CHM1_1.1 PACBIO 54X BLASR REVIEW (CG-1839)

GOAL

Use PacBio read alignments to CHM1_1.1 to identify regions of misassembly introduced as a consequence of using a reference-guided approach to assembly. Mis-assemblies may be errors propagated from GRCh37 or reflect structural variation.

METHODS

PacBio reads: http://datasets.pacb.com/2014/Human54x/fast.html

BLASR: https://github.com/PacificBiosciences/blasr

Used blasr-update provided directly from Mark Chaisson on 2013-12-10

Used compute farm with pre-computed files:

```
sawriter  GCF_000306695.2.fasta.sa.8  GCF_000306695.2.fasta   -blt 8

sa2bwt  GCF_000306695.2.fasta GCF_000306695.2.fasta.sa.8
GCF_000306695.2.fasta.bwt

printTupleCountTable  chm1.reads.tuple.table 12 reads/m13*.fasta


blasr
m131012_060411_42215_c100597762550000001823105905221422_s1_p0.1.subrea
ds.fasta GCF_000306695.2.fasta

 -bwt GCF_000306695.2.fasta.bwt   -ctab  chm1.reads.tuple.table

 -nproc 4 -sam  -clipping soft  -bestn 2 -minMatch 12  -affineAlign -
sortRefinedAlignments
```

BENCHMARKING

INDIVIDUAL JOBS

- Took between 7.967G and 10.403G of ram, as reported by the farm 'job finished' report.
- Between 00:06:44 and 01:53:35 'User Time'. (User Time is roughly Wall Time * 4, due to '-nproc 4') Average closer to the ~1:30 range.

FULL JOBS

- Limited to 25 simultaneous Farm Jobs, ran from 'Feb 19 15:34' to 'Feb 20 02:43'. A little over 12 hours.
- Farm Jobs needed to be given 'h_vmem=12G, mem_free=12G' to have enough RAM space.
- Farm Jobs need to be given LD_LIBRARY_PATH='/usr/local/hdf5/1.8.10/lib/', so that blasr can find its HDF5 library, even though we aren't using HDF5 data.

APPROACH

We sought to identify errors in the CHM1_1.1 genome assembly (GCF_000306695.2) introduced as a consequence of errors in the GRCh37 primary assembly unit, which was used to guide its assembly. It was hypothesized that alignments of CHM1 PacBio reads in such regions of CHM1_1.1 would exhibit one or more of the following characteristics:

1. Low coverage with respect to coverage in surrounding regions
    a. Low coverage is often associated with highly fragmented assembly regions, which are themselves hallmarks of assembly problems (though they may not necessarily reflect errors introduced by GRCh37).
2. Sharp boundaries at which alignment coverage dropped off
    a. These could occur at component boundaries (indicative of GRCh37 tiling path errors) or within assembly components (indicative of component assembly errors in GRCh37)
3. Inversions

Although other assembly features (i.e. repeats or structural variation) can also result in read alignments having similar characteristics, such regions should be enriched for assembly errors.

To identify CHM1_1.1 assembly errors corresponding to unrecognized GRCh37 errors, we focused on CHM1_1.1 assembly sites where alignment coverage dropped off sharply (exhibited characteristic #2). To this end, we produced a list of CHM1_1.1 regions where there were PacBio aligned reads that met the following criteria. We refer to these reads as "cliffs".

- PacBio read must have 2 alignments on CHM1_1.1 (no more, no less)
- Both alignments must be on the same CHM1_1.1 sequence
- One of the two alignments must meet the criteria of "Score ≤ -2.0*ReadLength"

    - This ensures that at least one of the alignments is "reasonable" and prevents us from declaring a read with two bad alignments as a cliff read.
- Query-coverage of smaller of 2 segments ≥ 10%
    - The smaller alignment must still involve at least 10% of the PacBio read
- Intersecting (query-only checked) coverage ≤ 10%
    - This means that the two alignments do not overlap each other by more than 10%
- Unique coverage ≥ 50% (refers to full query)
    - Max coverage can be 110%, due to the intersection rule
- Coverage drop-offs that occurred within 1 Kbp of a CHM1_1.1 boundary were flagged

The PacBio reads used for this analysis aligned to the CHM1_1.1 assembly at an average coverage depth of 54x. As expected, coverage at regions containing repetitive sequence was notably higher. To improve our likelihood of detecting examples of mis-assemblies, we restricted our review of this list to sites where surrounding coverage did not indicate the presence of repetitive sequence and the drop-off in coverage was roughly equivalent to surrounding coverage.

RESULTS:

Report File: cliff_report_54_full (posted on GRC private FTP site: ../ human/MISC/CHM1_1.1)

BAM files for cliff reads only: cliffs_54.bam, cliffs_54.bam.bai (posted on GRC private FTP site: ../ human/MISC/CHM1_1.1). These can be loaded to Genome Workbench).

Column	Descriptor	Comment
1	Chromosome	Chromosome
2	Coordinate bin	Starting coordinate of 1Kb sequence bin containing reads meeting cliff criteria are found
3	Cliff count	Count of alignment ends meeting cliff criteria that fall the coordinate bin.
4	Cliff type	C: cliff is within 1K of component boundary M: cliff is >1K from component boundary
5	Read depth	count of reads with >50% coverage in the bin

The full report has no filtering. Any 1Kb sequence bin that has even 1 alignment meeting cliff criteria in it is reported. Likewise, no filtering is applied with respect to read depth. In repetitive regions, the read depth greatly exceeds the expected 54x coverage.

Coordinate bins of interest are going to have a "reasonable" cliff count coupled with an average cliff depth. Initial review of the report has suggested that bins where the cliff count is ≥10 and the depth is less than 2x the coverage (<108) would likely yield interesting results and would be good for first pass review (File: cliff_report_54_gte10_lt108). Several of these cases have been reviewed for this report (see below)

Inversions or indels between PacBio CHM1 reads and the CHM1_1.1 assembly can generally be recognized as pairs of closely located cliff bins in the report. **It is important to look at graphical views of**

individual reads in addition to graphical views of the CHM1_1.1 assembly. Doing so gives you a better understanding of the relationship between the assembly and the read sequence.

