Metadata and Data Standards for NIDDK Research Data - The ATLAS-D2K Experience

M. Todd Valerius, Ph.D.

Brigham and Women's Hospital / Harvard Medical School

June, 2023

The ATLAS-D2K Center

A kidney and lower urinary tract-focused data discovery hub with access to visualizations and analysis tools.

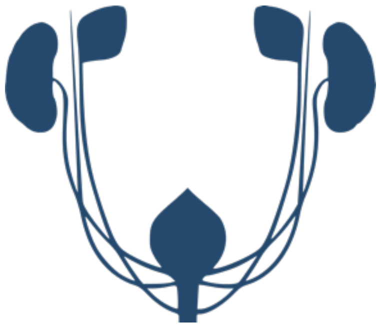
Bringing GUDMAP & RBK data under one ATLAS that embraces open science and provides links to related consortiums.

GUDMAP

RBK
ReBuilding a Kidney

Overall Aims:

Our long-term goal is to bring complex data into an accessible form for our research community.



Establish connections between molecular data of kidney and lower urinary tract present in GUDMAP, RBK, KPMP, HuBMAP, and the HCA.

Enable researchers of various levels of experience by providing tools to interact with the data.

GenitoUrinary Development Molecular Anatomy Project (GUDMAP) & (Re)Building a Kidney (RBK): Overview

Overarching Program Goal:

- **GUDMAP:** high resolution molecular anatomy of the developing and mature genitourinary system (mouse, human, rat, dog)
- **RBK:** optimize differentiation of human kidney cell types in defined structures, and determine methods to promote kidney repair, to generate or repair nephrons that can function within the kidney (human, human iPSCs, zebrafish)

Number of investigators involved: GUDMAP: 9 RBK: 25

Technology Focus: array of gene expression techniques on tissues, differentiation of stem cells, a range of imaging techniques

Current Gaps? robust anatomical ontologies broadly implemented, metadata standardization, interactive tools for data analysis/data annotation (e.g., cluster data)

ATLAS-D2K Research goals

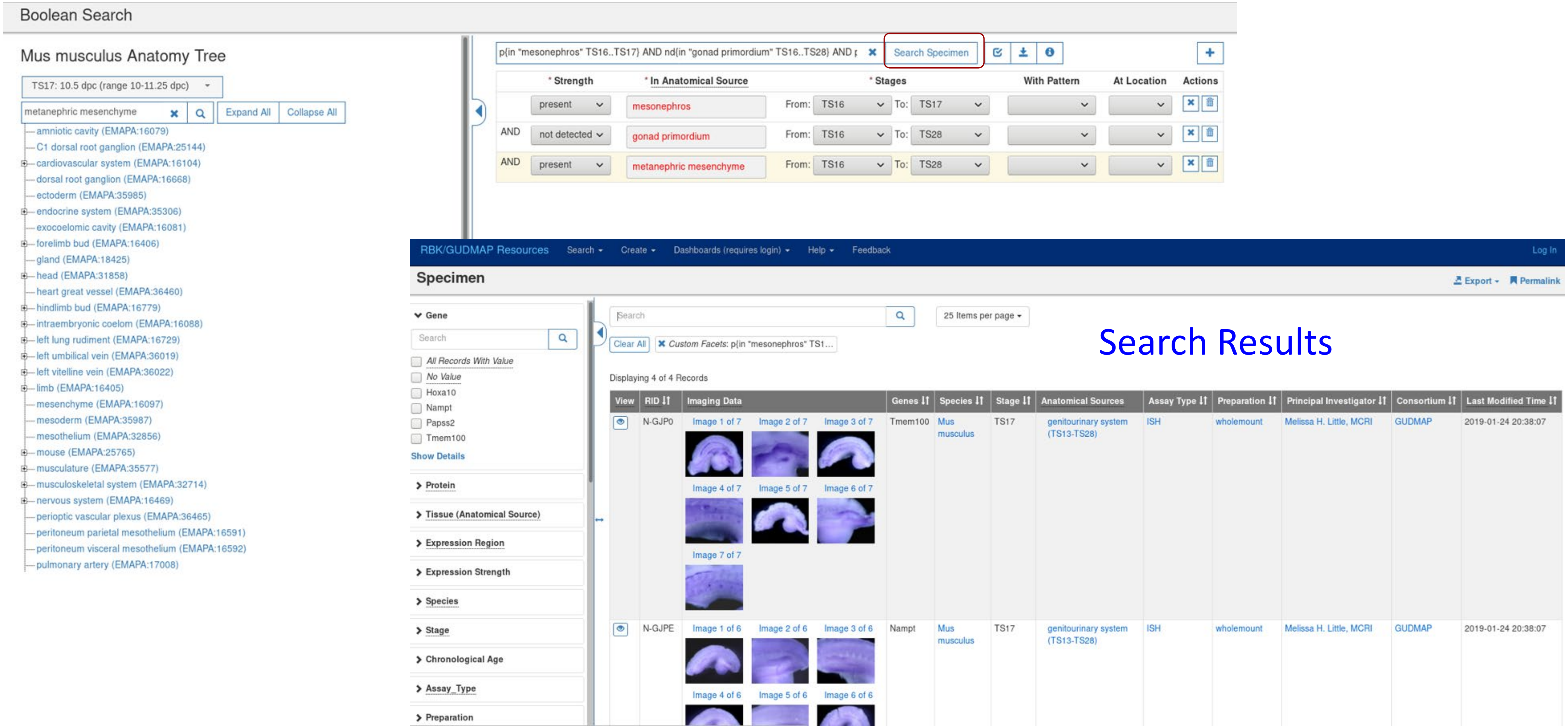
- Example queries:
  - a) scRNA-Seq data analysis (rookie/veteran)
  - b) GWAS gene list mapped to expression.
- Graphical Tools for Genitourinary Data
- Integrating molecular and imaging data
- Establishing reference datasets
- Data harmonization across consortiums
- Bioinformatic pipelines and visualization tools

Ontologies and controlled vocabularies

- Why is this important?
  - Gene expression and function occurs in tissues. Consistent use of names removes confusion amongst researchers and ***enables*** computation of complex queries.
  - Quickly apparent when trying to connect data
    - GUDMAP had generated thousands of wholemount & section *in situ* hybridizations, scored for expression, from two groups. -> ***anatomical ontology*** needed to connect.

Abler, L.L. *et al.* (2011) *Developmental dynamics : an official publication of the American Association of Anatomists*. https://doi.org/10.1002/dvdy.22730.
Georgas, K.M. *et al.* (2015) *Development (Cambridge, England)*. https://doi.org/10.1242/dev.117903.
Harding, S.D. *et al.* (2011) *Development (Cambridge, England)*. https://doi.org/10.1242/dev.063594.
Henry, G.H. *et al.* (2018) *Cell Rep*. https://doi.org/10.1016/j.celrep.2018.11.086.
Little, M.H. *et al.* (2007) *Gene expression patterns : GEP*. https://doi.org/10.1016/j.modgep.2007.03.002.

Boolean Search on Scored Expression



Search Results

Boolean Search

Mus musculus Anatomy Tree

TS17: 10.5 dpc (range 10-11.25 dpc)

metanephric mesenchyme

- amniotic cavity (EMAPA:16079)
- C1 dorsal root ganglion (EMAPA:25144)
- cardiovascular system (EMAPA:16104)
- dorsal root ganglion (EMAPA:16668)
- ectoderm (EMAPA:35985)
- endocrine system (EMAPA:35306)
- exocoelomic cavity (EMAPA:16081)
- forelimb bud (EMAPA:16406)
- gland (EMAPA:18425)
- head (EMAPA:31858)
- heart great vessel (EMAPA:36460)
- hindlimb bud (EMAPA:16779)
- intraembryonic coelom (EMAPA:16088)
- left lung rudiment (EMAPA:16729)
- left umbilical vein (EMAPA:36019)
- left vitelline vein (EMAPA:36022)
- limb (EMAPA:16405)
- mesenchyme (EMAPA:16097)
- mesoderm (EMAPA:35987)
- mesothelium (EMAPA:32856)
- mouse (EMAPA:25765)
- musculature (EMAPA:35577)
- musculoskeletal system (EMAPA:32714)
- nervous system (EMAPA:16469)
- perioptic vascular plexus (EMAPA:36465)
- peritoneum parietal mesothelium (EMAPA:16591)
- peritoneum visceral mesothelium (EMAPA:16592)
- pulmonary artery (EMAPA:17008)

p{in "mesonephros" TS16..TS17} AND nd{in "gonad primordium" TS16..TS28} AND p

	* Strength	* In Anatomical Source	* Stages		With Pattern	At Location	Actions
	present	mesonephros	From: TS16	To: TS17			
AND	not detected	gonad primordium	From: TS16	To: TS28			
AND	present	metanephric mesenchyme	From: TS16	To: TS28			

Specimen

Gene

- □ *All Records With Value*
- □ *No Value*
- □ Hoxa10
- □ Nampt
- □ Papss2
- □ Tmem100

Custom Facets: p{in "mesonephros" TS1...

