Progress and Challenges in Developing a Web-based Platform for Computational Biomedical Research

Jeremy Goecks
Depts. of Biology and Math & Computer Science
Emory University

EMORY

Enis Afgan

Dannon Baker

Dave Clements

Jeremy Goecks

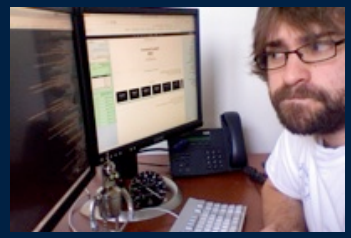
Kanwei Li

James Taylor

PENNSTATE

1855

Dan Blankenberg



Nate Coraor

Jennifer Jackson

Greg von Kuster

Kelly Vincent

Anton Nekrutenko

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

Overview

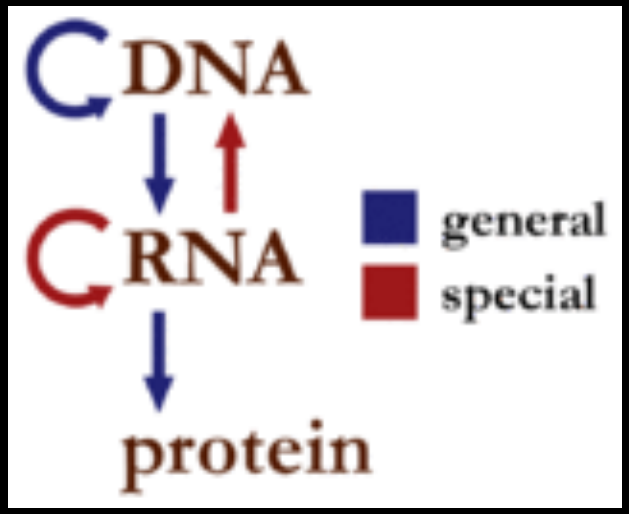
Genomics

Galaxy

- accessible, reproducible, and transparent science
- on the cloud
- visual analytics

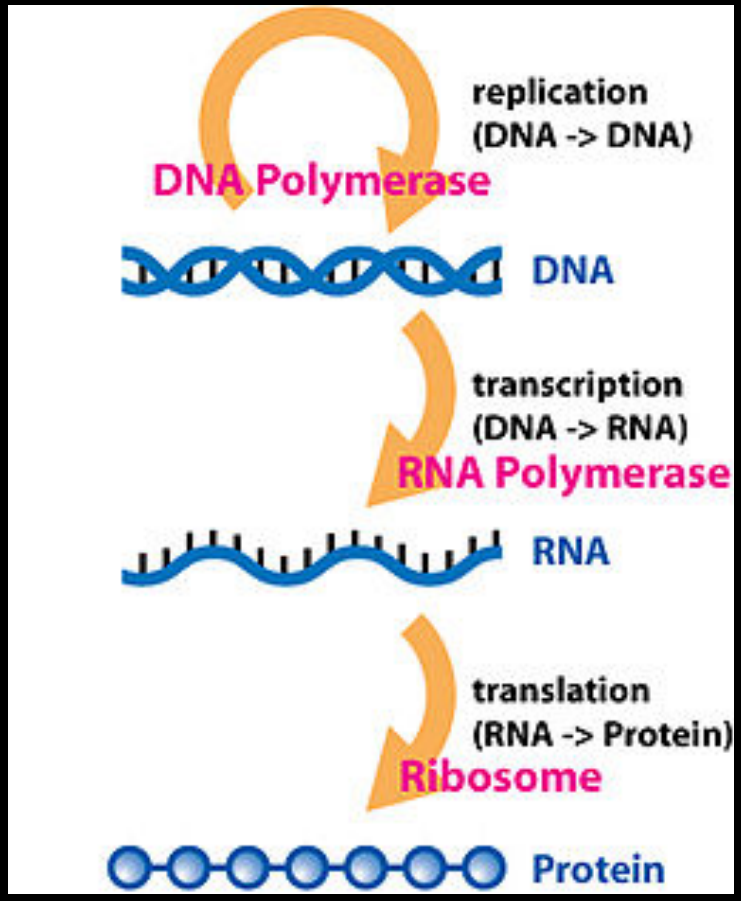
Reflections on Galaxy

Molecular Biology Primer



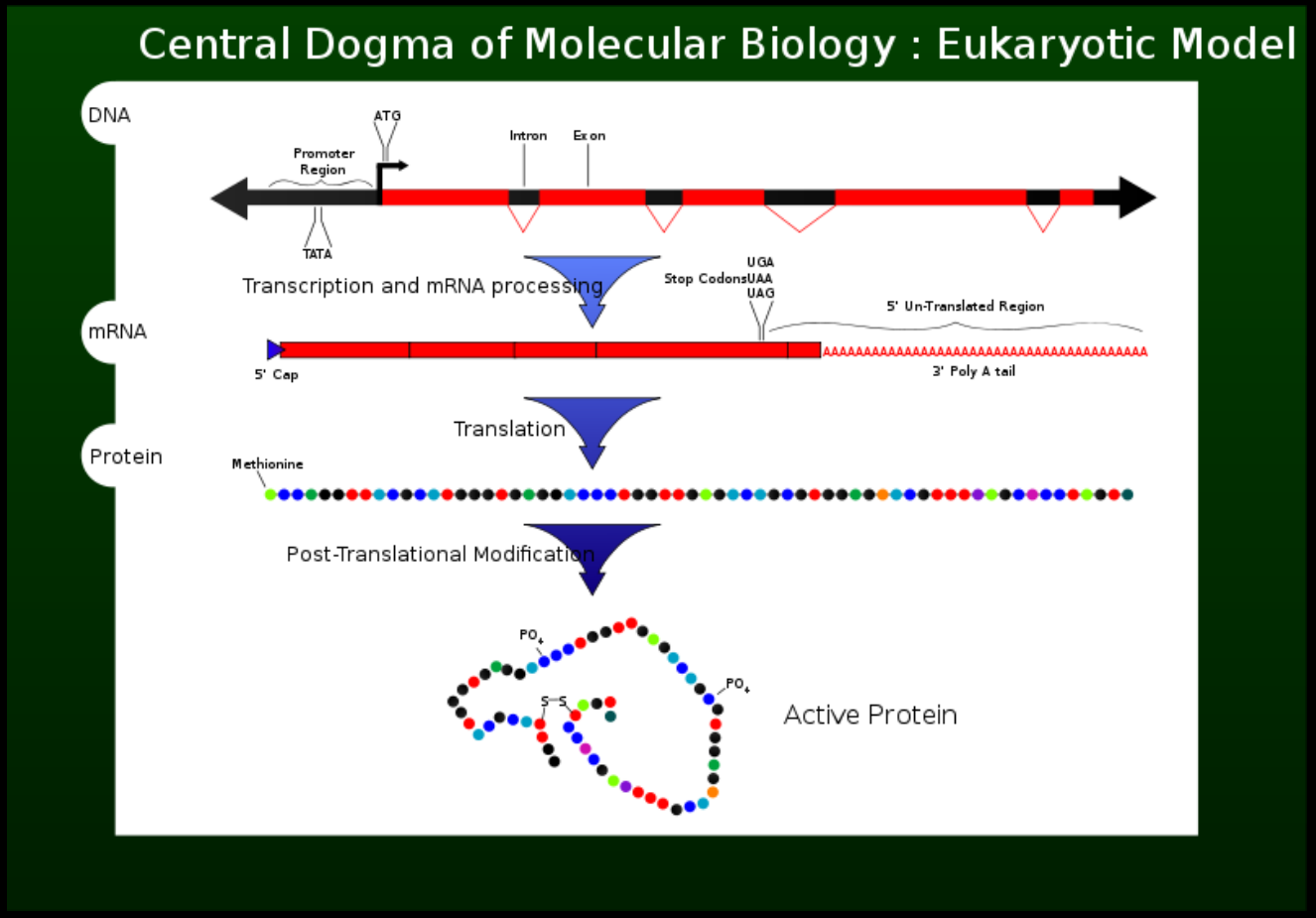
http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

Molecular Biology Primer



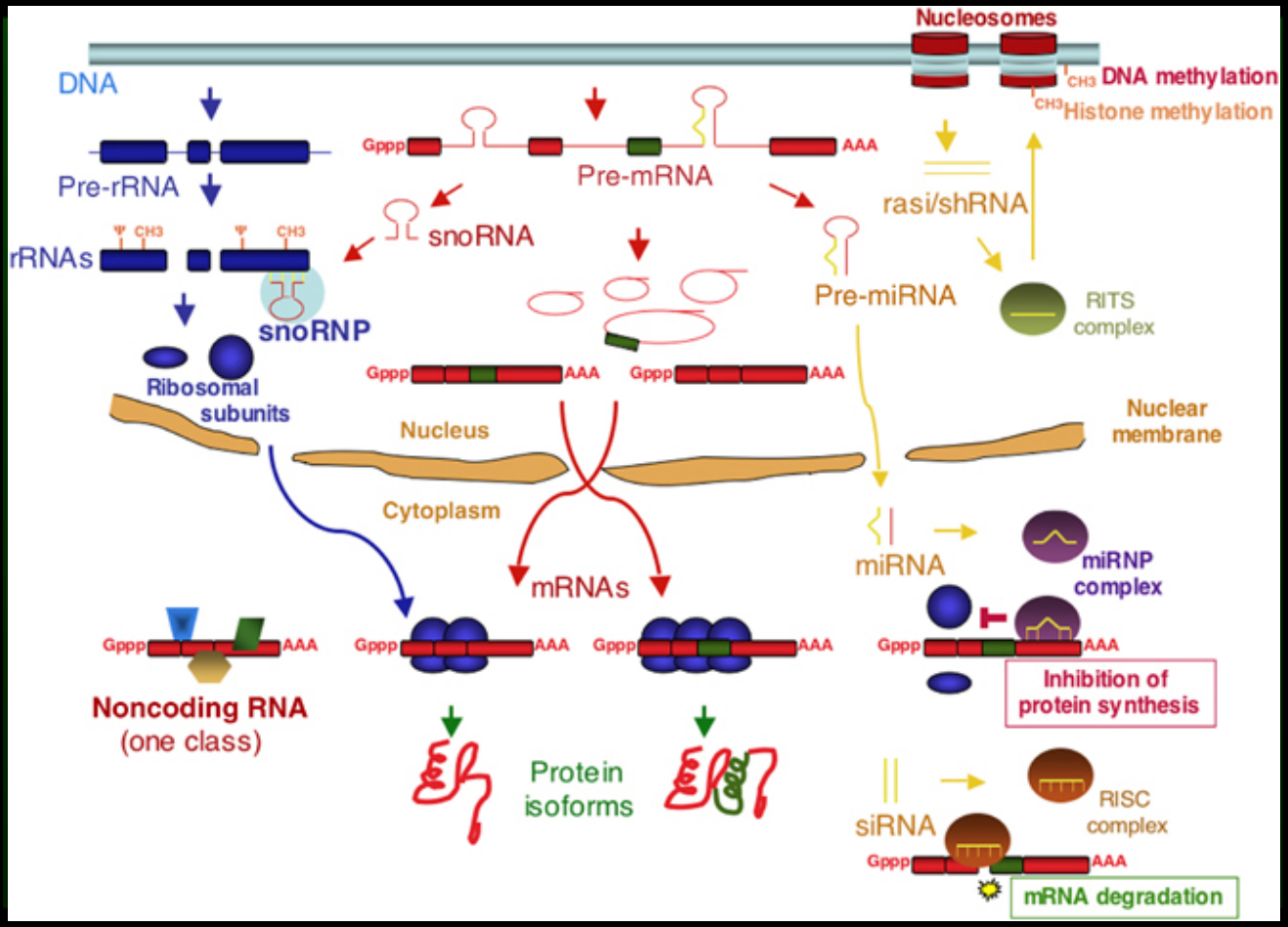
http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

Molecular Biology Primer



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

Molecular Biology Primer



Goals of Genomics

Identify and annotate all functional genomic elements

- genes, promoters, enhancers, silencers, epigenetic modifications

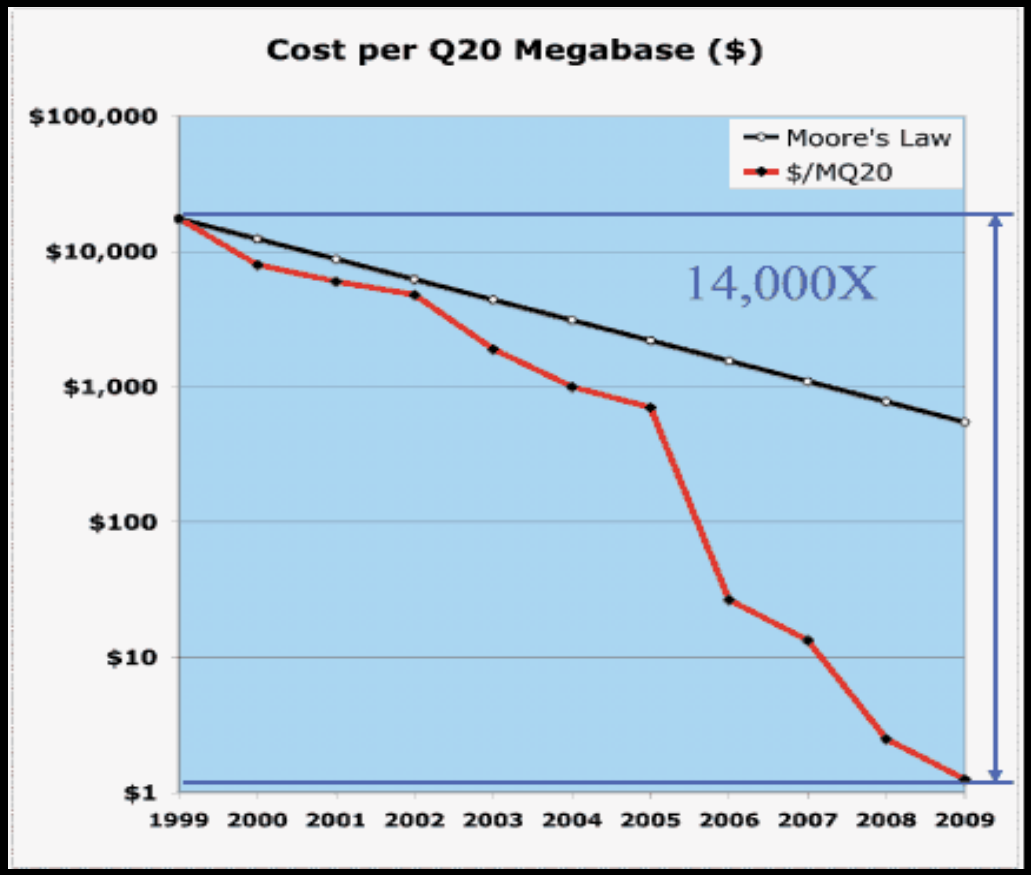
Understand genomic regulation

- interactions, networks, feedback systems

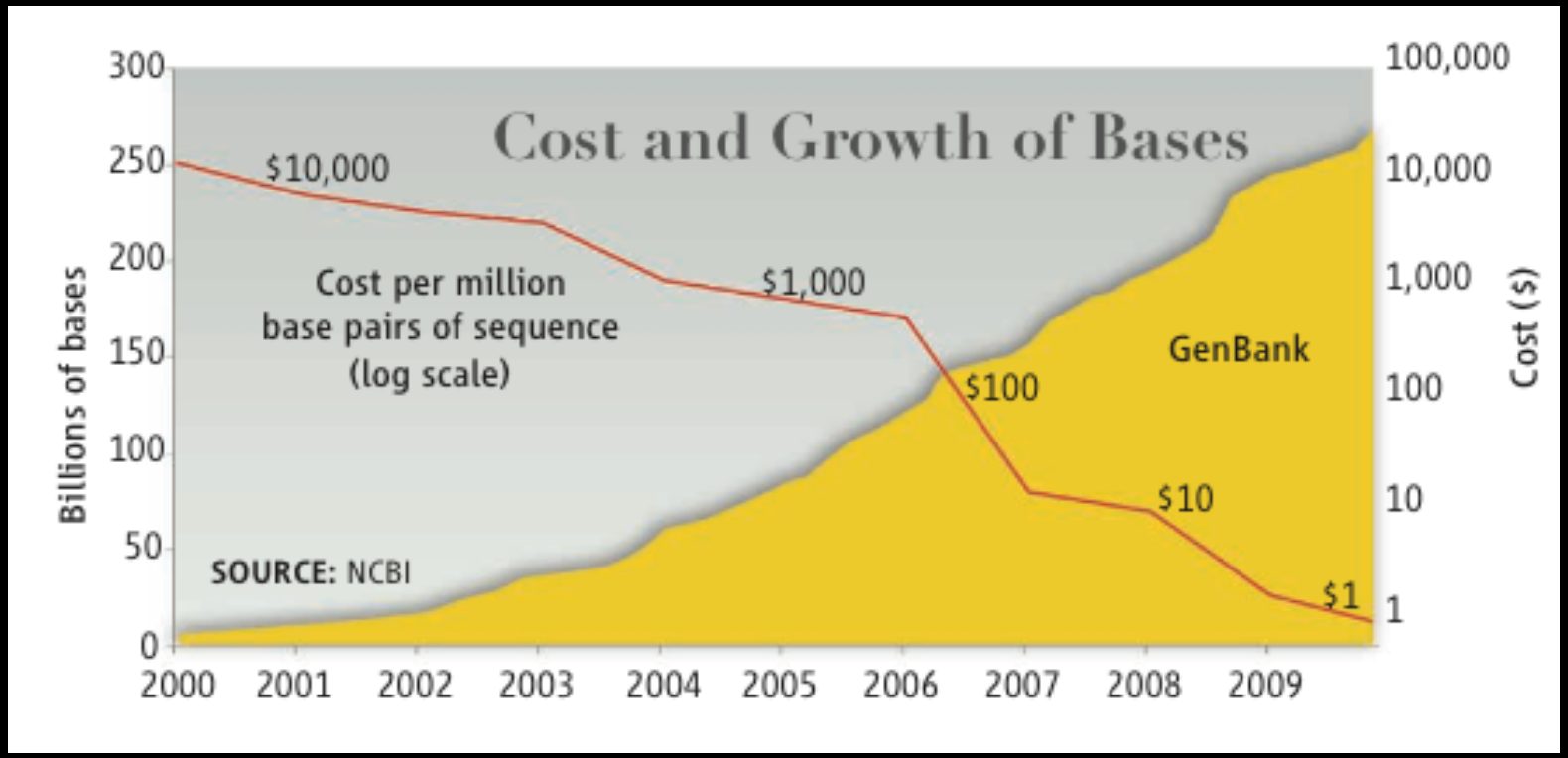
Apply knowledge of genome to address biomedical challenges

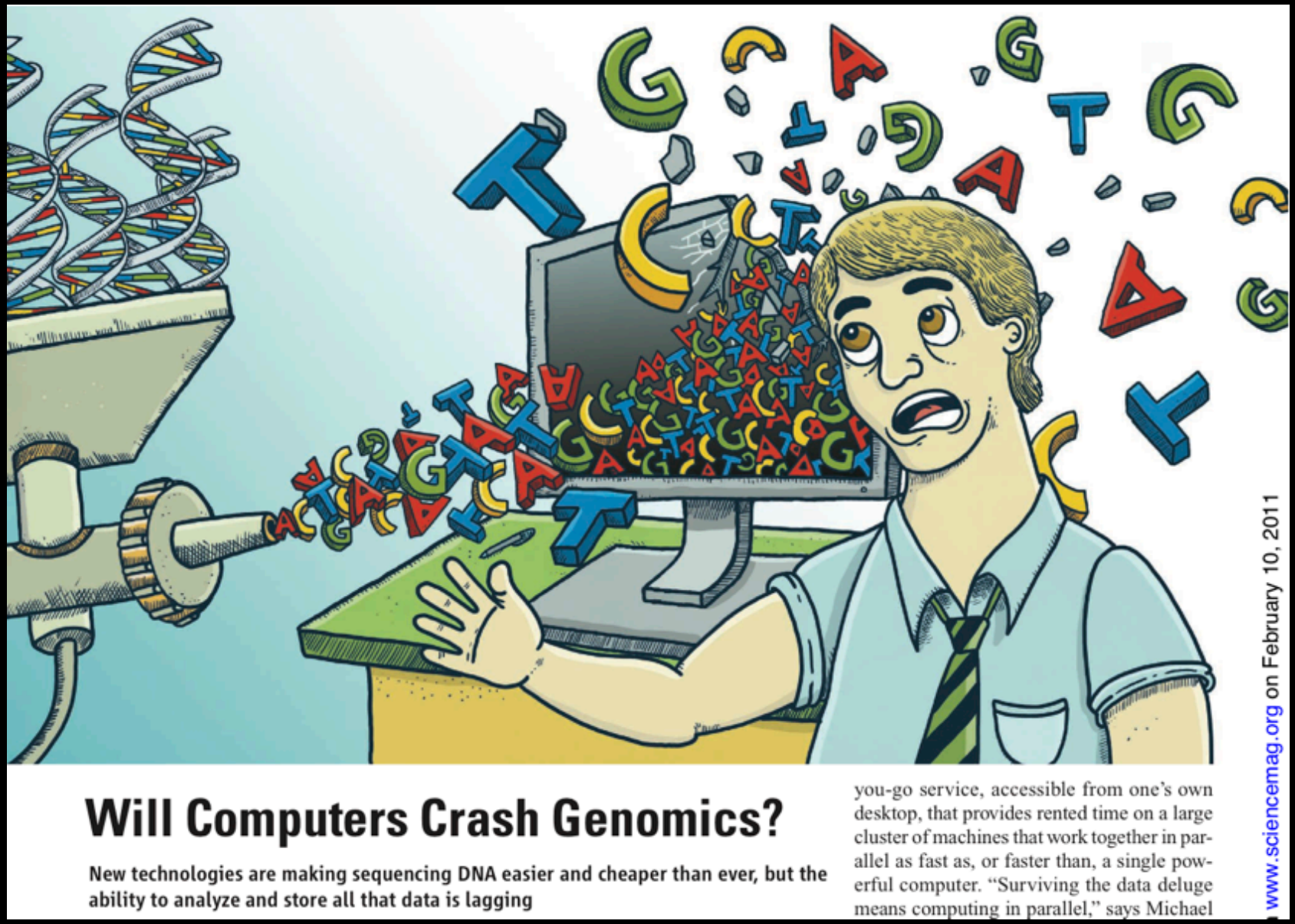
- personalized medicine
- aging
- environment interactions
- pathogen analysis
- ...

Trends in Genomics (1)



Trends in Genomics (2)





Will Computers Crash Genomics?

New technologies are making sequencing DNA easier and cheaper than ever, but the ability to analyze and store all that data is lagging

you-go service, accessible from one's own desktop, that provides rented time on a large cluster of machines that work together in parallel as fast as, or faster than, a single powerful computer. "Surviving the data deluge means computing in parallel," says Michael

"Will Computers Crash Genomics?", Pennisi, E., *Science*, Feb 11, 2011

Challenges in Genomics

Generating data is easy

- high-throughput sequencing (HTS) technologies improving rapidly
- datasets are hundreds of MBs to GBs

Analyzing data is THE bottleneck

- computation is essential due to dataset size
- $1,000 genome, $1,000,000 interpretation?
  http://www.bio-itworld.com/2010/10/01/interpretation.html

Using Computation in Science?

