MoBEDAC -- Integrated data and analysis for the indoor and built environment

Folker Meyer

Argonne National Laboratory

GSC 13 – Shenzhen, China

ALFRED P. SLOAN FOUNDATION
1934

NGS is causing paradigm shift

Environmental clone libraries (“functional metagenomics”)

- $250 / 96 clones/reads (prep + sequencing)

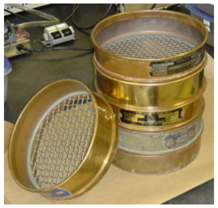
Amplicon studies (single gene studies, 16s rDNA)

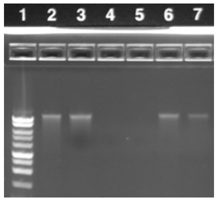
- $17 / 100,000 reads (PCR, barcoding, sequencing)

Shotgun metagenomics

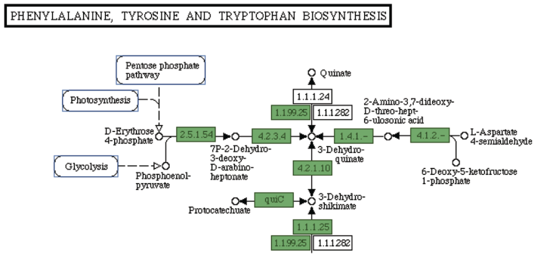
Cost for library, barcoding and sequencing

- $1200 / 10GBp / 100 million reads (single ended)
- $2400 / 20GBp / 100 million fragments (paired ends)

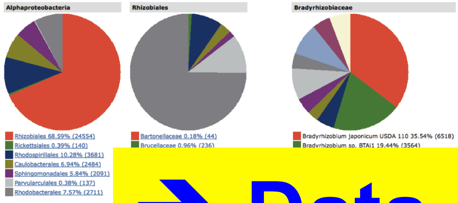




What are they doing?



Who are they?

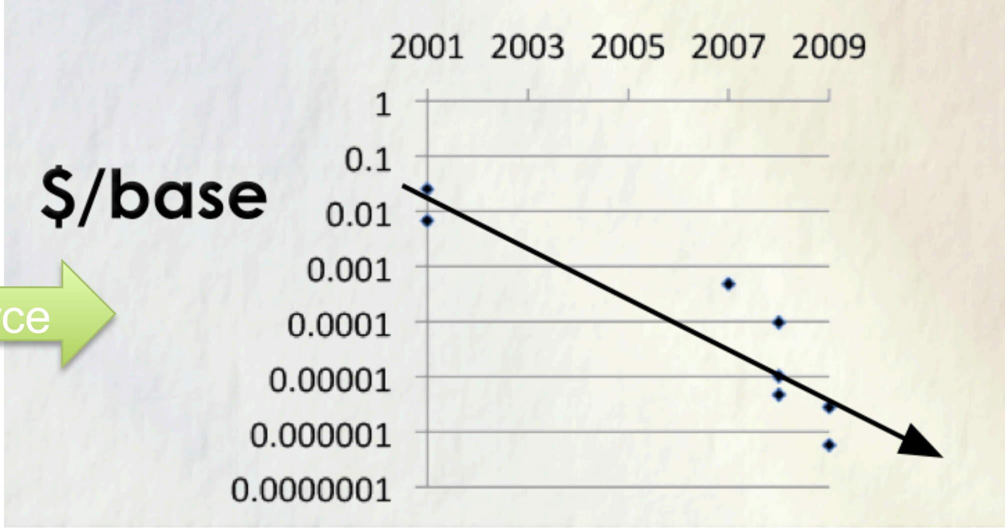


→ Data is cheap!

Background: Metagenomics data challenge

- Data growing fast:
  - 2004: C. Venter’s GOS with **600MBp** (or 0.6GBp)
  - 2011: HMP with **6TBp** (or 6,000GBp)
  - 2012: MG-RAST hits 11TBp (10 *10^12 bases)
- Sequencing cost will continue to drop
- Analysis needs to speed up 10x annually
- Analysis cost is 10x of sequencing cost

Driving force



Source: Rob Knight, UColorado

Background: Numerous data sources

- In the past just a few genome centers produced data, now hundreds of groups
- MG-RAST alone has 2500 data submitters
- Metadata coverage is sparse

- MAP OF Submissions

Background: Integration is missing

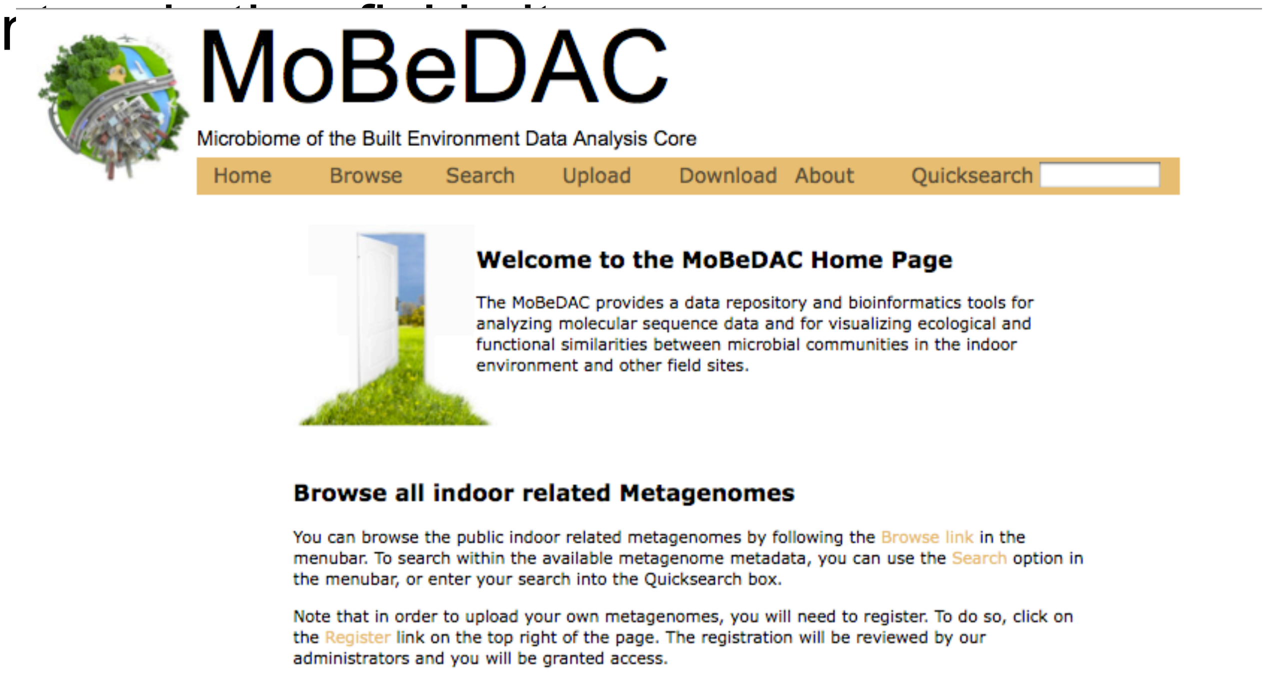
- There is no Genbank for metagenomes
- SRA is not functioning in that role
- Even if it did, it would be raw data only

- We lack an integration of data, analysis and pre-analyzed data!

Microbiome of the Built Environment Data Analysis Core (MoBeDAC)

What is MoBEDAC?

The MoBEDAC provides a data repository and bioinformatics tools for analyzing molecular sequence data and for visualizing ecological and functional similarities between microbial communities in the indoor environmen



MoBeDAC

Microbiome of the Built Environment Data Analysis Core

Welcome to the MoBeDAC Home Page

The MoBeDAC provides a data repository and bioinformatics tools for analyzing molecular sequence data and for visualizing ecological and functional similarities between microbial communities in the indoor environment and other field sites.

Browse all indoor related Metagenomes

You can browse the public indoor related metagenomes by following the Browse link in the menubar. To search within the available metagenome metadata, you can use the Search option in the menubar, or enter your search into the Quicksearch box.

Note that in order to upload your own metagenomes, you will need to register. To do so, click on the Register link on the top right of the page. The registration will be reviewed by our administrators and you will be granted access.

What is MoBEDAC?

- FungiDB
- QIIME
- MG-RAST
- VAMPS

Common

- Submission
- API
- Analysis (BIOM format)

Metadata standard working group

The MoBDAC PIs

Folker Meyer,
University of Chicago
ANL

Mitch Sogin
MBL

Rob Knight
University of
Colorado
Boulder

Jason
Stajich
University
of California
Riverside

BE minimal metadata working group

Argonne: Elizabeth Glass, Folker Meyer, Andreas Wilke
Colorado: Rob Knight, Doug Wendel, Bob Van Pelt
microBEnet: Hal Levin
UC Davis: Jonathan Eisen
UMD-SOM,IGS: Lynn Schriml
MBL: Mitch Sogin, Anna Shipunova
Sloan: Paula Olsiewski

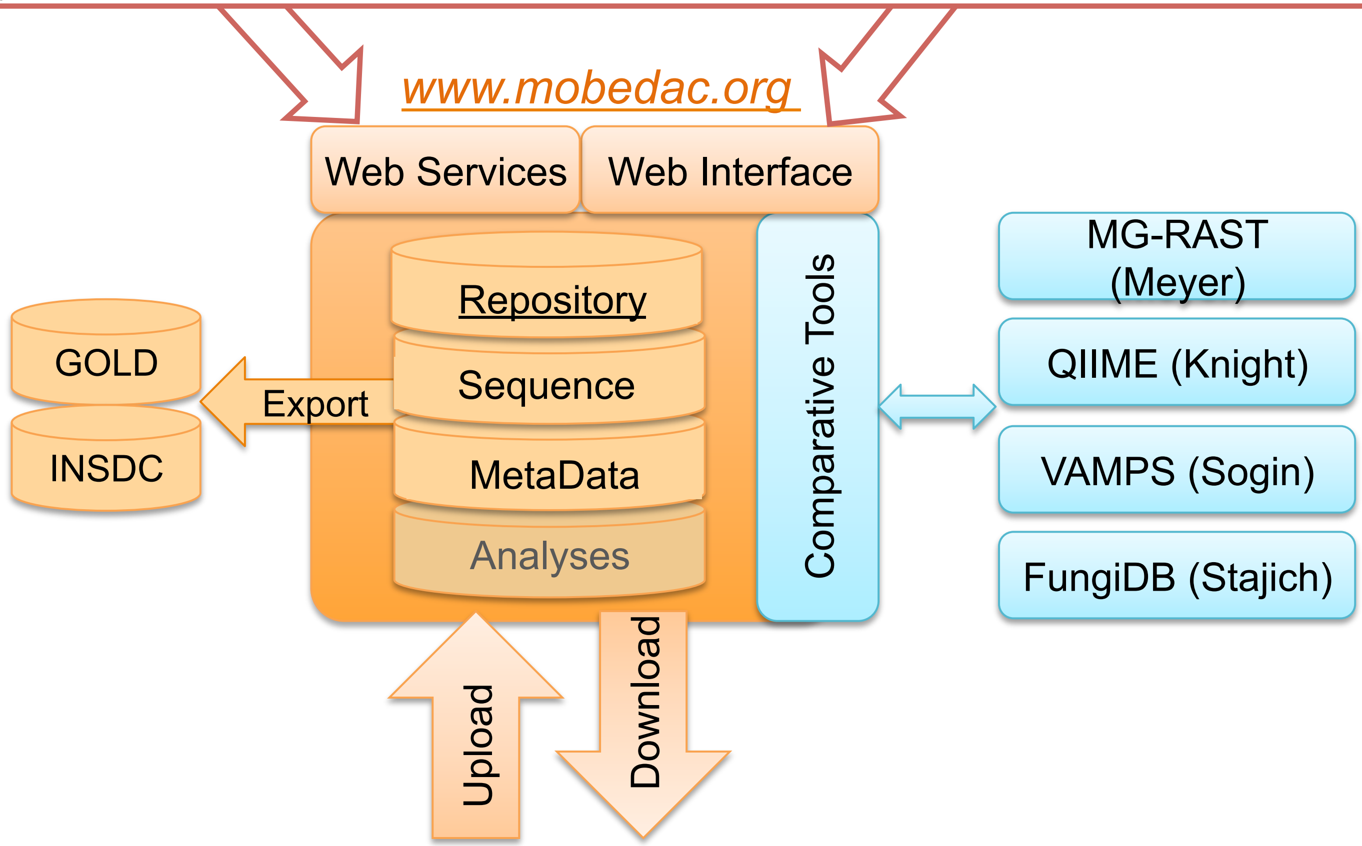
Complex Queries and Analysis

Retrieve and compare all 16s sequences and meta data from sample from industrial buildings.

Retrieve the set of samples for which both the V2 and V6 regions have been sequenced (for comparisons of primer bias).

Retrieve and compare a set of samples from cities in which both drinking water and sewage have been sampled (to allow comparisons of contamination levels and source tracking).

Retrieve and compare metabolic profiles of samples from waste water that were sequenced using HISEQ.



www.mobedac.org

Web Services

Web Interface

Repository

Sequence

MetaData

Analyses

Comparative Tools

GOLD

INSDC

Export

Upload

Download

MG-RAST (Meyer)

QIIME (Knight)

VAMPS (Sogin)

FungiDB (Stajich)

Data Portal and Repository

- No single analysis tool could satisfy all researchers across metagenomics; a federated approach to analysis is required.
  - At the same time, the size of data sets from next-generation sequencing platforms have made these data sets difficult to move and share.
- The MoBEDAC will act as an archive for all sequence data (plus metadata) and analysis generated in the Sloan IE program, allowing PIs easy upload directly or via one of the tools participating in the MoBEDAC project.
- We will provide unified access to all sequence data created in the program, as well as from other relevant IE programs.

Repository and Data Synchronization

- MoBEDAC will include mechanisms to automatically retrieve pertinent datasets from various websites and archives, including data relevant to the indoor environment from INSDC, KEGG, SEED, VAMPS, GOLD, SRA, QIIME, MG-RAST, FungiDB, and IMG/M as well as corresponding metadata.
- We will accommodate existing exchange and data formats for inclusion in the repository.
- Sequence data collected and integrated will be provided in various formats and made available via FTP download or web services.
- Metadata will be available in GCDML format.

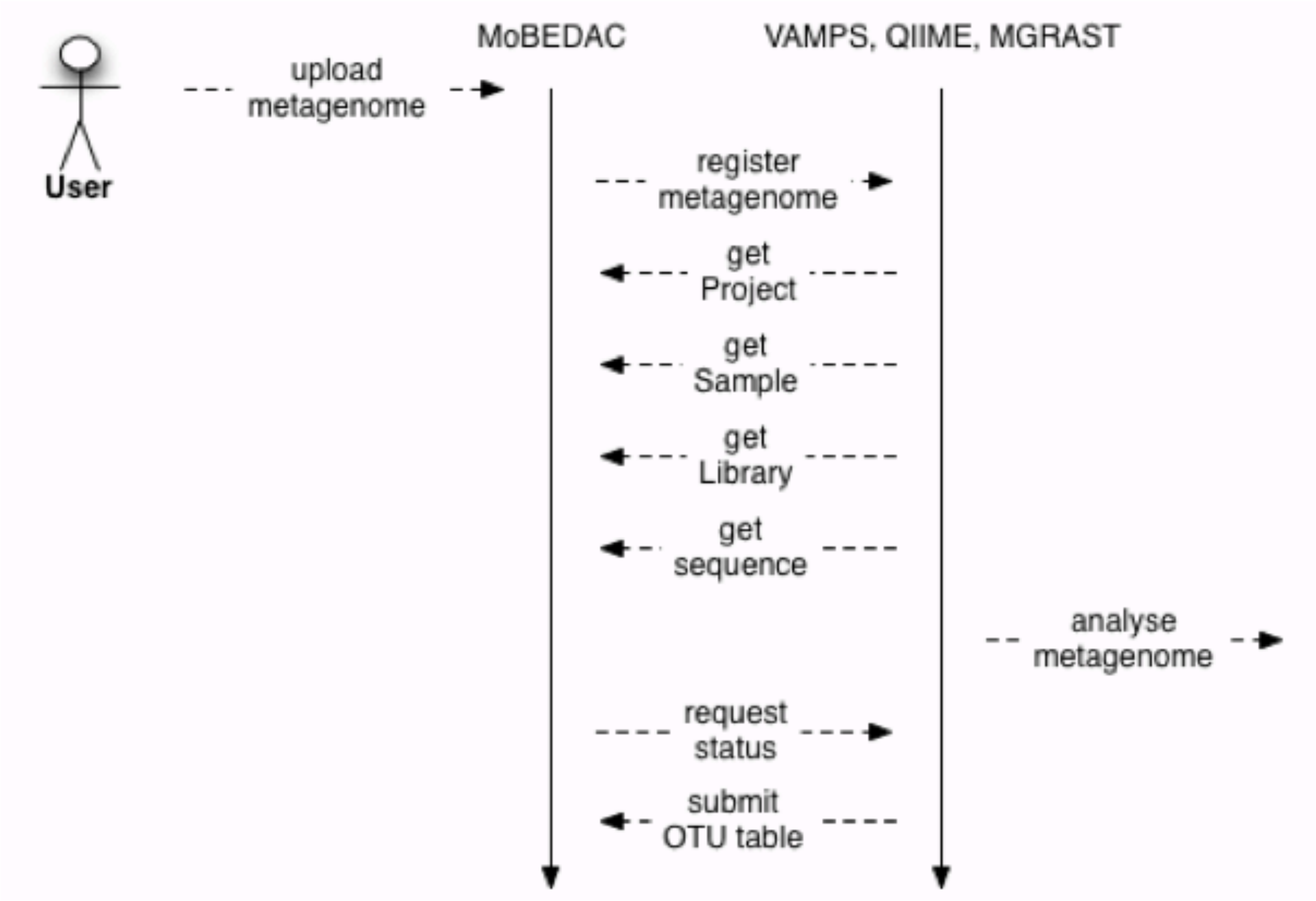
Metadata

- Metadata provides an essential complement to sequence data, helping answer questions about its source, mode of collection, and reliability.
- Metadata collection and interpretation have become vital to the IE community, but considerable challenges remain, including exchange, curation, distribution, and IE-specific standards.
- Communication and feedback from the IE community is vital.
- We have developed a GSC-compliant BE minimal metadata package (Glass and Schriml).

Mechanisms Enabling Metadata-driven Queries for Sequence Data

- Mechanism to enable download from the MoBEDAC and linking to analysis results on existing analysis servers (VAMPS, QIIME, MG-RAST, and FungiDB).
- The query results can be of two kinds:
  - datasets for download or
  - links to the analysis of those datasets in existing tools.
- Enables researchers to obtain an overview of microbial communities for existing data sets with various tools.
- The query results returned via web pages or web services.
- The MoBEDAC team is also developing data management capabilities for the core. These will support prepublication project creation and data sharing by PIs via web-based tools.

APIs


User
MoBEDAC
VAMPS, QIIME, MGRAST
upload metagenome
register metagenome
get Project
get Sample
get Library
get sequence
analyse metagenome
request status
submit OTU table

When will this be available?

Timeline

1st Beta testing
March 2012

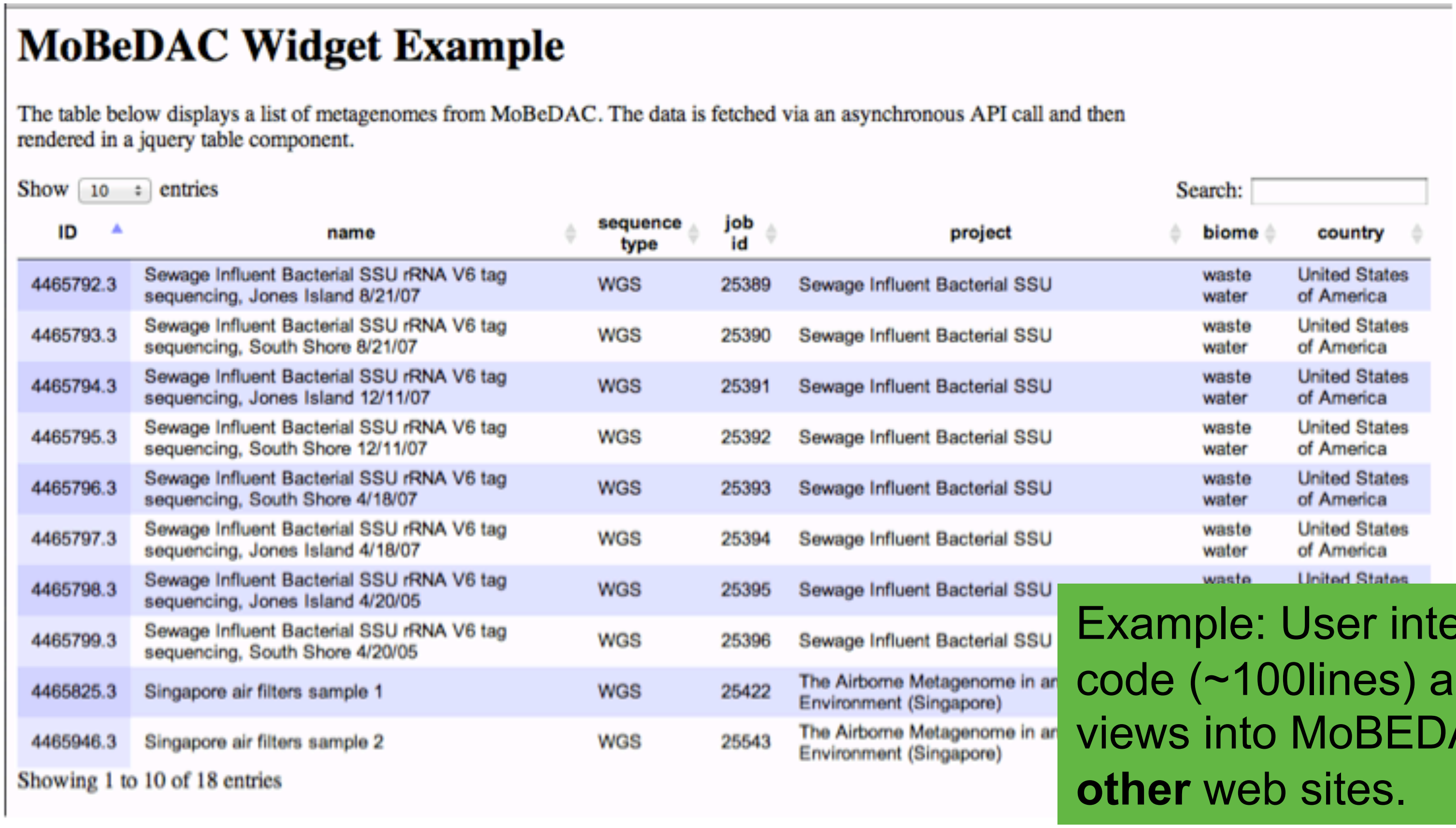
2nd Beta testing
March/April 2012

Integration of Feedback
April 2012

MoBEDAC
Public Launch
May 2012

Web integration – Widgets!

- Next phase: Widget to allow integration of views into MoBEDAC integration (prototype)



MoBeDAC Widget Example

The table below displays a list of metagenomes from MoBeDAC. The data is fetched via an asynchronous API call and then rendered in a jquery table component.

ID	name	sequence type	job id	project	biome	country
4465792.3	Sewage Influent Bacterial SSU rRNA V6 tag sequencing, Jones Island 8/21/07	WGS	25389	Sewage Influent Bacterial SSU	waste water	United States of America
4465793.3	Sewage Influent Bacterial SSU rRNA V6 tag sequencing, South Shore 8/21/07	WGS	25390	Sewage Influent Bacterial SSU	waste water	United States of America
4465794.3	Sewage Influent Bacterial SSU rRNA V6 tag sequencing, Jones Island 12/11/07	WGS	25391	Sewage Influent Bacterial SSU	waste water	United States of America
4465795.3	Sewage Influent Bacterial SSU rRNA V6 tag sequencing, South Shore 12/11/07	WGS	25392	Sewage Influent Bacterial SSU	waste water	United States of America
4465796.3	Sewage Influent Bacterial SSU rRNA V6 tag sequencing, South Shore 4/18/07	WGS	25393	Sewage Influent Bacterial SSU	waste water	United States of America
4465797.3	Sewage Influent Bacterial SSU rRNA V6 tag sequencing, Jones Island 4/18/07	WGS	25394	Sewage Influent Bacterial SSU	waste water	United States of America
4465798.3	Sewage Influent Bacterial SSU rRNA V6 tag sequencing, Jones Island 4/20/05	WGS	25395	Sewage Influent Bacterial SSU	waste	United States
4465799.3	Sewage Influent Bacterial SSU rRNA V6 tag sequencing, South Shore 4/20/05	WGS	25396	Sewage Influent Bacterial SSU		
4465825.3	Singapore air filters sample 1	WGS	25422	The Airborne Metagenome in ar Environment (Singapore)		
4465946.3	Singapore air filters sample 2	WGS	25543	The Airborne Metagenome in ar Environment (Singapore)		

Showing 1 to 10 of 18 entries

Example: User interface code (~100lines) allows views into MoBEDAC from **other** web sites.