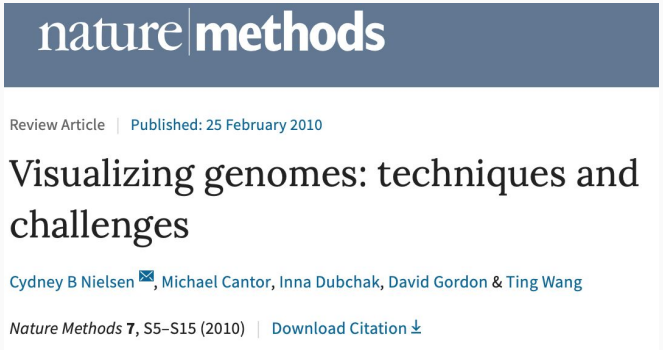
Sequence Visualization

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:

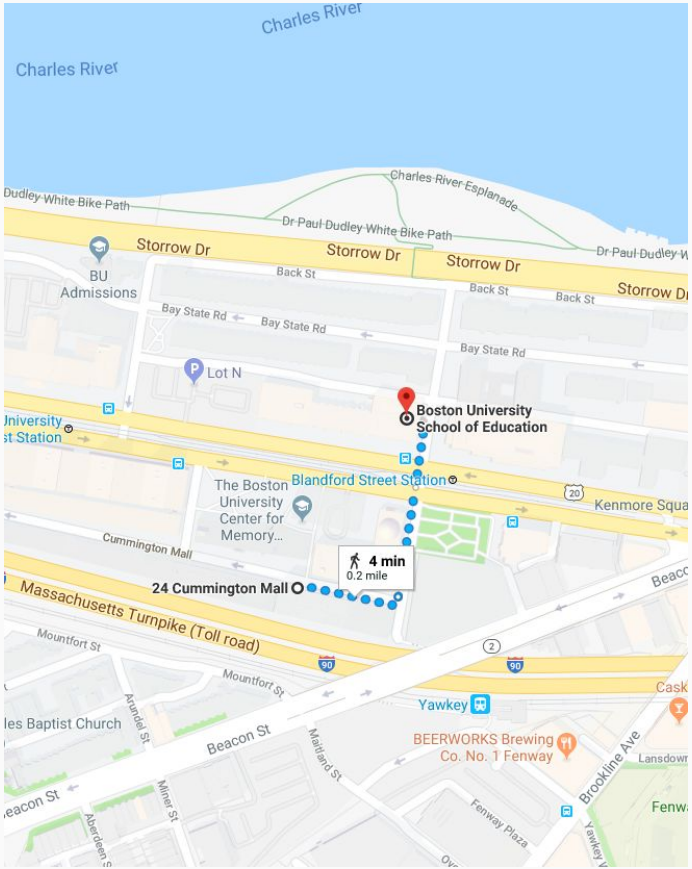


Sequence Visualization - Motivation

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

- High-throughput genomics – daunting at first
- Files contains millions of reads – go through each one?
- Our favorite file formats are easy for machines to read, not us

Google Maps Comparison

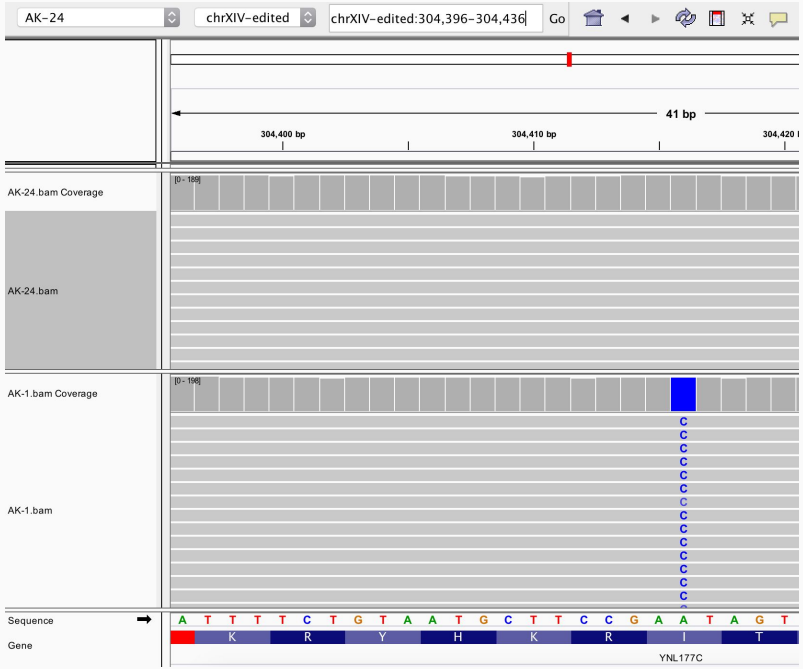


- Would google maps be effective if it just spat out minimally formatted sets of coordinates?
- The map to the left is a **human-centered visual summary** of how to get from LSEB to SED
- Additional layers beyond start, stop, and directions provide additional context
- Genome browsers (like IGV) provide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

Integrative Genomics Viewer

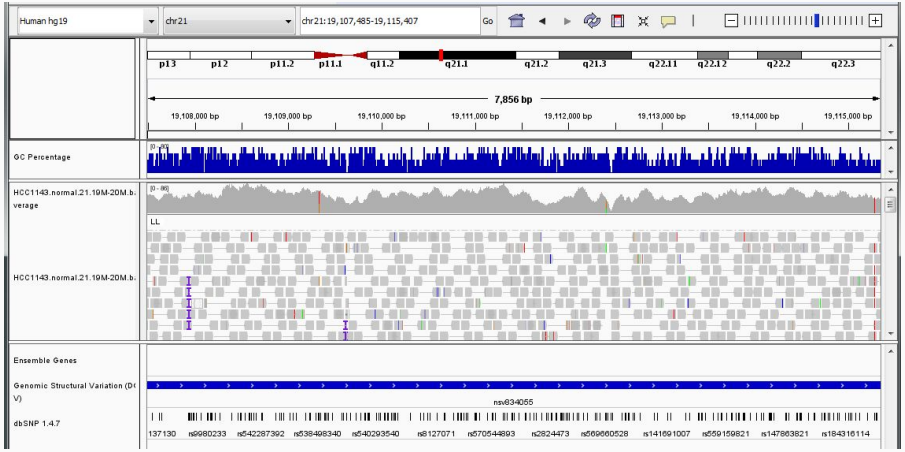
Genomic ”address”

Why use a genome browser?

