Approaching an Error-Free Diploid Human Genome Assembly of East Asian Origin

Yanan Chu[1,6], Zhuo Huang[1,2,6], Changjun Shao[1,6], Shuming Guo[3,6], Yiji Yang[1,2], 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[1,2,7*].

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

6. These authors contributed equally

7. Lead contact

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

SUMMARY

Achieving an error-free diploid human genome remains a major challenge. We report T2T-YAO v2.0, a near-perfect assembly of a Han Chinese individual. To systematically evaluate assembly accuracy, we developed Sufficient Alignment Support (SAS), an alignment-based metric that detects structural and base-level errors in windows lacking adequate read support. Building on this, we established a "structural error–first" polishing strategy, correcting misassemblies with ultra-long ONT reads, then refining bases with PWC (Platform-integrated Window Consensus), a multi-platform, window-based polishing pipeline. Together, SAS and PWC provide a rigorous framework for

genome assessment and correction, resolving all detectable structure and non-homopolymer-related errors outside ribosomal DNA clusters. The resulting assembly contains no erroneous 21-mers absent from sequencing data, meeting k-mer–based criteria for an error-free genome. T2T-YAO v2.0 delivers the most perfect East Asian reference to date, enabling precise annotation, benchmarking, and variant discovery—laying a stronger foundation for human genomics and precision medicine.

KEYWORDS

Error-free genome, assembly quality assessment, near-perfect genome, SAS (Sufficient Alignment Support), erroneous k-mer, human diploid genome, Han Chinese, T2T-YAO, ultra-long ONT reads, homopolymer, PWC (Platform-integrated Window Consensus)

INTRODUCTION

Advancements in sequencing technologies and assembly algorithms have rapidly improved genome assembly quality, 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(Church, 2022; Nurk et al., 2022), which achieved unprecedent completeness and correctness. Nonetheless, assembly errors persist, arising from sources such as base-calling inaccuracies, incorrect haplotype assignment of contigs, imperfect read correction, and homopolymer compression during the assembly process(Mc Cartney et al., 2022; Rhie et al., 2021). 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(Aganezov et al., 2022; Formenti and Howe, 2022; Jarvis et al., 2022; Olson et al., 2023), which is still a major challenge. While the last few years have seen dramatic improvements in base-calling accuracy across all major sequencing platforms of ONT, PacBio, and short-read sequencing, an interesting question arises: how close the current technologies can bring us to a perfect or error-free human genome assembly, whose accuracy ensures reliable downstream annotation and analysis?

The remaining errors in an initial assembly can be typically corrected *post hoc* by realigning sequencing reads to the assembly and polishing regions of inconsistency. To achieve this, a variety of structural and single-nucleotide variant (SNV) callers currently available can be applied in parallel to identify and correct assembly errors based on alternative allele detection. Then 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(Rhie et al., 2020). 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 of shorter, high-accuracy reads for subsequent base-level corrections, then introduced PWC (Platform-integrated Window Consensus), a multi-platform, window-based polishing pipeline to refine bases. Starting from our previously published T2T-YAO v1.1, an initial complete diploid genome

assembly of a Han Chinese individual(He et al., 2023), we incorporated additional sequencing data from high-accuracy platforms and iteratively applied SAS evaluation and PWC polishing with SE-First principle. 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, satisfying k-mer–based criteria for an error-free 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 (**Figure 1A**).

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-dependent, 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 mean depth; 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 non-redundant BE and SE windows, respectively, L is the base length of the assembly.

To accurately evaluate the T2T-YAO assembly, we upgraded its raw data to high-accuracy sequencing platforms: 260 Gb Q28 ultra-long ONT reads(Ni et al., 2023) (filtered for >30 kb; N50 >120 kb), 172 Gb Pore-C(Deshpande et al., 2022), 306 Gb HiFi reads (PacBio SPRQ; filtered >3 kb), and 320 Gb Element AVITI(Arslan et al., 2024) Q50 reads (**Table S1**). To reduce ambiguity in haplotype assignment during mapping, ONT and SPRQ reads were phased in advance using haplotype-specific k-mers (hapmers) and mapped in a haplotype-aware manner (see Methods).

We applied the SAS metric to evaluate the previously released T2T-YAO v1.1 assembly and compared the results to those obtained using Merqury, the most widely adopted k-

mer–based quality assessment tool (**Table 1**). SAS identified 21,511 base-level error (BE) windows and 4,438 structural error (SE) windows, yielding QV_{SAS} values of 53.12 and 54.13 for the maternal and paternal haplotypes, respectively. Chromosome-specific QV_{SAS} scores ranged from 45.36 to 61.37 (**Table S2**). In contrast, Merqury using 21-mers reported only 22,556 erroneous k-mers, corresponding to 1,074 estimated errors and Q67.42 across the diploid genome. This quality is notably lower than initially reported(He et al., 2023), as the use of more accurate Element and SPRQ reads revealed assembly errors that were previously obscured by sequencing errors, demonstrating the dependency of k-mer based QV on platform accuracy. Nonetheless, the much more errors identified by SAS than Merqury highlights the markedly greater sensitivity of SAS. Many BE windows detected by SAS lacked corresponding erroneous k-mers, even when more sensitive 31-mers were used, which identified 123,085 erroneous k-mers corresponding to 3,970 estimated errors and Q61.74 (**Table 1**). These results underscore the limitations of k-mer–based methods in capturing all types of assembly errors. As a representative example, we showcased the paternal chromosome 1 to illustrate the differences in error detection between SAS and other metrics (**Figure 2A**).

Closer inspection of BE windows revealed a frequent association with homopolymer sequences, where reads often disagreed on the number of repeated nucleotides. To classify these regions, we defined homopolymer sites as stretches of ≥10 A/T, ≥7 G/C mononucleotide repeats, or ≥10 dinucleotide repeats. Accordingly, BEs were divided into homopolymer-associated (h-BEs) and non-homopolymer-associated (nh-BEs), accounting for 17.17% (3,694) and 82.83% (17,817) of all BEs, respectively (**Table 1**). While nh-BEs were relatively enriched in centromeric regions and corresponded with k-mer–derived errors identified by Merqury, h-BEs were more broadly distributed outside centromeres and were largely undetected by k-mer-based methods (**Figure 2A**). Conversely, some erroneous k-mers reported by Merqury—using the Element-SPRQ hybrid k-mer set, following the protocol in T2T-CHM13(Mc Cartney et al., 2022)—were found outside SAS-flagged BE windows, clustering instead within degenerative GnA or CnT repeat regions. These loci typically had sufficient ONT coverage and were not

flagged as errors by SAS, but dropped out by SPRQ and Element reads due to platform-specific sequencing bias. These discrepancies underscore the fundamental limitations 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 latter arises from Merqury's reliance on the hybrid k-mer sets derived exclusively from high-accuracy Illumina (Element in this study) and PacBio platforms, while excluding ONT data to avoid introducing large number of sequencing error-related k-mers given its relatively higher sequencing error rate.

Given Merqury's limited ability to detect structural errors, we compared SAS with other tools, including Sniffles2(Smolka et al., 2024) (ONT), the widely used structural variant (SV) callers, and coverage-based structural error detectors— Flagger(Liao et al., 2023) (ONT) and nucflag(Vollger et al., 2019) (SPRQ). Sniffles2 identified only 263 confident SV-like errors (supported by ≥60% of local reads), demonstrating limited sensitivity, particularly in repetitive regions (**Figure S1**), where it relies on repeat annotations often unavailable in just completed 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**). However, 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. Notably, nucflag frequently misclassifies misjoins in GnA/CnT repeats and other regions dropped out by SPRQ due to sequencing bias (**Figure 2A**), which obvious damages the reliability of SE calling based on PacBio reads. 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 either classified as SEs if flanking clips are present or as BEs if disagree with phased read 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, 1,662 low-depth regions, 331 contained indels, 61 were marked by clipped alignments, and 1,929 were consecutive windows without spanning reads. Note that individual windows could be assigned to multiple categories. Errors related to clipped alignments and abnormal depth were particularly concentrated in the short arms of acrocentric chromosomes and the centromeres of maternal chromosomes 6 and 7, whereas indels were more evenly distributed across the genome (**Figure 3A**).

Given the critical importance of accurate read alignment for subsequent base-level error correction, we adopt a "Structural Error-first" (SE-first) strategy. This approach prioritizes correction of structural errors using ONT reads before applying PWC base-level polishing with high-accuracy sequencing data. By resolving large-scale mis-assemblies upfront, the SE-first strategy reduce the risk of overcorrection due to artificial alignments in shorter reads, thereby enhancing the overall fidelity of the assembly.

We employed a comprehensive correction pipeline combining alternate variant calls from Sniffles2, local reassemblies using ONT reads, and targeted telomere extension to eliminate all detectable SEs. After each correction round, reads were re-aligned, and SAS was reapplied to reassess error content (**Figure 1B**). Most SEs were resolved after several iterations of this correction–realignment–evaluation loop, though certain sites of clipped alignment required manual inspection and correction. When local assembly fails to resolve mis-assemblies, incorporation of Pore-C chromosomal configuration data often enables correction. The resulting contact map validates the structural integrity of the final assembly, confirming accurate haplotype phasing and the correct pairing of short and long arms across the ten acrocentric chromosomes, a particularly challenging structural feature to resolve (**Figure S2**). In telomeric and subtelomeric regions, ONT

reads often exhibited heterogeneity that impeded successful reassembly. In such cases, stratifying ONT reads based on characteristic SNVs and assembling the dominant haplotype group enabled resolution. Following these polishing efforts, all telomeres were extended to an average length of 8.7 ± 2.6 kb, with all exceeding 5 kb except for three in the q arms of maternal chromosome 12 and paternal chromosomes 21 and 22 (**Figure 4A**). The telomeres in v2.0 are closer to conventionally accepted lengths, and far more complete than those in T2T-CHM13 (2.7 ± 0.7 kb).

Ultimately, T2T-YAO v2.0 becomes free of detectable SEs outside of ribosomal DNA (rDNA) clusters (**Figure 2B, 3A**). Centromeric collapses, particularly in the higher-order repeat (HOR) arrays of maternal chromosomes 6 and 7, are also resolved (**Supplemental information file 1**, **Figure 4B, Figure S3A**). One potential exception is observed in a 12 kb high-depth region upstream of the *TSPAN33* gene on maternal chromosome 7q32. Although the local ONT depth (84–89×) marginally exceeds the global cutoff (84×), the local SPRQ depth (45×) is quite near mean depth (51.2 ± 13.5), and the sequences there and of flanks are fully supported by ONT and SPRQ reads without signs of clip or misjoin. Thus, the reason for the elevated ONT depth in this particular region is still unknown. Further inspection reveals that this region is composed of L1M and Alu elements which sequences are widely present in human genome with millions of copies(Belancio et al., 2009). We suspect that the repetitive nature of this region may contribute to the elevated depth where 22.8% of mapped reads are shorter than 60 kb and exhibiting MAPQ =1 (**Figure S4**).

