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).

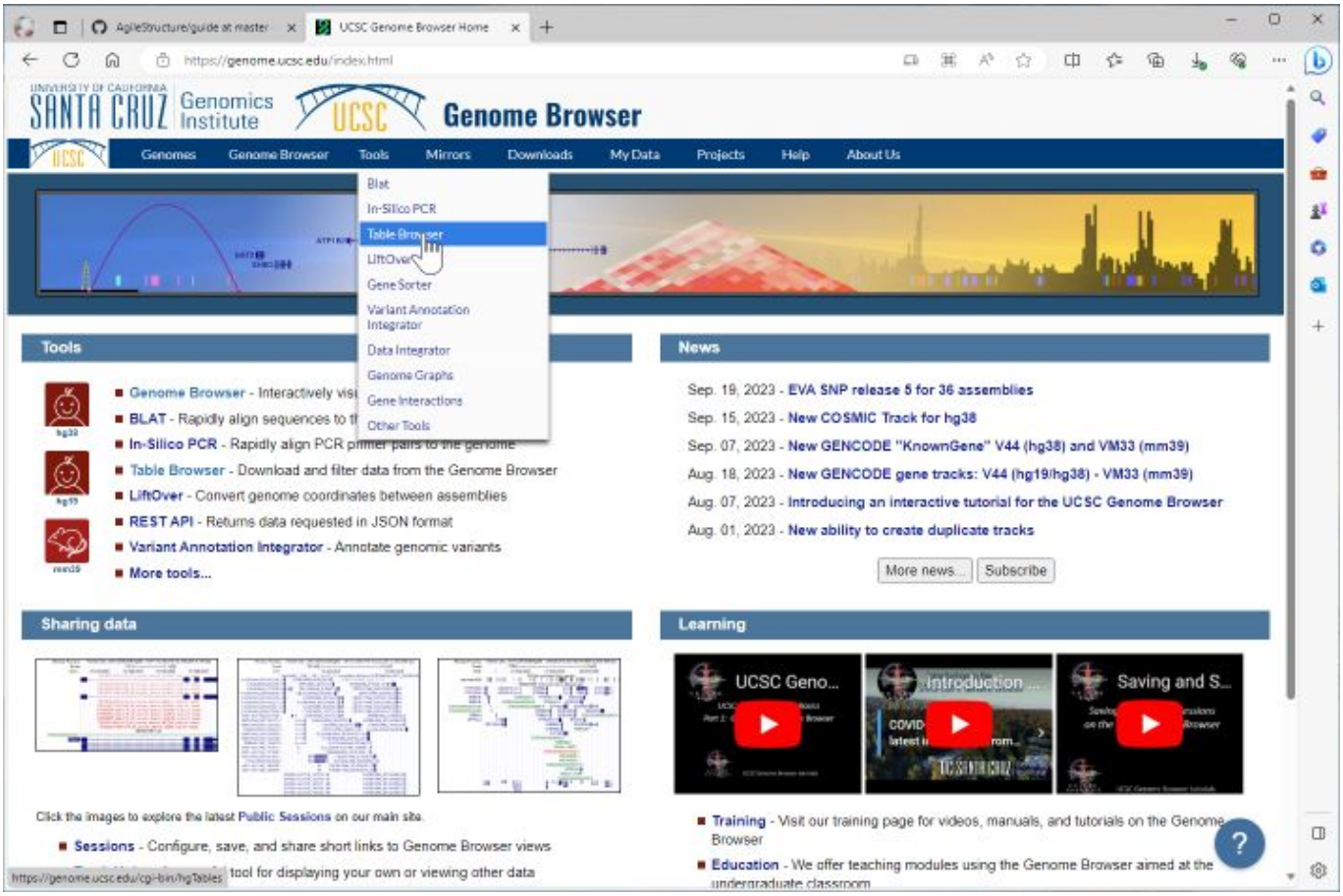


Figure 1

The web page contains of 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.

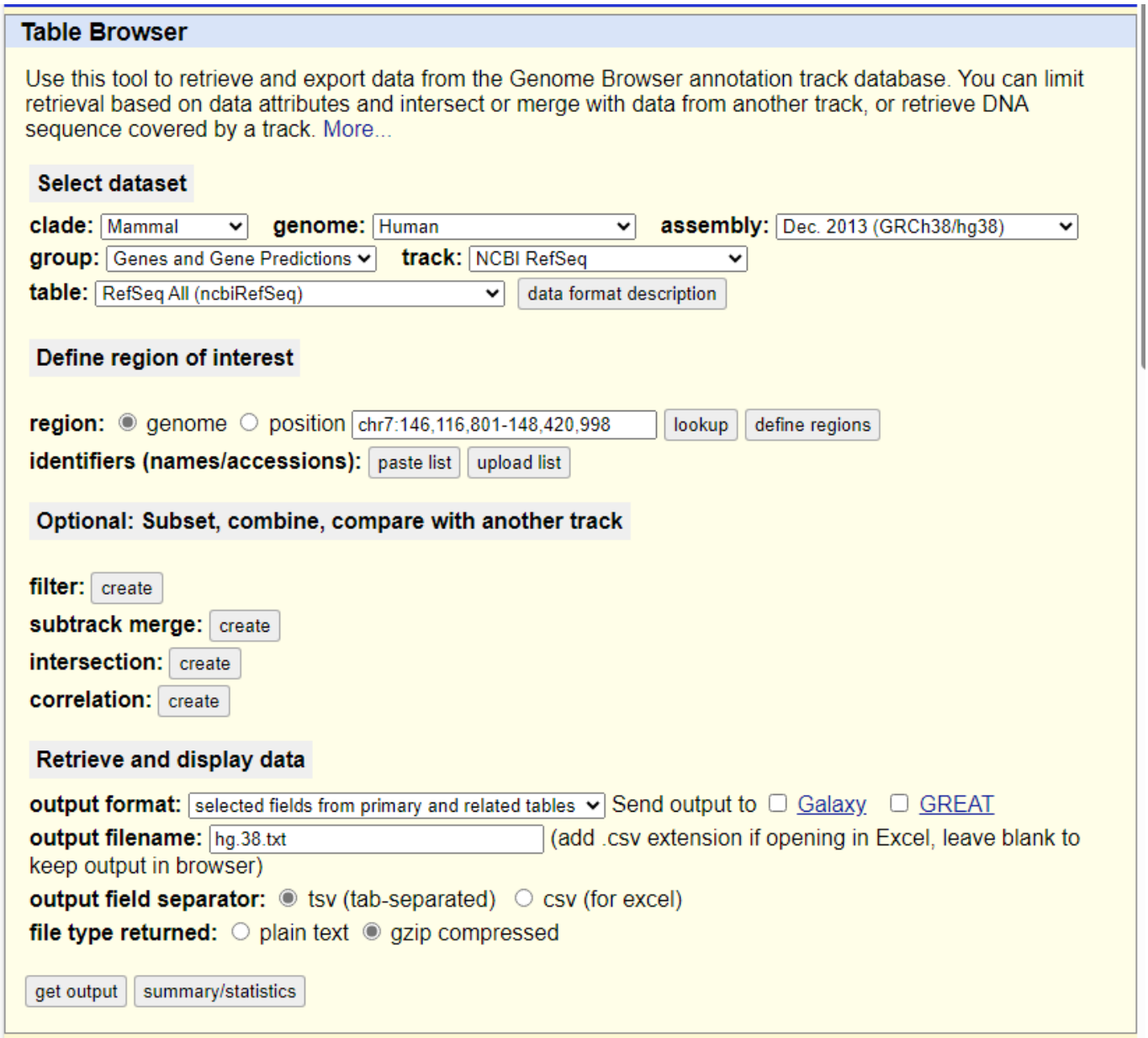
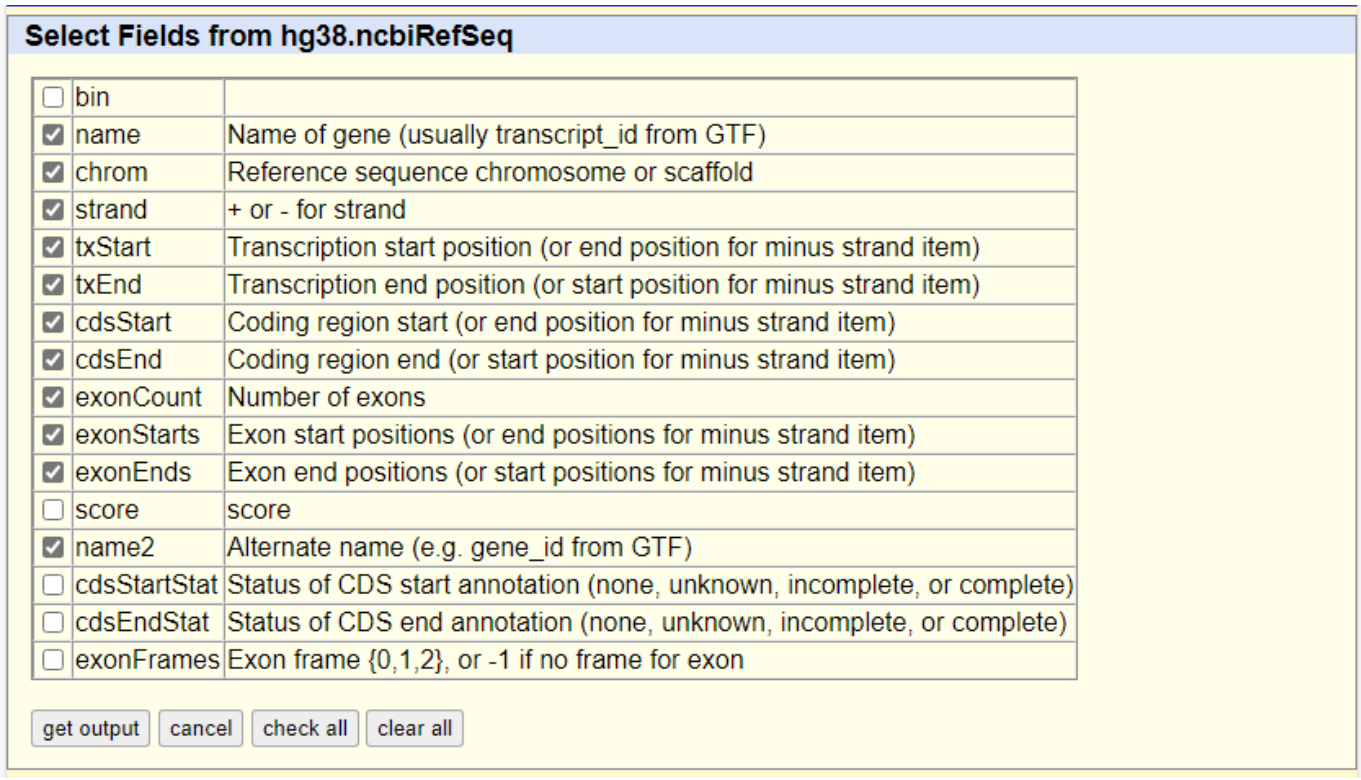


Figure 2a: Downloading gene coordinates

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.

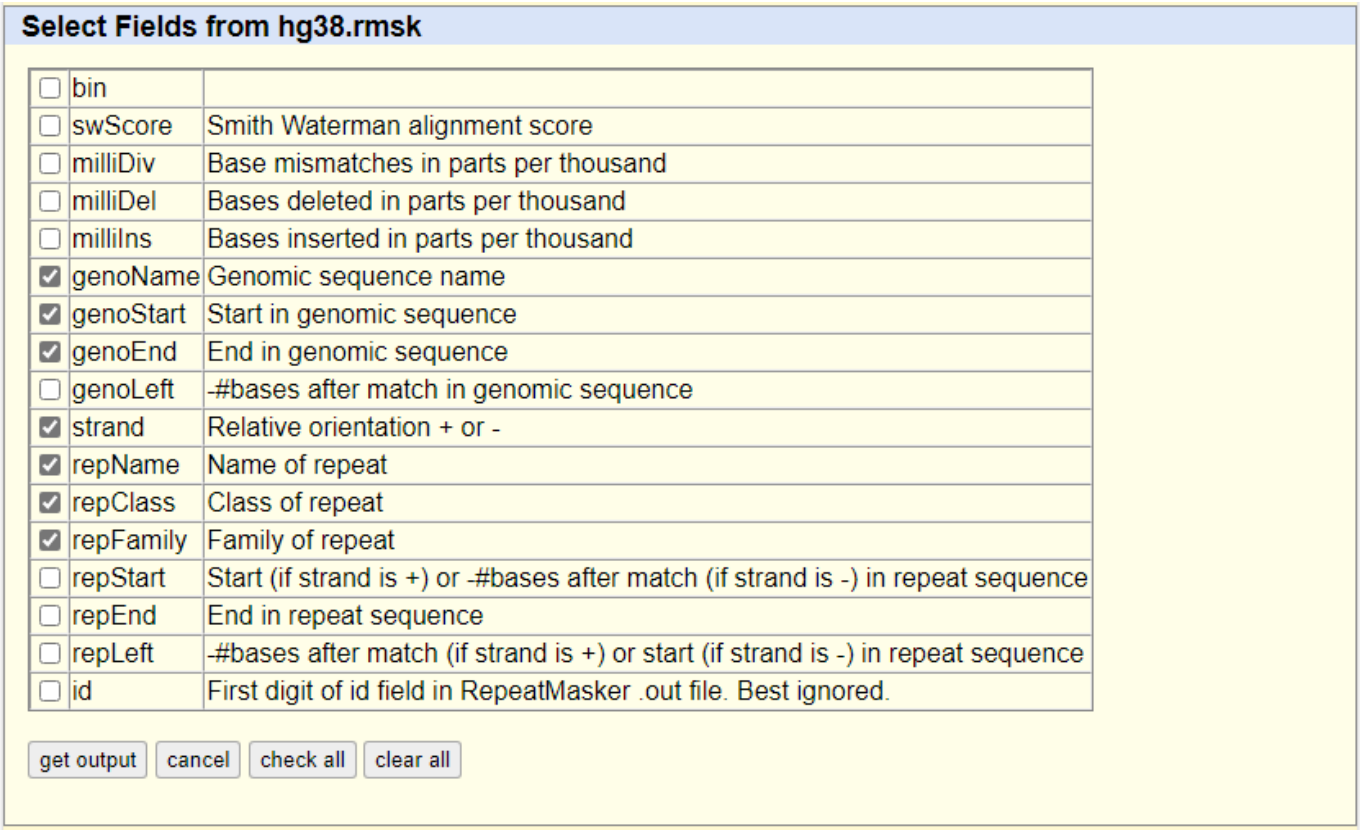


Select Fields from hg38.ncbiRefSeq

☐	bin	
☑	name	Name of gene (usually transcript_id from GTF)
☑	chrom	Reference sequence chromosome or scaffold
☑	strand	+ or - for strand
☑	txStart	Transcription start position (or end position for minus strand item)
☑	txEnd	Transcription end position (or start position for minus strand item)
☑	cdsStart	Coding region start (or end position for minus strand item)
☑	cdsEnd	Coding region end (or start position for minus strand item)
☑	exonCount	Number of exons
☑	exonStarts	Exon start positions (or end positions for minus strand item)
☑	exonEnds	Exon end positions (or start positions for minus strand item)
☐	score	score
☑	name2	Alternate name (e.g. gene_id from GTF)
☐	cdsStartStat	Status of CDS start annotation (none, unknown, incomplete, or complete)
☐	cdsEndStat	Status of CDS end annotation (none, unknown, incomplete, or complete)
☐	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

☐	bin	
☐	swScore	Smith Waterman alignment score
☐	milliDiv	Base mismatches in parts per thousand
☐	milliDel	Bases deleted in parts per thousand
☐	milliIns	Bases inserted in parts per thousand
☑	genoName	Genomic sequence name
☑	genoStart	Start in genomic sequence
☑	genoEnd	End in genomic sequence
☐	genoLeft	-#bases after match in genomic sequence
☑	strand	Relative orientation + or -
☑	repName	Name of repeat
☑	repClass	Class of repeat
☑	repFamily	Family of repeat
☐	repStart	Start (if strand is +) or -#bases after match (if strand is -) in repeat sequence
☐	repEnd	End in repeat sequence
☐	repLeft	-#bases after match (if strand is +) or start (if strand is -) in repeat sequence
☐	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)