Bios 785 Project

Wenbo Wang, Anil Anderson, Kevin Zhang

April 2023

1 Introduction

In our analysis we are using the data sets from “Targeting a Braf/Mapk pathway rescues podocyte lipid peroxidation in CoQ-deficiency kidney disease” by Sidhom et. al. The data set consists of a set of six scRNA-seq data sets corresponding to six five-month old mice with three controls and three mice with kd/kd mutation relevant to kidney disease. We sought to determine which genes and biological pathways were most highly associated with the kidney disease phenotype in mice and whether there were implications in human studies for kidney disease. For humans, kidney failure occurs when monogeneic PDSS2 mutations occur which is similar to the process of kd/kd kidney failure in mice. A mutation in PDSS2 in mice results in global CoQ deficiency. The BRAF/Mapk pathway is driven by CoQ deficiency, and modulating the pathway with the targeting compound GDC-0879 rescued podocyte injury and kidney filter function. The GDC-0879 compound have some effects; treatment with the compound restores an enzyme that protects cells from lipid per-oxidation, and relevant pathways that affect several human kidney diseases.

2 Methods

2.1 Imputation

The UMI expression count data across all 6 datasets were sparse, with around 5% of count data being non-zero. This observation is supported in the literature (van Dijk and others, 2018). Due to the lack of count data found in our datasets due to “dropout” or potential undersampling of mRNA molecules, we imputed the dataset using SAVER and MAGIC. Although we are aware of the transfer learning methods of DCA and SAVER X being an improvement on the two imputation methods, loading Tensorflow as required for these new methods on the local computer led to errors related to package dependencies and file paths, and can be considered future work.

We will briefly describe the two imputation methods in the context of our work. Please also note that the imputation step was completed before quality control, due to concerns of the effect of reducing the count of podocyte cells following quality control on imputation accuracy.

1. MAGIC (Markov affinity-based graph imputation of cells)

   This imputation method considers cell-by-cell distances, computing an affinity matrix that are normalized into a transition matrix M that can be exponentiated by t. The exponentiated matrix M^t represents a random walk of length t that represent cellular states at each time, where the entries are the probability of a cell i will reach cell j. This is also known as diffusion.

   In our computation of the imputed UMI matrix, the cell-by-cell distance matrices are computed using nearest neighbor graphs of 5 nearest neighbors defined using euclidean distance. The level of diffusion parameter to be the default of 3.

2. SAVER (Single-cell analysis via expression recovery)

   A second approach we applied to impute our data is SAVER, which addresses potential oversmoothing from normalization across cells and genes as completed by MAGIC by maintaining biological variation through a gamma distributed uncertainty term in a poisson-gamma mixture, or negative binomial model.

   By modeling UMI count data as two-level hierarchical model with gamma prior and poisson posterior, the poisson distribution represents cellular noise which can be isolated from the UMI expression count estimates.

$$\begin{aligned} Y_{gc} &\sim \textit{Poisson}\left(s_c \lambda_{gc}\right) \\ \lambda_{gc} &\sim \textit{Gamma}\left(\alpha_{gc}, \beta_{gc}\right) \end{aligned}$$

   For λ_{gc} being the normalized true expression data, the parameters α_{gc}, β_{gc} are reparameterized as $\alpha_{gc} = \mu_{gc}^2/v_{gc}, \beta_{gc} = \mu_{gc}/v_{gc}$ that can be obtained from the data via a poisson LASSO regression model on the log-likelihood function (2) in the paper (Huang and others, 2018). Finally, the estimated true expression parameter $\hat{\lambda}_{gc}$ can be represented by the estimated prior parameters along with scaled terms Y_{gc}, s_c as seen below.

$$\hat{\lambda}_{gc} = \frac{Y_{gc} + \hat{\alpha}_{gc}}{s_c + \hat{\beta}_{gc}} = \frac{s_c}{s_c + \hat{\beta}_{gc}} \frac{Y_{gc}}{s_c} + \frac{\hat{\beta}_{gc}}{s_c + \hat{\beta}_{gc}} \hat{\mu}_{gc}$$

   In our code, we keep default parameters of cell size normalization factor of 1 and set the number of cores to 12.

Although we do not expect downstream results to vary too much, it would be of interest to see how differential expression and gene set enrichment analysis would differ between the original and imputed datasets.

