



# **GALAXY**

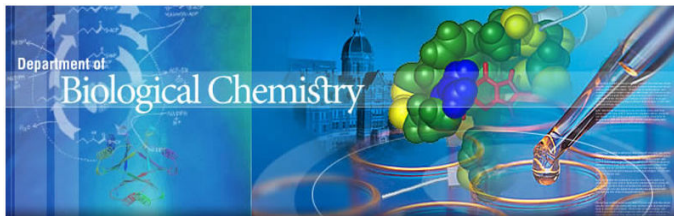
## COMMUNITY CONFERENCE

**BALTIMORE, MD | JUNE 30 - JULY 2, 2014**

Homewood Campus  
Johns Hopkins University  
Baltimore, Maryland  
United States

<http://galaxyproject.org/gcc2014>

#usegalaxy @galaxyproject



Platinum Sponsor

ion torrent



by *life* technologies™

Gold Sponsors



EMC ISILON



Silver Sponsors



Training Day Sponsor



Bronze Sponsor



# Welcome

Welcome to the 2014 Galaxy Community Conference (GCC2014). This is the fifth annual gathering of the Galaxy Community and every year the gatherings continue to grow. GCC2014 features three days of training sessions, talks, posters, vendor exhibits, birds-of-a-feather (BoF) gatherings, and lightning talks, all about high-throughput biology research and its supporting compute infrastructure. This event also features plenty of time for networking and impromptu gatherings. This year, for the first time, GCC will be preceded by a multi-day Galaxy Hackathon, focused on extending Galaxy's capabilities in novel ways.

The 2014 Training Day continues the 5 track, 3 sessions per track format introduced in 2013. However, this year, each session is half hour longer than last year (and a full hour longer than in 2012). Training topics were nominated and selected by the Galaxy community and reflect the truly wide range of interests in the community. This is again an excellent opportunity to get hands-on experience while learning from the experts.

We would like to give an enormous thanks to our sponsors and hosts, the Training Day instructors, the Scientific Committee, BoF organizers, speakers and poster presenters, and to anyone else who helped contribute to making this event a success. We would like to especially thank Stacey Hooker, the JHU Conference Coordinator for this event, for all her efforts (and patience!) while planning this event.

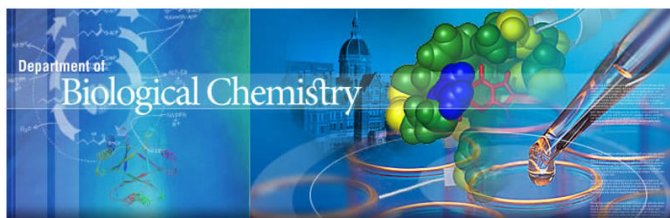
Thanks you,

The GCC2014 Organizing Committee



**GALAXY**  
COMMUNITY  
CONFERENCE  
BALTIMORE, MD | JUNE 30 - JULY 2, 2014

## Many Thanks to Our Hosts



## Network

Networking at GCC2014 is setup to accommodate those needing robust wired connections, and those needing wireless connections. Wifi is supported, and there are also lots of wired connections, as well. Where you are seated determines your networking options.

## E Tables / Ethernet Connections

A number of tables in the main ballroom (Salons A, B, and C) will have a big letter "E" on them. Ethernet connections are available at these tables, and if your laptop/device can use an Ethernet connection than we encourage you to consider sitting at these tables. There are no login credentials.

We choose to claim that this setup provides *more high-speed wired connections than any other conference you have been to, ever!*

## Wifi Connections & Credentials

All other tables in the ballroom and all tables in other rooms are considered wifi tables. If your device only supports wifi connections, then you are encouraged to sit at these tables (although wifi will work anywhere in the conference facility).

We also ask that you avoid large data transfers and streaming over the wireless network.

## One Device Limit!

Please limit yourself to having at most one device connected at a time! We are tech-savvy people at this event and if even a fraction of us connect all our devices at once we will swamp the network.



<http://penguincomputing.com>

## Meals

A continental breakfast is provided in the main ballroom (Salons A-C) every day from 8 to 9am. Lunches will also be catered in those rooms every day. Each lunch is sponsored by one of our Gold Level Sponsors:

Monday: Sponsored by EMC Isilon  
Tuesday: Sponsored by the BioTeam  
Wednesday: Sponsored by SGI

The official conference dinner is Wednesday evening. It will be held in the Mattin Center courtyard, directly across Charles Ave. from Charles Commons. The conference dinner is sponsored by Ion Torrent.

Please thank these sponsors profusely for their support!

## Breaks

Coffee, other beverages, and light snacks will be available during breaks. On Tuesday and Wednesday break refreshments will be served in the Sponsor Room (Barber Room 302) and the two Poster Rooms (Multipurpose Room and East Room 304).

## Social Media

Unless requested by the speaker, Tweeting and other social media activity are actively encouraged. Post early, post often.

#usegalaxy

@galaxyproject

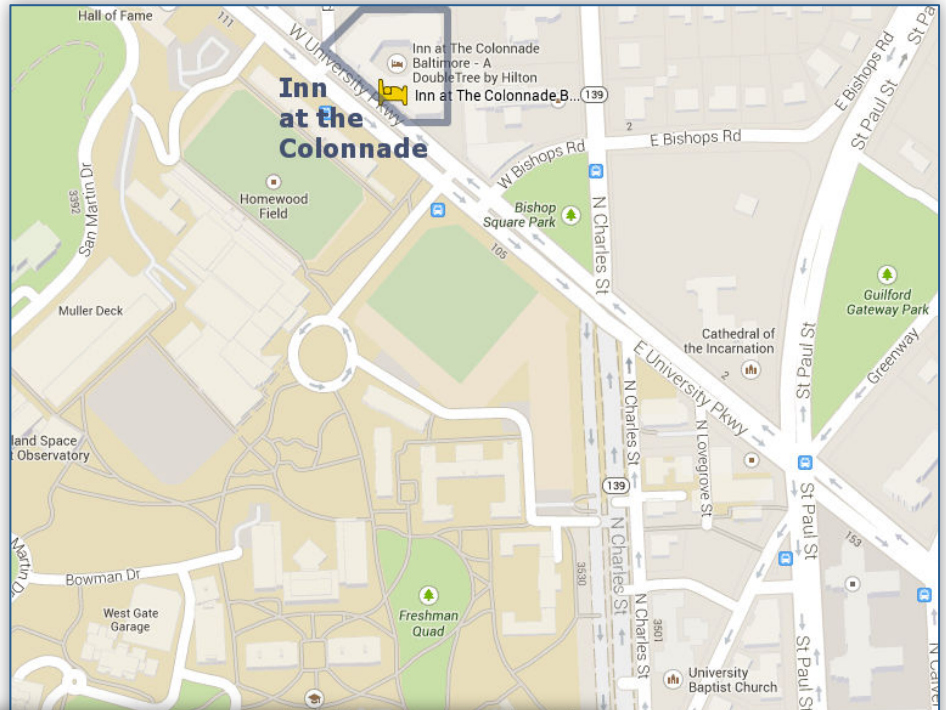


# Getting Around

GCC2014 is being held in *Charles Commons* on the Johns Hopkins University Homewood Campus. Charles Commons is also an official lodging option for GCC2014 and many participants are staying in the rooms in the floors above. Charles Commons is located in the vibrant neighborhood just east of campus.

The *Inn at the Colonnade* is another lodging option located within walking distance on the north end of campus.

The conference dinner will be held in the *Mattin Center* courtyard, directly east of Charles Commons.





I'm a  
**Bioinformagician**

Well, that's what everyone thinks. Thanks to Ion Torrent™ next-generation sequencing solutions, I can do plenty of amazing things. Like quickly and easily discover *de novo* mutations in my research, and follow up on relevant variants using Sanger sequencing in a streamlined workflow. Maybe there is a little magic to it after all.



Find out more at [lifetechnologies.com/seqmagic](http://lifetechnologies.com/seqmagic)

*life*  
technologies

A Thermo Fisher Scientific Brand

For Research Use Only. Not for use in diagnostic procedures. © 2014 Thermo Fisher Scientific Inc. All rights reserved. All trademarks are the property of Thermo Fisher Scientific and its subsidiaries unless otherwise specified. CO29108 0614

# Program



## Day 0: June 30, Monday, Training Day

Descriptions and prerequisites for these workshops are listed in the *Training Day* section of this program.

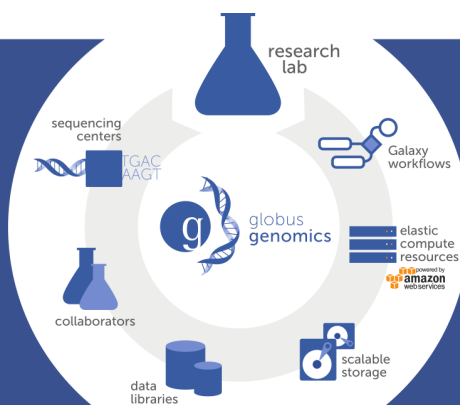
Time	Barber Room 302	Salon A	Salon B	Salon C	Multipurpose Room 324
8:00	Registration Opens and Catered Breakfast				
9:00	<b>Visualization of NGS data</b> Jeremy Goecks & Sam Guerler	<b>Raisins &amp; Rabbit Turds: NGS Quality Control with Galaxy</b> Tom Bair & Jennifer Jackson	<b>Galaxy Internals: Flow control within Galaxy</b> James Taylor	<b>Galaxy installation and administration</b> Nate Coraor & John Chilton	<b>Training with Galaxy: a Genome Assembly Example</b> Simon Gladman & Andrew Lonie
11:30	Lunch Sponsored by EMC Isilon				
12:30	<b>Galaxy on a Cluster - User and Project Management</b> Nikolay Vazov & Katerina Michalickova	<b>Galaxy Automation: Using the API</b> Dannon Baker & Carl Eberhard	<b>Tool Development from bright idea to toolshed - Designing a Galaxy Tool</b> Greg Von Kuster, Björn Grüning & Peter Cock	<b>RNA-Seq Analysis with Galaxy and the Tuxedo Suite</b> Saskia Hiltemann, Youri Hoogstrate & Hailiang (Leon) Mei	<b>3D Genome Analysis with Galaxy</b> Jonas Paulsen, Tonje Lien Gulbrandsen, Morten Johansen & Karen Reddy
3:00	Break				
3:30	<b>RNA-Seq Analysis with Galaxy and Alternative Tools</b> Saskia Hiltemann, Youri Hoogstrate & Hailiang (Leon) Mei	<b>Tool Development from bright idea to toolshed - Data Managers</b> JJ Johnson & Dan Blankenberg	<b>Visualization of NGS data</b> Jeremy Goecks and Sam Guerler	<b>Scriptable Bioinformatics Cloud Infrastructures with Cloud BioLinux, CloudMan &amp; Galaxy</b> Ntino Krampis, Enis Afgan, Ravi Sanka, Brad Chapman	<b>Galaxy on a Cluster - User and Project Management</b> Nikolay Vazov & Katerina Michalickova
6:00	Break				
6:15	Dinner (on your own) / Birds-of-a-Feather Flock I				
10:00	Finish				

Biologist-centric

Developer-centric

Flexible, scalable, affordable genomics analysis for all biologists.

[globus.org/genomics](http://globus.org/genomics)



“ At ICBI, we are working very closely with leading researchers to advance the frontiers of genomic science. By adopting Globus Genomics, we are much better positioned to deliver on our mission to enhance clinical and translational research at the medical center. ”

**Dr. Subha Madhavan**  
 Director of the Innovation Center for Biomedical Informatics  
 Georgetown University Medical Center

# Program: Day 1: July 1, Tuesday morning

8:00	Registration Opens and Catered Breakfast
9:00	Welcome & Opening
9:15	<p align="center"><b>Session 1</b></p> <p align="center">Moderator: Karen Reddy, Johns Hopkins University</p> <p>9:15 <b>Keynote: Transcriptomes and Exomes: Computational Challenges of NGS Data</b> Steven Salzberg, Biostatistics and Computer Science at the Johns Hopkins University School of Medicine</p> <p>10:00 <i>The Galaxy framework as a unifying bioinformatics solution for multi-omic data analysis</i> Pratik D. Jagtap, University of Minnesota</p> <p>10:15 <i>iReport: HTML Reporting in Galaxy</i> Saskia Hiltemann, Erasmus University Medical Center</p>
10:30	Break
11:00	<p align="center"><b>Session 2</b></p> <p align="center">Moderator: Hailiang (Leon) Mei, Leiden University Medical Center</p> <p>11:00 <i>Galaxy Deployment on Heterogenous Hardware</i> Carrie Ganote, National Center for Genome Analysis Support</p> <p>11:20 <i>Connecting Galaxy to tools with alternative storage and compute models</i> Brad Chapman, Bioinformatics Core, Harvard School of Public Health</p> <p>11:35 <i>A journal's experiences of reproducing published data analyses using Galaxy</i> Peter Li, GigaScience</p> <p>11:55 <i>Enabling Dynamic Science with Flexible Infrastructure</i> Anushka Brownley and Aaron Gardner, BioTeam</p>
12:15	Lunch - Sponsored by BioTeam



INTEL® XEON® E5  
PROCESSORS (16 CORES)

384 GB RAM

100GB SOLID STATE DRIVE

16TB INTEGRATED  
HIGH-SPEED STORAGE

FLEXIBLE LINUX-BASED  
SERVER ARCHITECTURE

CONFIGURATION  
AS SHOWN  
\$29,995.00 USD



## GET RESULTS FASTER WITH A DEDICATED GALAXY SERVER

Official Appliance Solution for the Galaxy Project

	NO WAIT TIMES	NO STORAGE QUOTAS	NO JOB SUBMISSION LIMITS	NO DATA TRANSFER BOTTLENECKS	NO REQUIRED TECHNICAL EXPERIENCE	NO REQUIRED INFRASTRUCTURE
GALAXY MAIN	✗	✗	✗	✗	✓	✓
LOCAL GALAXY	?	?	?	✓	✗	✗
CLOUD GALAXY	✓	✓	✓	✗	✗	✓
SLIPSTREAM GALAXY	✓	✓	✓	✓	✓	✓

**BE AN EARLY ACCESS PARTNER.**

For more information, visit [www.bioteam.net/slipstream/galaxy-edition](http://www.bioteam.net/slipstream/galaxy-edition)



## Program: Day 1: July 1, Tuesday afternoon

	<b>Session 3</b> Moderator: Liisa Koski, BASF Plant Science
1:15	<i>State of the Galaxy</i> Anton Nekrutenko and James Taylor, Galaxy Project
	1:50 <i>Update on Ion Torrent Sequencing – Accurate, Long Reads</i> Mike Lelivelt, Director of Bioinformatics and Software Products, Ion Torrent, part of Life Technologies
2:30	<b>Sponsor and Vendor Exhibition and Poster Session 1</b> (odd numbered posters)
	<b>Session 4</b> Moderator: Ravi Madduri, Argonne National Laboratory, and University of Chicago
4:00	4:00 <i>The Galaxy Tool Shed: A Framework for Building Galaxy Tools</i> Greg von Kuster, Penn State University
	4:20 <i>Integrating the NCBI BLAST+ suite into Galaxy</i> Peter Cock, The James Hutton Institute
	4:35 <i>deepTools: a flexible platform for exploring deep-sequencing data</i> Sarah Diehl, Max Planck Institute of Immunobiology and Epigenetics
	4:50 <b>Lightning Talks, Group 1</b>
5:30	<b>Break</b>
5:45	<b>Dinner (on your own) and Birds-of-a-Feather Flock 2</b>
10:00	<b>Finish</b>

# DATA IN MOTION

## WHEN YOU NEED IT, WHERE YOU NEED IT

**LIFE SCIENCES. IT'S ON ISILON.**

Over 200+ Life Sciences organizations rely on Isilon to manage data across next generation sequencing, proteomics and imaging technologies, high performance computing environments and end users. **Are you ready to deliver?**

- Extreme performance
- Massive scalability
- Unmatched efficiency
- Remarkable ease of use
- Continuous availability

[www.emc.com/isilon](http://www.emc.com/isilon)

**EMC<sup>2</sup>**

## Program: Day 2: July 2, Wednesday morning

8:00	Registration Opens and Catered Breakfast
9:00	<b>Welcome</b>
	<b>Session 5</b> Moderator: Mohammad Heydarian, Johns Hopkins University
9:10	<i>The GCC2014 Hackathon</i> Dannon Baker, Brad Chapman, John Chilton, Kyle Ellrott, and GCC2014 Hackathon Participants
9:30	<i>More Options, Less Time: Streamlining Access to Reference Datasets</i> Daniel Blankenberg, Penn State University
9:45	<i>Building More Powerful Galaxy Workflows with Dataset Collections</i> John Chilton, Penn State University
10:05	<i>An Appliance for Life Science Research: Isilon, Penguin and Galaxy</i> Patrick Combes, Senior Solution Architect for Life Sciences, EMC Isilon
10:25	<b>Break</b>
	<b>Session 6</b> Moderator: Tom Bair, Univeristy of Iowa
10:55	<i>Lab Specimen Tracking with Galaxy</i> Martin Čech, Penn State University
11:10	<i>The Munich NGS-FabLab for medical sequence data</i> Sebastian Schaaf, German Cancer Consortium (DKTK), and Ludwig Maximilians University (LMU)
11:25	<i>Galaxydx - A Web-server dedicated to diagnosis data analysis</i> Vivien DESHAIES and Alban LERMINE, Institut Curie and Mines ParisTech
11:40	<i>Using Galaxy and Globus to deliver Science as a Service</i> Ravi K Madduri, Argonne National Laboratory, and University of Chicago
11:55	<i>SGL UV: Harnessing the Big Brain Platform for Galaxy</i> James Reaney, Senior Director, Research Markets, SGI