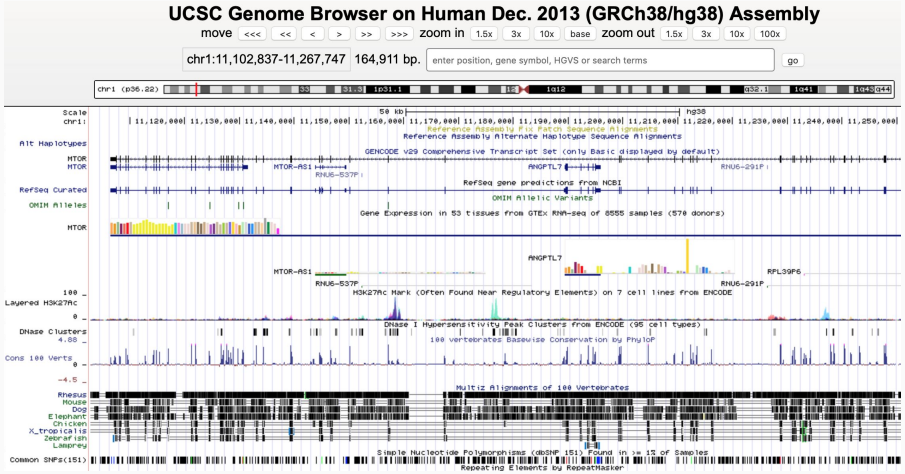


- **Visually confirm phenomena** from sequencing experiments (seeing is believing)
    - Left: Visualization of a SNP identified in a lab-evolved strain of yeast
- **Integration of multiple experiments** on the same coordinate system – collapsing several files
- **Communication of key findings** from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

Commonly Used Genome Visualization Tools



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



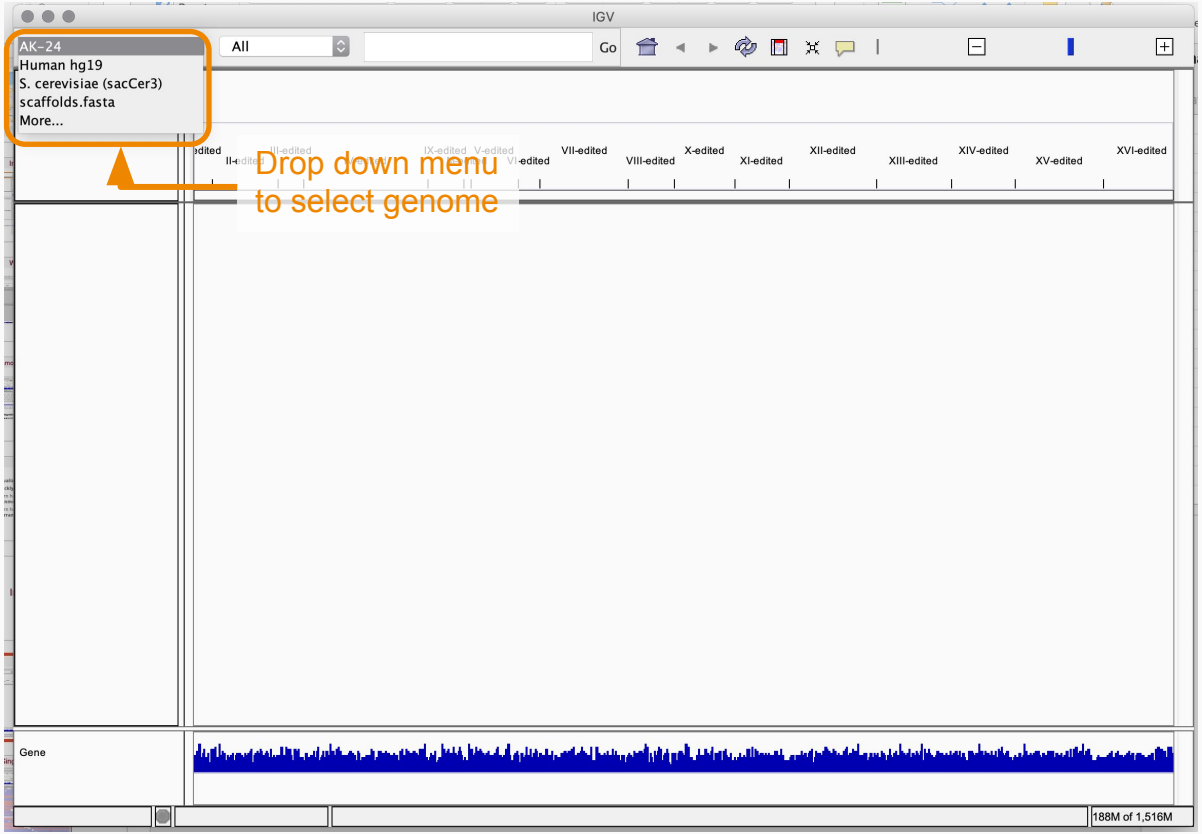
UCSC Genome Browser
https://genome.ucsc.edu

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


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

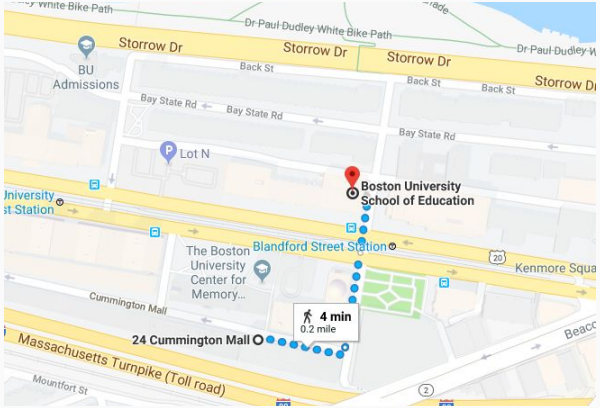
```
gatk HaplotypeCaller \
     -L chr22:29268316-29300343 \
     -R ../06/ref/chr22.fa \
     -I NA12878_phased_chr22.bam \
     -O NA12878_chr22_gatk_EWSR1.vcf.gz
\
     -ERC GVCF # BP_RESOLUTION
```