Some examples of assembly issues that can be detected via review of the cliff report are shown below.

NC_018934.2 (CHRX): ~8MB: FALSE GAP IN CHM1_1.1

This is an example where PacBio alignments suggest that a 50kb gap was erroneously introduced in the CHM1_1.1 assembly. This looks like an assembly error in the CHM1_1.1 assembly, rather than structural variation since there is no corresponding gap in the GRCh37 chromosome where it aligns (assembly-assembly alignments not shown; NC_000023.10: ~8Mb), and no errors in the GRCh37 assembly are suspected at this location. The reason for this apparent assembly error has not yet been ascertained.

Report shows pair of cliffs at:

#chr	1kb_bin	cliff_edges	comp	depth
chrX	8032000	36	C	46
chrX	8082000	39	C	46

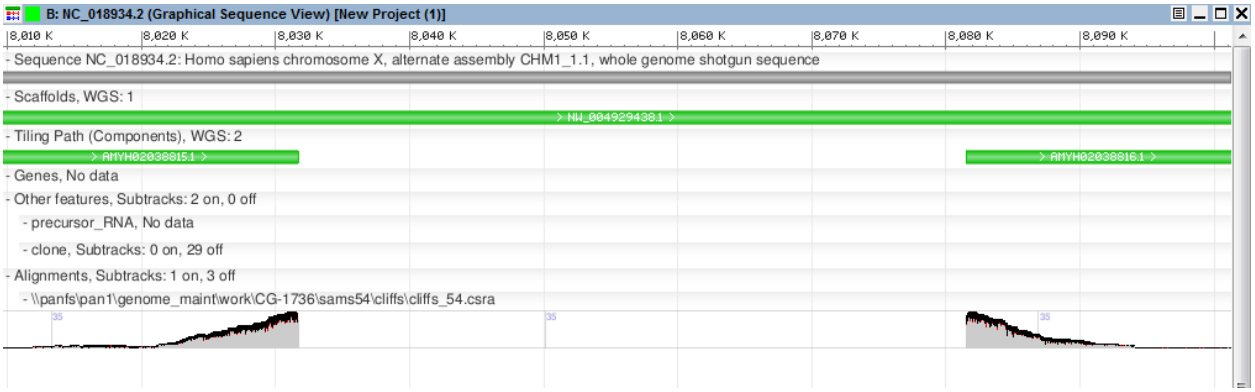

Figure 1: Zoomed out view showing PacBio reads meeting cliff criteria aligned to CHM1_1.1 at this region. Note how the read count increases rapidly from background as you come near the gap edges.