2.2 Quality Control

We did not conduct too much quality control as the nature of the data allowed us to conduct analysis fairly similar to what they did in the paper. However, the datasets were very large and conducting analysis such as clustering, differential expression, and gene set enrichment analysis consumed too much time and often we did not have enough computing power to conduct these analysis. Instead, we took a random sample of 3000 cells for each of the six mice but kept all the podocyte cells to improve power in differential expression. We then further subsetted the data using the percent of mitochondrial genes in each cell. In our quality control step, our goal was to retain as much information as possible while also making our analysis more computationally feasible.

2.3 Batch Correction

The first step we did was perform data integration, we splitted the dataset into a list of six seurat objects. The method that we did this was through the SplitObject function on the seurat object that was created from the metadata. We then normalized the seurat objects and selected features that are repeatedly variable across datasets for integration through the FindIntegrationAnchors function. Finally to create the integrated data assay we used the IntegrateData command on the combined seurat objects.

2.4 Dimensionality Reduction

After integrating all six datasets together, we performed dimensionality reduction using Seurat's RunPCA with 30 principal components (PCs). This PCA was based on the top 2000 variable genes. We found that the first 15 PCs captured most of the variability in the data, so when running Seurat's RunUMAP and RunTNSE, we set the dims parameter to 1:15. We compared the UMAPs with and without SAVER imputation. Though ideally we would have also generated UMAPs using MAGIC imputation, the extremely large size of the generated MAGIC count matrices and time limitations prevented us from doing so.

2.5 Clustering

We used Louvain clustering on the integrated data, implemented in Seurat as FindClusters. We used all default settings except for resolution, which we set to 0.2 after experimenting with several values. This resolution produced 14 clusters, which we then matched to the celltypes assigned by Sidhom et. al., observing high concordance between the two groupings despite using only a subset of the cells in the count matrix. Lower resolutions produced fewer clusters, and higher resolutions produced more clusters, but we used a resolution of 0.2 primarily because it produced nearly identical clusters to those in Sidhom et. al.

2.6 Differential Expression

To determine which genes were most strongly associated with kidney disease, we conducted differential expression (DE) analysis. We switched back to the unintegrated data stored in the "RNA" assay of the Seurat object. Though integration was valuable in the dimensionality reduction and clustering steps, the procedure introduces dependence between data points, violating the assumptions of most statistical tests used for DE analysis. Admittedly we then reintroduce potential batch effects into the data, but we preferred this to violations of the assumptions of tests. We also subsetted the Seurat object to include only the podocyte cells in the DE analysis. Using only podocyte cells helped narrow the focus of this section to the key celltype of interest when studying kidney disease and also aided in computational speed.

To implement this analysis, we used Seurat's FindMarkers, setting ident.1 to "CTRL" and ident.2 to "KDKD"—labels set in the metadata element of the Seurat object corresponding to control and kidney disease samples, respectively. We also adjusted the p-values using the Benjamini & Hochberg procedure—a less conservative method than the Bonferroni correction—and sorted by adjusted p-value.

2.7 Gene Set Enrichment Analysis

While the above DE analysis identified individual genes that were significantly associated with kidney disease, we also sought to determine the biological pathways associated with kidney disease. To this end, we conducted gene set enrichment analysis (GSEA) as introduced by Subramanian et. al. We accomplished this using the "fgsea" package in R and the collection of gene sets found in the UC San Diego-Broad Institute gene set database. The specific set of pathways we used can be obtained at this website: https://www.gsea-msigdb.org/gsea/msigdb/mouse/genesets.jsp?collection=CP

We obtained a full list of the 18,945 genes ranked by

$$-\log10(\text{pval}) \cdot \text{sign}(\text{avgLogFC}),$$

chosen to incorporate both the significance and direction of the association (up/down-regulation). Because of the way we set up the FindMarkers function to obtain the ranked list of genes, a negative rank metric indicated higher expression in kidney disease mice and a positive rank metric indicated higher expression in control mice. We found this to be intuitive, since a negative rank metric corresponded to an undesirable phenotype.

We then ran GSEA using the function “fgsea” with eps set to 0.0, minSize set to 15, and maxSize set to 500. Setting eps to 0.0 removes the boundary for calculating the p-values, and minSize and maxSize set boundaries on the size of gene sets in the set of pathways to consider. We created an enrichment plot for the pathway with the largest absolute normalized enrichment score and identified the leading edge subset of genes for this pathway. Finally, we created a table of the top 6 independent pathways identified using fgsea’s collapsePathways function.

