Galaxy

Daniel Blankenberg
The Galaxy Team
http://GalaxyProject.org

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the Cloud
- tool shed/contributing tools

The Vision

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

What is Galaxy?

GUI for genomics

- for complete analyses: analyze, visualize, share, publish

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

Overview

What is Galaxy?

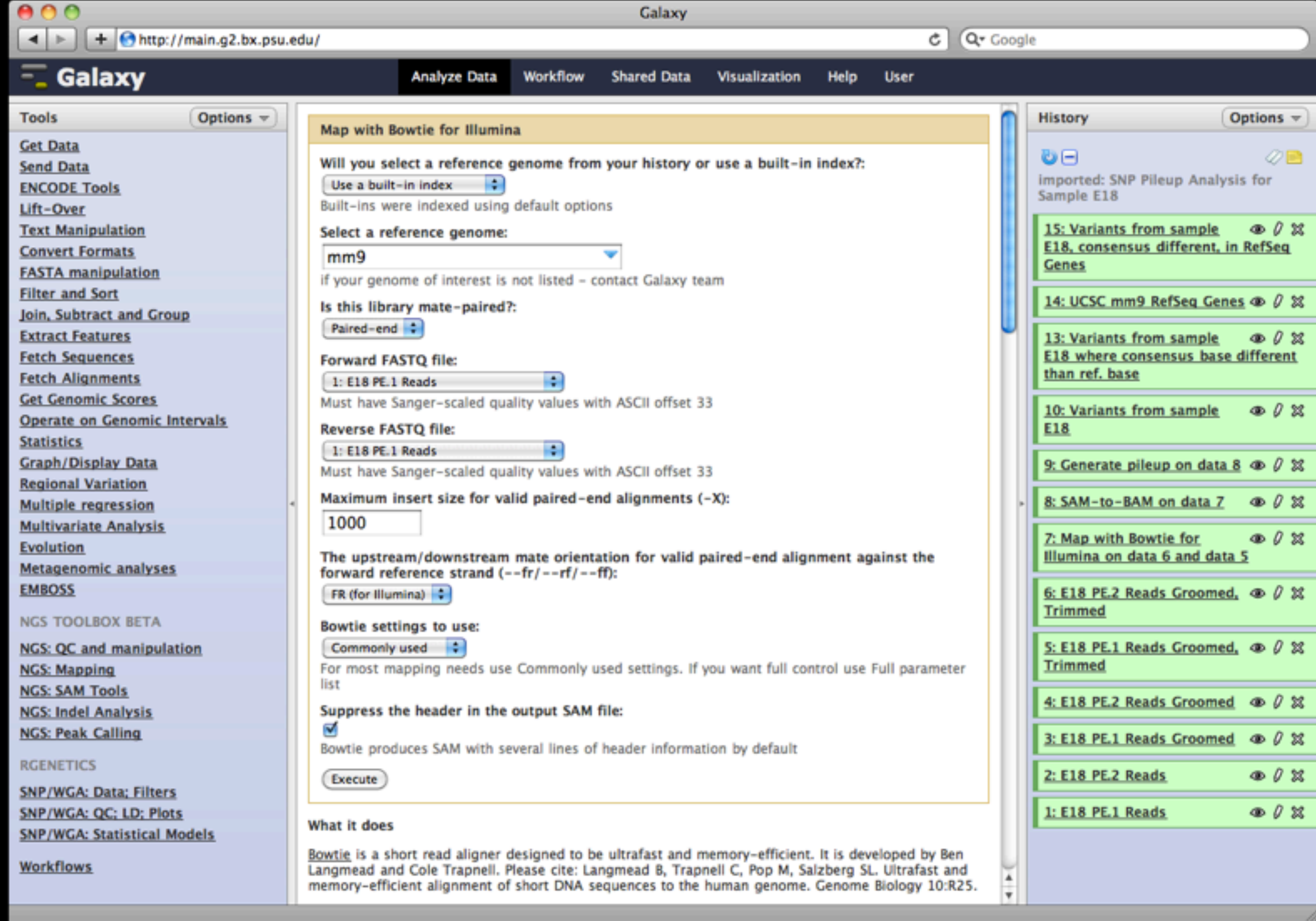
What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

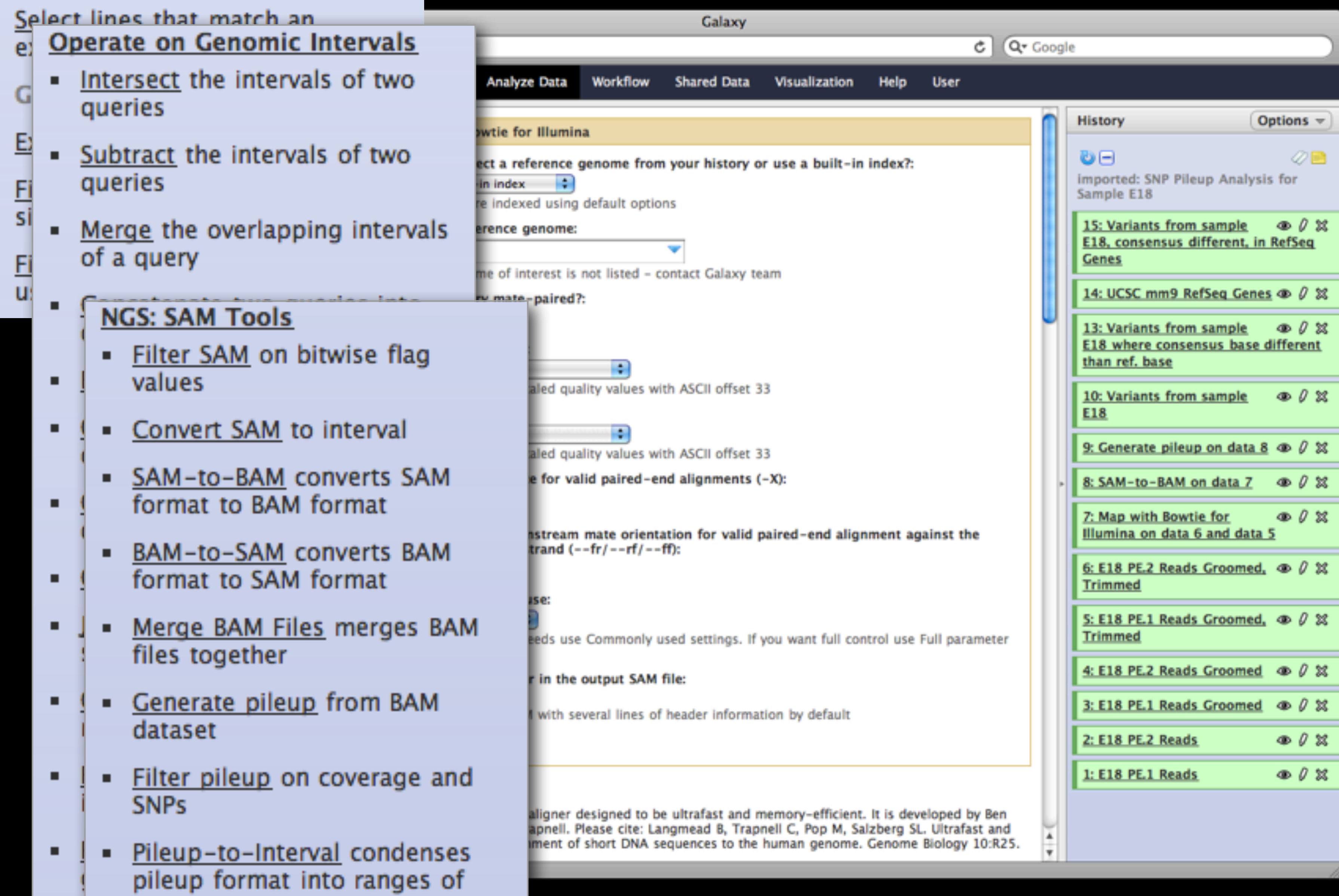
Where you can use and build Galaxy

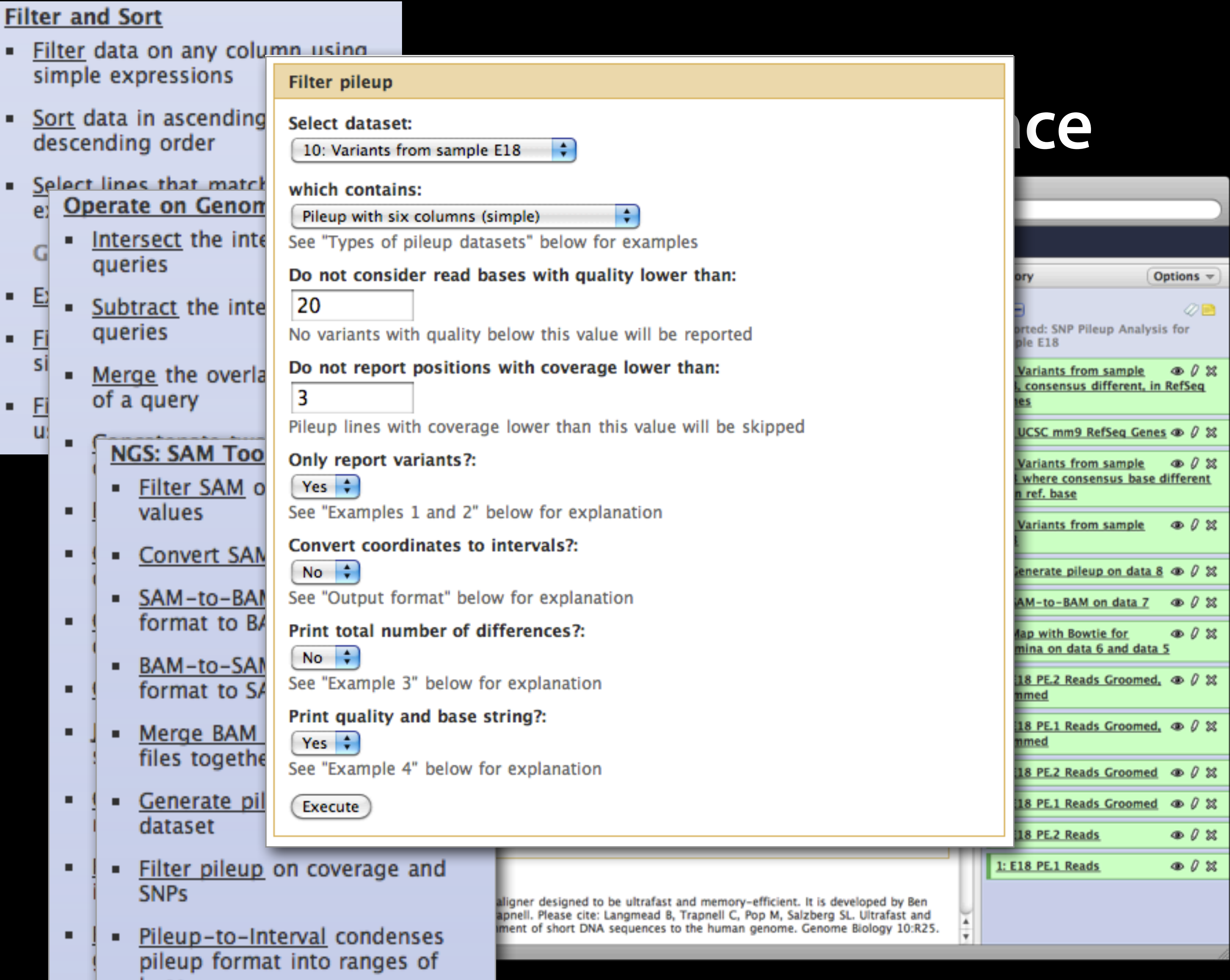
- public website
- local instance
- on the Cloud
- tool shed/contributing tools

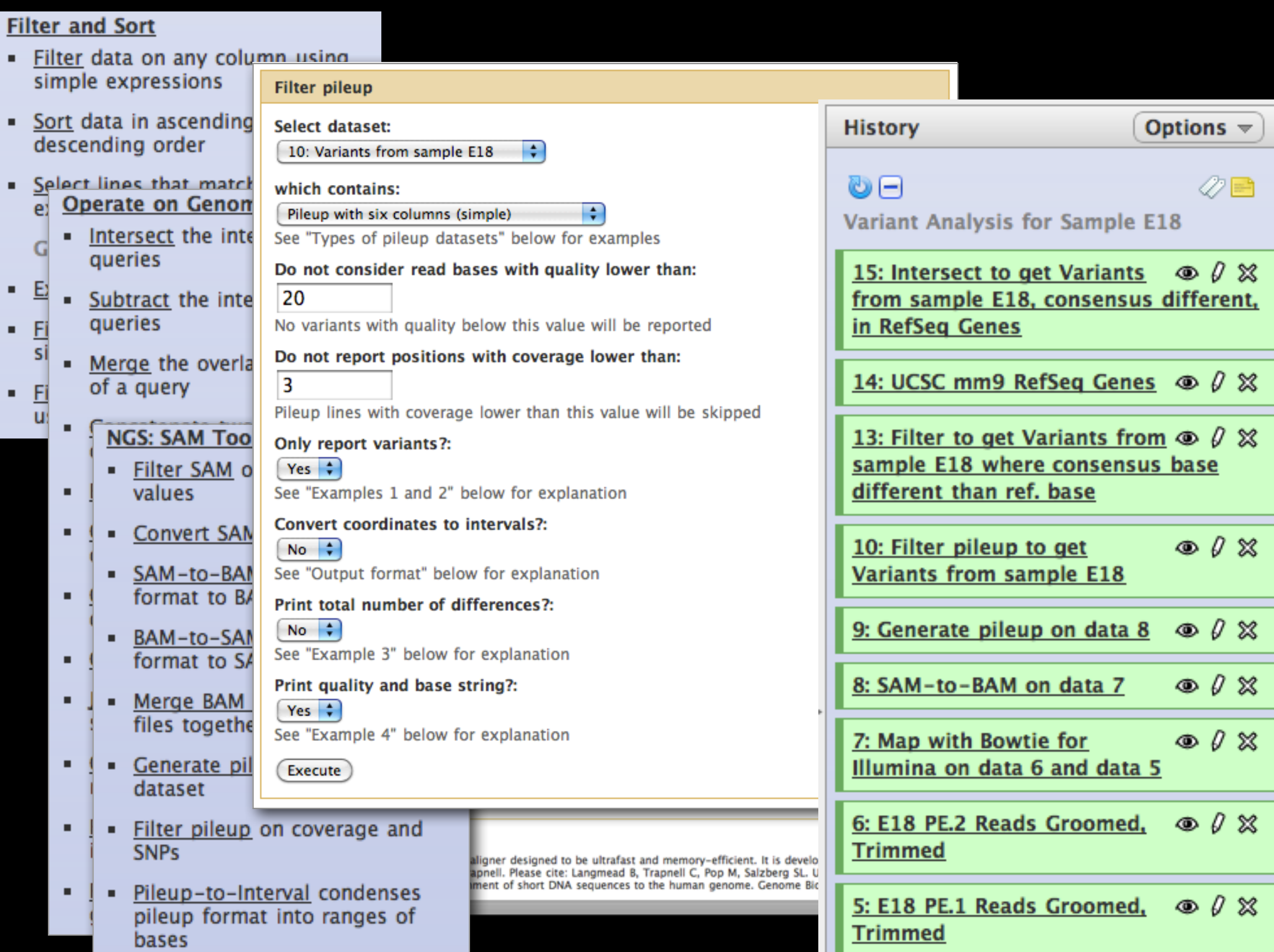
Galaxy Analysis Workspace



xy Analysis Workspace




Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending or descending order
Operate on Genom
Intersect the inte queries
Subtract the inte queries
Merge the overla of a query
NGS: SAM Too
Filter SAM o values
Convert SAM
SAM-to-BAM format to B
BAM-to-SAM format to S
Merge BAM files togethe
Generate pil dataset
Filter pileup on coverage and SNPs
Pileup-to-Interval condenses pileup format into ranges of bases
Filter pileup
Select dataset:
10: Variants from sample E18
which contains:
Pileup with six columns (simple)
See "Types of pileup datasets" below for examples
Do not consider read bases with quality lower than:
20
No variants with quality below this value will be reported
Do not report positions with coverage lower than:
3
Pileup lines with coverage lower than this value will be skipped
Only report variants?:
Yes
See "Examples 1 and 2" below for explanation
Convert coordinates to intervals?:
No
See "Output format" below for explanation
Print total number of differences?:
No
See "Example 3" below for explanation
Print quality and base string?:
Yes
See "Example 4" below for explanation
Execute
Options
UCSC mm9 RefSeq Genes
1: E18 PE.1 Reads


Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending descending order
Operate on Genom
Intersect the inte queries
Subtract the inte queries
Merge the overla of a query
NGS: SAM Too
Filter SAM o values
Convert SAM
SAM-to-BAM format to BA
BAM-to-SAM format to SA
Merge BAM files togethe
Generate pil dataset
Filter pileup on coverage and SNPs
Pileup-to-Interval condenses pileup format into ranges of bases
Filter pileup
Select dataset:
10: Variants from sample E18
which contains:
Pileup with six columns (simple)
See "Types of pileup datasets" below for examples
Do not consider read bases with quality lower than:
20
No variants with quality below this value will be reported
Do not report positions with coverage lower than:
3
Pileup lines with coverage lower than this value will be skipped
Only report variants?:
Yes
See "Examples 1 and 2" below for explanation
Convert coordinates to intervals?:
No
See "Output format" below for explanation
Print total number of differences?:
No
See "Example 3" below for explanation
Print quality and base string?:
Yes
See "Example 4" below for explanation
Execute
History
Options
Variant Analysis for Sample E18
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes
14: UCSC mm9 RefSeq Genes
13: Filter to get Variants from sample E18 where consensus base different than ref. base
10: Filter pileup to get Variants from sample E18
9: Generate pileup on data 8
8: SAM-to-BAM on data 7
7: Map with Bowtie for Illumina on data 6 and data 5
6: E18 PE.2 Reads Groomed, Trimmed
5: E18 PE.1 Reads Groomed, Trimmed

Filter and Sort

- Filter data on any co
  simple expressions
- Sort data in ascendir
  descending order
- Select lines that mat
  e

Operate on Geno

- Intersect the in
  queries
- Subtract the int
  queries
- Merge the over
  of a query

NGS: SAM To

- Filter SAM
  values
- Convert SA
- SAM-to-BA
  format to
- BAM-to-SA
  format to
- Merge BAM
  files togeth
- Generate p
  dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases

This dataset is large and only the first megabyte is shown below.
Show all | Save

chr10	6882036	6882037	A	A	107	0	60	32	.$..,.....C.......
chr10	14243075	14243076	G	G	96	0	60	35	t.
chr10	14243079	14243080	C	C	106	0	60	35	..
chr10	14465082	14465083	T	K	173	176	60	35	GG
chr10	14465083	14465084	G	K	144	144	60	35	..
chr10	14465084	14465085	T	T	117	0	60	38	..
chr10	14465085	14465086	G	G	70	0	60	38	..
chr10	14465257	14465258	C	C	79	0	60	42	..
chr10	14465258	14465259	A	A	137	0	60	46	..
chr10	14465263	14465264	A	A	136	0	60	61	..
chr10	14465366	14465367	A	A	101	0	60	38	g$
chr10	14465371	14465372	G	G	137	0	60	50	.$
chr10	14465410	14465411	G	G	184	0	60	69	.$
chr10	14465447	14465448	T	T	186	0	60	65	.$
chr10	14465456	14465457	G	G	193	0	60	70	..
chr10	14465465	14465466	T	T	177	0	60	63	.$
chr10	14465485	14465486	C	T	129	129	60	34	t$
chr10	14465569	14465570	T	T	219	0	60	84	.,
chr10	14465581	14465582	G	G	240	0	60	84	,$
chr10	14465586	14465587	C	C	248	0	60	82	.$
chr10	14465621	14465622	C	C	134	0	60	49	.,
chr10	14465658	14465659	C	C	134	0	60	49	,,
chr10	14465660	14465661	T	T	153	0	60	55	,,
chr10	14465691	14465692	G	G	128	0	60	42	,$
chr10	14465778	14465779	C	C	89	0	60	34	,$
chr10	14465791	14465792	G	G	104	0	60	33	,$
chr10	14465881	14465882	G	G	110	0	60	41	..
chr10	17445088	17445089	A	A	103	0	60	34	..
chr10	17445271	17445272	A	A	55	0	60	34	..
chr10	17731269	17731270	T	T	113	0	60	42	,$
chr10	19928287	19928288	G	A	135	135	60	36	AA
chr10	19928468	19928469	C	T	132	132	60	35	T$
chr10	19928488	19928489	A	A	119	0	60	44	..
chr10	19928494	19928495	C	T	138	138	60	37	TT
chr10	19928527	19928528	A	A	134	0	60	45	,,
chr10	19928538	19928539	G	G	144	0	60	52	,$
chr10	19928543	19928544	A	G	147	147	60	40	Gg
chr10	19928741	19928742	T	T	80	0	60	30	,,
chr10	20799826	20799827	G	G	117	0	60	37	,$
chr10	28750217	28750218	C	T	138	138	60	37	TT
chr10	28750397	28750398	A	C	154	211	60	64	C$
chr10	28750401	28750402	A	A	128	0	60	47	,$
chr10	28750423	28750424	C	T	113	113	60	35	T$
chr10	28750438	28750439	A	A	95	0	60	36	.$
chr10	28750446	28750447	A	G	165	165	60	46	G$
chr10	28750487	28750488	A	A	80	0	60	31	,,
chr10	28750512	28750513	G	G	220	0	60	72	.$
chr10	28750548	28750549	G	C	255	255	60	97	C$
chr10	28750574	28750575	T	T	237	0	60	83	.$
chr10	28750577	28750578	T	T	234	0	60	82	,$
chr10	28750578	28750579	T	T	242	0	60	76	,$
chr10	28750593	28750594	G	G	220	0	60	75	.$
chr10	28750640	28750641	T	C	165	165	60	46	Cc
chr10	28750746	28750747	G	A	202	202	60	58	AA
chr10	28750766	28750767	A	G	205	205	60	59	G$
chr10	28750769	28750770	T	C	175	175	60	49	cc

aligner designed to be ultrafast and memory-efficient. It is develo
apnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. U
ment of short DNA sequences to the human genome. Genome Bio

