**Data Set Introduction:**

NanoString GeoMx DSP dataset of diabetic kidney disease (DKD) vs healthy kidney tissue. Seven slides were analyzed, 4 DKD and 3 healthy. Regions of Interest (ROI) were focused two different parts of a kidney’s structure: tubules or glomeruli. One glomerulus ROI contains the entirety of a glomerulus. Individual glomeruli were identified by a pathologist as either behaving relatively healthy or diseased regardless on if the tissue was DKD or healthy. Tubule ROIs were segmented into distal (PanCK) and proximal (neg) tubules. While both distal and proximal tubules are called tubules, they perform very different functions in the kidney. Segmented areas of an ROI are called Areas of Interest (AOI).

**Image Origin:** top left corner with the y-axis extending down and the x-axis extending right as per image editing software

**Limit of Quantitation (LOQ) Calculation**:

(GeoMean of 139 Negative Probes) x (SD of 139 Negative Probes)2

**Filtering Methodologies:**

ROI QC:

1. removed ROIs with Raw Reads < 5000
   1. Raw Read count indicates coverage of ROI sequencing
2. removed ROIs with Sequencing Saturation < 50%
   1. Sequencing Saturation indicates confidence in sequencing

Probe QC:

1. removed probes that failed two-tailed outlier test in >50% of ROIs
2. removed probes that fell below background in >99% of ROIs

**Description of files:**

All .txt files are tab delimited

**Start Here:**

1. Sample Annotations: Kidney\_Sample\_Annotations.txt
   * Annotations of ROIs (i.e. samples) in files 3-5.
   * .rda files contain these annotations in the NanoStringGeoMxSet variable.
     + kidney\_norm@phenoData@data
   * Explanation of columns:
     + **SlideName**: Name of slide
     + **ScanName**: Name of GeoMx scan used to match with image names
     + **ROILabel**: ROI label used to match with image names
     + **SegmentLabel**: segmentation of tubule ROIs in to distal (PanCK) and proximal (neg) tubules. Glomeruli segmentation is Geometric Segment.
     + **SegmentDisplayName**: unique identifier of an AOI
     + **Sample\_ID**: unique identifier of an AOI that matches DCC file, used in NanoStringGeoMxSet
     + **AOISurfaceArea**: area of AOI in μm2
     + **AOINucleiCount**: number of nuclei (cell estimate) in the AOI
     + **ROICoordinateX**: X coordinate for center of mass of ROI
     + **ROICoordinateY**: Y coordinate for center of mass of ROI
     + **Reads**: Sequencing values
     + **SequencingSaturation**: % non-unique reads, confidence in sequencing
     + **UMIQ30**: % bases >Q30 in UMI, sequencing QC metric
     + **RTSQ30**: % bases >Q30 in RTS sequence, sequencing QC metric
     + **disease\_status**: normal or DKD, disease state for tissue
     + **pathology**: pathological annotation of each glomerulus as relatively healthy or abnormal regardless of disease status, disease state of individual glomerulus
     + **region**: glomerulus or tubule, kidney structure
     + **LOQ**: limit of quantitation calculated using negative probes, estimation of background noise level
     + **NormalizationFactor**: 75th percentile normalization factor
2. Feature Annotations: Kidney\_Feature\_Annotations.txt
   * Annotations:
     + **RTS\_ID:**
       - RNA sequencing tag ID (another ID used for probe ID)
     + **TargetName:**
       - Target (i.e. gene) name
     + **ProbeID:**
       - a.k.a. ProbeName used for probe identification
     + **Negative:**
       - Boolean indicating if the probe is a negative control
3. Probe Expression: Kidney\_Raw\_BioProbeCountMatrix.txt & kidney\_probe.rda
   * Raw counts of targets and negative probes. All probes, not filtered, negative probes NOT collapsed. Only ROIs that passed QC are included.
   * Rows are probes and columns are samples
4. Target Expression: Kidney\_Raw\_TargetCountMatrix.txt & kidney\_target.rda
   * Raw target counts. All probes, not filtered. Targets with multiple probes get aggregated using the geometric mean across all probes for a target. In this panel, only the negative has multiple probes, and therefore is the only target count aggregated. Only ROIs that passed QC are included.
   * Rows are targets and columns are samples
5. Normalized Expression: Kidney\_Q3Norm\_TargetCountMatrix.txt & kidney\_norm.rda
   * 75th percentile normalized counts. Only ROIs and genes that passed QC are included.
   * Rows are targets and columns are samples

R Objects:

1. Kidney R Dataset: KidneyDatasetR.zip
   * R data package
   * Three GeoMxSet class objects which include sample, feature, and expression data within each object
     + kidney\_probe.rda (#3)
     + kidney\_target.rda (#4)
     + kidney\_norm.rda (#5)
2. Example Analysis: KidneyDatasetRVignette.html
   * Example analysis on the kidney dataset GeoMxSet object

Images:

1. Slide Images: ROI\_reports.zip
   * Scans and ROI images for 7 scans
   * File name structure using [Annotation Column Names] in brackets
     + Glomeruli
       1. [ScanName] /[ScanName] – [ROILabel].png
     + Tubule
       1. [ScanName] /[ScanName] – [ROILabel] – [SegmentLabel].png

