Galaxy for SNP and Variant Data Analysis

Plant and Animal Genome XXIV (PAG 2016)
January 12, 2016

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

#usegalaxy @galaxyproject

Agenda

Minimum Information About Galaxy to Get Going (MIAGGG)

Learning Galaxy with SNP/Variation Analysis

Galaxy Ecosystem (time allowing)

http://galaxyproject.org

What is Galaxy?

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

http://galaxyproject.org

What is Galaxy?

Keith Bradnam's definition:

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

From

"13 Questions You May Have About Galaxy"

http://bit.ly/13questions

Galaxy is **available** several ways ...

http://galaxyproject.org

As a free for everyone service on the web: usegalaxy.org

https://usegalaxy.org

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 3%

Tools

search tools

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- NGS: QC and manipulation
- NGS: Mapping
- NGS: RNA-seq
- NGS: SAMtools
- NGS: BAM Tools
- NGS: Picard
- NGS: VCF Manipulation
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association
- snpEff
- BEDTools
- Genome Diversity
- EMBOSS

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

Galaxy 101

Start small

The very first tutorial you need

Tweets

Follow

NIH BD2K @NIH_BD2K 1h
Submit #BD2K #Hackathon Proposals to the BD2K Centers Coord. Center! Due OCT15 Read more at ow.ly/SlUkm pic.twitter.com/2bUDJh1tJZ
Retweeted by Galaxy Project
Show Photo

Dawei Lin @iGenomics 23h
@mike_schatz My former group at UC Davis has been update an AMI with Galaxy bioinformatics.ucdavis.edu/software/
Retweeted by Galaxy Project
Expand

Tweet to @galaxyproject

PENNSTATE 1855 | JOHNS HOPKINS UNIVERSITY | TACC | iPlant Collaborative

A free for everyone web service:

http://usegalaxy.org

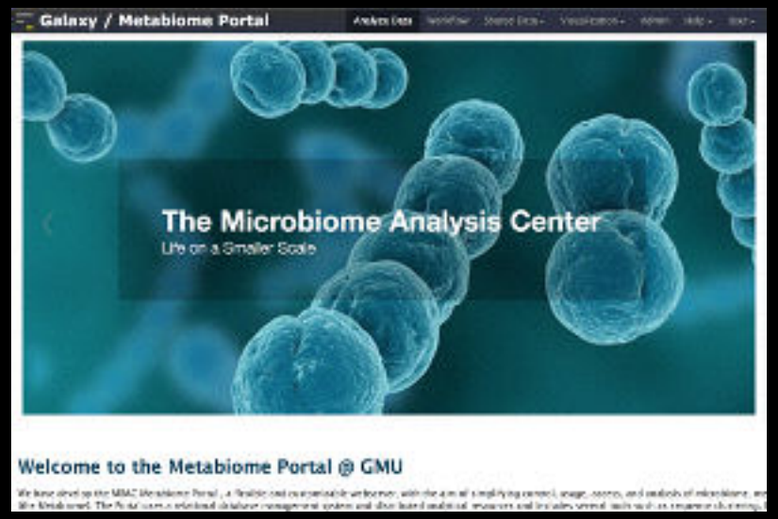
A free (for everyone) web server integrating a wealth of tools, compute resources, petabytes of reference data and permanent storage

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

Explore the Galaxy with

RNA-Rocket

PATHOGENPORTAL
THE BIOINFORMATICS RESOURCE CENTERS PORTAL



香港中文大學 - 華大基因跨組學創新研究院
CUHK-BGI Innovation Institute of Trans-Omics

(GIGA)^n Galaxy
by CBIIT

Integrated publishing of workflows from (GIGA)^n SCIENCE

Cistrome

A Galaxy Server dedicated to ChIP-* analysis

070+

Public Galaxy Servers
and *still* counting

The Genomic HyperBrowser

Powered by Galaxy

SCDE
STEM CELL DISCOVERY ENGINE

Experiments Connected

South Green
bioinformatics platform

Genomic analysis tools for southern and Mediterranean plants

bit.ly/gxyServers

Galaxy is available as Open Source Software

Galaxy is installed in locations around the world.

http://getgalaxy.org

Galaxy is available on the Cloud

OpenNebula.org

The Open Source Toolkit for Cloud Computing

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

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

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

CLOUDMAN

Agenda

Minimum Information About Galaxy to Get Going (MIAGGG)

Learning Galaxy with SNP/Variation Analysis

Galaxy Ecosystem (time allowing)

http://galaxyproject.org

Quick Poll: Are you ...

1. A bioinformatics novice
2. A bioinformatics apprentice
3. A bioinformatics guru

Yes, those are your only choices.

http://galaxyproject.org

Demo Goals

Provide a basic introduction to using Galaxy for bioinformatic analysis using SNP calling as the driving example.

Demonstrate how Galaxy can help you explore and learn options, perform analysis, and then share, repeat, and reproduce your analyses.

