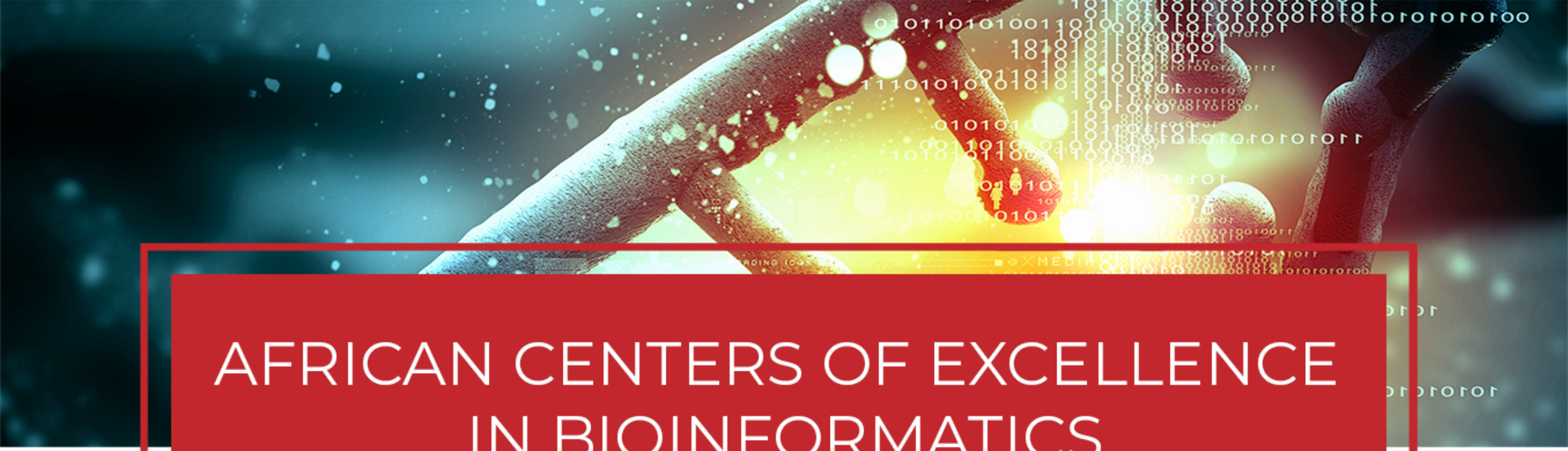


Genome Browsers:
UCSC Genome Browser Workshop
February 23, 2021

Today's Instructor

Andrew Oler, PhD

Senior Bioinformatics Scientist

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: bioinformatics@niaid.nih.gov
  - Instructors:
    - andrew.oler@nih.gov

Topics

- Genome Browser basic concepts
- Highlighting Various Genome Browsers
- UCSC Genome Browser Tutorial
  - Basic Navigation
  - Using tracks
  - Table Browser
  - Other tools

Why Genome Browsers?

Sequence is not enough

Full human genome sequenced printed at 4pt font comprises 130 volumes – double-sided, 43K characters per page

https://www2.le.ac.uk/news/blog/2012/december/you-in-130-volumes-entire-human-genome-printed-for-exhibition

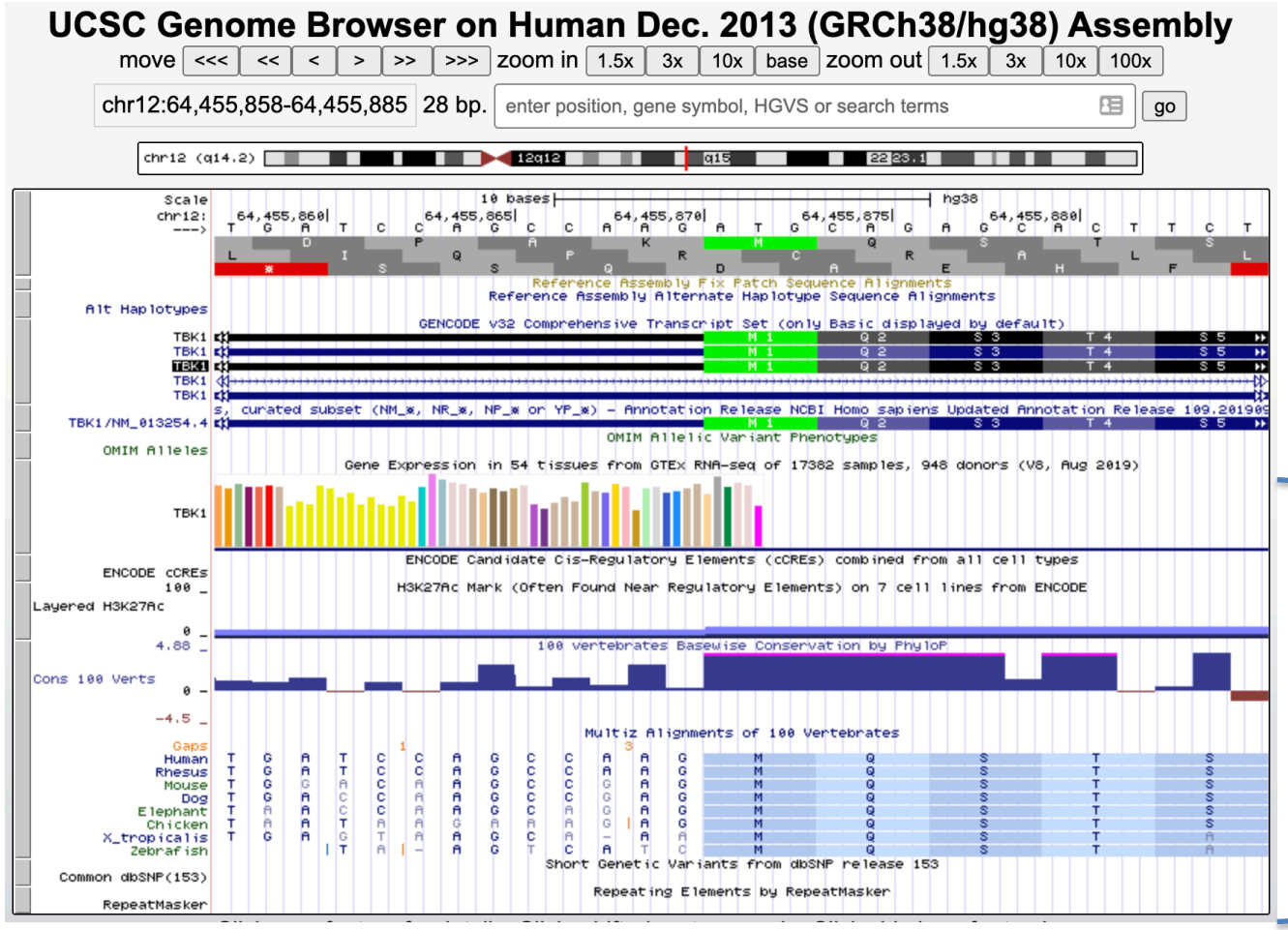
https://www.bio-itworld.com/news/2012/12/28/encyclopedia-genomica-uk-scientists-print-the-book-of-life-in-130-volumes