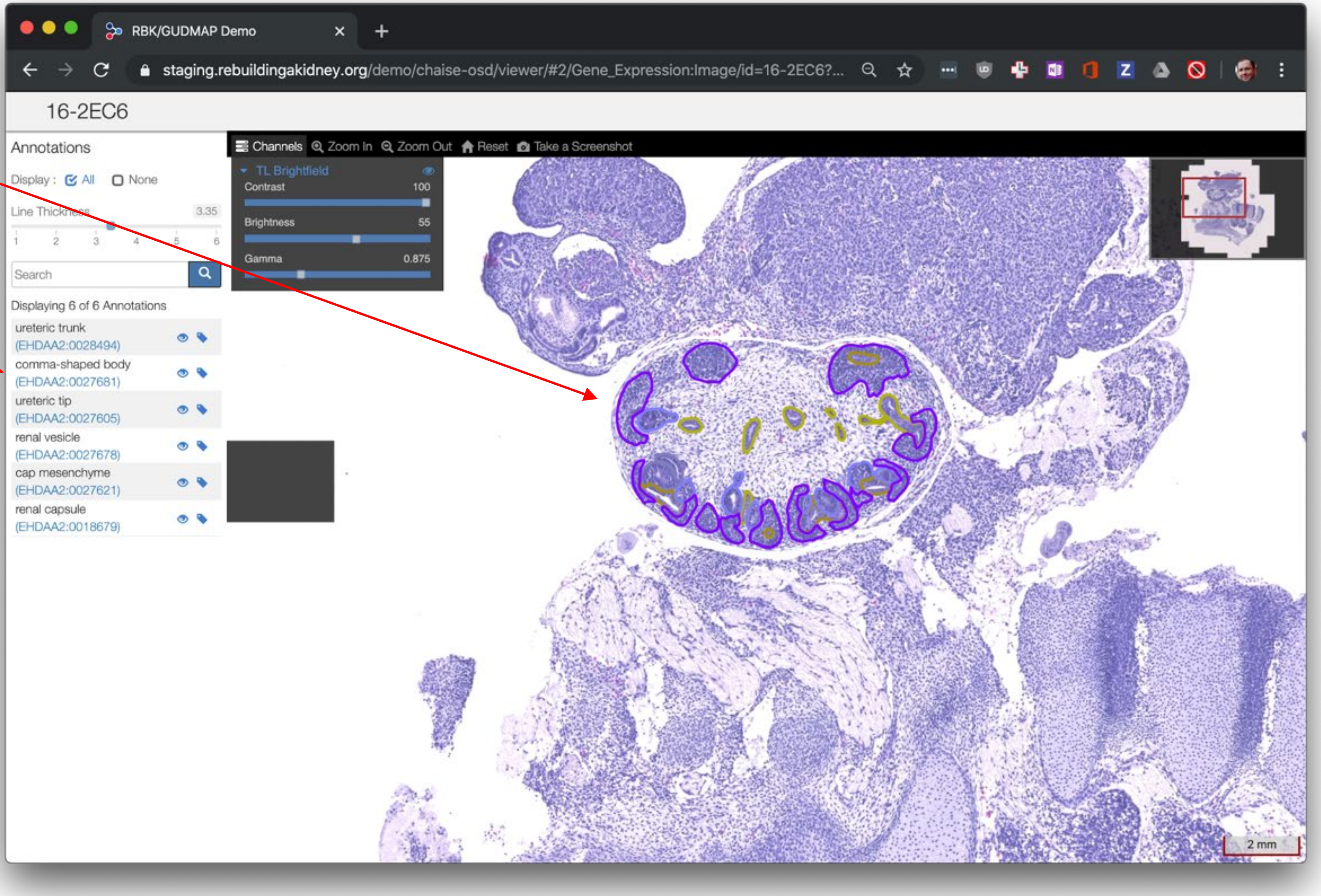
Displaying 4 of 4 Records

View	RID	Imaging Data	Genes	Species	Stage	Anatomical Sources	Assay Type	Preparation	Principal Investigator	Consortium	Last Modified Time
	N-GJP0	Image 1 of 7 Image 2 of 7 Image 3 of 7 Image 4 of 7 Image 5 of 7 Image 6 of 7 Image 7 of 7	Tmem100	Mus musculus	TS17	genitourinary system (TS13-TS28)	ISH	wholemount	Melissa H. Little, MCRI	GUDMAP	2019-01-24 20:38:07
	N-GJPE	Image 1 of 6 Image 2 of 6 Image 3 of 6 Image 4 of 6 Image 5 of 6 Image 6 of 6	Nampt	Mus musculus	TS17	genitourinary system (TS13-TS28)	ISH	wholemount	Melissa H. Little, MCRI	GUDMAP	2019-01-24 20:38:07

Anatomical annotation using established terms

Manual annotation of structures

Links to other data through anatomy


RBK/GUDMAP Demo
staging.rebuildingakidney.org/demo/chaise-osd/viewer/#2/Gene_Expression:Image/id=16-2EC6?...
Channels
Zoom In
Zoom Out
Reset
Take a Screenshot
Search
2 mm

16-2EC6

Annotations

Display : ☑ All □ None

Line Thickness 3.35

TL Brightfield

Contrast 100

Brightness 55

Gamma 0.875

Displaying 6 of 6 Annotations

ureteric trunk (EHDAA2:0028494)

comma-shaped body (EHDAA2:0027681)

ureteric tip (EHDAA2:0027605)

renal vesicle (EHDAA2:0027678)

cap mesenchyme (EHDAA2:0027621)

renal capsule (EHDAA2:0018679)

Ontologies and controlled vocabularies

- Gene expression and function occurs in tissues. Consistent use of names removes confusion amongst researchers and ***enables*** computation of complex queries.
- Quickly apparent when trying to connect data
  - GUDMAP had generated thousands of wholemount & section *in situ* hybridizations, scored for expression, from two groups. -> ***anatomical ontology*** needed to connect.
- To accommodate cross-species data, we use Uberon and Cell Ontology
  - Uberon multi-species anatomy ontology
  - The Cell Ontology
- Healthy adult human tissue focus in **HuBMAP ASCT+B Tables**
  - Data-driven effort lead by Sanjay and enhanced by many.

Recommendation:

Select a source of anatomical and cell type terms that fit your research, use them as a standard, and capture the source of those terms from established ontologies.

Data formats: Is anything as future proof as plain text?

RAW sequence data is, but there are privacy issues.

