MEDS5420 Lec13

Using the UCSC Genome Browser

March 1, 2023

Lead creators of UCSC genome browser: Jim Kent and David Haussler

email Jim

Jim Kent's Web Page

I'm a research scientist at UCSC. I work primarily on web tools to help understand the human genome. Before becoming a bioinformatician I got a PhD in biology working with with the Zahler lab. Before that I wrote computer art and animation programs. I live in Santa Cruz, CA and have three children, Mira, Tisa, and Maia.

Links to Stuff I Work On:

- The UCSC Genome Browser - also known as 'the golden path'.
- Parasol - A Job Control System for Computer Clusters. It's fast and it's free.
- Papers - PubMed index of my papers.
- Presentations - slides from presentations. Sadly no voice-over.
- The Intronerator - to look at C. elegans genes and splicing patterns.
- *cis*-Site Seeker - Look for regulatory regions in RNA or DNA sequences with the Improbizer.
- Cross Species Alignments - Program and some samples using the not yet famous WABA algorithm.
- Commands - Catalog of command line driven programs.
- Source Code - free for academic, non-profit or personal use.
- Executables - commonly requested compiled versions also free for academic, non-profit and personal use.

Other Useful Links

- PubMed - to search the scientific literature.
- Blast - homology searching.
- Google Scholar - search scientific literature the Google way.
- UCSC Schedule of Classes - er, when was that final?

https://cbse.soe.ucsc.edu/people/kent
https://users.soe.ucsc.edu/~kent/

https://hausslergenomics.ucsc.edu/people/david-haussler/

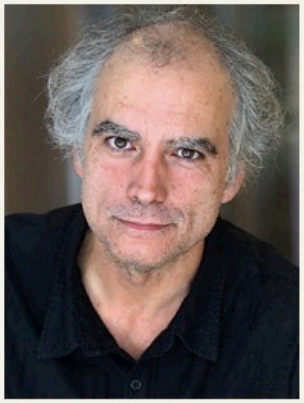
Lead creator of UCSC genome browser: Jim Kent

UNIVERSITY OF CALIFORNIA SANTA CRUZ CENTER FOR BIOMOLEC SCIENCE AND ENGINEE

ABOUT US | RESEARCH | EDUCATION

Home

Jim Kent



Director, UCSC Genome Browser Project

Research Scientist, Department of Biomolecular Engineering, UC

Jim Kent

Research Projects

Jim Kent directs the genome br
assurance staff of the UCSC Ge
created the computer program
of the human genome sequence
sequencing centers worldwide
informatics associated with the
Bioinformatics Group participat
produce, assemble, and annota

UCSC Genome Browser

The UCSC Genome Browser pr
metazoan genome sequences.
wide annotation in a web-basec

START HERE | EXPLORE | RESEARCH TALKS | STORIES | C

Home » Speakers » David Haussler

David Haussler

University of California, Santa Cruz

Share

David Haussler is Scientific D
California Santa Cruz (UCSC)
Investigator of the Howard Hu
(HHMI). Haussler uses mathe
biology to study the genomes
understanding disease and evolution. As part of the Human Genor
that published the first publicly available draft of the human genom
several large-scale projects, including the Genome 10K Project, th
Hub (CGHub), and the Global Alliance for Genomics and Health.

https://www.ibiology.org/speakers/david-haussler/

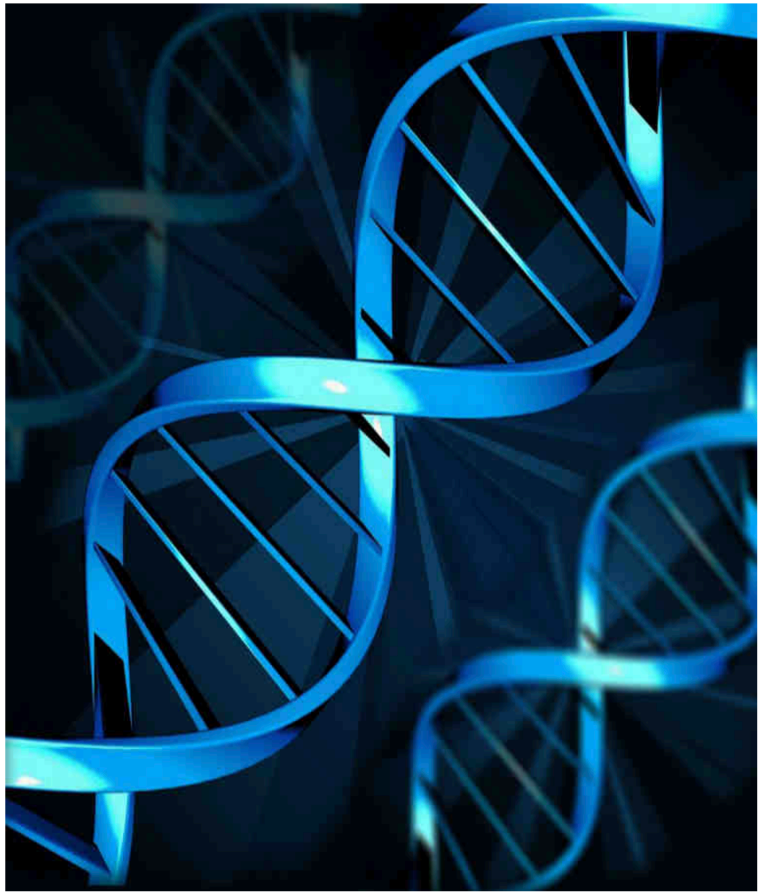
Some useful capabilities of the browser

- Some resources:
  - Genome build files (fasta format)
  - Converting genome coordinates between versions (e.g. hg19 to hg38)
  - Gene annotation files
    - Convert gene IDs from different sources (GENCODE-refSeq)
  - Obtaining DNA sequences from regions of interest
  - In-silico tools
    - BLAT- fast alignment of short sequences
    - PCR
- Custom tracks and resources:
  - Uploading custom data or annotations
  - Storing them in a user profile
  - Sharing / viewing track hubs
  - Configure Display for optimal clarity (e.g. for making figures)

UCSC homepage

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute | UCSC **Genome Browser**

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

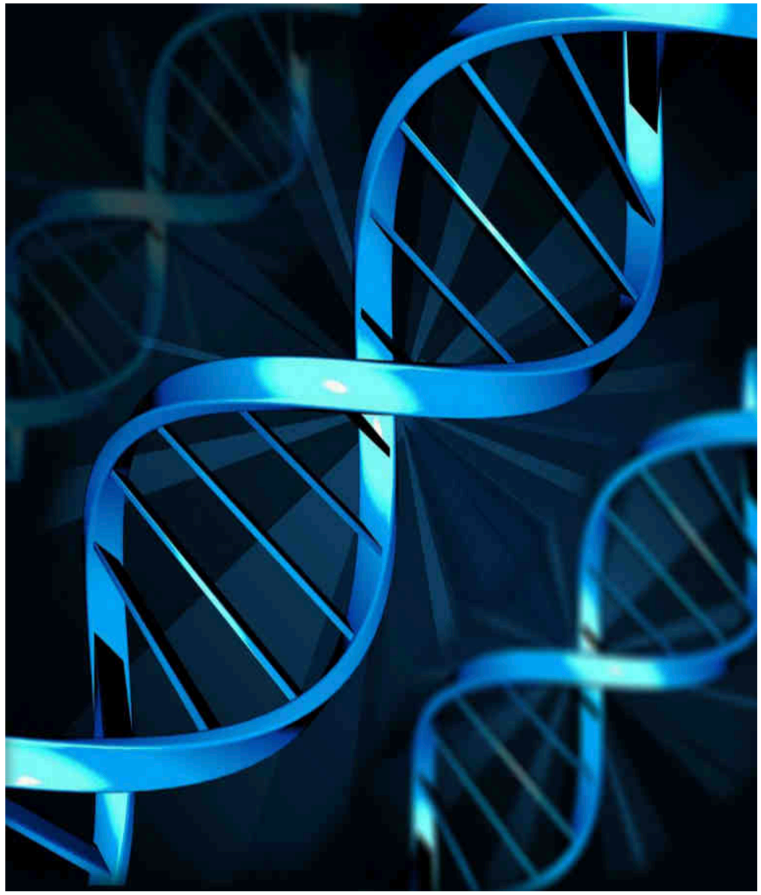


Our tools

- **Genome Browser**
  interactively visualize genomic data
- **COVID-19 Research**
  use the SARS-CoV-2 genome browser and explore coronavirus datasets
- **BLAT**
  rapidly align sequences to the genome
- **Table Browser**
  download data from the Genome Browser database
- **Variant Annotation Integrator**
  get functional effect predictions for variant calls
- **Data Integrator**
  combine data sources from the Genome Browser database
- **Genome Browser in a Box (GBiB)**
  run the Genome Browser on your laptop or server
- **In-Silico PCR**
  rapidly align PCR primer pairs to the genome
- **LiftOver**
  convert genome coordinates between assemblies
- **Track Hubs**
  import and view external data tracks
- **REST API**
  returns data in JSON format

More tools...

Click on Genomes



Our tools

- **Genome Browser**
  interactively visualize genomic data
- **COVID-19 Research**
  use the SARS-CoV-2 genome browser and explore coronavirus datasets
- **BLAT**
  rapidly align sequences to the genome
- **Table Browser**
  download data from the Genome Browser database
- **Variant Annotation Integrator**
  get functional effect predictions for variant calls
- **Data Integrator**
  combine data sources from the Genome Browser database
- **Genome Browser in a Box (GBiB)**
  run the Genome Browser on your laptop or server
- **In-Silico PCR**
  rapidly align PCR primer pairs to the genome
- **LiftOver**
  convert genome coordinates between assemblies
- **Track Hubs**
  import and view external data tracks
- **REST API**
  returns data in JSON format

More tools...

Available genomes and info

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute | UCSC **Genome Browser Gateway**

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Browse/Select Species

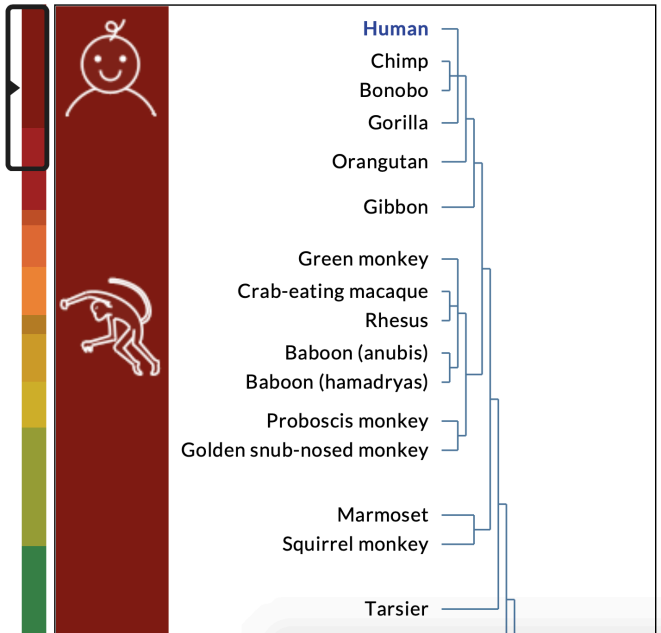
POPULAR SPECIES

Human | Mouse | Rat | Zebrafish | Fruitfly | Worm | Yeast

Enter species, common name or assembly ID

Can't find a genome assembly?

REPRESENTED SPECIES



Find Position

Human Assembly: Dec. 2013 (GRCh38/hg38)

GO

Position/Search Term: Enter position, gene symbol or search terms

Current position: chr1:1,300,932-3,320,931

Human Genome Browser - hg38 assembly | view sequences

UCSC Genome Browser assembly ID: hg38
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p13 (GCA_000001405.28)
Assembly date: Dec. 2013 initial release; Dec. 2017 patch release 13
Assembly accession: GCA_000001405.28
NCBI Genome ID: 51 (Homo sapiens (human))
NCBI Assembly ID: GCF_000001405.39 (GRCh38.p13, GCA_000001405.28)
BioProject ID: PRJNA31257

Homo sapiens
(Graphic courtesy of CBSE)

Search the assembly:

- **By position or search term:** Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. More information, including sample queries.
- **By gene name:** Type a gene name into the "search term" box, choose your gene

Click on the Genome Browser

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute — UCSC **Genome Browser Gateway**

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Browse/Select Species

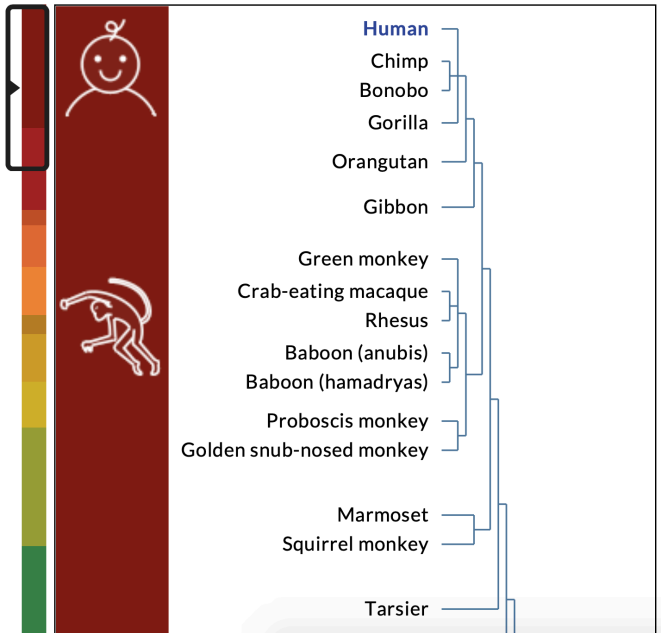
POPULAR SPECIES

Human | Mouse | Rat | Zebrafish | Fruitfly | Worm | Yeast

Enter species, common name or assembly ID

Can't find a genome assembly?

