UPPSALA
UNIVERSITET

Single-cell type profiling of fetal kidney tissue

Santhosh Rajakumar

Programme name – Master's in Bioinformatics

Course: Research training (1MB841) 15 hp 2024-2025

Department of Immunology Genetics and Pathology, Dr Cecilia Lindskog group , Uppsala university, Uppsala, Sweden

Project Supervisor: Cecilia Lindskog, Loren Méar

1. ABSTRACT

The fetal kidney, a dynamic site of nephron formation known as nephrogenesis, makes an ideal model for studying protein expression during human development. This study aimed to map protein expression in fetal kidney tissues and compare it to adult kidney tissue using a multi-omics approach that integrates transcriptomics and antibody-based proteomics. Bulk mRNA sequencing data were analyzed from fetal kidney samples to identify genes with significantly higher expression levels compared to adult kidney tissue. The functional roles of these genes were validated through Gene Ontology (GO) enrichment analysis, which highlighted key biological processes associated with kidney development. Protein localization and expression patterns were further examined and confirmed using Immunohistochemistry (IHC).

2. INTRODUCTION

The kidney is a vital organ responsible for maintaining homeostasis by regulating blood composition, including water balance, electrolyte levels, and pH. It plays a critical role in excreting metabolic waste products and organic compounds. (Habuka, Fagerberg, Hallström, Kampf, et al., 2015). The fetal kidney undergoes a dynamic and tightly regulated process of development, marked by cellular differentiation and organogenesis. As the site of nephrogenesis, where functional nephrons are formed, the fetal kidney serves as a unique model for studying the protein expression driving kidney development.

Human kidney development begins during the first three months of pregnancy and progresses through three stages: pronephros (around weeks 3–4), mesonephros (around weeks 4–8), and metanephros (beginning around week 9 and continues throughout fetal development). Among these, the pronephros and mesonephros are transient structures that form and then subsequently regress, while the metanephros evolves into the functional kidney. Each stage represents a critical phase in the transformation of primitive structures into fully functional kidneys, which play a vital role in maintaining fluid and electrolyte balance and excreting metabolic waste (Rosenblum et al., 2017).

The Human Protein Atlas (HPA) project stands out as a groundbreaking initiative for mapping all human protein-coding genes. The remarkable advancements in proteomics and transcriptomics have fueled global efforts to map human genes with unprecedented resolution, exemplified by HPA(Thul et al., 2017). It offers a comprehensive database integrating spatial antibody-based proteomics with transcriptomics technologies, accessible at (www.proteinatlas.org).

The HPA has made significant advancements in proteomic research, particularly in adult tissues, targeting over 80% of the ~20,000 protein-coding genes in the human genome Despite these advancements there are hundreds of proteins that are not detected in any of the tissues or cell types based on mRNA, and we hope that some of these may be expressed in fetal tissues (Uhlén et al., 2015), (Karlsson et al., n.d.).

This study seeks to systematically map protein expression in fetal kidney tissue by integrating transcriptomic data analysis with antibody-based proteomics approaches to shed light on the

cell type-specific localization of proteins that are uniquely expressed or overexpressed in fetal tissues compared to adult tissues, providing a unique opportunity to address critical gaps in our knowledge of the human proteome and uncover valuable insights. The findings will help enhance our understanding of protein expression in kidney development and identifying biomarkers specific to fetal kidney development.

3. AIM

The aim of this project is to identify and prioritize relevant protein targets specific to fetal kidney tissue by leveraging a multi-omics approach. This involves the comprehensive analysis of transcriptomics data to uncover gene expression patterns unique to the fetal kidney, followed by the integration and validation of these findings with antibody-based proteomics data.

