

Galaxy for NGS Data Analysis

Matt Shirley

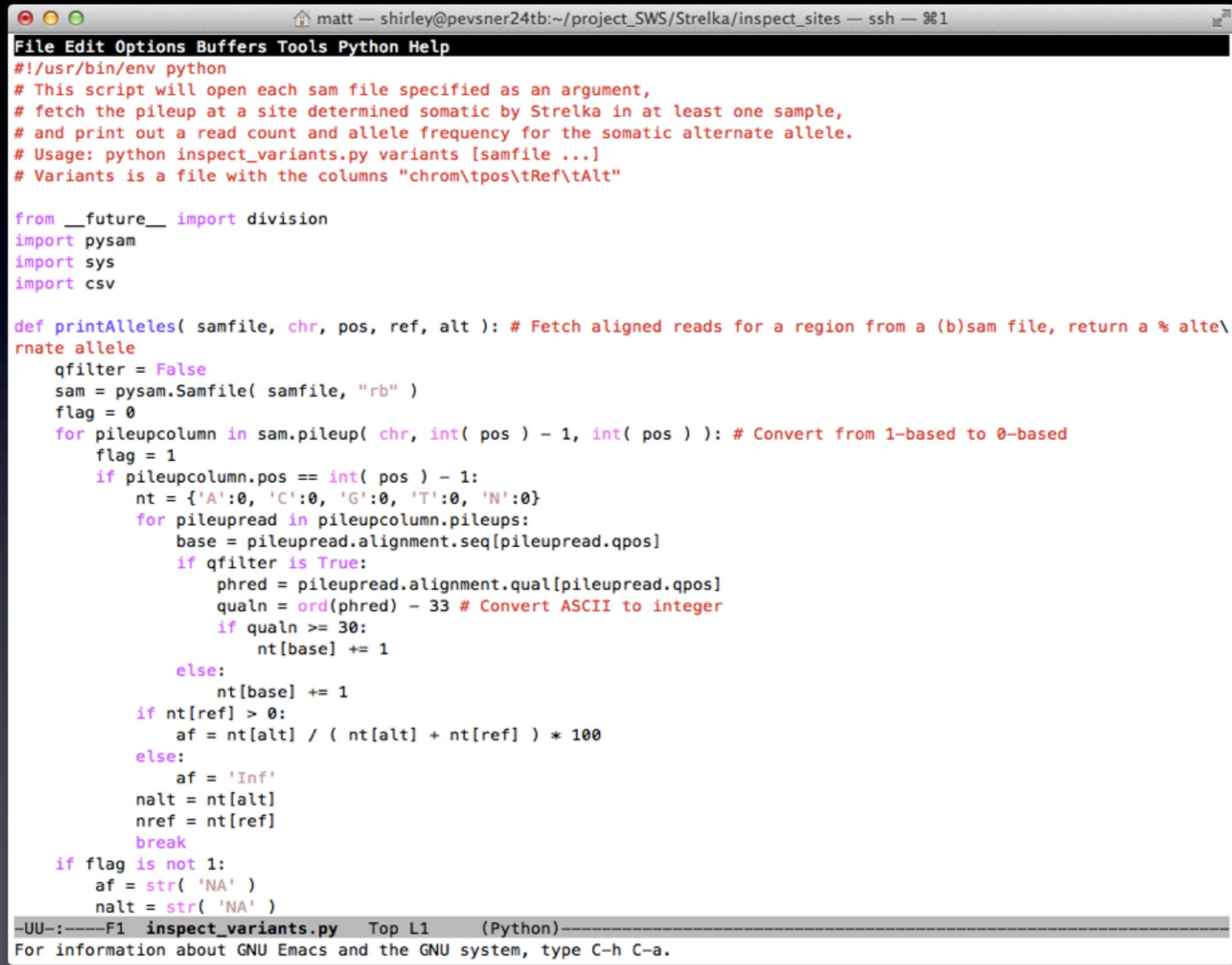
Johns Hopkins Medicine
Pevsner Lab

Slides available at <http://mattshirley.com/presentations>

Contents

- What is Galaxy?
- Interface elements
- Retrieving data
- Creating and running workflows
- A FASTQ quality statistics workflow
- Galaxy on Amazon Web Services (AWS)
 - Automatic configuration through cloudlaunch
 - Monitoring your AWS charges
 - (optional) Manual configuration through AWS console

Who wants to do this? :(



```
matt — shirley@pevsner24tb:~/project_SWS/Strelka/inspect_sites — ssh — #1
File Edit Options Buffers Tools Python Help
#!/usr/bin/env python
# This script will open each sam file specified as an argument,
# fetch the pileup at a site determined somatic by Strelka in at least one sample,
# and print out a read count and allele frequency for the somatic alternate allele.
# Usage: python inspect_variants.py variants [samfile ...]
# Variants is a file with the columns "chrom\tpos\tRef\tAlt"

from __future__ import division
import pysam
import sys
import csv

def printAlleles( samfile, chr, pos, ref, alt ): # Fetch aligned reads for a region from a (b)sam file, return a % alternate allele
    qfilter = False
    sam = pysam.Samfile( samfile, "rb" )
    flag = 0
    for pileupcolumn in sam.pileup( chr, int( pos ) - 1, int( pos ) ): # Convert from 1-based to 0-based
        flag = 1
        if pileupcolumn.pos == int( pos ) - 1:
            nt = {'A':0, 'C':0, 'G':0, 'T':0, 'N':0}
            for pileupread in pileupcolumn.pileups:
                base = pileupread.alignment.seq[pileupread.qpos]
                if qfilter is True:
                    phred = pileupread.alignment.qual[pileupread.qpos]
                    qualn = ord(phred) - 33 # Convert ASCII to integer
                    if qualn >= 30:
                        nt[base] += 1
                else:
                    nt[base] += 1
            if nt[ref] > 0:
                af = nt[alt] / ( nt[alt] + nt[ref] ) * 100
            else:
                af = 'Inf'
            nalt = nt[alt]
            nref = nt[ref]
            break
        if flag is not 1:
            af = str( 'NA' )
            nalt = str( 'NA' )
-UU-----F1 inspect_variants.py Top L1 (Python)-----
For information about GNU Emacs and the GNU system, type C-h C-a.
```

Wouldn't you rather do this?

facebook

Introducing Graph Search

Find more of what you're looking for through your friends and connections. [Learn More ▾](#)

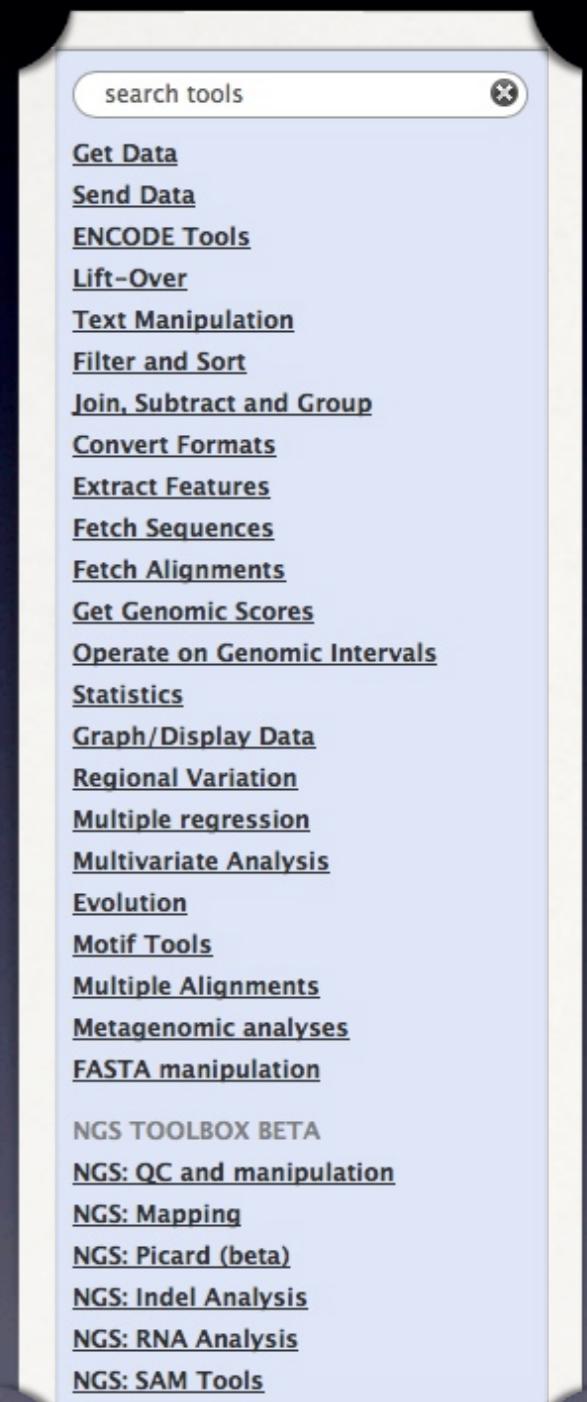


A photograph of a young boy with blonde hair, wearing a light-colored jacket, standing outdoors. He is holding up two small fish by their tails. A large black play button is overlaid in the center of the image, suggesting it is a video thumbnail.

What is Galaxy?

Galaxy is a collection of bioinformatics tools for:

- data conversion and manipulation
- statistical analysis
- next generation sequencing analysis
- much, much more!....



matt.shirley@jhmi.edu

What is Galaxy?

Citation

If you use this tool, please cite [Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A; Galaxy Team. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15;26\(14\):1783-5.](#)

