**A.5 Comparison with Older Reference (CHIMPANZEE)**

We compared our assembly to the chimpanzee genome assembly panTro5 (GenBank: GCA\_000001515.7). Normally, a more contiguous assembly adds sequence as it replaces N’s with bases and extends and joins contigs. This was the case with our gorilla and orangutan assemblies and with our chimpanzee assembly compared with the 2010 panTro3 reference. However, the two newer chimpanzee panTro5 assemblies (which are nearly identical to each other) are an exception, having 30 to 60 Mbps of additional sequence relative to our mainly long-read assemblies (see Table DG How many bases were added or removed from older assemblies).

**Table DG How Many Bases Were Added or Removed from Older Assemblies.** These figures do not include N’s so inaccuracies in the number of N’s have no effect. For example, if the old assembly had 100 N’s, and our assembly replaced those 100 N’s by 150 bases, that would contribute 150 bases to the “Bases added to old genome” number.

|  |  |  |  |
| --- | --- | --- | --- |
| Old genome | Eichler lab genome | Bases added to old genome (Mbp) | Bases removed from old genome (Mbp) |
| panTro5.1 | Clint\_PTRv1 | 3.5 | 33.3 |
| panTro5.0 | Clint\_PTRv1 | 6.9 | 27.2 |
| panTro3 (Oct 2010) | Clint\_PTRv1 | 45.1 | 2.6 |
| ponAbe2 | Susie\_PABv1 | 54.5 | 3.8 |

To determine the amount of sequence added (or removed) by Clint\_PTRv1 relative to panTro5, we divided all panTro5 chromosomes into 10 Mbp sequences and aligned each 10 Mbp sequence to the corresponding Clint\_PTRv1 chromosome and unplaced contigs yielding a set of alignment blocks for that 10 Mbp sequence. (BLASR git hash 7cc3379a Nov 2016 version was run with parameters -bestn 1000 -clipping soft -alignContigs -sam -minMapQV 30 -minPctIdentity 50.) If each alignment block is C Clint\_PTRv1 bases long and P panTro5 bases long and contains N N’s within the aligned panTro5 bases, then the number of bases added is B = C - (P-N), the increase in number of non-N’s. This formula is insensitive to panTro5 over or underestimating the number of N’s of a gap. To avoid counting the same bases twice, the alignment blocks were sorted by size, the largest analyzed first, and alignment blocks were only analyzed if they did not overlap panTro5 bases of any of the previously considered alignment blocks. This resulted in ignoring less than 4% of the 2.6 Gbp of aligned panTro5 bases.

Adding all negative B's gives the amount of sequence removed by Clint\_PTRv1 with respect to panTro5, 27.8 Mbp (not including N’s), and adding all the positive B's gives the amount of sequence added, 6.9 Mbp, implying that panTro5 has more than three times as much extra sequence than it has missing sequence. This is different from orangutan (below) as well as in a comparison between Clint\_PTRv1 and the 2010 panTro3 which all fit the expected pattern. Furthermore, these numbers should be considered lower limits: if, for example, a single alignment block added 10 kbp of sequence to panTro5 in one place and removed 15 kbp in another place, the net removal of 5 kbp would be used in the sum.

A separate cruder analysis of just indels >=1 kbp found ~60 Mbp (including N’s) inserted in panTro5 relative to Clint\_PTRv1. (It was crude in that it counted, for example, a cigar of 1I1D1I1D as a 2 bp insertion, ignoring the neighoring deletions.)

We visualized 100 randomly chosen alignment blocks in MUMmer. In this sample there were roughly 10 times as many alignment blocks with more panTro5 bases than Clint\_PTRv1 bases.

****

**Fig DG 8: Example of duplicated sequence in panTro5.** The size of the panTro5 sequence is 2,307 bp larger than that of the aligned Clint\_PTRv1 sequence due to the duplicated sequence as indicated. (SeeFig DG False gap for more explanation.) There are 25 N’s inserted at this location. panTro5 CM000315.3:35656976-35711619 is aligned using MUMmer against Clint\_PTRv1 000121F\_1\_7613180:3755729-3808040.

****

**Fig DG 8.1: Example of misassembled and duplicated sequence in panTro5.** The size of the panTro5 sequence is 11,583 bp larger (not including N’s) than the aligned Clint\_PTRv1 sequence due to the duplicated sequence (indicated) and due to the duplicated and inverted sequence (red lines, indicated). This region contains 475 N’s in seven gaps. panTro5 CM000332.3:28173-80754 is aligned using MUMmer against Clint\_PTRv1 001655F\_1\_57369:11230-51753.