4. MATERIALS AND METHODS

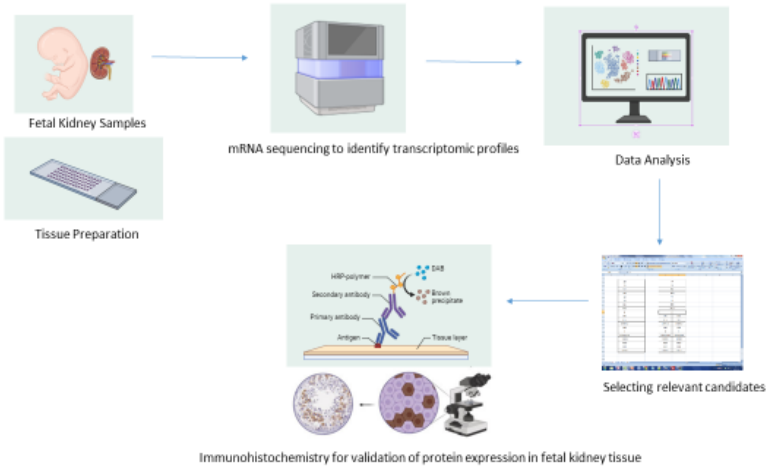

Figure 1. *Schematic representation of the workflow for identifying fetal kidney-specific proteins.The data is analyzed to identify genes that are elevated in fetal tissues compared to adult tissues. Candidate proteins are selected and validated using antibody-based proteomics approaches. Finally,IHC is performed to validate protein expression within fetal kidney tissue, enabling spatial localization and confirmation of findings.*

4.1 Dataset Description

This study utilized bulk mRNA sequencing data provided by HPA, derived from frozen tissue samples of fetal kidneys collected at two developmental time points: week 15 and week 18. The mRNA sequencing was conducted using Illumina HiSeq2000 and 2500 platforms (Illumina, San Diego, CA), adhering to the standard Illumina RNA sequencing protocol employed by the HPA for all normal tissue samples (Wang et al., 2009). Gene expression levels in these samples were quantified as normalized transcripts per million (nTPM), a standard metric for comparing transcript abundance across samples. The samples were obtained through collaboration with the biobank infrastructure at Mount Sinai Hospital. In addition to the fetal kidney data, adult kidney gene expression profiles were obtained from a combination of HPA and Genotype-Tissue Expression (GTEx) project datasets. These adult datasets were integrated to create a unified resource for comparative analysis. It is important to note that the RNA sequencing data were derived exclusively from frozen tissue samples

curated by the HPA. In contrast, paraffin-embedded tissue samples used for immunohistochemistry were obtained from different individuals, ensuring optimization of assays for their respective methodologies.

4.2 Data Processing and Bioinformatics Analysis

4.2.1 Data Structuring

To facilitate direct comparison between fetal and adult tissue expression datasets, the data structures were carefully transformed and standardized ensuring compatibility. The analysis focused on identifying expression patterns and differences in gene expression between fetal tissues at weeks 15 and 18 compared to the adult tissue. To enable comparative analysis, the data structure was transformed from a long to a wide format. This restructuring was essential for organizing nTPM values into a format suitable for side-by-side comparisons across the adult and fetal tissues. In the adult tissue expression data, gene identifiers were standardized to match the format used in the fetal dataset, allowing seamless integration of the two datasets.

4.2.2 Comparative Analyses

Comparisons between fetal tissue expression and adult consensus levels were performed by calculating the ratio of average fetal kidney expression to adult kidney consensus expression highlighting genes with differential expression patterns between them.We also focused on the two fetal kidney samples representing distinct developmental stages. Expression ratios were calculated between these samples to quantify differential expression dynamics between week 15 and week 18, provided insight into genes that might drive stage-specific developmental processes in the fetal kidney.

A extensive literature review was conducted to identify well-established marker genes specific to fetal kidney tissues. These marker genes along with their associated cell types information were incorporated into the dataset providing a resource for subsequent analyses. Summary metrics such as consensus distribution and maximum expression were included to provide additional context for gene expression patterns. Several binary columns were added to the dataset to flag genes with ratios exceeding a predefined threshold of two, including ratios such as fetal kidney to adult kidney expression and comparisons between fetal weeks 15 and 18 . These binary indicators enriched the dataset by highlighting genes with significant tissue-specific expression patterns which help in refining the list of potential protein targets.

4.2.3 Validation

The Annotated Protein Expression (APE) score, developed within the HPA project, provides a high-confidence estimate of on-target protein expression. This score uses two or more antibodies targeting non-overlapping epitopes on a single protein. By integrating staining patterns, published literature, expression data, and bioinformatic predictions, the APE score enhances the reliability of protein expression analysis.