Aside from this, ONT read depth is within the 20%~200% mean depth range and follows a Gaussian distribution, with extremely few regions tallied by Flagger as false duplication (2.81 Mb, 0.047%) or collapsed (1.55 Mb, 0.026%) (**Figure 5A**), with ~90% present in rDNA clusters. All overlapping issue regions identified by Flagger and nucflag are confined to rDNA arrays or telomeric regions, supporting the structural integrity 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% coverage by non-split ONT reads when rDNA is masked, confirming the absence of unresolved SEs outside rDNA.

Depth fluctuations in v1.1 are much more pronounced in Element and SPRQ compared to ONT (53.5 ± 42.0 and 51.2 ± 13.5 vs. 42.4 ± 8.1, averaged in 2 kb windows), reflecting their greater susceptibility to artificial alignments due to shorter read lengths. Such falsely collapsed alignments frequently introduce incorrect variant calls and promote overcorrection during base-level polishing. In T2T-YAO v1.1, alignment of SPRQ and Element reads revealed widespread regions of abnormally high depth (>2-fold mean depth), totaling 4.71 Mb and 2.52 Mb, respectively, enriched for artificial alignments (**Figure 5B**). These misaligned reads often contain consensus mismatches to the reference and represent a major source of overcorrections. After SE correction, the collapsed regions were reduced to 3.43 Mb (SPRQ) and 0.93 Mb (Element), including those within rDNA (**Figure 5B**). Thus, SE correction reduced false alignments in high-accuracy data, mitigating the risk of over-correction and improving 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 (**Figure 3B**). Each flagged window is corrected using PWC pipeline, which generates consensus sequence (supported by ≥60% reads) from spanning reads (depth ≥10) for each flagged window. PWC follows an across-platform prioritization scheme, where consensus supported by all platforms is directly applied. In case where the three sequencing platforms disagree, corrections supported by two platforms are adopted for nh-BEs, otherwise, platform-specific reliability is applied in the order: Element > SPRQ > ONT. If no consensus is obtainable, ONT reads spanning the region are locally reassembled to correct the sequence, which often enables consensus

generation by mitigating artificial alignments (**Figure 1C**). For h-BEs, local ONT-based consensus or assembly is omitted due to the known limitations of ONT sequencing in accurately resolving homopolymer lengths (**Figure 6A**). The stringent conditions for generating consensus minimized overcorrection caused by alignment artifacts during the polishing rounds. Read alignment and polishing are repeated iteratively until the number of BEs is stabilized.

To correct additional base-level errors potentially missed by SAS, we applied multiple variant callers: Deeppolisher(Mastoras et al., 2024b) (SPRQ), DeepVariant(Poplin et al., 2018a; Shafin et al., 2021) (ONT and Element), and GATK HaplotypeCaller(Poplin et al., 2018b) (Element) to call variants by mapping reads to the T2T-YAOv1.1 diploid genome. These tools typically reported several thousand to over twenty thousand homozygous variants (*i.e.* supported by the majority of aligned reads from the corresponding platform), yielding a total of 63,398 non-redundant calls with genotype quality (GQ) ≥20, with a little overlap among tools. Because obvious false positives were present, we first filtered out variants whose reference sequence, along with 10 bp flanking sequences, were supported by ≥60% local depth and ≥10 exact-matching reads from any platforms. This filtering step yielded 43,444 potential SNV-like errors across both haplotypes, of which 40.21% were associated with homopolymer regions, an even higher proportion than the observed rate of h-BEs (**Figure 5C**). The elevated frequency of homopolymer-associated errors likely stems from the homopolymer compression step used in the initial T2T-YAO assembly with Verkko(Antipov et al., 2025; Rautiainen et al., 2023). Of these SNV-like errors, 43.90% of non-homopolymer-associated variants overlapped with SAS-detected nh-BEs, whereas only 7.08% of homopolymer-related calls overlapped with SAS h-BEs. This discrepancy highlights the challenge of identifying true errors via alignment, particularly those associated with homopolymers. To avoid overcorrection from alternative sequences proposed by SNV callers, we applied the same PWC pipeline to potential SNV-like errors outside BE windows: replacing the variant site and its 10 bp flanking sequence with a read-derived consensus sequence rather than the variant caller's prediction. In many cases, the

consensus sequence was identical to the reference, meaning that only a small proportion of sites beyond BE windows are truly errors could be corrected.

After several rounds of iterative polishing, T2T-YAO v2.0 contains only 336 nh-BEs (all within rDNA regions) and 837 h-BEs located in repetitive regions (including 489 in rDNA) that remain unresolved due to the inherent limitations of current sequencing technologies (**Table 1, Figure 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 (**Figure 5C**). 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 supported by reads of other platforms. The sharp drop in SNV-like errors without extensive corrections suggests that many of them, particularly outside BE windows are likely false calls derived from artificially aligned reads bearing variant sequences.

An illustrative example of unresolved h-BEs can be found in the 3.3 kb D4Z4 repeat array on chromosome 4q35(Lemmers et al., 2018). In this region, the polyC tracts between DUX4-like genes can not be confidently resolved: ONT and SPRQ reads struggled with homopolymer sequencing, while Element reads suffered from sequencing bias and mapping ambiguity (**Figure 6B**). Nonetheless, the remaining h-BEs account for only 0.04% of the total 2,412,242 homopolymer sites genome-wide. Although not completely resolved, T2T-YAOv2.0 provides an accurate count and distribution of homopolymer sites in a human genome. 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. Overall, the distribution of homopolymers clearly avoids centromeric regions and exhibit very similar pattern across haplotypes (**Figure S5**). Notably, polyC/polyG and poly-dinucleotide repeats are approximately 10-fold enriched in rDNA sequences (up to ~600/Mb), whereas

polyA/polyT tracts are much sparser in the short arms of acrocentric chromosomes, encompassing rDNA.

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 as described above) 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 1, Table S2, Supplemental information file 1**).

T2T-YAO v2.0 contains tens of kilobases of GnA/CnT repeat sequences that are poorly covered 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-mers are encompassed by the erroneous 31-mers, we focused our inspection on the 31-mers. These were consistently found in regions with near-zero coverage (depth ≤1) in Element and SPRQ data but showed strong support from ONT reads. The 31-mers clustered into 1,127 discrete regions (after merging overlaps), collectively spanning 76,248 bp. Of these, 70% regions (totaling 53,307 bp in 674 regions) had extremely high GA/CT content (>90%), while the rest were composed of non-canonical 6-mers found in telomeric sequences or GC-rich tandem repeats, patterns consistent with known sequencing biases(Freeman et al., 2020; Ross et al., 2013) (**Table S4**). Given the presence of these sequences in human genome that are dropped out by PacBio and short-read

sequencing (SRS), achieving Q100 quality is unattainable with the current PacBio/SRS hybrid k-mer set, even with the complete correction of all assembly errors.

The ONT data used in this study were generated using the R10 flow cell, known for high base-calling accuracy (Q28; >99.8%). To leverage this accuracy, we incorporated k-mers from the ONT dataset into our analysis, stringently filtering out k-mers of occurrence ≤4 to exclude those of potential sequencing errors, compared to ≤1 for Element and SPRQ reads. We first use this expanded trinary k-mer set to assess T2T-YAO v1.1, which QV only shows a slight improvement over the binary set, indicating minimal inclusion of k-mers caused by ONT sequencing errors. When we applied the trinary k-mer set to assess T2T-YAO v2.0, the QV for 21-mers increased to "infinite," representing an Merqury-defined error-free diploid genome. The QV for 31-mers also improved significantly, reaching 96.29 for the maternal haplotype and 88.73 for the paternal, with only 143 erroneous 31-mers remaining (**Table 1**). All paternal erroneous 31-mers are confined to rDNA regions, while those of maternal are located in two h-BE loci of GnA repeat regions containing several polyG tracts (≥8 bases), corresponding precisely with the residual errors detected by the SAS metric.

Ensuring the accuracy of annotation

We generated the gene annotation of T2T-YAO v2.0 by transferring annotations from T2T-CHM13 v2.0 using Liftoff(Shumate and Salzberg, 2021). rDNA clusters and Ig/TCR VDJ segments, which are highly variable among individuals, were annotated separately and excluded from gene count statistics. Of the 57,377 genes annotated in T2T-CHM13, 56,601 (98.65%) were successfully lifted to T2T-YAO, including most protein-coding genes, lncRNAs, and pseudogenes. A total of 115 protein-coding genes (0.59%) failed to lift over in either haplotype, while lncRNAs and pseudogenes showed higher loss rates, 272 (1.66%) and 339 (1.99%), respectively (**Table 2, Table S5**). To recover the missed protein-coding genes, we applied miniprot (Li, 2023) which enabled the annotation of additional genes and extra gene copies based on MANE(Morales et al., 2022)(v1.4) protein-to-genome alignments, adding 14 genes (11 absent in T2T-CHM13)

and 30 extra copies. In total, 773 genes, including 112 protein-coding genes, were absent in T2T-YAO compared to T2T-CHM13. Notably, all missing protein-coding genes were either uncharacterized ("LOC" genes) or members of gene families such as *FAM90* and *DEFB*, which are often confused among paralogues when lifting over due to high sequence similarity. The overall lift-over rate was slightly lower than that of T2T-HG002(Hansen and al, 2025), where only 39 distinct genes (3 protein-coding) were missing across both haplotypes. Since both HG002 and CHM13 are of European origin, this discrepancy may reflect population-related differences in gene content, particularly in the number and sequence of paralogues. As expected, protein-coding genes are the most conserved, whereas lncRNAs and pseudogenes exhibit greater variability between any genomes (**Table 2**). Most lifted-over genes in T2T-YAO v2.0 retain synteny with their counterparts in T2T-CHM13, with a few exceptions, particularly on chromosome Y (**Figure S6**). In total, 4,143 extra copies of unique genes were identified in the diploid T2T-YAO, of which only 9.5% are protein-coding (**Table 2**, **Table S5**). Notably, about one-third of the extra lncRNA copies are located on chromosome Y, with their biological significance largely unknown (**Figure S7**).

For the protein-coding genes successfully lifted in T2T-YAO, we identified 254,704 transcripts in both haplotypes. Among these, 194,042 (76.18%), including 374 corrected by Lifton (Chao et al., 2025), encode protein sequences identical to their T2T-CHM13 counterparts (**Table S6**). Of the transcripts with altered protein sequences, 458 (0.75%) showed markedly reduced sequence identity (<80%), suggesting potential impairment of protein function. Considering both alleles, 363 non-redundant transcripts corresponding to 200 genes exhibit either decreased identity (<80%) or absence in both haplotypes. These findings highlight ancestry-related differences in gene content and function, emphasizing the need to refine and complete the annotation of T2T-YAO v2.0 through additional strategies beyond liftover, such as *de novo* gene prediction, transcriptome sequencing from diverse East Asian tissues and multiple lineages differentiated from iPSC-YAO cells.

