Coverage Anomaly Polishing Instructions, February 12, 2023

This document is a guide to curating read coverage anomalies in HG002 diploid assemblies (currently the v0.7 verkko/rukki assembly), and describes the methods that we use to determine whether these anomalies are due to technical issues like sequencing bias (e.g., in HiFi regions with low coverage due to high GA/CT content) or actual errors in the assembly (e.g., duplicated or collapsed regions in the consensus on one or both alleles).

This document is a work in progress, and available to edit at https://docs.google.com/document/d/1lef3T5wdFdw4_m8oxCNkvRicckC8qZfoCbrM6Hjz9z8/edit . Please feel free to suggest changes or additions as they occur to you.

HOW ARE READ COVERAGE ANOMALIES DETECTED?

Our current set of coverage "issues" (as of February, 2023) was created by running the T2T-Polish workflow (https://github.com/arangrhie/T2T-Polish/tree/master/coverage) on Winnowmap2 primary alignments of ONT and HiFi reads to the v0.7 HG002 assembly. This run tagged 352 regions, 10 of which heavily overlap the rDNA regions of the acrocentric p-arms. The 352 T2T-Polish regions have been added to the issue section of our github repository (https://github.com/marbl/HG002-issues/issues), and various labels have been added to annotate them. For each issue, the status (open/closed), the people assigned for curation, and the labels are mirrored in a Google spreadsheet visible at: https://docs.google.com/spreadsheets/d/1eRpT0fXYmODoA2A9YYK4Z3CR_o6NTJBJnYHOem5LRjo. *Note: this google spreadsheet will be updated regularly FROM the github site, so any changes made to columns that mirror github will be overwritten!*

In addition to the T2T-Polish calls, Mobin Asri has also provided us with Flagger (Liao et al., 2022) calls for the v0.7 assembly chromosomes, which mark regions that appear to be duplications, collapses, or errors. Nancy Hansen has compared these calls to the T2T-Polish regions with bedtools. Since Flagger predicts a large number of regions/bases to be erroneous, the polishing group's first round of curation will focus only on the intersection of Flagger's "ALT-removed" calls on the HiFi reads with ALT-removed calls from the ONT data (a.k.a., "flagger intersect" calls). On github, "flagger_intersect" labels have been added to all T2T-Polish coverage issues which intersect with these Flagger calls, but once we have a better handle on which Flagger calls are likely to be correct, we'll likely import regions that are called only by Flagger to github as well.

VIEWING ASSIGNED ISSUES

If you have volunteered to curate assembly issues and we have your github username, a number of issues will have been assigned to you for curation. You can view them at

https://github.com/marbl/HG002-issues/issues (assuming you are logged in to github) by using the “Assignee” drop down menu to select yourself.

COVERAGE LABELS

The labels (in colored bubbles) attached to issues on the github site give context to why each particular region was reported as an issue. For coverage issues, the table below gives descriptions:

Label	Meaning	Source
coverage_pri	Anomaly in long read coverage for reads aligned to both haplotypes	T2T-Polish pipeline
low_cov_hifi, low_cov_ont	Lower than expected coverage of HiFi or ONT reads, respectively	T2T-Polish pipeline
high_cov_hifi, high_cov_ont	Higher than expected coverage of HiFi or ONT reads, respectively	T2T-Polish pipeline
hsat2/hsat3	Region has been annotated as containing HSat sequence	Julian Lucas
alpha_sat	Region has been annotated as containing alpha satellite sequence	Julian Lucas
ga_tc, gc, at	Indicates high content of the label’s nucleotides	T2T-Polish pipeline
error_kmer	Consensus contains kmers that are not observed in the read data	T2T-Polish pipeline
clipped	Region where read alignments show a high level of clipping	T2T-Polish pipeline
flagger_intersect	A T2T-Polish issue which overlaps with Flagger’s “intersected” calls	Flagger v0.2

VIEWING THE DATA IN IGV

Most of the data that we will use for validation is hosted on the human-pangenomics aws server at: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/HG002/assemblies/polishing/HG002/. A PDF with a list of the various track descriptions and their aws URLs is maintained at

https://github.com/marbl/HG002-issues/blob/main/manuals/DescriptionOfAWSHostedIGVTracks.pdf, and these tracks can be loaded into IGV using their URLs, or using any of the IGV session files that are maintained at https://github.com/marbl/HG002-issues/tree/main/igv_sessions. These directories, session files, and the tracks on aws will be updated frequently as the project progresses, and we'll try to post to the T2T Slack HG002 channel when new tracks are available.

