Phase Switch Anomaly Polishing Instructions, February 12, 2023

This document is a guide to curating anomalies in HG002 diploid assemblies (currently the v0.7 verkko/rukki assembly) which suggest potential phase switch errors. The document describes the methods that we use when we try to determine whether these anomalies are due to alignment or binning issues in the methods that suggest phase switches (currently analysis of Strand-Seq data and kmer-based methods as implemented in Merqury), or whether there is an actual phase switch error or errors in the assembly.

This document is a work in progress, and available to edit at https://docs.google.com/document/d/1_gHoxjoaDOavROWHtTXFVr5F2E6xw-bwxFAIlk39724/edit#heading=h.8otnq3fgc0qt . Please feel free to suggest changes or additions as they occur to you.

HOW ARE PHASE SWITCH ANOMALIES DETECTED?

Our current set of phase switch "issues" (as of February, 2023) was created by using Bedtools to intersect two sets of phase switch "calls": The first set of calls are the regions for which Peter Ebert's Strand-Seq variant calls predicted a maternal allele that disagrees with the assembly's consensus for at least two SNVs spanning at least 1000 bases with no intervening non-discordant SNVs. The second set of regions are assembly locations where Merqury called a phase block on the opposite haplotype. There are 58 regions in the intersection between the Strand-Seq-derived regions and the Merqury wrong-haplotype phase blocks. These have been added to the issue section of our github repository (https://github.com/marbl/HG002-issues/issues) with the label "phase_switch". 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/109hui0CYDkkuyWHeHBbjuXHWEGFIk2X3A-gMsAULY9s. *Note: this google spreadsheet will be updated regularly FROM the github site, so any changes made to columns that mirror github will be overwritten!*

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
phase_switch	Evidence exists that the assembly maternal hap. consensus comes from the paternal haplotype and/or vice versa	Strand-Seq analyses (Peter Ebert), Merqury kmer analysis
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
strandseq_evidence	Strand-Seq shows evidence that there is a phase switch error in the assembly here	Peter Ebert

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 PHASE SWITCH ISSUES IN IGV

The best way to determine the underlying cause of a phase switch issue (is it due to errors in the strand-seq calls, the merqury kmer analysis, or the assembly?) 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 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_phase_switch_view.xml	Tags regions marked as potential phase switches and supportive info	
hg002v0.7_aligned_reads.xml	Different types of alignments	