Getting Bioinformatics Done With Galaxy

J Fass, M Britton, N Joshi, R Feltstykket
UCD Genome Center Bioinformatics Core
bioinformatics.core@ucdavis.edu
May 13, 2015

The Bioinformatics Core

Nik Joshi

Monica Britton

Joe Fass

Keith Bradnam

Blythe Durbin-Johnson

Mike Lewis

Adam Schaal

Richard Feltstykket

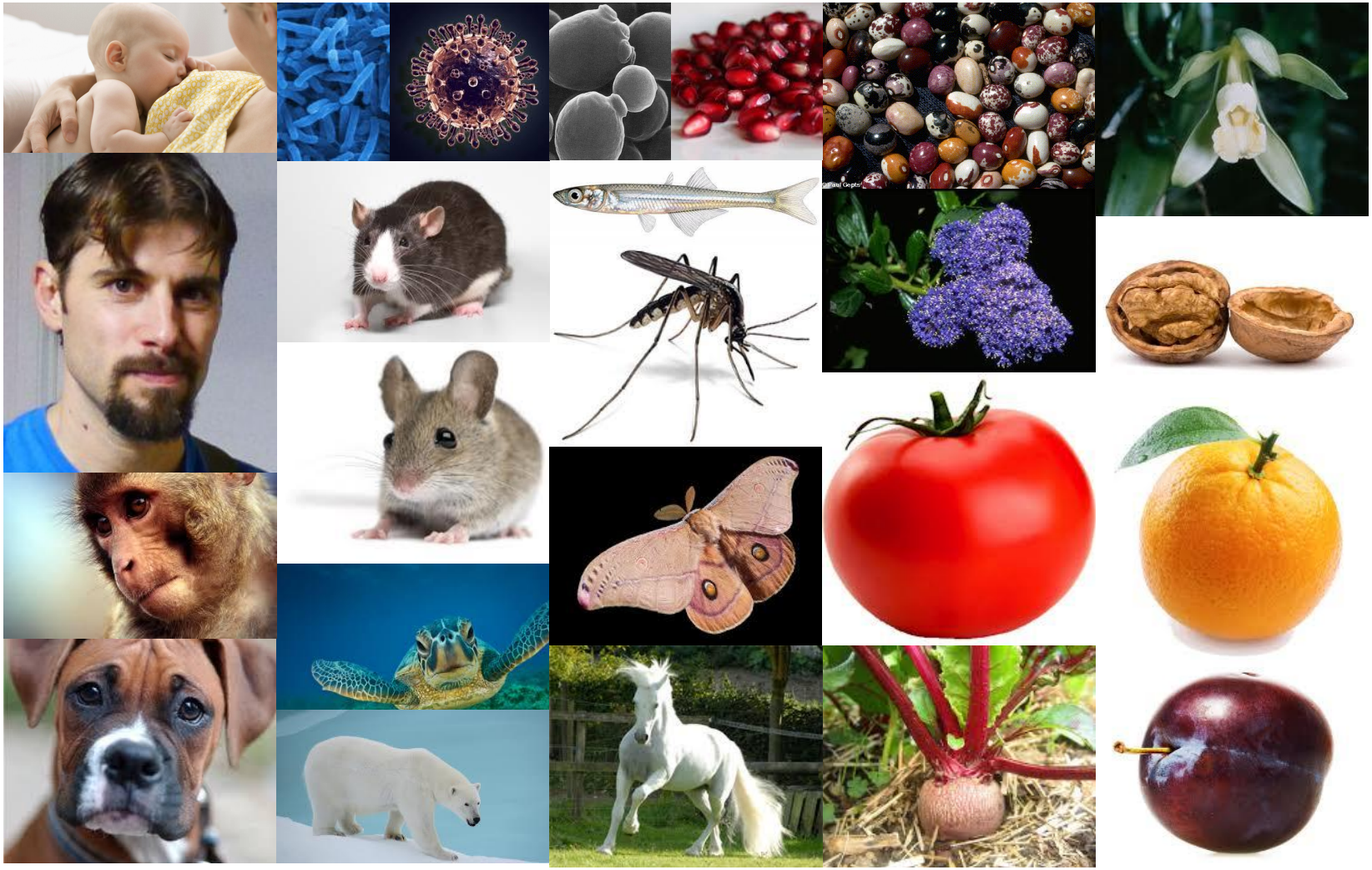
The Genome Center

The tall building across from the football stadium (you may have passed by on your way here)
We're in Room 1300 (under the clock)

What We Do

- Project Work
  - grant planning, methods, letters of support
  - data analysis, consulting (experimental design)
  - manuscript methods
- Training
  - workshops (Galaxy, command line)
- Systems Administration
  - HPC spec, setup
  - cluster, instrument, lab server maintenance
  - build custom tools
- Web Applications, DB
  - build, connect custom applications

