Research Objects for Everyday Use

Carl Kesselman

Dean’s Professor, Industrial and Systems Engineering
University of Southern California

USC Viterbi
School of Engineering

Information Sciences Institute

What does it mean to have a scientific "result"

- Others have to "know" about it
- Others have to be able to validate it
  - Reproduce the method and achieve the same result
  - Achieve the same result via a different method
  - Reuse the result in a new method

"Non-reproducible single occurrences are of no significance to science."

– Karl Popper, 1959. The logic of scientific discovery. Hutchinson, London, United Kingdom.

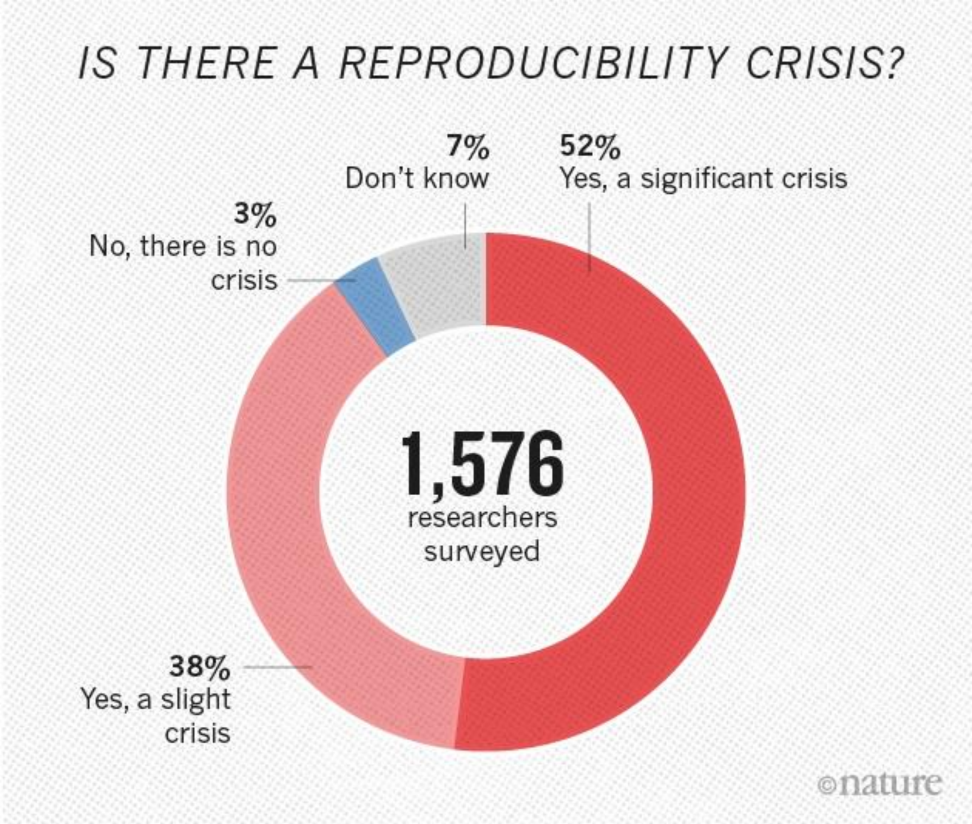
Science must be reproducible..

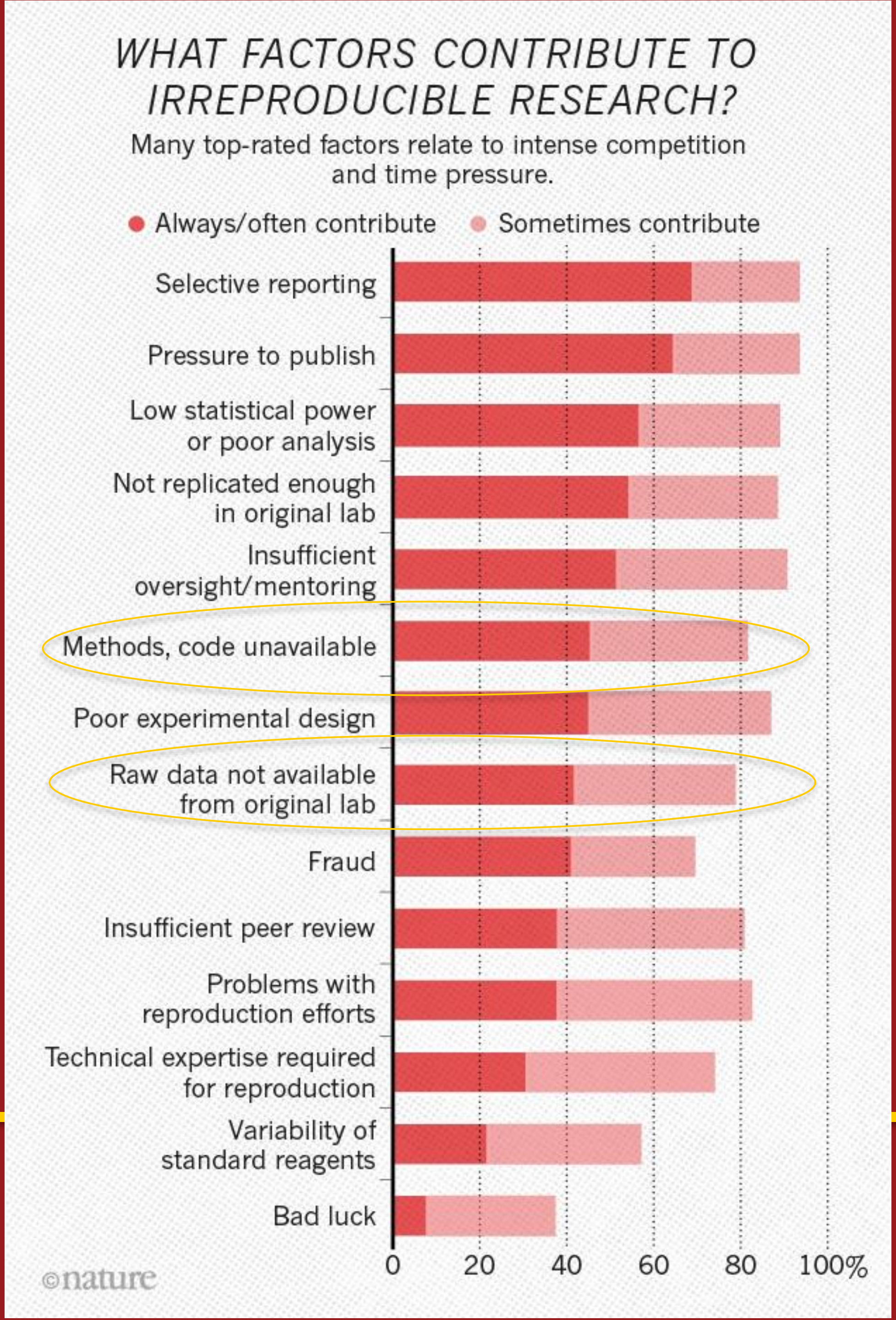
Nature May 25, 2016

- “More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from *Nature*'s survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research."

Where are we now?

Only 10% of published results are reproducible




WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?
Many top-rated factors relate to intense competition and time pressure.
● Always/often contribute
● Sometimes contribute
Selective reporting
Pressure to publish
Low statistical power or poor analysis
Not replicated enough in original lab
Insufficient oversight/mentoring
Methods, code unavailable
Poor experimental design
Raw data not available from original lab
Fraud
Insufficient peer review
Problems with reproduction efforts
Technical expertise required for reproduction
Variability of standard reagents
Bad luck
0
20
40
60
80
100%
©nature

Lets focus on the data…

Reproducibility of data

- Data is only useful if we can figure out:
  - Where it is
  - What it is
  - How it was created
- Process and product: the data lifecycle
  - Initial data acquisition (perhaps from experiment)
  - Computing, and analysis
  - Publication

Reproducible Data is FAIR

- Findable
  - – identified by a unique identifier, characterized by rich metadata
- Accessible
  - – standard protocol with access control, metadata accessible even when the data is not,
- Interoperable
  - – by standardized terms to describe it
- Reusable
  - – Accurate and relevant attributes.

GOAL: FAIR Collaboration

- Lets make all data produced in an investigation FAIR
  - Not just final “published” result, but all results
  - Need to scale application of FAIRness principals down to level of daily practice
- Requires accurate descriptions of data
  - Characteristics of data element
  - Relationships between data elements
- Requires robust naming of the data products

Knowledge Turns and Publication

- Publication
  - Slow turn around, polished content and presentation
  - Human processes and audience
- µPublication
  - Rapid turn around, incremental raw data exchange
  - Human and machine processes and audience

What does it mean to have research objects in daily use

Get users thinking about lots of publication events

- Prepare information for sharing
- Package it up so that its accessible
- Name the package so that it can be identified.

What can we do to make these things easier?

Identifiers are important, but too hard!!

