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# . 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

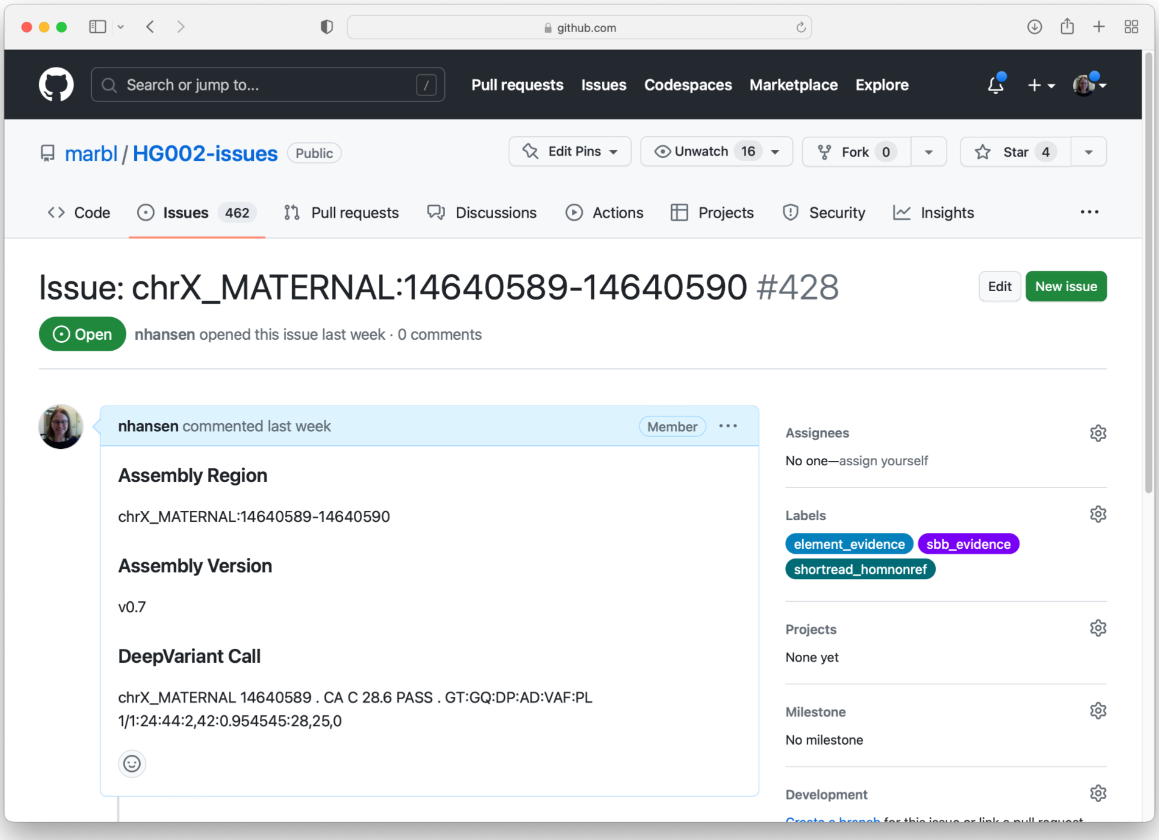
If you have volunteered to curate assembly issues and we have your github username, a number of issues will have been assigned to you for curation. You can view them at https://github.com/marbl/HG002-issues/issues (assuming you are logged in to github) by using the "Assignee" drop down menu to select 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

As an example, let's look at the DeepVariant call at chrX_MATERNAL:14640589-14640590, shown below in the GitHub HG002-issues repository. The DeepVariant call is a deletion of one base (reference "CA" becomes "C"), suggesting that the assembly's chrX_MATERNAL haplotype has an erroneous additional "A" at this position.



