

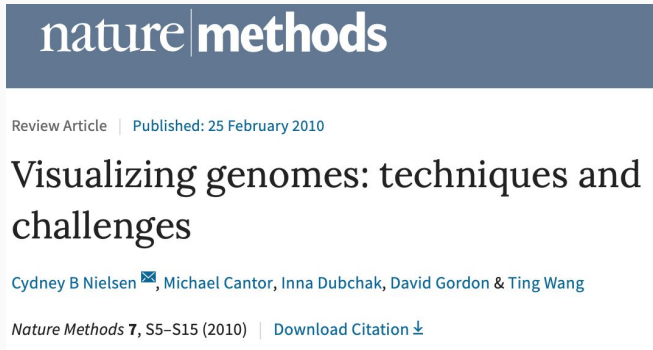
Sequence Visualization

Tutorials and References

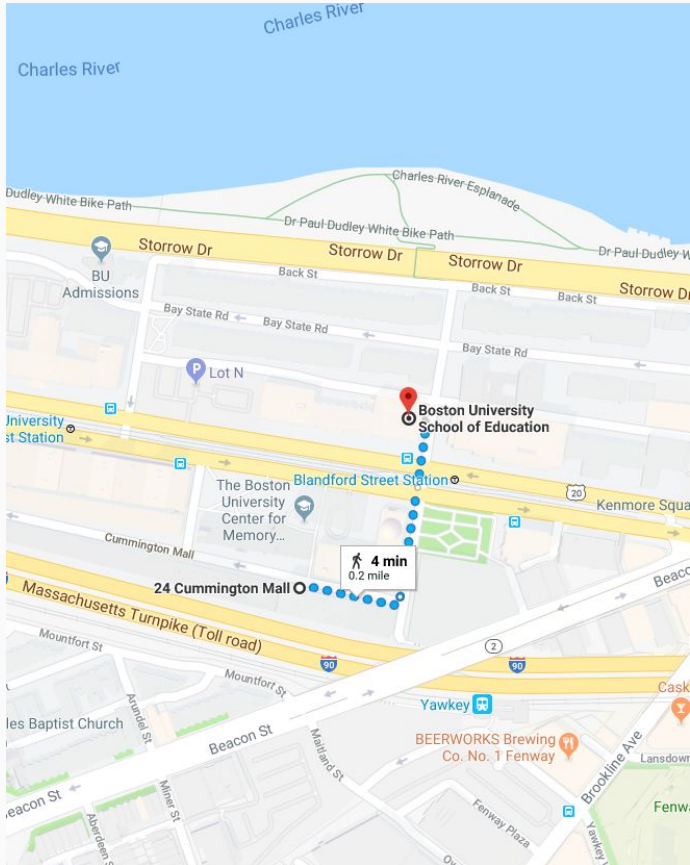
Tutorials:

- IGV □ Griffith Lab Tutorials (https://github.com/griffithlab/rnaseq_tutorial/)
- Broad Institute of MIT & Harvard (<http://software.broadinstitute.org/software/igv/>)

Additional Reading:



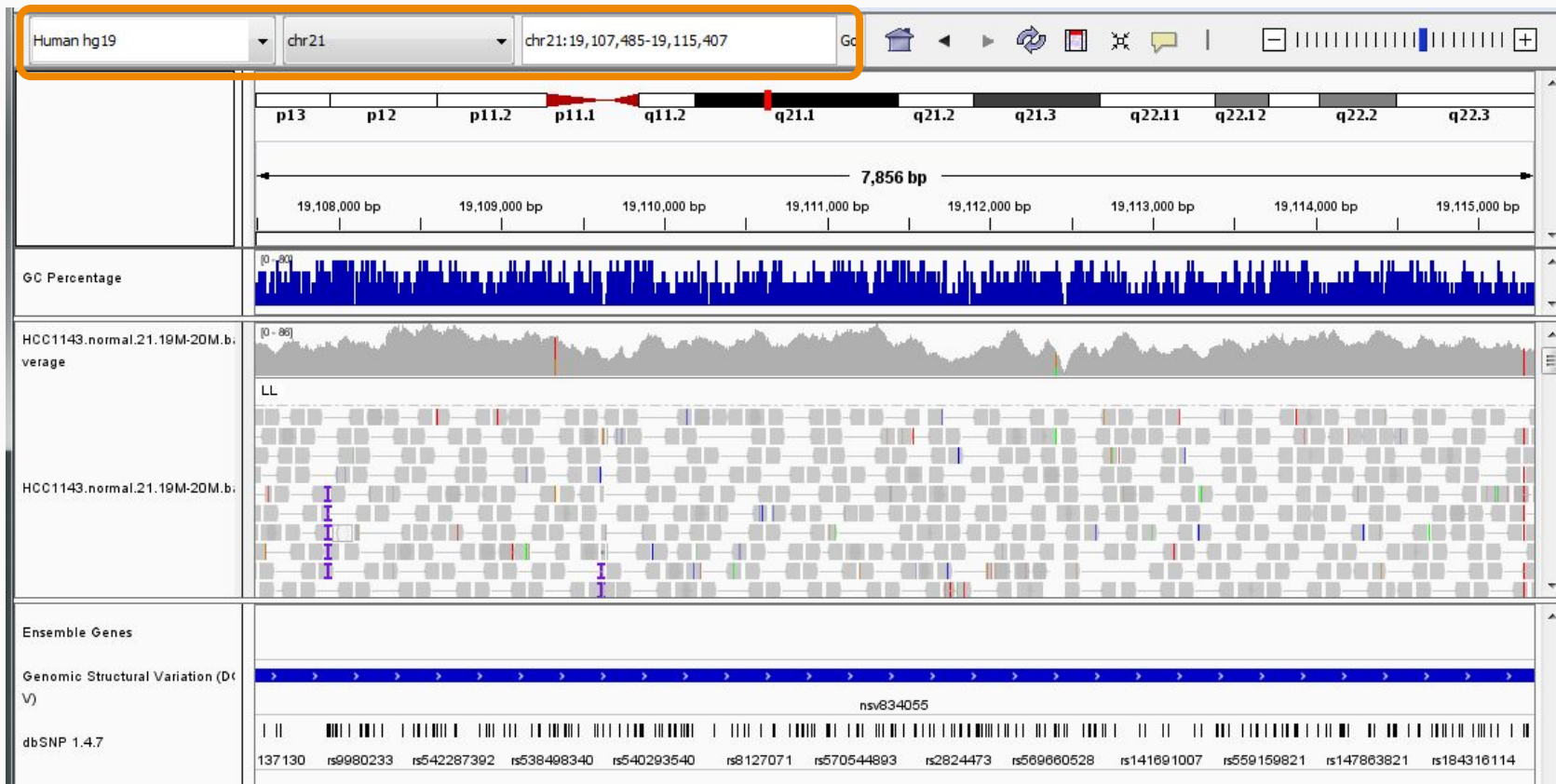
Google Maps Comparison



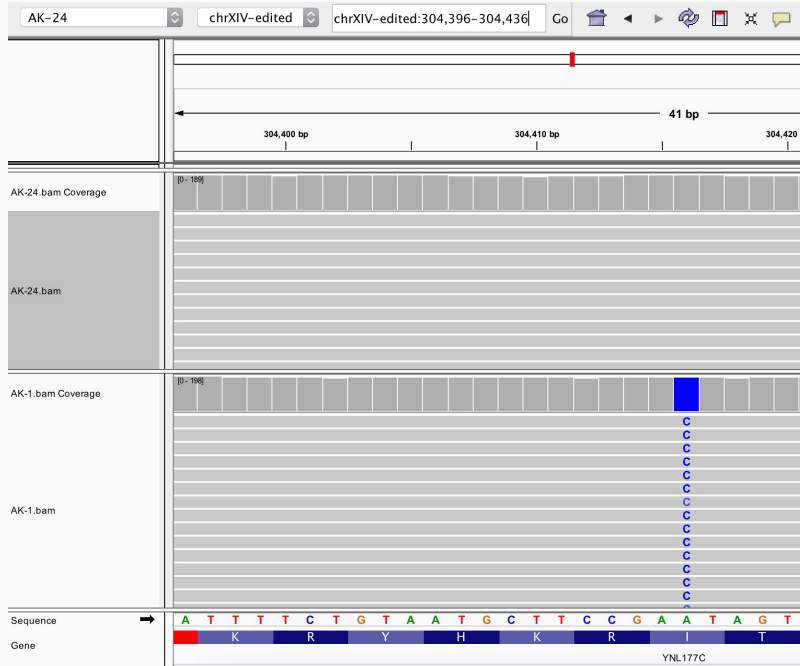
- Would google maps be effective if it just spat out minimally formatted sets of coordinates?
- The map to the left is a **human-centered visual summary** of how to get from LSEB to SED
- Additional layers beyond start, stop, and directions provide additional context
- Genome browsers (like IGV) provide a **human-centered visual summary** of one/many sequencing experiments

Integrative Genomics Viewer

Genomic "address"

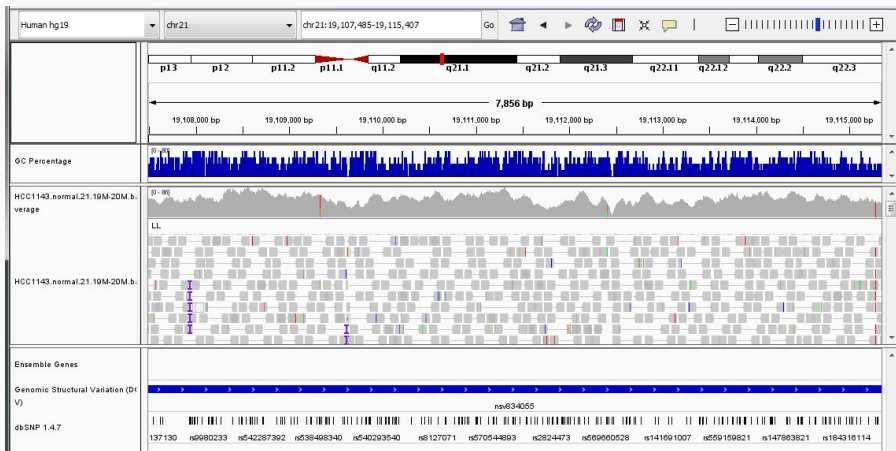


Why use a genome browser?

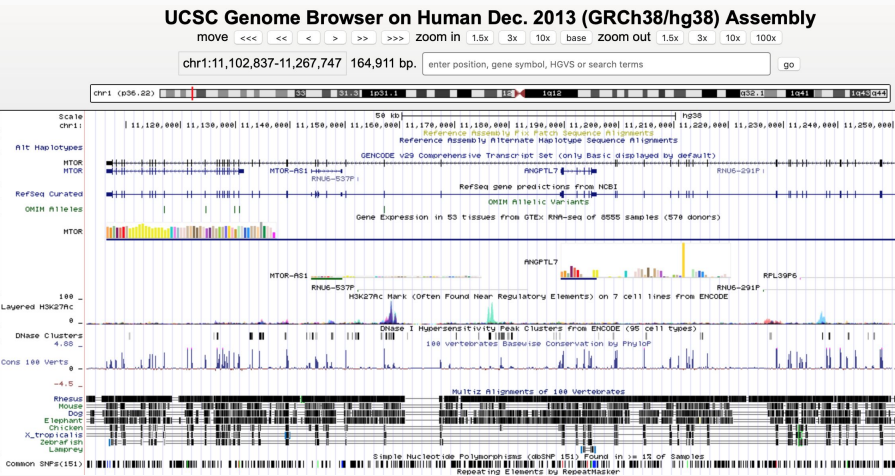


- **Visually confirm phenomena** from sequencing experiments (seeing is believing)
 - Left: Visualization of a SNP identified in a lab-evolved strain of yeast
- **Integration of multiple experiments** on the same coordinate system – collapsing several files
- **Communication of key findings** from sequencing experiments

