



AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

**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

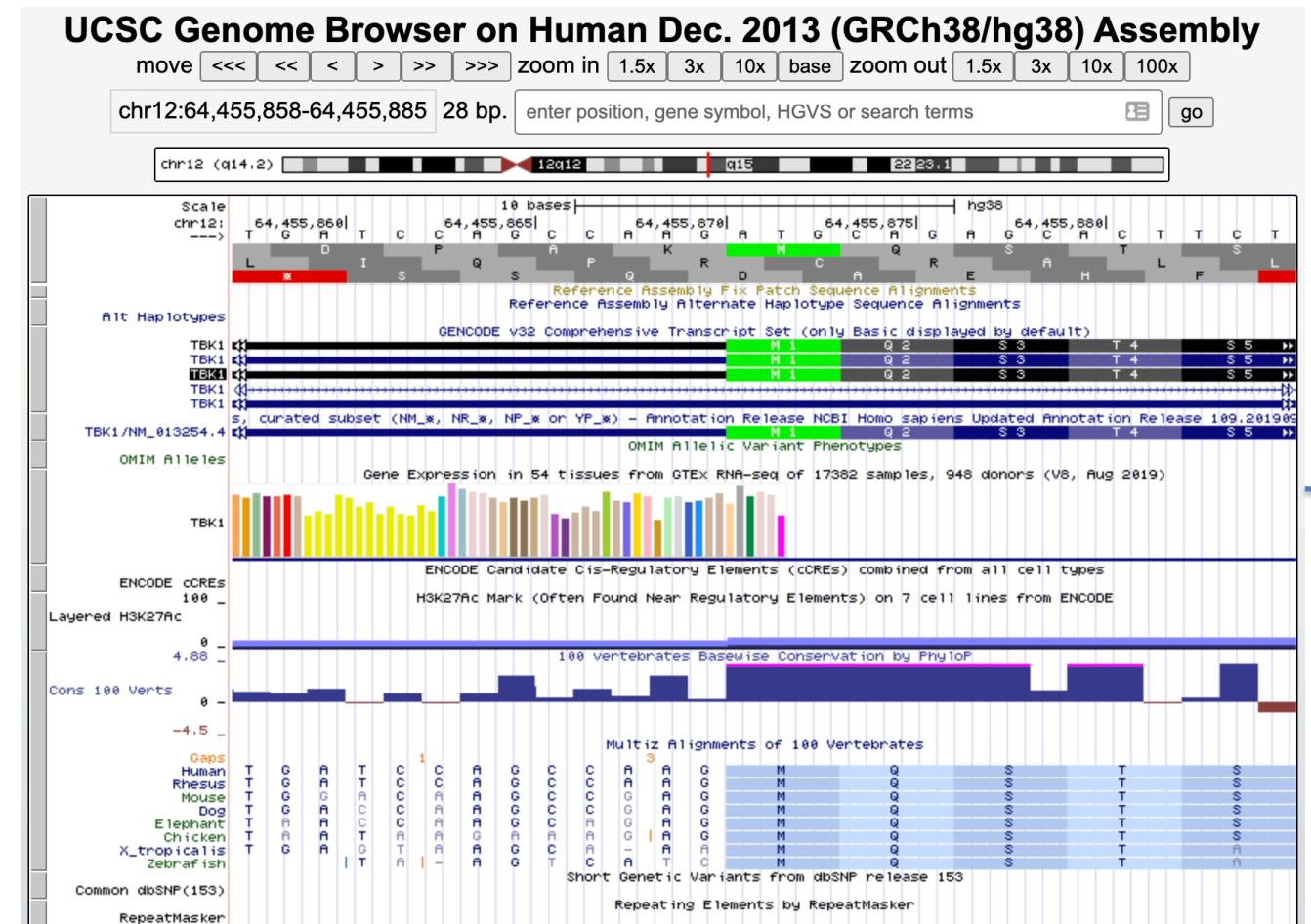
<https://www2.le.ac.uk/news/blog/2012/december/you-in-130-volumes-entire-human-genome-printed-for-exhibition>

<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/Epigentetics
 - Variation
 - Conservation
- Annotations allow researchers to understand and interpret the sequence in its biological context

Genome browser basics



Genome species and version

Chromosomal coordinates

Sequence (and translation)

Genes

Variants

Other Annotations

Genome Browsers: Ensembl

- www.ensembl.org

The screenshot shows the Ensembl homepage for the Human genome (GRCh38.p13). At the top, there's a search bar with placeholder text "Search all species...", a login/register link, and a species dropdown set to "Human (GRCh38.p13)". Below the header, there's a navigation bar with links to BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and a Blog.

Search Human (Homo sapiens)

Search all categories ▾ Search... Go

e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis

Genome assembly: GRCh38.p13 (GCA_000001405.28)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

Other assemblies
GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾ Go

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

Pax6 INS FOXP2 DMD ssh
Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

- More about comparative analysis
- Download alignments (EMF)

Example gene tree

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes.

