Running mapping commands and viewing the results

Ester Feldmesser

SAM format

HWI-ST808:87:C068VACXX:2:1101:1234:2199 16 chr5 177482768 2 100M * 0 0
TGACGGTCCATTCCCGGGCTCGATGCCGGAAAAACCCCTTGGCCCGCCGGAAGGGCAGGCACATGGGCATAGGTAAGCGGAAGGGTACAGCCAATGCACG
######@CA?5&DBB@@9BA99<7@98:(?@?5)5(@?<807DCBHFHGBIIIHHHEF@@FB?3GIIIIGGGGEIGFIIIHEFBIGFFFBHFEDDAA:=:
AS:i:-29 XS:i:-32 XN:i:0 XM:i:6 XO:i:0 XG:i:0 NM:i:6 MD:Z:35A26G3C7G3A18C2 YT:Z:UU

Alignments are reported in a compact representation: SAM format

```
0        61G9EAAXX100520:5:100:10095:16477  (read name)
1        83  (FLAGS stored as bit fields; 83 = 00001010011 )
2        chr1  (alignment target)
3        51986 (position alignment starts)
4        38
5        46M  (Compact description of the alignment in CIGAR format)
6        =
7        51789
8        -264
9        CCCAAACAAGCCGAACTAGCTGATTTGGCTCGTAAAGACCCGGAAA
10       ###CB?=ADDBCBCDEEFFDEFFFDEFFGDBEFGEDGCFGFGGGGG
11       MD:Z:67
12       NH:i:1
13       HI:i:1
14       NM:i:0
15       SM:i:38
16       XQ:i:40
17       X2:i:0
```

(read sequence, oriented according to the forward alignment)

(base quality values)

(Metadata)

SAM format specification: http://samtools.sourceforge.net/SAM1.pdf

Mapping reads

- Bowtie
- Star
- Why do we need 2 software?
- What are the differences between them?

Bowtie and its options (parameters)

What is the minimum info we need to tell to a mapping software?

- Reads, sequences or clusters
- Indexed genome

- More…
  - how many mismatches
  - how many times each read is allowed to map

Star and its options (parameters)

- Reads, sequences or clusters
- Indexed genome

- More…
  - how many mismatches
  - how many times each read is allowed to map

Quantification with Star

Counting reads in genes with STAR:

- Use uniquely mapped reads

--outFilterMultimapNmax 1

- Use gene information to know where the exons are

--sjdbGTFfile /shareDB/iGenomes/Arabidopsis_thaliana/NCBI/TAIR10/Annotation/Genes/genes.gtf

- Count reads per gene

--quantMode GeneCounts

GTF file

The Gene transfer format (**GTF**) is a **file** format used to hold information about gene structure.

```
1	unknown exon	3631	3913	.	+	.	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown CDS	3760	3913	.	+	0	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown start_codon	3760	3762	.	+	.	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown CDS	3996	4276	.	+	2	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown exon	3996	4276	.	+	.	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown CDS	4486	4605	.	+	0	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown exon	4486	4605	.	+	.	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown CDS	4706	5095	.	+	0	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown exon	4706	5095	.	+	.	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown CDS	5174	5326	.	+	0	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown exon	5174	5326	.	+	.	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown CDS	5439	5627	.	+	0	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown exon	5439	5899	.	+	.	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown stop_codon	5628	5630	.	+	.	gene_id "NAC001"; gene_name "NAC001"; p_id "P20197"; transcript_id "NM_099983.2"; tss_id "TSS18986";
1	unknown exon	5928	6263	.	-	.	gene_id "ARV1"; gene_name "ARV1"; p_id "P11536"; transcript_id "NM_099984.5"; tss_id "TSS12104";
1	unknown exon	6437	7069	.	-	.	gene_id "ARV1"; gene_name "ARV1"; p_id "P11536"; transcript_id "NM_099984.5"; tss_id "TSS12104";
1	unknown exon	6790	7069	.	-	.	gene_id "ARV1"; gene_name "ARV1"; p_id "P31963"; transcript_id "NM_001035846.1"; tss_id "TSS12104";
1	unknown stop_codon	6915	6917	.	-	.	gene_id "ARV1"; gene_name "ARV1"; p_id "P11536"; transcript_id "NM_099984.5"; tss_id "TSS12104";
1	unknown CDS	6918	7069	.	-	2	gene_id "ARV1"; gene_name "ARV1"; p_id "P11536"; transcript_id "NM_099984.5"; tss_id "TSS12104";
1	unknown CDS	7157	7232	.	-	0	gene_id "ARV1"; gene_name "ARV1"; p_id "P11536"; transcript_id "NM_099984.5"; tss_id "TSS12104";
1	unknown exon	7157	7232	.	-	.	gene_id "ARV1"; gene_name "ARV1"; p_id "P11536"; transcript_id "NM_099984.5"; tss_id "TSS12104";
1	unknown exon	7157	7450	.	-	.	gene_id "ARV1"; gene_name "ARV1"; p_id "P31963"; transcript_id "NM_001035846.1"; tss_id "TSS12104";
1	unknown stop_codon	7315	7317	.	-	.	gene_id "ARV1"; gene_name "ARV1"; p_id "P31963"; transcript_id "NM_001035846.1"; tss_id "TSS12104";
```

Output format

SAM format

```
HWI-ST808:87:C068VACXX:2:1101:1234:2199        16       chr5     177482768     2       100M     *       0 0
TGACGGTCCATTCCCGGGCTCGATGCCGGAAAAACCCCTTGGCCCGCCGGAAGGGCAGGCACATGGGCATAGGTAAGCGGAAGGGTACAGCCAATGCACG
######@CA?5&DBB@@9BA99<7@98:(?@?5)5(@?<807DCBHFHGBIIIHHHEF@@FB?3GIIIIGGGGEIGFIIIHEFBIGFFFBHFEDDAA:=:
AS:i:-29          XS:i:-32          XN:i:0   XM:i:6  XO:i:0   XG:i:0   NM:i:6 MD:Z:35A26G3C7G3A18C2    YT:Z:UU
```

BAM format – Binary format of the SAM file

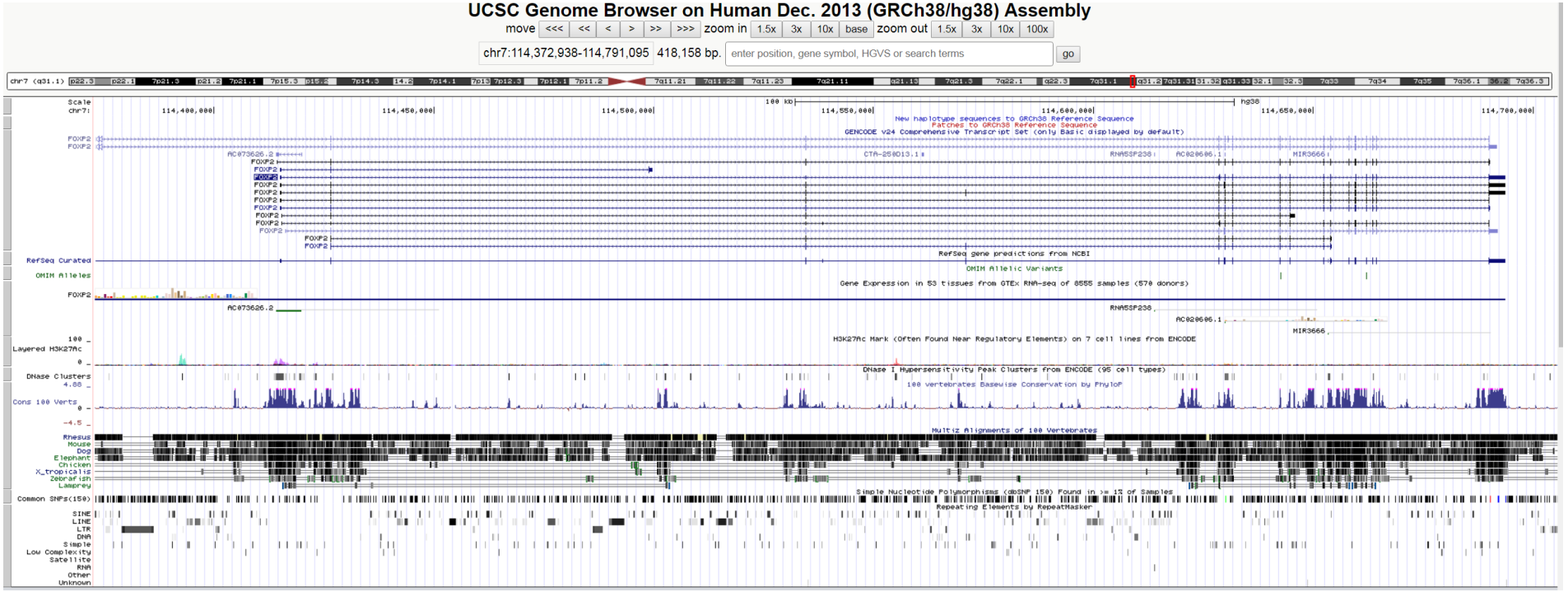
BAM.BAI format – Index for the bam format

name_log.txt – Mapping statistics

```
25000 reads; of these:
  25000 (100.00%) were unpaired; of these:
    3612 (14.45%) aligned 0 times
    14794 (59.18%) aligned exactly 1 time
    6594 (26.38%) aligned >1 times
85.55% overall alignment rate
```

Genome browsers

The basic paradigm of display is to show the genome sequence in the horizontal dimension, and show graphical representations of the locations of genome features such as mRNAs, gene predictions or any feature that coordinates can be assigned to it.

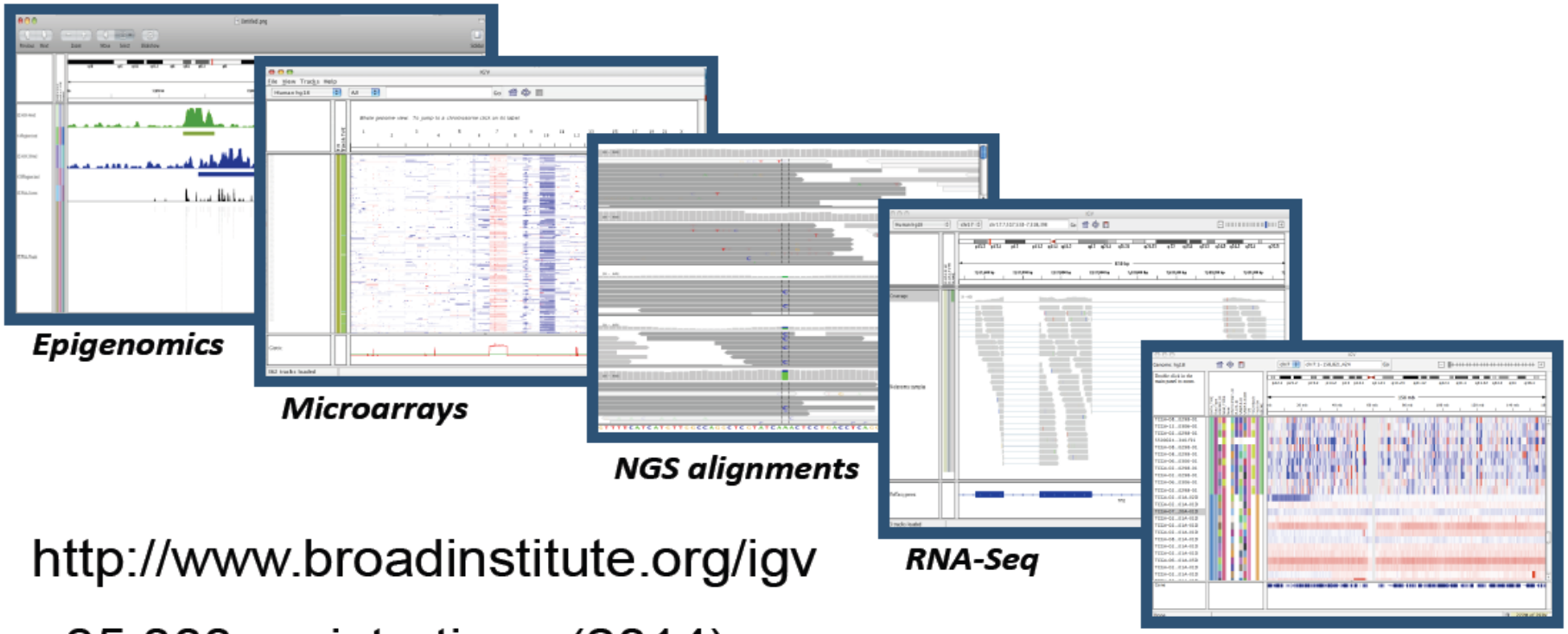


Integrated Genomics Viewer (IGV)

- The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.
- It was developed at the Broad Institute (http://software.broadinstitute.org/software/igv/home).
- It supports a wide variety of data types, including next-generation sequence data, and genomic annotations.
- It supports a wide variety of file formats to upload data (http://software.broadinstitute.org/software/igv/FileFormats).

Integrative Genomics Viewer (IGV)

Desktop application for the interactive visual exploration of integrated genomic datasets

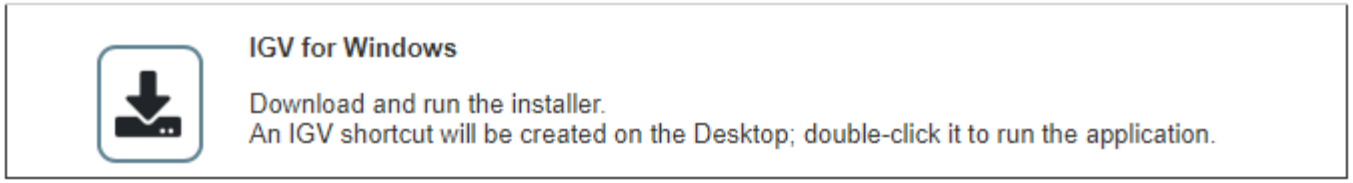


http://www.broadinstitute.org/igv

>85,000 registrations (2014)

Opening IGV

1. Go http://software.broadinstitute.org/software/igv/download
2. You might be requested to register if you are not
3. Click on the arrow to download the IGV installer

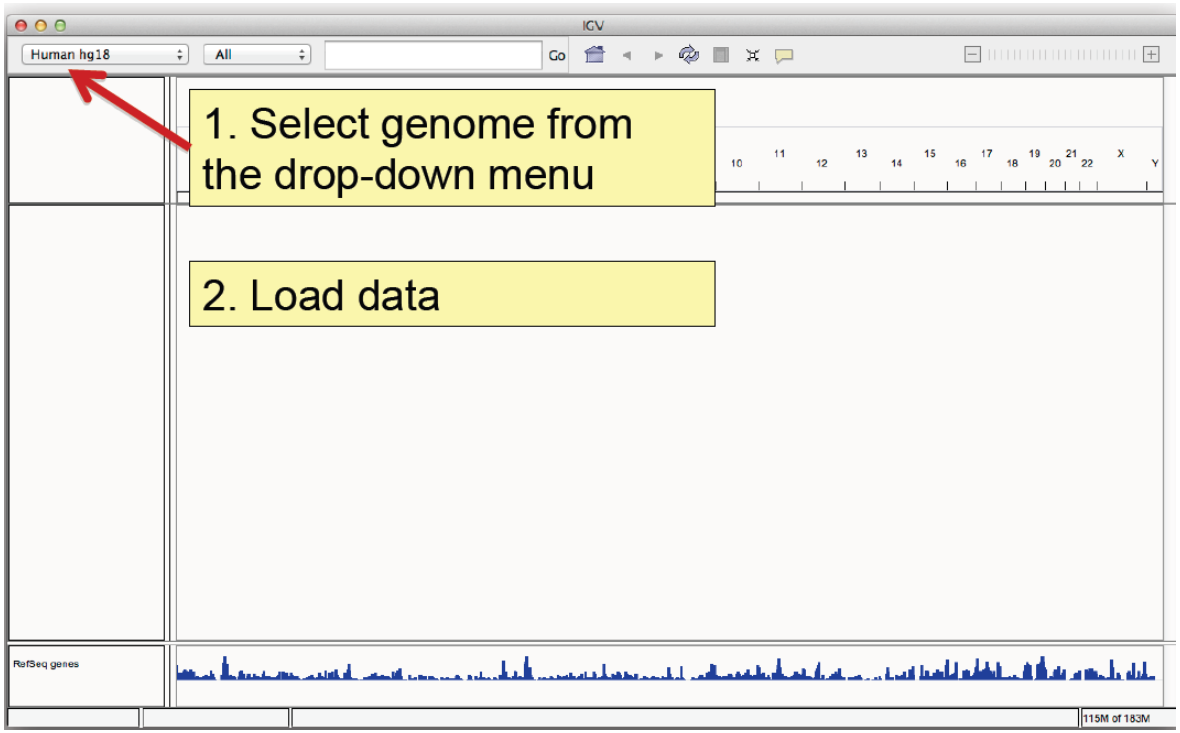


4. Double-click the installer to run the installation
5. An IGV shortcut will be created on the Desktop
6. Double-click it to run the application.
7. Allow Java to run if requested
8. Note: All the files that you load need to be indexed

Using IGV

User guide:
http://software.broadinstitute.org/software/igv/book/export/html/6

A genome will be loaded when you open IGV. You can change the genome by clicking the drop down menu in the upper-left.

