JOHNS HOPKINS
UNIVERSITY

Zanvyl Krieger School of Arts and Sciences
Advanced Academic Programs

410.734.81 and 410.734.82
Practical Introduction to Metagenomics

Topic: Reference datasets and resources

Instructor: Joshua Orvis

Introduction

In this week's lesson we cover reference datasets and general metagenomics resources. These are often linked, as we'll see, since many of the web-based resources also either host reference sets or offer tools to help generate them.

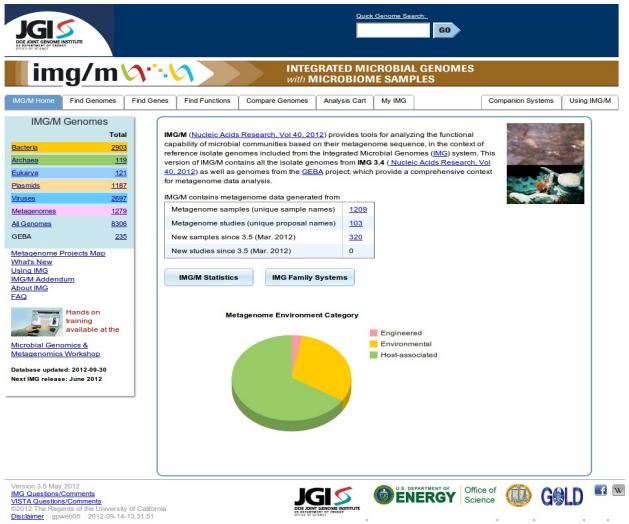
Two resources we focus on are also described in your assigned readings for this week:

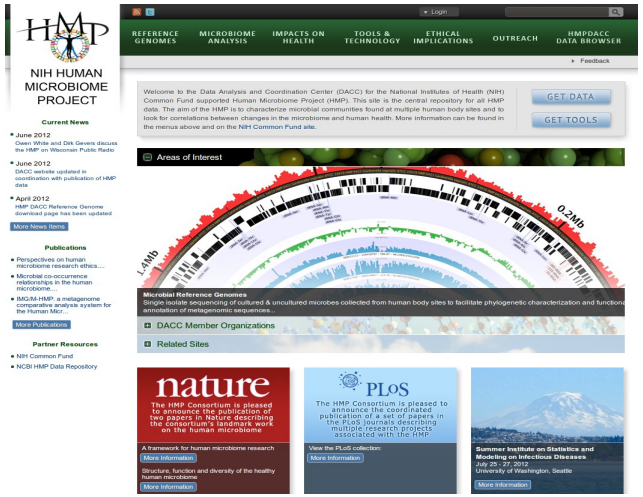
- Genomes OnLine Database (GOLD)
- Human Microbiome Project's Reference Genome Catalog

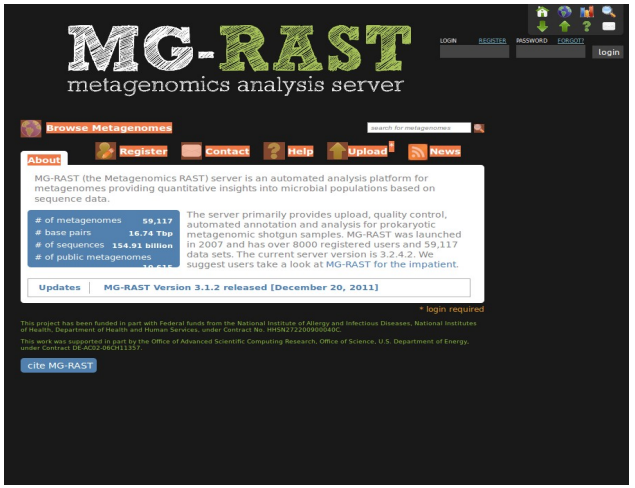
We'll also briefly cover a few other resources this week, some of which we go deeper into later:

- MG-RAST
- IMG/M
- CAMERA
- METAREP

← Metagenomics resources apparently have a CAPS LOCK problem.







Meta

You won't get very far in reading any of these papers or site descriptions without hearing the word **metadata**, so we'll start with a quick explanation of it and a few exampes.

P.S. I have to admit, when I first heard the term I was annoyed that the trendy kids in bioinformatics wanted a prefix to go along with their favorite suffix (-omics).

P.P.S.: Please, don't ever taunt me with the word 'metaomics'

Metadata

Without metadata, data themselves are often useless. So what are metadata?

Metadata is the descriptive information about data that explains the measured attributes, their names, units, precision, accuracy, data layout and ideally a great deal more. Most importantly, metadata includes the data lineage that describes how the data was measured, acquired or computed. (Gray, et al. 2005)

In metagenomic context, they start with the descriptions of sampling sites and habitats that provide the context for sequence information. Metadata are of great importance for metagenomic sequence data for two reasons:

- Only by fully describing the samples from which metagenomics sequences have been obtained can one have any possibility of replicating a study.
- Metadata are essential for the analysis of metagenomic sequence data which, without an environmental context, have no value.

"New science of metagenomics." 2007

GSC – MIGS & MIMS

Storing and sharing metadata isn't enough. It's critical that metadata be stored in a consistent way across all submited metagenomes in order for researches to parse, mine and generally compare samples.

The Genome Standards Consortium (GSC) was formed in 2005 with the goal of standardizing the description of genomes/microbiomes and helping to promote the exchange and integration of genomic metadata.