IGV: Visualize SNPs Identified From Variant Calling

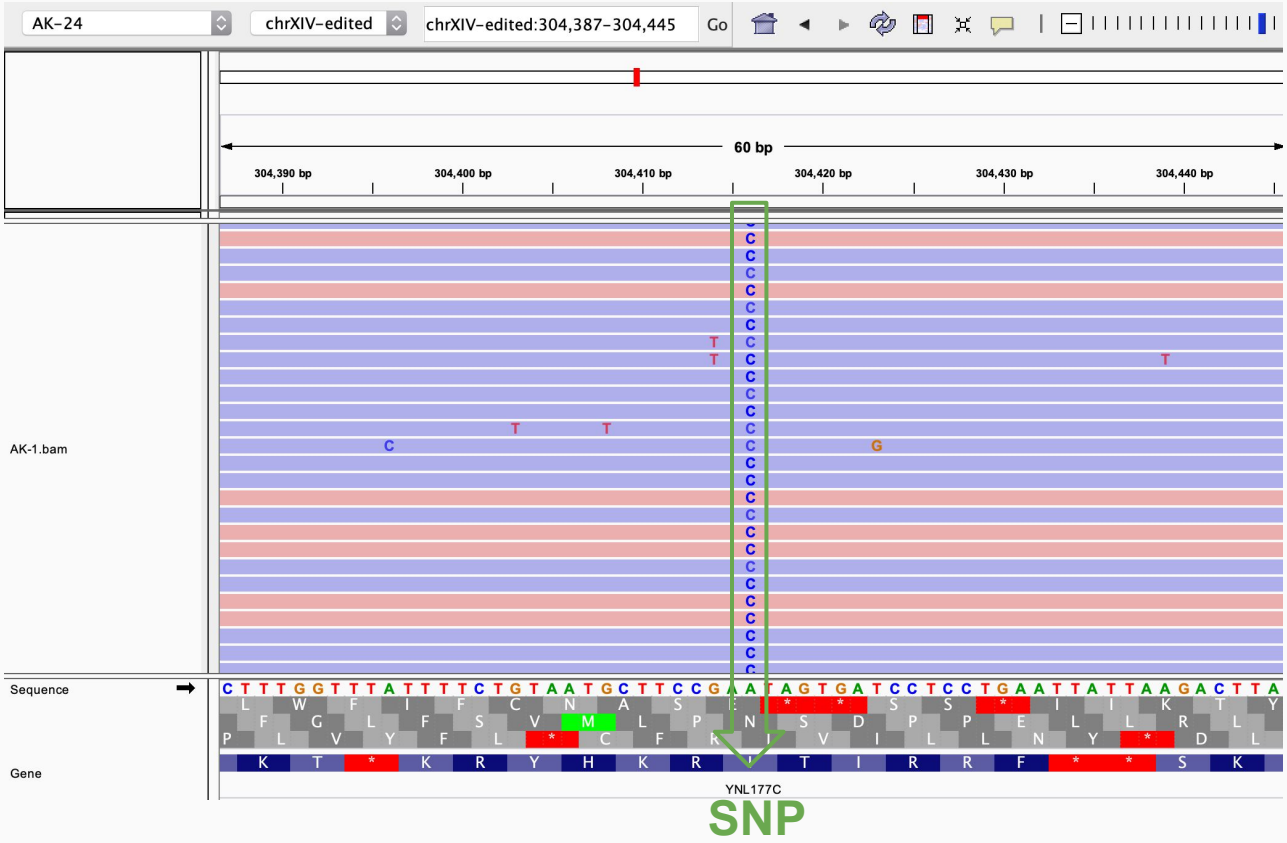
How do we go from a set of **labelled coordinates** to a **human-centered visual summary**?

(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;Bas
eQRankSum=-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
                                                                                0/1:5,5:10:91:91,0,142  0/1:7,2:9:2:2,0,239
    0/1:3,3:6:21:21,0,79    0/1:3,2:5:5:5,0,107
NODE_961_length_155_cov_2566.05_ID_185286       123     .       C       A       13.61   .       AC=1;AF=0.125;AN=8;Bas
eQRankSum=-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;Bas
eQRankSum=-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;Bas
eQRankSum=-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



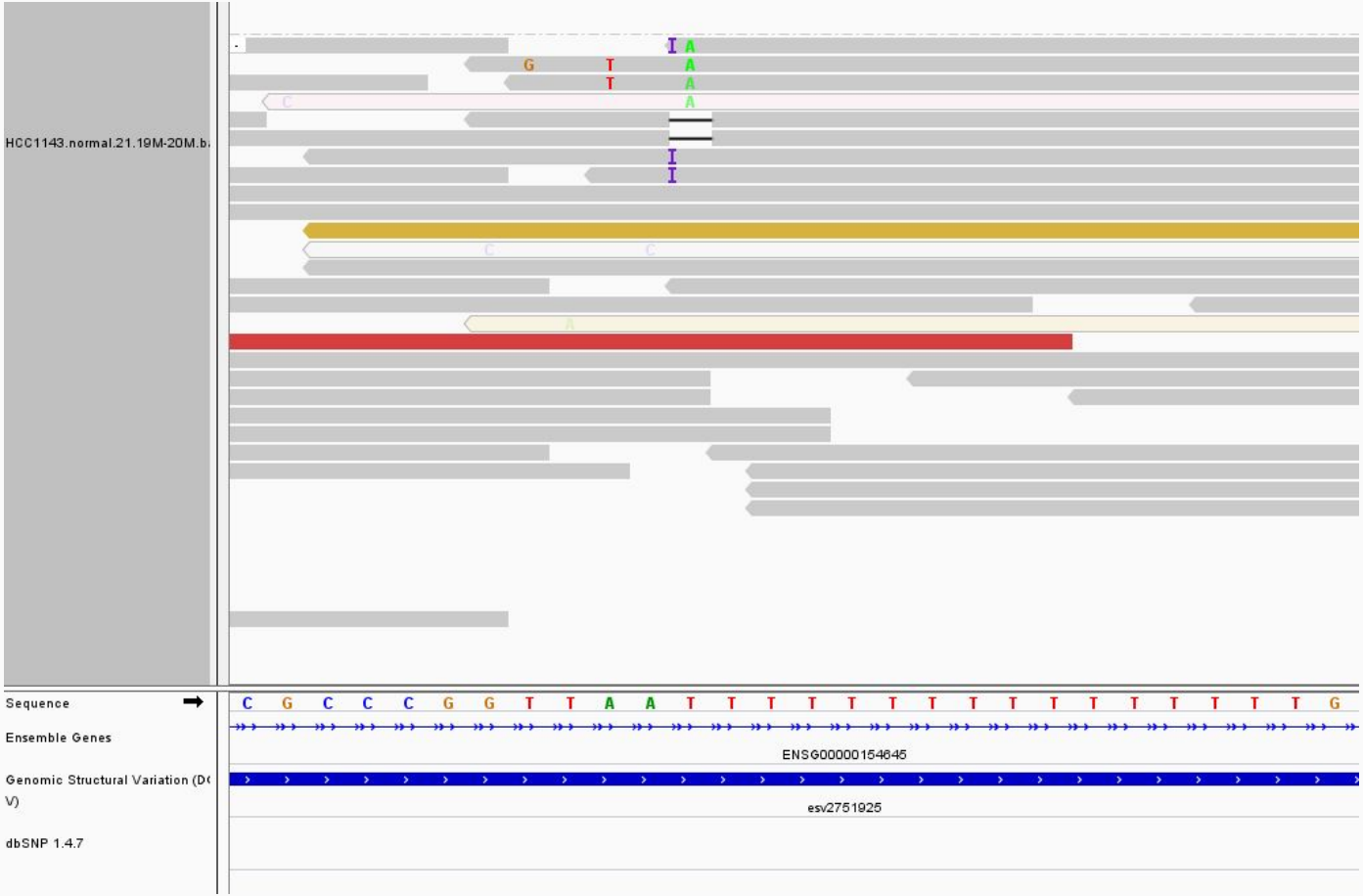
1. Load tracks (.BAM files, .VCF files, etc.). **Here:** Alignment file for 1 sample