Scientists often not trained in computation

Reproducibility hindered by complexity: systems, scripts, tools, parameters

Collaboration and publishing difficult because current media do not support computational artifacts well

Overview

Genomics

Galaxy

- accessible, reproducible, and transparent science
- on the cloud
- visual analytics

Reflections on Galaxy

Galaxy Project: Fundamental Questions

When Biology (or any science) becomes dependent on computational methods:

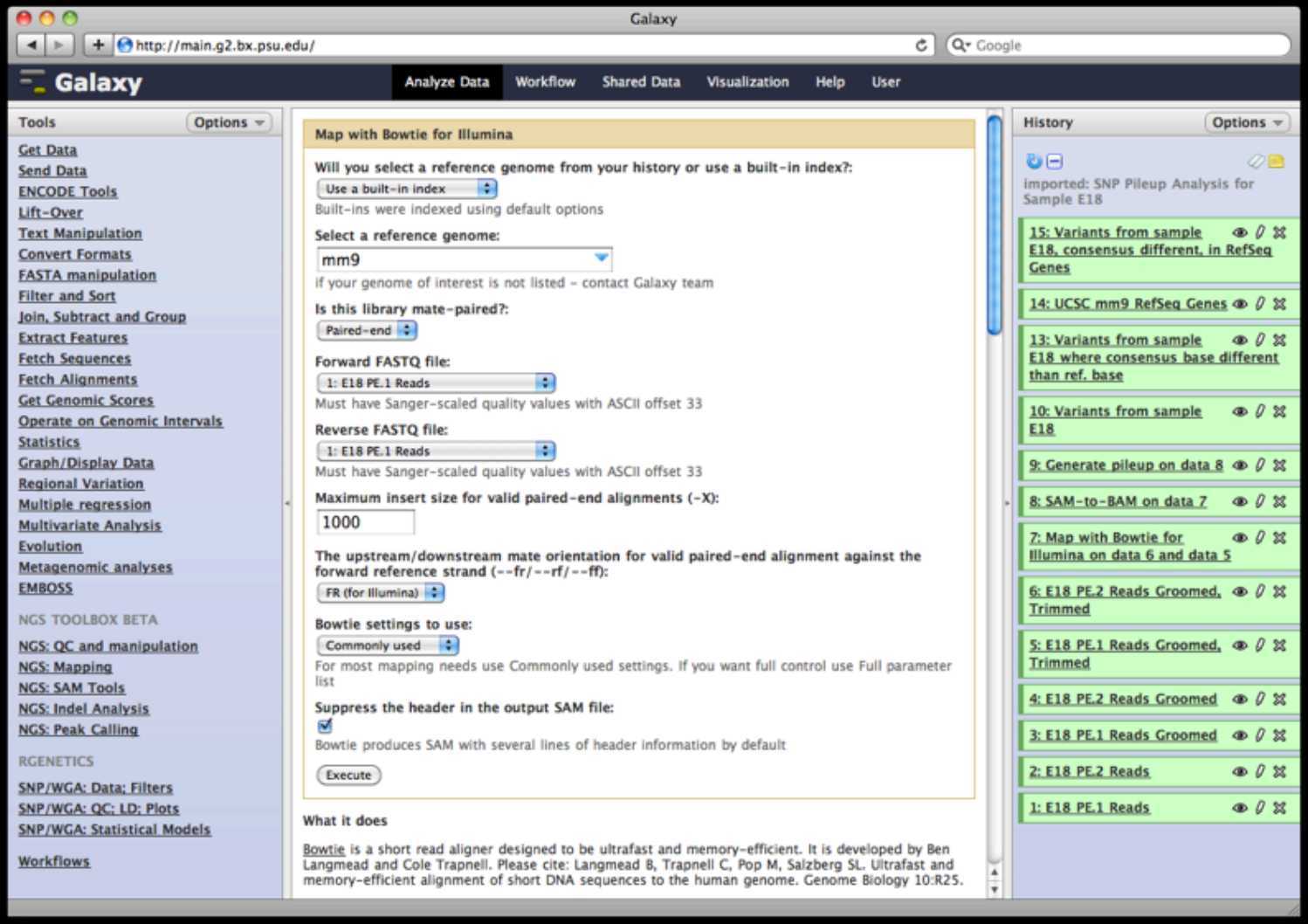
- how can those methods best be made accessible to scientists?
- how best to ensure that analyses are reproducible?
- how best to facilitate transparent communication and reuse of analyses?

Vision

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

Accessibility

Connecting Users with Tools



Accessibility

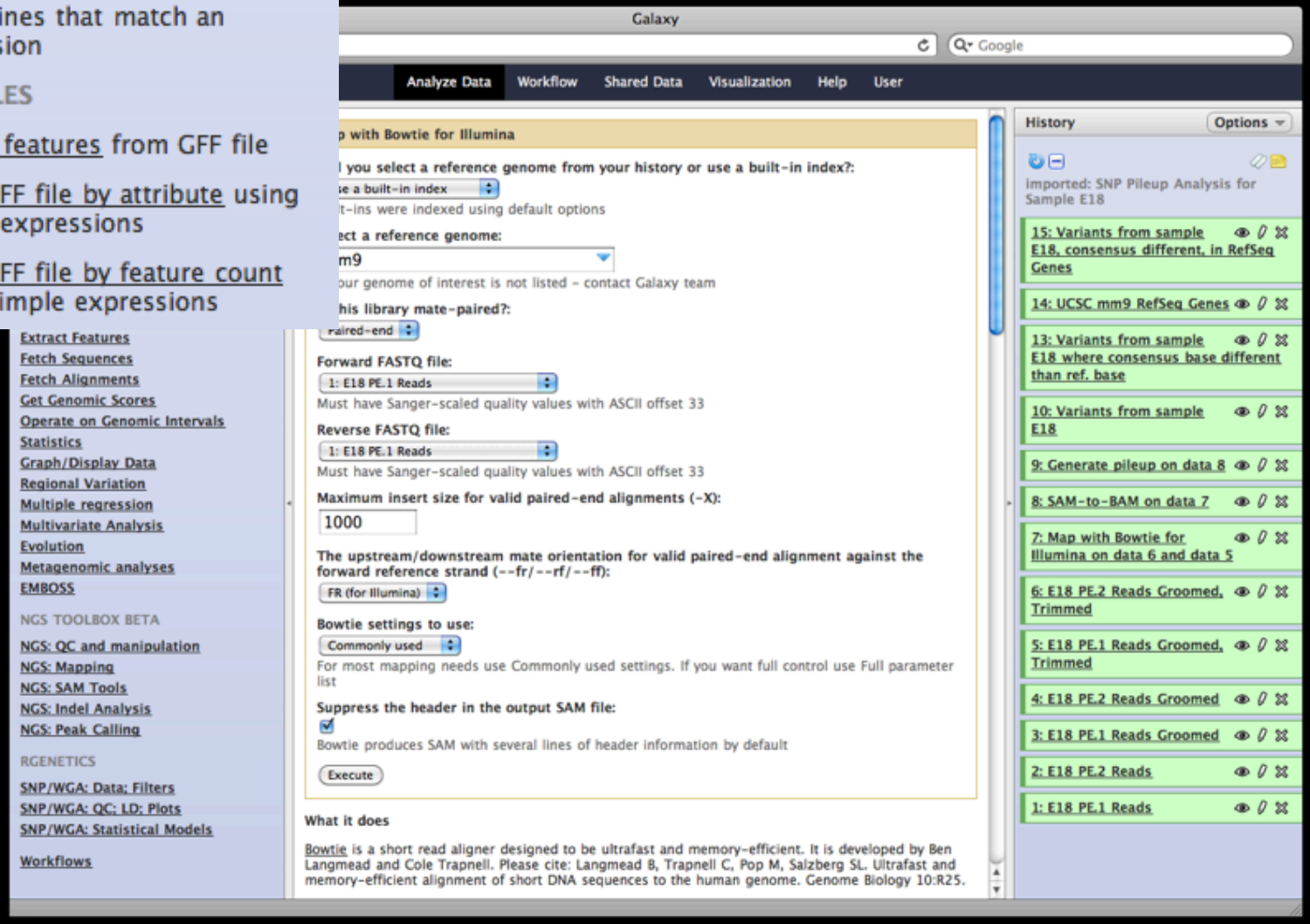
ecting Users with Tools

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression

GFF FILES

- Extract features from GFF file
- Filter GFF file by attribute using simple expressions
- Filter GFF file by feature count using simple expressions



Accessibility

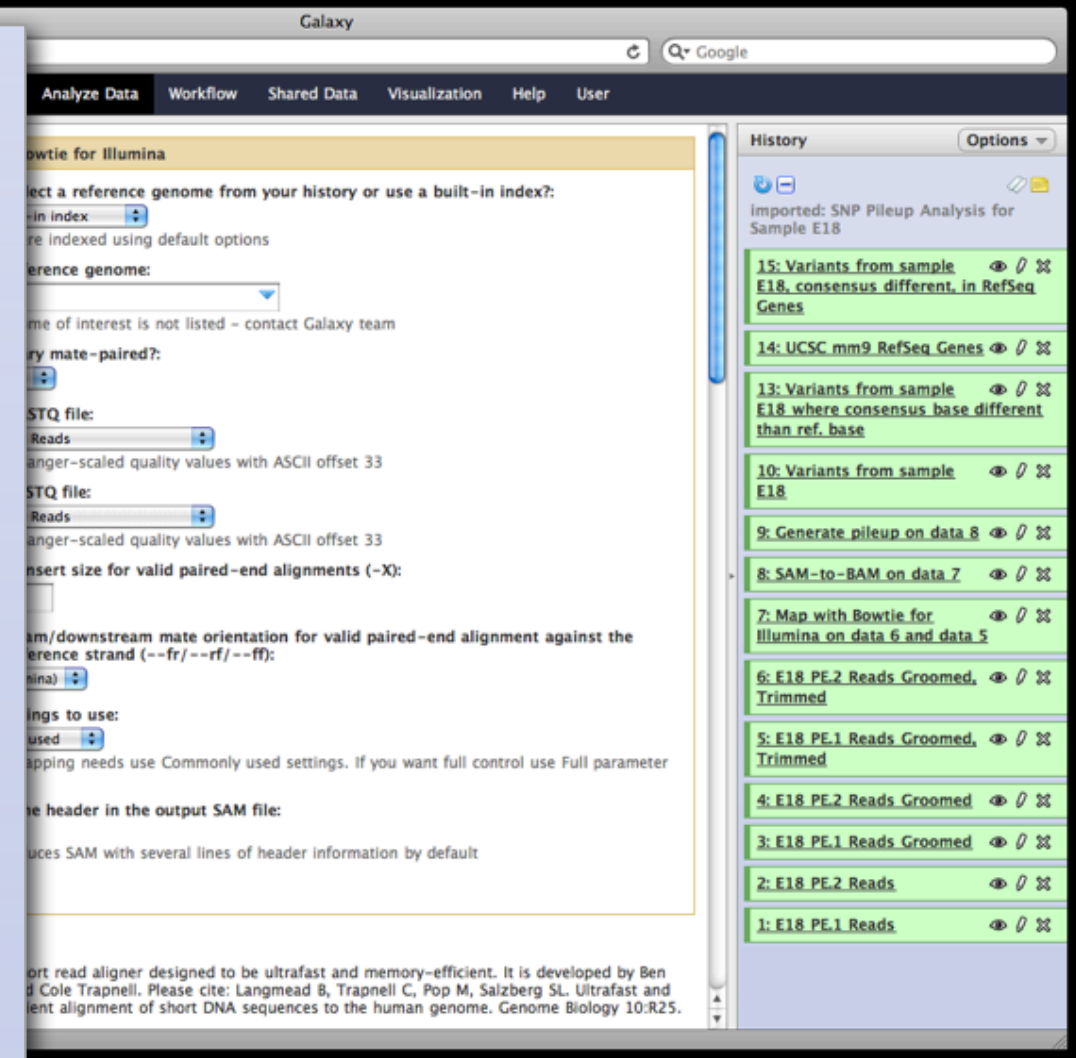
ecting Users with Tools

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an e
- E
- Fi si
- Fi u

Operate on Genomic Intervals

- Intersect the intervals of two queries
- Subtract the intervals of two queries
- Merge the overlapping intervals of a query
- Concatenate two queries into one query
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a query
- Cluster the intervals of a query
- Join the intervals of two queries side-by-side
- Get flanks returns flanking region/s for every gene
- Fetch closest feature for every interval
- Profile Annotations for a set of genomic intervals



Accessibility

ecting Users with Tools

Filter and Sort

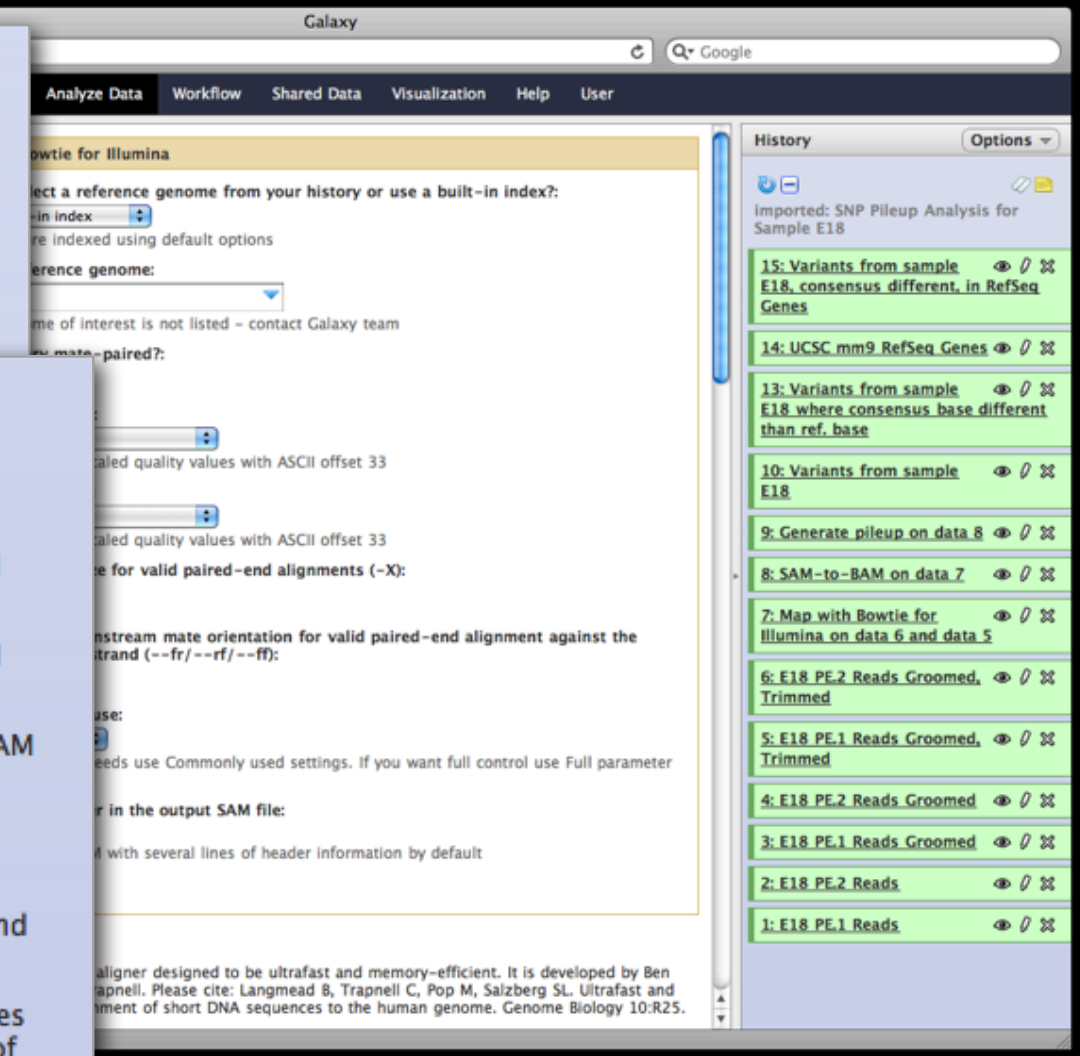
- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an e
- E
- Fi si
- Fi u

Operate on Genomic Intervals

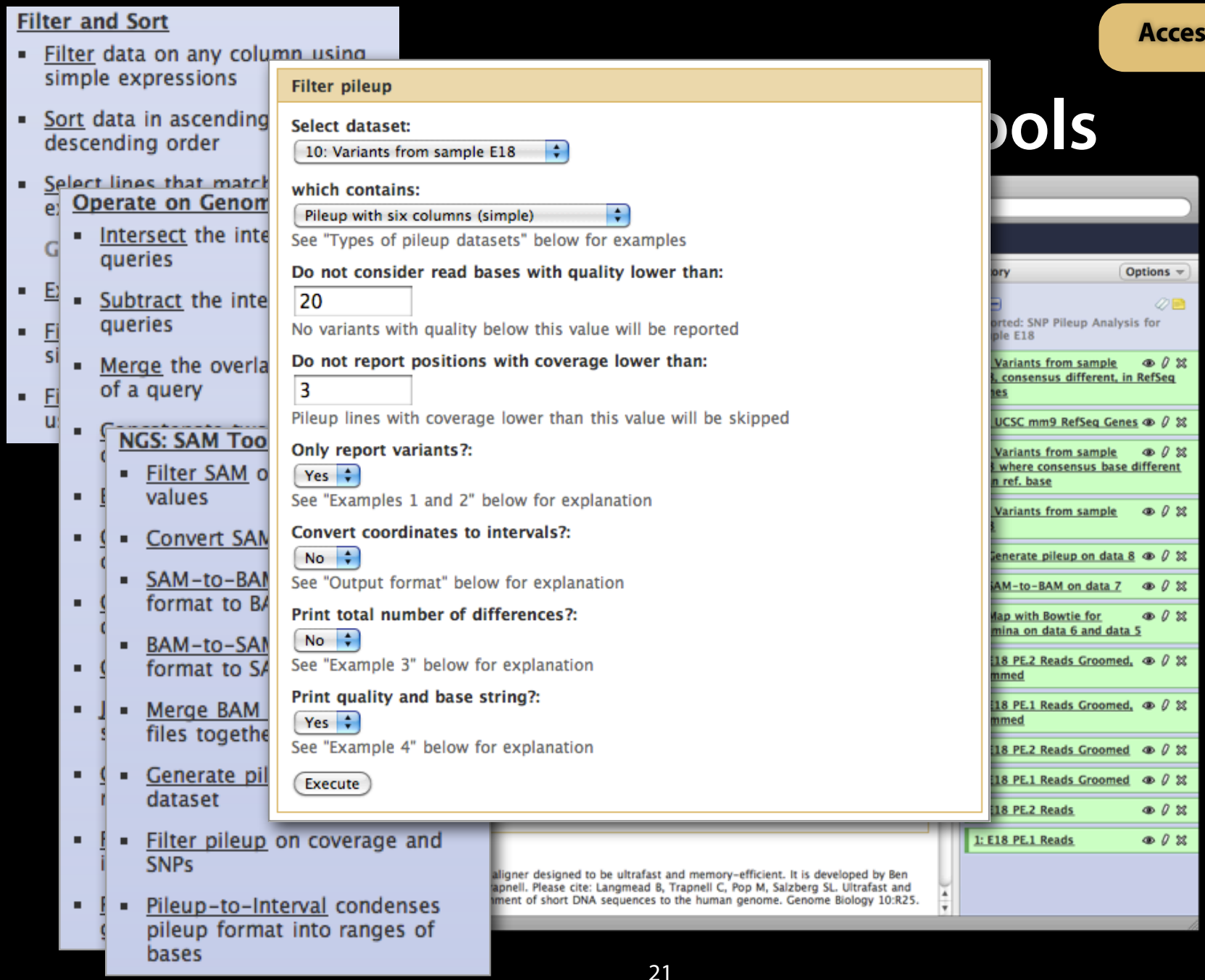
- Intersect the intervals of two queries
- Subtract the intervals of two queries
- Merge the overlapping intervals of a query
-
-
-
-
-
-
-

