UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

HMPDACC
Human Microbiome Project
Data Analysis and Coordination Center

Anup Mahurkar

April 1st 2015

Outline

- Background
- Data types
- How do we store the data now (OSDF)
- Lessons learnt

BACKGROUND

Human **M**icrobiome **P**roject (HMP)

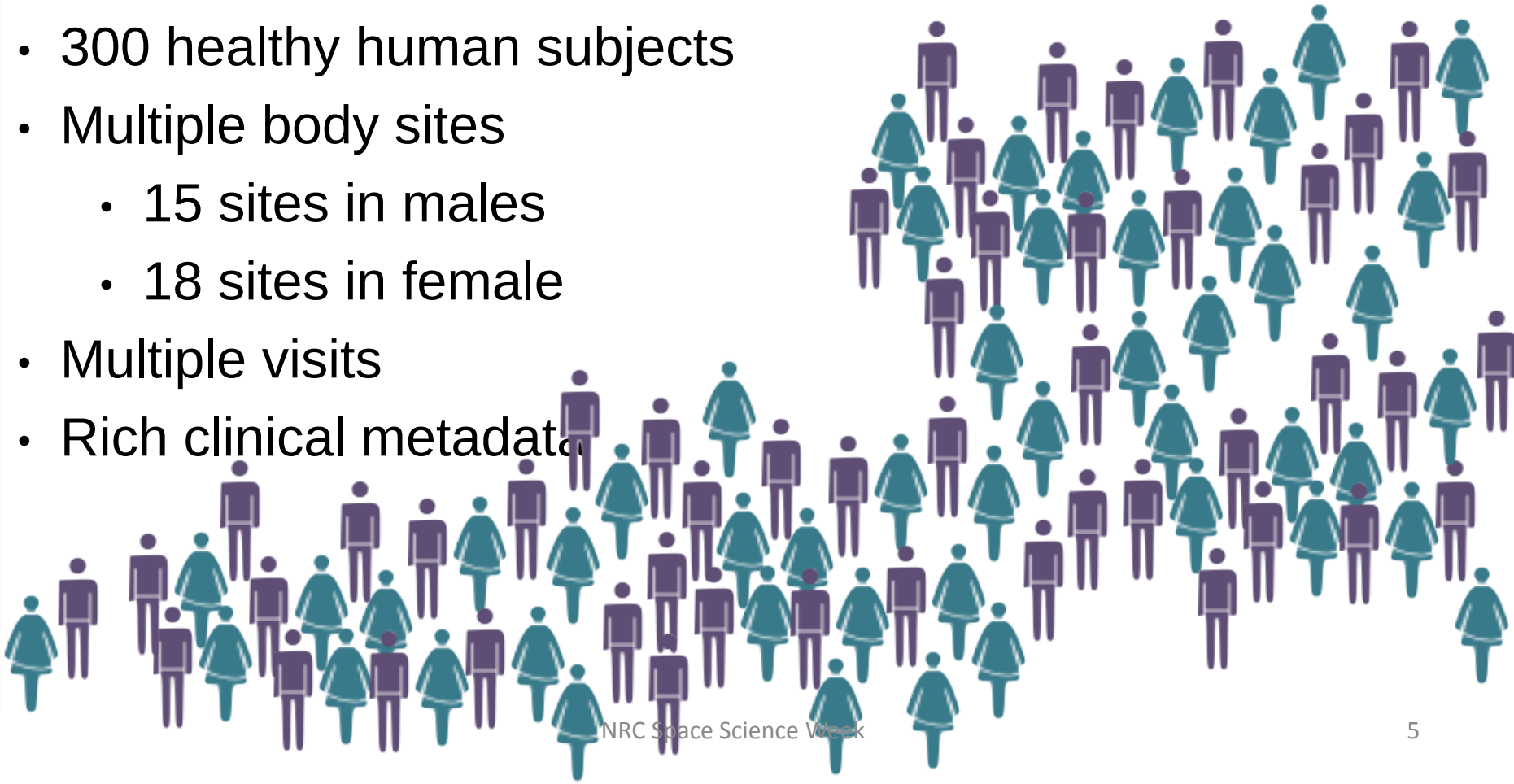
A comprehensive microbial survey

- Phase I (2007 – 2013)
  - Healthy Human Subjects
  - Demo Projects
  - Human Associated Reference Genomes
- Phase II (2014 – Ongoing)
  - Integrative Human Microbiome Project (iHMP)

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

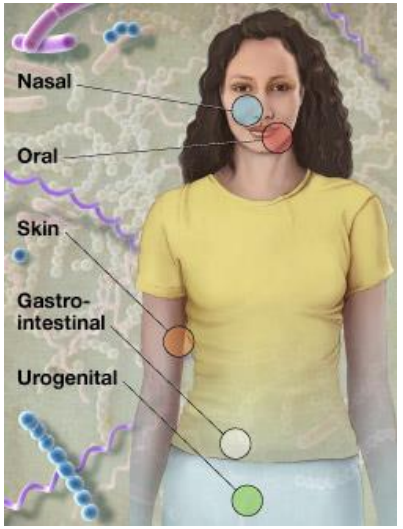
Healthy **H**uman **S**ubjects (HHS) Project

- ***What is a "normal" human microbiome?***
- 300 healthy human subjects
- Multiple body sites
  - 15 sites in males
  - 18 sites in female
- Multiple visits
- Rich clinical metadata



Healthy Human Subjects Project

- 300 healthy adults, 18-40
- 16S RNA + whole genome shotgun
- 5 sites/18 samples + blood
  - **Oral cavity:** saliva, tongue, palate, buccal mucosa, gingiva, tonsils, throat, teeth
  - **Skin:** ears, inner elbows
  - **Nasal cavity**
  - **Gut:** stool
  - **Vagina**
- Reference genomes (~1300)



