Genome Browsers

Peter FitzGerald. PhD

Head Genome Analysis Unit. CCR. NCI

Talk Outline

- Overview of Genome Browsers
- Different typed of Browsers
- Highlighting two specific Browsers
  - UCSC Genome Browser
  - IGV - Integrative Genomics Viewer
- Browsers for geeks/pros
- Conclusion

Overview of Genome Browsers

- **Tools that provide a graphical view of genomic data**
  - Enable Biological insight

    1) Provide a view of the spatial relationship between “genes”

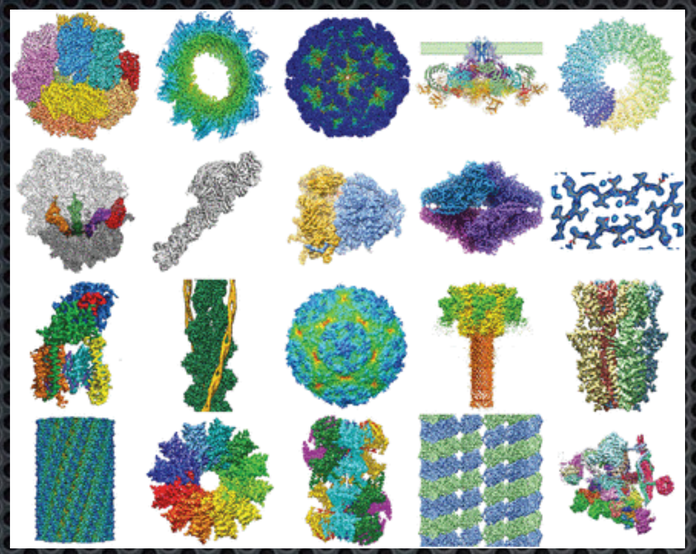
    2) Good for comparing and integrating different data sets

  - Check of data integrity

    3) Drilling down into the raw data (see what cutoffs are doing)

    4) Enable one to perform sanity checks on the data

The 2016 Nucleic Acids Research



Database issue
Volume 44 Issue D1 04 January 2016
https://nar.oxfordjournals.org/content/44/D1.toc
Database Summaries
https://nar.oxfordjournals.org/content/44/D1/D1/suppl/DC1

Web Server issue
Volume 44 Issue W1 08 July 2016
http://nar.oxfordjournals.org/content/44/W1.toc

Wikipedia

https://en.wikipedia.org/wiki/Genome_browser

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

Web-based Browsers

- UCSC Genome Browser
  - **http://genome.ucsc.edu**
- NCBI Map Viewer
  - **http://www.ncbi.nlm.nih.gov/mapview/**
- Ensemble
  - **http://www.ensembl.org/**

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

Hybrid Browsers

- Integrated Genome Browser (IGB)
  http://genoviz.sourceforge.net/

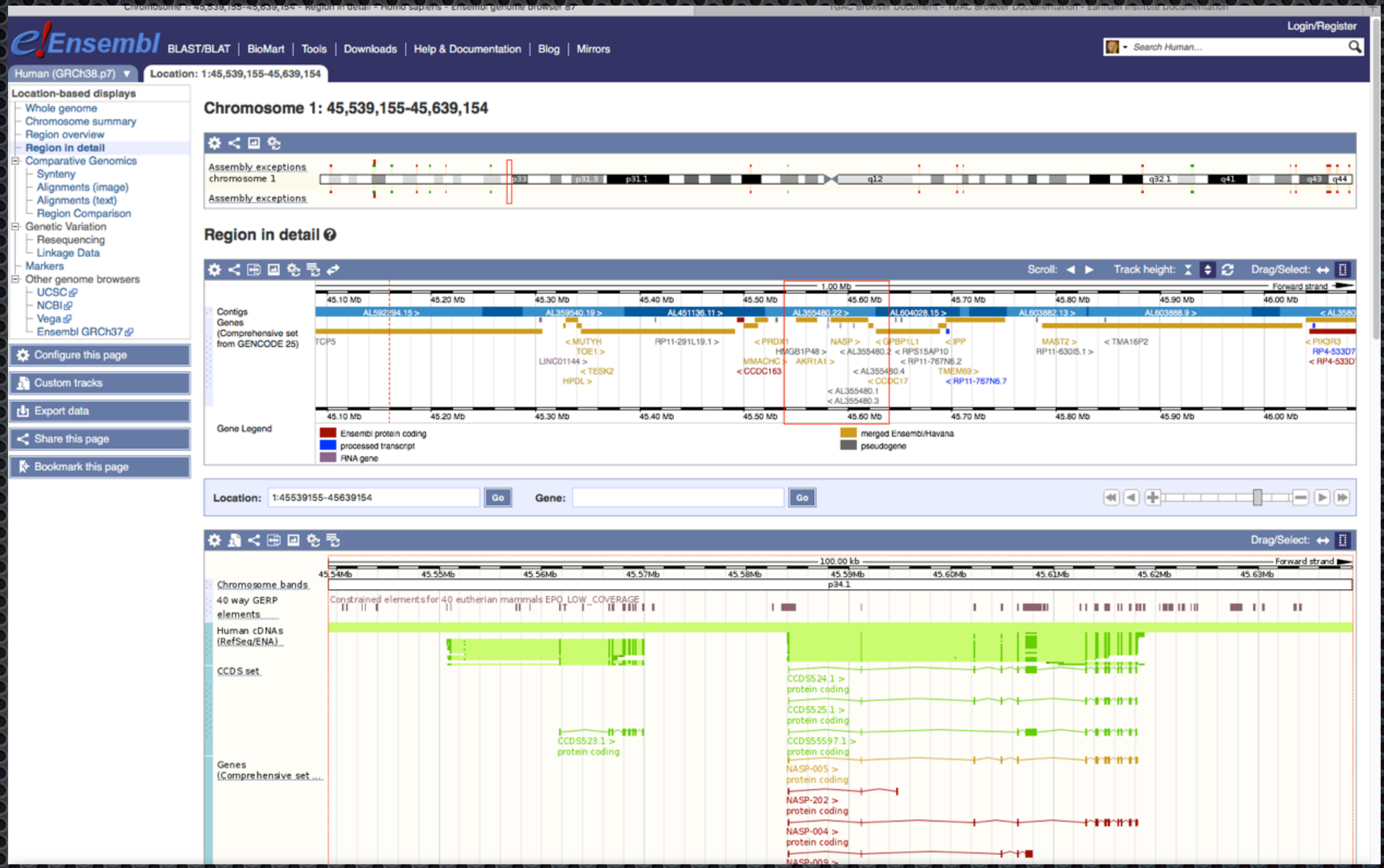
- The Integrative Genomics Viewer (IGV)
  http://www.broadinstitute.org/igv/

Commercial Browsers

- Partek Genomic Suite & Partek FLOW
  http://www.partek.com
- Genomatix
  https://www.genomatix.de
- Geneious
  https://www.geneious.com
- CLC Genomic Workbench (Biomedical Workbench)
  http://www.clcbio.com