Commonly Used Genome Visualization Tools



Integrative Genomics Viewer
<http://software.broadinstitute.org/software/igv/>



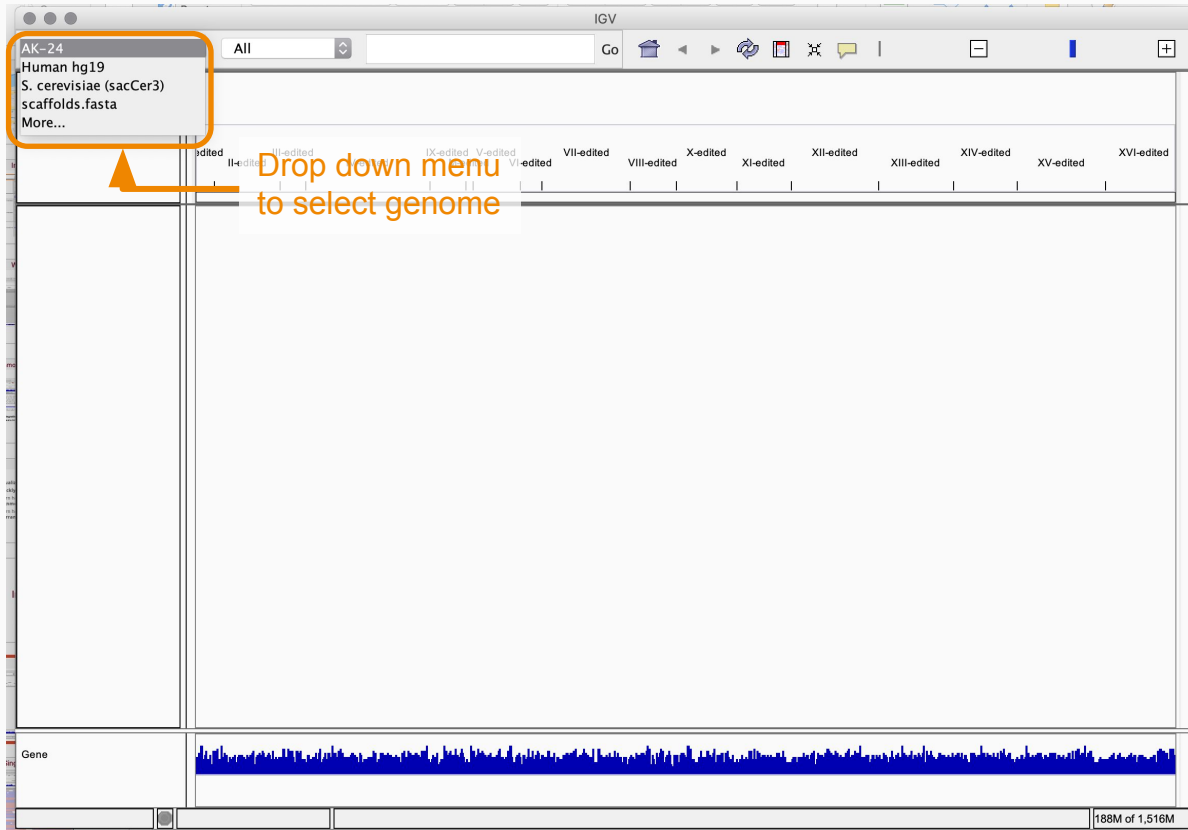
UCSC Genome Browser
<https://genome.ucsc.edu>

Goals for this Lecture

- **Visualize** a variety of **genomic data**
- **Quickly navigate** around the genome
- Learn how to be able to **visualize your own read alignments**
- Learn how to **recognize SNPs and structural rearrangements**

Integrative Genomics Viewer (IGV)

IGV: Introduction to Usage



1. Download software from: <http://software.broadinstitute.org/software/igv/download>
2. Open up the application
3. Choose genome (e.g. Hg38, Mm10, or a custom genome)

IGV: Introduction to Usage



1. Download software from: <http://software.broadinstitute.org/software/igv/download>
2. Open up the application
3. Choose genome (e.g. Hg38, Mm10, or a custom genome)
4. Load alignment file(s)
5. Visualize alignments:
 - Coverage plot shows distribution of alignment
 - Each elongated pentagon is a read
 - Colored lines = differences from reference
 - Reference sequence, amino acid sequences, and gene

SNPs

reference: AA-TACGG**A**CGGACTT**T**A

read1: AA**C**TACGG-CGGACTT**T**A

read2: AA**C**TACGG-CGGACTT**T**A

read4: AA**C**TACGG-CGGACTT**G**A

read5: AA**C**TACGG-CGGACTT**G**A

Insertion **D**ELetion **S**NP

```
samtools mpileup -u -v -r
chr22:29268316-29300343 -d 150 -f
../06/ref/chr22.fa
NA12878_phased_chr22.bam >
NA12878_chr22_samtools_EWSR1.vcf
```

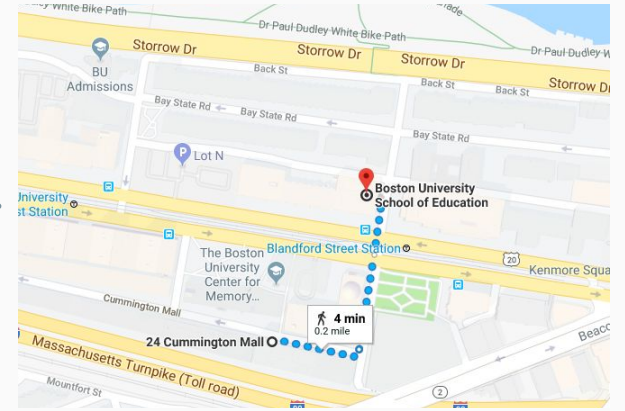
```
gatk HaplotypeCaller \
-L chr22:29268316-29300343 \
-R ../06/ref/chr22.fa \
-I NA12878_phased_chr22.bam \
-O NA12878_chr22_gatk_EWSR1.vcf.gz
\
-ERC GVCF # BP_RESOLUTION
```