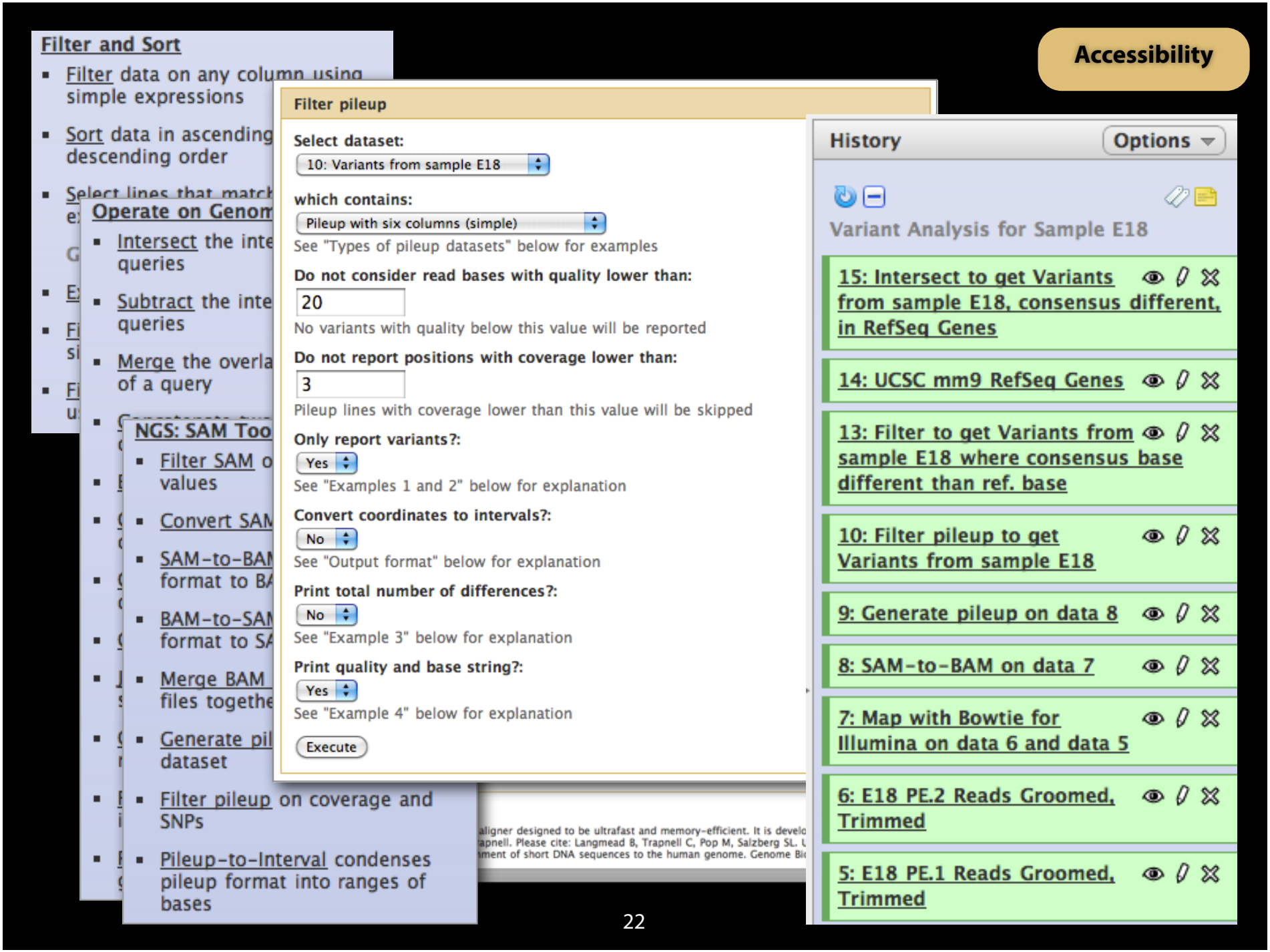
NGS: SAM Tools

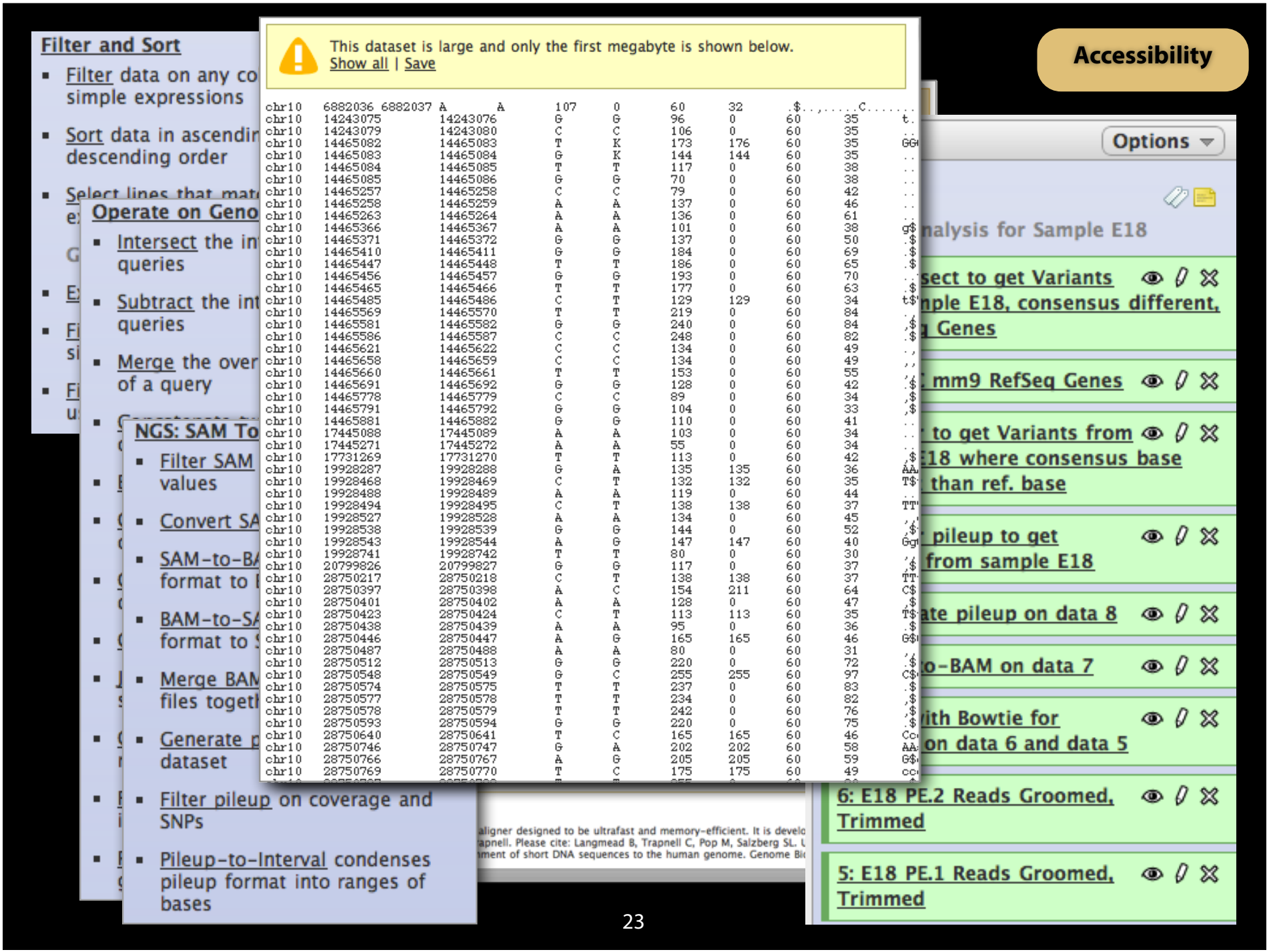
- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases



Accessibility

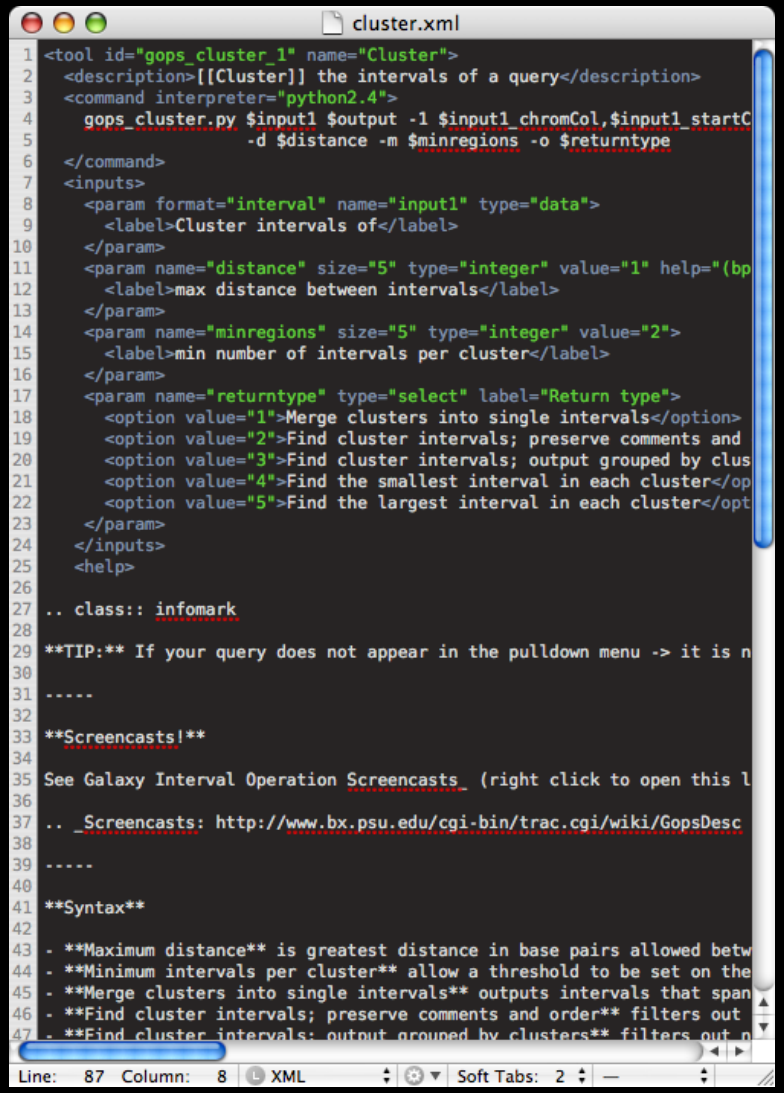



Accessibility
Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending descending order
Operate on Genom
Intersect the inte queries
Subtract the inte queries
Merge the overla of a query
NGS: SAM Too
Filter SAM o values
Convert SAM
SAM-to-BAM format to BA
BAM-to-SAM format 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


Accessibility
Filter and Sort
Filter data on any co
simple expressions
Sort data in ascendir
descending order
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
Options
nalysis for Sample E18
sect to get Variants
nple E18, consensus different,
Genes
mm9 RefSeq Genes
to get Variants from
E18 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

Accessibility

A Tool in Galaxy



Defined via abstract interface:

- inputs & outputs
- parameters
- how to generate command line

As simple as possible but allows for rigorous reasoning

Reproducibility in Genomics

18 *Nat. Genetics* experiments in microarray gene expression

<50% of reproducible

Problems

- missing data (38%)
- missing software, hardware details (50%)
- missing method, processing details (66%)

Ioannidis, J.P.A. et al. "Repeatability of published microarray gene expression analyses." Nat Genet 41, 149-155 (2009)

Reproducibility in Genomics

18 *Nat. Genetics* experiments in microarray gene expression

<50% of reproducible

Problems

- missing data (38%)
- missing software, hardware details (50%)
- missing method, processing details (66%)

Ioannidis, J.P.A. et al. "Repeatability of published microarray gene expression analyses." Nat Genet 41, 149-155 (2009)

14 re-sequencing experiments in *Nat. Genetics*, *Nature*, *Science*

0% reproducible?

Problems

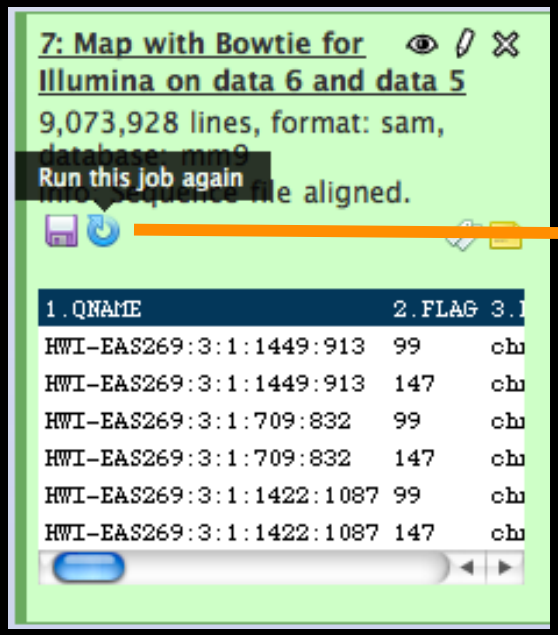
- missing primary data (50%)
- tools unavailable (50%)
- missing parameter setting, tool versions (100%)

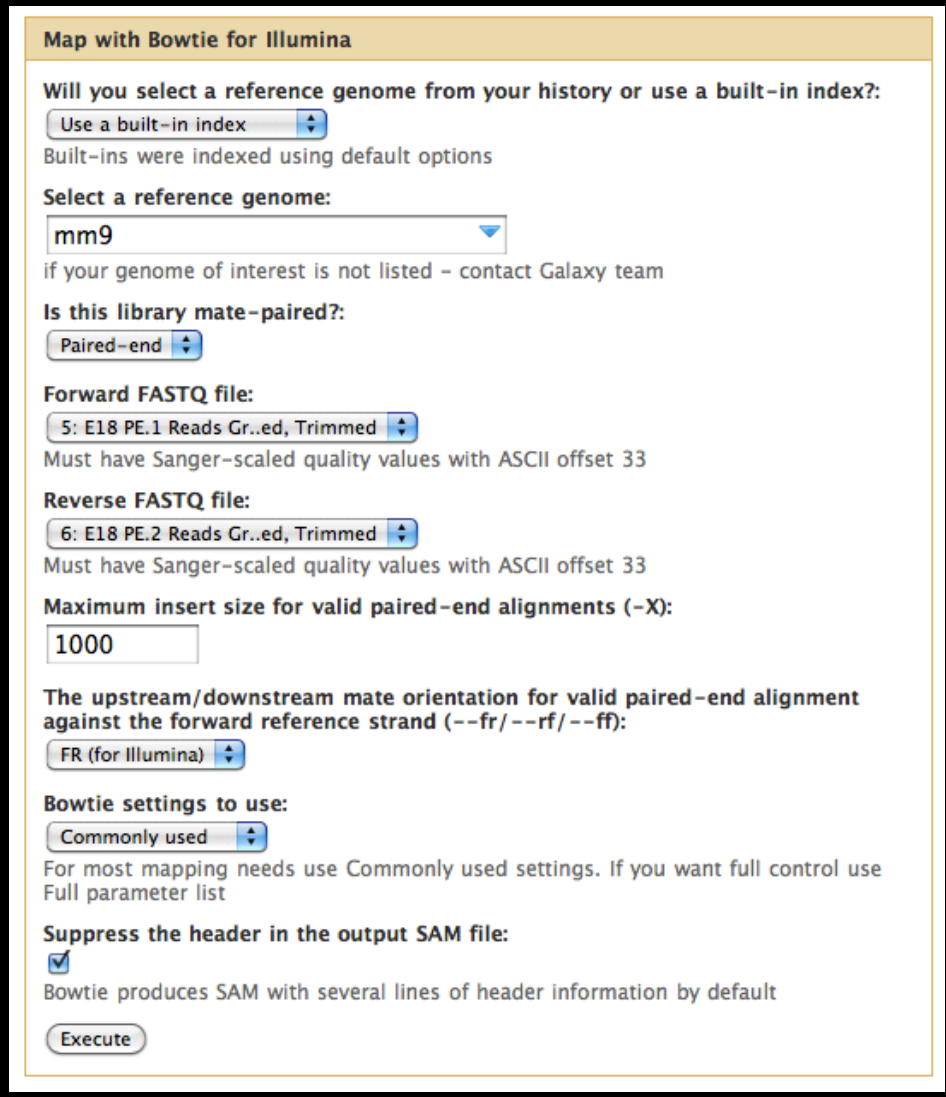
"Devil in the details," Nature, vol. 470, 305-306 (2011).

Metadata = Reproducibility

Reproducibility

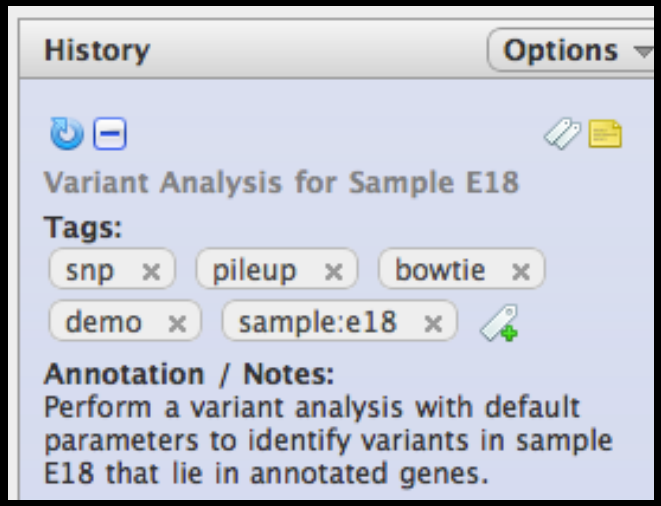
Automatic Metadata

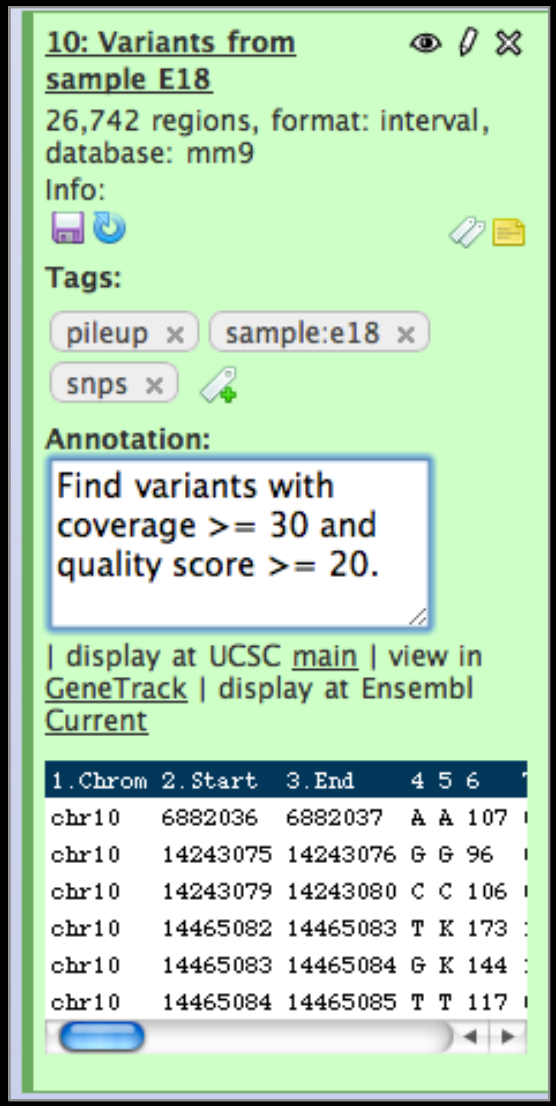




Reproducibility

User Metadata





Data Provenance

Datasets are immutable in Galaxy

- associations between *Dataset* objects and their usage: *HistoryDatasetAssociation*, *LibraryDatasetAssociation*
- metadata associated with associations, not datasets

Exporting from Galaxy

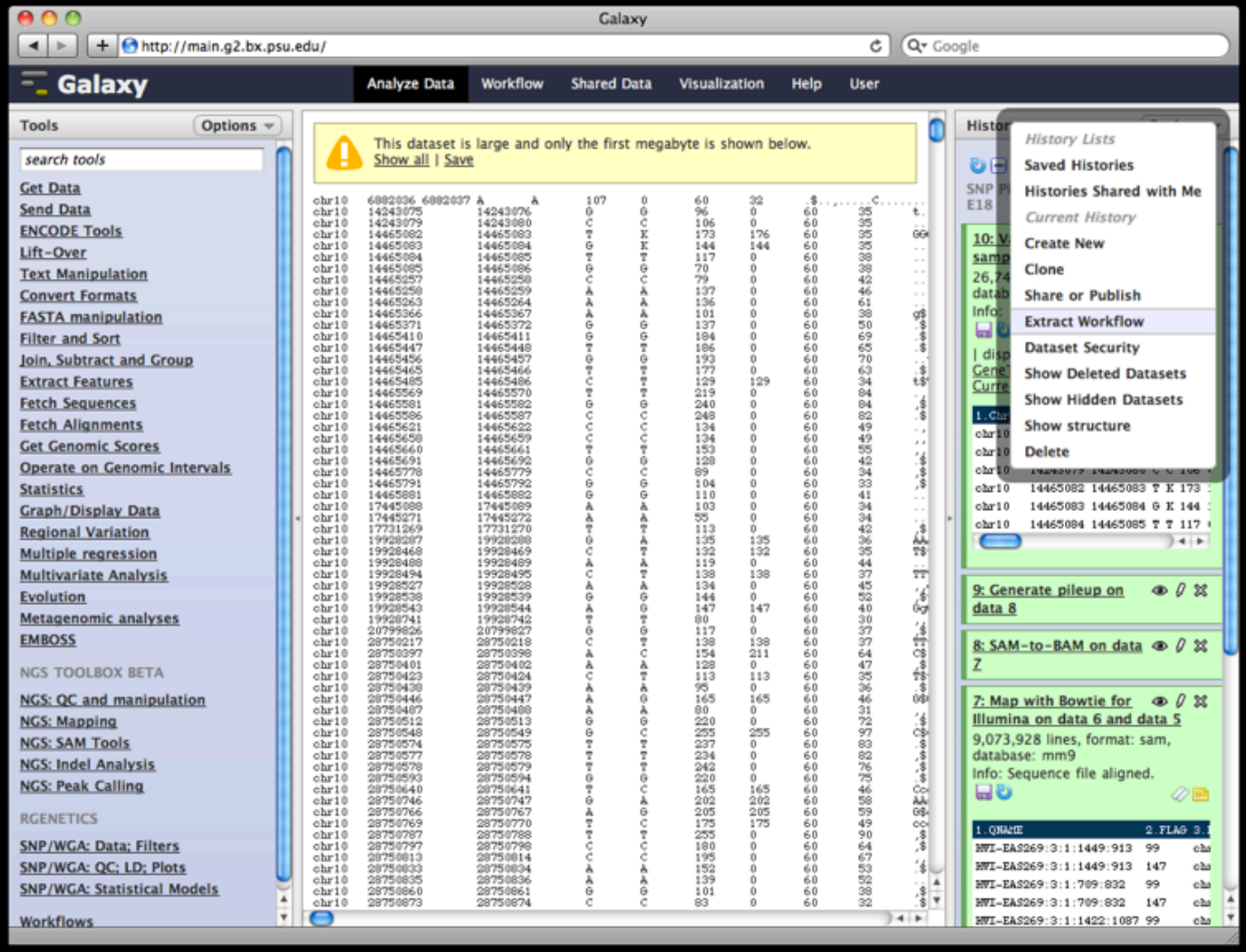
- histories and workflows can be exported in JSON format

Metadata

- Datatype defined in Python code
- Job/tool metadata has official
- JSON format in database and when exported