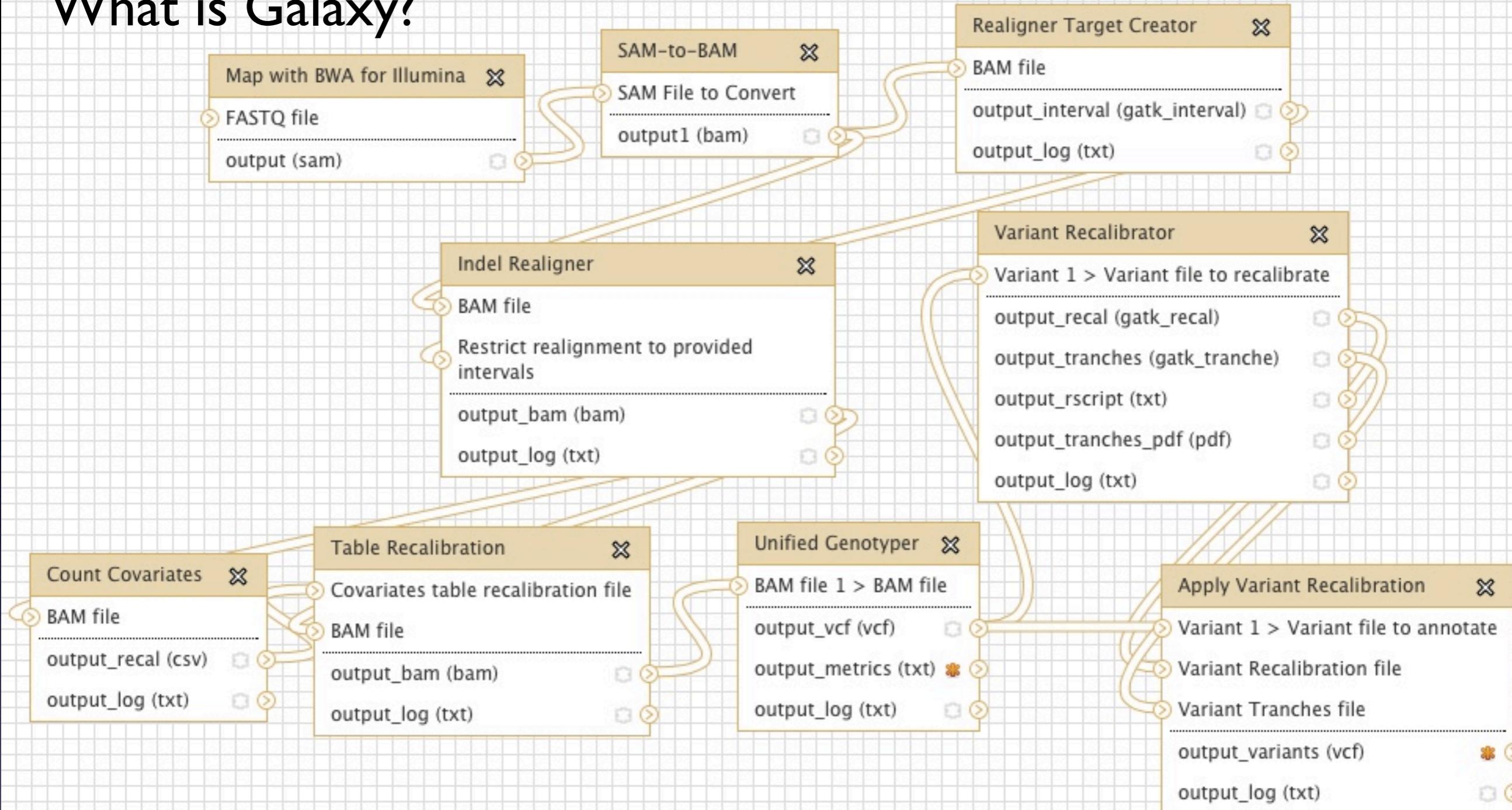
- Based on peer-reviewed and open-source implementations of each tool
- Galaxy provides integration with useful tools, targeted toward “bench” scientists
- Unified and consistent interface for easy exploration

What is Galaxy?

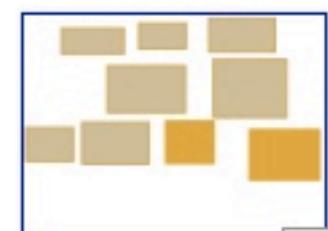
- Data library: management and sharing for collaborative analysis
- Sequencer interface for direct intake of raw data

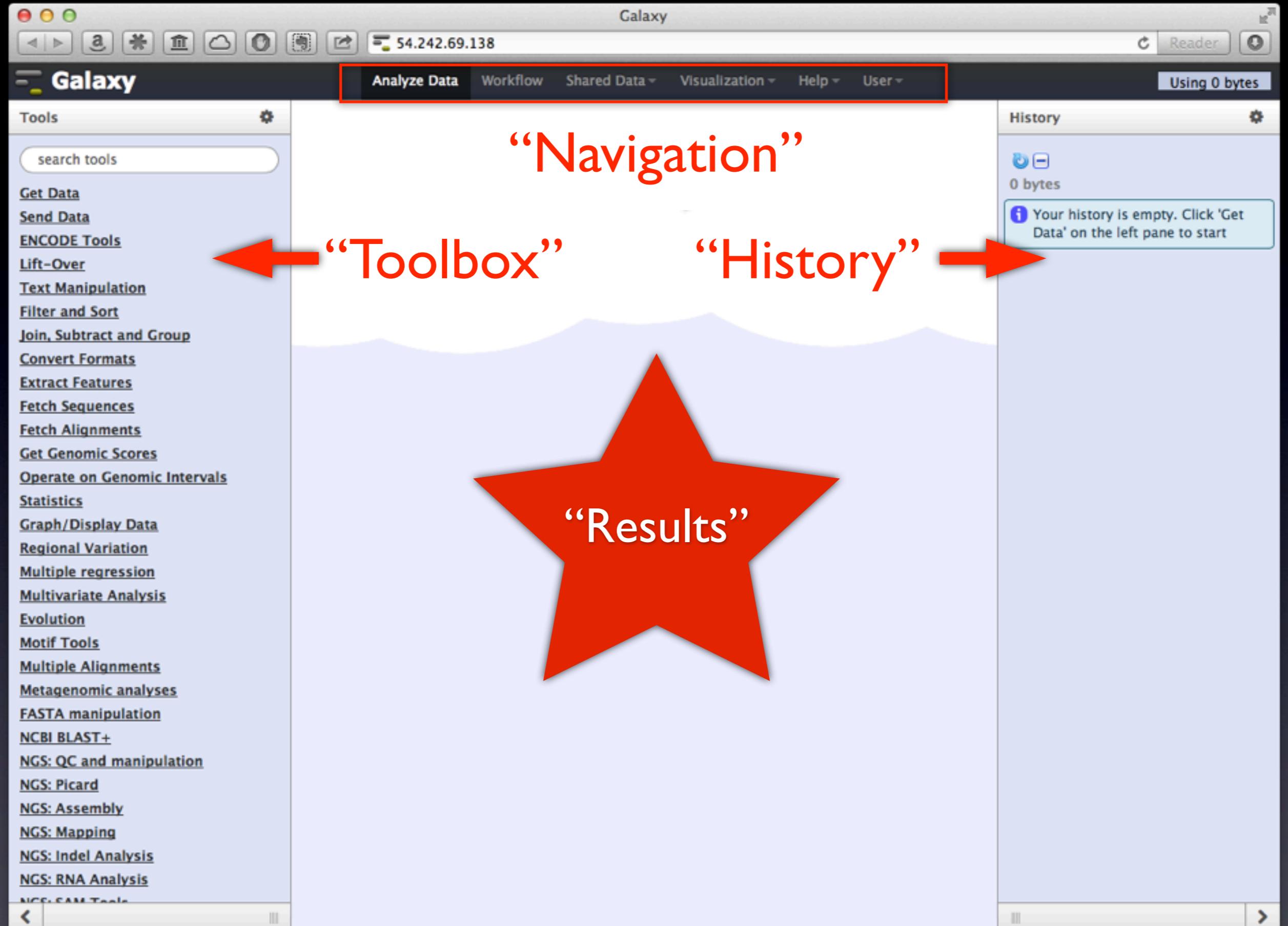
Data Libraries	
<input type="text" value="search dataset name, info, message, dbke"/> <input type="button" value=""/>	
Advanced Search	
Data library name ↓	Data library description
1000 Genomes	Data from the 1000 Genomes Project FTP site
AC-exome	
Bushman	Data for Nature Letter "Complete Khoisan and
ChIP-Seq Mouse Example	Data used in examples that demonstrate analy
Chobi	
CloudMap	Contains userguide, reference files, and config
Codon Usage Frequencies	
Coleman	IonPGM
Denisovan sequences	Files from 'A high-coverage genome sequence
Erythroid Epigenetic Landscape	Dynamics of the epigenetic landscape during e
Evolutionary Trajectories in a Phage	Experimental evolution (Illumina)
GATK	
GCAT	Consortium
Genome Diversity	Nucleotide polymorphisms for several threatener
guru 1000GP	
He-2010	
Heteroplasmy	Data for Genome Biology 2011 manuscript
iGenomes	Selected files from Illumina iGenomes collectio

What is Galaxy?



Workflows that enable
reproducible research





The “toolbox”



Contains links for :

- retrieving (“get”) data
- manipulating data (lift-over, filter, sort, set operations, format conversions)
- data analysis (statistics, sequence alignment, variant calling and annotation)

“Get” data

In addition to uploading files from your computer, you may:

- Choose a file in the “shared data” library
- Import from UCSC, EBI SRA, BioMart, CBI Rice Map, modENCODE, Ratmine, Flymine, YeastMine, WormBase, EuPath, Microbial Genome Project, EncodeDB, EpiGRAPH, HbVar, GenomeSpace

Galaxy

54.242.69.138

Analyze Data Workflow Shared Data Visualization Help User Using 0 bytes

Tools

search tools

Get Data

- [Upload File from your computer](#)
- [UCSC Main table browser](#)
- [UCSC Test table browser](#)
- [UCSC Archaea table browser](#)
- [BX main browser](#)
- [Get Microbial Data](#)
- [BioMart Central server](#)
- [BioMart Test server](#)
- [CBI Rice Mart rice mart](#)
- [GrameneMart Central server](#)
- [modENCODE fly server](#)
- [Flymine server](#)
- [Flymine test server](#)
- [modENCODE modMine server](#)
- [Ratmine server](#)
- [YeastMine server](#)
- [metabolicMine server](#)
- [modENCODE worm server](#)
- [WormBase server](#)
- [Wormbase test server](#)
- [EuPathDB server](#)
- [EncodeDB at NHGRI](#)
- [EpiGRAPH server](#)
- [EpiGRAPH test server](#)

Upload File (version 1.1.3)

File Format:

Auto-detect

Which format? See help below

File:

Choose File no file selected

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq.gz

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
------	------	------

Please [create](#) or [log in](#) to a Galaxy account to view files uploaded via FTP.

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at localhost using your Galaxy credentials (email address and password).

Convert spaces to tabs:

Yes

Use this option if you are entering intervals by hand.

Genome:

Click to Search or Select

Execute

Auto-detect

The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabular, Wiggle, Bed and Interval (Bed with headers) formats. If your file is not detected properly as one of the known formats, it will likely need to be converted to one of these formats before it can be used.

History

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

Galaxy

Analyze Data Workflow Shared Data Visualization Help User Using 2.1 Mb

Tools

- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation
 - FASTQC: FASTQ/SAM/BAM
 - Fastqc: Fastqc QC using FastQC from Babraham
 - ILLUMINA FASTQ
 - FASTQ Groomer convert between various FASTQ quality formats
 - FASTQ splitter on joined paired end reads
 - FASTQ joiner on paired end reads
 - FASTQ Summary Statistics by column
 - ROCHE-454 DATA
 - Build base quality distribution
 - Select high quality segments
 - Combine FASTA and QUAL into FASTQ
 - AB-SOLID DATA
 - Convert SOLID output to fastq
 - Compute quality statistics for SOLID data
 - Draw quality score boxplot for SOLID data

History

1: https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq 2.1 Mb

Edit Attributes

Name: https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq

Info: uploaded fastq file

Database/Build: Click to Search or Select

Save Auto-detect

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

Convert to new format

Convert FASTQ files to seek locations

This will create a new dataset with the contents of this dataset converted to a new format.

