Role of Data Standards in Quality and Harmonization

Metadata and Data Standards for NIDDK Research Data

June 27th, 2023

Sanjay Jain MD, PhD
Professor: Medicine (Nephrology), Pathology and Pediatrics
Director: Kidney Translational Research Center
sanjayjain@wustl.edu

Washington University in St. Louis
School of Medicine

HuBMAP
Human BioMolecular Atlas Program

KPMP

Objective: Create a high-resolution integrated single cell and spatial multimodal atlas of the human kidney in health and disease across lifespan

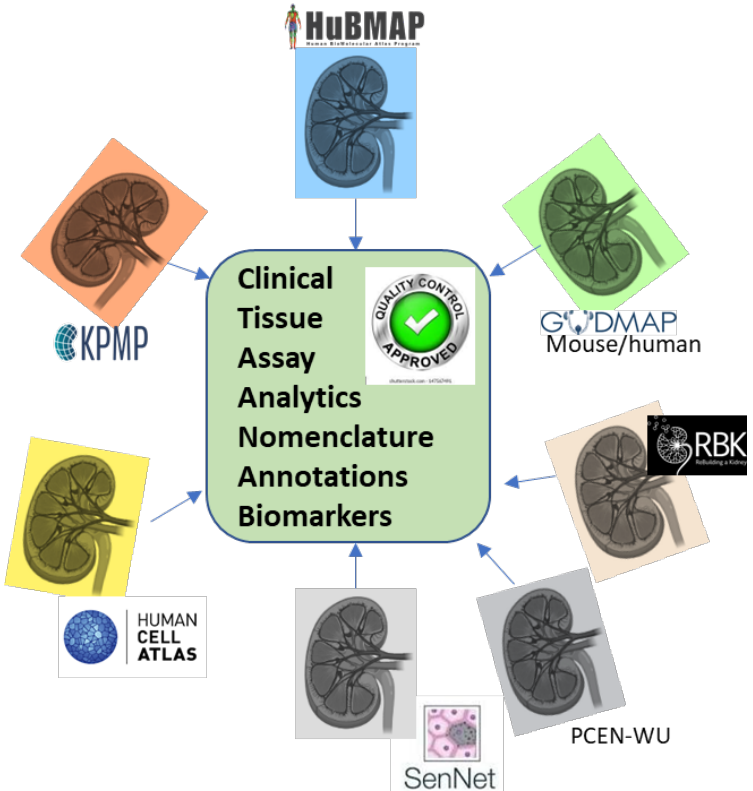
PCEN

A collection of morphological and molecular maps

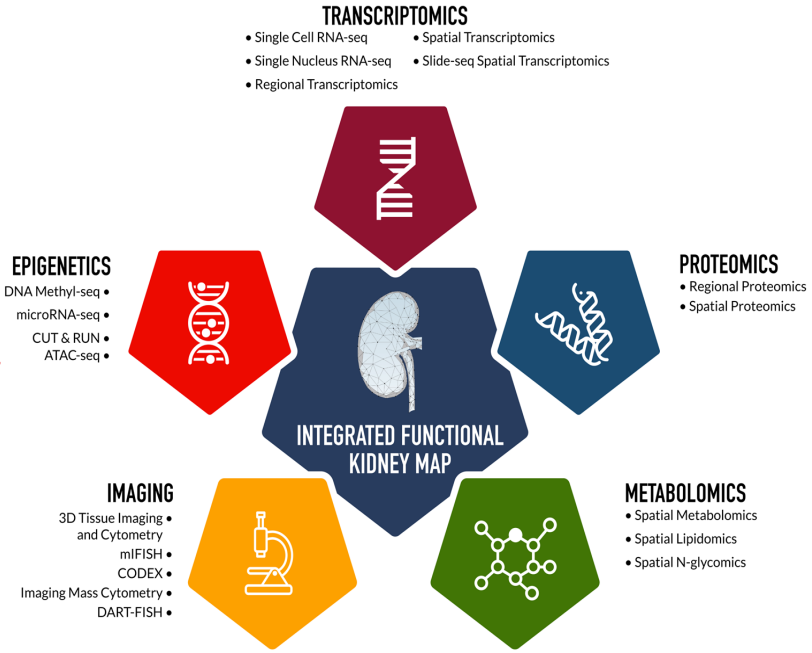
- ***understand vital functions of the kidney***
- ***create knowledge to prevent kidney injury and recover from it***

Overview

- Quality assurance and control
  - Follow the tissue pipeline
- Assay / Data Harmonization
  - Bridges across data types and rigor
- Examples showing cross-species integration
  - Cellular diversity and injury time course



Many assays,
Many institutions



Metadata and Nomenclature: Quality Control

Process

Choices

Challenges

Tissue

Isolation

Separation

RNA libraries

Sequence

Expression matrix

Data analysis

Validate

Share

Resection, donor, biopsy

Mechanical Enzymatic

FACS, Manual, Limited dil Automated

Plate Droplets (barcodes, UMI)

