Complete Methods

Conserved and divergent features of mesenchymal progenitor cell types within the cortical nephrogenic niche of the mouse and human kidney.

Lindström, N.O.[1*], Guo, J.[1*], Kim, A.D.[1*], Tran, T.[1†], Guo, Q.[1†], De Sena Brandine, G.[2†], Ransick, A.[1], Parvez, R. K.,[1] Thornton, M.E. [3], Basking, L. [4], Grubbs, B. [3], McMahon, J.A. [1], Smith, A.D.[2], McMahon, A.P.[1]

1. Department of Stem Cell Biology and Regenerative Medicine, Broad-CIRM Center, Keck School of Medicine, University of Southern California, Los Angeles, CA90089, USA.

2. Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California.

3. Maternal Fetal Medicine Division, University of Southern California, Los Angeles, California.

4. Department of Urology and Pediatrics, University of California San Francisco, San Francisco, California.

[*]These authors contributed equally.

[†]These authors contributed equally.

Corresponding author: Andrew P. McMahon, telephone: 323-442-7847, email: amcmahon@med.usc.edu

Complete Methods

Animals

All animal work was reviewed and institutionally approved by Institutional Animal Care and Use Committees (IACUC) at the University of Southern California and performed according to institutional guidelines. Timed matings were set up to recover embryos and neonates at the appropriate age. The Foxd1-GCE strain (B6;129S4-Foxd1tm2(GFP/cre/ERT2)Amc/J) was generated as previously described (Humphreys et al., 2010). The Rosa26tdTomato reporter line (B6.Cg-Gt(ROSA)26Sortm14(CAG-tdTomato)Hze/J) was obtained from JAX (Madisen et al., 2010). Heterozygous Foxd1-GCE animals were crossed with female Rosa26tdTomato homozygous females. Pregnant females were injected with 3mg Tamoxifen per 40g at E13.3 and kidneys collected at E14.5. The Six2TGC line was generated as previously described [3]. Male heterozygous Six2TGCs animals were crossed with female Swiss Webster mice and embryos collected at E16.5.

Image and sample quantification

2D immunofluorescent analyses

Frozen and sectioned samples were stained as described previously (Lindström et al., 2017a) to detect SIX2, SIX2, LEF1, FOXD1, and CITED1. KRT8 and β-laminin were used as structural markers to determine the location of the ureteric epithelium and nephrons. Images were captured with a 63x objective on a Leica SP8. Data were captured as 8-bit images. IMARIS 8.2 (Bitplane) was used for quantification of nuclear antibody signals. The Spot function was used to manually add circular spots to mark all nuclei on the image frame using DAPI-highlighted nuclei as a reference. Due to the convoluted shape of nuclei in 2D sections and the circular shape of the spot function, we used multiple smaller spots to represent single nuclei to ensure accurate quantitation and coverage across the nuclei. Spots were grouped into three cell populations: (1) cap mesenchyme ($SIX2^+$ cells), (2) cortical interstitium ($FOXD1^+$ cells), (3) all other cells. To compare the mean intensity of spots we first normalized the mean intensity values taking into consideration the background signal and the maximum signal for each channel. To do this we measured the intensity for each channel throughout all spots and identified the lower 5^{th} percentile intensity mean (background), as well as the maximum value. Each spot's intensity mean was thereafter normalized as follows:

$$\left(\frac{Intensity\ mean\ of\ spot - 5th\ percentile\ intensity}{Maximum\ intensity - 5th\ percentile\ intensity}\right)x100$$

This transforms the intensity mean of each spot onto a 0-100 scale with the 5^{th} percentile equaling 0 and the maximum being 100, respectively. To plot the normalized intensity of spots against their position within the cap

mesenchyme population we marked the most cortical point of the cap mesenchyme and utilized this as point 0. A line was extended in a medullary direction parallel to the ureteric epithelium around which the cap mesenchyme was located.

RNA-sequencing data

All RNA sequencing data is provided at the Gene Expression Omnibus; GEO accession numbers: GSE102378 (mouse RNAseq data), GSE102230 (human RNAseq data), (human single-cell RNA seq data: GSE102596). Full details for the number of samples can be found for each submission. In brief: MARIS or conventional RNA sequencing was performed on 5 kidneys for SIX2 MARIS, 2 kidneys for MEIS1/SIX2 MARIS, 3 kidneys for mouse Six2 MARIS, 3 kidneys for mouse Six2-GFP sequencing, 2 kidneys for mouse Foxd1 sequencing. One kidney was dissociated for single-cell RNA sequencing. The GEO submission comprises 39 RNA-seq libraries for RNA-seq and 1 multi-cell library for single-cell RNA sequencing.