Convert

Change data type

New Type: fastqsanger

This will change the datatype of the existing dataset but *not* modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

Save

54.242.69.138

Galaxy

Analyze Data Workflow Shared Data Visualization Help User Using 2.1 Mb

Tools

- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation
 - FASTQC: FASTQ/SAM/BAM
 - Fastqc: Fastqc QC using FastQC from Babraham
 - ILLUMINA FASTQ
 - FASTQ Groomer convert between various FASTQ quality formats
 - FASTQ splitter on joined paired end reads
 - FASTQ joiner on paired end reads
 - FASTQ Summary Statistics by column
 - ROCHE-454 DATA
 - Build base quality distribution
 - Select high quality segments
 - Combine FASTA and QUAL into FASTQ
 - AB-SOLID DATA
 - Convert SOLID output to fastq
 - Compute quality statistics for SOLID data
 - Draw quality score boxplot for SOLID data
- GENERIC FASTQ MANIPULATION

FASTQ Summary Statistics (version 1.0.0)

FASTQ File:
1: https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq

Execute

This tool creates summary statistics on a FASTQ file.

TIP: This statistics report can be used as input for the Boxplot and Nucleotides Distribution tools.

The output file will contain the following fields:

column = column number (1 to 36 for a 36-cycles read Solexa file)
count = number of bases found in this column.
min = Lowest quality score value found in this column.
max = Highest quality score value found in this column.
sum = Sum of quality score values for this column.
mean = Mean quality score value for this column.
Q1 = 1st quartile quality score.
med = Median quality score.
Q3 = 3rd quartile quality score.
IQR = Inter-Quartile range (Q3-Q1).
lW = 'Left-Whisker' value (for boxplotting).
rW = 'Right-Whisker' value (for boxplotting).
outliers = Scores falling beyond the left and right whiskers (comma separated list).
A_Count = Count of 'A' nucleotides found in this column.
C_Count = Count of 'C' nucleotides found in this column.
G_Count = Count of 'G' nucleotides found in this column.
T_Count = Count of 'T' nucleotides found in this column.
N_Count = Count of 'N' nucleotides found in this column.
Other_Nucs = Comma separated list of other nucleotides found in this column.
Other_Count = Comma separated count of other nucleotides found in this column.

For example:

#column	count	min	max	sum	mean	Q1	med	Q3	IQR	lW	rW	outliers	A_Count	C_Count
1	14336356	2	33	450600675	31.4306281875	32.0	33.0	33.0	33.0	1.0	31	33		
2	14336356	2	34	441135033	30.7703737965	30.0	33.0	33.0	33.0	3.0	26	34		
3	14336356	2	34	433659182	30.2489127642	29.0	32.0	33.0	33.0	4.0	23	34		
4	14336356	2	34	433635331	30.2472490917	29.0	32.0	33.0	33.0	4.0	23	34		
5	14336356	2	34	432498583	30.167957813	29.0	32.0	33.0	33.0	4.0	23	34		

History

Unnamed history 2.1 Mb

1: https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Using 2.1 Mb

The following job has been successfully added to the queue:

2: FASTQ Summary Statistics on data 1

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

Unnamed history 2.1 Mb

2: FASTQ Summary Statistics on data 1

1: https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq

This screenshot shows the Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A message in the center indicates a successful job submission: 'The following job has been successfully added to the queue: 2: FASTQ Summary Statistics on data 1'. Below this, instructions advise checking the status in the History pane. The left sidebar lists various tools under categories like 'Multivariate Analysis', 'Evolution', 'Motif Tools', etc. The right sidebar shows the 'History' pane with an 'Unnamed history' entry for '2.1 Mb' and a detailed entry for '2: FASTQ Summary Statistics on data 1' with a red border, containing a link to an S3 bucket: '1: https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq'.

Galaxy

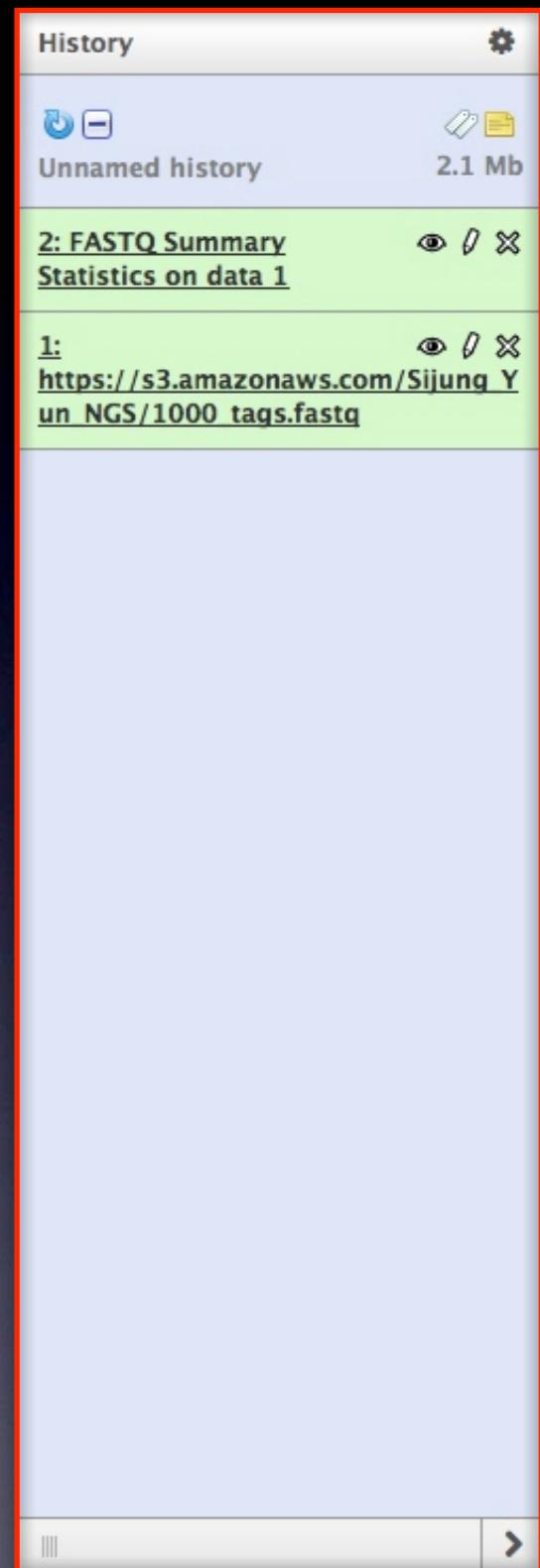
54.242.69.138

Analyze Data Workflow Shared Data Visualization Help User Using 2.1 Mb

Tools	#column	count	min	max	sum	mean	Q1	med	Q3	IQR	lW	rW	History
<u>Multivariate Analysis</u>	1	10000	2	40	358994	35.8994	35.0	38.0	39.0	4.0	29	40	
<u>Evolution</u>	2	10000	2	40	356075	35.6075	35.0	38.0	39.0	4.0	29	40	
<u>Motif Tools</u>	3	10000	2	40	355450	35.545	35.0	38.0	39.0	4.0	29	40	
<u>Multiple Alignments</u>	4	10000	2	40	354268	35.4268	35.0	38.0	39.0	4.0	29	40	
<u>Metagenomic analyses</u>	5	10000	2	40	355691	35.5691	35.0	38.0	39.0	4.0	29	40	
<u>FASTA manipulation</u>	6	10000	2	40	357066	35.7066	35.0	38.0	39.0	4.0	29	40	
<u>NCBI BLAST+</u>	7	10000	2	40	357335	35.7335	35.0	38.0	39.0	4.0	29	40	
<u>NGS: QC and manipulation</u>	8	10000	2	40	356799	35.6799	35.0	38.0	39.0	4.0	29	40	
FASTQC: FASTQ/SAM/BAM	9	10000	2	40	355891	35.5891	35.0	38.0	39.0	4.0	29	40	
▪ <u>Fastqc: Fastqc QC using FastQC from Babraham</u>	10	10000	2	40	356335	35.6335	35.0	38.0	39.0	4.0	29	40	
ILLUMINA FASTQ	11	10000	2	40	357109	35.7109	35.0	38.0	39.0	4.0	29	40	
▪ <u>FASTQ Groomer</u> convert between various FASTQ quality formats	12	10000	2	40	357572	35.7572	35.0	38.0	39.0	4.0	29	40	
▪ <u>FASTQ splitter</u> on joined paired end reads	13	10000	2	40	358956	35.8956	35.0	38.0	39.0	4.0	29	40	
▪ <u>FASTQ joiner</u> on paired end reads	14	10000	2	40	360000	36.0	35.0	38.0	39.0	4.0	29	40	
▪ <u>FASTQ Summary Statistics</u> by column	15	10000	2	40	356901	35.6901	35.0	38.0	39.0	4.0	29	40	
ROCHE-454 DATA	16	10000	2	40	350223	35.0223	34.0	38.0	39.0	5.0	27	40	
▪ <u>Build base quality distribution</u>	17	10000	2	40	351943	35.1943	34.0	38.0	39.0	5.0	27	40	
▪ <u>Select high quality segments</u>	18	10000	2	40	351757	35.1757	34.0	38.0	39.0	5.0	27	40	
▪ <u>Combine FASTA and QUAL</u> into FASTQ	19	10000	2	40	351216	35.1216	34.0	38.0	39.0	5.0	27	40	
AB-SOLID DATA	20	10000	2	40	351567	35.1567	34.0	38.0	39.0	5.0	27	40	
▪ <u>Convert SOLID output to fastq</u>	21	10000	2	40	349414	34.9414	34.0	38.0	39.0	5.0	27	40	
▪ <u>Compute quality statistics</u> for SOLID data	22	10000	2	40	350206	35.0206	34.0	38.0	39.0	5.0	27	40	
▪ <u>Draw quality score boxplot</u> for SOLID data	23	10000	2	40	350242	35.0242	34.0	38.0	39.0	5.0	27	40	
GENERIC FASTQ MANIPULATION	24	10000	2	40	351490	35.149	34.0	38.0	39.0	5.0	27	40	
	25	10000	2	40	349360	34.936	34.0	37.0	39.0	5.0	27	40	
	26	10000	2	40	353087	35.3087	35.0	38.0	39.0	4.0	29	40	
	27	10000	2	40	351412	35.1412	35.0	38.0	39.0	4.0	29	40	
	28	10000	2	40	352640	35.264	35.0	38.0	39.0	4.0	29	40	
	29	10000	2	40	352542	35.2542	35.0	38.0	39.0	4.0	29	40	
	30	10000	2	40	349193	34.9193	34.0	38.0	39.0	5.0	27	40	
	31	10000	2	40	354332	35.4332	35.0	38.0	39.0	4.0	29	40	
	32	10000	2	40	347471	34.7471	34.0	38.0	39.0	5.0	27	40	
	33	10000	2	40	354691	35.4691	35.0	38.0	39.0	4.0	29	40	
	34	10000	2	40	352285	35.2285	35.0	38.0	39.0	4.0	29	40	
	35	10000	2	40	348150	34.815	34.0	38.0	39.0	5.0	27	40	
	36	10000	2	40	350222	35.0222	35.0	38.0	39.0	4.0	29	40	

The “history”

- Displays a list of your analysis steps
- Allows interaction with analysis results
- Each item in the history is a “data-set”
- Multiple concurrent histories allowed
- Maintains the order of analysis steps, allowing extraction of workflows on-the-fly



Extracting workflows from histories

The screenshot shows a software interface with a context menu open over a history item. The menu includes options like 'Saved Histories', 'Create New', 'Clone', 'Copy Datasets', 'Share or Publish', and 'Extract Workflow'. A red arrow points from the 'Extract Workflow' option to a workflow diagram on the right. The workflow diagram consists of three nodes: 'Input dataset' (with an 'output' edge), 'FASTQ Summary Statistics' (with an 'output_file (tabular)' edge), and 'FASTQ File'. The nodes are connected by arrows indicating data flow.

