Bionimbus Protected Data Cloud

Allison Heath

University of Chicago

OSDC Edinburgh Workshop 2013

OPEN CLOUD CONSORTIUM

OPEN SCIENCE DATA CLOUD

Institute for Genomics & Systems Biology

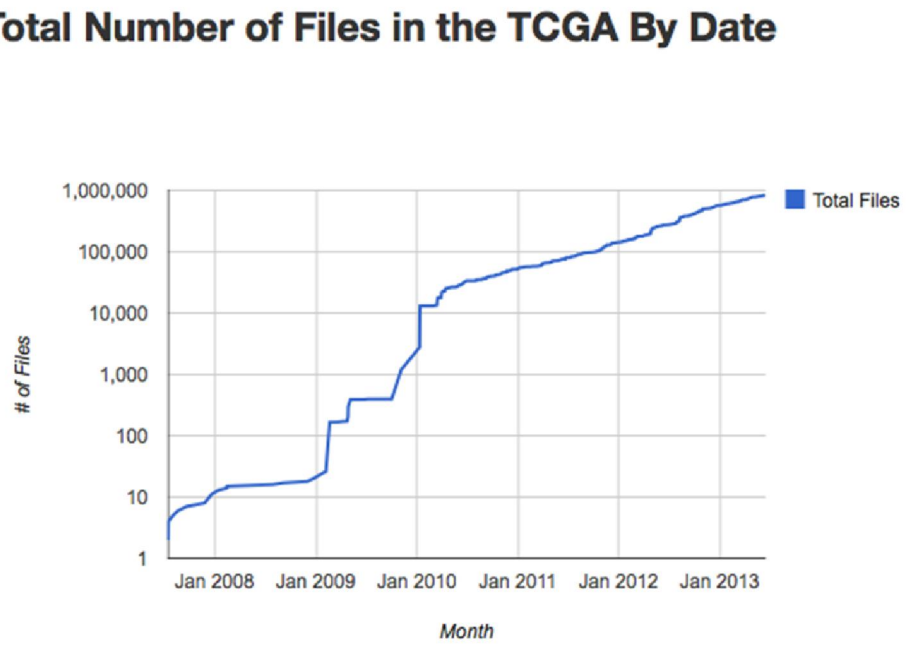
THE UNIVERSITY OF CHICAGO

Genomic Data

- Sequencing technology getting faster and cheaper
  - 1,000 genomes project currently 464 TB (still growing)
- "The $1,000 genome, the $100,000 analysis"
- Raw sequence data is noisy
  - New methods, important to keep the raw data and have ability to reprocess

The Cancer Genome Atlas (TCGA)

- Began in 2005 as an effort by NIH/NCI to catalogue genomic changes in cancer
  - 27 cancer types selected
  - Multiple samples from the same patient
  - Tumor/normal pairs
- Currently 525 TB
  - Projected to grow to >2 PB in the next two years
- Distribute data (cghub / genetorrent) but no computational facilities



http://tcga.github.io/Roadmap/

OPEN CLOUD CONSORTIUM

OPEN SCIENCE DATA CLOUD

Institute for Genomics & Systems Biology

THE UNIVERSITY OF CHICAGO

Account Requirements

- eRA Commons username – granted by NIH
- dbGaP access to TCGA – granted by NIH
  - Database for genotypes and phenotypes
- bionimbus-pdc.opensciencedatacloud.org/apply

Bionimbus-PDC Overview

Initial Equipment – 1 Rack

39 1U Servers

1 Head Node

1 Starlight Connected Node

1 Cloud Cloud Controller Node

35 Compute Nodes

- 8 cores
- 32GB RAM
- 4 x 2TB SATA
- 10 Gbps NIC

168TB Usable GlusterFS

OpenStack Essex

Base VM Image Ubuntu 12.04 LTS

Second Phase Equipment - 1 rack

8 4U Servers

32 Compute Nodes

- 16 cores
- 128GB RAM
- 7 x 4TB SATA
- 1 x 120 GB SSD

704TB Usable GlusterFS

1 rack automatically provisioned and integrated to existing cloud with “Yates” in ~60 minutes