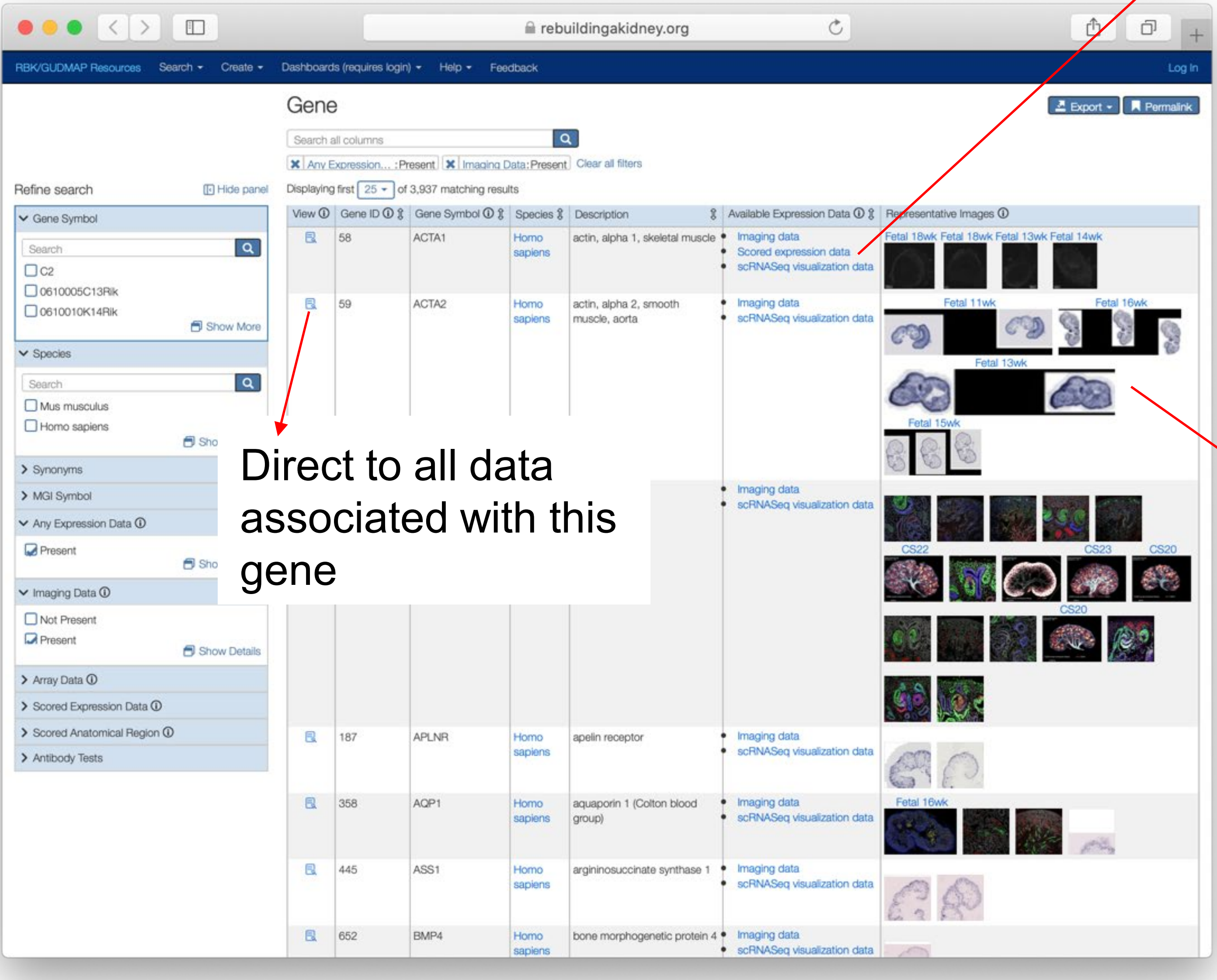
- Detailed *sequencing metadata* is produced and captured by computational tools.
- *Biosample metadata* needs to be captured well at the time of experiment.
- Protocols should be well referenced. (consortiums rely on self hosting OR commercial repositories like Protocols.io

Image data is a mature, poor example

- Center on "open" established standards like OME-TIFF and related formats (e.g., Zarr and OME-NGFF) that capture microscopy/imaging metadata.
  - The benefit: these formats work well with open-source software like ImageJ and QuPath.
  - Images adjusted for publication and presentation <u>are not useful for downstream reuse and quantitative analysis</u>.
- The "biosample metadata" associated with an image needs to be captured EARLY in the process.

Data Curation/Interaction UI: Gene

Search for presence of different types of expression data, scored exp. region, etc.



Gene

Search all columns

☒ Any Expression... : Present ☒ Imaging Data: Present Clear all filters

Displaying first 25 ▾ of 3,937 matching results

Refine search Hide panel

- Gene Symbol
  - Search
  - □ C2
  - □ 0610005C13Rik
  - □ 0610010K14Rik
  - Show More
- Species
  - Search
  - □ Mus musculus
  - □ Homo sapiens
  - Sho
- Synonyms
- MGI Symbol
- Any Expression Data ⓘ
  - ☑ Present
  - Sho
- Imaging Data ⓘ
  - □ Not Present
  - ☑ Present
  - Show Details
- Array Data ⓘ
- Scored Expression Data ⓘ
- Scored Anatomical Region ⓘ
- Antibody Tests

View ⓘ	Gene ID ⓘ	Gene Symbol ⓘ	Species	Description	Available Expression Data ⓘ	Representative Images ⓘ
	58	ACTA1	Homo sapiens	actin, alpha 1, skeletal muscle	• Imaging data • Scored expression data • scRNASeq visualization data	Fetal 18wk Fetal 18wk Fetal 13wk Fetal 14wk
	59	ACTA2	Homo sapiens	actin, alpha 2, smooth muscle, aorta	• Imaging data • scRNASeq visualization data	Fetal 11wk Fetal 16wk Fetal 13wk Fetal 15wk
					• Imaging data • scRNASeq visualization data	CS22 CS23 CS20 CS20
	187	APLNR	Homo sapiens	apelin receptor	• Imaging data • scRNASeq visualization data	
	358	AQP1	Homo sapiens	aquaporin 1 (Colton blood group)	• Imaging data • scRNASeq visualization data	Fetal 16wk
	445	ASS1	Homo sapiens	argininosuccinate synthase 1	• Imaging data • scRNASeq visualization data	
	652	BMP4	Homo sapiens	bone morphogenetic protein 4	• Imaging data • scRNASeq visualization data	

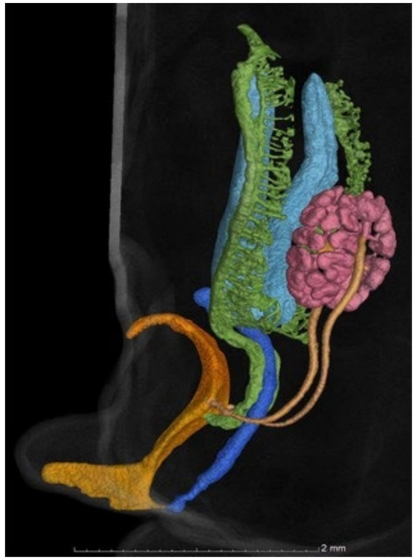
Direct to all data associated with this gene

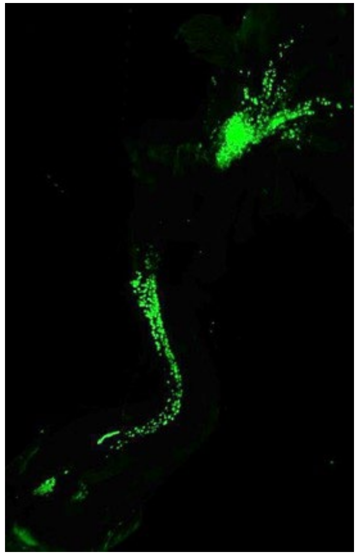
Direct to expression data associated with this gene

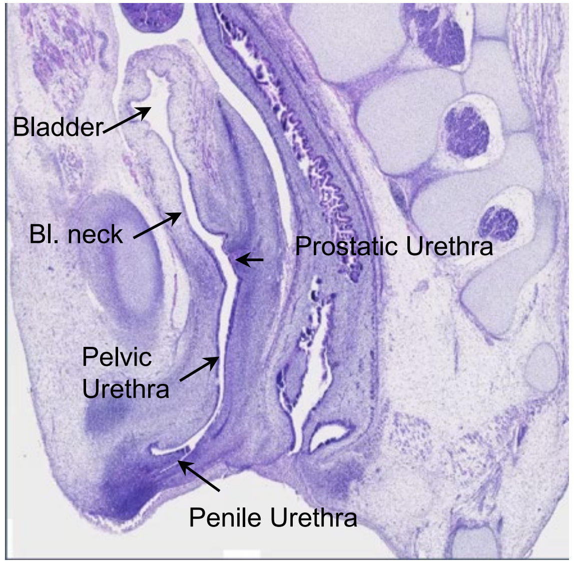
Direct to different specimens associated with this gene

Normal anatomical and structural changes

3-D Mapping of Tissues







Data formats: Is anything as future proof as plain text?

RAW sequence data is, but there are privacy issues.

- Detailed *sequencing metadata* is produced and captured by computational tools.
- *Biosample metadata* needs to be captured well at the time of experiment.
- Protocols should be well referenced. (consortiums rely on self hosting OR commercial repositories like Protocols.io

Image data is a mature, poor example

- Center on "open" established standards like OME-TIFF and related formats (e.g., Zarr and OME-NGFF) that capture microscopy/imaging metadata.
  - The benefit: these formats work well with open-source software like ImageJ and QuPath.
  - Images adjusted for publication and presentation are not useful for downstream reuse and quantitative analysis.
- The "biosample metadata" associated with an image needs to be captured EARLY in the process.