The Protein Existence (PE) score, metric for validating protein expression in cells and tissues, provides evidence confirming the actual detection and presence of a protein within a tissue based on experimental data. As highlighted, this data is primarily derived from the UniProt database (www.uniprot.org), a comprehensive resource containing both manually reviewed

and automatically annotated protein sequences. Protein existence data from the HPA LIMS database and APE scores were integrated into the dataset to validate the candidate genes.

To understand the functional roles of the genes, the ClusterProfiler package (v4.14.4) and the org.Hs.eg.db package (v3.2.0) was used to perform Gene Ontology (GO) enrichment analysis (Wu et al., 2021). Biological process (BP) was the sub-ontology applied to retrieve comprehensive GO term information. The results were adjusted for multiple testing using the Benjamini-Hochberg method (pAdjustMethod = "BH") and filtered for significance with a q-value cutoff of 0.2. Throughout the analysis, data visualization techniques, including summary tables, were generated to depict gene expression patterns. Gene expression patterns were visualized using scatter plots generated with the ggplot2 R package (version 3.5.1).

4.2.4 Immunohistochemistry (IHC)

Automated immunohistochemistry (IHC) was performed for both adult and fetal tissue samples to see protein expression patterns, employing the Autostainer 480 Instruments (Lab Vision, Fremont, CA). This process was strictly done to the standardized protocols established by the Human Protein Atlas (HPA), which include optimized incubation times, and washing steps to maximize specificity and minimize background staining. The dilution factors were optimized or each antibody to ensure the best possible staining and detection during the IHC process. Following the IHC staining procedure, high-resolution digital images of the stained slides were acquired using the Scanscope AT2 slide scanner (Aperio, Vista, CA), equipped with a 20× objective lens. This advanced imaging system ensured accurate image of tissue morphology and staining patterns at a cellular resolution.

4.2.5 Code availability

The analysis was conducted using R (version 4.4.1) using various bioinformatics packages.The scripts used for data processing have been uploaded to my GitHub repository https://github.com/SanthoshRajakumar/Human-protein-atlas-project and are available upon request.

5. RESULTS

The dataset derived from bioinformatics analysis of fetal and adult kidney tissues, comprised of 20,163 genes, which were filtered using specific criteria to identify potential candidates for further analysis. Six separate lists were created based on distinct selection parameters detailed in Table 1.

Once the genes were selected in each sheet, further refinement was performed. For a gene to be considered for final analysis, genes that were expressed highly in fetal kidney tissue compared to adult and had evidence from HPA LIMS (laboratory information management system) confirming RNA expression in kidney tissue were prioritized. Genes that were expressed in fetal kidney tissue but showed RNA expression in tissues other than kidney were also considered. Genes that exhibited high intensity staining in adult kidney tissue were excluded, as no significant difference in expression was expected in fetal kidney tissue.

Sheet	APE reliability score	Cell type/Gene	Expression criteria	Protein existence	No of genes
1	High and Medium	Gene- no known gene/celltype were included	Overexpressed in fetal kidney compared to adult tissue	Experimental evidence at transcript level, Protein inferred from homology	6
2	High and Medium	Gene- no known gene/celltype were included	Overexpressed in early fetal kidney tissue compared to late fetal kidney tissue	Experimental evidence at transcript level, Protein inferred from homology	38
3	High and Medium	Gene- no known gene/celltype were included	Overexpressed in late fetal kidney tissue compared to early fetal kidney tissue	Experimental evidence at transcript level, Protein inferred from homology	34
4	High	Gene-Included known gene/celltype	Overexpressed in kidney compared to adult tissue	Not applicable	18
5	High	Gene- Included known gene/celltype	Overexpressed in early fetal kidney tissue compared to late fetal kidney tissue	Not applicable	18
6	High	Gene- Included known gene/celltype	Overexpressed in late fetal kidney tissue compared to early fetal kidney tissue	Not applicable	12

Table 1*- The table summarizes the gene selection criteria and the number of genes identified for further analysis based on expression and reliability parameters. Each sheet corresponds to a specific combination of these parameters.*

The data analysis identified a total of 2,355 genes with significantly higher expression levels in fetal kidneys compared to adult kidneys, filtered based on a threshold ratio of fetal to adult expression greater than 2. These genes are active in kidney development during the fetal stages, showing distinct expression profiles that are significantly less expressed in adult kidneys.