Demo Projects

- Individual PI driven projects
- Comparing disease to healthy state

Data Generators

Baylor College of Medicine
Texas Medical Center, Houston, Texas

Washington University in St. Louis
School of Medicine

BROAD INSTITUTE

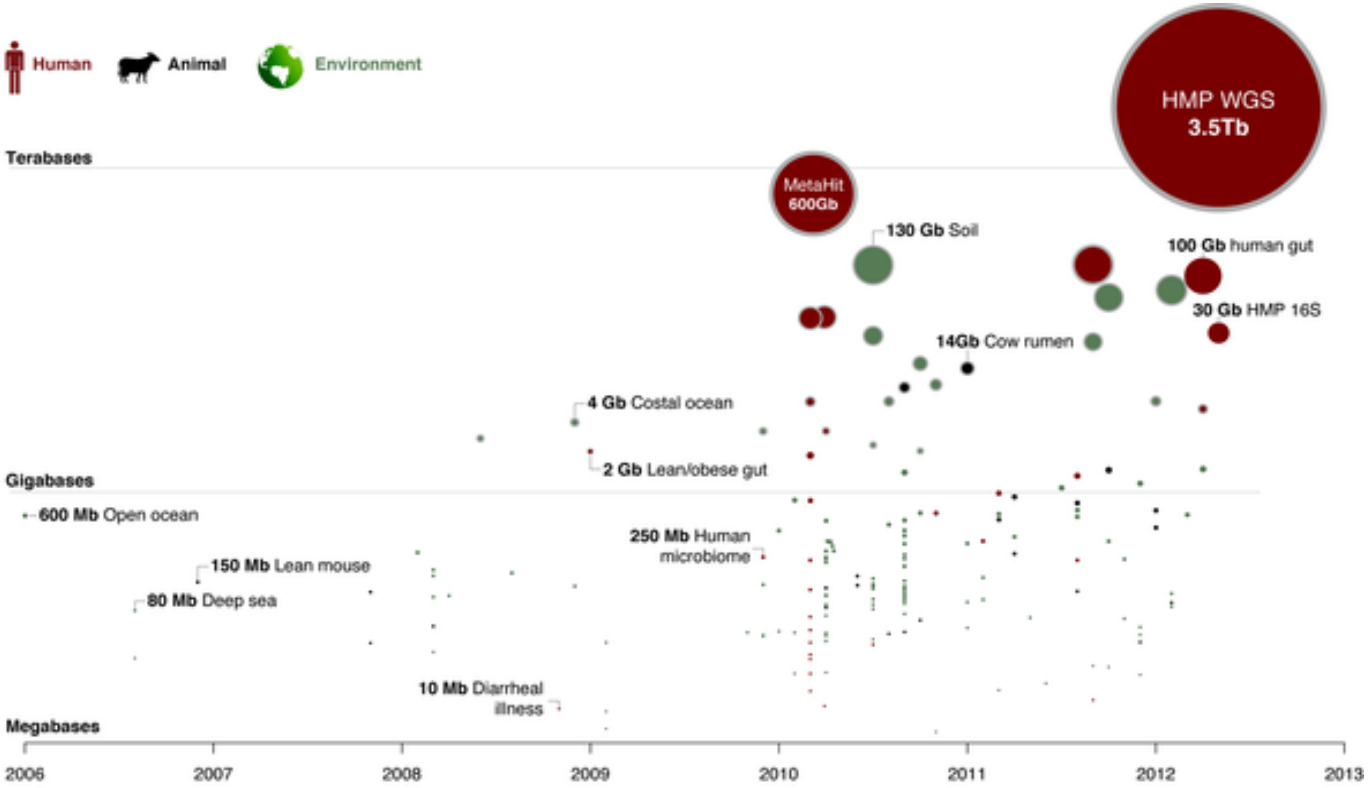
J. Craig Venter INSTITUTE

Data Repository

NCBI
National Center for Biotechnology Information

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Timeline of microbial community studies using high-throughput sequencing.

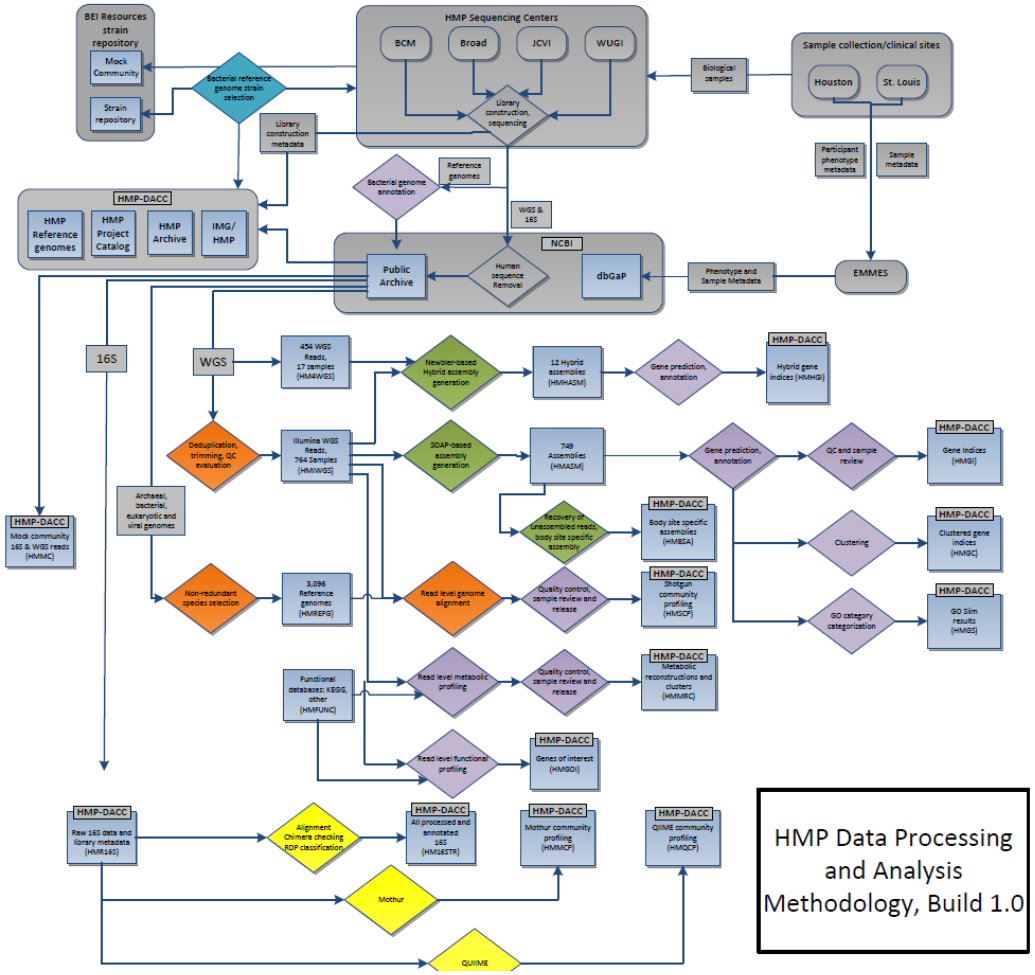


Gevers D, Knight R, Petrosino JF, Huang K, et al. (2012) The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. PLoS Biol 10(8): e1001377. doi:10.1371/journal.pbio.1001377
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001377


