# Viewing AWS-Hosted Data Tracks in IGV, March 7, 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 |
| ONT R10 duplex reads | hg002v0.7_ont_r10_duplex.pri.bam | ONT R10 duplex | Winnowmap2/whole v0.7 assembly |
| 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 |

| | | | |
|---|---|---|---|
| ONT Guppy6.1.2 remora phased with whatshap | hg002v0.7_ont_guppy_6.1.2_remora.pri.pmdv_wh.phased.bam | ONT Guppy6.1.2 Remora with phasing in read tags | Winnowmap2 + whatshap |
| ONT R10 duplex reads phased | hg002v0.7pat_ont_r10_duplex_dorando_0.11.1_splitduplex.pri.phased.haplotagged.bam | ONT R10 duplex UL reads with phasing in read tags | |
| HiFi DCv1.1 veritymap alignments | hg002v0.7_hifi_veritymap.merged.bam | HiFi DeepConsensus v1.1 | Veritymap to entire 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** |
|---|---|---|---|
| DeepVariant calls on all **HiFi** DCv1.1 reads vs. **single** haplotypes | hg002v0.7_hifi_dcv1.1.DV_1.5.vcf.gz | HiFi DeepConsensus v1.1, DeepVariant | Winnowmap2/whole v0.7 assembly |
| DeepVariant calls on all **ONT R10 simplex** reads vs. **single** haplotypes | hg002v0.7_matpat_r10_simplex_DV_1.5.vcf.gz | ONT R10 simplex DeepVariant v1.5 | Winnowmap2/maternal or paternal assembly |
| DeepVariant calls on all **ONT R10 duplex** reads vs. **single** haplotypes | hg002v0.7_mat_only_r10_duplex_dorado_DV_1.5.vcf.gz | ONT R10 duplex DeepVariant v1.5 | Winnowmap2/maternal or paternal assembly |
| DeepVariant calls on all **HiFi** DCv1.1 reads vs. **both** haplotypes | hg002v0.7_hifi_dcv1.1.DV_1.5.vcf.gz | HiFi DeepConsensus v1.1, DeepVariant | Winnowmap2/whole v0.7 assembly |

| | | | |
|---|---|---|---|
| DeepVariant calls on all **ONT R9** reads vs. **both** haplotypes | hg002v0.7_ont_r9_pmdv_0.8.vcf.gz | ONT Guppy6.1.2 DeepVariant | Winnowmap2/whole v0.7 assembly |
| DeepVariant calls on all **ONT R10 simplex** reads vs. **both** haplotypes | hg002v0.7_ont_r9_pmdv_0.8.vcf.gz | ONT R10 simplex DeepVariant v1.5 | Winnowmap2/whole v0.7 assembly |
| **Element** PCR free 2x150 100x reads to **single** haplotypes | hg002v0.7_mat_element_PCR_free_2x150_100X_DV_1.5.vcf.gz | Element 2x150, DeepVariant v1.5 | Maternal or paternal v0.7 assembly |
| **Intersection** of HiFi, R10 duplex, and R10 simplex DeepVariant calls | hg002v0.7_mat_only_hifi_DC1.1_DV_1.5.duplex_simplex_tp.pass_only.vcf.gz | Various, DeepVariant v1.5 | Winnowmap2, maternal or paternal 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 |
| Marker | hg002v0.7.k21.marker.bw | Locations of 21-mers that are 1-copy k-mers (they will be either correctly placed hap-mers or uncategorized k-mers) |
| MAT | hg002v0.7_k21_mat_hapmer.bw | Locations of maternal hap-mers (if on maternal chromosomes, correctly placed, or if on paternal chromosomes, incorrectly placed) |
| PAT | hg002v0.7_k21_pat_hapmer.bw | Locations of paternal hap-mers (correctly placed, if on paternal chromosomes) |

| | | |
|---|---|---|
| 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.cont_1000.bed | Strand-seq + bedtools |

## ANNOTATION TRACKS FOR THE ASSEMBLY

Wherever possible, we'll try to upload various annotation tracks to aws so they can be viewed as well. Right now, only the locations of microsatellite tracts are available, but in the future we may add more, so check back.

| Name | File | Description |
|---|---|---|
| Microsat. AT, GA, GC, TC | v0.7.microsatellite.AT.128.bw<br>v0.7.microsatellite.GA.128.bw<br>v0.7.microsatellite.GC.128.bw<br>v0.7.microsatellite.TC.128.bw | Microsatellite tract locations |

IGV tips and tricks for evaluation can be found at Evaluating with IGV.