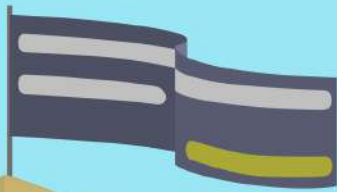


GCC 2015

Galaxy Community Conference

6-8th July 2015

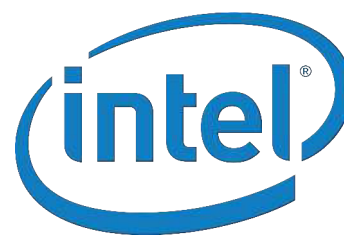


The Sainsbury Laboratory
Norwich, UK

<http://gcc2015.tsl.ac.uk>

#usegalaxy @galaxyproject

Platinum Sponsorship



Gold Sponsorship



Silver Sponsorships



Training Day Sponsor



Hackathons Sponsor



Johns Hopkins Data and Genomic Data
Science Specializations

Welcome

Welcome to the 2015 Galaxy Community Conference (GCC2015), the sixth annual gathering of the Galaxy Community. GCC2015 features two days of training, two days of hackathons, and a two day meeting with accepted and keynote talks, posters, vendor exhibits, birds-of-a-feather (BoF) gatherings, and lightning talks - all about high-throughput biological research and its supporting compute infrastructure. This event also features plenty of time for networking and impromptu gatherings.

GCC2015 features two hackathons: The GCC Coding Hackathon, now in its second year and focused on extending Galaxy in novel ways; and the new GCC Data Wrangling Hackathon, focused on challenging analysis and integration problems.

Training has also expanded this year. GCC2015 includes two days of training, starting on Sunday with a single track, with 3 sessions on using Galaxy. Monday features 5 tracks with 3 sessions per track covering the full spectrum of usage, development, and administration. Training topics were nominated and selected by the Galaxy community and reflect the community's wide range of interests. This is an excellent opportunity to get hands-on experience while learning from the experts.

We would like to give an enormous thank you to our sponsors and hosts, the Training Days instructors, the Scientific Committee, BoF organisers, speakers and poster presenters, and to anyone else who helped contribute to making this event a success.

Thank you,

The GCC2015 Organising Committee



Host

The Sainsbury Laboratory
TSL



MORE HEADROOM

BIG THINKERS TRUST SGI

www.sgi.com



Built with Intel®
Xeon® processors

sgi

Wifi Connections & Credentials

Wifi is available throughout all conference facilities. If you have EDUROAM, then please use it.

If you don't have EDUROAM, then you can connect to wireless SSID gcc2015 with password gcc2015psk

Social Media

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

#usegalaxy

@galaxyproject

Meals

If you are staying in conference lodging in Nelson Court, then breakfast is included for each day of your lodging and will be served from 7:30-9:00.

Lunches are provided as part of all events. Wednesday's lunch is sponsored by BioTeam.

The official conference dinner is Wednesday evening. It will be held in the Sainsbury Centre for the Visual Arts on the University of East Anglia campus. The conference dinner is sponsored by SGI, Kelway, and Intel.

Please thank these sponsors profusely for their support.

Breaks

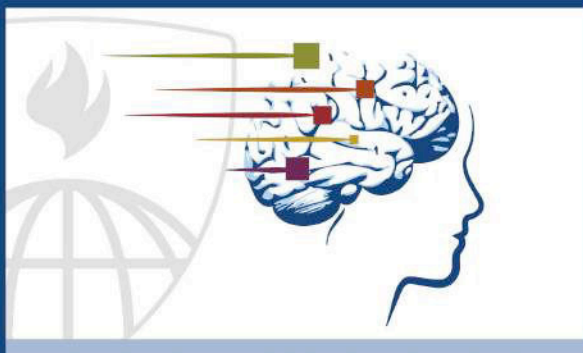
Coffee, other beverages, and light snacks will be available during breaks.

Johns Hopkins Data Science

on **coursera**

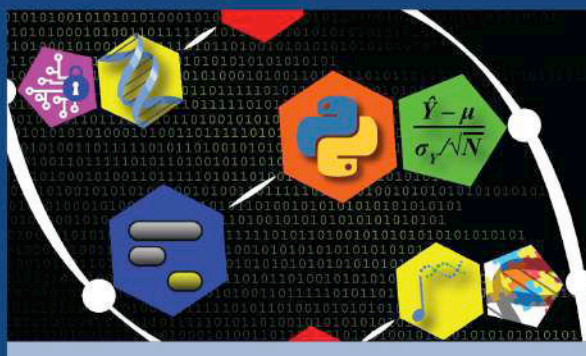
Data Science Specialization

Learn the concepts & tools you'll need throughout the entire data science pipeline, from asking the right kinds of questions to making inferences & publishing results.



Genomic Data Science Specialization

Learn to understand, analyze, & interpret data from next-generation sequencing experiments by using common tools of genomic data science, including Python, R, Bioconductor, & Galaxy.



Jeff Leek
Roger Peng
Brian Caffo
Steven Salzberg
Kasper Hansen

Mihaela Pertea
James Taylor
Liliana Florea
Ben Langmead
Jacob Pritt



coursera.org/jhu

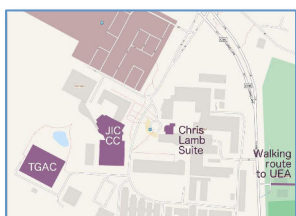
The Genome Analysis Centre

Building Excellence in Genomics and Computational Bioscience



Visit us at www.tgac.ac.uk

Getting Around



GCC2015 is being held in the Norwich Research Park, Norwich, UK.

Most GCC2015 events are being held in the *John Innes Centre Conference Centre (JICCC)*, including most training events, the GCC Conference, the poster and sponsor sessions, birds-of-a-feather meetups, lunches and breaks.

The hackathons are being held in The Genome Analysis Centre (TGAC), also in the Norwich Research Park.

See the *full Norwich Research Park map on the second to last page.*

Conference lodging and the conference dinner are both on the University of East Anglia (UEA) campus, within easy walking distance of conference events.

Conference lodging is in Nelson Court. Parking is available in the main campus car park. The conference dinner will be held at The Sainsbury Centre for Visual Arts, on the west end of the UEA campus.

See the *full UEA campus map on the last page.*

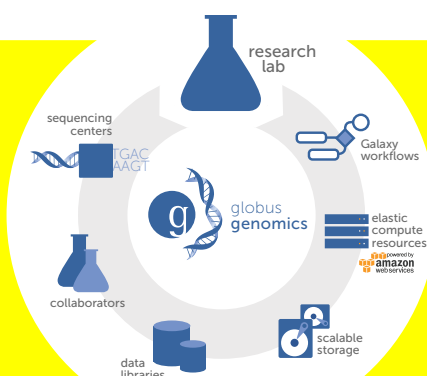


There is also a custom Google Map with all conference venues and lodging, and the Norwich airport, and train and bus stations.

See <http://bit.ly/gcc2015map>

Flexible, scalable,
affordable genomics
analysis for all
biologists.

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

GCC2015 Events

Sat 4 July	Sun 5 July	Mon 6 July	Tue 7 July	Wed 8 July
Coding Hackathon Data Wrangling Hackathon 2 intense days of coding & data analysis		Training Day 5 tracks, 15 sessions: covering the full spectrum of Galaxy usage, administration & development	GCC2015 Meeting 2 days of accepted talks, poster sessions, birds-of-a-feather, lightning talks, exhibits, keynotes, and networking.	
BioJS Hackathon Hack on BioJS and Galaxy	Training SunDay Full day, single track, on using Galaxy			
Pre-GCC2015		2015 Galaxy Community Conference		

Sunday 5 July: Training SunDay

Another day of training is being offered for the first time at a Galaxy Community Conference. Training SunDay features a single track offering three popular *Using Galaxy* topics.

Training SunDay will be in the John Innes Centre Conference Centre (JICCC) in the combined Watson and Crick rooms.



Time	Topic
08:00	Registration
08:45	Introduction to Galaxy Daniel Blankenberg
11:00	Break
11:15	RNA-Seq Analysis with Galaxy, Part 1 Andrew Stubbs, Saskia Hiltemann, Youri Hoogstrate
12:45	Catered Lunch
13:45	RNA-Seq Analysis with Galaxy, Part 2 Andrew Stubbs, Saskia Hiltemann, Youri Hoogstrate
14:45	Visualisation of NGS Data, Part 1 Carl Eberhard, Jeremy Goecks, Aysam Guerler
15:30	Break
15:50	Visualisation of NGS Data, Part 2 Carl Eberhard, Jeremy Goecks, Aysam Guerler
17:35	Done

Monday 6 July: Training MonDay

The second day of training features 15 topics in 5 parallel tracks spanning a full range of topics. Each session is two and half hours long.

Many sessions are full, so please attend only the sessions you registered for.



Time	Auditorium	Watson G34	Crick G35	Wilkins G36	Franklin G37
08:00	Registration				
09:15	Setting up a Galaxy instance as a service Hans-Rudolf Hotz, Nikolay Vazov, Jochen Bick	Scripting Galaxy using the API and BioBlend Nicola Soranzo, Dannon Baker, Carl Eberhard	Galaxy Interactive Environments Björn Grüning, Eric Rasche, Cameron Smith, John Chilton <i>In the Chris Lamb Lounge</i>	Introduction to Galaxy Daniel Blankenberg	Finding causative mutations in genomes with a Candidate SNP approach Dan MacLean
11:45	Catered Lunch				
13:00	Introduction to Writing Galaxy Tools & Publishing in Galaxy ToolShed Martin Čech, Björn Grüning, Dan Blankenberg, Dave Bouvier, John Chilton, Peter Cock, Eric Rasche, Nicola Soranzo	Advanced Workflows and Variables Jennifer Hillman-Jackson	Running Galaxy on Docker and StarCluster Gaurav Kaul, Robert Sugar	The Galaxy Database Schema Dave Clements, Nitesh Turaga	RNA-Seq Analysis with Galaxy Andrew Stubbs, Saskia Hiltmann, Youri Hoogstrate
15:30	Break				
16:00	Test-Driven Development of Galaxy Tools with Planemo & Advanced Topics in Tool Creation John Chilton, Martin Čech, Björn Grüning, Dan Blankenberg, Dave Bouvier, Peter Cock, Eric Rasche, Nicola Soranzo	Visualisation of NGS Data Jeremy Goecks, Aysam Guerler, Carl Eberhard	Variant Analysis with Galaxy Andrew Lonie, Clare Sloggett	Mass Spectrometry-based Proteomics Data Analysis using Galaxy-P Tim Griffin, Pratik Jagtap, James Johnson	Galaxy Architecture James Taylor, Nate Coraor
18:30	Training Sessions Done				
19:00	Dinner (on your own) Birds-of-a-Feather Flocking				
22:00	Finish				


Tuesday 7 July: GCC2015 Meeting Day 1

The first day of the meeting will feature accepted and lightning talks, and birds-of-a-feather meetups, poster, and sponsor sessions. Formal content will run from approximately 9:00-18:00, with birds-of-a-feather (and Hackathon followup meetings) running in to the evening.




Time	Content
07:30	Breakfast opens in Nelson Court for those staying in UEA Lodging
08:00	Registration Opens
09:00	Welcome and Opening Dan MacLean, The Sainsbury Lab
09:15	Session 1 Moderator: Dan MacLean, The Sainsbury Lab
09:15	<i>Keynote Address: Modeling molecular heterogeneity between individuals and single cells</i> Oliver Stegle, Statistical genomics and systems genetic research group, European Bioinformatics Institute (EBI)
10:00	<i>Galaxy as backend for TralT genotype to phenotype studies</i> Youri Hoogstrate, Erasmus MC Rotterdam
10:15	<i>Enabling large scale Genotype-Tissue Expression studies using Galaxy</i> Genna Gliner, Princeton University
10:30	Break
11:00	Session 2 Moderator: Anne Pajon, Cancer Research UK
11:00	<i>BioJS2Galaxy: Automatic Conversion of BioJS Visualisation Components into Galaxy Plugins</i> Sebastian Wilzbach, Technical University of Munich
11:20	<i>Proteomics Visualization in Galaxy</i> James E Johnson, Minnesota Supercomputing Institute, University of Minnesota
11:40	<i>Integration and visualization of sequence results across experiments for method development and quality control</i> Bradley W. Langhorst, New England Biolabs
12:00	<i>GSuite Tools – efficiently manage and analyze collections of genomic data</i> Boris Simovski, University of Oslo
12:15	<i>Reproducible galaxy: Improved development and administration</i> Aarif Mohamed Nazeer Batcha, Ludwig-Maximilians-University Munich
12:30	Lunch
13:30	Session 3 Moderator: Manuel Corpas, The Genome Analysis Centre
13:30	<i>State of the Galaxy</i> Anton Nekrutenko, Penn State University and James Taylor, Johns Hopkins University
14:05	<i>Galaxy and the RNA Bioinformatics Center</i> Cameron Smith and Torsten Houwaart, University of Freiburg
14:25	<i>Data-Driven Science: Advanced Storage Systems for Genomics Analysis</i> James Reaney, SGI Corp
15:00	Sponsor Exhibition and Poster Session I Odd numbered posters

16:20	<p>Session 4</p> <p>Moderator: Jeremy Goecks, George Washington University</p> <p>16:20 <i>Galaxy Tool Shed: Tool Discovery and Repository Management</i> Martin Čech, Penn State University</p> <p>16:40 <i>ReGaTE, Registration of Galaxy Tools in Elixir</i> Olivia Doppelt-Azeroual, Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur</p> <p>17:00 <i>A curated Domain centric shared Docker registry linked to the Galaxy toolshed</i> François Moreews, Genscale Team, IRISA, and Yvan le Bras, Genouest Bioinformatics facility, INRIA/IRISA</p> <p>17:15 Lightning Talks Session I</p>
18:00	Break
18:20	Dinner (on your own) Birds-of-a-Feather Flocking
22:00	Finish




APPLIANCE
BIOTEAM GALAXY EDITION


made easy.

GET RESULTS FASTER WITH A DEDICATED GALAXY SERVER

A push-button, out of the box solution designed by BioTeam

	No Wait Times	No Storage Quotas	No Job Submission Limits	No Data Transfer Bottlenecks	No Required Infrastructure	No Required Technical Experience
Galaxy Main	✗	✗	✗	✗	✓	✓
Local Galaxy	?	?	?	✓	✗	✗
Cloud Galaxy	✓	✓	✓	✗	✗	✓
BioTeam Appliance	✓	✓	✓	✓	✓	✓



Intel Xeon E5 Processors 20 Cores
384GB RAM [Upgrade to 512GB]
2x 100GB SSD [Up to 2x 400GB]
32TB Storage [Upgrade to 96TB]
10Gb Ethernet
 Configuration as shown

BE AN EARLY ACCESS PARTNER

For more information, visit bioteam.net/bioteam-appliance/galaxy-edition

Wednesday 8 July: GCC Meeting Day 2

The final day continues the program from the first day, and ends with the Conference Dinner. Formal meeting content will be finished by 18:00.



Time	Content
07:30	Breakfast opens in Nelson Court for those staying in UEA Lodging
08:00	Registration Open
09:00	Welcome
09:10	Session 5 Moderator: Graham Etherington, The Genome Analysis Centre <div> <div>09:10</div> <div> <i>A Galaxy metagenomic workflow for reference-tree based phylogenetic placement (MG-RTPP)</i> Ambrose Andongabo, Rothamsted Research </div> </div> <div> <div>09:30</div> <div> <i>Less Click, More Quick: Unattended Installation of Galaxy's Built-in Reference Data</i> Daniel Blankenberg, Pennsylvania State University </div> </div> <div> <div>09:50</div> <div> <i>Galaxy flavours – shipped by a whale</i> Björn Grüning, University of Freiburg </div> </div> <div> <div>10:10</div> <div> <i>Planemo – A Galaxy Tool SDK</i> John Chilton, Penn State University </div> </div>
10:30	Break
11:00	Session 6 Moderator: Karen Reddy, Johns Hopkins University <div> <div>11:00</div> <div> <i>Galaxy Interactive Environments – a new way to interact with your data</i> Eric Rasche, Texas A&M University; and Björn Grüning, University of Freiburg </div> </div> <div> <div>11:20</div> <div> <i>Opening Galaxy to script execution by everyone</i> Marius van den Beek, Institut de Biologie Paris-Seine </div> </div> <div> <div>11:40</div> <div> <i>Using Galaxy resources from the command line</i> Clare Sloggett, University of Melbourne </div> </div> <div> <div>11:55</div> <div> <i>Integrating Galaxy and Tripal: Cyberinfrastructure for the Genome Community Database</i> Emily Grau </div> </div> <div> <div>12:15</div> <div> <i>Simplifying IT for Local Galaxy</i> Anushka Brownley, BioTeam </div> </div>
12:30	Lunch Sponsored by BioTeam
13:30	Session 7 Moderator: Vicky Schneider, The Genome Analysis Centre <div> <div>13:30</div> <div> <i>Creating dynamic tools with Galaxy ProTo</i> Sveinung Gundersen, University of Oslo </div> </div> <div> <div>13:45</div> <div> <i>Beyond Galaxy: portable workflows and tool definitions with the CWL</i> Michael R. Crusoe, University of California, Davis </div> </div> <div> <div>14:05</div> <div> <i>Extending Galaxy's reach: recent progress towards complete multi-omic data analysis workflows</i> Timothy J Griffin, University of Minnesota </div> </div> <div> <div>14:25</div> <div> <i>A Genomics Virtual Laboratory in practice</i> Andrew Lonie, University of Melbourne </div> </div>

	14:45	<i>IRIDA: A Genomic Epidemiology Platform Built on top of Galaxy</i> Aaron Petkau, National Microbiology Laboratory, Winnipeg, Canada
15:00	Sponsor Exhibition and Poster Session II Even numbered posters	
16:20	Session 8 Moderator: Carrie Ganote, NCGAS, Indiana University	
	16:20	<i>Building Galaxy Community VM</i> Ryota Yamanaka, Genome Science Division, The University of Tokyo
	16:40	<i>An initiative to federate the galactic community in France: the IFB Galaxy Working Group</i> Olivier INIZAN, INRA URGI Versailles; Gildas LE CORGUILLE, CNRS-UPMC Station Biologique de Roscoff; Alban LERMINE, Institut Curie Paris
	17:00	Lightning Talks Session II
17:50	Closing	
18:00	Break	
19:00	Conference Dinner At the Sainsbury Centre for Visual Arts Sponsored by SGI, Kelway, and Intel	

