An Introduction to Galaxy

Daniel Blankenberg
The Galaxy Team | Penn State
http://GalaxyProject.org

Goals

What is Galaxy

- Mission, Terminology, and Capabilities

Using Galaxy

- Play along with a basic Analysis

How to get involved with Galaxy

- The community and how to participate

Adding Tools

- Making your code Accessible and Reproducible

Bonus: Brief Intro to Sequence Analysis in Galaxy

Goals

What is Galaxy

- Mission, Terminology, and Capabilities

Using Galaxy

- Play along with a basic Analysis

How to get involved with Galaxy

- The community and how to participate

Adding Tools

- Making your code Accessible and Reproducible

Bonus: Brief Intro to Sequence Analysis in Galaxy

Overview

What is Galaxy?

Analysis Interface, Tools, and Datasources

Workflows

Visualizations

Sharing, Publishing, and Galaxy Pages

ToolShed

Galaxy Project Mission

Galaxy is a data integration and analysis platform emphasizing accessibility, reproducibility, and transparency

Accessible: Users without programming experience can easily specify parameters and run tools and workflows.

Reproducible: Galaxy captures information so that any user can repeat and understand a complete computational analysis.

Transparent: Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

What is Galaxy?

A **data analysis and integration** tool

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

There are several **ways to use Galaxy**

Using Galaxy - 4 ways

- **Public Main** Galaxy web instance: ***usegalaxy.org***
- **Local** instance: ***getgalaxy.org***
- **Cloud** instance: ***usegalaxy.org/cloud***
- **Other Public** Galaxy web instances hosted by various groups: wiki.galaxyproject.org/PublicGalaxyServers

Galaxy as a *Genomics WorkBench*

Dataset:

Any input, output or intermediate set of data + metadata. A record of a specific data or analysis step.

History:

A series of inputs, analysis steps, intermediate datasets, and outputs. A record of a group of data and analysis steps.

Tool:

An operation within Galaxy that acts upon dataset(s) as an analysis step. May be developed by Galaxy team or a 3rd party program that has been "wrapped" for Galaxy.

Workflow:

A series of analysis steps executed as a unit.

More Galaxy Terminology

Share:

Make something available to someone else

Publish:

Make something available to everyone

Galaxy Page:

Analysis documentation within Galaxy; easy to embed and link to any Galaxy object (histories, datasets, workflows, visualization) or external resource (video, graphics, publications).

Visualize:

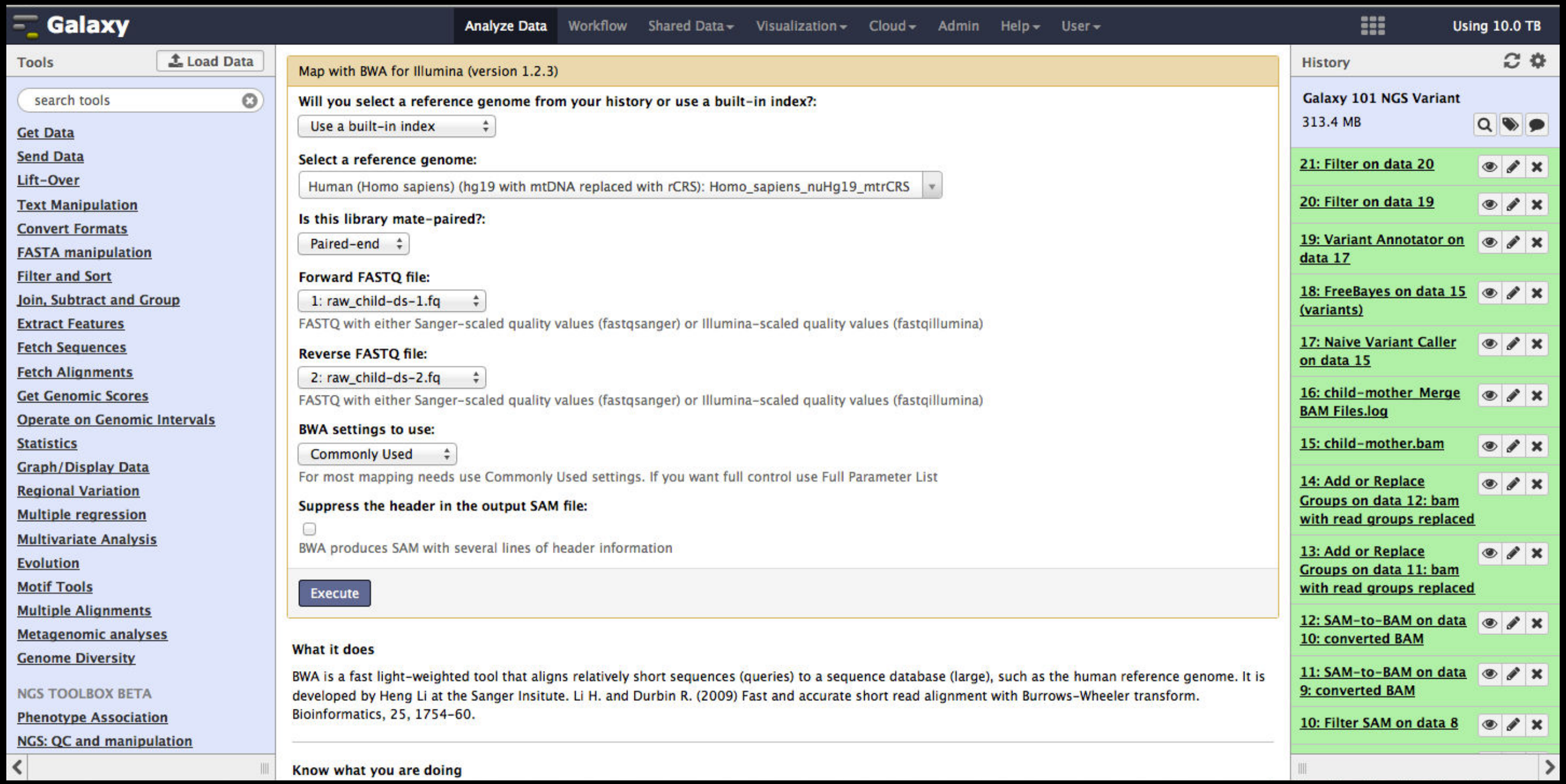
External resources. Trackster. Galaxy Charts & Visualizations.

Overview

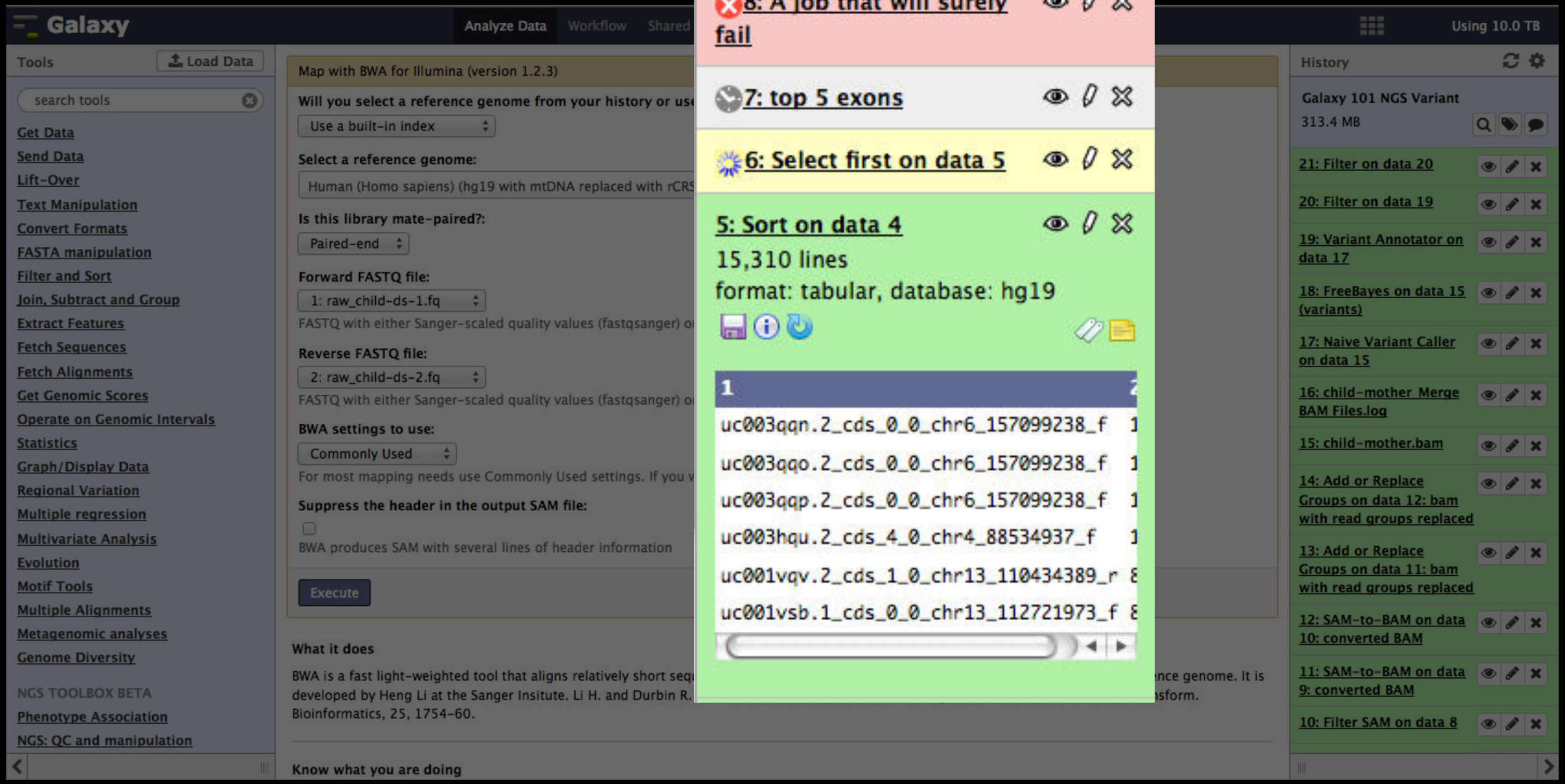
What is Galaxy?

- Analysis Interface, Tools, and Datasources
- Workflows
- Visualizations
- Sharing, Publishing, and Galaxy Pages
- ToolShed

Galaxy Analysis Workspace



Galaxy Analysis Workspace



Tool Suites

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

Regional Variation
EMBOSS
Evolution / Phylogeny
RNA-seq
ChIP-seq
GATK
Picard
Metagenomics
Motif Tools
...and more

Datasources

Upload file from your computer

- Import via URL
- FTP support for large datasets

UCSC table browser

EBI SRA

BioMart

interMine / modMine

GenomeSpace

add more!

Overview

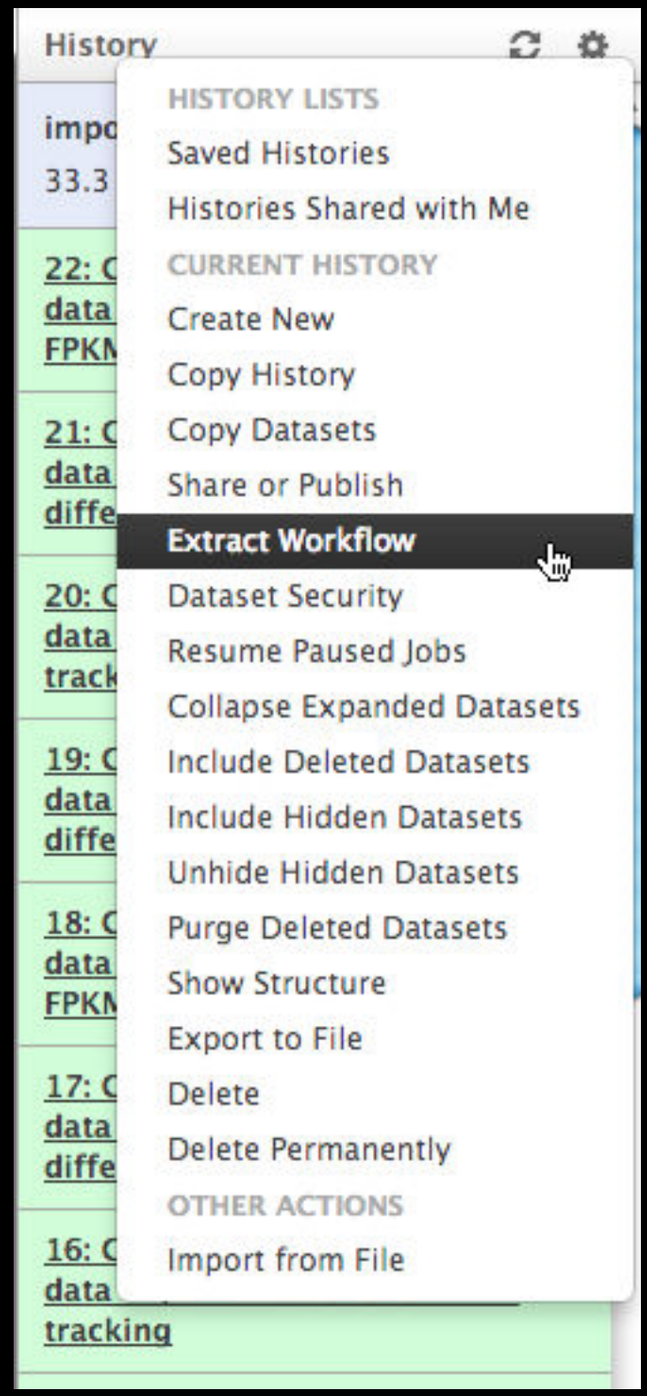
What is Galaxy?

- Analysis Interface, Tools, and Datasources
- Workflows
- Visualizations
- Sharing, Publishing, and Galaxy Pages
- ToolShed