Comprehensive correction of SEs and nh-BEs in T2T-YAO v2.0 eliminated all detectable errors in centromeres, which naturally lack homopolymer sites, providing unprecedented resolution of centromeric sequences. We refined the annotation of centromeres, higher-order repeats (HORs), and other satellite elements across all 46 chromosomes in T2T-YAO v2.0 (**Figure 4B, Supplement Information file 1**) (Altemose et al., 2022; Logsdon et al., 2024). Compared to v1.1, subtle but significant differences can be observed, exemplified by updates to the paternal chromosome 1 assembly (**Figure S3B**), highlighting the importance of high-fidelity polishing for accurate centromere characterization. In v2.0, we identified 536 active HOR monomer types, comprising 320,424 sequences. These sequences showed limited within-type variance, with 86.23% of sequences being 100% identical to their type consensus and 98.61% showing <5 bp variance (identity >97%). Comparison of HOR variants between maternal vs. paternal autosomes identified three major haplotype-specific patterns (>90 repeats), two in paternal chromosome 17 and one in maternal chromosome 6. Besides these, T2T-YAO has five additional major patterns absent from both T2T-CHM13 and CHM1, but lacks two compared to them (**Table S7**). The presence of these patterns suggests population-specific differences in HOR variants, warranting further investigation into centromere diversity across populations.

Using the recently developed Genomic Centromere Profiling (GCP) pipeline, a method analogous to fragment length analysis for characterizing the position, orientation, and organization of centromere protein B (CENP-B) binding motifs(Corda and Giunta, 2025), we constructed structural models for centromeric repeat clusters in T2T-YAO. These models confirmed the correctness of the v2.0 centromere assemblies and revealed extensive variation in centromere size, structure, and repeat composition among chromosomes(Logsdon et al., 2024). Although maternal and paternal centromeres of the same chromosome generally exhibited similar organization, clear inter-haplotype differences were observed in repeat size. For instance, the HORs of maternal chromosomes 6 and 7 were substantially expanded compared to their paternal counterparts in T2T-YAO (**Figure S3A**).

Implications in accurate benchmark and variant calling

High-quality, ancestry-matched references are known to enhance variant discovery by improving read mapping (Aganezov et al., 2022; Guo et al., 2024). With T2T-YAO, the completely assembled centromeres now enable systematic analysis of variation in these previously inaccessible regions. We first examined the small variants between T2T-YAO's paternal and maternal haplotypes in 100 kb windows of alignable regions across autosomes. Genome-wide, the density of small indels (1-3bp) was 0.14/kb and SNP density was 0.89/kb with a Ti/Tv ratio of 1.78. Centromeric regions showed strikingly different patterns, reflecting extensive haplotype divergence as illustrated in paternal chromosome 1(**Figure 7A**). The amount of alignable sequence dropped sharply, especially in satellite DNA outside HORs. Meanwhile SNP density was markedly elevated, 4.2-fold in HORs (3.77/kb) and 3.0-fold in other satellites (2.65/kb). Most variants in HORs arose from inactive or divergent HOR units, accompanied by a reduced Ti/Tv ratio of 0.72, consistent with previous report (Logsdon et al., 2024). Indels were rare, occurring at 0.08/kb in HORs and 0.06/kb in other satellites, compared with 0.17/kb in other repeat regions (**Table S8**).

When aligned T2T-YAO to the ancestry-unmatched T2T-CHM13, SNP and indel densities rose across the genome, with the strongest effects in centromeres, which showed further reduced alignability and higher variant density (**Figure 7B, Table S8**). While overall Ti/Tv kept relatively stable, the distribution of variant types across repeats shifted with the reference, highlighting the importance of ancestry-matched references for accurate variant calling in complex regions.

Coding regions were strongly constrained. SNP density in CDS was 0.46/kb—half the genome average—and indel density was 7.65/Mb, about one-twentieth. The Ti/Tv ratio rose to 2.65, reflecting purifying selection, and the missense-to-synonymous ratio (dN/dS) of SNPs was reduced, both annotated using VEP(McLaren et al., 2016).

However, using unmatched or lower-quality references of CHM13 and GRCh38, inflated variant density in CDS reduced dN/dS due to inflated synonymous calls from divergent alignments and selective depletion of missense variants (**Figure 7C**). Predictions with LOFTEE (Karczewski et al., 2020), which relies on population variant data coordinated in GRCh38, identified only 91 SNPs and 35 indels as high-confidence loss-of-function (LOF) variants, none in CDS regions (**Figure 7D**). Most LOF variants were homozygous and affected uncharacterized or gene family members, consistent with genes missed in liftover, implying potential confusions in variant calling among paralogs.

Finally, telomeres in T2T-YAO v2.0 were extended and corrected to error-free status across all chromosomes. Most contain non-canonical 6-mer repeats interspersed with canonical (TTAGGG)n motifs, especially at subtelomeric transitions (Lyčka et al., 2023) (**Figure 4B**). Motif composition and repeat length varied widely, without haplotype-, chromosome-, or arm-specific patterns, confirming earlier reports (Stephens and Kocher, 2024) but now with direct sequence evidence and strong read support, intriguing further investigation into telomere variations and their biological significance.

DISCUSSION

An error-free diploid human genome holds immense value for calibrating variant calling pipelines and enhancing the accuracy of clinical genetic diagnoses(Jarvis et al., 2022; Wagner et al., 2022). 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(Telomere-to-telomere-consortium, 2023). Given the approximate size of the human diploid genome (~6 billion bases), achieving Q100 implies constructing a truly error-free genome. The validation of T2T-YAO v2.0 by 100% coverage of non-split ONT reads except at rDNA regions and free of erroneous 21-mers absent from the trinary k-mer set, thus qualifying for an "infinite" QV under Merqury standards, further demonstrate the power of the SAS framework, PWC pipeline, and our "Structural Error-first" strategy. The resulting near perfect genome enables accurate annotation of genome components and benchmarks of variant calling. However, technical limitations

in both genome evaluation and error correction exist, preventing it from being deemed absolutely error-free.

Achieving an error-free genome requires assembly quality metrics capable of evaluating accuracy at such level. These metrics encompasses multiple dimensions, including completeness, correctness, and continuity. Of these, completeness is relatively straightforward to assess via k-mer-based approaches(Rhie et al., 2020). However, accurate evaluation of correctness and continuity remains more challenging and typically relies on alignment-based methods(Li et al., 2023; Mastoras et al., 2024a). Most alignment-based tools identify discrepancies between the genome assembly and realigned reads and confined to a single sequencing platform. These discrepancies often include both false positives arising from artificial alignments of heterogeneous reads, as well as false negatives missed due to sequencing dropout. Consequently, integrating alignment discrepancies across platforms is not a trivial problem as simply merging or intersecting them does not reliably identify true assembly errors.

To overcome these challenges, we developed SAS, which evaluate alignment support rather than discrepancies with the assembly. This reverse strategy facilitates integration of regions supported across multiple platforms, avoiding the need to set arbitrary cutoffs to balance false positives and negatives. Assessing supportive alignments incurs a greater computational burden than simply examining discrepancies embedded in the output files of aligners. To improve efficiency, we first quickly filtered for regions perfectly covered (depth ≥5) by Element reads without mismatches or clips, which account for 99.2% of the T2T-YAO genome, thereby greatly reducing the overall computational load. SAS implemented a window-based approach, which offers higher sensitivity in identifying BEs than variation callers of single base resolution which read support is occasionally unreliable particularly related to indels in repetitive regions. A window (50-mer) is considered supported only if at least five reads (representing >10% of the mean sequencing depth in our case) cover it exactly, a criterion more stringent than k-mer-based methods like Merqury, which typically require only two occurrences of

21- or 31-mers. Structural errors can likewise be detected in the window-based strategy, by flagging 2 kb windows lacking sufficient spanning reads, and scanning 30 kb windows without spanning reads to verify continuity in broader genomic context.

Theoretically, an alignment-based approach can identify all assembly errors and correct them according to the consensus of aligned reads—provided that the authentic sequence is fully covered by reads and the reads are correctly mapped to their original locations. However, artificial alignments are inevitable, especially with shorter reads, compromising the accuracy of error detection and correction. Therefore, minimizing alignment artifacts is crucial. Previously, MapQ (mapping quality, reflecting the confidence that a read is placed in a unique place of the genome) was commonly used to filter ambiguous alignments that are typically false alignments. However, MapQ is ineffective in diploid genomes, as most reads receive a zero MapQ score due to the high similarity between haplotypes as well as recent duplications within a haplotype. This limitation explains why MapQ-dependent methods, such as GCI(Chen et al., 2024) and DeepVariant(Shafin et al., 2021), often produce false calls. In this work, we apply three strategies to reduce artificial alignments:1) A “SE–first” approach that leverages ultra-long ONT reads to define genome structure, similar to the proprietary strategy used in the chromosome X project(Miga et al., 2020). 2) Phasing ONT and SPRQ reads into haplotypes based on the hapmers they contain prior to mapping. We use stringent criteria—such as larger k-mers and significant hapmer differences—to avoid false assignments caused by noise. 3)Applying filters to remove secondary alignments and multi-split reads instead of MapQ. All together, these measures are more effective then MapQ in reducing false alignment while preserving the correct ones.

Based on the near-perfect T2T-YAO v2.0, our results in variants between haplotypes establish that ancestry-matched references are critical for accurate variant calling in complex regions. When using unmatched genomes, SNP and indel densities were artificially inflated, alignable fractions collapsed, and functional interpretations were distorted. These effects were most pronounced in centromeres and paralogous genes,

where divergence across haplotypes exaggerated spurious variants and obscured true biological differences. Even in coding regions, unmatched references skewed dN/dS by inflating synonymous calls and underestimating missense variation, particularly in gene family members. These findings underscore a broader principle: a single reference genome cannot serve all populations. The ancestry-matched assemblies are more powerful in reducing mapping artifacts, resolving repeats, and sharpening functional inference than uniformed reference. As population-specific T2T references emerge, they will provide not only more accurate baselines for variant discovery but also essential resources for precise human genomics.

Limitations of the study

Despite its sensitivity, SAS still misses some base-level errors in repetitive regions where misaligned short 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(Vollger et al., 2019), 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(Li and Durbin, 2024). 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.

RESOURCE AVAILABILITY

Lead contact

- Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Yu Kang (kangy@big.ac.cn).

Materials availability

- This study did not generate new unique reagents.

