KBase

PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

Adam Arkin, KBase PI

2023 Spring BERAC

COMMUNITY-DRIVEN

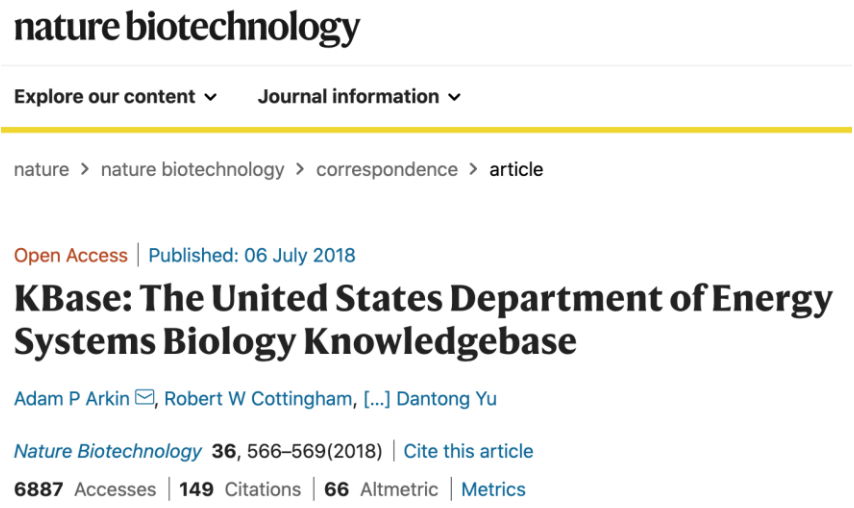
PREDICTIVE BIOLOGY

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

What is KBase?

KBase enables users to analyze, share, and collaborate using data and tools designed to help build increasingly realistic models for biological function.



nature biotechnology

Explore our content ⌄ Journal information ⌄

nature › nature biotechnology › correspondence › article

Open Access | Published: 06 July 2018

KBase: The United States Department of Energy Systems Biology Knowledgebase

Adam P Arkin ✉, Robert W Cottingham, [...] Dantong Yu

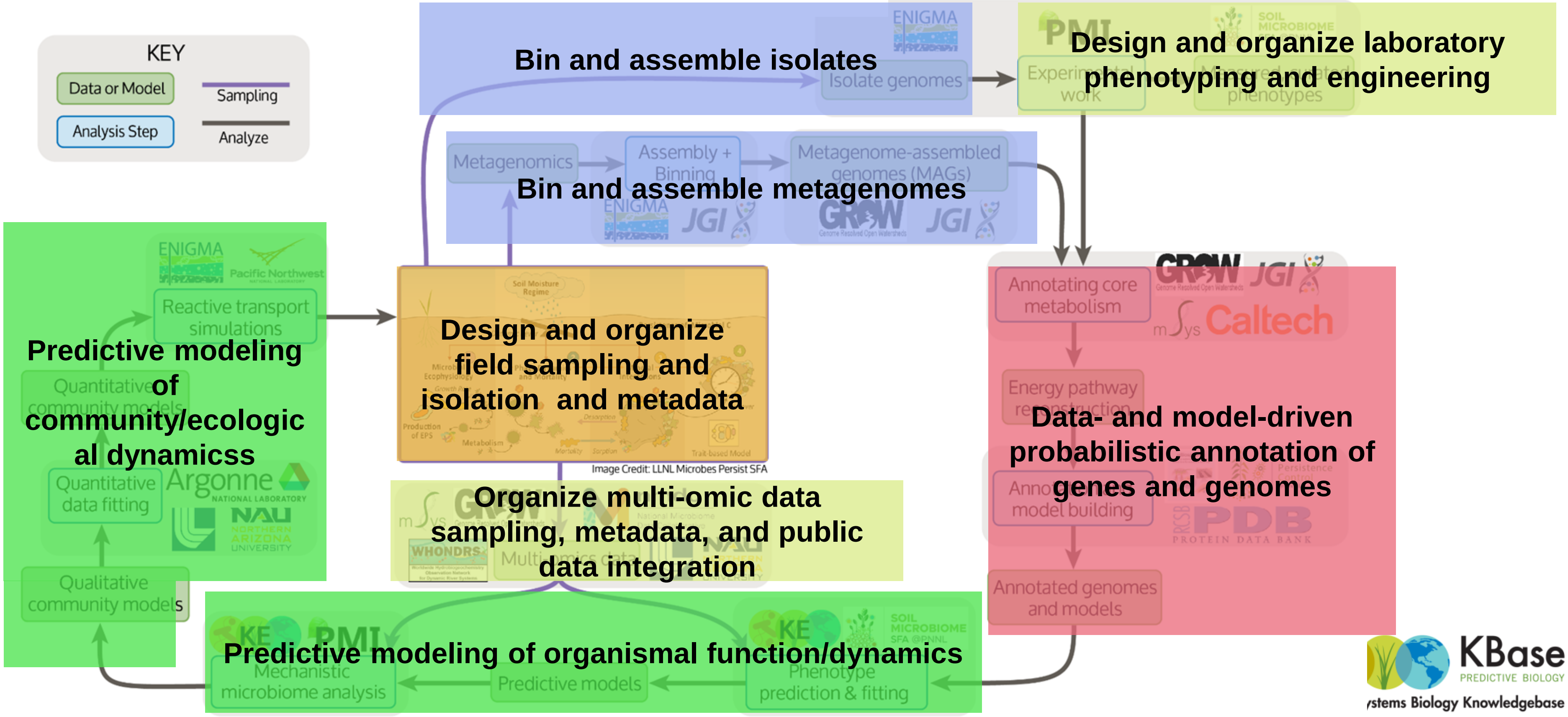
Nature Biotechnology **36**, 566–569(2018) | Cite this article

6887 Accesses | **149** Citations | **66** Altmetric | Metrics

U.S. DEPARTMENT OF ENERGY
Office of Science

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

A Virtual Cycle for Genotype->Phenotype Prediction from Genes to Ecosystems



KEY

Data or Model	Sampling
Analysis Step	Analyze

Bin and assemble isolates

Design and organize laboratory phenotyping and engineering

Bin and assemble metagenomes

Predictive modeling of community/ecological dynamicss

Design and organize field sampling and isolation and metadata

Image Credit: LLNL Microbes Persist SFA

Data- and model-driven probabilistic annotation of genes and genomes

Organize multi-omic data sampling, metadata, and public data integration

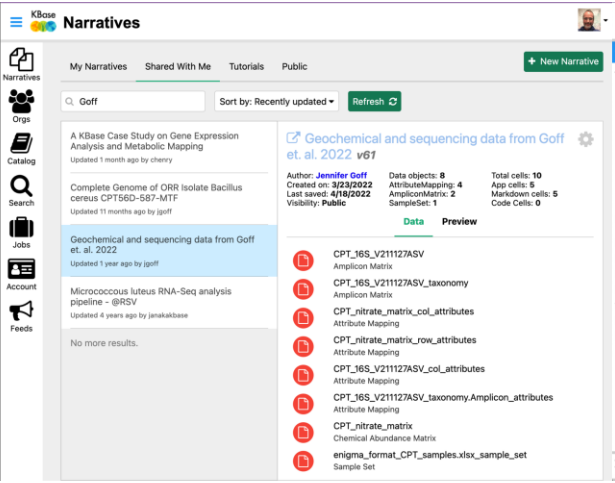
Predictive modeling of organismal function/dynamics

