Genome Browsers

Peter FitzGerald. PhD

Head Genome Analysis Unit, CCR, NCI

Talk Outline

- Overview of Genome Browsers
- Different typed of Browsers
- Highlighting three specific Browsers
  - UCSC Genome Browser
  - IGB - Integrated Genome Browser
  - IGV - Integrative Genomics Viewer
- Conclusion

Overview of Genome Browsers

- Tools that provide a graphical view of genomic data
- Provide a view of the spatial relationship between “genes”
- Good for comparing and integrating different data sets
- Drilling down into the raw data (see what cutoffs are doing)

Three different types of browsers based on the way they handle data

- Web-based - Numerous examples but all have these common factors:
  - User interface is though a web browser
  - The data resides remotely (with local hooks possible)
  - Processing and data manipulation is done remotely
- Client-server model - Hybrid (IGB,IGV)
  - Browser runs locally but some the data (genome annotation etc.) resides remotely and some is local and processing and graphics is local.
- Local Client (Vendor Browsers)
  - All data and processing is done locally

Other Web-based Browsers

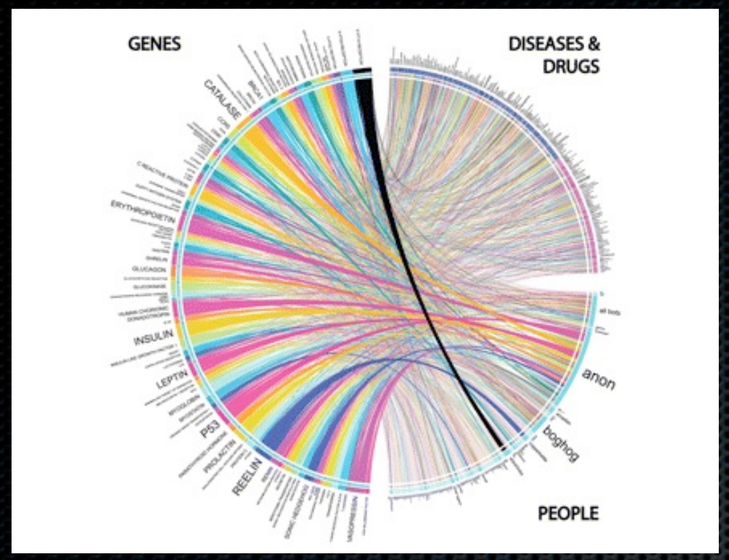
- NCBI Map Viewer
  - http://www.ncbi.nlm.nih.gov/mapview/
- Ensemble
  - http://www.ensembl.org/
- Genome Specific

Different browsers have the same underlying genomic sequence (**within the same version**), but annotations are frequently different

Beware of Version Changes

- Different browsers update on different schedules
- Different browsers have different archiving policies
- Annotations are more variable than sequence data
- Nomenclature is different

When comparing data from different sources always make sure your are looking at the same version.
UCSC “liftover” files can interconvert



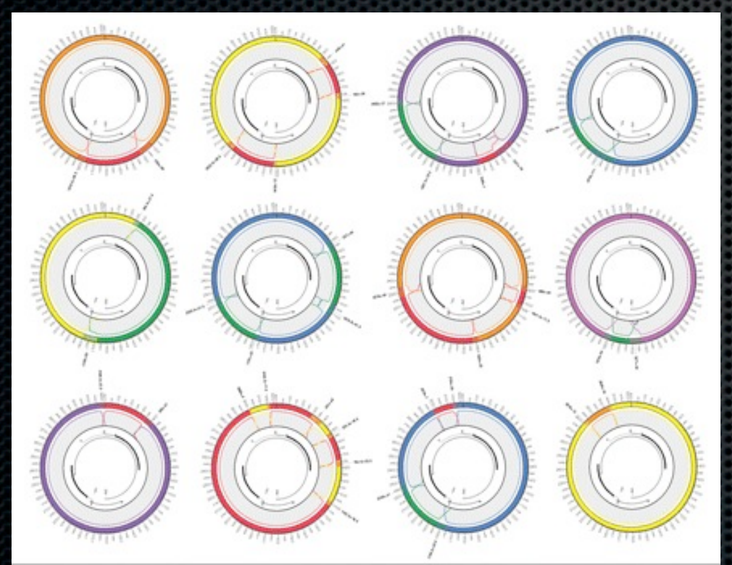
Volume 40 Issue D1 January 2012

The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection

http://nar.oxfordjournals.org/content/40/D1.toc

Database Summaries

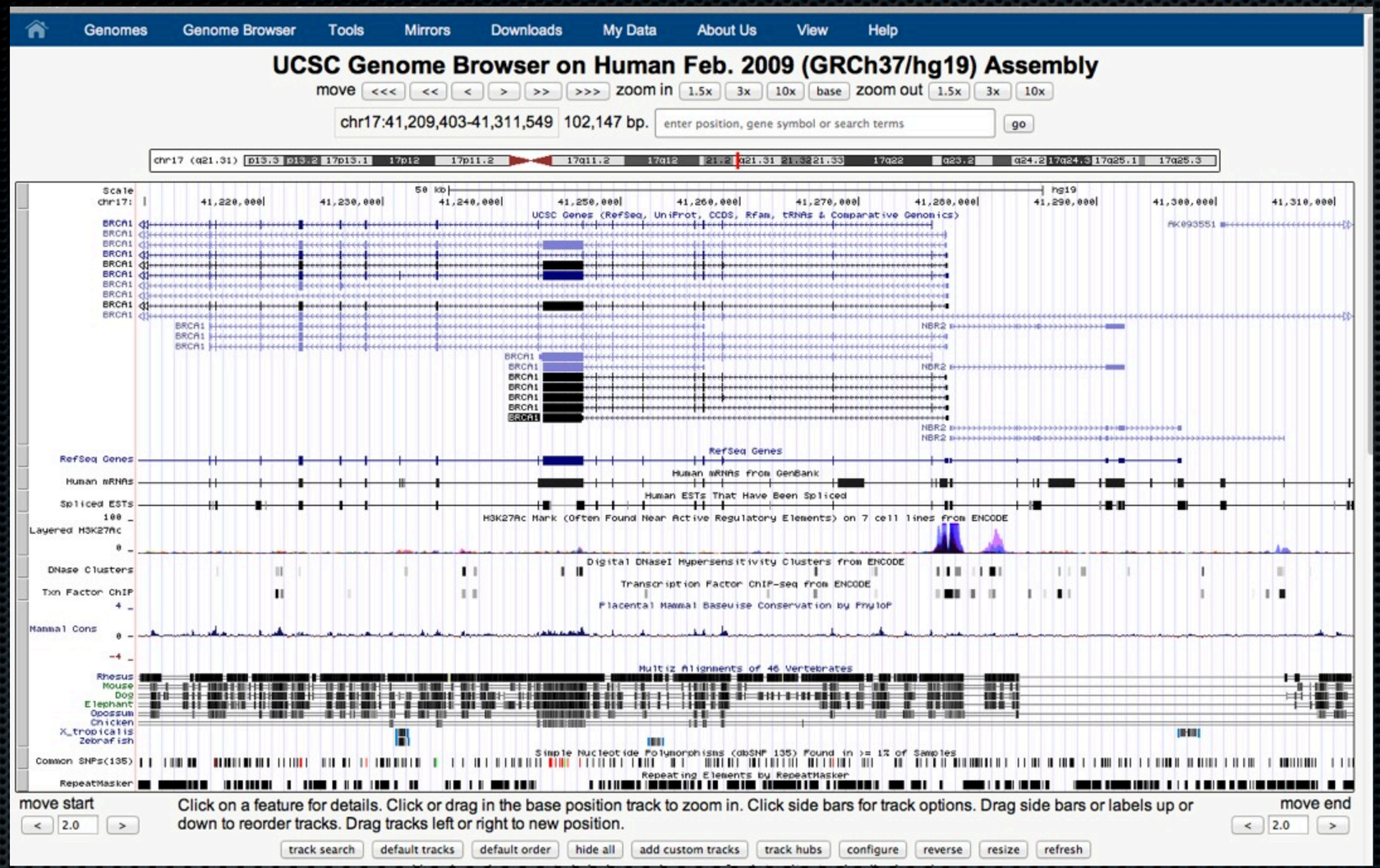
http://nar.oxfordjournals.org/content/40/D1/D1/suppl/DC1



Volume 40 Issue W1 July 2012

Web Server issue

http://nar.oxfordjournals.org/content/40/W1.toc



UCSC Genome Browser

http://genome.ucsc.edu/

Cover of *Nature* human genome issue, published on 15 February 2001.

Browser was launched in 2000 coinciding with the release of the Human Genome Sequence. Initially built and still managed by Jim Kent, then a graduate student,and David Haussler, professor of Computer Science (now Biomolecular Engineering) at the University of California, Santa Cruz.

