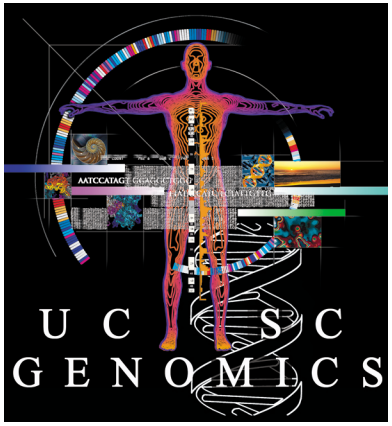
Visualizing ENCODE Data in the UCSC Genome Browser



Pauline Fujita, Ph.D.
UCSC Genome Bioinformatics Group

Training Resources

genome@soe.ucsc.edu

- Genomewiki: *genomewiki.ucsc.edu*
- Mailing list archives: *genome.ucsc.edu/FAQ/*
- Training page:

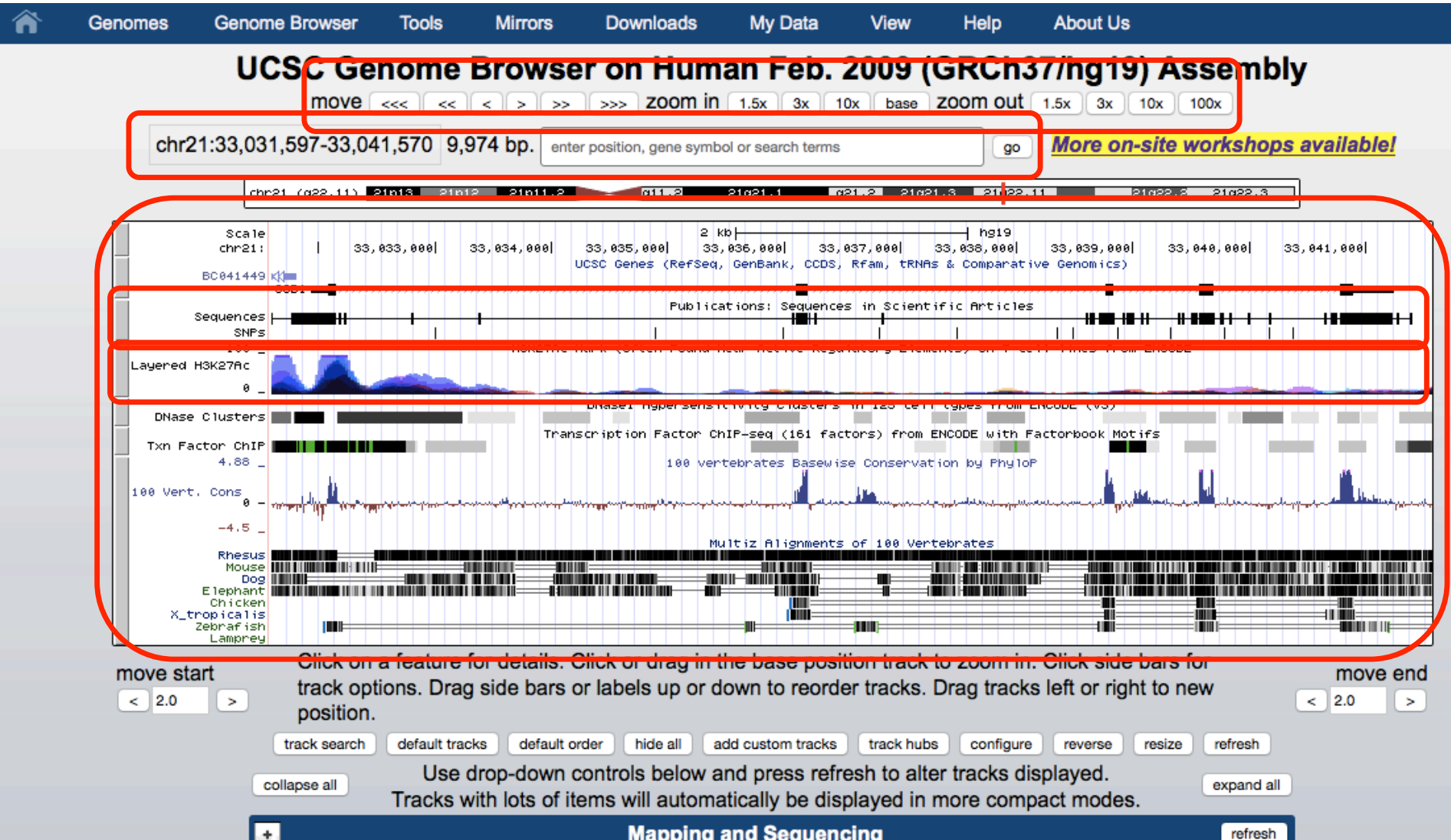
  genome.ucsc.edu/training.html
- Twitter @GenomeBrowser
- Tutorial videos: YouTube channel
- Open Helix: *openhelix.com/ucsc*

Outline

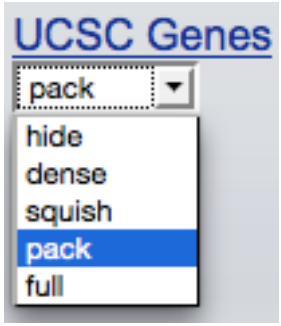
- Basics: search, display, more info
- Tools for finding ENCODE data
- Annotating a BED file: RNAseq example
- Annotating a VCF file
- Track Hubs: What are they? How do I make one?
- Exercises

Basic Navigation: Main Display

genome.ucsc.edu/cgi-bin/hgTracks?db=hg19

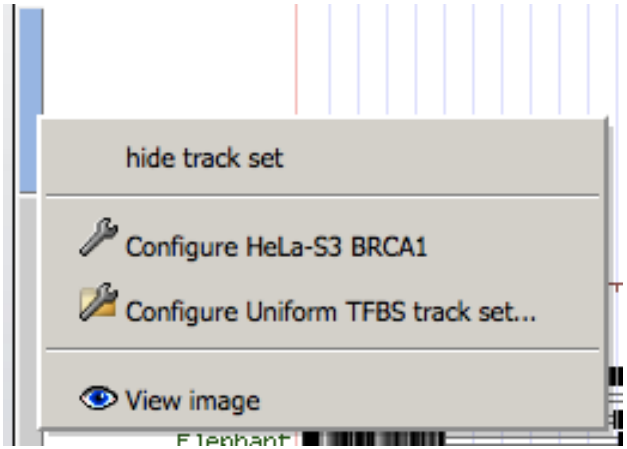


Display Configuration

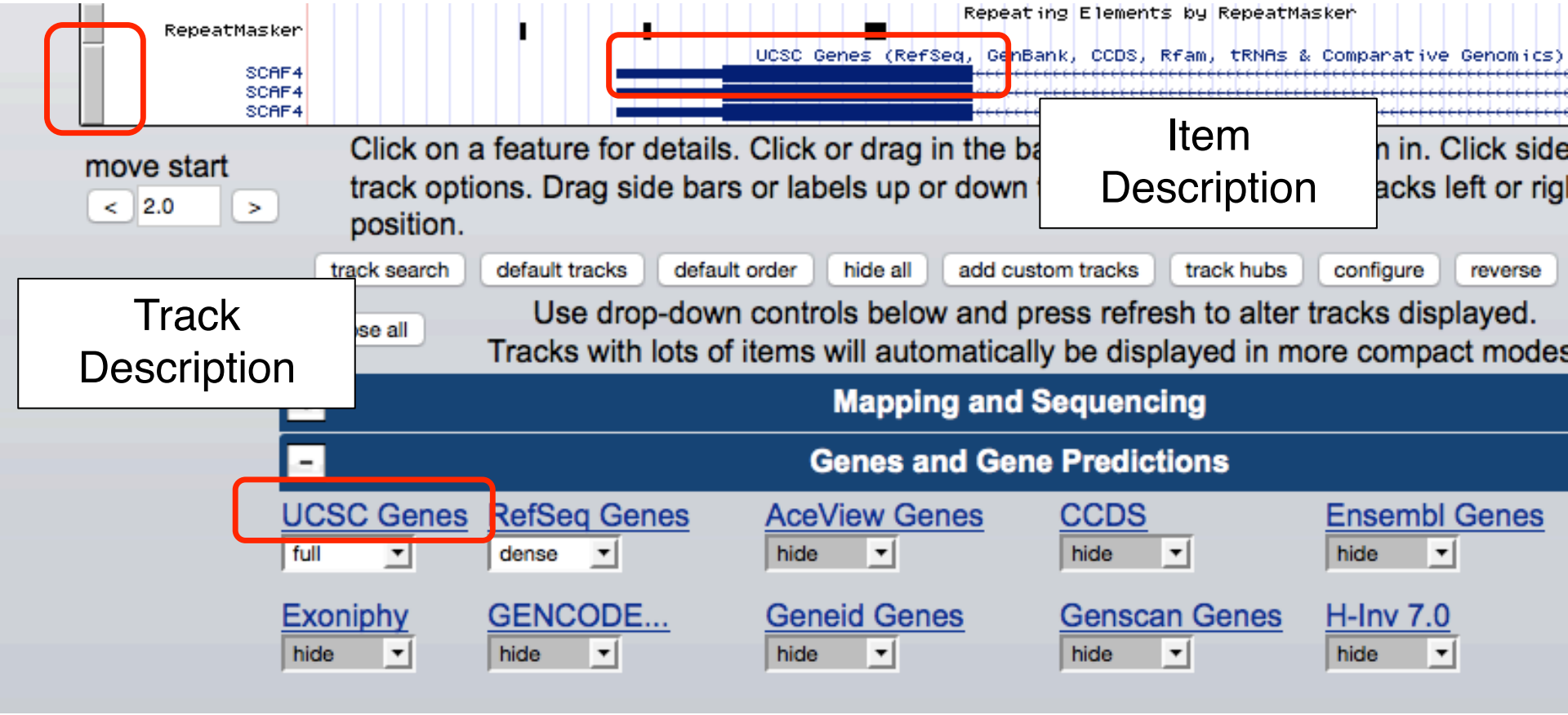


- Visibility:

  hide, dense, squish, pack, full
- Track ordering: drag and drop
- Drag and zoom/highlighting
- Configuration page
- Right click menu



How to find more info



More info: Track Description

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

UCSC Genes Track Settings

UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)

(▲All Genes and Gene Predictions tracks)

Display mode: full Submit

Label: gene symbol UCSC Known Gene ID UniProt Display ID OMIM ID

Show: non-coding genes splice variants

Color track by codons: genomic codons Help on codon coloring

Show codon numbering:

View table schema

Data last updated: 2013-06-14

Description

The UCSC Genes track is a set of gene predictions based on data from RefSeq, GenBank, CCDS, Rfam, and the tRNA Genes track. The track includes both protein-coding genes and non-coding RNA genes. Both types of genes can produce non-coding transcripts, but non-coding RNA genes do not produce protein-coding transcripts. This is a moderately conservative set of predictions. Transcripts of protein-coding genes require the support of one RefSeq RNA, or one GenBank RNA sequence plus at least one additional line of evidence. Transcripts of non-coding RNA genes require the support of one Rfam or tRNA prediction. Compared to RefSeq, this gene set has generally about 10% more protein-coding genes, approximately four times as many putative non-coding genes, and about twice as many splice variants.

More info: Item Description

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

Human Gene SCAF4 (uc002ypd.2) Description and Page Index

Description: Homo sapiens SR-related CTD-associated factor 4 (SCAF4), transcript variant 1, mRNA.
RefSeq Summary (NM_020706): This gene likely encodes a member of the arginine/serine-rich splicing factor family. A similar protein in Rat appears to bind the large subunit of RNA polymerase II and provide a link between transcription and pre-mRNA splicing. Alternatively spliced transcript variants have been described. [provided by RefSeq, Feb 2009].
Transcript (Including UTRs)
Position: chr21:33,043,313-33,104,431 **Size:** 61,119 **Total Exon Count:** 20 **Strand:** -
Coding Region
Position: chr21:33,043,712-33,104,005 **Size:** 60,294 **Coding Exon Count:** 20

Page Index	Sequence and Links	UniProtKB Comments	CTD	Gene Alleles	RNA Structure
Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Other Names	Model Information
Methods					

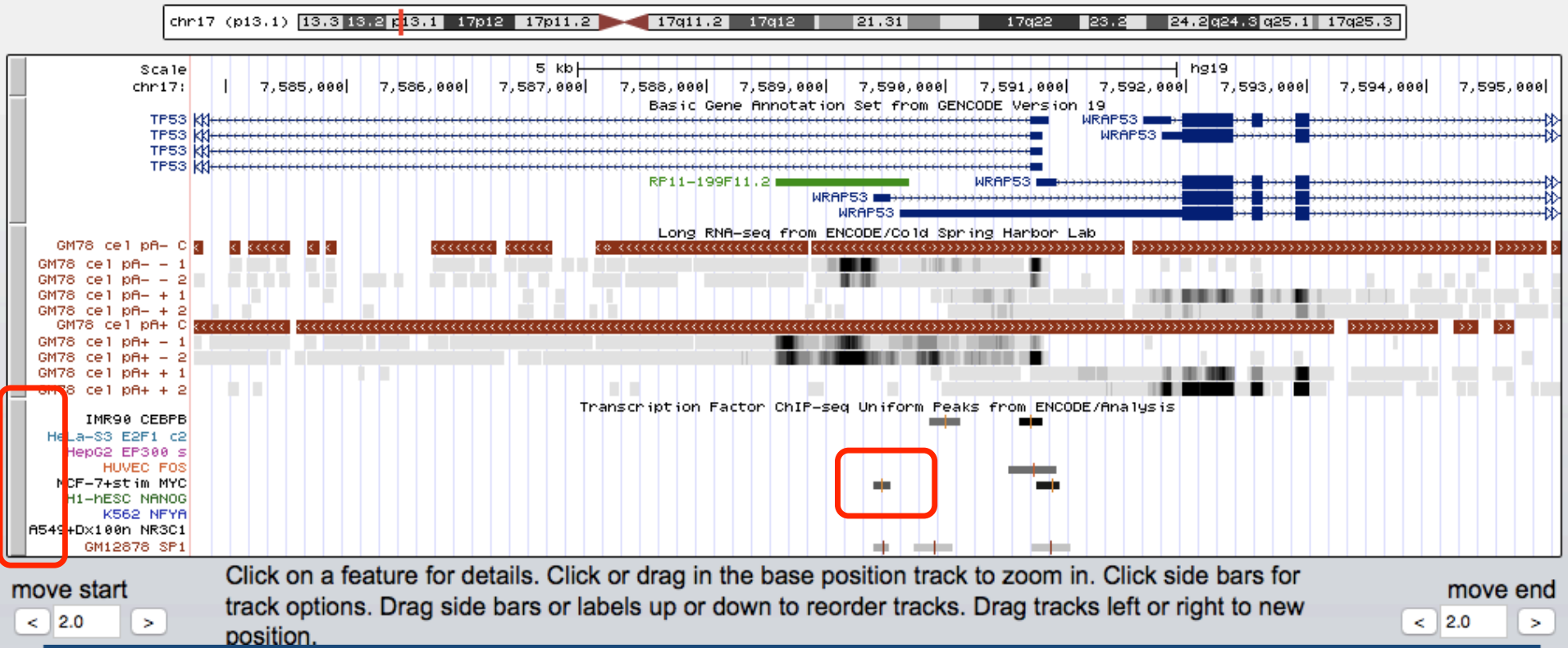
Data last updated: 2013-06-14

− Sequence and Links to Tools and Databases

Genomic Sequence (chr21:33,043,313-33,104,431)			mRNA (may differ from genome)		Protein (1147 aa)
Gene Sorter	Genome Browser	Protein FASTA	VisiGene	Table Schema	BioGPS
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneNetwork
Gepis Tissue	H-INV	HGNC	HPRD	Jackson Lab	Lynx
MOPED	neXtProt	OMIM	PubMed	Stanford SOURCE	UniProtKB

