# Viewing AWS-Hosted Data Tracks in IGV, February 14, 2023

This document is a comprehensive description of the data tracks available on aws to polishers of the HG002 diploid assemblies (currently the v0.7 verkko/rukki assembly), as well as notes on how to use IGV to view them. *This document is a work in progress, and available to edit at https://docs.google.com/document/d/19jhy19crbqwewexQ0UoknsPXYEs_XjNI7GwCQO5TEns/ . Please feel free to suggest changes or additions as they occur to you.*

## CATEGORIES OF DATA TRACKS AND THEIR LOCATIONS

There are various types of data available for viewing in IGV using URLs hosted on the project's aws "human-pangenomics" S3 endpoint. Most of the available bam, bed, bigBed, and wig files for curating the v0.7 HG002 assemblies will be in subdirectories of https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/HG002/assemblies/polishing/HG002/v0.7/ (it may be helpful to bookmark this link). This aws prefix will be referred to in this document as "AWS_POL_PREFIX". The prefixes within AWS_POL_PREFIX are currently organized into a set of categories ("mapping", "wigfiles", "variants", "haplotypes"), which each have README files with up-to-date information about their contents. If there is no README file in a subprefix, or something is unclear about a particular section, feel free to post to our T2T #hg002 channel and tag @Nancy F. Hansen, and we'll try to add to or edit it as needed.

## ADDING DATA TRACKS TO IGV BY LOADING URLS

To display any of the data tracks described in this document in IGV, you can first copy the URL to the aws object by navigating to AWS_POL_PREFIX and following the appropriate links to it within your browser. Then within IGV, you can select "File…Load from URL", and paste the copied aws URL into the IGV popup.You can also copy URLs from this document or from the prepared session files (see next section).

## ADDING DATA TRACKS TO IGV USING SESSION FILES

An easier way to load groups of tracks into IGV is to make use of IGV session files. There is a useful set of session files in the HG002-issues github repository at https://github.com/marbl/HG002-issues/tree/main/igv_sessions/. In addition to being easier to load and grouped into useful categories, the tracks within the prepared session files are also given more descriptive names, which are displayed in the leftmost panel of IGV.

## RUNNING MULTIPLE INSTANCES OF IGV

Single instances of IGV can become very slow, especially if you are viewing a large genomic region or are loading bam files or other tracks with lots of data. One way to keep the program from slowing to a crawl is to run multiple instances of it. By launching IGV from the command line of your computer, it's possible to bring up one IGV window to view read alignments, for example, and another to view less intensive annotation tracks.

**ALIGNED READ TRACKS (READ BAM FILES)**

Tracks are available to display aligned reads from various platforms, binned by parental haplotype or not. They might be aligned to the entire v0.7 assembly, the maternal or paternal haplotype only, or to a "squashed" haplotype including one copy of each autosome + chrX, chrY, chrM, and chrEBV.

If the assembly is correct, accurate reads properly aligned should have sequence completely in agreement with the assembly, and in the absence of sequencing bias, read coverage should be uniformly random. If reads show uneven coverage or discrepancies with the consensus, it could indicate structural or consensus errors in the assembly, or it could be due to sequencing error and/or misalignment of the reads.

| **Name** | **File/URL** | **Platform/Caller** | **Aligner/Reference** |
|---|---|---|---|
| HiFi DCv1.1 primary | hg002v0.7_hifi_dcv1.1.pri.bam | HiFi DeepConsensus v1.1 | Winnowmap2/whole v0.7 assembly |
| ONT Guppy6.1.2 remora primary | hg002v0.7_ont_guppy_6.1.2_remora.pri.bam | ONT Guppy6.1.2 Remora | Winnowmap2/whole v0.7 assembly |
| Hifi DCv1.1 Linked Binned primary | hg002v0.7_hifi_dc1.1_mergebinned.pri.sort.rg.bam | HiFi DeepCons v1.1 Trio- and linked-marker-binned | Winnowmap2/corresponding haplotype |
| SSR | hg002v0.7matY_SSR.bam | 1x200, 40x coverage | bwa mem/maternal+Y+EBV |
| HiFi DCv1.1 all vs. maternal+Y+EBV | hg002v0.7.mat.Y.EBV_hifi_dc1.1.pri.bam | HiFi DeepConsensus v1.1 | Winnowmap2/maternal+Y+EBV |
| ONT Guppy6.1.2 remora all vs. maternal+Y+EBV | hg002v0.7.mat.Y.EBV_ont_guppy_6.1.2.pri.bam | ONT Guppy6.1.2 Remora | Winnowmap2/maternal+Y+EBV |
| 100X Element reads | HG002T2Tv0.7_HG002-element-PCR-free | Element Biosciences PCR WGS | Whole v0.7 assembly |

| | | | |
|---|---|---|---|
| | _2x150_100X.bam | | |

**VARIANT TRACKS (VCF FILES)**

Variant callers indicate places where the read data indicate a sequence different from the consensus. Depending on the quality of the calls and the read alignments used to generate them, these “variants” may indicate errors in the assembly, or just be false positives.