History

HISTORY LISTS

- Saved Histories
- Histories Shared with Me

CURRENT HISTORY

- Create New
- Clone
- Copy Datasets
- Share or Publish
- Extract Workflow**
- Dataset Security
- Show Deleted Datasets
- Show Hidden Datasets
- Purge Deleted Datasets
- Show Structure
- Export to File
- Delete
- Delete Permanently

OTHER ACTIONS

- Import from File

2: FASTQ Statistics

1: https://s3.amazonaws.com/NGS/1

Input dataset ×

output

FASTQ Summary Statistics ×

FASTQ File

output_file (tabular)

19 matt.shirley@jhmi.edu

Histories and workflows result in
reproducible research

NGS analysis in Galaxy

- QC and manipulation: filter, trim, mask, and convert fastq files
- Picard: a Java implementation of many samtools functions
- Mapping: align to reference genome with BWA, Bowtie, Bowtie2, BFAST, PerM, Mosaik, Lastz
- RNA: Tophat, Cufflinks (gapped alignment and transcript assembly)
- GATK: advanced analysis tools from BROAD
- Peak Calling: ChIP-Seq analysis tools

[NGS: QC and manipulation](#)
[NGS: Picard \(beta\)](#)
[NGS: Mapping](#)
[NGS: Indel Analysis](#)
[NGS: RNA Analysis](#)
[NGS: SAM Tools](#)
[NGS: GATK Tools \(beta\)](#)
[NGS: Peak Calling](#)

Visualizations

Trackster linear genome browser supports most interval, continuous, and discreet data formats

Circster “circos” style connectivity browser with interactive zooming

Visual parametric optimization allows the user to pick the most optimum local parameters, then optionally apply these globally

Strengths and Weaknesses

Strengths:

- Each tool has similar user interface elements, leading to a much lower learning curve
- Histories and workflows allow reproducibility
- Cluster and cloud compute-compatible
- Extensible tool set via Python scripting

Weaknesses:

- Inefficient resource management (for now)
- Limited set of parameters for some tools

Local vs. Public

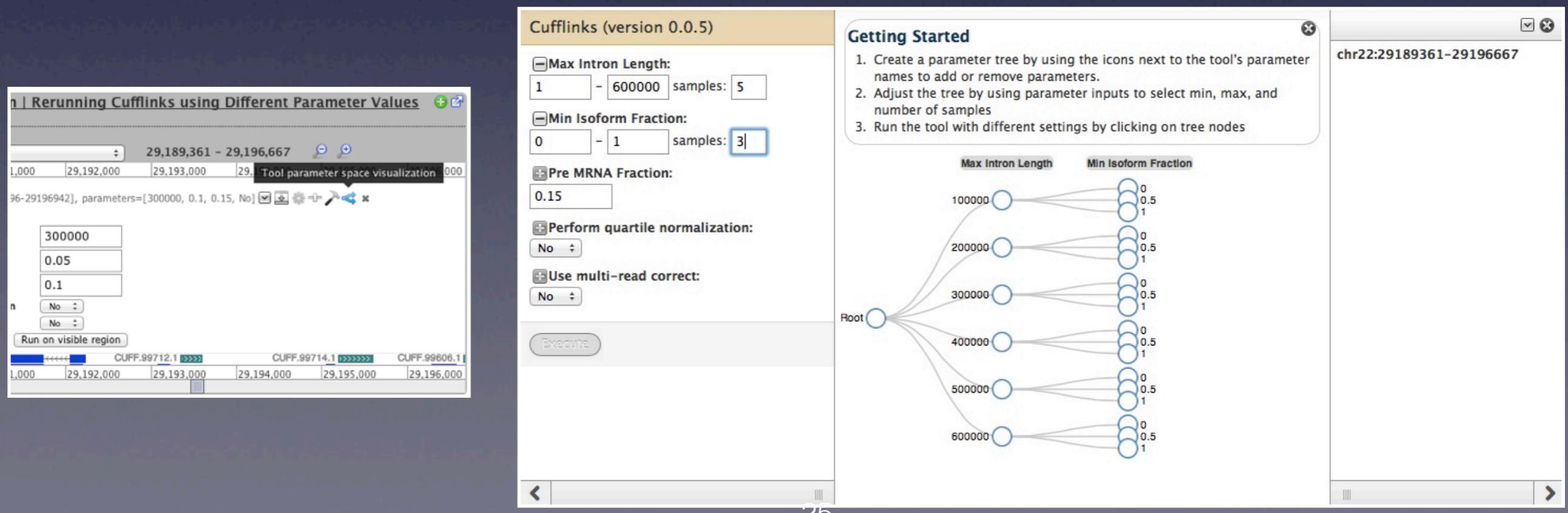
- Public Galaxy server is accessible at
<http://usegalaxy.org>
- Learn about installing local instances at
<http://getgalaxy.org>
- NGS analysis involves *large* data, and long compute times.
- For NGS analysis, a local (or cloud) installation of Galaxy is recommended.

Questions?

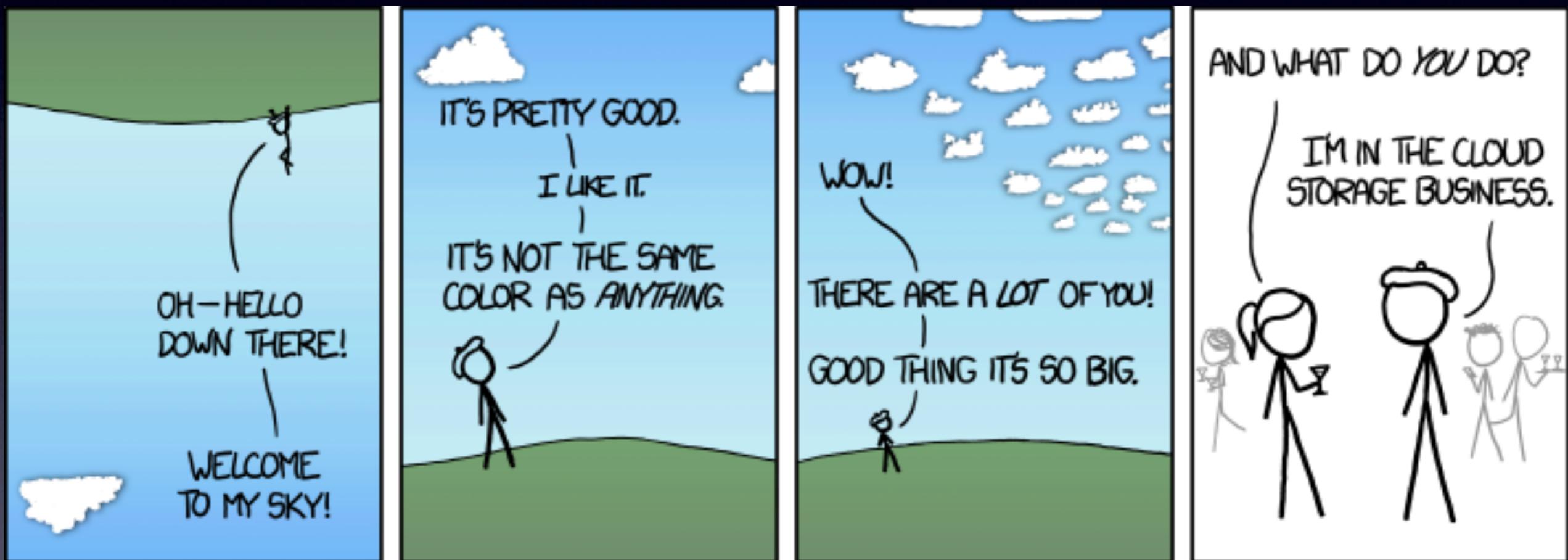
Slides available at <http://mattshirley.com/presentations>

Examples

- Basic protocols for Galaxy: Using Galaxy to Perform Large-Scale Interactive Data Analyses
- Parameter-space visualization: TopHat/CuffLinks RNA-seq optimization



Galaxy on AWS (“the cloud”)



<http://xkcd.com/1117/>

New! Two options for cluster initialization

1. Use the new cloud launch tool from the main public instance.
2. Manually configure a cluster through Amazon Web Services management console.

Using the “cloud launch” tool at Galaxy Main

I. Log in to AWS EC2 management console

<http://console.aws.amazon.com/ec2>

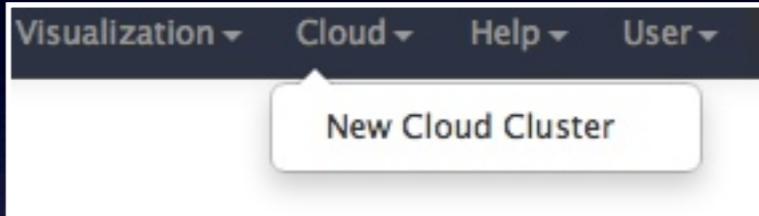
- Access your *Security Credentials* page
- Save your Access Key ID and Secret Access Key



Your Access Keys				
Created	Access Key ID	Secret Access Key	Status	
October 3, 2012	AKIAJRBU4D36GXYROQWQ	Show	Active (Make Inactive)	

Automatic Galaxy cloud initialization

- I. Click “New Cloud Cluster” from “Cloud” toolbar of the main public instance.



Alternative mirror (please use sparingly)

2. Enter your AWS access key ID and secret key

Launch a Galaxy Cloud Instance

To launch a Galaxy Cloud Cluster, enter your AWS Secret Key ID, and Secret Key. Galaxy will use these to present appropriate options for launching your cluster. Note that using this form to launch computational resources in the Amazon Cloud will result in costs to the account indicated above. See [Amazon's pricing](#) for more information. options for launching your cluster.

Key ID
 This is the text string that uniquely identifies your account, found in the [Security Credentials](#) section of the AWS Console.

Secret Key
 This is your AWS Secret Key, also found in the [Security Credentials](#) section of the AWS Console.