What We Do - Species Snapshot



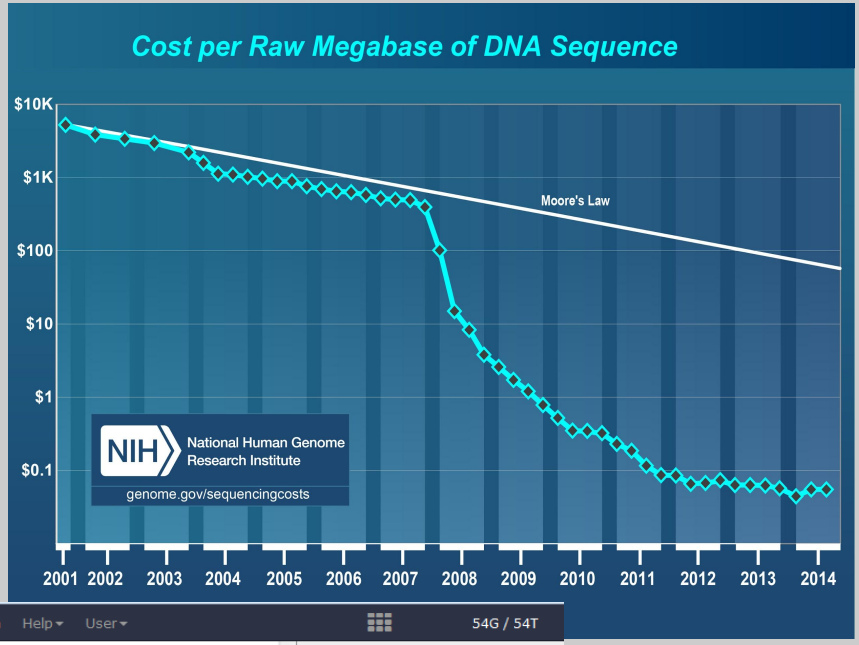
What We Do - Project Snapshot

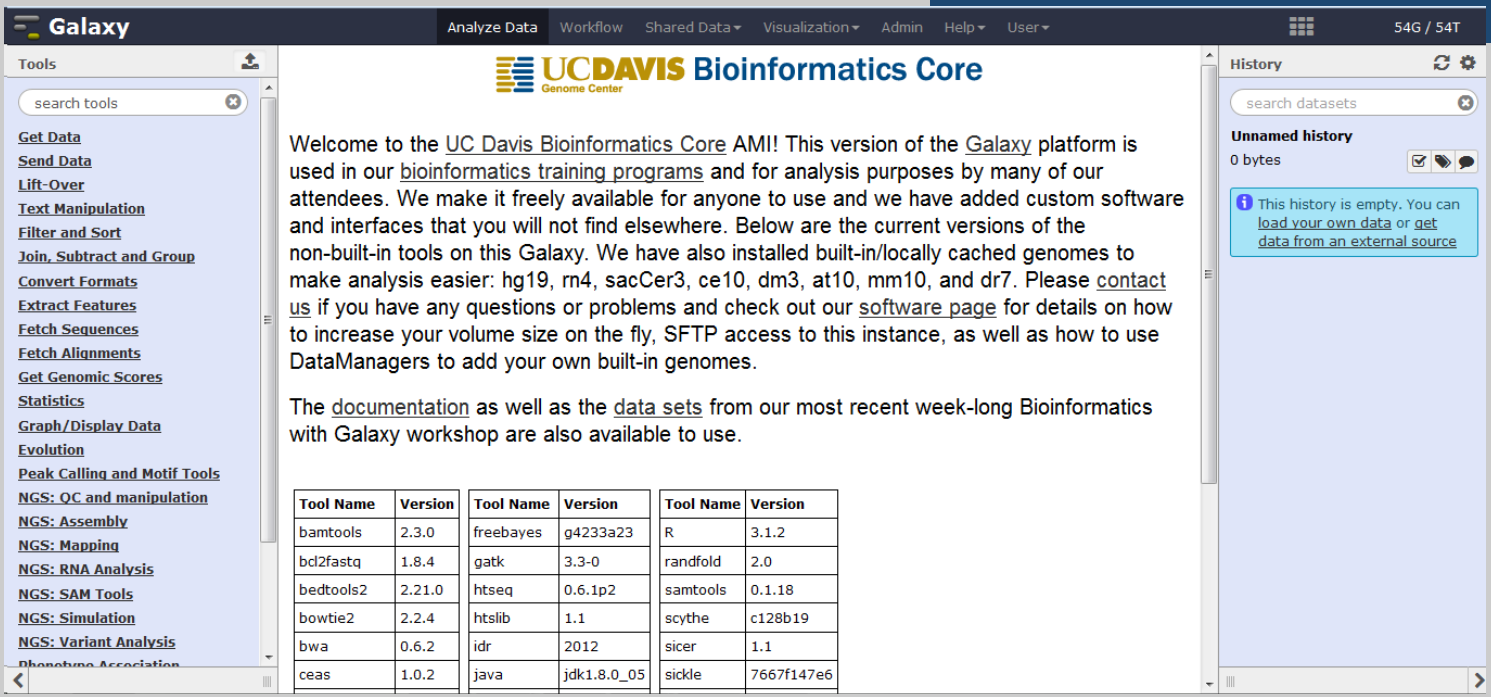
- ***RNA-Seq*** (differential expression, annotation)
- … GO-, pathway enrichment
- … Transcriptome assembly and annotation
- Exome sequencing (variant discovery)
- ***Genome sequencing*** (assembly, variant discovery)
- miRNA-Seq (differential expression)
- ChIP-Seq (peak finding, motifs, differential binding)
- 16S bacterial population profiling, metagenomics
- Pathogen discovery
- …