Data and code 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, genome annotation, and associated datasets can be freely downloaded without application and are also publicly accessible at https://github.com/KANGYUlab/T2T-YAO-resources.
- Code for SAS, error correction, and all other workflows are available at https://github.com/KANGYUlab/sas-pipeline and https://github.com/KANGYUlab/PWC-polishing-pipeline.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

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. Y.Y.,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.

DECLARATION OF INTERESTS

All authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the preparation of this work, the author(s) used Deepseek and chatGPT in order to improve language fluency, clarity, and readability. After using this tool or service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

SUPPLEMENTAL INFORMATION

Document S1.

Table S1-S8

Supplemental Information file 1: S1_Figures 1-45.

Supplemental Information file 2: S2_Figures 1-24.

FIGURE TITLES AND LEGENDS

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

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.

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 is 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, which is full of repeat sequences.

Figure 5. Read depth distribution and SNV-like errors in T2T-YAO v1.1 and v2.0

(A) Depth distribution of ONT reads aligned to T2T-YAO v2.0, along with misassemblies predicted by Flagger. (B) Comparison of read depth distributions for SPRQ and Element datasets in T2T-YAO v1.1 and v2.0. (C) Quality score distribution of SNV-like errors identified by SNV callers in T2T-YAO v1.1 and v2.0.

Figure 6. Unresolved homopolymer-related base-level errors

(**A**) Examples of discordance among ONT and SPRQ reads at homopolymer sites. (**B**) Example of unresolved homopolymer sites in the D4Z4 repetitive region on chromosome 4q35. The sequencing depth and read alignment from multiple sequencing platforms (ONT, SPRQ and Element) are displayed with peaks above each track representing the sequencing depth at each position. The gray horizontal bars indicate the sequencing reads, with colored vertical bars denoting mismatches in these reads. The bottom tracks highlight the position of DUX4L genes and unresolved h-BE.

Figure 7. Impact of reference genome on variant calling

(A) Density of SNPs and small indels, and Ti/Tv ratio in 100 kb windows along paternal chromosome 1. Upper panel: whole chromosome; lower panel: centromere and flanking regions, with repeat annotations indicated in the background of the peak map. **(B)** Comparison of SNP and indel densities, and Ti/Tv ratios across different repeat classes using ancestry-matched (inter-haplotype) versus unmatched (T2T-CHM13) references. **(C)** Comparison of SNP and indel densities, Ti/Tv ratios, and missense-to-synonymous variant ratios in CDS regions across different references. **(D)** Counts of high-confidence loss-of-function variants per chromosome, shown separately for SNPs (upper panel) and small indels (lower panel). Heterozygous variants, present on a single haplotype, are indicated by hatched bars stacked above the homozygous variants.

SUPPLEMENTAL FILES

Figure S1. Example of a structural error in a repetitive region missed by Sniffles2.

The vertical bar in top track indicates the structural error detected by Sniffles2 (missed in panels A and B, present in panel C but incomplete). The blue peaks above middle track represents the ONT sequencing depth at each position. The gray horizontal bars indicate the ONT sequencing reads, with colored vertical bars indicating mismatches. Black short lines within the reads indicate sequence deletions (i.e., base pair deletions relative to the reference genome). The purple vertical bars and the numbers above them indicate insertions and their length. The bottom SAS-SE track and dashed box highlight the location of structural errors detected by SAS pipeline.

Figure S2. Contact map of pore-c reads aligned to the T2T-YAO v2.0.

Figure S3. Comparison of centromere infrastructure of T2T-YAO v1.1 and v2.0.

(**A**) Comparison of centromere structures in T2T-YAO v1.1 and v2.0 via Genomic Centromere Profiling (GCP). (**B**) Patterns of higher-order repeats (HOR) and satellite sequences in the centromere of paternal chromosome 1 in T2T-YAO v1.1 and v2.0.

Figure S4. Repetitive elements in the SE (structural error) region on maternal chromosome 7 with abnormal ONT read depth.

The sequencing depth (blue peak for ONT and pink peak for SPRQ) and read alignment from ONT, SPRQ platforms are displayed across the SE region in maternal chromosome 7. The SAS-SE track

highlight the location of abnormal ONT read depth (dark red) detected by SAS pipeline. The bottom tracks highlight the genomic features, including repeat motif and genes in this region.

Figure S5. Distribution of three major types of homopolymers in each chromosome of T2T-YAO v2.0.

Figure S6. Synteny of the unique genes between T2T-CHM13 and T2T-YAO haplotype.

Synteny maps of (A) Maternal haplotype (MAT) and (B) paternal haplotype (PAT). Each dot represents a gene shared between both genomes in the panel, colored according to gene transcript sequence identity. The continuous diagonal line in panels indicates strong conservation of gene order. Sparse off-diagonal dots suggest possible local rearrangements or transpositions, typically accompanied by a slight reduction in sequence identity.

Figure S7. **Relative locations of extra gene copies in T2T-YAO v2.0 to their corresponding genes in T2T-CHM13.**

Each line connects the position of an extra-copy gene in Maternal haplotype (MAT) **(A)** and paternal haplotype (PAT) **(B)** of T2T-YAO to its corresponding gene position in T2T-CHM13, colored with gene types.

Table S1. Sequencing datasets and phasing statistics for the evaluation of the T2T-YAO assembly.

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 the remaining erroneous 31-mers in T2T-YAO v2.0.

Table S5. The copy number and location of each gene in T2T-CHM13, maternal and paternal haplotype of T2T-YAO v2.0.

Table S6. Transcript variants and sequence identity between T2T-YAO v2.0 and T2T-CHM13

Table S7. Specific HOR units in T2T-YAO v2.0

Table S8. SNP and small indel in repetitive regions.

Supplemental Information file 1: S1_Figures 1-45. Genomic distribution of assembly errors on each chromosome except paternal chromosome 1 before and after polishing

Supplemental Information file 2: S2_Figures 1-24. The composition of centromeric α-satellite HOR units in each chromosome.

TABLES AND TEXT BOXES

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-binary	**QV-21mer**	**66.48**	**68.68**	**67.42**	**74.84**	**78.23**	**76.18**
	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- trinary	**QV-21mer**	**67.7888**	**69.854**	**68.6823**	**inf**	**inf**	**inf**
	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

Table 2. Gene annotation of T2T-YAO v2.0 liftoff from T2T-CHM13

	Gene Type	CHM13	Unique	Success %	Extra copy	Failures	Failure %	Total	MANE Unique	MANE Extra
Maternal	Protein coding	19987	19842	99.27%	182	145	0.73%	20024	11	4
	LncRNA	15511	15291	98.58%	625	220	1.42%	15916	--	--
	Pseudogene	16599	16197	97.58%	674	402	2.42%	16871	--	--
	Others	3899	3806	97.61%	302	93	2.39%	4108	--	--
	Total	55996	55136	98.46%	1783	860	1.54%	56919	56930	56934
Paternal	Protein coding	19238	19087	99.22%	213	151	0.78%	19300	9	20
	LncRNA	16072	15773	98.14%	1258	299	1.86%	17031	--	--
	Pseudogene	16059	15651	97.46%	649	408	2.54%	16300	--	--
	Others	3733	3675	98.45%	240	58	1.55%	3915	--	--
	Total	55102	54186	98.34%	2360	916	1.66%	56546	56555	56575
Diploid	Protein coding	20090	19975	99.43%	--	115	0.57%	39324	14	30
	LncRNA	16388	16116	98.34%	--	272	1.66%	32947	--	--
	Pseudogene	16993	16654	98.01%	--	339	1.99%	33171	--	--
	Others	3906	3855	98.69%	--	51	1.31%	8023	--	--
	Total	57377	56600	98.65%	--	777	1.35%	113465	113479	113509

Box 1. Near-Perfect Genome

Thanks to advances in sequencing and computing technologies, human genomes can now be assembled almost exactly as they exist in nature. Apart from the challenging ribosomal DNA (rDNA) arrays—long stretches of nearly identical ~45 kb repeats—and the exact lengths of a few homopolymer runs, every other basepaire of the diploid human genome T2T-YAO can now be resolved from telomere to telomere. This achievement marks a major step beyond what was possible only a few years ago with the advent of T2T-CHM13.

Using diploid human genome as an example, we can see how close the current technologies bring us to an error-free or perfect eukaryotic genome:

1. **K-mer completeness**: >99.8%
2. **K-mer switch error rate** (diploid genome): <0.01%

In addition, the following criteria are met with rDNA excluded:

3. **Coverage by non-split long reads** (N50 >100 kb): 100%

4. **Number of 30 kb windows not spanned by long reads**: 0

5. **Non-homopolymer sequences lacking sufficient read support**: 0

6. **Homopolymer sites lacking sufficient read support**: <0.02% of total homopolymer sites

Together, these metrics define the current benchmark for what can be considered a *near-perfect* eukaryotic genome.

REFERENCES

Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D.J., Shafin, K., Shumate, A., Xiao, C.*, et al.* (2022). A complete reference genome improves analysis of human genetic variation. Science (New York, NY) *376*, eabl3533.

Altemose, N., Logsdon, G.A., Bzikadze, A.V., Sidhwani, P., Langley, S.A., Caldas, G.V., Hoyt, S.J., Uralsky, L., Ryabov, F.D., Shew, C.J.*, et al.* (2022). Complete genomic and epigenetic maps of human centromeres. Science (New York, NY) *376*, eabl4178.

Antipov, D., Rautiainen, M., Nurk, S., Walenz, B.P., Solar, S.J., Phillippy, A.M., and Koren, S. (2025). Verkko2 integrates proximity-ligation data with long-read De Bruijn graphs for efficient telomere-to-telomere genome assembly, phasing, and scaffolding. Genome Res *35*, 1583–1594.

Arslan, S., Garcia, F.J., Guo, M., Kellinger, M.W., Kruglyak, S., LeVieux, J.A., Mah, A.H., Wang, H., Zhao, J., Zhou, C.*, et al.* (2024). Sequencing by avidity enables high accuracy with low reagent consumption. Nature Biotechnology *42*, 132–138.

Belancio, V.P., Deininger, P.L., and Roy-Engel, A.M. (2009). LINE dancing in the human genome: transposable elements and disease. Genome Medicine *1*, 97.

Chao, K.H., Heinz, J.M., Hoh, C., Mao, A., Shumate, A., Pertea, M., and Salzberg, S.L. (2025). Combining DNA and protein alignments to improve genome annotation with LiftOn. Genome Res *35*, 311–325.

Chen, Q., Yang, C., Zhang, G., and Wu, D. (2024). GCI: a continuity inspector for complete genome assembly. bioRxiv, 2024.2004.2006.588431.

Church, D.M. (2022). A next-generation human genome sequence. Science (New York, NY) *376*, 34–35.