− Comments and Description Text from UniProtKB

ENCODE



move start < 2.0 >

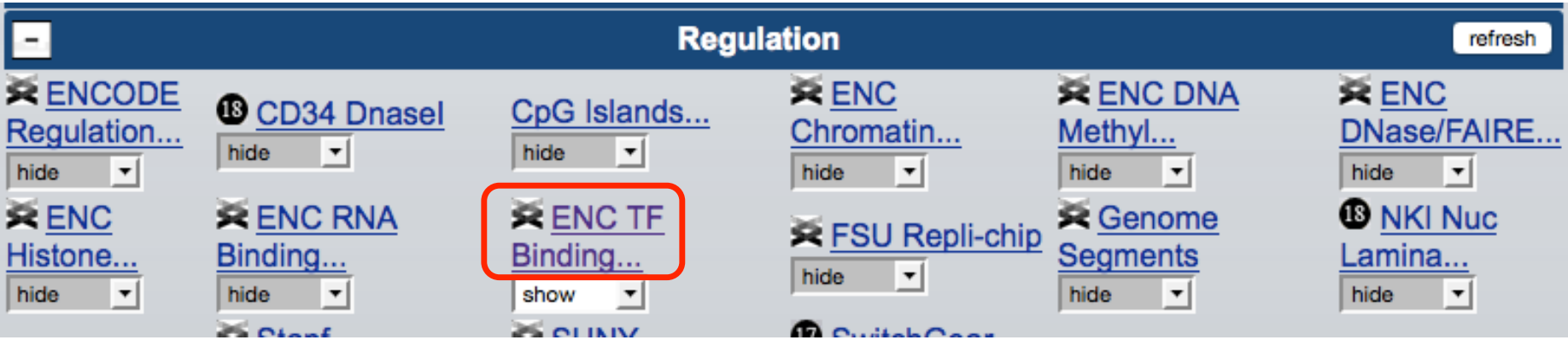
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

move end < 2.0 >

Regulation refresh

ENCODE Regulation... hide	CD34 Dnase1 hide	CpG Islands... hide	ENC Chromatin... hide	ENC DNA Methyl... hide	ENC DNase/FAIRE... hide
ENC Histone... hide	ENC RNA Binding... hide	ENC TF Binding... show	FSU Repli-chip hide	Genome Segments hide	NKI Nuc Lamina... hide

ENCODE: Super-track Settings



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

ENC TF Binding Super-track Settings

ENCODE Transcription Factor Binding Tracks (▲All Regulation tracks)

Display mode: show Submit

All

dense	Uniform TFBS	Transcription Factor ChIP-seq Uniform Peaks from ENCODE/Analysis *ENCODE March 2012 Freeze*
hide	HAIB TFBS	Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
hide	SYDH TFBS	Transcription Factor Binding Sites by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard
hide	UChicago TFBS	Transcription Factor Binding Sites by Epitope-Tag from ENCODE/UChicago
hide	UTA TFBS	Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin) *ENCODE July 2011 Freeze*
hide	UW CTCF Binding	CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington

Description

ENCODE: Track Settings

Uniform TFBS Track Settings ENCODE Downloads Subtracks↓ Description↓ Contact↓

Transcription Factor ChIP-seq Uniform Peaks from ENCODE/Ana

Display mode: dense Submit Cancel Reset to defaults

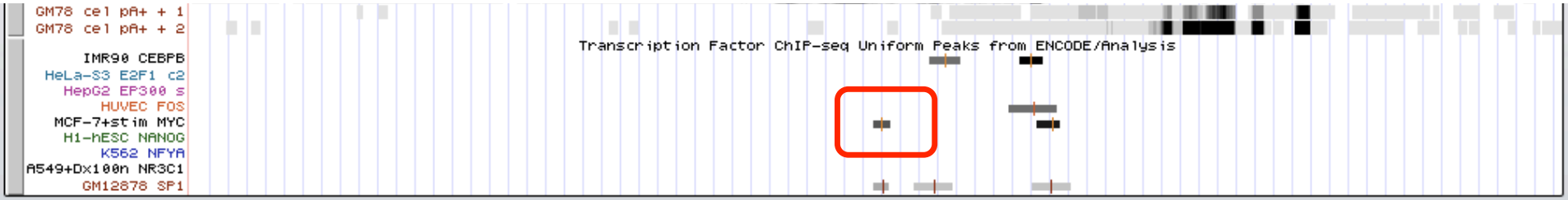
Score range: min: 0 (0 to 1000)

Select subtracks by cell line and factor: (help)

All

Factor	Cell Line GM12878 (Tier 1)	H1-hESC (Tier 1)	K562 (Tier 1)	HeLa-S3 (Tier 2)	HepG2 (Tier 2)	HUVEC (Tier 2)	IMR90 (Tier 2*)	A549 (Tier 2*)	MCF-7 (Tier 2*)	SK-N-SH (Tier 2*)	AG04449	AG04450	AG09309	AG09319	AG10803	AoAF	BE2 C	BJ	Caco-2	Dnd41	ECC-1	Fibrobl	Gliobla	GM06990	GM08714	GM10847	GM12801	GM12864	GM12865	GM12872	GM12873	GM12874	GM12
ARID3A			☐		☐																												
ATF1			☐																														
ATF2	☐	☐																															

ENCODE: Item Details



move start < 2.0 >

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.

move end < 2.0 >

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

MCF-7 (serum_stimulated) TFBS Uniform Peaks of c-Myc from ENCODE/UT-A/Analysis

Position: chr17:7589391-7589530
Peak point: 7589461
Score: 688
Signal value: 107.768
Q-value (FDR): 4.212

View table: schema, downloads, metadata

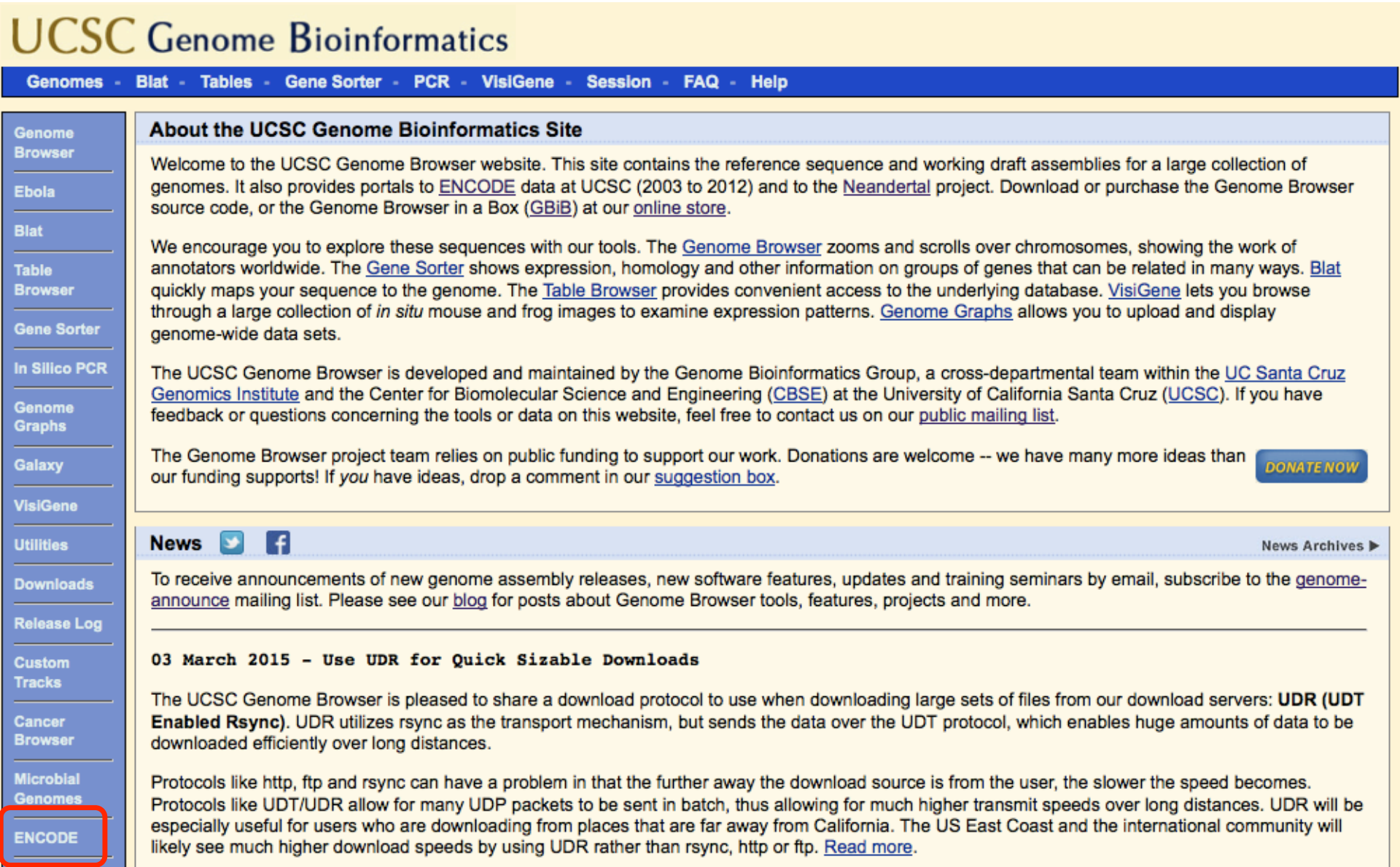
Go to Uniform TFBS track controls

Data version: ENCODE Mar 2012 Freeze
Data last updated: 2013-04-12

Description

This track represents a comprehensive set of human transcription factor binding sites based on ChIP-seq experiments generated by production in the ENCODE Consortium from the inception of the project in September 2007, through the March 2012 internal data freeze. The track repres

ENCODE Tools



ENCODE

ENCODE → genome.ucsc.edu/ENCODE/

Encyclopedia of DNA Elements at UCSC 2003 - 2012

- ***Human Data at UCSC***
- **Downloads**
- **Experiment Matrix**
- **Search**
- **Genome Browser (hg19)**
- **Experiment List**
- **Cell Types**
- ***Mouse Data at UCSC***
- **Downloads**
- **Experiment Matrix**
- **Search**
- **Genome Browser (mm9)**

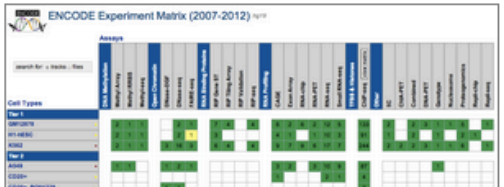
About

The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

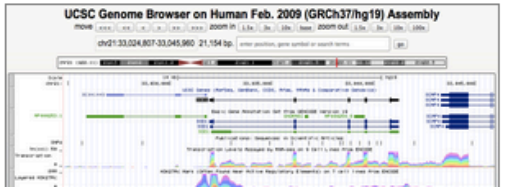
UCSC coordinated data for the ENCODE Consortium from its inception in 2003 (Pilot phase) to the end of the first 5 year phase of whole-genome data production in 2012. All data produced by ENCODE investigators and the results of ENCODE analysis projects from this period are hosted in the UCSC Genome browser and database. Explore ENCODE data using the image links below or via the left menu bar. ***All ENCODE data at UCSC are freely available for download and analysis***.

ENCODE results from 2013 and later are available from the ENCODE Project Portal, encodeproject.org. The ENCODE Project Portal also hosts ENCODE data from the first production phase, additional ENCODE access tools, and ENCODE project pages including up-to-date information about data releases, publications, and upcoming tutorials.