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 gcc2015-org@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

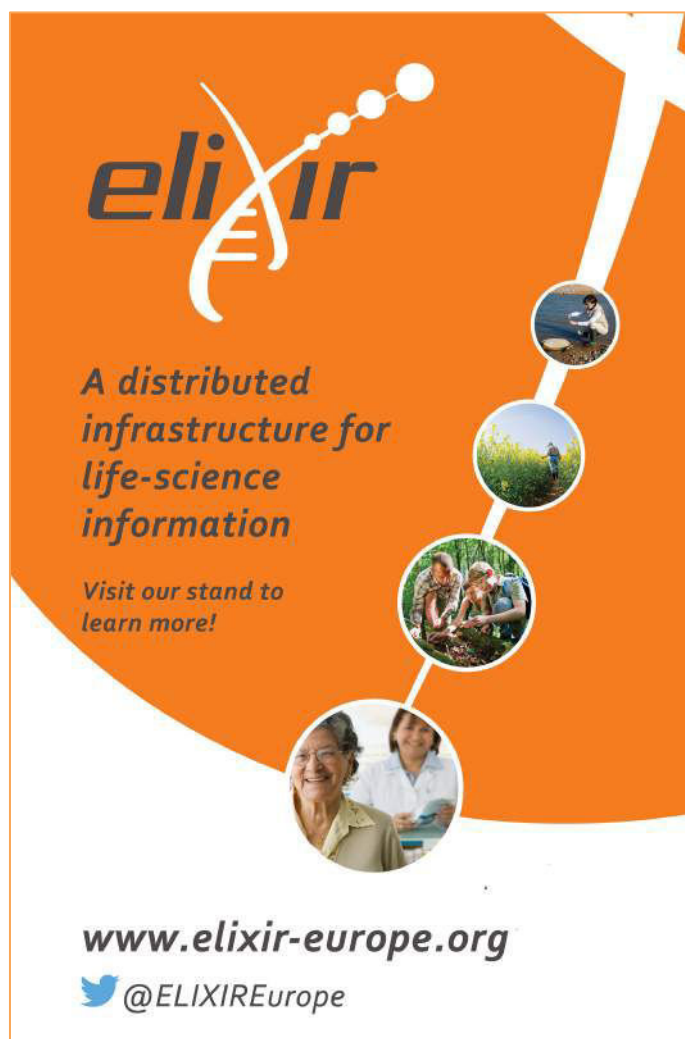
Lightning talks are 5 minutes followed by 2 minutes for questions.

- 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/gcc2015lightning> for the current list of lightning talks.



elixir

A distributed infrastructure for life-science information

Visit our stand to learn more!

www.elixir-europe.org

@ELIXIREurope

The graphic features a large orange circle with a white DNA helix structure. Along the helix, there are five circular inset images showing people in various settings: a person in a lab, a person in a field, a group of people, and two people smiling. The text is arranged around the circle, with the ELIXIR logo at the top and the website and Twitter handle at the bottom.

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. GCC2015 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 GCC2015, 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/gcc2015bofs> 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/gcc2015bofs> for BoFs that have been added since.

Galaxy Training Network

The Galaxy Training Network (GTN) is a network of trainers who teach bioinformatics using Galaxy, or teach about Galaxy itself. GTN makes it easy to find Galaxy trainers, and to share and discover the wealth of training resources available for Galaxy. This includes training materials, a trainer directory, best practices, and guidance on computing platforms for teaching with Galaxy. The Galaxy Training Network is accessible to the entire community. If you teach with Galaxy, then please consider joining us at this BoF.

See <http://bit.ly/gcc2015gtnbof>

Meeting at Monday, 6 July, 18:50, Franklin Room

GalaxyAdmins

GalaxyAdmins is for people that are responsible for administering Galaxy instances. We meet online and at events like GCC2015, where a lot of us happen to be. This BoF was very well attended in previous years and each has resulted in several actions items that were implemented in the following year. 2015 has seen a resurgence in GalaxyAdmins participation and interest.

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

See <http://bit.ly/gcc2015adminsbof>

Meeting on Tuesday, 7 July, 18:20, Franklin Room

ELIXIR-Galaxy Birds of a Feather

As part of ELIXIR's efforts to understand Galaxy usage across Europe and globally we have set up a questionnaire (<http://bit.ly/elixir-galaxy-survey>) with the objective to identify how Galaxy is used across different institutions. Regardless of national origin, any Galaxy community user or developer is welcome to fill it in.

Questionnaire results will be made public and fed into a recommendation that will influence the technical strategy across the ELIXIR pan-European bioinformatics infrastructure with regards the provision of Galaxy services and infrastructure nationally and internationally.

We would thus like to invite you to fill in this questionnaire by no later than 12pm (noon) BST on July 7th. Initial results will be discussed and presented while at the ELIXIR-Galaxy Workshop held at the Genome Seminar Room at the TGAC building (next door to the conference centre in the same Norwich Research Campus as the Galaxy Community Conference). Please add your name to the roster (<http://bit.ly/gcc2015elixirbof>) if you are interested in attending.

See <http://bit.ly/gcc2015elixirbof>

Meeting on Tuesday, 7 July, 18:00, Genome Seminar Room, TGAC

See <http://bit.ly/gcc2015bofs> for more

Organisers

Organising Committee

Dave Clements
Johns Hopkins University

Dan MacLean
The Sainsbury Laboratory

Martin Page
The Sainsbury Laboratory

Anne Pajon
University of Cambridge

Robert Ping
Indiana University

Gabriella Rustici
University of Cambridge

Vicky Schneider
The Genome Analysis Centre

Christian Schudoma
The Sainsbury Laboratory

Paul Fretter
The Norwich BioScience
Institutes

Scientific Committee

Dan MacLean
The Sainsbury Laboratory

Manuel Corpas
The Genome Analysis Centre

Gianmauro Cuccuru
CRS4

Hailiang (Leon) Mei
Leiden University Medical
Center, and the Dutch
Techcentre for Life Sciences

Katerina Michalickova
University of Oslo

Karen Reddy
Johns Hopkins University

Nicola Soranzo
The Genome Analysis Centre

James Taylor
Johns Hopkins University

Code Hackathon Organisers

Dannon Baker
Johns Hopkins University

Dan Blankenberg
Penn State University

Carrie Ganote
NCGAS, Indiana University

Martin Page
The Sainsbury Laboratory

Data Wrangling Hackathon Organisers

Jennifer Jackson
Penn State University

Karen Reddy
Johns Hopkins University

Christian Schudoma
The Sainsbury Laboratory

Integrating workflows with papers

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

- **Quick publication**— average time to first decision in 2013/14 less than 25 days
- **15% Article Processing Charge discount** (£200) to all submissions from GCC2015
- **Free deposition and curation** of your data in **GigaDB database** with no size limit
- **All data and tools can be hosted** with the journal's **gigagalaxy.net** server
- **A home & citeable DOI for data & workflows**

(GIGA)ⁿ
SCIENCE

(GIGA)ⁿ Galaxy

Editor-in-Chief: Laurie Goodman
Executive Editor: Scott Edmunds

华大基因
BGI

 **BioMed Central**
The Open Access Publisher

www.gigasciencejournal.com/series/galaxy

Training Topics

Training sessions are divided into three broad groups: *Using Galaxy*, *Using Galaxy Programmatically*, and *Deploying, Administering, and Wrapping Tools for Galaxy*.



Prerequisites

All sessions are hands on and participants should bring a wifi-enabled, fully charged laptop to participate in each session. Each session also has additional prerequisites as well.

Using Galaxy

Introduction to Galaxy

Daniel Blankenberg, Penn State University

New to Galaxy? This will introduce you to the Galaxy Project, the Galaxy Community, and walk you through a simple use case demonstrating what Galaxy can do. This session is recommended for anyone who has not used, or only rarely uses Galaxy.

Prerequisites:

- Little or no knowledge of Galaxy

Finding causative mutations in genomes with a Candidate SNP approach

Dan MacLean, The Sainsbury Laboratory

Mapping mutations by position, either using classical methods or whole genome high-throughput sequencing (HTS), largely relies on the analysis of genome-wide polymorphisms in F2 recombinant populations.

We will study high-throughput genomic sequence from genomes of back-and out-crossed bulks of plants to identify a genetic mutation caused by EMS mutagenisation of bulk segregants. The workflow demonstrated and implemented by the attendees will QC paired Illumina reads and align them against the Arabidopsis reference genome using BWA, generate a BAM file, identify SNPs using SAMtools and separate SNPs by allele frequency. We will then use SNPeff to annotate SNPs as to their effect and location in genes and generate plots that will allow us to compare the relative densities of SNP classes across the genome and reveal the candidate positions of the causative mutation.

Prerequisites:

- General knowledge of Galaxy, or attendance at the “Introduction to Galaxy” session.
- Basic understanding of genetics.

RNA-Seq Analysis with Galaxy

Andrew Stubbs, Erasmus MC
Saskia Hiltemann, Erasmus MC,
Youri Hoogstrate, Erasmus MC

This hands-on workshop will demonstrate basic RNA-Seq analysis pipelines including quality control, alignment, and differential expression analysis in Galaxy.

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.

Prerequisites:

- General knowledge of Galaxy, or attendance at the “Introduction to Galaxy” session.

Advanced Workflows and Variables

Jennifer Hillman-Jackson, Penn State University

This workshop will teach participants all they need to know in order to create their own publication and/or production quality Galaxy Workflows.

1. Basic and Advanced Workflow Editor functions.
2. Demystify the magic variables defined by the Workflow’s engine with a special emphasis on how to track data inputs and outputs: utilize labels inherited from existing datasets, prompt for user-defined labels, and/or create custom-specified labels (or portions of labels) within the Workflow itself.
3. Hands-on examples for batch processing, including how to execute using multiple input streams or Dataset Collections.
4. Tips for preparing a Workflow so it may be used effectively by others: annotation options, run-time parameter changes, and proper input selection.
5. Best Practices for Sharing or Publishing a Workflow on a Galaxy instance, be it stand-alone or embedded within a Page.

Prerequisites:

- General knowledge of Galaxy, or attendance at the “Introduction to Galaxy” session.
- A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.

Visualisation of NGS Data

Jeremy Goecks, George Washington University
Aysam Guerler, Johns Hopkins University
Carl Eberhard, Johns Hopkins University

This workshop will cover visualisation of both primary NGS analyses —alignments, variants, annotations — as well as downstream options such as heat maps, charts, and graphs.

Prerequisites:

1. General knowledge of Galaxy, or attendance at the “Introduction to Galaxy” session.
2. A wi-fi enabled laptop with a modern web browser. Google Chrome, Firefox and Safari will work best.

Variant Analysis with Galaxy

Andrew Lonie, University of Melbourne
Clare Sloggett, University of Melbourne

This tutorial is designed to introduce the tools, datatypes and workflow of variation detection using human genomic DNA using a small set of sequencing reads from chromosome 20.

In this session we will:

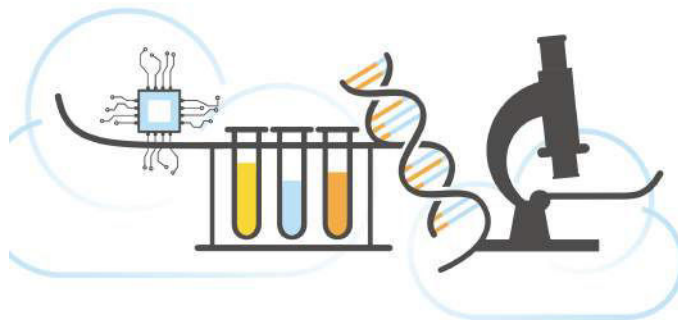
- Evaluate the quality of the short data.
- Map individual reads in the sample FASTQ readsets to a reference genome, so that we can then identify the sequence changes with respect to the reference genome. Some variant callers need extra information regarding the source of reads in order to identify the correct error profiles to use in their statistical variant detection model, so we add more information into the alignment step.
- Calling Variants using the GATK Unified Genotyper. The GATK Unified Genotyper is a Bayesian variant caller and genotyper from the Broad Institute. Many users consider the GATK to be best practice in human variant calling.
- Try an alternative caller: Mpileup
- Evaluate known variations. We know a lot about variation in humans from many empirical studies, including the 1000Genomes project, so we have some expectations on what we should see when we detect variants in a new sample.
- Annotate the detected variants against the ensemble database and interpret the annotation output.

Prerequisites:

- General knowledge of Galaxy, or attendance at the “Introduction to Galaxy” session.

Mass Spectrometry-based Proteomics Data Analysis using Galaxy-P

Tim Griffin, University of Minnesota
Pratik Jagtap, University of Minnesota
James Johnson, University of Minnesota



Accelerate Time to Science with AWS

Scientific Computing Engineered for Genomics Research

At Amazon Web Services (AWS), we want to improve our world and accelerate the pace of genomics research without costly, space-consuming IT infrastructure.

With AWS Scientific Computing solutions, you can focus on making breakthroughs while we securely provide the cloud infrastructure necessary to analyze genomic data pipelines, store petabytes of data and share your results with collaborators around the world.

Launch your first computing machine in minutes and discover why leading researchers around the world rely on:

- Amazon EC2 and S3 for on demand compute and storage that easily scales without a hardware footprint
- High Performance Computing (HPC) for simulation and engineering
- EC2 Spot Instances that let you provision one or thousands of server instances at prices typically far lower than On Demand prices
- High Throughput Computing (HTC) with Amazon Elastic MapReduce (EMR) for large data analytics and Hadoop workloads
- End-to-end surveillance tools and systems designed to provide data security and privacy

Get Started Now

Access free AWS usage credits that help you explore, invent and teach.

Apply at: www.aws.amazon.com/grants.

Read the *Architecting for Genomic Data Security and Compliance in AWS* whitepaper at:

<http://aws.amazon.com/whitepapers/>

This hands-on workshop will take participants through the essential steps for using Galaxy for the analysis of mass spectrometry (MS)-based proteomics data, focusing protein identification from large-scale datasets. After a short introduction on the basics of MS-based proteomics data types and concepts that underlie protein identification from this data, the workshop will be organized around three integrated modules, presented in this order:

1. Basic proteomic workflows for protein identification

Attendees will be taken on a tour of MS-based proteomics tools available in the Tool Shed; using some of these tools, attendees will learn methods for protein sequence database construction and manipulation, available Galaxy-based tools for sequence database searching, outputted data types and tools for collating results

2. Advanced proteomic workflows

Building on knowledge gained in module 1, attendees will learn about advanced applications in protein identification, focusing on applications that integrate genomic/transcriptomic data with proteomics data. Attendees will learn methods to construct protein databases from RNA-seq data, and downstream tools designed to evaluate the quality of protein identifications matching to

genomic/transcriptomic-derived protein sequences.

3. Visualization and interpretation of results

Attendees will gain exposure to the mechanics of visualization in Galaxy, a variety of tools in place for visualizing outputted protein identifications from upstream workflows. These include tools for data quality control. Visualization tools for interpreting results from proteogenomics applications, via mapping of identified peptides to reference genomes, will also be demonstrated.

At the end of the workshop, attendees will have working knowledge of MS-based proteomics tools in the Tool Shed, experience in setting up basic workflows for protein identification, as well as more advanced applications in proteogenomics. An understanding of available tools for results visualization and interpretation will also be gained. Participants will be given temporary accounts to local Galaxy instance at the University of Minnesota to participate in hands-on workshop activities.

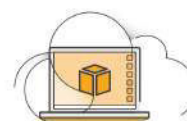
Prerequisites:

- General knowledge of Galaxy, or attendance at the "Introduction to Galaxy" session.



AWS Educate is Amazon's global initiative provides students and educators with the resources needed to greatly accelerate cloud-related learning and to help power researchers of tomorrow.

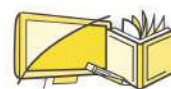
Apply at www.awseducate.com for grant-based access to:



Grants for free usage of AWS technology and services



Labs, tutorials and training on cloud computing topics and AWS products



Open source course content provided by experts professors and AWS



Communities that share best practices, both virtually and in person

Not a student or educator?

Help us extend grants to more students by sharing this on social media: #awseducate

Using Galaxy Programmatically

Scripting Galaxy using the API and BioBlend

Nicola Soranzo, The Genome Analysis Centre (TGAC)
Dannon Baker, Johns Hopkins University
Carl Eberhard, Johns Hopkins University

Galaxy has a growing API that allows for external programs to upload and download data, manage histories and datasets, run tools and workflows, and even perform admin tasks. This session will cover programmatic access of the API either by direct REST web requests or by using the BioBlend Python library.

Prerequisites:

- Basic understanding of Galaxy from a developer point of view.
- Python programming.

Deploying, Administering & Wrapping Tools for Galaxy

Setting up a Galaxy instance as a service

Hans-Rudolf Hotz, Friedrich Miescher Institute for Biomedical Research
Nikolay Vazov, University of Oslo
Jochen Bick, ETH Zürich

The premise: You are given the task to set up a Galaxy instance for others (i.e. as a core service in your institute) and you are not really familiar with Galaxy.

In this workshop, you will learn what is important when you set up a Galaxy server from scratch, what are the pitfalls you might run into, how to interact with the potential users of the service you're going to offer, and how to make sure the Galaxy instance you have set up is really used in the end. After a general introduction, several Galaxy installations are presented. The session will finish with a panel discussion, where we intend to discuss questions from the workshop participants.

Prerequisites:

- Basic knowledge of the Unix/Linux command line interface
- Familiar with the Bioinformatics problems (and their solutions) that wet lab scientists run into.

Galaxy Interactive Environments

Björn Grüning, University of Freiburg
Eric Rasche, Texas A&M University
Cameron Smith, University of Freiburg
John Chilton, Penn State University

In this session you will get an introduction to Interactive Environments (IE) as an easy and powerful way to integrate arbitrary interactive web services into Galaxy.

We will demonstrate the IPython Galaxy project and the general concept of IE's. Moreover, we will create an IE on-the-fly to get you started!