Genome browser basics

- Genome sequence alone doesn't explain the biology
- Genome Browsers give researchers access to **Annotations**
  - Genes/Protein
  - Gene Expression
  - Regulation/Epigenetics
  - Variation
  - Conservation
- Annotations allow researchers to understand and interpret the sequence in its biological context

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

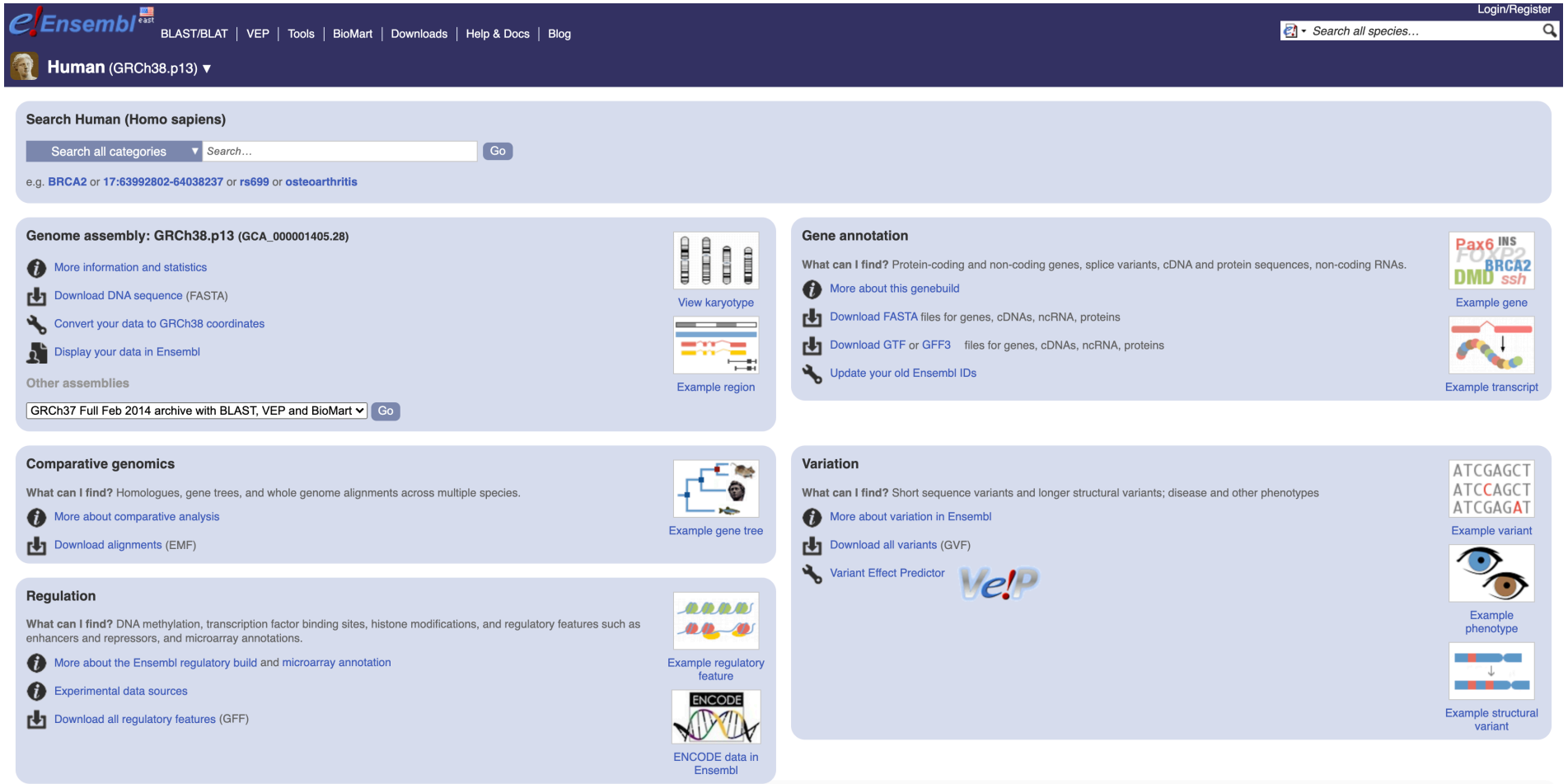
Genome browser basics



- Genome species and version
- Chromosomal coordinates
- Sequence (and translation)
- Genes
- Variants
- Other Annotations

Genome Browsers: Ensembl

- www.ensembl.org

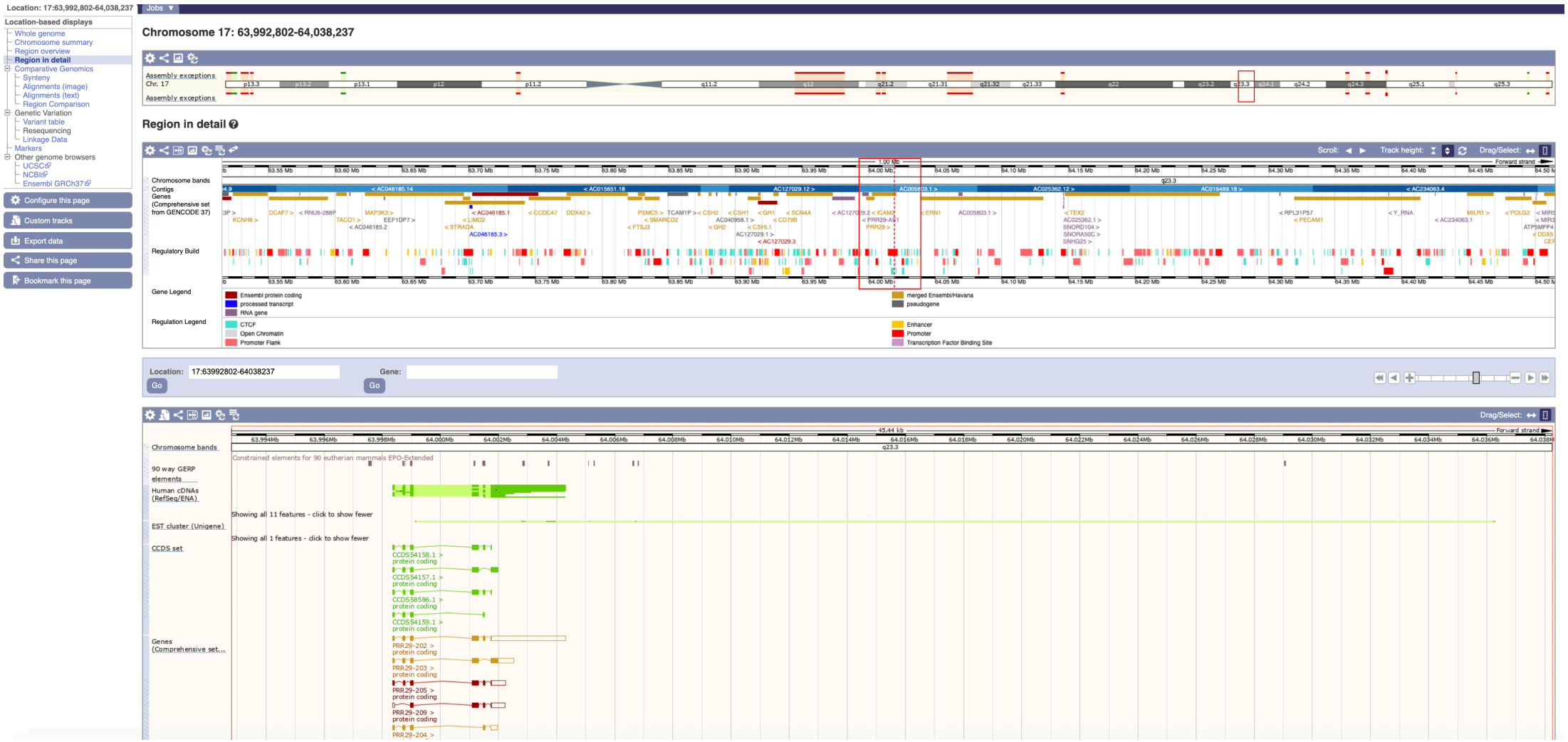


"Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species."