UCSC Genome Browser

Other Flavors

- **UCSC Archive**
- **UCSC Microbial**
- **Local mirror of UCSC at NIH (restricted)**
  - NIH local Mirror (Subset) maintained by the Helix Sytems (CIT) http://genome.cit.nih.gov/

UCSC Genome Browser

Where to get help

- http://genome.ucsc.edu/FAQ/ **FAQ**
- http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html **Help**
  - **file formats**
- http://genome.ucsc.edu/contacts.html **Mailing Lists**
- http://genomewiki.ucsc.edu/ **Wiki**

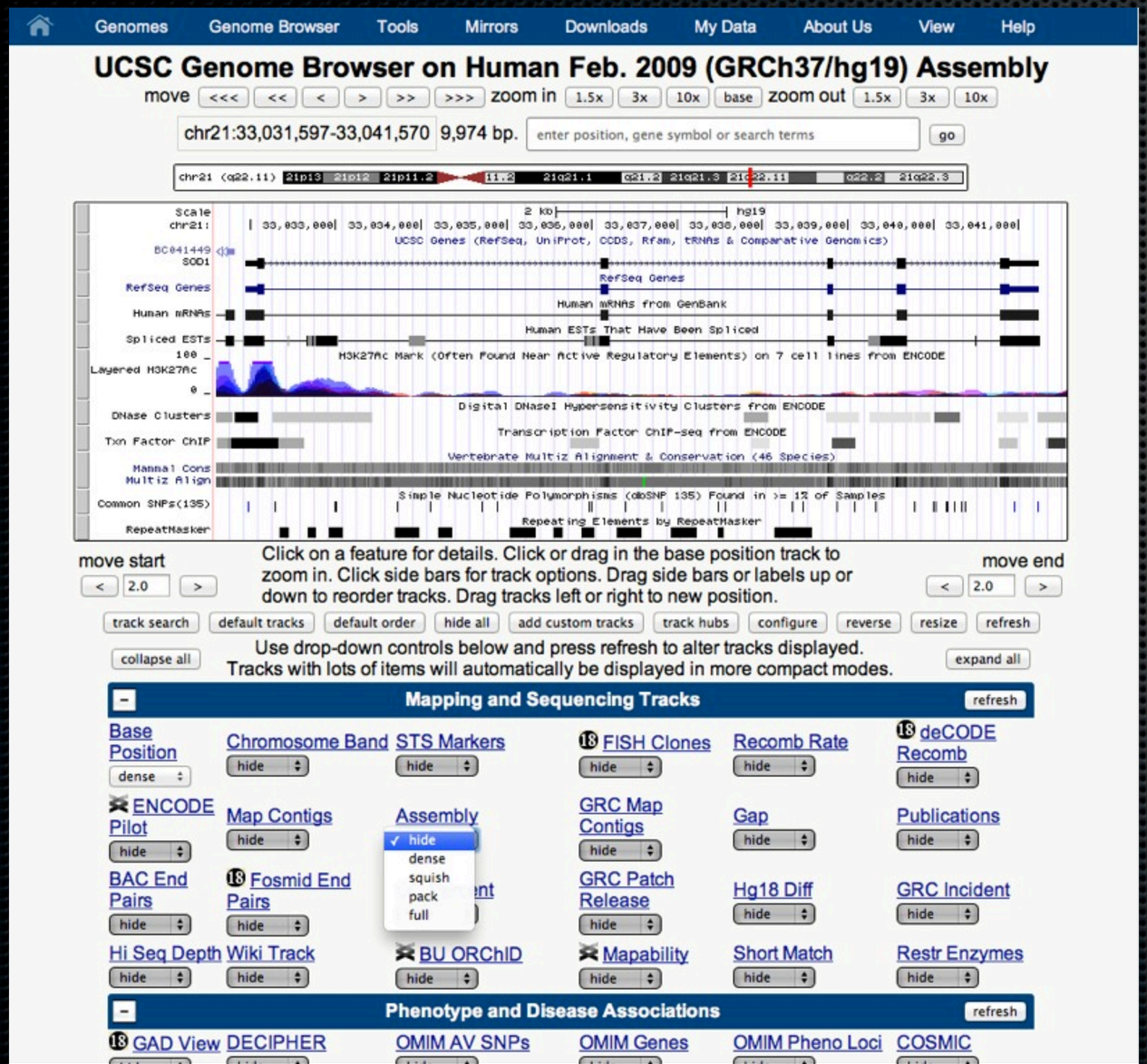
UCSC Genome Browser

Genomes 6-63

- ***Mammals***
- Human
- Chimp
- Gorilla
- Orangutan
- Gibbon
- Rhesus
- Marmoset
- Mouse
- Rat
- Naked mole-rat
- Guinea pig
- Rabbit
- Pig
- Sheep
- Cow
- Horse
- Cat
- Dog
- Panda
- Microbat
- Tenrec
- Elephant
- Opossum
- Tasmanian devil
- Wallaby
- Platypus
- ***Vertebrate***
- Chicken
- Turkey
- Zebra finch
- Medium ground finch
- Lizard
- Painted turtle
- X. tropicalis
- Zebrafish
- Tetraodon
- Fugu
- Stickleback
- Medaka
- Lamprey
- ***Deuterostome***
- Lancelet
- C.intestinalis
- S.purpuratus
- ***Insect***
- D. melanogaster
- D. simulans
- D. sechellia
- D. yakuba
- D. erecta
- D. ananassae
- D. pseudoobscura
- D. persimilis
- D. virilis
- D. mojavensis
- D. grimshawi
- A. gambiae
- A. mellifera
- ***Nematode***
- C. elegans
- C. brenneri
- C. briggsae
- C. remanei
- C. japonica
- P. pacificus
- ***Other***
- Sea Hare
- S.cerevisiae

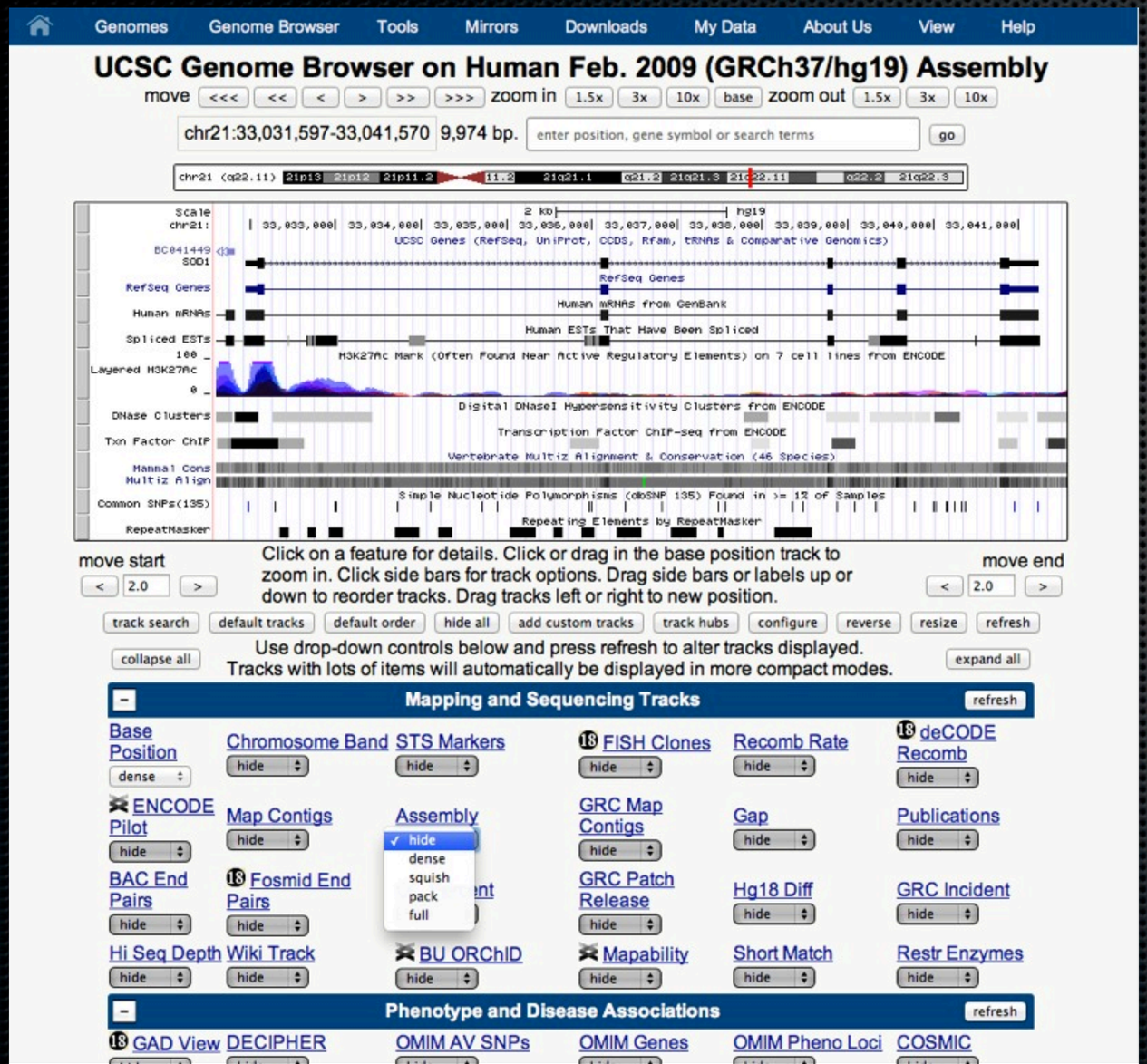
UCSC Genome Browser

Overview



UCSC Genome Browser

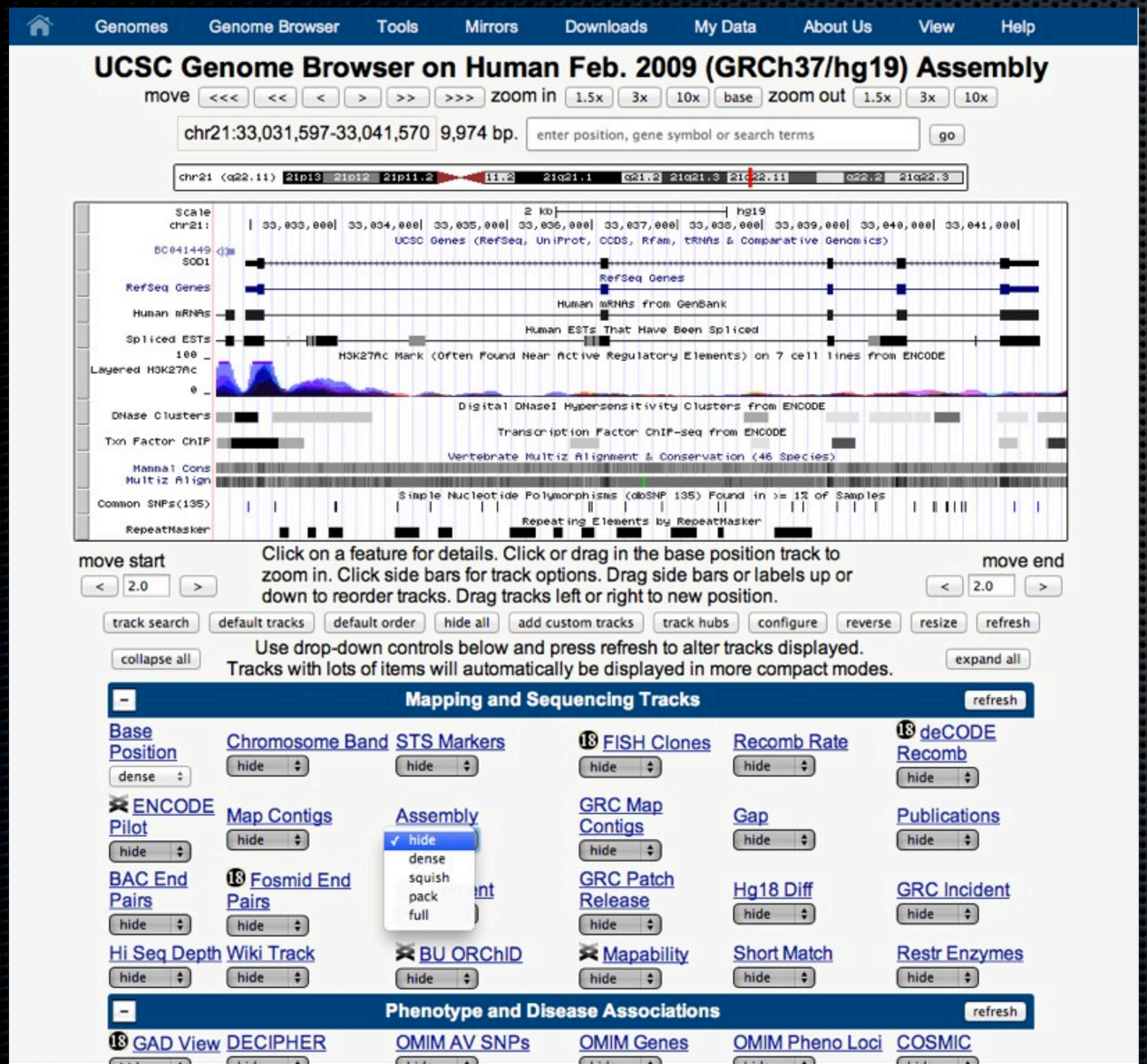
Overview



← Zoom level

UCSC Genome Browser

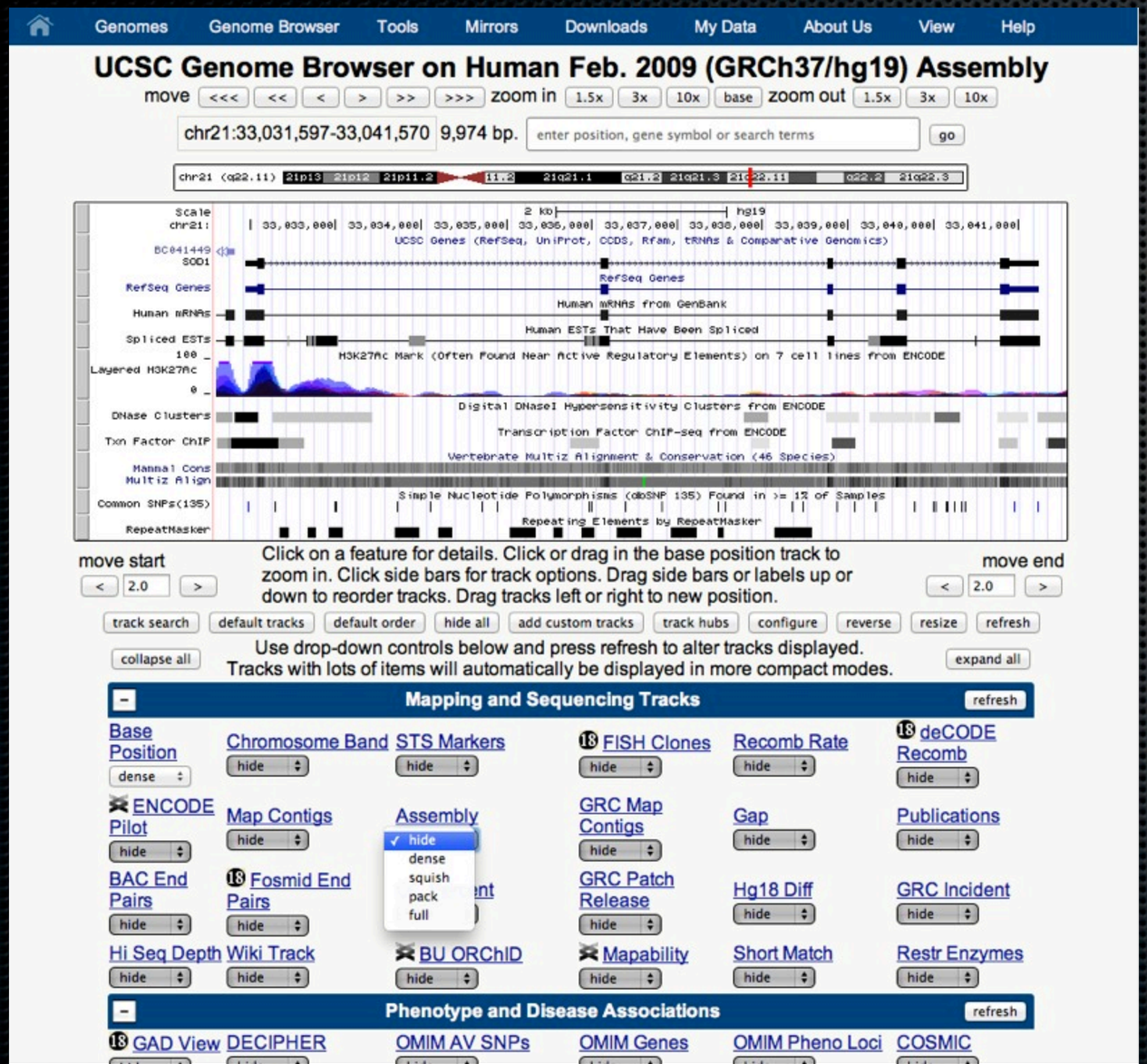
Overview



← Coordinates and search

UCSC Genome Browser

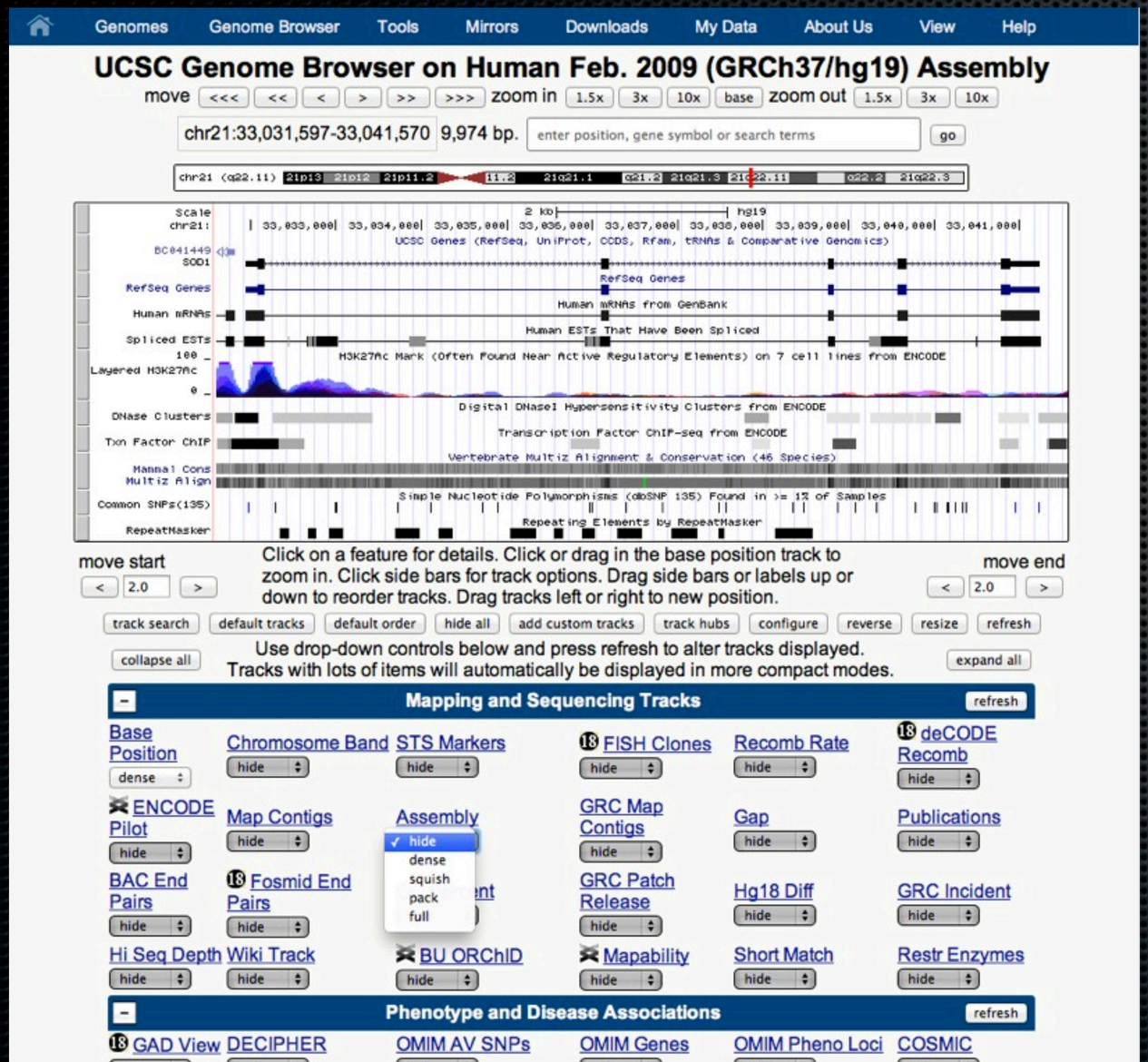
Overview



← Chromosome position and coordinates

UCSC Genome Browser

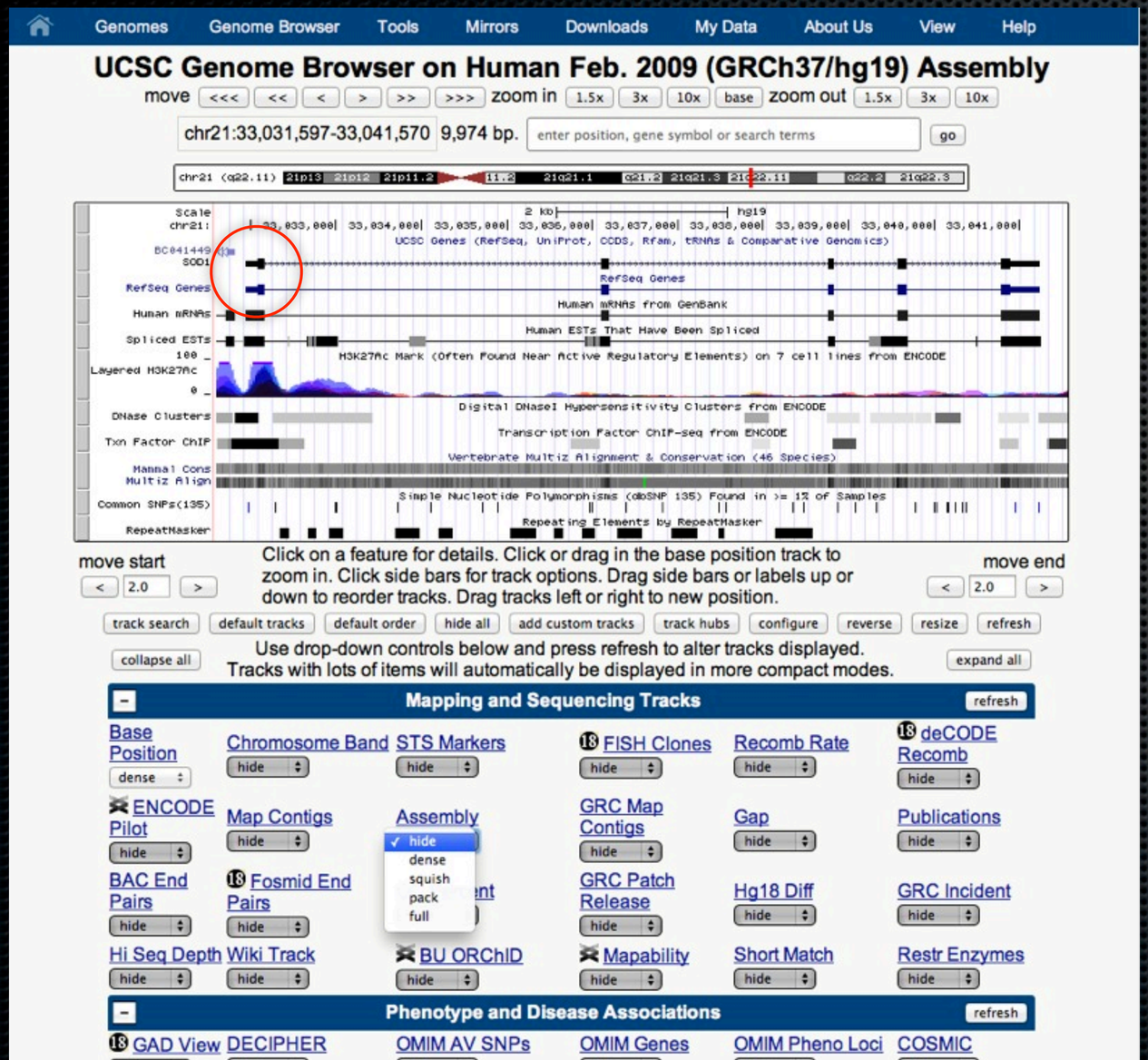
Overview



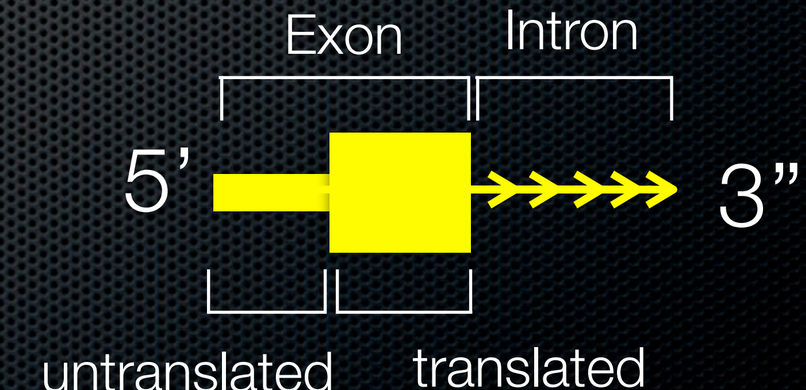
← Main Display (Image)

UCSC Genome Browser

Overview

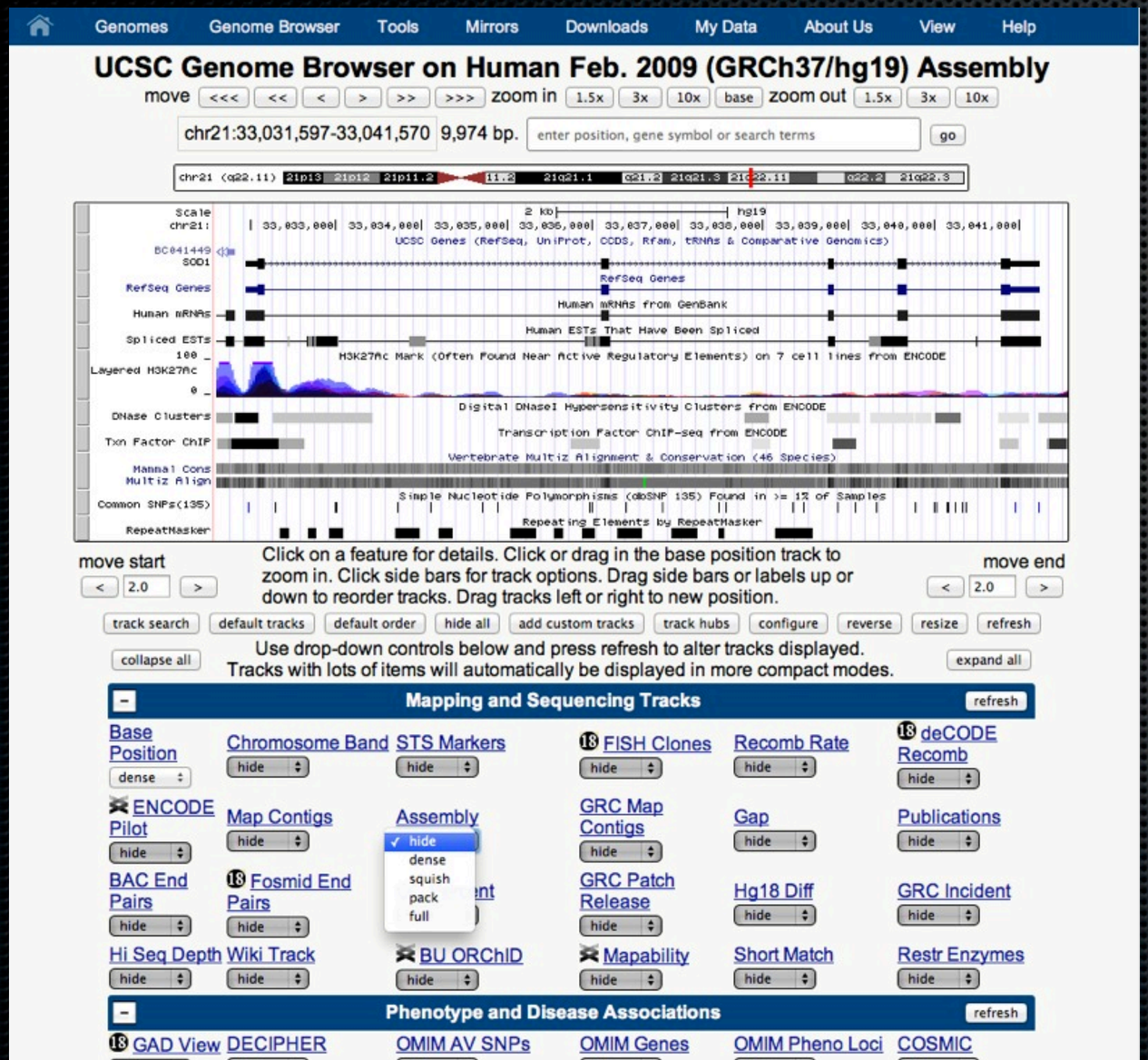


← Main Display (Image)



UCSC Genome Browser

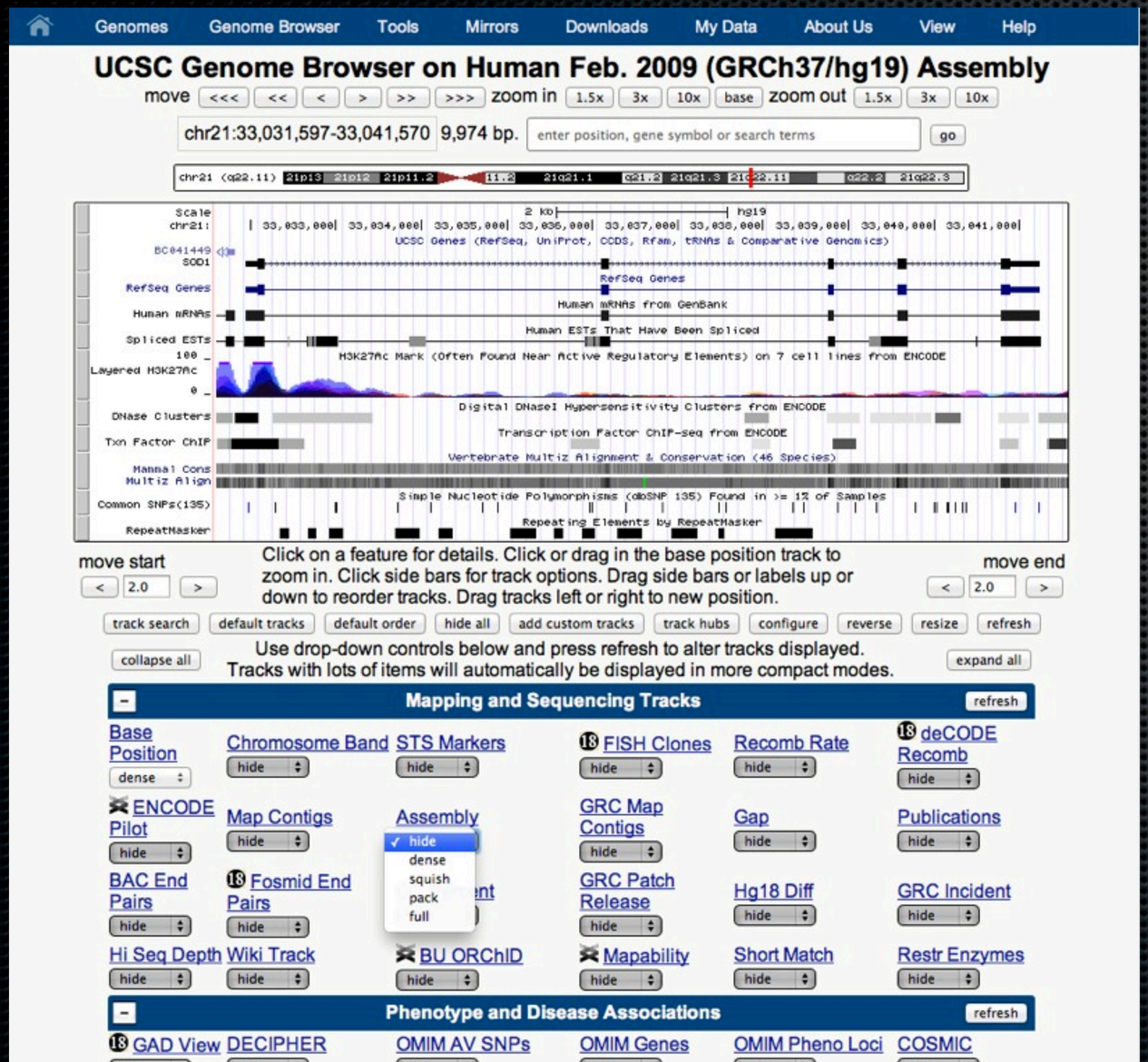
Overview



← Select controls

UCSC Genome Browser

Overview



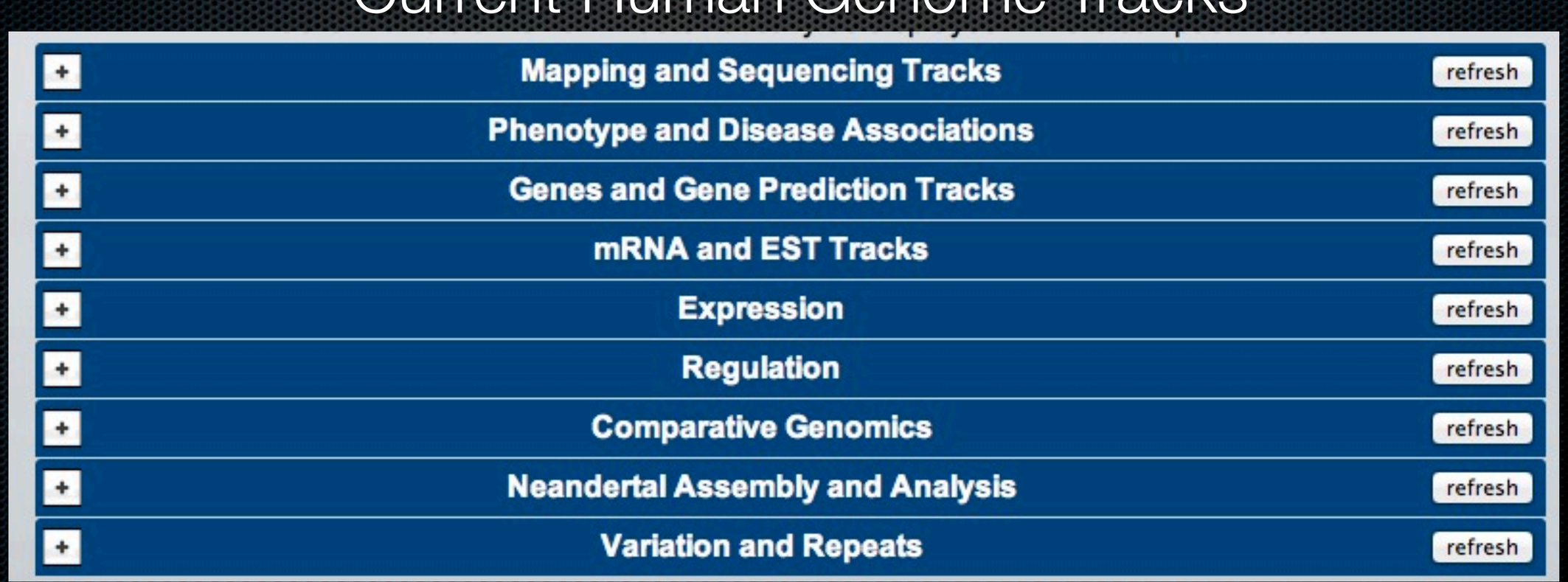
← Track controls

UCSC Genome Browser

Tracks

- Vary from Genome to Genome and possibly version
- Many current versions may lack tracks that you expect

Current Human Genome Tracks



UCSC Genome Browser

Tracks

Mapping and Sequencing Tracks — refresh

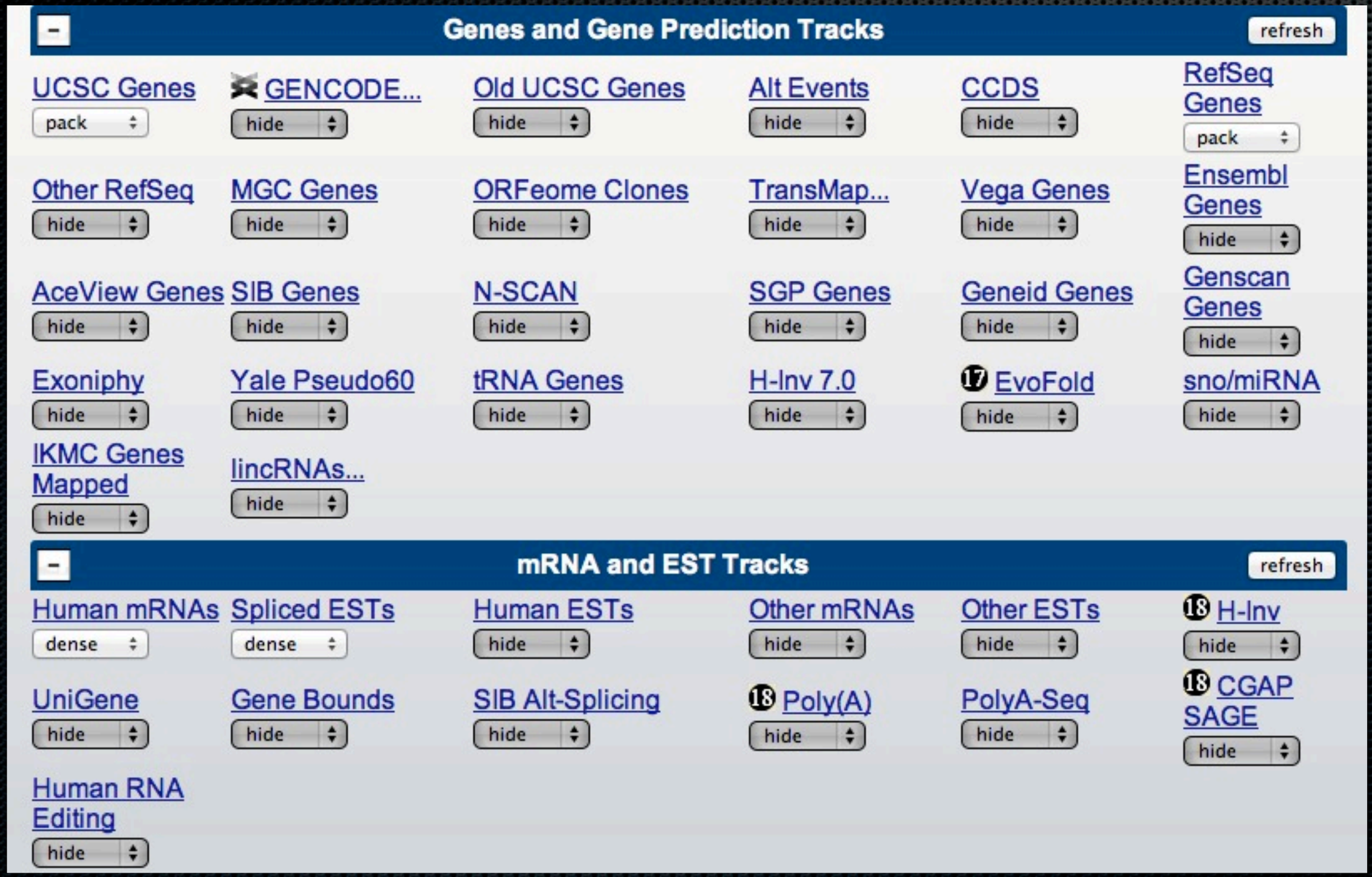
Base Position: dense	Chromosome Band: hide	STS Markers: hide	FISH Clones: hide	Recomb Rate: hide	deCODE Recomb: hide
ENCODE Pilot: hide	Map Contigs: hide	Assembly: hide	GRC Map Contigs: hide	Gap: hide	Publications: hide
BAC End Pairs: hide	Fosmid End Pairs: hide	GC Percent: hide	GRC Patch Release: hide	Hg18 Diff: hide	GRC Incident: hide
Hi Seq Depth: hide	Wiki Track: hide	BU ORChID: hide	Mapability: hide	Short Match: hide	Restr Enzymes: hide

Phenotype and Disease Associations — refresh

GAD View: hide	DECIPHER: hide	OMIM AV SNPs: hide	OMIM Genes: hide	OMIM Pheno Loci: hide	COSMIC: hide
GWAS Catalog: hide	ISCA: hide	RGD Human QTL: hide	RGD Rat QTL: hide	MGI Mouse QTL: hide	GeneReviews: hide

UCSC Genome Browser

Tracks



UCSC Genome Browser

Tracks

Genes and Gene Prediction Tracks — refresh

UCSC Genes: pack	GENCODE...: hide	Old UCSC Genes: hide	Alt Events: hide	CCDS: hide	RefSeq Genes: pack
Other RefSeq: hide	MGC Genes: hide	ORFeome Clones: hide	TransMap...: hide	Vega Genes: hide	Ensembl Genes: hide
AceView Genes: hide	SIB Genes: hide	N-SCAN: hide	SGP Genes: hide	Geneid Genes: hide	Genscan Genes: hide
Exoniphy: hide	Yale Pseudo60: hide	tRNA Genes: hide	H-Inv 7.0: hide	17 EvoFold: hide	sno/miRNA: hide
IKMC Genes Mapped: hide	lincRNAs...: hide				

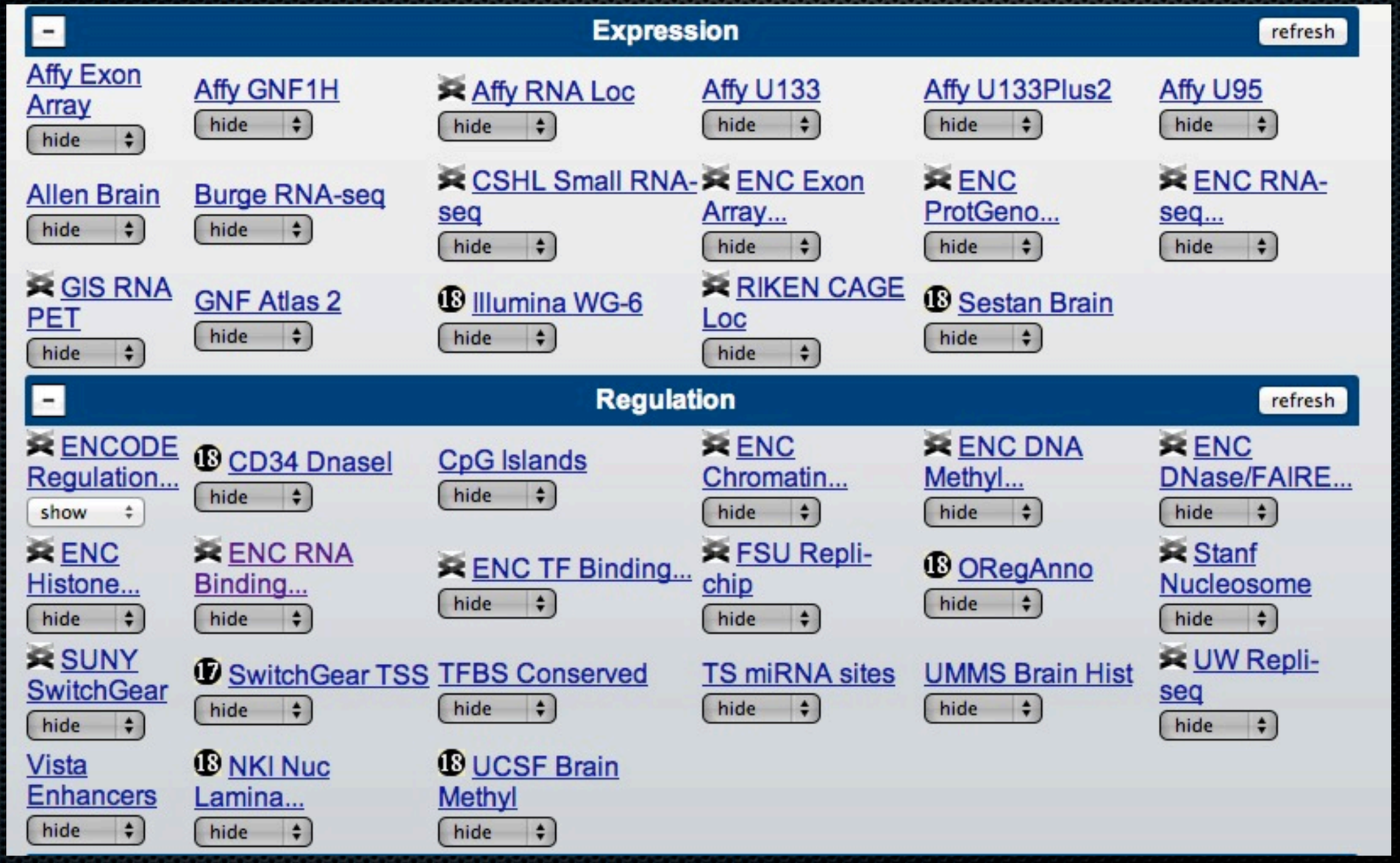
mRNA and EST Tracks — refresh

Human mRNAs: dense	Spliced ESTs: dense	Human ESTs: hide	Other mRNAs: hide	Other ESTs: hide	18 H-Inv: hide
UniGene: hide	Gene Bounds: hide	SIB Alt-Splicing: hide	18 Poly(A): hide	PolyA-Seq: hide	18 CGAP SAGE: hide
Human RNA Editing: hide					