Systems Administration
- keeping it all running

- HPC for Bioinformatics Core, Genome Center, campus, and beyond
- standardizing around Ubuntu, modules
- custom VMs for various needs (including Galaxy)
- rent time on high memory compute nodes (up to 500 GB RAM, 48 cpus)

Big Data Meets Galaxy





What Does “Big Data” Really Mean?

- As the cost of sequencing decreases, it’s much easier to generate *huge* amounts of data.
- Rapid technology change means software has to keep up with *changing* data characteristics.
- Both features encourage *centralized* approach to data storage and analysis.
- So, where can the data be stored and accessed, with relative ease, from my laptop?

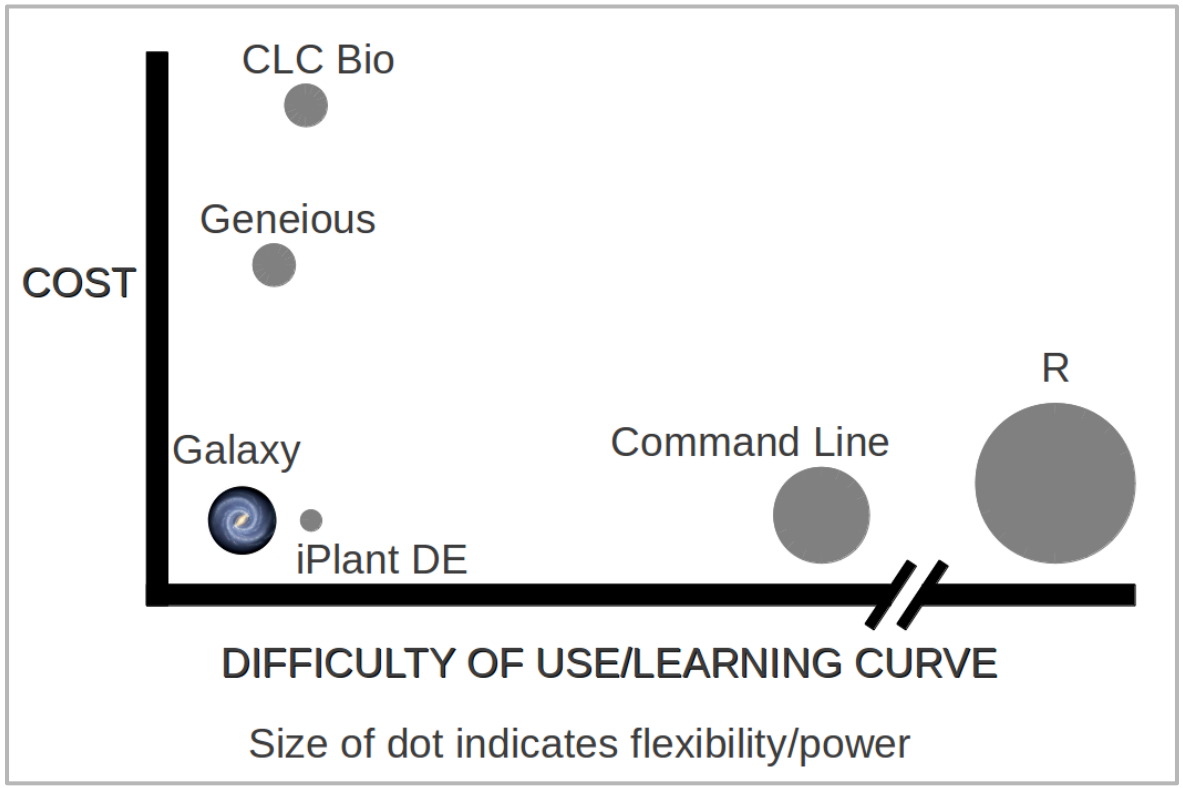
Bioinformatics Platforms

- Linux command line
  - cutting edge, flexible
  - highest learning curve (~months to years)
- Galaxy
  - deployed on many public servers, lots of history
  - open source and relatively easy to develop for
- iPlant Collaborative (link)
  - younger than Galaxy, but more funding, personnel
  - not open source yet
- Commercial products (Geneious, CLC, etc.)
  - black boxes (no public review for bugs)
  - they develop new tools for you … or, they don't

Linux Command Line

It can be challenging and frustrating to learn concepts, tools, workflows, and command line at the same time.