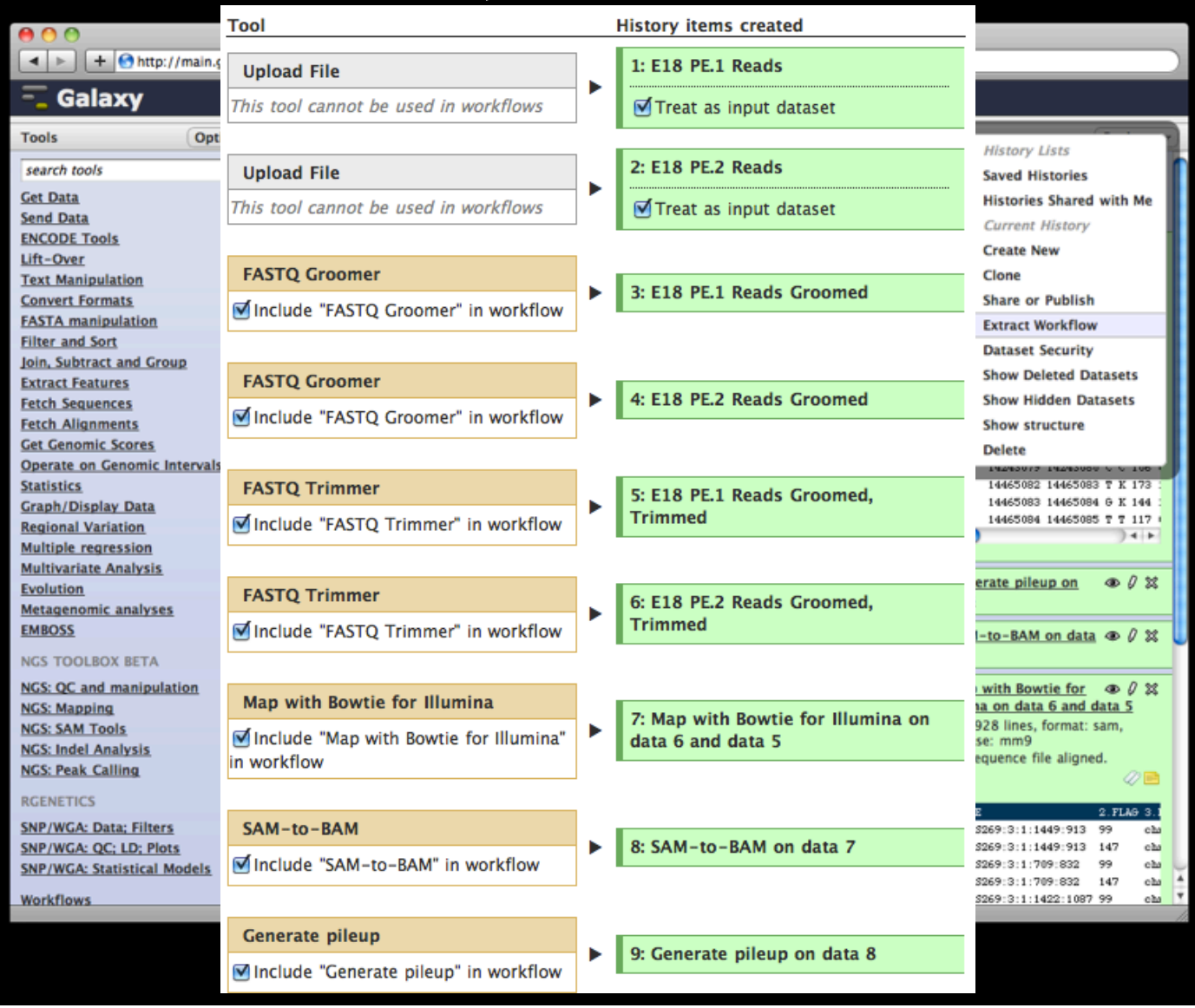
Reproducibility

Galaxy Workflows



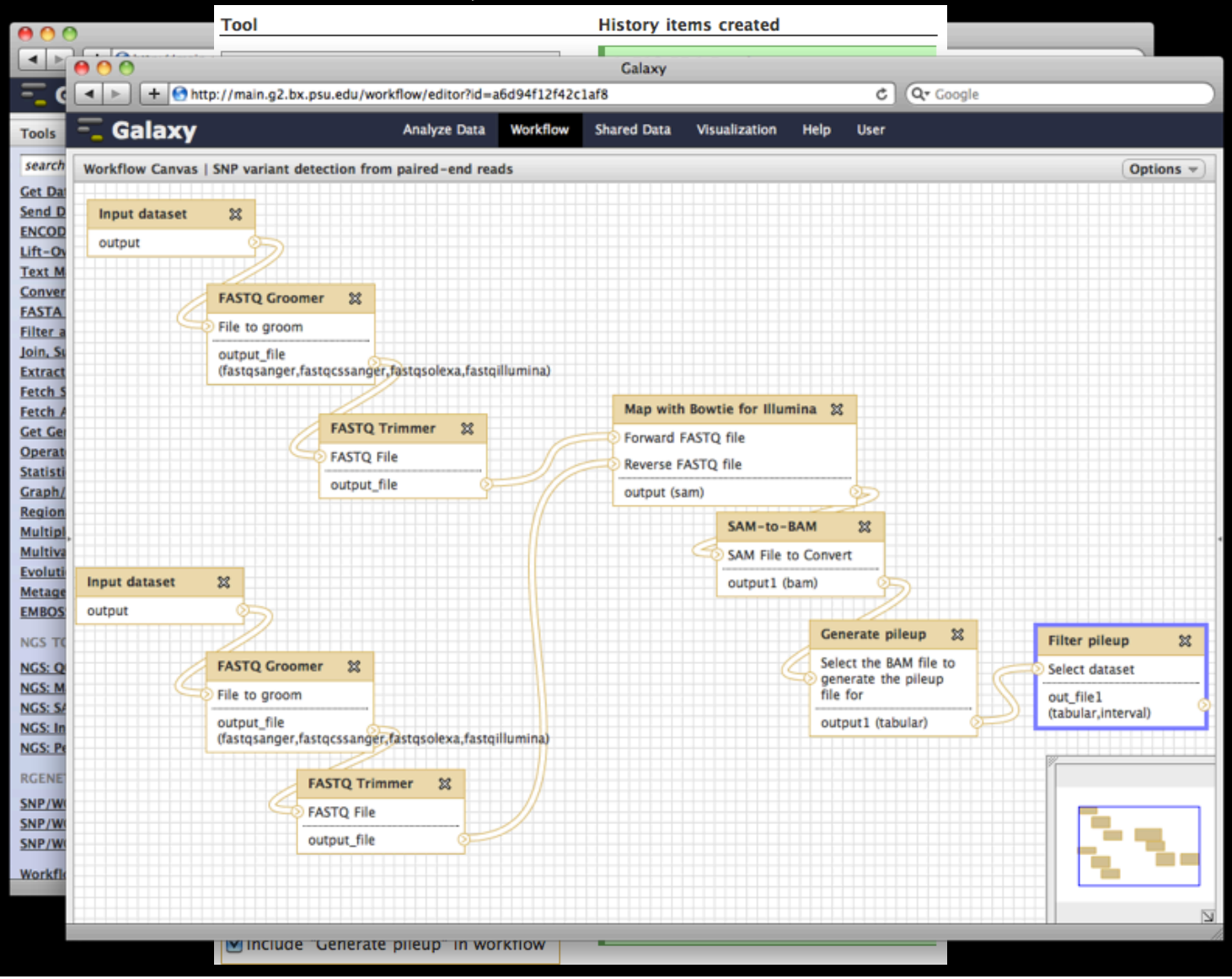
Reproducibility

Galaxy Workflows



Reproducibility

Galaxy Workflows

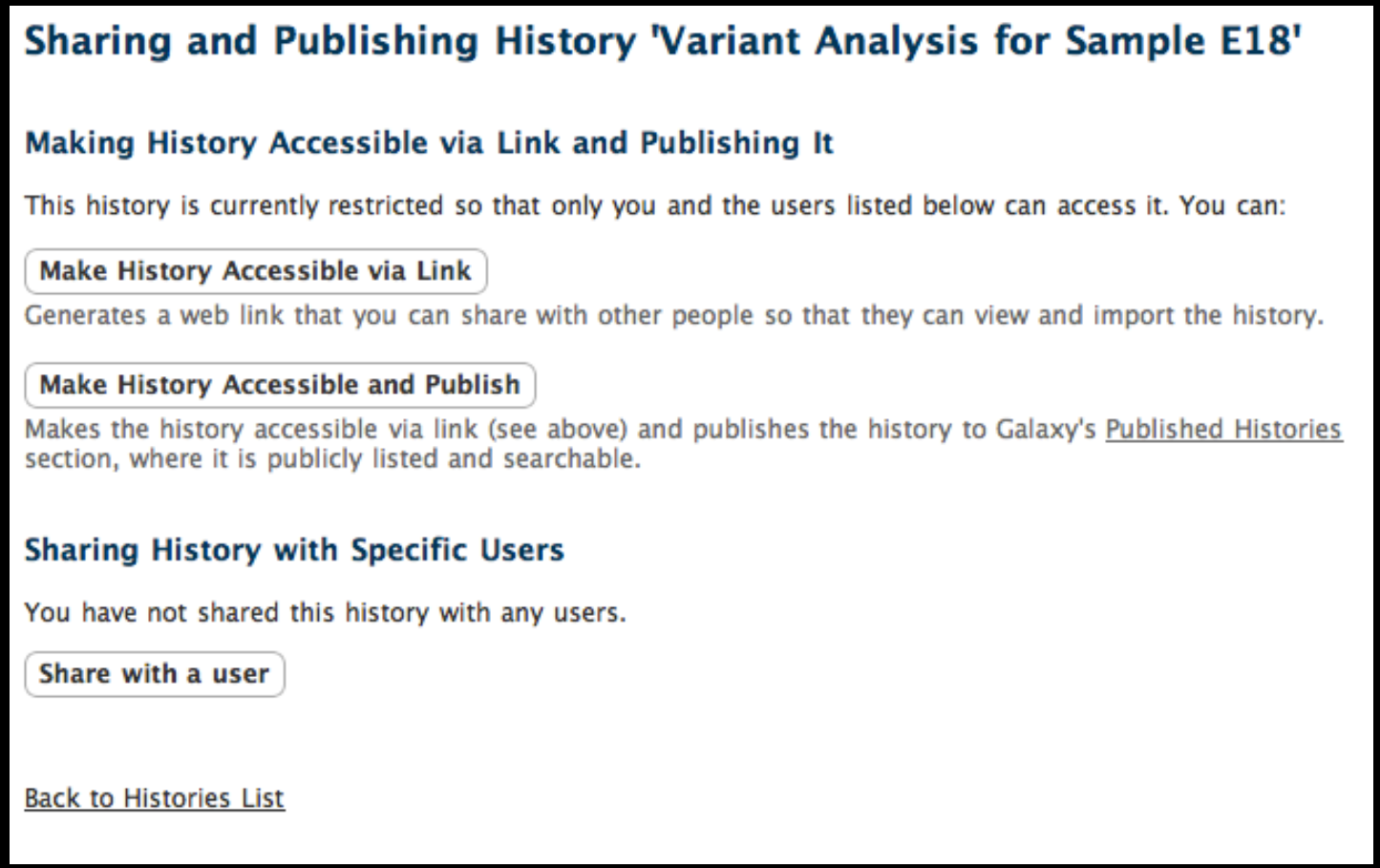


Transparency

Sharing, Collaborating, and Publishing with Galaxy

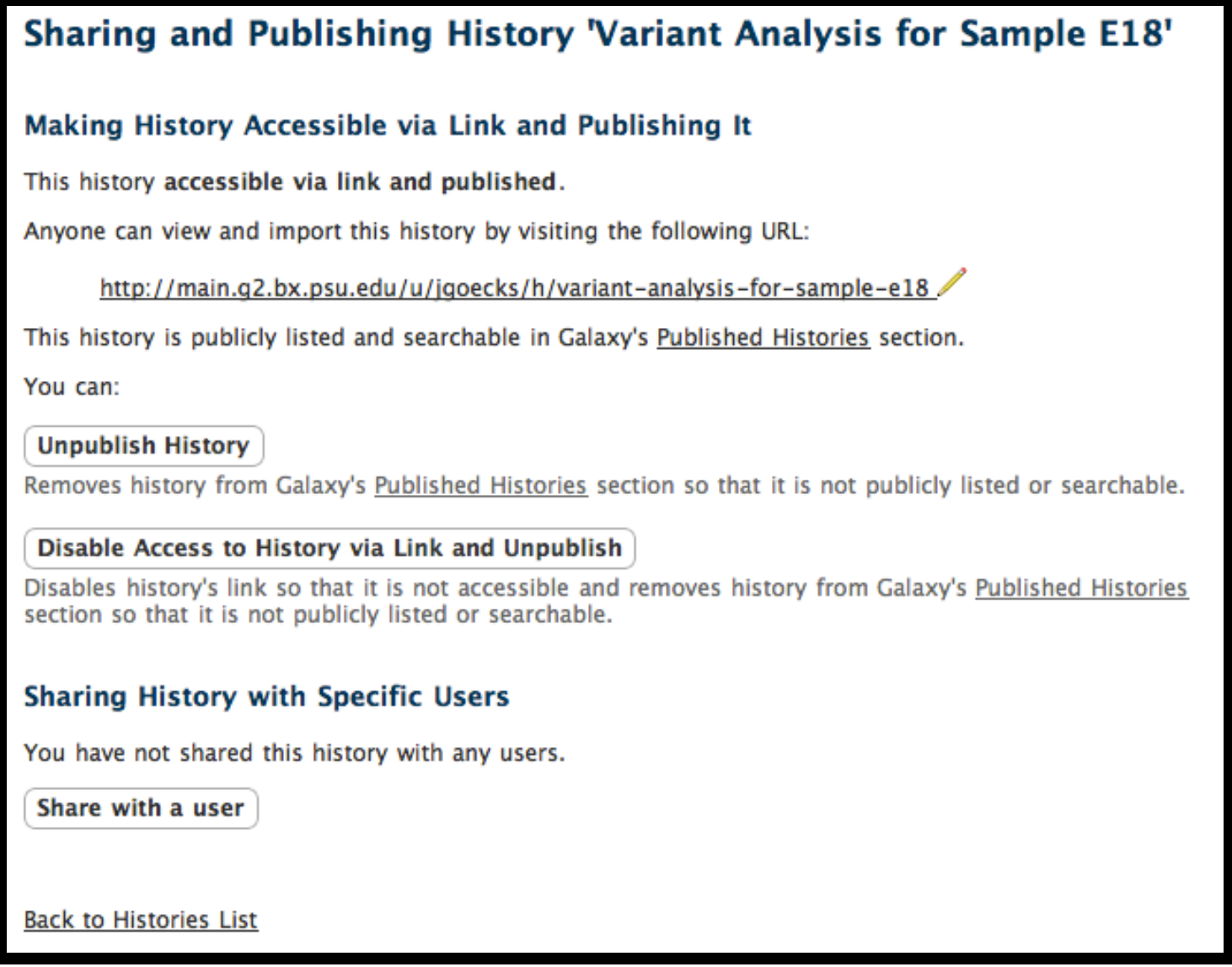
Transparency

Sharing and Publishing

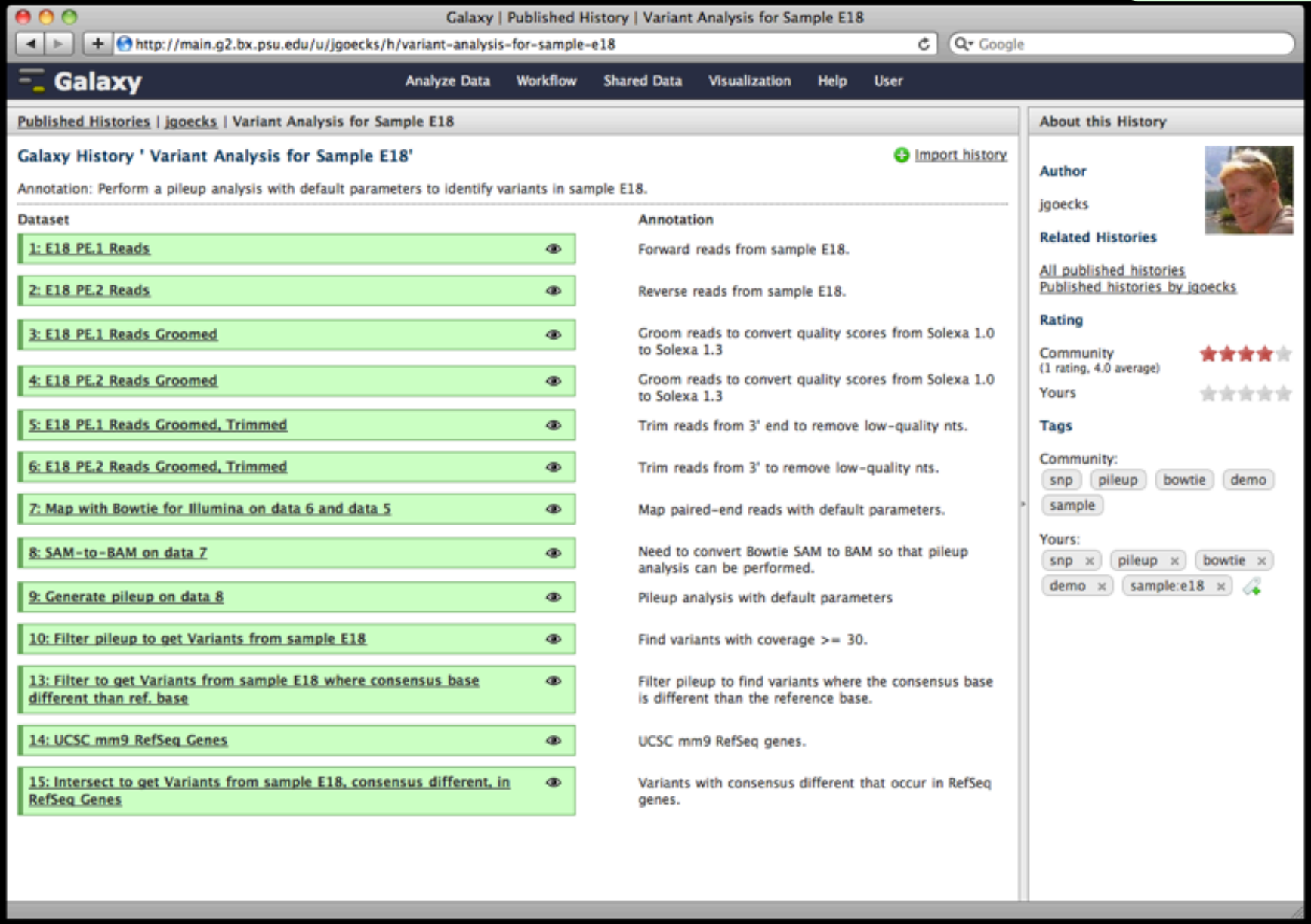


Transparency

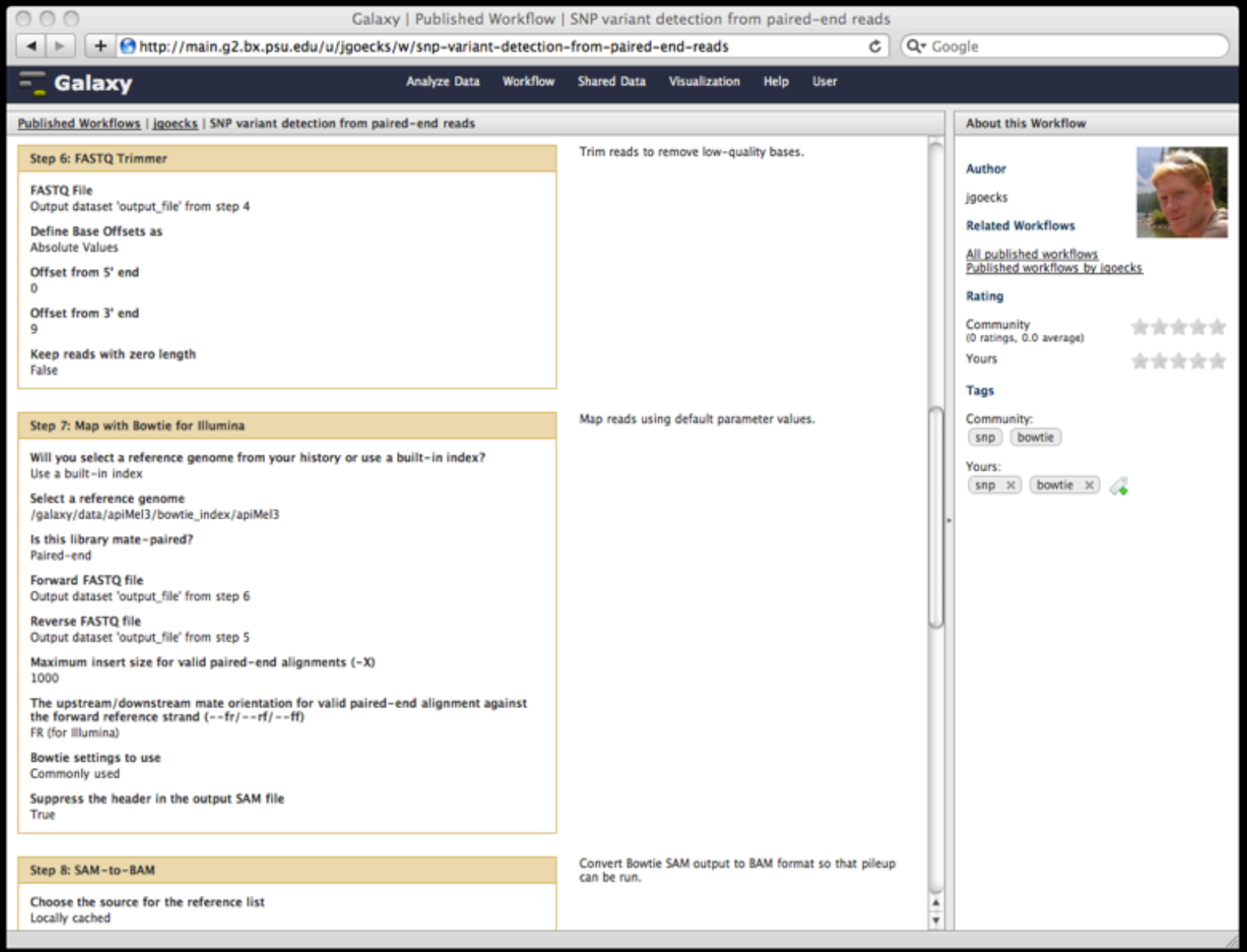
Sharing and Publishing



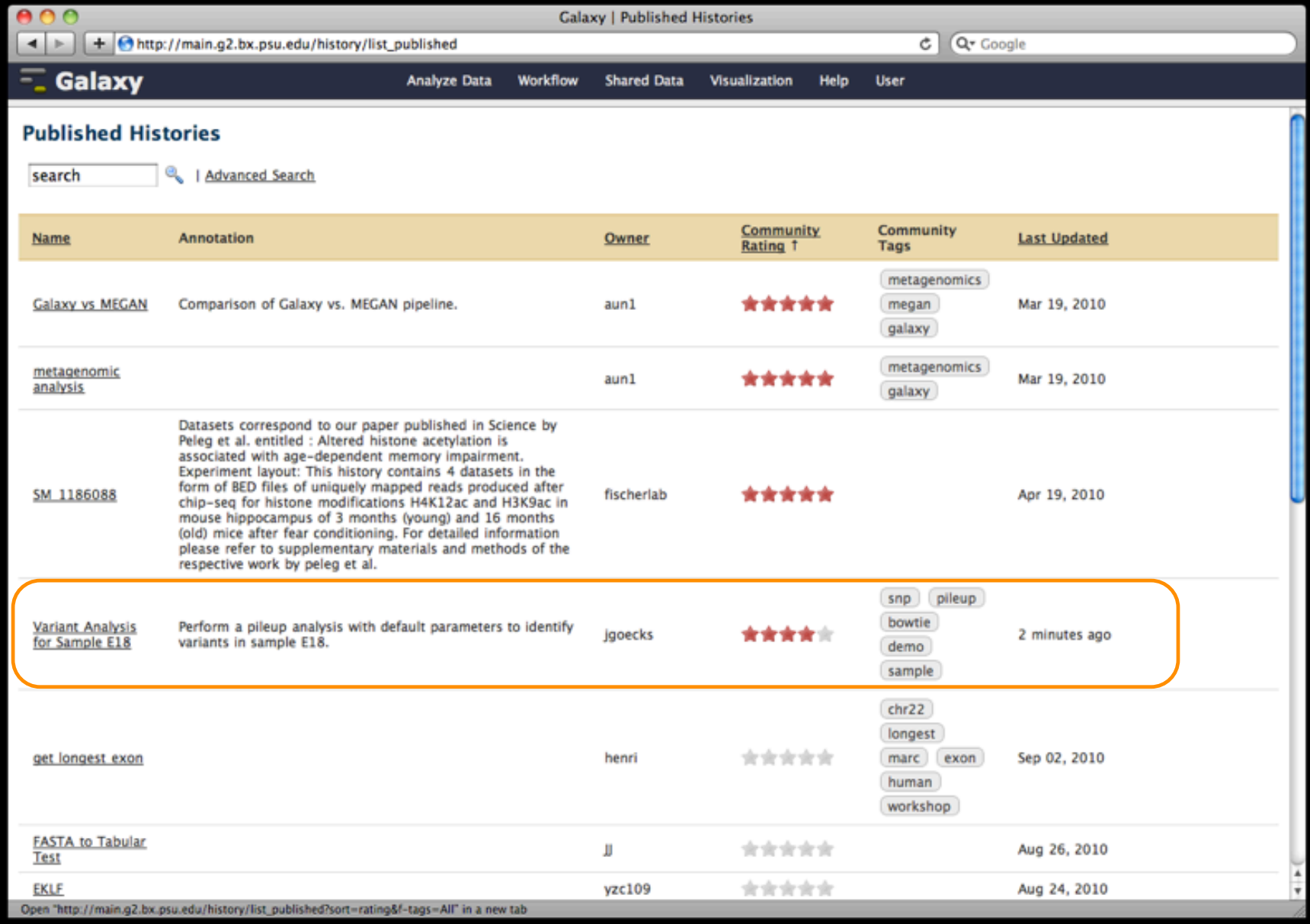
Transparency



Transparency



Transparency

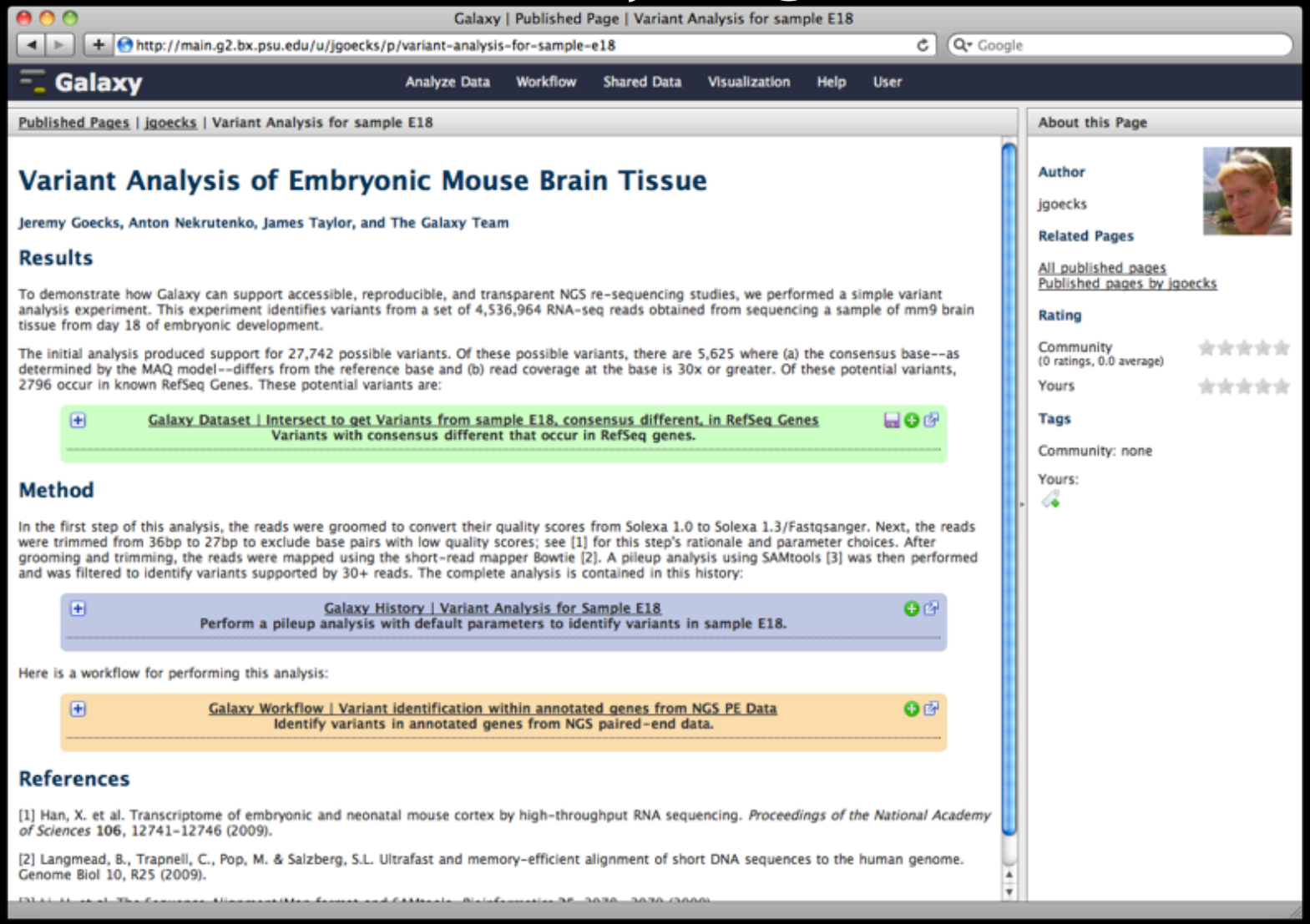


Galaxy Pages

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

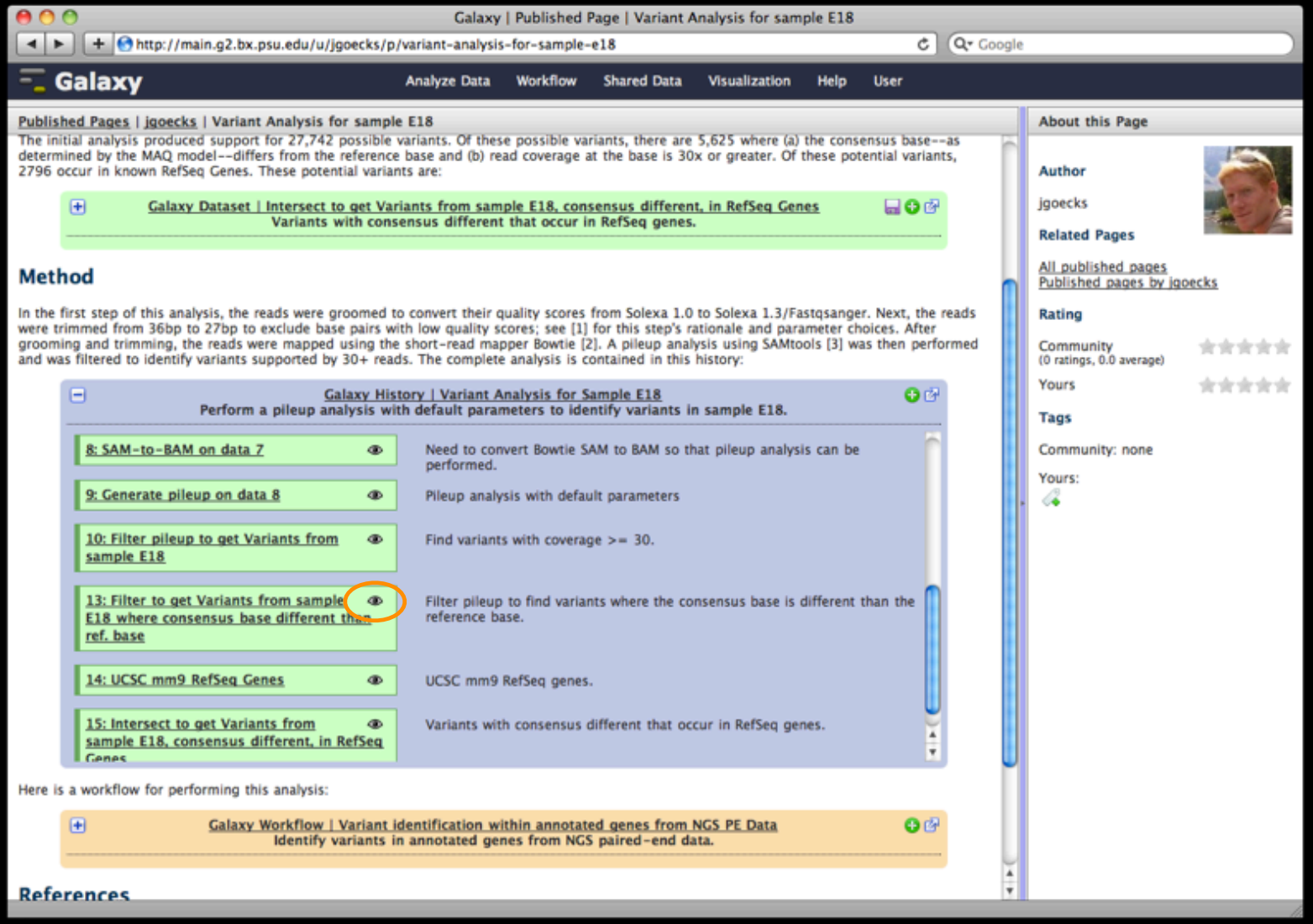
Transparency

Galaxy Pages



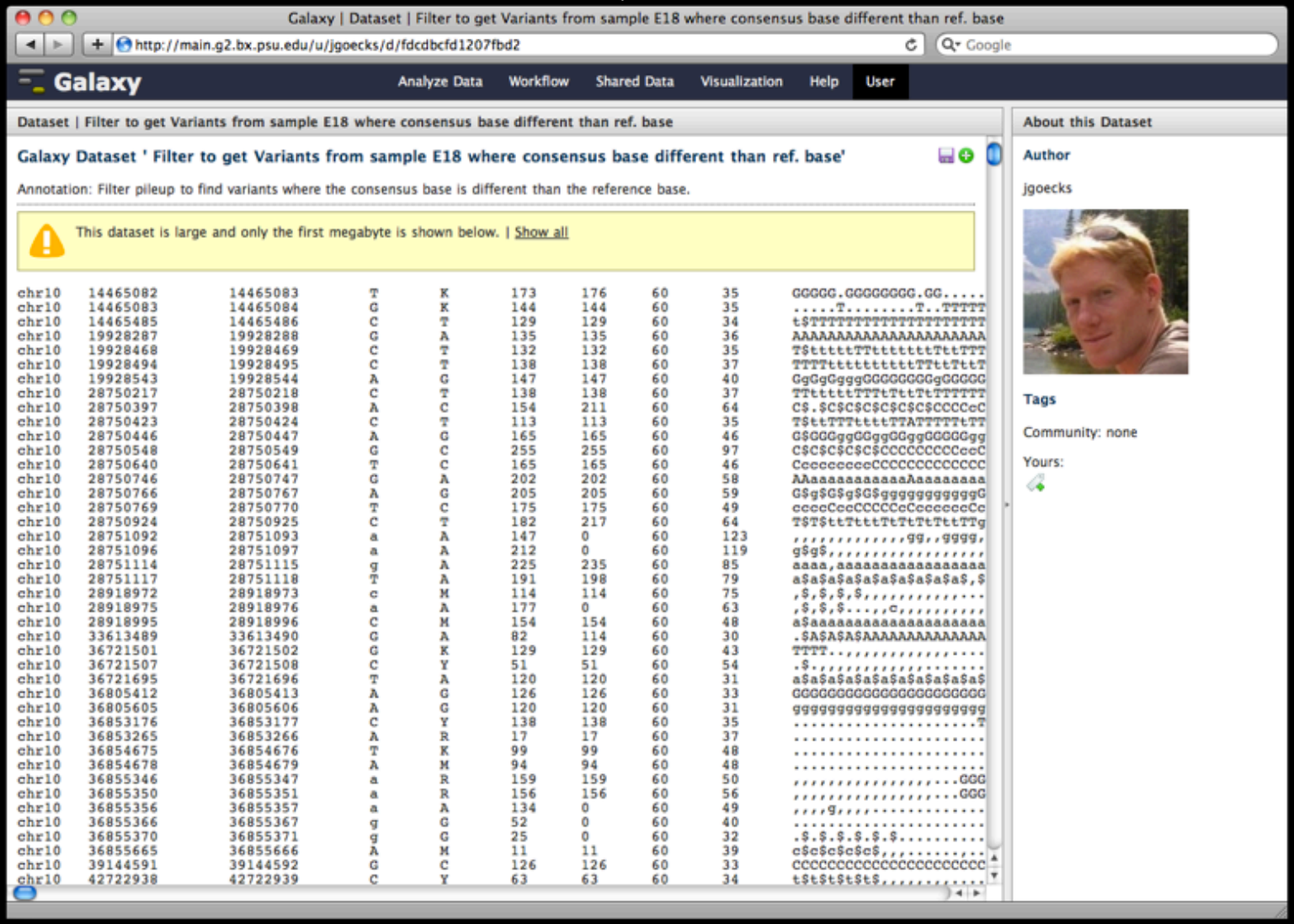
Transparency

Galaxy Pages



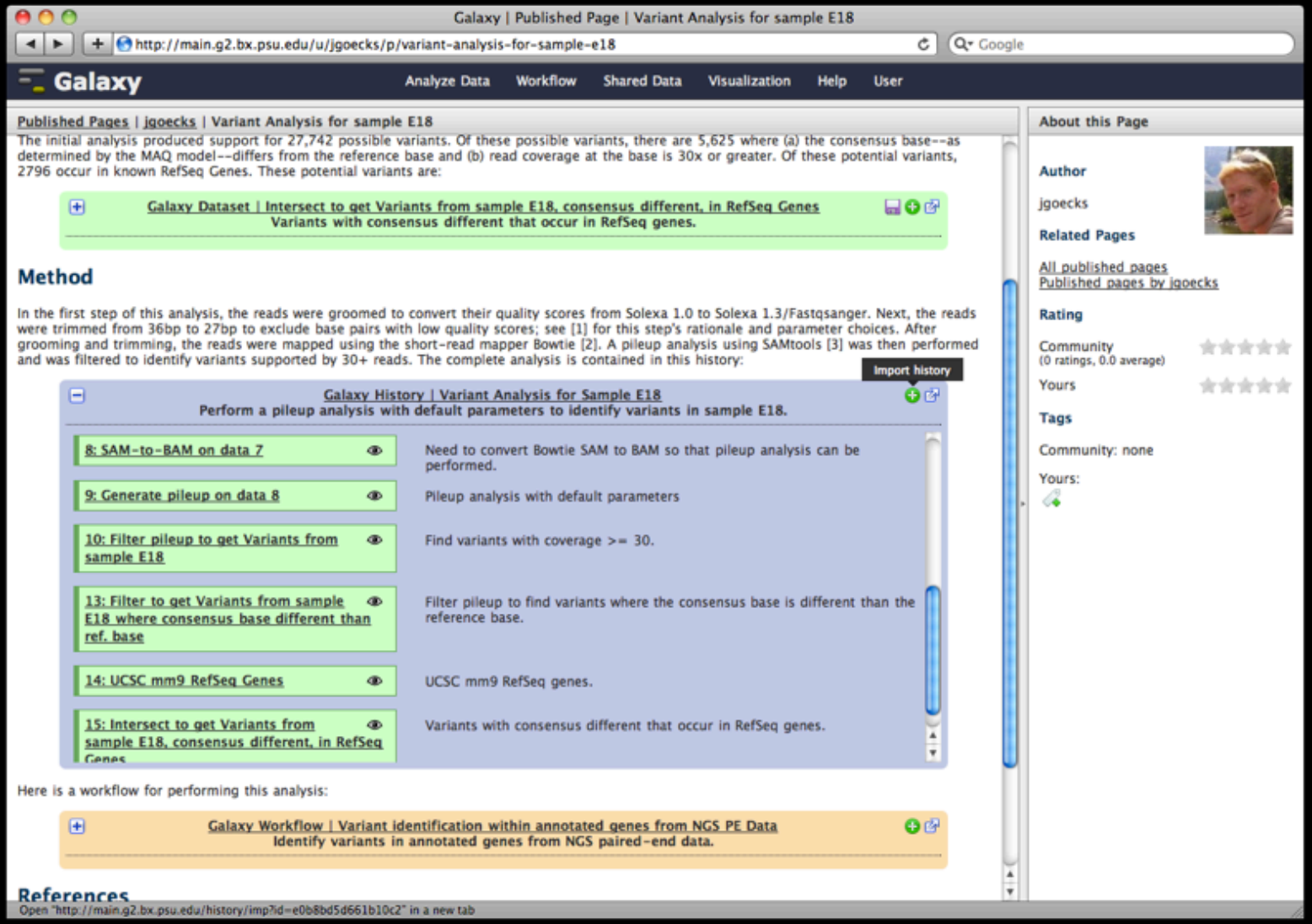
Transparency

Galaxy Pages



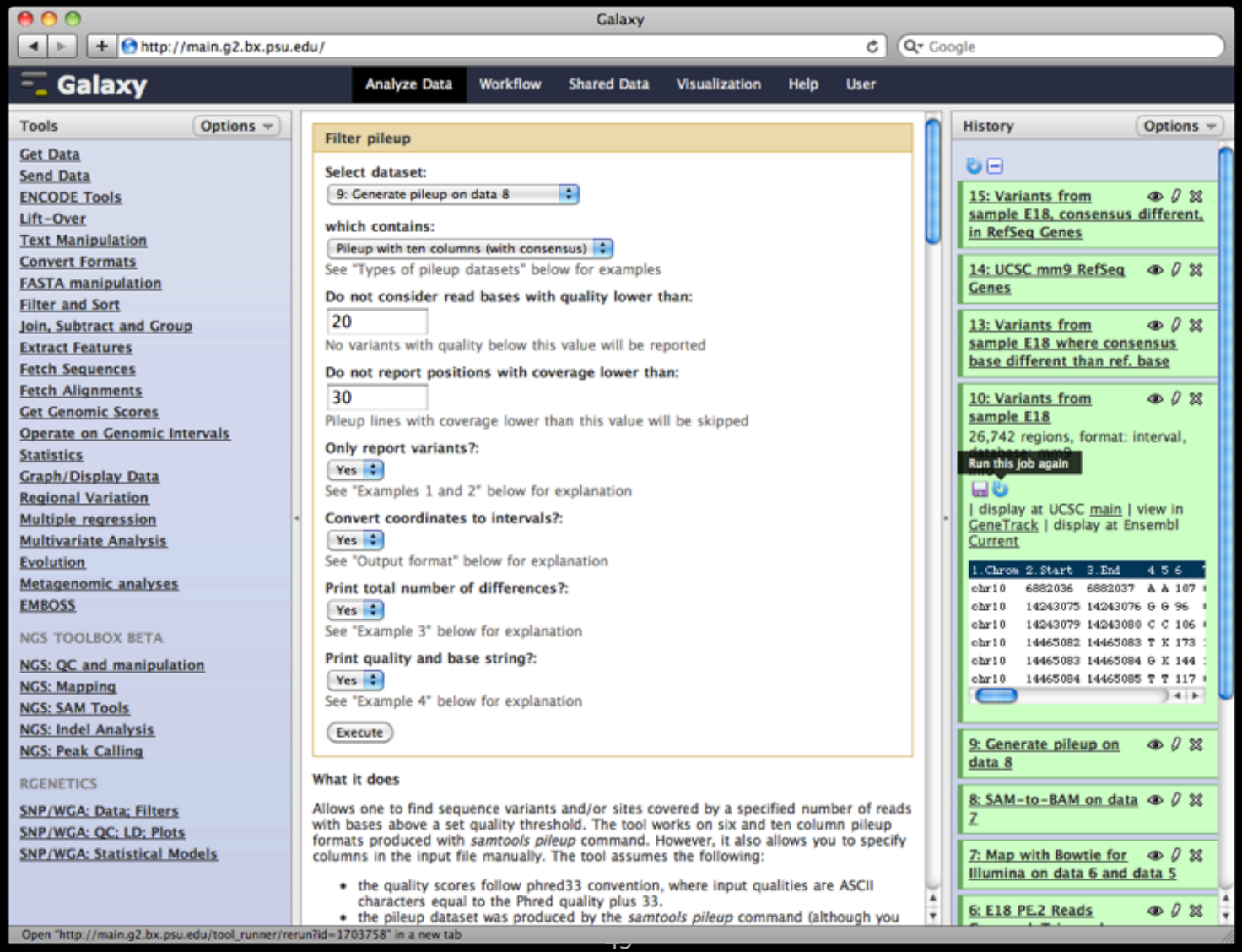
Transparency

Galaxy Pages



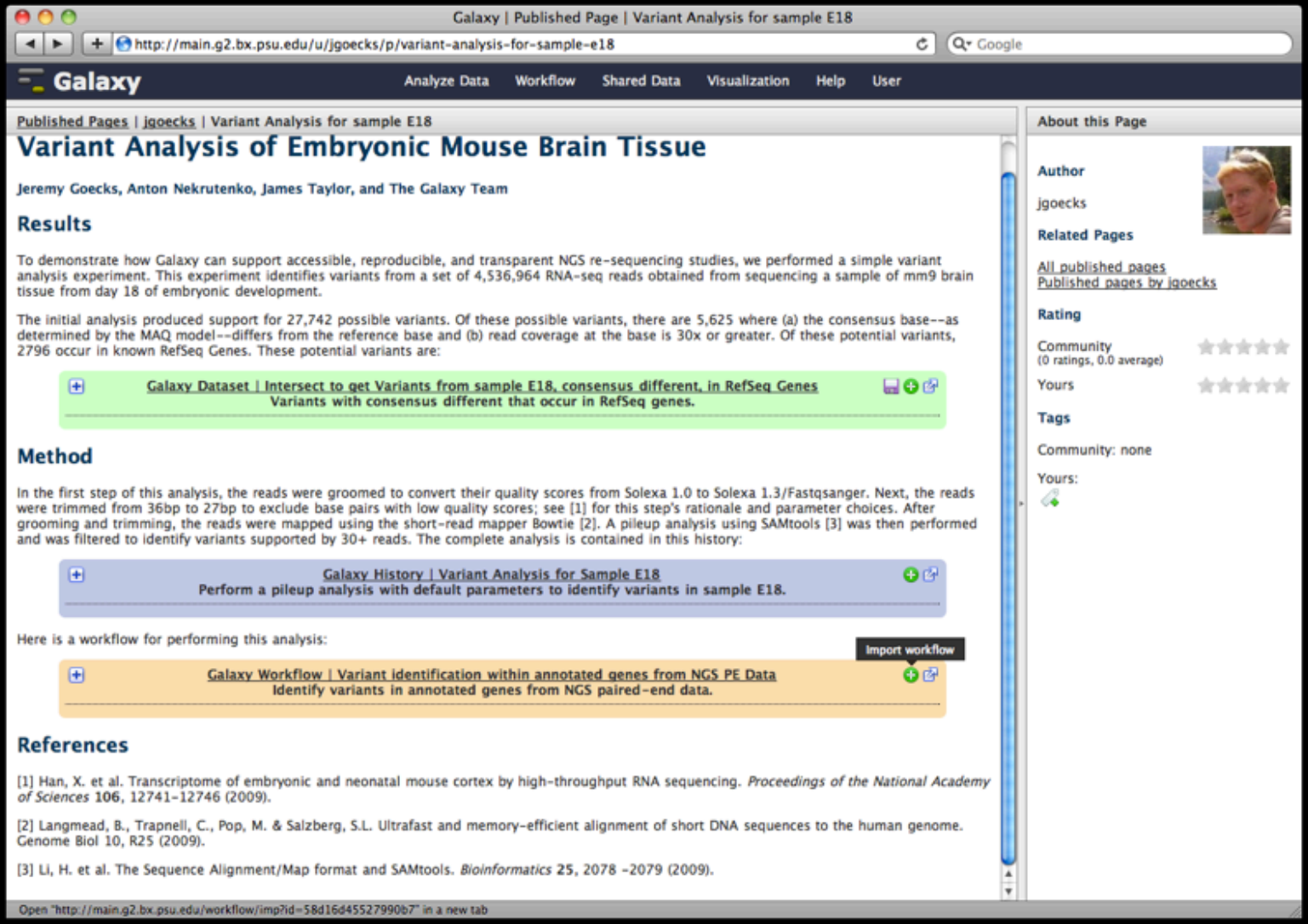
Transparency

Galaxy Pages