Ensembl

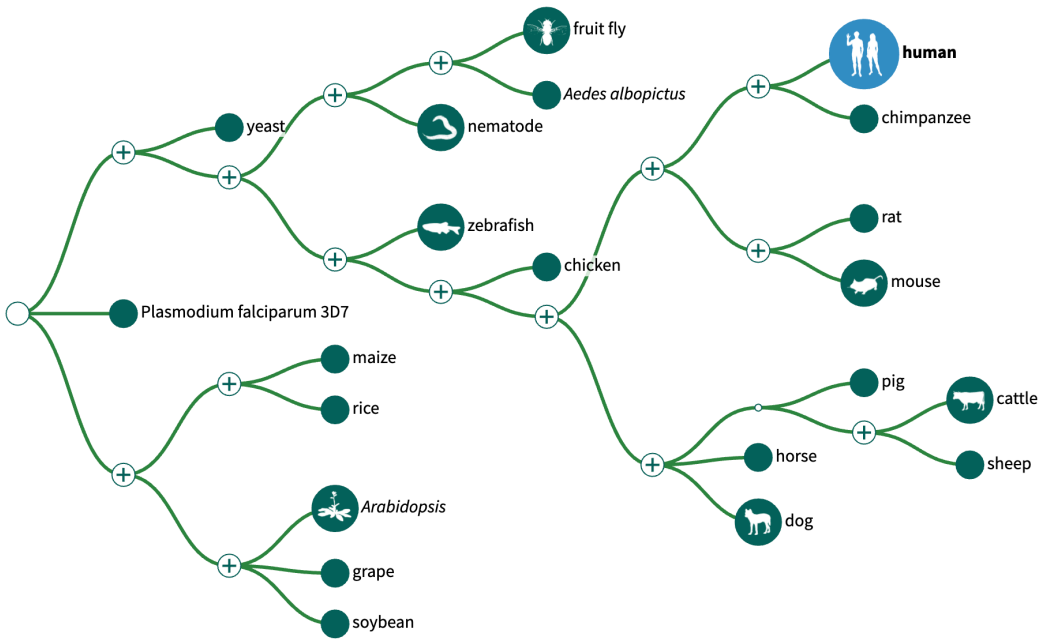
- 310 species available (vertebrates)
  - https://www.ensembl.org/info/about/species.html
- Gene annotations
- Tools available for analysis, e.g., Variant Effect Predictor (VEP)

