments where you are deleting genes, where one does not always expect a 100% deletion, we considered this good enough to proceed with the differential expression analysis.

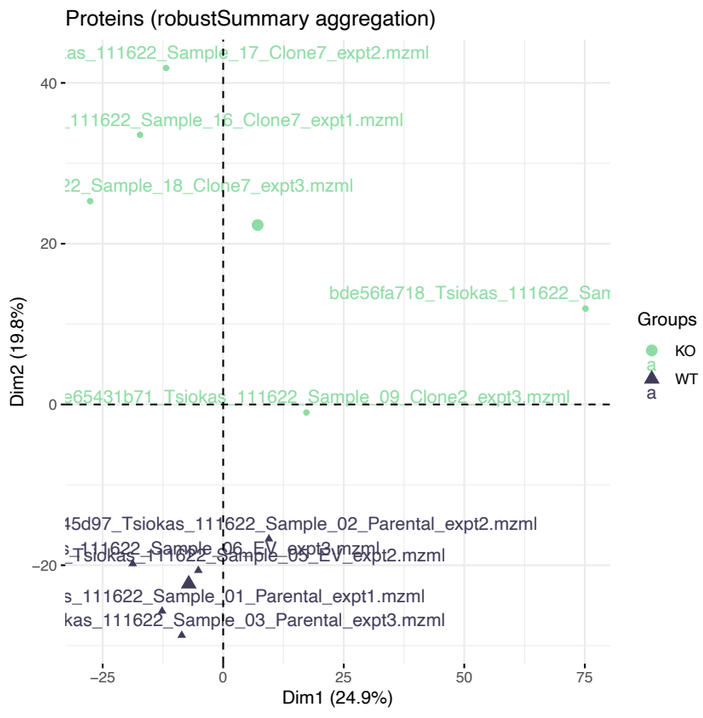

Figure 7. PCA on the aggregated protein data.

We performed the Differential Expression Analysis (DEA) using the package msqrob2, with default parameters, on the protein level data. This fits a model using robust linear regression on the protein expression values that have been aggregated from the peptide intensities. As contrasts, we used the KO samples (in green in Figure 7) vs the wild type samples (in purple in Figure 7). For the analysis, we pooled together the parental and the EV controls to have a balanced data set (5 KO and 5 WT), since we had to exclude from the EncyclopeDIA analysis some of the EV samples that could not be read by the software. In the original publication, they used the EV controls to compare to the KO samples. We repeated the DEA using only the remaining EV samples (n = 2) and only the parental samples (n=3) and the results did not change much, so we decided to include the 5 vs 5 set-up in the report.

Using an adjusted p-value < 0.05 as a threshold, we identify 8 differentially expressed proteins in the dataset (Figure 8, volcano plot). We plotted those proteins in a heatmap where we can see that the expression levels are clearly different from WT to KO samples.