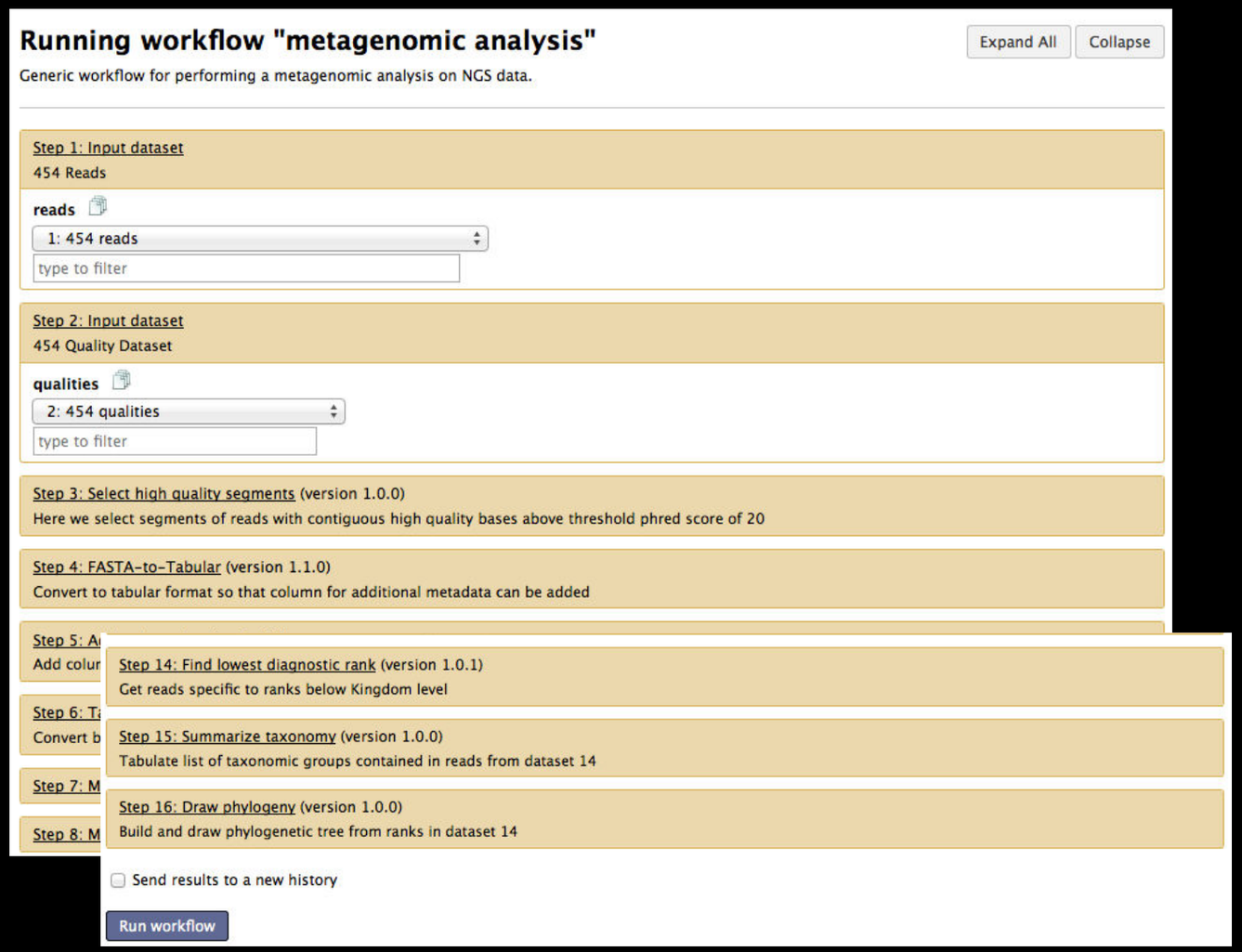
Create Workflow Automatically

Extract Workflow from History

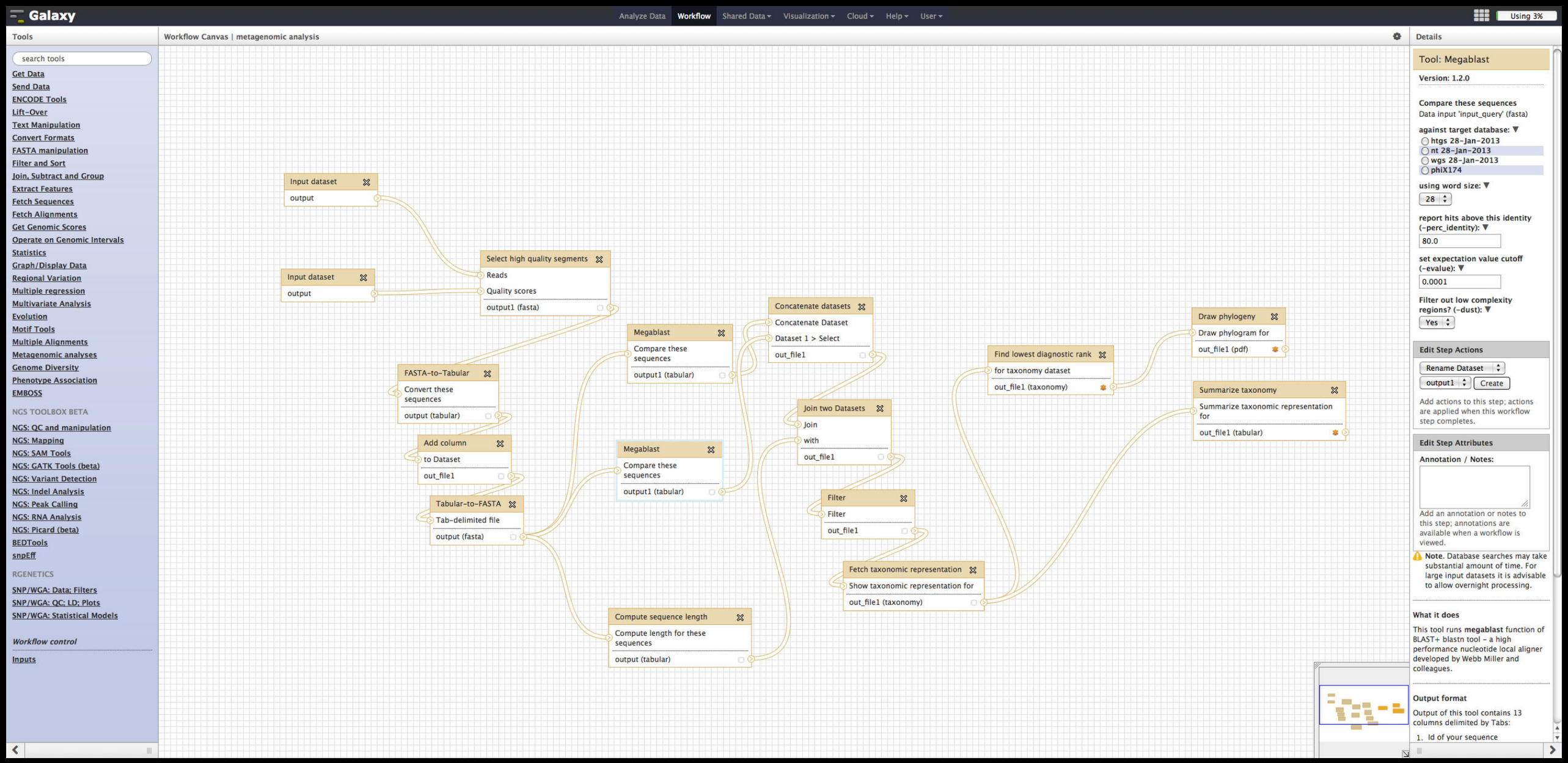
Create a workflow from a History that you created interactively.



Run it



Workflow Editor



Overview

What is Galaxy?

- Analysis Interface, Tools, and Datasources
- Workflows
- Visualizations
- Sharing, Publishing, and Galaxy Pages
- ToolShed

Visualizing Genomics

Supported external browsers

- UCSC
- Ensembl
- GBrowse
- IGB
- IGV

Traditional browser strengths:

- Showing what is nearby
- what else is happening here
- highlighting correlations
- integrating many datasets

UCSC Genome Browser on Mouse July 2007 (NCBI37/

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom

position/search chr12:57,795,963-57,815,592 gene jump clear

chr12 (qC1) 12qA1.1 qA2 12qA3 qB1 12qB3 12qC1 qC2 12qC3 qD1 D2 12qD3 12qE 12qF1 qF2

14: Tag Counts (bigWig)

2.4 Gb, format: bigwig, database: mm9

Info:

| display at UCSC main

Binary UCSC BigWig file

Scale
chr12: 57800000 5 kb 57805000 57810000 57815000

Tag Counts (bigWig)

g Counts (bigWig)

STS Markers on Genetic and Radiation Hybrid Maps

STS Markers

UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics

Pax9 Slc25a21

RefSeq Genes

RefSeq Genes

Non-Mouse RefSeq Genes

Other RefSeq

Ensembl Gene Predictions

Ensembl Genes

Human Proteins Mapped by Chained tBLASTn

PAX9 SLC25A21

Mouse mRNAs from GenBank

AK132404
AK034985
X84000
BC005794
BC006930
AK167535
AK162379
AK045180
AK041694
AK036988
EU180236
BC057980

Mouse ESTs That Have Been Spliced

Spliced ESTs

30-Way Multiz Alignment & Conservation

Mammal Cons

Rat
Human
Orangutan
Dog
Horse
Opossum
Chicken
Stickleback

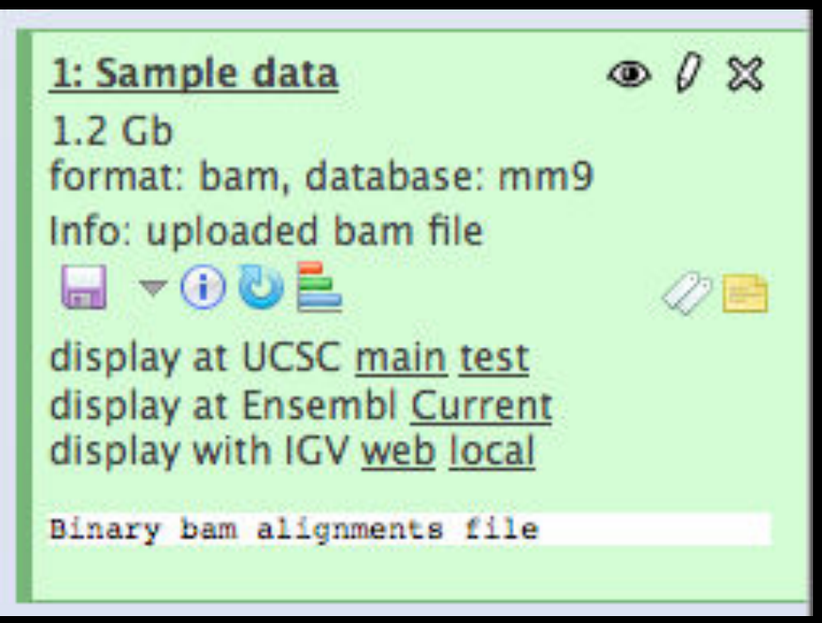
Simple Nucleotide Polymorphisms (dbSNP build 128)

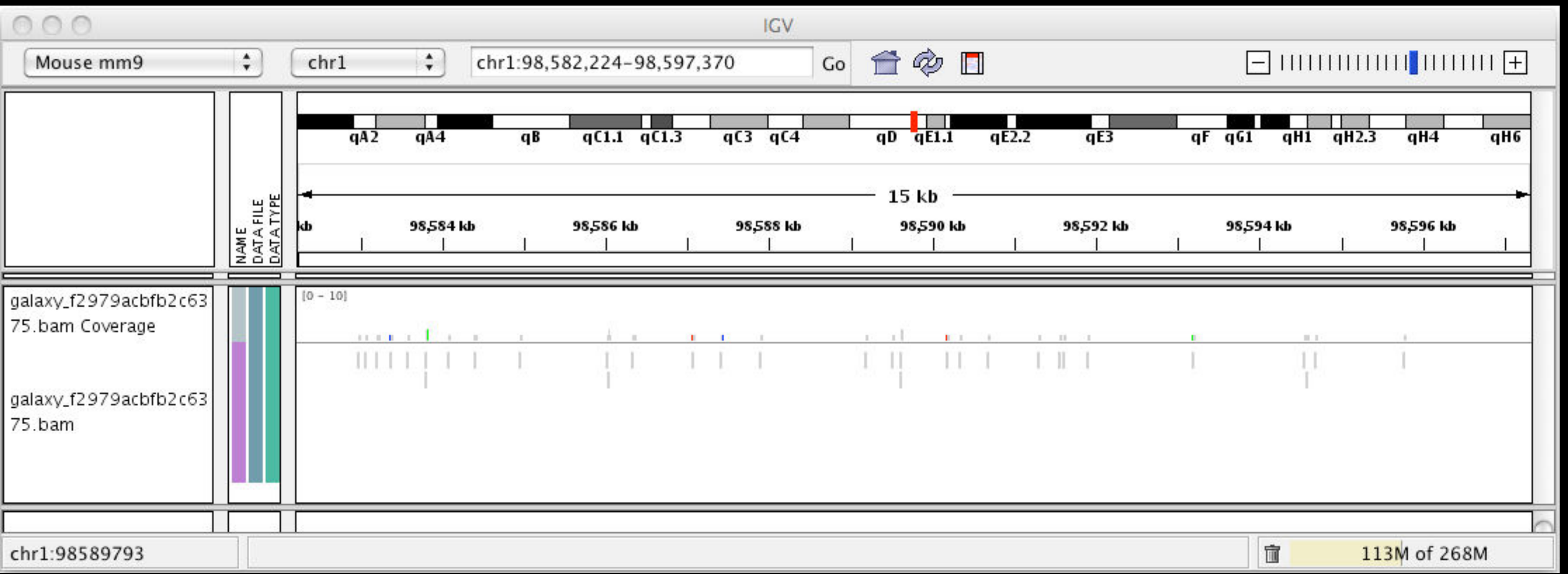
SNPs (128)

Repeating Elements by RepeatMasker

RepeatMasker

Integrative Genomics Viewer (IGV)