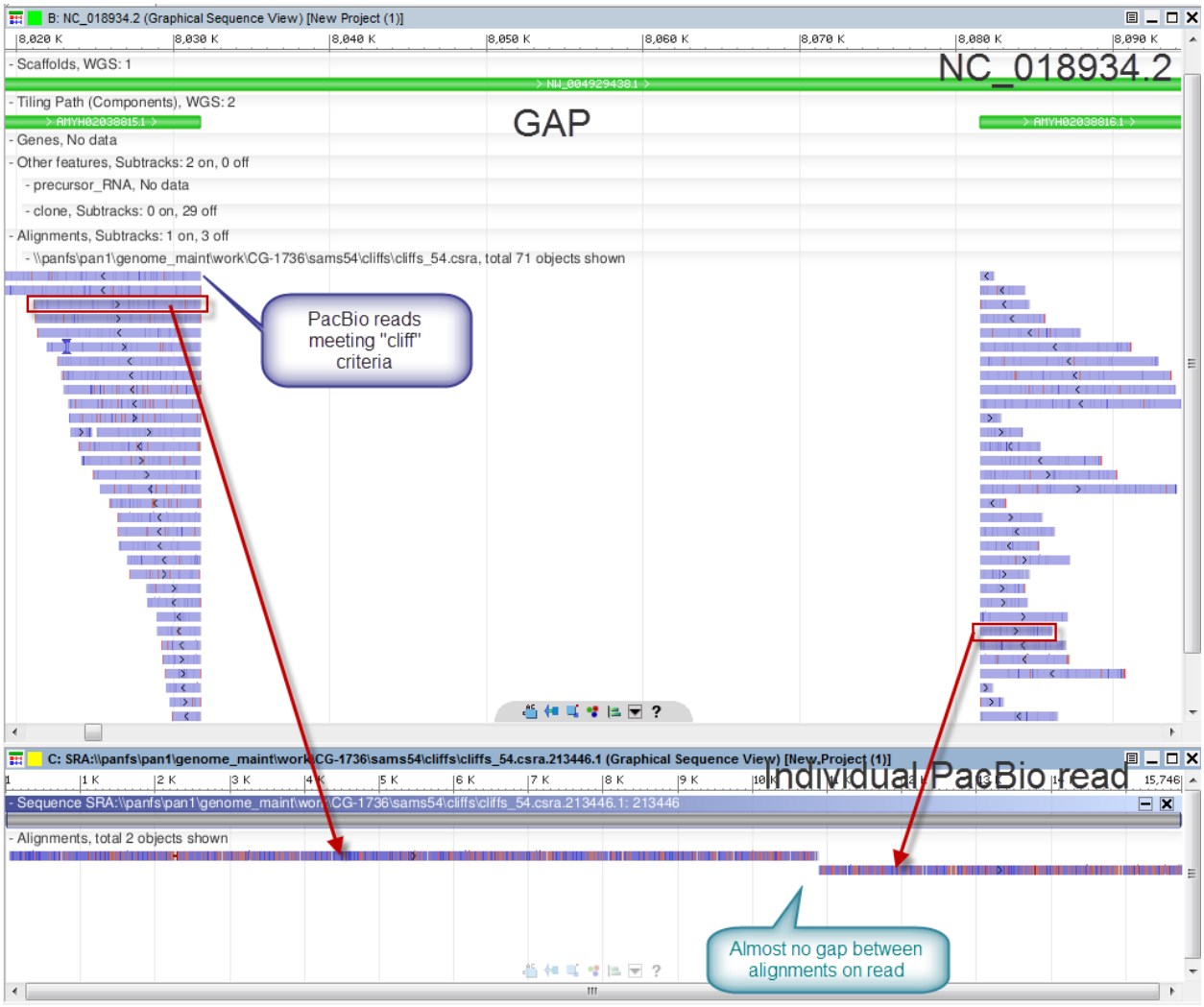

Figure 2: (**Top panel**) View of same region of NC_018934.2 that shows individual PacBio alignments at this gap. Two alignments corresponding to one read have been selected (red arrows/boxes). The *lower panel* shows the PacBio read whose alignments are highlighted in the top panel. Note that the alignments do not overlap on the read, but are very close.

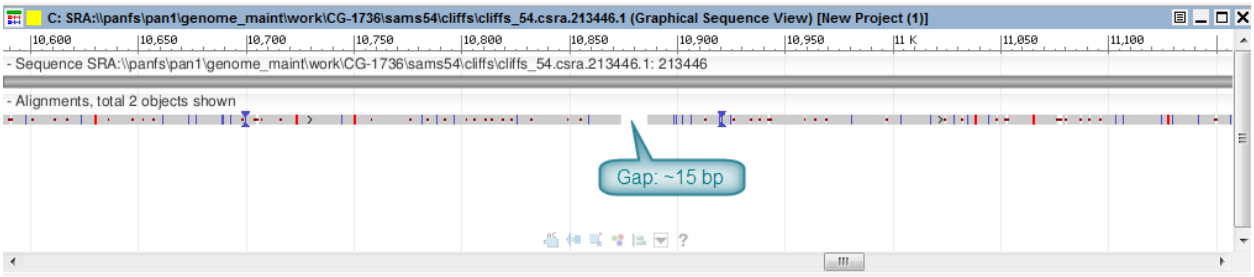

Figure 3: Zoomed in view of PacBio read shown in the bottom panel of Figure 2. This image demonstrates that the gap between the alignments on the read is only ~15 bp, as opposed to the 50Kb that is in the CHM1_1.1 assembly.

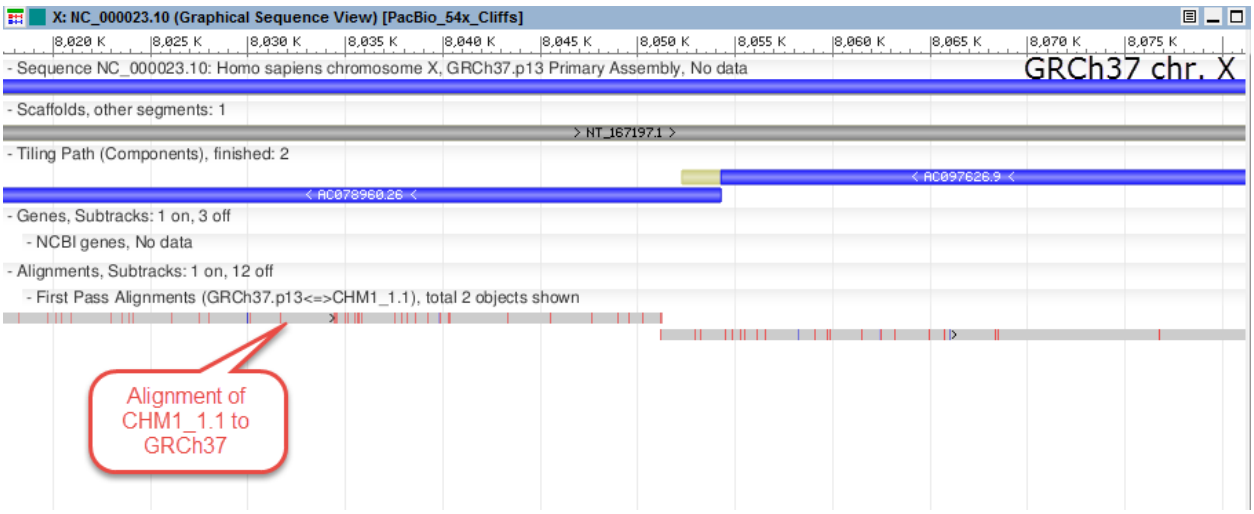

Figure 4: No gap is observed in the GRCh37 assembly at the corresponding location. Thus, it is not clear why a gap of this size was introduced in the CHM1_1.1 assembly.

Review shows that this gap's size was defined during the assembly process, and was not modified during the subsequent gap clean-up that was performed to identify inter-scaffold gaps. Further analysis will be needed to determine why such a large gap was placed into the assembly between these two WGS contigs.

CHM1_1.1 AGP assembly (pre-gap manipulation and submission):

chrX 7348085 7941283 373 W all.251651.1 1 593199 +

chrX 7941284 7991094 374 N 49811 scaffold no na

chrX 7991095 8025977 375 W all.251672.1 1 34883 +

CHM1_1.1 AGP (post-gap manipulation and submission):

CM001631.2 7438512 8031710 375 W AMYH02038815.1 1 593199 +

CM001631.2 8031711 8081521 376 N 49811 scaffold yes align_genus

CM001631.2 8081522 8116404 377 W AMYH02038816.1 1 34883 +

NC_018912.2 (CHR. 1) 153.95 MB: OVERLY LARGE GAP DUE TO LIKELY STRUCTURAL VARIATION BETWEEN CHM1 AND. GRCH37

This is another example where PacBio alignments suggest that a 32Kb gap in NC_018912.2 is also overly large, but in this case likely reflects structural variation between the CHM1 and GRCh7 genomes. Alignment of CHM1_1.1 to the GRCh37 assembly shows that the reference assembly has sequence in this region, not a gap. Thus, sequence is only "missing" from the CHM1_1.1 assembly. However, CHM1_1.1 aligns to the PacBio reads without a gap, indicating that no CHM1 sequence is missing from the assembly. Review of the annotation shows that there is a family of LCE3 genes here. The data suggests that there is structural variation between CHM1 and the reference here, with the CHM1 haplotype carrying a 32 Kb deletion relative to the reference. Two genes fall in the indel: LCE3C and LCE3B. The gap in the CHM1_1.1 assembly appears to have been introduced because GRCh37 guided the assembly. A similar situation is observed at NC_018912.2 72.9Mb.