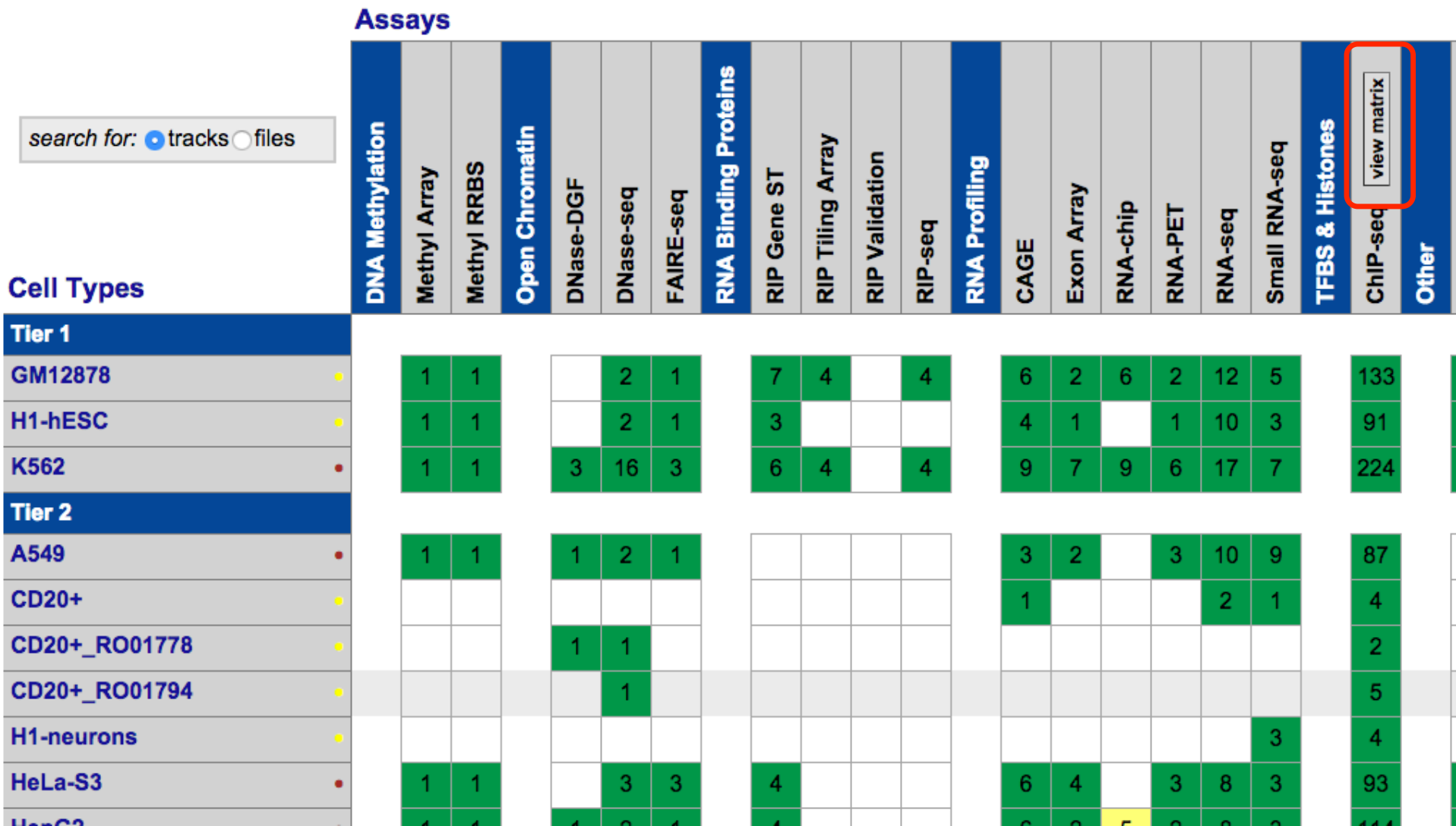
Explore ENCODE data at UCSC



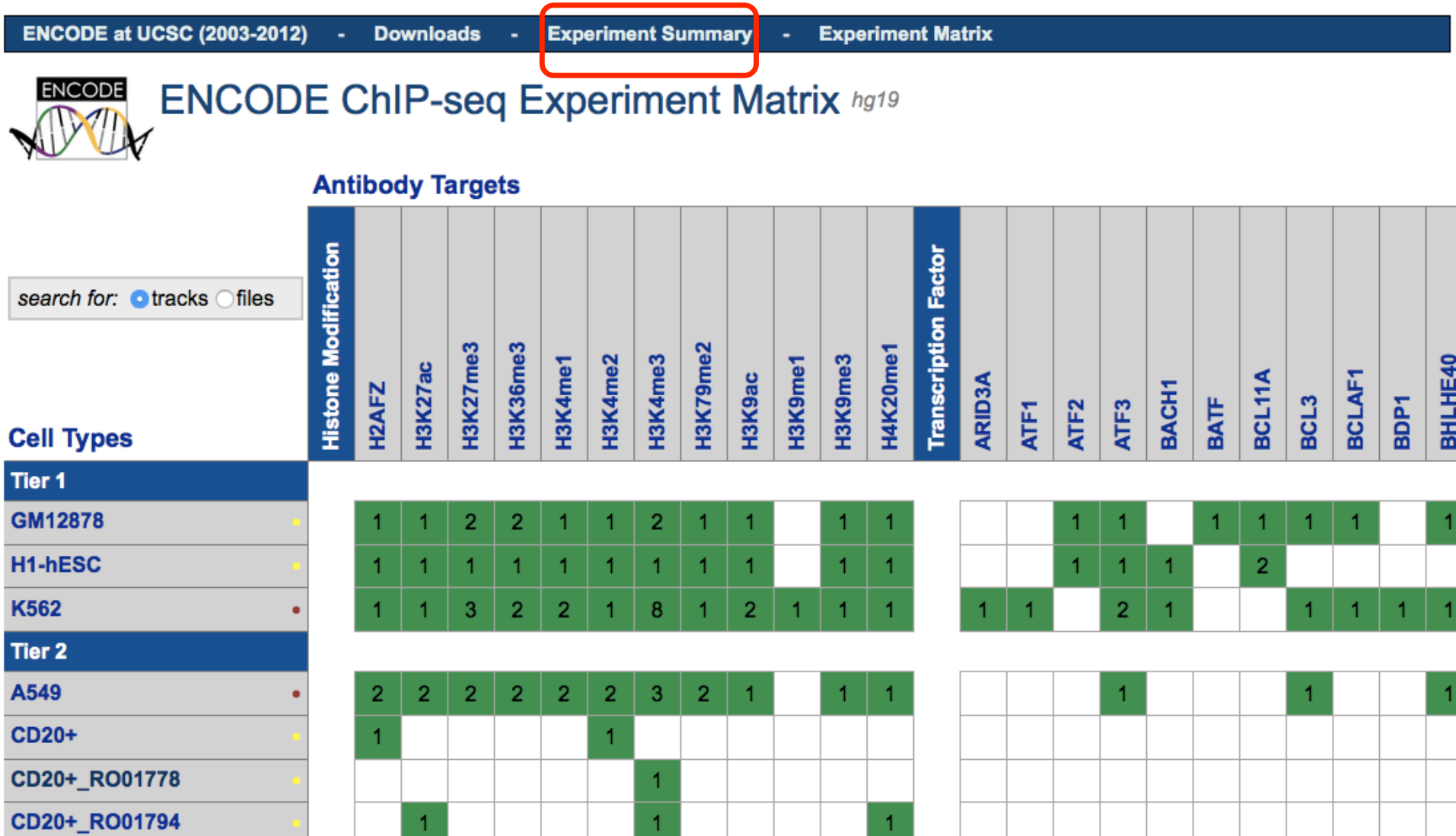
View ENCODE data in the UCSC Genome Browser



ENCODE: Experiment Matrix



ENCODE: ChIP-Seq Matrix



ENCODE: Experiment Summary

ENCODE Experiment Summary *hg19* 2007 - 2012

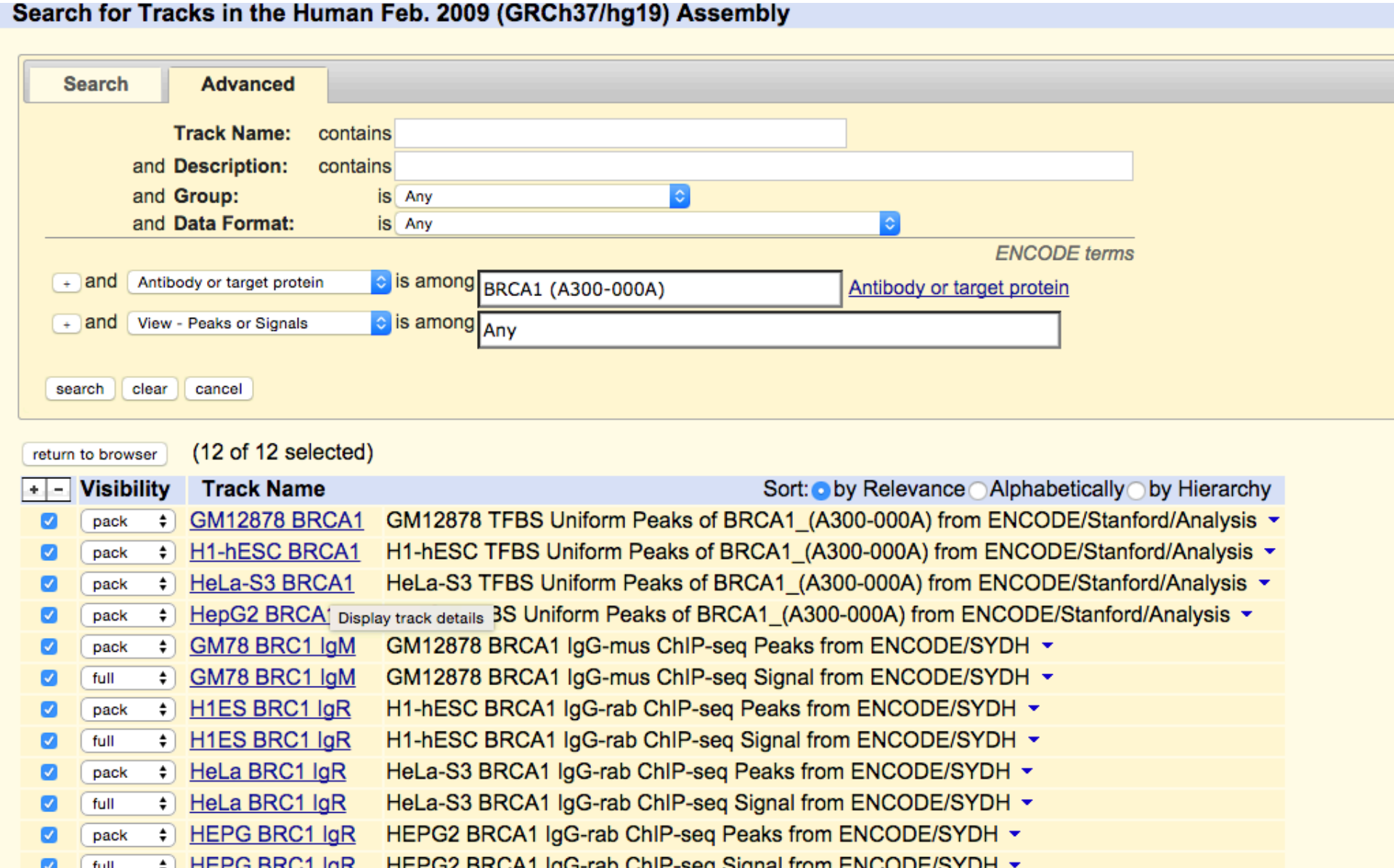
search for: tracks files

Genome Annotation	
GENCODE	4
Mappability	10
ORChID	1
Total: 3	**15**

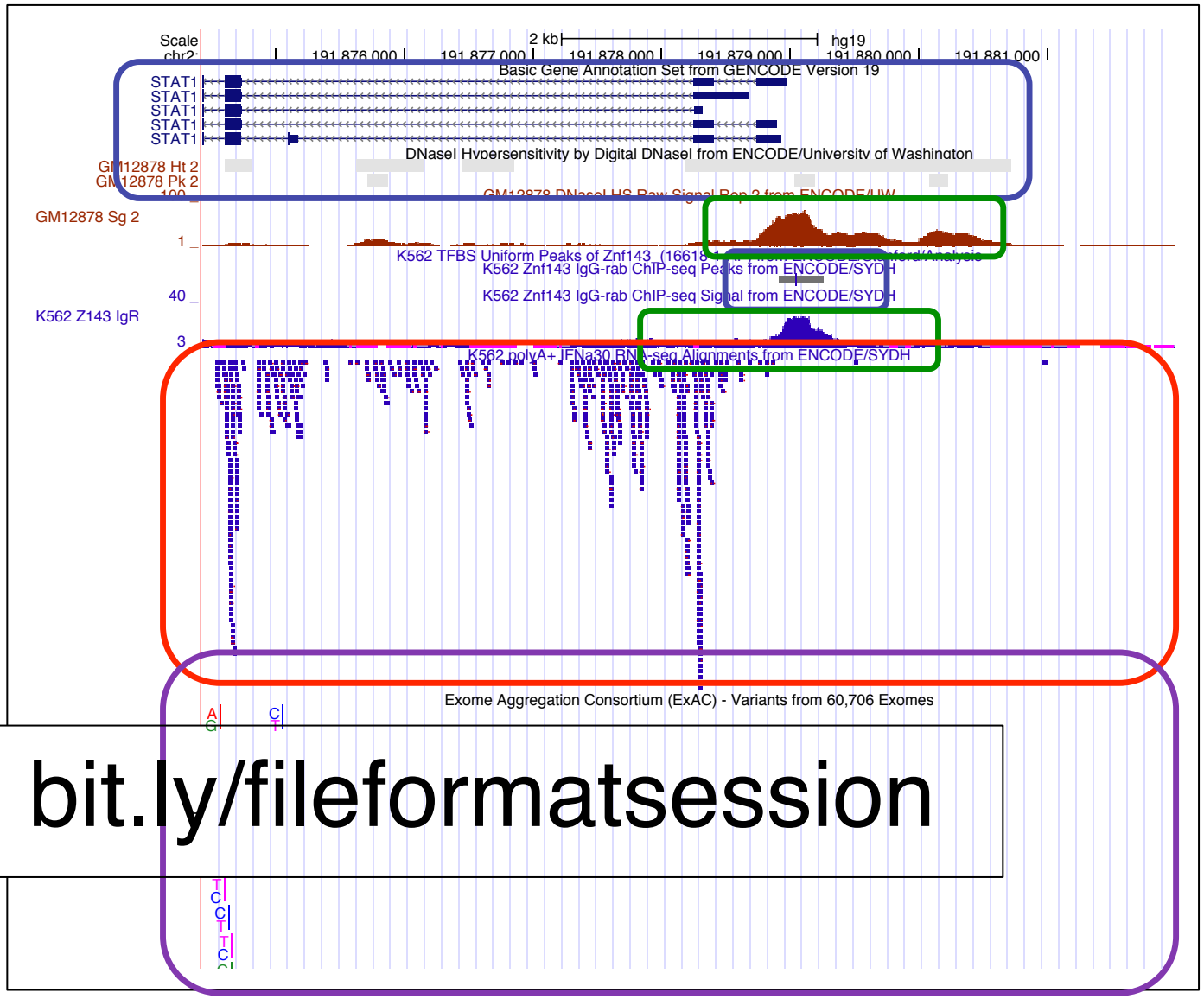
All Experiments	view matrix by cell line
5C	13
CAGE	78
ChIA-PET	8
ChIP-seq	1479
Combined	26
DNA-PET	6
DNase-DGF	56
DNase-seq	219
Exon Array	158
FAIRE-seq	37
Genotype	64
Methyl Array	62

ChIP-seq Experiments	view matrix by cell line
ARID3A	2
ATF1	1
ATF2	2
ATF3	7
BACH1	2
BATF	1
BCL11A	3
BCL3	3
BCLAF1	2
BDP1	2
BHLHE40	5
BRCA1	4

ENCODE: Track Search



File Formats



BED

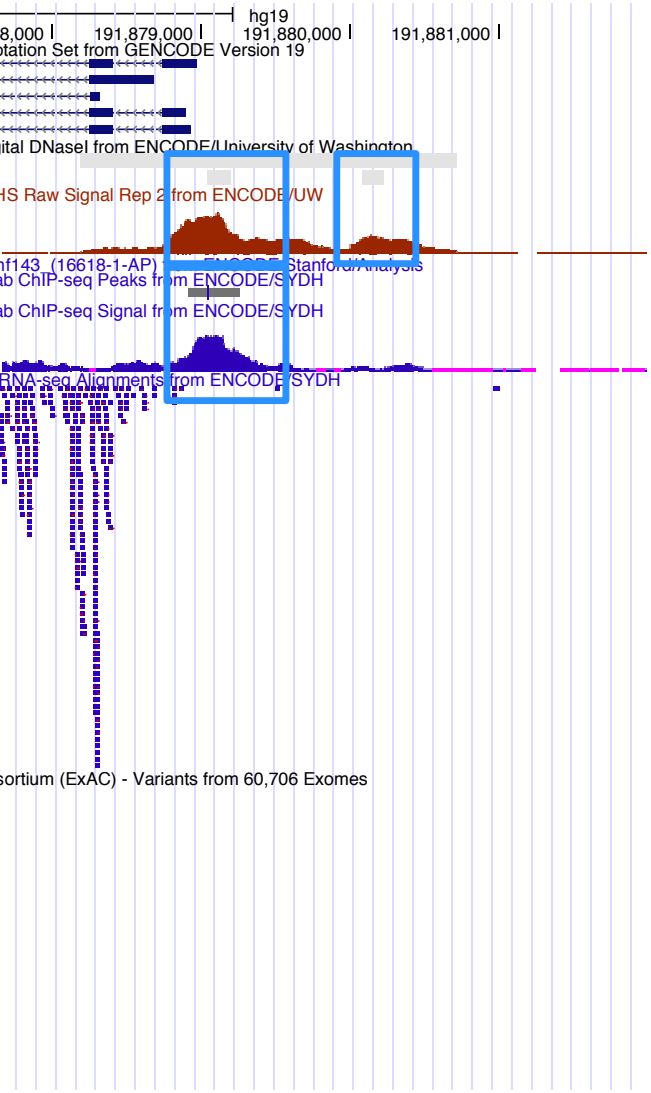
wig(gle)

BAM

VCF

bit.ly/fileformatsession

File Formats



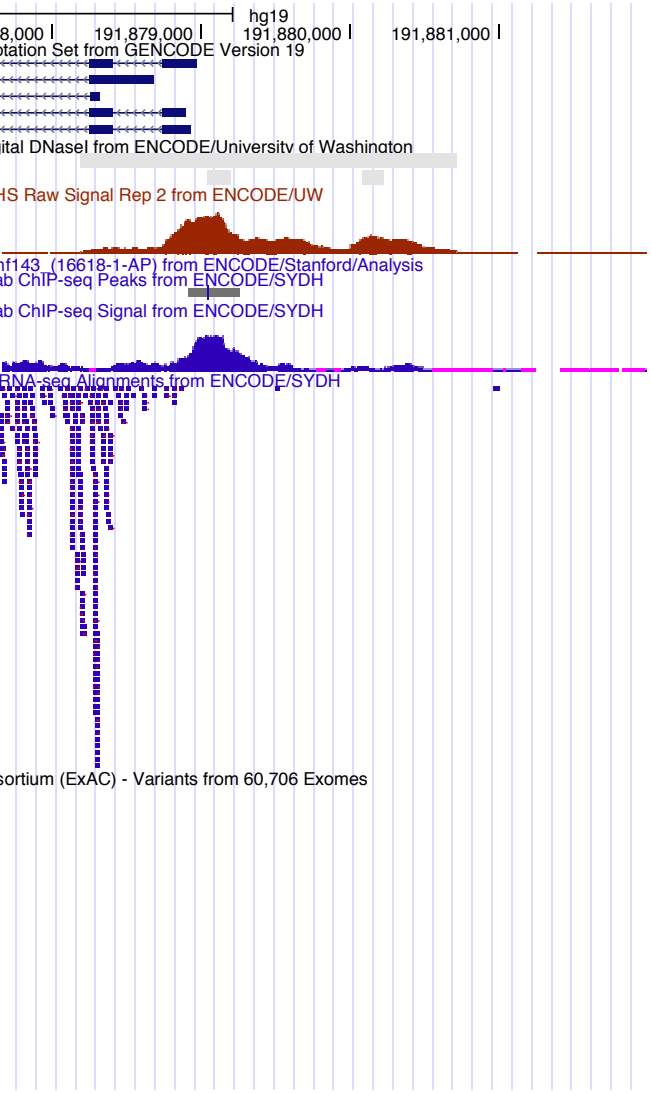
BED Positional annotations. (ex. Regions w/: enriched ChIP-seq signal for TF binding, Δ'l methylation, splice jxns from RNA-seq)

wig(gle) Continuous signal data. # of reads (ex. DNase I HS and ChIP-seq signals)

BAM Alignments of seq. reads, mapped to genome (ex. RNA-seq alignments)

VCF Variation data: SNPs, indels, Copy Number Variants, Structural Variants (ex. ExAC data)

Indexed File Formats



BED ⇨ **bigBed**

wig(gle) ⇨ **bigWig**

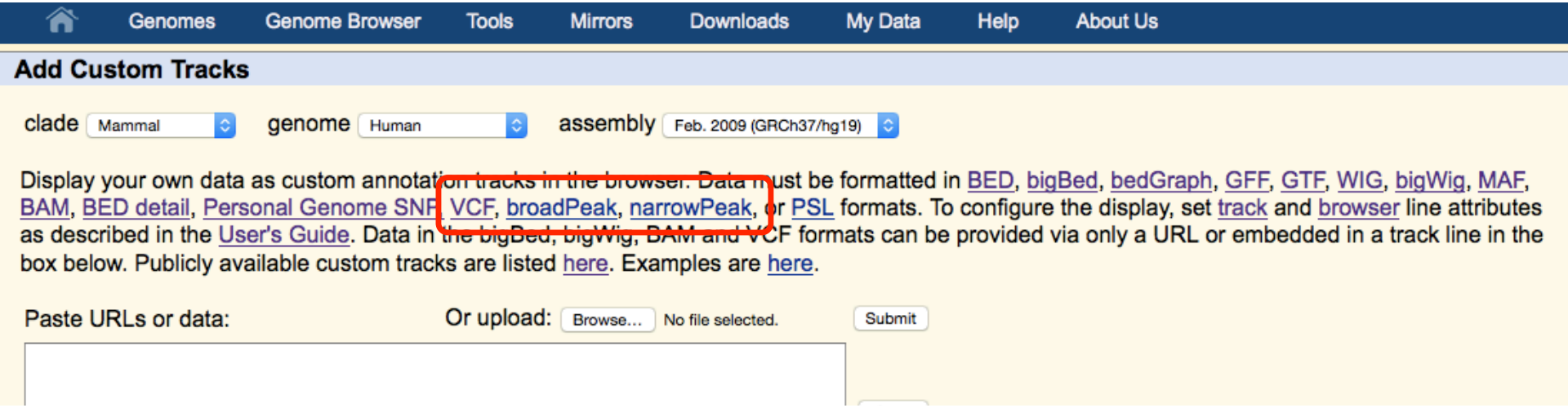
BAM

VCF

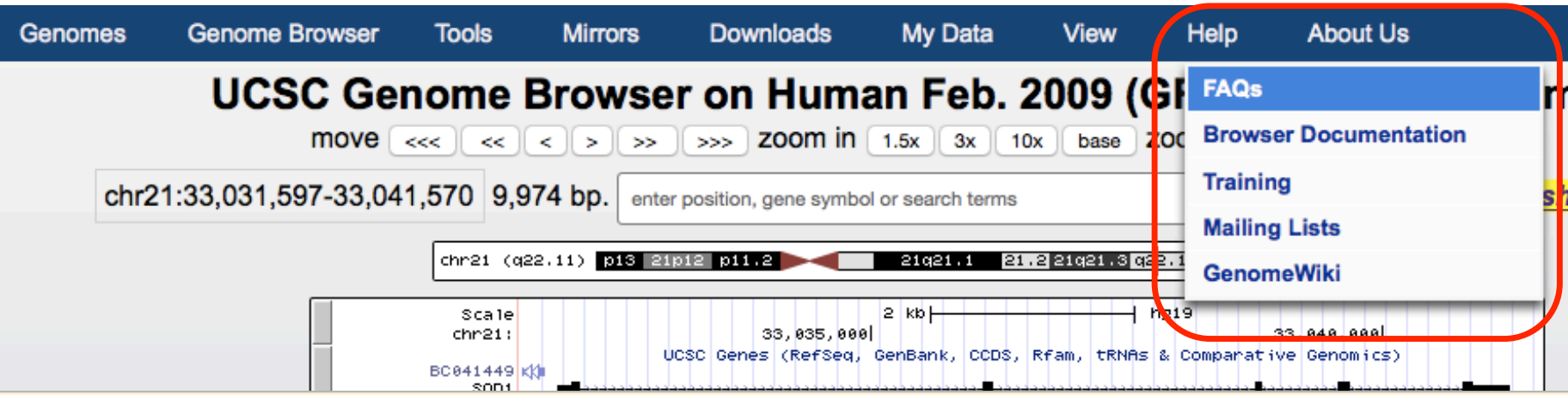
Indexed File Formats

- Only displayed portions of files transferred to UCSC
- Display large files (would time out)
- File + index on your web-accessible server (http, https, or ftp)
- Faster display
- More user control

File Formats



File Formats



UCSC Genome Bioinformatics

Home - Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - Help

FAQ Table of Contents

This page contains responses to questions frequently asked by our user community and subscribers to the Genome Browser maili

- Display Problems
- Assembly Releases and Versions
- Data and Downloads
- Genome Browser Tracks
- Custom Annotation Tracks
- Data File Formats
- Blat
- ENCODE Resources and FAQ

File Formats

UCSC Genome Bioinformatics

Home - Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - Help

Frequently Asked Questions: Data File Formats

General formats:

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- bigBed format
- bigGenePred table format
- bigWig format

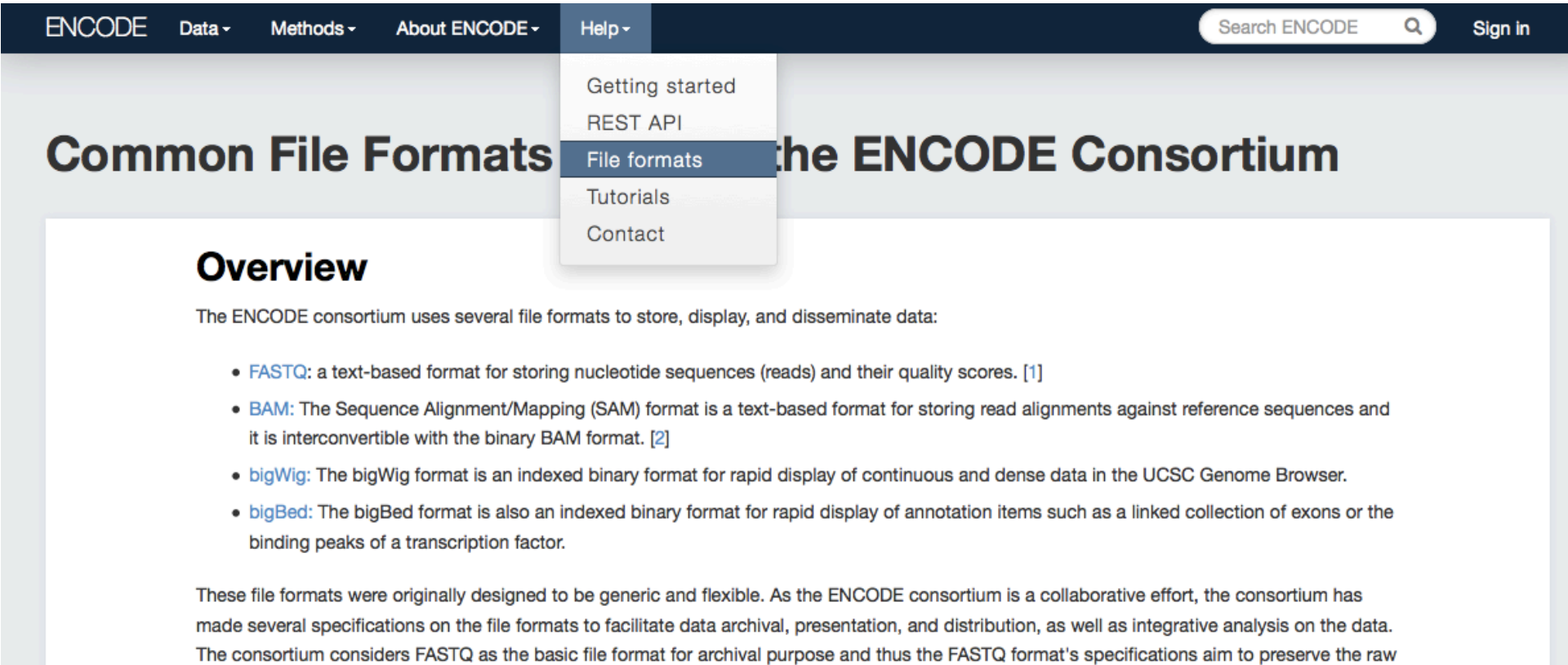
ENCODE-specific formats:

- ENCODE broadPeak format
- ENCODE gappedPeak format
- ENCODE narrowPeak format
- ENCODE pairedTagAlign format
- ENCODE peptideMapping format
- ENCODE RNA elements format
- ENCODE tagAlign format