Isolate genomes
Experimental work
Measured/related phenotypes
Metagenomics
Assembly + Binning
Metagenome-assembled genomes (MAGs)
Annotating core metabolism
Energy pathway reconstruction
Annotation & model building
Annotated genomes and models
Reactive transport simulations
Quantitative community models
Quantitative data fitting
Qualitative community models
Multi-omics data
Mechanistic microbiome analysis
Predictive models
Phenotype prediction & fitting
Soil Moisture Regime
Production of EPS
Metabolism
Trait-based Model

KBase
PREDICTIVE BIOLOGY
ystems Biology Knowledgebase

KBase ensures all data & analyses are FAIR *and* credited

Team member shares an analysis



A KBase Case Study on Gene Expression Analysis and Metabolic Mapping
Updated 1 month ago by chenry

Complete Genome of ORR Isolate Bacillus cereus CPT56D-587-MTF
Updated 11 months ago by jgoff

Geochemical and sequencing data from Goff et. al. 2022
Updated 1 year ago by jgoff

Micrococcous luteus RNA-Seq analysis pipeline - @RSV
Updated 4 years ago by janakakbase

No more results.

Geochemical and sequencing data from Goff et. al. 2022 *v61*

Author: Jennifer Goff	Data objects: 8	Total cells: 10
Created on: 3/23/2022	AttributeMapping: 4	App cells: 5
Last saved: 4/18/2022	AmpliconMatrix: 2	Markdown cells: 5
Visibility: Public	SampleSet: 1	Code Cells: 0

CPT_16S_V211127ASV
Amplicon Matrix

CPT_16S_V211127ASV_taxonomy
Amplicon Matrix

CPT_nitrate_matrix_col_attributes
Attribute Mapping

CPT_nitrate_matrix_row_attributes
Attribute Mapping

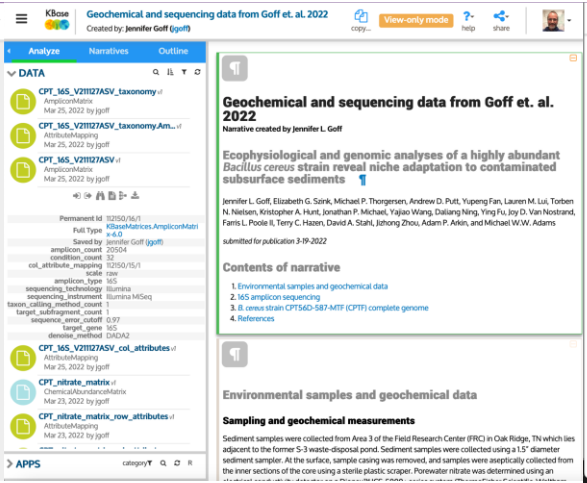
CPT_16S_V211127ASV_col_attributes
Attribute Mapping

CPT_16S_V211127ASV_taxonomy.Amplicon_attributes
Attribute Mapping

CPT_nitrate_matrix
Chemical Abundance Matrix

enigma_format_CPT_samples.xlsx_sample_set
Sample Set

'Narrative' of analysis has everything



Geochemical and sequencing data from Goff et. al. 2022
Created by: Jennifer Goff (jgoff)

CPT_16S_V211127ASV_taxonomy v1
AmpliconMatrix
Mar 25, 2022 by jgoff

CPT_16S_V211127ASV_taxonomy.Am... v1
AttributeMapping
Mar 25, 2022 by jgoff

CPT_16S_V211127ASV v1
AmpliconMatrix
Mar 25, 2022 by jgoff

Permanent Id	112150/16/1
Full Type	KBaseMatrices.AmpliconMatrix-6.0
Saved by	Jennifer Goff (jgoff)
amplicon_count	20504
condition_count	32
col_attribute_mapping	112150/15/1
scale	raw
amplicon_type	16S
sequencing_technology	Illumina
sequencing_instrument	Illumina MiSeq
taxon_calling_method_count	1
target_subfragment_count	1
sequence_error_cutoff	0.97
target_gene	16S
denoise_method	DADA2

CPT_16S_V211127ASV_col_attributes v1
AttributeMapping
Mar 25, 2022 by jgoff

CPT_nitrate_matrix v1
ChemicalAbundanceMatrix
Mar 23, 2022 by jgoff

CPT_nitrate_matrix_row_attributes v1
AttributeMapping
Mar 23, 2022 by jgoff

Geochemical and sequencing data from Goff et. al. 2022
Narrative created by Jennifer L. Goff

Ecophysiological and genomic analyses of a highly abundant *Bacillus cereus* strain reveal niche adaptation to contaminated subsurface sediments

Jennifer L. Goff, Elizabeth G. Szink, Michael P. Thorgersen, Andrew D. Putt, Yupeng Fan, Lauren M. Lui, Torben N. Nielsen, Kristopher A. Hunt, Jonathan P. Michael, Yajiao Wang, Daliang Ning, Ying Fu, Joy D. Van Nostrand, Farris L. Poole II, Terry C. Hazen, David A. Stahl, Jizhong Zhou, Adam P. Arkin, and Michael W.W. Adams

submitted for publication 3-19-2022

Contents of narrative

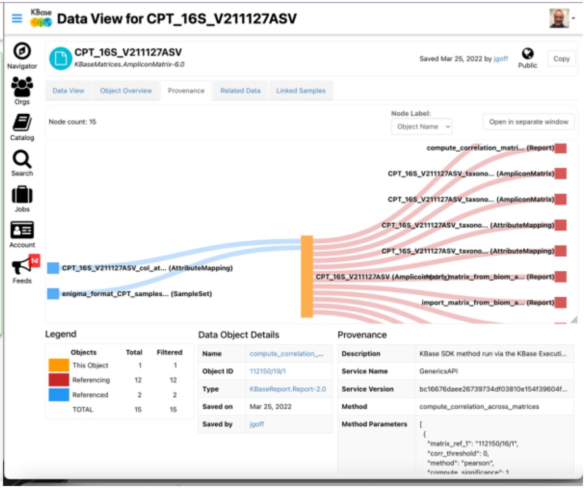
1. Environmental samples and geochemical data
2. 16S amplicon sequencing
3. *B. cereus* strain CPT56D-587-MTF (CPTF) complete genome
4. References