Issue: chrX_MATERNAL:14640589-14640590 #428

nhansen opened this issue last week · 0 comments

nhansen commented last week

Assembly Region

chrX_MATERNAL:14640589-14640590

Assembly Version

v0.7

DeepVariant Call

chrX_MATERNAL 14640589 . CA C 28.6 PASS . GT:GQ:DP:AD:VAF:PL 1/1:24:44:2,42:0.954545:28,25,0

Assignees
No one—assign yourself

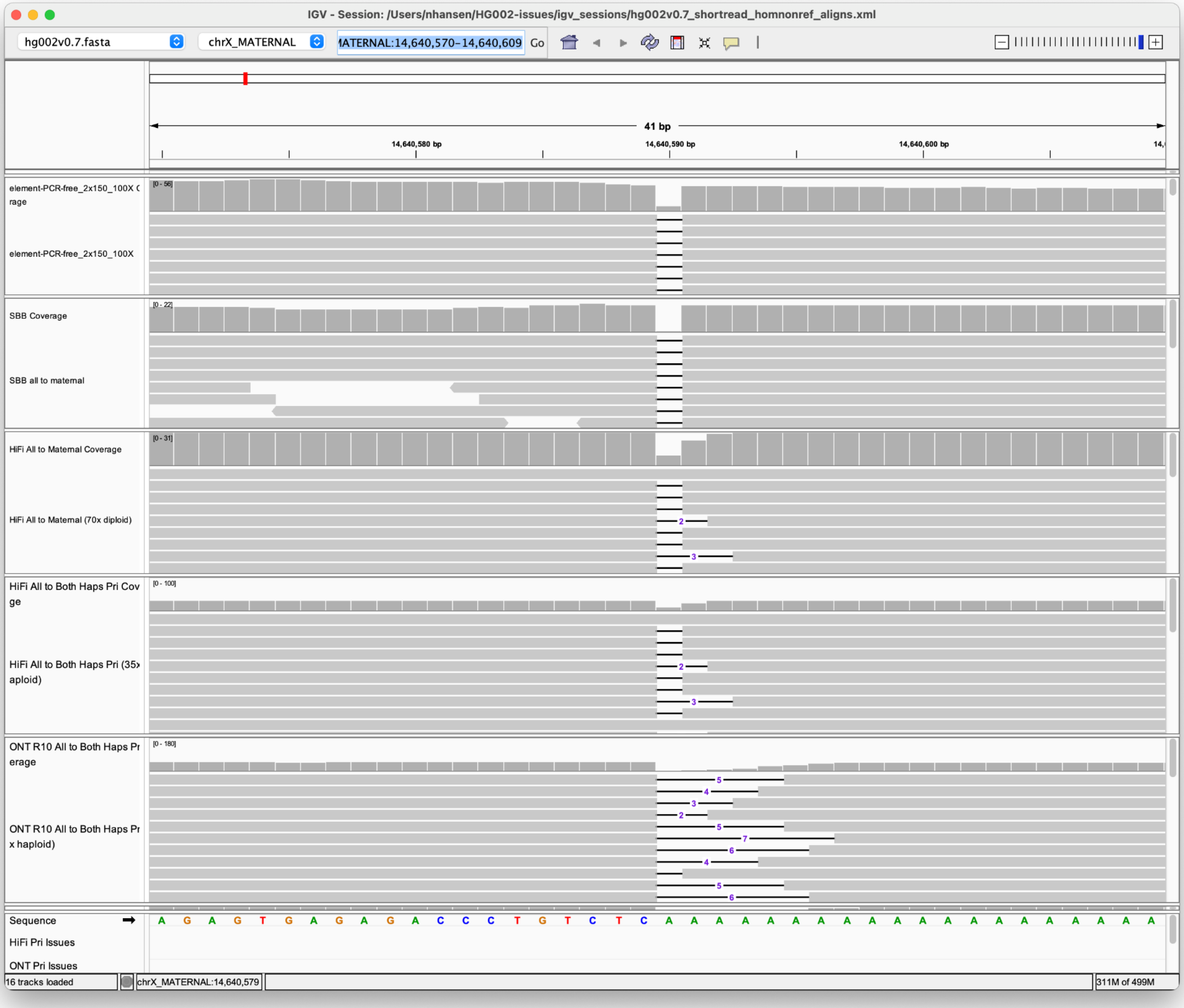
Labels
element_evidence sbb_evidence shortread_homnonref

