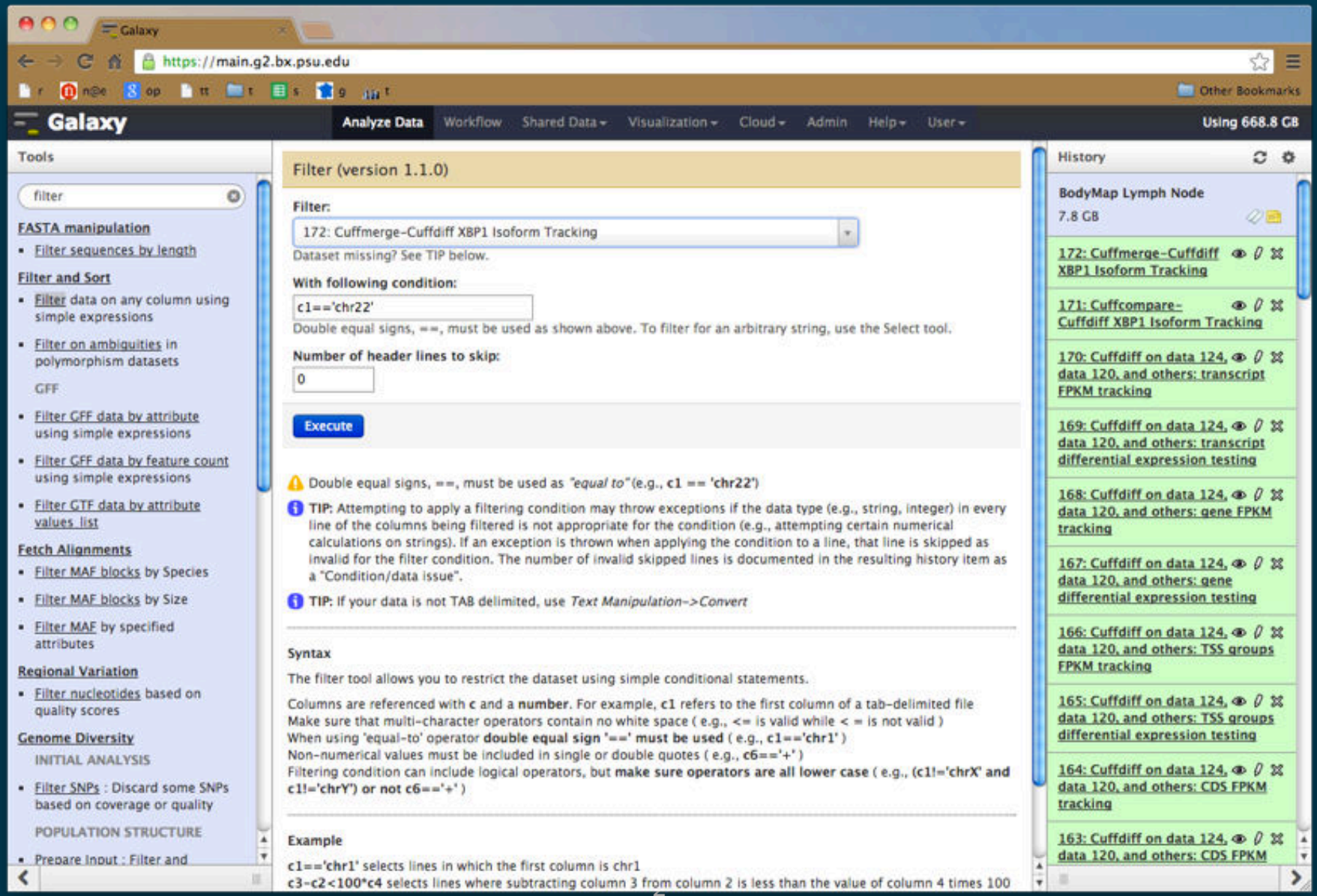
Supporting Multiple Community Networks with Galaxy

Jeremy Goecks

Department of Biology
Department of Math and Computer Science
Emory University

EMORY
UNIVERSITY

~~Short Demo~~


Galaxy
https://main.g2.bx.psu.edu
Other Bookmarks
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Admin
Help
User
Using 668.8 GB
Tools
filter
FASTA manipulation
Filter sequences by length
Filter and Sort
Filter data on any column using simple expressions
Filter on ambiguities in polymorphism datasets
GFF
Filter GFF data by attribute using simple expressions
Filter GFF data by feature count using simple expressions
Filter GTF data by attribute values list
Fetch Alignments
Filter MAF blocks by Species
Filter MAF blocks by Size
Filter MAF by specified attributes
Regional Variation
Filter nucleotides based on quality scores
Genome Diversity
INITIAL ANALYSIS
Filter SNPs : Discard some SNPs based on coverage or quality
POPULATION STRUCTURE
Prepare Input : Filter and
Filter (version 1.1.0)
Filter:
172: Cuffmerge-Cuffdiff XBP1 Isoform Tracking
Dataset missing? See TIP below.
With following condition:
c1=='chr22'
Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.
Number of header lines to skip:
0
Execute
Double equal signs, ==, must be used as "equal to" (e.g., c1 == 'chr22')
TIP: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".
TIP: If your data is not TAB delimited, use Text Manipulation->Convert
Syntax
The filter tool allows you to restrict the dataset using simple conditional statements.
Columns are referenced with c and a number. For example, c1 refers to the first column of a tab-delimited file
Make sure that multi-character operators contain no white space (e.g., <= is valid while < = is not valid)
When using 'equal-to' operator double equal sign '==' must be used (e.g., c1=='chr1')
Non-numerical values must be included in single or double quotes (e.g., c6=='+')
Filtering condition can include logical operators, but make sure operators are all lower case (e.g., (c1!='chrX' and c1!='chrY') or not c6=='+')
Example
c1=='chr1' selects lines in which the first column is chr1
c3-c2<100*c4 selects lines where subtracting column 3 from column 2 is less than the value of column 4 times 100
History
BodyMap Lymph Node
7.8 GB
172: Cuffmerge-Cuffdiff XBP1 Isoform Tracking
171: Cuffcompare-Cuffdiff XBP1 Isoform Tracking
170: Cuffdiff on data 124, data 120, and others: transcript FPKM tracking
169: Cuffdiff on data 124, data 120, and others: transcript differential expression testing
168: Cuffdiff on data 124, data 120, and others: gene FPKM tracking
167: Cuffdiff on data 124, data 120, and others: gene differential expression testing
166: Cuffdiff on data 124, data 120, and others: TSS groups FPKM tracking
165: Cuffdiff on data 124, data 120, and others: TSS groups differential expression testing
164: Cuffdiff on data 124, data 120, and others: CDS FPKM tracking
163: Cuffdiff on data 124, data 120, and others: CDS FPKM

Vision

Galaxy is an **open, Web-based platform** for accessible, reproducible, and collaborative computational genomics

What is Galaxy?

GUI for high-throughput, high-performance genomics

1. get and integrate public, private data
2. analyze data and create workflows
3. visualization and visual analysis, sharing, publication

Customizable open-source software on various HPC resources

- public website — http://usegalaxy.org
- local instance
- on the cloud

Galaxy is Very Popular

Local installations all over the world

Public Website (**http://usegalaxy.org**), anybody can use:

- ~500 new users per month, ~100 TB of user data, ~130,000 analysis jobs per month

Used and cited in more than 1000 publications

Galaxy Communities

End Users

Deployers/Maintainers

Developers

- plug-ins: tools, visualizations
- core framework

Goals

A community connected to the project and each other

A well-informed community

Grow the community

A community that contributes:

- analyses, curation, and data (=knowledge)
- tools and enhancements
- support
- outreach and training

How?