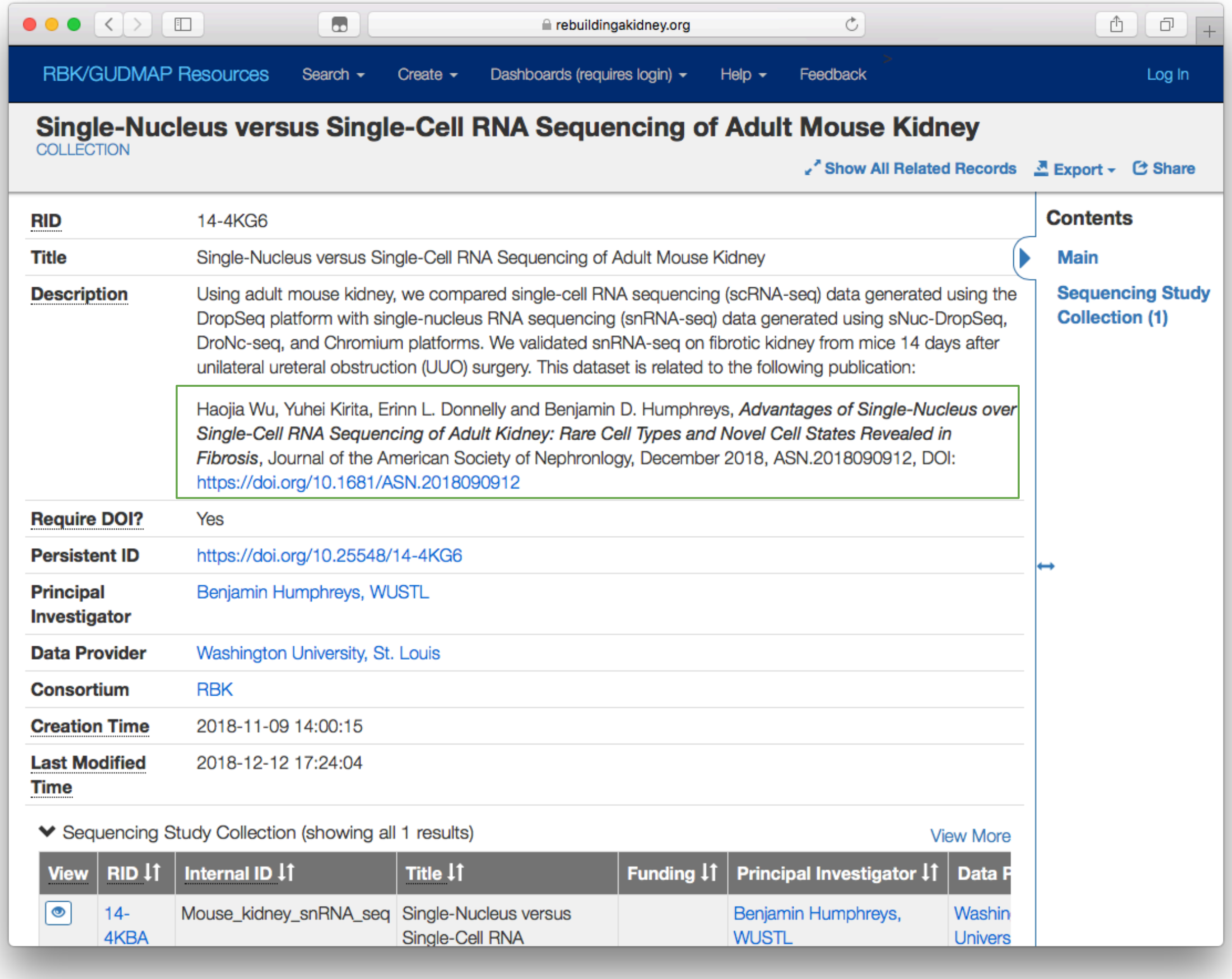
Recommendations:

1. Develop a plan to capture biosample metadata and protocol with the sequencing data.
2. Capture and associate biosamples metadata and consolidate on a lossless image file standard.

Data storage and levels of sharing, accessibility to open-source tools.

- **Sequence data – a layered approach**
  - "Processed" data is useful to a wider range of researchers.
    - Count files are ready for analysis without computationally intensive genome aligners in HPCs. More researchers can use such data immediately.
  - Privacy of participants - what is reasonable to share even with full consent?
    - The ability to de-identify participants from limited sequencing data expands before thoughtful policy will catch up. Think beyond the contractual protection to anticipate while maintaining data availability.
  - What intermediate products are available?
    - R objects like Seurat capture analysis decisions for a data generators fine analysis and can be used by less experienced researchers subsequently.
  - Processed data more freely shareable, but what happens when references change?
    - As data ages, re-alignment may be necessary as reference genomes change and improve.

Direct linking to data for efficiency with large datasets



Single-Nucleus versus Single-Cell RNA Sequencing of Adult Mouse Kidney
COLLECTION

RID	14-4KG6
Title	Single-Nucleus versus Single-Cell RNA Sequencing of Adult Mouse Kidney
Description	Using adult mouse kidney, we compared single-cell RNA sequencing (scRNA-seq) data generated using the DropSeq platform with single-nucleus RNA sequencing (snRNA-seq) data generated using sNuc-DropSeq, DroNc-seq, and Chromium platforms. We validated snRNA-seq on fibrotic kidney from mice 14 days after unilateral ureteral obstruction (UUO) surgery. This dataset is related to the following publication: Haojia Wu, Yuhei Kirita, Erinn L. Donnelly and Benjamin D. Humphreys, *Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis*, Journal of the American Society of Nephronlogy, December 2018, ASN.2018090912, DOI: https://doi.org/10.1681/ASN.2018090912
Require DOI?	Yes
Persistent ID	https://doi.org/10.25548/14-4KG6
Principal Investigator	Benjamin Humphreys, WUSTL
Data Provider	Washington University, St. Louis
Consortium	RBK
Creation Time	2018-11-09 14:00:15
Last Modified Time	2018-12-12 17:24:04

Sequencing Study Collection (showing all 1 results)

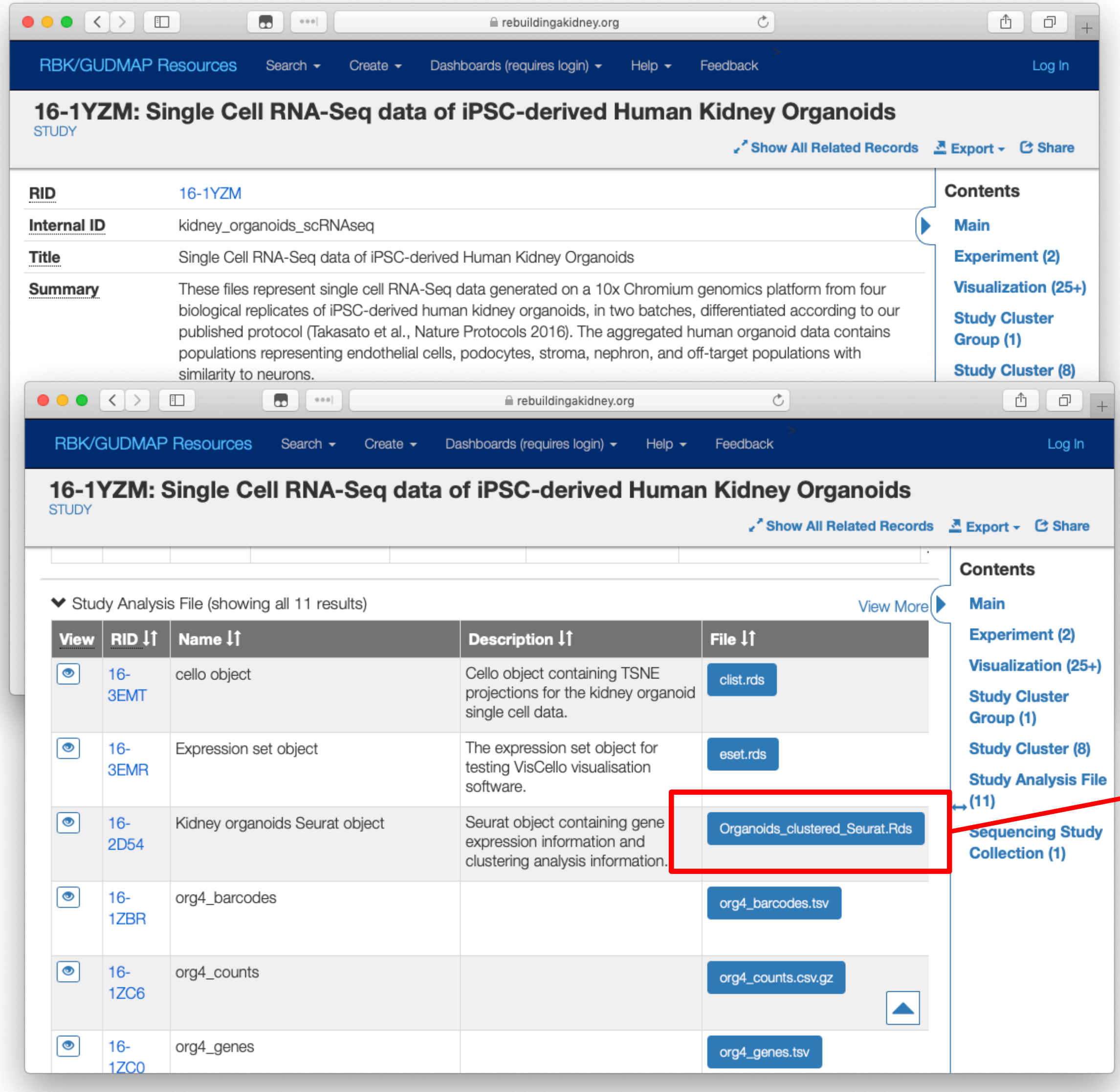
View	RID	Internal ID	Title	Funding	Principal Investigator	Data P
👁	14-4KBA	Mouse_kidney_snRNA_seq	Single-Nucleus versus Single-Cell RNA		Benjamin Humphreys, WUSTL	Washin Univers

