Utilizing the Galaxy Analysis Framework at Core Facilities

Western Association of Core Directors (WACD)
September 18, 2015

Dave Clements
Galaxy Team
Johns Hopkins University
http://galaxyproject.org/

#usegalaxy @galaxyproject

Talk plan when I got off the train yesterday morning

1/3 What is Galaxy and what can it do?
1/3 Help clients to do their own data analysis w/ Galaxy!
1/6 Using Galaxy for in-house data pipelines
1/6 Q & A

http://galaxyproject.org

Talk plan after yesterday

1/3 ~~What is Galaxy and what can it do?~~
1/3 ~~Help clients to do their own data analysis w/ Galaxy!~~
1/6 ~~Using Galaxy for in-house data pipelines~~
1/6 ~~Q & A~~

2/10 What is Galaxy and what can it do?
1/10 Using Galaxy for in-house data pipelines
2/10 Help clients to do their own data analysis w/ Galaxy!
5/10 Open discussion:
What is the role of cores in supporting client data analysis?
Should this be part of your value proposition?

http://galaxyproject.org

---

1) The discussions yesterday were great. This is a great group. People aren't shy.
2) I really liked Ron and Jim's emphasis on "what's your value proposition"
3) Most of my outline presupposed that supporting your client's data analysis should be part of your core's value proposition.

I think that's actually an important question worth discussing

What is Galaxy?

Data integration and analysis platform that emphasizes accessibility, reproducibility, and transparency

http://galaxyproject.org

Says everything!

What is Galaxy?

Keith Bradnam's definition:

"A web-based platform that provides a simplified interface to many popular bioinformormatics tools."

From

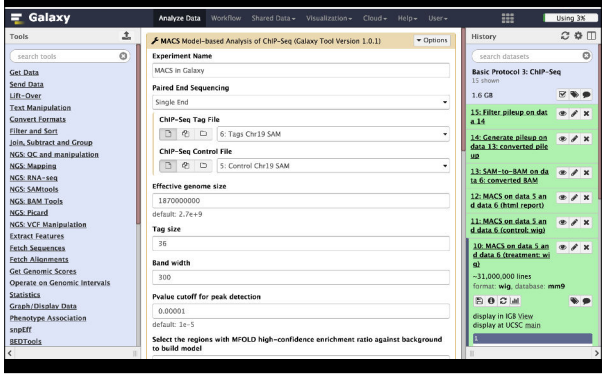
"13 Questions You May Have About Galaxy"

http://bit.ly/13questions

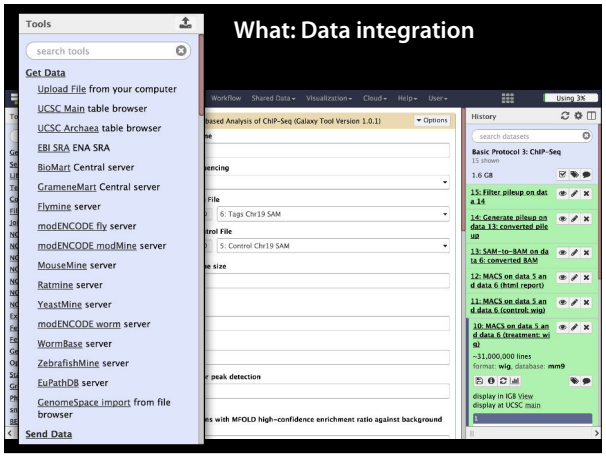
More more informative and concrete, but also somewhat too narrow.

This is a brilliant talk that I wish I had written.

What: A web based platform



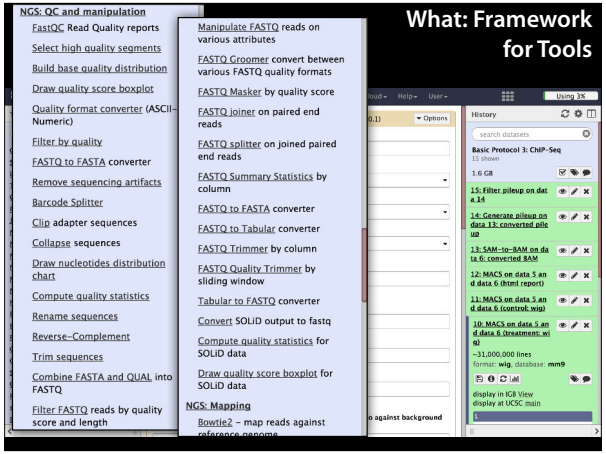
It's a web based platform.
Here's a screenshot showing the typical three panel design
Introduce each of those panels



Data integration: Can define data sources to Galaxy. This particular server has these defined datasources. These can talk directly to this Galaxy instance.