Report shows a pair of cliffs:

#chr	1kb_bin	cliff_edges	comp	depth
chr1	153951000	36	C	49
chr1	153983000	34	C	57

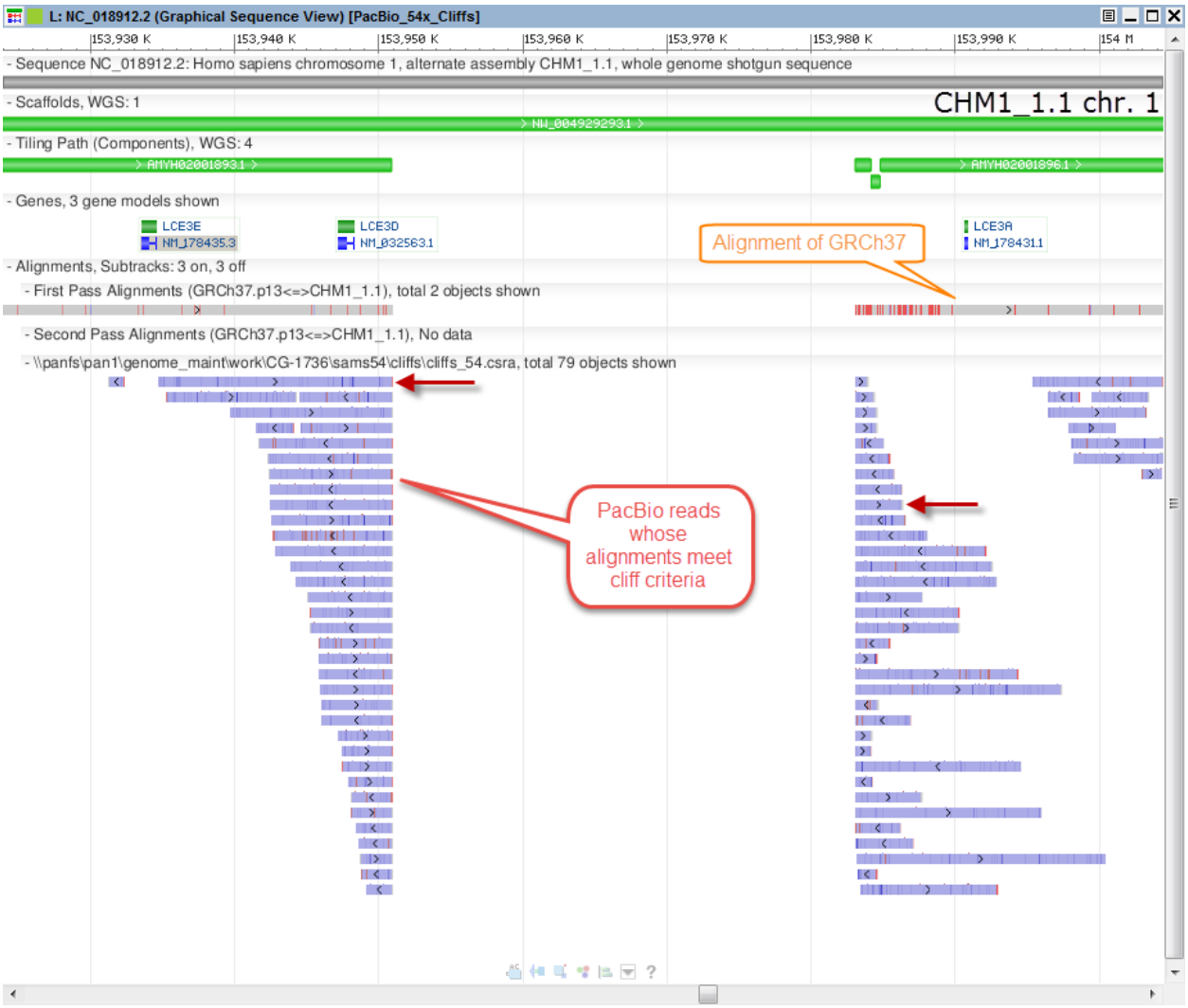

Figure 5: PacBio reads aligning on either side of a gap in the CHM1_1.1 assembly. The red arrows highlight 2 non-overlapping alignments of the same PacBio read (shown in Fig. 6). The many other alignments on either side of this gap also reflect alignment pairs.

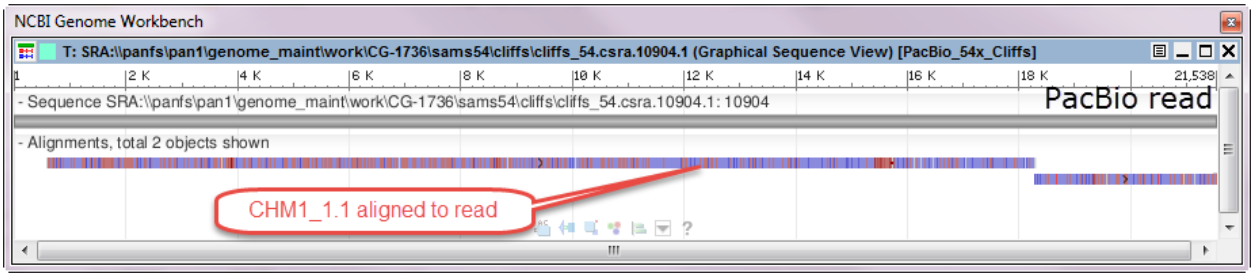

Figure 6: CHM1_1.1 assembly aligned to the PacBio read whose 2 alignments are highlighted in Fig. 5. Note that there is no gap between the alignments. The lack of unaligned sequence suggests that the gap in the CHM1_1.1 assembly can be closed.