Projects
None yet

Milestone
No milestone

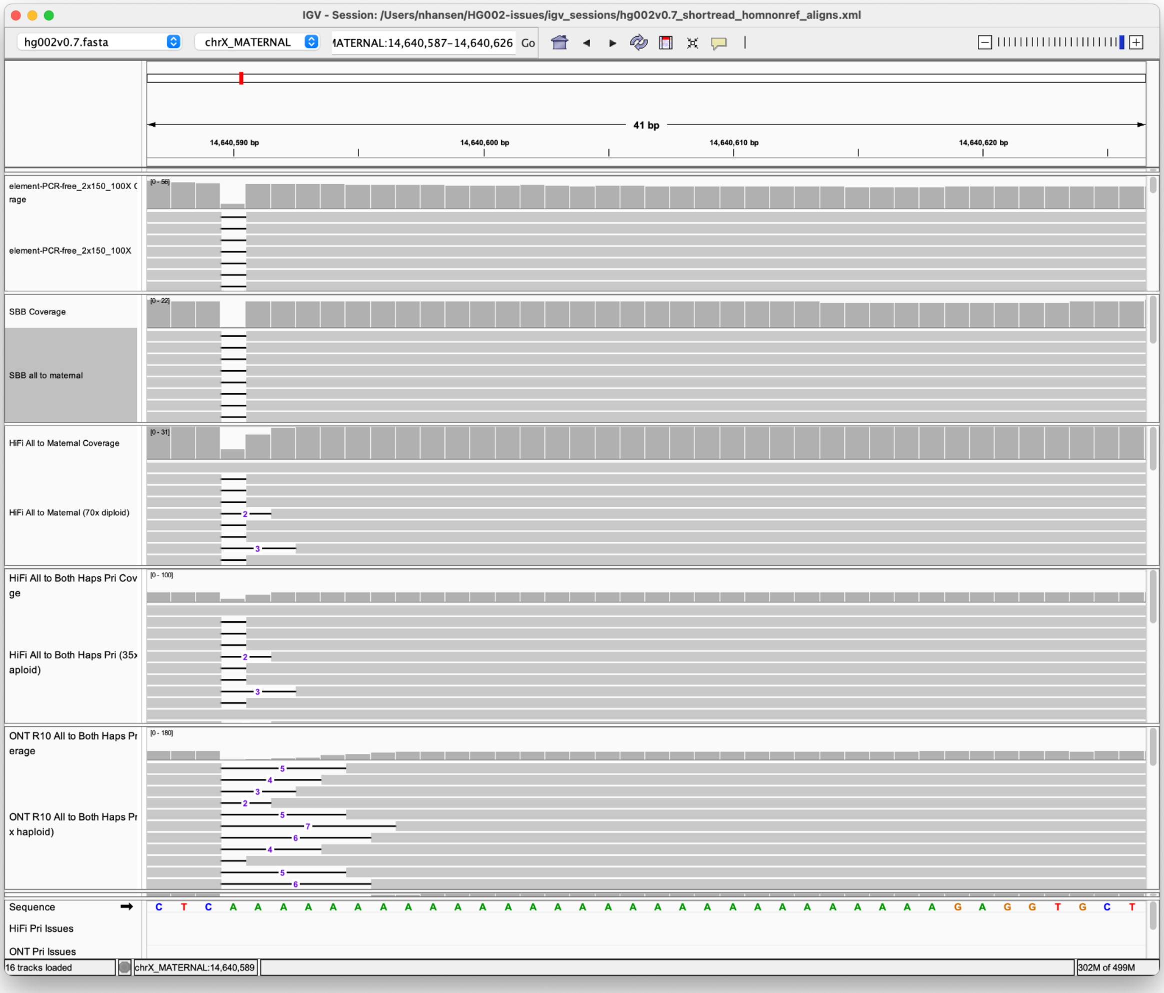
Development

To walk through this example, load the hg002v0.7_shortread_homnonref_aligns.xml IGV session file, and 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:

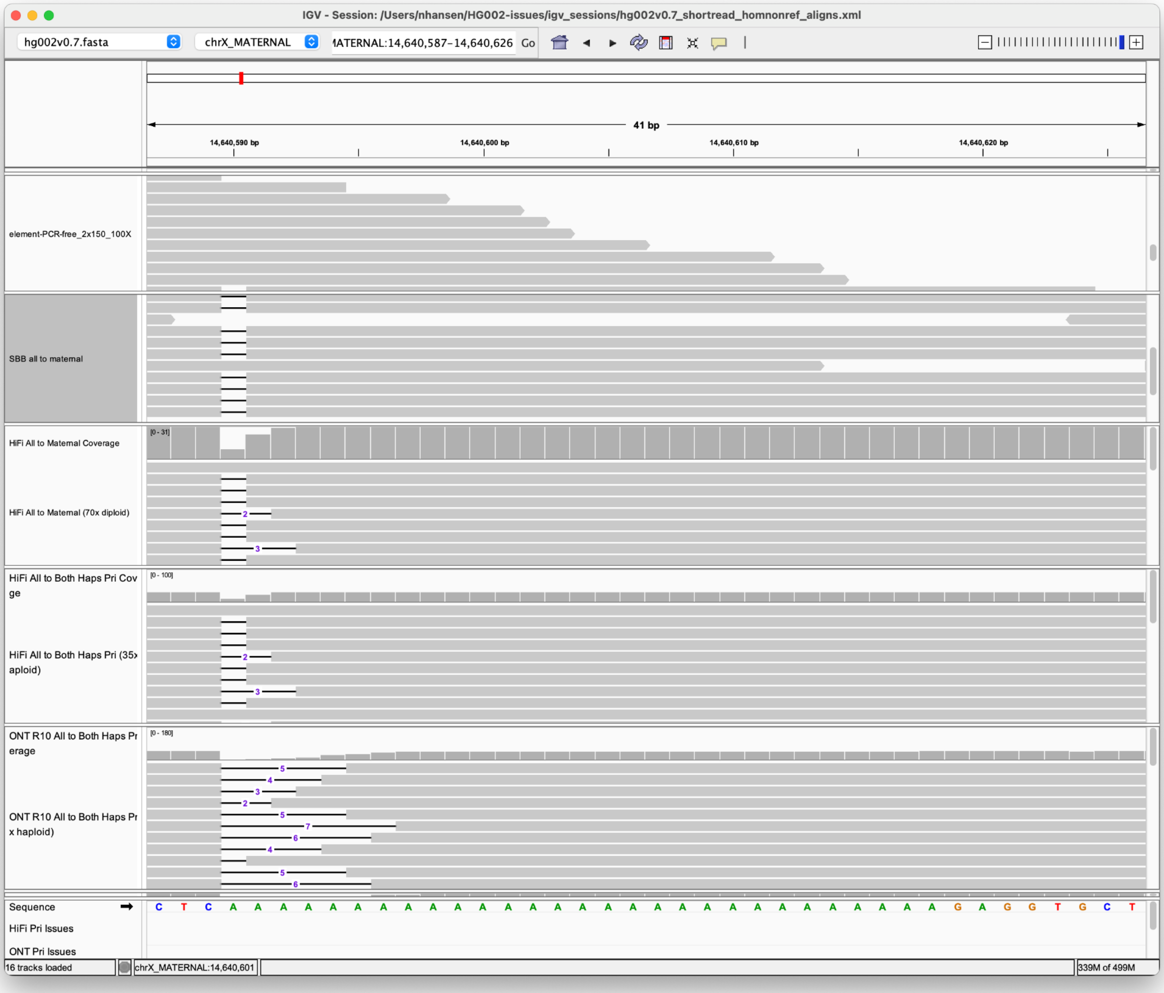

IGV - Session: /Users/nhansen/HG002-issues/igv_sessions/hg002v0.7_shortread_homnonref_aligns.xml
hg002v0.7.fasta
chrX_MATERNAL
IATERNAL:14,640,570-14,640,609
Go
41 bp
14,640,580 bp
14,640,590 bp
14,640,600 bp
element-PCR-free_2x150_100X (
rage
[0 - 56]
element-PCR-free_2x150_100X
SBB Coverage
[0 - 22]
SBB all to maternal
HiFi All to Maternal Coverage
[0 - 31]
HiFi All to Maternal (70x diploid)
2
3
HiFi All to Both Haps Pri Cov
ge
[0 - 100]
HiFi All to Both Haps Pri (35x
aploid)
2
3
ONT R10 All to Both Haps Pr
erage
[0 - 180]
ONT R10 All to Both Haps Pr
x haploid)
5
4
3
2
5
7
6
4
5
6
Sequence
A G A G T G A G A G A C C C T G T C T C A A A A A A A A A A A A A A A A A A A A A
HiFi Pri Issues
ONT Pri Issues
16 tracks loaded
chrX_MATERNAL:14,640,579
311M of 499M

We can clearly see the called variant (the deletion of an A at position 14,640,590) because our tracks have all been configured not to “hide small indels”. If you load tracks yourself while curating rather than using the provided IGV session file, **be sure that you also have not set each track to “hide small indels”**.

Looking at this spot of the genome, we can see clearly that the deleted A is just one of a long mononucleotide stretch of A’s, so long, in fact, that we can’t see the 3’ end of the run. To center our screen on the entire run of A’s, we can click on reads within a track and drag the entire view to the left until the string of A’s is centered on our screen. If the stretch you happen to be looking at is longer than this example, you may have to use the “-” box on the upper right of your IGV window to zoom out a bit. After moving the screen’s view, it looks like this:


IGV - Session: /Users/nhansen/HG002-issues/igv_sessions/hg002v0.7_shortread_homnonref_aligns.xml
hg002v0.7.fasta
chrX_MATERNAL
ıATERNAL:14,640,587-14,640,626
Go
41 bp
14,640,590 bp
14,640,600 bp
14,640,610 bp
14,640,620 bp
element-PCR-free_2x150_100X C rage
[0 - 56]
element-PCR-free_2x150_100X
SBB Coverage
[0 - 22]
SBB all to maternal
HiFi All to Maternal Coverage
[0 - 31]
HiFi All to Maternal (70x diploid)
2
3
HiFi All to Both Haps Pri Cov ge
[0 - 100]
HiFi All to Both Haps Pri (35x aploid)
2
3
ONT R10 All to Both Haps Pr erage
[0 - 180]
ONT R10 All to Both Haps Pr x haploid)
5
4
3
2
5
7
6
4
5
6
Sequence
C T C A A A A A A A A A A A A A A A A A A A A A A A A A A A A A G A G G T G C T
HiFi Pri Issues
ONT Pri Issues
16 tracks loaded
chrX_MATERNAL:14,640,589
302M of 499M

Our next step will be to determine whether **all** of the read data are consistent with the alternative consensus suggested by the variant call. In this case, we are testing the hypothesis that rather than the 29 A’s that the assembler has predicted, the actual HG002 sample has only 28 A’s at this spot on its maternal chromosome X. Without scrolling through the reads that are hidden, we can see that the Element and Onso reads within view all agree that there should be one less A than the assembly has, but the gray coverage track above the Element reads shows that at least some of the read alignments have aligned an A to the position of the additional (wrong) base, so we should investigate why that is. By scrolling down in both the Element and the Onso read alignment tracks, we can view the read alignments that don’t contain the one-base deletion:


IGV - Session: /Users/nhansen/HG002-issues/igv_sessions/hg002v0.7_shortread_homnonref_aligns.xml
hg002v0.7.fasta
chrX_MATERNAL
IATERNAL:14,640,587-14,640,626
Go
41 bp
14,640,590 bp
14,640,600 bp
14,640,610 bp
14,640,620 bp
element-PCR-free_2x150_100X
SBB all to maternal
HiFi All to Maternal Coverage
[0 - 31]
HiFi All to Maternal (70x diploid)
HiFi All to Both Haps Pri Cov
ge
[0 - 100]
HiFi All to Both Haps Pri (35x
aploid)
ONT R10 All to Both Haps Pr
erage
[0 - 180]
ONT R10 All to Both Haps Pr
x haploid)
Sequence
C T C A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A G A G G T G C T
HiFi Pri Issues
ONT Pri Issues
16 tracks loaded
chrX_MATERNAL:14,640,601
339M of 499M

At first glance, the read alignments that don't contain the deletion might appear to suggest that the shorter stretch of A's is not in every read, and that these reads therefore give support to the assembly's consensus. But looking more closely, we can see that none of the alignments without the deletion extend to the end of the run of A's, and because of that, these reads are unable to tell us anything about which length of A's is correct. It's important to remember when curating that **alignments which fail to extend entirely across a mono- or di-nucleotide run are not informative about the length of that run.**

The HiFi and ONT long reads also have alignments that fail to support the assembly or the suggested correction, but this is a result of the two platforms' inaccuracy in base-calling mono- and di-nucleotide stretches that are this long. It's worth looking at the read data from as many examples of these long sequence stretches as you can find, in order to get a feel for what a "normal" distribution of length discrepancies is for each platform. In general, ONT especially can vary very widely in the number of single- or two-base repeats it calls.

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.