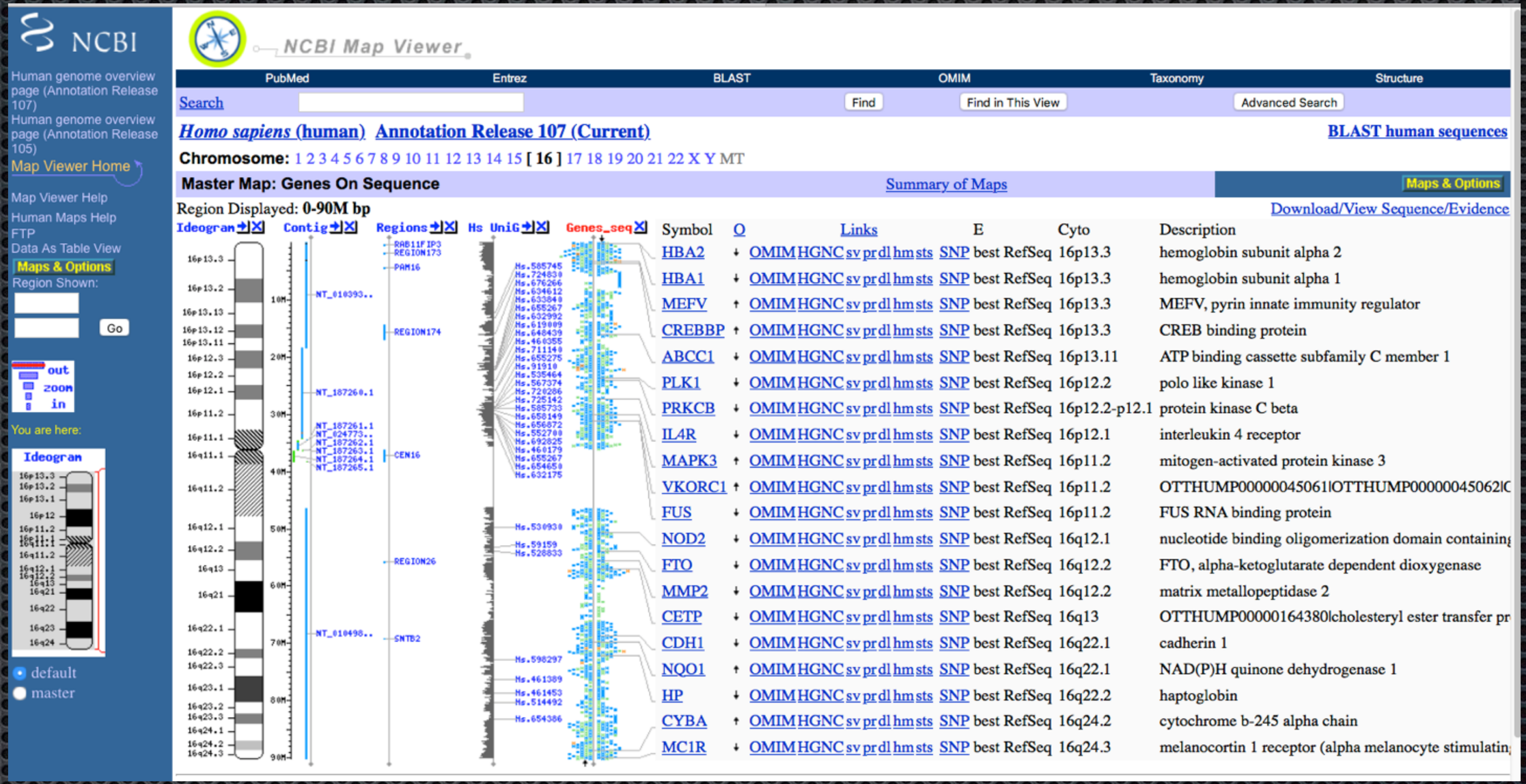
Ensembl Genome Browser

http://www.ensembl.org/Homo_sapiens/Location/View?r=1



NCBI Map Viewer

https://www.ncbi.nlm.nih.gov/projects/mapview/



Browsers of Note

- **Cancer Biology**
  - Cancer (Cosmic) Genome Browser
    COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.
    http://cancer.sanger.ac.uk/cosmic
  - Cancer Browser - now replaced with Xena Browser
    A collection of UCSC-hosted public databases such as TCGA, ICGC, TARGET, GTEx, CCLE, and others. Databases are normalized so they can be combined, linked, filtered, explored and downloaded.
    http://xena.ucsc.edu/

Browsers of Note

- **Epigenetic**
  - WashU EpiGenome Browser
    http://epigenomegateway.wustl.edu/browser/
- **Other**
  - 1000 Genomes (NCBI)
    https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/
  - 1000 Genomes (EMBI-EBI)
    http://browser.1000genomes.org/
  - VISTA (Tools for Comparative Genomics)
    http://genome.lbl.gov/vista/index.shtml

Browsers of Note

- **Standalone (Custom)**
  - Biodalliance - a fast, interactive, genome visualization tool that's easy to embed in web pages and applications **http://www.biodalliance.org**
  - GBrowse **http://gmod.org/wiki/GBrowse**
  - JBrowse **http://jbrowse.org**
  - MochiView - Java software that integrates browsing of genomic sequences, features, and data with DNA motif visualization and analysis. **http://www.johnsonlab.ucsf.edu/mochi/**

Beware of Version Changes

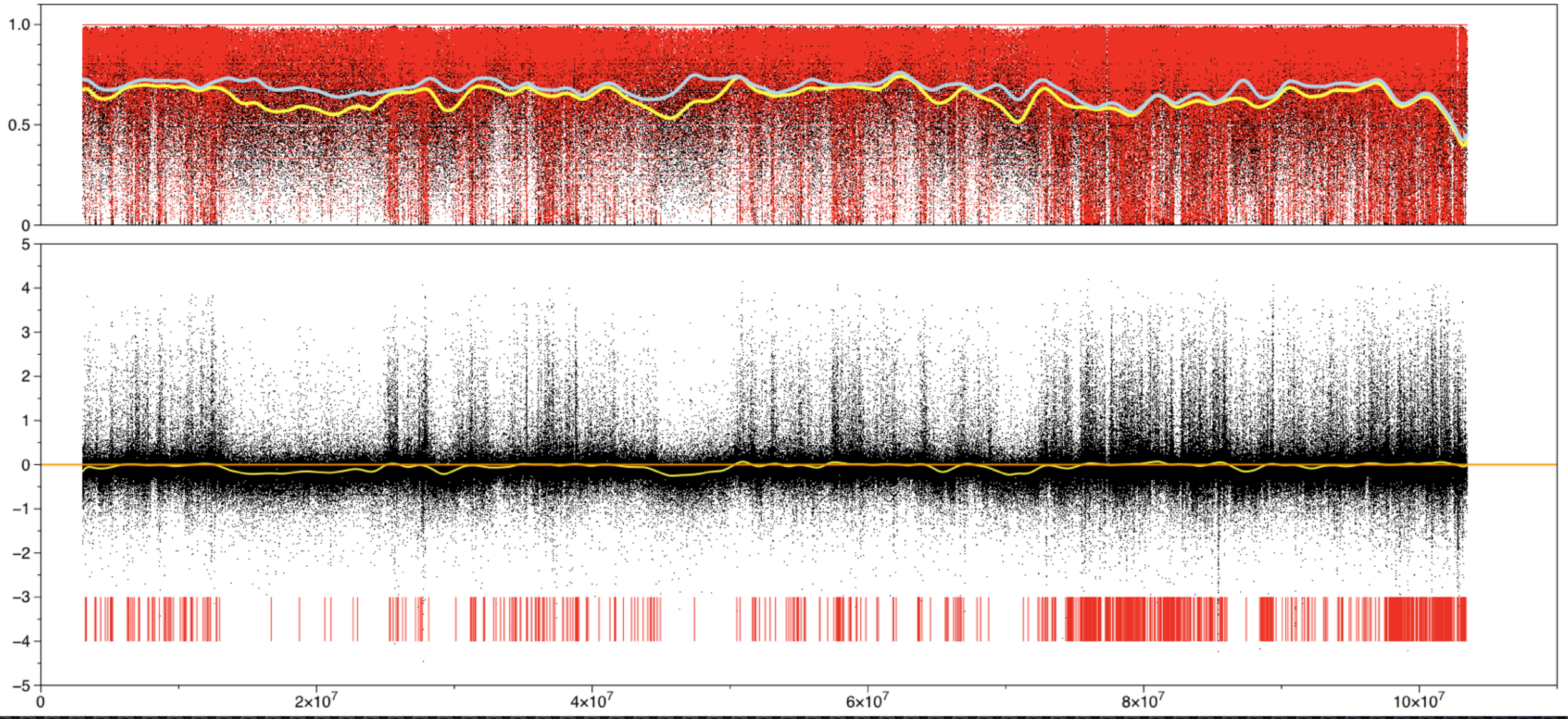
- Different browsers update on different schedules
- Different browsers have different archiving policies
- Annotations are more variable than sequence data
- Nomenclature is different

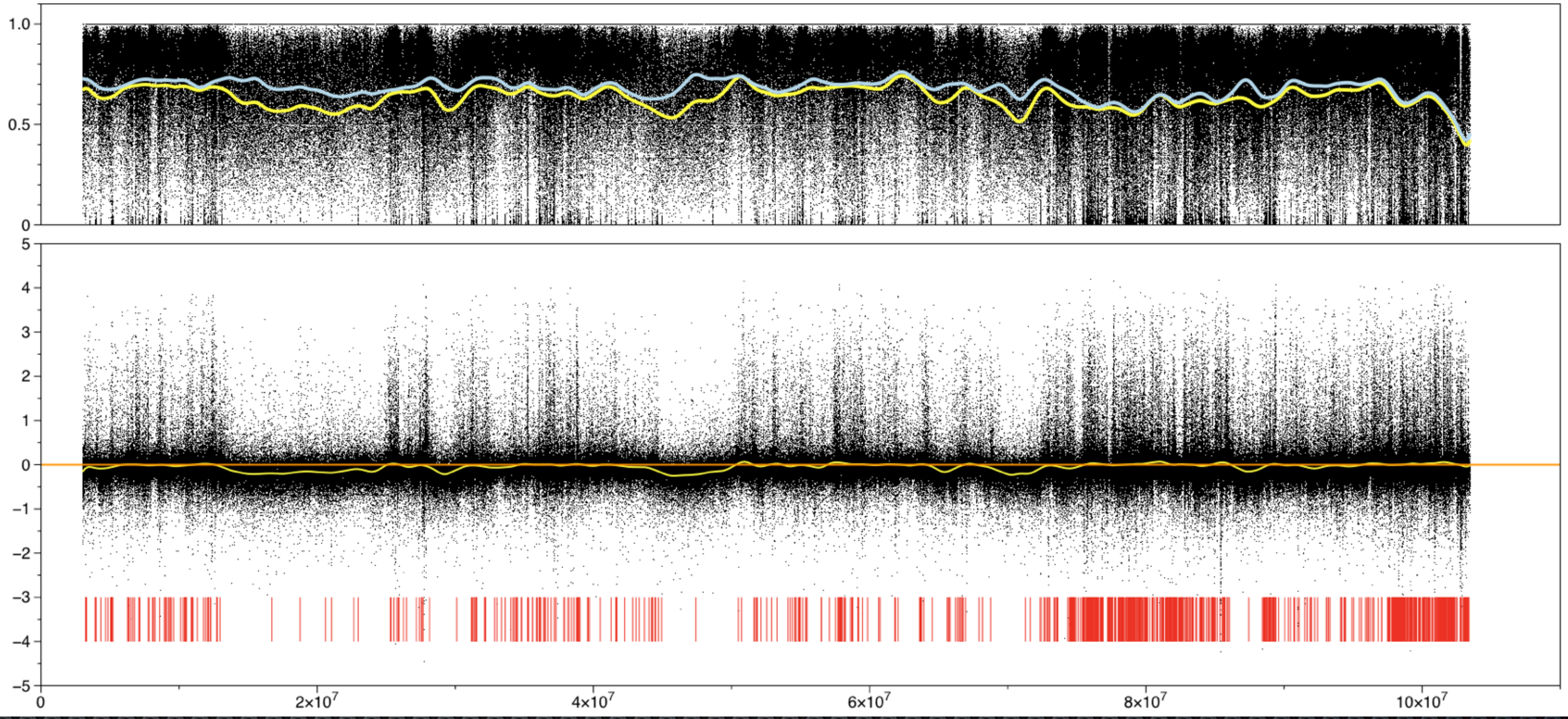
When comparing data from different sources always make sure you are looking at the same version.