Prerequisites:

- Basic understanding of Galaxy from a developer point of view.

Running Galaxy on Docker and StarCluster

Gaurav Kaul, Intel Corporation
Robert Sugar, Intel Corporation

Two different methods of running Galaxy would be covered

1. As a Docker container : here we will cover the fundamentals of Docker containers and why would you want to use them for running your pipeline. After the overview we will have a hands on session of running Docker Galaxy image and running the deepTools pipeline
2. Managing Galaxy using Starcluster : The STAR (Software Tools for Academics and Researchers) program at MIT provides a command-line tool called StarCluster. This tool has a number of sub commands, which can be used to create, manage, login to, stop, and destroy clusters of one or more VM instances on EC2. Although StarCluster does not natively support Galaxy (yet), it provides convenient command tool chain to manage EC2 AMI (which could be the CloudMan instances running Galaxy servers). The real utility of StarCluster comes when doing development on Galaxy ToolShed, whose workflow we will demonstrate as part of the hands on.

Prerequisites:

- Python
- Linux Shell Scripting

Introduction to Writing Galaxy Tools and Publishing in Galaxy ToolShed

Martin Čech, Penn State University
Björn Grüning, University of Freiburg
Dan Blankenberg, Penn State University
Dave Bouvier, Penn State University
John Chilton, Penn State University
Peter Cock, The James Hutton Institute
Eric Rasche, Texas A&M University
Nicola Soranzo, The Genome Analysis Centre (TGAC)

This tutorial will teach developers and bioinformaticians how to take a working script or application and turn it into a Galaxy tool. It will cover the basics of wrapping, common parameters, tool linting, best practices, loading tools into Galaxy, add citations, and publishing tools to the Github and Galaxy Tool Shed. Common tips and tricks will be discussed as well as insights from some of the best tool developers out there.

Prerequisites:

- General knowledge of Galaxy, or attendance at the “Introduction to Galaxy” session.
- Familiarity with Unix command line and text editors

Test-Driven Development of Galaxy Tools with Planemo & Advanced Topics in Tool Creation

John Chilton, Penn State University
Martin Čech, Penn State University
Björn Grüning, University of Freiburg
Dan Blankenberg, Penn State University
Dave Bouvier, Penn State University
Peter Cock, The James Hutton Institute
Eric Rasche, Texas A&M University
Nicola Soranzo, The Genome Analysis Centre (TGAC)

This tutorial is aimed at people with some experience developing tools and will cover more advanced topics in tool development, more complex tools, and recent enhancements to the Galaxy tool development process including:

- Using Planemo, a new command-line application to aid Galaxy tool development, to develop Galaxy tools using a test driven development methodology.
- Designing tools for use with the dataset collections.
- Publishing complex tools to the Galaxy Tool Shed.
- Maintaining Galaxy Tools.

Prerequisites:

- Basic Knowledge of Galaxy Tools, or attendance at the Introduction to Writing Galaxy Tools and Publishing in Galaxy Tool Shed session.

The Galaxy Database Schema

Dave Clements, Johns Hopkins University
Nitesh Turaga, Johns Hopkins University

Running a production Galaxy server, you some times end up in with a situation, where you manually need to interact with the database. e.g. you need to change the state of a job to ‘error’. This is always a very risky adventure. Or a not-at-all risky situation: you want to extract usage information, which can not be gathered using the given report tools. For both cases, you need a good understanding of the Galaxy database schema.

Learn some of the design concepts of the database, which parts of the schema are stable, and which will be changing in the foreseeable future.

Prerequisites:

- Experience maintaining a production Galaxy server (recommended)
- Basic knowledge of relational databases and SQL statements

Galaxy Architecture

James Taylor, Johns Hopkins University
Nate Coraor, Penn State University

Want to know the big picture about what is going on *inside* Galaxy? This workshop will introduce participants to the high-level architecture of Galaxy internals, and to the project’s coding practices and standards.

Prerequisites:

- General knowledge of Galaxy, or attendance at the “Introduction to Galaxy” session.
- Knowledge of programming or a scripting language.

Accepted and Keynote Talks

Session 1: 09:15 Tuesday 7 July

Modeling molecular heterogeneity between individuals and single cells

Oliver Stegle¹

¹ The European Bioinformatics Institute (EBI), Hinxton, United Kingdom

The analysis of large-scale expression datasets is frequently compromised by hidden structure between samples. In the context of genetic association studies, this structure can be linked to differences between individuals, which can reflect their genetic makeup (such as population structure) or be traced back to environmental and technical factors.

In this talk, I will discuss statistical methods to reconstruct this structure from the observed data to account for it in genetic analyses.

In the second part of this talk I will extend the introduced class of latent variable models to model biological and technical sources of heterogeneity in single-cell transcriptome datasets. In applications to a T helper cell differentiation study, we show how this model allows for dissecting expression patterns of individual genes and reveals new substructure between cells that is linked to cell differentiation.

I will finish with an outlook of modeling challenges and initial solutions that enable combining multiple omics layers that are profiled in the same set of single cells.

Galaxy as backend for TraIT genotype to phenotype studies

Youri Hoogstrate¹, Freek de Bruijn², Ruslan Forostianov³, Wim van der Linden⁴

¹ Erasmus MC Rotterdam

² VUmc Amsterdam

³ The Hyve NL

⁴ Philips

The Center for Translation and Molecular Medicine Translational Research IT project (TraIT) aims to facilitate an IT infrastructure for translation research, and to enable multi-domain access to clinical, imaging, biobanking and experimental data.

TraIT offers a public Galaxy server (<http://galaxy-demo.ctmm-trait.nl/>) for general use and a private Galaxy server (<http://galaxy.ctmm-trait.nl/>) that can be securely used by anyone participating in collaborating biomedical studies. Several Galaxy tools and workflows have been created specifically for CTMM projects which include CGtag, RNA-Seq EdgeR, QDNAseq Copynumber Aberration Tool and iReport.

For the current release of our TraIT platform, -omics results and experimental meta-data are integrated in a

datawarehouse, tranSMART, whilst the analytical workflows are delivered to the end user from TraIT Galaxy. Results from user cohort selection in tranSMART can be analysed using our tranSMART to Galaxy API service which processes the data on the galaxy server and returns the resultant output (tables, visuals, etc ..) back to tranSMART.

To extend our current TraIT analytical infrastructure to other genotype to phenotype resources we plan to develop, in collaboration with the European Bioinformatics Institute, a generalised European Genome-phenome Archive (EGA) Galaxy connector with functionality similar to the existing Galaxy-European Nucleotide Archive connector. This connectivity with EGA will deliver an “end to end” analytical environment for genotype to phenotype analysis with the TraIT platform (Galaxy & tranSMART).

Enabling large scale Genotype-Tissue Expression studies using Galaxy

Genna Gliner¹, Ian McDowell², Barbara E Engelhardt³

¹ Operations Research and Financial Engineering Department, Princeton University

² Computational Biology and Bioinformatics, Duke University

³ Computer Science Department and Center for Statistics and Machine Learning, Princeton University

The Princeton BEEHIVE Group develops statistical models and methods for high-dimensional genomic data. As part of the Genotype-Tissue Expression (GTEx) consortium, we are involved in processing vast quantities of RNA-sequencing and whole genome sequence data for different types of statistical and functional genomics studies, including cis- and trans-eQTLs, non-coding RNA regulation, and allele specific expression studies. The creation, testing, and deployment of the processing pipelines for each of these different study types require comprehensive analysis of large datasets through a dedicated pipeline used by all members of the group. With the ability to create custom tools and share and modify workflows, Galaxy provides a robust framework to develop this pipeline for use across our lab, but incorporating our diverse set of analysis tools into Galaxy is a non-trivial task.

In this talk we chronicle the evolution of the Princeton BEEHIVE Galaxy Pipeline. We illustrate our vision for a flexible, scalable, and streamlined pipeline using Galaxy for statistical genomics studies. We explore how our pipeline evolved by highlighting how our lab addressed the challenges of tool creation and integration, data processing and organization, and training lab members to use our Galaxy instance.

BioJS2Galaxy: Automatic Conversion of BioJS Visualisation Components into Galaxy Plugins

Sebastian Wilzbach¹, Manuel Corpas²

¹ Technical University of Munich, BioJS Project

² The Genome Analysis Centre

BioJS (<http://biojs.net>) is an open source library for visualisation of biological data on the web. Using the latest JavaScript technologies, BioJS provides interactive modular components that can be reused, combined and extended. To date the BioJS registry (<http://biojs.io>) provides 93 reusable components applicable to biological data in a variety of -Omics domains. Many of these BioJS components are suitable for integration within the Galaxy framework. The advantages for integrating such BioJS components into Galaxy are multiple as they can: a) complement existing data exploration functionality, b) provide additional visual analytics capabilities and c) enhance human cognition when interpreting results. As most users interact with the Galaxy platform via their web interface, the Galaxy project has already recognised the need to allow the community to contribute custom-made visualisation tools. A specification has thus been proposed for contribution of “Galaxy visualisation plugins”. Making use of this specification, we have developed BioJS2Galaxy, a tool that automatically converts BioJS components into Galaxy visualisation plugins. In this presentation we will showcase BioJS2Galaxy with BioJS’s “Multiple Sequence Aligner Viewer” (<http://msa.biojs.net>), integrated within the Galaxy framework as proof of concept. BioJS2Galaxy is entirely written in JavaScript, licensed under Apache 2 and is freely available on GitHub at <https://github.com/biojs/biojs2galaxy>.

Proteomics Visualization in Galaxy

Thomas McGowan¹, James E Johnson¹, Ira Cooke², Pratik D Jagtap^{3,4}, Timothy Griffin^{3,4}

¹ Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota, United States

² Life Sciences Computation Centre, La Trobe University, Melbourne, Australia

³ Center for Mass Spectrometry and Proteomics, University of Minnesota, Minneapolis, Minnesota, United States

⁴ Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, Minnesota, United States

The Galaxy-P project incorporated proteomics tools into the Galaxy framework, enabling multi-omics analysis from a single framework. A centralized Galaxy server can provide the computational and storage capacity required to manage the large and diverse datasets and applications required for that analysis, and in addition manage collaborative access.

A researcher may want to investigate results in a proteomics analysis using interactive visualization, for example to verify peptide spectral matches (PSMs). While there are excellent applications for visualizing PSMs, many require downloading large data files to the user’s computer. In addition, many proteomics applications record cross-references to input data using absolute file paths. This limits portability of the resulting output datasets.

The Galaxy Visualization plugin framework provides a lightweight solution for visualization that allows the big data to remain on the server. Visualization of a Galaxy dataset

requires a REST-like dataset dataprovider that can respond interactively to client requests.

To achieve interactive access for proteomics data, we added the SQLite datatype into Galaxy along with dataproviders to return results for client queries. We developed Galaxy tools to consolidate proteomics datasets into a sqlite dataset.

Our ProtViz Galaxy visualization plugin offers tabular views of the proteomics data from which to inspect and filter results, and visualization components, such as the lorikeet viewer, to analyze individual PSMs. We demonstrate the use of these novel visualization tools in interpreting and filtering results from MS-based proteomics data.

Integration and visualization of sequence results across experiments for method development and quality control

Bradley W. Langhorst¹, Erbay Yigit¹, Eileen T. Dimalanta¹, Theodore B. Davis¹

¹ New England Biolabs

Manual synthesis of results across experiments is error-prone and laborious. We have constructed a galaxy tool, database, and visualization solution, SeqResults, to capture results and metadata and use it to understand how changes to library preparation affect sequence data. This extensible system includes modules to extract information from bam files, fastq files, coverage bed files, GC bias, and other summary metrics. Results and metadata are acquired in galaxy and sent to a custom relational database where they can be edited and deleted using a simple web front end. Finally we have constructed dynamic visualization tools to allow users to select data by date, flowcell, sample name, run type, read group, etc and compare sequence quality metrics, artifacts, coverage depths, etc. The SeqResults system has captured millions of data points generated from more ~1700 sequence experiments so far and continues to grow.

GSuite Tools – efficiently manage and analyze collections of genomic data

Boris Simovski¹, Sveinung Gundersen¹, Abdulrahman Azab¹, Diana Domanska², Eivind Hovig¹, Geir Kjetil Sandve¹

¹ University of Oslo

² University of Silesia

Advances in sequencing technologies provide abundance of genomic data, often in the form of genomic tracks. There is already a multitude of tools that can handle and analyze single tracks, but not many that allow one to efficiently manage and meaningfully analyze large collections of tracks, even though it is the natural next step in genome analysis. We here propose a simple, extensible tabular format called GSuite (Genomic Suite) for representing collections of datasets, along with a set of tools that allow efficient retrieval of collections of datasets from public repositories like ENCODE, convenient manipulation of each dataset in a collection, as well as novel analyses involving the full collections. The toolkit is openly available at <http://hyperbrowser.uio.no/gsuite> as an extension to the existing Genomic HyperBrowser, which is powered by Galaxy. Dataset lists in Galaxy provide a similar concept to GSuite, allowing analysis on each track (or track pair) in a list and downloading of all tracks in a list, but has limited

features and requires manual list compilation in its present form. GSuite Tools provides various forms of automated compilation of GSuite files, provides a simple means to include metadata with each dataset, and provides greater ease of manipulation on both collections and individual datasets. To explore features and potential use cases freely, we have developed GSuite independently of the existing dataset lists functionality in Galaxy, but will work towards a tighter integration or even a merger of the two.

Reproducible galaxy: Improved development and administration

Aarif Mohamed Nazeer Batcha¹, Sebastian Schaaf, Guokun Zhang, Sandra Fischer, Ashok Varadharajan, Ulrich Mansmann

¹ Ludwig-Maximilians-Universität München, Germany

Ever faced the issue not to turn the need for reproducibility into reality? The Munich NGS-Fablab wasn't an exception. Creating an NGS infrastructure dedicated for clinicians and research scholars in order to perform some experimental diagnostic procedures and basic biomedical research was the

task at hand. Reproducing the results and its working environment is as important in medicine as in other fields. Understanding the issues of reproducibility, intra- and inter-compatibility among instances within and outside our institute, we came up with our own shell setup scripts introduced in GCC2014. Later, we converted our scripts into more elegant Ansible-playbooks.

Ansible is one of the easy-to-script, open-source software platforms for configuration and management of computers. Ansible from our point of view should be regarded as revolutionary for administration and development as Galaxy turned out to be for scientific users in a bioinformatics flavored setting: it disburdens from technical 'house-keeping' work and thus enables more sophisticated work. It manages nodes over SSH. These ansible playbooks have also been used at the Galaxy Main. Our Ansible-playbook scripts can setup an orderly and clean working environment completely within few minutes by providing an inifile and a blank UNIX. Although developed in SLES, the scripts were modularized for further developments to work on other linux systems. We would like to present a short review on the experiences we gained and the flow towards ansible scripting and finally provide answers to the question "what DevOps take home from their daily work?"

Session 3: 13:30 Tuesday 7 July

State of the Galaxy

Anton Nekrutenko¹, James Taylor²

¹ Penn State University

² Johns Hopkins University

A review of what's happened and what's coming in the Galaxy Project.

Galaxy and the RNA Bioinformatics Center

Cameron Smith¹, Torsten Houwaart¹, Björn Grüning¹

¹ Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany

The recently launched German Network for Bioinformatics Infrastructure aims to provide comprehensive bioinformatics services to users in life sciences research, industry and medicine. Within this network, the RNA Bioinformatics Center (RBC) is responsible for supporting RNA related research in Germany, such as the detection of noncoding RNAs and RNA structure prediction. In this talk we will present the RBC and the RNA workbench in more detail.

The RNA workbench is a ready-to-run Docker based Galaxy instance, bundled with a variety of RNA analysis tools, sample data and teaching material. This image has already been proven to be useful as a platform for training users in Galaxy driven bioinformatics analysis.

Support of RNA research also includes enabling seamless access to diverse data sources. As a first step towards this

goal, the RBC has extended the Galaxy documentation by providing a base example for including external databases as Galaxy accessible sources. Our experience with data sources and the communication with different database administrators will be outlined.

The RNA-workbench provides a well documented interface for creating new Galaxy flavours, allowing users to easily include their chosen toolset, define desired indices and provide custom data. We would like to raise the awareness of the importance of RNA related research and to kickstart an RNA focused Galaxy community.

Data-Driven Science: Advanced Storage Systems for Genomics Analysis

James Reaney¹

¹ SGI Corp

A brief perspective of computational solutions for genomics analysis with an eye towards how the generation and manipulation of genomics data has both enabled and constrained the science. An overview of a few SGI customers and their workflows in the genomics research space is presented. With ever-expanding genomics workflows in mind, we will introduce the SGI UV system with NVMe storage as a tool capable of addressing both present and especially future workflows, enabling the science in ways not possible with other architectures.

Galaxy Tool Shed: Tool Discovery and Repository Management

Martin Čech¹, Galaxy Team²

¹ Department of Biochemistry and Molecular Biology, PSU, USA, <http://galaxyproject.org/>

² <https://wiki.galaxyproject.org/GalaxyTeam>

Galaxy uses the Tool Shed (TS) as an App Store-like platform for tool exploration and deployment with support for sharing reproducible workflows. This talk will review the current state of the TS, and recent and upcoming work.

Today the TS contains over 3000 tools in many areas of computational research, and a vibrant community is updating and improving these tools, led by the efforts of the Intergalactic Utilities Commission (IUC).

The Fall 2014 questionnaire identified tool discovery and repository management as priorities areas for the TS. We have rewritten search from scratch to allow deployers to easily identify high quality repositories. A review by the IUC, significant traffic, good ratings, and the number of downloads are all indications of high quality repositories, and can also be used to increase these repositories' visibility. Moreover it is now possible to search for individual tools directly, rather than just repositories (which may contain multiple tools).

Groups have also been introduced in the TS ecosystem. This feature aims at unified presentation of labs and development teams and their consolidated work. To streamline the process of tool development authors will soon be able to work on new additions to their repositories in private mode, affording more control over what is visible to users and what is still work in progress.