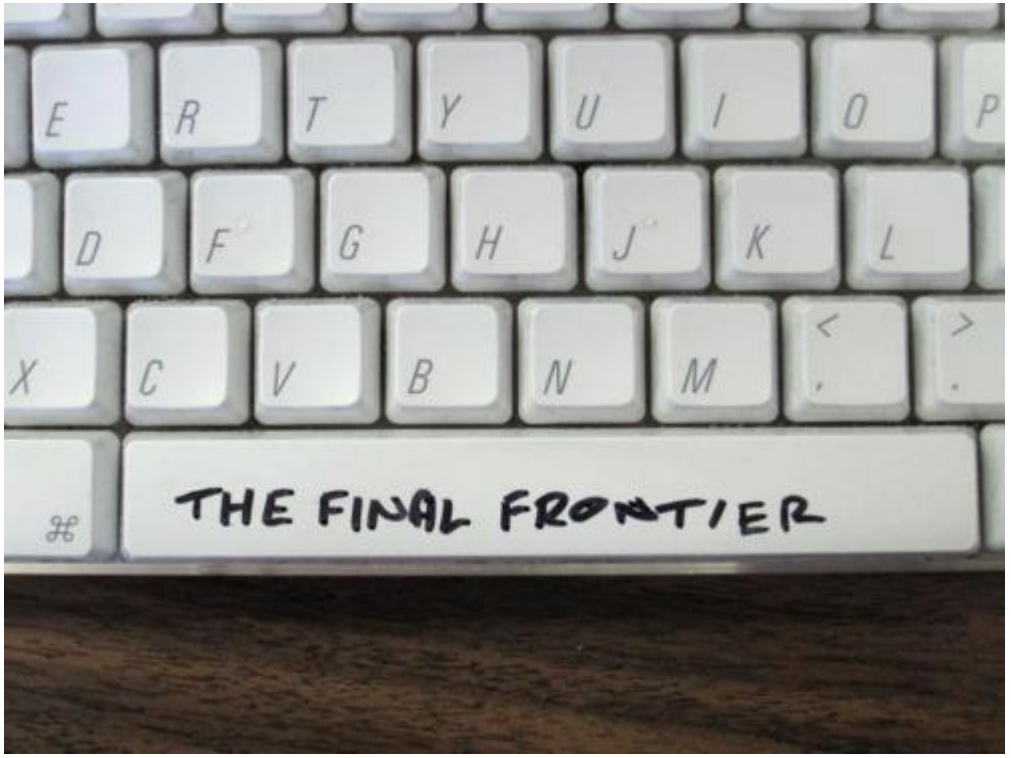
Most importantly, the GSC has the support of the major sequencing archives (Genbank, EMBL, and DDBJ) who agreed to support the recommendations of the GSC.

The GSC initially created the Minimum Information about a Genome Sequence (MIGS) specification, which was then extended in the Minimum Information about a Metagenome Sequence (MIMS) spec. Both of these contains standard formats for recording environmental and experimental data.

The official MIMS specification:
http://gensc.org/gc_wiki/index.php/MIGS/MIMS

Field, et al. 2008

+1 for nerdiness, -1 for ergonomics



GSC – MIMS in GBK records

The most common place you might find MIMS data is within Genbank entries. But the GBK sequence records consist of sequence data, organism info, and features located on that sequence. These are all based on a controlled list of organism modifiers. Unwilling to absorb the MIMS descriptors into their controlled list, how do they fit in the GBK records?

They hacked it, but attempted to keep some dignity about it by calling it a **Structured Comment**.

"The comment consists of tag-value pairs that are contained within START and END tags that function as delimiters for easy parsing. These comments can be incorporated from a tab-delimited table into submission files using either Sequin or tbl2asn."

```
COMMENT     ##MIENS-Data-START##
            collection_date    :: 2009-10-15
            collection_time    :: 08:35:00
            lat_lon            :: 55.01575 8.43785
            geodetic_datum     :: WGS84
            lat_long_details   :: 7 m recorded accuracy
            site               :: German Wadden Sea, Sylt ...
            depth              :: -0.09 m
            samp_size          :: 85.0 ml
            temperature        :: 10.0 degrees Celsius
            container          :: 100 ml glass bottle
            environment        :: Temperate shelf and sea biome
                                  [ENVO:00000895], coastal water
                                  body
                                  [ENVO:02000049], coastal water
                                  [ENVO:00002150]
            alt_elev           :: 0 m
            country            :: Germany
            investigation_type :: miens-survey
            project_name       :: Marine Microbiology (MarMic)
                                  class 2013 field excursion to
                                  Sylt, 2009
            sequencing_meth    :: Sanger
            target_gene        :: 16S rRNA
            MetaBar_barcode    :: 1000009000015
            ##MIENS-Data-END##
```

http://www.ncbi.nlm.nih.gov/genbank/structuredcomment#MIGS (testy comments are my own)

GSC - Ontologies

We've seen so far that the GSC proposed what information about metaganomic sequences to store and how they are stored in sequence entries like Genbank. Wisely (I believe), they also chose to use controlled vocabularies (ontologies) to standardize the set of values possible in some of the MIGS/MIMS value fields.

We'll discuss ontologies in greater depth in our lesson on functional annotation. For now, From the Sequence Ontology site:

"The Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequence. SO includes different kinds of features which can be located on the sequence. Biological features are those which are defined by their disposition to be involved in a biological process. Examples are binding_site and exon."

This means any value can have a rigid definition and be a part of a functional hierarchy. Given that there isn't even agreement on the formal definition of a 'gene', this is critical when systematically mining these datafiles.

We'll cover this more in a later lesson, but you can see the URL below if you're desperate for more information now.