Analysis

Options

nalysis for Sample E18

sect to get Variants
nple E18, consensus different,
Genes

mm9 RefSeq Genes

to get Variants from
18 where consensus base
than ref. base

pileup to get
from sample E18

ate pileup on data 8

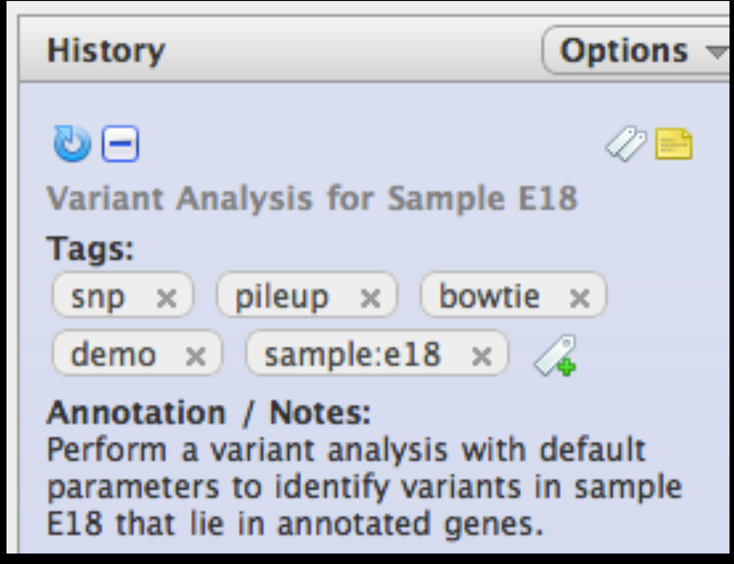
o-BAM on data 7

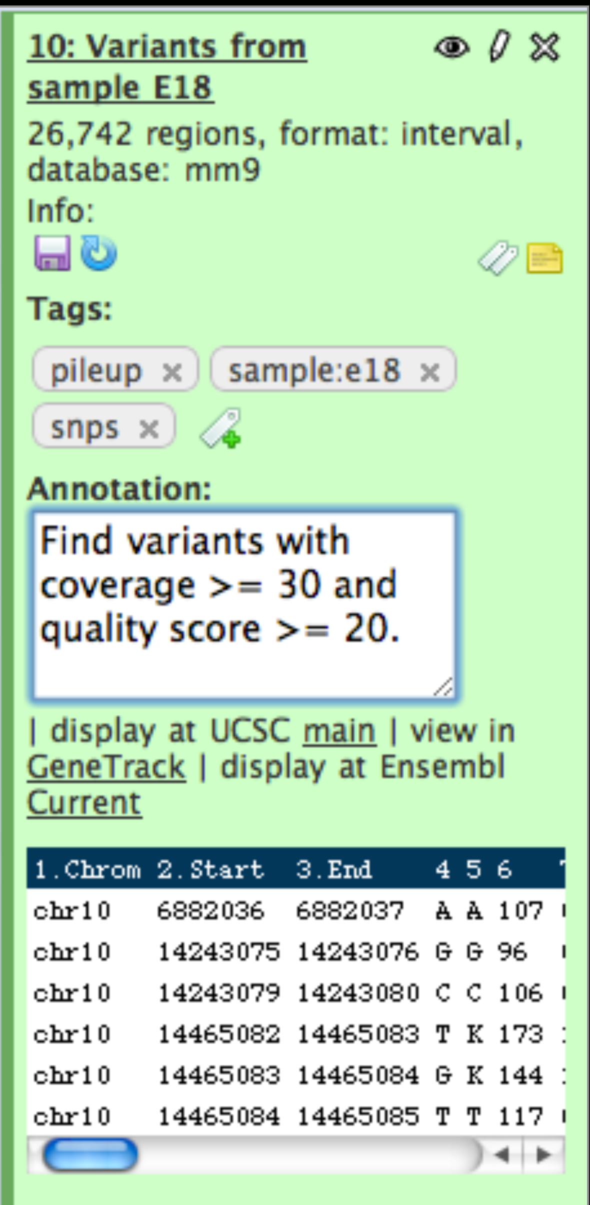
ith Bowtie for
on data 6 and data 5

6: E18 PE.2 Reads Groomed, Trimmed

5: E18 PE.1 Reads Groomed, Trimmed

User Metadata





Datasources

Upload file from your computer

- FTP support for large datasets

Files directly from a sequencer

- Sample Tracking System

UCSC table browser

BioMart

interMine / modMine

EuPathDB server

EncodeDB at NHGRI

EpiGRAPH server

Tool Suites

Text Manipulation
Format Converters
Filtering and Sorting
Join, Subtract, Group
Sequence Tools
Multi-species Alignment Tools
Genomic Interval Operations
Summary Statistics
Graphing / Plotting

Regional Variation
EMBOSS
Evolution / Phylogeny
RNA-seq
ChIP-seq
GATK
Picard
RGenetics
...and more

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the Cloud
- tool shed/contributing tools

Data Library "Bushman"

Library Actions

These are the data underlying the analyses reported in the paper "Complete Khoisan and Bantu genomes from southern Africa" by S. C. Schuster et al., published in the journal Nature, February 18, 2010. Each data set can be downloaded and/or imported into a Galaxy history. Data will be updated as the project progresses.

Name	Information	Uploaded By	Date	File Size
All SNPs in personal genomes	Summary table of SNPs in all individuals	greg@bx.psu.edu	2010-01-28	676.8 Mb
Alu insertions in KB1		greg@bx.psu.edu	2010-02-10	14.9 Kb
Alu insertions in NB1		greg@bx.psu.edu	2010-02-10	6.5 Kb
KB1_microsatellites.txt		greg@bx.psu.edu	2010-02-15	3.5 Mb
NB1_microsatellites.txt		greg@bx.psu.edu	2010-02-15	828.5 Kb
amino acid differences with functional predictions		greg@bx.psu.edu	2010-02-05	1.1 Mb
gene copy numbers in KB1 and other personal genomes		greg@bx.psu.edu	2010-02-15	2.1 Mb
indels in ABT		greg@bx.psu.edu	2010-02-03	105.3 Kb
indels in KB1		greg@bx.psu.edu	2010-02-03	14.2 Mb
indels in MD8		greg@bx.psu.edu	2010-02-03	109.8 Kb
indels in NB1		greg@bx.psu.edu	2010-02-03	[illegible] Kb
indels in TK1		greg@bx.psu.edu	2010-02-03	123.2 Kb
novel SNPs in ABT		greg@bx.psu.edu	2010-02-09	9.4 Mb
novel SNPs in KB1		greg@bx.psu.edu	2010-02-09	16.9 Mb
novel SNPs in MD8		greg@bx.psu.edu	2010-02-09	594.1 Kb
novel SNPs in NB1		greg@bx.psu.edu	2010-02-09	4.1 Mb
novel SNPs in TK1		greg@bx.psu.edu	2010-02-09	722.6 Kb
sequenced exon-containing intervals		greg@bx.psu.edu	2010-02-03	3.1 Mb

For selected items: Import into your current history Go

http://usegalaxy.org/bushman

Managing Libraries

Loading Data

- Upload a single file
- Import datasets from a Galaxy history
- Upload a directory of files
- Directly from Sequencer using Sample Tracking System

Accessing Data

- Data contents on disk are not copied
- Dataset security: public, Role-based access control (RBAC)

Annotating Library Data: Library Templates

- Build user fillable forms
- Associate at Library, Folder or Dataset level

Overview

What is Galaxy?

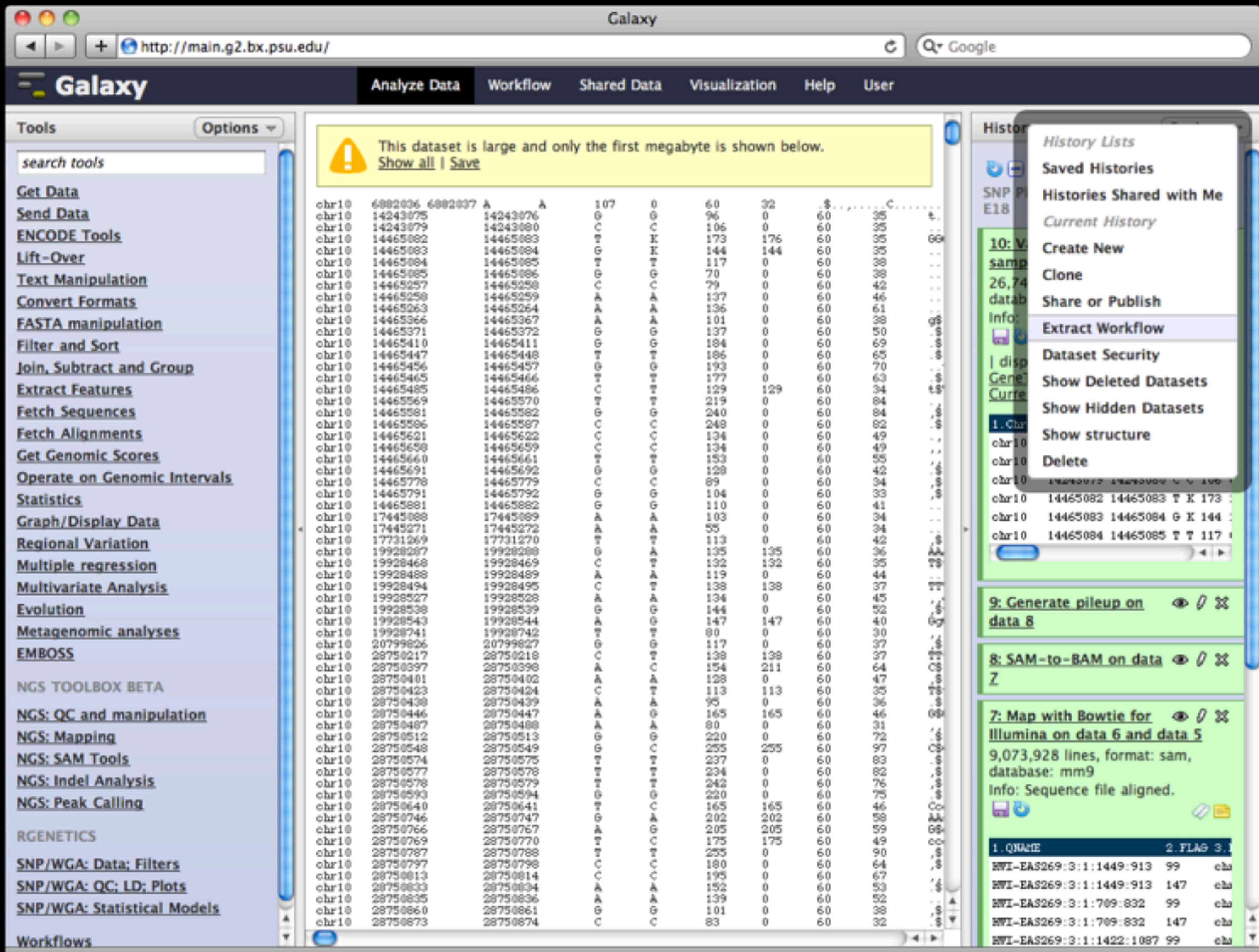
What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