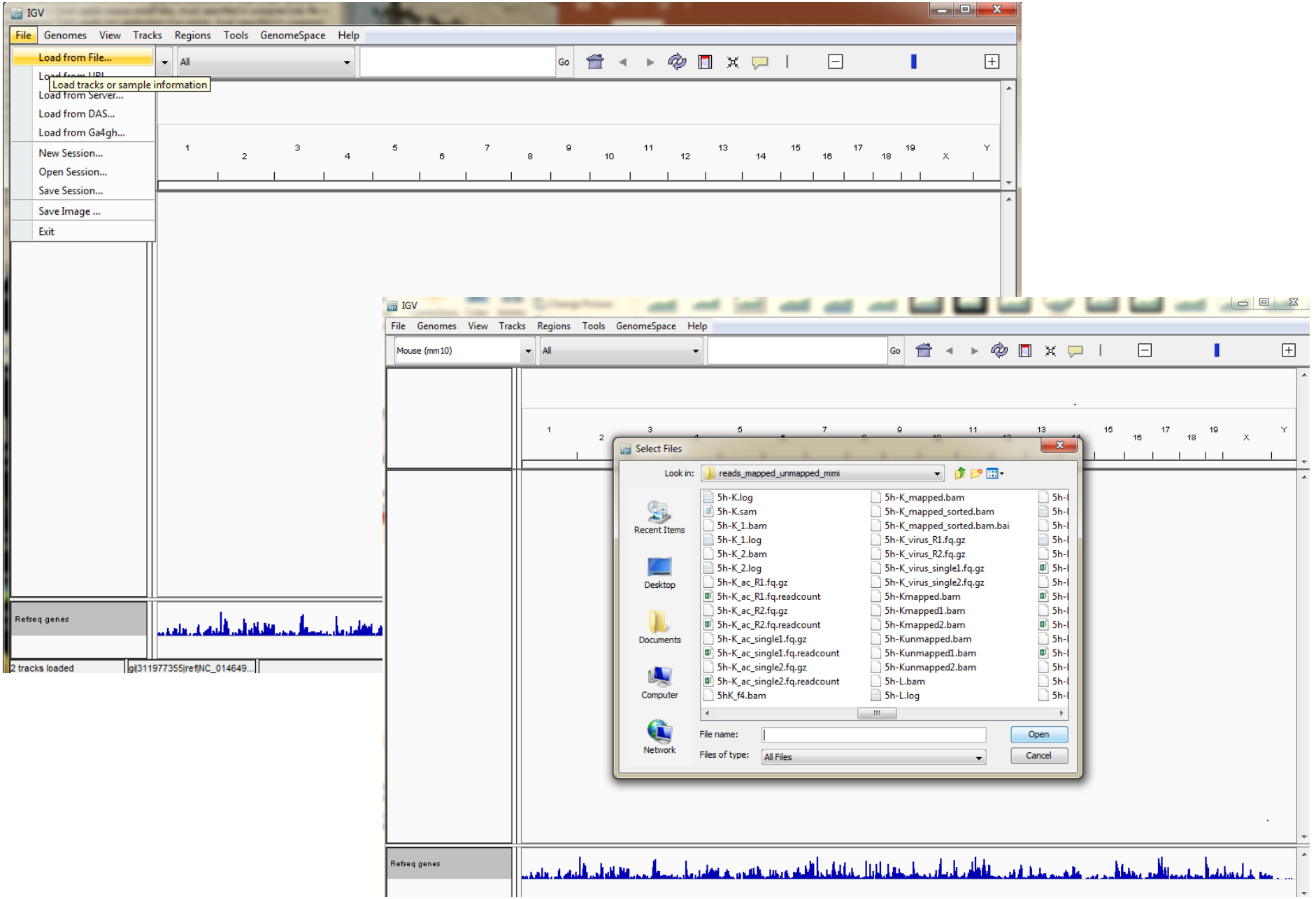


File Formats

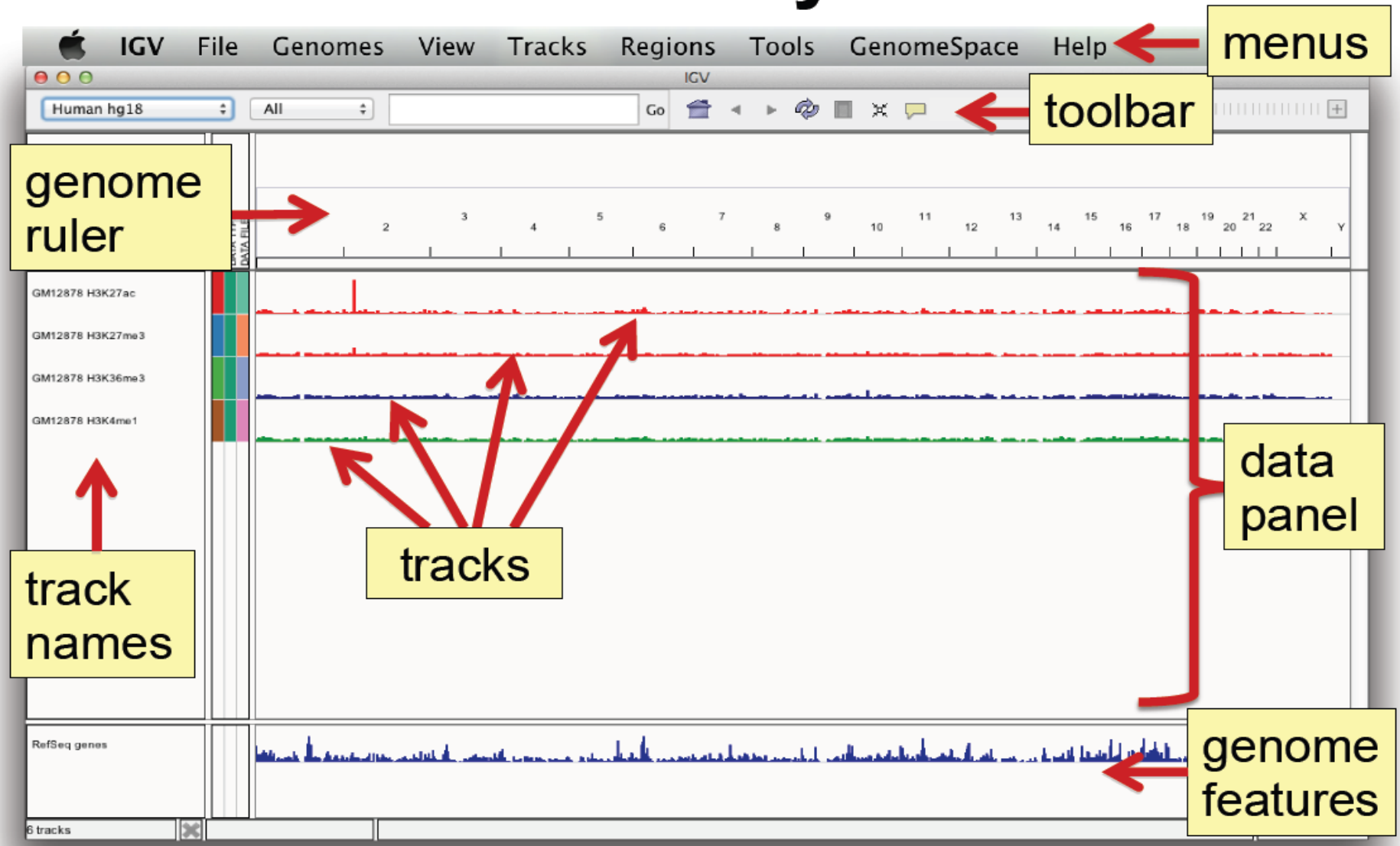
Most files formats need an index

- BAM
- BED
- broadPeak
- FASTA
- GFF/GTF
- Merged BAM File
- narrowPeak
- PSL
- SNP
- TAB
- TDF
- VCF