IGV: Visualize SNPs Identified From Variant Calling

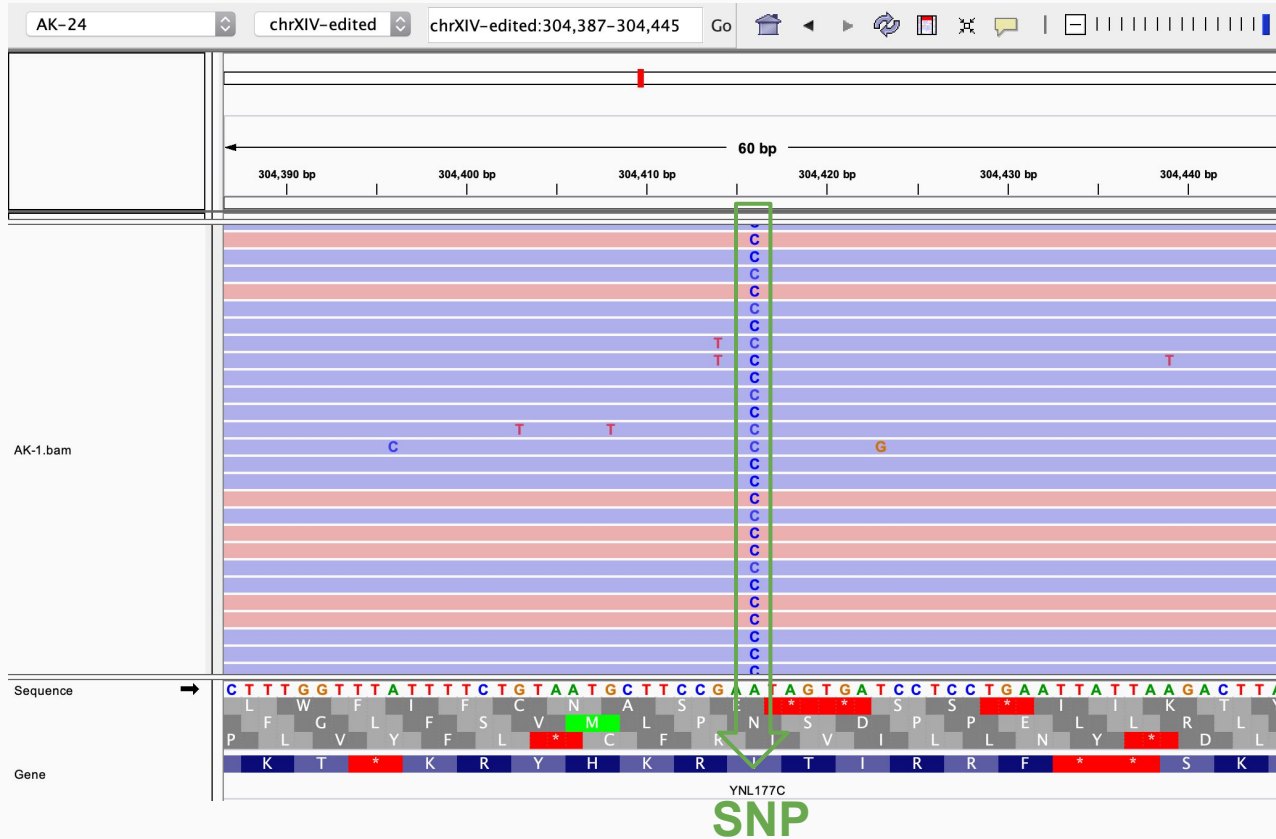
How do we go from a set of **labelled coordinates** to a **human-centered visual summary**?

(e.g. a VCF file)

```
NODE_961_length_155_cov_2566.05_ID_185286 111 G T 89.32 AC=4;AF=0.500;AN=8;BaseQRankSum=-2.590;ClippingRankSum=0.000;DP=1643;ExcessHet=9.4201;FS=46.557;MLEAC=4;MLEAF=0.500;MQ=43.90;MQRankSum=1.564;QD=2.98;ReadPosRankSum=-1.720;SOR=5.131 GT:AD:DP:GQ:PL
0/1:3,3:6:21:21,0,79 0/1:3,2:5:5:5,0,107 0/1:5,5:10:91:91,0,142 0/1:7,2:9:2:2,0,239
NODE_961_length_155_cov_2566.05_ID_185286 123 C A 13.61 AC=1;AF=0.125;AN=8;BaseQRankSum=-0.138;ClippingRankSum=0.000;DP=1375;ExcessHet=3.0103;FS=26.107;MLEAC=1;MLEAF=0.125;MQ=43.46;MQRankSum=1.084;QD=1.70;ReadPosRankSum=-0.220;SOR=3.585 GT:AD:DP:GQ:PL
0/1:4,4:8:43:43,0,130 0/0:6,0:6:18:0,18,16
7 0/0:2,0:2:6:0,6,86 0/0:3,1:4:2:0,2,94
NODE_962_length_155_cov_197.282_ID_185102 75 A T 97.92 AC=3;AF=0.375;AN=8;BaseQRankSum=-0.773;ClippingRankSum=0.000;DP=164;ExcessHet=5.4407;FS=1.636;MLEAC=3;MLEAF=0.375;MQ=45.75;MQRankSum=-1.367;QD=2.13;ReadPosRankSum=2.644;SOR=0.399 GT:AD:DP:GQ:PL
0/0:9,0:9:27:0,27,427 0/1:8,2:10:60:60,0,3
43 0/1:17,2:19:33:33,0,728 0/1:15,2:17:39:39,0,647
NODE_962_length_155_cov_197.282_ID_185102 78 T A 97.93 AC=3;AF=0.375;AN=8;BaseQRankSum=-0.474;ClippingRankSum=0.000;DP=164;ExcessHet=5.4407;FS=1.653;MLEAC=3;MLEAF=0.375;MQ=45.75;MQRankSum=-1.605;QD=2.13;ReadPosRankSum=2.882;SOR=0.389 GT:AD:DP:GQ:PL
0/0:9,0:9:27:0,27,427 0/1:8,2:10:60:60,0,2
94 0/1:17,2:19:33:33,0,728 0/1:15,2:17:39:39,0,647
```

