Analysis of Fetal Kidney Differentiation from Progenitor Cells using Single Cell Nuclear RNA-Seq

Guilherme de Sena Brandine

March 29, 2017

The transcriptome of 1600 fetal kidney cells were assembled by using molecular (UMI) and cellular barcode filtering of a drop-seq run of the dissolved tissue. For each individual cellular barcode (corresponding to a single cell) we had the number of UMIs mapped for each gene in the UCSC human annotation (hg19). These counts were used to further attempt to identify the cell types and the driver genes that characterize differences between said cell types.

Initially, cells were normalized using the sum factors method proposed by Lun et al. This method accounts for potential dropouts in the reverse transcription phase and potentially different amplification differences and total amount of RNA in each single cell. Although Drop-Seq data is generated in a single PCR run, there may be intrinsic biases during amplification for specific genes or cells, so normalization is a fundamental step to put transcriptomes in a common baseline.

We then used Principal Component Analysis on the normalized dataset and plotted the cumulative standard deviation (sum of standard deviation of the first principal components) to visually infer how many PCs should be used for further analysis. We chose to keep the first 50 PCs based on the cumulative sum of the PC's eigenvalues

We further used the approach proposed by Shekhar et al to identify cell subpopulations by clustering the k-nearest-neighbors graph, where the distance measure used was the Euclidean distance between the transcriptome vectors. This method minimizes the Louvain modularity in a graph where edge weights between any pair of cells is given by the ratio between number of common nearest-neighbors between the two cells and the size union of both cells' neighbors (Jaccard index). This approach has been shown to be more robust to different parameters (in this case, the number of neighbors) than other commonly used clustering methods for single cell data such as k-means, infomaps or DBSCAN. We chose k = 10 neighbors for this dataset, as this allows small subpopulations of outlier cells to be equally identifiable.

A total of 11 cell clusters were identified and their spatial distribution was plotted using t-SNE, a stochastic dimensionality reduction technique that models the high-dimensional data as a combination of two independent t-distributed random variables and optimizes the variable parameters to minimize the Kullback-Leibler divergence between the high and low dimensional signals. If the high-dimensional clustering using the principal components makes sense, then we should expect cells in the same cluster to have similar t-SNE coordinates, which is confirmed by figure XXXX. The fact that some clusters intersect, that is, some outlier cells are spatially located closer to another cluster rather than the one it belongs to, suggests that some pairs of clusters may be correlated by a continuous process of differentiation.

To further understand what are the driver genes that characterize the identity of each cluster, we performed a differential expression test between each individual cluster and the rest of the dataset, also as proposed by Shekhar, where we model the number of transcripts in the raw data as a binomial distribution and calculate a p-value under the hypothesis that the read count inside the cluster can be explained by the binomial parameters estimated by the data outside of the cluster. This gives a sorted list of genes whose up or down regulation characterizes the cluster phenotype.

Motivated by the strong intersection between clusters in the t-SNE plot, we interrogated what would be the most likely order of transition between cells, and if the clusters previously found correlated to specific developmental points in differentiation. For this we used the method proposed by Haghverdi et al to infer such correlations. The method uses diffusion maps, that is, a transition matrix between cells whose value is the Hellinger distance between two gaussian functions centered at the cell points, and uses the first two eigenvectors of this matrix to infer the most likely point in which a random walk would reach a specific cell. By plotting the first two eigenvalues, we see that the first eigenvalue identifies cluster 11 as an outlier, whereas both the second and third eigenvalues sort the cells in a coinciding way, both indicating that clusters 3,2,8 and 9 are sorted in this developmental order, as well as clusters 10, 6, 1, 5 and 4. Cluster 7 appears to represent either fully matured distal renal vesicles or distal pretubular aggregates, as evidenced by its overexpression of LEF1 and low expression of WT1.