Final steps before initialization

3. Enter a name for your cluster

4. Enter a password you can remember

5. Either choose an existing keypair or let the tool generate one for you

6. Select at least a “Large” instance type

7. Submit

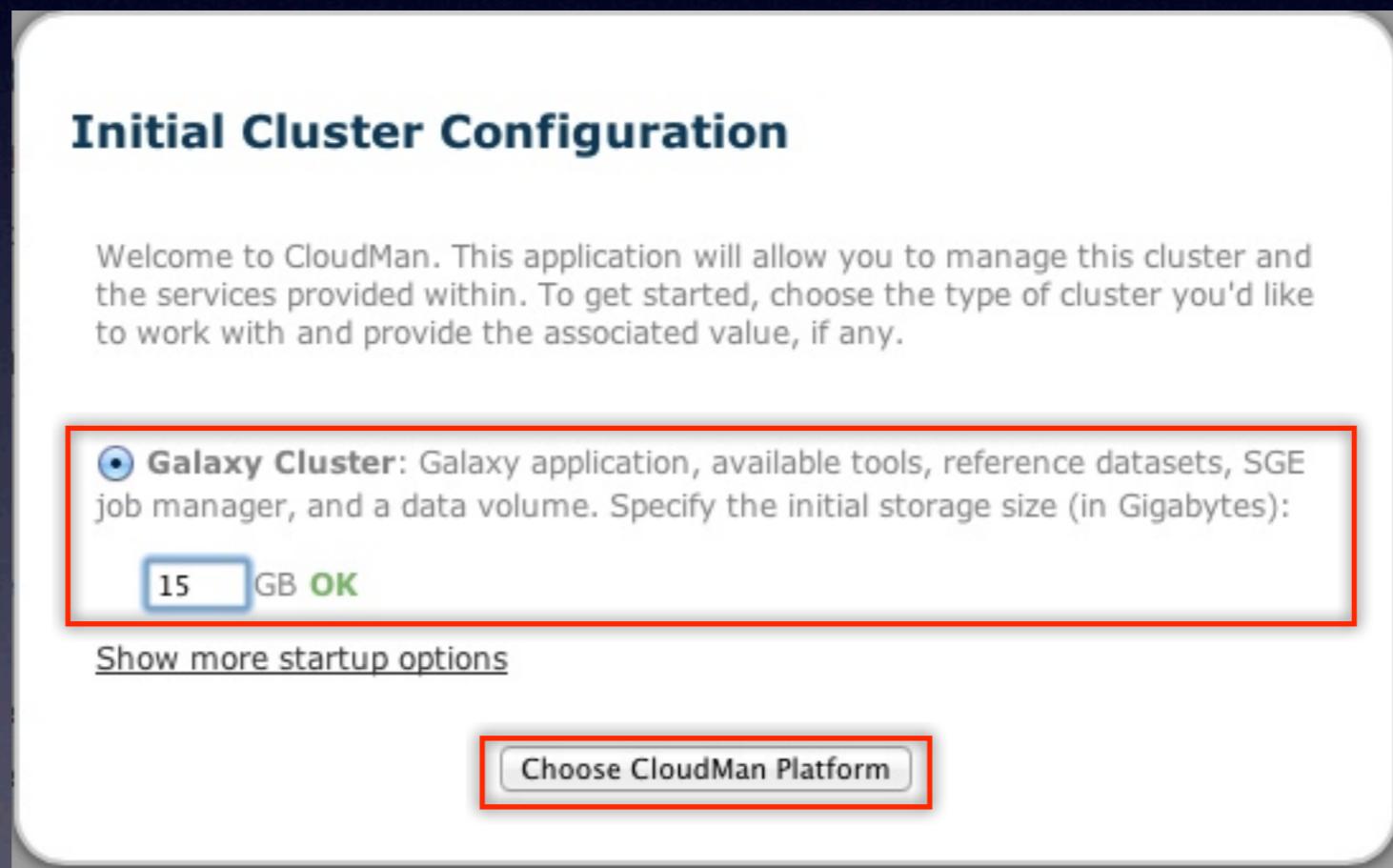
The screenshot shows a configuration form for creating a new AWS Lambda function. The fields are as follows:

- Key ID:** Your Access Key ID
This is the text string that uniquely identifies your account.
- Secret Key:** Your Secret Key
This is your AWS Secret Key, also found in the [Security Credentials](#) section of the AWS Management Console.
- Cluster Name:** NGS Cluster
This is the name for your cluster. You'll use this when you want to invoke your Lambda function from a Lambda@Edge trigger.
- Cluster Password:**
A redacted password field.
- Cluster Password - Confirmation:**
A redacted confirmation password field.
- Key Pair:** clouzman_keypair
The selected key pair for the Lambda function.
- Instance Type:** Large
The selected instance type for the Lambda function.
- Submit:** A button to submit the form.

A message at the bottom of the form says: Requesting the instance may take a moment, please wait.

Galaxy on AWS (“the cloud”)

8. After logging in using the previously specified “cluster name” and “password”, specify the initial storage for the Galaxy cluster



Galaxy on AWS (“the cloud”)

9. After a few minutes, the *Access Galaxy* button will become accessible, signaling success

- Note that performance will be improved if autoscaling is turned on

CloudMan Console

Welcome to [CloudMan](#). This application allows you to manage this cloud cluster and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

[Terminate cluster](#) [Add nodes ▾](#) [Remove nodes](#) [Access Galaxy](#)

Status

Cluster name: plato

Disk status: 51M / 15G (1%)

Worker status: Idle: 0 Available: 0 Requested: 0

Service status: Applications Data

[Cluster status log](#)

You're ready to analyze some data!

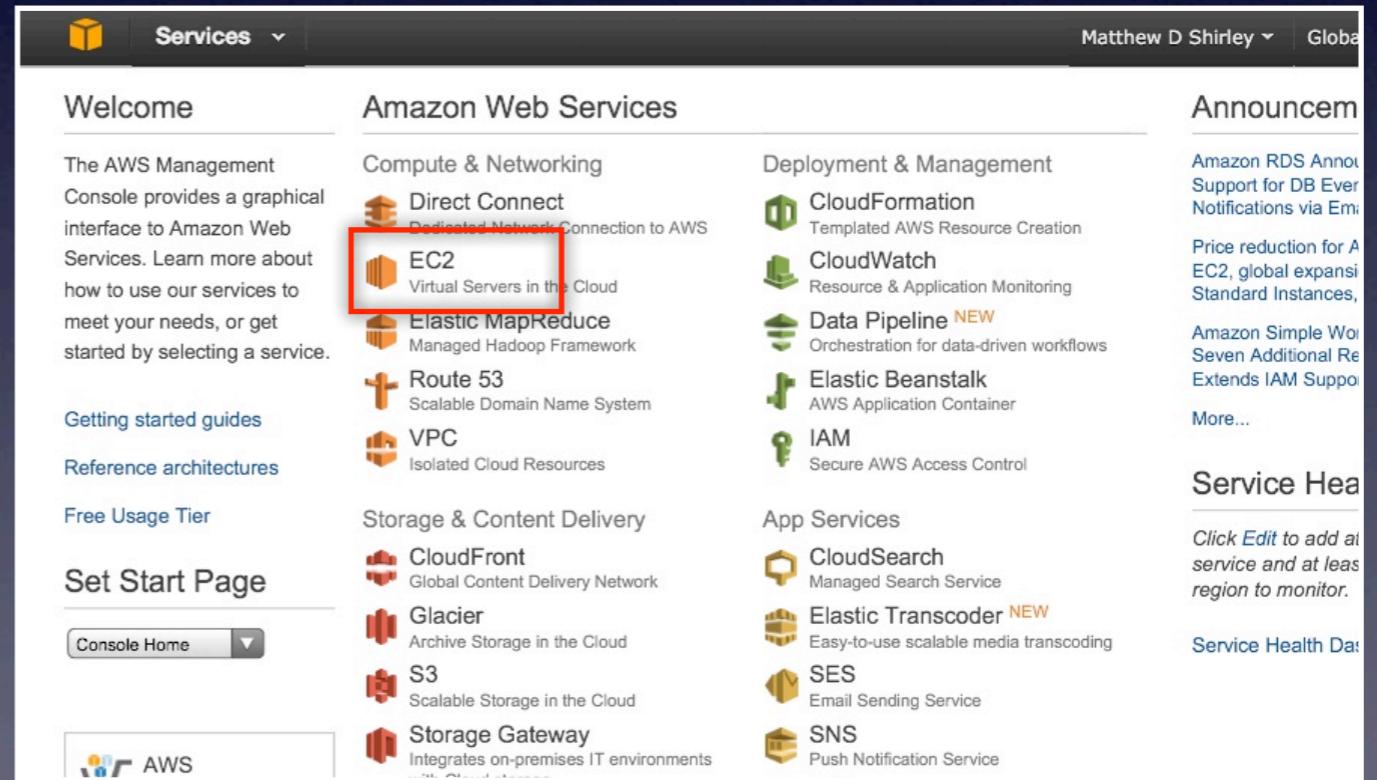
Next:

1. Learn how to shut down your cluster when you have finished.
2. Learn how to monitor your AWS usage.
3. Something didn't work? Try the hard way.

Shutting down your cluster

I. Log in to your AWS console

2. Select EC2



Shutting down your cluster

3. Select "instances" on the left and terminate any running EC2 instances

The screenshot shows the AWS EC2 Instances page. On the left, the navigation menu is open, with the 'Instances' option under 'INSTANCES' highlighted with a red box. In the center, the 'My Instances' section displays two EC2 instances:

Name	Type	Status	Status Checks	Alarm Status	Monit
empty	m1.large	terminated		none	basic
empty	m1.large	running	2/2 checks p	none	b

Below the instances, the 'Instance Actions' dropdown is open, showing options: Terminate (highlighted with a red box), Reboot, Stop, and Start.

On the right, detailed information for the running instance is shown:

- Description:** EC2 Instance selection
- AMI:** galaxy-cloudman
- Zone:**
- Type:**
- Scheduled Events:** No scheduled events
- VPC ID:** -
- CloudWatch Monitoring:** Enable Detailed Monitoring, Disable Detailed Monitoring, Add/Edit Alarms
- Alarm Status:** none
- Security Groups:** galaxy, view rules
- State:** running
- Owner:** 994862681730
- Subnet ID:** -