- More about variation in Ensembl
- Download all variants (GVF)
- Variant Effect Predictor

Ve!P

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

- More about the Ensembl regulatory build and microarray annotation
- Experimental data sources
- Download all regulatory features (GFF)

Example regulatory feature

ENCODE data in Ensembl

Example variant

Example phenotype

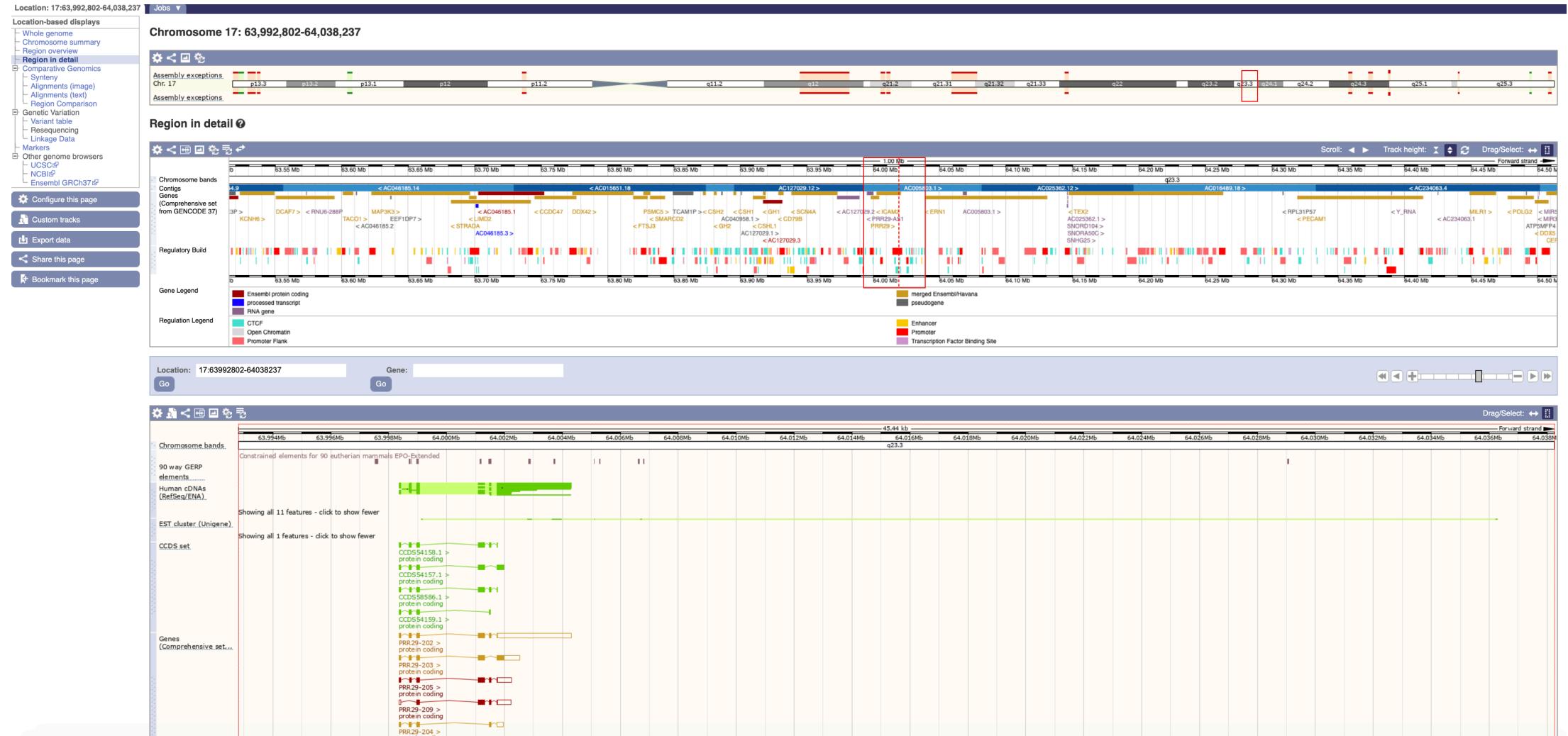