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

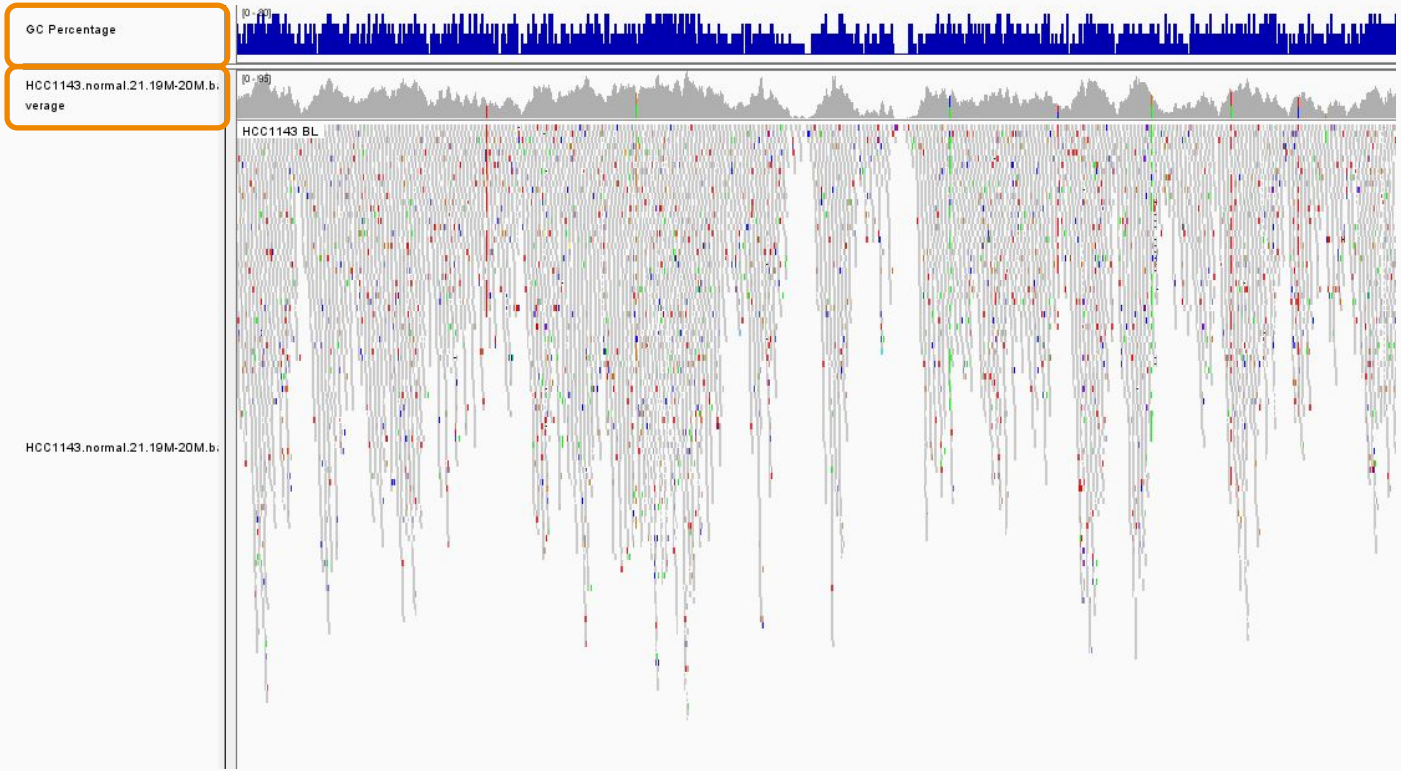
4. Use annotation (.GTF file) to identify which gene SNP is in

IGV: A Homopolymer Run



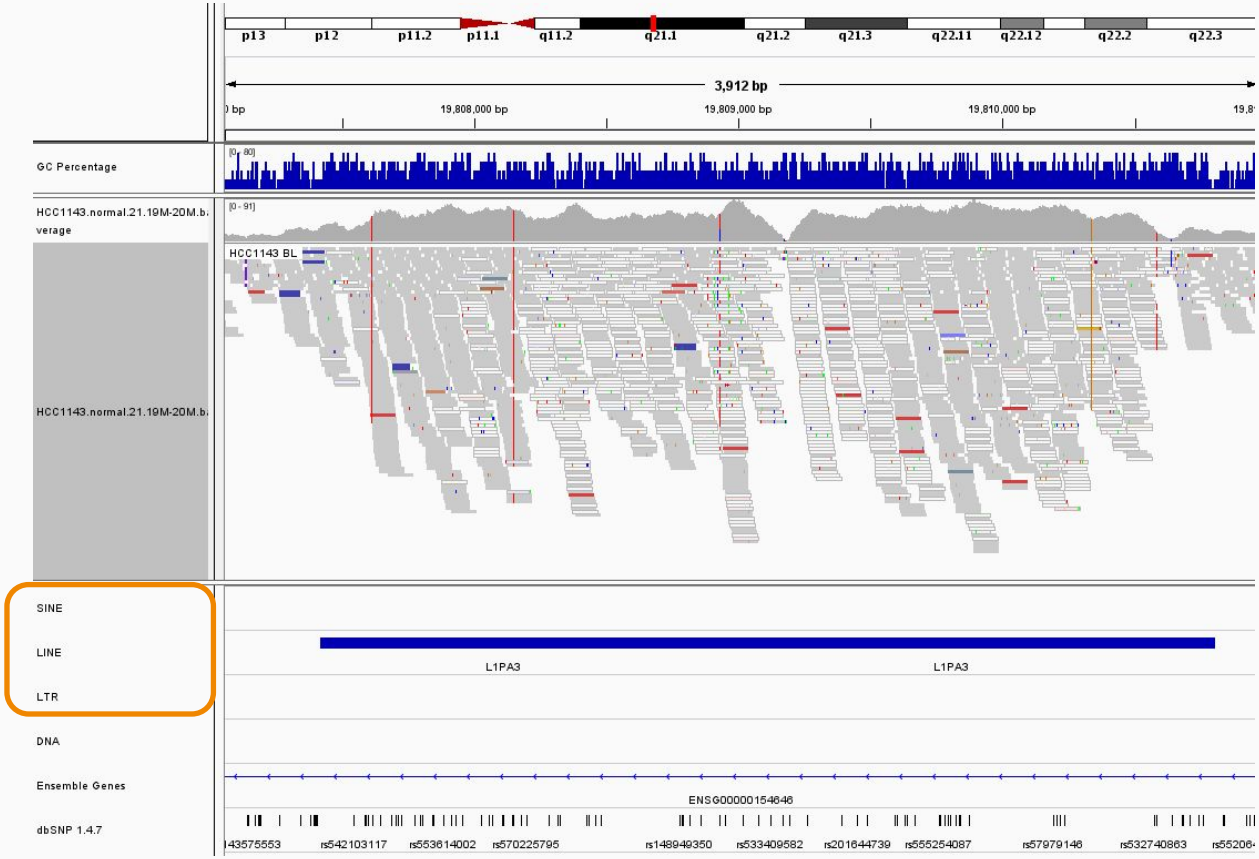
- A long stretch consisting of a single base
- You want to be looking at the sequence here (all those Ts)
- Difficult to map against, particularly at ends of reads

IGV: Coverage by GC percent



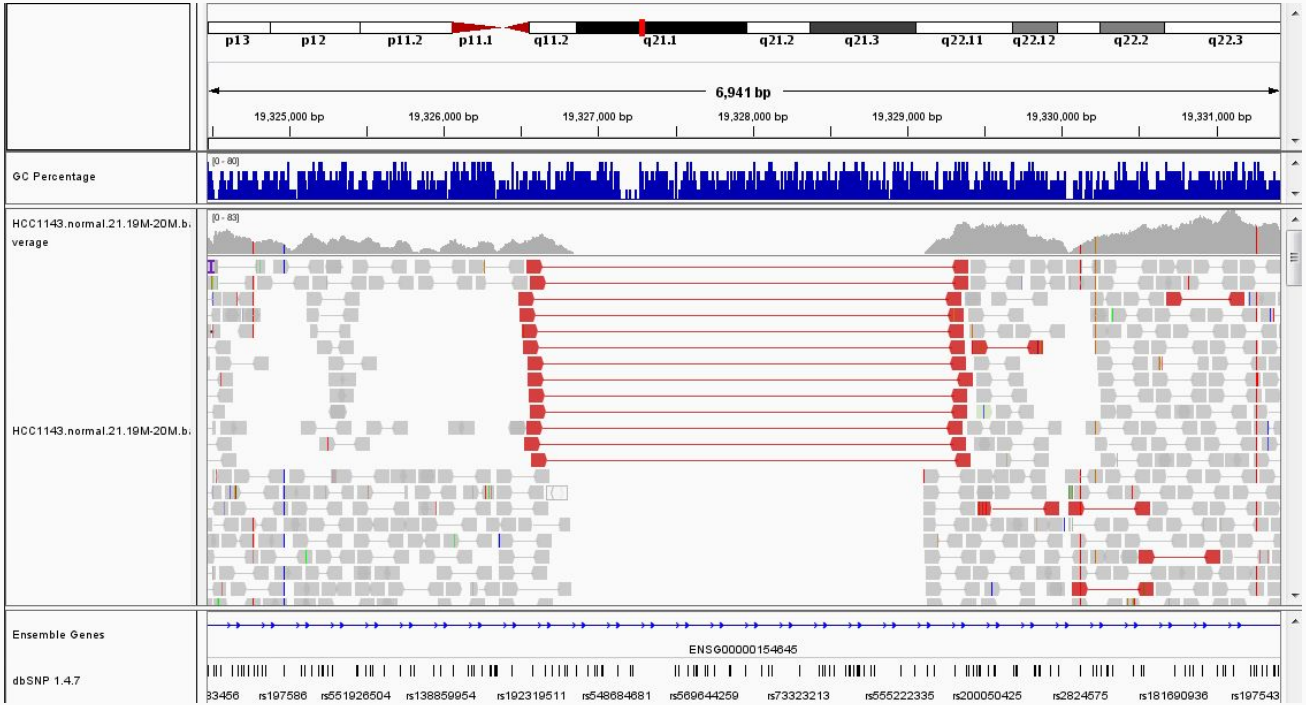
- Benjamini & Speed (2012) proposed that **PCR step** generates this **GC bias**
- Severity differs from experiment to experiment

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).