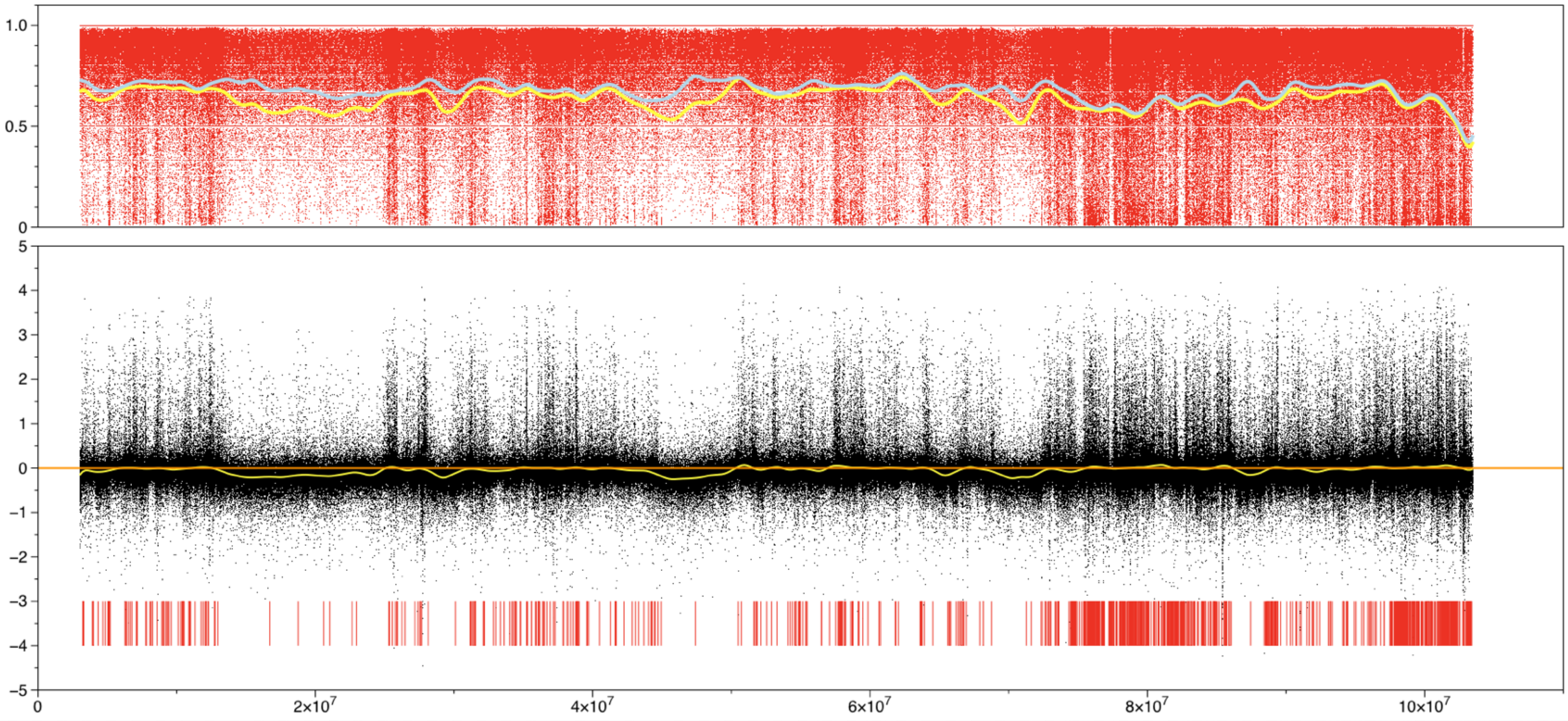
UCSC “liftover” allows files to be interconvert between versions

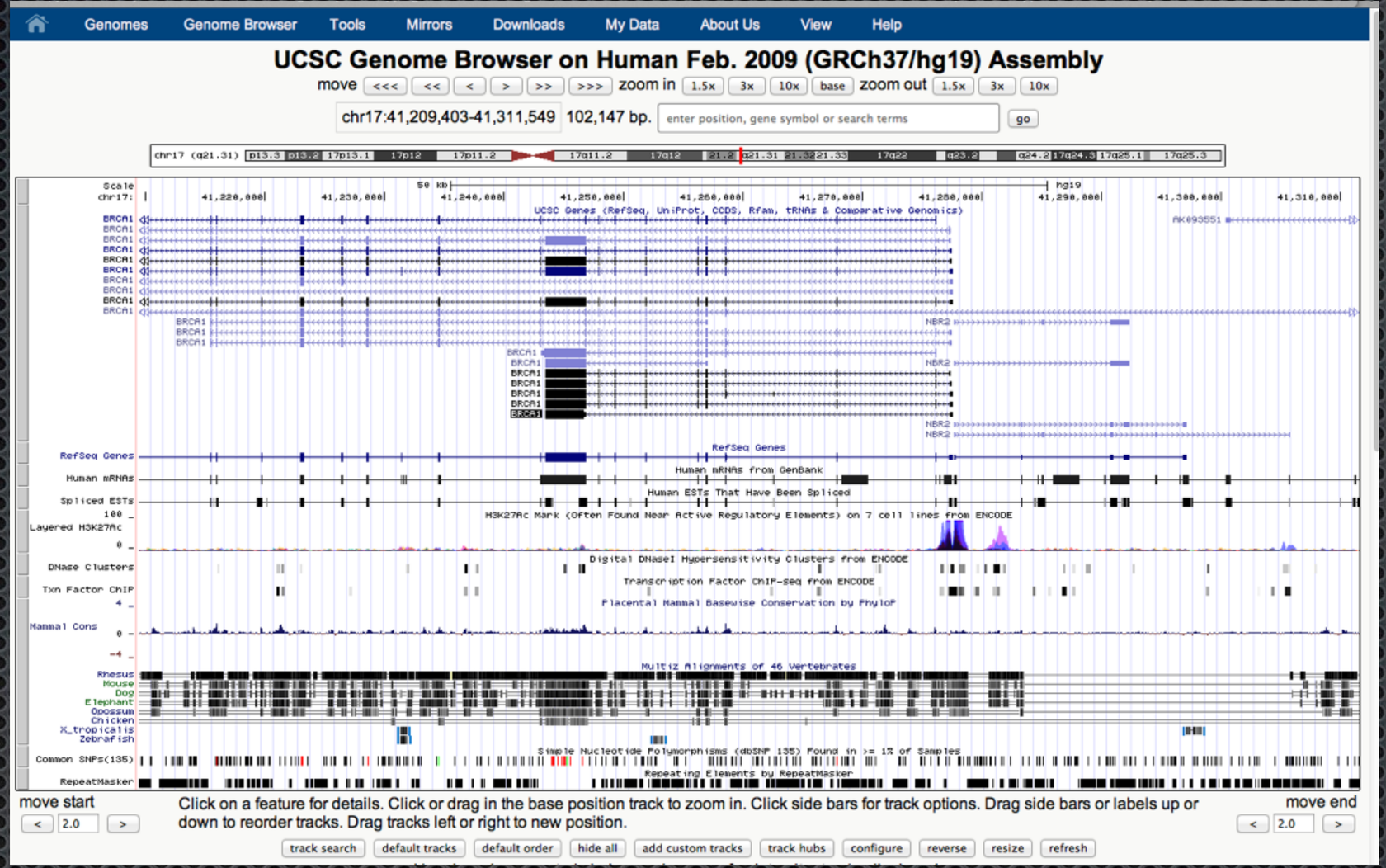
The Non-Browser browser

- For mammalian genomes most browsers to not provide a good method of viewing an entire chromosome - the really big picture.
- In such cases a **good graphing program** (capable of dealing with millions of points) may be more useful.


