Workflows (and more!) with Galaxy

Expression Dynamics of Human XBP1 Revealed using an Integrative Approach

Jeremy Goecks
The Galaxy Team & Emory University

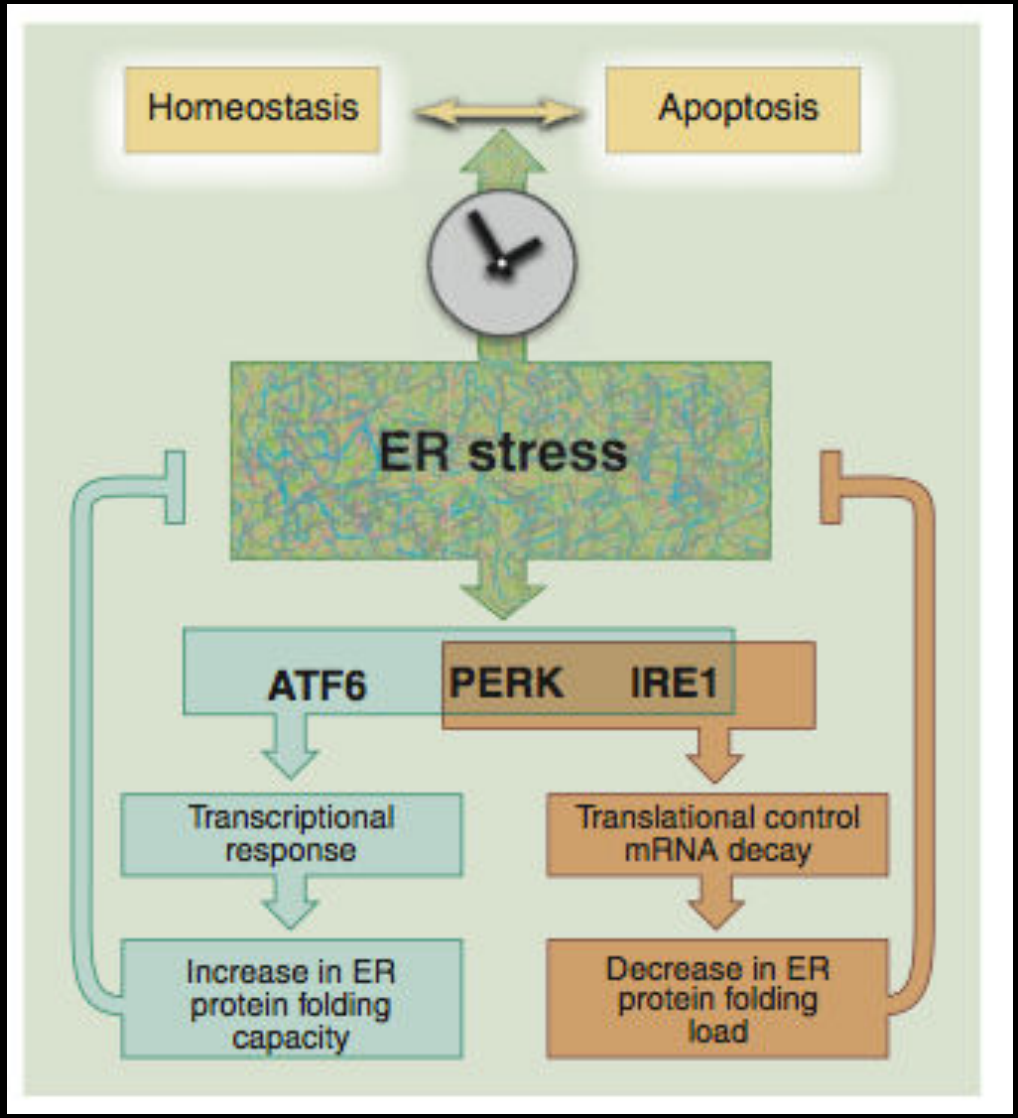
Approach

Describe how Galaxy was used to complete an NGS biological analysis, and highlight Galaxy features along the way

The Investigation

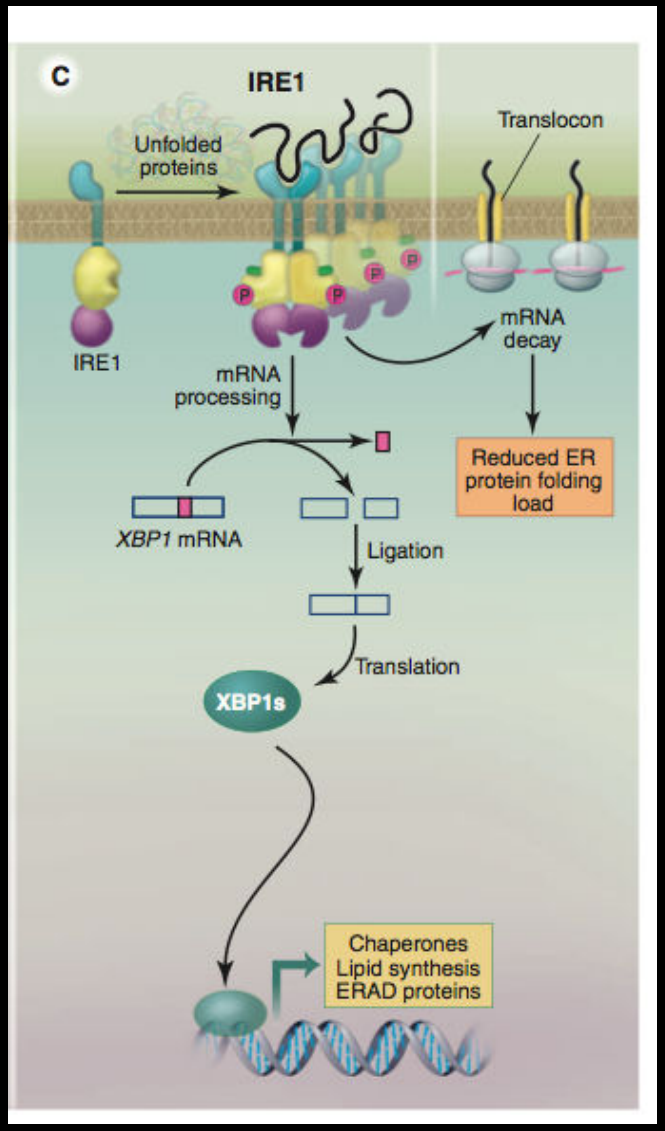
Understanding human *XBP1* expression dynamics

Unfolded Protein Response



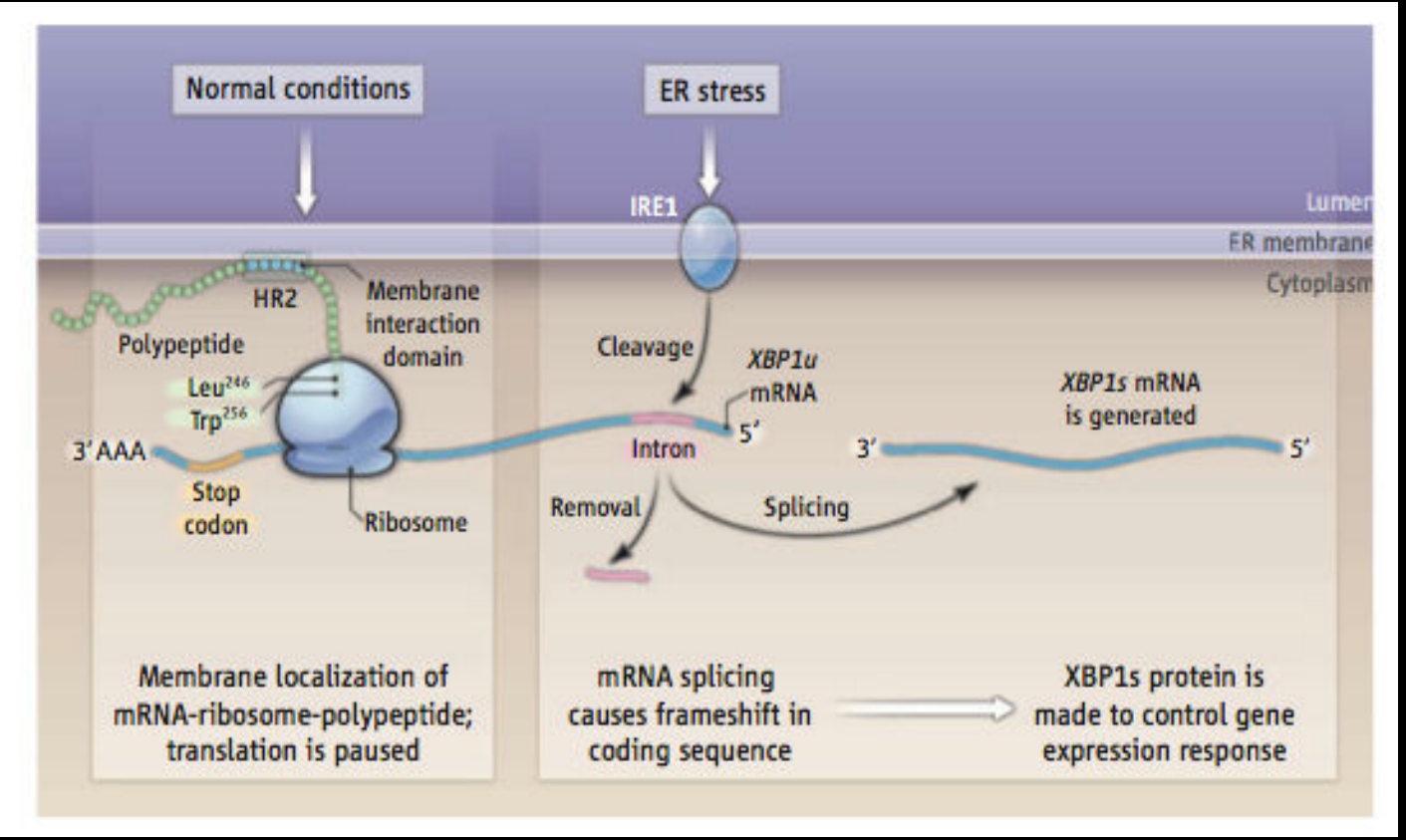
Walter & Ron, *Science* 2011

IRE1 Pathway



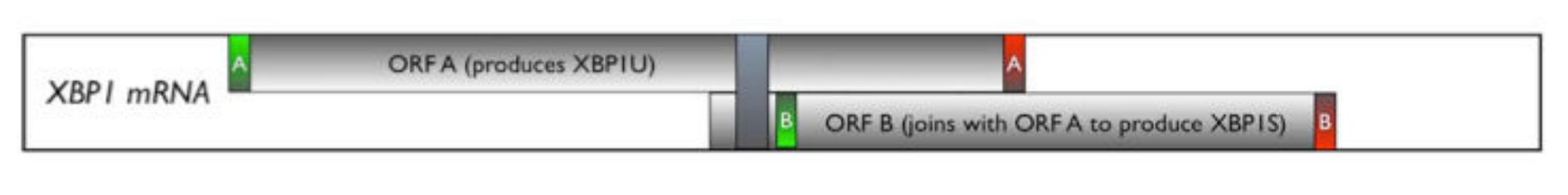
Walter & Ron, *Science* 2011

Translational Pausing



Ron & Ito, *Science* 2011

XBP1 Isoforms



very similar isoforms

frameshift between isoforms

Chung et al., *PLoS CB* 2007

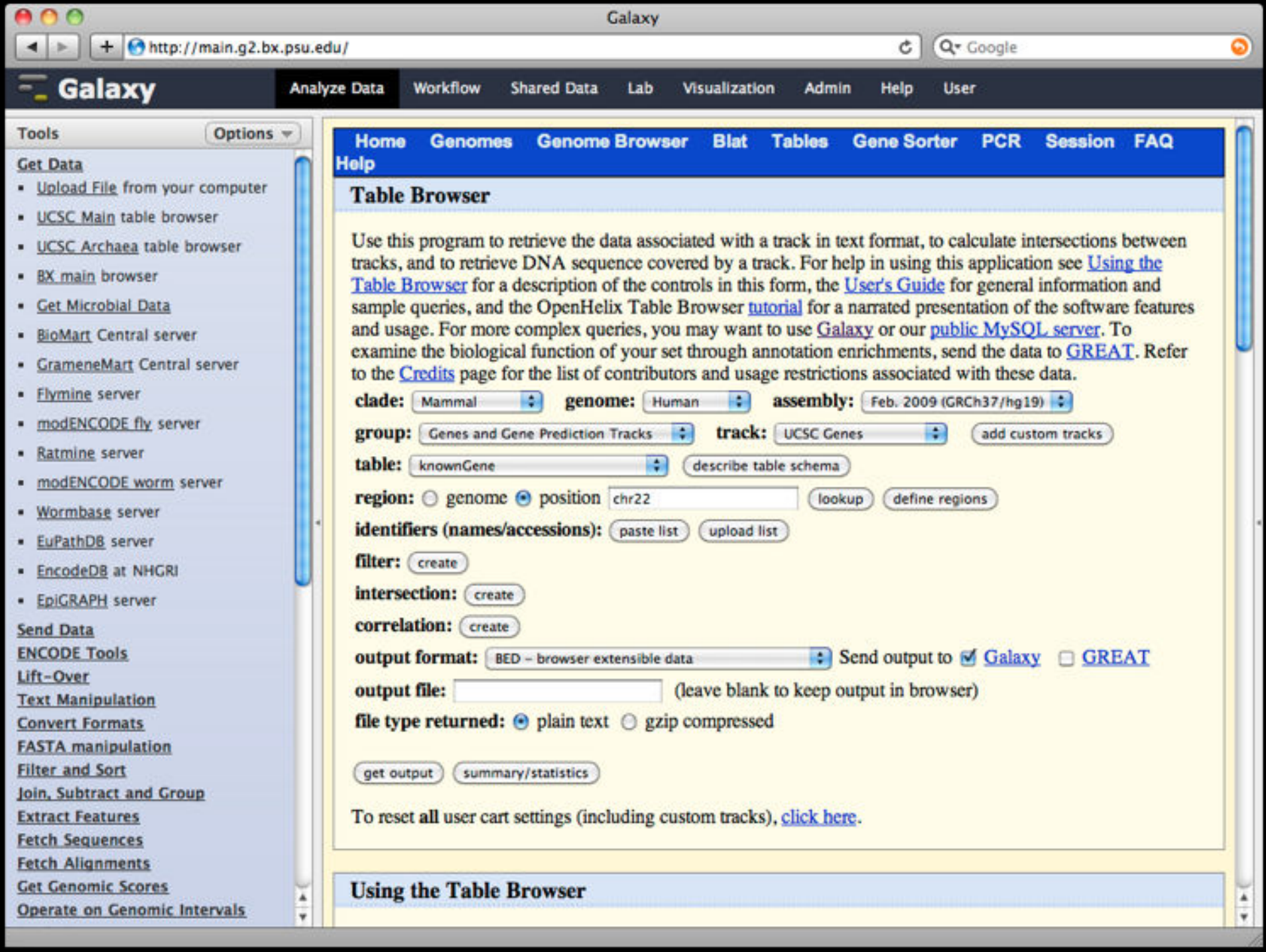
Questions

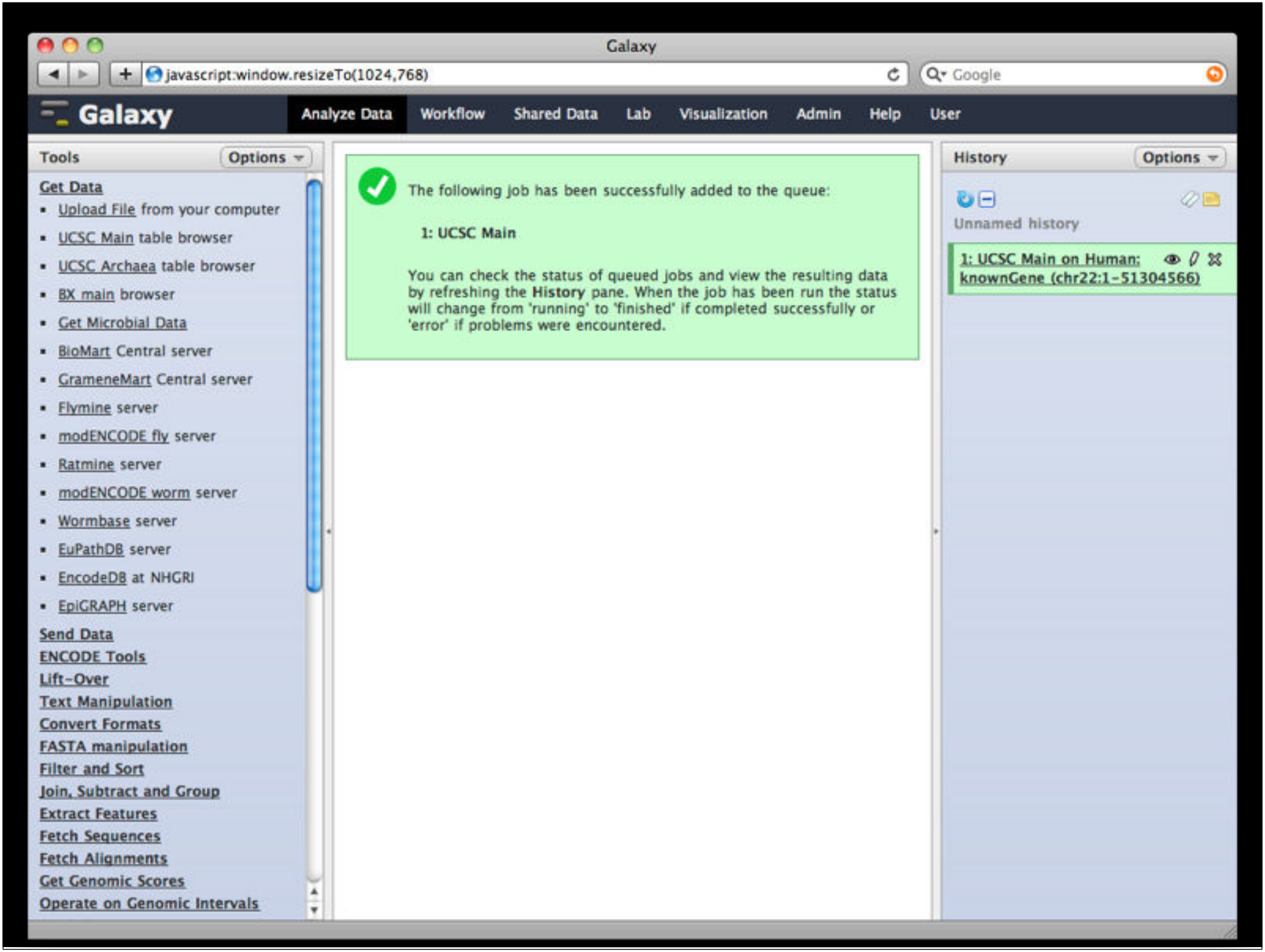
Can current RNA-seq tools be used to identify and quantify *XBP1's* highly similar isoforms?

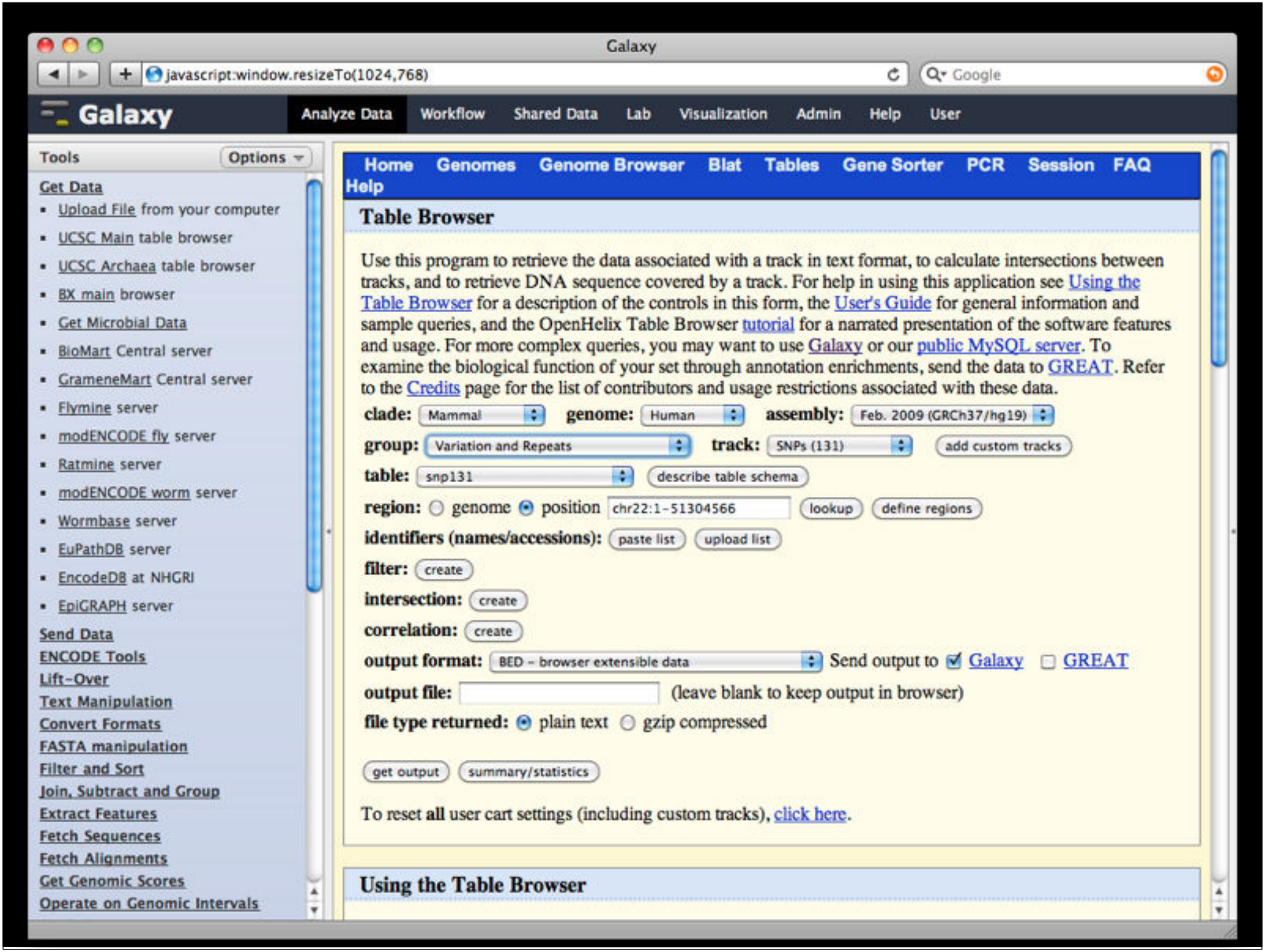
Can *XBP1's* translational pause point be identified?

