

Real-Time Next-Generation Sequencing (NGS) in the Classroom using Galaxy

Josephine Palencia
Alex Ropelewski

March 17, 2015

Acknowledgement

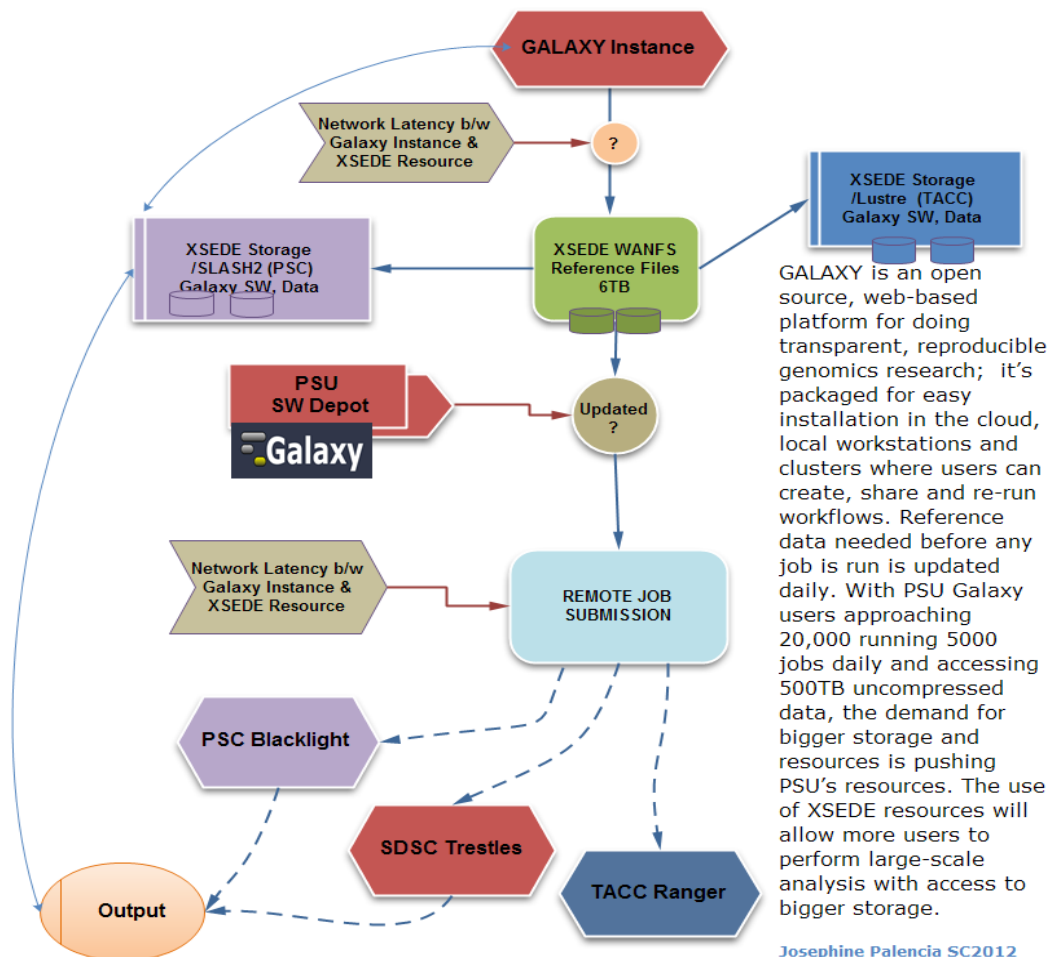
- Philip Blood
- Sergiu Sanieleveci
- Robert Light
- Brian Johanson
- Joel McManus



Content

GALAXY Using XSEDE Compute & Storage Resources

- **Galaxy System Setup**
 - VM
 - Shared filesystem
 - Accounts management
 - Job submission
 - Scaling and Load balancing
- **Class Session**
 - Tools and Workflow
 - Preparation
 - Actual sessions
- **Discussion/Future**



GALAXY is an open source, web-based platform for doing transparent, reproducible genomics research; it's packaged for easy installation in the cloud, local workstations and clusters where users can create, share and re-run workflows. Reference data needed before any job is run is updated daily. With PSU Galaxy users approaching 20,000 running 5000 jobs daily and accessing 500TB uncompressed data, the demand for bigger storage and resources is pushing PSU's resources. The use of XSEDE resources will allow more users to perform large-scale analysis with access to bigger storage.