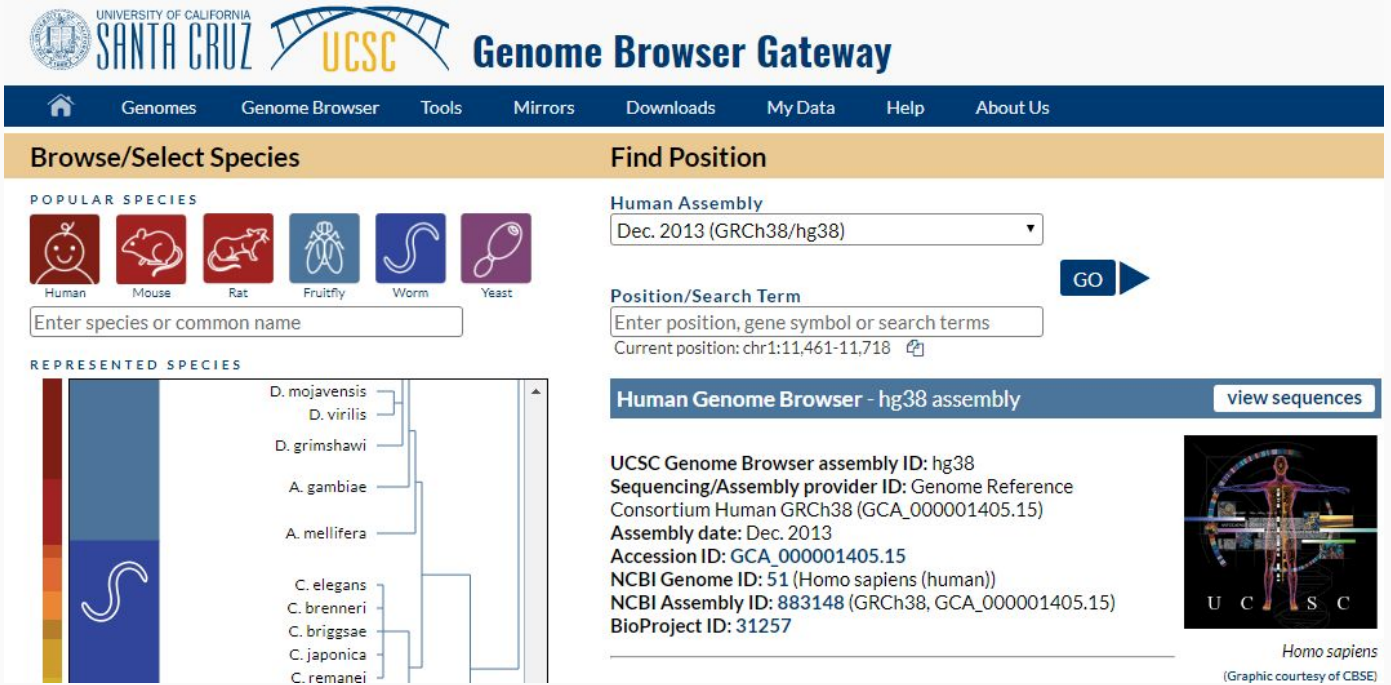
Usage:

```
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>
```

