



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 Microbiome Project (HMP)

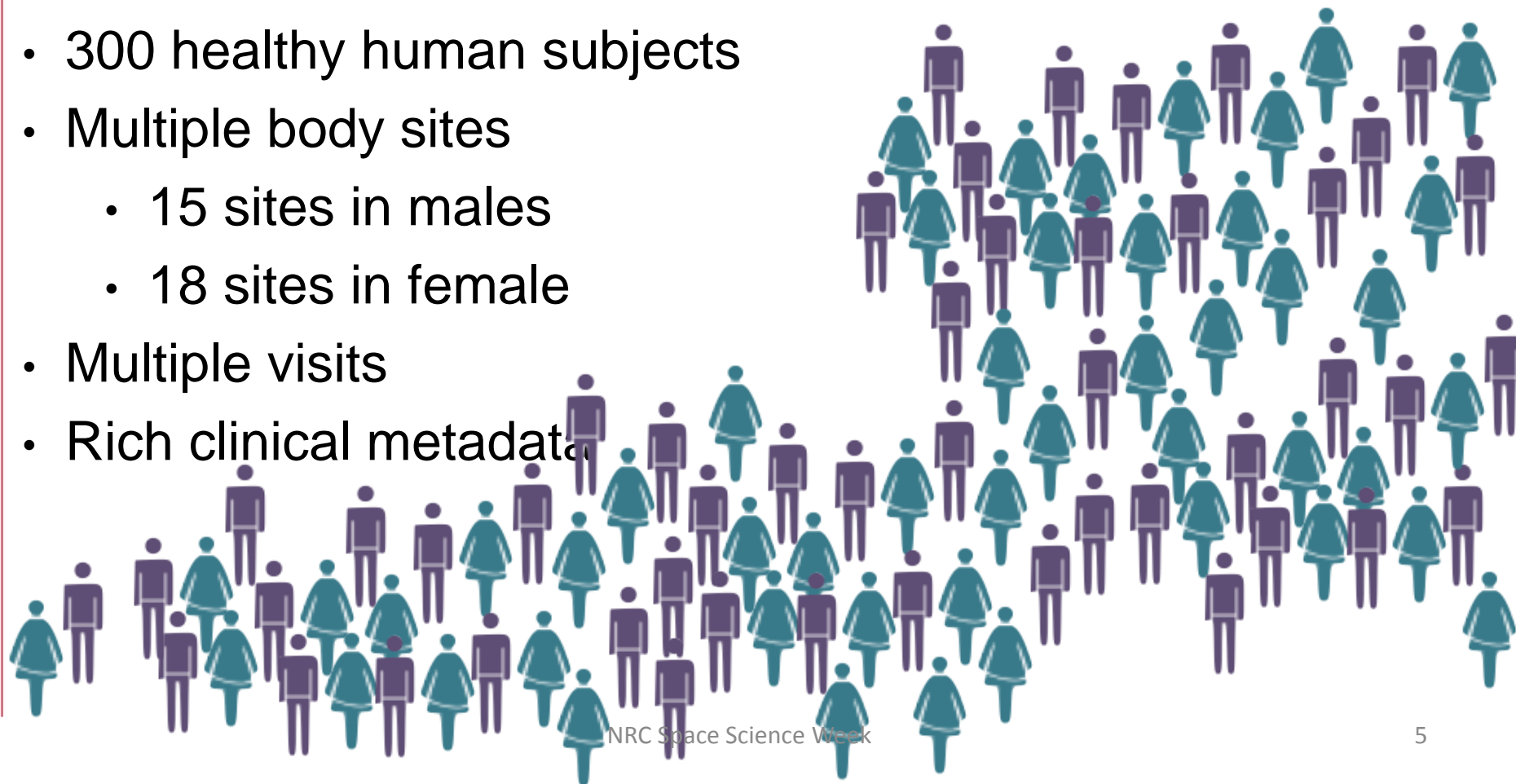
A comprehensive microbial survey

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



Healthy Human Subjects (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

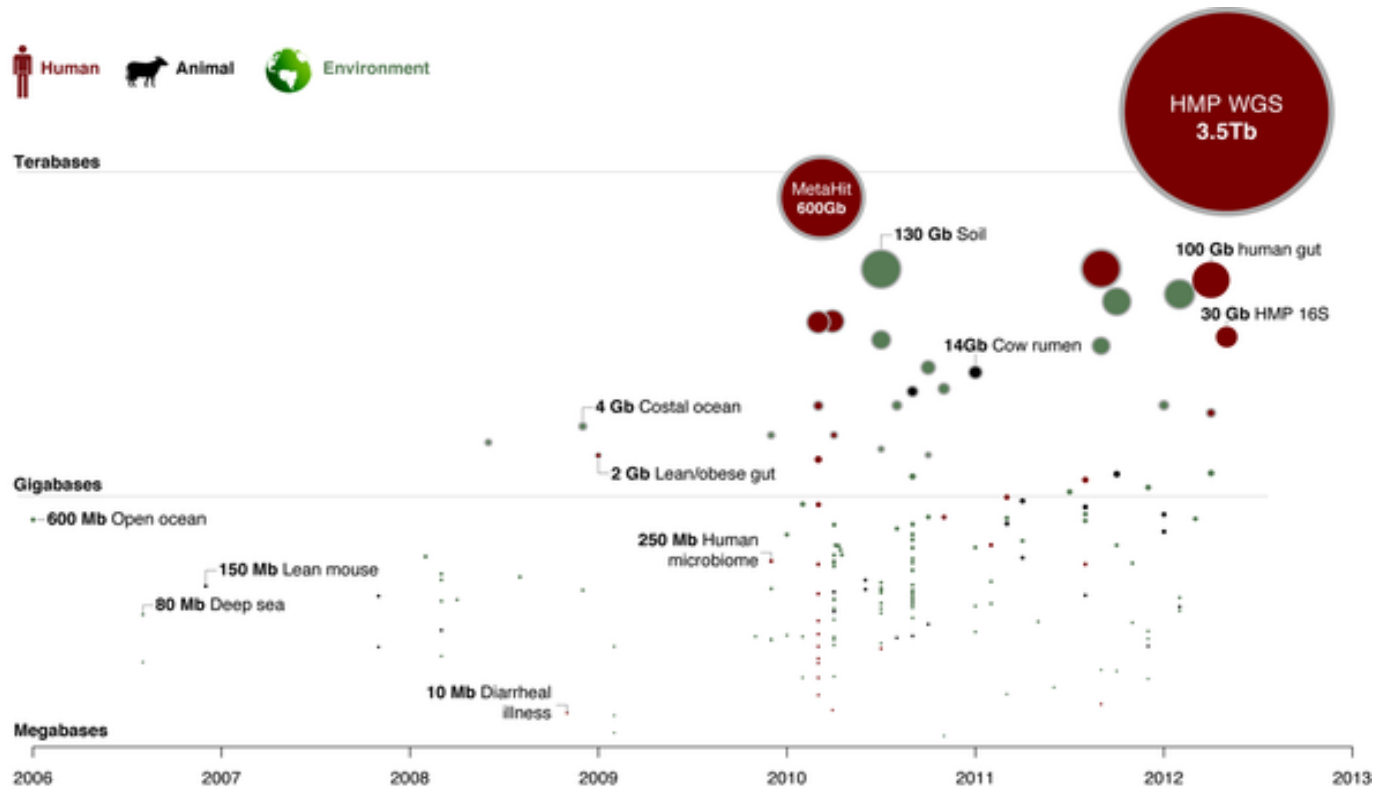


Data Repository





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

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

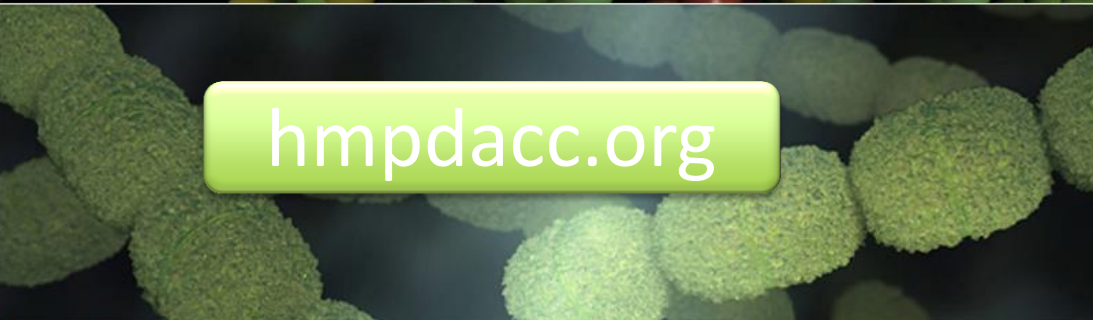


Welcome to the Data Analysis and Coordination Center (DACC) for the National Institutes of Health (NIH) Common Fund supported Human Microbiome Project (HMP). This site is the central repository for all HMP data. The aim of the HMP is to characterize microbial communities found at multiple human body 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](#)

Areas of Interest



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](#)

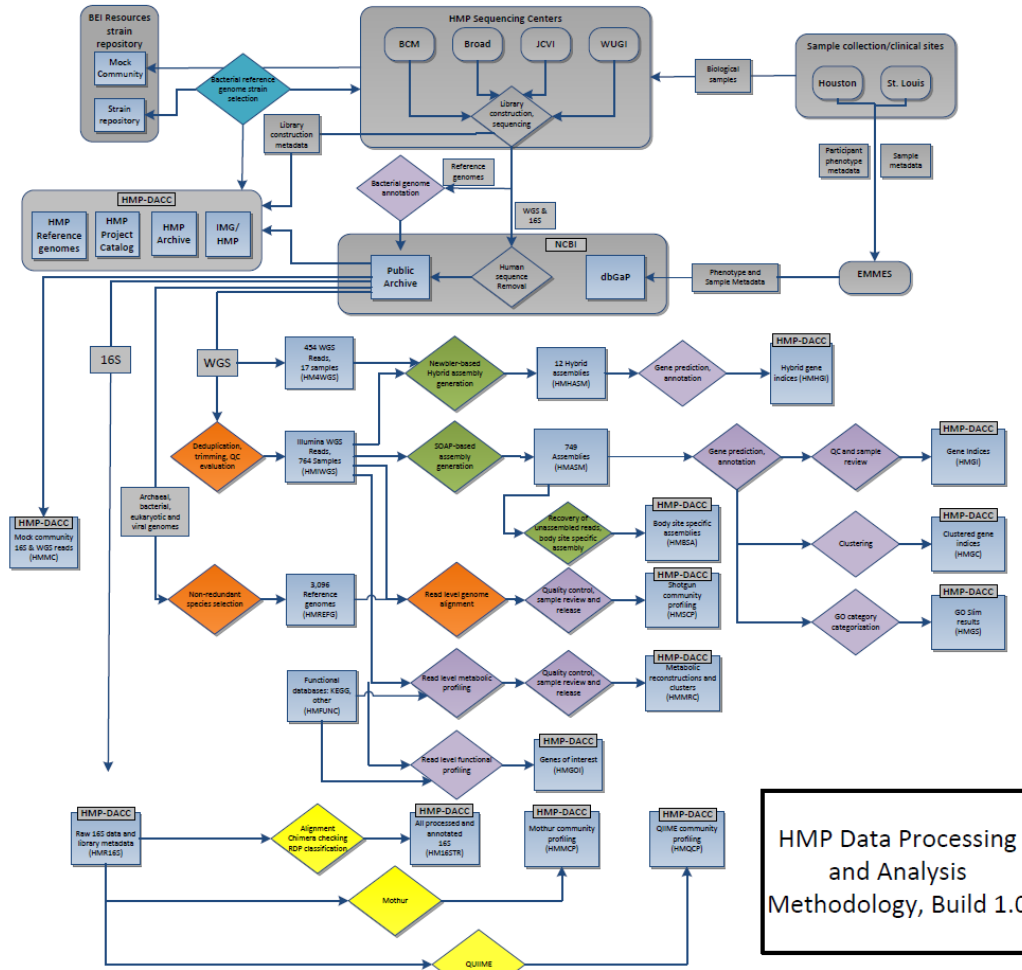




DATA TYPES

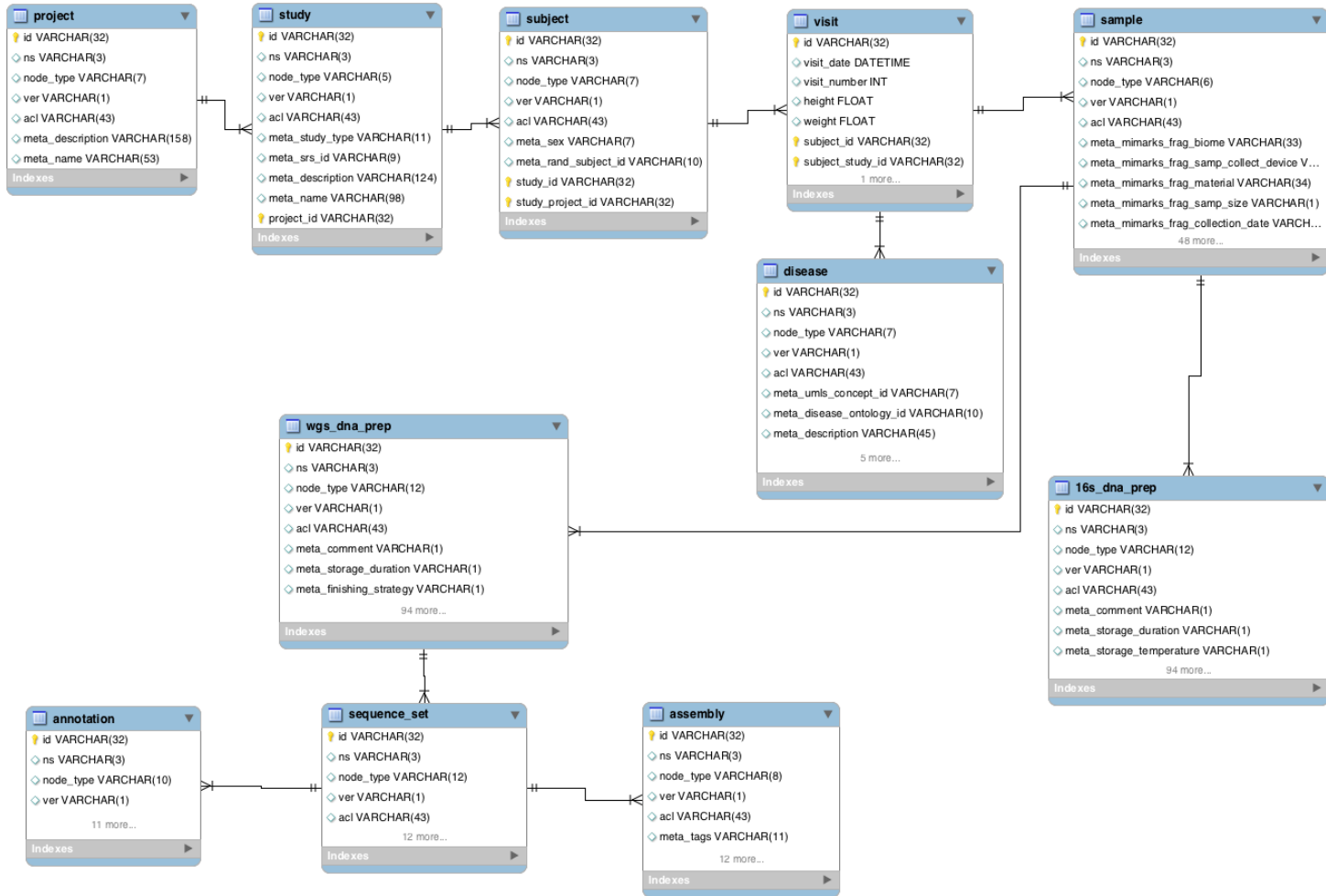


Data Processing Flow





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



Current HMP DACC Site



NIH HUMAN MICROBIOME PROJECT

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

[Feedback](#)

HMSCP - Shotgun community profiling

Reads generated by Illumina wgs sequencing were mapped on to a [database of reference genomes](#) in order to calculate organism abundance.

For each sample, we provide three files:

- A tab delimited abundance table, indicating depth and breadth of mapping to each reference
- A metrics file, summarizing the number of reads mapped versus the number that aligned to a reference
- Mapping alignment files in bam format

- [Data Table](#)
- [Protocols and Tools](#)
- [Related Pages](#)

HMSCP			
Description ▲	Download	Size	MD5
⊕ SRS011061 (3 Rows)			
Abundance Table		6.4 KB	21f28e1f2b5e2359b025eebdb37a4dfe
BAM File		7.5 GB	38b89e68b52bfdd1db4cd2165c7e3cf9
Metrics File		64.0 bytes	5633a925bfe1d142fd31fde215fc5924
⊕ SRS011086 (3 Rows)			
⊖ SRS011090 (3 Rows)			
Abundance Table		3.1 KB	eeef2d2eda1358a6745d93c51a1c3217c
BAM File		145.7 MB	7ea19443bd981e67e59e63a8fd6d45a1
Metrics File		162.0 bytes	1f4b8ecab4453e7dcaa7299f034d691a



Data Discovery

The screenshot shows the NIH Human Microbiome Project (HMP) website. The main navigation bar includes: REFERENCE GENOMES, MICROBIOME ANALYSIS, IMPACTS ON HEALTH, TOOLS & TECHNOLOGY, ETHICAL IMPLICATIONS, OUTREACH, and HMPDACC DATA BROWSER. A search bar contains the query 'female fornix'. Below the search bar, a table displays the following statistics:

Projects	0	Subjects	0
Samples	0	Assemblies	0
WGS DNA Preps	45	16S DNA Preps	330
Reference Genomes	0	Alignments	0
Sequence Sets	0	Annotations	0

On the left side of the page, there are sections for 'Current News' (listing events from September 2014, May 2014, and April 2014), 'Publications' (listing research on pressure stabilizer, vaginal microbiome, and microbiome/immunity), and 'Partner Resources' (listing NIH Common Fund and NCBI HMP Data Repository).



Data Discovery

The screenshot shows the NIH Human Microbiome Project website. The header includes navigation links for Reference Genomes, Microbiome Analysis, Impacts on Health, Tools & Technology, Ethical Implications, Outreach, and HMPDACC Data Browser. The main content area displays 'OSDF Query Results' with a table of query results. The table has columns for Type, Version, View OSDF JSON, Sequencing Method, Sex, Random Subject ID, SRS ID, and HMP Body Site. The results list 30 entries, all of which are '16S DNA Prep' for 'posterior_fornix' samples.

OSDF Query Results

Type	Version	View OSDF JSON	Sequencing Method	Sex	Random Subject ID	SRS ID	HMP Body Site
16S DNA Prep	4	View OSDF Data		female	158458797	SRS011110	posterior_fornix
16S DNA Prep	4	View OSDF Data		female	158883629	SRS011236	posterior_fornix
16S DNA Prep	4	View OSDF Data		female	158944319	SRS011355	posterior_fornix
16S DNA Prep	4	View OSDF Data		female	158984779	SRS011403	posterior_fornix
16S DNA Prep	4	View OSDF Data		female	338793263	SRS044653	posterior_fornix
16S DNA Prep	4	View OSDF Data		female	706846339	SRS055349	posterior_fornix
16S DNA Prep	3	View OSDF Data		female	393523607	SRS053365	posterior_fornix
16S DNA Prep	4	View OSDF Data		female	158742018	SRS011268	posterior_fornix
16S DNA Prep	4	View OSDF Data		female	159733294	SRS011618	posterior_fornix
16S DNA Prep	4	View OSDF Data		female	675950834	SRS056796	posterior_fornix
16S DNA Prep	9	View OSDF Data		female	158458797	SRS011111	posterior_fornix
16S DNA Prep	9	View OSDF Data		female	158742018	SRS011269	posterior_fornix
16S DNA Prep	9	View OSDF Data		female	159005010	SRS011449	posterior_fornix
16S DNA Prep	9	View OSDF Data		female	159085930	SRS011526	posterior_fornix
16S DNA Prep	9	View OSDF Data		female	159713063	SRS011685	posterior_fornix
16S DNA Prep	4	View OSDF Data		female	161473083	SRS021832	posterior_fornix
16S DNA Prep	3	View OSDF Data		female	159005010	SRS023647	posterior_fornix
16S DNA Prep	3	View OSDF Data		female	159247771	SRS024073	posterior_fornix
16S DNA Prep	3	View OSDF Data		female	414519462	SRS043799	posterior_fornix
16S DNA Prep	3	View OSDF Data		female	638754422	SRS043922	posterior_fornix
16S DNA Prep	3	View OSDF Data		female	764366428	SRS044902	posterior_fornix
16S DNA Prep	3	View OSDF Data		female	159005010	SRS011450	posterior_fornix
16S DNA Prep	3	View OSDF Data		female	159733294	SRS011619	posterior_fornix

Return to Home Contact Policy



Structured Query Interface

OSDF GUI

← → ↻ jcrabtreevm-lx.igs.umaryland.edu/osdf-gui/

OSDF namespace: <http://osdf-devel.igs.umaryland.edu:8123/namespaces/hmp/>

View Schema

View ▾ Help

```
graph TD; clustering[clustering 14] --> annotation[annotation 702]; clustering --> assembly[assembly 779]; clustering --> sequence_set[sequence_set 21321]; annotation --> sequence_set; assembly --> sequence_set; sequence_set --> wgs_dna_prep[wgs_dna_prep 764]; sequence_set --> 16s_dna_prep[16s_dna_prep 9991]; sequence_set --> uncategorised[uncategorised_reference_genome 1]; sequence_set --> sample[sample 10492]; sequence_set --> subject[subject 1331]; sequence_set --> disease[disease 8]; sequence_set --> study[study 12]; sequence_set --> project[project 6]; sample --> subject; sample --> disease; sample --> study; subject --> disease; subject --> study; study --> project;
```

0 nodes selected

Clear All Build New Query

Query Builder

Click on an OSDF node type to see options, double click to select/deselect.

Download schema overview: [JPEG](#) [PDF](#) [PNG](#) [SVG](#)



Structured Query Interface

OSDF namespace: <http://osdf-devel.igs.umaryland.edu:8123/namespaces/hmp/>

View Schema Query 1

Query Nodes
study, subject

ElasticSearch Query JSON

```
study: {"bool": {"must": [{"term": {"node_type": "study"}}, {"term": {"srs_id": "srp002426"}}]}}
subject: {"bool": {"must": [{"term": {"node_type": "subject"}}, {"term": {"sex": "male"}}]}}
```

Build Query

add	delete	subject.sex	matches	male
add	delete	study.srs_id	matches	SRP002426

Run Query



Structured Query Interface

OSDF namespace: <http://osdf-devel.igs.umaryland.edu:8123/namespaces/hmp/>

View Schema Query 1 Query 1 Result 1

Query 1 Result 1

subject.meta.rand_subject_id	subject.meta.sex	study.meta.name	study.meta.study_type
UC200202	male	The Role of the Gut Microbiota in Ulcerative Colitis, Targeted Gene Survey.	16S
UC200204	male	The Role of the Gut Microbiota in Ulcerative Colitis, Targeted Gene Survey.	16S
UC200203	male	The Role of the Gut Microbiota in Ulcerative Colitis, Targeted Gene Survey.	16S
UC200200	male	The Role of the Gut Microbiota in Ulcerative Colitis, Targeted Gene Survey.	16S



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
Metatranscriptomes	4,980
Whole Metagenomes	4,980
16S	32,340
Other	7,960



LESSONS LEARNT

Data Harmonization

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

IHMC Variable	Total Fraction Mappable (Identical+Mappable)	Fraction Identical	Fraction Mappable	Fraction Not mappable	Fraction Not present	Fraction Microbiome In Development & Esophageal Adenocarcinoma	Urbal microbiome of adolescent males	The Oral Microbiome: The Role of the Gut Microbiota in Obesity in the Adult	Metagenomic Analysis of the Structure and Function of the Human Gut Microbiota in Crohn's Disease	Effect of Crohn's Disease Risk Alleles of Ethnic Microbiota	Metagenomic study of the human skin microbiome associated with acne	The Microbial Ecology of Bacterial Vaginitis: A Pilot Study Resolving Metagenomic Analysis	NH3 Human Skin Microbiome Consortium Study of Acne Dermatitis and Other Primary Immunodeficiencies	The Human Virome in Children and Its Relationship to Feline Infection	The Neonatal Microbiome and Neurozing Enterococci	The Human Gut Microbiome and Recurrent Abdominal Pain in Children	The Vaginal Microbiome and Recurrent Abdominal Pain	Ulcerative Colitis Human Microbiome Project	Human Microbiome Demographic Evaluation of the C
SUB_ID	1.00	0.88	0.13			SUB_ID	SUB_ID	SUB_ID	SUB_ID	SUB_ID	SUB_ID	SUB_ID	SUB_ID	SUB_ID	Subject_ID	SUB_ID	Patient ID	Subject ID	
Gender	0.94	0.94			0.06	Gender	SEX	Sex	SEX	sex		Gender	GENDER	Gender		Gender	Gender	Gender	
Age	0.88	0.81	0.06	0.06	0.06	Age_at_first_visit	AgeAtEnrollment	AGE	DOB	AGE	age	age	Age_status	Age_years	AGE	Age	age		Age
Race	0.81	0.44	0.38		0.19	Race	Place_Other_Test		RACE	ethnicity	ethnicity	Face	Face	RACE	Ethnicity	racial_background		Face	
Other Race	0.56	0.31	0.25		0.44	Other Race	Place_Other			ethnicity	ethnicity		Face			racial_background		Face	
Smoking	0.38	0.31	0.06		0.63	Smoking_status			SMOKING_STAT	smoke_ever						smoking_status	Tob		
Lab	0.31	0.19	0.13	0.06	0.63	Diagnosis	TID		CLOSTRIDIA	nugent_score			VITAL_STATUS					PASI	
Smoking_duration	0.31	0.19	0.13		0.69	Smoking_status				smoke_ever						age_startsmoking	Tob Year		
Drugs	0.31	0.19	0.13		0.69	Antacids, Steroids, Antibiotics		Antibiotics		Currently Treated?	treatment1, treatment2							ABX	
Weight_kg	0.25	0.25		0.06	0.69						Weight	Weight_kg	BIRTHWT						
BP	0.19	0.19			0.81		SBP, DBP					S_BP, D_BP							
Weight_lbs	0.19	0.19			0.81						Weight								
Height	0.19	0.19			0.81								FEET, INCHES						
Disease	0.19	0.06	0.13		0.81		Status		face, nose									Family Hist	
Institution	0.13	0.00	0.13		0.88					highest_grade									
Dose	0.13	0.06	0.06		0.88								FEEDING_START						
Duration	0.13	0.06	0.06		0.88								FEEDING_ACHIEVED			days_since_most_r			
Start_date	0.13	0.13			0.88	TID		Disease duration											
Finish_date	0.13	0.13			0.88	TID		Disease duration											
Location	0.13	0.13		0.06	0.81	Other Country												Body site	
Drug_name	0.06		0.06	0.06	0.88			DURATION										Duration	



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