In week 15, a wide range of genes with high expression was observed, as seen by numerous outliers with extremely high nTPM values in the scatter plot (Image 1). In contrast, week 18 displayed fewer genes with extremely high nTPM values compared to week 15, suggesting a continuation of developmental activity during the later fetal stage.

Figure 2. *A scatter plot illustrating the comparison of Consensus Kidney (nTPM) values, representing adult kidney expression levels (x-axis), with fetal kidney expression levels at week 15 (blue points) and week 18 (red points) (y-axis).*

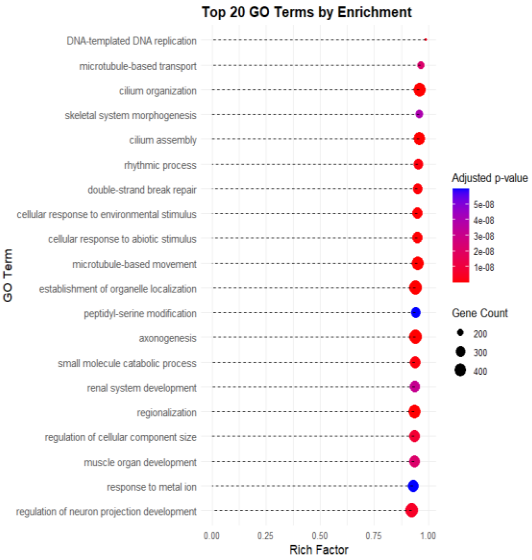

Figure 3. *Visualizing enrichment results using ggplot2- A lollipop chart to visualize the top 20 most significant GO terms on p-value.*

In figure 2 green points represent adult kidney samples plotted against themselves, serving as a reference for consistency. The plot highlights differentially expressed genes across the adult and fetal kidney stages, emphasizing transcriptional differences important to renal development.

The top 20 most significant GO terms were selected based on adjusted p-values. These terms were visualized in a lollipop plot (Image 3) that highlighted the rich factor (calculated as the

ratio of genes in the cluster associated with a GO term to the total number of genes associated with that term), adjusted p-value (represented by a color gradient from red to blue), and gene count (indicated by the size of the points). The plot provided a clear depiction of the enriched biological processes, focusing on the most significant terms with higher gene representation and lower adjusted p-values.

Key findings included specific developmental processes such as axonogenesis, renal system development, muscle organ development, and regulation of neuron projection development, indicating their critical roles in the biological systems under kidney development. The lollipop plot effectively summarized these results, providing a comprehensive overview of the enriched biological processes and their relevance to kidney development.

After filtering and selecting the most relevant genes for fetal kidney tissue analysis, the following genes were identified as the ideal candidates.The antibodies were carefully selected to target proteins that are expected to show specific expression patterns in fetal kidney tissues. The details of the selected genes, antibodies, and their respective concentrations are in Table 2.

ENSG ID	**Gene name**	**Antibody ID**	**Volume (ul)**	**Dilution**
ENSG00000131747	TOP2A	CAB002448	641	1:30
ENSG00000149294	NCAM1	HPA039835	348	1:75
ENSG00000116183	PAPPA2	HPA018412	204	1:50
ENSG00000100373	UPK3A	HPA018415	338	1:250
ENSG00000105668	UPK1A	HPA049879	316	1:35
ENSG00000164363	SLC6A18	HPA011885	326	1:450

Table 2- *The table summarizes the ideal candidates along with their HPA ID, dilution and volume.*

The dilution of each antibody was optimized based on strict validation criteria taking into consideration the expected staining pattern based on available literature as well as correlation between RNA expression levels and protein expression levels. Before performing IHC on fetal kidney tissues, the antibodies used for detecting target proteins were first validated for reliability and specificity. The validation process involved testing the antibodies on adult kidney tissue to ensure they effectively recognize the proteins of interest. This step was crucial for confirming that the antibodies would yield reliable results when used on fetal tissue. High-resolution digital images were obtained through slide scanning, and the resulting images are presented in figure 5.