Corda, L., and Giunta, S. (2025). Chromosome-specific centromeric patterns define the centeny map of the human genome. Science (New York, NY) *389*, eads3484.

Deshpande, A.S., Ulahannan, N., Pendleton, M., Dai, X., Ly, L., Behr, J.M., Schwenk, S., Liao, W., Augello, M.A., Tyer, C.*, et al.* (2022). Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nat Biotechnol *40*, 1488–1499.

Formenti, G., and Howe, K. (2022). An assembly line for an improved human reference genome. Nature *611*, 519–+.

Freeman, T.M., Wang, D., and Harris, J. (2020). Genomic loci susceptible to systematic sequencing bias in clinical whole genomes. Genome Res *30*, 415–426.

Guo, S., Huang, Z., Zhang, Y., He, Y., Chen, X., Wang, W., Li, L., Kang, Y., Gao, Z., Yu, J.*, et al.* (2024). Enhancing Variant Calling in Whole-exome Sequencing Data Using Population-matched Reference Genomes. Genomics Proteomics Bioinformatics *22*.

Hansen, N., and al, e. (2025). A complete diploid human genome benchmark for personalized genomics. bioRxiv (to appear).

He, Y., Chu, Y., Guo, S., Hu, J., Li, R., Zheng, Y., Ma, X., Du, Z., Zhao, L., Yu, W.*, et al.* (2023). T2T-YAO: A Telomere-to-telomere Assembled Diploid Reference Genome for Han Chinese. Genomics Proteomics

Bioinformatics *21*, 1085–1100.

Jarvis, E.D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., Tracey, A., Thibaud-Nissen, F., Vollger, M.R., Porubsky, D.*, et al.* (2022). Semi-automated assembly of high-quality diploid human reference genomes. Nature *611*, 519–531.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P.*, et al.* (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Lemmers, R., van der Vliet, P.J., Vreijling, J.P., Henderson, D., van der Stoep, N., Voermans, N., van Engelen, B., Baas, F., Sacconi, S., Tawil, R.*, et al.* (2018). Cis D4Z4 repeat duplications associated with facioscapulohumeral muscular dystrophy type 2. Hum Mol Genet *27*, 3488–3497.

Li, H. (2023). Protein-to-genome alignment with miniprot. Bioinformatics (Oxford, England) *39*.

Li, H., and Durbin, R. (2024). Genome assembly in the telomere-to-telomere era. Nat Rev Genet *25*, 658–670.

Li, K., Xu, P., Wang, J., Yi, X., and Jiao, Y. (2023). Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. Nature Communications *14*, 6556.

Liao, W.W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J.*, et al.* (2023). A draft human pangenome reference. Nature *617*, 312–324.

Logsdon, G.A., Rozanski, A.N., Ryabov, F., Potapova, T., Shepelev, V.A., Catacchio, C.R., Porubsky, D., Mao, Y., Yoo, D., Rautiainen, M.*, et al.* (2024). The variation and evolution of complete human centromeres. Nature *629*, 136–145.

Lyčka, M., Bubeník, M., Závodník, M., Peska, V., Fajkus, P., Demko, M., Fajkus, J., and Fojtová, M. (2023). TeloBase: a community-curated database of telomere sequences across the tree of life. Nucleic Acids Research *52*, D311–D321.

Mastoras, M., Asri, M., Brambrink, L., Hebbar, P., Kolesnikov, A., Cook, D.E., Nattestad, M., Lucas, J., Won, T.S., Chang, P.-C.*, et al.* (2024a). Highly accurate assembly polishing with DeepPolisher. bioRxiv, 2024.2009.2017.613505.

Mastoras, M., Asri, M., Brambrink, L., Hebbar, P., Kolesnikov, A., Cook, D.E., Nattestad, M., Lucas, J., Won, T.S., Chang, P.C.*, et al.* (2024b). Highly accurate assembly polishing with DeepPolisher. bioRxiv.

Mc Cartney, A.M., Shafin, K., Alonge, M., Bzikadze, A.V., Formenti, G., Fungtammasan, A., Howe, K., Jain, C., Koren, S., Logsdon, G.A.*, et al.* (2022). Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. Nature methods *19*, 687–695.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome biology *17*, 122.

Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A.*, et al.* (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature *585*, 79–84.

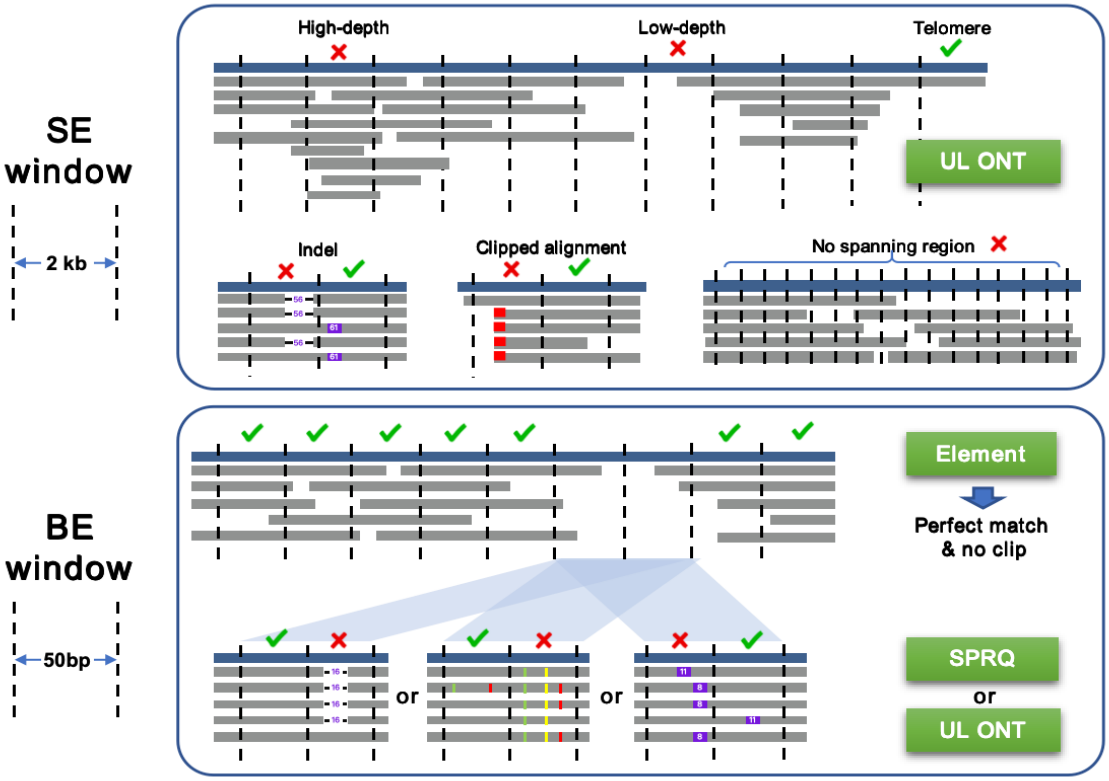
Morales, J., Pujar, S., Loveland, J.E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C.,

897 Ermolaeva, O., Farrell, C.M., *et al.* (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics
898 and research. Nature *604*, 310–315.
899 Ni, Y., Liu, X., Simeneh, Z.M., Yang, M., and Li, R. (2023). Benchmarking of Nanopore R10.4 and R9.4.1
900 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Comput
901 Struct Biotechnol J *21*, 2352–2364.
902 Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N.,
903 Uralsky, L., Gershman, A., *et al.* (2022). The complete sequence of a human genome. Science (New
904 York, NY) *376*, 44–53.
905 Olson, N.D., Wagner, J., Dwarshuis, N., Miga, K.H., Sedlazeck, F.J., Salit, M., and Zook, J.M. (2023).
906 Variant calling and benchmarking in an era of complete human genome sequences. Nature Reviews
907 Genetics *24*, 464–483.
908 Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J.,
909 Nguyen, N., Afshar, P.T., *et al.* (2018a). A universal SNP and small-indel variant caller using deep neural
910 networks. Nature Biotechnology *36*, 983–987.
911 Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling,
912 D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., *et al.* (2018b). Scaling accurate genetic variant
913 discovery to tens of thousands of samples. bioRxiv, 201178.
914 Rautiainen, M., Nurk, S., Walenz, B.P., Logsdon, G.A., Porubsky, D., Rhie, A., Eichler, E.E., Phillippy,
915 A.M., and Koren, S. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nature
916 Biotechnology.
917 Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W.,
918 Fungtammasan, A., Kim, J., *et al.* (2021). Towards complete and error-free genome assemblies of all
919 vertebrate species. Nature *592*, 737–746.
920 Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality,
921 completeness, and phasing assessment for genome assemblies. Genome biology *21*, 245.
922 Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B.
923 (2013). Characterizing and measuring bias in sequence data. Genome biology *14*, R51.
924 Shafin, K., Pesout, T., Chang, P.C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M.,
925 Eizenga, J.M., Miga, K.H., *et al.* (2021). Haplotype-aware variant calling with PEPPER-Margin-
926 DeepVariant enables high accuracy in nanopore long-reads. Nature methods *18*, 1322–1332.
927 Shumate, A., and Salzberg, S.L. (2021). Liftoff: accurate mapping of gene annotations. Bioinformatics
928 (Oxford, England) *37*, 1639–1643.
929 Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E.,
930 Gandhi, M., Hong, K., Pehlivan, D., *et al.* (2024). Detection of mosaic and population-level structural
931 variants with Sniffles2. Nature Biotechnology *42*, 1571–1580.
932 Stephens, Z., and Kocher, J.-P. (2024). Characterization of telomere variant repeats using long reads
933 enables allele-specific telomere length estimation. BMC Bioinformatics *25*, 194.