Load the data

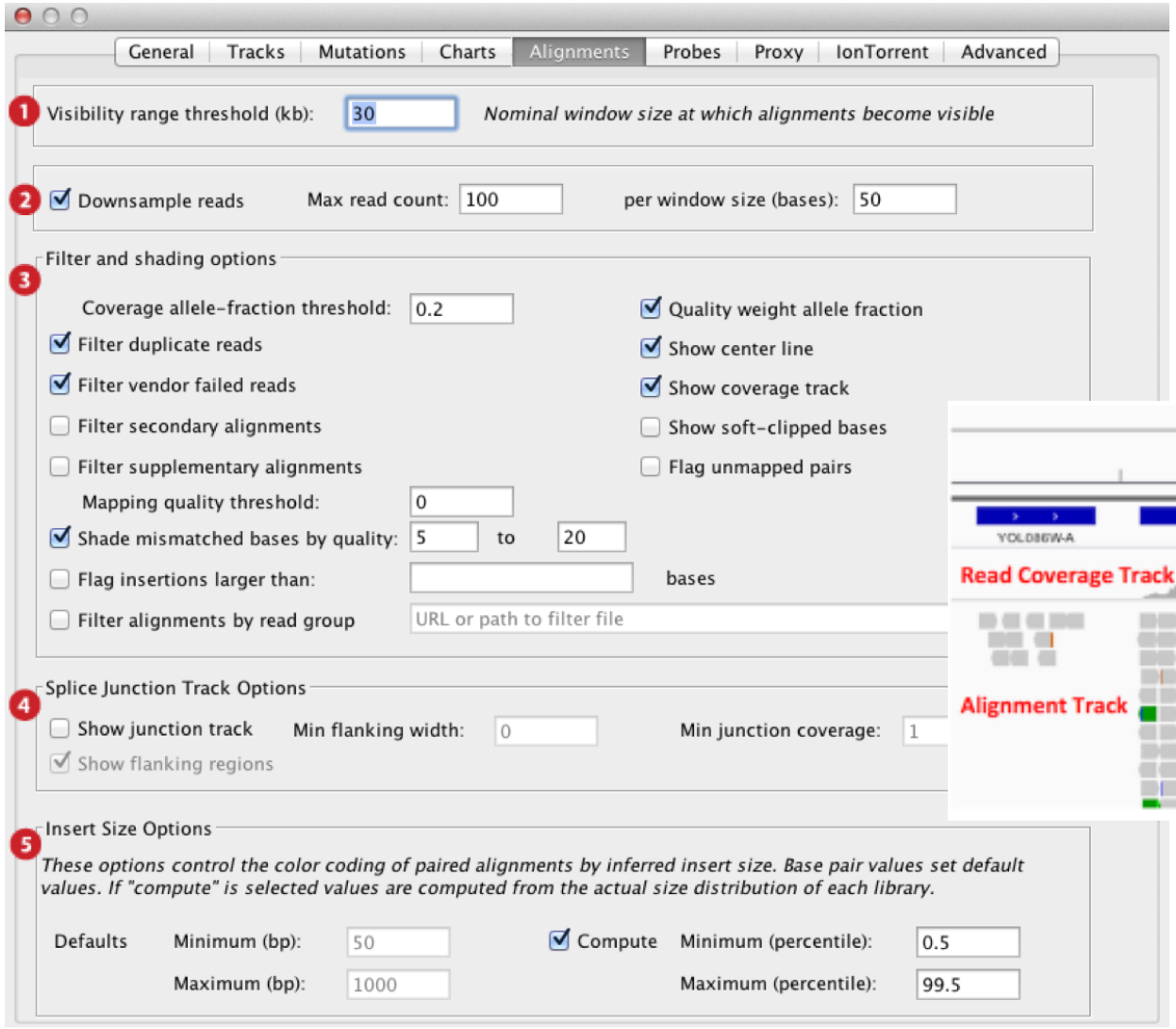


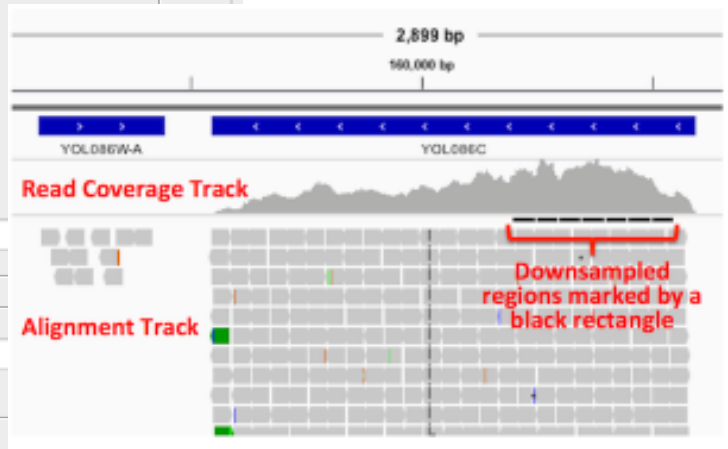
Screen layout



Preferences

To display the Preferences window, click *View>Preferences*.

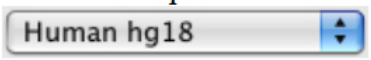




http://software.broadinstitute.org/software/igv/Preferences#Alignments

Tool Bar

Genome drop-down box

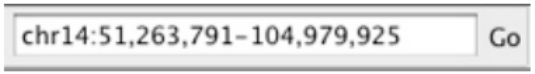


Loads a genome. more...

Chromosome drop-down box

Zooms to a chromosome. more...

Search box



Displays the chromosome location being shown. To scroll to a different location, enter the gene name, locus, or track name and click Go. more...

Whole genome view

Zooms to whole genome view. more...

Moves backward and forward through views of the genome like the back and forward buttons in a web browser.

Refresh

Refreshes the display.

Define a region

Defines a region of interest on the chromosome. more...

Reduces the row height on all tracks to fit all data for the region in view into the window; will also expand tracks (to their maximum preferred size) to fill the view, if needed.

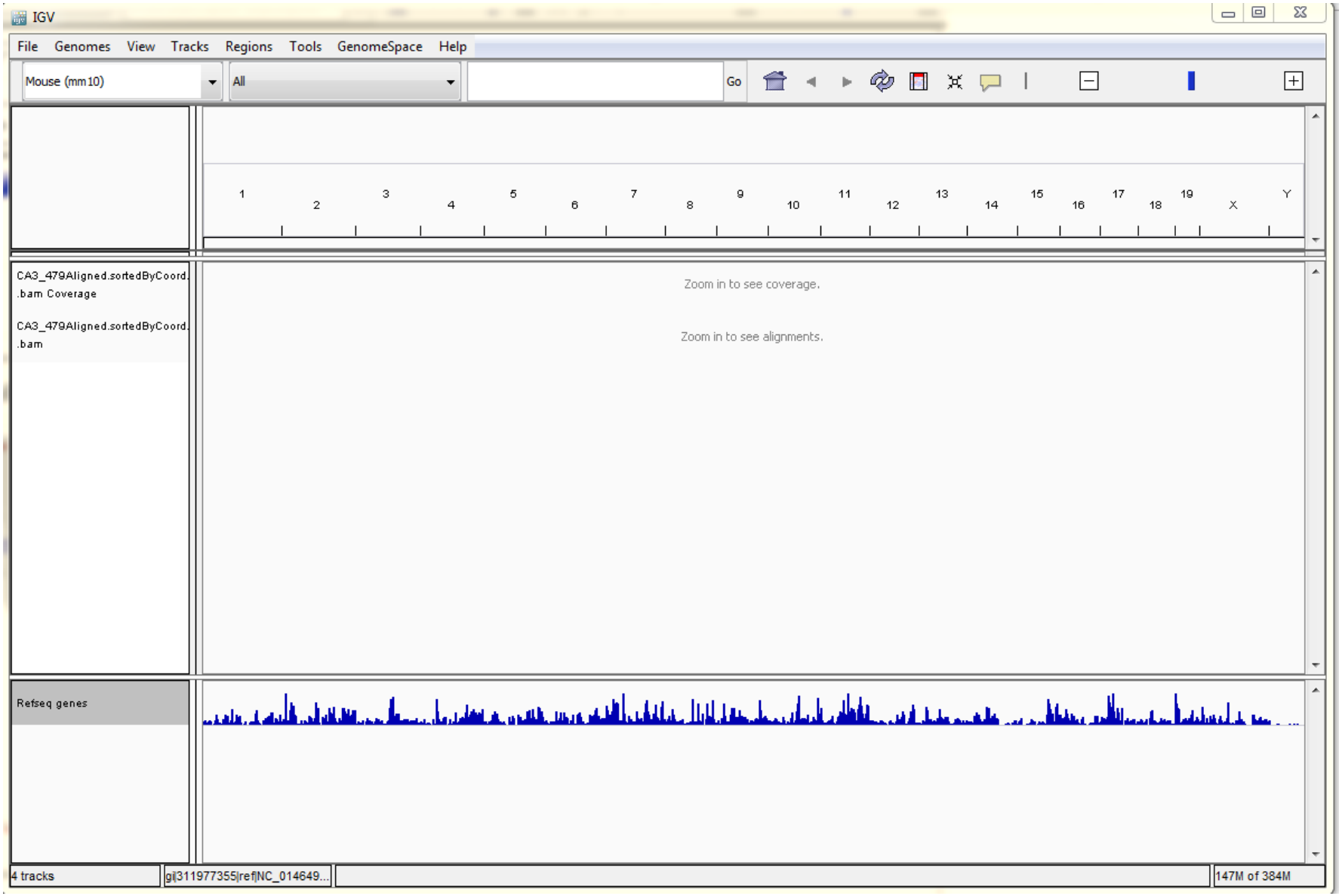
Controls the popup information behavior. Options include displaying the information as the cursor hovers over an item, or when the item is clicked. The popup can also be disabled.