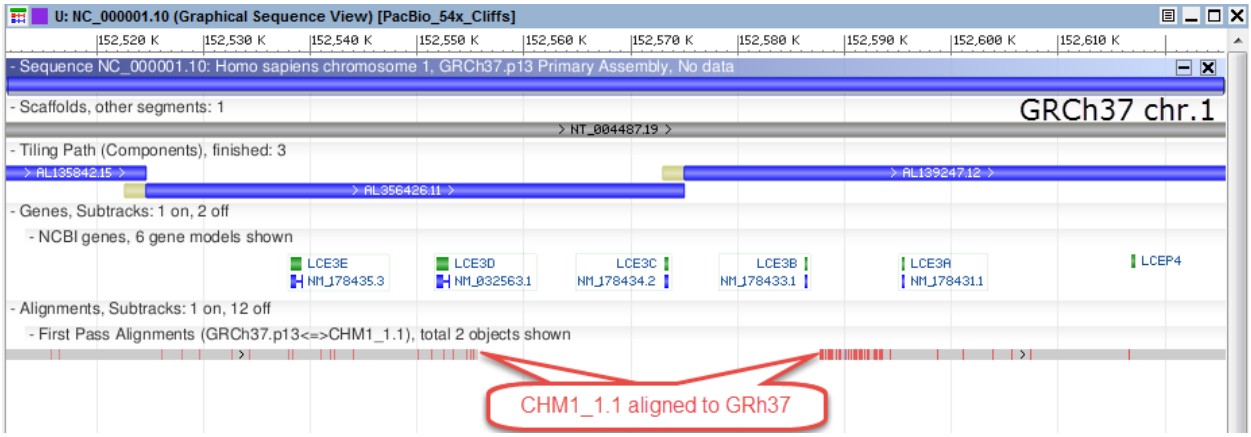

Figure 7: Alignment of CHM1_1.1 assembly to GRCh37 chr. 1. There is a ~32Kb region of the GRCh37 that does not align to the CHM1_1.1 genome (alignment gap). There are 2 members from a family of LCE3 genes that are annotated in this region.

NC_018192.2 (CHR. 1) 245.29MB: COLLAPSED REPEAT IN CHM1_1.1 ASSEMBLY

This is an example where the PacBio read alignments suggest that a repeat in the CHM1 genome has been collapsed in the CHM1_1.1 assembly. There is no annotation in the GRCh37 assembly (AL592151.13: RP11-351N5) to suggest an assembly error. This suggests that the PacBio read alignments may be indicating structural variation in this region.

There is a single cliff in this region:

#chr	1kb_bin	cliff_edges	comp	depth
chr1	245288000	47	M	88

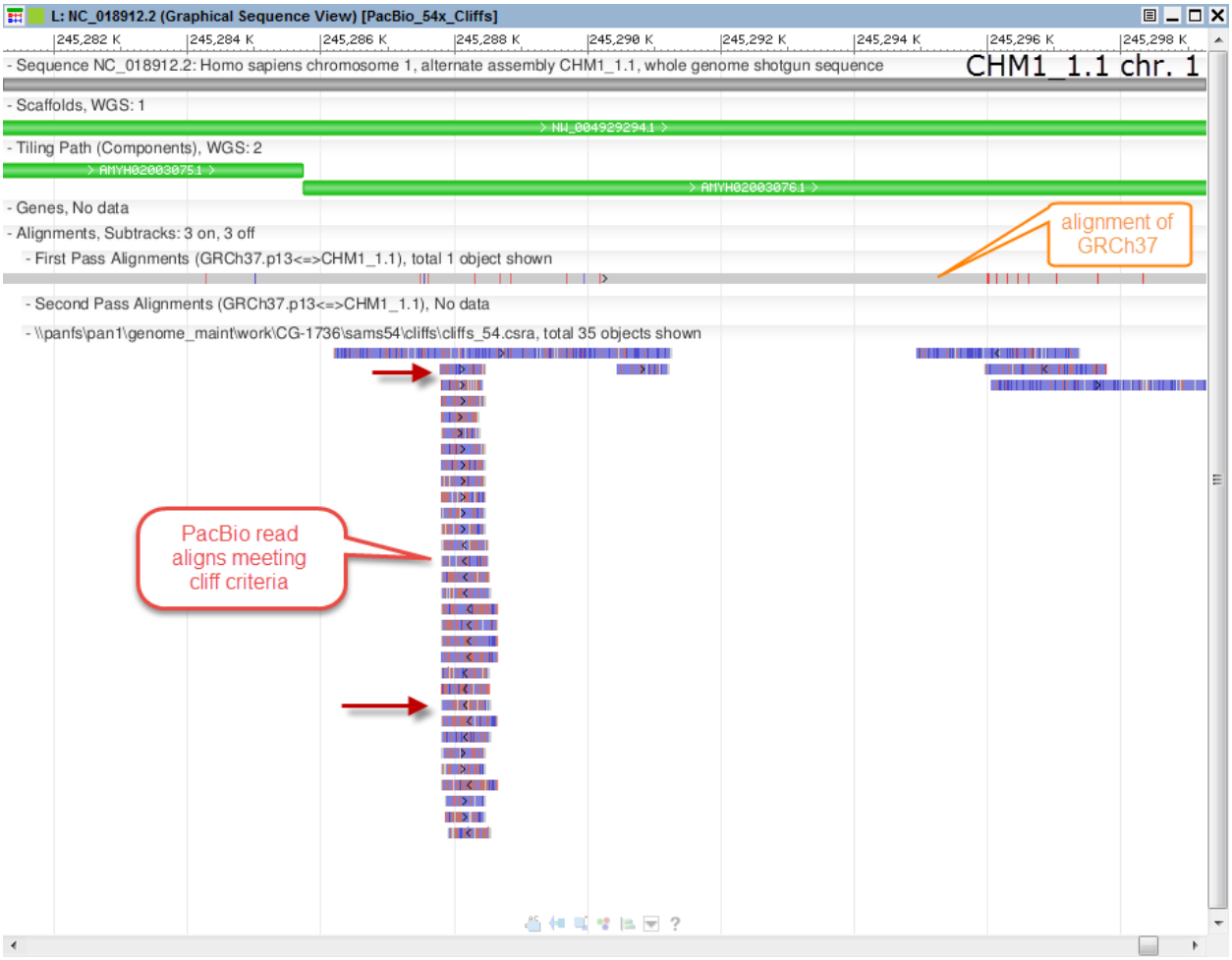

Figure 8: PacBio reads aligned to CHM1_1.1. There are a large number of ends meeting cliff criteria. This location occurs within a single WGS contig. The red arrows highlight 2 overlapping alignments from the same PacBio read. Review demonstrates that the other overlapping alignments at this location are also pairs, and do not overlap on the PacBio read (not shown).

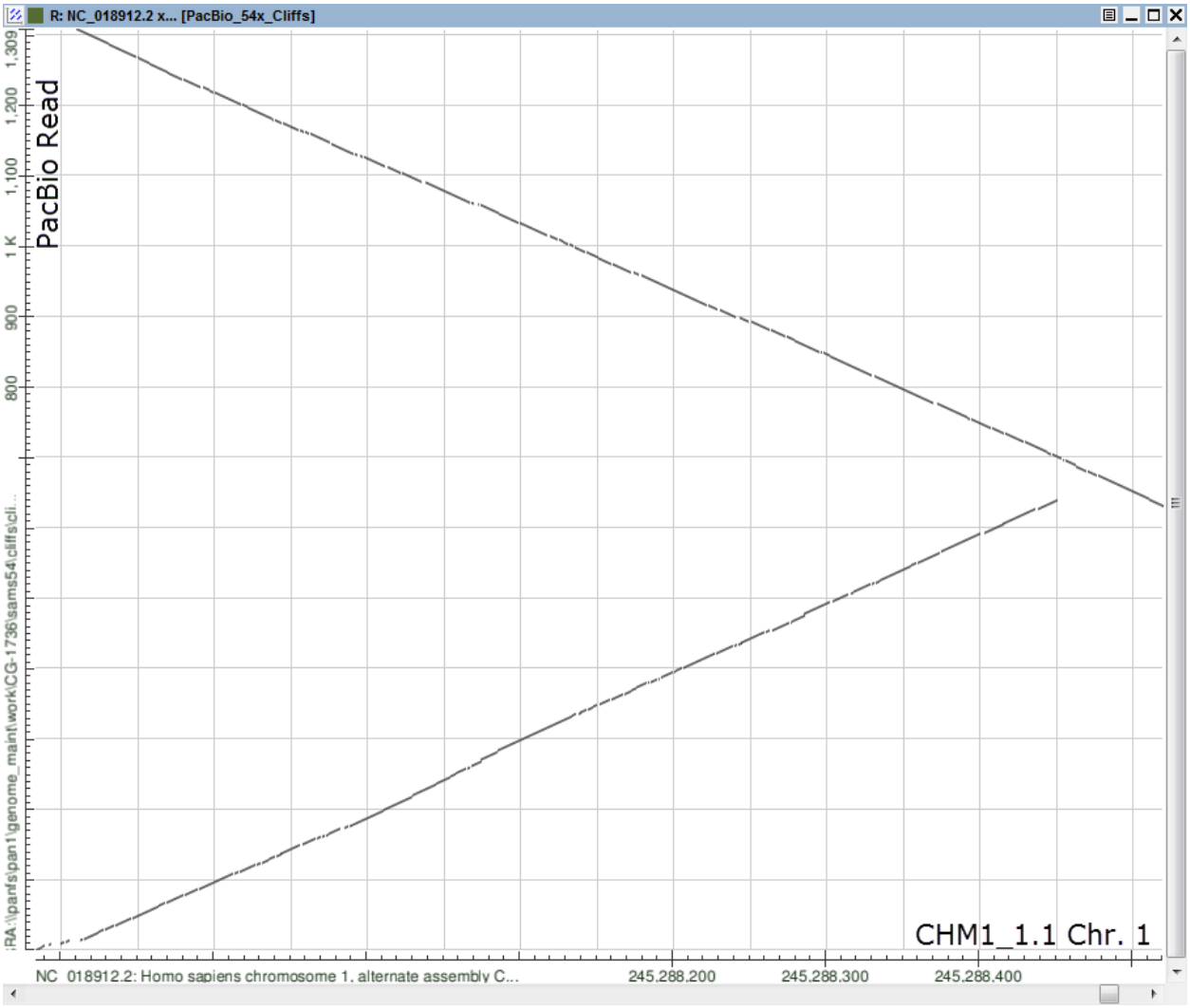

Figure 9: Dot matrix view of the PacBio read whose alignments are highlighted in Figure 8 aligned to CHM1_1.1 chr. 1. The alignment suggests that the CHM1 genome contains an inverted duplication that was collapsed in the assembly of the CHM1_1.1 WGS contig.

NC_018922.2 (CHR.11): 1.9MB GRCH37 PROBABLE INVERSION ERROR PROPAGATED TO CHM1_1.1 (HG-28)

In this example, there are 3 bins with large numbers of reads meeting cliff criteria located within 50 Kb of one another near NC_018922.2 1.9 Mb. These bins all coincide with WGS contig boundaries in the CHM1_1.1 assembly, suggesting there may be a path problem. Review of the PacBio read alignments in this region reveals that the 3 bins can be resolved into 3 non-adjacent groups of split reads, 2 of which are inverted relative to one another.