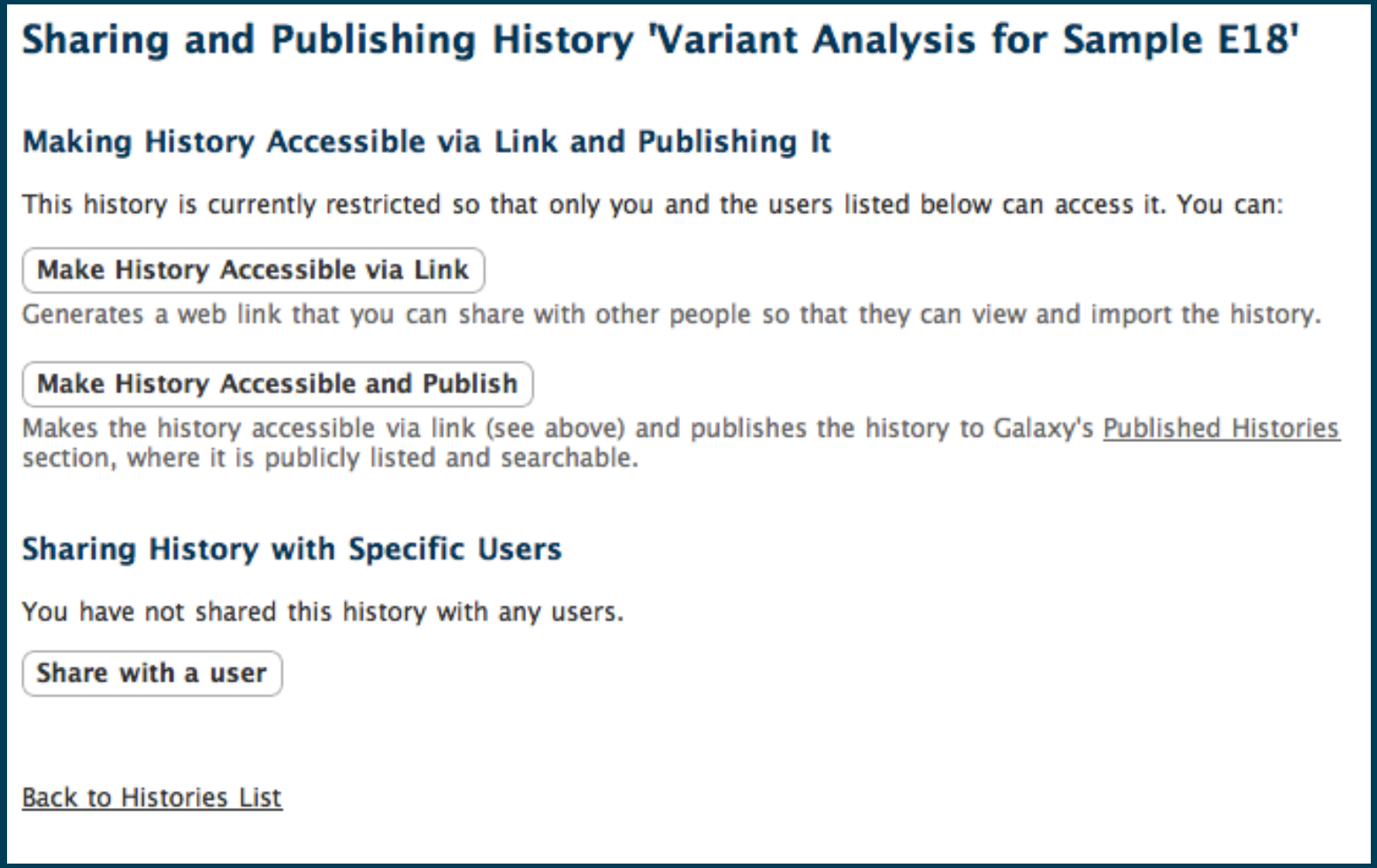
Enable your community

Communicate with your community

Reward your community

Sharing, Collaborating, and Publishing Analyses with Galaxy

Sharing and Publishing



Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history **accessible via link and published.**

Anyone can view and import this history by visiting the following URL:

http://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-for-sample-e18

This history is publicly listed and searchable in Galaxy's Published Histories section.

You can:

Unpublish History

Removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Disable Access to History via Link and Unpublish

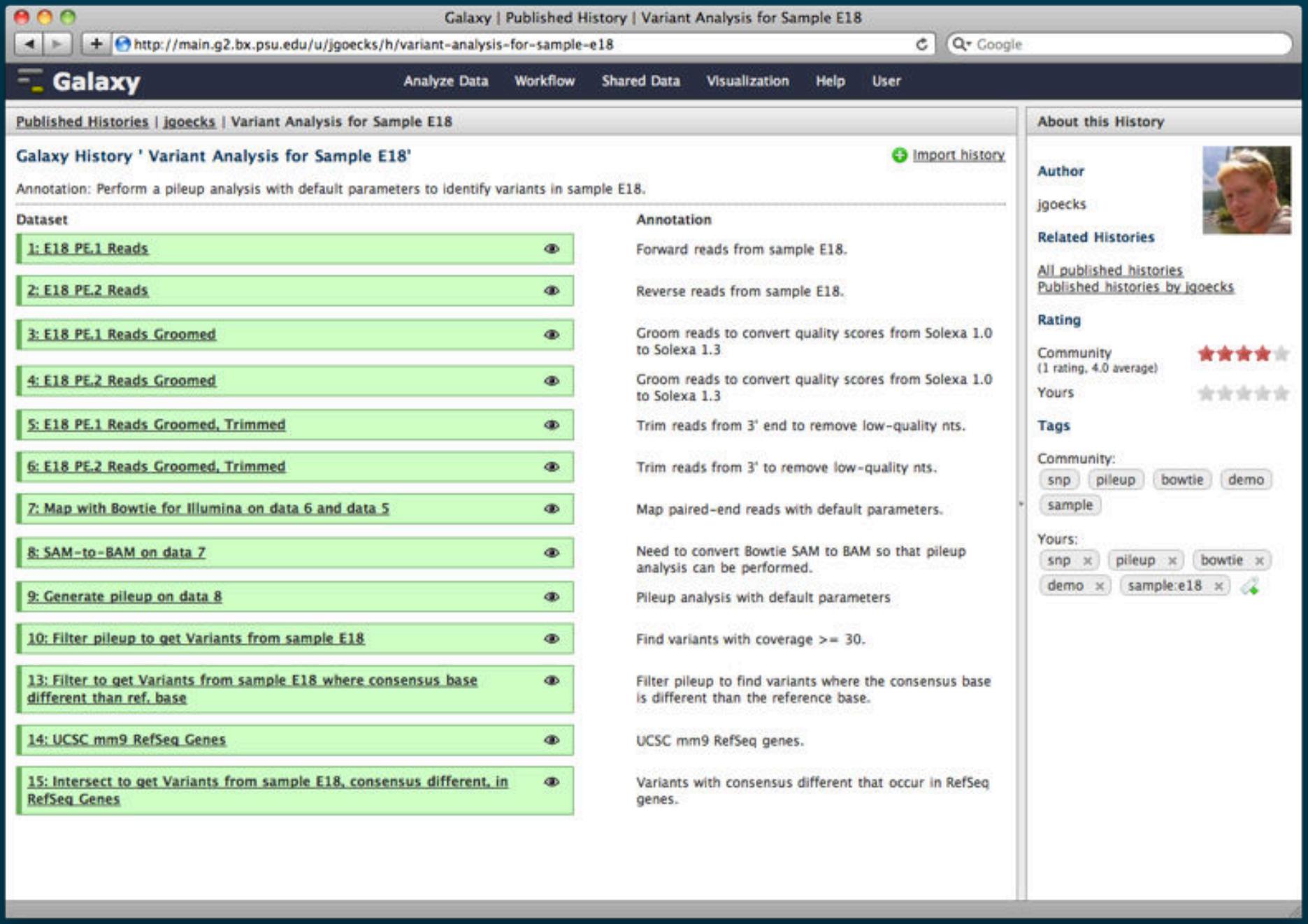
Disables history's link so that it is not accessible and removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

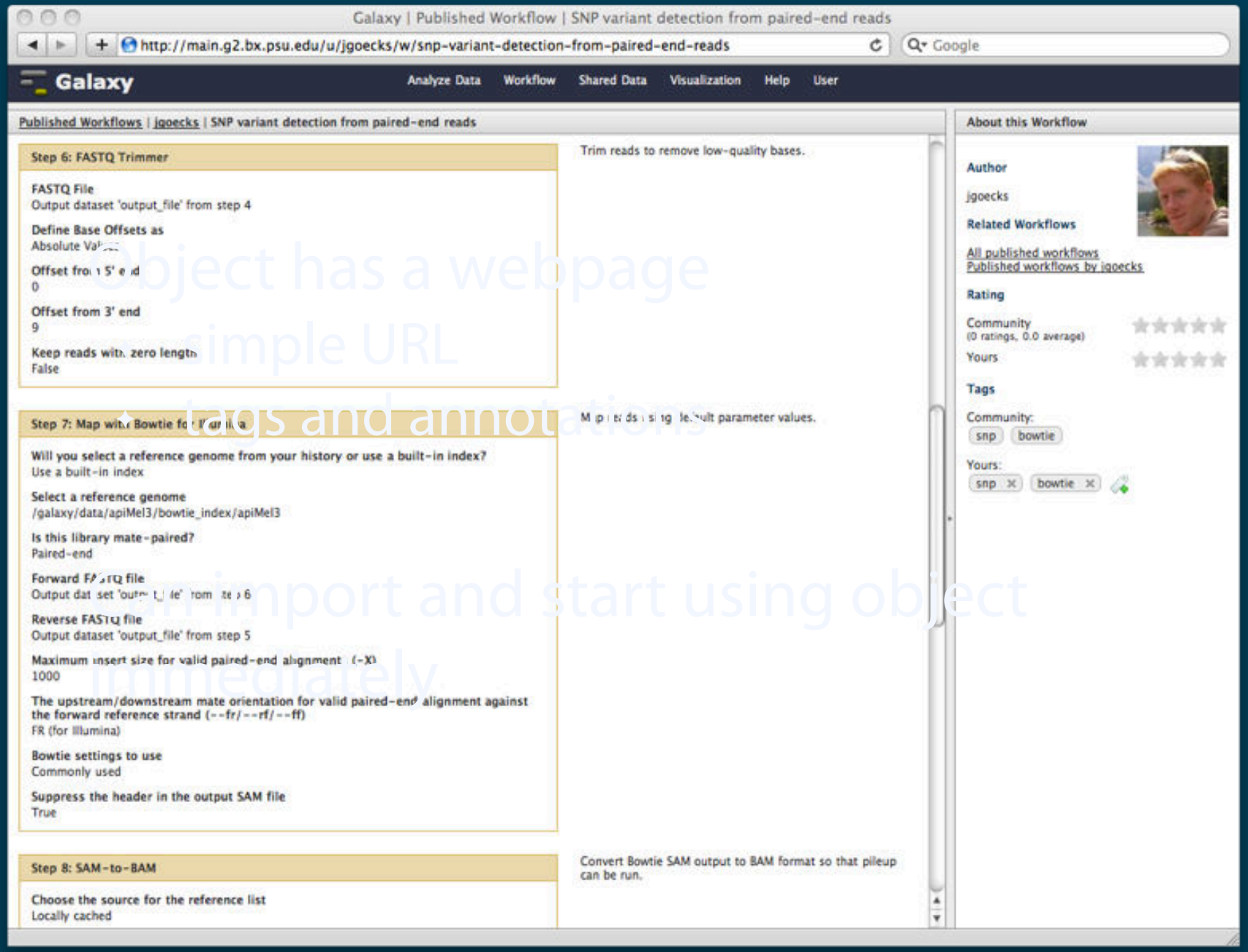
Sharing History with Specific Users

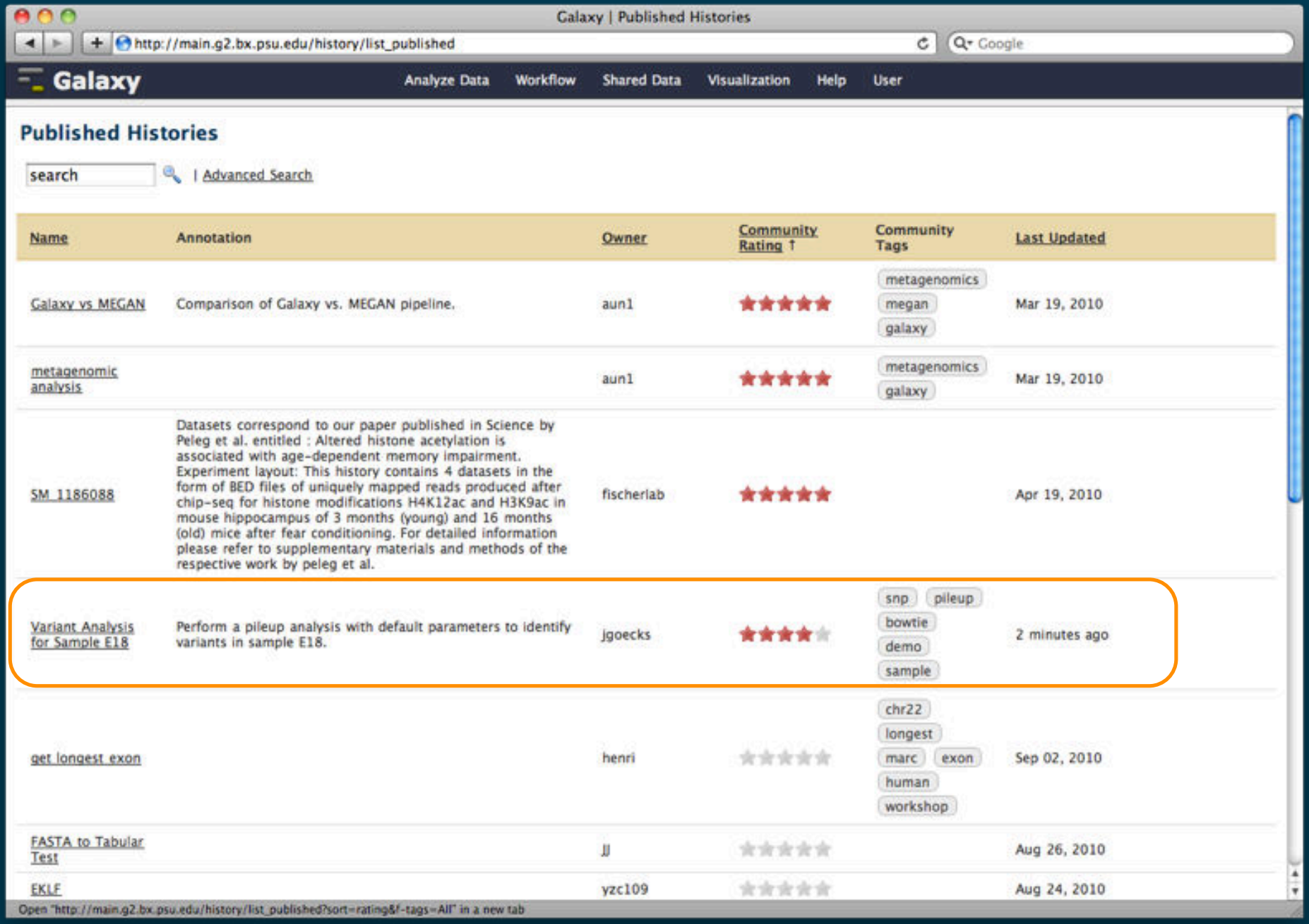
You have not shared this history with any users.

Share with a user

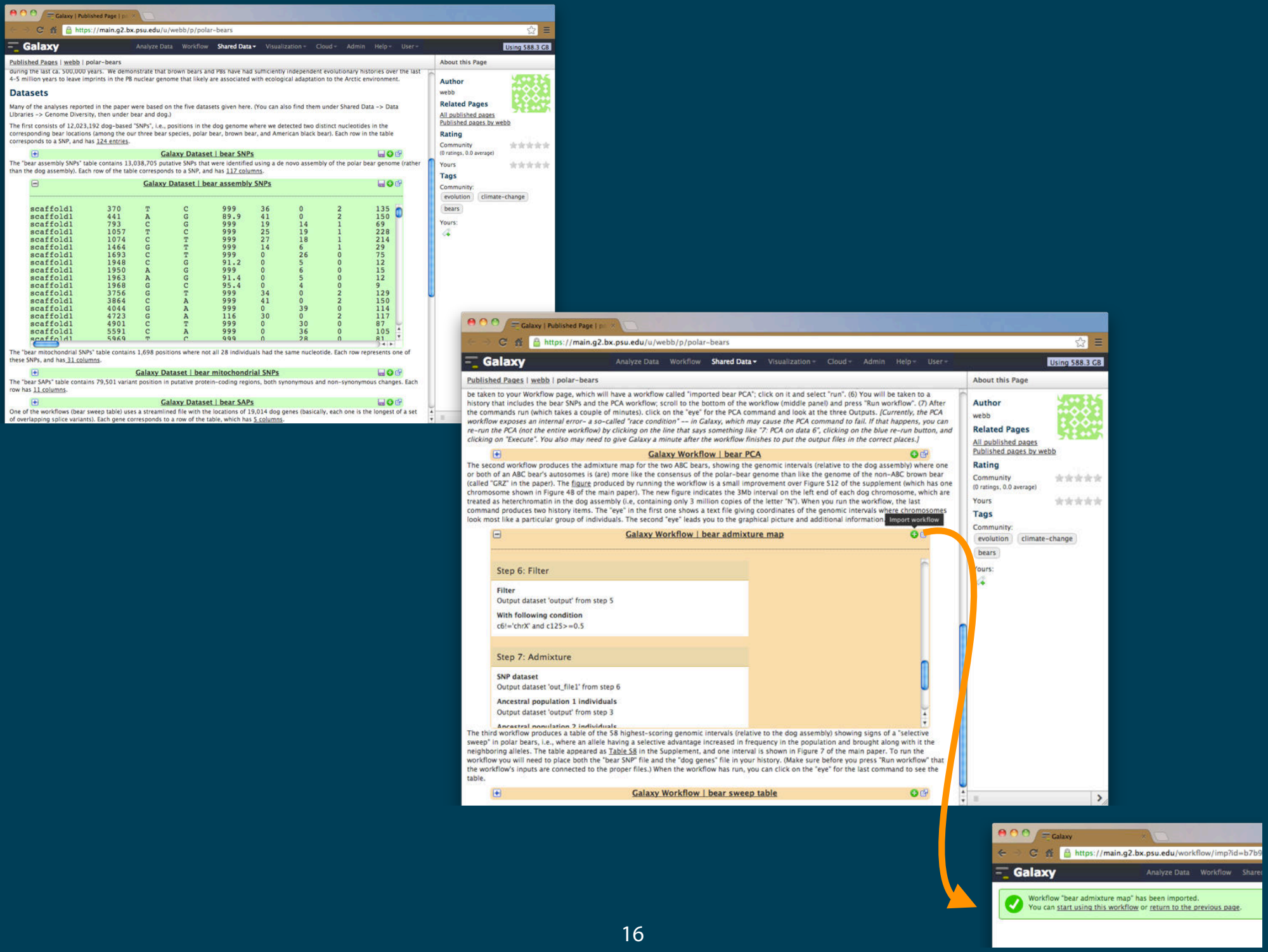
Back to Histories List


Galaxy | Published History | Variant Analysis for Sample E18
http://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-for-sample-e18
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Histories | jgoecks | Variant Analysis for Sample E18
Galaxy History ' Variant Analysis for Sample E18'
Import history
Annotation: Perform a pileup analysis with default parameters to identify variants in sample E18.
Dataset
Annotation
1: E18 PE.1 Reads
Forward reads from sample E18.
2: E18 PE.2 Reads
Reverse reads from sample E18.
3: E18 PE.1 Reads Groomed
Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3
4: E18 PE.2 Reads Groomed
Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3
5: E18 PE.1 Reads Groomed, Trimmed
Trim reads from 3' end to remove low-quality nts.
6: E18 PE.2 Reads Groomed, Trimmed
Trim reads from 3' to remove low-quality nts.
7: Map with Bowtie for Illumina on data 6 and data 5
Map paired-end reads with default parameters.
8: SAM-to-BAM on data 7
Need to convert Bowtie SAM to BAM so that pileup analysis can be performed.
9: Generate pileup on data 8
Pileup analysis with default parameters
10: Filter pileup to get Variants from sample E18
Find variants with coverage >= 30.
13: Filter to get Variants from sample E18 where consensus base different than ref. base
Filter pileup to find variants where the consensus base is different than the reference base.
14: UCSC mm9 RefSeq Genes
UCSC mm9 RefSeq genes.
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes
Variants with consensus different that occur in RefSeq genes.
About this History
Author
jgoecks
Related Histories
All published histories
Published histories by jgoecks
Rating
Community
(1 rating, 4.0 average)
Yours
Tags
Community:
snp
pileup
bowtie
demo
sample
Yours:
snp
pileup
bowtie
demo
sample:e18


Galaxy | Published Workflow | SNP variant detection from paired-end reads
http://main.g2.bx.psu.edu/u/jgoecks/w/snp-variant-detection-from-paired-end-reads
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Workflows | jgoecks | SNP variant detection from paired-end reads
Step 6: FASTQ Trimmer
Trim reads to remove low-quality bases.
FASTQ File
Output dataset 'output_file' from step 4
Define Base Offsets as
Absolute Va
Offset fro 5' e d
0
Offset from 3' end
9
Keep reads with zero length
False
Step 7: Map wi Bowtie fo
M p ds s g le ult parameter values.
Will you select a reference genome from your history or use a built-in index?
Use a built-in index
Select a reference genome
/galaxy/data/apiMel3/bowtie_index/apiMel3
Is this library mate-paired?
Paired-end
Forward FA TQ file
Output dat set 'outp t_ le' rom te 6
Reverse FASTQ file
Output dataset 'output_file' from step 5
Maximum nsert size for valid paired-end alignment (-X)
1000
The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff)
FR (for Illumina)
Bowtie settings to use
Commonly used
Suppress the header in the output SAM file
True
Step 8: SAM-to-BAM
Convert Bowtie SAM output to BAM format so that pileup can be run.
Choose the source for the reference list
Locally cached
About this Workflow
Author
jgoecks
Related Workflows
All published workflows
Published workflows by jgoecks
Rating
Community
(0 ratings, 0.0 average)
Yours
Tags
Community:
snp
bowtie
Yours:
snp
bowtie
Object has a webpage
Simple URL
Tags and annotations
Import and start using object immediately


Galaxy | Published Histories
http://main.g2.bx.psu.edu/history/list_published
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Histories
search
| Advanced Search
Name
Annotation
Owner
Community Rating ↑
Community Tags
Last Updated
Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.
aun1
metagenomics
megan
galaxy
Mar 19, 2010
metagenomic analysis
aun1
metagenomics
galaxy
Mar 19, 2010
SM_1186088
Datasets correspond to our paper published in Science by Peleg et al. entitled : Altered histone acetylation is associated with age-dependent memory impairment. Experiment layout: This history contains 4 datasets in the form of BED files of uniquely mapped reads produced after chip-seq for histone modifications H4K12ac and H3K9ac in mouse hippocampus of 3 months (young) and 16 months (old) mice after fear conditioning. For detailed information please refer to supplementary materials and methods of the respective work by peleg et al.
fischerlab
Apr 19, 2010
Variant Analysis for Sample E18
Perform a pileup analysis with default parameters to identify variants in sample E18.
jgoecks
snp
pileup
bowtie
demo
sample
2 minutes ago
get longest exon
henri
chr22
longest
marc
exon
human
workshop
Sep 02, 2010
FASTA to Tabular Test
JJ
Aug 26, 2010
EKLF
yzc109
Aug 24, 2010
Open "http://main.g2.bx.psu.edu/history/list_published?sort=rating&f-tags=All" in a new tab

Interactive Research Documents

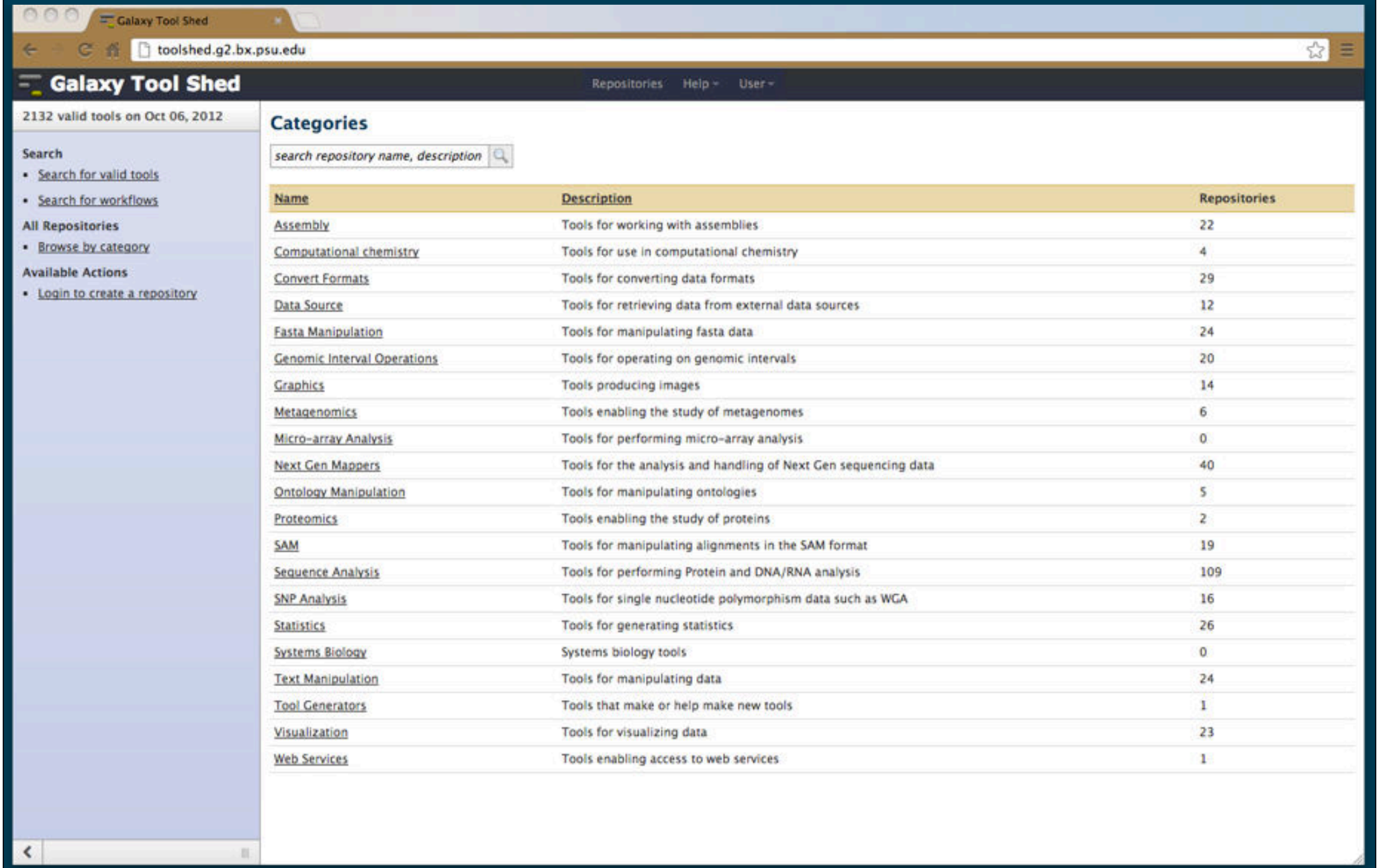

Galaxy | Published Page
https://main.g2.bx.psu.edu/u/webb/p/polar-bears
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Admin
Help
User
Using 588.3 GB
Published Pages | webb | polar-bears
Datasets
Many of the analyses reported in the paper were based on the five datasets given here. (You can also find them under Shared Data -> Data Libraries -> Genome Diversity, then under bear and dog.)
The first consists of 12,023,192 dog-based "SNPs", i.e., positions in the dog genome where we detected two distinct nucleotides in the corresponding bear locations (among the our three bear species, polar bear, brown bear, and American black bear). Each row in the table corresponds to a SNP, and has 124 entries.
Galaxy Dataset | bear SNPs
The "bear assembly SNPs" table contains 13,038,705 putative SNPs that were identified using a de novo assembly of the polar bear genome (rather than the dog assembly). Each row of the table corresponds to a SNP, and has 117 columns.
Galaxy Dataset | bear assembly SNPs
scaffold1 370 T C 999 36 0 2 135
scaffold1 441 A G 89.9 41 0 2 150
scaffold1 793 C G 999 19 14 1 69
scaffold1 1057 T C 999 25 19 1 228
scaffold1 1074 C T 999 27 18 1 214
scaffold1 1464 G T 999 14 6 1 29
scaffold1 1693 C T 999 0 26 0 75
scaffold1 1948 C G 91.2 0 5 0 12
scaffold1 1950 A G 999 0 6 0 15
scaffold1 1963 A G 91.4 0 5 0 12
scaffold1 1968 G C 95.4 0 4 0 9
scaffold1 3756 G T 999 34 0 2 129
scaffold1 3864 C A 999 41 0 2 150
scaffold1 4044 G A 999 0 39 0 114
scaffold1 4723 G A 116 30 0 2 117
scaffold1 4901 C T 999 0 30 0 87
scaffold1 5591 C A 999 0 36 0 105
The "bear mitochondrial SNPs" table contains 1,698 positions where not all 28 individuals had the same nucleotide. Each row represents one of these SNPs, and has 31 columns.
Galaxy Dataset | bear mitochondrial SNPs
The "bear SAPs" table contains 79,501 variant position in putative protein-coding regions, both synonymous and non-synonymous changes. Each row has 11 columns.
Galaxy Dataset | bear SAPs
One of the workflows (bear sweep table) uses a streamlined file with the locations of 19,014 dog genes (basically, each one is the longest of a set of overlapping splice variants). Each gene corresponds to a row of the table, which has 5 columns.
About this Page
Author
webb
Related Pages
All published pages
Published pages by webb
Rating
Community
(0 ratings, 0.0 average)
Yours
Tags
Community:
evolution
climate-change
bears
Yours:
be taken to your Workflow page, which will have a workflow called "imported bear PCA"; click on it and select "run". (6) You will be taken to a history that includes the bear SNPs and the PCA workflow; scroll to the bottom of the workflow (middle panel) and press "Run workflow". (7) After the commands run (which takes a couple of minutes). click on the "eye" for the PCA command and look at the three Outputs. [Currently, the PCA workflow exposes an internal error- a so-called "race condition" -- in Galaxy, which may cause the PCA command to fail. If that happens, you can re-run the PCA (not the entire workflow) by clicking on the line that says something like "7: PCA on data 6", clicking on the blue re-run button, and clicking on "Execute". You also may need to give Galaxy a minute after the workflow finishes to put the output files in the correct places.]
Galaxy Workflow | bear PCA
The second workflow produces the admixture map for the two ABC bears, showing the genomic intervals (relative to the dog assembly) where one or both of an ABC bear's autosomes is (are) more like the consensus of the polar-bear genome than like the genome of the non-ABC brown bear (called "GRZ" in the paper). The figure produced by running the workflow is a small improvement over Figure S12 of the supplement (which has one chromosome shown in Figure 4B of the main paper). The new figure indicates the 3Mb interval on the left end of each dog chromosome, which are treated as heterchromatin in the dog assembly (i.e, containing only 3 million copies of the letter "N"). When you run the workflow, the last command produces two history items. The "eye" in the first one shows a text file giving coordinates of the genomic intervals where chromosomes look most like a particular group of individuals. The second "eye" leads you to the graphical picture and additional information.
Import workflow
Galaxy Workflow | bear admixture map
Step 6: Filter
Filter
Output dataset 'output' from step 5
With following condition
c6!='chrX' and c125>=0.5
Step 7: Admixture
SNP dataset
Output dataset 'out_file1' from step 6
Ancestral population 1 individuals
Output dataset 'output' from step 3
The third workflow produces a table of the 58 highest-scoring genomic intervals (relative to the dog assembly) showing signs of a "selective sweep" in polar bears, i.e., where an allele having a selective advantage increased in frequency in the population and brought along with it the neighboring alleles. The table appeared as Table S8 in the Supplement, and one interval is shown in Figure 7 of the main paper. To run the workflow you will need to place both the "bear SNP" file and the "dog genes" file in your history. (Make sure before you press "Run workflow" that the workflow's inputs are connected to the proper files.) When the workflow has run, you can click on the "eye" for the last command to see the table.
Galaxy Workflow | bear sweep table
Workflow "bear admixture map" has been imported.
You can start using this workflow or return to the previous page.

Uses

Biomedical collaboration

Publication support

Teaching and demonstration


Galaxy Tool Shed
toolshed.g2.bx.psu.edu
Galaxy Tool Shed
Repositories
Help
User
2132 valid tools on Oct 06, 2012
Search
Search for valid tools
Search for workflows
All Repositories
Browse by category
Available Actions
Login to create a repository
Categories
search repository name, description
Name
Description
Repositories
Assembly
Tools for working with assemblies
22
Computational chemistry
Tools for use in computational chemistry
4
Convert Formats
Tools for converting data formats
29
Data Source
Tools for retrieving data from external data sources
12
Fasta Manipulation
Tools for manipulating fasta data
24
Genomic Interval Operations
Tools for operating on genomic intervals
20
Graphics
Tools producing images
14
Metagenomics
Tools enabling the study of metagenomes
6
Micro-array Analysis
Tools for performing micro-array analysis
0
Next Gen Mappers
Tools for the analysis and handling of Next Gen sequencing data
40
Ontology Manipulation
Tools for manipulating ontologies
5
Proteomics
Tools enabling the study of proteins
2
SAM
Tools for manipulating alignments in the SAM format
19
Sequence Analysis
Tools for performing Protein and DNA/RNA analysis
109
SNP Analysis
Tools for single nucleotide polymorphism data such as WGA
16
Statistics
Tools for generating statistics
26
Systems Biology
Systems biology tools
0
Text Manipulation
Tools for manipulating data
24
Tool Generators
Tools that make or help make new tools
1
Visualization
Tools for visualizing data
23
Web Services
Tools enabling access to web services
1

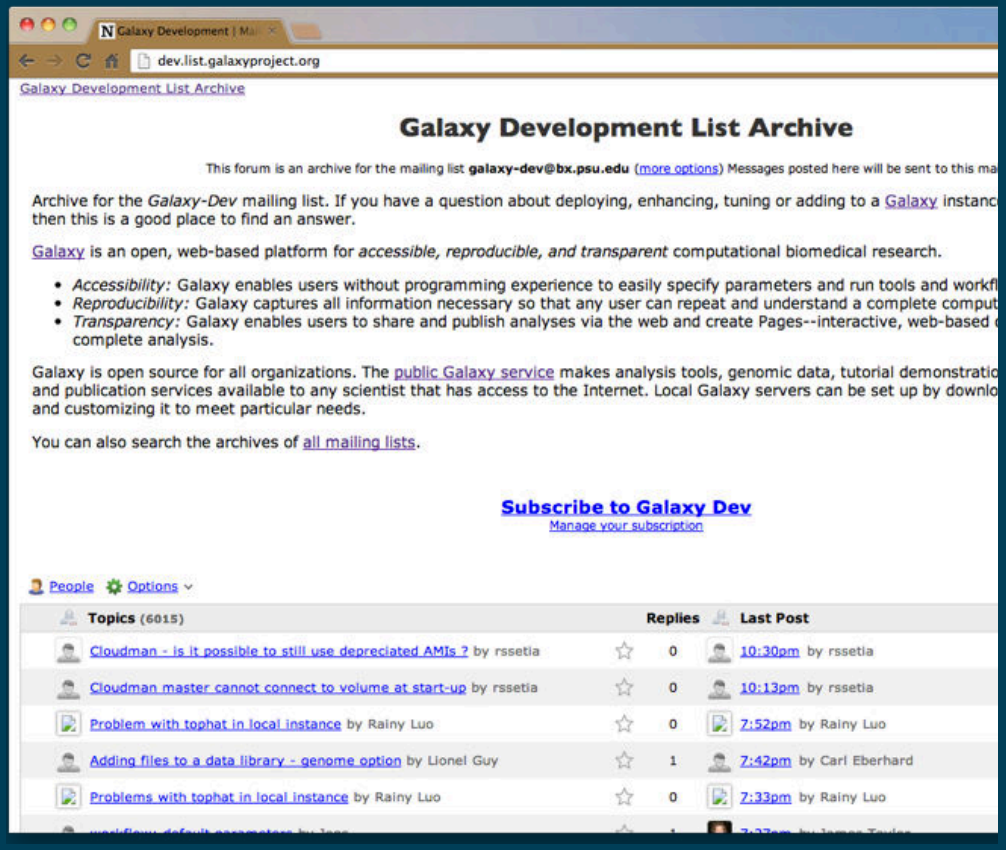
How?

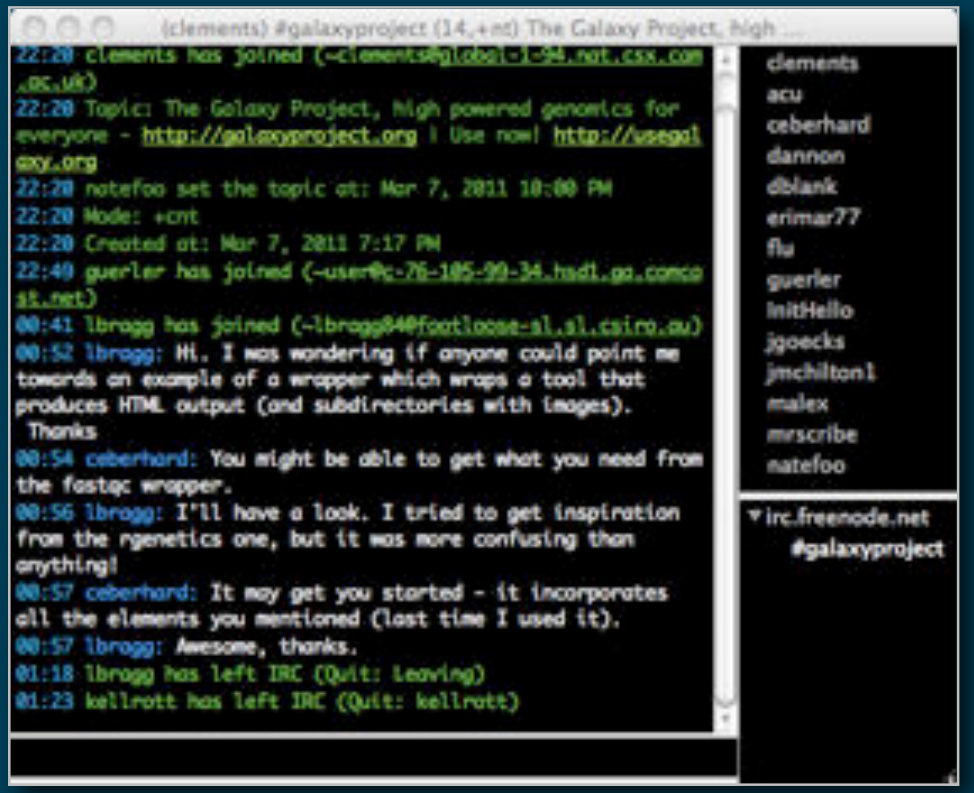
Enable your community

Communicate with your community

Reward your community

Provide Open Support Channels





Open channels support anyone posting questions

and answers

Mailing Lists vs Forums?

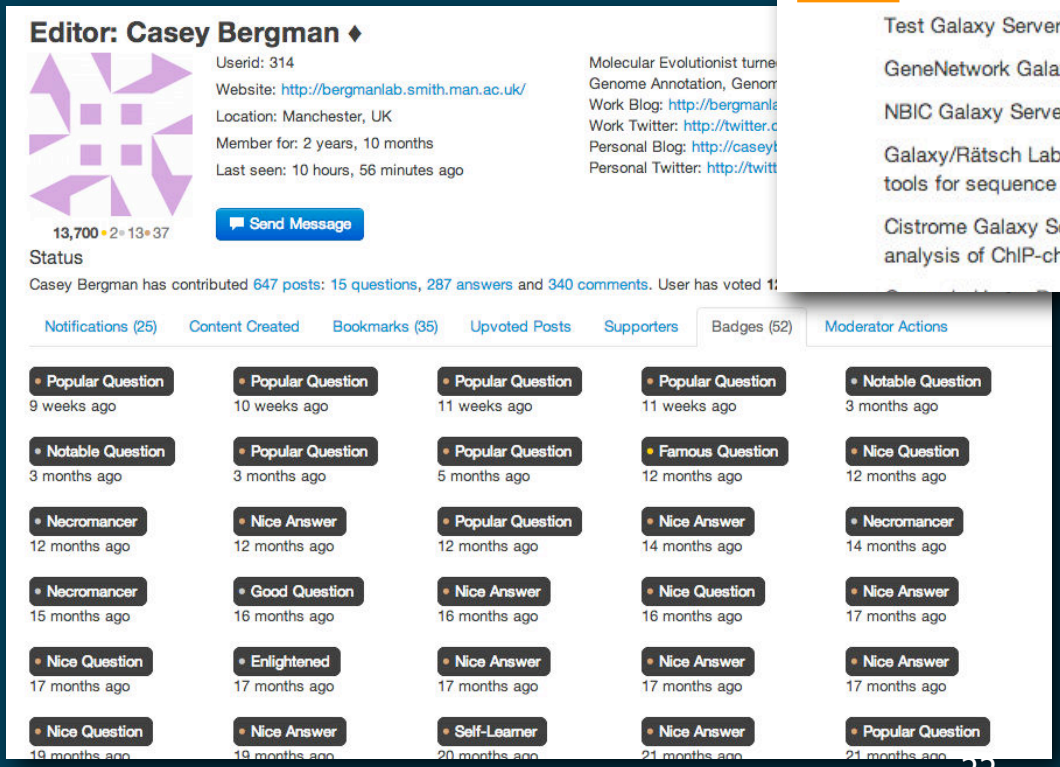
We are replacing Galaxy-User mailing list with a forum.

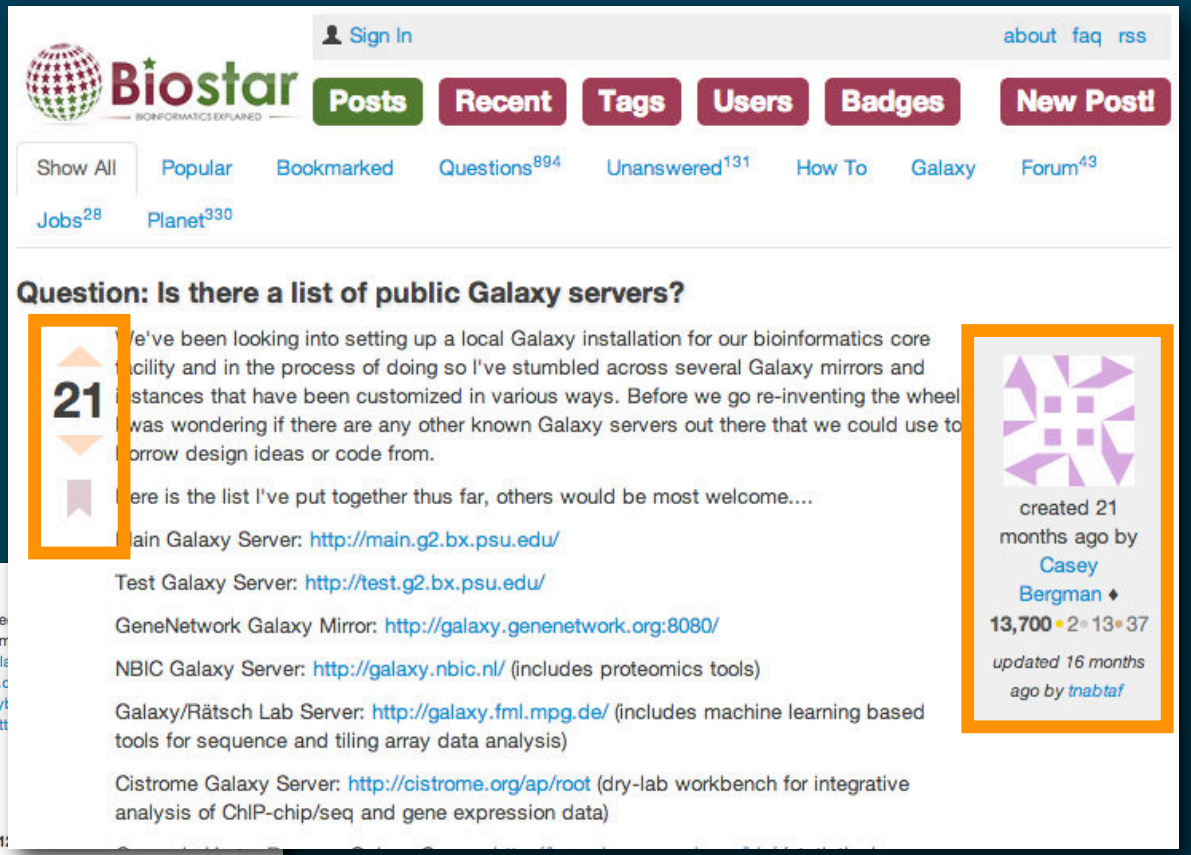
We will be using Biostar (http://biostars.org) directly

Incentivize contributions

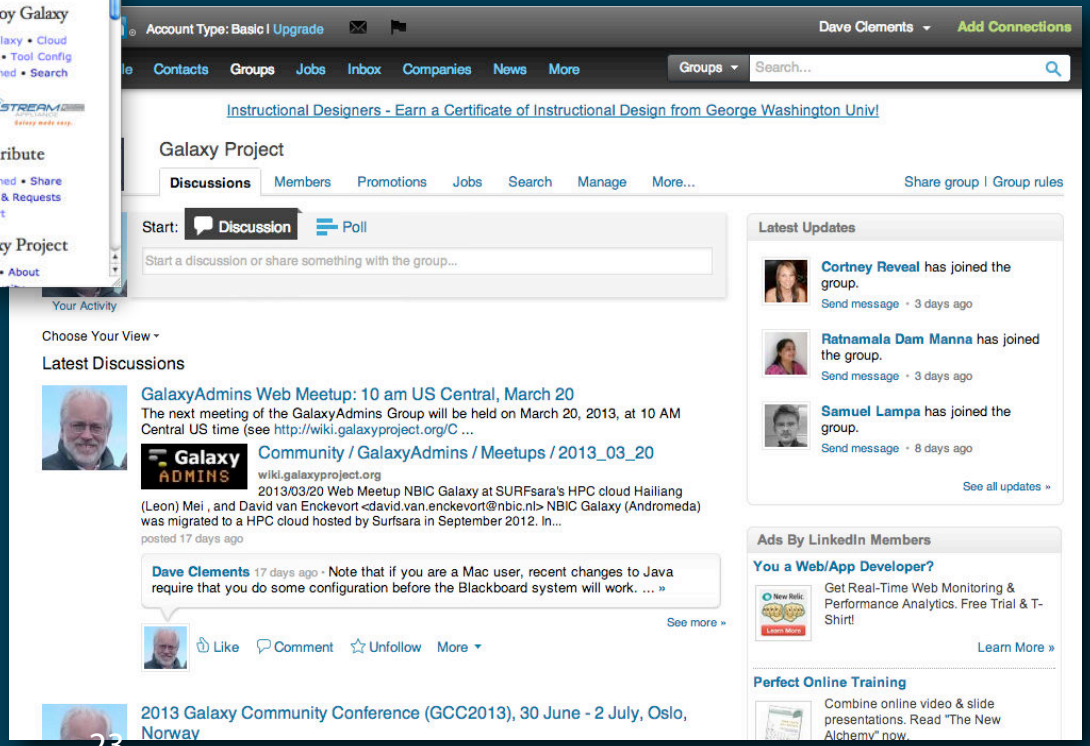
BioStar is a gamified online forum.

Points, badges, voting, ...



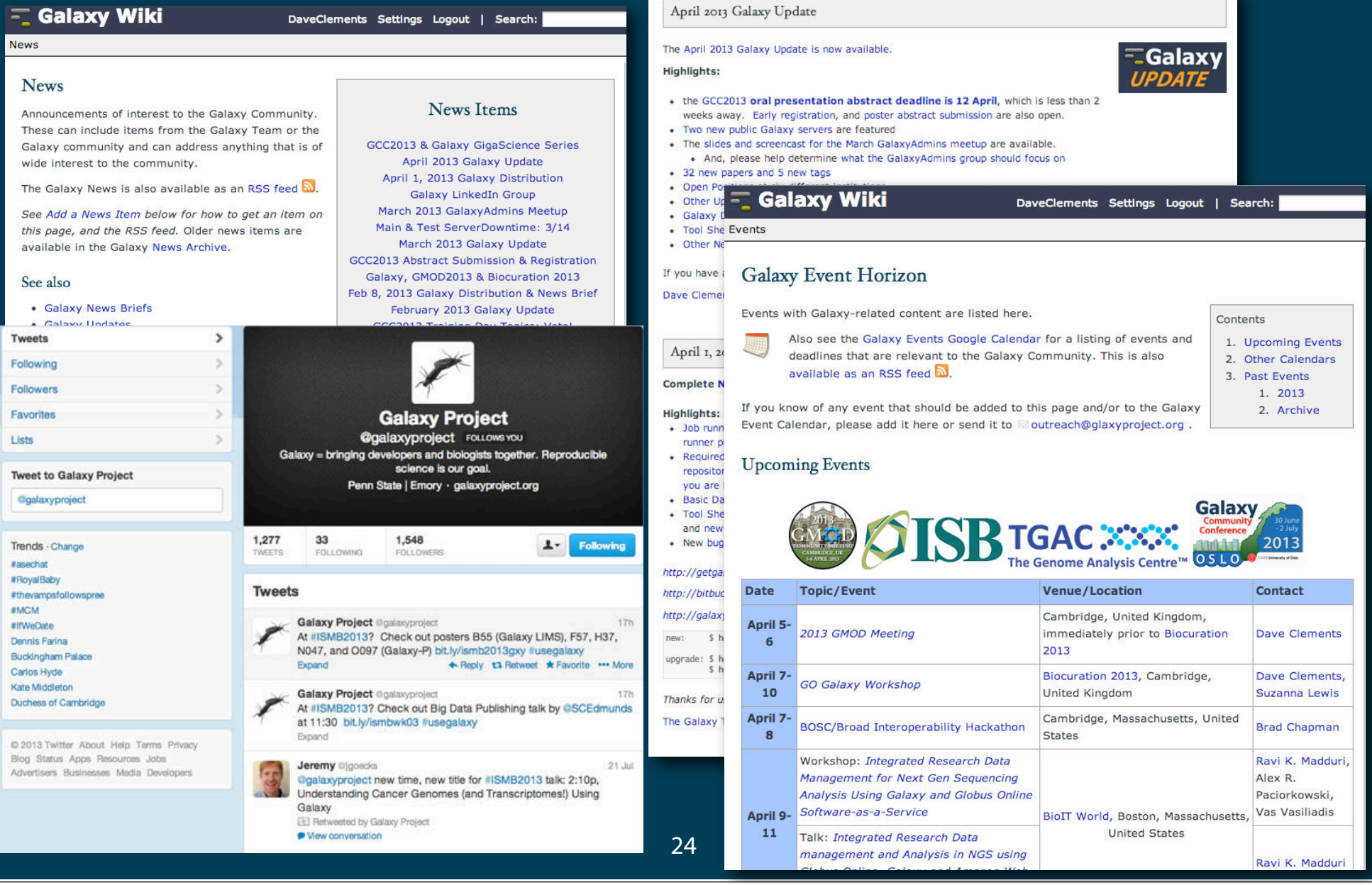


Community Resources and Hubs



Wikis and social sites:
communicate, enable, reward.

Communicate!



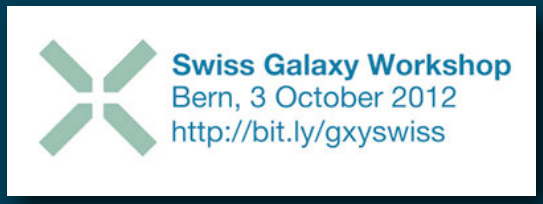
How?

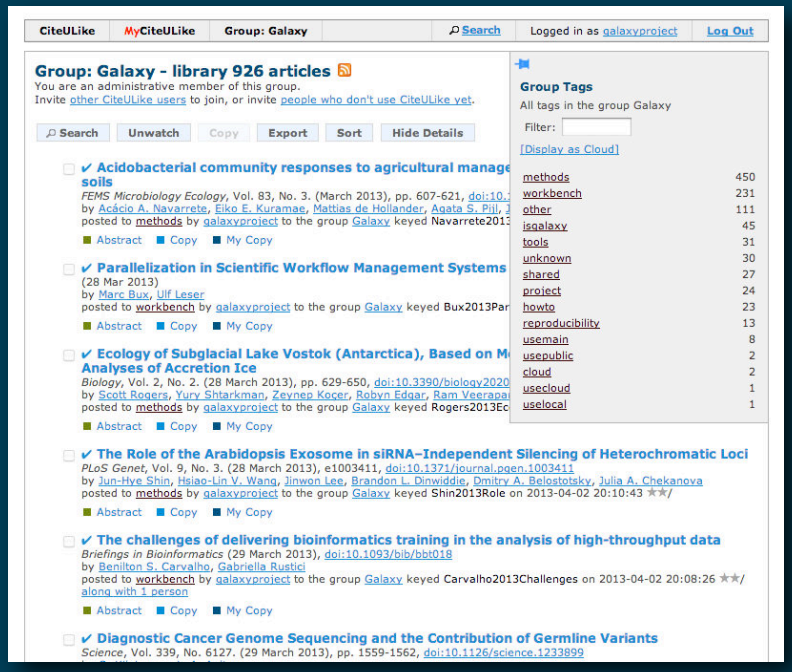
Enable your community

Communicate with your community

Reward your community

Reward Community Efforts





Publicly Accessible Galaxy Servers

Galaxy's public server (UseGalaxy.org, *Main*) can meet many needs, but it is not suitable for everything (see Big Picture/Choices for why) and cannot possibly scale to meet the entire world's needs.

Fortunately the Galaxy Community is helping out by installing Galaxy at their institutions and then making those installations either publicly available or open to their organizations or community.

This page lists such public or semi-public Galaxy servers.

To add your public Galaxy server to this list, please either just add it (hey, *it's a wiki*), or contact Galaxy Outreach

Contents

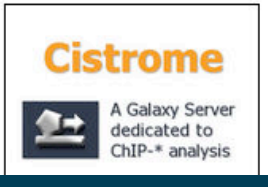
Andromeda

- *Link:*
  Andromeda
- *Domain/Purpose:*
  This is a fully populated Galaxy instance.
- *Comments:*
  Andromeda is hosted at the SURFsara High Performance Computing (HPC) cloud. The installation is supported by Enis Afgan (CloudMan project) and Mattias de Hollander (NIOO).
- *Quotas:*
  Registered users: 10GB; Anonymous users: 10MB
- *Sponsor(s):*
  Netherlands Bioinformatics Centre (NBIC) and BiG Grid SURFsara

ballaxy

- *Link:*
  ballaxy
- *Domain/Purpose:*
  Hosts the BALL (Biochemical Algorithms Library) Project tools, i.e. computer aided drug design and molecular modelling based on protein and ligand structure data.
- *Comments:*
  ballaxy is a workflow framework for structure based computational biology based on the Galaxy workflow engine and the BALL (Biochemical Algorithms Library) application framework. It is tailored to handle structural molecular data (pdb, mol, mol2, hin, xyz, smiles) and to offer tools for modelling tasks like chemical shift prediction "NightShift" or optimal bond order assignment of ligands "BOA Constructor".
- *Quotas:*
  A login is required and everybody can create a login, *but there is no guarantee how long any data will be preserved.*
- *Sponsor(s):*
  The groups of Hans-Peter Lenhof (Saarland University, Saarbrücken, Germany), Oliver Kohlbacher (University of Tübingen, Germany), and Andreas Hildebrandt (University of Mainz, Germany).

Cistrome Analysis Pipeline



- *Link:*
  Cistrome Analysis Pipeline
- *Domain/Purpose:*
  ChIP-chip/seq and gene expression data
- *Comments:*
  The Cistrome Analysis Pipeline has the standard Galaxy tools, plus 29 additional ChIP-chip and ChIP-seq specific tools, including preliminary peak calling and correlation analyses, downstream genome feature association, gene expression analyses, and motif discovery.

Gather and Reward!

Summary

Enable

- Galaxy sharing and publication
- toolshed
- provide opportunities to contribute: open mailing lists, conference talks, subgroups

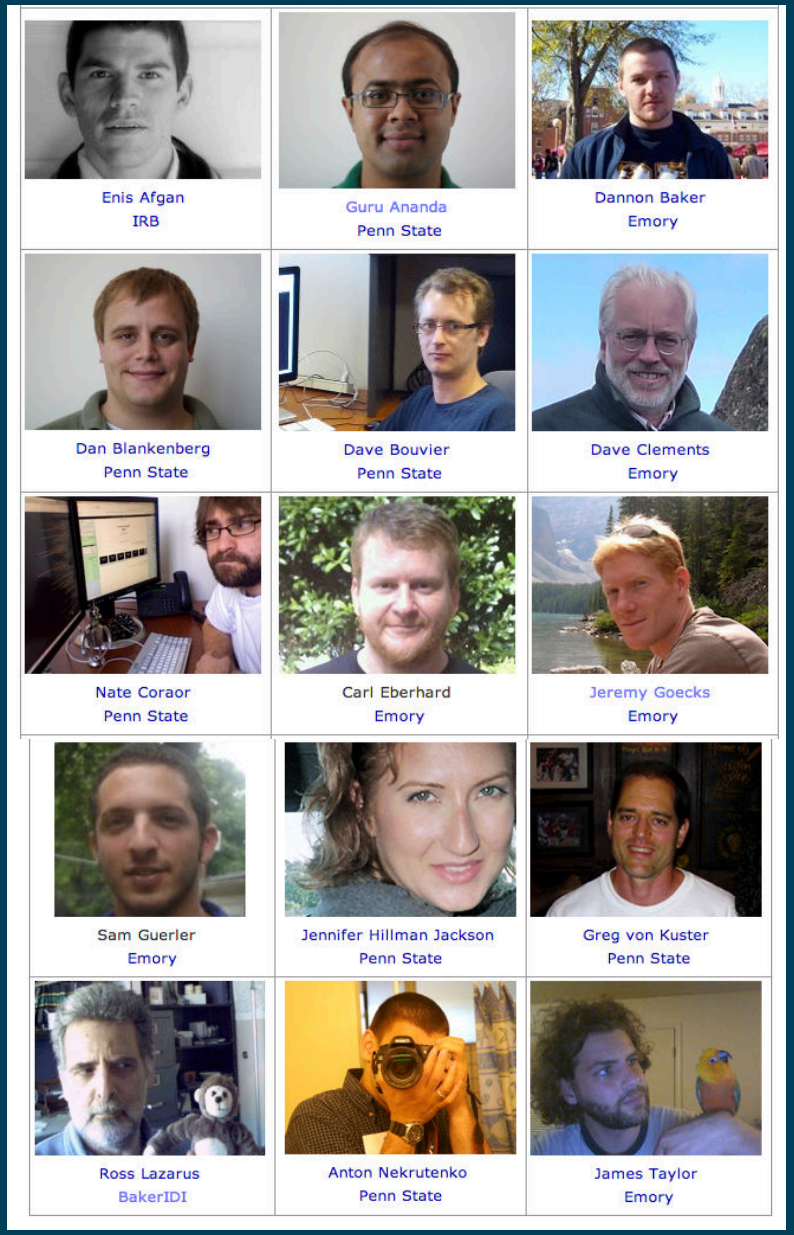
Communicate

- social media: mailing lists, wiki, twitter

Reward

- publicize and praise contributions
- use social media

Galaxy



http://galaxyproject.org
http://wiki.galaxyproject.org
http://usegalaxy.org

http://bitbucket.org/galaxy/galaxy-central

clements@galaxyproject.org
jeremy.goecks@emory.edu

EMORY UNIVERSITY

PENNSTATE 1855

genome.gov
National Human Genome Research Institute
National Institutes of Health

NSF National Science Foundation
WHERE DISCOVERIES BEGIN