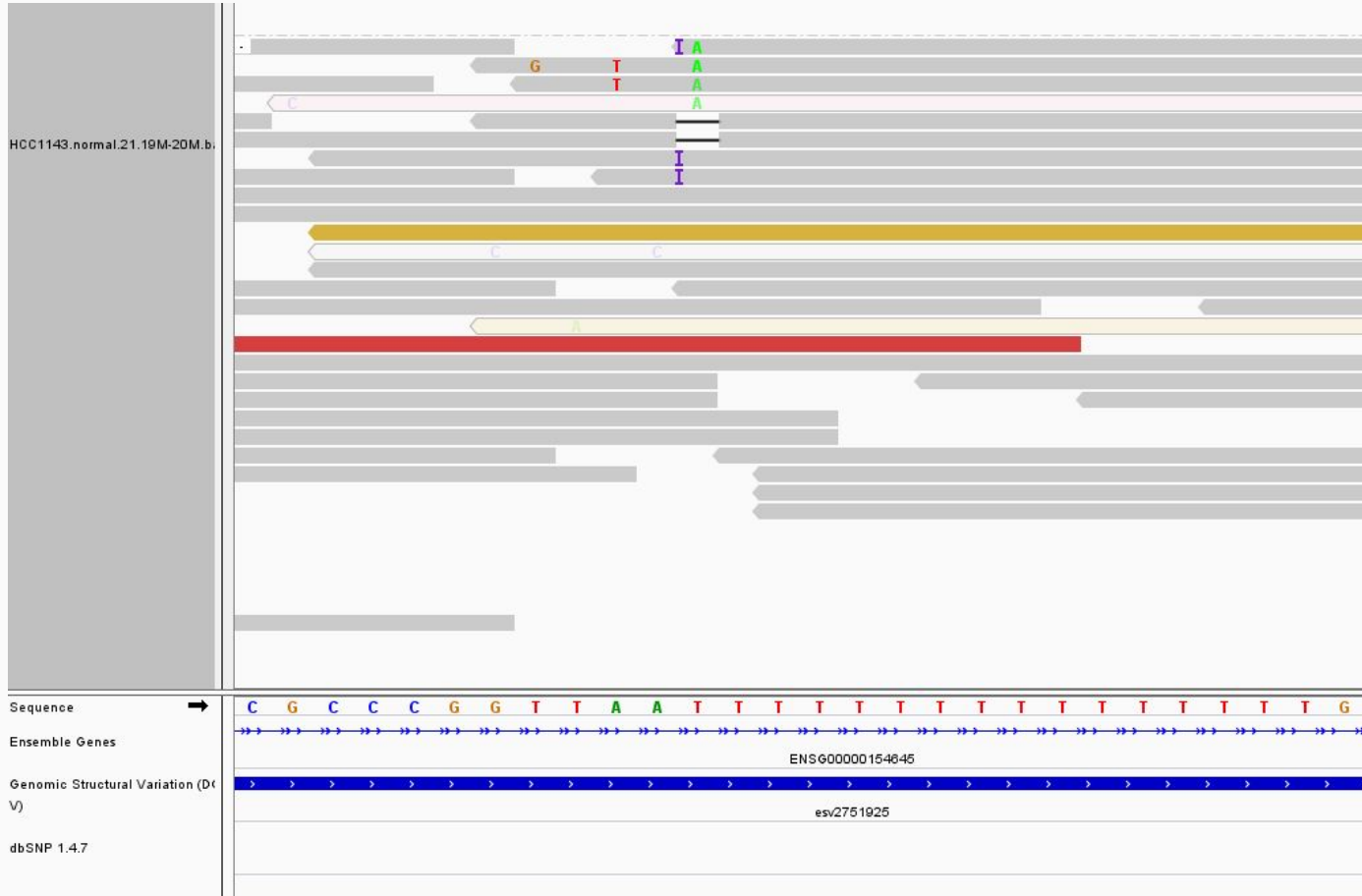


IGV: Visualize SNPs Identified From Variant Calling



1. Load tracks (.BAM files, .VCF files, etc.). **Here:** Alignment file for 1 sample
2. Zoom into locus of interest. **Here:** chrXIV of our custom genome
3. Set visualization parameters (colors, shading, etc.). **Here:** paired-end reads colored by forward (red) or reverse (blue) read
4. Use annotation (.GTF file) to identify which gene SNP is in

IGV: A Homopolymer Run



- A long stretch consisting of a single base
- You want to be looking at the sequence here (all those Ts)
- Difficult to map against, particularly at ends of reads

IGV: Coverage by GC percent

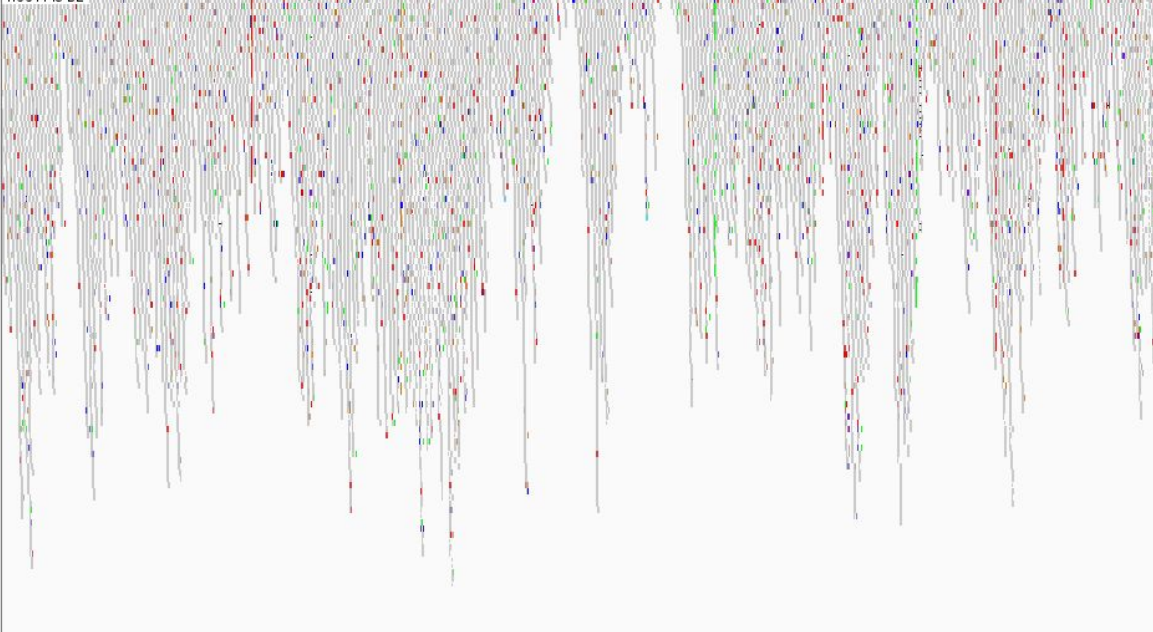
GC Percentage



HCC1143.normal.21.19M-20M.b.
verage



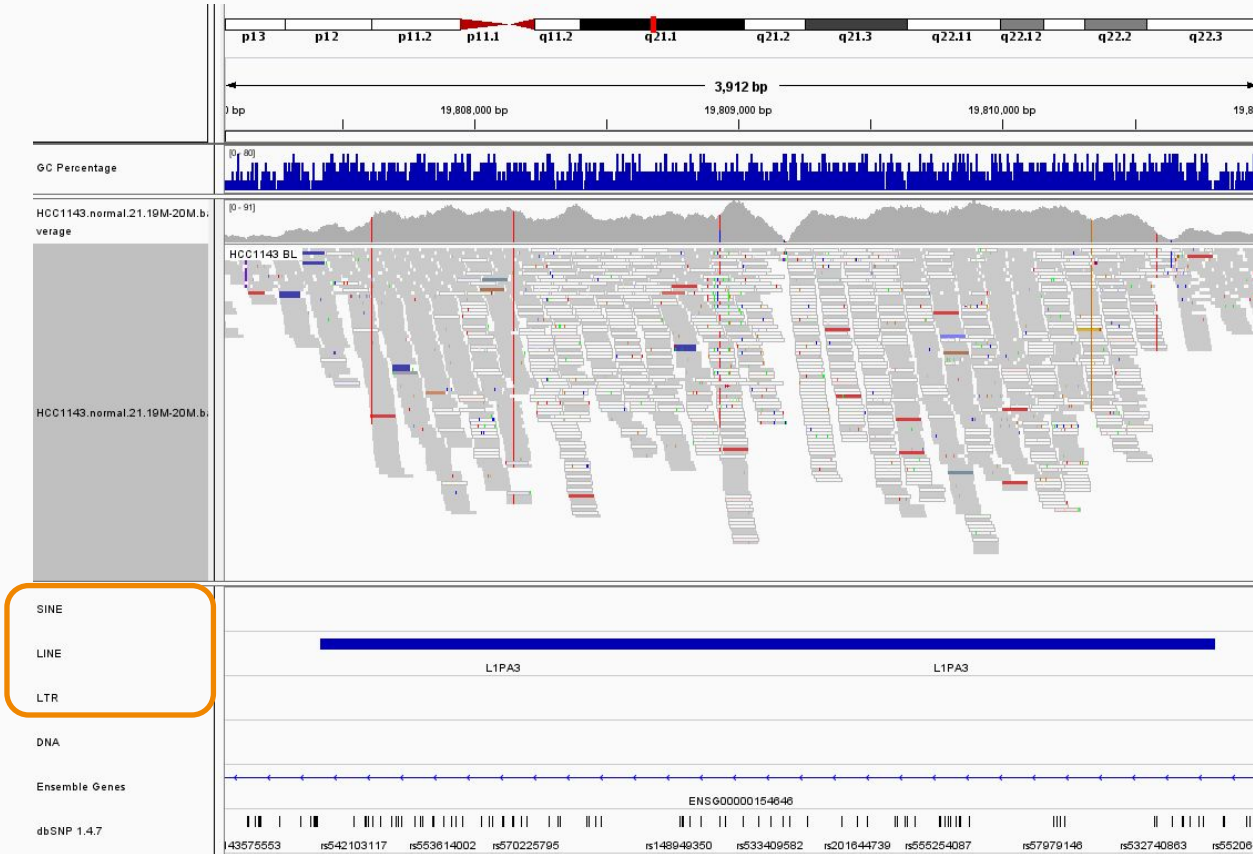
HCC1143 BL



HCC1143.normal.21.19M-20M.b.

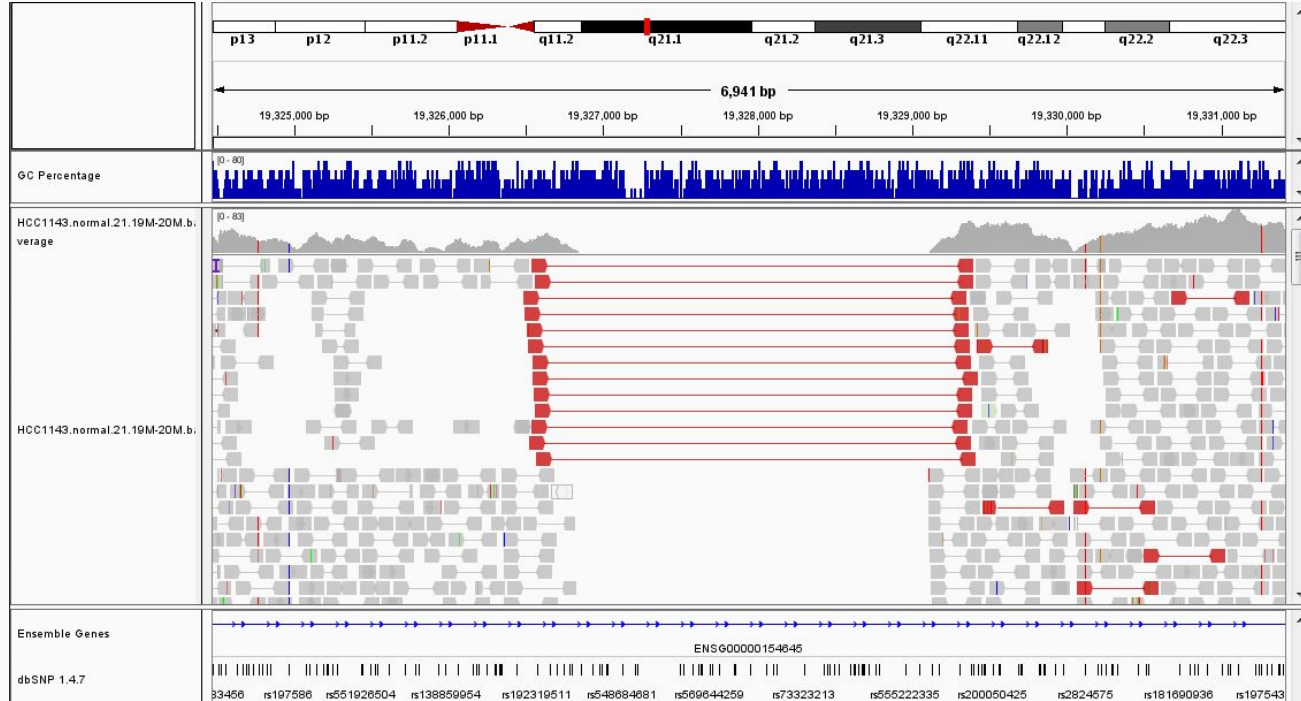
- Benjamini & Speed (2012) proposed that **PCR step** generates this **GC bias**
- Severity differs from experiment to experiment

IGV: Low Mapping Quality



- Repetitive elements (tandem repeats, LINEs, SINEs, etc.) can have multiple nearly identical copies in the genome
- Reads will map to multiple versions in the genome
- Referred to as “low mapping quality” (reads visualized as white, not grey)

IGV: Homozygous Deletion



- All mate pairs that map here span the deletion
- Visually, the reference contains an “insert” of ~3kb
- Look at the sizes of other fragments

Automating Tasks

- IGV has its own set of common commands that it recognizes
- You can load a bunch of tracks for example using successive “load” commands in a script file
- The commands can be harnessed to do cool things (like sweep through a bed file and create snapshots of all the regions):



David Jenkins
dfjenkins3

bedToIgv

After creating the BED file, the `bedtools igv` (`bedToIgv`) tool can be used to convert the bed file into an IGV batch script. At a minimum, you should supply a BED file to `bedToIgv` , but you may want to specify an output directory (`-path`), how to sort the screenshot (`-sort`), or how much padding to add to the records (`-slop`).

Usage:

```
Tool:    bedtools igv (aka bedToIgv)
Version: v2.21.0
Summary: Creates a batch script to create IGV images
         at each interval defined in a BED/GFF/VCF file.

Usage:   bedtools igv [OPTIONS] -i <bed/gff/vcf>
```

UCSC genome browser

Selecting which species to browse

The screenshot displays the UCSC Genome Browser Gateway interface. At the top, the University of California Santa Cruz logo and the UCSC Genome Browser Gateway title are visible. A navigation bar includes links for Home, Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us.