Josephine Palencia SC2012
Galaxy Eposter 9/1/2012

Galaxy System



GalaxyVM-1



GalaxyVM-2



GalaxyVM-3..

VM FARM: *minimal resource, lightweight*
XEN 4core, 4GB mem, 10GB disk

NFS

`/usr/local/packages/galaxy/{tools, tool-data}`

Tools: *xml wrappers & interfaces per tool*

Tool-data: *.loc files point to ref-data*

Tool_dependencie: *sets environment*

LUSTRE

`/usr/local/packages/galaxy/{ref_files,/database}`

Database: *user data files, jobs, histories, metadata*

Ref_files: *genome annotations, sequence databases;
each tool with multiple ref files; pointed to by .loc files*



Galaxy System



Remote Job Submission
via SSH (Secure Shell/CLI)

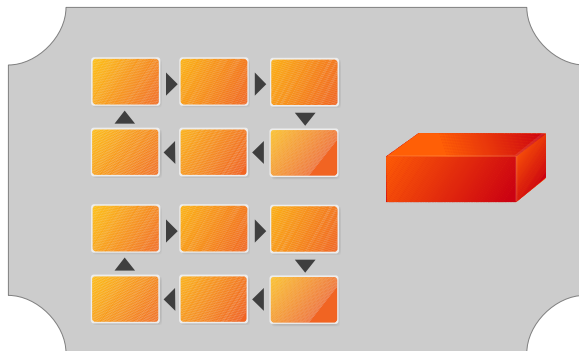


Blacklight SGI UV 1000
30 blades; 1 blade (16core, 128GB)

Standalone Paste-based: simpler, not as resilient, load balancing- round-robin; no dynamic scaling

UWSGI: better {performance, scalability, fault tolerance}, easier process management, server restart

Paste-based Scaling/Load Balancing: 10 web handlers:1 job handler



```
<Proxy balancer://galaxy>
BalancerMember http://galaxy-dev.psc.xsede.org:8080
BalancerMember http://galaxy-dev.psc.xsede.org:8081

cmd="cd $RUN_IN && sh run.sh --server-name=web0 --daemon"
cmd="cd $RUN_IN && sh run.sh --server-name=web1 --daemon"
cmd="cd $RUN_IN && sh run.sh --server-name=job0 --daemon"
```

PSC LOGIN

Login here and you will be redirected to your destination:

USERNAME

PASSWORD

[Forgot my password](#)

Galaxy-dev.psc.xsede.org



AuthType
mod_auth_pubtkt

PSC User auth DB

The application you are visiting can generate charges to a PSC allocation. Please select the award that should be charged if you choose to utilize features that generate charges:

Select	Grant Number	PI	Title	Balance	End Date
<input type="radio"/>	SYS120007P	Josephine Palencia	Galaxy test grant	-752	11/15/2015
<input type="radio"/>	SYS110010P	J Ray Scott	new xsede staff project	899,955	09/20/2015
<input checked="" type="radio"/>	SYS110018P	Nancy Wilkins-Diehr	XSEDE 1.5 AUSS-Communities	10,894,887	09/20/2015
<input type="radio"/>	ASC110007P	Anirban Jana	MATLAB testing	3,926	07/15/2015

Select Checked Award

- Simple approach to Web SSO
- Cookie (session)-based
- Tickets generated by central server and passed to web servers

Galaxy System

Tools

search tools

BLACKLIGHT TOOLS

[PBS Monitor](#)

[Brashear File Tools](#)

[Trinity: NGS RNA Seq](#)

GALAXY TOOLS

[Get Data](#)

[Send Data](#)