**Example of Typical Analyses:**

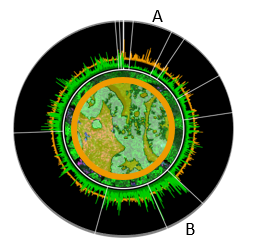
Cell Deconvolution:

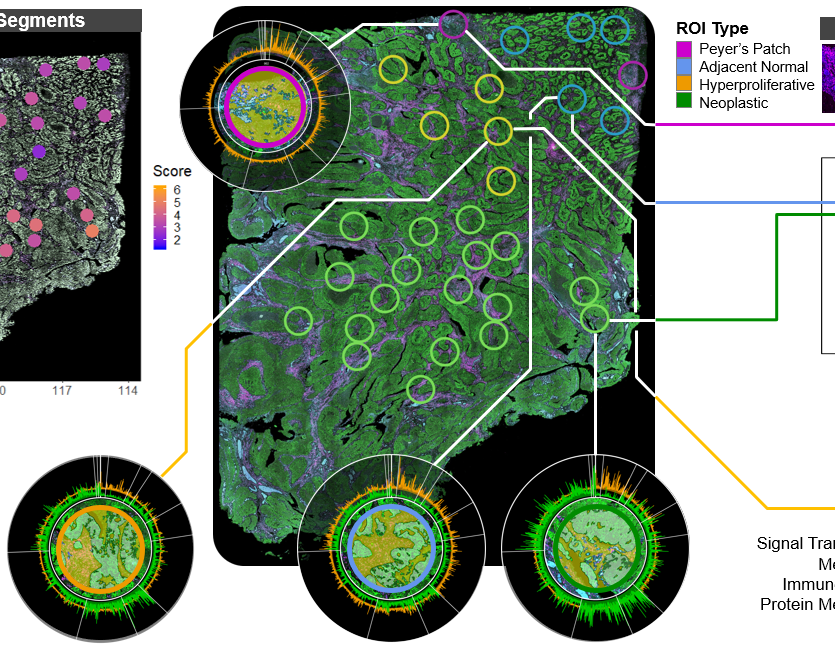
1. Decon Results: Kidney\_Spatial\_Decon
   * Results of [cell deconvolution](https://www.biorxiv.org/content/10.1101/2020.08.04.235168v1) analysis on the kidney dataset (cell type x AOI)
     + cell matrix expression in #11 was used to estimate the cell populations and abundance within an AOI.
   * Rows are cell types and columns are samples
   * Values are abundance scores of cell types in AOIs.
2. Cell Annotations: Cell\_Types\_for\_Spatial\_Decon.txt
   * Known cell types pertinent to the kidney dataset from previous single cell RNA-seq (scRNA-seq) data
     + Young, M.D., et al. (2018). Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science 361, 594–599. [Link to paper](https://science.sciencemag.org/content/361/6402/594).
     + Subset of Supplemental Table 2, Cluster info
   * Explanation of columns:
     + **ClusterID**: ClusterID from scRNA-seq datasets either Normal (N) or Immune (IN)
     + **Alias**: shorthand ID for cluster based on given Cell Type
     + **Data\_set**: scRNA-seq dataset cluster was derived from
     + **Number\_of\_cells**: number of single cells in each cluster
     + **Cell\_type1:** cell type of cluster label 1
     + **Cell\_type2:** cell type of cluster label 2, if necessary
     + **Cell\_type3:** cell type of cluster label 3, if necessary
     + **cell\_type\_specific:** specific cell type label
     + **cell\_type\_general:** generalized cell type label, used to bucket clusters
     + **cluster\_name:** Unique identifier of cluster, cell\_type\_specific & Alias
   * Marker genes in clusters were compared to canonical markers from literature.
     + See Table S3 in paper
3. Average Gene Expression: Young\_kidney\_cell\_profile\_matrix.csv
   * Average gene expression for each cell type cluster in #10
   * Rows are cell types and columns are samples
   * Values are gene counts expected in a cell type X.

Gene Set Enrichment Analysis (GSEA):

1. Single Sample GSEA Results: Kidney\_ssGSEA.txt
   * Single-sample Gene Set Enrichment Analysis (ssGSEA) analysis results (pathway x AOI)
     + known groups of genes that work concordantly to perform a biological function are surveyed for enrichment of over- or under-expressed genes within each sample.
   * Rows are known gene sets and columns are samples
   * Values are abundance scores of biological pathways in AOIs.

**Standard Analysis Performed on GeoMx Data**:

1. Dimensional Reduction
   1. PCA
   2. tSNE
   3. UMAP
2. Expression Circos plots across segmentation
   1. Expression is indicated by the height of the spikes
   2. Colors
      1. Green = Segment 1
      2. Orange = Segment 2
   3. Genes are arranged by pathway around the plot. Pathway A is enriched in Segment 2 where Pathway B is enriched in Segment 1.
   4. These plots can be overlayed on the image and compared across ROI types.



1. Cell Deconvolution using SpatialDecon
   1. Calculating the proportion of cell types in an ROI based on the gene expression
   2. Danaher, P., et al. (2020). Advances in mixed cell deconvolution enable quantification of cell types in spatially-resolved gene expression data. [Manuscript available on BioRxiv](https://www.biorxiv.org/content/10.1101/2020.08.04.235168v1)
2. Differential expression
3. Pathway analysis
   1. Gene Set Enrichment Analysis (GSEA)
   2. Single sample GSEA
   3. Over enrichment analysis (ORA)