**Fig DG 8.2: Clint\_PTRv1 adds bases to panTro5.** The additional sequence is the size of the displacement between the 2 black lines: 9298 bp. The panTro5 assembly only had a 25 N’s at this location. PanTro5’s CM000314.3 position 5087202-5307586 is aligned by MUMmer against Clint\_PTRv1’s contig 000193F\_1\_4106580 position 3878731-4108388.

 

**Fig DG Histogram of insertion sizes in Clint\_PTRv1 and in panTro5**. Blue (nearly always higher) represents insertions due to excess bases (including N’s) in panTro5 relative to Clint\_PTRv1. Red indicates insertions due to excess bases in Clint\_PTRv1. For example, if a particular location in panTro5 has 5,000 excess N’s and 5,000 excess bases relative to Clint\_PTRv1, then the alignment would have a 10 kbp insertion in panTro5 and would make the blue bar 1 higher at x position 10. Insertions of length <100 bp are excluded.



**Fig DG Histogram of insertion sizes in Clint\_PTRv1 and the 2010 panTro3.** Compare this with fig DG Histogram of Alignment Gap Sizes Between Clint\_PTRv1 and panTro5 (above) which has many more insertions in panTro5. (If N’s were not included, then the red bars would always be higher, as indicated by Table DG How Many Bases Were Added or Removed from Older Assemblies.)

Two spot checks were made in which there was an insertion of over 20 kbp in panTro5. Two kbp of Clint\_PTRv1 sequence consisting of 1 kbp on either side of such an insertion was compared against the sequence at the corresponding location in the 2010 PanTro3 assembly. In both cases the sequences matched with 99.6% accuracy over all but the terminal 60 bp of the alignment—the panTro5 inserted sequence was not present. Clint\_PTRv1 had 70 and 100X (respectively) of PacBio reads spanning the location of the panTro5 insertion.

Examining several BLASR alignments at the base-pair level (e.g., Fig DG Insertion type BNB) confirms that the BLASR alignments are generally accurate and that panTro5 has extra sequence.

The insertions in panTro5 were put into categories as shown by Fig DG Categories of panTro5 insertions. The most common was BNB, probably where two contigs were joined, indicating that much of the additional sequence in panTro5 was at the ends of contigs.



**Fig DG Categories of panTro5 insertions.** BNB means BB..BBNN..NNBB..BB (a sequence of non-N’s followed by a sequence of N’s followed by a sequence of non-N’s). NB means NN.NNBB..BB. NBN means NN..NNBB..BB. complex means some other pattern including, for example, BB..BBNN..NNBB..BBNN..NNBB..BB See Fig DG Insertion Type BNB.



**Fig DG Insertion type BNB.** An alignment of a contig from panTro5 (top sequence) and from Clint\_PTRv1 (bottom sequence) with matching, mismatching, and alignment gaps annotated by "|", "\*", and "-", respectively. Within the panTro5 assembly, there is missing sequence denoted by 25 N's, where two contigs were likely joined. The insertion continues many lines upwards off this figure. This could indicate that the panTro5 contig before the N’s has extraneous bases on its right side. Shown is part of the alignment between panTro5 CM000336.3:44,101,847-46,821,596 and Clint\_PTRv1 000228F\_1\_3426367:13-2,714,711. Clint\_PTRv1 has read depth >18X everywhere in this region.

We looked for regions in which the panTro5 assembly differed from Clint\_PTRv1 due to mismatching bases. Spot checks of 29,853 non-overlapping 1 kbp base windows showed that only 17 such windows (0.06%) had similarity of less than 80% (excluding differences due to alignment gaps). Either the two assemblies matched very well, or else they differed by a large insertion or deletion.

**Gaps in panTro5**

panTro5 contains 27,797 gaps (runs of N’s) with a mean number of N’s of 3.4 kbp. We aligned panTro5 to Clint\_PTRv1 to identify gap closures. A panTro5 gap is considered closed if a Clint\_PTRv1 contig is aligned at that location in panTro5, including both sides of the run of N’s. In total, 52% (14,518) of the 27,797 panTro5 gaps were closed in Clint\_PTRv1. This is markedly different from the 96.8% gaps closed by our assembly of orangutan. The lower figure in chimpanzee is possibly due to incorrect sequence in panTro5 near runs of N’s preventing Clint\_PTRv1 contigs from aligning on both sides of the run of N’s.