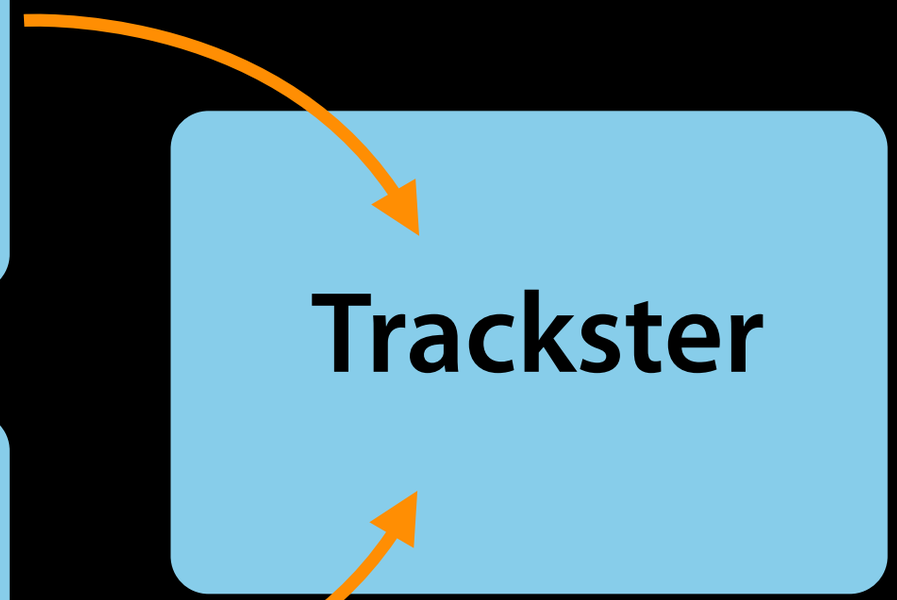


Galaxy

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

Genome Browser

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



Trackster

View your data from within Galaxy

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

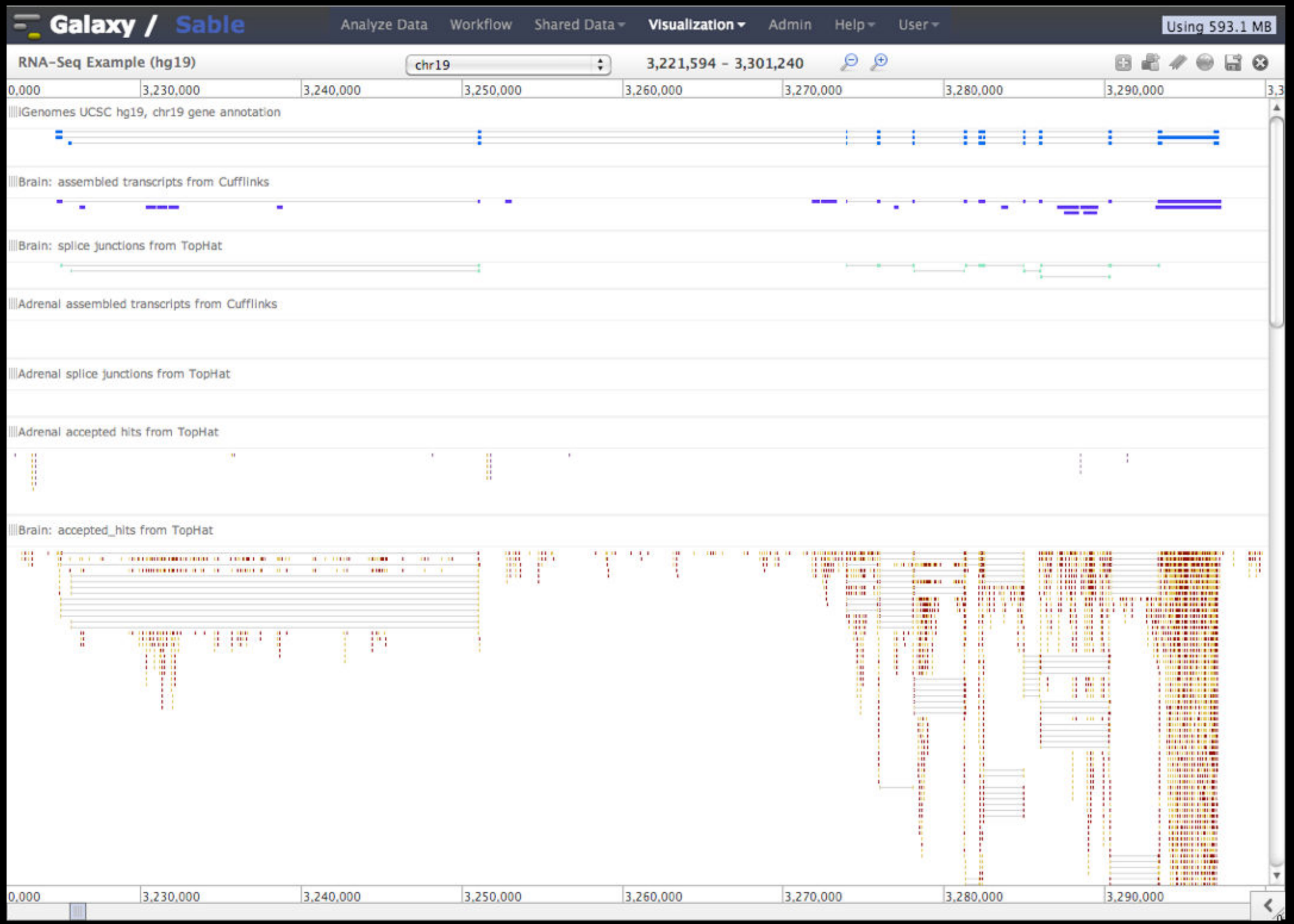
Supports common filetypes

- BAM, BED, GFF/GTF, WIG

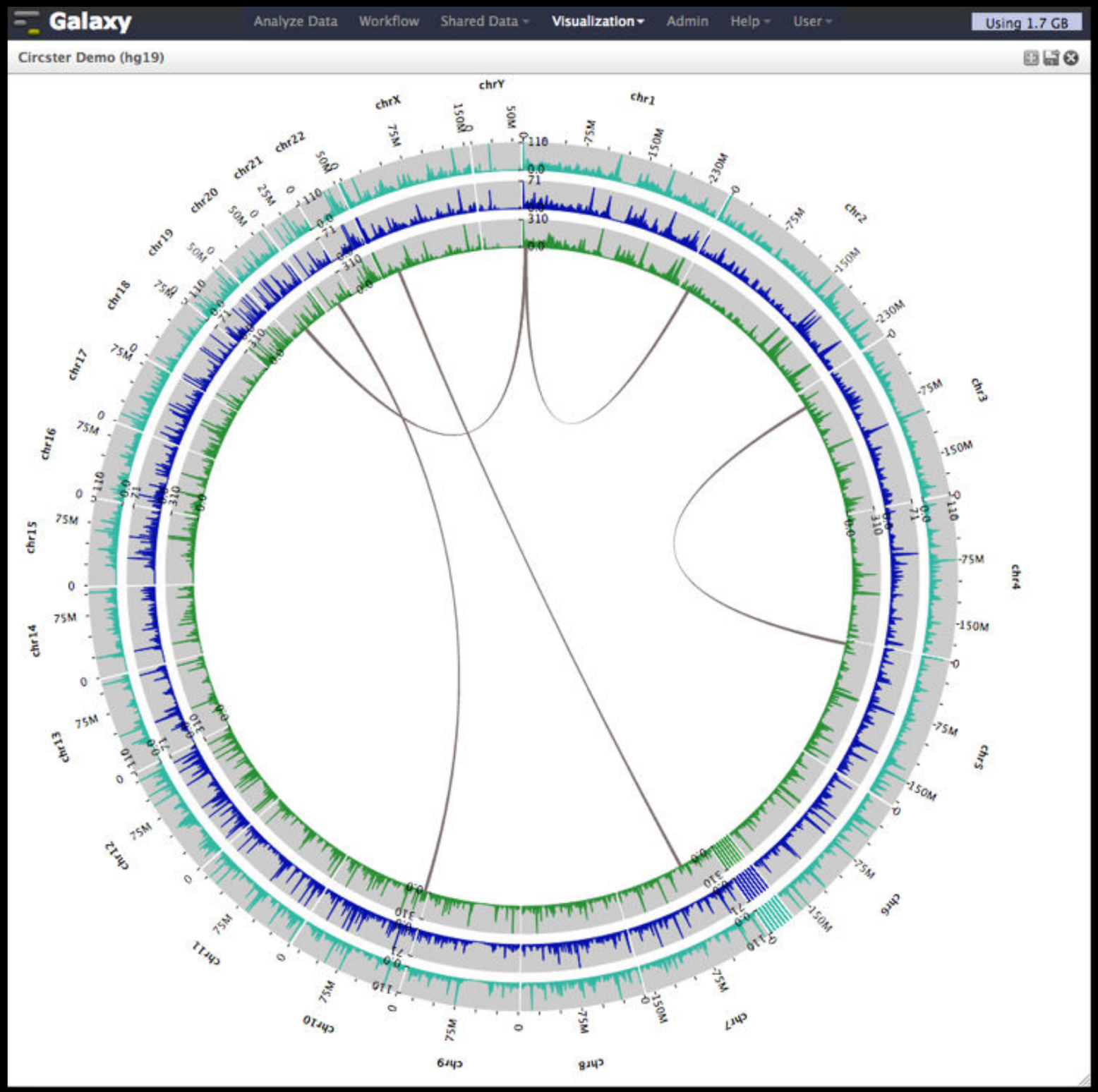
Unique features

- custom genomes
- highly interactive

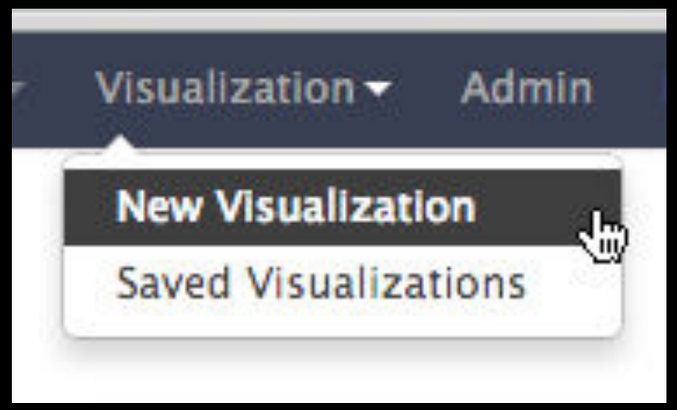
Trackster: Galaxy's embedded track browser



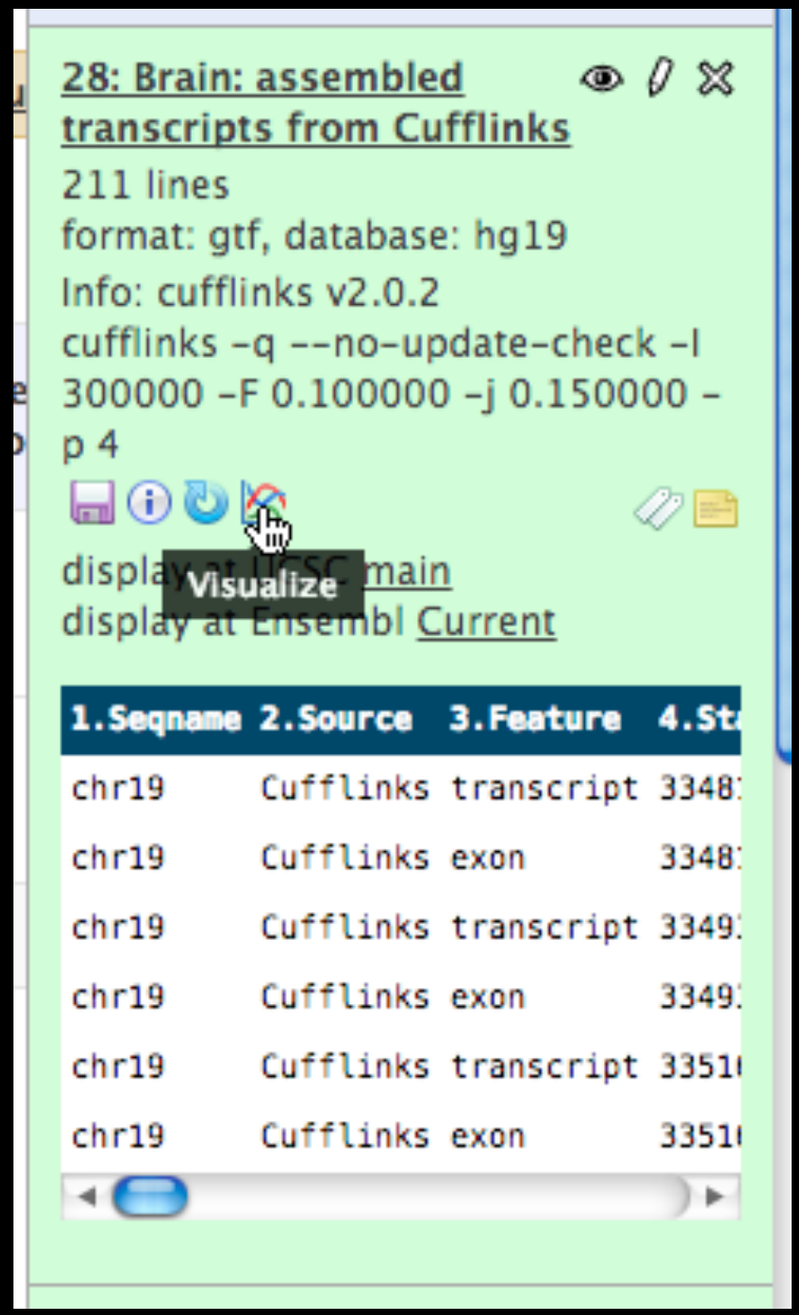
Circster


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 1.7 GB
Circster Demo (hg19)
chrY
chrX
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22

Create a visualization in Galaxy



or



Visualization inside Galaxy

- Leverages visualization to evaluate and refine analyses
- Exposes basic analyses in visualization to make it more informative
- Makes that analyze-visualize-refine loop seamless and fast
- Enables learning tools and exploring their parameter space
- Supports custom genome browsers, without a predefined reference genome

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
26.1
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
2.1
CUFF.4.1
CUFF.6.1
CUFF.3.2
CUFF.5.1
CUFF.7.1
CUFF.3.3
CUFF.8.1

Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.001, No]

CUFF.3.1
2.1
CUFF.4.1
CUFF.6.1
CUFF.3.2
CUFF.5.1
CUFF.7.1
CUFF.3.3
CUFF.8.1

0 | 1,000,000

GM12878 Cufflinks assembled transcripts BEST

h1-hESC Cufflinks assembled transcripts BEST

Score [387-1000]
exon_number [1-1]
frac [0-1]
cov [1-1248]
conf_lo [2227-178180]
FPKM [215-223870]
conf_hi [304-269560]

Run on complete dataset

Overview

What is Galaxy?

Analysis Interface, Tools, and Datasources

Workflows

Visualizations

Sharing, Publishing, and Galaxy Pages

ToolShed

Sharing & Publishing enables Reproducibility

Reproducibility: Everybody talks about it, but ...

Galaxy aims to push the goal of reproducibility from the bench to the bioinformatics realm

All analysis in Galaxy is recorded without any extra effort from the user.

Histories, workflows, visualizations and ***pages*** can be shared with others or published to the world.

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

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

Make History Accessible via Link

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

Make History Accessible and Publish

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

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

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

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

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

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

You can:

Unpublish History

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

Disable Access to History via Link and Unpublish

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

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List

Galaxy | Published History | Variant Analysis for Sample E18

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

Google

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User

Published Histories | jgoecks | Variant Analysis for Sample E18

Galaxy History ' Variant Analysis for Sample E18'

Import history

Annotation: Perform a pileup analysis with default parameters to identify variants in sample E18.

Dataset	Annotation
1: E18 PE.1 Reads	Forward reads from sample E18.
2: E18 PE.2 Reads	Reverse reads from sample E18.
3: E18 PE.1 Reads Groomed	Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3
4: E18 PE.2 Reads Groomed	Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3
5: E18 PE.1 Reads Groomed, Trimmed	Trim reads from 3' end to remove low-quality nts.
6: E18 PE.2 Reads Groomed, Trimmed	Trim reads from 3' to remove low-quality nts.
7: Map with Bowtie for Illumina on data 6 and data 5	Map paired-end reads with default parameters.
8: SAM-to-BAM on data 7	Need to convert Bowtie SAM to BAM so that pileup analysis can be performed.
9: Generate pileup on data 8	Pileup analysis with default parameters
10: Filter pileup to get Variants from sample E18	Find variants with coverage >= 30.
13: Filter to get Variants from sample E18 where consensus base different than ref. base	Filter pileup to find variants where the consensus base is different than the reference base.
14: UCSC mm9 RefSeq Genes	UCSC mm9 RefSeq genes.
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes	Variants with consensus different that occur in RefSeq genes.

About this History

Author

jgoecks

Related Histories

All published histories
Published histories by jgoecks

Rating

Community
(1 rating, 4.0 average)

Yours

Tags

Community:
snp | pileup | bowtie | demo | sample

Yours:
snp × | pileup × | bowtie × | demo × | sample:e18 ×

Galaxy | Published Workflow | SNP variant detection from paired-end reads

http://main.g2.bx.psu.edu/u/jgoecks/w/snp-variant-detection-from-paired-end-reads

Google

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Published Workflows | jgoecks | SNP variant detection from paired-end reads

Step 6: FASTQ Trimmer

Trim reads to remove low-quality bases.

FASTQ File
Output dataset 'output_file' from step 4

Define Base Offsets as
Absolute Values

Offset from 5' end
0

Offset from 3' end
9

Keep reads with zero length
False

Step 7: Map with Bowtie for Illumina

Map reads using default parameter values.

Will you select a reference genome from your history or use a built-in index?
Use a built-in index

Select a reference genome
/galaxy/data/apiMel3/bowtie_index/apiMel3

Is this library mate-paired?
Paired-end

Forward FASTQ file
Output dataset 'output_file' from step 6

Reverse FASTQ file
Output dataset 'output_file' from step 5

Maximum insert size for valid paired-end alignments (-X)
1000

The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff)
FR (for Illumina)

Bowtie settings to use
Commonly used

Suppress the header in the output SAM file
True

Step 8: SAM-to-BAM

Convert Bowtie SAM output to BAM format so that pileup can be run.

Choose the source for the reference list
Locally cached

About this Workflow

Author

jgoecks

Related Workflows

All published workflows
Published workflows by jgoecks

Rating

Community
(0 ratings, 0.0 average)

Yours

Tags

Community:
snp | bowtie

Yours:
snp | bowtie

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

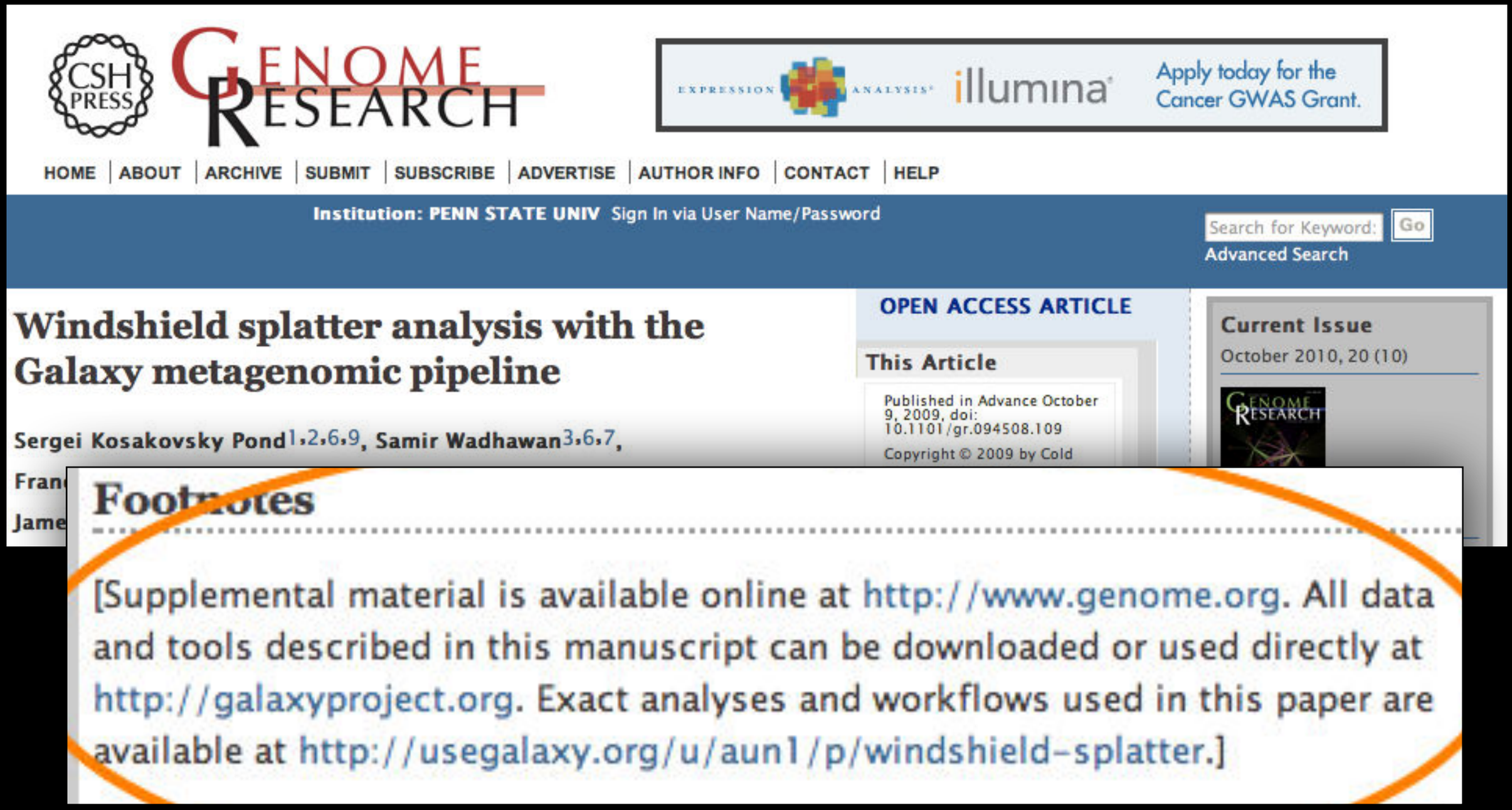
Galaxy's publishing features facilitate access and reproducibility without any extra leg work

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

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

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

Sharing and Publishing Your Work



Histories, workflows, visualizations and ***pages*** can be shared with others or published to the world.

http://usegalaxy.org/u/aun1/p/windshield-splatter

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) – a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Galaxy History | Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):

Galaxy History | metagenomic analysis

This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

Galaxy Workflow | metagenomic analysis
Generic workflow for performing a metagenomic analysis on NGS data.

Accessing the Data

Windshield Splatter datasets analyzed in this manuscript can be accessed through this Galaxy Library. From

Author

aun1

Related Pages

All published pages
Published pages by aun1

Rating

Community
(6 ratings, 5.0 average)

Tags

Community:
paper galaxy megan

http://usegalaxy.org/u/aun1/p/windshield-splatter

Overview

What is Galaxy?