SESSION FILES AND VIEWING ISSUES IN IGV

The best way to determine the underlying cause of a coverage issue (is it due to an error in the assembly, is it due to bias in the experimental platform, is it both?) is to view as much relevant data as possible for that region. One easy way to view groups of relevant data tracks in IGV is to load IGV session files.

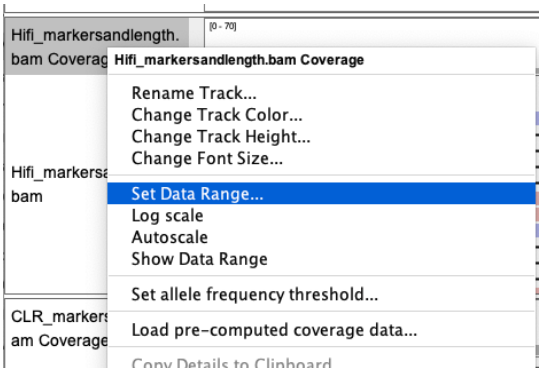
Session files contain specifications for the tracks you view in IGV, including where to obtain the data (generally an AWS URL for this project), what label should appear next to it, as well as data ranges for wiggle tracks, whether to display a coverage track for BAM-formatted alignments, etc. There are some example session files for this project on the github site at https://github.com/marbl/HG002-issues/tree/main/igv_sessions. Here are brief descriptions of what's included in each:

Session Name	Description	Included Tracks
hg002v0.7_validation.xml	Tracks produced by the T2T-polish pipeline	
hg002v0.7_aligned_reads.xml	Different types of alignments	

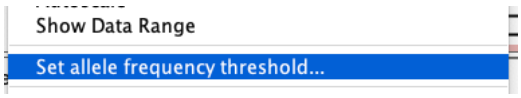
SCREENSHOTS

It may help to take screenshots of data for each issue region. Navigate to the region in IGV, 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 the issue ID as file name (e.g. Issue249.png).