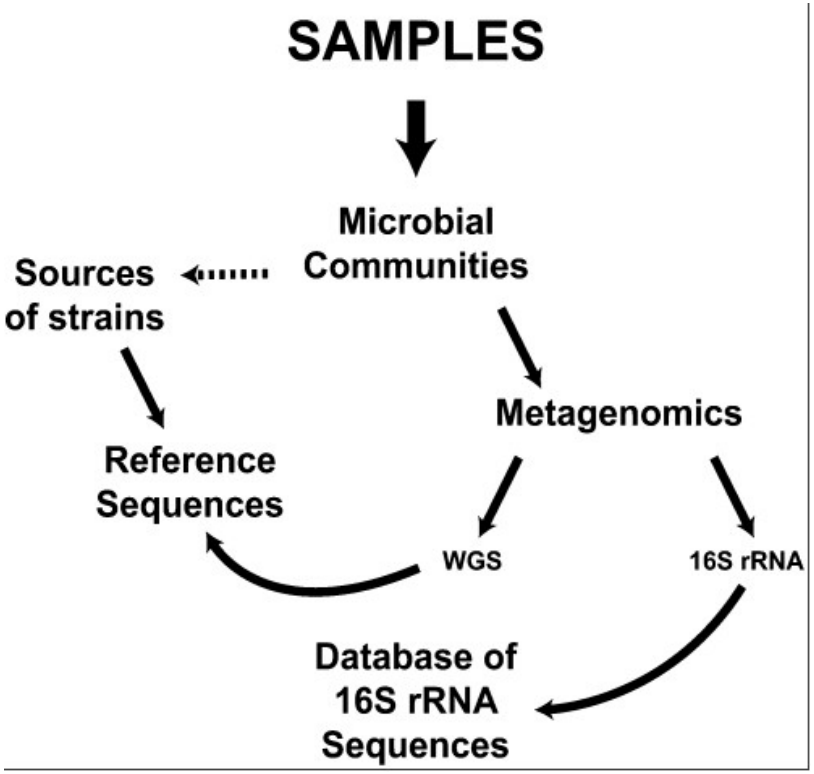
http://www.sequenceontology.org/

Reference genomes

It is quite common, when exploring something new and unusual, to make comparisons with past experience and reference. In the sequencing world, when we have a pool of sequence from an unknown community we'll start by comparing these sequences with those we've previously characterized.

It should be no surprise then that one of the first steps in the planning of a metagenomics experiment often includes reference genome data sets. These can be amassed by any combination of mining existing metadata, curation of previously-sequenced individual genomes suggested by experts in your habitat of interest, or even targeted isolate sampling and sequencing.

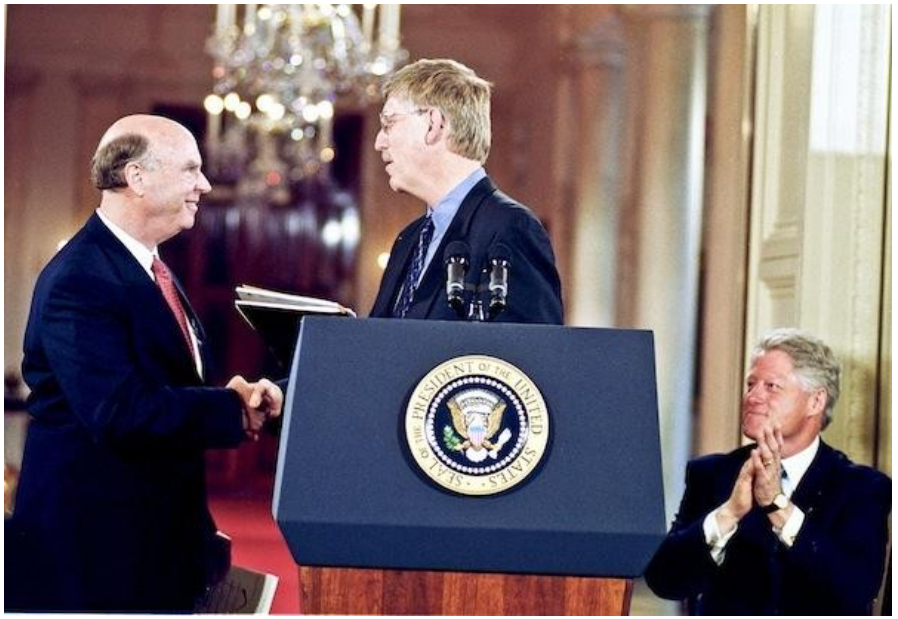
In the Human Microbiome Project, for example, an initial set of 800 individual genomes were chosen for sampling and sequencing. These were used in addition to the already-published public genes and served as the anchor point for many different types of analysis, from pan-genome studies, fragment recruitment, diversity measurement within genera, etc.



A scheme showing the role of references in a microbiome study. The microbial communities in the sample are analyzed using metagenomic sequencing approaches. 16S rRNA sequences are compared to a database of 16S sequences, while WGS sequences are compared to existing reference strains. (NIH HMP WG, et al. 2009)

The Human Microbiome Project (HMP)

In the summer of 2000 President Clinton gathered the (feuding) Francis Collins and J. Craig Venter at the White House to announce the joint draft completion of the Human Genome Project.



Not even a year after this resounding achievement, an opinion paper was published in *Science* calling for us to "count the microbes, too." (Davies, PMID: 11269298)

Given that our bodies carry 10x as many microbial cells than human ones, and 100x as many genes, it's fair to consider out microbial makeup our "other genome."