We sought to identify the new sequence in Clint\_PTRv1 that bridges gaps (runs of N’s) in panTro5. For this purpose, we extracted 2 kbp of sequence from each side of a gap (see Fig DG 9.0) and attempted to align this sequence to Clint\_PTRv1 using BLASR. We applied a series of filters (details in table SM 6.1 and Fig DG 9.0 below) to converge on a set of gaps that could be reliably tracked to the Clint\_PTRv1 assembly. This method resulted in only 9% (2,508) of the total 27,797 gaps passing all filters (see Table SM 6.1 Flank alignment filter and Fig DG 9.0 Method of determining the sequence of runs of N’s in panTro5). In the other 91% of gaps, if the run of N’s was bridged by a Clint\_PTRv1 contig, the sequence flanking the N’s in panTro5 did not match Clint\_PTRv1 so not only did Clint\_PTRv1 remove or replace the N’s but the surrounding sequence was replaced as well. True gaps are the only runs of N’s in which the N’s are replaced by the correct nucleotides (possibly more or less than the number of N’s). In the 20% of gaps labelled “false gaps” the flanking sequence mapped but in reverse order and same strand (see Fig Gordon False gaps). In most of these cases there was neighboring duplicated sequence and no missing bases, contrary to the presence of N’s (see Fig Gordon Common Cause of False Gaps). All true gaps lie in the BioNano scaffolded regions.

****

**Fig DG 9.0: Method of determining the sequence of runs of N’s in panTro5.** In this case the “old assembly” is panTro5 and the “new assembly” is Clint\_PTRv1 but the same method is used for orangutan. 2 kbp of the old assembly (green) on each side of the run of N’s is aligned to the new assembly. The sequence in the new assembly (red) between those two alignments is the sequence of the run of N’s, providing the alignments pass the filters listed in Table SM 6.1. In panTro5 only 9% of runs of N’s fit this model due to incorrect sequence neighboring the N’s.

****

**Fig Gordon false gaps.** Compare with Fig DG 9.0 Method of determining the sequence of runs of N’s in panTro5. For the gaps that pass all of the filters (see Table SM 6.1), the flanking sequence may map in the same order in the new assembly as in the old assembly. Such gaps, “true gaps”, will have B'-A', the gap size, positive, as in Fig DG 9.0. But as in the case shown here, some gaps, “false gaps,” (the majority in panTro5) map in reverse order to the Falcon contig, and B'-A' is negative. (There are corresponding rules if the new assembly contig and the old assembly contig are opposite strands.)

****

**Fig DG Common cause of false gaps.** An examination of 98 sample false gaps showed that sequence was erroneously duplicated in the reference in 87% of cases. However, many of those cases were more complicated, such as having many duplications, duplications both in the old and in the new assemblies, palindromic sequence, or a background of dozens of repeats. See Fig DG False Gap for a simple example.

**Table SM 6.1 Flank alignment filters.** These filters are necessary to obtain gaps (runs of N’s) whose size can be determined. For each alignment as shown in Fig DG 9.0, the filters listed in the table are successively applied giving the number of gaps and the percent of all gaps passing all filters up to that point. The “whole contig alignment” filter means that the contig in Clint\_PTRv1 bridging the run of N’s must be the same contig (and at the same position) as the contig mapped to by the 2 kbp flanks of N’s in panTro5. “\*” indicates not a filter.

|  |  |  |
| --- | --- | --- |
| Alignment Filter condition | Recovery (gaps) | Recovery (%) |
| number of gaps | 27,797 | 100 |
| both flanks are mapped | 26,982 | 97 |
| both flanks map to the same new assembly contig | 22,042 | 79 |
| both flanks map to the same strand of contig (above) | 11,193 | 40 |
| \*both flanks do not map to the same strand | 10,849 | 39 |
| whole contig alignment |  8,137  | 29 |
| \*false gaps | 5,629 | 20 |
| true gaps | 2,508 | 9 |