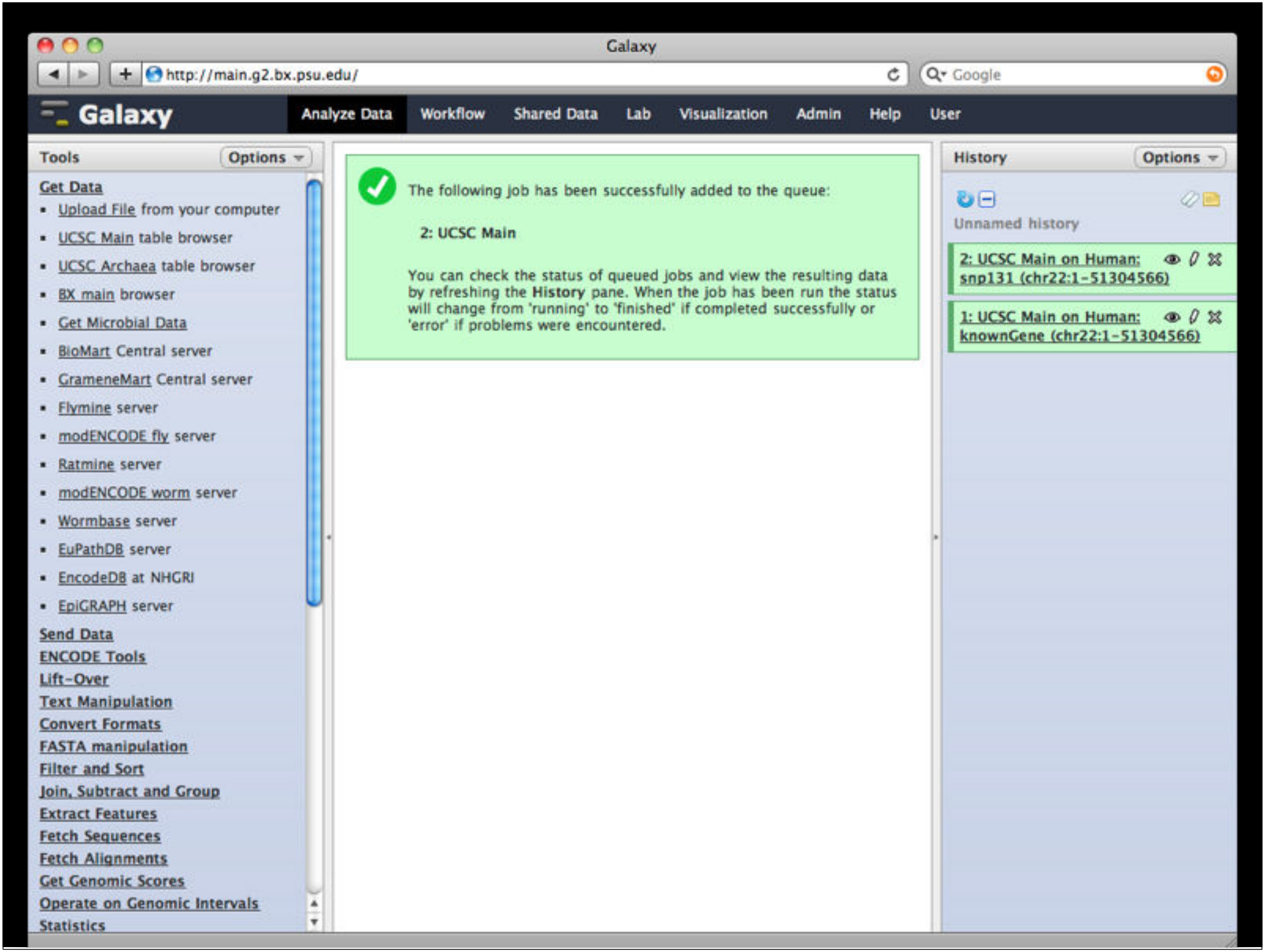
Galaxy 101
(the *really* short version)

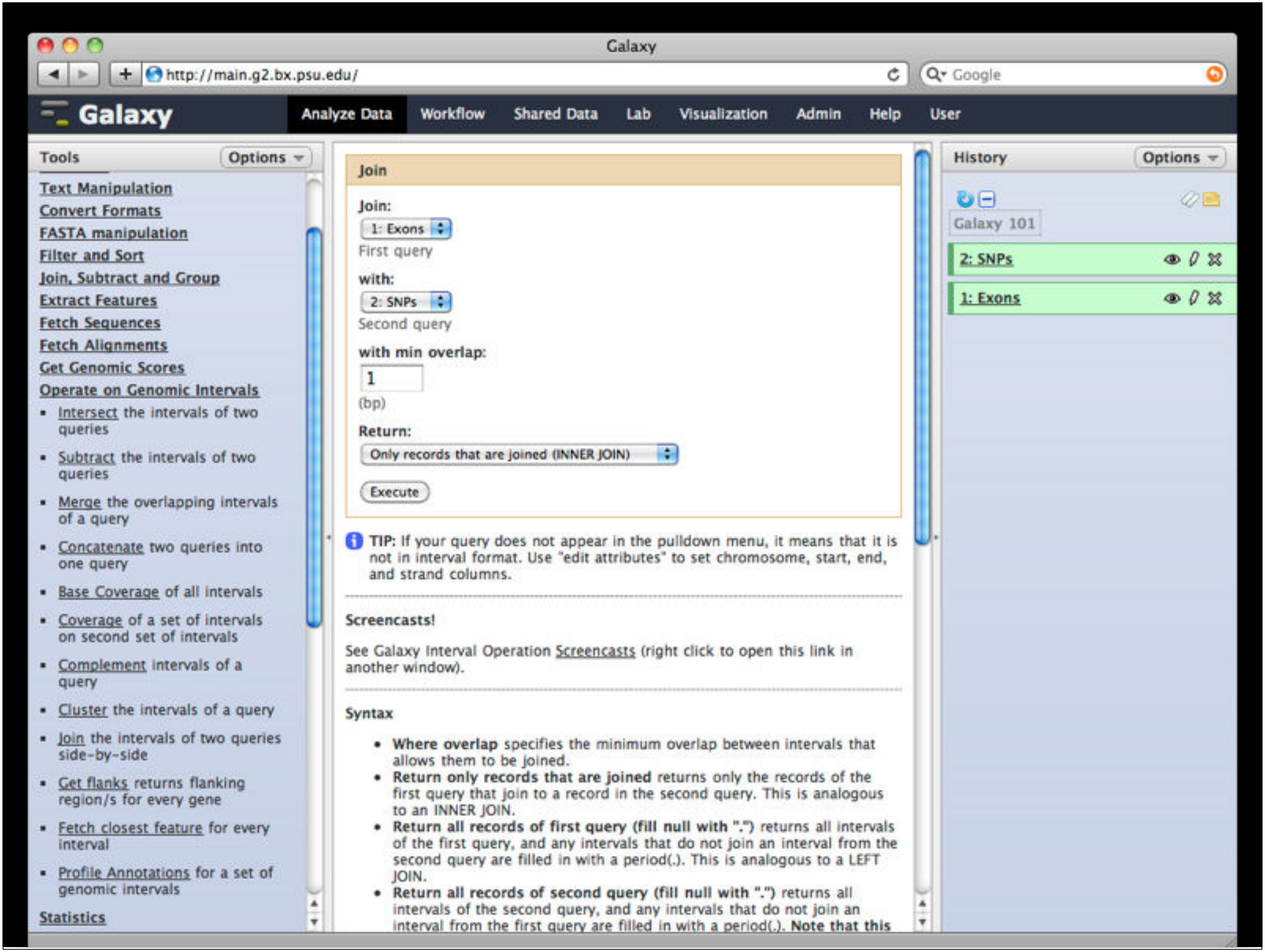
https://main.g2.bx.psu.edu/galaxy101

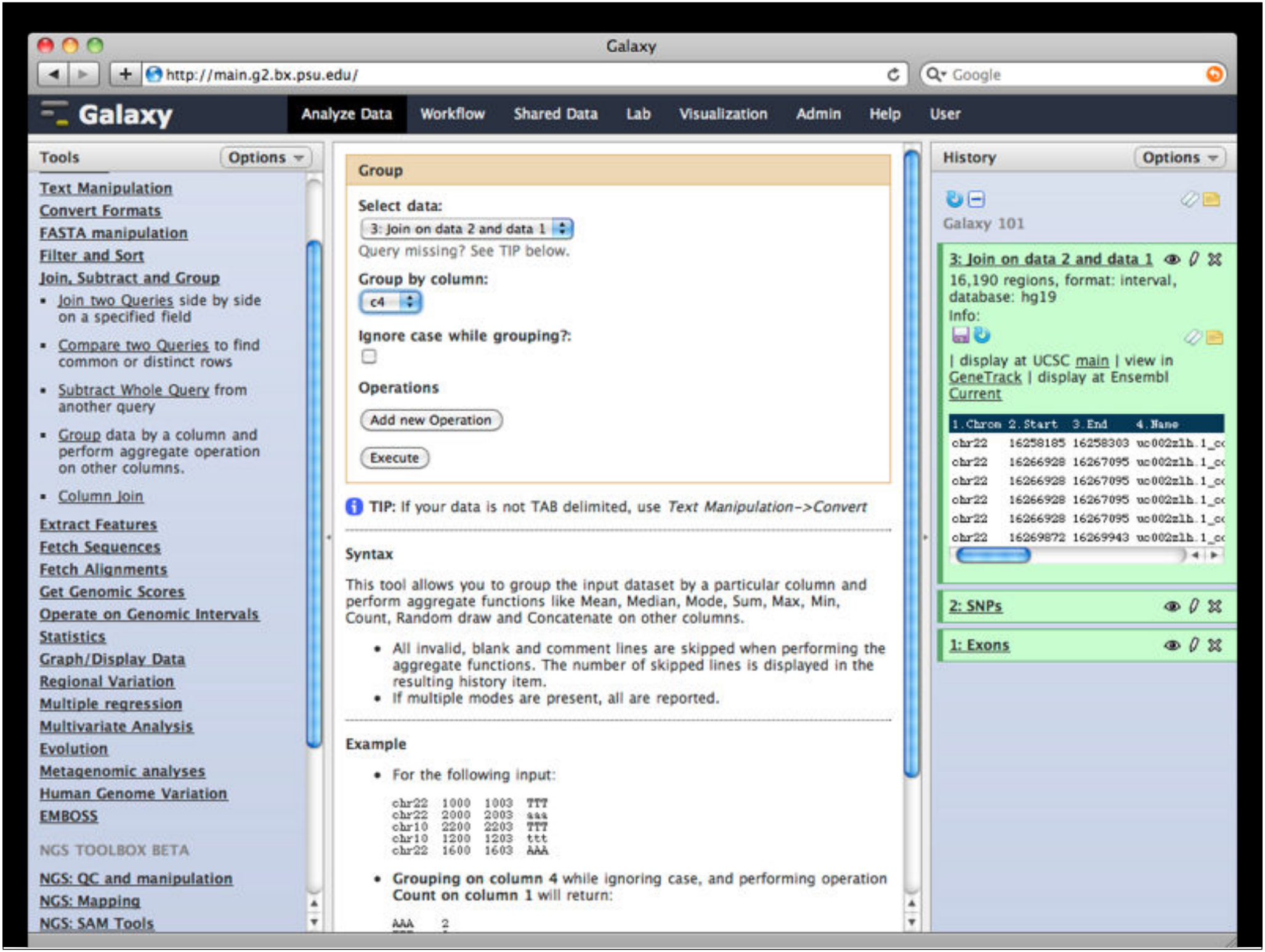

Galaxy
http://main.g2.bx.psu.edu/
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Get Data
Upload File from your computer
UCSC Main table browser
UCSC Archaea table browser
BX main browser
Get Microbial Data
BioMart Central server
GrameneMart Central server
Flymine server
modENCODE fly server
Ratmine server
modENCODE worm server
Wormbase server
EuPathDB server
EncodeDB at NHGRI
EpiGRAPH server
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Home
Genomes
Genome Browser
Blat
Tables
Gene Sorter
PCR
Session
FAQ
Help
Table Browser
Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.
clade: Mammal
genome: Human
assembly: Feb. 2009 (GRCh37/hg19)
group: Genes and Gene Prediction Tracks
track: UCSC Genes
add custom tracks
table: knownGene
describe table schema
region: genome position chr22
lookup
define regions
identifiers (names/accessions):
paste list
upload list
filter:
create
intersection:
create
correlation:
create
output format: BED - browser extensible data
Send output to Galaxy GREAT
output file:
(leave blank to keep output in browser)
file type returned: plain text gzip compressed
get output
summary/statistics
To reset all user cart settings (including custom tracks), click here.
Using the Table Browser

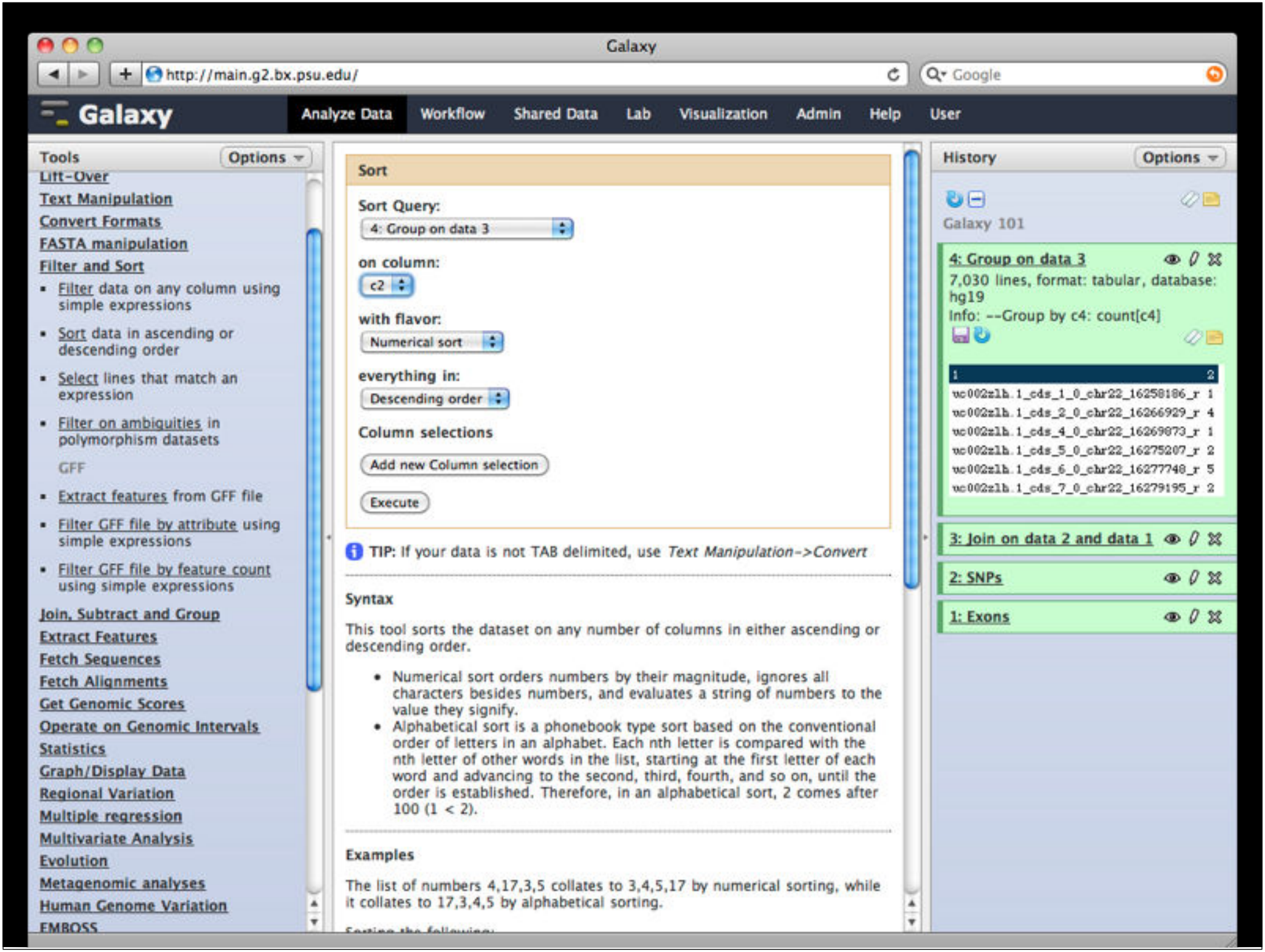

Galaxy
javascript:window.resizeTo(1024,768)
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Get Data
Upload File from your computer
UCSC Main table browser
UCSC Archaea table browser
BX main browser
Get Microbial Data
BioMart Central server
GrameneMart Central server
Flymine server
modENCODE fly server
Ratmine server
modENCODE worm server
Wormbase server
EuPathDB server
EncodeDB at NHGRI
EpiGRAPH server
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
The following job has been successfully added to the queue:
1: UCSC Main
You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
History
Options
Unnamed history
1: UCSC Main on Human: knownGene (chr22:1-51304566)

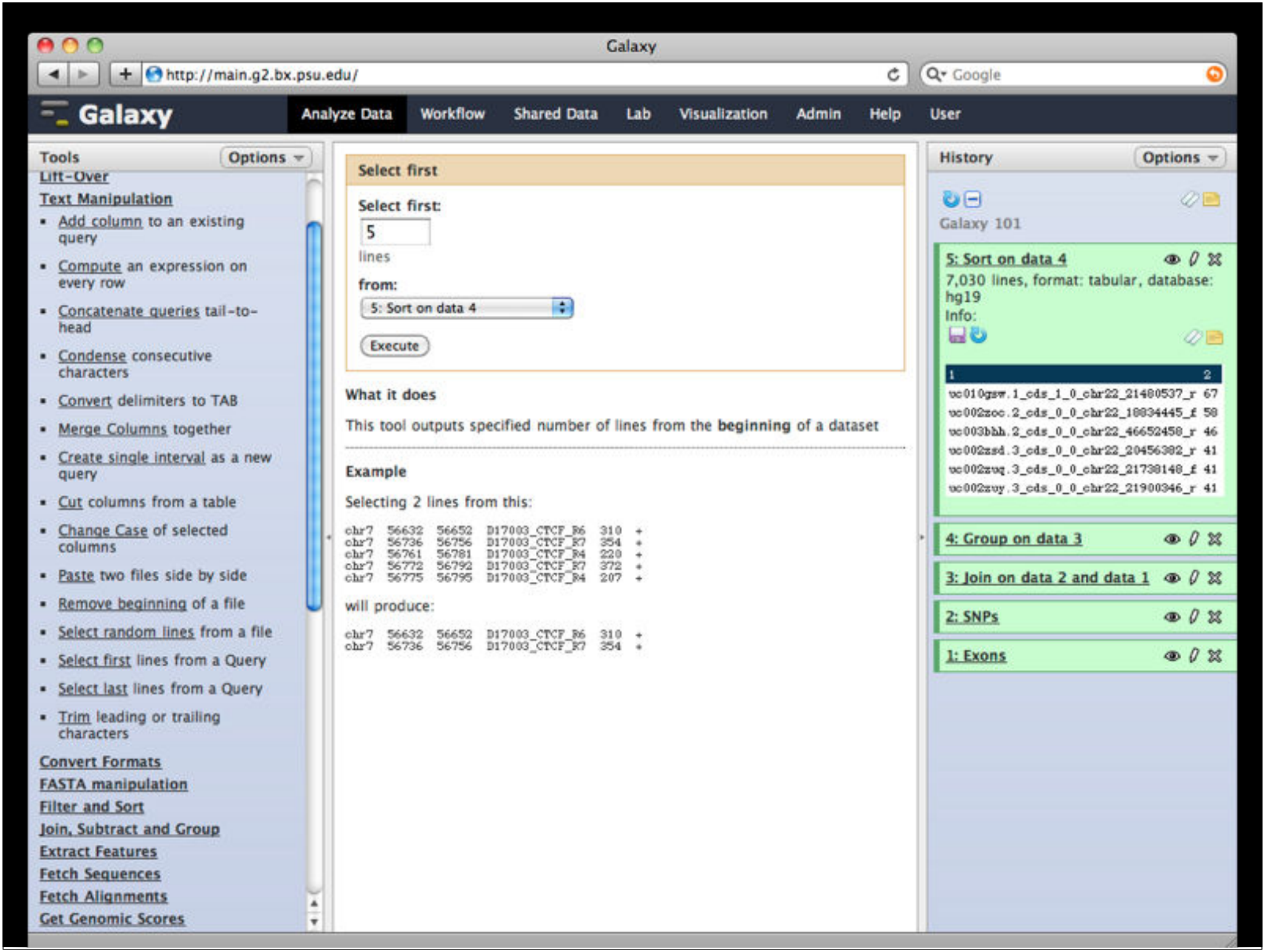

Galaxy
javascript:window.resizeTo(1024,768)
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Get Data
Upload File from your computer
UCSC Main table browser
UCSC Archaea table browser
BX main browser
Get Microbial Data
BioMart Central server
GrameneMart Central server
Flymine server
modENCODE fly server
Ratmine server
modENCODE worm server
Wormbase server
EuPathDB server
EncodeDB at NHGRI
EpiGRAPH server
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Home
Genomes
Genome Browser
Blat
Tables
Gene Sorter
PCR
Session
FAQ
Help
Table Browser
Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.
clade: Mammal
genome: Human
assembly: Feb. 2009 (GRCh37/hg19)
group: Variation and Repeats
track: SNPs (131)
add custom tracks
table: snp131
describe table schema
region: genome position chr22:1-51304566
lookup
define regions
identifiers (names/accessions):
paste list
upload list
filter:
create
intersection:
create
correlation:
create
output format: BED - browser extensible data
Send output to Galaxy GREAT
output file:
(leave blank to keep output in browser)
file type returned: plain text gzip compressed
get output
summary/statistics
To reset all user cart settings (including custom tracks), click here.
Using the Table Browser


Galaxy
http://main.g2.bx.psu.edu/
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Get Data
Upload File from your computer
UCSC Main table browser
UCSC Archaea table browser
BX main browser
Get Microbial Data
BioMart Central server
GrameneMart Central server
Flymine server
modENCODE fly server
Ratmine server
modENCODE worm server
Wormbase server
EuPathDB server
EncodeDB at NHGRI
EpiGRAPH server
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
The following job has been successfully added to the queue:
2: UCSC Main
You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
History
Options
Unnamed history
2: UCSC Main on Human: snp131 (chr22:1-51304566)
1: UCSC Main on Human: knownGene (chr22:1-51304566)


Galaxy
http://main.g2.bx.psu.edu/
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
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
Statistics
Join
Join:
1: Exons
First query
with:
2: SNPs
Second query
with min overlap:
1
(bp)
Return:
Only records that are joined (INNER JOIN)
Execute
TIP: If your query does not appear in the pulldown menu, it means that 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
Where overlap specifies the minimum overlap between intervals that allows them to be joined.
Return only records that are joined returns only the records of the first query that join to a record in the second query. This is analogous to an INNER JOIN.
Return all records of first query (fill null with ".") returns all intervals of the first query, and any intervals that do not join an interval from the second query are filled in with a period(.). This is analogous to a LEFT JOIN.
Return all records of second query (fill null with ".") returns all intervals of the second query, and any intervals that do not join an interval from the first query are filled in with a period(.). Note that this
History
Options
Galaxy 101
2: SNPs
1: Exons


Galaxy
http://main.g2.bx.psu.edu/
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Join two Queries side by side on a specified field
Compare two Queries to find common or distinct rows
Subtract Whole Query from another query
Group data by a column and perform aggregate operation on other columns.
Column Join
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
Group
Select data:
3: Join on data 2 and data 1
Query missing? See TIP below.
Group by column:
c4
Ignore case while grouping?:
Operations
Add new Operation
Execute
TIP: If your data is not TAB delimited, use Text Manipulation->Convert
Syntax
This tool allows you to group the input dataset by a particular column and perform aggregate functions like Mean, Median, Mode, Sum, Max, Min, Count, Random draw and Concatenate on other columns.
All invalid, blank and comment lines are skipped when performing the aggregate functions. The number of skipped lines is displayed in the resulting history item.
If multiple modes are present, all are reported.
Example
For the following input:
chr22 1000 1003 TTT
chr22 2000 2003 aaa
chr10 2200 2203 TTT
chr10 1200 1203 ttt
chr22 1600 1603 AAA
Grouping on column 4 while ignoring case, and performing operation Count on column 1 will return:
AAA 2
History
Options
Galaxy 101
3: Join on data 2 and data 1
16,190 regions, format: interval, database: hg19
Info:
| display at UCSC main | view in GeneTrack | display at Ensembl Current
1.Chrom 2.Start 3.End 4.Name
chr22 16258185 16258303 uc002zlh.1_c
chr22 16266928 16267095 uc002zlh.1_c
chr22 16266928 16267095 uc002zlh.1_c
chr22 16266928 16267095 uc002zlh.1_c
chr22 16266928 16267095 uc002zlh.1_c
chr22 16269872 16269943 uc002zlh.1_c
2: SNPs
1: Exons