Environmental samples and geochemical data

Sampling and geochemical measurements

Sediment samples were collected from Area 3 of the Field Research Center (FRC) in Oak Ridge, TN which lies adjacent to the former S-3 waste-disposal pond. Sediment samples were collected using a 1.5" diameter sediment sampler. At the surface, sample casing was removed, and samples were aseptically collected from the inner sections of the core using a sterile plastic scraper. Porewater nitrate was determined using an

Every object fully 'provenanced'



KBase
Navigator
Orgs
Catalog
Search
Jobs
Account
Feeds
14
Public
Copy
Data View
Object Overview
Provenance
Related Data
Linked Samples
Node Label:
Object Name
Open in separate window
compute_correlation_matri... (Report)
CPT_16S_V211127ASV_taxono... (AmpliconMatrix)
CPT_16S_V211127ASV_taxono... (AmpliconMatrix)
CPT_16S_V211127ASV_taxono... (AttributeMapping)
CPT_16S_V211127ASV_taxono... (AttributeMapping)
CPT_16S_V211127ASV_col_at... (AttributeMapping)
enigma_format_CPT_samples... (SampleSet)
CPT_16S_V211127ASV (AmpliconMatrix)
import_matrix_from_biom_a... (Report)
import_matrix_from_biom_a... (Report)

Data View for CPT_16S_V211127ASV

CPT_16S_V211127ASV
KBaseMatrices.AmpliconMatrix-6.0
Saved Mar 25, 2022 by jgoff

Node count: 15

Legend

	Objects	Total	Filtered
🟧	This Object	1	1
🟥	Referencing	12	12
🟦	Referenced	2	2
	TOTAL	15	15

Data Object Details

Name	compute_correlation_...
Object ID	112150/19/1
Type	KBaseReport.Report-2.0
Saved on	Mar 25, 2022
Saved by	jgoff

Provenance

Description	KBase SDK method run via the KBase Executi...
Service Name	GenericsAPI
Service Version	bc16676daee26739734df03810e154f39604f...
Method	compute_correlation_across_matrices
Method Parameters	[{ "matrix_ref_1": "112150/16/1", "corr_threshold": 0, "method": "pearson", "compute_significance": 1

U.S. DEPARTMENT OF ENERGY
Office of Science

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Why do people use KBase?

- Combine different types of data from many resources
- Create unique, complex, reproducible analyses
- Contribute new knowledge back to the research community

ESS-DIVE
Deep Insight for Earth Science Data

NIH

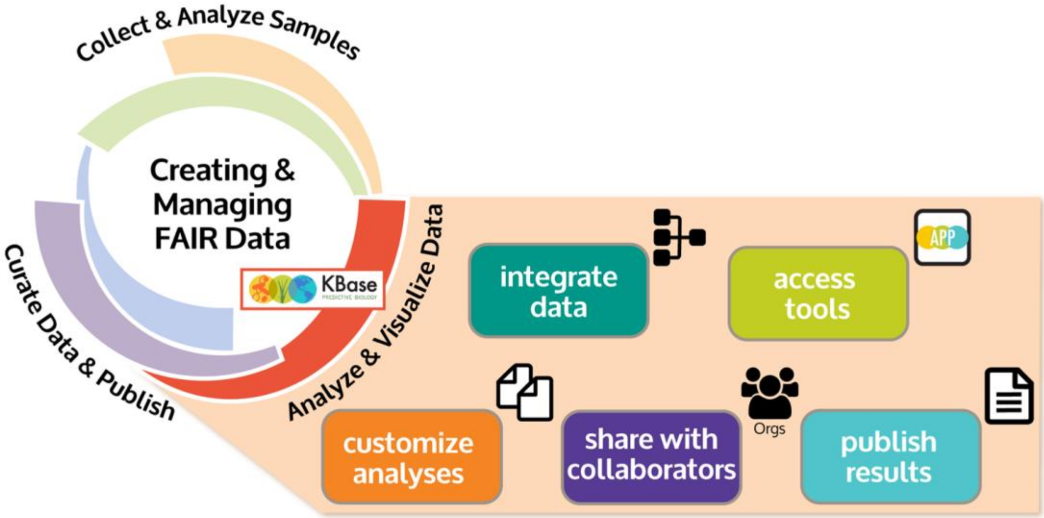
RCSB PDB
PROTEIN DATA BANK

EMSL
Environmental Molecular Sciences Laboratory
A DOE OFFICE OF SCIENCE USER FACILITY

nmdc
National Microbiome Data Collaborative

JGI

Data Portal

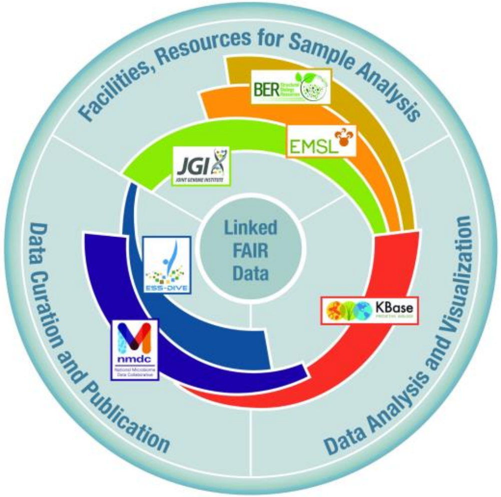


U.S. DEPARTMENT OF ENERGY
Office of Science

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

But it takes a (BER) village to enable FAIR data at scale

- Data (and all research products) must be FAIR across programs
- FAIR data must have provenance across the data life cycle
- BER program can create the infrastructure, but culture change requires *trust*
  - Where the data came from
  - How it was processed
  - Ability to explore quality
  - Ability to see impact and effect of combined power
- People *want* to be FAIR, but it takes support to do well, and they *must get credit for it*
- **Designed KBase to connect everything *and* give credit**


Facilities, Resources for Sample Analysis
Data Curation and Publication
Data Analysis and Visualization
Linked FAIR Data
BER
EMSL
JGI
ESS-DIVE
nmdc
KBase

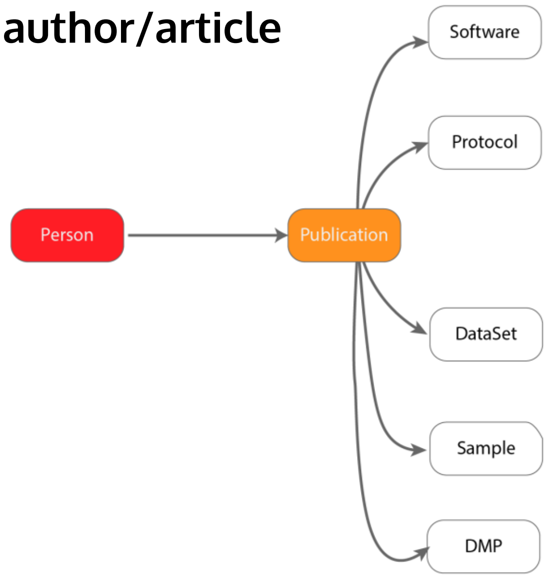
U.S. DOE. 2021. Biological Systems Science Division Strategic Plan, DOE/SC-0205. U.S. Department of Energy Office of Science

U.S. DEPARTMENT OF ENERGY Office of Science

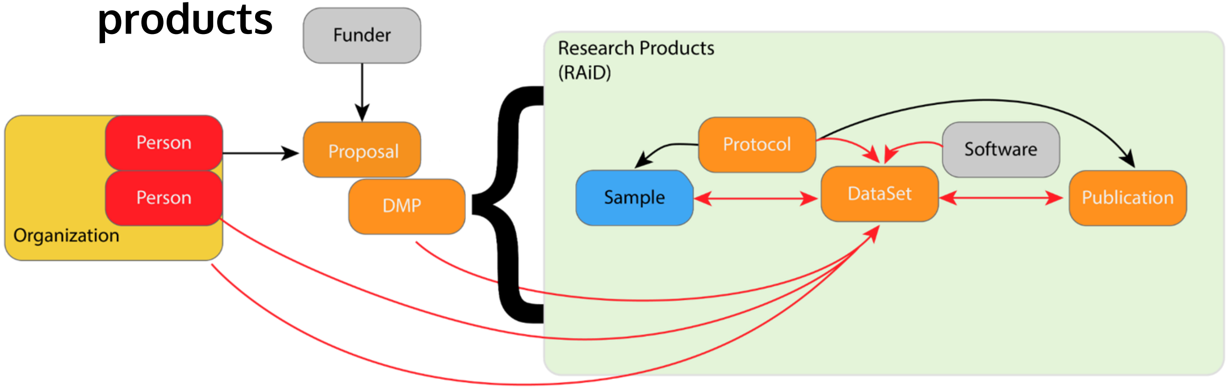
KBase PREDICTIVE BIOLOGY DOE Systems Biology Knowledgebase

Goal: FAIR, trackable research outputs

Now: Focus on author/article



Future: Credit to author/funder for all research products



Red arrows: KBase connections between data sets and other PIDs

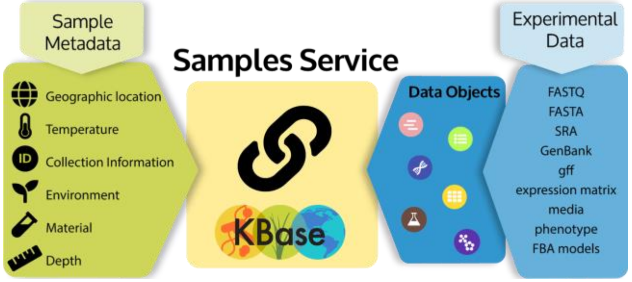
Legend

DOI	IGSN	ROR
RAiD	ORCiD	Standard PID TBD

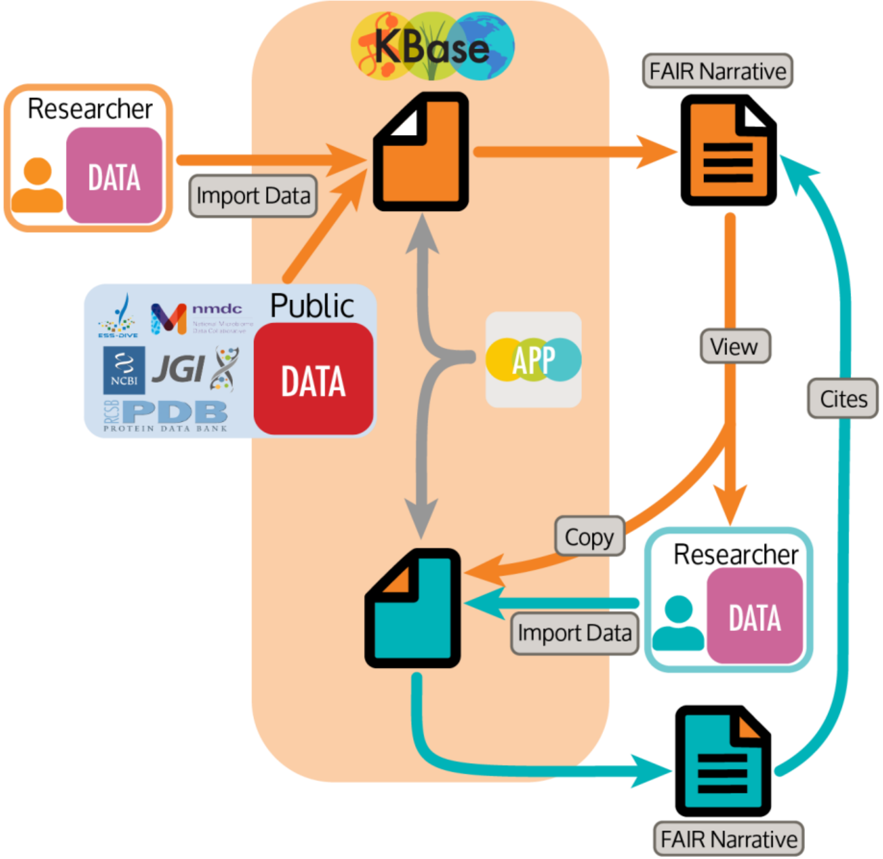
Glossary of Persistent Identifiers (PIDs)

- DOI - digital object identifier
- RAiD - research activity identifier
- IGSN - international geo/general sample number
- ORCiD - open researcher and contributor identifier
- ROR - research organization registry

Sample Sets Organize Complex Data



Tracking impact requires provenance and PIDs



- Provenance has always been integral to KBase's design
- We are not alone! KBase is positioned to effectively leverage external resources to ensure BER research products are FAIR, visible, and trackable.
- FAIR data is best reused when truly understood - knowing what it is, who generated it, and how it was generated.

U.S. DEPARTMENT OF ENERGY Office of Science

KBase PREDICTIVE BIOLOGY DOE Systems Biology Knowledgebase

Moving FAIR data requires alignment between platforms

JGI-KBase Co-development "Data Transfer Service"

Data transfer
- -file
- -type
- -md5

KBase Credit Engine Metadata Fields
- citation (person, title, date, version, etc)
- license
- funding
- related IDs
    - DOIs (proposal, protocols, data, etc)
    - Sample ID (IGSN, GOLD, etc)
    - ORCIDs
    - etc.

OSTI.GOV

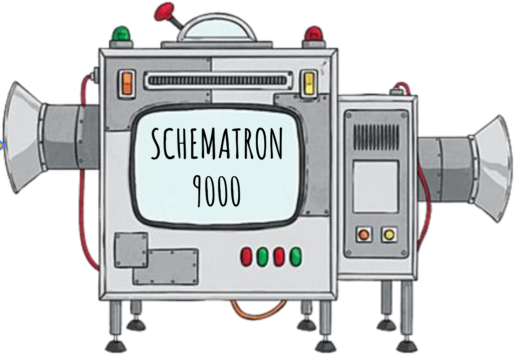
RefSeq:

GROW Genome Resolved Open Watersheds

ESS-DIVE Deep Insight for Earth Science Data

EMSL Environmental Molecular Sciences Laboratory A DOE OFFICE OF SCIENCE USER FACILITY

JGI Data Portal



KBase Data Object

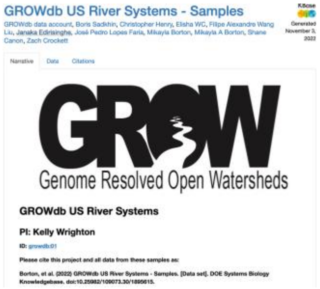
Document of Provenance

This is a genuine KBase object, imported into the narrative at 03:57:19 PST on Thursday October 20th in the year of our lord 2022. Day after day, day after day, We stuck, nor breath nor motion; As idle as a painted ship Upon a painted ocean. Water, water, every where, And all the boards did shrink; Water, water, every where, Nor any drop to drink. The very deep did rot: O Christ! That ever this should be! Yea, slimy things did crawl with legs Upon the slimy sea. About, about, in reel and rout The death-fires danced at night; The water, like a witch's oils, Burnt green, and blue and white. And some in dreams assurèd were Of the Spirit that plagued us so; Nine fathom deep he had followed us From the land of mist and snow. And every tongue, through utter drought, Was withered at the root; We could not speak, no more than if We had been choked with soot. Ah! well a-day! what evil looks Had I from old and young! Instead of the cross, the Albatross About my neck was hung.

Credit Metadata

There passed a weary time. Each throat Was parched, and glazed each eye. A weary time! a weary time! How glazed each weary eye, When looking westward, I beheld A something in the sky. At first it seemed a little speck, And then it seemed a mist; It moved and moved, and took at last A certain shape,

FAIR Narrative



GROWdb US River Systems - Samples

GROWdb data account, Boris Sadkhin, Christopher Henry, Elisha WC, Filipe Alexandre Wang Liu, Janaka Edirisinghe, José Pedro Lopes Faria, Mikayla Borton, Mikayla A Borton, Shane Canon, Zach Crockett

KBase Generated November 3, 2022

GROWdb US River Systems

PI: Kelly Wrighton

ID: growdb:01

Please cite this project and all data from these samples as:

Borton, et al. (2022) GROWdb US River Systems - Samples. [Data set]. DOE Systems Biology Knowledgebase. doi:10.25982/109073.30/1895615.