If you happen to learn a little bit of bioinformatics and variant detection along the way, *then that's a bonus*.

SNP and Variation Analysis Live Demo

Demonstrate a variant analysis workflow

- get a public dataset
- check and maybe fix quality concerns
- map it
- identify variants
- determine effects

https://test.galaxyproject.org

Our data

- *Oryza sativa*
- Paired end DNA reads from an exome study
- Illumina HiSeq 2000
- From the UC Davis Genome Center
- Get our copy from EBI

- Using the full dataset, but it's relatively small
- No real science going on today!

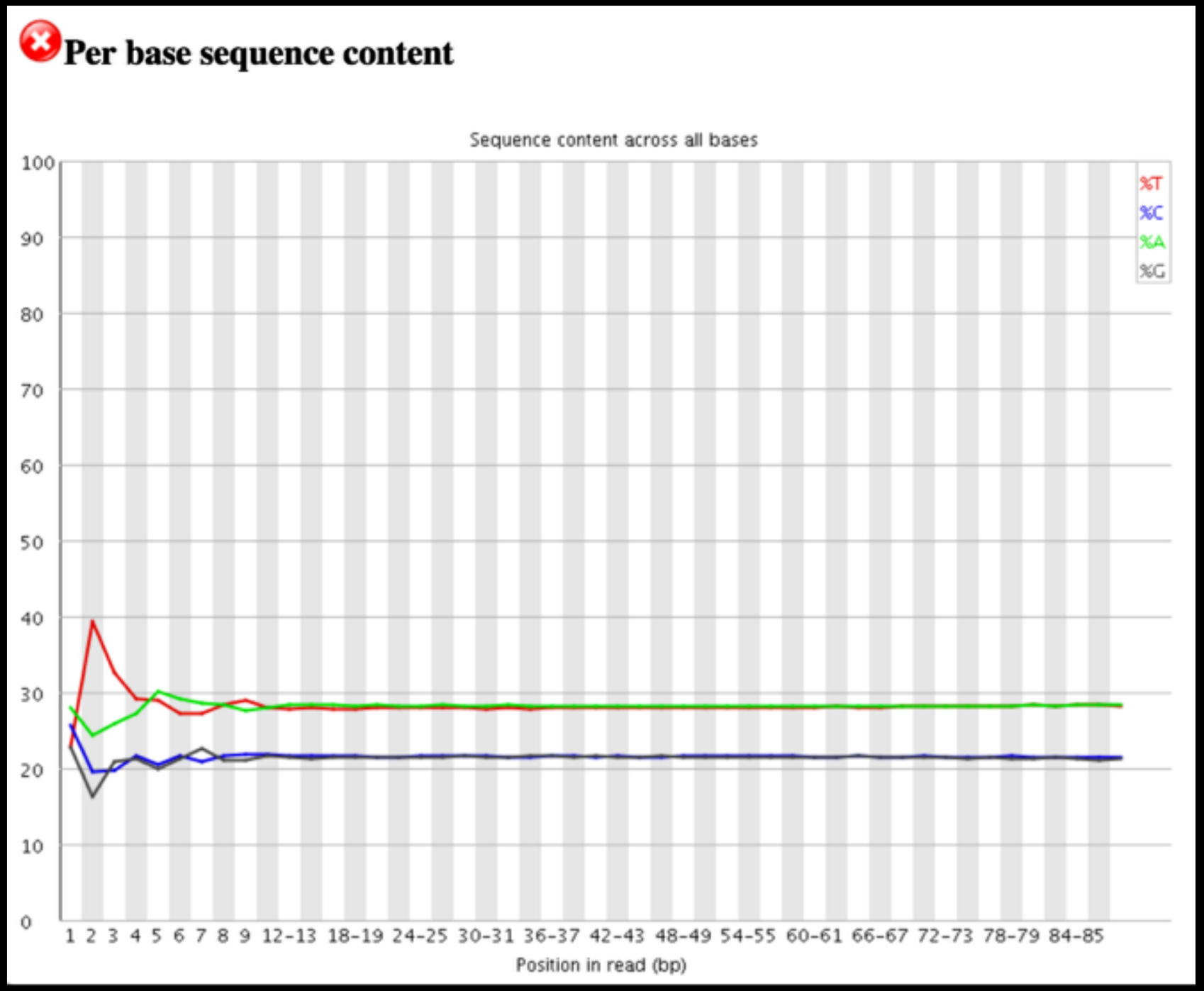
http://www.ncbi.nlm.nih.gov/sra/SRX376532
http://www.ebi.ac.uk/ena/data/view/SRR1028565

SNP and Variation Analysis Live Demo

Lets do it.

https://test.galaxyproject.org

NGS Data Quality: Sequence bias at front of reads?



From a sequence specific bias that is caused by use of random hexamers in Illumina library preparation.

Hansen, *et al.*, "Biases in Illumina transcriptome sequencing caused by random hexamer priming" *Nucleic Acids Research*, Volume 38, Issue 12 (2010)

SNP and Variation Analysis: What we did

Get data from ENA
Examine quality with FastQC
Clean it up with Trimmomatic

Map it with Bowtie2
Removed unmapped and PCR dups with BAM Filter
Looked at mapped data with FastQC & IdxStats

Called variants with FreeBayes
Calculated effects with the Variant Effect Predictor @ EBI

https://test.galaxyproject.org

Agenda

Minimum Information About Galaxy to Get Going (MIAGGG)

Learning Galaxy with SNP/Variation Analysis

Galaxy Ecosystem (time allowing)

http://galaxyproject.org

2016 Galaxy Community Conference (GCC2016)

June 25-29, 2016
Bloomington, Indiana

galaxyproject.org/GCC2016

Galaxy Resources and Community

Mailing Lists (very active)
Unified Search
Issues Board
Events Calendar, News Feed
Community Wiki
GalaxyAdmins
Screencasts
Tool Shed
Public Installs
CiteULike group, Mendeley mirror
Annual Community Meting

http://wiki.galaxyproject.org

Galaxy Community Resources: Galaxy Biostar

Tens of thousands of users leads to a lot of questions.

Absolutely have to encourage community support.

Project traditionally used mailing list

Moved the user support list to Galaxy Biostar, an online forum, that uses the Biostar platform

https://biostar.usegalaxy.org/

Galaxy Community Resources: Mailing Lists

http://wiki.galaxyproject.org/MailingLists

Galaxy-Dev

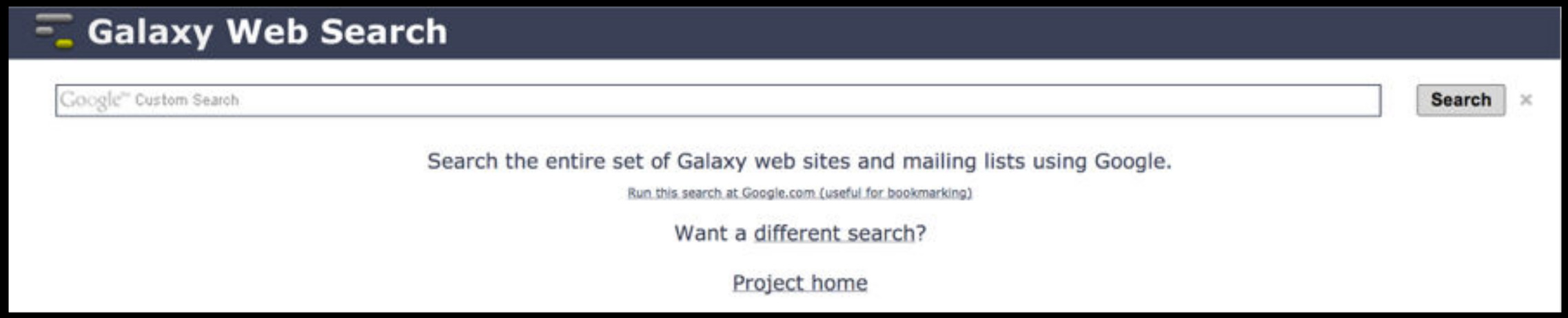
Questions about developing for and deploying Galaxy
High volume (2336 posts in 2015, 1000+ members)