Galaxy
http://main.g2.bx.psu.edu/
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
Filter on ambiguities in polymorphism datasets
GFF
Extract features from GFF file
Filter GFF file by attribute using simple expressions
Filter GFF file by feature count using simple expressions
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS
Sort
Sort Query:
4: Group on data 3
on column:
c2
with flavor:
Numerical sort
everything in:
Descending order
Column selections
Add new Column selection
Execute
TIP: If your data is not TAB delimited, use Text Manipulation->Convert
Syntax
This tool sorts the dataset on any number of columns in either ascending or descending order.
Numerical sort orders numbers by their magnitude, ignores all characters besides numbers, and evaluates a string of numbers to the value they signify.
Alphabetical sort is a phonebook type sort based on the conventional order of letters in an alphabet. Each nth letter is compared with the nth letter of other words in the list, starting at the first letter of each word and advancing to the second, third, fourth, and so on, until the order is established. Therefore, in an alphabetical sort, 2 comes after 100 (1 < 2).
Examples
The list of numbers 4,17,3,5 collates to 3,4,5,17 by numerical sorting, while it collates to 17,3,4,5 by alphabetical sorting.
History
Options
Galaxy 101
4: Group on data 3
7,030 lines, format: tabular, database: hg19
Info: --Group by c4: count[c4]
1 2
uc002zlh.1_cds_1_0_chr22_16250186_r 1
uc002zlh.1_cds_2_0_chr22_16266929_r 4
uc002zlh.1_cds_4_0_chr22_16269073_r 1
uc002zlh.1_cds_5_0_chr22_16275207_r 2
uc002zlh.1_cds_6_0_chr22_16277748_r 5
uc002zlh.1_cds_7_0_chr22_16279195_r 2
3: Join on data 2 and data 1
2: SNPs
1: Exons


Galaxy
http://main.g2.bx.psu.edu/
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Lift-Over
Text Manipulation
Add column to an existing query
Compute an expression on every row
Concatenate queries tail-to-head
Condense consecutive characters
Convert delimiters to TAB
Merge Columns together
Create single interval as a new query
Cut columns from a table
Change Case of selected columns
Paste two files side by side
Remove beginning of a file
Select random lines from a file
Select first lines from a Query
Select last lines from a Query
Trim leading or trailing characters
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Select first
Select first:
5
lines
from:
5: Sort on data 4
Execute
What it does
This tool outputs specified number of lines from the beginning of a dataset
Example
Selecting 2 lines from this:
chr7 56632 56652 D17003_CTCF_R6 310 +
chr7 56736 56756 D17003_CTCF_R7 354 +
chr7 56761 56781 D17003_CTCF_R4 220 +
chr7 56772 56792 D17003_CTCF_R7 372 +
chr7 56775 56795 D17003_CTCF_R4 207 +
will produce:
chr7 56632 56652 D17003_CTCF_R6 310 +
chr7 56736 56756 D17003_CTCF_R7 354 +
History
Options
Galaxy 101
5: Sort on data 4
7,030 lines, format: tabular, database: hg19
Info:
1 2
uc010gsw.1_cds_1_0_chr22_21480537_r 67
uc002zoc.2_cds_0_0_chr22_18834445_f 58
uc003bhh.2_cds_0_0_chr22_46652458_r 46
uc002zsd.3_cds_0_0_chr22_20456382_r 41
uc002zvq.3_cds_0_0_chr22_21738148_f 41
uc002zvy.3_cds_0_0_chr22_21900346_r 41
4: Group on data 3
3: Join on data 2 and data 1
2: SNPs
1: Exons

Galaxy 101 Highlights

Requires only a Web browser

Integrated data sources, including UCSC

Collection of many tools that can be combined into complex analyses

Datasets grouped together into a history

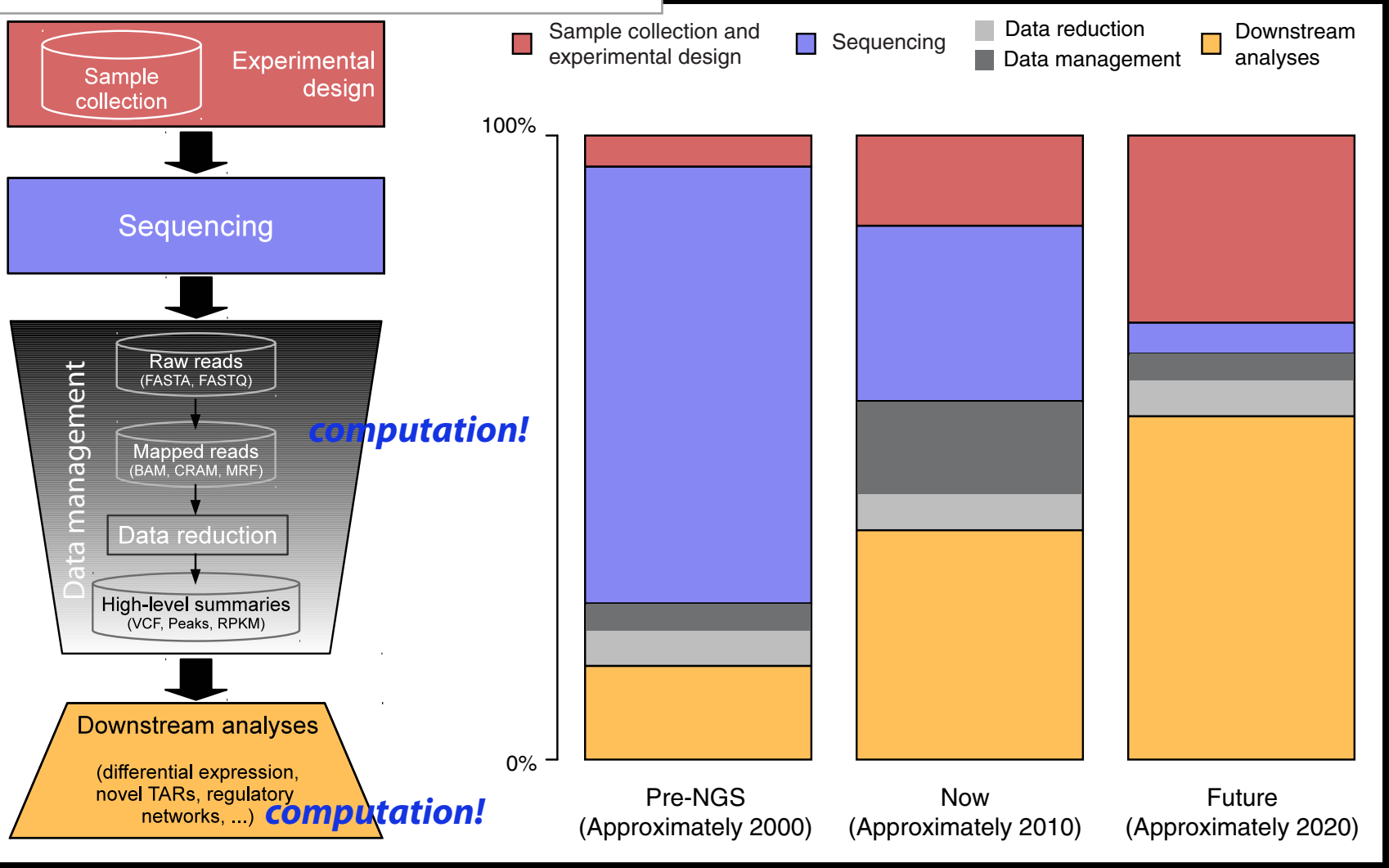
Sboner et al. Genome Biology 2011, 12:125
http://genomebiology.com/2011/12/8/125

Genome Biology

OPINION

The real cost of sequencing: higher than you think!

Andrea Sboner[1,2], Xinmeng Jasmine Mu[1], Dov Greenbaum[1,2,3,4,5], Raymond K Auerbach[1] and Mark B Gerstein*[1,2,6]



Computation in Science?

Scientists unfamiliar with computation

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

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

Galaxy Project: Fundamental Questions

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

- make tools and methods accessible to scientists?
- ensure that analyses are reproducible?
- enable transparent communication and reuse of analyses?

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

Amplification

Many tools available in a single place means that tools can be combined in novel ways

Everything is open source: framework, protocols, and libraries

- reused and extended by anyone

Users, developers, community benefit

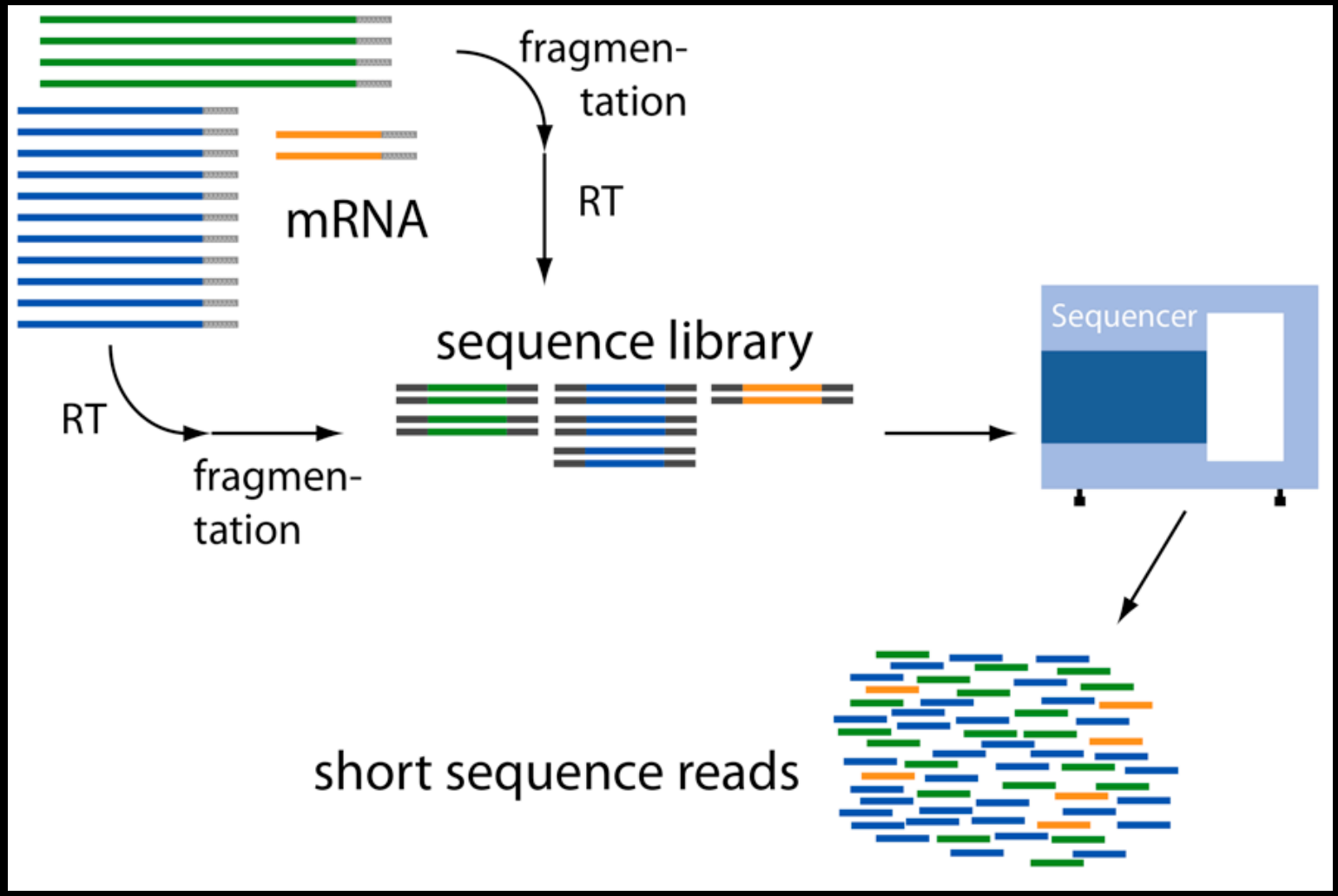
Questions

Can current RNA-seq tools be used to identify and quantify *XBP1*'s highly similar isoforms?

Can *XBP1*'s translational pause point be identified?

RNA-seq

next generation gene expression data



http://www.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics

Illumina BodyMap 2.0

RNA-seq (Hi-Seq) data from 16 human tissues

Two datasets for each tissue

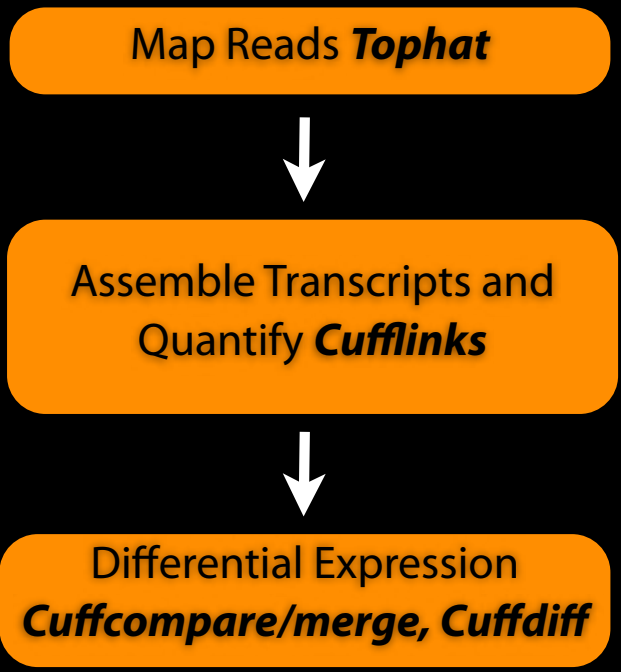
- 50bp paired-end
- 75bp single-end

~80 million reads per dataset, ~160 million reads per tissue

In a public data library on usegalaxy.org

http://www.ensembl.info/blog/2011/05/24/human-bodymap-2-0-data-from-illumina/

Differential Expression Analysis in Galaxy



NGS: Expression Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffmerge merge together several Cufflinks assemblies
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

DE NOVO ASSEMBLY

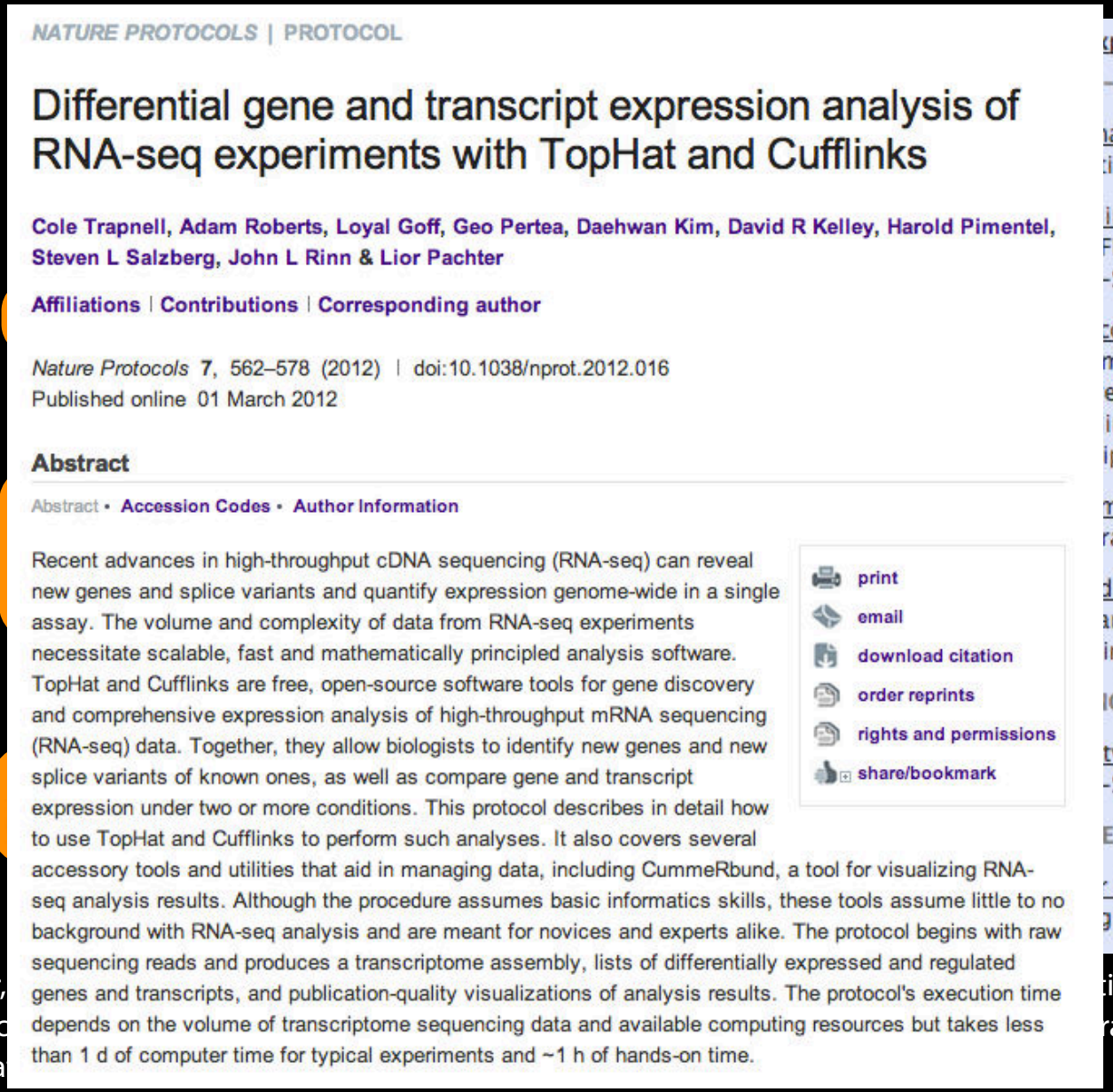
- Trinity De novo assembly of RNA-Seq data

FILTERING

- Filter Combined Transcripts using tracking file

1. Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111 (2009).
2. Trapnell et al. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nature Biotechnology doi:10.1038/nbt.1621

Differential Expression Analysis in



NATURE PROTOCOLS | PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

Affiliations | Contributions | Corresponding author

Nature Protocols **7**, 562–578 (2012) | doi:10.1038/nprot.2012.016
Published online 01 March 2012

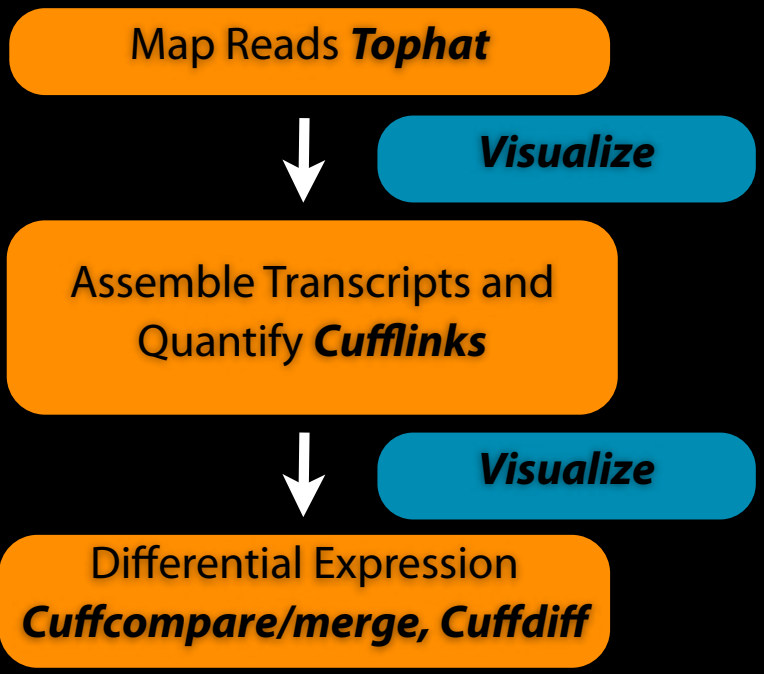
Abstract

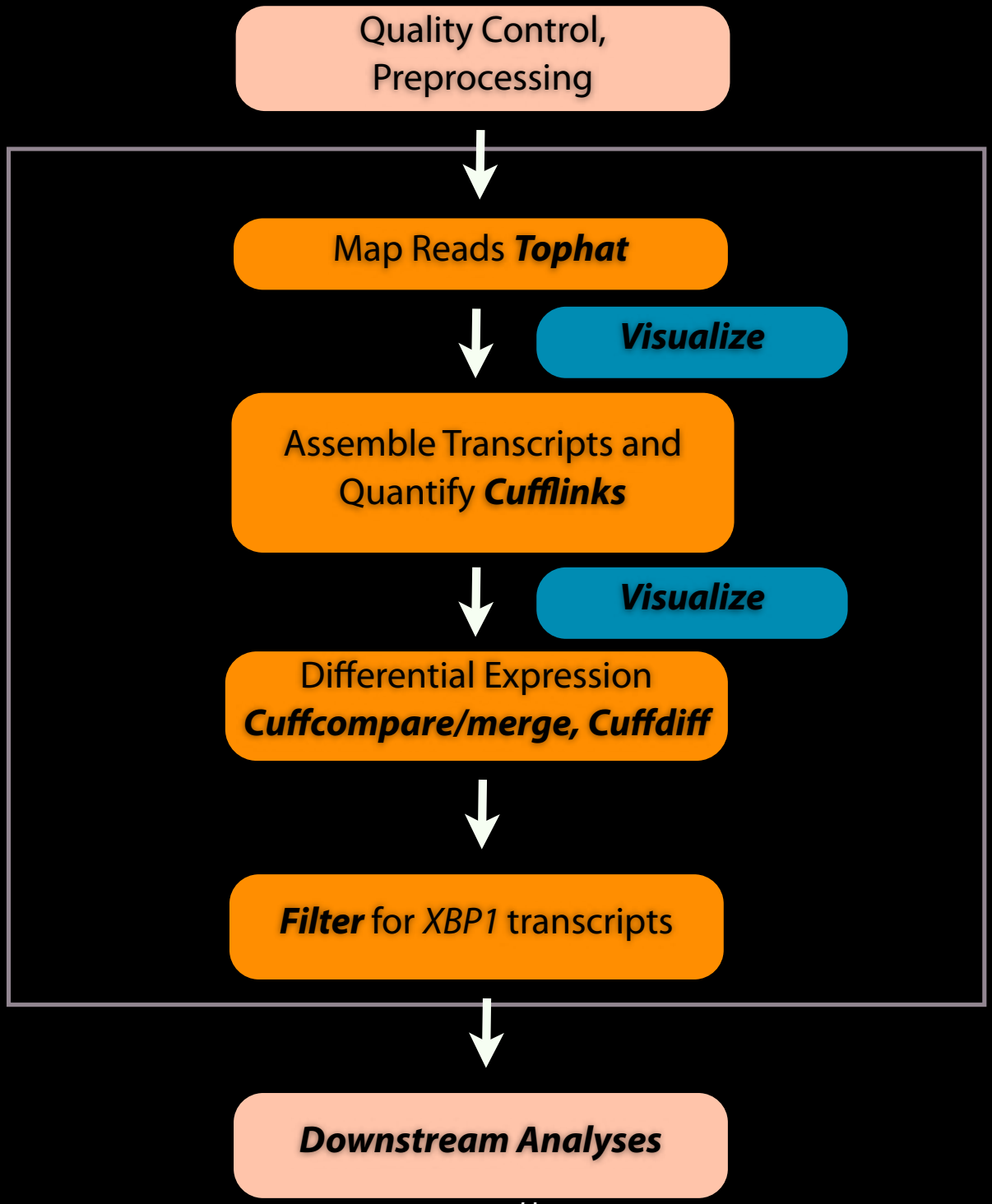
Abstract • Accession Codes • Author Information