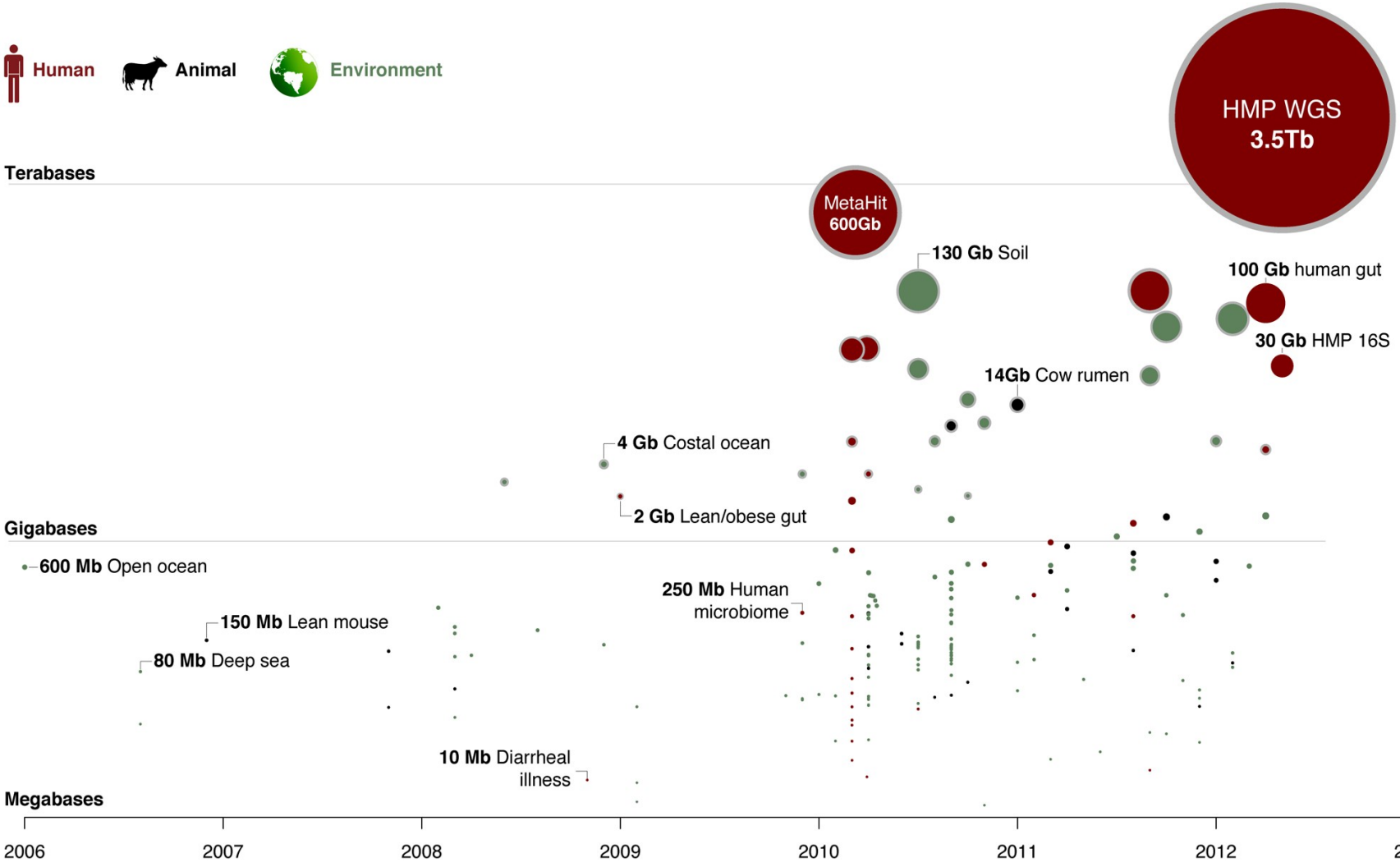
In 2007 the time was right to invest in a concerted study of the microbial communities associated with the human body and the metabolic capabilities they provide and NIH launched the $115-million Human Microbiome Project.

Five years later the initial results were published in an array of papers in *Nature* and *PLoS One.*

doi:10.1371/journal.pbio.1001377

Scale of the HMP

Timeline of the scale of microbial community studies: each circle represents a high-throughput sequence-based 16S or shotgun metagenomic bioproject in NCBI (May 2012), indicating the amount of sequence data produced for each project (circle area and y-coordinate) at the time of publication/registration (x-coordinate). The 'SRS' samples we've used in this course so far are from the largest of these circles, the HMP Whole-Genome Shotgun (WGS)



