UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

OSDF
A cloud enabled system to store, access, and analyze scientific data

Anup Mahurkar

Institute for Genome Sciences

BOSC 2013

Motivation

- Work based on NIH Human Microbiome Project Data Analysis and Coordination Center
- 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

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, Subject information

What is OSDF?

- Includes a mechanism for
  - Defining data model for elements
  - Reliance on Ontologies, and controlled vocabularies
  - Defining relationships between elements
  - Generic RESTful API for accessing and placing data
  - Domain specific API with Perl/Python/Java bindings
  - Versioning and history
  - Access control

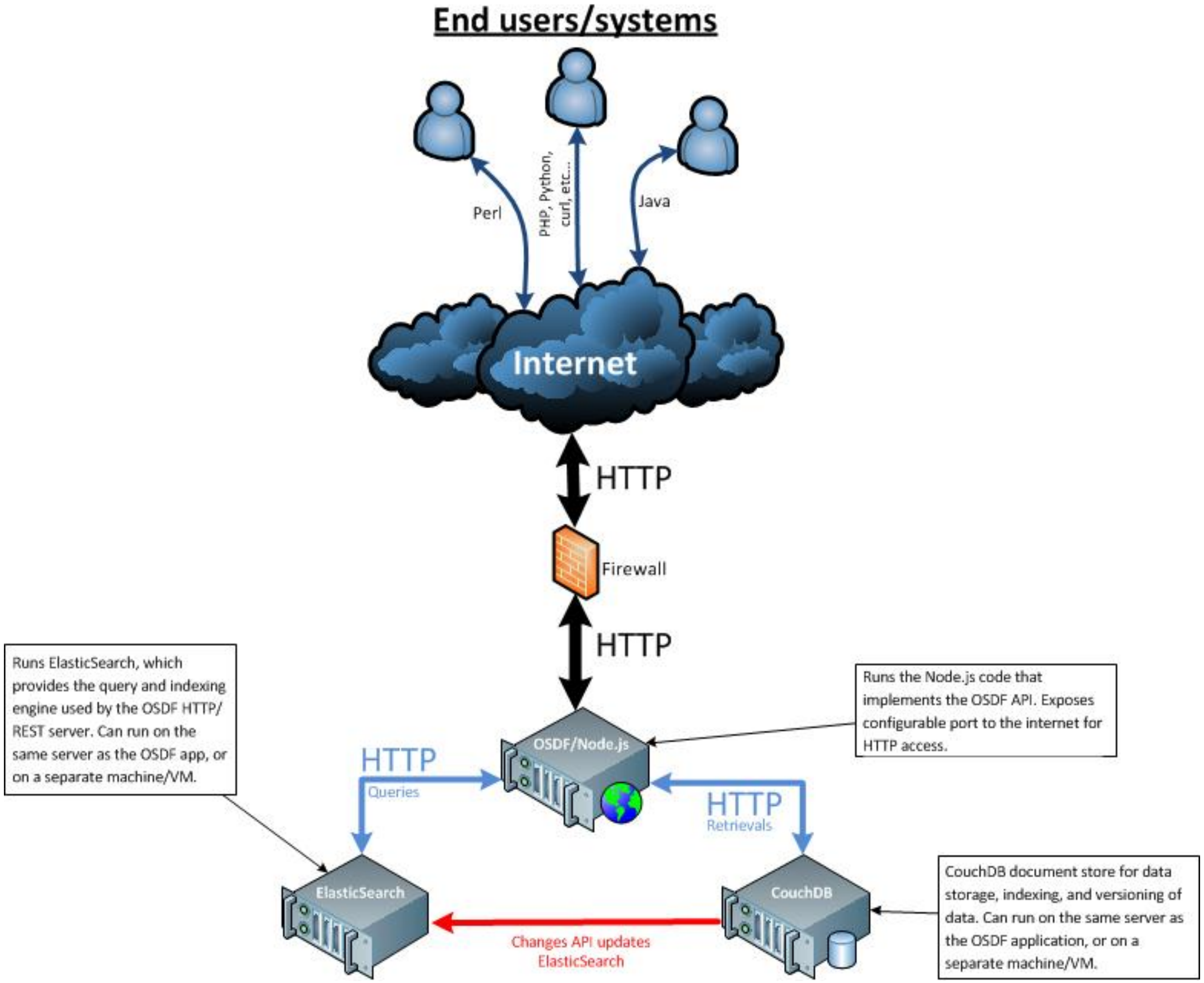
Technologies Used

- JavaScript Object Notation (JSON) objects for modeling data elements
  - Lightweight data interchange format
  - Allows sparse data in elements
  - Easy to generate and parse
  - Compact and human readable
  - JSON Schema for validation
- CouchDB for storing JSON objects
  - Document-oriented database
  - Apache project
  - RESTful JSON API out of the box
  - Real-time bi-directional replication

