Viewing AWS-Hosted Data Tracks in IGV, February 10, 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_2x150_100X.bam	Element Biosciences PCR WGS	Whole v0.7 assembly

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	To come	HiFi DeepConsensus v1.1, ONT guppy6.1.2 remora	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 strand calls	hg002.v0.7.mat.strandseq.sort.bb	Strand-seq vs. mat-Y-EBV, Peter Ebert
Strand-seq regions with wrong strand	hg002.v0.7.mat.strandseq.gt1000.ws.bed	Strand-seq + bedtools

SCREENSHOTS

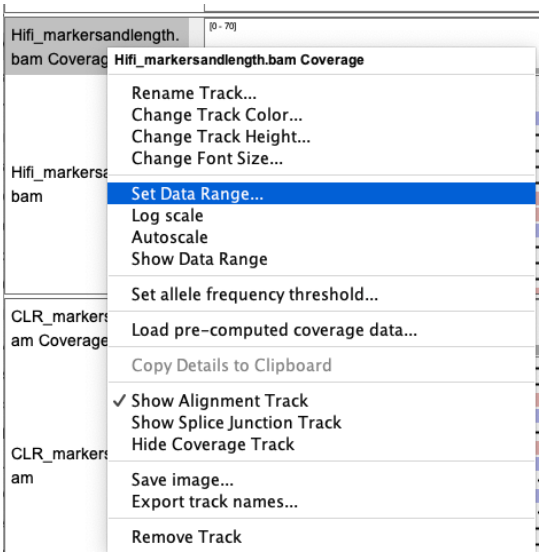
Please take a screenshot of each issue region. Navigate to the region, then zoom out until at least one single-copy marker k-mer is within view, up until ~100kbp. Then take a screenshot, save in .png or .pdf with SV ID as file name (e.g. DEL006957SUR.png), and upload to folder sv_validation under project “T2T collaboration 0.9 to 1.0” on DNAnexus.

REQUIRED TRACKS

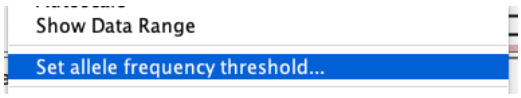
The following tracks will not be a part of the IGV session file, but you should download them and display them locally.

You may download files outside the session file using wget on files below here.

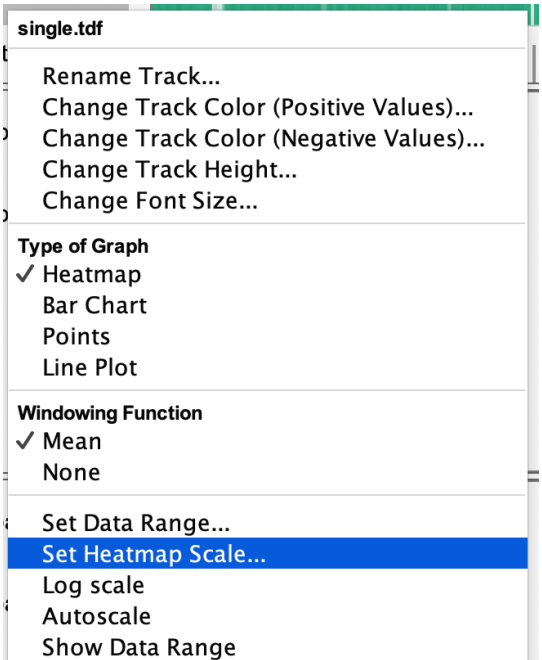
Format bam Coverage tracks

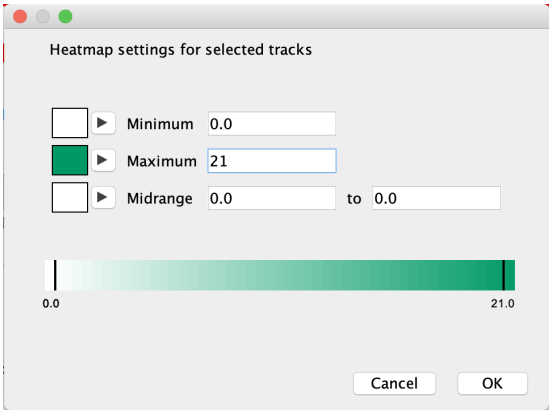


1) Set Data Range
    a) HiFi: Max to 100
    b) ONT: Max to 180
2) Set allele frequency threshold… to 0.3