| Name | File | Platform/Caller | Aligner/Reference |
|---|---|---|---|
| Sniffles SV calls | hg002v0.7_hifi_dcv1.1.pri.bam | HiFi DeepConsensus v1.1 | Winnowmap2/whole v0.7 assembly |
| DeepVariant calls on all reads vs. both haplotypes | hg002v0.7_hifi_dcv1.1.DV_1.5.vcf.gz | HiFi DeepConsensus v1.1 | Winnowmap2/whole v0.7 assembly |

**TRACKS CALLED BY MERQURY AND THE T2T POLISH PIPELINE**

In addition to the trio- and linked- haplotype-binned alignments above, merqury and the T2T-Polish pipeline output many bed-formatted files highlighting issues in the assembly.

| Name | File/URL | Description |
|---|---|---|
| HiFi Pri Coverage | hg002v0.7_hifi_dcv1.1.pri.cov.wig | HiFi DeepConsensus v1.1 read coverage |
| HiFi Pri Issues | hg002v0.7_hifi_dcv1.1.pri.issues.bed | Regions with anomalous HiFi read coverage |
| ONT Pri Coverage | hg002v0.7_ont_guppy_remora.pri.cov.wig | ONT Guppy 6.1.2/Remora read coverage |
| ONT Pri Issues | hg002v0.7_ont_guppy_remora.pri.issues.bed | Regions with anomalous ONT read coverage |
| Error kmers | hg002v0.7_k21_hybrid_error.bed | Locations of consensus kmers not present in HiFi/Illumina reads |
| Linked Hapmer-based Switches | v0.7_illumina_ext2.v0.7.both.100_20000.phased_block.switch.bed | Locations of linked hapmers in stretches of wrong parent’s haplotype |

## HAPLOTYPE COMPARISON TRACKS (BAM AND BIGBED FILES)

To give polishers a sense of what the other parental haplotype looks like for the region of the assembly they are examining, the two haplotypes have been aligned to each other with Winnowmap2 and Nucmer, and tracks are available with the BAM files. Because the alignments in the BAM files are too long to determine the coordinate of the alternate haplotype location in the middle of an alignment, a windowed bigBed file is also available to give the coordinates of the other haplotype that aligned to your location.

In addition to comparing each chromosome in the current assembly to its alternate haplotype, the assembly graph nodes have also been aligned to the assembly, and are available in a bed file so that regions can easily be viewed in Bandage using available gfa files.

| **Name** | **File** | **Aligner** |
|---|---|---|
| Alt hap nucmer alignment | hg002v0.7.haplotypemapping.nucmer.bam | Nucmer |
| Alt hap WM2 alignment | hg002v0.7.haplotypemapping.pri.wm.bam | Winnowmap2 |
| Alt hap nucmer coordinates | hg002v0.7.haplotypemapping.nucmer.withsimscores.bb | Nucmer |
| Alt hap WM2 coordinates | hg002v0.7.haplotypemapping.pri.wm.withsimscores.bb | Winnowmap2 |
| Assembly graph nodes | v0.7_combined_graph_nodes.corr.bed | Mashmap on HPC coords, then lifted to uncompressed |

