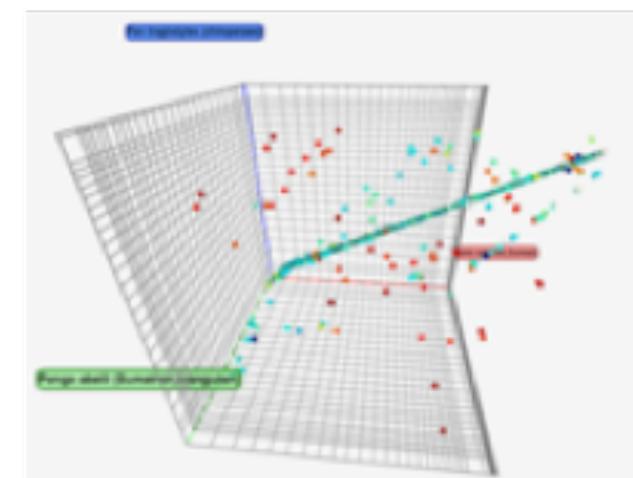
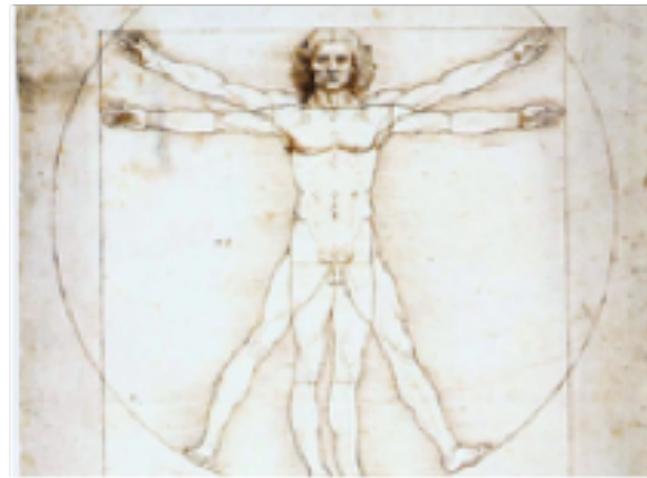
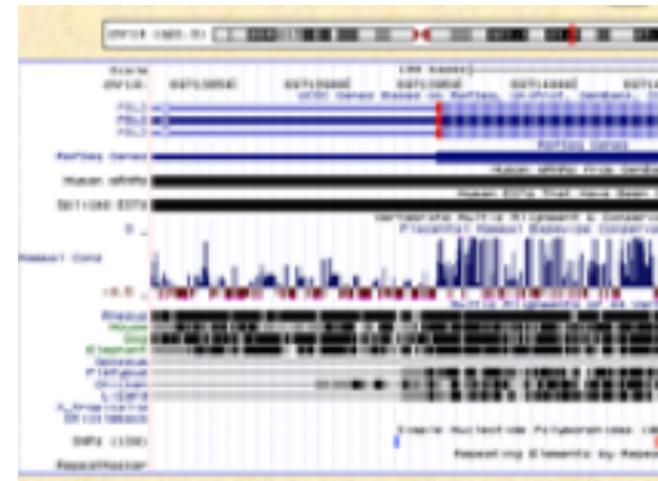


Computational Genomics

Computational Arithmetics I



Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome
Retrieve all Gencode Exons from Human Genome (hg38)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

Tools

search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

Mapping

NCBI BLAST+

Assembly

Annotation

FASTA/FASTQ

Datamash

EMBOSS

Multiple Alignments

Picard

BED

SAM/BAM

Proteomics

History

search datasets

Unnamed history

(empty)

Best Practices for Kaiser Galaxy

- Kaiser Galaxy (docs, slides) is configured for teaching purposes so all users have a file quota of 1TB. How to permanently delete nonessential files.
- FTP uploads are removed from ftp directory 48 hours after uploading so import your ftp files into Galaxy the same day as you upload to ftp
- Only certain tools that support multi-core processing have the Job Resources Parameters option which allow you to select cores, memory and time.
- The default job resource parameters for all tools is 1 core with 2GB memory for 24 hours (24 SUs).
 - Configuring a job to use 28 cores for 1 hour requires 28 SUs. (28 cores for 168 hours = 4704 SUs).
 - Configuring a job to use 54GB memory for 1 hour requires 28 SUs. (54GB memory for 168 hours = 4704 SUs).
 - If a tool you used failed because it needs the Job Resource Parameters option added, contact the HPRC helpdesk.



COVID-19 related research on Galaxy: [training](#), [tutorials](#), [documents](#)

Current known issues: Some tools cannot be removed from favorites. Choose your favorites wisely.

Uploading data from your computer from Galaxy Project

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Retrieve all Gencode Exons from Human Genome (hg38)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾  

Tools    

HPRC
Get Data
Send Data
Collection Operations
Lift-Over
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
Operate on Genomic Intervals
Statistics
Graph/Display Data
Phenotype Association
FASTQ Quality Control
Mapping
NCBI BLAST+
Assembly
Annotation
FASTA/FASTQ
Datamash
EMBOSS
Multiple Alignments
Picard
BED
SAM/BAM
Proteomics

Best Practices for Kaiser Galaxy

Name Your History!

time.