3 Results

3.1 Dimensionality Reduction

We first generated a UMAP plot using the integrated non-imputed Seurat object, color-coding by mouse ID. We see almost complete overlap between mice, showing that the integration step worked and that the data do not cluster by mouse.

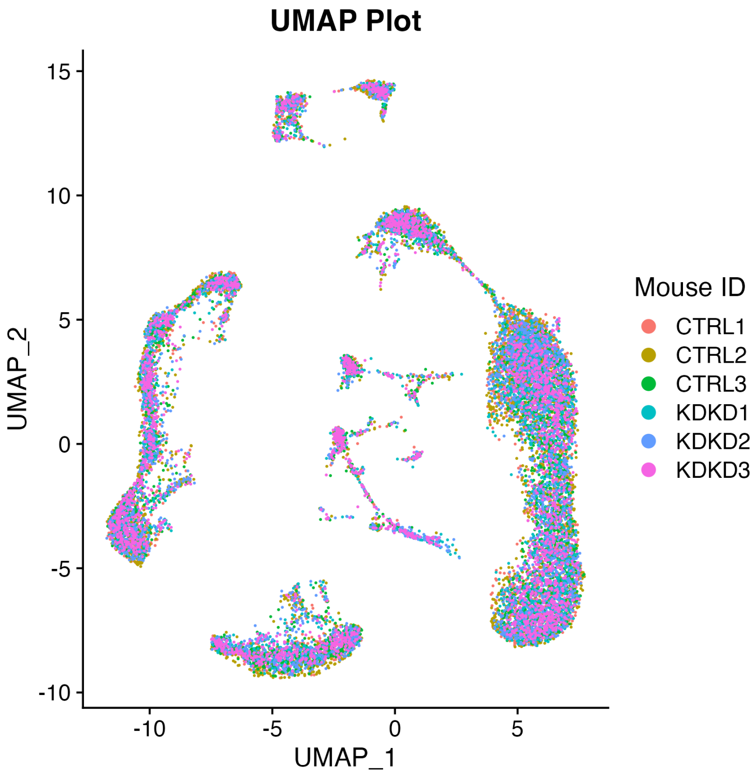
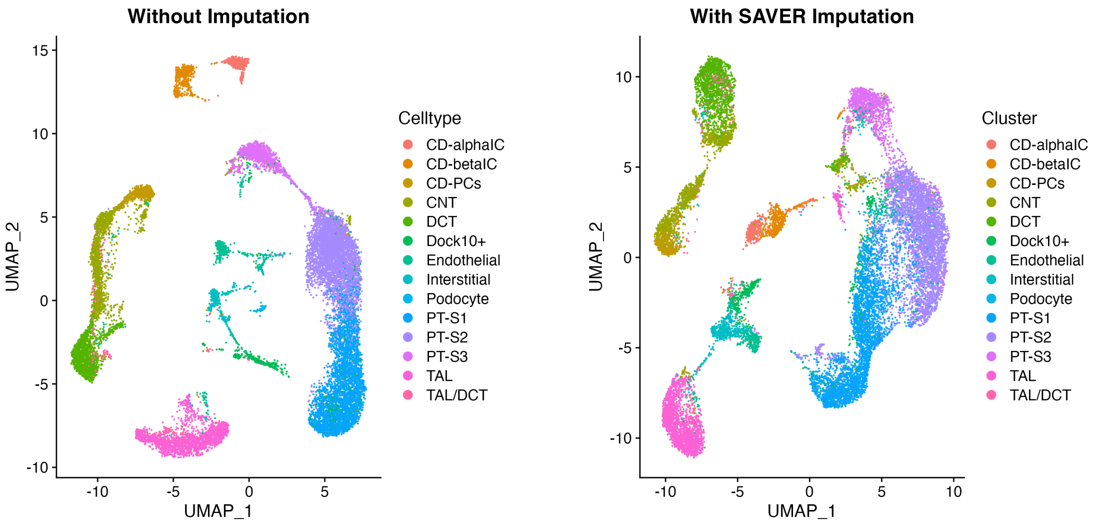


Figure 1: UMAP color-coded by mouse ID

We then color-coded by celltype and compared the UMAP generated using non-imputed data to the UMAP generating using the SAVER-imputed data. Overall the results are quite similar, but we do see slightly better separation between the PT-S1 and PT-S2 cells using the imputed data. Because the UMAPs were quite similar, however, we chose to proceed with the non-imputed data for the downstream analysis for the sake of efficiency. The non-imputed Seurat object was roughly



(a)

(b)

Figure 2: UMAPs color-coded by celltype

300 Mb large, whereas the SAVER-imputed Seurat object was roughly 2.7 Gb large, so it would have taken much longer to run our code using the imputed object.

