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 challenging. We report T2T-YAO v2.0, a telomere-to-telomere complete assembly of a Han Chinese individual, polished to near-perfect base-level and structural accuracy. To systematically identify assembly errors, we developed Sufficient Alignment Support (SAS), an automatic method that flags structural and base-level errors in windows lacking sufficient read support. Building on this, we established a "structural–error–first" polishing strategy, correcting misassemblies using ultra-long ONT reads, followed by base-level refinement with PWC (Platform-integrated Window Consensus). Using these approaches, we resolved all detectable

structure and non-homopolymer-related errors outside ribosomal DNA (rDNA) regions. The resulting assembly contains no unsupported 21-mers across sequencing platforms, meeting k-mer–based criteria for an error-free genome. T2T-YAO v2.0 delivers the most perfect East Asian reference to date, with limited issues confined to rDNA arrays and homopolymer tracks, enabling precise genome annotation, benchmarking, and variant discovery—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, reference genome, 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[1,2], which achieved unprecedented 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[3,4]. 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[5-8], 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 or insufficient support. In practice, multiple structural and single-nucleotide variant (SNV) callers are often applied in parallel to identify candidate errors based on alternative alleles inferred from local reads evidence. The accuracy of the polished assembly is then 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-and-evaluation strategy remains insufficiently efficient or accurate to achieve a truly error-free human genome by now. Two key components are required to break this limitation: First, a comprehensive and reliable method is needed to identify assembly errors with sufficient sensitivity and specificity to effectively guide correction. K-mer–based approaches are intrinsically limited for high-quality assemblies, as they tend to miss genuine errors that do not introduce novel k-mers, or obscured by sequencing errors, making quality values sensitive to sequencing depth and platform-specific error profiles. Second, a streamlined polishing pipeline is required. Existing SNV callers, such as Deeppolisher [10] and DeepVariant [11,12], are typically optimized for single-platform datasets or haploid references, 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 automatic assembly error detector called Sufficient Alignment Support (SAS), which flag regions of the assembly lacking adequate read support, the 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[13], 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 detectable

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 automatic assembly error detector

To systematically identify assembly errors, we developed Sufficient Alignment Support (SAS), an automated framework that flags genomic windows lacking sufficient support from mapped reads. SAS is motivated by the single-molecule, PCR-free nature of current sequencing technologies, in which each read that perfectly matches the assembly provides independent evidence that the corresponding sequence originates from the biological sample. Genomic windows supported by sufficient independent reads are therefore considered reliable. Because sequencing platforms differ in their error profiles and then efficiencies for detecting base-level errors (BEs; <50 bp) and structural errors (SEs; ≥50 bp), SAS evaluate these two error classes, based on their typical cutoff[14], independently using appropriate sequencing data (**Figure 1A**).

For BE detection, the diploid assembly is partitioned into 50 bp, non-overlapping windows. A window is flagged as erroneous if it is not supported by enough reads without mismatches. Because each read constitutes independent evidence, SAS requires a minimal number of fully matching reads, which can be user-defined. By default, this threshold is set to > 10% of the mean sequencing depth for any platform. In regions with systematically reduced coverage due to sequencing bias, it can be relaxed to >10% of the local depth, specifically in: (1) regions within 10 kb of chromosomal termini (telomeres); (2) sequences of degenerative GnA or CnT repeats (GA/CT content ≥90% and GC content ≥60%); and (3) GC-rich regions (GC content ≥90%). These contexts are known to be underrepresented in polymerase-dependent, high-accuracy sequencing platforms

(*e.g.*, Element, Illumina, and PacBio), and therefore rely predominantly on ONT reads for coverage which are more error-prone and provide fewer fully matching reads.

SEs are assessed using larger windows (2 kb by default) and exclusively ultra-long ONT reads, leveraging their extended read-length. A window is considered free of SEs if it satisfies all the following criteria: (1) The sequencing depth falls within the acceptable range. The upper and lower outlier cutoffs were calculated using the adjusted boxplot method (based on MedCouple), while the abnormal depth thresholds correspond to approximately 20–200% of the mean depth (**Figure S1A**); and (2) a simple majority ($\geqslant$ 50%) of local reads span the entire window without indels ≥50 bp indels. In chromosomal termini regions (within 10 kb), where coverage gradients are expected, the requirement is relaxed to ≥10% of the local-depth. To reduce false signals from misalignment, reads exhibiting excessive fragmentation (by default, >4 split in ultra-long reads) are excluded from analysis. To improve detection of misjoin events in larger genomic scales, SAS additionally evaluating sliding SE windows of increased size. If no reads span a given sliding window, the corresponding consecutive SE windows are flagged as no-span SEs. We empirically set sliding window size to 30 kb, consistent with the filter we used for ONT reads, such that every authentic genomic segment of this size should, in principle, be fully spanned by at least one read.