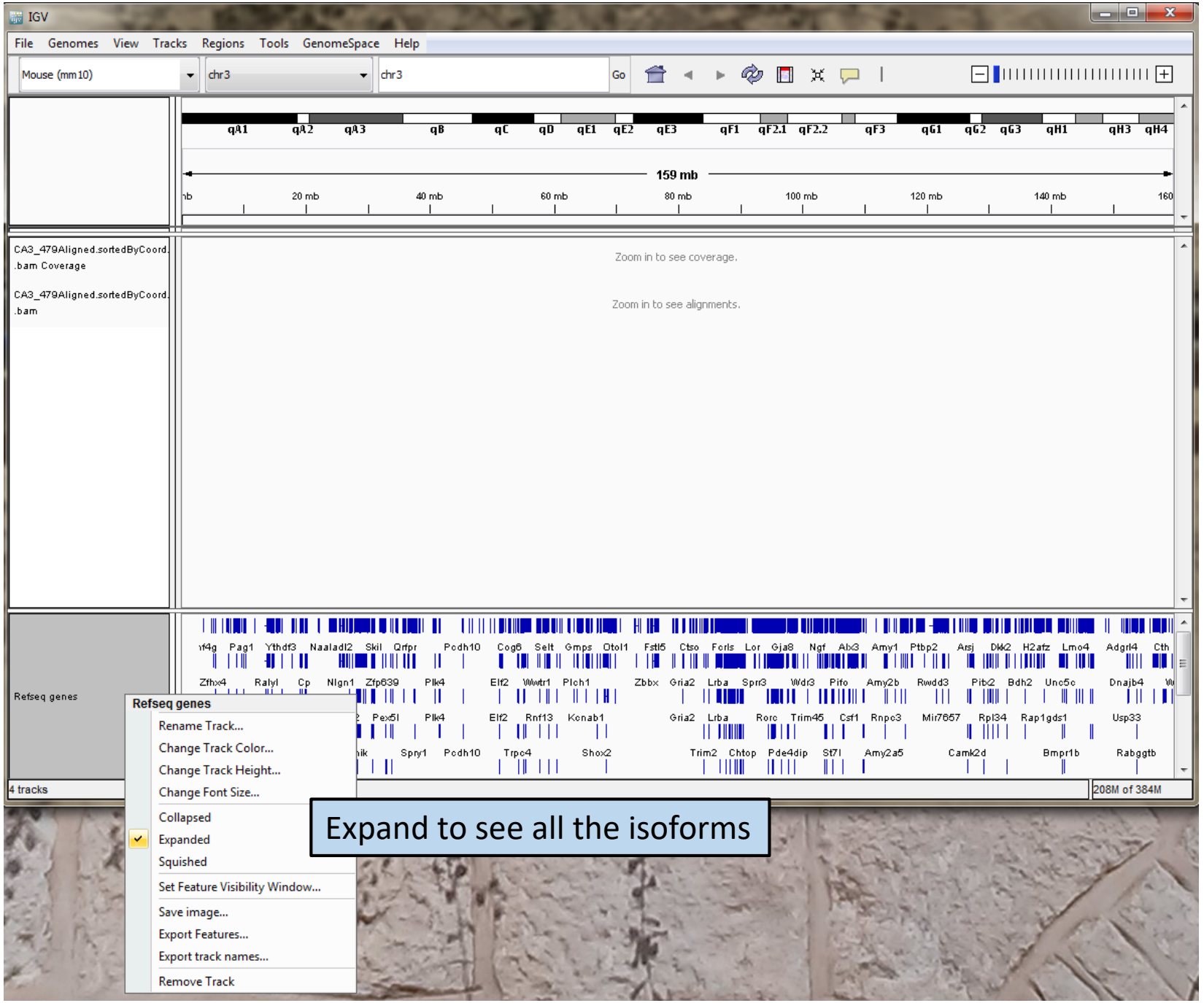
Zoom slider

Zooms in and out on a chromosome. Sometimes referred to as the "railroad track." more...

http://software.broadinstitute.org/software/igv/book/export/html/6

Viewing alignments – Zoom in




IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Mouse (mm10)
chr3
chr3
Go
159 mb
CA3_479Aligned.sortedByCoord.bam Coverage
CA3_479Aligned.sortedByCoord.bam
Zoom in to see coverage.
Zoom in to see alignments.
Refseq genes
Refseq genes
Rename Track...
Change Track Color...
Change Track Height...
Change Font Size...
Collapsed
Expanded
Squished
Set Feature Visibility Window...
Save image...
Export Features...
Export track names...
Remove Track
4 tracks
208M of 384M
Expand to see all the isoforms

Feature Track Options

Viewing Options for the Feature Track

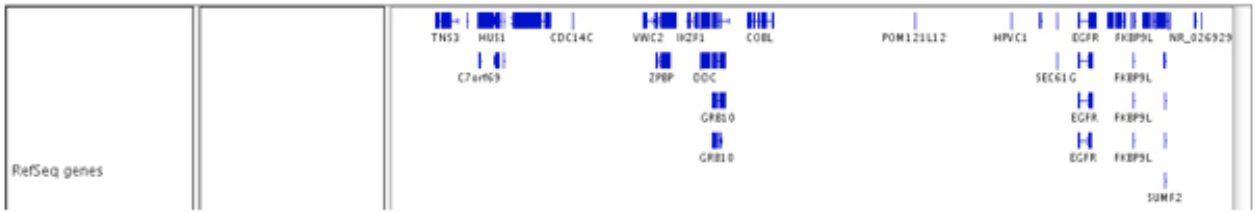
There are 3 different options for viewing the feature track. These allow you to display overlapping features, such as different transcripts of a gene, on one line or multiple lines

To change the view of the feature track, right-click on the feature track and select one of the options:

Collapsed:

Squished:

Expanded:



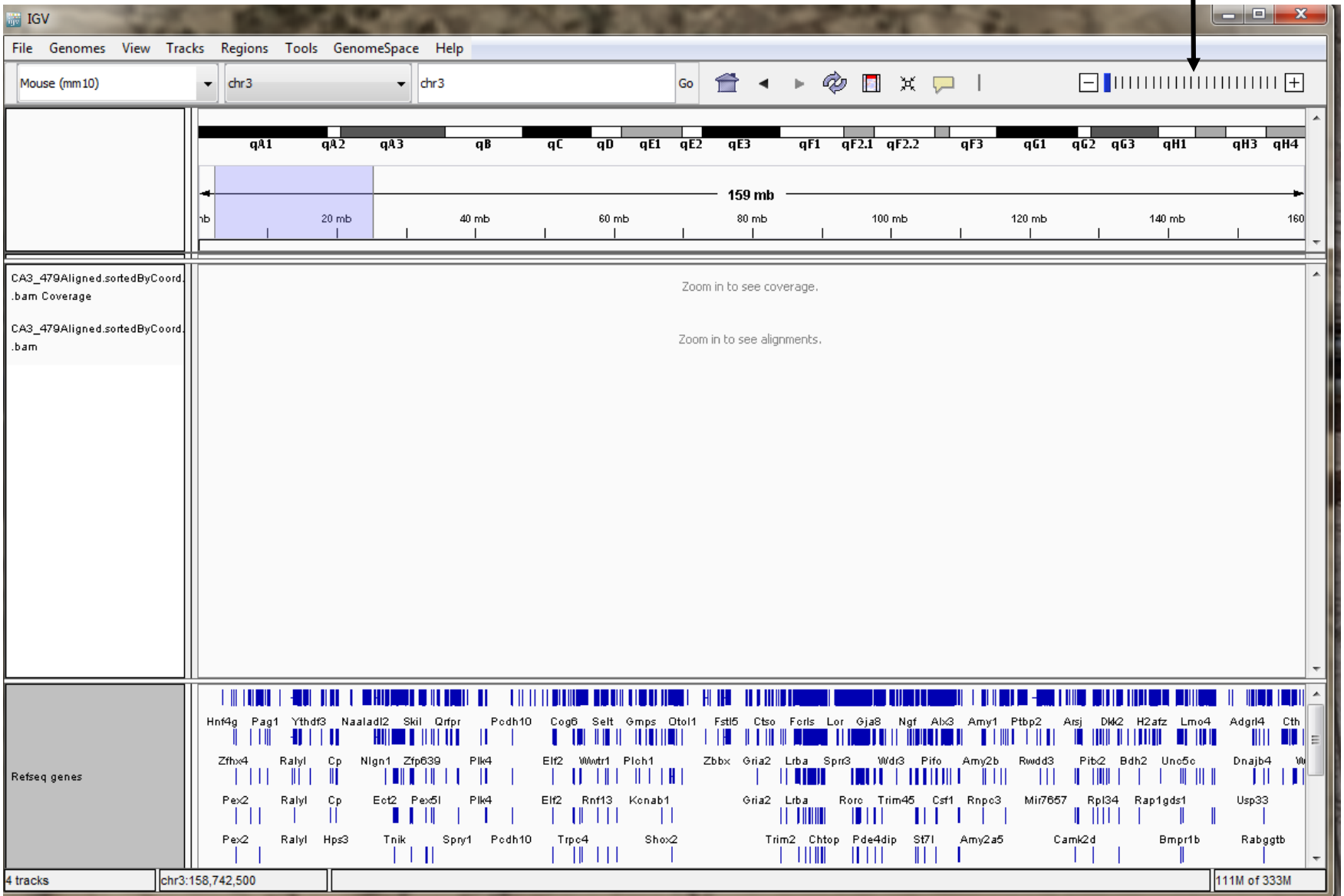
Exon Jumping

This feature is similar to feature jumping. To feature-jump, you select a feature track and press Ctrl-F for forward, Ctrl-B for back. To exon-jump, you select a feature track and press SHIFT-Ctrl-F to center the next exon in your view, SHIFT-Ctrl-B to move back one exon.

