A Call for Standardized Classification of Metagenome Projects

Natalia Ivanova, Susannah G. Tringe, Konstantinos Liolios,
Wen-Tso Liu[1], Philip Hugenholtz and Nikos C. Kyrpides

DOE Joint Genome Institute, Walnut Creek, CA 94598, USA
Department of Civil and Environmental Engineering,
University of Illinois at Urbana-Champaign, IL 61801, USA

July 2010

ACKNOWLEDGMENT

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

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.

Genomics Update

A Call for Standardized Classification of Metagenome Projects

Natalia Ivanova, Susannah G. Tringe, Konstantinos Liolios, Wen-Tso Liu[1], Philip Hugenholtz and Nikos C. Kyrpides

DOE Joint Genome Institute, Walnut Creek, CA 94598, USA, [1]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, IL 61801, USA

Everyone would agree that metagenomics has been a great boon to the field of environmental microbiology. Fueled by major advances in sequencing technology, the number of metagenome projects has exploded in recent years, with hundreds of environmental samples having been interrogated by shotgun sequencing (Markowitz et al., 2008; Meyer et al., 2008; Liolios et al., 2009). As a result, while just a few years ago it was possible for an individual investigator to be familiar with the major shotgun metagenomic datasets, today there are far too many to easily recite. Therefore we argue that the time is ripe for developing and implementing a metagenome classification system.

Why classify metagenomes? The ability to extract, study and understand information from genomic data depends heavily on comparative analysis, and metagenomic data is no exception. Yet the appropriate comparisons to make are much less clear for metagenomes than for genomes, where the choice of comparison can be guided by phylogenetic classification. Moreover, even if the type of environmental studies one would want to compare to is known, it still remains difficult to know how many and which are available given the lack of systematic nomenclature describing these projects (i.e. standardized naming) or categorization. For example, if you were looking for metagenomes from organisms in the digestive tracts of various animals, they might be named “gut” but could also be “rumen”, “forestomach”, “cecum” or “fecal” communities.

Currently metagenomic projects are not systematically classified. NCBI’s metagenomic project catalog has implemented a simple and general project type distinction between “environmental” and “host-associated” projects (named correspondingly as Ecological and Organismal). This shallow classification is a starting point but does not address the many other environmental features potentially of interest for comparison. In order to circumvent the present difficulty in identifying appropriate metagenomic projects for comparative analysis, we present here a five-tiered metagenome naming and classification scheme. The top level includes the broad NCBI categories, but we also add a third “engineered” category that separates out manipulated communities such as bioreactors or treatment plants from natural environmental communities (Figure 1). Each of these is then subcategorized according to a variety of criteria, taking into account knowledge of key variables that influence community composition (e.g. salinity (Lozupone and Knight, 2007) or soil pH (Lauber et al., 2009)). Where possible, we have taken advantage of existing classification systems such as the Environment Ontology (EnvO; http://www.environmentontology.org/). Environmental communities are separated by the ecosystem category (aquatic, terrestrial, air) and ecosystem type (e.g. freshwater, marine) with more detailed categorizations based on specific features (e.g. salinity, pH). Host-associated communities are defined by host phylogeny, then sampling site; and finally engineered communities are classified by their function (e.g. bioremediation or food production) with further levels based on specific substrates or features. In some cases an individual "project" may span multiple categories because it includes samples from different habitat types. A sampling of the higher-level categories is shown in Table 1, and the complete proposed schema is available from GOLD (Genomes OnLine Database, www.genomesonline.org) and IMG/M (http://img.jgi.doe.gov/m/). Although we developed this schema to address an immediate need within these databases, we hope that it will provide the basis for a broadly sanctioned classification system coordinated by the Genomics Standards Consortium (GSC). We are also working to standardize naming of projects, with names that incorporate information not only on habitat (as defined in the classification schema) but also community type (e.g. microbial or viral), project type (e.g. metatranscriptome), geographical location (e.g. Yellowstone or Southern Ocean), and project-specific identifiers (e.g. proctodeal segment 1).

Constructing a classification-based "tree" and populating it with the metagenome project data collected in GOLD allows one to see what sort of environments have been well studied and which are unexplored (Figure 1). Much like the case in genomics, metagenomes have been chosen for sequencing based on idiosyncratic criteria rather than any systematic approach, and therefore the "tree" has not been evenly sampled. Within the host-associated category, not surprisingly, human studies dominate and digestive system communities are the primary target for all animal studies as this is the niche most heavily colonized by microorganisms. Within the environmental category, aquatic environments are much more heavily studied than terrestrial, perhaps due to the perceived intractability of complex soil communities.

Categorization and naming systems go hand-in-hand with efforts to standardize metadata collection for metagenome samples (Garrity et al., 2008) and cannot exist without them. Many published metagenome datasets cannot be readily classified based on available data; in some cases the relevant information may have been collected but there is simply no forum for capturing it. When investigators submit their sequence data to comparative metagenomics databases, such as IMG/M (Markowitz et al., 2008) and MG-RAST (Meyer et al., 2008), we recommend first registering the project in GOLD and providing appropriate metadata to facilitate the goal of comprehensive metadata dissemination. To this end, the JGI registers metagenome projects upon initiation, and we encourage other investigators to do the same. Ultimately this will increase the power of metagenomics by enabling meaningful comparisons.

Table 1. A sampling of the proposed 5-tiered metagenome classification schema

Ecosystem	Ecosystem-category	Ecosystem type	Ecosystem subtype	Specific Ecosystem
Environmental	Air	Indoor Air	Unclassified	Unclassified
Environmental	Aquatic	Freshwater	Lentic	Limnetic zone
Environmental	Aquatic	Freshwater	Groundwater	Cave water
Environmental	Aquatic	Freshwater	Drinking water	Filters
Environmental	Aquatic	Marine	Intertidal zone	Estuary
Environmental	Aquatic	Marine	Hydrothermal vents	Black smokers
Environmental	Aquatic	Non-marine Saline and Alkaline	Saline	Athalassic
Environmental	Aquatic	Thermal springs	Near-boiling (>90C)	Acidic
Environmental	Terrestrial	Surface soil	Neutral	Clay
Environmental	Terrestrial	Surface soil	Neutral	Sand
Environmental	Terrestrial	Surface soil	Acid	Silt
Host-associated	Mammals	Digestive system	Large intestine	Fecal
Host-associated	Birds	Digestive system	Crop	Lumen
Host-associated	Human	Reproductive system	Female	Vagina
Host-associated	Fish	Respiratory system	Gills	Filaments
Host-associated	Arthropoda	Digestive system	Hindgut	P1 segment
Host-associated	Annelida	Integument	Subcuticular space	Extracellular symbionts
Host-associated	Microbial	Archaea	Viriome	Unclassified
Engineered	Food production	Dairy products	Unclassified	Unclassified
Engineered	Bioremediation	Persistent organic pollutants (POP)	Polycyclic aromatic hydrocarbons	Bioreactor
Engineered	Solid waste	Landfill	Unclassified	Unclassified
Engineered	Solid waste	Composting	Grass	Bioreactor
Engineered	Wastewater	Nutrient removal	Dissolved organics (aerobic)	Activated sludge
Engineered	Modeled	Simulated communities (sequence read mixture)	Sanger	Sanger

Figure 1: Five-tiered hierarchical metagenome classification schema collapsed into groups at level 3. The size of terminal nodes reflects the number of projects in GOLD for each grouping. Branches that do not extend to the outer edge indicate categories for which there are no current metagenome projects in GOLD (e.g. amphibia).

References

Garrity, G.M., Field, D., Kyrpides, N., Hirschman, L., Sansone, S.A., Angiuoli, S. et al. (2008) Toward a standards-compliant genomic and metagenomic publication record. *Omics* 12: 157-160.

Lauber, C.L., Hamady, M., Knight, R., and Fierer, N. (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* 75: 5111-5120.

Liolios, K., Chen, I.M., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V.M., and Kyrpides, N.C. (2009) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 38: D346-354.

Lozupone, C.A., and Knight, R. (2007) Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A* 104: 11436-11440.

Markowitz, V.M., Ivanova, N.N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D. et al. (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* 36: D534-538.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M. et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386.