Ensembl

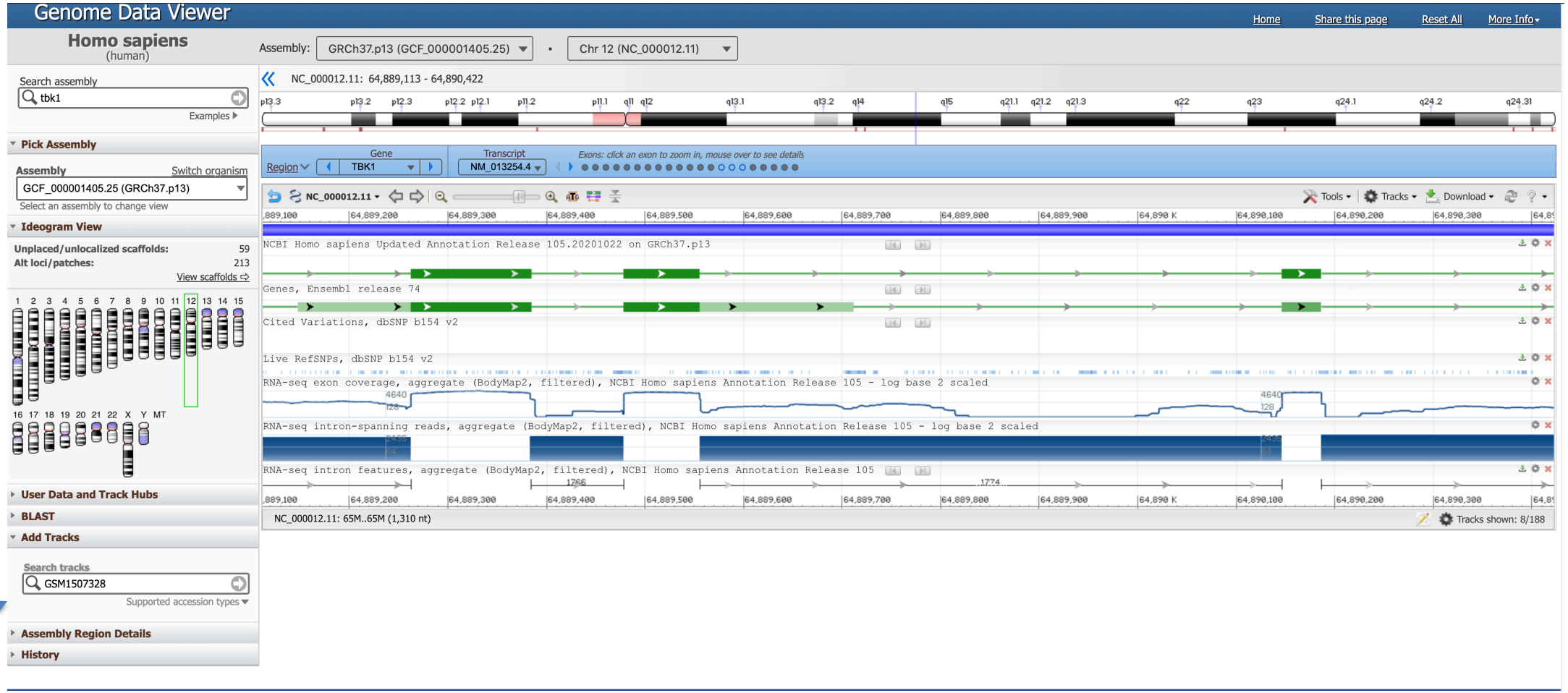

Location: 17:63,992,802-64,038,237
Jobs
Location-based displays
Whole genome
Chromosome summary
Region overview
Region in detail
Comparative Genomics
Synteny
Alignments (image)
Alignments (text)
Region Comparison
Genetic Variation
Variant table
Resequencing
Linkage Data
Markers
Other genome browsers
UCSC
NCBI
Ensembl GRCh37
Configure this page
Custom tracks
Export data
Share this page
Bookmark this page
Chromosome 17: 63,992,802-64,038,237
Assembly exceptions
Chr. 17
Region in detail
Scroll:
Track height:
Drag/Select:
Chromosome bands
Contigs
Genes
(Comprehensive set from GENCODE 37)
Regulatory Build
Gene Legend
Ensembl protein coding
processed transcript
RNA gene
merged Ensembl/Havana
pseudogene
Regulation Legend
CTCF
Open Chromatin
Promoter Flank
Enhancer
Promoter
Transcription Factor Binding Site
Location:
17:63992802-64038237
Gene:
Go
Constrained elements for 90 eutherian mammals EPO-Extended
90 way GERP elements
Human cDNAs (RefSeq/ENA)
Showing all 11 features - click to show fewer
EST cluster (Unigene)
Showing all 1 features - click to show fewer
CCDS set
CCDS54158.1 > protein coding
CCDS54157.1 > protein coding
CCDS58586.1 > protein coding
CCDS54159.1 > protein coding
Genes (Comprehensive set...
PRR29-202 > protein coding
PRR29-203 > protein coding
PRR29-205 > protein coding
PRR29-209 > protein coding
PRR29-204 >

Genome Browsers: NCBI Genome Data Viewer (GDV)

- https://www.ncbi.nlm.nih.gov/genome/gdv/
- Currently supports 721 eukaryotic genome assemblies (> 1000 genome versions)



NCBI Genome Data Viewer

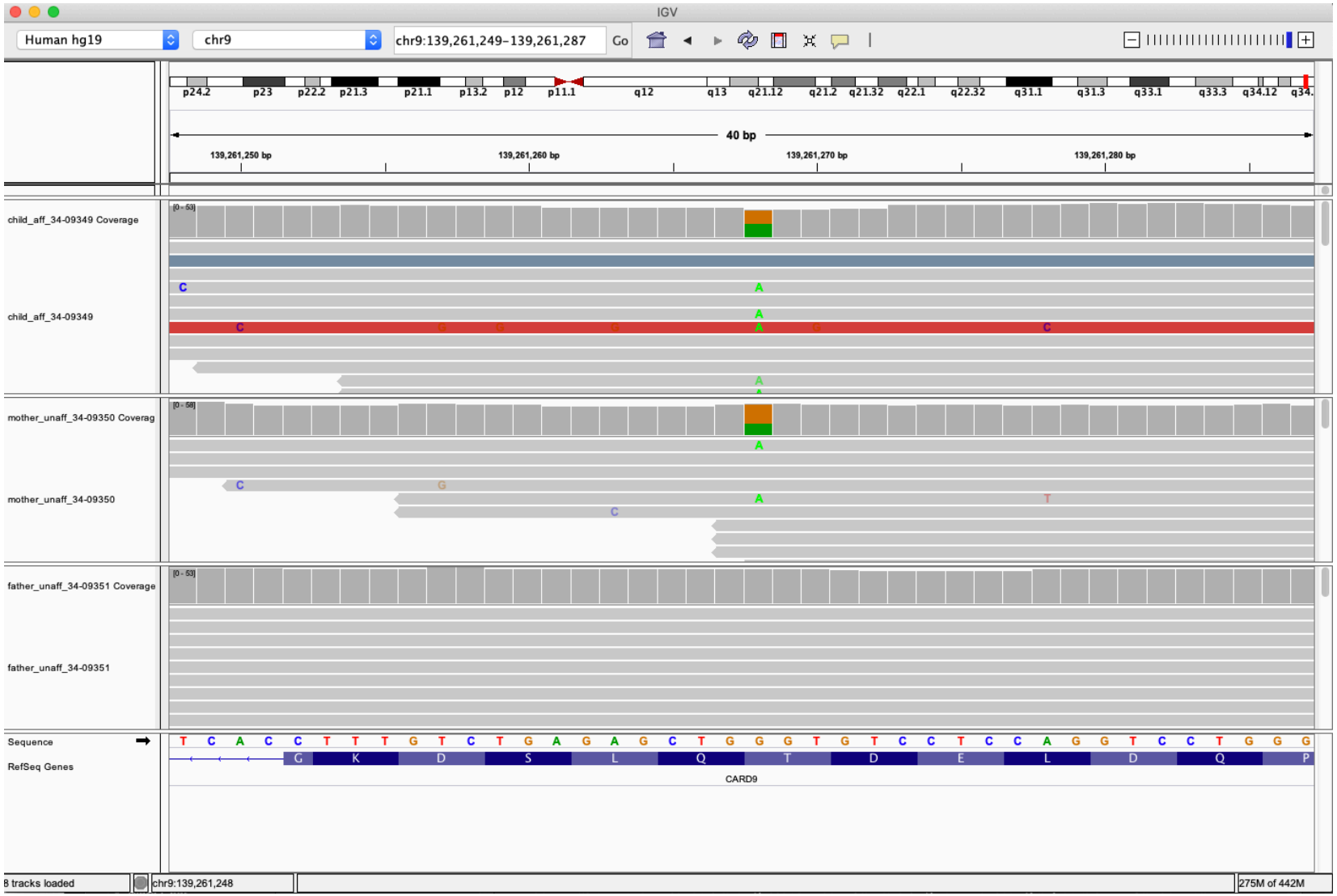


- Useful for loading aligned data available in NCBI (SRA, GEO)

