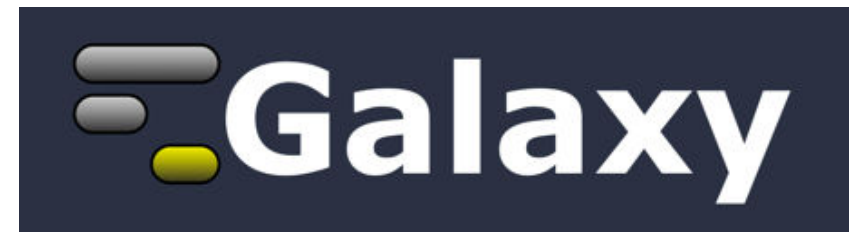


Galaxy for SNP and Variant Data Analysis

Plant and Animal Genome XXIII
San Diego, California, United States
January 13, 2015

Dave Clements
Galaxy Project
Johns Hopkins University

Galaxy Team
<http://bit.ly/gxyteam>



Outline

4:00 Introduction: The Galactic Landscape

4:10 Variant Analysis: A worked example

5:40 Variant Analysis: Other options

5:50 Galaxy: Resources and Community

6:10 Done

Slides at <http://bit.ly/>

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.

Not Goals

This workshop will *not* cover

- details of how tools are implemented, or
- new algorithm designs, or
- which SNP caller or mapper or peak caller or ... is best for you.

While this workshop does cover Galaxy *you won't become a Galaxy expert in the next two hours.*

What is Galaxy?

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

A free (for everyone) web server

Open source software

These options result in several ways to use Galaxy

<http://galaxyproject.org>

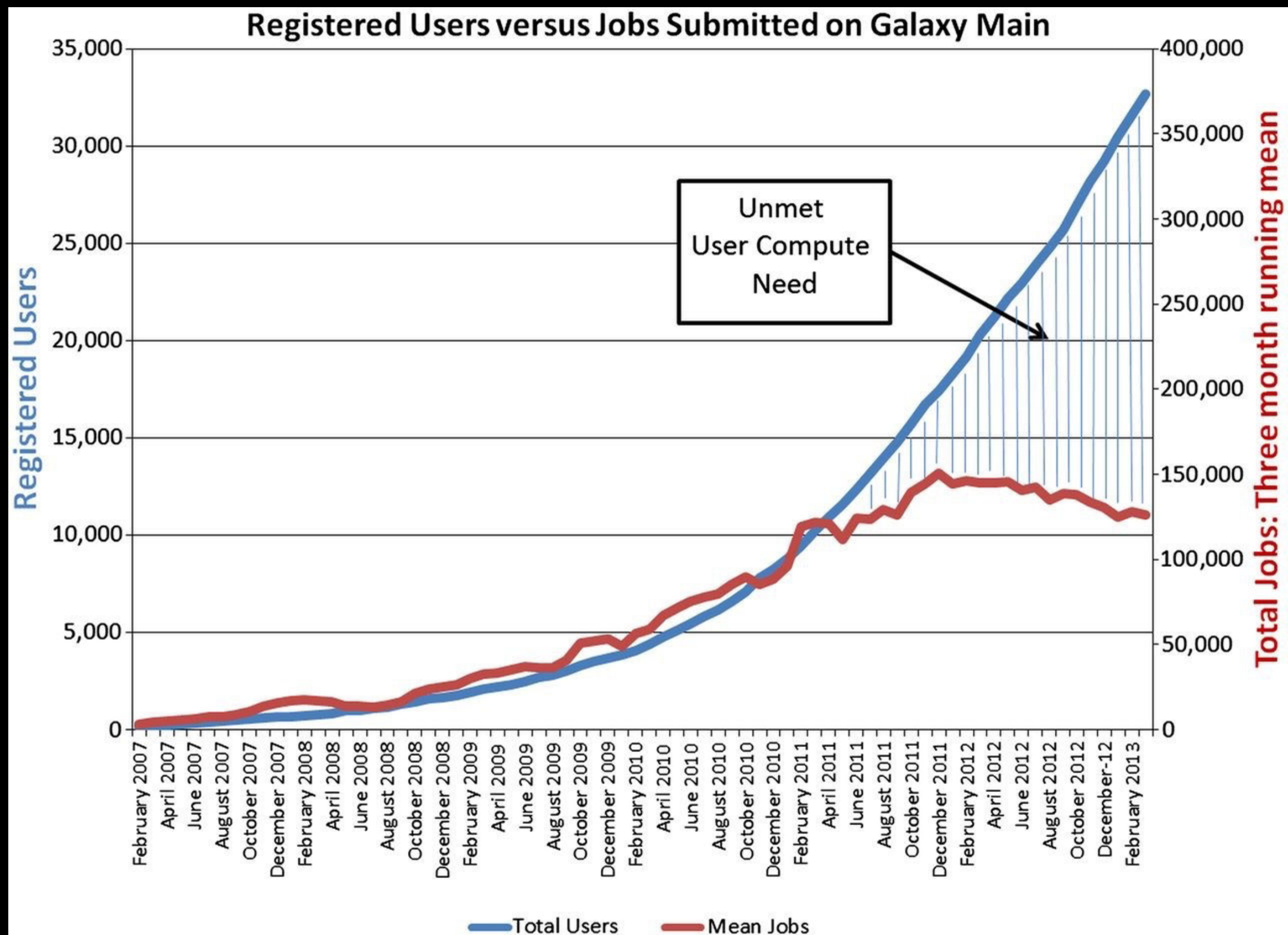
Galaxy is available ...

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

<http://usegalaxy.org>



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



Leveraging the national cyberinfrastructure for biomedical research

LeDuc, et al. *J Am Med Inform Assoc* doi:10.1136/amiajnl-2013-002059

Galaxy is available ...

- As a free (for everyone) web service

<http://usegalaxy.org>

- As open source software

<http://getgalaxy.org>

It is installed in locations around the world

Galaxy is available ...



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

<http://globus.org/>

<http://wiki.galaxyproject.org/Cloud>

We are using the cloud today.

Galaxy is available: **With Commercial Support**

A ready-to-use appliance
(BioTeam)

Cloud-based solutions
(ABgenomica, AIS,
GenomeCloud)

Consulting & Customization
(Arctix, BioTeam, Deena
Bioinformatics)



Galaxy Project: Further reading & Resources

<http://galaxyproject.org>

<http://usegalaxy.org>

<http://getgalaxy.org>

<http://wiki.galaxyproject.org/Cloud>

<http://bit.ly/gxychoices>

Outline

4:00 Introduction: The Galactic Landscape

4:10 Variant Analysis: A worked example

5:40 Variant Analysis: Other options

5:50 Galaxy: Resources and Community

6:10 Done

SNP Calling

Basic SNP Calling Exercise
Tutorial created by Dan Bolser
of EMBL-EBI



Datasets from Institute of
Population Genetics, University
of Veterinary Medicine, Vienna
(Vetmeduni Vienna)



<http://bit.ly/bolserSNP2014>

Variant Calling: A General Plan

- Get to Galaxy
- Get data to look for variation in
- Get some reference
- Align reads to reference
- Call variants

<http://bit.ly/bolserSNP2014>

Get started with Galaxy

usegalaxy.org

<http://bit.ly/bolserSNP2014>

Variant Calling: A General Plan

- Get to Galaxy
- Get data to look for variants in
- Align reads to reference
- Call variants

<http://bit.ly/bolserSNP2014>

Galaxy

Tools

search tools

Get Data

Lift-Over

Text Manipulation

Filter and Sort

NGS: QC and manipulation

NGS: Mapping

NGS: RNA Analysis

NGS: SAM Tools

NGS: BAM Tools

NGS: Picard

Galaxy

Tools

search tools

Get Data

Venn Diagram from lists

Upload File from your computer

UCSC Main table browser

UCSC Test table browser

UCSC Archaea table browser

EBI SRA ENA SRA

Get Microbial Data

BioMart Central server

Get Some Data (*but don't*)

Using ENA read set ERR358492

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User Using 0%

EMBL-EBI

Services Research Training About us

ENA
European Nucleotide Archive

Home Search & Browse Submit & Update About ENA Support

Please subscribe to ena-announce mailing list here: listserv.ebi.ac.uk/mailman/listin... to receive alerts about ENA services.

ERR358492

Examples: [BN000065](#), [histone](#)

Search

[Advanced Sequence](#)

SNP Calling: Get some data

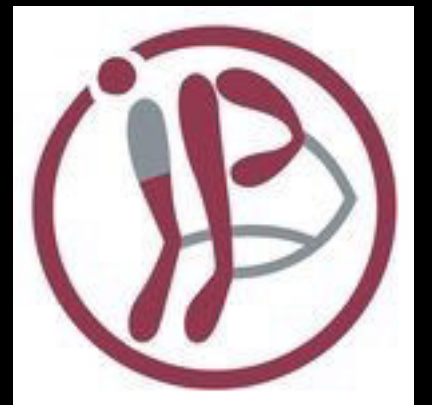
ENA read set ERR358492

Wolbachia endosymbiont of Drosophila melanogaster; wol.F15ColdReplicate d(7)

Illumina HiSeq 2000 paired end sequencing

<http://www.ebi.ac.uk/ena/data/view/ERR358492>

<http://bit.ly/ERR358492>




Wolbachia endosymbiont of Drosophila melanogaster; wol.F15ColdReplicate d(7)

Illumina HiSeq 2000 paired end sequencing

Illumina HiSeq 2000 paired end sequencing

View: XML

<http://bit.ly/ERR358492>

Send Feedback 
Download: XML

Submitting Centre Institute of Population Genetics, University of Veterinary Medicine, Vienna (Vetmeduni Vienna)	Run Date	Platform ILLUMINA	Model Illumina HiSeq 2000	Read Count 522,080	Base Count 103,658,531
Library Layout PAIRED	Library Strategy WGS	Library Source GENOMIC	Library Selection RANDOM PCR	Library Name unspecified	

Navigation Read Files

This table contains the files for run ERR358492
[Download files](#)

View: TEXT

[Select columns](#)

Showing results 1 - 1 of 1 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Fastq files (ftp)	Fastq files (galaxy)	Submitted files (ftp)	Submitted files (galaxy)
PRJEB4648	ERP003956	SAMEA2242609	ERS362605	ERX331264	ERR358492	File 1	File 1	BAM File 1	BAM File 1
						File 2	File 2		
						File 3	File 3		

But we'll just cheat.

Create new history



(cog) → Create New, name it

Import:

Shared Data → Data Libraries

→ Wolbachia Example

→ Reference (select all)

→ Reads (select all)

Import to current history → Go

<http://bit.ly/SRR488253>

NGS Data Quality Control

- FASTQ format
- Examine quality
- Trim/filter as we see fit

Quality Control is not sexy.
It is vital.

What is FASTQ?

Specifies sequence (FASTA) and quality scores (PHRED)

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
! ' ' * ( ( ( ( * * * + ) ) % % % + + ) ( % % % % ) . 1 * * * - + * ' ' ) ) * * 55CCF>>>>>CCCCCCC65
```

FASTQ is such a cool standard, there are 3 (or 5) of them!

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMN...
|          |         |         |                                |
33          59        64        73                               104                                  126
0.....26...31.....40
               -5....0.....9.....40
                   0.....9.....40
                       3.....9.....40
0.2.....26...31.....41
```

S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

http://en.wikipedia.org/wiki/FASTQ_format

FastQC: NGS Data Quality Assessment

NGS QC and Manipulation → **FastQC**

Gives you a lot of information but little control over how it is calculated or presented.

<http://bit.ly/FastQCBoxPlot>

FASTQ Groomer: FASTQ Translation

NGS QC and Manipulation → **FASTQ Groomer**

→ **Input FASTQ: Illumina 1.3 - 1.7**

→ **Execute**

Variant Calling: A General Plan

- Get to Galaxy
- Get data to look for SNPs in
- Align reads to reference
- Call Variants

<http://bit.ly/bolserSNP2014>

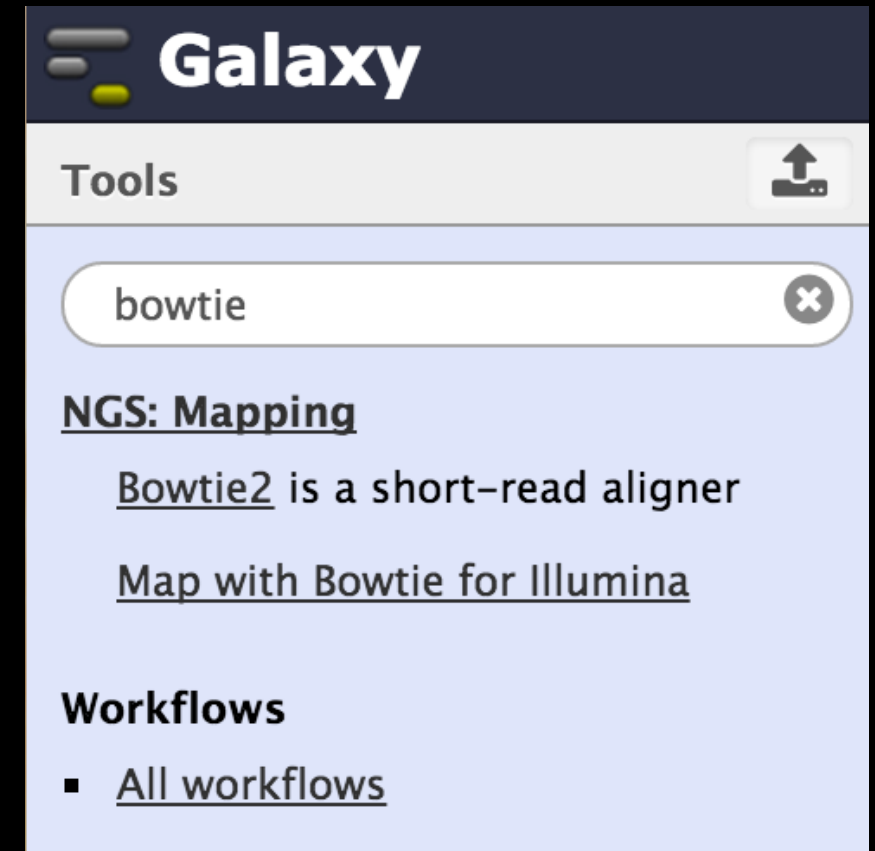
Mapping

Two most (?) common choices:

BWA and **Bowtie(2)**

Both supported by many Galaxy
instances

I'll use **Bowtie2**



But first a word about naming ...

... and tediousness

Chromosome?

chr?

ENA|AE017196|AE017196.1 ???

GFF3: Chromosome → chr

Text manipulation → Add column

Text manipulation → Cut

ENA|AE017196|AE017196.1 ???

Reference: ENA|AE017196|AE017196.1 → chr

Get data → Upload File → Paste/Fetch data

Paste:

```
chr ENA|AE017196|AE017196.1 Wolbachia  
endosymbiont of Drosophila melanogaster, complete  
genome
```

Reference: ENA|AE017196|AE017196.1 → chr

Text manipulation

→ **Line/Word/Character count**

Count number of lines in sequence

→ **Select last**

Select the last $n-1$ lines.

→ **Concatenate datasets**

Concatenate the pasted sequence name
with

the last $n-1$ lines of the reference sequence

→ Give the dataset a meaningful name

Bowtie2

From FASTQC

Reads are 50-100 bp

From the SRA Sample

Accession:

average fragment size is
... ?

Is this library mate paired?

→ Paired-end

Minimum insert size

→ 0

Maximum insert size

→ 5000 ?

Reference Genome?

→ Use one from the history

Remove PCR duplicates

NGS: SAM Tools → **rmdup**

Find variation

NGS: Variant Analysis → **Naive Variant Caller**

Source for the reference list → **History**

Minimum number of reads to consider a REF/ALT

→ **20**

Minimum base quality → **20**

Minimum mapping quality → **20**

...

Find variation

NGS: Variant Analysis → **Naive Variant Caller**

...

Ploidy → **1**

- ✓ Only write out positions with alternate alleles
- ✓ Report counts by strand

Additional Filtering: VCFfilter

NGS: VCF Manipulation → **VCFfilter**

Which ones have enough evidence to be called?

Try several tests:

-f "DP > 20" Read depth is more than 20

-f "QUAL > 30" Phred score likelihood of variant

-f "QUAL > 30" -f "DP > 20"

Filter on quality *and* read depth

VCF Filtering and Annotation w/ features

NGS: VCF Manipulation

VCFannotate

Add information about overlapping features

VCF-BEDintersect

Filter out variation that does not overlap with a genomic feature.

Some Galaxy Terminology

Dataset:

Any input, output or intermediate set of data + metadata

History:

A series of inputs, analysis steps, intermediate datasets, and outputs

Workflow:

A series of analysis steps

Can be repeated with different data

Wolbachia Variation *History* → Reusable *Workflow*?

If we are running 20 or 50 or 2000 of these experiments, we can create a **workflow** to rerun this entire analysis on each of those datasets.

Create a Workflow from a History

Extract Workflow from history

Create a workflow from this history.
Edit it to make some things clearer.



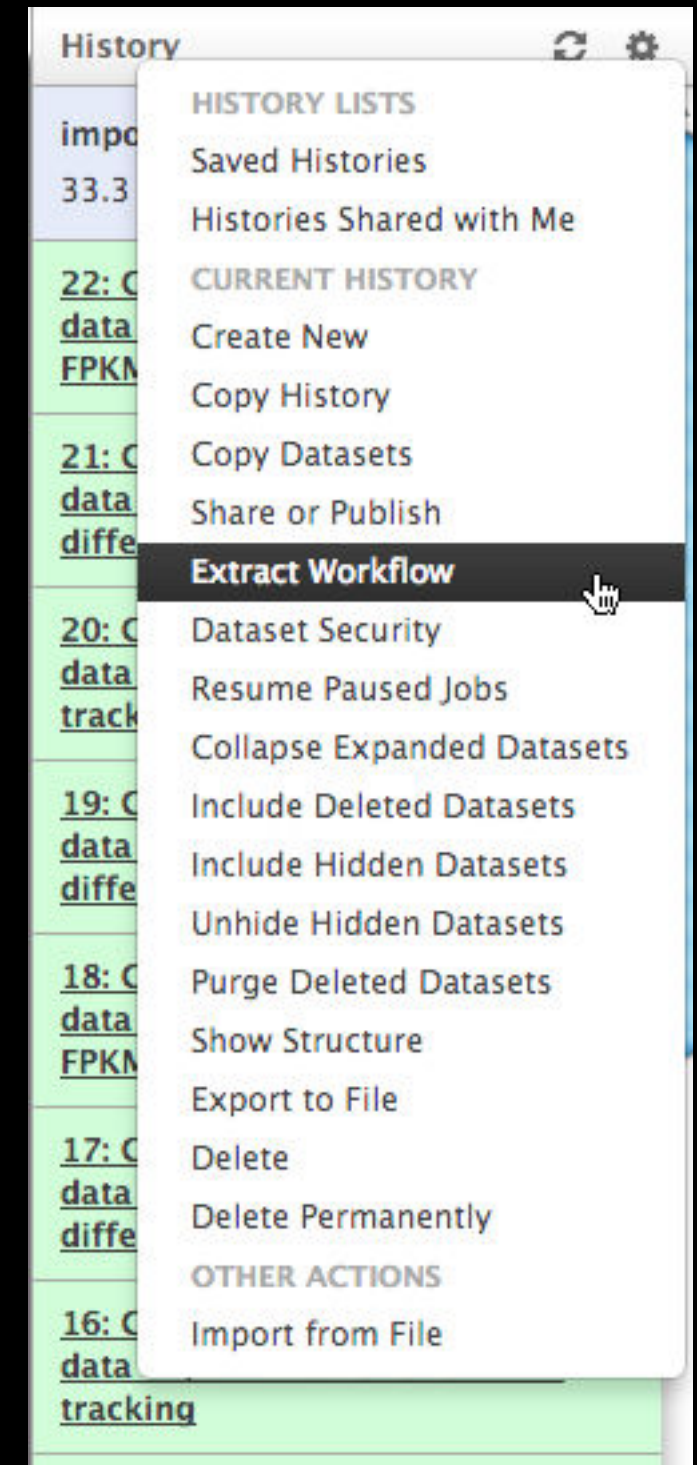
(cog) → Extract Workflow

Run / test it

Rerun with same inputs

Did that work?

When might this not work?



More Galaxy Terminology

Share:

Make something available to someone else

Publish:

Make something available to everyone

Galaxy Page:

Analysis documentation within Galaxy; easy to embed any Galaxy object

Let's all share...

Sharing & Publishing enables **Reproducibility**

Reproducibility: Everybody talks about it, but ...

Galaxy aims to push the goal of reproducibility from the bench to the bioinformatics realm

All analysis in Galaxy is recorded without any extra effort from the user.

Histories, workflows, visualizations and *pages* can be shared with others or published to the world.

Sharing & Publishing enables **Reproducibility**





Apply today for the
Cancer GWAS Grant.

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP

Institution: PENN STATE UNIV Sign In via User Name/Password

Search for Keyword:

Advanced Search

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond^{1,2,6,9}, Samir Wadhawan^{3,6,7},
Francesca Chiaromonte⁴, Guruprasad Ananda^{1,3}, Wen-Yu Chung^{1,3,8},
James Taylor^{1,5,9}, Anton Nekrutenko^{1,3,9} and The Galaxy Team¹

OPEN ACCESS ARTICLE

This Article

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109
Copyright © 2009 by Cold Spring Harbor Laboratory Press

- » Abstract **Free**
- » Full Text (PDF) **Free**

Current Issue

October 2010, 20 (10)



Footnotes

[Supplemental material is available online at <http://www.genome.org>. All data and tools described in this manuscript can be downloaded or used directly at <http://galaxyproject.org>. Exact analyses and workflows used in this paper are available at <http://usegalaxy.org/u/aun1/p/windshield-splatter>.]

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND^{1,2,*}, SAMIR WADHAWAN^{3,6*}, FRANCESCA CHIAROMONTE⁴, GURUPRASAD ANANDA^{1,3}, WEN-YU CHUNG^{1,3,7}, JAMES TAYLOR^{1,5}, ANTON NEKRUTENKO^{1,3} and THE GALAXY TEAM^{1*}

Correspondence should addressed to [SKP](#), [JT](#), or [AN](#).

How to use this document

This document is a live copy of supplementary materials for [the manuscript](#). It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must [create a Galaxy account](#) (unless you already have one) – a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:




[Galaxy History | Galaxy vs MEGAN](#)


Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):


[Galaxy History | metagenomic analysis](#)


This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

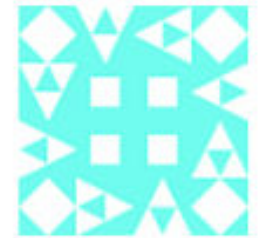

[Galaxy Workflow | metagenomic analysis](#)


Generic workflow for performing a metagenomic analysis on NGS data.

Accessing the Data

Windshield Splatter datasets analyzed in this manuscript can be accessed through this [Galaxy Library](#). From there, they can be analyzed through Galaxy using the shown workflows or downloaded.

About this Page



Author

aun1

Related Pages

[All published pages](#)
[Published pages by aun1](#)

Rating

Community
 (6 ratings, 5.0 average)



Tags

Community:

paper galaxy
 megan

<http://usegalaxy.org/u/aun1/p/windshield-splatter>

Outline

4:00 Introduction: The Galactic Landscape

4:10 Variant Analysis: A worked example

5:40 Variant Analysis: Other options

5:50 Galaxy: Resources and Community

6:10 Done

Galaxy	Galaxy	Galaxy	Galaxy
Tools	Tools	Tools	Tools
<p>NGS: Variant Analysis</p> <p><u>BamLeftAlign</u> indels in BAM datasets</p> <p><u>FreeBayes</u> – Bayesian genetic variant detector</p> <p><u>Slice VCF</u> to get data from selected regions</p> <p><u>MAF boxplot</u> Minor Allele Frequency Boxplot</p> <p><u>Phylorelatives</u> Relatedness of minor allele sequences in NJ tree</p> <p><u>CloudMap: in silico complementation</u> Perform in silico complementation analysis on multiple tabular snpEff output files</p> <p><u>CloudMap: Variant Discovery Mapping with WGS data</u> Map a mutation using in silico bulk segregant linkage analysis using variants that are already present in the mutant strain of interest (rather than those introduced by a cross to a polymorphic strain).</p> <p><u>CloudMap: Hawaiian Variant Mapping with WGS data</u> Map a mutation by plotting recombination frequencies resulting from crossing to a highly polymorphic strain</p> <p><u>CloudMap: EMS Variant Density</u></p>	<p><u>CloudMap: EMS Variant Density Mapping</u> Map a mutation by linkage to regions of high mutation density using WGS data</p> <p><u>Variant Annotator</u> process variant counts</p> <p><u>FASTA from allele counts</u> Generate major and minor allele sequences from alleles table</p> <p><u>Naive Variant Caller</u> – tabulate variable sites from BAM datasets</p> <p><u>Annotate</u> a VCF dataset with custom filters</p> <p><u>Varscan</u> for variant detection</p> <p><u>ANNOVAR</u> Annotate VCF with functional information using ANNOVAR</p> <p>NGS: VCF Manipulation</p> <p><u>VCFfixup</u>: Count the allele frequencies across alleles present in each record in the VCF file</p> <p><u>VCFprimers</u>: Extract flanking sequences for each VCF record</p> <p><u>VCFcombine</u>: Combine multiple VCF datasets</p> <p><u>VCFbreakCreateMulti</u>: Break multiple alleles into multiple records, or combine overallpoing alleles into a single record</p>	<p><u>VcfAllelicPrimitives</u>: Split allelic primitives (gaps or mismatches) into multiple VCF lines</p> <p><u>VCFfilter</u>: filter VCF data in a variety of attributes</p> <p><u>VCFrandomSample</u>: Randomly sample sites from VCF dataset</p> <p><u>VCFgenotype-to-haplotype</u>: Convert genotype-based phased alleles into haplotype alleles</p> <p><u>VCFcommonSamples</u>: Output records belonging to samples common between two datasets</p> <p><u>VCFselectsamples</u>: Select samples from a VCF dataset</p> <p><u>VCFleftAlign</u>: Left-align indels and complex variants in VCF dataset</p> <p><u>VCFtoTab-delimited</u>: Convert VCF data into TAB-delimited format</p> <p><u>VCFaddinfo</u>: Adds info fields from the second dataset which are not present in the first dataset</p> <p><u>VCFannotate</u>: Intersect VCF records with BED annotations</p> <p><u>VCF-VCFintersect</u>: Intersect two VCF datasets</p> <p><u>VCFannotateGenotypes</u>:</p>	<p><u>VCFannotateGenotypes</u>: Annotate genotypes in a VCF dataset using genotypes from another VCF dataset</p> <p><u>VCFcheck</u>: Verify that the reference allele matches the reference genome</p> <p><u>VCFsort</u>: Sort VCF dataset by coordinate</p> <p><u>VCFgenotypes</u>: Convert numerical representation of genotypes to allelic</p> <p><u>VCFhetHomAlleles</u>: Count the number of heterozygotes and alleles, compute het/hom ratio</p> <p><u>VCF-BEDintersect</u>: Intersect VCF and BED datasets</p> <p><u>VCFdistance</u>: Calculate distance to the nearest variant</p> <p>snpEff</p> <p><u>CloudMap: Check snpEff Candidates</u> Marks up a snpEff output file with matches to a gene candidate list.</p> <p><u>SnpSift Filter</u> Filter variants using arbitrary expressions</p> <p><u>SnpEff</u> Variant effect and annotation</p> <p>BEDTools</p> <p>EMBOSS</p>

2857 valid tools on Dec 16, 2014

Search

- Search for valid tools
- Search for workflows

Valid Galaxy Utilities

- Tools
- Custom datatypes
- Repository dependencies
- Tool dependency

All Repositories

- Browse by category

Available Actions

- Login to create a repository

Repositories by Category

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	36
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	10
Transcriptomics	Tools for use in the study of Transcriptomics.	6
Variant Analysis	Tools for single nucleotide polymorphism data such as WGA	147
Visualization	Tools for visualizing data	39
Data Source	Tools for retrieving data from external data sources	17
Fasta Manipulation	Tools for manipulating fasta data	42

Getting help on Variant Calling

- <http://biostars.org/>
- <http://seqanswers.com/>
- <http://galaxyproject.org/search>



Outline

4:00 Introduction: The Galactic Landscape

4:10 SNP Analysis: A worked example

5:40 SNP Analysis: Other options

5:50 Galaxy: Resources and Community

6:10 Done

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 (5200 posts in 2013, 900+ members)

(3246 posts in 2014, 1000+ members)

Galaxy-Announce

Project announcements, low volume, moderated

Low volume (47 posts in 2013, 3400+ members)

(34 posts in 2014, 4400+ members)


Galaxy-User (discontinued 2014/05)

Questions about using Galaxy and usegalaxy.org


High volume (1328 posts in 2013, 2600+ members)

(358 posts in 2014, 2600+ members)

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

 **Galaxy Web Search**

Google™ Custom Search

Search 

Search the entire set of Galaxy web sites and mailing lists using Google.

[Run this search at Google.com \(useful for bookmarking\)](#)

Want a [different search](#)?

[Project home](#)

Find

Everything on ...

Tools for ...

Email about ...


Source code for ...

Published Histories, Pages, Workflows, about ...

Documentation on ...

Papers using Galaxy for ...

Related feature requests

 **Galaxy Web Search**

chip-seq

All Tools Email Source code Shared Documentation Abstracts Requests

About 444 results (0.06 seconds)

[Galaxy | Accessible Page | ChIP-seq exercise](#)



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.



Community & Project

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

- [Community](#)

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](#)



Contribute


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



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

Events

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 in the 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, Virginia
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 Sequencing , 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 , Pacific Conference, San Francisco, California
	Large-Scale NGS data Analysis on Amazon Web Services Using Globus Genomic	Genomics & Sequencing Data Integration , of Molecular Medicine Tri-Conference, San Francisco, California

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

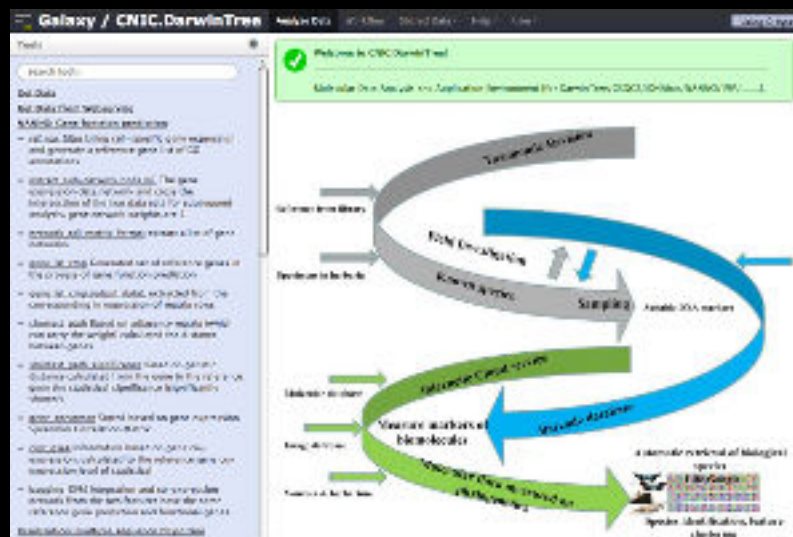
[Max Planck Institute of Immunobiology and Epigenetics](#) in Freiburg, Germany has an opening for a Bioinformatician for an initial period of two years. The successful candidate will work at the interface between an in-house deep-sequencing facility (HiSeq-2500) and the various research groups at the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



Cistrome

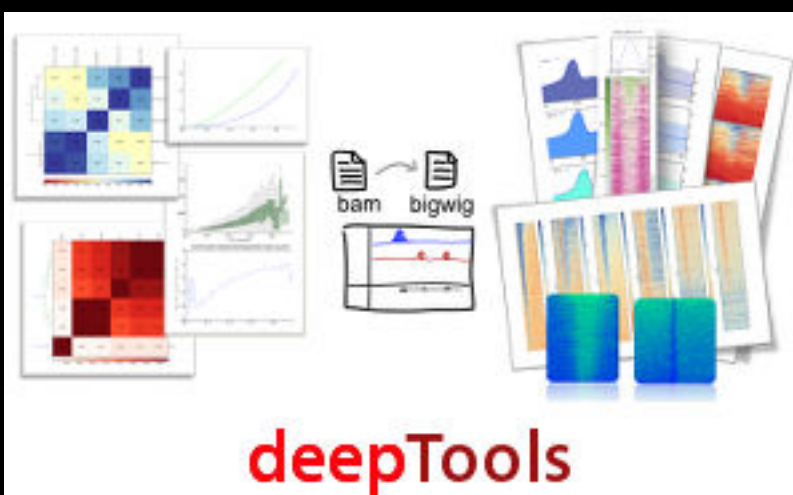


A Galaxy Server
dedicated to
ChIP-* analysis



ballaxy

Powered by the
Biochemical
Algorithms
Library
Project



deepTools



CoSSci
Galaxy for Complex Social Sciences



galaxy.berkeleybop.org



GWIS: Online exhaustive bivariate GWAS
in minutes...

David R. Krawinkel, Hanyu Pan, Andrew Krawinkel, Qian Wang, Ben Ouyang, Adam Krawinkel

Correspondence: gwis@nicta.com.au

GWIS (Genome-Wide Interactive Search) is a fast method for detecting
bivariate association between genotype and phenotype in GWAS data.
The algorithm used in GWIS was recently evaluated against
conventional methods (1) on 7 Wellcome Trust Case-Control Consortium
datasets (2).

Not only was it shown that GWIS methods were faster than all other
exhaustive algorithms, but they explicitly search for a well-defined proxy of
epistasis. An improvement in association for SNP pairs, over the
association for each individual SNP.

Web Service

We have now developed a fast online interface to GWIS, based on an instance of the GalaxyProject server (3).
Users can upload GWAS datasets in PLINK format (4) for processing with a history of previous computational tasks
for documentation, or using the 3 basic scripts in GWIS (5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100).

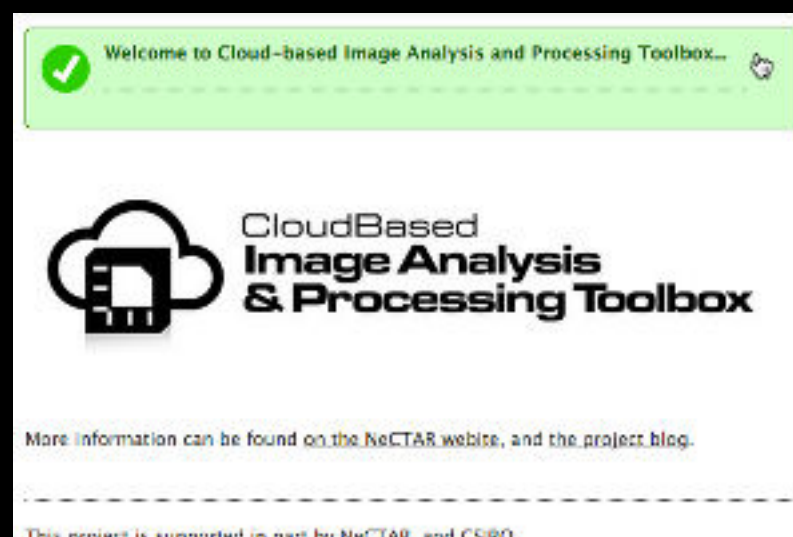
The server is free for public use as a
demonstration of the method. It is hosted on
a single desktop machine, but exhaustive
searches are performed on a cluster of
servers in the NICTA network.

For more detailed test, the server returns a
summary list of the most significant SNPs
pairs along with the source coordinates. Up to 1
million pairs can be tested.



Figure 1. Screenshot of the GWIS web service interface.

Processing Pipeline



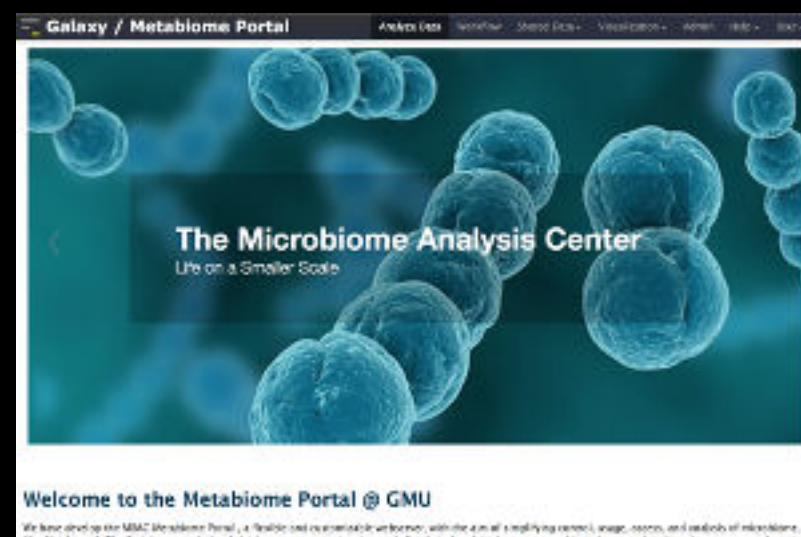
Welcome to Cloud-based Image Analysis and Processing Toolbox...



CloudBased
Image Analysis
& Processing Toolbox

More information can be found on the NeCTAR website, and the project blog.

This project is supported in part by NeCTAR, and CSIRO.



Galaxy / Metabiome Portal



The Microbiome Analysis Center
Life on a Smaller Scale

Welcome to the Metabiome Portal @ GMU

We have developed the Metabiome Portal, a flexible and customizable framework with the aim of simplifying access, usage, storage, and analysis of microbiome data. The Portal is a web-based data management system and also a cloud-based analysis and visualization tool such as a metagenomic analysis.

bit.ly/gxyServers

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

Tool Requests

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

Bug Reports

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

Ideas

- 697: Workflow job control functions
10 votes 9 comments
- User Metrics and Analytics
3 votes 3 comments 1/2
CE
- Tuxedo RNA-seq tools: report command-line
2 votes 3 comments
- Tools: Incorporate key Cuffdiff output files for Cummerbund
2 votes 1 comment 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
1 vote

Pull Requests

- 665: P... issue
Custom...
2 votes
- Tools: Reque...
3 votes
- add m...
downs...
2 votes
- Please wrapp...
mappi...
2 votes
- [galaxy...
libxml...
1 vote
- Pull Re...
manag...

Menu

Members

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

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

<http://bit.ly/gxytrello>



GALAXY

COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Slides, posters & videos now online
<http://bit.ly/gcc2014>





GCC 2015

Galaxy Community Conference

6-8th July 2015

The Sainsbury Laboratory
Norwich, UK

galaxypproject.org

Galaxy Australasia Workshop

2
0
1
4

We also support
community
organized efforts
and events.



Galaxy Resources & Community: Videos

The screenshot shows the Vimeo channel for the Galaxy Project. The header includes the Vimeo logo and navigation links: Me, Videos, Create, Watch, Tools, Upload. A search bar is located on the right. The channel name 'Galaxy Project' is displayed with a 'PLUS' badge and a note 'Joined 1 month ago'. Below this, a statistics bar shows: 54 Videos, 0 Likes, 0 Following, 1 Group, 6 Channels, and 0 Albums. The 'Recently Uploaded' section features four video thumbnails. The first two are titled 'Using Galaxy protocol 3' and 'Using Galaxy protocol 2', both by 'CPB Using Galaxy' and uploaded 5 days ago. The third is 'Using Galaxy protocol 1' by 'CPB Using Galaxy 1', also uploaded 5 days ago. The fourth is 'FASTQ Prep Illumina' by 'FASTQ Prep - Illumina', uploaded 1 week ago. A sidebar on the left contains a 'Settings' button and a paragraph of text about the Galaxy project.

Galaxy Project PLUS
Joined 1 month ago

54 Videos | 0 Likes | 0 Following | 1 Group | 6 Channels | 0 Albums

Recently Uploaded + See all 54 videos

- Using Galaxy protocol 3**
Calling Peaks For ChIP-seq Data
CPB Using Galaxy 3
5 days ago
- Using Galaxy protocol 2**
Loading Data and Understanding Datatypes
CPB Using Galaxy 2
5 days ago
- Using Galaxy protocol 1**
Finding Human Coding Exons with Highest SNP Density
CPB Using Galaxy 1
5 days ago
- FASTQ Prep Illumina**
usegalaxy.org
FASTQ Prep
Illumina
FASTQ Prep - Illumina
1 week ago

Settings

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for

“How to”
screencasts on
using and
deploying
Galaxy

Talks from
previous
meetings.

<http://vimeo.com/galaxyproject>

Galaxy Resources & Community: CiteULike Group



[CiteULike](#) [MyCiteULike](#) [Group: Galaxy](#) [Search](#) Logged in as [galaxyproject](#) [Log Out](#)

Group: Galaxy - library 1437 articles

You are an administrative member of this group.
Invite [other CiteULike users](#) to join, or invite [people who don't use CiteULike yet](#).

[Search](#) [Unwatch](#) [Copy](#) [Export](#) [Sort](#) [Hide Details](#)

☐ **✓ Life science data analysis workflow development using the bioextract server leveraging the iPlant collaborative cyberin**
Concurrency Computat.: Pract. Exper. (1 February 2014), pp. n/a-n/a, [doi:10.1002/cpe.3237](#)
by [Carol M. Lushbough](#), [Etienne Z. Gnimpieba](#), [Rion Dooley](#)
posted to [workbench](#) by [galaxyproject](#) to the group [Galaxy](#) keyed Lushbough2014Life on 2014-03-04 19:10:09 ★★/
[Abstract](#) [Copy](#) [My Copy](#)

☐ **✓ Workshops: A Great Way to Enhance and Supplement a Degree**
PLoS Comput Biol, Vol. 10, No. 2. (27 February 2014), e1003497, [doi:10.1371/journal.pcbi.1003497](#)
by [Segun Fatumo](#), [Sayane Shome](#), [Geoff Macintyre](#)
posted to [other](#) by [galaxyproject](#) to the group [Galaxy](#) keyed Fatumo2014Workshops on 2014-03-04 19:08:20 ★★/
[Abstract](#) [Copy](#) [My Copy](#)

☐ **✓ Wrangling Galaxy's Reference Data**
Bioinformatics (28 February 2014), [doi:10.1093/bioinformatics/btu119](#)
by [Daniel Blankenberg](#), [James E. Johnson](#), [James Taylor](#), [Anton Nekrutenko](#)
posted to [project](#) by [galaxyproject](#) to the group [Galaxy](#) keyed Blankenberg2014Wrangling on 2014-03-04 18:55:14 ★★★★★/
[Abstract](#) [Copy](#) [My Copy](#)

☐ **✓ Detection of PIWI and piRNAs in the mitochondria of mammalian cancer cells**
Biochemical and Biophysical Research Communications (March 2014), [doi:10.1016/j.bbrc.2014.02.112](#)
by [ChangHyuk Kwon](#), [Hyosun Tak](#), [Mina Rho](#), [et al.](#)
posted to [methods](#) by [galaxyproject](#) to the group [Galaxy](#) keyed Kwon2014Detection on 2014-03-04 18:53:07 ★★/ [along with 1 person](#)
[Copy](#) [My Copy](#)

☐ **✓ CanSNPer: a hierarchical genotype classifier of clonal pathogens**
Bioinformatics (25 February 2014), [doi:10.1093/bioinformatics/btu113](#)
by [Adrian Lärkeryd](#), [Kerstin Myrtenäs](#), [Edvin Karlsson](#), [et al.](#)
posted to [tools](#) by [galaxyproject](#) to the group [Galaxy](#) keyed Larkeryd2014CanSNPer on 2014-03-04 18:51:21 ★★/
[Abstract](#) [Copy](#) [My Copy](#)

☐ **✓ Web-based Workflow Planning Platform Supporting the Design and Execution of Complex Multiscale Cancer Models**
pp. 1-1, [doi:10.1109/jbhi.2013.2297167](#)

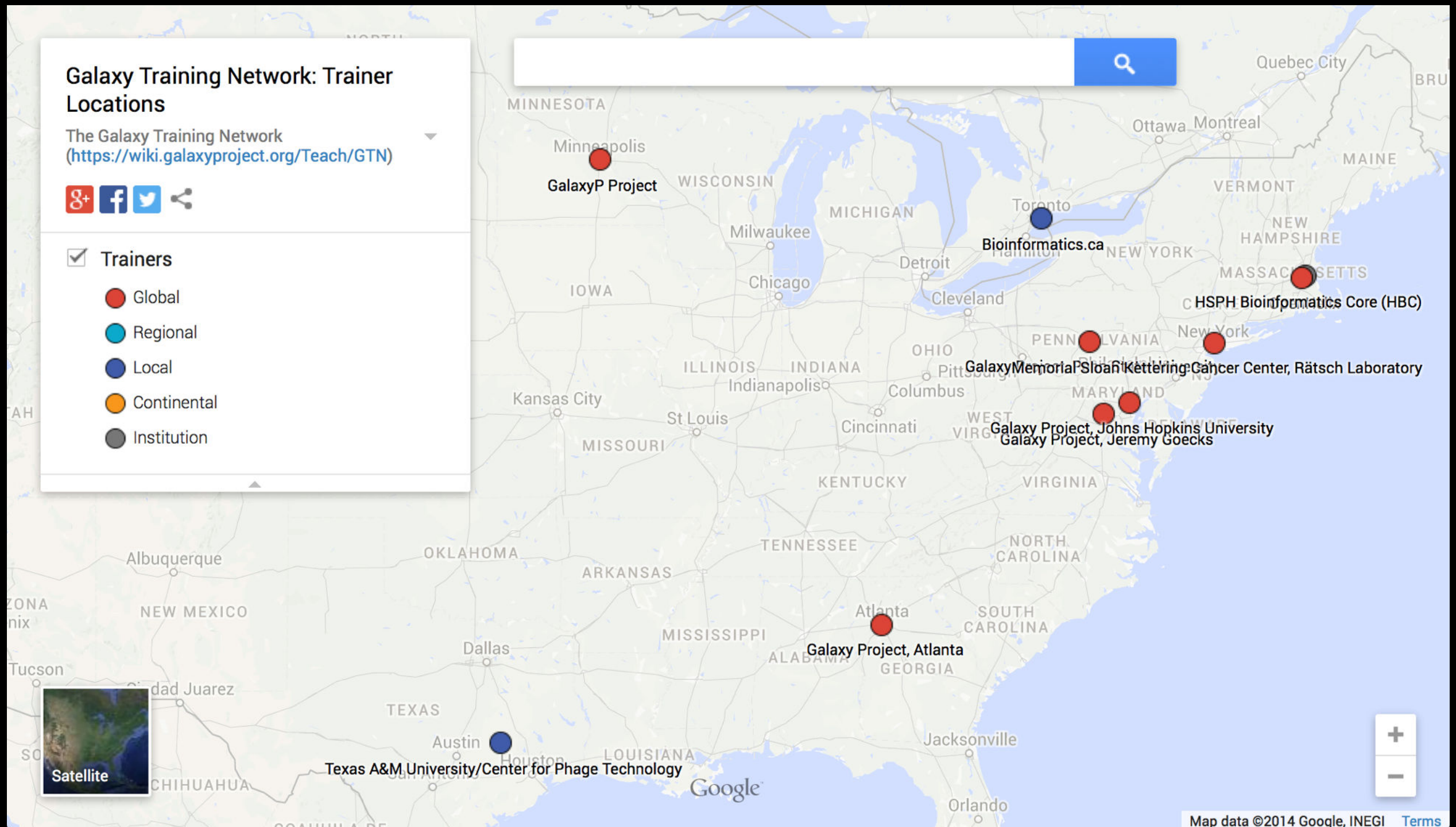
Group Tags
All tags in the group Galaxy
Filter:
[\[Display as Cloud\]](#)

methods	697
workbench	466
usemain	108
tools	91
isgalaxy	80
cloud	50
shared	50
unknown	47
uselocal	37
project	32
howto	30
reproducibility	28
other	23
usepublic	19
refpublic	12
visualization	7
usecloud	3

Over
2000
papers

<http://bit.ly/gxycul>

Scaling the Project: Training



Galaxy Training Network launched In October.

bit.ly/gxygtn

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



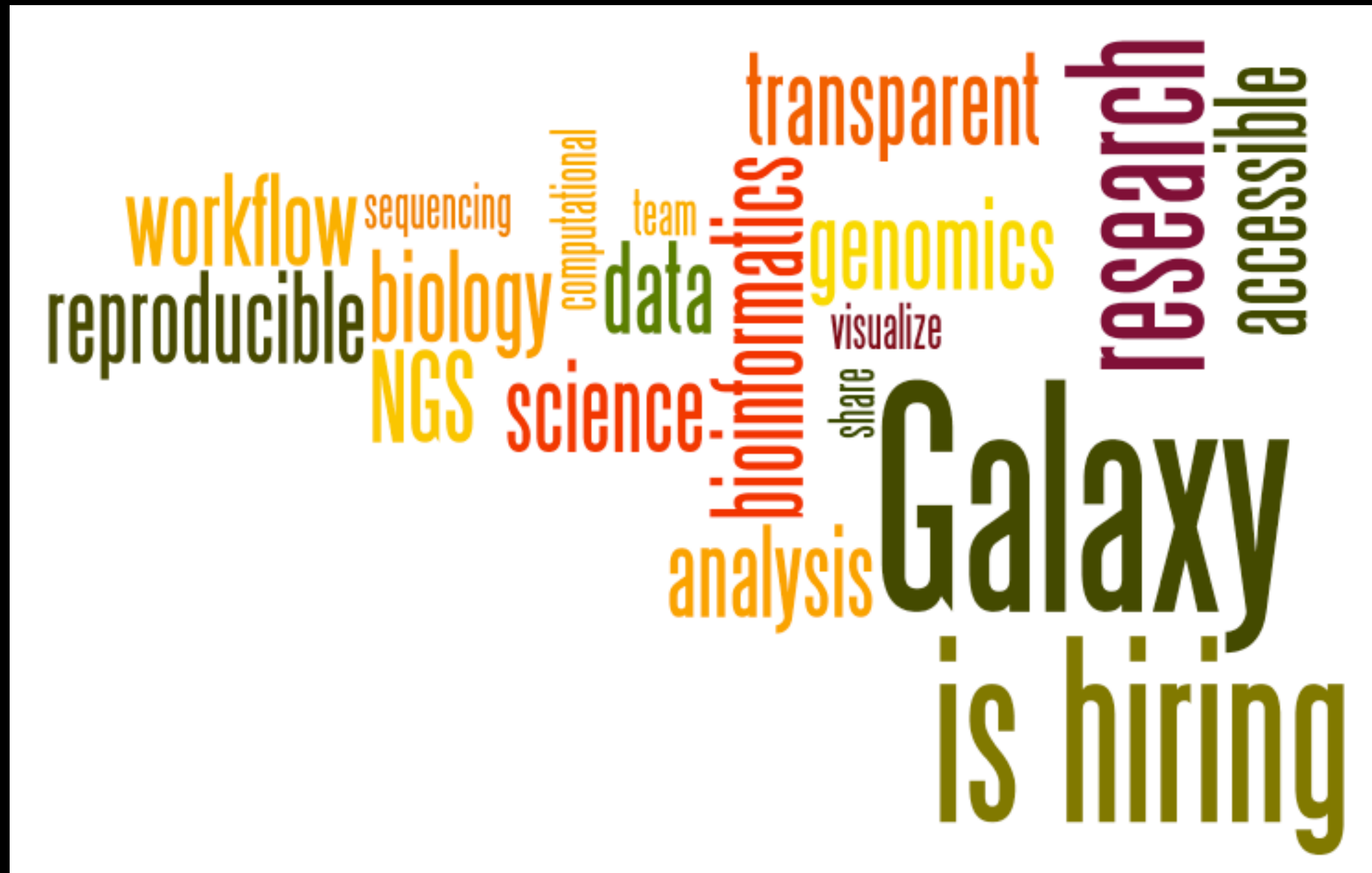
James Taylor



Nitesh Turaga

<http://wiki.galaxyproject.org/GalaxyTeam>

Galaxy is hiring post-docs and software engineers



Please help.

<http://wiki.galaxyproject.org/GalaxyIsHiring>

Thanks



Dave Clements

Galaxy Project
Johns Hopkins University
outreach@galaxyproject.org