Lastly, we generated a t-SNE plot using the non-imputed Seurat object and color-coding by celltype.

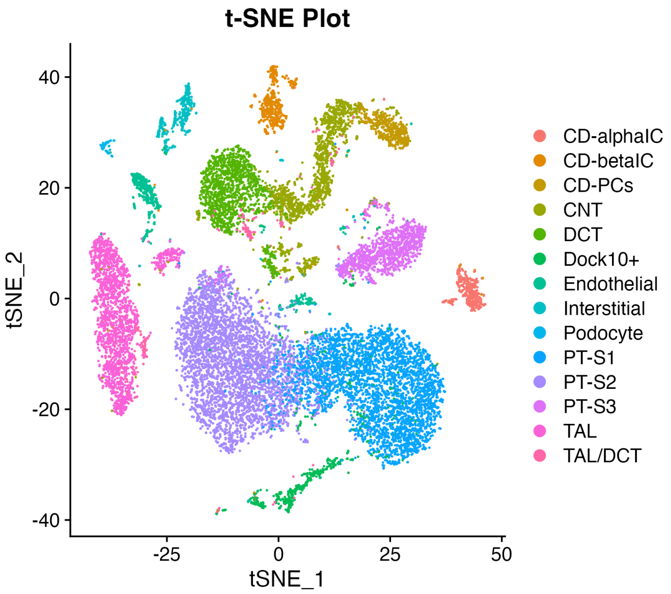
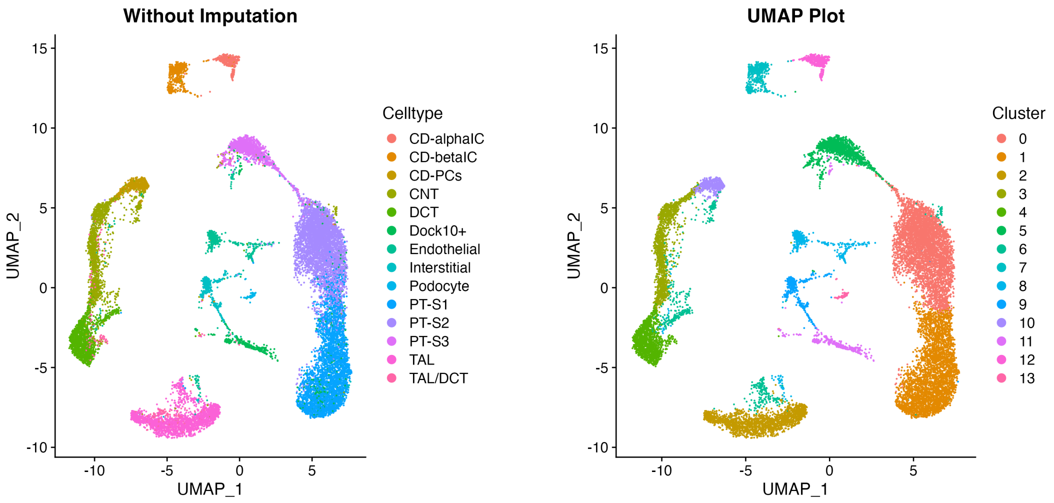


Figure 3: t-SNE color-coded by celltype

We observe separation between celltypes, but the separation is not as clear as in the UMAPs.

3.2 Clustering

After using FindClusters as described in the Methods section, we compared the celltype groups given by Sidhom et. al (fig. 4(a)) to our assigned clusters (fig. 4(b)) by a visual examination of the UMAP plots. Ignoring the difference in colors, we observe high concordance between groupings.