Platforms depth, cells, 3`, 5`

Normalization, Filtering

Commercial PCA, tSNE, UMAP

ISH, IF, Orthogonal

GEO, dbGAP websites

Metadata, Limited sample

Integrity, Viability, Stress Logistics - time

Cell loss Doublets Ambient RNA

Dropouts Failed amp

Standardization Cost

Batch effects, data quality

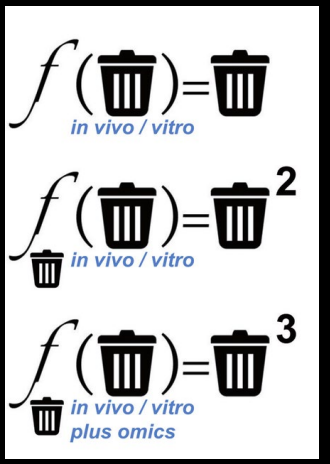
Variability Clustering Annotation

Reproducibility, availability

Big Data, Visualization

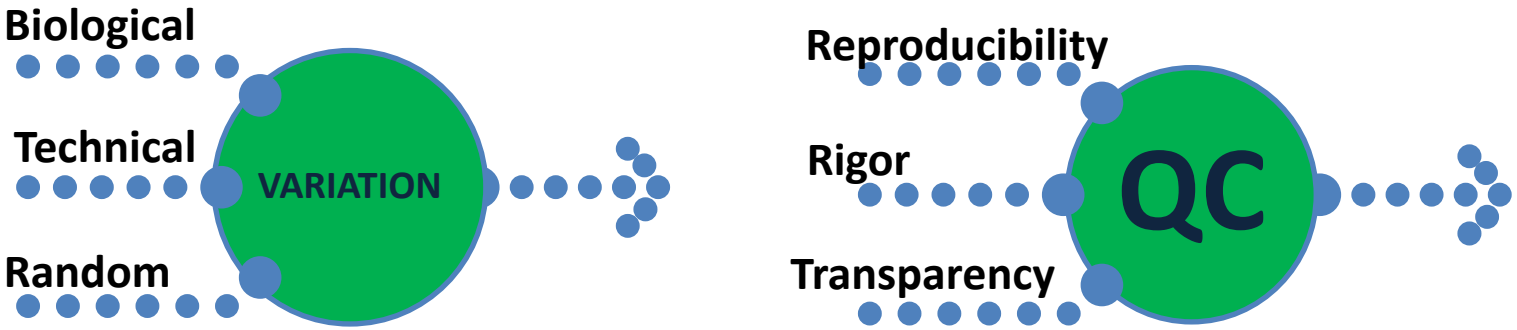
QA/QC at each step

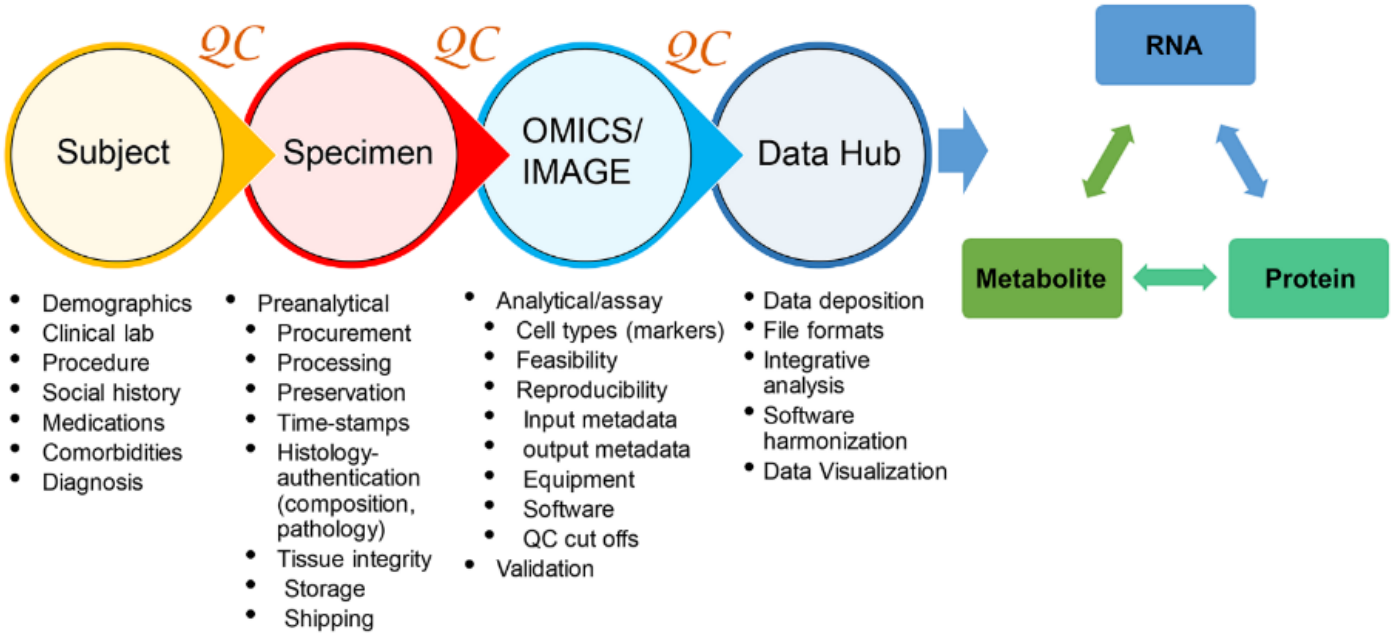
F.A.I.R



Jain, Current Opinion in Nephrology & Hypertension, 2023

Process: QC Pipeline, Standardization



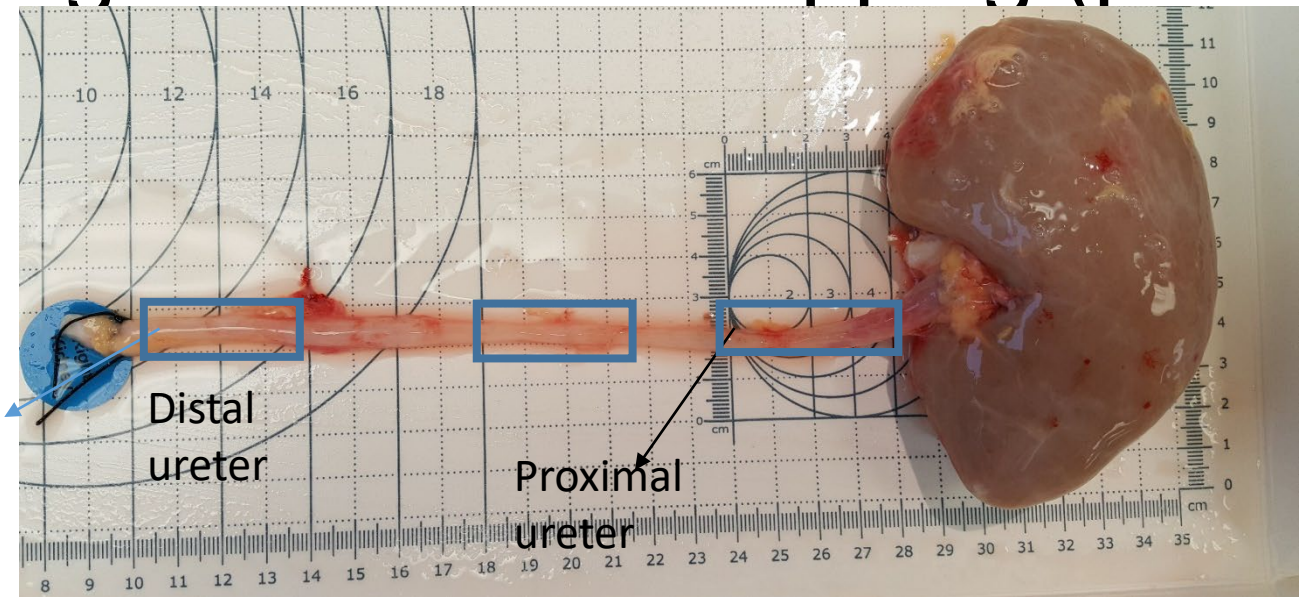


QUALITY CONTROL

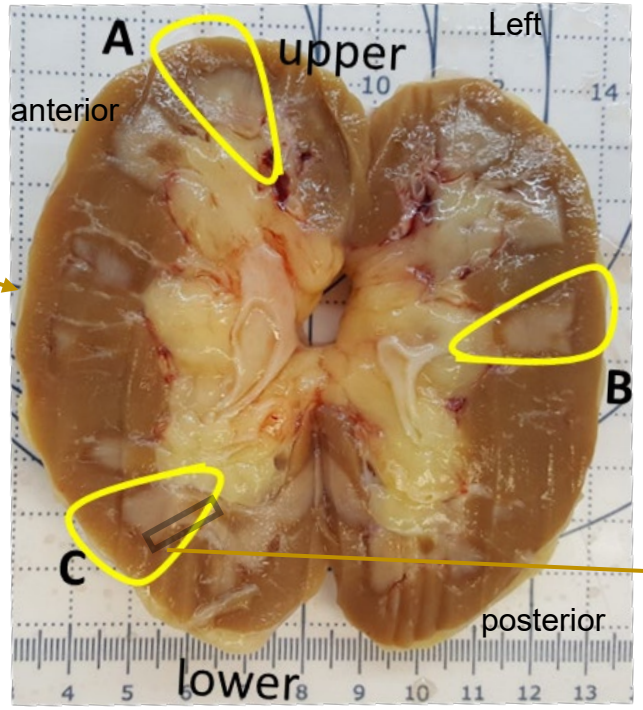
El-Achkar…Jain, AJP, Phys Genomics, 2021

KPMP

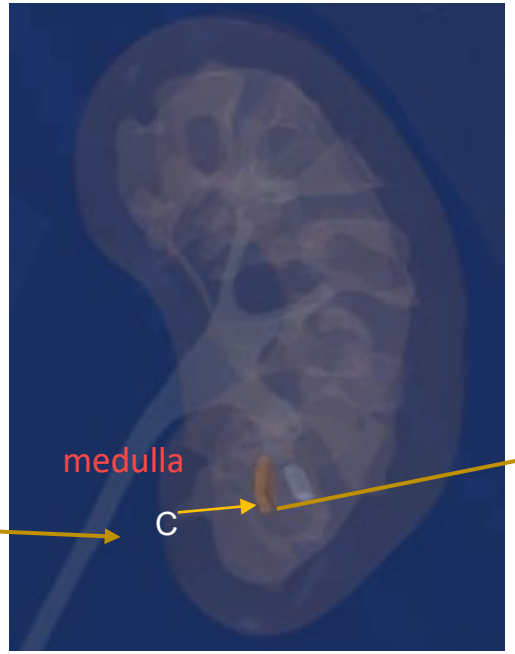
Tissue: Registration and Mapping (provide location)