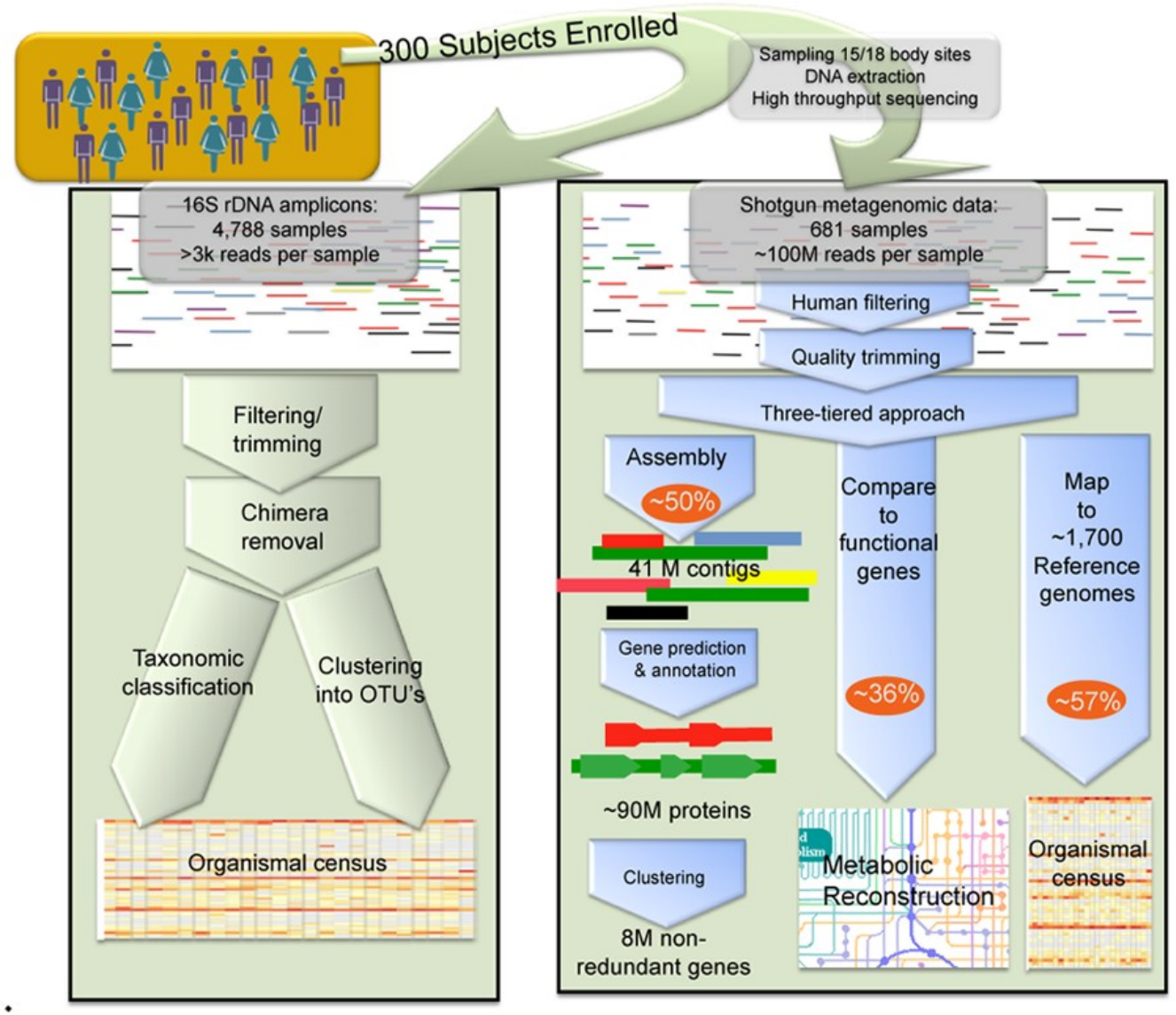
HMP analysis model

This diagram illustrates the general analysis plan of the HMP.

300 subjects were enrolled and 681 samples were sequenced, filtered to remove human data, quality trimmed, then processed along three different analysis paths in parallel.

The figure on the left illustrates these analysis paths at a very high-level.

The HMP DACC also provides an interactive data flow diagram which gives greater detail and immediate access to data at any stage of the flow.



doi:10.1371/journal.pbio.1001377

A fix for open-toed shoes in the lab

HMP reference genomes

"A catalog of reference genomes from the human microbiome."
Human Microbiome Jumpstart Reference Strains Consortium
Science. 2010 May 21;328(5981):994-9.
PMID: 20489017

One of the assigned readings for this week, this paper provides an initial report of the reference genomes isolated and sequenced for the HMP.

In it, different analyses are performed using the reference genomes to evaluate their utility when compared with the bacterial genomes that were already present in GenBank.

At the time ony 178 ref genomes had been completed, yet still read recruitment efforts showed that 20-40% of the reads were recruited only because of the presence of the HMP genomes.

RESEARCH ARTICLE

A Catalog of Reference Genomes from the Human Microbiome

The Human Microbiome Jumpstart Reference Strains Consortium*†

The human microbiome refers to the community of microorganisms, including prokaryotes, viruses, and microbial eukaryotes, that populate the human body. The National Institutes of Health launched an initiative that focuses on describing the diversity of microbial species that are associated with health and disease. The first phase of this initiative includes the sequencing of hundreds of microbial reference genomes, coupled to metagenomic sequencing from multiple body sites. Here we present results from an initial reference genome sequencing of 178 microbial genomes. From 547,968 predicted polypeptides that correspond to the gene complement of these strains, previously unidentified ("novel") polypeptides that had both unmasked sequence length greater than 100 amino acids and no BLASTP match to any nonreference entry in the nonredundant subset were defined. This analysis resulted in a set of 30,867 polypeptides, of which 29,987 (~97%) were unique. In addition, this set of microbial genomes allows for ~40% of random sequences from the microbiome of the gastrointestinal tract to be associated with organisms based on the match criteria used. Insights into pan-genome analysis suggest that we are still far from saturating microbial species genetic data sets. In addition, the associated metrics and standards used by our group for quality assurance are presented.

The human microbiome is the enormous community of microorganisms occupying the habitats of the human body. Different microbial communities are found in each of the varied environments of human anatomy. The aggregate microbial gene tally surpasses that of the human genome by orders of magnitude. Understanding the relationship of the microbial content to human health and disease is one of the primary goals of human microbiome studies. Determining the structure and function of any microbial community requires a detailed definition of the genomes that it encompasses and the prediction and annotation of their genes.

In 2007, the National Institutes of Health (NIH) initiated the Human Microbiome Project (HMP) as one of its Roadmap initiatives (*1*) to provide resources and build the research infrastructure. One component of the HMP is the production of reference genome sequences for at least 900 bacteria from the human microbiome, which will catalog the microbial genome sequences from the human body and aid researchers conducting human metagenomic sequencing in assigning species to sequences in their metagenomic data sets.

The HMP catalog of reference sequences is being produced by the NIH HMP Jumpstart Consortium of four genome centers: the Baylor College of Medicine Human Genome Sequencing Center, the Broad Institute, the J. Craig Venter Institute, and the Genome Center at Washington University. The challenges for the Jumpstart Consortium include selecting strains to sequence and identifying sources, creating standards for sequencing and annotation to ensure consistency and quality, and the rapid release of information to the community.