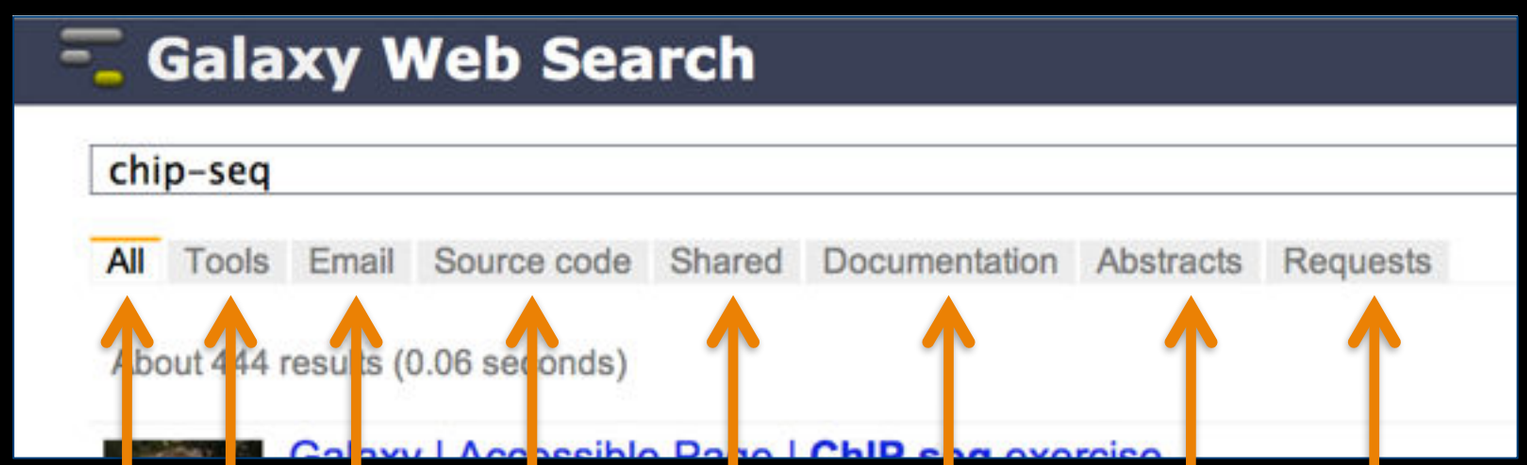
Galaxy-Announce

Project announcements, low volume, moderated
Low volume (36 posts in 2015, 6500+ members)

Also Galaxy-UK, -France, -Proteomics, -Training, ...

Unified Search: http://galaxyproject.org/search





Find

Everything on …

Tools for …

Email about …

Source code for …

Published Histories, Pages, Workflows, about …

Documentation on …

Papers using Galaxy for …

Related feature requests

http://wiki.galaxyproject.org

Galaxy Wiki DaveClements Settings Logout | Search: Titles Text

FrontPage Edit History Actions

Galaxy is an open, web-based platform for *accessible, reproducible*, and *transparent* computational biomedical research.

- **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible:** Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent:** Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

This is the Galaxy Community Wiki. It describes all things Galaxy.

Use Galaxy

Galaxy's public web server usegalaxy.org makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist. Extensive user documentation applicable to any public or local Galaxy instance is available.

usegalaxy.org

Deploy Galaxy

Galaxy is a free and open source project available to all. Local Galaxy servers can be set up by downloading the Galaxy application.

- Admin
- Cloud

getgalaxy.org

Community & Project

Galaxy has a large and active user community and many ways to get involved.

- Community

Contribute

- **Users:** Share your histories, workflows, visualizations, data libraries, and Galaxy Pages, enabling others to use and learn from them.

Galaxy web search

Use Galaxy

Servers • Learn
Main • Choices
Share • Search

Communicate

Support • Biostar
Events • Mailing Lists
News • Twitter

Deploy Galaxy

Get Galaxy • Cloud
Admin • Tool Config
Tool Shed • Search

Contribute

Develop • Tools
Issues & Requests
Logs • Deployments
Teach

Galaxy Project

Home • About • Cite
Community
Big Picture

Events

News

Galaxy Wiki

DaveClements Settings Logout | Search: Titles Text

Events

Edit History Actions

Galaxy Event Horizon

Events with Galaxy-related content are listed here.

Also see the Galaxy Events Google Calendar for a listing of events and deadlines that are
Galaxy Community. This is also available as an RSS feed.

If you know of any event that should be added to this page and/or to the Galaxy Event Calendar,
send it to outreach@glaxyproject.org.

For events prior to this year, see the **Events Archive**.

Upcoming Events

Date	Topic/Event	Venue/Location
December 12	*Introduction to Galaxy Workshop*	Virginia State University, Petersburg, Virgin
December 16-19	*RNA-Seq and ChIP-Seq Analysis with Galaxy*	UC Davis, California, United States
		2015
January 10-14	*Galaxy for SNP and Variant Data Analysis*	Plant and Animal Genome XXIII (PAG2014), States
January 19-20	*NGS pipelines with Galaxy*	e-Infrastructures for Massively Parallel Sequ Sweden
February 9-13	*Analyse bioinformatique de séquences sous Galaxy*	Montpellier, France
February 16-18	*Accessible and Reproducible Large-Scale Analysis with Galaxy*	Genome and Transcriptome Analysis, p Conference, San Francisco, Cali
	Large-Scale NGS data Analysis on Amazon Web Services Using Globus Genomic	Genomics & Sequencing Data Integration, of Molecular Medicine Tri-Conference, Sa States
	iReport: An Integrative "omics"	

News Items

Opening at McMaster University

The McArthur Lab in the McMaster University Department of Biochemistry & Biomedical Sciences is seeking a Systems Administrator / Information Technologist to help establish a new bioinformatics laboratory at McMaster, plus develop the next generation of the Comprehensive Antibiotic Resistance Database (CARD).

From the job announcement on EvolDir:

> The candidate will configure BLADE and other hardware for general bioinformatics analysis, development of a GIT version control system, **construction of an in house Galaxy server (usegalaxy.org)**, and development of a new interface, stand-alone tools, APIs, and algorithms for the CARD (based on Chado).

See the full announcement for details.

Posted to the Galaxy News on 2014-12-05

December 2014 Galaxy Newsletter

As always there's a lot going on in the Galaxy this month. "Like what?" you say. Well, read the dang December Galaxy Newsletter we say! Highlights include:

- Galaxy Day! In Paris! This Wednesday!
- Near Richmond, Virginia? There's a Galaxy Workshop at Virginia State U on December 12.
- GCC2015 needs sponsors!
- Other upcoming events on two continents
- **96 new papers**, including 6 highlighted papers, referencing, using, extending, and implementing Galaxy.
- Job openings at 7+ organizations
- A new mailing list: Galaxy-Training
- 15 new ToolShed repositories from 10 contributors
- And, 10 other juicy (well maybe not *juicy*, but certainly not *crunchy*) bits of news

Dave Clements and the *crisp* Galaxy Team

Posted to the Galaxy News on 2014-12-01

Bioinformaticians, Freiburg

Max Planck Institute of Immunobiology and Epigenetics in Freiburg, Germany has an opening for a Bioinformatician for an initial period of two years. The successful candidate will work at the interface between an in-house deep-sequencing facility (HiSeq-2500) and the various research groups at the institute. Main responsibilities include

- primary analysis of deep sequencing data and quality controls

Community can create, vote and comment on issues

HOME TOUR GOLD BUSINESS CLASS BLOG Trello Sign Up Log In

Want to subscribe, vote or comment on these cards? Sign up for free or learn more about Trello

Galaxy: Development Public

Inbox

- To add cards, use http://galaxyproject.org/trello — 4 votes, 2
- To request reference genome, comment on this card. — 1 vote, 5, 0/6
- Toolshed installation fails silently — 3 votes, 1
- Handle cluster job preemption — 2 votes, 1
- Return code 271 causes traceback for PBS torque — 1 vote, 2
- BUG: Tool shed repository export to capsule does not always capture all dependencies — 1 vote, 1
- Remove manual_builds.txt from source control and replace with a .sample version — 1 vote, 1

Tool Requests

- 595: Add SAMTools "Sort" — 4 votes, 13
- 601: SAM-to-BAM tool enhancements — 2 votes, 1
- Tools: Add tool to generate simulated reads to Main — 3 votes, 1
- default max insert size of Bowtie2 should be increased — 2 votes, 5
- 307: A tool to produce a set of random intervals. — 2 votes, 2
- Converter Tool: SAM to BAM enhancements — 2 votes
- New Tool: convert IUPAC chars to N — 1 vote, 7, 1

Bug Reports

- Usability: expanding datasets near the bottom of panel — CE
- Bug: SICER on Main dependency issue — 2 votes, 20, 3/5
- Profile Annotations bad values when "select all" — 1 vote, 5
- Filter pileup tool doesn't recognize pileup output data — 1 vote, 2
- Bug: Odd Fetch Taxonomy tool behavior — 1 vote, 1
- Strip message after pause jobs resumed — 1 vote, 1

Ideas

- 697: Workflow job control functions — 10 votes, 9
- User Metrics and Analytics — 3 votes, 3, 1/2 — CE
- Tuxedo RNA-seq tools: report command-line — 2 votes, 3
- Tools: Incorporate key Cuffdiff output files for Cummerbund — 2 votes, 1, 0/3
- Moving objects between Galaxy instances, data federation, distributed storage, and data locality — 2 votes
- Workflow Editor: Provide explicit access to implicit datatype converter tools

Pull Re

- 665: P issue
- Custo — 2 vo
- Tools: Reque — 3 vo
- add m downs — 2 vo
- Please wrapp mappi — 2 vo
- [galax libxml — 1 vo
- Pull R manag

Menu

Members: CE, NS, G

Activity

Lance Parsons on Add or update wrappers for SamTools 1.0

> I see that @peterjc has a wrapper for idxstats already and that it's listed on this card as "done" but I don't see it in the github repo. Will idxstats become part of this devteam collection or should I just start using the wrapper from @peterjc (Thanks Peter!)

today at 3:52 pm

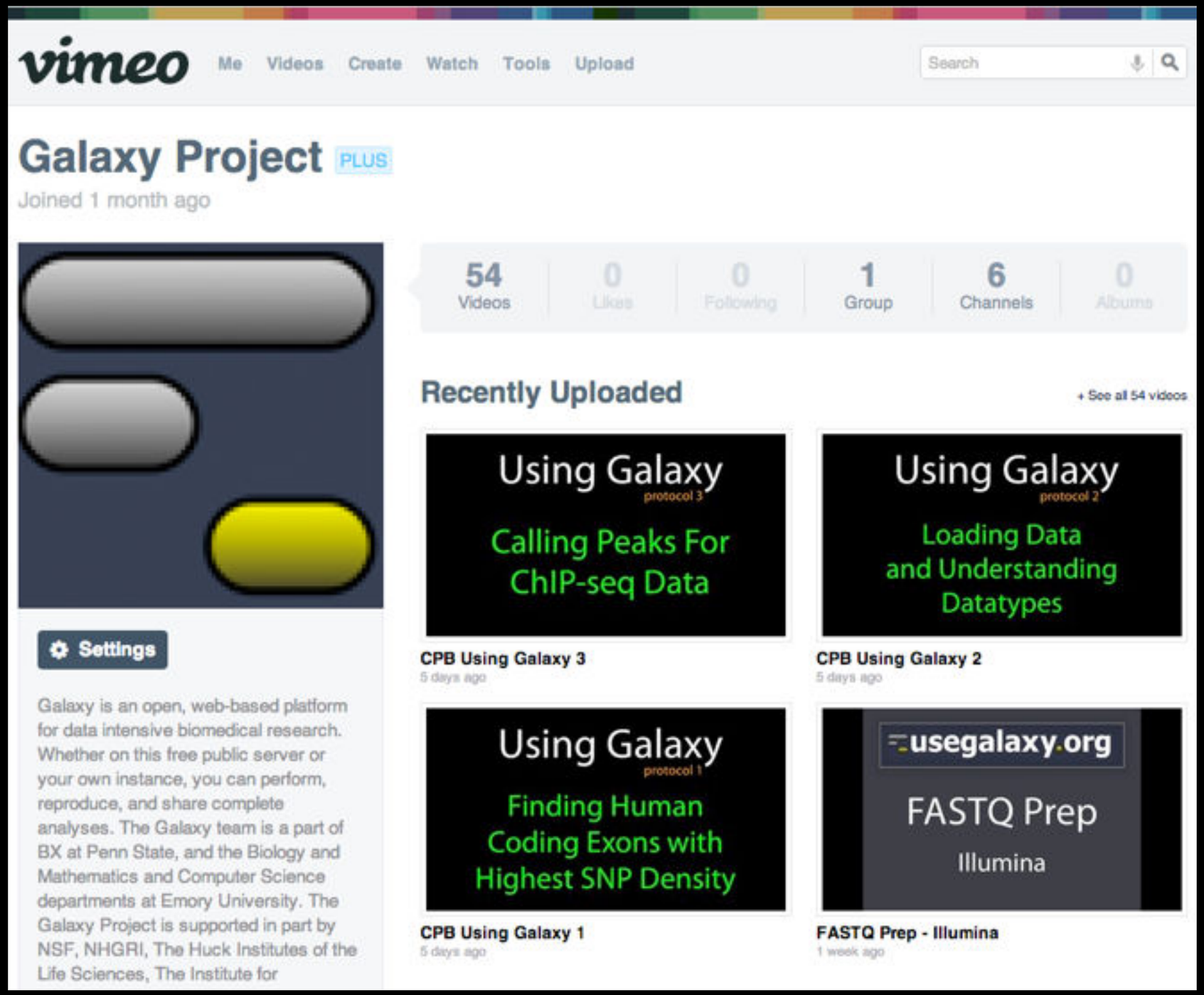
g2roboto added Pull Request #606 - [STABLE] Escape instances of message passed in through kwd before pushing them back out to

http://bit.ly/gxytrello

GCUK IS LIVE!

We also support community organized efforts and events.

Galaxy Resources & Community: Videos



"How to" screencasts on using and deploying Galaxy

Talks from previous meetings.

http://vimeo.com/galaxyproject

Galaxy Resources & Community: CiteULike Group

citeulike

CiteULike | Group: Galaxy | Search | Register | Log in

Group: Galaxy - library 2336 articles

Search | Copy | Export | Sort | Hide Details

✔ **Adaptation of the targeted capture Methyl-Seq platform for the mouse genome identifies novel tis**
methylation patterns of genes involved in neurodevelopment
Epigenetics (18 May 2015), pp. 00-00, doi:10.1080/15592294.2015.1045179
by Benjamin Hing, Enrique Ramos, Patricia Braun, et al.
posted to methods by galaxyproject to the group Galaxy on 2015-05-28 21:46:38 ★★
■ Abstract

✔ **Genomic and experimental evidence for multiple metabolic functions in the RidA/YjgF/YER057c/U**
BMC Genomics, Vol. 16, No. 1. (15 May 2015), 382, doi:10.1186/s12864-015-1584-3
by Thomas D. Niehaus, Svetlana Gerdes, Kelsey Hodge-Hanson, et al.
posted to methods usemain by galaxyproject to the group Galaxy on 2015-05-28 21:41:14 ★★
■ Abstract

✔ **NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data**
Nat. Protocols, Vol. 10, No. 6. (07 June 2015), pp. 823-844, doi:10.1038/nprot.2015.052
by Jianguo Xia, Erin E. Gill, Robert E. W. Hancock
posted to visualization by galaxyproject to the group Galaxy on 2015-05-28 21:37:43 ★★ along with 2 people and

✔ **Repression by H-NS of genes required for the biosynthesis of the Vibrio cholerae biofilm matrix is**
messenger cyclic diguanylic acid
Molecular Microbiology (1 May 2015), pp. n/a-n/a, doi:10.1111/mmi.13058
by Julio C. Ayala, Hongxia Wang, Anisia J. Silva, Jorge A. Benitez
posted to methods usemain by galaxyproject to the group Galaxy on 2015-05-28 21:30:30 ★★
■ Abstract

✔ **A Sleeping Beauty forward genetic screen identifies new genes and pathways driving osteosarcoma development and**

Group Tags

All tags in the group Galaxy

Filter:

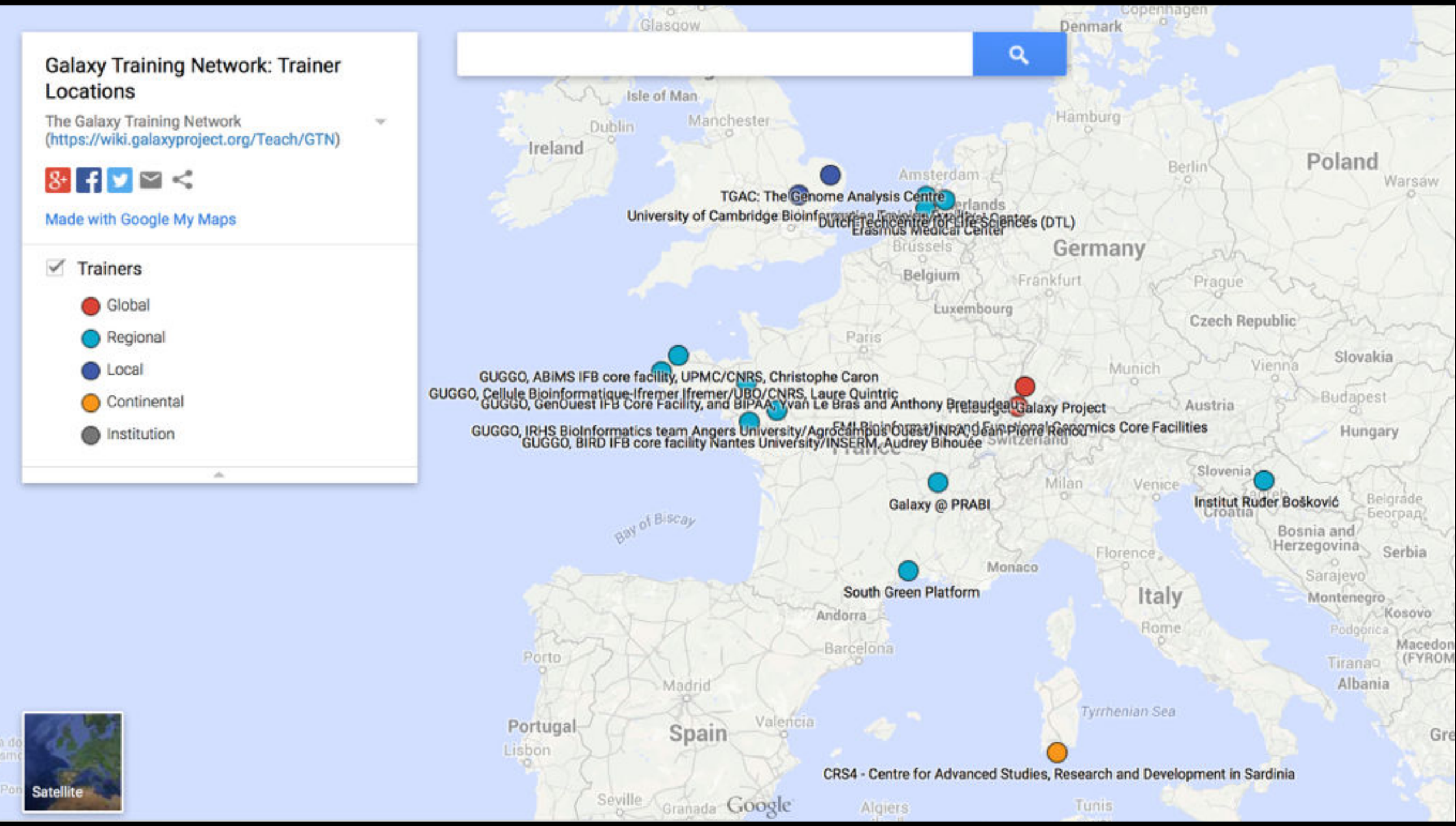
[Display as Cloud]

methods	1149
workbench	702
usemain	233
tools	169
usepublic	129
isgalaxy	124
uselocal	90
cloud	89
shared	81
other	68
refpublic	57
unknown	53
reproducibility	51
howto	45
project	43
visualization	15
usecloud	4

Now almost 3000 papers

http://bit.ly/gxycul

Scaling Training



Galaxy Training Network launched In October.

bit.ly/gxygtn

Galaxy Project: Further reading & Resources

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

http://wiki.galaxyproject.org/Cloud

http://bit.ly/gxychoices

Further adventures in Galaxy

Galaxy Community Update

Wednesday 11:25, in Golden West

Covering recent enhancements and activity in the Galaxy community.

Part of the GMOD workshop that starts @ 10:30

http://bit.ly/gmodpag16

The Galaxy Team


Enis Afgan


Dannon Baker


Dan Blankenberg


Dave Bouvier


Marten Cech


John Chilton


Dave Clements


Nate Coraor


Carl Eberhard


Jeremy Goecks


Sam Guerler


Jen Jackson


Ross Lazarus


Anton Nekrutenko


Nick Stoler

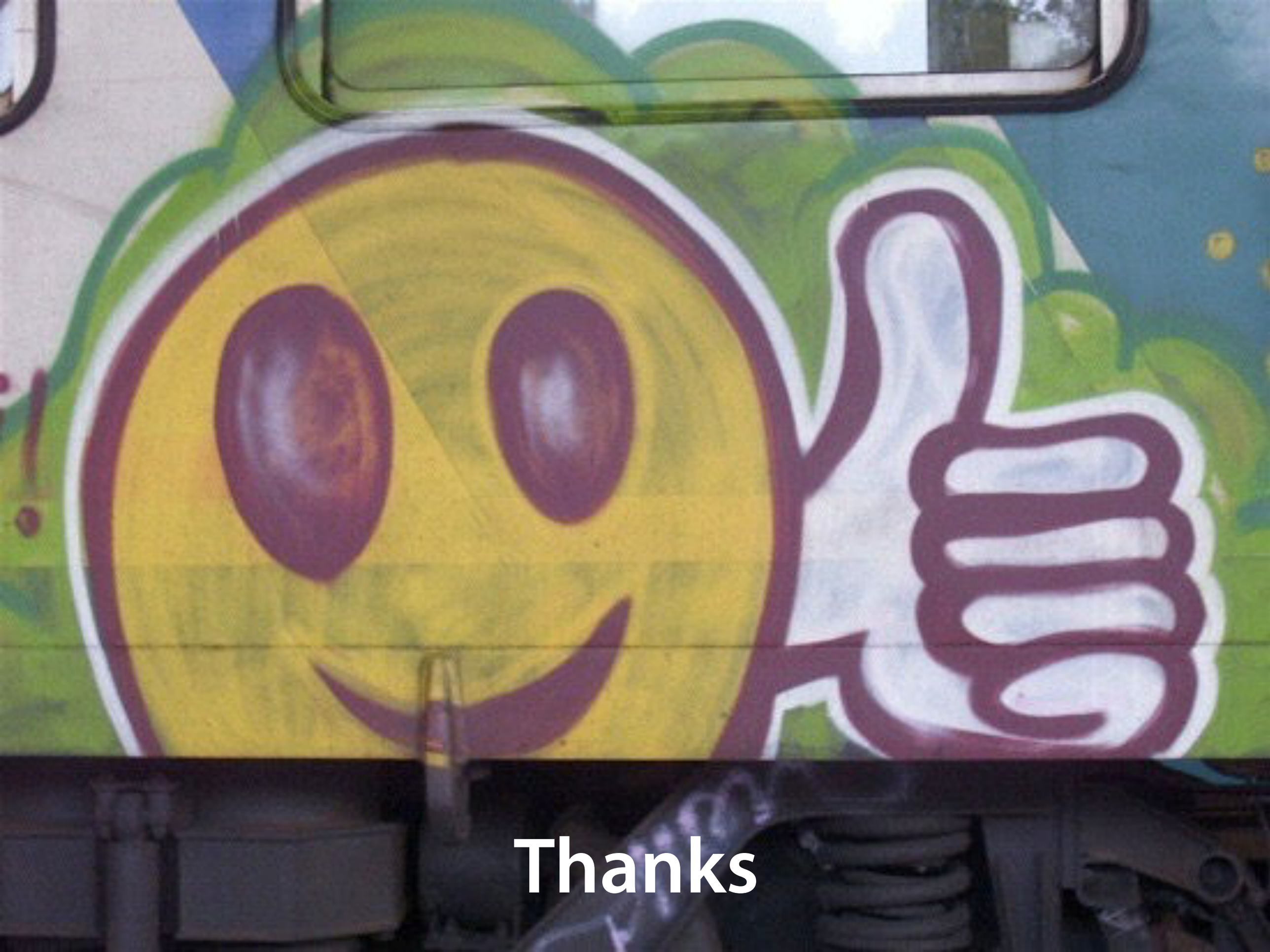

James Taylor


Nitesh Turaga

http://wiki.galaxyproject.org/GalaxyTeam

Acknowledgements

You
Anthony Bolger
Nate Coraor
PAG
NIH
Johns Hopkins University
Penn State University



Thanks