Beyond local error detection, SAS provides a unified metric for global assembly quality. An overall SAS-based quality value (QV_{SAS}) can be calculated by summing non-redundant BE and SE windows then computing the predicted error rate (E) as:

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

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

Evaluation of T2T-YAO v1.1 with SAS and complementary metrics

To improve the quality of T2T-YAO assembly, we upgraded its raw data to multiple high-accuracy sequencing platforms: 260 Gb Q28 ultra-long ONT reads[15] (filtered to >30 kb; N50 >120 kb), 172 Gb Pore-C[16], 306 Gb HiFi reads (SPRQ chemistry; filtered to >3 kb), and 320 Gb Element AVITI[17] Q50 reads (**Table S1**). All sequencing data were generated from DNA extracted from an induced pluripotent stem cell (iPSC) line (<10 passages) to minimize cellular heterogeneity and preserve germline configurations of intact immunoglobin loci. ONT and SPRQ reads were phased in advance using haplotype-specific k-mers (hapmers) and mapped in a haplotype-aware manner (Methods), reducing ambiguity in haplotype assignment.

We applied SAS to evaluate the previously released T2T-YAO v1.1 assembly and compared its performance with Merqury, the most widely adopted k-mer–based assembly quality metric (**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} values ranged from 45.36 to 61.37 (**Table S2**). In contrast, Merqury —using the Element-SPRQ hybrid k-mer set following the T2T-CHM13 protocol[3]—reported only 1,074 estimated errors (Q67.42) based on 21-mers and 3,970 errors (Q61.74) based on 31-mers across the diploid genome (**Table 1**). Although the use of more high-accuracy Element and SPRQ reads uncovered more errors than originally reported[13], the substantially larger number of errors detected by SAS highlights its markedly greater sensitivity. Many BE windows detected by SAS due to insufficient reads support were not detceted by Merqury because they do not introduce novel k-mers. Closer inspection revealed that a substantial fraction of BE windows were association with homopolymer tracts, where reads often disagreed on repeat length. We therefore defined homopolymer sites as stretches of ≥10 A/T, ≥7 G/C mononucleotide repeats, or ≥10 dinucleotide repeats, where SPRQ/ONT reads exhibit elevated variability (measured by median absolute deviation, MAD, **Figure S1B**). 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**).

Conversely, some erroneous k-mers reported by Merqury occurred outside SAS-flagged BE windows and were enriched in degenerative GnA or CnT repeats. These loci were not flagged by SAS when adequately supported by ONT reads but lacked SPRQ and Element coverage due to platform-specific sequencing bias. These discrepancies illustrate fundamental limitations of k-mer–based assessment: their inability to detect errors that do not generate novel k-mers (e.g., homopolymer length discrepancies), sensitivity to platform-specific error profile and sequencing depth, and susceptibility to false positives in regions with biased or absent coverage. These issues arise from Merqury's reliance on hybrid k-mer sets derived exclusively from high-accuracy Illumina (or Element) and PacBio data, while excluding ONT reads to avoid inflation by sequencing-error-derived k-mers. As a representative example, we highlight paternal chromosome 1 to illustrate the constrasting error profile detected by SAS and other metrics (**Figure 2A**).

Given Merqury's limited ability to detect structural errors, we further compared SAS with additional tools, including the ONT-based structural variant (SV) caller Sniffles2[18] and two coverage-based structural error detectors—Flagger[19] (ONT) and nucflag[20] (SPRQ). Sniffles2 identified 263 confident SV-like disruptions (supported by ≥60% of local reads), in contrast to the 307 events associated to 2,247 clip-, indel-, or no-span–SE windows detected by SAS. Flagger and nucflag showed substrantial inconsistency, identifying 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—representing higher confidence calls—were either within or adjacent (<10 kb) to SAS-identified SE windows. Furthermore, 78.50% of low-coverage regions jointly flagged by Flagger ("erroneous" and "duplication" calls) and nucflag ("misjoin" calls) were also detected by SAS. In constrast, nucflag frequently misclassifies GnA/CnT repeat regions and loci affected by SPRQ dropout due to sequencing bias as "misjoins" (**Figure 2A**), undermining the reliability of SE detection based on PacBio coverage. Haplotypic switch errors, reported as heterogeneous regions by nucflag and as switch k-mers by Merqury, are not treated as a distinct error class in SAS. Instead, depending on the supporting read evidence, they are