- New approach to scientific rigor and reproducibility
  - Data followed from slide to database image
  - ”Largest possible” supplementary data
- Collections designed around specific structures

Multiple layers with scRNA-seq

1. RAW sequencing files (fastq)
2. Processed gene expression matrix files (txt)
3. R objects of analysis (Rds, e.g., Seurat)
4. Static visualization tools
5. Interactive visualization tools

Direct linking to data for efficiency with large datasets



16-1YZM: Single Cell RNA-Seq data of iPSC-derived Human Kidney Organoids
STUDY

RID	16-1YZM
Internal ID	kidney_organoids_scRNAseq
Title	Single Cell RNA-Seq data of iPSC-derived Human Kidney Organoids
Summary	These files represent single cell RNA-Seq data generated on a 10x Chromium genomics platform from four biological replicates of iPSC-derived human kidney organoids, in two batches, differentiated according to our published protocol (Takasato et al., Nature Protocols 2016). The aggregated human organoid data contains populations representing endothelial cells, podocytes, stroma, nephron, and off-target populations with similarity to neurons.

16-1YZM: Single Cell RNA-Seq data of iPSC-derived Human Kidney Organoids
STUDY

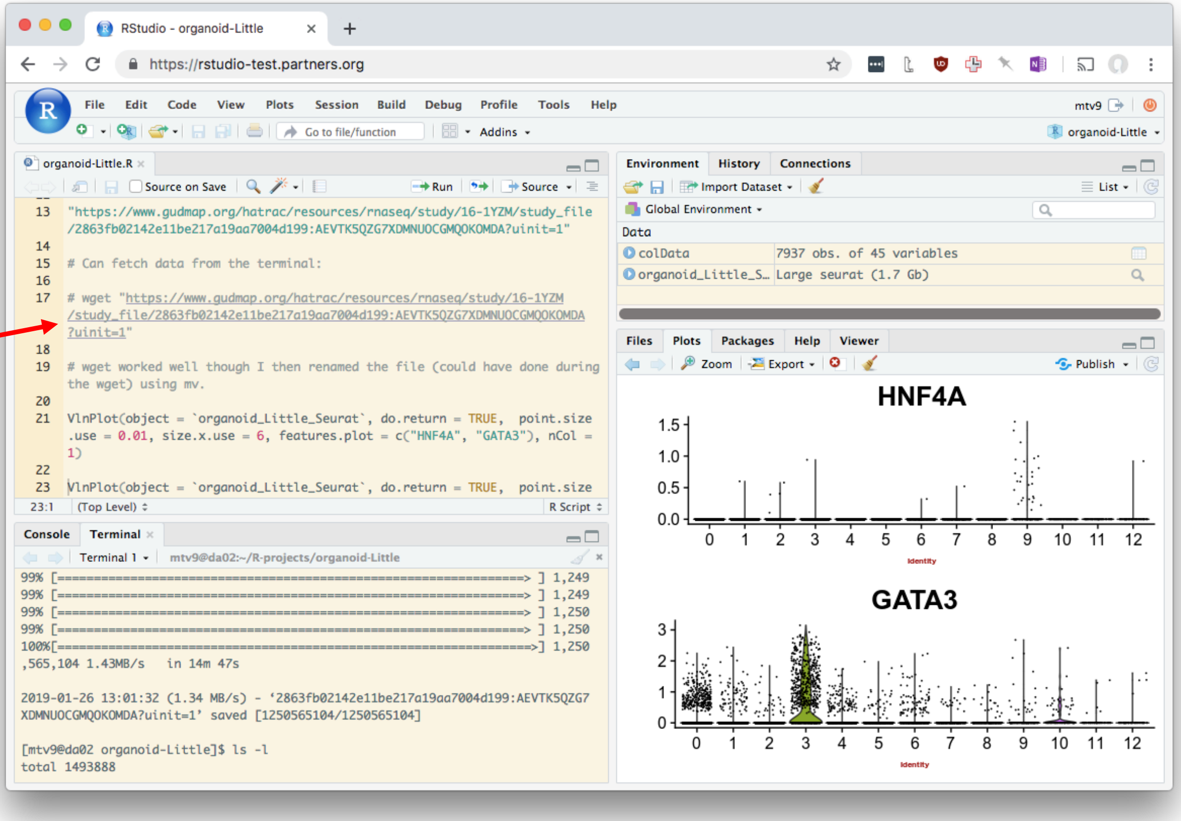
Study Analysis File (showing all 11 results) View More

View	RID	Name	Description	File
👁	16-3EMT	cello object	Cello object containing TSNE projections for the kidney organoid single cell data.	clist.rds
👁	16-3EMR	Expression set object	The expression set object for testing VisCello visualisation software.	eset.rds
👁	16-2D54	Kidney organoids Seurat object	Seurat object containing gene expression information and clustering analysis information.	Organoids_clustered_Seurat.Rds
👁	16-1ZBR	org4_barcodes		org4_barcodes.tsv
👁	16-1ZC6	org4_counts		org4_counts.csv.gz
👁	16-1ZC0	org4_genes		org4_genes.tsv

- New approach to scientific rigor and reproducibility
  - Data followed from slide to database image
  - "Largest possible" supplementary data
- Collections designed around specific structures

Multiple layers with scRNA-seq

1. RAW sequencing files (fastq)



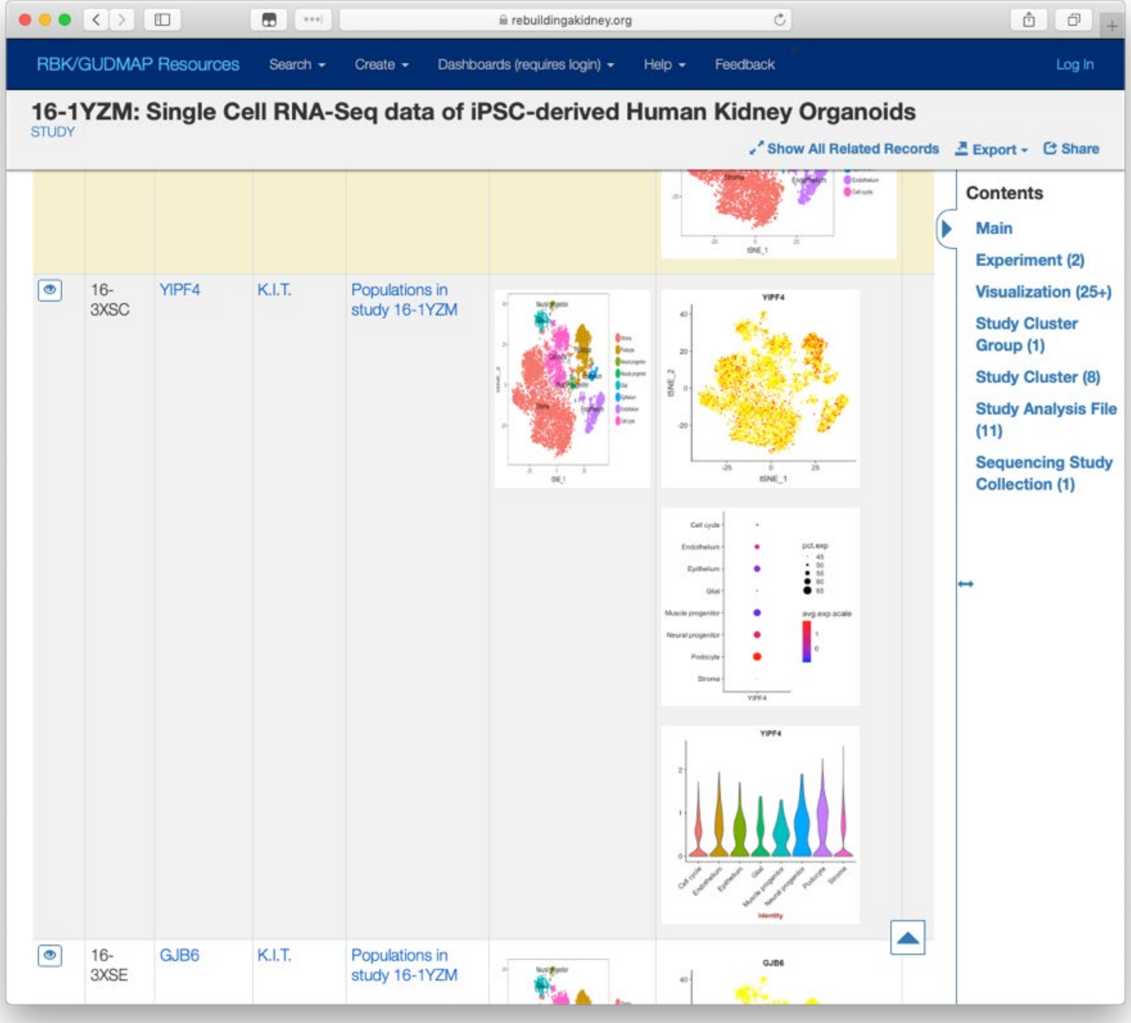
RStudio - organoid-Little
https://rstudio-test.partners.org
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
mtv9
organoid-Little
organoid-Little.R
Source on Save
Run
Source
Environment History Connections
Import Dataset
List
Global Environment
Data
Files Plots Packages Help Viewer
Zoom
Export
Publish
HNF4A
GATA3
Identity
0 1 2 3 4 5 6 7 8 9 10 11 12
0.0 0.5 1.0 1.5
0 1 2 3
Console Terminal
Terminal 1
mtv9@da02:~/R-projects/organoid-Little
23:1 (Top Level) R Script