1.0
0.5
0
5
4
3
2
1
0
−1
−2
−3
−4
−5
0
2×10^7
4×10^7
6×10^7
8×10^7
10×10^7


1.0
0.5
0
5
4
3
2
1
0
−1
−2
−3
−4
−5
0
2×10^7
4×10^7
6×10^7
8×10^7
10×10^7


1.0
0.5
0
5
4
3
2
1
0
−1
−2
−3
−4
−5
0
2×10^7
4×10^7
6×10^7
8×10^7
10×10^7



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 Mirrors**
  - *European Mirror: http://genome-euro.ucsc.edu*
  - *Asian Mirror: http://genome-asia.ucsc.edu*
  - NIH local Mirror (Subset) maintained by the Helix Sytems (CIT) **http://genome.cit.nih.gov/**
- Cancer Browser - now replaced with Xena Browser **http://xena.ucsc.edu/**
  - A collection of UCSC-hosted public databases such as TCGA, ICGC, TARGET, GTEx, CCLE, and others. Databases are normalized so they can be combined, linked, filtered, explored and downloaded.

Genome Browser (UCSC) in a Box (GBiB)

http://genome.ucsc.edu/goldenpath/help/gbib.html

Genome Browser in a Box (GBiB) is a "virtual machine" of the entire UCSC Genome Browser website that is designed to run on most PCs (Windows, Mac OSX or Linux). GBiB allows you to access much of the UCSC Genome Browser's functionality from the comfort of your own computer. It is particularly directed at individuals who want to use the Genome Browser toolset to view protected data.

GBiB depends upon remote connections to various UCSC servers for much of its functionality and data. GBiB is currently optimized for use with the hg19 assembly.

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**
- http://genome.ucsc.edu/training/vids/ **Videos**

UCSC Genome Browser

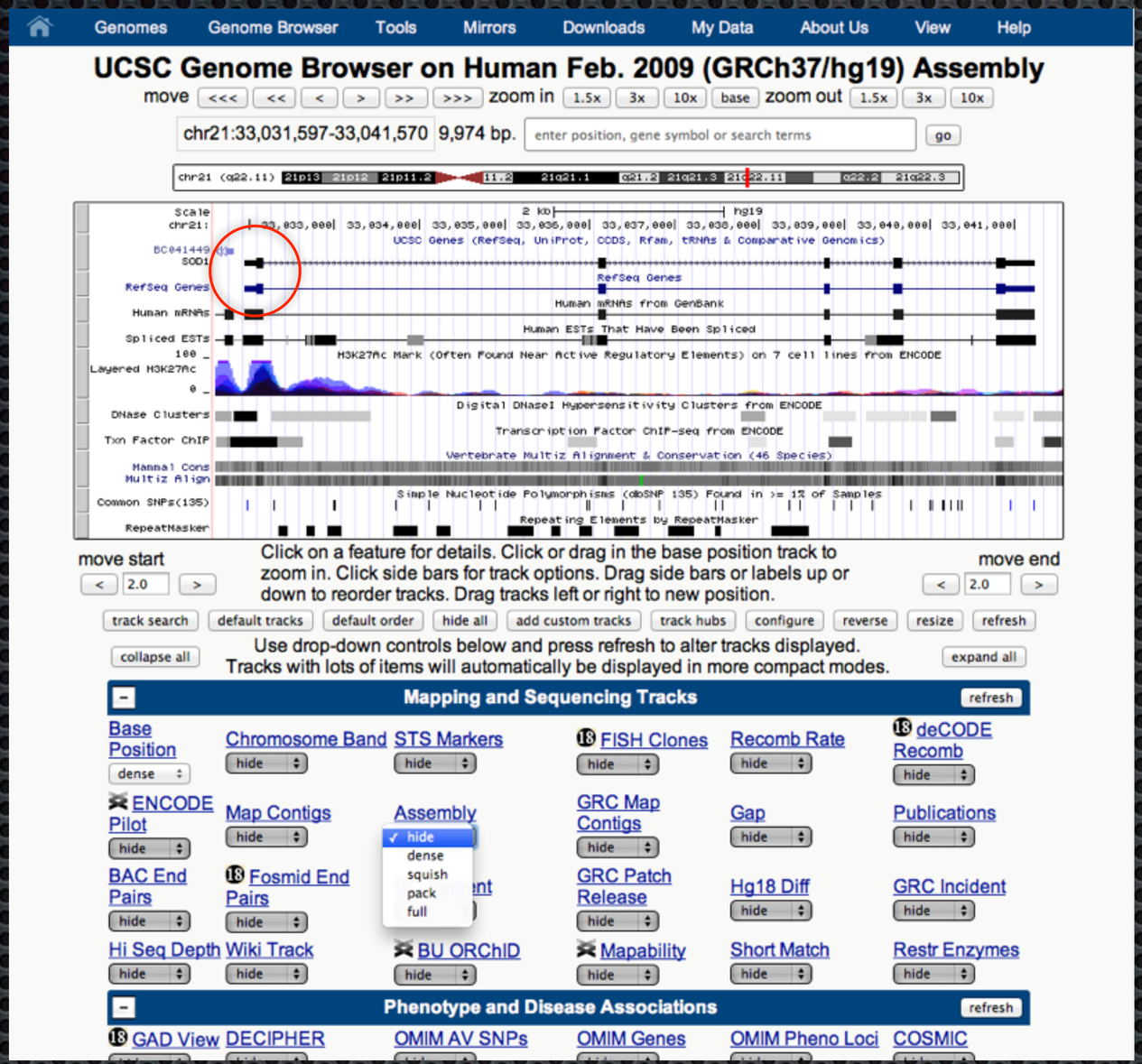
Genomes 98

- Human
- Chimp
- Bonobo
- Gorilla
- Orangutan
- Gibbon
- Green monkey
- Crab-eating macaque
- Rhesus
- Baboon (anubis)
- Baboon (hamadryas)
- Marmoset
- Squirrel monkey
- Tarsier
- Mouse lemur
- Bushbaby
- Mouse
- Rat
- Chinese hamster
- Kangaroo rat
- Squirrel
- Naked mole-rat
- Guinea pig
- Rabbit
- Pika
- Malayan flying lemur
- Tree shrew
- Hedgehog
- Shrew
- Pig
- Cow
- Sheep
- Dolphin
- Minke whale
- Alpaca
- Horse
- White rhinoceros
- Dog
- Ferret
- Panda
- Cat
- Megabat
- Microbat
- Elephant
- Manatee
- Rock hyrax
- Tenrec
- Armadillo
- Sloth
- Wallaby
- Tasmanian devil
- Opossum
- Platypus
- Chicken
- Turkey
- Zebra finch
- Medium ground finch
- Budgerigar
- Brown kiwi
- American alligator
- Painted turtle
- Lizard
- X. tropicalis
- Coelacanth
- Zebrafish
- Stickleback
- Fugu
- Tetraodon
- Medaka
- Nile tilapia
- Atlantic cod
- Elephant shark
- Lamprey
- Lancelet
- C. intestinalis
- S. purpuratus
- D. melanogaster
- D. erecta
- D. sechellia
- D. simulans
- D. yakuba
- D. ananassae
- D. persimilis
- D. pseudoobscura
- D. mojavensis
- D. virilis
- D. grimshawi
- A. gambiae
- A. mellifera
- C. elegans
- C. brenneri
- C. briggsae
- C. japonica
- C. remanei
- P. pacificus
- Sea hare
- S. cerevisiae
- Ebola viru

BTEP

UCSC Genome Browser

Overview

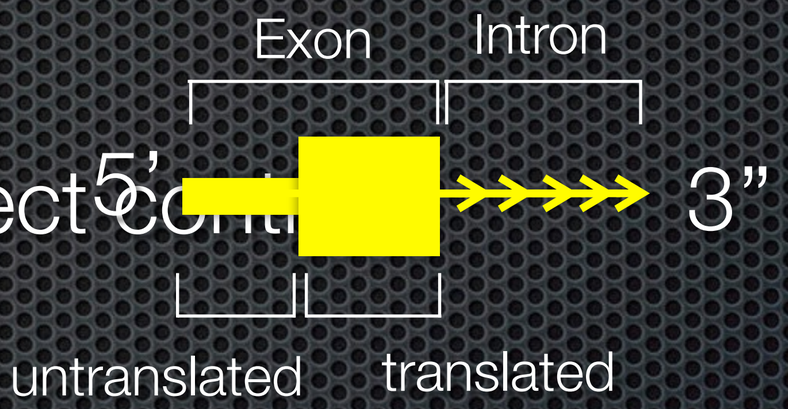


Zoom level

Coordinates and search

Chromosome position and coordinates

Main Display (image)



Select control

Track controls

UCSC Genome Browser

Tracks

- Vary from Genome to Genome and between versions
- Most current versions may lack tracks that you expect

Current Human Genome Tracks

Human Feb. 2009 (GRCh37/hg19) Assembly

Mapping and Sequencing	refresh
Genes and Gene Predictions	refresh
Phenotype and Literature	refresh
mRNA and EST	refresh
Expression	refresh
Regulation	refresh
Comparative Genomics	refresh
Neandertal Assembly and Analysis	refresh
Denisova Assembly and Analysis	refresh
Variation	refresh
Repeats	refresh

Human Dec. 2013 (GRCh38/hg38) Assembly

Mapping and Sequencing	refresh
Genes and Gene Predictions	refresh
Phenotype and Literature	refresh
mRNA and EST	refresh
Expression	refresh
Regulation	refresh
Comparative Genomics	refresh
Variation	refresh
Repeats	refresh