[ENCODE Tools](#)

[Lift-Over](#)

[Text Manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Convert Formats](#)

[Extract Features](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

[Get Genomic Scores](#)

[Operate on Genomic Intervals](#)

[Statistics](#)

[Wavelet Analysis](#)

[Graph/Display Data](#)

[Regional Variation](#)

[Multiple regression](#)

[Multivariate Analysis](#)

[Evolution](#)

[Motif Tools](#)

[Multiple Alignments](#)

[Metagenomic analyses](#)

[FASTA manipulation](#)

[NGS: QC and manipulation](#)

[NGS: Mapping](#)

GALAXY Using XSEDE Compute & Storage Resources

The diagram illustrates the architecture of GALAXY using XSEDE resources. It shows the flow from the GALAXY Instance to XSEDE Storage and XSEDE WANFS Reference Files. The PSU SW Depot provides updates to the Remote Job Submission process. Network latency is noted between the GALAXY Instance and XSEDE Resources, and between the GALAXY Instance and the Remote Job Submission process. The PSC Blacklight is also shown as a component of the system.

GALAXY is an open source, web-based platform for doing transparent, reproducible genomics research; it's packaged for easy installation in the cloud, local workstations and clusters where users can create, share and re-run workflows. Reference data needed before any job is run is updated daily. With PSU Galaxy users approaching 20,000 running 5000 jobs daily and accessing 500TB uncompressed data, the demand for bigger storage and resources is pushing PSU's resources. The use of XSEDE resources is a key part of the solution.

History

post-test3
3.2 GB

10: : Varscan variants

9: MPileup on data 5 (log)

8: MPileup on data 5

7: : alignment summary

6: MarkDups X_oc_sample_name_cc : mapped reads, no dups.html

5: MarkDups X_oc_sample_name_cc : mapped reads, no dups.bam

4: : mapped reads

3: Map with BWA for Illumina on data 2 and data 2: mapped reads

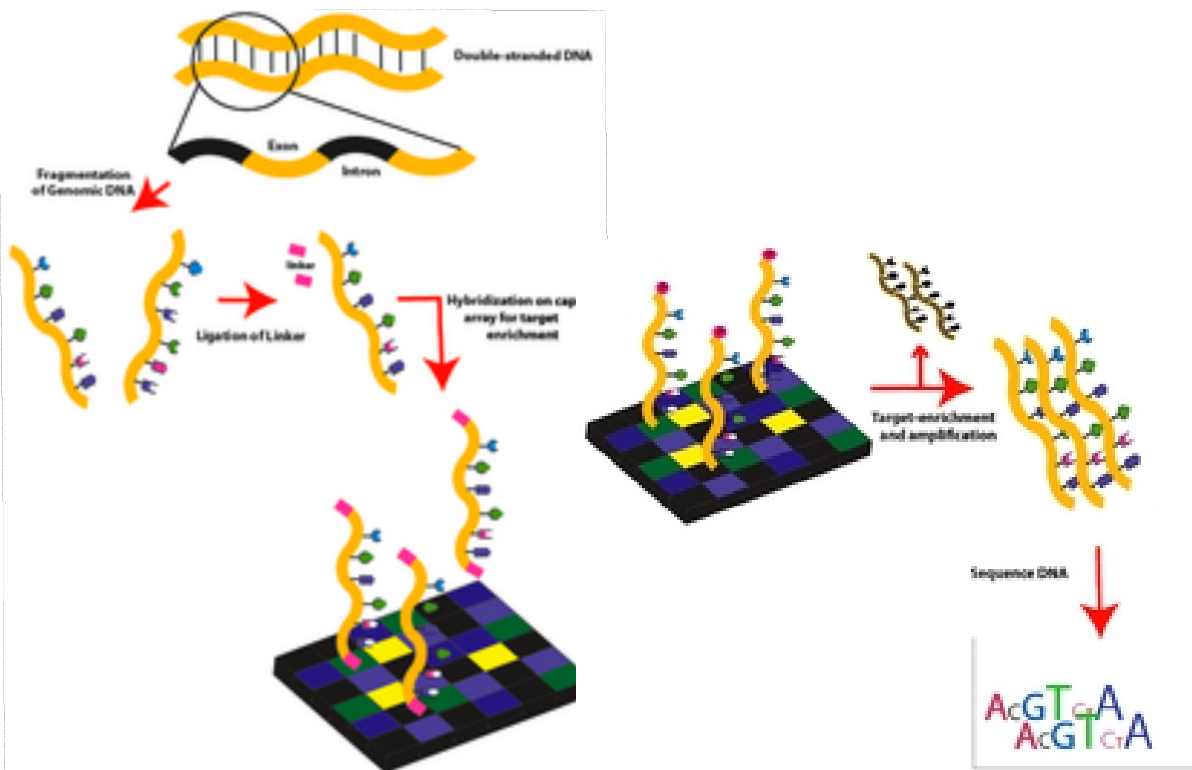
2: Brashear /brashear/mcmanus/exomeData/HG 01967_chr20_R2.fastq