REPRESENTED SPECIES



Find Position

Human Assembly

Dec. 2013 (GRCh38/hg38)

GO

Position/Search Term

Enter position, gene symbol or search terms

Current position: chr1:1,300,932-3,320,931

Human Genome Browser - hg38 assembly — view sequences

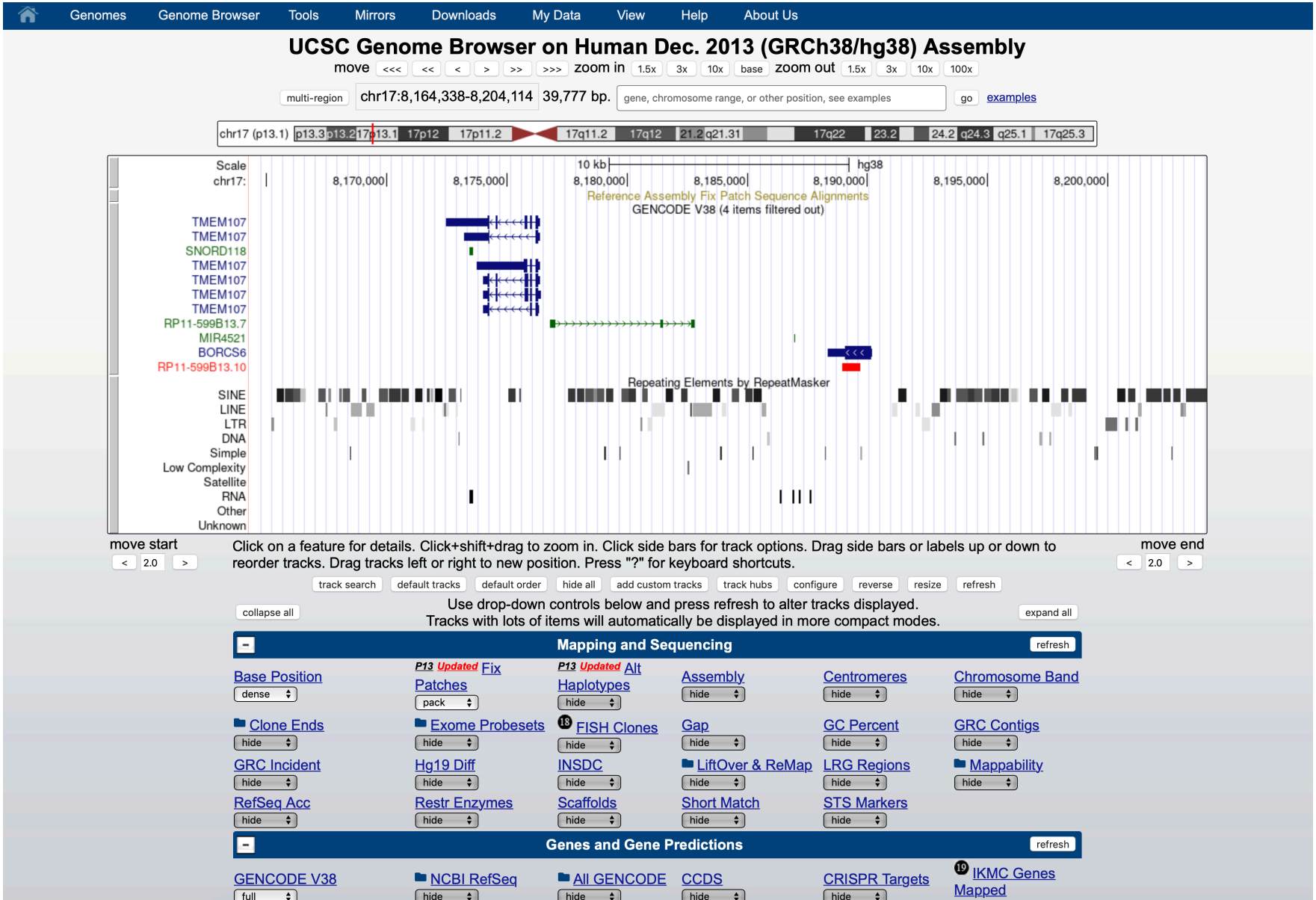
UCSC Genome Browser assembly ID: hg38
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p13 (GCA_000001405.28)
Assembly date: Dec. 2013 initial release; Dec. 2017 patch release 13
Assembly accession: GCA_000001405.28
NCBI Genome ID: 51 (Homo sapiens (human))
NCBI Assembly ID: GCF_000001405.39 (GRCh38.p13, GCA_000001405.28)
BioProject ID: PRJNA31257

Homo sapiens
(Graphic courtesy of CBSE)

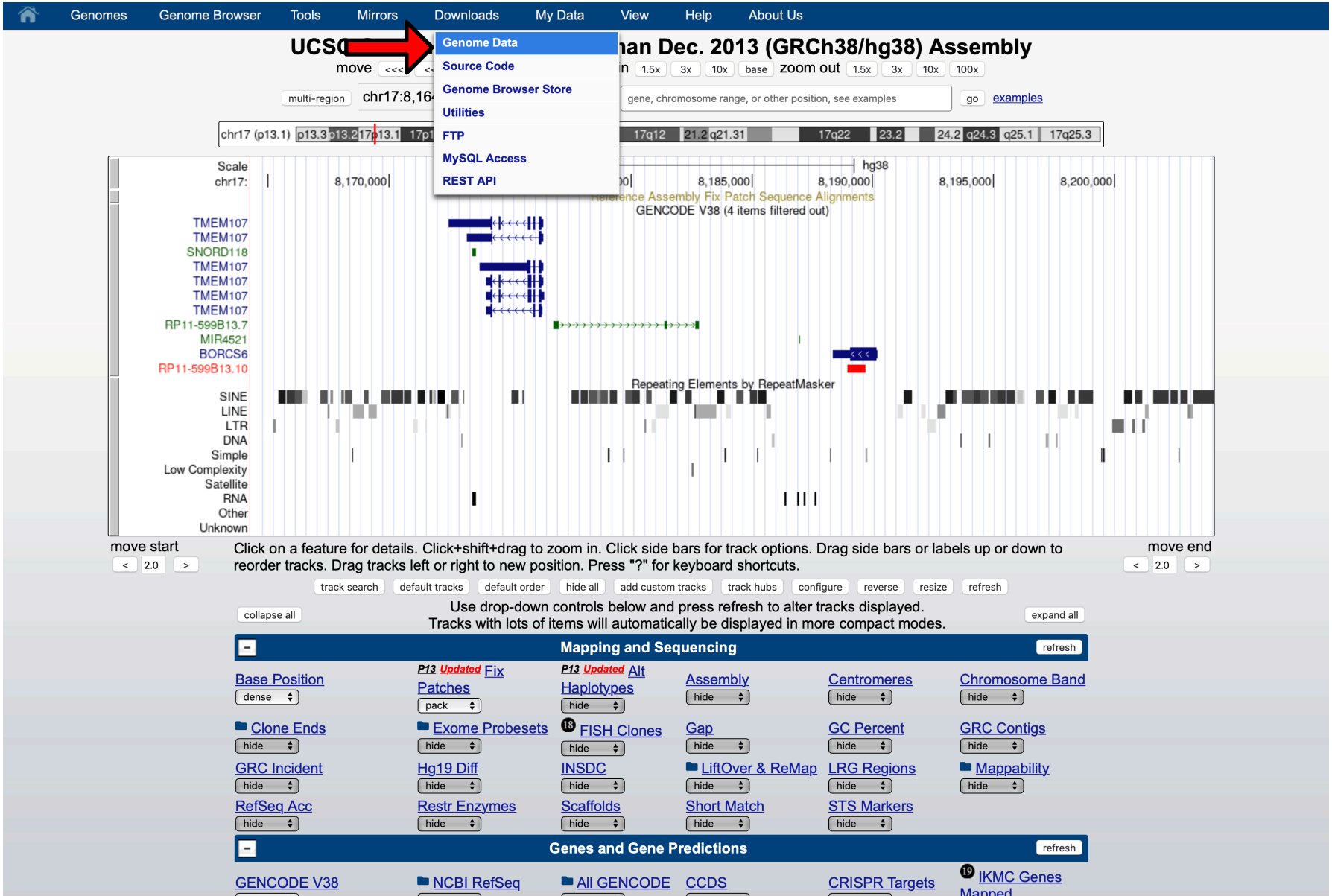
Search the assembly:

- **By position or search term:** Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. More information, including sample queries.
- **By gene name:** Type a gene name into the "search term" box, choose your gene

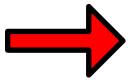
Browser view



Browser view



Downloading Genomes and source code for utilities and browser



UCSC | Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Sequence and Annotation Downloads

This page contains links to sequence and annotation downloads for the genome assemblies featured in the UCSC Genome Browser. Downloads are also available via the Genome Browser **FTP server**. For access to the most recent assembly of each genome, see the **current genomes** directory. To query and download data in JSON format, use our **JSON API**. To view descriptions of annotations, use the "describe table schema" button in the **Table Browser**. Previous versions of certain data are available from our **track archive**. For data hosted in **Public Hubs** the files exist on external sites, with **GenArk** (Genome Archive) Public Hub species found **here**.

All tables in the Genome Browser are freely usable for any purpose except as indicated in the README.txt files in the download directories. To view restrictions specific to a particular data set, click on the corresponding download link and review the README text. These data were contributed by many researchers, as listed on the Genome Browser **credits** page. Please acknowledge the contributor(s) of the data you use.

Human

Mouse

Mammals

Other vertebrates

Deuterostomes

Insects

Nematodes

Other genomes

Other downloads

Source and utilities downloads

The source for the Genome Browser, Blat, liftOver and other utilities is free for non-profit academic research and for personal use. For information on commercial licensing, see the **Genome Browser** and **Blat** licensing requirements. The source and executables for several of these products can be downloaded or purchased from our **online store**.

Mirroring the Genome Browser

Precompiled executable binaries are available for installing a local mirrored copy of the Genome Browser website on your web server, eliminating the need to compile the entire source tree.

- **Mirror instructions**
- **Build instructions**
- **Genome Browser hgcentral tables**
- **Genome Browser source code downloads**

If you encounter difficulties with slow download speeds, try using UDT Enabled Rsync (UDR), which improves the throughput of large data transfers over long distances. The 32-bit and 64-bit versions can be downloaded **here**.

Utilities

The **utilities directory** offers downloads of pre-compiled standalone binaries for:

- LiftOver (which may also be accessed via the **web version**). The over.chain liftOver conversion files are located in the individual assembly download sections.
- **Blat**
- Other command-line utilities

All our command line tools can be obtained as pre-built binaries from **our downloads server.** The following command can be used to copy all command line utilities into a directory with the correct permission bits set:

```
$ rsync -aP hgdownload.soe.ucsc.edu::genome/admin/exe/linux.x86_64/ ./
```

Note about 'permission denied' error when downloading with curl or wget:
In order for your computer to run a freshly downloaded utility, you will need to update the file system permissions to allow your operating system to run the program.
To make utilities usable, download a tool and turn on its 'executable' bit:

```
$ wget https://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/liftOver
```

```
$ chmod +x ./filePath/utility_name
```

Downloading Genomes

Human genome

Dec. 2013 (GRCh38/hg38)

- Genome sequence files and select annotations (2bit, GTF, GC-content, etc)
  - Standard genome sequence files and select annotations (2bit, GTF, GC-content, etc)
  - Analysis set sequence files (See: **What is the analysis set?**)
- Sequence data by chromosome
- Annotations
  - SQL table dump annotations
  - Fileserver (bigBed, maf, fa, etc) annotations Also see **Data Access**
- SNP-masked fasta files
- LiftOver files
- Pairwise alignments
- Multiple alignments
- Patches
- Data archive

Downloading Genomes and more

Top of page

```
Introduction
^^^^^^^^^^^^

The Dec. 2013  assembly of the human genome (GRCh38 Genome Reference Consortium
Human Reference 38), is called hg38 at UCSC. This directory contains the genome
as released by UCSC, selected annotation files and updates. The directory
"genes/" contains GTF/GFF files for the main gene transcript sets.

For more information about this assembly, see these NCBI resources:
    http://www.ncbi.nlm.nih.gov/genome/51
    http://www.ncbi.nlm.nih.gov/genome/assembly/883148
    http://www.ncbi.nlm.nih.gov/bioproject/31257

These files are used by the UCSC Genome Browser for display and analysis. If you
want to do analysis and show it later on the browser, it is usually easiest to
run your analysis on the UCSC hg38 file. For most users, this will be the file
"latest/hg38.fa.gz" in this directory. However, if you need a genome file for
alignment or variant calling, please read the section "Analysis set" below.

The sequences of the main chromosomes are identical to the genome files distributed
by NCBI and the EBI, but the sequence names are different. For example, the
name of chromosome 1 is called "chr1" at UCSC, "NC_000001.11" at NCBI, and "1"
at the EBI.  Also, the lowercasing in the files is not exactly identical, as
UCSC, NCBI and EBI run Repeatmasker with slightly different settings.

The NCBI accession of the UCSC hg38 genome is GCA_000001405.15. The version
that includes the updates for patch release 13 GRCh38.p13 has the NCBI
accession GCA_000001405.28.

Analysis set
^^^^^^^^^^^^

The GRCh38 assembly contains more than just the chromosome sequences, but also
a mitochondrial genome, unplaced sequences, centromeric sequences
and alternates. To better capture variation in the human genome across the world
it contains more copies of some loci than hg19. Some of these additions, like
the EBV genome, are mostly relevant for genomic analysis, i.e. alignment.
For an overview of the different types and reasons for the additions see
https://software.broadinstitute.org/gatk/documentation/article?id=11010

This means that if you want to use the genome sequence for alignment and
especially for variant calling, you should use the optimal genome file for your
aligner. The genome file can make a big difference, especially for variant
calling.  In most cases, the authors of your alignment program will provide
advice on which hg38 genome version to use and usually they recommend one of
the files in our analysisSet/ directory, like the GATK link above. These
special genome files sometimes remove the alternate sequences, sometimes they
add decoys or change single nucleotides towards the major allele, but they never
insert or delete sequences, so the annotation coordinates remain the same.

- for BWA see also https://lh3.github.io/2017/11/13/which-human-reference-genome-to-use
- for Novoalign see its manual at http://www.novocraft.com/userfiles/file/Novocraft.pdf
- For Bowtie, see the different versions of the human genome that the Bowtie authors
  provide: http://bowtie-bio.sourceforge.net/index.shtml

Also see analysisSet/README.txt for further details

Patches
^^^^^^^

Like hg19, hg38 has been updated with patches since its release in 2013. GRC
patch releases do not change any previously existing sequences; they simply add
small, new sequences for fix patches or alternate haplotypes that correspond to
specific regions of the main chromosome sequences (see below). For most users,
the patches are unlikely to make a difference and may complicate the analysis
as they introduce more duplication. If you want a version of the genome
without these complexities, look at the analysisSet/ subdirectory.

The initial/ subdirectory contains files for the initial release of GRCh38,
which includes the original alternate sequences (261) and no fix sequences.

The p11/ subdirectory contains files for GRCh38.p11 (patch release 11).

The p12/ subdirectory contains files for GRCh38.p12 (patch release 12).

The p13/ subdirectory contains files for GRCh38.p13 (patch release 13).

The "latest/" symbolic link points to the subdirectory for the most recent
patch version.
```