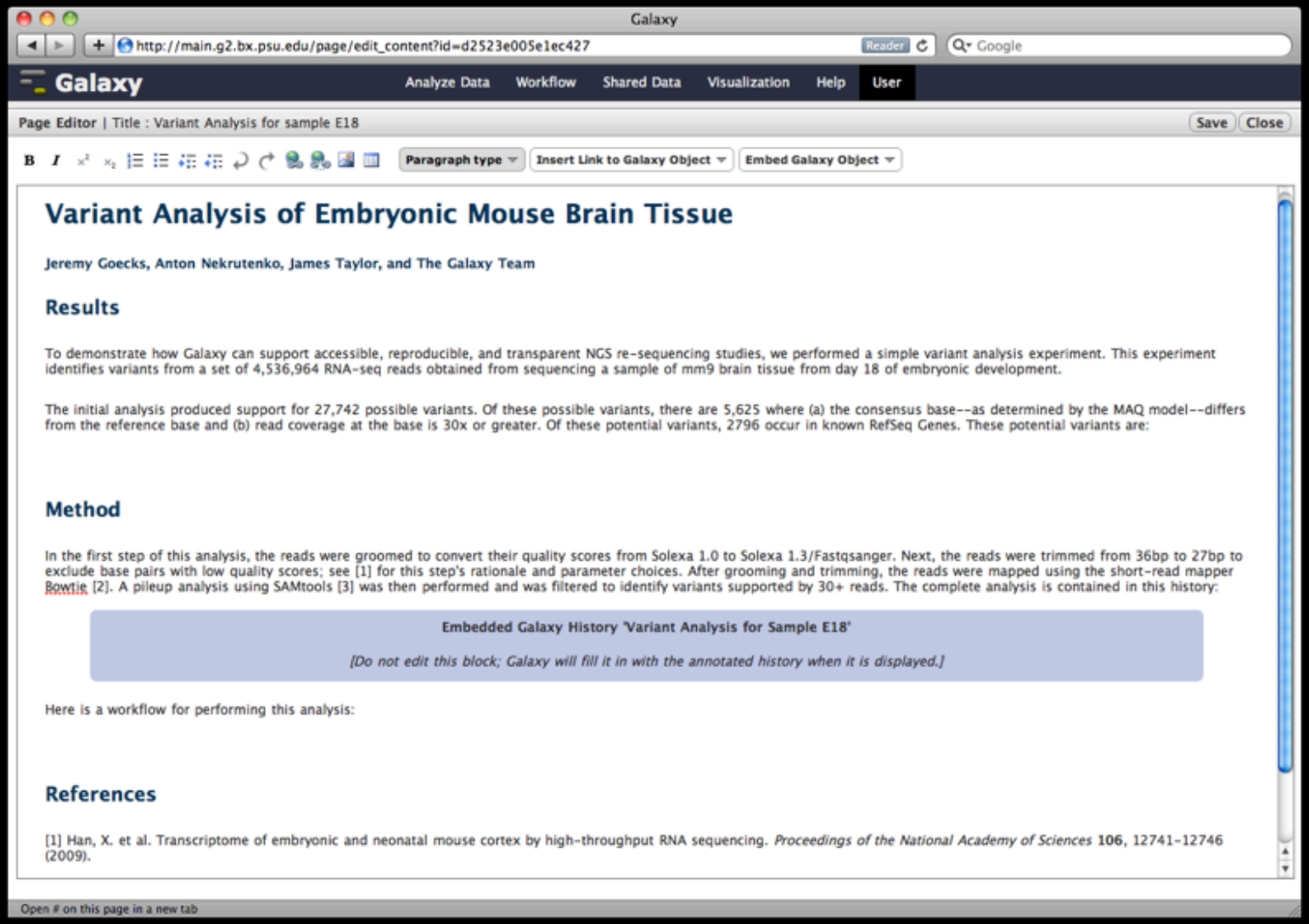
Transparency

Galaxy Pages



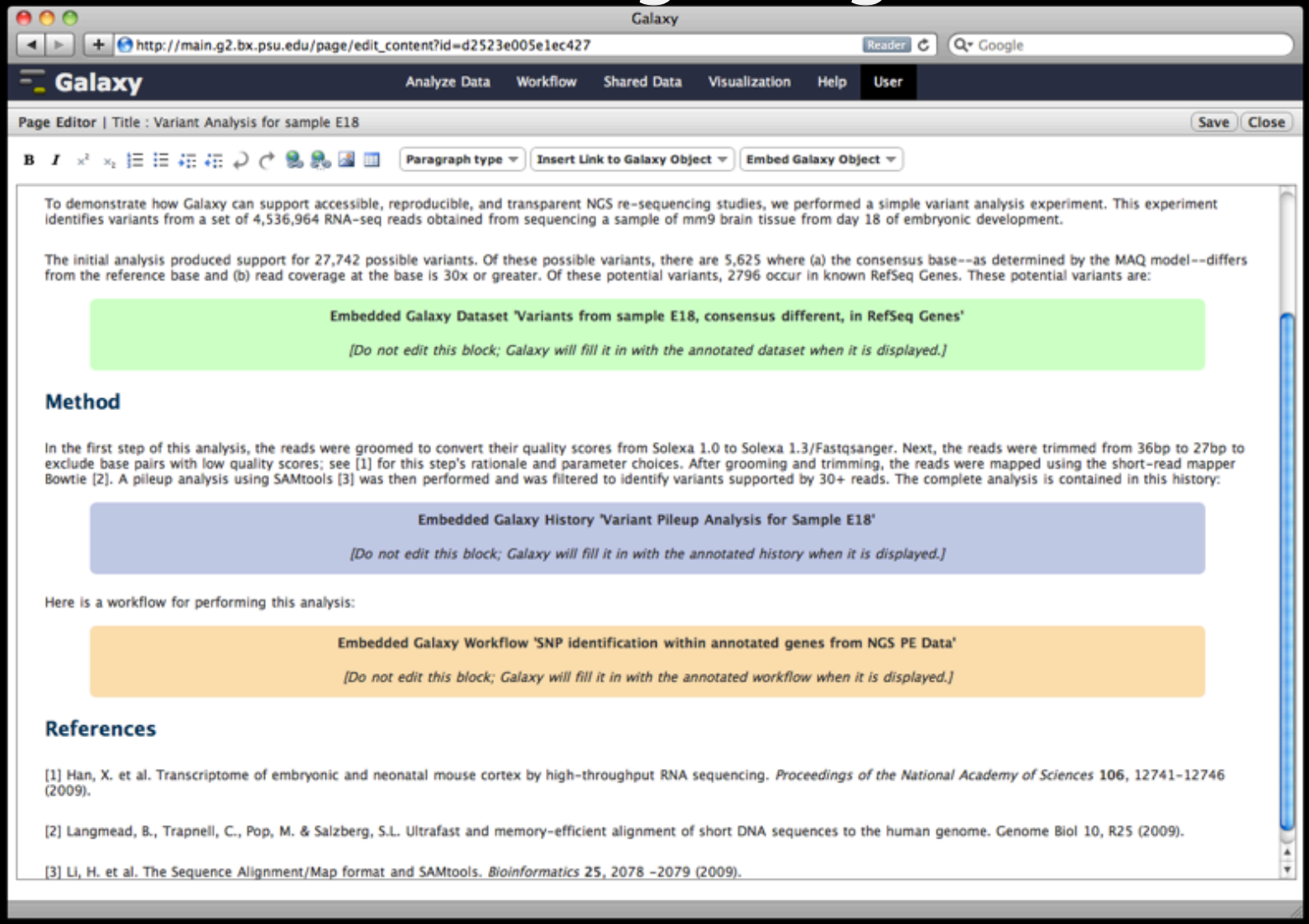
Transparency

Creating a Page



Transparency

Creating a Page



AI Opportunity: "Now What?"

What can I do with this dataset?

When should I use this tool?

What should be the next step in my analysis?

What are the "best practice" workflows for my analysis?

“Now What?” Factors

Past work

- what have I already done? (personal history)
- what have other people already done? (community history)

Approach

- exploration vs. focused analysis
- Google approach vs. scientist approach?

Overview

Genomics

Galaxy

- accessible, reproducible, and transparent science
- on the cloud
- visual analytics

Reflections on Galaxy

Three Ways to Use Galaxy

1. Download and Run Locally

2. Public Website (http://usegalaxy.org)

3. Run on the Cloud

1. Download and Run Locally

No configuration needed but everything can be configured

Prominent local installations at:

- Cold Spring Harbor Lab
- Dept. of Energy's Joint Genome Institute
- Harvard School of Public Health