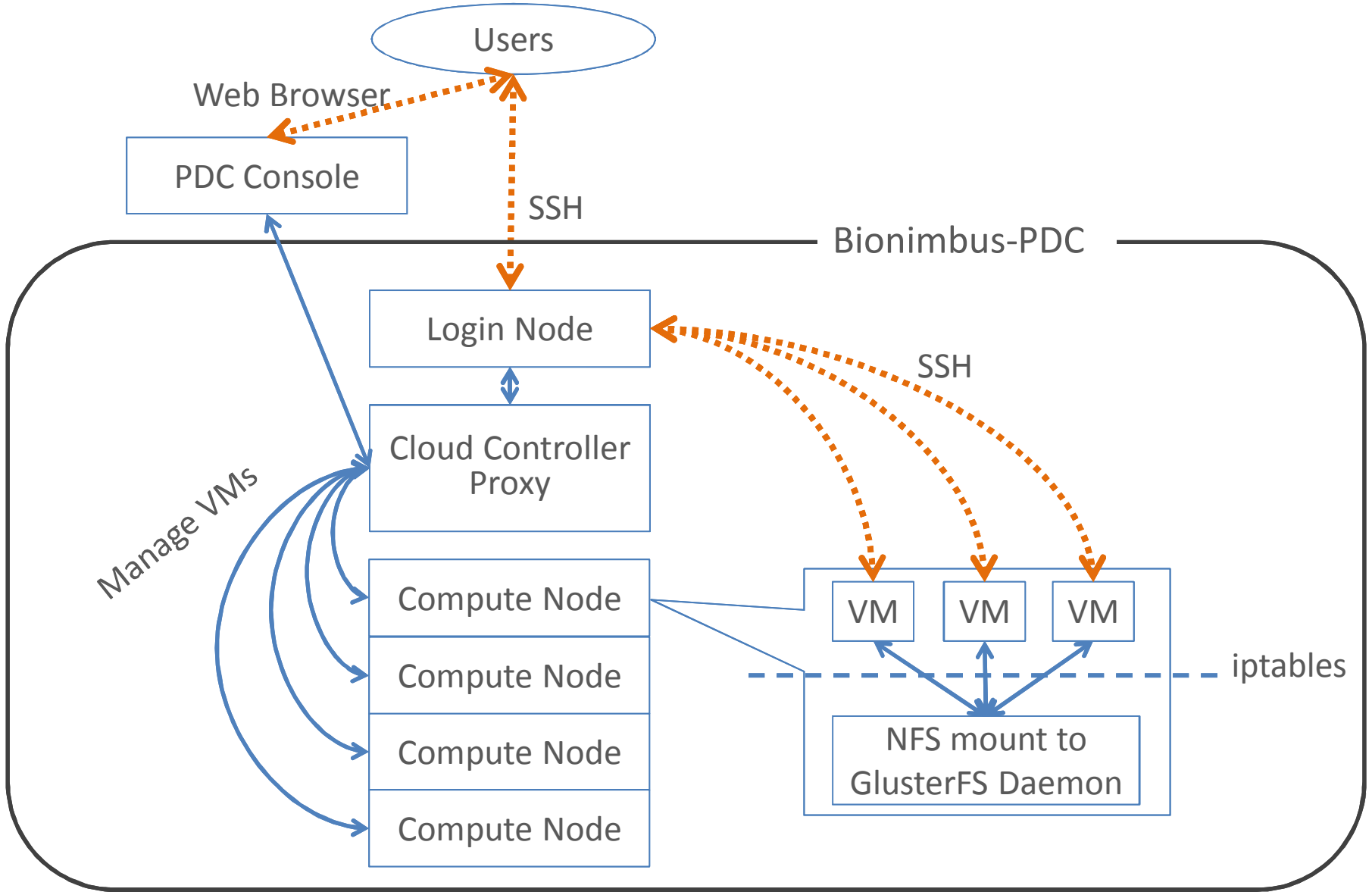
OPEN CLOUD CONSORTIUM

OPEN SCIENCE DATA CLOUD

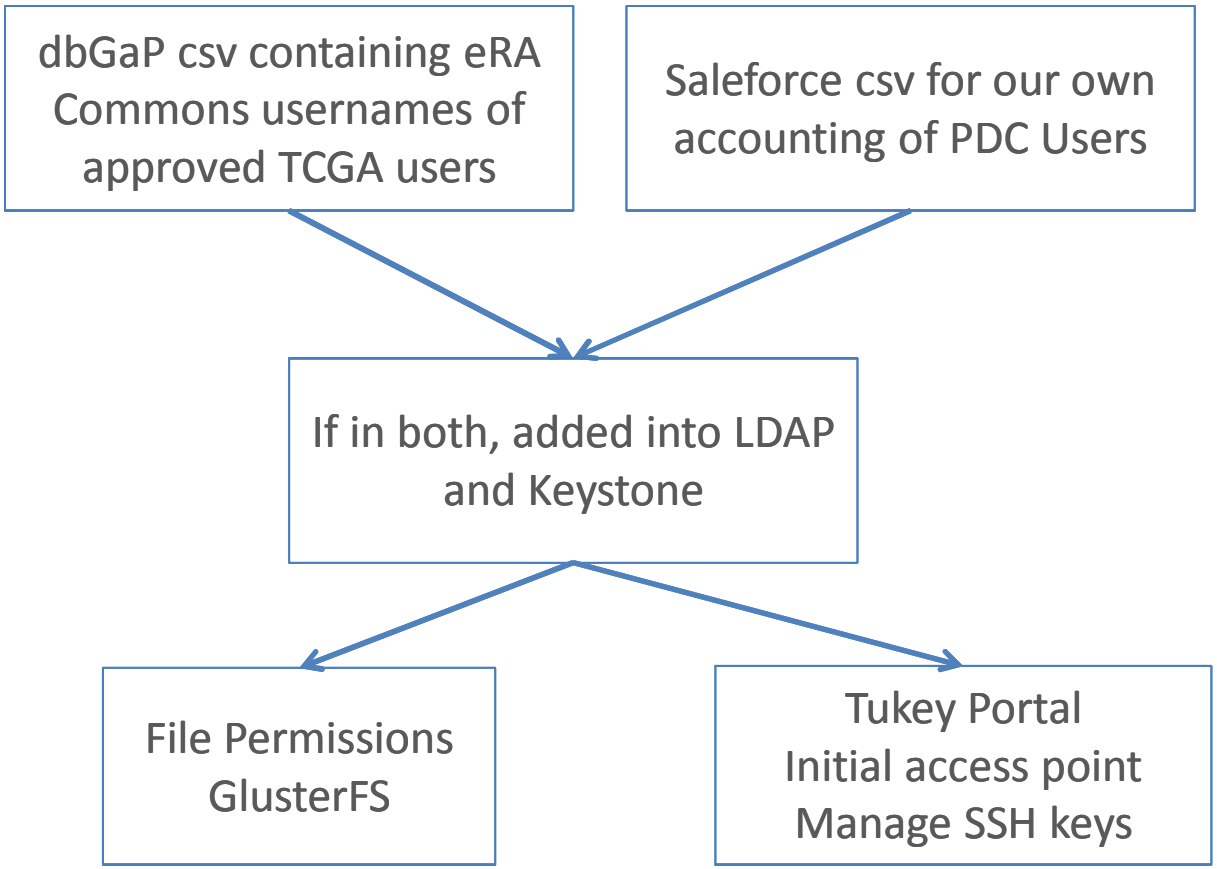
Institute for Genomics & Systems Biology

THE UNIVERSITY OF CHICAGO

Bionimbus-PDC Overview



Authentication Overview



Current Basic Workflow

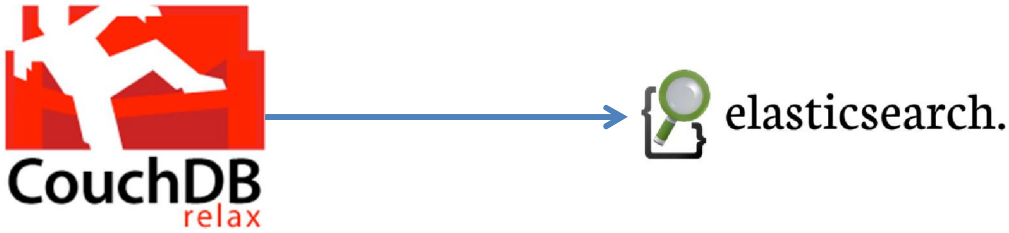
- Similar to other OSDC resources
- Obtain an account
  - TCGA data so requires eRA commons account and dbGaP access
- Login into Tukey Web Console with eRA commons credentials
- Set up key pairs in Tukey
- Start VMs
  - Plain Ubuntu or custom images
  - Automatically mounts your home dir and shared data inside VM
- Login (ssh) to bionimbus-pdc.opensciencedatacloud.org
  - Home directories and shared data stored on GlusterFS
- Login (ssh) to VMs and perform analysis
  - Can install software packages
  - Can save VM image for future use

Security

- Documented in a System Security Plan (SSP)
- Operates at a FISMA moderate level
- Some highlights:
  - All user commands and files touched are recorded
  - All traffic goes through a monitored proxy
  - No root access on VMs, selected sudo access
  - State monitored by Nagios, automated alerts
  - Strict use of key pairs for SSH access

Metadata Services

<u>JSON Store</u> → <u>Indexing/Searching</u> → <u>User Access/Security</u>



Tukey

TCGA Metadata
- disease, sequencing center, etc.