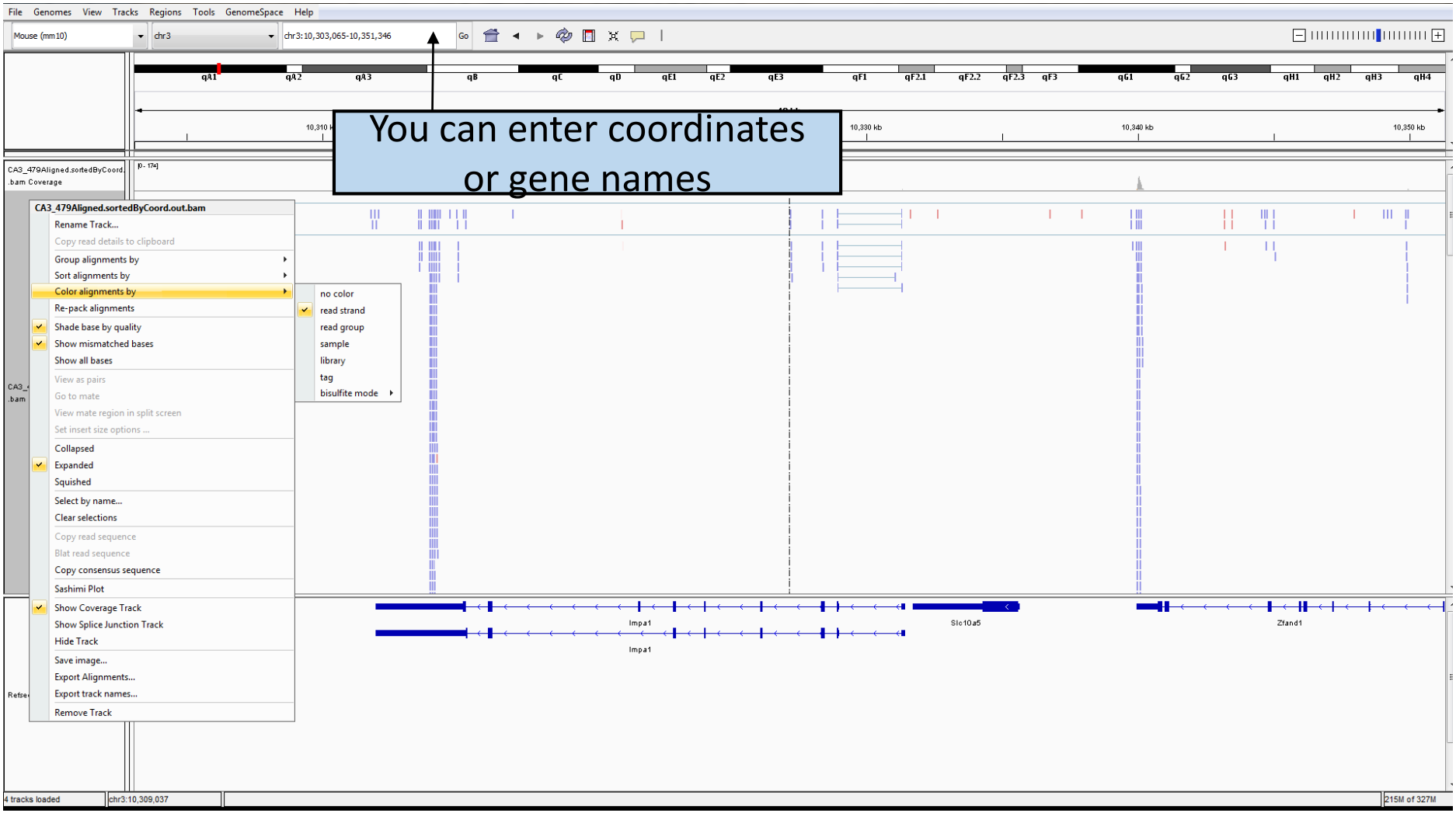
GFF style tags for BED files

The "name" field (column 4) of a BED file can contain GFF3 style key-value attribute tags by specifying "gffTags=on" on the track line. These attributes will be displayed in the mouse hover popup text.

Zooming in


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Mouse (mm10)
chr3
chr3
Go
qA1 qA2 qA3 qB qC qD qE1 qE2 qE3 qF1 qF2.1 qF2.2 qF3 qG1 qG2 qG3 qH1 qH3 qH4
159 mb
20 mb 40 mb 60 mb 80 mb 100 mb 120 mb 140 mb
CA3_479Aligned.sortedByCoord.
.bam Coverage
CA3_479Aligned.sortedByCoord.
.bam
Zoom in to see coverage.
Zoom in to see alignments.
Refseq genes
Hnf4g Pag1 Ythdf3 Naaladl2 Skil Qrfpr Pcdh10 Cog6 Selt Gmps Otol1 Fstl5 Ctso Fcrls Lor Gja8 Ngf Alx3 Amy1 Ptbp2 Arsj Dlk2 H2afz Lmo4 Adgrl4 Cth
Zfhx4 Ralyl Cp Nlgn1 Zfp639 Plk4 Elf2 Wwtr1 Plch1 Zbbx Gria2 Lrba Sprr3 Wdr3 Pifo Amy2b Rwdd3 Pibx2 Bdh2 Unc5c Dnajb4
Pex2 Ralyl Cp Ect2 Pex5l Plk4 Elf2 Rnf13 Kcnab1 Gria2 Lrba Rorc Trim45 Csf1 Rnpc3 Mir7657 Rpl34 Rap1gds1 Usp33
Pex2 Ralyl Hps3 Tnik Spry1 Pcdh10 Trpc4 Shox2 Trim2 Chtop Pde4dip St7l Amy2a5 Camk2d Bmpr1b Rabggtb
4 tracks
chr3:158,742,500
111M of 333M

Viewing reads aligned to genes

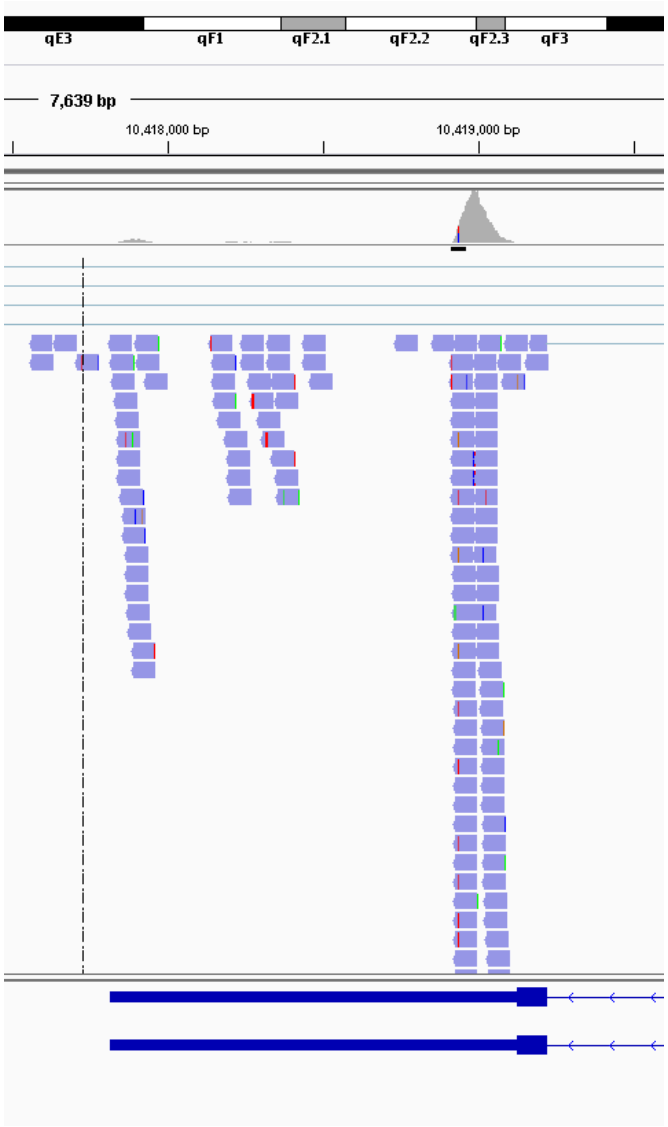


Gene model

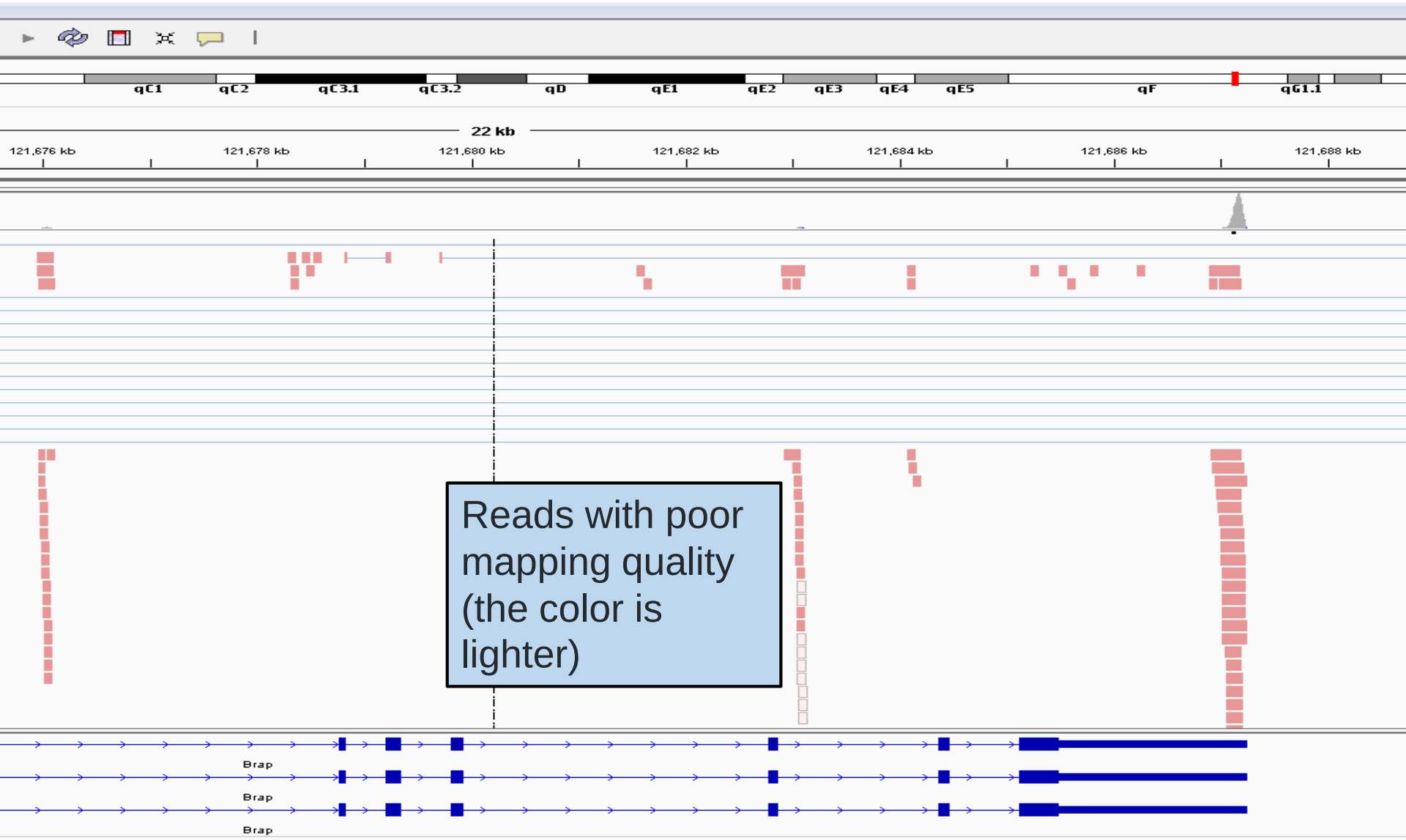
Genes are represented as lines and boxes. Lines represent intronic regions, and boxes represent exonic regions. The arrows indicate the direction/strand of transcription for the gene. When an exon box become narrower in height, this indicates a UTR.

