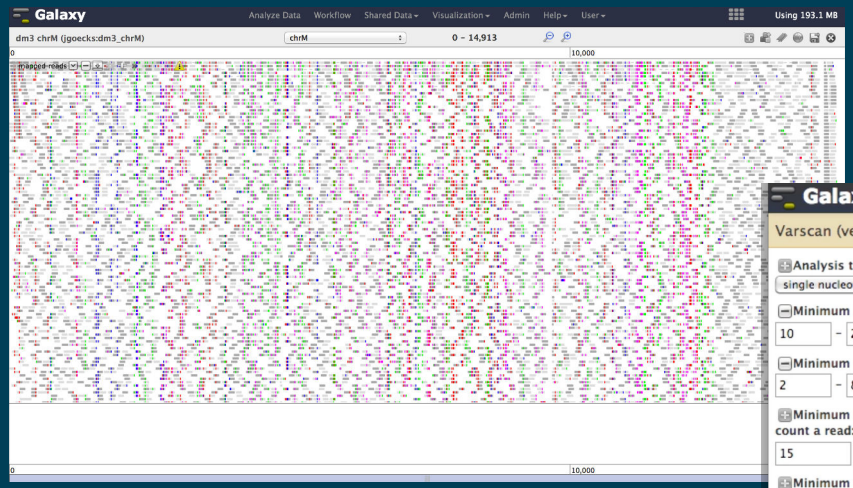
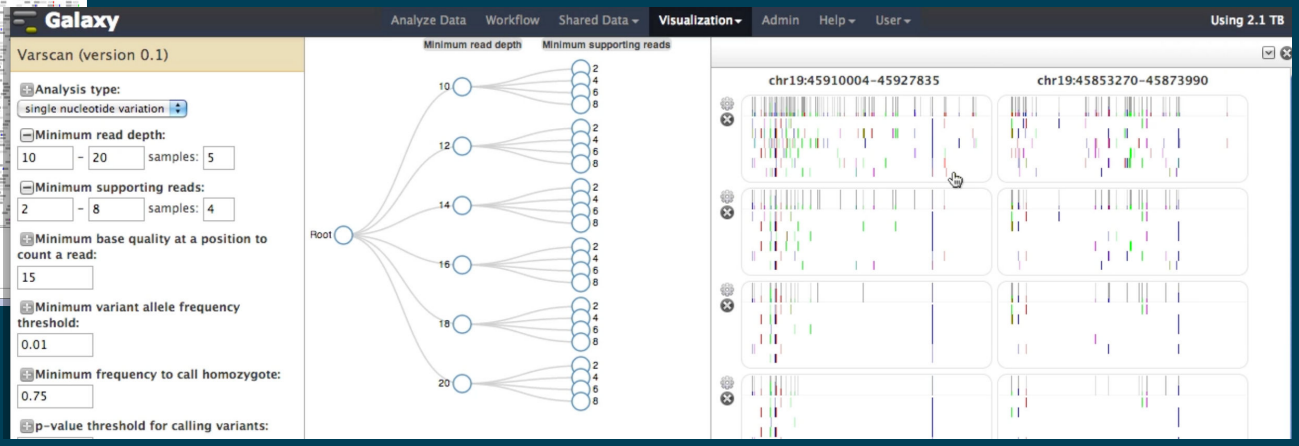




Web-Based Visualization and Visual Analysis for High-Throughput Genomics with Galaxy

Jeremy Goecks

Computational Biology Institute

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Topics

Galaxy

Visualization framework

Large-scale visualization

Integrated visual analysis

Galaxy Project: Fundamental Questions

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

- make tools and workflows **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 collaborative computational genomics

Goecks et al. (2010) *Genome Biology*

Galaxy Demo

Galaxy

tachylite01.bx.mathcs.emory.edu/g/jeremy/root

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 2.1 TB

Tools

search tools

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
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
NGS: GATK Tools (beta)
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis

Hello world! It's running...

To customize this page edit `static/welcome.html`

This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

History

PanCan P3
9.3 GB

22: Tophat Fusion Post on data 10 and data 14: html results
21: Tophat Fusion Post on data 10 and data 14: text results
20: MarkDups_Dupes Marked.html
19: MarkDups_Dupes Marked.bam
17: Cufflinks on data 9 and data 4: assembled transcripts
16: Cufflinks on data 9 and data 4: transcript expression
15: Cufflinks on data 9 and data 4: gene expression
14: Tophat2 on data 3, data 4, and data 2: accepted_hits
13: Tophat2 on data 3, data 4, and data 2: splice junctions
12: Tophat2 on data 3, data 4, and data 2: deletions
11: Tophat2 on data 3, data 4, and data 2: insertions

What is Galaxy?

Platform for high-throughput genomics

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

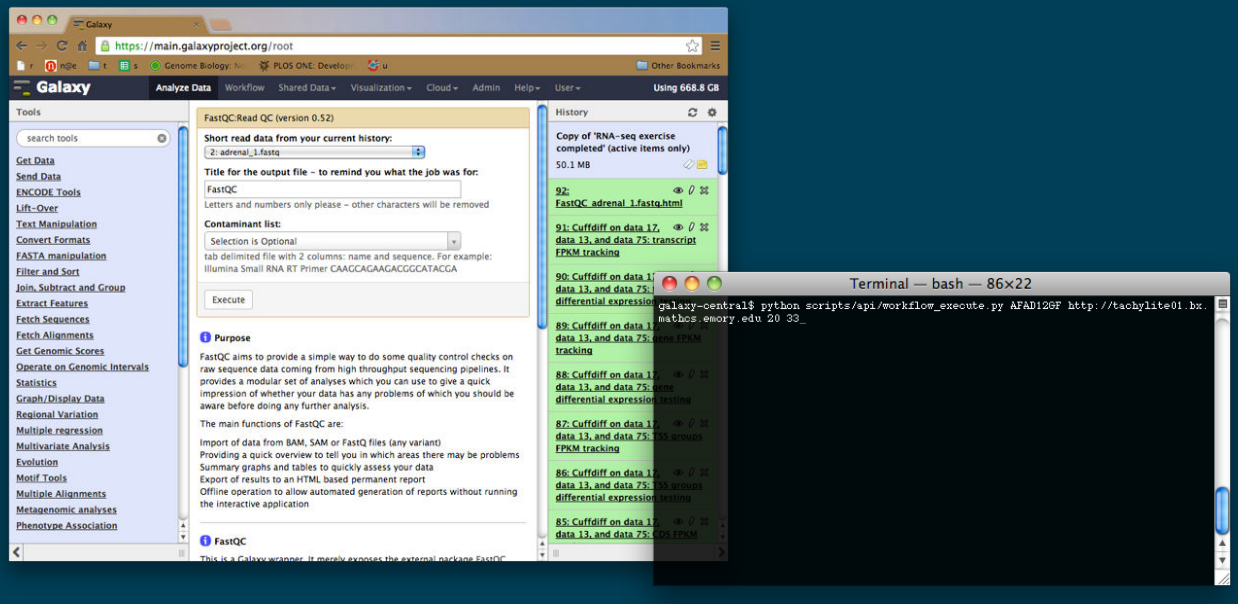
Customizable open-source software for various HPC resources

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

Galaxy

RESTful API

Plug-in framework



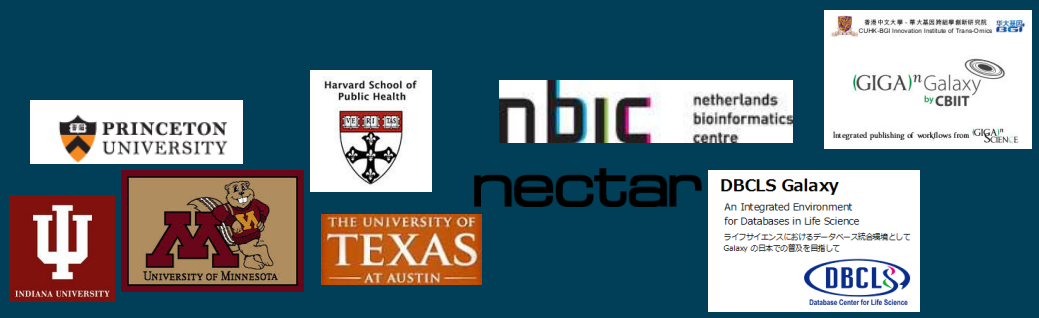
Galaxy is Very Popular

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

- ~500 new users per month, ~200 TB of user data, ~130,000 analysis jobs per month
- 7000 visualizations created

Used and cited in more than 1000 publications

50+ local servers all over the world (http://bit.ly/gxyservers)



Topics

Galaxy

Visualization framework

Large-scale visualization

Integrated visual analysis

Opportunities and Challenges for Web-based Visualization

Benefits

- no data or software downloads
- the Web is built for sharing

Challenges

- New client/server infrastructure required
- Scaling, scaling, scaling

Galaxy Visualizations

Visualizations first-class objects in Galaxy, just like tools

A visualization can be added to Galaxy via a configuration file that specifies:

- datasets that can be used
- location of visualization code (client-side or on server)

Galaxy handles visualization integration and data management, so developer focus is on visualization

Visualizations are 1st class Galaxy objects

Can be saved and versioned for reproducibility

Have a human-readable URL for sharing a fully interactive visualization:
`http://usegalaxy.org/u/jgoecks/v/tumor-mutations`

Can embed interactive visualizations in online supplementary materials via Galaxy Pages

Galaxy Data Providers

Support for data management in both the Web browser (JavaScript) and on the server (Python)

Data providers make it easy to request data from datasets

- filtering and subsetting options: by chromosome, column, or custom filter
- *indexing using best-practice tools is key*

Data providers make it much easier to build new visualizations by handling much of the data fetching