Analysis Interface, Tools, and Datasources

Workflows

Visualizations

Sharing, Publishing, and Galaxy Pages

ToolShed

Enables sharing of Galaxy Utilities:

- tools
- proprietary datatypes
- exported Galaxy workflows

Automatically install tools and tool suites, and their dependencies, into a Galaxy instance

Galaxy Utilities can be created and shared by any member of the community

https://wiki.galaxyproject.org/ToolShed

Galaxies on private clouds

Galaxies on public clouds

http://usegalaxy.org

Galaxy Tool Shed

1 2 3 ... ∞

private Galaxy installations

2464 valid tools on May 27, 2014

Search

- Search for valid tools
- Search for workflows

Valid Galaxy Utilities

- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

All Repositories

- Browse by category

Available Actions

- Login to create a repository

Categories

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	31
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	7
Computational chemistry	Tools for use in computational chemistry	18
Convert Formats	Tools for converting data formats	31
Data Managers	Utilities for Managing Galaxy's built-in data cache	4
Data Source	Tools for retrieving data from external data sources	16
Fasta Manipulation	Tools for manipulating fasta data	51
Fastq Manipulation	Tools for manipulating fastq data	21
Genome-Wide Association Study	Utilities to support Genome-wide association studies	1
Genomic Interval Operations	Tools for operating on genomic intervals	37
Graphics	Tools producing images	22
Imaging	Utilities to support imaging	
Metabolomics	Tools for use in the study of Metabolomics	3
Metagenomics	Tools enabling the study of metagenomes	25
Micro-array Analysis	Tools for performing micro-array analysis	7
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	58
Ontology Manipulation	Tools for manipulating ontologies	7
Phylogenetics	Tools for performing phylogenetic analysis	6
Proteomics	Tools enabling the study of proteins	50

toolshed.g2.bx.psu.edu

Installing from the ToolShed

Requires Admin access on Galaxy Server

Point and Click to Install

Minimal additional configuration needed

- Dependency directory

Accessible Galaxy tool sheds

Galaxy main tool shed

Galaxy test tool shed

- Browse valid repositories
- Search for valid tools
- Search for workflows

Search repositories for valid tools

Valid tools are those that properly load in Galaxy. Enter any combination of the following tool attributes to find repositories that contain valid tools matching the search criteria.

Comma-separated strings may be entered in each field to expand search criteria. Each field must contain the same number of comma-separated strings if these types of search strings are entered in more than one field.

Tool id:

Tool name:

bwa

Tool version:

Exact matches only:

Check the box to match text exactly (text case doesn't matter as all strings are forced to lower case).

Search repositories

Repositories with matching tools

tool id:
tool name: **bwa**
tool version:
exact matches only: **False**

	Repository name	Synopsis	Revision	Owner
☐	bwa_base ▾ Install to Galaxy	bwa_base	35f793d775ea	greg
☐		bwa_color	17da959633b0	greg
☐	bwa_mappers ▾	Galaxy BWA short read aligner	3365c42b076d	devteam
☐	bwa_mappers ▾	Galaxy BWA short read aligner	95d5133dd241	devteam
☐	complex_repository_dependency_on_bwa_059 ▾	Demonstrates a tool_dependencies.xml file that defines a complex repository dependency on bwa_059	96a15cf82c50	greg
☐	data_manager_bwa_index_builder ▾	Builds bwa indexes	fe6508204acc	blankenberg

For 0 selected items: Install to Galaxy

The core Galaxy development team does not maintain the contents of many Galaxy tool shed repositories. Some repository tools may include code that produces malicious behavior, so be aware of what you are installing.

If you discover a repository that causes problems after installation, contact Galaxy support, sending all necessary information, and appropriate action will be taken.

Contact the repository owner for general questions or concerns.

Confirm dependency installation

These dependencies can be automatically handled with the installed repository, providing significant benefits, and Galaxy includes various features to manage them.

Handle repository dependencies?

☑

Un-check to skip automatic installation of these additional repositories required by this repository.

▶ Repository dependencies – *requires bwa color*

Handle tool dependencies?

☑

Un-check to skip automatic handling of these tool dependencies.

▶ Tool dependencies – *repository tools require handling of these dependencies*

Choose the tool panel section to contain the installed tools (optional)

Shed tool configuration file:

/home/g2test/var/shed_tool_conf.xml

Your Galaxy instance is configured with 1 shed-related tool configuration file, so repositories will be installed using it's **tool_path** setting.

Add new tool panel section:

Add a new tool panel section to contain the installed tools (optional).

Select existing tool panel section:

- Get Data
- BEDtools
- SNP Eff
- Send Data
- ENCODE Tools

Monitor installing tool shed repositories

Name	Description	Owner	Revision	Status
bwa_base	bwa_base	greg	35f793d775ea	New
bwa_color	bwa_color	greg	17da959633b0	Setting tool versions

Installed tool shed repositories

Update tool shed status

search repository name

Advanced Search

Name↓	Description	Owner	Revision	Installation Status	Tool shed
allele_counts_1	Parse the output of the Naive Variant Detector tool, count alleles	nick	db6f217dc45a	Installed	testtoolshed
allele_counts_1	Parse the output of the Naive Variant Detector tool, count alleles	nick	eca1ea054d0d	Installed	testtoolshed
bwa_base	bwa_base	greg	35f793d775ea	Installed	testtoolshed
bwa_color	bwa_color	greg	17da959633b0	Installed	testtoolshed

Goals

What is Galaxy

- Mission, Terminology, and Capabilities

Using Galaxy

- Play along with a basic Analysis

How to get involved with Galaxy

- The community and how to participate

Adding Tools

- Making your code Accessible and Reproducible

Bonus: Brief Intro to Sequence Analysis in Galaxy

Basic Analysis

Which 5 coding exons have most overlapping SNPs?

Use Human, hg19, Chromosome 22

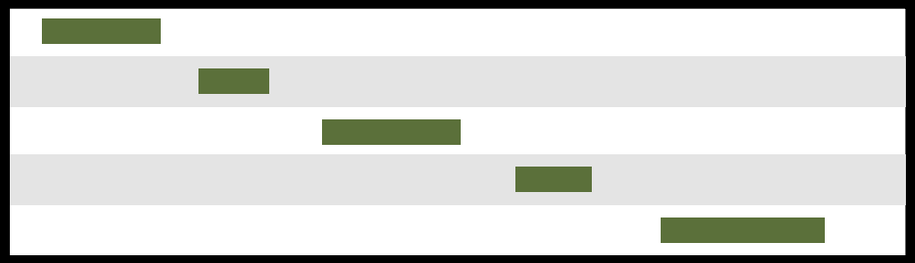
gcc2015-X.dblankenberg.org
(where *X* is 1 through 3)

(~ http://usegalaxy.org/galaxy101)

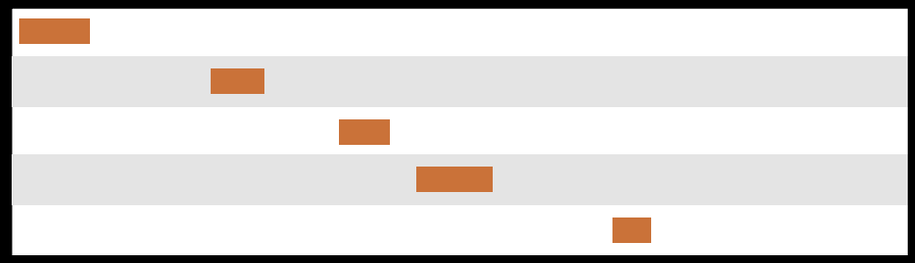
Exons & Repeats: A General Plan

- Get some data
  - Get Data → UCSC Table Browser
- Identify which exons have SNPs
- Count SNPs per exon
- Visualize, save, download, ... exons with most SNPs

(~ http://usegalaxy.org/galaxy101)

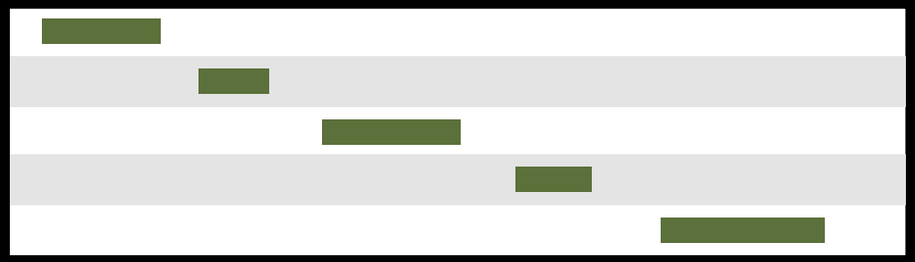


Exons

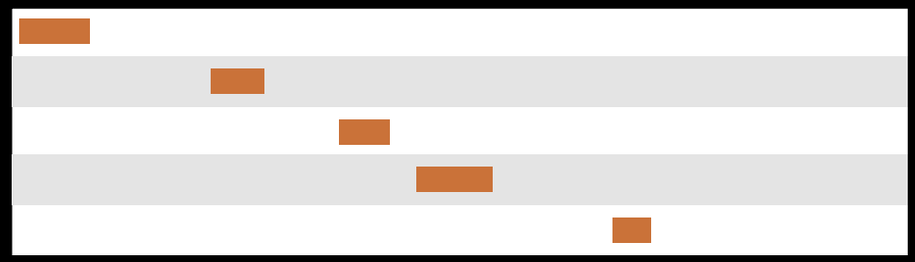


SNPs

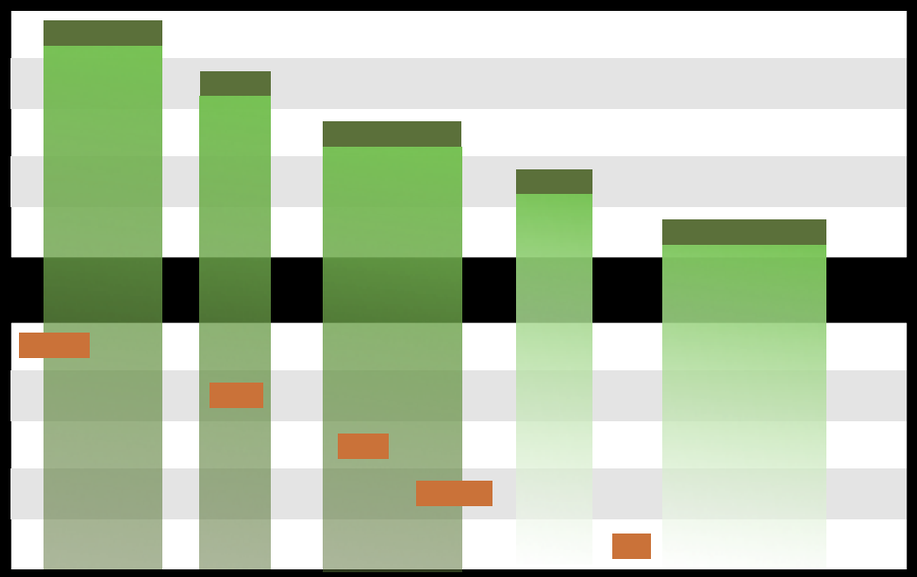
(Identify which exons have SNPs)



Exons



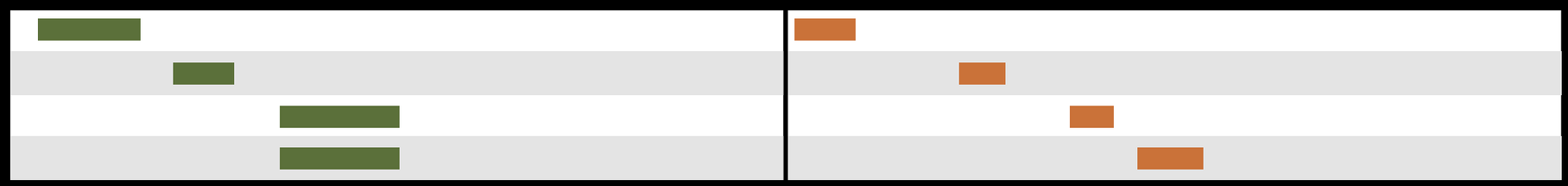
SNPs



Exons

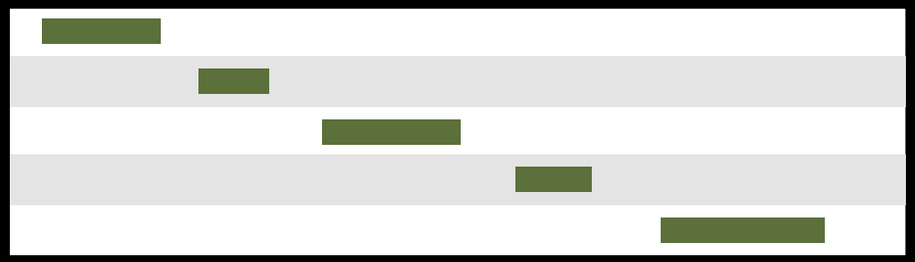
SNPs

Overlap pairings

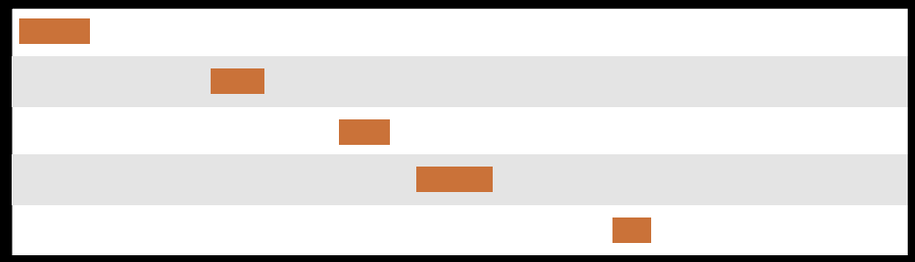


Operate on Genomic Intervals → **Join**

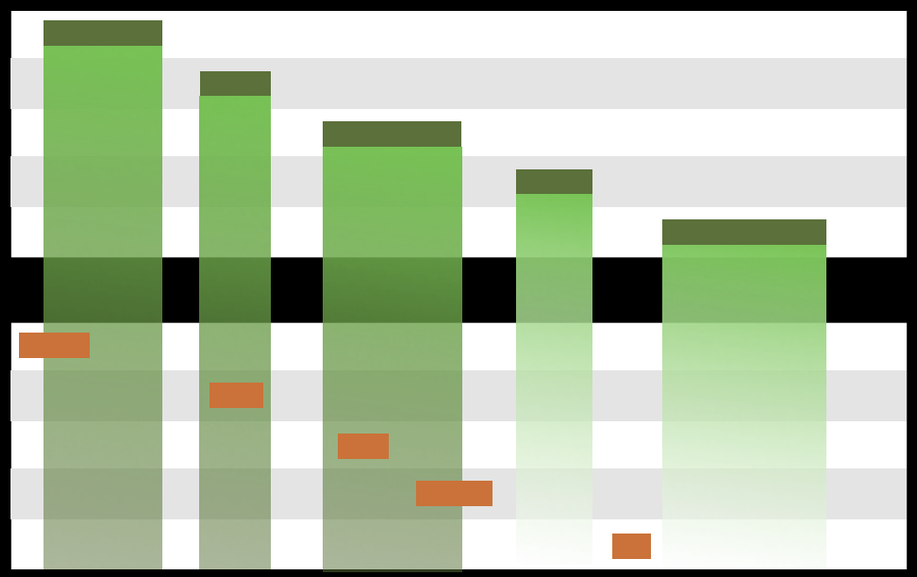
(Identify which exons have SNPs)



Exons



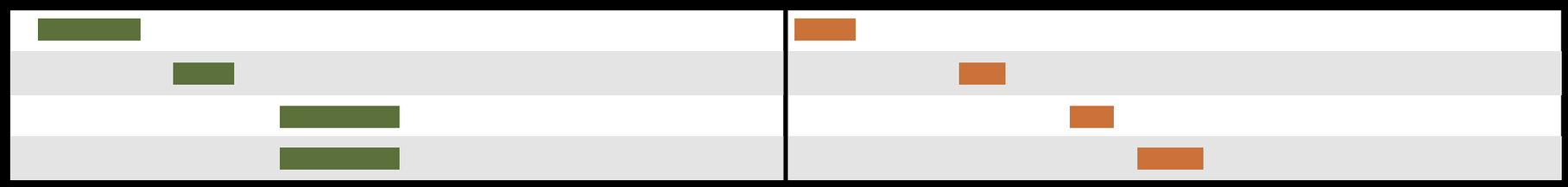
SNPs



Exons

SNPs

Overlap pairings



Exon overlap counts

Join, Subtract, and Group → Group

(Count SNPs per exon)

Find Top 5

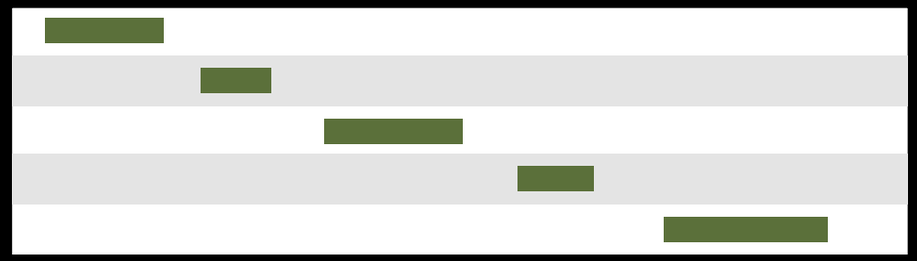
Sort by counts

Filter and Sort -> Sort

Select top 5

Text Manipulation -> Select First

Exon overlap counts

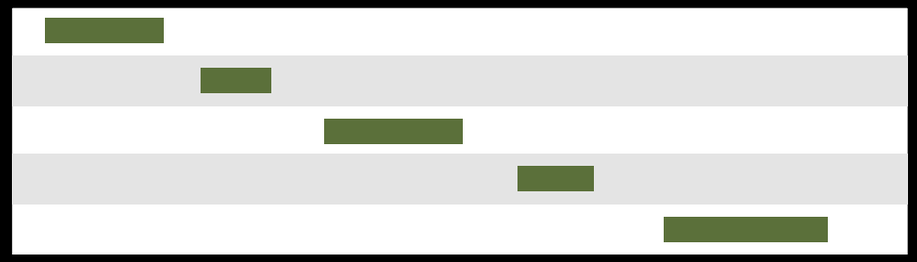


Exons

We’ve answered our question, but we can do better.