Mismatches

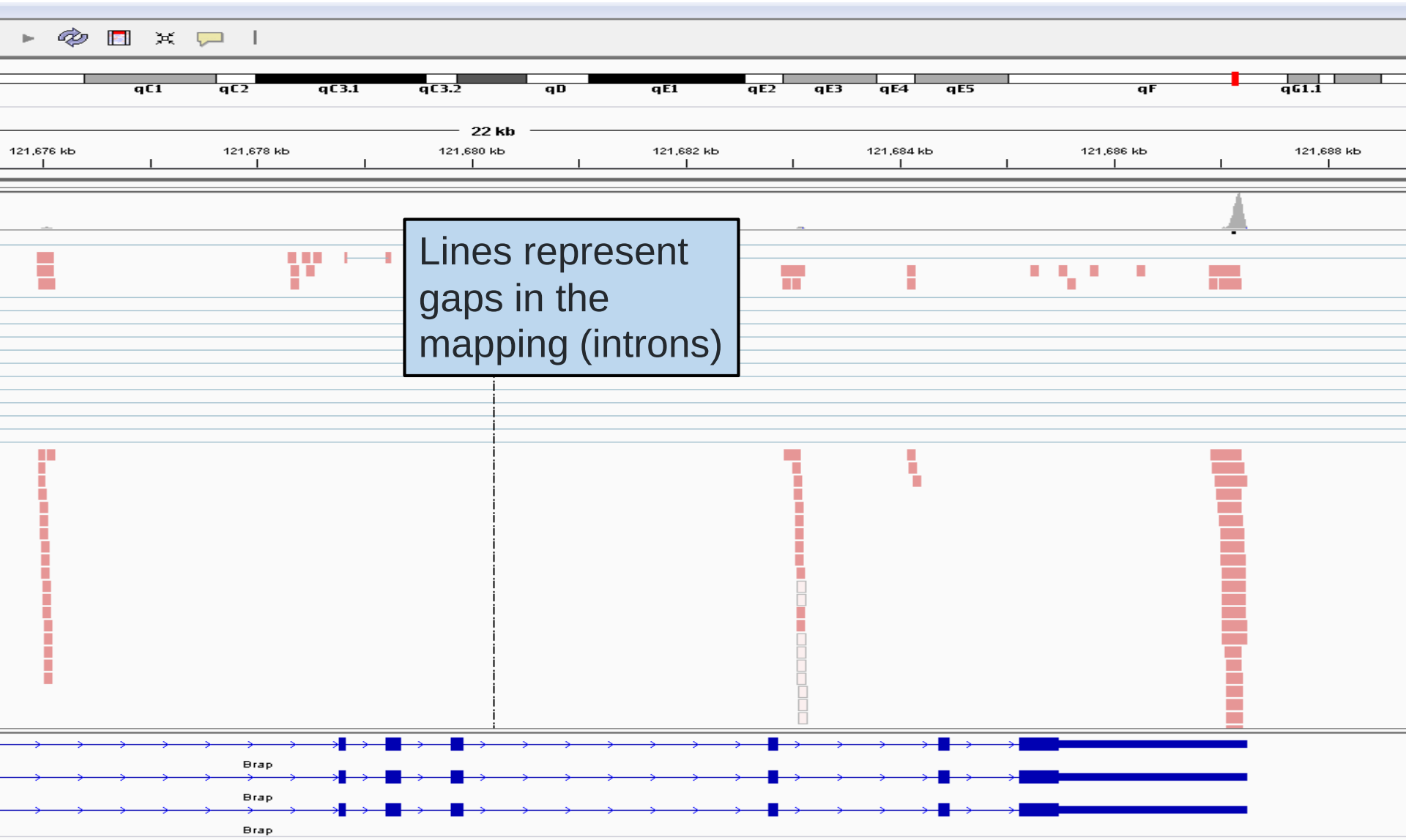
Bases that do not match the reference sequence are highlighted by color



Mapping quality



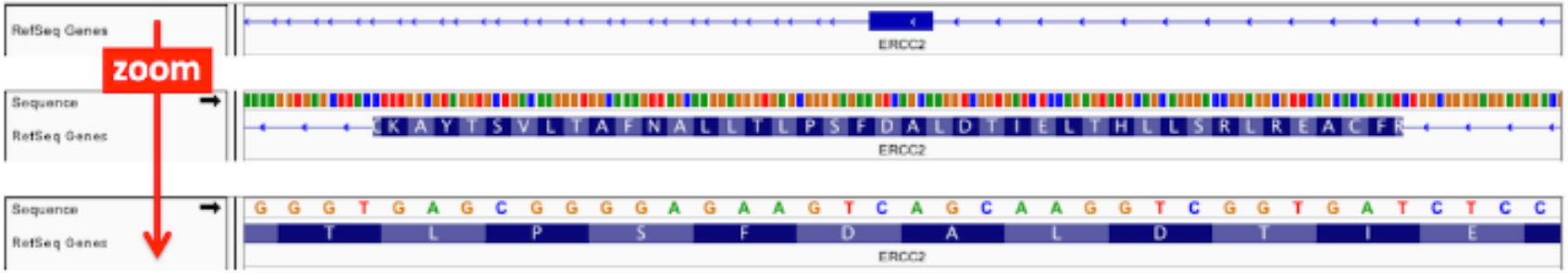
Spliced reads



Sequence Track Options

When zoomed in sufficiently, the reference genome *Sequence* track appears at the top of the lower panel above the *Genes* track, if any, in the IGV display as shown in the **Screenshot** (2015.04.01). The sequence is represented by colored bars or colored letters, depending on zoom level, with adenine in green, cytosine in blue, guanine in yellow, and thymine in red (**A**, **C**, **G**, **T**).

- To change this default nucleotide coloring scheme see the Modify the prefs.properties file page.



Flipping the Strand

You can change the strand that is displayed by clicking on the arrow in the title to the left of the track. Note that the sequence and the arrow are only displayed when zoomed in to a sufficiently small region.

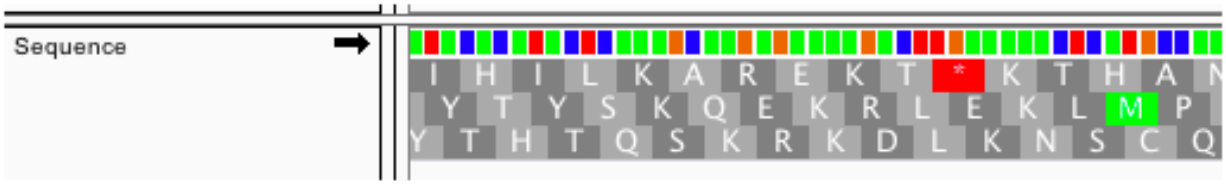
- Alternatively, right-click on *Sequence* track to select *Flip strand* from the pop-up menu.

The direction of the arrow indicates which strand is currently displayed. An arrow pointing left indicates that the negative strand is showing. This strand will show the complement nucleotides and reverse complement translations.

Sequence Translation

With the reference genome sequence track, you can optionally display a 3-band track that shows a 3-frame translation of the amino acid sequence for the corresponding nucleotide sequence. The translation is shown for the strand indicated.

- Right-click on *Sequence* track to select *Show translation* from the pop-up menu and to select a *Translation Table*.
- Selecting *Save image* from the right-click pop-up menu save the lower display panel containing the *Sequence* track as an image.



Amino acids are displayed as blocks colored in alternating shades of gray. Methionines are colored green, and all stop codons are colored red. When you zoom all the way in, the amino acid symbols will appear.

You can toggle the display of this translation track by clicking once, anywhere in the sequence or translation track, or by toggling *Show Translation* in the track popup menu.

