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

Alex Thomas¹*, Tatiparthi Reddy¹, Michelle Isbandi¹, Jyothi Mallajosyula¹, Dimitrios Stamatis¹, Jonathan Bertsch¹, Nikos Kyrpides¹

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.

DISCLAIMER:

LBNL: This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.
**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

### Genomic Study

**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.

### Metagenomic Study

**Organism Information**
- **Proposal Name**: Dietzia cinnamia P4
- **Display Name**: Dietzia cinnamia P4
- **NCBI Taxon ID**: 91994
- **NCBI Kingdom**: Bacteria
- **NCBI Phylum**: Actinobacteria
- **NCBI Class**: Actinobacteria
- **NCBI Order**: Actinomycetales
- **NCBI Family**: Dietziaceae
- **NCBI Genus**: Dietzia
- **NCBI Species**: Dietzia cinnamia

**Environment Metadata**
- **Good Example**
  - Isolation Site: Intestinal tract of child
  - Strain Habitat: Human gastrointestinal tract

**Bad Example: Missing Information**
- Isolation Site: Intestinal tract
- Strain Habitat: Human gastrointestinal tract

**Environment Metadata**
- **Bad Example: Misinformation**
  - A common mistake is to misplace the negative sign on the Latitude or Longitude coordinates. The sign does matter!

**Environment Metadata**
- **Incorrect Coordinates**
  - Latitude: 29.402
  - Longitude: -51.629

**Correct Coordinates**
- Latitude: 29.402
- Longitude: 51.629

### 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 IMG is mandatory for using IMG. Metadata from GOLD enhance the results of analyses in IMG (Markowitz et al. 2014).

**https://img.jgi.doe.gov/**

### Literature Cited

---

**Utilizing GOLD Metagenome Metadata in IMG/M tools**

1. **Find genomes from similar locations**
   - Using “metadata category operation”
   - Search by: 1) Species Habitat = soil 2) Temperature Range = Mesophile 3) Display Isolation 4) Filter Isolation fields containing “soil”
   - Results: 8 genomes identified

2. **Missing Metadata**
   - Unfortunately, ~300,000 projects (out of ~40,000) have no value for temperature range. How many more genomes could have been found?

---

**Gold Study Name**
- Hydrocarbon Resource Environments Microbial Communities from Canada and USA

**Sample Description**
- Sediment core from a heavy oil reservoir, Alberta, Canada

**Sample Collection Date**
- Jul-08

**Sampling Strategy**
- Scuba diving

**Geographic Location**
- Alberta, Canada

**Latitude**
- 56.04

**Longitude**
- -118.13

**Sample Site Isolation**
- Sediment core from a heavy oil reservoir, Alberta, Canada

**Sequencing Center**
- McGill Univer.

**Sequencing Methods**
- 454 GS-FLX, Titanium, Illumina HiSeq 2000

---

**Genomes OnLine Database**

**GOLD**

**IMG-GOLD**

**img**

**Flowchart**

**COCs**
- Enzymes
- KO
- Plasmid

**Image**
- Microbial Communities from Canada and USA

**Comparison**
- Filter Genome Name/ Sample Name containing “Alberta”
- Results: 18 metagenomes identified

**Query metagenomes**

1. **Browse genomes**
   - 1) Filter Genome Name/ Sample Name containing “Alberta”
   - 2) Display Ecosystem Type
   - 3) Select “Oil Reservoir” and “Soil” studies

2. **Compare the phylogenetic distribution of genes from the five metagenomes with soil in Alberta, Canada.**