Example structural variant

"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

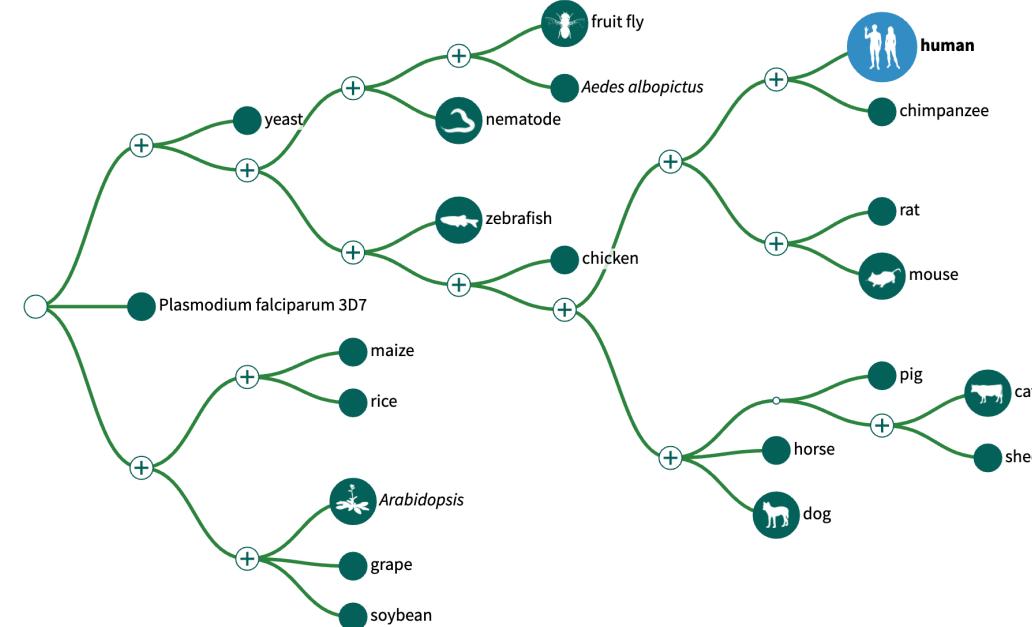
- 310 species available (vertebrates)
 - <https://www.ensembl.org/info/about/species.html>
- Gene annotations
- Tools available for analysis, e.g., Variant Effect Predictor (VEP)

Ensembl

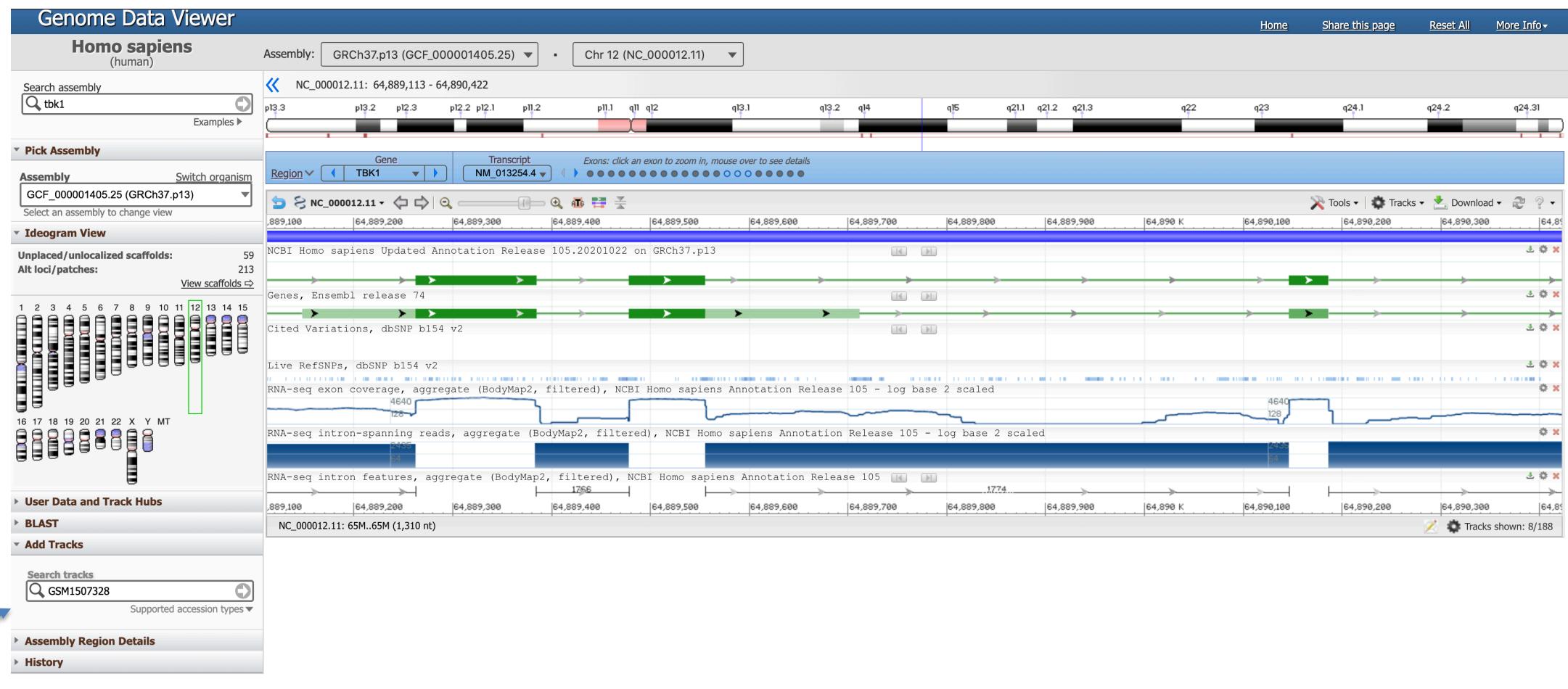


Genome Browsers: NCBI Genome Data Viewer (GDV)