Lets consider a range of identifier user cases

- We need to separate out issues of naming from persistence
- Digital Object Identifiers (DOIs) are designed for archival objects
  - – You don't mind a DOI unless you really "mean it"
- What about identifying intermediate and temporary data
  - – Can also benefit from unambiguous naming
- Two options:
  - – Local identifiers, e.g. assession numbers (often not actionable)
  - – Alternative identifier systems such as ARK

Minimal Identifiers (Minids)

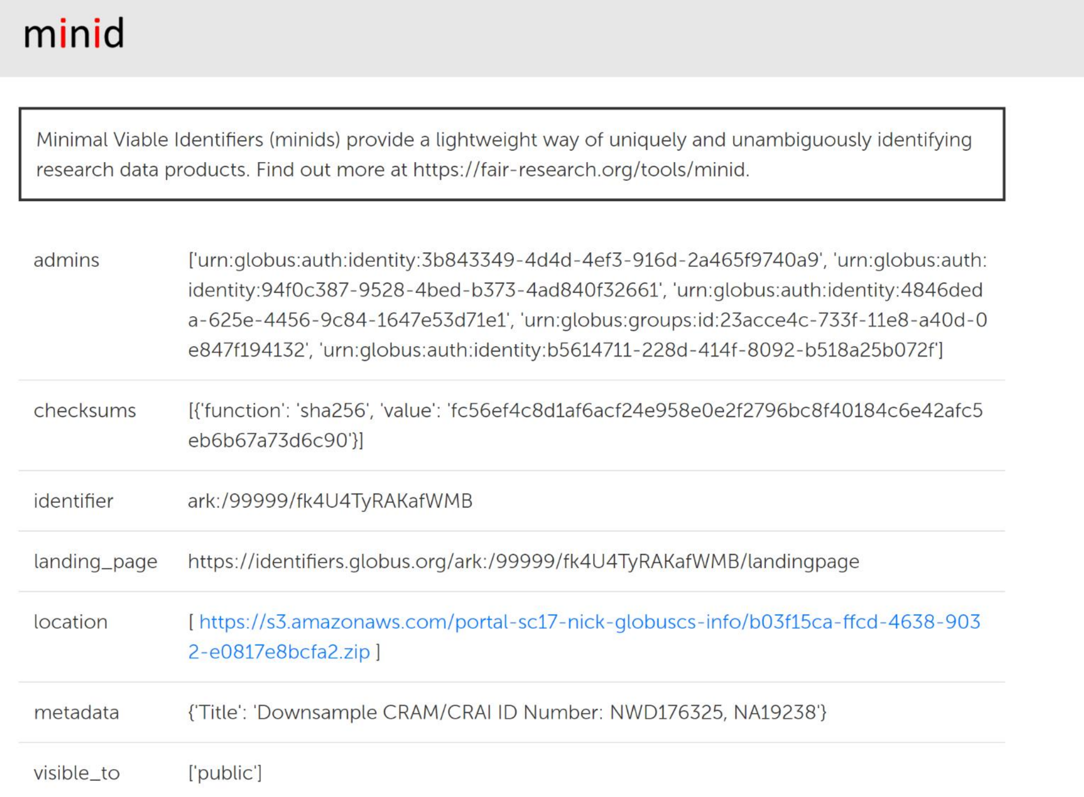
Lightweight identifiers that support simple creation and use

- Unique identifier (ARK)
  - E.g., /ark:/57799/b9040f
  - Or compact identifier (minid:b9040f)
- Minimal metadata (creator, date, name)
- Checksum ensures data is verifiable
- Service to provide the landing page
- Easy to use: CLI, Python SDK, R SDK, JSON-based REST API

Why and When use Minids?

Naming intermediate data

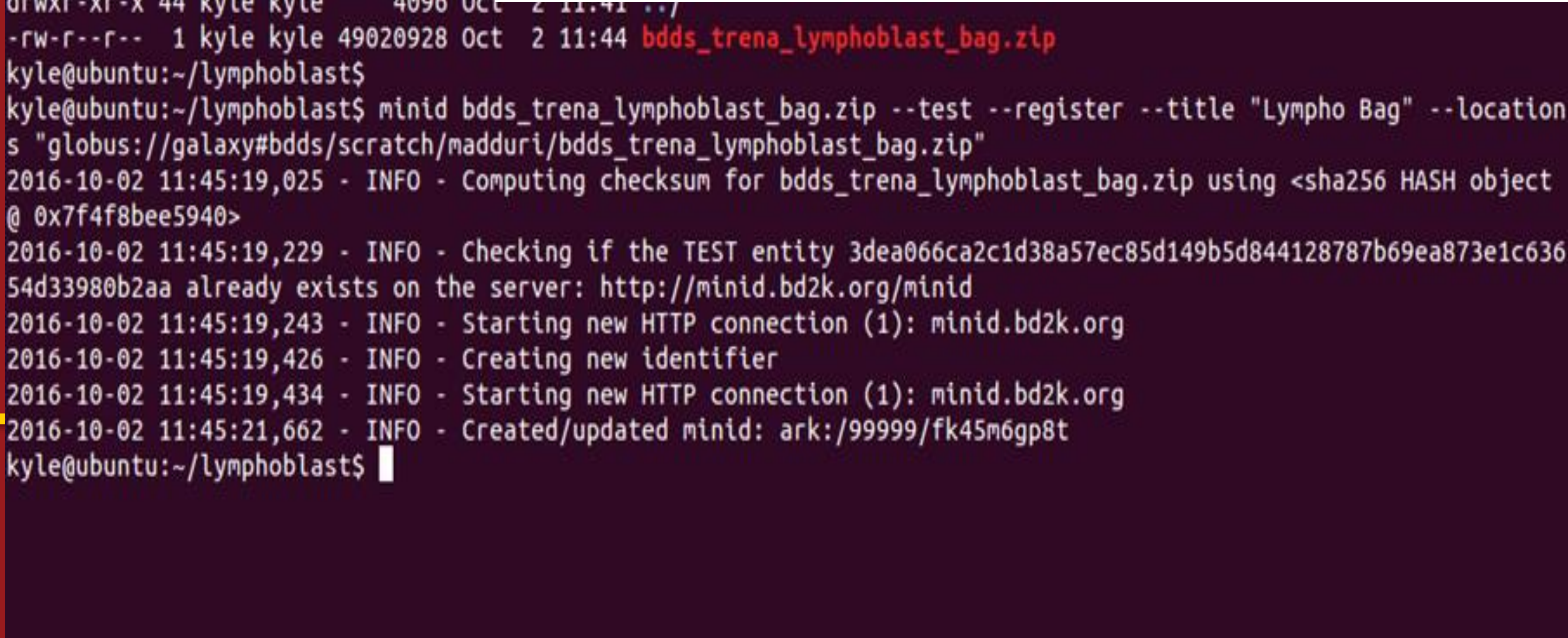
- Quickly associate a lightweight identifier
- Validate data integrity
- Lookup identifiers based on checksum
- CLI or Python/R client



minid

Minimal Viable Identifiers (minids) provide a lightweight way of uniquely and unambiguously identifying research data products. Find out more at https://fair-research.org/tools/minid.

admins	['urn:globus:auth:identity:3b843349-4d4d-4ef3-916d-2a465f9740a9', 'urn:globus:auth:identity:94f0c387-9528-4bed-b373-4ad840f32661', 'urn:globus:auth:identity:4846deda-625e-4456-9c84-1647e53d71e1', 'urn:globus:groups:id:23acce4c-733f-11e8-a40d-0e847f194132', 'urn:globus:auth:identity:b5614711-228d-414f-8092-b518a25b072f']
checksums	[{'function': 'sha256', 'value': 'fc56ef4c8d1af6acf24e958e0e2f2796bc8f40184c6e42afc5eb6b67a73d6c90'}]
identifier	ark:/99999/fk4U4TyRAKafWMB
landing_page	https://identifiers.globus.org/ark:/99999/fk4U4TyRAKafWMB/landingpage
location	[https://s3.amazonaws.com/portal-sc17-nick-globuscs-info/b03f15ca-ffcd-4638-9032-e0817e8bcfa2.zip]
metadata	{'Title': 'Downsample CRAM/CRAI ID Number: NWD176325, NA19238'}
visible_to	['public']



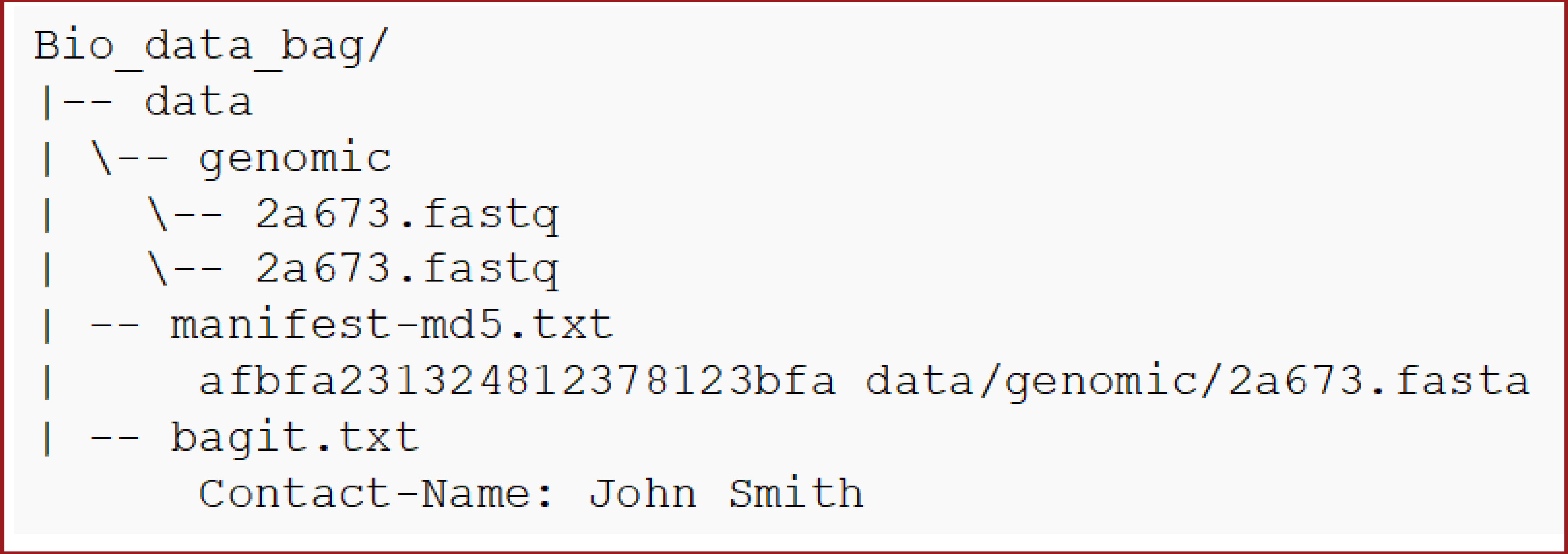
kyle@ubuntu: ~/lymphoblast

```
kyle@ubuntu:~/lymphoblast$ ll
total 47880
drwxrwxr-x  2 kyle kyle     4096 Oc
drwxr-xr-x 44 kyle kyle     4096 Oct
-rw-r--r--  1 kyle kyle 49020928 Oct  2 11:44 bdds_trena_lymphoblast_bag.zip
kyle@ubuntu:~/lymphoblast$
kyle@ubuntu:~/lymphoblast$ minid bdds_trena_lymphoblast_bag.zip --test --register --title "Lympho Bag" --locations "globus://galaxy#bdds/scratch/madduri/bdds_trena_lymphoblast_bag.zip"
2016-10-02 11:45:19,025 - INFO - Computing checksum for bdds_trena_lymphoblast_bag.zip using <sha256 HASH object @ 0x7f4f8bee5940>
2016-10-02 11:45:19,229 - INFO - Checking if the TEST entity 3dea066ca2c1d38a57ec85d149b5d844128787b69ea873e1c63654d33980b2aa already exists on the server: http://minid.bd2k.org/minid
2016-10-02 11:45:19,243 - INFO - Starting new HTTP connection (1): minid.bd2k.org
2016-10-02 11:45:19,426 - INFO - Creating new identifier
2016-10-02 11:45:19,434 - INFO - Starting new HTTP connection (1): minid.bd2k.org
2016-10-02 11:45:21,662 - INFO - Created/updated minid: ark:/99999/fk45m6gp8t
kyle@ubuntu:~/lymphoblast$
```

USC Viterbi
School of Engineering

Large Multi-File Datasets: Big Data Bags

- Profile on the BagIt specification
  - Payload: arbitrary content
  - Tags: metadata describing payload
  - Checksums: to verify content
- Content may be "missing"
  - Missing content must be listed in "fetch.txt"
  - Fetch entries list local name in data directory, and URL of where to fetch data
- Enhancements to support big data
  - Different data access protocols (Globus, HTTP, iRODS, S3)
  - Research Object metadata



```
Bio_data_bag/
|-- data
| \-- genomic
|   \-- 2a673.fastq
|   \-- 2a673.fastq
| -- manifest-md5.txt
|     afbfa231324812378123bfa data/genomic/2a673.fasta
| -- bagit.txt
      Contact-Name: John Smith
```

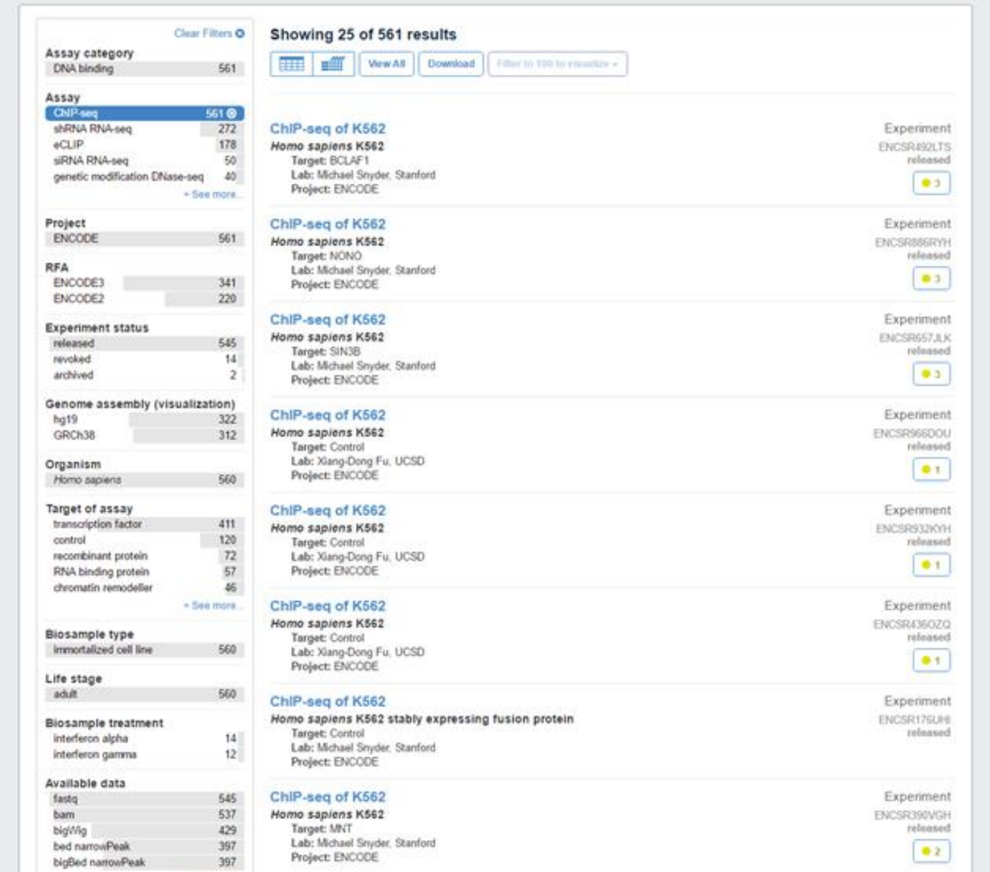
Research Objects: rich metadata for bags

```
{
  "@context": {
      "@vocab": "http://purl.org/dc/terms/",
      "dcmi": "http://purl.org/dc/dcmitype/Dataset"
  },
  "@id": "../../data/numbers.csv",
  "@type": "dcmi:Dataset",
  "title": "CSV files of beverage consumption",
 ”description": "A CSV file listing the number of cups consumed per person.”
}
```

Why and When use Minids?

Large, multi-file datasets

- Integration with BDBag allows for content independent verification (i.e., holey bags)
- Streamline download and references to complex multi-file datasets
- Example: Encode2Bag



Assay category	
DNA binding	561
Assay	
ChIP-seq	561
shRNA RNA-seq	272
eCLIP	178
siRNA RNA-seq	50
genetic modification DNase-seq	40
Project	
ENCODE	561
RFA	
ENCODE3	341
ENCODE2	220
Experiment status	
released	545
revoked	14
archived	2
Genome assembly (visualization)	
hg19	322
GRCh38	312
Organism	
Homo sapiens	560
Target of assay	
transcription factor	411
control	120
recombinant protein	72
RNA binding protein	57
chromatin remodeller	46
Biosample type	
immortalized cell line	560
Life stage	
adult	560
Biosample treatment	
interferon alpha	14
interferon gamma	12
Available data	
fastq	545
bam	537
bigWig	429
bed narrowPeak	397
bigBed narrowPeak	397

Showing 25 of 561 results

ChIP-seq of K562
Homo sapiens K562
Target: BCLAF1
Lab: Michael Snyder, Stanford
Project: ENCODE
Experiment
ENCSR492LTS
released
● 3

ChIP-seq of K562
Homo sapiens K562
Target: NONO
Lab: Michael Snyder, Stanford
Project: ENCODE
Experiment
ENCSR886RYH
released
● 3

ChIP-seq of K562
Homo sapiens K562
Target: SIN3B
Lab: Michael Snyder, Stanford
Project: ENCODE
Experiment
ENCSR657JLK
released
● 3

ChIP-seq of K562
Homo sapiens K562
Target: Control
Lab: Xiang-Dong Fu, UCSD
Project: ENCODE
Experiment
ENCSR966DOU
released
● 1

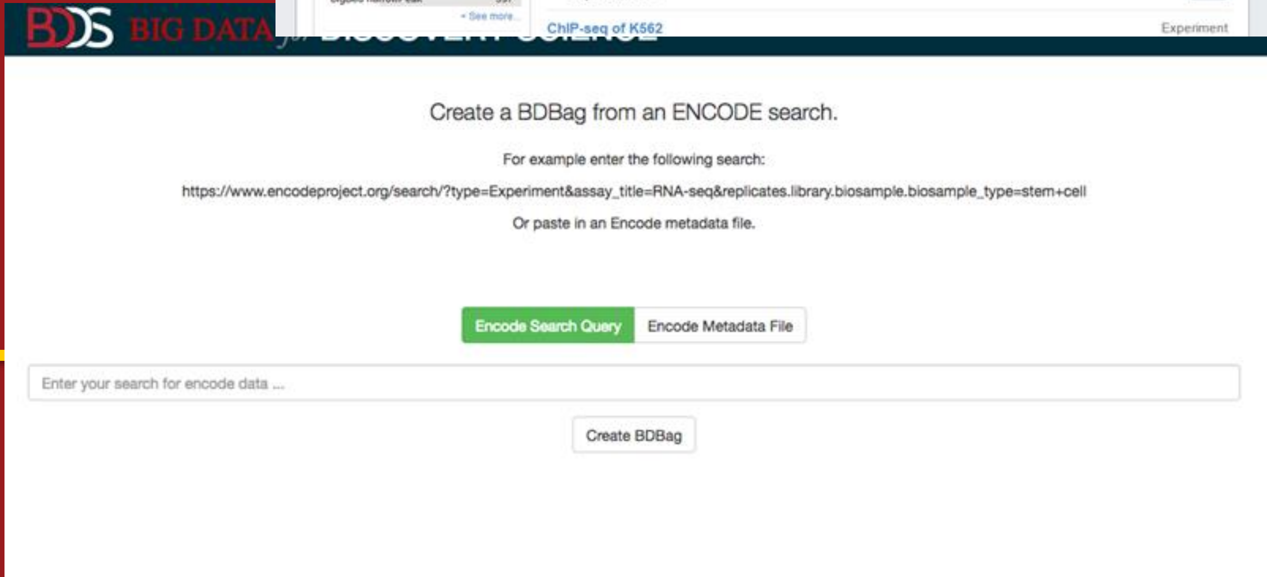
ChIP-seq of K562
Homo sapiens K562
Target: Control
Lab: Xiang-Dong Fu, UCSD
Project: ENCODE
Experiment
ENCSR932KYH
released
● 1

ChIP-seq of K562
Homo sapiens K562
Target: Control
Lab: Xiang-Dong Fu, UCSD
Project: ENCODE
Experiment
ENCSR436OZQ
released
● 1

ChIP-seq of K562
Homo sapiens K562 stably expressing fusion protein
Target: Control
Lab: Michael Snyder, Stanford
Project: ENCODE
Experiment
ENCSR176UHI
released

ChIP-seq of K562
Homo sapiens K562
Target: MNT
Lab: Michael Snyder, Stanford
Project: ENCODE
Experiment
ENCSR390VGH
released
● 2

ChIP-seq of K562
Experiment



Create a BDBag from an ENCODE search.