- U of Texas System
- U of Oklahoma
- Netherlands Bioinformatics Center
- Oxford College

Requires computing resources, technical expertise, and maintenance

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

Public Website, 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

3. Galaxy on the Cloud

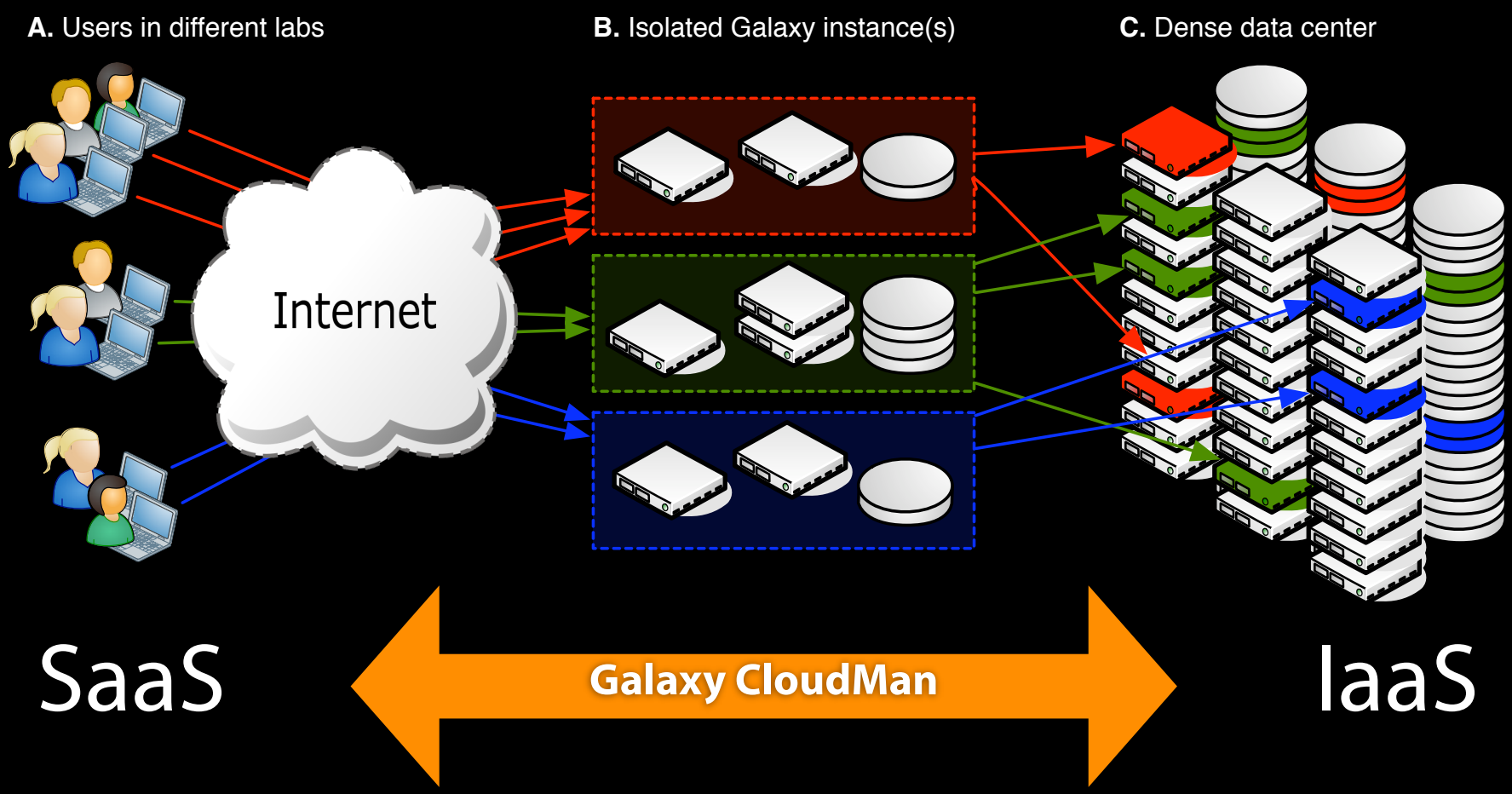
For extended or particular resource needs

- customization necessary
- oscillating data volume

Limited informatics expertise or infrastructure

Data production and (no?) sharing

The big picture



Galaxy CloudMan

Complete solution for instantiating, running and scaling cloud resources with automatically configured Galaxy

- Tools and reference datasets exceed Galaxy Main

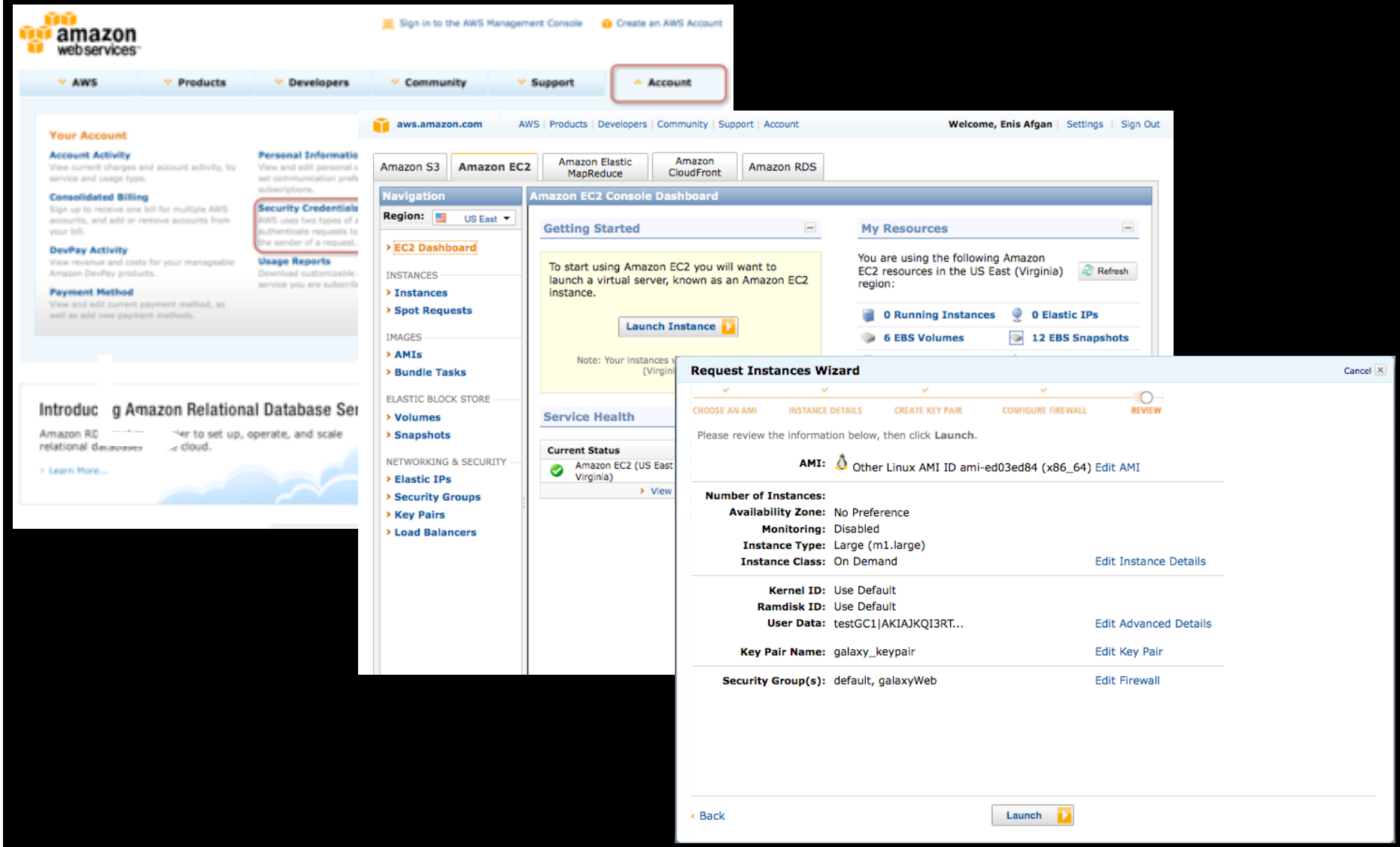
No computational expertise needed

- no configuration needed but completely configurable

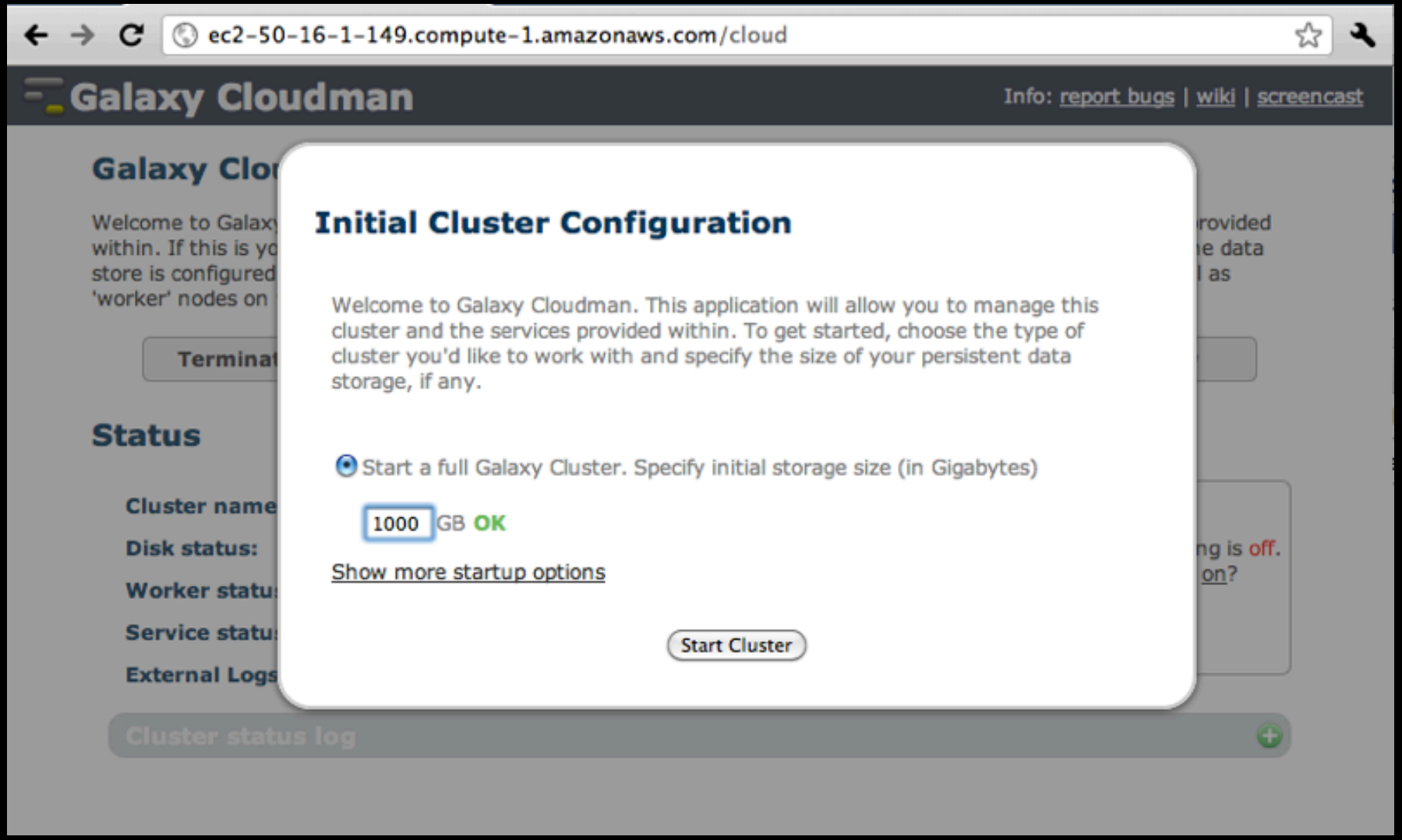
Reproducibility ensured

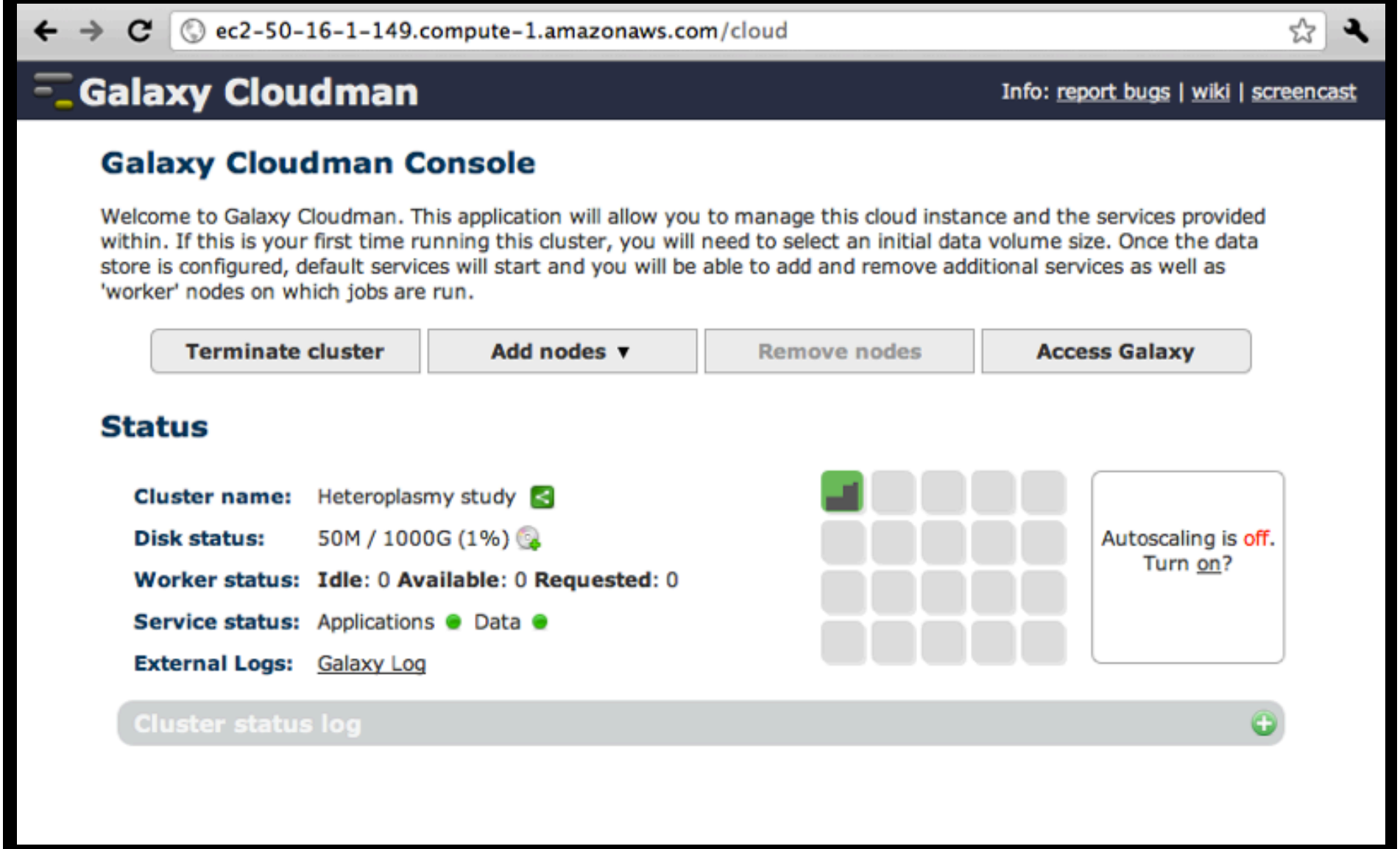
- Automated configuration for machine image, tools, and data
- Dynamic but persistent storage