#chr	1kb_bin	cliff_edges	comp	depth
chr11	1914000	27	C	73
chr11	1936000	33	C	72
chr11	1960000	26	C	69

Analysis suggests the CHM1_1.1 path should be reordered and reoriented as follows in order to be consistent with the PacBio alignments:

Existing:

AMYH02024179.1 +
AMYH02024180.1 +
AMYH02024181.1 +
AMYH02024182.1 +
AMYH02024183.1 +

Corrected:

AMYH02024179.1 +
AMYH02024182.1 +
AMYH02024181.1 -
AMYH02024180.1 -
AMYH02024183.1 +

The inverted contigs are approximately 20 kb in length. These findings corroborate reports of an inversion/translocation in GRCh37 assembly component AC051649.21 (RP11-534I22). This was reported in HG-28 (https://ncbijira.ncbi.nlm.nih.gov/browse/HG-28) and originally identified by paired end mapping. The CHM1 PacBio data supports the HuRef path in this region, rather than GRCh37. **NOTE:** This assembly region has not been corrected in GRCh38.

In this example, it appears that the individual CHM1_1.1 WGS contigs are correct, but it is their order and orientation in the assembly that is incorrect.

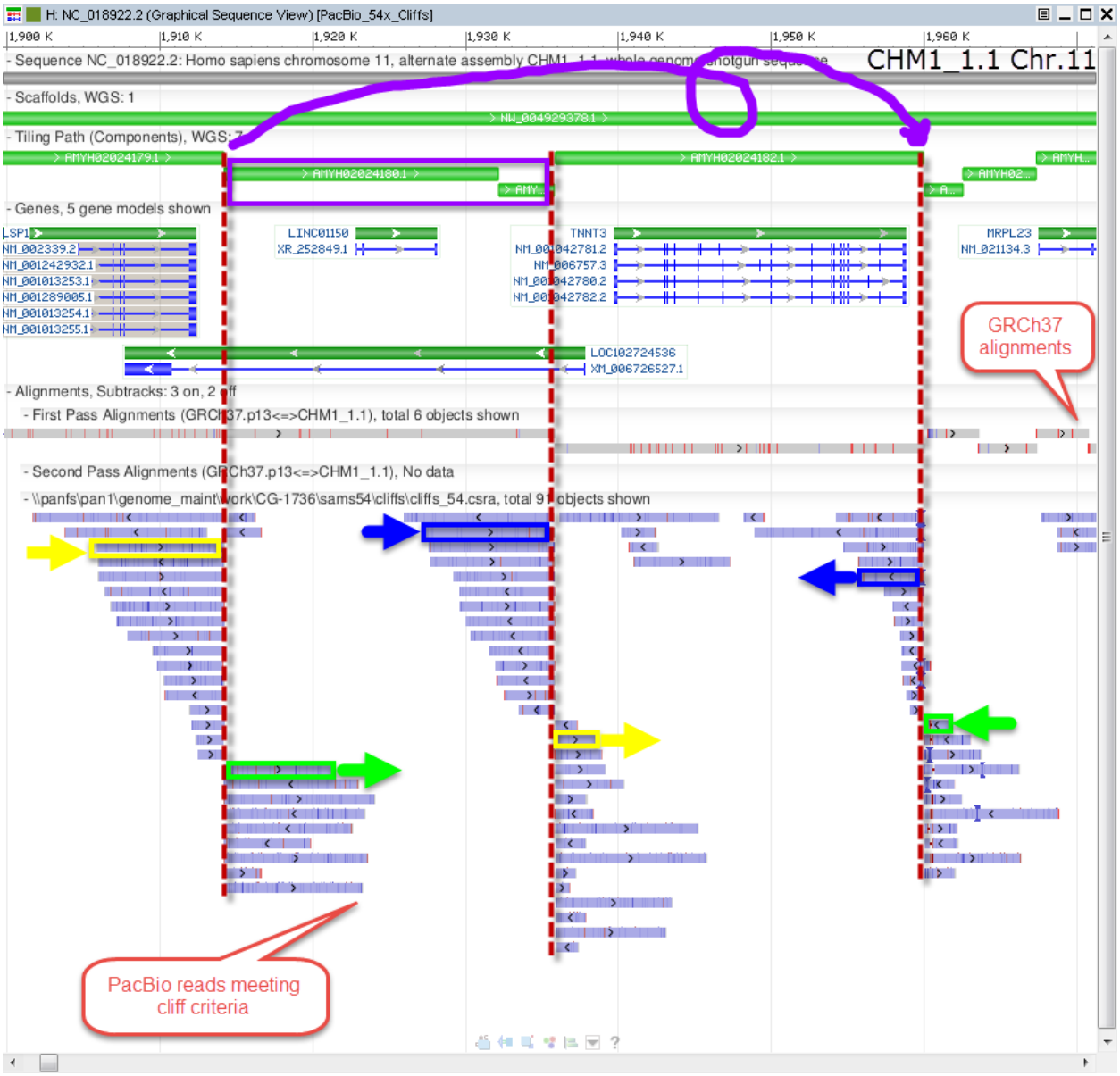

Figure 10: Overview of the NC_018922.2 1.9Mb region, exhibiting three alignment bins with a large number of "cliff" reads. WGS component boundaries flanked by such reads are marked with red dashed lines. Pairs of alignments corresponding to 3 different PacBio reads are marked in yellow, green and blue. These alignments overlap by <10% on each the reads (not shown). The split alignments for these 3 reads suggest that the two WGS components marked in purple should be inverted and translocated as indicated by the arrow at the top of the image. The other PacBio reads in these bins exhibit the same pattern of split alignments, which supports the proposed reordering and orientation of the WGS components.

NC_018934.2 (CHR. X) 149.45 MB: ERROR IN MISASSEMBLED AMPLICONIC REGION PROPAGATED TO CHM1_1.1 (HG-1459)

Another triple of bins containing a large number of cliff reads occurs on NC_018934.2 (chr. X) at 149.45 Mb. The edges of alignments in these bins do not correspond to WGS component boundaries.