Recent advances in high-throughput cDNA sequencing (RNA-seq) can reveal new genes and splice variants and quantify expression genome-wide in a single assay. The volume and complexity of data from RNA-seq experiments necessitate scalable, fast and mathematically principled analysis software. TopHat and Cufflinks are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. Together, they allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. This protocol describes in detail how to use TopHat and Cufflinks to perform such analyses. It also covers several accessory tools and utilities that aid in managing data, including CummeRbund, a tool for visualizing RNA-seq analysis results. Although the procedure assumes basic informatics skills, these tools assume little to no background with RNA-seq analysis and are meant for novices and experts alike. The protocol begins with raw sequencing reads and produces a transcriptome assembly, lists of differentially expressed and regulated genes and transcripts, and publication-quality visualizations of analysis results. The protocol's execution time depends on the volume of transcriptome sequencing data and available computing resources but takes less than 1 d of computer time for typical experiments and ~1 h of hands-on time.

pression Analysis
-SEQ
at for Illumina Find splice
ions using RNA-seq data
inks transcript assembly
PKM (RPKM) estimates for
-Seq data
compare compare
mbled transcripts to a
ence annotation and track
inks transcripts across
iple experiments
merge merge together
ral Cufflinks assemblies
diff find significant changes
anscript expression,
ing, and promoter use
OVO ASSEMBLY
ty De novo assembly of
-Seq data
ERING
Combined Transcripts
g tracking file

1. Trapnell, C., Pachter, ics 25, 1105-1111 (2009).
2. Trapnell et al. Transc ranscripts and switching among isoforms. Na


Map Reads *Tophat*
Visualize
Assemble Transcripts and
Quantify *Cufflinks*
Visualize
Differential Expression
Cuffcompare/merge, Cuffdiff


Quality Control, Preprocessing
Map Reads Tophat
Visualize
Assemble Transcripts and Quantify Cufflinks
Visualize
Differential Expression Cuffcompare/merge, Cuffdiff
Filter for XBP1 transcripts
Downstream Analyses

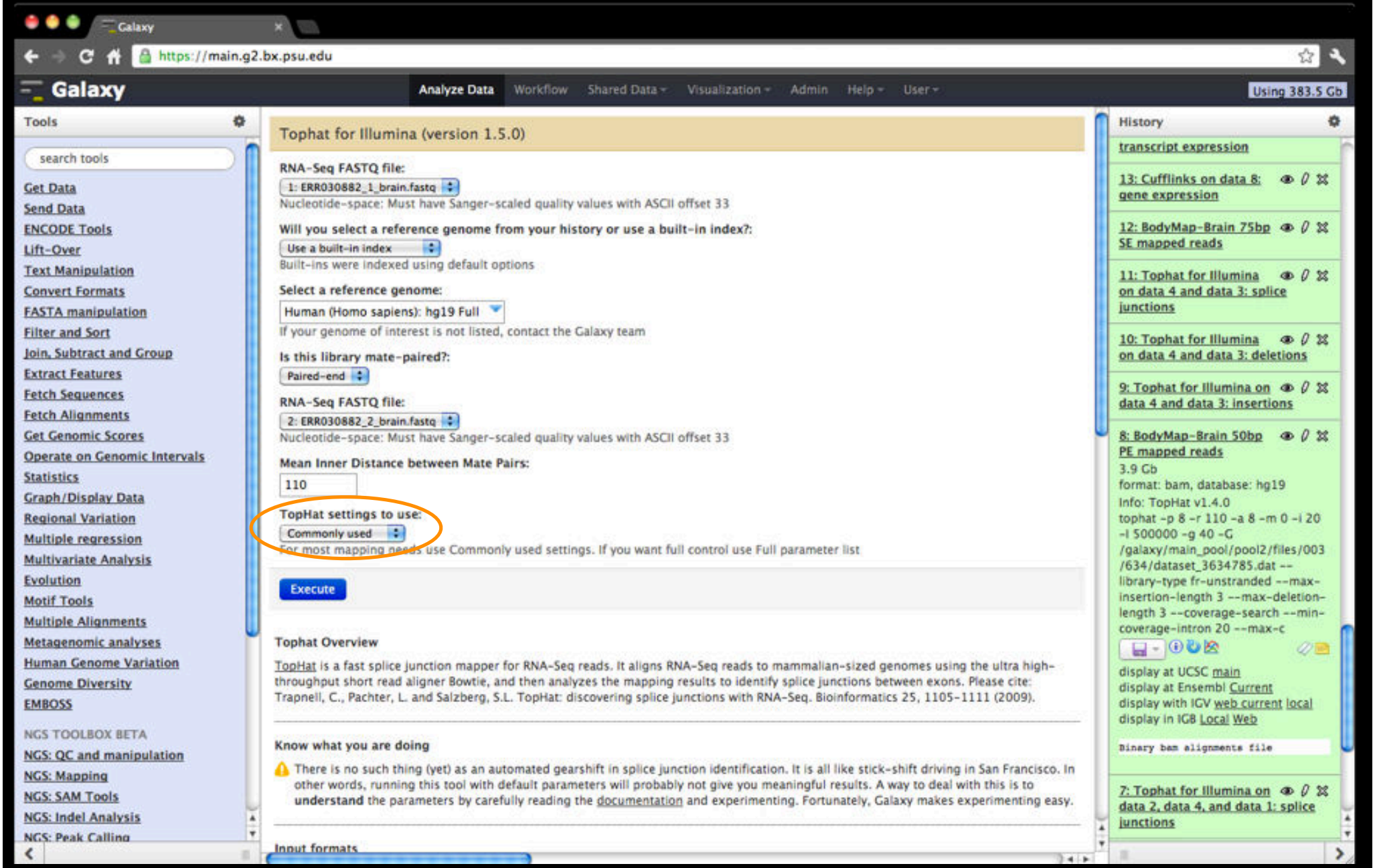
Working with NGS Tools

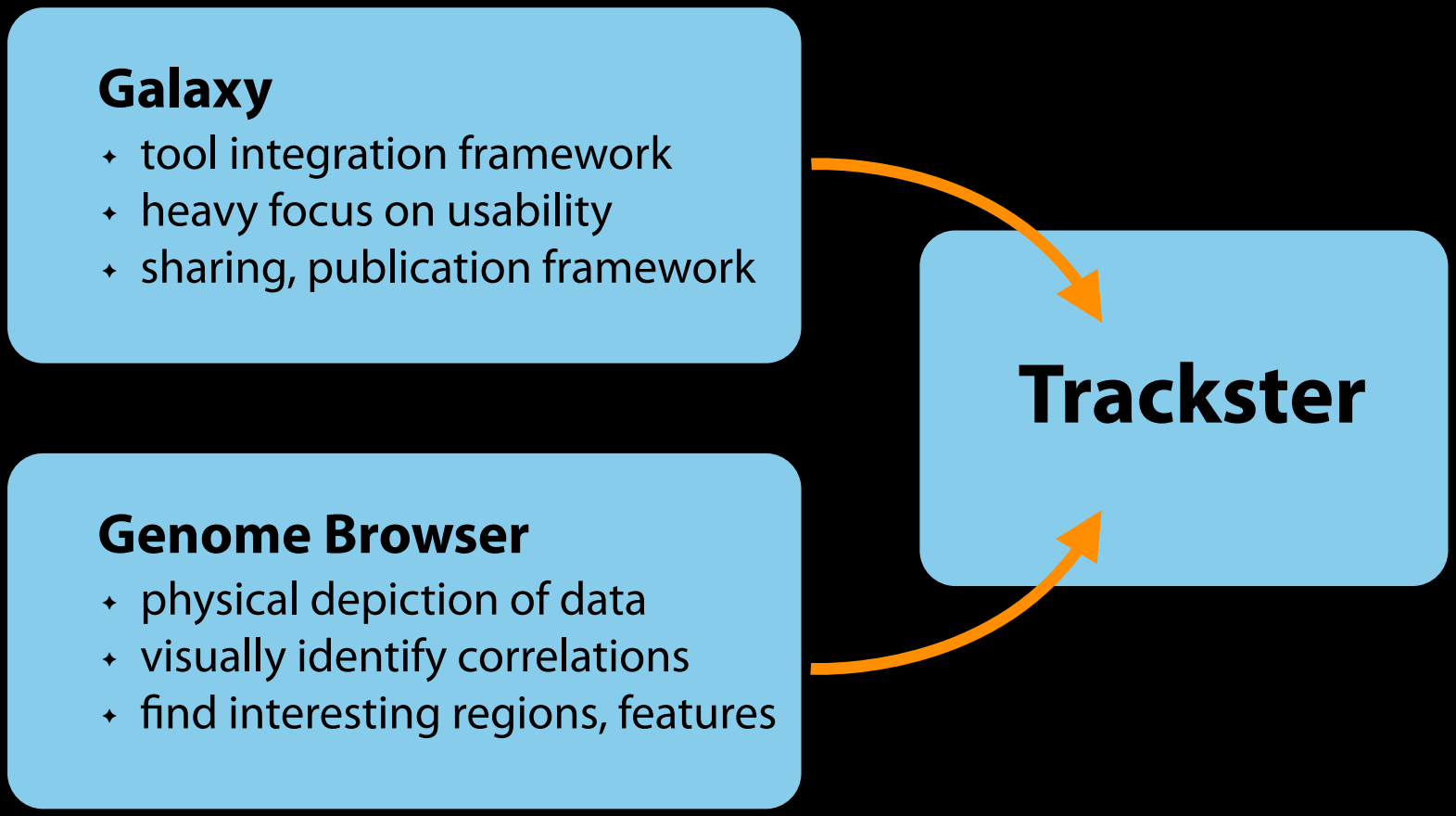
Often challenging

- many parameters
- time intensive
- evaluating results difficult

Three ways Galaxy can help:

- experimentation: can rerun tools, workflows
- visualization: Trackster, display applications
- reproducibility: parameters tracked, workflows available


Galaxy
https://main.g2.bx.psu.edu
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 383.5 Gb
Tools
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
Genome Diversity
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
Tophat for Illumina (version 1.5.0)
RNA-Seq FASTQ file:
1: ERR030882_1_brain.fastq
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33
Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Built-ins were indexed using default options
Select a reference genome:
Human (Homo sapiens): hg19 Full
If your genome of interest is not listed, contact the Galaxy team
Is this library mate-paired?:
Paired-end
RNA-Seq FASTQ file:
2: ERR030882_2_brain.fastq
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33
Mean Inner Distance between Mate Pairs:
110
TopHat settings to use:
Commonly used
For most mapping needs use Commonly used settings. If you want full control use Full parameter list
Execute
Tophat Overview
TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Please cite: Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111 (2009).
Know what you are doing
There is no such thing (yet) as an automated gearshift in splice junction identification. It is all like stick-shift driving in San Francisco. In other words, running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to understand the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.
Input formats
History
transcript expression
13: Cufflinks on data 8: gene expression
12: BodyMap-Brain 75bp SE mapped reads
11: Tophat for Illumina on data 4 and data 3: splice junctions
10: Tophat for Illumina on data 4 and data 3: deletions
9: Tophat for Illumina on data 4 and data 3: insertions
8: BodyMap-Brain 50bp PE mapped reads
3.9 Gb
format: bam, database: hg19
Info: TopHat v1.4.0
tophat -p 8 -r 110 -a 8 -m 0 -i 20 -I 500000 -g 40 -G /galaxy/main_pool/pool2/files/003/634/dataset_3634785.dat --library-type fr-unstranded --max-insertion-length 3 --max-deletion-length 3 --coverage-search --min-coverage-intron 20 --max-c
display at UCSC main
display at Ensembl Current
display with IGV web current local
display in IGB Local Web
Binary bam alignments file
7: Tophat for Illumina on data 2, data 4, and data 1: splice junctions


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

Trackster

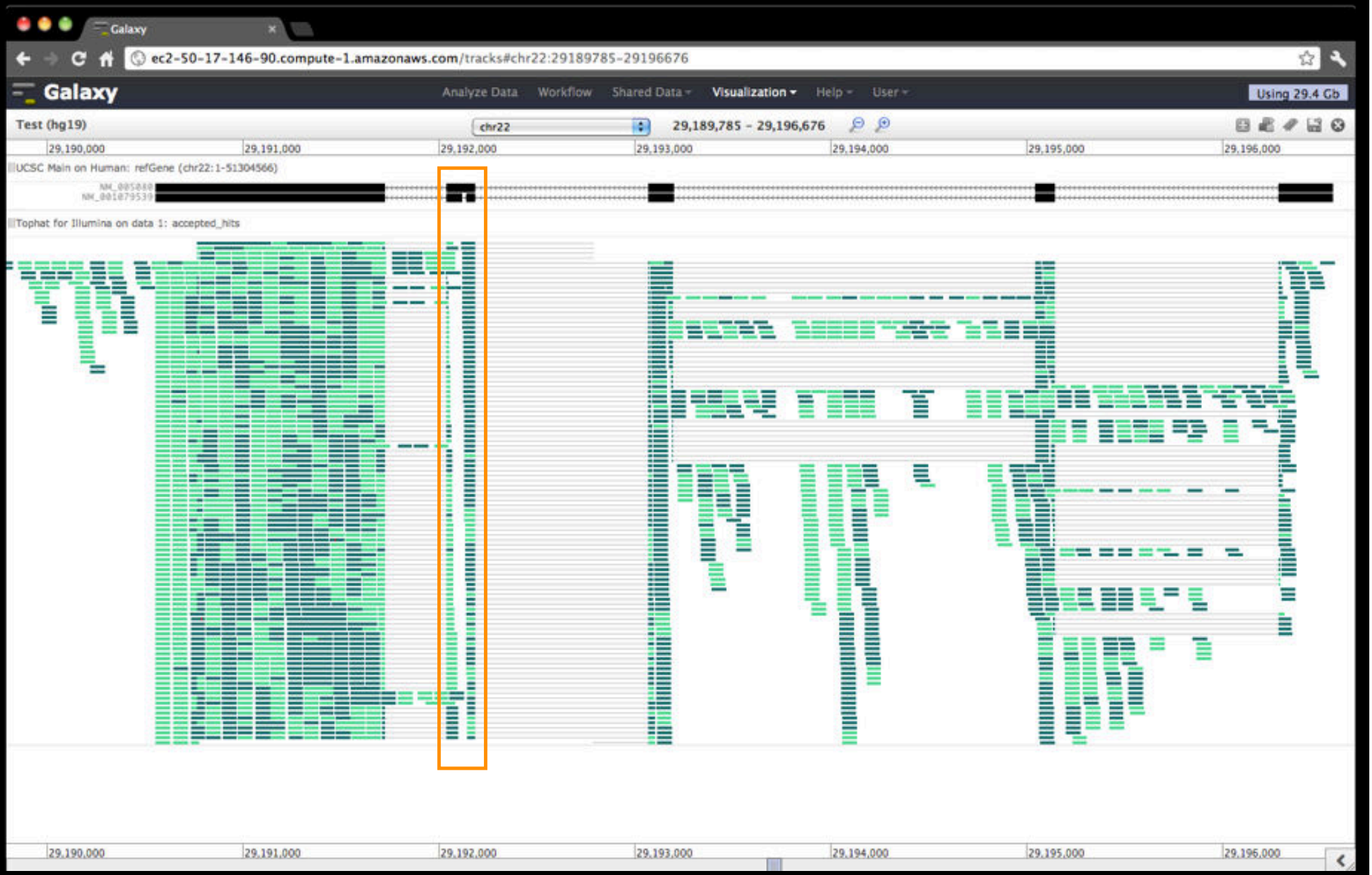
Web-based visualization of your NGS data

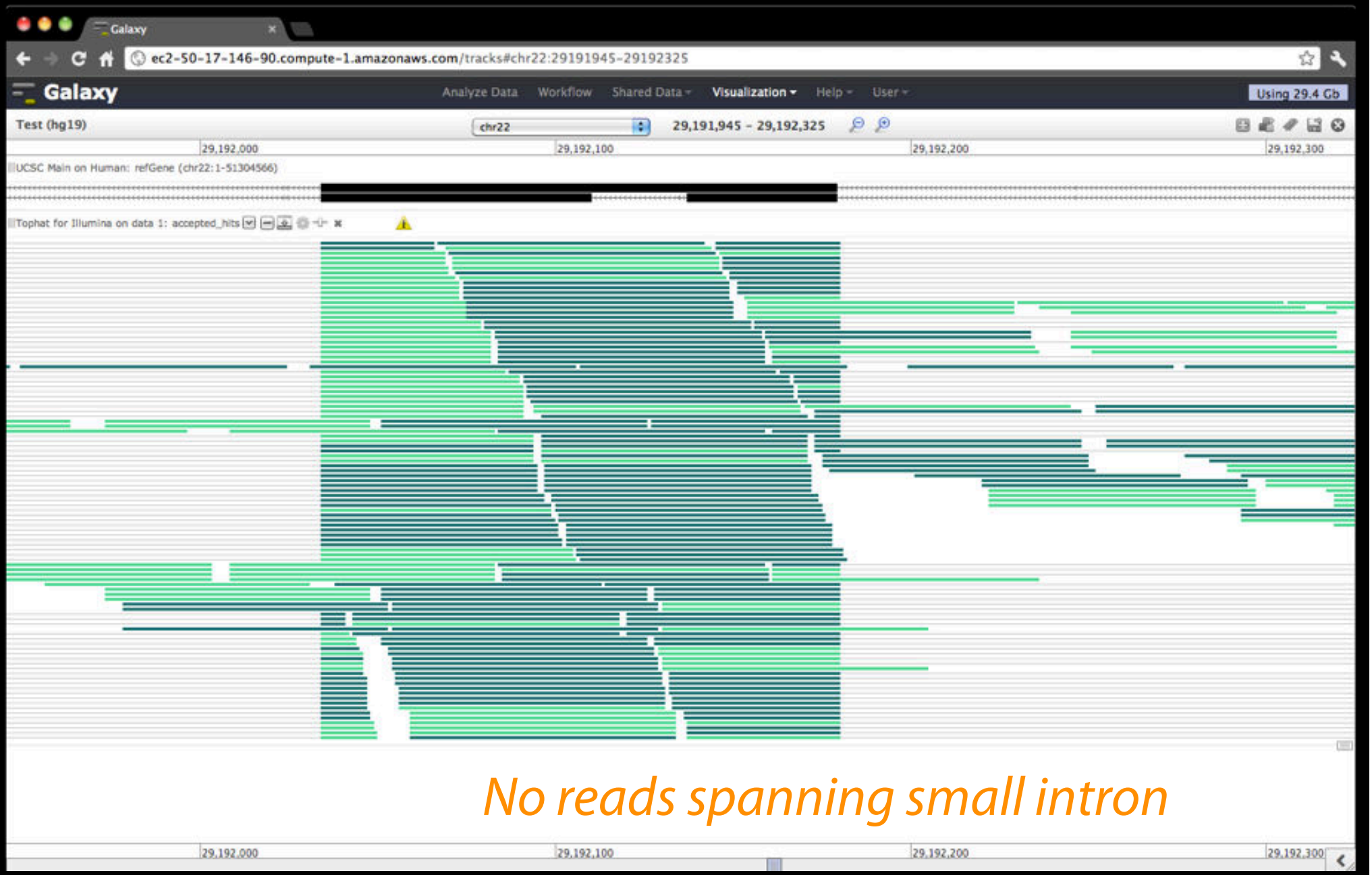
- requires only a Web browser
- dynamic and customizable
- supports BAM, BED, GFF/GTF, WIG, VCF

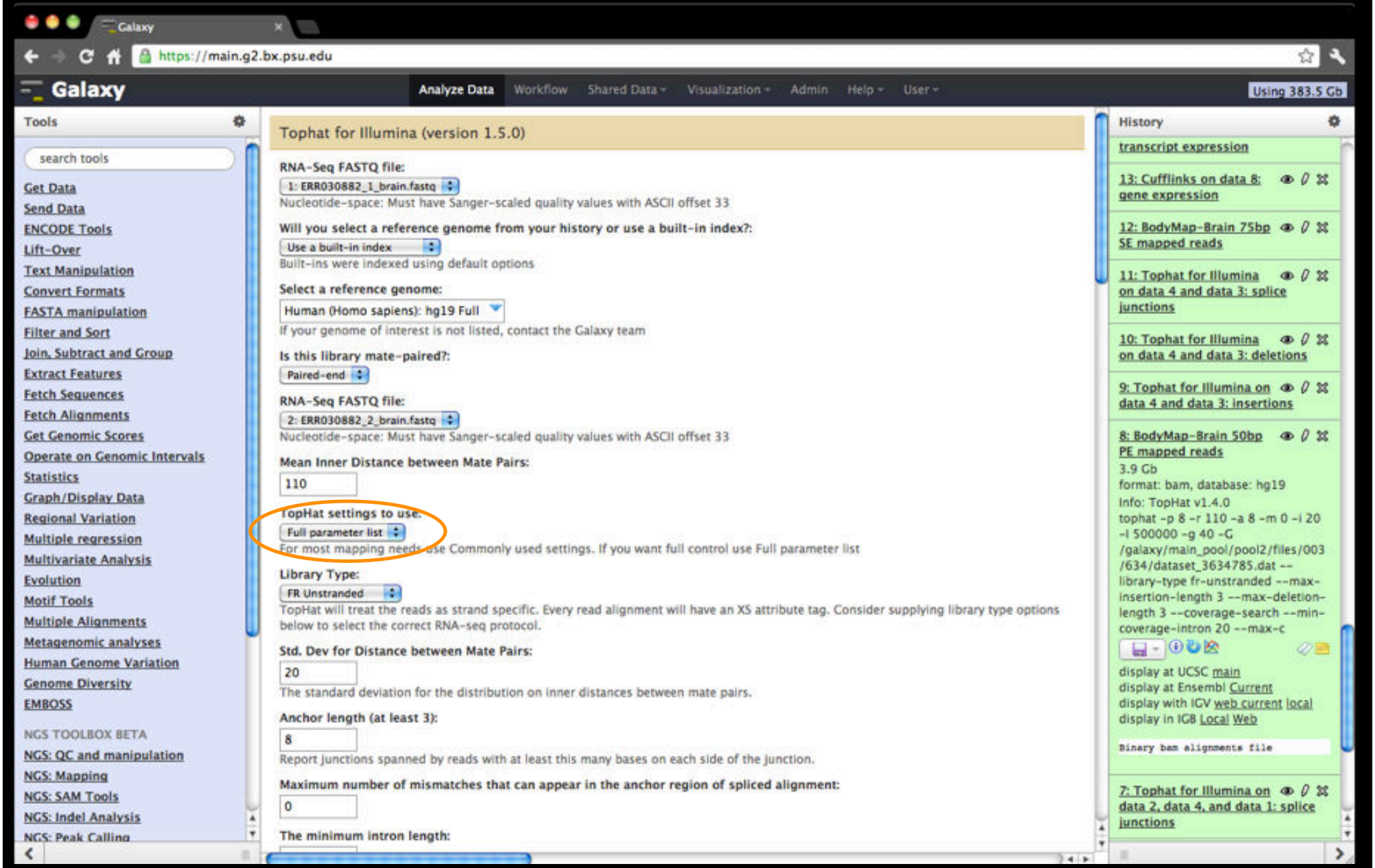
Platform for visual analysis

- tools integrated with visualization

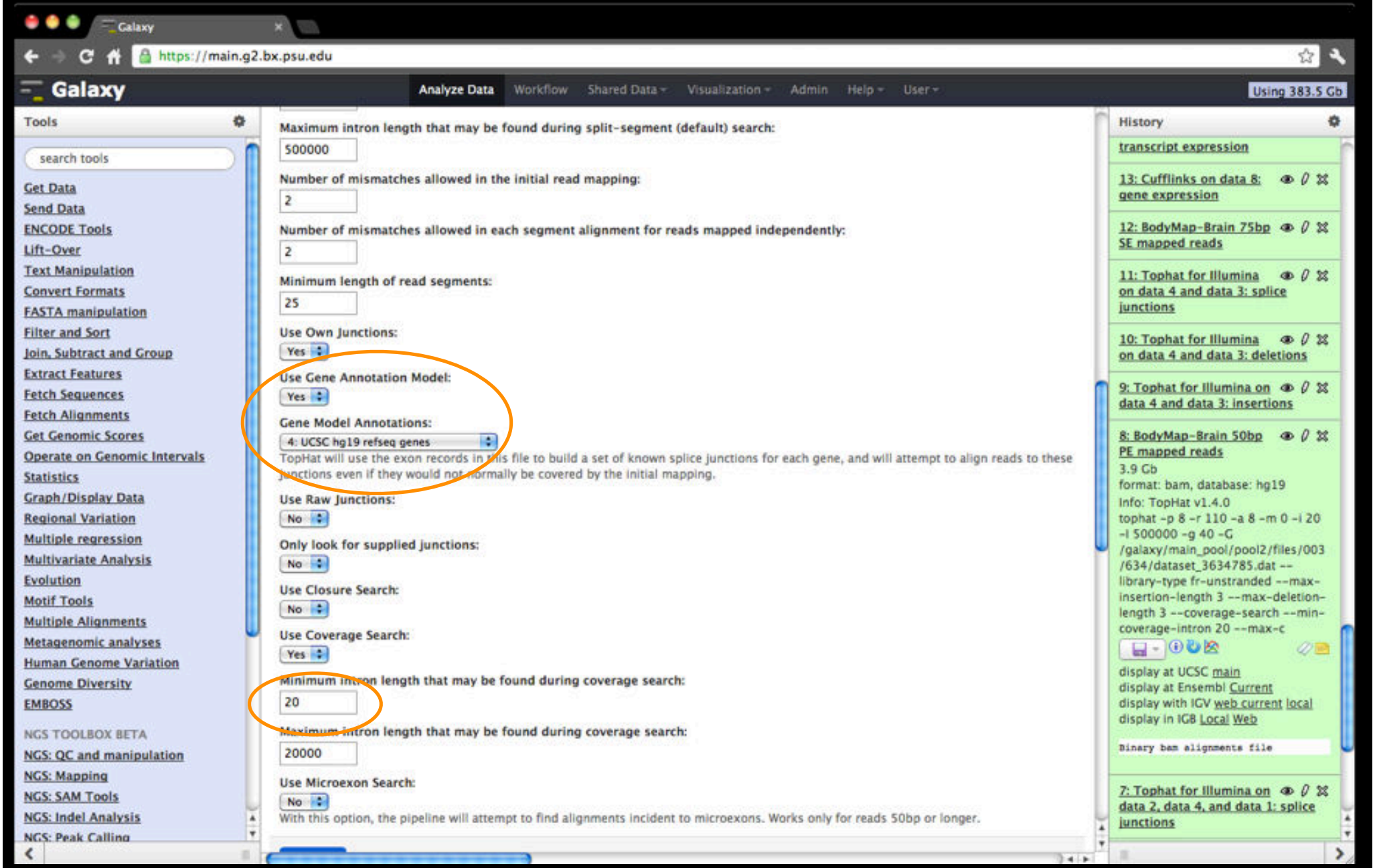
Can share & publish fully-functional visualizations