Human Feb. 2009 (GRCh37/hg19) Assembly

Mapping and Sequencing — refresh

Expand this group

Genes and Gene Predictions — refresh

UCSC Genes: hide	RefSeq Genes: dense	AceView Genes: hide	AUGUSTUS: hide	CCDS: hide	CRISPR...: hide
Ensembl Genes: hide	17 EvoFold: hide	Exoniphy: hide	GENCODE...: hide	Geneid Genes: hide	Genscan Genes: hide
H-Inv 7.0: hide	IKMC Genes Mapped: hide	lincRNAs...: hide	LRG Transcripts: hide	MGC Genes: hide	N-SCAN: hide
Old UCSC Genes: hide	ORFeome Clones: hide	Other RefSeq: hide	Pfam in UCSC Gene: hide	Retroposed Genes: hide	SGP Genes: hide
SIB Genes: hide	sno/miRNA: hide	TransMap...: hide	tRNA Genes: hide	UCSC Alt Events: hide	UniProt: hide
Vega Genes: hide	Yale Pseudo60: hide				

Phenotype and Literature — refresh

mRNA and EST — refresh

Expression — refresh

Regulation — refresh

Comparative Genomics — refresh

Neandertal Assembly and Analysis — refresh

Denisova Assembly and Analysis — refresh

Variation — refresh

Repeats — refresh

Human Dec. 2013 (GRCh38/hg38) Assembly

Mapping and Sequencing — refresh

Genes and Gene Predictions — refresh

GENCODE v24: hide	RefSeq Genes: pack	RetroGenes V9: hide	All GENCODE...: hide	AUGUSTUS: hide	CCDS: hide
CRISPR...: hide	Geneid Genes: hide	Genscan Genes: hide	19 IKMC Genes Mapped: hide	LRG Transcripts: hide	MGC Genes: hide
Non-coding RNA...: hide	Old UCSC Genes: hide	ORFeome Clones: hide	Other RefSeq: hide	Pfam in UCSC Gene: hide	SGP Genes: hide
SIB Genes: hide	UCSC Alt Events: hide	UniProt: hide			

Phenotype and Literature — refresh

mRNA and EST — refresh

Expression — refresh

Regulation — refresh

Comparative Genomics — refresh

Variation — refresh

Repeats — refresh

Mouse Dec. 2011 (GRCm38/mm10) Assembly

Mapping and Sequencing — refresh

Genes and Gene Predictions — refresh

UCSC Genes: hide	GENCODE (Ensembl)...: show	Other RefSeq: dense	RefSeq Genes: dense	RetroGenes V6: pack	AUGUSTUS: hide
CCDS: hide	CRISPR...: hide	Ensembl Genes: hide	Geneid Genes: hide	Genscan Genes: hide	MGC Genes: hide
Old UCSC Genes: hide	ORFeome Clones: hide	Pfam in UCSC Gene: hide	SGP Genes: hide	tRNA Genes: hide	UCSC Alt Events: hide

Literature — refresh

mRNA and EST — refresh

Expression and Regulation — refresh

Comparative Genomics — refresh

Variation and Repeats — refresh

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

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

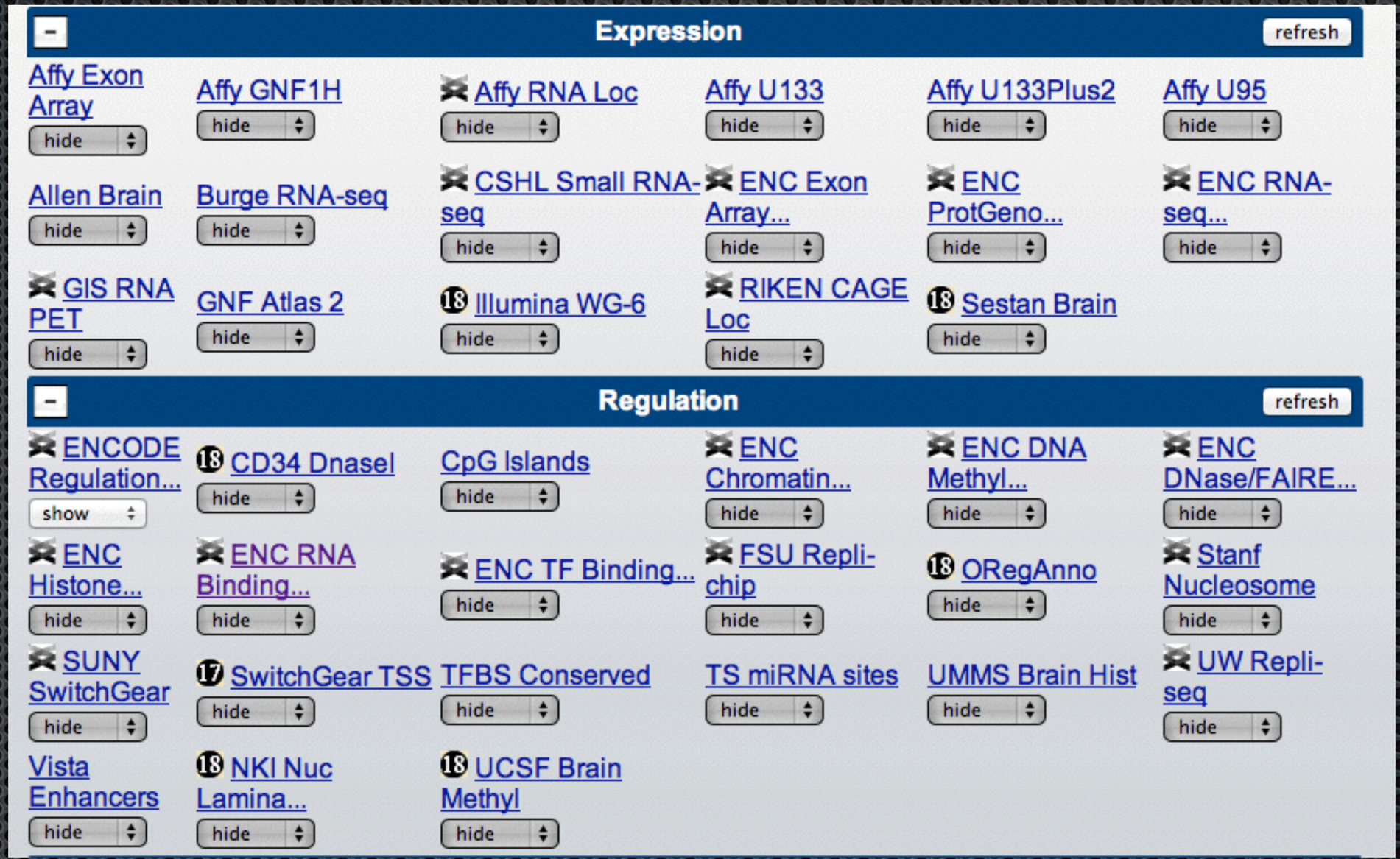
Phenotype and Disease Associations

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

Comparative Genomics — refresh

Conservation: full	18 Cons Indels MmCf: hide	GERP: hide	18 Evo Cpg: hide	Primate Chain/Net: hide	Placental Chain/Net: hide
Vertebrate Chain/Net: hide					

Neandertal Assembly and Analysis — refresh

18 H-C Coding Diffs: hide	18 Sel Swp Scan (S): hide	18 5% Lowest S: hide	18 S SNPs: hide	18 Cand. Gene Flow [No data-chr21]	Neandertal Seq: hide
18 Neandertal Mito [No data-chr21]					

Variation and Repeats — refresh

Common SNPs(135): dense	Flagged SNPs(135): hide	Mult. SNPs(135): hide	All SNPs(135): hide	Common SNPs(132): hide	Flagged SNPs(132): hide
Mult. SNPs(132): hide	All SNPs(132): hide	SNPs (131): hide	GIS DNA PET: hide	HAIB Genotype: hide	SNP/CNV Arrays: hide
HGDP Allele Freq: hide	18 HapMap SNPs: hide	DGV Struct Var: hide	Segmental Dups: hide	RepeatMasker: dense	Interrupted Rpts: hide
Simple Repeats: hide	Microsatellite: hide	Self Chain: hide	18 Genome Variants: hide	NumtS Sequence: hide	