Incorporate the overlap count with rest of Exon information

Exon overlap counts



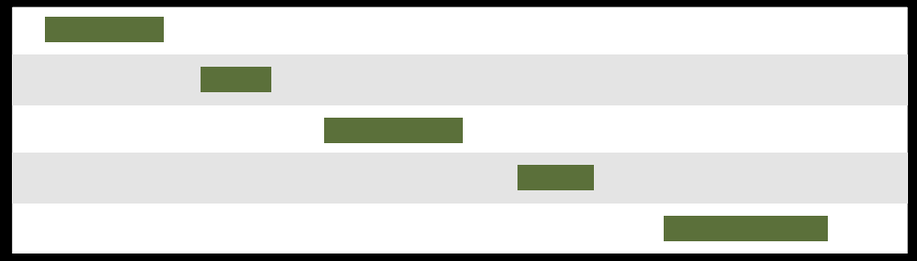
Exons

Join on exon name

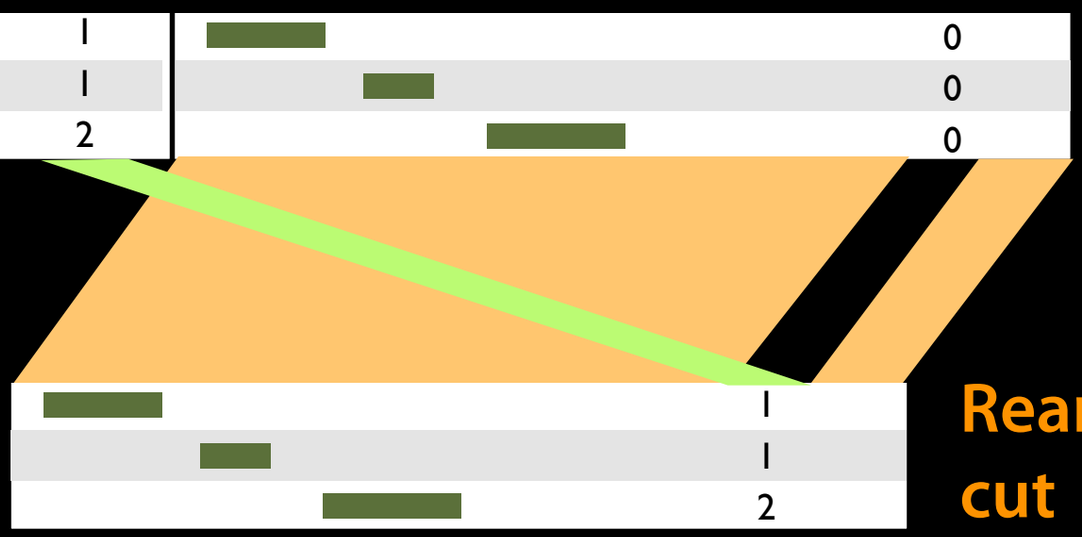
Join, Subtract, and Group → Join

(Incorporate the overlap count with rest of Exon information)

Exon overlap counts



Exons



Join on exon name

Rearrange columns w/ cut

Text Manipulation → Cut

(Incorporate the overlap count with rest of Exon information)

Visualize within a Genome Browser

Click display at UCSC link

Zoom, scroll, examine

Do it again on Repeats

Extract Workflow

Edit Workflow to include input labels, rename outputs

Get Genome-wide repeat and exon information

Run Workflow

Drink Coffee

Goals

What is Galaxy

- Mission, Terminology, and Capabilities

Using Galaxy

- Play along with a basic Analysis

How to get involved with Galaxy

- The community and how to participate

Adding Tools

- Making your code Accessible and Reproducible

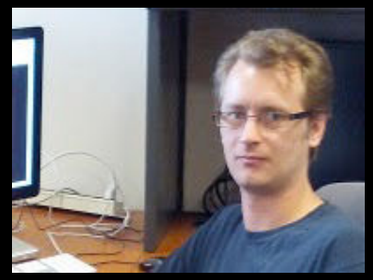
Bonus: Brief Intro to Sequence Analysis in Galaxy

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg



Dave Bouvier

Marten Čech

John Chilton

Dave Clements

Nate Coraor

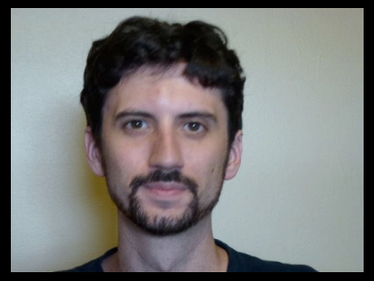
Carl Eberhard

Jeremy Goecks

Sam Guerler

Jen Jackson

Ross Lazarus



Nick Stoler

Anton Nekrutenko

James Taylor

Nitesh Turaga

http://wiki.galaxyproject.org/GalaxyTeam

The Extended Galaxy Team & Contributors

Peter Cock
(TJHI)

Greg Von Kuster
(PSU)

Eric Rasche
(CPT)

Kyle Ellrott
(UCSC)

50+ More Not Pictured

Yousef Kowsar
(VLSCI)

Future You?

Nicola Soranzo
(TGAC)

Nuwan Goonasekera
(VeRSI)

Björn Grüning
(Uni Freiburg)

Questions about Using Galaxy

LATEST 20 | OPEN 12 | NEWS 1 | TUTORIAL | FORUM | PLANET | ALL »

Biostars
GALAXY EXPLAINED

Daniel Blankenberg ♦♦ 1.5k | Logout — about • faq • rss

Community | Messages 23 | Votes | My Posts | My Tags | Following | Bookmarks | New Post

Live search: start typing... or Classic search

Limit to: all time — <prev • 4,271 results • page 1 of 123 • next > — Sort by: update

votes	answer/answers	views		
0	1 answer	9	Disk usage over quota after permanent delete	disk usage — written 23 hours ago by j.norris • 0 • updated 3 hours ago by Martin Čech ♦♦ 1.9k
0	1 answer	13	Help with running gi2taxonomy.py	gi2taxonomy, run python script — written 1 day ago by ibis • 0 • updated 16 hours ago by Bjoern Gruening ♦ 2.7k
0	1 answer	12	How to query the Galaxy database for time/resource data for users and/or workflows?	workflows, database, galaxy, admin — written 2 days ago by mrals89 • 50 • updated 16 hours ago by Bjoern Gruening ♦ 2.7k
0	0 answers	7	Metadata missing error when using lastz	software error — written 1 day ago by bruce.bejcek • 0
0	0 answers	14	bwa-mem installation and use	galaxy, software error, bwa — written 1 day ago by noè • 0
0	1 answer	20	cleanup datasets script python	dataset, clean, galaxy — written 2 days ago by k.mamoud • 20 • updated 1 day ago by Dannon Baker ♦ 2.5k
0	0 answers	15	Problems with login into Galaxy (local instance)	galaxy, login, local instance — written 1 day ago by FT • 0

the last hour

Recent Votes

- C: Disk quota at 100% when I believe it should be less than 50%
- Attaching annotations to Sequences
- A: Disk Usage do not decrease after deleting/purging files. Real use 37%, shows 85
- Adding read group information in a bam header
- Creating a Workflow for Variant Detection for 200 samples
- A: Is it possible for workflow parameters to take information from the names of inp
- Is it possible for workflow parameters to take information from the names of input datasets? e.g. ${Read_group } is set to be the first three characters of input dataset name.

Recent Locations • All »

- United States, just now
- United States, 1 hour ago
- European Union, 1 hour ago
- United States, 3 hours ago

Recent Awards • All »

https://biostar.usegalaxy.org/

How to Help

Report your bugs and request Features

galaxyproject.org/trello

galaxy-bugs@lists.galaxyproject.org

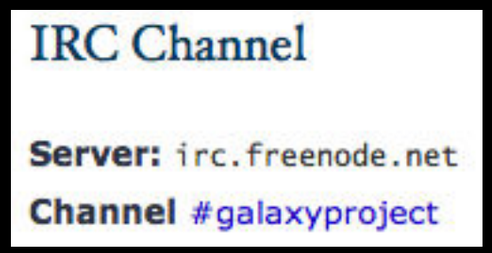
Contribute (check trello)

https://github.com/galaxyproject

write and publish tools

Edit wiki.galaxyproject.org

Chat



Twitter:
@galaxyproject
#usegalaxy

Contributing Code

All code is add via Pull Requests

Info for contributors

https://github.com/galaxyproject/galaxy/blob/dev/

CONTRIBUTING.md

Contributing

This document briefly describes how to contribute to the core galaxy project - also checkout our 2013 Galaxy Community Conference presentation on the topic (video, presentation). For information on contributing more broadly to the Galaxy ecosystem and a deeper discussion of some of these points - please see the Develop section of the Galaxy Wiki.

Before you Begin

If you have an idea for a feature to add or an approach for a bugfix - it is best to communicate with Galaxy developers early. The most common venue for this is the Galaxy Trello board. Browse through existing cards here and if one seems related comment on it. If no existing cards seem appropriate, a new issue can be opened using this form. Galaxy developers are also generally available via IRC and on the development mailing list.

The Intergalactic Utilities Commission (IUC)

Community Group tasked with

Developing Tool Best Practices

Policing the ToolShed

http://galaxy-iuc-standards.readthedocs.org/en/latest/best_practices.html

2013

Goals

What is Galaxy

- Mission, Terminology, and Capabilities

Using Galaxy

- Play along with a basic Analysis

How to get involved with Galaxy

- The community and how to participate

Adding Tools

- Making your code Accessible and Reproducible

Bonus: Brief Intro to Sequence Analysis in Galaxy

Creating Your Own Tools

The Problem

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

The Galaxy Solution

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

Steps:

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

Adding your Own

Write or download a command-line executable

Determine number and kind of

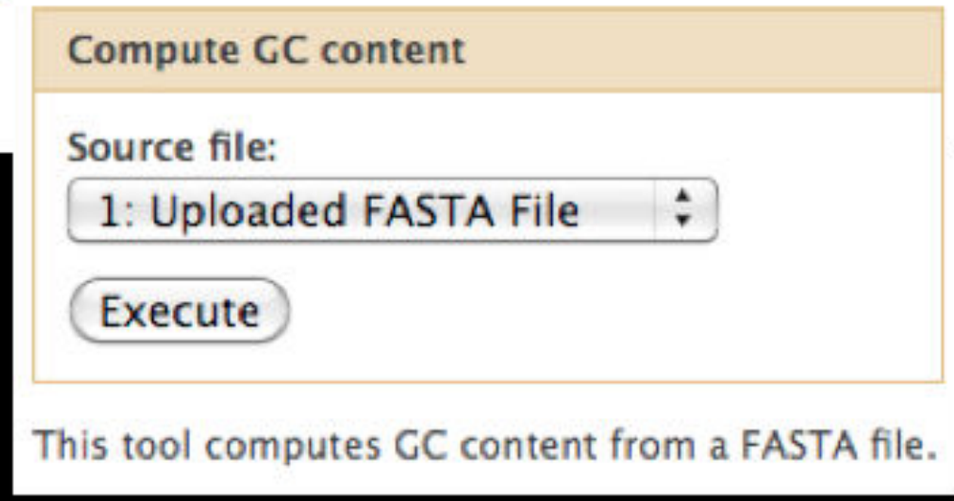
- Input and Output Datasets
- Input Parameters

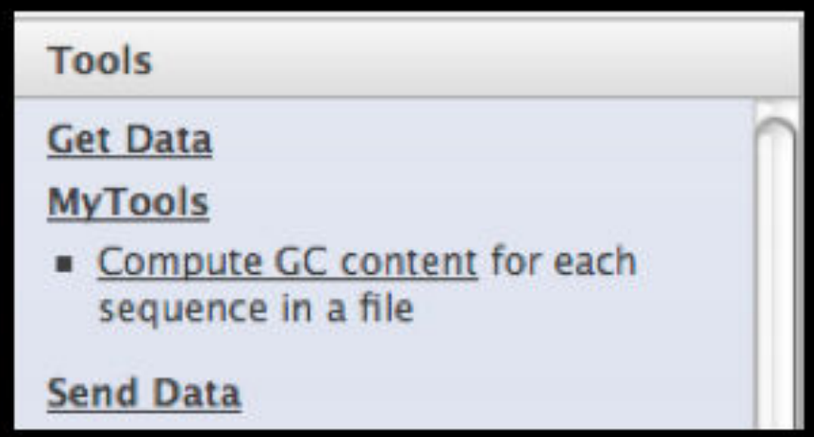
Construct a descriptive tool configuration XML file

- Write a wrapper script, only if required

A Basic Tool

```
<tool id="fa_gc_content_1" name="Compute GC content">
  <description>for each sequence in a file</description>
  <command interpreter="perl">toolExample.pl $input $output</command>
  <inputs>
    <param format="fasta" name="input" type="data" label="Source file'
  </inputs>
  <outputs>
    <data format="tabular" name="output" />
  </outputs>

  <tests>
    <test>
      <param name="input" value="fa_gc_content_input.fa"/>
      <output name="out_file1" file="fa_gc_content_output.txt"/>
    </test>
  </tests>

  <help>
This tool computes GC content from a FASTA file.
  </help>

</tool>
```






```
<section name="MyTools" id="mTools">
    <tool file="myTools/toolExample.xml" />
 </section>
```

tool_conf.xml

Cluster

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

max distance between intervals: 1 (bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

Execute

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

---

Screencasts!

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

Syntax

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

Example

cluster.xml

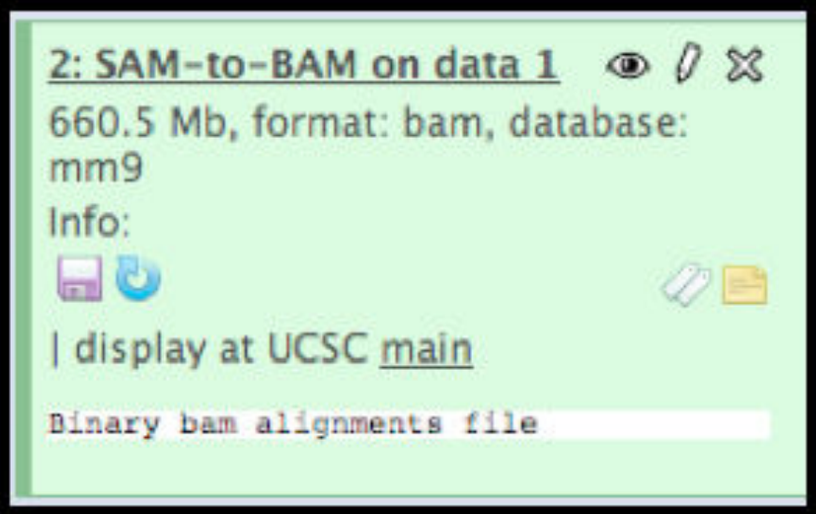
```
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="python2.4">
    gops_cluster.py $input1 $output -1 $input1_chromCol,$input1_startC
                   -d $distance -m $minregions -o $returntype
  </command>
  <inputs>
    <param format="interval" name="input1" type="data">
      <label>Cluster intervals of</label>
    </param>
    <param name="distance" size="5" type="integer" value="1" help="(bp
      <label>max distance between intervals</label>
    </param>
    <param name="minregions" size="5" type="integer" value="2">
      <label>min number of intervals per cluster</label>
    </param>
    <param name="returntype" type="select" label="Return type">
      <option value="1">Merge clusters into single intervals</option>
      <option value="2">Find cluster intervals; preserve comments and
      <option value="3">Find cluster intervals; output grouped by clus
      <option value="4">Find the smallest interval in each cluster</op
      <option value="5">Find the largest interval in each cluster</opt
    </param>
  </inputs>
  <help>

.. class:: infomark

**TIP:** If your query does not appear in the pulldown menu -> it is n

-----

**Screencasts!**

See Galaxy Interval Operation Screencasts_ (right click to open this l

.. _Screencasts: http://www.bx.psu.edu/cgi-bin/trac.cgi/wiki/GopsDesc

-----

**Syntax**

- **Maximum distance** is greatest distance in base pairs allowed betw
- **Minimum intervals per cluster** allow a threshold to be set on the
- **Merge clusters into single intervals** outputs intervals that span
- **Find cluster intervals; preserve comments and order** filters out
- **Find cluster intervals; output grouped by clusters** filters out n
```

Line: 87 Column: 8 XML Soft Tabs: 2

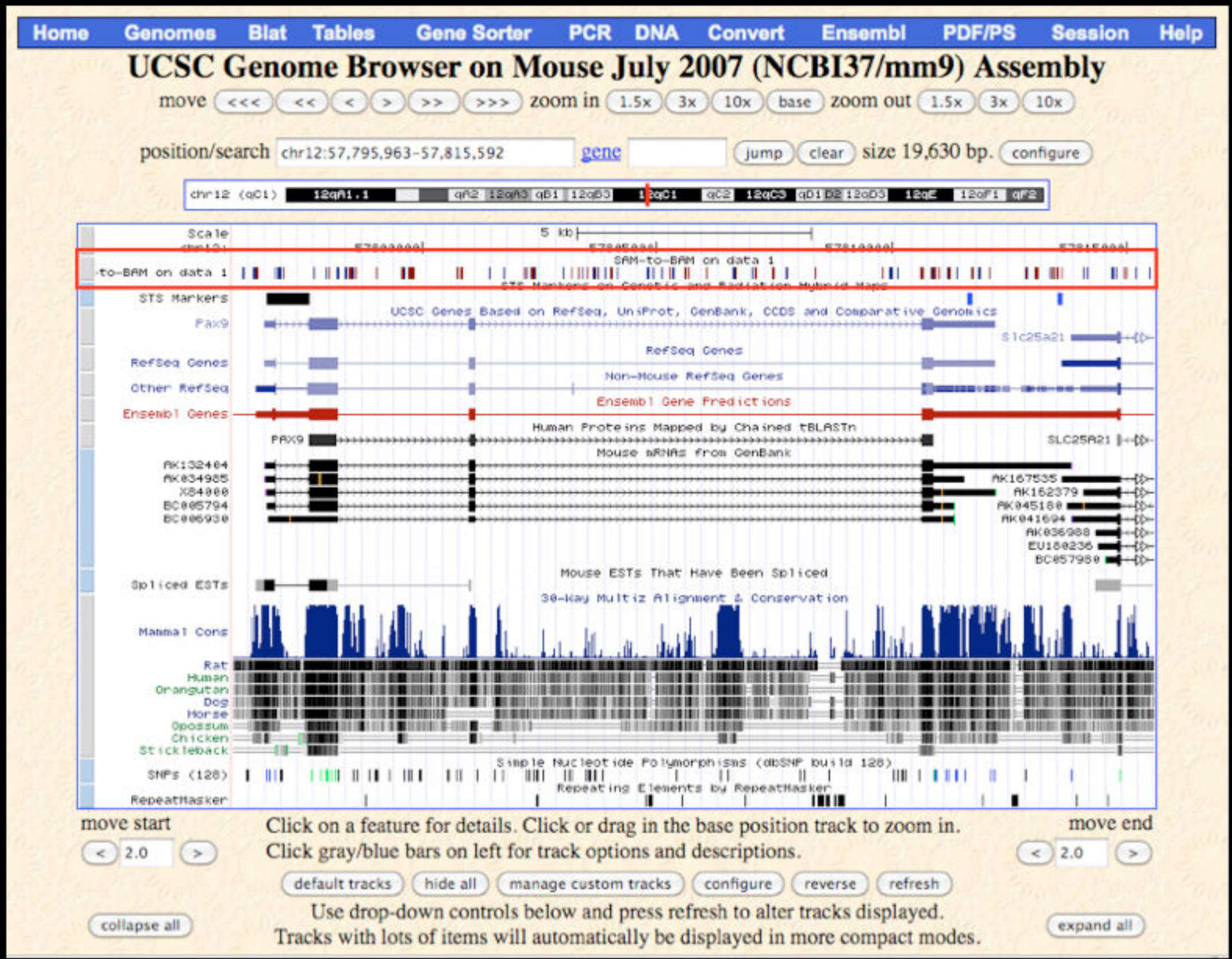
External Display Application

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

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



BAM at UCSC


Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl PDF/PS Session Help
UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr12:57,795,963-57,815,592 gene jump clear size 19,630 bp. configure
chr12 (qC1) 12qA1.1 qA2 12qA3 qB1 12qB3 12qC1 qC2 12qC3 qD1 D2 12qD3 12qE 12qF1 qF2
Scale
5 kb
SAM-to-BAM on data 1
STS Markers
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Pax9
S1c25a21
RefSeq Genes
Non-Mouse RefSeq Genes
Other RefSeq
Ensembl Gene Predictions
Ensembl Genes
Human Proteins Mapped by Chained tBLASTn
PAX9
SLC25A21
Mouse mRNAs from GenBank
AK132404
AK034985
X84000
BC005794
BC006930
AK167535
AK162379
AK045180
AK041694
AK036988
EU160236
BC057959
Mouse ESTs That Have Been Spliced
Spliced ESTs
30-Way Multiz Alignment & Conservation
Mammal Cons
Rat
Human
Orangutan
Dog
Horse
Opossum
Chicken
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 128)
SNPs (128)
Repeating Elements by RepeatMasker
RepeatMasker
move start
< 2.0 >
Click on a feature for details. Click or drag in the base position track to zoom in.
Click gray/blue bars on left for track options and descriptions.
move end
< 2.0 >
default tracks hide all manage custom tracks configure reverse refresh
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.
collapse all
expand all

Goals

What is Galaxy

- Mission, Terminology, and Capabilities

Using Galaxy

- Play along with a basic Analysis

How to get involved with Galaxy

- The community and how to participate

Adding Tools

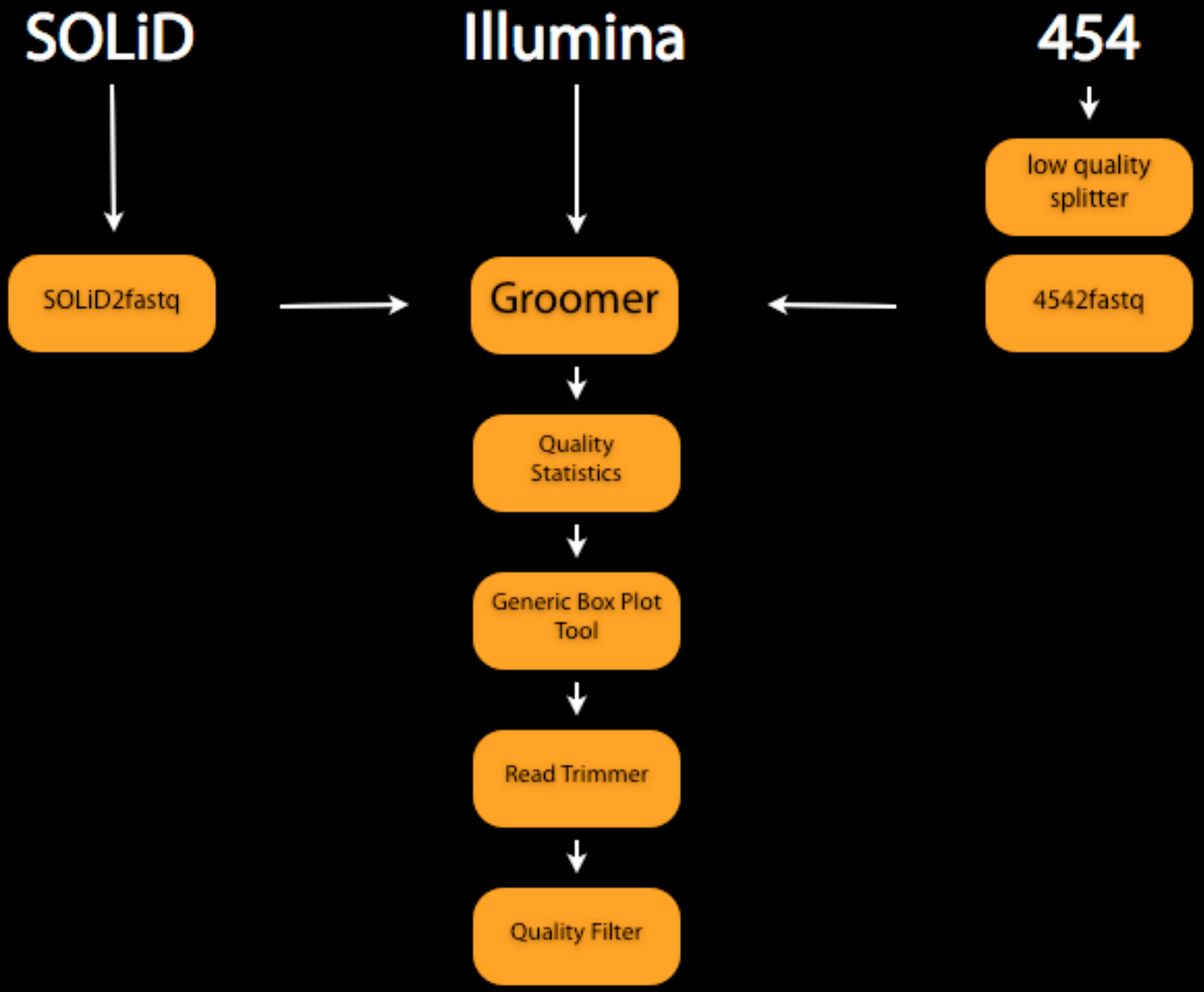
- Making your code Accessible and Reproducible

Bonus: Brief Intro to Sequence Analysis in Galaxy

Using Galaxy to Analyze NGS Data

- Prepare, quality control and manipulate reads
- Read Mapping
- ChIP-Seq: Binding sites analysis and peak calling

Prepare and Quality Check



Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A; Galaxy Team. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15;26(14):1783-5.

What is FASTQ?

- Specifies sequence (FASTA) and quality scores (PHRED)
- Text format, 4 lines per entry

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- FASTQ is such a cool standard, there are 3 (or 5) of them!

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
  ............................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger       Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
 I - Illumina 1.3 Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
 X - Solexa       Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

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

Combining Sequences and Qualities

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

Tools — Options

- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

- Filter FASTQ rea… score and length…
- FASTQ Trimmer …
- FASTQ Quality T… sliding window
- FASTQ Masker b…

Combine FASTA and QUAL

FASTA File: 1: 454.fasta

Quality Score File: 2: 454.qual

Force Quality Score encoding: ASCII

Execute

What it does

This tool joins a FASTA file to a Quality Score file, creating a single FASTQ block for each read.

Specifying a set of quality scores is optional; when not provided, the output will be fastqsanger or fastqcssanger (when a csfasta is provided) with each quality score being the maximal allowed value (93).

Use this tool, for example, to convert 454-type output to FASTQ.

History — Options

Combine QUAL and Sequence

2: 454.qual
52 lines
format: qual454, database: ?
Info: uploaded qual454 file

```
>EYKX4VC01B65GS length=54 xy=0784_1
33 23 34 25 28 28 28 32 23 34 27 4
>EYKX4VC01BNCSP length=187 xy=0558_
27 35 26 25 37 28 37 28 25 28 27 36
22 9 23 19 28 28 28 28 26 28 39 32
26 27 37 29 28 26 28 36 28 26 24 38
```

1: 454.fasta
18 sequences
format: fasta, database: ?
Info: uploaded fasta file

```
>EYKX4VC01B65GS length=54 xy=0784_1
CCGGTATCCGGGTGCCGTGATGAGCGCCACCGAA
>EYKX4VC01BNCSP length=187 xy=0558_
CTTACCGGTCACCACCGTGCCTTCAGGATTGATCG
GGTGACATCGCCCACCACGGTACTCACTGGCTGGC
CACCACGTTGAGGGTATTCCCCTCGGTTTGTGGCT
```

```
@EYKX4VC01B65GS length=54 xy=0784_1754 region=1 run=R_2007_11_07_16_15_57_
CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAACGAATTCGACTATGCCGAA
+
B8C:===A8C<%==@6=<<=======B8=B9E<@6==B;B9<=======A8=C:
@EYKX4VC01BNCSP length=187 xy=0558_3831 region=1 run=R_2007_11_07_16_15_57_
CTTACCGGTCACCACCGTGCCTTCAGGATTGATCGCCAGATCGGTCGGTGCGTCAGGCGGGGTGACATCGCCCACCACGGTACTCACTGGCTGGCTCTGGTTCCCGGCGGCATCGGAG
+
<D;:F=F=:=<E<=E<==<E<?4<=E=8E<<=<<=F><;<99E<;=E=9:6=9=;C:;LE7*84====;=HA-<E==;F==;===<=;E<<<E=<<==E<E=HA-D=;F>====F>=E
@EYKX4VC01CD9FT length=115 xy=0865_1719 region=1 run=R_2007_11_07_16_15_57_
GGGGGCTTTGGCCTGTCGTCCGGCACCTCGCAAGAGCTACAGCAGGCGCGGCTGGCGATCATCGGCGGCACGCCGGCCTATATGTCGCCGGAACACACCACCCGCACCCAACGCG
+
D91*#<HB.E<E<===<==B8F===E<====E<===<=======F====F=;=E<========F==D;<==<E<D:A7=========C:E<C:=<==E<=D>'===F?)B9=<<<
@EYKX4VC01B8FW0 length=95 xy=0799_0514 region=1 run=R_2007_11_07_16_15_57_
TAAATTTCAAGGAATGCAAATCAGGGTCGTGTGTTTAGACTTCGGCTTTAGAGACCTGAATACGTCAAAAACATAACTTCATGATATCTTGCAGT
+
=IC0D='<B8C9A7===JC2===F?*=====<=F?)==<=D;<D;=F?*=<===C:==A7;====<LE8-"=6=<1=A8<=<==<===A7=;;<=
@EYKX4VC01BCGYW length=115 xy=0434_3926 region=1 run=R_2007_11_07_16_15_57_
GGCCAGCCGGGACAGCGTTGTTGGGCTGCATGGCGACGAGCTAAAAGTCGCCATCACCGCCCCGCCGGTTGATGGGCAGGCTAATGCCCATCTGGTAAAAACTTTCTCGCCAAAC
+
=';0<=F=JD2=6=86<E<9E=IC/7:=9<=F=;=<<====<LE7)=;=<;/=:5=C9:IB3"4<1E=E=6<:JC17=F>;;D<=;JC1==<=F>:LE8-",HA-=25==2E>(9
@EYKX4VC01AZXC6 length=116 xy=0292_0280 region=1 run=R_2007_11_07_16_15_57_
GGGGGCGTTTGGCCTGTCGTCCGGCACCTCGCAAGAGCTACAGCAGGCGCGGCTGGCGATCATCGGCGGCACGCCGGCCTATATGTCGCCGGAACACACCACCCGCACCCAACGCG
+
```

Grooming --> Sanger

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

Tools Options

NGS TOOLBOX BETA

NGS: QC and manipulation

ILLUMINA DATA

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ

FASTQ Groomer

File to groom:
3: Combine FASTA and.. and data 2

Input FASTQ quality scores type:
Sanger
Solexa
Illumina 1.3+
Sanger
Color Space Sanger
Execute

What it does

This tool offers several conversions options relating to the FA

When using *Basic* options, the output will be *sanger* formatte Space Sanger).

When converting, if a quality score falls outside of the target able value (i.e. the minimum or maximum).

When converting between Solexa and the other formats, qual scales using the equations found in Cock PJ, Fields CJ, Goto N, Heuer ML, ces with quality scores, and the Solexa/Illumina FASTQ variants. Nucle

When converting between color space (csSanger) and base/se adapter bases are lost or gained; if gained, the base 'G' is used as the adapt space if there is no adapter present in the color space sequence. Any masked or ambiguous nucleotides in base space will be converted to 'N's when determining color space encoding.

4: FASTQ Groomer on data 3

18 sequences
format: fastqsanger, database: ?
Info: Groomed 18 sanger reads into sanger reads.
Based upon quality and sequence, the input data is valid for: sanger
Input ASCII range: '!'(33) - 'L'(76)
Input decimal range: 0 - 43

```
@EYKX4VC01B65GS length=54 xy=0784_1
CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAA
+
B8C:===A8C<%==@6=<<=======B8=B9E<@6
@EYKX4VC01BNCSP length=187 xy=0558_
CTTACCGGTCACCACCGTGCCTTCAGGATTGATCG
```

History Options

Combine QUAL and Sequence

3: Combine FASTA and QUAL on data 1 and data 2

18 sequences
format: fastqsanger, database: ?
Info:
Combined 18 of 18 sequences with quality scores (100.00%).

```
@EYKX4VC01B65GS length=54 xy=0784_1
CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAA
+
B8C:===A8C<%==@6=<<=======B8=B9E<@6
@EYKX4VC01BNCSP length=187 xy=0558_
CTTACCGGTCACCACCGTGCCTTCAGGATTGATCG
```

2: 454.qual

52 lines
format: qual454, database: ?
Info: uploaded qual454 file

```
>EYKX4VC01B65GS length=54 xy=0784_1
33 23 34 25 28 28 28 32 23 34 27 4
>EYKX4VC01BNCSP length=187 xy=0558_
27 35 26 25 37 28 37 28 25 28 27 36
```

Quality Score Comparison

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger       Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
I - Illumina 1.3 Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
X - Solexa       Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

Diagram adapted from http://en.wikipedia.org/wiki/FASTQ_format

Quality Statistics and Box Plot Tool

NGS TOOLBOX BETA

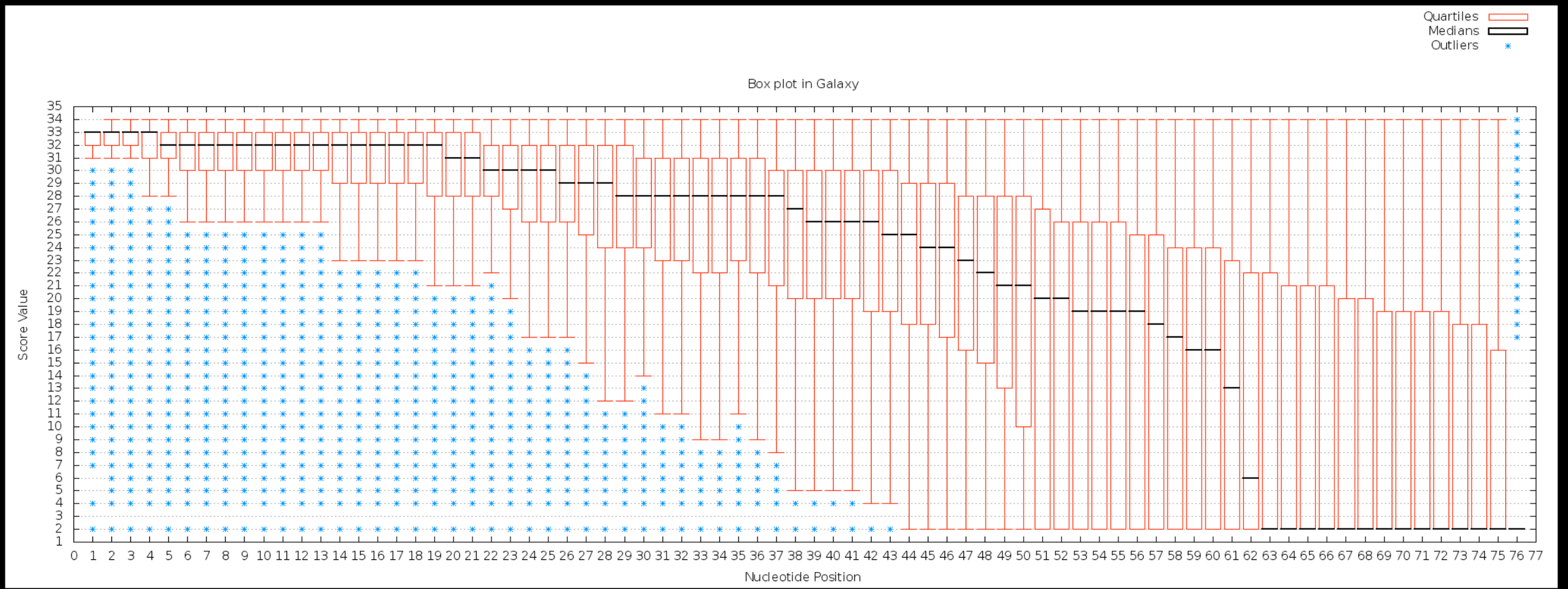
NGS: QC and manipulation

ILLUMINA DATA

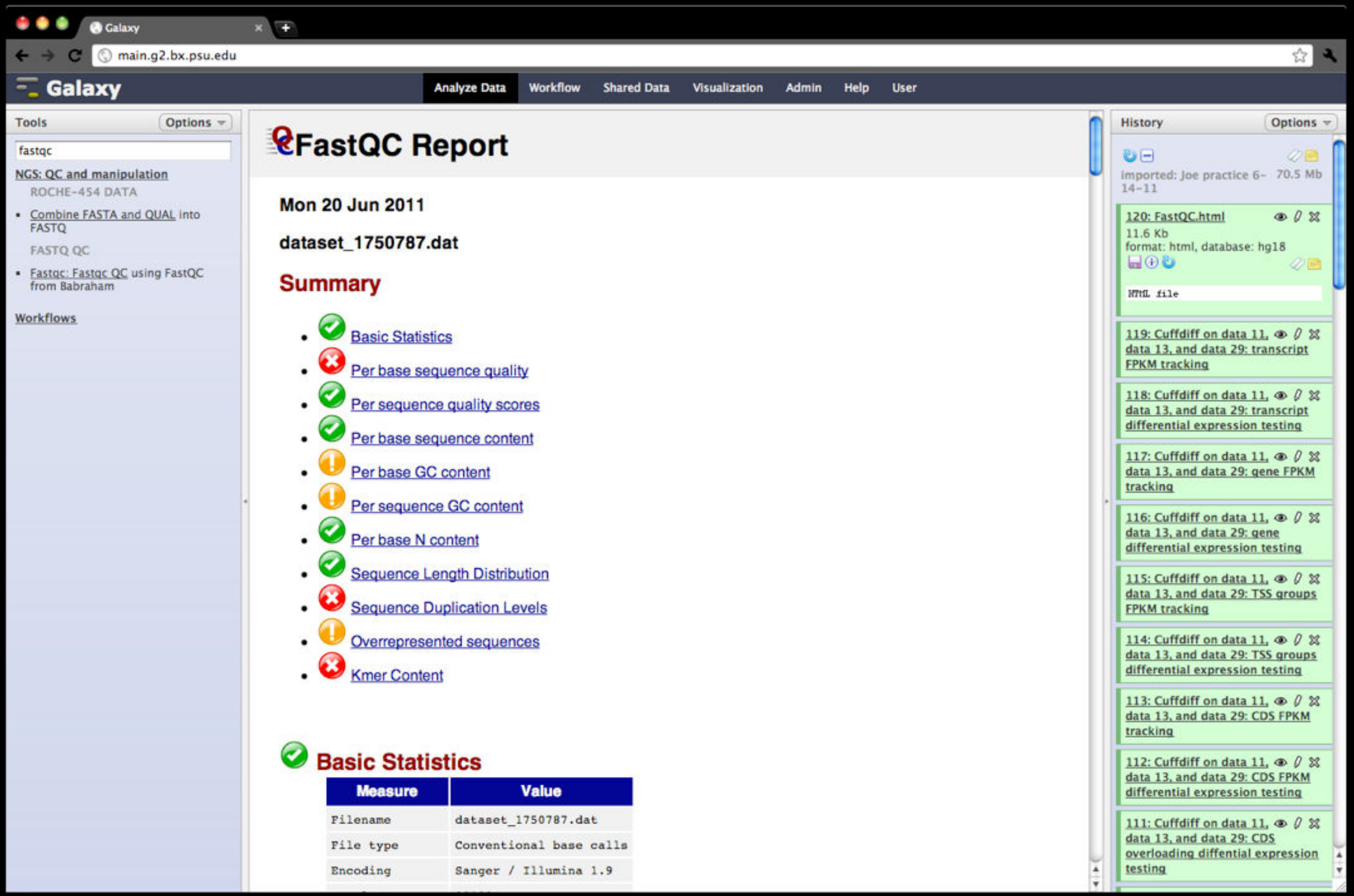
- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

Graph/Display Data

- Histogram of a numeric column
- Scatterplot of two numeric columns
- Plotting tool for multiple series and graph types
- Boxplot of quality statistics



FastQC


Galaxy
main.g2.bx.psu.edu
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Tools
Options
fastqc
NGS: QC and manipulation
ROCHE-454 DATA
Combine FASTA and QUAL into FASTQ
FASTQ QC
Fastqc: Fastqc QC using FastQC from Babraham
Workflows
FastQC Report
Mon 20 Jun 2011
dataset_1750787.dat
Summary
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content
Basic Statistics
Measure
Value
Filename
dataset_1750787.dat
File type
Conventional base calls
Encoding
Sanger / Illumina 1.9
History
Options
imported: Joe practice 6-14-11
70.5 Mb
120: FastQC.html
11.6 Kb
format: html, database: hg18
HTML file
119: Cuffdiff on data 11, data 13, and data 29: transcript FPKM tracking
118: Cuffdiff on data 11, data 13, and data 29: transcript differential expression testing
117: Cuffdiff on data 11, data 13, and data 29: gene FPKM tracking
116: Cuffdiff on data 11, data 13, and data 29: gene differential expression testing
115: Cuffdiff on data 11, data 13, and data 29: TSS groups FPKM tracking
114: Cuffdiff on data 11, data 13, and data 29: TSS groups differential expression testing
113: Cuffdiff on data 11, data 13, and data 29: CDS FPKM tracking
112: Cuffdiff on data 11, data 13, and data 29: CDS FPKM differential expression testing
111: Cuffdiff on data 11, data 13, and data 29: CDS overloading diffential expression testing

Read Trimming

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

Tools | Options

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter

FASTX-TOOLKIT FOR FASTQ DATA

- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- Remove sequencing artifacts

FASTQ Trimmer

FASTQ File: 2: imported: GM12878..ple Dataset

Define Base Offsets as: Absolute Values

Use Absolute for fixed length reads (Illumina, SOLiD
Use Percentage for variable length reads (Roche/45

Offset from 5' end: 0

Values start at 0, increasing from the left

Offset from 3' end: 16

Values start at 0, increasing from the right

Keep reads with zero length: ☐

Execute

This tool allows you to trim the ends of reads.

You can specify either absolute or percent-based offs
trimmed. When using the percent-based method, offs

For example, if you have a read of length 36:

```
@Some FASTQ Sanger Read
CAATATGTNCTCACTGATAAGTGGATATNAGCNCCA
+
=@@.@;B-%78>CBA@>7@7BBCA4-48%<;;%<B@
```

And you set absolute offsets of 2 and 0:

FASTQ Quality Trimmer

FASTQ File: 7: FASTQ Trimmer on data 2

Keep reads with zero length: ☐

Trim ends: 5' and 3'

Window size: 1

Step Size: 1

Maximum number of bases to exclude from the window during aggr

0

Aggregate action for window: min score

Trim until aggregate score is: >=

Quality Score: 0.0

Execute

Filter FASTQ

FASTQ File:

7: FASTQ Trimmer on data 2

Requires groomed data: if your data does not appear here try using the FASTQ groomer.

Minimum Size:

0

Maximum Size:

0

A maximum size less than 1 indicates no limit.

Minimum Quality:

0.0

Maximum Quality:

0.0

A maximum quality less than 1 indicates no limit.

Maximum number of bases allowed outside of quality range:

0

This is paired end data:

☐

Quality Filter on a Range of Bases

Add new Quality Filter on a Range of Bases

Execute

Quality Filter on a Range of Bases

Quality Filter on a Range of Bases 1

Define Base Offsets as:

Absolute Values

Use Absolute for fixed length reads (Illumina, SOLiD)
Use Percentage for variable length reads (Roche/454)

Offset from 5' end:

0

Values start at 0, increasing from the left

Offset from 3' end:

0

Values start at 0, increasing from the right

Aggregate read score for specified range:

min score

Keep read when aggregate score is:

>=

Quality Score:

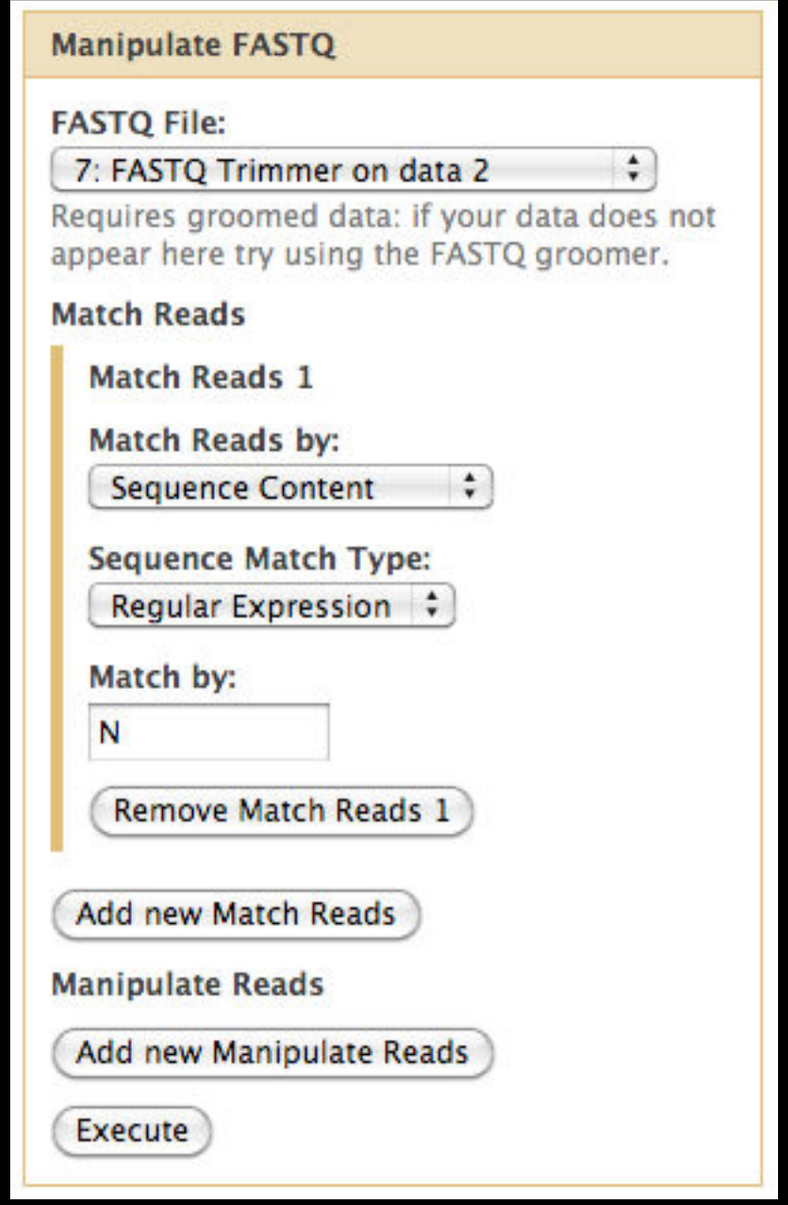
0.0

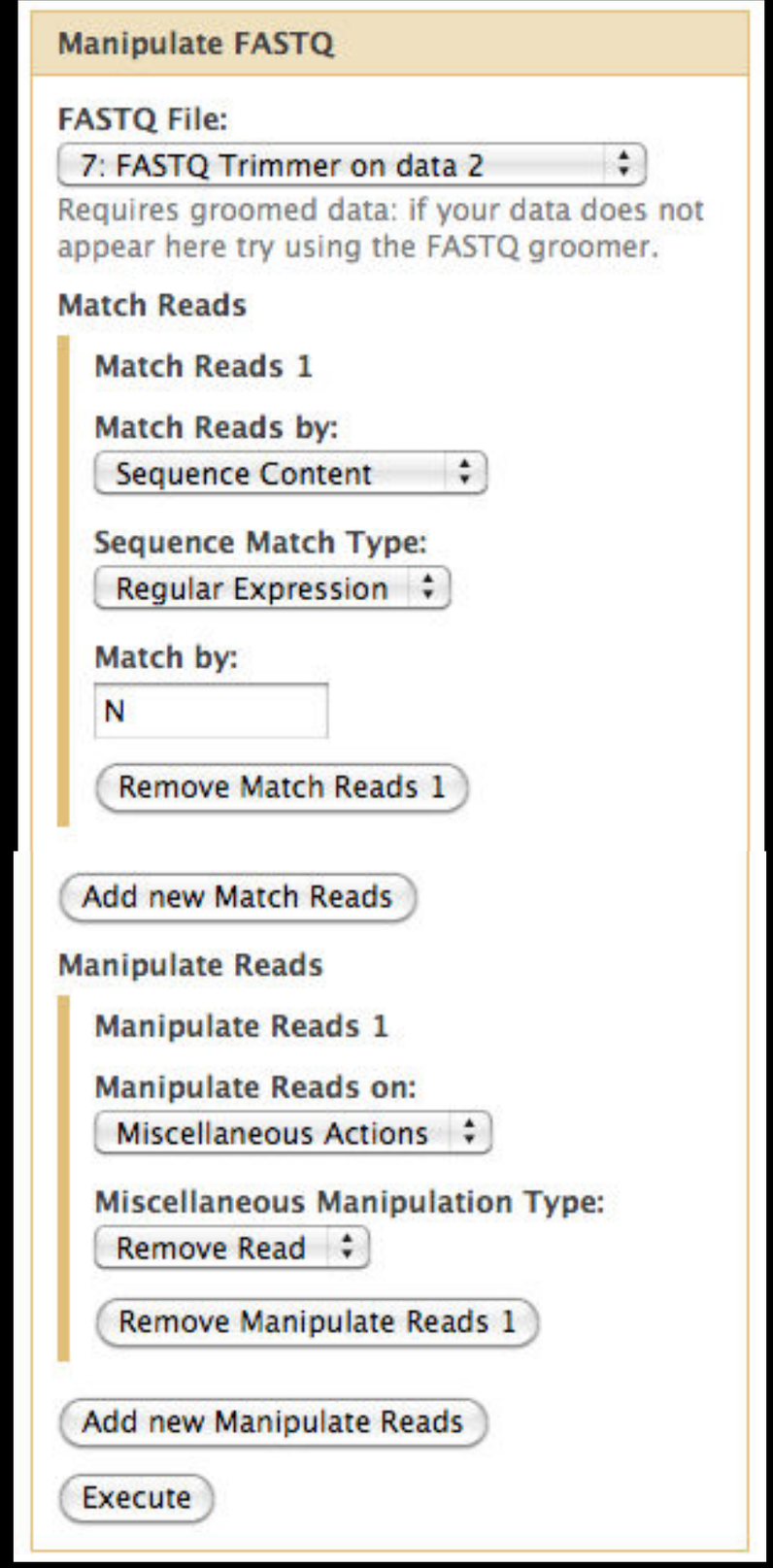
Remove Quality Filter on a Range of Bases 1

Add new Quality Filter on a Range of Bases

Execute

Manipulate FASTQ





Using Galaxy to Analyze NGS Data

- Prepare, quality control and manipulate reads
- Read Mapping
- ChIP-Seq: Binding sites analysis and peak calling

Mapping NGS Data

Collection of interchangeable mappers

- accept fastq format, produce SAM/BAM

Mappers for

- DNA
- RNA

Mappers

DNA

- short reads: Bowtie, BWA, BFAST, PerM
- longer reads: LASTZ

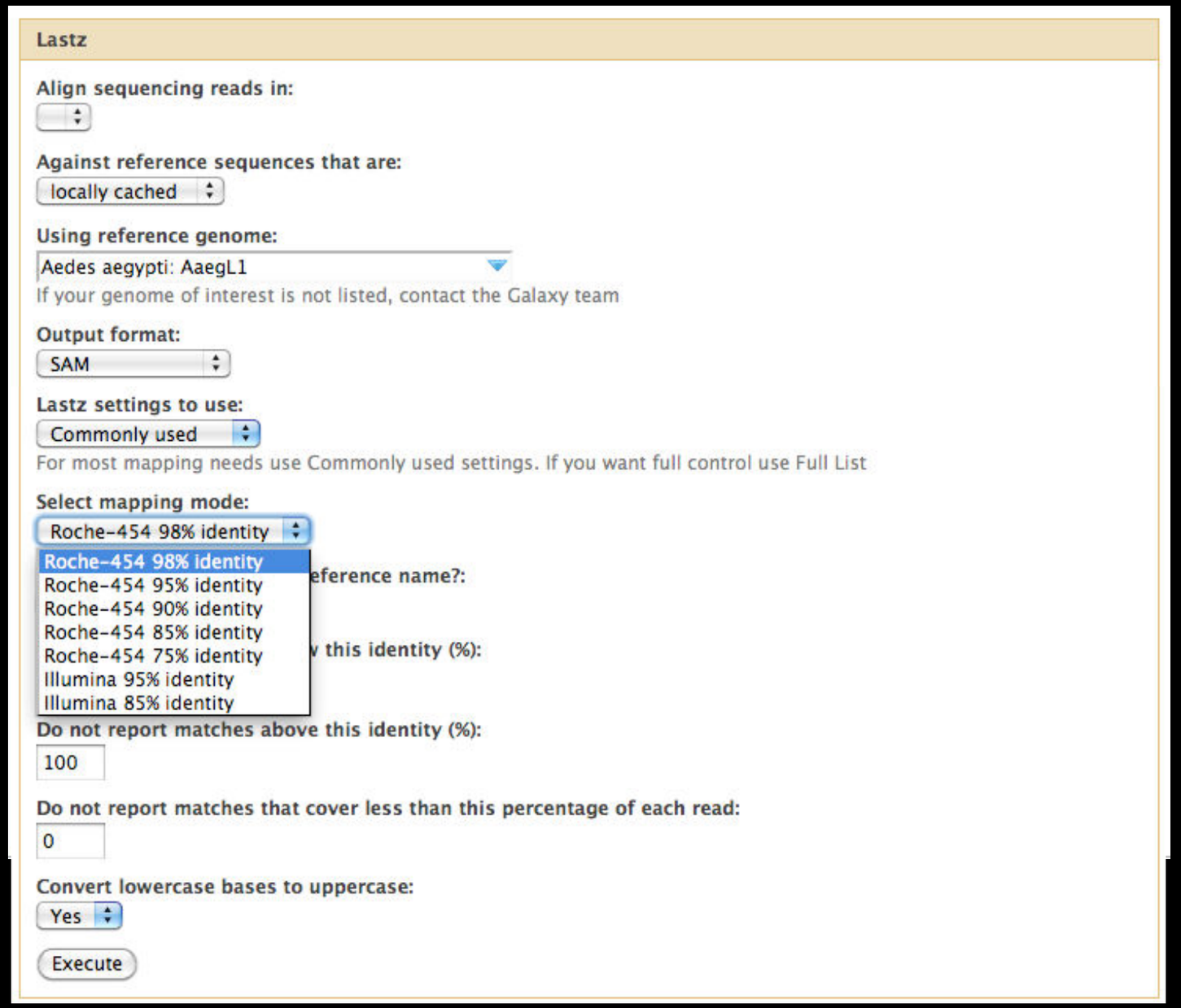
Metagenomics

- Megablast

RNA / gapped-reads mapper

- Tophat

Commonly Used/Default Parameters



Full Parameter List

Lastz

Align sequencing reads in:
53: FASTQ to FASTA on data 7

Against reference sequences that are:
locally cached

Using reference genome:
Aedes aegypti: AaegL1
If your genome of interest is not listed, contact the Galaxy team

Output format:
SAM

Lastz settings to use:
Full Parameter List
Commonly used
Full Parameter List
use Commonly used settings. If you want full control use Full List

Which strand to search?:
Both

Select seeding settings:
Seed hits require a 19 bp word with matches i
allows you set word size and number of mismatches

Select transition settings:
Allow one transition in each seed hit
affects the number of allowed transition substitutions

Perform gap-free extension of seed hits to HSPs (high scoring segment pairs)?:
No

Perform chaining of HSPs?:
No

Gap opening penalty:
400

Gap extension penalty:
30

X-drop threshold:
910

Y-drop threshold:
9370

Set the threshold for HSPs (ungapped extensions scoring lower are discarded):
3000

Set the threshold for gapped alignments (gapped extensions scoring lower are discarded):
3000

Involve entropy when filtering HSPs?:
No

Do you want to modify the reference name?:
No

Do you want to modify the reference name?:
No

Do not report matches below this identity (%):
0

Do not report matches above this identity (%):
100

Do not report matches that cover less than this percentage of each read:
0

Convert lowercase bases to uppercase:
Yes

Execute

What it does

LASTZ is a high performance pairwise sequence aligner derived from BLASTZ. It is written by Bob Harris in Webb Miller's laboratory at Penn State University. Special scoring sets were derived to improve runtime performance and quality. This Galaxy version of LASTZ is geared towards aligning short (Illumina/Solexa, AB/SOLiD) and medium (Roche/454) reads against a reference sequence. There is excellent, extensive documentation on LASTZ available here.

Input formats

LASTZ accepts reference and reads in FASTA format. However, because Galaxy supports implicit format conversion the tool will recognize fastq and other method specific formats.

Using Galaxy to Analyze NGS Data

- Prepare, quality control and manipulate reads
- Read Mapping
- ChIP-Seq: Binding sites analysis and peak calling

Peak Calling / ChIP-seq analysis

Punctate binding

- transcription factors

Diffuse binding

- histone modifications
- PolII

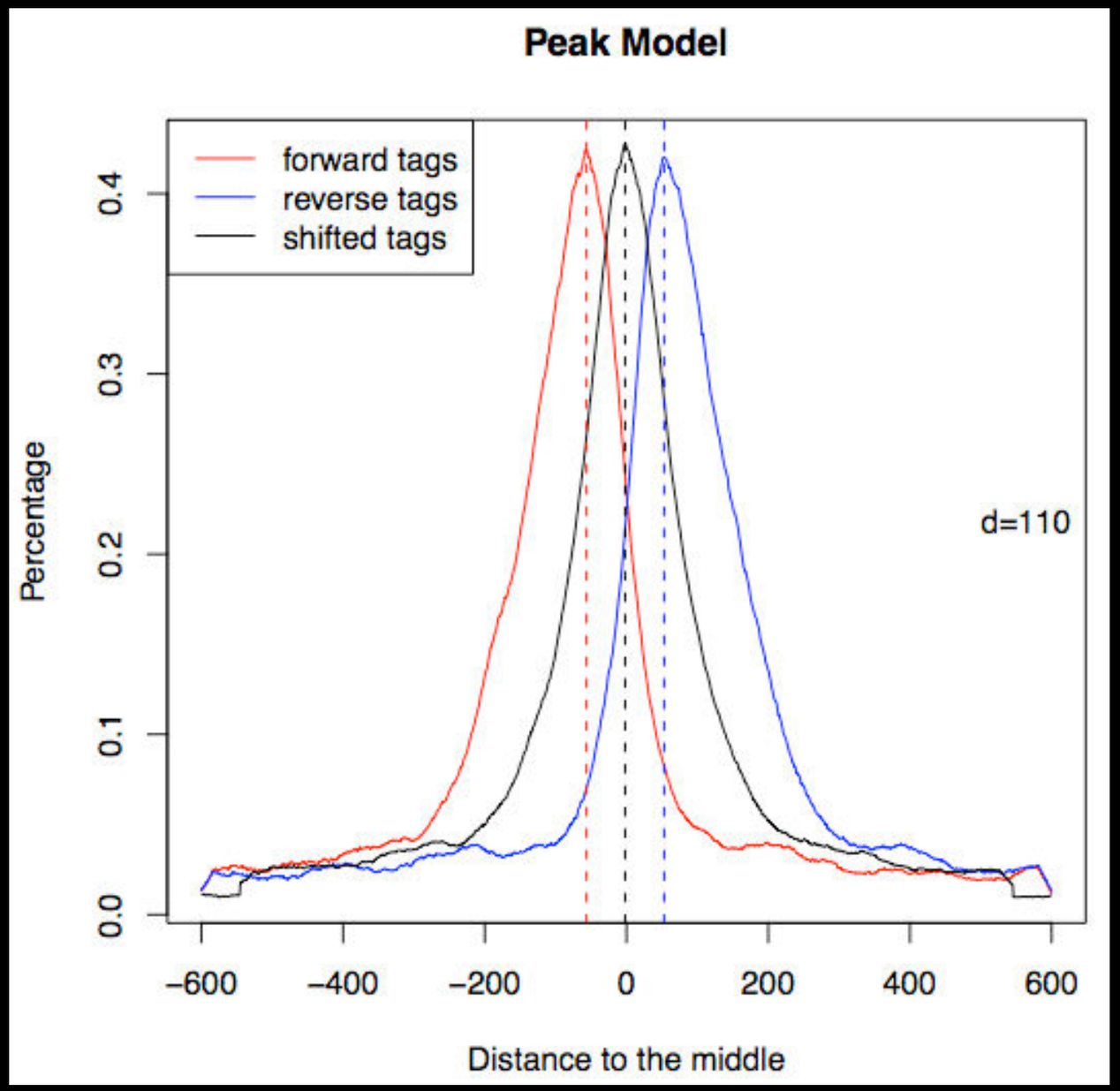
Punctate Binding --> MACS

Inputs

- Enriched Tag file
- Control / Input file (optional)

Outputs

- Called Peaks
- Negative Peaks (when control provided)
- Shifted Tag counts (wig, convert to bigWig for visualization)



Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

Diffuse Binding

SICER (version 0.0.1)

ChIP-Seq Tag File:
1: Mapped Tags

ChIP-Seq Control File:
2: Mapped Control

Fix off-by-one errors in output files:
☑
SICER creates non-standard output files, this option will fix these coordinates

Redundancy Threshold:
1
The number of copies of identical reads allowed in a library

Window size:
200
Resolution of SICER algorithm. For histone modifications, one can use 200 bp

Fragment size:
150
for determination of the amount of shift from the beginning of a read to the center of the DNA fragment represented by the read. FRAGMENT_SIZE=150 means the shift is 75.

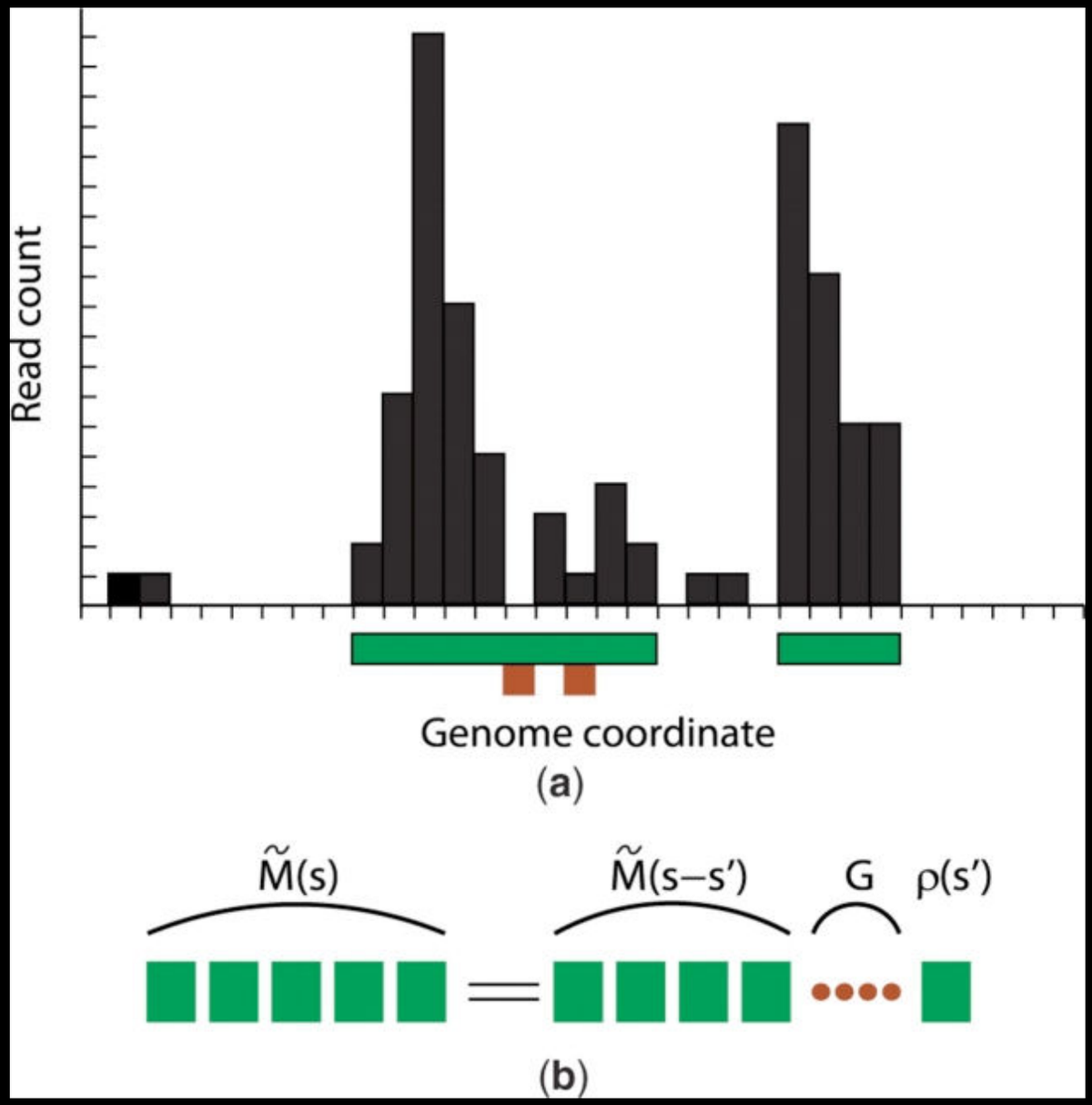
Effective genome fraction:
0.74
Effective Genome as fraction of the genome size. It depends on read length.

Gap size:
600
Needs to be multiples of window size. Namely if the window size is 200, the gap size should be 0, 200, 400, 600, ...

Statistic threshold value:
0.01
FDR (with control) or E-value (without control)

Execute

SICER



Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009 Aug 1;25(15):1952-8. Epub 2009 Jun 8.

Hands on NGS: ChIP-Seq

This time we will be using Data from a
Shared Data Library

gcc2015-X.dblankenberg.org
(where *X* is 1 through 3)