hide
dense
squish
pack
full

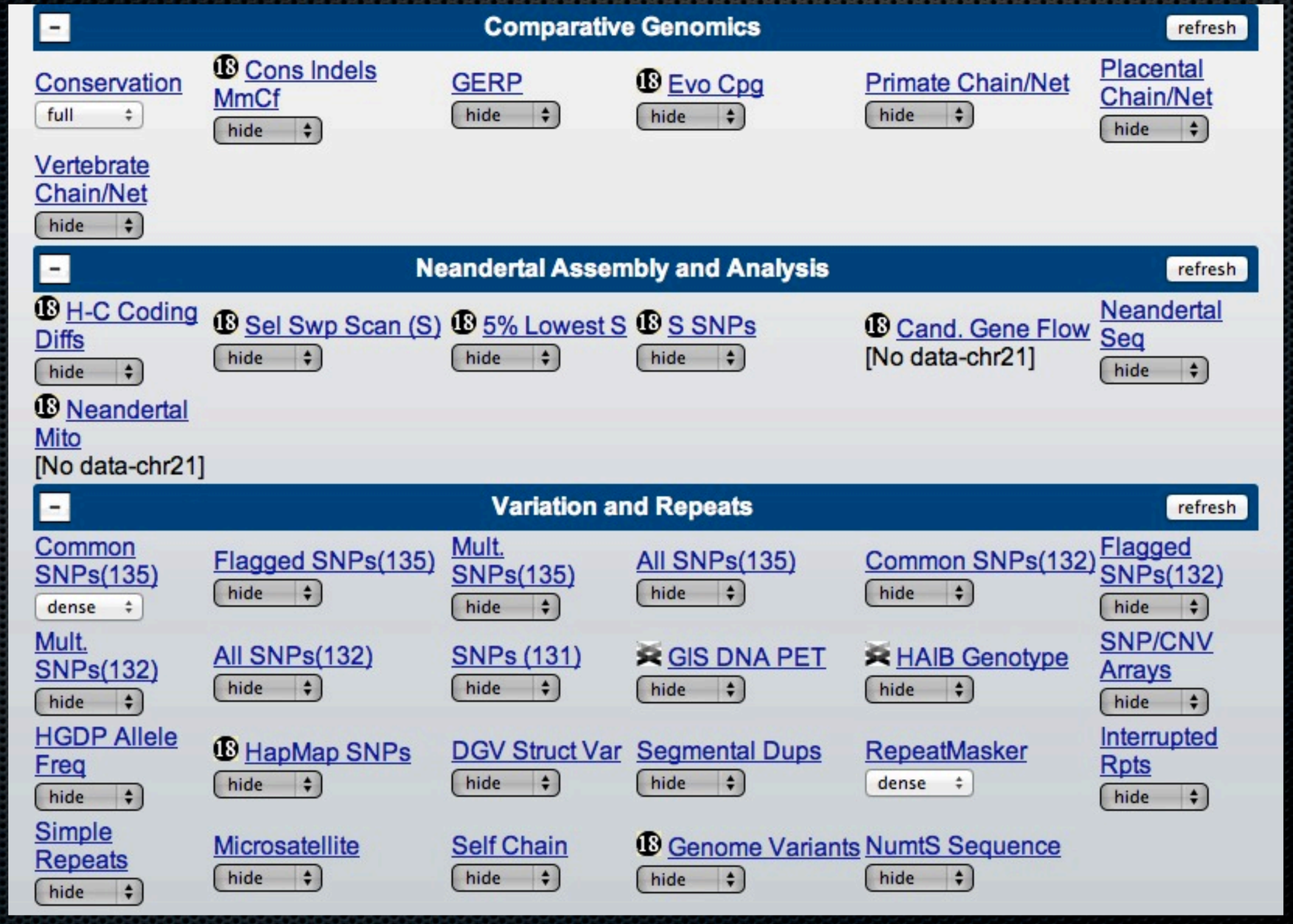
UCSC Genome Browser

Tracks



UCSC Genome Browser

Tracks



UCSC Genome Browser

“Genes”

- Different annotation tracks for the same “gene” have different amounts of information
  - UCSC track typically shows the most extensive data
  - Sequence, crossreferences, structure, expression, gene ontology, pathways, **orthologs**
  - Different amounts of data are available for different genes

UCSC Genome Browser

Extract Sequence

- Download specific Sequence Data
  - Corresponding to the display
    - Corresponding to a feature
- Download Genomic DNA,RNA,Protein,Promoter sequence
- Download feature via theTable Browser

UCSC Genome Browser

Extract Sequence

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

Get DNA in Window (hg19/Human)

Get DNA for

Position chr21:32,537,884-33,535,283

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.

Sequence Retrieval Region Options:

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

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)

get DNA | extended case/color options

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

Options for get DNA from View menu or mouse-over feature

UCSC Genome Browser

Extract Sequence

Sequence and Links to Tools and Databases

Genomic Sequence (chr21:32,490,736-32,649,224)			mRNA (may differ from genome)		Protein (1452 aa)
Gene Sorter	Genome Browser	Protein FASTA	VisiGene	Table Schema	BioGPS
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneNetwork
Gepis Tissue	H-INV	HGNC	HPRD	Human Cortex Gene Expression	Jackson Lab
MOPED	OMIM	PubMed	Reactome	Stanford SOURCE	Treefam
UniProtKB	Wikipedia				

Options for getting sequence data from Gene view

UCSC Genome Browser

Extract Sequence

Sequence and Links to Tools and Databases

Genomic Sequence (chr21:32,490,736-32,649,224)			mRNA (may differ from genome)		Protein (1452 aa)
Gene Sorter	Genome Browser	Protein FASTA	VisiGene	Table Schema	BioGPS
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneNetwork
Gepis Tissue	H-INV	HGNC	HPRD	Human Cortex Gene Expression	Jackson Lab
MOPED	OMIM	PubMed	Reactome	Stanford SOURCE	Treefam
UniProtKB	Wikipedia				

Options for getting sequence data from Gene view

UCSC Genome Browser

Extract Sequence

Genomic Sequence Near Gene

Get Genomic Sequence Near Gene

Note: if you would prefer to get DNA for more than one feature of this track at a time, try the Table Browser using the output format sequence.

Sequence Retrieval Region Options:

- ☐ Promoter/Upstream by 1000 bases
- ☑ 5' UTR Exons
- ☑ CDS Exons
- ☑ 3' UTR Exons
- ☑ Introns
- ☐ Downstream by 1000 bases
- ◉ One FASTA record per gene.
- ○ One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (5') and 0 extra downstream (3')
  - ☐ Split UTR and CDS parts of an exon into separate FASTA records

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:

- ◉ Exons in upper case, everything else in lower case.
- ○ CDS in upper case, UTR in lower case.
- ○ All upper case.
- ○ All lower case.
- ☐ Mask repeats: ◉ to lower case ○ to N