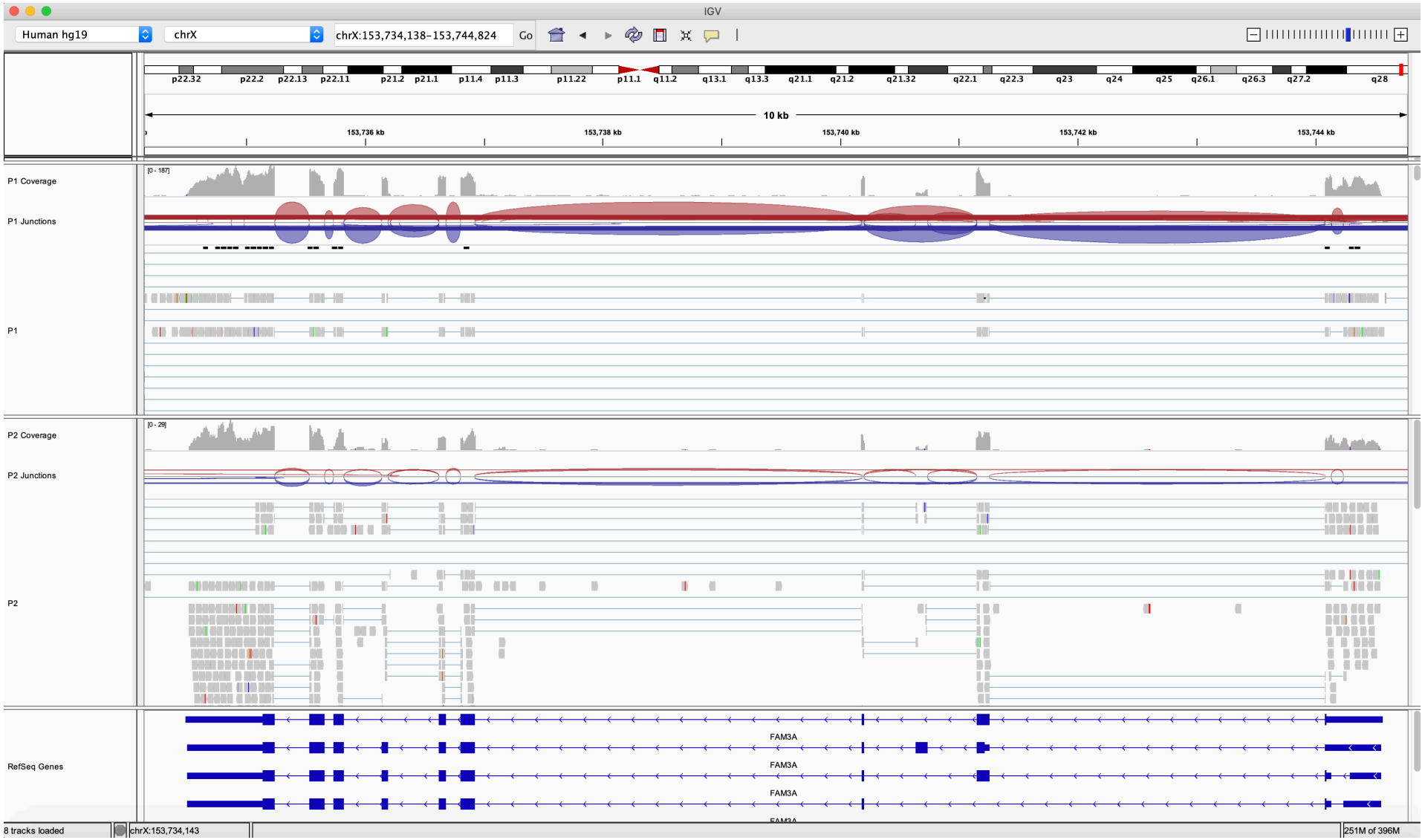
Genome Browsers: Integrated Genome Viewer (IGV)

- Desktop genome browser, developed by the Broad Institute
  - http://software.broadinstitute.org/software/igv/
- "The **Integrative Genomics Viewer (IGV)** is a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data. It supports flexible integration of all the common types of genomic data and metadata, investigator-generated or publicly available, loaded from local or cloud sources."
- Helpful for visualization of raw sequencing data stored **locally**, e.g., BAM files
- 87 genomes/versions supported currently
  - Includes many vertebrates (human, mouse, chicken, cow, etc.) as well as many model organisms and pathogens (*C. elegans, P. falciparum, E. coli, M. tuberculosis, S. cerevisiae, A. gambia*, etc.)
- Supports a wide variety of file types and views
  - BAM/CRAM (read alignment file – DNA sequencing or RNA-seq)
  - BED (intervals, including chromosome, start, end, name, score, strand)
  - bigWig (coverage representation of read depth)

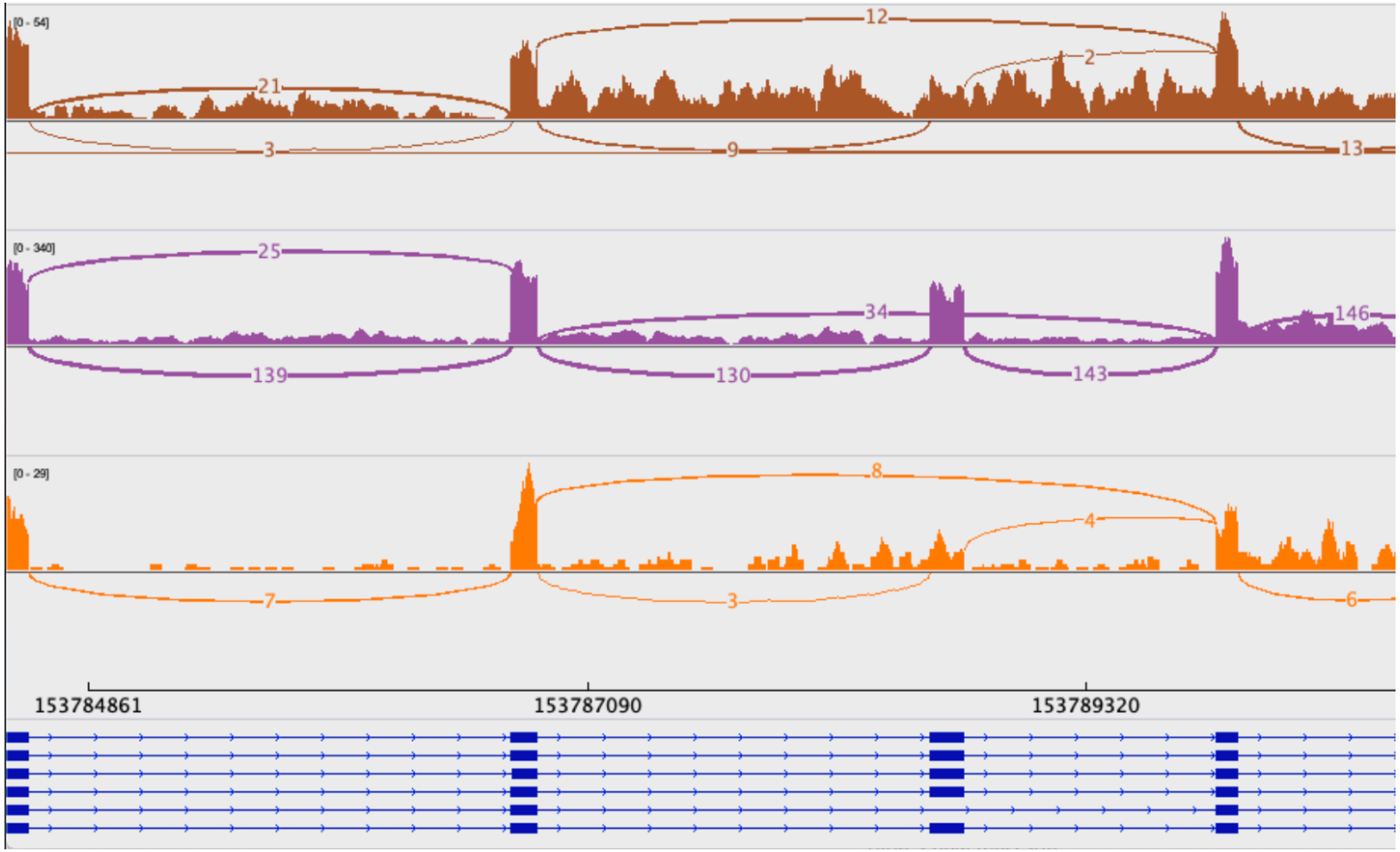
IGV – DNA-seq BAM files to visualize variants



IGV RNA-seq BAM files to visualize expression and splicing



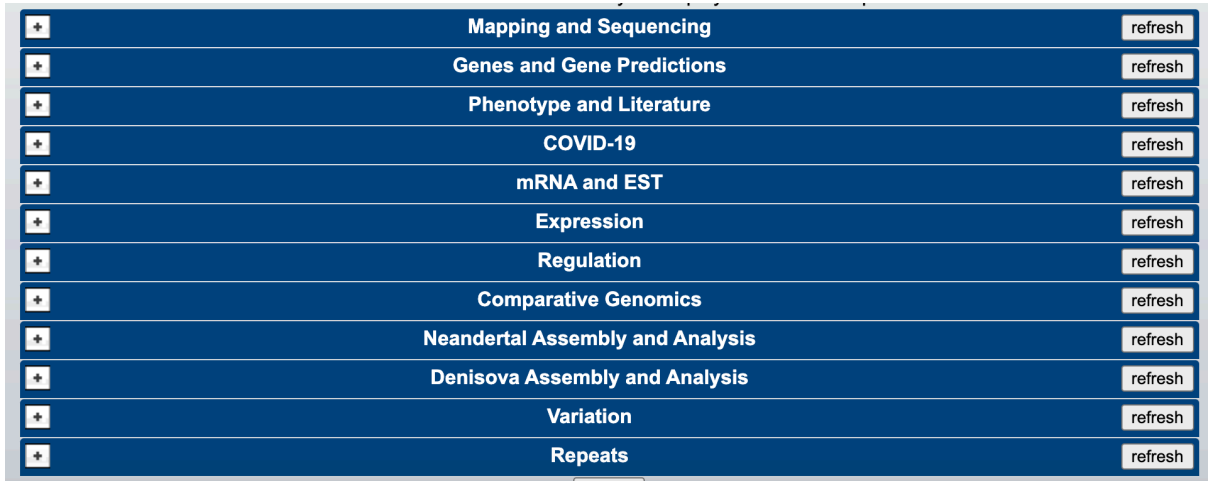
IGV – Sashimi Plots



http://software.broadinstitute.org/software/igv/Sashimi

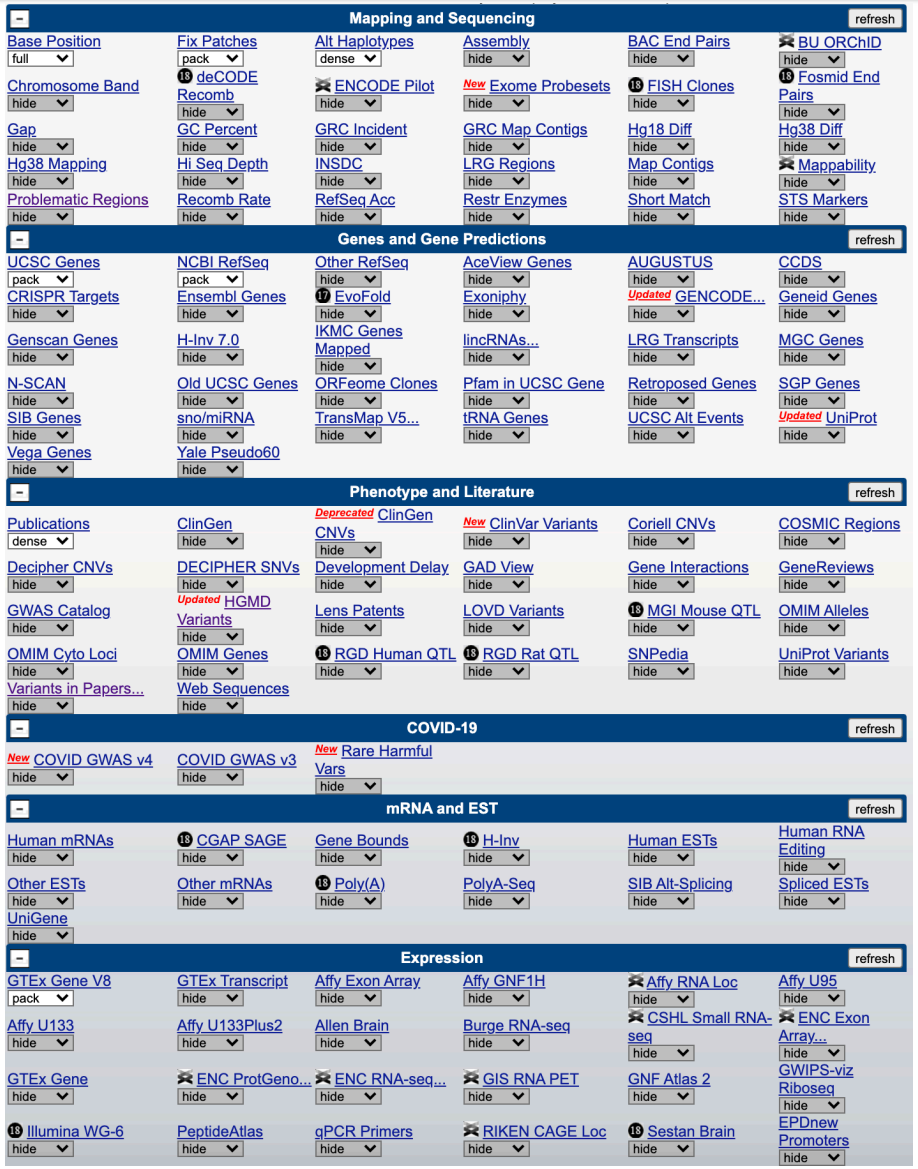
Genome Browsers: UCSC Genome Browser

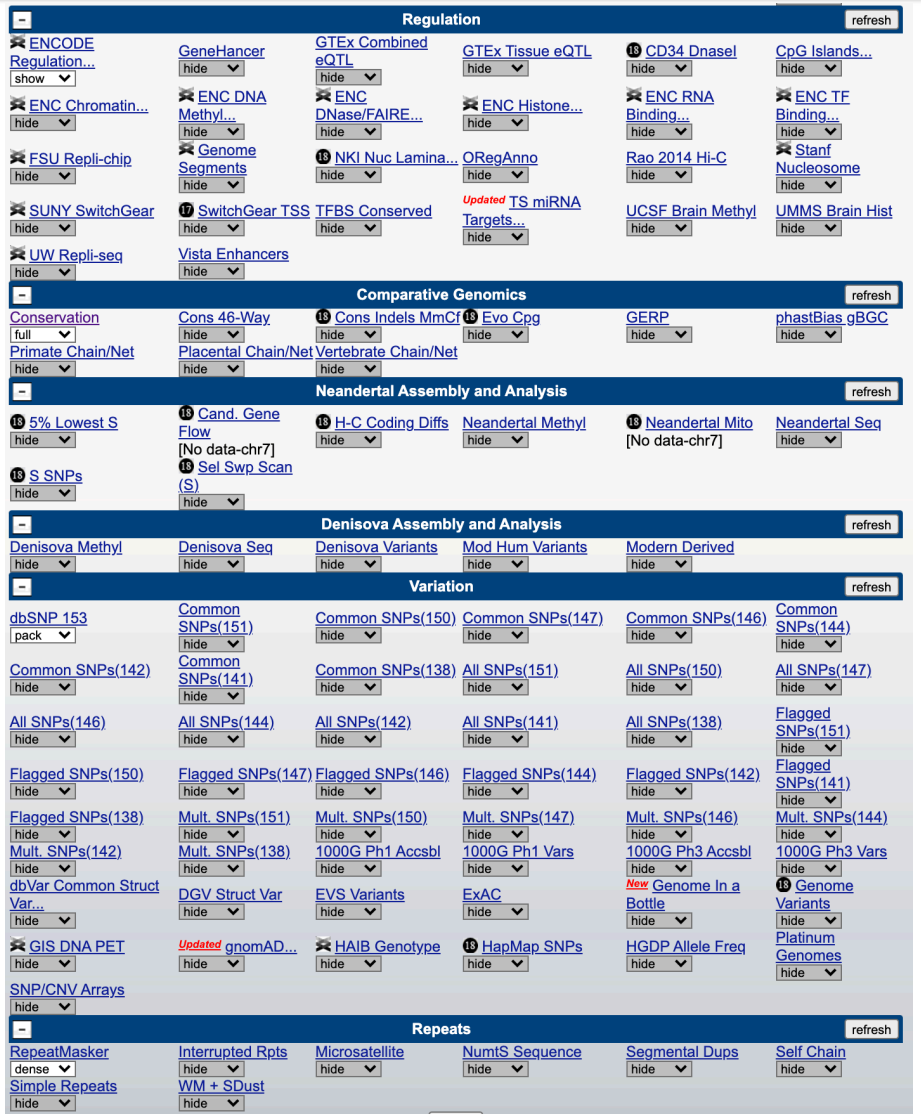
- https://genome.ucsc.edu/
- Supports a variety of species and model organisms – 110 species (240 versions total)
  - http://genome.ucsc.edu/FAQ/FAQreleases.html#release1
- Gene annotations
- Large number of annotations for human genomes
  - ENCODE data from 2003-2012 (https://genome.ucsc.edu/ENCODE/index.html)
  - Disease variants
  - Gene expression
  - Links to external websites
- Table browser
- Liftover
- More…



Annotation track categories

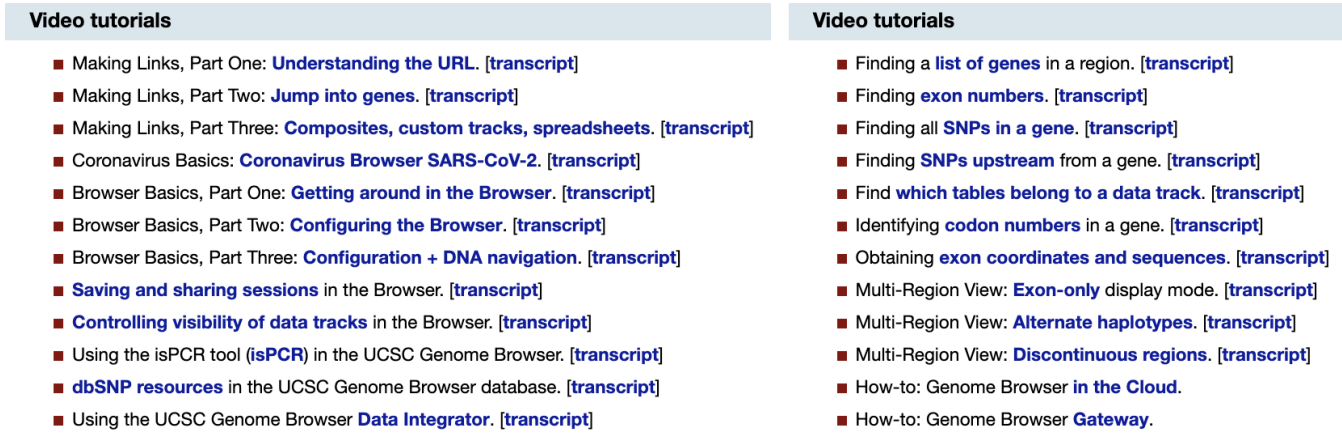
UCSC Genome Browser – Human hg19/GRCh37 annotations





UCSC Genome Brower - Online Training and Tutorials

- https://genome.ucsc.edu/training/
- Video Tutorials



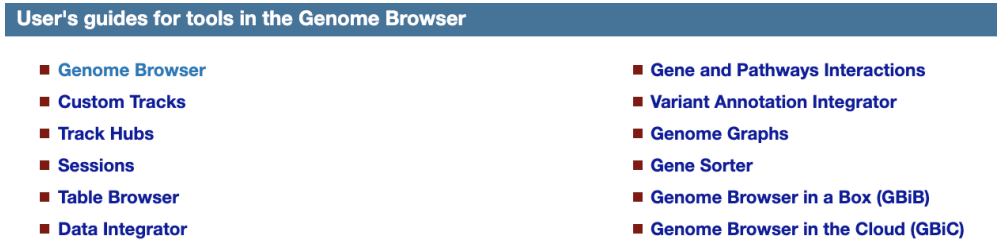
Video tutorials

- Making Links, Part One: **Understanding the URL**. [transcript]
- Making Links, Part Two: **Jump into genes**. [transcript]
- Making Links, Part Three: **Composites, custom tracks, spreadsheets**. [transcript]
- Coronavirus Basics: **Coronavirus Browser SARS-CoV-2**. [transcript]
- Browser Basics, Part One: **Getting around in the Browser**. [transcript]
- Browser Basics, Part Two: **Configuring the Browser**. [transcript]
- Browser Basics, Part Three: **Configuration + DNA navigation**. [transcript]
- **Saving and sharing sessions** in the Browser. [transcript]
- **Controlling visibility of data tracks** in the Browser. [transcript]
- Using the isPCR tool (**isPCR**) in the UCSC Genome Browser. [transcript]
- **dbSNP resources** in the UCSC Genome Browser database. [transcript]
- Using the UCSC Genome Browser **Data Integrator**. [transcript]

Video tutorials

- Finding a **list of genes** in a region. [transcript]
- Finding **exon numbers**. [transcript]
- Finding all **SNPs in a gene**. [transcript]
- Finding **SNPs upstream** from a gene. [transcript]
- Find **which tables belong to a data track**. [transcript]
- Identifying **codon numbers** in a gene. [transcript]
- Obtaining **exon coordinates and sequences**. [transcript]
- Multi-Region View: **Exon-only** display mode. [transcript]
- Multi-Region View: **Alternate haplotypes**. [transcript]
- Multi-Region View: **Discontinuous regions**. [transcript]
- How-to: Genome Browser **in the Cloud**.
- How-to: Genome Browser **Gateway**.