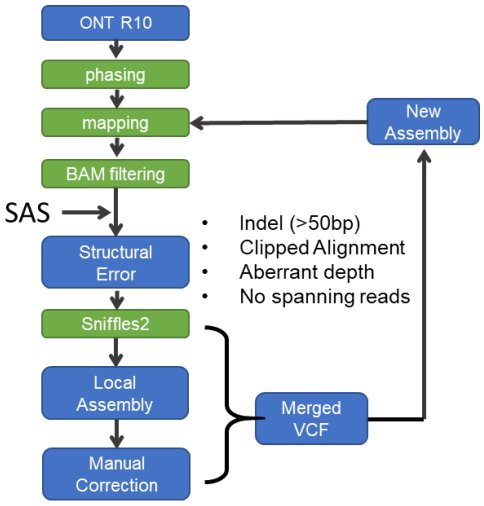
934 Telomere-to-telomere-consortium (2023). HG002 Q100 project. In https://githubcom/marbl/hg002.
935 Vollger, M.R., Dishuck, P.C., Sorensen, M., Welch, A.E., Dang, V., Dougherty, M.L., Graves-Lindsay, T.A.,
936 Wilson, R.K., Chaisson, M.J.P., and Eichler, E.E. (2019). Long-read sequence and assembly of
937 segmental duplications. Nature methods *16*, 88–94.
938 Wagner, J., Olson, N.D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y.C., Gupta, R.,
939 Wenger, A.M., Rowell, W.J.*, et al.* (2022). Curated variation benchmarks for challenging medically
940 relevant autosomal genes. Nat Biotechnol *40*, 672–680.
941
942

Figure 1

A


SE window
2 kb
High-depth
Low-depth
Telomere
UL ONT
Indel
Clipped alignment
No spanning region
BE window
50bp
Element
Perfect match & no clip
or
or
SPRQ
or
UL ONT

B


ONT R10
phasing
mapping
BAM filtering
SAS
Structural Error
Sniffles2
Local Assembly
Manual Correction
New Assembly
Merged VCF
Indel (>50bp)
Clipped Alignment
Aberrant depth
No spanning reads

C

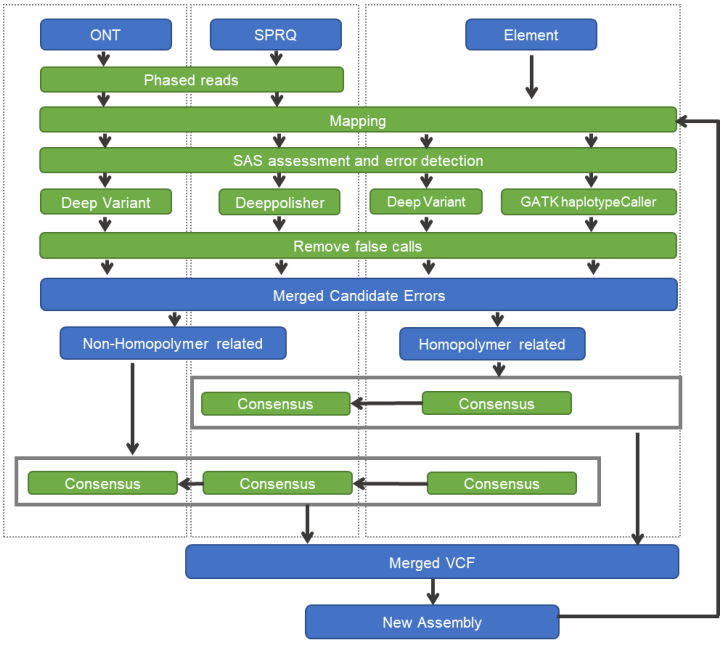
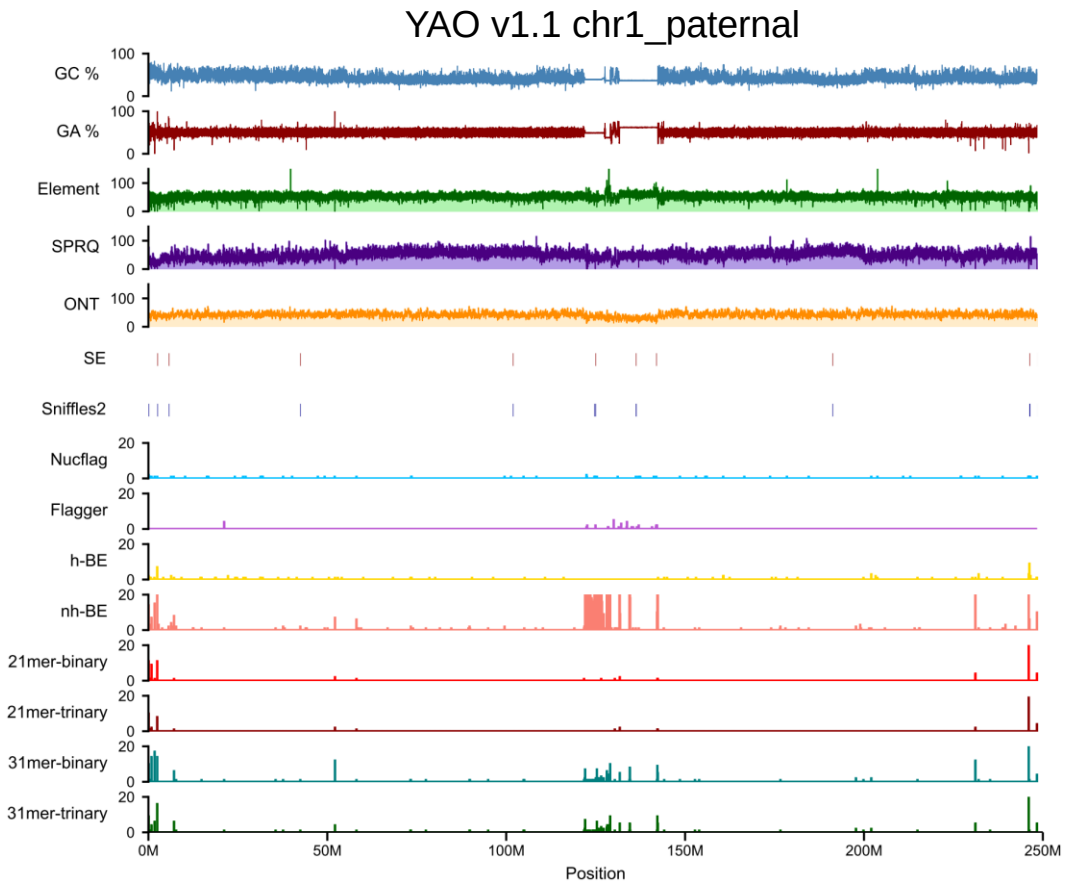

ONT
SPRQ
Element
Phased reads
Mapping
SAS assessment and error detection
Deep Variant
Deeppolisher
Deep Variant
GATK haplotypeCaller
Remove false calls
Merged Candidate Errors
Non-Homopolymer related
Homopolymer related
Consensus
Consensus
Consensus
Consensus
Consensus
Merged VCF
New Assembly

Figure 2

A


YAO v1.1 chr1_paternal
GC %
GA %
Element
SPRQ
ONT
SE
Sniffles2
Nucflag
Flagger
h-BE
nh-BE
21mer-binary
21mer-trinary
31mer-binary
31mer-trinary
100
0
20
0
0M
50M
100M
150M
200M
250M
Position

B

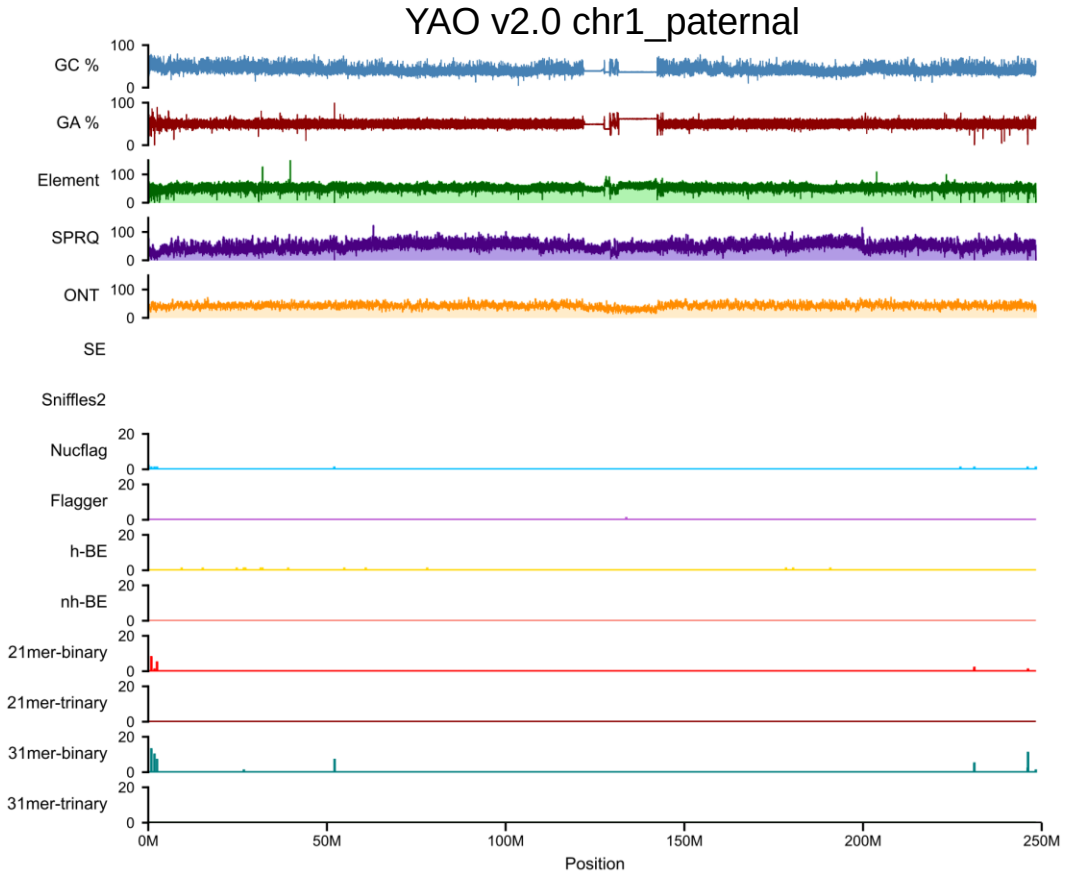

YAO v2.0 chr1_paternal
GC %
GA %
Element
SPRQ
ONT
SE
Sniffles2
Nucflag
Flagger
h-BE
nh-BE
21mer-binary
21mer-trinary
31mer-binary
31mer-trinary
100
0
20
0
0M
50M
100M
150M
200M
250M
Position

Figure 3

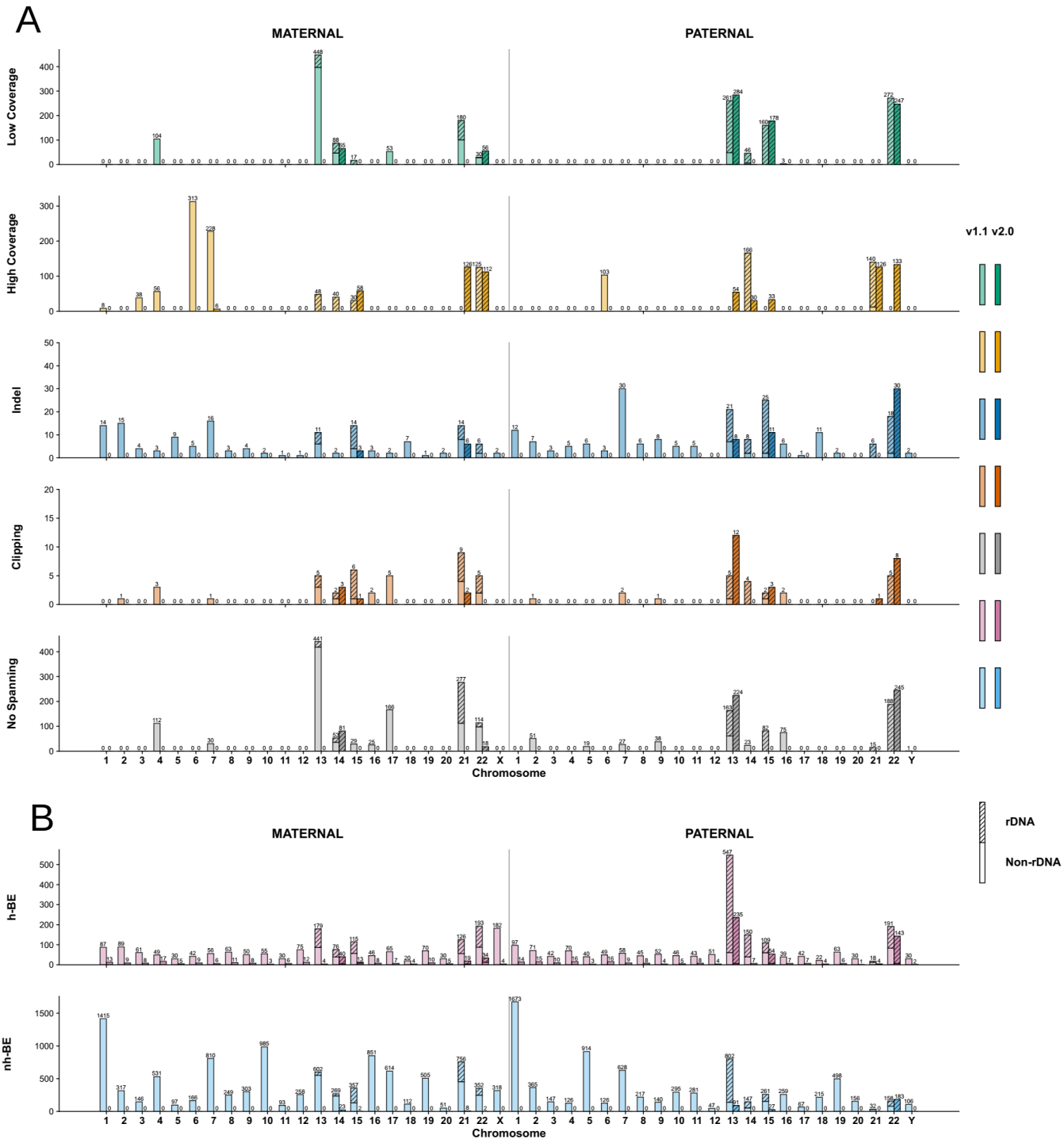
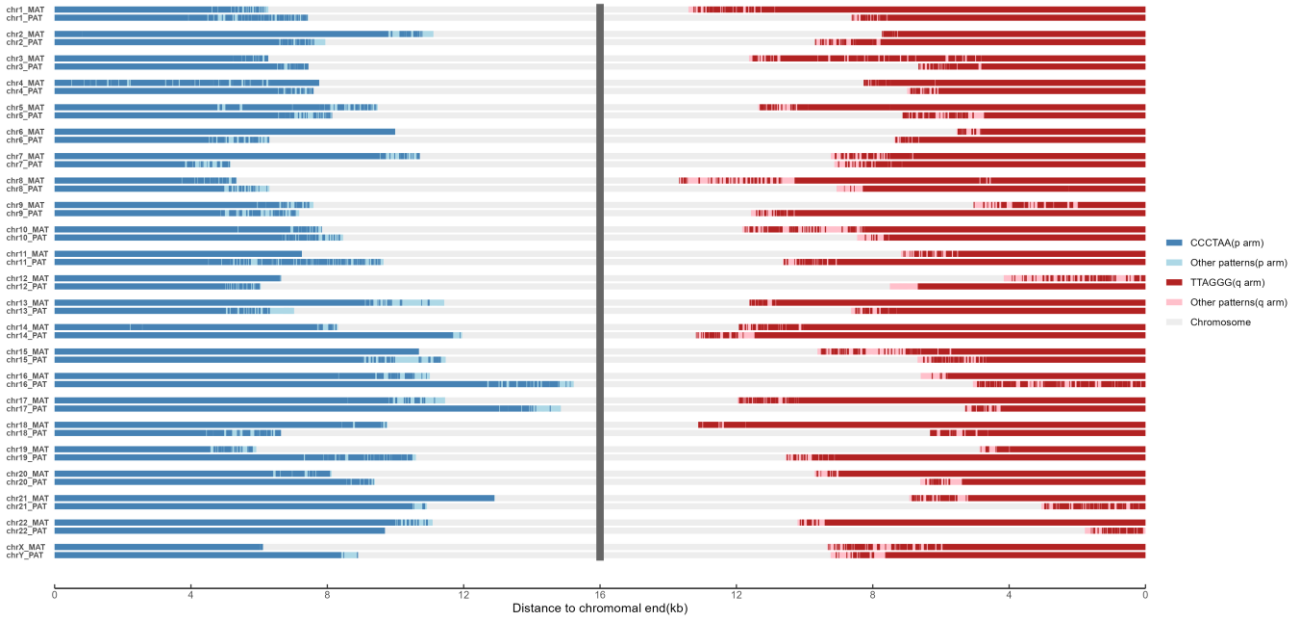

A
MATERNAL
PATERNAL
Low Coverage
High Coverage
Indel
Clipping
No Spanning
v1.1 v2.0
Chromosome
B
MATERNAL
PATERNAL
h-BE
nh-BE
rDNA
Non-rDNA
Chromosome

Figure 4

A

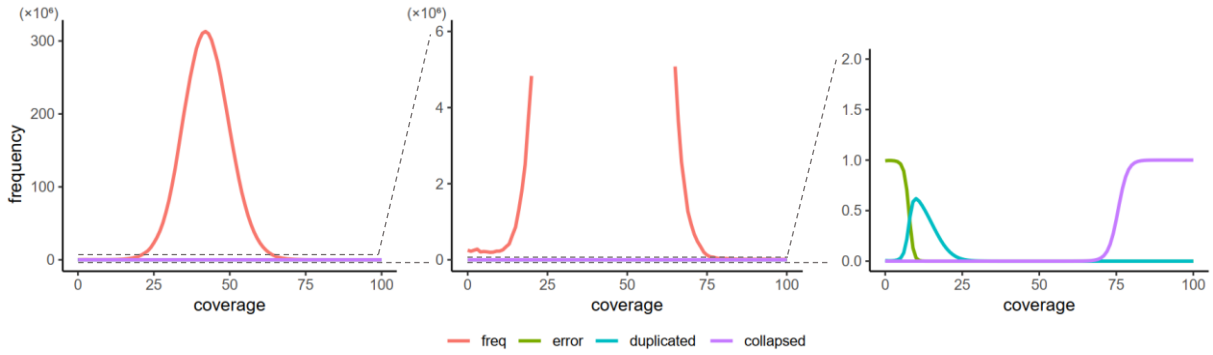

CCCTAA(p arm)
Other patterns(p arm)
TTAGGG(q arm)
Other patterns(q arm)
Chromosome
Distance to chromomal end(kb)

B

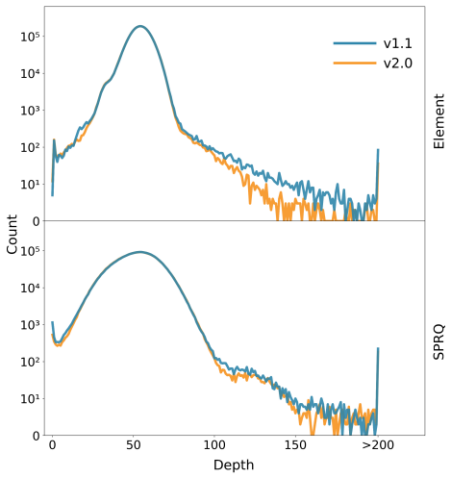
Chr1_MAT
Chr1_PAT
Chr2_MAT
Chr2_PAT
Chr3_MAT
Chr3_PAT
Chr4_MAT
Chr4_PAT
Chr5_MAT
Chr5_PAT
Chr6_MAT
Chr6_PAT
Chr7_MAT
Chr7_PAT
Chr8_MAT
Chr8_PAT
Chr9_MAT
Chr9_PAT
Chr10_MAT
Chr10_PAT
Chr11_MAT
Chr11_PAT
Chr12_MAT
Chr12_PAT
Chr13_MAT
Chr13_PAT
Chr14_MAT
Chr14_PAT
Chr15_MAT
Chr15_PAT
Chr16_MAT
Chr16_PAT
Chr17_MAT
Chr17_PAT
Chr18_MAT
Chr18_PAT
Chr19_MAT
Chr19_PAT
Chr20_MAT
Chr20_PAT
Chr21_MAT
Chr21_PAT
Chr22_MAT
Chr22_PAT
ChrX_MAT
ChrY_PAT
0
5Mb
Repeat Family
αSat-HOR-active
αSat-HOR-div
αSat-HOR-inactive
αSat-HOR-SF
β-Sat
gsat
HSat1
HSat2
HSat3
otherRepeat
rDNA
transitionRegion
0
2Mb