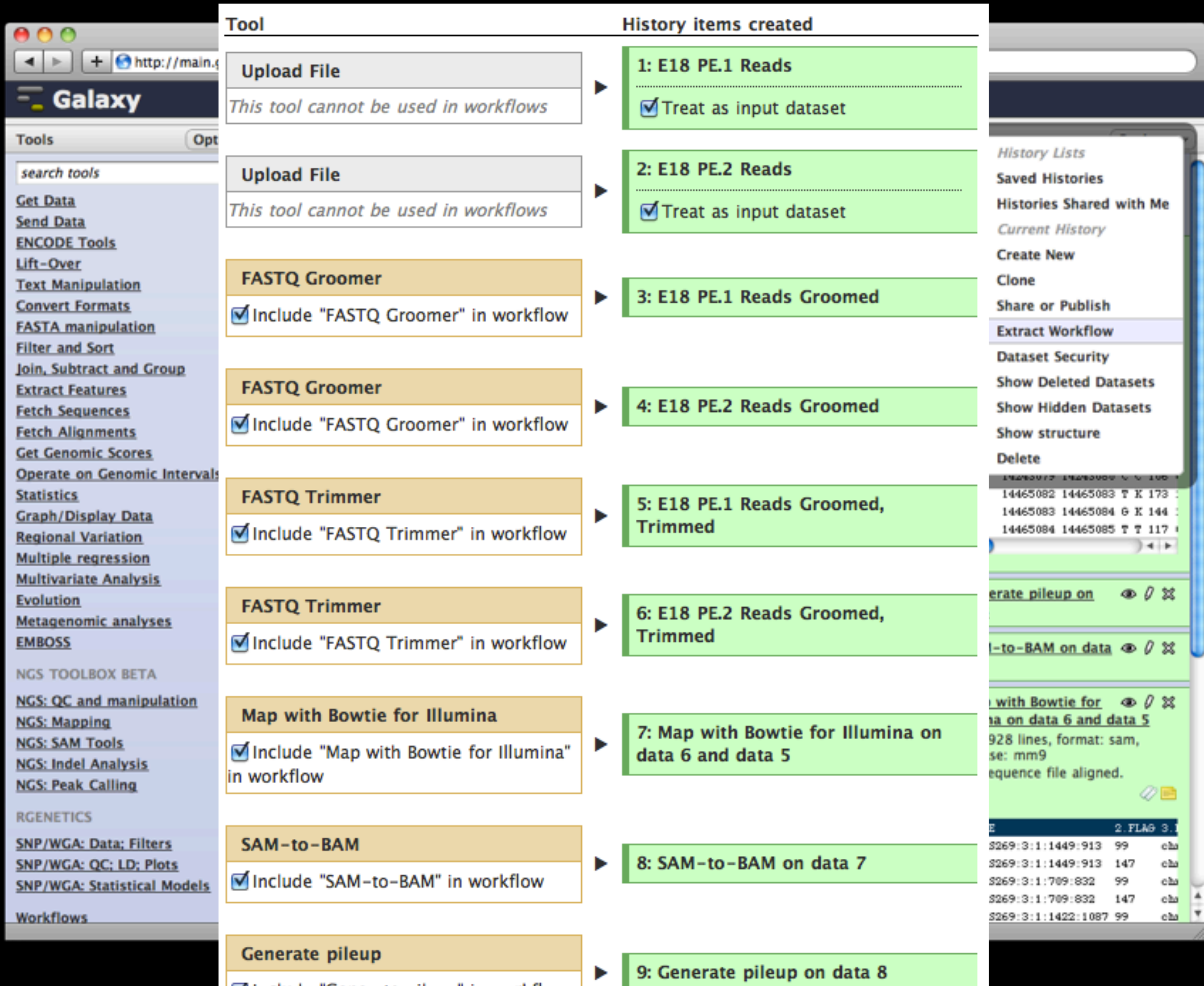
Where you can use and build Galaxy

- public website
- local instance
- on the Cloud
- tool shed/contributing tools

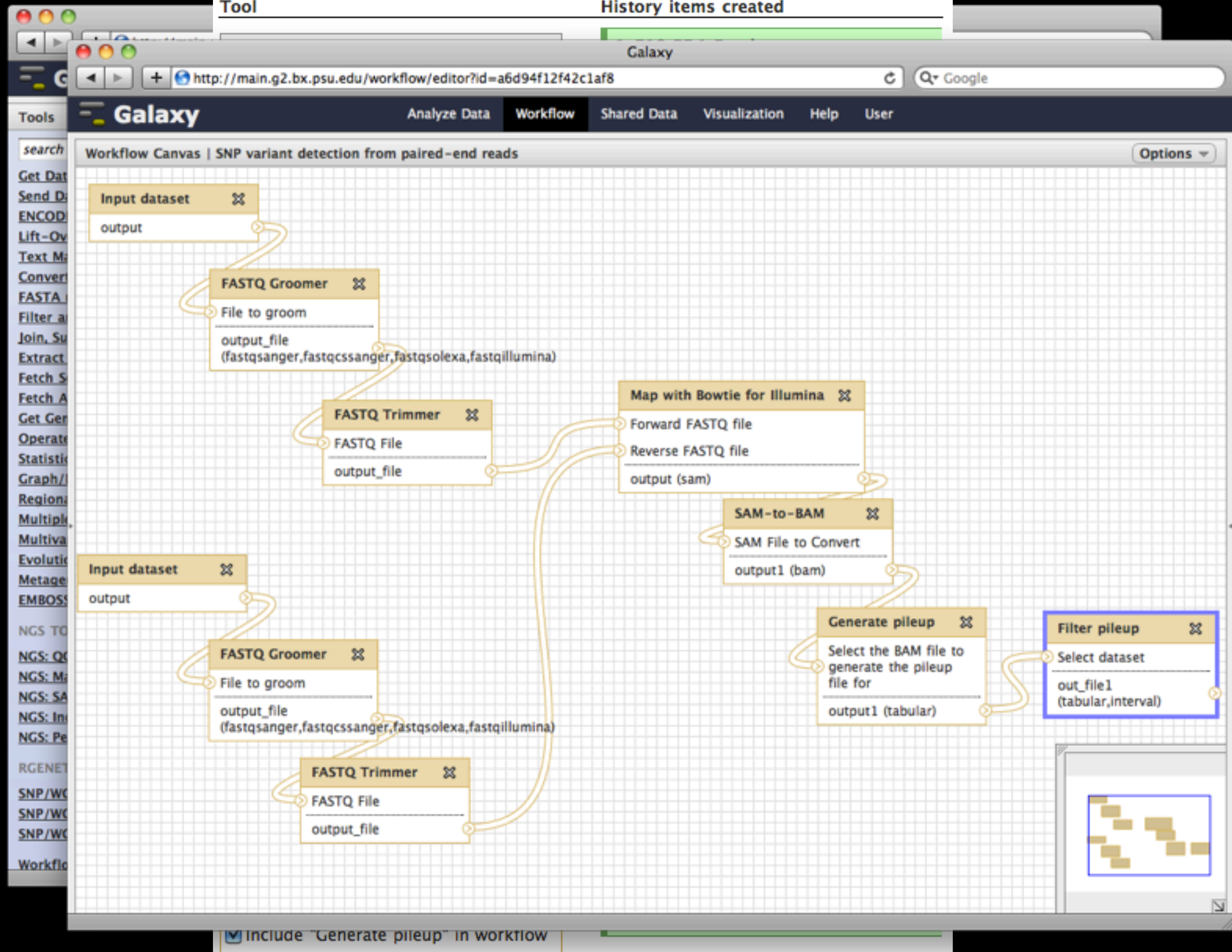
Galaxy Workflows



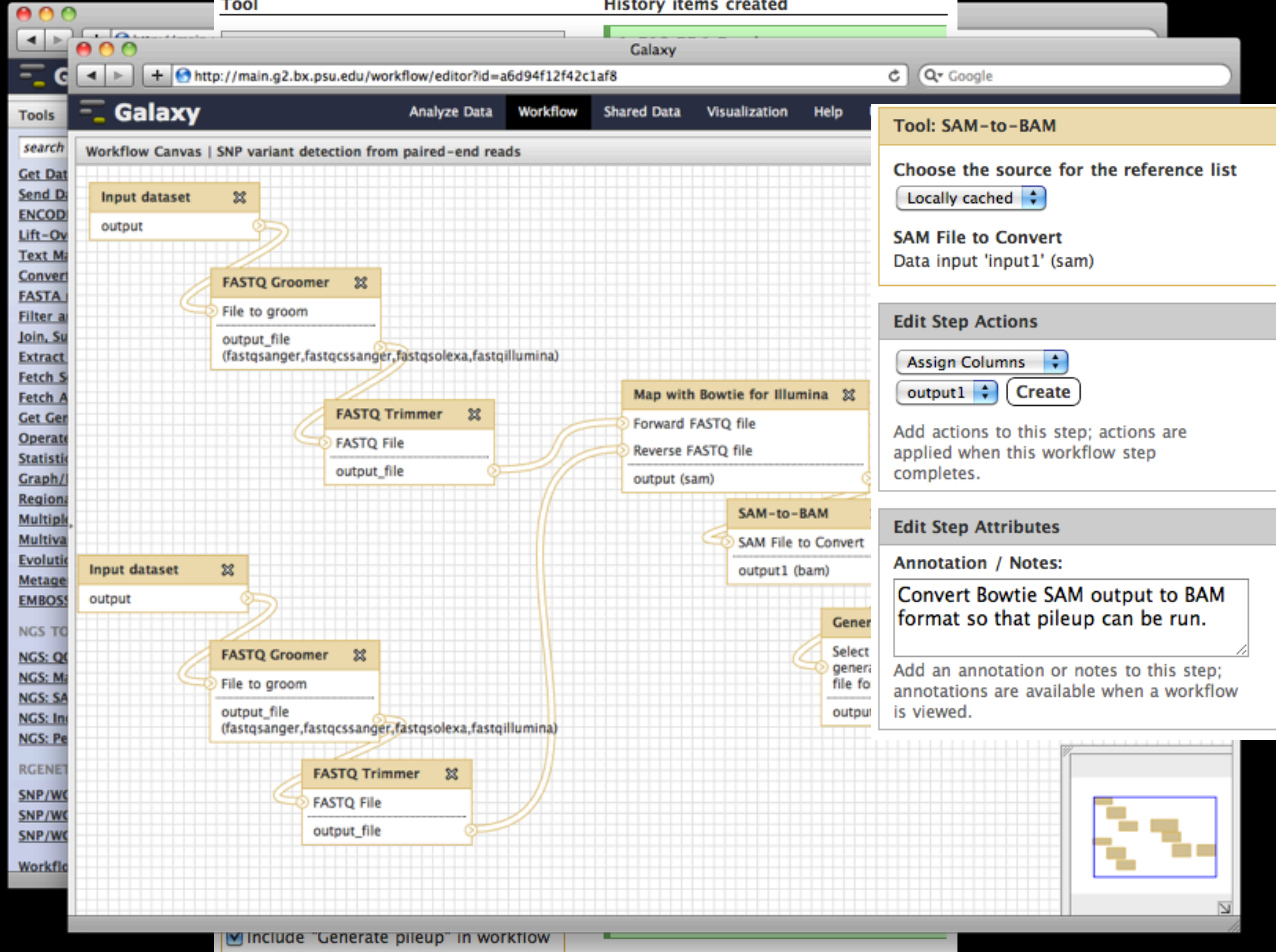
Galaxy Workflows



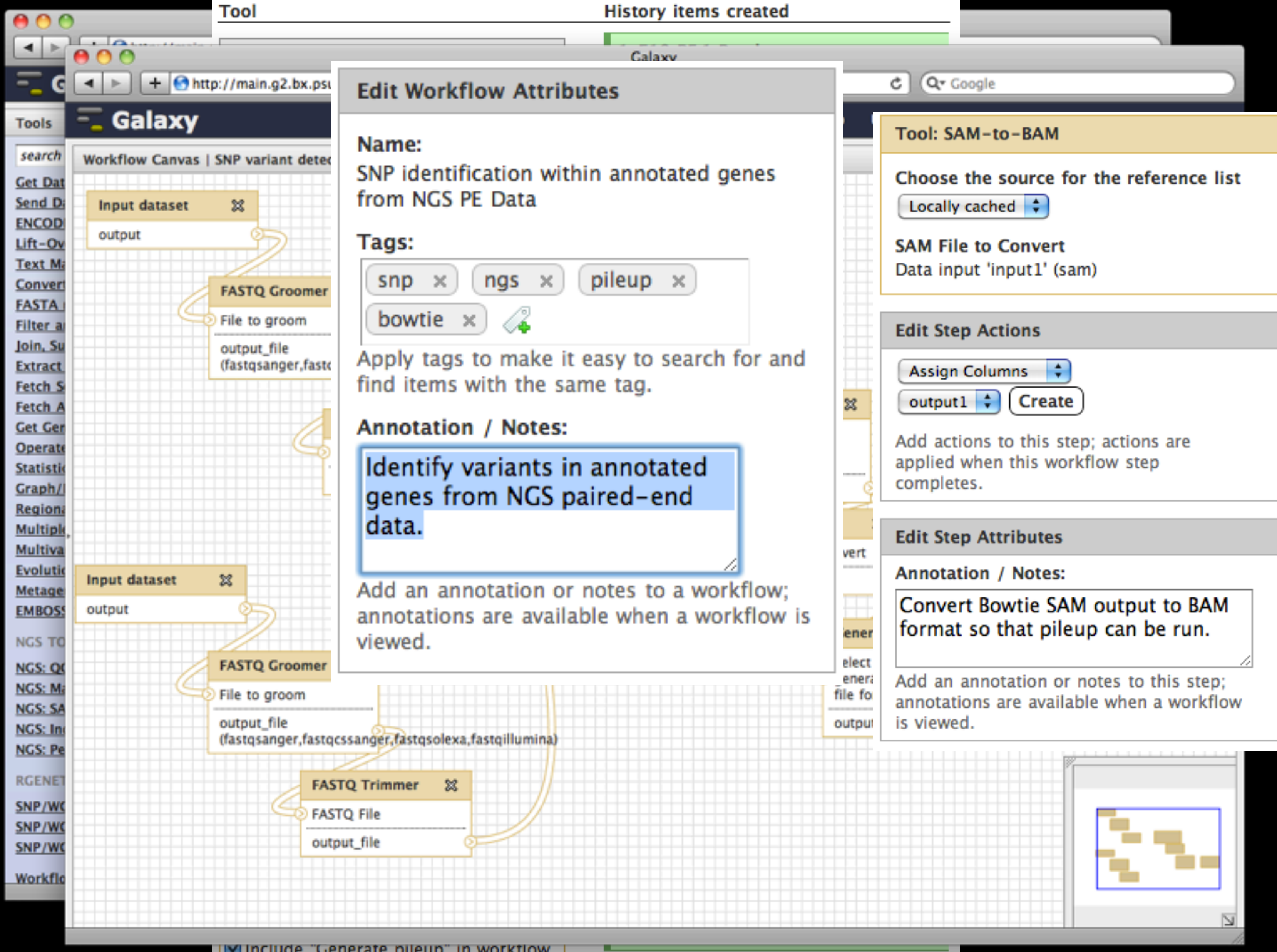
Galaxy Workflows

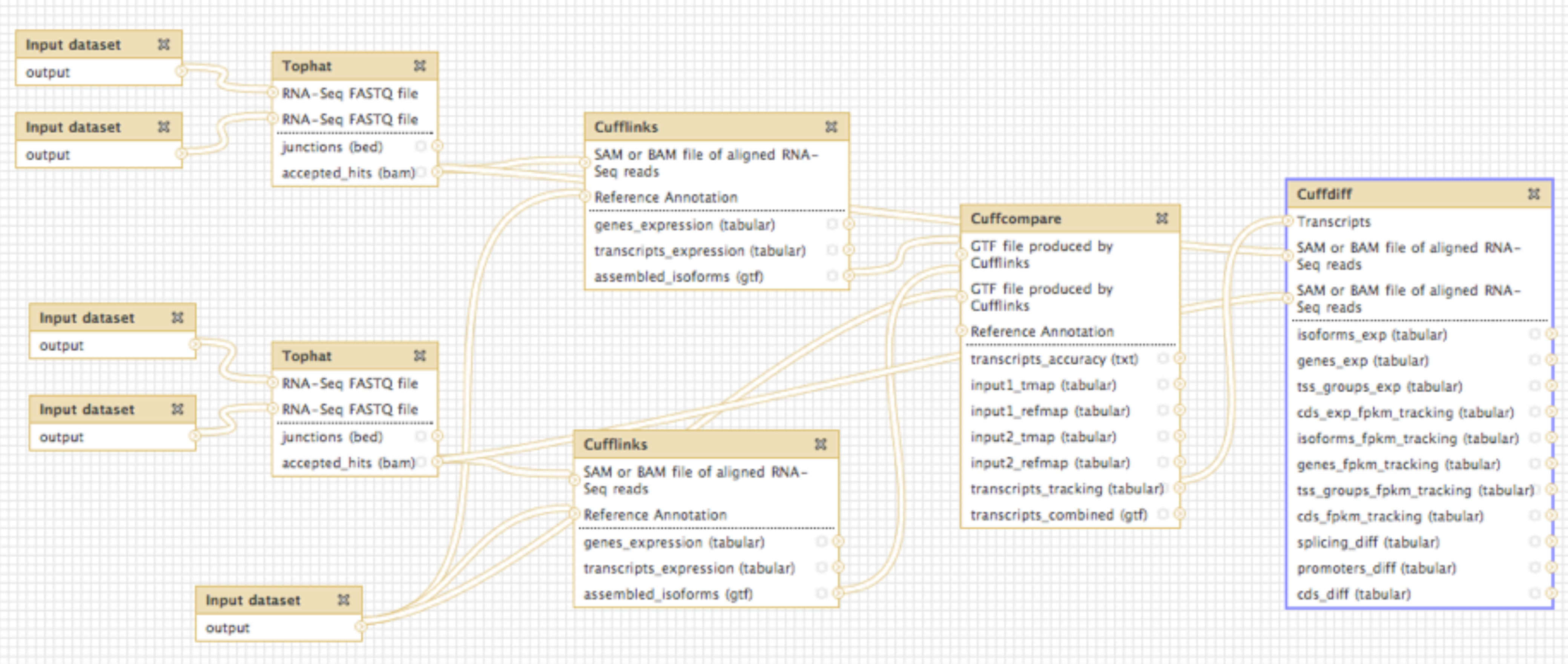


Galaxy Workflows

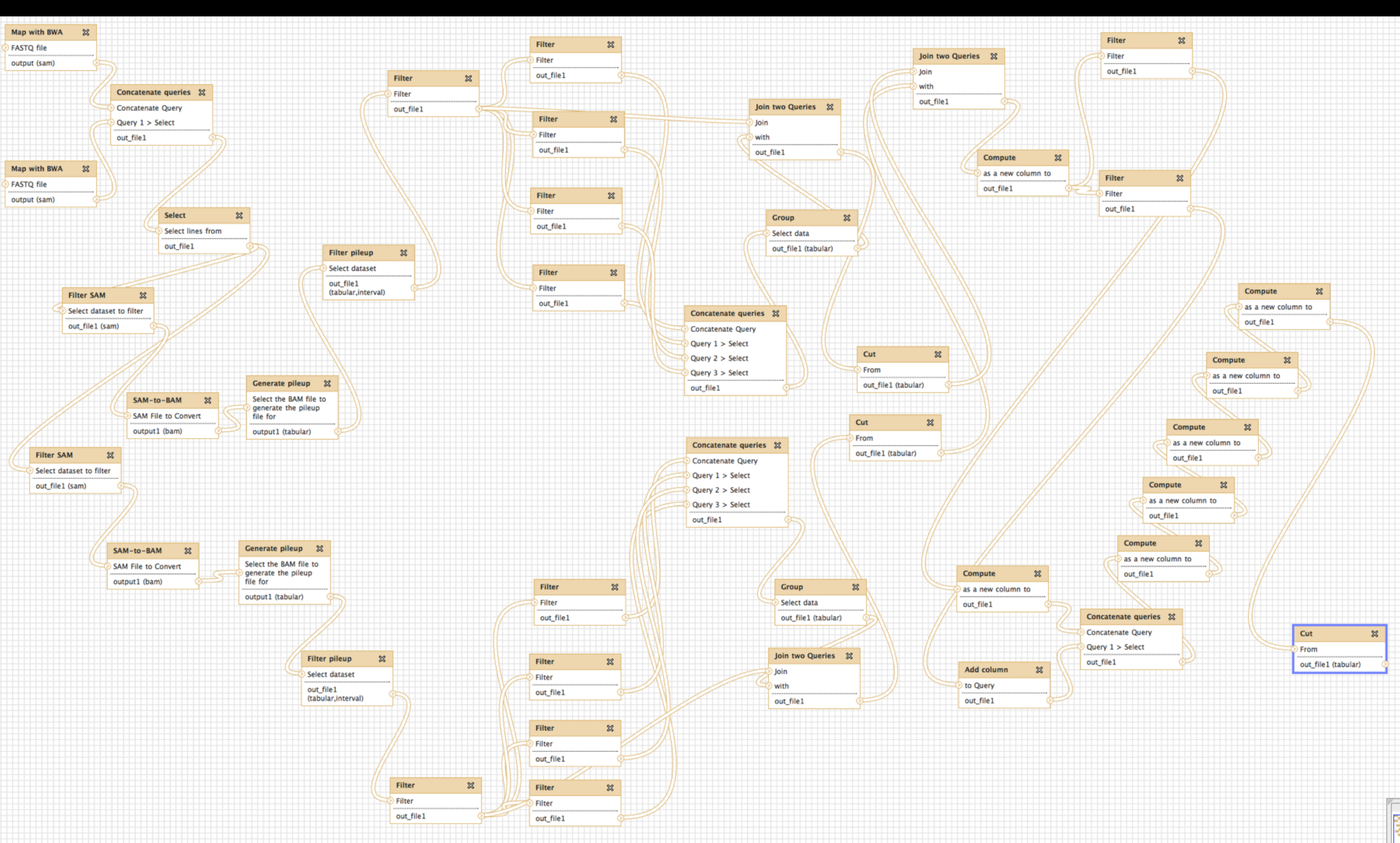


Galaxy Workflows





Example: Workflow for differential expression analysis of RNA-seq using Tophat/ Cufflinks tools



Example: Diagnosing low-frequency heterosplasmic sites in two tissues from the same individual

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the Cloud
- tool shed/contributing tools

Visualize

Send data results to external genome browsers

Trackster: Galaxy's genome browser

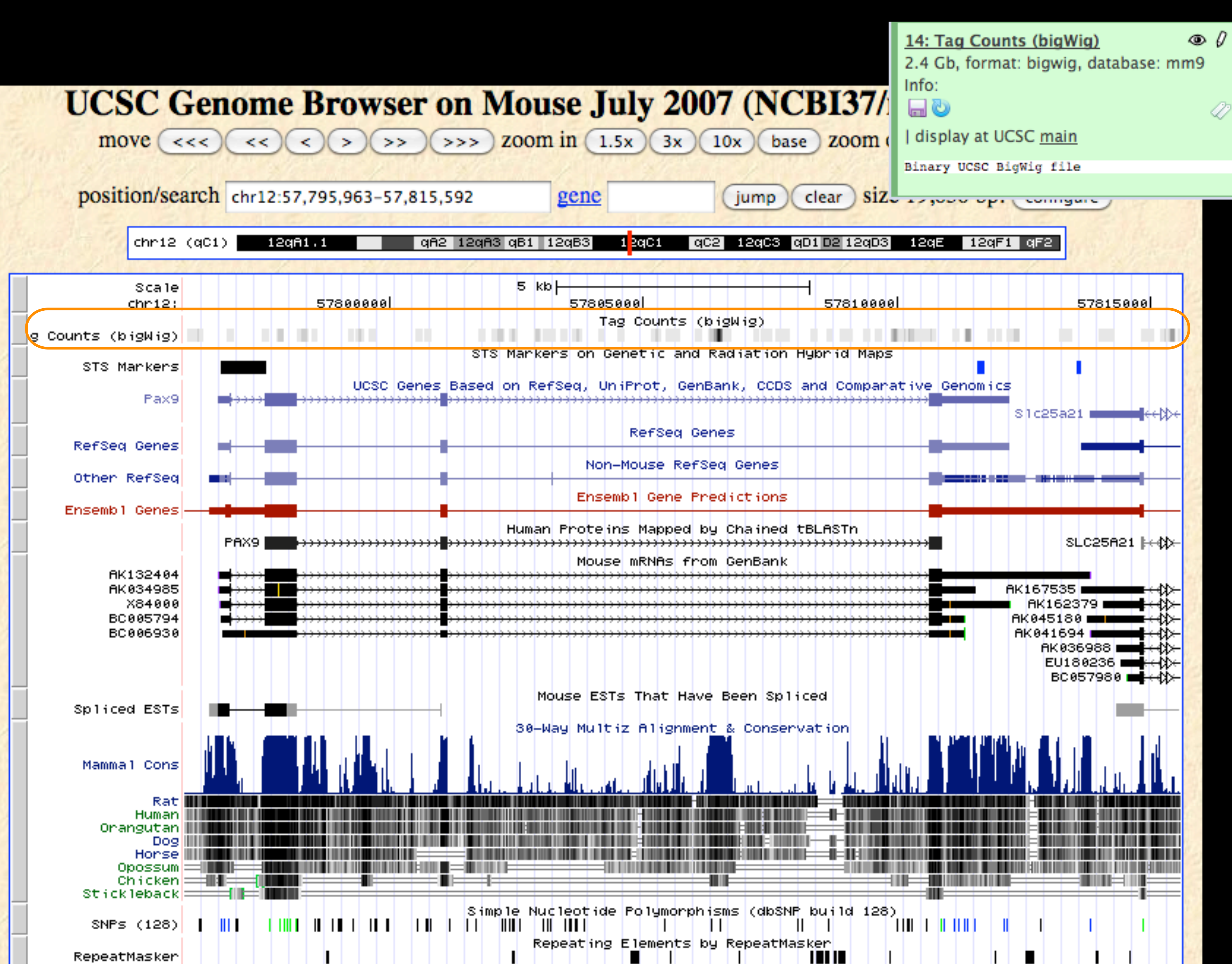
External Genome Browsers

UCSC

Ensembl

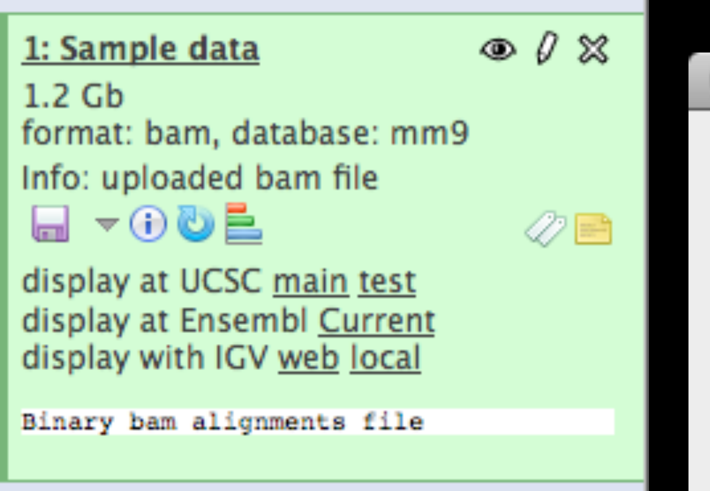
GBrowse

IGV



UCSC Genome Browser on Mouse July 2007 (NCBI37/
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom
position/search chr12:57,795,963-57,815,592 gene jump clear
chr12 (qC1)
14: Tag Counts (bigWig)
2.4 Gb, format: bigwig, database: mm9
Info:
| display at UCSC main
Binary UCSC BigWig file
Tag Counts (bigWig)
STS Markers on Genetic and Radiation Hybrid Maps
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
RefSeq Genes
Non-Mouse RefSeq Genes
Ensembl Gene Predictions
Human Proteins Mapped by Chained tBLASTn
Mouse mRNAs from GenBank
Mouse ESTs That Have Been Spliced
30-Way Multiz Alignment & Conservation
Simple Nucleotide Polymorphisms (dbSNP build 128)
Repeating Elements by RepeatMasker