Galaxy
ec2-50-17-146-90.compute-1.amazonaws.com/tracks#chr22:29189785-29196676
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 29.4 Gb
Test (hg19)
chr22
29,189,785 - 29,196,676
29,190,000
29,191,000
29,192,000
29,193,000
29,194,000
29,195,000
29,196,000
UCSC Main on Human: refGene (chr22:1-51304566)
Tophat for Illumina on data 1: accepted_hits

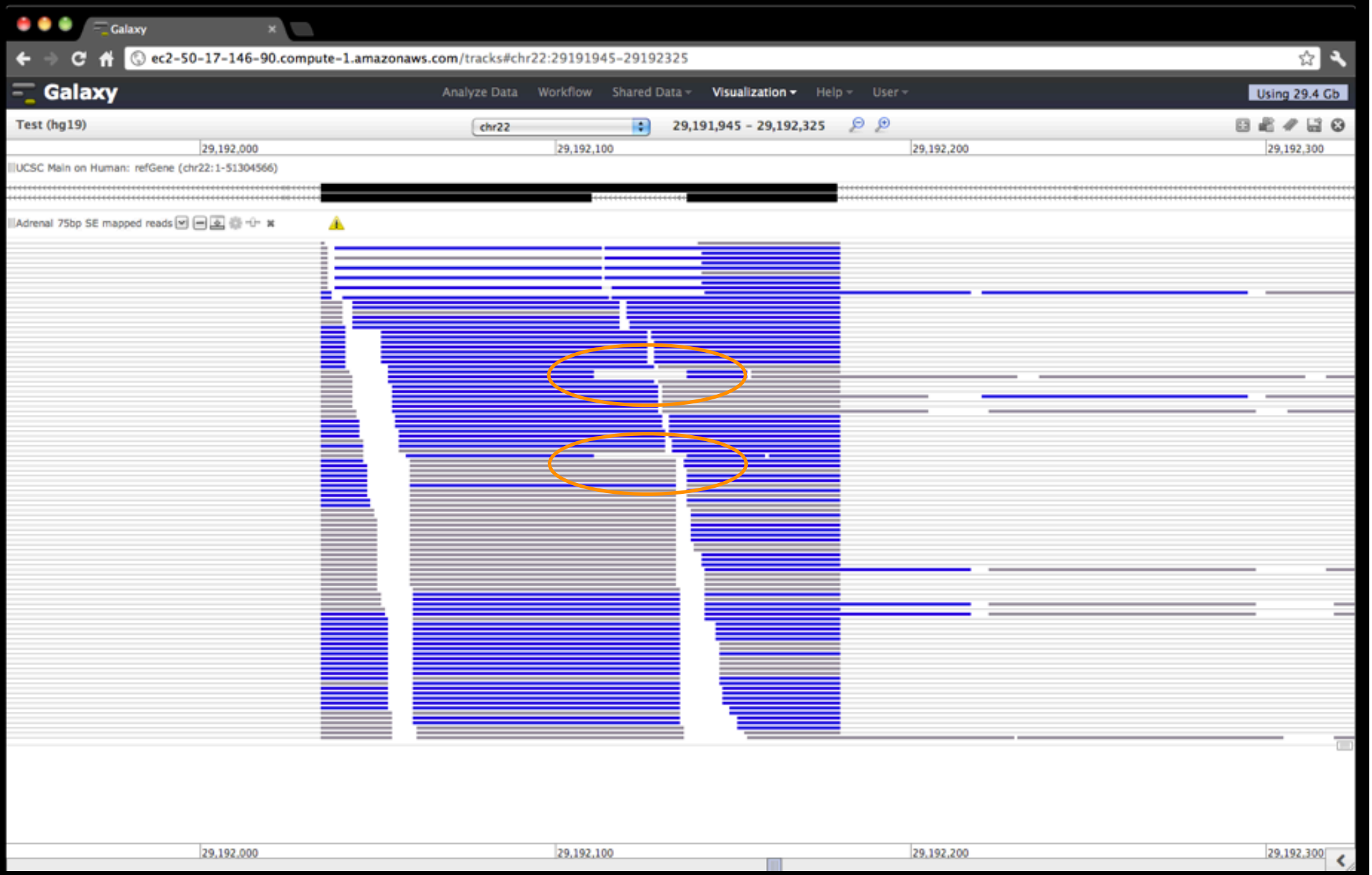

Galaxy
ec2-50-17-146-90.compute-1.amazonaws.com/tracks#chr22:29191945-29192325
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 29.4 Gb
Test (hg19)
chr22
29,191,945 - 29,192,325
29,192,000
29,192,100
29,192,200
29,192,300
UCSC Main on Human: refGene (chr22:1-51304566)
Tophat for Illumina on data 1: accepted_hits
No reads spanning small intron



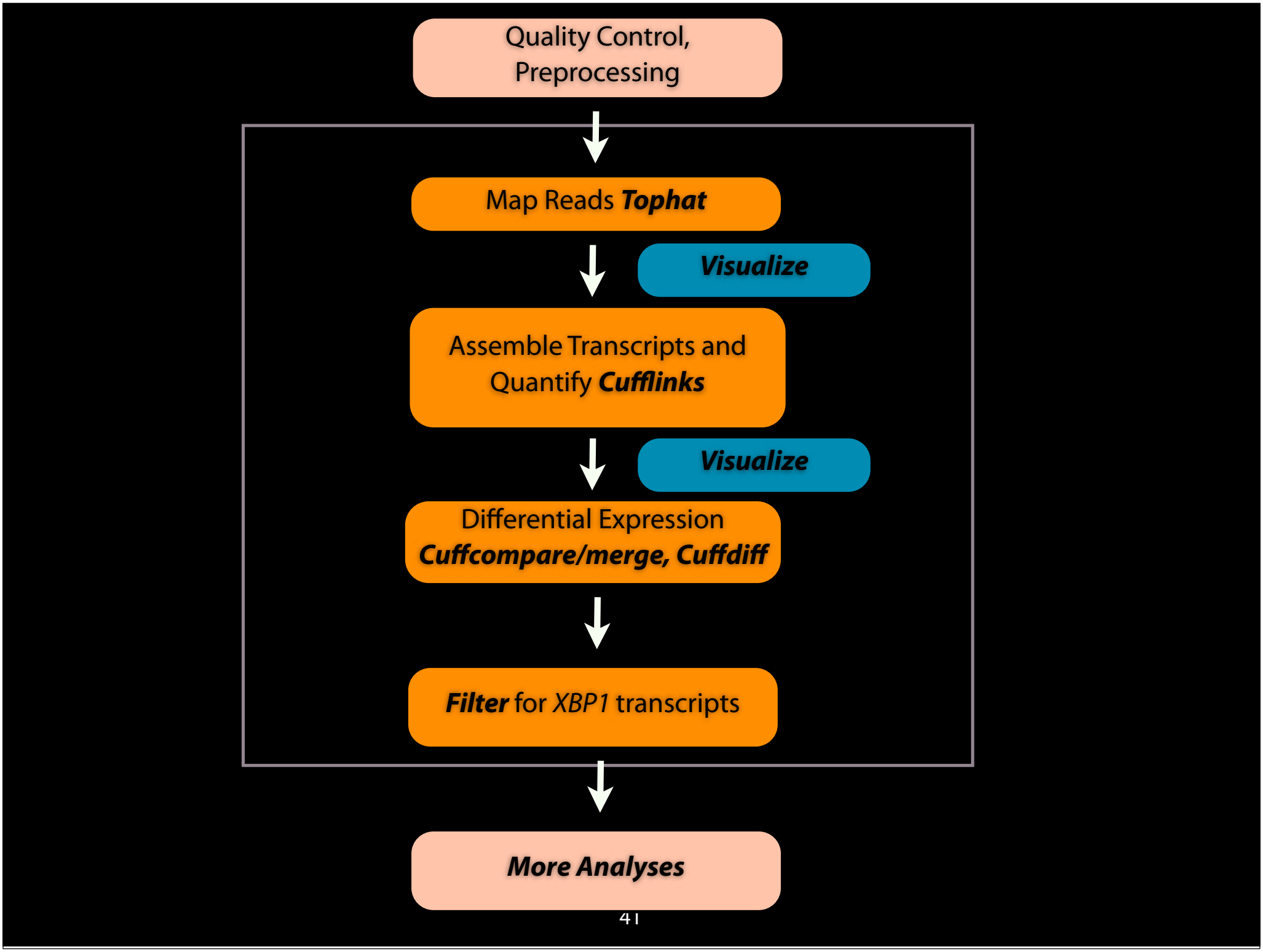
experimentation

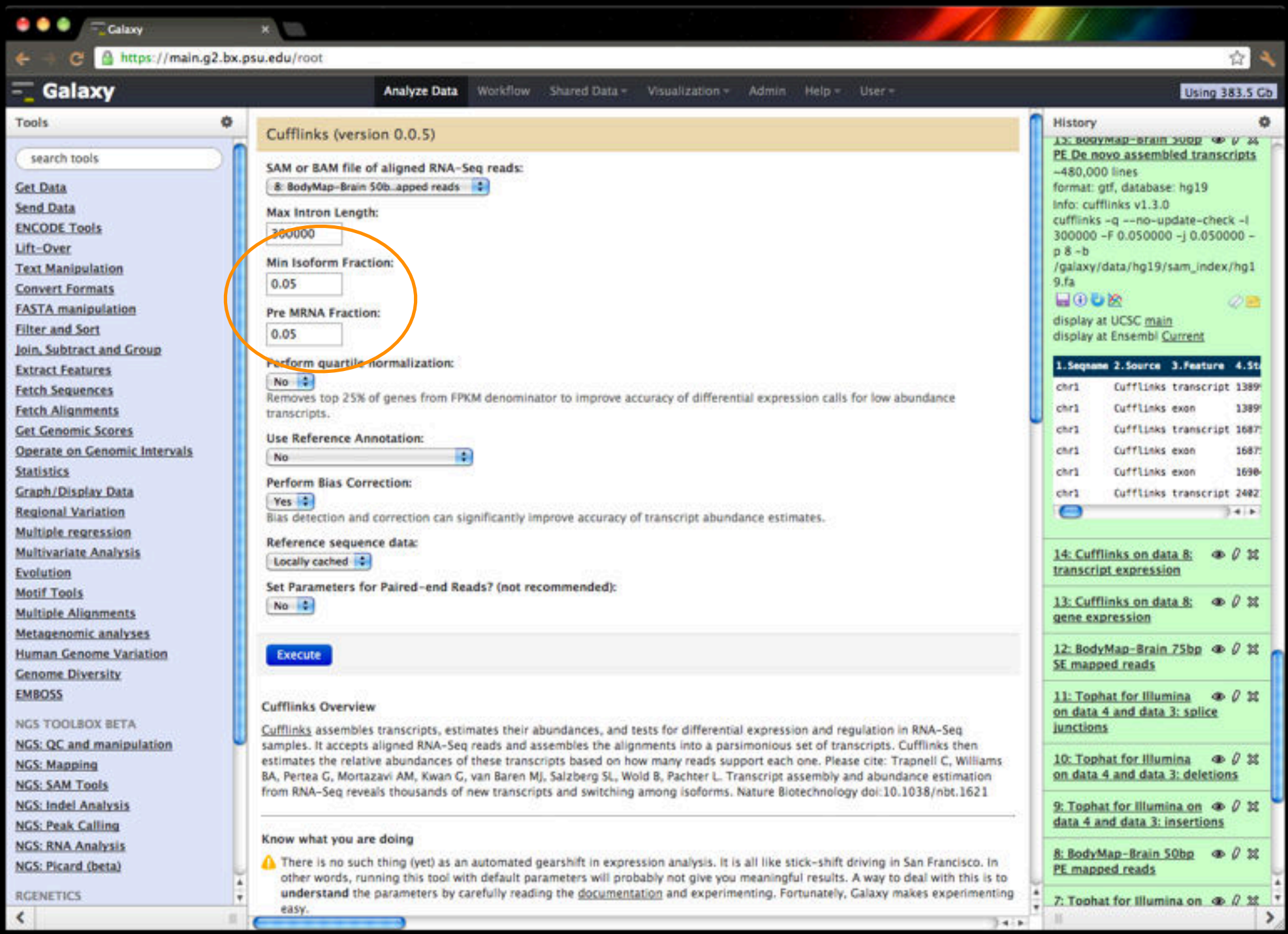


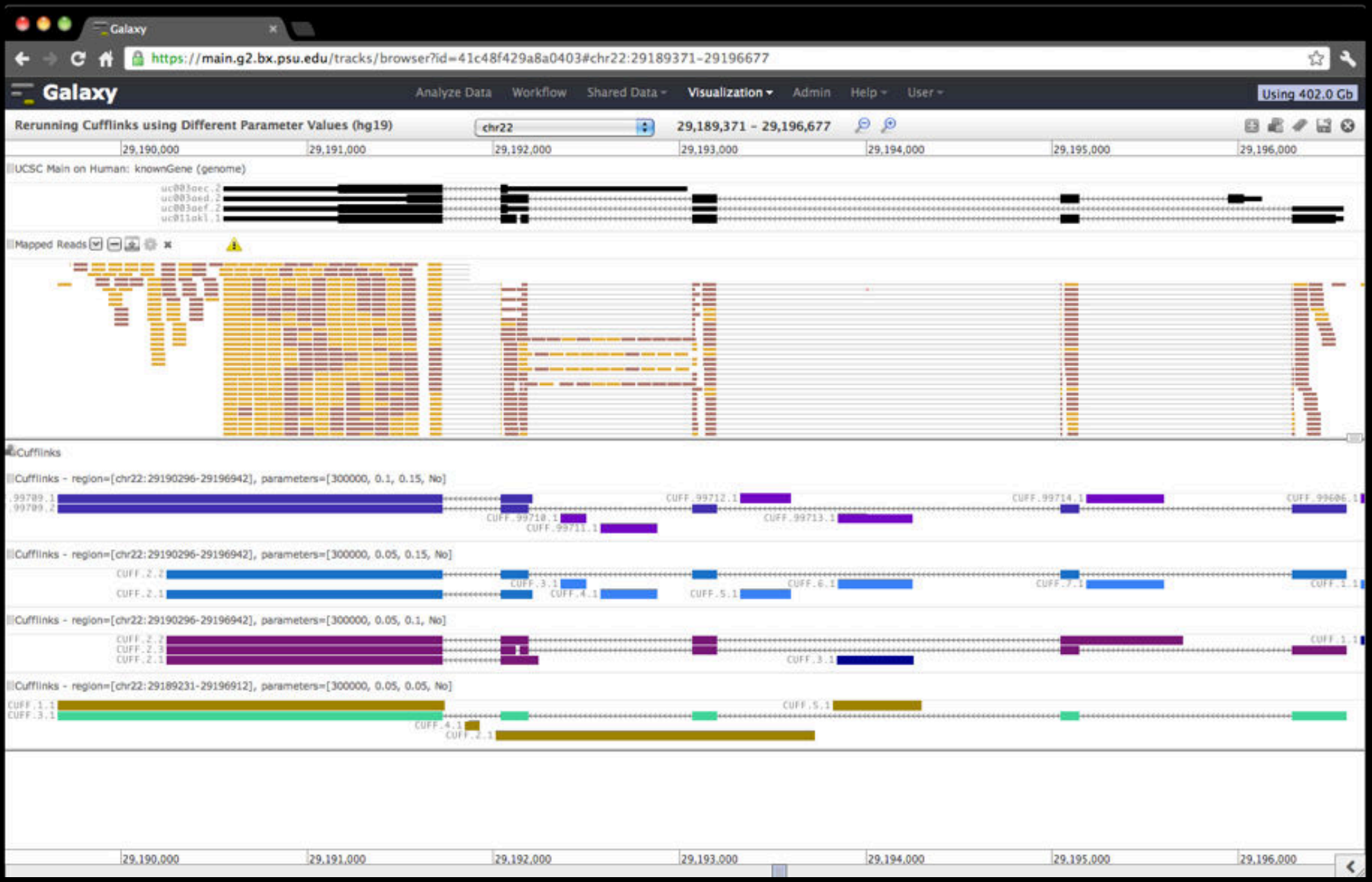
experimentation


Galaxy
ec2-50-17-146-90.compute-1.amazonaws.com/tracks#chr22:29191945-29192325
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 29.4 Gb
Test (hg19)
chr22
29,191,945 - 29,192,325
29,192,000
29,192,100
29,192,200
29,192,300
UCSC Main on Human: refGene (chr22:1-51304566)
Adrenal 75bp SE mapped reads

success


Quality Control, Preprocessing
Map Reads Tophat
Visualize
Assemble Transcripts and Quantify Cufflinks
Visualize
Differential Expression Cuffcompare/merge, Cuffdiff
Filter for XBP1 transcripts
More Analyses


Galaxy
https://main.g2.bx.psu.edu/root
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 383.5 Gb
Tools
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
Genome Diversity
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
NGS: Picard (beta)
RGENETICS
Cufflinks (version 0.0.5)
SAM or BAM file of aligned RNA-Seq reads:
8: BodyMap-Brain 50b..apped reads
Max Intron Length:
300000
Min Isoform Fraction:
0.05
Pre MRNA Fraction:
0.05
Perform quartile normalization:
No
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.
Use Reference Annotation:
No
Perform Bias Correction:
Yes
Bias detection and correction can significantly improve accuracy of transcript abundance estimates.
Reference sequence data:
Locally cached
Set Parameters for Paired-end Reads? (not recommended):
No
Execute
Cufflinks Overview
Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one. Please cite: Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nature Biotechnology doi:10.1038/nbt.1621
Know what you are doing
There is no such thing (yet) as an automated gearshift in expression analysis. It is all like stick-shift driving in San Francisco. In other words, running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to understand the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.
History
PE De novo assembled transcripts
~480,000 lines
format: gtf, database: hg19
Info: cufflinks v1.3.0
cufflinks -q --no-update-check -I 300000 -F 0.050000 -j 0.050000 -p 8 -b /galaxy/data/hg19/sam_index/hg19.fa
display at UCSC main
display at Ensembl Current
1.Seqname 2.Source 3.Feature 4.St
chr1 Cufflinks transcript 1389
chr1 Cufflinks exon 1389
chr1 Cufflinks transcript 1687
chr1 Cufflinks exon 1687
chr1 Cufflinks exon 1690
chr1 Cufflinks transcript 2402
14: Cufflinks on data 8: transcript expression
13: Cufflinks on data 8: gene expression
12: BodyMap-Brain 75bp SE mapped reads
11: Tophat for Illumina on data 4 and data 3: splice junctions
10: Tophat for Illumina on data 4 and data 3: deletions
9: Tophat for Illumina on data 4 and data 3: insertions
8: BodyMap-Brain 50bp PE mapped reads
7: Tophat for Illumina on