Topics

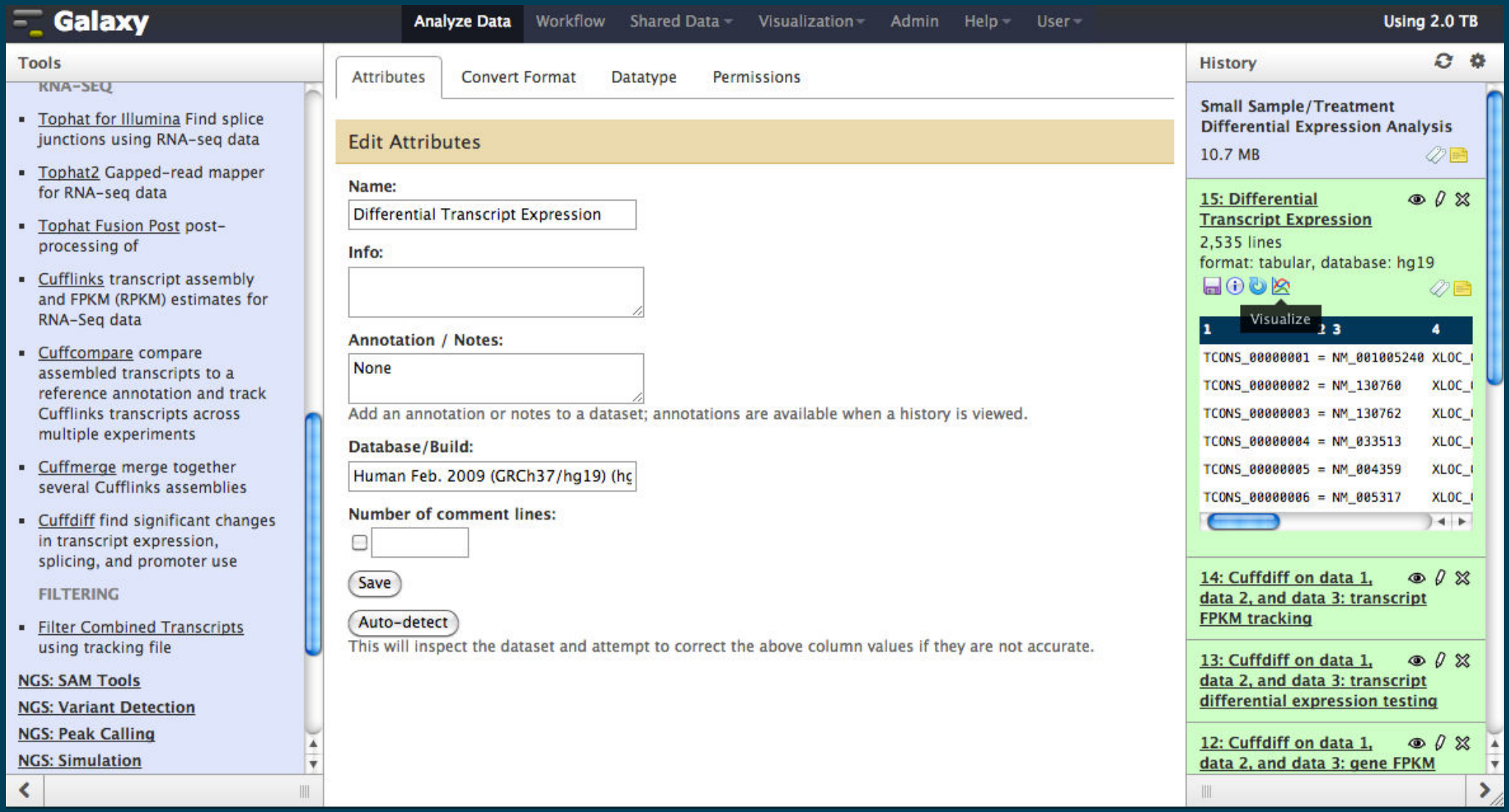
Galaxy

Visualization framework

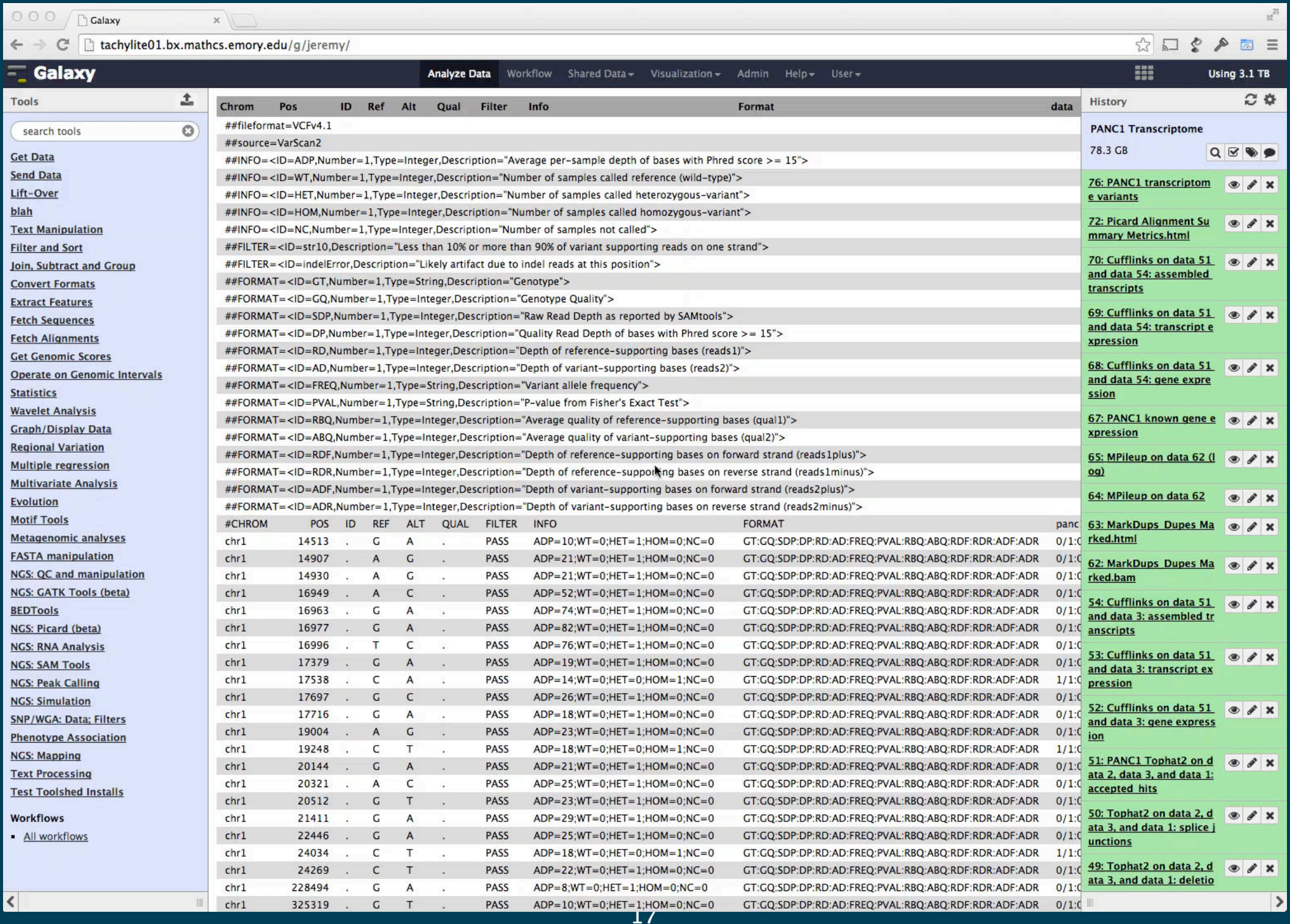
Large-scale visualization

Integrated visual analysis

Visualization Integration



Demo: Trackster—Galaxy's Genome Browser


Galaxy
tachylite01.bx.mathcs.emory.edu/g/jeremy/
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 3.1 TB
Tools
search tools
Get Data
Send Data
Lift-Over
blah
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
Motif Tools
Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
NGS: GATK Tools (beta)
BEDTools
NGS: Picard (beta)
NGS: RNA Analysis
NGS: SAM Tools
NGS: Peak Calling
NGS: Simulation
SNP/WGA: Data; Filters
Phenotype Association
NGS: Mapping
Text Processing
Test Toolshed Installs
Workflows
All workflows
Chrom Pos ID Ref Alt Qual Filter Info Format data
##fileformat=VCFv4.1
##source=VarScan2
##INFO=<ID=ADP,Number=1,Type=Integer,Description="Average per-sample depth of bases with Phred score >= 15">
##INFO=<ID=WT,Number=1,Type=Integer,Description="Number of samples called reference (wild-type)">
##INFO=<ID=HET,Number=1,Type=Integer,Description="Number of samples called heterozygous-variant">
##INFO=<ID=HOM,Number=1,Type=Integer,Description="Number of samples called homozygous-variant">
##INFO=<ID=NC,Number=1,Type=Integer,Description="Number of samples not called">
##FILTER=<ID=str10,Description="Less than 10% or more than 90% of variant supporting reads on one strand">
##FILTER=<ID=indelError,Description="Likely artifact due to indel reads at this position">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=SDP,Number=1,Type=Integer,Description="Raw Read Depth as reported by SAMtools">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Quality Read Depth of bases with Phred score >= 15">
##FORMAT=<ID=RD,Number=1,Type=Integer,Description="Depth of reference-supporting bases (reads1)">
##FORMAT=<ID=AD,Number=1,Type=Integer,Description="Depth of variant-supporting bases (reads2)">
##FORMAT=<ID=FREQ,Number=1,Type=String,Description="Variant allele frequency">
##FORMAT=<ID=PVAL,Number=1,Type=String,Description="P-value from Fisher's Exact Test">
##FORMAT=<ID=RBQ,Number=1,Type=Integer,Description="Average quality of reference-supporting bases (qual1)">
##FORMAT=<ID=ABQ,Number=1,Type=Integer,Description="Average quality of variant-supporting bases (qual2)">
##FORMAT=<ID=RDF,Number=1,Type=Integer,Description="Depth of reference-supporting bases on forward strand (reads1plus)">
##FORMAT=<ID=RDR,Number=1,Type=Integer,Description="Depth of reference-supporting bases on reverse strand (reads1minus)">
##FORMAT=<ID=ADF,Number=1,Type=Integer,Description="Depth of variant-supporting bases on forward strand (reads2plus)">
##FORMAT=<ID=ADR,Number=1,Type=Integer,Description="Depth of variant-supporting bases on reverse strand (reads2minus)">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT panc
chr1 14513 . G A . PASS ADP=10;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 14907 . A G . PASS ADP=21;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 14930 . A G . PASS ADP=21;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 16949 . A C . PASS ADP=52;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 16963 . G A . PASS ADP=74;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 16977 . G A . PASS ADP=82;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 16996 . T C . PASS ADP=76;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 17379 . G A . PASS ADP=19;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 17538 . C A . PASS ADP=14;WT=0;HET=0;HOM=1;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 1/1:0
chr1 17697 . G C . PASS ADP=26;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 17716 . G A . PASS ADP=18;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 19004 . A G . PASS ADP=23;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 19248 . C T . PASS ADP=18;WT=0;HET=0;HOM=1;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 1/1:0
chr1 20144 . G A . PASS ADP=21;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 20321 . A C . PASS ADP=25;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 20512 . G T . PASS ADP=23;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 21411 . G A . PASS ADP=29;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 22446 . G A . PASS ADP=25;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 24034 . C T . PASS ADP=18;WT=0;HET=0;HOM=1;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 1/1:0
chr1 24269 . C T . PASS ADP=22;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 228494 . G A . PASS ADP=8;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
chr1 325319 . G T . PASS ADP=10;WT=0;HET=1;HOM=0;NC=0 GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR 0/1:0
History
PANC1 Transcriptome
78.3 GB
76: PANC1 transcriptome variants
72: Picard Alignment Summary Metrics.html
70: Cufflinks on data 51 and data 54: assembled transcripts
69: Cufflinks on data 51 and data 54: transcript expression
68: Cufflinks on data 51 and data 54: gene expression
67: PANC1 known gene expression
65: MPileup on data 62 (log)
64: MPileup on data 62
63: MarkDups Dupes Marked.html
62: MarkDups Dupes Marked.bam
54: Cufflinks on data 51 and data 3: assembled transcripts
53: Cufflinks on data 51 and data 3: transcript expression
52: Cufflinks on data 51 and data 3: gene expression
51: PANC1 Tophat2 on data 2, data 3, and data 1: accepted_hits
50: Tophat2 on data 2, data 3, and data 1: splice junctions
49: Tophat2 on data 2, data 3, and data 1: deletio

Trackster—Galaxy's Genome Browser

Genome browsers are a foundational tool for more complex visualization and analysis

Addresses challenges in the high-throughput sequencing era

- supports many very large datasets
- data rendered on client to provide for maximal customization

Demo: Scaling to Large Datasets

Saved Visualizations

Create new visualization

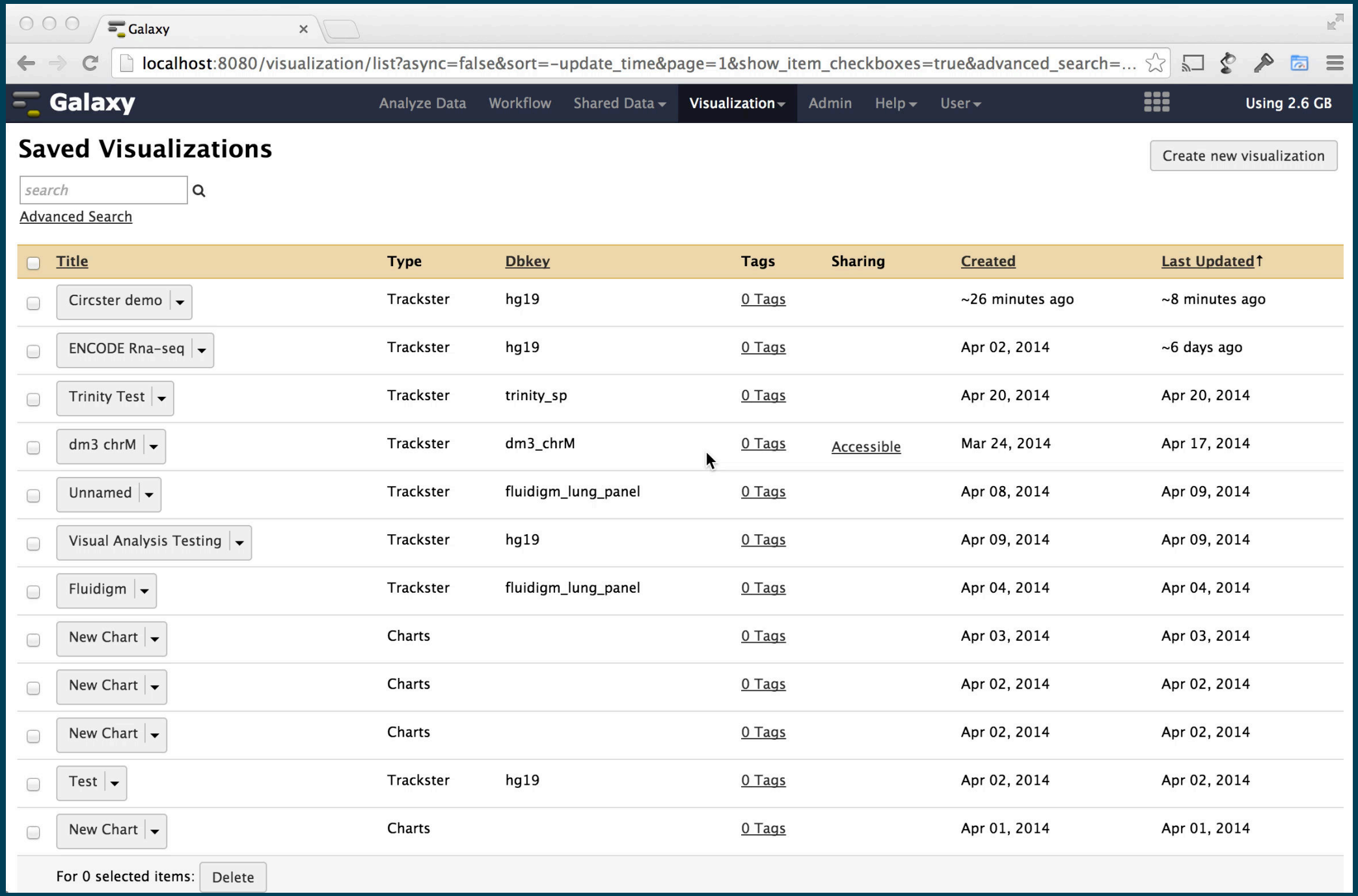
search

Advanced Search

Title	Type	Dbkey	Tags	Sharing	Created	Last Updated↑
PAAC Cell Lines	Trackster	hg19	0 Tags		~5 hours ago	~8 minutes ago
Heteroplasmy	Trackster	chrM	0 Tags		Apr 02, 2014	~1 day ago
PANC1	Trackster	hg19	0 Tags		~2 days ago	~2 days ago
S1	Trackster	hg19	0 Tags		~3 days ago	~3 days ago
Lung Sample 1	Trackster	fluidym_lung_panel	0 Tags		Apr 04, 2014	Apr 04, 2014
P1	Trackster	hg19	0 Tags		Sep 05, 2013	Apr 02, 2014
Patient Mutations	Trackster	hg19	0 Tags	Accessible	May 20, 2013	Mar 09, 2014
Freebayes vs. VarScan	Trackster	hg19	0 Tags		Feb 27, 2014	Feb 27, 2014
AML Deleterious Mutations	Trackster	hg19	0 Tags		Feb 25, 2014	Feb 25, 2014
MiaPaCa2	Trackster	hg19	0 Tags		Feb 07, 2014	Feb 11, 2014
Mutations in Expressed Genes	Trackster	hg19	0 Tags		Feb 07, 2014	Feb 07, 2014
Tumor 2	Trackster	hg19	0 Tags		Feb 07, 2014	Feb 07, 2014
Test	Trackster	hg19	0 Tags		Jan 09, 2014	Jan 10, 2014
P1	Trackster	hg19	0 Tags		Dec 16, 2013	Dec 17, 2013
MiaPaCa2	Trackster	hg19	0 Tags		Nov 18, 2013	Nov 18, 2013

For 0 selected items: Delete

Demo: Circster—An Interactive Circos Plot


Galaxy
localhost:8080/visualization/list?async=false&sort=-update_time&page=1&show_item_checkboxes=true&advanced_search=...
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 2.6 GB
Saved Visualizations
Create new visualization
search
Advanced Search
Title
Type
Dbkey
Tags
Sharing
Created
Last Updated
Circster demo
Trackster
hg19
0 Tags
~26 minutes ago
~8 minutes ago
ENCODE Rna-seq
Trackster
hg19
0 Tags
Apr 02, 2014
~6 days ago
Trinity Test
Trackster
trinity_sp
0 Tags
Apr 20, 2014
Apr 20, 2014
dm3 chrM
Trackster
dm3_chrM
0 Tags
Accessible
Mar 24, 2014
Apr 17, 2014
Unnamed
Trackster
fluidigm_lung_panel
0 Tags
Apr 08, 2014
Apr 09, 2014
Visual Analysis Testing
Trackster
hg19
0 Tags
Apr 09, 2014
Apr 09, 2014
Fluidigm
Trackster
fluidigm_lung_panel
0 Tags
Apr 04, 2014
Apr 04, 2014
New Chart
Charts
0 Tags
Apr 03, 2014
Apr 03, 2014
New Chart
Charts
0 Tags
Apr 02, 2014
Apr 02, 2014
New Chart
Charts
0 Tags
Apr 02, 2014
Apr 02, 2014
Test
Trackster
hg19
0 Tags
Apr 02, 2014
Apr 02, 2014
New Chart
Charts
0 Tags
Apr 01, 2014
Apr 01, 2014
For 0 selected items:
Delete

Topics

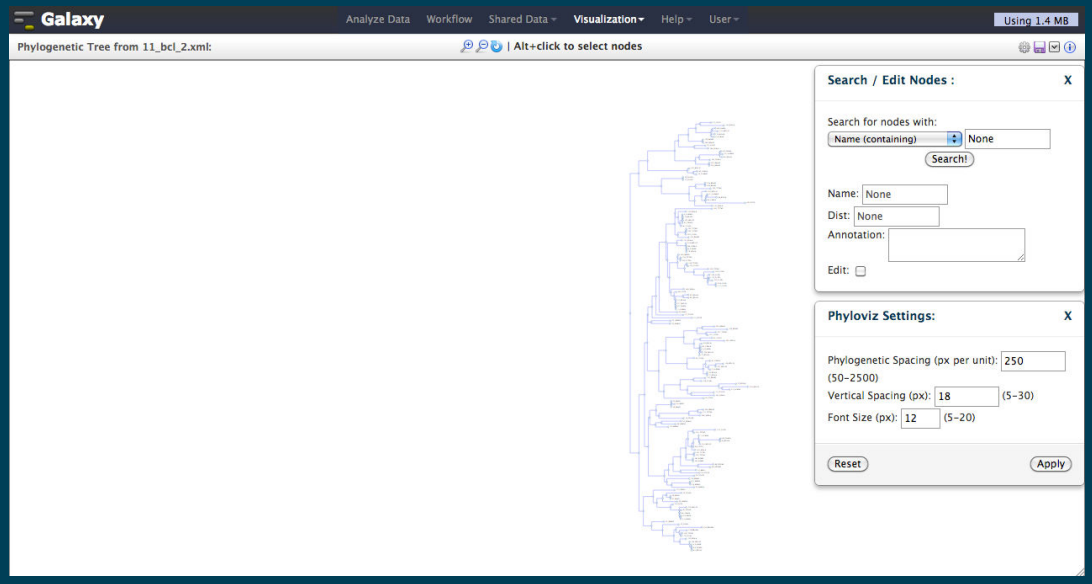
Galaxy

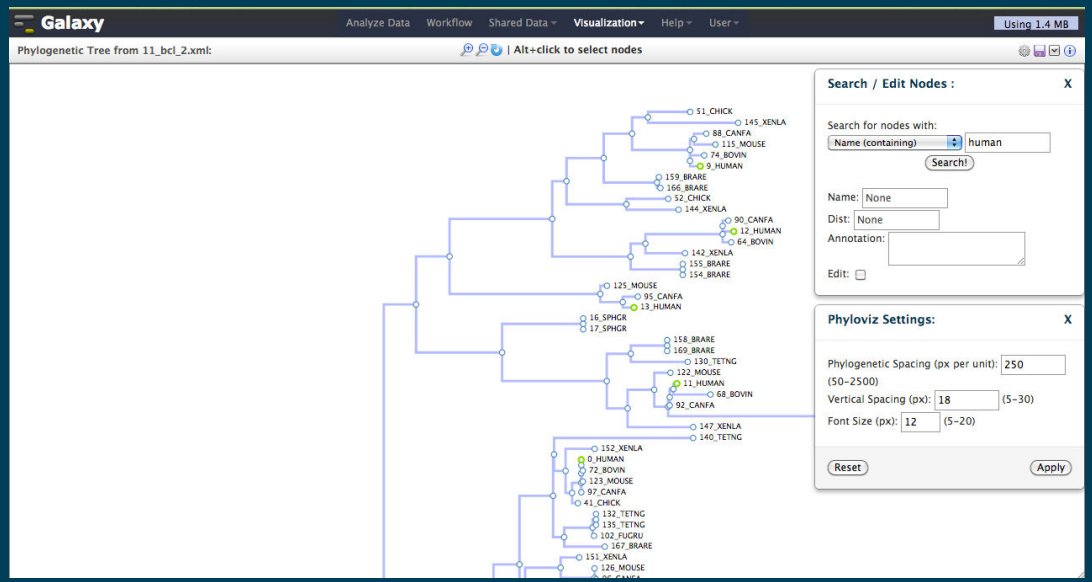
Visualization framework

Large-scale visualization

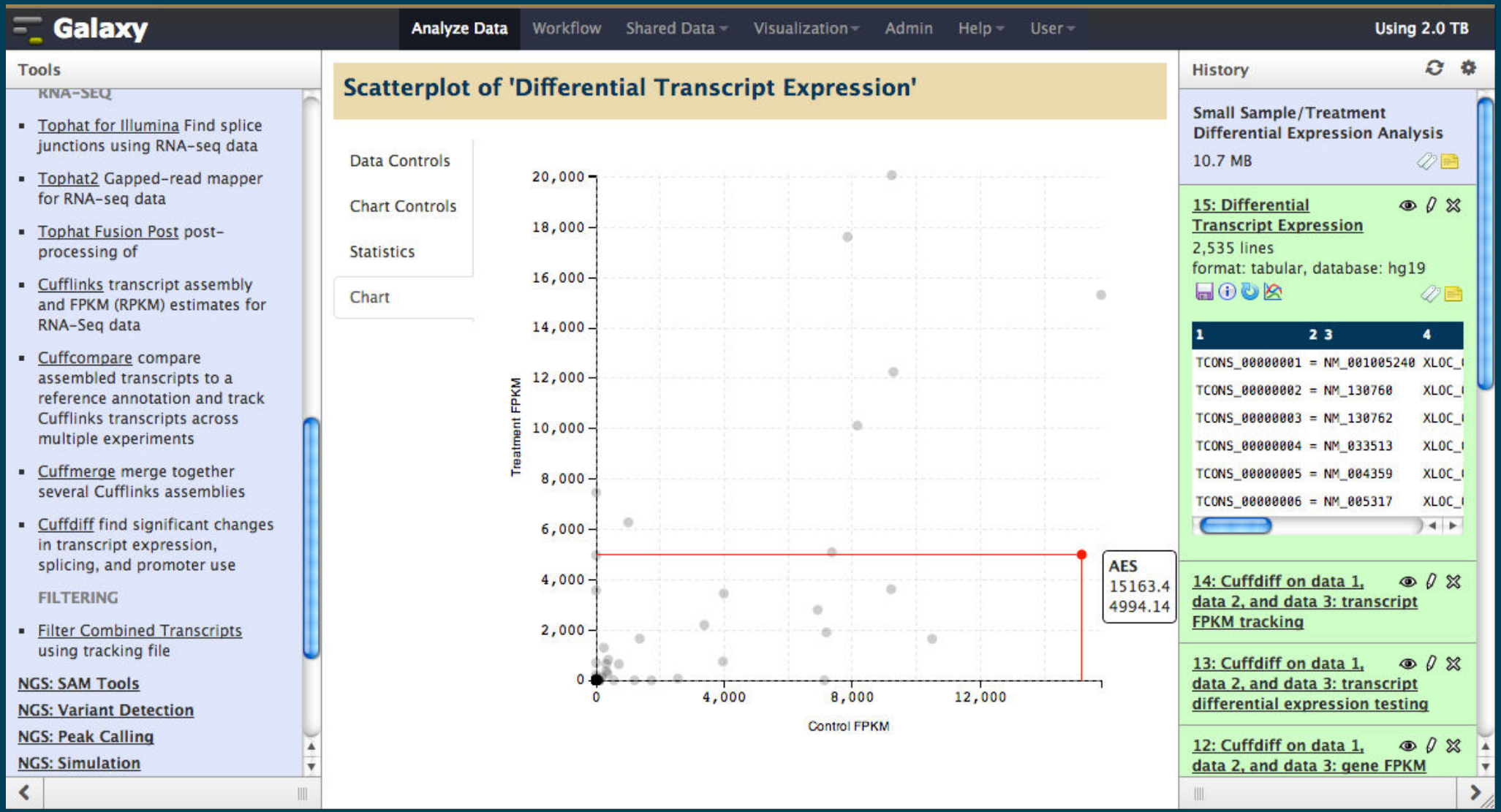
Integrated visual analysis

PhyloViz Tree Viewer

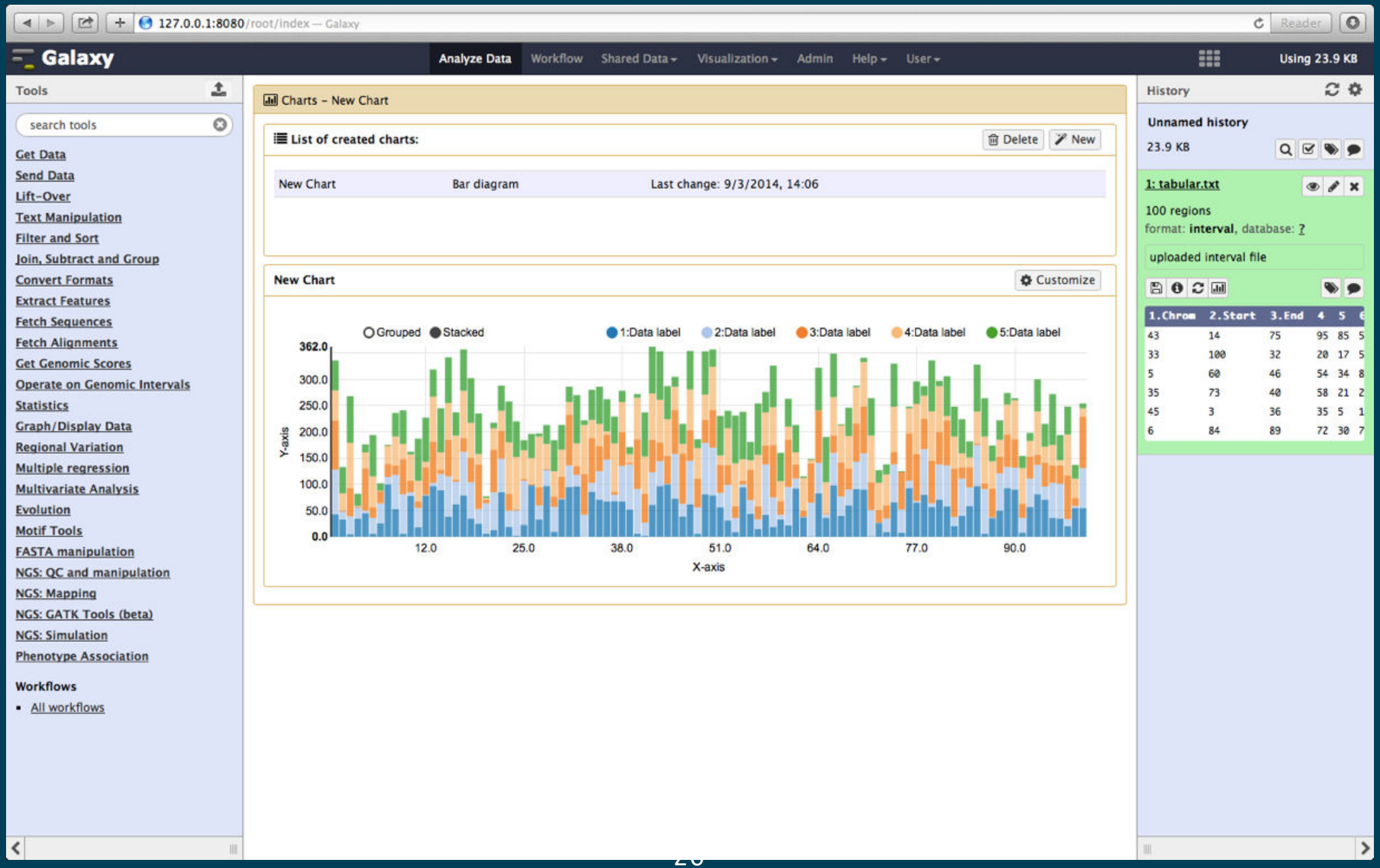




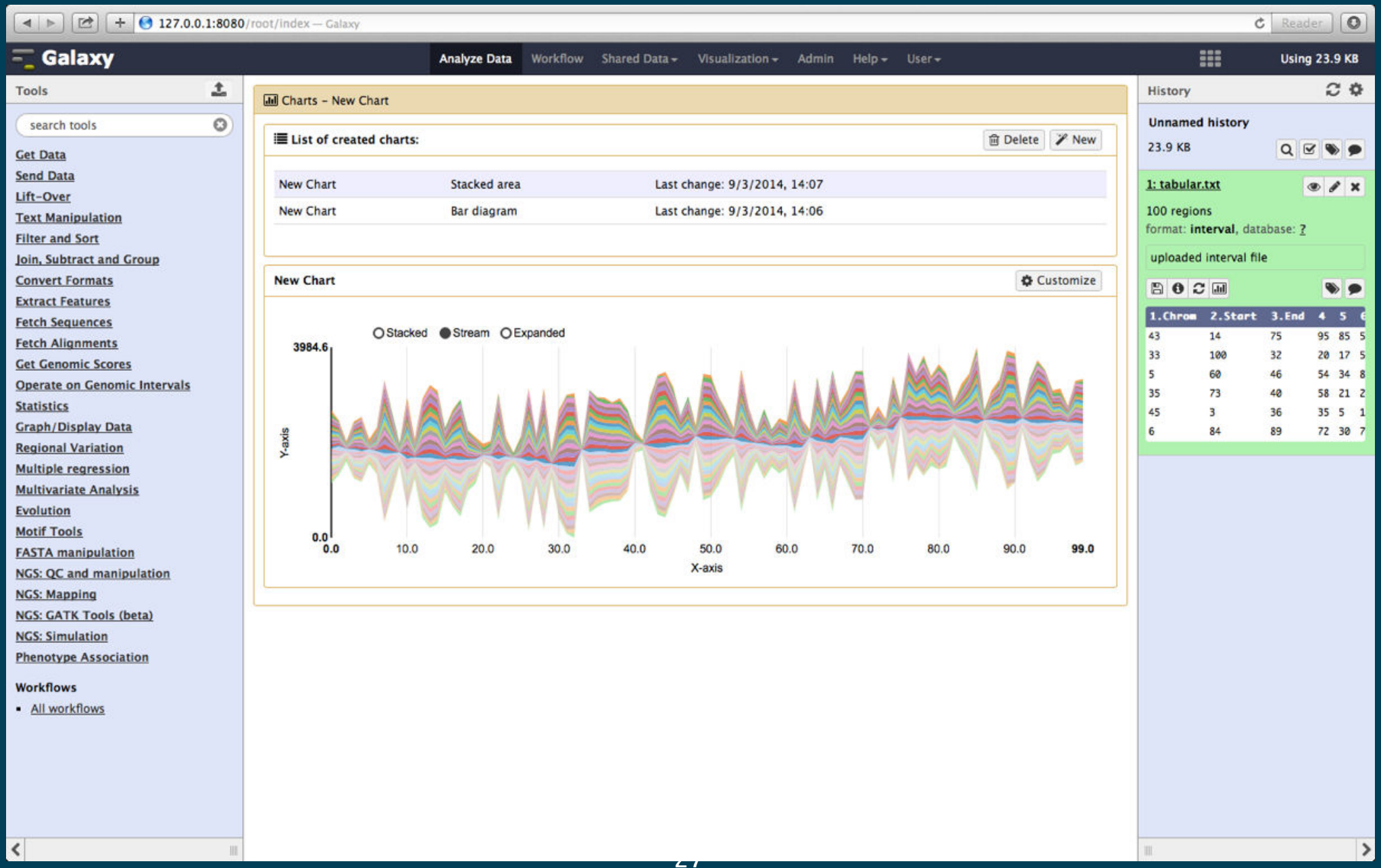
Scatterplot



NVD3 Charts


127.0.0.1:8080/root/index — Galaxy
Reader
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 23.9 KB
Tools
search tools
Get Data
Send Data
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
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
FASTA manipulation
NGS: QC and manipulation
NGS: Mapping
NGS: GATK Tools (beta)
NGS: Simulation
Phenotype Association
Workflows
All workflows
Charts - New Chart
List of created charts:
Delete
New
New Chart
Bar diagram
Last change: 9/3/2014, 14:06
New Chart
Customize
Grouped
Stacked
1:Data label
2:Data label
3:Data label
4:Data label
5:Data label
362.0
300.0
250.0
200.0
150.0
100.0
50.0
0.0
Y-axis
12.0
25.0
38.0
51.0
64.0
77.0
90.0
X-axis
History
Unnamed history
23.9 KB
1: tabular.txt
100 regions
format: interval, database: ?
uploaded interval file
1.Chrom 2.Start 3.End 4 5
43 14 75 95 85
33 100 32 20 17
5 60 46 54 34
35 73 40 58 21
45 3 36 35 5
6 84 89 72 30

NVD3 Charts


127.0.0.1:8080/root/index — Galaxy
Reader
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 23.9 KB
Tools
search tools
Get Data
Send Data
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
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
FASTA manipulation
NGS: QC and manipulation
NGS: Mapping
NGS: GATK Tools (beta)
NGS: Simulation
Phenotype Association
Workflows
All workflows
Charts - New Chart
List of created charts:
Delete
New
New Chart
Stacked area
Last change: 9/3/2014, 14:07
New Chart
Bar diagram
Last change: 9/3/2014, 14:06
New Chart
Customize
Stacked
Stream
Expanded
3984.6
0.0
Y-axis
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0
90.0
99.0
X-axis
History
Unnamed history
23.9 KB
1: tabular.txt
100 regions
format: interval, database: ?
uploaded interval file
1.Chrom 2.Start 3.End 4 5
43 14 75 95 85 5
33 100 32 20 17 5
5 60 46 54 34 8
35 73 40 58 21 2
45 3 36 35 5 1
6 84 89 72 30 7

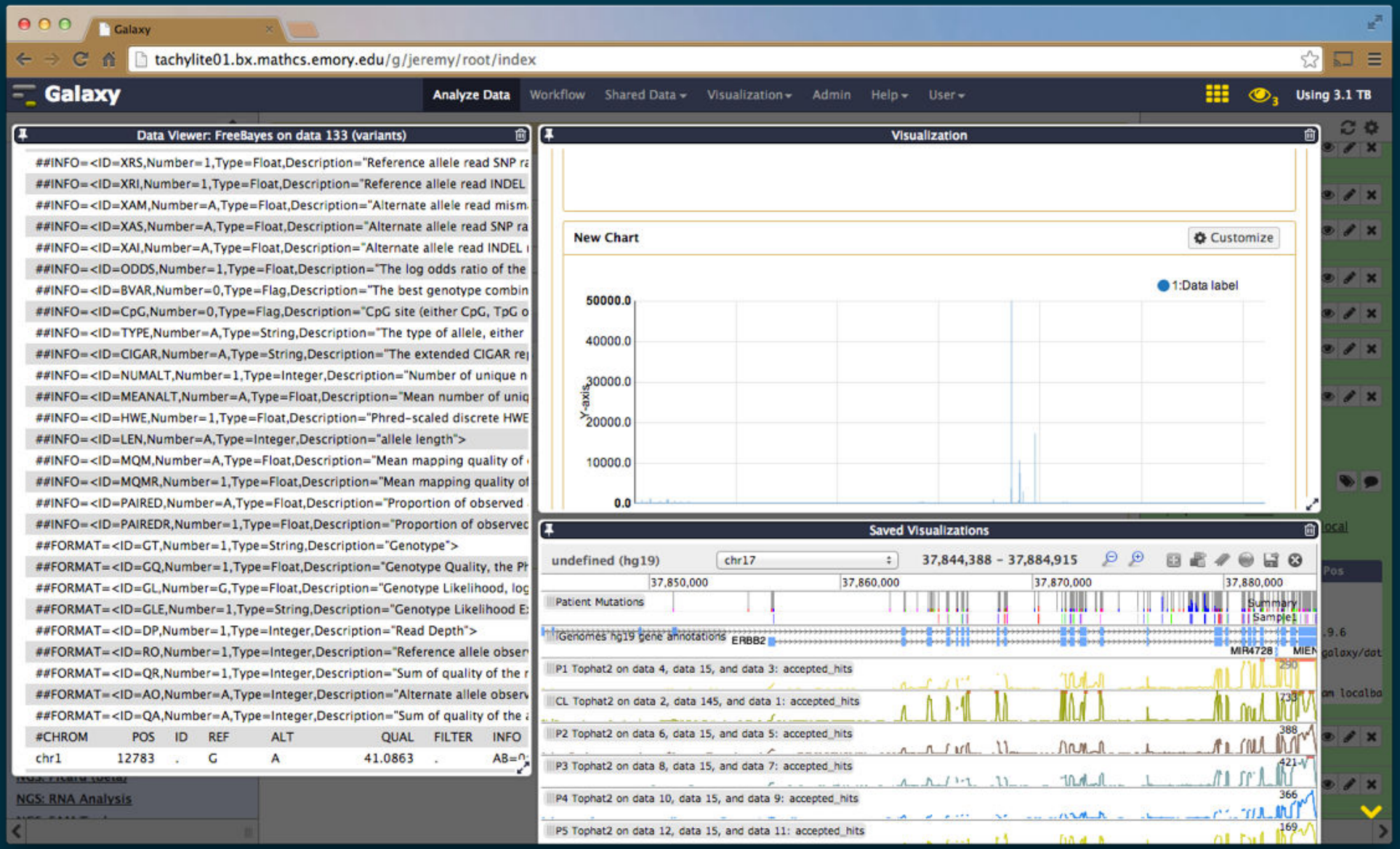
Coupling and Visual Analysis

Coupling visualizations together leads to more compelling applications

Visual analysis is the meaningful coupling of analysis tools and visualization

Galaxy enables both loose and tight coupling

Coupling Visualizations and Analysis



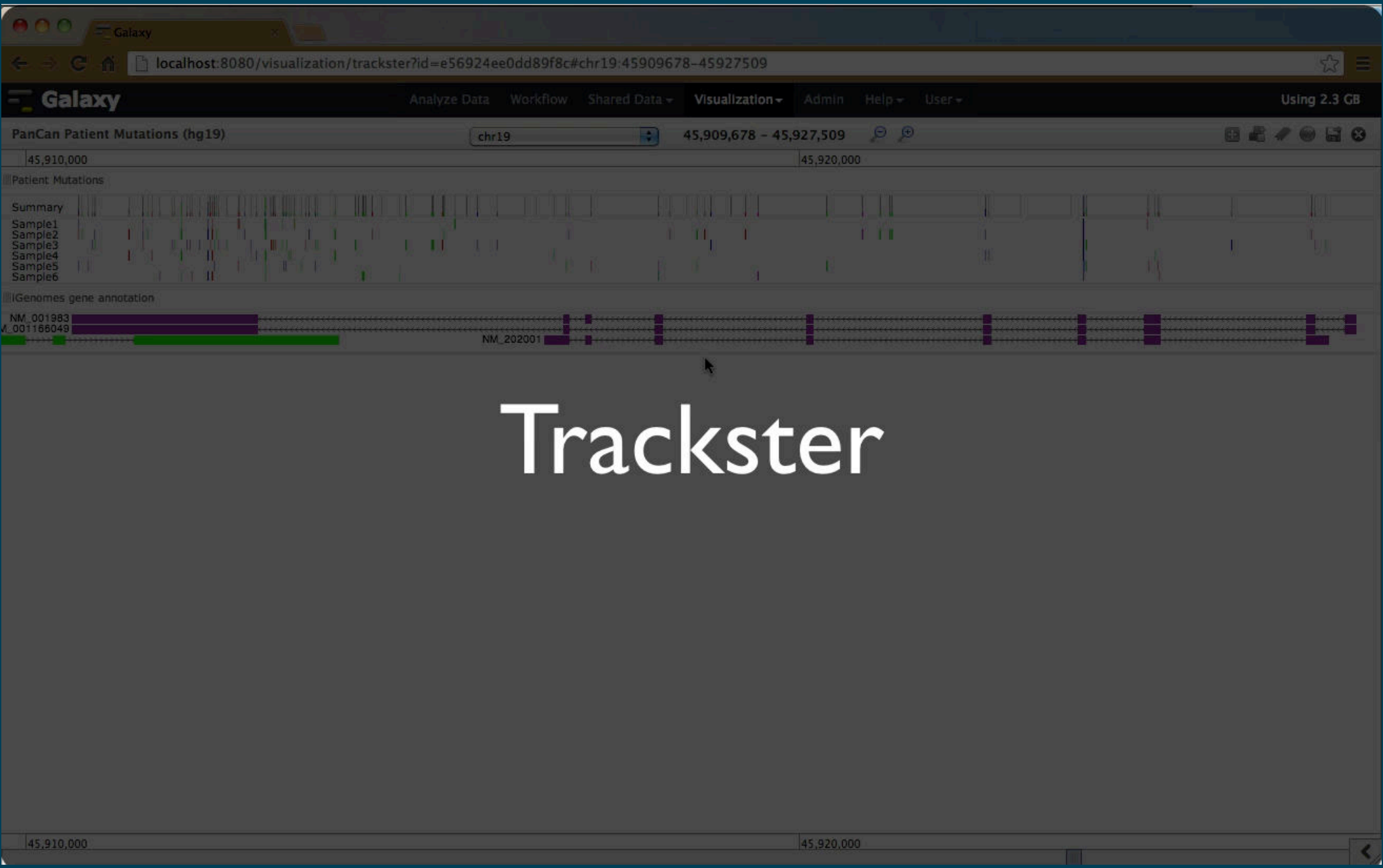
Visual Analysis for Improving Pipeline Development

Challenges for genomic tools

- time and compute intensive
- many parameters
- parameter sensitive

Approach

- repeatedly run tools on selected regions
- visualize outputs side by side
- compare to find best parameters


Galaxy
localhost:8080/visualization/trackster?id=e56924ee0dd89f8c#chr19:45909678-45927509
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 2.3 GB
PanCan Patient Mutations (hg19)
chr19
45,909,678 - 45,927,509
45,910,000
45,920,000
Patient Mutations
Summary
Sample1
Sample2
Sample3
Sample4
Sample5
Sample6
Genomes gene annotation
NM_001983
NM_001166049
NM_202001
45,910,000
45,920,000

Trackster

Summary

Visualizations are 1st class Galaxy objects

- any Web-based visualization can be added to Galaxy
- strong support for data indexing
- creating, saving, sharing, and versioning visualizations all done through a Web browser

Galaxy enables visualizations + data + tools = powerful integrated visual analysis tools

The Galaxy Team


Enis Afgan


Dannon Baker


Dan Blankenberg


Dave Bouvier


Marten Cech


Dave Clements


Nate Coraor


Carl Eberhard

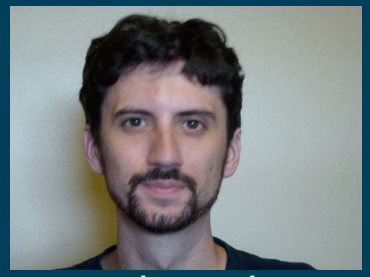

Dorine Francheteau


Sam Guerler


Jen Jackson


Greg von Kuster


Ross Lazarus


Nick Stoler


Jeremy Goecks


Anton Nekrutenko


James Taylor

JOHNS HOPKINS UNIVERSITY

PENNSTATE

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Thanks!
Questions?

Galaxy

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

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

Postdoc and software engineer positions available in
Interactive Genomics Lab @ THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

http://jeremygoecks.com
jgoecks@gwu.edu