DataCite

ORCID

Crossref

Google

Index by schema.org

U.S. DEPARTMENT OF ENERGY Office of Science

KBase PREDICTIVE BIOLOGY DOE Systems Biology Knowledgebase

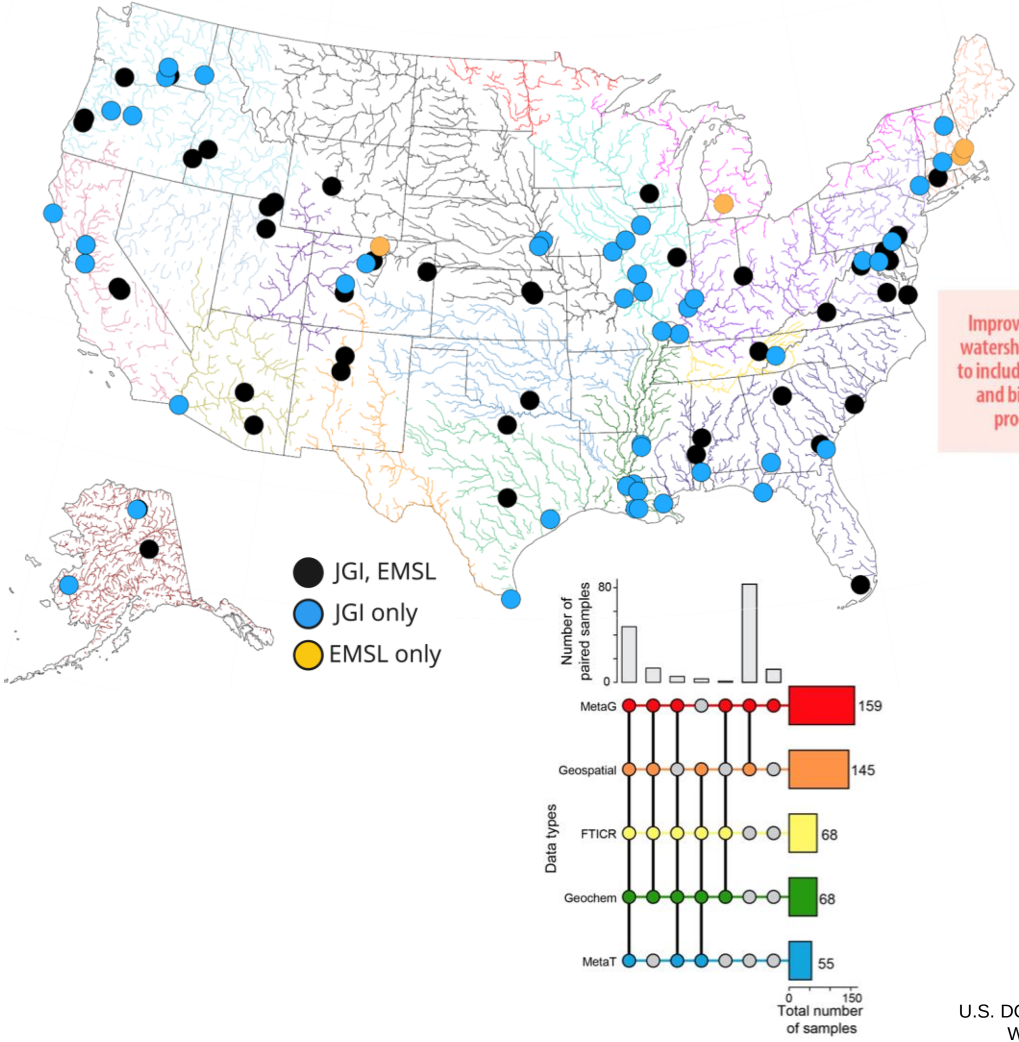
GROW

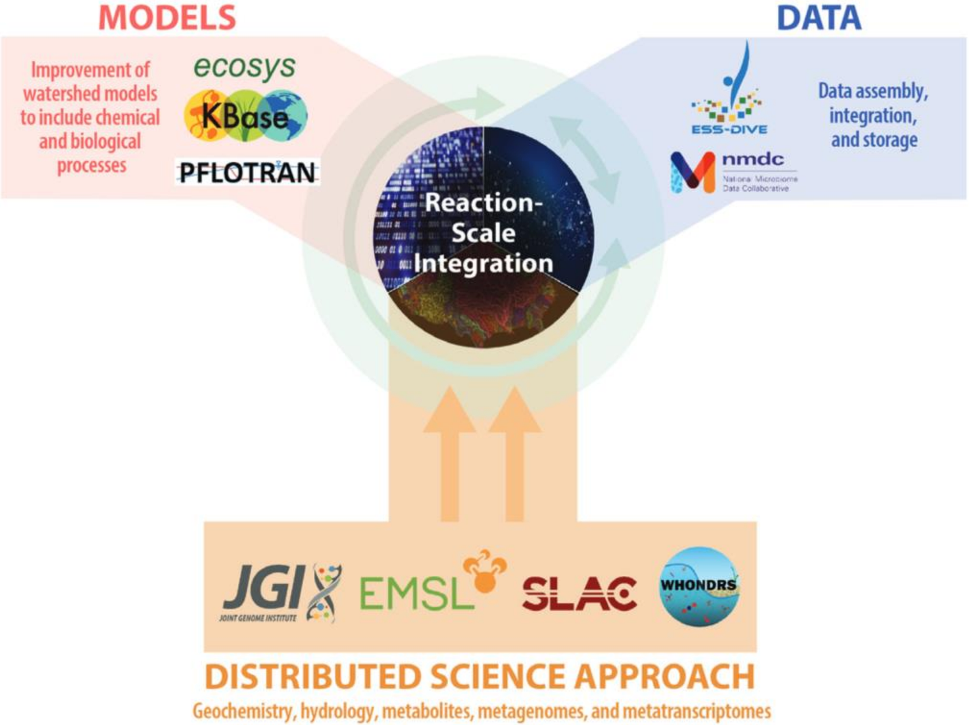
Genome Resolved Open Watersheds

Mikayla Borton

Kelly Wrighton

Microbial Ecosystems Lab
Colorado State University





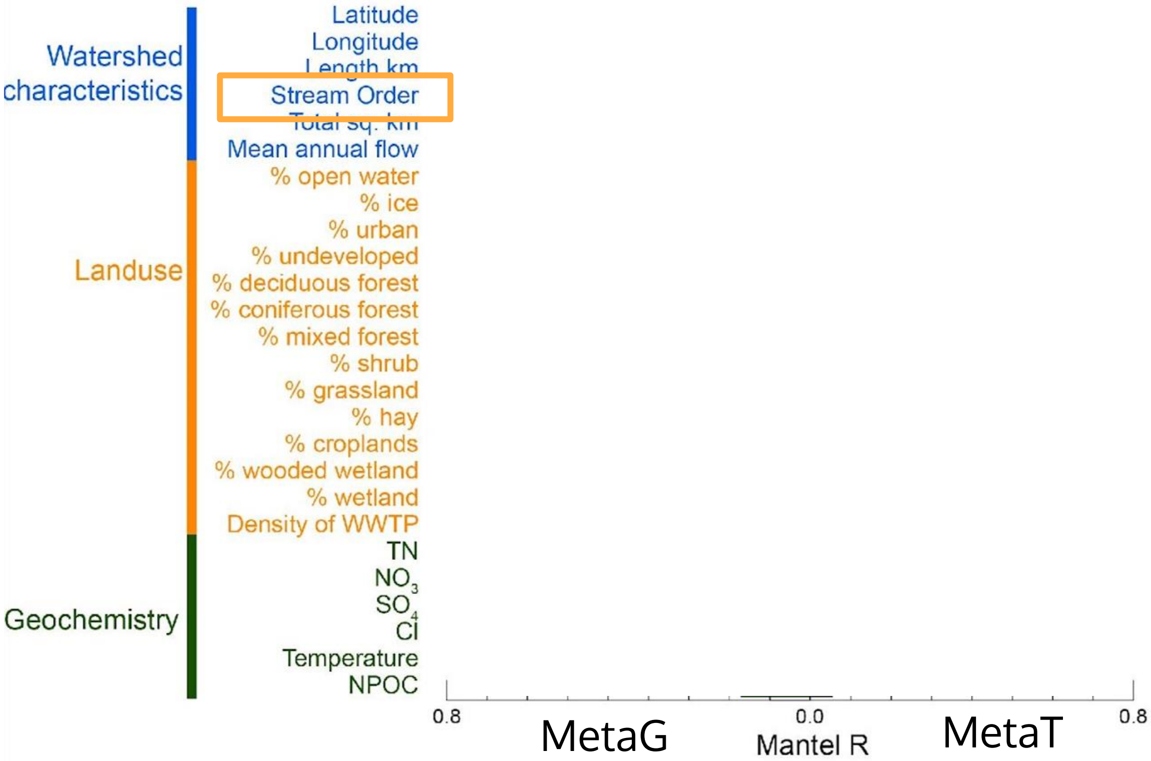
MODELS
Improvement of watershed models to include chemical and biological processes
ecosys
KBase
PFLOTRAN
DATA
ESS-DIVE
nmdc
National Microbiome Data Collaborative
Data assembly, integration, and storage
Reaction-Scale Integration
JGI
JOINT GENOME INSTITUTE
EMSL
SLAC
WHONDRS
DISTRIBUTED SCIENCE APPROACH
Geochemistry, hydrology, metabolites, metagenomes, and metatranscriptomes

U.S. DOE. 2019. Open Watershed Science by Design: Leveraging Distributed Research Networks to Understand Watershed Systems Workshop Report, DOE/SC-0200, U.S. Department of Energy Office of Science.