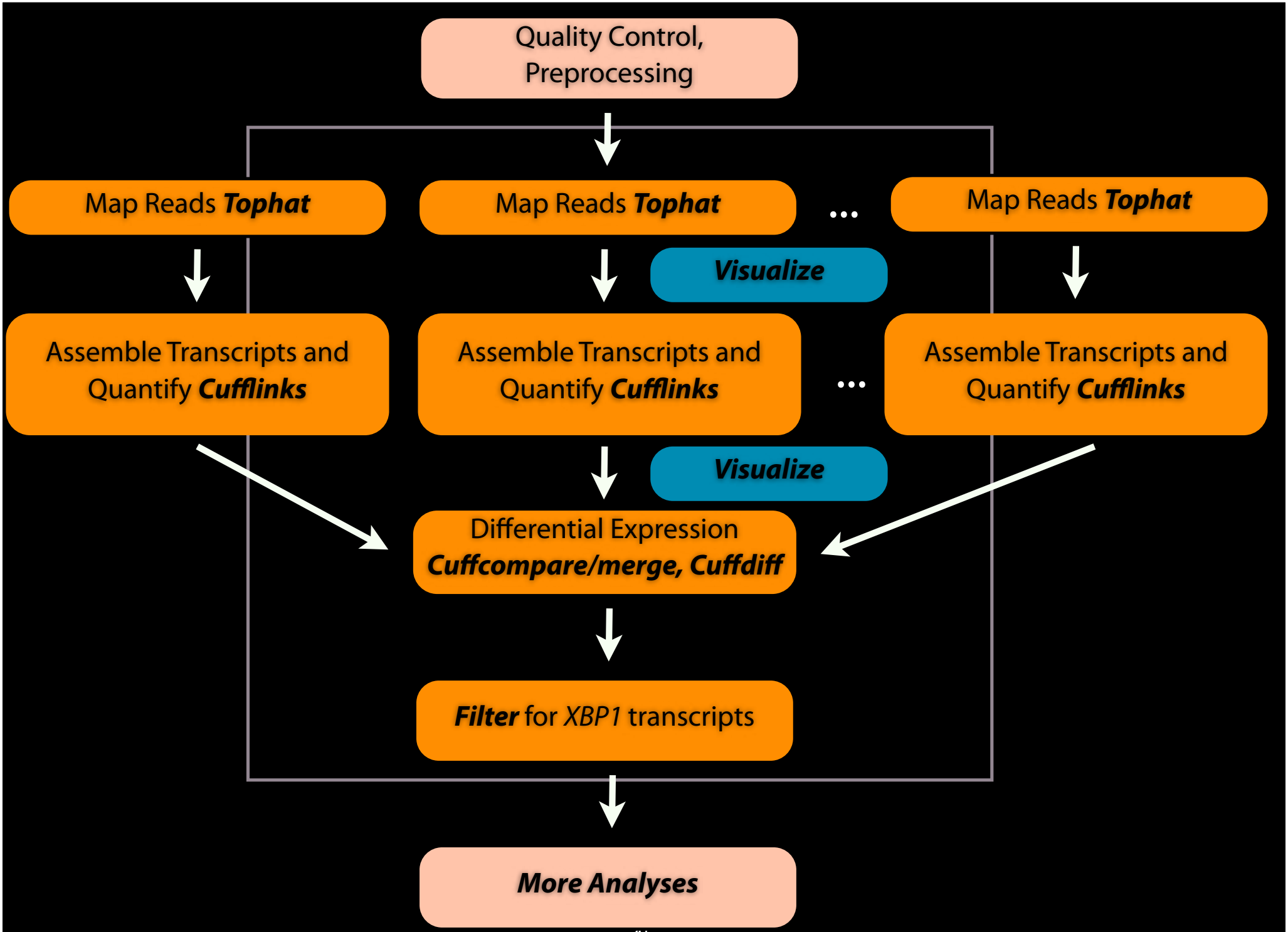

Galaxy
https://main.g2.bx.psu.edu/tracks/browser?id=41c48f429a8a0403#chr22:29189371-29196677
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 402.0 Gb
Rerunning Cufflinks using Different Parameter Values (hg19)
chr22
29,189,371 - 29,196,677
29,190,000
29,191,000
29,192,000
29,193,000
29,194,000
29,195,000
29,196,000
UCSC Main on Human: knownGene (genome)
Mapped Reads
Cufflinks
Cufflinks - region=[chr22:29190296-29196942], parameters=[300000, 0.1, 0.15, No]
Cufflinks - region=[chr22:29190296-29196942], parameters=[300000, 0.05, 0.15, No]
Cufflinks - region=[chr22:29190296-29196942], parameters=[300000, 0.05, 0.1, No]
Cufflinks - region=[chr22:29189231-29196912], parameters=[300000, 0.05, 0.05, No]

Waypoint

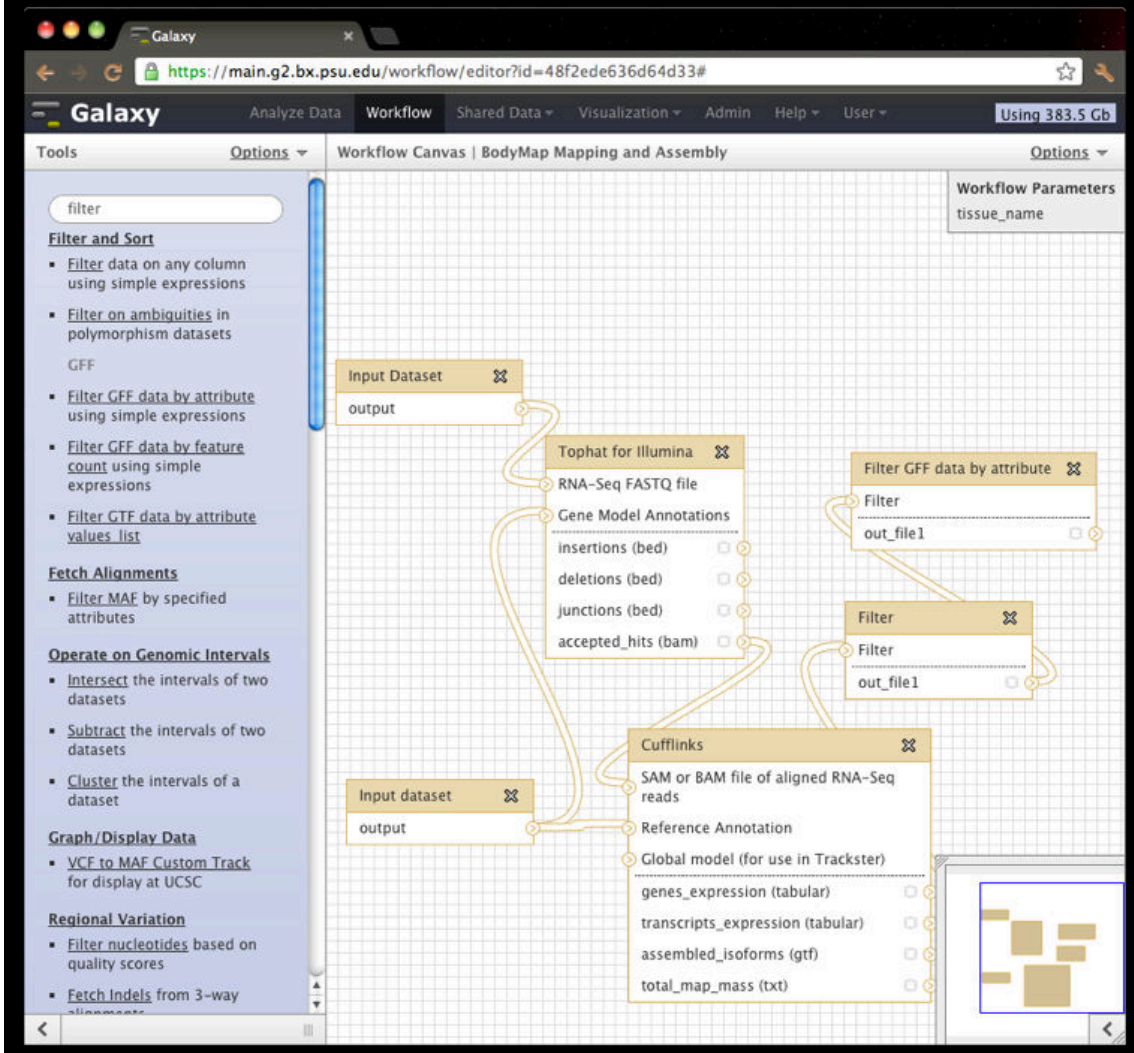
Found good parameter values via experimentation

- can map and assemble RNA-seq reads from a single dataset

But there are 16 tissues, each with 3 RNA-seq datasets


Quality Control, Preprocessing
Map Reads Tophat
Map Reads Tophat
...
Map Reads Tophat
Visualize
Assemble Transcripts and Quantify Cufflinks
Assemble Transcripts and Quantify Cufflinks
...
Assemble Transcripts and Quantify Cufflinks
Visualize
Differential Expression Cuffcompare/merge, Cuffdiff
Filter for XBP1 transcripts
More Analyses

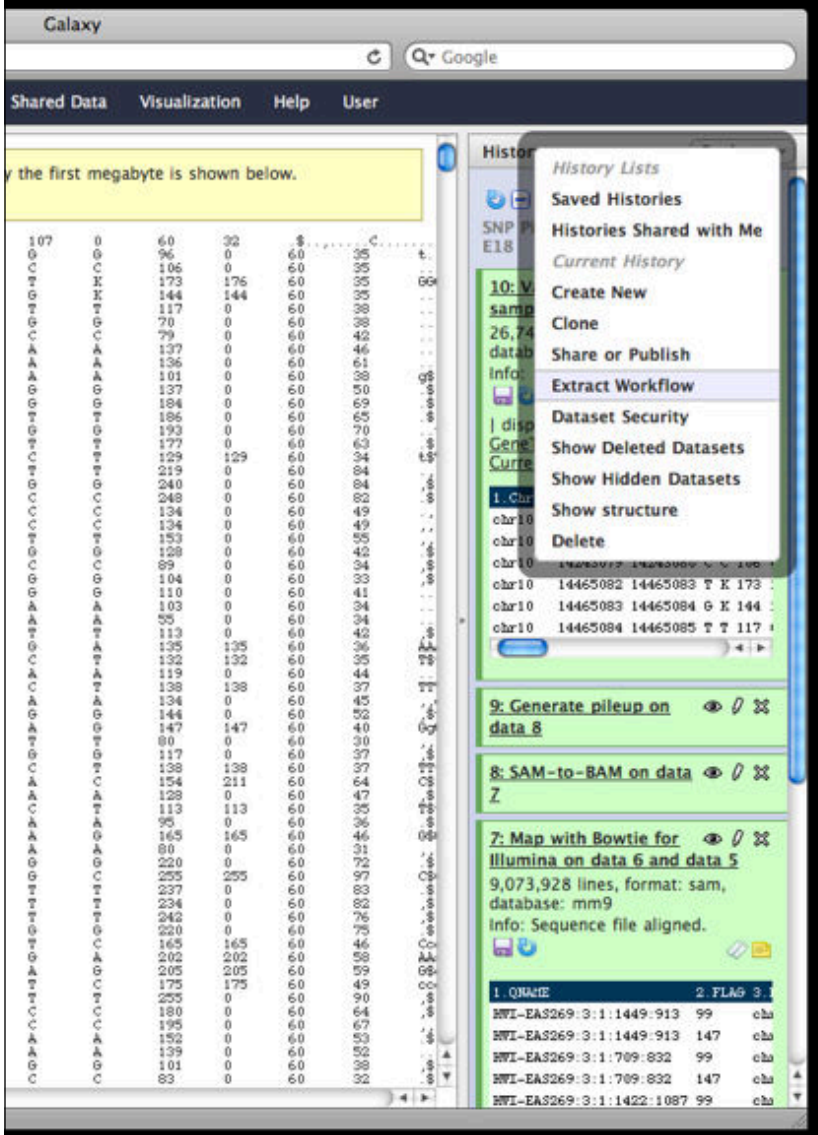
Galaxy Workflow System

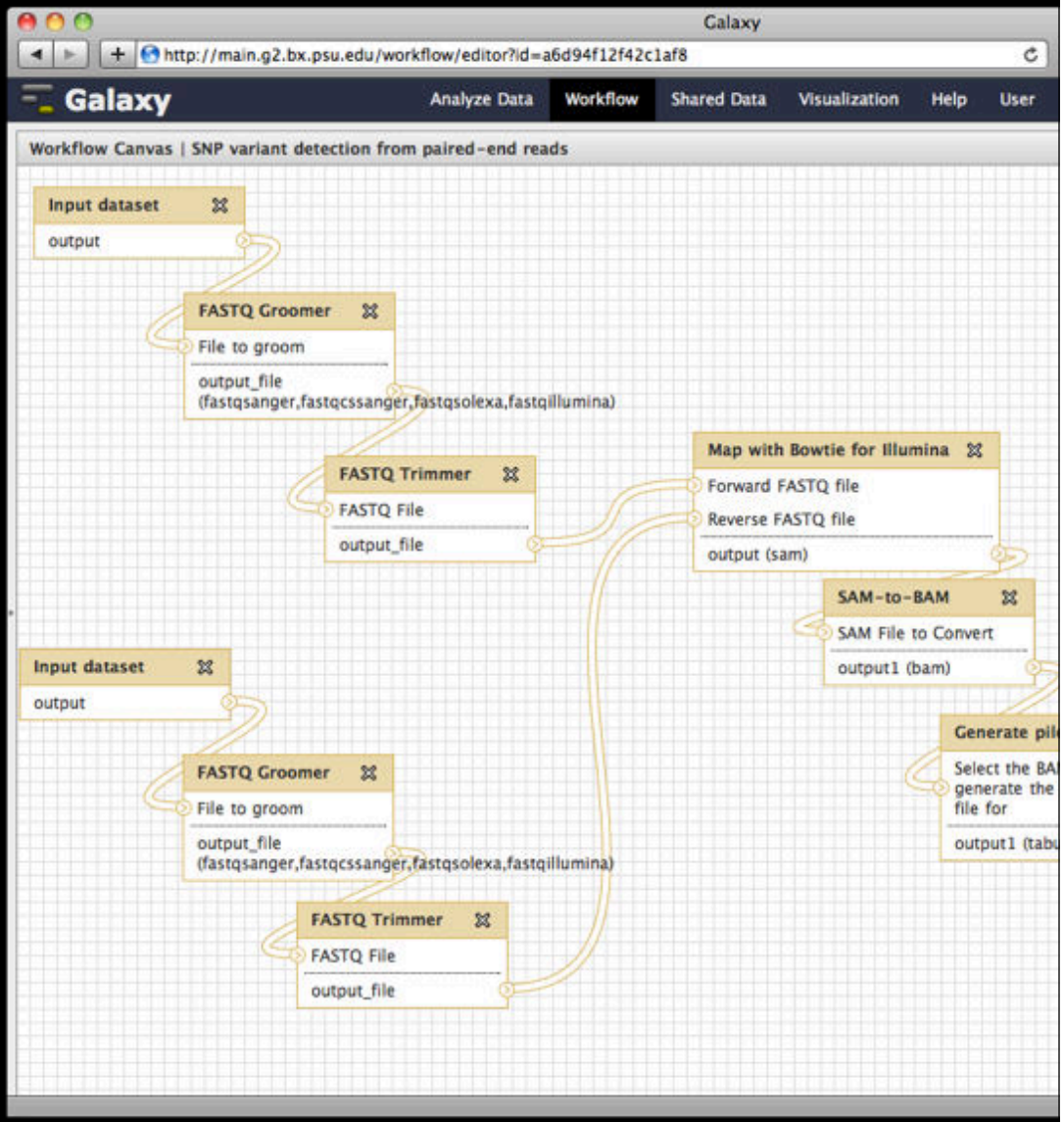


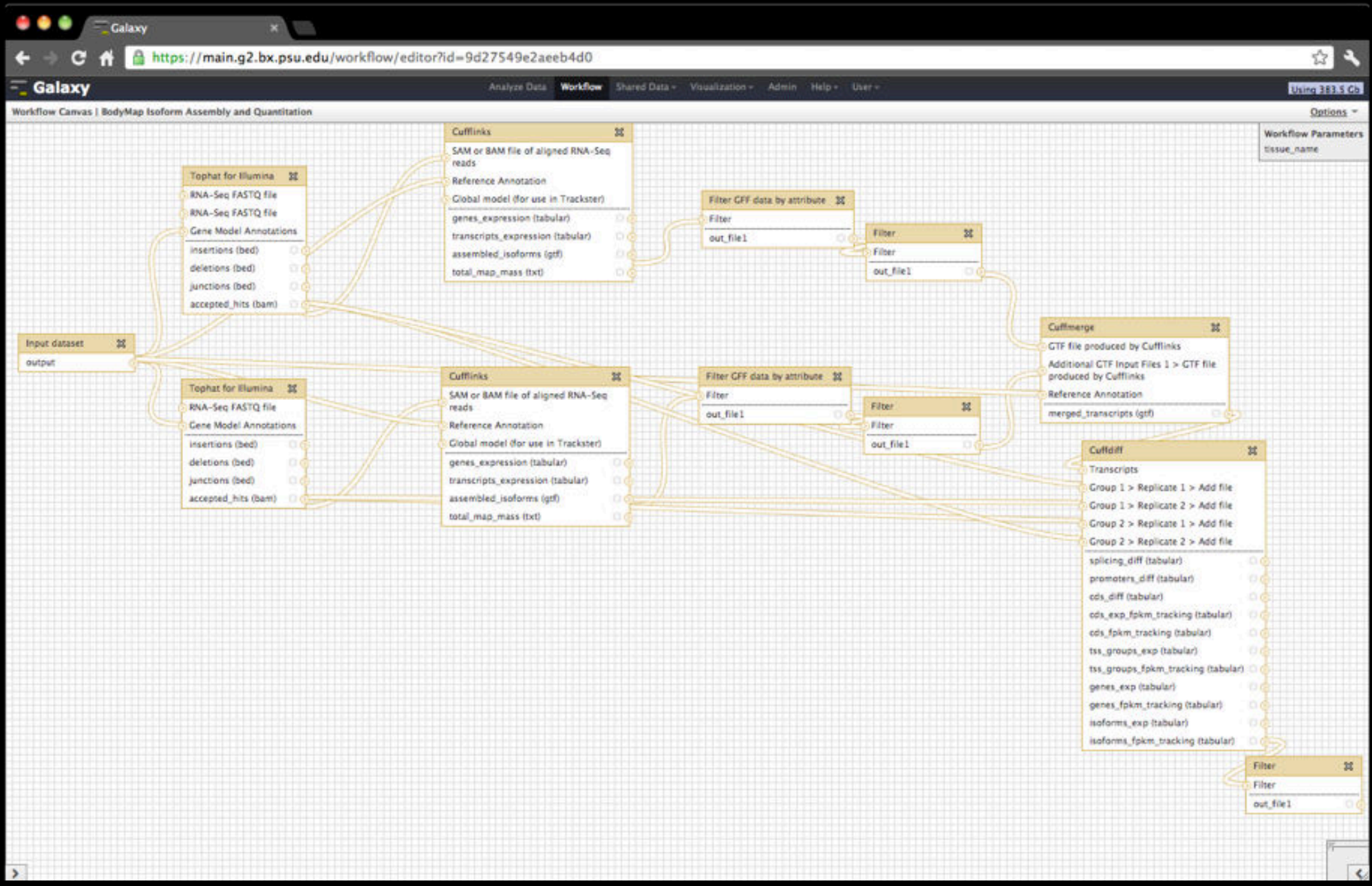
Workflows can be constructed from scratch *or* extracted from existing analysis histories

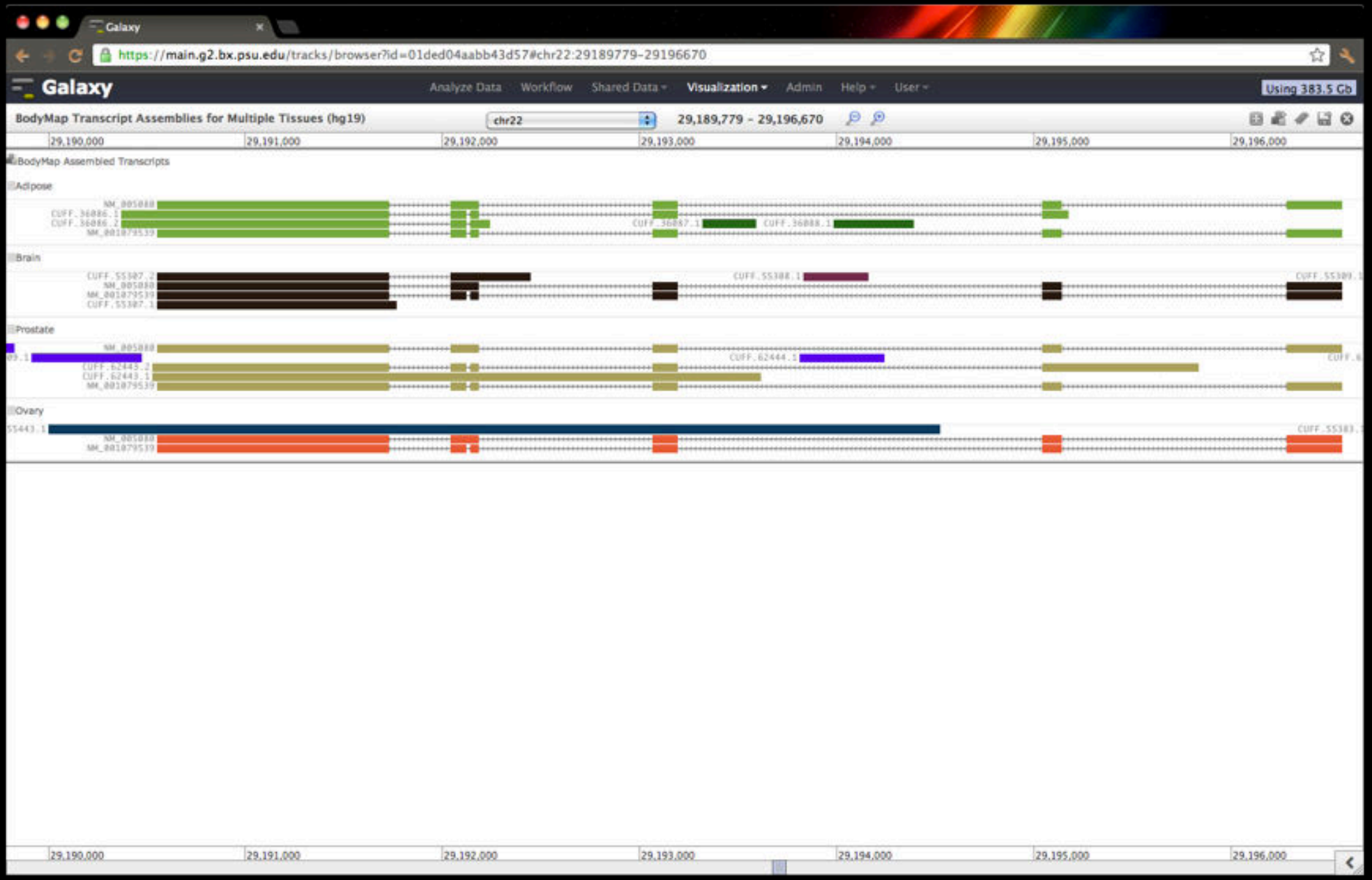
Facilitate reuse and provide precise reproducibility of a complex analysis