For example enter the following search:

https://www.encodeproject.org/search/?type=Experiment&assay_title=RNA-seq&replicates.library.biosample.biosample_type=stem+cell

Or paste in an Encode metadata file.

Scientific Digital Asset Management

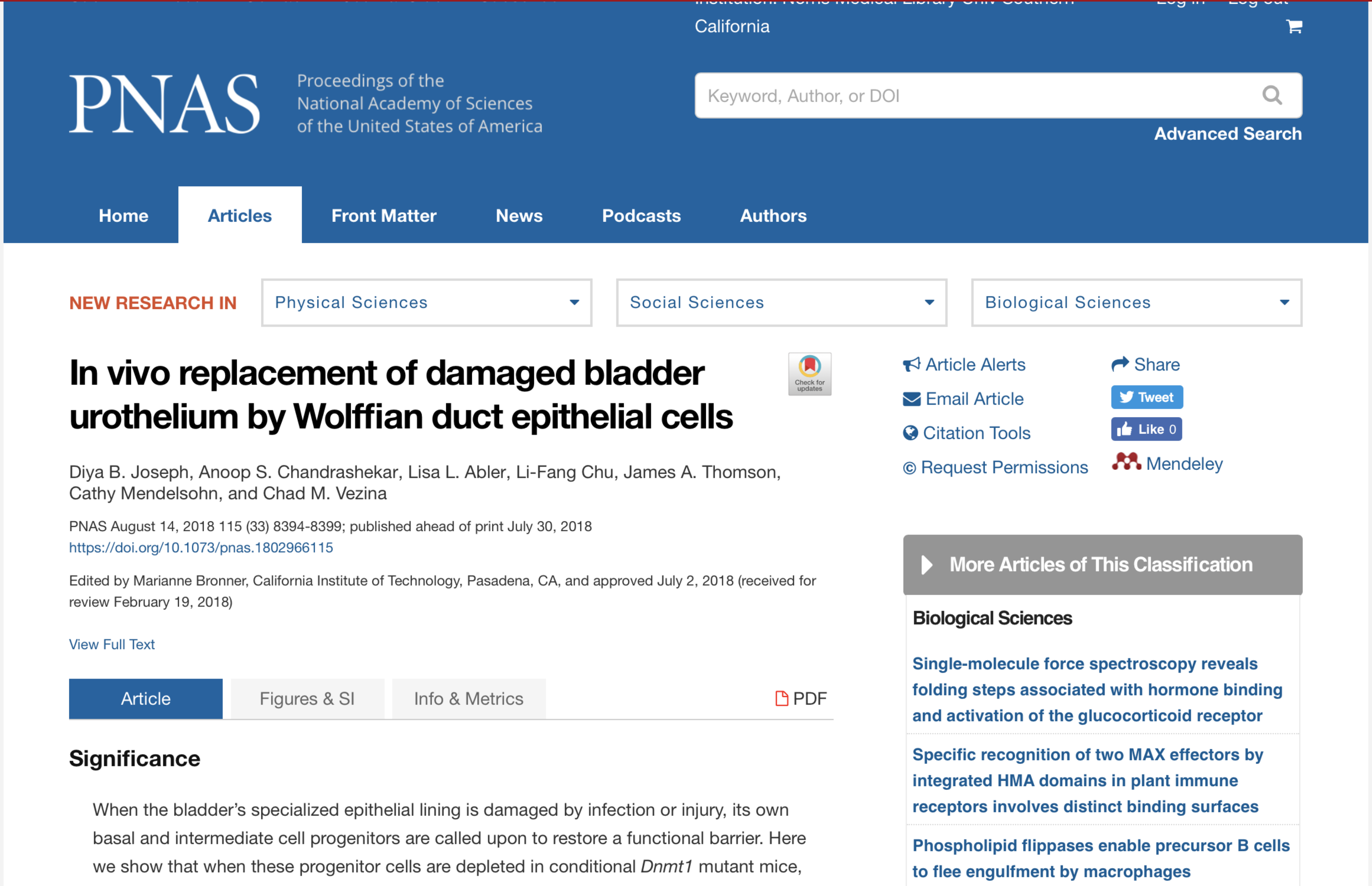
Discovery Environment for Relational Information and Versioned Assets (DERIVA)

- Discovery as process of creating and updating contextualized digital assets.
- Adaptive and extensible

Scientific Asset Management System

Discovery Environment for Relational Information and Versioned Assets (DERIVA)

- DERIVA promotes FAIR data production by:
  - **F**: providing rich metadata using an Entity-Relationship model to express relationships between diverse data elements;
  - **A**: offering rich access control and access to metadata via standard HTTP web service interfaces;
  - **I**: integrating with standardized terms defined by collaborators, consortium or communities; and
  - **R**: supporting dynamic model evolution so that the data presented accurately represents the current structure and state of knowledge within an investigation.



PNAS Proceedings of the National Academy of Sciences of the United States of America

In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

Diya B. Joseph, Anoop S. Chandrashekar, Lisa L. Abler, Li-Fang Chu, James A. Thomson, Cathy Mendelsohn, and Chad M. Vezina

PNAS August 14, 2018 115 (33) 8394-8399; published ahead of print July 30, 2018
https://doi.org/10.1073/pnas.1802966115

Edited by Marianne Bronner, California Institute of Technology, Pasadena, CA, and approved July 2, 2018 (received for review February 19, 2018)

Biological Sciences

Single-molecule force spectroscopy reveals folding steps associated with hormone binding and activation of the glucocorticoid receptor

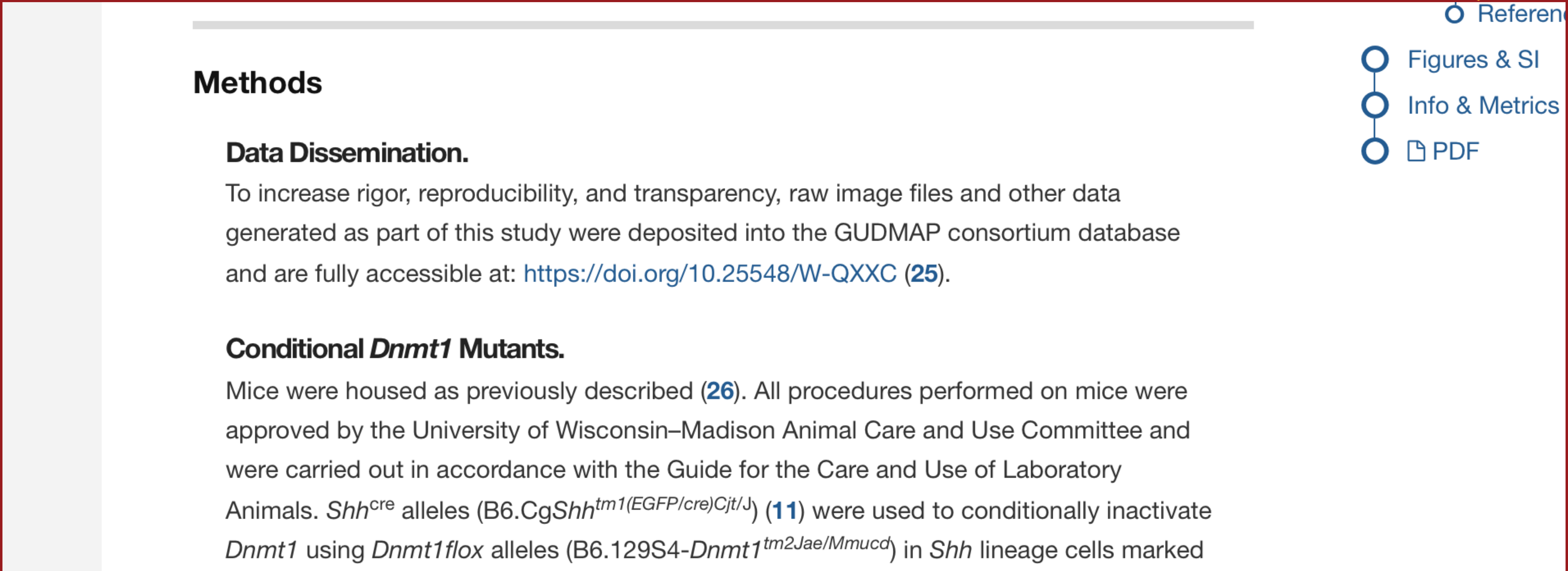
Specific recognition of two MAX effectors by integrated HMA domains in plant immune receptors involves distinct binding surfaces

Phospholipid flippases enable precursor B cells to flee engulfment by macrophages