classified either as SEs when flanking clipping or spanning disruptions are present, or as BEs when phased reads disagree 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 in sliding 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 reduces 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**). Telomeric and subtelomeric regions are also challenging due to high sequence similarity among chromosomes and haplotypes, for which ONT reads provide the only reliable support. In cases where heterogeneity among ONT reads impeded successful reassembly, stratifying ONT reads by 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, approaching conventionally accepted telomere lengths and substantially exceeding those in T2T-CHM13 (2.7 ± 0.7 kb; **Figure 4A**).

Ultimately, T2T-YAO v2.0 becomes free of detectable SEs outside of ribosomal DNA (rDNA) clusters (**Figure 2B, 3A**). Centromeric collapses, particularly within the higher-order repeat (HOR) arrays of the 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, composed of L1M and Alu elements[21], which shows normal SPRQ depth and no evidence of clipping or misjoins (**Figure S4**). Aside from this, ONT read depth across the genome falls within 20–200% of the mean and follows a near-Gaussian distribution, with only 2.81 Mb (0.047%) and 1.55 Mb (0.026%) tallied by Flagger as false duplications and collapses, respectively (**Figure 5A**), ~90% of which are confined to rDNA clusters. All regions jointly identified by Flagger and nucflag are restricted to rDNA arrays or telomeric regions, validating the structural integrity of T2T-YAO v2.0 (**Table S3**). To further assess continuity, we quantified the fraction of the assembly covered by non-split ONT reads (NScov%), defined as reads aligning continuously without splits and thus providing direct evidence that the full sequence is present in the sample. In T2T-YAO v1.1, NScov% is 99.98%, leaving ~0.91 Mb (excluding rDNA) left uncovered. In contrast, T2T-YAO v2.0 accomplishes 100% NScov% outside rDNA, confirming the absence of unresolved SEs beyond rDNA clusters.

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 SAS together with 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**). Then we developed PWC pipeline to correct each flagged BE window. The principle of PWC is to generate a reliable consensus sequence for each flagged window (supported by ≥60% of spanning reads at depth ≥10) independently for each platform and applies a cross-platform prioritization scheme: (1) consensus supported by all platforms is directly applied; (2) for nh-BEs, corrections supported by two platforms are adopted, otherwise following the order Element > SPRQ > ONT; and (3) for h-BEs, ONT-based consensus is excluded due to the known difficulty of ONT sequencing in resolving homopolymer lengths (**Figure S5A**). When no consensus can be obtained for nh-BEs, ONT reads spanning the region are locally reassembled, which often resolves alignment artifacts and enables consensus generation (**Figure 1C**). Read mapping and polishing were performed iteratively until the number of detected BEs stabilized. The stringent consensus criteria of PWC, together with its restriction to SAS-defined BE windows, minimized overcorrection from alignment

artifacts and allowed repeated polishing, culminating in 47 rounds to produce T2T-YAO v2.0.

To identify additional base-level errors potentially missed by SAS, we applied multiple variant callers—Deeppolisher[10] (SPRQ), DeepVariant[11,12] (ONT and Element), and GATK HaplotypeCaller[22] (Element)—by mapping reads to the T2T-YAOv1.1 diploid assembly. These callers reported thousand to tens of thousands of homozygous SNV-like issues with genotype quality (GQ) ≥20, yielding a total of 63,398 non-redundant calls in total. However, these calls showed limited overlap across platforms (**Figure 5C**). Among them, 59.79% were non-homopolymer-associated, of which 43.90% overlapped SAS-detected nh-BEs; while 40.21% were homopolymer-associated, with only 7.08% overlapped SAS h-BEs. This discrepancy highlights the difficulty of identifying true base-level errors using single-platform alignment, particularly in homopolymer contexts, likely reflecting artifacts introduced during homopolymer compression in the initial Verkko-based assembly[23,24].

After iterative PWC 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[25], where the polyC tracts between DUX4-like genes cannot be confidently resolved (**Figure S5B**). Nonetheless, the remaining h-BEs account for only 0.04% of the total 2,412,242 homopolymer sites genome-wide, enabling an accurate census of homopolymer content in a human genome. Specifically, polyA/polyT tracts (≥10 bp) comprise 81.70% of total sites, while polyC/polyG tracts (≥7 bp) and poly-dinucleotid (≥10 repeats) account 9.20% and 9.10%, respectively. Homopolymers are depleted in centromeric regions and exhibit highly concordant distribution across haplotypes. Notably, polyC/polyG and poly-dinucleotide repeats are enriched ~10-fold in rDNA sequences (up to ~600 sites /Mb), whereas polyA/polyT tracts are markedly reduced in the short arms of acrocentric chromosomes containing rDNA (**Figure S6**).

Toward an Error-Free Human Genome Assembly

We re-evaluated the quality of the T2T-YAO v2.0 assembly using both Merqury and SAS metrics based on the same sequencing datasets (**Table 1**). Merqury reported substantially improved QV of 76.18 and 68.03 using 21-mers and 31-mers, respectively, compared with v1.1. The remaining erroneous k-mer were consistently confined to regions with near-zero coverage (depth ≤1) in Element and SPRQ data. In total, 28,956 erroneous 31-mers can be merged into 1,127 discrete regions spanning 76,248 bp, encompassing all erroneous 21-mers. Among these regions, 70% length (totaling 53,307 bp) exhibited >90% GA/CT content, while the remainder consisted of non-canonical telomeric 6-mers or GC-rich tandem repeats—sequences contexts known to be affected by platform-specific sequencing biases[26,27] (**Table S4**). Given the presence of these sequences in human genome that are dropped out by PacBio and short-read sequencing (SRS), achieving Q100 using conventional k-mer-based metrics is unattainable even after complete correction of assembly errors.

To completement these drop-out regions, we incorporated the high-accuracy ONT-derived k-mer set (Q28; >99.8%). After filtering k-mers with occurrence ≤4, the resulting set was smaller than those derived from Element and SPRQ data, effectively excluding

most sequencing-error-related k-mers (**Figure S7**). This expanded trinary k-mer set did not inflate the QV of T2T-YAO v1.1, indicating minimal inclusion of sequencing-error-derived k-mers from ONT data. When applied to T2T-YAO v2.0, the trinary 21-mer QV became “infinite,” satisfying Merqury definition of an error-free diploid genome, while 31-mer QVs reached 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 were confined to rDNA regions, whereas those of maternal localized to two h-BE loci within GnA repeats containing several polyG tracts (≥8 bases).

Consistent with these results, SAS reassessment of T2T-YAO v2.0 identified 1,773 SE windows, all confined to rDNA regions except for the six high-depth windows in maternal chromosome 7 described above, and 1,175 BE windows, comprising 827 within rDNA and 348 h-BEs outside rDNA, precisely matching residual erroneous k-mers detected by the Merqury. The resulting whole-genome QV_{SAS} values were 66.00 (maternal) and 61.25 (paternal). When rDNA regions were masked, leaving only the 348 residual h-BEs, the QV_{SAS} increased to 71.62 and 71.81, respectively (**Table 1, Table S2, Supplemental information file 1**).

Evaluation of SAS and PWC

We evaluated SAS (default parameters) against Sniffles2 and Deeppolisher using a spike-in version of YAO v2.0 containing 4,950 SEs and 20,000 BEs, with error profiles and genomic distributions matching those observed in v1.1, including subsets in centromeric (~4%) and rDNA (~0.17%) regions. Assuming YAO v2.0 to be free of SEs, both SAS and Sniffles2 recovered most spike-in events, with SAS achieving higher accuracy (F1 = 0.987 vs. 0.978). Excluding residual SEs led to only modest improvement, suggesting a low pre-existing SE burden (**Figure. 6A**, **Table S5**). For BE detection, SAS outperformed Deeppolisher (F1 = 0.988 genome-wide; 0.990 outside centromeres/rDNA), whereas Deeppolisher reached a substantially lower peak F1 (0.487) at GQ = 10 and showed strong sensitivity to GQ thresholds. After accounting for residual BEs, SAS performance further improved (F1 = 0.995 genome-wide; 0.997 outside

centromeres/rDNA; **Figure. 6B, Table S5**). SAS false negatives were enriched at loci with ambiguous Element read alignments. