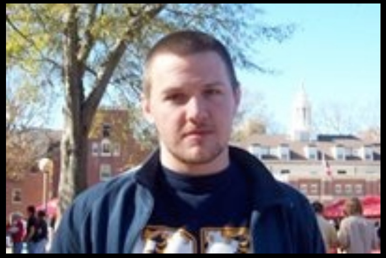
Enhancing the Galaxy Experience through Community Involvement

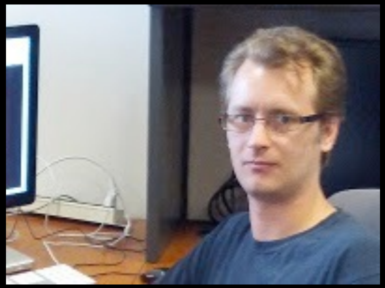
Daniel Blankenberg
Postdoctoral Research Associate
The Galaxy Team | Nekrutenko Lab@Penn State
http://UseGalaxy.org

The Galaxy Team


Enis Afgan


Dannon Baker


Dan Blankenberg


Dave Bouvier


Marten Čech


John Chilton


Dave Clements


Nate Coraor


Carl Eberhard


Jeremy Goecks


Sam Guerler


Jen Jackson


Ross Lazarus


Anton Nekrutenko


Nick Stoler


James Taylor


Greg Von Kuster

http://wiki.galaxyproject.org/GalaxyTeam

Overview

What is Galaxy?

Usage / Contributions

Virtual Interactions

ToolShed

Events and Outreach

Overview

What is Galaxy?

Usage / Contributions

Virtual Interactions

ToolShed

Events and Outreach

Galaxy Project Mission

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.

What is Galaxy?

A **data analysis and integration** tool

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

Open source software that makes integrating your own tools and data and customizing for your own site simple

There are several **ways to use Galaxy**

Using Galaxy - 4 ways

- **Public Main** Galaxy web instance: ***usegalaxy.org***
- **Local** instance: ***getgalaxy.org***
- **Cloud** instance: ***usegalaxy.org/cloud***
- **Other Public** Galaxy web instances hosted by various groups: wiki.galaxyproject.org/PublicGalaxyServers

Galaxy as a *Genomics WorkBench*

Dataset:

Any input, output or intermediate set of data + metadata. A record of a specific data or analysis step.

History:

A series of inputs, analysis steps, intermediate datasets, and outputs. A record of a group of data and analysis steps.

Tool:

An operation within Galaxy that acts upon dataset(s) as an analysis step. May be developed by Galaxy team or a 3rd party program that has been “wrapped” for Galaxy.

Workflow:

A series of analysis steps executed as a unit.

More Galaxy Terminology

Share:

Make something available to someone else

Publish:

Make something available to everyone

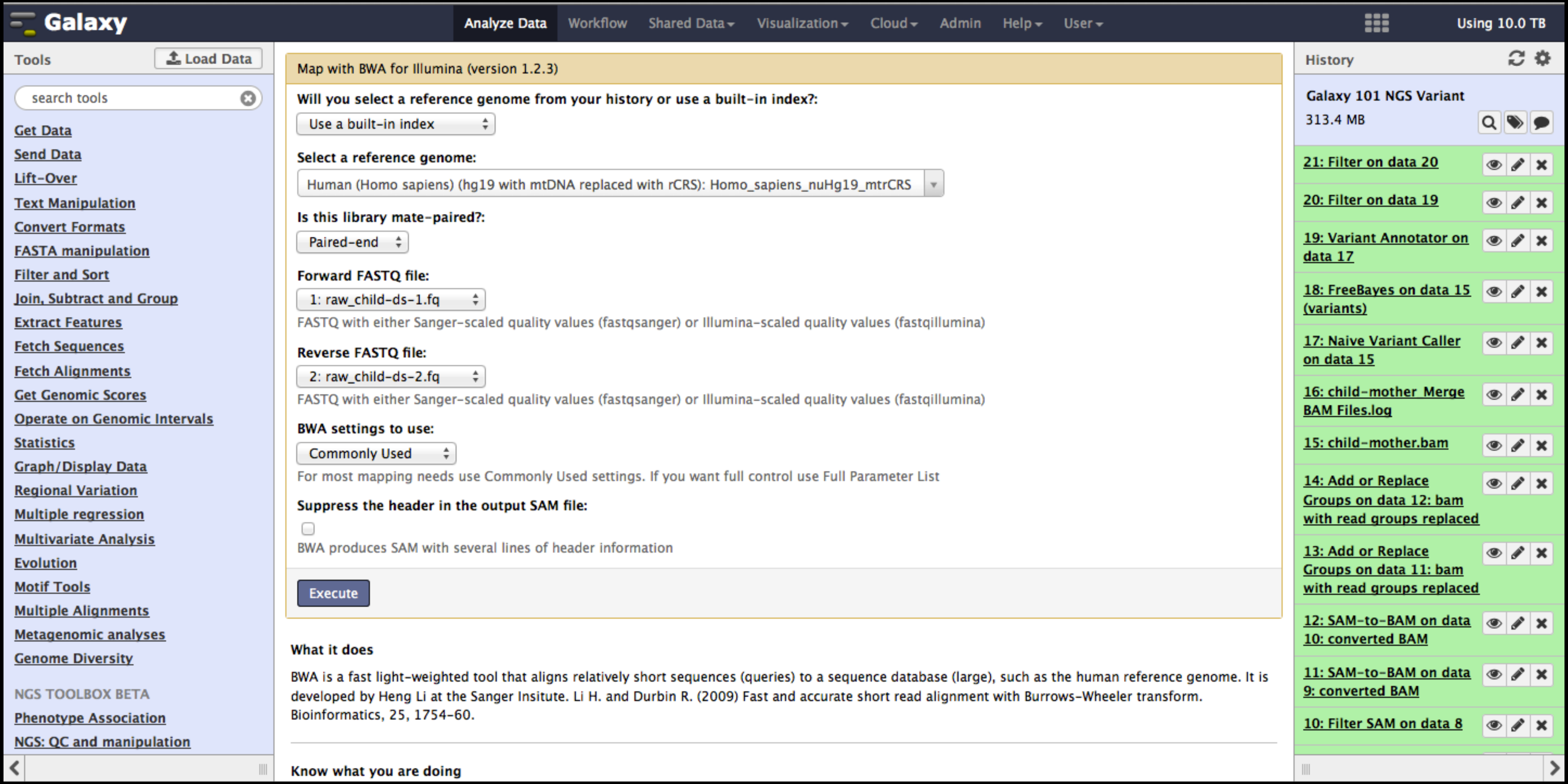
Galaxy Page:

Analysis documentation within Galaxy; easy to embed and link to any Galaxy object (histories, datasets, workflows, visualization) or external resource (video, graphics, publications).

Visualize:

External resources. Trackster. Galaxy Charts (D3/NVD3).

Galaxy Analysis Workspace



Map with BWA for Illumina (version 1.2.3)

Will you select a reference genome from your history or use a built-in index?:
Use a built-in index

Select a reference genome:
Human (Homo sapiens) (hg19 with mtDNA replaced with rCRS): Homo_sapiens_nuHg19_mtrCRS

Is this library mate-paired?:
Paired-end

Forward FASTQ file:
1: raw_child-ds-1.fq
FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)

Reverse FASTQ file:
2: raw_child-ds-2.fq
FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)

BWA settings to use:
Commonly Used
For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List

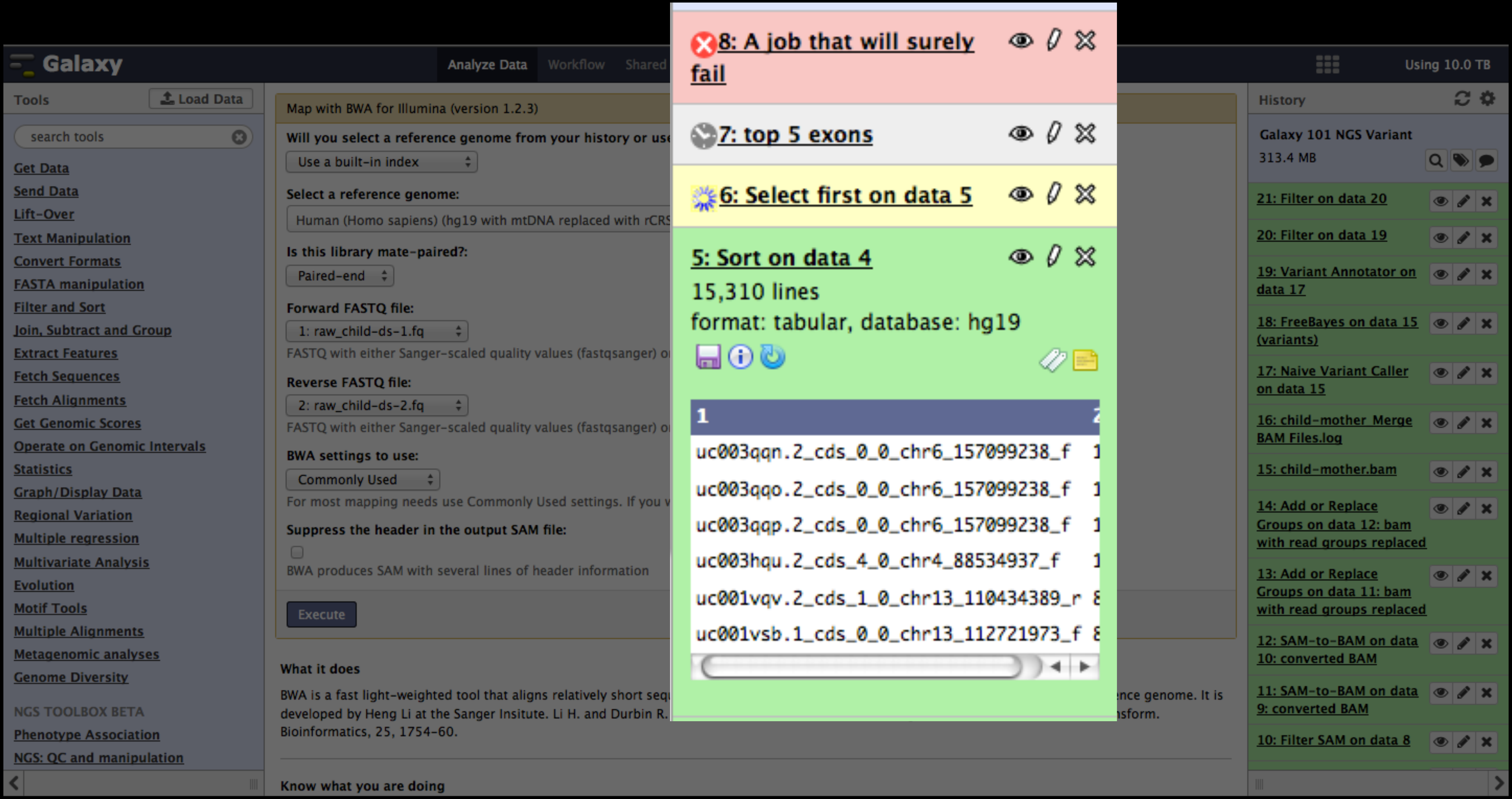
Suppress the header in the output SAM file:
☐
BWA produces SAM with several lines of header information

What it does

BWA is a fast light-weighted tool that aligns relatively short sequences (queries) to a sequence database (large), such as the human reference genome. It is developed by Heng Li at the Sanger Insitute. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-60.

Know what you are doing

Galaxy Analysis Workspace



Map with BWA for Illumina (version 1.2.3)

Will you select a reference genome from your history or use

Use a built-in index

Select a reference genome:

Human (Homo sapiens) (hg19 with mtDNA replaced with rCRS

Is this library mate-paired?:

Paired-end

Forward FASTQ file:

1: raw_child-ds-1.fq

FASTQ with either Sanger-scaled quality values (fastqsanger) o

Reverse FASTQ file:

2: raw_child-ds-2.fq

FASTQ with either Sanger-scaled quality values (fastqsanger) o

BWA settings to use:

Commonly Used

For most mapping needs use Commonly Used settings. If you v

Suppress the header in the output SAM file:

☐

BWA produces SAM with several lines of header information

What it does

BWA is a fast light-weighted tool that aligns relatively short seq … ence genome. It is developed by Heng Li at the Sanger Insitute. Li H. and Durbin R. … nsform. Bioinformatics, 25, 1754-60.

Know what you are doing

8: A job that will surely fail

7: top 5 exons

6: Select first on data 5

5: Sort on data 4

15,310 lines

format: tabular, database: hg19

1
uc003qqn.2_cds_0_0_chr6_157099238_f
uc003qqo.2_cds_0_0_chr6_157099238_f
uc003qqp.2_cds_0_0_chr6_157099238_f
uc003hqu.2_cds_4_0_chr4_88534937_f
uc001vqv.2_cds_1_0_chr13_110434389_r
uc001vsb.1_cds_0_0_chr13_112721973_f

Galaxy 101 NGS Variant

313.4 MB

21: Filter on data 20

20: Filter on data 19

19: Variant Annotator on data 17

18: FreeBayes on data 15 (variants)

17: Naive Variant Caller on data 15

16: child-mother Merge BAM Files.log

15: child-mother.bam

14: Add or Replace Groups on data 12: bam with read groups replaced

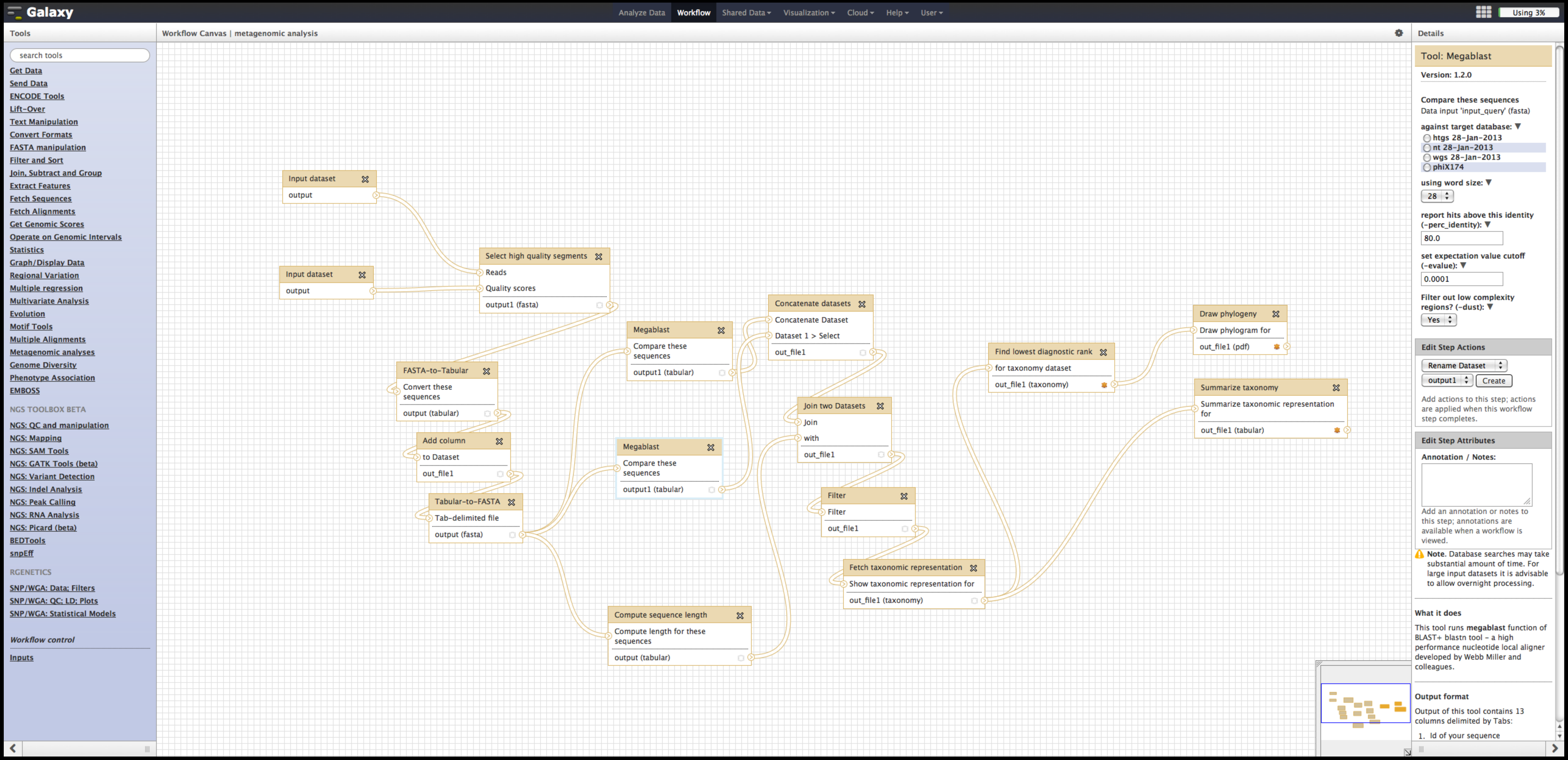
13: Add or Replace Groups on data 11: bam with read groups replaced

12: SAM-to-BAM on data 10: converted BAM

11: SAM-to-BAM on data 9: converted BAM

10: Filter SAM on data 8

Workflow Editor



Tool: Megablast

Version: 1.2.0

Compare these sequences
Data input 'input_query' (fasta)

against target database:
○ htgs 28-Jan-2013
○ nt 28-Jan-2013
○ wgs 28-Jan-2013
○ phiX174

using word size: 28

report hits above this identity (-perc_identity): 80.0

set expectation value cutoff (-evalue): 0.0001

Filter out low complexity regions? (-dust): Yes

Edit Step Actions

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Note. Database searches may take substantial amount of time. For large input datasets it is advisable to allow overnight processing.

What it does

This tool runs **megablast** function of BLAST+ blastn tool - a high performance nucleotide local aligner developed by Webb Miller and colleagues.

Output format

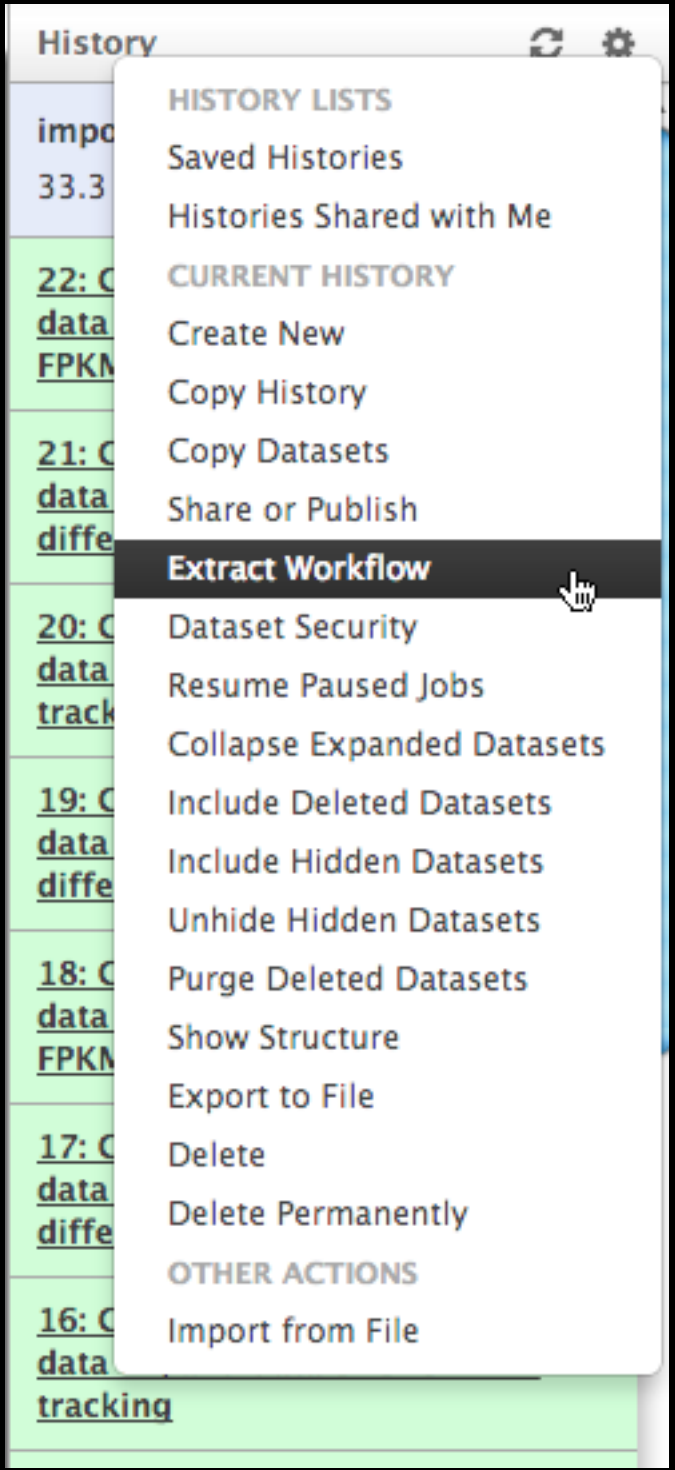
Output of this tool contains 13 columns delimited by Tabs:

1. Id of your sequence

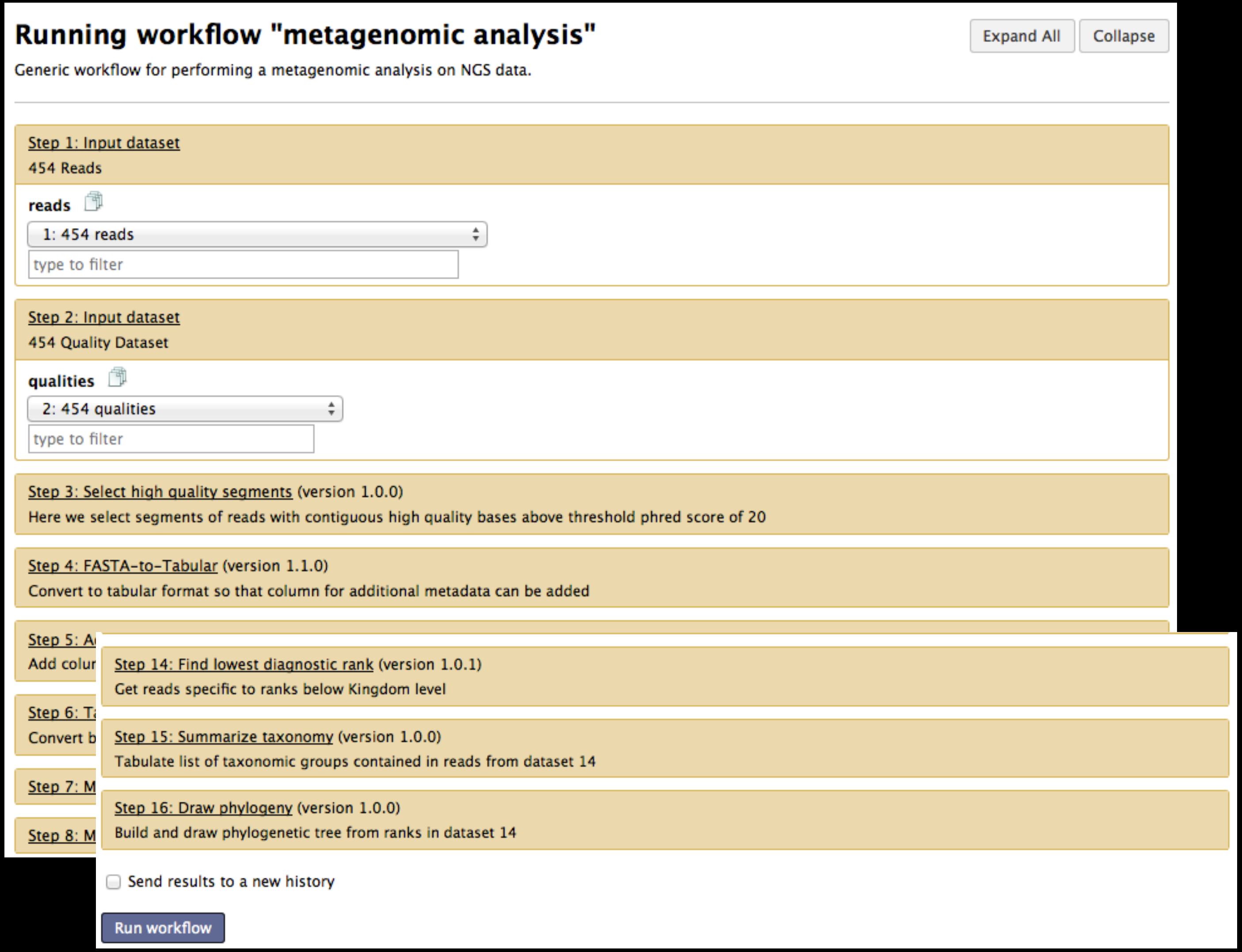
Create Workflow Automatically

Extract Workflow from History

Create a workflow from a History that you created interactively.



Run it



Running workflow "metagenomic analysis"

Generic workflow for performing a metagenomic analysis on NGS data.

Step 1: Input dataset
454 Reads

reads
1: 454 reads
type to filter

Step 2: Input dataset
454 Quality Dataset

qualities
2: 454 qualities
type to filter

Step 3: Select high quality segments (version 1.0.0)
Here we select segments of reads with contiguous high quality bases above threshold phred score of 20

Step 4: FASTA-to-Tabular (version 1.1.0)
Convert to tabular format so that column for additional metadata can be added

Step 5: A
Add colur

Step 6: T
Convert b

Step 7: M

Step 8: M

Step 14: Find lowest diagnostic rank (version 1.0.1)
Get reads specific to ranks below Kingdom level

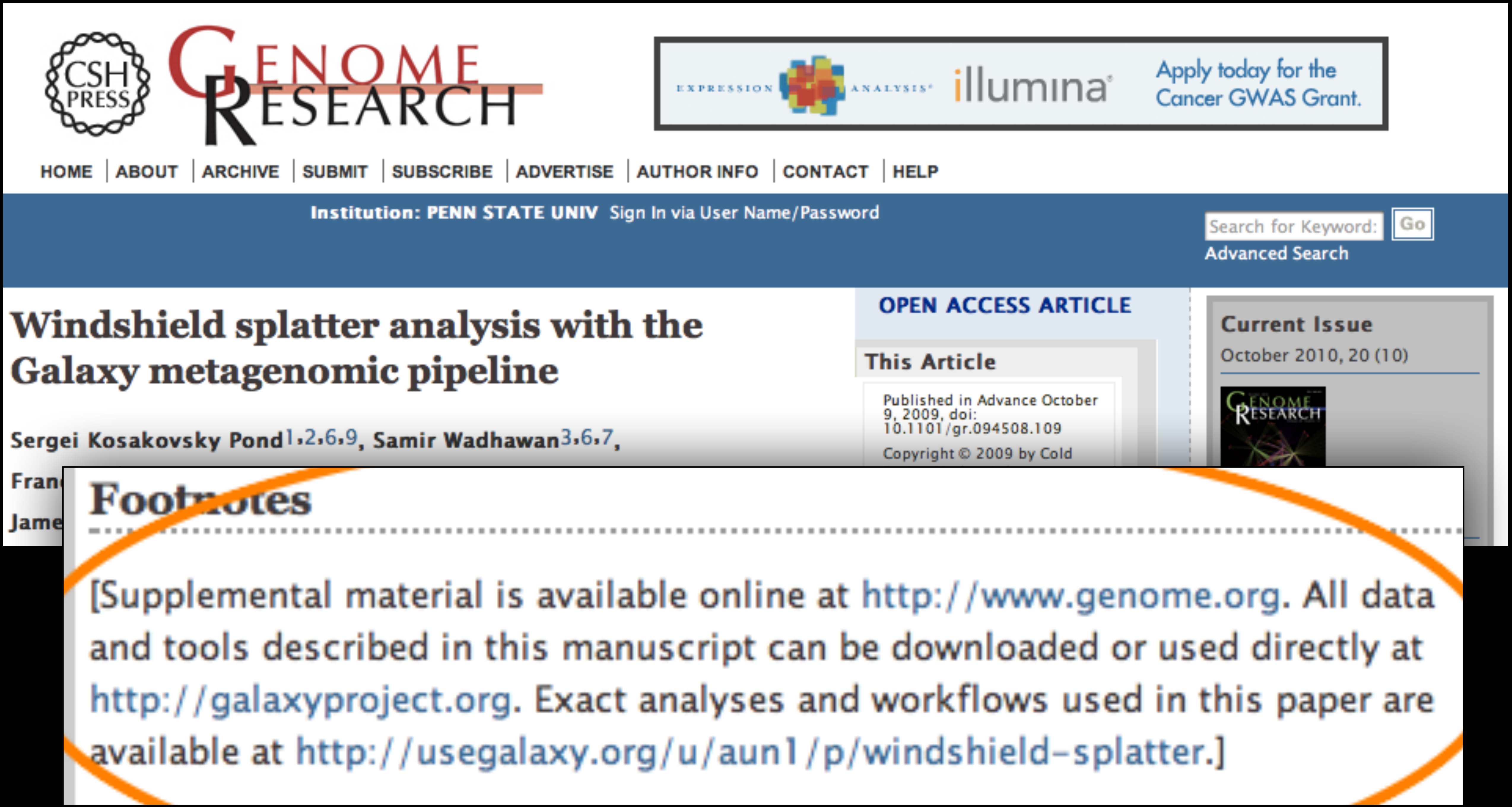
Step 15: Summarize taxonomy (version 1.0.0)
Tabulate list of taxonomic groups contained in reads from dataset 14

Step 16: Draw phylogeny (version 1.0.0)
Build and draw phylogenetic tree from ranks in dataset 14

☐ Send results to a new history

Run workflow

Sharing and Publishing Your Work



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

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], **Samir Wadhawan**[3,6,7],

Fran

Jame

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

Copyright © 2009 by Cold

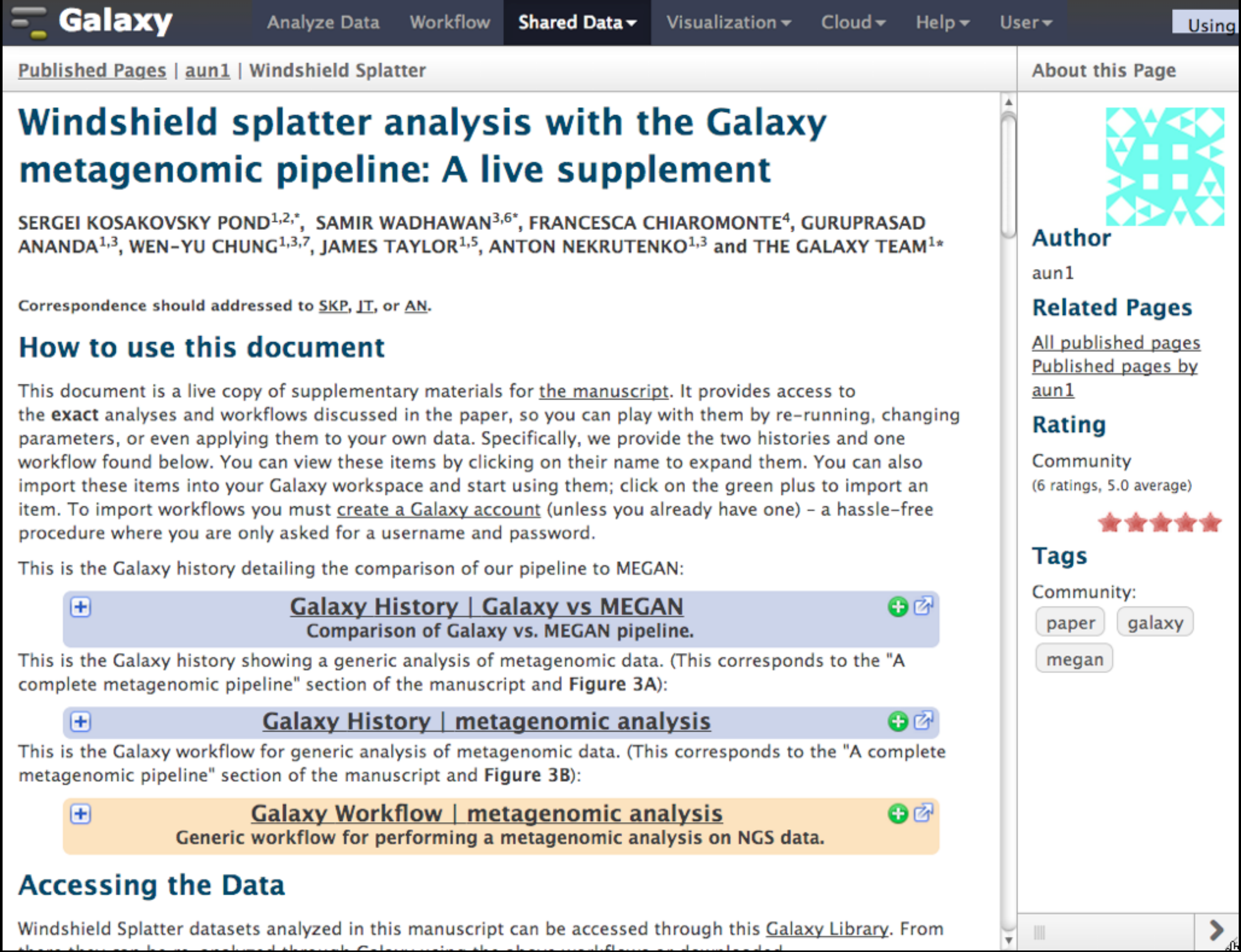
October 2010, 20 (10)

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.]

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

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 CHIAROMONTE^{4}, 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

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

Overview

What is Galaxy?

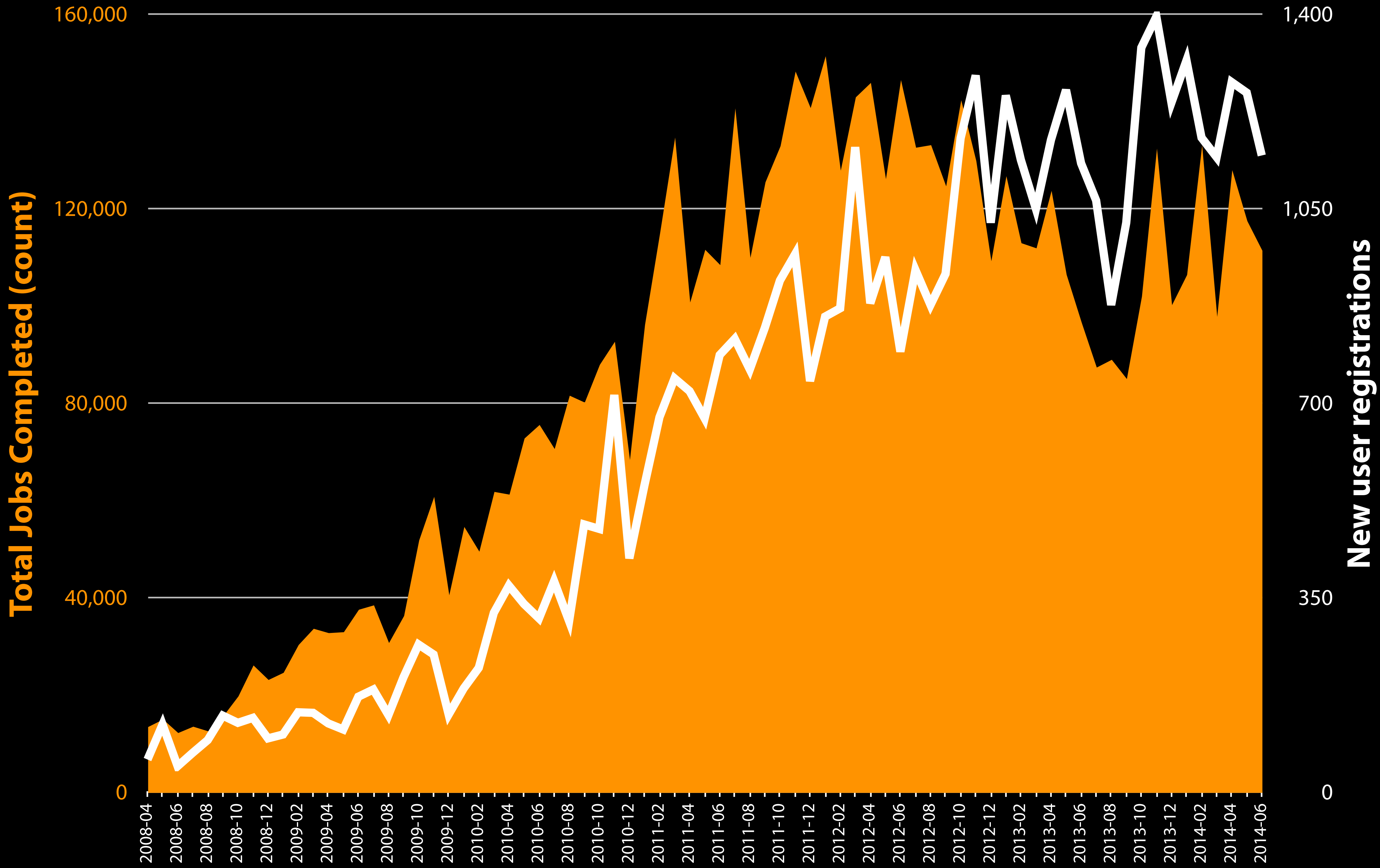
Usage / Contributions

Virtual Interactions

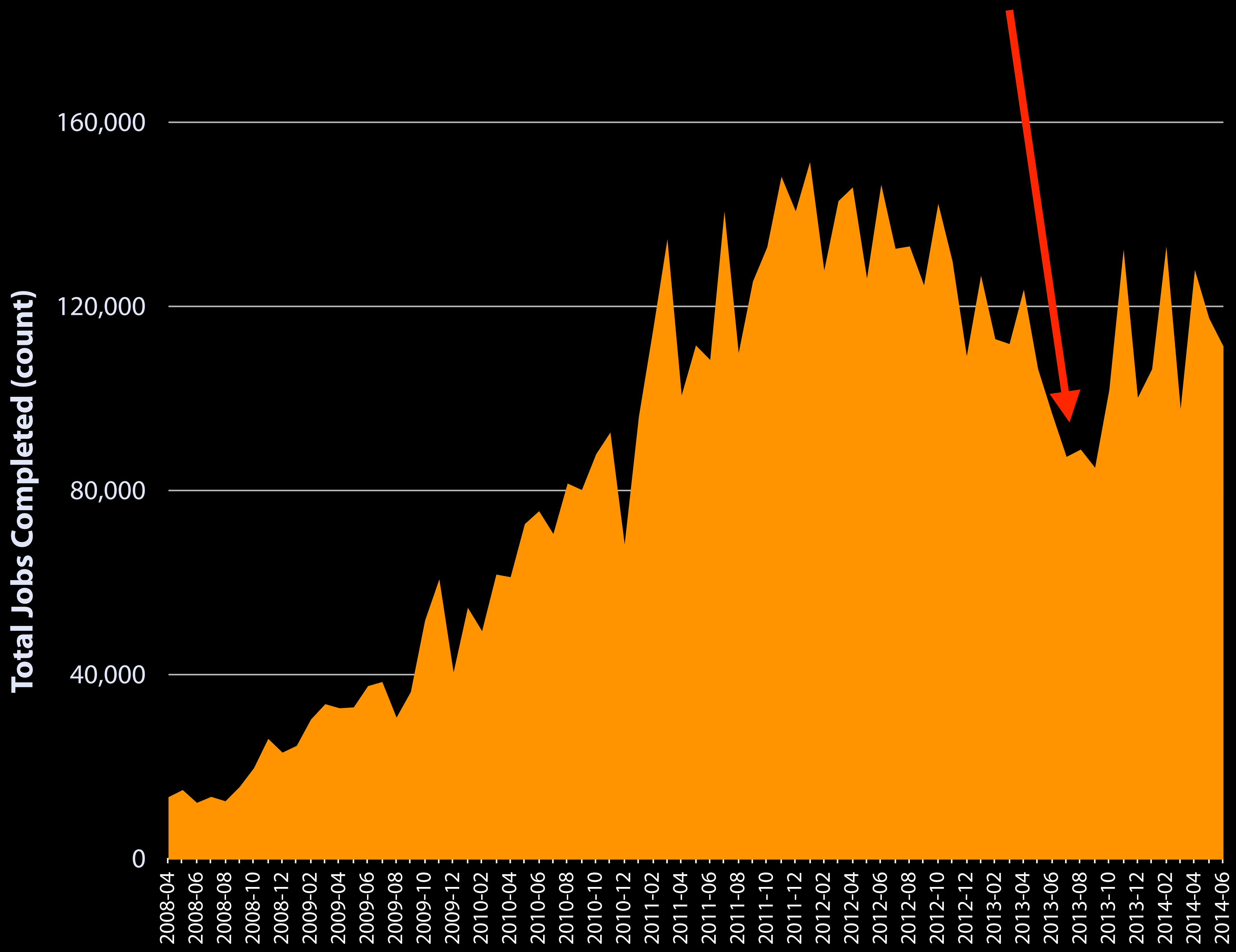
ToolShed

Events and Outreach

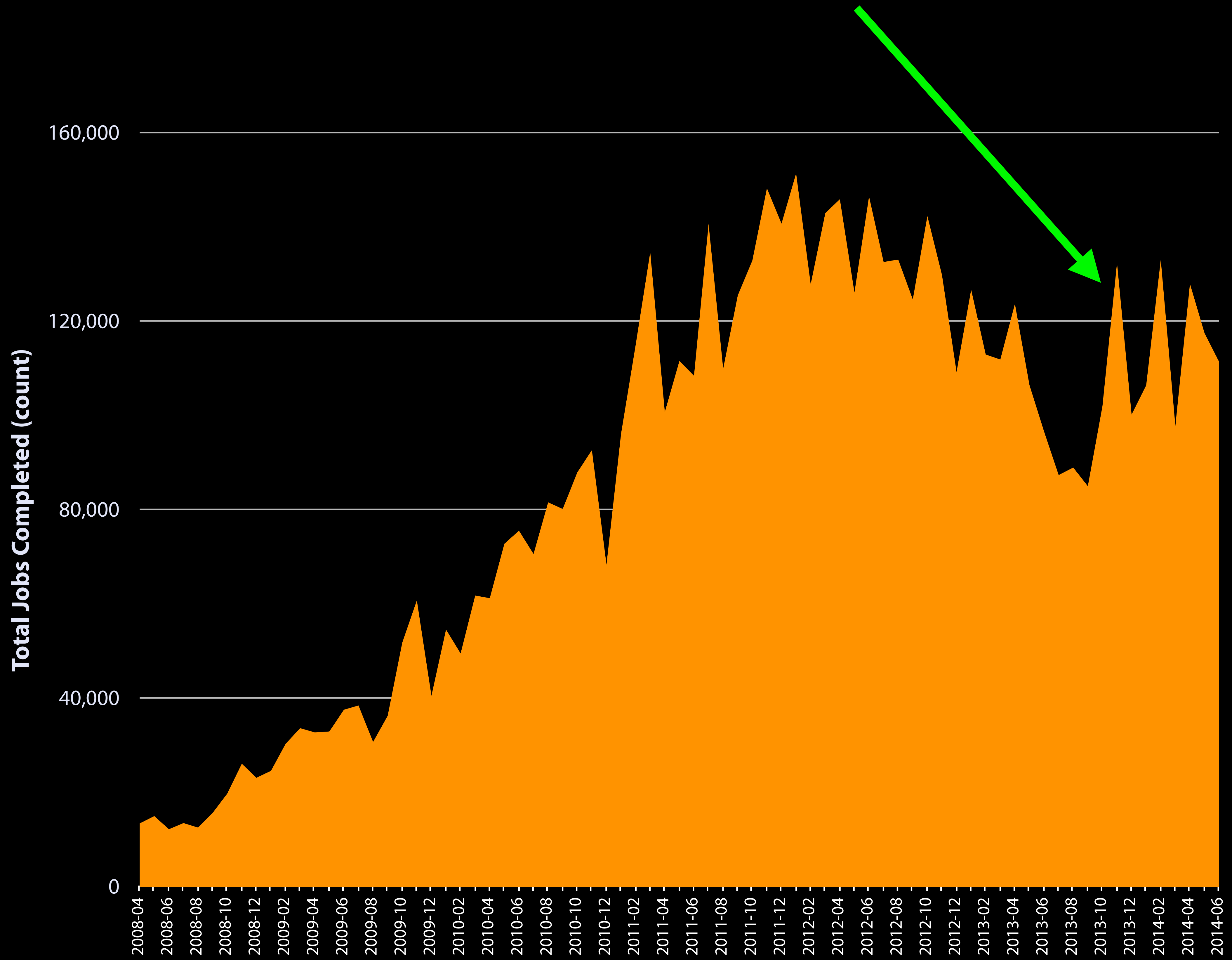
user dynamics at usegalaxy.org



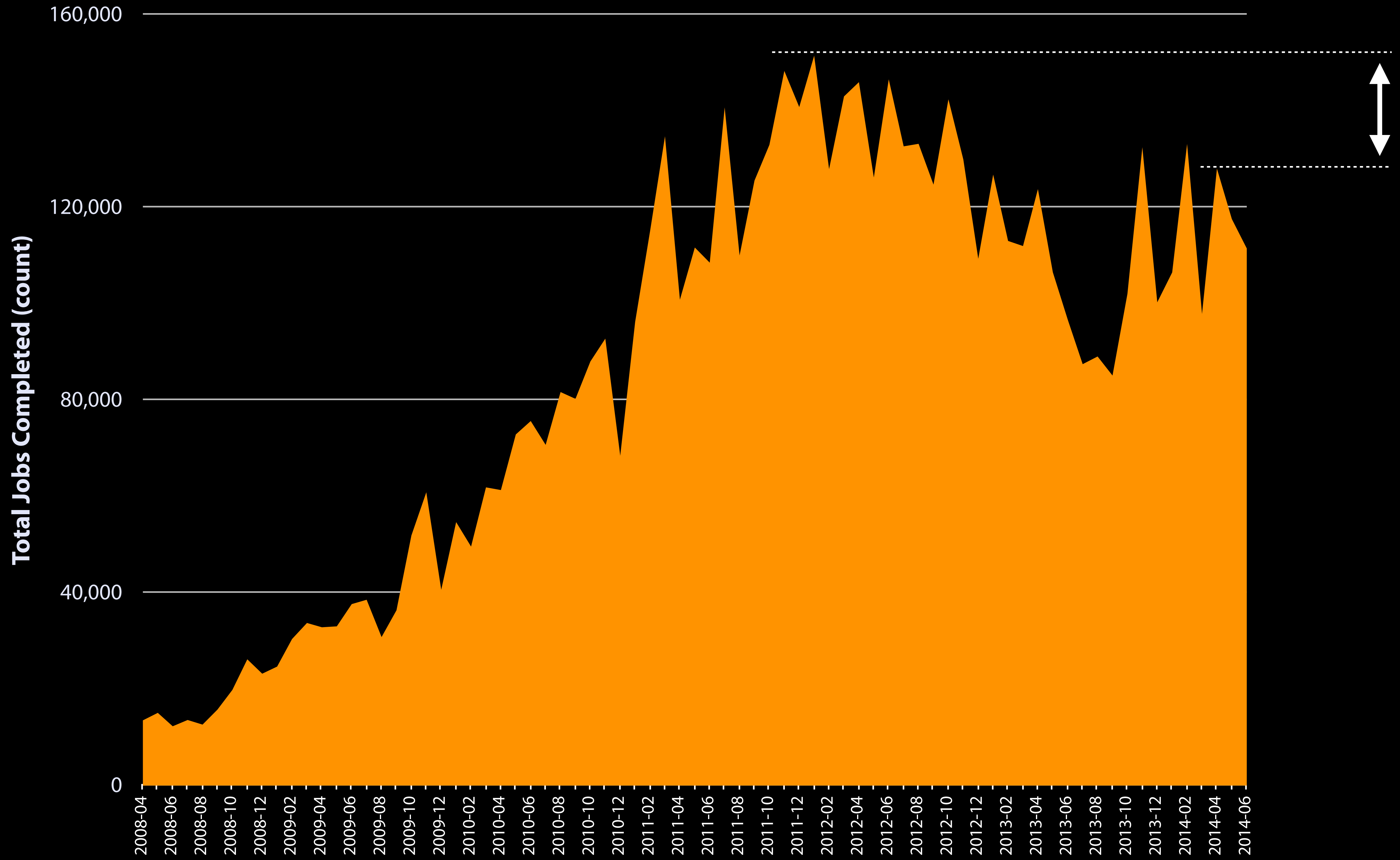
severe resource bottleneck



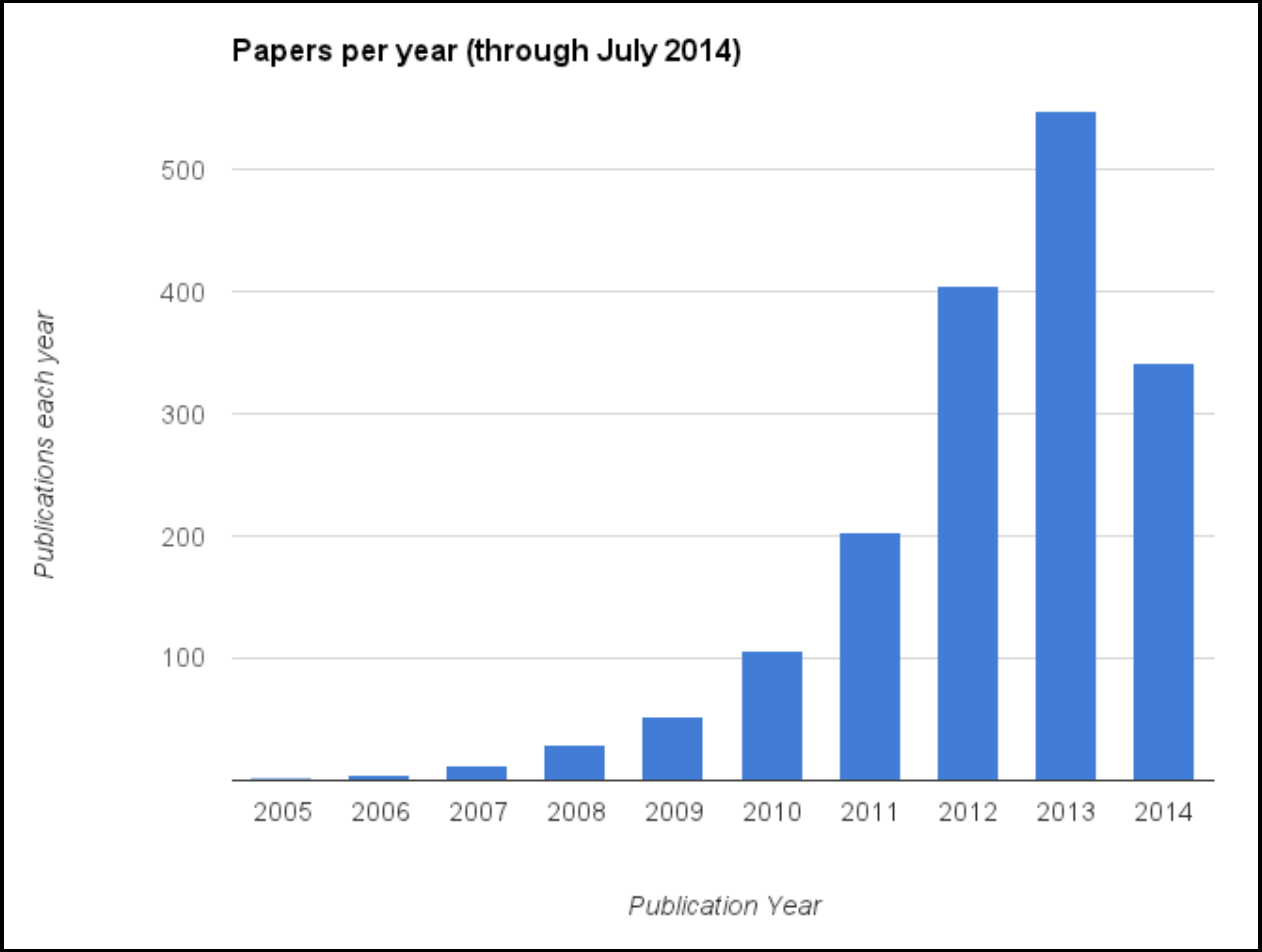
move to TACC


Total Jobs Completed (count)
160,000
120,000
80,000
40,000
0
2008-04
2008-06
2008-08
2008-10
2008-12
2009-02
2009-04
2009-06
2009-08
2009-10
2009-12
2010-02
2010-04
2010-06
2010-08
2010-10
2010-12
2011-02
2011-04
2011-06
2011-08
2011-10
2011-12
2012-02
2012-04
2012-06
2012-08
2012-10
2012-12
2013-02
2013-04
2013-06
2013-08
2013-10
2013-12
2014-02
2014-04
2014-06

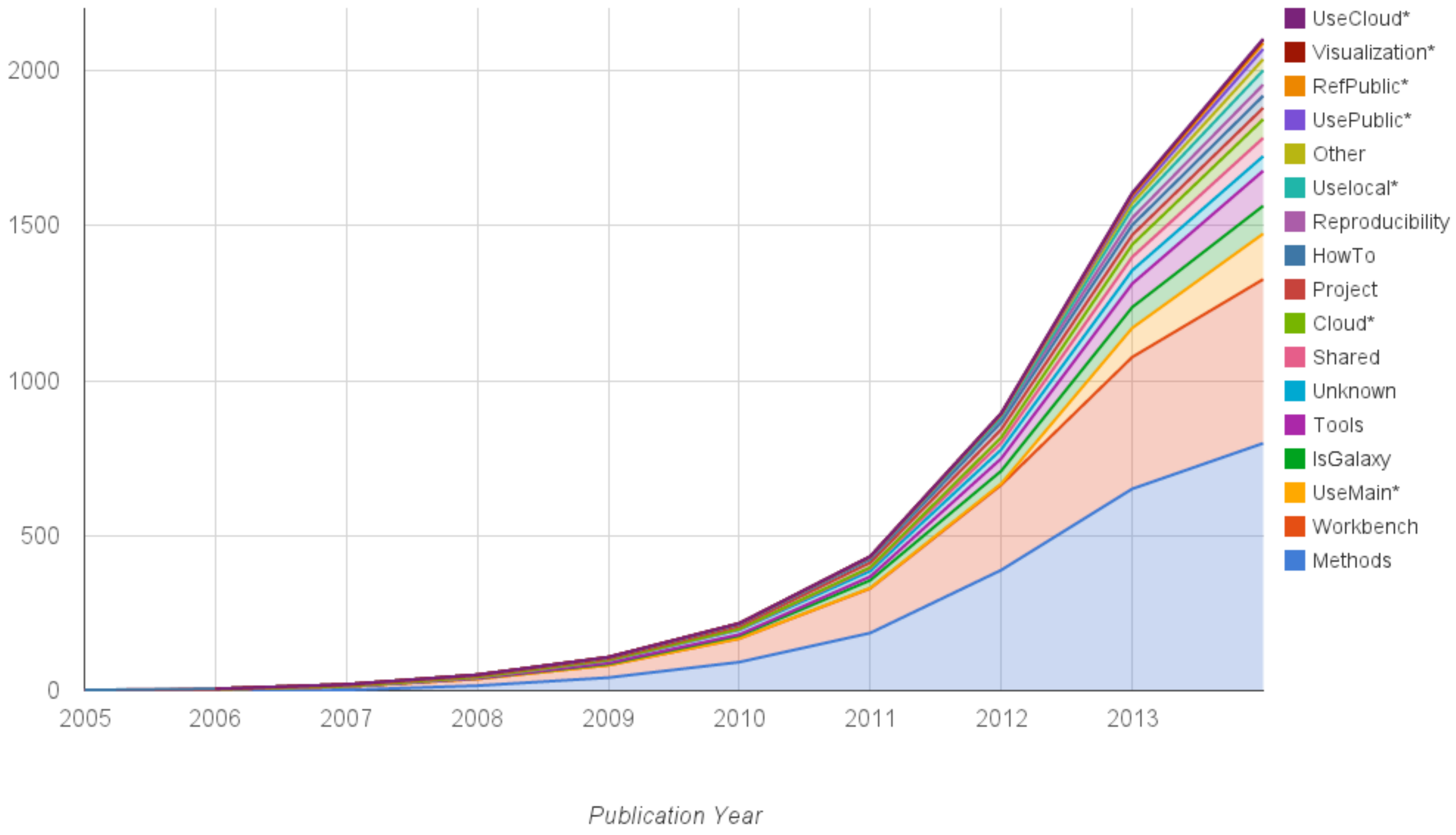
data is bigger = jobs are longer



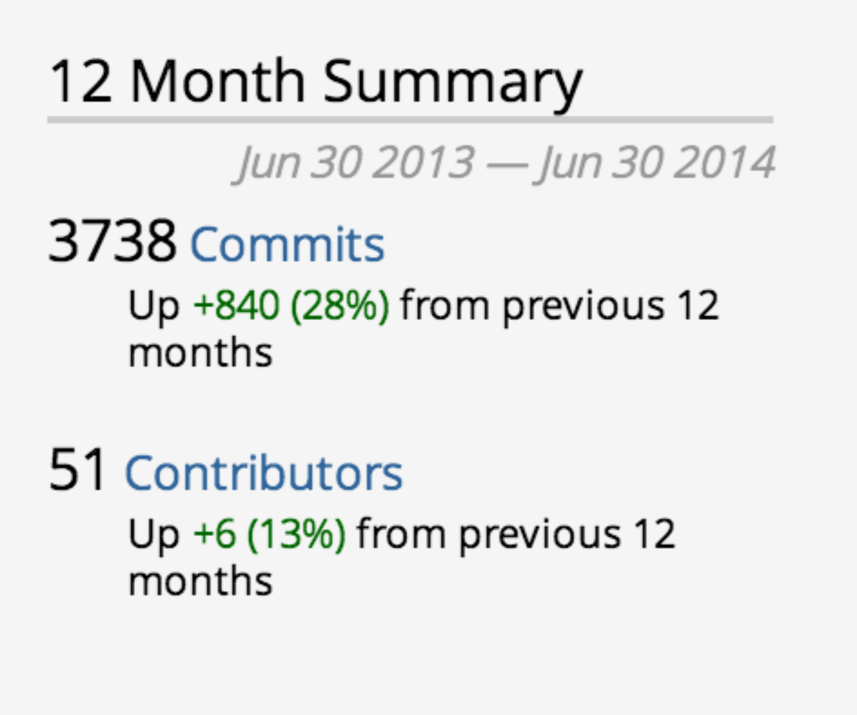
Citations / Publications



Publications: Cumulative Tags per Year (through July 2014)



Community Code Contributions



12 Month Summary

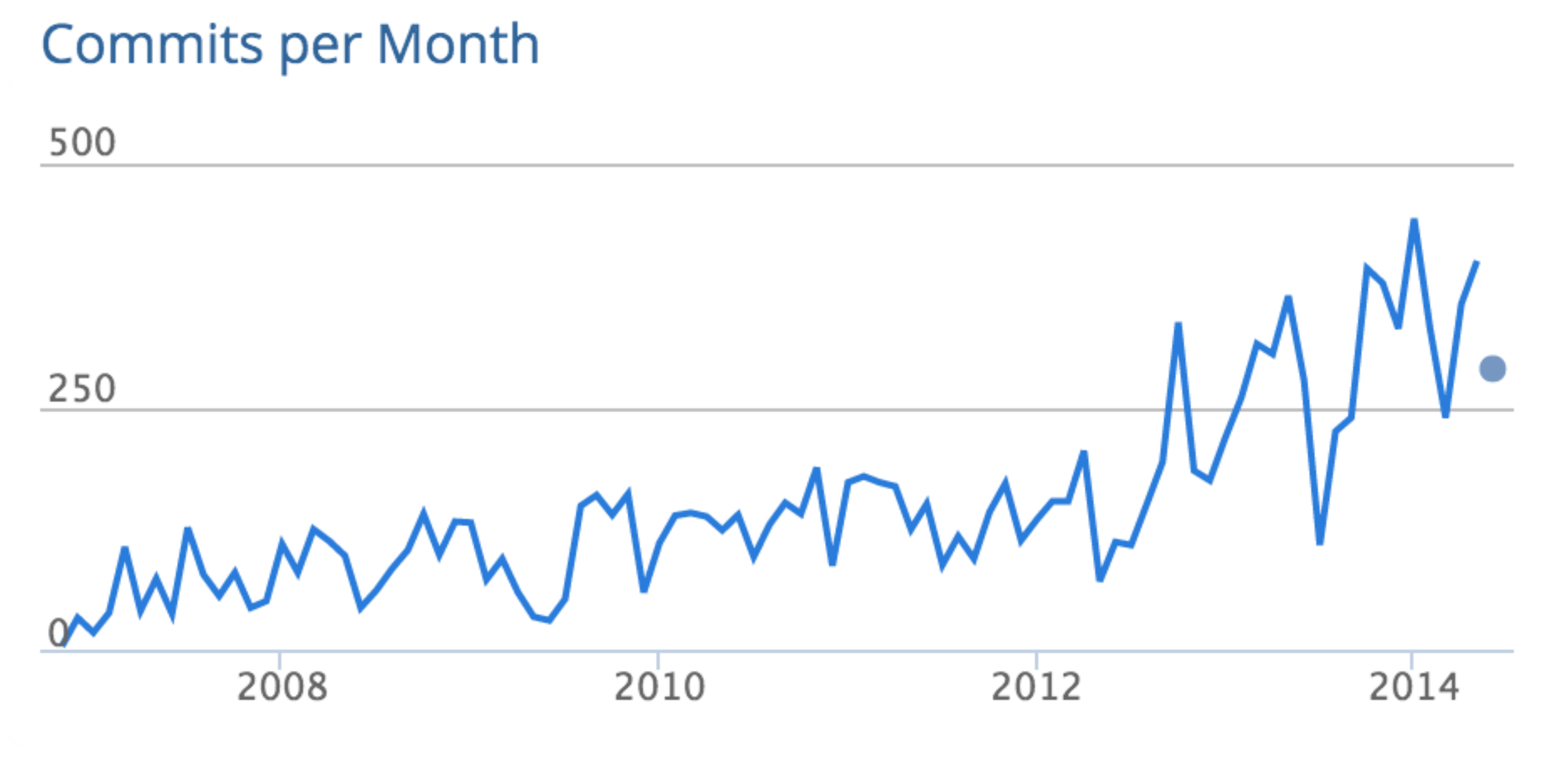
Jun 30 2013 — Jun 30 2014

3738 Commits

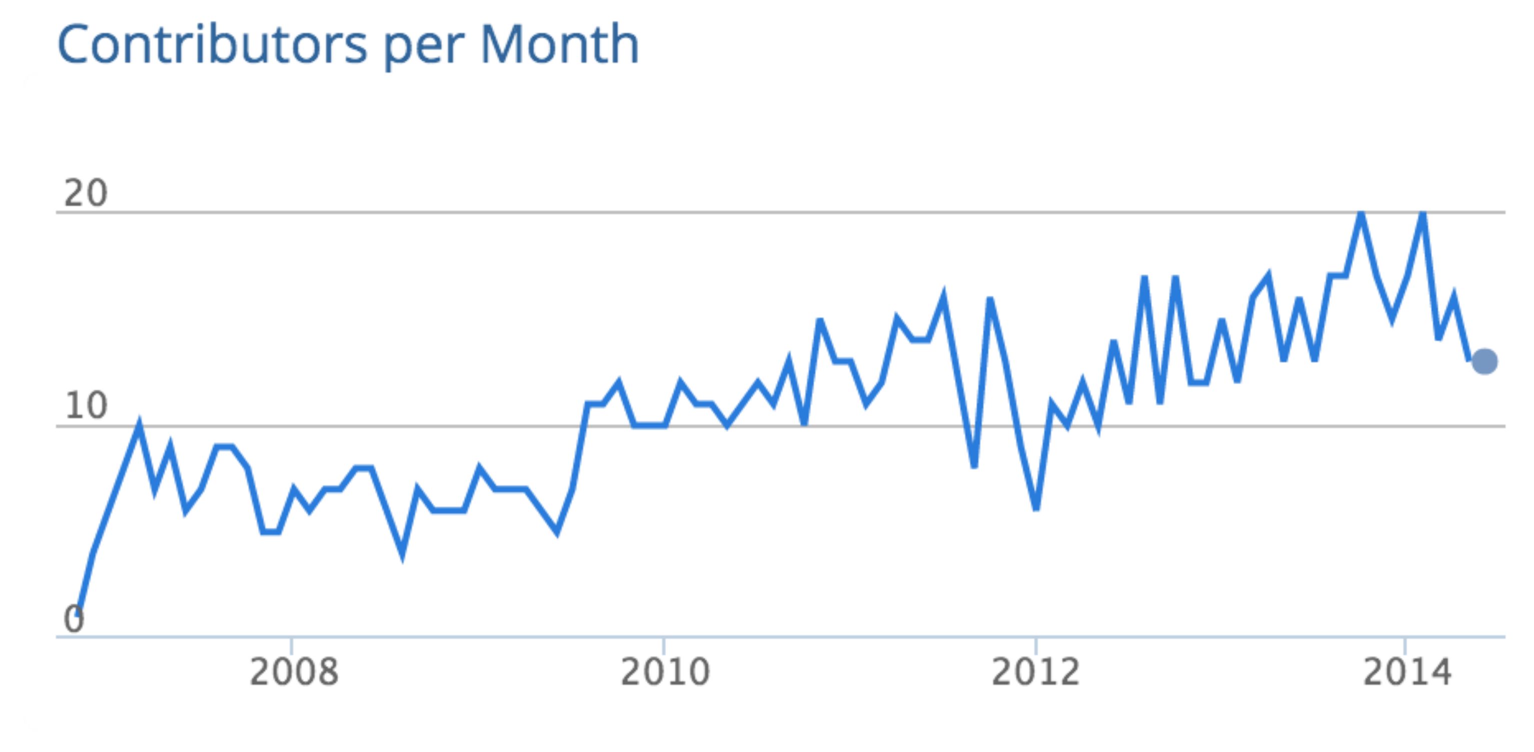
Up +840 (28%) from previous 12 months

51 Contributors

Up +6 (13%) from previous 12 months



Commits per Month



Contributors per Month

https://www.ohloh.net/p/galaxybx

Overview

What is Galaxy?

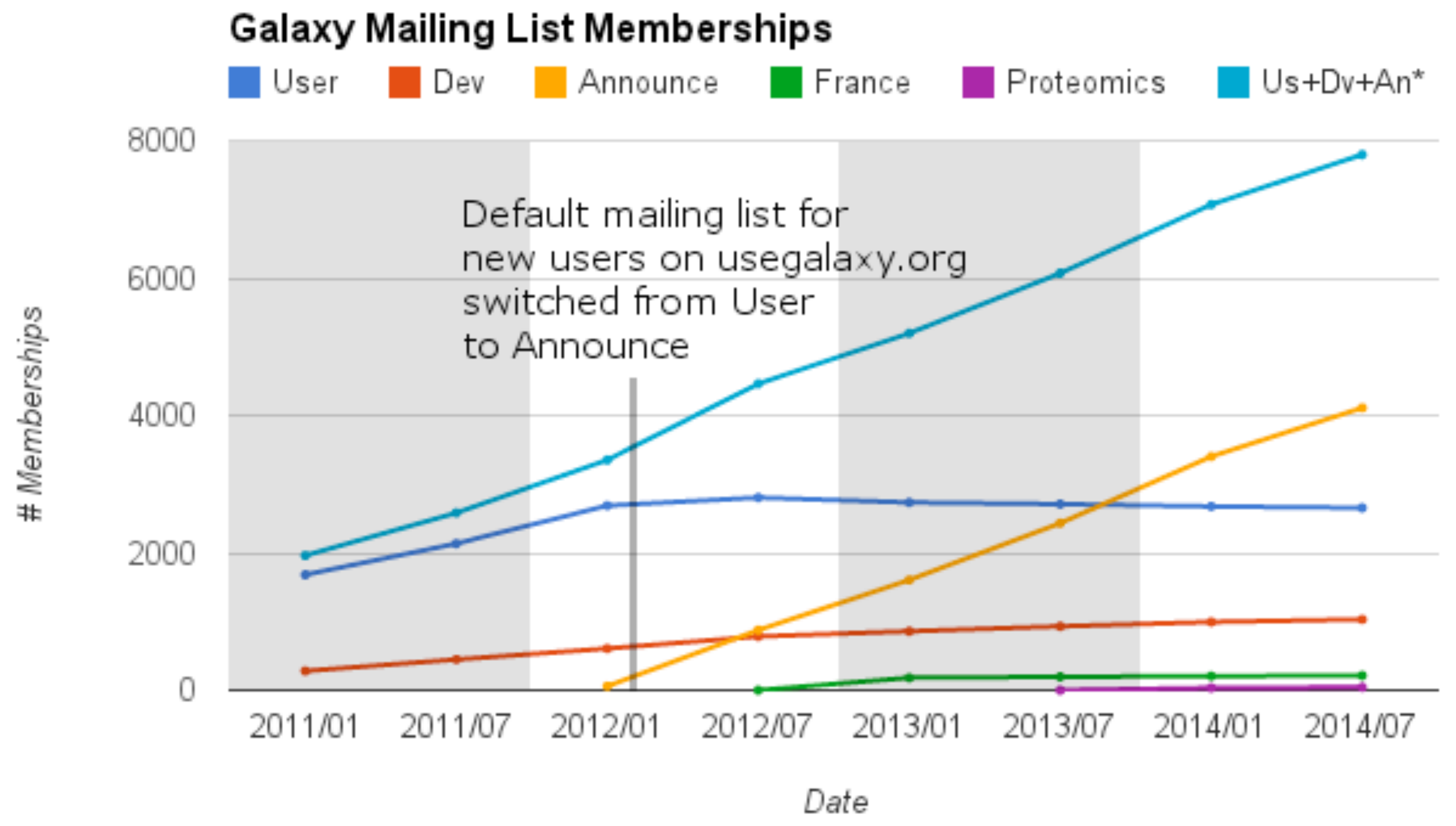
Usage / Contributions

Virtual Interactions

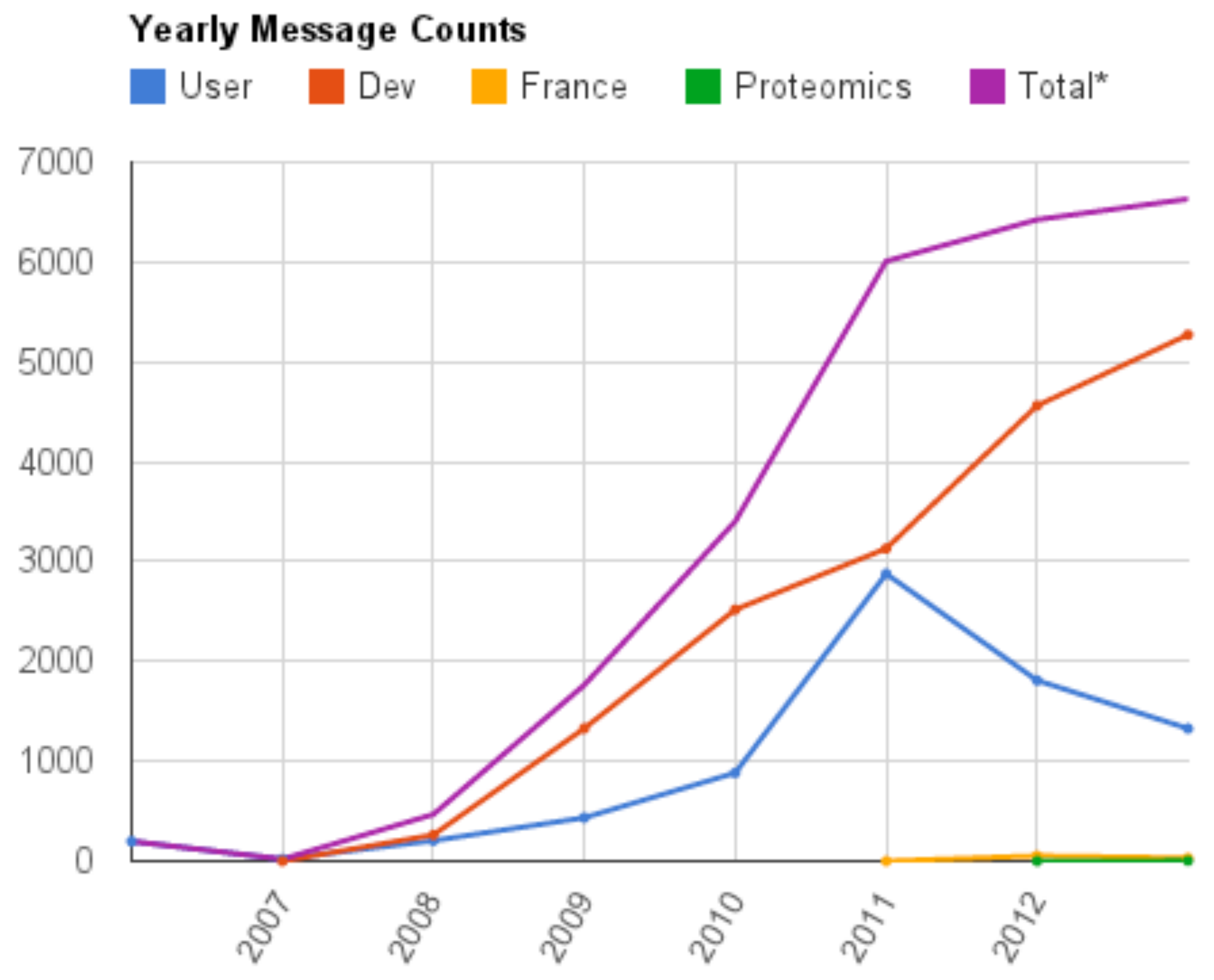
ToolShed

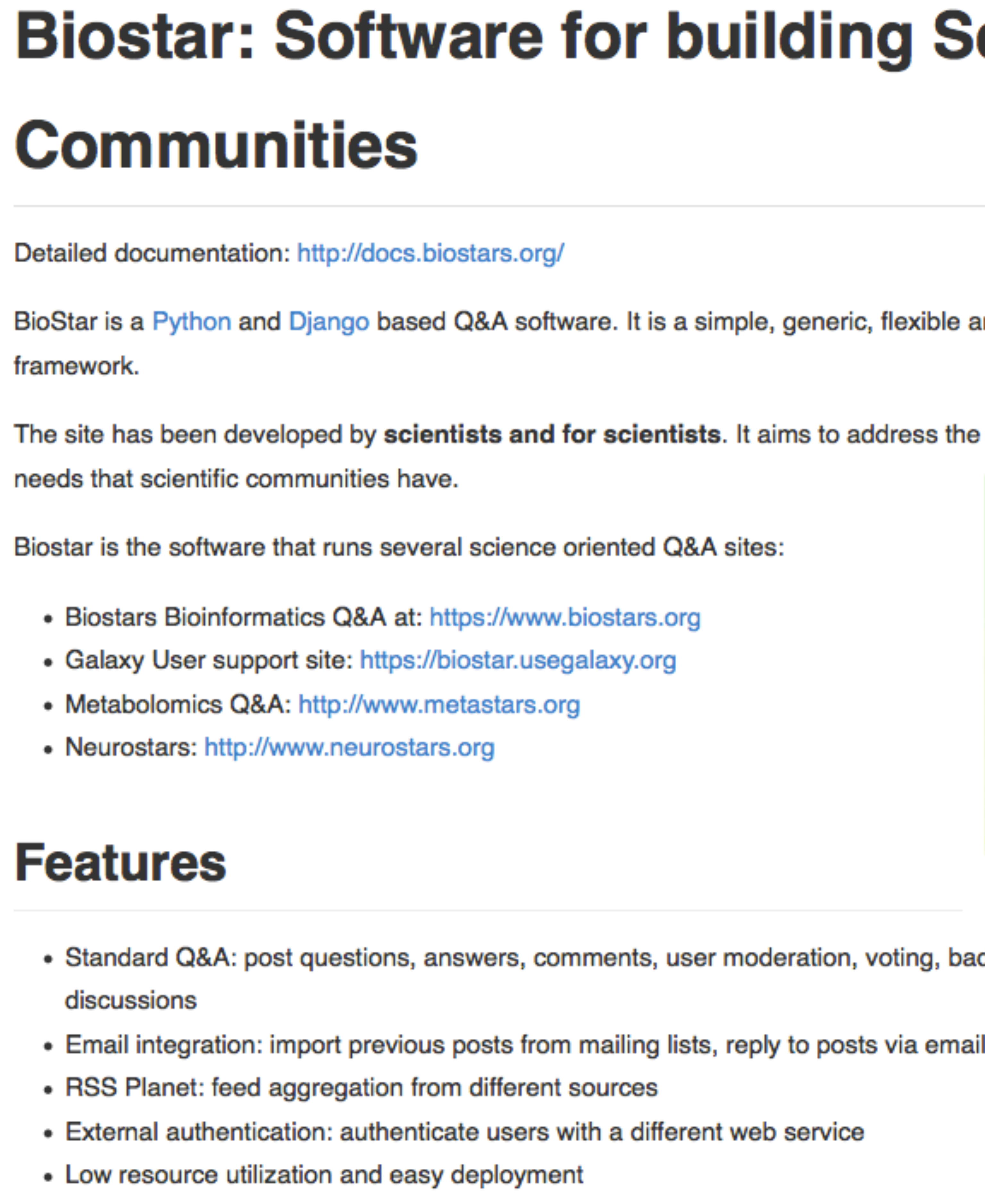
Events and Outreach

Mailing Lists



Mailing Lists





Biostar: Software for building Scientific Communities

Detailed documentation: http://docs.biostars.org/

BioStar is a Python and Django based Q&A software. It is a simple, generic, flexible and extensible Q&A framework.

The site has been developed by **scientists and for scientists**. It aims to address the requirements and needs that scientific communities have.

Biostar is the software that runs several science oriented Q&A sites:

- Biostars Bioinformatics Q&A at: https://www.biostars.org
- Galaxy User support site: https://biostar.usegalaxy.org
- Metabolomics Q&A: http://www.metastars.org
- Neurostars: http://www.neurostars.org

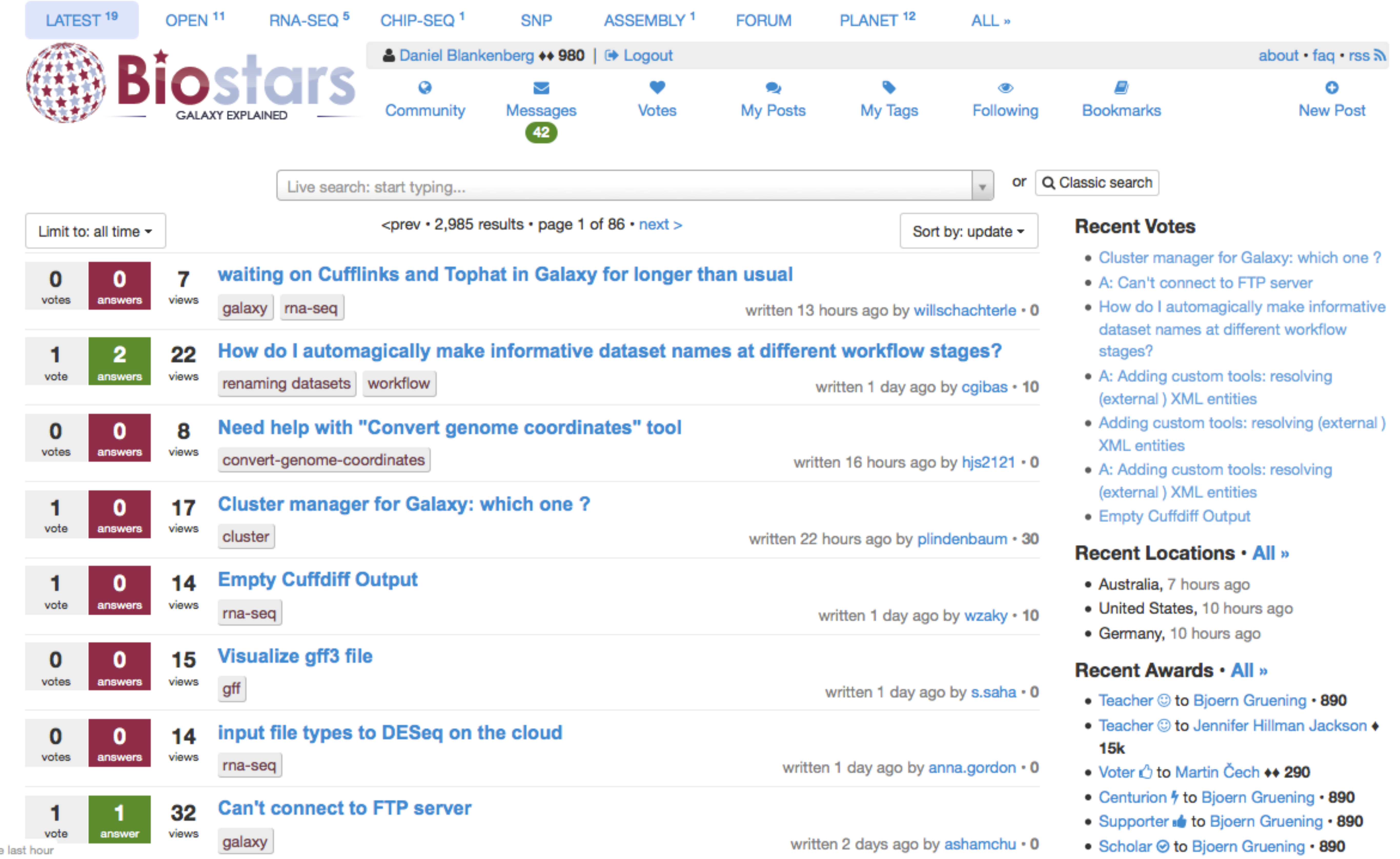
Features

- Standard Q&A: post questions, answers, comments, user moderation, voting, badges, threaded discussions
- Email integration: import previous posts from mailing lists, reply to posts via email
- RSS Planet: feed aggregation from different sources
- External authentication: authenticate users with a different web service
- Low resource utilization and easy deployment

Istvan Albert

https://github.com/ialbert/biostar-central

Biostar



<prev • 2,985 results • page 1 of 86 • next >

0 votes · 0 answers · 7 views
waiting on Cufflinks and Tophat in Galaxy for longer than usual
galaxy · rna-seq
written 13 hours ago by willschachterle • 0

1 vote · 2 answers · 22 views
How do I automagically make informative dataset names at different workflow stages?
renaming datasets · workflow
written 1 day ago by cgibas • 10

0 votes · 0 answers · 8 views
Need help with "Convert genome coordinates" tool
convert-genome-coordinates
written 16 hours ago by hjs2121 • 0

1 vote · 0 answers · 17 views
Cluster manager for Galaxy: which one ?
cluster
written 22 hours ago by plindenbaum • 30

1 vote · 0 answers · 14 views
Empty Cuffdiff Output
rna-seq
written 1 day ago by wzaky • 10

0 votes · 0 answers · 15 views
Visualize gff3 file
gff
written 1 day ago by s.saha • 0

0 votes · 0 answers · 14 views
input file types to DESeq on the cloud
rna-seq
written 1 day ago by anna.gordon • 0

1 vote · 1 answer · 32 views
Can't connect to FTP server
galaxy
written 2 days ago by ashamchu • 0

e last hour

Recent Votes

- Cluster manager for Galaxy: which one ?
- A: Can't connect to FTP server
- How do I automagically make informative dataset names at different workflow stages?
- A: Adding custom tools: resolving (external) XML entities
- Adding custom tools: resolving (external) XML entities
- A: Adding custom tools: resolving (external) XML entities
- Empty Cuffdiff Output

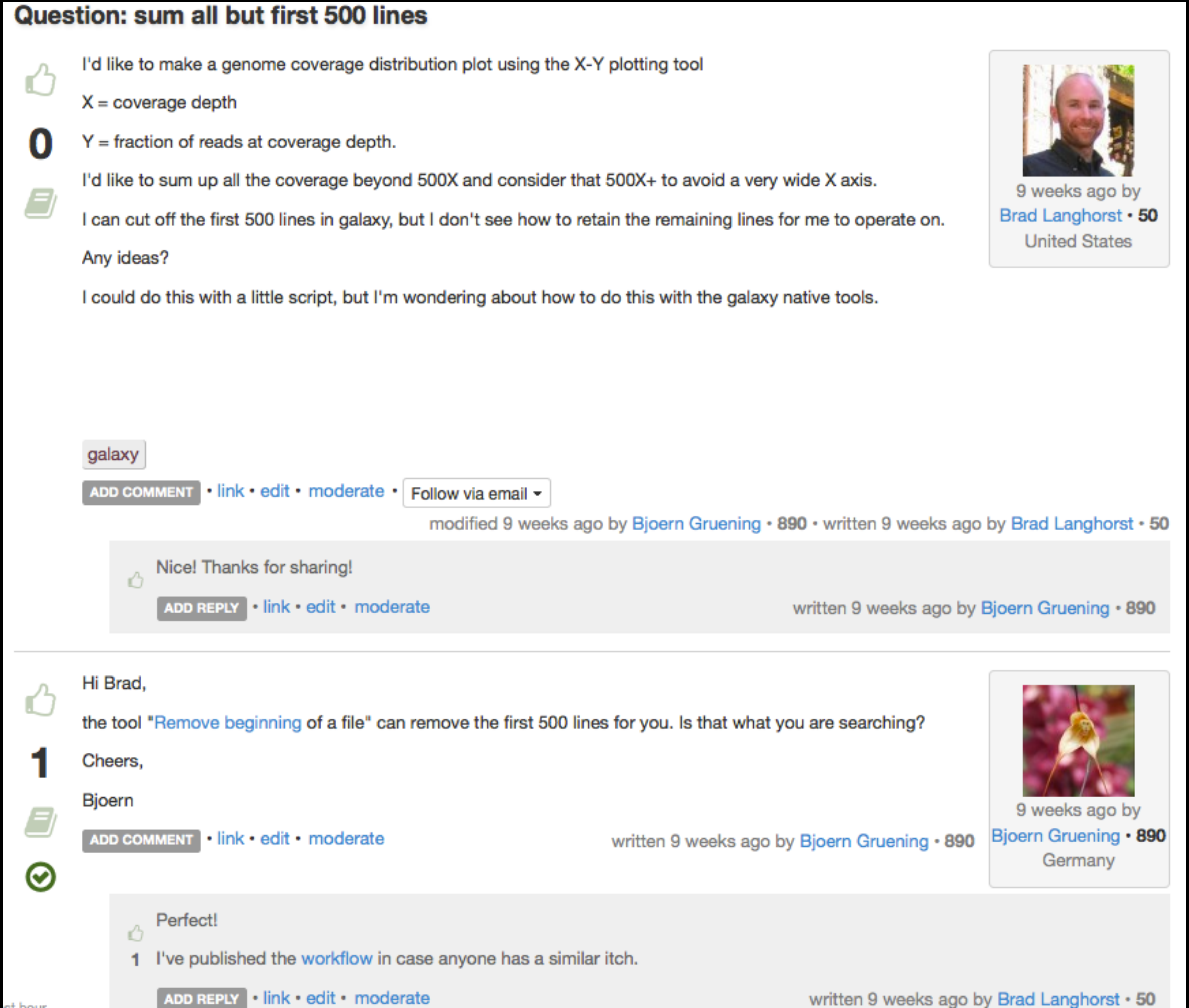
Recent Locations • All »

- Australia, 7 hours ago
- United States, 10 hours ago
- Germany, 10 hours ago

Recent Awards • All »

- Teacher ☺ to Bjoern Gruening • 890
- Teacher ☺ to Jennifer Hillman Jackson ♦ 15k
- Voter 👍 to Martin Čech ♦♦ 290
- Centurion ϟ to Bjoern Gruening • 890
- Supporter 👍 to Bjoern Gruening • 890
- Scholar ✓ to Bjoern Gruening • 890

https://biostar.usegalaxy.org/



Question: sum all but first 500 lines

0

I'd like to make a genome coverage distribution plot using the X-Y plotting tool

X = coverage depth

Y = fraction of reads at coverage depth.

I'd like to sum up all the coverage beyond 500X and consider that 500X+ to avoid a very wide X axis.

I can cut off the first 500 lines in galaxy, but I don't see how to retain the remaining lines for me to operate on.

Any ideas?

I could do this with a little script, but I'm wondering about how to do this with the galaxy native tools.

9 weeks ago by
Brad Langhorst • **50**
United States

modified 9 weeks ago by Bjoern Gruening • **890** • written 9 weeks ago by Brad Langhorst • **50**

Nice! Thanks for sharing!

written 9 weeks ago by Bjoern Gruening • **890**

1

Hi Brad,

the tool "Remove beginning of a file" can remove the first 500 lines for you. Is that what you are searching?

Cheers,

Bjoern

written 9 weeks ago by Bjoern Gruening • **890**

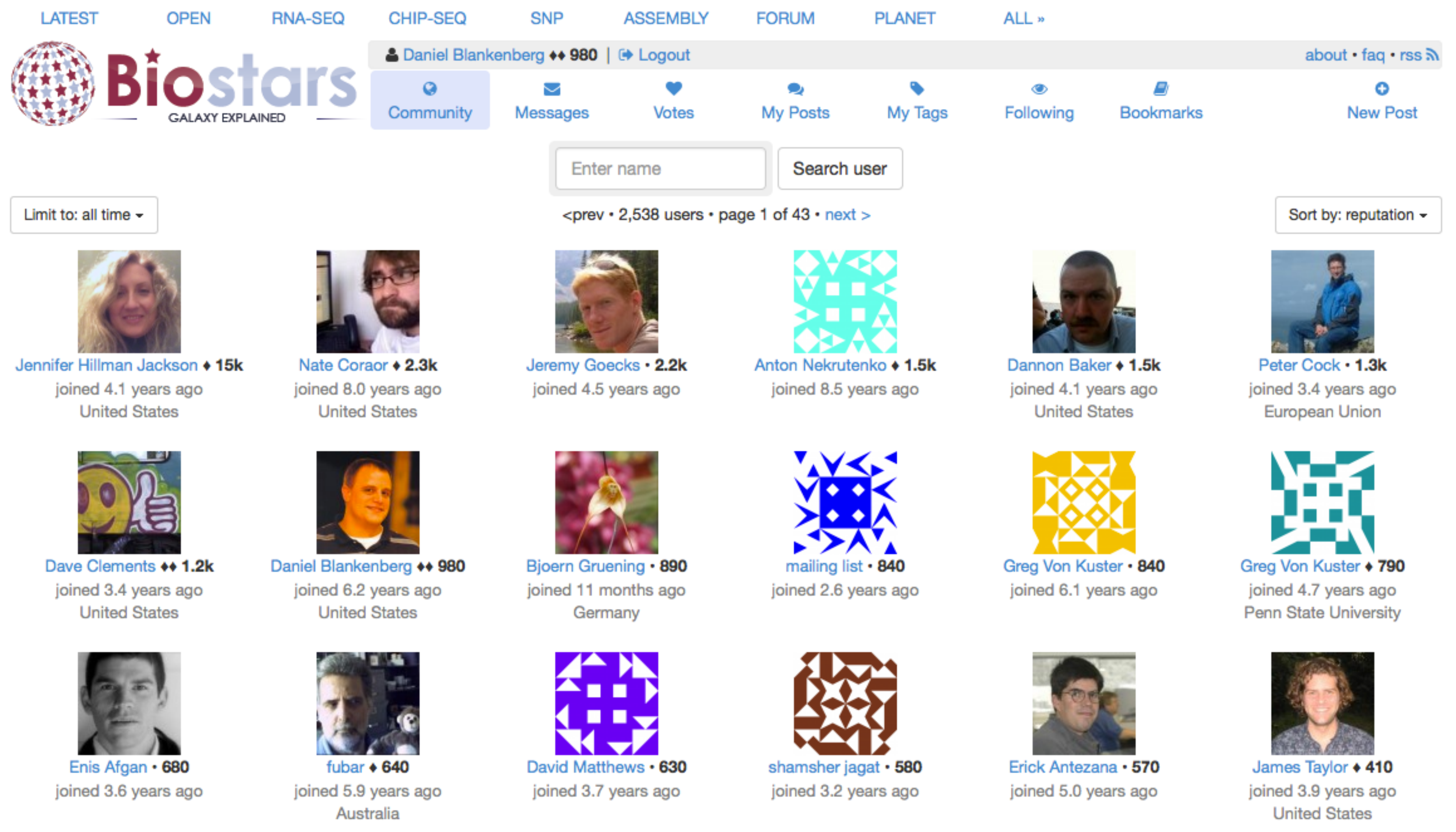
9 weeks ago by
Bjoern Gruening • **890**
Germany

Perfect!

1 I've published the workflow in case anyone has a similar itch.

written 9 weeks ago by Brad Langhorst • **50**

Gamification



<prev • 2,538 users • page 1 of 43 • next >

Jennifer Hillman Jackson ♦ 15k
joined 4.1 years ago
United States

Nate Coraor ♦ 2.3k
joined 8.0 years ago
United States

Jeremy Goecks • 2.2k
joined 4.5 years ago

Anton Nekrutenko ♦ 1.5k
joined 8.5 years ago

Dannon Baker ♦ 1.5k
joined 4.1 years ago
United States

Peter Cock • 1.3k
joined 3.4 years ago
European Union

Dave Clements ♦♦ 1.2k
joined 3.4 years ago
United States

Daniel Blankenberg ♦♦ 980
joined 6.2 years ago
United States

Bjoern Gruening • 890
joined 11 months ago
Germany

mailing list • 840
joined 2.6 years ago

Greg Von Kuster • 840
joined 6.1 years ago

Greg Von Kuster ♦ 790
joined 4.7 years ago
Penn State University

Enis Afgan • 680
joined 3.6 years ago

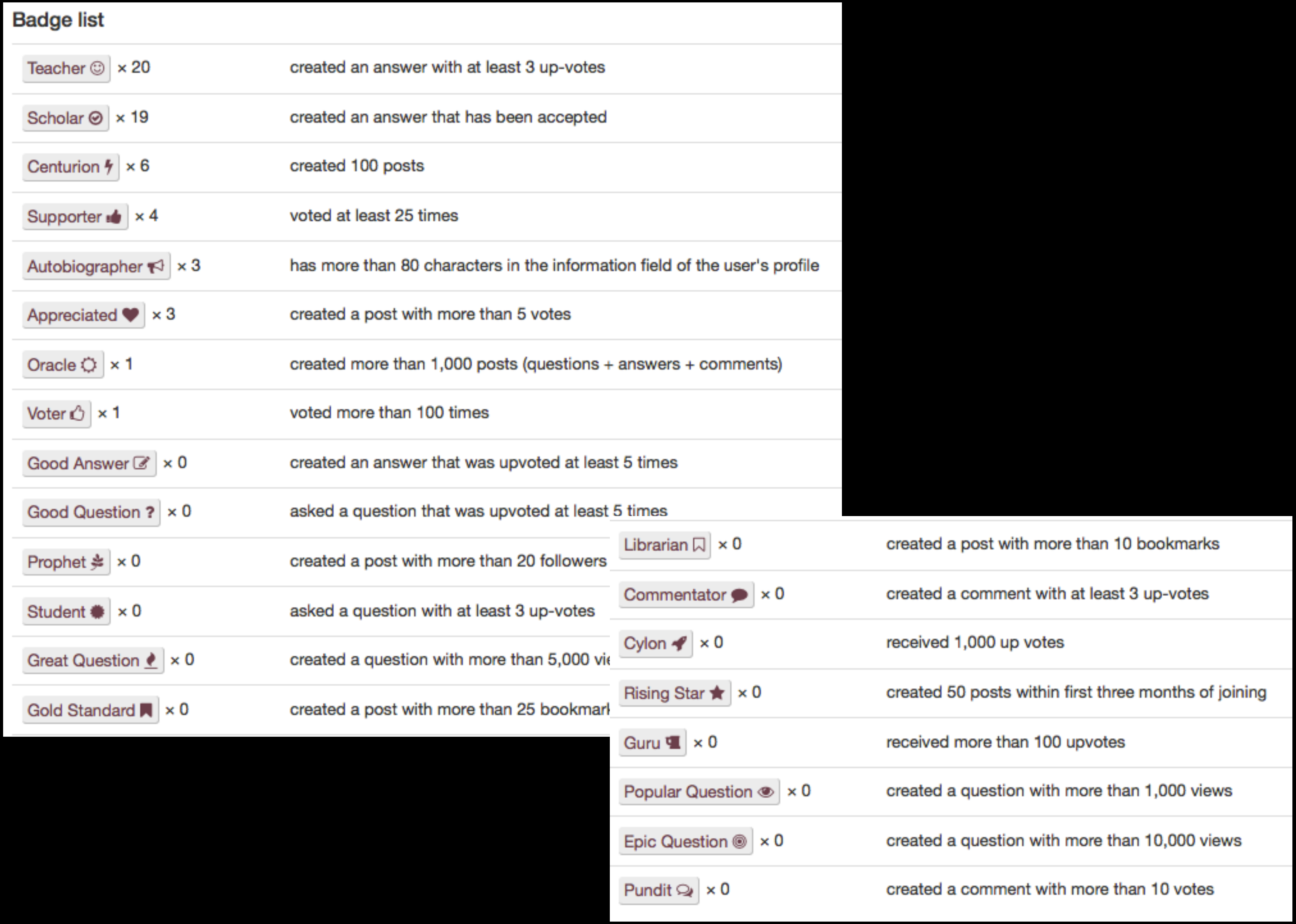
fubar ♦ 640
joined 5.9 years ago
Australia

David Matthews • 630
joined 3.7 years ago

shamsher jagat • 580
joined 3.2 years ago

Erick Antezana • 570
joined 5.0 years ago

James Taylor ♦ 410
joined 3.9 years ago
United States

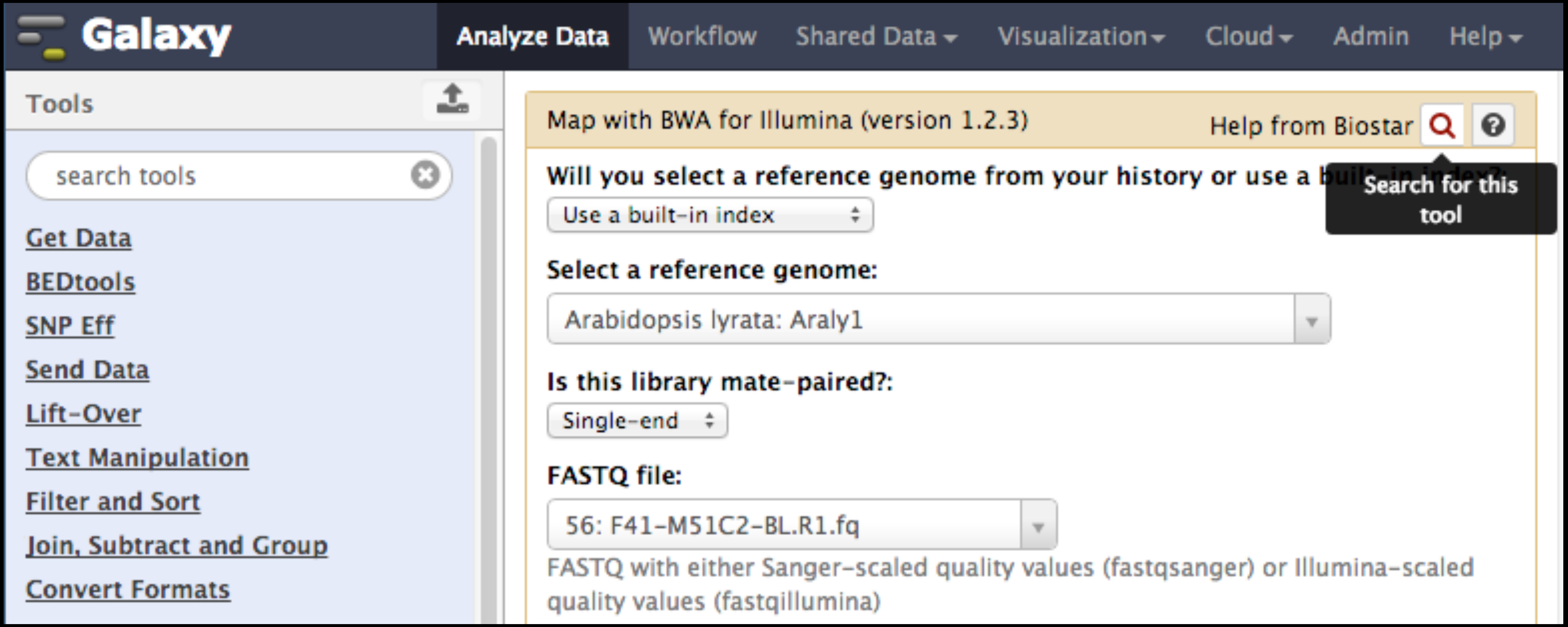


Badge list

Teacher	× 20	created an answer with at least 3 up-votes
Scholar	× 19	created an answer that has been accepted
Centurion	× 6	created 100 posts
Supporter	× 4	voted at least 25 times
Autobiographer	× 3	has more than 80 characters in the information field of the user's profile
Appreciated	× 3	created a post with more than 5 votes
Oracle	× 1	created more than 1,000 posts (questions + answers + comments)
Voter	× 1	voted more than 100 times
Good Answer	× 0	created an answer that was upvoted at least 5 times
Good Question	× 0	asked a question that was upvoted at least 5 times
Prophet	× 0	created a post with more than 20 followers
Student	× 0	asked a question with at least 3 up-votes
Great Question	× 0	created a question with more than 5,000 vi
Gold Standard	× 0	created a post with more than 25 bookmark

Librarian	× 0	created a post with more than 10 bookmarks
Commentator	× 0	created a comment with at least 3 up-votes
Cylon	× 0	received 1,000 up votes
Rising Star	× 0	created 50 posts within first three months of joining
Guru	× 0	received more than 100 upvotes
Popular Question	× 0	created a question with more than 1,000 views
Epic Question	× 0	created a question with more than 10,000 views
Pundit	× 0	created a comment with more than 10 votes

Accessing Biostar through Galaxy



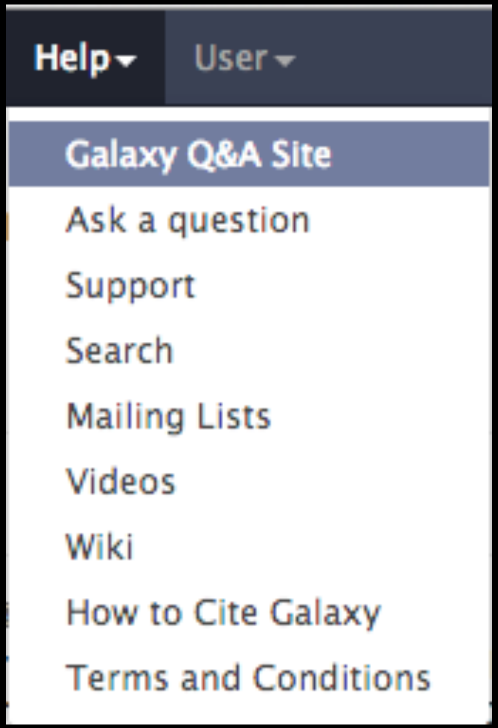
Map with BWA for Illumina (version 1.2.3)

Will you select a reference genome from your history or use a built-in index?:
Use a built-in index

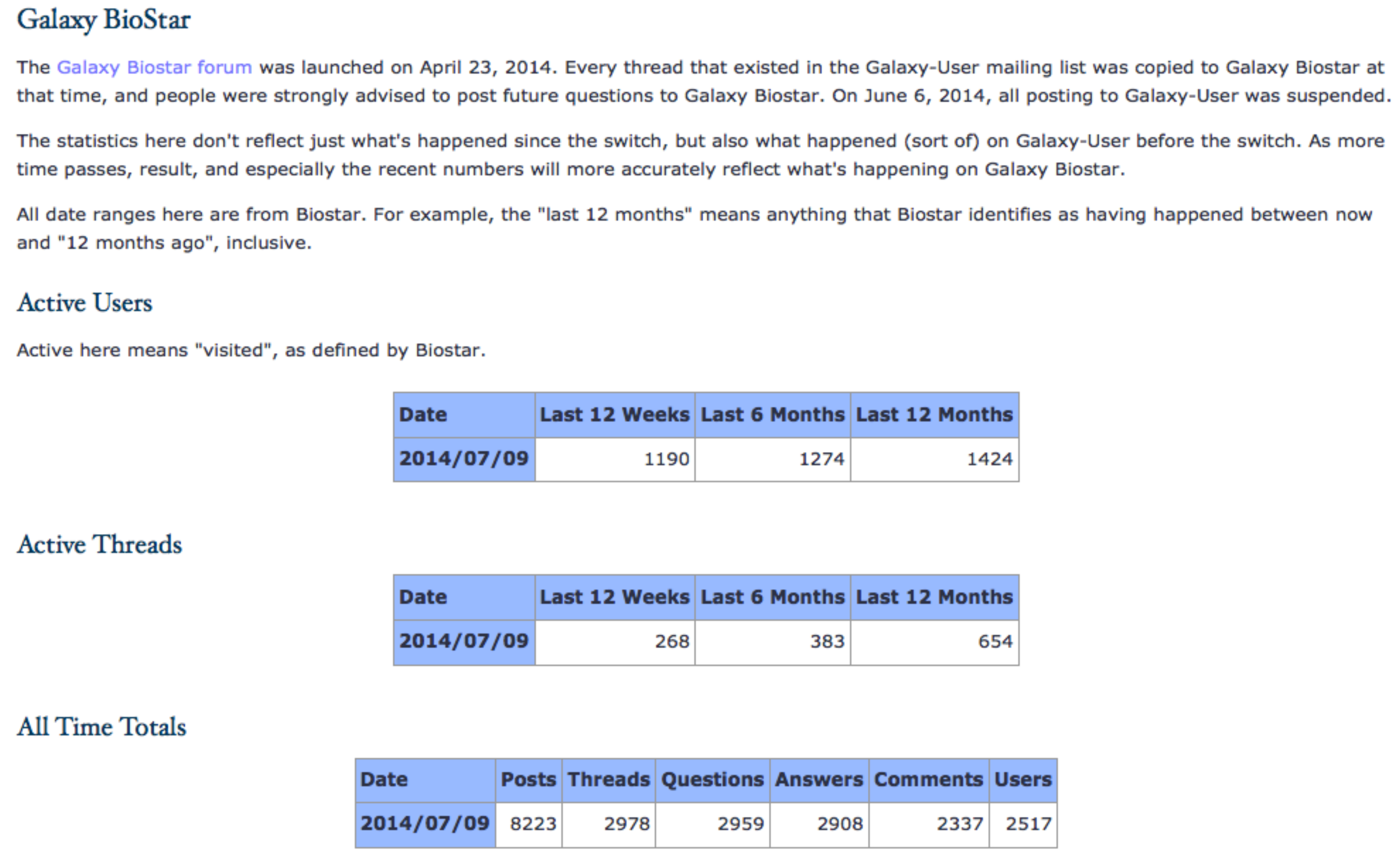
Select a reference genome:
Arabidopsis lyrata: Araly1

Is this library mate-paired?:
Single-end

FASTQ file:
56: F41-M51C2-BL.R1.fq
FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)



Registered Galaxy Users are automatically logged in



Galaxy BioStar

The Galaxy Biostar forum was launched on April 23, 2014. Every thread that existed in the Galaxy-User mailing list was copied to Galaxy Biostar at that time, and people were strongly advised to post future questions to Galaxy Biostar. On June 6, 2014, all posting to Galaxy-User was suspended.

The statistics here don't reflect just what's happened since the switch, but also what happened (sort of) on Galaxy-User before the switch. As more time passes, result, and especially the recent numbers will more accurately reflect what's happening on Galaxy Biostar.

All date ranges here are from Biostar. For example, the "last 12 months" means anything that Biostar identifies as having happened between now and "12 months ago", inclusive.

Active Users

Active here means "visited", as defined by Biostar.

Date	Last 12 Weeks	Last 6 Months	Last 12 Months
2014/07/09	1190	1274	1424

Active Threads

Date	Last 12 Weeks	Last 6 Months	Last 12 Months
2014/07/09	268	383	654

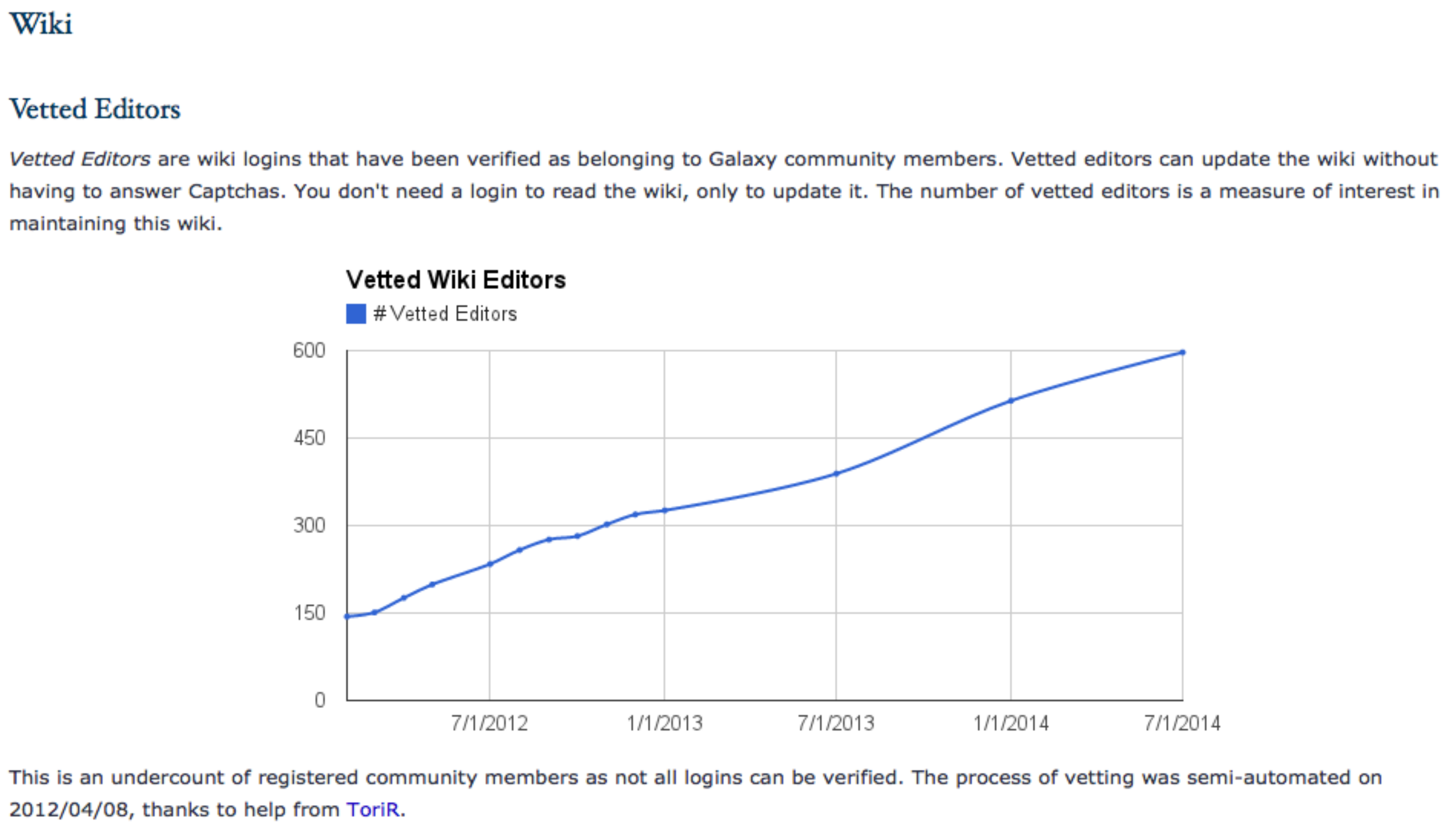
All Time Totals

Date	Posts	Threads	Questions	Answers	Comments	Users
2014/07/09	8223	2978	2959	2908	2337	2517

Biostar is a success, but…

- Users haven’t quite mastered the Q&A format
  - Reuse a post for multiple questions
  - Not accepting Answers
  - Comments are being Posted as Answers

wiki.galaxyproject.org



Wiki

Vetted Editors

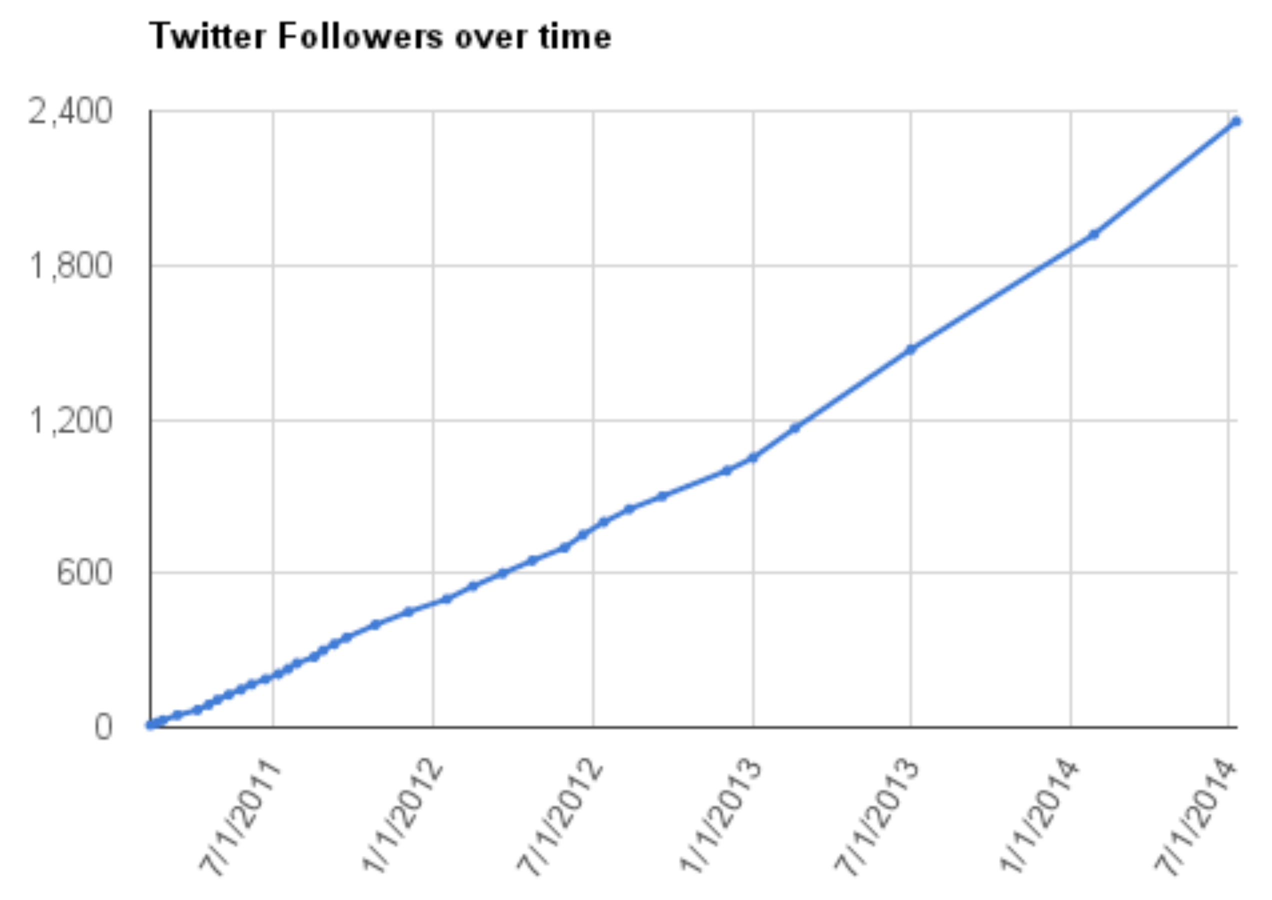
Vetted Editors are wiki logins that have been verified as belonging to Galaxy community members. Vetted editors can update the wiki without having to answer Captchas. You don't need a login to read the wiki, only to update it. The number of vetted editors is a measure of interest in maintaining this wiki.

This is an undercount of registered community members as not all logins can be verified. The process of vetting was semi-automated on 2012/04/08, thanks to help from ToriR.

Additional Channels

https://wiki.galaxyproject.org/GetInvolved

@galaxyproject #usegalaxy



https://twitter.com/galaxyproject

IRC Channel

Server: `irc.freenode.net`

Channel #galaxyproject

Overview

What is Galaxy?

Usage / Contributions

Virtual Interactions

ToolShed

Events and Outreach

Galaxy
Tool Shed

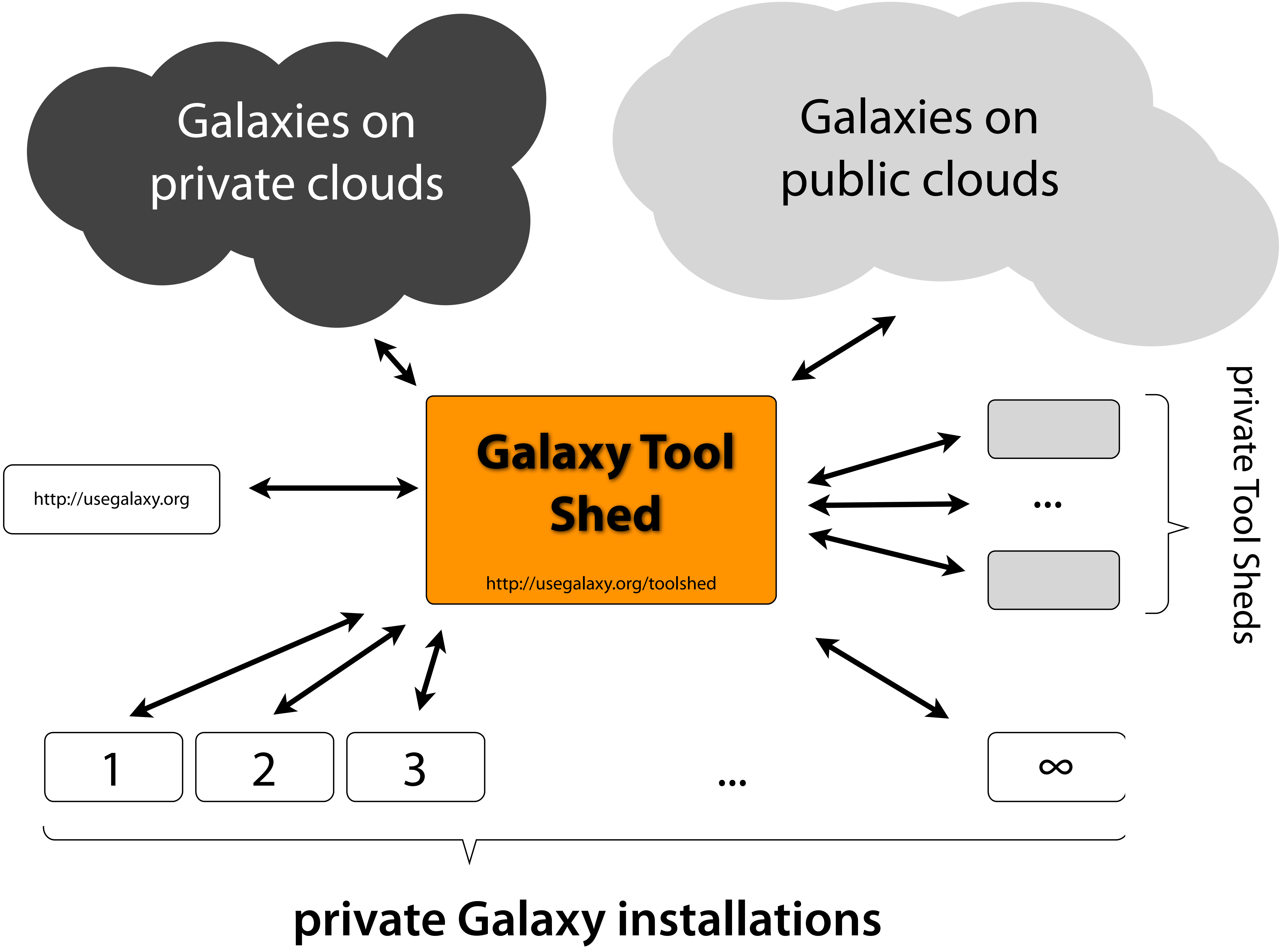
Enables sharing of Galaxy Utilities:

tools
proprietary datatypes
exported Galaxy workflows

Automatically install tools and tool suites, and their dependencies, into a Galaxy instance

Galaxy Utilities can be created and shared by any member of the community

https://wiki.galaxyproject.org/ToolShed



Galaxies on private clouds

Galaxies on public clouds

http://usegalaxy.org

Galaxy Tool Shed

http://usegalaxy.org/toolshed

...

private Tool Sheds

1

2

3

...

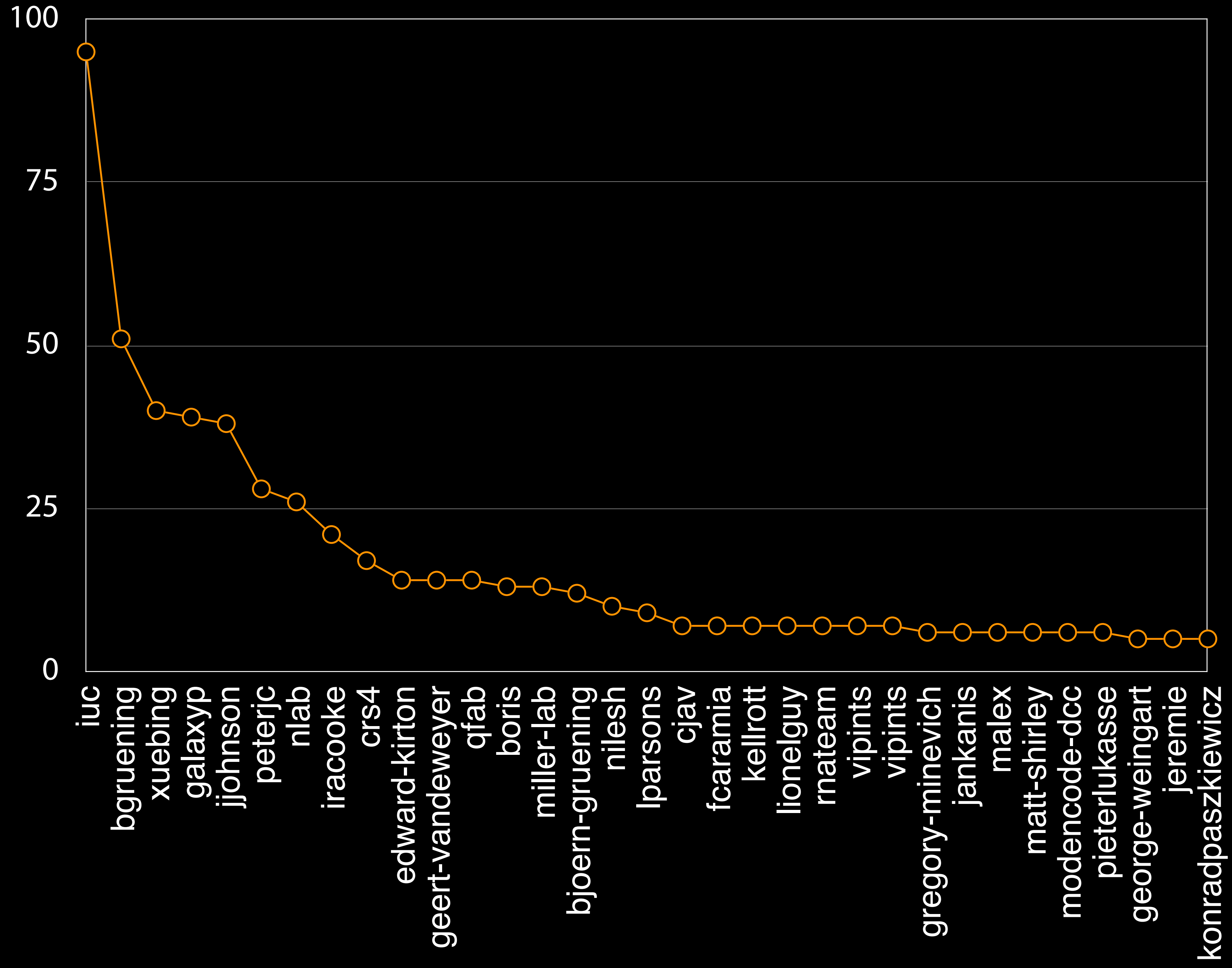
∞

private Galaxy installations

toolshed statistics

- 897 repositories
- 222 unique owners
- 176 Tool dependency package installation recipes
- 2,330 valid tools
- 3,420 valid versions of tools
- 54 exported Galaxy workflows
- 455 custom datatypes
- 62,021 total repository installations

toolshed contributions



Overview

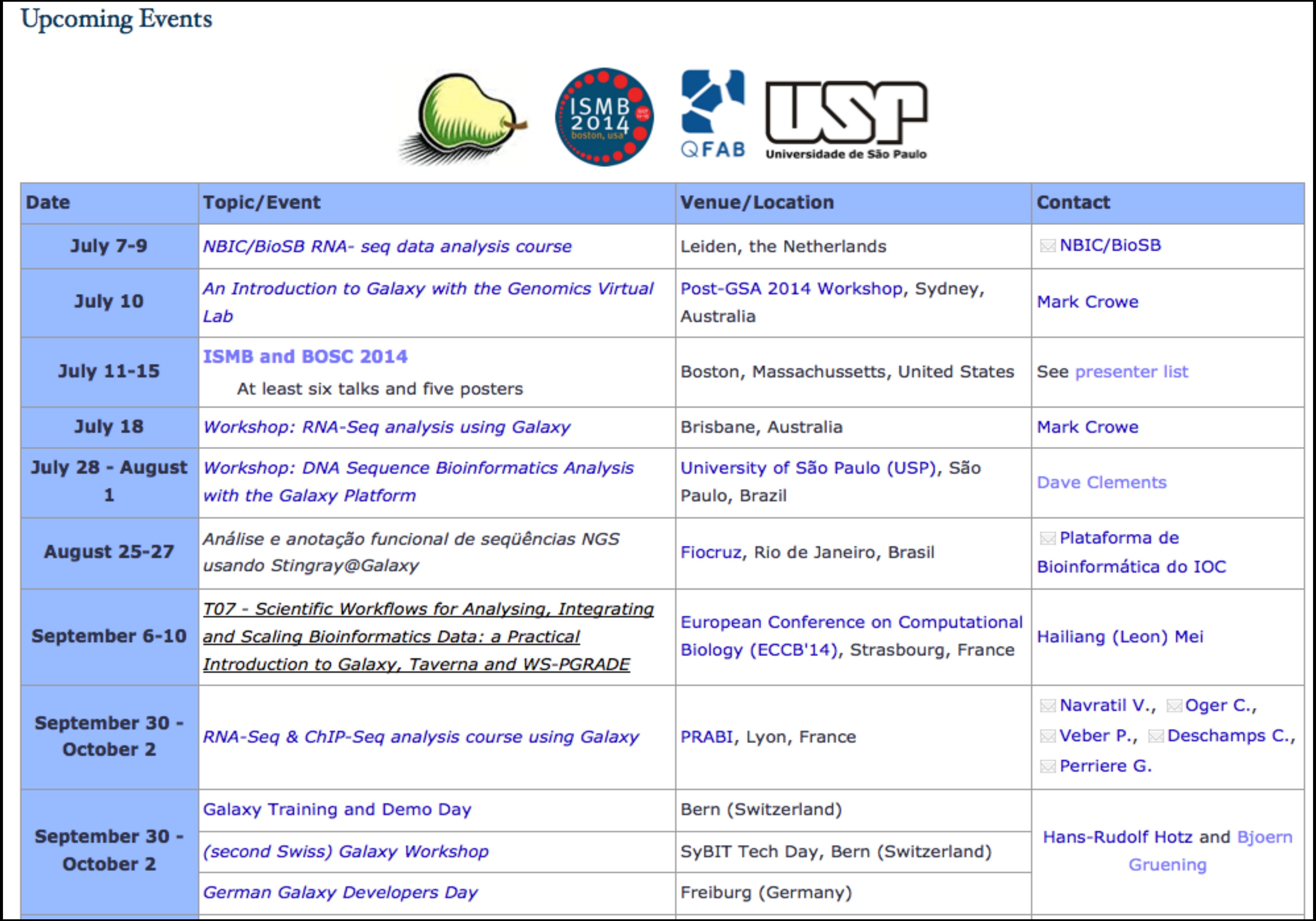
What is Galaxy?

Usage / Contributions

Virtual Interactions

ToolShed

Events and Outreach



Upcoming Events

Date	Topic/Event	Venue/Location	Contact
July 7-9	*NBIC/BioSB RNA- seq data analysis course*	Leiden, the Netherlands	✉ NBIC/BioSB
July 10	*An Introduction to Galaxy with the Genomics Virtual Lab*	Post-GSA 2014 Workshop, Sydney, Australia	Mark Crowe
July 11-15	**ISMB and BOSC 2014** At least six talks and five posters	Boston, Massachussetts, United States	See presenter list
July 18	*Workshop: RNA-Seq analysis using Galaxy*	Brisbane, Australia	Mark Crowe
July 28 - August 1	*Workshop: DNA Sequence Bioinformatics Analysis with the Galaxy Platform*	University of São Paulo (USP), São Paulo, Brazil	Dave Clements
August 25-27	*Análise e anotação funcional de seqüências NGS usando Stingray@Galaxy*	Fiocruz, Rio de Janeiro, Brasil	✉ Plataforma de Bioinformática do IOC
September 6-10	*T07 - Scientific Workflows for Analysing, Integrating and Scaling Bioinformatics Data: a Practical Introduction to Galaxy, Taverna and WS-PGRADE*	European Conference on Computational Biology (ECCB'14), Strasbourg, France	Hailiang (Leon) Mei
September 30 - October 2	*RNA-Seq & ChIP-Seq analysis course using Galaxy*	PRABI, Lyon, France	✉ Navratil V., ✉ Oger C., ✉ Veber P., ✉ Deschamps C., ✉ Perriere G.
September 30 - October 2	Galaxy Training and Demo Day	Bern (Switzerland)	Hans-Rudolf Hotz and Bjoern Gruening
	(second Swiss) Galaxy Workshop	SyBIT Tech Day, Bern (Switzerland)	
	German Galaxy Developers Day	Freiburg (Germany)	

https://wiki.galaxyproject.org/Events/

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

Fifth Annual Conference

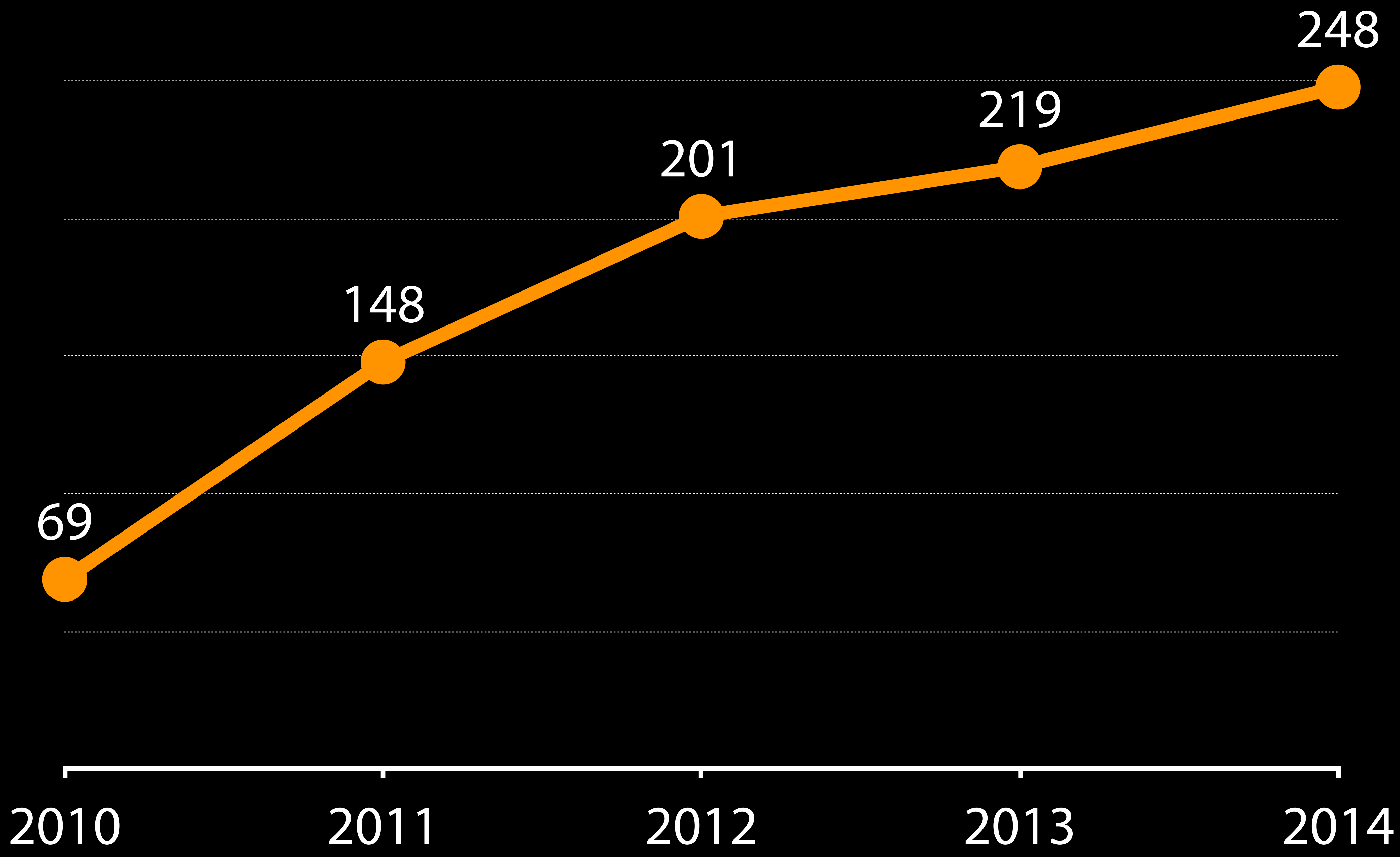
Two day Hackathon

One full day of Training Workshops

Two full days of presentations, posters, lightning talks, Birds of a Feather

https://wiki.galaxyproject.org/Events/GCC2014

GCC participation

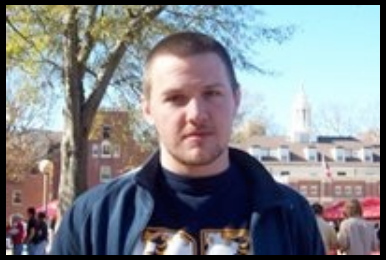


GCC2015

2015 Galaxy Community Conference will be held in Norwich, United Kingdom, at The Sainsbury Lab, 6-8th July

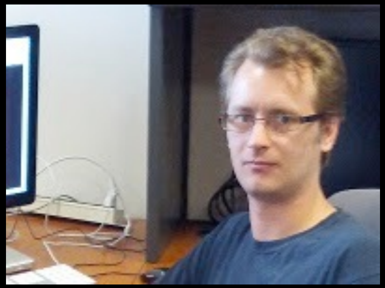
The Galaxy Team

Enis Afgan



Dannon Baker

Dan Blankenberg



Dave Bouvier

Marten Čech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Jeremy Goecks

Sam Guerler

Jen Jackson

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

Greg Von Kuster

http://wiki.galaxyproject.org/GalaxyTeam