Browse/Select Species

POPULAR SPECIES

Human, Mouse, Rat, Fruitfly, Worm, Yeast

Enter species or common name

REPRESENTED SPECIES

D. mojavensis
D. virilis
D. grimshawi
A. gambiae
A. mellifera
C. elegans
C. brenneri
C. briggsae
C. japonica
C. remanei

Find Position

Human Assembly
Dec. 2013 (GRCh38/hg38)

Position/Search Term
Enter position, gene symbol or search terms
Current position: chr1:11,461-11,718

Human Genome Browser - hg38 assembly [view sequences](#)

UCSC Genome Browser assembly ID: hg38
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38 (GCA_000001405.15)
Assembly date: Dec. 2013
Accession ID: GCA_000001405.15
NCBI Genome ID: 51 (Homo sapiens (human))
NCBI Assembly ID: 883148 (GRCh38, GCA_000001405.15)
BioProject ID: 31257

Homo sapiens
(Graphic courtesy of CBSE)

- A wide variety of species/references are available

UCSC genome browser interface

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

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr1:11,461-11,718 258 bp.

chr1 (p36.33) | 35 | 31.3 | 1p31.1 | 1q12 | 32.1 | 1q41 | 44

Scale chr1: 11,500 | 100 bases | 11,600 | hg38 | 11,700

RefSeq Curated: GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default), RefSeq gene predictions from NCBI

OMIM Alleles: OMIM Allelic Variants

Layered H3K27Ac: Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors), H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE

DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)

Cons 100 Verts: 100 vertebrates Basewise Conservation by PhyloP

Multiz Alignments of 100 Vertebrates: Rhesus, Mouse, Dog, Elephant, Chicken, X_tropicalis, Zebrafish, Lamprey

Common SNPs (150): Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples

Repeating Elements by RepeatMasker: SINE, LINE, LTR, DNA, Simple, Low Complexity, Satellite, RNA, Other, Unknown

move start > 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. Press "?" for keyboard shortcuts. < > move end

Where UCSC beats IGV

track search default tracks default order hide all add custom tracks track hubs configure multi-region reverse resize refresh

collapse all Use drop-down controls below and press refresh to alter tracks displayed. expand all
Tracks with lots of items will automatically be displayed in more compact modes.

<input type="checkbox"/>	Mapping and Sequencing						refresh
<input type="checkbox"/>	Genes and Gene Predictions						refresh
<input type="checkbox"/>	Phenotype and Literature						refresh
<input type="checkbox"/>	mRNA and EST						refresh
<input type="checkbox"/>	Expression						refresh
<input type="checkbox"/>	Regulation						refresh
<input type="checkbox"/>	Comparative Genomics						refresh
<input type="checkbox"/>	Variation						refresh
	Common SNPs(150)	Common SNPs(147)	Common SNPs(146)	Common SNPs(144)	Common SNPs(142)	Common SNPs(141)	
	dense ▾	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾	
	All SNPs(150)	All SNPs(147)	All SNPs(146)	All SNPs(144)	All SNPs(142)	All SNPs(141)	
	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾	
	Flagged SNPs(150)	Flagged SNPs(147)	Flagged SNPs(146)	Flagged SNPs(144)	Flagged SNPs(142)	Flagged SNPs(141)	
	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾	
	Mult. SNPs(150)	Mult. SNPs(147)	Mult. SNPs(146)	Mult. SNPs(144)	Mult. SNPs(142)	Mult. SNPs(141)	
	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾	
	DGV Struct Var						
	hide ▾						
<input type="checkbox"/>	Repeats						refresh

Options for viewing your own data

clade genome assembly

Display your own data as custom annotation tracks in the browser. Data must be formatted in [bigBed](#), [bigChain](#), [bigGenePred](#), [bigMaf](#), [bigPsl](#), [bigWig](#), [barChart](#), [bigBarChart](#), [BAM](#), [VCF](#), [BED](#), [BED detail](#), [bedGraph](#), [broadPeak](#), [CRAM](#), [GFF](#), [GTF](#), [MAF](#), [narrowPeak](#), [Personal Genome SNP](#), [PSL](#), or [WIG](#) formats. To configure the display, set [track](#) and [browser](#) line attributes as described in the [User's Guide](#). Data in the bigBed, bigWig, bigGenePred, BAM and VCF formats can be provided via only a URL or embedded in a track line in the box below. Examples are [here](#). If you do not have web-accessible data storage available, please see the [Hosting](#) section of the Track Hub Help documentation.

Please note a much more efficient way to load data is to use [Track Hubs](#), which are loaded from the [Track Hubs Portal](#) found in the menu under My Data.

Paste URLs or data: Or upload: No file chosen

Online:

- Individual tracks can be loaded using the “add custom tracks” option (not recommended)
- Paste link to a track or track hub hosted elsewhere

Options for viewing your own data

```
genomes.txt trackDb.txt hub.txt
1 hub THP1_TF_hub
2 shortLabel THP-1 TF hub
3 longLabel Transcription factor binding sites in THP-1 cells
4 genomesFile genomes.txt
5 email bray@bu.edu
6
```

```
genomes.txt trackDb.txt hub.txt
1 genome hg19
2 trackDb hg19/trackDb.txt
3
```

```
genomes.txt trackDb.txt hub.txt
1 track PU1_BR1_bw
2 bigDataUrl NM-L1_S1_L001_R1_001_filt_sorted.bw
3 shortLabel PU.1 (repl. 1)
4 longLabel PU.1 read coverage for biological replicate 1
5 type bigWig
6 color 187,208,229
7 priority 1
8
```

Local:

- Version of the UCSC genome browser can be downloaded (VirtualBox + GBiB)
- Supports viewing custom tracks, local track hub configurations
- Left: Text files that configure a local track hub

Other Fun Things from UCSC

Dog genome

Sep. 2011 (Broad CanFam3.1/[canFam3](#))

- [Full data set](#)
- [Annotation database](#)
- [LiftOver files](#)
- [Pairwise alignments](#) ▶

BLAT Search Genome

Genome: Search ALL Assembly: Query type: Sort output: Output type:

Mouse July 2007 (NCBI/37/mm9) BLAT's guess query.score hyperlink

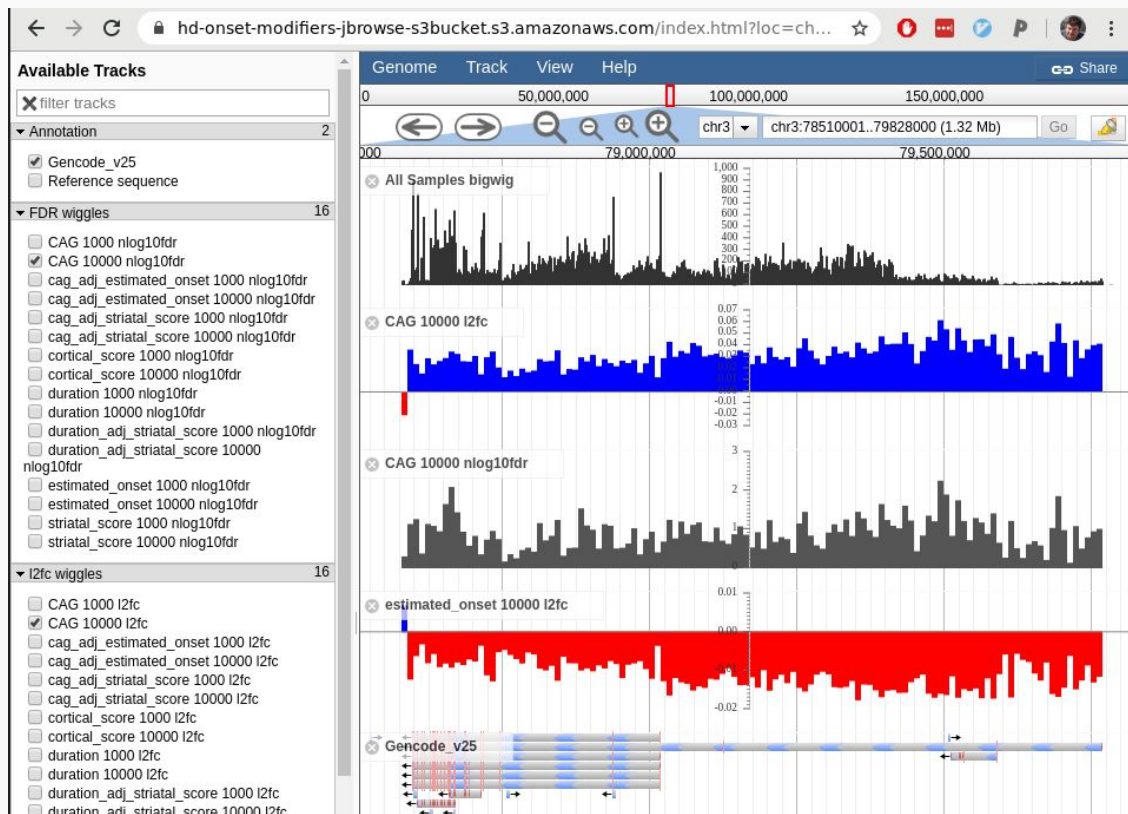
Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

IGV vs. UCSC

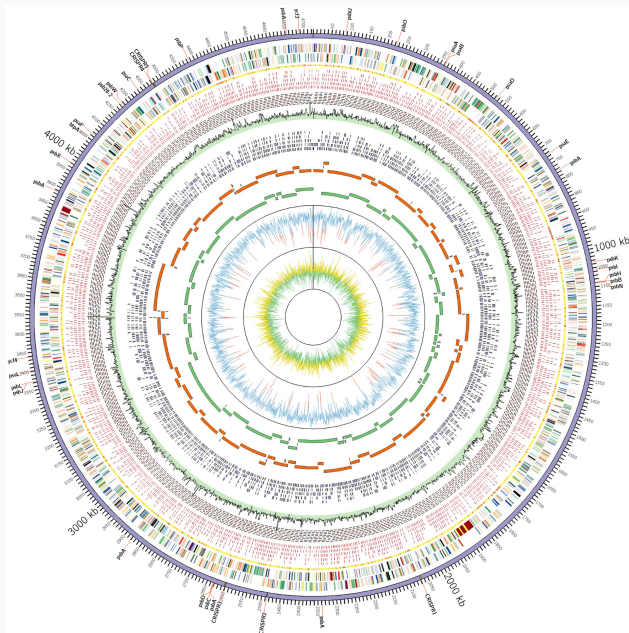
IGV	UCSC
Java application	Run in web browser
Run on your local computer	Hosted on the web
Data stored locally	Data stored on the web
Handles all major file types	Handles all major file types
Must load most tracks manually	Vast number of preloaded track types

Honorable Mention: JBrowse

- Javascript-based genome browser software
- Open source
- Highly extensible, customizable
- Must be hosted on a web server

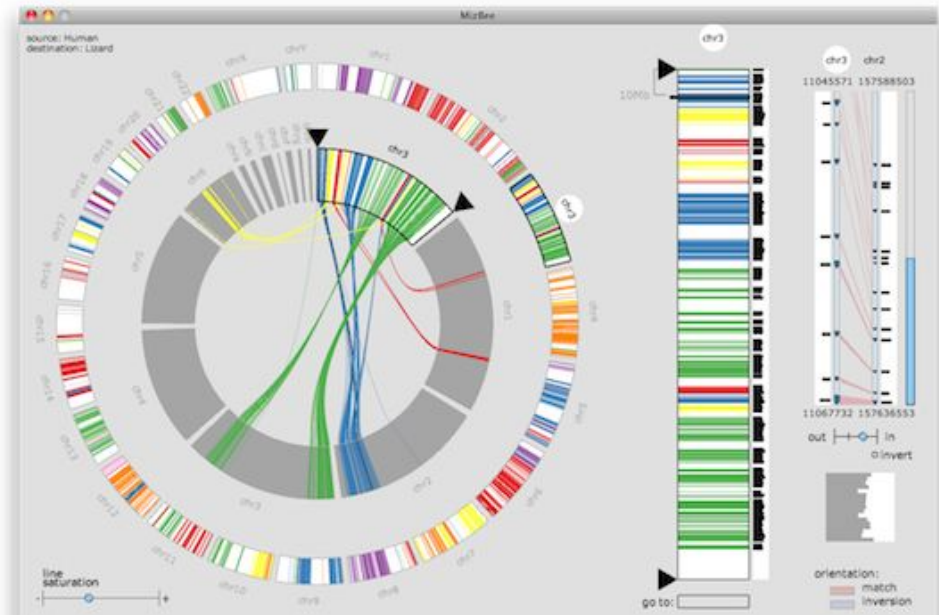


Other Genome Visualization Tools



Circos

<http://circos.ca/software/>



MizBee

(A Multiscale Synteny Browser)

<http://www.cs.utah.edu/~miriah/mizbee/Overview.html>