Comparison of Bioinformatics Interfaces



Some of these are easier to transition between than others.

Galaxy is a Collection of Tools

- Advantages
  - User-friendly interface with drop-down menus
  - Workflows facilitate processing data
  - All Galaxy tools are available for command line
  - Not a “Black Box”
- Disadvantages
  - All parameters may not be available for some tools
  - Installation of new software and writing new tools usually requires use of the command line.

Galaxy is a Workflow Manager

- Some examples from Galaxy Main (requires identical tool / wrapper versions)
- *All* workflows must be adjusted for:
  - “weird” data
  - new technologies
  - new tools
- The workflow / pipeline can sometimes make a *big* difference to the outcome!
- Workflows can be published to document your methods and allow reproducibility

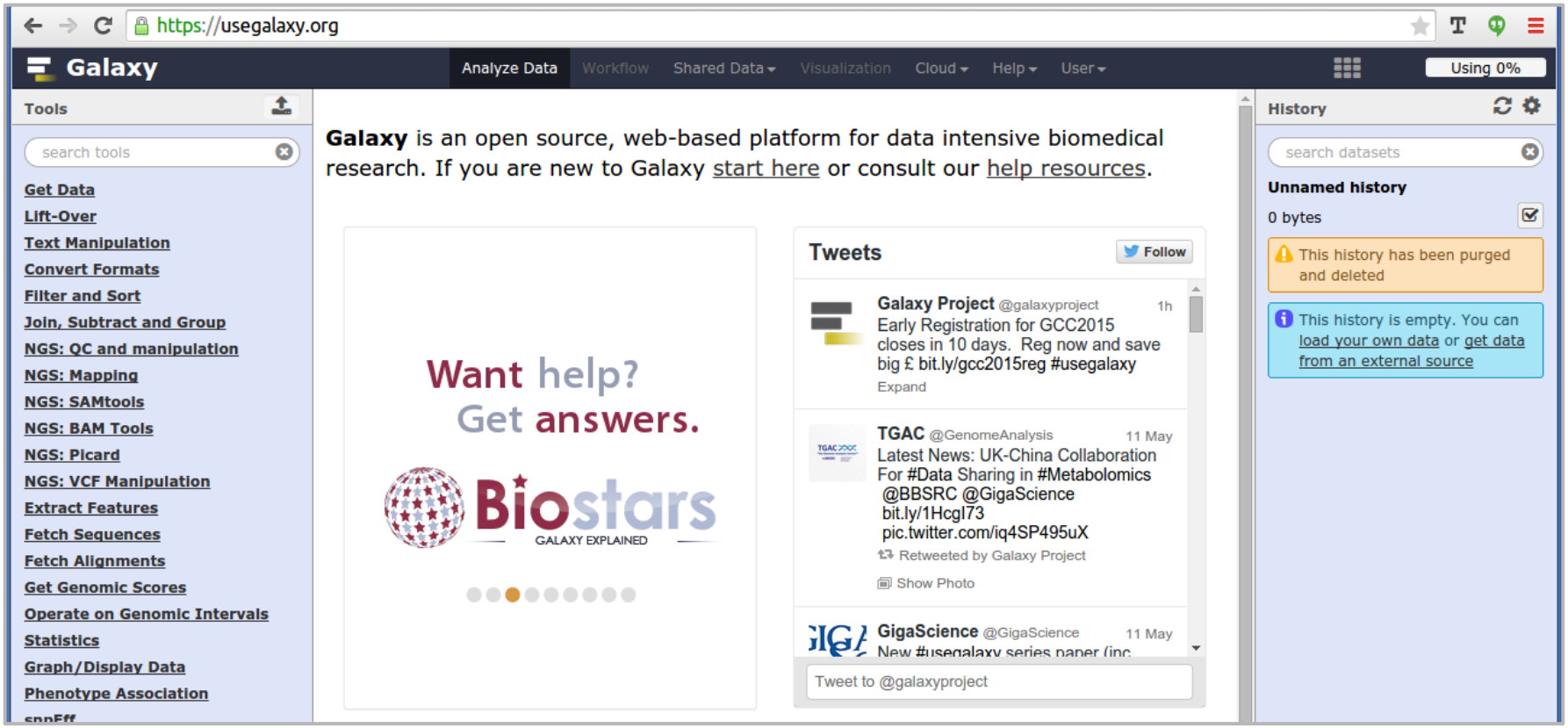
Open Source & Reproducibility

- Galaxy and command-line tools are open source
- Open source does not always mean *free*
- Open source tools don't guarantee reproducibility! The pipeline, commands, environment (hardware, OS), data must *also* be open.
  - public data archives - NCBI (various), GigaDB, etc.
  - code / commands in public, *stable* repositories (GitHub, ~~Google Code~~, etc.) or workflow management solutions (Galaxy, iPlant)

Where Does the Computing Power Reside?

- Private Galaxy (shared within lab, or personal)
  - On your own server
    - Best for heavy compute user
    - Fully customizable (with experience)
  - In the Cloud
    - Best for lighter use (30% rule)
    - On single servers, you get what you pay for
    - Some turnkey(-*ish*) cluster solutions: Galaxy's "Cloud" menu item, MIT's StarCluster (command line only)
    - data transfer may be an issue
- Public Galaxy (shared with everyone!)
  - Often limited compute power, scheduling
  - Not *readily* customizable (request changes from administrators)

Galaxy Main (usegalaxy.org)



Let’s Explore Galaxy Main ...

The Pinwheel galaxy, aka Messier 101; infrared image from NASA's Spitzer Space Telescope. Image credit: NASA/JPL-Caltech/STScI

Galaxy Main Resources

Users' Support (biostar.usegalaxy.org)

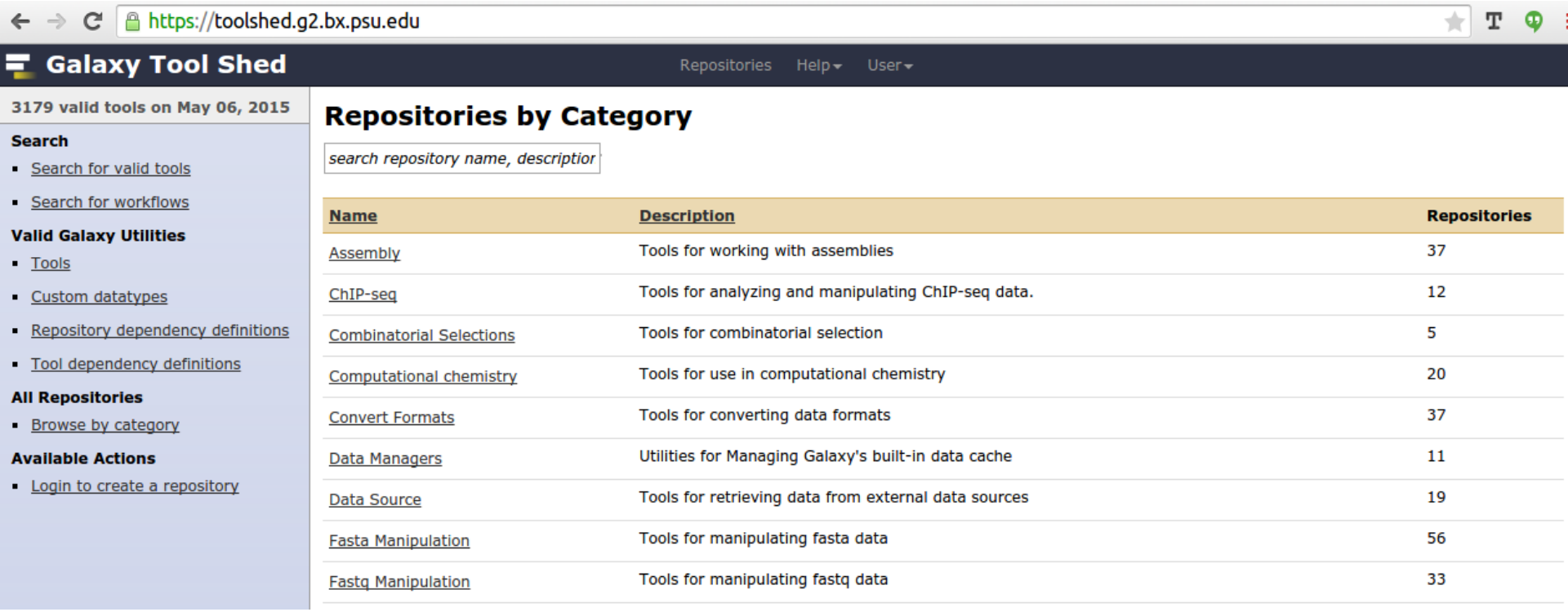
Training

- On-line (wiki.galaxyproject.org/Learn)
- Galaxy Training Network (wiki.galaxyproject.org/Teach/GTN)

Public Galaxy Disadvantages

- Occasionally, jobs may wait in queue for days
- Not every tool you want is there
- Some tools are not the most recent version
- Some tools may be missing menu choices for the options you want to run
- Storage quotas

Galaxy Tool Shed



Tool sheds contain repositories -- collections of tools and tool interfaces written for Galaxy

Galaxy Tool Shed

- Tools can be added to a custom Galaxy installation *by an admin user*.
- Most are only “wrappers” that need software to be installed separately.
- Caveat downloader: Some tools may be cutting-edge, some may be outdated. Some may be broken.
- There may be multiple wrappers for the same software.
- We have contributed to the Tool Shed (Sickle, Scythe, edgeR).

Custom Galaxy Installations

- There are over 60 public galaxies
  - List at wiki.galaxyproject.org/PublicGalaxyServers
  - Often have specialized software and/or databases
  - May not be well-maintained
- Install Galaxy on your own server
  - Some command line needed to get it running and maintain
  - We can provide this expertise
- Run a customized Galaxy “in the cloud”
  - Some commercial options at wiki.galaxyproject.org/Cloud
  - We have built a custom Galaxy (virtual machine) that is easy to run on Amazon Web Services (AWS)

Galaxy on AWS - Customized by UCD!

Home | Rates | Projects | Software | Training | People | Contact Us

UC Davis Bioinformatics Core Galaxy and Command-Line AMI

The Bioinformatics Core uses **Galaxy** and the command-line for our training workshops and courses, running in the **Amazon Cloud**. We make the Amazon Machine Image (AMI) publicly available so that the community can use it for their projects. In addition to the standard software, our AMI contains customized software and interfaces that you will not find elsewhere; these tools are available through the Galaxy interface, or via the command-line under /software. The AMI also contains all of the training materials from our week-long workshop as well as pre-indexed model genomes in Galaxy and under /data/refs. The current Bioinformatics Core AMI ID is **ami-ad46a9e9** and is located in the N. California Region. There is no charge for using this AMI to launch your own instances in the Amazon Cloud, but you *will* need an AWS account, and Amazon will charge you for running instances and storing/transferring data. The default admin login for this Galaxy AMI is **galaxyadmin@galaxyadmin.edu** and the password is **galaxy**.

Notes on starting our AMI: If you are using Galaxy, you need to use the following rules in your security group:

- SSH
- HTTP
- HTTPS
- Custom TCP Rule, port range 8080
- Custom TCP Rule, port range 2200
- Custom TCP Rule, port range 20-21

Increasing the disk storage on an instance

You can increase the size of your data partition (Device /dev/sdg) on launch by simply increasing the size of the data volume from 200Gb to whatever you want (up to 16384 Gb) in the "Add Storage" step. There are also instructions on how to **increase the data volume size of an existing instance** when you want to increase your capacity. We have added code in our AMI to automatically detect if the volume has increased capacity and to expand the filesystem to that capacity. The steps to increase to maximum capacity can take a while depending upon the size increase, which means you may need to wait a while before the AMI is up and running.

Using FTP and DataManagers

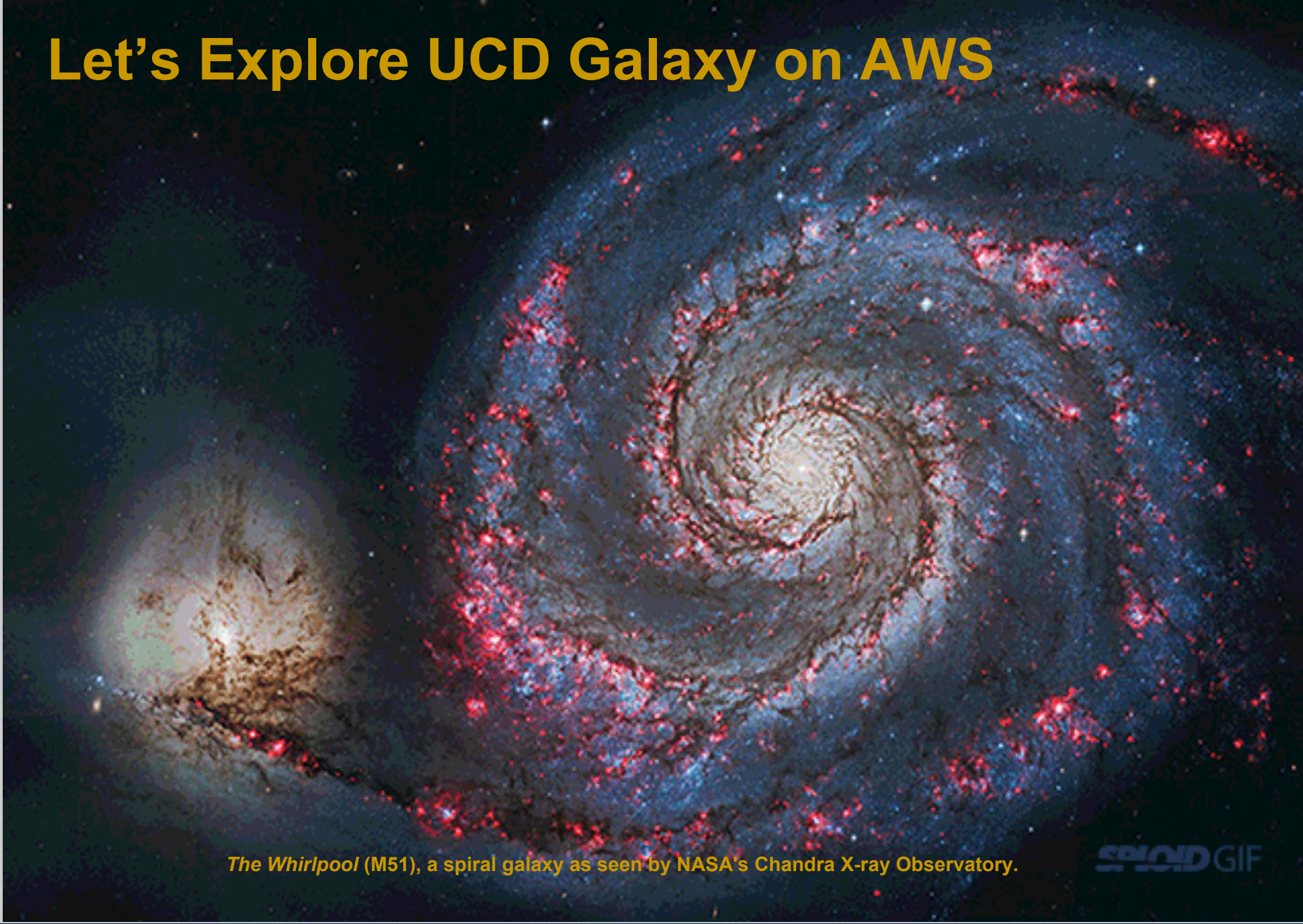
If you wish to use FTP to transfer files to your Galaxy instance, we have **instructions on how to use FileZilla** to do so. FTP transfer is recommended for large files, however, you can use whatever FTP client you want. We also have instructions on **how to use DataManagers** to create your own built-in/locally cached indexed genomes.

