Computing on Genomes, I

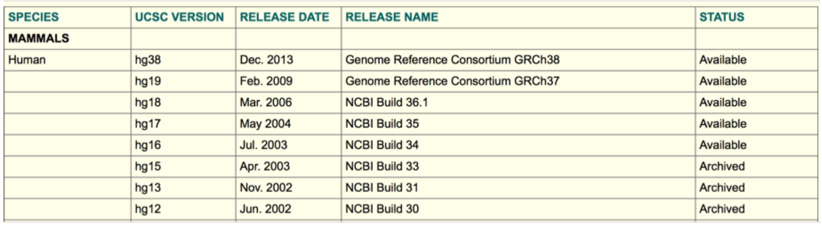
Biol4230 Friday, March 23, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

- Versions of Genomes, genome coordinates
- Genome data types:
  - sequence
  - alignment
  - features on locations
  - quantitative data
- Looking at a genome (IGV)
- incorporating data
  - The UCSC table viewer
  - importing into IGV

To learn more:

- UCSC FAQs: http://genome.ucsc.edu/FAQ/
- Genome releases:
  http://genome.ucsc.edu/FAQ/FAQreleases.html
- Genome data file formats:
  http://genome.ucsc.edu/FAQ/FAQformat.html
- Integrated Genome Viewer:
  https://www.broadinstitute.org/igv/

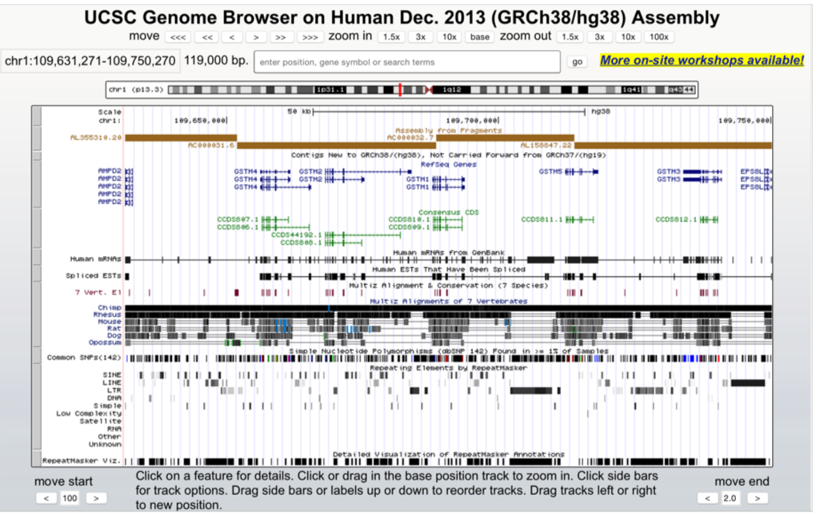
Genome Versions:



SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
MAMMALS				
Human	hg38	Dec. 2013	Genome Reference Consortium GRCh38	Available
	hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
	hg18	Mar. 2006	NCBI Build 36.1	Available
	hg17	May 2004	NCBI Build 35	Available
	hg16	Jul. 2003	NCBI Build 34	Available
	hg15	Apr. 2003	NCBI Build 33	Archived
	hg13	Nov. 2002	NCBI Build 31	Archived
	hg12	Jun. 2002	NCBI Build 30	Archived

release	chr	start	stop
hg18	1	110,031,941	110,037,890
hg19	1	110,230,418	110,236,367
hg38/	1	109,687,796	109,693,745

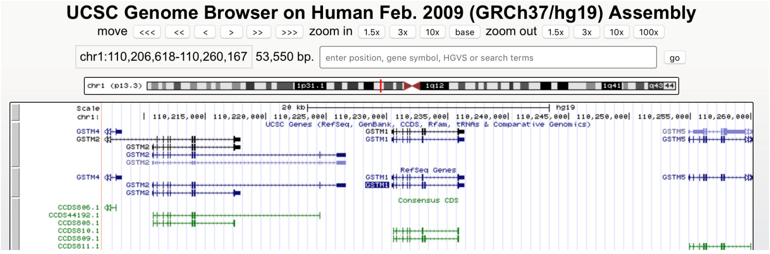
Genome Browsers (UCSC)



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

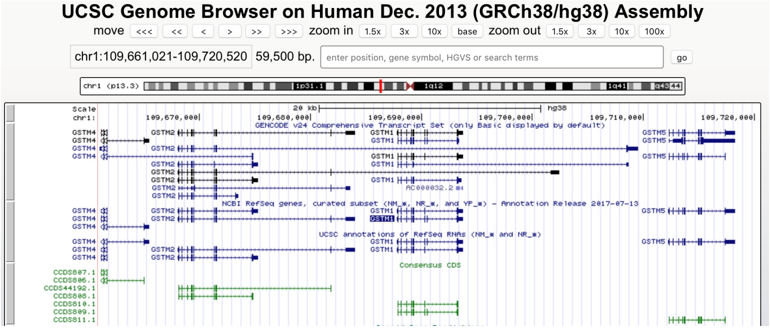
chr1:109,631,271-109,750,270 119,000 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr1:110,206,618-110,260,167 53,550 bp.



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr1:109,661,021-109,720,520 59,500 bp.

Genome data (UCSC table browser)

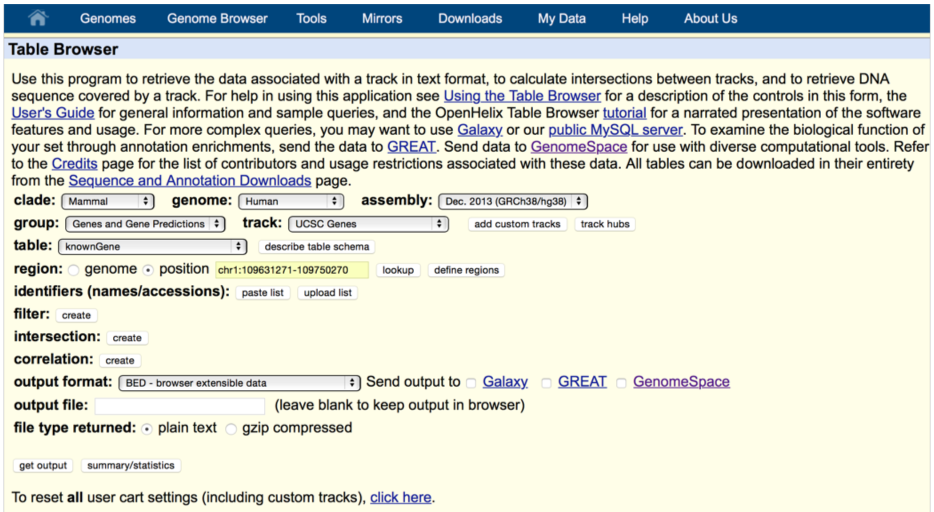


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

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

group: Genes and Gene Predictions **track:** UCSC Genes add custom tracks track hubs

table: knownGene describe table schema

region: ○ genome ● position chr1:109631271-109750270 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: BED - browser extensible data Send output to ☐ Galaxy ☐ GREAT ☐ GenomeSpace

output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Genome data (UCSC table browser)

```
chr1	hg38_refGene st_codon 109620269 109620271	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109620269 109620278	0.000000	+	0	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene exon	109619813 109620278	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109625303 109625433	0.000000	+	2	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene exon	109625303 109625433	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109625662 109625792	0.000000	+	0	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene exon	109625662 109625792	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109626160 109626228	0.000000	+	1	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene exon	109626160 109626228	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109626319 109626427	0.000000	+	1	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene exon	109626319 109626427	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109626726 109626912	0.000000	+	0	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene exon	109626726 109626912	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109627175 109627316	0.000000	+	2	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene exon	109627175 109627316	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109627429 109627518	0.000000	+	1	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene exon	109627429 109627518	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109627774 109627903	0.000000	+	1	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene exon	109627774 109627903	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109628083 109628277	0.000000	+	0	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene exon	109628083 109628277	0.000000	+	.	gene_id "NM_139156"; transcript_id "NM_139156";
chr1	hg38_refGene CDS	109628364 109628495	0.000000	+	0	gene_id "NM_139156"; transcript_id "NM_139156";
```

GFF/GTF format – Coordinate blocks

Genome data (UCSC table browser)

```
track type=wiggle_0 name="GM12878" description="H3K4Me1 Mark (Often Found Near
Regulatory Elements) on GM12878 Cells from ENCODE"
#bedGraph section chr1:110215375-110255450
chr1 110230400	110230425	2.72
chr1 110230425	110230450	3
chr1 110230450	110230475	3
chr1 110230475	110230500	3
chr1 110230500	110230525	3
chr1 110230525	110230550	3
chr1 110230550	110230575	3
chr1 110230575	110230600	1.36
chr1 110230925	110230950	0.72
chr1 110230950	110230975	1
```

Wiggle format – continuous data