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



NCBI Genome Data Viewer

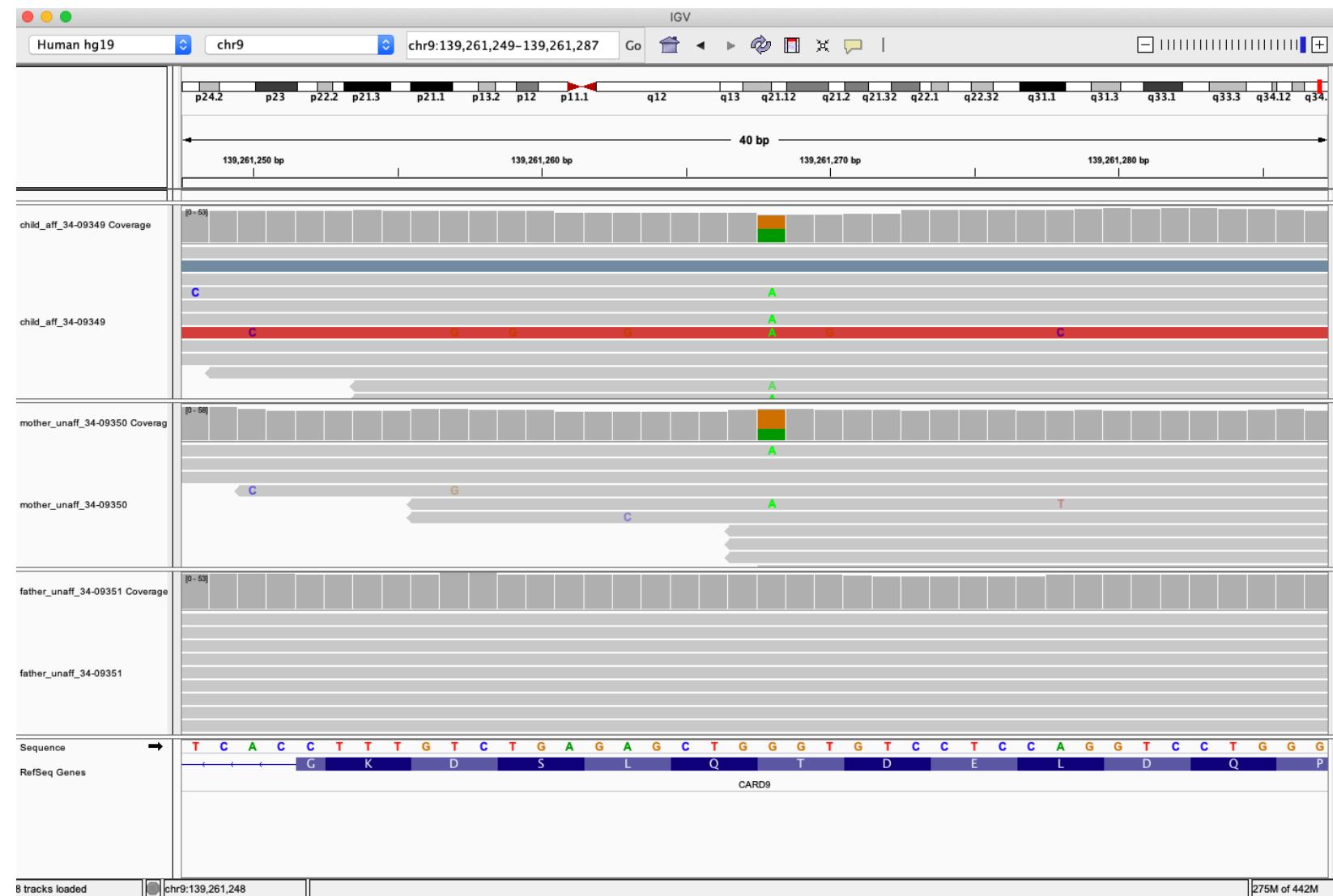


- Useful for loading aligned data available in NCBI (SRA, GEO)