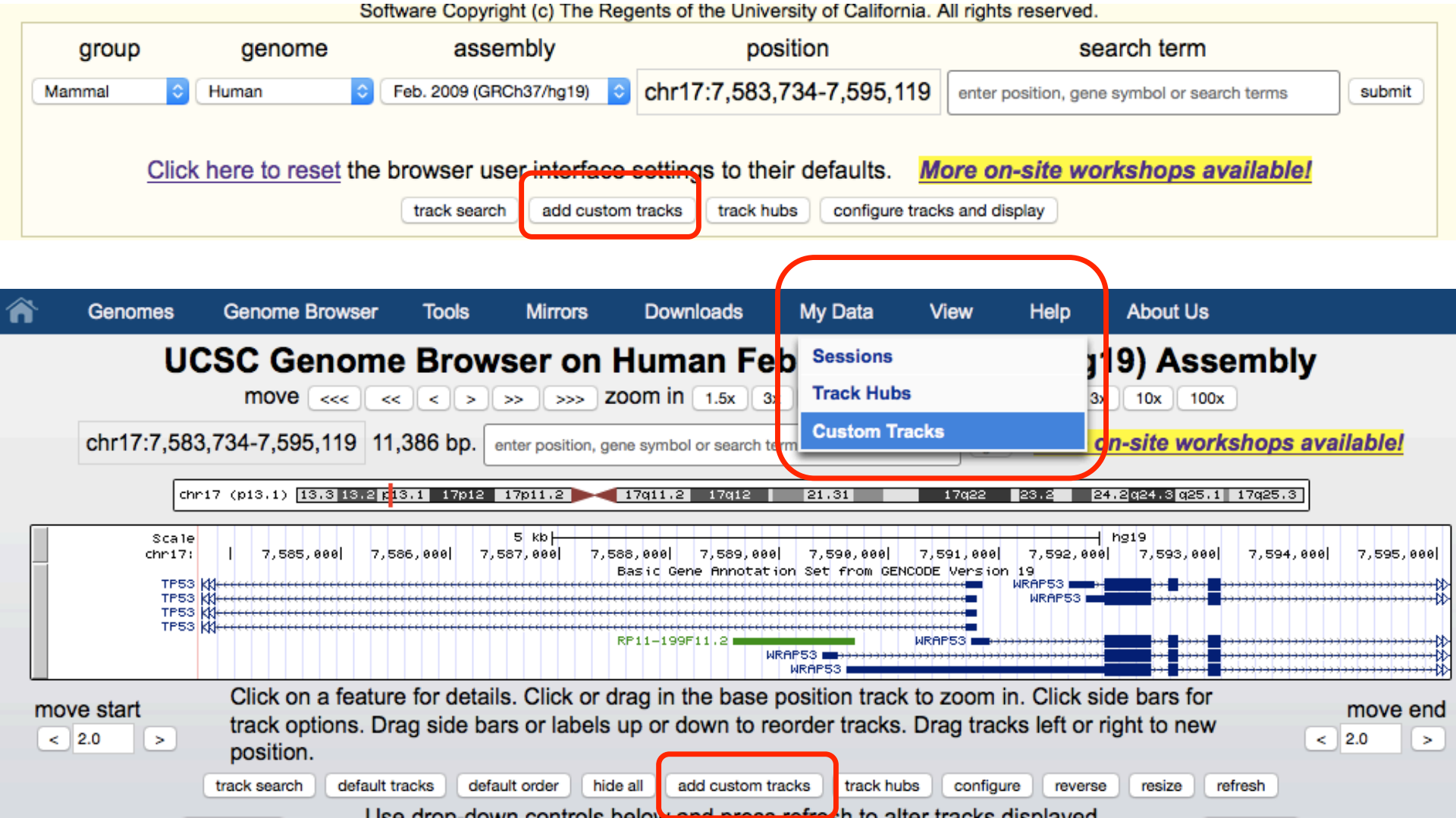
File Formats

www.encodeproject.org/help/file-formats/

Help → File formats

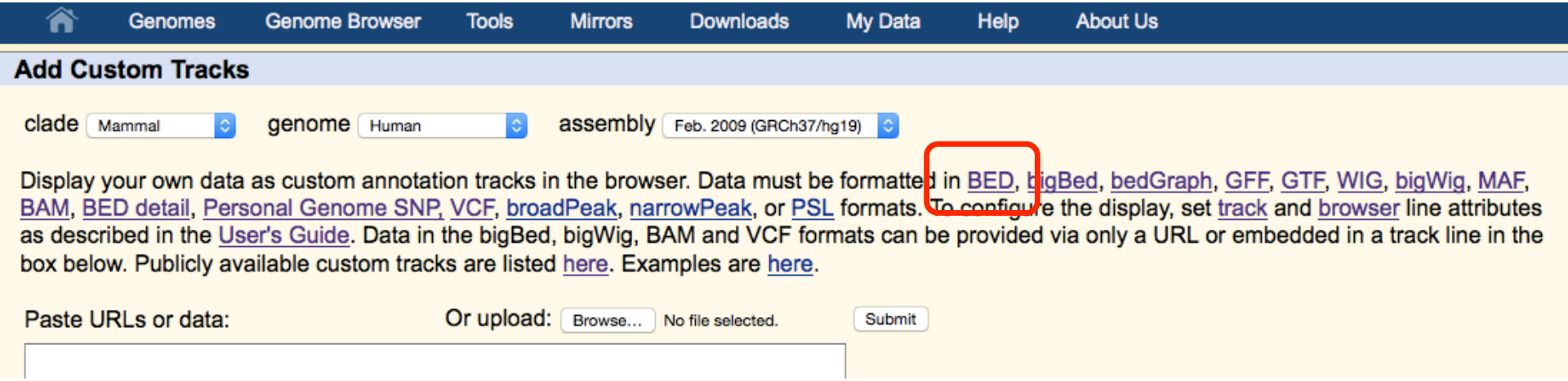


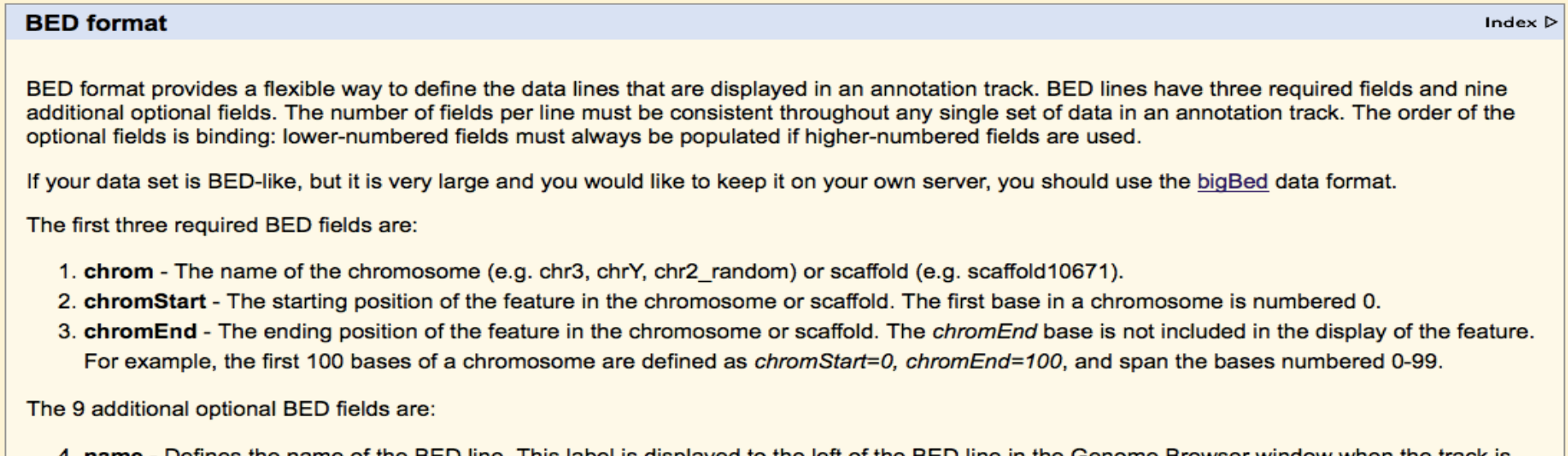
Custom Tracks



Custom Tracks

genome.ucsc.edu/cgi-bin/hgCustom





Custom Tracks

genome.ucsc.edu/cgi-bin/hgCustom

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

Add Custom Tracks

clade Mammal genome Human assembly Feb. 2009 (GRCh37/hg19)

Display your own data as custom annotation tracks in the browser. Data must be formatted in BED, bigBed, bedGraph, GFF, GTF, WIG, bigWig, MAF, BAM, BED detail, Personal Genome SNP, VCF, broadPeak, narrowPeak, or PSL formats. To configure the display, set track and browser line attributes as described in the User's Guide. Data in the bigBed, bigWig, BAM and VCF formats can be provided via only a URL or embedded in a track line in the box below. Publicly available custom tracks are listed here. Examples are here.

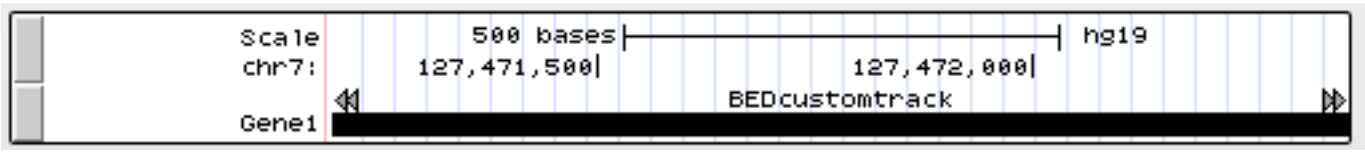
Paste URLs or data: Or upload: Browse... No file selected. Submit

```
track name="BED_custom_track"
chr7    127471196  127472363  Gene1
```

Clear

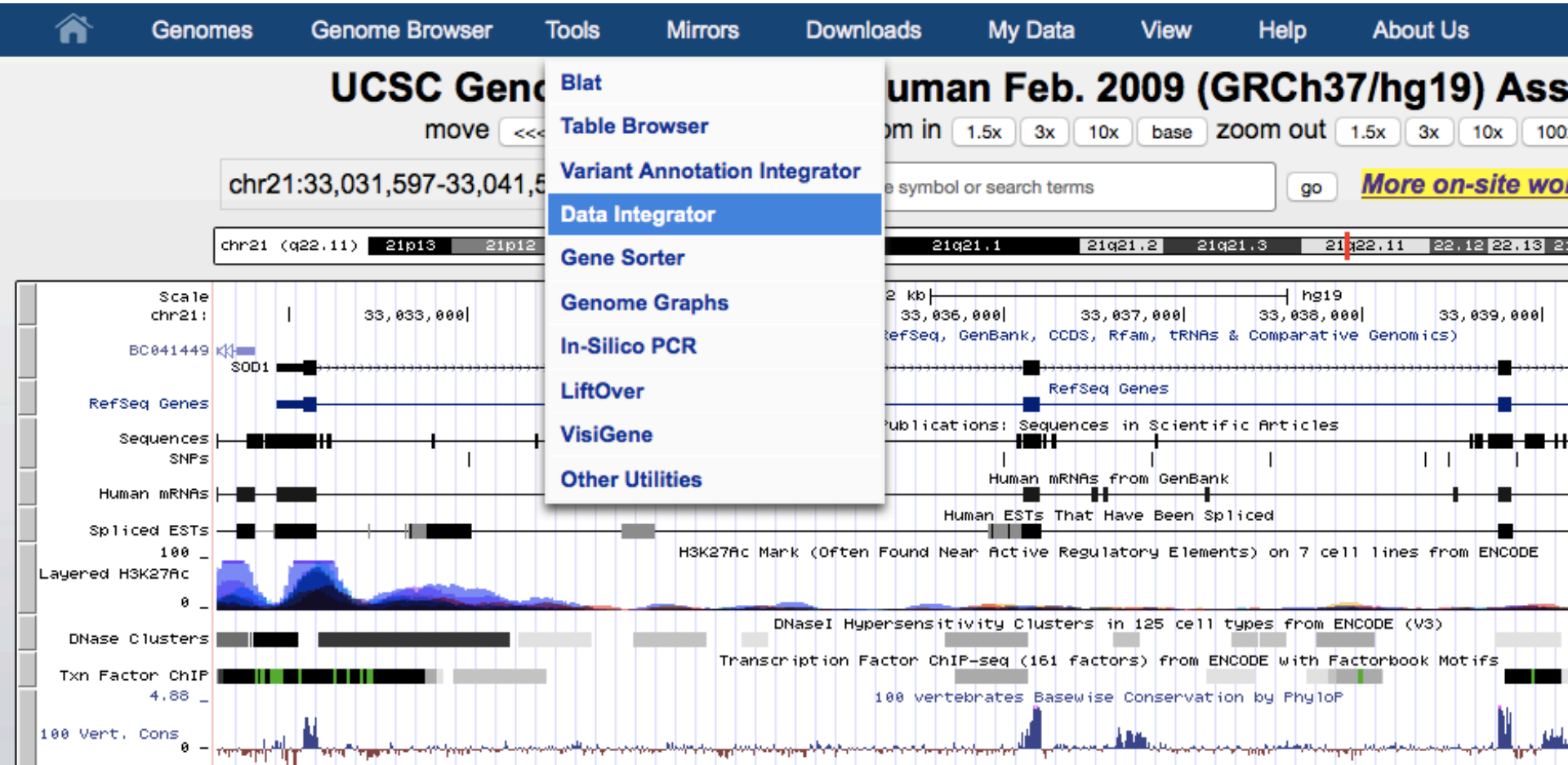
Optional track documentation: Or upload: Browse... No file selected.

Clear



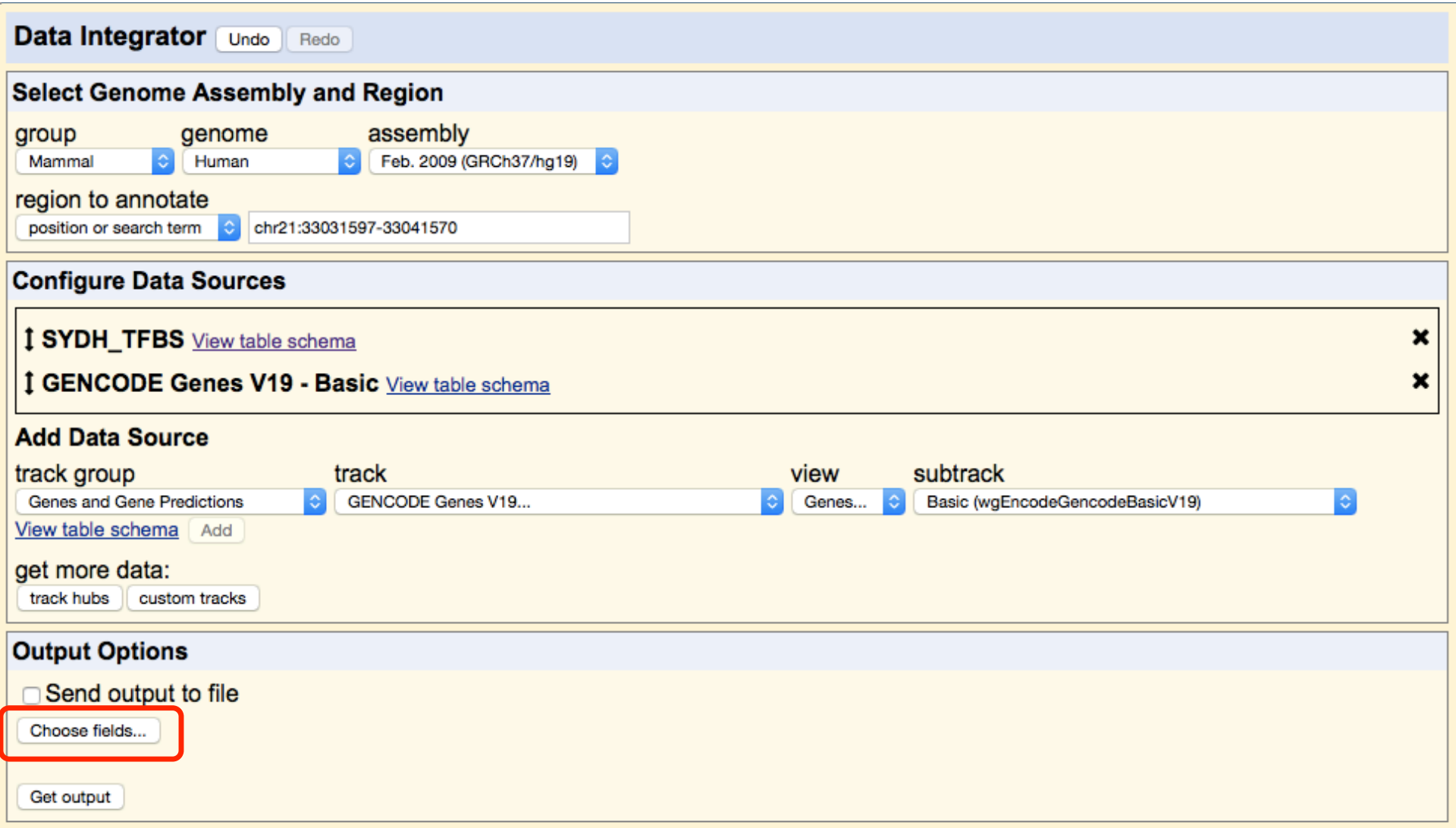
Annotating your data: BED

Tools → Data Integrator

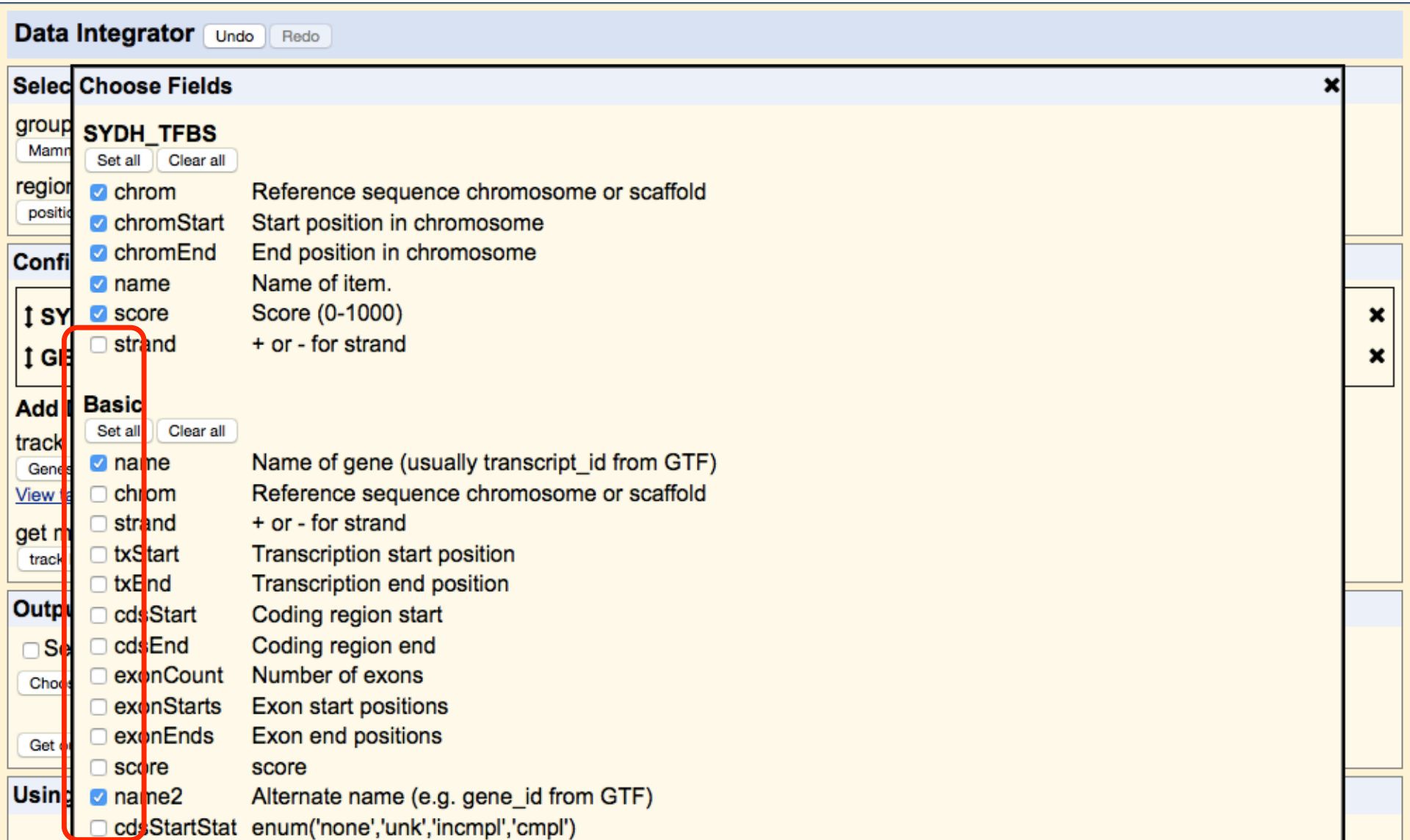


Data Integrator

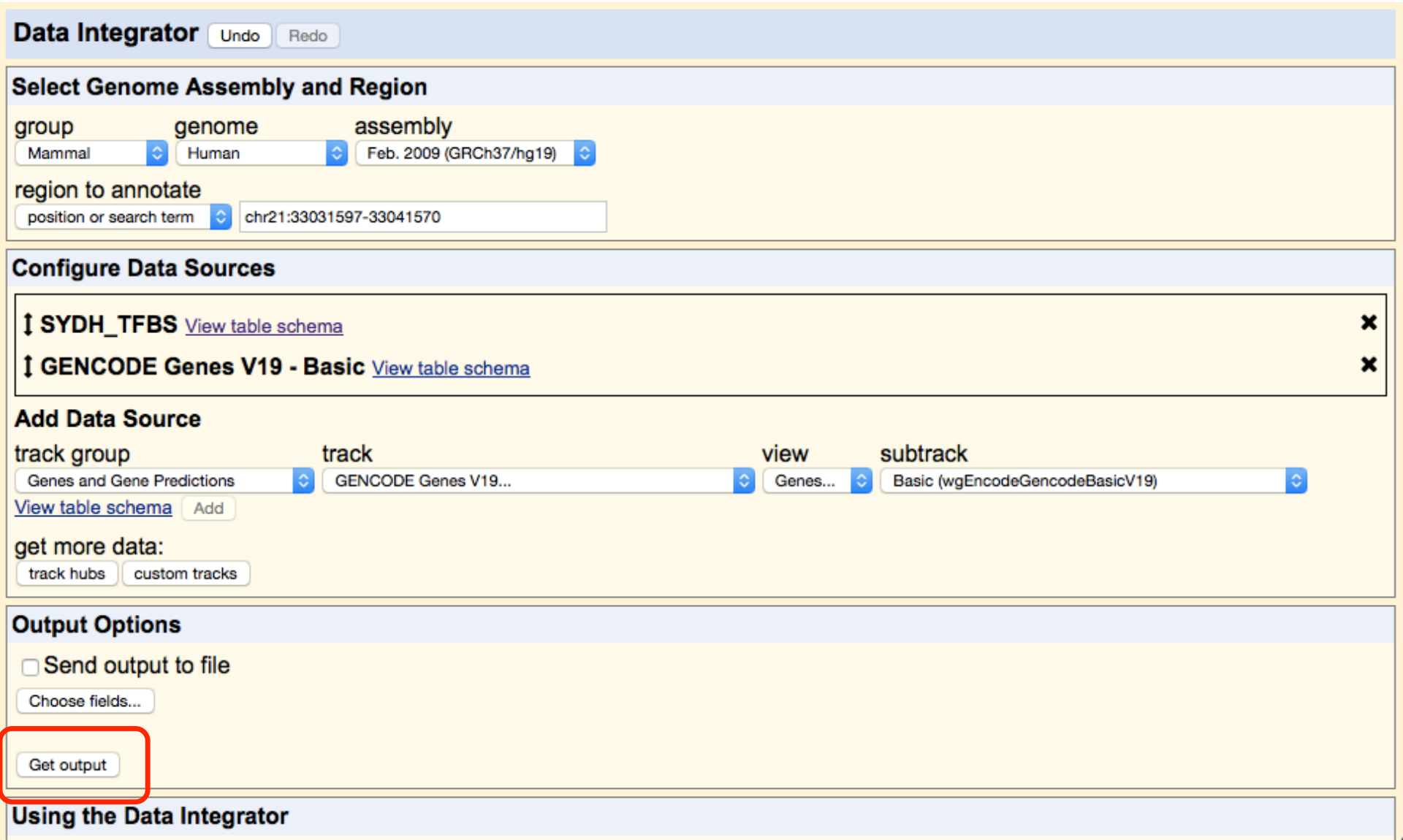
genome.ucsc.edu/cgi-bin/hgIntegrator



Data Integrator



Data Integrator



Data Integrator

```
#ct_SYDHTFBS_4733.chrom ct_SYDHTFBS_4733.chromStart
ct_SYDHTFBS_4733.chromEnd       ct_SYDHTFBS_4733.name   ct_SYDHTFBS_4733.score
wgEncodeGencodeBasicV19.name    wgEncodeGencodeBasicV19.name2
chr21   33031473        33032186        .       608     ENST00000449339.1
AP000253.1
chr21   33031473        33032186        .       608     ENST00000270142.6       SOD1
chr21   33031473        33032186        .       608     ENST00000389995.4       SOD1
chr21   33031473        33032186        .       608     ENST00000470944.1       SOD1
```

Annotating your VCF file

1. Make a VCF custom track
2. Go to the Variant Annotation Integrator
3. Choose your track
4. Add annotations

Remotely Hosted Custom Tracks

- Put data file (bigBed/bigWig/BAM/VCF, etc) in internet accessible location
- Must have: 1. track info, 2. bigDataUrl
- VCF example:

```
track type=vcfTabix
name="VCF_Example"
description="VCF Ex. 1: 1000 Genomes phase 1
interim SNVs"
bigDataUrl=
http://hgwdev.cse.ucsc.edu/~pauline/presentations/
vcfExample.vcf.gz
```

Variant Annotation Integrator

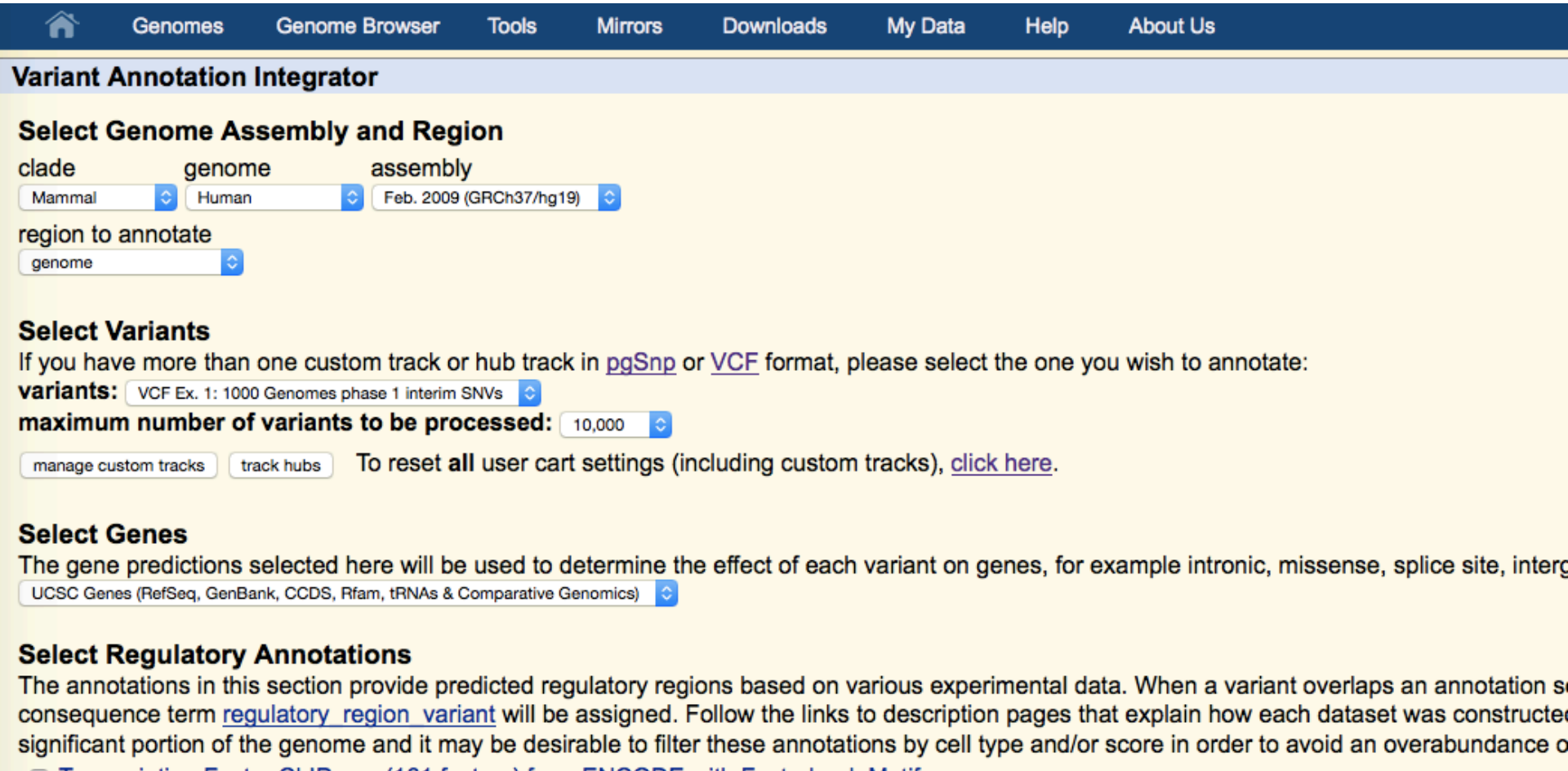
- Upload pgSnp or VCF custom track
- Associate UCSC annotations with your uploaded variant calls
- Add dbSNP info if dbSNP identifier found
- Select custom track and VAI options

Variant Annotation Integrator

Tools → Variant Annotation Integrator

Variant Annotation Integrator

genome.ucsc.edu/cgi-bin/hgVai



Track Data Hubs

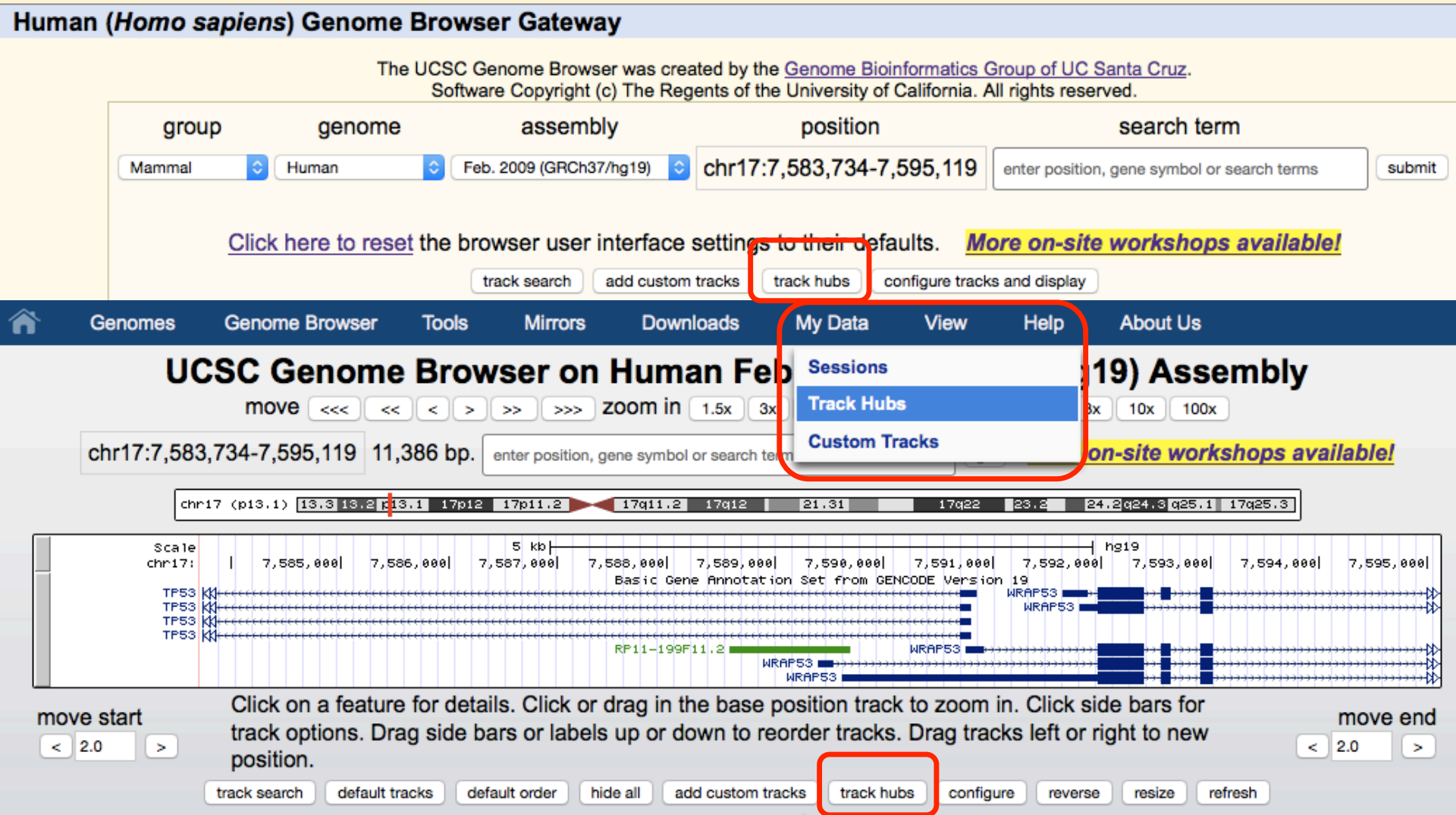
- Remotely hosted
- Data persistence
- File formats:

  bigBED, bigWig, BAM, VCF
- Track organization:

  groups, supertracks
- multiWigs
- Assembly hubs

Track Hubs

My Data → **Track Hubs**



Track Hubs

genome.ucsc.edu/cgi-bin/hgHubConnect

My Data → Track Hubs

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

Track Data Hubs

Track data hubs are collections of external tracks that can be imported into the UCSC Genome Browser. Hub tracks show up under the
label bar on the main browser page, as well as on the configure page. For more information, see the User's Guide.To import a public hu
"Connect" button below.

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

Public Hubs | My Hubs

Enter search terms to find in public track hub description pages:

Search Public Hubs

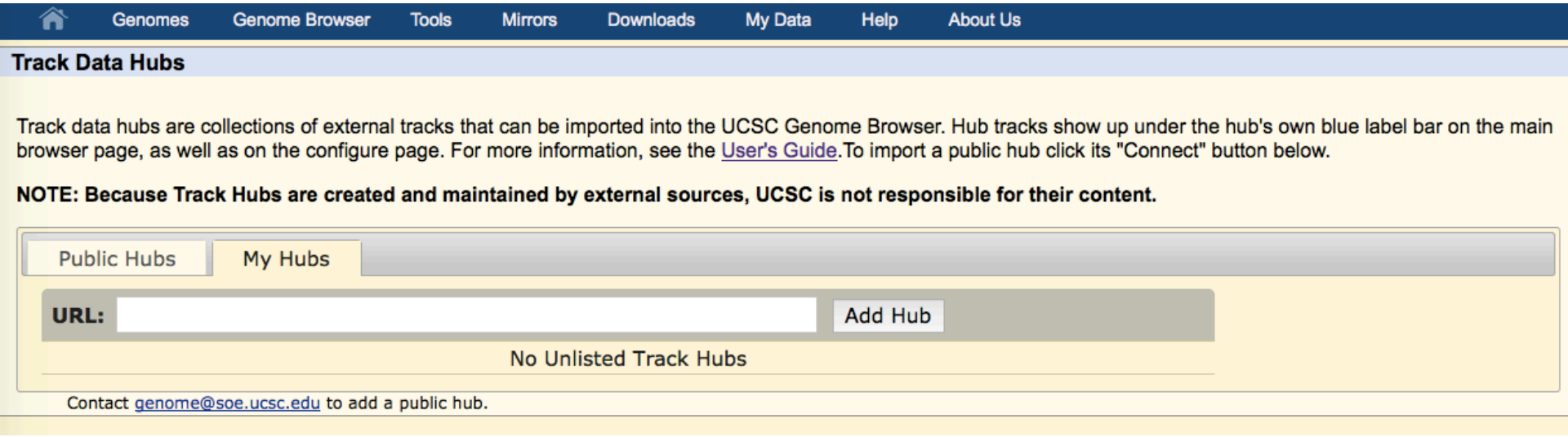
Clicking Connect redirects to the gateway page of the selected hub's default assembly.

Display	Hub Name	Description	Assemblies
Connect	Roadmap Epigenomics Data Complete Collection at Wash U VizHub	Roadmap Epigenomics Human Epigenome Atlas Data Complete Collection, VizHub at Washington University in St. Louis	hg19
		UMassMed H3K4me3 ChIP-seq data for	

My Hubs

genome.ucsc.edu/cgi-bin/hgHubConnect

My Data → Track Hubs



Make Your Own Track Hub

You will need:

- Data (compressed binary index formats: bigBed, bigWig, BAM, VCF)
- Text files to define properties of the track hub
- Internet-enabled web/ftp server
- *Assembly Hubs:*

  a twoBit sequence file

Track Hubs

genome.ucsc.edu/cgi-bin/hgHubConnect

My Data → Track Hubs

myHub/ - directory containing track hub files

hub.txt - a short description of hub properties

genomes.txt - list of genome assemblies included

hg19/ - directory of data for the hg19 human assembly

Data files! BAM, bigBed, bigWig, VCF

An Example Assembly Hub

An Arabidopsis hub:

http://genome-test.cse.ucsc.edu/~pauline/hubs/Plants/hub.txt

Acknowledgements

UCSC Genome Browser team

- David Haussler – co-PI
- Jim Kent – Browser Concept, BLAT, Team Leader, PI
- Bob Kuhn – Associate Director, Outreach – co-PI
- Donna Karolchik, Ann Zweig – Project Management

Engineering

Angie Hinrichs
Kate Rosenbloom
Hiram Clawson
Galt Barber
Brian Raney
Max Haeussler

QA, Docs, Support

Katrina Learned
Pauline Fujita
Luvina Guruvadoo
Steve Heitner
Brian Lee
Jonathan Caspar
Matt Speir

Sys-admins

Jorge Garcia
Erich Weiler
Gary Moro

THE GB TEAM

CENTER FOR BIOMOLECULAR SCIENCE & ENGINEERING

Baskin Engineering UC SANTA CRUZ

UNIVERSITY OF CALIFORNIA SANTA CRUZ | THE ORIGINAL AUTHORITY ON QUESTIONING AUTHORITY

Funding Sources

National Human Genome Research Institute (NHGRI)

National Cancer Institute (NCI)

National Institute for Dental and Cranio-Facial Research (NIDCR)

National Institute for Child Health and Human Development (NICHD)

QB3 (UCBerkeley, UCSF, UCSC)

American Recovery and Reinvestment Act (ARRA) stimulus funds

UNIVERSITY OF CALIFORNIA SANTA CRUZ | Genomics Institute

genome.ucsc.edu

THANK YOU!

Exercises

1. Load example BED and VCF tracks via url
2. Look at custom track data by pasting url into a web browser.
3. Annotate the TFBS custom track using the Data Integrator.
4. Annotate the VCF custom track using the Variant Annotation Integrator.

Exercise 1

Load example BED and VCF tracks via url

1. Go to the Custom tracks menu
   - My Data -> Custom Tracks
2. Input this url: http://bit.ly/customtracks (note that you must include the ”http” part of this url or you will get an error) and click [submit].
3. Click the [Go to genome browser] button.
4. Once in the main Browser, jump to this position:
   - chr21:33,034,804-33,037,719
5. See if you can drag your 2 custom tracks to the top of the display

Exercise 2

Exploring your BED and VCF tracks

1. Now that you have 2 custom tracks loaded, take a look at the data by pasting that same url into a web browser:
2. These custom tracks are actually data copied from some existing tracks, see if you can find them, turn them on, and observe that the original tracks and custom tracks look the same in the browser:
   - Track 1 (BED format): Group (Regulation), Super Track (ENC TF Binding), Track (SYDH TFBS)
   - Track 2 (VCF format): Group (Variation), Track (1000G Ph1 Vars)
3. Navigate to this position for best comparison (esp. for the VCF track): chr21:33,034,804-33,037,719

Exercise 3

Annotate your BED with the Data Integrator

1. Go to the Data Integrator
2. Once there select:
    1. Region to annotate: chr21:33031597-33041570
    2. Add data source: group (custom tracks), track (SYDH…) [click add]
3. Now choose which annotations you want to add by [add]ing more tracks to the list – ex:
    1. *Find the genes that overlap with your regions:* group (Genes and Gene Prediction), track (GENCODE V19), view (Genes), subtrack (Basic) [add]
    2. Find the SNPs that overlap with your regions: group (Variation), track (Common SNPs) [add]

    Choose which fields to include in your output:

    Output options -> Choose fields [Done] -> [get output]

Exercise 4

Annotate your VCF with the Variant Annotation Integrator

1. Go to the Variant Annotation Integrator
   - Tools -> V.A.I.
2. Select Variants:
   - Variants: “VCF Ex. 1…”
3. Now choose which annotations you want to add:
   - To determine which gene regions your variants fall into, select a gene track (Select Genes = “Basic Gene Annotation Set… GENCODE”)
   - Add regulatory annotations: Under “Select Regulatory Annotations” click the “+” button to choose which TFs to include (or select none to include all binding sites)

Bonus Material!

Where to search

genome.ucsc.edu/cgi-bin/hgGateway

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

(*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

group	genome	assembly	position	search term
mal	Human	Feb. 2009 (GRCh37/hg19)	chr21:33,031,597-33,041,570	brc

submit

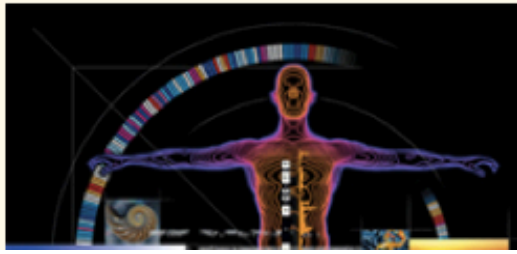
Click here to reset the browser user interface settings to their defaults. ***More on-site workshops available!***

track search | add custom tracks | track hubs | configure tracks and display

Genome Browser – hg19 assembly (sequences)

ruary 2009 human reference sequence (GRCh37) was produced by the Genome Reference
um. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

e position queries



Where to search

genome.ucsc.edu/cgi-bin/hgGateway

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

(*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

group | genome | assembly | position | search term

ammal | Human | Feb. 2009 (GRCh37/hg19) | chr21:33,031,597-33,041,570 | brc | submit

BRCA1 (Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 2, mRNA.)
BRCA2 (Homo sapiens breast cancer 2, early onset (BRCA2), mRNA.)
BRCC3 (Homo sapiens BRCA1/BRCA2-containing complex, subunit 3 (BRCC3), transcript variant 1, mRNA.)

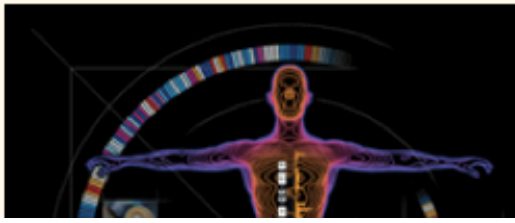
Click here to reset the browser user interface settings to their defaults.

track search | add custom tracks | track hubs | configure tracks

Genome Browser – hg19 assembly (sequences)

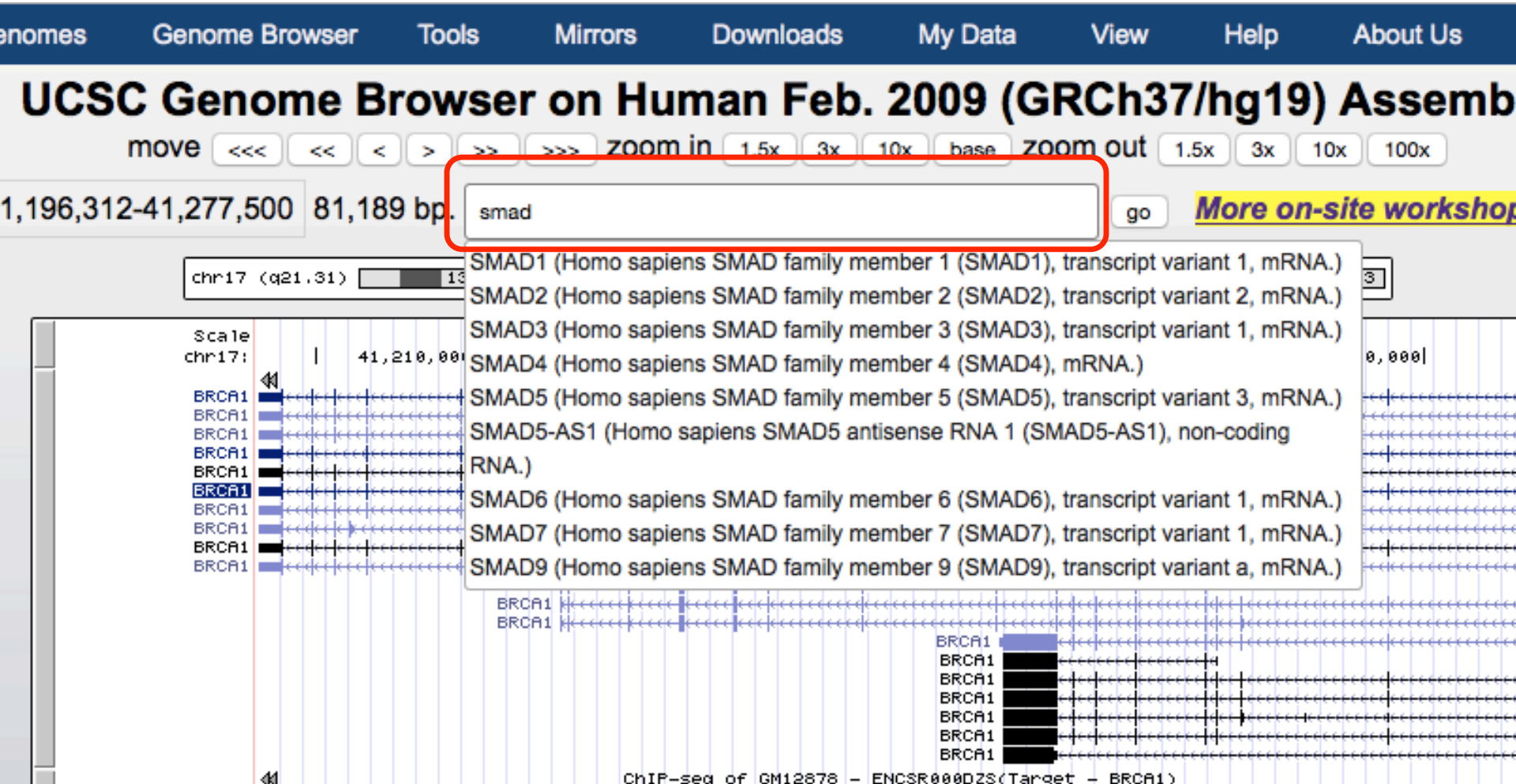
bruary 2009 human reference sequence (GRCh37) was produced by the Genome Reference tium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