Shutting down your cluster

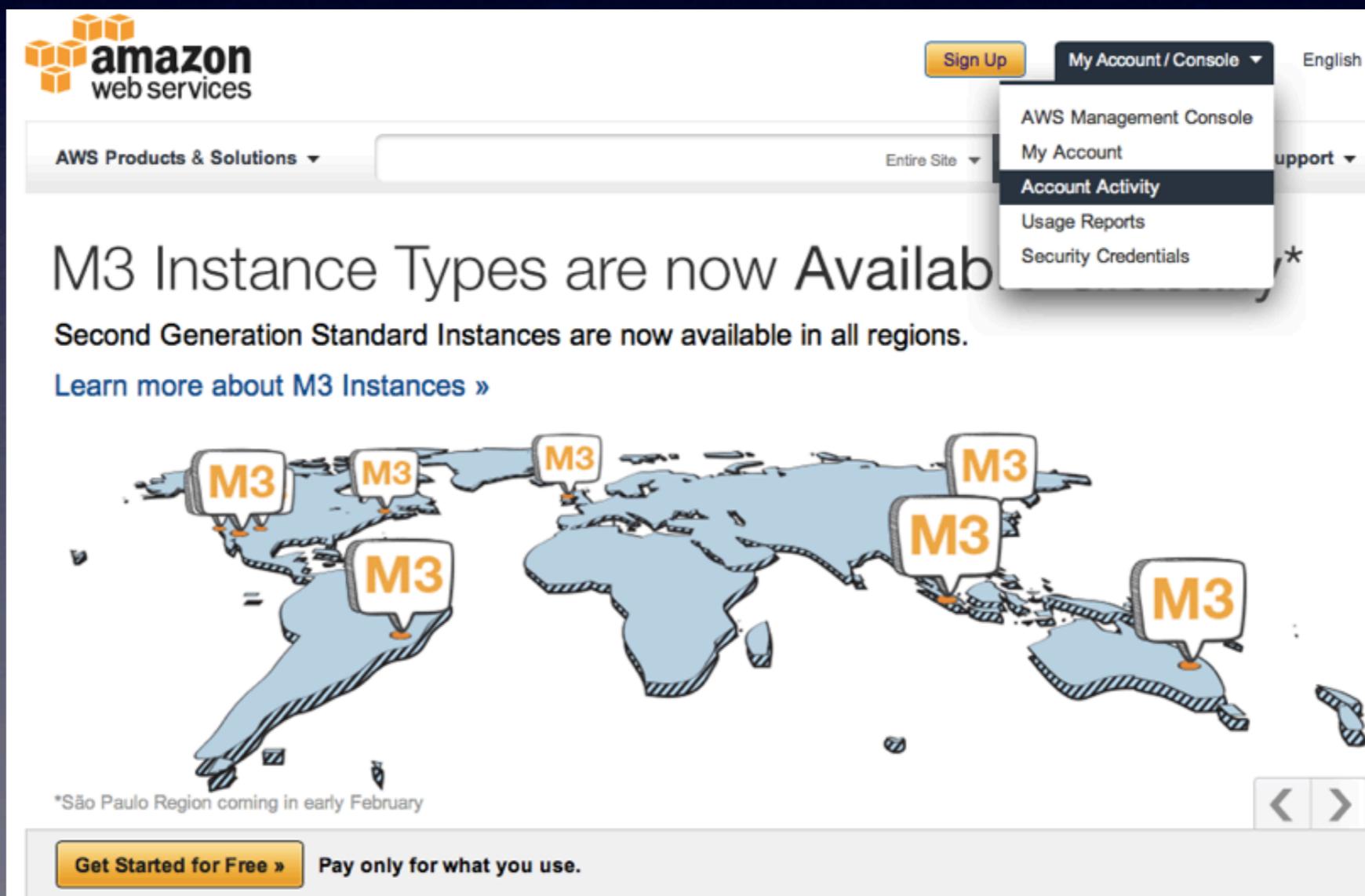
4. Also remember to delete any EBS volumes that persist

The screenshot shows the AWS Management Console interface for managing EBS volumes. The left sidebar has a 'Navigation' section with a 'Region' dropdown set to 'US East (N. Virginia)'. Below it are links for EC2 Dashboard, Events, INSTANCES, IMAGES, and ELASTIC BLOCK STORE. Under EBS, the 'Volumes' link is highlighted with a red box. The main content area is titled 'EBS Volumes' and displays a table of four volumes. The table columns are: Name, Volume ID, Capacity, Volume Type, Snapshot, Created, Zone, State, and Alarm. The volumes listed are all named 'empty' and are in the 'in-use' state. At the bottom of the table, it says '0 Volumes selected' and 'Select a volume above'.

Name	Volume ID	Capacity	Volume Type	Snapshot	Created	Zone	State	Alarm
empty	vol-f2abf088	15 GiB	standard	snap-a95003c5	2012-10-07T00:02:01	us-east-1d	in-use	none
empty	vol-c5da81bf	700 GiB	standard	snap-5b030634	2012-10-07T00:10:58	us-east-1d	in-use	none
empty	vol-5dd88327	10 GiB	standard	snap-cf3746b3	2012-10-07T00:11:16	us-east-1d	in-use	none
empty	vol-16d8836c	15 GiB	standard	-	2012-10-07T00:11:29	us-east-1d	in-use	none

Monitoring your usage!

I. Go to aws.amazon.com and select “Account Activity”



Monitoring your usage!

2. On your account activity page, select “Set your first billing alert”

Account Activity

Welcome Matthew D Shirley | Sign Out
Account Number 8638-7837-2088

Your account is enabled for monitoring estimated charges. Set your first billing alert to receive an e-mail when charges reach a threshold you define. [Learn More](#)

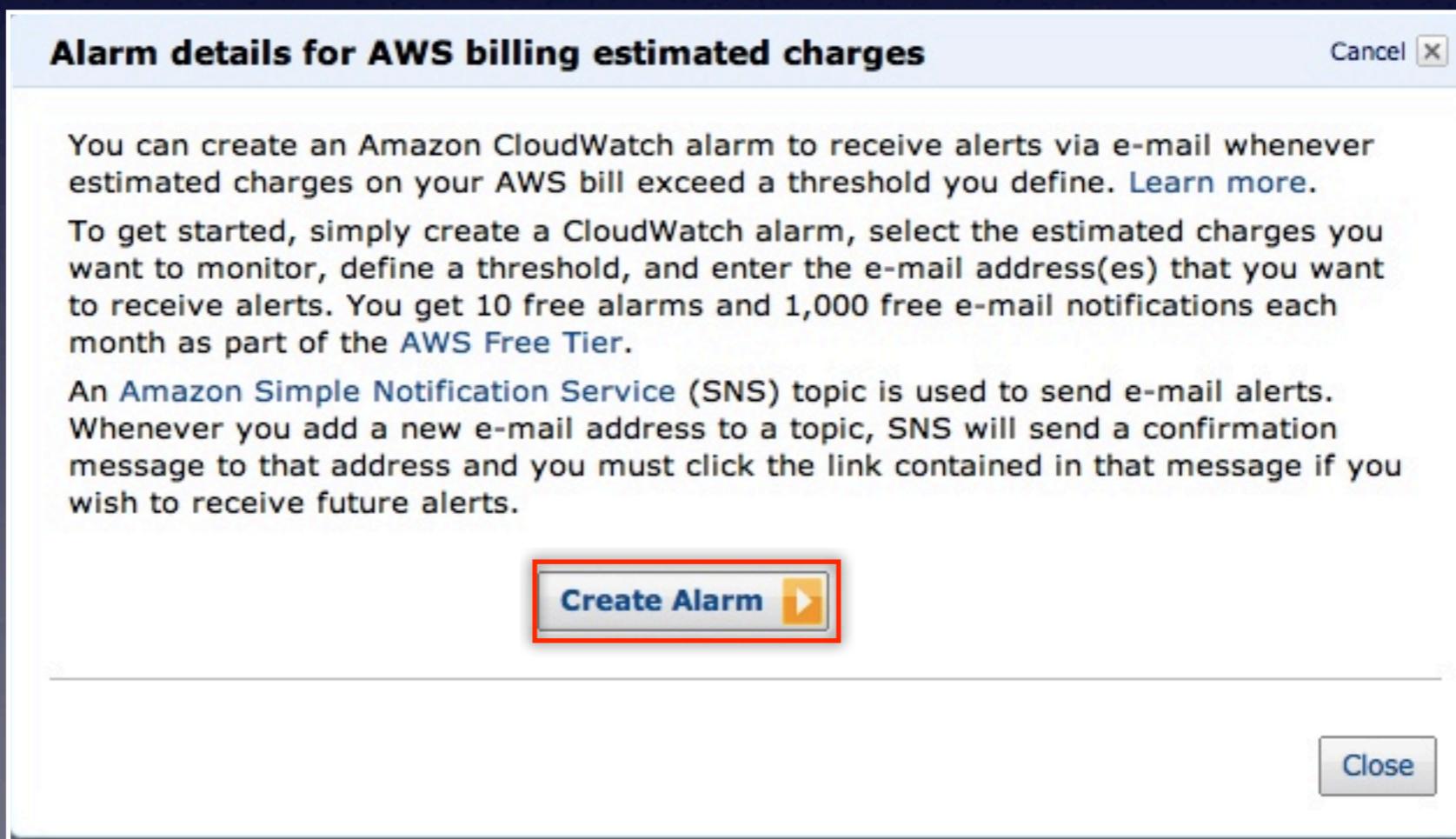
This Month's Activity as of February 6, 2013

The statement period for this report is February 1 - February 28, 2013. The charges on this page currently show activity through approximately 02/06/2013 09:43 GMT.

Select a different statement:

Monitoring your usage!

3. Select “Create Alarm”



Monitoring your usage!

4. Select an email address to send notifications to, and enter a threshold of total AWS service charges above which you wish to be notified.

Create Billing Alarm

Create an Amazon CloudWatch alarm to receive alerts via e-mail whenever estimated charges on your AWS bill exceed a threshold you define. The actual charges you will be billed in this statement period may differ from the charges shown on the notification. [Learn more](#).

To create an alarm, first choose whom to notify and then define when the notification should be sent

Send a notification to: NotifyMe (mdshw5@gmail.com) [create topic](#)

With these recipients: mdshw5@gmail.com

Whenever charges for: AWS Service Charges (total)

Exceed: USD \$ 100

Last statement period: USD 1.67 - AWS Service Charges (total)

Name this alarm: awsbilling-AWS-Service-Charges-total

EstimatedCharges (None)

1/26 1/29 2/1 2/4
00:00 00:00 00:00 00:00

[Cancel](#) [Create Alarm](#)

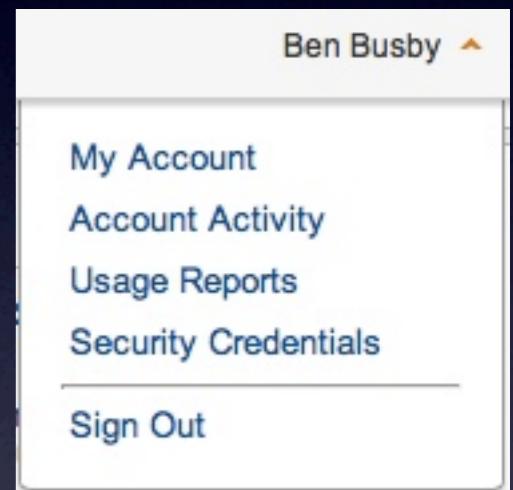
Manually configure a cluster through AWS management console

Steps adapted from <http://wiki.g2.bx.psu.edu/CloudMan>

I. Log in to AWS EC2 management console

<http://console.aws.amazon.com/ec2>

- Access your *Security Credentials* page
- Save your Access Key ID and Secret Access Key



Your Access Keys				
Created	Access Key ID	Secret Access Key	Status	
October 3, 2012	AKIAJRBU4D36GXYROQWQ	Show	Active (Make Inactive)	

Galaxy on AWS (“the cloud”)

2. Create a *Security Group* called “galaxy”, description “galaxy AMI”

- Choose Key Pairs

- Create a key pair named “galaxy” and download it to your computer

The screenshot shows the AWS EC2 Dashboard with the navigation menu on the right. The 'Security Groups' and 'Key Pairs' items are highlighted with red boxes. A 'Create Security Group' dialog is open, showing 'Name: galaxy' and 'Description: galaxy AMI'. A 'Create Key Pair' dialog is also open, showing 'Key Pair Name: galaxy'. The main dashboard area has a dark background with white text.

