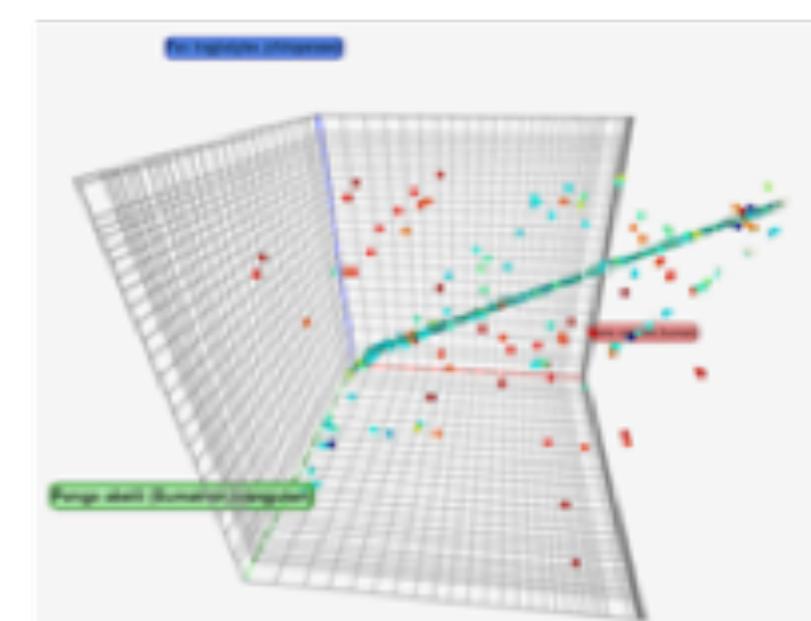
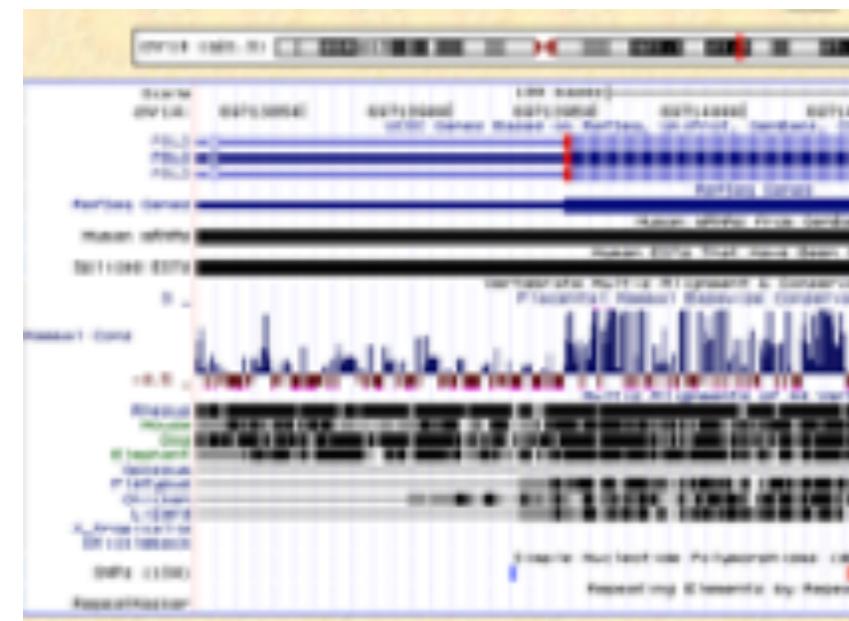


Computational Genomics

Working with Genome Files in Galaxy



Introduction to Regular Expressions

Go to [regexr.com](https://regextester.com)

The screenshot shows the RegExr interface. On the left is a sidebar with links: Menu, Pattern Settings, My Patterns, Cheatsheet, RegEx Reference, Community Patterns, and Help. The main area has tabs for Expression, Text, and Tests (which is selected). The expression entered is `/([A-Z])\w+/g`. The text area contains two paragraphs of explanatory text. A sidebar on the right lists various tools: Replace, List, Details, and Explain. Below the sidebar, there's a section for capturing groups with examples for character sets, ranges, word characters, and quantifiers.

RegExr is an online tool to **learn**, **build**, & **test** Regular Expressions (RegEx / RegExp).

- Supports **JavaScript & PHP/PCRE** RegEx.
- Results update in **real-time** as you type.
- Roll over** a match or expression for details.
- Validate patterns with suites of **Tests**.
- Save & share expressions with others.
- Use **Tools** to explore your results.
- Full **RegEx Reference** with help & examples.
- Undo & Redo with cmd-Z / Y in editors.
- Search for & rate **Community Patterns**.

Tools

Roll-over elements below to highlight in the Expression above. Click to open in Reference.

(Capturing group #1. Groups multiple tokens together and creates a capture group for extracting a substring or using a backreference.

[Character set. Match any character in the set.

A-Z Range. Matches a character in the range "A" to "Z" (char code 65 to 90). Case sensitive.

)

\w Word. Matches any word character (alphanumeric & underscore).

+ Quantifier. Match 1 or more of the preceding token.

Introduction to Regular Expressions

Regex: `^(chr.*)\t([0-9]+)\t+([0-9]+)\t+ENST([0-9]+)`

Expression PCRE ▾ Flags ▾

/`^(chr.*)\t([0-9]+)\t+([0-9]+)\t+ENST([0-9]+)`/gm

Text Tests 12 matches (0.3ms)

```
chr1→65564→65573→ENST00000641515.2_cds_1_0_chr1_65565_f→0→+→
chr1→69036→70008→ENST00000641515.2_cds_2_0_chr1_69037_f→0→+→
chrX→284187→284314→ENST00000429181.6_cds_2_0_chrX_284188_f→0→+→
chrX→288732→288787→ENST00000429181.6_cds_3_0_chrX_288733_f→0→+→
chrY→284187→284314→ENST00000429181.6_cds_2_0_chrY_284188_f→0→+→
chrY→288732→288787→ENST00000429181.6_cds_3_0_chrY_288733_f→0→+→
chr10→48054→48114→ENST00000562809.1_cds_2_0_chr10_48055_r→0→!-→
chr10→48614→48725→ENST00000562809.1_cds_1_0_chr10_48615_r→0→!-→
chr11→168957→169052→ENST00000410108.5_cds_4_0_chr11_168958_r→0→!-→
chr11→180208→180404→ENST00000410108.5_cds_3_0_chr11_180209_r→0→!-→
chr12→66882→67436→ENST00000538872.6_cds_0_0_chr12_66883_f→0→+→
chr12→99145→99214→ENST00000538872.6_cds_1_0_chr12_99146_f→0→+→
```

Introduction to GREP in Galaxy

GREP (Global Regular Expression Print) Where GREP Came From - Computerphile

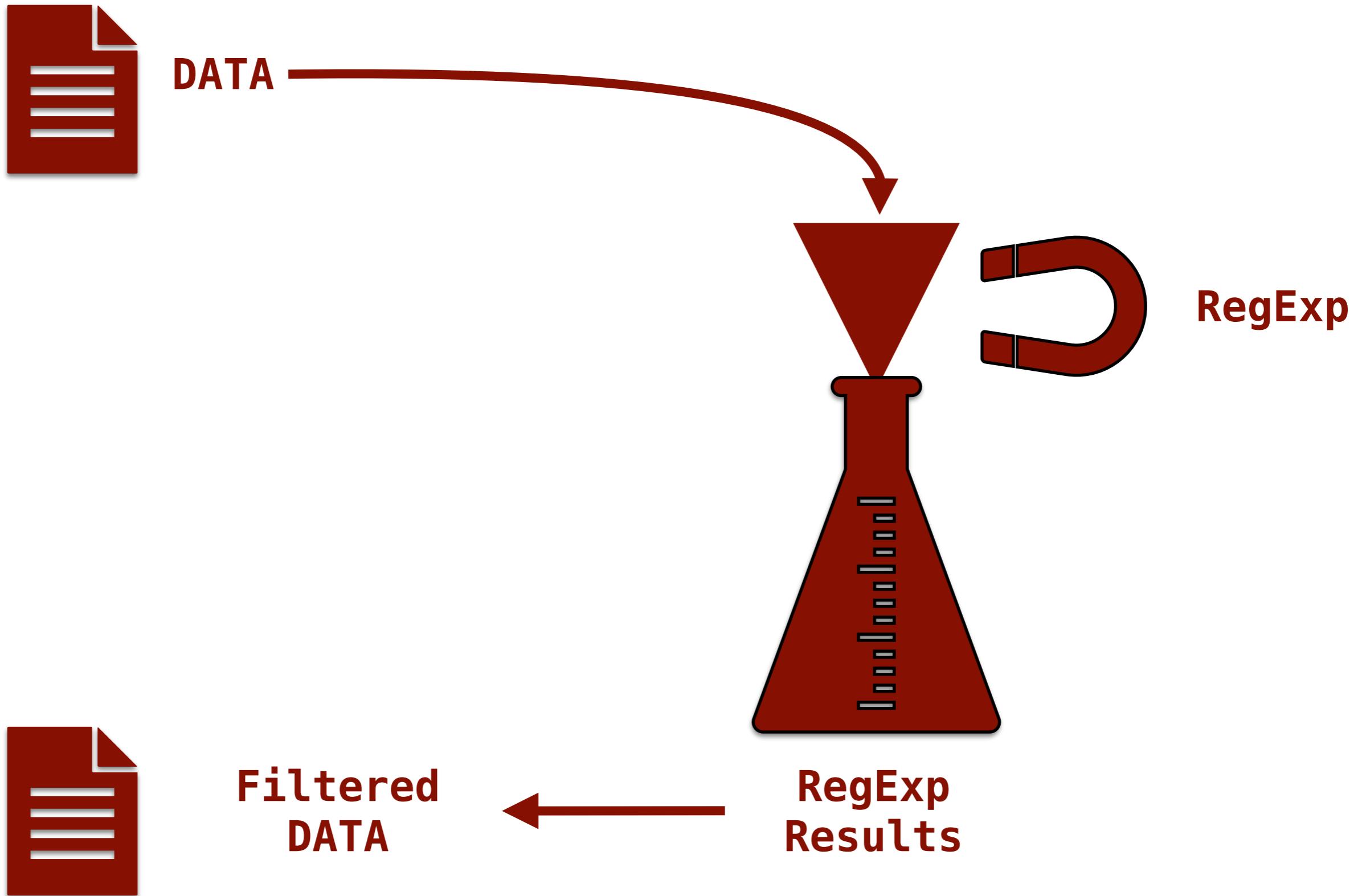
- **Regular Expressions (regex)**
 - Describes text and text patterns
 - Do not have to contain literal text
 - Comprised of metacharacters
 - Metacharacters are processed by ‘parsing’
- **Text Searching versus Grep Searching**
 - Text searching is literal, whereas GREP searching is abstract and conditional
 - Text is finite - GREP is flexible
 - Text looks for characters (what) - GREP looks for locations (where and what)
 - History of GREP
 - Also known as Regular Expression Parser
 - Original command-line text search utility
 - In sed (stream editor):

g/re/p or global/regular expression/print

- GREP remembers what it found and can be directed to re-use it
- GREP searches for patterns and most text can be described as a pattern

Introduction to GREP in Galaxy

GREP (Global Regular Expression Print)



Introduction to SED: Stream Editor

Regular SED

Typical Command: `sed 's/a/b/'`

s=substitution

a=search string

b=replacement string



DATA



SED



Transformed DATA



To Remove a String:

```
's/string//'  
's/;Parent=gene//'  
's/>//'
```

To Replace a String:

```
's/string/new_string/'
```

Working With Genome Files In Galaxy

BED Format

BED format

BED (Browser Extensible Data) format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

BED information should not be mixed as explained above (BED3 should not be mixed with BED4), rather additional column information must be filled for consistency, for example with a "." in some circumstances, if the field content is to be empty. BED fields in custom tracks can be whitespace-delimited or tab-delimited. Only some variations of BED types, such as [bedDetail](#), require a tab character delimitation for the detail columns.

Please note that only in custom tracks can the first lines of the file consist of header lines, which begin with the word "browser" or "track" to assist the browser in the display and interpretation of the lines of BED data following the headers. Such annotation track header lines are not permissible in downstream utilities such as bedToBigBed, which convert lines of BED text to indexed binary files.

If your data set is BED-like, but it is very large (over 50MB) and you would like to keep it on your own server, you should use the [bigBed](#) data format. Read a [blog post](#) for step-by-step instructions.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The **chromEnd** base is not included in the display of the feature, however, the number in [position format](#) will be represented. For example, the first 100 bases of chromosome 1 are defined as *chrom=1, chromStart=0, chromEnd=100*, and span the bases numbered 0-99 in our software (not 0-100), but will represent the position notation chr1:1-100. Read more [here](#).
chromStart and *chromEnd* can be identical, creating a feature of length 0, commonly used for insertions. For example, use *chromStart=0, chromEnd=0* to represent an insertion before the first nucleotide of a chromosome.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the score value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

shade									
score in range	≤ 166	167-277	278-388	389-499	500-611	612-722	723-833	834-944	≥ 945

6. **strand** - Defines the strand. Either "." (=no strand) or "+" or "-".
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RGB value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

In BED files with block definitions, the first *blockStart* value must be 0, so that the first block begins at *chromStart*. Similarly, the final *blockStart* position plus the final *blockSize* value must equal *chromEnd*. Blocks may not overlap.

```
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

Working With Genome Files In Galaxy

BED Format

BED (Browser Extensible Data) format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature, however, the number in **position format** will be represented. For example, the first 100 bases of chromosome 1 are defined as *chrom=1, chromStart=0, chromEnd=100*, and span the bases numbered 0-99 in our software (not 0-100), but will represent the position notation chr1:1-100. Read more [here](#).
chromStart and *chromEnd* can be identical, creating a feature of length 0, commonly used for insertions. For example, use *chromStart=0, chromEnd=0* to represent an insertion before the first nucleotide of a chromosome.

Working With Genome Files In Galaxy

GTF Format

GTF format

GTF (Gene Transfer Format, GTF2.2) is an extension to, and backward compatible with, GFF. The first eight GTF fields are the same as GFF. The *feature* field is the same as GFF, with the exception that it also includes the following optional values: *5UTR*, *3UTR*, *inter*, *inter_CNS*, and *intron_CNS*. The *group* field has been expanded into a list of *attributes*. Each attribute consists of a type/value pair. Attributes must end in a semi-colon, and be separated from any following attribute by exactly one space.

The attribute list must begin with the two mandatory attributes:

- **gene_id value** - A globally unique identifier for the genomic source of the sequence.
- **transcript_id value** - A globally unique identifier for the predicted transcript.

Example:

Here is an example of the ninth field in a GTF data line:

```
gene_id "Em:U62317.C22.6.mRNA"; transcript_id "Em:U62317.C22.6.mRNA"; exon_number 1
```

The Genome Browser groups together GTF lines that have the same *transcript_id* value. It only looks at features of type *exon* and *CDS*.

```
gene_id "Em:U62317.C22.6.mRNA"; transcript_id "Em:U62317.C22.6.mRNA"; exon_number 1
```

Working With Genome Files In Galaxy

GTF Format

GTF stands for Gene transfer format. It borrows from [GFF](#), but has additional structure that warrants a separate definition and format name.

Structure is as [GFF](#), so the fields are:

<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]

Here is a simple example with 3 translated exons. Order of rows is not important.

```
381 Twinscan CDS      380    401    .    +    0    gene_id "001"; transcript_id "001.1";
381 Twinscan CDS      501    650    .    +    2    gene_id "001"; transcript_id "001.1";
381 Twinscan CDS      700    707    .    +    2    gene_id "001"; transcript_id "001.1";
381 Twinscan start_codon 380    382    .    +    0    gene_id "001"; transcript_id "001.1";
381 Twinscan stop_codon 708    710    .    +    0    gene_id "001"; transcript_id "001.1";
```

The whitespace in this example is provided only for readability. In GTF, fields must be separated by a single TAB and no white space.

Working With Genome Files In Galaxy

GTF Format

Here is an example of a gene on the negative strand including UTR regions. Larger coordinates are 5' of smaller coordinates. Thus, the start codon is 3 bp with largest coordinates among all those bp that fall within the CDS regions. Note that the stop codon lies between the 3UTR and the CDS

```
140 Twinscan inter 5141 8522 . - . gene_id ""; transcript_id "";
140 Twinscan inter_CNS 8523 9711 . - . gene_id ""; transcript_id "";
140 Twinscan inter 9712 13182 . - . gene_id ""; transcript_id "";
140 Twinscan 3UTR 65149 65487 . - . gene_id "140.000"; transcript_id "140.000.1";
140 Twinscan 3UTR 66823 66992 . - . gene_id "140.000"; transcript_id "140.000.1";
140 Twinscan stop_codon 66993 66995 . - 0 gene_id "140.000"; transcript_id "140.000.1";
140 Twinscan CDS 66996 66999 . - 1 gene_id "140.000"; transcript_id "140.000.1";
140 Twinscan intron_CNS 70103 70151 . - . gene_id "140.000"; transcript_id "140.000.1";
140 Twinscan CDS 70207 70294 . - 2 gene_id "140.000"; transcript_id "140.000.1";
140 Twinscan CDS 71696 71807 . - 0 gene_id "140.000"; transcript_id "140.000.1";
140 Twinscan start_codon 71805 71806 . - 0 gene_id "140.000"; transcript_id "140.000.1";
140 Twinscan start_codon 73222 73222 . - 2 gene_id "140.000"; transcript_id "140.000.1";
140 Twinscan CDS 73222 73222 . - 0 gene_id "140.000"; transcript_id "140.000.1";
140 Twinscan 5UTR 73223 73504 . - . gene_id "140.000"; transcript_id "140.000.1";
```

Note the frames of the coding exons. For example:

1. The first CDS (from 71807 to 71696) always has frame zero.
2. Frame of the 1st CDS =0, length =112. $(3 - ((\text{length} - \text{frame}) \bmod 3)) \bmod 3 = 2$, the frame of the 2nd CDS.
3. Frame of the 2nd CDS=2, length=88. $(3 - ((\text{length} - \text{frame}) \bmod 3)) \bmod 3 = 1$, the frame of the terminal CDS.
4. Alternatively, the frame of terminal CDS can be calculated without the rest of the gene. Length of the terminal CDS=4. $\text{length} \bmod 3 = 1$, the frame of the terminal CDS.

Note the split start codon. The second start codon region has a frame of 2, since it is the second base, and has an accompanying CDS feature, since CDS always includes the start codon.

Working With Genome Files In Galaxy

GFF Format

GFF format

GFF (General Feature Format) lines are based on the Sanger [GFF2 specification](#). GFF lines have nine required fields that *must* be tab-separated. If the fields are separated by spaces instead of tabs, the track will not display correctly. For more information on GFF format, refer to Sanger's [GFF page](#).

Note that there is also a GFF3 specification that is not currently supported by the Browser. All GFF tracks must be formatted according to Sanger's GFF2 specification.

If you would like to obtain browser data in GFF (GTF) format, please refer to [Genes in gtf or gff format](#) on the Wiki.

Here is a brief description of the GFF fields:

1. **seqname** - The name of the sequence. Must be a chromosome or scaffold.
2. **source** - The program that generated this feature.
3. **feature** - The name of this type of feature. Some examples of standard feature types are "CDS" "start_codon" "stop_codon" and "exon".
4. **start** - The starting position of the feature in the sequence. The first base is numbered 1.
5. **end** - The ending position of the feature (inclusive).
6. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the score value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). If there is no score value, enter ":".
7. **strand** - Valid entries include "+", "-", or "." (for don't know/don't care).
8. **frame** - If the feature is a coding exon, *frame* should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be ".".
9. **group** - All lines with the same group are linked together into a single item.

Example:

Here's an example of a GFF-based track. This data format require tabs and some operating systems convert tabs to spaces. If pasting doesn't work, this [example's](#) contents or the url itself can be pasted into the custom track text box.

```
browser position chr22:10000000-10025000
browser hide all
track name=regulatory description="TeleGene(tm) Regulatory Regions" visibility=2
chr22 TeleGene enhancer 10000000 10001000 500 + . touch1
chr22 TeleGene promoter 10010000 10010100 900 + . touch1
chr22 TeleGene promoter 10020000 10025000 800 - . touch2
```

chr22	TeleGene	enhancer	10000000	10001000	500	+	.	touch1
chr22	TeleGene	promoter	10010000	10010100	900	+	.	touch1
chr22	TeleGene	promoter	10020000	10025000	800	-	.	touch2

Working With Genome Files In Galaxy

For Gene **ENSG00000139618**

Which file would you use:?

- To determine in which Chromosome this gene is located
- To extract all the transcripts IDs associated with gene
- To extract all the proteins IDs associated with gene
- To extract all the Exons IDs associated with gene
- To determine if the gene lives in either the Watson or the Crick strand

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

grep

Search in textfiles (grep)

obigrep Filters sequence file

FASTA Width formatter

Advanced Grep

newcpgreport Report CpG rich areas

cpgreport Reports all CpG rich regions

Text reformatting with awk

WORKFLOWS

All workflows

History

search datasets

L09-A

4 shown
1.97 GB

4: Homo_sapiens.GRCh3 8.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh3.8.109.gtf

1: Homo_sapiens.GRCh3.8.109.gff3

Search in textfiles (grep) (Galaxy Version 1.1.1)

Select lines from

4: Homo_sapiens.GRCh38.pep.all.fa
3: Hsapiens_BED
2: Homo_sapiens.GRCh38.109.gtf
1: Homo_sapiens.GRCh38.109.gff3

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

that

Match

Type of regex

Perl

Regular Expression

See below for more details

Match type

(-i)

Show lines preceding the matched line

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

Show lines trailing the matched line

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

Output

Job Resource Parameters

Memory (GB)

Maximum Job Memory

Time (hours)

Maximum job time

Execute

Working With Genome Files In Galaxy

For Gene ENSG00000139618

GFF3

Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
13	ensembl_havana	gene	32315086	32400268	.	+	.	ID=gene:ENSG00000139618;Name=BRCA2;biotype=protein_coding;description=BRCA2, breast cancer type 2 susceptibility protein
13	havana	mRNA	32315086	32400268	.	+	.	ID=transcript:ENST00000544455;Parent=gene:ENSG00000139618;Name=BRCA2 transcript
13	havana	mRNA	32315505	32333291	.	+	.	ID=transcript:ENST00000530893;Parent=gene:ENSG00000139618;Name=BRCA2 transcript
13	ensembl_havana	mRNA	32315508	32400268	.	+	.	ID=transcript:ENST00000380152;Parent=gene:ENSG00000139618;Name=BRCA2 transcript
13	havana	lnc_RNA	32315583	32316889	.	+	.	ID=transcript:ENST00000700199;Parent=gene:ENSG00000139618;Name=BRCA2 lncRNA
13	havana	lnc_RNA	32315583	32326971	.	+	.	ID=transcript:ENST00000700200;Parent=gene:ENSG00000139618;Name=BRCA2 lncRNA
13	havana	mRNA	32315585	32333290	.	+	.	ID=transcript:ENST00000700201;Parent=gene:ENSG00000139618;Name=BRCA2 transcript
13	havana	mRNA	32316072	32400268	.	+	.	ID=transcript:ENST00000680887;Parent=gene:ENSG00000139618;Name=BRCA2 transcript
13	havana	mRNA	32316461	32400268	.	+	.	ID=transcript:ENST00000614259;Parent=gene:ENSG00000139618;Name=BRCA2 transcript
13	havana	mRNA	32356428	32398233	.	+	.	ID=transcript:ENST00000665585;Parent=gene:ENSG00000139618;Name=BRCA2 transcript
13	havana	mRNA	32356526	32398770	.	+	.	ID=transcript:ENST00000700202;Parent=gene:ENSG00000139618;Name=BRCA2 transcript
13	havana	mRNA	32370971	32379495	.	+	.	ID=transcript:ENST00000528762;Parent=gene:ENSG00000139618;Name=BRCA2 transcript
13	havana	lnc_RNA	32375911	32398918	.	+	.	ID=transcript:ENST00000700203;Parent=gene:ENSG00000139618;Name=BRCA2 lncRNA
13	havana	mRNA	32379840	32398272	.	+	.	ID=transcript:ENST00000470094;Parent=gene:ENSG00000139618;Name=BRCA2 transcript
13	havana	mRNA	32380007	32394933	.	+	.	ID=transcript:ENST00000666593;Parent=gene:ENSG00000139618;Name=BRCA2 transcript

Working With Genome Files In Galaxy

For Gene ENSG00000139618

GTF

Working With Genome Files In Galaxy

For Gene ENSG00000139618

BED

Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGB	BlockCount	BlockSizes	BlockStarts
-------	-------	-----	------	-------	--------	------------	----------	---------	------------	------------	-------------

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

remove x

Upload Data

Show Sections

Remove beginning of a file

Remove, rearrange and/or rename columns in txt-converted FCS files

bedtools SubtractBed remove intervals based on overlaps

Remove sequencing artifacts

Cutadapt Remove adapter sequences from FASTQ/FASTA

unique_line remove duplicate lines

cutseq Removes a specified section from a sequence

degapseq Removes gap characters from sequences

noreturn Removes carriage return from ASCII files

Filter MAF blocks by Species

Filter empty datasets

Filter collection

Filter failed datasets

Trim leading or trailing characters

Filter MAF by specified attributes

RevertSam revert SAM/BAM datasets to a previous state

Subtract the intervals of two datasets

Join MAF blocks by Species

CD-HIT PROTEIN Cluster a protein dataset into representative sequences

CD-HIT-EST Cluster a nucleotide dataset into representative sequences

WORKFLOWS

All workflows

History

search datasets x

L09-A
7 shown
1.97 GB checkbox eye edit comment

7: Search in textfiles on data 3 eye edit x

6: Search in textfiles on data 2 eye edit x

5: Search in textfiles on data 1 eye edit x

4: Homo_sapiens.GRCh38.pep.all.fa eye edit x

3: Hsapiens_BED eye edit x

2: Homo_sapiens.GRCh38.109.gtf eye edit x

1: Homo_sapiens.GRCh38.109.gff3 eye edit x

Remove beginning of a file (Galaxy Version 1.0.0)

Remove first 1 lines from

7: Search in textfiles on data 3
6: Search in textfiles on data 2
5: Search in textfiles on data 1
4: Homo_sapiens.GRCh38.pep.all.fa
3: Hsapiens_BED
2: Homo_sapiens.GRCh38.109.gtf
1: Homo_sapiens.GRCh38.109.gff3

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Job Resource Parameters

Specify job resource parameters

Memory (GB) 7 Maximum Job Memory

Time (hours) 24 Maximum job time

✓ Execute

What it does

This tool removes a specified number of lines from the beginning of a dataset.

Example

Input File:

```
chr7 56632 56652 D17003_CTCF_R6 310 +
chr7 56736 56756 D17003_CTCF_R7 354 +
chr7 56761 56781 D17003_CTCF_R4 220 +
chr7 56772 56792 D17003_CTCF_R7 372 +
chr7 56775 56795 D17003_CTCF_R4 207 +
```

After removing the first 3 lines the dataset will look like this:

```
chr7 56772 56792 D17003_CTCF_R7 372 +
chr7 56775 56795 D17003_CTCF_R4 207 +
```

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

remove

Remove beginning of a file

Remove, rearrange and/or rename columns in txt-converted FCS files

bedtools SubtractBed remove intervals based on overlaps

Remove sequencing artifacts

Cutadapt Remove adapter sequences from FASTQ/FASTA

unique_line remove duplicate lines

cutseq Removes a specified section from a sequence

degapseq Removes gap characters from sequences

noreturn Removes carriage return from ASCII files

Filter MAF blocks by Species

Filter empty datasets

Filter collection

Filter failed datasets

Trim leading or trailing characters

Filter MAF by specified attributes

RevertSam revert SAM/BAM datasets to a previous state

Subtract the intervals of two datasets

Join MAF blocks by Species

CD-HIT PROTEIN Cluster a protein dataset into representative sequences

CD-HIT-EST Cluster a nucleotide dataset into representative sequences

WORKFLOWS

All workflows

Executed Remove beginning and successfully added 2 jobs to the queue.

The tool uses 2 inputs:

- 5: Search in textfiles on data 1
- 6: Search in textfiles on data 2

It produces 2 outputs:

- 8: Remove beginning on data 5
- 9: Remove beginning on data 6

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.

History	<input type="button" value="refresh"/> <input type="button" value="+"/> <input type="button" value="□"/> <input type="button" value="⚙"/>
<input type="button" value="search datasets"/> <input type="button" value="?"/> <input type="button" value="x"/>	
L09-A	
9 shown	<input type="checkbox"/> <input type="button" value="eye"/> <input type="button" value="edit"/> <input type="button" value="comment"/>
1.97 GB	
9: Remove beginning on data 6 <input type="button" value="eye"/> <input type="button" value="edit"/> <input type="button" value="x"/>	
8: Remove beginning on data 5 <input type="button" value="eye"/> <input type="button" value="edit"/> <input type="button" value="x"/>	
7: Search in textfiles on data 3 <input type="button" value="eye"/> <input type="button" value="edit"/> <input type="button" value="x"/>	
6: Search in textfiles on data 2 <input type="button" value="eye"/> <input type="button" value="edit"/> <input type="button" value="x"/>	
5: Search in textfiles on data 1 <input type="button" value="eye"/> <input type="button" value="edit"/> <input type="button" value="x"/>	
4: Homo_sapiens.GRCh3 8.pep.all.fa <input type="button" value="eye"/> <input type="button" value="edit"/> <input type="button" value="x"/>	
3: Hsapiens_BED <input type="button" value="eye"/> <input type="button" value="edit"/> <input type="button" value="x"/>	
2: Homo_sapiens.GRCh3 8.109.gtf <input type="button" value="eye"/> <input type="button" value="edit"/> <input type="button" value="x"/>	
1: Homo_sapiens.GRCh3 8.109.gff3 <input type="button" value="eye"/> <input type="button" value="edit"/> <input type="button" value="x"/>	

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

remove

Remove beginning of a file
Remove, rearrange and/or rename columns in txt-converted FCS files
bedtools SubtractBed remove intervals based on overlaps
Remove sequencing artifacts
Cutadapt Remove adapter sequences from FASTQ/FASTA
unique_line remove duplicate lines
cutseq Removes a specified section from a sequence
degapseq Removes gap characters from sequences
noreturn Removes carriage return from ASCII files
Filter MAF blocks by Species
Filter empty datasets
Filter collection
Filter failed datasets
Trim leading or trailing characters
Filter MAF by specified attributes
RevertSam revert SAM/BAM datasets to a previous state
Subtract the intervals of two datasets
Join MAF blocks by Species
CD-HIT PROTEIN Cluster a protein dataset into representative sequences
CD-HIT-EST Cluster a nucleotide dataset into representative sequences
WORKFLOWS
All workflows

Edit Dataset Attributes

Attributes Convert Datatypes Permissions

Name: ENSG00000139618.gff3

Info:

Annotation:

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:

unspecified (?)

Number of comment lines:

History

search datasets

L09-A

9 shown
1.97 GB

9: Remove beginning on data 6

8: Remove beginning on data 5

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh3 8.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh3 8.109.gtf

1: Homo_sapiens.GRCh3 8.109.gff3

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

remove

Remove beginning of a file
Remove, rearrange and/or rename columns in txt-converted FCS files
bedtools SubtractBed remove intervals based on overlaps
Remove sequencing artifacts
Cutadapt Remove adapter sequences from FASTQ/FASTA
unique_line remove duplicate lines
cutseq Removes a specified section from a sequence
degapseq Removes gap characters from sequences
noreturn Removes carriage return from ASCII files
Filter MAF blocks by Species
Filter empty datasets
Filter collection
Filter failed datasets
Trim leading or trailing characters
Filter MAF by specified attributes
RevertSam revert SAM/BAM datasets to a previous state
Subtract the intervals of two datasets
Join MAF blocks by Species
CD-HIT PROTEIN Cluster a protein dataset into representative sequences
CD-HIT-EST Cluster a nucleotide dataset into representative sequences
WORKFLOWS
All workflows

Edit Dataset Attributes

Attributes Convert Datatypes Permissions

Name: ENSG00000139618.gtf

Info:

Annotation:

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:

unspecified (?)

Number of comment lines:

History

search datasets

L09-A

9 shown
1.97 GB

9: Remove beginning on data 6

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh3 8.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh3 8.109.gtf

1: Homo_sapiens.GRCh3 8.109.gff3

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

remove

Remove beginning of a file
Remove, rearrange and/or rename columns in txt-converted FCS files
bedtools SubtractBed remove intervals based on overlaps
Remove sequencing artifacts
Cutadapt Remove adapter sequences from FASTQ/FASTA
unique_line remove duplicate lines
cutseq Removes a specified section from a sequence
degapseq Removes gap characters from sequences
noreturn Removes carriage return from ASCII files
Filter MAF blocks by Species
Filter empty datasets
Filter collection
Filter failed datasets
Trim leading or trailing characters
Filter MAF by specified attributes
RevertSam revert SAM/BAM datasets to a previous state
Subtract the intervals of two datasets
Join MAF blocks by Species
CD-HIT PROTEIN Cluster a protein dataset into representative sequences
CD-HIT-EST Cluster a nucleotide dataset into representative sequences
WORKFLOWS
All workflows

Edit Dataset Attributes

Attributes updated.

Name
ENSG00000139618.gtf

Info

Annotation

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build
unspecified (?)

Number of comment lines

History

search datasets

L09-A

9 shown
1.97 GB

9: ENSG00000139618.gtf

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

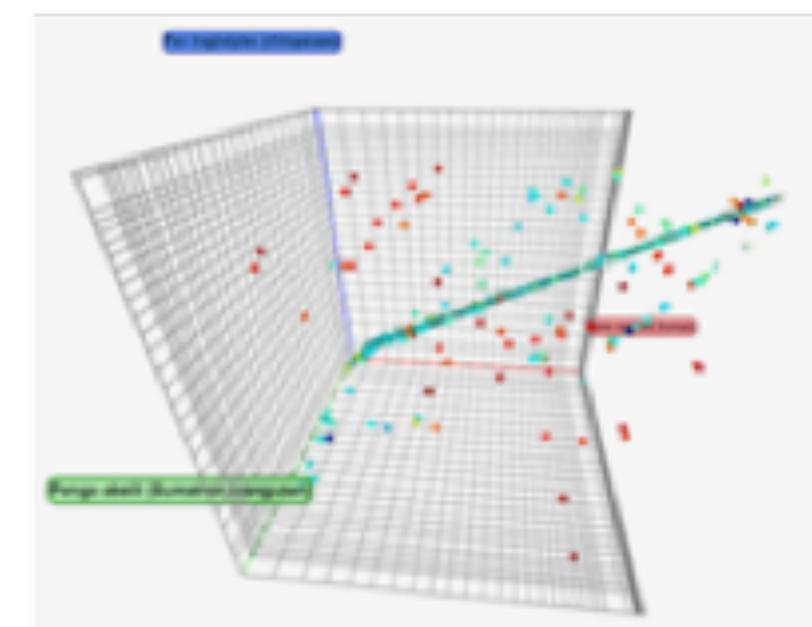
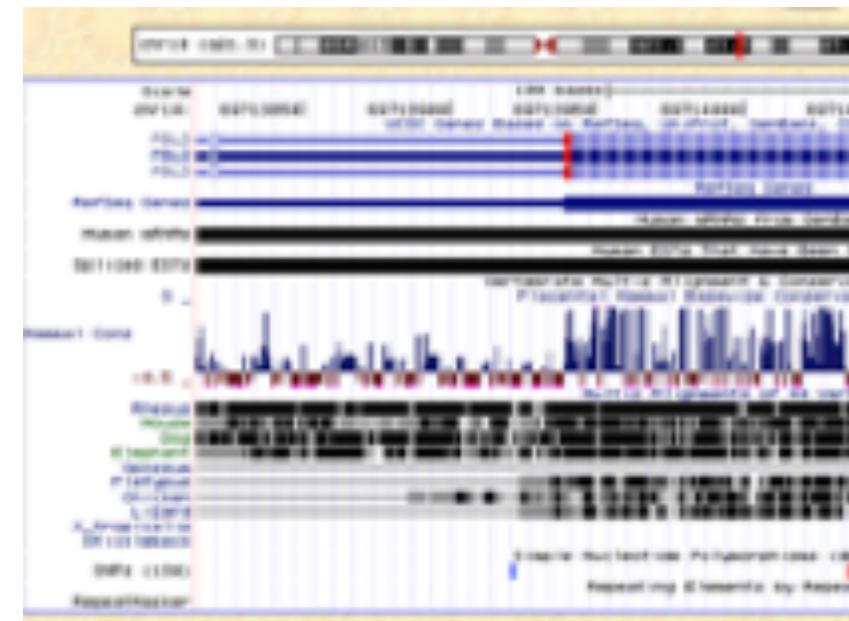
2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Computational Genomics

Introduction to Genome Browsers

IGV



Integrative Genome Viewer

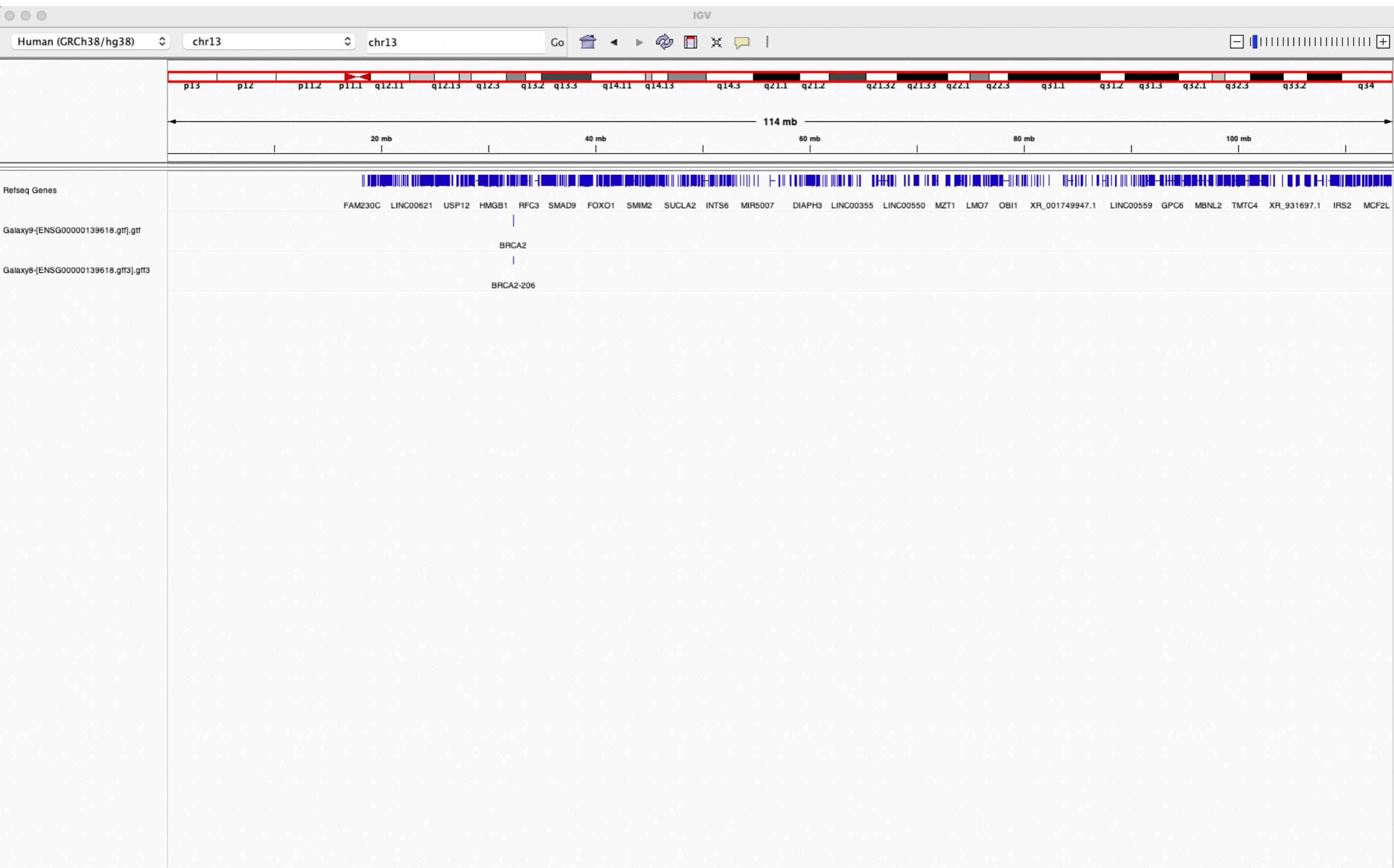
Integrative Genomics Viewer



[Download IGV](#)

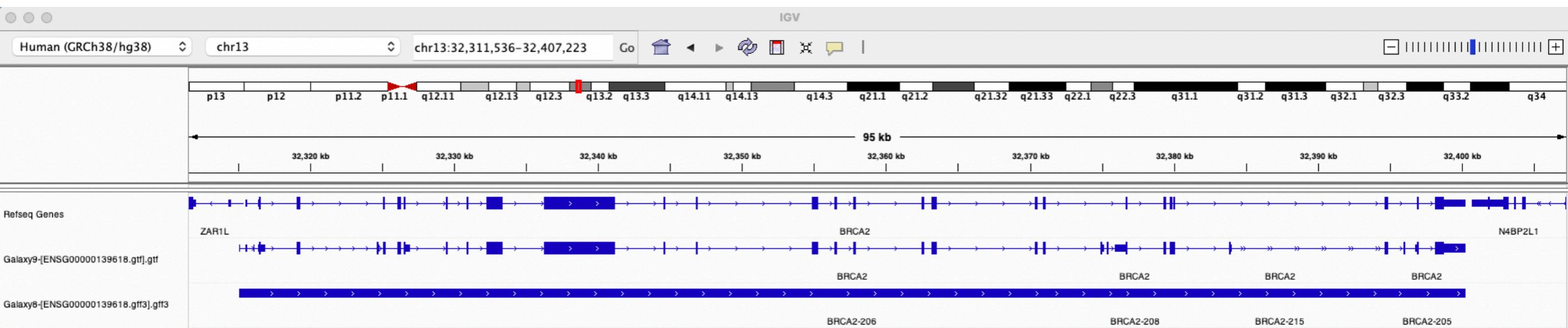
Working With Genome Files In Galaxy

For Gene ENSG00000139618



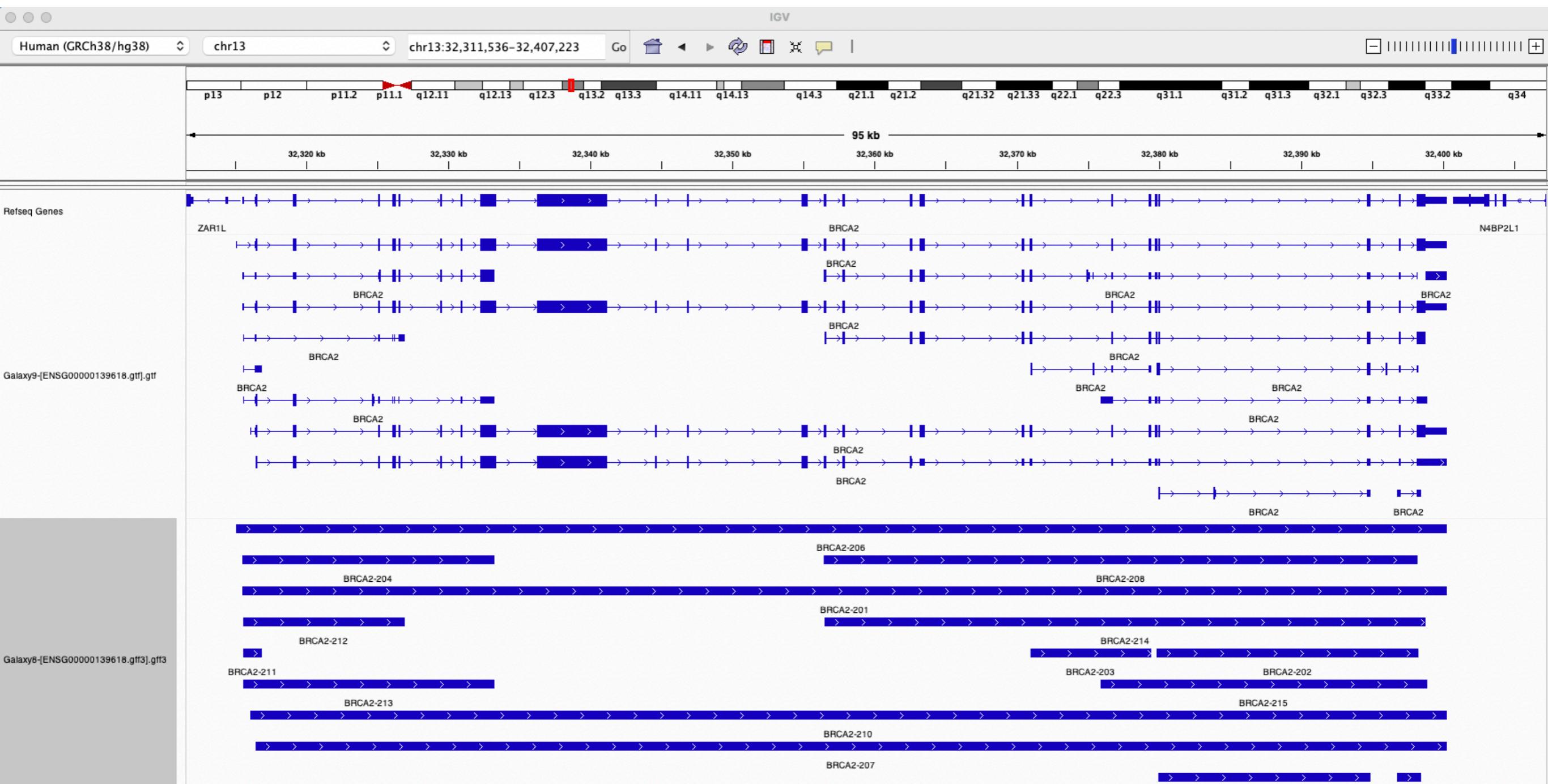
Working With Genome Files In Galaxy

For Gene ENSG00000139618



Working With Genome Files In Galaxy

For Gene ENSG00000139618



Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

- cut
- Upload Data
- Show Sections

Cut columns from a table

Advanced Cut columns from a table (cut)

Cutadapt Remove adapter sequences from FASTQ/FASTA

Filter by quality

Trimmomatic flexible read trimming tool for Illumina NGS data

Trim sequences

cutseq Removes a specified section from a sequence

WORKFLOWS

- All workflows

Cut columns from a table (Galaxy Version 1.0.2)

Cut columns
c9

Delimited by
Tab

From
8: ENSG00000139618.gff3

Job Resource Parameters
Specify job resource parameters

Memory (GB)
7

Maximum Job Memory

Time (hours)
24

Maximum job time

Execute

WARNING: This tool breaks column assignments. To re-establish column assignments run the tools and click on the pencil icon in the latest history item.

i The output of this tool is always in tabular format (e.g., if your original delimiters are commas, they will be replaced with tabs). For example:
Cutting columns 1 and 3 from:

apple,is,good
windows,is,bad

will give:

apple good
windows bad

What it does

This tool selects (cuts out) specified columns from the dataset.

- Columns are specified as **c1**, **c2**, and so on. Column count begins with **1**
- Columns can be specified in any order (e.g., **c2,c1,c6**)
- If you specify more columns than actually present - empty spaces will be filled with dots

Example

Input dataset (six columns: c1, c2, c3, c4, c5, and c6):

```
chr1 10 1000 gene1 0 +
```

History

search datasets

L09-A

9 shown

1.97 GB

9: ENSG00000139618.gtf

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Fetching Transcripts IDs to GREP BED File

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

convert x

Upload Data

Show Sections

Convert delimiters to TAB

Convert delimiters to TAB

NCBI BLAST+ convert2blastmask

Convert masking information in lower-case masked FASTA input to file formats suitable for makeblastdb

Convert genome coordinates between assemblies and genomes

Convert SAM to interval

SFF converter

Quality format converter (ASCII-Numeric)

MAF to FASTA Converts a MAF formatted file to FASTA format

BED-to-bigBed converter

Tabular-to-FASTA converts tabular file to FASTA format

Wig/BedGraph-to-bigWig converter

MAF to Interval Converts a MAF formatted file to the Interval format

GFF-to-BED converter

BED-to-GFF converter

MAF to BED Converts a MAF formatted file to the BED format

FASTA-to-Tabular converter

FASTQ to FASTA converter from FASTX-toolkit

BAM-to-SAM convert BAM to SAM

SAM-to-BAM convert SAM to BAM

WORKFLOWS

All workflows

Convert delimiters to TAB (Galaxy Version 1.0.1)

Convert all

Colons

in Query

10: Cut on data 8

Job Resource Parameters

Specify job resource parameters

Memory (GB)

Maximum Job Memory

Time (hours)

Maximum job time

Execute

What it does

Converts all delimiters of a specified type into TABs. Consecutive characters are condensed. For example, if columns are separated by 5 spaces they will be converted into 1 tab.

Example

- Input file:

```
chrX|151283558|151283724|NM_000808_exon_8_0_chrX_151283559_r|0|-  
chrX|151370273|151370486|NM_000808_exon_9_0_chrX_151370274_r|0|-  
chrX|151559494|151559583|NM_018558_exon_1_0_chrX_151559495_f|0|+  
chrX|151564643|151564711|NM_018558_exon_2_0_chrX_151564644_f|||0|+
```
- Converting all pipe delimiters of the above file to TABs will get:

```
chrX 151283558 151283724 NM_000808_exon_8_0_chrX_151283559_r 0 -  
chrX 151370273 151370486 NM_000808_exon_9_0_chrX_151370274_r 0 -  
chrX 151559494 151559583 NM_018558_exon_1_0_chrX_151559495_f 0 +  
chrX 151564643 151564711 NM_018558_exon_2_0_chrX_151564644_f 0 +
```

Requirements: ?

- python (Version 3.8)

History

search datasets ? x

L09-A

10 shown

1.97 GB checkbox eye edit comment

10: Cut on data 8

9: ENSG00000139618.gtf

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Fetching Transcripts IDs to GREP BED File

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

search tools

Upload Data

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

BED

Annotation

Multiple Alignments

NCBI BLAST+

Mapping

SAM/BAM

Assembly

FASTQ Quality Control

FASTA/FASTQ

RNA-seq

CD-HIT

Datamash

EMBOSS

MUMmer4

Nanopore

ID	Type	ENST ID	Parent gene	Name	Biotype	CCDS ID	Tag	Transcript ID	Transcript Support Level
ID=transcript		ENST00000544455	Parent=gene	ENSG00000139618;Name=BRCA2-206;biotype=protein_coding;ccdsid=CCDS9344.1;tag=basic;transcript_id=ENST00000544455;version=1					
ID=transcript		ENST00000530893	Parent=gene	ENSG00000139618;Name=BRCA2-204;biotype=protein_coding;transcript_id=ENST00000530893;transcript_support_level=1;version=1					
ID=transcript		ENST00000380152	Parent=gene	ENSG00000139618;Name=BRCA2-201;biotype=protein_coding;ccdsid=CCDS9344.1;tag=basic,Ensembl_canonical,MANE_Select;transcript_id=ENST00000380152;transcript_support_level=1;version=1					
ID=transcript		ENST00000700199	Parent=gene	ENSG00000139618;Name=BRCA2-211;biotype=retained_intron;transcript_id=ENST00000700199;version=1					
ID=transcript		ENST00000700200	Parent=gene	ENSG00000139618;Name=BRCA2-212;biotype=retained_intron;transcript_id=ENST00000700200;version=1					
ID=transcript		ENST00000700201	Parent=gene	ENSG00000139618;Name=BRCA2-213;biotype=nonsense-mediated_decay;transcript_id=ENST00000700201;version=1					
ID=transcript		ENST00000680887	Parent=gene	ENSG00000139618;Name=BRCA2-210;biotype=protein_coding;ccdsid=CCDS9344.1;transcript_id=ENST00000680887;version=1					
ID=transcript		ENST00000614259	Parent=gene	ENSG00000139618;Name=BRCA2-207;biotype=nonsense-mediated_decay;transcript_id=ENST00000614259;transcript_support_level=1					
ID=transcript		ENST00000665585	Parent=gene	ENSG00000139618;Name=BRCA2-208;biotype=nonsense-mediated_decay;transcript_id=ENST00000665585;version=1					
ID=transcript		ENST00000700202	Parent=gene	ENSG00000139618;Name=BRCA2-214;biotype=protein_coding;transcript_id=ENST00000700202;version=1					
ID=transcript		ENST00000528762	Parent=gene	ENSG00000139618;Name=BRCA2-203;biotype=nonsense-mediated_decay;transcript_id=ENST00000528762;transcript_support_level=1					
ID=transcript		ENST00000700203	Parent=gene	ENSG00000139618;Name=BRCA2-215;biotype=retained_intron;transcript_id=ENST00000700203;version=1					
ID=transcript		ENST00000470094	Parent=gene	ENSG00000139618;Name=BRCA2-202;biotype=nonsense-mediated_decay;transcript_id=ENST00000470094;transcript_support_level=1					
ID=transcript		ENST00000666593	Parent=gene	ENSG00000139618;Name=BRCA2-209;biotype=nonsense-mediated_decay;transcript_id=ENST00000666593;version=1					
ID=transcript		ENST00000533776	Parent=gene	ENSG00000139618;Name=BRCA2-205;biotype=retained_intron;transcript_id=ENST00000533776;transcript_support_level=3;version=1					

History

search datasets

L09-A

11 shown

1.97 GB

11: Convert on data 10

10: Cut on data 8

9: ENSG00000139618.gtf

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh3 8.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh3 8.109.gtf

1: Homo_sapiens.GRCh3 8.109.gff3

Fetching Transcripts IDs to GREP BED File

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

search tools

Upload Data

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

BED

Annotation

Multiple Alignments

NCBI BLAST+

Mapping

SAM/BAM

Assembly

FASTQ Quality Control

FASTA/FASTQ

RNA-seq

CD-HIT

Datamash

EMBOSS

MUMmer4

Nanopore

ID=transcript ENST00000544455;Parent=gene ENSG00000139618;Name=BRCA2-206;biotype=protein_coding;ccdsid=CCDS9344.1;tag=basic;transcript_id=ENST00000544455

ID=transcript ENST00000530893;Parent=gene ENSG00000139618;Name=BRCA2-204;biotype=protein_coding;transcript_id=ENST00000530893;transcript_support_level=1

ID=transcript ENST00000380152;Parent=gene ENSG00000139618;Name=BRCA2-201;biotype=protein_coding;ccdsid=CCDS9344.1;tag=basic,Ensembl_canonical,Modestly_supported

ID=transcript ENST00000700199;Parent=gene ENSG00000139618;Name=BRCA2-211;biotype=retained_intron;transcript_id=ENST00000700199;version=1

ID=transcript ENST00000700200;Parent=gene ENSG00000139618;Name=BRCA2-212;biotype=retained_intron;transcript_id=ENST00000700200;version=1

ID=transcript ENST00000700201;Parent=gene ENSG00000139618;Name=BRCA2-213;biotype=nonsense-mediated_decay;transcript_id=ENST00000700201;version=1

ID=transcript ENST00000680887;Parent=gene ENSG00000139618;Name=BRCA2-210;biotype=protein_coding;ccdsid=CCDS9344.1;transcript_id=ENST00000680887

ID=transcript ENST00000614259;Parent=gene ENSG00000139618;Name=BRCA2-207;biotype=nonsense-mediated_decay;transcript_id=ENST00000614259;version=1

ID=transcript ENST00000665585;Parent=gene ENSG00000139618;Name=BRCA2-208;biotype=nonsense-mediated_decay;transcript_id=ENST00000665585;version=1

ID=transcript ENST00000700202;Parent=gene ENSG00000139618;Name=BRCA2-214;biotype=protein_coding;transcript_id=ENST00000700202;version=1

ID=transcript ENST00000528762;Parent=gene ENSG00000139618;Name=BRCA2-203;biotype=nonsense-mediated_decay;transcript_id=ENST00000528762;transcript_support_level=1

ID=transcript ENST00000700203;Parent=gene ENSG00000139618;Name=BRCA2-215;biotype=retained_intron;transcript_id=ENST00000700203;version=1

ID=transcript ENST00000470094;Parent=gene ENSG00000139618;Name=BRCA2-202;biotype=nonsense-mediated_decay;transcript_id=ENST00000470094;transcript_support_level=1

ID=transcript ENST00000666593;Parent=gene ENSG00000139618;Name=BRCA2-209;biotype=nonsense-mediated_decay;transcript_id=ENST00000666593;version=1

ID=transcript ENST00000533776;Parent=gene ENSG00000139618;Name=BRCA2-205;biotype=retained_intron;transcript_id=ENST00000533776;transcript_support_level=1

History

search datasets

L09-A

11 shown

1.97 GB

11: Convert on data 10

10: Cut on data 8

9: ENSG00000139618.gtf

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Fetching Transcripts IDs to GREP BED File

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

sed

Upload Data

Show Sections

Text transformation with sed

Replace Text in entire line

Search in textfiles (grep)

CollectBaseDistributionByCycle

charts the nucleotide distribution per cycle in a SAM or BAM dataset

biosed Replace or delete sequence sections

WORKFLOWS

All workflows

Text transformation with sed (Galaxy Version 1.1.1)

File to process

11: Convert on data 10

SED Program

```
s;/Parent=gene//
```

Advanced Options

Hide Advanced Options

Job Resource Parameters

Specify job resource parameters

Memory (GB)

7

Maximum Job Memory

Time (hours)

24

Maximum job time

Execute

What it does

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

TIP: This tool uses the **extended regular expression** syntax (same as running 'sed -r').

Further reading

- Short sed tutorial (http://www.linuxhowtos.org/System/sed_tutorial.htm)
- Long sed tutorial (<http://www.grymoire.com/Unix/Sed.html>)
- sed faq with good examples (<http://sed.sourceforge.net/sedfaq.html>)
- sed cheat-sheet (<http://www.catonmat.net/download/sed.stream.editor.cheat.sheet.pdf>)

Sed commands

The most useful sed command is **s** (substitute).

Examples

- s/hsa//** will remove the first instance of 'hsa' in every line.
- s/hsa/g** will remove all instances (because c)
- s/A{4,}--&--/g** will find sequences of 4 or more place holder for 'whatever matched the regular expression'

History

search datasets

L09-A

11 shown

1.97 GB

11: Convert on data 10

10: Cut on data 8

9: ENSG00000139618.gtf

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

search tools

Upload Data

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

BED

Annotation

Multiple Alignments

NCBI BLAST+

Mapping

SAM/BAM

Assembly

FASTQ Quality Control

FASTA/FASTQ

RNA-seq

CD-HIT

Datamash

EMBOSS

MUMmer4

Nanopore

ID=transcript ENST00000544455 ENSG00000139618;Name=BRCA2-206;biotype=protein_coding;ccdsid=CCDS9344.1;tag=basic;transcript_id=ENST00000544455;transcript_support_level=1;version=1

ID=transcript ENST00000530893 ENSG00000139618;Name=BRCA2-204;biotype=protein_coding;transcript_id=ENST00000530893;transcript_support_level=1;version=1

ID=transcript ENST00000380152 ENSG00000139618;Name=BRCA2-201;biotype=protein_coding;ccdsid=CCDS9344.1;tag=basic,Ensembl_canonical,MANE_Select;transcript_id=ENST00000380152;transcript_support_level=1;version=1

ID=transcript ENST00000700199 ENSG00000139618;Name=BRCA2-211;biotype=retained_intron;transcript_id=ENST00000700199;version=1

ID=transcript ENST00000700200 ENSG00000139618;Name=BRCA2-212;biotype=retained_intron;transcript_id=ENST00000700200;version=1

ID=transcript ENST00000700201 ENSG00000139618;Name=BRCA2-213;biotype=nonsense-mediated_decay;transcript_id=ENST00000700201;version=1

ID=transcript ENST00000680887 ENSG00000139618;Name=BRCA2-210;biotype=protein_coding;ccdsid=CCDS9344.1;transcript_id=ENST00000680887;version=1

ID=transcript ENST00000614259 ENSG00000139618;Name=BRCA2-207;biotype=nonsense-mediated_decay;transcript_id=ENST00000614259;transcript_support_level=1

ID=transcript ENST00000665585 ENSG00000139618;Name=BRCA2-208;biotype=nonsense-mediated_decay;transcript_id=ENST00000665585;version=1

ID=transcript ENST00000700202 ENSG00000139618;Name=BRCA2-214;biotype=protein_coding;transcript_id=ENST00000700202;version=1

ID=transcript ENST00000528762 ENSG00000139618;Name=BRCA2-203;biotype=nonsense-mediated_decay;transcript_id=ENST00000528762;transcript_support_level=1

ID=transcript ENST00000700203 ENSG00000139618;Name=BRCA2-215;biotype=retained_intron;transcript_id=ENST00000700203;version=1

ID=transcript ENST00000470094 ENSG00000139618;Name=BRCA2-202;biotype=nonsense-mediated_decay;transcript_id=ENST00000470094;transcript_support_level=1

ID=transcript ENST00000666593 ENSG00000139618;Name=BRCA2-209;biotype=nonsense-mediated_decay;transcript_id=ENST00000666593;version=1

ID=transcript ENST00000533776 ENSG00000139618;Name=BRCA2-205;biotype=retained_intron;transcript_id=ENST00000533776;transcript_support_level=3;version=1

History

search datasets

L09-A

12 shown

1.97 GB

12: Text transformation on data 11

11: Convert on data 10

10: Cut on data 8

9: ENSG00000139618.gtf

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Fetching Transcripts IDs to GREP BED File

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

cut x

Upload Data

Show Sections

Cut columns from a table

Advanced Cut columns from a table (cut)

Cutadapt Remove adapter sequences from FASTQ/FASTA

Filter by quality

Trimmomatic flexible read trimming tool for Illumina NGS data

Trim sequences

cutseq Removes a specified section from a sequence

WORKFLOWS

All workflows

Cut columns from a table (Galaxy Version 1.0.2)

Cut columns
c2

Delimited by
Tab

From
12: Text transformation on data 11

Job Resource Parameters
Specify job resource parameters

Memory (GB)
7

Maximum Job Memory

Time (hours)
24

Maximum job time

Execute

WARNING: This tool breaks column assignments. To re-establish column assignments run the tools and click on the pencil icon in the latest history item.

The output of this tool is always in tabular format (e.g., if your original delimiters are commas, they will be replaced with tabs). For example:

Cutting columns 1 and 3 from:

apple,is,good
windows,is,bad

will give:

apple good
windows bad

What it does

This tool selects (cuts out) specified columns from the dataset.

- Columns are specified as c1, c2, and so on. Column count begins with 1
- Columns can be specified in any order (e.g., c2,c1,c6)
- If you specify more columns than actually present - empty spaces will be filled with dots

Example

Input dataset (six columns: c1, c2, c3, c4, c5, and c6)
chr1 10 1000 gene1 0 +

History

search datasets x

L09-A

12 shown
1.97 GB

12: Text transformation on data 11

11: Convert on data 10

10: Cut on data 8

9: ENSG00000139618.gtf

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Fetching Transcripts IDs to GREP BED File

Working With Genome Files In Galaxy

For Gene ENSG00000139618

The screenshot shows the Galaxy web interface with the following components:

- Tools Panel:** On the left, a sidebar titled "Tools" contains a search bar "search tools" and a button "Upload Data". Below are categorized links:
 - 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
 - BED
 - Annotation
 - Multiple Alignments
 - NCBI BLAST+
 - Mapping
 - SAM/BAM
 - Assembly
 - FASTQ Quality Control
 - FASTA/FASTQ
 - RNA-seq
 - CD-HIT
 - Datamash
 - EMBOSS
 - MUMmer4
 - Nanopore
- History Panel:** On the right, a "History" section titled "L09-A" shows 13 items with details:
 - 13 shown
 - 1.97 GB
 - 13: Cut on data 12
 - 12: Text transformation on data 11
 - 11: Convert on data 10
 - 10: Cut on data 8
 - 9: ENSG00000139618.gtf
 - 8: ENSG00000139618.gff3
 - 7: Search in textfiles on data 3
 - 6: Search in textfiles on data 2
 - 5: Search in textfiles on data 1
 - 4: Homo_sapiens.GRCh38.pep.all.fa
 - 3: Hsapiens_BED
 - 2: Homo_sapiens.GRCh38.109.gtf
 - 1: Homo_sapiens.GRCh38.109.gff3
- Content Area:** The central area displays a list of ENST IDs:

ENST ID
ENST00000544455
ENST00000530893
ENST00000380152
ENST00000700199
ENST00000700200
ENST00000700201
ENST00000680887
ENST00000614259
ENST00000665585
ENST00000700202
ENST00000528762
ENST00000700203
ENST00000470094
ENST00000666593
ENST00000533776

Fetching Transcripts IDs to GREP BED File

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

grep x

Search in textfiles (grep)

obigrep Filters sequence file

FASTA Width formatter

Advanced Grep

newcpgreport Report CpG rich areas

cpgreport Reports all CpG rich regions

Text reformatting with awk

WORKFLOWS

All workflows

Advanced Grep (Galaxy Version 0.0.1)

Input file: 3: Hsapiens_BED

text/tabular/fasta/sam/... file to extract the matches from

Fetch Extra Lines: No

Also fetch lines following the matched pattern (eg for fastq extraction based on readname)

Pattern Source: Pattern File

Pattern File: 13: Cut on data 12

A text file with one pattern per line

Matching type: Exact string matching

Job Resource Parameters: Specify job resource parameters

Memory (GB): 7

Maximum Job Memory

Time (hours): 24

Maximum job time

What it does

This tool extends the grepping options available in galaxy. It allows extracting lines following a match, grepping from a file of patterns, and perl-based matching.

Requirements: ?

- perl_module_threads (Version 1.92)
- perl_module_threads_shared (Version 1.46)
- perl_module_Thread_Queue (Version 3.02)

History

search datasets x

L09-A

13 shown
1.97 GB

13: Cut on data 12

12: Text transformation on data 11

11: Convert on data 10

10: Cut on data 8

9: ENSG00000139618.gtf

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Fetching Transcripts IDs to GREP BED File

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

search tools x

Upload Data

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

BED

Annotation

Multiple Alignments

NCBI BLAST+

Mapping

SAM/BAM

Assembly

FASTQ Quality Control

FASTA/FASTQ

RNA-seq

CD-HIT

Datamash

EMBOSS

MUMmer4

Nanopore

Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGB	BlockCount	BlockSizes
chr13	32315085	32400268	ENST00000544455.6	0	+	32316460	32398770	0,0,0	27	60,106,249,109,50,41,115,50,112
chr13	32315504	32333291	ENST00000530893.6	0	+	32325128	32333291	0,0,0	10	163,102,249,109,50,41,115,50,11:
chr13	32315507	32400268	ENST00000380152.8	0	+	32316460	32398770	0,0,0	27	160,106,249,109,50,41,115,50,11:
chr13	32315582	32316889	ENST00000700199.1	0	+	32315582	32315582	0,0,0	2	85,468,
chr13	32315582	32326971	ENST00000700200.1	0	+	32315582	32315582	0,0,0	6	85,106,105,50,41,473,
chr13	32315584	32333290	ENST00000700201.1	0	+	32316460	32324731	0,0,0	11	83,106,249,169,109,50,41,115,50
chr13	32316071	32400268	ENST00000680887.1	0	+	32316460	32398770	0,0,0	27	86,106,249,109,50,41,115,50,112
chr13	32316460	32400268	ENST00000614259.2	0	+	32316460	32362659	0,0,0	26	67,249,109,50,41,115,50,112,1116
chr13	32356427	32398233	ENST00000665585.1	0	+	32356427	32375080	0,0,0	15	182,188,171,355,156,145,249,64
chr13	32356525	32398770	ENST00000700202.1	0	+	32356525	32398770	0,0,0	13	84,188,171,355,156,145,122,199,
chr13	32370970	32379495	ENST00000528762.1	0	+	32370970	32376670	0,0,0	4	130,64,122,179,
chr13	32375910	32398918	ENST00000700203.1	0	+	32375910	32375910	0,0,0	7	881,199,164,139,245,147,757,
chr13	32379839	32398272	ENST00000470094.1	0	+	32379839	32396092	0,0,0	6	74,139,245,126,147,111,
chr13	32380006	32394933	ENST00000666593.1	0	+	32380006	32383930	0,0,0	3	139,139,245,
chr13	32396808	32398448	ENST00000533776.1	0	+	32396808	32396808	0,0,0	2	236,287,

History

search datasets x

L09-A
14 shown
1.97 GB

14: Advanced Grep on data 13 and data 3 x

13: Cut on data 12 x

12: Text transformation on data 11 x

11: Convert on data 10 x

10: Cut on data 8 x

9: ENSG00000139618.gtf x

8: ENSG00000139618.gff3 x

7: Search in textfiles on data 3 x

6: Search in textfiles on data 2 x

5: Search in textfiles on data 1 x

4: Homo_sapiens.GRCh38.pep.all.fa x

3: Hsapiens_BED x

2: Homo_sapiens.GRCh38.109.gtf x

1: Homo_sapiens.GRCh38.109.gff3 x

Fetching Transcripts IDs to GREP BED File

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

search tools x

Upload Data

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

BED

Annotation

Multiple Alignments

NCBI BLAST+

Mapping

SAM/BAM

Assembly

FASTQ Quality Control

FASTA/FASTQ

RNA-seq

CD-HIT

Datamash

EMBOSS

MUMmer4

Nanopore

Edit Dataset Attributes

Name: ENSG00000139618.bed

Info: Running time:00:00:01

Annotation: Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build: Human Dec. 2013 (GRCh38/hg38) (hg38)

Number of comment lines:

Chrom column: 1

Start column: 2

End column: 3

Strand column (click box & select):

Name/Identifier column (click box & select):

Score column for visualization: Select/Unselect all x 4

Save Auto-detect

History

search datasets x

L09-A

14 shown
1.97 GB ✓ 🔗 💬

14: Advanced Grep on data 13 and data 3 eye edit x

13: Cut on data 12 eye edit x

12: Text transformation on data 11 eye edit x

11: Convert on data 10 eye edit x

10: Cut on data 8 eye edit x

9: ENSG00000139618.gtf eye edit x

8: ENSG00000139618.gff3 eye edit x

7: Search in textfiles on data 3 eye edit x

6: Search in textfiles on data 2 eye edit x

5: Search in textfiles on data 1 eye edit x

4: Homo_sapiens.GRCh38.pep.all.fa eye edit x

3: Hsapiens_BED eye edit x

2: Homo_sapiens.GRCh38.109.gtf eye edit x

1: Homo_sapiens.GRCh38.109.gff3 eye edit x

Fetching Transcripts IDs to GREP BED File

Working With Genome Files In Galaxy

For Gene ENSG00000139618

Tools

search tools x

Upload Data +

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

BED

Annotation

Multiple Alignments

NCBI BLAST+

Mapping

SAM/BAM

Assembly

FASTQ Quality Control

FASTA/FASTQ

RNA-seq

CD-HIT

Datamash

EMBOSS

MUMmer4

Nanopore

Edit Dataset Attributes

Attributes updated.

Attributes Convert Datatypes Permissions

Name: ENSG00000139618.bed

Info: Running time:00:00:01

Annotation: Add an annotation or notes

Database/Build: Human Dec. 2013 (GRCh38)

Number of comment lines: 0

Chrom column: 1

Start column: 2

End column: 3

Strand column (click box & select):

Name/Identifier column (click box & select):

Score column for visualization: Select/Unselect all x 4

Save As: Galaxy14-[ENSG00000139618.bed] Tags:

Downloads Cancel Save

History

search datasets ? x

L09-A

14 shown
1.97 GB edit comment more

14: ENSG00000139618.bed

15 regions
format: bed, database: hg38
Running time:00:00:01

display in IGB View
display with IGV local Human hg38
display at UCSC main test

1. Chrom 2. Start 3. End 4 5 6 7

chr13	32315085	32400268	ENST00000544455.6	0 +	32316460
chr13	32315504	32333291	ENST00000530893.6	0 +	32325128
chr13	32315507	32400268	ENST00000380152.8	0 +	32316460
chr13	32315582	32316889	ENST00000700199.1	0 +	32315582
chr13	32315582	32326971	ENST00000700200.1	0 +	32315582

13: Cut on data 12

12: Text transformation on data 11

11: Convert on data 10

10: Cut on data 8

9: ENSG00000139618.gtf

8: ENSG00000139618.gff3

7: Search in textfiles on data 3

6: Search in textfiles on data 2

5: Search in textfiles on data 1

4: Homo_sapiens.GRCh38.pep.all.fa

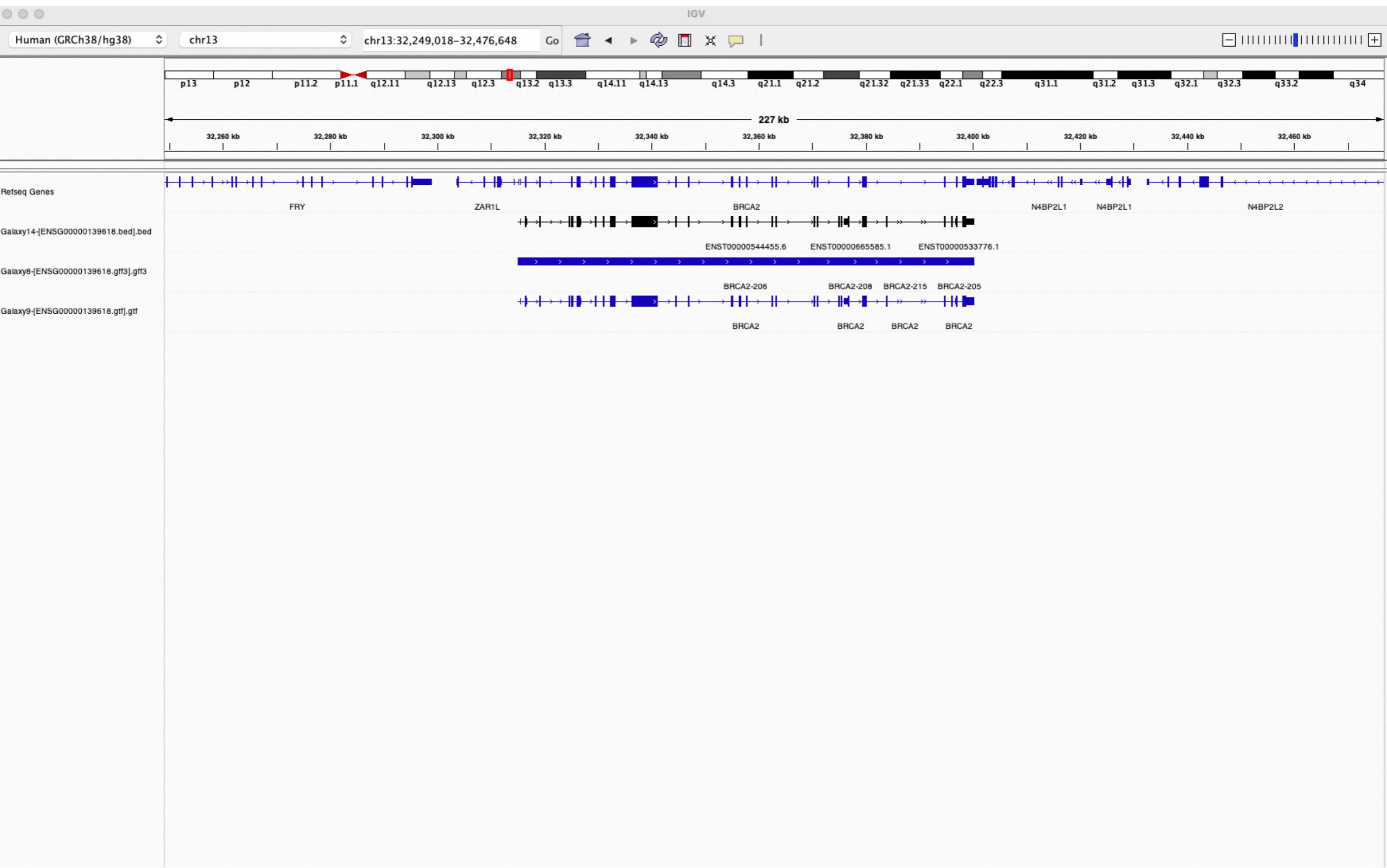
3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Working With Genome Files In Galaxy

For Gene ENSG00000139618



Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

Tools

search tools x

Upload Data

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

BED

Annotation

Multiple Alignments

NCBI BLAST+

Mapping

SAM/BAM

Assembly

FASTQ Quality Control

FASTA/FASTQ

RNA-seq

CD-HIT

Datamash

EMBOSS

MUMmer4

Nanopore

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.
- 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 7GB memory for 24 hours (24 SUs). Configuring a job to use 48 cores for 1 hour requires 48 SUs. (48 cores for 168 hours = 7872 SUs).
 - Configuring a job to use 360GB memory for 1 hour requires 48 SUs. (360GB memory for 168 hours = 7872 SUs).
 - If a tool you used failed because it needs the Job Resource Parameters option added, [contact](#) the HPRC helpdesk.



Take an interactive tour: [Galaxy UI](#) [History](#) [Window Manager](#) [Deferred Datasets](#)

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

History

search datasets ? x

L09-B

4 shown

1.97 GB checkbox trash comment

4: Homo_sapiens.GRCh38.pep.all.fa	eye edit x
3: Hsapiens_BED	eye edit x
2: Homo_sapiens.GRCh38.109.gtf	eye edit x
1: Homo_sapiens.GRCh38.109.gff3	eye edit x

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

Tools

grep x

Upload Data

Show Sections

Search in textfiles (grep)

obigrep Filters sequence file

FASTA Width formatter

Advanced Grep

newcpgreport Report CpG rich areas

cpgreport Reports all CpG rich regions

Text reformatting with awk

WORKFLOWS

All workflows

History

search datasets

L09-B

4 shown

1.97 GB

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Search in textfiles (grep) (Galaxy Version 1.1.1)

Select lines from
4: Homo_sapiens.GRCh38.pep.all.fa

that

Match

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)

Job Resource Parameters

Specify job resource parameters

Memory (GB)
7

Maximum Job Memory

Time (hours)
24

Maximum job time

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

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

The screenshot shows the Galaxy web interface with a tool panel on the left, a main workspace in the center, and a history panel on the right.

Tools Panel:

- search tools
- Upload Data
- 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
- BED
- Annotation
- Multiple Alignments
- NCBI BLAST+
- Mapping
- SAM/BAM
- Assembly
- FASTQ Quality Control
- FASTA/FASTQ
- RNA-seq
- CD-HIT
- Datamash
- EMBOSS
- MUMmer4
- Nanopore

Main Workspace (Central Area):

This dataset is large and only the first megabyte is shown below.
Show all | Save

```
>ENSP00000488240.1 pep chromosome:GRCh38:CHR_HSCHR7_2_CTG6:142847306:142847317:1 gene:ENSG00000282253.1 transcript:ENST00000631435.1 gene_biotype:TR_
>ENSP00000451042.1 pep chromosome:GRCh38:14:22438547:22438554:1 gene:ENSG00000223997.1 transcript:ENST00000415118.1 gene_biotype:TR_D_gene transcript
>ENSP00000452494.1 pep chromosome:GRCh38:14:22449113:22449125:1 gene:ENSG00000228985.1 transcript:ENST00000448914.1 gene_biotype:TR_D_gene transcript
>ENSP00000451515.1 pep chromosome:GRCh38:14:22439007:22439015:1 gene:ENSG00000237235.2 transcript:ENST00000434970.2 gene_biotype:TR_D_gene transcript
>ENSP00000487941.1 pep chromosome:GRCh38:7:142786213:142786224:1 gene:ENSG00000282431.1 transcript:ENST00000632684.1 gene_biotype:TR_D_gene transcript
>ENSP00000488695.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105866322:105866332:-1 gene:ENSG00000282455.1 transcript:ENST00000632524.1 gene_biotype:I
>ENSP00000488000.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105881805:105881824:-1 gene:ENSG00000282323.1 transcript:ENST00000633009.1 gene_biotype:I
>ENSP00000488392.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105882310:105882327:-1 gene:ENSG00000282724.1 transcript:ENST00000634070.1 gene_biotype:I
>ENSP00000488113.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105884674:105884693:-1 gene:ENSG00000282674.1 transcript:ENST00000632963.1 gene_biotype:I
>ENSP00000488168.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105885641:105885659:-1 gene:ENSG00000282640.1 transcript:ENST00000633030.1 gene_biotype:I
>ENSP00000488711.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105886802:105886832:-1 gene:ENSG00000282396.1 transcript:ENST00000633765.1 gene_biotype:I
>ENSP00000487599.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105889322:105889349:-1 gene:ENSG00000281984.1 transcript:ENST00000632619.1 gene_biotype:I
>ENSP00000487812.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105891962:105891978:-1 gene:ENSG00000282592.1 transcript:ENST00000632968.1 gene_biotype:I
>ENSP00000487789.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105892470:105892490:-1 gene:ENSG00000282487.1 transcript:ENST00000633159.1 gene_biotype:I
>ENSP00000488201.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105894313:105894332:-1 gene:ENSG00000282346.1 transcript:ENST00000631871.1 gene_biotype:I
>ENSP00000488261.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105895279:105895294:-1 gene:ENSG00000282274.1 transcript:ENST00000633010.1 gene_biotype:I
>ENSP00000487787.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105896405:105896441:-1 gene:ENSG00000282232.1 transcript:ENST00000633379.1 gene_biotype:I
>ENSP00000487993.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105898728:105898758:-1 gene:ENSG00000282818.1 transcript:ENST00000632473.1 gene_biotype:I
>ENSP00000488522.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105901409:105901425:-1 gene:ENSG00000282736.1 transcript:ENST00000631884.1 gene_biotype:I
>ENSP00000488592.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105901913:105901933:-1 gene:ENSG00000282042.1 transcript:ENST00000632859.1 gene_biotype:I
>ENSP00000487922.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105903420:105903442:-1 gene:ENSG00000282102.1 transcript:ENST00000631895.1 gene_biotype:I
>ENSP00000488735.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105904387:105904402:-1 gene:ENSG00000281940.1 transcript:ENST00000634154.1 gene_biotype:I
>ENSP00000488475.2 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105905268:105905298:-1 gene:ENSG00000282373.1 transcript:ENST00000632609.1 gene_biotype:I
>ENSP00000487775.2 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105905452:105905482:-1 gene:ENSG00000281939.1 transcript:ENST00000632911.1 gene_biotype:I
>ENSP00000488083.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105907982:105908012:-1 gene:ENSG00000282132.1 transcript:ENST00000633504.1 gene_biotype:I
>ENSP00000488720.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105910678:105910694:-1 gene:ENSG00000282495.1 transcript:ENST00000632304.1 gene_biotype:I
>ENSP00000488589.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105911181:105911198:-1 gene:ENSG00000282010.1 transcript:ENST00000632542.1 gene_biotype:I
>ENSP00000487937.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105913028:105913047:-1 gene:ENSG00000282769.1 transcript:ENST00000633968.1 gene_biotype:I
>ENSP00000488889.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105913993:105914008:-1 gene:ENSG00000282227.1 transcript:ENST00000634085.1 gene_biotype:I
>ENSP00000487903.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105915130:105915160:-1 gene:ENSG00000282754.1 transcript:ENST00000633353.1 gene_biotype:I
>ENSP00000487604.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105917597:105917627:-1 gene:ENSG00000282578.1 transcript:ENST00000631803.1 gene_biotype:I
>ENSP00000488840.1 pep chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105920273:105920289:-1 gene:ENSG00000282714.1 transcript:ENST00000633210.1 gene_biotype:I
>ENSP00000473787.1 pep chromosome:GRCh38:15:20011153:20011169:-1 gene:ENSG00000271336.1 transcript:ENST00000605284.1 gene_biotype:IG_D_gene transcript
>ENSP00000473849.1 pep chromosome:GRCh38:15:20003840:20003862:-1 gene:ENSG00000270961.1 transcript:ENST00000604642.1 gene_biotype:IG_D_gene transcript
>ENSP00000474065.2 pep chromosome:GRCh38:15:20008402:20008432:-1 gene:ENSG00000282599.1 transcript:ENST00000603077.1 gene_biotype:IG_D_gene transcript
>ENSP00000473700.1 pep chromosome:GRCh38:15:21010494:21010516:-1 gene:ENSG00000270824.1 transcript:ENST00000604446.1 gene_biotype:IG_D_gene transcript
>ENSP00000474017.2 pep chromosome:GRCh38:15:21015048:21015078:-1 gene:ENSG00000282268.1 transcript:ENST00000604102.1 gene_biotype:IG_D_gene transcript
>ENSP00000475053.2 pep chromosome:GRCh38:15:21011451:21011469:-1 gene:ENSG00000270451.1 transcript:ENST00000603693.1 gene_biotype:IG_D_gene transcript
>ENSP00000474133.2 pep chromosome:GRCh38:15:20005905:20005935:-1 gene:ENSG00000282520.1 transcript:ENST00000604950.1 gene_biotype:IG_D_gene transcript
>ENSP00000474222.1 pep chromosome:GRCh38:15:21017800:21017816:-1 gene:ENSG00000270185.1 transcript:ENST00000604838.1 gene_biotype:IG_D_gene transcript
>ENSP00000474573.2 pep chromosome:GRCh38:15:21012559:21012589:-1 gene:ENSG00000282089.1 transcript:ENST00000603935.1 gene_biotype:IG_D_gene transcript
>ENSP00000474693.2 pep chromosome:GRCh38:15:20004797:20004815:-1 gene:ENSG00000271317.1 transcript:ENST00000603326.1 gene_biotype:IG_D_gene transcript
>ENSP00000418639.1 pep chromosome:GRCh38:14:105865551:105865561:-1 gene:ENSG00000236597.1 transcript:ENST00000439842.1 gene_biotype:IG_D_gene transcript
>ENSP00000420733.1 pep chromosome:GRCh38:14:105881034:105881053:-1 gene:ENSG00000211907.1 transcript:ENST00000390567.1 gene_biotype:IG_D_gene transcript
>ENSP00000417751.1 pep chromosome:GRCh38:14:105881539:105881556:-1 gene:ENSG00000225825.1 transcript:ENST00000452198.1 gene_biotype:IG_D_gene transcript
>ENSP00000419139.1 pep chromosome:GRCh38:14:105883903:105883922:-1 gene:ENSG00000211909.1 transcript:ENST00000390569.1 gene_biotype:IG_D_gene transcript
>ENSP00000430248.1 pep chromosome:GRCh38:14:105884870:105884888:-1 gene:ENSG00000227196.1 transcript:ENST00000437320.1 gene_biotype:IG_D_gene transcript
>ENSP00000429952.1 pep chromosome:GRCh38:14:105886031:105886061:-1 gene:ENSG00000211911.1 transcript:ENST00000390571.1 gene_biotype:IG_D_gene transcript
>ENSP00000429324.1 pep chromosome:GRCh38:14:105888551:105888578:-1 gene:ENSG00000211912.1 transcript:ENST00000390572.1 gene_biotype:IG_D_gene transcript
>ENSP00000420556.1 pep chromosome:GRCh38:14:105891191:105891207:-1 gene:ENSG00000237020.1 transcript:ENST00000450276.1 gene_biotype:IG_D_gene transcript
>ENSP00000418010.1 pep chromosome:GRCh38:14:105891699:105891719:-1 gene:ENSG00000211914.1 transcript:ENST00000390574.1 gene_biotype:IG_D_gene transcript
>ENSP00000417555.1 pep chromosome:GRCh38:14:105893542:105893561:-1 gene:ENSG00000211915.1 transcript:ENST00000390575.1 gene_biotype:IG_D_gene transcript
>ENSP00000431089.1 pep chromosome:GRCh38:14:105894508:105894523:-1 gene:ENSG00000227800.1 transcript:ENST00000431870.1 gene_biotype:IG_D_gene transcript
>ENSP00000428366.1 pep chromosome:GRCh38:14:105895634:105895670:-1 gene:ENSG00000211917.1 transcript:ENST00000390577.1 gene_biotype:IG_D_gene transcript
>ENSP00000427969.1 pep chromosome:GRCh38:14:105897957:105897987:-1 gene:ENSG00000211918.1 transcript:ENST00000390578.1 gene_biotype:IG_D_gene transcript
>ENSP00000419265.1 pep chromosome:GRCh38:14:105899628:105899651:-1 gene:ENSG00000227108.1 transcript:ENST00000390579.1 gene_biotype:IG_D_gene transcript
```

History Panel:

5 shown
2 GB

5: Search in textfiles on data 4
4: Homo_sapiens.GRCh38.pep.all.fa
3: Hsapiens_BED
2: Homo_sapiens.GRCh38.109.gtf
1: Homo_sapiens.GRCh38.109.gff3

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

Tools

convert x

Upload Data

Show Sections

Convert delimiters to TAB

Convert delimiters to TAB

NCBI BLAST+ convert2blastmask
Convert masking information in lower-case masked FASTA input to file formats suitable for makeblastdb

Convert genome coordinates between assemblies and genomes

Convert SAM to interval

SFF converter

Quality format converter (ASCII-Numeric)

MAF to FASTA Converts a MAF formatted file to FASTA format

BED-to-bigBed converter

Tabular-to-FASTA converts tabular file to FASTA format

Wig/BedGraph-to-bigWig converter

MAF to Interval Converts a MAF formatted file to the Interval format

GFF-to-BED converter

BED-to-GFF converter

MAF to BED Converts a MAF formatted file to the BED format

FASTA-to-Tabular converter

FASTQ to FASTA converter from FASTX-toolkit

BAM-to-SAM convert BAM to SAM

SAM-to-BAM convert SAM to BAM

WORKFLOWS

All workflows

Convert delimiters to TAB (Galaxy Version 1.0.1)

Convert all

Colons

in Query

5: Search in textfiles on data 4

Job Resource Parameters

Specify job resource parameters

Memory (GB) Maximum Job Memory

Time (hours) Maximum job time

✓ Execute

What it does

Converts all delimiters of a specified type into TABs. Consecutive characters are condensed. For example, if columns are separated by 5 spaces they will converted into 1 tab.

Example

- Input file:

```
chrX|151283558|151283724|NM_000808_exon_8_0_chrX_151283559_r|0|-  
chrX|151370273|151370486|NM_000808_exon_9_0_chrX_151370274_r|0|-  
chrX|151559494|151559583|NM_018558_exon_1_0_chrX_151559495_f|0|+  
chrX|151564643|151564711|NM_018558_exon_2_0_chrX_151564644_f|||0|+
```
- Converting all pipe delimiters of the above file to TABs will get:

```
chrX 151283558 151283724 NM_000808_exon_8_0_chrX_151283559_r 0 -  
chrX 151370273 151370486 NM_000808_exon_9_0_chrX_151370274_r 0 -  
chrX 151559494 151559583 NM_018558_exon_1_0_chrX_151559495_f 0 +  
chrX 151564643 151564711 NM_018558_exon_2_0_chrX_151564644_f 0 +
```

Requirements: ?
- python (Version 3.8)

History

search datasets ? x

L09-B
5 shown
2 GB

5: Search in textfiles on data 4

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

The screenshot shows the Galaxy web interface with a workflow titled "L09-B" for extracting gene, transcript, and protein IDs from FASTA headers.

Tools:

- search tools
- Upload Data

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

BED

Annotation

Multiple Alignments

NCBI BLAST+

Mapping

SAM/BAM

Assembly

FASTQ Quality Control

FASTA/FASTQ

RNA-seq

CD-HIT

Datamash

EMBOSS

MUMmer4

Nanopore

History:

- 6: Convert on data 5
- 5: Search in textfiles on data 4
- 4: Homo_sapiens.GRCh38.pep.all.fa
- 3: Hsapiens_BED
- 2: Homo_sapiens.GRCh38.109.gtf
- 1: Homo_sapiens.GRCh38.109.gff3

Data View:

Header	GRCh38	CHR_HSCHR7_2_CTG6	142847306	142847317	1 gene	ENSG00000282253.1 transcript	ENST00
>ENSP00000488240.1 pep chromosome							
>ENSP00000451042.1 pep chromosome	GRCh38	14	22438547	22438554	1 gene	ENSG00000223997.1 transcript	ENST00
>ENSP00000452494.1 pep chromosome	GRCh38	14	22449113	22449125	1 gene	ENSG00000228985.1 transcript	ENST00
>ENSP00000451515.1 pep chromosome	GRCh38	14	22439007	22439015	1 gene	ENSG00000237235.2 transcript	ENST00
>ENSP00000487941.1 pep chromosome	GRCh38	7	142786213	142786224	1 gene	ENSG00000282431.1 transcript	ENST00
>ENSP00000488695.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105866322	105866332	-1 gene	ENSG00000282455.1 transcript	ENST00
>ENSP00000488000.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105881805	105881824	-1 gene	ENSG00000282323.1 transcript	ENST00
>ENSP00000488392.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105882310	105882327	-1 gene	ENSG00000282724.1 transcript	ENST00
>ENSP00000488113.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105884674	105884693	-1 gene	ENSG00000282674.1 transcript	ENST00
>ENSP00000488168.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105885641	105885659	-1 gene	ENSG00000282640.1 transcript	ENST00
>ENSP00000488711.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105886802	105886832	-1 gene	ENSG00000282396.1 transcript	ENST00
>ENSP00000487599.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105889322	105889349	-1 gene	ENSG00000281984.1 transcript	ENST00
>ENSP00000487812.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105891962	105891978	-1 gene	ENSG00000282592.1 transcript	ENST00
>ENSP00000487789.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105892470	105892490	-1 gene	ENSG00000282487.1 transcript	ENST00
>ENSP00000488201.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105894313	105894332	-1 gene	ENSG00000282346.1 transcript	ENST00
>ENSP00000488261.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105895279	105895294	-1 gene	ENSG00000282274.1 transcript	ENST00
>ENSP00000487787.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105896405	105896441	-1 gene	ENSG00000282232.1 transcript	ENST00
>ENSP00000487993.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105898728	105898758	-1 gene	ENSG00000282818.1 transcript	ENST00
>ENSP00000488522.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105901409	105901425	-1 gene	ENSG00000282736.1 transcript	ENST00
>ENSP00000488592.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105901913	105901933	-1 gene	ENSG00000282042.1 transcript	ENST00
>ENSP00000487922.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105903420	105903442	-1 gene	ENSG00000282102.1 transcript	ENST00
>ENSP00000488735.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105904387	105904402	-1 gene	ENSG00000281940.1 transcript	ENST00
>ENSP00000488475.2 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105905268	105905298	-1 gene	ENSG00000282373.1 transcript	ENST00
>ENSP00000487775.2 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105905452	105905482	-1 gene	ENSG00000281939.1 transcript	ENST00
>ENSP00000488083.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105907982	105908012	-1 gene	ENSG00000282132.1 transcript	ENST00
>ENSP00000488720.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105910678	105910694	-1 gene	ENSG00000282495.1 transcript	ENST00
>ENSP00000488589.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105911181	105911198	-1 gene	ENSG00000282010.1 transcript	ENST00
>ENSP00000487937.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105913028	105913047	-1 gene	ENSG00000282769.1 transcript	ENST00
>ENSP00000488889.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105913993	105914008	-1 gene	ENSG00000282227.1 transcript	ENST00
>ENSP00000487903.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105915130	105915160	-1 gene	ENSG00000282754.1 transcript	ENST00
>ENSP00000487604.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105917597	105917627	-1 gene	ENSG00000282578.1 transcript	ENST00
>ENSP00000488840.1 pep chromosome	GRCh38	CHR_HSCHR14_3_CTG1	105920273	105920289	-1 gene	ENSG00000282714.1 transcript	ENST00
>ENSP00000473787.1 pep chromosome	GRCh38	15	20011153	20011169	-1 gene	ENSG00000271336.1 transcript	ENST00
>ENSP00000473849.1 pep chromosome	GRCh38	15	20003840	20003862	-1 gene	ENSG00000270961.1 transcript	ENST00
>ENSP00000474065.2 pep chromosome	GRCh38	15	20008402	20008432	-1 gene	ENSG00000282599.1 transcript	ENST00
>ENSP00000473700.1 pep chromosome	GRCh38	15	21010494	21010516	-1 gene	ENSG00000270824.1 transcript	ENST00
>ENSP00000474017.2 pep chromosome	GRCh38	15	21015048	21015078	-1 gene	ENSG00000282268.1 transcript	ENST00
>ENSP00000475053.2 pep chromosome	GRCh38	15	21011451	21011469	-1 gene	ENSG00000270451.1 transcript	ENST00
>ENSP00000474133.2 pep chromosome	GRCh38	15	20005905	20005935	-1 gene	ENSG00000282520.1 transcript	ENST00
>ENSP00000474222.1 pep chromosome	GRCh38	15	21017800	21017816	-1 gene	ENSG00000270185.1 transcript	ENST00
>ENSP00000474573.2 pep chromosome	GRCh38	15	21012559	21012589	-1 gene	ENSG00000282089.1 transcript	ENST00
>ENSP00000474693.2 pep chromosome	GRCh38	15	20004797	20004815	-1 gene	ENSG00000271317.1 transcript	ENST00
>ENSP00000418639.1 pep chromosome	GRCh38	14	105865551	105865561	-1 gene	ENSG00000236597.1 transcript	ENST00
>ENSP00000420733.1 pep chromosome	GRCh38	14	105881034	105881053	-1 gene	ENSG00000211907.1 transcript	ENST00

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

Tools

cut x

Upload Data

Show Sections

Cut columns from a table

Advanced Cut columns from a table (cut)

Cutadapt Remove adapter sequences from FASTQ/FASTA

Filter by quality

Trimmomatic flexible read trimming tool for Illumina NGS data

Trim sequences

cutseq Removes a specified section from a sequence

WORKFLOWS

All workflows

Cut columns from a table (Galaxy Version 1.0.2)

Cut columns

c1,c7,c8

Delimited by

Tab

From

6: Convert on data 5

Job Resource Parameters

Specify job resource parameters

Memory (GB)

7

Maximum Job Memory

Time (hours)

24

Maximum job time

Execute

WARNING: This tool breaks column assignments. To re-establish column assignments run the tools and click on the pencil icon in the latest history item.

i The output of this tool is always in tabular format (e.g., if your original delimiters are commas, they will be replaced with tabs). For example:

Cutting columns 1 and 3 from:

```
apple,is,good  
windows,is,bad
```

will give:

```
apple good  
windows bad
```

What it does

This tool selects (cuts out) specified columns from the dataset.

- Columns are specified as **c1**, **c2**, and so on. Column count begins with **1**
- Columns can be specified in any order (e.g., **c2,c1,c6**)
- If you specify more columns than actually present - empty spaces will be filled with dots

Example

Input dataset (six columns: c1, c2, c3, c4, c5, and c6):

History

search datasets

L09-B

6 shown

2.04 GB

6: Convert on data 5

5: Search in textfiles on data 4

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

The screenshot shows the Galaxy web interface with a tool workflow for extracting gene, transcript, and protein IDs from FASTA headers.

Tools Panel: On the left, a sidebar lists various tool categories: 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, BED, Annotation, Multiple Alignments, NCBI BLAST+, Mapping, SAM/BAM, Assembly, FASTQ Quality Control, FASTA/FASTQ, RNA-seq, CD-HIT, Datamash, EMBOSS, MUMmer4, and Nanopore.

Tool Workflow: The main area displays a tool workflow consisting of several steps:

- Step 1: Homo_sapiens.GRCh38.109.gff3** (highlighted in green)
- Step 2: Homo_sapiens.GRCh38.109.gtf**
- Step 3: Hsapiens_BED**
- Step 4: Homo_sapiens.GRCh38.pep.all.fa**
- Step 5: Search in textfiles on data 4**
- Step 6: Convert on data 5**
- Step 7: Cut on data 6**

History Panel: On the right, the History panel shows the dataset details for "L09-B": 7 shown, 2.05 GB, with options to search datasets, edit, or delete.

Header	Sequence	Description
>ENSP0000488240.1 pep chromosome	ENSG0000282253.1 transcript	ENST0000631435.1 gene_biotype
>ENSP0000451042.1 pep chromosome	ENSG0000223997.1 transcript	ENST0000415118.1 gene_biotype
>ENSP0000452494.1 pep chromosome	ENSG0000228985.1 transcript	ENST0000448914.1 gene_biotype
>ENSP0000451515.1 pep chromosome	ENSG0000237235.2 transcript	ENST0000434970.2 gene_biotype
>ENSP0000487941.1 pep chromosome	ENSG0000282431.1 transcript	ENST0000632684.1 gene_biotype
>ENSP0000488695.1 pep chromosome	ENSG0000282455.1 transcript	ENST0000632524.1 gene_biotype
>ENSP0000488000.1 pep chromosome	ENSG0000282323.1 transcript	ENST0000633009.1 gene_biotype
>ENSP0000488392.1 pep chromosome	ENSG0000282724.1 transcript	ENST0000634070.1 gene_biotype
>ENSP0000488113.1 pep chromosome	ENSG0000282674.1 transcript	ENST0000632963.1 gene_biotype
>ENSP0000488168.1 pep chromosome	ENSG0000282640.1 transcript	ENST0000633030.1 gene_biotype
>ENSP0000488711.1 pep chromosome	ENSG0000282396.1 transcript	ENST0000633765.1 gene_biotype
>ENSP0000487599.1 pep chromosome	ENSG0000281984.1 transcript	ENST0000632619.1 gene_biotype
>ENSP0000487812.1 pep chromosome	ENSG0000282592.1 transcript	ENST0000632968.1 gene_biotype
>ENSP0000487789.1 pep chromosome	ENSG0000282487.1 transcript	ENST0000633159.1 gene_biotype
>ENSP0000488201.1 pep chromosome	ENSG0000282346.1 transcript	ENST0000631871.1 gene_biotype
>ENSP0000488261.1 pep chromosome	ENSG0000282274.1 transcript	ENST0000633010.1 gene_biotype
>ENSP0000487787.1 pep chromosome	ENSG0000282232.1 transcript	ENST0000633379.1 gene_biotype
>ENSP0000487993.1 pep chromosome	ENSG0000282818.1 transcript	ENST0000632473.1 gene_biotype
>ENSP0000488522.1 pep chromosome	ENSG0000282736.1 transcript	ENST0000631884.1 gene_biotype
>ENSP0000488592.1 pep chromosome	ENSG0000282042.1 transcript	ENST0000632859.1 gene_biotype
>ENSP0000487922.1 pep chromosome	ENSG0000282102.1 transcript	ENST0000631895.1 gene_biotype
>ENSP0000488735.1 pep chromosome	ENSG0000281940.1 transcript	ENST0000634154.1 gene_biotype
>ENSP0000488475.2 pep chromosome	ENSG0000282373.1 transcript	ENST0000632609.1 gene_biotype
>ENSP0000487775.2 pep chromosome	ENSG0000281939.1 transcript	ENST0000632911.1 gene_biotype
>ENSP0000488083.1 pep chromosome	ENSG0000282132.1 transcript	ENST0000633504.1 gene_biotype
>ENSP0000488720.1 pep chromosome	ENSG0000282495.1 transcript	ENST0000632304.1 gene_biotype
>ENSP0000488589.1 pep chromosome	ENSG0000282010.1 transcript	ENST0000632542.1 gene_biotype
>ENSP0000487937.1 pep chromosome	ENSG0000282769.1 transcript	ENST0000633968.1 gene_biotype
>ENSP0000488889.1 pep chromosome	ENSG0000282227.1 transcript	ENST0000634085.1 gene_biotype
>ENSP0000487903.1 pep chromosome	ENSG0000282754.1 transcript	ENST0000633353.1 gene_biotype
>ENSP0000487604.1 pep chromosome	ENSG0000282578.1 transcript	ENST0000631803.1 gene_biotype
>ENSP0000488840.1 pep chromosome	ENSG0000282714.1 transcript	ENST0000633210.1 gene_biotype
>ENSP0000473787.1 pep chromosome	ENSG0000271336.1 transcript	ENST0000605284.1 gene_biotype
>ENSP0000473849.1 pep chromosome	ENSG0000270961.1 transcript	ENST0000604642.1 gene_biotype
>ENSP0000474065.2 pep chromosome	ENSG0000282599.1 transcript	ENST0000603077.1 gene_biotype
>ENSP0000473700.1 pep chromosome	ENSG0000270824.1 transcript	ENST0000604446.1 gene_biotype
>ENSP0000474017.2 pep chromosome	ENSG0000282268.1 transcript	ENST0000604102.1 gene_biotype
>ENSP0000475053.2 pep chromosome	ENSG0000270451.1 transcript	ENST0000603693.1 gene_biotype
>ENSP0000474133.2 pep chromosome	ENSG0000282520.1 transcript	ENST0000604950.1 gene_biotype
>ENSP0000474222.1 pep chromosome	ENSG0000270185.1 transcript	ENST0000604838.1 gene_biotype
>ENSP0000474573.2 pep chromosome	ENSG0000282089.1 transcript	ENST0000603935.1 gene_biotype
>ENSP0000474693.2 pep chromosome	ENSG0000271317.1 transcript	ENST0000603326.1 gene_biotype
>ENSP0000418639.1 pep chromosome	ENSG0000236597.1 transcript	ENST0000439842.1 gene_biotype
>ENSP0000420733.1 pep chromosome	ENSG0000211907.1 transcript	ENST0000390567.1 gene_biotype

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

Tools

convert x

Upload Data

Show Sections

Convert delimiters to TAB

Convert delimiters to TAB

NCBI BLAST+ convert2blastmask
Convert masking information in lower-case masked FASTA input to file formats suitable for makeblastdb

Convert genome coordinates between assemblies and genomes

Convert SAM to interval

SFF converter

Quality format converter (ASCII-Numeric)

MAF to FASTA Converts a MAF formatted file to FASTA format

BED-to-bigBed converter

Tabular-to-FASTA converts tabular file to FASTA format

Wig/BedGraph-to-bigWig converter

MAF to Interval Converts a MAF formatted file to the Interval format

GFF-to-BED converter

BED-to-GFF converter

MAF to BED Converts a MAF formatted file to the BED format

FASTA-to-Tabular converter

FASTQ to FASTA converter from FASTX-toolkit

BAM-to-SAM convert BAM to SAM

SAM-to-BAM convert SAM to BAM

WORKFLOWS

All workflows

Convert delimiters to TAB (Galaxy Version 1.0.1)

Convert all

Whitespaces

in Query

7: Cut on data 6

Job Resource Parameters

Specify job resource parameters

Memory (GB)

7

Maximum Job Memory

Time (hours)

24

Maximum job time

Execute

What it does

Converts all delimiters of a specified type into TABs. Consecutive characters are condensed. For example, if columns are separated by 5 spaces they will converted into 1 tab.

Example

- Input file:

```
chrX||151283558||151283724||NM_000808_exon_8_0_chrX_151283559_r|0|-  
chrX||151370273||151370486||NM_000808_exon_9_0_chrX_151370274_r|0|-  
chrX||151559494||151559583||NM_018558_exon_1_0_chrX_151559495_f|0|+  
chrX||151564643||151564711||NM_018558_exon_2_0_chrX_151564644_f|||0|+
```
- Converting all pipe delimiters of the above file to TABs will get:

```
chrX 151283558 151283724 NM_000808_exon_8_0_chrX_151283559_r 0 -  
chrX 151370273 151370486 NM_000808_exon_9_0_chrX_151370274_r 0 -  
chrX 151559494 151559583 NM_018558_exon_1_0_chrX_151559495_f 0 +  
chrX 151564643 151564711 NM_018558_exon_2_0_chrX_151564644_f 0 +
```

Requirements: ?

- python (Version 3.8)

History

search datasets

L09-B
7 shown
2.05 GB

7: Cut on data 6

6: Convert on data 5

5: Search in textfiles on data 4

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

The screenshot shows the Galaxy web interface with the following components:

- Tools Panel:** On the left, a sidebar titled "Tools" lists various bioinformatics categories: 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, BED, Annotation, Multiple Alignments, NCBI BLAST+, Mapping, SAM/BAM, Assembly, FASTQ Quality Control, FASTA/FASTQ, RNA-seq, CD-HIT, Datamash, EMBOS, MUMmer4, and Nanopore.
- Search Bar:** A search bar labeled "search tools" with a magnifying glass icon and a close button "x".
- Upload Data:** A button labeled "Upload Data" with an upward arrow icon.
- Data Preview:** The main area displays a table of genomic data extracted from FASTA headers. The columns are: ID, Type, Chromosome, Ensembl ID, Type, Ensembl ID, and Biotype. The data consists of 40 rows, each starting with a header like ">ENSP00000488240.1".
- History Panel:** On the right, a "History" section titled "L09-B" shows the following steps:
 - 8 shown
 - 2.06 GB
 - 8: Convert on data 7
 - 7: Cut on data 6
 - 6: Convert on data 5
 - 5: Search in textfiles on data 4
 - 4: Homo_sapiens.GRCh38.pep.all.fa
 - 3: Hsapiens_BED
 - 2: Homo_sapiens.GRCh38.109.gtf
 - 1: Homo_sapiens.GRCh38.109.gff3
- Search Bar:** A search bar labeled "search datasets" with a magnifying glass icon and a close button "x".
- Tool Buttons:** Icons for search, add, edit, and delete operations.

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

Tools

sed

File to process

8: Convert on data 7

SED Program

```
s/>//
```

Advanced Options

Hide Advanced Options

Job Resource Parameters

Specify job resource parameters

Memory (GB)

7

Maximum Job Memory

Time (hours)

24

Maximum job time

Execute

What it does

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

TIP: This tool uses the **extended regular expression syntax** (same as running 'sed -r').

Further reading

- Short sed tutorial (http://www.linuxhowtos.org/System/sed_tutorial.htm)
- Long sed tutorial (<http://www.grymoire.com/Unix/Sed.html>)
- sed faq with good examples (<http://sed.sourceforge.net/sedfaq.html>)
- sed cheat-sheet (<http://www.catonmat.net/download/sed.stream.editor.cheat.sheet.pdf>)

Sed commands

The most useful sed command is **s** (substitute).

Examples

- s/hsa//** will remove the first instance of 'hsa' in every line.
- s/hsa/g** will remove all instances (because of the **g**) of 'hsa' in every line.
- s/A{4,}--&--/g** will find sequences of 4 or more consecutive A's, and once found, will surround them with two dashes from each side. The **&** marker is a place holder for whatever matched the regular expression!

History

search datasets

L09-B

8 shown

2.06 GB

8: Convert on data 7

7: Cut on data 6

6: Convert on data 5

5: Search in textfiles on data 4

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

The screenshot shows the Galaxy web interface with a workflow for extracting gene, transcript, and protein IDs from FASTA headers.

Tools Panel: On the left, a sidebar lists various tool categories: 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, BED, Annotation, Multiple Alignments, NCBI BLAST+, Mapping, SAM/BAM, Assembly, FASTQ Quality Control, FASTA/FASTQ, RNA-seq, CD-HIT, Datamash, EMBOSS, MUMmer4, and Nanopore.

Workflow Data Table: The main area displays a table of genomic data. Each row contains the following columns: ID, Type, Chromosome, Reference ID, Type, Reference ID, and Category. The data consists of 40 rows of entries, such as ENSP00000488240.1 (pep, chromosome, ENSG00000282253.1, transcript, ENST00000631435.1, gene_biotype), followed by ENSP00000451042.1 through ENSP00000420733.1.

History Panel: On the right, the History panel shows a list of 9 steps in the workflow:

- 9: Text transformation on data 8
- 8: Convert on data 7
- 7: Cut on data 6
- 6: Convert on data 5
- 5: Search in textfiles on data 4
- 4: Homo_sapiens.GRCh38.pep.all.fa
- 3: Hsapiens_BED
- 2: Homo_sapiens.GRCh38.109.gtf
- 1: Homo_sapiens.GRCh38.109.gff3

Each step has an edit icon (pencil) and a delete icon (X).

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

Tools

cut x

Upload Data

Show Sections

Cut columns from a table

Advanced Cut columns from a table (cut)

Cutadapt Remove adapter sequences from FASTQ/Fasta

Filter by quality

Trimmomatic flexible read trimming tool for Illumina NGS data

Trim sequences

cutseq Removes a specified section from a sequence

WORKFLOWS

All workflows

Cut columns from a table (Galaxy Version 1.0.2)

Cut columns
c4,c6,c1

Delimited by
Tab

From
9: Text transformation on data 8

Job Resource Parameters

Specify job resource parameters

Memory (GB)
7

Maximum Job Memory

Time (hours)
24

Maximum job time

Execute

WARNING: This tool breaks column assignments. To re-establish column assignments run the tools and click on the pencil icon in the latest history item.

i The output of this tool is always in tabular format (e.g., if your original delimiters are commas, they will be replaced with tabs). For example:

Cutting columns 1 and 3 from:

```
apple,is,good  
windows,is,bad
```

will give:

```
apple    good  
windows bad
```

What it does

This tool selects (cuts out) specified columns from the dataset.

- Columns are specified as **c1**, **c2**, and so on. Column count begins with **1**
- Columns can be specified in any order (e.g., **c2,c1,c6**)
- If you specify more columns than actually present - empty spaces will be filled with dots

Example

Input dataset (six columns: c1, c2, c3, c4, c5, and c6):

History

search datasets

L09-B

9 shown

2.07 GB

9: Text transformation on data 8

8: Convert on data 7

7: Cut on data 6

6: Convert on data 5

5: Search in textfiles on data 4

4: Homo_sapiens.GRCh38.pep.all.fa

3: Hsapiens_BED

2: Homo_sapiens.GRCh38.109.gtf

1: Homo_sapiens.GRCh38.109.gff3

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

The screenshot shows the Galaxy web interface with a workflow titled "L09-B".

Tools:

- cut** (selected)
- Upload Data**
- Show Sections**

Cut columns from a table

Advanced Cut columns from a table (cut)

Cutadapt Remove adapter sequences from FASTQ/FASTA

Filter by quality

Trimmomatic flexible read trimming tool for Illumina NGS data

Trim sequences

cutseq Removes a specified section from a sequence

WORKFLOWS

All workflows

L09-B

10 shown
2.08 GB

History

- 10: Cut on data 9
- 9: Text transformation on data 8
- 8: Convert on data 7
- 7: Cut on data 6
- 6: Convert on data 5
- 5: Search in textfiles on data 4
- 4: Homo_sapiens.GRCh38.pep.all.fa
- 3: Hsapiens_BED
- 2: Homo_sapiens.GRCh38.109.gtf
- 1: Homo_sapiens.GRCh38.109.gff3

The main workspace displays a table of genomic identifiers:

ENSG00000282253.1	ENST00000631435.1	ENSP00000488240.1
ENSG00000223997.1	ENST00000415118.1	ENSP00000451042.1
ENSG00000228985.1	ENST00000448914.1	ENSP00000452494.1
ENSG00000237235.2	ENST00000434970.2	ENSP00000451515.1
ENSG00000282431.1	ENST00000632684.1	ENSP00000487941.1
ENSG00000282455.1	ENST00000632524.1	ENSP00000488695.1
ENSG00000282323.1	ENST00000633009.1	ENSP00000488000.1
ENSG00000282724.1	ENST00000634070.1	ENSP00000488392.1
ENSG00000282674.1	ENST00000632963.1	ENSP00000488113.1
ENSG00000282640.1	ENST00000633030.1	ENSP00000488168.1
ENSG00000282396.1	ENST00000633765.1	ENSP00000488711.1
ENSG00000281984.1	ENST00000632619.1	ENSP00000487599.1
ENSG00000282592.1	ENST00000632968.1	ENSP00000487812.1
ENSG00000282487.1	ENST00000633159.1	ENSP00000487789.1
ENSG00000282346.1	ENST00000631871.1	ENSP00000488201.1
ENSG00000282274.1	ENST00000633010.1	ENSP00000488261.1
ENSG00000282232.1	ENST00000633379.1	ENSP00000487787.1
ENSG00000282818.1	ENST00000632473.1	ENSP00000487993.1
ENSG00000282736.1	ENST00000631884.1	ENSP00000488522.1
ENSG00000282042.1	ENST00000632859.1	ENSP00000488592.1
ENSG00000282102.1	ENST00000631895.1	ENSP00000487922.1
ENSG00000281940.1	ENST00000634154.1	ENSP00000488735.1
ENSG00000282373.1	ENST00000632609.1	ENSP00000488475.2
ENSG00000281939.1	ENST00000632911.1	ENSP00000487775.2
ENSG00000282132.1	ENST00000633504.1	ENSP00000488083.1
ENSG00000282495.1	ENST00000632304.1	ENSP00000488720.1
ENSG00000282010.1	ENST00000632542.1	ENSP00000488589.1
ENSG00000282769.1	ENST00000633968.1	ENSP00000487937.1
ENSG00000282227.1	ENST00000634085.1	ENSP00000488889.1
ENSG00000282754.1	ENST00000633353.1	ENSP00000487903.1
ENSG00000282578.1	ENST00000631803.1	ENSP00000487604.1
ENSG00000282714.1	ENST00000633210.1	ENSP00000488840.1
ENSG00000271336.1	ENST00000605284.1	ENSP00000473787.1
ENSG00000270961.1	ENST00000604642.1	ENSP00000473849.1
ENSG00000282599.1	ENST00000603077.1	ENSP00000474065.2
ENSG00000270824.1	ENST00000604446.1	ENSP00000473700.1
ENSG00000282268.1	ENST00000604102.1	ENSP00000474017.2
ENSG00000270451.1	ENST00000603693.1	ENSP00000475053.2
ENSG00000282520.1	ENST00000604950.1	ENSP00000474133.2
ENSG00000270185.1	ENST00000604838.1	ENSP00000474222.1
ENSG00000282089.1	ENST00000603935.1	ENSP00000474573.2
ENSG00000271317.1	ENST00000603326.1	ENSP00000474693.2
ENSG00000236597.1	ENST00000439842.1	ENSP00000418639.1
ENSG00000211907.1	ENST00000390567.1	ENSP00000420733.1

Working With Genome Files In Galaxy

Extracting Gene, Transcript and Protein IDs From Fasta Headers

The screenshot shows the Galaxy web interface with a workflow for extracting gene, transcript, and protein IDs from FASTA headers.

Tools Panel: On the left, a sidebar lists various tools categorized under "HPRC".

Workflow Data: The main area displays a table of extracted data with three columns: Gene ID, Transcript ID, and Protein ID.

Gene ID	Transcript ID	Protein ID
ENSG00000282253.1	ENST00000631435.1	ENSP00000488240.1
ENSG00000223997.1	ENST00000415118.1	ENSP00000451042.1
ENSG00000228985.1	ENST00000448914.1	ENSP00000452494.1
ENSG00000237235.2	ENST00000434970.2	ENSP00000451515.1
ENSG00000282431.1	ENST00000632684.1	ENSP00000487941.1
ENSG00000282455.1	ENST00000632524.1	ENSP00000488695.1
ENSG00000282323.1	ENST00000633009.1	ENSP00000488000.1
ENSG00000282724.1	ENST00000634070.1	ENSP00000488392.1
ENSG00000282674.1	ENST00000632963.1	ENSP00000488113.1
ENSG00000282640.1	ENST00000633030.1	ENSP00000488168.1
ENSG00000282396.1	ENST00000633765.1	ENSP00000488711.1
ENSG00000281984.1	ENST00000632619.1	ENSP00000487599.1
ENSG00000282592.1	ENST00000632968.1	ENSP00000487812.1
ENSG00000282487.1	ENST00000633159.1	ENSP00000487789.1
ENSG00000282346.1	ENST00000631871.1	ENSP00000488201.1
ENSG00000282274.1	ENST00000633010.1	ENSP00000488261.1
ENSG00000282232.1	ENST00000633379.1	ENSP00000487787.1
ENSG00000282818.1	ENST00000632473.1	ENSP00000487993.1
ENSG00000282736.1	ENST00000631884.1	ENSP00000488522.1
ENSG00000282042.1	ENST00000632859.1	ENSP00000488592.1
ENSG00000282102.1	ENST00000631895.1	ENSP00000487922.1
ENSG00000281940.1	ENST00000634154.1	ENSP00000488735.1
ENSG00000282373.1	ENST00000632609.1	ENSP00000488475.2
ENSG00000281939.1	ENST00000632911.1	ENSP00000487775.2
ENSG00000282132.1	ENST00000633504.1	ENSP00000488083.1
ENSG00000282495.1	ENST00000632304.1	ENSP00000488720.1
ENSG00000282010.1	ENST00000632542.1	ENSP00000488589.1
ENSG00000282769.1	ENST00000633968.1	ENSP00000487937.1
ENSG00000282227.1	ENST00000634085.1	ENSP00000488889.1
ENSG00000282754.1	ENST00000633353.1	ENSP00000487903.1
ENSG00000282578.1	ENST00000631803.1	ENSP00000487604.1
ENSG00000282714.1	ENST00000633210.1	ENSP00000488840.1
ENSG00000271336.1	ENST00000605284.1	ENSP00000473787.1
ENSG00000270961.1	ENST00000604642.1	ENSP00000473849.1
ENSG00000282599.1	ENST00000603077.1	ENSP00000474065.2
ENSG00000270824.1	ENST00000604446.1	ENSP00000473700.1
ENSG00000282268.1	ENST00000604102.1	ENSP00000474017.2
ENSG00000270451.1	ENST00000603693.1	ENSP00000475053.2
ENSG00000282520.1	ENST00000604950.1	ENSP00000474133.2
ENSG00000270185.1	ENST00000604838.1	ENSP00000474222.1
ENSG00000282089.1	ENST00000603935.1	ENSP00000474573.2
ENSG00000271317.1	ENST00000603326.1	ENSP00000474693.2
ENSG00000236597.1	ENST00000439842.1	ENSP00000418639.1
ENSG00000211907.1	ENST00000390567.1	ENSP00000420733.1

History Panel: On the right, the history panel shows the steps of the workflow:

- 10: Gene-Transcript-Protein-Table
- 9: Text transformation on data 8
- 8: Convert on data 7
- 7: Cut on data 6
- 6: Convert on data 5
- 5: Search in textfiles on data 4
- 4: Homo_sapiens.GRCh38.pep.all.fa
- 3: Hsapiens_BED
- 2: Homo_sapiens.GRCh38.109.gtf
- 1: Homo_sapiens.GRCh38.109.gff3

Working With Genome Files In Galaxy

For a BED File

- Can we identify Supercontigs?
- Can we identify Chromosomes?
- Can we identify Genes?
- Can we identify transcripts?
- Can we identify Proteins?
- Can we identify/extract All the Features being declared?
- Can we separate them by those present in the Watson and/or Crick DNA strands?

Working With Genome Files In Galaxy

For a GTF File

- Can we identify Supercontigs?
- Can we identify Chromosomes?
- Can we identify Genes?
- Can we identify transcripts?
- Can we identify Proteins?
- Can we identify/extract All the Features being declared?
- Can we separate them by those present in the Watson and/or Crick DNA strands?

Working With Genome Files In Galaxy

For a GFF3 File

- Can we identify Supercontigs?
- Can we identify Chromosomes?
- Can we identify Genes?
- Can we identify transcripts?
- Can we identify Proteins?
- Can we identify/extract All the Features being declared?
- Can we separate them by those present in the Watson and/or Crick DNA strands?