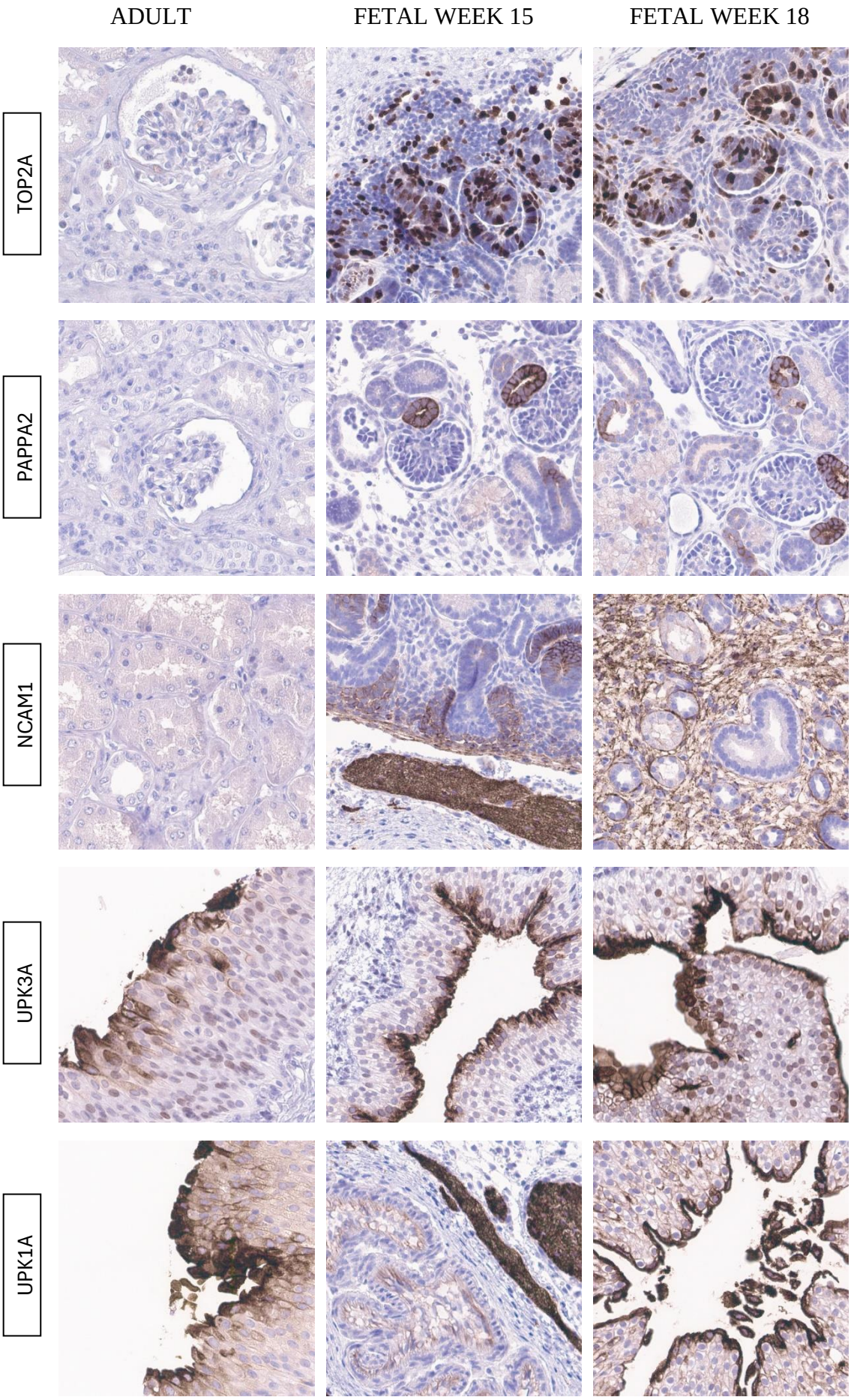
ADULT
FETAL WEEK 15
FETAL WEEK 18
TOP2A
PAPPA2
NCAM1
UPK3A
UPK1A

Figure 5. *The IHC images highlights the expression of TOP2A, PAPPA2,NCAM1,UPK3A,UPK1A across adult kidney(left), fetal kidney at 15 weeks(centre), and fetal kidney at 18 weeks(right).*

6. **DISCUSSION**

Through transcriptomic data analysis, we identified 6 key protein targets specific to fetal kidney tissue. These 6 findings, validated through IHC, provide important insights into expression profile and functional roles of these proteins during kidney development. The GO enrichment analysis further validated these findings, revealing biological processes involved in kidney development, including cell proliferation, morphogenesis, and extracellular matrix organization. Using antibody-based resources from HPA and the genes which we found as markers from the extensive research from literature helped in localizing proteins within specific cell types.