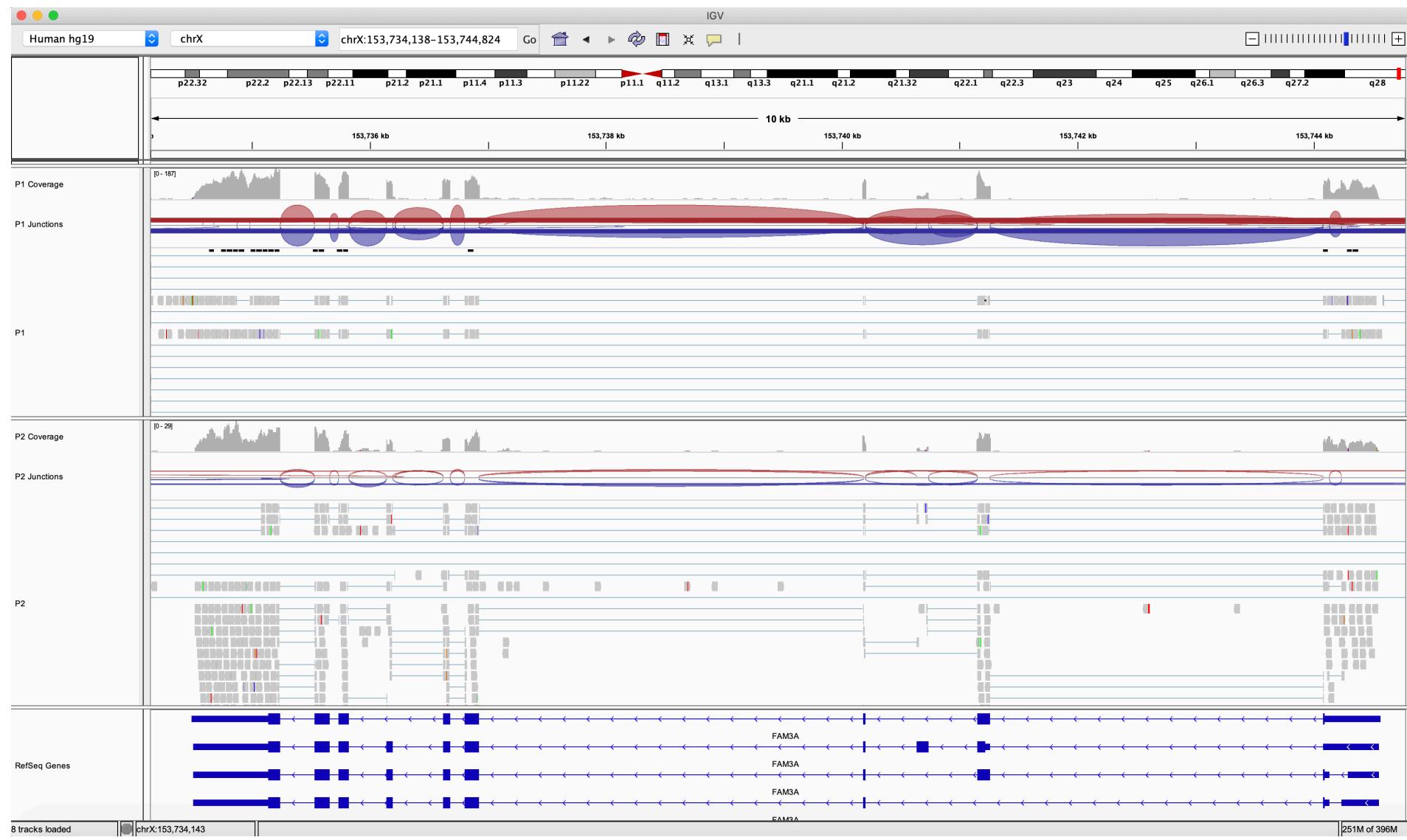
Genome Browsers: Integrated Genome Viewer (IGV)

- 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)