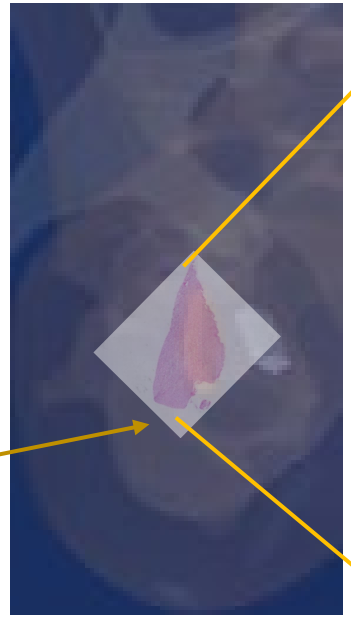


RUI

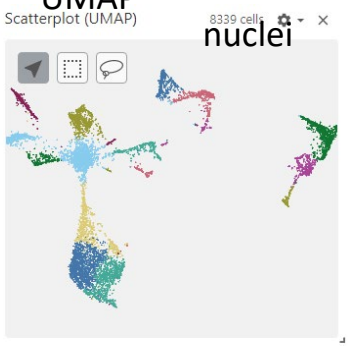


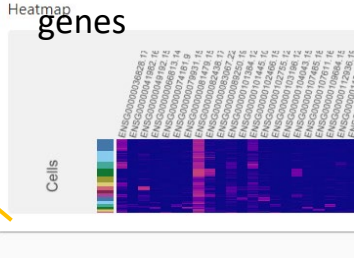
RUI





UMAP





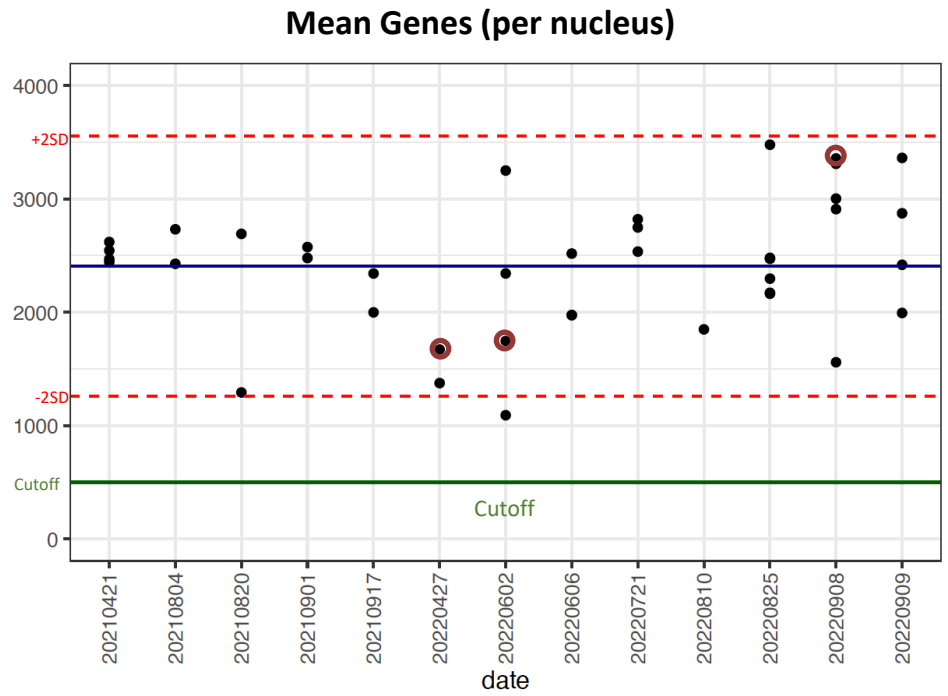
Assay Integration and Data Harmonizaton

Assay and Analytic Standardizations: Metadata

Participant clinical data	Demographics, comorbidities, social history (alcohol, tobacco, rec drugs), active medication categories (diurectics, ACE/ARB inhibitors), general health prior to current circumstances							
Tissue quality	gross pathology, microscopic composition (%cortex, % outer and inner medulla, papilla, non kidney components), age appropriate glomerulosclerosis, eGFR > 60, HbA1C <6.5, mild to no tubulointe							
Tissue data	warm ischemia (<30min Nx, Bx < 10min, Deceased donor <1h down time); cold ischemia time (<36h, perfused), Cold thaw cycles (-80 to -23 C), Storage temp tracking, quantity usage log							
Technologies	**snRNAseq**	**snATACseq**	**smFISH**	**DART-FISH**	**10X Visium**	**CODEX**	**3D IF**	**Light sheet (LS)**
Assay	nuclei > 1000; cDNA > 20ng; library size 200 – 1000bp		Instrument callibration Pass		RIN>7, Tissue affixment without folds or overlap, library size >400 bp	Visual tissue quality assessed via brightfield imaging (10X), adjacent PAS staining. Microscope QA with ArgoLight based microscope assessments (lamp or laser stability, SNR, resolution), labelled antibody lot testing.		Ab test each lot on Nx using confocal IF
Sequencing / Imaging	Q30 score > 80%; > 60% reads mapped; phiX spike (mapping errors, phasing/rephasing, reagents)		Housekeeping gene feature counts	RNA staining intensity, backgroud baseline	Q30 score > 90%; Reads mapped to Exons >30%; image inspection	Signal quality (dynamic range), probe carry-over, artifact detection	Signal quality (dynamic range), spectral unmixing accuracy, SHG and autofluorescence imaging	Verification of clearing by confocal
Post sequencing cells, genes, clusters / imaging features	400>genes<7500 non mitochondrial	>1000peak count per cell<100,000	# features / mm2	10^4 rolonies / mm2	Reads >50% map to spots under tissue	Identifiying ~3-10k cells per 1 mm2. Consistent segmentation and neigborhoods in reference tissue.		Confirm antibody penetration with prescan
	# cells (nuclei) passing QC > 100	fraction reads in peaks >0.2	# features / nucleus	5<#rolony<30 / nucleus	Fiducial alignment	Spatial, cell-type and neighborhood correlation between sequential sections of CODEX, 3D IF and 10X Visium.		
	minimum #cells to call a cluster - 30	>200 accessible peaks in >49 nuclei		2<#genes<20 / nucleus	reads/spot >20K	Nuclear segmentation (F1 score), cellular density and distribution on experimental tissue.		check for data loss post compression of raw data
	genes/cell QA > 500	TSS enrichment > 2	Fraction decoded	60% rolonies decoded	genes/spot QA > 500			
	doublet detection rate QA < 10%		# features / spot / gene	# features / spot / gene	Total genes detected >10K			
	# clusters or cell types > 5		Nuclear segmentation	Nuclear segmentation, poly-T (total cDNA)	# clusters or cell types mapped > 5	cell type classification/cluster predicition (Shilouette analysis,Elbow plots)		
Assay drift controls	Nephrectomy tissue every 8th run, L-J plots		Nx every run, house keeping genes		house keeping genes	Image intensity, segmentation and classification of Nx tissue visualized with L-J plots		psfs confirm clearing and instrumentation
Software used	Cellranger arc; HIVE snRNA-seq Pipeline	Cellranger arc; HIVE snRNA-seq Pipeline	Esper Spatial Studio software	in house python code	cloupe browser, R, giotto, vitessce	Leica LASX or CODEX processor and ImageJ with VTEA		Arevis, ZEN2 (Blue), Imaris, MFB
Data or file types	FASTQ	FASTQ		tif, tsv	FASTQ, BAM, tif	tif	tif, lif	CZI, JP2000, OME-TIFF, xml

HuBMAP
Human BioMolecular Atlas Program

Assay Performance Controls (**drift over time)**



Experimental Design: Assay QC (Levey-Jennings plots)