ReGaTE, Registration of Galaxy Tools in Elixir

Olivia Doppelt-Azeroual¹, Fabien Mareuil¹, Eric Deveau¹, Matus Kalas², Hervé Menager¹

¹ Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France

² Computational Biology Unit, University of Bergen, Norway

ReGaTE is a software component enabling the automated publication of Galaxy tools and workflows into the ELIXIR Tools and Data Services Registry (<https://elixir-registry.cbs.dtu.dk/#/>). This registry is a web portal for the exploration of bioinformatics resources, such as software packages, web services, websites, or reference databases. Through a dedicated interface, its users can search and locate relevant tools and data resources, and bioinformatics resource providers can enhance the visibility of their services. The registration of resources in the registry can be performed either manually, by filling a form on a web user interface, and providing the required description elements, or automatically by using the registry API.

ReGaTE uses the BioBlend API and the Registry API to completely automate the registration of the tools installed on any given Galaxy portal.

Central to the development of this tool is the mapping of the Galaxy datatype system to the EDAM Ontology. EDAM provides a controlled vocabulary for the description of scientific topics, software operations, types of data and data formats and it is used to describe the contents of the ELIXIR Registry. This mapping enables the automation of the registration of Galaxy tools by describing the format of their

input and output data in the controlled vocabulary of the registry.

This mapping is being developed in collaboration with members of the Galaxy team, the EDAM ontology and the Common Workflow Language project.

ReGaTE is available at <http://github.com/bioinfo-center-pasteur-fr/ReGaTE>.

A curated Domain centric shared Docker registry linked to the Galaxy toolshed

François Moreews¹, Olivier Sallou², Yvan le Bras², Marie Grosjean³, Cyril Monjeaud², Thomas Darde⁴, Olivier Collin², Christophe Blanchet³

¹ Genscale team -IRISA -Rennes, France

² Genouest Bioinformatics facility – INRIA/IRISA – Rennes, France

³ French Institute of Bioinformatics – CNRS IFB-Core UMS3601 – Gif-sur-Yvette, France

⁴ INSERM U625 – Rennes France

Nowadays, Docker containers are used to ease application deployment, from command lines tools to cluster management¹. This technology has a strong impact in bioinformatics where specialized software can often require multiple dependencies. It is a long term preservation solution for legacy and unmaintained tools and it enables a better process isolation in a multi-user environment. Docker as a way to quickly integrate new tools is already used with Galaxy. We have setup a functional prototype of a web registry of Docker images, BioShaDock,² dedicated to bioinformatics tools and utilities. We created a set of tools descriptors based on Docker images available in our toolshed³. Even if a general purpose registry can be used to hold shared Docker containers, we think that a domain centric registry, e.g. for the French life science community through a registry linked to the cloud of the French Institute of Bioinformatics (IFB⁸), would have a significant impact on bioinformatician productivity and help to spread best practices. With a clear open source and domain orientation, it could federate container providers^{4,5} more easily. It would also be able to include validation and curation to eliminate redundant tools, organize versioning and standardize documentation. Future works will concern advanced searching capabilities, possible referencing within the ELIXIR Tools and Data Services Registry⁶ and in the IFB one (as the ELIXIR French node). We want also to contribute to standardize containers⁷ and evaluate if benchmarks⁵ could be produced from a meta-data enriched, Docker registry.

References:

¹ Google Kubernetes, Docker container cluster management : kubernetes.io

² BioShaDock, a Bioinformatics Shared Docker registry : <http://docker-ui.genouest.org>

³ GUGGO Galaxy Tooshed : <http://toolshed.genouest.org>

⁴ Hexabio Docker repository : <http://biodocker.github.io>

⁵ Nucleotid.es, continuous, objective and reproducible evaluation of genome assemblers using docker containers: <http://nucleotid.es>

⁶ ELIXIR Tools and Data Services Registry : <https://elixir-registry.cbs.dtu.dk>

⁷ Bioboxes, a standard for creating interchangeable bioinformatics software containers : <http://bioboxes.org>

⁸ IFB academic Cloud : <http://www.france-bioinformatique.fr/?q=en/core/e-infrastructure-team/ifb-cloud>

A galaxy metagenomic workflow for reference-tree based phylogenetic placement (MG-RTPP)

Ambrose Andongabo^{1*}, Ian M. Clark^{1*}, Dariush Rowlands¹, Keywan Hassani-Pak¹, Penny R. Hirsch¹, Elisa Izoa¹, Andy Neal^{1*}

¹ Rothamsted Research, Harpenden, United Kingdom
* contributed equally

Background: High-throughput sequencing of environmental nucleic acids is revolutionizing and dramatically expanding our understanding of the diversity and functionality of complex microbial communities. There are a number of tools which allow community structure to be surveyed using metagenomics or meta-transcriptomics at the rRNA level, or by using COG- or KEGG-based functional assignments. However, there are limited complementary approaches to investigate the phylogenetic diversity of functionally important individual genes in large sequence databases.

Results: We have designed a workflow for reference-tree based phylogenetic placement (MG-RTPP) of metagenomics and meta-transcriptomics samples. The inputs to the workflow are unassembled reads, a multiple sequence alignment (MSA) of the genes of interest and large public sequence databases. Reference nucleotide profile hidden Markov models (pHMMs) are built from the MSA and are used as queries. Homologous reads are checked for accuracy before being placed on a reference phylogenetic tree, maximising phylogenetic likelihood. The workflow retains considerable flexibility, allowing for tuning of redundancy in the nucleotide pHMMs used as queries to recover as many true hits as possible.

Conclusions: MG-RTPP facilitates fast interrogation of sequence databases in a flexible and robust fashion. It avoids misidentification of false positives while pHMM tuning allows for maximum recovery of sequences. Phylogenetic placement provides unique visualization approaches which reveal the phylogenetic relationships between environment-derived sequences and sequenced organisms and between samples. The approach compliments tools such as QIIME, MG-RAST and MEGAN in allowing interrogation of individual gene abundance and diversity in samples.
Keywords: metagenome, metatranscriptome, assembly-free, community analysis, functional genes, phylogeny.

Less Click, More Quick: Unattended Installation of Galaxy's Built-in Reference Data

Daniel Blankenberg^{1,2}, The Galaxy Team²

¹ Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802

² <https://www.galaxyproject.org>

Once a Galaxy administrator has completed the configuration of their Galaxy instance as a production server, they will have a well-tuned machine that is capable of many things, but will actually do very little. In order to allow users to perform useful analyses, the administrator will need to install the desired set of tools. While this formidable task can be accomplished using the Galaxy ToolShed, it only solves part of the problem. These tools lack the reference datasets needed to make them really useful. Data Managers allow an administrator to install built-in datasets through a web-based interface. Traditionally, an administrator performs each part manually in a step-wise fashion: obtaining genomes directly from their source repositories and then building indexes on the retrieved genomes. Although this

solves many technical hurdles, it is time consuming and repetitive. This no longer needs to be the case.

Here, we demonstrate new enhancements to the Data Manager framework that greatly eases the burden of configuring large amounts of reference data. By harnessing the Galaxy Project's rysnc server, we allow Galaxy administrators to quickly and effortlessly fetch and configure the pre-computed reference data utilized at UseGalaxy.org. Administrators can filter by dbkey or Data Table name, or simply grab it all. Additionally, we provide a set of utilities to streamline the entire process. Using these utilities, an administrator can perform all the steps needed for populating pre-built data, from defining new dbkeys to building any number of mapping indexes or other reference data, with a single command.

Galaxy flavours – shipped by a whale

Björn Grüning¹, Eric Rasche², John Chilton³, Dannon Baker⁴

¹ Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany

² Center for Phage Technology, Texas A&M University, USA

³ Department of Biochemistry and Molecular Biology, PSU, USA

⁴ Department of Biology, Johns Hopkins University, USA

Project: <https://registry.hub.docker.com/u/bgruening/galaxy-stable/>

Code: <https://github.com/bgruening/docker-galaxy-stable>

License: MIT

For years Galaxy has made advanced bioinformatics software accessible to biologists directly by providing an intuitive web interface to these applications while fostering reproducibility through the automatic creation of re-runnable protocols of each analysis. With the Tool Shed, Galaxy gained a flexible deployment platform enabling identical software installations across Galaxies.

A major hurdle in using Galaxy today is simply finding an instance with the correct set of tools and with enough computational power and storage necessary for a particular analysis. An interesting solution to this challenge (and one required by certain data usage policies) is to move Galaxy and tools to the data instead of the more traditional approach of uploading the data to a remote server running Galaxy.

Galaxy is using Docker to solve this problem in a way that achieves an even greater level of reproducibility by delivering the entire software stack in a container. Each new release of Galaxy is now available as a production-ready Docker container. Additionally, this Docker image can be extended to build personalized Galaxy flavours, with site-specific sets of tools. For example, a Galaxy Docker flavour containing all necessary tools for RNA-seq analysis, or a genome annotation flavour with the NCBI BLAST suite. These flavours are simple to create and can be easily deployed on Linux, OS-X and Windows.

For more traditional Galaxy deployments, tools may now be configured to run securely in Docker containers. The isolation provided by running Galaxy jobs in this fashion provides a much higher degree of security than running them as native processes allowing both process and data isolation.

Planemo – A Galaxy Tool SDK

John Chilton¹, Björn Grüning², Eric Rasche³, Kyle Ellrott⁴, Galaxy Team

¹ Department of Biochemistry and Molecular Biology, PSU, USA

² Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany

³ Center for Phage Technology, Texas A&M University, USA

⁴ Center for Biomolecular Science & Engineering, University of California Santa Cruz, USA

This talk will summarize a year's worth of focus on making tools more expressive as well as easier to develop and test. These efforts will be highlighted in part through the prism of the new command-line toolkit Planemo.

Last fall the Galaxy team solicited community feedback via questionnaire and identified testing as the largest hurdle in the tool development process. So Planemo was built to vastly simplify tool testing and Galaxy now features many new testing facilities allowing more expressive tests.

Additionally, Planemo features “linting” functionality to catch tool and Tool Shed artifact problems before even running tests.

This is but one process improvement of many made over the last year, the Galaxy Intergalactic Utilities Commission and core development team have moved all tool development to GitHub and written a best practice guide for tool development. This talk will discuss the benefits of GitHub as well as Jenkins build scripts leveraging Planemo and Tool Shed API enhancements developed to automate testing and publishing tool repositories en masse to the Tool Shed.

In addition to making current developers more productive, Planemo lowers the barriers to entry. Planemo has great documentation and utilities to bootstrap new best-practice tools quickly using example commands as templates. Developed in part for entrants to the DREAM SMC-Het challenge (which will introduce many new developers to the ecosystem by including a reproducibility focused sub-challenge requiring submission of Galaxy tools and workflows) Planemo virtual appliances package complete development environments for Galaxy tools.

Session 6: 13:30 Wednesday 8 July

Galaxy Interactive Environments – a new way to interact with your data

Eric Rasche¹, Björn Grüning², John Chilton³, Dannon Baker⁴

¹ Texas A&M University, College Station, Texas, United States

² University of Freiburg, Germany;

³ Penn State University, Pennsylvania, United States

⁴ Johns Hopkins University, Maryland, United States

A common complaint leveled at Galaxy by bioinformaticians is that it lacks the flexibility and interactivity of the Unix shell and scripting languages. Heavy use of the command line often results in homemade scripts and a non-portable and non-transparent analysis, which is hard for biologists to understand, and hard for bioinformaticians to reproduce.

Here, we present a new concept in Galaxy called Interactive Environments (IEs). IEs are perfectly suited for bioinformaticians and can offer the missing flexibility of a modern scripting language – even shell access if desired – enabling rapid, iterative, and interactive bioinformatics analysis and software prototyping directly in Galaxy, next to your big data. IEs reduce the barriers that bioinformaticians often encounter while using Galaxy.

We will present one IE in detail that integrates the popular IPython environment in a secure manner in Galaxy. IPython is a platform providing a web-based interactive computing and visualization environment. Galaxy IPython allows Galaxy users to run IPython inside Galaxy and access it via their web browser. Additionally, it extends the default IPython environment by providing easy, secure access to Galaxy, its API, and the user's data. As Galaxy IPython is deployed on the Galaxy server, it removes the overhead of big-data downloads and uploads during analysis. Galaxy has long been a great platform for bioinformatics education, but Galaxy IPython makes it a great platform to teaching bioinformatics programming as well.

Opening Galaxy to script execution by everyone

Marius van den Beek¹, Christophe Antoniewski¹

¹ Institut de Biologie Paris-Seine, Bioinformatics analysis platform

Currently galaxy users are limited to tools that are already wrapped and installed in a Galaxy instance. While important in ensuring accessibility and reproducibility, tool wrapping remains a hurdle for users and developers not familiar to Galaxy's tool wrapping process. In addition complex workflows that involve loops and/or conditions have not been implemented in Galaxy to date.

To circumvent these limitations we extended Ross Lazarus' Galaxy Tool Factory into the Docker toolfactory. This tool sends script execution into an isolated docker container that only has access to the script, the input and the output data. We will demonstrate that the docker toolfactory opens up the possibility for bioinformaticians to run and store their scripts within a history entry, side by side with its input and output data. Other applications include execution of complex workflows through API scripts that are not possible solely by using galaxy's UI and the possibility to run and store very specific scripts that were required to generate figures in a publication.

In combination with interactive environments, such as IPython and Rstudio, our tool improves the attractiveness of Galaxy as a development platform for any bioinformatician/data-analyst. It also reduces the barrier to learn writing scripts, as one can still use all the features of galaxy, such as pre-installed tools, workflows, libraries, cluster and job management, while focusing on the script. The docker toolfactory is available in the testtoolshed and <https://bitbucket.org/mvdbEEK/dockertoolfactory>.

Using Galaxy resources from the command line

Clare Sloggett¹, Nuwan Goonasekera¹, David Powell², Simon Gladman¹, Enis Afgan³, Andrew Lonie¹

¹ University of Melbourne, Australia

² Monash University, Australia

³ John Hopkins University, USA

As a part of the Genomics Virtual Laboratory project¹, we have built CloudMan-enabled, scalable machine images providing bioinformatics researchers with Galaxy, RStudio, IPython Notebook, and the linux command line in one server. This allows users to work in different environments, and to move between platforms as appropriate – for instance, carrying out parts of an analysis in Galaxy and RStudio seamlessly.

To handle the technical challenge of maintaining bioinformatics resources for multiple platforms, we have exploited Galaxy and the Galaxy Toolshed. The Toolshed² has developed into a comprehensive management interface for bioinformatics tools, with the ability to install the underlying tool dependencies. More recently, Data Managers have been added to the Toolshed, allowing management of reference data and genome indices through Galaxy.

We have implemented a set of scripts which, in part:

create environment modules³ for bioinformatics tools that have been installed through the Toolshed. This approach allows access to multiple versions of a tool,

give the ability to mount Galaxy Datasets as appropriately-named files via FUSE, for direct read-only access from the command line, RStudio, or IPython Notebook,

provide convenient symlinks to Galaxy reference genomes and indices.

In addition, the BioBlend library⁴ is installed into all GVL machine images, providing programmatic access to the Galaxy workflow engine.

These scripts are run as part of the setup of a GVL instance. They are implemented as an Ansible playbook, allowing them to be easily adapted to other Galaxy servers.

Notes:

¹ GVL project website: <http://genome.edu.au/>

² Blankenberg et al. (2014) Dissemination of scientific software with Galaxy ToolShed. *Genome Biology* 15: 403

³ Environment modules

website: <http://modules.sourceforge.net/> ; Furlani, J.L. : Modules: Providing a Flexible User Environment, Proceedings of the Fifth Large Installation Systems Administration Conference (LISA V), pp. 141-152, San Diego, CA, September 30 – October 3, 1991

⁴ Sloggett, C., Goonasekera, N., and Afgan, E. (2013) BioBlend: automating pipeline analyses within Galaxy and CloudMan. *Bioinformatics* 29: 1685-1686

Integrating Galaxy and Tripal: Cyberinfrastructure for the Genome Community Database

Emily Grau¹, Connor Wytko², Brian Soto², Sook Jung², Kuangching Wang³, Nick Watts⁴, Margaret Staton⁵, Doreen Main², Jill Wegrzyn¹, F. Alex Feltus⁶, Stephen P. Ficklin²

¹ University of Connecticut Department of Ecology and Evolutionary Biology, Storrs, CT 06269, USA

² Washington State University Department of Horticulture, Pullman, WA 99164, USA

³ Clemson University Department of Electrical & Computer Engineering, Clemson, SC, 29634, USA

⁴ Clemson University, Clemson Computing and Information Technology, Anderson, SC 29625 USA

⁵ University of Tennessee Institute of Agriculture Department of Entomology and Plant Pathology, Knoxville, TN 37996, USA

⁶ Clemson University Department of Genetics & Biochemistry, Clemson, SC, 29634, USA

Model or clade organism databases (i.e. community research databases) enable both basic and applied research by offering curated data, visualization, and analytical tools. Tripal is a widely adopted, open-source toolkit for construction of online genomic and genetic databases. Tripal combines the power of Chado, an open-source database schema and Drupal, an open-source content management system, to facilitate construction of genomic and genetic websites while allowing complete customization. Advances in sequencing technology create new opportunities and challenges in genomics research for all organisms. Access, sharing, and analysis of these large data sets is hindered by transfer speeds, incompatible file formats, and insufficient metadata. The NSF DIBBs-funded Tripal Gateway project (ACI-1443040) is aimed at addressing these issues through development of three new Tripal modules that 1) improves data transfer by exploring software defined networking technologies (Tripal SDN module); 2) provides a RESTful web service framework with the goal of cross-database querying (Tripal Exchange module); and 3) integrates with Galaxy workflows to seamlessly provide commonly used analytical workflows to site patrons (Tripal Galaxy module). Development of the Tripal Galaxy module will include the creation of PHP bindings for the Galaxy API (usable outside of Tripal), integration of Galaxy workflows into Tripal, and coordination of data transfer for use in workflows to computational facilities and back to the community database. The Tripal Gateway project will be implemented for the legume, grains, cotton and tree crop communities but will be available for use by any Tripal site.

Session 7: 13:30 Wednesday 8 July

Creating dynamic tools with Galaxy ProTo

Morten Johansen¹, Sveinung Gundersen², Abdulrahman Azab², Eivind Hovig¹, Geir Kjetil Sandve¹

¹ Oslo University Hospital

² University of Oslo

Creating a Galaxy tool is not straightforward and has limitations. One has to write a XML file defining the inputs and outputs of a tool. This is practical when one has a predefined number of input fields with static options, but becomes complex when the options can change dynamically, and even impossible if the number of input fields can change (e.g. depending on what the user selected in a previous selection box).

The Galaxy Prototyping Tool API (Galaxy ProTo) is a new tool building methodology, introduced by the Genomic HyperBrowser project. Galaxy ProTo is an unofficial alternative for defining Galaxy tools. Instead of XML files, Galaxy ProTo supports defining the user interface of a tool as a Python class. Each input box is defined in a method that provides a high level of dynamicity. For instance one could read the beginning of an input file and provide dynamic options based on the file contents.

Beyond Galaxy: portable workflows and tool definitions with the CWL

Peter Amstutz¹, Nebojša Tijanić², Stian Soiland-Reyes³, John Kern⁴, Luka Stojanovic², Tim Pierce¹, John Chilton⁵, Maxim Mikheev⁶, Samuel Lampa⁷, Hervé Ménager⁸, Scott Frazer⁹, Venkat S. Malladi¹⁰, Michael R. Crusoe¹¹

¹ Curoverse Inc.

² Seven Bridges Genomics, Inc.

³ University of Manchester, School of Computer Science

⁴ AccuraGen Inc.

⁵ Penn State University, The Galaxy Project

⁶ BioDatomics LLC.

⁷ Uppsala University, Department of Pharmaceutical Biosciences; BILS (Bioinformatics Infrastructure for Life Sciences)

⁸ Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France

⁹ The Broad Institute

¹⁰ Stanford University

¹¹ University of California, Davis; School of Veterinary Medicine; Lab for Data Intensive Biology

With Galaxy one gets all the benefits of bioinformatics workflow platforms: provenance tracking, execution and data management, repeatability, and an environment for data exploration and visualization. But what are the options when we want to move to another platform?

To address this four engineers started working together at the BOSC 2014 Codefest with an initial focus on developing a portable means of representing, sharing, and invoking command line tools and a secondary focus on portable workflow descriptions.

On March 31st, 2015 the group released their second draft of the Common Workflow Language specification. Descriptions are a YAML document: validated by an Apache Avro schema and can be interpreted as an RDF graph using JSON-LD. The documents are also valid Wf4Ever 'wfdesc' descriptions after a simple transformation. Future drafts will include the use of the EDAM ontology to describe the tools enabling discovery via the ELIXIR tool registry.

Seven Bridges Genomics, the Galaxy Project, and the organization behind Arvados (Curoverse) have started to implement support for the Common Workflow Language, with interest from other projects and organizations like Apache Taverna, BioDatomics and the Broad Institute. Developers on the Galaxy Team are exploring adding CWL tool description support with plans to add support for the CWL workflow descriptions. Tool authors and other community members will benefit as they will only have to describe their tool and workflow interfaces once. This will enable scientists, researchers and other analysts to share their workflows and pipelines in an interoperable and yet human readable manner.

Extending Galaxy's reach: recent progress towards complete multi-omic data analysis workflows

Timothy J Griffin¹, James Johnson¹, Getiria Onsongo¹, Pratik D Jagtap¹, Candace R Guerrero¹, Kevin Murray¹, Ira Cooke², Bjoern Gruening³, Lennart Martens⁴, Marc Vaudel⁵, Harald Barsnes⁵

¹ University of Minnesota, USA

² La Trobe University, AUSTRALIA,

³ University of Freiburg, GERMANY;

⁴ Ghent University, BELGIUM;

⁵ University of Bergen, NORWAY

Integrative analysis of different 'omic data types, also known as multi-omics, is gaining momentum as a powerful biological discovery tool. Galaxy offers an ideal platform for these types of data analysis applications, which require sophisticated workflow development utilizing disparate tools for different data types (e.g. genomic, transcriptomic, proteomic data). Here, we will present recent progress from our global research team in this area, focusing on proteogenomic applications. Proteogenomics utilizes genomic and/or transcriptomic data as a template to translate in-silico possible encoded protein products, including novel sequences arising from genomic variation (splice isoforms, mutations, frameshifts etc). Mass spectrometry (MS)-based proteomics data is matched against these protein sequences, confirming known protein products, as well as novel sequences. We have built a unique Galaxy-based workflow offering complete proteogenomic analysis. Our workflow utilizes well-known Galaxy tools for working with transcriptomic and genomic data to identify potentially novel protein coding sequences such as splice isoforms and non-synonymous indels (e.g. TopHat, SamTools, SNPeff). We are developing the powerful SearchGUI/PeptideShaker platform, implemented in Galaxy, to match proteomics data to the generated protein sequences. This platform enables combined use of several proteomic database searching algorithms to provide more confident matches of data to novel protein sequences, and flexible outputs for further downstream analysis and evaluation to ensure high confident reporting of novel protein sequences. Finally, the results are compatible with visualization and interpretation using the popular Integrated Genome Viewer. We will demonstrate the use of this powerful proteogenomic workflow in the analysis of several biologically-relevant datasets.

A Genomics Virtual Laboratory in practice

Enis Afgan¹, Clare Sloggett², Nuwan Goonasekera², Igor Manukin³, Derek Benson³, Mark Crowe⁴, Simon Gladman², Yousef Kowsar², Michael Pheasant³, Ron Horst³, Andrew Lonie²

¹ John Hopkins University, USA

² University of Melbourne, Australia

³ University of Queensland, Australia

⁴ Queensland Facility for Advanced Bioinformatics, Australia

Over the last 4 years we have designed and implemented the Genomics Virtual Laboratory (GVL:<http://genome.edu.au>) as a middleware layer of machine images, cloud management tools, and online services that enable researchers to build arbitrary sized Galaxy compute clusters on demand, pre-populated with fully configured bioinformatics tools, reference datasets and workflow and visualisation options. The platform is flexible in that users can conduct analyses through multiple web-based (Galaxy, RStudio, IPython Notebook) or command-line interfaces, and add/remove compute nodes and data resources as required. Best practice tutorials and protocols provide a path from introductory training to practice. The GVL is available on the

OpenStack-based Australian Research Cloud (<http://nectar.org.au>) and the Amazon Web Services cloud via a dedicated web-based launcher application (<http://launch.genome.edu.au>).

We now have GVL implementations at major Australian research institutes including the Universities of Queensland, Melbourne, Monash and Western Australia, and the Peter MacCallum Cancer Centre; plus many hundreds of individual launches by researchers and students across the country. We have learned a great deal about the usage patterns of the platform, including scalability, reliability, and accessibility. This presentation will discuss progress on the GVL project, lessons learned in architecting for the cloud, and uptake and usage by the Australian research community.

The Genomics Virtual Laboratory project is funded by the federal NeCTAR and ANDS programs.

IRIDA: A Genomic Epidemiology Platform Built on top of Galaxy

Aaron Petkau¹, Franklin Bristow¹, Thomas Matthews¹, Josh Adam¹, Philip Mabon¹, Eric Enns¹, Jennifer Cabral^{1,2}, Joel Thiessen^{1,2}, Cameron Sieffert¹, Natalie Knox¹, Damion Dooley³, Emma Griffiths⁵, Geoff Winsor⁵, Matthew Laird⁵, Mélanie Courtot^{3,5}, Peter Kruczkiewicz⁶, Alex Keddy⁷, Robert G. Beiko⁷, William Hsiao^{3,4}, Gary Van Domselaar^{1,2}, Fiona Brinkman⁵

¹ National Microbiology Laboratory, Winnipeg, Canada

² University of Manitoba, Winnipeg, Canada

³ BC Public Health Microbiology and Reference Laboratory, Vancouver, Canada

⁴ University of British Columbia, Vancouver, Canada

⁵ Simon Fraser University, Burnaby, Canada

⁶ Laboratory for Foodborne Zoonoses, Lethbridge, Canada

⁷ Dalhousie University, Halifax, Canada

Whole genome sequencing (WGS) is revolutionizing epidemiological methods for identification and investigation of infectious disease outbreaks. However, the routine use of WGS has been hindered due to the complexity in data management and the lack of pipelines supporting quality control and data analysis standards. While an increasing number of pipelines for genomic epidemiology are being developed, each typically has different installation and execution requirements. This leads to a difficulty in the integration of these pipelines into a single genomic epidemiology system.

Galaxy offers a solution by providing a system to integrate, execute, and maintain data analysis pipelines. In addition, Galaxy provides a community of developers who contribute and maintain the bioinformatics tools used for genomic epidemiology. Our project, IRIDA (Integrated Rapid Infectious Disease Analysis), builds on top of Galaxy a platform for genomic epidemiology. IRIDA provides a system for the storage and management of sequencing data and sample metadata, an interface for the execution of data analysis pipelines, and the storage, auditing and visualization of results. Within IRIDA, we provide standard pipelines for genomic epidemiology including SNVPhyl, our SNV (Single Nucleotide Variant) phylogeny pipeline. These pipelines are executed using a Galaxy instance internal to IRIDA and additional support is provided for exporting genomic sequence data to external Galaxy instances.

By building on top of Galaxy we hope to simplify the process of pipeline integration, to share our pipelines with the bioinformatics community, and to contribute to the development of standards for genomic epidemiology. More information can be found at <http://irida.ca>.

Session 8: 16:20 Wednesday 8 July

Building Galaxy Community VM

Ryota Yamanaka¹, Tazro Ohta², Manami Kato³, Hiroyuki Aburatani¹

¹ Genome Science Division, The University of Tokyo

² Database Center for Life Science, ROIS

³ Laboratory for Disease Systems Modeling, IMS, RIKEN

For biomedical researchers, there are two barriers when they start using Galaxy. First, while Galaxy tools and workflows are shared in public repositories, new users can hardly get the information on how other research institutes use those tools and design workflows. Second, they may often not be able to reproduce the workflows they used before, since it is difficult for individual researchers or small laboratories to maintain their systems, so their Galaxy environments get often unrecoverable when they change the settings or reset their computers. To solve these problems, Galaxy Community Japan holds a monthly meet-up to share our workflows. We also distribute a virtual machine image, on which we configured Galaxy with necessary tools and workflows, and make available our practical know-how about these tools and workflows on our website. Users can download and run the virtual machine on their own PC or launch it on AWS, so they can immediately try pre-installed analysis workflows with their own data. The latest version of this virtual machine is running on our public test site, while the older versions are kept downloadable too. As a result, users can run the same workflows on different computational infrastructures and always reconstruct the

Galaxy environments they have used before. This will also help developers advertise their new tools to potential users. We would like to introduce several newly developed unique tools on our Galaxy, as well as our experiences in the local activities such as Galaxy Workshop Tokyo.

An initiative to federate the galactic community in France: the IFB Galaxy Working Group

Gwendoline ANDRES¹, Loraine BRILLET-GUEGUEN¹, Christophe CARON¹, Alexis DEREPPER², Sandra DEROZIER³, Olivia DOPPELT-AZEROUAL⁴, Jean François DUFAYARD⁵, Franck GIACOMONI⁶, Olivier INIZAN⁷, Gildas LE CORGUILLE¹, Alban LERMINE⁸, Valentin LOUX³, Sarah MAMAN⁹, Fabien MAREUIL⁴, Mishari MONSOOR¹

¹ CNRS-UPMC Station Biologique de Roscoff

² IRD Southgreen Montpellier

³ INRA MaIAGE Jouy en Josas

⁴ Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France

⁵ CIRAD Southgreen Montpellier

⁶ INRA PFEM Clermont Ferrand

⁷ INRA URGI Versailles

⁸ Institut Curie Paris

⁹ INRA Genotoul/SIGENAE Toulouse

As the Galaxy “tour de france” showed in 2012, the Galaxy platform aroused a great interest in France. Today this platform meet a great success throughout the different bioinformatics infrastructures in the country. In March 2013 a working group dedicated to Galaxy and supported by national infrastructure “Institut Français de Bioinformatique” (French Institute of Bioinformatics) has been set up: IFB Galaxy Working Group (IFB GWG). This Working Group has been built upon several national and regional bioinformatics platforms which use and deploy Galaxy (for training sessions, for analysis, ...). The IFB has given to the Working Group the mission to federate the french Galaxy

community (biologist and bioinformaticians). This was based on three main actions: animation, training and technology. After two years of activity we would like to present the more significant results in the animation of the french community: events (Galaxy Days), training sessions (Galaxy4Bioinformatics), development (Toolshed, Best Practice Guides). We will also present how we are using Galaxy as a Hub to build some federating projects (IFB/France Genomique & IFB/MetaboHUB) between different communities addressing scientific/technological challenges.

Accepted Posters

Odd-numbered posters will be presented Tuesday from 15:00 to 16:20, and even-numbered posters will be presented on Wednesday from 15:00 to 16:20.

Posters are size A0, 1189 mm wide by 841 mm tall, and will be hung with push-pins.

P01: Towards Bioinformatics for All: Galaxy at UoM

Peter Briggs¹, Ian Donaldson¹, Sarah Griffiths¹, Leo Zeef¹

¹ University of Manchester

As part of the Bioinformatics Core Facility (BCF) at the University of Manchester (UoM) we have developed a number of bespoke Galaxy tools to support local researchers conducting next generation sequencing (NGS) analyses. The tools are accessible via a private local Galaxy instance maintained by the BCF, but are also available to the wider Galaxy community via the public Galaxy toolshed.

In collaboration with researchers we were able to help improve the detection of microsatellites by implementing Trimmomatic and PALfinder as Galaxy tools. This now allows non-bioinformaticians to analyse their own data, circumventing installation and use of command line programs. Additionally we have developed a set of ChIP-seq analysis tools (Trimmomatic, MACS2, CEAS, Weeder2, RnaChIPIntegrator) that allows our users to further explore their data after it has left the BCF. The tools also provide a framework for tutorials about ChIP-seq analysis.

Our ongoing aim is to maintain and develop a local Galaxy instance that provides researchers with the means to run bioinformatics tools that they would not otherwise be able to use, and provide a means of easily rerunning analyses.

P02: The Galaxy framework as a concept for a national system for monitoring and surveillance of infectious disease

Arnold Knijn¹, Massimiliano Orsini², Valeria Michelacci¹, Stefano Morabito¹

¹ Istituto Superiore di Sanità, Rome, Italy

² Istituto Zooprofilattico Sperimentale dell'Abruzzo e Molise, Teramo, Italy

A proposal has been submitted to a national call of the Italian Ministry of health concerning the creation of a National Information System for the collection of genomic data in the field of veterinary public health, with the aim of deploying a state of the art molecular epidemiology approach to the surveillance of food-borne zoonoses and infectious diseases at the human and animal interface. The concept described revolves around the creation of a nationwide distributed cluster of pathogen-specific databases of NGS and epidemiological data hosted on servers present at each of the participating institutes. A common framework for the comparison

of such data will complete the system with the aim of detecting clusters of cases as well as to provide convincing evidence to link cases of disease and sources of infection. The databases will be replicated on each server constituting the DB-cluster. The redundancy originated by the replication process is meant to guarantee a distributed access to the Information System, a high availability of the data hosted, a geographically distributed disaster recovery capability and to enable load-balancing of queries at each node, increasing the performance of access to the analytical pipelines in case of heavy traffic on any of the servers. All the nodes of the network will use the same Information System implemented into the open source framework Galaxy.

P03: Gene identifier matching to join publicly available databases for the generation of a Mammalian Ortholog and Annotation Database with access from Galaxy-server

Jochen Bick¹, Mark Robinson², Susanne E. Ulbrich¹, Stefan Bauersachs¹

¹ Animal Physiology at ETH Zurich

² Institute of Molecular Life Sciences at University of Zurich

So far there is a number of well-organized databases that contain useful information regarding orthologous genes, e.g., EnsemblCompara ortholog database (EcoDb). The main problem when using information derived from different databases is to correctly assign different gene, transcript or protein identifiers. For example, EcoDb does not provide NCBI EntrezGene identifiers and the assignment available in BioMart is incomplete and contains errors. However, because NCBI annotation is for most species the most comprehensive, we need to map information from other databases to EntrezGene IDs. This is an important issue for the generation of a Mammalian Ortholog and Annotation Database (MOA-DB) which will be partially based on information from publicly available databases, which needs to be collected, analyzed, and connected. Since each public source database uses own unique identifiers, it is necessary to assign the corresponding database-specific identifiers. Existing lists that assign corresponding genes, e.g. between Ensembl and EntrezGene are incomplete and/or contain errors. Therefore, missing information needs to be calculated and duplicates need to be handled. R BioConductor packages were used to find overlapping gene and exon positions which were integrated as a lookup table into the MySQL database to handle the comparison of different database sources. Finally, this database will be integrated into our local Galaxy-server to give easy access to all our research groups and provide a useful interface

with various options to parse information via SQL queries. The MOA-DB provides a basis for optimal across-species comparisons of transcriptome datasets from different mammalian species accessible within a Galaxy-server.

P04: GenAP: A platform to provide Biomedical tools throughout Canadian HPCS

David Anderson de Lima Morais¹, Michel Barrette¹, David Bujold², Carol Gauthier¹, Kuang Chung Chen², Simon Nderitu², Maxime Levesque¹, Bryan Caron², Alain Veilleux¹, Pierre-Etienne Jacques¹, Guillaume Bourque²

¹ Université de Sherbrooke, Sherbrooke, Quebec, Canada

² McGill University, Montreal, Quebec, Canada

The Genetics and Genomics Analysis Platform (GenAP) is a computing platform for life sciences researchers. GenAP offers three components: a web portal from which users have access to tools (UCSC browser) and platforms (Galaxy); bioinformatics software and libraries, distributed via CERN Virtual Machine File System (CVMFS); and bioinformatics software pipelines.

In Galaxy-GenAP, we use a hybrid system involving cloud images on an HPC facility, to provide private Galaxy instances to our users. These private instances are only available to a project Principal Investigator (PI), his group members, and any external member that the PI chooses to add. GenAP is fully integrated with Compute Canada (CC) and all Galaxy jobs are computed toward the users' CC resource allocation in any HPC cluster.

GenAP was designed to be portable to any HPC center in Canada and in our second phase we will increase the number of hosts. To facilitate the installation of the platform we are currently integrating Galaxy and the CERN Virtual Machine File System (CVMFS). In this case Galaxy will be installed on the main CVMFS repository (stratum 0) and any HPC facility running a mirror client (stratum 1) will receive the Galaxy code, tools and updates automatically.

Through GenAP, Galaxy has been integrated to curricular courses at McGill and Sherbrooke Universities and is a fundamental part of several workshops. We aim to have GenAP and Galaxy integrated in most major HPC centers.

P05: Galaxy in teaching computational methods of genome analysis for master degree students in Medical Genetics program at the Faculty of Medicine, Vilnius University

Erinija Pranckeviciene¹, Laima Ambrozaityte¹, Ingrida Uktveryte¹, Algirdas Utkus¹, Vaidutis Kucinskas¹

¹ Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University

Master program in Medical Genetics is offered by the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University. In this program a computational analysis of genomic data constitutes a considerable part of practical exercises. In "Genome analysis" seminar and "Biotechnology and fundamentals of bioinformatics analysis" course students are introduced to a computational pipeline of next generation sequencing (NGS) data analysis starting by quality assessment of raw exome sequencing data and ending by interpretation of the identified genomic variants. In class students use data from scientific articles and have to reproduce some of the published results.

For these courses Galaxy runs on a Hardware-as-a-Service server (2 Hexa core Intel Xeon CPU E5-2630L, 8 processing units, 8192 Mb RAM and 320 Gb disk space).

Using Galaxy in teaching and learning is novel approach at the Department of Human and Medical Genetics. Noted benefit of this approach is that students without previous exposure to bioinformatics are efficiently grasping complex concepts and share "know how" of tools. A little effort is needed to get used to Galaxy interface, its visualization capabilities. Practical computations are

evaluated directly in students named histories in a process of step by step inspection. Benefits of using Galaxy in teaching at the end of the course will be evaluated by qualitative analysis (interview of students).

P06: Galaxy in Public Health: the Microbial Genomics Virtual Laboratory

Simon Gladman¹, Nuwan Goonasekera¹, Clare Sloggett¹, Dieter Bulach¹, Torsten Seemann¹, Andrew Lonie¹

¹ VLSCI, University of Melbourne, Australia,

The uptake of genomics in public health and clinical microbiology laboratories is being slowed by the perceived requirement that each laboratory needs to, counterproductively, establish and evaluate their own tools and infrastructure which will result in a lack of standardisation of methods.

An easily instantiated computer image based around Galaxy with a defined set of microbial-specific tools and reference data is an ideal solution for enabling standardisation between laboratories. We have established the Genomics Virtual Laboratory [GVL: <http://genome.edu.au>] to empower laboratories to establish their own private operating environment to securely analyse their own data using software and analysis methods that are widely used for microbial genomics in a reproducible manner suited to government accreditation

The GVL consists of a set of machine images for performing genomics analyses in a scalable, reproducible manner, plus web tools for instantiating and managing the images on multiple cloud architectures. The images incorporate a number of pre-configured genomic analyses platforms including Galaxy, the Linux command line, RStudio and IPython Notebook.

The GVL images are constructed from Ansible scripts which make it straightforward to customise. Here we present a flavour of the GVL fully tailored to microbial genomics (MGVL) by incorporating various microbial analysis pipelines and tools for both the Galaxy environment and the command line.

The Genomics Virtual Laboratory project is funded by the federal NeCTAR and ANDS programs (<http://nectar.org.au>; <http://ands.org.au>).

P07: 16S rDNA amplicon sequencing data analysis in Galaxy

Loïc Bourgeois¹, Amalia Soenens¹, Nuria Lozano¹, Juan Imperial¹

¹ Centro de Biotecnología y Genómica de Plantas (CBGP), Universidad Politécnica de Madrid, Campus de Montegancedo, 28223 Pozuelo de Alarcón, Madrid, Spain

Most biologists can easily access NGS technologies and data in order to characterize the microbial diversity of a sample with 16S rDNA amplicon sequencing. However, the output of this kind of experiment can be challenging to handle. We assessed the different options to address 16S rDNA amplicon data analysis in Galaxy, and will highlight the benefits and drawbacks of the existing solutions. Indeed, even if the bioinformatics community now provides numerous tools allowing treatment of this sort of data, determining which software best fits the user's needs is not trivial for several reasons. To begin with, some of this software is not easy to install, which can be a first barrier. In line with this, most tools do not provide any GUI, which can be tedious for people not used to the UNIX environment. Finally, a critical point is that even if the available software usually provide similar core steps to perform the analysis of 16S rDNA amplicon data, they do not always use the same approaches. Moreover there are a lot of different algorithms that can be used for each step of the analysis. The choice of the software and the algorithms one should use is important, as it will impact the output of the experiment and relies on the characteristics of the data and the user experience. Galaxy can handle the tool installation and GUI barriers on top of other

intrinsic benefits of using Galaxy, which allows users to focus on the data analysis itself.

P08: A Galaxy approach to microbial data integration: the USMI Galaxy Demonstrator

Daniele Pierpaolo Colobraro¹, Paolo Romano¹

¹ IRCCS AOU San Martino IST

Many application domains, such as health, food, energy and waste management, exploit research on micro-organisms, which information is distributed in many heterogeneous repositories.

The Microbial Resource Research Infrastructure (MIRRI) aims to orchestrate European microBiological Resource Centers (mBRCs) with the goal of providing improved and extended services and integrated access to data. In this context, the aims are i) integrating the information on microorganisms, ii) assessing available information, iii) pointing out discrepancies, errors and gaps, iv) carrying out in-silico analyses, and v) curating mBRC catalogues' data.

USMI Galaxy Demonstrator, which is under active development, is available at <http://galaxy.nettab.org:8088/>. All tools are written in Python.

The tools menu includes a section devoted to MIRRI tools, where three categories are shown, related to retrieval of data from MIRRI catalogues, extension of catalogues contents with data from external resources, and data integration applications.

Tools of the "data_source" type are available for importing both full catalogues and single strain data in Galaxy. Information is archived by using an extended version of the Microbiological Common Language (MCL, <http://www.straininfo.net/projects/mcl/reference>).

The external data sources that have been already taken into account are NCBI Taxonomy, BRENDA, Pubmed, UNIPROT and ENA, which are respectively queried in order to retrieve information on taxon identifiers, EC numbers, Pubmed identifiers and DOIs, UniProt identifiers, and rRNA sequences. These are linked by using either the strain numbers, or the enzyme and species names, or the bibliographic references.

Outputs are provided in tabular form, allowing both for human and machine readable.

P09: A Galaxy metagenomic workflow for reference-tree based phylogenetic placement (MG-RTPP)

Ambrose Andongabo^{1*}, Ian M. Clark^{1*}, Dariush Rowlands¹, Keywan Hassani-Pak¹, Penny R. Hirsch¹, Elisa Izoa¹, Andy Neal^{1*}

¹ Rothamsted Research, Harpenden, United Kingdom

* Contributed equally

Background: High-throughput sequencing of environmental nucleic acids is revolutionizing and dramatically expanding our understanding of the diversity and functionality of complex microbial communities. There are a number of tools which allow community structure to be surveyed using metagenomics or meta-transcriptomics at the rRNA level, or by using COG- or KEGG-based functional assignments. However, there are limited complementary approaches to investigate the phylogenetic diversity of functionally important individual genes in large sequence databases.

Results: We have designed a workflow for reference-tree based phylogenetic placement (MG-RTPP) of metagenomics and meta-transcriptomics samples. The inputs to the workflow are unassembled reads, a multiple sequence alignment (MSA) of the genes of interest and large public sequence databases. Reference nucleotide profile hidden Markov models (pHMMs) are built from the MSA and are used as queries. Homologous reads are checked for accuracy before being placed on a reference phylogenetic tree, maximising phylogenetic likelihood. The workflow retains considerable flexibility, allowing for tuning of redundancy in the

nucleotide pHMMs used as queries to recover as many true hits as possible.

Conclusions: MG-RTPP facilitates fast interrogation of sequence databases in a flexible and robust fashion. It avoids misidentification of false positives while pHMM tuning allows for maximum recovery of sequences. Phylogenetic placement provides unique visualization approaches which reveal the phylogenetic relationships between environment-derived sequences and sequenced organisms and between samples. The approach compliments tools such as QIIME, MG-RAST and MEGAN in allowing interrogation of individual gene abundance and diversity in samples. Keywords: metagenome, metatranscriptome, assembly-free, community analysis, functional genes, phylogeny.

P10: IRIDA: A Genomic Epidemiology Platform Built on top of Galaxy

Aaron Petkau¹, Franklin Bristow¹, Thomas Matthews¹, Josh Adam¹, Philip Mabon¹, Eric Enns¹, Jennifer Cabral^{1,2}, Joel Thiessen^{1,2}, Cameron Sieffert¹, Natalie Knox¹, Damion Dooley³, Emma Griffiths⁵, Geoff Winsor⁵, Matthew Laird⁵, Mélanie Courtot^{3,5}, Peter Kruczkiewicz⁶, Alex Keddy⁷, Robert G. Beiko⁷, William Hsiao^{3,4}, Gary Van Domselaar^{1,2}, Fiona Brinkman⁵

¹National Microbiology Laboratory, Winnipeg, Canada

²University of Manitoba, Winnipeg, Canada

³BC Public Health Microbiology and Reference Laboratory, Vancouver, Canada

⁴University of British Columbia, Vancouver, Canada

⁵Simon Fraser University, Burnaby, Canada

⁶Laboratory for Foodborne Zoonoses, Lethbridge, Canada

⁷Dalhousie University, Halifax, Canada

Whole genome sequencing (WGS) is revolutionizing epidemiological methods for identification and investigation of infectious disease outbreaks. However, the routine use of WGS has been hindered due to the complexity in data management and the lack of pipelines supporting quality control and data analysis standards. While an increasing number of pipelines for genomic epidemiology are being developed, each typically has different installation and execution requirements. This leads to a difficulty in the integration of these pipelines into a single genomic epidemiology system.

Galaxy offers a solution by providing a system to integrate, execute, and maintain data analysis pipelines. In addition, Galaxy provides a community of developers who contribute and maintain the bioinformatics tools used for genomic epidemiology. Our project, IRIDA (Integrated Rapid Infectious Disease Analysis), builds on top of Galaxy a platform for genomic epidemiology. IRIDA provides a system for the storage and management of sequencing data and sample metadata, an interface for the execution of data analysis pipelines, and the storage, auditing and visualization of results. Within IRIDA, we provide standard pipelines for genomic epidemiology including SNVPhyl, our SNV (Single Nucleotide Variant) phylogeny pipeline. These pipelines are executed using a Galaxy instance internal to IRIDA and additional support is provided for exporting genomic sequence data to external Galaxy instances.

By building on top of Galaxy we hope to simplify the process of pipeline integration, to share our pipelines with the bioinformatics community, and to contribute to the development of standards for genomic epidemiology. More information can be found at <http://irida.ca>.

PI1: Galaxy – a platform for teaching the analysis and interpretation of clinical NGS data

Ang Davies¹, Jan Taylor², Mike Cornell¹, Peter Briggs¹, Sanjeev Bhaskar³ Andy Brass¹

¹ The University of Manchester

² St James' Hospital Leeds, The University of Manchester

³ St Mary's Hospital Manchester

The University of Manchester delivers a masters programme in Clinical Bioinformatics which provides the education for trainee clinical bioinformaticians on the NHS Scientist Training Programme, training to become registered healthcare scientists. This programme is lead under the direction of Manchester Academy for Healthcare Scientist Education (MAHSE). Clinical bioinformaticians within the NHS are at the forefront of genomic medicine in their roles, often responsible for building and validating bioinformatic workflows that are used in the analysis and interpretation of clinical Next Generation Sequencing (NGS) data. Within the diagnostic genomic medical centres across the UK bioinformaticians are building and using NGS pipelines to analyse sequencing data from gene panels, whole exomes and now whole genomes for those centres involved in the 100000 Genomes Project. Within the masters programme we used Galaxy to teach the trainees how to analyse anonymised clinical NGS gene panel data, kindly provided by the Manchester Centre for Genomic Medicine for teaching purposes. The pipeline the trainees built included quality control, alignment, annotation, interpretation and viewing on a genome browser, enabling trainees to identify a potential causal pathogenic variant from the original Fastq file. The analysis was undertaken on a local installation of Galaxy configured by the Bioinformatics Core Facility at the university. For more information contact angela.davies@manchester.ac.uk.

PI2: Bioinformatics Evolving at Canada's National Microbiology Laboratory

Eric Enns¹, Philip Mabon¹, Jennifer Cabral^{1,2}, Mariam Iskander^{1,2}, Cameron Sieffert¹, Natalie Knox¹, Heather Kent¹, Shane Thiessen¹, Paul Williams¹, Brian Yeo², Joel Thiessen^{1,2}, Josh Adam¹, Aaron Petkau¹, Thomas Matthews¹, Franklin Bristow¹, Gary Van Domselaar^{1,2}

¹ National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada

² Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada

The National Microbiology Laboratory (NML) is Canada's leading public health laboratory, responsible for the identification, control and prevention of infectious diseases. The bioinformatics core facility at the NML deployed our first instance of Galaxy in 2010. The introduction of the Galaxy platform has revolutionized bioinformatics at the NML by bridging the gap between bioinformaticians and biologists.

Prior to Galaxy, most of our in-house tools and pipelines required an extensive background in UNIX command line and high performance computing to operate. This requirement demanded that bioinformaticians be intimately involved in projects with significant computational requirements. Galaxy was selected to be the bioinformatics analysis platform at the NML, as it made our tools and pipelines accessible to biologists. Bioinformaticians are able to focus more time on tool and pipeline development, as their project involvement has been reduced. Biologists are able to perform analyses on their own as Galaxy lowers the barrier to carrying out complex analysis in a high performance computing environment. As a result NGS (Next-generation sequencing) projects are able to progress at a much faster rate. Moving forward, we are developing a Galaxy-powered infectious disease analysis platform for our standardized analyses while retaining our traditional Galaxy environment for ad hoc pipeline development and bioinformatics analysis.

PI3: Galaxy Flavours – your highly portable, configurable local Galaxy distributions with preinstalled workflows – for Linux, MacOSX and Windows

Christian Rausch¹, Jeroen Galle², Stef van Lieshout¹, Wim van Crielinge², Björn Grüning³

¹ Cancer Center Amsterdam, VU University Medical Center, Amsterdam, The Netherlands

² Biobix, Lab of Bioinformatics and Computational Genomics, Ghent University, Ghent, Belgium

³ Chair of Bioinformatics, University of Freiburg, Germany

Galaxy makes it easy for biologists to use advanced bioinformatics software through graphical web-browser-based user interfaces. However, when using one of the public Galaxy servers like at usegalaxy.org is not an option (e.g. in the case of sensitive data), setting-up a local Galaxy installation still requires Linux administrator skills.

Therefore we are developing installers for the Linux, Macintosh and Windows operating systems that make use of portable Docker software containers.

Another aspect that makes the usage of Galaxy actually increasingly difficult especially for the less advanced user is the growing number of available tools – how to find the right tool for a given task? Here we want to help by providing a useful selection of tools and workflows for typical problems in biomedical data analysis, preconfigured in the Galaxy Flavours we provide.

On the poster we present the current status of our work, future plans and further ideas like e.g. a configurator for tailored Galaxy Docker images. Please join the discussion on the prioritisation of future Galaxy developments at our poster and at the conference in general.

PI4: GIO: Standards-compliant Galaxy workflows for proteomics informed by transcriptomics

Jun Fan¹, Shyamasree Saha¹, Adelyne Sue Li Chan¹, David A Matthews², Conrad Bessant¹

¹ School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS. UK

² School of Cellular and Molecular Medicine, University of Bristol, University Walk, Bristol. BS8 1TD. UK.

The most common method of identifying proteins in a complex sample is to perform liquid chromatography tandem mass spectrometry (LC-MS/MS) then search the acquired spectra against a reference proteome downloaded from a database such as UniProt. This approach has the major drawback of not being able to identify gene products that are not already known. The recently developed proteomics informed by transcriptomics (PIT) methodology tackles this problem by using RNA-seq to generate sample-specific protein databases that the LC-MS/MS data can be searched against. This allows the detection and quantitation of previously unknown proteins, protein variants and other exotic translated genomic elements. This is of particular utility when studying non-model organisms and samples with very dynamic proteomes, e.g. stem cells, cancer cells and virus-infected cells. The analysis of PIT data is complex and computationally intensive, requiring the integration of multiple third party tools from the proteomics, transcriptomics and genomics communities. To make this analysis tractable and repeatable we have produced GIO (Galaxy Integrated Omics) – a Galaxy-based framework containing the key tools and workflows needed to analyse data from PIT experiments in a reliable and repeatable way.

PI5: Reproducible Galaxy: Administration and Development

Aarif Mohamed Nazeer Batcha¹, Sebastian Schaaf, Guokun Zhang, Sandra Fischer, Ashok Varadharajan, Ulrich Mansmann

¹ Ludwig-Maximilians-University Munich

Establishing a structured IT infrastructure for processing NGS data is a challenge on multiple levels. To deal with such challenges in the field of molecular diagnostics and medical research, which often demands reproducibility, makes it much more interesting. An user-friendly, open source and modular galaxy framework was of great help, in facing those challenges although the reproducibility part was a bit questionable. Over three years of dedicated efforts, the Munich NGS-FabLab was build up as a running IT system, based on an assessment of requirements, constraints and given structural conditions. Aiming for a structured approach in resolving reproducibility issues and improving cross-connections between hospitals and research institutes associated with us, we came up with ansible-playbooks setup scripts to recreate our IT infrastructure. The scripts include setting up dedicated file servers, creating production, test and development environments, postgres database setup, apache configurations and grid engine for distributed management systems along with galaxy installation procedures. Although the playbooks were developed in SLES, blank unix systems with SSL connectivity and an infile is all that is necessary for the scripts to run. The scripts can be used to create and recreate an IT infrastructure and a reproducible environment for processing NGS data which is in high demand in medical research and diagnostics. We also hope to return our playbooks to the community that offered a great deal of support in developing our NGS-FabLab for processing medical sequencing data.

PI6: Mass spectrometry proteomics analysis with diverse tools for hundreds of runs

Jorrit Boekel^{1,2}, Rui Mamede-Branca², Henric Zazzi^{1,3}, Yafeng Zhu², Matthew The⁴, Lukas Käll⁴, and Janne Lehtiö²

¹ Bioinformatics Infrastructure for Life Sciences (BILS), Sweden

² Department of Oncology-Pathology, Science for Life Laboratory, Karolinska Institute, Stockholm, Sweden

³ PDC Center for High Performance Computing, Royal Institute of Technology – KTH, Stockholm, Sweden

⁴ School of Biotechnology, Science for Life Laboratory, Royal Institute of Technology – KTH, Stockholm, Sweden

The mass spectrometry (MS) field is currently undergoing rapid growth and is seeing an increasing amount of datasets per experiment. The growth is caused by sample size increase, meta-analyses and prefractionation, but analysis is constrained by MS computing environments which often lean towards proprietary software on Windows systems. Transition of typical analysis platforms to more powerful and flexible infrastructure is necessary to support availability of large scale analysis to users without access or in-depth knowledge to powerful bioinformatics tools and platforms.

We have combined a number of tools for spectra search (MSGF+), quantification (OpenMS, Hardklör/Krönik) and statistical scoring (Percolator) in the Galaxy framework. Since freely available MS tools do not always interact in all sought-after combinations, we have written software called mstitch to manipulate input and output files for a number of tools, including doing protein grouping and keeping an SQLite database of results. The resulting pipeline is under continuous development and can currently deliver data-repository-ready mzIdentML, and PSM, peptide and protein tables for end-users.

PI7: Read Between the Lines: Closing Gaps of Materials and Methods to Build Workflow from the Publication

Tazro Ohta¹, Osamu Ogasawara², Yoshinobu Masatani³, Shigetoshi Yokoyama³, Kento Aida³

¹ Database Center for Life Science

² DNA Data Bank of Japan

³ National Institute of Informatics

Publishing and sharing data analysis workflow using the galaxy platform has spectacularly reduced the cost of reproducing one's research, but following the description of data analysis which had been performed by other researchers to get the exact same result is still a big challenge. To evaluate the cost of data analysis workflow from the natural language description, we have performed to rebuild the workflow of CAGE sequencing data processing done by FANTOM5 team on the galaxy platform. Though the project has already published a set of papers with a lot of supplementary of methods and online protocols, it was not that straightforward to get the same result from the raw sequencing data available in the public data repository. The results processed by the rebuilt workflow are compared with the results published online by FANTOM5 team. This case study showed that some of the important information to rebuild the workflow is missing even in the well-described documents, for example, the location of the older source code, or the parameters for command execution. As the speed of biological data production increases, it will be more important to build the framework of cost-effective research reproducibility such as an automated evaluation process of published workflow. We will provide the details of our case study, and discuss how we can assure the reproducibility with the galaxy and other possible ways to perform, share, and publish the workflow as it is "executable materials and methods".

PI8: Galaxy-M: A galaxy workflow for processing and analysing direct infusion and liquid chromatography mass spectrometry-based metabolomics data

Riccardo Di Guida¹, Ralf J. M. Weber¹, Robert L. Davidson², Haoyu Liu¹, Archana Sharma-Oates¹, Warwick B. Dunn¹, Mark R. Viant¹

¹ School of Biosciences, University of Birmingham, Birmingham, B15 2TT, United Kingdom

² GigaScience, BGI-Hong Kong Co. Ltd, 16 Dai Fu Street, Tai Po Industrial Estate, NT, Hong Kong

Motivation: Metabolomics is increasingly recognised as an invaluable tool in the biological, medical and environmental sciences yet lags behind the methodological bioinformatics maturity of other 'omics fields, specifically genomics and transcriptomics. To achieve its full potential, standardisation and reproducibility of computational tools must be improved significantly. Here we report the development of Galaxy-M and describe further work to validate pre-processing methods for implementation in to Galaxy-M.

Development of Galaxy-M: We have developed an end-to-end mass spectrometry metabolomics pipeline in Galaxy for direct infusion mass spectrometry (DIMS) and liquid chromatography mass spectrometry (LC-MS) metabolomics. The range of tools presented spans from the processing of raw data, e.g. peak picking and alignment, and proceeds through data pre-processing to principal components analysis (PCA) and the associated statistical evaluation. To aid accessibility, the tools, Galaxy and data will all be provided via download. Additionally, source code, executables and installation instructions are available from Github.

Validation of pre-processing methods for LC-MS metabolomics: To provide a robust module for liquid chromatography-mass spectrometry (LC-MS) data pre-processing we have assessed the influence of different missing value imputation, normalisation, scaling and transformation methods on univariate and multivariate analysis. We show that normalisation by sum or PQN provides the

most robust results for univariate analysis while further KNN missing value imputation and log transformation are optimal for multivariate analysis. These methods are currently being implemented in to Galaxy-M.

P19: Integrating Galaxy in the Mr.SymBioMath Cloud Infrastructure

Óscar Torreño¹, Johan Karlsson², Alex Upton³, Michael Krieger¹, Oswaldo Trelles³

¹ RISC Software GmbH, 4232 Hagenberg, Austria

² Integromics S.L., 18100 Armilla Granada, Spain

³ University of Malaga, 29071 Malaga, Spain

Workflows are an increasingly important paradigm in bioinformatics and biomedicine; complex analyses are often performed by separate software packages that are later connected to form a complete pipeline. A number of workflows in both the bioinformatics and biomedicine domains are being developed in the Mr.SymBioMath project. GECKO¹, a biological sequence comparison workflow that studies the similarities between two or more genomes, and its post-processing steps, is the main development in the bioinformatics use case of the project. Genome wide association studies (GWAS) of SNPs² and Multi-SNPs

(epistatic interactions)³ are the principal implementations in the biomedicine use case. However, these workflows are command-line based, making their exploitation difficult for inexperienced users. Consequently, we have decided to use Galaxy in the project in order to facilitate their execution and distribution, whilst ensuring that all the experiments are reproducible. Our current architecture is comprised of 3 nodes deployed in a cloud infrastructure: 1) Gateway – which proxies the client requests to the web server; 2) Web server – which runs the galaxy web page contained in nginx; 3) DB server – which contains the meta-data queried from the galaxy web server. The present configuration executes the tasks in the second node, but we are currently working on the execution of the tasks in a separate Torque cluster which will be auto-scaled depending on the system load. The customised Mr.SymBioMath Galaxy configuration ensures that a wide spectrum of end users is able to obtain results as quickly and easily as possible.

Notes:

¹ Andres Rodriguez Moreno, Oscar Torreno Tirado, and Oswaldo Trelles Salazar. Out of core computation of hsp for large biological sequences. In *Advances in Computational Intelligence*, pages 189–199. Springer, 2013.

² P. Heinzlreiter, J.R. Perkins, O. Torreño, J. Karlsson, J.A. Ranea, A. Mitterecker, M. Blanca, O.Trelles: A Cloud-based GWAS Analysis Pipeline for Clinical Researchers In Proc. of the 4th International Conference on Cloud Computing and Services Science (CLOSER 2014), ISBN 978-989-758-019-2, Barcelona, Spain, pp. 387-394, April 2014, DOI 10.5220/0004802103870394

³ Alex Upton, Oswaldo Trelles, James Perkins, *Epistatic Analysis of Clarkson Disease*, *Procedia Computer Science*, Volume 51, 2015, Pages 725-734, ISSN 1877-0509, <http://dx.doi.org/10.1016/j.procs.2015.05.191>.

P20: Deep Proteome Coverage Through Ribosome Profiling and MS Integration

Elvis Ndash¹, Jeroen Crappé¹, Alexander Koch¹, Sandra Steyaert¹, Gerben Menschaert¹, Petra V. Damme²

¹ Biobix, University of Gent

² VIB Department of Medical Protein Research, University of Gent

The novel ribosome profiling (RIBO-seq) approach provides genome-wide information about protein synthesis by monitoring mRNA entering the translation machinery, while highly sensitive mass spectrometry (MS) provides information about the protein composition of a sample. Integrating these technologies provides

more intuitive information about the protein synthesis and the identification of novel translation products as well as a better understanding of the translation mechanism.

We developed a proteogenomic pipeline, called PROTEOFORMER, that automatically processes data from RIBO-seq experiments, resulting in the genome-wide visualization of ribosome occupancy. The tool includes pre-processing, mapping to a reference genome, sequence variation analysis and identification of translation initiation sites, allowing the delineation of the open reading frames of all translation products. A complete protein synthesis-based sequence database can thus be compiled for MS-based identification from shotgun proteomics and N-terminomics experiments. The tool is freely available as a stand-alone pipeline and has been implemented in the GALAXY framework allowing easy integration with available proteomics tools such as SearchGUI and PeptideShaker in a multi-omics setting.

To evaluate the pipeline we performed matching RIBO-seq, gel-free shotgun and N-terminal COFRADIC proteomics experiments on mouse and human cell samples. We were able to observe an overall increase in protein identification rates, detection of 5'-extended proteoforms, upstream ORF translation and near-cognate (non-AUG) translation start sites. Furthermore, integration through the PROTEOFORMER pipeline of RIBO-seq and N-terminomics data evidenced the translation of non-coding genes in the Arabidopsis genome indicative of mis-annotation in The Arabidopsis Information Resource (TAIR10).

P21: The de.NBI RNA Bioinformatics Center

Cameron Smith¹, Torsten Houwaart¹, Anika Erxleben¹, Sebastian Will³, Altuna Akalin², Uwe Ohler², Nikolaus Rajewsky², Peter F. Stadler³, Björn Grüning¹, Rolf Backofen¹

¹ Universität Freiburg

² MDC Berlin

³ Universität Leipzig

Genome-wide sequencing revealed pervasive transcription, where the majority of the DNA encodes non-coding RNAs. Non-coding RNAs and RNA-protein interactions play a fundamental role in cellular regulation; consequently they have received increasing attention over the past decade. Recent advances in high-throughput sequencing as well as in the genome-wide identification of miRNAs and RNA-protein interactions have shown that the complexity of post-transcriptional gene regulation is equivalent to that of transcriptional gene regulation.

The recently launched German Network for Bioinformatics Infrastructure (de.NBI) aims to provide comprehensive bioinformatics services to users in life sciences research, industry and medicine. Within this network, the RNA Bioinformatics Center (RBC) is responsible for supporting RNA related research in Germany, such as the detection of noncoding RNAs and RNA structure prediction. The RBC aspires to build and advance a movement of Galaxy based RNA bioinformatics and help foster a community of users and developers in this field.

This poster details the infrastructure, services and methods the RBC will employ to meet this challenge and how Galaxy is used to provide an integrated workbench for RNA analysis.

P22: VAPoR: A Visual web pipeline for Annotation of host/pathogen interactions in Plant Resistance

Benedikt Rauscher¹, Benjamin White¹, Manuel Corpas¹, Burkhard Rost²

¹ The Genome Analysis Centre, Norwich, UK

² Department for Bioinformatics and Computational Biology, Technical University of Munich

Plants are engaged in a continuous co-evolutionary struggle for dominance with their pathogens. The outcomes of these interactions are of particular importance to human activities as they can have dramatic effects on agricultural systems. Agricultural systems such as those conferring Effector Triggered Immunity (ETI)

allow recognition of specific pathogen effectors (i.e., proteins secreted by pathogens into host cells to enhance infection). R (Resistance) genes play a crucial role in controlling a broad set of disease resistance responses whose introduction is often sufficient to stop further pathogen growth and spread.

We introduce VAPoR, a novel tool specifically tailored to annotation of resistance genes in uncharacterised genomes. Taking as input a putative genome sequence for an R gene, it gathers relevant information about experimentally annotated homologues as well as their evolutionary relationship with the candidate gene from UniProt and STRING. The information is then displayed in an interactive and intuitive way.

We tested VAPoR with two datasets: 1) a set of known R genes in *Brachypodium* spp. and 2) a putative set of R genes for *Dioscorea alata*, a species of yam. Our application is written purely in Javascript, using the BioJS and Galaxy platforms. By exploiting Galaxy's powerful data transformation facilities and the variety and interactivity of BioJS components, we are able to display an abundance of relevant information in a concentrated and intuitive way.

P23: Enabling large scale Genotype-Tissue Expression studies using Galaxy

Genna Gliner¹, Ian McDowell², Barbara E Engelhardt³

¹ Operations Research and Financial Engineering Department, Princeton University

² Computational Biology and Bioinformatics, Duke University

³ Computer Science Department and Center for Statistics and Machine Learning, Princeton University

The Princeton BEEHIVE Group develops statistical models and methods for high-dimensional genomic data. As part of the Genotype-Tissue Expression (GTEx) consortium, we are involved in processing vast quantities of RNA-sequencing and whole genome sequence data for different types of statistical and functional genomics studies, including cis- and trans-eQTLs, non-coding RNA regulation, and allele specific expression studies. The creation, testing, and deployment of the processing pipelines for each of these different study types require comprehensive analysis of large datasets through a dedicated pipeline used by all members of the group. With the ability to create custom tools and share and modify workflows, Galaxy provides a robust framework to develop this pipeline for use across our lab, but incorporating our diverse set of analysis tools into Galaxy is a non-trivial task.

In this poster we chronicle the evolution of the Princeton BEEHIVE Galaxy Pipeline. We illustrate our vision for a flexible, scalable, and streamlined pipeline using Galaxy for statistical genomics studies. We explore how our pipeline evolved by highlighting how our lab addressed the challenges of tool creation and integration, data processing and organization, and training lab members to use our Galaxy instance.

P24: A French Galaxy Tool Shed to federate the national infrastructures and offering quality assessed tools

Loraine BRILLET-GUÉGUEN¹, Christophe CARON¹, Valentin LOUX², the French Galaxy Working Group³

Presented by Olivier Inizan

¹ ABIMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France

² URI404 Mathématiques et Informatique Appliquées du Génome à l'Environnement, INRA, F-78352 Jouy-en-Josas, France

³ Institut Français de Bioinformatique [ANR-11-INBS-0013], France Génomique [ANR-10-INBS-0009] and MetaboHUB [ANR-11-INBS-0010]

The Galaxy environment, notably dedicated to bio-analyses, is finding a growing success within bioinformatics and biology communities. The "Institut Français de Bioinformatique" (IFB)

commissioned in 2013 a Working Group around the Galaxy platform. This group gathers several national platforms, and manages animation actions (Galaxy Day, thematic schools, etc.) and actions to structure (training, good practices guides, etc.) users and developers communities.

Besides, as part of the bioinformatics work packages funded by the "France Génomique" project, the community has developed or evaluated many tools and set up analysis workflows. Exploitation and diffusion of these pipelines dedicated to people unfamiliar with the command line instructions now lies on using a common platform (Galaxy) and on creating a common repository (Tool Shed). From this perspective and in the Working Group dynamic, the IFB offers a reference repository to centralize and promote the bio-analyses tools of the French community. The scope of this repository, initially dedicated to "France Génomique" NGS pipelines, is now extending to other national infrastructures (MetaboHUB, etc.) and to training actions (e.g. "Ecole NGS AVIESAN").

The IFB Tool Shed is part of a strategy to federate the community around good practices for integrating tools into Galaxy and training of engineers from concerned platforms. A special effort is made on the quality of tools and workflows integration, with functional tests and validation procedures.

P25: Statistical method for filtering sequencing error from minor clonal mutation in sequencing data and implementation

Vojtech Kulvait¹, Katerina Machova Polakova², Tomas Stopka¹

¹ Charles University in Prague

² The Institute of Hematology and Blood Transfusion

When analyzing data from current NGS technologies one have to deal with sequencing and amplification errors. For clonal disorders (we study mainly cancer and leukemia) in the patient sample there may be present subclones in low relative amounts (~1%). These subclones do have individual mutational profile. Since NGS data contains technical errors we present statistical method to distinguish biologically relevant mutations in subclones from technical errors. This method is based on fitting negative binomial distribution to the sequencing data from control samples to obtain null distribution. Then the distribution is used to detect mutations in samples. Method is implemented in Java. I agree to these terms and conditions.

P26: Yet another Galaxy Genome viewer

Thomas Darde¹, François Moreews², Yvan le Bras³, Cyril Monjeaud³, Frédéric Chalmel¹

¹ INSERM U625 – Rennes, France

² Genscale team – IRISA -Rennes, France

³ Genouest Bioinformatics facility – INRIA/IRISA – Rennes, France

Galaxy owns its own genome viewer¹. Another alternate popular genome viewer is JBrowse². We developed a server application that acts as a gateway between Galaxy and the JBrowse viewer.

Unlike Trackster, which benefits of a strong GALAXY integration, we used a loosely coupled service architecture. Our gateway application exposes services than can retrieve the configuration of a genome view or update it, using json data produced by a set of scripts wrapped as GALAXY tools. These dedicated Galaxy Tools perform the pre-processing of BAM, SAM, BED, GTF or GFF files to produce the json configuration files used by JBrowse to display new tracks.

Our application includes a session mechanism allowing one user to restore, display or update the configuration of an already existing JBrowse genome view.

A feature allows for each session, an easy exportation of the corresponding configuration files. By this way, we combine both Galaxy and JBrowse systems to be able to download and redeploy

any user-defined custom genome view, independently of the processing environment, locally or within any web server. This option is particularly useful when data are processed in a cloud-based Galaxy instance. Unlike other integration of JBrowse within Galaxy³, we provide a generic way to display data produced by Galaxy in JBrowse.

It was successfully used⁴ to interpret multifaceted “omic” data. Thus, we consider this work as i) a contribution to improve bioinformatics open source software interoperability ii) a way to deploy and spread pre-populated genome browsers with minimum technical skills.

Notes:

¹ Goecks, J., Coraor, N., Nekrutenko, A. and Taylor, J. (2012) NGS analyses by visualization with Trackster. *Nat. Biotechnol.*, 30, 1036–1039.

² Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, 19, 1630–1638.

³ Venter Institute Cloud Viral Browser : <https://github.com/JCVI-Cloud/VICVB>

⁴ The ReproGenomics Viewer: an integrative cross-species toolbox for the reproductive science community, Thomas A. Darde; Olivier Sallou; Emmanuelle Becker; Bertrand Evrard; Cyril Monjeaud; Yvan Le Bras; Bernard Jegou; Olivier Collin; Antoine D. Rolland; Frederic Chalmel *Nucleic Acids Research* 2015; doi: 10.1093/nar/gkv345

P27: Integration of Mechanical Testing Process in the Galaxy Environment

R. Créac'Hcadec¹, E. Poupart², Y. Le Bras³, O. Collin³, D. Malfondet⁴, Y. Quéré⁵

¹ LBMS EA 4325 – ENSTA-Bretagne / Université de Brest / ENIB, France

² Université Européenne de Bretagne, Pôle Système d'information – Direction des usages et services numériques, France

³ e-Biogouest project, CNRS UMR 6074 IRISA-INRIA, France

⁴ Université de Bretagne Occidentale – UFR Sciences, France

⁵ Lab-Sticc CNRS UMR 6074, Université de Bretagne Occidentale – UFR Sciences, France

The aim of integrating mechanical testing process in galaxy environment is to play a part in the enhancement of the overall experimental process efficiency.

The concepts dealing with the experimental life cycle are similar both in biological and mechanical fields. Thus, as Galaxy presents flexible setups, it appears that adapting this tool to the needs of the mechanical approach could be relevant.

The initial step is to manage raw data provided by devices such as connected measurement instruments or specific personal computers. It relies on various sub-processes such as collecting, flagging, synchronizing, preprocessing and storing data.

As raw data are produced by expensive methods, they should be automatically linked with metadata, with treatment scripts in order to generate numerical studies, and finally with a project entity so that we get a reliable bottom up management. All these associations should be stored in order to be able to replay or improve previous numerical studies as well as to match to publishing criteria.

A demonstrator will show some points mentioned above and help to schedule further developments.

P28: BioMAJ2Galaxy: automatic update of reference data in Galaxy using BioMAJ

Anthony Bretaudeau^{1,2}, Cyril Monjeaud², Yvan Le Bras², Fabrice Legeai^{1,3}, Olivier Collin²

¹ INRA, UMR Institut de Génétique, Environnement et Protection des Plantes (IGEPP), Bioinformatics Platform for Agroecosystems Arthropods (BIPAA), Campus Beaulieu, 35042 Rennes, France

² INRIA, IRISA, GenOuest Core Facility, Campus de Beaulieu, 35042 Rennes, France

³ INRIA, IRISA, GenScale, Campus de Beaulieu, 35042 Rennes, France

Many bioinformatics tools use reference data, such as genome assemblies or sequence databanks. Galaxy offers multiple ways to give access to this data through its web interface. However, the process of adding new reference data was customarily manual and time consuming, even more so when this data needed to be indexed in a variety of formats (e.g. Blast, Bowtie, BWA, or 2bit).

BioMAJ is a widely used and stable software that is designed to automate the download and transformation of data from various sources. This data can be used directly from the command line in more complex systems, such as Mobyle, or by using a REST API.

To ease the process of giving access to reference data in Galaxy, we have developed the BioMAJ2Galaxy module, which enables the gap between BioMAJ and Galaxy to be bridged. With this module, it is now possible to configure BioMAJ to automatically download some reference data, to then convert and/or index it in various formats, and then make this data available in a Galaxy server using data libraries or data managers.

The developments presented in this paper allow us to integrate the reference data in Galaxy in an automatic, reliable, and disk-space-saving way. The code is freely available on the GenOuest GitHub account (<https://github.com/genouest/biomaj2galaxy>).

P29: Colib' read on Galaxy: A tools suite dedicated to biological information extraction from raw NGS reads

Yvan Le Bras¹, Olivier Collin¹, Cyril Monjeaud¹, Vincent Lacroix², Eric Rivals³, Claire Lemaitre⁴, Vincent Miele², Gustavo Sacomoto², Camille Marchet², Bastien Cazaux³, Amal Makrini³, Leena Salmela⁵, Susete Alves-Carvalho⁴, Alexan Andrieux⁴, Raluca Uricaru⁶, Pierre Peterlongo⁴

¹ GenOuest Core Facility, UMR6074 IRISA CNRS/INRIA/Université de Rennes I, France

² BAMBOO team, INRIA Grenoble Rhône-Alpes & Laboratoire Biométrie et Biologie Évolutive, UMR5558 CNRS

³ MAB team, UMR5506 CNRS

⁴ INRIA/IRISA, Genscale team, UMR6074 IRISA CNRS/INRIA/Université de Rennes I

⁵ Department of Computer Science and Helsinki Institute for Information Technology HIIT

⁶ University of Bordeaux, LaBRI/CNRS & CBIb

With NGS technologies, life sciences face a raw data deluge. Classical analysis processes of such data often begin with an assembly step, needing large amounts of computing resources, and potentially removing or modifying parts of the biological information contained in the data. Our approach proposes to directly focus on biological questions, by considering raw unassembled NGS data, through a suite of six command-line tools.

Dedicated to “whole genome assembly-free” treatments, the Colib' read tools suite uses optimized algorithms for various analyses of NGS datasets, such as variant calling or read set comparisons. Based on the use of {textit{de Bruijn}} graph and bloom filter, such analyses can be performed in few hours, using small amounts of memory. Applications on real data demonstrate the good accuracy of these tools compared to classical approaches. To facilitate data analysis and tools dissemination, we developed Galaxy tools and tool shed repositories.

With the Colib' read Galaxy tools suite, we give the possibility to a broad range of life scientists to analyze raw NGS data. More importantly, our approach allows to keep the maximum of biological information from data and use very low memory footprint.

P30: Galaxy for biological image analysis

Sylvain Prigent¹, Yvan Le Bras¹

¹ Biogenouest

Imaging technologies for biology are manifold (photos, photonic microscopy, electron microscopy, scanners, MRI...) and evolve rapidly. This provides a huge amount of data that are not anymore possible to be analyzed manually. The image analysts community has developed software for image analysis tasks. We can distinguish generic softwares like ImageJ or Icy and specific tools developed by researchers in Matlab, Java, R, Python or C. Most of these softwares aim at processing a dedicated image analysis task and do not work together. This is the reason why, the bio-analysts community tends to develop a unique interface to merge all these tools and make them easy to use for a biologist. A solution is Galaxy. We made a first attempt to use galaxy in the context of a participative project using imaging.

In the north-West of France population of rays are evaluated by picking up and identifying the rays eggs capsules on the beaches. To make this process easier, an application has been developed to allow people to take pictures of the egg capsules on the beaches and to upload them to a dedicated website. Images can then be analyzed to identify the eggs, species by specialists or any citizen. To automate this identification step, we develop a Galaxy workflow based on ImageJ tools that extract the egg shape features.

P31: Cancer Genomics in Galaxy

Marco Albuquerque¹, Bruno Grande¹, Dr. Ryan Morin¹

¹ Simon Fraser University

An inherent difficulty in data-driven biology is the multi-disciplinary skill set required of the scientist to draw meaningful inferences from complex data sets. Cancer genomics epitomizes this problem with the advent of next-generation sequencing (NGS) and the concomitant need for computational analysis. Despite the myriad available algorithms, a bottleneck in data analysis remains because of cryptic command-line parameters, inflexible system environments with difficult installations, and demanding hardware requirements. Our project directly addresses these issues by building a cancer genomics toolbox consisting of parallelized tools and workflows for the cloud-ready Galaxy platform. Our toolbox will contain some 50 new Galaxy tools spanning several sub-categories, notably variant calling, visualization and additional helper tools for integrating and summarizing results. These will be assembled with existing tools to form Galaxy workflows. Following Galaxy best practices, users will be able to seamlessly install our tools automatically. To ensure optimal accuracy, workflow design and tool parameterization will be informed by the benchmarking results from the ICGC-TCGA DREAM challenges. All tools and workflows will be developed to ensure optimal parallelism on a cluster environment. The incomplete Map-Reduce parallelization framework offered by Galaxy will be expanded, including new

merge and split functions for NGS data types used by our tools. Ultimately, this will provide a competitive graphical user interface for performing cancer genome analyses and hopefully find a home in clinics around the world, advancing the field of personalized medicine.

P32: Trinity Galaxy Portal

Carrie Ganote¹, Ben Fulton², Brian Haas³

¹ National Center for Genome Analysis Support

² Indiana University

³ Broad Institute of MIT and Harvard

Large memory requirements, long running times and large-scale CPU consumption are some of the barriers to providing bioinformatics services through Galaxy, especially when limited hardware is available to power these services. We will outline a brief overview of our hardware and software setup and provide benchmarking data that we have collected using the Trinity RNA-Seq Assembler through the Trinity Galaxy portal at <https://galaxy.ncgas-trinity.indiana.edu>. This should provide a starting point for other institutions who may want to implement similar workflows for their users.

P33: NeLS: Norwegian e-Infrastructure for Life Sciences

Sveinung Gundersen¹, Christian Andreetta², Abdulrahman Azab¹, Kjetil Klepper³, Inge Alexander Raknes⁴, Jeevan Karloss⁵, Teshome D. Mulugeta⁵, Xiaxi Li², Patcharee Thongtra⁵, Kai Trengereid¹, Tim Kahlke⁴, Erik Semb⁴, Kidane M. Tekle²

¹ University of Oslo

² University of Bergen

³ Norwegian University of Science and Technology

⁴ University of Tromsø

⁵ Norwegian University of Life Sciences

NeLS is one of the packages of the ELIXIR.NO project and aims to provide a national Norwegian e-infrastructure allowing users within the life sciences community to efficiently and safely store, share, analyse and publish their genomics scale data. The e-infrastructure maintains a web portal at <https://nels.bioinfo.no> that functions as the central point of access to the NeLS resources, which include data storage and analysis pipelines. NeLS relies on Galaxy as its primary platform for data analysis. Each of the five participating universities hosts its own Galaxy server with different types of analysis tools and workflows reflecting the research focus of the hosting groups. Workflows include differential gene expression analysis of RNA-seq data, variant calling in somatic and germline cells, and taxonomic classification of shotgun metagenomic sequences. For analysis of human patient data, NeLS collaborates closely with the Norwegian "services for sensitive data (TSD)". The servers are either running on dedicated hardware or as a front-end to shared computer clusters. Authentication of users is done with the common electronic identity provider for the Norwegian educational sector (FEIDE), with an alternative identity provider soon in production.

Note: Poster 33 will be presented on Wednesday rather than Tuesday, due to a time conflict of the presenter.



GALAXY Community Conference

Hosted by Indiana University June 28–29, 2016



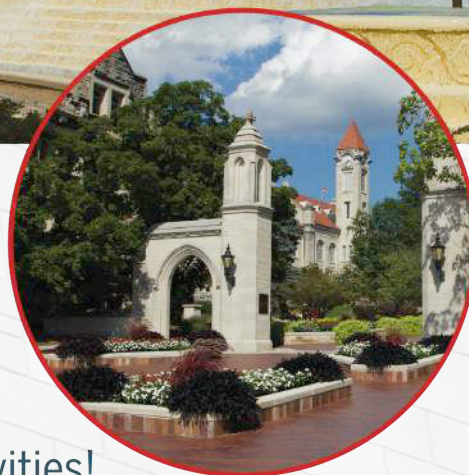
Join us in beautiful

Bloomington, Indiana

for the 2016 Galaxy
Community Conference
and pre-conference activities!

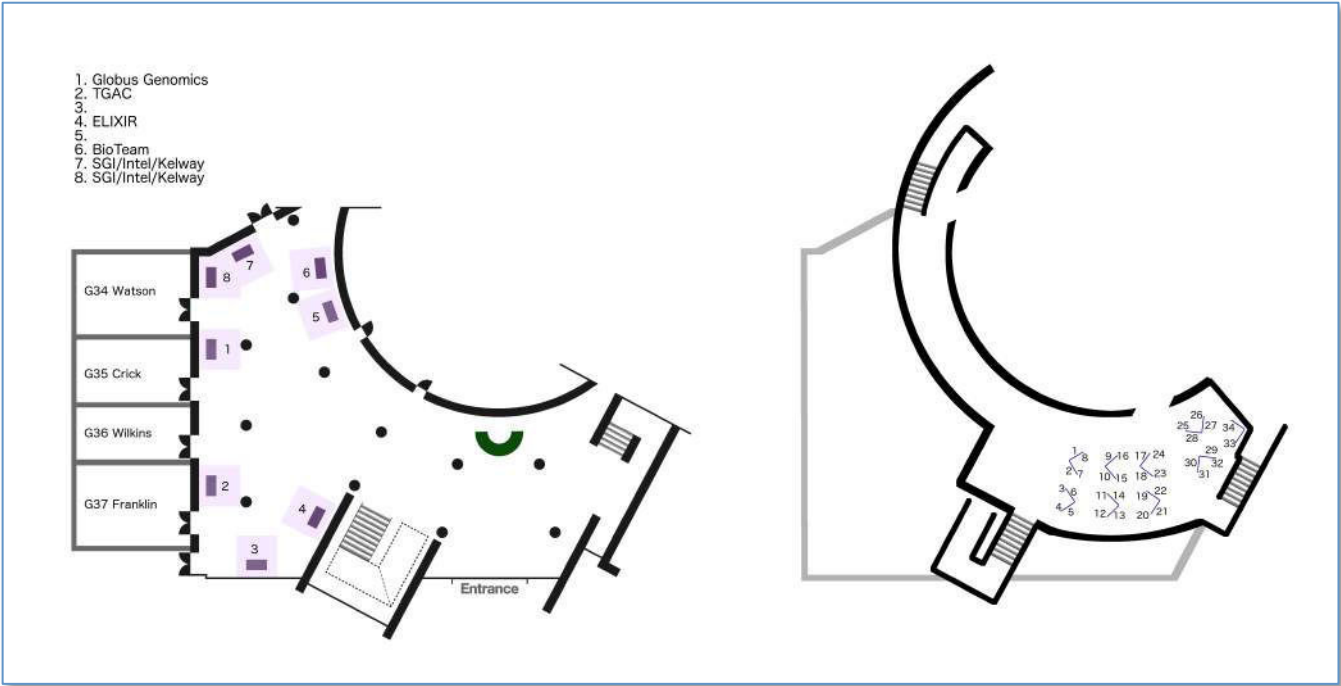
June 25–29, 2016

Considered one of the five
prettiest campuses in the US,
Indiana University is one of
the major public research
universities in the nation, and
home to the National Center
for Genome Analysis Support.

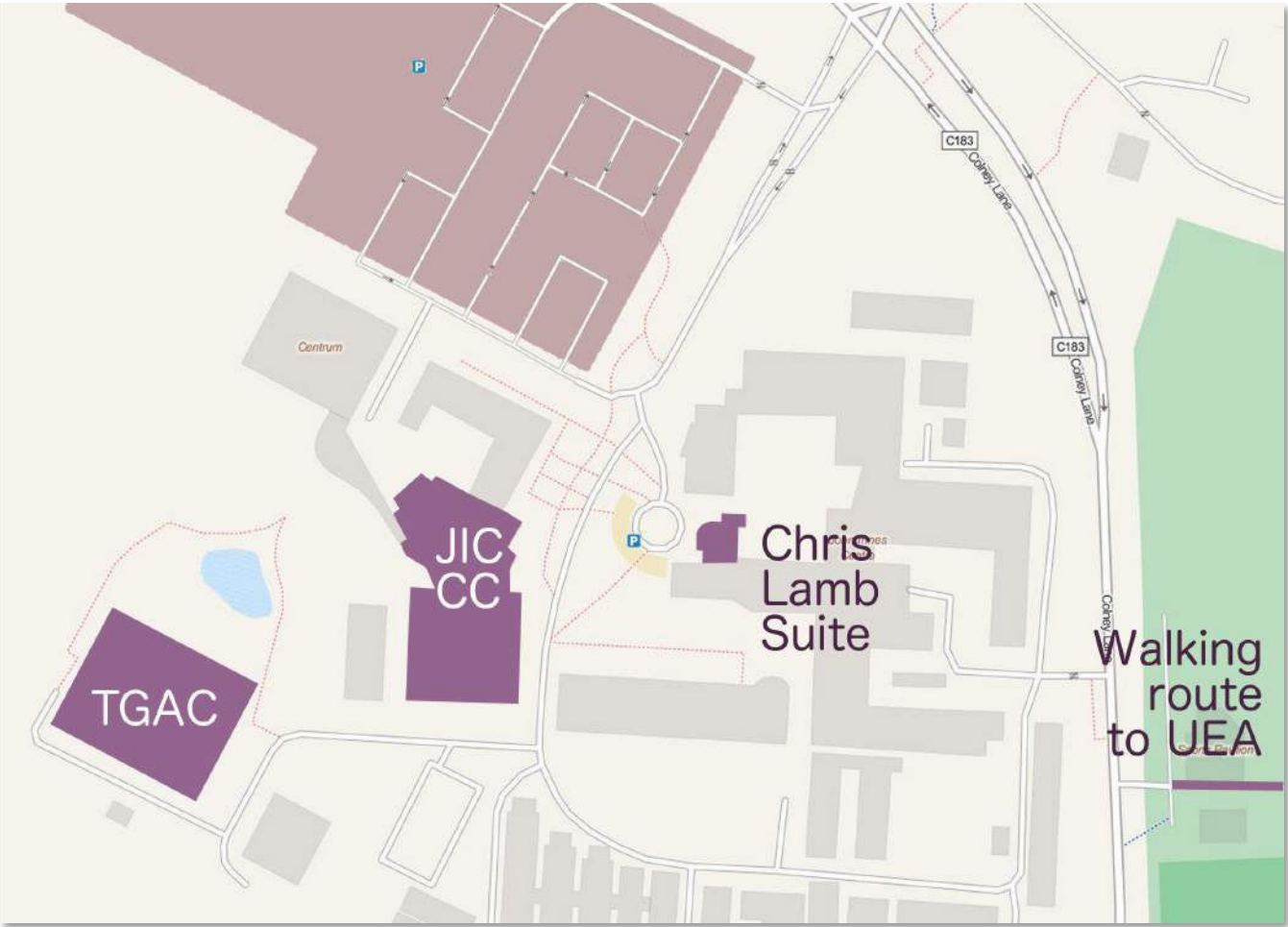


galaxyproject.org/gcc2016

John Innes Centre Conference Center



Norwich Research Park



UEA Campus