Copying AMIs to different Regions

To copy our AMI to a different Region, you'll need to launch an instance using our AMI and then **make your own AMI** from that instance. Then follow **these instructions** to copy the AMI to another Region. Make sure to terminate the instance, **deregister the AMI, and delete the snapshots** of your AMI from the N. California region.

bioinformatics.ucdavis.edu

Let's Explore UCD Galaxy on AWS



The Whirlpool (M51), a spiral galaxy as seen by NASA's Chandra X-ray Observatory.

Biologists and Bioinformatics

Galaxy is a great place to start, and can be all you will ever need.

But, once you become comfortable with workflows and analyzing your own data, you may want to explore “beyond the Galaxy”.

Command line won’t be as intimidating, once you know what tools you need to accomplish the task.

Run the same analysis in Galaxy and command line and compare the results!

Learning to Run Analyses With Galaxy

Today we've explored Galaxy as in interface for Bioinformatics tools and software.

For your analysis you will need to learn:

- Which tools are applicable to your data
- How to best combine those tools
- What parameters are important
- How to assess the quality of data
- What is the useful information in the output
- … and many more things

We can help you ...

UCD Bioinformatics Training Program (training.bioinformatics.ucdavis.edu)

UC Davis Bioinformatics Training Program

Bioinformatics courses, boot camps and workshops presented by the University of California, Davis Bioinformatics Core

Home | Training Events | Documentation | News | Accommodations | FAQ | Contact Us

Training Events

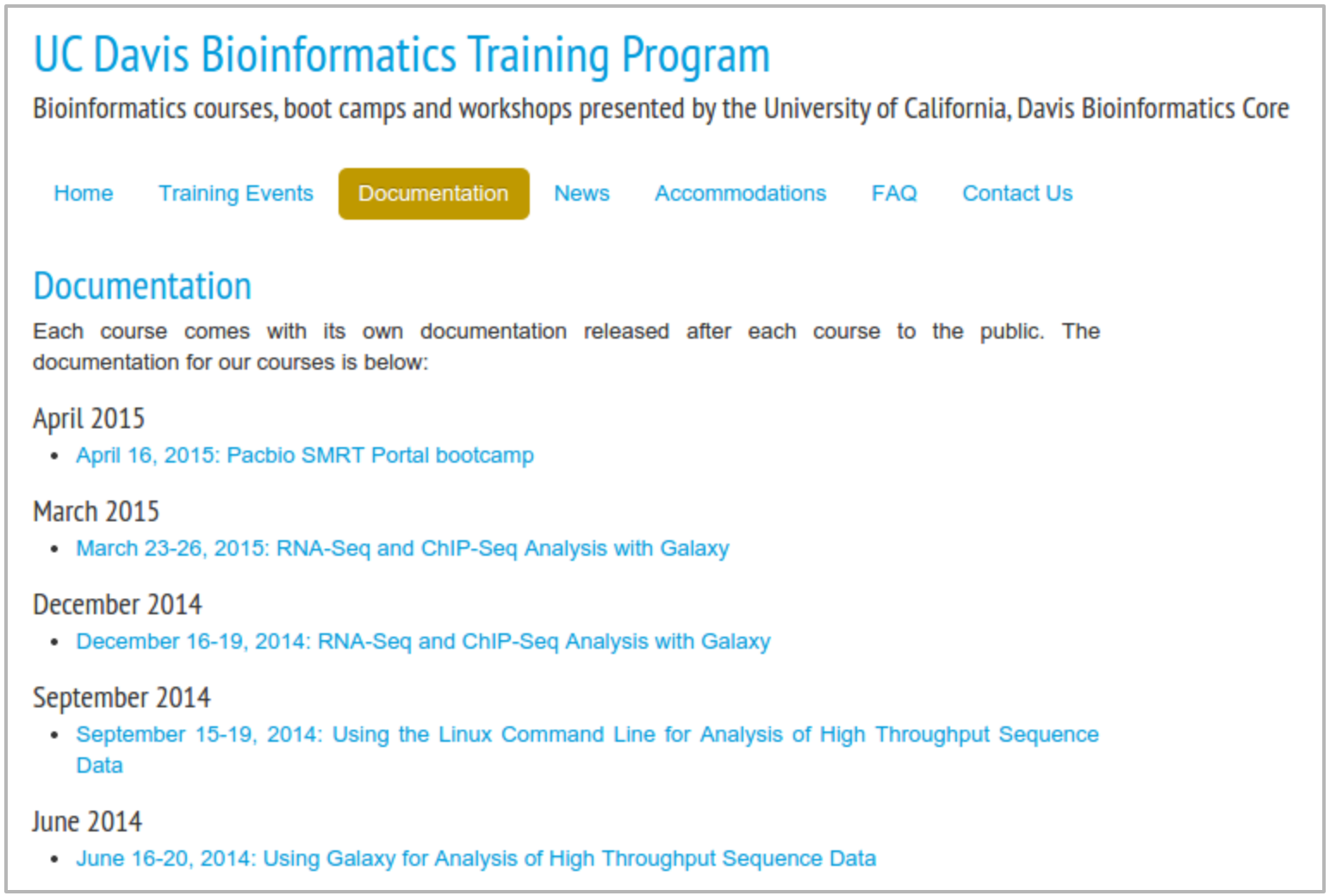
Upcoming

- Using the Linux Command Line for Analysis of High Throughput Sequence Data, June 15-19, 2015
- Using Galaxy for Analysis of High Throughput Sequence Data, September 14-18, 2015

Recent events

- RNA-Seq and ChIP-Seq Analysis with Galaxy, March 23-26, 2015
- Using the Linux Command Line for Analysis of High Throughput Sequence Data, September 15-19, 2014
- Using Galaxy for Analysis of High Throughput Sequence Data, June 16-20, 2014
- UC Davis Mission Critical Bioinformatics Workshop – iPlant/iAnimal Data to Publication, March 26-27, 2014
- Bootcamp: Introduction to Next Generation Sequence Analysis with Galaxy, December 10, 2013
- Bootcamp: Next Generation Sequence Alignment and Variant Discovery, December 11, 2013
- Bootcamp: Introduction to the Amazon Cloud for Galaxy and the Command-Line, December 13, 2013
- Bootcamp: Genome Assembly using Next Generation Sequence Data, December 12, 2013
- The 2013 RNA-Seq Workshop: From Pipette to P-value! (Sept 9-11, 2013)
- Bioinformatics Short Course 2013, September 16-20

Check Out the Documentation!



UC Davis Bioinformatics Training Program

Bioinformatics courses, boot camps and workshops presented by the University of California, Davis Bioinformatics Core

Home | Training Events | Documentation | News | Accommodations | FAQ | Contact Us

Documentation

Each course comes with its own documentation released after each course to the public. The documentation for our courses is below:

April 2015

- April 16, 2015: Pacbio SMRT Portal bootcamp

March 2015

- March 23-26, 2015: RNA-Seq and ChIP-Seq Analysis with Galaxy

December 2014

- December 16-19, 2014: RNA-Seq and ChIP-Seq Analysis with Galaxy

September 2014

- September 15-19, 2014: Using the Linux Command Line for Analysis of High Throughput Sequence Data

June 2014

- June 16-20, 2014: Using Galaxy for Analysis of High Throughput Sequence Data

training.bioinformatics.ucdavis.edu

But Wait …

There's More to Come!

Coming Soon to a Universe Near You ...

We're looking for beta testers!
email us at bioinformatics.core@ucdavis.edu

How To Contact the Bioinformatics Core

- email: bioinformatics.core@ucdavis.edu (best!)
- phone: 530-752-2698 (please consider emailing)
- stop by GBSF 1300, center aisle, under the clock (consider emailing first)

- and sign up for our announcements mailing list