## TRACKS WITH INFORMATION FROM OTHER PLATFORMS

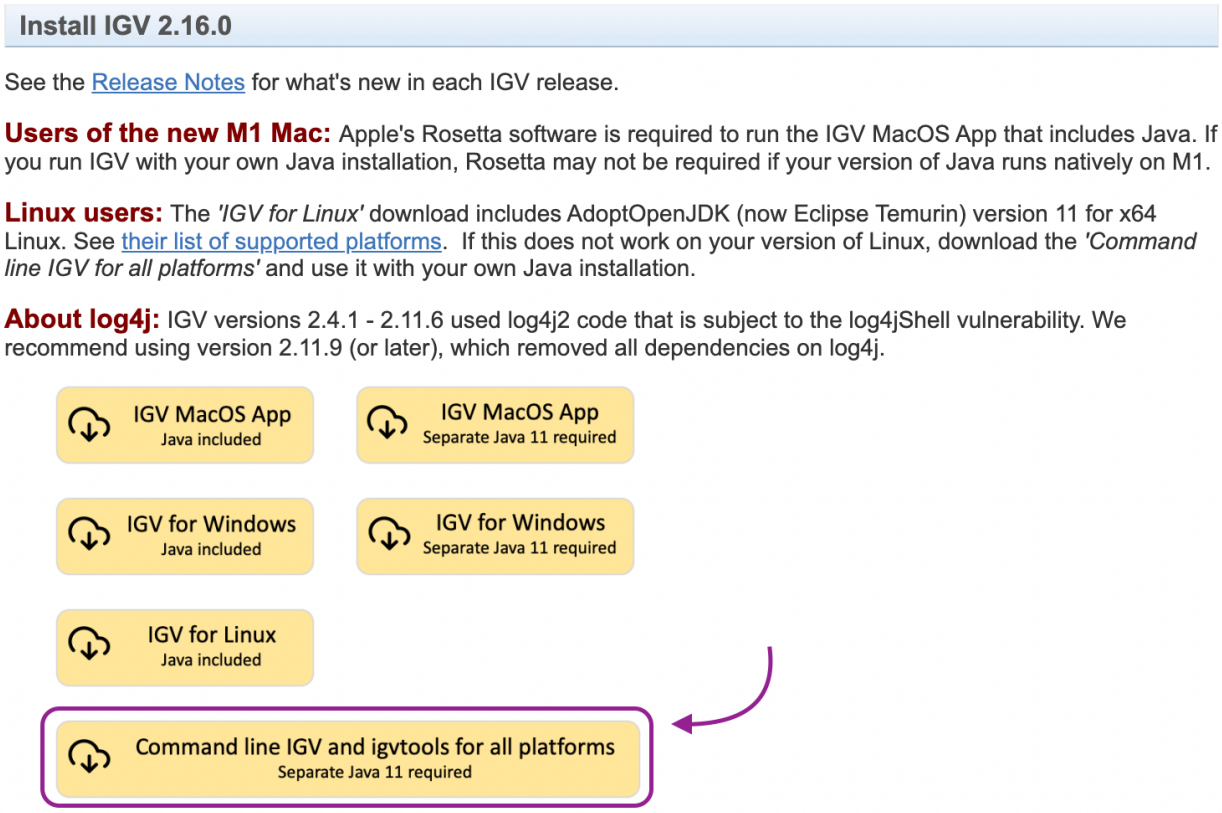
Tracks from platforms like strand-seq can give supplemental information that can help to determine the source of assembly issues.

| **Name** | **File** | **Platform** |
|---|---|---|
| Strand-seq wrong | hg002.v0.7.mat.strandseq.sort.bb | Strand-seq vs. |

| | | |
|---|---|---|
| strand calls | | mat-Y-EBV, Peter Ebert |
| Strand-seq regions with wrong strand | hg002.v0.7.mat.strandseq.cont_1000.bed | Strand-seq + bedtools |