Galaxy on AWS (“the cloud”)

3. Add *Inbound Rules* for the services you want to access on your AMI
 - HTTP, SSH, “Custom TCP Rule” (42284) (20-21) (30000-30100), “All TCP” source: galaxy

 **Security Group: galaxy**

[Details](#) **Inbound***

Create a new rule: Custom TCP rule

Port range: (e.g., 80 or 49152-65535)

Source: 0.0.0.0/0 (e.g., 192.168.2.0/24, sg-47ad482e, or 1234567890/default)

Add Rule

Your changes have not been applied yet.

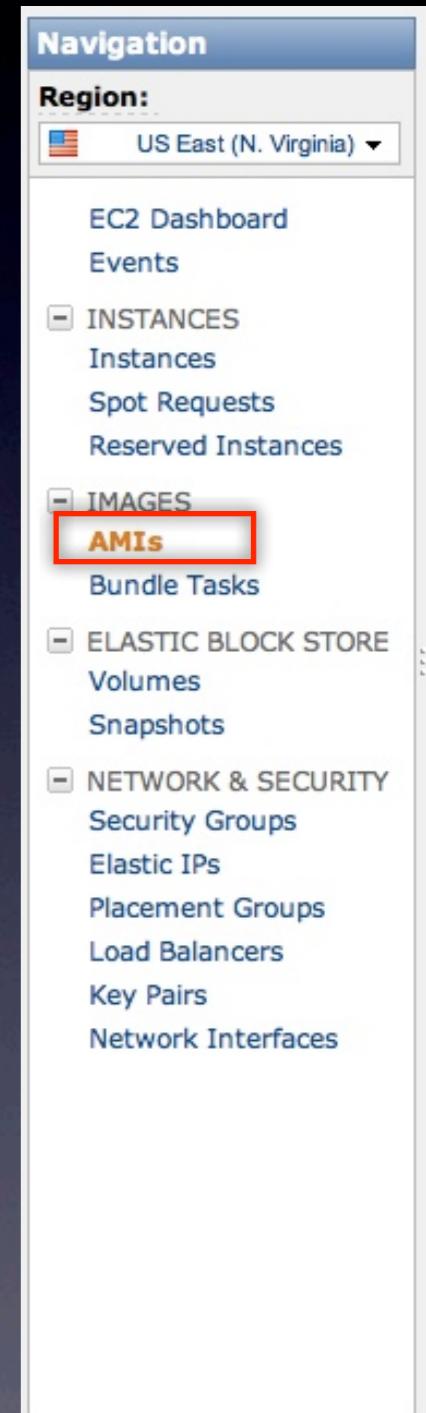
TCP Port (Service)	Source	Action
80 (HTTP)	0.0.0.0/0	Delete
22 (SSH)	0.0.0.0/0	Delete
42284	0.0.0.0/0	Delete
0 - 65535	sg-da5918b2	Delete
20 - 21	0.0.0.0/0	Delete
30000 - 30100	0.0.0.0/0	Delete

Galaxy on AWS (“the cloud”)

4. From the EC2 dashboard, select AMIs, and search for “galaxy” under *Public Images*

- Choose “galaxy-cloudman-2011-03-22” and click *Launch*

Amazon Machine Images			
Launch Spot Request Register New AMI De-register Permissions			
Viewing: Public Images All Platforms <input type="text" value="galaxy"/>			
Name	AMI ID	Source	
<input type="checkbox"/> empty	 ami-46b9412f	pb_secondary_ami/pacbio_galaxy_1_2_1.manifest.xml	
<input type="checkbox"/> empty	 ami-561bc93f	072133624695/galaxy-cloudman-2012-02-26	
<input type="checkbox"/> empty	 ami-5a759d33	115971652512/galaxy-cloud-2010-07-01	
<input type="checkbox"/> empty	 ami-78a00411	861460482541/galaxy-cloudman-2012-05-10	
<input type="checkbox"/> empty	 ami-9a7485f3	861460482541/galaxy-cloudman-2011-01-12	
<input checked="" type="checkbox"/> empty	 ami-da58aab3	861460482541/galaxy-cloudman-2011-03-22	
<input type="checkbox"/> empty	 ami-df60bfb6	072133624695/CBL_NoGalaxyUser_32bit	



matt.shirley@jhmi.edu

Galaxy on AWS (“the cloud”)

Request Instances Wizard

CHOOSE AN AMI **INSTANCE DETAILS** CREATE KEY PAIR CONFIGURE FIREWALL REVIEW

Provide the details for your instance(s). You may also decide whether you want to launch your instances as "on-demand" or "spot" instances.

Number of Instances: **Instance Type:** Large (m1.large, 7.5 GiB)

Launch as an EBS-Optimized instance (additional charges apply):

Launch Instances
EC2 Instances let you pay for compute capacity by the hour with no long term commitments. This transforms what are commonly large fixed costs into much smaller variable costs.
Launch into: EC2 VPC
Availability Zone: No Preference

Request Spot Instances

*Set Number of Instances = 1
Instance Type = “Large”
Availability Zone may be arbitrary*

Back **Continue**

Galaxy on AWS (“the cloud”)

Request Instances Wizard

CHOOSE AN AMI **INSTANCE DETAILS** CREATE KEY PAIR CONFIGURE FIREWALL REVIEW

Number of Instances: 1 **Availability Zone:** No Preference

Advanced Instance Options

Here you can choose a specific **kernel** or **RAM disk** to use with your instances. You can also choose to enable CloudWatch Detailed Monitoring or enter data that will be available from your instances once they launch.

Kernel ID: **RAM Disk ID:**

Monitoring: Enable CloudWatch detailed monitoring for this instance
(additional charges will apply)

User Data:
 as text as file

```
cluster_name: plato
password: eu_a-mousoi
access_key: <Access Key ID>
secret_key: <Secret Access Key>
```

base64 encoded

Termination Protection: Prevention against accidental termination.

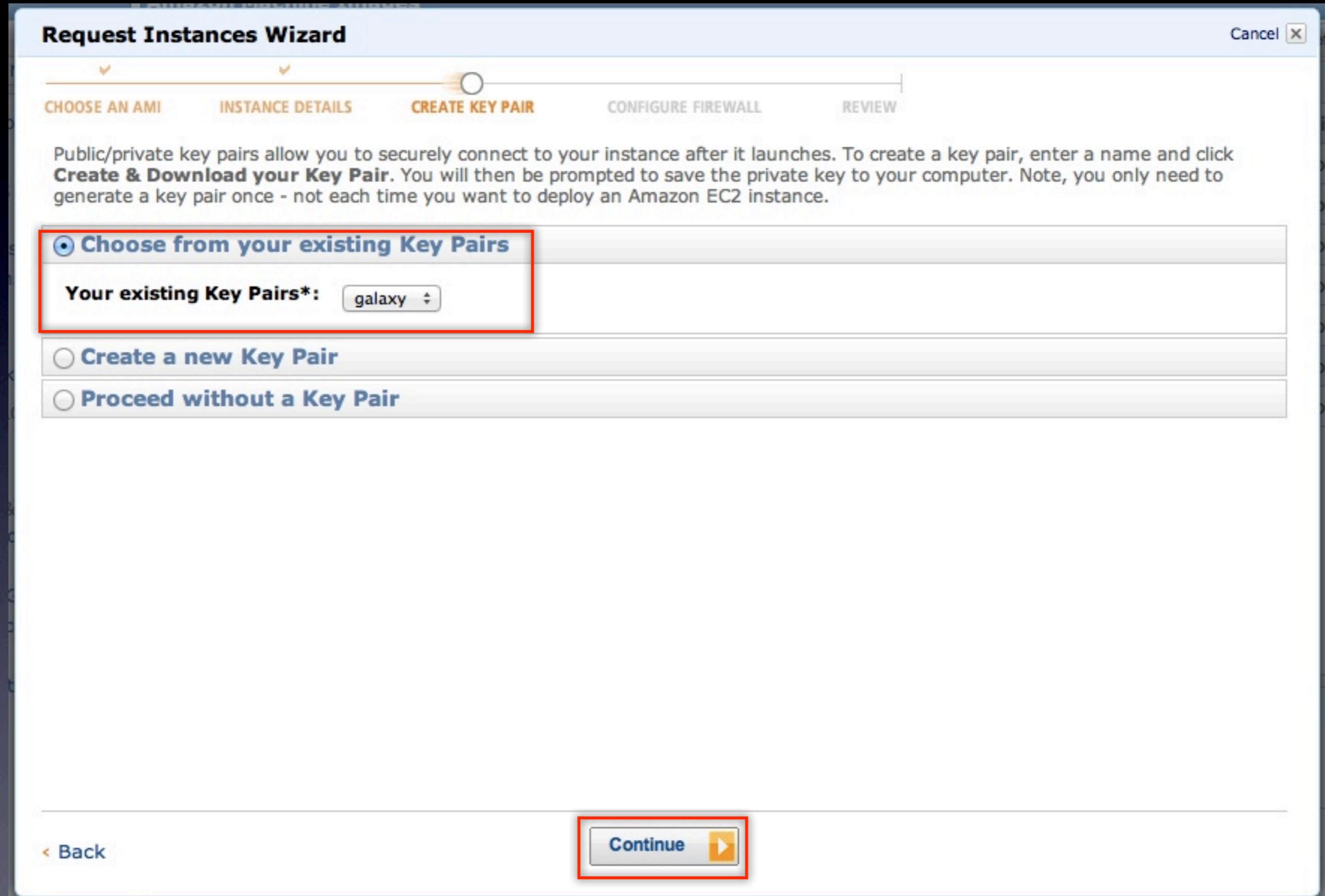
Shutdown Behavior:

IAM Role:

Continue 

Fill in User Data with information previously saved

Galaxy on AWS (“the cloud”)



Galaxy on AWS (“the cloud”)

The screenshot shows the 'Request Instances Wizard' on the 'CONFIGURE FIREWALL' step. The wizard has five steps: CHOOSE AN AMI, INSTANCE DETAILS, CREATE KEY PAIR, CONFIGURE FIREWALL, and REVIEW. The 'CONFIGURE FIREWALL' step is currently active. It displays instructions about security groups and lists two existing security groups: 'sg-ae3b7cc6 - default' and 'sg-da5918b2 - galaxy'. The 'sg-da5918b2 - galaxy' group is selected and highlighted with a blue background. A red box surrounds the 'sg-da5918b2 - galaxy' entry. Below the list, a note says '(Selected groups: sg-da5918b2)'. There is also an option to 'Create a new Security Group'. At the bottom right, a large red box surrounds the 'Continue' button, which has a yellow arrow icon.

Choose your “galaxy” security group

Galaxy on AWS (“the cloud”)

Request Instances Wizard

[Cancel](#)

CHOOSE AN AMI INSTANCE DETAILS CREATE KEY PAIR CONFIGURE FIREWALL **REVIEW**

Please review the information below, then click **Launch**.