e position queries



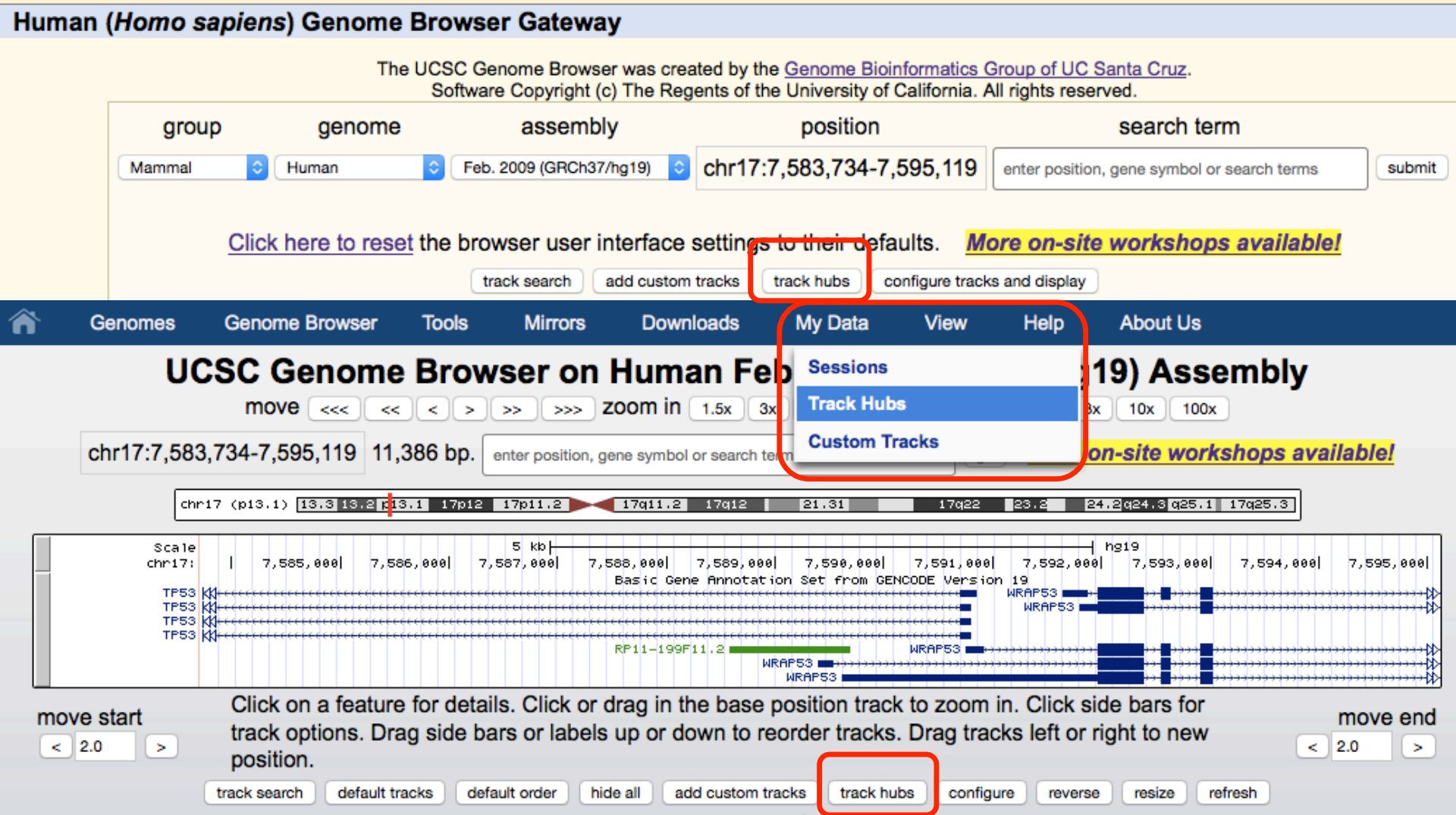
Where to search: Main Browser

genome.ucsc.edu/cgi-bin/hgTracks



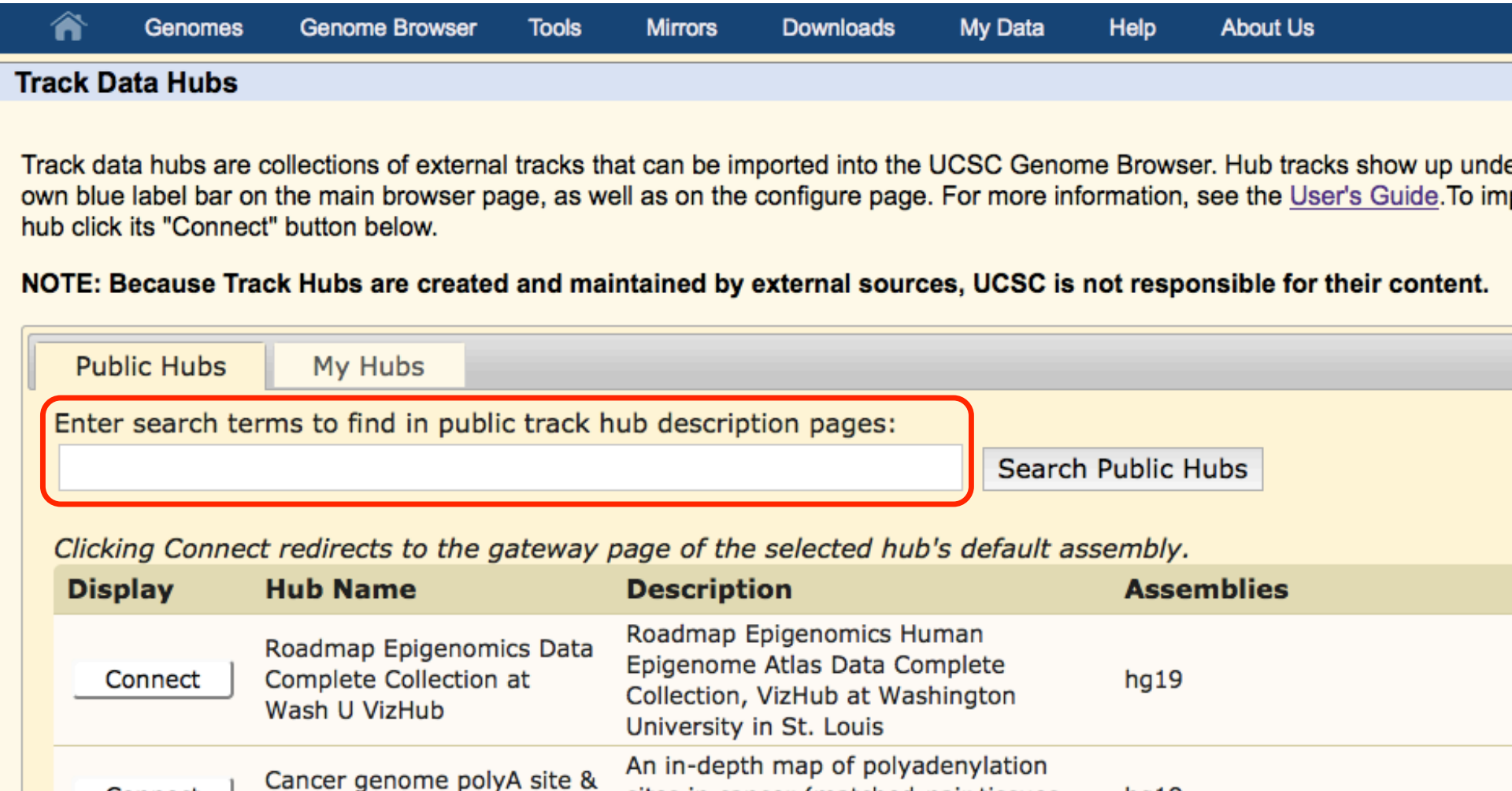
Public Hubs

My Data → **Track Hubs**

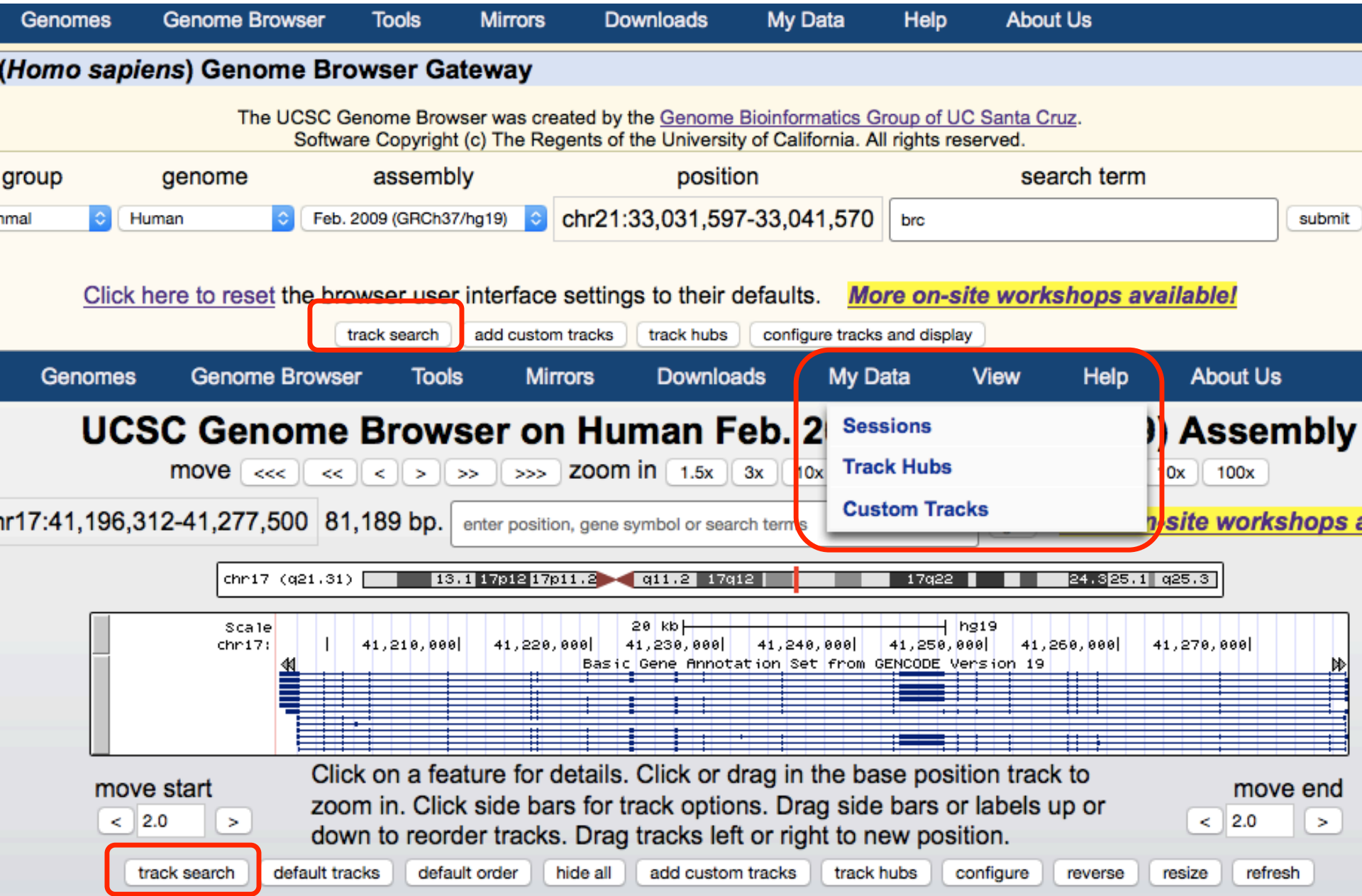


Where to search

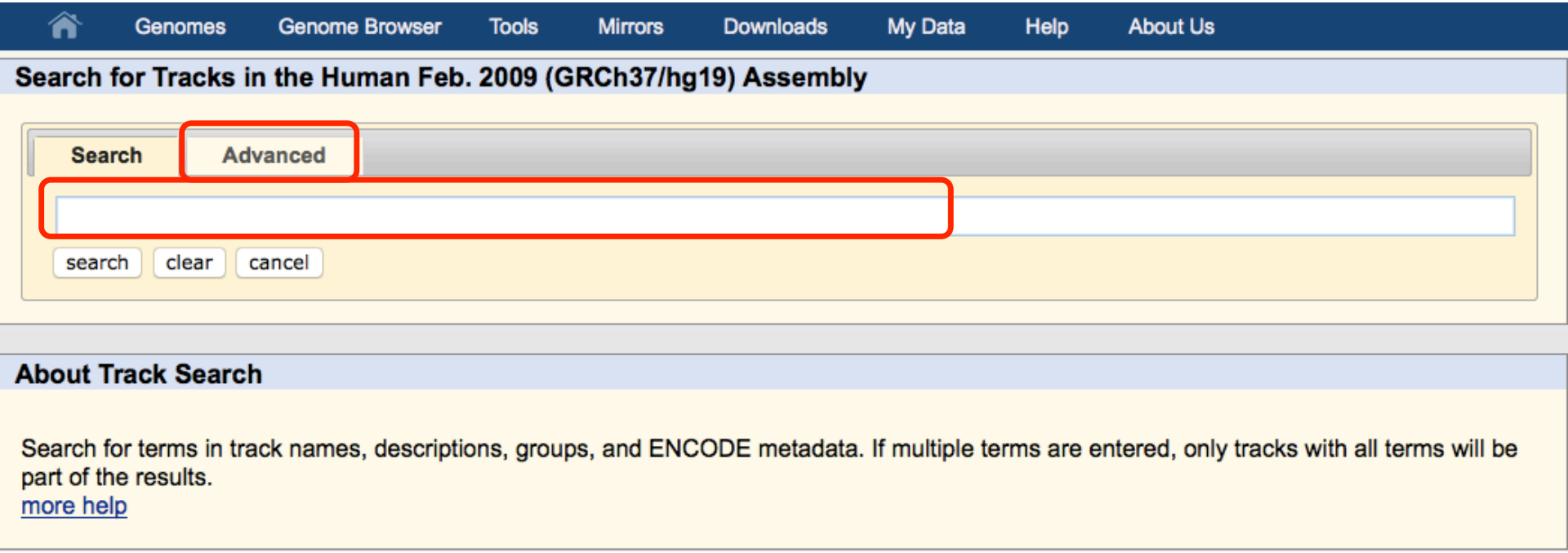
genome.ucsc.edu/cgi-bin/hgHubConnect



Track search



Track search



Track search