GROW Science

Hypothesis: Spatial and geochemical features influence river microbiomes

Results: Stream order, geochemistry, and temperature correlate with community structure



Data integration - provenance, credit

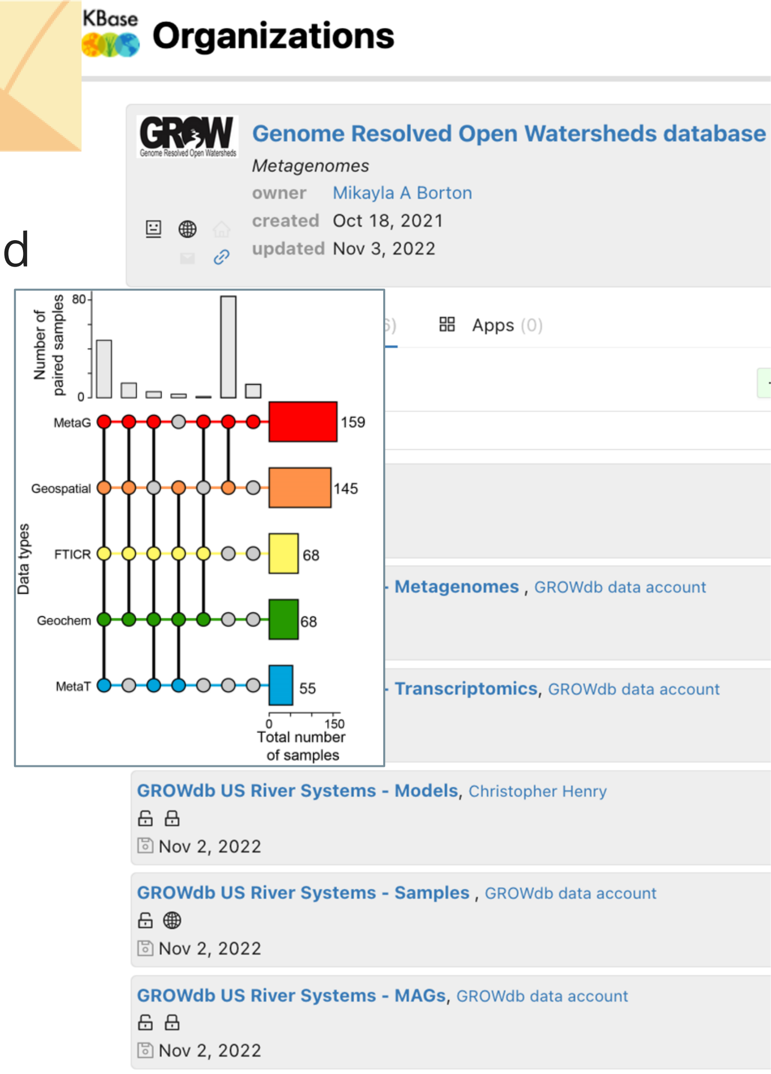
ESS-DIVE: sample metadata and biochemistry

JGI: metaG and metaT

EMSL: metaB

NMDC: study information, standardized analyses

KBase: combines everything to generate community models



KBase Organizations

Genome Resolved Open Watersheds database

Metagenomes

owner Mikayla A Borton

created Oct 18, 2021

updated Nov 3, 2022

Apps (0)

- Metagenomes , GROWdb data account

- Transcriptomics, GROWdb data account

GROWdb US River Systems - Models, Christopher Henry

Nov 2, 2022

GROWdb US River Systems - Samples , GROWdb data account

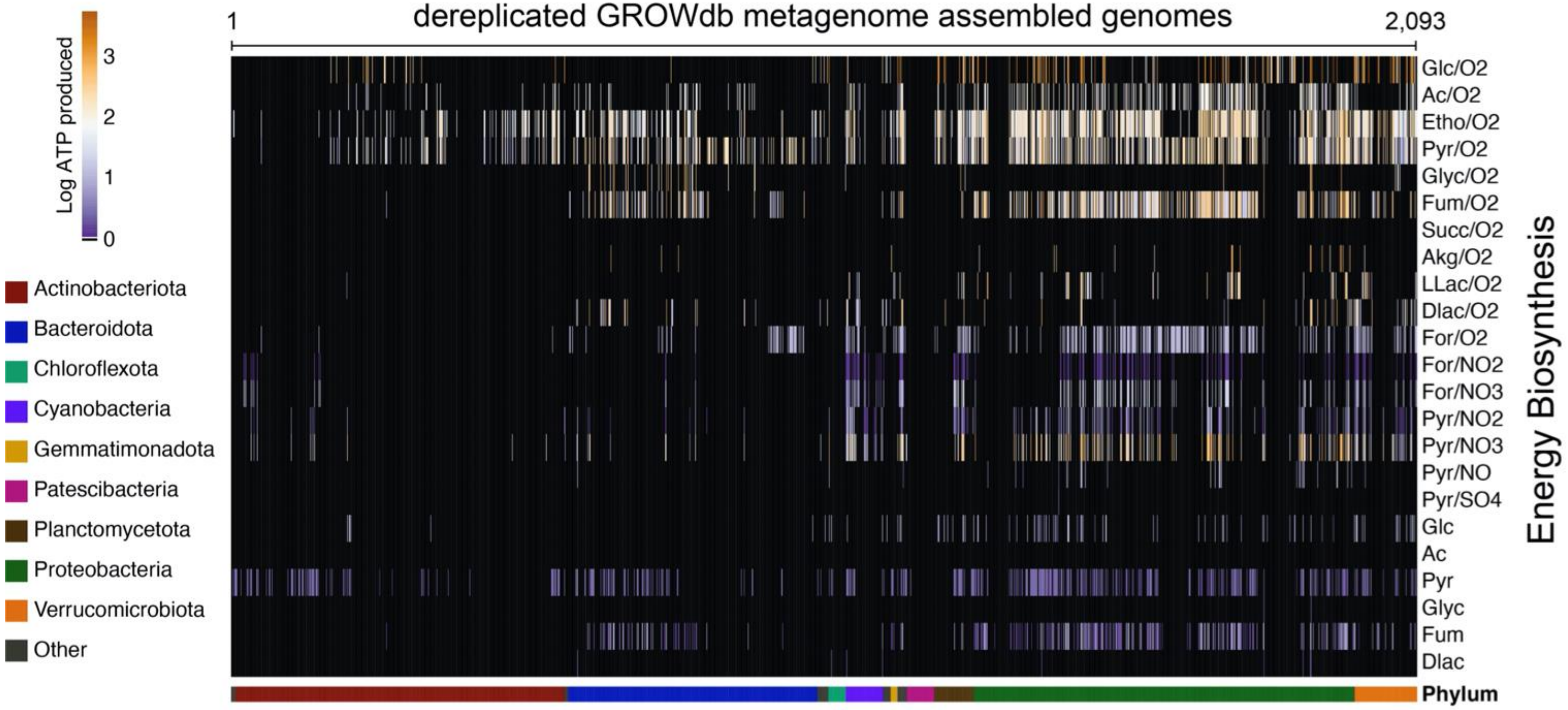
Nov 2, 2022

GROWdb US River Systems - MAGs, GROWdb data account

Nov 2, 2022

GROW Science

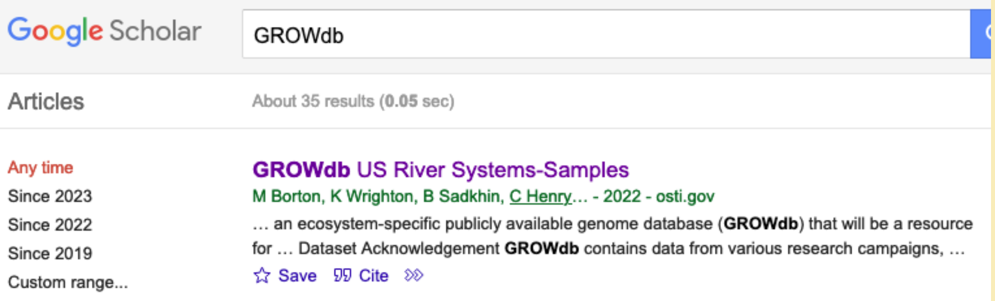
Metabolic models of metagenome-assembled genomes (MAGs) explore energy biosynthesis across major river phyla



Publishing to share and get credit

Samples, data, analyses - all FAIR and free!

Citation: Borton M, *et al.* (2022) GROWdb US River Systems - Samples. [Data set]. DOE Systems Biology Knowledgebase. doi:10.25982/109073.30/1895615.



GROWdb

Articles

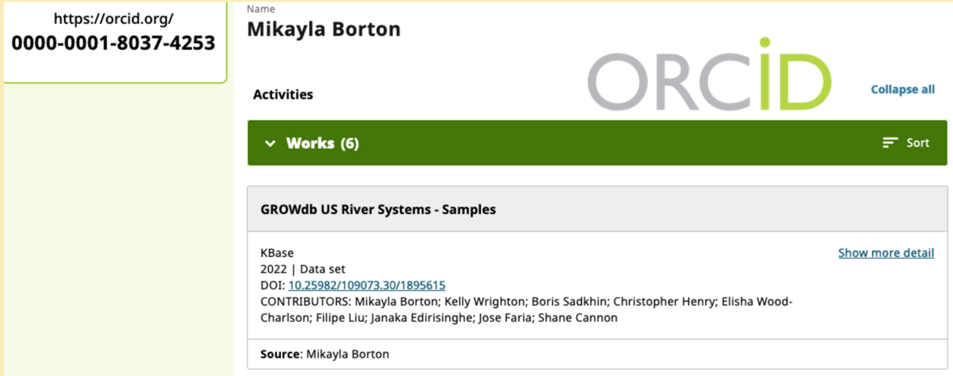
About 35 results (**0.05** sec)

Any time
Since 2023
Since 2022
Since 2019
Custom range...

GROWdb US River Systems-Samples

M Borton, K Wrighton, B Sadkhin, C Henry… - 2022 - osti.gov

… an ecosystem-specific publicly available genome database (**GROWdb**) that will be a resource for … Dataset Acknowledgement **GROWdb** contains data from various research campaigns, …



https://orcid.org/
0000-0001-8037-4253

Name
Mikayla Borton

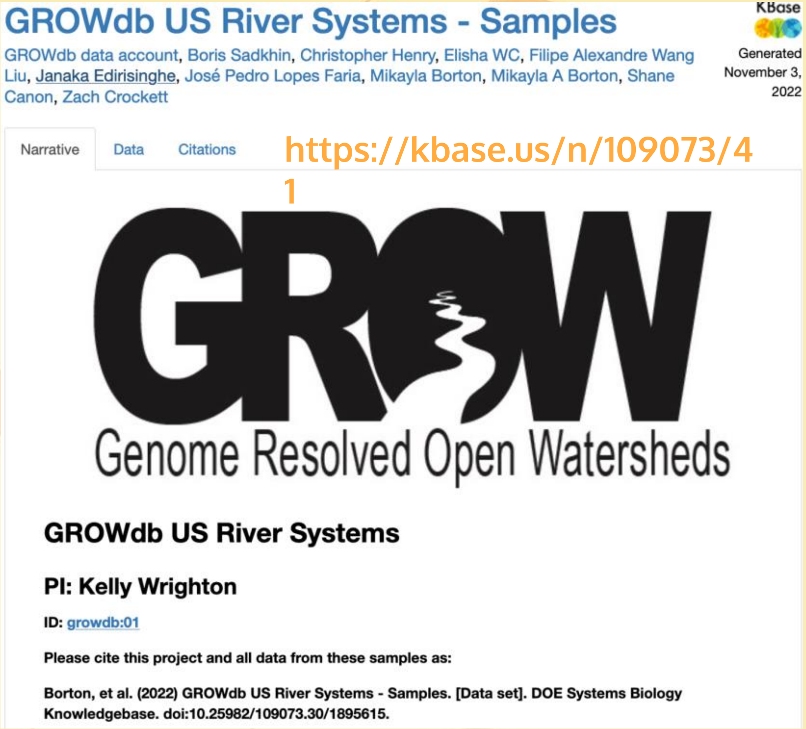
Activities

Works (6)

GROWdb US River Systems - Samples

KBase
2022 | Data set
DOI: 10.25982/109073.30/1895615
CONTRIBUTORS: Mikayla Borton; Kelly Wrighton; Boris Sadkhin; Christopher Henry; Elisha Wood-Charlson; Filipe Liu; Janaka Edirisinghe; Jose Faria; Shane Cannon

Source: Mikayla Borton



GROWdb US River Systems - Samples

GROWdb data account, Boris Sadkhin, Christopher Henry, Elisha WC, Filipe Alexandre Wang Liu, Janaka Edirisinghe, José Pedro Lopes Faria, Mikayla Borton, Mikayla A Borton, Shane Canon, Zach Crockett

Generated November 3, 2022

https://kbase.us/n/109073/41

GROWdb US River Systems

PI: Kelly Wrighton

ID: growdb:01

Please cite this project and all data from these samples as:

Borton, et al. (2022) GROWdb US River Systems - Samples. [Data set]. DOE Systems Biology Knowledgebase. doi:10.25982/109073.30/1895615.

U.S. DEPARTMENT OF
ENERGY
Office of Science

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Current and Future Efforts

Completing the KBase Credit Engine Inside and Out

Maturing our connection to the publishing infrastructure including ‘pushing’ MRA to ASM journals.

Working with large DOE team on a more universal data transfer system with long term goals of:

- Unified authorization validation
- Universal query with key terms across the resources
- Common identifiers, credit, etc.
- Common Data Transfer System: to uniformly move data among systems and validate on both ends. (near term)

BERKELEY LAB

OAK RIDGE National Laboratory

Argonne NATIONAL LABORATORY

Brookhaven National Laboratory

The KBase Team

U.S. DEPARTMENT OF ENERGY
Office of Science

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase