
Supplementary information

Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies

In the format provided by the
authors and unedited

Supplementary Tables

Supplementary Table 1..... 2
Supplementary Table 2..... 3
Supplementary Table 3..... 4

Supplementary Table 1. Correcting invalid ORFs via polishing. This considers the classification of GRCh38 transcripts with valid ORFs.

Query Assembly	True Negatives	False Positives	True Positives	False Negatives
CHM13v1.0	69123	0	16	611
Racon	69101	22	19	608
Racon+Merfin	69123	0	14	613

True/False Positives/Negatives for each query assembly were assessed as follows: If a transcript has a valid ORF in the CHM13v0.9 assembly and it remains valid in the query assembly, it is a "True Negative". If a transcript has a valid ORF in the CHM13v0.9 assembly but it has an invalid ORF in the query assembly, it is a "False Positive". If a transcript has an invalid ORF in the CHM13v0.9 assembly and it remains invalid in the query assembly, it is a "False Negative", and if a transcript has an invalid ORF in the CHM13v0.9 assembly but it has a valid ORF in the query assembly, it is a "True Positive". Only transcripts that mapped to all assemblies were considered.

Supplementary Table 2 | Remaining issues identified from both PacBio HiFi and ONT

CHM13 v1.0			CHM13 v1.1			Reason
Chr.	Start	End	Chr.	Start	End	
11	54243938	54285164	11	54258631	54260218	Low consensus quality
16	36735111	36775613	16	36753902	36757498	Low consensus quality
	n/a		15	3642490	3643193	GA sequencing bias in model rDNA sequence
	n/a		15	3732691	3732910	GA sequencing bias in model rDNA sequence
22	39136546	39145169	22	39107778	39116401	Collapsed low complexity sequence
19	28509243	28519604	19	28509246	28519607	Chimeric consensus of two haplotypes
19	28527178	28538917	19	28527181	28538920	Chimeric consensus of two haplotypes

Supplementary Table 3. Low coverage regions detected only from HiFi alignments. Regions with <7x Winnowmap primary read alignments were collected and categorized given the mapping quality, alignment identity, and sequence context (% microsatellites within 10 kb).

Cause of low coverage	Num. of regions (%)
Likely low consensus quality	11 5.0%
Low consensus quality	30 13.8%
AT biases	7 3.2%
GA or TC biases	170 78.0%