System Metadata
- md5 sums, existence on gluster, etc.

Harnessing search engine technology
Specific queries as well as ranked / fuzzy queries

GUI and APIs

OPEN CLOUD CONSORTIUM

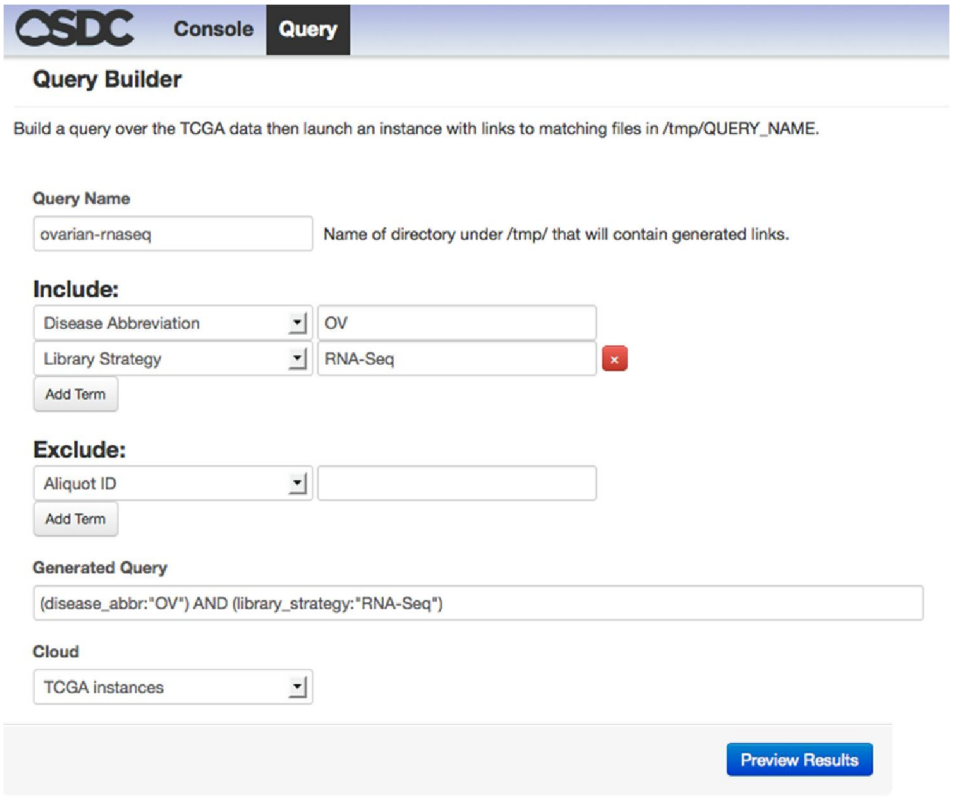
OPEN SCIENCE DATA CLOUD

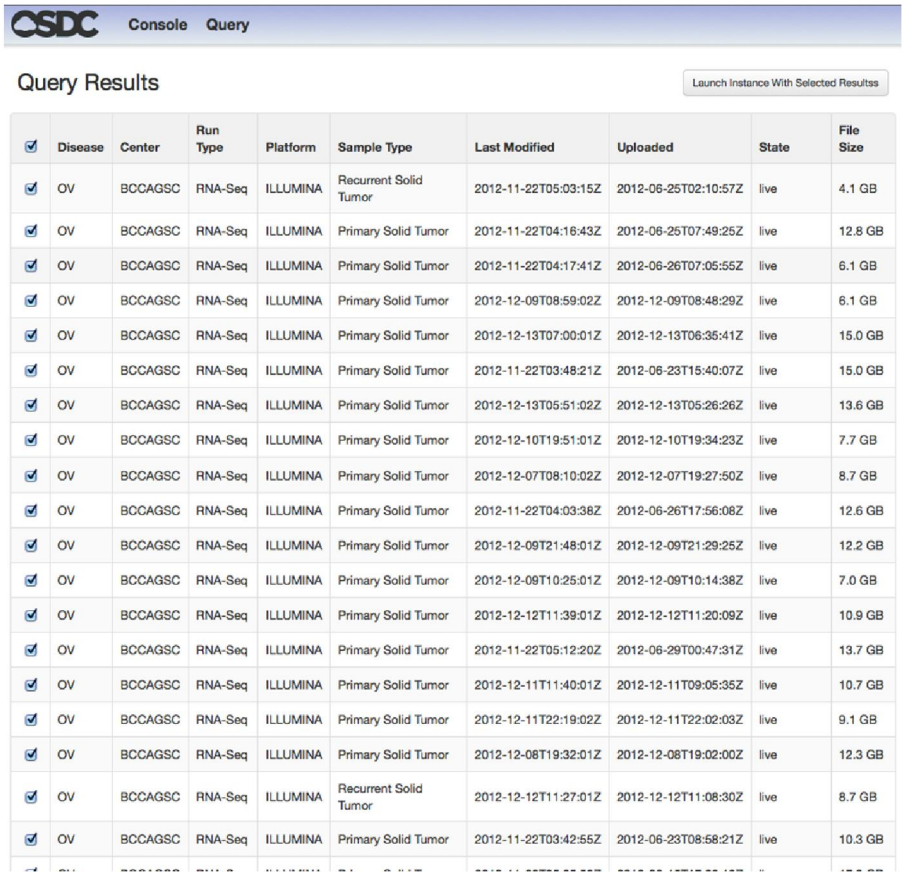
Institute for Genomics & Systems Biology

THE UNIVERSITY OF CHICAGO

Tukey: Metadata Query Tool

- Prototype with TCGA data





Disease	Center	Run Type	Platform	Sample Type	Last Modified	Uploaded	State	File Size
OV	BCCAGSC	RNA-Seq	ILLUMINA	Recurrent Solid Tumor	2012-11-22T05:03:15Z	2012-06-25T02:10:57Z	live	4.1 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-11-22T04:16:43Z	2012-06-25T07:49:25Z	live	12.8 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-11-22T04:17:41Z	2012-06-26T07:05:55Z	live	6.1 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-09T08:59:02Z	2012-12-09T08:48:29Z	live	6.1 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-13T07:00:01Z	2012-12-13T06:35:41Z	live	15.0 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-11-22T03:48:21Z	2012-06-23T15:40:07Z	live	15.0 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-13T05:51:02Z	2012-12-13T05:26:26Z	live	13.6 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-10T19:51:01Z	2012-12-10T19:34:23Z	live	7.7 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-07T08:10:02Z	2012-12-07T19:27:50Z	live	8.7 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-11-22T04:03:38Z	2012-06-26T17:56:08Z	live	12.6 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-09T21:48:01Z	2012-12-09T21:29:25Z	