colData	7937 obs. of 45 variables
organoid_Little_S…	Large seurat (1.7 Gb)

```
"https://www.gudmap.org/hatrac/resources/rnaseq/study/16-1YZM/study_file
/2863fb02142e11be217a19aa7004d199:AEVTK5QZG7XDMNUOCGMQOKOMDA?uinit=1"

# Can fetch data from the terminal:

# wget "https://www.gudmap.org/hatrac/resources/rnaseq/study/16-1YZM
/study_file/2863fb02142e11be217a19aa7004d199:AEVTK5QZG7XDMNUOCGMQOKOMDA
?uinit=1"

# wget worked well though I then renamed the file (could have done during
the wget) using mv.

VlnPlot(object = `organoid_Little_Seurat`, do.return = TRUE,  point.size
.use = 0.01, size.x.use = 6, features.plot = c("HNF4A", "GATA3"), nCol =
1)

VlnPlot(object = `organoid_Little_Seurat`, do.return = TRUE,  point.size
```

```
99% [=====================================================================> ] 1,249
99% [=====================================================================> ] 1,249
99% [=====================================================================> ] 1,250
99% [=====================================================================> ] 1,250
100%[======================================================================>] 1,250
,565,104 1.43MB/s   in 14m 47s

2019-01-26 13:01:32 (1.34 MB/s) - ‘2863fb02142e11be217a19aa7004d199:AEVTK5QZG7
XDMNUOCGMQOKOMDA?uinit=1’ saved [1250565104/1250565104]

[mtv9@da02 organoid-Little]$ ls -l
total 1493888
```

scRNA-seq Visualizations - Accessibility



16-1YZM: Single Cell RNA-Seq data of iPSC-derived Human Kidney Organoids

STUDY

16-3XSC	YIPF4	K.I.T.	Populations in study 16-1YZM
16-3XSE	GJB6	K.I.T.	Populations in study 16-1YZM

- New approach to scientific rigor and reproducibility
  - Data followed from slide to database image
  - ”Largest possible” supplementary data
- Collections designed around specific structures

Multiple layers with scRNA-seq

1. RAW sequencing files (fastq)
2. Processed gene expression matrix files (txt)
3. R objects of analysis (Rds, e.g. Seurat)
4. Static visualization tools
5. Interactive visualization tools

Data storage and levels of sharing, accessibility to open-source tools.

- **Sequence data – a layered approach**
  - "Processed" data is useful to a wider range of researchers.
    - Count files are ready for analysis without computationally intensive genome aligners in HPCs. More researchers can use such data immediately.
  - Privacy of participants - what is reasonable to share even with full consent?
    - The ability to de-identify participants from limited sequencing data expands before thoughtful policy will catch up. Think beyond the contractual protection to anticipate while maintaining data availability.
  - What intermediate products are available?
    - R objects like Seurat capture analysis decisions for a data generators fine analysis and can be used by less experienced researchers subsequently.
  - Processed data more freely shareable, but what happens when references change?
    - As data ages, re-alignment may be necessary as reference genomes change and improve.

mRNA-Seq Reanalysis Progress

# Studies	# Experiments	# Replicates	# Files
49	184	602	851

First Round of execution

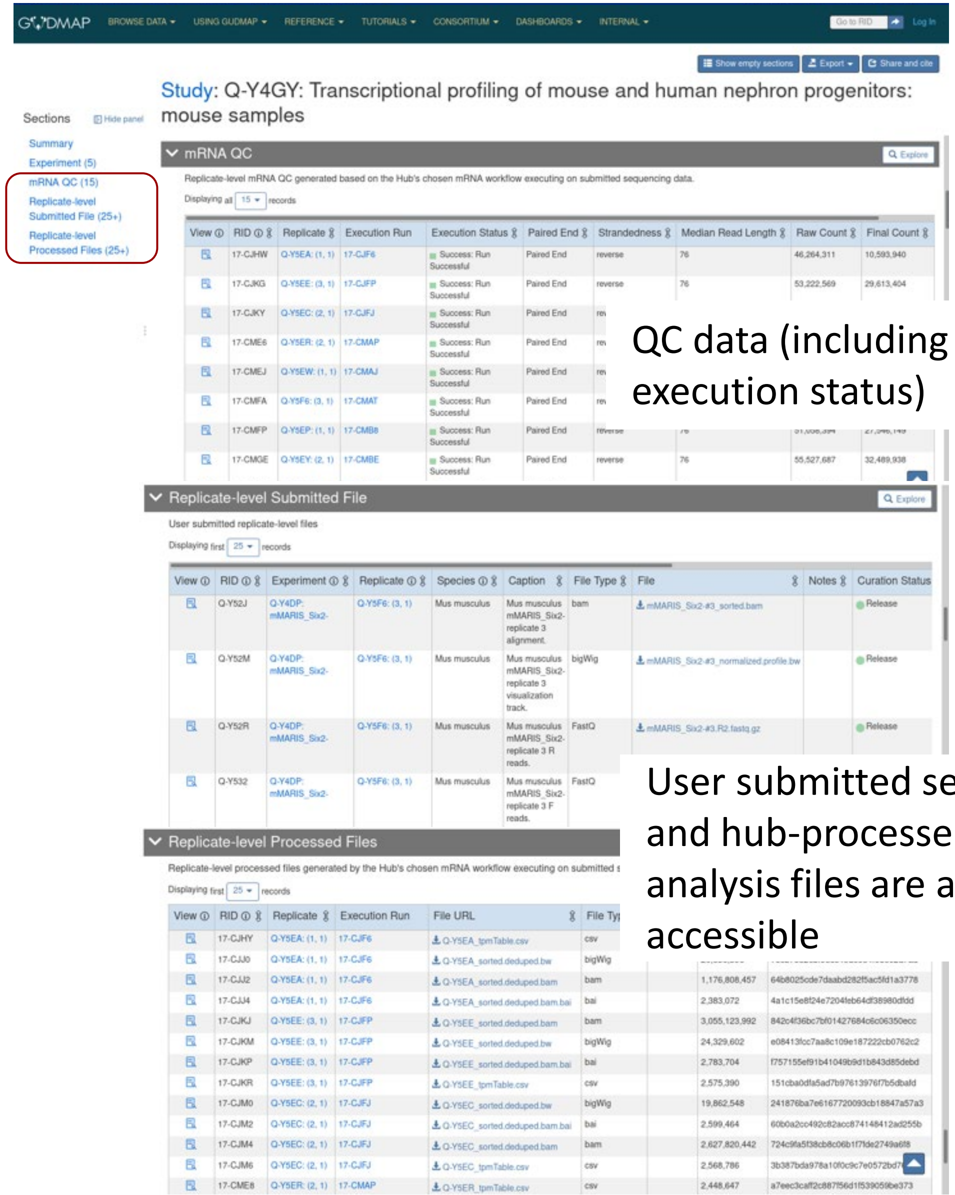
Execution Status	Count (# Replicates)	Description
Success	110 (18.3%)	
Error	492 (81.7%)	
- Metadata	311 (51.7%)	- Validate: Species, Paired-End, Strandedness, Spikes-in - No metadata or mismatched - Incorrect sequencing type (e.g. ChiP-Seq instead of mRNA-Seq)
- File	181 (30.0%)	- Mismatched #Reads of R1 and R2, multiple runs, missing files - Not fastq structure (e.g. fastq+bam)

After a few rounds of resolution and execution

Success	- 549 mRNA-Seq replicates (45 studies) - Re-labeled 3 mRNA-Seq replicates to ChiP-Seq (1 study)
Outstanding issues	- Missing files: 3 replicates - Conflicting files: 50 replicates (3 studies)

As of 08/10/2021

mRNA-Seq QC, Processed Files, and Visualization



Study: Q-Y4GY: Transcriptional profiling of mouse and human nephron progenitors: mouse samples

mRNA QC

Replicate-level mRNA QC generated based on the Hub's chosen mRNA workflow executing on submitted sequencing data.