Integrative Genomics Viewer (IGV)



Galaxy

- tool integration framework
- heavy focus on usability
- sharing, publication framework

Genome Browser

- physical depiction of data
- visually identify correlations
- find interesting regions, features

Galaxy

- tool integration framework
- heavy focus on usability
- sharing, publication framework

Genome Browser

- physical depiction of data
- visually identify correlations
- find interesting regions, features

Trackster

View your data from within Galaxy

- No data transfers to external site
- Use it locally, even without internet access

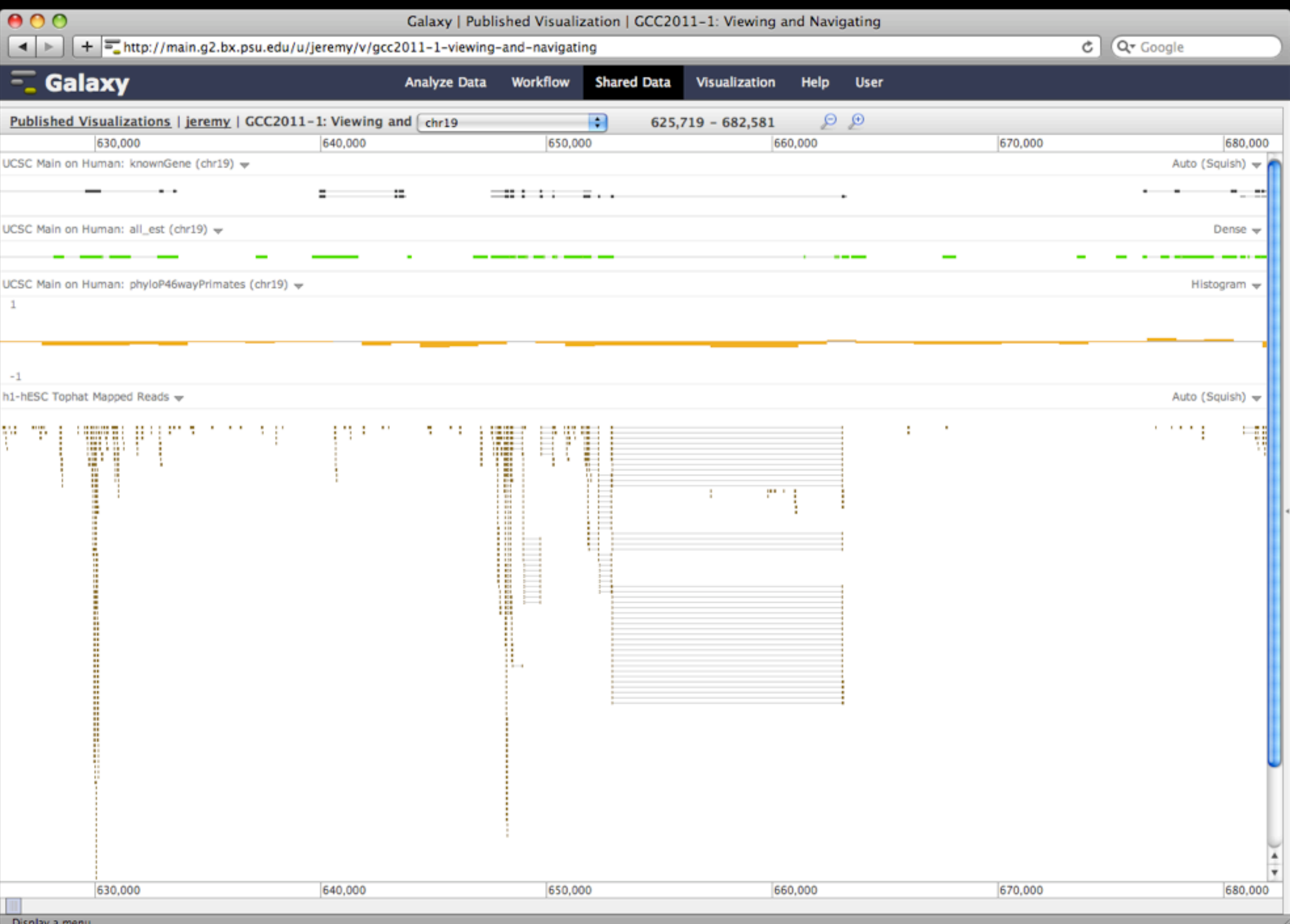
Supports common filetypes

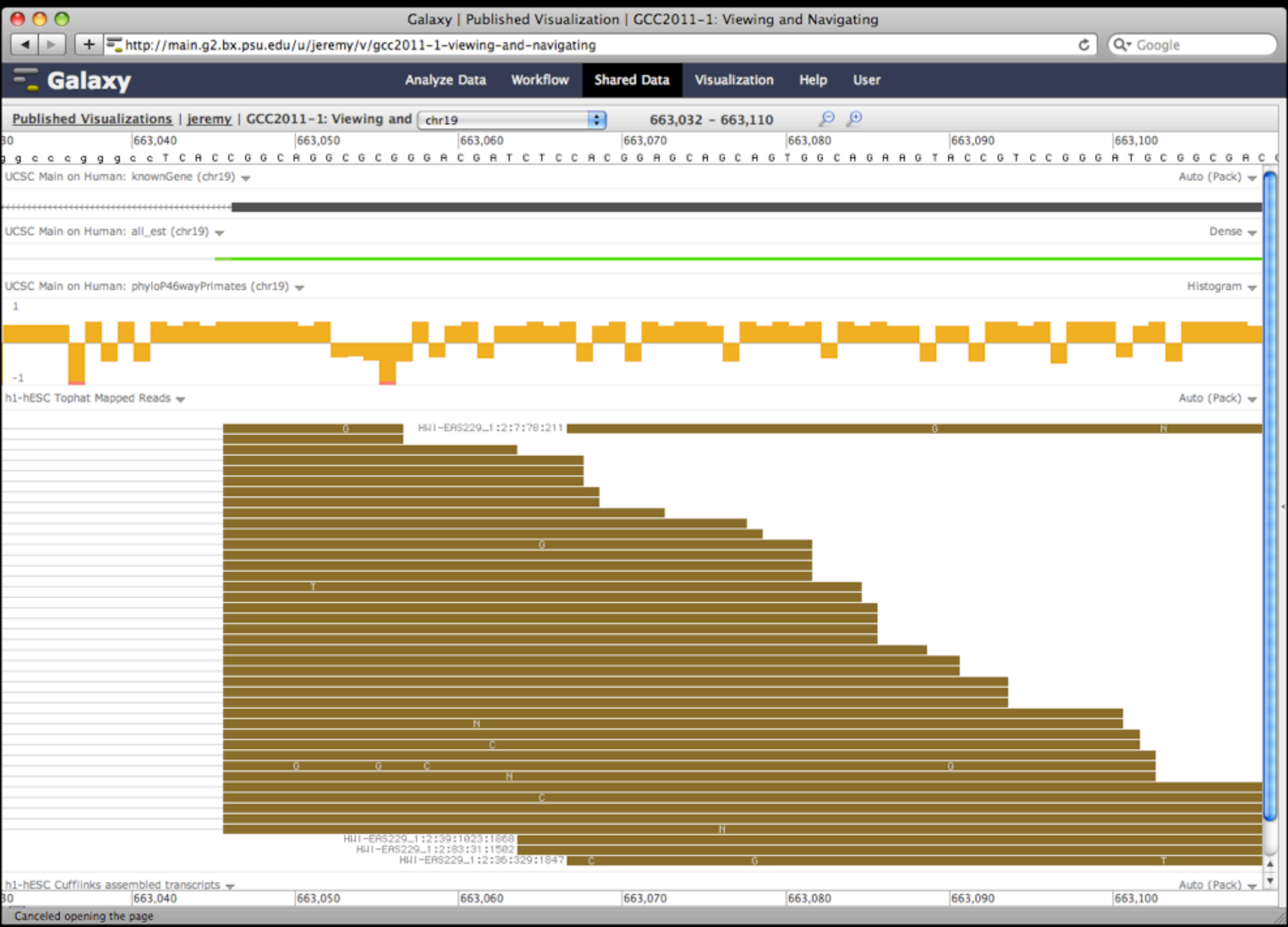
- BAM, BED, GFF/GTF, WIG

Unique features

- custom genomes
- highly interactive

Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating
http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-1: Viewing and
chr19
1,290 - 4,168,475
0
1,000,000
2,000,000
3,000,000
UCSC Main on Human: knownGene (chr19)
Auto (Squish)
UCSC Main on Human: all_est (chr19)
Auto (coverage histogram)
11431
UCSC Main on Human: phyloP46wayPrimates (chr19)
Histogram
1
-1
h1-hESC Tophat Mapped Reads
Auto (coverage histogram)
8732
h1-hESC Cufflinks assembled transcripts
Auto (Squish)
Display a menu


Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating
http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-1: Viewing and
chr19
625,719 - 682,581
630,000
640,000
650,000
660,000
670,000
680,000
UCSC Main on Human: knownGene (chr19)
Auto (Squish)
UCSC Main on Human: all_est (chr19)
Dense
UCSC Main on Human: phyloP46wayPrimates (chr19)
Histogram
1
-1
h1-hESC Tophat Mapped Reads
Auto (Squish)
Display a menu



Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating
http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-1: Viewing and
chr19
663,032 - 663,110
UCSC Main on Human: knownGene (chr19)
Auto (Pack)
UCSC Main on Human: all_est (chr19)
Dense
UCSC Main on Human: phyloP46wayPrimates (chr19)
Histogram
h1-hESC Tophat Mapped Reads
Auto (Pack)
h1-hESC Cufflinks assembled transcripts
Auto (Pack)
Canceled opening the page

But really, why *another* genome browser

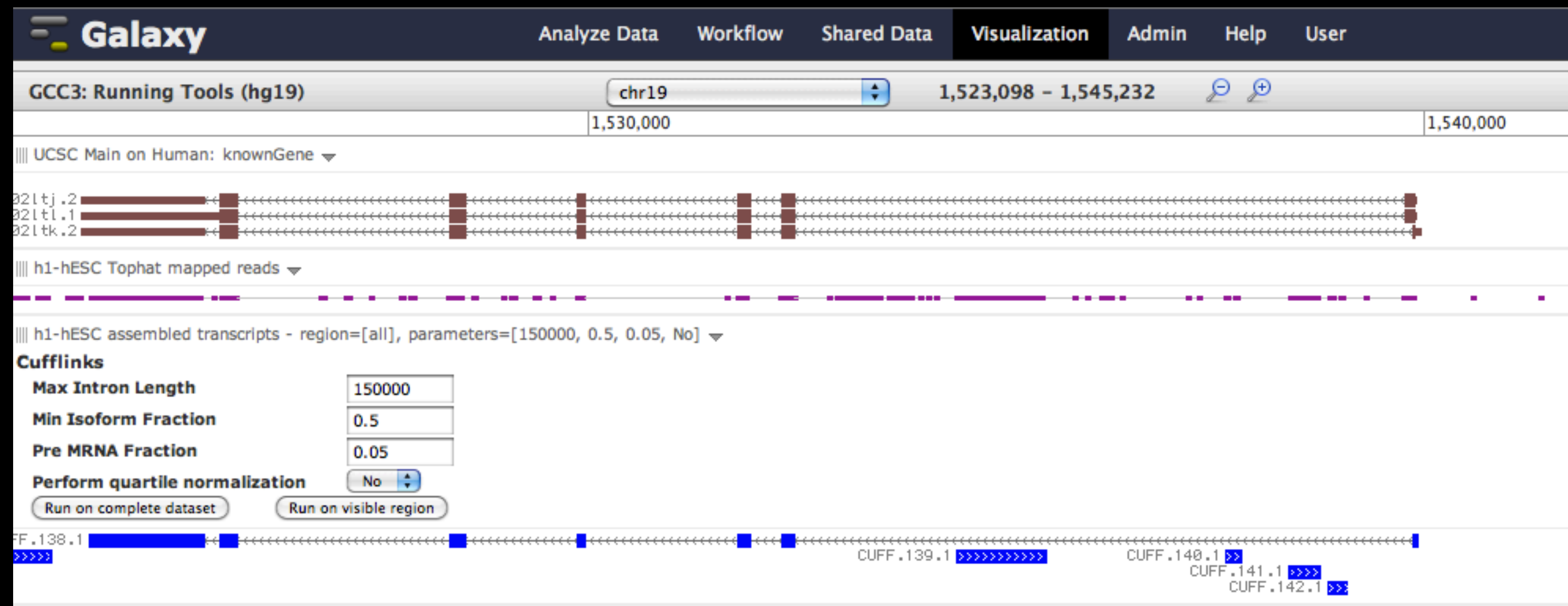
From static browsing to visual analysis

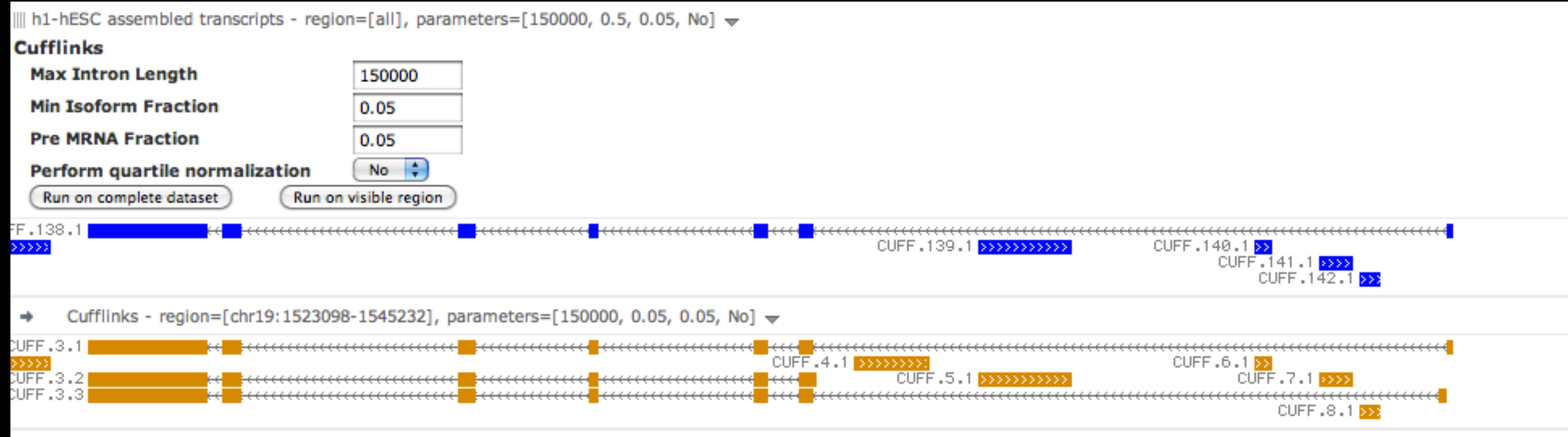
Visual feedback and experimentation needed for complex tools with many parameters

Leverage Galaxy strengths: a very sound model for abstracting interfaces to analysis tools and already integrates an enormous number

Dynamic Filtering

Integrating Tools and Visualization




h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No]
Cufflinks
Max Intron Length
150000
Min Isoform Fraction
0.05
Pre MRNA Fraction
0.05
Perform quartile normalization
No
Run on complete dataset
Run on visible region
CUFF.139.1
CUFF.140.1
CUFF.141.1
CUFF.142.1
Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.05, No]
CUFF.4.1
CUFF.5.1
CUFF.6.1
CUFF.7.1
CUFF.8.1

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the Cloud
- tool shed/contributing tools

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history is currently restricted so that only you and the users listed below can access it. You can:

Make History Accessible via Link

Generates a web link that you can share with other people so that they can view and import the history.

Make History Accessible and Publish

Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, where it is publicly listed and searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List

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

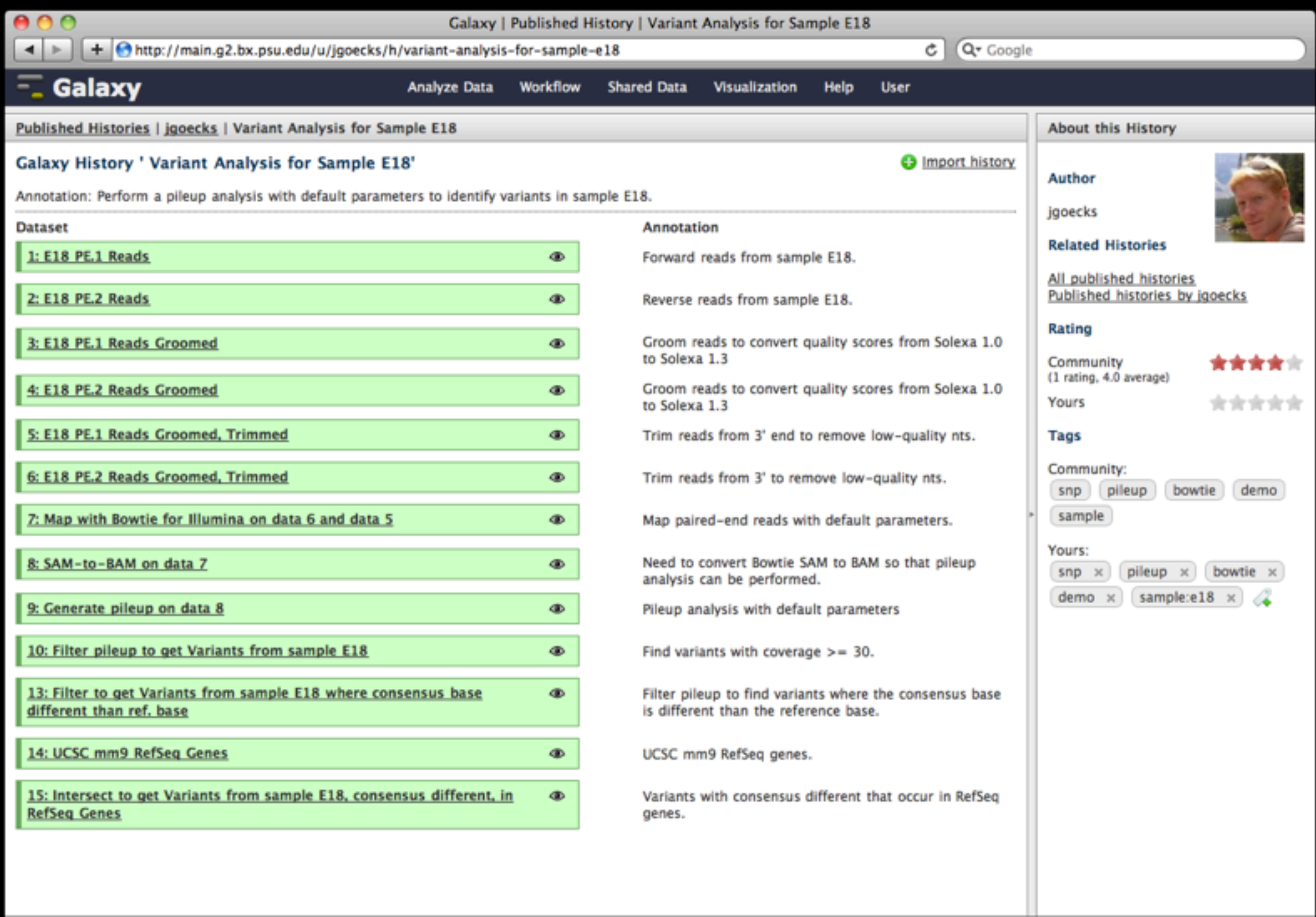
Disables history's link so that it is not accessible and removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

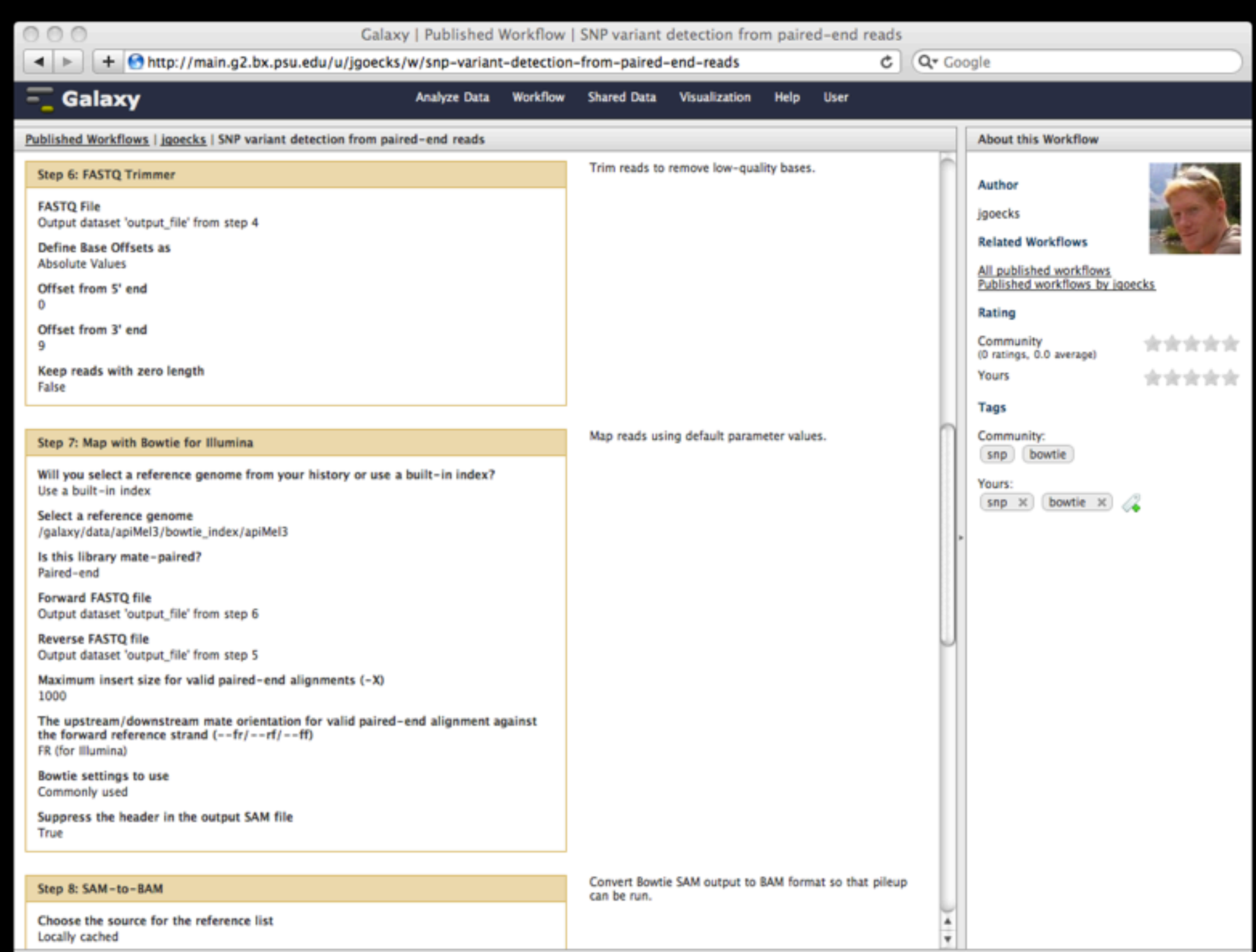
Sharing History with Specific Users

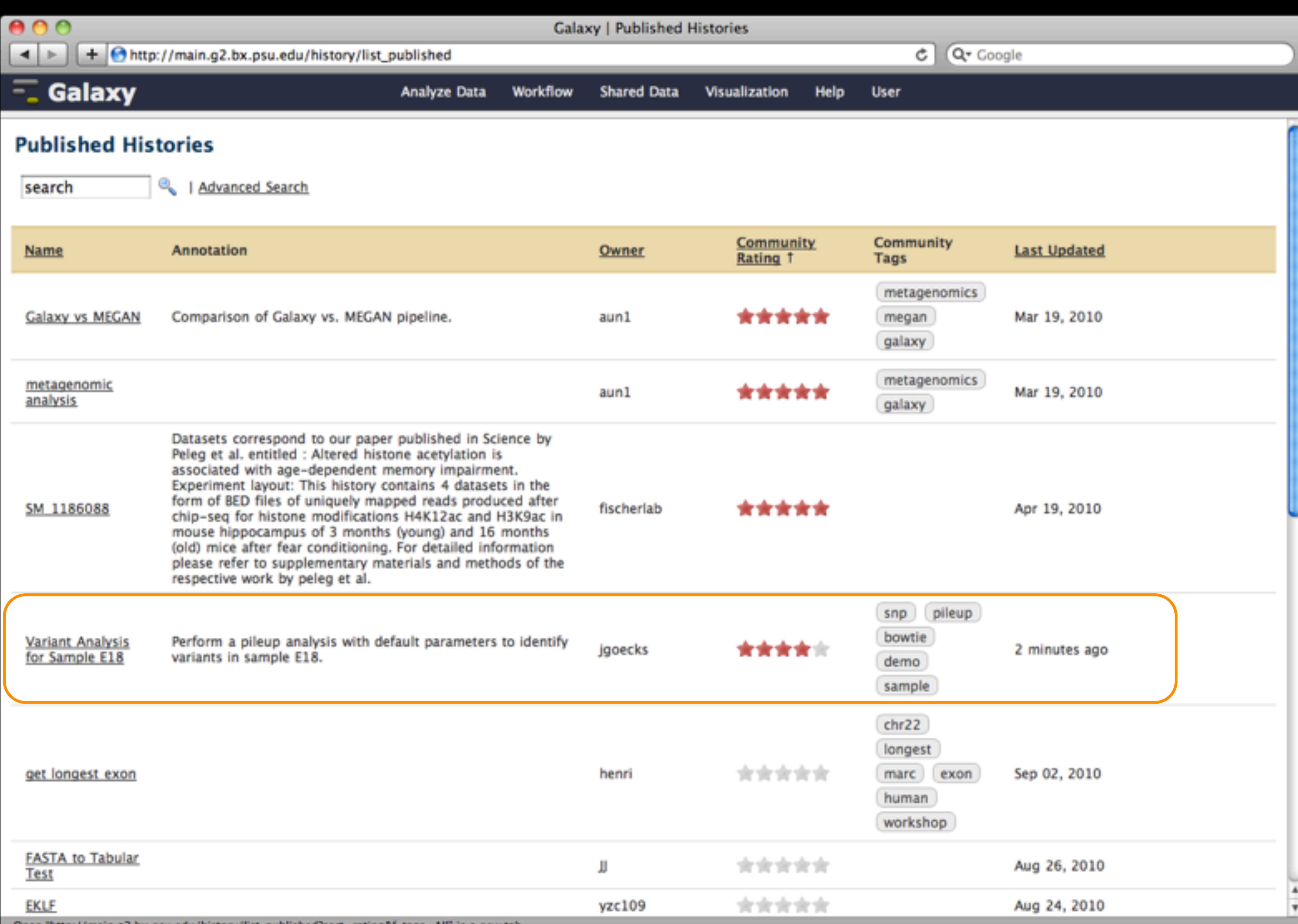
You have not shared this history with any users.

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
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Histories | jgoecks | Variant Analysis for Sample E18
About this History
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.
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
FASTQ File
Output dataset 'output_file' from step 4
Define Base Offsets as
Absolute Values
Offset from 5' end
0
Offset from 3' end
9
Keep reads with zero length
False
Trim reads to remove low-quality bases.
Step 7: Map with Bowtie for Illumina
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 FASTQ file
Output dataset 'output_file' from step 6
Reverse FASTQ file
Output dataset 'output_file' from step 5
Maximum insert size for valid paired-end alignments (-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
Map reads using default parameter values.
Step 8: SAM-to-BAM
Choose the source for the reference list
Locally cached
Convert Bowtie SAM output to BAM format so that pileup can be run.
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


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

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

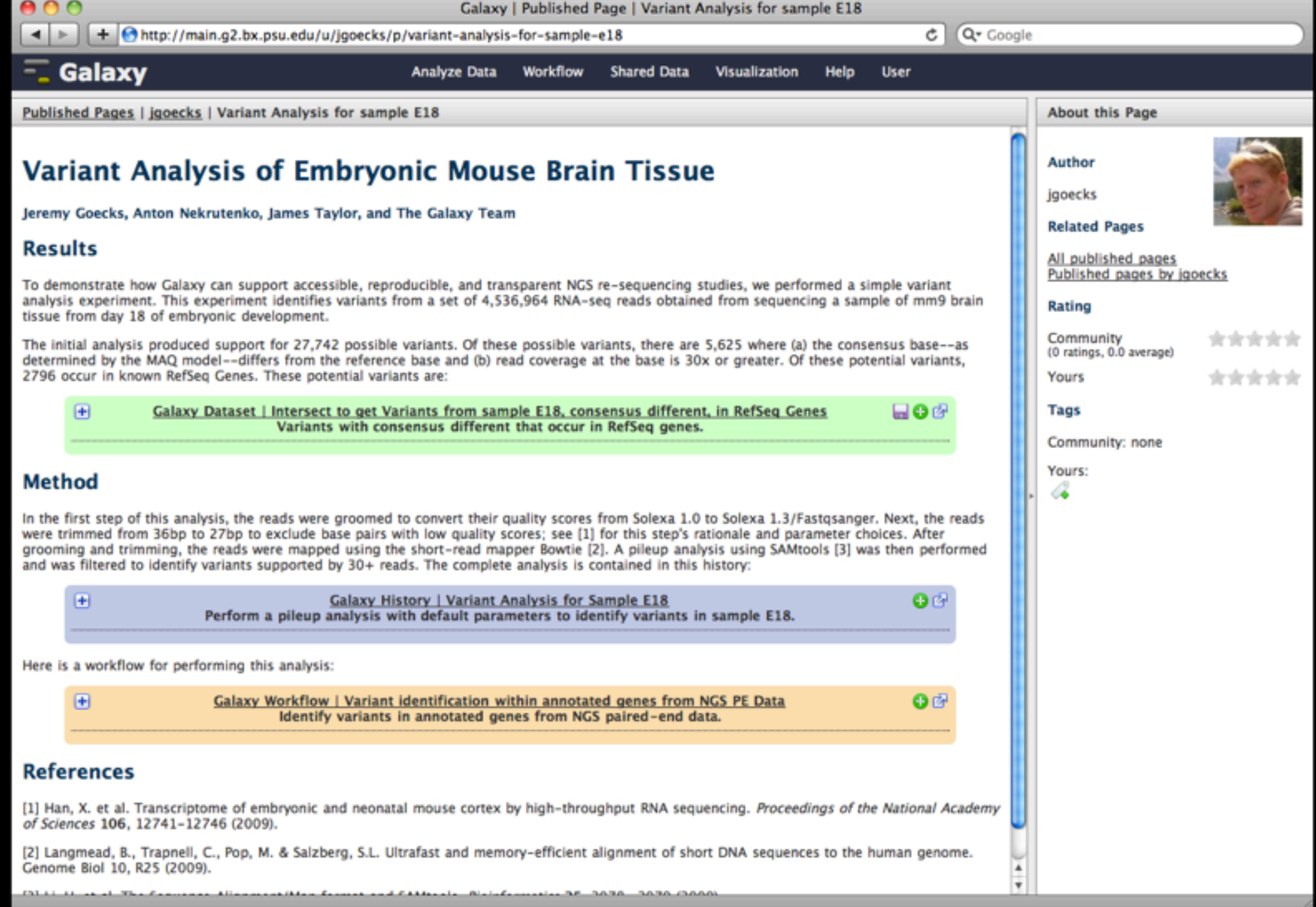
Where you can use and build Galaxy

- public website
- local instance
- on the Cloud
- tool shed/contributing tools

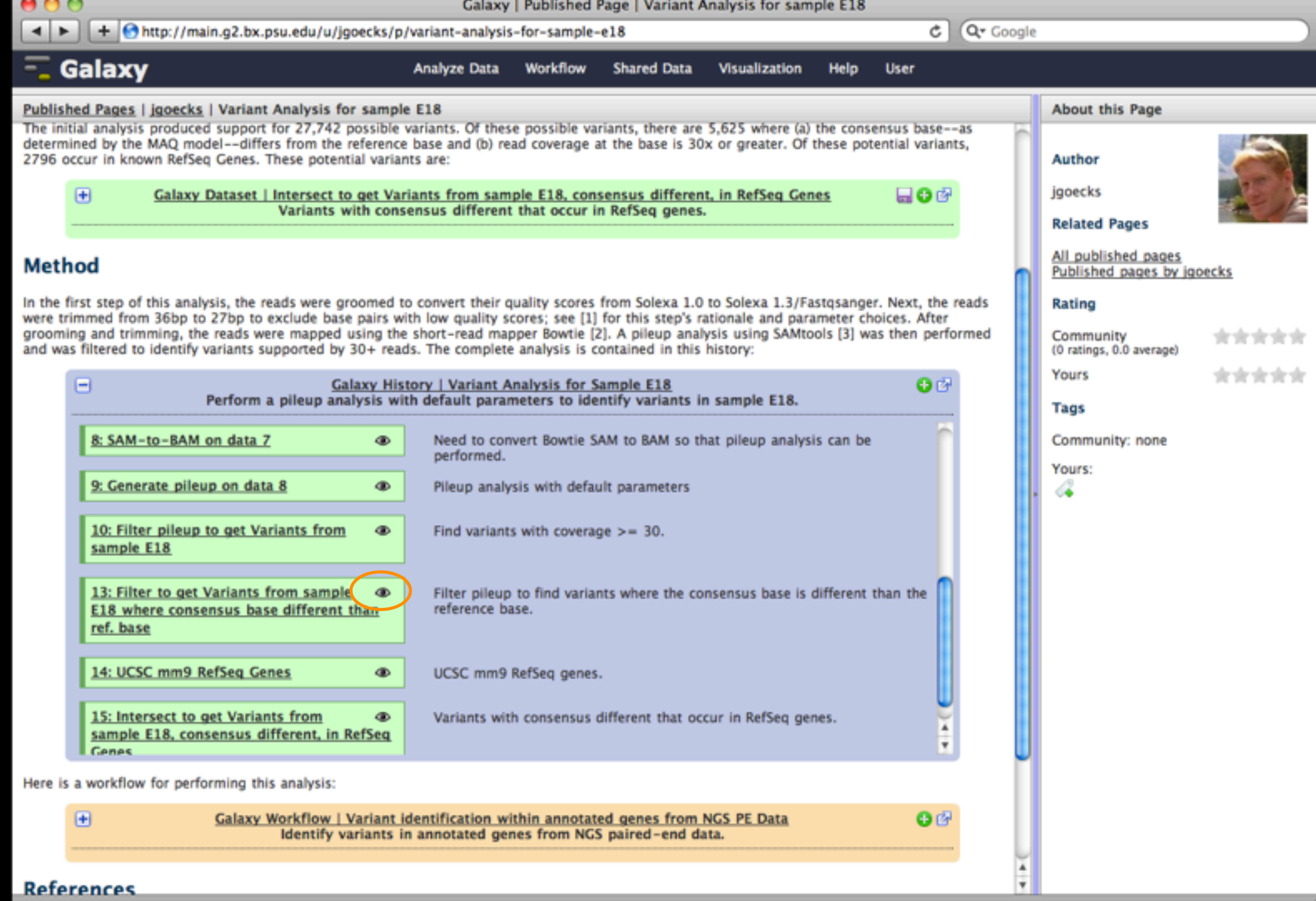
Galaxy Pages

A web-based, interactive medium for presenting all aspects of an analysis: data, methods, and results