live	12.2 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-09T10:25:01Z	2012-12-09T10:14:38Z	live	7.0 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-12T11:39:01Z	2012-12-12T11:20:09Z	live	10.9 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-11-22T05:12:20Z	2012-06-29T00:47:31Z	live	13.7 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-11T11:40:01Z	2012-12-11T09:05:35Z	live	10.7 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-11T22:19:02Z	2012-12-11T22:02:03Z	live	9.1 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-12-08T19:32:01Z	2012-12-08T19:02:00Z	live	12.3 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Recurrent Solid Tumor	2012-12-12T11:27:01Z	2012-12-12T11:08:30Z	live	8.7 GB
OV	BCCAGSC	RNA-Seq	ILLUMINA	Primary Solid Tumor	2012-11-22T03:42:55Z	2012-06-23T08:58:21Z	live	10.3 GB

OPEN CLOUD CONSORTIUM

OPEN SCIENCE DATA CLOUD

Institute for Genomics & Systems Biology

THE UNIVERSITY OF CHICAGO

Torque Cluster Launch

- Launch a “elastic” cluster
- Specify the number of compute nodes and VM image to use
- Launches a small VM with a headnode image
- Launches the number of nodes specified as compute nodes

Conclusions and Future Work

- Created a secure cloud computing environment for human genomics research
- More Data
- VMs with analysis pipelines
- System to produce daily analysis based as new data is ingested
  - How do we reanalyze petabytes daily?
  - Scalability

Thank You

OPEN CLOUD CONSORTIUM

OPEN SCIENCE DATA CLOUD

Institute for Genomics & Systems Biology

THE UNIVERSITY OF CHICAGO