Displaying all 15 records

View	RID	Replicate	Execution Run	Execution Status	Paired End	Strandedness	Median Read Length	Raw Count	Final Count
	17-CJHW	Q-Y5EA: (1, 1)	17-CJF6	Success: Run Successful	Paired End	reverse	76	46,264,311	10,593,940
	17-CJKG	Q-Y5EE: (3, 1)	17-CJFP	Success: Run Successful	Paired End	reverse	76	53,222,569	29,613,404
	17-CJKY	Q-Y5EC: (2, 1)	17-CJFJ	Success: Run Successful	Paired End	rev			
	17-CME6	Q-Y5ER: (2, 1)	17-CMAP	Success: Run Successful	Paired End	rev			
	17-CMEJ	Q-Y5EW: (1, 1)	17-CMAJ	Success: Run Successful	Paired End	rev			
	17-CMFA	Q-Y5F6: (3, 1)	17-CMAT	Success: Run Successful	Paired End	rev			
	17-CMFP	Q-Y5EP: (1, 1)	17-CMB8	Success: Run Successful	Paired End	reverse	76	[illegible]	[illegible]
	17-CMGE	Q-Y5EY: (2, 1)	17-CMBE	Success: Run Successful	Paired End	reverse	76	55,527,687	32,489,938

Replicate-level Submitted File

User submitted replicate-level files

Displaying first 25 records

View	RID	Experiment	Replicate	Species	Caption	File Type	File	Notes	Curation Status
	Q-Y52J	Q-Y4DP: mMARIS_Six2-	Q-Y5F6: (3, 1)	Mus musculus	Mus musculus mMARIS_Six2- replicate 3 alignment.	bam	mMARIS_Six2-#3_sorted.bam		Release
	Q-Y52M	Q-Y4DP: mMARIS_Six2-	Q-Y5F6: (3, 1)	Mus musculus	Mus musculus mMARIS_Six2- replicate 3 visualization track.	bigWig	mMARIS_Six2-#3_normalized.profile.bw		Release
	Q-Y52R	Q-Y4DP: mMARIS_Six2-	Q-Y5F6: (3, 1)	Mus musculus	Mus musculus mMARIS_Six2- replicate 3 R reads.	FastQ	mMARIS_Six2-#3.R2.fastq.gz		Release
	Q-Y532	Q-Y4DP: mMARIS_Six2-	Q-Y5F6: (3, 1)	Mus musculus	Mus musculus mMARIS_Six2- replicate 3 F reads.	FastQ			

Replicate-level Processed Files

Replicate-level processed files generated by the Hub's chosen mRNA workflow executing on submitted s

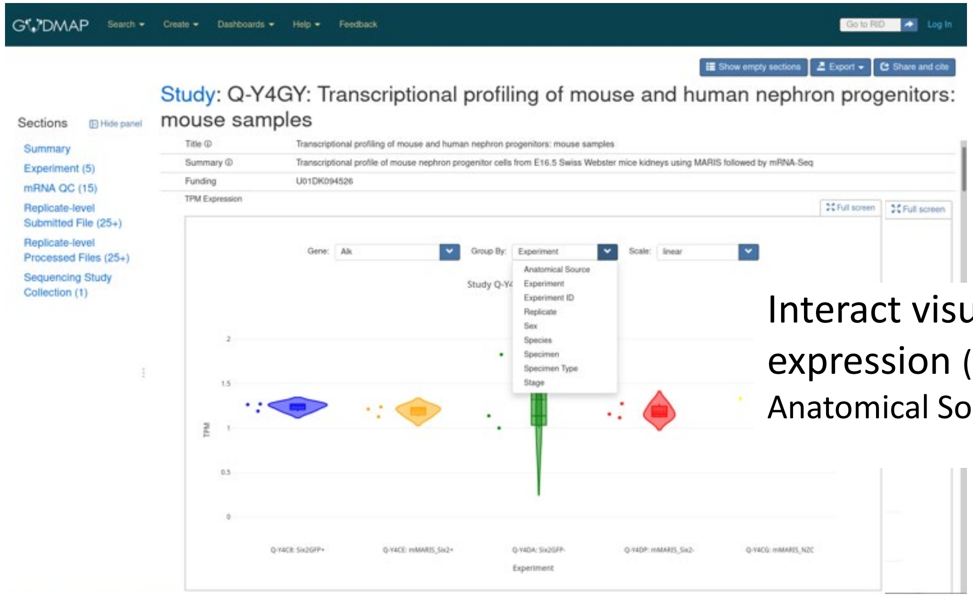
Displaying first 25 records

View	RID	Replicate	Execution Run	File URL	File Typ		
	17-CJHY	Q-Y5EA: (1, 1)	17-CJF6	Q-Y5EA_tpmTable.csv	csv		
	17-CJJ0	Q-Y5EA: (1, 1)	17-CJF6	Q-Y5EA_sorted.deduped.bw	bigWig		
	17-CJJ2	Q-Y5EA: (1, 1)	17-CJF6	Q-Y5EA_sorted.deduped.bam	bam	1,176,808,457	64b8025cde7daabd282f5ac5fd1a3778
	17-CJJ4	Q-Y5EA: (1, 1)	17-CJF6	Q-Y5EA_sorted.deduped.bam.bai	bai	2,383,072	4a1c15e8f24e7204feb64df38980dfdd
	17-CJKJ	Q-Y5EE: (3, 1)	17-CJFP	Q-Y5EE_sorted.deduped.bam	bam	3,055,123,992	842c4f36bc7bf01427684c6c06350ecc
	17-CJKM	Q-Y5EE: (3, 1)	17-CJFP	Q-Y5EE_sorted.deduped.bw	bigWig	24,329,602	e08413fcc7aa8c109e187222cb0762c2
	17-CJKP	Q-Y5EE: (3, 1)	17-CJFP	Q-Y5EE_sorted.deduped.bam.bai	bai	2,783,704	f757155ef91b41049b9d1b843d85debd
	17-CJKR	Q-Y5EE: (3, 1)	17-CJFP	Q-Y5EE_tpmTable.csv	csv	2,575,390	151cba0dfa5ad7b97613976f7b5dbafd
	17-CJM0	Q-Y5EC: (2, 1)	17-CJFJ	Q-Y5EC_sorted.deduped.bw	bigWig	19,862,548	241876ba7e6167720093cb18847a57a3
	17-CJM2	Q-Y5EC: (2, 1)	17-CJFJ	Q-Y5EC_sorted.deduped.bam.bai	bai	2,599,464	60b0a2cc492c82acc874148412ad255b
	17-CJM4	Q-Y5EC: (2, 1)	17-CJFJ	Q-Y5EC_sorted.deduped.bam	bam	2,627,820,442	724c9fa5f38cb8c06b1f7fde2749a6f8
	17-CJM6	Q-Y5EC: (2, 1)	17-CJFJ	Q-Y5EC_tpmTable.csv	csv	2,568,786	3b387bda978a10f0c9c7e0572bd7
	17-CME8	Q-Y5ER: (2, 1)	17-CMAP	Q-Y5ER_tpmTable.csv	csv	2,448,647	a7eec3caff2c887f56d1f539059be373

QC data (including execution status)

User submitted seq files and hub-processed analysis files are all accessible

https://www.gudmap.org/id/Q-Y4GY

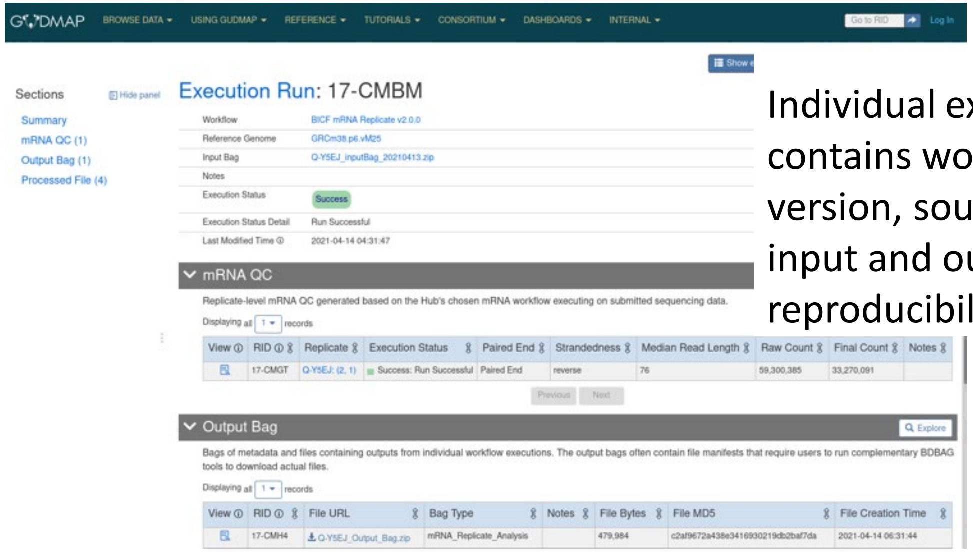


Study: Q-Y4GY: Transcriptional profiling of mouse and human nephron progenitors: mouse samples

Title	Transcriptional profiling of mouse and human nephron progenitors: mouse samples
Summary	Transcriptional profile of mouse nephron progenitor cells from E16.5 Swiss Webster mice kidneys using MARIS followed by mRNA-Seq
Funding	U01DK094526
TPM Expression	

Interact visualization of TPM expression (group by Experiment, Anatomical Source, Stage, etc)

https://dev.gudmap.org/id/Q-Y4GY (dev server only)



Execution Run: 17-CMBM

Workflow	BICF mRNA Replicate v2.0.0
Reference Genome	GRCm38.p6.vM25
Input Bag	Q-Y5EJ_inputBag_20210413.zip
Notes	
Execution Status	Success
Execution Status Detail	Run Successful
Last Modified Time	2021-04-14 04:31:47

mRNA QC

Replicate-level mRNA QC generated based on the Hub's chosen mRNA workflow executing on submitted sequencing data.

Displaying all 1 records

View	RID	Replicate	Execution Status	Paired End	Strandedness	Median Read Length	Raw Count	Final Count	Notes
	17-CMGT	Q-Y5EJ: (2, 1)	Success: Run Successful	Paired End	reverse	76	59,300,385	33,270,091	

Output Bag

Bags of metadata and files containing outputs from individual workflow executions. The output bags often contain file manifests that require users to run complementary BDBAG tools to download actual files.

Displaying all 1 records

View	RID	File URL	Bag Type	Notes	File Bytes	File MD5	File Creation Time
	17-CMH4	Q-Y5EJ_Output_Bag.zip	mRNA_Replicate_Analysis		479,984	c2af9672a438e3416930219db2baf7da	2021-04-14 06:31:44

Individual execution run contains workflow definition, version, source code URL, input and output for reproducibility

https://www.gudmap.org/id/17-BMCM

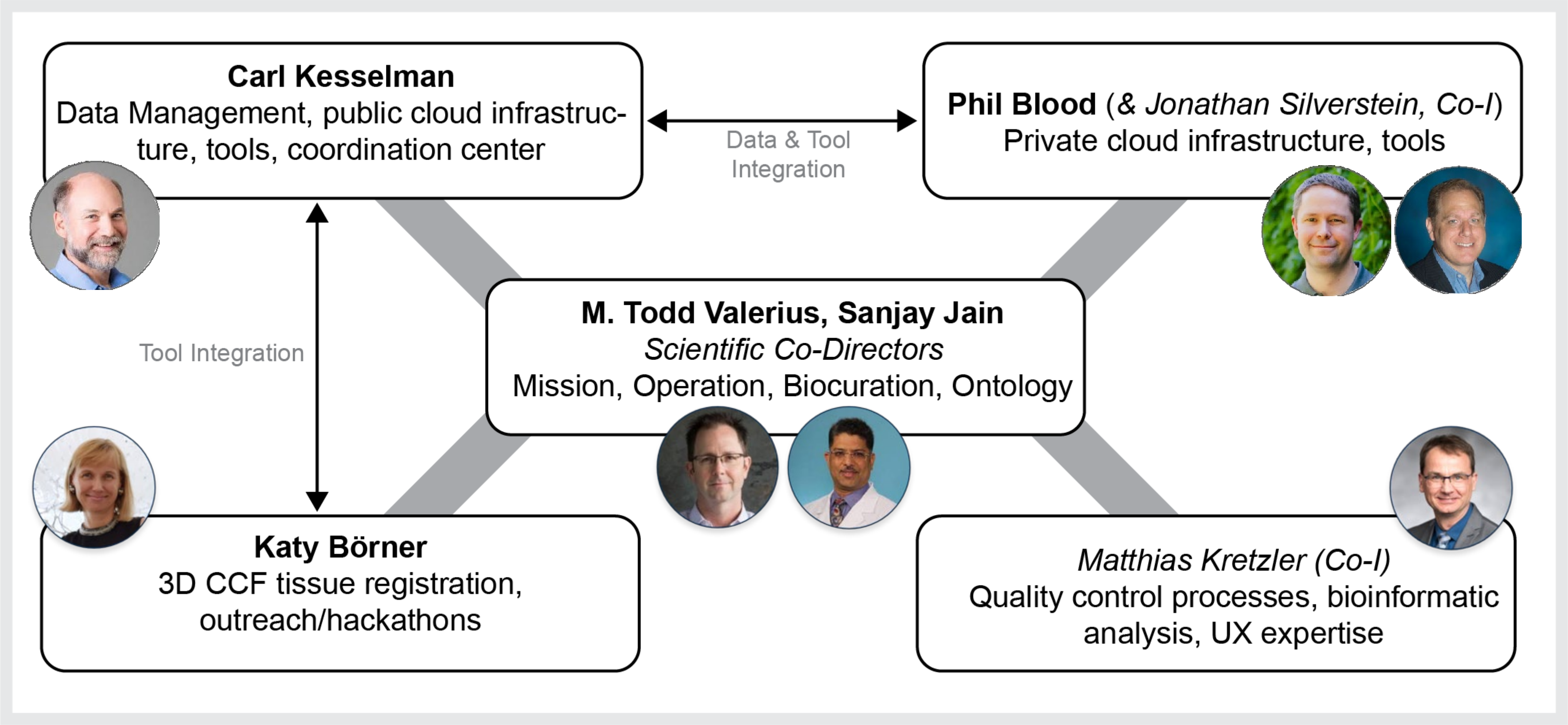
Data storage and levels of sharing, accessibility to open-source tools.

- **Sequence data – a layered approach**
  - "Processed" data is useful to a wider range of researchers.
    - Count files are ready for analysis without computationally intensive genome aligners in HPCs. More researchers can use such data immediately.
  - Privacy of participants - what is reasonable to share even with full consent?
    - The ability to de-identify participants from limited sequencing data expands before thoughtful policy will catch up. Think beyond the contractual protection to anticipate while maintaining data availability.
  - What intermediate products are available?
    - R objects like Seurat capture analysis decisions for a data generators fine analysis and can be used by less experienced researchers subsequently.
  - Processed data more freely shareable, but what happens when references change?
    - As data ages, re-alignment may be necessary as reference genomes change and improve.

Recommendations:

1. Consider asking your core to use programmatically published base processing pipelines.
2. Publish these AND analysis code to a coding repository. (Contemporaneous documentation saves time later.)
3. Consider what intermediate layers should be protected (RAW versus counts).
4. Publish analysis intermediates.

ATLAS-D2K Team



Carl Kesselman
Data Management, public cloud infrastructure, tools, coordination center

Phil Blood (*& Jonathan Silverstein, Co-I*)
Private cloud infrastructure, tools

M. Todd Valerius, Sanjay Jain
Scientific Co-Directors
Mission, Operation, Biocuration, Ontology

Katy Börner
3D CCF tissue registration, outreach/hackathons

Matthias Kretzler (Co-I)
Quality control processes, bioinformatic analysis, UX expertise

NIH National Institute of Diabetes and Digestive and Kidney Diseases

DK135157, DK133090