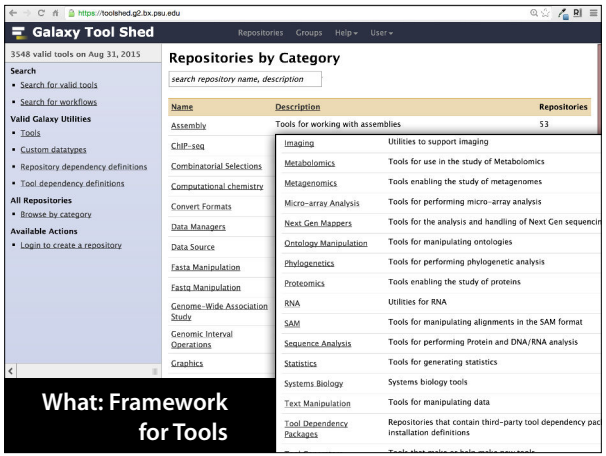
Can also upload data from computer or via a URL.

Will remember this indefinitely



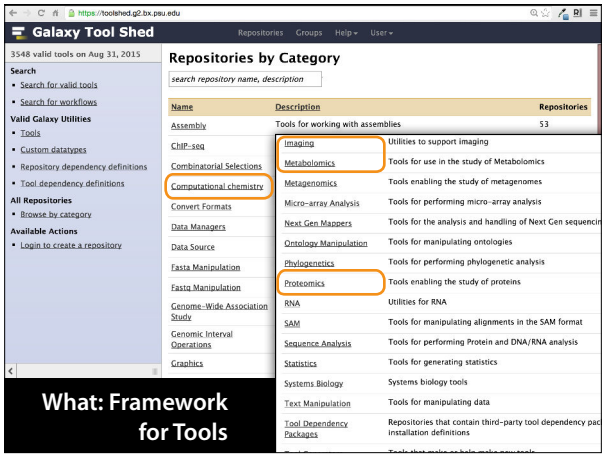
Platform for making tools easy to use.

Particular tool set on any Galaxy instance is determined by the site admin.

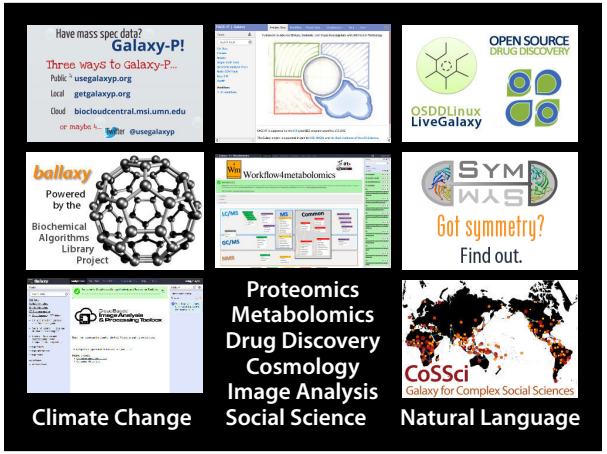


The tools that are on any given Galaxy instance are controlled by the administrators of that Galaxy Instance.

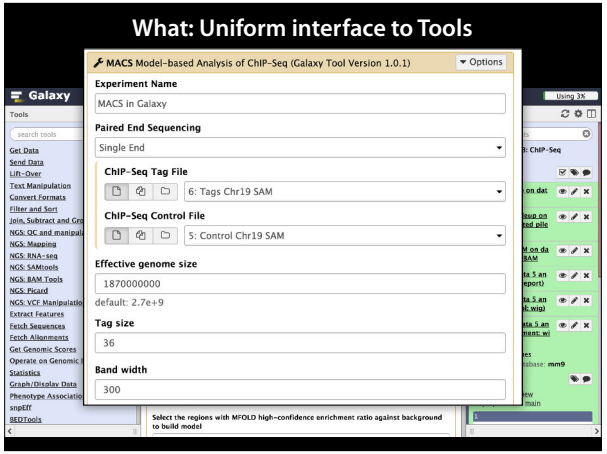
Where do those tools come from? Admins can wrap tools themselves, or they can get tool definitions from the community on the Galaxy Tool Shed.



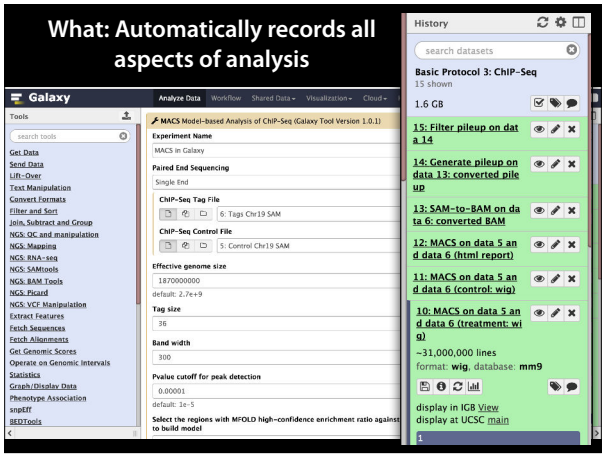
Note that there are a couple of categories here that may surprise people. Galaxy was originally developed for genomic data analysis, but the platform is actually domain agnostic.



In fact Galaxy is used in all sorts of domains, some of them having nothing to do with life sciences.

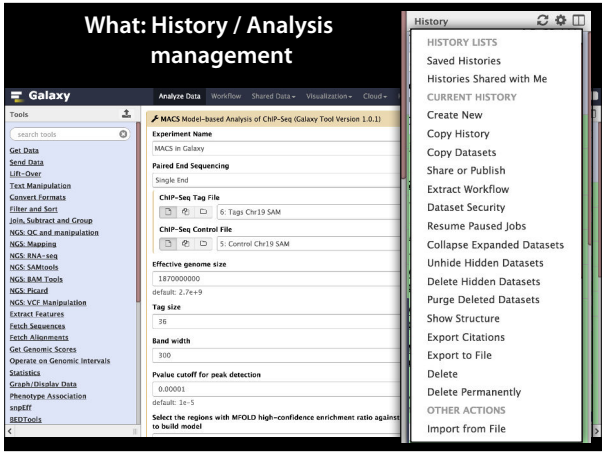


Uniform graphical interface to data analysis tools
Tools are often available only via command line otherwise

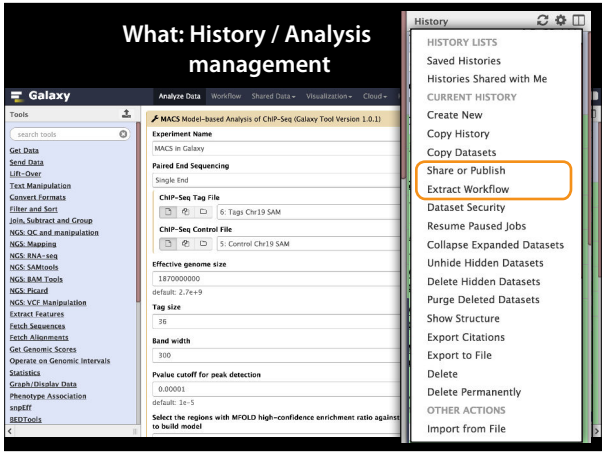


Has a built in history mechanism that automatically tracks all analysis, including datasets and tools used, and settings used with each tool

Galaxy will remember this indefinitely

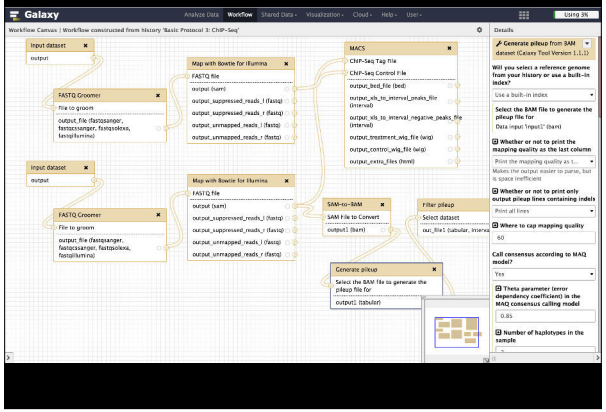


Histories also have a whole set of options.



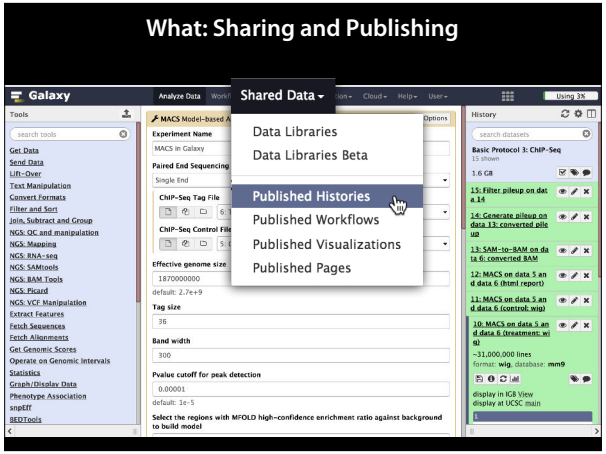
Two in particular I want to focus on

What: Create reusable workflows



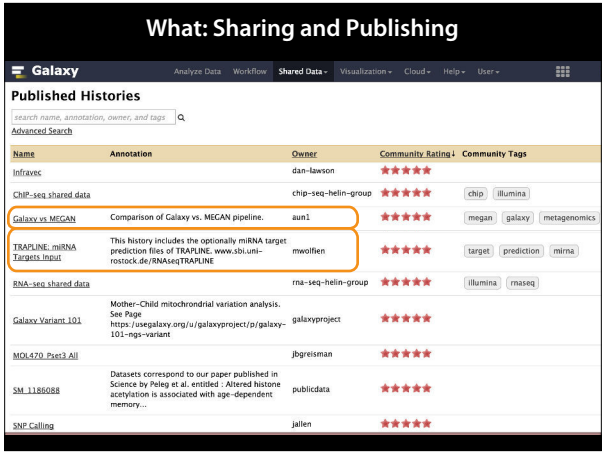
Create reusable workflows.
Can be created de novo, or by extracting pipelines from analysis.

What: Sharing and Publishing



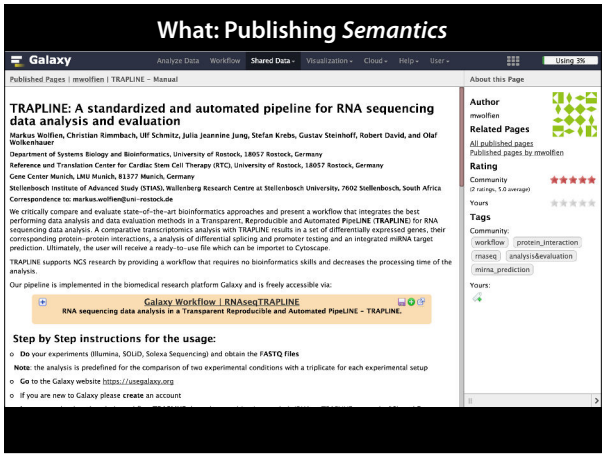
Anything on Galaxy can be shared or published.
If you publish a history it will show up in the published histories list and be visible to anyone who can access the server.
Shared/published histories can be examined in full detail and be imported into your own workspace.

What: Sharing and Publishing



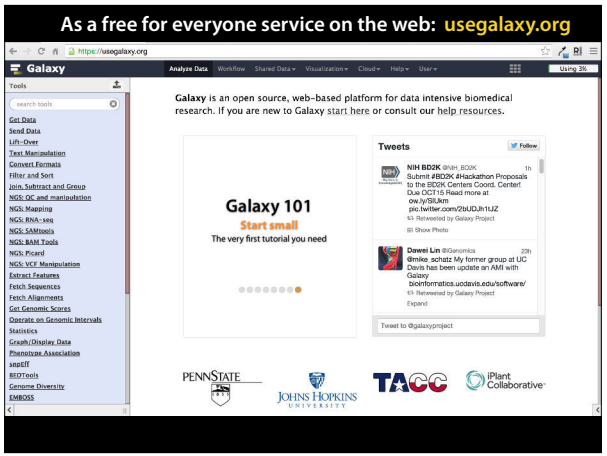
Any analysis on a Galaxy instance can be shared with specific other users on that instance, or published, and shared with everyone who has access to that server.

The 1st highlighted one is actually pointed to by a paper from a few years ago.
The 2nd highlighted one is being presented at the German Conference on Bioinformatics later this month.
People can look at these histories and see all parameters and tool versions used.
Can also import them into their workspace and rerun or tweak the



You can share histories (which are particular analysis), and workflows, which are reusable pipelines. These both describe the syntax of the analysis.

You can also create a Galaxy Page to describe the semantics of your analysis. Possible to embed any Galaxy Object such as a history in this.



All of the screen shots we've seen are from usegalaxy.org, the galaxy project's public Galaxy server.

A free for everyone web service:

http://usegalaxy.org

A free (for everyone) web server integrating a wealth of tools, compute resources, petabytes of reference data and permanent storage

TACC iPlant Collaborative

However, *a centralized solution cannot support the different analysis needs of the entire world.*

bit.ly/gxyServers

UseGalaxy.org is not the only publicly accessible server.

There are over 70 of them, we'll probably be over 80 by the end of the year.

Galaxy is available as Open Source Software

Galaxy is installed in locations around the world.

http://getgalaxy.org

Galaxy is available on the Cloud

http://aws.amazon.com/education
http://globus.org/
http://wiki.galaxyproject.org/Cloud

Galaxy on the Cloud: Galaxy CloudMan
http://usegalaxy.org/cloud

- Start with a **fully configured and populated** (tools and data) Galaxy instance.
- Allows you to scale up and down your compute assets as needed.
- Someone else manages the data center

Why Galaxy
inside Core Facilities?

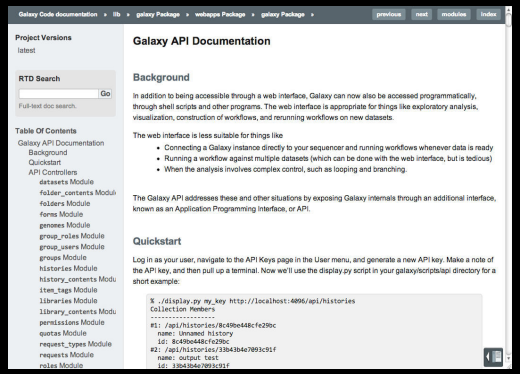
Get the advantages of Galaxy in your core

Job tracking
Analysis histories
Reusability
Reproducibility
Data Management

What does it take to host a Galaxy instance?

- Someone who knows how to admin and maintain it.
  - This is a non-trivial task
- Access to compute resources
  - Most institutions will have a compute core
    - Galaxy plays well with these
- Someone to do training
- Data retention policies, and a sustainable model.
- Can seem daunting, but there is huge community and ecosystem out there to help.

Galaxy API



https://galaxy-dist.readthedocs.org/

The Galaxy API allows you to access Galaxy from programs or scripts.

Why Galaxy
for Core Facility Clients?

The rest of the talk addresses this point.

Why Galaxy for Core Facilities

Empower your clients to actually use the data you generate for them **without**

learning a programming language, command line / shell interfaces, Linux package management, ...

or **extensive hand-holding** from core facility staff

I've been teaching NGS data analysis with Galaxy for 4+ years.

I can tell you that a common situation is researchers get that terabyte of data and they know there is insight in it, but many have no idea how to get to it. And even getting started is daunting.

Empower your clients *with Galaxy*:
Low hanging fruit

Point them at a Galaxy server for their research domain when you give them their data.

bit.ly/gxyServers

At minimum, suggest that when you hand over the deliverable, that you point any genomics users at UseGalaxy.org and any proteomics users at usegalaxyp.org and any epigenetic users at cistrome, and ...

Empower your clients with Galaxy:
Moderate

Deliver data inside a Galaxy instance with appropriate tools and reference datasets, inside

a virtual machine image,
a Docker container,
or an Amazon Machine Image (AMI)

For this you need a ladder

Gives them a platform with their data, the reference genomes they need, the tools they need to explore their data.

Empower your clients with Galaxy:
High

Deliver data inside
a core hosted, or institution hosted
Galaxy instance.

For this you need a cherry picker

What does it take to host a Galaxy instance?

- Someone who knows how to admin and maintain it.
  - This is a non-trivial task
- Access to compute resources
  - Most institutions will have a compute core
    - Galaxy plays well with these
- Someone to do training
- Data retention policies, and a sustainable model.
- Can seem daunting, but there is huge community and ecosystem

---

Open discussion:

What is the role of cores in supporting client data analysis?

Should this be part of your value proposition?

2016 Galaxy Community Conference (GCC2016)

June 25-29, 2016
Bloomington, Indiana

galaxyproject.org/GCC2016

Training day topic voting ends a week from today.

The Galaxy Team

http://wiki.galaxyproject.org/GalaxyTeam

Acknowledgements

Matt Settles
Ann Norton
Bridget the AV Guru

WACD
ABRF

NIH
Johns Hopkins University
Penn State University
Huck Institute