1: Brashear /brashear/mcmanus/exomeData/HG 01967_chr20_R2.fastq

Sequencing Tools

Exome Sequencing:

- technique for sequencing all protein-coding genes in a genome called **exomes**



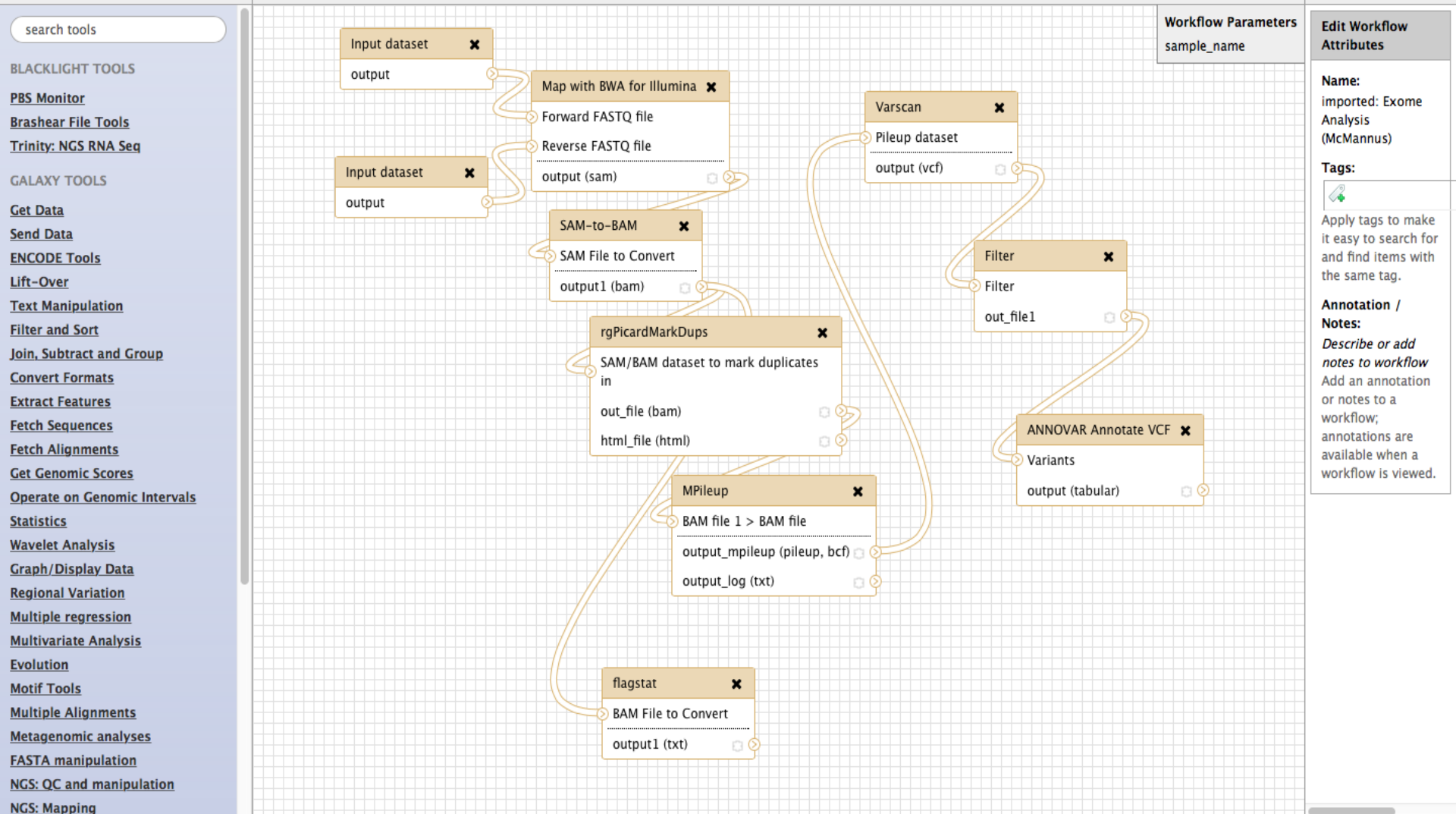
Exomes

- regions of genome that get transcribed and spliced into protein coding RNA transcripts
- represents only 1% of human genome yet contains 85% of mutations that cause diseases
- cost effective way to search for disease causing mutations for very rare genetic disorders

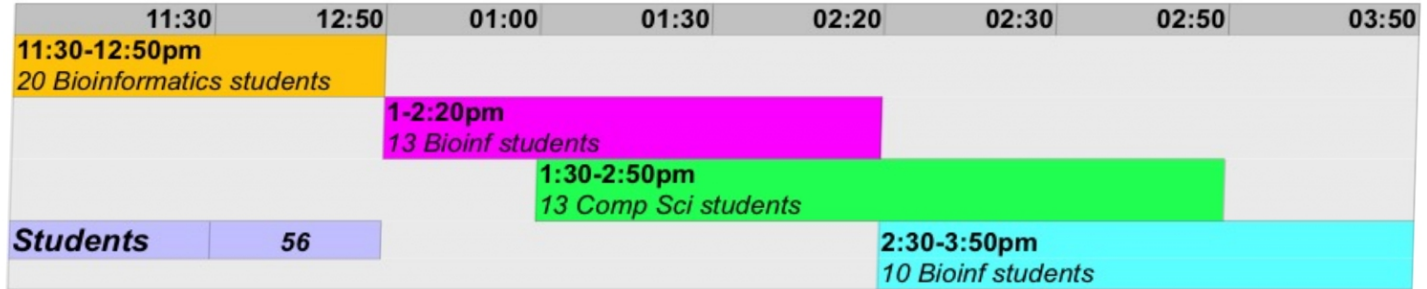
Tools and Wrappers

- annovar/2014-11-12
- bowtie/1.1.0
- bwa/0.7.10
- picard-tools/1.119
- samtools/1.1.0
- varscan/2.3.630

Workflow



Preparation



- 30 Blacklight blades reserved → Up to 30 students with overlap
- Simple paste based method for web/job scaling/load balancing
- Data used for exome sequencing underwent **3 size reductions**
Run times (1hr → 30min → 15min)
- Complete workflow took **15min** to complete
- Ran **2 simulations** of 50 workflows submitted with 30 running on Blacklight
- Scaling and Load balancing went well with workflows/jobs finishing on time

Actual Class Sessions



- 15min workflows → 45min - 1 hour to run
- Jobs running from previous class overlapped with next class
- 100+ remote workflows submitted, 30 running, rest are queued
- VM was steady- handled scaling web/job scaling and load balancing well
- 2nd session week- Wanted to use data 10x larger → won't finish on time

Discussion/Future

- 「 Need faster compute system:
15min → < 5min
- 「 Need faster, dedicated HPC filesystem for the class
vs shared
- 「 Use uwsgi for production use
 - dynamic scaling, load balancing
- 「 Use local job submission to reduce overhead