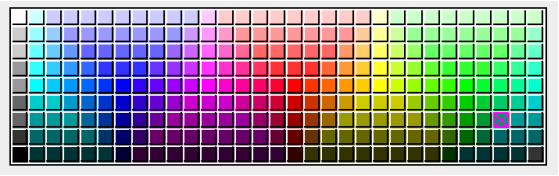


Format single.tdf track

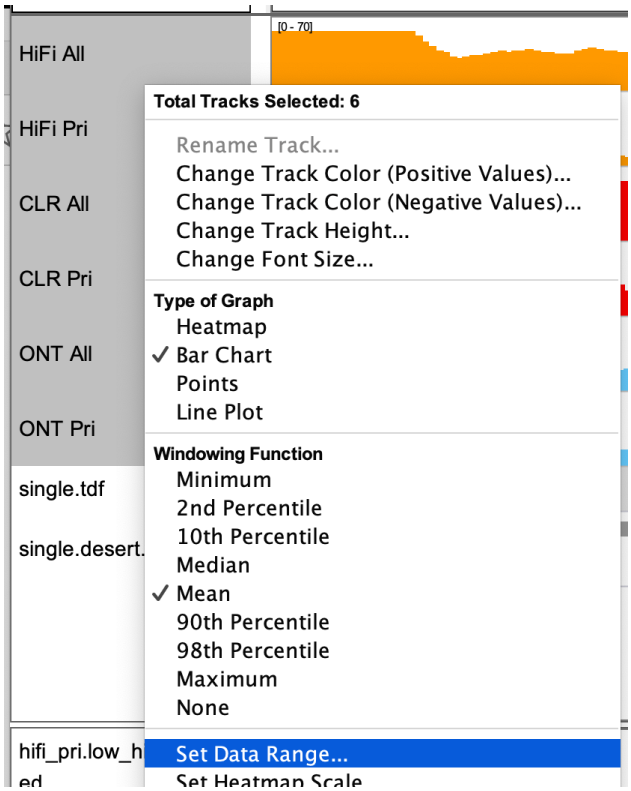


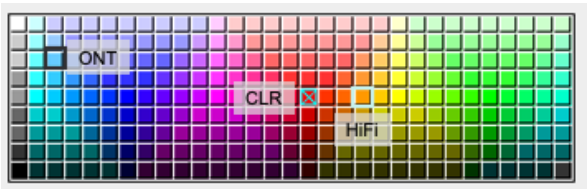


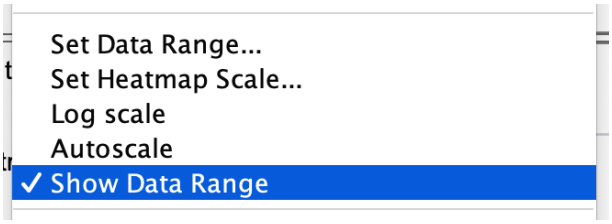
1) Select “Heatmap” as Type of Graph
2) Set Heatmap Scale
3) Set Maximum to 21
4) Select color and hit OK
5) Change Track Height to 30