### IGV Downloading and Settings

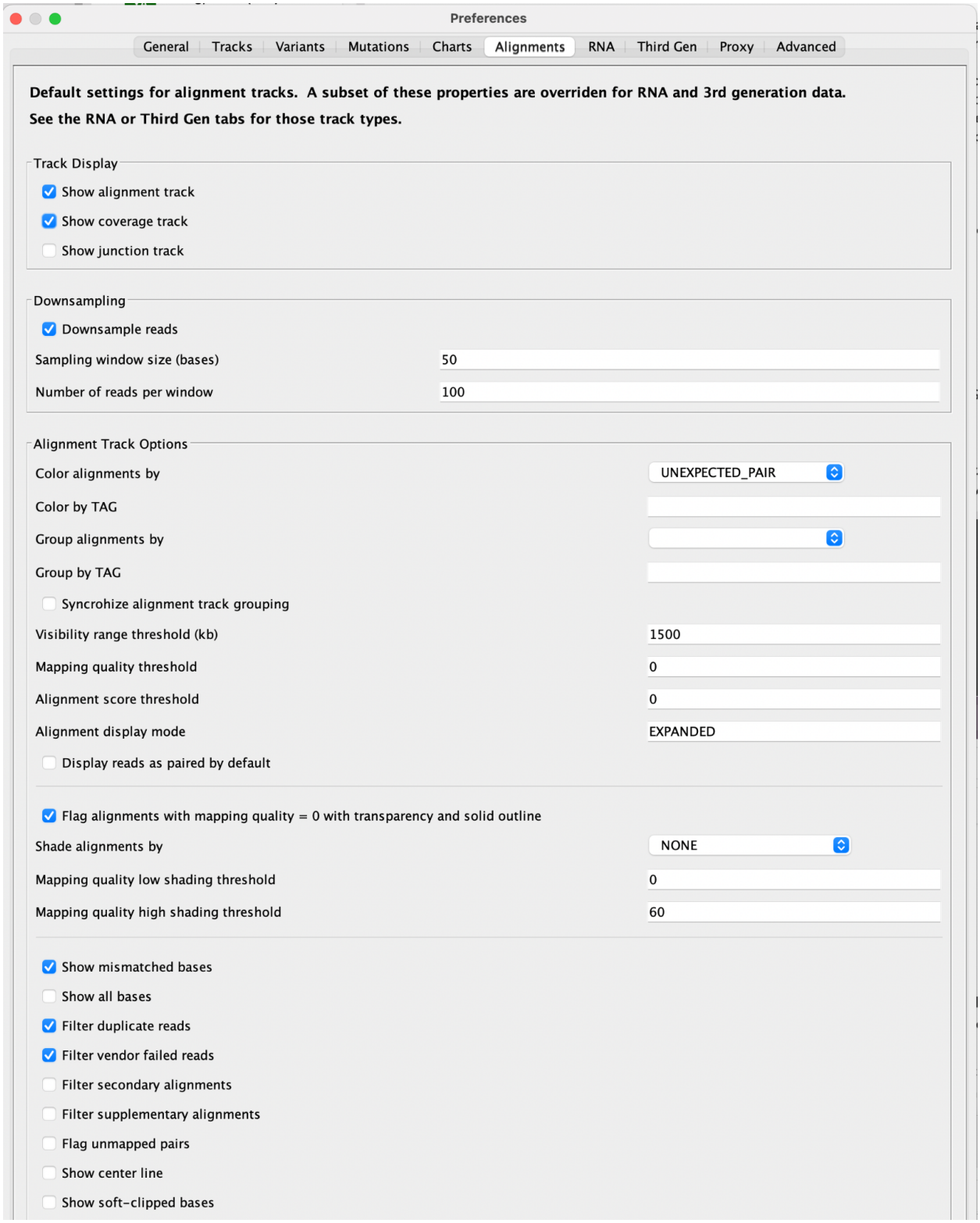
1. Download the “Command line IGV and igvtools for all platforms”



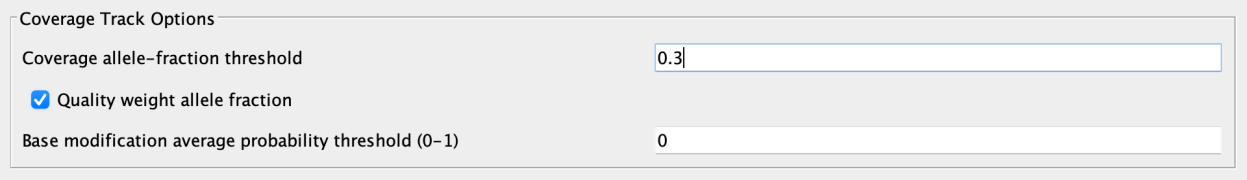

2. Edit igv.sh “-Xmx8g” part: -Xmx8g indicates 8 gb of memory (this is the default).
   There are 2 places in IGV 2.16.0 where you can edit.
   As we are usually opening 2 IGVs, depending on your laptop memory, two 8G apps might overkill and cause some strange issues.

```
#-Xmx8g indicates 8 gb of memory.
#To adjust this (or other Java options), edit the "$HOME/.igv/java_arguments"
#file.  For more info, see the README at
#https://raw.githubusercontent.com/igvteam/igv/master/scripts/readme.txt
#Add the flag -Ddevelopment = true to use features still in development
#Add the flag -Dsun.java2d.uiScale=2 for HiDPI displays
prefix=`dirname $(readlink -f $0 || echo $0)`

# Check whether or not to use the bundled JDK
if [ -d "${prefix}/jdk-11" ]; then
    echo echo "Using bundled JDK."
    JAVA_HOME="${prefix}/jdk-11"
    PATH=$JAVA_HOME/bin:$PATH
else
    echo "Using system JDK."
fi

# Check if there is a user-specified Java arguments file
if [ -e "$HOME/.igv/java_arguments" ]; then
    java -showversion --module-path="${prefix}/lib" -Xmx4g \
        @"${prefix}/igv.args" \
        -Dapple.laf.useScreenMenuBar=true \
        -Djava.net.preferIPv4Stack=true \
        -Djava.net.useSystemProxies=true \
        @"$HOME/.igv/java_arguments" \
        --module=org.igv/org.broad.igv.ui.Main "$@"
else
    java -showversion --module-path="${prefix}/lib" -Xmx4g \
        @"${prefix}/igv.args" \
        -Dapple.laf.useScreenMenuBar=true \
        -Djava.net.preferIPv4Stack=true \
        -Djava.net.useSystemProxies=true \
        --module=org.igv/org.broad.igv.ui.Main "$@"
fi
~
```