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

BTEP

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

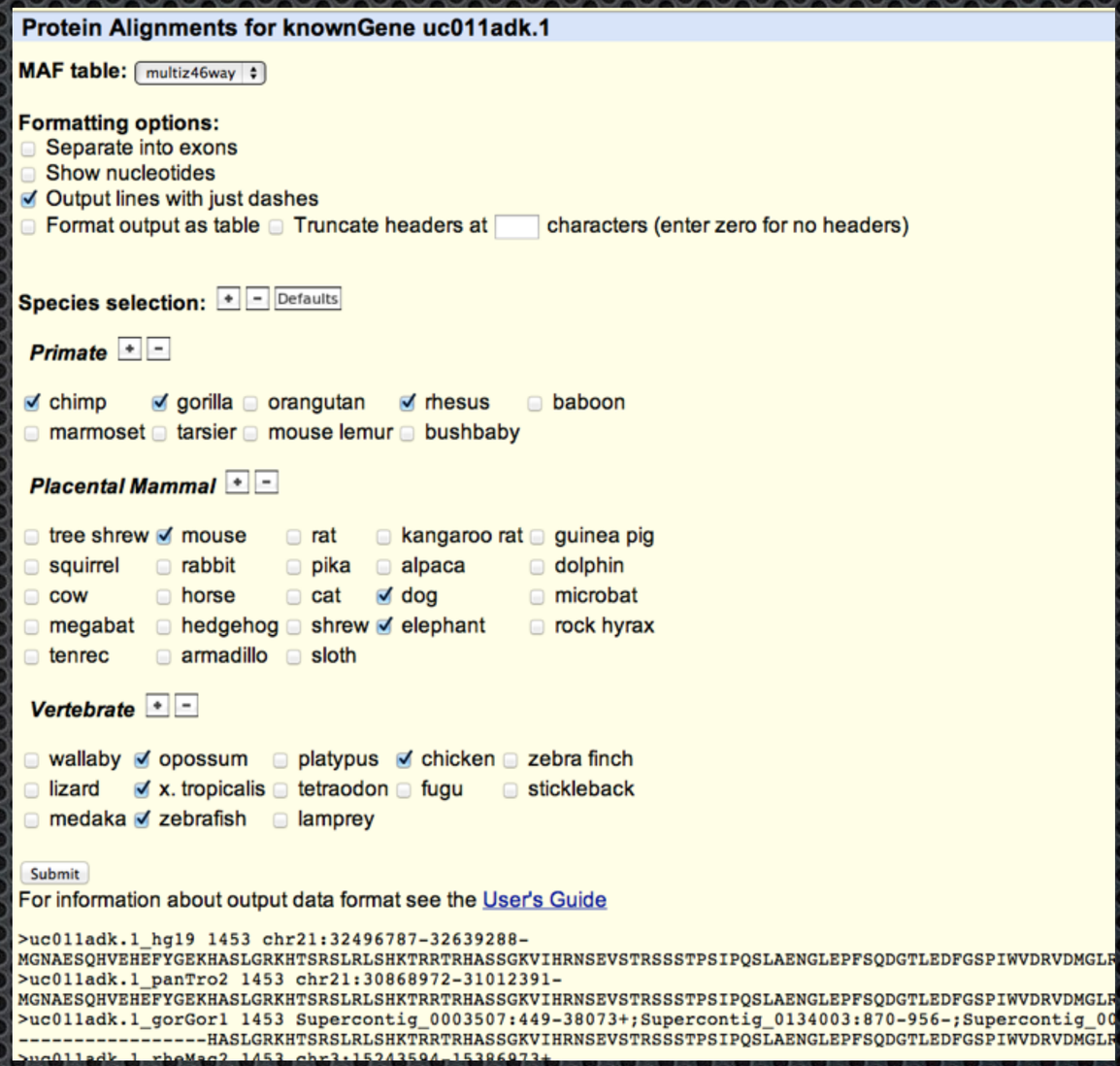
Options for getting Genomic sequence

BTEP

UCSC Genome Browser

Extract Sequence

Options for getting sequence data from Protein Fasta item



Protein Alignments for knownGene uc011adk.1

MAF table: multiz46way

Formatting options:
- ☐ Separate into exons
- ☐ Show nucleotides
- ☑ Output lines with just dashes
- ☐ Format output as table ☐ Truncate headers at ☐ characters (enter zero for no headers)

Species selection: + − Defaults

Primate + −

☑ chimp ☑ gorilla ☐ orangutan ☑ rhesus ☐ baboon
☐ marmoset ☐ tarsier ☐ mouse lemur ☐ bushbaby

Placental Mammal + −

☐ tree shrew ☑ mouse ☐ rat ☐ kangaroo rat ☐ guinea pig
☐ squirrel ☐ rabbit ☐ pika ☐ alpaca ☐ dolphin
☐ cow ☐ horse ☐ cat ☑ dog ☐ microbat
☐ megabat ☐ hedgehog ☐ shrew ☑ elephant ☐ rock hyrax
☐ tenrec ☐ armadillo ☐ sloth

Vertebrate + −

☐ wallaby ☑ opossum ☐ platypus ☑ chicken ☐ zebra finch
☐ lizard ☑ x. tropicalis ☐ tetraodon ☐ fugu ☐ stickleback
☐ medaka ☑ zebrafish ☐ lamprey

Submit

For information about output data format see the User's Guide

```
>uc011adk.1_hg19 1453 chr21:32496787-32639288-
MGNAESQHVEHEFYGEKHASLGRKHTSRSLRLSHKTRRTRHASSGKVIHRNSEVSTRSSSTPSIPQSLAENGLEPFSQDGTLEDFGSPIWVDRVDMGLR
>uc011adk.1_panTro2 1453 chr21:30868972-31012391-
MGNAESQHVEHEFYGEKHASLGRKHTSRSLRLSHKTRRTRHASSGKVIHRNSEVSTRSSSTPSIPQSLAENGLEPFSQDGTLEDFGSPIWVDRVDMGLR
>uc011adk.1_gorGor1 1453 Supercontig_0003507:449-38073+;Supercontig_0134003:870-956-;Supercontig_00
----------------HASLGRKHTSRSLRLSHKTRRTRHASSGKVIHRNSEVSTRSSSTPSIPQSLAENGLEPFSQDGTLEDFGSPIWVDRVDMGLR
>uc011adk.1_rheMac2 1453 chr3:15243594-15386873+
```

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 - is a virtual microscope for viewing in situ images. .

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

WIG format

1)fixedStep

fixedStep	chrom=chr1 start=3001 step=1
24	
56	
100	

variableStep	chrom=chr1
3001	24
3002	56
3003	100

2)variableStep

variableStep	chrom=chr1
3001	24
3003	56
3010	100

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
chr2	106000	108000	0.75

File Formats

FASTA

```
>HWI-ST398_0092:1:1:5372:2486#0/1
TTTTTCGTTCTTTTCATGTACCGCTTTTTGTTCGGTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAT
```

FASTQ

```
@HWI-ST398_0092:6:73:5372:2486#0/1
TTTTTCGTTCTTTTCATGTACCGCTTTTTGTTCGGTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAT
+HWI-ST398_0092:1:1:5372:2486#0/1
ffffeedfcedffffeffdefff_fffffdccfdZdeeadefecZedaecdbRdTY^ZYT``_T`_^bc_Wceaa[
```

6 - Flowcell lane

73 - Tile number

5372:2486 - 'x','y'-coordinates of the cluster within the tile

#0 - index number for a multiplexed sample (0 for no indexing)

/1 - the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

File Formats- SAM

```
8_100_10000_12419	163	chrVII	271183	255	40M	=	271294	151	TGGTGTATTATACGCTACCGTGCGGTGCCGGGGGCAACCG	bbbabbbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbbb	XA:i:0	MD:Z:40	NM:i:0
```

8_100_10000_12419	163	chr7	271183	255	40M	=	271294	151	TGGTGTATTATACGCTACCGTGCGGTGCCGGGGGCAACCG	bbbabbbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbbb	XA:i:0 MD:Z:40 NM:i:0
QNAME	**FLAG**	**RNAME**	**POS**	**MAPQ**	**CIGAR**	**MRNM**	**MPOS**	**TLEN**	**SEQ**	**QUAL**	**OPT**

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe (‘=’ if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

BTEP

File Formats BAM

BAM (*.bam) is the compressed binary version of the Sequence Alignment/Map (SAM) format, a compact and index-able representation of nucleotide sequence alignments. **BAM** is compressed in the **BGZF** format. BGZF files support random access through the BAM file index (*.bam.bai).

***BGZF** is block compression implemented on top of the standard gzip file format. The goal of BGZF is to provide good compression while allowing efficient random access to the BAM file for indexed queries. The BGZF format is 'gunzip compatible', in the sense that a compliant gunzip utility can decompress a BGZF compressed file.*

File Formats GFF/GTF

- GFF (General Feature Format)
- GTF (Gene Transfer Format)

1. **seqname** - The name of the sequence. Must be a chromosome or scaffold.
2. **source** - The program that generated this feature.
3. **feature** - The name of this type of feature. Some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
4. **start** - The starting position of the feature in the sequence. The first base is numbered 1.
5. **end** - The ending position of the feature (inclusive).
6. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). If there is no score value, enter ".".
7. **strand** - Valid entries include '+', '-', or '.' (for don't know/don't care).
8. **frame** - If the feature is a coding exon, *frame* should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
9. **group** - All lines with the same group are linked together into a single item.

GTF is a refined form of the GFF with group attributes

- **gene_id *value*** - A globally unique identifier for the genomic source of the sequence.
- **transcript_id *value*** - A globally unique identifier for the predicted transcript.