(Multiome Only)

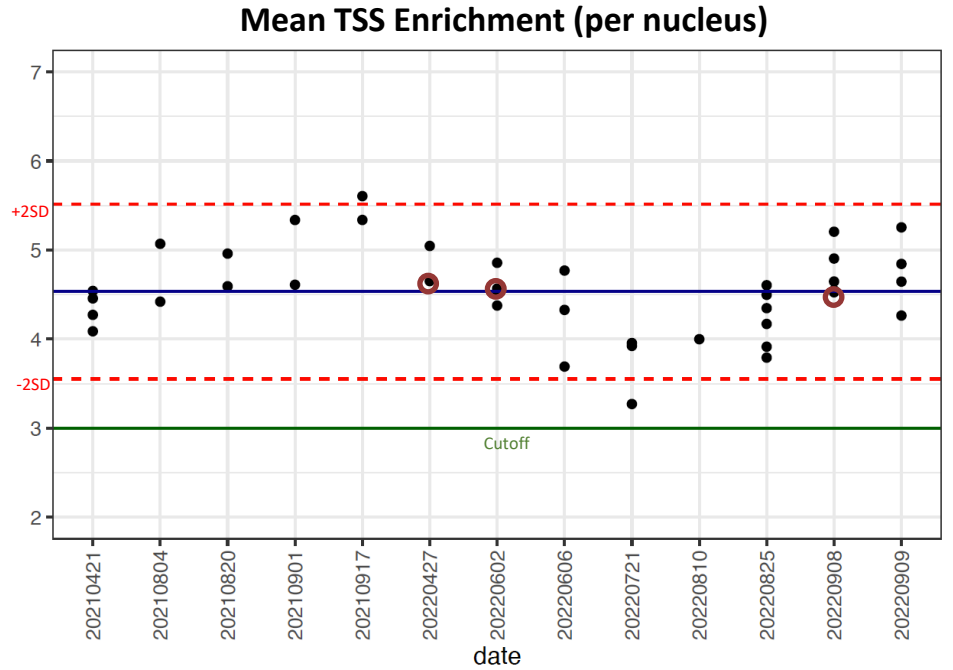
Multiome (RNA + ATAC):

229,780 nuclei

27 individuals

40 samples

Add a control sample every batch, or at any change in the pipeline



Assay and Analytic Standardizations: Cell Nomenclature (ASCT+B Tables)

Anatomical Structures, Cell Types and Biomarkers Table for kidney v1.3																		
Author Name(s):	Sanjay Jain; M. Todd Valerius; Yongqun He; Tarek El Achkar; Michael Eadon																	
Author ORCID(s):	0000 0003 2804 127X; 0000 0001 8143 9231; 0000 0001 9189 9661																	
Reviewer(s):	Blue Lake; Becky Steck; Rachel Dull; Ellen M. Quardokus; Daria Barwinska																	
Reviewer ORCID(s):	0000 0002 8637 8044;0000 0001 7565 4904;0000 0002 0864 7007;0000 0001 7655 4833																	
General Publication(s):	DOI: 10.1152/physiolgenomics.00104.2020, doi: https://doi.org/10.1101/2021.07.28.454201; DOI: 10.1038/s41467 019 10861 2; DOI: 10.1172/jci.insight.133267; DOI: 10.1038/s41586 020 2941 1; DOI: 10.1016/j.devcel.2019.10.005; DOI: 10.1073/pnas.2005477117; DOI: 10.1681/ASN.2020010071; DOI: 10.1126/science.aat1699; DOI: 10.3389/fimmu.2019.02035; DOI: 10.1126/science.aat5031; DOI: 10.1084/jem.20030323; DOI: 10.1016/j.exphem.2016.10.008																	
Data DOI:																		
Date:	revise for October 2022																	
Version Number:	v1.3																	
AS/1	AS/1/LABEL	AS/1/ID	AS/2	AS/2/LABEL	AS/2/ID	AS/3	AS/3/LABEL	AS/3/ID	AS/4	AS/4/LABEL	AS/4/ID	AS/5	AS/5/LABEL	AS/5/ID	AS/6	AS/6/LABEL	AS/6/ID	
kidney	kidney	UBERON:0002113	kidney capsule	kidney capsule	UBERON:000201													
kidney	kidney	UBERON:0002113	cortex of kidney	cortex of kidney	UBERON:00012 25	outer cortex of kidney	outer cortex of kidney	UBERON:0002189				Glomerular Epithelium	glomerular epithelium	UBERON:00041 88	visceral epithelial layer	glomerular visceral epithelium	UBERON:00068 52	
kidney	kidney	UBERON:0002113	cortex of kidney	cortex of kidney	UBERON:00012 25	juxtamedullary cortex	juxtamedullary cortex	UBERON:0005271				Glomerular Epithelium	glomerular epithelium	UBERON:00041 88	visceral epithelial layer	glomerular visceral epithelium	UBERON:00068 52	
kidney	kidney	UBERON:0002113	cortex of kidney	cortex of kidney	UBERON:00012 25	medullary ray	medullary ray	UBERON:0009883				Collecting Duct (Cortex)	cortical collecting duct	UBERON:000420				
kidney	kidney	UBERON:0002113	renal medulla	renal medulla	UBERON:00003 62	kidney outer medulla outer stripe	kidney outer medulla outer stripe	UBERON:0004202				Loop of Henle (Thick Limb/Medulla)	thick ascending limb of loop of Henle	UBERON:000129				
kidney	kidney	UBERON:0002113	renal medulla	renal medulla	UBERON:00003 62	kidney outer medulla inner stripe	kidney outer medulla inner stripe	UBERON:0004201				Loop of Henle (Thin Limb)	descending thin limb of loop of Henle	UBERON:00012 89	descending thin limb of loop of Henle 2			
kidney	kidney	UBERON:0002113	renal medulla	renal medulla	UBERON:00003 62	inner medulla of kidney	inner medulla of kidney	UBERON:0001294				Descending Vasa Recta	vasa recta descending limb	UBERON:000920				
kidney	kidney	UBERON:0002113	renal medulla	renal medulla	UBERON:00003 62	inner medulla of kidney	inner medulla of kidney	UBERON:0001294				renal papilla	renal papilla	UBERON:00012 28	papillary tip	tip of renal papilla	UBERON:00090 95	
kidney	kidney	UBERON:0002113	renal cortex	cortex of kidney	UBERON:00012 25	renal column	renal column (column of Bertin)	UBERON:0001284				Glomerular Epithelium	glomerular epithelium	UBERON:00041 88	visceral epithelial layer	glomerular visceral epithelium	UBERON:00068 52	
kidney	kidney	UBERON:0002113	renal medulla	renal medulla	UBERON:00003 62	renal pyramid	kidney pyramid	UBERON:0004200				Descending Vasa Recta	vasa recta descending limb	UBERON:000920				

Reviewers

Evidence - DOI

Anatomical Structures

Uberon

Cell Types names

Cell Type ontology

Biomarkers

HUGO (standardize)