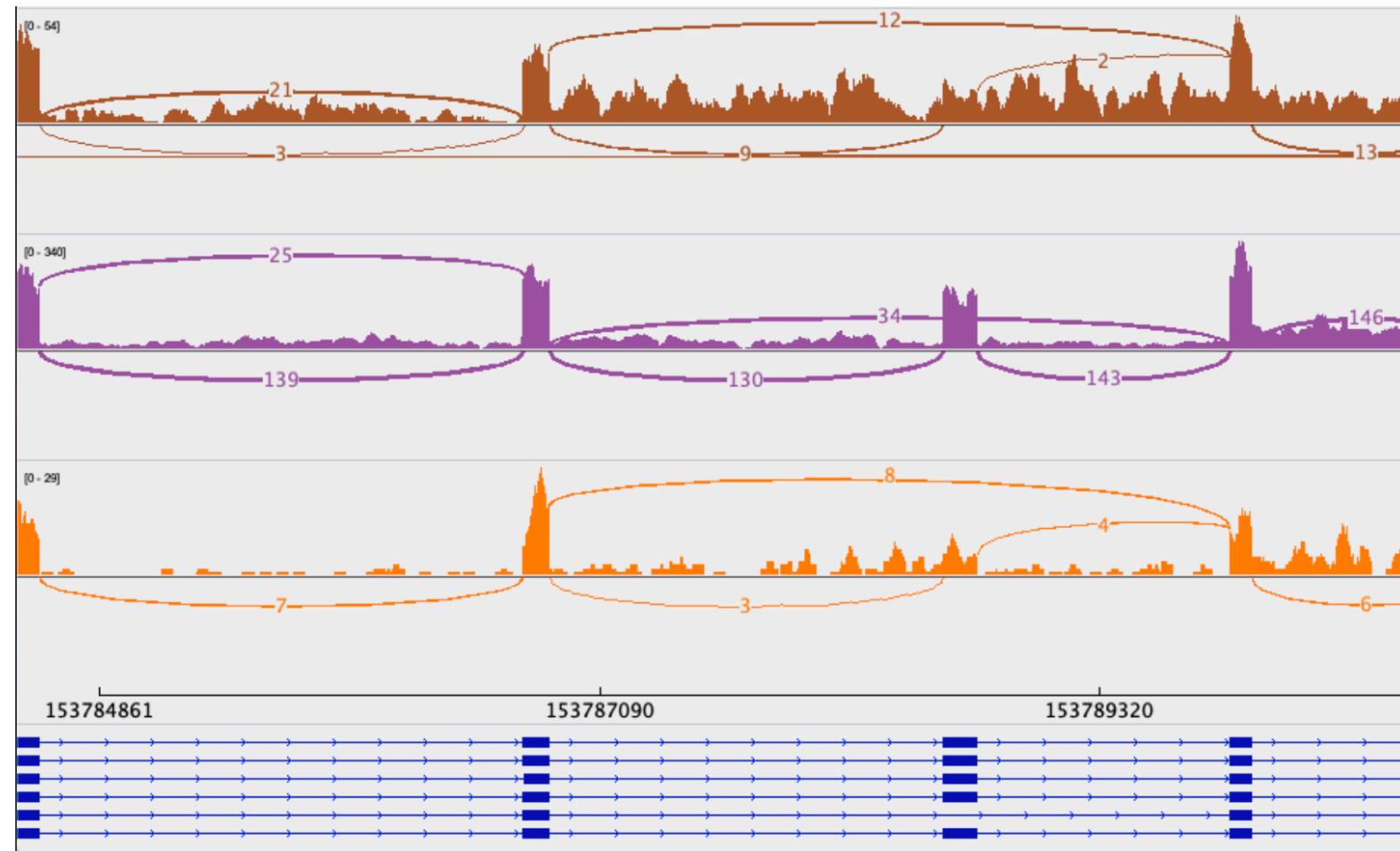
IGV – DNA-seq BAM files to visualize variants



IGV RNA-seq BAM files to visualize expression and splicing



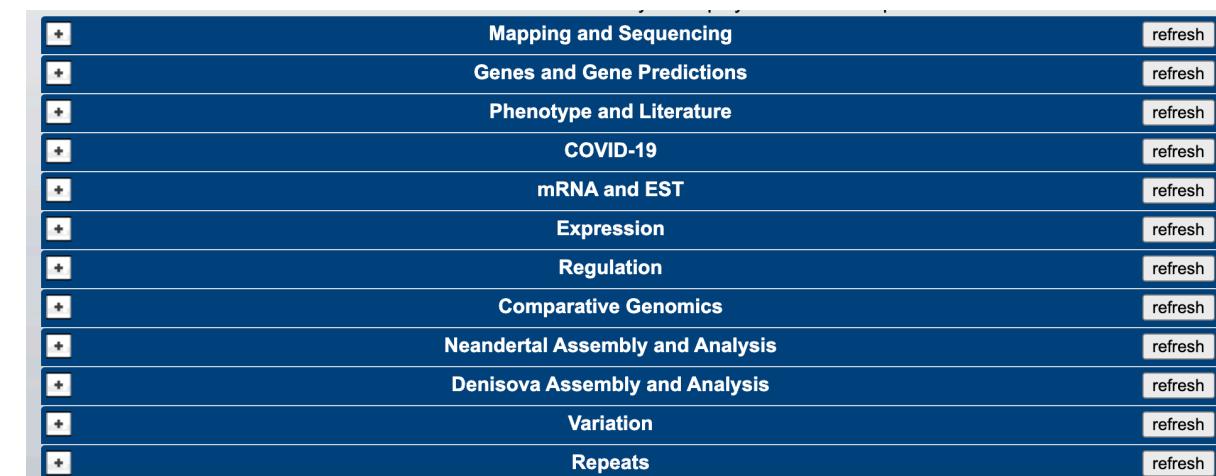
IGV – Sashimi Plots



<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

Mapping and Sequencing

Base Position	Fix Patches	Alt Haplotypes	Assembly	BAC End Pairs	BU ORCHID
Chromosome Band	deCODE	ENCODE Pilot	New Exome Probesets	FISH Clones	hide
Gap	Recomb	GC Percent	GRC Incident	Hg18 Diff	hide
Hg38 Mapping	Hi Seq Depth	INSDC	GRC Map Contigs	Hg38 Diff	hide
Problematic Regions	Hi Seq Depth	INSDC	LRG Regions	Map Contigs	Mapability
	Recomb Rate	RefSeq Acc	Resid Enzymes	Short Match	STS Markers