Technologies Used

- ElasticSearch
  - Rapid indexing on all keys and attributes of JSON
  - Query language for complex ad-hoc queries
  - 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

OSDF Architecture



Runs ElasticSearch, which provides the query and indexing engine used by the OSDF HTTP/ REST server. Can run on the same server as the OSDF app, or on a separate machine/VM.

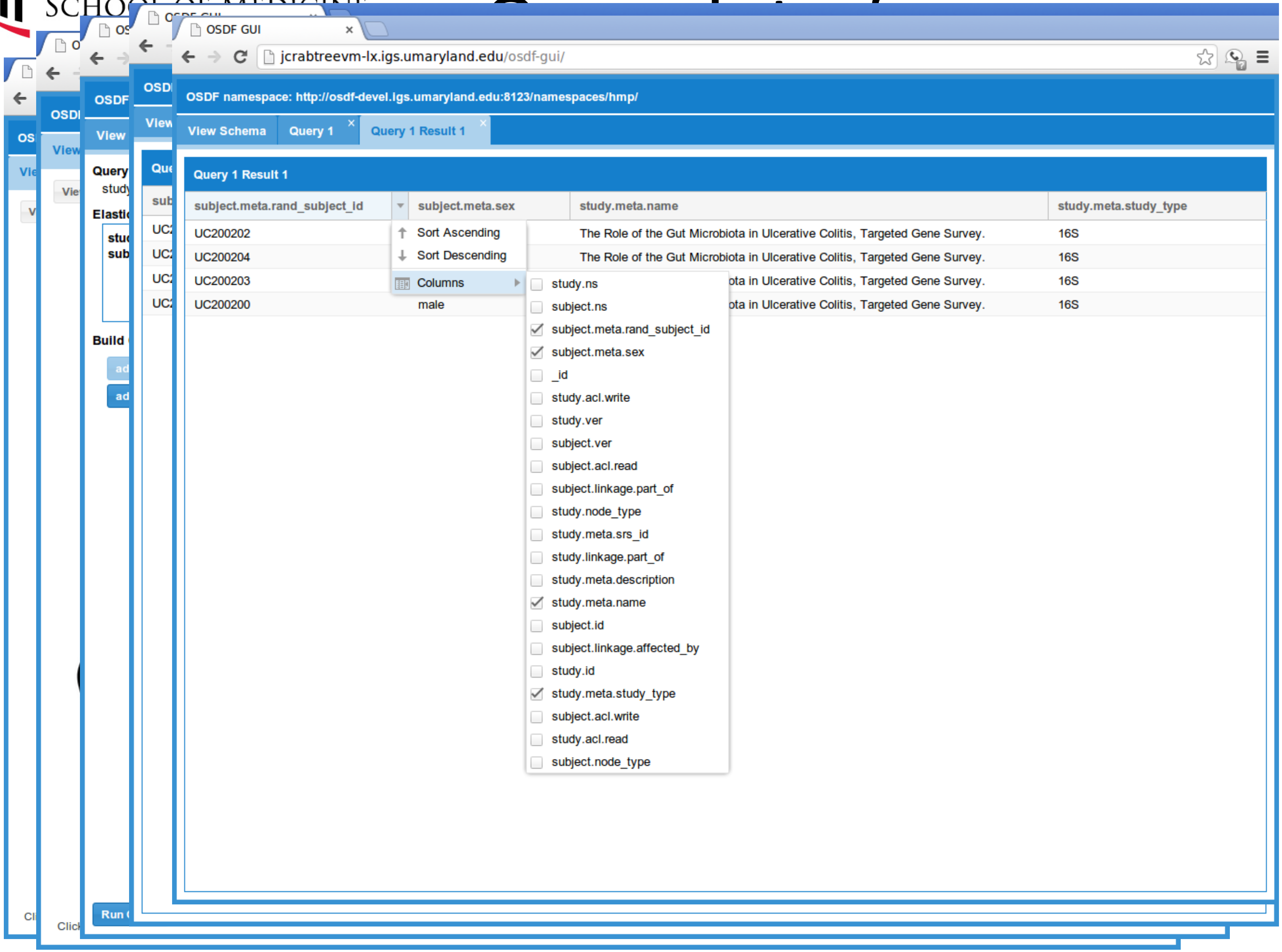
Runs the Node.js code that implements the OSDF API. Exposes configurable port to the internet for HTTP access.

CouchDB document store for data storage, indexing, and versioning of data. Can run on the same server as the OSDF application, or on a separate machine/VM.

Schema and Node Example

```
{
  "type": "object",
  "description": "Sample nodes describe physical samples.",
  "properties": {
    "hmp_body_site": {
      "$ref": "body_sites"
    },
    "hmp_supersite": {
      "$ref": "supersites"
    },
    "mimarks_frag": {
      "$ref": "mimarks_frag"
    },
    "mims_frag": {
      "$ref": "mims_frag"
    },
    "visit_number": {
      "type": "integer",
      "minimum": 1,
      "required": true
    },
    "fma_body_site": {
      "title": "Typically a term from the FMA ontology.",
      "type": "string",
      "required": true
    },
    "body_product": {
      "type": "string",
      "required": true
    }
  },
  "additionalProperties": false
}
```

```
{
  "linkage": {
   "part_of": ["c8550ef8d3ea8c9b980650de1c6e86cf”],
   "collected_from": ["c8550ef8d3ea8c9b980650de1c87e0e1”]
  },
  "node_type": "sample",
  "meta": {
   "body_product": "",
   "hmp_supersite": "skin",
   "visit_number": 1,
   "mimarks_frag": {
    "geo_loc_name": "United States of America",
    "samp_mat_process": "",
    "lat_lon": "Unknown",
    "samp_collect_device": "Catch-All sample collection swab",
    "biome": "terrestrial biome [ENVO:00000446]",
    "samp_size": "",
    "feature": "human-associated habitat [ENVO:00009003]",
    "collection_date": "Unknown",
    "env_package": "human-associated",
    "investigation_type": "mimarks-survey",
    "material": "biological product [ENVO:02000043]",
    "rel_to_oxygen": ""
   },
   "fma_body_site": "",
   "hmp_body_site": "left_retroauricular_crease"
  },
  "id": "091f03831014afe9b2da67d698000c51",
  "ver": 1
 },
```



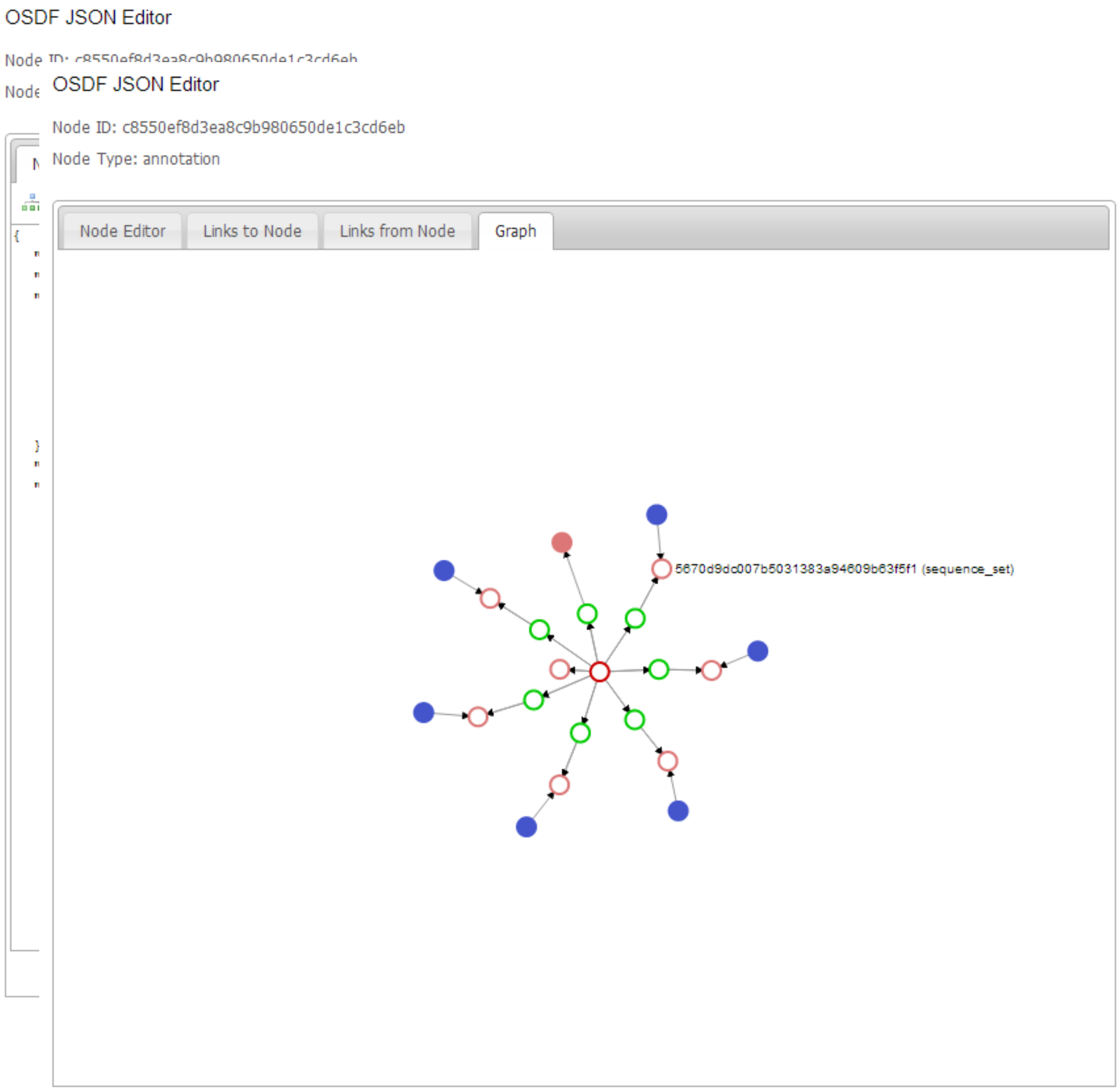
OSDF namespace: http://osdf-devel.igs.umaryland.edu:8123/namespaces/hmp/

Query 1 Result 1

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

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Node Editor



OSDF JSON Editor

Node ID: c8550ef8d3ea8c9b980650de1c3cd6eb

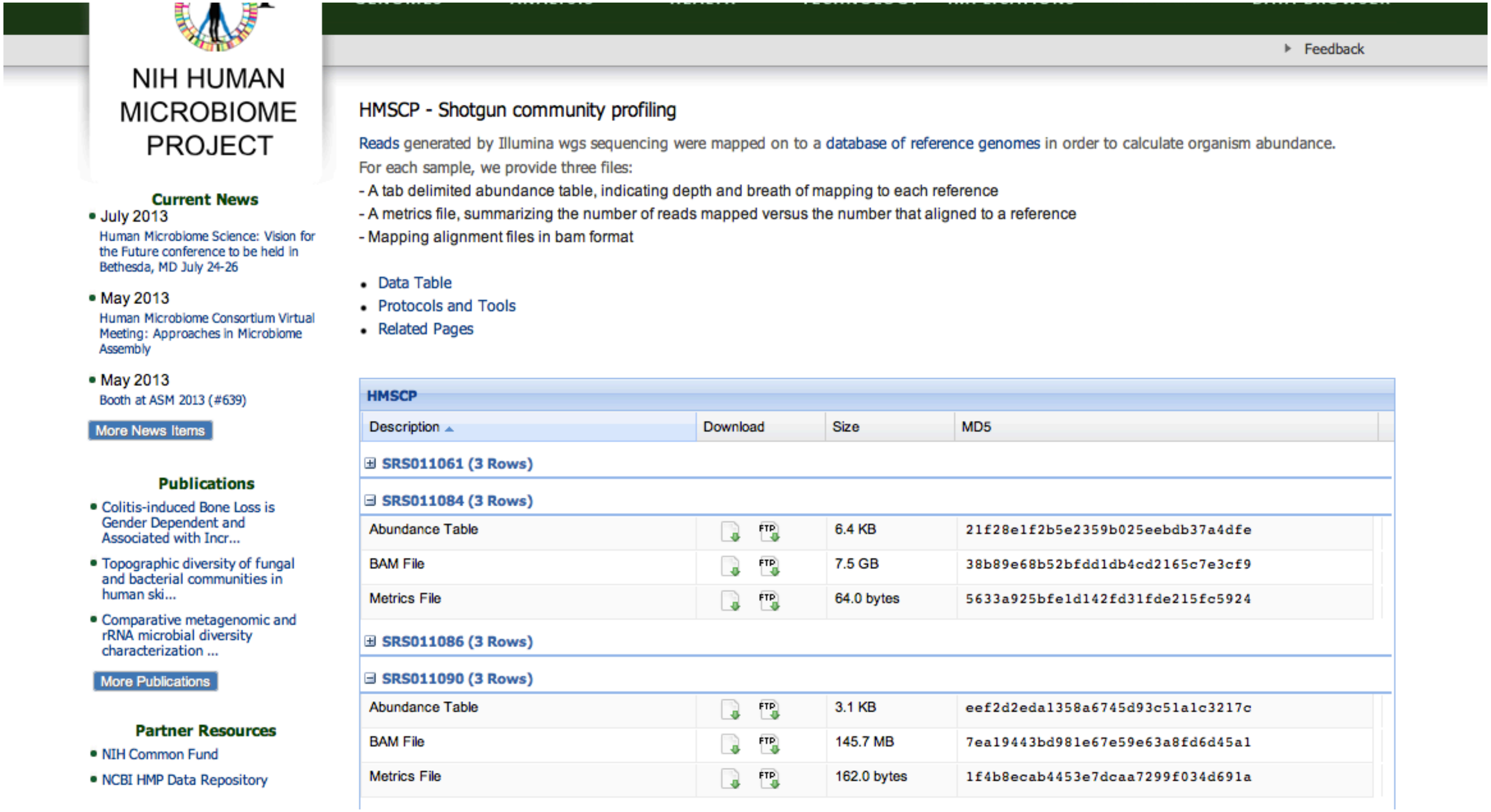
Node

OSDF JSON Editor

Node ID: c8550ef8d3ea8c9b980650de1c3cd6eb

Node Type: annotation

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)

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 ...

Partner Resources

- NIH Common Fund
- NCBI HMP Data Repository

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 breath 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)**			
⊟ **SRS011084 (3 Rows)**			
Abundance Table	FTP	6.4 KB	21f28e1f2b5e2359b025eebdb37a4dfe
BAM File	FTP	7.5 GB	38b89e68b52bfdd1db4cd2165c7e3cf9
Metrics File	FTP	64.0 bytes	5633a925bfe1d142fd31fde215fc5924
⊞ **SRS011086 (3 Rows)**			
⊟ **SRS011090 (3 Rows)**			
Abundance Table	FTP	3.1 KB	eef2d2eda1358a6745d93c51a1c3217c
BAM File	FTP	145.7 MB	7ea19443bd981e67e59e63a8fd6d45a1
Metrics File	FTP	162.0 bytes	1f4b8ecab4453e7dcaa7299f034d691a

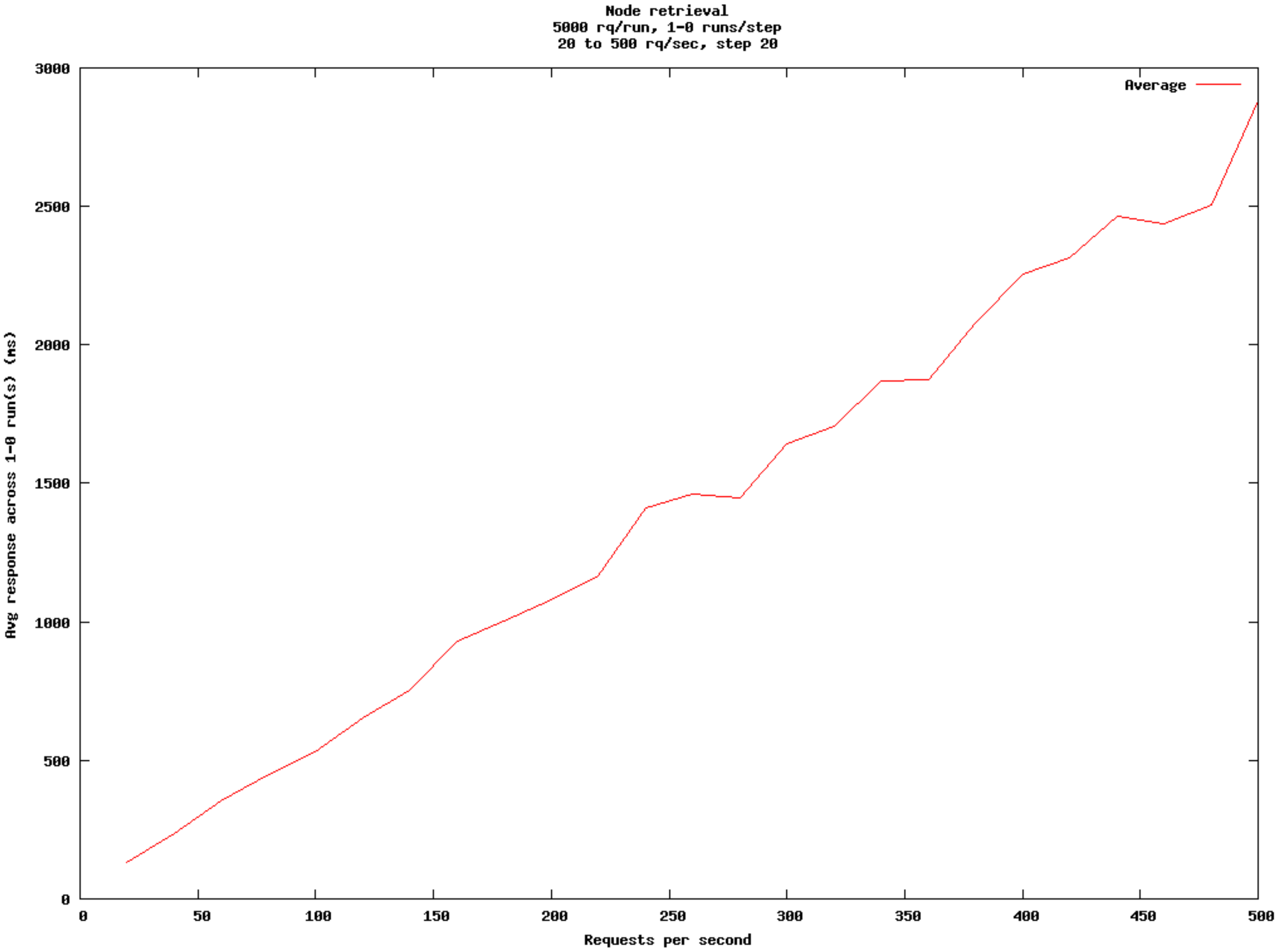
Programmatic Access

- Generic RESTful API
  - Get
  - Put
  - Update
  - Delete
- Domain Specific API
  - Metagenomics Perl API
    - Sequence data R operations
    - Assembly CR operations
    - Compress files
    - Upload files
    - Link nodes

Performance Testing

- Server specifications
  - VM running on VMWare ESX server
  - Ubuntu 12.04
  - 4 vCPUs, 16GB RAM
  - Node.js v0.10
  - Elastic Search v0.90
  - CouchDB v1.2
- Client specifications
  - 8 cores, 6GB RAM
- Benchmark
  - Use ‘ab’ Apache Server Benchmarking Tool
  - Run 5000 operations with increasing concurrency from 20 to 500 operations
  - { "query": {"term": {"node_type": "sample" }}}

Performance Numbers – Node Retrieval



Future Development

- Deployment in the cloud
- Replication
- Distributed deployment
- Performance enhancements
- Richer microbiome domain API
- Language bindings in Perl/Python/Java
- Host 1000 Genomes Data

Acknowledgements

- Principal Investigator
  - Owen White
- Institute for Genome Sciences
  - Sonia Agrawal
  - Cesar Arze
  - Jonathan Crabtree
  - Noam Davidovics
  - Victor Felix
  - Aaron Miller
  - Mike Schor
- University of Colorado Boulder
  - Rob Knight
- University of Chicago/Argonne National Lab
  - Folker Meyer
  - Narayan Desai
  - Jared Wilkening

Funding provided by:

NHGRI

U01 HG006537

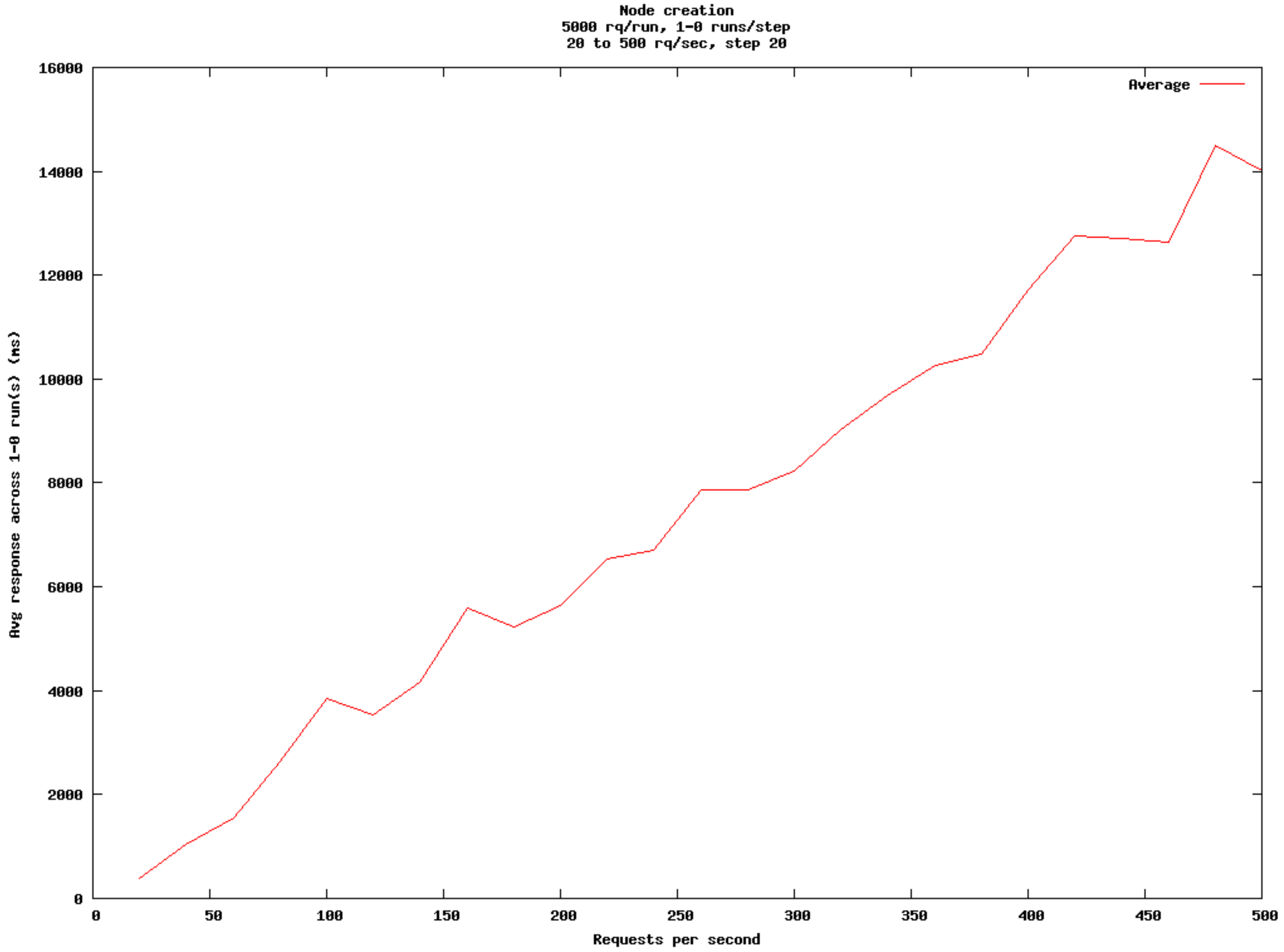
Resources

- 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

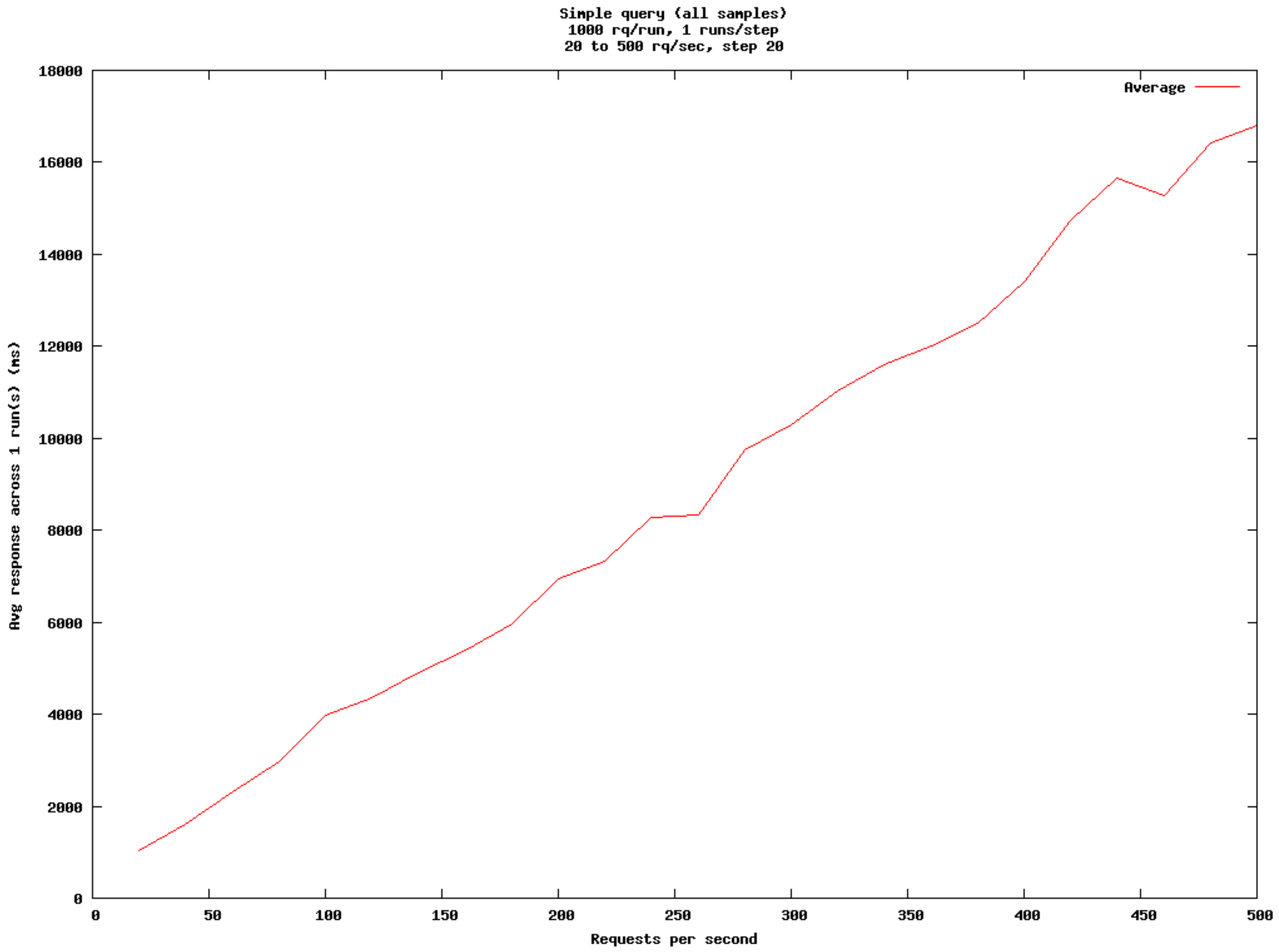
We invite you to take a look and participate

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Performance Numbers – Node Creation



Performance Numbers – Node Query


Simple query (all samples)
1000 rq/run, 1 runs/step
20 to 500 rq/sec, step 20
Average
Avg response across 1 run(s) (ms)
Requests per second