An important outcome of this study is the identification of proteins uniquely or highly expressed in fetal kidneys but absent in adult tissues. The first three proteins show such expression. TOP2A gene encodes a nuclear enzyme a DNA topoisomerase that controls DNA topological structure, chromosome segregation, and cell cycle progression.(Panvichian et al., 2015). In the adult kidney, the staining of the TOP2A protein is weak and scattered within a small subset of cells, likely within their nuclei, indicating low expression of TOP2A. In contrast to the adult kidney, expression in fetal kidney week 15 is more intense within developing glomeruli and tubules, which are regions of active growth and differentiation. In fetal kidney week 18, the staining remains intense but demonstrates a more localized distribution compared to week 15 reflecting transition from widespread proliferation to more targeted cellular growth as the kidney matures.

PAPPA2 is a metalloproteinase known to modulate insulin-like growth factor (IGF) availability by cleaving IGF-binding proteins, thereby regulating cell proliferation and differentiation (Barrios et al., 2021). Its elevated expression in the fetal kidney tissue aligns with its role in promoting mesenchymal-to-epithelial transition and nephron formation. PAPPA2 expression is mostly in developing nephron structures especially above the glomeruli, they look like developing distal tubule like structures, indicating its potential role in early nephrogenesis. This can be seen in both the development stages , where there is no expression in adult kidney , thereby correlating with the transcriptomic data that we analysed.

NCAM1 encodes a cell adhesion protein involved in cell-to-cell interactions as well as cell-matrix interactions during development and differentiation. It plays a role in the development of the nervous system by regulating neurogenesis. This protein plays a role in signal transduction by interacting with fibroblast growth factor receptors (Buzhor et al., 2013). NCAM1 expression in the adult kidney is faint but strong in the fetal kidney, with cytoplasmic and membranous staining localized primarily to proximal and distal tubules. Developing glomeruli and tubular structures involved in signaling are the sites of active morphogenesis , where we can see strong expression of NCAM1,emphasizing its critical role in nephron formation.

UPK3A and UPK1A encodes proteins which contributes to the structural integrity and regulating bladder membrane permeability (Habuka, Fagerberg, Hallström, Pontén, et al.,

2015). Its expression is highly tissue-specific, as seen by its presence in the urinary bladder and ureter-like structures in the fetal kidney. These proteins were observed to be expressed in urinary bladder tissue but absent in adult kidney tissue. Interestingly, while bulk RNA sequencing data suggested expression in fetal kidney tissues, the IHC results highlighted the protein's expression in urinary bladder tissue, highlighting the importance of combining antibody-based imaging with RNA sequencing data for validation.

Despite the significant findings, this study has certain limitations. The use of bulk mRNA sequencing may overshadow the contributions made by specific cell types. Single-cell RNA sequencing and expanded datasets on other fetal tissues could address this limitation in future enhancing our understanding of fetal proteome.

In conclusion, this research highlights the importance of mapping the fetal proteome for understanding the molecular mechanisms underlying human development and the unique proteins expressed during fetal development.The findings of this project will contribute to a more comprehensive understanding of the human proteome, offering insights into protein function during early kidney development.