#chr	1kb_bin	cliff_edges	comp	depth
chrX	149445000	16	M	88
chrX	149446000	14	M	79
chrX	149460000	21	M	69

Review of the PacBio read alignments suggests that a 14 kb region, encompassing portions of both AMYH02040764.1 and AMYH02040765.1 (but neither contig in its entirety) should be inverted. Additionally it appears that a 1kb region found on both contigs may be a false duplication and should be collapsed (purple triangles in figure).

Analysis of the GRCh37 assembly alignments shows that this region falls within one of the ampliconic regions on chr. X that was subsequently updated by an RP11 single haplotype fix patch (NW_004070890.2). For more information, see https://ncbijira.ncbi.nlm.nih.gov/browse/HG-1459. The first pass alignments in this region correspond to the GRCh37 reference chromosome, consistent with this sequence being used as the guide for the CHM1_1.1 assembly. However, there is a second pass alignment in this region that belongs to the fix patch. The inversion detected by the CHM1 PacBio reads corresponds to a similarly located inversion found in the fix patch alignment (light blue circle in figure). These data suggest that the GRCh37 assembly was incorrectly assembled in this region; this error was propagated to the CHM1 assembly. Had the fix patch been used to guide the assembly, it would be more consistent with the PacBio read sequences. Of note, this misassembly occurs *within* WGS components. However, this is not wholly surprising, as the highly repetitive nature of this genome region might be expected to confound the WGS assembly process.

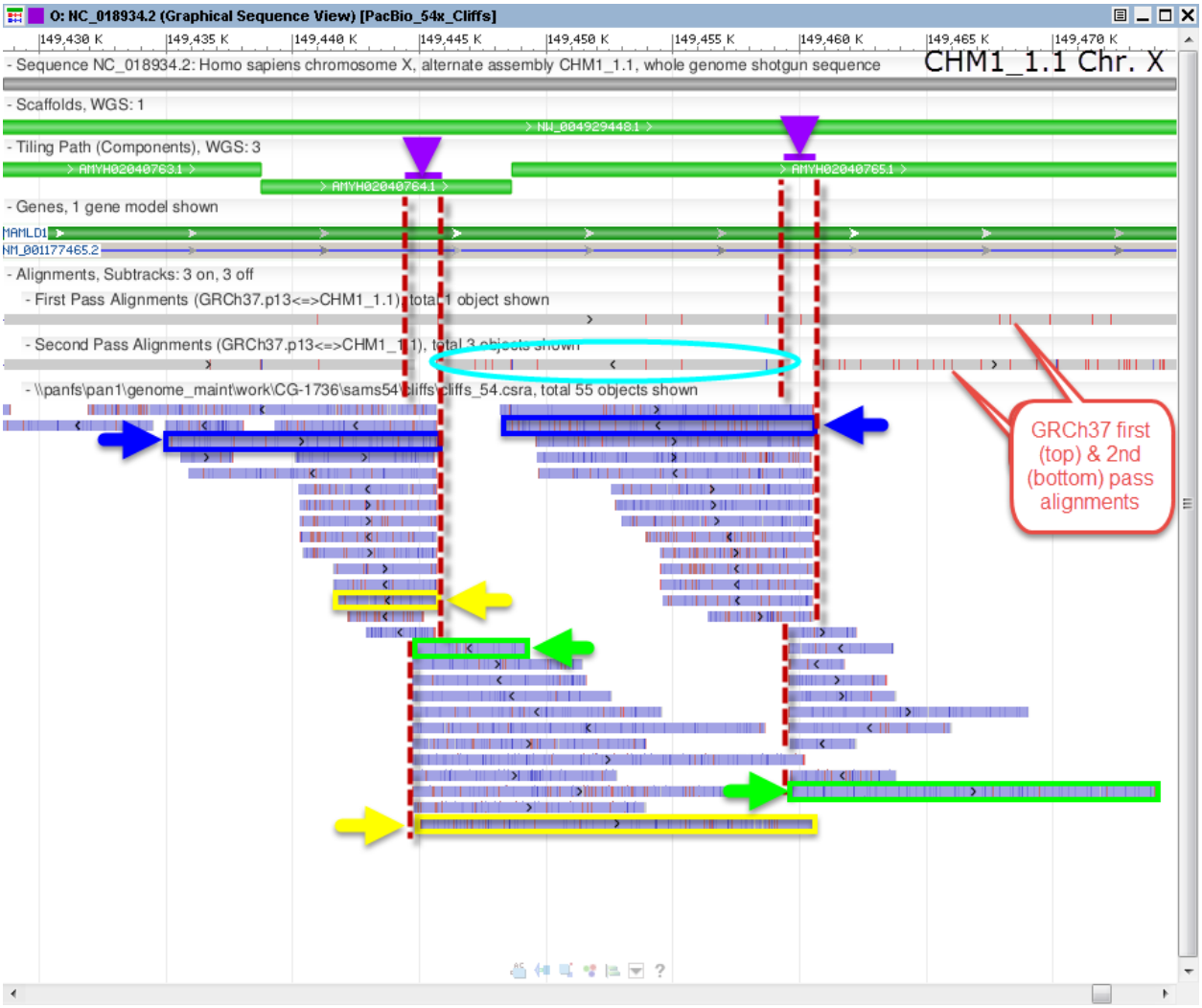

Figure 11: The NC_018934.2 region at 149.45Mb. Three alignment bins with large numbers of reads exhibiting cliff properties were identified. Alignment pairs for three different PacBio reads with non-overlapping alignments have been marked in blue, yellow and green. These alignments overlap by <10% on the PacBio read (not shown). However, there is some overlap (~1 kb), which is highlighted on the chromosome with purple triangles. Other reads in these bins exhibited the same pattern of alignments. The alignment patterns suggest that the central region should be inverted and the duplicated regions (purple) collapsed. First and second pass alignments of GRCh37 to the CHM1_1.1 assembly are shown.