Bottom of page

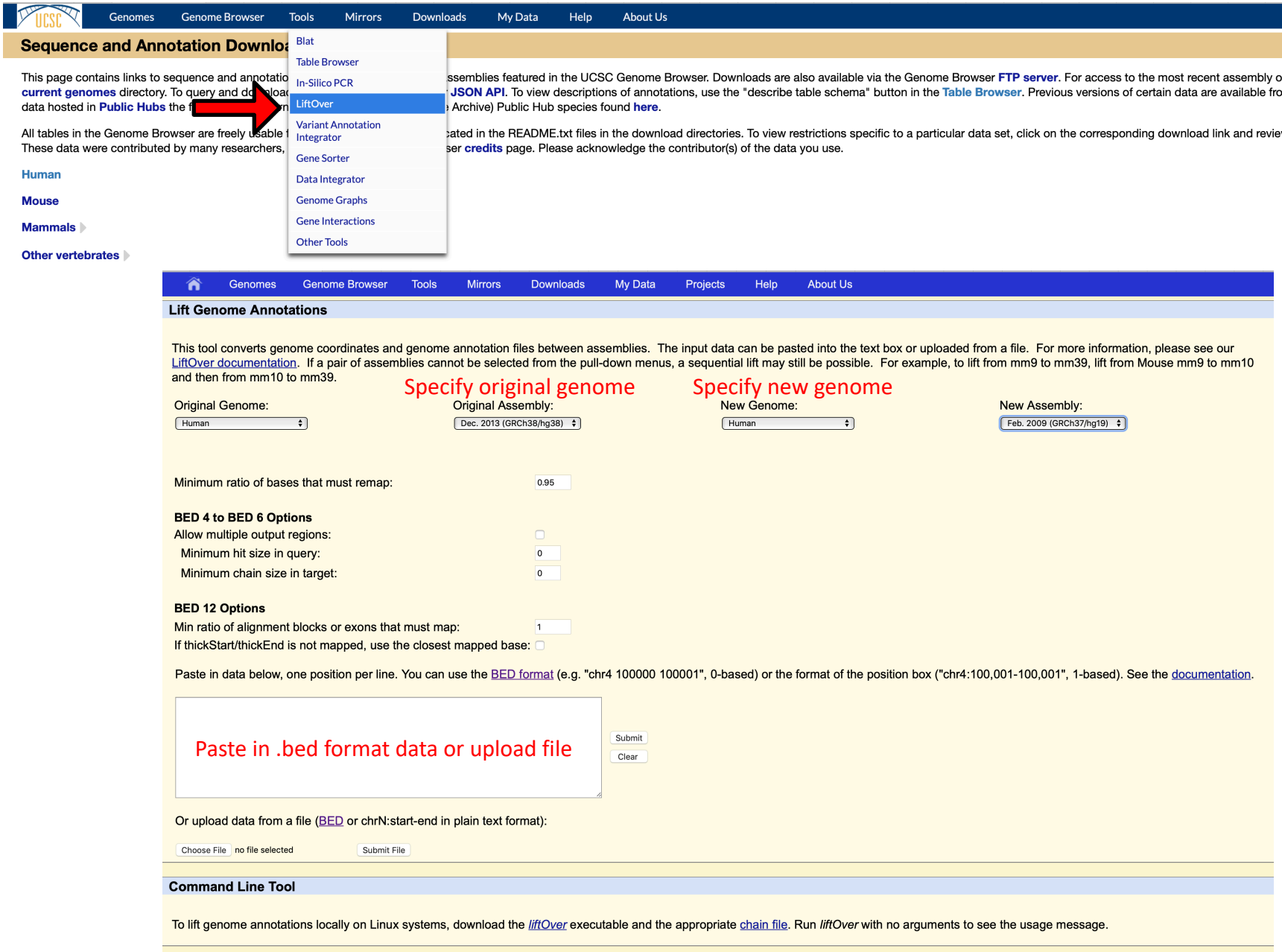
Name	Last modified	Size	Description
Parent Directory		-	
analysisSet/	2021-10-07 16:48	-	
est.fa.gz	2020-09-09 11:27	1.5G	
est.fa.gz.md5	2020-09-09 11:27	44	
genes/	2021-12-09 03:03	-	
hg38.2bit	2015-04-30 16:16	797M	
hg38.agp.gz	2014-01-15 20:55	842K	
hg38.chrom.sizes	2013-12-24 21:06	11K	
hg38.chromAlias.txt	2021-10-06 13:44	27K	
hg38.chromFa.tar.gz	2014-01-23 17:18	938M	
hg38.chromFaMasked.tar.gz	2014-01-23 17:10	487M	
hg38.fa.align.gz	2014-01-08 23:43	2.4G	
hg38.fa.gz	2014-01-15 21:14	938M	
hg38.fa.masked.gz	2014-01-15 21:24	487M	
hg38.fa.out.gz	2014-01-15 20:56	172M	
hg38.gc5Base.bw	2013-12-24 21:28	1.6G	
hg38.gc5Base.wib	2019-01-17 14:50	591M	
hg38.gc5Base.wig.gz	2019-01-17 14:50	11M	
hg38.gc5Base.wigVarStep.gz	2013-12-24 21:14	1.5G	
hg38.trf.bed.gz	2014-01-15 20:56	7.9M	
initial/	2021-10-07 16:50	-	
latest/	2021-10-07 16:57	-	
md5sum.txt	2021-10-06 14:07	667	
mrna.fa.gz	2020-09-09 11:08	372M	
mrna.fa.gz.md5	2020-09-09 11:08	45	
p11/	2021-10-07 17:01	-	
p12/	2021-10-07 16:51	-	
p13/	2021-10-07 16:53	-	
refMrna.fa.gz	2020-09-09 11:28	89M	
refMrna.fa.gz.md5	2020-09-09 11:28	48	
upstream1000.fa.gz	2020-09-09 11:29	11M	
upstream1000.fa.gz.md5	2020-09-09 11:29	53	
upstream2000.fa.gz	2020-09-09 11:29	20M	
upstream2000.fa.gz.md5	2020-09-09 11:29	53	
upstream5000.fa.gz	2020-09-09 11:30	52M	
upstream5000.fa.gz.md5	2020-09-09 11:30	53	
xenoMrna.fa.gz	2020-09-09 11:19	6.7G	
xenoMrna.fa.gz.md5	2020-09-09 11:19	49	
xenoRefMrna.fa.gz	2020-09-09 11:27	255M	
xenoRefMrna.fa.gz.md5	2020-09-09 11:27	52	

FASTA formatted genome

We use this for input in bedtools

Descriptions of file content are in the middle

Converting genome coordinates



Downloading annotation tracks: gene lists

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute — UCSC Genome Browser

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

- Blat
- In-Silico PCR
- Table Browser
- LiftOver
- Gene Sorter
- Variant Annotation Integrator

...ools

...ome Browser
...ctively visualize genomic data

...ID-19 Research
...e SARS-CoV-2 genome browser and explore coronavirus datasets

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Table Browser

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

Select dataset

clade: Mammal | genome: Human | assembly: Dec. 2013 (GRCh38/hg38)
group: Genes and Gene Predictions | track: GENCODE V38 → Track type
table: knownGene | describe table schema → Sub-table

Define region of interest

region: genome / position chr17:8,164,338-8,204,114 | lookup | define regions
identifiers (names/accessions): paste list | upload list

Optional: Subset, combine, compare with another track → Choose fields or columns from table

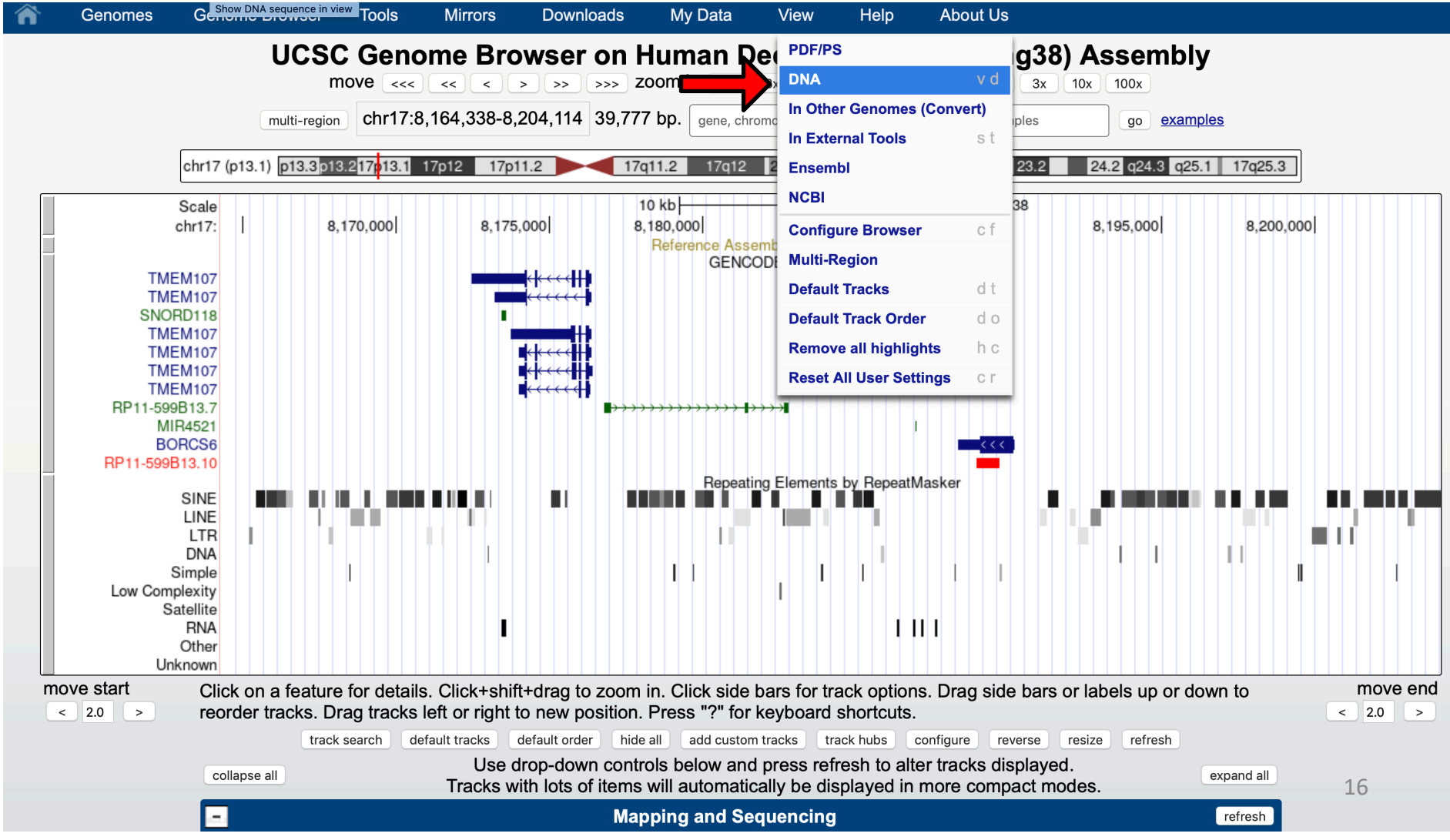
filter: create
intersection: create

Retrieve and display data

output format: BED - browser extensible data | Send output to Galaxy | GREAT
output filename: GENCODE_V38_intersect_XXX.bed (leave blank to keep output in browser) → Nameoutput file
file type returned: plain text / gzip compressed

get output | summary/statistics

Downloading DNA sequence from browser



Downloading DNA from browser

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Get DNA in Window (hg38/Human)

Get DNA for

Position chr17:8,164,338-8,204,114

Note: This page retrieves genomic DNA for a single region. If you would prefer to get DNA for many items in a particular track, or get DNA with formatting options based on gene structure (introns, exons, UTRs, etc.), try using the Table Browser with the "sequence" output format. You can also use the REST API with the **/getData/sequence** endpoint function to extract sequence data with coordinates.

Sequence Retrieval Region Options:

Add 0 extra bases upstream (5') and 0 extra downstream (3')

Options to get flanking DNA as well

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

Sequence Formatting Options:

- (•) All upper case.
- ( ) All lower case.
- [ ] Mask repeats: (•) to lower case ( ) to N
- [ ] Reverse complement (get '-' strand sequence)

Options for case and masking of repeats

get DNA | extended case/color options

Note: The "Mask repeats" option applies only to "get DNA", not to "extended case/color options".

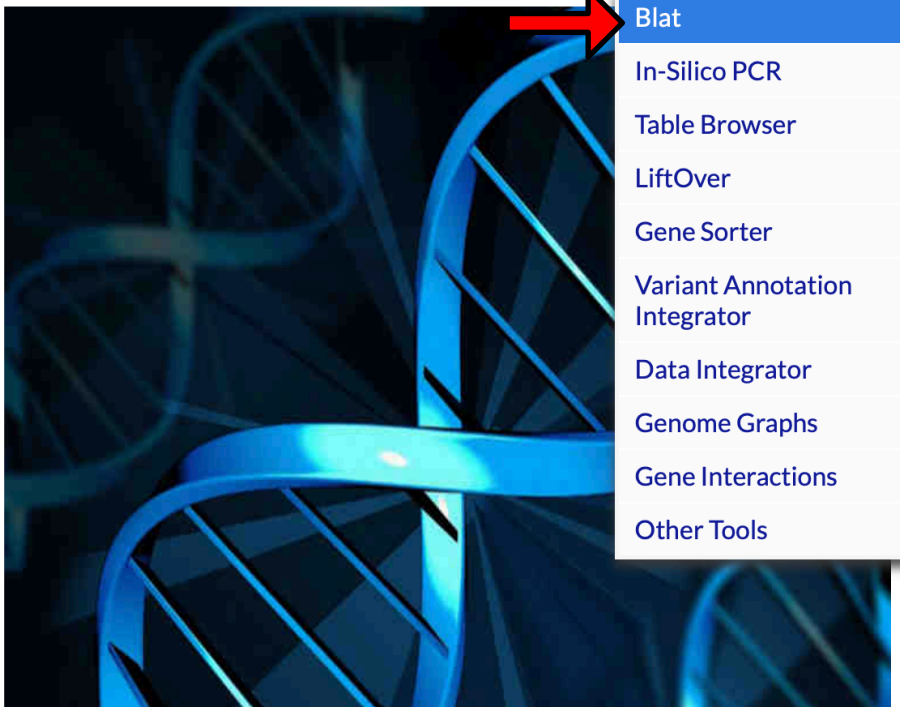
Fasta output

```
>hg38_dna range=chr17:8164338-8204114 5'pad=0 3'pad=0 strand=+ repeatMasking=none
GGCCTTGATCTCTAAGACAAACAGGAACCAGCTTCCTCCCTCCCCCACCC
TGGCTTCCCAGGGCCTCCGGGTGTGAGATCCTCCCCCACTGCAGTGCCCC
ACCCGCTCCCCACAGAAGCCCGGAGAGTGGCTCTGTCACCAGAGGTGTCA
TTTCCCAGCTGTCTGTGGGAGGTGAGTGAGCAGGGAATGTGTGTGCTGGG
TGTGGGAACTCAGCCCAATCTAAGAGAAGATACTCTTGGCTTCCTCCCCC
TCAGAGGAGCAGCCGCGTCCCTGGTCCTTGTGCACTGACAGTCCCGGGTG
GCTCAGCCTATGTGATGAGAAAGAAAAGAGCAGCTTCCACGGGGGTCTCA
GATCATGCTGTGAGATGCTGGCTCCTGCTTTTTGGTCTTCAAATGATTCC
TCTTTCTAGTCAGAGAACAAAGATGCCAAATTCCAGCCTTTTAAGTTCAA
AACCCTCAGTGTCTGTCAGGGTATAAATAAGACATGGAGGATTGGAAGAG
GCAGGCACAGGAAGCCCTTCCGCCCCCAGTTGGACACCCTGCACTAACTG
AGGGGAGCCTAGTGCAGAGCCAGACAAGACATTTGATGCTGGTTTGGGGG
TGATGTGGAAGGGACAGAGCCCCAGATACACGGTCAAGGCCAGGGGTCGT
GCTATTGCTATTGGATCGTGGGGACGGCAGCAGCAGAGTAGAAATGAGGG
CATGTGTGACCCACGGAGATGAAGGGCCCAGGAGGATTCATCTCAGAAAG
AGGACTCTCATCCAGCAGCCCCTTCCCACCCCACCCCACCTAGGCTGTGG
CTCAGAGGAACTCCCCAAATTTAGATCAATGGTGGAGGTGGGAGAAGGAG
TGGGATGACGGAGTTCACCTGGAAGTGACTGAGAGAATGTTTTCTGCAGG
CAGGTAGAATGGAGACCCAAATAGAGGTAAGTTGAACAATAGAAAAATAA
TGTATTTGCATAACTCAGATTTGGCTTCTACATGTACTGGCTTCGGGAAT
TGGCCAATGATTTTCTAGGGCACACCCCCAAGACTCATTCAAAGGATGTA
GGACTCAAGGTGATGGTGATGTAGGCAACAGAGACCACACTTTCTTTTTC
TTTTTTTTTTCTTTTTTTTTCTTTTTTGAGACAGAGTCTCTGTCTGTCAC
CCAGGCTGGAGTGCAATGGCGCGATCTCTGCTCACTGCAAACTCCGCCTC
CCAGGTTCACGCCATTCTCCTCCCTCAGCCTCCCGAGTAGCTGGGACTAC
AGGCGCCCACCACCCGGCCCGGCTAATTTTTTGTATTTTTAGTAGAGACA
GGGTTTCACCGTGTTAGCCAGGGTGGTCTTGATCTCCTGACCCCGTGATC
```

Fast DNA search: BLAT

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute | UCSC **Genome Browser**

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us



- Blat
- In-Silico PCR
- Table Browser
- LiftOver
- Gene Sorter
- Variant Annotation Integrator
- Data Integrator
- Genome Graphs
- Gene Interactions
- Other Tools

ools

ome Browser
ctively visualize genomic data

ID-19 Research
e SARS-CoV-2 genome browser and explore coronavirus datasets

T
y align sequences to the genome

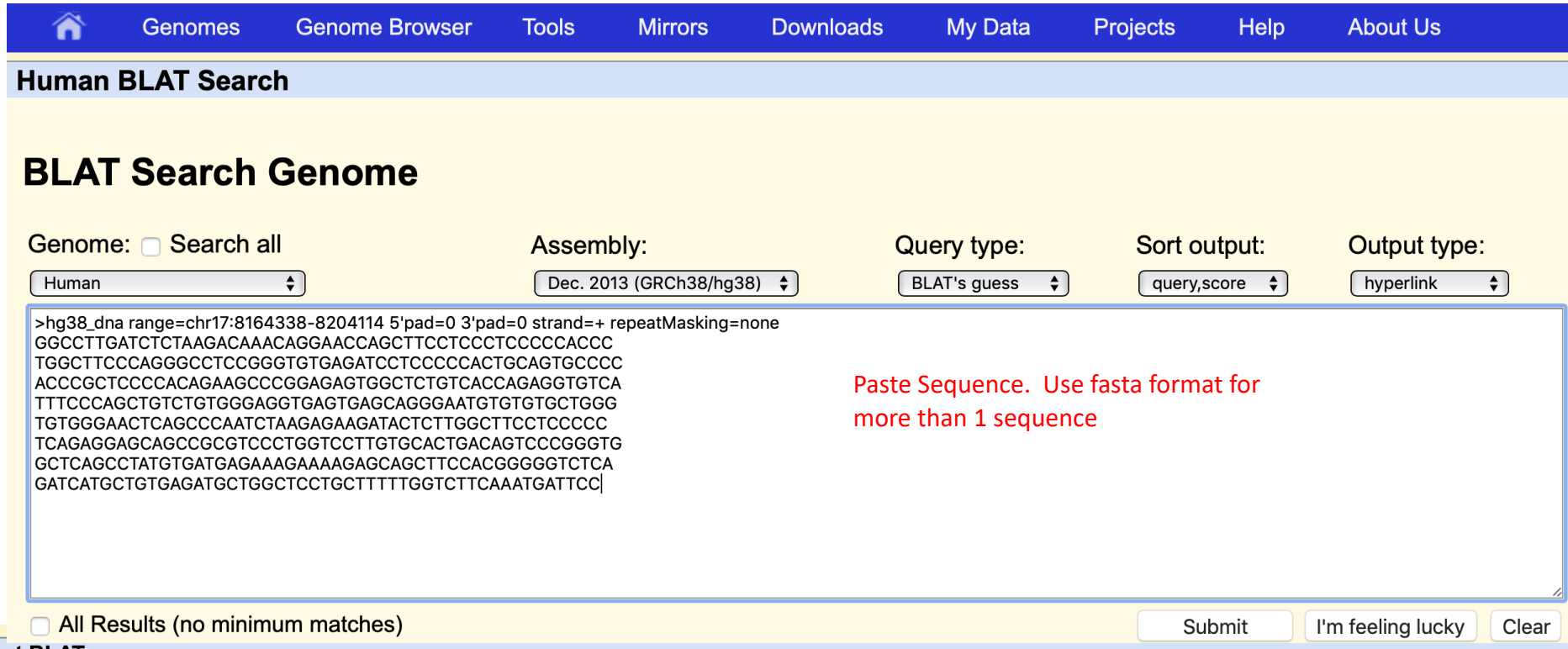
e Browser
load data from the Genome Browser database

ant Annotation Integrator
nctional effect predictions for variant calls

Integrator
combine data sources from the Genome Browser database

- **Genome Browser in a Box (GBiB)**
  run the Genome Browser on your laptop or server

Fast DNA search with BLAT



About BLAT

BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 25 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 20 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein BLAT on land vertebrates.

BLAT is not BLAST. DNA BLAT works by keeping an index of the entire genome in memory. The index consists of all overlapping 11-mers stepping by 5 except for those heavily involved in repeats. The index takes up about 2 gigabytes of RAM. RAM can be further reduced to less than 1 GB by increasing step size to 11. The genome itself is not kept in memory, allowing BLAT to deliver high performance on a reasonably priced Linux box. The index is used to find areas of probable homology, which are then loaded into memory for a detailed alignment. Protein BLAT works in a similar manner, except with 4-mers rather than 11-mers. The protein index takes a little more than 2 gigabytes.

BLAT was written by Jim Kent. Like most of Jim's software, interactive use on this web server is free to all. Sources and executables to run batch jobs on your own server are available free for academic, personal, and non-profit purposes. Non-exclusive commercial licenses are also available. See the Kent Informatics website for details.

For more information on the graphical version of BLAT, click the Help button on the top menu bar or see the Genome Browser FAQ.

Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64. PMID: 11932250

Fast DNA search with BLAT

BLAT Search Results

Go back to chr17:8164338-8204114 in the Genome Browser

Custom track name: blat hg38_dna

Custom track description: blat on hg38_dna

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	hg38_dna	400	1	400	400	100.0%	chr17	+	8164338	8164737	400
browser details	hg38_dna	25	29	54	400	100.0%	chr19	-	10310700	10310726	27
browser details	hg38_dna	25	349	379	400	88.9%	chr1	+	247443829	247443858	30
browser details	hg38_dna	25	15	49	400	89.3%	chr1	+	204244803	204244836	34
browser details	hg38_dna	23	34	58	400	96.0%	chr14	-	20262748	20262772	25
browser details	hg38_dna	23	321	344	400	100.0%	chr1	-	59157007	59157032	26
browser details	hg38_dna	23	42	65	400	100.0%	chr14	+	72549910	72549939	30
browser details	hg38_dna	22	34	58	400	96.0%	chr6	-	166516416	166516441	26
browser details	hg38_dna	22	32	56	400	95.9%	chr13	+	51135725	51135753	29
browser details	hg38_dna	21	32	52	400	100.0%	chr10	-	5966101	5966121	21
browser details	hg38_dna	20	26	47	400	95.5%	chr18	-	44760280	44760301	22
browser details	hg38_dna	20	367	386	400	100.0%	chr15	+	93348003	93348022	20
browser details	hg38_dna	20	360	381	400	95.5%	chr11	+			
browser details	hg38_dna	20	32	51	400	100.0%	chr1	+			
browser details	hg38_dna	20	196	215	400	100.0%	chr1	+			
browser details	hg38_dna	20	32	51	400	100.0%	chr1	+			

Help

Alignment of hg38_dna and chr17:8164338-8164737

Click on links in the frame to the left to navigate through the alignment. Matching bases in cDNA and genomic sequences are colored blue and capitalized. Light mark the boundaries of gaps in either sequence (often splice sites).

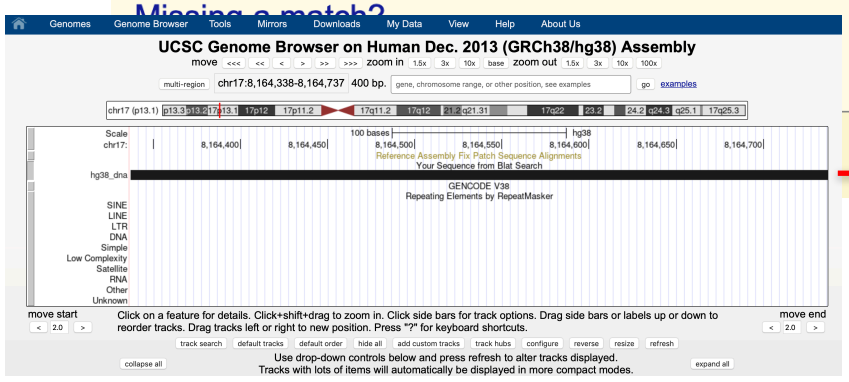
cDNA hg38_dna

```
GGCCTTGATC TCTAAGACAA ACAGGAACCA GCTTCCTCCC TCCCCCACCC   50
TGGCTTCCCA GGGCCTCCGG GTGTGAGATC CTCCCCCACT GCAGTGCCCC  100
ACCCGCTCCC CACAGAAGCC CGGAGAGTGG CTCTGTCACC AGAGGTGTCA  150
TTTCCCAGCT GTCTGTGGGA GGTGAGTGAG CAGGGAATGT GTGTGCTGGG  200
TGTGGGAACT CAGCCCAATC TAAGAGAAGA TACTCTTGGC TTCCTCCCCC  250
TCAGAGGAGC AGCCGCGTCC CTGGTCCTTG TGCACTGACA GTCCCGGGTG  300
GCTCAGCCTA TGTGATGAGA AAGAAAAGAG CAGCTTCCAC GGGGGTCTCA  350
GATCATGCTG TGAGATGCTG GCTCCTGCTT TTTGGTCTTC AAATGATTCC  400
```

Genomic chr17 :

```
caaagcaagc ctcaaggtgtg caggaggcat tttaggcgag gccctttggc  8164287
cctgggctgg gggtgctcac tacatccctg gaaggtgcag ggaggctgat  8164337
GGCCTTGATC TCTAAGACAA ACAGGAACCA GCTTCCTCCC TCCCCCACCC  8164387
TGGCTTCCCA GGGCCTCCGG GTGTGAGATC CTCCCCCACT GCAGTGCCCC  8164437
ACCCGCTCCC CACAGAAGCC CGGAGAGTGG CTCTGTCACC AGAGGTGTCA  8164487
TTTCCCAGCT GTCTGTGGGA GGTGAGTGAG CAGGGAATGT GTGTGCTGGG  8164537
TGTGGGAACT CAGCCCAATC TAAGAGAAGA TACTCTTGGC TTCCTCCCCC  8164587
TCAGAGGAGC AGCCGCGTCC CTGGTCCTTG TGCACTGACA GTCCCGGGTG  8164637
GCTCAGCCTA TGTGATGAGA AAGAAAAGAG CAGCTTCCAC GGGGGTCTCA  8164687
GATCATGCTG TGAGATGCTG GCTCCTGCTT TTTGGTCTTC AAATGATTCC  8164737
tctttctagt cagagaacaa agatgccaaa ttccagcctt ttaagttcaa  8164787
aaccctcagt gtctgtcagg gtataaataa gacatggagg attggaagag  8164837
```

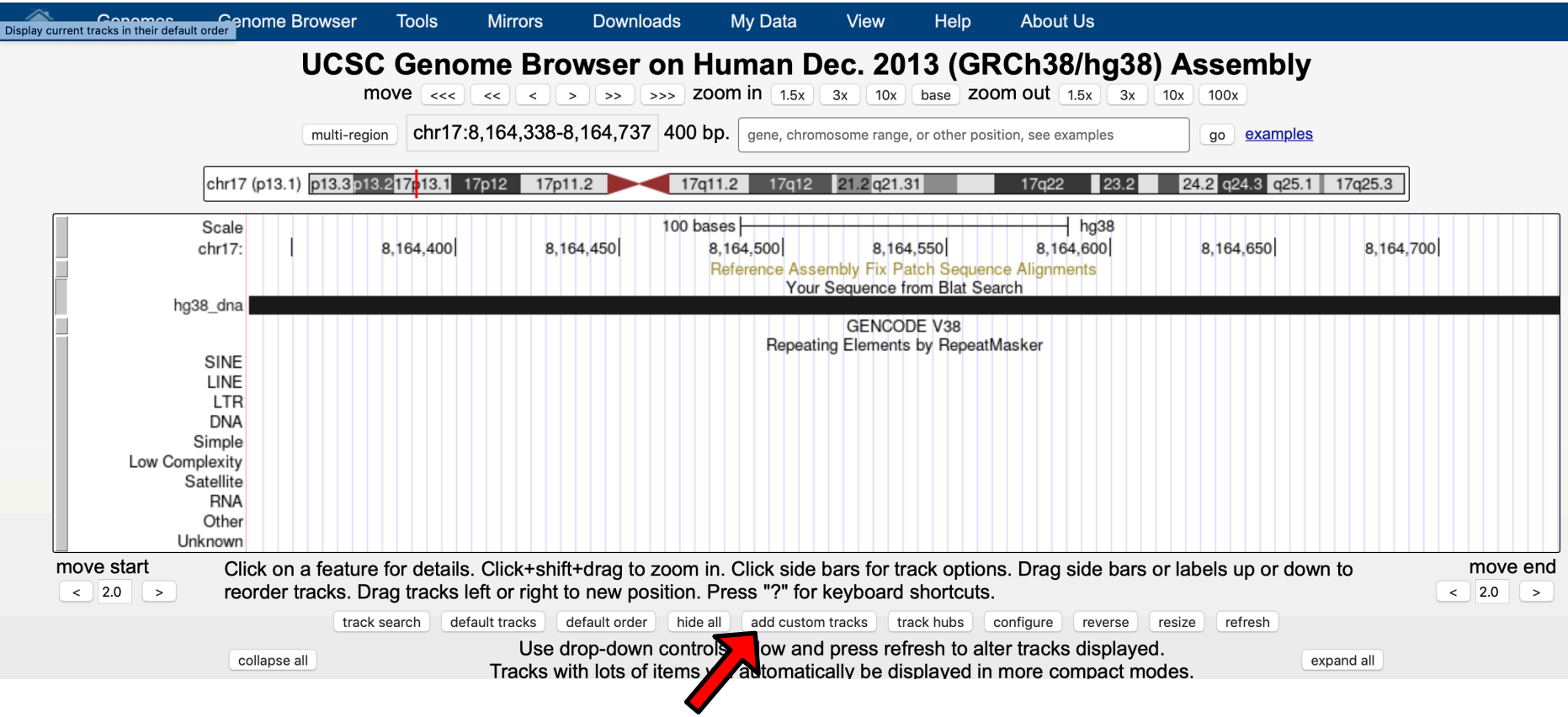
Side-by-Side Alignment



"Your Sequence From Blat Search"

Matched region shown as track in browser

Custom tracks



Custom tracks: uploading

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Add Custom Tracks

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

Display your own data as custom annotation tracks in the browser. Data must be formatted in bigBed, bigBarChart, bigChain, bigGenePred, bigInteract, bigLolly, bigMaf, bigPsl, bigWig, BAM, barChart, VCF, BED, BED detail, bedGraph, broadPeak, CRAM, GFF, GTF, hic, interact, MAF, narrowPeak, Personal Genome SNP, PSL, or WIG formats.

Select File

- You can paste just a URL to the file, without a "track" line, for bigBed, bigWig, bigGenePred, BAM and VCF.
- To configure the display, set track and browser line attributes as described in the User's Guide.

Examples are here. If you do not have web-accessible data storage available, please see the Hosting section of the Track Hub Help documentation.

Please note a much more efficient way to load data is to use Track Hubs, which are loaded from the Track Hubs Portal found in the menu under My Data.

Paste URLs or data: Or upload: Choose File no file selected Submit

upload

Clear

Optional track documentation: Or upload: Choose File no file selected

Clear

Click here for an HTML document template that may be used for Genome Browser track descriptions.

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Manage Custom Tracks

genome: Human **assembly:** Dec. 2013 (GRCh38/hg38) [hg38]

Name	Description	Type	Doc	Items	Pos	delete
ATF1_ChIP_10mil	ATF1_ChIP_10mil_bedGraph	bedGraph		8443869	chr11:	
ATF1_ChIP_bed	ATF1_ChIP_bed	bed		7252908	chr1:	

view in Genome Browser go

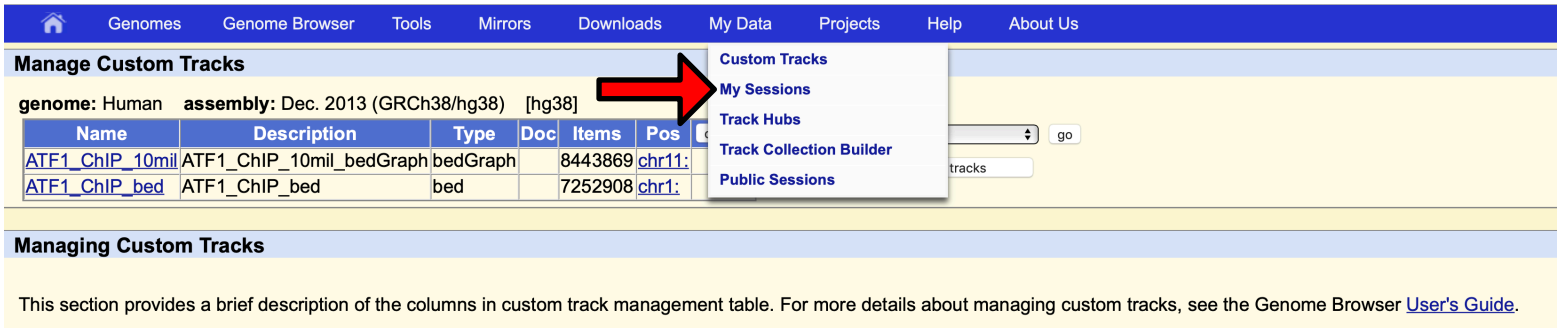
add custom tracks

Managing Custom Tracks

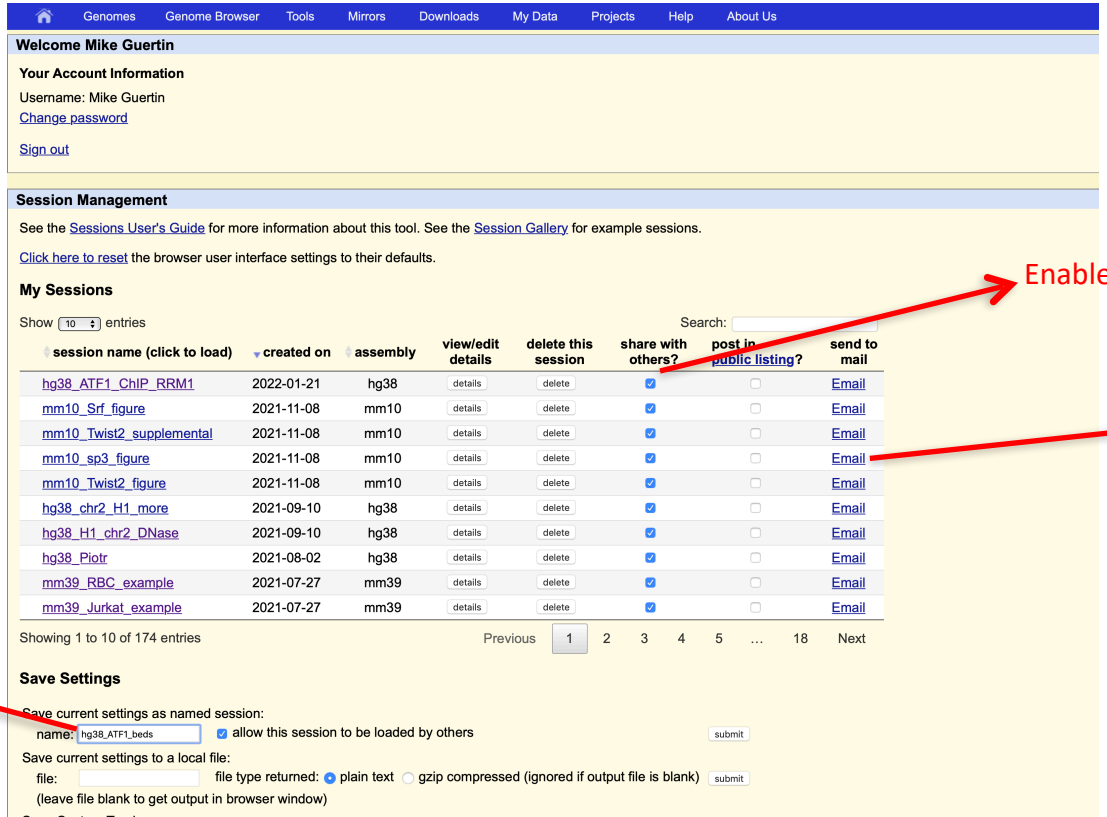
This section provides a brief description of the columns in custom track management table. For more details about managing custom tracks, see the Genome Browser User's Guide.

- **Name** - a hyperlink to the update page where you can edit your track data.
- **Description** - the value of the "description" attribute from the track line, if present. If no description is included in the input file, this field contains the track name.
- **Type** - the track type, determined by the Browser based on the format of the data.
- **Doc** - displays "Y" (Yes) if a description page has been uploaded for the track; otherwise the field is blank.
- **Items** - the number of data items in the custom track file. An item count is not displayed for tracks lacking individual items (e.g. wiggle format data).
- **Pos** - the default chromosomal position defined by the track file in either the browser line "position" attribute or the first data line. Clicking this link opens the Genome Browser or Table Browser at the specified position (note: only the chromosome name is shown in this column). The Pos column remains blank if the track lacks individual items (e.g. wiggle format data) and the browser line "position" attribute hasn't been set.

Custom tracks: saving and sharing



Sign in or register

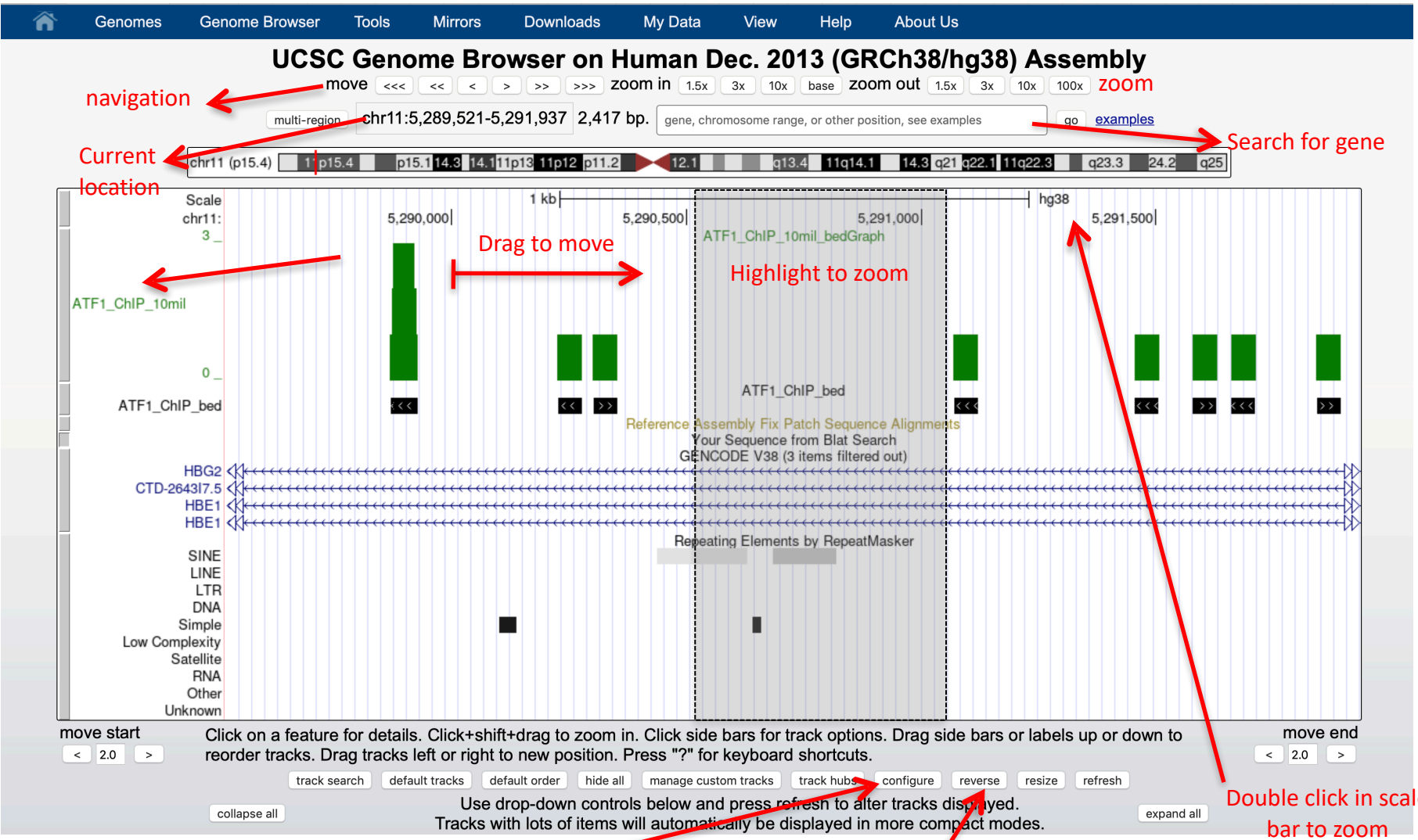


Enable sharing

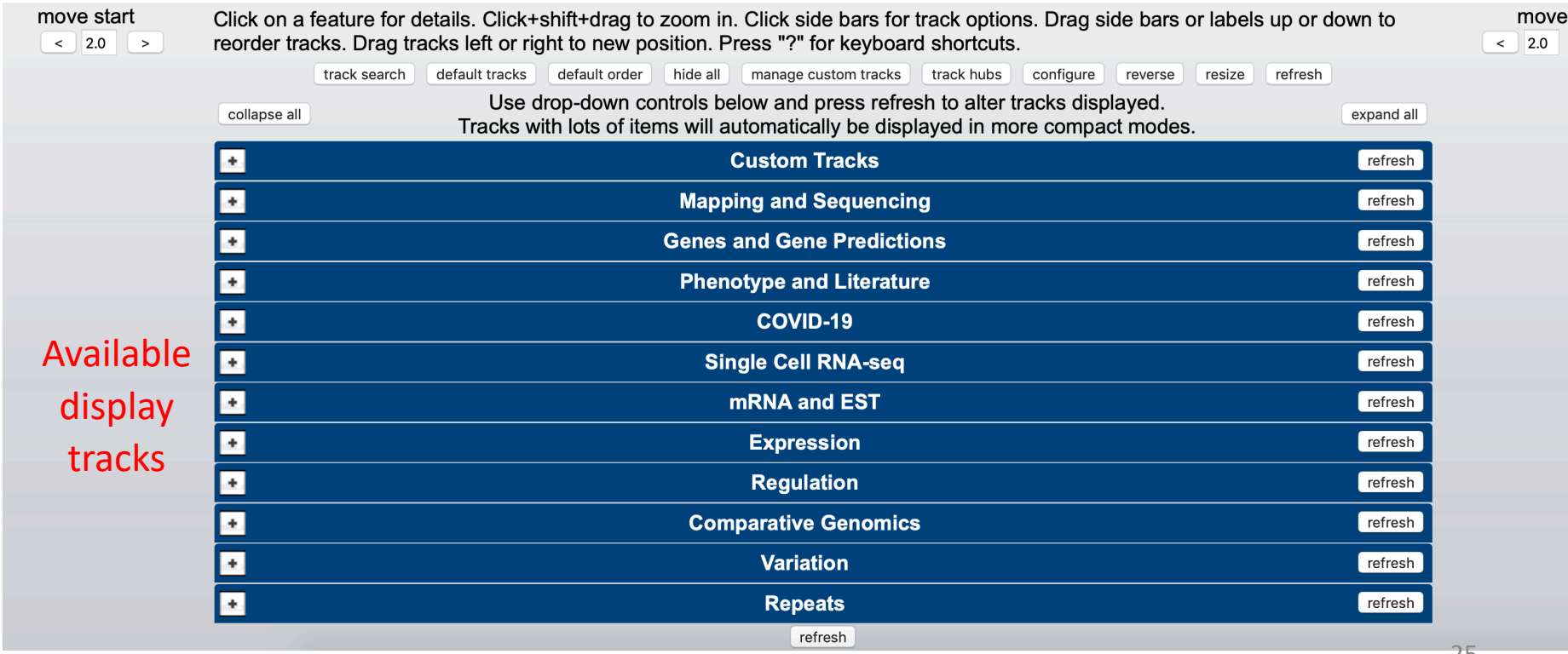
Create sharing link in email

Save current as:

Cruising the genome: Displays and Tracks

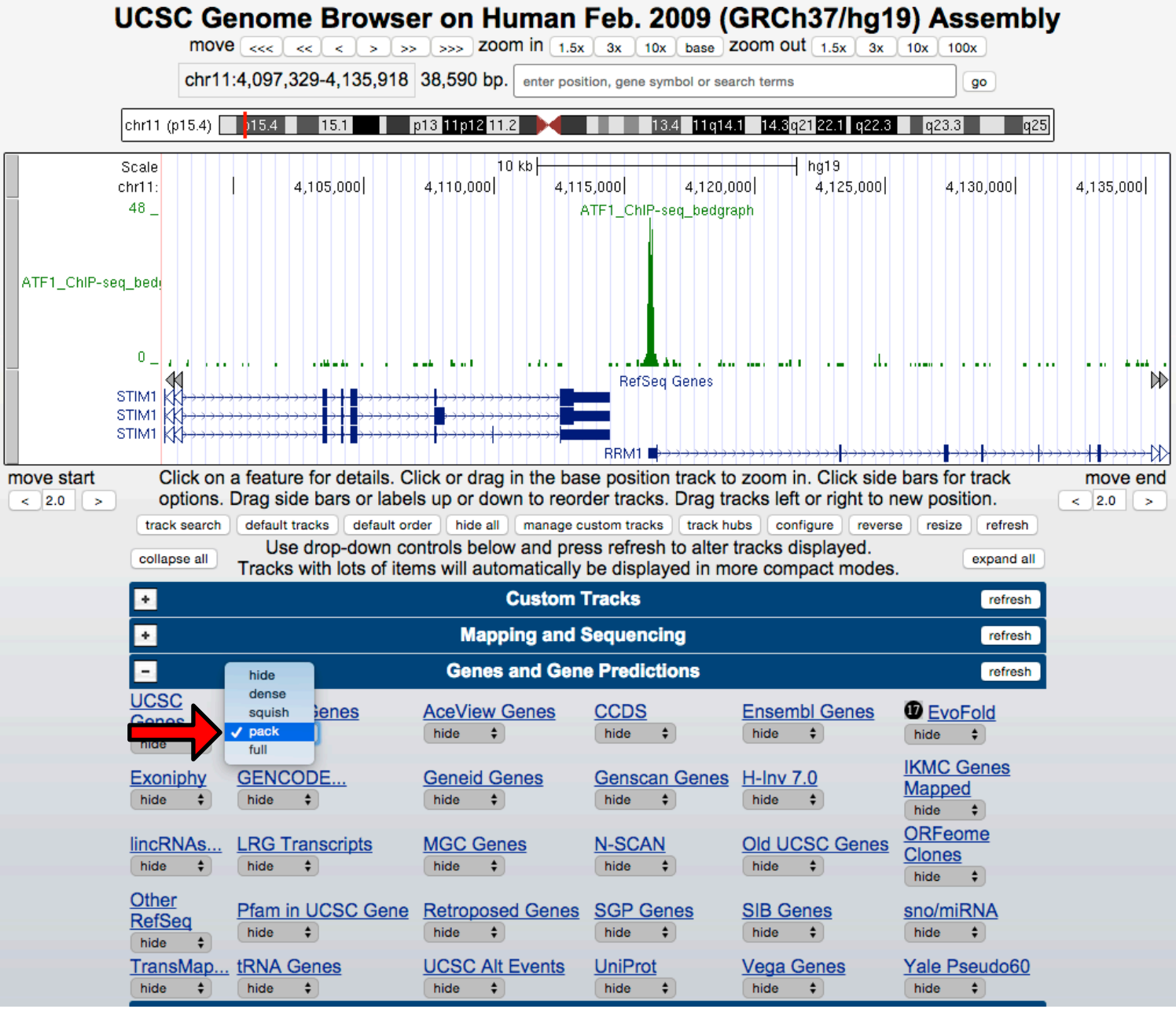


Track Display:

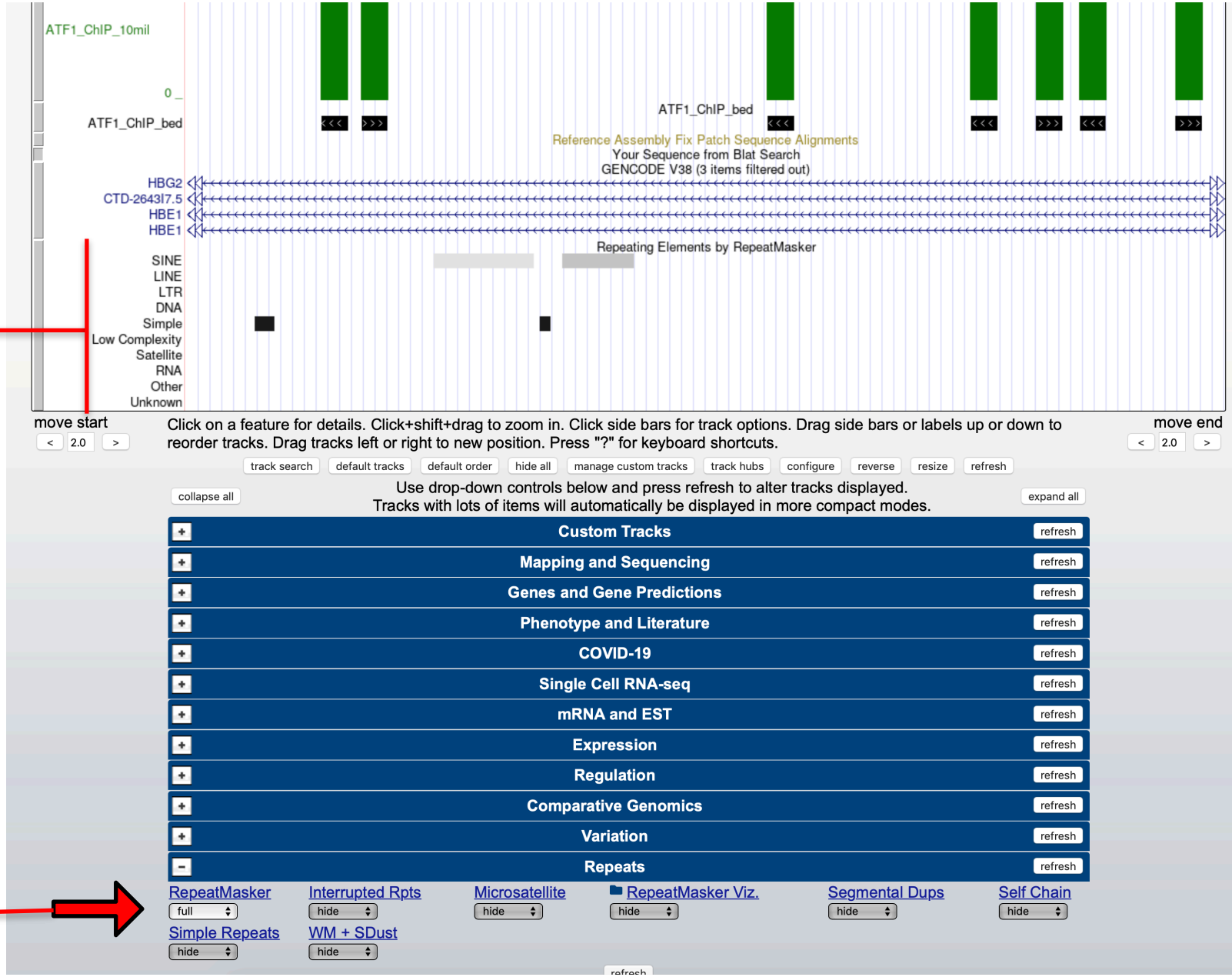


Available display tracks

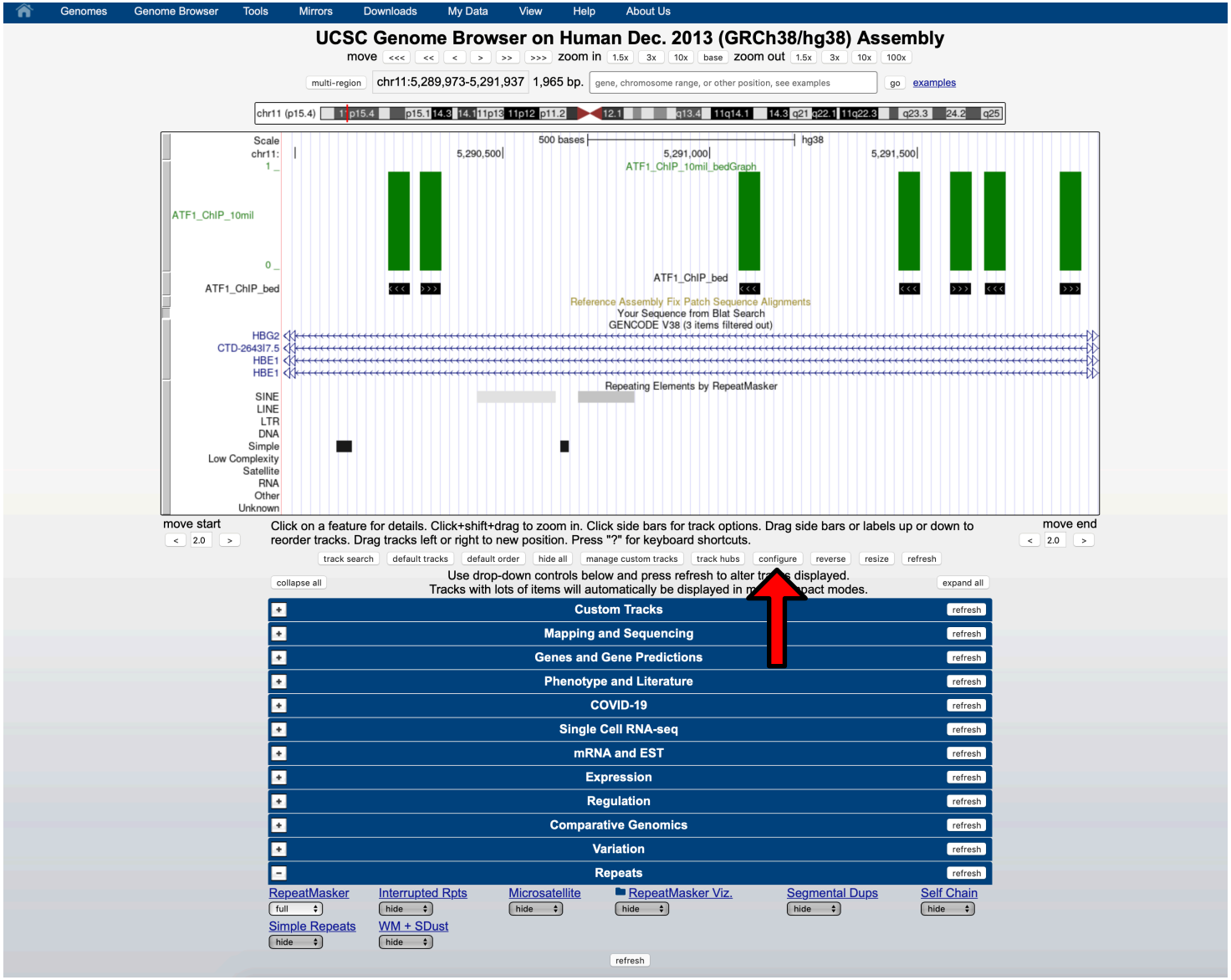
Track Display:



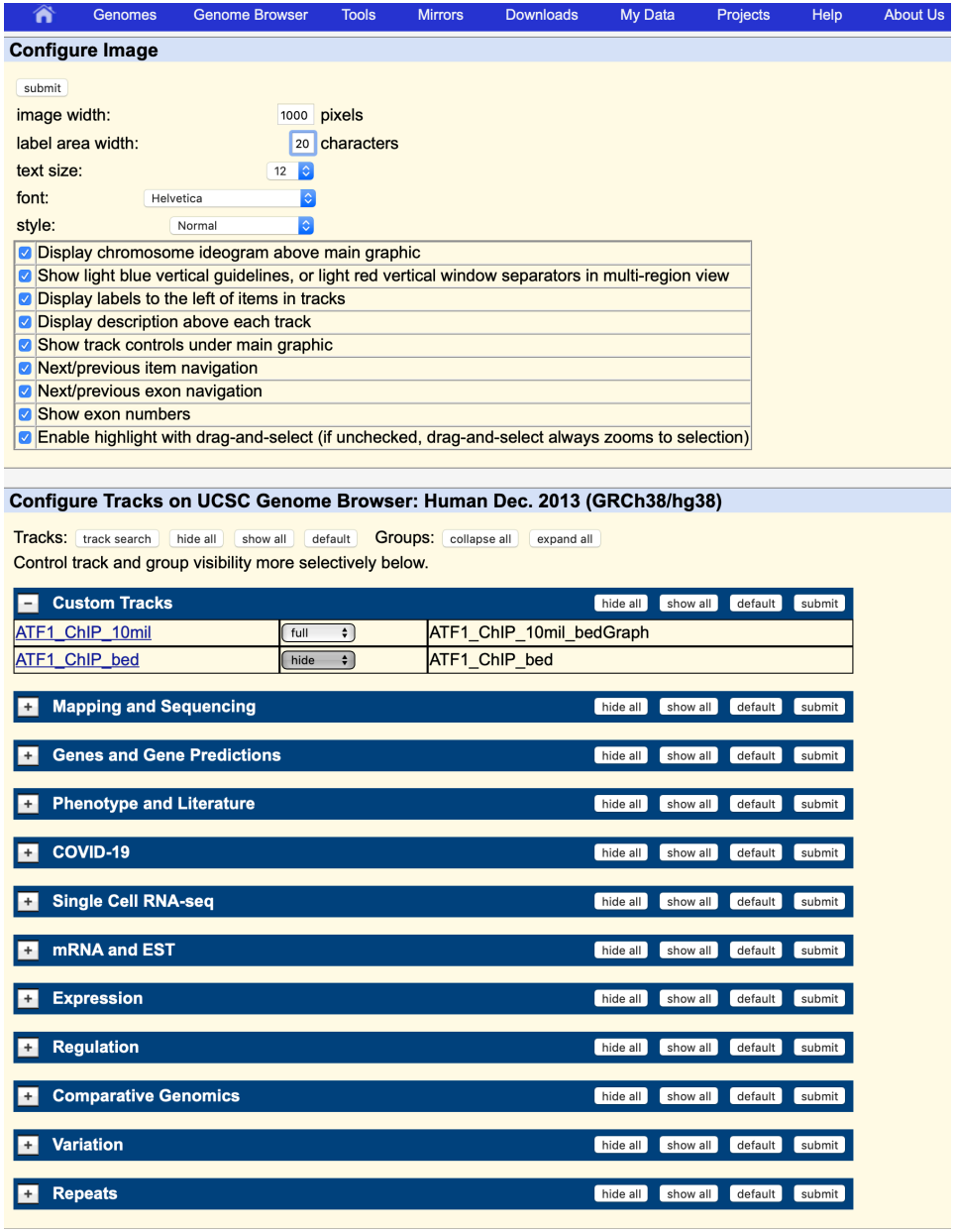
Track Display:



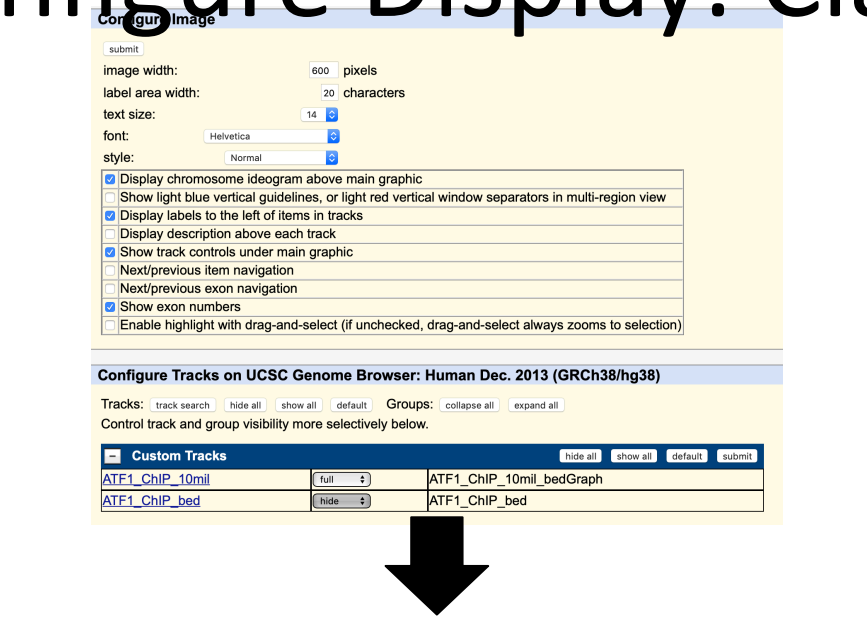
Configure Display

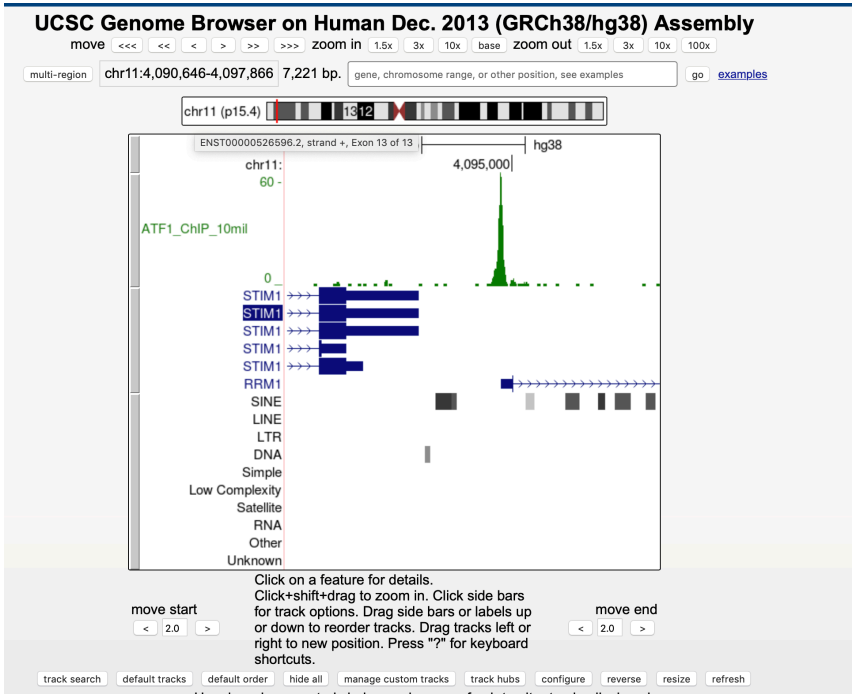


Configure Display Options

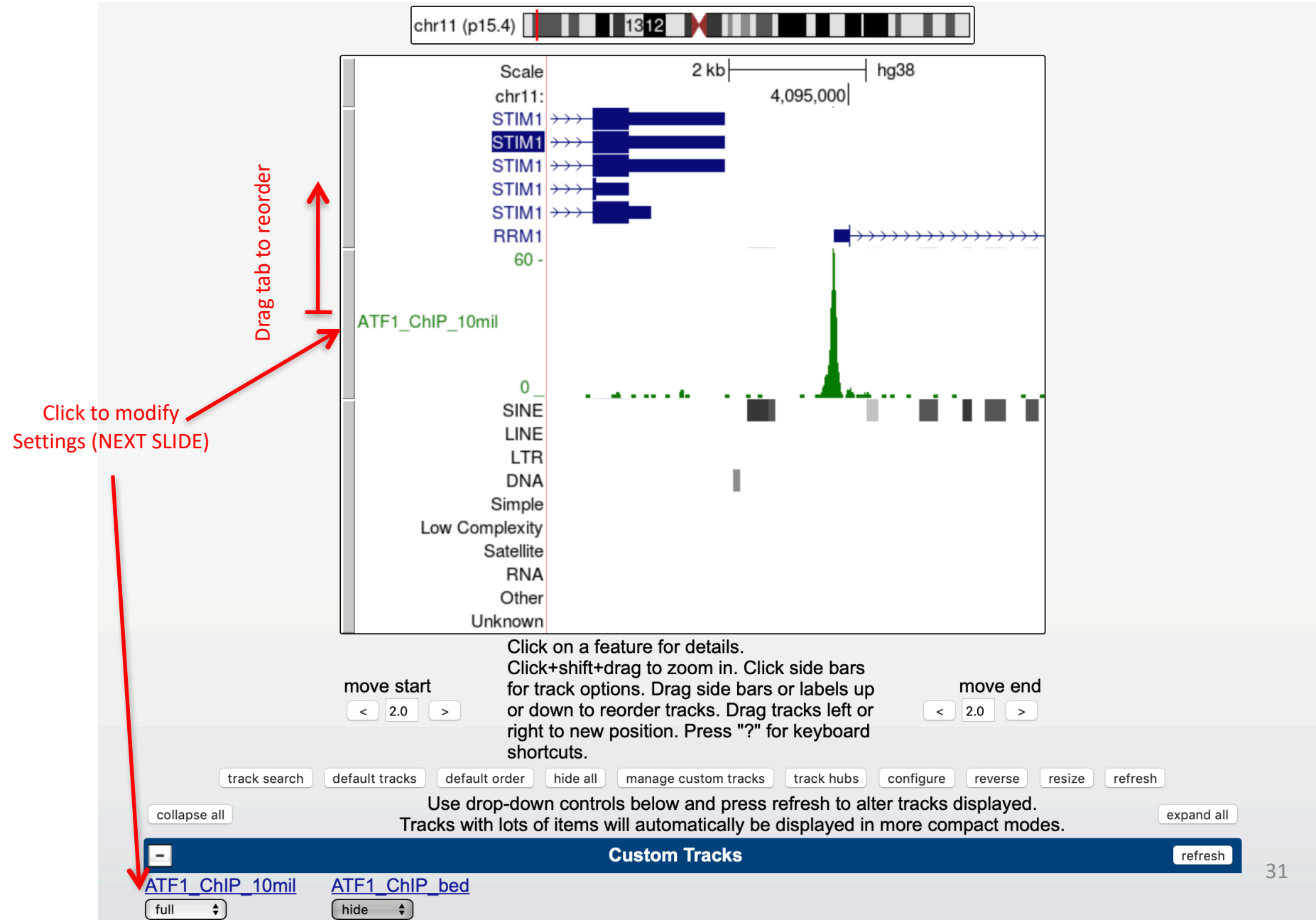


Configure Display: Clarity

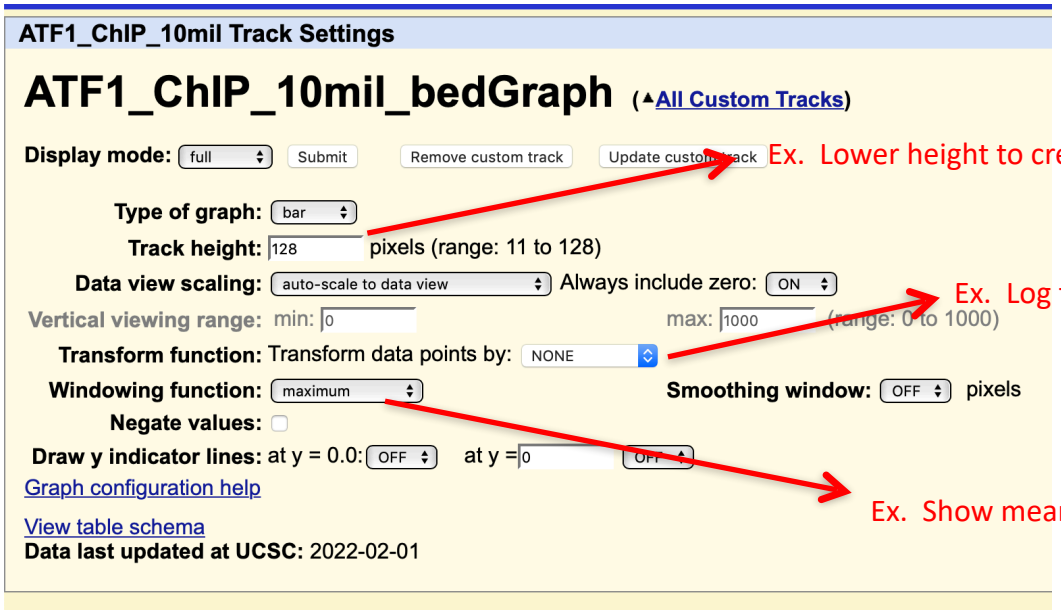




Configure Display: track order and options



Configure Display: track order and options



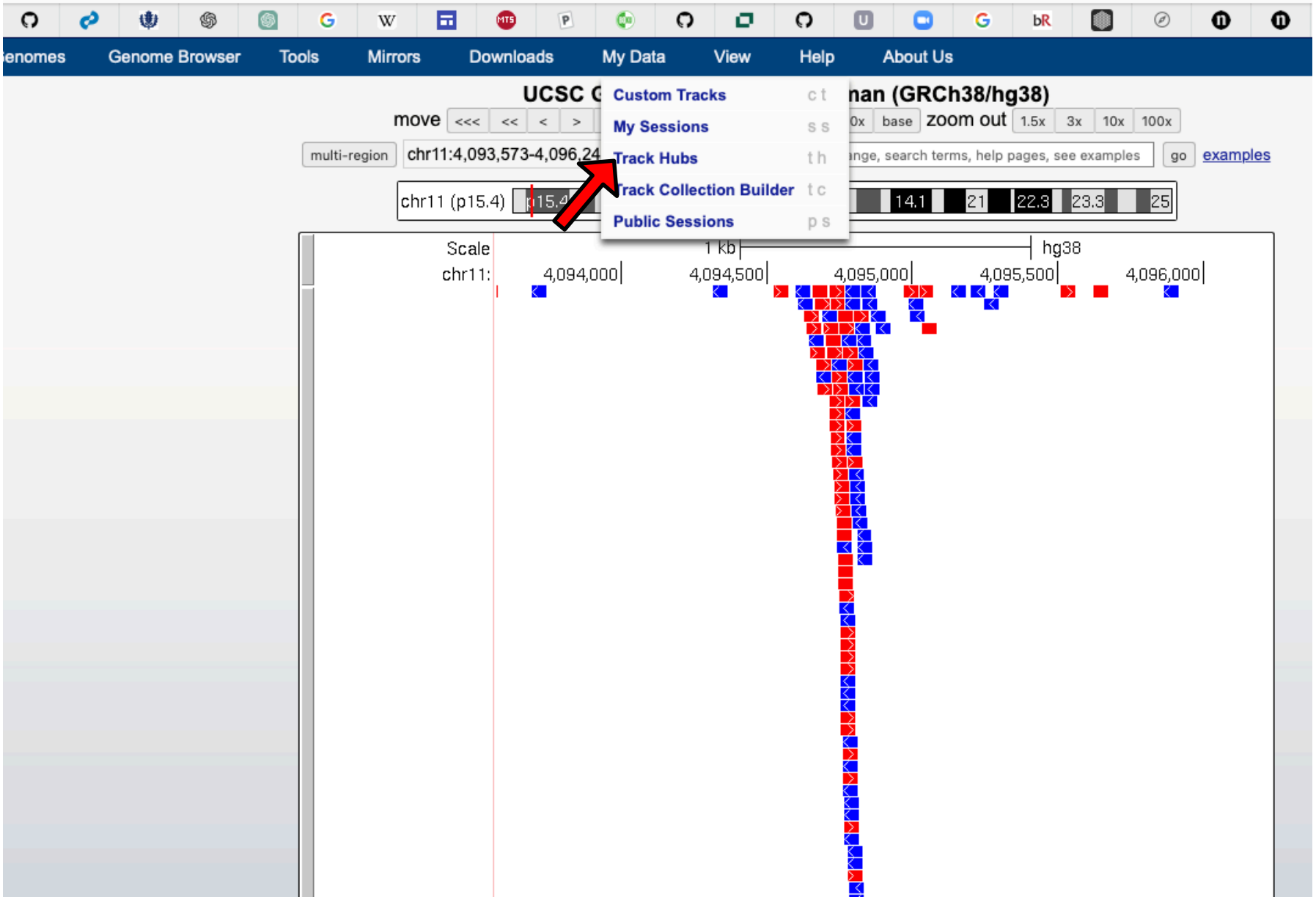
Ex. Lower height to create space with multiple tracks

Ex. Log transform the data

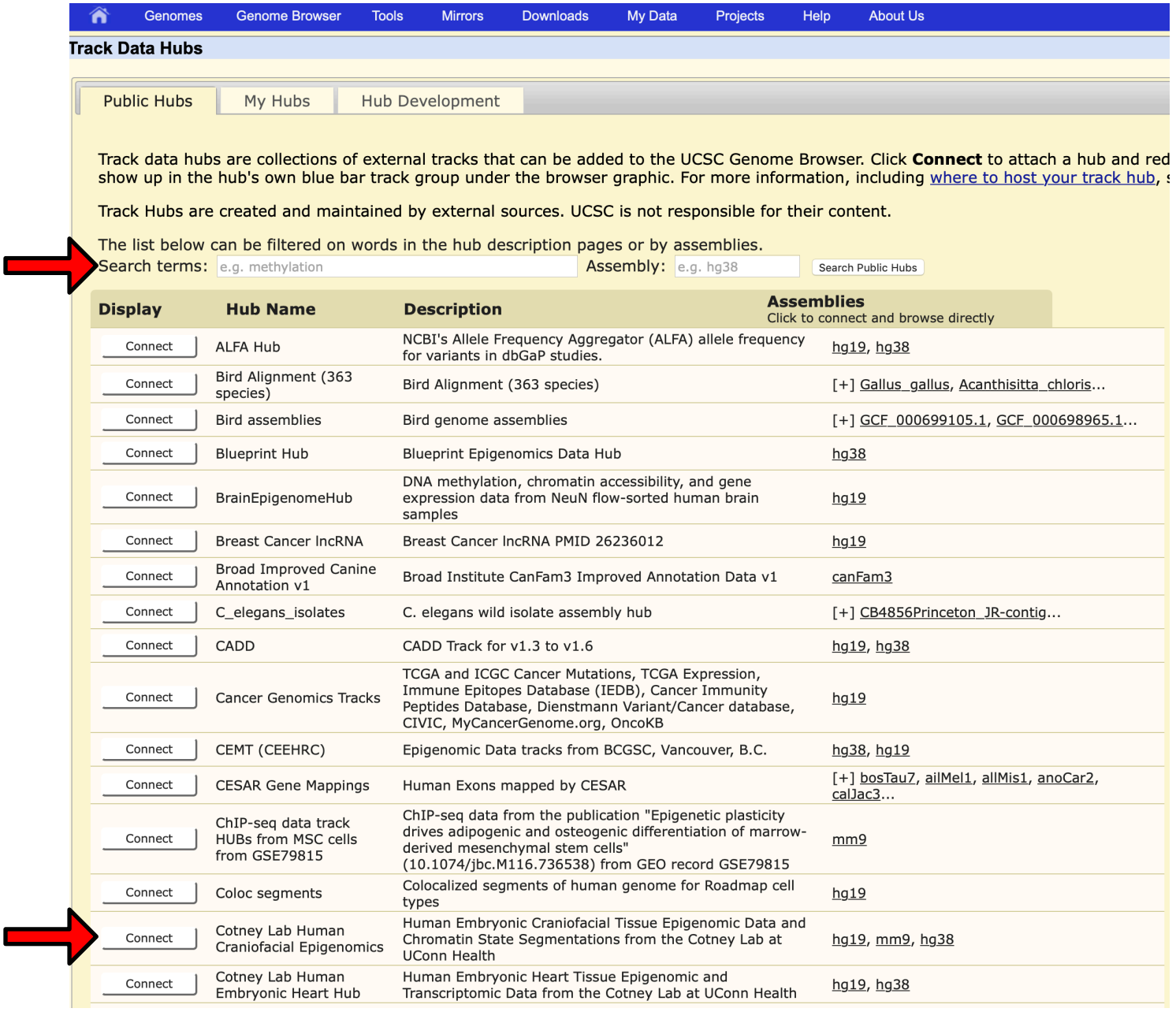
Ex. Show mean, max, SD

These are all track line settings that can be set when you upload data.
You can upload without any settings and then modify them here

Track hubs: view public data



Track hubs: view public data



Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Track Data Hubs

Public Hubs | My Hubs | Hub Development

Track data hubs are collections of external tracks that can be added to the UCSC Genome Browser. Click **Connect** to attach a hub and red
show up in the hub's own blue bar track group under the browser graphic. For more information, including where to host your track hub, s

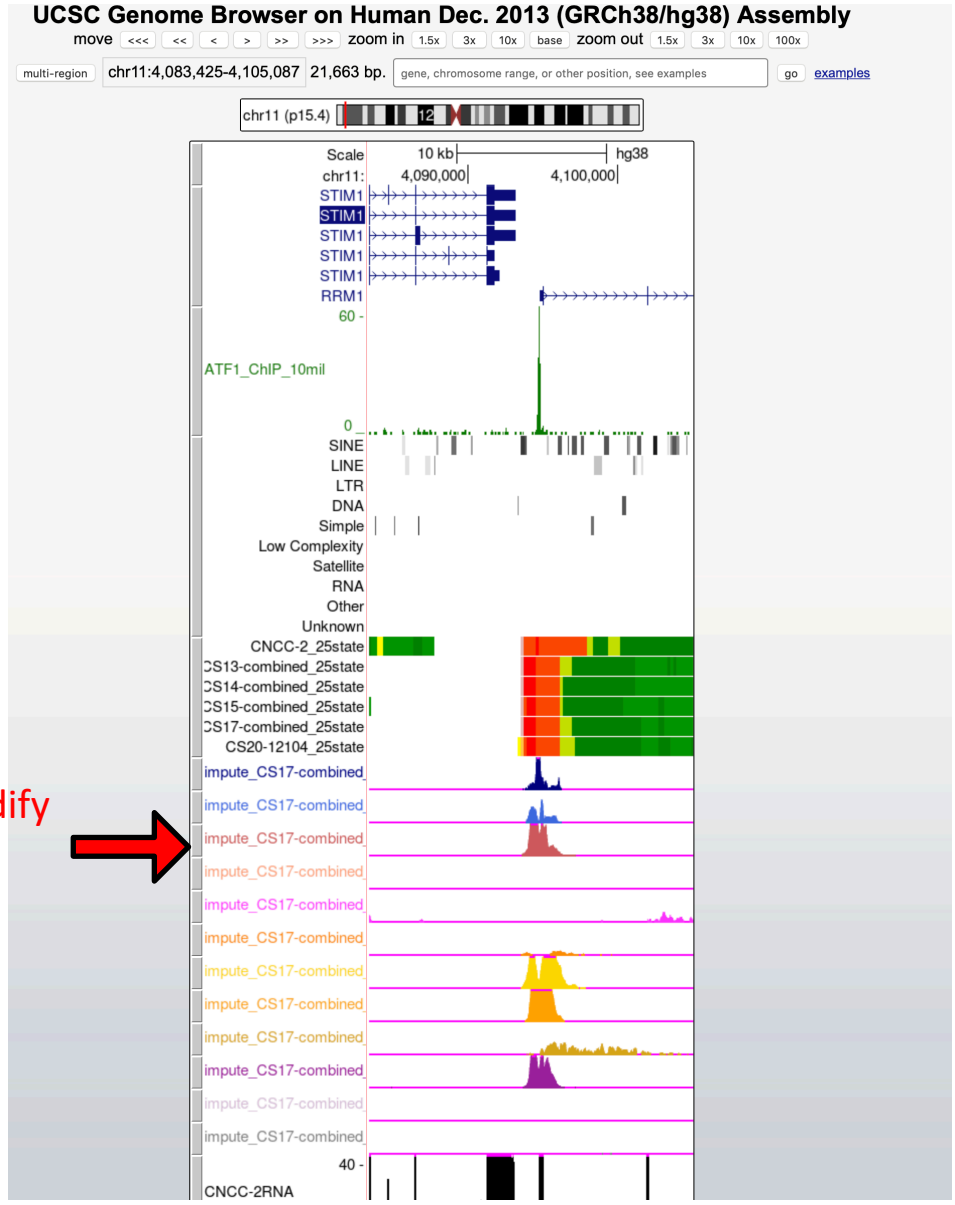
Track Hubs are created and maintained by external sources. UCSC is not responsible for their content.

The list below can be filtered on words in the hub description pages or by assemblies.

Search terms: e.g. methylation Assembly: e.g. hg38 Search Public Hubs

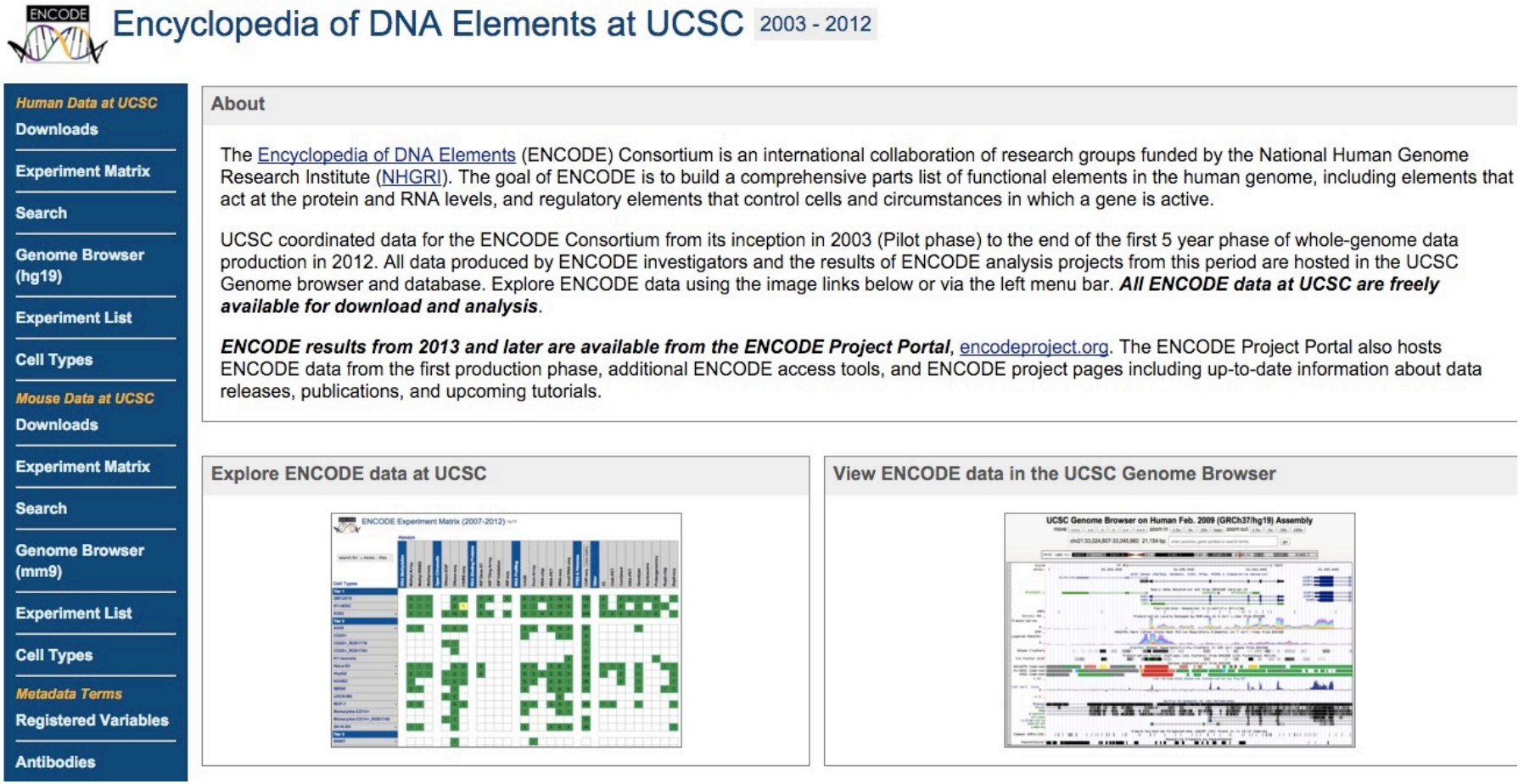
Display	Hub Name	Description	Assemblies Click to connect and browse directly
Connect	ALFA Hub	NCBI's Allele Frequency Aggregator (ALFA) allele frequency for variants in dbGaP studies.	hg19, hg38
Connect	Bird Alignment (363 species)	Bird Alignment (363 species)	[+] Gallus_gallus, Acanthisitta_chloris...
Connect	Bird assemblies	Bird genome assemblies	[+] GCF_000699105.1, GCF_000698965.1...
Connect	Blueprint Hub	Blueprint Epigenomics Data Hub	hg38
Connect	BrainEpigenomeHub	DNA methylation, chromatin accessibility, and gene expression data from NeuN flow-sorted human brain samples	hg19
Connect	Breast Cancer lncRNA	Breast Cancer lncRNA PMID 26236012	hg19
Connect	Broad Improved Canine Annotation v1	Broad Institute CanFam3 Improved Annotation Data v1	canFam3
Connect	C_elegans_isolates	C. elegans wild isolate assembly hub	[+] CB4856Princeton_JR-contig...
Connect	CADD	CADD Track for v1.3 to v1.6	hg19, hg38
Connect	Cancer Genomics Tracks	TCGA and ICGC Cancer Mutations, TCGA Expression, Immune Epitopes Database (IEDB), Cancer Immunity Peptides Database, Dienstmann Variant/Cancer database, CIVIC, MyCancerGenome.org, OncoKB	hg19
Connect	CEMT (CEEHRC)	Epigenomic Data tracks from BCGSC, Vancouver, B.C.	hg38, hg19
Connect	CESAR Gene Mappings	Human Exons mapped by CESAR	[+] bosTau7, ailMel1, allMis1, anoCar2, calJac3...
Connect	ChIP-seq data track HUBs from MSC cells from GSE79815	ChIP-seq data from the publication "Epigenetic plasticity drives adipogenic and osteogenic differentiation of marrow-derived mesenchymal stem cells" (10.1074/jbc.M116.736538) from GEO record GSE79815	mm9
Connect	Coloc segments	Colocalized segments of human genome for Roadmap cell types	hg19
Connect	Cotney Lab Human Craniofacial Epigenomics	Human Embryonic Craniofacial Tissue Epigenomic Data and Chromatin State Segmentations from the Cotney Lab at UConn Health	hg19, mm9, hg38
Connect	Cotney Lab Human Embryonic Heart Hub	Human Embryonic Heart Tissue Epigenomic and Transcriptomic Data from the Cotney Lab at UConn Health	hg19, hg38

Track hubs: view public data



Click to modify settings

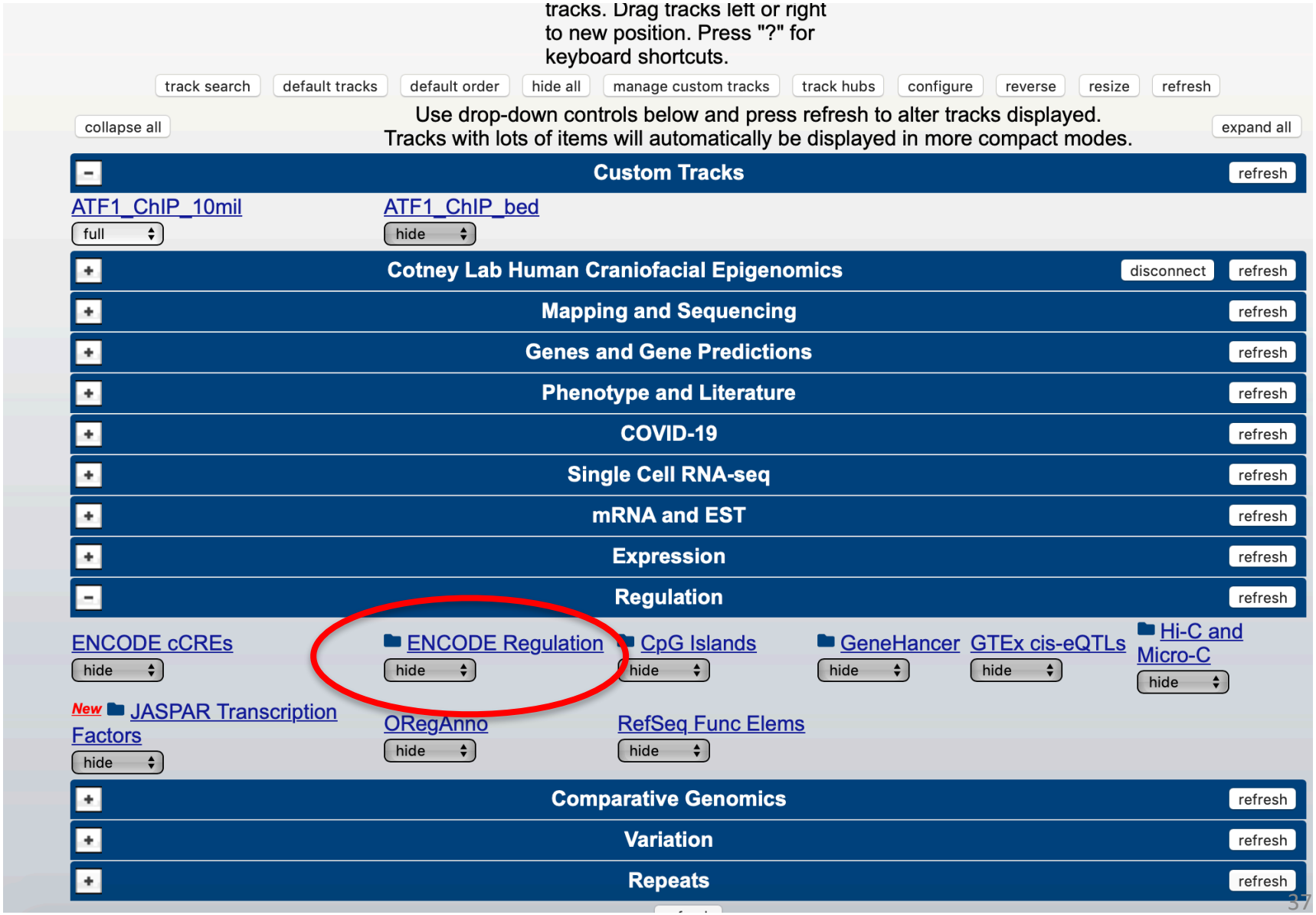
ENCODE data is available from several organisms



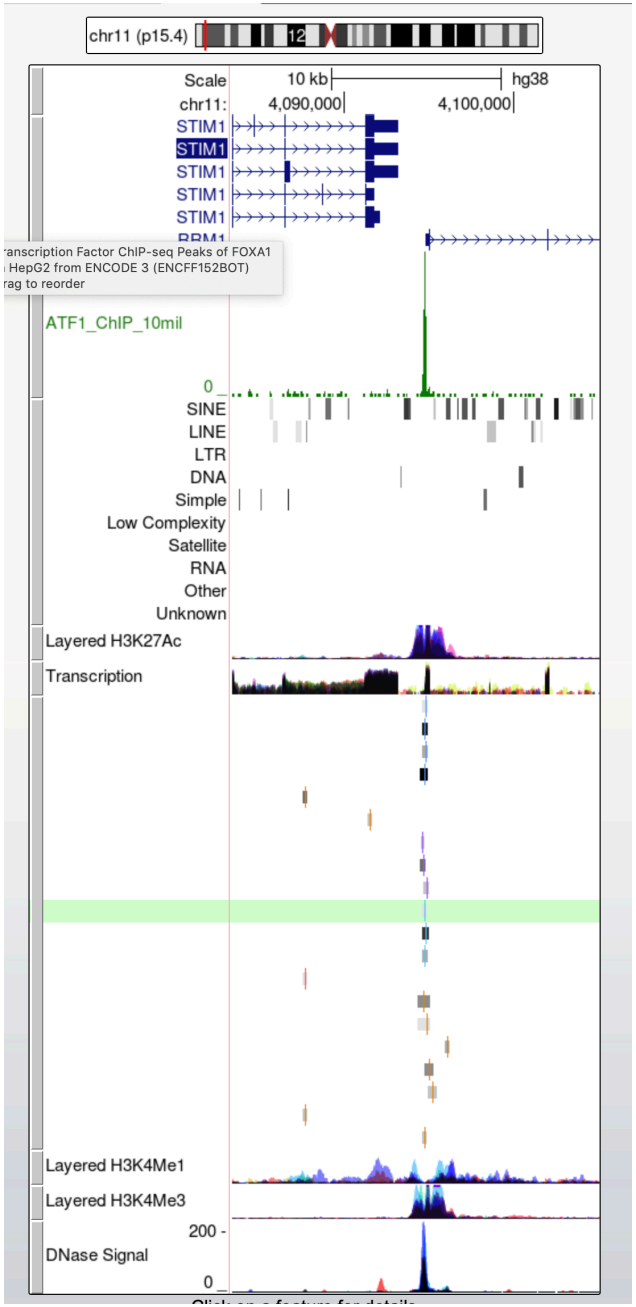
Experiment matrix

Preloaded browser tracks

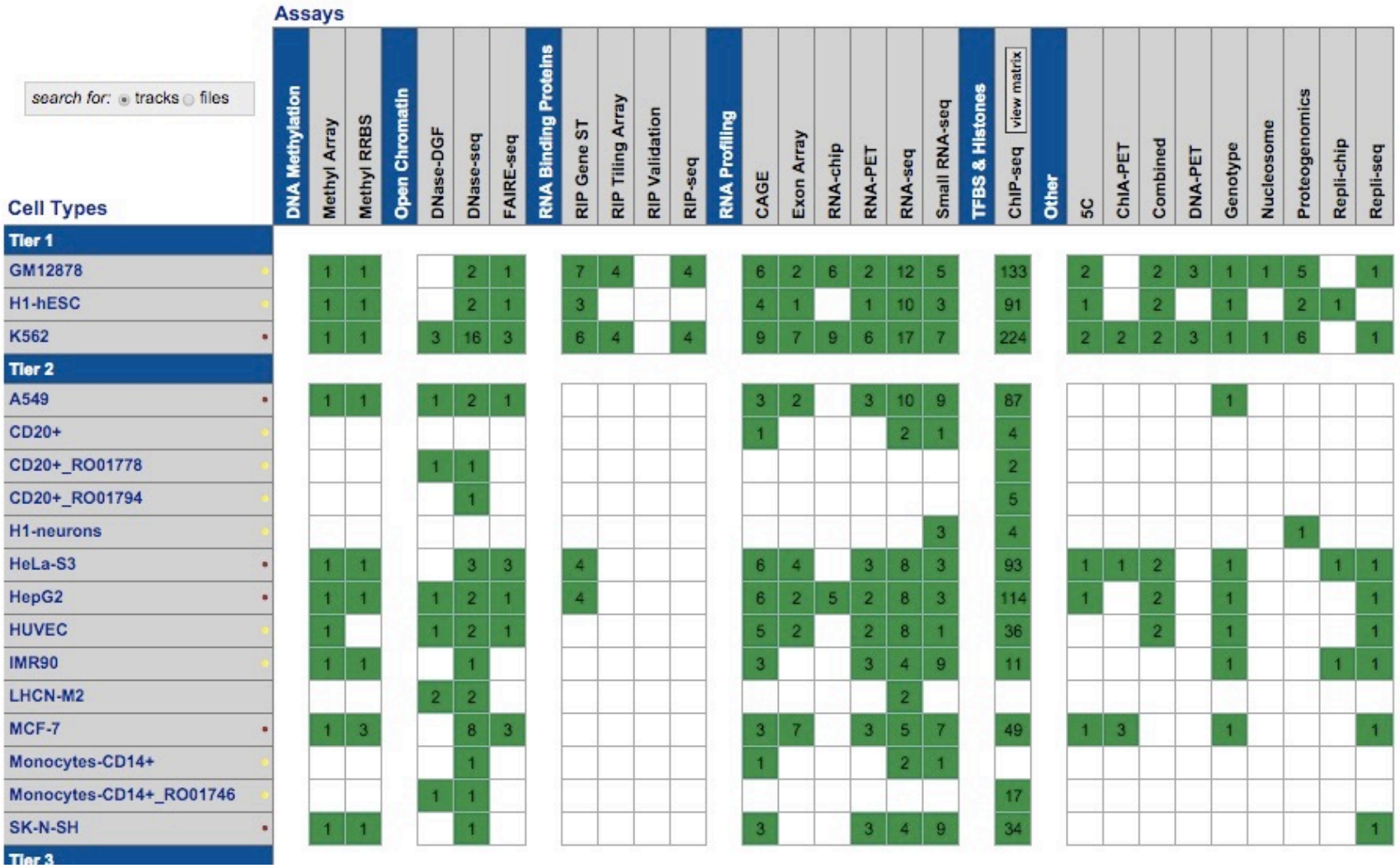
ENCODE data tracks



ENCODE data browser



ENCODE experiment matrix



Can view in browser or download raw and processed data

Resources / links

Browser:

- http://www.sciencedirect.com/science/article/pii/S0888754308000451
- http://genome.ucsc.edu/training/vids/
- http://www.nature.com/scitable/ebooks/guide-to-the-ucsc-genome-browser-16569863

ENCODE:

- http://genome.ucsc.edu/ENCODE/index.html
- http://www.genome.gov/encode/