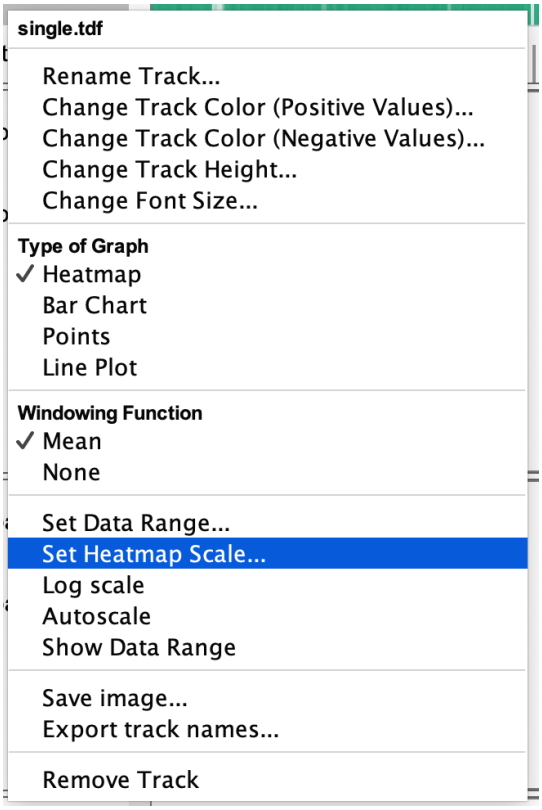
Format bam Coverage tracks

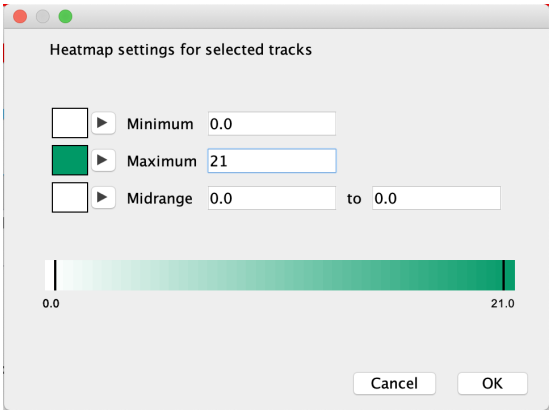


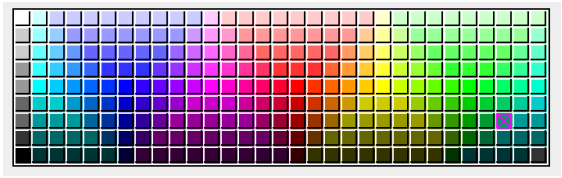
1) Set Data Range
    a) HiFi and CLR: Max to 70
    b) ONT: Max to 250
2) Set allele frequency threshold… to 0.3



Format single.tdf track

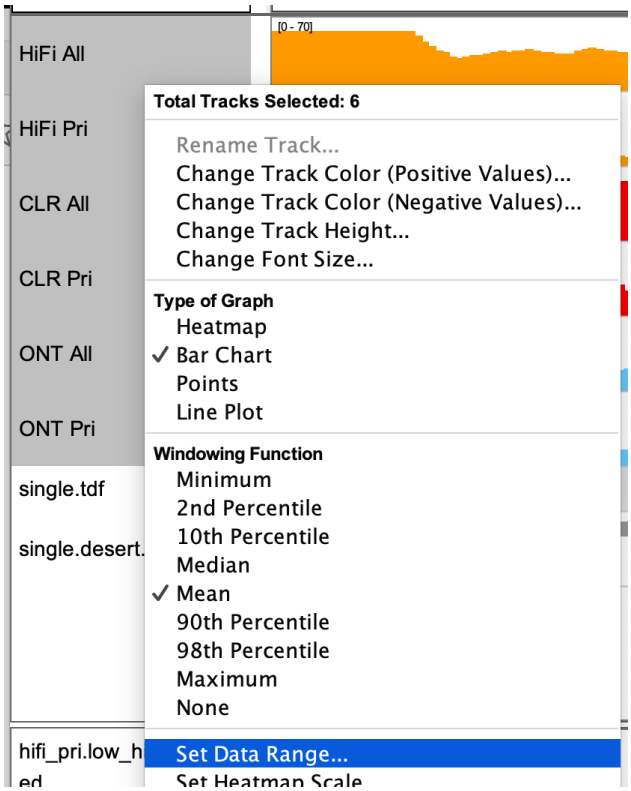


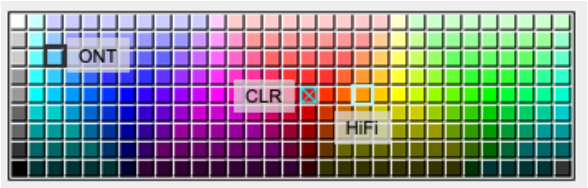


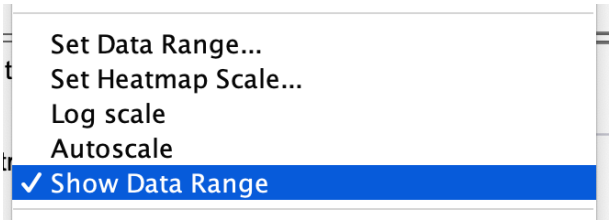


1) Select “Heatmap” as Type of Graph
2) Set Heatmap Scale
3) Set Maximum to 21
4) Select color and hit OK
5) Change Track Height to 30

