

# Instructions for curating issues discovered from short read alignments, April 17, 2023

This document is a guide to curating anomalies in HG002 diploid assemblies (currently the v0.7 verkko/rukki assembly) which suggest falsely heterozygous sites, usually alterations in the length of mono- and di-nucleotide repeats. The document describes how these anomalies were detected, and the methods that we use to determine whether they are genuine errors in the assembly or false calls by DeepVariant.

*This document is a work in progress, and available to edit at [https://docs.google.com/document/d/1ysCOBmjpc1A0GC4ynFuLAvpbWms-KpVX\\_RU20\\_WIVE8/edit#](https://docs.google.com/document/d/1ysCOBmjpc1A0GC4ynFuLAvpbWms-KpVX_RU20_WIVE8/edit#). Please feel free to suggest changes or additions as they occur to you.*

## HOW ARE THESE ANOMALIES DETECTED?

Like many assembly issues (as of April, 2023), these potentially false heterozygous sites are discovered as high quality DeepVariant (DV) calls made on HG002 reads aligned to the HG002 assembly. In this case, the DV calls were made on alignments of short reads (100x of Element 2x150 reads aligned to the maternal haplotypes of the assembly only, with DV calls [here](#), and PacBio Onso “SBB” short reads, also aligned to the maternal assembly, with DV calls [here](#)). The calls being considered are homozygous non-reference (HNR) calls against the assembly. If the assembly consensus were correct in this region, we would expect at least some of the short reads to agree with it, so HNR calls can be inferred to be an indication of errors in the assembly. To assure a high true positive rate for these potential issues, we are considering only the intersection of calls from the Element and the Onso reads, filtered to include calls where the Element genotype quality (GQ) of 10 or greater. As of April 2023, our plan is to curate 15 of these variants to determine the reliability of five calls in each of three categories: Element GQ between 10 and 20, Element GQ between 20 and 30, and Element GQ between 30 and 40. This information will aid us in deciding what corrections to apply to the current and/or future assemblies that we are working to polish.

## ACCESSING YOUR 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. To restrict to the issues described in this document, filter the issues to include only those with the label “**short\_read\_homnonref**”.

## **SESSION FILES AND VIEWING SHORT READ HOMOZYGOUS NONREF CALLS IN IGV**

The best way to determine the underlying evidence for a DeepVariant call's reliability 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. For curating this type of anomaly, we have created an IGV session file and put it on the HG002-issues github site at [hg002v0.7\\_shortread\\_homnonref\\_aligns.xml](#). Once the file is on your computer (if you clone the HG002-issues repository, you can update it to new versions easily), its tracks can be loaded into IGV by selecting "File...Open Session" and selecting the file's location on your computer.

### **A CURATION EXAMPLE**

Once you have loaded the IGV session file, copy and paste the location of one of your assigned issues (here we will walk through the example of chrX\_MATERNAL:14640589-14640590) into the position window in IGV. IGV will automatically widen any region less than 41 base pairs to 41 base pairs to show the surrounding region. For our example region, our IGV window looks like this:







Finally, in addition to viewing read data, it's sometimes useful to view the surrounding area by zooming out a few thousand bases and looking for coverage anomalies, inconsistent read pairs (colored blue), or ambiguously mapped (colored white) reads. Another useful thing to do for issues on the autosomes is to view the same region on the alternate haplotype. For DeepVariant homozygous non-reference calls on the autosomes, the opposite haplotype should match all the reads.

## **LABELING AN ISSUE WITH YOUR DECISION**

Once you've convinced yourself that an issue is either a true error in the assembly (i.e., the reads show a true difference from the consensus) or that the DeepVariant call was wrong and the assembly is correct, you should add a **priority** or a **false\_positive** label to the issue on github, and add a few comments with screenshots of what you saw that made you make the decision you did. You will, in general, have a "co-curator" who may have already made comments or applied labels, but even if they did, it's helpful for you to add your independent assessment as well, so thank you for your efforts!

See [📖 Evaluating with IGV](#) for further tips & tricks regarding IGV.