CT/1	CT/1/LABEL	CT/1/ID	CT/1/ABBR	CT/1/NOTES	BGene/1	BGene/1/LABEL	BGene/1/ID	BGene/1/NOTES	BGene/2	BGene/2/LABEL	BGene/2/ID	BGene/2/NOTES	BGene/3
capsule mesenchymal stromal cell					Foxd1	forkhead box D1	HGNC:3802		VSX2		HGNC:1975		
Podocyte	podocyte	CL:0000653			NPHS2	NPHS2 stomatin family member, podocin	HGNC:13394				HGNC:9171		NPHS1
Podocyte	glomerular visceral epithelial cell	CL:0000653			NPHS2	NPHS2 stomatin family member, podocin	HGNC:13394		PODXL		HGNC:9171	also see in endothelial cells	NPHS1
Cortical Collecting Duct Principal Cell	kidney cortex collecting duct principal cell	CL:1000714			AQP2	aquaporin 2	HGNC:634		AQP3		HGNC:636		FXYD4
Medullary Thick Ascending Limb Cell	kidney loop of Henle medullary thick ascending limb epithelial cell	CL:1001108			SLC12A1	solute carrier family 12 member 1	HGNC:10910		UMOD		HGNC:12559		CASR
Descending Thin Limb Cell Type 2	kidney loop of Henle long descending thin limb outer medulla epithelial cell	CL:4030013	DTL2		CRYAB	crystallin alpha B	HGNC:2389		SLC39A8		HGNC:20862		TACSTD2
Descending Vasa Recta Endothelial Cell	vasa recta descending limb cell	CL:1001285			SERPINE2	serpin family E member 2	HGNC:8951		TM4SF1		HGNC:11853		PALMD
Inner Medullary Collecting Duct Cell	kidney inner medulla collecting duct principal cell	CL:1000718			AQP2	aquaporin 2	HGNC:634		SLC14A2		HGNC:10919		
Podocyte	glomerular visceral epithelial cell	CL:0000653			NPHS2	NPHS2 stomatin family member, podocin	HGNC:13394		PODXL		HGNC:9171	also see in endothelial cells	NPHS1
Descending Vasa Recta Endothelial Cell	vasa recta descending limb cell	CL:1001285			SERPINE2	serpin family E member 2	HGNC:8951		TM4SF1		HGNC:11853		PALMD

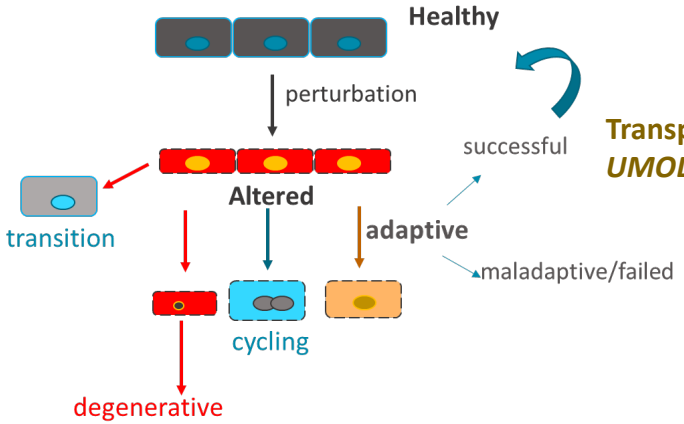
HuBMAP Human BioMolecular Atlas Program

KPMP

Integrated Atlas across Technologies, Institutions and Consortia

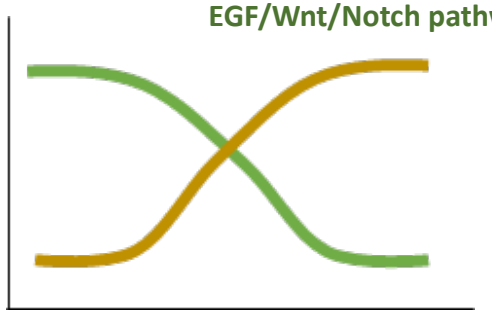
Data supported definitions: How to identify healthy cells?

Mature Functional State → **Altered state Dedifferentiation**



Transporters
***UMOD*, *EGF*, ESRRB TF**

EMT genes
TGF-β/SMAD pathway
MAPK (JNK) pathway
JAK/STAT pathway
Rel/NF-kB pathway
EGF/Wnt/Notch pathways



An atlas of healthy and injured cell states and niches in the human kidney

Blue B. Lake[1*], Rajasree Menon[2*], Seth Winfree[3*], Qiwen Hu[4*], Ricardo Melo Ferreira[5*], Kian Kalhor[1], Daria Barwinska[5], Edgar A. Otto[6], Michael Ferkowicz[5], Dinh Diep[1], Nongluk Plongthongkum[1], Amanda Knoten[7], Sarah Urata[1], Abhijit S. Naik[6], Sean Eddy[6], Bo Zhang[7], Yan Wu[1], Diane Salamon[7], James C. Williams[5], Xin Wang[4], Karol S. Balderrama[8], Paul Hoover[8], Evan Murray[8], Anitha Vijayan[7], Fei Chen[8], Sushrut S. Waikar[9], Sylvia Rosas[10], Francis P. Wilson[11], Paul M. Palevsky[12], Krzysztof Kiryluk[13], John R. Sedor[14], Robert D. Toto[15], Chirag Parikh[16], Eric H. Kim[17], Evan Z. Macosko[8], Peter V. Kharchenko[4], Joseph P. Gaut[18], Jeffrey B. Hodgin[6], Michael T. Eadon[5], Pierre C. Dagher[5,‡], Tarek M. El-Achkar[5,‡], Kun Zhang[1,‡], Matthias Kretzler[6,‡], Sanjay Jain[7,18,‡], for the KPMP consortium[19].

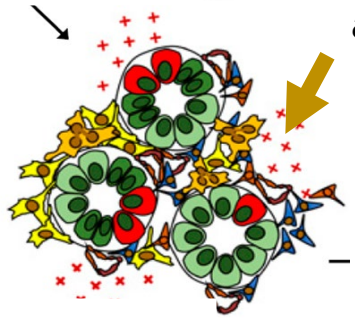
DOI: 10.1101/2021.07.28.454201; Nature in press

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

Not healthy reference

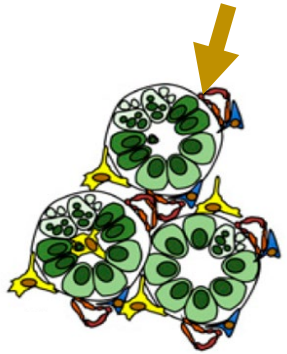
degenerative / damaged / dysfunctional

marked loss of differentiation markers, and/or increased %ERT, %MT, and/or marked decrease in genes detected. These cells could represent an early injury state or cells that will not recover function due to severe injury



adaptive / activated - stromal

represented by cells that are differentiating or have differentiated into high ECM producing myofibroblasts or activated fibroblasts



adaptive / successful or failed repair - epithelial injury

represented by cells that retain differentiation markers of reference states, albeit at lower levels, but also show expression of known injury associated genes, mesenchymal markers or factors promoting inflammation or fibrosis

Basile et al. (2016). J Am Soc Nephrol, 27: 687–697
Docherty et al. (2019). JASN, 30, 726-736
Ferenbach & Bonventre (2015). Nat Rev Nephrol 11, 264–276
Miller and Zachary. (2017). Path. Basis of Vet. Dis.,2-43.e19
Ronco et. al. (2019). Lancet, 394(10212):1949-1964

doi:10.1101/2021.05.31.446440
doi: 10.48539/HBM325.PTQS.258

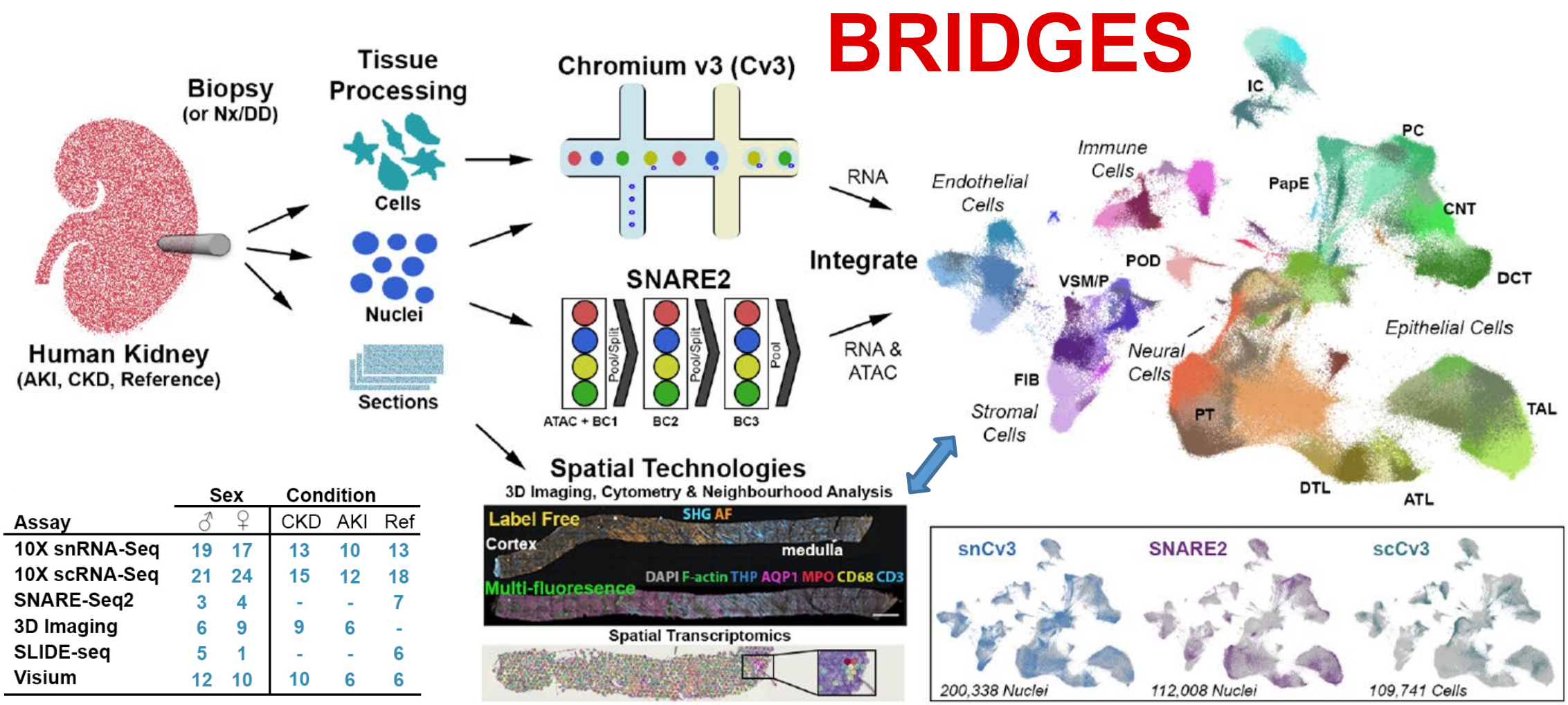
Collection of multi-modal maps of human kidney

Biopsy
(or Nx, DD)

Assay & Data Integration
(Layering of emerging information)

Atlas v1
Cell types/states
Cellular Neighborhoods

Kidney Cell Atlas Overview: Technologies + Data Integration



	Sex		Condition		
Assay	♂	♀	CKD	AKI	Ref
10X snRNA-Seq	19	17	13	10	13
10X scRNA-Seq	21	24	15	12	18
SNARE-Seq2	3	4	-	-	7
3D Imaging	6	9	9	6	-
SLIDE-seq	5	1	-	-	6
Visium	12	10	10	6	6

KPMP HuBMAP Human BioMolecular Atlas Program

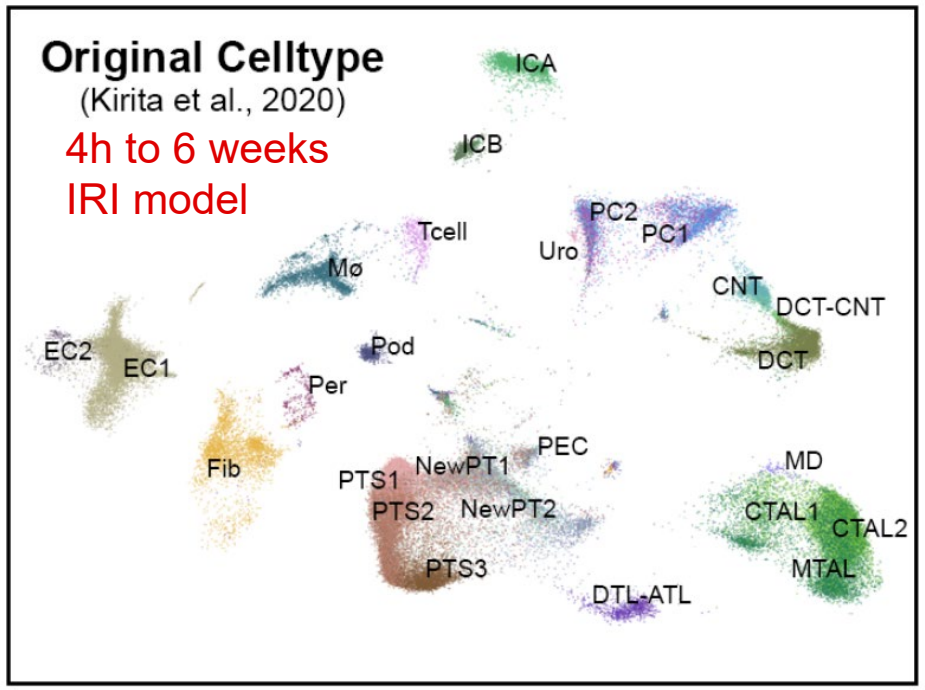
Standardized annotations as a bridge

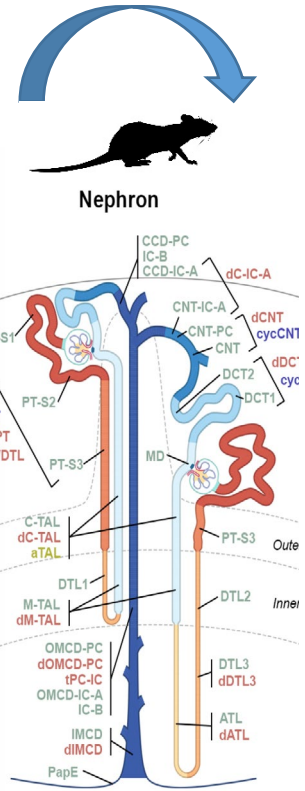
Cross-species mechanistic insights

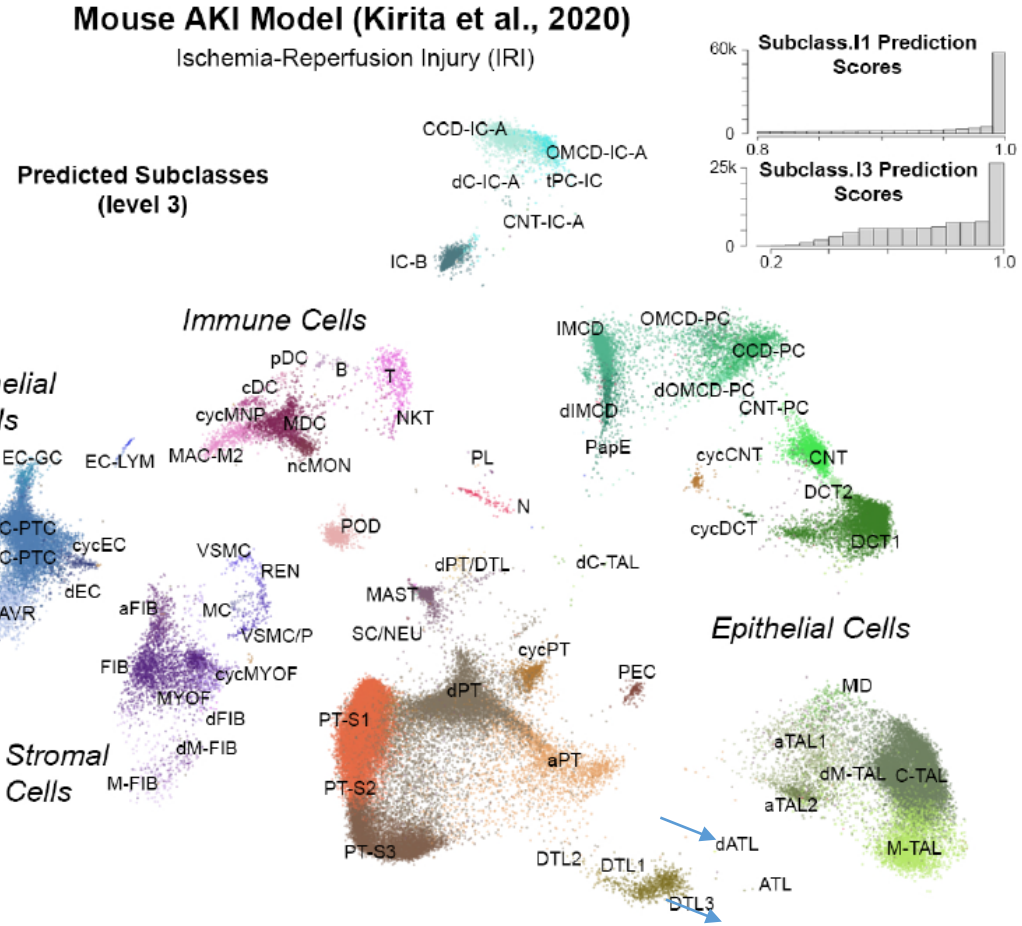
Cell profiling of mouse acute kidney injury reveals conserved cellular responses to injury

Yuhei Kirita[a,b], Haojia Wu[a], Kohei Uchimura[a], Parker C. Wilson[c], and Benjamin D. Humphreys[a,d,1]

15874–15883 | PNAS | July 7, 2020 | vol. 117 | no. 27







Mapped to snCv3 using Seurat v4

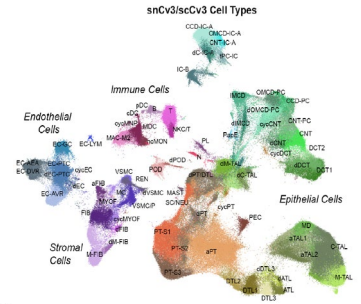
Can resolve many more cell types using human atlas data

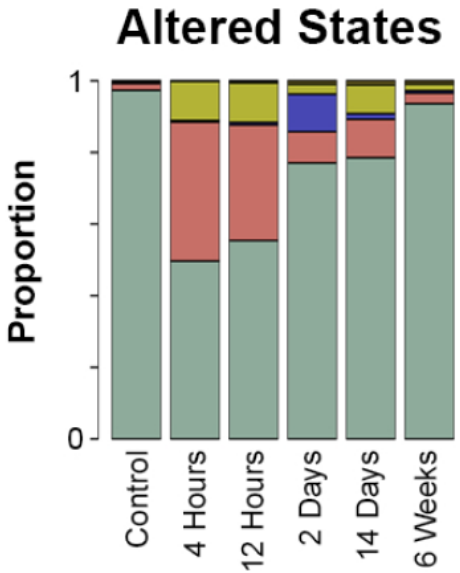
Blue Lake

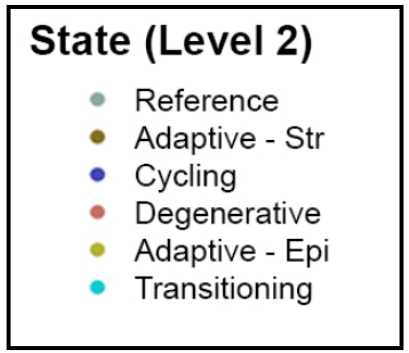
KPMP

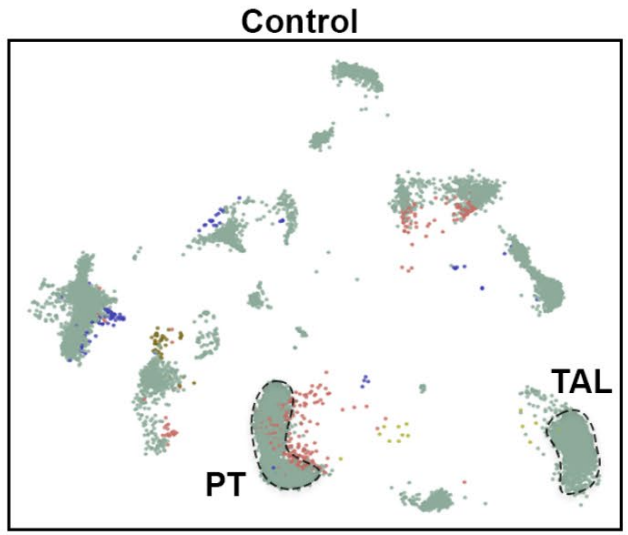
HuBMAP
Human BioMolecular Atlas Program

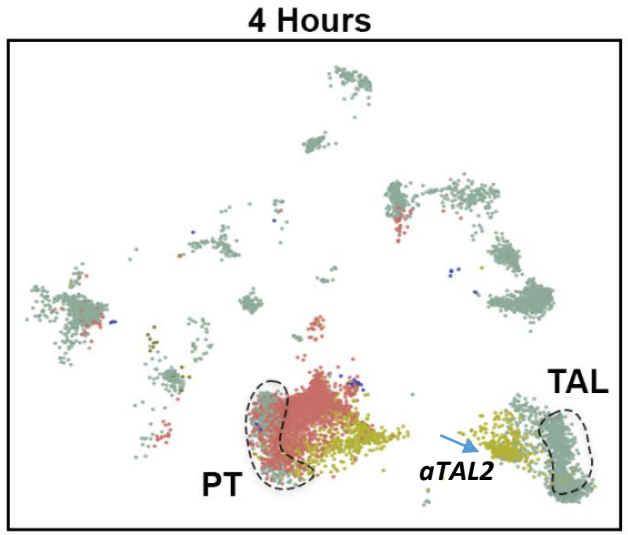
Integrating mouse IRI time course with human atlas

