JAX: Exploring The Galaxy

Glen Beane, Senior Software Engineer

The Jackson Laboratory
Bar Harbor, Maine

- Non-profit genetics research
- Founded in 1929
- 36 principal investigators
- 1,300+ employees
- $200 million budget
- NCI-designated Cancer Center

Mission

We discover the genetic basis for preventing, treating and curing human disease, and we enable research and education for the global biomedical community.

Vision

Our mouse models and genetics research lead the world to solutions for cancer and other complex and intractable diseases.

Scientific Computing Group: Who We Are

- Part of core software engineering and statistical analysis service (not IT)
  - scientific software development
  - High Performance Computing
  - not Linux/Unix system administrators
  - domain expertise

The Jackson Laboratory

Why Galaxy?

- Needed a HTS analysis platform
  - make routine analysis accessible to scientists
  - preferred local installation vs. hosted
  - wanted to integrate with existing HPC resources (using TORQUE/Moab)
  - looked at GenomeQuest, GenePattern and others
- Open Source (no license cost, customizable)
- Out of the box support for HTS tools
- Active community (users and developers)
- Facilitates collaboration
  - Share Histories, Data, Workflows

The Jackson Laboratory

Why HTS?

- RNA-seq
  - greater fidelity of expression levels
  - unbiased by microarray spot sequences
  - alternative splicing / RNA editing
- ChIP-seq
  - unbiased
  - new approaches for epigenetics
- Targeted re-sequencing
  - mutagenesis projects
  - spontaneous mutations in the production colony

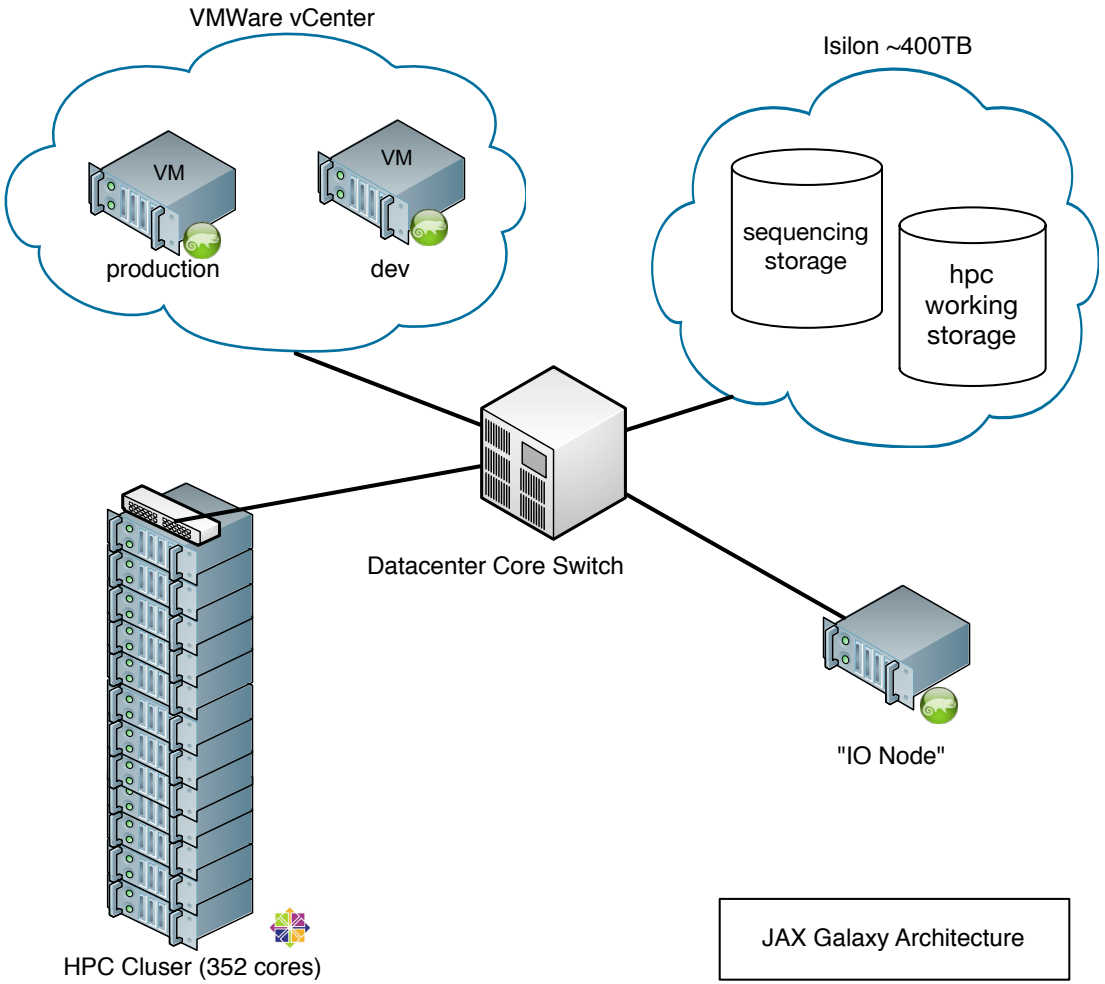
The Jackson Laboratory

What we are doing with Galaxy

- High throughput sequencing analysis
  - RNA-Seq
  - DNA-Seq
  - ChIP-Seq
- Other Genomic Analysis
  - e.g., Array Genotyping (Diversity Array, MUGA)
  - developing/wrapping new tools