MARIS Staining and FACS

The MARIS staining and FACS procedure was performed as described in Hrvatin et al., 2014 [45] with the following modifications. Human and mouse cortical nephrogenic zone cells were digested from E16.5 embryonic mouse kidneys or 13-18 week fetal human kidneys using 10 mg/ml pancreatin (Sigma, P1625) and 2.5 mg/ml collagenase A (Roche, 11 088 793 001) enzyme mixture and filtered through 40 µm filter (BD Falcon 352340) as described in [70]. Cell fixation, washing, permeabilization, and centrifugation were performed as described in Hrvatin et al., 2014 [45] using the following solutions with all subsequent steps performed at 4° C. Fix buffer: 4% PFA (Electron Microscopy Sciences), 0.1% saponin (Sigma-Aldrich 47036) in molecular grade PBS (Ambion) supplemented with 1:100 RNasin Plus RNase Inhibitor (Promega, N2615) Wash Buffer: PBS containing 0.2% BSA (Gemini Bio-Products), 0.1% saponin (Sigma-Aldrich), 1:100 RNasin Plus RNase Inihibitor. SIX2 Primary antibody (Mybiosource, MBS610128) and MEIS1/2/3 Primary antibody (Active Motif, 39795) staining of cells at 1:5000 dilution was carried out while rocking for overnight at 4° C in staining buffer containing PBS with 1% BSA, 0.1% saponin and 1:25 RNasin Plus RNase Inhibitor. Cells were washed and stained with donkey anti-rabbit Alexa 488 (Thermofischer, A-21206) and goat anti-mouse IgG1 555 (Thermofischer, A21127) secondary antibody for 45 minutes. Subsequent washing and FACS sorting were performed at a concentration of 5-10 M cells/ml with sort buffer containing PBS, 0.5% BSA, and 1:25 RNasin Plus RNase Inhibitor. Cells were sorted on the FACSAria I and II (BD Biosciences) using FACSDiva software. Sorting gates were set with reference to negative controls with no primary antibody stain. The sorting efficiency was maintained at above 90%. Cells were collected in tubes that were coated with a small amount of sorting buffer.

For FACS sorting of Six2TGC and TdTomato positive cells, the mouse kidneys were dissected and dissociated in the same enzymatic solution as described above. The cells were not fixed but instead immediately FAC sorted for GFP or tdTomato.

RNA isolation of MARIS

The RNA isolation was performed as described in Hrvatin et al., 2014 with the following modifications. FACS collected SIX2+, SIX2-, MEIS1+SIX2+, MEIS1+SIX2-, and MEIS1-SIX2- cells were pelleted by centrifugation at 3000 g for 10′ at 4° C. Total RNA was isolated using the RecoverAll Total Nucleic Acid Isolation kit (Ambion, AM1975), starting at protease digestion step with protease incubation time of 1 hour at 50°C and inactivated at 80° C for 15 minutes. Cell lysates were frozen at −80° C overnight and extracted for RNA according to the manufacturers recommended protocol.

For the non-fixed mouse cells, the RNA was isolated as previously described (Lindstrom et al., 2017).

RNA-seq analysis

mRNA-Seq libraries were synthesized with Kapa stranded mRNA-Seq kit, and were sequenced on Illumina NextSeq500 platform at USC Epigenome Center. All mRNA-Seq reads were aligned to the mouse or human reference genome (mm10 or hg38) using the TopHat2 (Trapnell et al., 2009). Quantification of RNA-Seq reads to generate RPKM was performed by Partek Genomics Suite software, version 6.6 (St. Louis, MO, USA). TPM was calculated by dividing RPKM value by ratio of sequencing reads from the corresponding library that were mapped to exon regions of the genome. We identified differentially expressed genes as those satisfying the following 3 criteria: 1) p value smaller than 0.05 from statistical tests performed by DESeq2 [71]; 2) more than 3-fold difference of average normalized read counts between the groups compared; 3) average TPM more than 5 in at least one of the groups. Gene ontology (GO) analysis was performed using PANTHER classification system [72]; http://pantherdb.org/). We ranked the relevance of GO terms by fold enrichment of number of observed genes over number of expected genes. The GO terms with binomial p value more than 0.01 were omitted due to statistical insignificance. We analyzed the variability between all the MARIS RNA-Seq data for NPCs from both the huSIX2+MARIS and the huSIX2+/MEIS1+ MARIS data and found that correlation between samples was high in 6/7 samples (R^2 range 0.929 to 0.977). Replicate 1 from the huSIX2+MARIS displayed lower correlation to the other samples (R^2=0.715 to 0.778). This variability may have arisen at various points: 1) each replicate RNA sample was extracted from a different human fetal kidney with no known, but presumed genetic variability, in addition to samples originating from a range of close developmental stages; 2) RNA from replicates 2-5 for the huSIX2+MARIS exhibited lower quality, as measured by RNA integrity, due to difficulties in library construction/sequencing consequent to low RNA content thereby