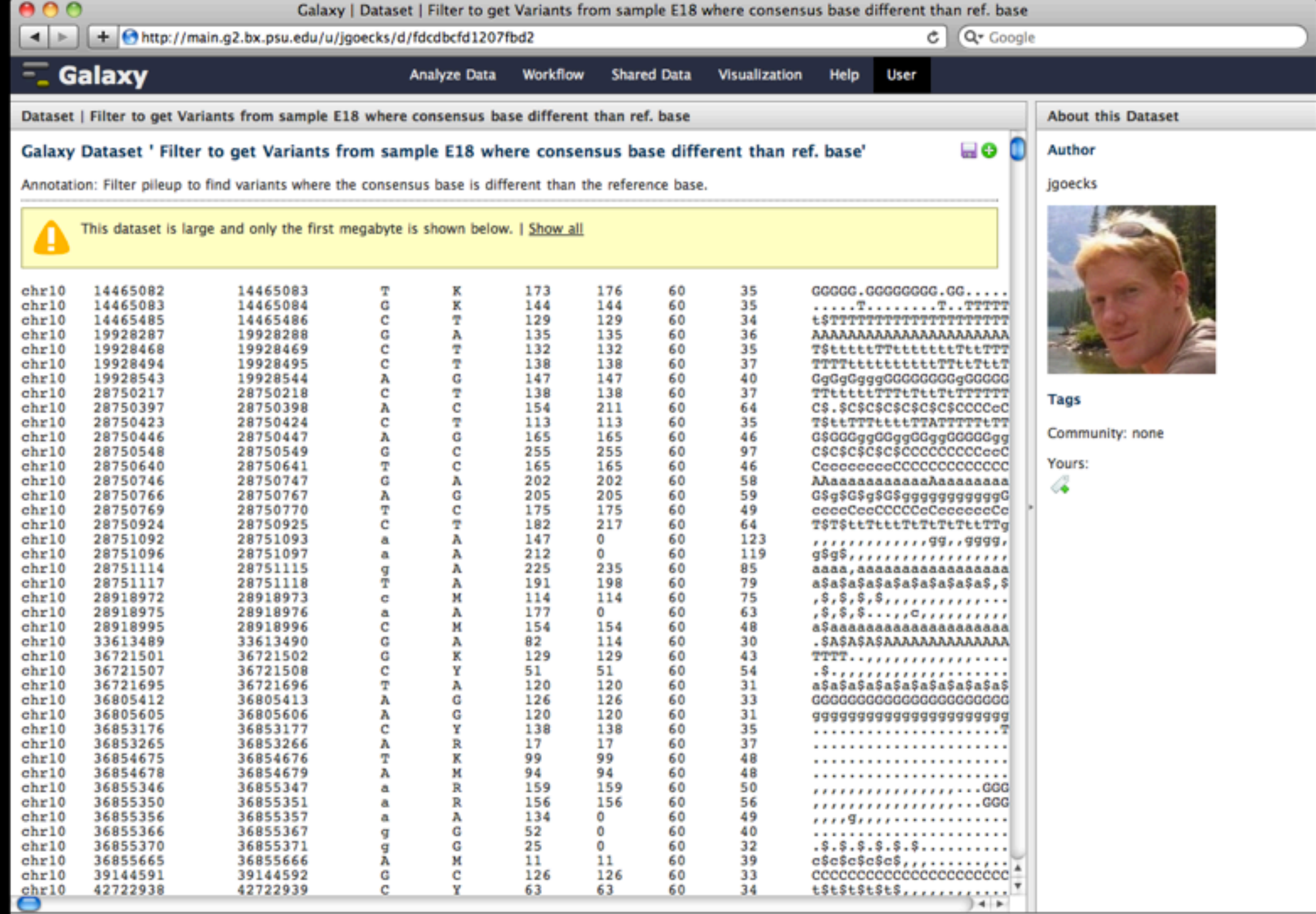
Galaxy Pages



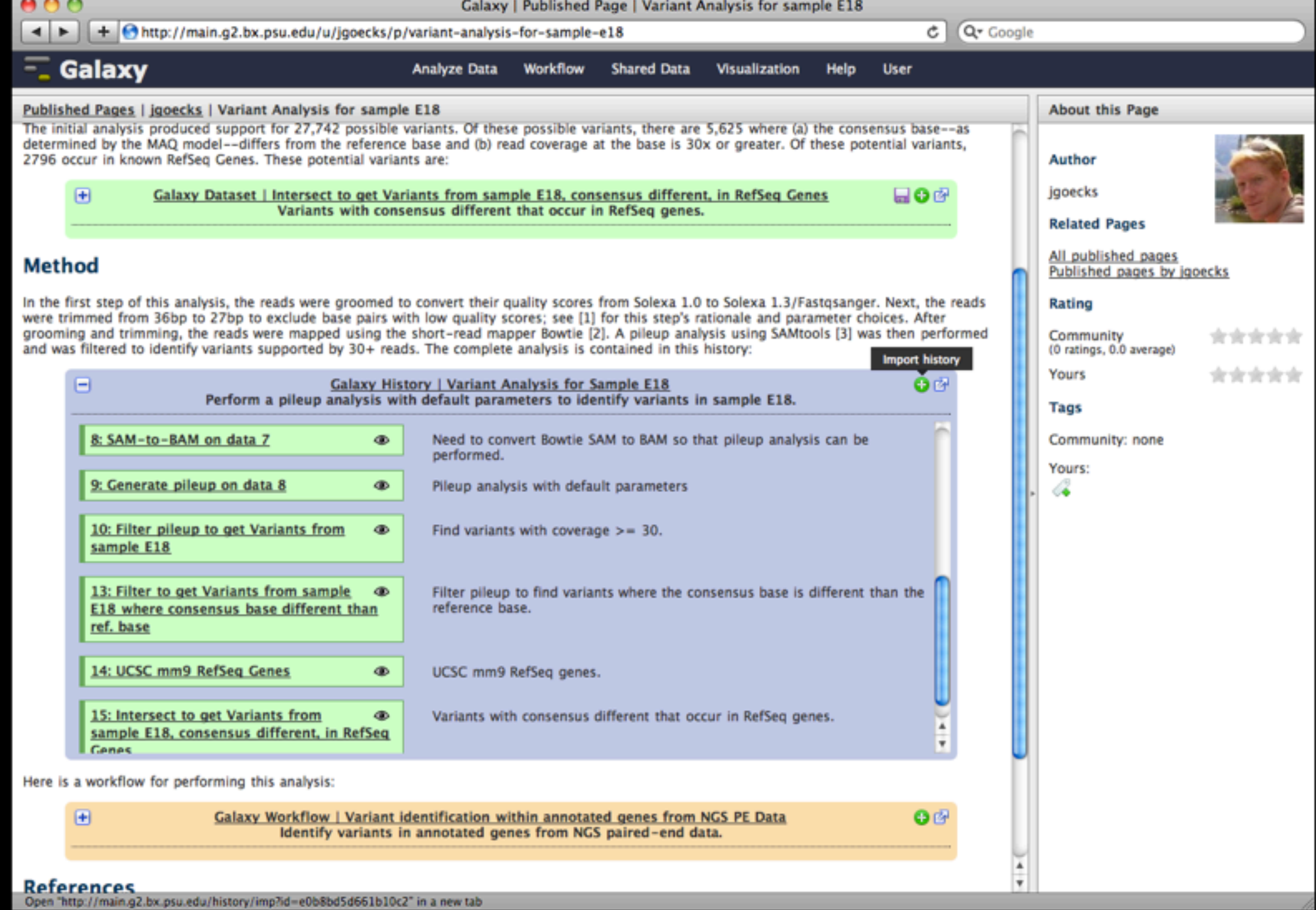
Galaxy Pages



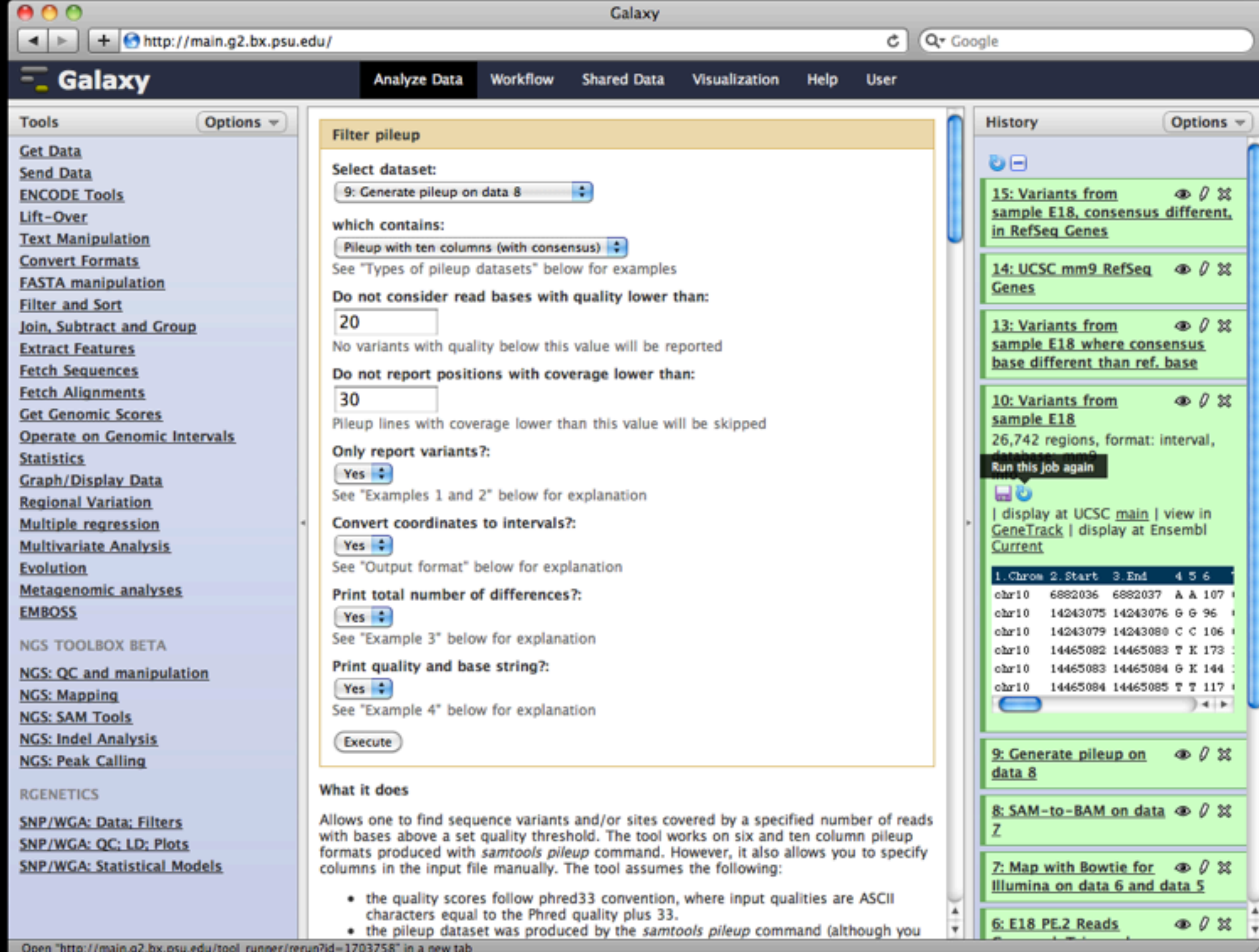
Galaxy Pages



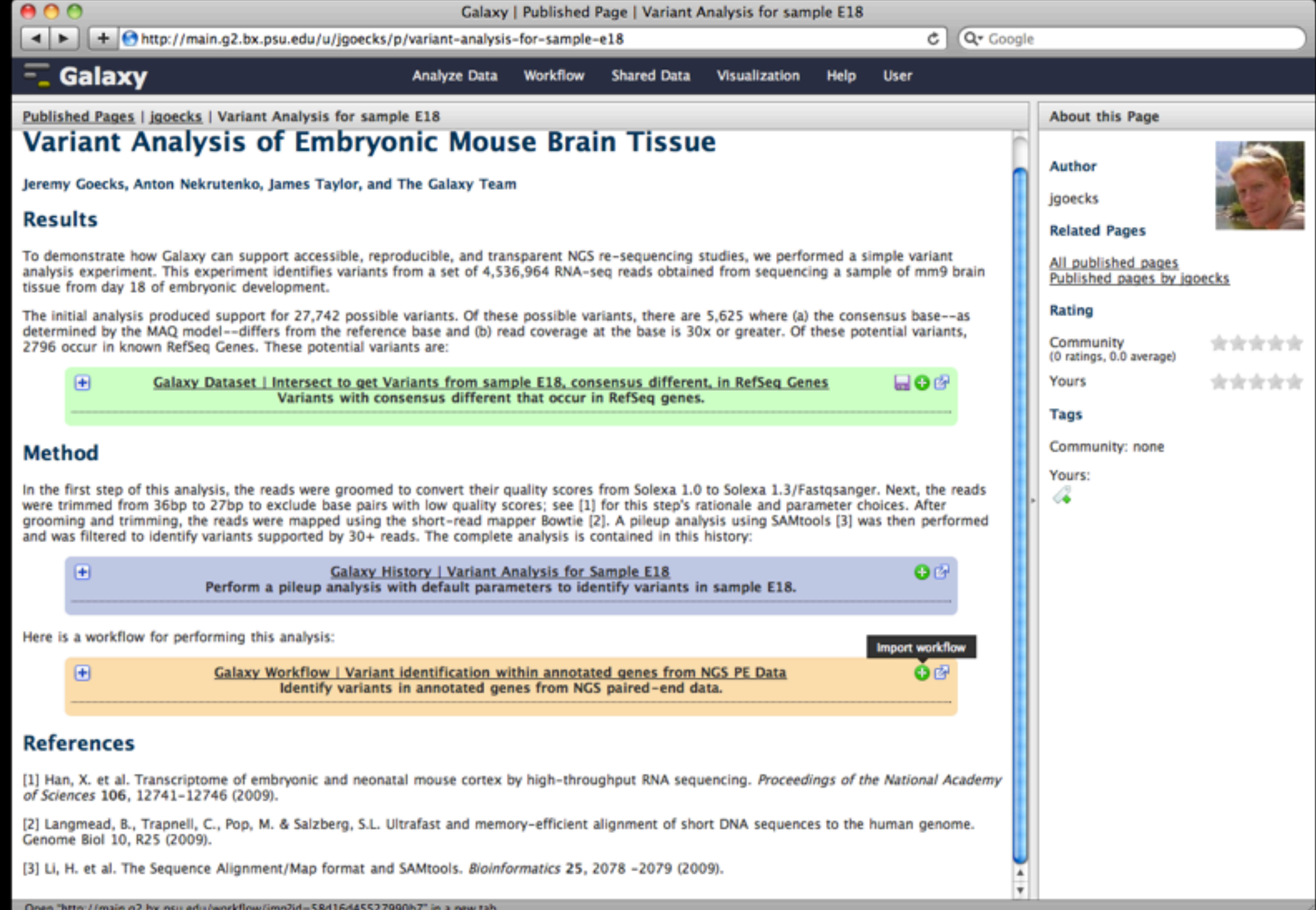
Galaxy Pages



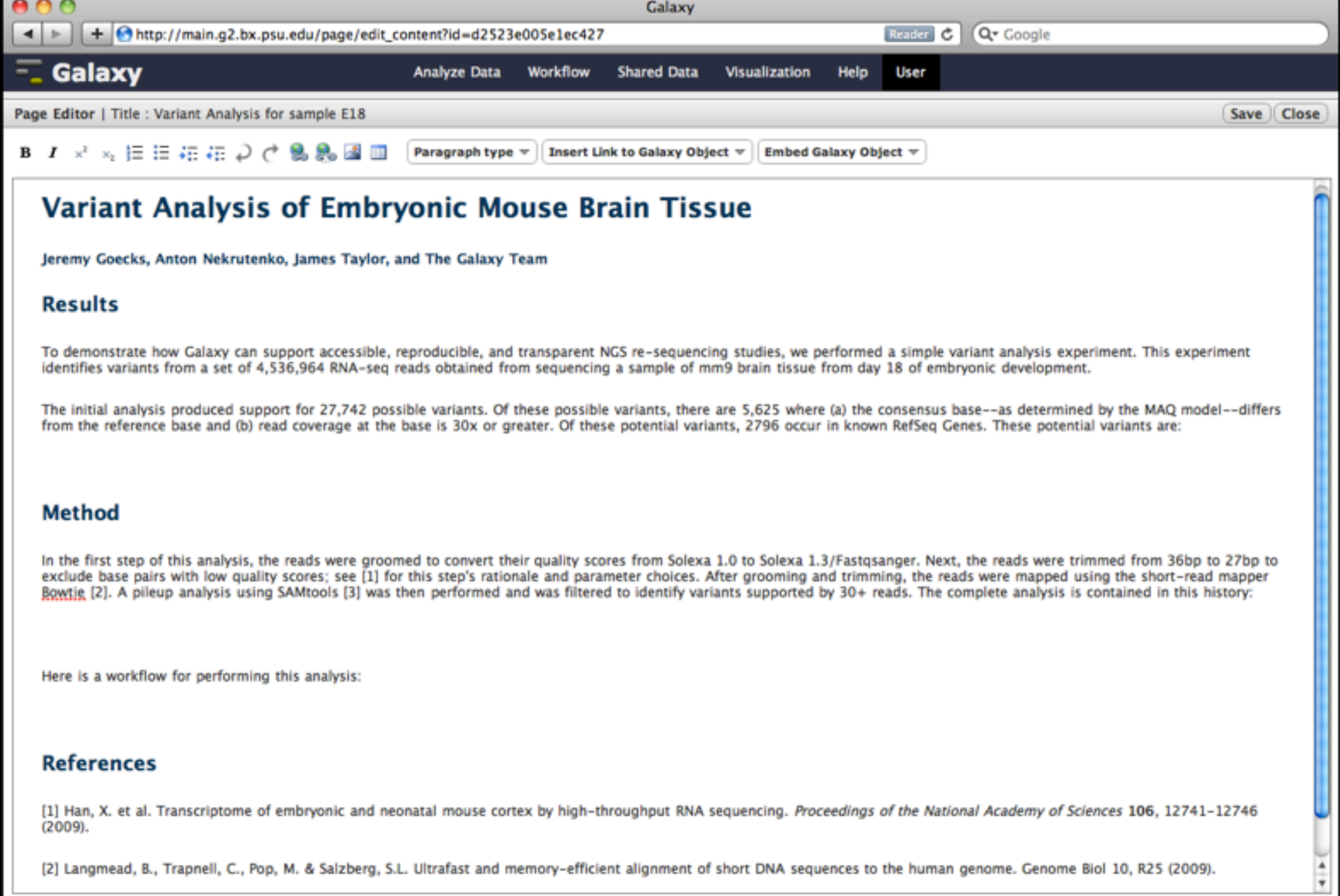
Galaxy Pages



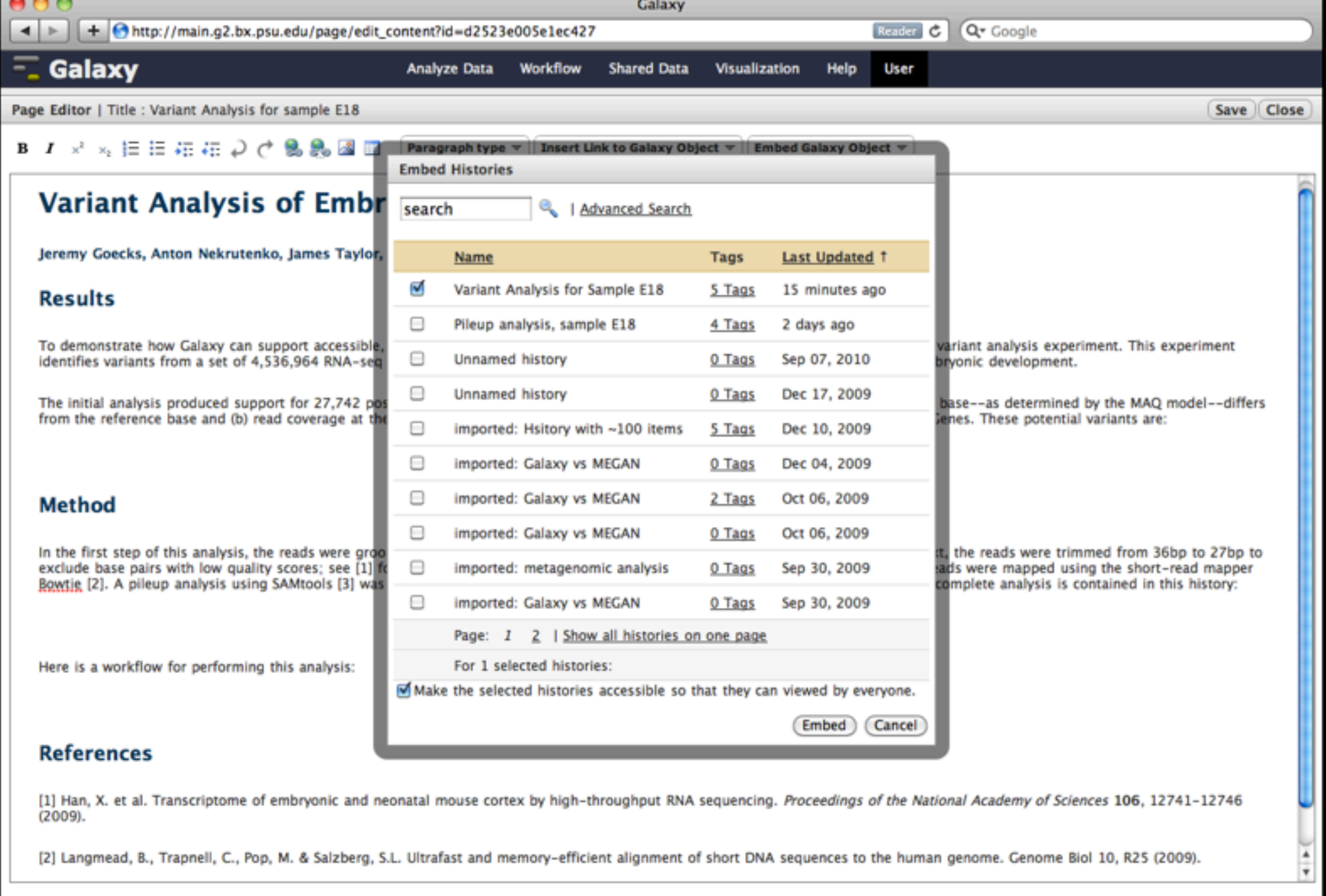
Galaxy Pages



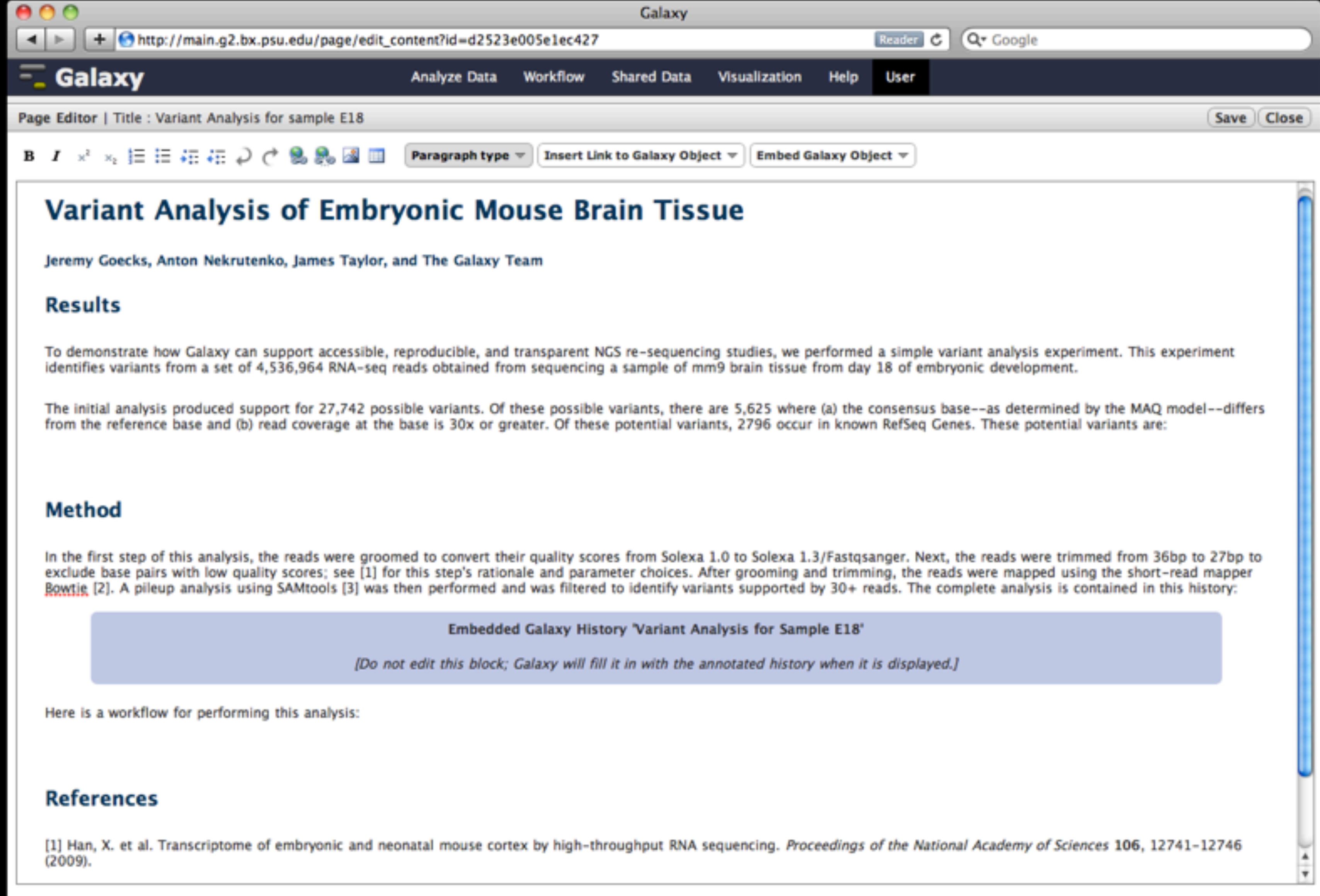
Creating a Page



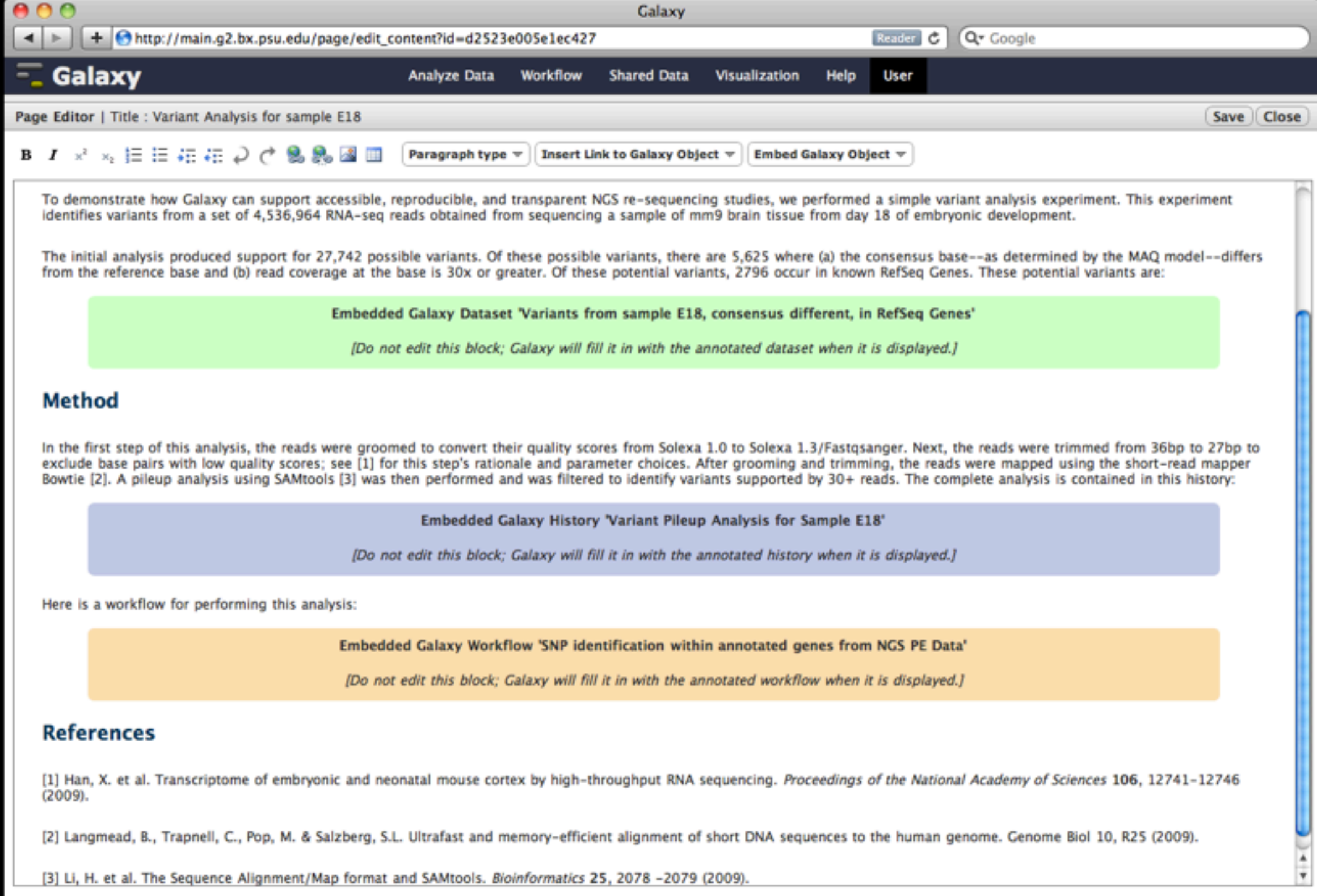
Creating a Page



Creating a Page



Creating a Page



The power of Galaxy publishing

Galaxy's publishing features facilitate access and reproducibility without any extra leg work

One click grants access to the *actual analysis* you performed to generate your original results

- Not just data access: the full pipeline
- Annotate each step
- Anyone can import your work and immediately reproduce or build on it

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the Cloud
- tool shed/contributing tools

Galaxy main site
(http://usegalaxy.org)

Public web site, anybody can use

~500 new users per month, ~100 TB of user data, ~130,000 analysis jobs per month, every month is our busiest month ever...

Will continue to be maintained and enhanced, but with limits and quotas

Centralized solution cannot scale to meet data analysis demands

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the Cloud
- tool shed/contributing tools

Local Galaxy instances (http://getgalaxy.org)

Galaxy is designed for local installation and customization

- Just download and run, completely self-contained
- Easily integrate new tools
- Easy to deploy and manage on nearly any (unix) system
- Run jobs on existing compute clusters

Especially useful for sensitive data

- can secure data and abide by regulations

Scale up on existing resources

Move intensive processing (tool execution) to other hosts

Frees up the application server to serve requests and manage jobs

Utilize existing resources

Supports any scheduler that supports DRMAA (most of them)

Running a Production Server

Use a real database server: PostgreSQL, MySQL

Run on compute cluster resources

External Authentication: LDAP, Kerberos, OpenID

Load balancing; proxy support

https://bitbucket.org/galaxy/galaxy-central/wiki/Config/ProductionServer

Lack IT knowledge or resources?

No problem, just use the Cloud

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

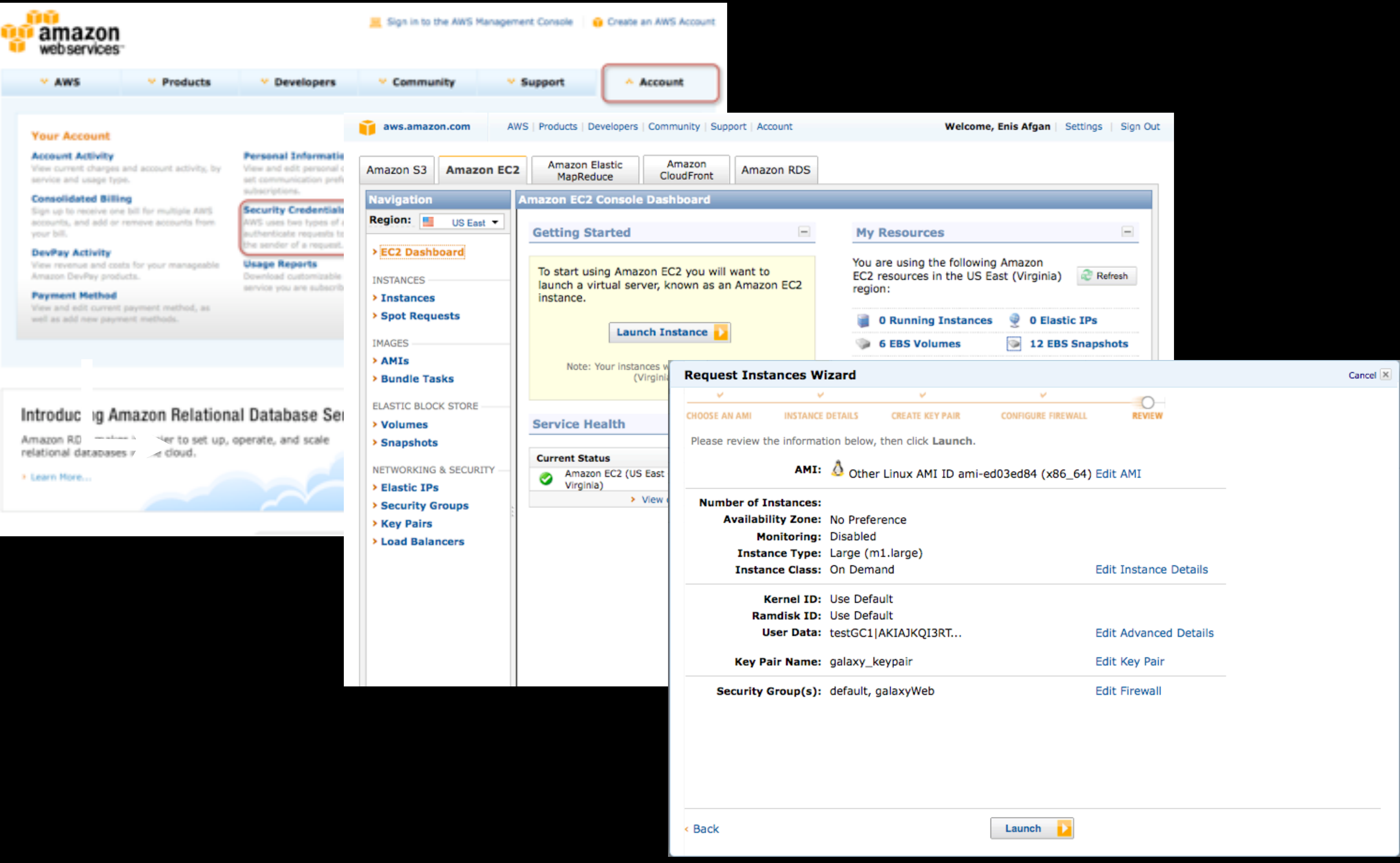
Where you can use and build Galaxy

- public website
- local instance
- on the Cloud
- tool shed/contributing tools

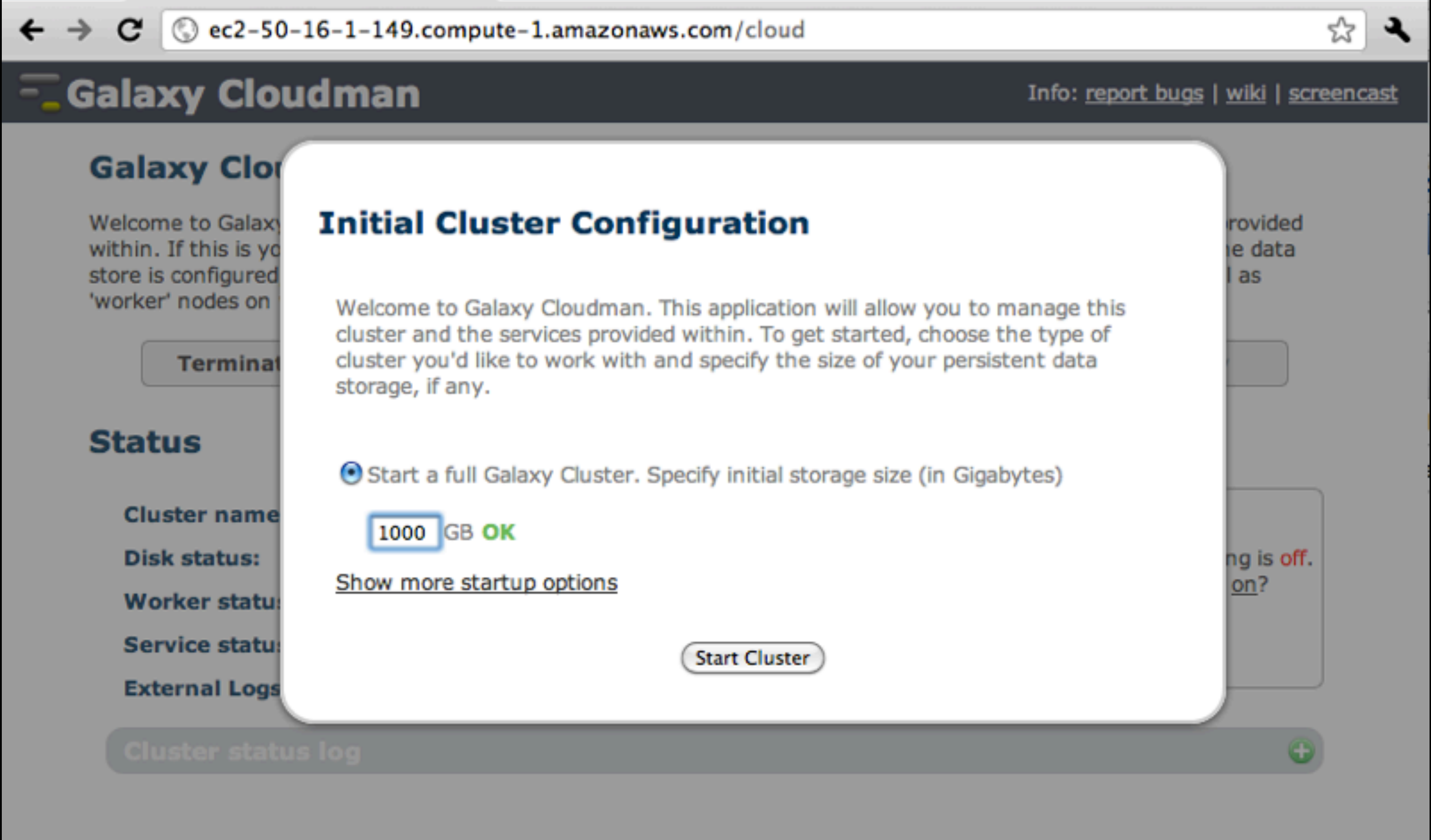
Deploying Galaxy on the AWS Cloud
http://usegalaxy.org/cloud

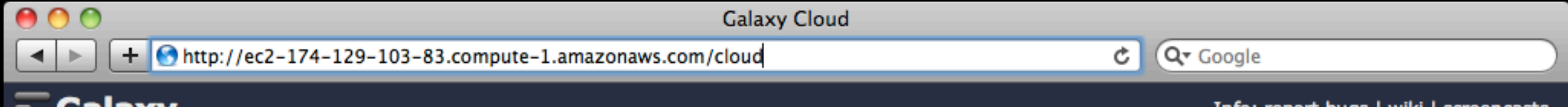
1. Open an AWS account (only once)
2. Use the AWS Management Console to start a master EC2 instance
3. Use the Galaxy CloudMan web interface on the master instance to manage the cluster

2. Start an EC2 Instance



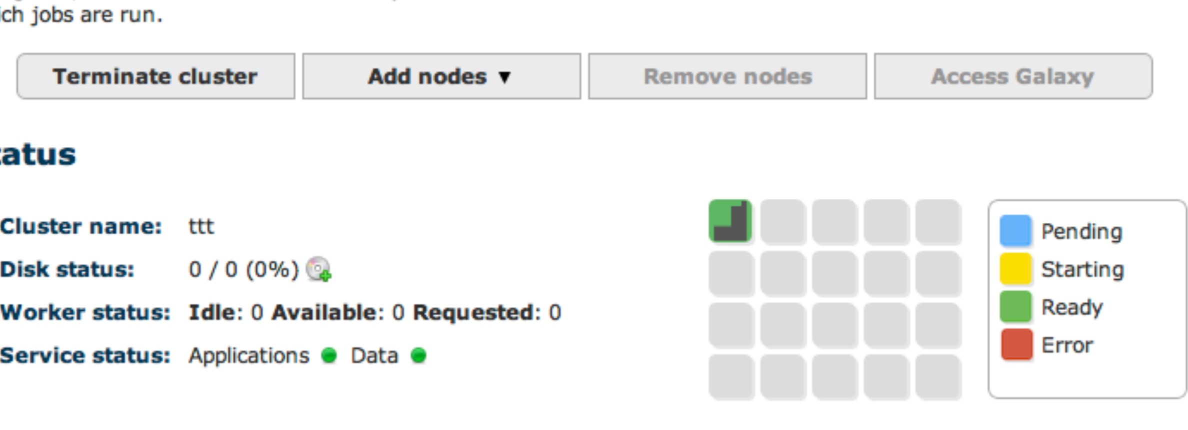
3. Configure Your Cluster



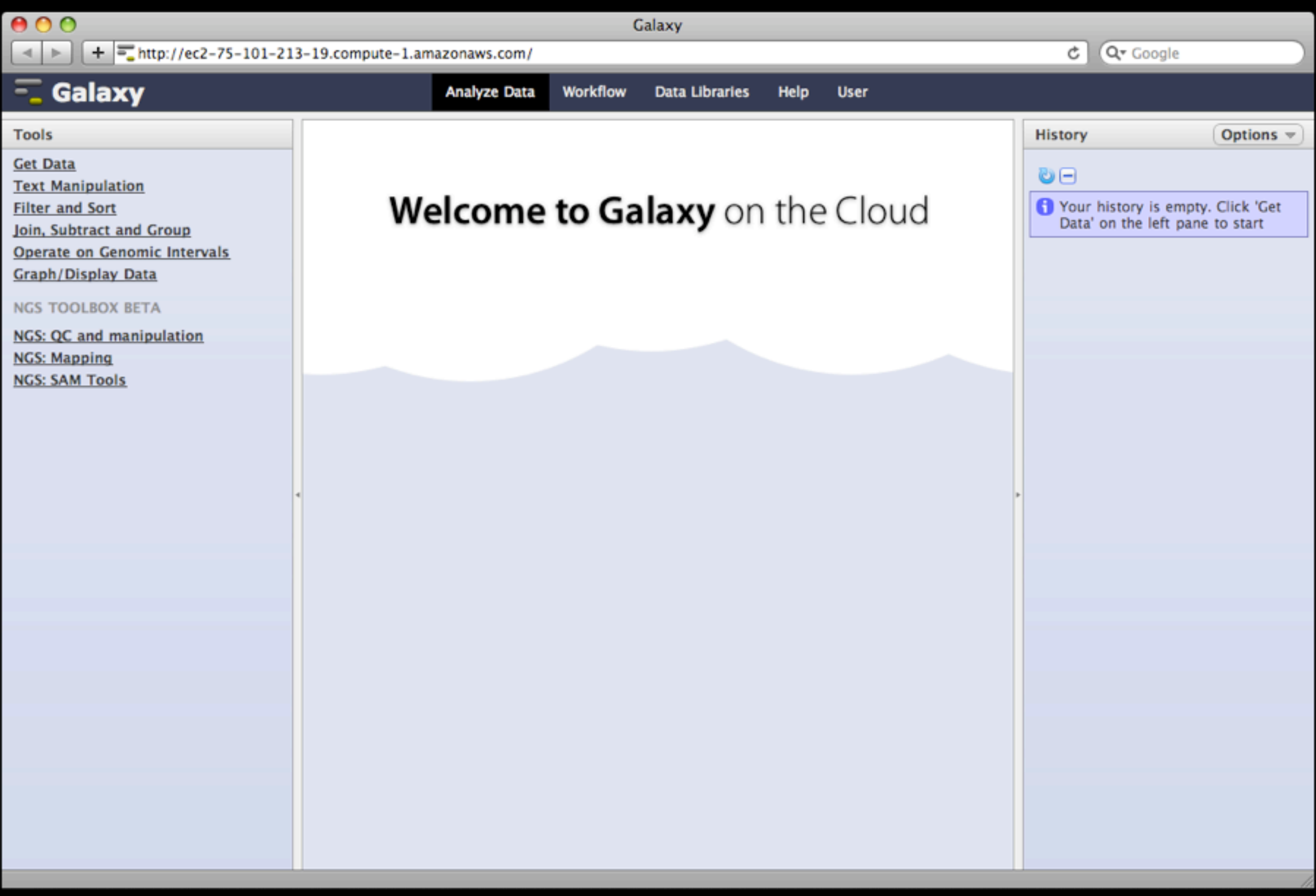


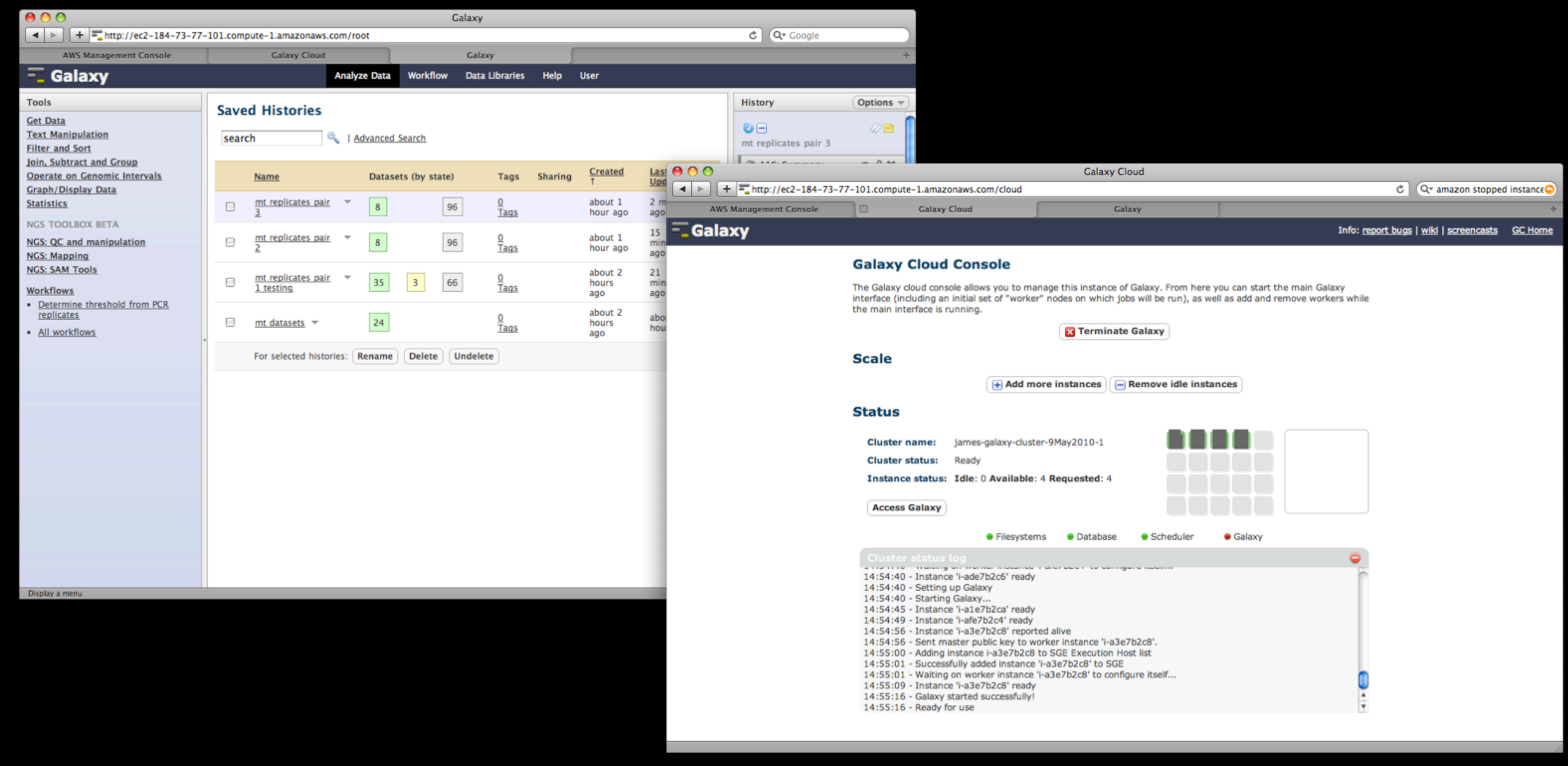
Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be add and remove additional services as well as 'worker' nodes on which jobs are run.



Cluster status log


Galaxy
http://ec2-75-101-213-19.compute-1.amazonaws.com/
Google
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Tools
Get Data
Text Manipulation
Filter and Sort
Join, Subtract and Group
Operate on Genomic Intervals
Graph/Display Data
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
Welcome to Galaxy on the Cloud
History
Options
Your history is empty. Click 'Get Data' on the left pane to start



Can use like any other Galaxy instance, with additional compute nodes acquired and released (*automatically*) in response to usage

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the Cloud
- tool shed/contributing tools

The Problem

You have written a Perl script to analyze genomic data and you want to share it with command-line averse colleagues

The Galaxy Solution

Solution: Integrate the script as a new Tool into your own Galaxy server

Steps:

- Obtain and install Galaxy source code (GetGalaxy.org)
- Write an XML file describing the inputs and outputs and how to execute the script
- Instruct Galaxy to load the tool

Adding your Own

Write or download a command-line executable

Determine number and kind of

- Input and Output Datasets
- Input Parameters

Construct a descriptive tool configuration XML file

- Write a wrapper script, only if required

Cluster

Cluster intervals of: 6: UCSC Main on Human: knownGene

max distance between intervals: 1 (bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

Execute

TIP: If your query does not appear in the pulldown menu -> it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns

---

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

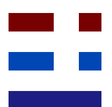
- **Maximum distance** is greatest distance in base pairs allowed between intervals that will be considered "clustered". **Negative** values for distance are allowed, and are useful for clustering intervals that overlap.
- **Minimum intervals per cluster** allow a threshold to be set on the minimum number of intervals to be considered a cluster. Any area with less than this minimum will not be included in the ouput.
- **Merge clusters into single intervals** outputs intervals that span the entire cluster.
- **Find cluster intervals; preserve comments and order** filters out non-cluster intervals while maintaining the original ordering and comments in the file.
- **Find cluster intervals; output grouped by clusters** filters out non-cluster intervals, but outputs the cluster intervals so that they are grouped together. Comments and original ordering in the file are lost.

Example

Query

Find clusters

Merge clusters



cluster.xml

```
1  <tool id="gops_cluster_1" name="Cluster">
2    <description>[[Cluster]] the intervals of a query</description>
3    <command interpreter="python2.4">
4      gops_cluster.py $input1 $output -1 $input1_chromCol,$input1_startC
5                      -d $distance -m $minregions -o $returntype
6    </command>
7    <inputs>
8      <param format="interval" name="input1" type="data">
9        <label>Cluster intervals of</label>
10     </param>
11     <param name="distance" size="5" type="integer" value="1" help="(bp
12       <label>max distance between intervals</label>
13     </param>
14     <param name="minregions" size="5" type="integer" value="2">
15       <label>min number of intervals per cluster</label>
16     </param>
17     <param name="returntype" type="select" label="Return type">
18       <option value="1">Merge clusters into single intervals</option>
19       <option value="2">Find cluster intervals; preserve comments and
20       <option value="3">Find cluster intervals; output grouped by clus
21       <option value="4">Find the smallest interval in each cluster</op
22       <option value="5">Find the largest interval in each cluster</opt
23     </param>
24    </inputs>
25    <help>
26
27 .. class:: infomark
28
29 **TIP:** If your query does not appear in the pulldown menu -> it is n
30
31 -----
32
33 **Screencasts!**
34
35 See Galaxy Interval Operation Screencasts_ (right click to open this l
36
37 .. _Screencasts: http://www.bx.psu.edu/cgi-bin/trac.cgi/wiki/GopsDesc
38
39 -----
40
41 **Syntax**
42
43 - **Maximum distance** is greatest distance in base pairs allowed betw
44 - **Minimum intervals per cluster** allow a threshold to be set on the
45 - **Merge clusters into single intervals** outputs intervals that span
46 - **Find cluster intervals; preserve comments and order** filters out
47 - **Find cluster intervals; output grouped by clusters** filters out n
```

Line: 87 Column: 8 XML Soft Tabs: 2

Adding your Own Display Application

Define An XML configuration which describes how and where to present the data to the External Web Application

- Static
- Dynamic - display options can be loaded from a file

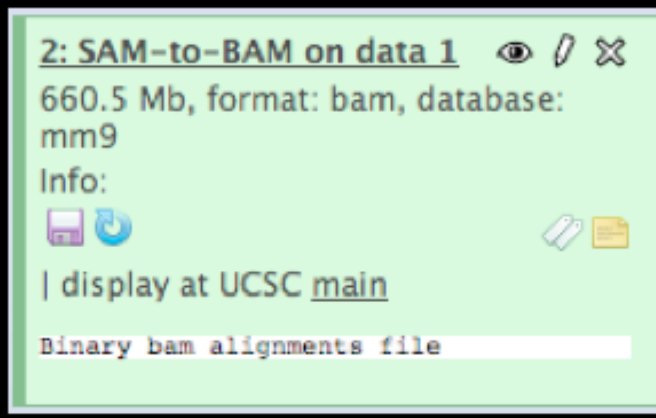
Inform Galaxy about the new display by adding to the appropriate datatype in datatypes_conf.xml

https://bitbucket.org/galaxy/galaxy-central/wiki/ExternalDisplayApplications/Tutorial

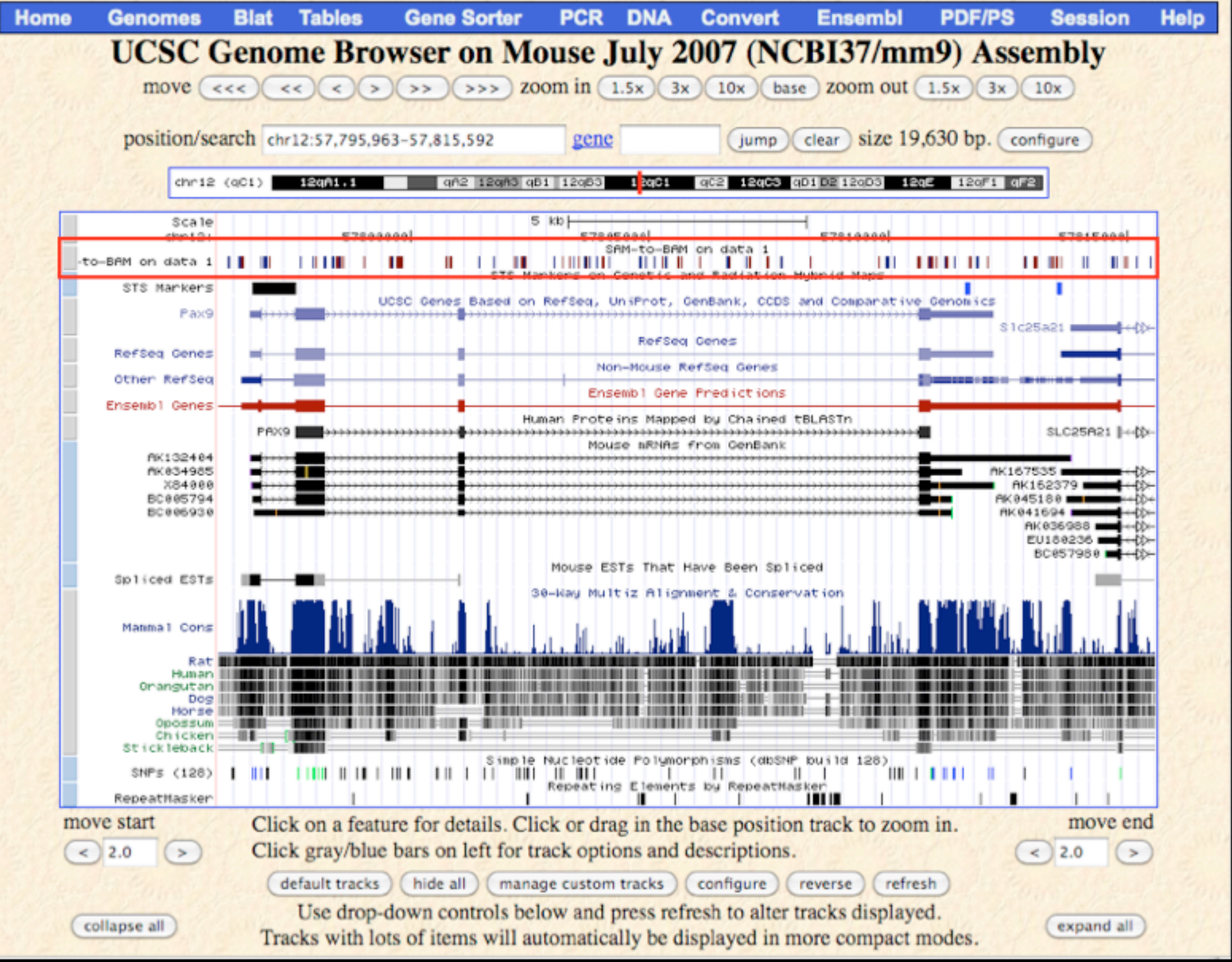
Static External Display Application

```
<display id="ucsc_bam" version="1.0.0" name="display at UCSC">
    <link id="main" name="main">
        <url>http://genome.ucsc.edu/cgi-bin/hgTracks?db=${qp($bam_file.dbkey)}&amp;hgt.customText=${qp($track.url)}</url>
        <param type="data" name="bam_file" url="galaxy.bam" strip_https="True" />
        <param type="data" name="bai_file" url="galaxy.bam.bai" metadata="bam_index" strip_https="True" />
        <param type="template" name="track" viewable="True" strip_https="True">
            track type=bam name="${bam_file.name}" bigDataUrl=${bam_file.url} db=${bam_file.dbkey}
        </param>
    </link>
</display>
```

```
<datatype extension="bam" type="galaxy.datatypes.binary:Bam"
    mimetype="application/octet-stream" display_in_upload="true">
    <display file="ucsc/bam.xml" />
</datatype>
```



BAM at UCSC



Dynamic External Display Application

```
<display id="ucsc_bam" version="1.0.0" name="display at UCSC">
    <!-- Load links from file: one line to one link -->
    <dynamic_links from_file="tool-data/shared/ucsc/ucsc_build_sites.txt" skip_startswith="#" id="0" name="0">

        <!-- Define parameters by column from file, allow splitting on builds -->
        <dynamic_param name="site_id" value="0"/>
        <dynamic_param name="ucsc_link" value="1"/>
        <dynamic_param name="builds" value="2" split="True" separator="," />

        <!-- Filter out some of the links based upon matching site_id to a Galaxy application configuration parameter and b
        <filter>${site_id in $APP.config.ucsc_display_sites}</filter>
        <filter>${dataset.dbkey in $builds}</filter>

        <!-- We define url and params as normal, but values defined in dynamic_param are available by specified name -->
        <url>${ucsc_link}db=${qp($bam_file.dbkey)}&amp;hgt.customText=${qp($track.url)}</url>
        <param type="data" name="bam_file" url="galaxy_${DATASET_HASH}.bam" strip_https="True" />
        <param type="data" name="bai_file" url="galaxy_${DATASET_HASH}.bam.bai" metadata="bam_index" strip_https="True" />
        <param type="template" name="track" viewable="True" strip_https="True">
            track type=bam name="${bam_file.name}" bigDataUrl=${bam_file.url} db=${bam_file.dbkey}
        </param>

    </dynamic_links>
</display>
```

```
#Harvested from http://genome.ucsc.edu/cgi-bin/das/dsn
main    http://genome.ucsc.edu/cgi-bin/hgTracks?    anoCar1,ce6,ce4,ce2,rn3,l
#Harvested from http://archaea.ucsc.edu/cgi-bin/das/dsn
archaea http://archaea.ucsc.edu/cgi-bin/hgTracks?   therSibi1,symbTher_IAM148
#Harvested from http://main.genome-browser.bx.psu.edu/cgi-bin/das/dsn
bx-main http://main.genome-browser.bx.psu.edu/cgi-bin/hgTracks? oviAri1,eriEu
```

You added a tool, now what?

Share it with the community!

Galaxy Tool Shed

- Upload and Download contributed tools
- Rate and provide comments and feedback

Galaxies on private clouds

Galaxies on public clouds

http://usegalaxy.org

Galaxy Tool Shed

http://usegalaxy.org/community

...

private Tool Sheds

1

2

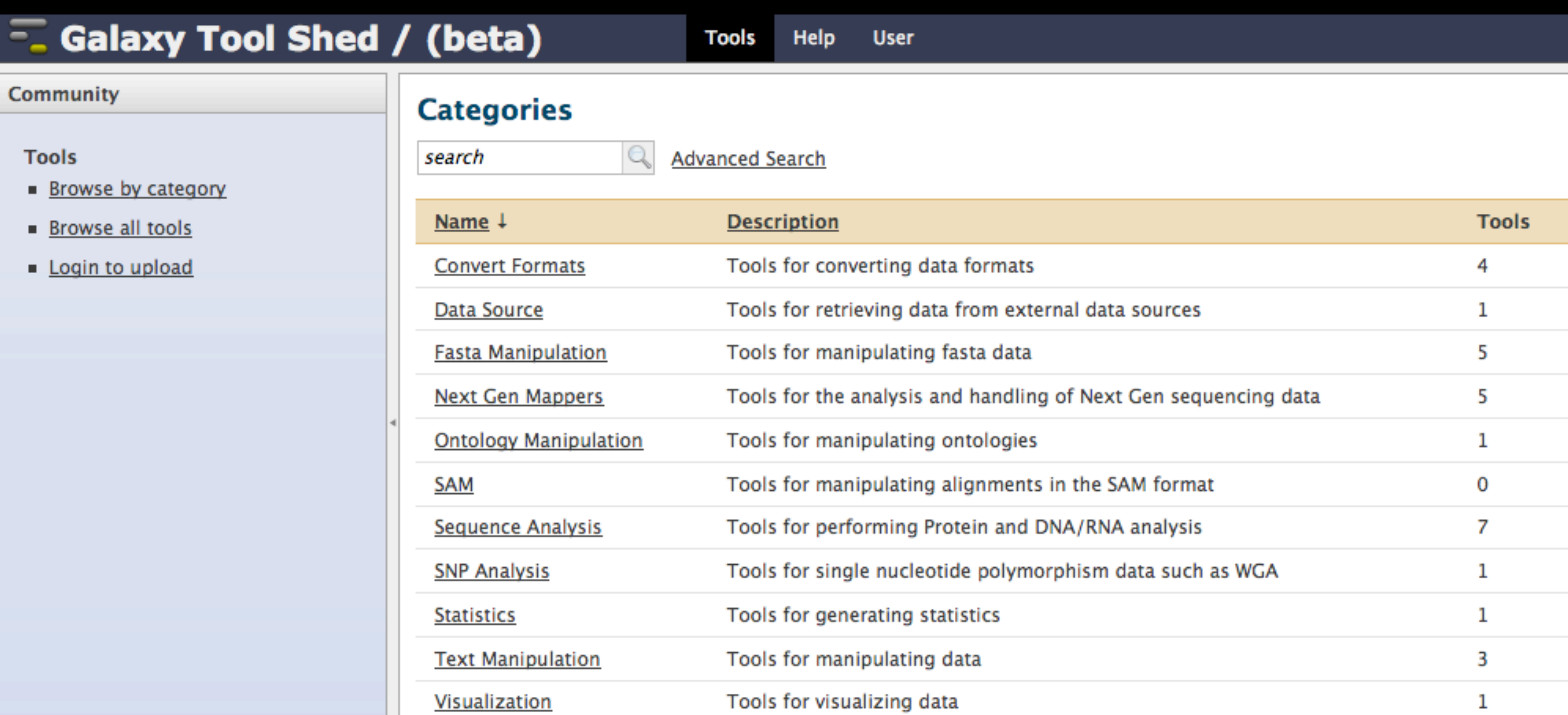
3

...

∞

private Galaxy installations

Get and Contribute Tools



Name ↓	Description	Tools
Convert Formats	Tools for converting data formats	4
Data Source	Tools for retrieving data from external data sources	1
Fasta Manipulation	Tools for manipulating fasta data	5
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	5
Ontology Manipulation	Tools for manipulating ontologies	1
SAM	Tools for manipulating alignments in the SAM format	0
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	7
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	1
Statistics	Tools for generating statistics	1
Text Manipulation	Tools for manipulating data	3
Visualization	Tools for visualizing data	1

http://usegalaxy.org/community

Some future challenges

- Capturing and automatically deploying tool dependencies, automatic tool acquisition in Galaxy instances
- Better interfaces for highly parallel analysis (e.g. running the same workflow across 192 individuals)
- Various workflow engine improvements, partial data streaming, combined experimental/computational workflows

Try it now: http://usegalaxy.org

Develop and deploy: http://getgalaxy.org

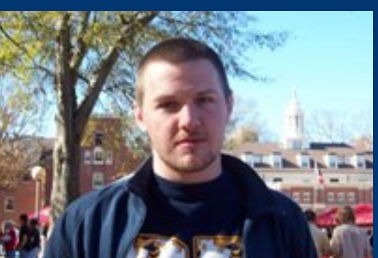
http://galaxyproject.org

Come do cool stuff, contact us at:

http://wiki.g2.bx.psu.edu/News/Galaxy is Hiring

Opportunities for collaboration, positions for postdocs, researchers, software engineers

Enis Afgan



Dannon Baker

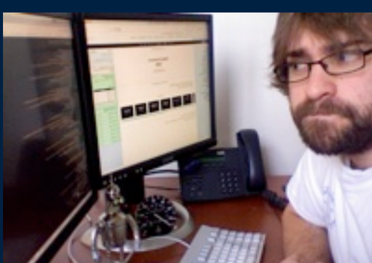
Dave Clements

Jeremy Goecks

Kanwei Li

James Taylor

Dan Blankenberg



Nate Coraor

Jennifer Jackson

Greg von Kuster

Guru Ananda

Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health