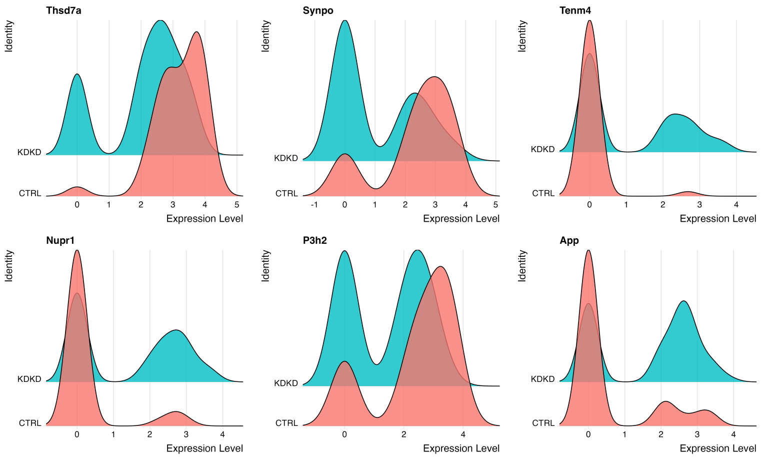
(a)

(b)

Figure 4: UMAPs color-coded by celltype and assigned cluster

Taking this high agreement into consideration, we decided to proceed using the annotated celltypes from Sidhom et. al. rather than our own clusters, as the authors were able to annotate celltypes with much more accuracy and expertise than us.

3.3 Differential Expression



Gene	Regulation
Thsd7a	↓
Synpo	↓
Tenm4	↑
Nupr1	↑
P3h2	↓
App	↑

Figure 5: Statistically significant genes associated with kidney disease in mice

Figure 5 shows the six genes we identified as significantly associated with kidney disease within the group of podocyte cells. These genes are Thsd7a, Synpo, Tenm4, Nupr1, P3h2, and App. Figure 5 also indicates whether the gene is upregulated or downregulated in kidney disease mice. We observe that all expression distributions here are bimodal. Interestingly, for the upregulated genes, the two peaks of the distribution occur at roughly the same expression level in control and kidney disease mice, but it is the mass at these peaks that changes. In this way, kidney disease can be considered to "turn on" these upregulated genes at a specific expression level. In contrast, the downregulated genes have roughly equally high peaks in the kidney disease, but the peaks occur at lower expression levels in kidney disease mice. Thus kidney disease seems to have the effect of "dimming" these downregulated genes at lower expression levels.

3.4 Gene Set Enrichment Analysis

For the last stage of our analysis, we conduced GSEA and identified the biological pathways most highly associated with kidney disease in mice. The pathway with the most negative normalized enrichment score was the eukaryotic small ribosomal subunit (40S), meaning this was the pathway containing the genes most positively associated with kidney disease. This makes sense, as the pathway contains mostly ribosomal protein encoding genes, The running sum of the enrichment score is shown in the plot below, and we can clearly see that the score is low because of the high concentration of genes towards the end of the ranked list of genes.