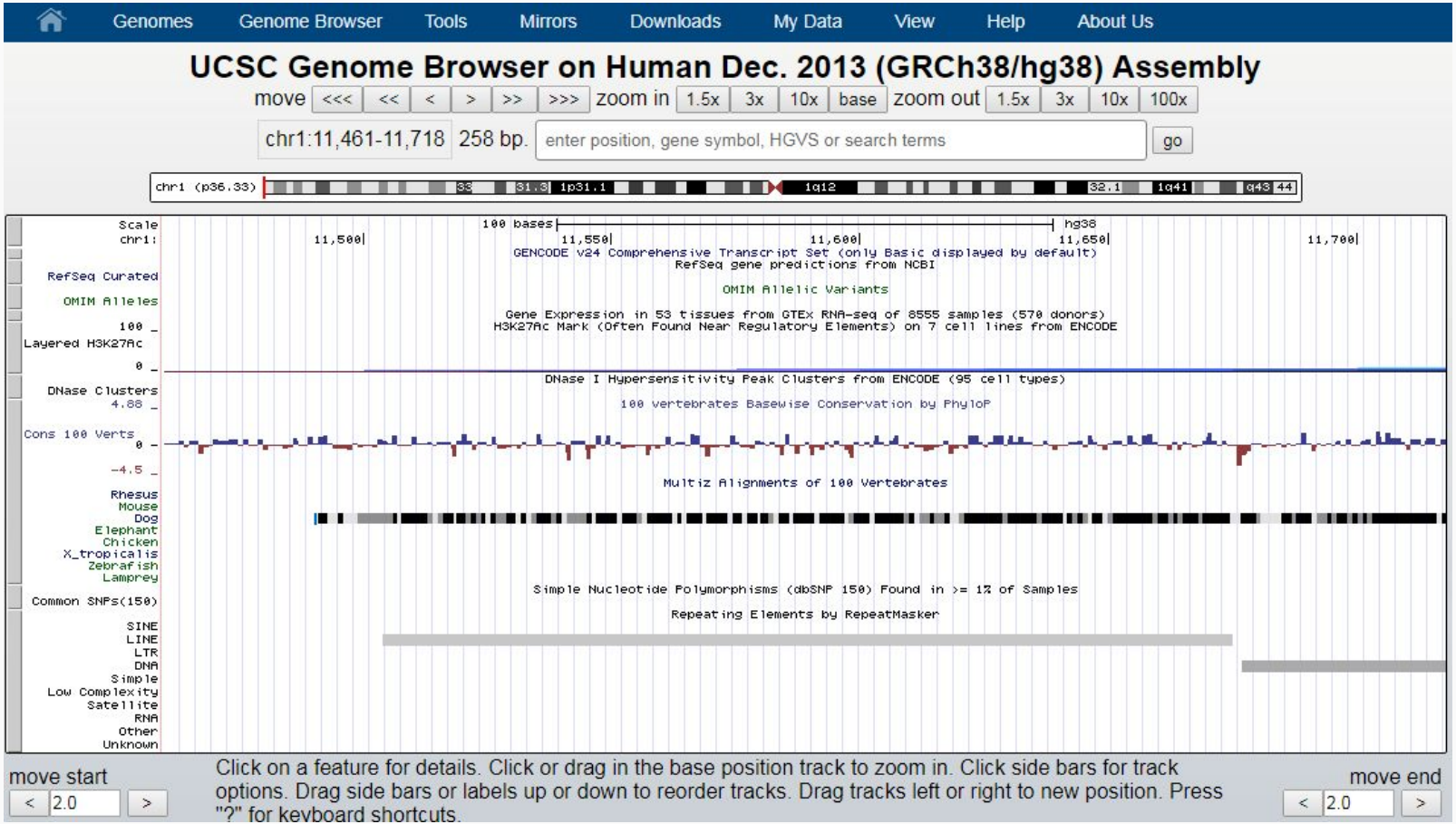
UCSC genome browser

Selecting which species to browse

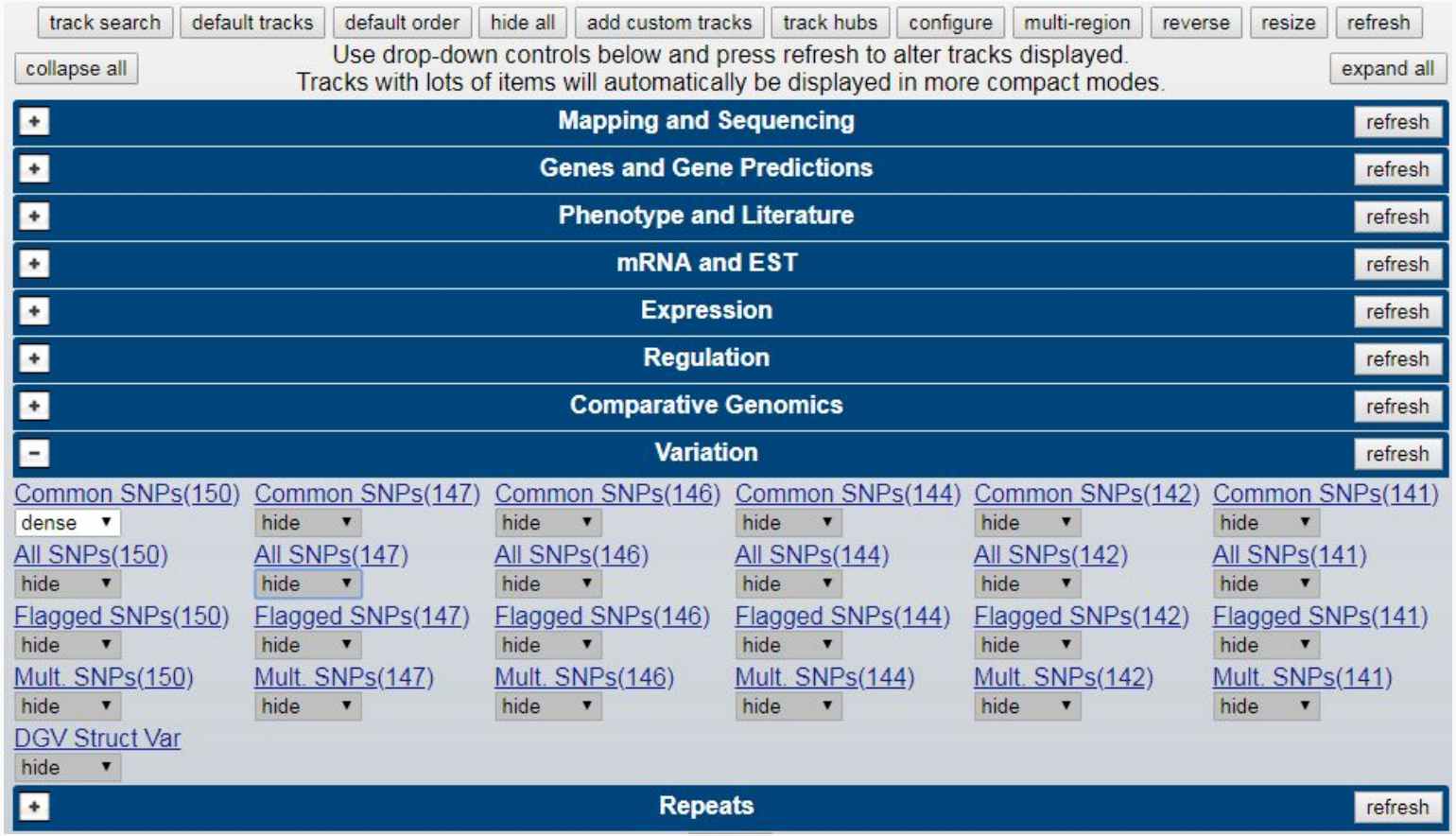


- A wide variety of species/references are available

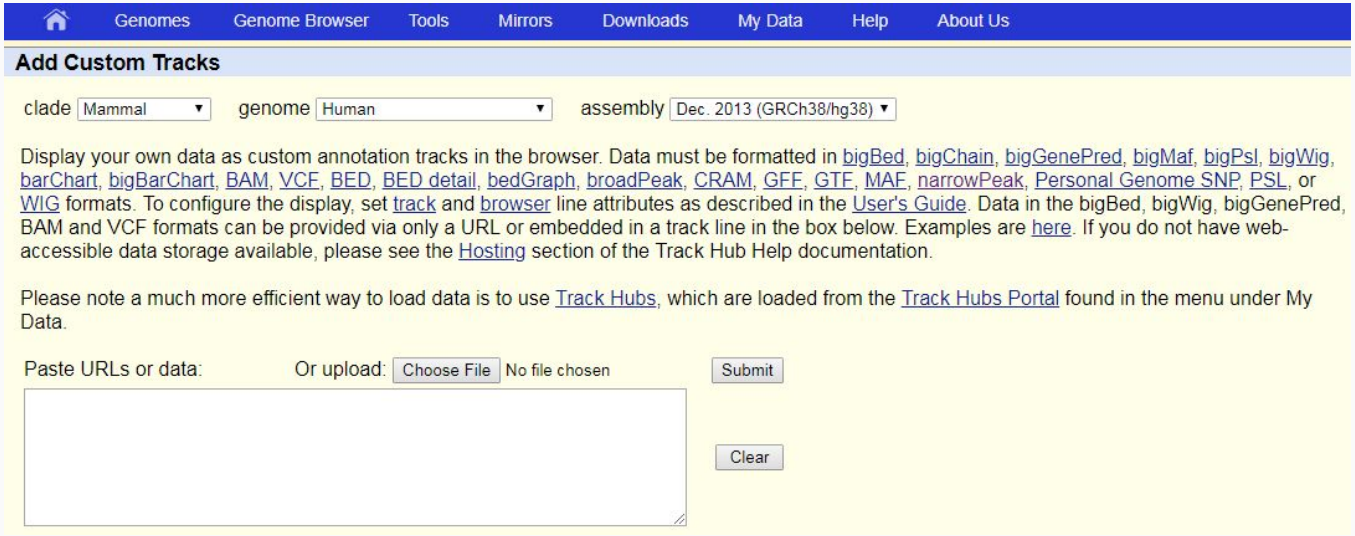
UCSC genome browser interface



Where UCSC beats IGV



Options for viewing your own data



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 bioligical 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 (NCBI37/mm9) | BLAT's guess | query,score | hyperlink

submit | I'm feeling lucky | clear

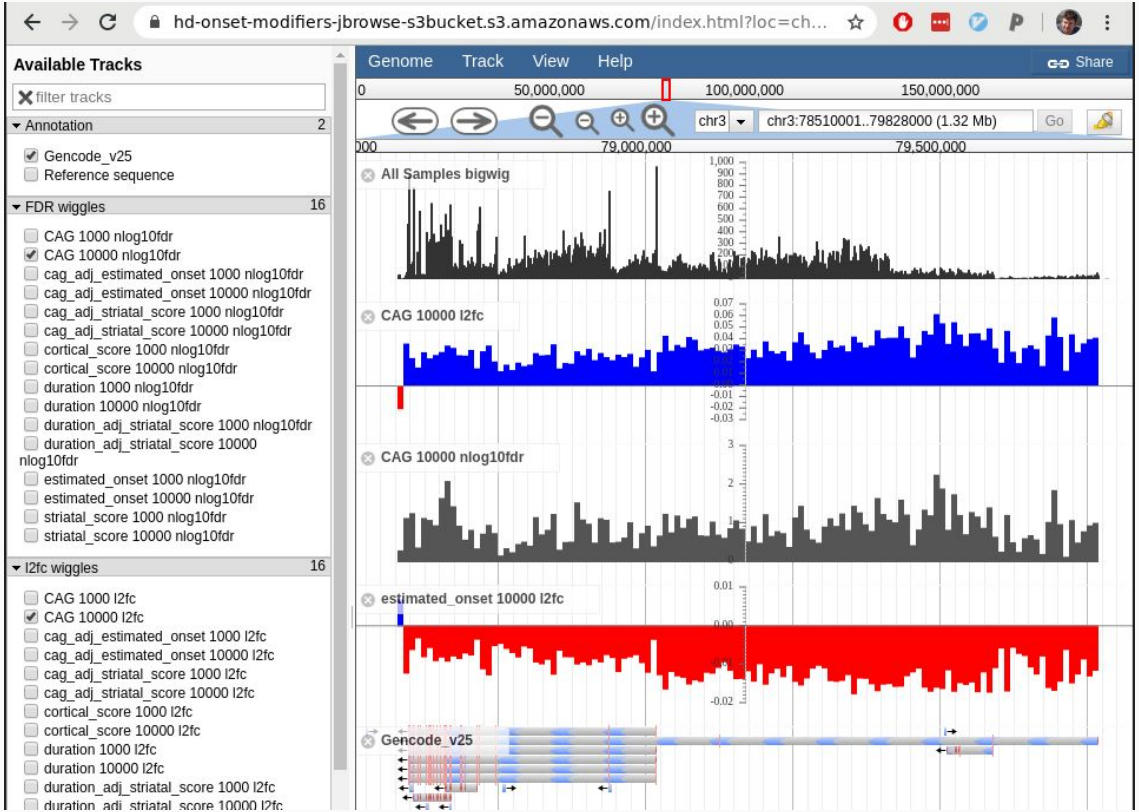
Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

IGV vs. UCSC

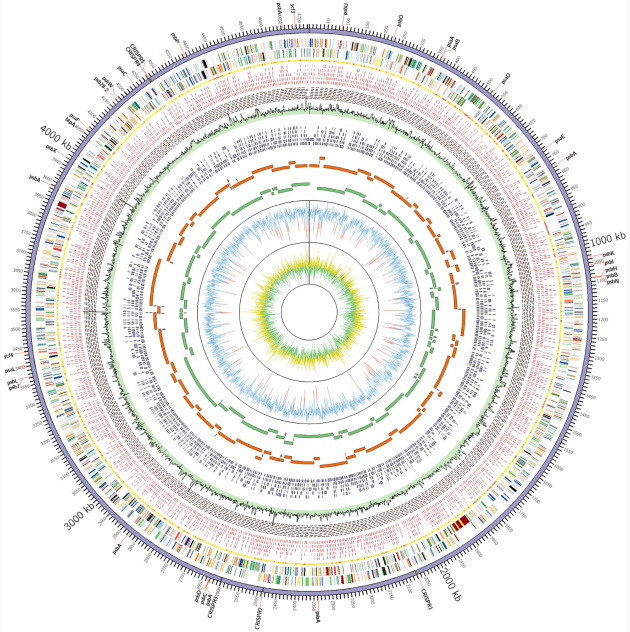
IGV	UCSC
Java application	Run in web browser
Run on your local computer	Hosted on the web
Data stored locally	Data stored on the web
Handles all major file types	Handles all major file types
Must load most tracks manually	Vast number of preloaded track types

Honorable Mention: JBrowse

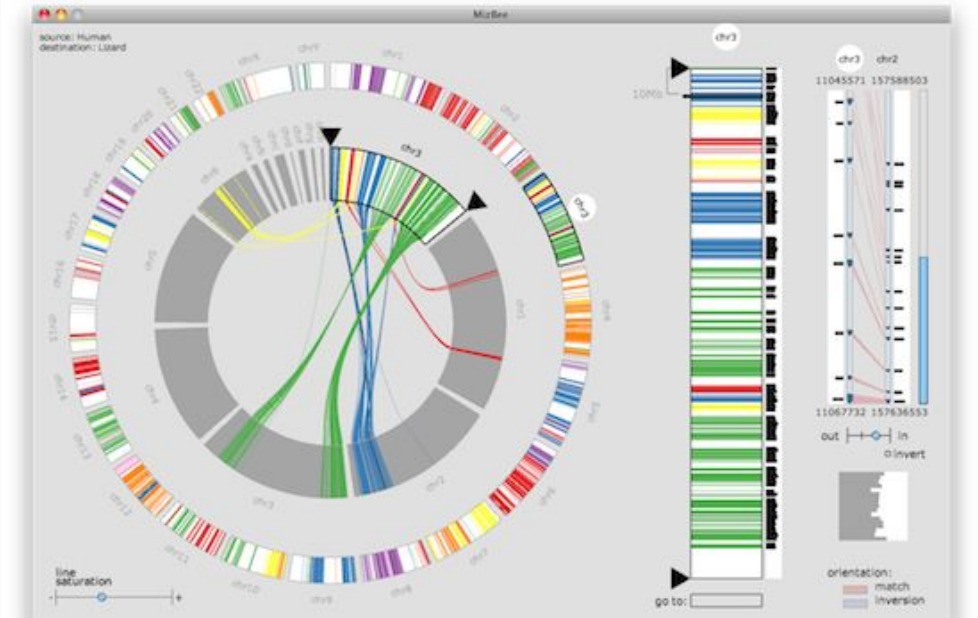
- Javascript-based genome browser software
- Open source
- Highly extensible, customizeable
- Must be hosted on a web server



Other Genome Visualization Tools



Circos
http://circos.ca/software/



MizBee
(A Multiscale Synteny Browser)
http://www.cs.utah.edu/~miriah/mizbee/Overview.html