Genes and Gene Predictions

UCSC Genes	NCBI RefSeq	Other RefSeq	AceView Genes	AUGUSTUS	CCDS
CRISPR Targets	Ensembl Genes	EvoFold	Exoniphy	Updated GENCODE...	Geneid Genes
Genscan Genes	H-Inv 7.0	IKMC Genes	IncRNAs...	LRG Transcripts	MGC Genes
N-SCAN	Old UCSC Genes	ORFeome Clones	Pfam in UCSC Gene	Retroposed Genes	SGP Genes
SIB Genes	sno/miRNA	TransMap V5...	tRNA Genes	UCSC Alt Events	Updated UniProt
Vega Genes	Yale Pseudo60				

Phenotype and Literature

Publications	ClinGen	Deprecated ClinGen	CNVs	New ClinVar Variants	Coriell CNVs
Decipher CNVs	DECIPHER SNVs	Development Delay	GAD View	Gene Interactions	GeneReviews
GWAS Catalog	Updated HGMD	Variants	Lens Patents	LOVD Variants	MGI Mouse QTL
OMIM Cyto Loci	OMIM Genes	RGD Human QTL	RGD Rat QTL	OMIM Alleles	UniProt Variants
Variants in Papers...	Web Sequences				

COVID-19

New COVID GWAS v4	COVID GWAS v3	New Rare Harmful	Vars		
-------------------	---------------	------------------	------	--	--

mRNA and EST

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

Expression

GTEx Gene V8	GTEx Transcript	Affy Exon Array	Affy GNF1H	Affy RNA Loc	Affy U95
Affy U133	Affy U133Plus2	Allen Brain	Burge RNA-seq	CSHL Small RNA-seq	ENC Exon Array...
GTEx Gene	ENC ProtGeno...	ENC RNA-seq...	GIS RNA PET	GNF Atlas 2	GWIPS-viz
Illumina WG-6	PeptideAtlas	gPCR Primers	RIKEN CAGE Loc	Sestan Brain	Riboseq
					EPDnew Promoters

Regulation

ENCODE Regulation...	GeneHancer	GTEX Combined eQTL	GTEX Tissue eQTL	CD34 DnaseI	CpG Islands...
ENC Chromatin...	ENC DNA Methyl...	ENC DNase/FAIRE...	ENC Histone...	ENC RNA Binding...	ENC TF Binding...
FSU Repli-chip	Genome Segments	NKI Nuc Lamina...	RegAnno	Rao 2014 Hi-C	Staf Nucleosome
SUNY SwitchGear	SwitchGear TSS	TFBS Conserved	Updated TS miRNA Targets...	UCSF Brain Methyl	UMMS Brain Hist
UW Repli-seq	Vista Enhancers				

Comparative Genomics

Conservation	Cons 46-Way	Cons Indels MmCf	Evo CpG	GERP	phastBias.gBGC
Primate Chain/Net	full	Placental Chain/Net	Vertebrate Chain/Net		

Neandertal Assembly and Analysis

5% Lowest S	Cand. Gene Flow	H-C Coding Diffs	Neandertal Methyl	Neandertal Mito	Neandertal Seq
[No data-chr7]	[No data-chr7]	[Sel Swap Scan (S)]			

Denisova Assembly and Analysis

Denisova Methyl	Denisova Seq	Denisova Variants	Mod Hum Variants	Modern Derived	
dbSNP 153	Common SNPs(151)	Common SNPs(150)	Common SNPs(147)	Common SNPs(146)	Common SNPs(144)
	Common SNPs(142)	Common SNPs(141)	All SNPs(151)	All SNPs(150)	All SNPs(147)
	All SNPs(146)	All SNPs(144)	All SNPs(142)	All SNPs(141)	All SNPs(138)
Flagged SNPs(150)	Flagged SNPs(147)	Flagged SNPs(146)	Flagged SNPs(144)	Flagged SNPs(142)	Flagged SNPs(151)
Flagged SNPs(138)	Mult. SNPs(151)	Mult. SNPs(150)	Mult. SNPs(147)	Mult. SNPs(146)	Mult. SNPs(144)
Mult. SNPs(142)	Mult. SNPs(138)	1000G Ph1 Accsbl	1000G Ph1 Vars	1000G Ph3 Accsbl	1000G Ph3 Vars
dbVar Common Struct Var...	DGV Struct Var	EVS Variants	ExAC	New Genome in a Bottle	Genome Variants
GIS DNA PET	Updated gnomAD...	HAIB Genotype	HapMap SNPs	HGDP Allele Freq	Platinum Genomes
SNP/CNV Arrays					