GFF3 http://www.sequenceontology.org/resources/gff3.html

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/**

BTEP

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 at 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

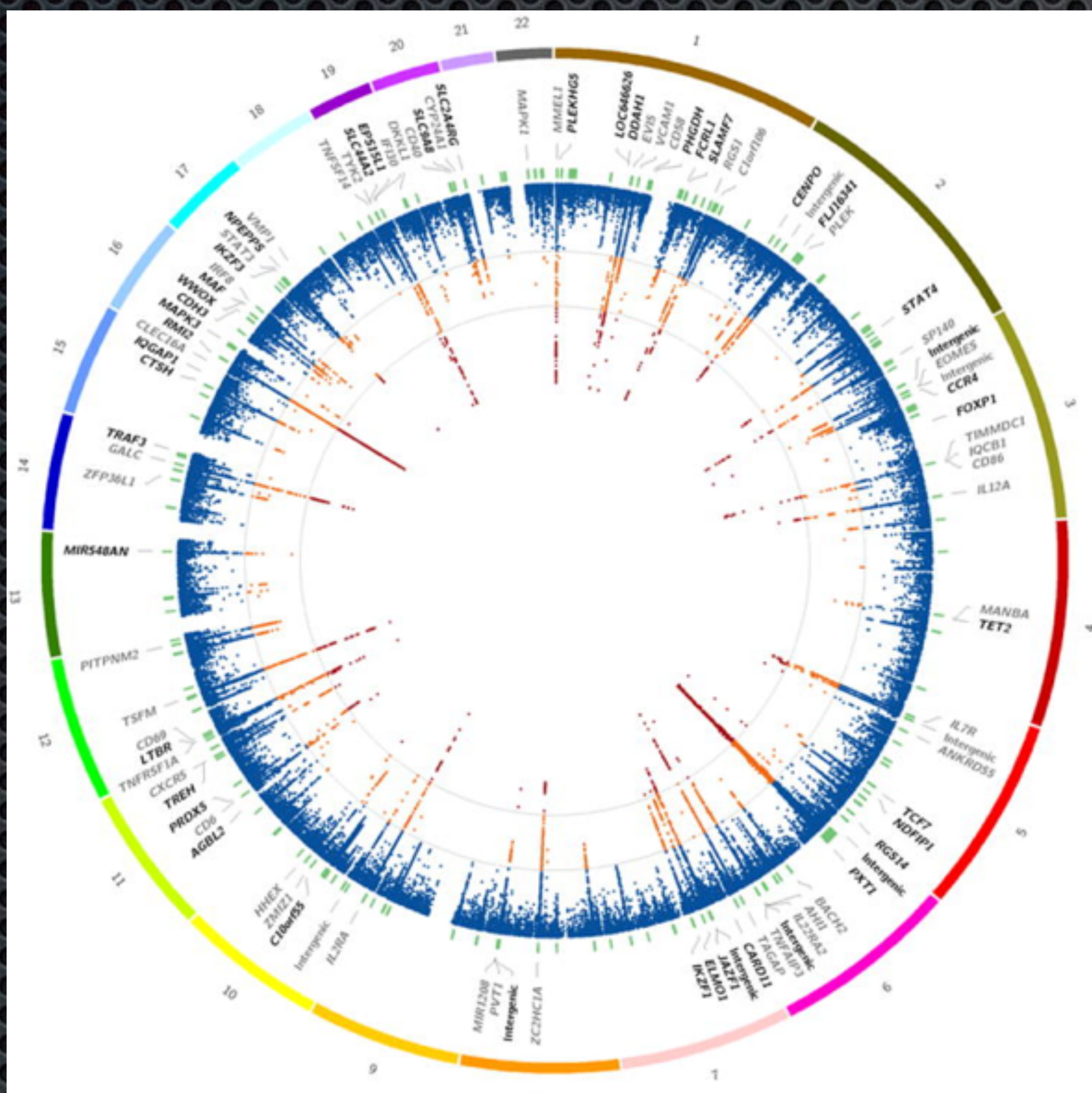
The Integrative Genomics Viewer IGV

- 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

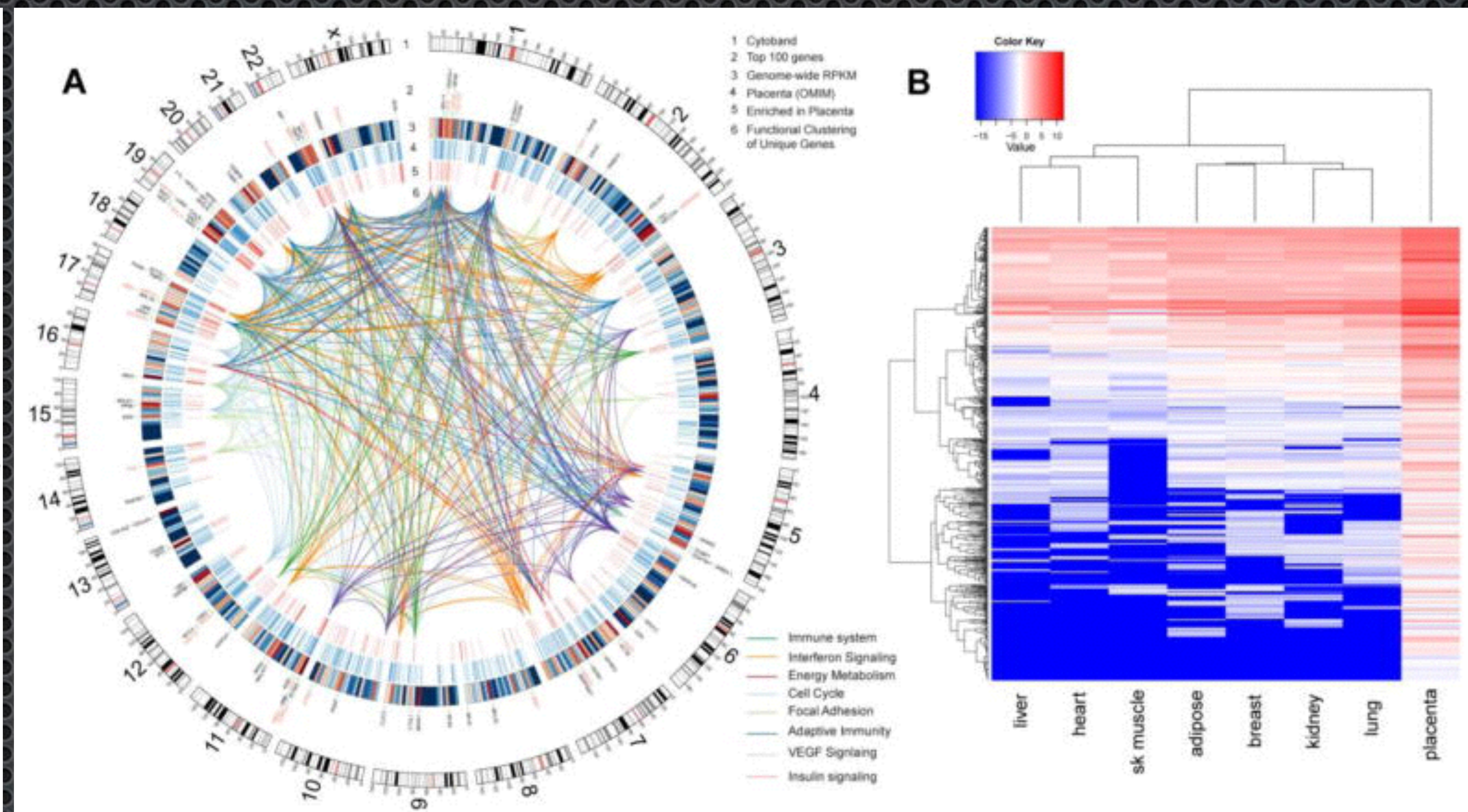
Circular Maps

- CGView - Circular Genome Viewer
  http://wishart.biology.ualberta.ca/cgview/index.html
- Circos
  http://circos.ca