NIH HUMAN MICROBIOME PROJECT
REFERENCE GENOMES
MICROBIOME ANALYSIS
IMPACTS ON HEALTH
TOOLS & TECHNOLOGY
ETHICAL IMPLICATIONS
OUTREACH
HMPDACC DATA BROWSER
Feedback
Welcome to the Data Analysis and Coordination Center (DACC) for the National Institutes of Health (NIH) Common Fund supported Human Microbiome Project (HMP). This site is the central repository for all HMP data. The aim of the HMP is to characterize microbial communities found at multiple human body sites and to look for correlations between changes in the microbiome and human health. More information can be found in the menus above and on the NIH Common Fund site.
GET DATA
GET TOOLS
Current News
July 2013
Human Microbiome Science: Vision for the Future conference to be held in Bethesda, MD July 24-26
May 2013
Human Microbiome Consortium Virtual Meeting: Approaches in Microbiome Assembly
May 2013
Booth at ASM 2013 (#639)
More News Items
Publications
Colitis-induced Bone Loss is Gender Dependent and Associated with Incr...
Topographic diversity of fungal and bacterial communities in human ski...
Comparative metagenomic and rRNA microbial diversity characterization ...
More Publications
Partner Resources
NIH Common Fund
NCBI HMP Data Repository
Areas of Interest
hmpdacc.org
Human Microbial Sampling
16S RNA and whole metagenome sequencing of samples collected from 300 healthy human participants, to characterize complexity of microbial communities at individual body sites and to provide insights into functions performed by the human microbiome...
DACC Member Organizations
Related Sites
img/hmp m
CloVR

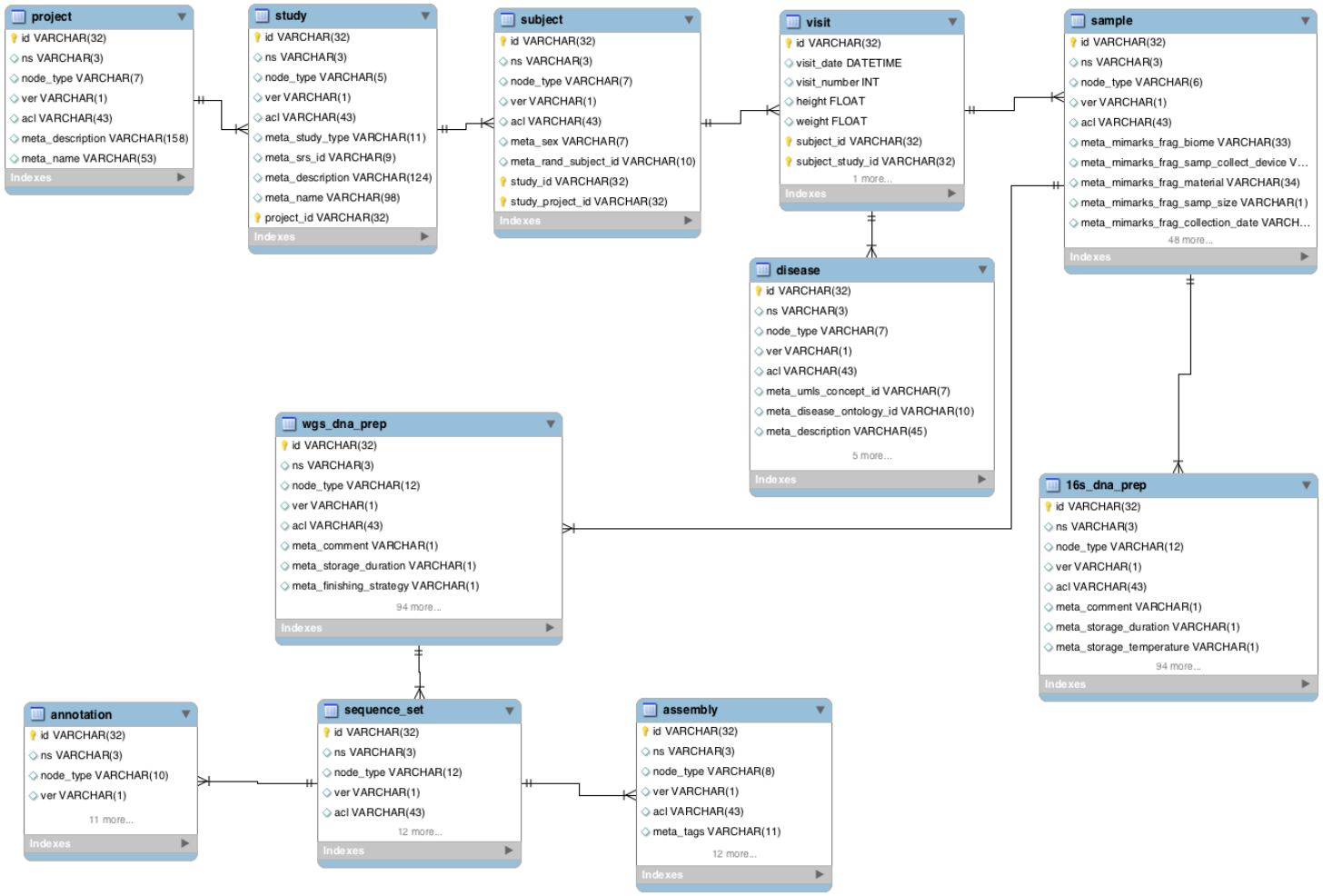
DATA TYPES

Data Processing Flow



UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Data Types



OSDF@DACC

- 800 Subjects
- 1,600 Samples
- 14,000 16S datasets
- 1,635 WGS datasets
- ~5,000 Derived datasets

How do we store the data?

