

Obtaining gene and repeat data

It is possible to view the location of genes and repeats with reference to the aligned data. The required data can be obtained from the UCSC genome browser 'Table Browser'. The genome browser is located here: <https://genome.ucsc.edu/index.html>, with the Table Browser accessed via the Tools > Table Browser menu option (Figure 1).

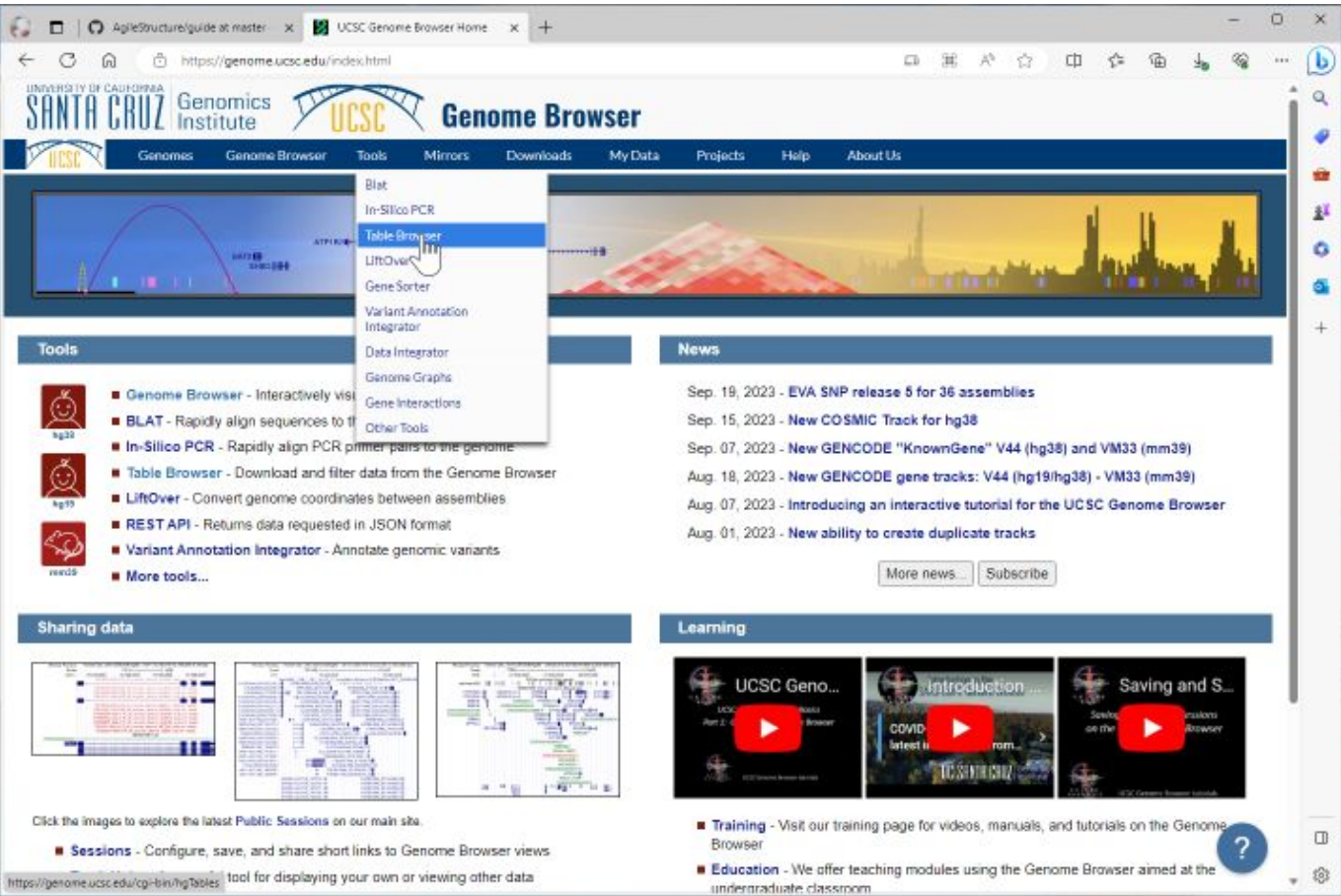


Figure 1

The web page contains a number of options used to select the genome to which the long read data was aligned, the type of data you want and its format. Figure 2a shows the options used to select the genomic coordinates for the genes in NCBI refseq data set for the human hg38 reference sequence. similarly , Figure 2b contains the settings for downloading the locations of the repeat sequences. The options for both datasets are very similar differing only by type of data to selected via the "group" and "track" options and the name of the file.

Table Browser

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

Select dataset

clade: Mammal genome: Human assembly: Dec. 2013 (GRCh38/hg38)

group: Genes and Gene Predictions track: NCBI RefSeq

table: RefSeq All (ncbiRefSeq)

data format description

Define region of interest

region: ☒ genome ☐ position chr7:146,116,801-148,420,998

lookup

define regions

identifiers (names/accessions):

paste list

upload list

Optional: Subset, combine, compare with another track

filter: create

subtrack merge: create

intersection: create

correlation: create

Retrieve and display data

output format: selected fields from primary and related tables

Send output to ☐ Galaxy ☐ GREAT

output filename: hg.38.txt

(add .csv extension if opening in Excel, leave blank to keep output in browser)

output field separator: ☒ tsv (tab-separated) ☐ csv (for excel)

file type returned: ☐ plain text ☒ gzip compressed

get output

summary/statistics

Figure 2a: Downloading gene coordinates

2 / 4

Table Browser

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

Select dataset

clade: Mammal

genome: Human

assembly: Dec. 2013 (GRCh38/hg38)

group: Repeats

track: RepeatMasker

table: rmsk

data format description

Define region of interest

region: ☒ genome ☐ position

chr7:146,116,801-148,420,998

lookup

define regions

identifiers (names/accessions):

paste list

upload list

Optional: Subset, combine, compare with another track

filter: create

intersection: create

Retrieve and display data

output format: selected fields from primary and related tables

Send output to ☐ Galaxy ☐ GREAT

output filename: hg38_repeats.txt

(add .csv extension if opening in Excel, leave blank to keep output in browser)

output field separator: ☒ tsv (tab-separated) ☐ csv (for excel)

file type returned: ☐ plain text ☒ gzip compressed

get output

summary/statistics

Figure 2b: Downloading repeat coordinates.

In both cases the genome option is selected to obtain data from the entire genome, while the format is set using the `selected fields from primary and related tables` and `tsv (tab-separated) text file` options. Finally, the data is compressed using the "gzip compressed" option. Pressing the `get output` directs the user to a 2nd page with which to selected what data fields are required (Figures 3a and 3b). Once the required fields have been set, pressing the `get output` button on this webpage will start the download.

3 / 4

Select Fields from hg38.ncbiRefSeq

<input type="checkbox"/>	bin	
<input checked="" type="checkbox"/>	name	Name of gene (usually transcript_id from GTF)
<input checked="" type="checkbox"/>	chrom	Reference sequence chromosome or scaffold
<input checked="" type="checkbox"/>	strand	+ or - for strand
<input checked="" type="checkbox"/>	txStart	Transcription start position (or end position for minus strand item)
<input checked="" type="checkbox"/>	txEnd	Transcription end position (or start position for minus strand item)
<input checked="" type="checkbox"/>	cdsStart	Coding region start (or end position for minus strand item)
<input checked="" type="checkbox"/>	cdsEnd	Coding region end (or start position for minus strand item)
<input checked="" type="checkbox"/>	exonCount	Number of exons
<input checked="" type="checkbox"/>	exonStarts	Exon start positions (or end positions for minus strand item)
<input checked="" type="checkbox"/>	exonEnds	Exon end positions (or start positions for minus strand item)
<input type="checkbox"/>	score	score
<input checked="" type="checkbox"/>	name2	Alternate name (e.g. gene_id from GTF)
<input type="checkbox"/>	cdsStartStat	Status of CDS start annotation (none, unknown, incomplete, or complete)
<input type="checkbox"/>	cdsEndStat	Status of CDS end annotation (none, unknown, incomplete, or complete)
<input type="checkbox"/>	exonFrames	Exon frame {0,1,2}, or -1 if no frame for exon

get output

cancel

check all

clear all

Figure 3a: Selecting the options for gene coordinates file

Select Fields from hg38.rmsk

<input type="checkbox"/>	bin	
<input type="checkbox"/>	swScore	Smith Waterman alignment score
<input type="checkbox"/>	milliDiv	Base mismatches in parts per thousand
<input type="checkbox"/>	milliDel	Bases deleted in parts per thousand
<input type="checkbox"/>	milliIns	Bases inserted in parts per thousand
<input checked="" type="checkbox"/>	genoName	Genomic sequence name
<input checked="" type="checkbox"/>	genoStart	Start in genomic sequence
<input checked="" type="checkbox"/>	genoEnd	End in genomic sequence
<input type="checkbox"/>	genoLeft	-#bases after match in genomic sequence
<input checked="" type="checkbox"/>	strand	Relative orientation + or -
<input checked="" type="checkbox"/>	repName	Name of repeat
<input checked="" type="checkbox"/>	repClass	Class of repeat
<input checked="" type="checkbox"/>	repFamily	Family of repeat
<input type="checkbox"/>	repStart	Start (if strand is +) or -#bases after match (if strand is -) in repeat sequence
<input type="checkbox"/>	repEnd	End in repeat sequence
<input type="checkbox"/>	repLeft	-#bases after match (if strand is +) or start (if strand is -) in repeat sequence
<input type="checkbox"/>	id	First digit of id field in RepeatMasker .out file. Best ignored.

get output

cancel

check all

clear all

Figure 3b Selecting the options for repeat coordinates file

Once downloaded, the files should be decompressed using a program such as 7zip (home page: <https://www.7-zip.org/> and download page: <https://www.7-zip.org/download.html>)