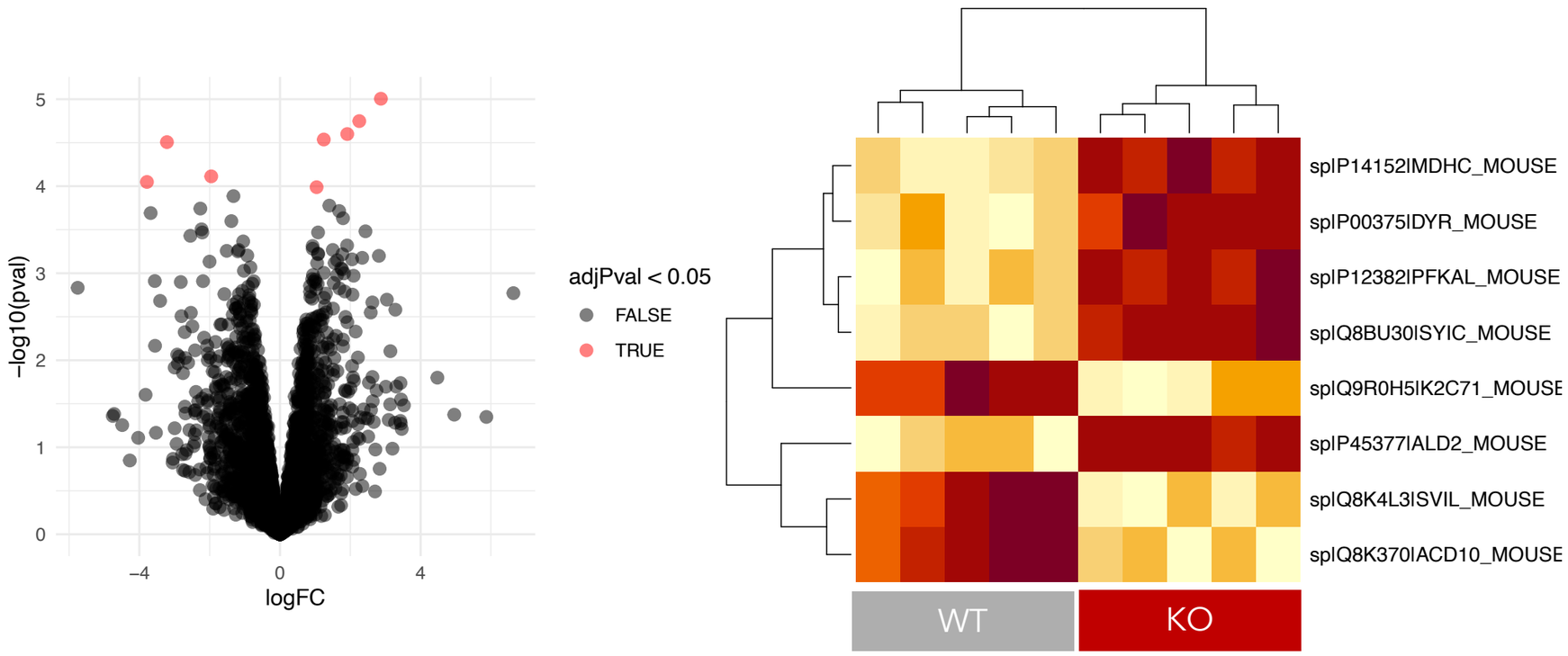

Figure 8. Volcano plot (left) highlighting (in red) the differentially expressed proteins in the dataset. Heatmap of those proteins (right).

The authors report TMEM237 in the original publication (Figure 1 A in this report). TMEM237 is the name of the gene that encodes for the protein TM237, with accession Q3V0J1 in UniProt. Hence, in the volcano plot the name should rather be TM237. We tried to query our results for both TM237 and Q3V0J1, but could not identify this proteins from our dataset.

Further, upon inspection of the results they provide in the. mztab on PRIDE, we are also not able to find this protein either by protein name or accession number. Because of this and since they are not reporting on the differentially expressed proteins they get form the analysis, a head-to-head comparison with their DEA is not possible. However, as can be seen from the PSM summary and from their reported uniquely identified peptides and proteins, our analysis identifies more, with the number of identified proteins being a bit too high for the instrument utilised for data collection. This indicates that we have most likely performed a bit less stringent filtering than they have. A possibility would have been to not only filter for rank 1 but to establish some sort of threshold for the Hyperscore parameter that would have discarded peptides identified with confidence < than “X”. Regardless, that does not affect our comparison with their differential expression analysis nor the fact that the protein/gene they report in the original publication does not seem to appear in their results either.

References

Aggarwal, S., Wang, Z., Rincon Fernandez Pacheco, D., Rinaldi, A., Rajewski, A., Callemeyn, J., Van Loon, E., Lamarthée, B., Covarrubias, A.E., Hou, J., Yamashita, M., Akiyama, H., Karumanchi, S.A., Svendsen, C.N., Noble, P.W., Jordan, S.C., Breunig, J.J., Naesens, M., Cippà, P.E., Kumar, S., 2024. SOX9 switch links regeneration to fibrosis at the single-cell level in mammalian kidneys. Science 383, eadd6371.

Gana, S., Serpieri, V., Valente, E.M., 2022. Genotype–phenotype correlates in Joubert syndrome: A review. American Journal of Medical Genetics Part C: Seminars in Medical Genetics 190, 72-88.

Gupta, S., Ozimek-Kulik, J.E., Phillips, J.K., 2021. Nephronophthisis-Pathobiology and Molecular Pathogenesis of a Rare Kidney Genetic Disease. Genes 12, 1762.

Maulin Mukeshchandra Patel, V.G., Bryan Lettenmaier, Eleni Petsouki, Kurt A Zimmerman, John A Sayer, and Leonidas Tsiokas, 2025. SOX9-dependent fibrosis drives renal function in nephronophthisis.

Patel, M.M., Gerakopoulos, V., Lettenmaier, B., Petsouki, E., Zimmerman, K.A., Sayer, J.A., Tsiokas, L., 2025. SOX9-dependent fibrosis drives renal function in nephronophthisis. EMBO Molecular Medicine, 1-21-21.

Petsouki, E., Gerakopoulos, V., Szeto, N., Chang, W., Humphrey, M.B., Tsiokas, L., 2021. FBW7 couples structural integrity with functional output of primary cilia. Communications Biology 4, 1066.

Shimizu, K., Nihira, N.T., Inuzuka, H., Wei, W., 2018. Physiological functions of FBW7 in cancer and metabolism. Cellular Signalling 46, 15-22.

Suryo Rahmanto, A., Savov, V., Brunner, A., Bolin, S., Weishaupt, H., Malyukova, A., Rosén, G., Čančer, M., Hutter, S., Sundström, A., Kawauchi, D., Jones, D.T.W., Spruck, C., Taylor, M.D., Cho, Y.J., Pfister, S.M., Kool, M., Korshunov, A., Swartling, F.J., Sangfelt, O., 2016. FBW7 suppression leads to SOX9 stabilization and increased malignancy in medulloblastoma. The EMBO Journal 35, 2192-2212-2212.

Searle, B.C., Pino, L.K., Egertson, J.D., Ting, Y.S., Lawrence, R.T., MacLean, B.X., Villén, J., MacCoss, M.J., 2018. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. Nat Commun 9, 5128. https://doi.org/10.1038/s41467-018-07454-w

Searle, B.C., Swearingen, K.E., Barnes, C.A., Schmidt, T., Gessulat, S., Küster, B., Wilhelm, M., 2020. Generating high quality libraries for DIA MS with empirically corrected peptide predictions. Nat Commun 11, 1548. https://doi.org/10.1038/s41467-020-15346-1