OPEN SCIENCE DATA FRAMEWORK

Design Goals

- Manage large collections of data sets
- Decorate data sets with rich metadata
- Develop a framework that could be reused for other collaborative or multi-center projects
- Ease of development using a language agnostic API
- Scalable and cloud-enabled
- Support federated data

What is OSDF?

- Generic extensible framework for associating data with metadata
- Examples of data
  - Reference Databases, Sequenced Reads, Assemblies, Alignments, Annotation
- Examples of metadata
  - Sequencing platform, Library preparation, Sequencing strategy, Assembly method, Alignment method, Alignment reference, Project, Subject information

What is OSDF?

- Includes a mechanism for
  - Defining data model for elements
  - Reliance on ontologies, and controlled vocabularies
  - Defining relationships between elements to build a relationship graph
  - Generic RESTful API for accessing and placing data
  - Versioning and history
  - Access control

Technologies Used

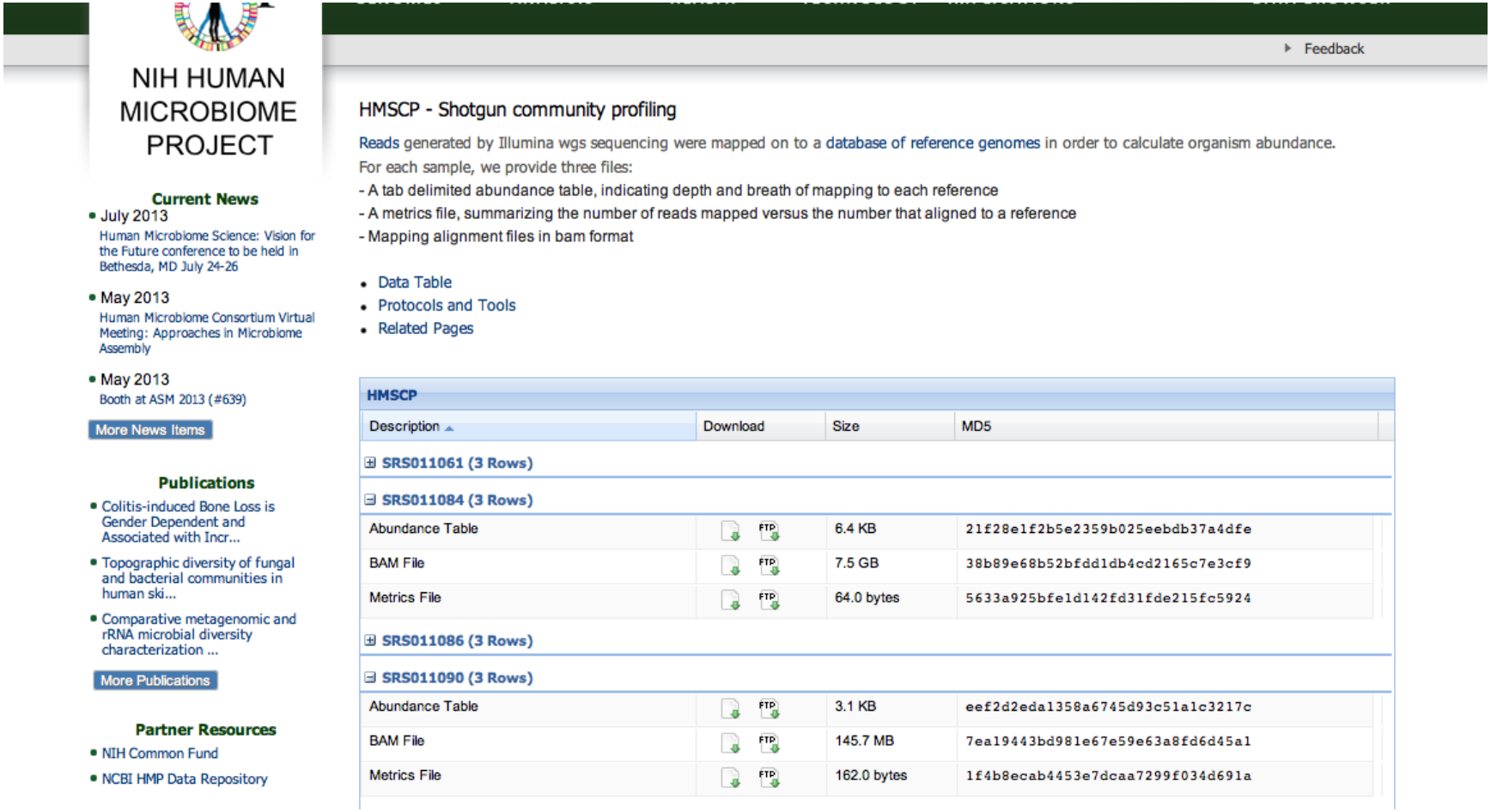
- JavaScript Object Notation (JSON) objects for modeling data elements
  - Lightweight data interchange format
  - JSON Schema for validation
- CouchDB for storing JSON objects
  - Document-oriented database
  - RESTful JSON API out of the box
- PostgreSQL for some queries
  - To drive advanced query builder