indicating that replicate 1 may be higher quality; 3) The total amount of mapped reads from replicates 2-5 are approximately 25% less than replicate 1, which might have contributed to decreased sample complexity.

Single-cell sequencing

Cell preparation

Cells were dissociated as described for the MARIS protocol from a week 16 kidney and live cells sorted by FACS using DAPI (Thermo Fisher Scientific) and DRAQ5 (Thermo Fisher Scientific) to select against dead cells and for live cells, respectively. 78 % of cells were live and intact, indicating robust isolation methods. Seven thousand live cells were input into a 10X Genomics Chromium device expecting the capture of 4000 cells. Illumina ready sequence-able libraries were then generated using the 10X Chromium single cell 3' RNASeq protocol. Subsequently, sequencing was carried out on the Illumina NextSeq 500/550 platform with the goal of obtaining at least 50,000 reads per cell. 3731 valid barcodes ('cells) were recovered after filtering and UMI counting.

Sequence Mapping

Mapping was performed using the CellRanger software version 1.3.1 through the CellRanger count command. We used STAR version 2.5.1b to map the second end of the FASTQ reads to the human genome version GRCh37.p13 and uniquely mapped reads were counted using the Ensembl GTF annotation as reference. A total of 70.1% of the reads had unique mapping, which corresponded to a total of 3,731 valid barcodes.

Quality Control

To filter out potential doublets and low quality cells we calculated 3 quality measures for each individual cell:

1. the Good-Turing estimate of observed expression [73] given by $S = 1-n_1/N$, where n_1 is the number of genes with one mapped read and N is the total number of reads in the cell. Saturation ranged from 40% to 100%. We chose to keep only cells with $S > 0.6$.

2. The percentage of mitochondrial gene expression. We filtered out any cell with more than 5% of the total expression mapped to genes annotated to come from mitochondrial DNA.

3. The deviation from a read-UMI fitted curve: We expect the number of observed genes to increase linearly with the number of reads for cells that have not attained full saturation. We fitted a line between the number of non-zero genes and number of reads and filtered out cells whose residuals were more than 5 standard deviations from the line.

A total of 2,750 cells were kept after filtering through these 3 criteria, indicating 73% of sequenced cells were of high quality.

Analysis of week 16 scRNAseq dataset

We used the Seurat R package [55] version 1.4 for further analysis of the remaining cells. The *MeanVarPlot* function with default parameters was used to find a subset of genes whose variability is above the expected technical noise. We found 582 such genes, which were further used for Principal Component Analysis (PCA function). To find significant PCs, we used the *JackStraw* test [74] and kept the first 24 PCs, which had $p < 10^{-4}$. These PCs were used for clustering using the *FindClusters* function with k = 30 nearest neighbors. We found 12 clusters, whose identities were further validated by the *AssessNodes* function, which builds a random forest classifier for each split node in the cluster hierarchy. The highest out of bag error (OOBE) we found was 9%, which indicates that all clusters have a clear identity. Differential expression was performed with the likelihood ratio proposed by McDavid [75] and implemented in the *FindAllMarkers* function, in which genes inside a cluster are compared with the expression in all cells outside of the cluster. We set the minimum average difference between inside and outside clusters to 0.15 and no minimum average expression threshold.

Analysis of Cluster 4

We repeated the aforementioned procedures with only the subset of 318 cells that were assigned to cluster 4. We found 685 variable genes and the 4 first principal components to be statistically significant ($p < 10^{-4}$). Clustering with 4 principal components yielded 4 sub-clusters whose maximum OOBE was 6.5%.