Repeats

RepeatMasker	Interrupted Rpts	Microsatellite	NumtS Sequence	Segmental Dups	Self Chain
Simple Repeats	WM + SDust				

UCSC Genome Brower - Online Training and Tutorials

- <https://genome.ucsc.edu/training/>

- Video Tutorials

- User Guides

- Upcoming workshops

April 13, 2021

June 13, 2021

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's guides for tools in the Genome Brower

- [Genome Brower](#)
- [Custom Tracks](#)
- [Track Hubs](#)
- [Sessions](#)
- [Table Brower](#)
- [Data Integrator](#)

- [Gene and Pathways Interactions](#)
- [Variant Annotation Integrator](#)
- [Genome Graphs](#)
- [Gene Sorter](#)
- [Genome Brower in a Box \(GBiB\)](#)
- [Genome Brower in the Cloud \(GBiC\)](#)

[ACMG 2021](#) Los Angeles --> virtual

[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

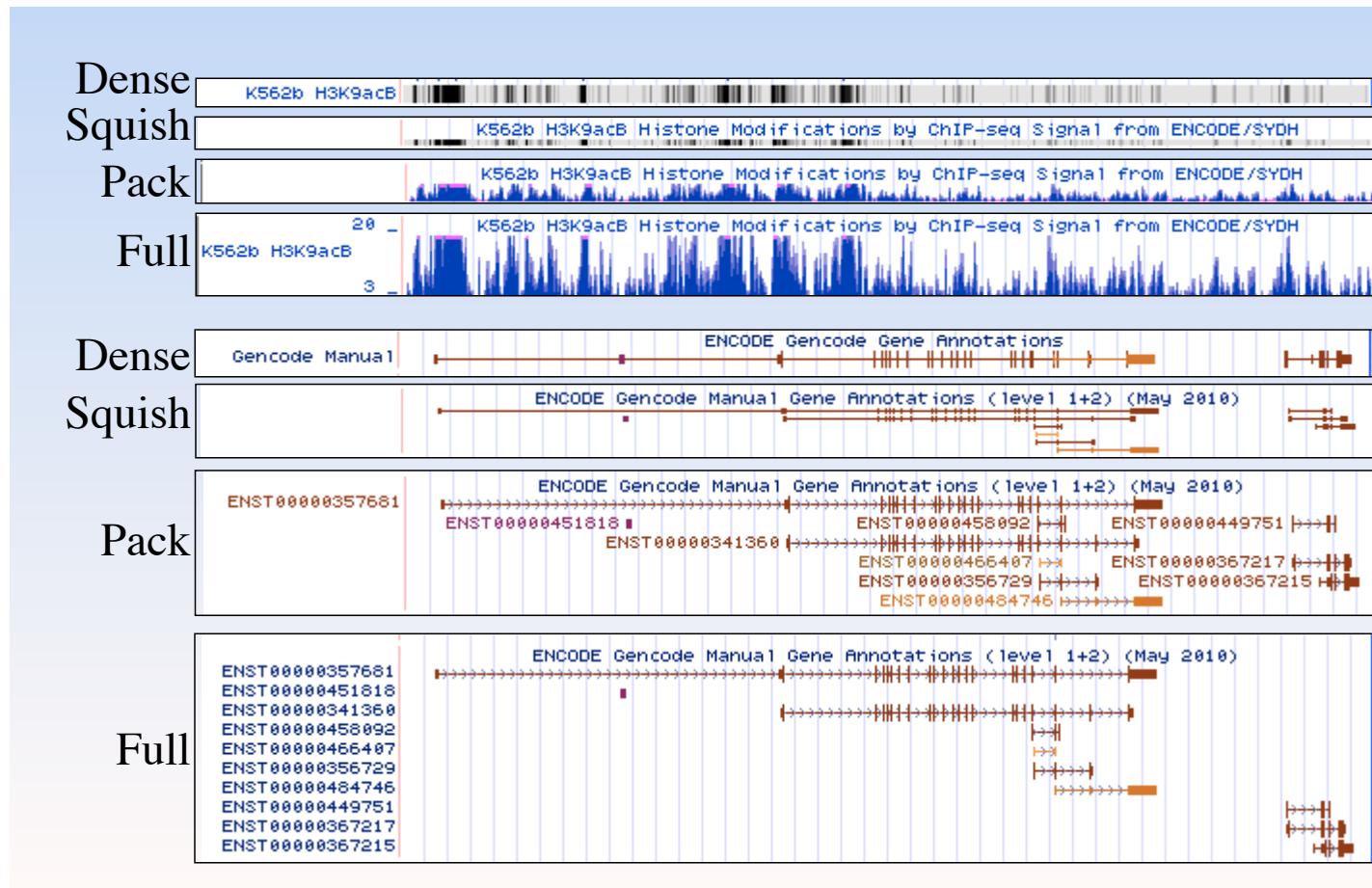
Annotation Tracks by Group

- 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