Technologies Used

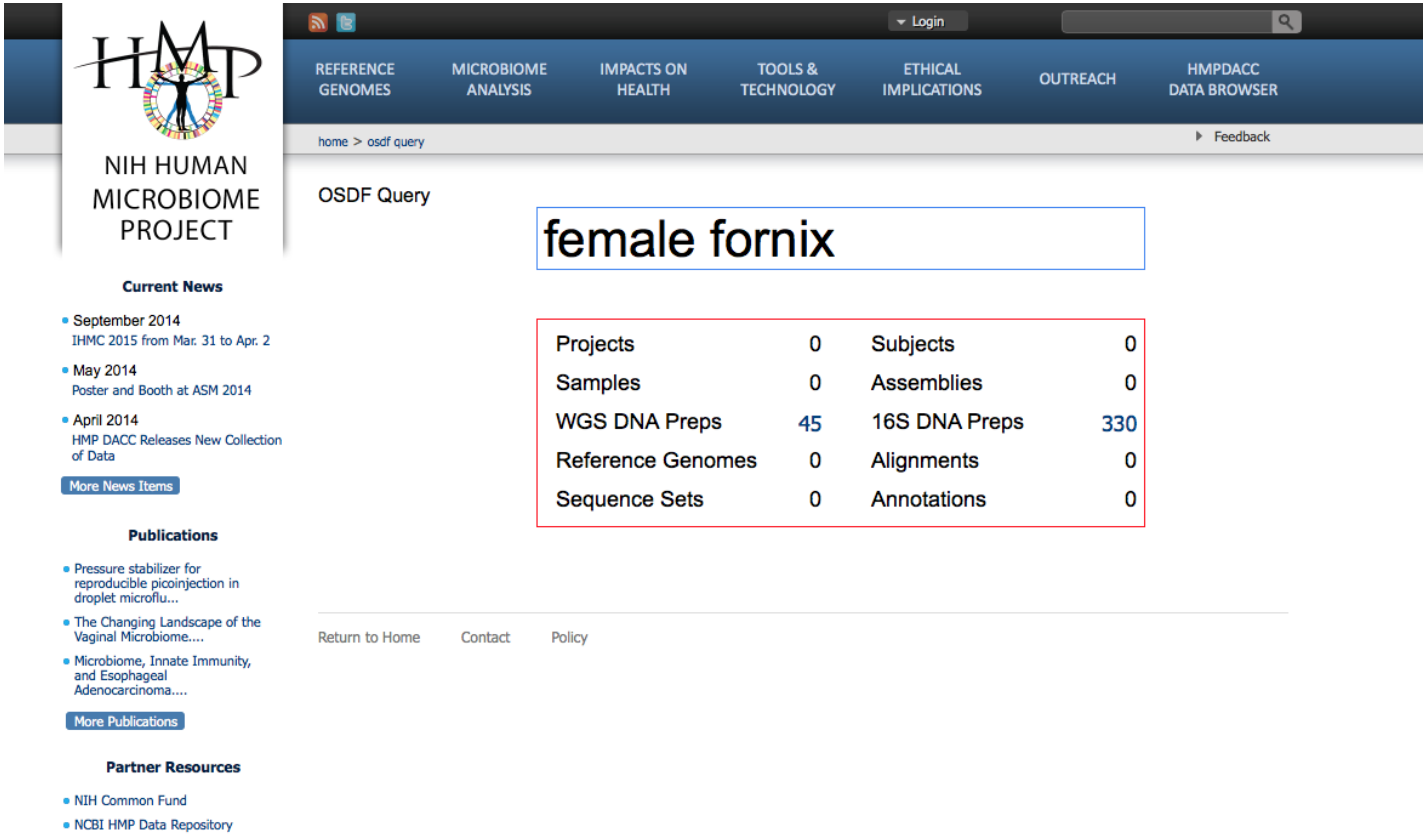
- ElasticSearch
  - Rapid indexing on all keys and attributes of JSON
  - Allows wild card, proximity, range, and Boolean operators
- Metadata modeled by CVs, ontologies, standards, and dictionaries
  - MIGS, MIMS, MIMARKS, GO, Relationship
- API implementation using node.js
  - Scalable server optimized for concurrency
  - Implement JSON validation using JSON Schema
- UI Implementation
  - ExtJS, GraphViz, jQuery, D3

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Current HMP DACC Site

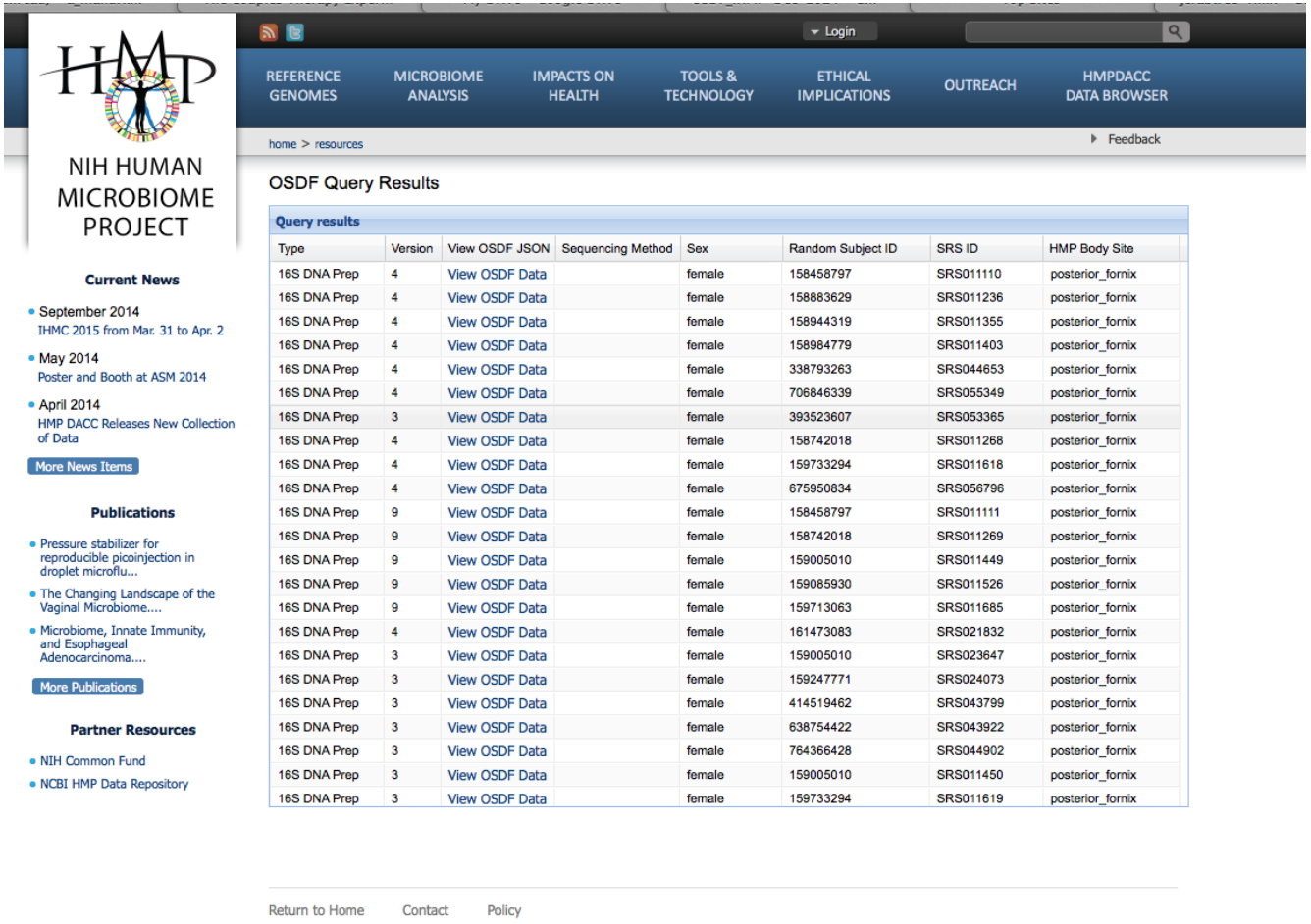


Data Discovery



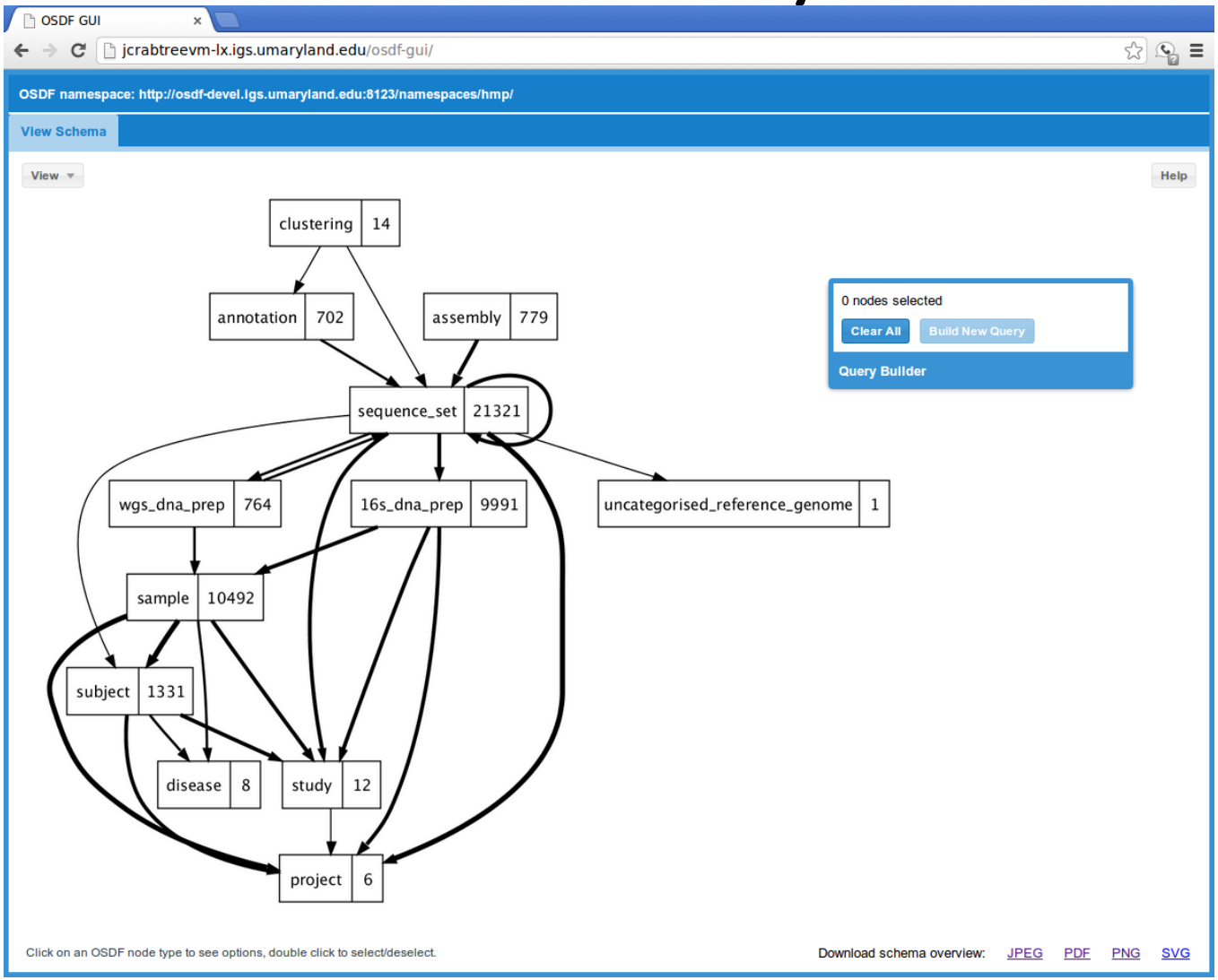
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Data Discovery



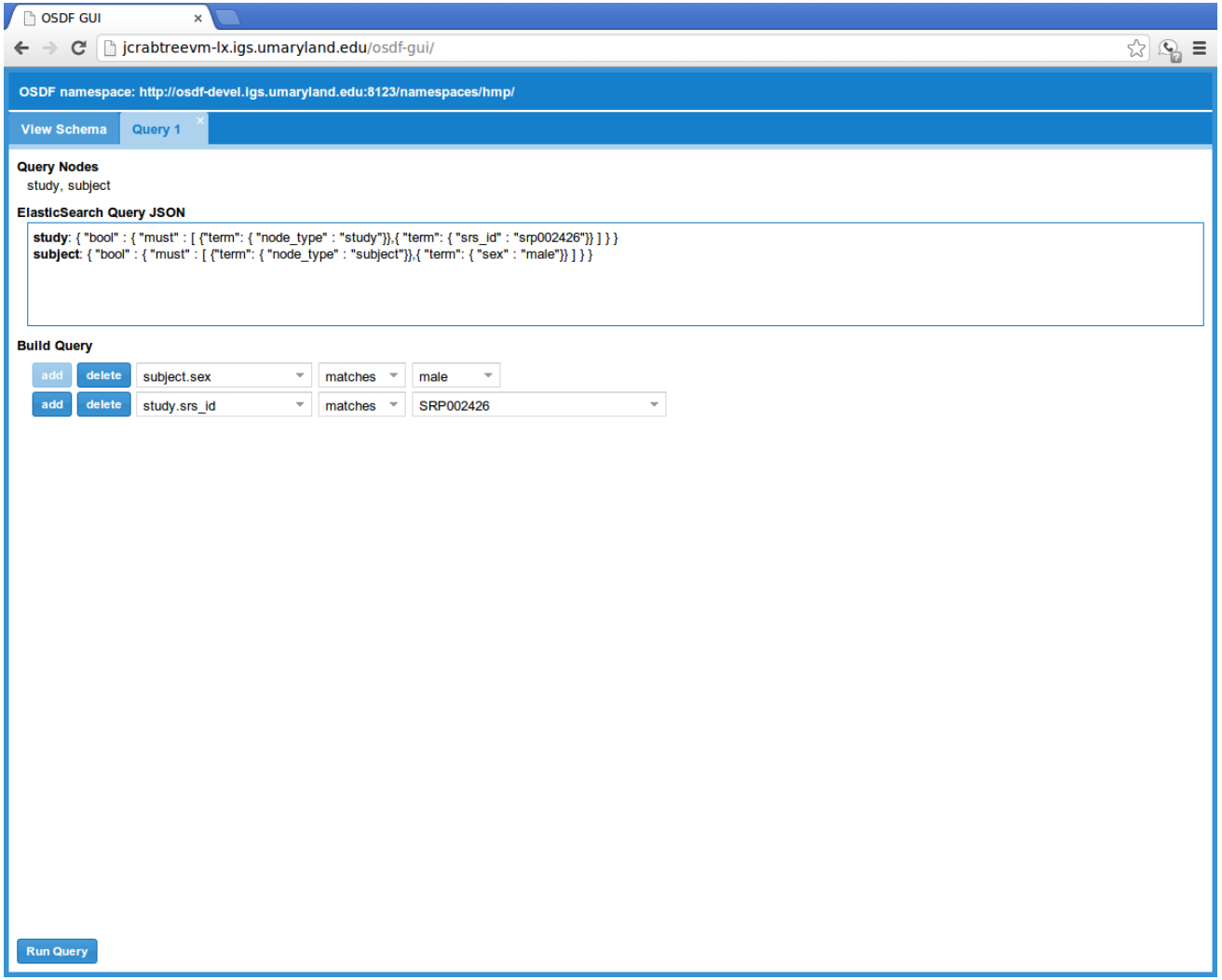
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Structured Query Interface



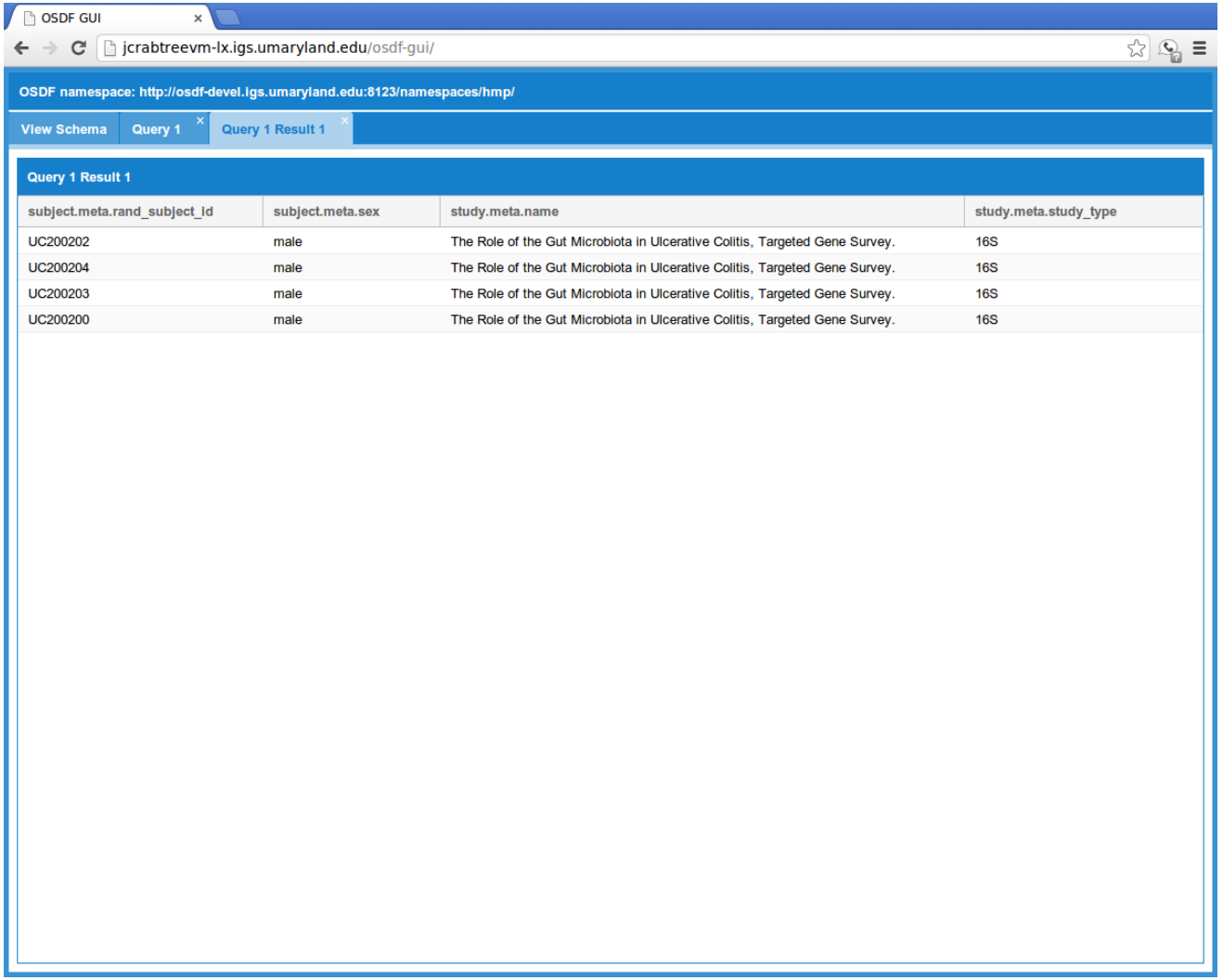
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Structured Query Interface



UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Structured Query Interface



NIAID Intramural Cloud Project

- Explore the use of computational clouds for storage and analysis
- All HMP data generated till 2012 was uploaded to Amazon S3 as a public dataset
- It has been downloaded over 13,000 times
- Building VM with common pipelines so data can stay within Amazon

iHMP DCC

- We will be managing the data for HMP2/iHMP
- Three PI driven projects
  - Virginia Commonwealth University
  - Harvard School of Public Health/The Broad Institute
  - Stanford University/Jackson Lab
- Multi-omic
  - host genomes
  - Transcriptomics
  - Lipidomics
  - Proteomics
- Longitudinal data
- 500 TB of raw and processed data

Anticipated iHMP Data

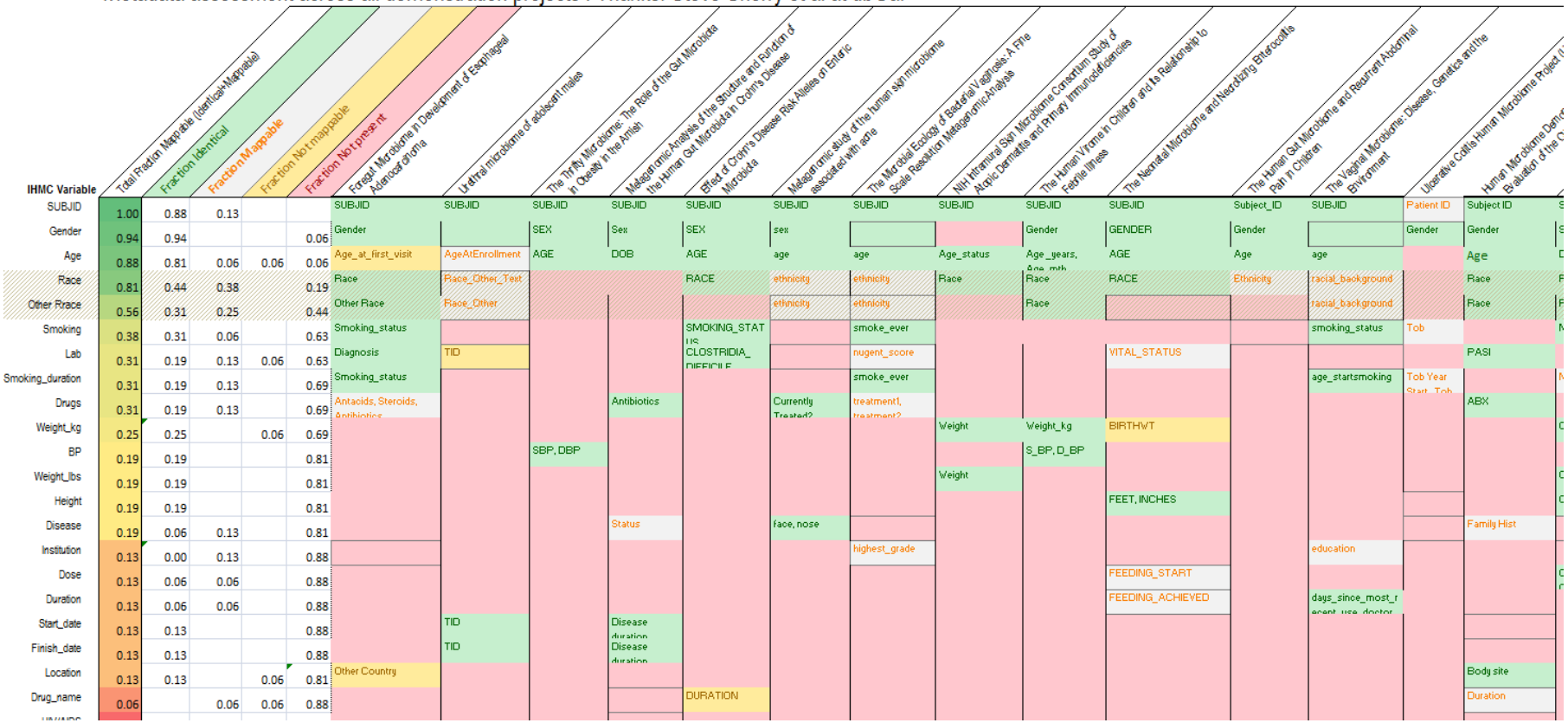
Data type	Count
Whole Genomes	150
Host Transcriptomes	1,440
Metatranscritptomes	4,980
Whole Metagenomes	4,980
16S	32,340
Other	7,960

LESSONS LEARNT

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Data Harmonization

Metadata assessment across all demonstration projects : Thanks: Steve Sherry et al at dbGaP



All Databases Are Not Made Equal

- CouchDB is great for returning entire documents
- But making ad-hoc structured queries are difficult and slow

Maintaining Analytical Engine is Hard

- Operating systems change
- Applications change
- Algorithms improve

User Support Cannot be Downplayed

- Make it easier for users to find and analyze data
- Helpdesk to support users
- Create multi-media tutorials

Links

- OSDF Website

  http://osdf.igs.umaryland.edu
- OSDF Code

  http://sourceforge.net/projects/osdf/
- Metagenome API site

  http://sourceforge.net/projects/metagenosdf/
- DACC Website

  http://hmpdacc.org

Acknowledgements

- Own White (PI)
- Michelle Giglio
- Victor Felix
- Jonathan Crabtree
- Heather Huot Creasy
- Michael Schor