Format .wig tracks

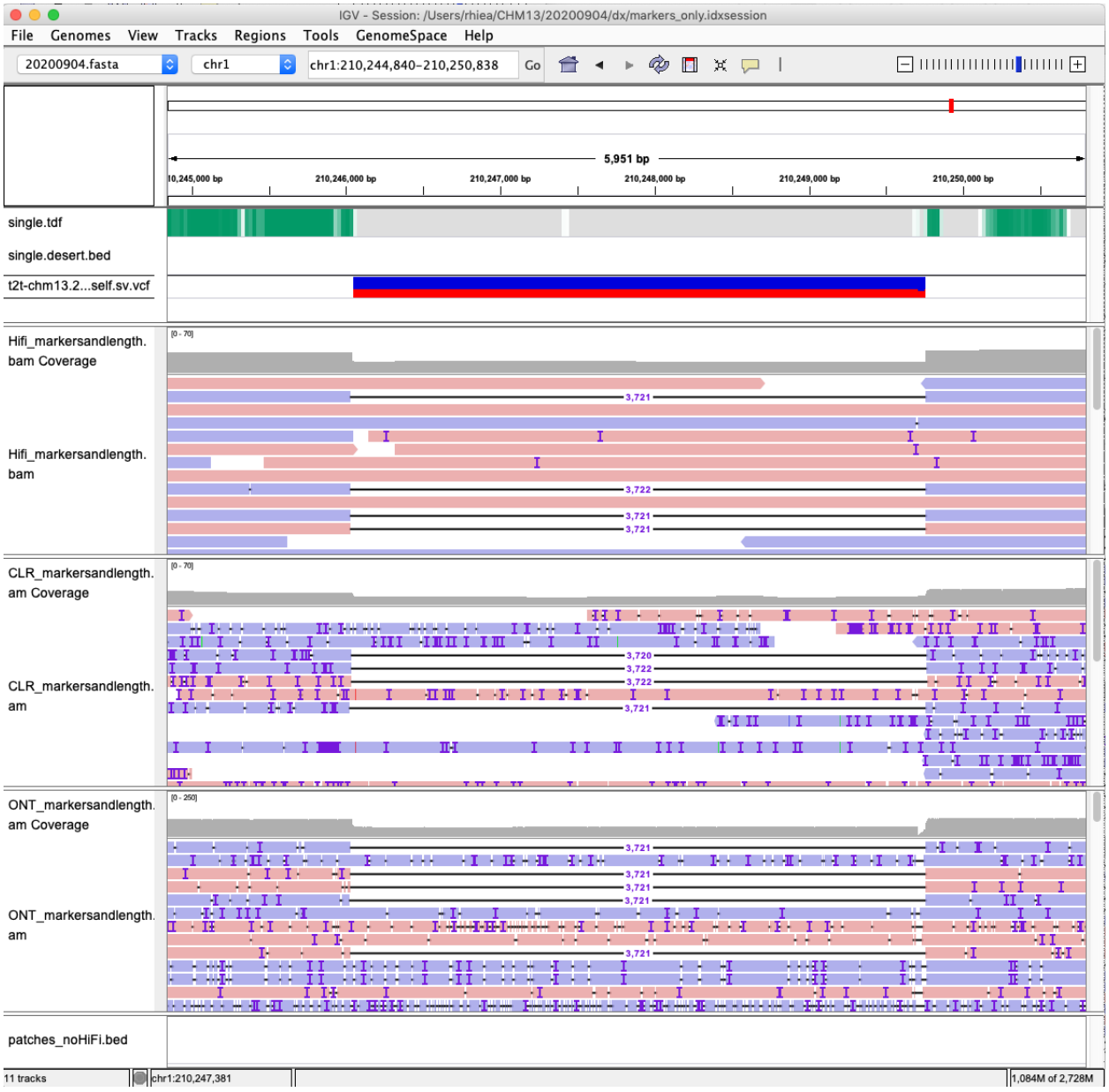






1) Set Data Range
    a) HiFi and CLR: Max to 70
    b) ONT: Max to 250
2) Change Track Color
    a) HiFi: 255, 153, 0
    b) CLR: 255, 0, 0

Example Screenshot



Checkpoints

-

Screenshot tip on MAC

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