Figure 5

A


(×10⁶)
frequency
coverage
freq
error
duplicated
collapsed

B


v1.1
v2.0
Element
SPRQ
Count
Depth

C

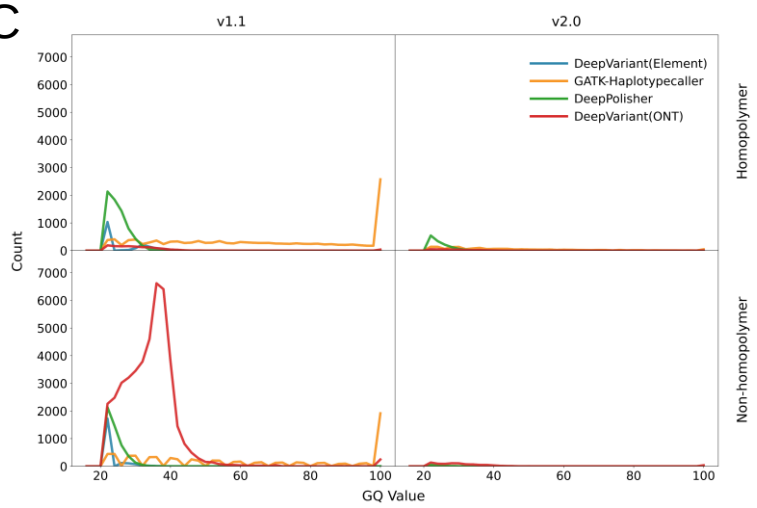
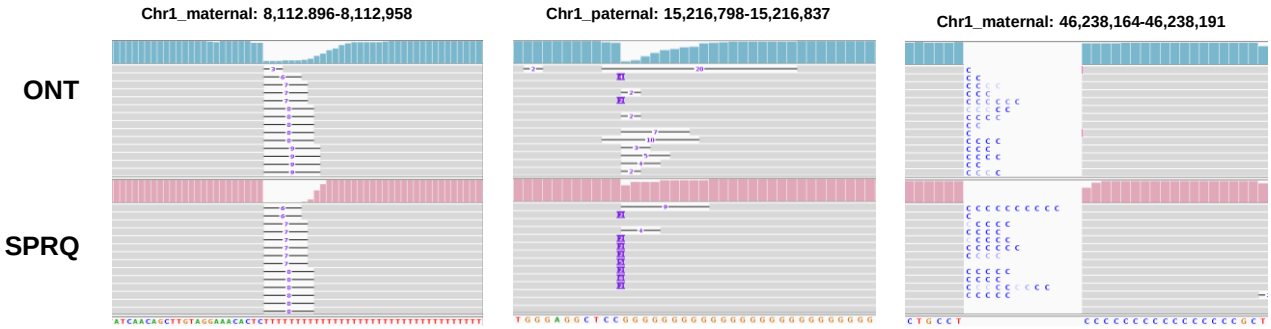

v1.1
v2.0
DeepVariant(Element)
GATK-Haplotypecaller
DeepPolisher
DeepVariant(ONT)
Homopolymer
Non-homopolymer
Count
GQ Value

Figure 6

A


Chr1_maternal: 8,112.896-8,112,958
Chr1_paternal: 15,216,798-15,216,837
Chr1_maternal: 46,238,164-46,238,191
ONT
SPRQ

B

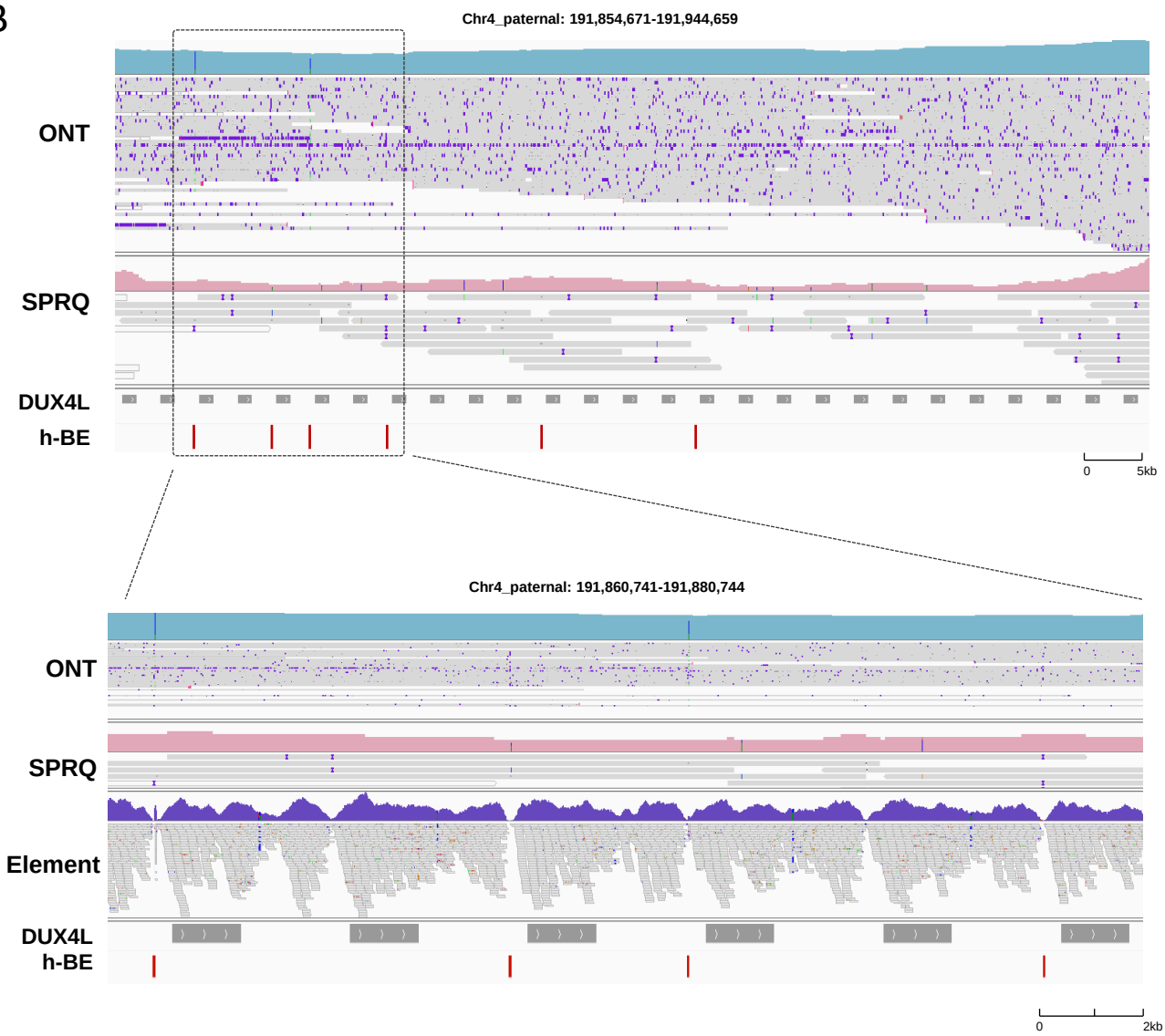
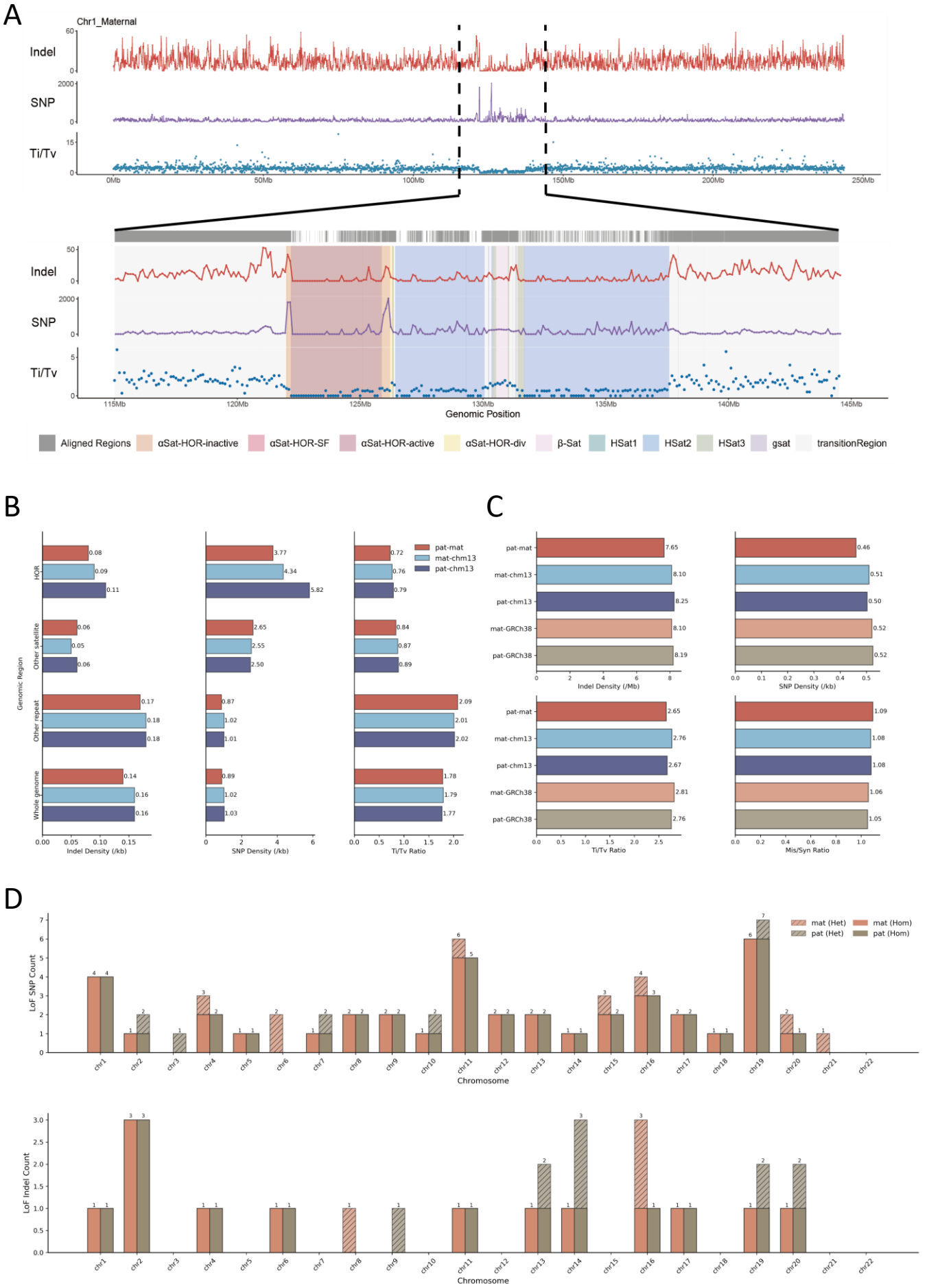

Chr4_paternal: 191,854,671-191,944,659
ONT
SPRQ
DUX4L
h-BE
0
5kb
Chr4_paternal: 191,860,741-191,880,744
ONT
SPRQ
Element
DUX4L
h-BE
0
2kb

Figure 7


A
Chr1_Maternal
Indel
SNP
Ti/Tv
0Mb
50Mb
100Mb
150Mb
200Mb
250Mb
115Mb
120Mb
125Mb
130Mb
135Mb
140Mb
145Mb
Genomic Position
Aligned Regions
αSat-HOR-inactive
αSat-HOR-SF
αSat-HOR-active
αSat-HOR-div
β-Sat
HSat1
HSat2
HSat3
gsat
transitionRegion
B
Genomic Region
HOR
Other satellite
Other repeat
Whole genome
Indel Density (/kb)
SNP Density (/kb)
Ti/Tv Ratio
pat-mat
mat-chm13
pat-chm13
C
pat-mat
mat-chm13
pat-chm13
mat-GRCh38
pat-GRCh38
Indel Density (/Mb)
SNP Density (/kb)
Ti/Tv Ratio
Mis/Syn Ratio
D
mat (Het)
pat (Het)
mat (Hom)
pat (Hom)
LoF SNP Count
LoF Indel Count
Chromosome