Promoting, understanding, recording and utilizing metadata in genomic/metagenomic studies

Alex Thomas[1*], Tatiparthi Reddy[1], Michelle Isbandi[1], Jyothi Mallajosyula[1], Dimitrios Stamatis[1], Jonathan Bertsch[1], Nikos Kyrpides[1]

[1] LBNL Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA

**To whom correspondence should be addressed*: Email: AlexanderThomas@lbl.gov

March 21, 2014

ACKNOWLEDGMENTS:

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Promoting, understanding, recording, and utilizing metadata in genomic/metagenomic studies

Alex Thomas*, T.B.K. Reddy, Michelle Isbandi, Jyothi Mallajosyula, Dimitrios Stamatis, Jonathon Bertsch, Nikos Kyrpides

DOE Joint Genome Institute, Walnut Creek, CA, USA

What is metadata?

Metadata is data about data (NISO, 2004). A genomic or metagenomic project's metadata includes the sequencing methods and statistics. Metadata also describe the taxonomy, physical characteristics, and environment of the sequence source organism.

Why is metadata important?

It is critical to ensure the quality of metadata for genome and metagenome projects to facilitate database queries, comparative analyses, and hypothesis testing. Missing or misleading metadata can reduce database search results and negatively impact interpretation of analyses.

Genomes OnLine Database

The Genomes OnLine Database (GOLD) is an online catalog of genome and metagenome project metadata. The ability to find projects in GOLD depends on the quantity and quality of metadata provided by users (Pagani et al. 2012).

www.genomesonline.org

Integrated Microbial Genomes

Integrated Microbial Genomes (IMG) is a data warehouse that provides genome analysis tools. Defining a project in GOLD is mandatory for using IMG. Metadata from GOLD enhance the results of analyses in IMG (Markowitz et al. 2014).

https://img.jgi.doe.edu/

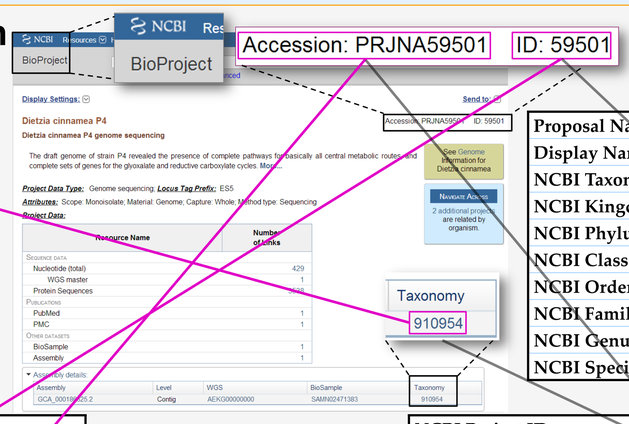
Genomic Study

Organism Information

Good Example

Proposal Name	Dietzia cinnamea P4
Display Name	Dietzia cinnamea P4
NCBI Taxon ID	910954
NCBI Kingdom	Bacteria
NCBI Phylum	Actinobacteria
NCBI Class	Actinobacteria
NCBI Order	Actinomycetales
NCBI Family	Dietziaceae
NCBI Genus	Dietzia
NCBI Species	Dietzia cinnamea

NCBI Project ID	59501
NCBI Bioproject Accession ID	PRJNA59501
NCBI Project Name	Dietzia cinnamea P4



Bad Example

Proposal Name	Dietzia cinnamea P4
Display Name	Dietzia cinnamea P4
NCBI Taxon ID	59501
NCBI Kingdom	Eukaryota
NCBI Phylum	Streptophyta
NCBI Class	
NCBI Order	Rosales
NCBI Family	Rosaceae
NCBI Genus	Rubus
NCBI Species	Rubus idaeus

NCBI Project ID	PRJNA59501
NCBI Bioproject Accession ID	910954
NCBI Project Name	Dietzia cinnamea P4

Environment Metadata

Good Example

More information about Dietzia cinnamea P4

Isolation Site	microcosms containing oil-contaminated soil collected from an environmentally protected area of a tropical Atlantic forest (Biological Reserve of Poco das Antas)
Strain Habitat	Oil polluted soil
Isolation Country	Brazil
Isolation Pubmed ID	17174505
Geographic Location	Biological Reserve of Poco das Antas
Latitude	-29.402
Longitude	-51.629
Lat/Long Information	Inferred

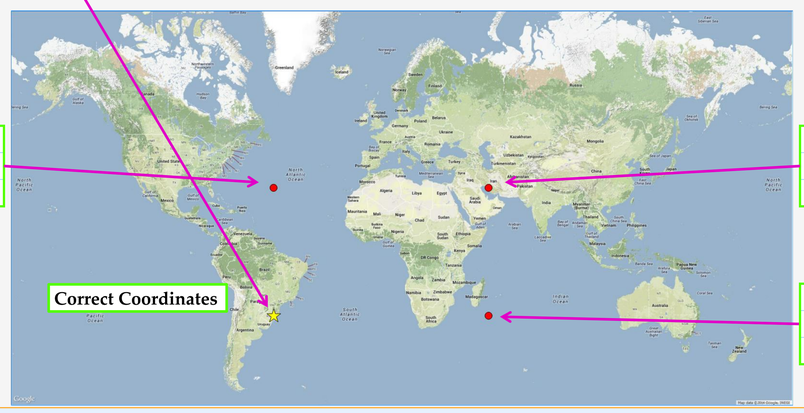
Bad Example: Missing Information

Isolation Site	intestinal tract of child
Strain Habitat	Human gastrointestinal tract

Missing information can take considerable effort to find, if it can be located at all

Bad Example: Misinformation

A common mistake is to misplace the negative sign on the Latitude or Longitude coordinates.
The sign does matter!



Incorrect Coordinates	
Latitude	29.402
Longitude	-51.629

Incorrect Coordinates	
Latitude	29.402
Longitude	51.629

Incorrect Coordinates	
Latitude	-29.402
Longitude	51.629

Utilizing GOLD Genome Metadata in IMG tools

Find genomes from similar environments

Query Genomes

Using "metadata category operation"
Search by:
1) Species Habitat = soil
2) Temperature Range = Mesophile
3) Display Isolation
4) Filter Isolation fields containing "oil"

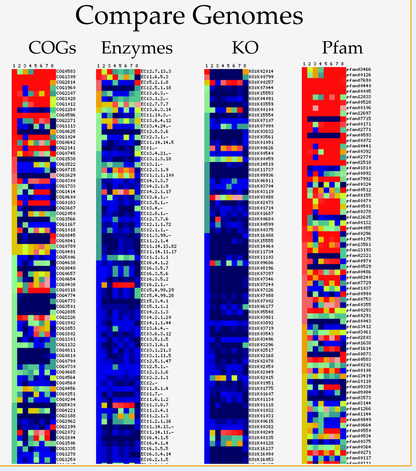
Results: 8 genomes identified

Missing Metadata

Unfortunately, ~30,000 projects (out of ~40,000) have no value for temperature range. How many more genomes could have been found?

Select	Domain	Status	Genome Name	Proposal Gold ID	Sample Gold ID	Submission ID	Habitat	Isolation	Temperature Range
	B	F	Acinetobacter calcoaceticus PHEA-2	Gc01672			Soil	industrial wastewater of an oil refinery in China	Mesophile
	B	F	Amycolicicoccus subflavus DQS3-9A1	Gc01773			Soil	crude oil-polluted saline soil in Daqing Oilfield in China	Mesophile
			Bacillus cereus Q1	Gc00895			Soil	deep-subsurface oil reservoir in Daqing oilfield, Northeastern China	Mesophile
			Dietzia cinnamea P4	Gi08236			Soil	microcosms containing oil-contaminated soil collected from an environmentally protected area of a tropical Atlantic forest (Biological Reserve of Po? o das Antas)	Mesophile
	B	D	Rhodococcus opacus M213	Gi17151		12143	Soil	Fuel-oil contaminated soil	Mesophile
			Rhodococcus pyridinivorans AK37	Gi17083		8884	Soil	Crude oil-contaminated site in Hungary	Mesophile
			Rhodococcus sp. DK17	Gi17726		24556	Soil	Crude oil-contaminated soil	Mesophile
			Shewanella putrefaciens 200 (Missing data)	Gc01637		6985	Soil	Crude-oil pipeline from Alberta, Canada	Mesophile

Compare Genomes



Literature Cited

NISO. Understanding Metadata. (NISO Press, 2004).

Pagani, I. *et al.* The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **40,** D571–9 (2012).

Markowitz, V. M. *et al.* IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* **42,** D560–7 (2014).

GOLD←"IMG-GOLD"→img

IMG-GOLD: interface for defining projects. Required for submitting projects to IMG
GOLD: public catalog of genome/metagenome project metadata
IMG: tools for comparative and functional genome analysis

Metagenomic Study

Metagenome Information

Good Example

Gold Study Name	Hydrocarbon Resource Environments Microbial Communities from Canada and USA
Submitter's Study Name	Sean M. Caffrey
NCBI Taxon ID (if any)	938273
Domain	MICROBIAL
Ecosystem (*)	Engineered
Ecosystem Category (*)	Wastewater
Ecosystem Type (*)	Industrial wastewater
Ecosystem Subtype	Petrochemical
Specific Ecosystem	Unclassified
NCBI Project Name	hydrocarbon metagenome
Habitat	Hydrocarbon Resource Environments
Community	Microbial communities
Location	from Canada and USA

Five levels of classification for metagenome studies

Bad Example

Gold Study Name	hkbic
Domain	MICROBIAL
Ecosystem (*)	Engineered
Ecosystem Category (*)	Air
Specific Ecosystem	Unclassified

Unfortunately many studies are given cryptic names, environmental information is omitted, and fields are populated with ambiguous terms.

Standardized naming convention based on:
- Habitat: ex. Hydrocarbon Resource Environments
- Community: ex. Microbial communities
- Location: ex. from Canada and USA
- Identifier: ex. N/A, would describe specific type of community such as thermophilic

Metagenome Sample Information

Good Example

ER Sample ID	2061
Sample Study Name	Hydrocarbon Resource Environments Microbial Communities from Canada and USA
GOLD Sample ID	Gs0005346
ER Study ID	2444
Sample Display Name	Microbes from Sediment core from a heavy oil reservoir,Alberta Canada Inniskillen 614.3
Biosample Name	Sediment ecosystem from Alberta, Canada
Submitter's Name	Inniskillen 614.3
NCBI Taxonomy ID	938273
IMG Object ID	3300001197
Ecosystem	Environmental
Ecosystem Category	Terrestrial
Ecosystem Type	Oil reservoir
Ecosystem Subtype	Oil reservoir
Sample Type	Metagenome

Metagenome samples are nested in projects

Samples can be from different locations, so information can be unique for each sample

Sample Description	Sediment core from a heavy oil reservoir, Alberta, Canada.
Sampling Site	Sediment core from a heavy oil reservoir, Alberta, Canada.
Sample Collection Date	Jul-08
Sampling Strategy	Scuba diving
Sample Isolation Country	Canada
Geographic Location	Alberta, Canada
Latitude	56.04
Longitude	-118.13
Sample Isolation Site	Sediment core from a heavy oil reservoir, Alberta, Canada.

It is also important to include additional information to describe each sample

Sequencing metadata are also extremely important

Sequencing Center	McGill Univ
Sequencing Methods	454-GS-FLX-Titanium, Illumina HiSeq 2000

Utilizing GOLD Metagenome Metadata in IMG/M tools

Find metagenomes from similar locations

Query metagenomes

Select	Domain	Status	Study Name	Genome Name / Sample Name	Sequencing Center	Ecosystem Type
	*	D	Hydrocarbon resource environments microbial communities from Canada and USA	Microbes from Sediment core from a heavy oil reservoir Alberta Canada Inniskillen 614.3 (Inniskillen 614.3: 454 sequencing assembly)	McGill Univ	Oil reservoir
	*	D	Hydrocarbon resource environments microbial communities from Canada and USA	sands outcrops, Fort McMurray, Alberta, Canada (H1C: 454 sequencing assembly)	McGill Univ	Soil
	*	D	Hydrocarbon resource environments microbial communities from Canada and USA	Microbes from Oil-contaminated ecosystem in Alberta, Canada Inniskillen 604.3 (Inniskillen 604.3: 454+illumina sequencing assembly)	McGill Univ	Soil
	*	D	Hydrocarbon resource environments microbial communities from Canada and USA	Microbes from Oil-contaminated ecosystem in Alberta, Canada Inniskillen 604.3 (Inniskillen 604.3: 454+illumina sequencing assembly)	McGill Univ	Oil reservoir

Browse genomes
1) Filter Genome Name/ Sample Name containing 'Alberta''

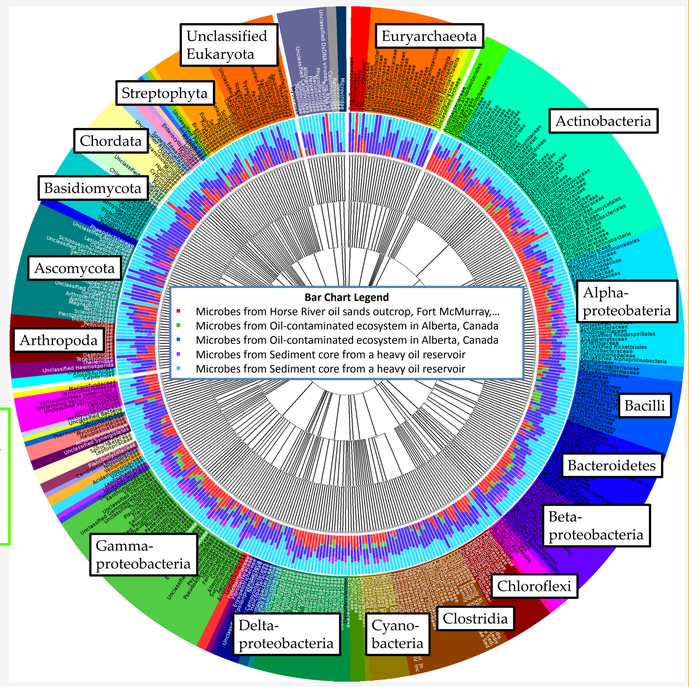
Results: 18 metagenomes identified

2) Display Ecosystem Type
3) Select "Oil Reservoir" and "Soil" studies

4) Compare the phylogenetic distribution of genes from the five metagenomes from oil reservoirs and soil in Alberta, Canada.

Missing Metadata

Unfortunately, 2844 metagenome samples (out of 4507) have no value for geographic location. How many more metagenomes could have been found?



JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231