Significance

When the bladder's specialized epithelial lining is damaged by infection or injury, its own basal and intermediate cell progenitors are called upon to restore a functional barrier. Here we show that when these progenitor cells are depleted in conditional *Dnmt1* mutant mice,

tute



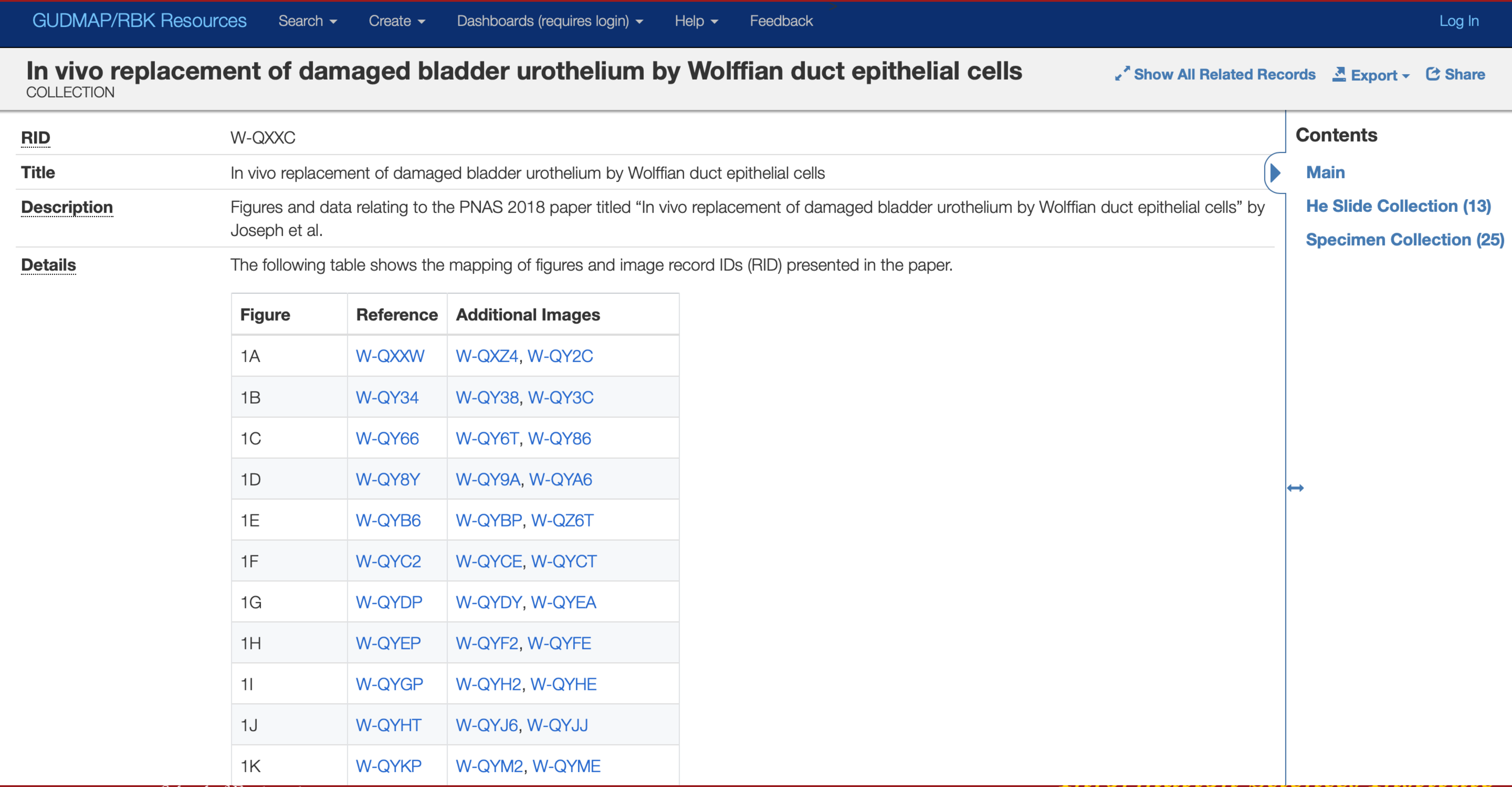
Methods

Data Dissemination.

To increase rigor, reproducibility, and transparency, raw image files and other data generated as part of this study were deposited into the GUDMAP consortium database and are fully accessible at: https://doi.org/10.25548/W-QXXC (**25**).

Conditional *Dnmt1* Mutants.

Mice were housed as previously described (**26**). All procedures performed on mice were approved by the University of Wisconsin–Madison Animal Care and Use Committee and were carried out in accordance with the Guide for the Care and Use of Laboratory Animals. *Shh*^{cre} alleles (B6.Cg*Shh*$^{tm1(EGFP/cre)Cjt/J}$) (**11**) were used to conditionally inactivate *Dnmt1* using *Dnmt1flox* alleles (B6.129S4-*Dnmt1*$^{tm2Jae/Mmucd}$) in *Shh* lineage cells marked



In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

COLLECTION

RID W-QXXC

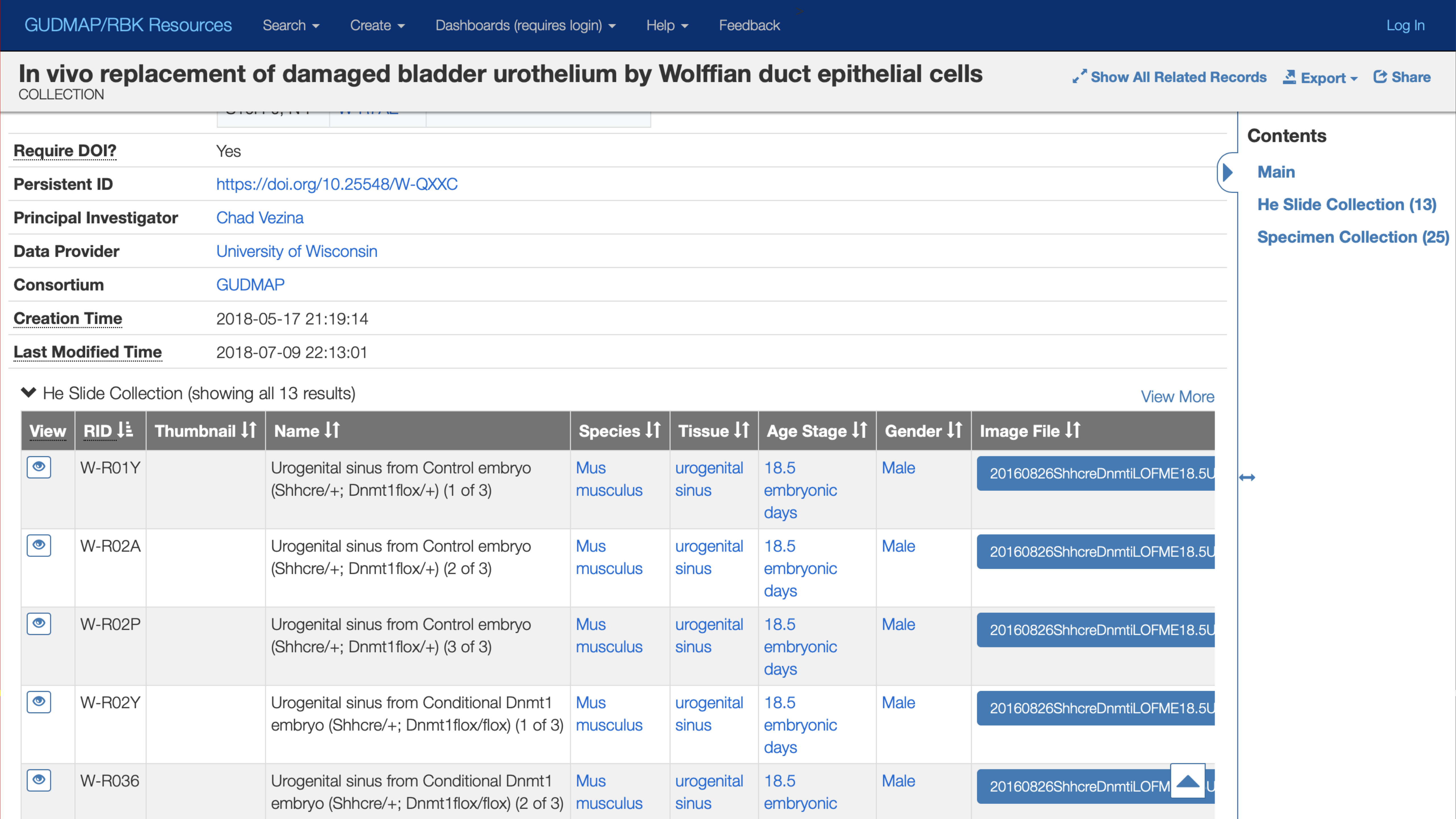
Title In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

Description Figures and data relating to the PNAS 2018 paper titled “In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells” by Joseph et al.

Details The following table shows the mapping of figures and image record IDs (RID) presented in the paper.

Figure	Reference	Additional Images
1A	W-QXXW	W-QXZ4, W-QY2C
1B	W-QY34	W-QY38, W-QY3C
1C	W-QY66	W-QY6T, W-QY86
1D	W-QY8Y	W-QY9A, W-QYA6
1E	W-QYB6	W-QYBP, W-QZ6T
1F	W-QYC2	W-QYCE, W-QYCT
1G	W-QYDP	W-QYDY, W-QYEA
1H	W-QYEP	W-QYF2, W-QYFE
1I	W-QYGP	W-QYH2, W-QYHE
1J	W-QYHT	W-QYJ6, W-QYJJ
1K	W-QYKP	W-QYM2, W-QYME

Contents



In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

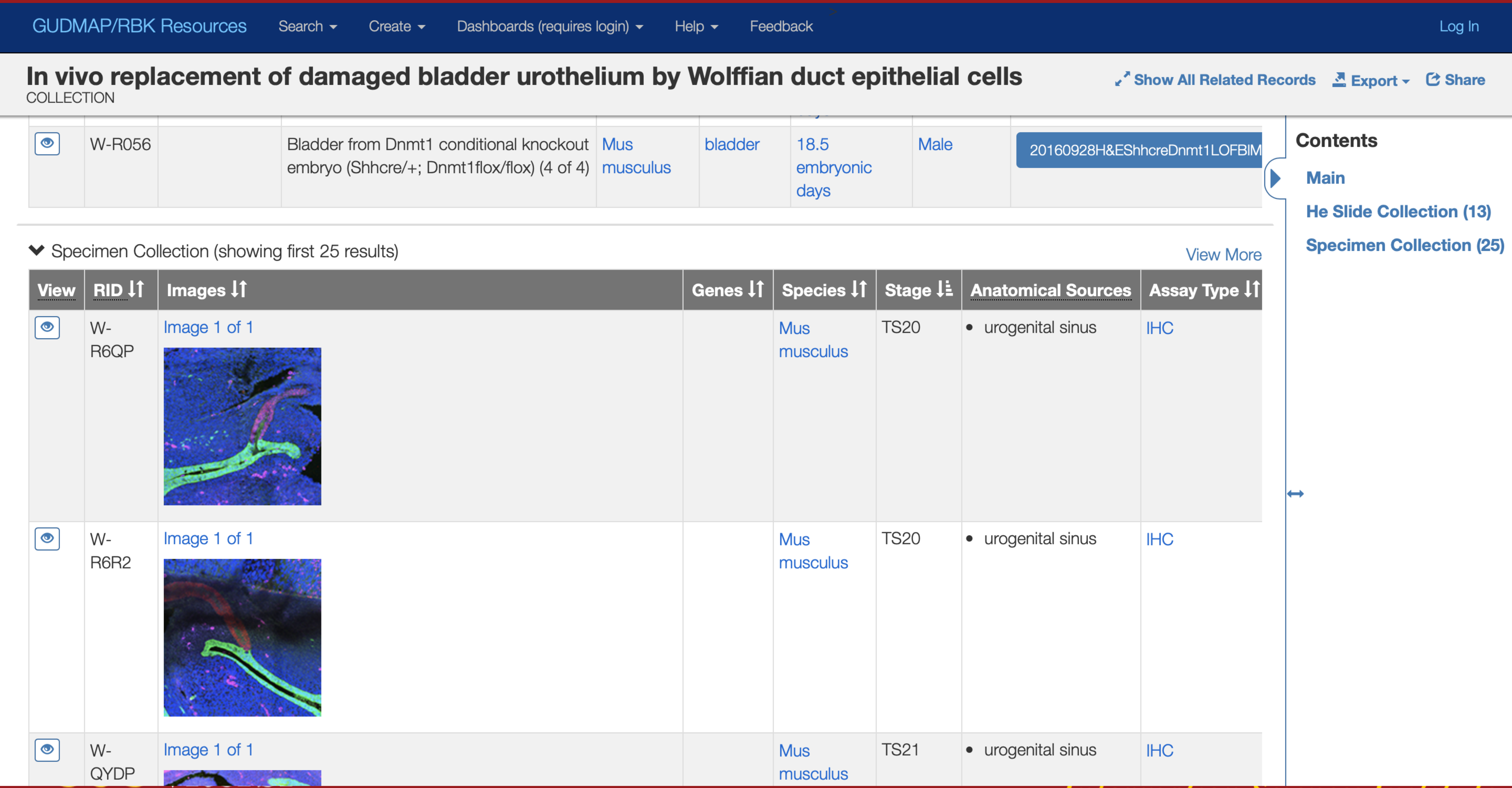
COLLECTION

Require DOI?	Yes
Persistent ID	https://doi.org/10.25548/W-QXXC
Principal Investigator	Chad Vezina
Data Provider	University of Wisconsin
Consortium	GUDMAP
Creation Time	2018-05-17 21:19:14
Last Modified Time	2018-07-09 22:13:01

He Slide Collection (showing all 13 results)

View	RID	Thumbnail	Name	Species	Tissue	Age Stage	Gender	Image File
👁	W-R01Y		Urogenital sinus from Control embryo (Shhcre/+; Dnmt1flox/+) (1 of 3)	Mus musculus	urogenital sinus	18.5 embryonic days	Male	20160826ShhcreDnmtiLOFME18.5U
👁	W-R02A		Urogenital sinus from Control embryo (Shhcre/+; Dnmt1flox/+) (2 of 3)	Mus musculus	urogenital sinus	18.5 embryonic days	Male	20160826ShhcreDnmtiLOFME18.5U
👁	W-R02P		Urogenital sinus from Control embryo (Shhcre/+; Dnmt1flox/+) (3 of 3)	Mus musculus	urogenital sinus	18.5 embryonic days	Male	20160826ShhcreDnmtiLOFME18.5U
👁	W-R02Y		Urogenital sinus from Conditional Dnmt1 embryo (Shhcre/+; Dnmt1flox/flox) (1 of 3)	Mus musculus	urogenital sinus	18.5 embryonic days	Male	20160826ShhcreDnmtiLOFME18.5U
👁	W-R036		Urogenital sinus from Conditional Dnmt1 embryo (Shhcre/+; Dnmt1flox/flox) (2 of 3)	Mus musculus	urogenital sinus	18.5 embryonic	Male	20160826ShhcreDnmtiLOFM U

Contents



GUDMAP/RBK Resources   Search ▾   Create ▾   Dashboards (requires login) ▾   Help ▾   Feedback   Log In

In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

COLLECTION

Show All Related Records   Export ▾   Share

👁	W-R056		Bladder from Dnmt1 conditional knockout embryo (Shhcre/+; Dnmt1flox/flox) (4 of 4)	Mus musculus	bladder	18.5 embryonic days	Male	20160928H&EShhcreDnmt1LOFBIM

Specimen Collection (showing first 25 results)

View More

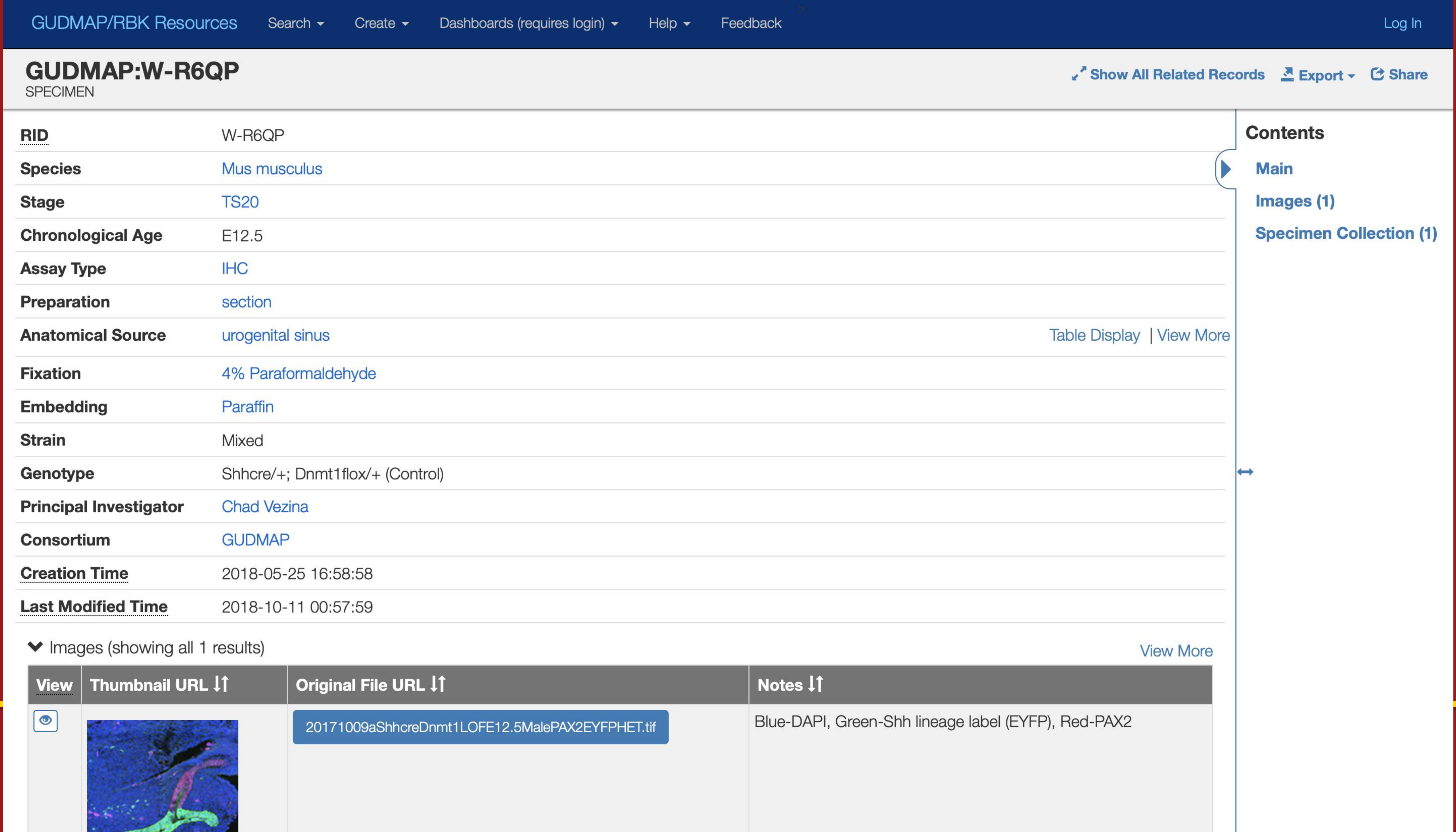
View	RID	Images	Genes	Species	Stage	Anatomical Sources	Assay Type
👁	W-R6QP	Image 1 of 1		Mus musculus	TS20	• urogenital sinus	IHC
👁	W-R6R2	Image 1 of 1		Mus musculus	TS20	• urogenital sinus	IHC
👁	W-QYDP	Image 1 of 1		Mus musculus	TS21	• urogenital sinus	IHC

Contents

Main

He Slide Collection (13)

Specimen Collection (25)



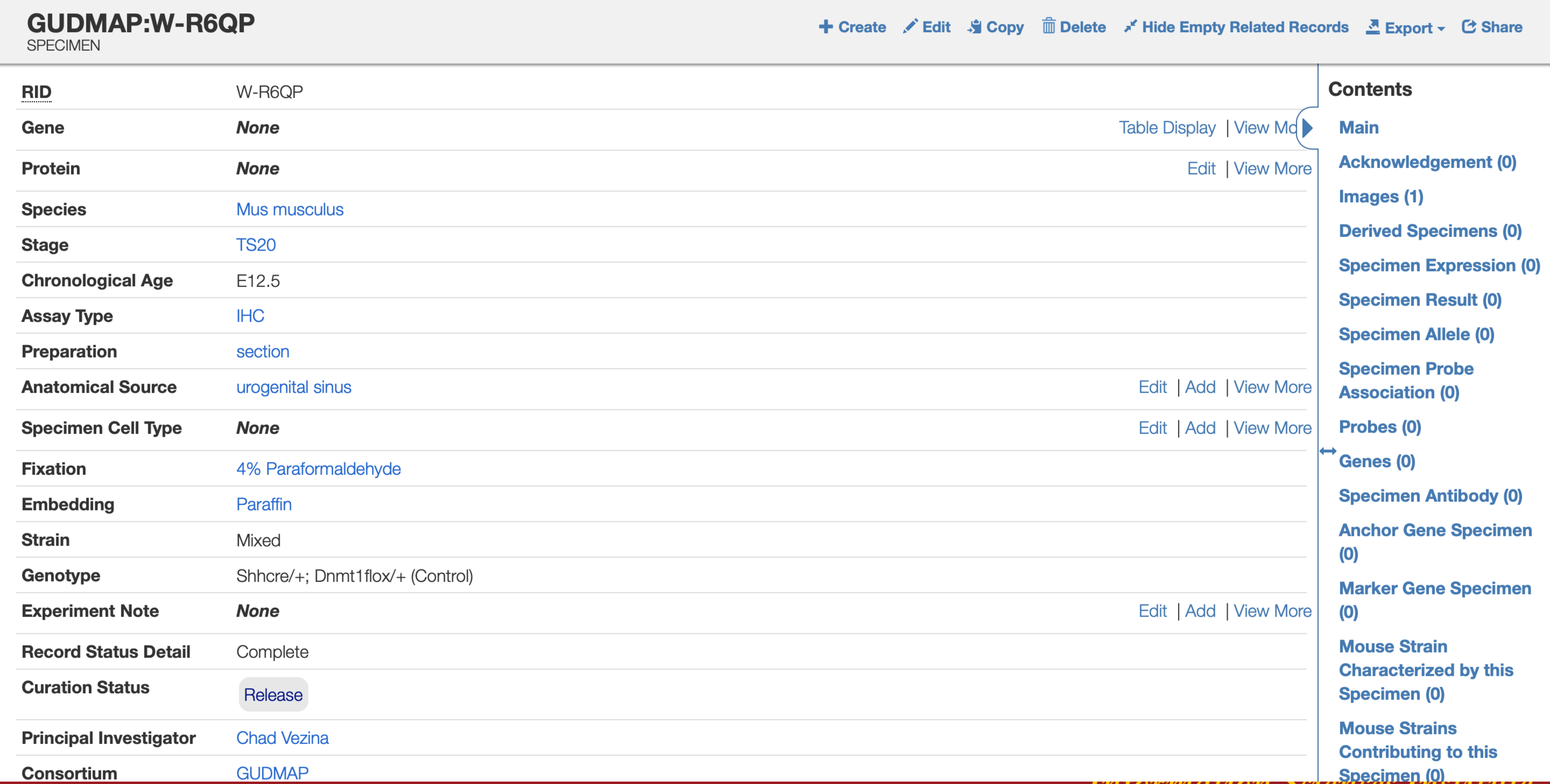
GUDMAP:W-R6QP

SPECIMEN

RID	W-R6QP
Species	Mus musculus
Stage	TS20
Chronological Age	E12.5
Assay Type	IHC
Preparation	section
Anatomical Source	urogenital sinus
Fixation	4% Paraformaldehyde
Embedding	Paraffin
Strain	Mixed
Genotype	Shhcre/+; Dnmt1flox/+ (Control)
Principal Investigator	Chad Vezina
Consortium	GUDMAP
Creation Time	2018-05-25 16:58:58
Last Modified Time	2018-10-11 00:57:59

Images (showing all 1 results)

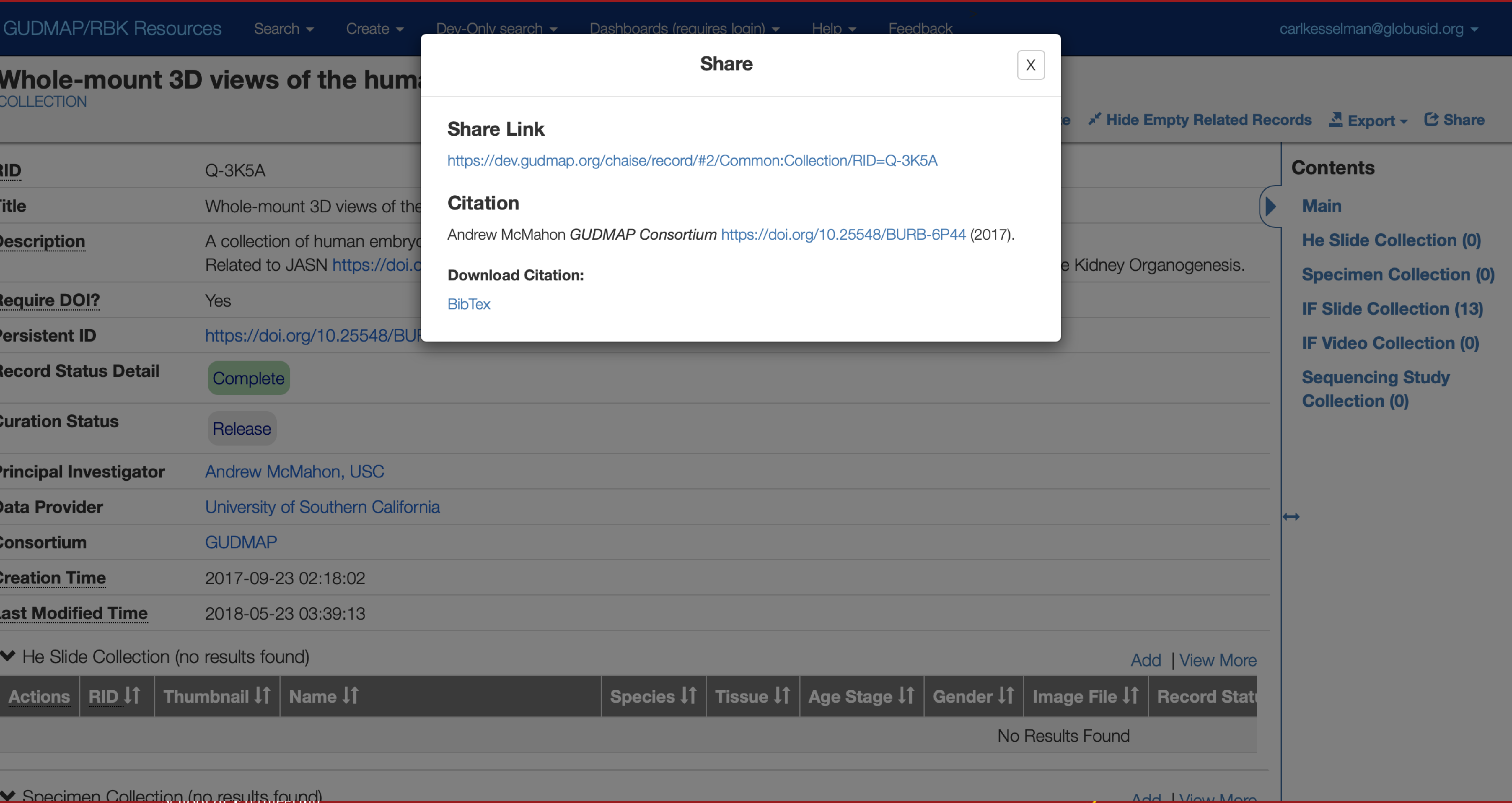
View	Thumbnail URL	Original File URL	Notes
👁		20171009aShhcreDnmt1LOFE12.5MalePAX2EYFPHET.tif	Blue-DAPI, Green-Shh lineage label (EYFP), Red-PAX2



GUDMAP:W-R6QP

SPECIMEN

RID	W-R6QP	
Gene	*None*	Table Display | View Mo
Protein	*None*	Edit | View More
Species	Mus musculus	
Stage	TS20	
Chronological Age	E12.5	
Assay Type	IHC	
Preparation	section	
Anatomical Source	urogenital sinus	Edit | Add | View More
Specimen Cell Type	*None*	Edit | Add | View More
Fixation	4% Paraformaldehyde	
Embedding	Paraffin	
Strain	Mixed	
Genotype	Shhcre/+; Dnmt1flox/+ (Control)	
Experiment Note	*None*	Edit | Add | View More
Record Status Detail	Complete	
Curation Status	Release	
Principal Investigator	Chad Vezina	
Consortium	GUDMAP	



GUDMAP/RBK Resources Search ▾ Create ▾ Dev-Only search ▾ Dashboards (requires login) ▾ Help ▾ Feedback carlkesselman@globusid.org ▾

Whole-mount 3D views of the huma

COLLECTION

e Hide Empty Related Records Export ▾ Share

Share

X

Share Link

https://dev.gudmap.org/chaise/record/#2/Common:Collection/RID=Q-3K5A

Citation

Andrew McMahon *GUDMAP Consortium* https://doi.org/10.25548/BURB-6P44 (2017).

Download Citation:

BibTex

RID	Q-3K5A
itle	Whole-mount 3D views of the
escription	A collection of human embryo Related to JASN https://doi.c … e Kidney Organogenesis.
equire DOI?	Yes
ersistent ID	https://doi.org/10.25548/BUF
ecord Status Detail	Complete
uration Status	Release
rincipal Investigator	Andrew McMahon, USC
ata Provider	University of Southern California
onsortium	GUDMAP
reation Time	2017-09-23 02:18:02
ast Modified Time	2018-05-23 03:39:13

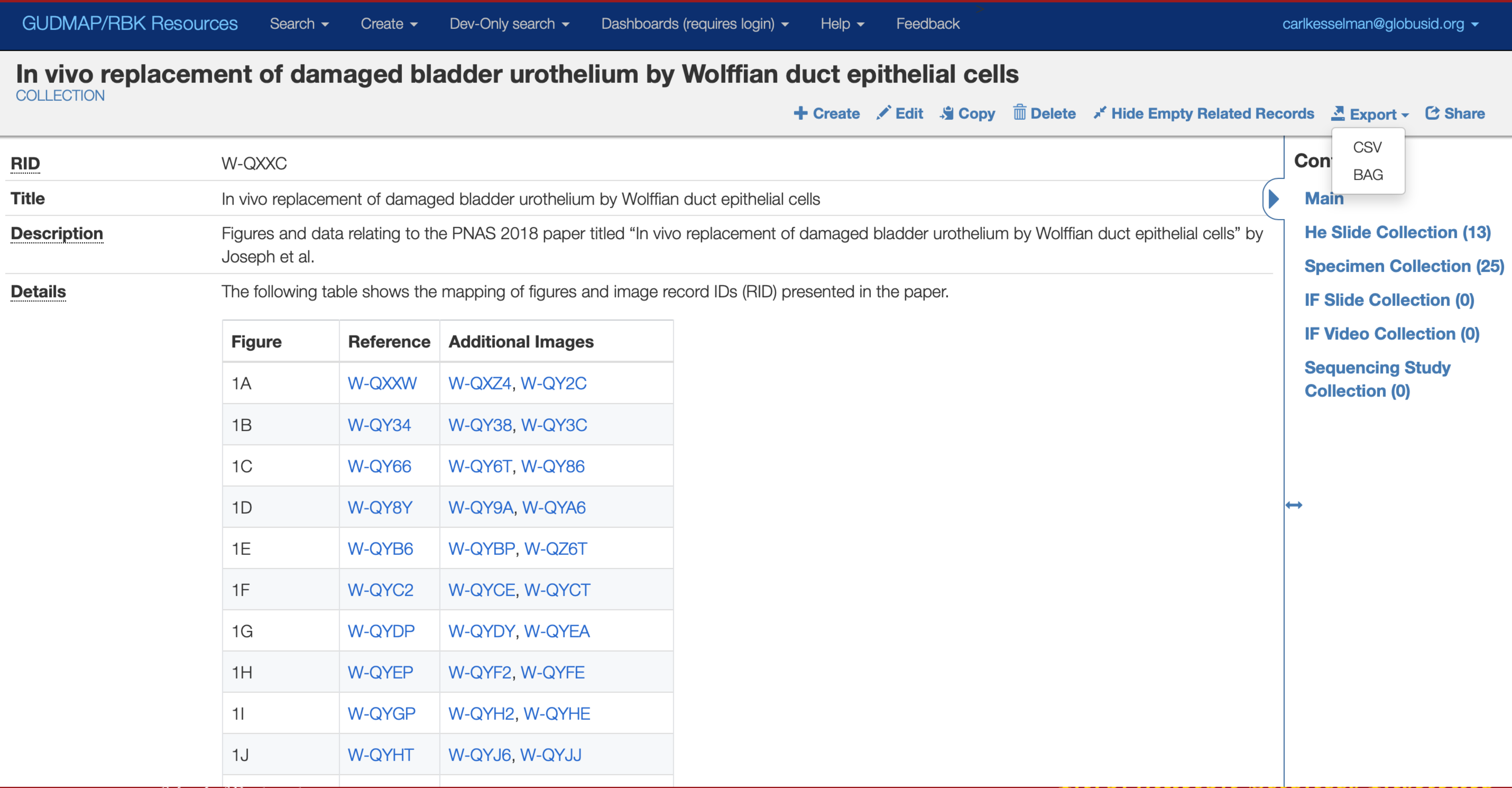
He Slide Collection (no results found) Add | View More

Actions	RID	Thumbnail	Name	Species	Tissue	Age Stage	Gender	Image File	Record Stat
No Results Found									

Specimen Collection (no results found) Add | View More

Contents

- Main
- He Slide Collection (0)
- Specimen Collection (0)
- IF Slide Collection (13)
- IF Video Collection (0)
- Sequencing Study Collection (0)



In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

COLLECTION

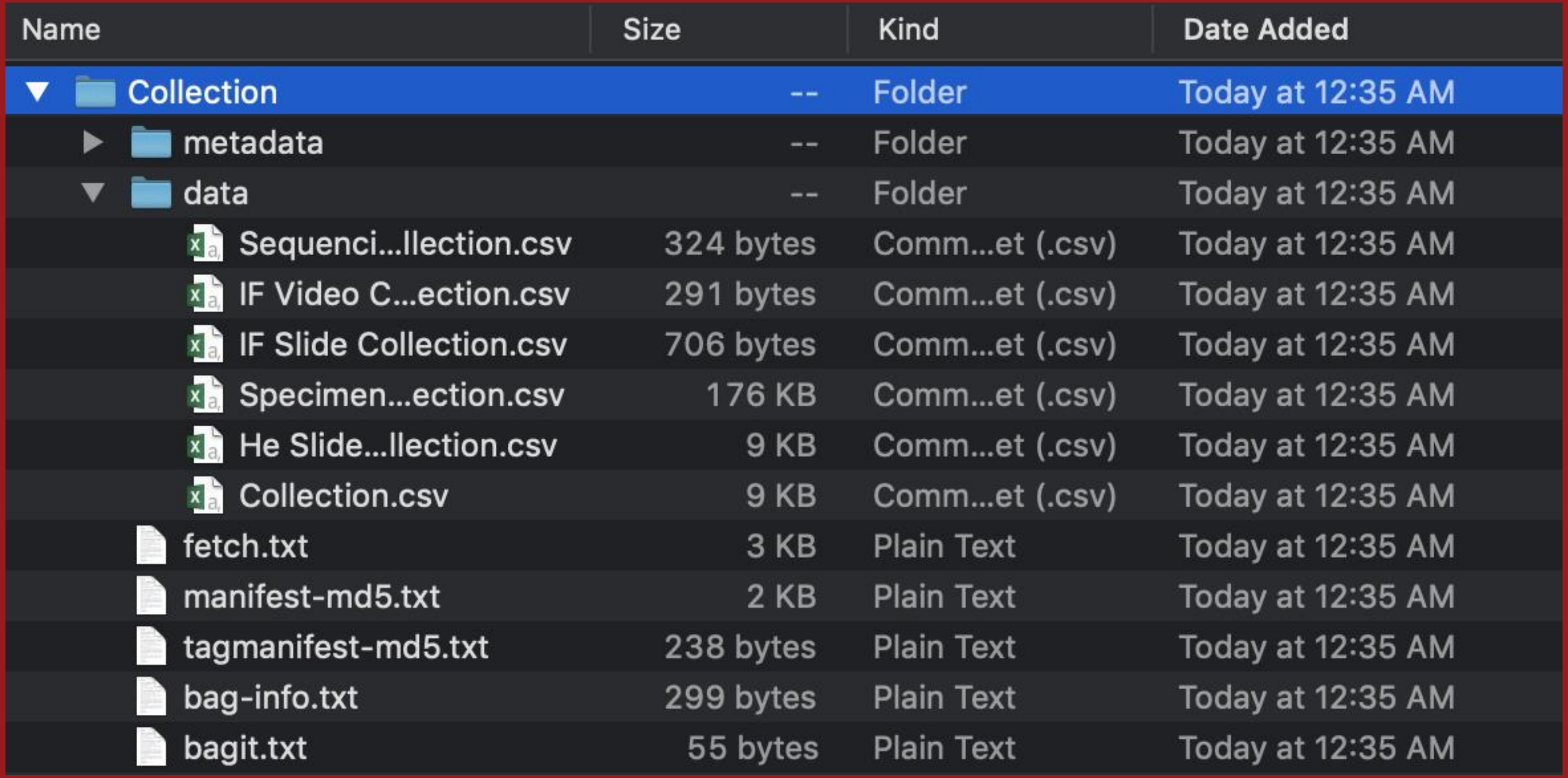
RID W-QXXC

Title In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

Description Figures and data relating to the PNAS 2018 paper titled “In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells” by Joseph et al.

Details The following table shows the mapping of figures and image record IDs (RID) presented in the paper.

Figure	Reference	Additional Images
1A	W-QXXW	W-QXZ4, W-QY2C
1B	W-QY34	W-QY38, W-QY3C
1C	W-QY66	W-QY6T, W-QY86
1D	W-QY8Y	W-QY9A, W-QYA6
1E	W-QYB6	W-QYBP, W-QZ6T
1F	W-QYC2	W-QYCE, W-QYCT
1G	W-QYDP	W-QYDY, W-QYEA
1H	W-QYEP	W-QYF2, W-QYFE
1I	W-QYGP	W-QYH2, W-QYHE
1J	W-QYHT	W-QYJ6, W-QYJJ



Name	Size	Kind	Date Added
▼ Collection	--	Folder	Today at 12:35 AM
▶ metadata	--	Folder	Today at 12:35 AM
▼ data	--	Folder	Today at 12:35 AM
Sequenci...llection.csv	324 bytes	Comm...et (.csv)	Today at 12:35 AM
IF Video C...ection.csv	291 bytes	Comm...et (.csv)	Today at 12:35 AM
IF Slide Collection.csv	706 bytes	Comm...et (.csv)	Today at 12:35 AM
Specimen...ection.csv	176 KB	Comm...et (.csv)	Today at 12:35 AM
He Slide...llection.csv	9 KB	Comm...et (.csv)	Today at 12:35 AM
Collection.csv	9 KB	Comm...et (.csv)	Today at 12:35 AM
fetch.txt	3 KB	Plain Text	Today at 12:35 AM
manifest-md5.txt	2 KB	Plain Text	Today at 12:35 AM
tagmanifest-md5.txt	238 bytes	Plain Text	Today at 12:35 AM
bag-info.txt	299 bytes	Plain Text	Today at 12:35 AM
bagit.txt	55 bytes	Plain Text	Today at 12:35 AM

Summary

- derivacloud.org
- https://fair-research.org

- Carl@isi.edu