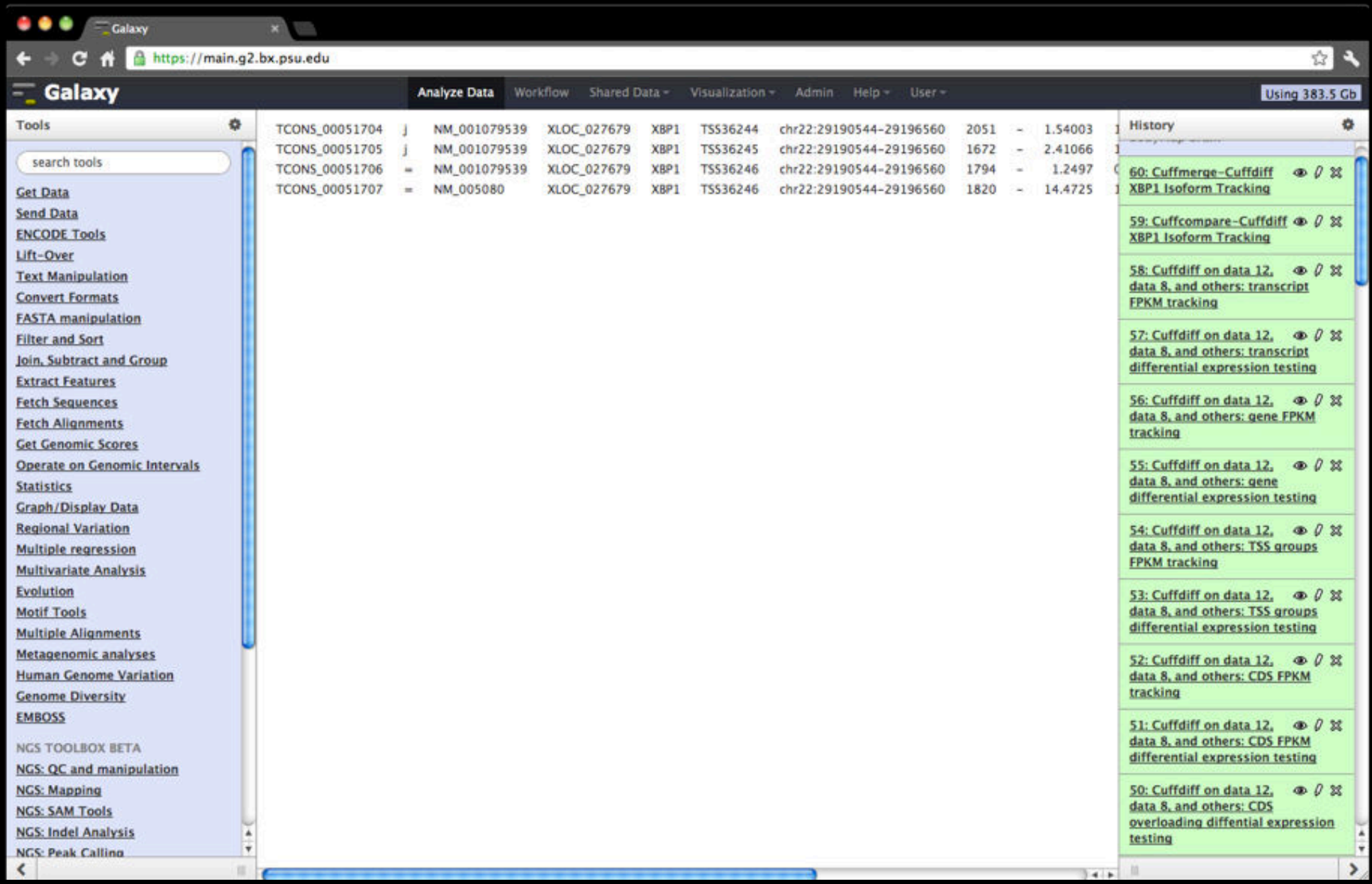
Galaxy Workflows






Galaxy
https://main.g2.bx.psu.edu/workflow/editor?id=9d27549e2aeeb4d0
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 383.5 Gb
Workflow Canvas | BodyMap Isoform Assembly and Quantitation
Options
Workflow Parameters
tissue_name
Input dataset
output
Tophat for Illumina
RNA-Seq FASTQ file
RNA-Seq FASTQ file
Gene Model Annotations
insertions (bed)
deletions (bed)
junctions (bed)
accepted_hits (bam)
Cufflinks
SAM or BAM file of aligned RNA-Seq reads
Reference Annotation
Global model (for use in Trackster)
genes_expression (tabular)
transcripts_expression (tabular)
assembled_isoforms (gtf)
total_map_mass (txt)
Filter GFF data by attribute
Filter
out_file1
Filter
Filter
out_file1
Cuffmerge
GTF file produced by Cufflinks
Additional GTF Input Files 1 > GTF file produced by Cufflinks
Reference Annotation
merged_transcripts (gtf)
Cuffdiff
Transcripts
Group 1 > Replicate 1 > Add file
Group 1 > Replicate 2 > Add file
Group 2 > Replicate 1 > Add file
Group 2 > Replicate 2 > Add file
splicing_diff (tabular)
promoters_diff (tabular)
cds_diff (tabular)
cds_exp_fpkm_tracking (tabular)
cds_fpkm_tracking (tabular)
tss_groups_exp (tabular)
tss_groups_fpkm_tracking (tabular)
genes_exp (tabular)
genes_fpkm_tracking (tabular)
isoforms_exp (tabular)
isoforms_fpkm_tracking (tabular)
Filter
Filter
out_file1


Galaxy
https://main.g2.bx.psu.edu/tracks/browser?id=01ded04aabb43d57#chr22:29189779-29196670
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 383.5 Gb
BodyMap Transcript Assemblies for Multiple Tissues (hg19)
chr22
29,189,779 - 29,196,670
29,190,000
29,191,000
29,192,000
29,193,000
29,194,000
29,195,000
29,196,000
BodyMap Assembled Transcripts
Adipose
Brain
Prostate
Ovary


Galaxy
https://main.g2.bx.psu.edu
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 383.5 Gb
Tools
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
Genome Diversity
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
TCONS_00051704 j NM_001079539 XLOC_027679 XBP1 TSS36244 chr22:29190544-29196560 2051 - 1.54003
TCONS_00051705 j NM_001079539 XLOC_027679 XBP1 TSS36245 chr22:29190544-29196560 1672 - 2.41066
TCONS_00051706 = NM_001079539 XLOC_027679 XBP1 TSS36246 chr22:29190544-29196560 1794 - 1.2497
TCONS_00051707 = NM_005080 XLOC_027679 XBP1 TSS36246 chr22:29190544-29196560 1820 - 14.4725
History
60: Cuffmerge-Cuffdiff XBP1 Isoform Tracking
59: Cuffcompare-Cuffdiff XBP1 Isoform Tracking
58: Cuffdiff on data 12, data 8, and others: transcript FPKM tracking
57: Cuffdiff on data 12, data 8, and others: transcript differential expression testing
56: Cuffdiff on data 12, data 8, and others: gene FPKM tracking
55: Cuffdiff on data 12, data 8, and others: gene differential expression testing
54: Cuffdiff on data 12, data 8, and others: TSS groups FPKM tracking
53: Cuffdiff on data 12, data 8, and others: TSS groups differential expression testing
52: Cuffdiff on data 12, data 8, and others: CDS FPKM tracking
51: Cuffdiff on data 12, data 8, and others: CDS FPKM differential expression testing
50: Cuffdiff on data 12, data 8, and others: CDS overloading diffential expression testing

Quantification

Read to transcript mapping difficult because read can map to multiple transcripts

FPKM = fragments per kilobase of exon per millions of reads

- normalize by exon length and sample size

1 FPKM ~ 1 transcript per cell in mouse

Results

	Adipose	Adrenal	Brain	Breast	Colon	Heart	Kidney	Liver	Lung
XBP1u	24.5 (CI: 22.0-27.0)	84.7 (77.6-91.7)	14.5 (11.9-17.0)	69.4 (63.0-75.8)	46.1 (41.0-51.1)	39.9 (34.8-45.0)	121.1 (111.2-131.0)	287.1 (253.8-320.3)	89.2 (82.2-96.3)
XBP1s	11.7 (10.1-13.3)	15.5 (13.3-17.7)	1.2 (0.8-1.7)	12.7 (10.4-14.9)	23.9 (20.8-27.0)	14.1 (11.2-16.9)	10.3 (8.7-11.8)	25.3 (20.7-29.8)	40.9 (37.2-44.7)
Non-coding Isoform (uc003aec.2)		6.9 (6.0-7.9)	1.5 (1.0-2.1)						

K	L	M	N	O	P	Q
Lymph Node	Ovary	Prostate	Skeletal Muscle	Testes	Thyroid	White Blood Cells
325.7 (300.4-350.9)	74.6 (66.9-82.4)	29.9 (27.4-32.4)	60.4 (55.0-65.8)	22.6 (20.5-24.8)	49.0 (44.9-53.1)	97.0 (88.9-105.1)
58.5 (53.3-63.7)	21.2 (18.0-24.4)	Not enough data	20.1 (17.6-22.5)	12.3 (10.8-13.8)	7.3 (5.9-8.8)	4.7 (3.4-6.0)

Can assemble *XBP1* isoforms and quantify expression

- Quick check: *XBP1u* > *XBP1s* in all tissues

Pervasive transcription of both *XBP1u* and *XBP1s* in 15 (16?) tissues

- very high in some tissues, lower in others

UPR is a continuous process rather than a discreet event

Questions

Can current RNA-seq tools be used to identify and quantify *XBP1's* highly similar isoforms?

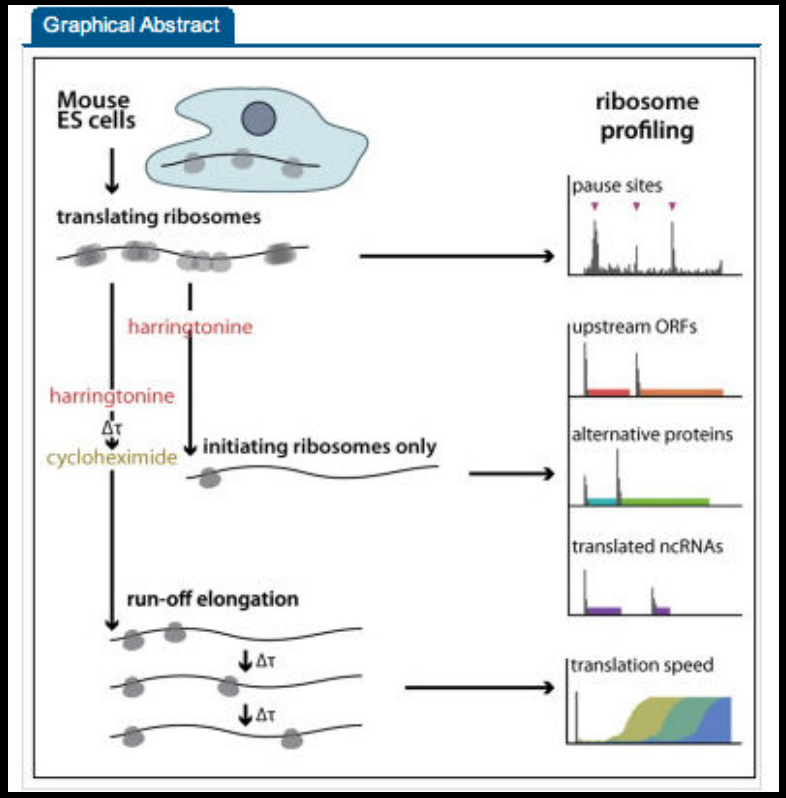
Can *XBP1's* translational pause point be identified?

Ribosome Profiling

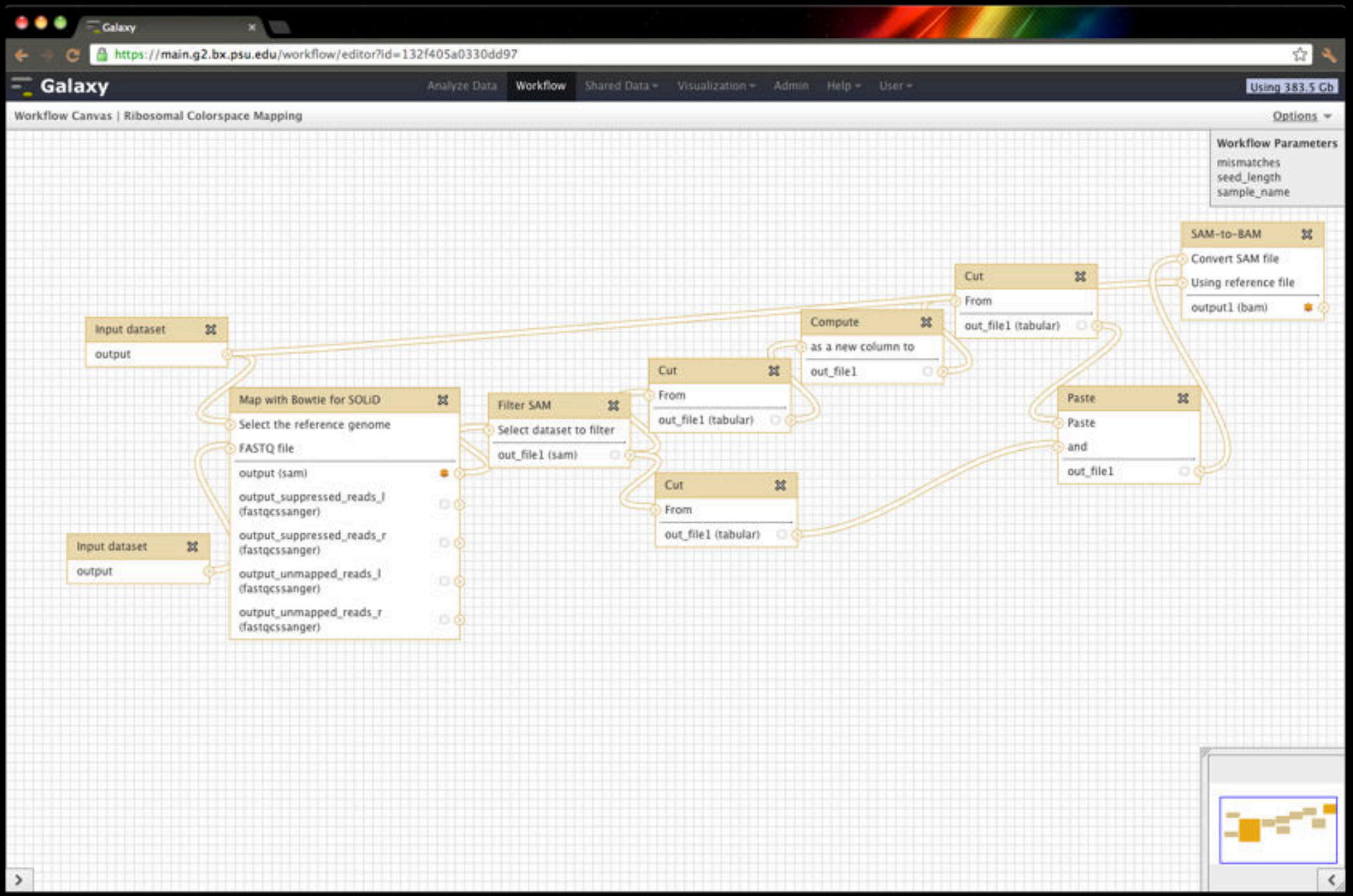
Sequence RNA undergoing translation

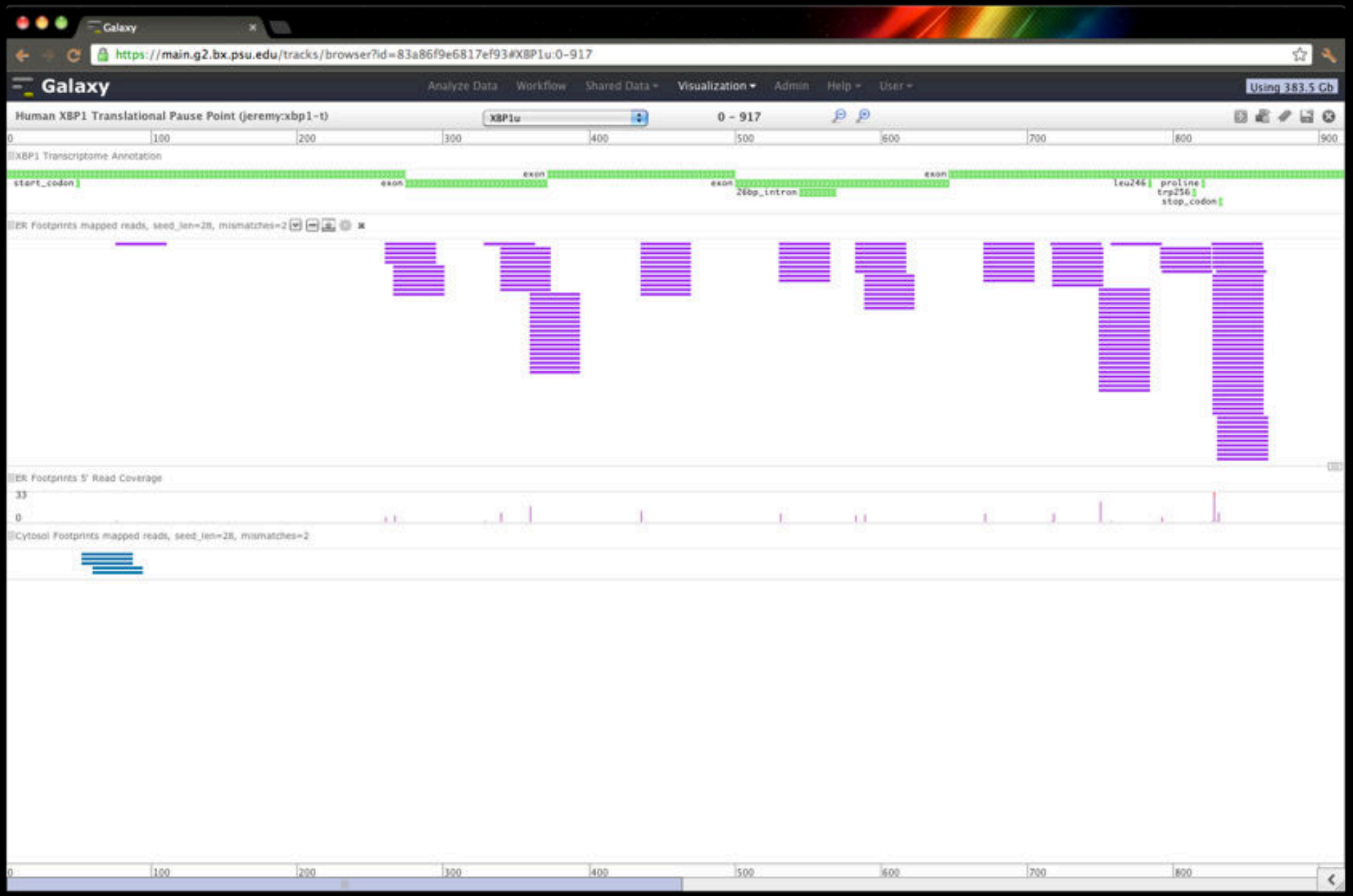
- ~35bp

In this work, data is colorspace

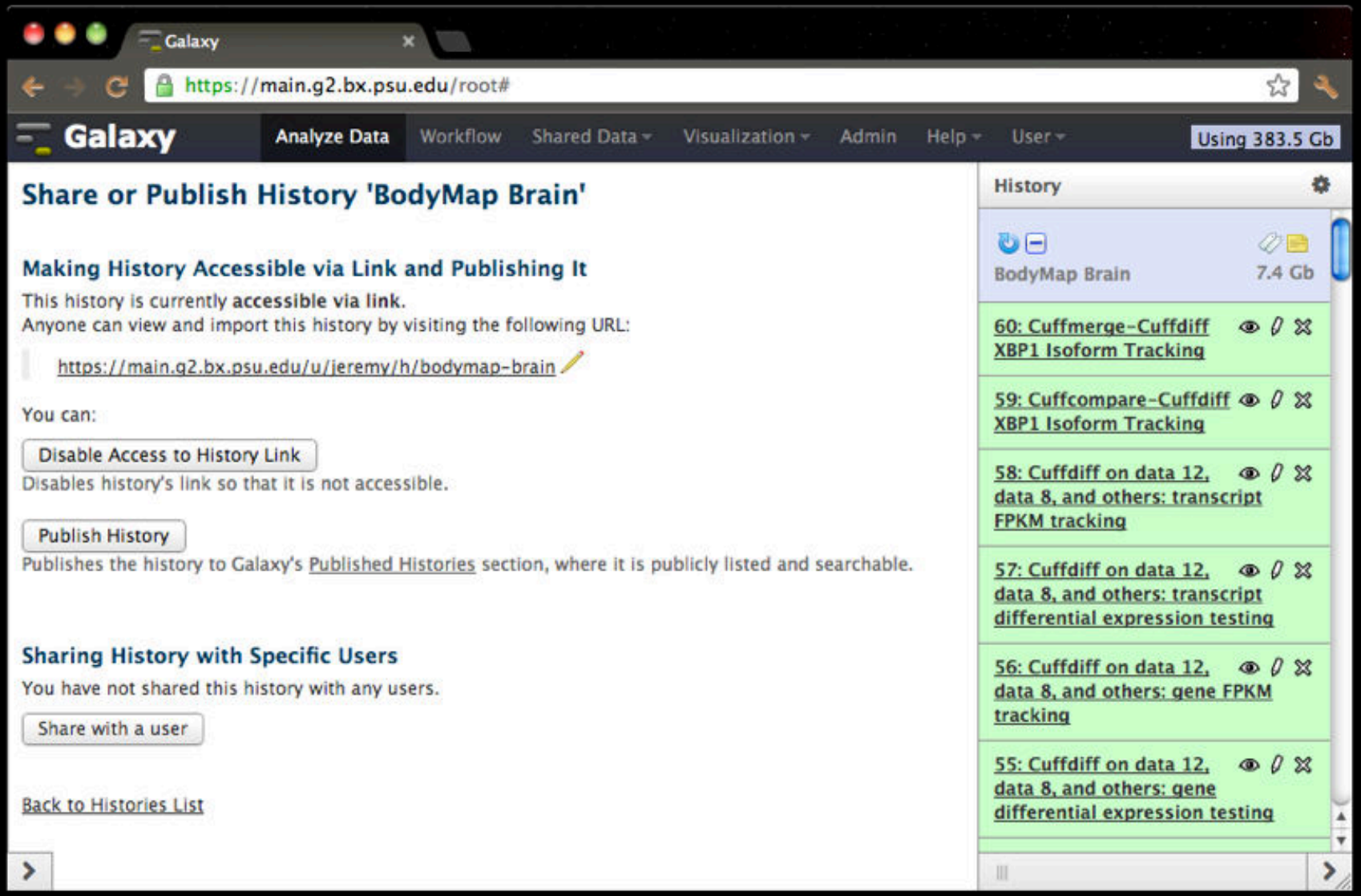


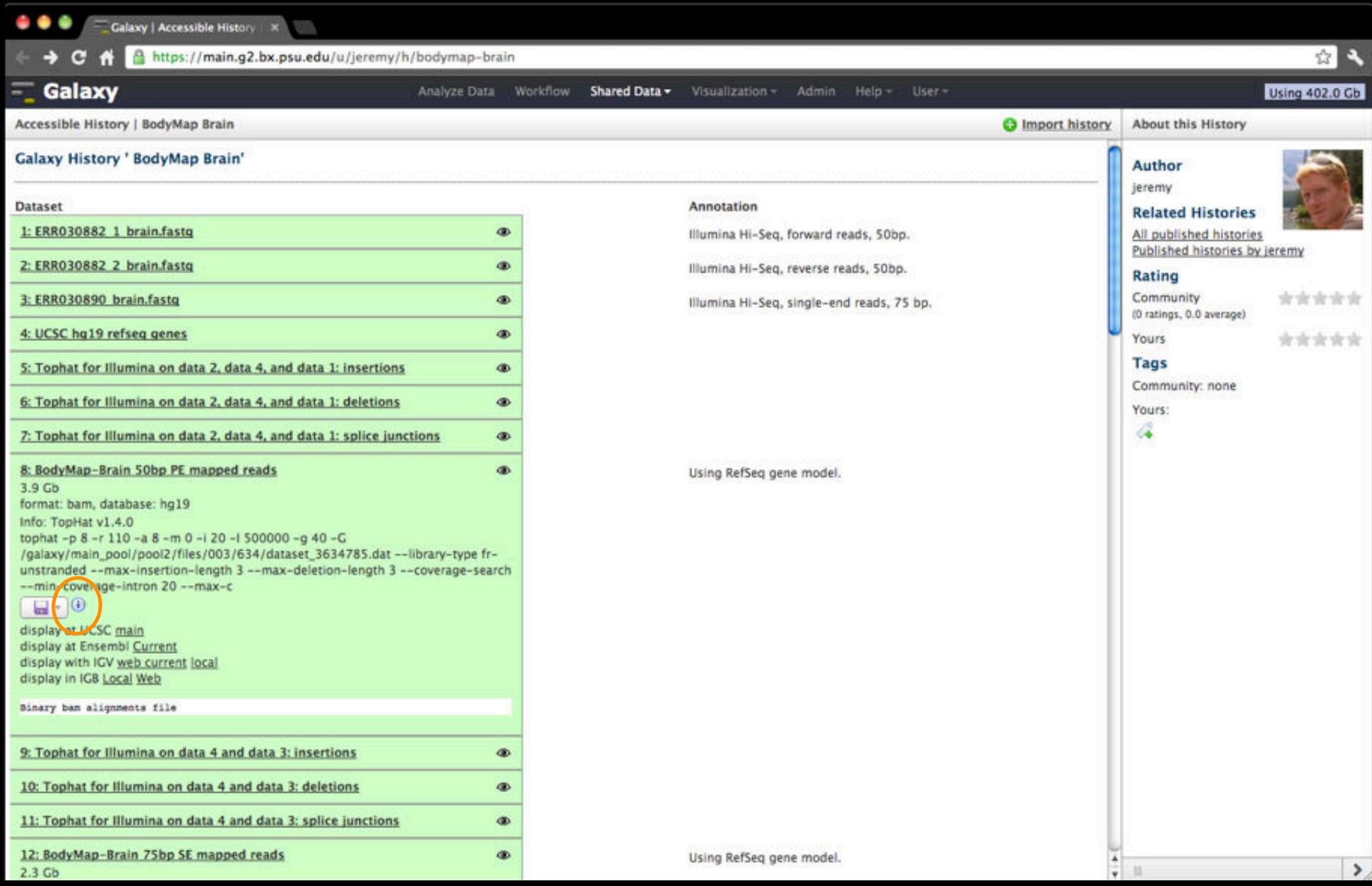
Ingolia et al., *Cell 2011*


Galaxy
https://main.g2.bx.psu.edu/workflow/editor?id=132f405a0330dd97
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 383.5 Gb
Workflow Canvas | Ribosomal Colorspace Mapping
Options
Workflow Parameters
mismatches
seed_length
sample_name
Input dataset
output
Input dataset
output
Map with Bowtie for SOLiD
Select the reference genome
FASTQ file
output (sam)
output_suppressed_reads_l (fastqcssanger)
output_suppressed_reads_r (fastqcssanger)
output_unmapped_reads_l (fastqcssanger)
output_unmapped_reads_r (fastqcssanger)
Filter SAM
Select dataset to filter
out_file1 (sam)
Cut
From
out_file1 (tabular)
Cut
From
out_file1 (tabular)
Compute
as a new column to
out_file1
Cut
From
out_file1 (tabular)
Paste
Paste
and
out_file1
SAM-to-BAM
Convert SAM file
Using reference file
output1 (bam)

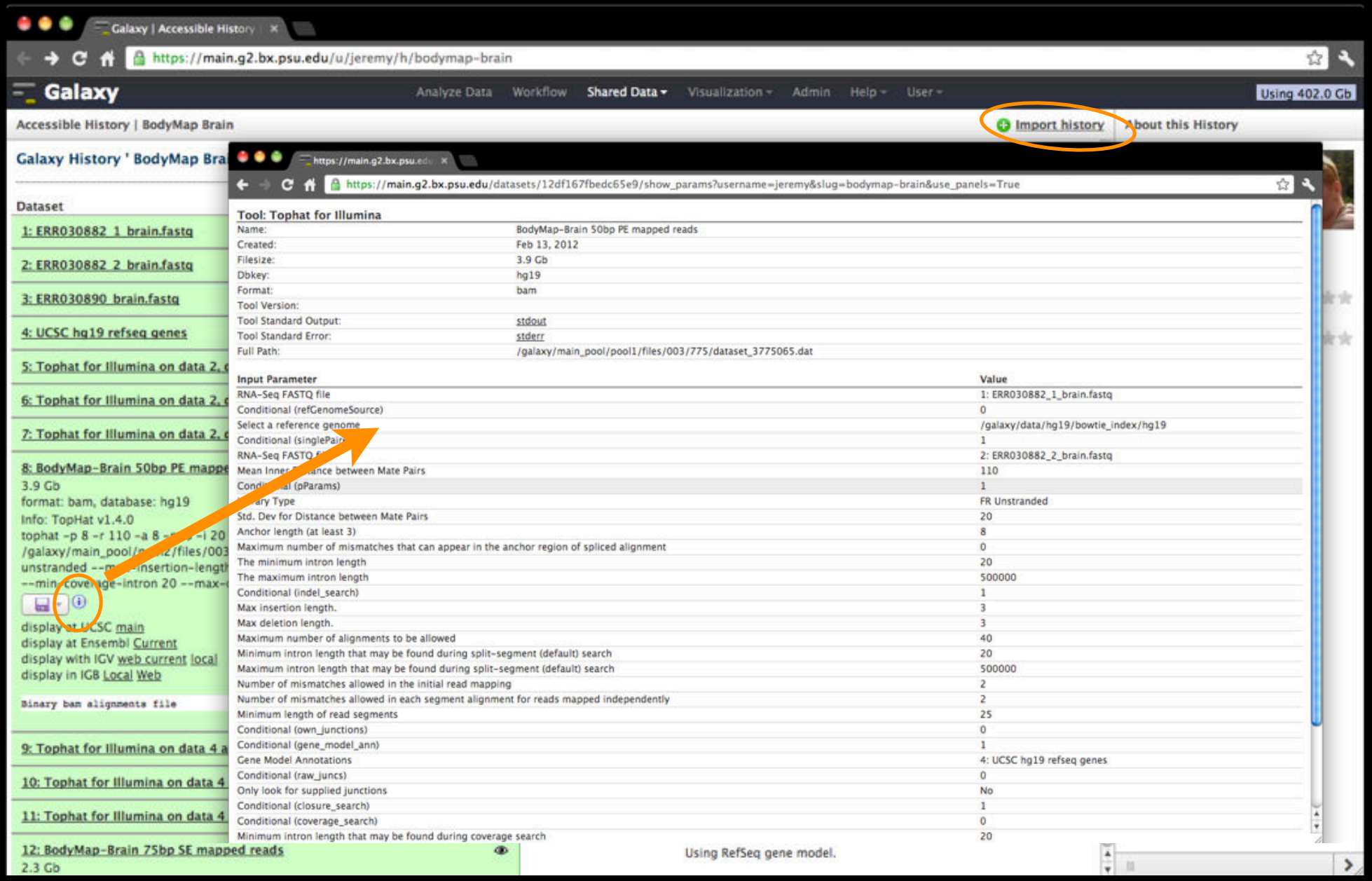

Galaxy
https://main.g2.bx.psu.edu/tracks/browser?id=83a86f9e6817ef93#XBP1u:0-917
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 383.5 Gb
Human XBP1 Translational Pause Point (jeremy:xbp1-t)
XBP1u
0 - 917
0
100
200
300
400
500
600
700
800
900
XBP1 Transcriptome Annotation
start_codon
exon
exon
exon
exon
26bp_intron
leu246
proline
trp256
stop_codon
ER Footprints mapped reads, seed_len=28, mismatches=2
ER Footprints 5' Read Coverage
33
0
Cytosol Footprints mapped reads, seed_len=28, mismatches=2

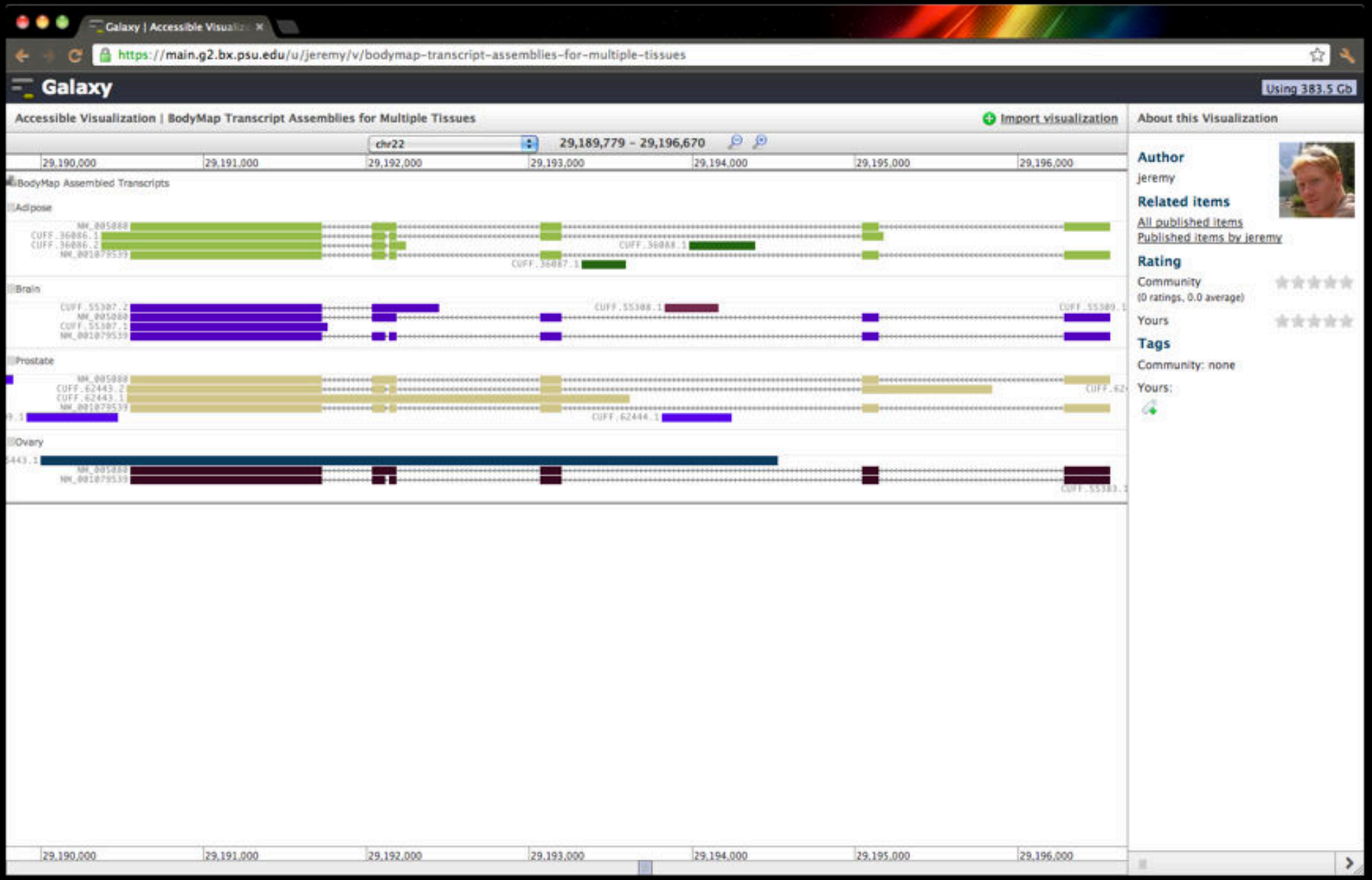
Everything can be Shared

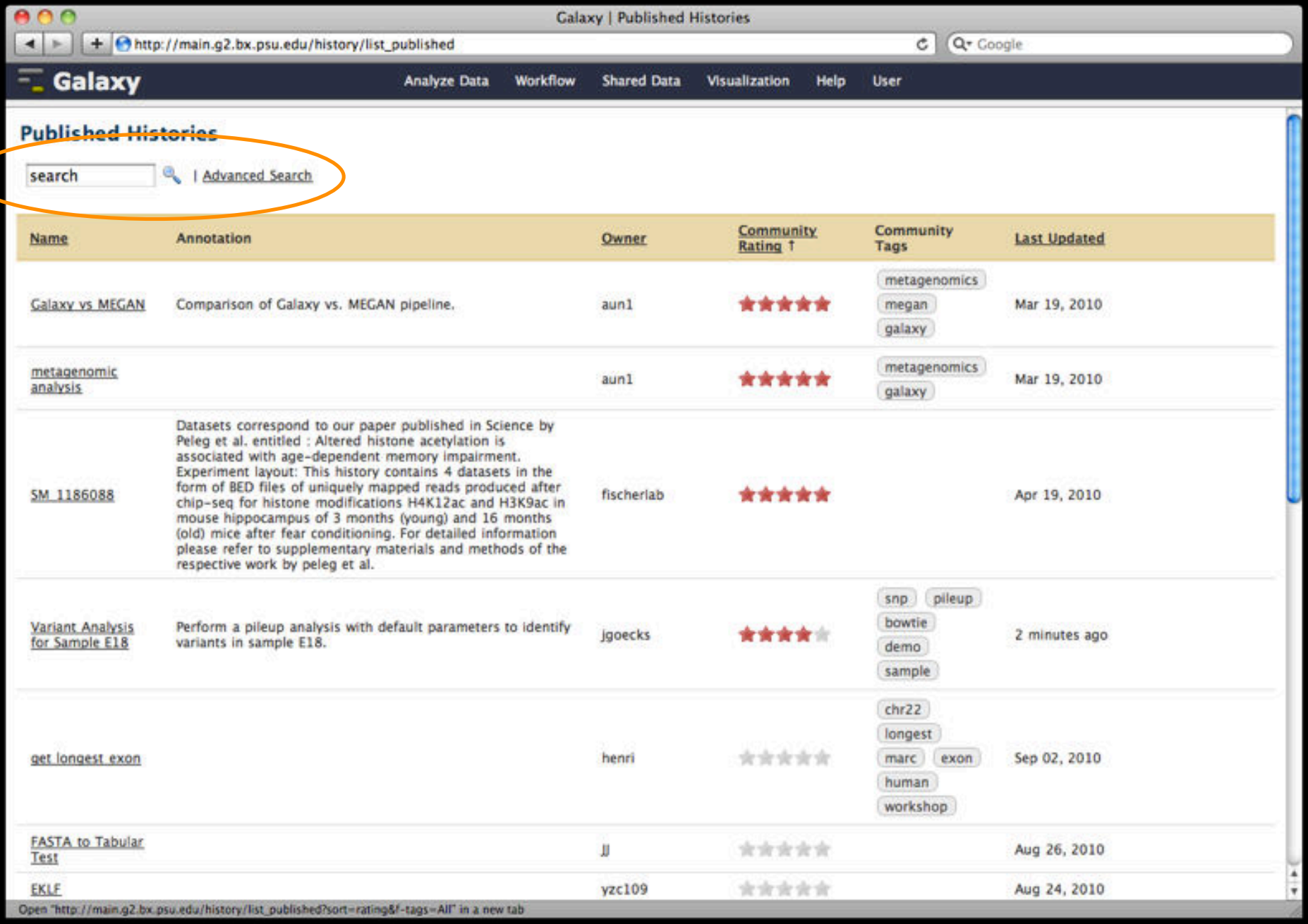



Galaxy | Accessible History
https://main.g2.bx.psu.edu/u/jeremy/h/bodymap-brain
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 402.0 Gb
Accessible History | BodyMap Brain
Import history
About this History
Galaxy History ' BodyMap Brain'
Dataset
Annotation
1: ERR030882_1_brain.fastq
Illumina Hi-Seq, forward reads, 50bp.
2: ERR030882_2_brain.fastq
Illumina Hi-Seq, reverse reads, 50bp.
3: ERR030890_brain.fastq
Illumina Hi-Seq, single-end reads, 75 bp.
4: UCSC hg19 refseq genes
5: Tophat for Illumina on data 2, data 4, and data 1: insertions
6: Tophat for Illumina on data 2, data 4, and data 1: deletions
7: Tophat for Illumina on data 2, data 4, and data 1: splice junctions
8: BodyMap-Brain 50bp PE mapped reads
Using RefSeq gene model.
3.9 Gb
format: bam, database: hg19
Info: TopHat v1.4.0
tophat -p 8 -r 110 -a 8 -m 0 -i 20 -I 500000 -g 40 -G
/galaxy/main_pool/pool2/files/003/634/dataset_3634785.dat --library-type fr-unstranded --max-insertion-length 3 --max-deletion-length 3 --coverage-search
--min-coverage-intron 20 --max-c
display at UCSC main
display at Ensembl Current
display with IGV web current local
display in IGB Local Web
Binary bam alignments file
9: Tophat for Illumina on data 4 and data 3: insertions
10: Tophat for Illumina on data 4 and data 3: deletions
11: Tophat for Illumina on data 4 and data 3: splice junctions
12: BodyMap-Brain 75bp SE mapped reads
Using RefSeq gene model.
2.3 Gb
Author
jeremy
Related Histories
All published histories
Published histories by jeremy
Rating
Community
(0 ratings, 0.0 average)
Yours
Tags
Community: none
Yours:

Exact Reproduction Possible




Galaxy | Accessible Visualiz
https://main.g2.bx.psu.edu/u/jeremy/v/bodymap-transcript-assemblies-for-multiple-tissues
Galaxy
Using 383.5 Gb
Accessible Visualization | BodyMap Transcript Assemblies for Multiple Tissues
Import visualization
About this Visualization
chr22
29,189,779 - 29,196,670
29,190,000
29,191,000
29,192,000
29,193,000
29,194,000
29,195,000
29,196,000
BodyMap Assembled Transcripts
Adipose
Brain
Prostate
Ovary
Author
jeremy
Related items
All published items
Published items by jeremy
Rating
Community
(0 ratings, 0.0 average)
Yours
Tags
Community: none
Yours:


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

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 + annotations
- anyone can import your work and immediately reproduce or build on it

Three Ways to Use Galaxy

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

2. Download and run locally

3. Run on the cloud

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

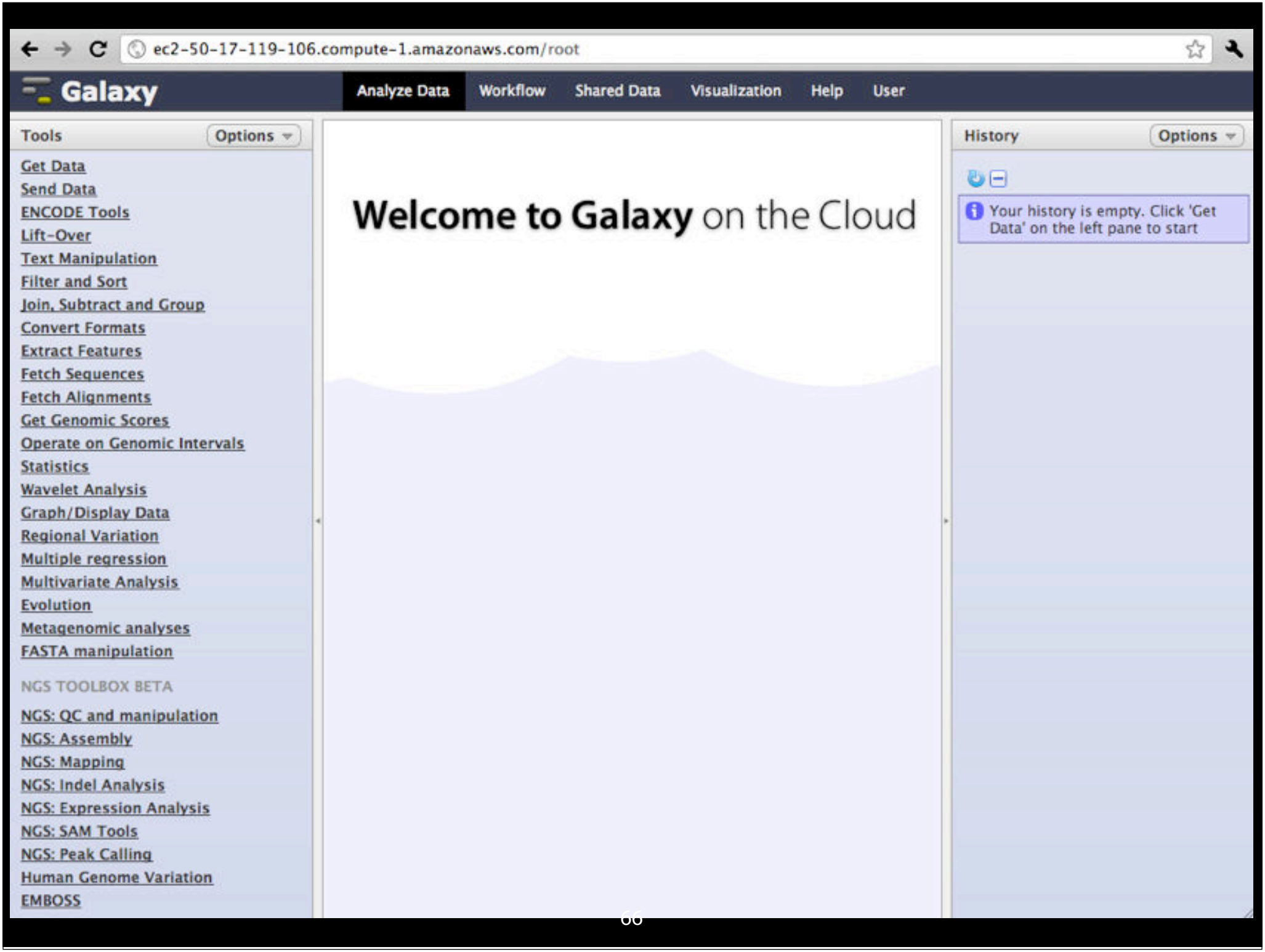
Centralized solution cannot scale to meet data analysis demands

Galaxy on the Cloud

For extended or particular resource needs

- customization necessary
- oscillating data volume

For when informatics expertise or infrastructure is limited


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

EMORY

Enis Afgan

Dannon Baker

Dave Clements

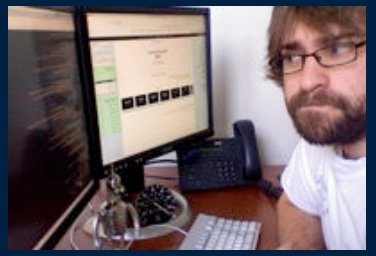
Jeremy Goecks

James Taylor

PENNSTATE.

1855

Dan Blankenberg



Nate Coraor

Jennifer Jackson

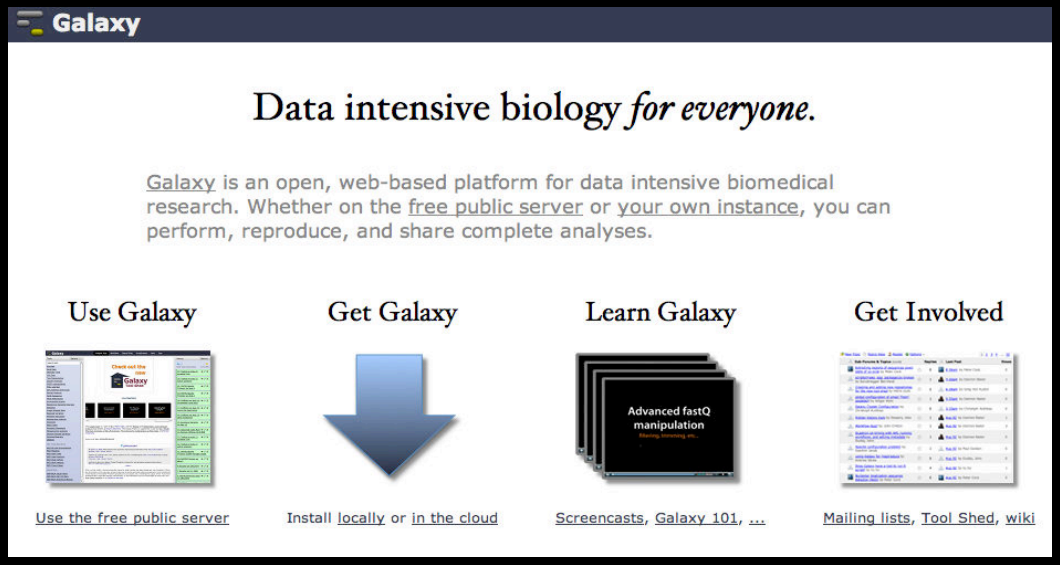
Greg von Kuster

Anton Nekrutenko

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

Thanks! Questions?

http://galaxyproject.org



Galaxy publications: http://galaxyproject.org/wiki/Citing

Galaxy is hiring! http://galaxyproject.org/wiki/GalaxyisHiring

jeremy.goecks@emory.edu