Circos

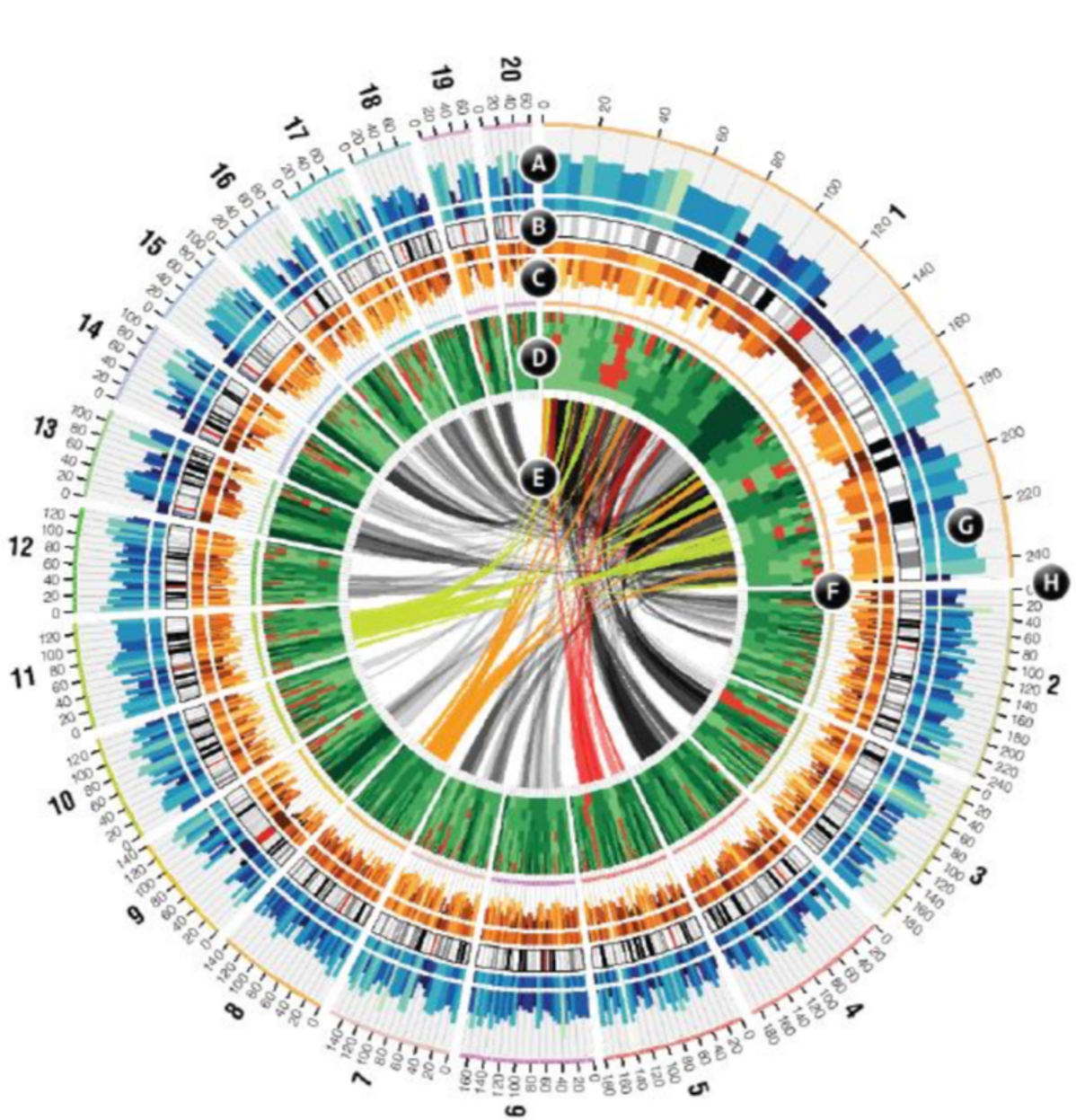


MIR548AN
STAT4
FOXP1
TET2



A
1 Cytoband
2 Top 100 genes
3 Genome-wide RPKM
4 Placenta (OMIM)
5 Enriched in Placenta
6 Functional Clustering of Unique Genes
Immune system
Interferon Signaling
Energy Metabolism
Cell Cycle
Focal Adhesion
Adaptive Immunity
VEGF Signaling
Insulin signaling
B
Color Key
Value
liver
heart
sk muscle
adipose
breast
kidney
lung
placenta

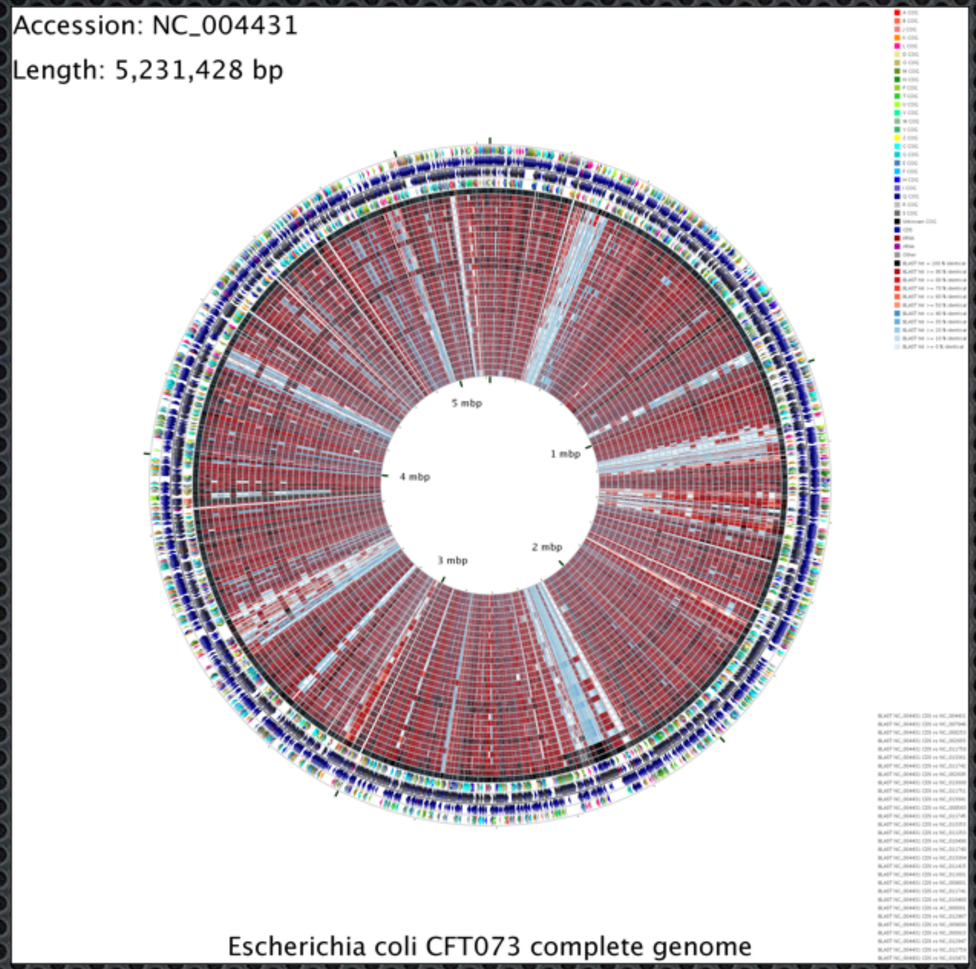
Circos



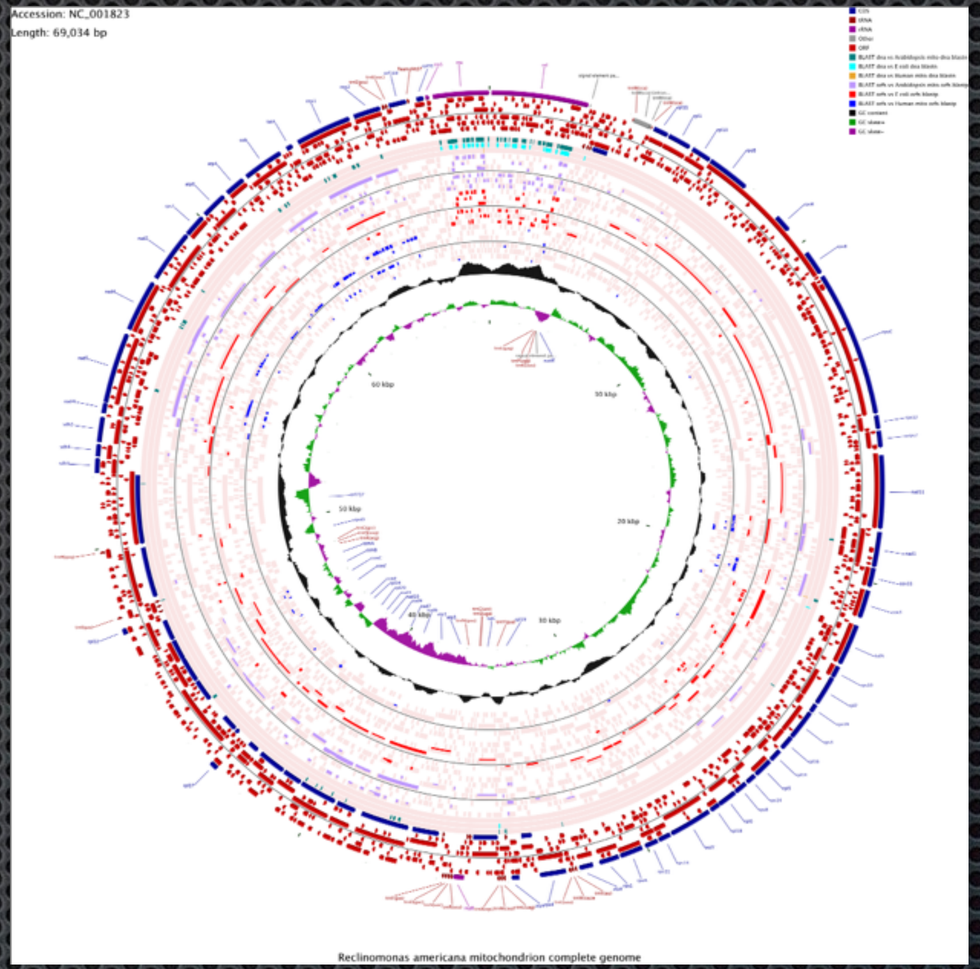
Available Tracks/Display:

- A: Histogram
- B: Ideogram
- C: Histogram (inverted)
- D: Heatmap
- E: Links
- F: Highlights
- G: Grid
- H: Ticks

CGView - Circular Genome Viewer



Escherichia coli CFT073 complete genome



Reclinomonas americana mitochondrion complete genome

Center for Cancer Research

https://bioinformatics.cancer.gov/btep

ncibtep@mail.nih.gov

Upcoming Events

- **iPathwayGuide Workshop**
  12/19/2016 - 9:30am to 4:00pm
  Cordelia Ziraldo, Andrew Olson

- **R/Bioconductor Basics Workshop**
  12/20/2016 - 9:30am to 4:00pm
  David Wheeler