Reference genome progress. To date, 356 genomes, including 117 genomes at various stages of upgrading, have been produced by the Jumpstart Consortium and released into public databases. At the time of manuscript preparation, 178 had been completely annotated and are presented in the analysis here. The process for the selection of these strains is described in (*2*). The strains sequenced to date are distributed among body sites as follows: gastrointestinal tract (151), oral cavity (28), urogenital/vaginal tract (33), skin (18), and respiratory tract (8). They also include one isolate from blood (*3*). These are the five major body sites targeted by the HMP.

The broad phylogenetic distribution of the sequenced strains is presented in Fig. 1, which represents a 16*S* ribosomal RNA (rRNA) overlay of HMP-sequenced genomes on 16*S* rRNA sequences from cultured organisms with sequenced genomes (*4*). HMP-sequenced genomes represent two kingdoms (Bacteria and Archaea), nine phyla, 18 classes, and 24 orders. Additional rRNA overlay figures broken down by individual body sites are available in (*5*).

To obtain high-quality draft genomes and a meaningful gene list, minimum standards were defined for the assembly and annotation of draft genomes. Three reference bacterial genome assemblies were evaluated for efficacy of gene predictions and genome completeness. Based on the analysis, metrics for assembly characteristics and annotation characteristics were defined [for more details, see (*2*)]. The quality of HMP genome assemblies is summarized in Table 1 and exceeds the Jumpstart Consortium standards described in (*2*), with the exception of some genomes produced before the standards were in place.

Genome improvement. As described in (*2*), there are justifications for upgrading these high-quality draft assemblies. The Jumpstart Consortium has completed initial improvement work on 26 bacterial genomes that differed significantly with respect to GC content and assembly metrics to explore the effort required and resulting benefits (Fig. 2). The average contig N50 increased 3.63-fold, from 109 kb at draft to 396 kb after improvement. *Bacteroides pectinophilus* displays substantial improvement in N50, from 163 kb in the draft sequence to 862 kb after improvement. *Lactobacillus reuteri* illustrates the opposite extreme, with improvement leading to a smaller contig N50 change, 56 kb to 72 kb. As more genomes improve and some graduate to higher levels of improvement, the assembly state or group of states most useful to the HMP scientific goals will be evaluated.

Pan-genome analysis. A bacterial species' pan-genome can be described as the sum of the core genes shared among all sequenced members of the species and the dispensable genes, or those genes unique to one or more strains studied. To start addressing questions about pan-genomes, we identified all species within our sequenced reference genome catalog for which there was more than one sequenced and annotated genome. Of the nine species identified, four of them have five or more annotated genomes that were generated either by the HMP or by external projects publicly available at the National Center for Biotechnology Information (NCBI); five genomes is the minimum number for which a curve can reliably be fit to pan-genome data. These are *L. reuteri*, *Bifidobacterium longum*, *Enterococcus faecalis*, and *Staphylococcus aureus*. The genomic data used for the analysis consisted of both complete and draft genomes, the only requirement being that >90% of the genome be represented in the available annotated contigs or scaffolds.

Pan-genome curves (*6*) of the gastrointestinal tract isolates *L. reuteri*, *B. longum*, and *E. faecalis* (figs. S3 to S5) are consistent with an open pan-genome model, suggesting that more genome sequencing needs to be undertaken to characterize the actual makeup of the species as a whole. Preliminary results suggest core genome sizes of approximately 1430 genes, 1800 genes, and 1600 genes for *B. longum*, *E. faecalis*, and *L. reuteri*, respectively. Based on the current core gene plots, *L. reuteri* (fig. S3) appears to be approaching a closed pan-genome model, with newly sequenced strains contributing very small numbers of new genes to the pan-genome; however, we see an interesting community substructure within this species. Our current *L. reuteri* pan-genome analysis of seven isolates suggests that four of the

*All authors with their affiliations and contributions are listed at the end of this paper.
†To whom correspondence should be addressed. E-mail: kenelson@jcvi.org

GOLD database

"The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata."
Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC.
Nucleic Acids Res. 2012 Jan;40:D571-9.
PMID: 22135293

The second of two paper assignments for this week, the GOLD project is an invaluable source for tracking genome and metagenomic projects and creating your own reference datasets.

Where several other resources store collections of reference genomes, GOLD is the only one whose mission includes gathering metadata and tracking sequencing projects **before** they are completed.

GOLD was started in 1997 and this report provides an update on status and future directions as of September, 2011.

It is an authoritative source on building and tracking reference data sets.

Published online 1 December 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D571–D579***
doi:10.1093/nar/gkr1100

The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata

Ioanna Pagani[1], Konstantinos Liolios[1,*], Jakob Jansson[1], I-Min A. Chen[2], Tatyana Smirnova[3], Bahador Nosrat[1], Victor M. Markowitz[2] and Nikos C. Kyrpides[1,*]

[1]Department of Energy Joint Genome Institute, Microbial Genomics and Metagenomics Program, 2800 Mitchell Drive, Walnut Creek, [2]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley and [3]Department of Energy Joint Genome Institute, Genome Portals Group, 2800 Mitchell Drive, Walnut Creek, CA, USA

Received September 28, 2011; Revised November 2, 2011; Accepted November 3, 2011

ABSTRACT

The Genomes OnLine Database (GOLD, http://www.genomesonline.org/) is a comprehensive resource for centralized monitoring of genome and metagenome projects worldwide. Both complete and ongoing projects, along with their associated metadata, can be accessed in GOLD through precomputed tables and a search page. As of September 2011, GOLD, now on version 4.0, contains information for 11 472 sequencing projects, of which 2907 have been completed and their sequence data has been deposited in a public repository. Out of these complete projects, 1918 are finished and 989 are permanent drafts. Moreover, GOLD contains information for 340 metagenome studies associated with 1927 metagenome samples. GOLD continues to expand, moving toward the goal of providing the most comprehensive repository of metadata information related to the projects and their organisms/environments in accordance with the Minimum Information about any (x) Sequence specification and beyond.