- The default job resource parameters for all tools is 1 core with 2GB memory for 24 hours (24 SUs).
 - Configuring a job to use 28 cores for 1 hour requires 28 SUs. (28 cores for 168 hours = 4704 SUs).
 - Configuring a job to use 54GB memory for 1 hour requires 28 SUs. (54GB memory for 168 hours = 4704 SUs).
 - If a tool you used failed because it needs the Job Resource Parameters option added, contact the HPRC helpdesk.



COVID-19 related research on Galaxy: [training](#), [tutorials](#), [documents](#)

Current known issues: Some tools cannot be removed from favorites. Choose your favorites wisely.

 Uploading data from your computer
from Galaxy Project

History      

JSG_History
(empty)  

This history is empty. You can load your own data or get data from an external source

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome
Retrieve all Gencode Exons from Human Genome (hg38)

HPRC Kaiser Galaxy Using 2%

Tools Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

History search datasets JSG_History (empty)

Best Practices for Kaiser Galaxy

- Kaiser Galaxy (docs, slides) is configured for teaching purposes so all users have a file quota of 1TB. How to permanently delete nonessential files.
- FTP uploads are removed from ftp directory 48 hours after uploading so import your ftp files into Galaxy the same day as you upload to ftp
- Only certain tools that support multi-core processing have the Job Resources Parameters option which allow you to select cores, memory and time.
- The default job resource parameters for all tools is 1 core with 2GB memory for 24 hours (24 SUs).
 - Configuring a job to use 28 cores for 1 hour requires 28 SUs. (28 cores for 168 hours = 4704 SUs).
 - Configuring a job to use 54GB memory for 1 hour requires 28 SUs. (54GB memory for 168 hours = 4704 SUs).
 - If a tool you used failed because it needs the Job Resource Parameters option added, contact the HPRC helpdesk.



UCSC Main table browser

UCSC Archaea table browser

SRA server

EBI SRA ENA SRA

modENCODE fly server

InterMine server

Flymine server

modENCODE modMine server

MouseMine server

Ratmine server

YeastMine server

modENCODE worm server

WormBase server

ZebrafishMine server

EuPathDB server

HbVar Human Hemoglobin Variants and Thalassemias

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Using 2% History search datasets JSG_History (empty)

This history is empty. You can load your own data or get data from an external source

COVID-19 related research on Galaxy: [training](#), [tutorials](#), [documents](#)

Current known issues: Some tools cannot be removed from favorites. Choose your favorites wisely.

Uploading data from your computer from Galaxy Project

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome
Retrieve all Gencode Exons from Human Genome (hg38)

[Genomes](#)[Genome Browser](#)[Tools](#)[Mirrors](#)[Downloads](#)[My Data](#)[Projects](#)[Help](#)[About Us](#)

Table Browser

Use this tool to retrieve and export data from the Genome Browser annotation track database. You can limit retrieval based on data attributes and intersect or merge with data from another track, or retrieve DNA sequence covered by a track. [More...](#)

Select dataset

clade: Mammal genome: Human assembly: Dec. 2013 (GRCh38/hg38)
group: Genes and Gene Predictions track: GENCODE V39
table: knownGene [describe table schema](#)

Define region of interest

region: genome position chrX:15,560,138-15,602,945 [lookup](#) [define regions](#) ←
identifiers (names/accessions): [paste list](#) [upload list](#)

Optional: Subset, combine, compare with another track

filter: [create](#)

intersection: [create](#)

Retrieve and display data

output format: BED - browser extensible data Send output to Galaxy GREAT
output filename: (leave blank to keep output in browser)
file type returned: plain text gzip compressed

[get output](#) [summary/statistics](#)



Using the Table Browser

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome
Retrieve all Gencode Exons from Human Genome (hg38)

[Genomes](#)[Genome Browser](#)[Tools](#)[Mirrors](#)[Downloads](#)[My Data](#)[Projects](#)[Help](#)[About Us](#)

Output knownGene as BED

Include [custom track header](#):

name=

description=

visibility= ▾

url=

Create one BED record per:

Whole Gene

Upstream by bases

Exons plus bases at each end 

Introns plus bases at each end

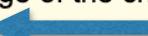
5' UTR Exons

Coding Exons

3' UTR Exons

Downstream by bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.



Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Retrieve all Gencode Exons from Human Genome (hg38)

HPRC Kaiser Galaxy Using 2%

Tools Analyze Data Workflow Visualize Shared Data Admin Help User

History search datasets JSG_History 1 shown (empty) Running 1: UCSC Main on Hum an: knownGene (genome)

Best Practices for Kaiser Galaxy

- Kaiser Galaxy (docs, slides) is configured for teaching purposes so all users have a file quota of 1TB. How to permanently delete nonessential files.
- FTP uploads are removed from ftp directory 48 hours after uploading so import your ftp files into Galaxy the same day as you upload to ftp
- Only certain tools that support multi-core processing have the Job Resources Parameters option which allow you to select cores, memory and time.
- The default job resource parameters for all tools is 1 core with 2GB memory for 24 hours (24 SUs).
 - Configuring a job to use 28 cores for 1 hour requires 28 SUs. (28 cores for 168 hours = 4704 SUs).
 - Configuring a job to use 54GB memory for 1 hour requires 28 SUs. (54GB memory for 168 hours = 4704 SUs).
 - If a tool you used failed because it needs the Job Resource Parameters option added, contact the HPRC helpdesk.



COVID-19 related research on Galaxy: [training](#), [tutorials](#), [documents](#)

Current known issues: Some tools cannot be removed from favorites. Choose your favorites wisely.

Uploading data from your computer from Galaxy Project

Navigation: Tools, search tools, HPRC, Get Data, Send Data, Collection Operations, Lift-Over, Text Manipulation, Convert Formats, Filter and Sort, Join, Subtract and Group, Fetch Alignments/Sequences, Operate on Genomic Intervals, Statistics, Graph/Display Data, Phenotype Association, FASTQ Quality Control, Mapping, NCBI BLAST+, Assembly, Annotation, FASTA/FASTQ, Datamash, EMBOSS, Multiple Alignments, Picard, BED, SAM/BAM, Proteomics.

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Retrieve all Gencode Exons from Human Genome (hg38)

HPRC Kaiser Galaxy Using 2%

Tools Analyze Data Workflow Visualize Shared Data Admin Help User

History search datasets JSG_History 1 shown 113.62 MB 1: UCSC Main on Human: knownGene (genome) Done!

Best Practices for Kaiser Galaxy

- Kaiser Galaxy (docs, slides) is configured for teaching purposes so all users have a file quota of 1TB. How to permanently delete nonessential files.
- FTP uploads are removed from ftp directory 48 hours after uploading so import your ftp files into Galaxy the same day as you upload to ftp
- Only certain tools that support multi-core processing have the Job Resources Parameters option which allow you to select cores, memory and time.
- The default job resource parameters for all tools is 1 core with 2GB memory for 24 hours (24 SUs).
 - Configuring a job to use 28 cores for 1 hour requires 28 SUs. (28 cores for 168 hours = 4704 SUs).
 - Configuring a job to use 54GB memory for 1 hour requires 28 SUs. (54GB memory for 168 hours = 4704 SUs).
 - If a tool you used failed because it needs the Job Resource Parameters option added, contact the HPRC helpdesk.



COVID-19 related research on Galaxy: [training](#), [tutorials](#), [documents](#)

Current known issues: Some tools cannot be removed from favorites. Choose your favorites wisely.

Uploading data from your computer from Galaxy Project

Navigation: Tools, Analyze Data, Workflow, Visualize, Shared Data, Admin, Help, User, History, search datasets, JSG_History, 1 shown, 113.62 MB, 1: UCSC Main on Human: knownGene (genome), Done!, Best Practices for Kaiser Galaxy, COVID-19 related research on Galaxy: training, tutorials, documents, Current known issues: Some tools cannot be removed from favorites. Choose your favorites wisely., Uploading data from your computer from Galaxy Project.

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome
Determine their lengths (using Compute an expression)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

Tools

compute an expression X

Show Sections

Replace parts of text

Mummer Align two or more sequences

Search in textfiles (grep)

Regex Find And Replace

Text transformation with sed

Column Regex Find And Replace

Compute an expression on every row

Text reformatting with awk

tac reverse a file (reverse cat)

Select lines that match an expression

Summary Statistics for any numerical column

Replace Text in a specific column

Replace Text in entire line

bedtools Compute both the depth and breadth of coverage of features in file B on the features in file A (bedtools coverage)

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules

Convert SAM to interval

MarkDuplicatesWithMateCigar examine aligned records in BAM datasets to locate duplicate molecules

EstimateLibraryComplexity assess sequence library complexity from read sequences

Filter GFF data by feature count using simple expressions

Filter GFF data by attribute using simple expressions

WORKFLOWS

Best Practices for Kaiser Galaxy

- Kaiser Galaxy (docs, slides) is configured for teaching purposes so all users have a file quota of 1TB. How to permanently delete nonessential files.
- FTP uploads are removed from ftp directory 48 hours after uploading so import your ftp files into Galaxy the same day as you upload to ftp
- Only certain tools that support multi-core processing have the Job Resources Parameters option which allow you to select cores, memory and time.
- The default job resource parameters for all tools is 1 core with 2GB memory for 24 hours (24 SUs).
 - Configuring a job to use 28 cores for 1 hour requires 28 SUs. (28 cores for 168 hours = 4704 SUs).
 - Configuring a job to use 54GB memory for 1 hour requires 28 SUs. (54GB memory for 168 hours = 4704 SUs).
 - If a tool you used failed because it needs the Job Resource Parameters option added, contact the HPRC helpdesk.



COVID-19 related research on Galaxy: [training](#), [tutorials](#), [documents](#)

Current known issues: Some tools cannot be removed from favorites. Choose your favorites wisely.

Uploading data from your computer from Galaxy Project

History

search datasets X

JSG_History

1 shown

113.62 MB X

1: UCSC Main on Human: knownGene (genome) X

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome
Determine their lengths (using Compute an expression)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

Tools

compute an expression x

Show Sections

Replace parts of text

Mummer Align two or more sequences

Search in textfiles (grep)

Regex Find And Replace

Text transformation with sed

Column Regex Find And Replace

Compute an expression on every row

Text reformatting with awk

tac reverse a file (reverse cat)

Select lines that match an expression

Summary Statistics for any numerical column

Replace Text in a specific column

Replace Text in entire line

bedtools Compute both the depth and breadth of coverage of features in file B on the features in file A (bedtools coverage)

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules

Convert SAM to interval

MarkDuplicatesWithMateCigar examine aligned records in BAM datasets to locate duplicate molecules

EstimateLibraryComplexity assess sequence library complexity from read sequences

Filter GFF data by feature count using simple expressions

Filter GFF data by attribute using simple expressions

WORKFLOWS

Compute an expression on every row (Galaxy Version 1.5)

Add expression: c3-c2

as a new column to: 1: UCSC Main on Human: knownGene (genome)

Dataset missing? See TIP below

Round result? Yes No

Avoid scientific notation Yes No

If yes, use fully expanded decimal representation when writing new columns (use only if expression produces decimal numbers).

Input has a header line with column names? No

Select Yes to be able to specify a name for the new column and have it added to the header line. If you select No, the first line will be treated as a regular line: If it is empty or starts with a # character it will be skipped, otherwise the tool will attempt to compute the specified expression on it.

✓ Execute

TIP: If your data is not TAB delimited, use *Text Manipulation->Convert*

What it does

This tool computes an expression for every row of a dataset and appends the result as a new column (field).

- Columns are referenced with **c** and a **number**. For example, **c1** refers to the first column of a tab-delimited file
- c3-c2** will add a length column to the dataset if **c2** and **c3** are start and end position

Example

If this is your input:

```
chr1 151077881 151077918 2 200 -
chr1 151081985 151082078 3 500 +
```

computing "c4*c5" will produce:

```
chr1 151077881 151077918 2 200 - 400.0
chr1 151081985 151082078 3 500 + 1500.0
```

if, at the same time, "Round result?" is set to YES results will look like this:

History

search datasets x

JSG_History

1 shown

113.62 MB x

1: UCSC Main on Human: knownGene (genome) x

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome
Determine their lengths (using Compute an expression)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

Tools

compute an expression

Show Sections

Replace parts of text

Mummer Align two or more sequences

Search in textfiles (grep)

Regex Find And Replace

Text transformation with sed

Column Regex Find And Replace

Compute an expression on every row

Text reformatting with awk

tac reverse a file (reverse cat)

Select lines that match an expression

Summary Statistics for any numerical column

Replace Text in a specific column

Replace Text in entire line

bedtools Compute both the depth and breadth of coverage of features in file B on the features in file A (bedtools coverage)

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules

Convert SAM to interval

MarkDuplicatesWithMateCigar examine aligned records in BAM datasets to locate duplicate molecules

EstimateLibraryComplexity assess sequence library complexity from read sequences

Filter GFF data by feature count using simple expressions

Filter GFF data by attribute using simple expressions

WORKFLOWS

History

search datasets

JSG_History

2 shown

113.62 MB

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)

Executed **Compute** and successfully added 1 job to the queue.

The tool uses this input:

- 1: UCSC Main on Human: knownGene (genome)

It produces this output:

- 2: Compute on data 1

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Running

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome
Determine their lengths (using Compute an expression)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

Tools

compute an expression x

Show Sections

Replace parts of text

Mummer Align two or more sequences

Search in textfiles (grep)

Regex Find And Replace

Text transformation with sed

Column Regex Find And Replace

Compute an expression on every row

Text reformatting with awk

tac reverse a file (reverse cat)

Select lines that match an expression

Summary Statistics for any numerical column

Replace Text in a specific column

Replace Text in entire line

bedtools Compute both the depth and breadth of coverage of features in file B on the features in file A (bedtools coverage)

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules

Convert SAM to interval

MarkDuplicatesWithMateCigar examine aligned records in BAM datasets to locate duplicate molecules

EstimateLibraryComplexity assess sequence library complexity from read sequences

Filter GFF data by feature count using simple expressions

Filter GFF data by attribute using simple expressions

WORKFLOWS

History

search datasets x

JSG_History

2 shown

236.08 MB x

2: Compute on data 1 x

1: UCSC Main on Human: knownGene (genome) x

Executed **Compute** and successfully added 1 job to the queue.

The tool uses this input:

- 1: UCSC Main on Human: knownGene (genome)

It produces this output:

- 2: Compute on data 1

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Done!

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Sort numeric descending on c7 (using Sort data in ascending or descending order)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

Tools

Sort data in ascending or descending order

Sort data in ascending or descending order

Samtools sort order of storing aligned sequences

Unique lines assuming sorted input file

bedtools SortBED order the intervals

Wig/BedGraph-to-bigWig converter

SortSam sort SAM/BAM dataset

SAM-to-BAM convert SAM to BAM

Aggregate datapoints Appends the average, min, max of datapoints per interval

MergeSamFiles merges multiple SAM/BAM datasets into one

SFF converter

bedtools ClusterBed cluster overlapping/nearby intervals

Download and Extract Reads in FASTA/Q format from NCBI SRA

Wiggle-to-Interval converter

Faster Download and Extract Reads in FASTQ format from NCBI SRA

Merge Columns together

Merge Columns together

Cluster the intervals of a dataset

CollectInsertSizeMetrics plots distribution of insert sizes

ReplaceSamHeader replace header in a SAM/BAM dataset

WORKFLOWS

All workflows

History

search datasets

JSG_History

2 shown

236.08 MB

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)

Executed **Compute** and successfully added 1 job to the queue.

The tool uses this input:

- 1: UCSC Main on Human: knownGene (genome)

This tool produces this output:

- 2: Compute on data 1

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Sort numeric descending on c7 (using Sort data in ascending or descending order)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize Shared Data Admin Help User

Tools

search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Filter data on any column using simple expressions

Sort data in ascending or descending order

Select lines that match an expression

GFF

Extract features from GFF data

Filter GFF data by attribute using simple expressions

Filter GFF data by feature count using simple expressions

Filter GTF data by attribute values_list

Join, Subtract and Group

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

Analyze Data Workflow Visualize Shared Data Admin Help User

Sort data in ascending or descending order (Galaxy Version 1.1.0)

Sort Dataset

2: Compute on data 1

on column

Column: 7

with flavor

Numerical sort

everything in

Descending order

Column selection

+ Insert Column selection

Number of header lines to skip

0

characters are already considered as comments and kept

✓ Execute

1: Select Column 7

2: Select Fast Numeric

3: Select Descending Order

4: Click Execute

History

search datasets

JSG_History

2 shown

243.72 MB

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Sort numeric descending on c7 (using Sort data in ascending or descending order)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

Tools

Sort data in ascending or descending order

Samtools sort order of storing aligned sequences

Unique lines assuming sorted input file

bedtools SortBED order the intervals

Wig/BedGraph-to-bigWig converter

SortSam sort SAM/BAM dataset

SAM-to-BAM convert SAM to BAM

Aggregate datapoints Appends the average, min, max of datapoints per interval

MergeSamFiles merges multiple SAM/BAM datasets into one

SFF converter

bedtools ClusterBed cluster overlapping/nearby intervals

Download and Extract Reads in FASTA/Q format from NCBI SRA

Wiggle-to-Interval converter

Faster Download and Extract Reads in FASTQ format from NCBI SRA

Merge Columns together

Merge Columns together

Cluster the intervals of a dataset

CollectInsertSizeMetrics plots distribution of insert sizes

ReplaceSamHeader replace header in a SAM/BAM dataset

WORKFLOWS

All workflows

History

search datasets

JSG_History

3 shown

358.53 MB

3: Sort on data 2

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)

Done!



Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Select first 1000 lines (using Select first lines from a dataset)

HPRC Kaiser Galaxy Using 2%

Tools

search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Remove, rearrange and/or rename columns in text files

Advanced Grep

Cut columns from a table

Merge Columns together

Convert delimiters to TAB

Regex Find And Replace

Column Regex Find And Replace

Change Case of selected columns

unique_line remove duplicate lines

Compute an expression on every row

Multi-Join (combine multiple files)

Unique occurrences of each record

Unfold columns from a table

Sort data in ascending or descending order

tac reverse a file (reverse cat)

Select first lines from a dataset (head)

Text reformatting with awk

Text transformation with sed

Create text file with recurring lines

Replace Text in a specific column

Replace parts of text

Advanced Cut columns from a table

History

search datasets

JSG_History

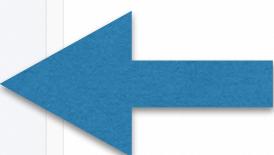
3 shown

358.53 MB

3: Sort on data 2

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)



The tool uses this input:

- 2: Compute on data 1

It produces this output:

- 3: Sort on data 2

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Select first 1000 lines (using Select first lines from a dataset)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

Tools

search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Remove, rearrange and/or rename columns in text files

Advanced Grep

Cut columns from a table

Merge Columns together

Convert delimiters to TAB

Regex Find And Replace

Column Regex Find And Replace

Change Case of selected columns

unique_line remove duplicate lines

Compute an expression on every row

Multi-Join (combine multiple files)

Unique occurrences of each record

Unfold columns from a table

Sort data in ascending or descending order

tac reverse a file (reverse cat)

Select first lines from a dataset (head)

Text reformatting with awk

Text transformation with sed

Create text file with recurring lines

Replace Text in a specific column

Replace parts of text

Advanced Cut columns from a table

File to select

3: Sort on data 2

Operation

Keep first lines

Number of lines

1000

These will be kept/discarded depending on 'operation'. (-)

✓ 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 +
```

will produce:

```
chr7 56632 56652 D17003_CTCF_R6 310 +
chr7 56736 56756 D17003_CTCF_R7 354 +
```

Citation

If you use this tool in Galaxy, please cite:

Bjoern A. Gruenig (2014), *Galaxy wrapper*

Assaf Gordon (gordon <at> cshl dot edu)

History

search datasets

JSG_History

3 shown

358.53 MB

3: Sort on data 2

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome
Search in textfiles (grep)

HPRC Kaiser Galaxy Using 2%

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

Tools

Search in textfiles (grep) X

Show Sections

Advanced Grep

Search in textfiles (grep)

tRNA prediction (tRNAscan)

FASTA Width formatter

NCBI BLAST+ blastp Search protein database with protein query sequence(s)

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s)

NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)

NCBI BLAST+ tblastn Search translated nucleotide database with protein query sequence(s)

NCBI BLAST+ tblastx Search translated nucleotide database with translated nucleotide query sequence(s)

Download and Extract Reads in BAM format from NCBI SRA

Download and Extract Reads in FASTA/Q format from NCBI SRA

Faster Download and Extract Reads in FASTQ format from NCBI SRA

tRNA and tmRNA prediction (Aragorn)

NCBI BLAST+ rpsblast Search protein domain database (PSSMs) with protein query sequence(s)

NCBI BLAST+ rpstblastn Search protein domain database (PSSMs) with translated nucleotide query sequence(s)

Create assemblies with Unicycler

KEGG pathway mapping and

Search in textfiles (grep) (Galaxy Version 1.1.1)

Select lines from

4: Select first on data 3

1: Search in textfiles (grep)

Type of regex

Perl

Regular Expression

See below for more details

Match type

case insensitive

(-i)

Show lines preceding the matched line

0

leave it at zero unless you know what you're doing. (-B)

Show lines trailing the matched line

0

leave it at zero unless you know what you're doing. (-A)

Output

text file (for further processing)

Execute

What it does

This tool runs the unix **grep** command on the selected data file.

TIP: This tool uses the **perl** regular expression syntax (same as running 'grep -P'). This is **NOT** the POSIX or POSIX-extended syntax (unlike the awk/sed tools).

Further reading

- Wikipedia's Regular Expression page (http://en.wikipedia.org/wiki/Regular_expression)
- Regular Expressions cheat-sheet (PDF) (<http://www.addedbytes.com/cheat-sheets/download/regular-expressions-cheat-sheet-v2.pdf>)
- Grep Tutorial (<http://www.panix.com/~elflord/unix/grep.html>)

History

search datasets X

JSG_History

4 shown

358.61 MB

4: Select first on data 3 X

3: Sort on data 2 X

2: Compute on data 1 X

1: UCSC Main on Human: knownGene (genome) X

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome
Search in textfiles (grep)

Using 2%

HPRC Kaiser Galaxy

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

Tools

Search in textfiles (grep) X

Show Sections

Advanced Grep

Search in textfiles (grep)

tRNA prediction (tRNAscan)

FASTA Width formatter

NCBI BLAST+ blastp Search protein database with protein query sequence(s)

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s)

NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)

NCBI BLAST+ tblastn Search translated nucleotide database with protein query sequence(s)

NCBI BLAST+ tblastx Search translated nucleotide database with translated nucleotide query sequence(s)

Download and Extract Reads in BAM format from NCBI SRA

Download and Extract Reads in FASTA/Q format from NCBI SRA

Faster Download and Extract Reads in FASTQ format from NCBI SRA

tRNA and tmRNA prediction (Aragorn)

NCBI BLAST+ rpsblast Search protein domain database (PSSMs) with protein query sequence(s)

NCBI BLAST+ rpstblastn Search protein domain database (PSSMs) with translated nucleotide query sequence(s)

Create assemblies with Unicycler

KEGG pathway mapping and

Search in textfiles (grep) (Galaxy Version 1.1.1)

4: Select first on data 3

Select lines from

Match

Type of regex

Perl

Regular Expression

See below for more details

1: Enter regex

^chr[\d]+\t|^chrx\t|^chry\t

Match type

case insensitive

(-i)

Show lines preceding the matched line

0

leave it at zero unless you know what you're doing. (-B)

Show lines trailing the matched line

0

leave it at zero unless you know what you're doing. (-A)

Output

text file (for further processing)

Execute

What it does

This tool runs the unix **grep** command on the selected data file.

TIP: This tool uses the **perl** regular expression syntax (same as running 'grep -P'). This is **NOT** the POSIX or POSIX-extended syntax (unlike the awk/sed tools).

Further reading

- Wikipedia's Regular Expression page (http://en.wikipedia.org/wiki/Regular_expression)
- Regular Expressions cheat-sheet (PDF) (<http://www.addedbytes.com/cheat-sheets/download/regular-expressions-cheat-sheet-v2.pdf>)
- Grep Tutorial (<http://www.panix.com/~elflord/unix/grep.html>)

History

search datasets X

JSG_History

4 shown

358.61 MB

4: Select first on data 3 X

3: Sort on data 2 X

2: Compute on data 1 X

1: UCSC Main on Human: knownGene (genome) X

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome Search in textfiles (grep)

HPRC Kaiser Galaxy Using 2%

Tools

Search in textfiles (grep)

Show Sections

Advanced Grep

Search in textfiles (grep)

tRNA prediction (tRNAscan)

FASTA Width formatter

NCBI BLAST+ blastp Search protein database with protein query sequence(s)

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s)

NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)

NCBI BLAST+ tblastn Search translated nucleotide database with protein query sequence(s)

NCBI BLAST+ tblastx Search translated nucleotide database with translated nucleotide query sequence(s)

Download and Extract Reads in BAM format from NCBI SRA

Download and Extract Reads in FASTA/Q format from NCBI SRA

Faster Download and Extract Reads in FASTQ format from NCBI SRA

tRNA and tmRNA prediction (Aragorn)

NCBI BLAST+ rpsblast Search protein domain database (PSSMs) with protein query sequence(s)

NCBI BLAST+ rpstblastn Search protein domain database (PSSMs) with translated nucleotide query sequence(s)

Create assemblies with Unicycler

KEGG pathway mapping and

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾  

History

search datasets

JSG_History

5 shown

358.68 MB

5: Search in textfiles on data 4

4: Select first on data 3

3: Sort on data 2

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Go to: 'Join_Subtract_and_Group' => Group

HPRC Kaiser Galaxy Using 2%

Tools

search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

Mapping

NCBI BLAST+

Assembly

Annotation

FASTA/FASTQ

Datamash

EMBOSS

Multiple Alignments

Picard

BED

SAM/BAM

Proteomics

Analyze Data Workflow Visualize Shared Data Admin Help User

History

search datasets

JSG_History

5 shown

358.68 MB

5: Search in textfiles on data 4

4: Select first on data 3

3: Sort on data 2

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)

Join_Subtract_and_Group



Computational Arithmetics

Join, Subtract and Group

About GROUPING

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
TTT	3

- Grouping on column 4 while not ignoring case, and performing operation Count on column 1 will return:

aaa	1
AAA	1
ttt	1
TTT	2

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Go to: 'Join_Subtract_and_Group' => Group

HPRC Kaiser Galaxy Using 2%

Tools Analyze Data Workflow Visualize Shared Data Admin Help User

History search datasets JSG_History 5 shown 358.68 MB

search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Join two Datasets side by side on a specified field

Compare two Datasets to find common or distinct rows

Group data by a column and perform aggregate operation on other columns.

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

Mapping

NCBI BLAST+

Assembly

Annotation

FASTA/FASTQ

Datamash

EMBOSS

Executed **Search in textfiles** and successfully added 1 job to the queue.

The tool uses this input:

- 4: Select first on data 3

It produces this output:

- 5: Search in textfiles on data 4

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Group



Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Go to: 'Join_Subtract_and_Group' => Group

HPRC Kaiser Galaxy Using 2%

Tools search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Join two Datasets side by side on a specified field

Compare two Datasets to find common or distinct rows

Group data by a column and perform aggregate operation on other columns.

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

Mapping

NCBI BLAST+

Assembly

Annotation

FASTA/FASTQ

Datamash

EMBOSS

Analyze Data Workflow Visualize Shared Data Admin Help User

Group data by a column and perform aggregate operation on other columns. (Galaxy Version 2.1.4)

5: Search in textfiles on data 4

Dataset missing? See TIP below.

Group by column

Column: 1

Ignore case while grouping?

Yes No

Ignore lines beginning with these characters

Select/Unselect all

> @ + < * - = | ? \$. : & % ^ #

lines beginning with these are not grouped

Operation

1: Operation

Type

Count

On column

Column: 1

Round result to nearest integer?

NO

Replace non numeric data

History

search datasets

JSG_History

5 shown

358.68 MB

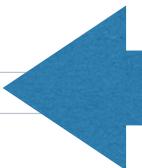
5: Search in textfiles on data 4

4: Select first on data 3

3: Sort on data 2

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)



Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Go to: 'Join_Subtract_and_Group' => Group

HPRC Kaiser Galaxy Using 2%

Tools

search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Join two Datasets side by side on a specified field

Compare two Datasets to find common or distinct rows

Group data by a column and perform aggregate operation on other columns.

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

Mapping

NCBI BLAST+

Assembly

Annotation

FASTA/FASTQ

Datamash

EMBOSS

Analyze Data Workflow Visualize Shared Data Admin Help User

Operation

1: Operation

Type: Count

On column: Column: 1

Round result to nearest integer?: NO

Replace non numeric data: leave empty for no replacements. Will replace, e.g., empty cells and text cells.

2: Operation

Type: Maximum

On column: Column: 7

Round result to nearest integer?: NO

Replace non numeric data: leave empty for no replacements. Will replace, e.g., empty cells and text cells.

3: Operation

Type: Minimum

On column: Column: 7

Round result to nearest integer?: NO

Replace non numeric data: leave empty for no replacements. Will replace, e.g., empty cells and text cells.

History

358.68 MB

Count by C01 (to determine the number of genes)

5: Search in textfiles on data 4

4: Select first on data 3

3: Sort on data 2

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Go to: 'Join_Subtract_and_Group' => Group

HPRC Kaiser Galaxy Using 2%

Tools

search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Join two Datasets side by side on a specified field

Compare two Datasets to find common or distinct rows

Group data by a column and perform aggregate operation on other columns.

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

Mapping

NCBI BLAST+

Assembly

Annotation

FASTA/FASTQ

Datamash

EMBOSS

Analyze Data Workflow Visualize Shared Data Admin Help User History

2: Operation

Type: Maximum

On column: Column: 7

Round result to nearest integer?: NO

Replace non numeric data: leave empty for no replacements. Will replace, e.g., empty cells and text cells.

3: Operation

Type: Minimum

On column: Column: 7

Round result to nearest integer?: NO

Replace non numeric data: leave empty for no replacements. Will replace, e.g., empty cells and text cells.

4: Operation

Type: Mean

On column: Column: 7

Round result to nearest integer?: NO

Replace non numeric data: leave empty for no replacements. Will replace, e.g., empty cells and text cells.

Maximum on C07 (contains lengths of exons)

History

5 shown

358.68 MB

5: Search in textfiles on data 4

4: Select first on data 3

3: Sort on data 2

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Go to: 'Join_Subtract_and_Group' => Group

Using 2%

HPRC Kaiser Galaxy

Analyze Data Workflow Visualize Shared Data Admin Help User

Tools

search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Join two Datasets side by side on a specified field

Compare two Datasets to find common or distinct rows

Group data by a column and perform aggregate operation on other columns.

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

Mapping

NCBI BLAST+

Assembly

Annotation

FASTA/FASTQ

Datamash

EMBOSS

3: Operation

Type: Minimum

On column: Column: 7

Round result to nearest integer?: NO

Replace non numeric data:

leave empty for no replacements. Will replace, e.g., empty cells and text cells.

4: Operation

Type: Mean

On column: Column: 7

Round result to nearest integer?: NO

Replace non numeric data:

leave empty for no replacements. Will replace, e.g., empty cells and text cells.

+ Insert Operation

✓ Execute

Minimum on C07 (contains lengths of exons)

History

search datasets

5 shown

358.68 MB

5: Search in textfiles on data 4

4: Select first on data 3

3: Sort on data 2

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)

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: Mean, Median, Mode, Sum, Max, Min, Count, Concatenate, and Randomly pick on any column(s).

The Concatenate function will take, for each group, each item in the specified column and build a comma delimited list. Concatenate Unique will do the same but will build a list of unique items with no repetition.

Count and Count Unique are equivalent to Concatenate and Concatenate Unique, but will only count the number of items and will return an integer.

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Go to: 'Join_Subtract_and_Group' => Group

HPRC Kaiser Galaxy Using 2%

Tools search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Join two Datasets side by side on a specified field

Compare two Datasets to find common or distinct rows

Group data by a column and perform aggregate operation on other columns.

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

Mapping

NCBI BLAST+

Assembly

Annotation

FASTA/FASTQ

Datamash

EMBOSS

Analyze Data Workflow Visualize Shared Data Admin Help User History search datasets

4: Operation

Type: Mean

On column: Column: 7

Round result to nearest integer?: NO

Replace non numeric data: leave empty for no replacements. Will replace, e.g., empty cells and text cells.

+ Insert Operation

✓ Execute

Mean on C07 (contains lengths of exons)

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: Mean, Median, Mode, Sum, Max, Min, Count, Concatenate, and Randomly pick on any column(s).

The Concatenate function will take, for each group, each item in the specified column and build a comma delimited list. Concatenate Unique will do the same but will build a list of unique items with no repetition.

Count and Count Unique are equivalent to Concatenate and Concatenate Unique, but will only count the number of items and will return an integer.

- 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
TTT 3
```

Computational Arithmetics

Join, Subtract and Group

Running a quick statistics on the ~ first 1000 Exons of the Human Genome

Go to: 'Join_Subtract_and_Group' => Group

HPRC Kaiser Galaxy Using 2%

Tools

search tools

HPRC

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Join two Datasets side by side on a specified field

Compare two Datasets to find common or distinct rows

Group data by a column and perform aggregate operation on other columns.

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

Mapping

NCBI BLAST+

Assembly

Annotation

FASTA/FASTQ

Datamash

EMBOSS

Analyze Data Workflow Visualize ▾ Shared Data ▾ Admin Help ▾ User ▾

	1	2	3	4	5
chr1	93	44880	8018	10841.7	
chr10	31	15220	8095	10270.8	
chr11	36	91667	8022	12354.4	
chr12	52	205012	7990	16011.7	
chr13	24	27561	7970	12402.2	
chr14	33	33290	8006	12653.9	
chr15	47	19780	7971	11216.4	
chr16	24	16058	8118	10320.8	
chr17	22	13836	7956	9781.41	
chr18	23	32994	8308	11027.2	
chr19	30	21693	7985	10468.8	
chr2	102	20552	8054	11648.7	
chr20	17	11579	8159	9208.53	
chr21	11	18112	8114	11756.4	
chr22	33	49287	8030	14466	
chr3	47	24927	8006	11795.4	
chr4	23	14165	8003	9796.83	
chr5	61	22753	7967	9508.16	
chr6	39	23264	8061	11022.9	
chr7	57	23777	8064	10611.7	
chr8	32	19865	8186	10867.5	
chr9	25	13761	8028	9434.04	
chrX	38	347300	7965	20242.6	

History

search datasets

JSG_History

6 shown

358.68 MB

Final Result

6: Group on data 5

5: Search in textfiles on data 4

4: Select first on data 3

3: Sort on data 2

2: Compute on data 1

1: UCSC Main on Human: knownGene (genome)