3. Launch IGV from terminal where it is unzipped: ./igv.sh
   This way you can see the logs in case IGV is just loading or an error occurred.
4. IGV Settings under **View** > **Preferences**
   Under [Alignments]

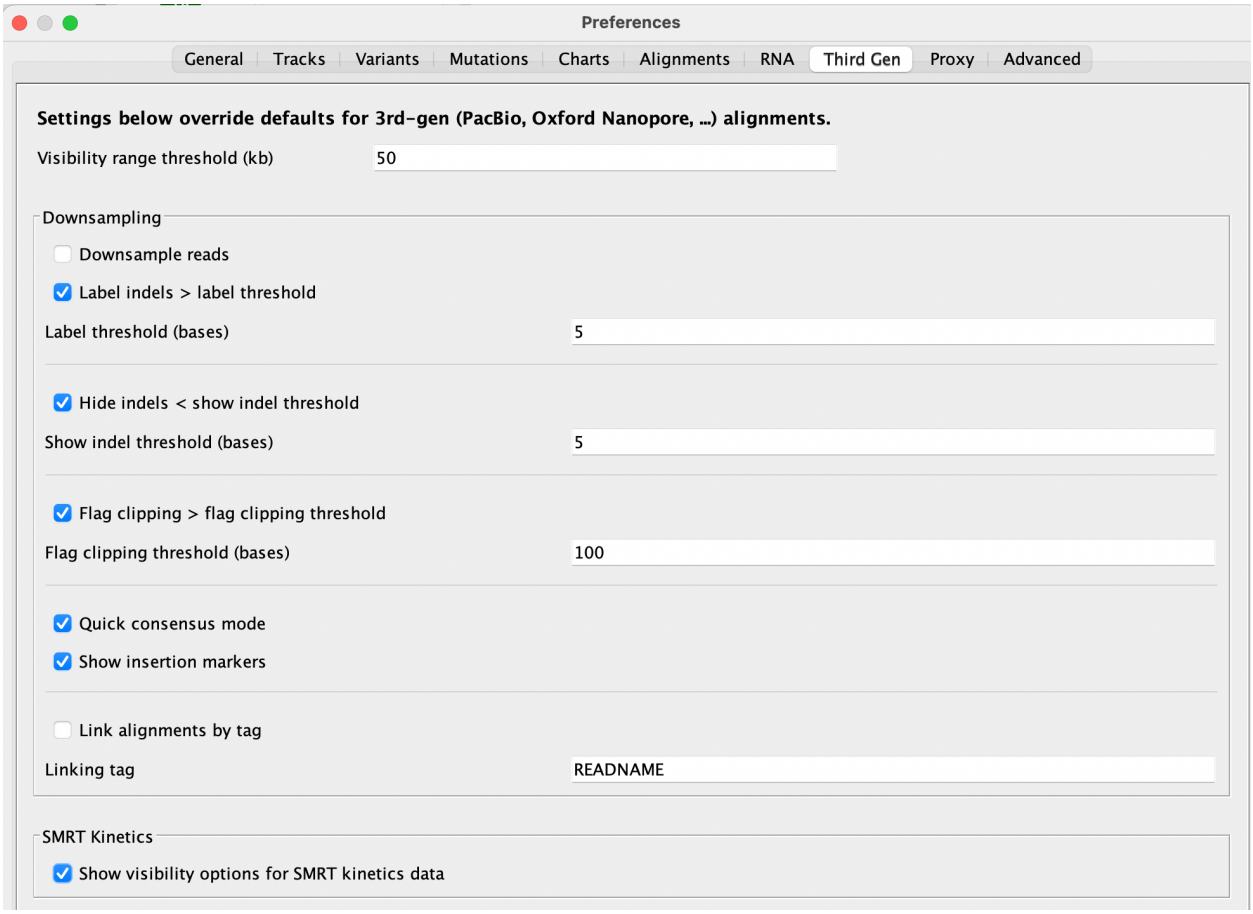

Preferences
General
Tracks
Variants
Mutations
Charts
Alignments
RNA
Third Gen
Proxy
Advanced
Default settings for alignment tracks. A subset of these properties are overriden for RNA and 3rd generation data.
See the RNA or Third Gen tabs for those track types.
Track Display
Show alignment track
Show coverage track
Show junction track
Downsampling
Downsample reads
Sampling window size (bases)
50
Number of reads per window
100
Alignment Track Options
Color alignments by
UNEXPECTED_PAIR
Color by TAG
Group alignments by
Group by TAG
Syncrohize alignment track grouping
Visibility range threshold (kb)
1500
Mapping quality threshold
0
Alignment score threshold
0
Alignment display mode
EXPANDED
Display reads as paired by default
Flag alignments with mapping quality = 0 with transparency and solid outline
Shade alignments by
NONE
Mapping quality low shading threshold
0
Mapping quality high shading threshold
60
Show mismatched bases
Show all bases
Filter duplicate reads
Filter vendor failed reads
Filter secondary alignments
Filter supplementary alignments
Flag unmapped pairs
Show center line
Show soft-clipped bases

...



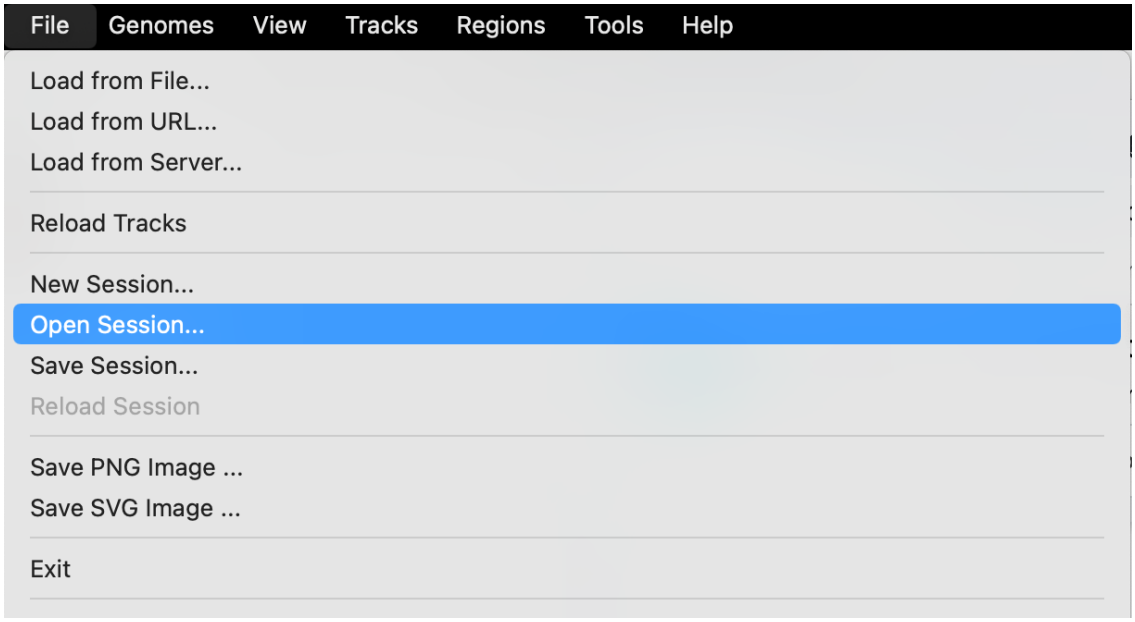

- This option is largely applied to non-bam views
- Uncheck the “Filter supplementary alignments”
- Adjust “Visibility range threshold (kb): 0 will display all ranges
- Set “Coverage allele-fraction threshold” to 0.3

Under [Third Gen]




- This option is applied to long-read bams
- Set “Visibility range threshold (kb) to something smaller than 100 kb (I use 60 kb)
- Depending on the variant, adjust indel or clipping base threshold to show or hide