- User Guides



User's guides for tools in the Genome Browser

- Genome Browser
- Custom Tracks
- Track Hubs
- Sessions
- Table Browser
- Data Integrator
- Gene and Pathways Interactions
- Variant Annotation Integrator
- Genome Graphs
- Gene Sorter
- Genome Browser in a Box (GBiB)
- Genome Browser in the Cloud (GBiC)

- Upcoming workshops

April 13, 2021	**ACMG 2021** ~~Los Angeles~~ --> virtual
June 13, 2021	**ESHG** ~~Glasgow~~ --> virtual

Using the Browser: Navigation and Configuration

- Tour through the Browser
  - Click on “Genome”
  - Choosing a species, genome version (e.g, hg19)
  - Searching for gene of interest or for Genbank accession numbers, etc. in search bar (examples below). E.g., MAF1.
  - Navigation:
    - Zooming in, zooming out
    - selecting region with chromosome image
    - panning to right/left with arrows or click and drag
    - Selecting region with “scale” bar (Rectangular selection)
    - or manually change coordinates
  - Configuration Buttons
    - Configuration to widen the view
    - “resize” button
    - “default tracks” button
      - Gene annotation, Repeats, Mammalian Conservation, some summary ENCODE data tracks

Using the Browser: Annotation Tracks

- Gene annotation
  - UCSC Genes track
  - Click on a gene, lots of information. E.g., MAF1, SHARPIN
  - Coding exons (thick), introns (wire), untranslated regions (thin). E.g., KIAA1875.
    - Arrows in intron show which strand is expressed
    - Exons match conservation
    - Repeats found mostly in introns
- Data Tracks
  - Peaks vs. Signals
    - Peaks (BED or BigBed) are intervals, i.e., chr, start, end
    - Signal (WIG or BigWig) are graphs.
  - Visibility levels (hide, dense, squish, pack, full), e.g., Mapability

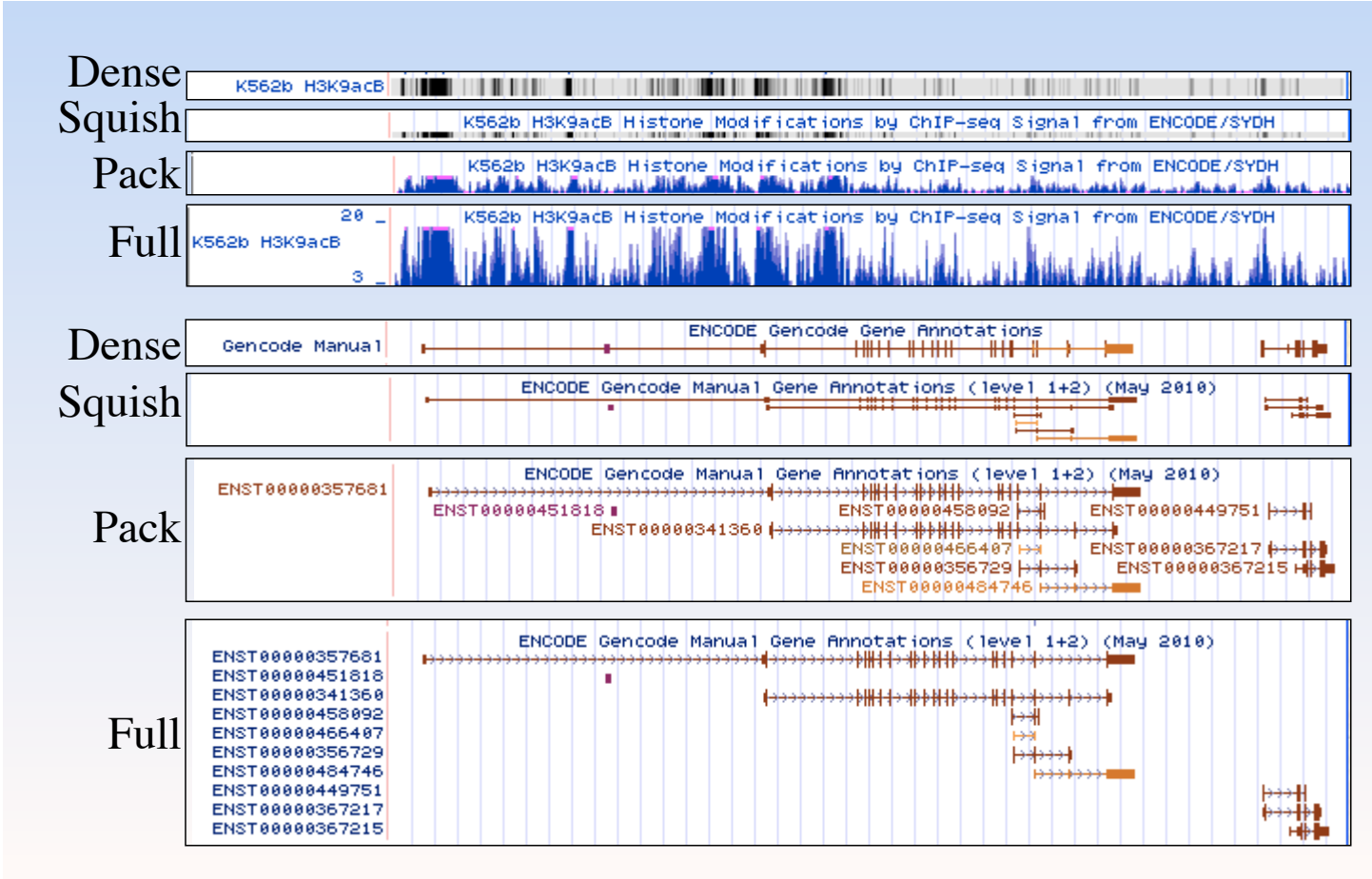
Annotation Tracks by Group

- Data Annotation tracks, grouped by type. *Look for NHGRI symbol for ENCODE data*
  - Mapping and Sequencing – need to be aware of context of the data, can affect your confidence in the data. (e.g., Mapability) Usually, unique regions are highly mappable, leading to highly confident data.
    - # Stop to demonstrate turning on and modifying tracks with Mapability as an example (next slide)
  - Phenotype and Disease Associations
  - Genes and Gene Prediction tracks. E.g., UCSC, Refseq, Ensembl, usually open by default.
  - mRNA and EST tracks – can turn on if desired.
  - Expression – some microarray, some RNA-seq. Anything with NHGRI symbol means ENCODE data.
    - Super-tracks
  - Regulation – ChIP-seq, DNaseI HS, histone modifications etc. Most of ENCODE data here.
  - Comparative Genomics – how conserved it is, e.g., peaks of conservation
  - Variation and Repeats – common SNPs (dbSNP137), Repeatmasker. -- Along with mappability, it is good to be aware of what repeats are present, as they can also affect mappability and the results. E.g., satellites are a common false positive in ChIP-seq data.

Modifying Track Visibility

- Turning tracks on/off down below.
  - Choose default by just choosing a visibility and click refresh
  - Click on link and configure, then click submit.
- Modifying tracks in the browser view.
  - Change visibility (multiple levels of visibility) – click or right-click and change.
  - Configure track vs. track set (group of tracks, turn on/off, other)
  - Change visibility of *peaks* track too (expands to see all overlapping peaks)

Display Visibility Settings



Additional files for demonstration

- https://github.com/niaid/Genome-Browsers
  - Slides: Genome_Browsers_2021.pdf
  - Exercise Handout: UCSC_Genome_Browser_exercises_2021-02-23.pdf
  - Input file for BLAT exercise: BEAS2B_Udorn_24h_TBK1_RNAseq_reads.txt