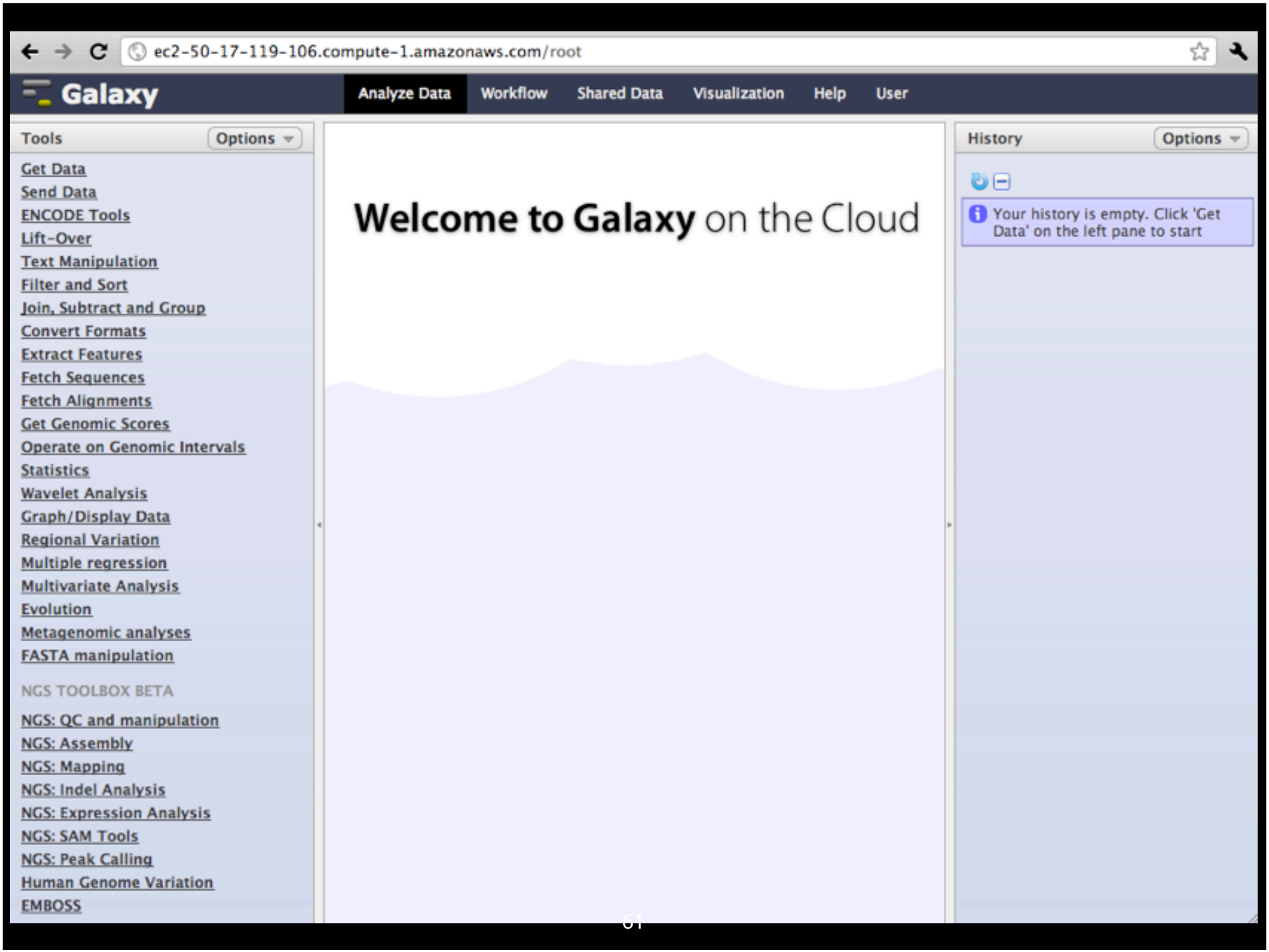
Start an Instance



Configure Your Instance




ec2-50-16-1-149.compute-1.amazonaws.com/cloud
Galaxy Cloudman
Info: report bugs | wiki | screencast
Galaxy Cloudman Console
Welcome to Galaxy Cloudman. This application will allow you to manage this cloud instance 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 able to add and remove additional services as well as 'worker' nodes on which jobs are run.
Terminate cluster
Add nodes ▾
Remove nodes
Access Galaxy
Status
Cluster name: Heteroplasmy study
Disk status: 50M / 1000G (1%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications Data
External Logs: Galaxy Log
Autoscaling is off.
Turn on?
Cluster status log


ec2-50-17-119-106.compute-1.amazonaws.com/root
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
FASTA manipulation
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Assembly
NGS: Mapping
NGS: Indel Analysis
NGS: Expression Analysis
NGS: SAM Tools
NGS: Peak Calling
Human Genome Variation
EMBOSS
Welcome to Galaxy on the Cloud
History
Options
Your history is empty. Click 'Get Data' on the left pane to start

The importance of sharing

Share entire Galaxy CloudMan cluster instances

- Fully automated solution

Publish an analysis

- In progress or otherwise

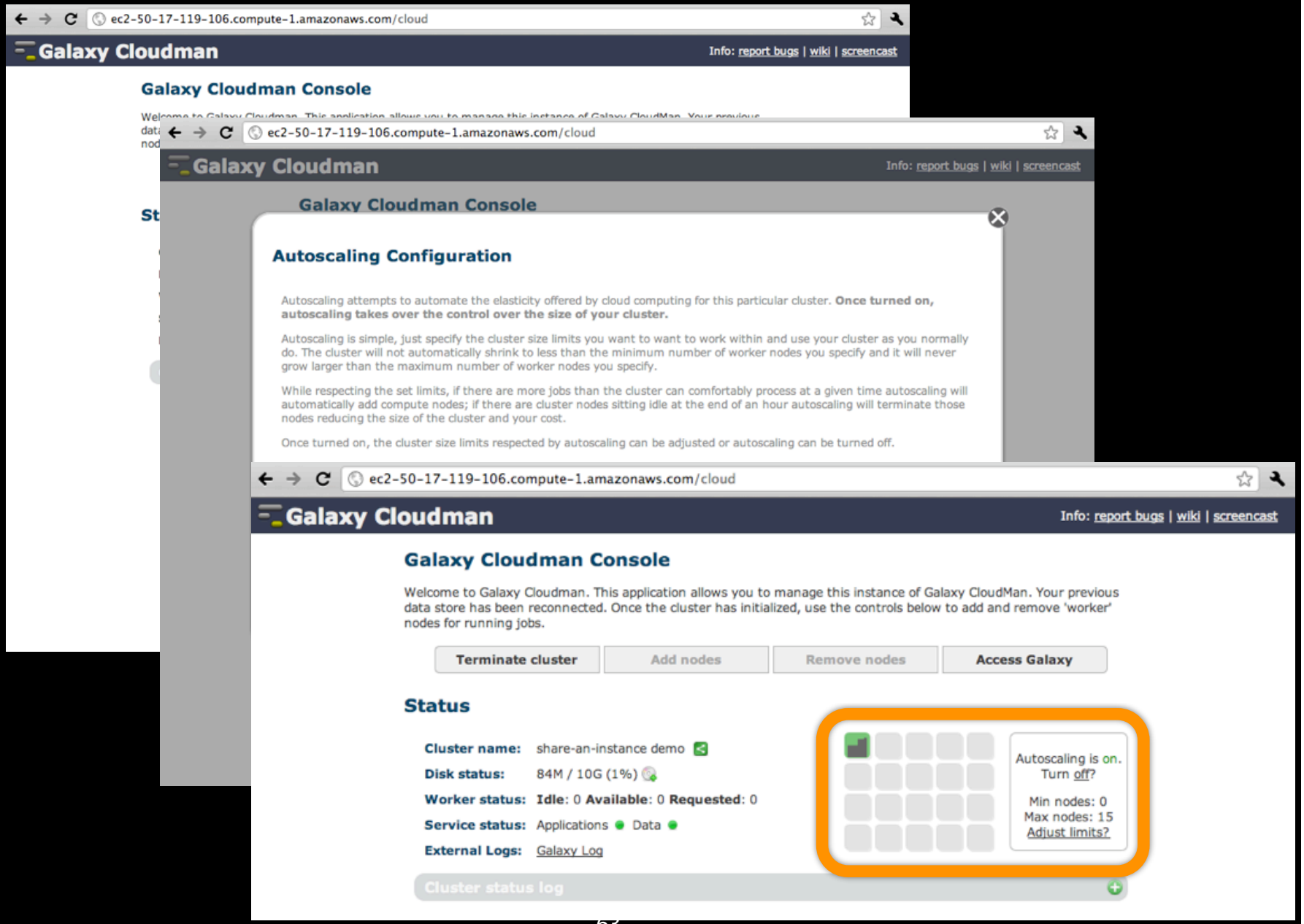
Use CloudMan as PaaS

- Deploy your own tool and make it available

Snapshot your instance

- Data
- Configuration

Scaling the infrastructure with the computation



Exercising elasticity

Fixed cluster size

5 nodes

Computation time: 9 hrs

Computation cost: $20

20 nodes

Computation time: 6 hrs

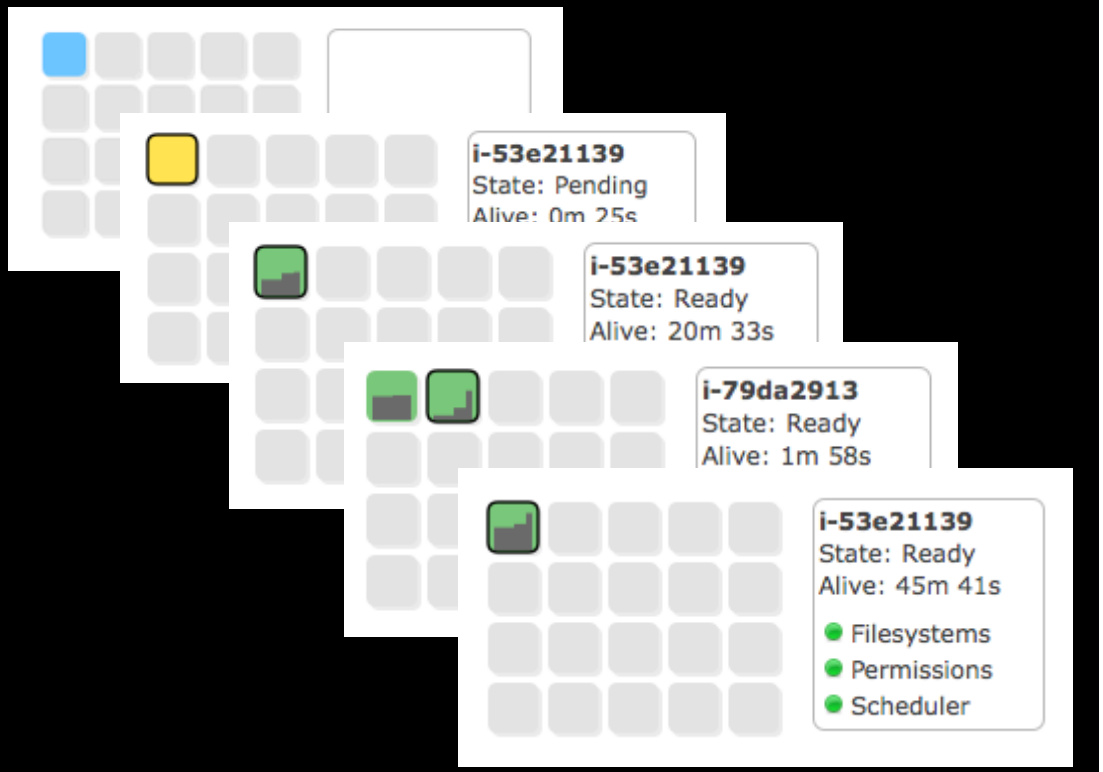
Computation cost: $50

Dynamic cluster size

1 to 16 nodes

Computation time: 6 hrs

Computation cost: $20

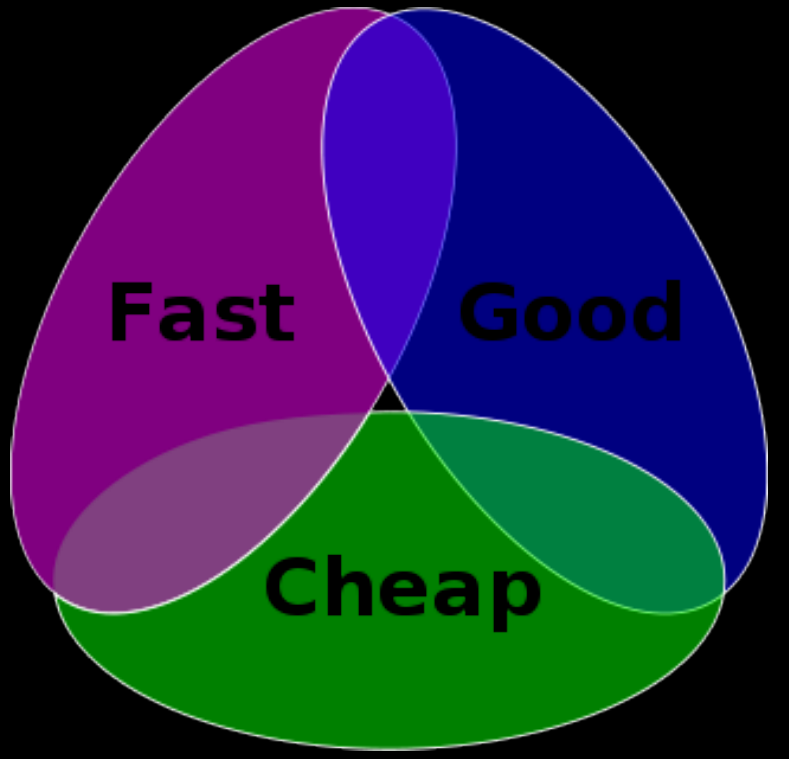


AI Opportunity: Smart Resource Usage

How long will this tool run?

How much will it cost?

How much computing power do I need for this analysis?



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

User Support vs. Automated

Users like (or require) control

- cognitive models of dynamic computing: parallelization and autoscaling?

Automation requires tool profiling

- local vs. global, parallelization
- linear vs. non-linear (e.g. graphs)