We noticed that the filter that removes the largest percent of gaps is requiring that both flanks map to the same strand (39% of the original gaps are eliminated, contrasting with orangutan in which this filter removed 0.19% of gaps). To confirm this, we examined the dotplots of 195 such cases (Figure DG 10.0: Dotplot of Clint\_PTRv1 versus panTro5 near a run of N’s) and saw the sequence adjoining a run of N’s is in the wrong orientation in panTro5 about 50% of the time. We passed this information to the submitters of panTro5 who found that 2,990 contigs and 1,505 pieces of contigs needed to be inverted, making version panTro5.1. Not all of the putative inversions were fixed in panTro5.1. A sequenced BAC spanned 2 gaps in panTro5 that were not fixed in panTro5.1. For each of these gaps, the left flank and right flank mapped by cross\_match to opposite strands of the BAC, indicating that sequence was still inverted in panTro5.1. Miropeats (Miropeats: graphical DNA sequence comparisons, J.D. Parsons, Bioinformatics, Volume 11, Issue 6, 1 December 1995, Pages 615–619, https://doi.org/10.1093/bioinformatics/11.6.615) (Fig DG Miropeats) also shows this inversion as well as an insertion of (probably incorrect) sequence in panTro5.





**Figure DG Miropeats of a BAC aligned against panTro5 (top panel) and aligned against Clint\_PTRv1 (bottom panel)**. BAC CH251-72H6 was PacBio sequenced to ~1770X (sic) coverage. The top panel, BAC CHM251-72H6:150000-187962 aligned against panTro5 CM000329.3:68020000-68064272, shows two types of common problems in panTro5 that are fixed by Clint\_PTRv1: inverted sequence and extraneous sequence. The inversion is flanked by N’s on both sides suggesting it was a contig that was inserted into the chromosome in the wrong orientation. The purple bar “copy2” is a ~8 kbp insertion in panTro5 which (according to a GenBank blast search against nr/nt) best matches to this very BAC at location “BAC copy” with 89.6% identity. “copy1” (the correct sequence) matches “BAC copy” at 99.6% identity. This suggests that “copy2” is artifactually duplicated sequence. Green arrows indicate LINES, purple SINES, and yellow LTRs. The bottom panel shows the same BAC region aligned very closely against Clint\_PTRv1 000109F\_1\_7982668\_quiver\_pilon:2067235-2104235. According to cross\_match, this alignment is 99.9% identical.

****

**Figure DG 10.0 Dot plot of Clint\_PTRv1 versus panTro5 near a run of N’s.** In this example 100 N’s lie between the two very close horizontal blue lines. The regions immediately flanking these N’s (the black and red dotted lines) are in opposite orientations from each other. There is another run of N’s (not indicated) above the black dotted line so the black line was likely a contig that was inserted in the correct location but incorrect orientation. An examination of a random sample 98 of the 10,849 runs of N’s in panTro5 in which flanks map in opposite orientation to the same Clint\_PTRv1 contig showed that about 56% of these clearly had such an inversion relative to the Clint\_PTRv1 contig. To increase confidence that the Clint\_PTRv1 assembly is correct at these locations, we checked the PacBio read depth. Out of these 10,849 locations, only 307 cases had a read depth less than 5 or greater than mean plus 2 standard deviations (252). The location shown is 000561F\_1\_298860\_quiver\_pilon:77108-89108 in Clint\_PTRv1 aligned against CM000329.3:3243549-3255549 in panTro5.



**Figure DG 11.0: Example of a True Gap in panTro5.** There are 10 N’s in panTro5 at the blue line. Notice the large breaks in the red line, indicating that the panTro5 sequence does not match the Clint\_PTRv1 sequence. The largest such break in the red line overlaps the blue line, indicating that the panTro5 sequence at the position of the green segment (which includes the 10 N's) should be replaced by the sequence in Clint\_PTRv1 indicated by the ~800 bp purple segment. Thus, rather than just replacing the N’s with correct nucleotides, Clint\_PTRv1 is correcting a misassembly in the ~800 bp region surrounding a 10 N’s. Reviewing a random sample shows this is typically the case with true gaps in panTro5. panTro5 CM000314.3:140035785-140041785 is aligned against Clint\_PTRv1 000083F\_1\_11068219\_quiver\_pilon:10701361-10707361.



**Fig DG False gap.** This is an example of a simple “false gap”. Even though there are 25 N’s in panTro3 (the blue line), the Clint\_PTRv1 assembly shows that there is in fact no gap at all but rather ~2500 bp is duplicated in panTro5. panTro5’s CM000316.3:66521488-66528346 is aligned against Clint\_PTRv1’s 000001F\_1\_57587592\_quiver\_pilon:44071726-44078584.