Interpreting Color by Insert Size

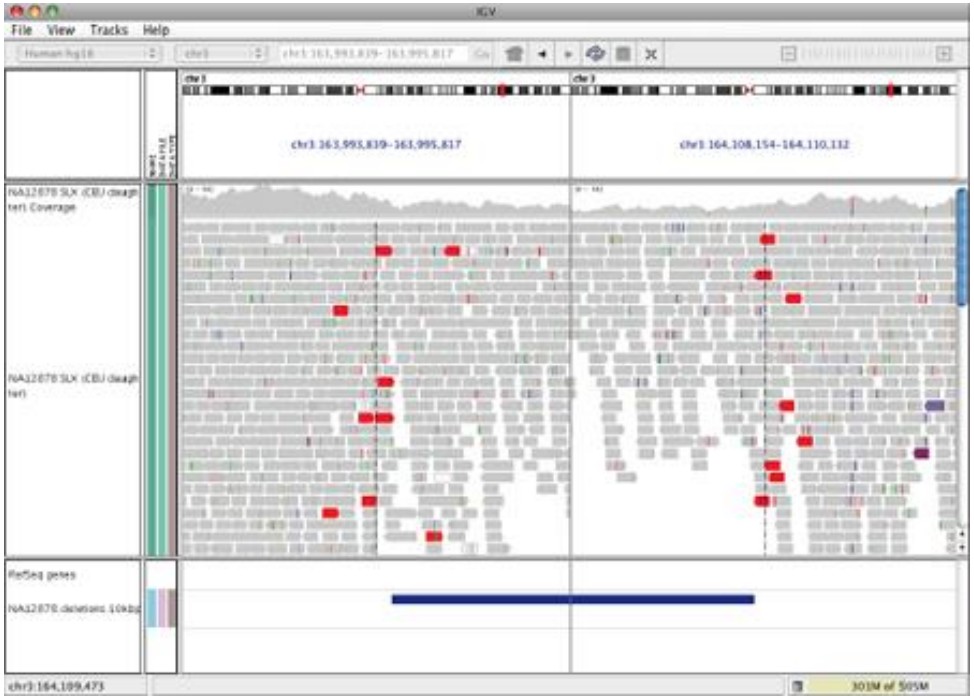
Coloring by insert size is for DNA alignments and is not designed to indicate RNA-Seq paired read mate distances. It is based on set base pair values or computed from the size distribution of a library against the reference genome as defined in the Alignment Preferences Panel.

The inferred insert size can be used to detect structural variants, such as:

- deletions
- insertions
- inter-chromosomal rearrangements

IGV uses color coding to flag anomalous insert sizes. When you select *Color alignments>by insert size* in the popup menu, the default coloring scheme is:

- for an inferred insert size that is larger than expected (possible evidence of a deletion)
- for an inferred insert size that is smaller than expected (possible evidence of an insertion)



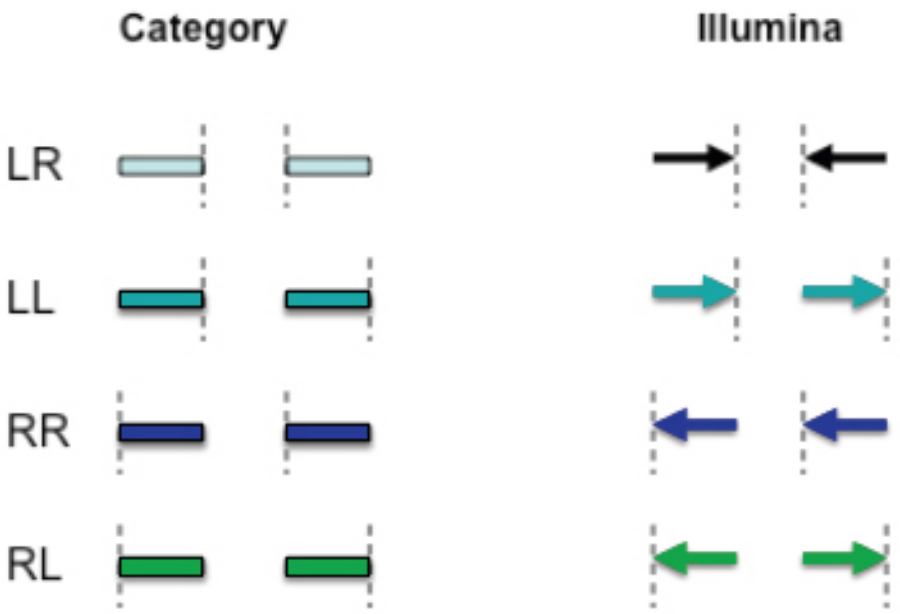
Interpreting Color by Pair Orientation

The orientation of paired reads can be used to detect structural events including:

- inversions
- duplications
- Translocations

By selecting *Color alignments>by pair orientation*, you can flag anomalous pair orientations in IGV.

Orientation is defined in terms of read-strand: left versus right, and first read versus second read of a pair.



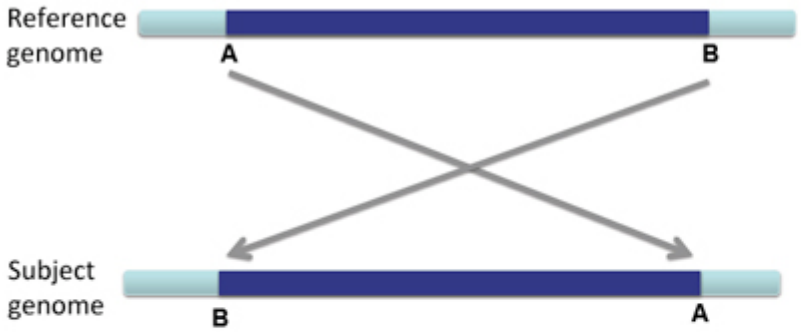
LR	Normal reads. The reads are left and right (respectively) of the unsequenced part of the sequenced DNA fragment when aligned back to the reference genome.
LL,RR	Implies inversion in sequenced DNA with respect to reference.
RL	Implies duplication or translocation with respect to reference.

(figure courtesy of Bob Handsaker)

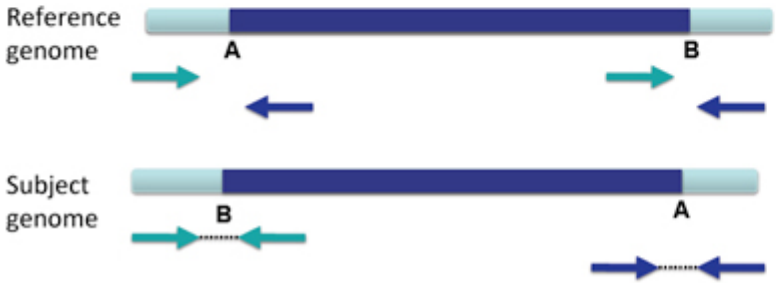
http://software.broadinstitute.org/software/igv/interpreting_pair_orientations

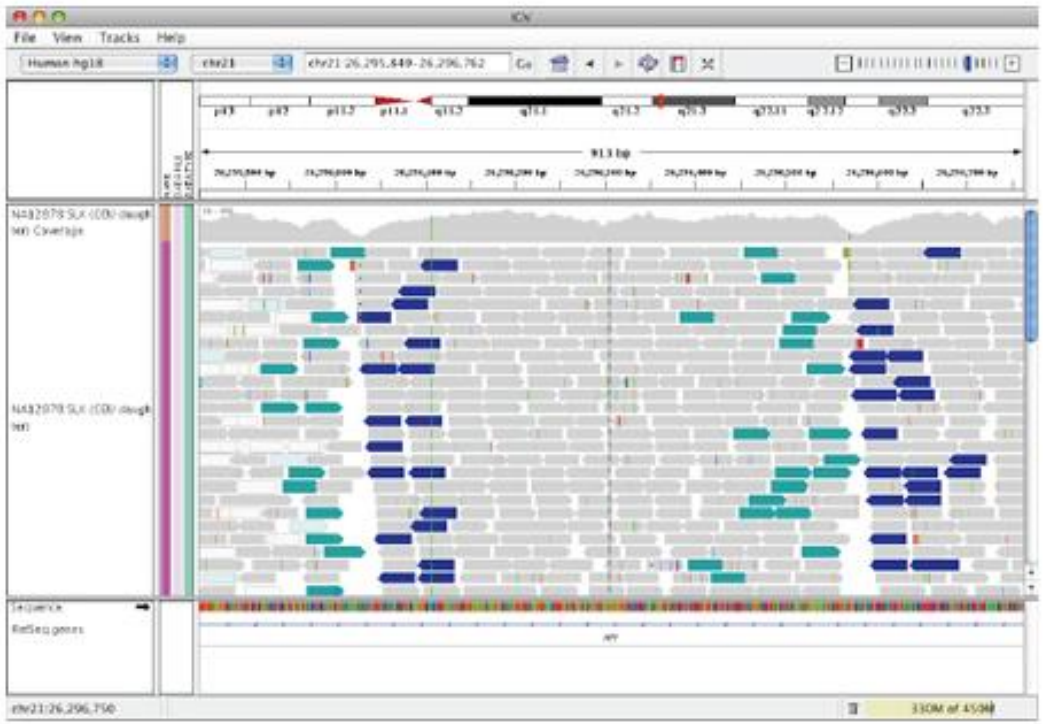
Inversions

An inversion is a large section of DNA that is reversed in the subject genome compared to the reference genome.



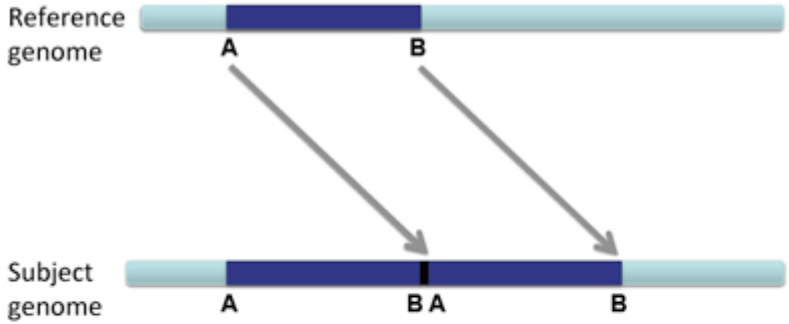
When an inversion shows up in paired-end reads, the reads are distinctively variant from the reference genome.





Tandem Duplication

When a large section of DNA is duplicated and inserted into the genome next to the original sequence, this is called a tandem duplication.



The reads will not only be duplicated, but also be arranged as shown below.