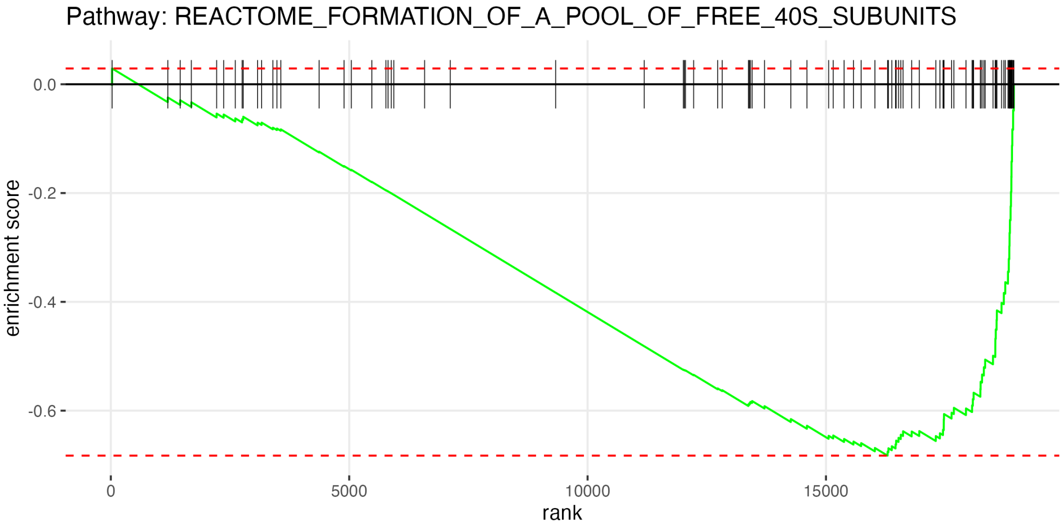


Figure 6: t-SNE color-coded by celltype

The leading edge subset contains the following genes: Rps19, Ubb, Rpl38, Rpl32, Rps24, Rps28, Rpl28, Rps8, Rps10, Rpl4, Rpl22, Rplp0, Rpl13a, Rps11, Rps6, Rps29, Rpl23, Rps15, Rpl6, Rpl35a, Rps16, Rps17, Eif2b4, Rpl36a, Rpl31, Rpl18, Rps23, Rpl10, Rpl5, Rpl19, Rps15a, Eif3g, Eif2s3x, Rps14, Rpl8, Eif2s2, Gm2000, Rpsa, Rps9, Rplp1, Eif4g1, Eif3c, Rpl26, Rplp2, Eif3i, Rps3a1, Rps4x, Rpl39, Rpl37, Rpl34, Rpl22l1, Eif5, Rps13, Rpl29

Finally, we present the top six pathways obtained by inputting our fgsea result into fgsea's collapsePathways. Of note is that all pathways have a negative normalized enrichment score, meaning all pathways are positively associated with kidney disease. This could be due to the set of pathways we decided to use for GSEA. It would have also been informative to determine which pathways were most negatively associated with kidney disease.

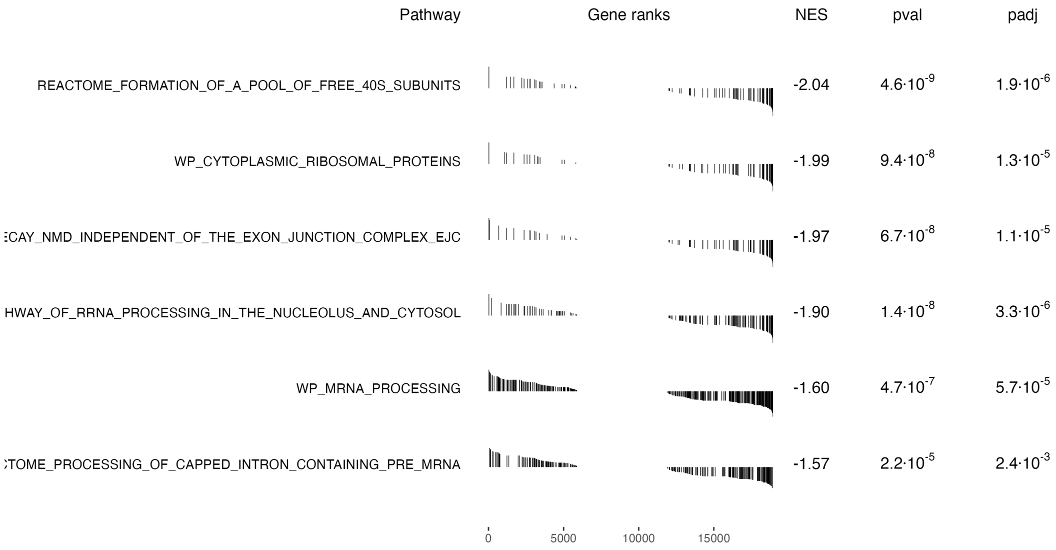


Figure 7: t-SNE color-coded by celltype

4 Discussion

In the completion of our analysis we were able to successfully execute all of the above methods. Our overarching goal was to identify differentially expressed genes in mice with kidney disease compared to the control group of mice without kidney disease. We were able to identify the top marker genes and biological pathways associated with kidney disease. The top pathway was the small ribosomal subunit (40S) which included mostly ribosomal protein genes. However, we were left with one question regarding this pathway and that is if this is a true biological effect or simply a result of higher expression of these genes.

5 Future Work

In order to retain all of the podocyte cells in the datasets, for the purpose of matching the results that were generated in Sidhom et. al., imputation was completed before quality control. For future work, we can consider reversing the order of imputation and quality control and see how that affects our results. We could also further refine imputation and conduct downstream analyses using imputed count matrices. Since this study was done with an emphasis on the podocyte cells we could also study the celltype-specific pathways in non-podocyte cells. The studies that were done on the mice samples could be further studied to determine if we could generalize these results to human samples.

6 Data and Code Availability

7 References

1. van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion.Cell174.

2. Huang, M., Wang, J., Torre, E. et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods 15, 539–542 (2018). https://doi.org/ 10.1038/s41592-018-0033-z