AMI:  Other Linux AMI ID ami-da58aab3 (x86_64) [Edit AMI](#)

Number of Instances: 1

Availability Zone: No Preference

Instance Type: Large (m1.large)

Instance Class: On Demand [Edit Instance Details](#)

EBS-Optimized: No

Monitoring: Disabled **Termination Protection:** Disabled

Tenancy: Default

Kernel ID: Use Default **Shutdown Behavior:** Stop

RAM Disk ID: Use Default

Network Interfaces:

Secondary IP Addresses:

User Data: cluster_name: plato...

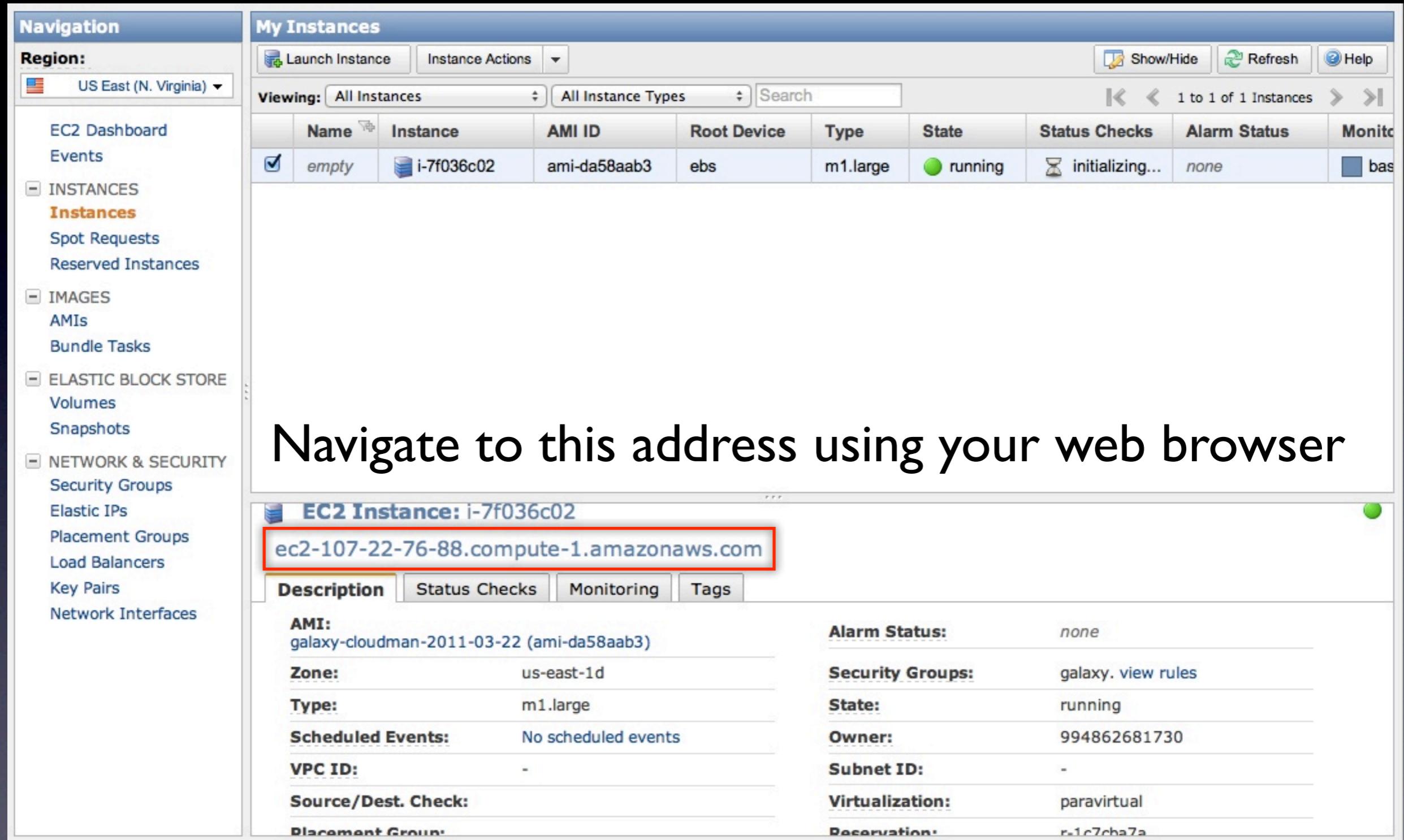
IAM Role: [Edit Advanced Details](#)

Key Pair Name: galaxy [Edit Key Pair](#)

Security Group(s): sg-da5918b2 [Edit Firewall](#)

[Back](#) **Launch** 

Galaxy on AWS (“the cloud”)



The screenshot shows the AWS Management Console interface for the EC2 service. The left sidebar contains a navigation menu with various AWS services like EC2 Dashboard, Events, Instances, Images, Elastic Block Store, Network & Security, and more. The main content area is titled "My Instances" and displays a table of running instances. One instance is listed: "empty" (ami-da58aab3, m1.large, running). Below this, a detailed view of the selected instance (i-7f036c02) is shown, including its AMI, zone, type, and network details. A red box highlights the public IP address "ec2-107-22-76-88.compute-1.amazonaws.com".

Navigation

Region: US East (N. Virginia)

My Instances

Viewing: All Instances | All Instance Types | Search | 1 to 1 of 1 Instances

Name	Instance	AMI ID	Root Device	Type	State	Status Checks	Alarm Status	Monitor
<input checked="" type="checkbox"/> empty	i-7f036c02	ami-da58aab3	ebs	m1.large	running	initializing...	none	basic

Instances

EC2 Instance: i-7f036c02

ec2-107-22-76-88.compute-1.amazonaws.com

Description Status Checks Monitoring Tags

AMI: galaxy-cloudman-2011-03-22 (ami-da58aab3)

Zone: us-east-1d

Type: m1.large

Scheduled Events: No scheduled events

VPC ID: -

Source/Dest. Check:

Placement Group:

Alarm Status: none

Security Groups: galaxy. view rules

State: running

Owner: 994862681730

Subnet ID: -

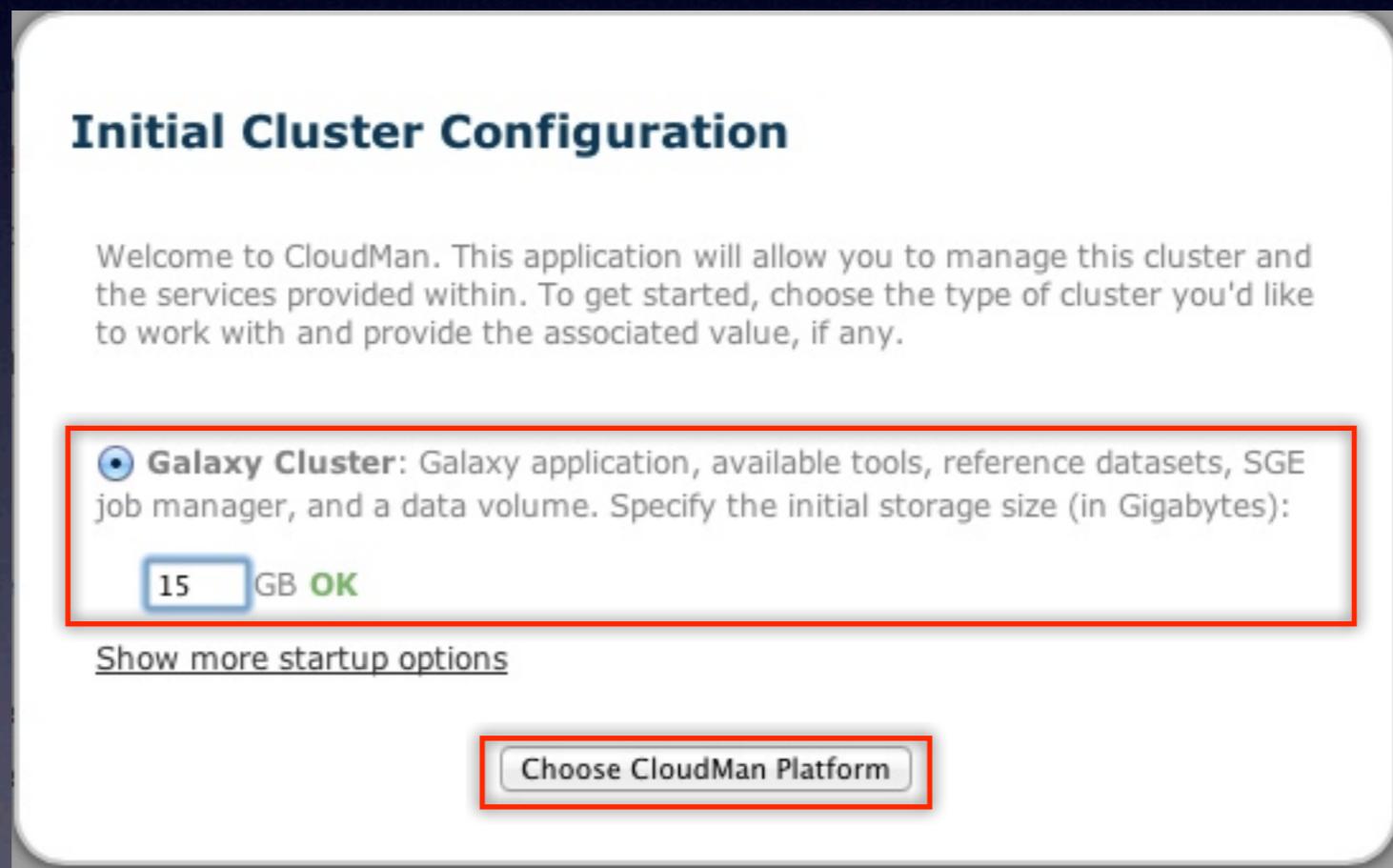
Virtualization: paravirtual

Reservation: r-1c7cha7a

Navigate to this address using your web browser

Galaxy on AWS (“the cloud”)

5. After logging in using the previously specified “cluster name” and “password”, specify the initial storage for the Galaxy cluster



Galaxy on AWS (“the cloud”)

6. After a few minutes, the *Access Galaxy* button will become accessible, signaling success
 - Note that performance will be improved if autoscaling is turned on

CloudMan Console

Welcome to CloudMan. This application allows you to manage this cloud cluster and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

[Terminate cluster](#) [Add nodes ▾](#) [Remove nodes](#) [Access Galaxy](#)

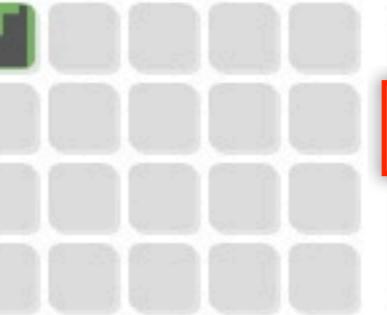
Status

Cluster name: plato 

Disk status: 51M / 15G (1%)  

Worker status: Idle: 0 Available: 0 Requested: 0

Service status: Applications  Data 



Autoscaling is **off**.
Turn on?

[Cluster status log](#) 