## GCC2014 Committees

### Organizing Committee

Dave Clements  
Johns Hopkins University

Mohammad Heydarian  
Johns Hopkins University

Dan MacLean  
The Sainsbury Laboratory

Karen Reddy  
Johns Hopkins University

### Scientific Committee

Jeremy Goecks  
George Washington University

Jessica Kissinger  
University of Georgia

Anton Nekrutenko  
Penn State University

Karen Reddy  
Johns Hopkins University

James Taylor  
Johns Hopkins University

### Hackathon Committee

Dannon Baker  
Johns Hopkins University

Brad Chapman  
Harvard University

John Chilton  
Penn State University

Kyle Ellrott  
University of California Santa Cruz (UCSC)

With essential and tremendous support from

**Stacey Hooker**  
Johns Hopkins University

**Paula Davis**  
Johns Hopkins University

## Program: Day 2: July 2, Wednesday afternoon

12:15	Lunch - Sponsored by SGI	
1:15	<p align="center"><b>Session 7</b></p> <p align="center">Moderator: Jessica Kissinger, University of Georgia</p> <p>1:15 <i>Building a virtual research environment with Galaxy</i> Olivier Inizan, Mikael Loaec, URGI-INRA</p> <p>1:30 <i>The Australian Genomics Virtual Laboratory</i> Andrew Lonie, University of Melbourne</p> <p>1:45 <i>Galaxy on the GenomeCloud : Yet another on-demand Galaxy cloud, but only powered by Apache CloudStack</i> Youngki Kim, GenomeCloud</p> <p>2:00 <i>Test-driven Evaluation of Galaxy Scalability on the Cloud</i> Nuwan Goonasekera, University of Melbourne</p> <p>2:15 <i>Bioinformatics on AWS: New and Noteworthy Features</i> Angel Pizarro, Senior Solutions Architect, Amazon Web Services</p>	
	2:35	Sponsor and Vendor Exhibition and Poster Session 2 (even numbered posters)
	4:05	<p align="center"><b>Session 8</b></p> <p align="center">Moderator: Jeremy Goecks, George Washington University</p> <p>4:05 Lightning Talks, Group 2</p> <p>5:20 Closing</p>
		5:30
	5:45	Birds-of-a-Feather Flock 3
7:00	Conference Dinner - Sponsored by Ion Torrent	

**MORE  
HEADROOM**

BIG THINKERS TRUST SGI

[www.sgi.com](http://www.sgi.com)

intel  
inside  
XEON

sgi

# Lightning Talks

Topics for lightning talks will be solicited during the meeting, and will be presented during Session 4, on Tuesday and Session 8 on Wednesday. If you wish to give a lightning talk, please send it to [outreach@galaxyproject.org](mailto:outreach@galaxyproject.org) before the start of Session 2 (Tuesday) or the start of Session 6 (Wednesday). The slides for all lightning talks will be made available on the conference web site, and the talks may be videotaped and also posted on the conference web site.

## Goals

This is your opportunity to give an impassioned and enthralling talk about something that you care about - but you only have 300 seconds. Make every one count, because your audience may include people suffering from limited attention spans this late in the proceedings.

## Timing

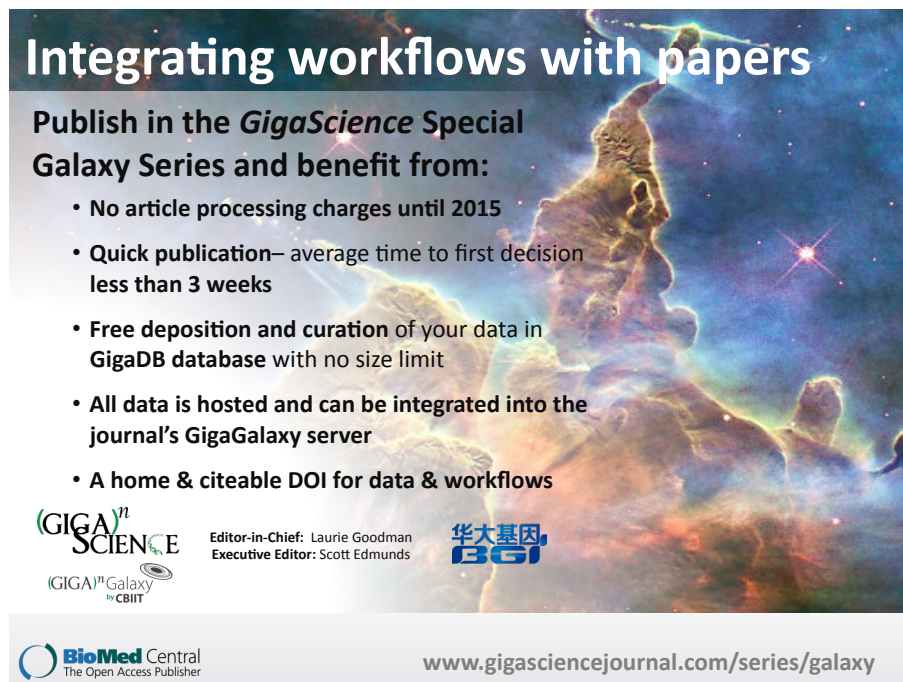
Lightning talks are 5 minutes followed by 2 minutes for questions.

- At 5 minutes in, thunder will be played
- At 6 minutes in we will take over the presentation laptop and start switching to the next set of slides.
- At 7 minutes the next talk will start, no matter what.

## Slides

- Your slides (as PDF or PowerPoint) should be on the presentation computer before the session starts (talk to Dave Clements)
- You can BYOD (your own computer or whatever) but you are advised not to.
- If you do BYOD, we'll start swapping out your device at 5 minutes, rather than 6.
- Connection and fiddling time comes out of your time and is painful, for everyone.

See <http://bit.ly/gcc2014lightning> for the list of lightning talks.



## Integrating workflows with papers

Publish in the *GigaScience* Special Galaxy Series and benefit from:

- No article processing charges until 2015
- Quick publication— average time to first decision less than 3 weeks
- Free deposition and curation of your data in GigaDB database with no size limit
- All data is hosted and can be integrated into the journal's GigaGalaxy server
- A home & citeable DOI for data & workflows

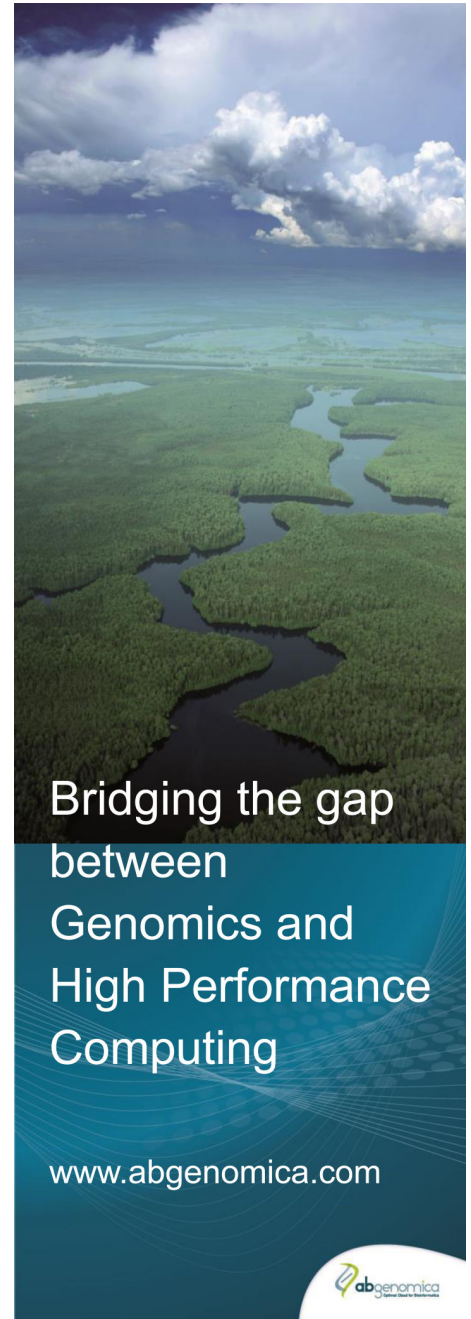
**(GIGA)<sup>n</sup> SCIENCE** Editor-in-Chief: Laurie Goodman  
Executive Editor: Scott Edmunds

**华大基因 BGI**

**(GIGA)<sup>n</sup> Galaxy** by CBIIT

**BioMed Central** The Open Access Publisher

[www.gigasciencejournal.com/series/galaxy](http://www.gigasciencejournal.com/series/galaxy)



## Bridging the gap between Genomics and High Performance Computing

[www.abgenomica.com](http://www.abgenomica.com)

**abgenomica**

# Feedback

Please give us feedback on the meeting at [bit.ly/gcc2014feedback](http://bit.ly/gcc2014feedback).

# Birds of a Feather Meetups



There is no better place than a Galaxy Community Conference to meet and learn from others doing high-throughput biology. GCC2014 continues this tradition by again including *Birds of a Feather (BoF)* meetups. Birds of a Feather are informal gatherings based on the participants' shared interests. BoFs are encouraged throughout GCC2014, particularly during the *flocking sessions* at the end of each day. These sessions are time set aside each evening specifically for Birds of a Feather gatherings.

If you are interested in a BoF then *just show up*. If you want to organize a BoF, see <http://bit.ly/gcc2014bofs> for how to get one going. It's never too late to start a BoF, and once one is proposed the organizers will get the word out about it.

The BoFs below were planned when this program was printed. See <http://bit.ly/gcc2014bofs> for BoFs that have been added since.

## Galaxy End-Users

This Birds-of-a-Feather session will serve as a forum for end-users of the Galaxy environment to share experiences and lessons learned, as well as address and discuss issues that hinder progress from the end-user perspective.

End-users of Galaxy who would like to share experiences (or listen to those of others) and developers interested in the perspective of the end-user should attend this BoF.

See <http://bit.ly/gcc2014usersbof>

**Meeting on Monday, June 30, 6:15pm**

## Doing the Branch, Release, and Merge Waltz

We will focus on branching and release management with regard to existing instances which implement customized code within Galaxy. This may create huge challenges in the future, especially for instances in production which require a lot of maintenance and which run older versions of Galaxy. All Clouds and Clusters which require multiple extensions like:

- huge file management (upload, etc)
- authentication issues
- cluster/cloud connectivity

And the customization of these issues is not easy and straightforward.

See <http://bit.ly/gcc2014mergebof>

**Meeting on Monday, June 30, 6:15pm**

## Using Galaxy with Heterogeneous and Remote Resources

We'll summarize recent efforts to enable Galaxy Main to send jobs to remote HPC resources (XSEDE) and invite others to share related experiences and requirements. We'll discuss a path to overcome challenges and better meet the needs of the research community.

Anyone (users or admins) interested in leveraging remote or heterogeneous hardware resources with their Galaxy instances.

See <http://bit.ly/gcc2014heterogeneousbof>

**Meeting at Wednesday, July 2, 5:45pm**

## GalaxyAdmins

GalaxyAdmins is for people that are responsible for administering large Galaxy instances. We meet online and at events like GCC2014, where a lot of us happen to be. GCC2014 coincides with the two-year anniversary of GalaxyAdmins starting up. This BoF was very well attended at GCC2013 and resulted in several actions items (now implemented). However, the past 12 months have been less successful at having this group meet online.

This meetup will discuss last year's action items, what we can do about meetups in the coming year, GalaxyAdmins leadership, and whatever else participants want to talk about.

See <http://bit.ly/gcc2014adminsbof>

**Meeting on Wednesday, July 2, 6:15pm**

See <http://bit.ly/gcc2014bofs> for more

**BIO X MARYLAND**  
FROM RESEARCH TO REALITY

**RESOURCES TO HELP YOU**

- Access capital
- Promote your business
- Refine your business strategy
- Develop partnerships
- Find a business location
- Expand your network
- Grow your workforce

[www.Bio.Maryland.gov](http://www.Bio.Maryland.gov)

Martin O'Malley, Governor  
Anthony G. Brown, Lt. Governor

# Training Day



## Day 0: June 30, Monday

Time	Barber Room 302	Salon A	Salon B	Salon C	Multipurpose Room 324
8:00	Registration Opens and Catered Breakfast				
9:00	Visualization of NGS data Jeremy Goecks & Sam Guerler	Raisins & Rabbit Turds: NGS Quality Control with Galaxy Tom Bair & Jennifer Jackson	Galaxy Internals: Flow control within Galaxy James Taylor	Galaxy installation and administration Nate Coraor & John Chilton	Training with Galaxy: a Genome Assembly Example Simon Gladman & Andrew Lonie
11:30	Lunch Sponsored by EMC Isilon				
12:30	Galaxy on a Cluster - User and Project Management Nikolay Vazov & Katerina Michalickova	Galaxy Automation: Using the API Dannon Baker & Carl Eberhard	Tool Development from bright idea to toolshed - Designing a Galaxy Tool Greg Von Kuster, Björn Grüning & Peter Cock	RNA-Seq Analysis with Galaxy and the Tuxedo Suite Saskia Hiltemann, Youri Hoogstrate & Hailiang (Leon) Mei	3D Genome Analysis with Galaxy Jonas Paulsen, Tonje Lien Gulbrandsen, Morten Johansen & Karen Reddy
3:00	Break				
3:30	RNA-Seq Analysis with Galaxy and Alternative Tools Saskia Hiltemann, Youri Hoogstrate & Hailiang (Leon) Mei	Tool Development from bright idea to toolshed - Data Managers JJ Johnson & Dan Blankenberg	Visualization of NGS data Jeremy Goecks and Sam Guerler	Scriptable Bioinformatics Cloud Infrastructures with Cloud BioLinux, CloudMan & Galaxy Ntino Krampis, Enis Afgan, Ravi Sanka, Brad Chapman	Galaxy on a Cluster - User and Project Management Nikolay Vazov & Katerina Michalickova
6:00	Break				
6:15	Dinner (on your own) / Birds-of-a-Feather Flock I				
10:00	Finish				

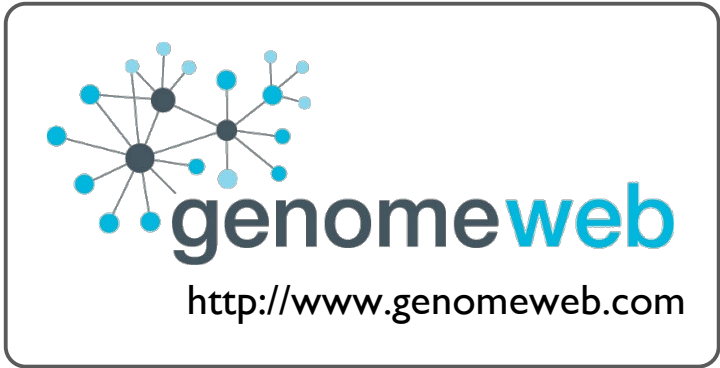
Biologist-centric

Developer-centric

## Prerequisites

These workshops are hands-on whenever possible. Since they are hands-on, they have prerequisites. If you don't meet the prerequisites, then you won't get much out of the workshop. If you don't have the prerequisites for most of the workshops, then you won't get much out of the day. Our goal is that when you walk out of a workshop, you will have meaningful experience with the workshop topic.

A wifi enabled laptop running a web browser is needed for all workshops.





AWS provides a comprehensive suite of tools to manage Scientific Computing workloads by utilizing services like: Amazon Elastic Compute Cloud (Amazon EC2) for scaling compute capacity up and down as needed, Amazon Simple Storage Service (Amazon S3) for storing data, and Amazon Elastic Map Reduce (Amazon EMR) to manage your Hadoop-based workflows. AWS allows you to increase the speed of research and to reduce costs by providing Cluster Compute or Cluster GPU servers on-demand so you do not have to wait in queues. Amazon EC2 Spot Instances in particular is a pricing model targeted for

batch processing use cases, providing your customers with the flexibility of ad-hoc provisioning while receiving significant price savings over other pricing models.

### Easy to use

AWS is designed to minimize the heavy lifting of setting up and managing your own IT infrastructure. You can get started with AWS by leveraging the AWS Management Console, a variety of third-party management tools, or the well-documented AWS web service APIs to manage and maintain your cloud infrastructure.

### Flexible

AWS enables you to select the operating system, programming language, software tools, application platform, and other services you need. This eases the migration process for existing applications while preserving options to build new ones.

### Sharing and collaboration

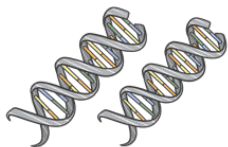
Creates a common space where you and your collaborators can share data, results and methods.

### Spot Pricing

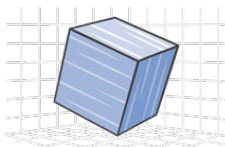
Spot Instances enable you to bid on unused Amazon EC2 capacity at whatever price customers choose. Customers whose bids exceed the Spot price gain access to the available Spot Instances and run as long as the bid exceeds the Spot Price. Historically, the Spot price has been 50% to 93% lower than the on-demand price. Customers whose bids exceed the Spot price gain access to the available Spot Instances and run as long as the bid exceeds the Spot Price. Spot Instances work with other services like Amazon S3 and Amazon EMR to help you manage all of your compute needs.

### Some example use cases that work well with Spot Instances include:

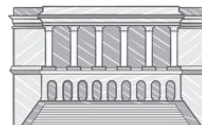
Genome processing



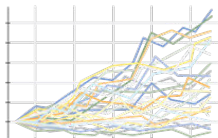
Modeling and Simulation



Government and Educational Research



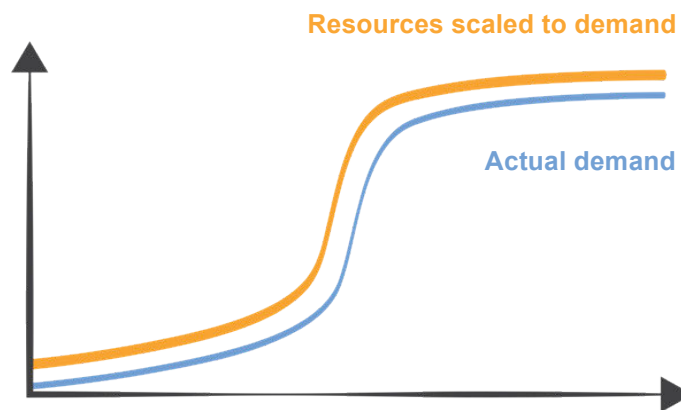
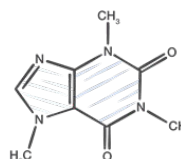
Monte Carlo Simulations



Transcoding and Encoding



Computational Chemistry




Elastic Cloud-Based Resources



# Biologist Centric Workshops




## Raisins and Rabbit Turds: NGS Quality Control with Galaxy

<b>Instructors</b>	Tom Bair, University of Iowa Jennifer Jackson, Penn State University	
<b>Content</b>	Often the first step in next generation sequencing data analysis is <i>quality control</i> . How reliable is the data? Does it have GC bias, or inaccuracies at the read ends, or contamination, or barcode corruption, or any number of other conditions that need to be detected and dealt with before the science begins. This workshop will provide hands-on experience performing quality control checks and how to get your data analysis-ready using Galaxy.  This workshop is also a good introduction to Galaxy for those who are not familiar with it.	  This workshop uses AWS-based compute infrastructure
<b>Prereqs</b>	A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.	

This title was inspired by Richard Smith's talk on "Experimental design: the importance of filtering" at the Iowa Institute for Human Genetics' Bioinformatics Short course

## Training with Galaxy: a Genome Assembly Example


<b>Instructors</b>	Simon Gladman, VLSCI Andrew Lonie, University of Melbourne	
<b>Content</b>	The Australian Genomics Virtual Laboratory (GVL) has developed a range of online tutorials based on Galaxy to aid in training and dissemination of bioinformatics expertise. The tutorials are completely self contained (data, workflows, rationale and background) and cover a range of introductory and advanced topics including genome assembly, variant detection and RNA-seq. This workshop will provide an overview of the available tutorials followed by a hands-on session based on a microbial genome assembly tutorial. To perform the analysis, participants will use cloud instances of the GVL platform.	  This workshop uses GVL-based compute infrastructure
<b>Prereqs</b>	<ul style="list-style-type: none"><li>A general knowledge of Galaxy (for example, you should be familiar with the material in Galaxy 101), or attendance at the "Raisins and Rabbit Turds: NGS Quality Control with Galaxy" session.</li><li>A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li></ul>	

## 3D Genome Analysis with Galaxy


<b>Instructors</b>	Jonas Paulsen, University of Oslo Tonje Lien Gulbrandsen, University of Oslo Morten Johansen, University of Oslo Karen Reddy, Johns Hopkins University	
<b>Content</b>	The session will introduce the basics of tracks and track types, and how these relate to hypothesis formulation and statistical analysis, using the Galaxy-based Genomic HyperBrowser. The emphasis will be on analysing, interpreting and integrating 3D genomic data (such as Hi-C), using the HiBrowse system. In addition to introducing the general concepts, the session will show examples on how 3D genome analyses can be combined with other HyperBrowser and Galaxy tools, in order to go from initial hypotheses to final results.	
<b>Prereqs</b>	<ul style="list-style-type: none"><li>A general knowledge of Galaxy (for example, you should be familiar with the material in Galaxy 101), or attendance at the "Raisins and Rabbit Turds: NGS Quality Control with Galaxy" session.</li><li>A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li></ul>	




## RNA-Seq Analysis with Galaxy and the Tuxedo Suite

<b>Instructors</b>	Saskia Hiltemann, Erasmus Medical Center Youri Hoogstrate, Erasmus Medical Center Hailiang (Leon) Mei, Leiden University Medical Center	
<b>Content</b>	<p>This hands-on workshop will demonstrate basic RNA-Seq transcript level comparison analysis using the Tophat (Bowtie), Cufflinks, Cuffmerge and Cuffdiff tools in Galaxy. We will compare the expression of genes under two conditions.</p> <p>We will demonstrate this analysis both with an installed reference genome and with a non-installed organism.</p> <p>Sample datasets small enough to be successfully processed during the course of the seminar will be provided. Participants will perform the analyses themselves on the provided cloud instance of Galaxy.</p>	 <p>This workshop uses AWS-based compute infrastructure</p>
<b>Prereqs</b>	<ul style="list-style-type: none"> <li>• A general knowledge of Galaxy and NGS quality control issues and tools, or attendance at the "Raisins and Rabbit Turds: NGS Quality Control with Galaxy" session.</li> <li>• A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li> </ul>	

## RNA-Seq Analysis with Galaxy and Alternative Tools

<b>Instructors</b>	Saskia Hiltemann, Erasmus Medical Center Youri Hoogstrate, Erasmus Medical Center Hailiang (Leon) Mei, Leiden University Medical Center	
<b>Content</b>	<p>The Tuxedo suite of RNA-Seq tools (Cuff*, Tophat, ...) are installed on many Galaxy instances, including Main and CloudMan installs. However, many other options are available. For example, Htseq, EdgeR and DESeq are also widely used, take a different approach to RNA-Seq analysis and return different results from the Tuxedo suite.</p> <p>This workshop would introduce alternative methods for RNA-Seq analysis, cover how to install them from the Tool Shed and to test they are properly installed. The workshop could finish by comparing results from these tools with those from the Tuxedo suite.</p>	 <p>This workshop uses AWS-based compute infrastructure</p>
<b>Prereqs</b>	<ul style="list-style-type: none"> <li>• A general knowledge of Galaxy and NGS quality control issues and tools, or attendance at the "Raisins and Rabbit Turds: NGS Quality Control with Galaxy" session.</li> <li>• Familiarity with the Tuxedo suite or attendance at the RNA-Seq Analysis with Galaxy and the Tuxedo Suitesession</li> <li>• A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li> </ul>	

## Visualization of NGS data

<b>Instructors</b>	Jeremy Goecks, George Washington University Sam Guerler, Johns Hopkins University	
<b>Content</b>	Different ways of visualizing NGS data more on downstream analysis such as heat maps, pathway networks and R based charts and graphs. This workshop will cover both primary NGS analyses --alignments, variants, annotations -- as well as downstream options.	 <p>This workshop uses AWS-based compute infrastructure</p>
<b>Prereqs</b>	<ul style="list-style-type: none"> <li>• A general knowledge of Galaxy (for example, you should be familiar with the material in Galaxy 101), or attendance at the "Raisins and Rabbit Turds: NGS Quality Control with Galaxy" session.</li> <li>• A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li> </ul>	


## Feedback

Please give us feedback on the meeting at [bit.ly/gcc2014feedback](http://bit.ly/gcc2014feedback).


# Deployment and Development Workshops



## Galaxy Installation and Administration

<b>Instructors</b>	Nate Coraor, Penn State University John Chilton, Penn State University	
<b>Content</b>	<p>Topics:</p> <ul style="list-style-type: none"><li>• Installing Galaxy on a standalone system</li><li>• Installing Galaxy in a cluster environment</li><li>• Common administrative tasks</li><li>• Tool installation (using Tool Shed and manually)</li><li>• Reference genome installation and configuration</li><li>• Misc. (user authentication, data libraries, other...)</li><li>• Upgrading</li><li>• Troubleshooting</li></ul>	 <p>This workshop will require that you have the VirtualBoxplayer (or VMwareplayer) installed on your laptop.</p>
<b>Prereqs</b>	<ul style="list-style-type: none"><li>• Knowledge and comfort with the Unix/Linux command line interface and a text editor. If you don't know what cd, mv, rm, mkdir, chmod, grep and so on can do then you will struggle in this workshop.</li><li>• Secure Shell (SSH) client software such as PuTTY for Windows, or the Terminal Application that comes with Mac OS.</li><li>• The virtual machine image (download from link to be provided) for this workshop.</li><li>• A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li></ul>	

## Galaxy on a Cluster - User and Project Management

<b>Instructors</b>	Nikolay Vazov, University of Oslo Katerina Michalickova, University of Oslo	
<b>Content</b>	<p>Galaxy is more and more often used as a front-end to huge HPC resources. At the same time, the HPC facilities require solid user authentication procedures and accounting mechanisms allowing to control the use of HPC resources. We will provide an overview of issues and several possible approaches the problem. Participants will then install a specific third party solution (GOLD) into a test Galaxy.</p>	 <p>This workshop will require that you have the VirtualBoxplayer (or VMwareplayer) installed on your laptop.</p>
<b>Prereqs</b>	<ul style="list-style-type: none"><li>• Experience maintaining a production Galaxy server (recommended)</li><li>• Secure Shell (SSH) client software such as PuTTY for Windows, or the Terminal Application that comes with Mac OS.</li><li>• The virtual machine image (download from link to be provided) for this workshop.</li><li>• A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li></ul>	

## Galaxy Internals: Flow control within Galaxy


<b>Instructors</b>	James Taylor, Johns Hopkins University	
<b>Content</b>	<p>Galaxy deployers often face problems in customizing the galaxy instance because of the lack of documentation that talks about how the control flows within Galaxy when job is run. This workshop will help deployers understand the Galaxy's internals.</p>	
<b>Prereqs</b>	<ul style="list-style-type: none"><li>• Knowledge and comfort with the Unix/Linux command line interface and a text editor. If you don't know what cd, mv, rm, mkdir, chmod, grep and so on can do then you will struggle in this workshop.</li></ul>	

Got a question?


<https://biostar.usegalaxy.org/>



## Galaxy Automation: Using the API

<b>Instructors</b>	Dannon Baker, Johns Hopkins University Carl Eberhard, Johns Hopkins University	
<b>Content</b>	Galaxy has a growing API that allows for external programs to control the system, search the resources, and issue work requests. The session would cover programmatic access of the API either by direct REST web calls or by using the BioBlend/blend4j APIs.	 <p>This workshop will require that you have the VirtualBoxplayer (or VMwareplayer) installed on your laptop.</p>
<b>Prereqs</b>	<ul style="list-style-type: none"> <li>• Knowledge and comfort with the Unix/Linux command line interface and a text editor. If you don't know what cd, mv, rm, mkdir, chmod, grep and so on can do then you will struggle in this workshop.</li> <li>• A knowledge of Python programming.</li> <li>• Secure Shell (SSH) client software such as PuTTY for Windows, or the Terminal Application that comes with Mac OS.</li> <li>• A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li> </ul>	


## Scriptable Bioinformatics Cloud Infrastructures with Cloud BioLinux, CloudMan & Galaxy

<b>Instructors</b>	Ntino Krampis, JCVI Enis Afgan, Ruđer Bošković Institute (RBI) Ravi Sanka, JCVI Brad Chapman, Harvard University	
<b>Content</b>	<p>This workshop will provide instruction on building bioinformatics infrastructures with Galaxy as front-end, combined with Cloud BioLinux for standardization and CloudMan for scalability in the back-end. It will be a technically-oriented workshop targeted to software developers, and will provide a tutorial how to jointly leverage the three systems for building bioinformatics applications on various cloud platforms including Amazon, OpenStack and Eucalyptus.</p> <p>The basics of deploying bioinformatics tools and pipelines on Galaxy running pre-configured on a Virtual Machine will be demonstrated. We will then move onto methods for standardizing deployment of complex bioinformatics pipelines through Galaxy by leveraging the Python Fabric scripts of Cloud BioLinux, in order to achieve interoperability and easy deployment across the various cloud platforms. The software blueprint of CloudMan for instantiating and using virtualized clusters connected to the Galaxy back-end will be presented, in addition to best practices for designing bioinformatics applications that leverage the distributed computing capabilities offered by the CloudMan framework.</p> <p>All concepts will be demonstrated through hands-on sessions where users will deploy tools through Galaxy, build VMs through Cloud BioLinux, instantiate clusters and data volumes and run distributed computing through CloudMan, using Amazon or Eucalyptus clouds.</p>	 <p>This workshop uses AWS-based compute infrastructure</p>
<b>Prereqs</b>	<ul style="list-style-type: none"> <li>• Knowledge and comfort with the Unix/Linux command line interface and a text editor. If you don't know what cd, mv, rm, mkdir, chmod, grep and so on can do then you will struggle in this workshop.</li> <li>• Secure Shell (SSH) client software such as PuTTY for Windows, or the Terminal Application that comes with Mac OS.</li> <li>• A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li> </ul>	


## Feedback

Please give us feedback on the meeting at [bit.ly/gcc2014feedback](http://bit.ly/gcc2014feedback).

## Tool Development from bright idea to toolshed - Designing a Galaxy Tool

<b>Instructors</b>	Greg Von Kuster, Penn State University Björn Grüning, University of Freiburg Peter Cock, James Hutton Institute	
<b>Content</b>	<p>Galaxy provides an easy way to create reproducible, sharable, easy-to-use analytical workflows... if every step of the analysis has a galaxy tool available to perform that application.</p> <p>The Galaxy Toolshed offers a place to share tools that can be imported into a Galaxy Server to complete an analysis workflow. Installation of a well-designed tool can be as simple as a couple button clicks by a Galaxy administrator.</p> <p>This session covers development process and the design considerations for stocking the toolshed with well-designed, easy-to-install tools. We will design a couple tools, determining how to lay out the inputs and parameters, generate the command line with the cheetah template, and add test cases. Then we'll submit them to a toolshed, and install them in our galaxy server.</p>	 <p>This workshop will require that you have the VirtualBoxplayer (or VMwareplayer) installed on your laptop.</p>
<b>Prereqs</b>	<ul style="list-style-type: none"> <li>• Knowledge and comfort with the Unix/Linux command line interface and a text editor. If you don't know what cd, mv, rm, mkdir, chmod, grep and so on can do then you will struggle in this workshop.</li> <li>• Secure Shell (SSH) client software such as PuTTY for Windows, or the Terminal Application that comes with Mac OS.</li> <li>• The virtual machine image (download from link to be provided) for this workshop.</li> <li>• A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li> </ul>	

## Tool Development from bright idea to toolshed - Data Managers

<b>Instructors</b>	JJ Johnson, University of Minnesota Dan Blankenberg, Penn State University	
<b>Content</b>	<p>Galaxy tools can require installed reference data in order to be used effectively. For example, Bowtie requires prebuilt indexes in order to efficiently map sequences to a genome.</p> <p>Data Managers enable a Galaxy administrator to add reference data to a Galaxy server via the admin webpage.</p> <p>This session covers the tool and toolshed requirements for using reference data within galaxy tools, and the design and development of tool data managers to install reference data on a Galaxy server.</p>	 <p>This workshop will require that you have the VirtualBoxplayer (or VMwareplayer) installed on your laptop.</p>
<b>Prereqs</b>	<ul style="list-style-type: none"> <li>• Knowledge and comfort with the Unix/Linux command line interface and a text editor. If you don't know what cd, mv, rm, mkdir, chmod, grep and so on can do then you will struggle in this workshop.</li> <li>• Secure Shell (SSH) client software such as PuTTY for Windows, or the Terminal Application that comes with Mac OS.</li> <li>• The virtual machine image (download from link to be provided) for this workshop.</li> <li>• A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.</li> </ul>	

Searching for something? <http://galaxyproject.org/search/>

Need just the right tool? <http://bit.ly/gxyshed>

# Talk Abstracts

## Session I: Tuesday, July 1, morning

### Transcriptomes and Exomes: Computational Challenges of NGS Data

Steven Salzberg<sup>1</sup>

<sup>1</sup> Professor of Medicine, Biostatistics, and Computer Science at the Johns Hopkins University School of Medicine. Director of the Center for Computational Biology at the McKusick-Nathans Institute of Genetic Medicine.

Steven Salzberg is a Professor of Medicine, Biostatistics, and Computer Science at the Johns Hopkins University School of Medicine where he is also Director of the Center for Computational Biology at the McKusick-Nathans Institute of Genetic Medicine. Steven has made many prominent contributions to open source software, including several of the most popular tools used on Galaxy Platforms. Recently he was awarded the 2013 Benjamin Franklin Award for Open Access in the Life Sciences, and the 2012 Balles Prize in Critical Thinking for his science column at Forbes.

### The Galaxy framework as a unifying bioinformatics solution for multi-omic data analysis

Pratik D. Jagtap<sup>1,3</sup>, James Johnson<sup>2</sup>, Getiria Onsongo<sup>2</sup>, Bart Gottschalk<sup>2</sup>, Timothy J. Griffin<sup>1,3</sup>

<sup>1</sup> Center for Mass Spectrometry and Proteomics, University of Minnesota, Minneapolis, Minnesota, United States

<sup>2</sup> Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota, United States

<sup>3</sup> Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, Minnesota, United States

Integration and correlation of multiple areas of 'omics' datasets (genomic, transcriptomic, proteomic) has potential to provide novel biological insights. Integration of these datasets is challenging however, involving use of multiple, domain-specific software in a sequential manner.

We describe extending the use of Galaxy for proteomics software, enabling novel, advanced multi-omic applications in proteogenomics and metaproteomics. Focusing on the perspective of a biological user, we will demonstrate the benefits of Galaxy for these analyses, as well as its value for software developers seeking to publish new software. We will also report on our experience in training non-expert biologists to use Galaxy for these advanced, multi-omic applications.

Working with biological collaborators, multiple proteogenomics and metaproteomics datasets representing a broad array of biological applications were used to develop workflows. Software required for sequential analytical steps such as database

generation (RNA-Seq derived and others), database search and genome visualization were deployed, tested and optimized for use in workflows.

Novel proteoforms (proteogenomic workflows, e.g., Galaxy Workflow: Integrated ProteoGenomics Workflow (ProteinPilot)) and microorganisms (metaproteomic workflows, e.g., Workflow for metaproteomics analysis - ProteinPilot') were reliably identified using shareable workflows. Tandem proteogenomic and metaproteomic analysis of datasets will be discussed using modular workflows. Sharing of datasets, workflows and histories on the [usegalaxy.org](http://usegalaxy.org) website and proteomic public repositories will also be discussed.

We demonstrate the use of Galaxy for integrated analysis of multi-omic data, in an accessible, transparent and reproducible manner. Our results and experiences using this framework demonstrate the potential for Galaxy to be a unifying bioinformatics solution for multi-omic data analysis.

### iReport: HTML Reporting in Galaxy

Saskia Hiltemann<sup>1</sup>, Youri Hoogstrate<sup>1</sup>, Hailiang Mei<sup>2</sup>, Guido Jenster<sup>1</sup>, Andrew Stubbs<sup>1</sup>

<sup>1</sup> ErasmusMC, Rotterdam, The Netherlands

<sup>2</sup> LUMC, Leiden, The Netherlands

Galaxy offers a number of great visualisation tools (Trackster, Circster), but currently lacks the ability to easily summarise the various outputs of a workflow into a single view. iReport is a Galaxy tool for the easy creation of HTML reports from Galaxy outputs. Rather than having a static HTML output, iFUSE2 uses javascript and jQuery to allow for interactivity in the form of searching and sorting of tables, automatic zooming of image data, tabbed view for organisation of outputs, etc. Users define the number and names of tabs for their report, and can add different types of content-items to these tabs (e.g. text, tabular data, image data, PDF files, links to datasets, and more).

We have previously implemented Galaxy-based data processing pipelines for next-generation sequencing (NGS) and for array based allelic copy number determination named CGtag (Hiltemann et al. 2014) and developed a web based fusion gene visualizer, iFUSE (Hiltemann 2013). We used the iReport tool to make iFUSE2, the next-step extension to support fusion gene determination within Galaxy, which runs as the last step of our workflow and combines the outputs of various Galaxy tools into a single view.

iReport is available from the DTL toolshed ([toolshed.dtl.nl](http://toolshed.dtl.nl)) and the main Galaxy toolshed.

## Session 2: Tuesday, July 1, late morning

### Galaxy Deployment on Heterogenous Hardware

**Carrie Ganote<sup>1</sup>**, Soichi Hayashi<sup>1</sup>

<sup>1</sup> National Center for Genome Analysis Support

Indiana University, like many institutions, houses a heterogenous mixture of compute resources. In addition to university resources, the National Center for Genome Analysis Support, the Extreme Science and Engineering Discovery Environment, and the Open Science Grid all provide resources to biologists with NSF affiliations. Such a diverse mixture of compute power and services could be applied to address the equally diverse set of problems and needs in the bioinformatics field.

Many software suites are well suited for large numbers of fast CPUs, such as phylogenetic tree building algorithms. *De novo* assembly problems really crave a machine with lots of RAM to spare. Alignment and mapping problems where each input is a separate invocation lend themselves perfectly to high-throughput, heavily distributed compute systems. Galaxy is a web interface that acts as a mediator between the biologist and the underlying hardware and software - in an ideal setup, Galaxy would be able to delegate work to the best suited underlying infrastructure.

We present an instance of Galaxy at Indiana University, installed and maintained by NCGAS, that takes advantage of a variety of compute resources to increase utilization and efficiency. The OSG is a distributed grid through which Blast jobs can be run. IU, NCGAS and XSEDE jointly support Mason, a 512Gb/node system. For IU users, Big Red 2 is the first university-owned petaFLOPS machine. Connecting these resources to Galaxy and using the best tool for the job results in the best performance and utilization - everyone wins.

### Connecting Galaxy to tools with alternative storage and compute models

**Brad Chapman<sup>1</sup>**, Rory Kirchner<sup>1</sup>, Oliver Hofmann<sup>1</sup>, Winston Hide<sup>1</sup>

<sup>1</sup> Bioinformatics Core, Harvard School of Public Health

The community developed bcbio-nextgen framework provides implementations of best-practice pipelines for variant calling and RNA-seq analysis. The framework handles computation, data storage and program connectivity in ways that parallel Galaxy's approaches, making it difficult to plug in as a standard tool. We'd like to be able to integrate with Galaxy by sharing the underlying implementation code for accessing data, rather than pushing and pulling large files. This talk will discuss ideas to access shared data on external object stores like S3 or HDFS in a consistent way that does not rely on data copying. It also will incorporate approaches to compartmentalize complex sets of tools inside containers using Docker. The goal is to stimulate discussion about ways to make Galaxy a modular component within complex analysis environments. Our ultimate vision is to have an Amazon based cloud implementation that uses CloudMan to run a Galaxy front end sending out jobs to tools like bcbio-nextgen.

### A journal's experiences of reproducing published data analyses using Galaxy

**Peter Li<sup>1</sup>**, Huayan Gao<sup>2</sup>, Tin-Lap Lee<sup>2</sup> and Scott C. Edmunds<sup>1</sup>

<sup>1</sup> *GigaScience*, BGI-Hong Kong Co., Ltd, Hong Kong <sup>2</sup> School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong

*GigaScience* is a journal with a focus on the publication of reproducible research. This is facilitated by its GigaDB databasewhere the data and the tools used for its analysis may be deposited by authors and made publicly available with citable DOIs. We have investigated the extent by which the results from articles published in *GigaScience* can be made reproducible using Galaxy in a pilot project based on a previously published paper reporting on SOAPdenovo2. The performance of this *de novo* genome assembler was compared with SOAPdenovo1 and ALL-PATHS-LG by Luo *et al.*, (2012) for its ability to assemble bacterial, insect and human genomes. After integrating the three genome assemblers, and their associated tools into Galaxy, workflows were implemented in a way that re-created the genome assembly pipelines used by the authors. However, our aim of reproducing the genome assembly statistics from Luo *et al.*, (2012) with the workflows was met with mixed success. Whilst the results generated by SOAPdenovo2 could be reproduced by our Galaxy workflows, we were less successful with SOAPdenovo1 and ALL-PATHS-LG. In this presentation, we will show how Galaxy was used, the problems that were encountered and the results of this reproducibility exercise.

#### Reference

Luo *et al.*, (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* 1:18.

### Enabling Dynamic Science with Flexible Infrastructure

**Anushka Brownley<sup>1</sup>**, Aaron Gardner<sup>1</sup>,

<sup>1</sup> BioTeam

As a trusted industry leader in designing and implementing effective scientific infrastructure for research and other organizations, BioTeam has partnered with the Galaxy Project to build and offer the SlipStream Galaxy Appliance, a commercially supported platform. With the increasing throughput of data generation instruments, the dynamic landscape of computational tools, and the variability in analysis processes, it is challenging for scientists to work within the confines of a static infrastructure. BioTeam will discuss some of these challenges and the technical advances we have been working on to build a more flexible Galaxy appliance to support the changing compute and analysis needs of the scientific researcher.

## Session 3: Tuesday, July 1, early afternoon

### State of the Galaxy

**Anton Nekrutenko<sup>1</sup> and James Taylor<sup>2</sup>**

<sup>1</sup> Penn State University

<sup>2</sup> Emory University

An overview of where the Galaxy Project is and where it is going.

### Update on Ion Torrent Sequencing – Accurate, Long Reads

**Mike Lelivel<sup>1</sup>**

<sup>1</sup> Director of Bioinformatics and Software Products, Ion Torrent, part of Life Technologies

## Session 4: Tuesday, July 1, late afternoon

### The Galaxy Tool Shed: A Framework for Building Galaxy Tools

**Greg von Kuster<sup>1</sup>** and the Galaxy Team

<sup>1</sup> Penn State University, State College, Pennsylvania, United States

The Tool Shed has become an integral part of the process for building and deploying Galaxy tools and other utilities. In addition to tools, the Tool Shed supports Galaxy Data Managers, custom data types and exported Galaxy workflows. This list will be extended to support additional utilities when appropriate. The Tool Shed provides the ability to define relationships between repositories, enabling complementary utilities to be installed together.

The Tool Shed assures reproducibility within Galaxy when utilities are installed from the Tool Shed using the streamlined installation process between the two applications. An underlying principle of this assurance is that all versions of utilities available in the Tool Shed will always be accessible to any Galaxy instance. This principle implies that a select development path should be followed to produce repositories that are optimal for sharing.

Here we'll examine the various components and steps that comprise this process. Development begins within a local environment that includes Galaxy and a Tool Shed, where a hierarchy of related repositories can be built. The Tool Shed allows the developer to export the related repositories into a capsule that can be imported into another Tool Shed. This mechanism streamlines the process of deploying utilities from a development environment to the test and main public Galaxy Tool Sheds where an automated install and test framework certifies the repositories for sharing. When installed together into Galaxy after certification, the related repositories provide complementary Galaxy utilities that function together.

### Integrating the NCBI BLAST+ suite into Galaxy

**Peter Cock<sup>1</sup>**, John Chilton<sup>2</sup>, Björn Grüning<sup>3</sup>, Jim Johnson<sup>4</sup>, Nicola Soranzo<sup>5</sup>

<sup>1</sup> The James Hutton Institute, Scotland, United Kingdom

<sup>2</sup> Department of Biochemistry and Molecular Biology, Penn State University, United States

<sup>3</sup> Pharmaceutical Bioinformatics, Institute of Pharmaceutical Sciences, Albert-Ludwigs-University, Freiburg, Germany

<sup>4</sup> Minnesota Supercomputing Institute, University of

Minnesota, Minneapolis, United States

<sup>5</sup> Bioinformatics Research Program, CRS4, Pula, Italy

NCBI BLAST is one of the best known computational tools in modern biology, and a common addition to Galaxy instances. This talk covers the history of the Galaxy wrappers for the NCBI BLAST+ command line tool suite, example use cases and workflows, and in particular our development process as a potential best practice model for Galaxy tool development - both technically and by showcasing Galaxy functionality, but also in terms of community building.

Initially included within the main Galaxy distribution, the BLAST+ wrappers are now run as a separate open source project using a dedicated repository on GitHub, combined with open discussion on the public Galaxy development mailing list.

The BLAST+ wrappers have grown to take advantage of most features offered by Galaxy and the ToolShed, including ToolShed dependencies, custom datatypes (including composite types for BLAST databases), configuration files for local databases, Galaxy tool XML macros to avoid duplication, and functional testing.

Automated testing is an important part of the development model and release process used. Integration with TravisCI provides continuous integration testing where any update to the code is automatically tested on a Virtual Machine. This is reinforced by a policy of staging updates to the Galaxy Test ToolShed for an additional round of automated testing, prior to release on the main Galaxy ToolShed.

Finally, an overview of how BLAST is setup on the Galaxy Instances we maintain will cover issues like job parallelization, thread and memory considerations, updating NCBI BLAST databases, and caching BLAST databases on cluster nodes.

### deepTools: a flexible platform for exploring deep-sequencing data

Fidel Ramírez<sup>1</sup>, Friederike Dündar<sup>1,2</sup>, Sarah Diehl<sup>1</sup>, **Björn A. Grüning<sup>3</sup>**, and Thomas Manke<sup>1</sup>

<sup>1</sup> Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany

<sup>2</sup> Faculty of Biology, University of Freiburg, Freiburg, Germany

<sup>3</sup> Department of Computer Science, University of Freiburg, Freiburg, Germany

We present a Galaxy based web server for processing and visualizing deeply sequenced data. The web server core functionality consists of a suite of newly developed tools, called deepTools, that enable users with little bioinformatic background to explore the results of their sequencing experiments in a standardized setting. Users can upload preprocessed files with continuous data in standard formats and generate heatmaps and summary plots in a straightforward, yet highly customizable manner. In addition, we offer several tools for the analysis of files containing aligned reads

and enable efficient and reproducible generation of normalized coverage files. As a modular and open-source platform, deepTools can easily be expanded and customized to future demands and developments. The deepTools webserver is freely available at <http://deeptools.ie-freiburg.mpg.de> and is accompanied by extensive documentation and tutorials aimed at conveying the principles of deepsequencing data analysis. The web server can be used without registration. deepTools is also available from the Galaxy toolshed, which allows an easy automated installation to any Galaxy instance.

## Session 5: Wednesday, July 2, morning

### The GCC2014 Hackathon

**Dannon Baker<sup>1</sup>, Brad Chapman<sup>2</sup>, John Chilton<sup>3</sup>, Kyle Ellrott<sup>4</sup>, and GCC2014 Hackathon Participants**

<sup>1</sup> Johns Hopkins University, Baltimore Maryland, United States

<sup>2</sup> Harvard University, Cambridge, Massachusetts, United States

<sup>3</sup> Penn State University, State College, Pennsylvania, United States

<sup>4</sup> University of California Santa Cruz (UCSC), Santa Cruz, California, United States

This year for the three days before GCC we are hosting a Galaxy Hackathon. Hackathons are events at which a group of developers with different backgrounds and skills collaborate hands-on and face-to-face to try to solve problems affecting a particular community, and in this case the Galaxy community. Gathering a diverse set of people in a single room where they can focus on code free of all the distractions that are inevitable back at the office has proven to be a great mechanism for not only getting interesting things done in a short amount of time, but also for community building. The hackathon goals include growing the Galaxy developer community and connecting existing developers who are interested in similar problems, giving them an in-person opportunity to code together and plan for future post-hackathon collaborations.

In this talk, we'll very briefly describe our Galaxy Hackathon goals and provide a general overview of progress made at the event. Since hackathons are by definition community driven, most of the talk will showcase the efforts of and be presented by the self-organizing groups that form during the event.

### More Options, Less Time: Streamlining Access to Reference Datasets

**Daniel Blankenberg<sup>1</sup> and the Galaxy Team<sup>2</sup>**

<sup>1</sup> Penn State University, State College, Pennsylvania, United States

<sup>2</sup> <http://galaxyproject.org/>

Recent enhancements to the Galaxy framework have introduced a new class of Galaxy Utilities, known as *Data Managers* (doi:10.1093/bioinformatics/btu119). Data Manager tools allow the Galaxy administrator to download, create and

install additional datasets for any type of built-in datasets using a web-based GUI in real time.

Despite these advances, populating a Galaxy instance with a set of built-in datasets can be quite time consuming, especially in cases where data not only needs to be downloaded, but additional computation, such as building indexes, is required. While this works quite well, it is wasteful to have each Galaxy installation build these datasets especially for common resources and genomes. It can take considerable amounts of time to populate a new Galaxy instance with needed datasets. Although the Galaxy Project provides a public rsync server with all of the built-in datasets that are used on the Main public site, utilizing this resource can be difficult and unwieldy, as there is a large amount of data and it lacks an accessible interface. While the individual location files are made available, they cannot be used as-is by an end user, unless the user has the exact same directory structure on their own machine that is hosting their Galaxy instance.

Here, we describe a new set of resources that aim to rectify this situation. These resources streamline the configuration of built-in data datasets for new and existing Galaxy instances and alleviate the technical barriers preventing many users from taking advantage of prebuilt reference datasets.

### Building More Powerful Galaxy Workflows with Dataset Collections

**John Chilton<sup>1</sup> and the Galaxy Team**

<sup>1</sup> Penn State University, State College, Pennsylvania, United States

Galaxy features the ability to extract a sample analysis histories out into reusable workflows as well as the ability to construct such workflows up from scratch or via modification to existing workflows. While these have been salient features of Galaxy for some time, the kinds of workflows that could be expressed by Galaxy have had critical limitations. Perhaps most glaring of these is that Galaxy workflows have required a fixed number of inputs. Many relatively basic biomedical analyses require running a variable number of inputs across identical processing steps ("mapping") and then combining or collecting these results into a merged output ("reducing"). This talk will present dataset collections - an extension to Galaxy that allows for the expression of these mapping, reducing workflows.

In particular, the concepts behind dataset collections will be covered including briefly discussing implementation details such



as data model modifications and API methods. Demonstration of how to “map” existing Galaxy tools across dataset collections to produce new collections and how to “reduce” these collections using other tools. Likewise, modification to the workflow extraction and editing interfaces to accommodate these new operations will be demonstrated.

Dataset collections are a powerful new feature that greatly enhance the expressivity of Galaxy workflows, but a lot work remains to do be done. The talk will conclude with a potential roadmap and timeline for dataset collection related development - including building UI components for digging into collections, building new collections, visualizing across

collections, and tool enhancements allowing tools to create collections.

## An Appliance for Life Science Research: Isilon, Penguin and Galaxy

**Patrick Combes**<sup>1</sup>

<sup>1</sup> Senior Solution Architect for Life Sciences, EMC Isilon

Isilon and Penguin Computing have paired to create a mid-size appliance for Galaxy by leveraging their respective strengths in storage and compute. This session will detail the architecture and projected use cases for the appliance.

## Session 6: Wednesday, July 2, late morning

### Lab Specimen Tracking with Galaxy

**Martin Čech**<sup>1</sup>, Pavel Švéda<sup>1</sup>, Ondřej Fabián<sup>1</sup> and the Galaxy Team

<sup>1</sup> Penn State University, State College, Pennsylvania, United States

No experiment begins with sequencing. Instead it commences with a collection of samples followed by DNA isolation (generation of cDNA, immunoprecipitation etc.), preparation of sequencing libraries, sequencing itself, and, finally, data analysis. In other words, during an NGS experiment a biological specimen undergoes transformation into a dataset to be analyzed. When an experiment involves a handful of samples, tracking the specimen-to-dataset metamorphosis is straightforward. However, low cost of sequencing enables individual single-PI laboratories to perform studies involving hundreds and even thousands of samples. At this scale tracking information about individual samples becomes challenging. Yet such tracking is essential for troubleshooting and ensuring a successful study. We have developed an open-source sample tracking system based on mobile devices carried by everyone in their pockets. The mobile application is able to communicate with a variety of sequencing instruments and trigger automated data analyses through the Galaxy system (<http://usegalaxy.org>).

### The Munich NGS-FabLab for medical sequence data

**Sebastian Schaaf**<sup>1,2</sup>, Aarif Mohamed Nazeer Batcha<sup>2</sup>, Sandra Fischer<sup>2</sup>, Guokun Zhang<sup>2</sup>, Ulrich Mansmann<sup>1,2</sup>

<sup>1</sup> German Cancer Consortium (DKTK), Heidelberg, Germany

<sup>2</sup> Department of Medical Informatics, Biometry and Epidemiology (IBE), Ludwig Maximilians University (LMU) Munich, Germany

Using NGS data in a clinical context comes along with a whole range of challenges, constraints and requirements, affecting all levels of an IT infrastructure dealing with that type of data – and related biomedical metadata. Especially in Germany, the restrictive data security laws play a key role. In 2010, the Munich regional area successfully applied for a grant ('Leading-

Edge Cluster Competition') dedicated to 'personalized medicine', supporting infrastructures for improving cross-connections between the medical faculties of both universities and associated institutions, their hospitals, independent research institutes (Helmholtz Centre, Max Planck Institutes) and industrial partners.

Aiming for a structured, biomedical metadata-driven organization of clinical NGS data, an interconnected, user-friendly, modular, broad-ranged and self-hosted open source analysis platform turned out to be crucial. Or in a nutshell: a Galaxy instance.

This talk is about the experiences of nearly three years of getting from blank to a conceptual Galaxy-driven NGS infrastructure, dedicated to scientist or clinicians from basic research up to experimental molecular diagnostics within a university medical center's environment. Topics will include experiences with core IT, faculty politics, project cooperations, software establishment etc. as well as derived Dos and Don'ts. Furthermore, some small software improvements will be presented, hopefully contributing back to the community. On top, we would like to draw connections to contents presented, discussed, improved since the last two GCC's in Chicago and Oslo - and also may have been forgotten. Over time, we had the impression to face several of them, pretty glad not to be in a minority of one.

### Galaxydx - A Web-server dedicated to diagnosis data analysis

**Vivien DESHAIES**<sup>1,2,3</sup>, **Alban LERMINE**<sup>1,2,3</sup>, Séverine LAIR<sup>1,2,3</sup>, Nicolas SERVANT<sup>1,2,3</sup>, Elodie GIRARD<sup>1,2,3</sup>, Julien TARABEUX<sup>4,5</sup>, Philippe HUPE<sup>1,2,3</sup>, Claude HOUDAYER<sup>4,5</sup>, Emmanuel BARILLOT<sup>1,2,3</sup>

<sup>1</sup> Institut Curie

<sup>2</sup> INSERM U900, Bioinformatics and Computational Systems Biology of Cancer, Paris, France

<sup>3</sup> Mines ParisTech, Fontainebleau, France

<sup>4</sup> INSERM U830, Génétique et biologie des cancers, Paris, France

<sup>5</sup> Biologie des Tumeurs, Paris, France

Early cancer diagnostic is a challenge that can dramatically improve cancer treatment efficiency. High throughput

sequencing technology is the more promising solution to reach this goal, but the analysis of their output is not straightforward and most of the time, need to launch software only available via command line interface.

Galaxy is a web platform that aim to: (1) make command line softwares accessible in an easy to use web interface, (2) construct personal workflows, (3) make analyses reproducible among time, (4) share know-how (workflow sharing) as well as data and annotations.

We built Galaxydx, an implementation of Galaxy containing a suite of softwares used for the analyses of diagnosis sequencing data (PGM torrent suite, BWA, GATK, VarScan, Annovar, ... etc). Galaxydx allows Clinicians as well as Biologists to be autonomous to perform a complete set of analyses such as: (1) mapping, (2) variant calling, (3) variant filtering, (4) variant annotation, (5) rearrangements calling and (6) visualization through diagnosis dedicated Genome browser (Alamut).

We also work on data integrity and confidentiality by modifying the Galaxy writing methodology. Analyses in Galaxydx are organized by project and user, output files are owned by the user who generates them. It allows us to systematically check system rights on data before any process (Can the current user read input data? Can the current user write in this project?)

## Using Galaxy and Globus to deliver Science as a Service

**Ravi K Madduri**<sup>1,2</sup>, Paul Dave<sup>2</sup>, Alex Rodriguez<sup>2</sup>, Vassily Trubetskoy<sup>3</sup>, Dinanath Sulakhe<sup>2</sup>, Lea Davis<sup>3</sup>, Nancy Cox<sup>3</sup> and Ian Foster<sup>1,2</sup>

<sup>1</sup> Argonne National Laboratory, Argonne, Illinois, United States

<sup>2</sup> Computation Institute, University of Chicago, Chicago, Illinois, United States

<sup>3</sup> Section of Genetic Medicine, University of Chicago, Chicago, Illinois, United States

At the Computation Institute, we originally posited the notion of science as a service in 2005 as a means of publishing and accessing scientific data and applications through well-defined and internet accessible services. Our vision of science as a service worked well in a world when computing resources were scarce; when we needed to federate heterogeneous

resources and make them accessible to researchers; when different tools and data were provided using different interfaces and representations; and when research problems involved datasets that could be hosted and analyzed on a single computer. In this talk we re-examine our vision of science as a service in a world in which computing resources are now commoditized; a world in which researchers are increasingly facing 'big data' challenges; a world in which Cloud providers, such as Amazon Web Services, have become viable alternatives to purchasing dedicated infrastructure; and a world in which building reliable infrastructure for solving scientific problems is only an API call away.

We will present our efforts on using Galaxy and Globus to create cloud-based services for scientific domains such as Genomics, Climate modeling, Cosmology, ECG Analysis and Material Sciences. We will present lessons learned, extensions we created to enable these communities adoption of Galaxy as an analysis engine. We will present a recent genomics usecase enabled using Galaxy based Globus Genomics on creating and running Consensus Genotyper for exome sequencing pipeline on large scale Tourette's Syndrome data set. (Joint work with Dr. Nancy Cox's group at UChicago.)

## SGI UV: Harnessing the Big Brain Platform for Galaxy

**James Reaney**<sup>1</sup>

<sup>1</sup> Senior Director, Research Markets, SGI  
GI UV scales to truly extraordinary levels – today up to 2,560 physical cores and 64TB of cache-coherent, globally shared memory in a single system. UV is also a developer's dream playground: standard Intel x86 architecture, standard Linux distros, support for large numbers of Nvidia GPU and Xeon® PHI®, and all those cores and memory at your disposal in a single OS. Run standard ISV applications or any open-source code just like any Linux instance, no recompiling necessary. The versatility, high performance, and extreme scale of UV makes it the ultimate "analysis supernode", but what if we used UV as an enabling platform for Galaxy workflows? How much more extensible might the tools become? What new scales might Galaxy workflows reach? What larger-scale research might be simply enabled in the first place by having a more effective computational architecture underlying the Galaxy workflow?

## Session 7: Wednesday, July 2, early afternoon

### Building a virtual research environment with Galaxy

**Olivier Inizan**<sup>1</sup>, **Mikael Loaec**<sup>1</sup>, Eric Rasche<sup>2</sup>, Hadi Quesneville<sup>1</sup>

<sup>1</sup> URGI-INRA, Versailles, France <sup>2</sup> Center for Phage Technology, Texas A&M University, College Station, Texas, United States

The democratization of virtualization techniques provide a new opportunity to improve bioinformatics analysis. Storing, sharing and reusing tools dedicated to an analysis is the goal of the galaxy toolshed project. With virtualization techniques, it is

now possible to expand their strategy to all the components required to perform a bioinformatic analysis such as the operating system, the software, the datasets, the dependencies, the user data, ...).

Integrating these components in a virtual machine provide a virtual research environment (VRE) that could be duplicated and shared. With the growing availability of infrastructures supporting virtualization (such as cloud computing infrastructures), VREs offer a new opportunity to improve bioinformatics analysis accessibility and reproducibility.

Accessibility and reproducibility are the building blocks of the Galaxy project and the Galaxy platform could play a significant

role in such environments. However, to become accessible and shareable, creating and updating a VRE should be automated as much as possible, from the virtual machine provisioning to tools deployment and tests.

Here we describe our progress towards an automation process for the deployment of a Galaxy instance. The current work is focused on virtual machine provisionment with Cobbler and automatic configuration with Puppet. The opportunities that such an approach provides to developers and biologists will be discussed, illustrated on the future French infrastructures dedicated to cloud computing: the IFB and INRA academic Clouds.

## The Australian Genomics Virtual Laboratory

**Andrew Lonie**<sup>1</sup>, Enis Afgan<sup>2,3</sup>, Ron Horst<sup>4</sup>, Simon Gladman<sup>5</sup>, Clare Sloggett<sup>1</sup>, Nuwan Goonasekera<sup>1</sup>, Igor Manukin<sup>4</sup>, Yousef Kowsar<sup>4</sup>

<sup>1</sup> Life Sciences Computation Centre, University of Melbourne, Australia

<sup>2</sup> University of Melbourne, Australia

<sup>3</sup> Ruđer Bošković Institute, Croatia

<sup>4</sup> University of Queensland, Australia

<sup>5</sup> Life Sciences Computation Centre, Monash University, Australia

The Australian Genomics Virtual Laboratory (GVL) is a national program aiming to provide the research community with an accessible, scalable genomics analysis platform on national compute infrastructure. The GVL leverages a significant investment in cloud infrastructure by the Australian government and existing cloud management tools to enable researchers to create on-demand genomics analyses environments based on the open source Galaxy workflow platform, linked through high speed networks to very large reliable data storage, and local instances of visualization engines like the UCSC browser.

This talk will discuss the technical and practical lessons learned during the development of the Genomics Virtual Lab, including considerations in defining and implementing a one-size-fits-all pre-configured Galaxy image, the constraints a cloud environment places on practical 'real data' genomics, identification of and interaction with the user base, and deliberations on the future of the Genomics Virtual Laboratory including architecting for the entire genomics analysis life cycle on the cloud.

## Galaxy on the GenomeCloud : Yet another on-demand Galaxy cloud, but only powered by Apache CloudStack

**Youngki Kim**<sup>1</sup>, CB Hong<sup>1</sup>, Kjoong Kim<sup>1</sup>, Daechul Choi<sup>1</sup>

<sup>1</sup> GenomeCloud, Seoul, Korea

Bioinformatics and genome data analysis in South Korea is at its early stage but getting busier. To keep pace with this trend of research, GenomeCloud was created at the end of 2012. GenomeCloud is an integrated platform for analysing, interpreting and storing genome data, based on KT's cloud computing infrastructure which uses Apache CloudStack software. GenomeCloud consists of g-Analysis

(automated genome analysis pipelines at your fingertips), g-Cluster (easy-of-use and cost-effective genome research infrastructure) and g-Storage (a simple way to store and share genome-specific data).

Because of flexible tool integration architecture and seamless workflow creation functionality, Galaxy was selected to achieve multi purpose goals such as agile pipeline development and bioinformatics education support. To provide on-demand and Apache CloudStack based Galaxy cluster, we have automated virtual machine creation, clustering and various software setup including Galaxy.

Furthermore, seamless integration with GenomeCloud helps researchers not only create and manage Galaxy through a convenient web interface but also fully utilizes genome data in g-Storage. g-Storage is powered by OpenStack Swift and specially designed genome file transfer protocol.

Galaxy on the GenomeCloud uses Grid Engine as a Cloud HPC Solutions, Ganglia as a distributed monitoring system and LVM over NFS as a large volume shared storage, all of which are setup automatically upon request. This talk will be about our experiences while integrating Galaxy with GenomeCloud and use cases of Galaxy such as scalable bioinformatics education system and request fulfillment of RNA-seq analysis.

## Test-driven Evaluation of Galaxy Scalability on the Cloud

Enis Afgan<sup>1,2</sup>, Derek Benson<sup>3</sup>, and **Nuwan Goonasekera**<sup>1</sup>

<sup>1</sup> VLSCI, University of Melbourne, Melbourne, Australia

<sup>2</sup> CIR, RBI, Zagreb, Croatia

<sup>3</sup> Research Computing Centre, University of Queensland

To verify the essential functions of a Galaxy instance are being provided correctly to the end-user, functional testing of typical Galaxy tasks is important. In addition, for groups which intend to deploy their own Galaxy instances (on the cloud or otherwise), knowing the scalability characteristics of the instance with respect to the number of users, machine size, storage solution and cloud provider, is also important. By combining both functional and performance testing into one common testing infrastructure, we assessed both of these aspects with the same underlying test code.

With respect to the first aspect of assessing whether the basic functions of Galaxy are working correctly from an end-user perspective, functional testing was performed via the browser automation tool Selenium, which can mimic the exact actions of an end-user interacting with the application. We then extended these tests to use the Selenium Grid, which converted the functional test into a performance test by running the tests in parallel, thus simulating multiple concurrent users.

This presentation will describe how these two aspects were used to determine the scalability characteristics of Galaxy on the cloud. The presentation will discuss the following:

Describe how the same infrastructure is reused for testing the functional and scalability characteristics of Galaxy, using CloudMan;

Analyse how a number of variables, such as the number of users, machine size and storage option, affects scalability;

Provide insights into how Galaxy scales on the cloud, and what factors to consider when deploying on your own infrastructure;

Provide a reusable suite of tests for functionally verifying and benchmarking private GVL/Galaxy instances

Data and results collected to obtain above conclusions will be made publicly available and can act as reference data points for others reusing the presented system on their own Galaxy instances.

## Bioinformatics on AWS: New and Noteworthy Features

**Angel Pizarro<sup>1</sup>**

<sup>1</sup> Senior Solutions Architect, Amazon Web Services

In this talk, we will cover recent service and feature releases from Amazon Web Services, and how they apply to bioinformatics and scientific computing.



At the **Institute for Basic Biomedical Sciences** at the Johns Hopkins University School of Medicine, we lay the foundations of future advances in prevention, diagnosis and treatment.

Virtually all mechanistic understanding of disease, current treatments and diagnostic tools have as their foundation a basic science discovery; the more basic the discovery, the more far-reaching its effects. Fundamental research touches upon everything from diagnosis to treatment, in addition to therapy for conditions ranging from cancer to autoimmune disease.

**Research:** Our nine basic science departments study all the fundamentals, from solving protein structures to dissecting cell movement, from analyzing chromosome structure to deconstructing biochemical pathways. The eight interdisciplinary IBBS centers bring together experts from a vast range of scientific and medical backgrounds to study metabolism and obesity, pain, autism and mental illness, sensory loss, and other medical conditions in new and innovative ways. We're adopting new technologies and building new tools, and we're using them to track cells and molecules, crack the codes that control how genetic material is read, and rebuild tissues and organs.

We have so much exciting, inspiring research going on here. To learn more, visit us at [www.hopkinsmedicine/institute\\_basic\\_biomedical\\_sciences/](http://www.hopkinsmedicine/institute_basic_biomedical_sciences/).

**Education:** The education and training of future leaders in biomedical research is deeply rooted in the mission of IBBS. More than half of the Ph.D. candidates at the Johns Hopkins University School of Medicine choose to do their research in IBBS labs through one of the 13 Ph.D.-granting graduate programs. Additionally, IBBS faculty members teach nearly all fundamental preclinical courses for medical students.

## Posters

Posters are presented in the East Room and the Multipurpose Room. There will be two poster sessions.

**1<sup>st</sup> Poster Session: Tuesday, July 1, 2:30-4:00**

*Odd* numbered posters will be presented during poster session 1.

**2<sup>nd</sup> Poster Session: Wednesday, July 2, 2:30-4:00**

*Even* numbered posters will be presented during poster session 2.

**PI: Lifeportal - web portal to high performance computing resources at University of Oslo**

**Nikolay Vazov<sup>1</sup>, Katerina Michalickova<sup>1</sup>**

<sup>1</sup> University of Oslo, HPC group

One of the main goals of the HPC (High Performance Computing) services at University of Oslo, Norway, is to make the complex HPC resources accessible to wide audience with a varied degree of experience. The Lifeportal ([lifeportal.uio.no](http://lifeportal.uio.no)) is currently geared towards the biomedical research with a special emphasis on the next generation sequencing data processing while a text mining instance is being finalized.

In addition to the existing Galaxy core facilities, the Lifeportal has a set of newly developed features that are essential for the Galaxy - HPC functionality. Our poster will discuss:

- user authentication and authorization with the integration of the National Academic IDP based on SAML technology

- integrated user/project management module for project applications, authorization and management
- project accounting module based on an external resource allocation manager (GOLD)
- module for project reporting and providing feedback to the funding agency
- big file upload based on Filesender technology allowing to upload files up to 250 GB into Galaxy
- details of cluster deployment via SLURM DRMAA including:
  - Galaxy code modification allowing for user-selected cluster job parameters such as queue, time, memory, number of nodes and cores
  - export of Galaxy libraries for deployment of the core Galaxy tools on the cluster
  - general changes to tool wrappers needed for cluster implementation

## P2: Building a scalable Galaxy cluster for biomedical research in The Netherlands

David van Enckevort<sup>1</sup>, Anthony Potappel<sup>2</sup>, Niek Bosch<sup>3</sup>, Jeroen Beliën<sup>4</sup>, Rita Azevedo<sup>5</sup>, Rob Hooft<sup>5</sup>, Sander Ruiter<sup>2</sup>, Sanne Abeln<sup>6</sup>, Irene Nooren<sup>3</sup>, Jan-Willem Boiten<sup>7</sup>

<sup>1</sup> University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

<sup>2</sup> Vancis, Amsterdam

<sup>3</sup> SURFsara, Amsterdam, The Netherlands

<sup>4</sup> VU university medical center, Amsterdam, The Netherlands

<sup>5</sup> Netherlands eScience Center, Amsterdam, The Netherlands

<sup>6</sup> VU university, Amsterdam, The Netherlands

<sup>7</sup> Center for Translational Molecular Medicine, Eindhoven, The Netherlands

### Introduction

For the national translational IT project CTMM/TraIT Galaxy has been selected as one of the tools in the experimental domain. The TraIT partners (among others NBIC and SURFsara) have developed a vision how to make Galaxy available to the research community in The Netherlands. The scalable Galaxy cluster on the SURFsara HPC Cloud will be transferred to Vancis to provide a sustainable production-level Galaxy cluster. In the design of this environment Vancis has made use of the knowledge and experience of NBIC and SURFsara hosting the public NBIC instance on the SURFsara HPC Cloud.

### Material & Methods

To assess the minimal requirements for the infrastructure we used metrics collected while running the NBIC Galaxy on the HPC Cloud. Next we drafted a set of use cases the infrastructure should be able to fulfil, such as the ability to run Omics-pipelines and the ability to scale to handle peak demand. We identified I/O performance as a major bottleneck, since many Galaxy tools are I/O intensive, while Galaxy has a shared data design. Memory was also recognized as a critical factor, since typical datasets are in the order of the tens of gigabytes. We also built upon the experiences from SURFsara in operating the HPC Cloud and other HPC. To accommodate for a full set of development, testing, acceptance & production environments, as well as private installations, the infrastructure should support multiple Galaxy clusters. The chosen architecture will use a Linux High Availability environment with OpenStack, which will run on two large-size blades. Storage is split into multiple tiers with different characteristics to support both high I/O workloads and a reliable large storage. The chosen setup is horizontally scalable in a cost-efficient manner.

### Results

From May to September 2014 we will pilot the new architecture within the TraIT project. For this pilot we have selected a few TraIT NGS tools and pipelines to stress test the system under different workload scenarios. Furthermore we have established a process to ensure the quality of the tools required for a stable production environment.

## P3: Practical experiences from the Munich NGS-FabLab - Tools, compatibility and pitfalls off the standard tracks

Aarif Mohamed Nazeer Batcha<sup>1</sup>, Sebastian Schaaf<sup>1,2</sup>, Guokun Zhang<sup>1</sup>, Sandra Fischer<sup>1</sup>, Ashok Varadharajan<sup>1</sup>, Ulrich Mansmann<sup>1,2</sup>

<sup>1</sup> Department of Medical Informatics, Biometry and Epidemiology (IBE), Ludwig Maximilians University (LMU) Munich, Germany

<sup>2</sup> German Cancer Consortium (DKTK), Heidelberg, Germany

Over three years, the Munich NGS-FabLab was built up first as a concept and later as a running IT system, based on an assessment of requirements, constraints and given structural conditions. Since some months it is in active use, although still under intense development.

As every developer knows, especially complex and broad open source software like Galaxy does not come error-free. Expected issues were due to non-standard elements like the operating system (SLES 11), hardware (x86 server not supported by the clinics IT, FPGA hybrid-

computer, network load, ...), computational requests (projects with special needs or proprietary software) and not to forget financing and politics. Apart from that, Galaxy itself and the associated software packages and/or the respecting wrappers surprisingly often turned out to be in need of corrections, although we assumed to use standard input data and perform simple jobs. Finally, those tools or computations which were needed, but are not yet supported by the Galaxy framework, most work invested deals with trouble-shooting, bug-hunting and code analysis.

Experiences, fixes, improvements and new integrations are subject to this poster, which may appear more like a collage of loosely connected sub-topics. While we did not return those code snippets to the community yet, we also hope to get into the process of submitting contents for public use and discuss them, in order to improve the framework as a whole.

## P4: e-Science in France, a Life science Western story

Yvan LE BRAS<sup>1</sup>, Aurélien ROULT<sup>1</sup>, Cyril MONJEAUD<sup>1</sup>, Mathieu BAHIN<sup>2</sup>, Olivier QUENEZ<sup>3,4</sup>, Claudia HERIVEAU<sup>1</sup>, Olivier SALLOU<sup>1</sup>, Anthony BRETAUDEAU<sup>1,5</sup> and Olivier COLLIN<sup>1</sup>

<sup>1</sup> GenOuest Core Facility, UMR6074 IRISA CNRS/INRIA/Université de Rennes I, Campus de Beaulieu, 35042, Rennes Cedex, France

<sup>2</sup> IGDR, UMR 6290-CNRS Université de Rennes I, 2 avenue Professeur Léon Bernard, Campus de Villejean, 35065, Rennes Cedex, France

<sup>3</sup> Inserm U1079, Institut de Recherche et d'Innovation Biomédicale (IRIB), Université de Rouen, France

<sup>4</sup> Centre National de Référence pour les Malades Alzheimer Jeune, CHU de Rouen, Lille et Paris-Salpêtrière, Rouen, France

<sup>5</sup> INRA IGEPP, UMR1349 Agrocampus-Ouest INRA Université Rennes I, domaine de la motte, 35653, Le Rheu, Cedex 35327, France

Research processes are evolving at a rapid pace. This evolution, mainly due to technological advances, offers powerful equipment and generalizes the digital aspect of the research data. If facing the actual data deluge context represents a challenge, it also offers an opportunity to change and enhance our manner to tackle research tasks and disseminate science. In Life Sciences, as in other domains, we are noting a sharp increase in storage and computing needs. Regularly adding hardware resources to the bioinformatics core facilities is no longer sustainable. Scientific data management and analysis have to be enhanced in order to offer services and developments matching the new uses.

Since 2 years, Galaxy platform is used in combination with ISATools and HUBzero to build a Life Sciences Virtual Research Environment. Each tool offers complementary functionalities: ISATools software suite for metadata management, HUBzero for scientific collaboration and Galaxy for computation. The resulting combination allows scientists to manage their project from collaboration to data management and analysis. This Virtual Research Environment (VRE) is tested in partnership with the scientific communities in Western France. The evaluation will give us insights on the usage and acceptance of new tools in a scientific field characterized by profound modification of its traditional processes.

Although the deployment of this kind of environment is challenging, it represents an opportunity to pave the way towards better research processes through enhanced collaboration, data management, analysis practices and resources optimization.

## P5: [drylab.nl.enabling.translational.research](http://drylab.nl.enabling.translational.research)

Christian Rausch<sup>1</sup>, Daoud Sie<sup>1</sup>, Jeroen Galle<sup>2</sup>, Jeroen Crape<sup>2</sup>, Gerben Menschaert<sup>2</sup>, Bauke Ylstra<sup>1</sup>, Wim Van Criel<sup>1,2</sup>

<sup>1</sup> Cancer Center Amsterdam, VU University Medical Center, Amsterdam, The Netherlands

<sup>2</sup> Biobix, Lab of Bioinformatics and Computational Genomics, Ghent University, Ghent, Belgium

The Cancer Center Amsterdam (CCA) of VU University Medical Center is a research center that performs internationally recognized research in the area of oncogenetics, immunopathogenesis, disease profiling, innovative therapy and quality of life.

We are currently establishing a Drylab that empowers both researchers and clinicians with state-of-the-art bioinformatics solutions. The Drylab is expected to contribute scientifically, which we want to make possible by building a team with diverse interdisciplinary backgrounds: Biology, statistics, experimental design, bioinformatics etc.

Establishing an organizational context with continued funding is an ongoing challenging task. First, we have built a scalable infrastructure. We established Drylab.nl as a custom Wordpress instance, expanded with a helpdesk and ticketing system and linked to a Galaxy based workflow system using a tool shed to (re)use and share internal and external workflows. In external collaborations (e.g. with Biobix in Ghent, Belgium) we are building/exchanging pipelines/workflows for RNAseq, proteogenomics (riboSeq) and methylome analysis (methylcapSeq). We are also implementing a workflow validation procedure using test data. In order to close the loop to the end user we are planning to visualize genomic data on different platforms.

Our initial measure for success will be the actual consolidation and integration of bioinformatics efforts in addition to (re)use of these workflows by non-experts. We do recognize that in order to mature we have to avoid getting caught in a "firefighting mode". Given the shared vision amongst all stakeholders and the embedded organizational context we hope to mature and become an innovation engine within translational medicine.

#### P6: Mississippi: a galaxy server centered on small RNA analysis

Marius van den Beek<sup>1</sup>, Christophe Antoniewski<sup>1</sup>

<sup>1</sup> Drosophile.org, CNRS and University Pierre-et-Marie-Curie, Paris

Non-coding small RNAs (miRNA, siRNA, piRNA, ...) are involved in the regulation of genes and transposable elements as well as in the defense against viral infections. Their discovery and their functional characterization rely heavily on high throughput RNA sequencing. The ~20:30nt length of small RNAs raises specific challenges for meaningful read mapping and analysis, so that standard RNAseq analysis methods have to be adapted. We provide an integrated set of galaxy tools that should streamline the most frequent small RNA analysis needs. This includes a modified bowtie-wrapper and workflows that allow users to quickly and reproducibly interrogate various aspects of small RNA biology. We provide tools for the discovery and differential expression analysis of miRNAs and a way for genome-wide visualization of miRNA precursors that complements Trackster. Furthermore we provide tools to detect the "ping-pong" biogenesis signature of piRNAs, to detect piRNA-producing loci in the genome and to study and visualize the impact of piRNAs and siRNAs on transposable elements.

#### P7: Bacterial and viral NGS analysis in a public health agency using Galaxy

Ulf Schaefer<sup>1</sup>, Anthony Underwood<sup>1</sup>, and Jonathan Green<sup>1</sup>

<sup>1</sup> Advanced Laboratory and Bio-Informatics, Microbiology Services, Public Health England, 61 Colindale Avenue, London NW9 5EQ, United Kingdom

Public Health England is home to the United Kingdom's national microbiology reference laboratories and deals with the surveillance and control of infectious disease. Assays for the investigation of selected pathogenic bacteria and viruses are being migrated from traditional wet lab based methodologies such as Multiple Loci VNTR Analysis to methods based on Next Generation Sequencing (NGS) data. Apart

from the set up of an NGS service and automated analysis of a small number of priority organisms, one of the key challenges in the management of this paradigm shift in public health is to enable microbiologists and epidemiologists with little to no bioinformatics knowledge and training to interact with and derive scientific value from NGS data. We maintain a local installation of Galaxy in an attempt to address this challenge. This local installation houses all specialised software required for public health microbiology and phylogenetics. Furthermore it provides bespoke workflows for standard analyses regularly employed in outbreak investigations, such the creation a SNP tree from multiple viral or bacterial NGS samples. In addition to an overview of our hardware and software setup, this presentation will highlight 1) An example of a public health specific workflow that can be used for routine reference microbiology services and 2) some of the soft issues around employing Galaxy in this context, such as user acceptance, training, and support.

#### P8: iReport: HTML Reporting in Galaxy

Saskia Hiltmann<sup>1</sup>, Yuri Hoogstrate<sup>1</sup>, Hailiang Mei<sup>2</sup>, Guido Jenster<sup>1</sup>, Andrew Stubbs<sup>1</sup>

<sup>1</sup> ErasmusMC, Rotterdam, The Netherlands

<sup>2</sup> LUMC, Leiden, The Netherlands

Galaxy offers a number of great visualisation tools (Trackster, Circster), but currently lacks the ability to easily summarise the various outputs of a workflow into a single view. iReport is a Galaxy tool for the easy creation of HTML reports from Galaxy outputs. Rather than having a static HTML output, iFUSE2 uses javascript and jQuery to allow for interactivity in the form of searching and sorting of tables, automatic zooming of image data, tabbed view for organisation of outputs, etc. Users define the number and names of tabs for their report, and can add different types of content-items to these tabs (e.g. text, tabular data, image data, PDF files, links to datasets, and more).

We have previously implemented Galaxy-based data processing pipelines for next-generation sequencing (NGS) and for array based allelic copy number determination named CGtag (Hiltmann et al. 2014) and developed a web based fusion gene visualizer, iFUSE (Hiltmann 2013). We used the iReport tool to make iFUSE2, the next-step extension to support fusion gene determination within Galaxy, which runs as the last step of our workflow and combines the outputs of various Galaxy tools into a single view.

iReport is available from the DTL toolshed (toolshed.dtls.nl) and the main Galaxy toolshed.

#### P9: workflow4metabolomics.org : Galaxy and the metabolomics analysis Universe

Mishar<sup>1</sup> MONSOOR<sup>1</sup>, Gildas LE CORGUILLE<sup>1</sup>, Marion LANDI<sup>2</sup>, Mélanie PETERA<sup>2</sup>, Pierre PERICARD<sup>1</sup>, Christophe DUPERIER<sup>2</sup>, Marie TREMBLAY-FRANCO<sup>3</sup>, Jean-François MARTIN<sup>3</sup>, Sophie GOULITQUER<sup>1</sup>, Etienne THEVENOT<sup>4</sup>, Franck GIACOMONI<sup>2</sup>, Christophe CARON<sup>1</sup>

<sup>1</sup> ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France

<sup>2</sup> PFEM, UMR1019 INRA, Centre Clermont-Ferrand-Teix, 63122, Saint Genes Champanelle, France

<sup>3</sup> PF MetaToul-AXIOM, UMR 1331 Toxalim INRA, 180 chemin de Tournefeuille, F-31027, Toulouse, France

<sup>4</sup> DRT/LIST/DM2I/LADIS, Saclay Center CEA, F-91191, Gif-sur-Yvette, France

Facing the emergence of new technologies in the field of metabolomics, treatment solutions adopted so far (XCMS, R scripts, etc.) clearly show their limits. Bottlenecks affect unified access to core applications as well as computing infrastructure and storage. In the context of collaboration between metabolomics (MetaboHUB French infrastructure) and bioinformatics platforms (IFB: Institut Français de Bioinformatique), we have developed a full pipeline using Galaxy

framework for data analysis including preprocessing, normalization, quality control, statistical analysis and annotation steps. This modular and extensible workflow is composed with existing components (XCMS and CAMERA functions, etc.) but also a whole suite of complementary statistical tools. This implementation is accessible through a web interface, which guarantees the parameters completeness. The advanced features of Galaxy have made possible the integration of components from different sources and of different types. Finally, an extensible environment is offered to metabolomics communities (platforms, end users, etc.), and enables preconfigured workflows sharing for new users, but also experts in the field.

### PI0: The Munich NGS-FabLab - A glimpse on an IT infrastructure for medical sequence data

Sebastian Schaaf<sup>1,2</sup>, Aarif Mohamed Nazeer Batcha<sup>2</sup>, Guokun Zhang<sup>2</sup>, Sandra Fischer<sup>2</sup>, Ashok Varadharajan<sup>2</sup>, Ulrich Mansmann<sup>1,2</sup>

<sup>1</sup> German Cancer Consortium (DKTK), Heidelberg, Germany

<sup>2</sup> Department of Medical Informatics, Biometry and Epidemiology (IBE), Ludwig Maximilians University (LMU) Munich, Germany

While NGS data becomes increasingly important in medical basic research and molecular diagnostics, dealing with it is a challenge in multiple aspects. Apart from 'classical' issues like high demands to hardware, the interconnectivity to resources of biomedical meta information for enriching sequence data is a central task. Users from various fields of study have to be enabled to work with a variety of bioinformatic tools off the command line (which currently do not offer any gold standard analyses), concentrating on contents instead of technical elements. On top, patient-related data is subject to strong restrictions by the German data security law, which also affects IT infrastructures on all levels. For medical genome informatics in Munich, the NGS-FabLab (including its admin round-table "NGS-ART") is the central hub for clinicians, researchers and developers, serving as data center, knowledge core, teaching unit and technical template for further instances. During development, the standard Galaxy distribution setup has been equipped with some features that we would like to present with this poster.

Apart from the operating system layer (VMWare, SLES 11), key features are fully automated scripts for proper development cycles and quick setups, distributed computing resources (SGE queue, Convey FPGA hybrid-core computer), highly integrated network structures and access controls. Furthermore, scientific broadness has been enhanced (e.g. via qiime toolbox, pathway analyses, additional and improved tools). Last but not least, archiving and sophisticated analysis are subject to improvements by using Bii as searchable and Galaxy-interconnected database, relying on biomedical ontologies.

### PI 1: Oqtans: Online quantitative transcriptome analysis

Vipin T. Sreedharan<sup>1</sup>, Yi Zhong and Gunnar Rätsch

<sup>1</sup> Memorial Sloan Kettering Cancer Center, New York City, NY-10065 USA

Powerful algorithmic techniques lead to software applications that can answer important biomedical questions that analyze massive and complex genomic data sets. Starting from 2009, oqtans has served the biological research community with state-of-the-art machine learning tools for sequence analysis and high-throughput experimental technologies like RNA sequencing.

We have been leveraging the oqtans codebase to withstand different RNA-seq downstream analysis directions. In particular, it has been utilized recently for translational research to understand the effect of anticancer therapeutics. To measure the translational efficiency change for protein coding genes from multiple samples (treated vs nontreated), we used the sequencing based transcriptome scale ribosome footprinting and RNA-seq data. Our approach allowed us to detect significant changes of the ribosome binding profile of mRNA

transcripts between two conditions using a non-parametric testing strategy.

Moving the Galaxy framework from academic to clinical research introduces a myriad of informatics challenges concerning the security of the data sets. In addition to developing new methods for oqtans components, it is equally important to handle the informatics complexities that come with scaling oqtans for clinical use. We have deployed our instance under ModSecurity and encrypted user authentication and subsequent session transmissions using Secure Sockets Layer (SSL). We have applied patches to the core codebase of the Galaxy framework to responsively address vulnerable redirection via URL injection, Reflected and stored Cross-site scripting (XSS) and properly sanitize and encode all potential user input and output.

Availability

- oqtans cloudman image - ami-65376a0c
- oqtans public compute server - galaxy.cbio.mskcc.org

### PI2: Locally managed Galaxy instances with access to external resources in a supercomputing environment

Nuria Lozano<sup>1,2</sup>, Oscar Lozano<sup>2</sup>, Beatriz Jorriñ<sup>1</sup>, Juan Imperial<sup>3</sup>, Vicente Martín<sup>2</sup>

<sup>1</sup> Center for Biotechnology and Genomics of Plants (CBGP), Technical University of Madrid, Spain

<sup>2</sup> Madrid Supercomputing and Visualization Center (CeSViMa), Technical University of Madrid, Spain

<sup>3</sup> Center for Biotechnology and Genomics of Plants (CBGP), Technical University of Madrid and CSIC, Spain

For a research lab, accessing shared resources like those available in supercomputer centers is a welcome addition to Galaxy capabilities. However, privacy or flexibility requirements might impose the need for a locally managed Galaxy installation. In these cases a way to communicate a local instance of Galaxy with the supercomputer would be a solution.

The Center for Biotechnology and Genomics of Plants (CBGP) and the Madrid Supercomputing and Visualization Center (CeSViMa) are located at Technical University of Madrid (UPM) Montegancedo Campus. CeSViMa manages the large heterogeneous Magerit cluster, with about 4,000 Power7 and 1,000 Intel cores, accessed in batch mode. The resource manager used is SLURM and scheduler is MOAB. Standard job runs in Magerit involve logging into one of the interactive nodes, preparing a job command file and then submitting them to one of the batch queues. The challenge was to be able to seamlessly use this system through a Galaxy front-end. The solution adopted was to set up a Virtual Private Server that runs Galaxy. The Galaxy instance has been installed in a filesystem shared between VPS and Magerit, which is under the control of Magerit GPFS filesystem.

Galaxy jobs are sent to Magerit through Command Line Interface. A Job Plug-In has been coded that creates the needed Jobfiles transparently submitted to the queuing system.

Using this approach, research group members are fully responsible for deploying and maintaining their own Galaxy Local Instance, while heavy work is offloaded to external computing resources.

### PI3: Argument Parsing Libraries for Automatic Galaxy XML Generation

Eric Rasche<sup>1</sup> and Dr. Ryland F. Young<sup>1</sup>

<sup>1</sup> Center for Phage Technology, Texas A&M University, College Station, TX

Addition of new software to Galaxy is currently a non-trivial task. Galaxy tools consist of many interdependent parts; packaged executables or scripts, tool data, and tool configuration in the form of XML files. This presents a problem in the form of a large codebase to

maintain, especially for groups that regularly produce tools to add to Galaxy.

With the goals of code deduplication, simplification of deployment workflow, and improved accessibility of the Galaxy platform for new developers, we have developed Python and Perl libraries that function to replace traditional methods of obtaining command line arguments like GetOpt and argparse. Our libraries are capable of automatically generating valid Galaxy XML tool description files that represent the full set of a tool's command line options. This removes the need to maintain the Python/Perl script and the XML file separately, as the XML files can be regenerated at any time from the Python/Perl script. We believe this will lead to significant reductions in time spent on maintenance of codebases and decreases turn around times for shipping new releases. These libraries will benefit anyone adding new custom tools to Galaxy by providing a convenient method to specify command line parameters, an easy way to access that data in their tools, and automatic Galaxy integration.

#### PI4: Advantages and Challenges of Using the Galaxy API within an Integrated Data Analysis and Visualization Platform

Ilya Sytchev<sup>1</sup>, Nils Gehlenborg<sup>2</sup>, Shannan Ho Sui<sup>1</sup>, Richard Park<sup>2,3</sup>, Psalm Haseley<sup>2</sup>, Winston Hide<sup>3</sup>, Peter Park<sup>1</sup>

<sup>1</sup> Center for Stem Cell Bioinformatics, Harvard Stem Cell Institute

<sup>2</sup> Center for Biomedical Informatics of Harvard Medical School

<sup>3</sup> Boston University Bioinformatics Program

The Refinery Platform (<http://refinery-platform.org>) is an integrated web-based data visualization and analysis system powered by an ISA-Tab-compatible data repository. Analyses are implemented as Galaxy workflows. As a result, Refinery makes extensive use of the Galaxy API to automate analyses, including such features as uploading datasets into Galaxy libraries, importing "workflow templates", exporting workflows back into Galaxy after initialization with user-selected inputs, running workflows, and downloading workflow results from Galaxy histories back into Refinery. Some of these features were implemented through custom extensions to the Galaxy API. We directly benefit by using key Galaxy features such as cluster deployment, progress monitoring, a large selection of tools, and the workflow editor.

The recent development of the BioBlend library (<http://bioblend.readthedocs.org>) motivated us to replace our existing custom Galaxy API client code with BioBlend library components. BioBlend encapsulates the underlying REST API of Galaxy in a way that is more suitable for programming and makes it easier to automate end-to-end large-data analyses. It has a more robust implementation and is maintained by the community to keep up-to-date with the changes in the Galaxy API. Extensions to the BioBlend library and the Galaxy API to enable the use of Galaxy in fully automated fashion will be contributed back to this community effort. We hope to use this opportunity to gain feedback and suggestions for improvements from the Galaxy developer community.

#### PI5: Resistance to Toxic Compounds in Metagenomic Fosmid Library from Mangrove Sediment in São Paulo State, Brazil

Lucélia Cabral<sup>1</sup>, Sanderson Tarciso Pereira de Sousa<sup>1</sup>, Gileno Vieira Lacerda Júnior<sup>1</sup>, Júlia Ronzella Ottoni<sup>1</sup>, Daniela Ferreira Domingos<sup>1</sup>, Valéria Maia de Oliveira<sup>1</sup>

<sup>1</sup> Divisão de Recursos Microbianos, Research Center for Chemistry, Biology and Agriculture (CPQBA), Campinas University - UNICAMP. Mailbox: 6171. CEP: 13081-970. Campinas. São Paulo. Brazil

The mangrove is a typically tropical ecosystem, located between land and sea, and very rich in biodiversity, including aquatic animals, birds, reptiles, mammals and microorganisms. Despite of this, mangroves

have been highly exposed to anthropic activities, including oil spills and industrial waste disposals that carry heavy metals. Microorganisms found in the environment can adapt to the presence of pollutants, thus developing survival mechanisms. However, traditional cultivation methods are not efficient for cultivation of most microorganisms present in nature. In this context, the aim of this study was to assess the presence of heavy metal resistance in a fosmid library constructed using metagenomic DNA from sediment samples collected from a mangrove area located in Bertioga, State of São Paulo, Brazil. The fosmid library comprised 13,000 clones and the sampling site was affected by oil spill. Next generation sequencing was performed using the 454 sequencing platform. Sequences associated with toxic compounds resistance were analyzed using MG-RAST V3.3.8. The annotations used were: Functional abundance, Hierarchical classification, level 1 (Virulence, Disease and Defense), level 2 (Resistance to antibiotics and toxic compounds), Level 3 (Resistance). The most abundant sequences involved in metal resistance in the dataset were cobalt-zinc-cadmium resistance detected by the presence of Cobalt-zinc-cadmium resistance protein and Cobalt-zinc-cadmium resistance protein CzcA (489 and 346 hits, respectively). Sequences related with copper and silver resistance were detected by the presence of cation efflux system protein CusA (330 hits). The functional screening of fosmid library will be performed and the positive clones will be selected for further studies on metal tolerance and degradation.

#### PI6: BlockClust: efficient clustering and classification of non-coding RNAs from short read profiles

Pavankumar Videm<sup>1</sup>, Dominic Rose<sup>1,5</sup>, Fabrizio Costa<sup>1</sup>, Rolf Backofen<sup>1-4</sup>

Presented by Björn Grüning<sup>1</sup>

<sup>1</sup> Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany

<sup>2</sup> Centre for Biological Signalling Studies (BLOSS), University of Freiburg, Germany

<sup>3</sup> Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Germany

<sup>4</sup> Centre for Non-coding RNA in Technology and Health, Bagsvaerd, Denmark

<sup>5</sup> Munich Leukemia Laboratory (MLL), Munich, Germany

Non-coding RNAs play a vital role in many cellular processes such as RNA splicing, translation, gene regulation. However the vast majority of ncRNAs still have no functional annotation. One prominent approach for putative function assignment is clustering of transcripts according to sequence and secondary structure. However sequence information is changed by post-transcriptional modifications, and secondary structure is only a proxy for the true three dimensional conformation of the RNA polymer. A different type of information that does not suffer from these issues and that can be used for the detection of RNA classes, is the pattern of processing and its traces in small RNA-seq reads data.

Here we introduce BlockClust, an efficient approach to detect transcripts with similar processing patterns. We propose a novel way to encode expression profiles in compact discrete structures, which can then be processed using fast graph kernel techniques. We perform both unsupervised clustering and develop family specific discriminative models; finally we show how the proposed approach is both scalable, accurate and robust across different organisms, tissues and cell lines.

BlockClust was tested and works with small RNA-seq data of eukaryotic organisms. It is the first tool of its kind, which is easily installable and usable on galaxy framework. To run BlockClust all you need is an alignment file of short reads in Sequence Alignment/Map (SAM/BAM) format. A complete workflow of BlockClust and its tool dependencies are now available at Galaxy ToolShed.



## P17: A Galaxy-Based framework for online streaming data analytics in Heart Rate Variability Analysis

Calogero Zarbo<sup>1</sup>, Andrea Bizzego<sup>1,2,3</sup>, Marco Mina<sup>1</sup>, Gianluca Esposito<sup>2,4</sup>, Cesare Furlanello<sup>1</sup>

<sup>1</sup> FBK - Fondazione Bruno Kessler, Trento, Italy

<sup>2</sup> University of Trento, Italy

<sup>3</sup> SKIL Telecom Italia, Trento, Italy

<sup>4</sup> RIKEN BSI, Wako-Shi, Japan

The emerging applications in physiological data processing, encouraged by the availability of wearable sensors for continuous self-monitoring and quantified self, require new platforms for time series analysis supporting real-time processing and fast prototyping capabilities. We recently proposed Physiolyze, a Galaxy-based web framework to support complex workflows for Heart Rate Variability (HRV) analysis. Here we extend Physiolyze by introducing scalable online processing capabilities.

The enhanced version still relies on Galaxy as core platform to design and manage the pipelines. In order to incrementally analyze the streams, a set of Python routines based on the Bioblend library works as middleware to trigger the pipelines as new data become available. A web interface based on the Django Python framework allows the user to control the execution of the pipelines, running them on new data streams. We tested our system on the task of predicting infant behavioral state from HRV patterns. We simulated a real-time scenario of 100 asynchronous data streams from data for 24 infants previously collected with a Light WP Holter ECG recorder (GE Healthcare). The system incrementally extracts 37 HRV indicators from each data stream and predicts the infant state (e.g. wake, sleep, cry) with a Random Forest regression model. The pipeline is modular and fully managed as a Galaxy workflow.

Our system can easily be adapted to other online streaming analytics applications, such as for the parallelized analysis of multiple data streams acquired from physiological sensors and wearable devices.

## P18: Implementing qDNAseq in Galaxy: a whole genome sequencing copy number analysis tool

Stef van Lieshout<sup>1</sup>, Ilari Scheinin<sup>1</sup>, Daoud Sie<sup>1</sup>, Remond J.A. Fijneman<sup>1</sup>, Bauke Ylstra<sup>1</sup>

<sup>1</sup> Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands

DNA copy number aberrations are a hallmark of cancer and can be quantified by shallow whole-genome sequencing (WGS). A robust method has been developed<sup>1</sup> that detects copy number aberrations by binning and counting sequence reads in non-overlapping windows (usually of 15kb). Then a combined LOESS correction for mappability and GC content is applied followed by excluding genomic regions from both ENCODE project blacklists and a novel blacklist based on sequence depth of 38 individuals from the 1000 Genomes project.

The procedure is available as a Bioconductor package, QDNAseq<sup>2</sup>. The accompanying Galaxy tool uses the popular BAM format as input and reports results in a clear and concise HTML based view within Galaxy itself. Various output formats can be downloaded, including an R data structure file for downstream analysis and a Zipped archive with all the output together.

Due to precalculated bin annotations, current limitations include the support for one genome build (GRCh37/hg19) and one sequencing type (50bp single read). Additional dedicated tools will handle these challenges and future plans include the addition of different strategies for segmenting and calling the copy number data.

Funding was supported by the Center for Translational Molecular Medicine, Translational research IT project (CTMM TraIT).

<sup>1</sup> Ilari Scheinin, Daoud Sie et al. DNA copy number analysis of fresh and formalin-fixed specimens by whole-genome sequencing: Improved correction of systematic biases and exclusion of problematic regions, (submitted).

<sup>2</sup> <http://www.bioconductor.org/packages/release/bioc/html/QDNAseq.html>

## P19: Integrating Integrated Genome Browser with Galaxy

Ann Loraine<sup>1</sup>, David Norris<sup>1</sup>, Kyle Suttlemyre<sup>1</sup>, Tarun Kanaparthi<sup>1</sup>

<sup>1</sup> University of North Carolina - Charlotte

Integrated Genome Browser is a fast, flexible and free Java-based desktop software tool that enables interactive exploration of genomic data sets. To accommodate large data sets, IGB featured a simple ReST-style interface that triggers incremental loading of data from local files or URLs. We used this ReST-style interface and the Galaxy viewers API to enable IGB visualization for Galaxy users. When Galaxy users create compatible data files, they now see a link labeled "View in IGB" upon clicking data file links in their Galaxy History. Clicking this link triggers delivery of data to IGB for display. This is a simple interaction from the user's perspective, but from an engineering point of view, it highlights a key extension point for Galaxy that enables integration with IGB or other visualization tools. By enabling access to data sets in a user-friendly, web-based interface, Galaxy offers many possibilities to enhance user interactions for data analysis and data sharing.

## P20: An approach for detecting structural variations from NGS paired end reads using Split Reads, Discordant Read Pairs and Local Alignment

Michael Ta<sup>1</sup>, Philip D. Cotter<sup>1</sup>, Mathew W. Moore<sup>1</sup>

<sup>1</sup> Bioinformatics Department, ResearchDx, Irvine CA, USA

A major challenge in Next Generation Sequencing is the development of efficient algorithms to detect structural variants present in the genome. Several different approaches for the detection of structural variants have been identified. Breakdancer searches for clusters of anomalous read pairs for sites to investigate. Similarly, another analysis tool, SoftSearch, uses the soft clipped read data from the aligner to determine sites of interest and heuristically report potential structural variants around them. Our algorithm, HardSearch, expands on the approach of SoftSearch to further identify the exact break points that support chromosomal structural variations. Paired end reads from DNA-seq with an unmapped mate are collected around each potential fusion site; the unmapped mates are realigned to the reference genome using a local aligner. The segment of each read that aligns with the highest alignment score without gaps is subtracted from the original and the remainder is realigned allowing for the identification of the breakpoint and breakpoint partners.

## P21: Synapse: Software infrastructure for collaborative reproducible research

J Christopher Bare<sup>1</sup>, Synapse Platform Team<sup>1</sup>, Michael R Kellen<sup>1</sup>, Stephen H Friend<sup>1</sup>

<sup>1</sup> Sage Bionetworks

Synapse (<http://www.synapse.org>) is a free and open source informatics platform for data-driven collaborative research. Built from the ground up for a rich data sharing experience, Synapse provides tools for versioning, annotating and citing data combined with provenance tracking and fine grained access control. Synapse operates under a complete governance process developed, approved and monitored by an independent ethics advisory board and the Western Institutional Review Board.

Synapse is designed to support Sage Bionetworks' mission to promote a scientific culture founded on broad and open collaboration. Sage

Bionetworks develops and operates Synapse as a public resource for the scientific community. For example, the Cancer Genome Atlas pan-cancer group published a total of 18 papers in Nature Publishing Group journals (<http://www.nature.com/tcga/>), using Synapse as a single point for sharing data, results and methods among 250 collaborators spread across 30 institutions. In partnership with DREAM, Synapse hosts predictive modeling challenges on a diverse array of topics including disease prognosis, drug response, toxicology and genetic variant analysis.

Galaxy and Synapse share many goals including transparency and reproducibility. Both enable sharing and reuse of research code and are cloud-native applications with similar models of computation including provenance, workflows, data sets and pages.

Bridging these two complementary services would benefit users of both. Synapse could act as a data source for Galaxy workflows and/or as a shared workspace for results and intermediate products. Other options to explore include exchanging workflows, provenance, and narrative pages. Integration between Synapse and Galaxy could enrich the ways in which data and analysis code can be presented, shared and reused.

## P22: Integration of Galaxy with IRIDA, a Genomic Epidemiology Platform

Aaron Petkau<sup>1</sup>, Franklin Bristow<sup>1</sup>, Thomas Matthews<sup>1</sup>, Josh Adam<sup>1</sup>, Damion Dooley<sup>2</sup>, Emma Griffiths<sup>4</sup>, Geoff Winsor<sup>4</sup>, Matthew Laird<sup>4</sup>, Melanie Courtot<sup>2,4</sup>, William Hsiao<sup>2,3</sup>, Gary Van Domselaar<sup>1</sup>, Fiona Brinkman<sup>4</sup>

<sup>1</sup> National Microbiology Laboratory, Public Health Agency of Canada, Canada

<sup>2</sup> BC Public Health Microbiology and Reference Laboratory, Canada

<sup>3</sup> University of British Columbia, Canada

<sup>4</sup> Simon Fraser University, Canada

The continuing decrease in the cost of genomic sequencing and the development of new data analysis methods has led to the increasing usage of whole genome sequencing as an epidemiological tool. Whole genome sequencing can provide a high-resolution snapshot of the relationship among pathogens and lead to a greater ability to identify and track infectious disease outbreaks. Initiatives, such as the Global Microbial Identifier, have already started the discussion on developing a system and standards for genomic epidemiology. In our project, IRIDA (Integrated Rapid Infectious Disease Analysis), we propose a platform for genomic epidemiology which provides a secure storage of whole genome sequence data, epidemiological metadata, data analysis pipelines, visualization of results, a RESTful API, and a federated data sharing model. Galaxy has already proven to be a useful application for integration of common bioinformatics tools and data, execution of data analysis pipelines, collection of results, and data sharing. In addition, Galaxy provides a RESTful API for programmatic access to running instances of Galaxy. We intend to leverage Galaxy as much as possible by interacting with locally installed Galaxy instances via the API to execute pre-defined data analysis pipelines, store data results and Galaxy histories, and manage installed bioinformatics tools. Direct export of whole genome sequencing data to instances of Galaxy will be provided for more complicated analysis. IRIDA will be released as free and open-source software and make use of common data standards to facilitate sharing with other genomic epidemiology platforms. More information will be made available at <http://irida.ca>.

## P23: Galaxy on the GenomeCloud : Yet another on-demand Galaxy cloud, but only powered by Apache CloudStack

Youngki Kim<sup>1</sup>, CB Hong<sup>1</sup>, Kjoong Kim<sup>1</sup>, Daechul Choi<sup>1</sup>

<sup>1</sup> GenomeCloud, Seoul, Korea

Bioinformatics and genome data analysis in South Korea is at its early stage but getting busier. To keep pace with this trend of

research, GenomeCloud was created at the end of 2012.

GenomeCloud is an integrated platform for analysing, interpreting and storing genome data, based on KT's cloud computing infrastructure which uses Apache CloudStack software. GenomeCloud consists of g-Analysis (automated genome analysis pipelines at your fingertips), g-Cluster (easy-of-use and cost-effective genome research infrastructure) and g-Storage (a simple way to store and share genome-specific data).

Because of flexible tool integration architecture and seamless workflow creation functionality, Galaxy was selected to achieve multi purpose goals such as agile pipeline development and bioinformatics education support. To provide on-demand and Apache CloudStack based Galaxy cluster, we have automated virtual machine creation, clustering and various software setup including Galaxy.

Furthermore, seamless integration with GenomeCloud helps researchers not only create and manage Galaxy through a convenient web interface but also fully utilizes genome data in g-Storage. g-Storage is powered by OpenStack Swift and specially designed genome file transfer protocol.

Galaxy on the GenomeCloud uses Grid Engine as a Cloud HPC Solutions, Ganglia as a distributed monitoring system and LVM over NFS as a large volume shared storage, all of which are setup automatically upon request. This talk will be about our experiences while integrating Galaxy with GenomeCloud and use cases of Galaxy such as scalable bioinformatics education system and request fulfillment of RNA-seq analysis.

## P24: GenomeSpace: An Environment for Frictionless Bioinformatics

Michael Reich<sup>1</sup>, John Liefeld<sup>1</sup>, Marco Ocana<sup>1</sup>, Donkeung Jang<sup>1</sup>, James Robinson<sup>1</sup>, Peter Carr<sup>1</sup>, Barbara Hill<sup>1</sup>, Thorin Tabor<sup>1</sup>, Helga Thorvaldsdottir<sup>1</sup>, Aviv Regev<sup>1</sup>, Jill P. Mesirov<sup>1</sup>

<sup>1</sup> Broad Institute, Cambridge, MA

Over the past several years, initiatives such as The Cancer Genome Atlas and 1000 Genomes Project have produced an explosion of genomic data. These efforts offer a new era of potential for the understanding of basic mechanisms of disease and identification of novel treatments. Comprehensive analysis of these datasets requires coordinated use of Web-based applications, data repositories, and desktop analysis tools. However, the effort required to transfer data between tools, convert between formats, and manage results often prevents researchers from utilizing the wealth of methods available. Many "bench to bedside" discoveries are possible with combinations of existing tools, but the necessary transitions between them puts them out of the reach of most researchers.

GenomeSpace is an environment that brings together diverse computational tools, enabling non-programmer scientists to easily combine their capabilities. It provides a space to create, manipulate and share a growing collection of genomic analysis tools. GenomeSpace features support for cloud-based data storage and analysis, automatic conversion of data formats, and ease of connecting new tools to the environment via a RESTful API.

The Galaxy main server is one of the first GenomeSpace-enabled tools, as well as the Galaxy-based Cistrome epigenetic analysis platform. These and the other GenomeSpace-enabled tools, including Cytoscape, GenePattern, Genomica, IGV, ArrayExpress, Genomica, and others, form a comprehensive environment for analysis of genomic data, with new resources being released regularly. We show how researchers can use GenomeSpace to combine the capabilities of these tools and how developers can add their tools to the GenomeSpace environment.

## P25: Less talking, more doing: crowd-sourcing the integration of Galaxy with a high-performance computing cluster

Dirk Colbry<sup>1</sup>, Michael R. Crusoe<sup>2</sup>, Andy Keen<sup>1</sup>, Greg Mason<sup>1</sup>, Jason Muffett<sup>1</sup>, Matthew Scholz<sup>1</sup>, Tracy K. Teal<sup>2</sup>

<sup>1</sup> Michigan State University, Institute for Cyber-Enabled Research

<sup>2</sup> Michigan State University, Department of Microbiology and Molecular Genetics

On March 5th, 2014 a team of system administrators and bioinformaticians conducted a hack-a-thon to integrate Galaxy on top of the high-performance computing cluster at Michigan State University complete with single-sign-on and the ability to run jobs as the submitting user. They elicited and received strong community support during the hack-a-thon and engaged Galaxy developers and users through IRC and Twitter. In eight hours this hack-a-thon was able to quickly navigate the various integration hurdles via real time assistance from the Galaxy community. The entire deployment was done as openly as possible with coordination of the various efforts via a separate public chat channel. While there were a couple person-days of prep and follow up, the scheduling of a single day to do the bulk of the installation proved to be critical in getting the job done and was far more effective than the many hours talking about the idea of deploying Galaxy prior. The format allowed for rapid progress as communication time was reduced and developers could modify or add components, receive prompt feedback and continue to build on the growing infrastructure. We advocate a similar recipe of using virtual machines, the Puppet configuration management system, and agile development enabled by the built-in implementations of various components of Galaxy to enable forward progress.

## P26: Galaxy Training Network

Dave Clements<sup>1,2</sup>, Vicky Schneider<sup>3</sup>, Nikhil Joshi<sup>4</sup>, Joseph Fass<sup>4</sup>, Monica Britton<sup>4</sup>, Andrew Lonie<sup>5,6</sup>, Simon Gladman<sup>5,7</sup>, Mark Crowe<sup>8</sup>

<sup>1</sup> Galaxy Project

<sup>2</sup> Johns Hopkins University, Baltimore, Maryland, United States

<sup>3</sup> The Genome Analysis Centre (TGAC), Norwich, United Kingdom

<sup>4</sup> Bioinformatics Core Facility, University of California, Davis, United States

<sup>5</sup> Life Sciences Computation Centre, Melbourne, Australia

<sup>6</sup> University of Melbourne, Melbourne, Australia

<sup>7</sup> Monash University, Australia <sup>8</sup> QFAB, Brisbane, Queensland, Australia

Scalability is a recurring challenge in all aspects of high-throughput biology, including training. There is far more demand for training than can be met by just in-person training by the core Galaxy Team.

This poster will highlight training resources that are available for teaching bioinformatics in Galaxy and for teaching using and administering Galaxy itself. This includes information about the new Galaxy Training Network. The Galaxy Training Network unifies core project and community training efforts under one umbrella so that existing training resources become more easily available, and it makes it easier for new arrivals to get up to speed with training in their locations and communities. We will also highlight directories of tutorials/worked exercises, including sample data, slide sets, videos, and computational resources such as shared virtual machine images and Amazon Web Service Machine Images (AMI's).

## P27: Integrating new visualization tool in Galaxy

Alexan ANDRIEUX<sup>1</sup>, Pierre PETERLONGO<sup>1</sup>, Yvan LE BRAS<sup>2</sup>, Cyril MONJEAUD<sup>2</sup>, Charles DELTEL<sup>3</sup>

<sup>1</sup> Genscale, INRIA, Campus de Beaulieu, 35042, Rennes Cedex, France

<sup>2</sup> GenOuest Core Facility, UMR6074 IRISA CNRS/INRIA/Université de Rennes I, Campus de Beaulieu, 35042, Rennes Cedex, France

<sup>3</sup> SED, INRIA, Campus de Beaulieu, 35042, Rennes Cedex, France

Galaxy supports adding tools, constructing workflows and analyzing diverse and large datasets. Galaxy offers some visualization tools, like Trackster and Phyloviz, but users can have difficulties finding the right visualizer to see the output of their own tools. To avoid the use of external tools, users may also want to integrate their own visualization tools.

In earlier versions of Galaxy, implementation of a new visualizer was complex because it required 1) to put each file type (JavaScript, Css, Mako, Python ...) of the new visualizer in the right place in the directories tree and also 2) edit several Galaxy source files. Recent Galaxy versions give the possibility to add visualizations more easily: You only have to give to the new visualizer the right structure and paste it. It's a good beginning even if some tasks are still difficult as for adding the Galaxy save function to the new visualizer.

The new visualization framework was tested to facilitate Mapsembler 2 outputs interpretation. This tool extends references sequences from each side with one or more sets of reads. Sometimes, several extensions are possible and Mapsembler 2, constructs a graph with each possible of extension. To view the output graph we have developed a graph viewer in JavaScript and jQuery. At the moment, this visualizer is compatible only with the Mapsembler 2 outputs, but further works will make it compatible with semantic web or Systems biology tools to visualize, for example, rdf files or biological networks. Finally, this work represents an important step towards visualization of data in Life Sciences Virtual Research Environment (introduce by the poster n°4).

## P28: Integrating GALAXY workflows in a metadata management environment

Francois MOREEWS<sup>1</sup>, Yvan LE BRAS<sup>2</sup>, Olivier Dameron<sup>3</sup>, Cyril MONJEAUD<sup>2</sup> and Olivier COLLIN<sup>2</sup>

<sup>1</sup> Genscale team, Irisa / INRA, Campus de Beaulieu, 35042, Rennes Cedex, France

<sup>2</sup> GenOuest Core Facility, UMR6074 IRISA CNRS/INRIA/Université de Rennes I, Campus de Beaulieu, 35042, Rennes Cedex, France

<sup>3</sup> Dyliss team, Irisa / Inria Rennes-Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France"

The Galaxy platform offers repositories of user data and related analysis processes (data histories and workflows). These repertories enable traceability and reproducibility of the processes within the platform. At a larger scale, to answer questions like "What protocol was used to analyze my data?" or "how were these data generated?", we could consider any protocol as a metadata set that annotates inputs and results.

We present a preliminary approach for integrating the GALAXY workflows in an extensible meta-data management environment.

Using ISA-tools, we have developed a formalism to describe an abstraction of data processing workflows. This specification, in the ISA-TAB format is named ISA-DATAFLOW.

A conversion tool extracts a structured dataflow representation in GRAPHML, a generic XML graph format, from GALAXY workflows. This intermediary format can then be normalized using controlled vocabularies and converted into ISA-TAB following our ISA-DATAFLOW specification.

We plan to integrate this work to propose advanced research functionalities within a virtual research environment (VRE) deployed on a geographically and thematically distributed infrastructure already using multiple Galaxy instances. Future developments will concern workflow meta-analysis and workflow composition assistance.

# GCC 2015

Galaxy Community Conference

6-8th July 2015

The Sainsbury Laboratory  
Norwich, UK



[galaxyproject.org](http://galaxyproject.org)



### Training Day

#### Barber Room #302

1. Visualization of NGS Data
2. Galaxy on a Cluster - User and Project Management
3. RNA-Seq Analysis with Galaxy and Alternative Tools

#### Salon A Room #303

1. Rains & Rabbit Turds: NGS Quality Control with Galaxy
2. Galaxy Automation: Using the API
3. Tool Development from Bright Idea to ToolShed - Data Managers

#### Salon B Room #303

1. Galaxy Internals: Flow Control within Galaxy
2. Tool Development from Bright Idea to ToolShed - Designing a Galaxy Tool
3. Visualization of NGS Data

#### Salon C Room #303

1. Galaxy Installation and Administration
2. RNA-Seq Analysis with Galaxy and the Tuxedo Suite
3. Scriptable Bioinformatics Cloud Infrastructures with Cloud BioLinux, CloudMan & Galaxy

#### Multipurpose Room #324

1. Training with Galaxy: a Genome Assembly Example
2. 3D Genome Analysis with Galaxy
3. Galaxy on a Cluster - User and Project Management

### Conference

Salons A, B & C  
Room #303

### Exhibitors

Barber  
Room #302

### Posters

East Room #304  
Posters 1-15

Multipurpose Room #324  
Posters 16-30

### BoFs

Everywhere

N Lovegrove St



E 33rd St

N Charles St

Mattin Center across N. Charles Street (conference dinner)

