Approaching an Error-Free Diploid Human Genome

Yanan Chu[1*], Zhuo Huang[1,2*], Changjun Shao[1*], Shuming Guo[3*], Xinyao Yu[1,2], Jian Wang[1], Yabin Tian[4], Jing Chen[1], Ran Li[5], Yukun He[5], Jun Yu[2#], Jie Huang[4#], Zhancheng Gao[3,5#], Yu Kang[1,2#].

1. Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Center for Bioinformation, Beijing, 100101, China.

2. University of Chinese Academy of Sciences, Beijing, 100049, China.

3. Linfen Clinical Medicine Research Center, Institute of Chest and Lung Diseases, Shanxi Medical University, Linfen, 041000, China

4. State Key Laboratory of Drug Regulatory Science, National Institutes for Food and Drug Control, Beijing, 100050, China

5. Department of Respiratory and Critical Care Medicine, Human Genome Research Center, Peking University People's Hospital, Beijing 100044, China

Correspondence to: Yu Kang, kangy@big.ac.cn, Zhancheng Gao, zcgao@bjmu.edu.cn, Jie Huang, jhuang5522@nifdc.org.cn, Jun Yu, junyu@big.ac.cn.

Abstract

Achieving a truly error-free human diploid genome remains a central challenge in genomics. Here, we present T2T-YAO v2.0, a nearly perfect diploid genome assembly for a Han Chinese individual. We first developed Sufficient Alignment Support (SAS), a sensitive alignment-based metric that identifies both structural and base-level assembly erroneous windows by quantifying inadequate read support across sequencing platforms. Unlike conventional methods that detect erroneous k-mers (*e.g.*, Merqury) or inconsistencies between assemblies and aligned reads from single platform, SAS integrates supportive information from various platform and avoids false positives caused by sequencing bias and artificial alignments. Employing a "Structural Error-first" polishing strategy, prioritizing structural error correction using ultra-long Oxford Nanopore reads, we iteratively refined the assembly with high-accuracy PacBio and Element sequencing data. This approach eliminates all detectable structural and non-homopolymer-related base-level errors outside

ribosomal DNA clusters. The final T2T-YAO v2.0 assembly exhibits complete non-split ONT read coverage and free of erroneous 21-mers outside sequencing data, representing an error-free human genome by Merqury standards. Our results not only highlight the limitations of current genome quality assessment methods but also provide a roadmap for constructing high-fidelity diploid references essential for precision medicine and genome-based diagnostics.

Main Text

Introduction

Advancements in sequencing technologies and assembly algorithms have rapidly improved the quality of genome assemblies, culminating in the release of the first telomere-to-telomere (T2T) human reference genome, T2T-CHM13, by the Telomere-to-Telomere (T2T) Consortium in 2022[1,2]. As precision medicine advances, there is a growing demand for a diploid genome assembly with perfect base-level accuracy to serve as a benchmark for both sequencing and variant calling[3-6]. However, assembly errors can still arise from multiple sources, including base-calling inaccuracies, incorrect haplotype assignment of contigs, imperfect read correction, and homopolymer compression during the assembly process[7,8]. These errors must be corrected *post hoc* by realigning sequencing reads to the assembly and polishing regions of inconsistency.

Currently, a variety of structural and single-nucleotide variant (SNV) callers can be applied in parallel to identify and correct assembly errors based on alternative allele detection. The accuracy of the resulting polished assembly is typically evaluated using k-mer–based metrics such as Merqury, which quantifies erroneous k-mers absent from the raw sequencing data[9]. However, this polishing-evaluation strategy remains not efficient or accurate enough to achieve a truly error-free human genome. Several limitations in current polishing strategies are apparent. First, there is no comprehensive and reliable method to identify assembly errors that can effectively guide correction. K-mer–based approaches are especially inadequate for high-quality assemblies, as they tend to miss errors that do not introduce novel k-mers, such as homopolymer length mismatches. Moreover, sequencing errors can obscure genuine erroneous k-mers, making QV values highly sensitive to sequencing depth and platform-specific error profiles. Second, uncorrected structural errors can

compromise read alignment, leading to failed corrections or overcorrection at the base level. Third, no streamlined and accurate pipeline exists for correcting base-level errors: false positives and false negatives are inevitable with current SNV callers. These tools are typically optimized for haploid references and single-platform datasets, often resulting in suboptimal performance on diploid genomes and posing challenges for integrating calls across multiple sequencing platforms.

To address these issues, we developed an alignment-based quality assessment metric called Sufficient Alignment Support (SAS), which directly identifies regions of the assembly lacking adequate read support, an intrinsic indicator of assembly error. Building on this, we implemented a "Structural Error-First" (SE-First) polishing strategy, which prioritizes the correction of structural errors using ultra-long Oxford Nanopore Technology (ONT) reads to improve the alignment quality for subsequent base-level corrections using shorter, high-accuracy reads. Starting from our previously published T2T-YAO v1.1, an initial complete diploid genome assembly of a Han Chinese individual[10], we incorporated additional sequencing data from high-accuracy platforms and iteratively applied SAS evaluation and SE-First polishing. This approach yielded T2T-YAO v2.0, a near-error-free diploid human genome, with the exception of residual issues in ribosomal DNA (rDNA) clusters and a small number of homopolymer-associated errors in repetitive regions. Based on Merqury analysis, T2T-YAO v2.0 contains no erroneous 21-mers absent from the raw sequencing data, yielding an "infinite" quality value (QV) by k-mer standards, the first such result for a diploid human genome.

Results

SAS: an alignment-based metric for assembly errors

To systematically evaluate the quality of genome assemblies, we developed a sensitive alignment-based metric termed Sufficient Alignment Support (SAS). Rather than detecting inconsistencies between the assembly and aligned sequencing reads, SAS identifies assembly errors by flagging genomic windows that lack sufficient support from mapped reads. In SAS, base-level errors (BEs; <50 bp) and structural errors (SEs; ≥50 bp) are quantified independently (**Fig. 1**).

For BEs, the diploid genome assembly is divided into 50 bp windows. Windows are flagged as

erroneous if they are not precisely supported by reads exceeding 10% of the overall mean sequencing depth across any platform. This threshold is relaxed to 10% of the local sequencing depth in specific low-coverage contexts: (1) regions within 10 kb of chromosomal ends (*i.e.*, telomeres); (2) sequences containing degenerative GnA or CnT repeats (GA/CT content ≥90% and GC content ≥60%); or (3) GC-rich regions (GC content ≥90%), which are known to be underrepresented in polymerase-dependant, high-accuracy sequencing platforms (*e.g.*, Element, Illumina, and PacBio platforms) and rely predominantly on ONT reads for coverage.

SEs are assessed in 2 kb non-overlapping windows using ultra-long ONT reads for its steady coverage without obvious sequencing bias. A window is considered structurally erroneous if it meets one of the following conditions: (1) sequencing depth is outside the 20% to 200% range of the global average; or (2) fewer than 50% of local-depth reads span the window without large indels (≥50 bp). This threshold is relaxed in chromosomal termini (within 10 kb), where ≥10% of the local-depth reads must provide contiguous support. Additionally, to improve detection of misjoin events, 30 kb sliding windows (stepped every 2 kb) are evaluated. If no reads span a given window, the corresponding 15 consecutive 2 kb windows are flagged as SEs. Mitochondrial sequence, which is small (~16 kb) and typically free of structural errors, are excluded from this analysis.

The overall SAS-based quality value (QV_{SAS}) is calculated by summing the BE and SE windows and computing the predicted error rate (E) using the formula:

$$QV_{sas} = -10 \cdot \log E$$
$$E = \frac{BE + SE}{L}$$

Where E is the estimated error rate of the assessed assembly, BE and SE are the number of BE and SE windows, respectively, L is the base length of the assembly.

To accurately evaluate the T2T-YAO assembly, we upgrated its raw data to high-accuracy sequencing platforms: 260 Gb Q28 ultra-long ONT reads[11] (filtered for >30 kb; N50 >120 kb), ~172 Gb Pore-C[12], 306 Gb HiFi reads (PacBio SPRQ; filtered >3 kb), and 320 Gb Element AVITI[13] Q50 reads (**Table S1**). To reduce ambiguity in haplotype assignment during mapping, ONT and SPRQ reads were phased in advance using haplotype-specific kmers (hapmers) and mapped in a haplotype-

118 aware manner.

119

120 We applied the SAS metric to evaluate the previously released T2T-YAO v1.1 assembly and
121 compared the results to those obtained using Merqury, a widely adopted k-mer–based quality
122 assessment tool (**Table 1**). SAS identified 21,511 base-level error (BE) windows and 4,438
123 structural error (SE) windows, yielding QV_{SAS} values of 53.12 and 54.13 for the maternal and
124 paternal haplotypes, respectively. Chromosome-specific QV_{SAS} scores ranged from 45.36 to 61.37
125 (**Table S2**). In contrast, Merqury using 21-mers reported only 22,556 erroneous k-mers,
126 corresponding to 1,074 estimated errors across the diploid genome. This number is notably higher
127 than initially reported[10], as the use of more accurate Element and SPRQ reads revealed assembly
128 errors that were previously obscured by sequencing errors. Nonetheless, the large discrepancy
129 between SAS and Merqury highlights the markedly greater sensitivity of SAS. Many BE windows
130 detected by SAS lacked corresponding erroneous k-mers, even when more sensitive 31-mers were
131 used, which identified 123,085 erroneous k-mers corresponding to 3,970 estimated errors (**Table 1**).
132 These results underscore the limitations of k-mer–based methods in capturing all types of assembly
133 errors. As a representative example, we showcased the paternal chromosome 1 to illustrate the
134 differences in error detection between SAS and other metrics (**Fig. 2a**).

135

136 Closer inspection of BE windows revealed a frequent association with homopolymer sequences,
137 where reads often disagreed on the number of repeated nucleotides. To classify these regions, we
138 defined homopolymer sites as stretches of ≥10 A/T, ≥7 G/C mononucleotide repeats, or ≥10
139 dinucleotide repeats. Accordingly, BEs were divided into homopolymer-associated (h-BEs) and
140 non-homopolymer-associated (nh-BEs), accounting for 17.17% (3,694) and 82.83% (17,817) of all
141 BEs, respectively (**Table 1**). While nh-BEs were relatively enriched in centromeric regions and
142 corresponded with k-mer–derived errors identified by Merqury, h-BEs were more broadly
143 distributed outside centromeres and were largely undetected by k-mer-based methods (**Fig.2**).
144 Conversely, some erroneous k-mers reported by Merqury were located outside SAS-flagged BE
145 windows and clustered within degenerative GnA or CnT repeat regions. These loci typically had
146 sufficient ONT coverage and were not flagged as errors by SAS, but lacked support from SPRQ and
147 Element reads due to platform-specific sequencing bias. This discrepancy underscores a

fundamental limitation of k-mer–based assessment tools: their inability to detect errors that do not introduce novel k-mers, such as homopolymer length discrepancies, and their susceptibility to false positives in regions with biased or zero coverage. The issue is particularly severe in the current version of Merqury due to its reliance on the hybrid k-mer sets derived solely from high-accuracy Illumina and PacBio platforms (Element in this study), excluding ONT data due to its relatively higher sequencing error rate.

Given that Merqury cannot detect structural errors effectively, we compared SAS against several structural variant (SV) callers: Sniffles2[14] (ONT), Flagger[15] (ONT), and nucflag[16] (SPRQ). Sniffles2 identified only 263 confident SV-like errors (supported by ≥60% of local reads), demonstrating limited sensitivity, particularly in repetitive regions (**Extended Fig. 1**), where it relies on repeat annotations often unavailable in *de novo* assemblies. Flagger and nucflag are widely inconsistent with each other, that they identify 4.03 Mb and 12.30 Mb of collapsed regions (>50 bp), respectively, with only 1.08 Mb of overlap (**Table S3**). Notably, 83.63% of their overlapping regions are either within or adjacent (<10 kb) to SE windows identified by SAS. Furthermore, 78.50% of low-coverage regions jointly flagged by Flagger ("erroneous" and "duplication" regions) and nucflag ("misjoin" regions) are also detectable by SAS. We also observed that nucflag frequently misclassifies misjoins in GnA/CnT repeats and other regions with zero SPRQ coverage due to sequencing bias (**Fig. 2**). The obvious bias of SPRQ in specific sequences damages the reliability of SPRQ-based SE callers. Haplotypic switch errors, which are flagged by nucflag as heterogeneous regions and by Merqury as switch k-mers, are not treated as a separate error class in SAS. Depending on the read evidence, such events are classified as SEs if flanking clips are present or as BEs if phased read support is available but without a clear breakpoint.

Correction of Structural Errors

Using SAS and ultra-long ONT reads, we identified 4,438 non-redundant SE windows in the T2T-YAO v1.1 assembly. These windows were further categorized based on their underlying features: 1,295 were high-depth regions (29.18%), 1,662 low-depth regions (37.45%), 331 contained indels (7.46%), 61 were marked by clipped alignments (1.37%), and 1,929 were consecutive windows without spanning reads (43.46%). Note that individual windows could be assigned to multiple

178 categories. Errors related to clipped alignments and abnormal depth were particularly concentrated
179 in the short arms of acrocentric chromosomes and the centromeres of maternal chromosomes 6 and
180 7, whereas indels were more evenly distributed across the genome (**Fig. 3a**).
181
182 Given the critical importance of accurate read alignment for subsequent base-level error correction,
183 we adopte a "Structural Error-first" (SE-first) strategy. This approach prioritizes correction of
184 structural errors using ONT reads before applying base-level polishing with high-accuracy
185 sequencing data. By resolving large-scale misassemblies upfront, we reduce the risk of
186 overcorrection due to artificial alignments in shorter reads, thereby enhancing the overall fidelity of
187 the assembly.
188
189 We employed a comprehensive correction pipeline combining variant calls from Sniffles2, local
190 reassemblies using ONT reads, and targeted telomere extension to eliminate all detectable SEs. After
191 each correction round, reads were re-aligned, and SAS was reapplied to reassess error content
192 (**Extended Fig. 2**). Most SEs were resolved after several iterations of this correction–realignment–
193 evaluation loop, though certain clipped alignment sites required manual inspection and correction.
194 In telomeric and subtelomeric regions, ONT reads often exhibited heterogeneity that impeded
195 successful reassembly. In such cases, stratifying ONT reads based on characteristic SNVs and
196 assembling the dominant haplotype group enabled resolution. Following these polishing efforts, all
197 telomeres were extended to an average length of 8.7 ± 2.6 kb, with all exceeding 5 kb except for
198 three in the q arms of maternal chromosome 12 and paternal chromosomes 21 and 22 (**Fig. 4a**).
199 Notably, most telomeres contained some non-canonical tandem 6-mer repeats similar to
200 (TTAGGG)n[17], often interspersed with canonical motifs, particularly at the transition from
201 subtelomeric to telomeric regions, as previously described[18]. The structural integrity of the assembly
202 is further validated by the Pore-C chromosomal configuration data, with its contact map confirming
203 the haplotype phasing and the accurate pairing of short and long arms across the ten acrocentric
204 chromosomes, a particularly challenging structural feature to resolve (**Extended Fig. 3**).
205
206 Ultimately, T2T-YAO v2.0 becomes free of detectable SEs outside of ribosomal DNA (rDNA)
207 clusters (**Fig. 2b, 3b**). Centromeric collapses, particularly in maternal chromosomes 6 and 7, are

also resolved. Their corresponding higher-order repeat (HOR) arrays in T2T-YAO v2.0 are successfully extended to near-native lengths (**Fig. 4b, Extended Fig. 4**). One potential exception of high-depth is observed in a 12 kb region upstream of the *TSPAN33* gene on maternal chromosome 7q32. Although the local ONT depth (84–89×) marginally exceedes the global cutoff (84×), the sequence there is fully supported by all platforms without signs of collapse or misjoin. Further inspection reveals that this region is composed of LIM and Alu elements which sequences are widely present in human genome with millions of copies[19]. The repetitive nature of this region may be the source of the elevated depth where 22.8% of mapped reads are shorter than 60 kb and exhibiting MAPQ scores of 0–1 (**Extended Fig. 5**).

Aside from this, ONT read depth followed a Gaussian distribution, with extremely few regions tallied by Flagger as false duplication or collapsed (**Extended Fig. 6**). All overlapping issue regions identified by Flagger and nucflag are confined to rDNA arrays or telomeric regions, supporting the structural completeness of T2T-YAO v2.0 (**Table S3**). To further validate continuity, we assessed the fraction of the assembly covered by non-split ONT reads (NScov%). In T2T-YAO v1.1, NScov% is 99.98%, with ~0.91 Mb (excluding rDNA) left uncovered. In contrast, T2T-YAO v2.0 accomplishes 100% NScov when rDNA is masked, confirming the absence of unresolved SEs.

In T2T-YAO v1.1, alignment of SPRQ and Element reads revealed widespread artificial collapses, totaling 4.71 Mb and 2.52 Mb, respectively, in regions with >2-fold mean depth (**Fig. 2a, Extended Fig. 7a**). After correction, these figures were reduced to 3.43 Mb (SPRQ) and 0.93 Mb (Element), with 22.99% and 35.37% of these remaining regions located within rDNA (**Fig. 2b, Extended Fig. 7b**). Reduced artifact-induced alignments in high-accuracy data minimized the risk of over-correction of errors and improved reliability in the polished assembly.

Correction of Base-level Errors

Using ONT, SPRQ, and Element sequencing data, we identified 21,511 base-level error (BE) windows in T2T-YAO v1.1, comprising 3,694 homopolymer-associated BEs (h-BEs) and 17,817 non-homopolymer-associated BEs (nh-BEs), broadly distributed across all chromosomes (**Fig. 3b**). To avoid overcorrection, each flagged window is replaced with a consensus sequence derived from

238 reads spanning that window. In cases where the three sequencing platforms disagree, we prioritize
239 corrections supported by at least two platforms or follow the order of reliability: Element > SPRQ >
240 ONT. If no consensus is obtainable, ONT reads spanning the region are locally reassembled to
241 correct the sequence, provided the error is not homopolymer-related (**Extended Fig. 8**). For h-BEs,
242 local ONT-based consensus or assembly is omitted due to the known limitations of ONT sequencing
243 in accurately resolving homopolymers (**Extended Fig. 9**). This platform-aware consensus strategy
244 minimized alignment artifacts and reduced overcorrection during subsequent polishing rounds.
245 Read alignment and polishing are repeated iteratively until the number of BEs is stabilized.
246
247 To correct additional base-level errors, we applied multiple variant callers: Deeppolisher[20] (SPRQ),
248 DeepVariant[21] (ONT and Element), and GATK HaplotypeCaller[22] (Element) to call variants in the
249 diploid genome. These tools yielded a total of non-redundant 63,398 homozygous variants with
250 genotype quality (GQ) ≥20. To eliminate false positives, we filtered out variants which original
251 sequences as well as 10 bp flanking sequences were supported by ≥60% local depth and ≥10 exact-
252 matching reads. This filtering step yielded 43,444 potential SNV-like errors across both haplotypes,
253 of which 40.21% were associated with homopolymer regions, an even higher proportion than the
254 observed rate of h-BEs (**Extended Fig. 10a**). The elevated frequency of homopolymer-associated
255 errors likely stems from the homopolymer compression step used in the initial T2T-YAO assembly
256 with Verkko[23,24]. Of these SNV-like errors, 43.90% of non-homopolymer-associated variants
257 overlapped with SAS-detected nh-BEs, whereas only 7.08% of homopolymer-related calls
258 overlapped with SAS h-BEs. This discrepancy highlights the challenge of identifying true
259 homopolymer-associated errors via alignment. All variant callers detect discrepancies between the
260 assembly and aligned reads; however, many such discrepancies are false positives arising from read
261 misalignment or homopolymer inaccuracy. Accordingly, we treated SNV-like variants with the same
262 correction strategy as BEs: replacing both the variant site and its 10 bp flanking sequence with the
263 read-derived consensus sequence, rather than using the sequence predicted by the variant caller.
264
265 After several rounds of iterative polishing, T2T-YAO v2.0 containes only 336 nh-BEs (all within
266 rDNA regions) and 837 h-BEs located in repetitive regions (including 489 in rDNA) that remain
267 unresolved due to the inherent limitations of current sequencing technologies (**Table 1, Fig. 3b**).

When we reassess the v2.0 assembly using the same SNV callers, the total number of filtered SNV-like potential errors (homozygous SNV with GQ ≥20) drops significantly to 2,354 (**Extended Fig. 10b**). Of these, only 38 are non-homopolymer-associated, with 36 (94.74%) located within rDNA units, consistent with the distribution of nh-BEs identified by SAS. The remaining two non-homopolymer calls outside rDNA are confirmed to be false positives.

An illustrative example of an unresolved h-BE can be found in the D4Z4 repeat array on chromosome 4q35[25]. In this region, the polyC tracts within DUX4-like genes can not be confidently resolved: ONT and SPRQ reads struggled with homopolymer sequencing, while Element reads suffered from mapping ambiguity (**Extended Fig. 11**). Nonetheless, the remaining h-BEs account for only 0.04% of the total 2,412,254 homopolymer sites genome-wide. Among these, polyA/polyT tracts (≥10 bp) comprised 81.70%, while polyC/polyG tracts (≥7 bp) and poly-dinucleotide repeats (≥10 bp) represented 9.20% and 9.10%, respectively. The distribution of these homopolymer types is avoid of centromere regions and consistent across haplotypes (**Extended Fig. 12**).

Toward an Error-Free Human Genome Assembly

We re-evaluated the quality of the T2T-YAO v2.0 assembly using both SAS and Merqury metrics, based on the same sequencing datasets (**Table 1**). The final SAS assessment identified 1,773 structural error (SE) windows (all confined to rDNA regions except for the six windows of high-depth in maternal chromosome 7) and 1,175 base-level error (BE) windows, comprising 827 within rDNA and 348 homopolymer-associated BEs (h-BEs) outside rDNA. The resulting QV_{SAS} scores were 66.00 (maternal) and 61.25 (paternal). When rDNA regions were masked, leaving only the 348 h-BEs, the QV_{SAS} values rose to 71.62 and 71.81, respectively. (**Table S2, Supplementary Figs. 1–45**).

T2T-YAO v2.0 contains tens of kilobases of GnA/CnT repeat sequences that are poorly supported by Element and SPRQ reads but well supported by ONT data. Due to the limited representation of these sequences in the k-mer set used by Merqury, a high number of false positives are called in these loci. Specifically, Merqury identified 3,006 erroneous 21-mers (corresponding to a QV of 76.18) and 28,956 erroneous 31-mers (QV 68.03) in the diploid assembly. Since all erroneous 21-

298 mers are encompassed by the erroneous 31-mers, we focused our inspection on the 31-mers. These
299 were consistently found in regions with near-zero coverage in Element and SPRQ data but showed
300 strong support from ONT reads. The 31-mers clustered into 1,127 discrete regions (after merging
301 overlaps), collectively spanning 76,248 bp. Of these, 674 regions (totaling 53,307 bp) had extremely
302 high GA/CT content (>90%), while the rest were composed of non-canonical 6-mers found in
303 telomeric sequences or GC-rich tandem repeats, patterns consistent with known sequencing
304 biases[26,27] (**Table S4**).
305
306 The ONT data used in this study were generated using the R10 flow cell, known for high base-
307 calling accuracy (Q28; >99.8%). To leverage this accuracy, we incorporated k-mers from the ONT
308 dataset into our analysis, applying a stringent filter to exclude potential sequencing errors, setting
309 the minimum occurrence threshold at ≤4 for ONT data, compared to ≤1 for Element and SPRQ
310 reads. We first use this expanded trinary k-mer set to assess T2T-YAO v1.1, which QV only shows
311 a slight improvement over the binary set, indicating minimal inclusion of ONT sequencing errors.
312 When we applied the trinary k-mer set to assess T2T-YAO v2.0, the QV for 21-mers increased to
313 "infinite," representing the first Merqury-defined error-free diploid genome. The QV for 31-mers
314 also improved significantly, reaching 96.29 for the maternal haplotype and 88.73 for the paternal,
315 with only 143 erroneous 31-mers remaining. All paternal errors are confined to rDNA regions, while
316 the maternal errors are located in two h-BE loci of GnA repeat regions containing several polyG
317 tracts (≥8 bases), corresponding precisely with the residual errors detected by the SAS metric.
318
319 The comprehensive correction of SEs and non-homopolymer BEs outside rDNA enabled
320 unprecedented resolution of human centromeres, regions that remain among the most challenging
321 to assemble aside from rDNA clusters[28]. We refined the annotation of centromeres, higher-order
322 repeats (HORs), and other satellite elements across all 46 chromosomes (**Fig. 4b**). Subtle but
323 significant differences between T2T-YAO v1.1 and v2.0 were observed, exemplified by updates to
324 the paternal chromosome 1 assembly, highlighting the importance of high-fidelity polishing for
325 accurate centromere characterization (**Extended Fig. 13**). Using the recently developed Genomic
326 Centromere Profiling (GCP) pipeline, a method analogous to fragment length analysis for
327 characterizing the position, orientation, and organization of centromere protein B (CENP-B)

binding motifs[29], we constructed structural models for centromeric repeat clusters in T2T-YAO. These models revealed that the sequence architecture of human centromeres varies substantially in size, structure, and composition among chromosomes[30]. Although maternal and paternal centromeres of the same chromosome exhibit similar patter, differences in the size of components are still obvious between haplotypes, such as the HOR of maternal chromosomes 6 and 7 were significantly expanded relative to their paternal counterparts in T2T-YAO (**Extended Fig. 4**).

Together, these advances, achieved through iterative structural and base-level polishing, demonstrate the power of the SAS framework and our "Structural Error-first" strategy. The resulting T2T-YAO v2.0 assembly represents one of the most accurate and complete diploid human genomes to date, nearing the status of an error-free reference genome.

Discussion

An error-free diploid human genome holds immense value for calibrating variant calling pipelines and enhancing the accuracy of clinical genetic diagnoses[5,31]. Following the completion of the T2T-CHM13 assembly, the "Q100" project was initiated with the ambitious goal of producing a diploid human genome of Phred quality score 100, equivalent to one error per 10 billion bases[32]. Given the approximate size of the human diploid genome (~6 billion bases), achieving Q100 implies constructing a truly error-free genome. While T2T-YAO v2.0 has been validated by non-split ONT reads and is free of erroneous 21-mers, thus qualifying for an "infinite" QV under Merqury standards, technical limitations in both genome evaluation and error correction prevent it from being deemed absolutely error-free.

Genome assembly quality encompasses multiple dimensions, including completeness, correctness, and continuity. Of these, completeness is relatively straightforward to assess via k-mer-based approaches[9]. However, accurate evaluation of correctness and continuity remains more challenging and often requires alignment-based methods[33,34]. Most alignment-based tools identify discrepancies between the genome assembly and realigned reads. These discrepancies often include both false positives, arising from artificial alignments of heterogeneous reads, and false negatives, which are missed due to sequencing bias. Moreover, integrating alignment discrepancies across platforms

remains difficult.

To overcome these challenges, we developed SAS, a window-based alignment metric that flags 50 bp segments not sufficiently supported by reads from any platform. In our implementation, a 50-mer is considered supported only if it is exactly supported by at least five reads (representing >10% of the mean sequencing depth) in its context. This approach offers higher stringency and sensitivity than k-mer-based methods such as Merqury, which typically examine 21- or 31-mers and require only two occurrences. Similarly, structural errors are detected in windows lacking sufficient spanning reads. The window-based design also improves computational efficiency: by initially filtering for regions perfectly covered by Element reads without mismatches or clips, which constitute 99.2% of the T2T-YAO genome, the overall computational burden is greatly reduced. This enables SAS to complete genome-wide assessment in under two hours after alignment.

Despite its sensitivity, SAS still misses some base-level errors in repetitive regions where misaligned reads from other loci create false support. It also fails to detect collapses or false duplications that fall within acceptable depth thresholds. These problematic regions often show heterogeneous SNP patterns in the aligned reads[16], complicating distinction from sequencing noise, particularly with ONT reads. A canonical example is the ribosomal DNA (rDNA) clusters, which remain the final hurdle to a fully error-free human genome[35]. In these regions, ONT reads are long enough to span complete rDNA units but lack the base-level accuracy to resolve unit-level SNVs, whereas SPRQ reads provide the needed accuracy but are not long enough to span the entire unit. Thus, the primary barriers to achieving a completely error-free diploid human genome are intrinsic to the architecture of highly repetitive sequences, particularly their unit length and sequence identity, and the current limitations of sequencing technologies in generating reads that are both sufficiently long and accurate. Bridging this gap remains the final frontier in complete human genome assembly.

Methods

iPSC Culture and DNA Sequencing

Induced pluripotent stem cells (iPSCs) used for the T2T-YAO assembly were cultured in mTeSR™ Plus medium on vitronectin-coated plates and passaged at a 1:10 ratio using 0.02% EDTA (Versene). High-molecular-weight genomic DNA for Oxford Nanopore Technologies (ONT) ultra-long read sequencing was extracted using the Grandomics BAC-long kit. DNA fragments >50 kb were enriched using SageHLS, and sequencing libraries were prepared with the SQK-LSK114 kit. Sequencing was performed on a PromethION platform with R10 flow cells. Basecalling was conducted using Dorado (v5.0.0, sup mode), and reads with quality scores <Q7 were excluded.

For PacBio Revio® sequencing, DNA was extracted via CTAB and purified using Grandomics kits. SMRTbell® libraries were prepared with the SMRTbell Prep Kit 3.0 and analyzed using SMRT Link v25.1. For Element AVITI sequencing, PCR-free libraries were generated using the Elevate Enzymatic Library Prep Kit (Cat. No. 830-00009) and Long UDI Adapter Kit Set A (Cat. No. 830-00010), following the manufacturer's protocols. Libraries were pooled, denatured, and sequenced using the AVITI 2×150 Sequencing Kit CB UltraQ (Cat. No. 860-00018) on the AVITI platform.

Detection of Structural Error (SE) Windows

ONT ultra-long reads were filtered to retain only those >30 kb and phased using haplotype-specific 31-mers (hapmers). A read was assigned to a haplotype if it contained >40 hapmers from one haplotype and at least three-fold more than from the opposite haplotype. Phased reads were aligned to the corresponding haplotype assemblies using Winnowmap2[36]; unassigned reads were aligned to the diploid reference. Only primary alignments were retained.

The assembly was scanned using non-overlapping 2 kb windows. A window was flagged as an SE if :

1. **INDELs and Clipping**: >40% of the local reads contained insertions or deletions ≥50 bp detected via CIGAR string analysis, or Presence of hard or soft clipping in aligned reads.
2. **Abnormal Depth**: Read depth <20% or >200% of the genome-wide mean (assessed using Mosdepth).
3. **Lack of Spanning Reads**: No spanning reads were observed within 15 consecutive 2 kb windows (*i.e.*, a 30 kb region).

Thresholds were relaxed near chromosome ends (<10 kb from termini), where a window was flagged only if no or fewer than 10% of local reads spanned the region and lacked indels.

Detection of Base-Level Error (BE) Windows

Element reads were aligned to the diploid assembly using BWA-MEM. Alignments were filtered to retain only perfect matches (CIGAR NM:i:0) without clipping. SPRQ reads were phased and aligned similarly to ONT reads.

Base-level errors were evaluated in sliding 50 bp windows. Windows were flagged as BEs if the exact sequence was supported by reads at <10% of the mean sequencing depth. Windows with sufficient (>10% mean depth) perfect-match Element reads were excluded. Remaining regions were further validated using SPRQ and ONT data. In low-coverage regions, thresholds were relaxed to 10% of **local** depth under the following conditions:

1. Within 10 kb of chromosomal termini (*i.e.*, telomeric regions);
2. Regions with degenerative GA/CT repeats (≥90% GA/CT and ≥60% GC content);
3. GC-rich regions (≥90% GC content).

Merqury-Based Assembly Evaluation

Merqury and Meryl were used to assess k-mer-based quality metrics (QV), completeness, and phasing accuracy using 21-mers and 31-mers. The evaluation followed the methods described by McCartney et al[7] for the binary k-mer set using Element and SPRQ data, and for the trinary k-mer set except combining the k-mer set of ONT reads of occurrence >4.

Variant Calling (SVs and SNVs)

Structural variants-like errors were called using Sniffles2 (ONT), Flagger (ONT), and NucFlag (SPRQ), with default parameters. Base-level SNV-like errors were identified using DeepVariant (ONT and Element), Deeppolisher (SPRQ), and GATK HaplotypeCaller (Element). Variants with genotype quality (GQ) <20 were excluded. Heterozygous calls were also filtered out except for haplotype-aware calls from Deeppolisher.

Error Correction

SE windows were first corrected using variant calls from Sniffles2 when available. In cases without confident calls, reads spanning the SE and 5–10 kb flanks were reassembled using Hifiasm or Flye. If required, flanking regions were extended to 50–100 kb for resolving large collapses or duplications. Reassembled contigs were aligned back to the reference, and the corrected sequences were substituted accordingly. Manual assessment and correction with IGV is often necessary for SEs of clipping or complex errors. Telomeric regions with misalignment or heterogeneity were extended using major-allele ONT reads. Corrections were followed by haplotype-aware re-alignment and re-evaluation with SAS. The process was repeated until no SEs were detected outside rDNA clusters.

BE windows were classified into homopolymer-associated (h-BE) and non-homopolymer-associated (nh-BE).

nh-BEs were corrected using consensus sequences from supporting reads in the order of platform priority: Element > SPRQ > ONT. Where consensus could not be reached, local ONT reassembly was performed.

h-BEs were polished using Element or SPRQ consensus only, due to ONT's known limitations in homopolymer resolution.

For SNV-like errors called by DeepVariant, Deeppolisher, and GATK HaplotypeCaller, false positives were removed if >55% and ≥10 reads from any platform exactly supported the reference allele sequence including ±10 bp flanks. Remaining SNV calls were corrected using consensus-based replacement as nh-BEs and h-BEs except within the ±10 bp flanks windows. Polishing iterations continued until the number of detected BEs plateaued.

Centromere and Telomere Annotation

Centromeric regions were annotated using RepeatMasker[37], with α-satellite arrays identified via HumAS-HMMER[30]. Telomeric repeats were detected and quantified using Telescope[38].

References

1 Church, D. M. A next-generation human genome sequence. *Science (New York, N.Y.)* **376**, 34-35 (2022). https://doi.org:10.1126/science.abo5367

2 Nurk, S. *et al.* The complete sequence of a human genome. *Science (New York, N.Y.)* **376**, 44-53 (2022). https://doi.org:10.1126/science.abj6987

3 Olson, N. D. *et al.* Variant calling and benchmarking in an era of complete human genome sequences. *Nature Reviews Genetics* **24**, 464-483 (2023). https://doi.org:10.1038/s41576-023-00590-0

4 Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science (New York, N.Y.)* **376**, eabl3533 (2022). https://doi.org:doi:10.1126/science.abl3533

5 Jarvis, E. D. *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519-531 (2022). https://doi.org:10.1038/s41586-022-05325-5

6 An assembly line for an improved human reference genome. *Nature* (2022). https://doi.org:10.1038/d41586-022-03151-3

7 Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nature methods* **19**, 687-695 (2022). https://doi.org:10.1038/s41592-022-01440-3

8 Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737-746 (2021). https://doi.org:10.1038/s41586-021-03451-0

9 Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology* **21**, 245 (2020). https://doi.org:10.1186/s13059-020-02134-9

10 He, Y. *et al.* T2T-YAO: A Telomere-to-telomere Assembled Diploid Reference Genome for Han Chinese. *Genomics, proteomics & bioinformatics* **21**, 1085-1100 (2023). https://doi.org:10.1016/j.gpb.2023.08.001

11 Ni, Y., Liu, X., Simeneh, Z. M., Yang, M. & Li, R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. *Comput Struct Biotechnol J* **21**, 2352-2364 (2023). https://doi.org:10.1016/j.csbj.2023.03.038

12 Deshpande, A. S. *et al.* Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nature biotechnology* **40**, 1488-1499 (2022). https://doi.org:10.1038/s41587-022-01289-z

13 Arslan, S. *et al.* Sequencing by avidity enables high accuracy with low reagent consumption. *Nature biotechnology* **42**, 132-138 (2024). https://doi.org:10.1038/s41587-023-01750-7

14 Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nature biotechnology* **42**, 1571-1580 (2024). https://doi.org:10.1038/s41587-023-02024-y

15 Liao, W. W. *et al.* A draft human pangenome reference. *Nature* **617**, 312-324 (2023). https://doi.org:10.1038/s41586-023-05896-x

16 Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nature methods* **16**, 88-94 (2019). https://doi.org:10.1038/s41592-018-0236-3

17 Lyčka, M. *et al.* TeloBase: a community-curated database of telomere sequences across the tree of life. *Nucleic acids research* **52**, D311-D321 (2023).

https://doi.org:10.1093/nar/gkad672

18 Stephens, Z. & Kocher, J.-P. Characterization of telomere variant repeats using long reads enables allele-specific telomere length estimation. *BMC Bioinformatics* **25**, 194 (2024). https://doi.org:10.1186/s12859-024-05807-5

19 Belancio, V. P., Deininger, P. L. & Roy-Engel, A. M. LINE dancing in the human genome: transposable elements and disease. *Genome Medicine* **1**, 97 (2009). https://doi.org:10.1186/gm97

20 Mastoras, M. *et al.* Highly accurate assembly polishing with DeepPolisher. *bioRxiv* (2024). https://doi.org:10.1101/2024.09.17.613505

21 Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature biotechnology* **36**, 983-987 (2018). https://doi.org:10.1038/nbt.4235

22 Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 201178 (2018). https://doi.org:10.1101/201178

23 Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature biotechnology* (2023). https://doi.org:10.1038/s41587-023-01662-6

24 Antipov, D. *et al.* Verkko2 integrates proximity-ligation data with long-read De Bruijn graphs for efficient telomere-to-telomere genome assembly, phasing, and scaffolding. *Genome research* **35**, 1583-1594 (2025). https://doi.org:10.1101/gr.280383.124

25 Lemmers, R. *et al.* Cis D4Z4 repeat duplications associated with facioscapulohumeral muscular dystrophy type 2. *Hum Mol Genet* **27**, 3488-3497 (2018). https://doi.org:10.1093/hmg/ddy236

26 Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome biology* **14**, R51 (2013). https://doi.org:10.1186/gb-2013-14-5-r51

27 Freeman, T. M., Wang, D. & Harris, J. Genomic loci susceptible to systematic sequencing bias in clinical whole genomes. *Genome research* **30**, 415-426 (2020). https://doi.org:10.1101/gr.255349.119

28 Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science (New York, N.Y.)* **376**, eabl4178 (2022). https://doi.org:doi:10.1126/science.abl4178

29 Corda, L. & Giunta, S. Chromosome-specific centromeric patterns define the centeny map of the human genome. *Science (New York, N.Y.)* **389**, eads3484 (2025). https://doi.org:10.1126/science.ads3484

30 Logsdon, G. A. *et al.* The variation and evolution of complete human centromeres. *Nature* **629**, 136-145 (2024). https://doi.org:10.1038/s41586-024-07278-3

31 Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nature biotechnology* **40**, 672-680 (2022). https://doi.org:10.1038/s41587-021-01158-1

32 consortium, T.-t.-t. HG002 Q100 project. *https://github.com/marbl/hg002* (2023).

33 Li, K., Xu, P., Wang, J., Yi, X. & Jiao, Y. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nature communications* **14**, 6556 (2023). https://doi.org:10.1038/s41467-023-42336-w

34 Mastoras, M. *et al.* Highly accurate assembly polishing with DeepPolisher. *bioRxiv*, 2024.2009.2017.613505 (2024). https://doi.org:10.1101/2024.09.17.613505

35 Li, H. & Durbin, R. Genome assembly in the telomere-to-telomere era. *Nat Rev Genet* **25**,

658-670 (2024). https://doi.org:10.1038/s41576-024-00718-w

36 Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nature methods* **19**, 705-710 (2022). https://doi.org:10.1038/s41592-022-01457-8

37 Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, 4.10.11-14.10.14 (2009). https://doi.org:10.1002/0471250953.bi0410s25

38 Formenti, G. *et al.* Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics (Oxford, England)* **38**, 4214-4216 (2022). https://doi.org:10.1093/bioinformatics/btac460

Data Availability

The raw sequencing data for T2T-YAO v2.0 generated in this study have been deposited in the GSA for Human database at the China National Center for Bioinformation (CNCB) under accession number HRA011075, publicly accessible at https://ngdc.cncb.ac.cn/gsa-human. The raw data for v1.1, released on August 16, 2023, are available from the same URL under accession number HRA004987. The T2T-YAO v1.1 and v2.0 genome assemblies are available in the Genome Warehouse at CNCB (https://ngdc.cncb.ac.cn/gwh) under the following accession numbers: v1.1: GWHDQZJ00000000 (maternal), GWHDOOG00000000 (paternal); v2.0: GWHGEYC00000000.1 (maternal), GWHGEYB00000000.1 (paternal). All genome sequences and associated datasets can be freely downloaded without application and are also publicly accessible at https://www.genomeyao.org.cn/v2.

Code Availability

Code for SAS, error correction, and all other workflows are available at https://github.com/KANGYUlab/sas-pipeline and https://github.com/KANGYUlab/genome-polish-pipeline.

Code for other software used in this work:

Meryl v1.3 (https://github.com/marbl/meryl) and Merqury v1.1 (https://github.com/marbl/merqury) for k-mer QV assessment. ONT and SPRQ reads were aligned to assembly using Winnowmap v2.03 (https://github.com/marbl/Winnowmap). Aligment between assemblies was using minimap2 2.27 (https://github.com/lh3/minimap2). Element reads were aligned with BWA-MEM 0.7.17

(https://github.com/lh3/bwa). SAMtools v1.19.2 (https://github.com/samtools/samtools/) and bedtools v2.25.0 (https://github.com/arq5x/bedtools2) were used for BAM files and BED files processing. Reads alignment were visualized with IGV v2.17.3 (https://github.com/igvteam/igv.js#igvjs). Genome-wide alignment depth was calculated using mosdepth (https://github.com/brentp/mosdepth). The SVs were detected using NucFlag v0.3.6 (https://github.com/logsdon-lab/NucFlag) from the phased alignmen of SPRQ, Flagger v0.3.2 (https://github.com/mobinasri/flagger/), Sniffles2 v2.4 (https://github.com/fritzsedlazeck/Sniffles) from the phased alignment of ONT. The SNVs were detected by GATK v4.3.0.0 (https://github.com/broadinstitute/gatk), Deepvariant v1.9.0 (https://github.com/google/deepvariant), Deepploisher r0.1.0 (https://github.com/google/deeppolisher). Local reassemblies were performed with Flye v2.9.5 (https://github.com/mikolmogorov/Flye) and Hifiasm v0.25.0 (https://github.com/chhylp123/hifiasm). Bcftools v1.20 (https://github.com/samtools/bcftools/) was utilized for processing VCF files and incorporating polishing edits into the new assembly. Centromere regions were identified using RepeatMasker v4.1.9 (https://www.repeatmasker.org/RepeatMasker), and the location of the α-satellite HOR arrays were identified using HumAS-HMMER (https://github.com/enigene/HumAS-HMMER). The telomere regions were identified using Telescope v0.1.1 (https://github.com/vgl-hub/teloscope). Haplotype phasing were validated with Pore-C contact maps using wf-pore-C v1.3.0 (https://github.com/epi2me-labs/wf-pore-c) and visualized in Juicebox v2.20 (https://github.com/aidenlab/Juicebox). The organization patterns of centromeres and telomeres were visualized using ggplot2 (v3.5.1) in R (v4.4.0). Statistical visualizations were generated using Matplotlib (v3.7.1) with NumPy (v1.24.3) and Pandas (v2.0.3) in Python (v3.9.16).

Acknowledgments

This study was supported by the grants of the National Key Research and Development Program of China (2024YFC3405701), the National Science Foundation of China (32371537).

Author contributions

Y.K. conceived and designed the project. Z.H. implemented the software SAS. Y.C. conducted the

correction of T2T-YAO. S.G., C.S., J.W. performed sample preparation and sequencing. X.Y., Y.T., J.C., R.L. and Y.H. participated in data analyses, and Y.K., Z.G., J.H., J.Y. wrote the manuscript. All the authors agree on the manuscript.

Competing interest declaration

All authors declare no competing interests.

Table

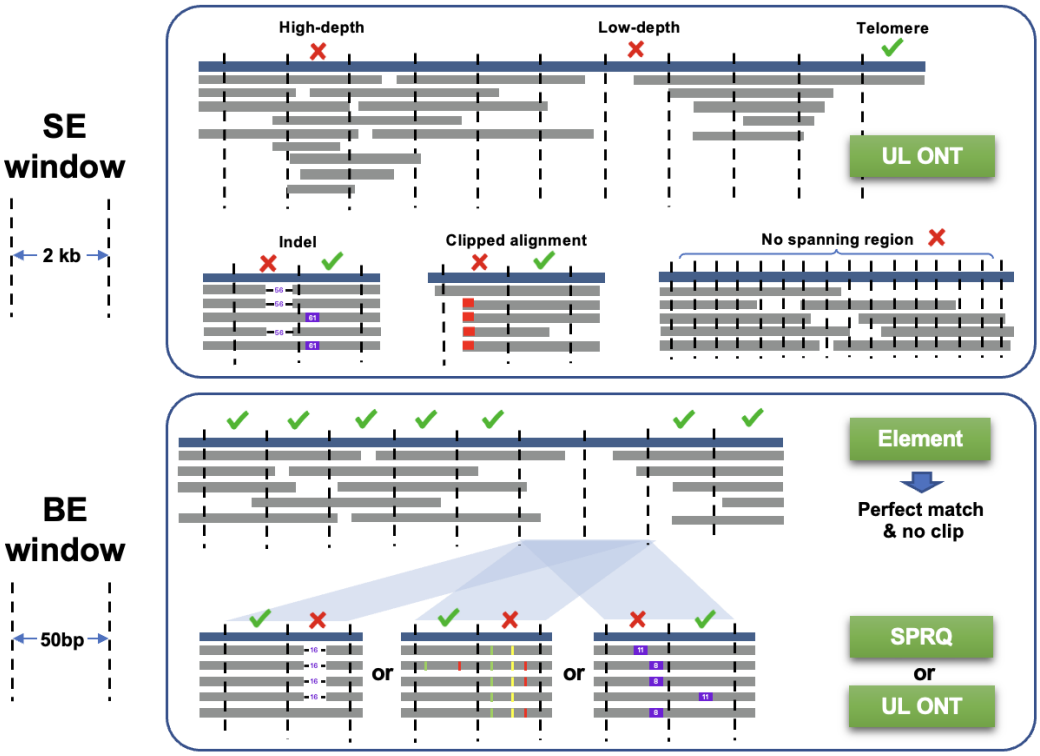
Table 1. Comparative Quality Metrics of the T2T-YAO Assembly Assessed by SAS and Merqury.

		v1.1			v2.0		
		mat	**pat**	**Both**	**mat**	**pat**	**Both**
SAS	**QV (total)**	**53.12**	**54.13**	**53.59**	**66.00**	**61.25**	**63.04**
	BE	**11946**	**9565**	**21511**	**291**	**884**	**1175**
	h-BE	1789	1905	3694	255	582	837
	nh-BE	10157	7660	17817	35	301	336
	SE	**2752**	**1686**	**4438**	**469**	**1304**	**1773**
	low_depth	920	742	1662	121	709	830
	high_depth	886	409	1295	302	376	678
	indel	141	190	331	9	49	58
	clip	39	22	61	6	24	30
	no spanning	1247	682	1929	99	469	568
	QV (rDNA masked)	**53.65**	**55.39**	**54.42**	**71.62**	**71.81**	**71.72**
	BE	**10871**	**7863**	**18734**	**179**	**169**	**348**
	h-BE	1426	1145	2571	179	169	348
	nh-BE	9445	6718	16163	0	0	0
	SE	**2147**	**572**	**2719**	**6**	**0**	**6**
	low_depth	730	57	787	0	0	0
	high_depth	643	115	758	6	0	6
	indel	116	125	241	0	0	0
	clip	23	8	31	0	0	0
	no spanning	1027	295	1322	0	0	0
Merqury-	**QV-21mer**	**66.48**	**68.68**	**67.42**	**74.84**	**78.23**	**76.18**
binary	err-kmer	14252	8304	22556	2085	921	3006
	total-kmer	3.02×10^9	2.92×10^9	5.93×10^9	3.02×10^9	2.92×10^9	5.93×10^9
	err rate	2.25×10^7	1.36×10^7	1.81×10^7	3.28×10^8	1.50×10^8	2.41×10^8
	switch err%	0.0107	0.0254	—	0.0017	0.0038	—
	Completeness%	97.770	93.708	99.839	97.777	93.710	99.850
	QV-31mer	**60.97**	**62.73**	**61.74**	**67.24**	**69.04**	**68.03**
	err-kmer	74864	48221	123085	17694	11262	28956
	total-kmer	3.02×10^9	2.92×10^9	5.93×10^9	3.02×10^9	2.92×10^9	5.93×10^9
	err rate	8.00×10^7	5.34×10^7	6.69×10^7	1.89×10^7	1.25×10^7	1.57×10^7
	switch err%	0.0219	0.0413	—	0.0050	0.0072	—
	Completeness%	96.780	92.607	99.806	96.792	92.610	99.826
Merqury-	**QV-21mer**	**67.7888**	**69.854**	**68.6823**	**inf**	**inf**	**inf**
trinary	err-kmer	10543	6331	16874	0	0	0
	total-kmer	3.02×10^9	2.92×10^9	5.93×10^9	3.02×10^9	2.92×10^9	5.93×10^9
	err rate	1.66×10^7	1.03×10^7	1.35×10^7	0	0	0
	switch err%	0.0107	0.0254	—	0.0017	0.0038	—
	Completeness%	97.7896	93.7264	99.8638	97.7966	93.7281	99.875
	QV-31mer	**62.4363**	**64.2025**	**63.2153**	**96.2941**	**88.7314**	**91.0965**

err-kmer	53376	34337	87713	22	121	143
total-kmer	3.02×10^{9}	2.92×10^{9}	5.93×10^{9}	3.02×10^{9}	2.92×10^{9}	5.93×10^{9}
err rate	5.71×10^{7}	3.80×10^{7}	4.77×10^{7}	2.35×10^{7}	1.34×10^{7}	7.77×10^{7}
switch err%	0.0219	0.0413	—	0.0050	0.0072	—
Completeness%	96.810	92.636	99.841	96.823	92.639	99.861

641 **Figures**

642

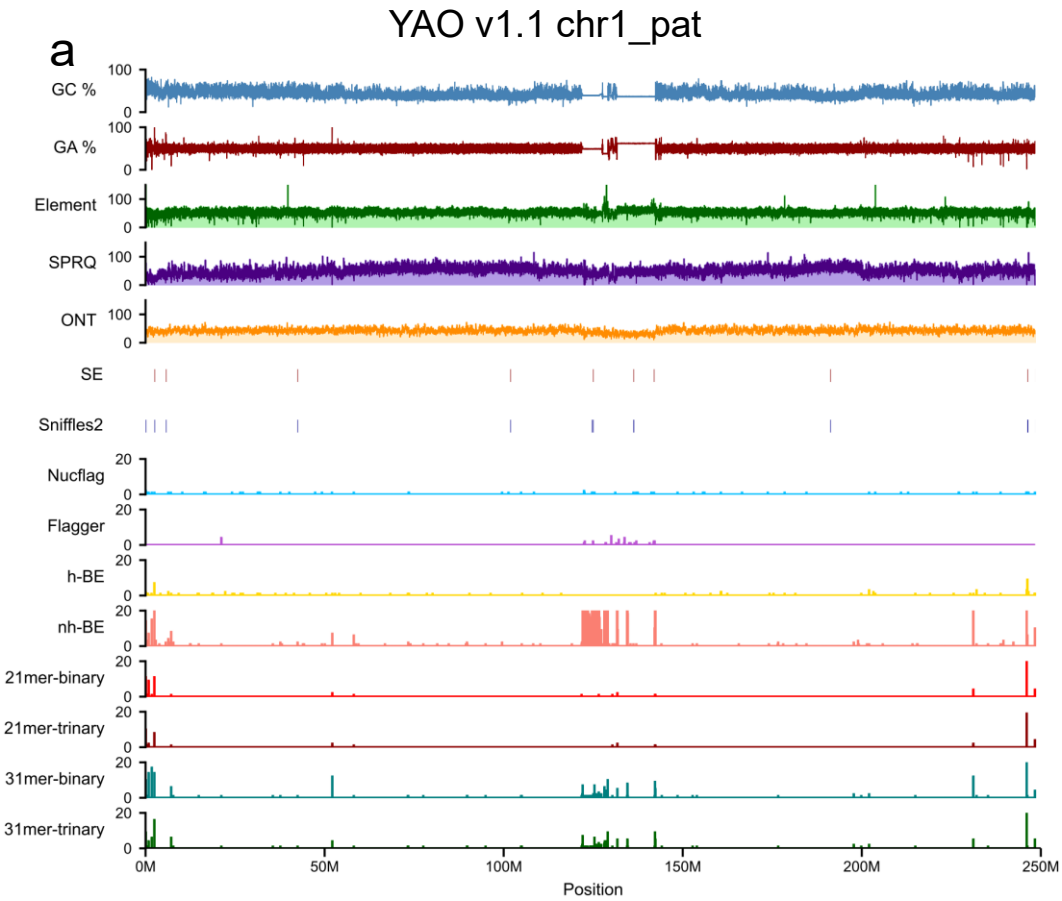


643

644 **Figure 1. Overview of the sufficient alignment support (SAS) metric for detecting assembly**
645 **errors.** SAS identifies structural errors (SEs) and base-level errors (BEs) based on read alignment
646 patterns across defined windows. Top panel (SE windows): Structural errors are detected in 2 kb
647 windows using ultra-long Oxford Nanopore (UL ONT) reads. Windows are flagged based on five
648 indicators: abnormally high or low depth, large insertions or deletions (indels), clipped alignments,
649 and lack of spanning reads. Telomeric regions with any supportive reads are not flagged. Bottom
650 panel (BE windows): Base-level errors are evaluated in 50 bp windows using reads from high-
651 accuracy sequencing platforms (Element, SPRQ, or UL ONT). A window is considered error-free
652 only if all bases are covered by perfectly matched, non-clipped reads. In regions not covered by
653 Element reads, support from SPRQ or UL ONT reads is used to validate sequence accuracy. SAS
654 flags windows lacking sufficient read support (*i.e.*, <10% of average sequencing depth), illustrating
655 example regions containing various assembly errors.

656

657


a
YAO v1.1 chr1_pat
GC %
GA %
Element
SPRQ
ONT
SE
Sniffles2
Nucflag
Flagger
h-BE
nh-BE
21mer-binary
21mer-trinary
31mer-binary
31mer-trinary
100
0
20
0M
50M
100M
150M
200M
250M
Position

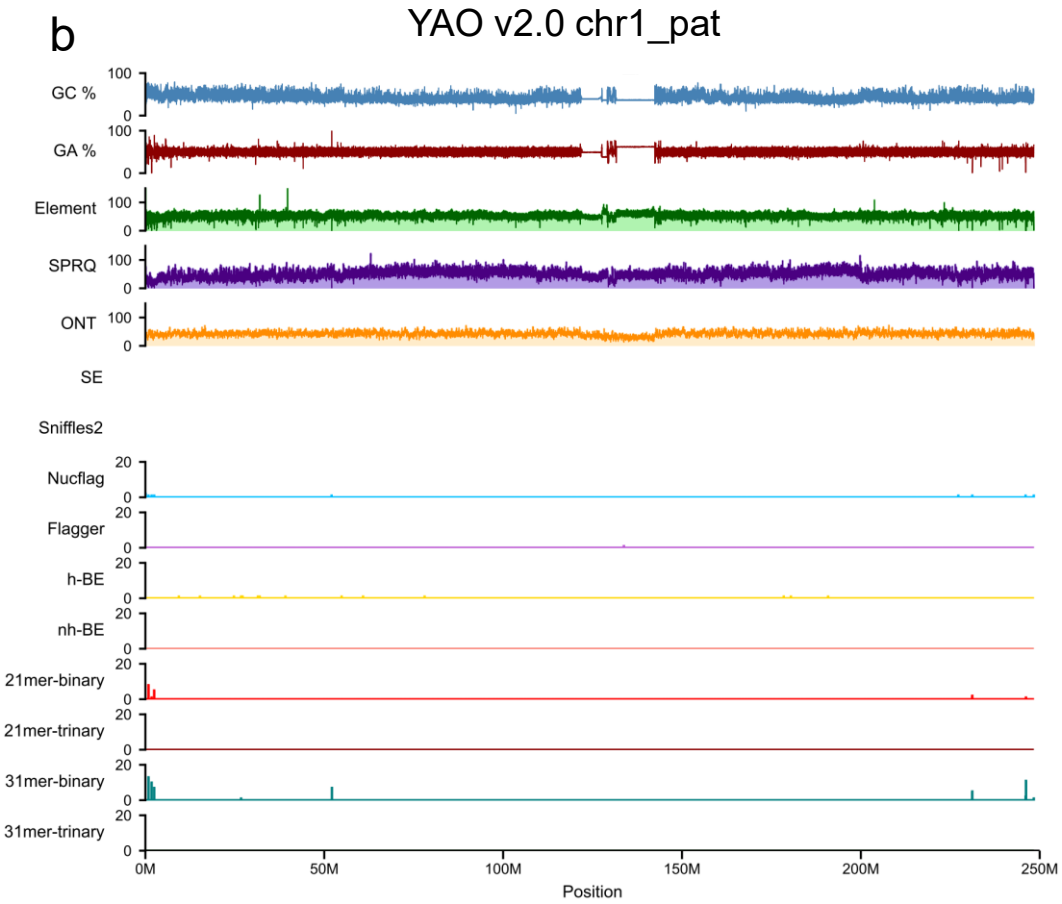

b
YAO v2.0 chr1_pat
GC %
GA %
Element
SPRQ
ONT
SE
Sniffles2
Nucflag
Flagger
h-BE
nh-BE
21mer-binary
21mer-trinary
31mer-binary
31mer-trinary
100
0
20
0M
50M
100M
150M
200M
250M
Position

Figure 2. Genomic distribution of assembly errors on paternal chromosome 1 before and after polishing. Genomic features and assembly error profiles for T2T-YAO v1.1(**a**) compared to the polished T2T-YAO v2.0 (**b**). Tracks from top to bottom: GC% and GA% content; normalized read depth from Element, PacBio Revio (SPRQ), and ONT ultra-long reads; structural error (SE) windows identified by SAS and Sniffles2; misjoins and collapses flagged by Nucflag and Flagger; base-level errors (BEs) categorized as homopolymer-associated (h-BE) and non-homopolymer-associated (nh-BE); erroneous 31-mers and 21-mers detected by Merqury with binary and trinary k-mer set.

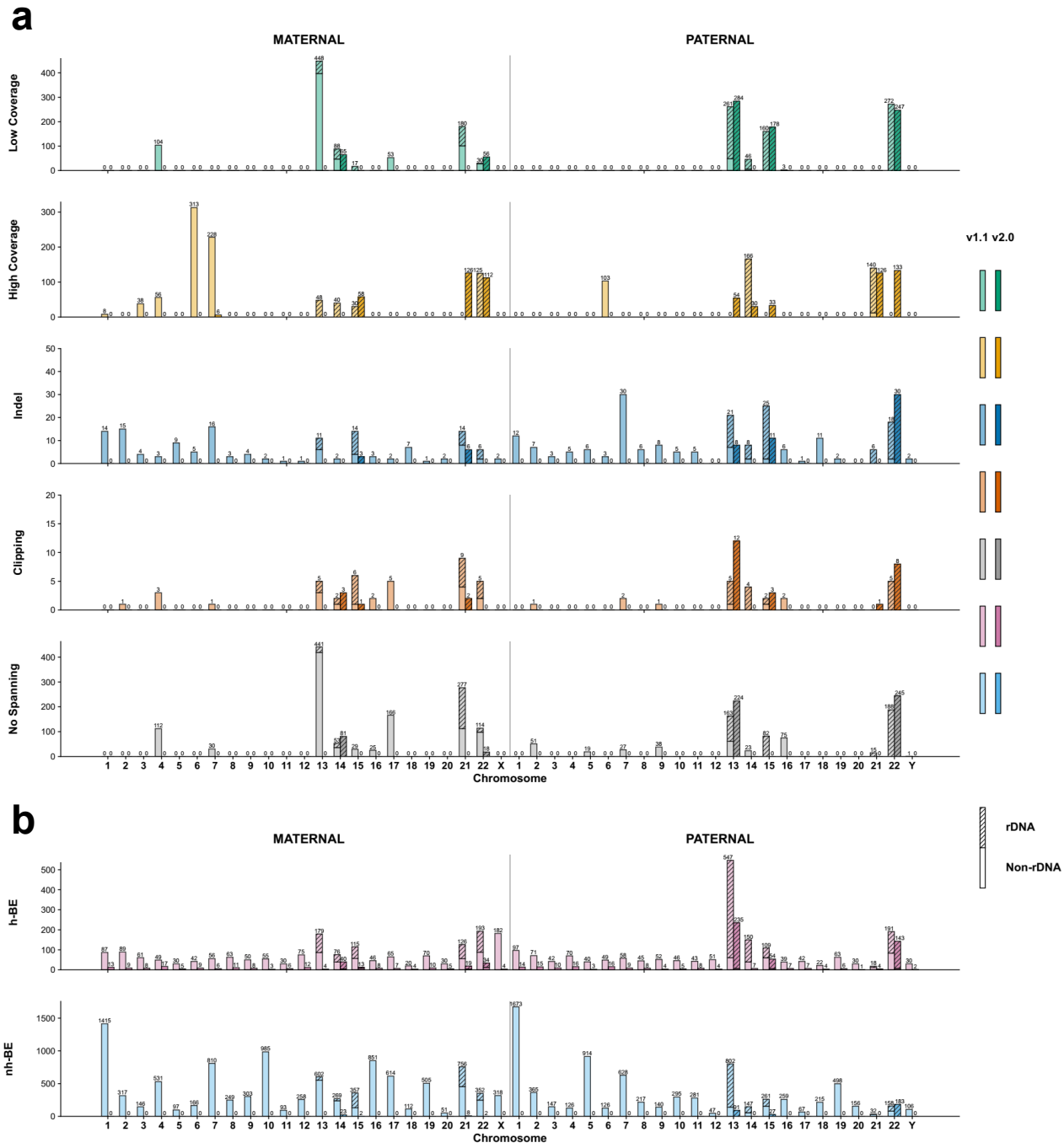
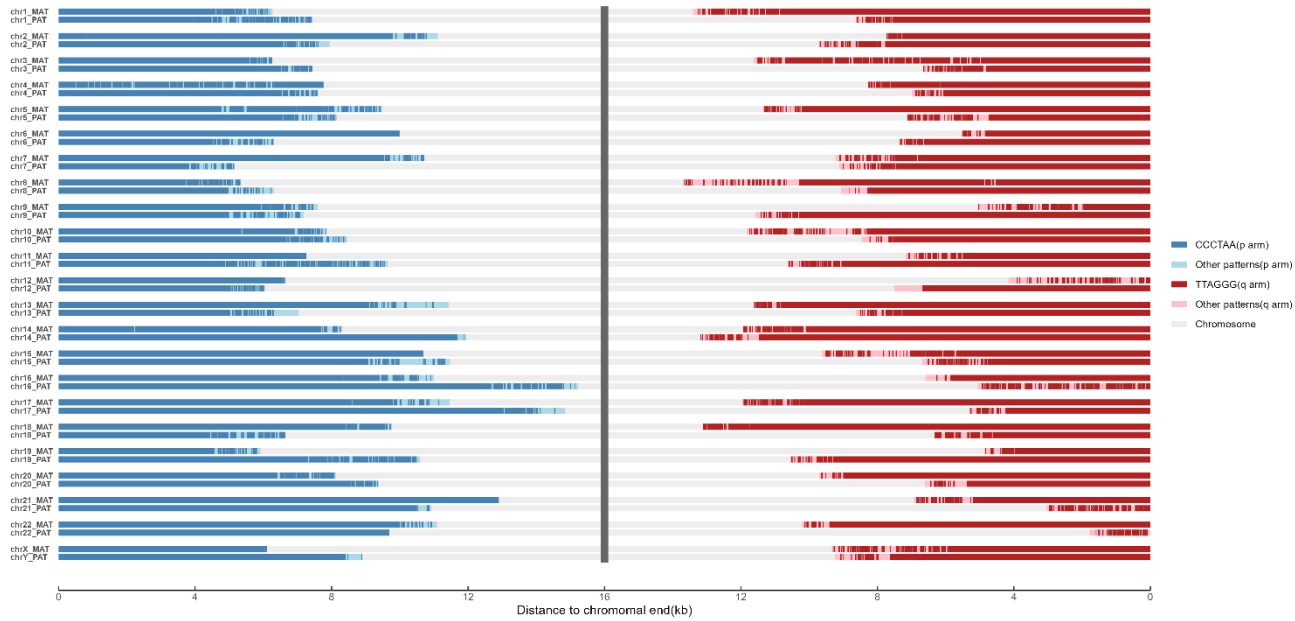


Figure 3. Counts of base-level and structural errors in T2T-YAO before and after iterative correction. Distribution of structural error (SE) windows (**a**) and base-level error (BE) windows (**b**) across chromosomes in maternal and paternal haplotypes of T2T-YAO v1.1 compared to v2.0. SEs were categorized into five types: low coverage, high coverage, indels (≥50 bp), clipped alignments, and non-spanning regions. BEs are divided into homopolymer-associated BEs (h-BEs) and non-homopolymer-associated BEs (nh-BEs). Bar plots show counts per chromosome, with the left bar

in each pair for v1.1 (light colors) and the right for v2.0 (dark colors). The rDNA-associated errors are shown as hatched bars.

680

a


CCCTAA(p arm)
Other patterns(p arm)
TTAGGG(q arm)
Other patterns(q arm)
Chromosome
Distance to chromomal end(kb)

b

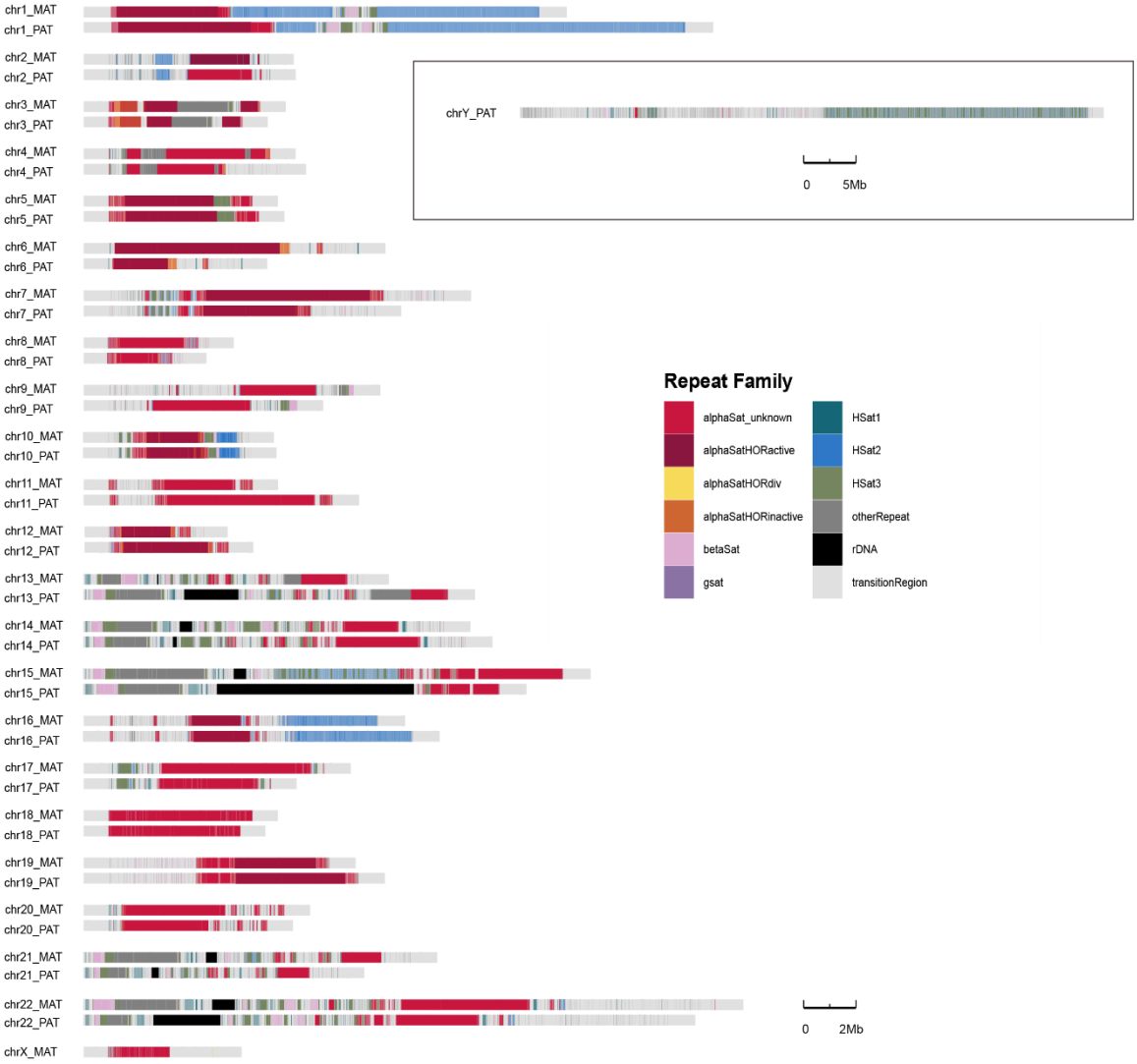

chrY_PAT
0
5Mb
Repeat Family
alphaSat_unknown
alphaSatHORactive
alphaSatHORdiv
alphaSatHORinactive
betaSat
gsat
HSat1
HSat2
HSat3
otherRepeat
rDNA
transitionRegion
0
2Mb

Figure 4. Structural characterization of telomeres and centromeres in the polished T2T-YAO v2.0 assembly.

(**a**) Telomeric repeat content and extension across all 46 chromosomal termini. Each bar represents the distance from the chromosomal terminus, with colored segments indicating specific 6-mer tandem repeats. Canonical (TTAGGG)n (in long arm) and complementary (CCCTAA)n (in short arm) motifs are shown in dark red and blue, along with shallow colors indicate non-canonical telomeric-like 6-mer repeats. (**b**) Annotation of centromeric regions in all the 46 chromosomes. Repeat content was visualized as stacked bars by annotated repeat family. Each pair of centromeres displays a unique repeat composition and organization, with notable differences between haplotypes. The inset shows a zoom-out view of the Y chromosome centromere, which is full of repeat sequences.

Extended Figures

Extended Figure 1. Example of a Structural Error in a Repetitive Region Missed by Sniffles2

Extended Figure 2. Schematic Overview of the Structural Error (SE) Correction Pipeline

Extended Figure 3. Contact Map of Pore-C Reads Aligned to the T2T-YAO v2.0 Assembly

Extended Figure 4. Comparison of Centromere Structures in T2T-YAO v1.1 and v2.0 via Genomic Centromere Profiling (GCP)

Extended Figure 5. Repetitive Elements in Maternal Chromosome 7 with Abnormal ONT Read Depth

Extended Figure 6. Length Distributions of Collapsed and Duplicated Regions Predicted by Flagger

Extended Figure 7. Read Depth Distributions of SPRQ and Element Data in T2T-YAO v1.1 and v2.0

Extended Figure 8. Schematic Overview of the Base-Level Error (BE) Correction Pipeline

Extended Figure 9. Example of Discordance Between ONT and SPRQ Reads at Homopolymer Sites

Extended Figure 10. Quality Score Distribution of SNV-like Errors in T2T-YAO v1.1 and v2.0

Extended Figure 11. Homopolymer Sites in the D4Z4 Region on Chromosome 4q35

Extended Figure 12. Distribution of Homopolymers Across Chromosomes in T2T-YAO v2.0

Extended Figure 13. Patterns of Higher-Order Repeats and Satellite Sequences in the Paternal Centromere of Chromosome 1 in T2T-YAO v1.1 and v2.0

Supplementary Tables

Table S1. Sequencing Datasets Utilized in This Study

Table S2. Chromosome-Wise Assembly Quality Metrics Before and After Polishing

Table S3. Structural Errors Identified by Multiple Tools Before and After Correction

Table S4. GC and GA/CT Content of Remaining Erroneous 31-mers in T2T-YAO v2.0

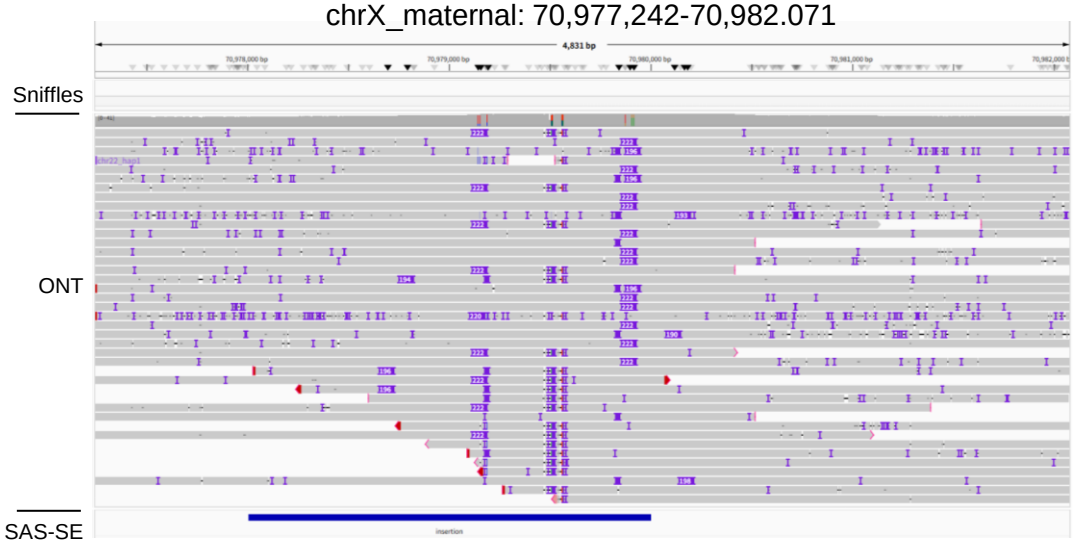
Supplement Methods

Supplement Figure 1-45 Distribution of assembly errors before and after correction in each chromosome except paternal chromosome 1.

Extended Fig. 1

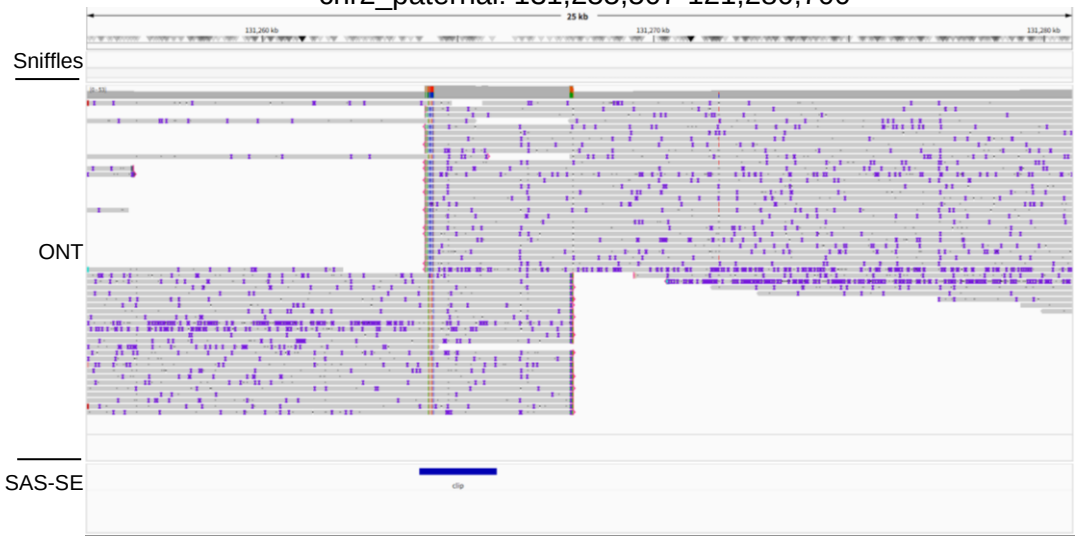
a

chrX_maternal: 70,977,242-70,982.071



b

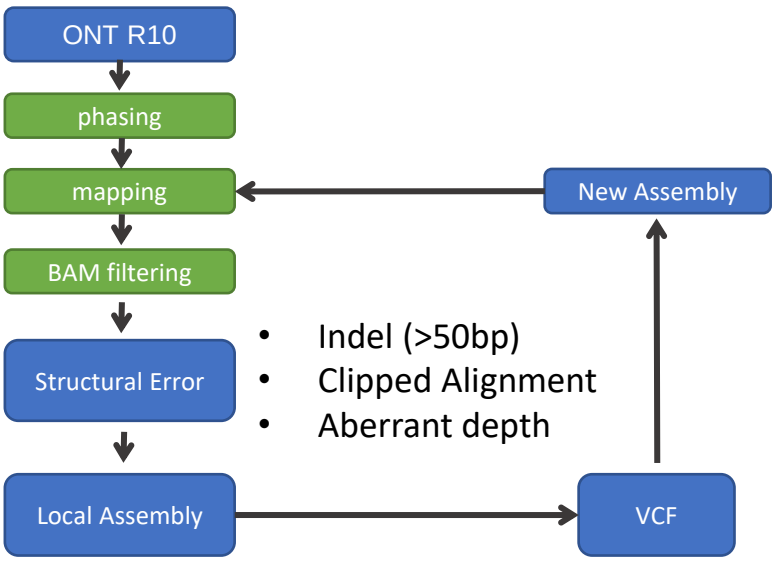
chr2_paternal: 131,255,507-121,280,700



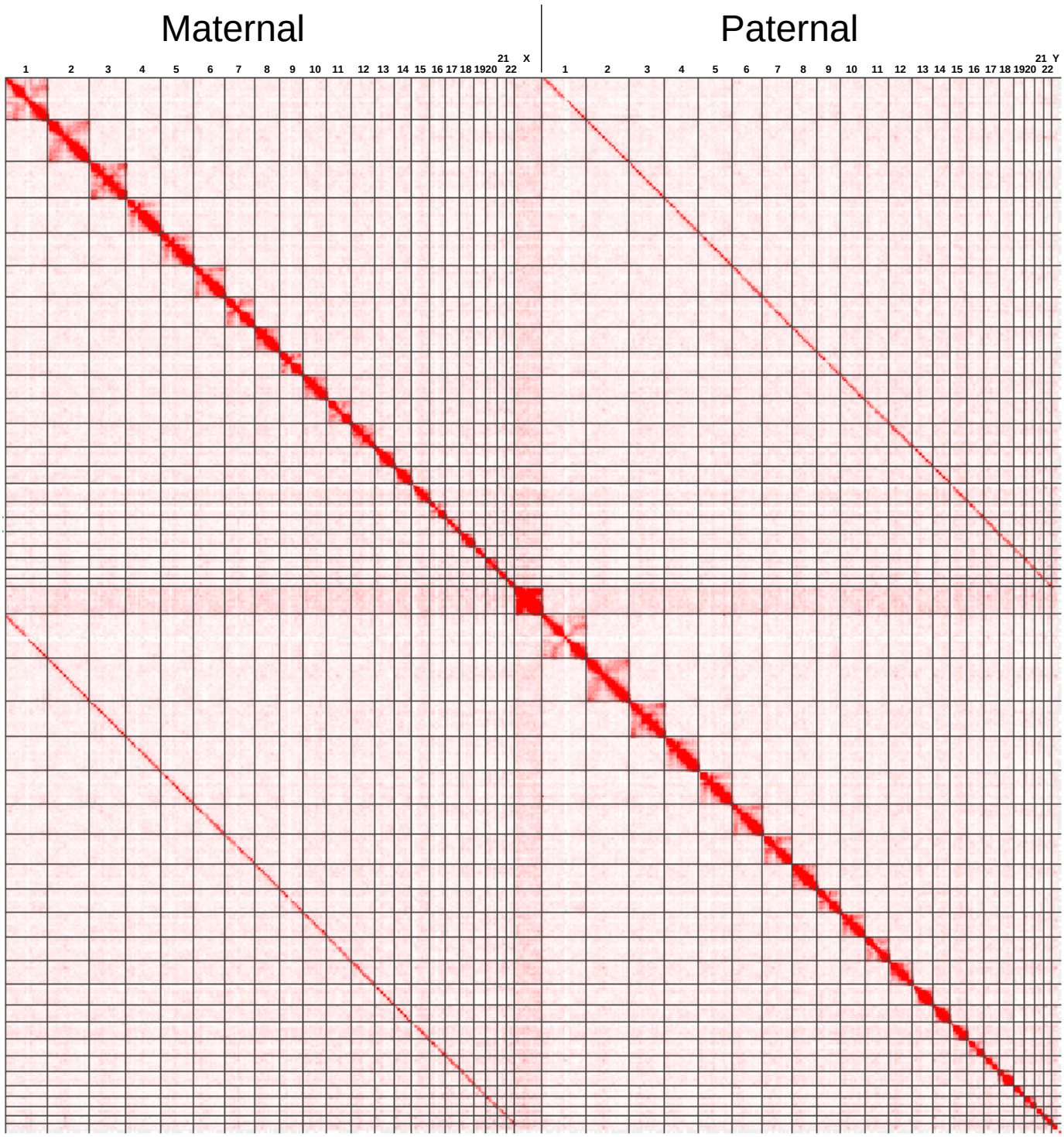
c

chr2_paternal: 87,402,206-87,496,228

Extended Fig. 2

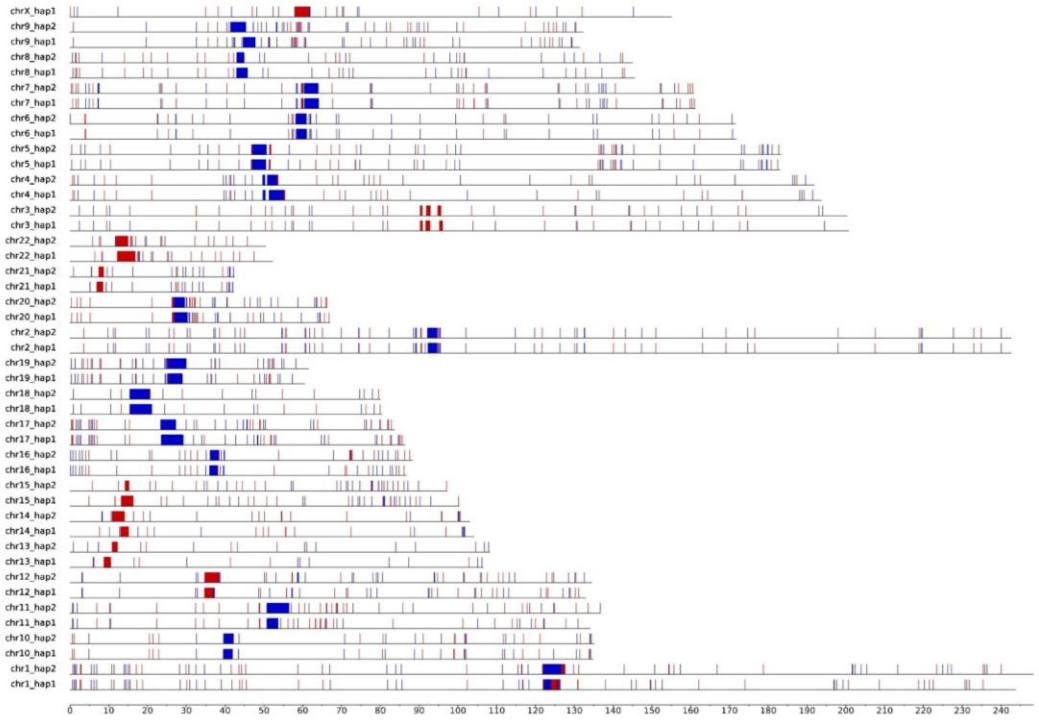


Extended Fig. 3

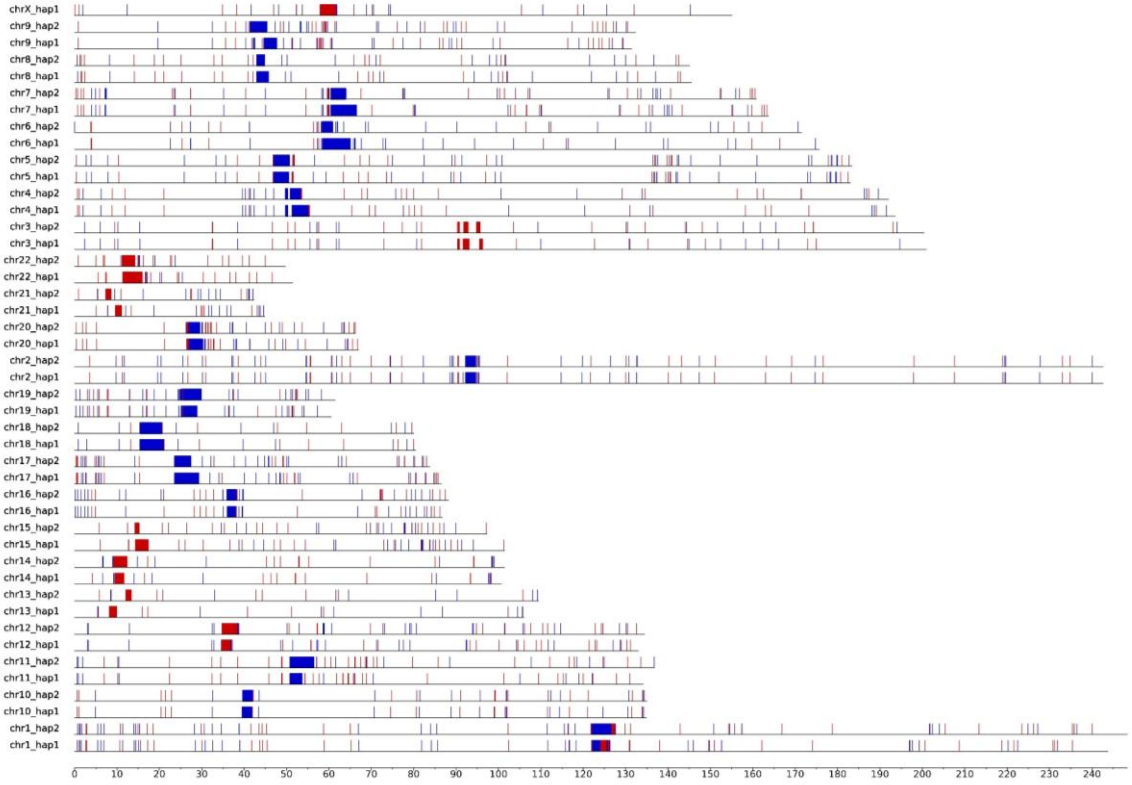


Extended Fig. 4

a. YAOv1.1

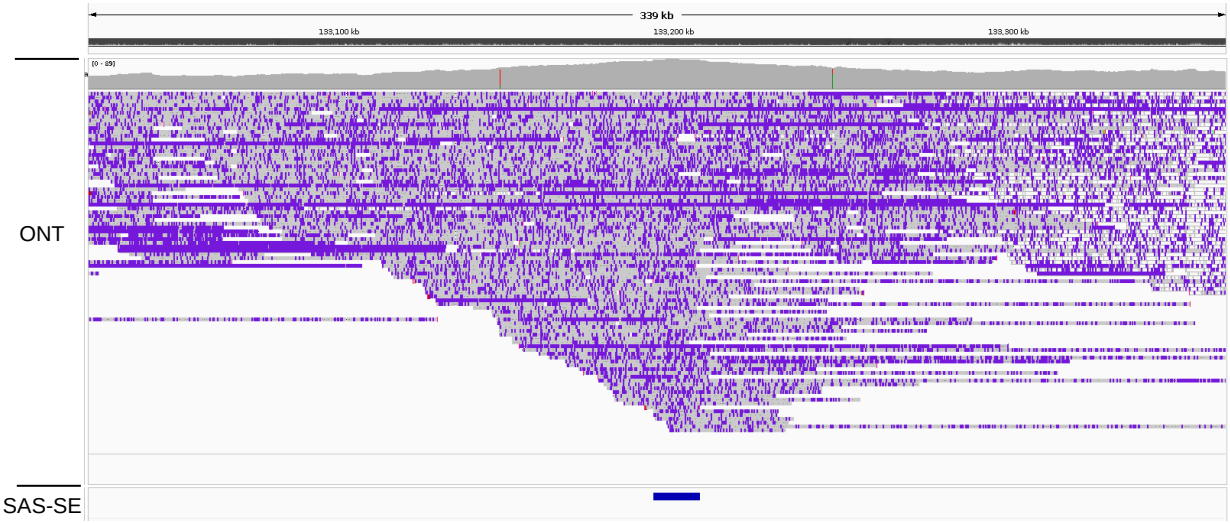


b. YAOv2.0

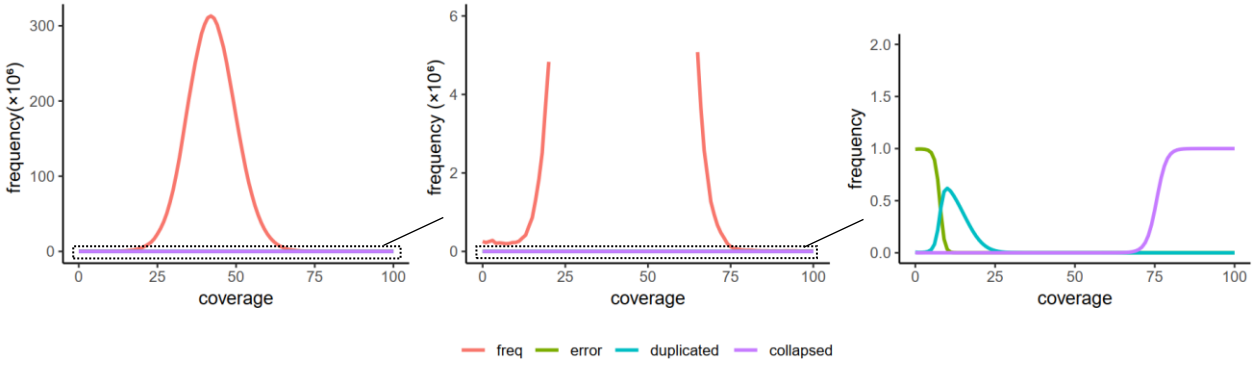


Extended Fig. 5

chr7_maternal: 133,025,150-133,364,953



Extended Fig. 6

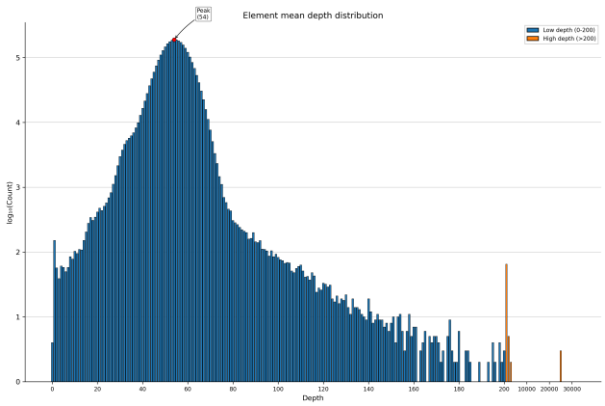


Extended Fig. 7

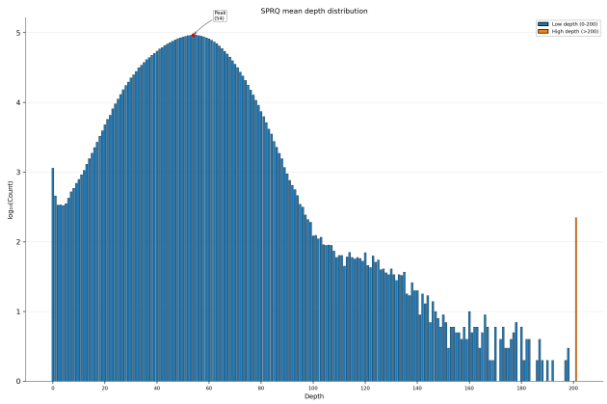
a

v1.1

Element



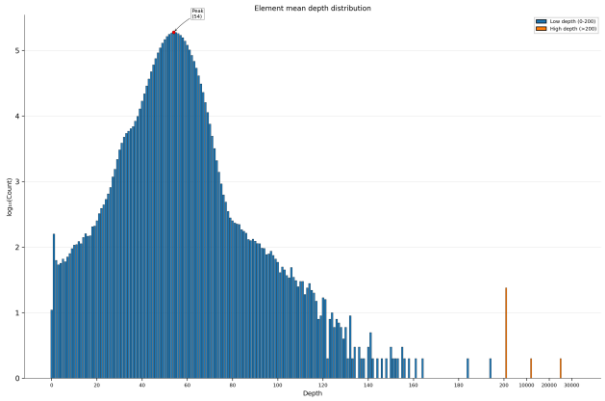
SPRQ



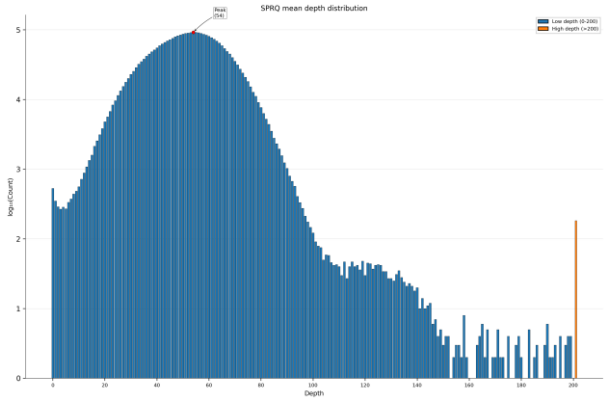
b

v2.0

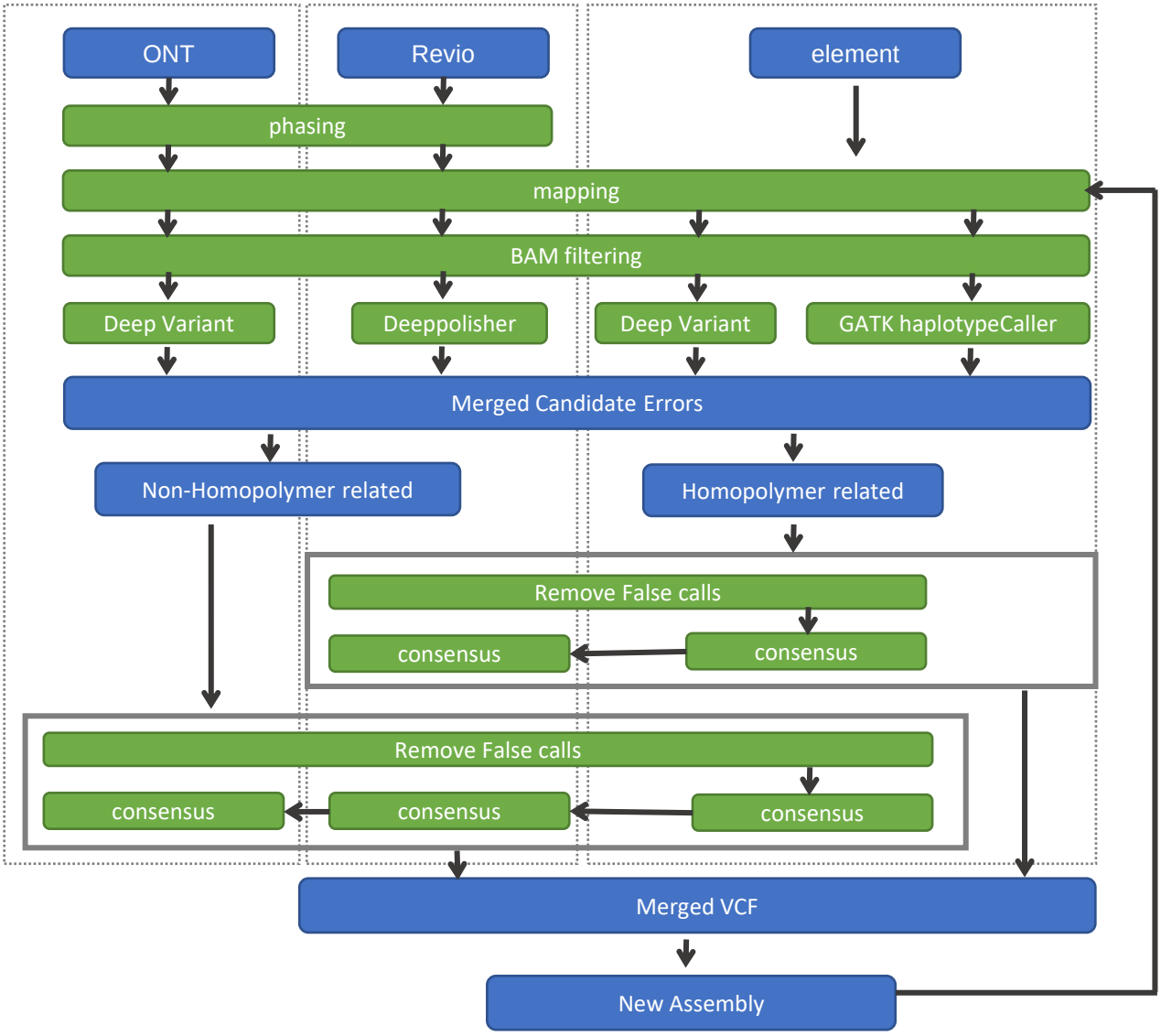
Element



SPRQ

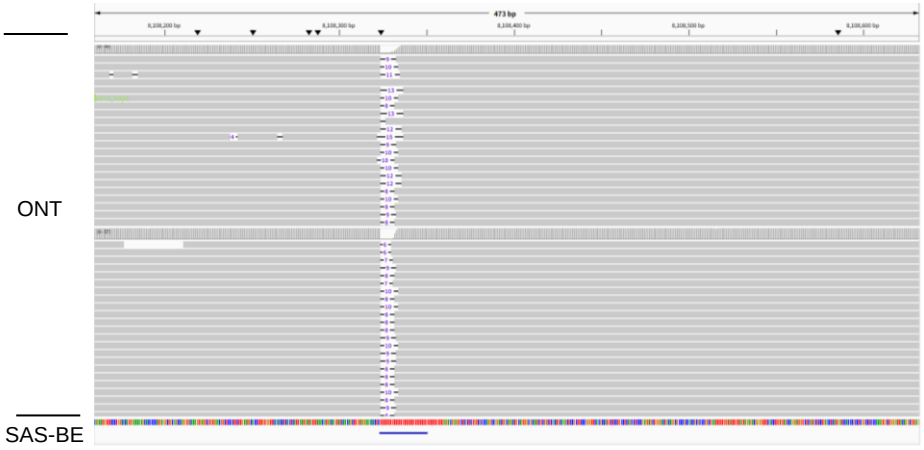


Extended Fig. 8

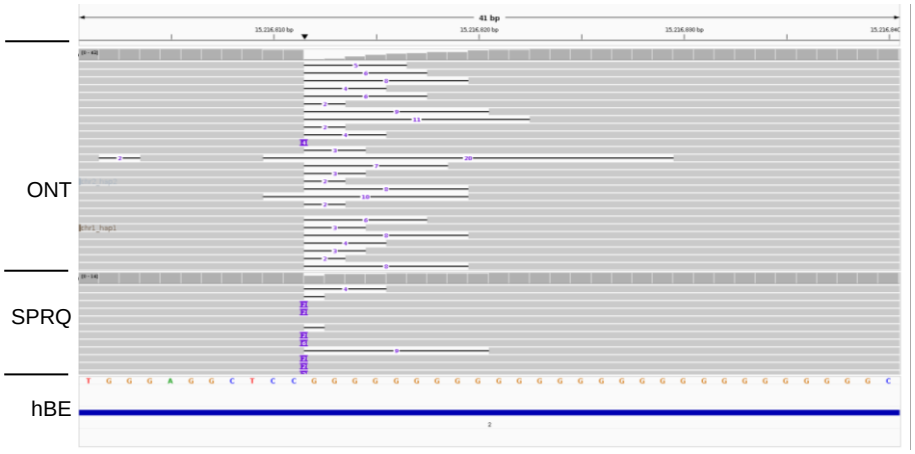


Extended Fig. 9

chr1_maternal: 8,108,160-8,108,631



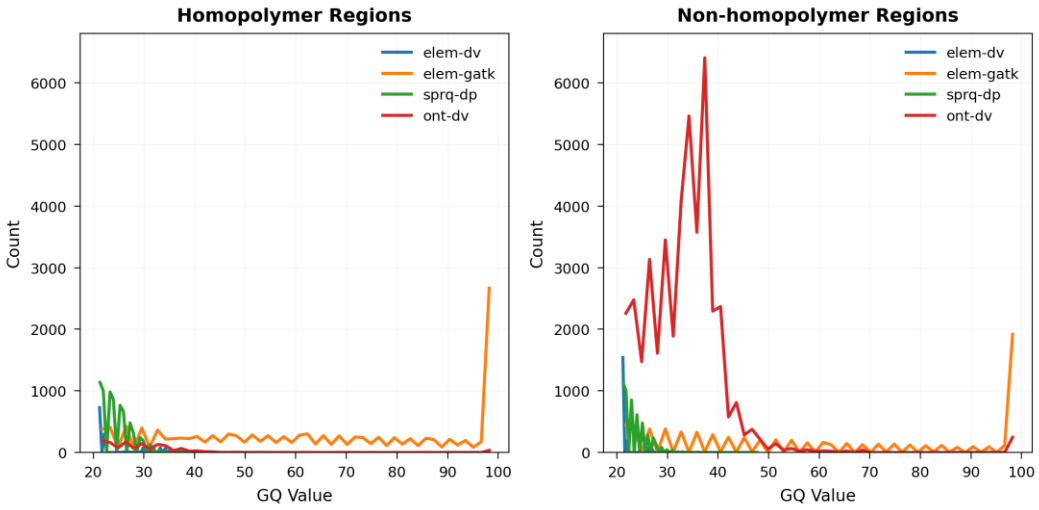
chr1_paternal: 15,216,801-15,216,840



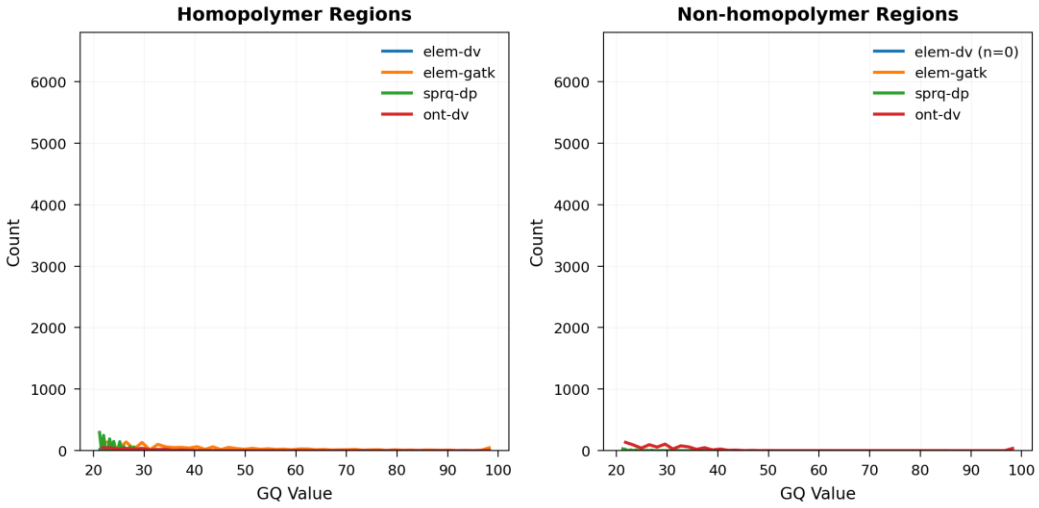
chr1_maternal: 46,238,167-46,238,198

Extended Fig. 10

a V1.1



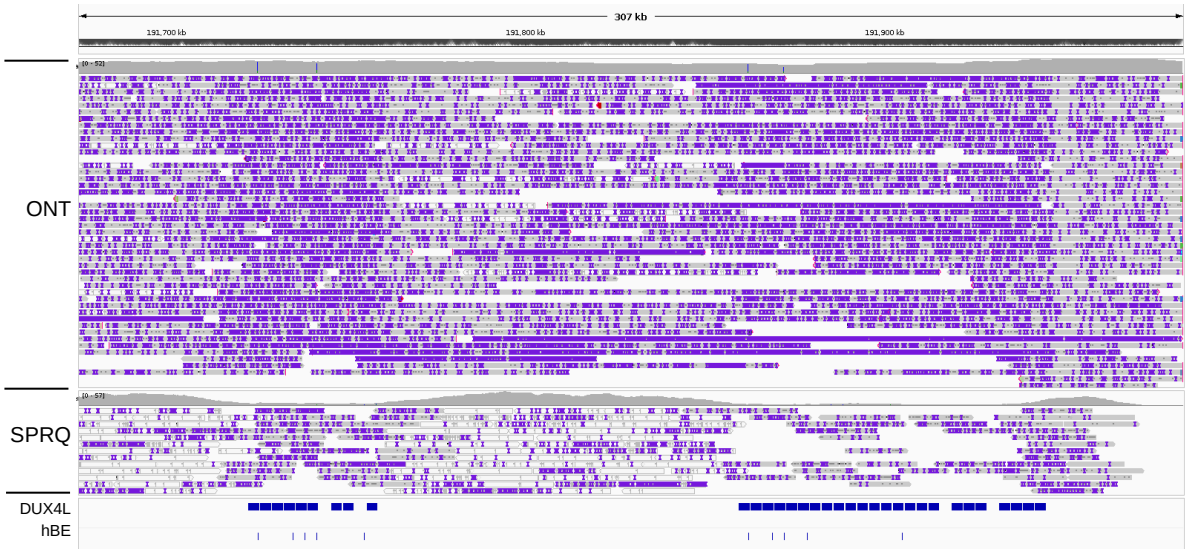
b V2.0



Extended Fig. 1

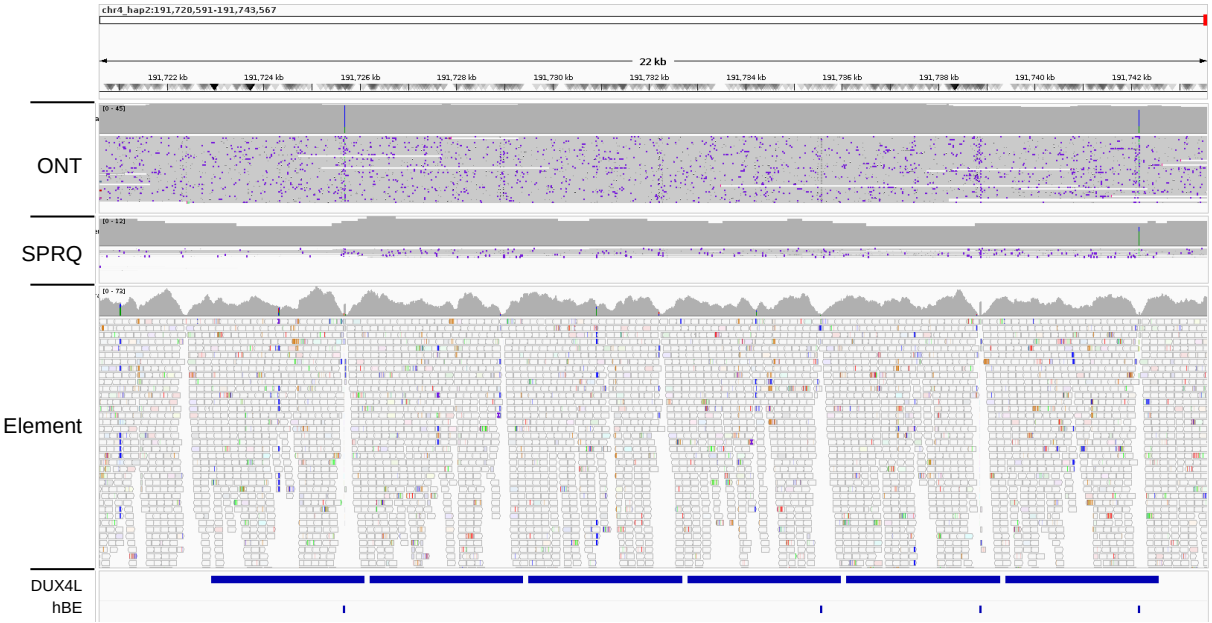
a

chr4_paternal: 191,675,743-191,983,215

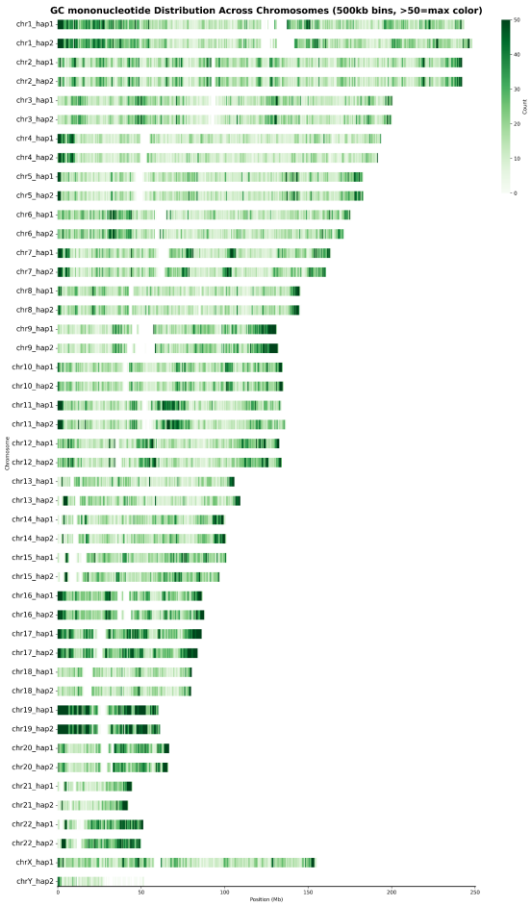


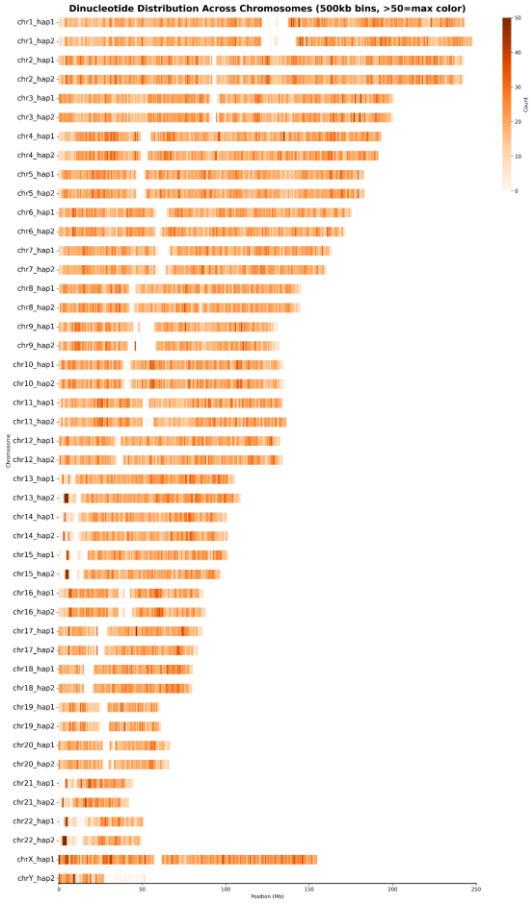
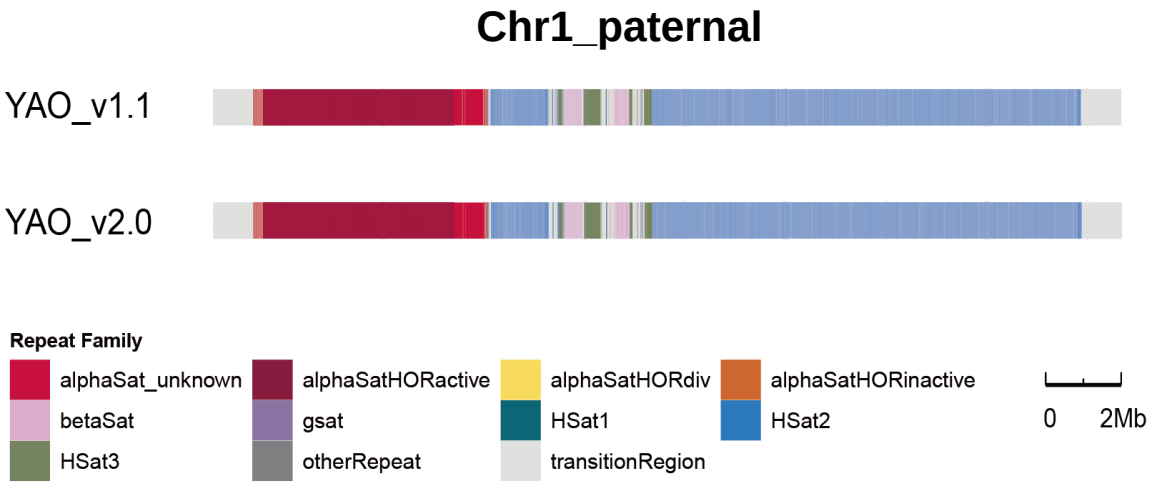
b

chr4_paternal: 191,720,591-191,743,567



Extended Fig. 12





Extended Fig. 13