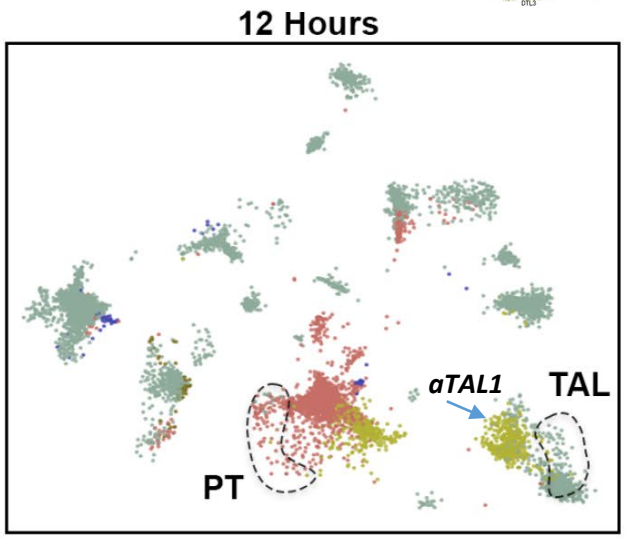


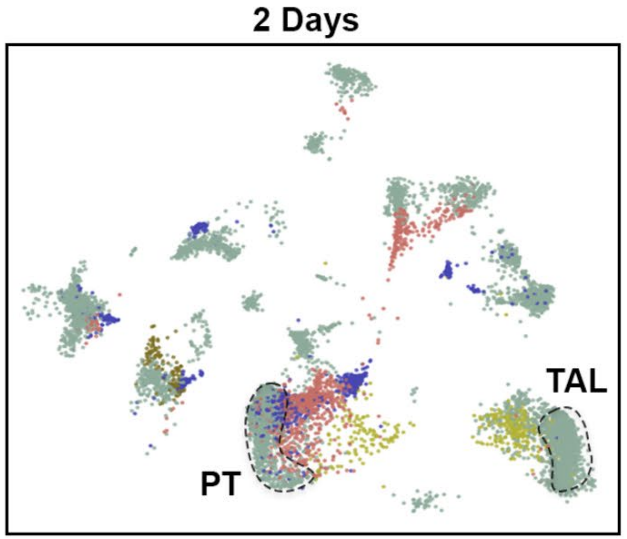


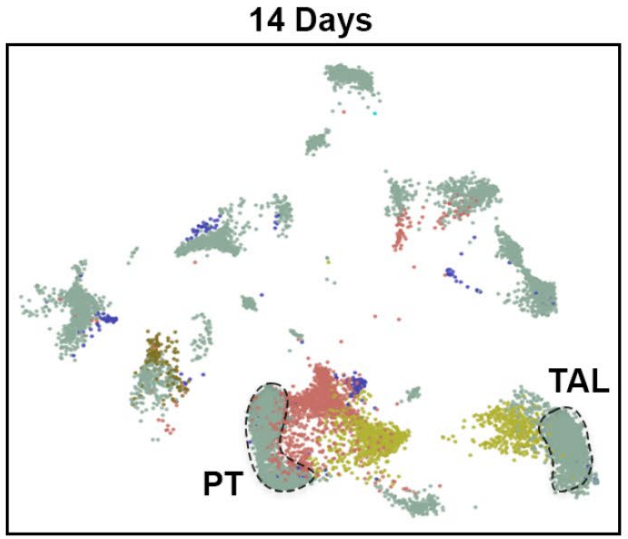


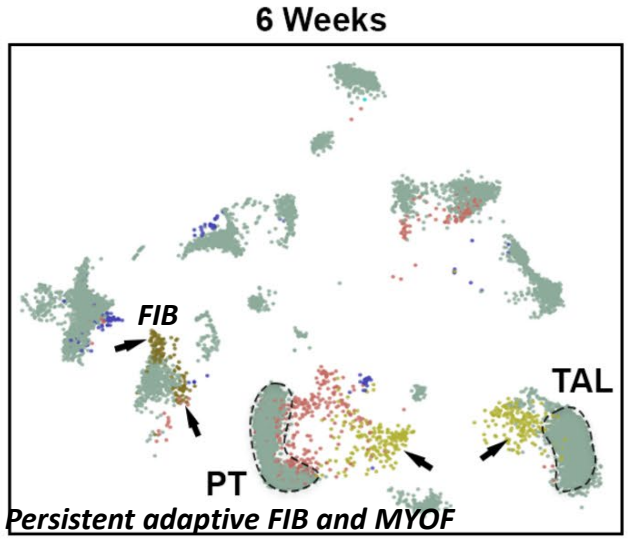










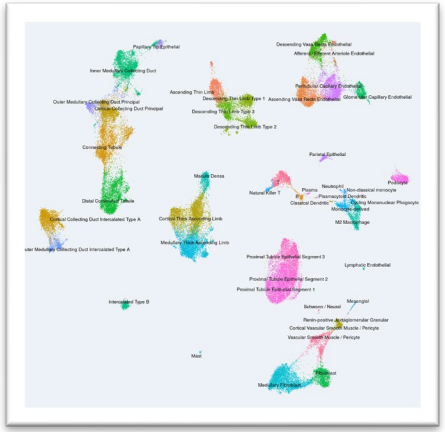


Persistent adaptive FIB and MYOF

Persistent repairing aPT and aTAL (Failed Repair?)

How to use it: Cross Collaborative Tools

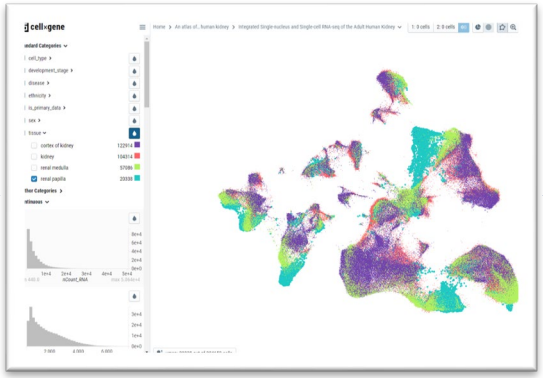
Azimuth reference



Identify cell types in your own dataset

https://azimuth.hubmapconsortium.org/

Cell Types and States

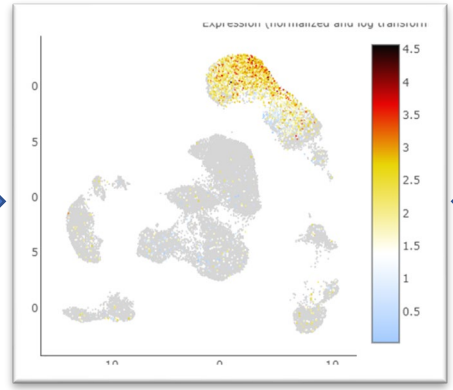


Explore genes, cell types, correlating with clinical data

> 200 downloads of integrated analytics before publishing!

https://cellxgene.cziscience.com/collections/bcb61471-2a44-4d00-a0af-ff085512674c

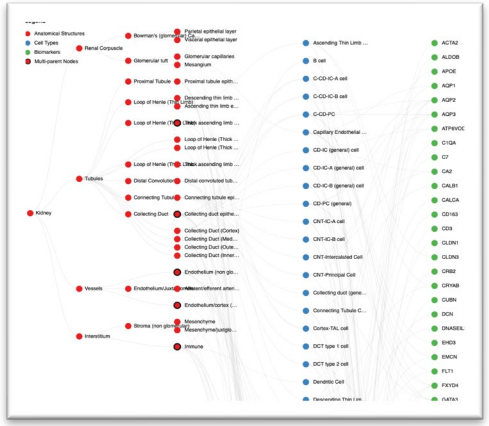
Atlas Explorer



Gene expression across multiple technologies and kidney space

https://atlas.kpmp.org/explorer/

ASCT+B



Standardizing cell type nomenclature and markers

https://doi.org/10.48539/HBM248.CBJV.556

https://hubmapconsortium.github.io/ccf/pages/ccf-anatomical-structures.html

Acknowledgements for Atlas v1 (Many – more than a village)

snCv3/SNARE2

Kun Zhang
Blue B. Lake
Kian Kalhor
Dinh Diep
Nongluk Plongthongkum
Sarah Urata
Yan Wu

Sanjay Jain
Joseph P. Gaut
Anitha Vijayan
Amanda Knoten
Kristy Conlon
Bo Zhang
Madhurima Kaushal
Gerald Nwanne
Reetika Ghag
Diane Salamon

scCv3

Matthias Kretzler
Jeffrey B. Hodgin
Rajasree Menon
Edgar A. Otto
Abhijit S. Naik
Sean Eddy
Paul Hoover

3D Imaging

Tarek M. El-Achkar
Pierre C. Dagher
Seth Winfree
Daria Barwinska
Michael Ferkowicz
James C. Williams

10X Visium

Michael T. Eadon
Pierre C. Dagher
Ricardo Melo Ferreira

Slide-seq2

Evan Z. Macosko
Fei Chen
Anna Greka
Evan Murray
Karol S. Balderrama
Jamie L. Marshall

SNARE2 Pipelines + Data Analysis/Integration

Peter V. Kharchenko
Qiwen Hu
Xin Wang

HuBMAP Human BioMolecular Atlas Program
KPMP
neptune
CHAN ZUCKERBERG INITIATIVE
ERCB
NIH The Common Fund
NIH National Institute of Diabetes and Digestive and Kidney Diseases

DK133081, DK133091, DK133092,DK133093, DK133095, DK1330971, DK114866, DK114908, DK133090, DK133113,DK133766, DK133768, DK114907, DK114920, DK114923, DK114933, DK114886.

Thank You!