Format .wig tracks

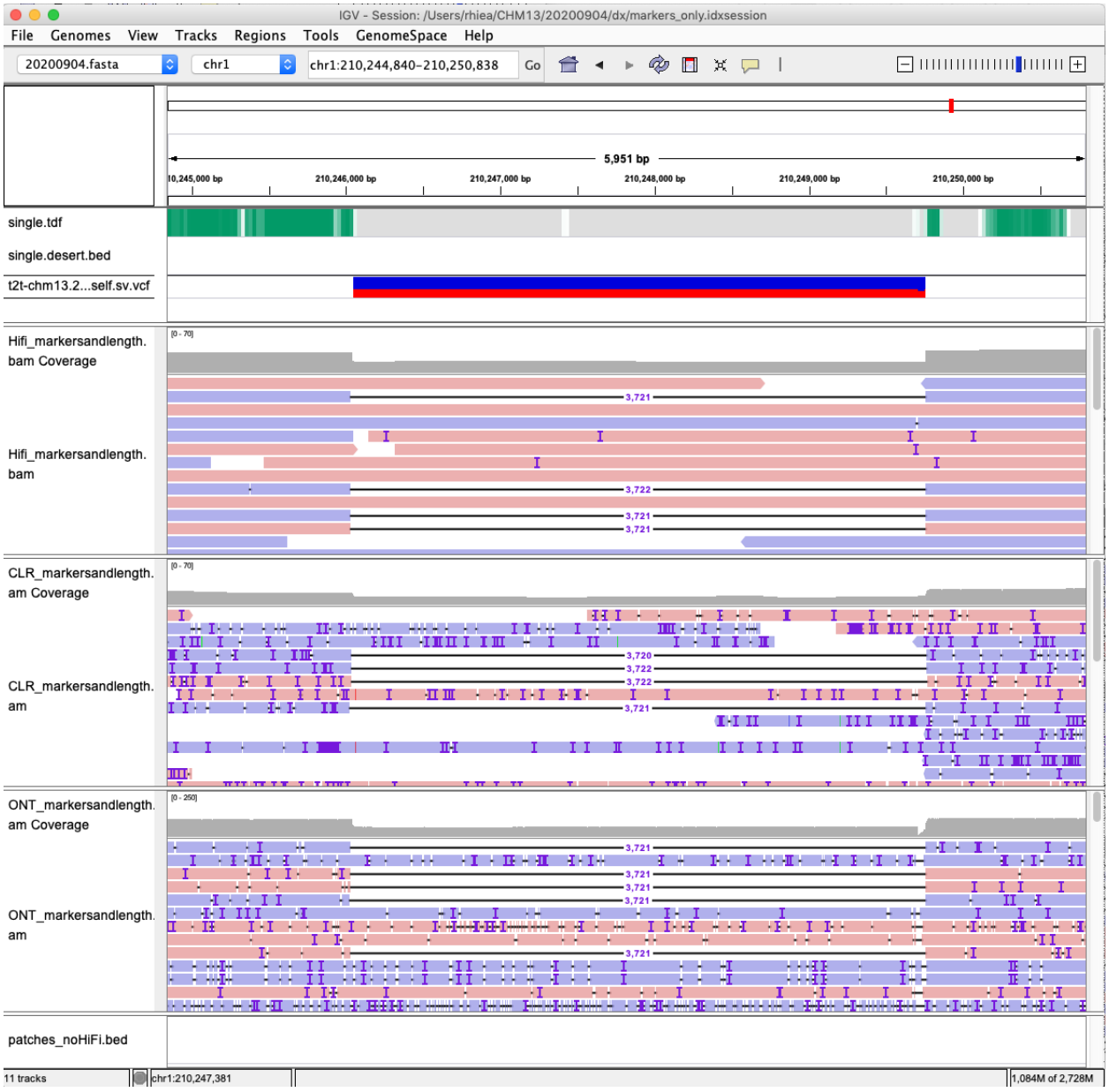






1) Set Data Range
    a) HiFi: Max to 100
    b) ONT: Max to 180
2) Change Track Color
    a) HiFi: 255, 153, 0
    b) CLR: 255, 0, 0
    c) ONT: 102, 204, 255
3) Show Data Range

Example Screenshot



Screenshot tip on MAC

- Ctrl + Shift + 4, select IGV area. Open the preview, hit Done.
- Rename the .png file generated on the desktop