**Using provided session files**

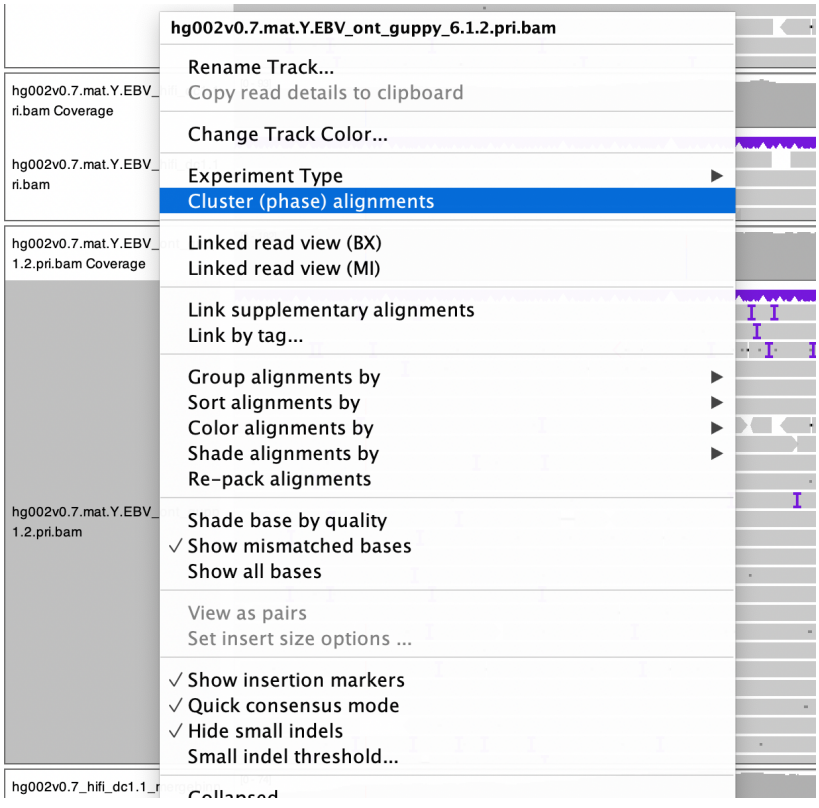



Some predefined session files (.xml) are available on https://github.com/marbl/HG002-issues/tree/main/igv_sessions.

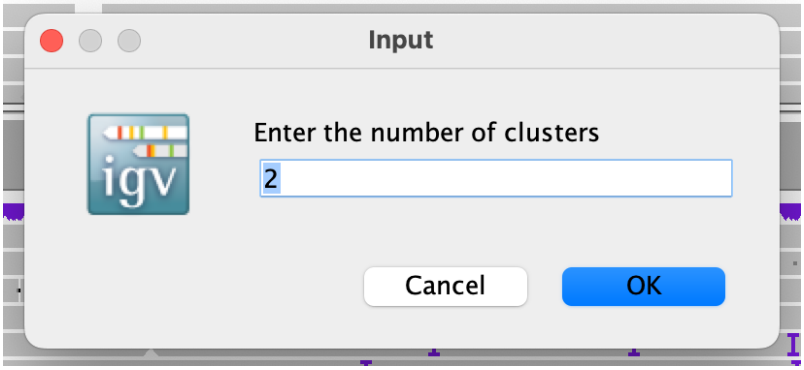
- **hg002v0.7_phase_switch_view.xml**: Overview. Contains coverage wiggles + markers, and bed formatted annotations. The only bam here is ALT haplotype. Suitable for zooming out over 100 kb or more.
- …

**Using the "Phase" feature**

It is possible to phase bam alignments on-the-fly on IGV, as long as the region contains at least 1 heterozygous variant visible on the coverage track.

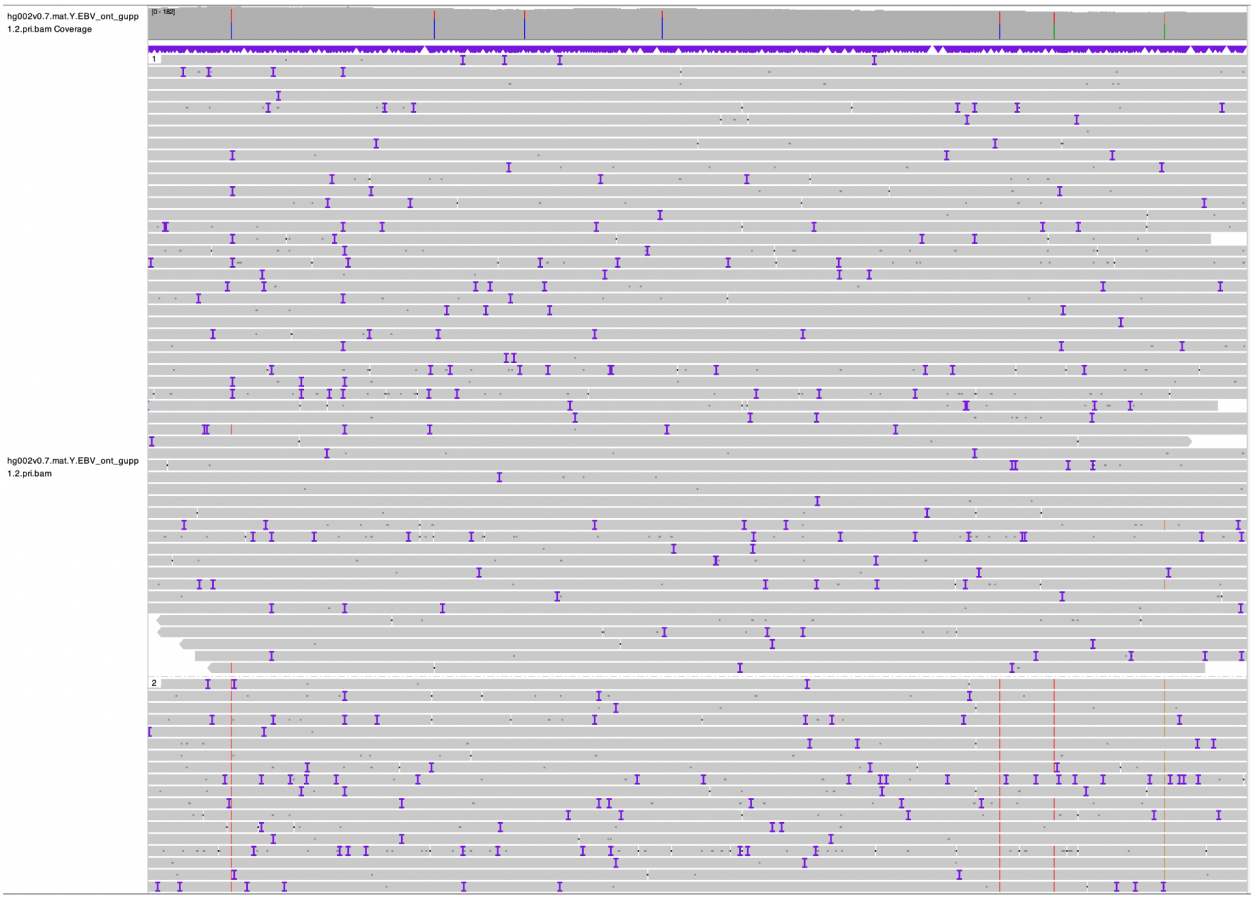



Right-click on the bam read pileup name panel
Select "Cluster (phase) alignments"




Setting it as 2 will cluster the reads to 1, 2, and “None”

This is how it looks after “Clustering”



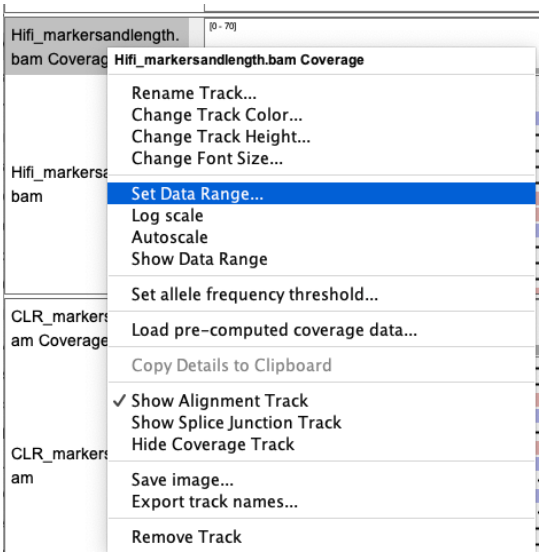

Setting the cluster to 0 will bin the reads back to group “None”.

**SCREENSHOTS**

Please take a screenshot of each issue region. Navigate to the region, then zoom out until at least one single-copy marker k-mer is within view, up until ~100kbp. Then take a screenshot, and upload to the assigned issue.

You may add tracks that are not in your session file by selecting “File…Load From URL” and entering any of the URLs in the tables in this document .

### Format bam Coverage tracks




1) Set Data Range
    a) HiFi: Max to 100
    b) ONT: Max to 200
2) Set allele frequency threshold… to 0.3