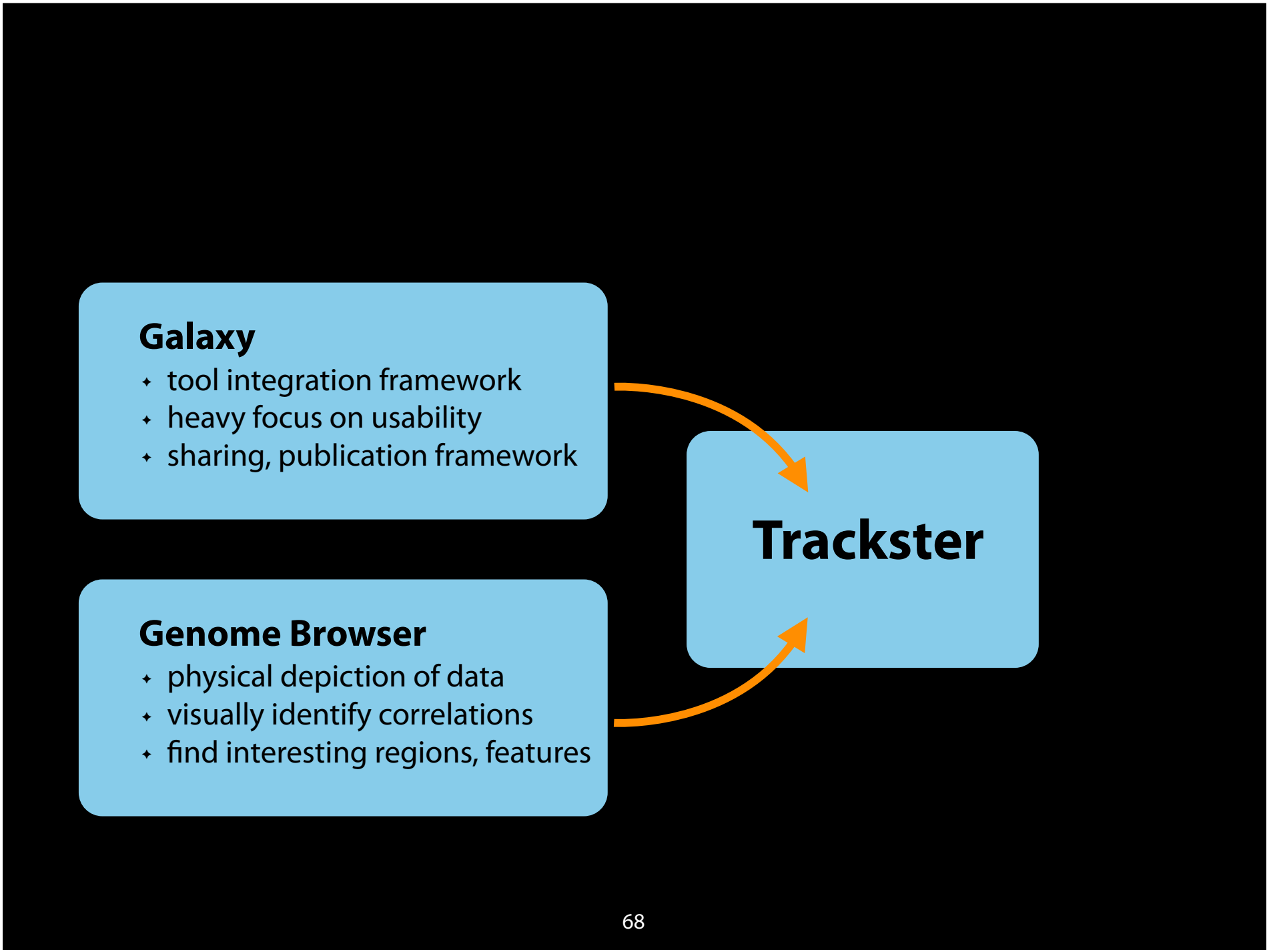
Overview

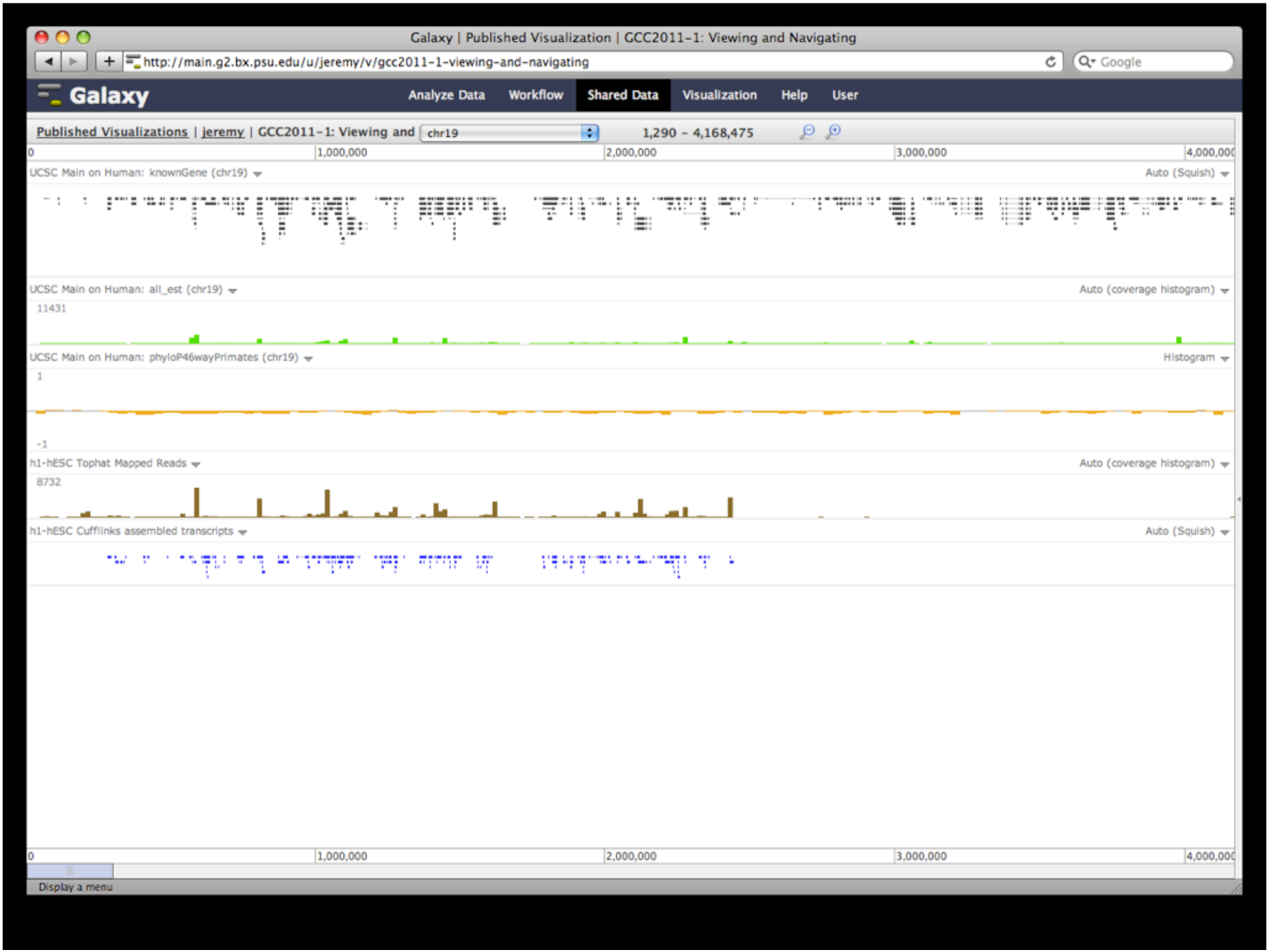
Genomics

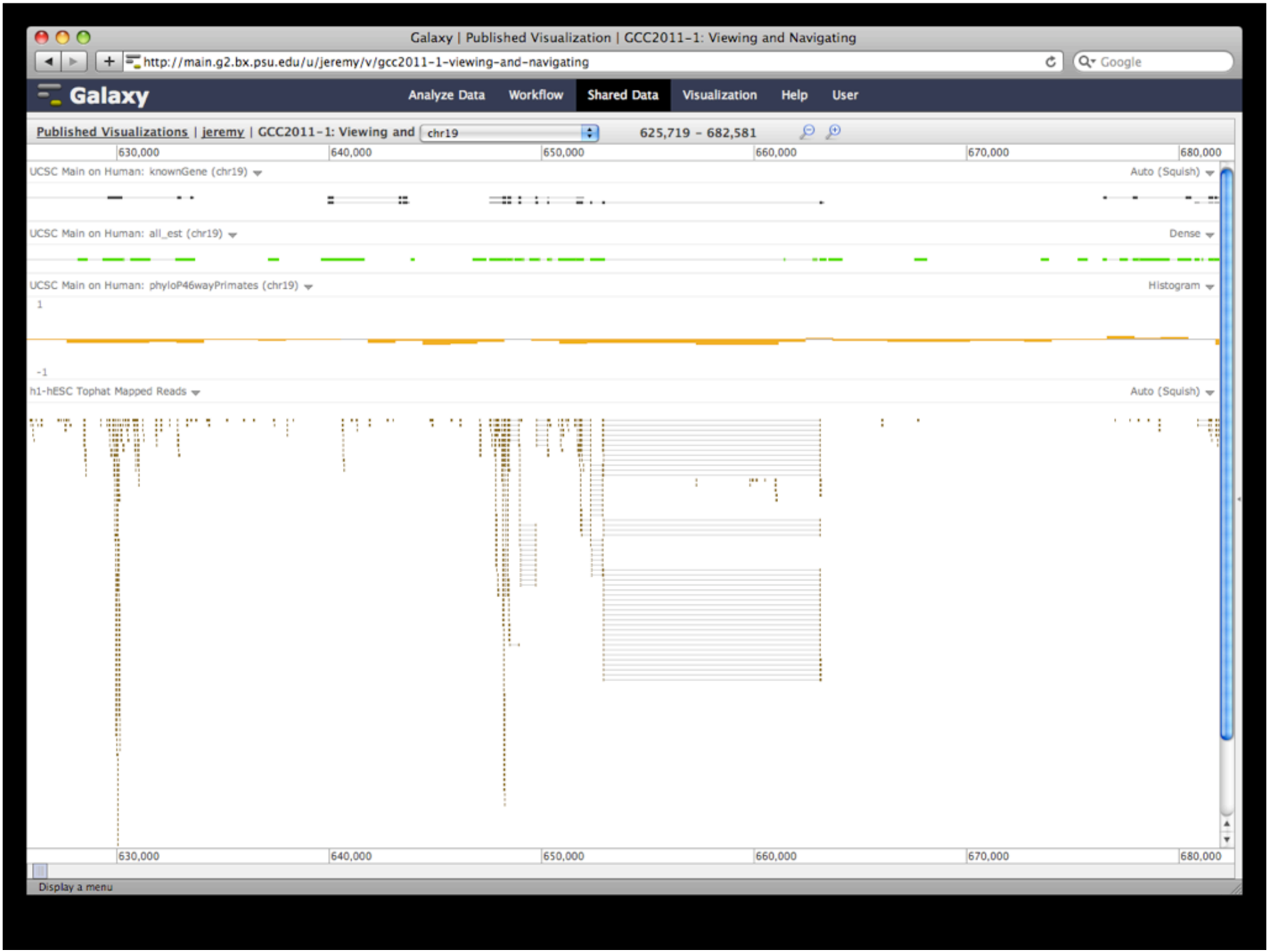
Galaxy

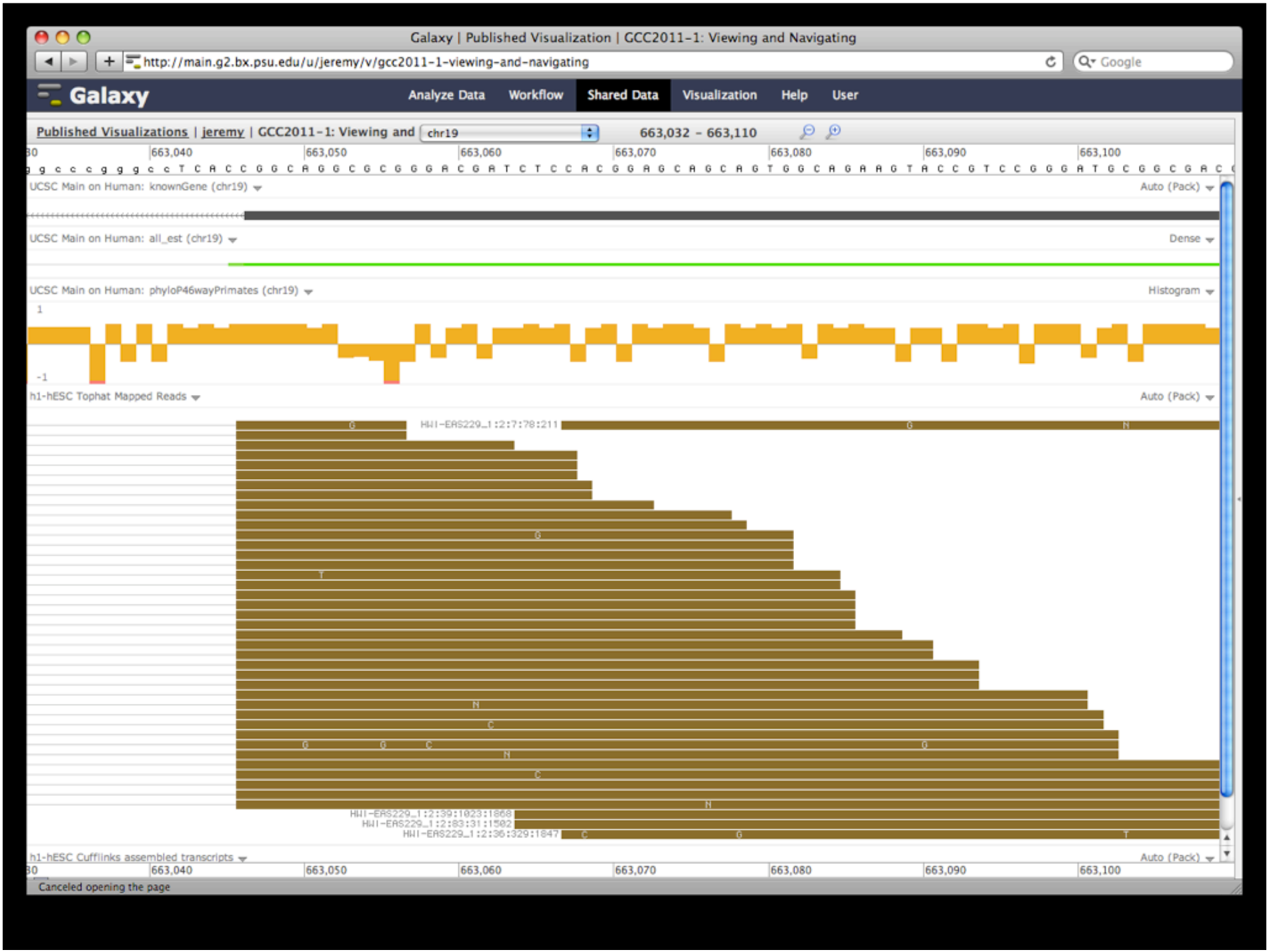
- accessible, reproducible, and transparent science
- on the cloud
- visual analytics

Reflections on Galaxy


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


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
4,000,000
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
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-1: Viewing and
chr19
663,032 - 663,110
663,040
663,050
663,060
663,070
663,080
663,090
663,100
UCSC Main on Human: knownGene (chr19)
Auto (Pack)
UCSC Main on Human: all_est (chr19)
Dense
UCSC Main on Human: phyloP46wayPrimates (chr19)
Histogram
1
-1
h1-hESC Tophat Mapped Reads
Auto (Pack)
HWI-EAS229_1:2:7:78:211
HWI-EAS229_1:2:39:1023:1868
HWI-EAS229_1:2:83:31:1502
HWI-EAS229_1:2:36:329:1847
h1-hESC Cufflinks assembled transcripts
Auto (Pack)
Canceled opening the page

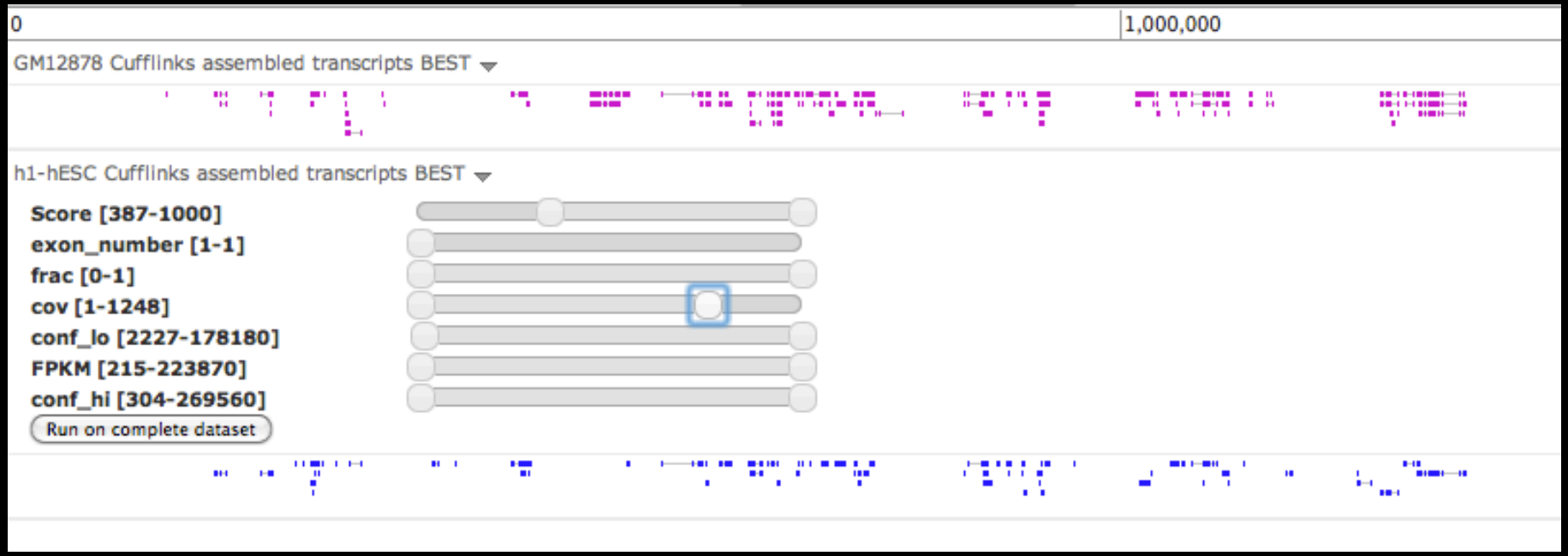
HTS Analysis Challenges

Complex tools, parameter dependent

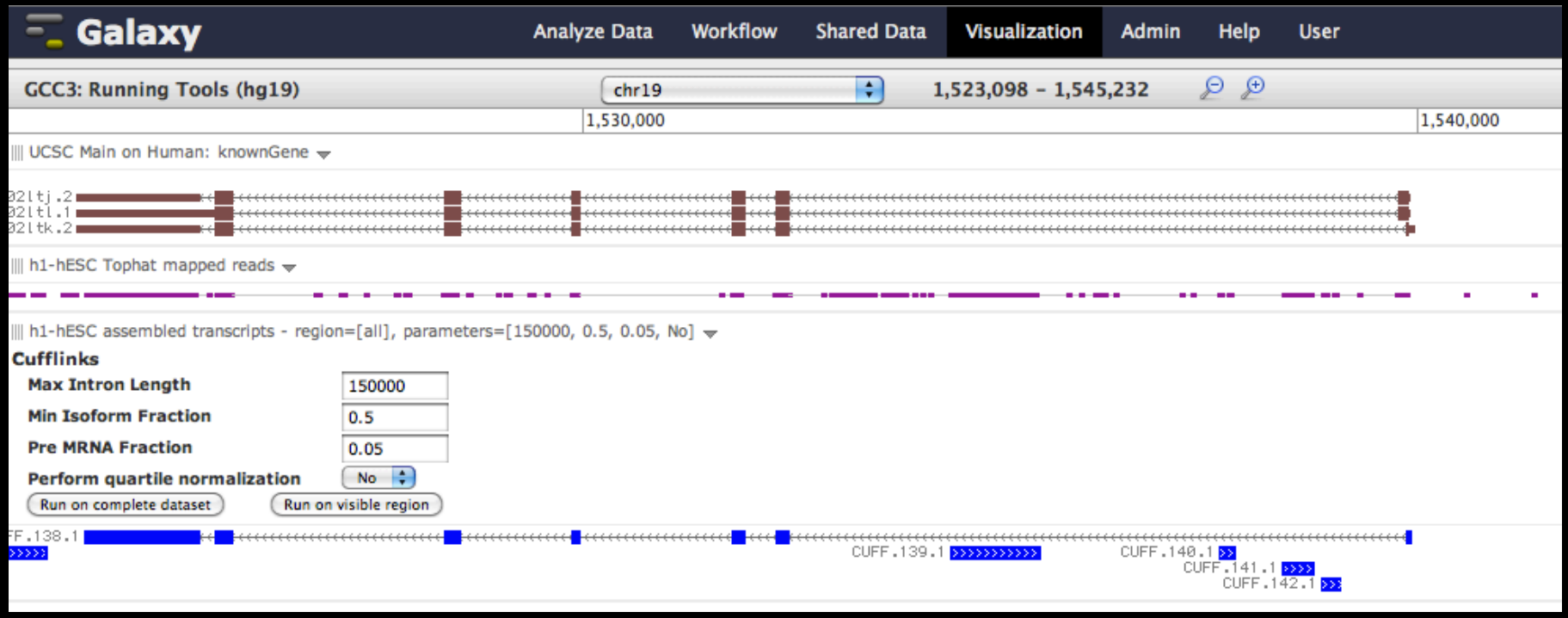
Analysis and visualization not integrated

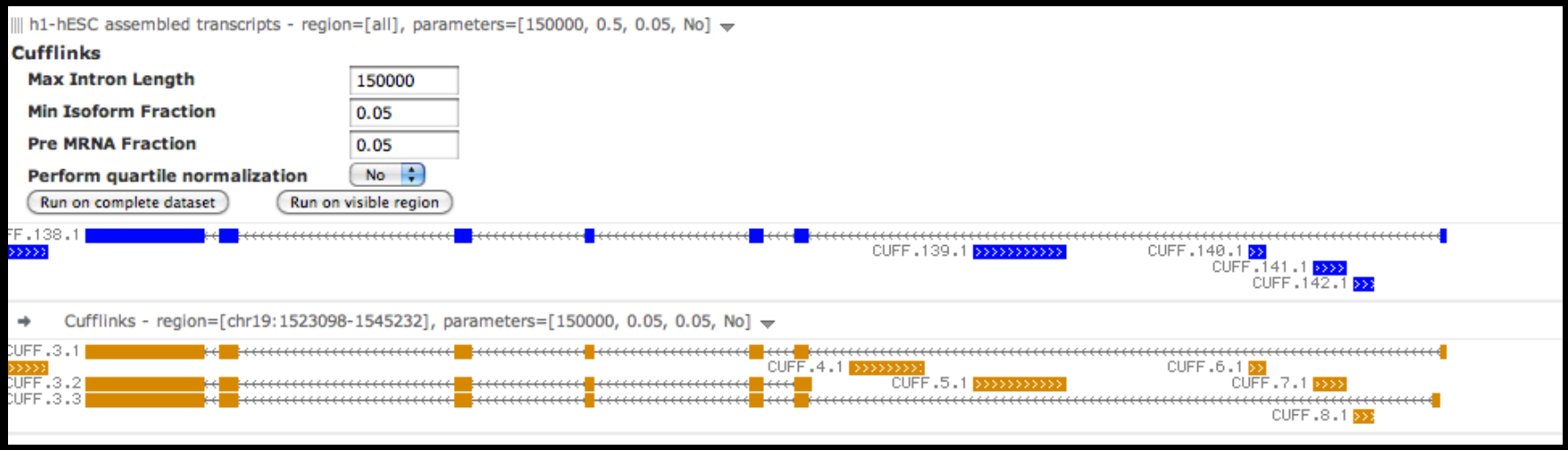
Want to be able to experiment

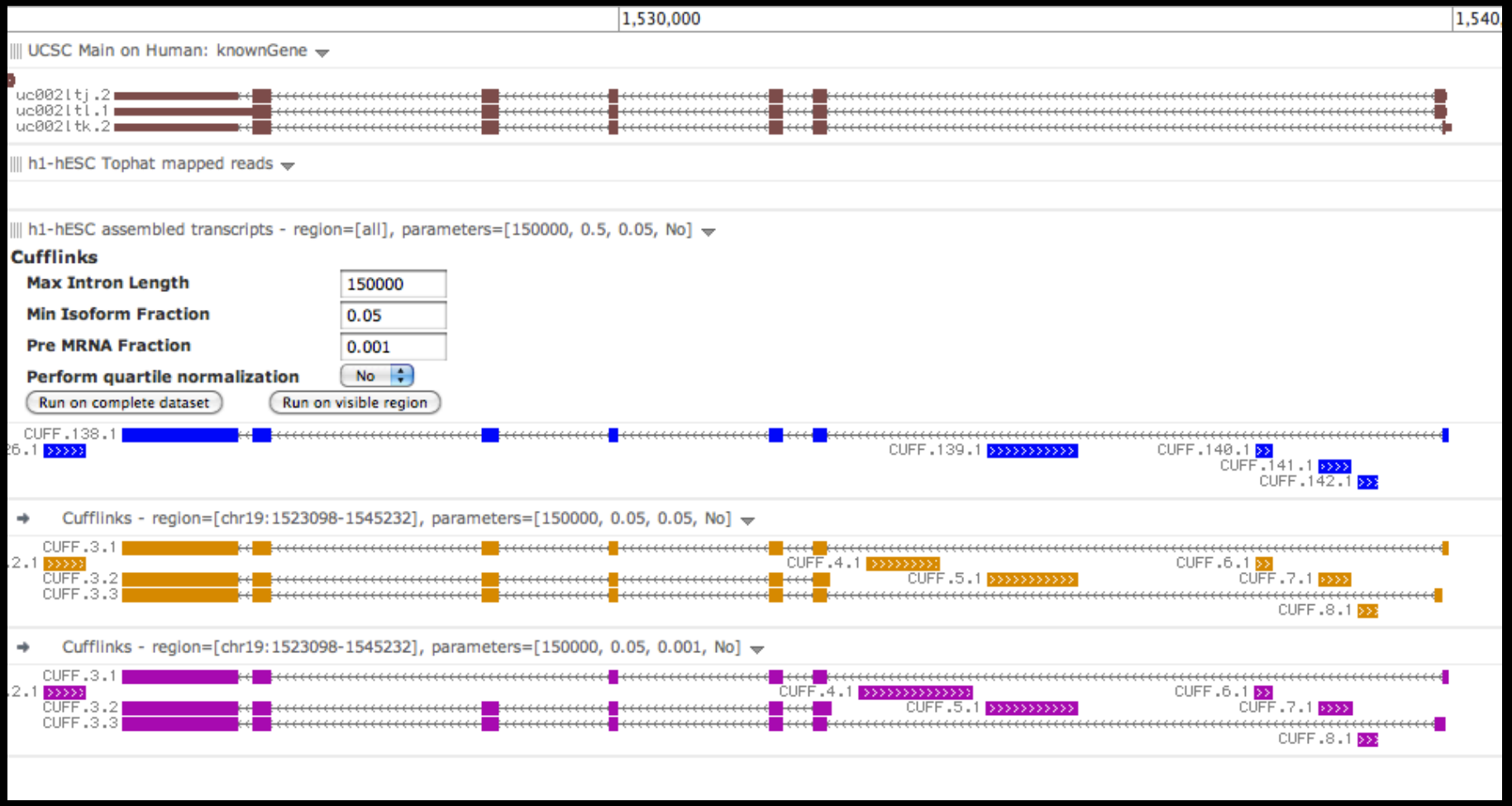
Dynamic Filtering



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
FF.138.1
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.3.1
CUFF.3.2
CUFF.3.3
CUFF.4.1
CUFF.5.1
CUFF.6.1
CUFF.7.1
CUFF.8.1


1,530,000
1,540
UCSC Main on Human: knownGene
uc002ltj.2
uc002ltl.1
uc002ltk.2
h1-hESC Tophat mapped reads
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
0.001
Perform quartile normalization
No
Run on complete dataset
Run on visible region
CUFF.138.1
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.3.1
CUFF.3.2
CUFF.3.3
CUFF.4.1
CUFF.5.1
CUFF.6.1
CUFF.7.1
CUFF.8.1
Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.001, No]
CUFF.3.1
CUFF.3.2
CUFF.3.3
CUFF.4.1
CUFF.5.1
CUFF.6.1
CUFF.7.1
CUFF.8.1

Experimentation in Trackster

Tools integrated in visualization environment

Dynamically filter:

- visually identify features that match ranges
- on whole dataset

Run tools:

- quickly on visible region
- on whole dataset

AI Opportunity:
Smart Visual Analytics

What datasets should I include in my visualization?

What are the interesting areas of my visualization?

What tools should I use in my visualization?

Combining Models

User modeling

- interests + past activities + community activities

Data modeling

- relationships amongst datasets

"Visual acuity" modeling: what can and should be seen in a visualization

Overview

Genomics

Galaxy

- accessible, reproducible, and transparent science
- on the cloud
- visual analytics

Reflections on Galaxy

Accessibility

Accessibility drives usage

- transparency and esp. reproducibility are secondary
- few incentives for reproducible research (for now)

A win-win for everyone

- tool developers: free GUI, more exposure
- users: easy to run tools, consistent interface
- everyone: amplification b/c tools can be chained to create complex analyses

Approach

Computer science research is driven by scientific needs

- we do biology research ourselves with Galaxy
- publish largely in biology journals

Don't need do everything, but what is done is done well

- software engineering, community support are priorities

Galaxy used in and cited by papers in *Science, Nature*, *Genome Research*, *Genome Biology,* and more

Challenges Going Forward

Promoting community involvement

- tools, assays, analyses growing too fast for us alone
- facilitate community contributions and usage of contributions

Scaling to many, many Galaxies

- moving objects between Galaxies while ensuring reproducibility
- enabling users to find useful "stuff"

Novel application areas

- genomics ideal application area -- what next?

Thanks! Questions, discussion?

https://bitbucket.org/galaxy/galaxy-central/wiki/Citations

Public server: http://usegalaxy.org

Download and run: http://getgalaxy.org

On the cloud: http://usegalaxy.org/cloud

jeremy.goecks@emory.edu