The Jackson Laboratory

Our Installation



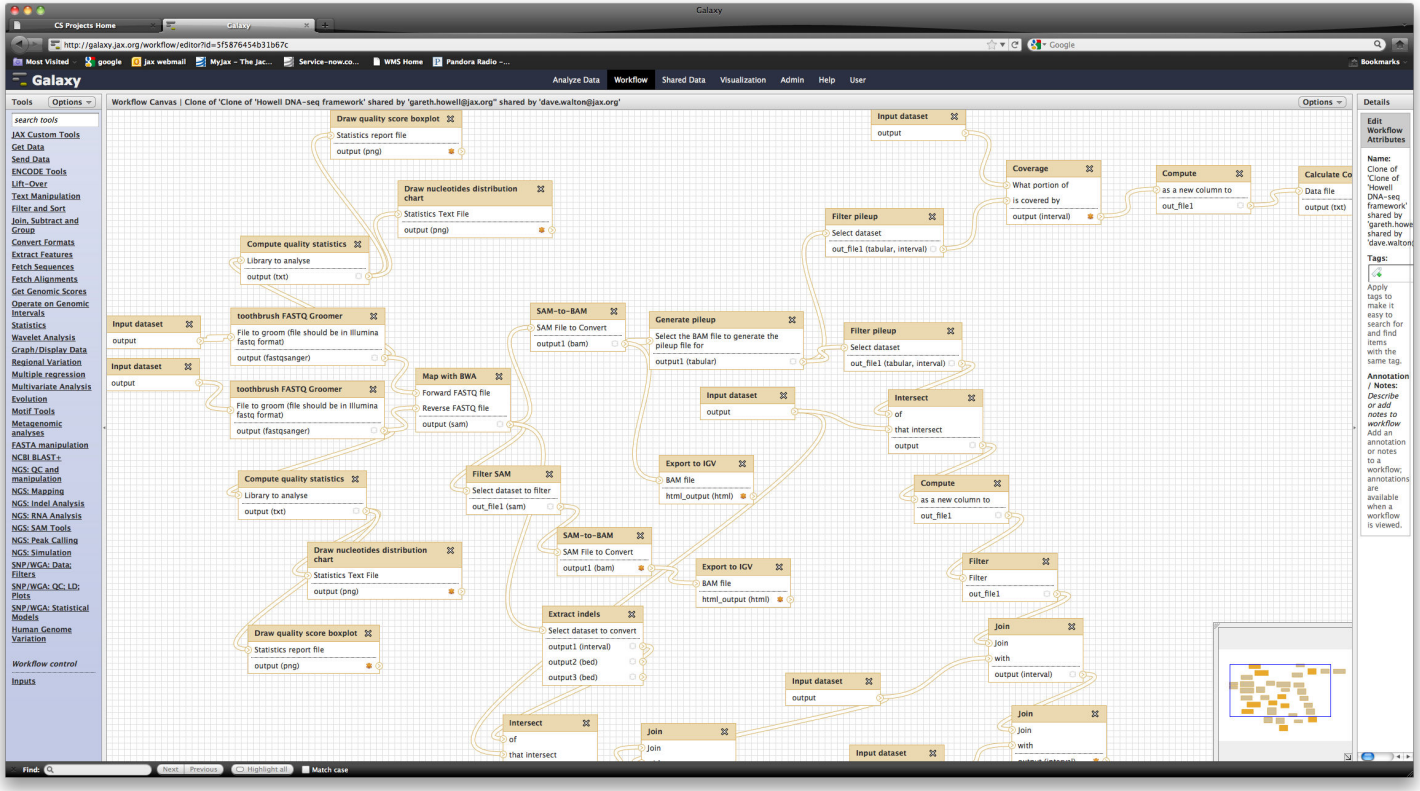
The Jackson Laboratory

What we've been up to so far

- Custom Tools & Workflows
  - e.g., Array Genotyping Workflow
    - custom "get data" tool
    - group by SNP probe set tool
    - genotyping tools (Alchemy, MDG)
    - EMMA (mixed-model association mapping)
- RNA-Seq and DNA-Seq workflows, Whole-Genome workflows
- "Toothbrush" (custom "FASTQ groomer" written in C)
- Search Mouse SNPs Tool (Sanger 17 strains)
- Tools for custom statistical calculations on tabular data files
- HDF5 support ("sniffable")

The Jackson Laboratory

Users creating non-trivial workflows



user would not have done this from the command line on our cluster

The Jackson Laboratory

Challenges

- Importing Data!
  - ftp uploads a big help!
  - using “upload directory of files” heavily
  - plan to automate uploads
- Sparse developer docs (e.g. API)
- Truncated error messages from tools
- difficulty managing experiments w/ large numbers of samples (e.g. run 40 samples through same workflow)
  - output file names difficult to match up with original sample names (get 40 “N toolX on Y” in history)
  - merging results from many workflows is manual
  - can’t automatically run multiple *pairs* of files through same workflow

The Jackson Laboratory

Wish List

- Input file name or parameter value as variable in workflow (we want to name output files based on initial input name)
- Auto delete intermediate files in WF (not just hide)
- Tools with associated roles
- Reduction
  - merge results from multiple WFs (with custom “Reduce” tool or something standard like simple concatenation)
- more developer documentation
- more reports (e.g. disk space per user, active/inactive data files, etc)
- “favorite tools”
- list tool versions

The Jackson Laboratory

Acknowledgements

Dave Walton, Manager Scientific Computing

Keith Sheppard & Matt Vincent, Software Engineers, Center For Genome Dynamics

Rich Brey & Michael Genrich, Linux Systems Administrators, IT

Matt Hibbs, PhD – Assistant Professor

Joel Graber, PhD – Associate Professor

Gary Churchill, PhD – Professor

Carol Bult, PhD – Professor

Gareth Howell, PhD – Research Scientist (workflow image)

The Jackson Laboratory