Computing and Data Management at The DOE Joint Genome Institute

Kjiersten Fagnan
NERSC-JGI Engagement Lead

October 27, 2014
DOE Joint Genome Institute

- Walnut Creek, CA
- facility opened in 1999
- 250 employees
- ~$69M annual expenses

Mission:

DOE JGI, Serving as a genomic user facility in support of the DOE missions:

bioenergy, carbon cycling, & biogeochemistry
• National Energy Research Scientific Computing Center

  – Established 1974, first unclassified supercomputer center

  – Original mission: to enable computational science as a complement to magnetically controlled plasma experiment

• Today’s mission: Accelerate scientific discovery at the DOE Office of Science through high performance computing and extreme data analysis
JGI’s Compute Cluster at NERSC

User Access
- Command Line
- Scheduler
- Service

ssh genepool.nersc.gov

http://...jgi-psf.org

login nodes

compute nodes
400 8-core GigE Nodes
225 16-core IB Nodes
17 High Memory Nodes

gpint nodes

high priority & interactive nodes

fpga

web services

database services

/filesystems

/homes

/scratch

/seqfs

/dataNarchive

/software
JGI Data Storage

/projectb
- 2.6PB
- SCRATCH/Sandboxes = "Wild West"
- Write here from compute jobs

WebFS
- small file system for web servers
- mounted on gpwebs and in xfer queue

Working Directories
2.6 PB

Web Services
100TB

DnA
- Project directories, finished products
- NCBI databases, etc
- Read-only on compute nodes, read-write in xfer queue

SeqFS
- 500 TB File system
- accessible to sequencers at JGI

Shared Data
1 Pb

Sequencer Data
500TB
The NERSC archive is a Hierarchical Storage Management system (HSM)

- Highest performance requirements and access characteristics at top level
- Lowest cost, greatest capacity at lower levels
- Migration between levels is automatic, based on policies
Path of Sample Processing

SDM → RQC → QC Assembly → QD (Quality Draft) → IMG M/ER

SDM

RQC

QC Assembly

QD (Quality Draft)

IMG M/ER

Collaborator

LANL

Microbial Finished
Microbial HQD
Microbial Std/Quick Draft
Fungal Std Draft
Fungal Minimal Draft
Plant
Metagenome
RNAseq

Phytozome

Mycocosm
Data Sharing Between Groups

- PMO
- SDM
- QAQC/RQC
- GAAG
- Plant
- MEP
- RnD
- Fungal
- IMG
- MGM

Web Services: Mycocosm, Phytozome, IMG M/ER

External Collaborators
New System – JGI Archive and Metadata Organizer

PMO

SDM

QAQC/RQC

Genome Portal

External Collaborators

Web Services (Mycocosm, Phytozome, IMG M/ER)

JAMO

GAAG

Plant

MEP

RnD

Fungal

IMG

MGM
Thank you!
de novo Assembly

Samples taken from the wild

Microbial DNA extracted and sequenced

Reads (short segments of DNA) generated by sequencer

Genome Assembly done by AllPaths-LG

Complete Genome
de novo Assembly

Microbial DNA extracted and sequenced

Reads (short segments of DNA) generated by sequencer

Genome Assembly done by AllPaths-LG

Samples taken from the wild

Contigs