7. REFERENCES

1. Barrios, V., Chowen, J. A., Martín-Rivada, Á., Guerra-Cantera, S., Pozo, J., Yakar, S., Rosenfeld, R. G., Pérez-Jurado, L. A., Suárez, J., & Argente, J. (2021). Pregnancy-Associated Plasma Protein (PAPP)-A2 in Physiology and Disease. *Cells*, *10*(12). https://doi.org/10.3390/cells10123576
2. Buzhor, E., Omer, D., Harari-Steinberg, O., Dotan, Z., Vax, E., Pri-Chen, S., Metsuyanim, S., Pleniceanu, O., Goldstein, R. S., & Dekel, B. (2013). Reactivation of NCAM1 defines a subpopulation of human adult kidney epithelial cells with clonogenic and stem/progenitor properties. *The American Journal of Pathology*, *183*(5), 1621–1633. https://doi.org/10.1016/j.ajpath.2013.07.034
3. Habuka, M., Fagerberg, L., Hallström, B. M., Kampf, C., Edlund, K., Sivertsson, Å., Yamamoto, T., Pontén, F., Uhlén, M., & Odeberg, J. (2015). The Kidney Transcriptome and Proteome Defined by Transcriptomics and Antibody-Based Profiling. *PLOS ONE*, *9*(12), 1–19. https://doi.org/10.1371/journal.pone.0116125
4. Habuka, M., Fagerberg, L., Hallström, B. M., Pontén, F., Yamamoto, T., & Uhlen, M. (2015). The Urinary Bladder Transcriptome and Proteome Defined by

Transcriptomics and Antibody-Based Profiling. *PLOS ONE*, *10*(12), 1–13. https://doi.org/10.1371/journal.pone.0145301

5. Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., Sjöstedt, E., Butler, L., Odeberg, J., Dusart, P., Edfors, F., Oksvold, P., von Feilitzen, K., Zwahlen, M., Arif, M., Altay, O., Li, X., Ozcan, M., Mardinoglu, A., … Lindskog, C. (n.d.). A single–cell type transcriptomics map of human tissues. *Science Advances*, *7*(31), eabh2169. https://doi.org/10.1126/sciadv.abh2169
6. Panvichian, R., Tantiwetrueangdet, A., Angkathunyakul, N., & Leelaudomlipi, S. (2015). TOP2A amplification and overexpression in hepatocellular carcinoma tissues. *BioMed Research International*, *2015*, 381602. https://doi.org/10.1155/2015/381602
7. Rosenblum, S., Pal, A., & Reidy, K. (2017). Renal development in the fetus and premature infant. *Seminars in Fetal & Neonatal Medicine*, *22*(2), 58–66. https://doi.org/10.1016/j.siny.2017.01.001
8. Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M., Bäckström, A., Danielsson, F., Fagerberg, L., Fall, J., Gatto, L., Gnann, C., Hober, S., Hjelmare, M., Johansson, F., … Lundberg, E. (2017). A subcellular map of the human proteome. *Science*, *356*(6340), eaal3321. https://doi.org/10.1126/science.aal3321
9. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419. https://doi.org/10.1126/science.1260419

10. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

11. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, *2*(3). https://doi.org/10.1016/j.xinn.2021.100141

ADDITIONAL INFORMATION:

I. **Background, where, when, and for how long.**
Research training at the Department of Immunology Genetics and Pathology, Dr Cecilia Lindskog group , Uppsala university, Uppsala, Sweden, supervised by Cecilia Lindskog and Loren Méar, took place from November 1st to January 10th. This computational study was conducted under the direct guidance of the supervisors in the lab.

II. **Describe the central activities of your workplace**.
Progress meetings on Thursday were conducted to review project advancements and analyze results. Additionally, zoom meetings offered diverse perspectives on problem-solving approaches for finding relevant candidates.

III. **Description of personnel, methods, equipment, and possible research results.** (*kindly refer to the* ***materials and methods*** *and* ***results*** *section of the report*)

IV. **A short description of a common work day.**
Daily tasks involved data segregation, cleaning and analysis as per the schedule. Utilizing R for data visualization guided subsequent analyses and determined research directions.

V. **A short description of group meetings, literature seminars, etc.**

Meetings with Cecilia and Loren, and other research assistants helped in resolving data analysis and visualization challenges while enhancing comprehension of the biological aspects underlying computational work.

VI. **Briefly summarize your theory task**

Initial training involved a two-week literature survey to grasp the research's background.

VII. **References to publications or similar.** (*refer to **references** section of the report*)

VIII. **Self-assessment of your experience during the research training**.

Utilizing R for data analysis and visualization seemed challenging at first, but improved as the training progressed. I gained some experience on analysis of data of transcriptomics and a bit of proteomics data , which I will definitely use in my master thesis.

IX. **What worked well and what could have been done better?**

(*kindly refer to **discussion** section of the report*)