INTRODUCTION

The Genomes OnLine Database (GOLD) provides a centralized resource for the continuous monitoring of genome and metagenome sequencing projects worldwide, uniquely integrated with their associated metadata and is currently in its fourth version since its launching in 1997 (1–5). The number of registered sequencing projects has almost doubled since the publication of the previous report 2 years ago (5). As of September 2011, 11 472 projects have been registered, versus 5843 in September 2009 (5), 2905 in September 2007 (4) and 1575 in September 2005 (3) (Figure 1A). This rapid growth is mainly attributed to decreasing costs due to advances in sequencing technologies, instigating several large-scale microbial genome sequencing initiatives, such as the Human Microbiome Project (HMP; http://www.hmpdacc.org/) (6) and the Genomic Encyclopedia of Bacteria and Archaea (GEBA; http://www.jgi.doe.gov/programs/GEBA/) (7). During this period, GOLD has also expanded its scope beyond standard genomic and metagenomic projects to now encompass data from the growing number of resequencing, transcriptome, metatranscriptome and single cell sequencing projects.

Among the most important developments of the database during the last 2 years are those coupled to the growth of the metadata and metagenome projects. These include the implementation of GOLD-specific controlled vocabularies (CVs) for the representation of the associated data, in coordination with the Genomics Standards Consortium (GSC) (8) complying with its recommendations for the Minimum Information about any (x) Sequence (MIxS) specifications (9). Additionally, GOLD has implemented the canonical metagenome naming and standardized classification for all metagenome projects, as it has been proposed in 2010 (10). Finally, GOLD has placed emphasis on the rapidly advancing field of metagenomics through (i) increasing the number of metadata fields associated with metagenomic samples,

*To whom correspondence should be addressed. Tel: +1 925 296 5718; Fax: +1 925 296 5666; Email: nckyrpides@lbl.gov
Correspondence may also be addressed to Konstantinos Liolios. Tel: +1 925 296 2582; Fax: +1 925 296 5666; Email: kliolios@lbl.gov

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Other resources

On the remaining slides you'll find short summaries of several other metagenomics resources you can try using as this class progresses. We'll cover some of them in greater detail in future lectures, but I wanted you to be aware of them now.

You'll find a lot of duplicated/copied features among the following resources but they each have their own unique parts that make them worth trying out.

Don't need to memorize their feature lists for anything – just get a feel for what their focus is, how well they are integrated with the community standards. Note how well each enables the creation of a reference data set (if at all) as well as the set of analysis tools they offer to enrich your collections.

Resource: CAMERA

Started in early 2006, the Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis is a data repository and bioinformatics tool resource.

The first released focused on the Global Ocean Survey data, but the resource has since expanded, with a stated goal to become “definitive repository for metagenomic data and metadata, focusing on enabling molecular microbial ecology.”

They have competition there.

Once registered and accepted, you can launch pipelines on their grid to do assembly, clustering, functional annotation, KEGG analysis and more.

http://camera.calit2.net

Resource: MG/RAST

This is a very popular annotation server that boasts good adoption numbers, has great documentation and an easy-to-use interface. You start by submitting either raw or assembled sequences and their associated metadata. The metadata must always be public but you can keep your primary data private. We'll cover their unique method of functional annotation using 'subsystems' in a later lesson.

Data summary:

# of Metagenomes: 59,186

Base pairs: 16.81 Tbp

Sequences: 155.55 billion

Public:

# of Metagenomes: 10,666

# of Projects: 270

Environments: 15

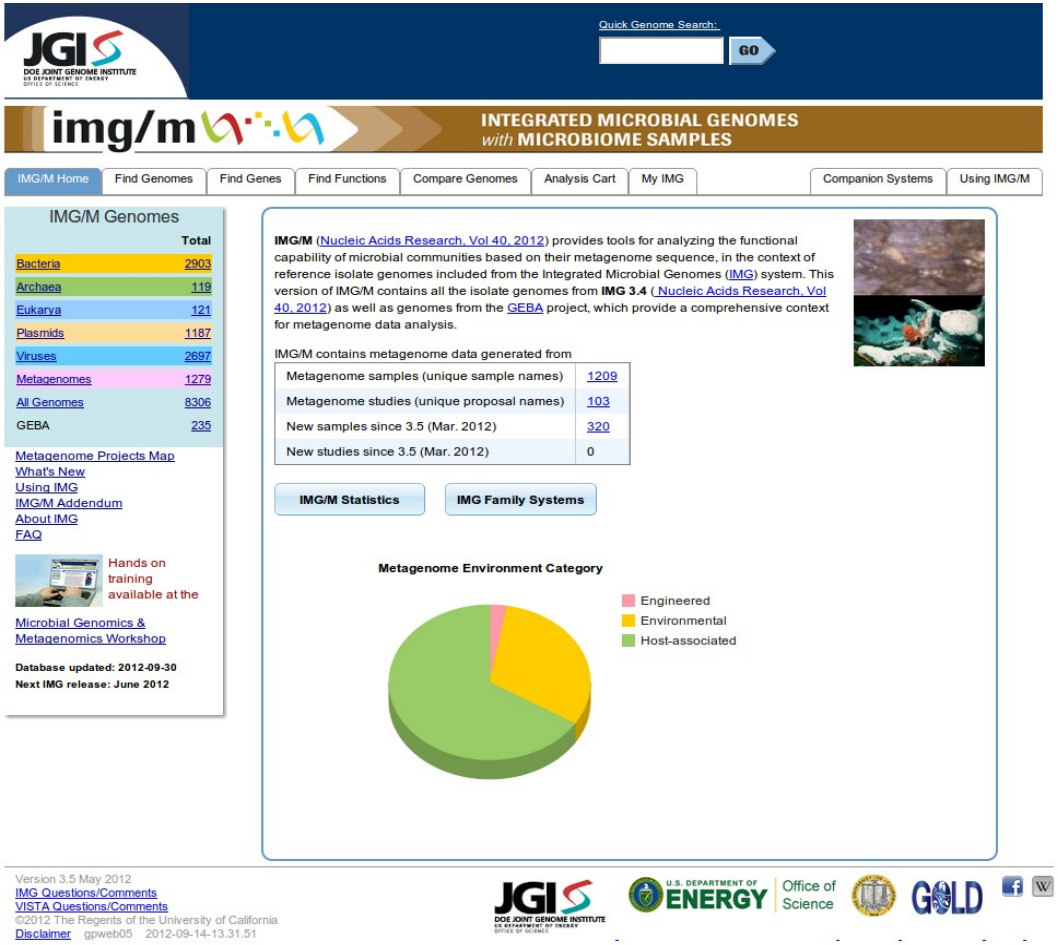
PI's: 87

Resource: IMG/M

JGI's IMG/M builds upon their Integrated Microbial Genomes (IMG) resource. IMG currently has around 8000 genomes from the 3 domains of life along with tools for interrogating and comparing these. IMG serves as a large reference collection in IMG/M, which provides tools for analyzing the functional capability of microbial communities in the context of a chosen set of ref genomes.

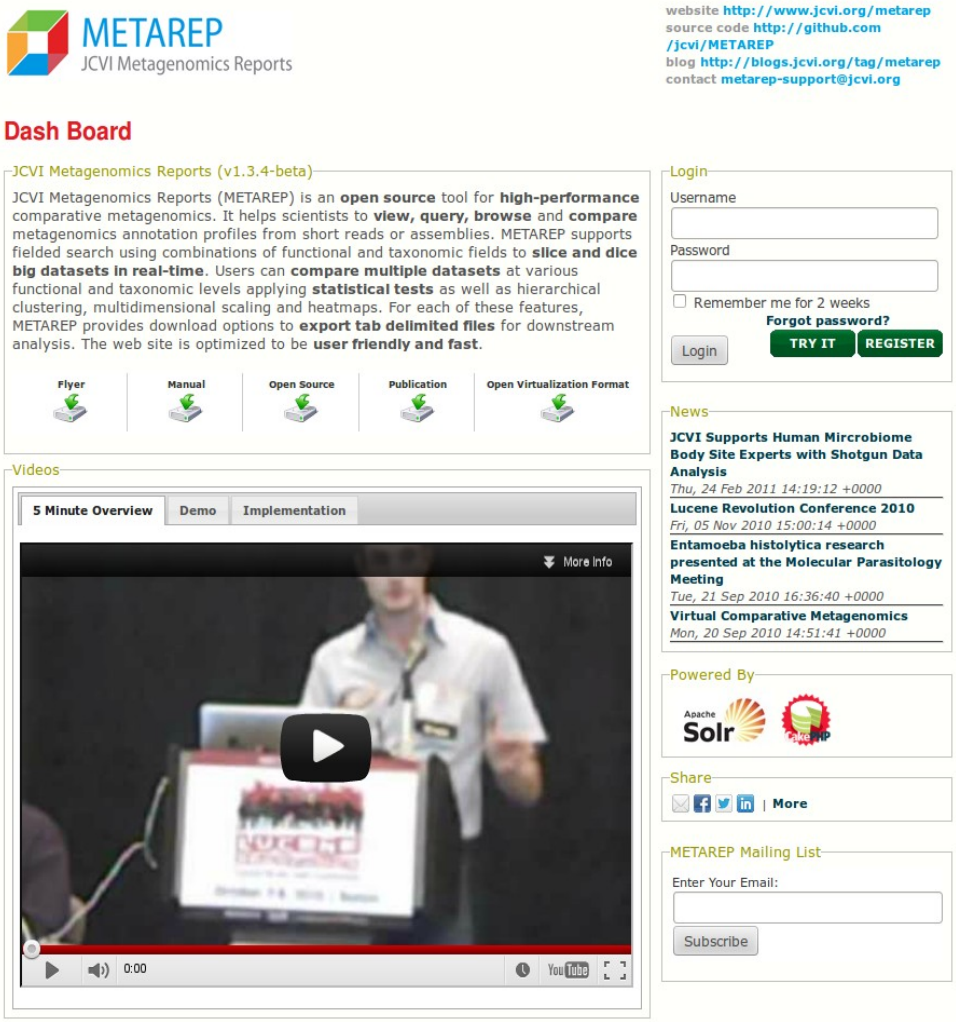
Whereas MG-RAST supports only prokaryotic sequences, IMG/M supports sequences from bacteria, archaea, eukaryotes and viruses.

You should at least try the site's Microbiome Projects Map – a Google Maps-driven browser of metagenomic sequences that are popular on resources like this. You can browse over 150 metagenomics projects and find the one nearest you or other parts of the world you're interested in.



Resource: METAREP

Like the CAMERA website, this JCVI-sourced resource requires registration, but is among the newest of these tools. It is primarily focused on comparative metagenomics, and supports analysis at the read or assembly levels.



It has a good UI and features include:

- High scalability
- Exports publication-ready graphics
- KEGG metabolic pathway analysis
- Multiple different points of entry:
  - Summaries by data type
  - SQL-like formal query syntax
  - NCBI taxonomy browser
  - GO browser
- Dataset comparison using plots and statistical tests.

Further reading

"The NIH human microbiome project"
NIH HMP working group, et al. 2009.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2792171

"Scientific data management in the coming decade"
Gray J, et al. 2005.
http://arxiv.org/pdf/cs/0502008.pdf

"The minimum information about a genome sequence (MIGS) specification"
Field D, et al. 2008
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2409278/