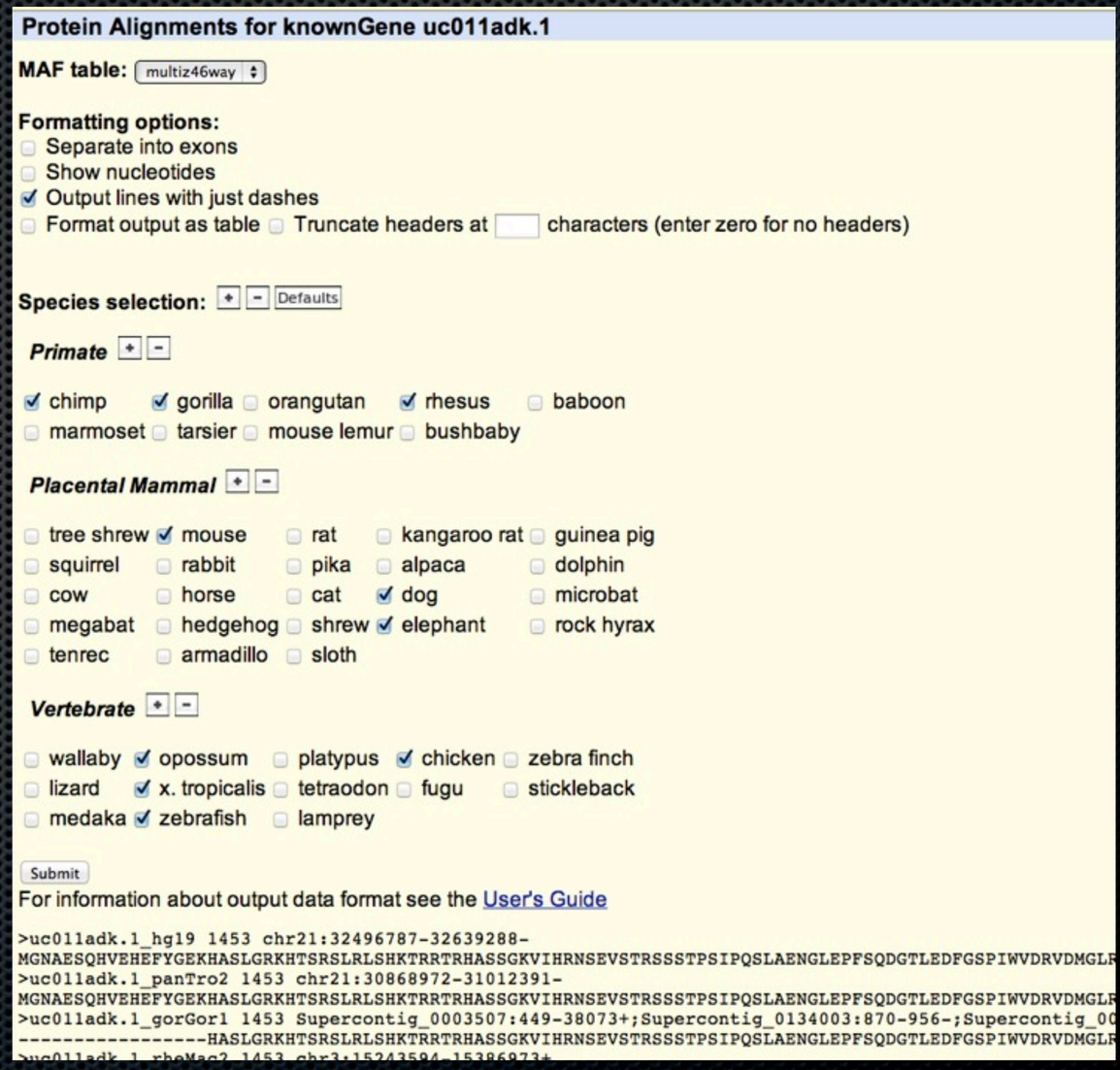
submit

Options for getting Genomic sequence

UCSC Genome Browser

Extract Sequence

Options for getting sequence data from Protein Fasta item



UCSC Genome Browser

Tools - BLAT

(BLAST-Like Alignment Tool)

Search by sequence **similarity**

- BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 25 (sometimes 20) bases or more.
  - It may miss more divergent or shorter sequence alignments.
  - BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more.

BLAT is not BLAST

Be cautious of matches, and no-match means little

UCSC Genome Browser

Tools - Table Browser

- Retrieve the data associated with a track in text format
- Calculate intersections between tracks
- Retrieve DNA sequence covered by a track

UCSC Genome Browser

Tools - GeneSorter

- Displays a sorted table of genes that are related to one another. The relationship can be one of several types, including protein-level homology, similarity of gene expression profiles, or genomic proximity.
- One of the most powerful features of the Gene Sorter is its filtering capabilities. Use the filter to fine-tune the list of displayed genes to a subset based on a selection of detailed and flexible criteria. For example, the filter may be used to select all human genes over-expressed in the cerebellum that have GO-annotated G-protein coupled receptor activity.
- The Gene Sorter offers two options for displaying and downloading sequence associated with the genes in the table.
  - **sequence button** will fetch associated protein, mRNA, promoter, or genomic sequence.
  - **text button** will dump the table into a simple tab-delimited format suitable for import into a spreadsheet or relational database.

UCSC Genome Browser

Tools - Others

- Genome Graphs - is a tool for displaying genome-wide data sets
- InSilico PCR - searches a sequence database with a pair of PCR primers
- Liftover - converts genome coordinates and genome annotation files between assemblies.
- VisiGene - This tool converts genome coordinates and genome

UCSC Genome Browser

MyData - Sessions

- Way of saving your “View” of the genome for later use or sharing
- Requires an account
- Saved for 4 months after last access
- Can be shared with others
- Can include Custom Tracks - Your Data

UCSC Genome Browser

MyData - Track Hubs

- Way of importing data from outside sources
- Generally consists of **collections** of tracks
- Publically created Hubs
- My Hubs - “your” data hosted on Web server or FTP site

UCSC Genome Browser

MyData - Custom Tracks

Way of importing data from an outside source as a TRACK.

- Upload annotation data from your own computer
  - Only visible from same location (lasts 48 hours)
- Upload data to a Web or FTP site
  - Can be shared with anyone (forever - or data life time)
- Create session with custom track and share
  - Can be shared with anyone (lasts 4 months or data life time)
  - Contribute track to UCSC

UCSC Genome Browser

MyData - Custom Tracks

Source of information about many different file formats

- **General Formats**
- Axt
- BAM
- BED
- BED detail
- bedGraph
- bigBed
- bigWig
- Chain
- GenePred table
- GFF
- GTF
- MAF
- Microarray
- Net
- Personal Genome SNP
- PSL
- VCF
- WIG
- **ENCODE-specific formats:**
- ENCODE broadPeak
- ENCODE gappedPeak
- ENCODE narrowPeak
- ENCODE pairedTagAlign
- ENCODE peptideMapping
- ENCODE RNA elements
- ENCODE tagAlign
- **Download only formats:**
- 2bit
- fasta format
- fastQ format
- nib format

UCSC Genome Browser

MyData - Custom Tracks

BED format

1. **chrom** - name of the chromosome
2. **chromStart** - Start of feature (0-based)
3. **chromEnd** - End of the feature (not included in display)

9 more optional columns

4. **name** - a label for the feature
5. **score** - a score (0-1000)
6. **strand** - which strand the feature on (+/-)

chr1	**15000**	**20000**	**gene1**	**50**	**+**
chr2	**106000**	**108000**	**gene2**	**400**	**-**

UCSC Genome Browser

MyData - Custom Tracks

BEDGraph format

1.**chrom** - name of the chromosome
2.**chromStart** - Start of feature (0-based)
3.**chromEnd** - End of the feature (not included in display)
4.**score** - a score (integer or real positive / negative number)

chr1	**15000**	**20000**	**1.00**
chr2	**106000**	**108000**	**0.75**

Format Conversion Utilities

- Galaxy (http://galaxy.psu.edu/ - http://galaxy.cit.nih.gov/)
  - Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.
- Samtools (http://samtools.sourceforge.net)
  - SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format. Also, note TABIX for indexing generic tab delimited files.
- Picard (http://picard.sourceforge.net/)
  - Picard comprises Java-based command-line utilities that manipulate SAM files, and a Java API (SAM-JDK) for creating new programs that read and write SAM files. Both SAM text format and SAM binary (BAM) format are supported.
- UCSC Utilities (http://hgdownload.cse.ucsc.edu/admin/exe/)

UCSC Genome Browser

MyData - Custom Tracks

Indexed formats

Indexed binary file formats are much more efficient.

Only the portions of the files needed for the region currently displayed are transferred and loaded into the Browser. Thus for large data sets they are considerably faster than regular files. (e.g. bigBED, bigWIG, BAM*indexed*)

UCSC Genome Browser

Downloads

- Genome Data - download any and all data behind the browsers in large chunks or specific files
- Source Code - download the code used to drive the browser (mirrors)
- Utilities- download the code for ancillary programs
- FTP - download data via our FTP site at **ftp://hgdownload.cse.ucsc.edu/**
- MYSQL Access - Direct MySQL access to the UCSC Browser database **genome-mysql.cse.ucsc.edu**.

Integrated Genome Browser IGB

- Integrated Genome Browser (IGB) is an easy-to-use, highly customizable genome browser you can use to visualize and explore genomic data and annotations, including RNA-Seq, ChIP-Seq, tiling array data, and more

- IGB (GenoViz) http://genoviz.sourceforge.net/
- Support Forum http://sourceforge.net/projects/genoviz/forums/forum/439787
- IGB User's Guide (PDF) http://genoviz.sourceforge.net/IGB_User_Guide.pdf

The Integrative Genomics Viewer IGV

- The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

- IGV http://www.broadinstitute.org/igv/
- FAQ http://www.broadinstitute.org/software/igv/FAQ
- User's Guide http://www.broadinstitute.org/software/igv/UserGuide
- Support Forum https://groups.google.com/forum/?fromgroups#!forum/igv-help