****

**Figure Gordon More Complex False Gap.** In this case there are dozens of little repeats in the vicinity of the N’s (the blue line). The diagonal red line above the blue line is offset from that below the blue line indicating there is an insertion in panTro5. CM000331.3:19474251-19480251 in panTro3 is aligned against 000087F\_1\_10942768\_quiver\_pilon:10638487-10644487 in Clint\_PTRv1. There are many varieties of complex false gaps.



**Fig DG Gap sizes.** The 29% of gaps (runs of N’s) whose actual size could be determined (see Table SM 6.1: Flank Alignment Filter) are shown: false gaps are in red, true gaps are in blue. The spike of false gaps of size -210 to -200 might be due to the panTro5 reference having extra copies of SINEs at these locations. There were 5,526 false gaps and 2,611 true gaps.

**B.4 Comparison with Older Reference (ORANGUTAN)**

We compared our assembly (Susie\_PABv1) to the latest Orangutan assembly on NCBI (GCF\_000001545.4), P\_pygmaeus\_2.0.2, “ponAbe2”. ponAbe2 has 315,124 gaps (stretches of N’s). Following the same procedure as for chimpanzee to determine whether a run of N’s in ponAbe2 has been resolved in Susie\_PABv1, 96.8% (305,069) of all runs of N’s were resolved.

To determine the amount of sequence added (or removed) by Susie\_PABv1 relative to ponAbe2 we followed the same procedure as with chimpanzee of dividing all ponAbe2 chromosomes into 10 Mbp sequences and aligning each 10 Mbp sequence to the Susie\_PABv1. Each such alignment yields several alignment blocks (see Fig. DG361). In total, Susie\_PABv1 added 54.5 Mbp and removed 3.8 Mbp from ponAbe2 (see Table DG How Many Bases Were Added or Removed from Older Assemblies).



**Fig. DG361: Alignment of ponAbe2 to Susie\_PABv1.** Light blue vertical lines are the locations of N’s in ponAbe2. Notice that Susie\_PABv1 has more sequence than ponAbe2 in this alignment, which we found to be true in most alignments. ponAbe2 NC\_012613.1:18652168-19193999 is aligned against Susie\_PABv1 000411F\_1\_851530\_s\_58669\_853372:6-604522.

To determine the sequence of each gap, we chose 2 kbp flanking sequence of each gap in ponAbe2, aligned it against the Susie\_PABv1 contigs (see a more detailed description under the Pan troglodytes gap analysis section) and applied a series of filters to identify those runs of N’s, 88.52% of all runs of N’s, whose corresponding sequence could be reliably located in Susie\_PABv1. Gaps recovered in each filter is shown in table DG366 below.

**Table DG366: Gap Alignment Filters.** These filters are necessary to obtain gaps (runs of N’s) that can be reliably located in Susie\_PABv1. The filters listed in the table are successively applied giving the number of gaps and the percent of all gaps passing all filters up to that point. The “whole contig alignment” filter means that the Susie\_PABv1 contig that the flanking sequence aligns to is the same Susie\_PABv1 contig that aligns at this location of the reference sequence including both sides of the gap. “\*” indicates not a filter.

|  |  |  |
| --- | --- | --- |
| Alignment Filter condition  | Recovery (gaps) | Recovery (%) |
| number of gaps | 315,123 | 100.00 |
| both flanks are mapped | 294,780 | 93.54 |
| both flanks map to the same falcon contig | 279,571 |  88.71 |
| both flanks map to the same strand  | 278,961  | 88.52 |
| \*both flanks do not map to the same strand | 610 | 0.19 |
| whole contig alignment | 270,447  | 85.82 |
| \*false gaps | 97,809  | 31.03 |
| true gaps |  172,638  | 54.78 |

As explained under chimpanzee section, in some cases the flanks map to the new assembly contig in reverse order (but same strand) indicating there is a misassembly rather than a gap. This occurred with 97,809 gaps (31.03%) while there were 172,638 (54.78%) true gaps. It was the latter that we used to access the gap sequence. Size distributions of the true and false gaps are shown in Fig DG367.



**Fig DG367: Orangutan Gap Sizes.** The 85.82% of gaps (runs of N’s) whose actual size could be determined (see Table DG366: Gap Alignment Filters) are shown: false gaps are in red, true gaps are in blue.

We determined that 52.51% of the sequence that closed gaps were made up of repeats.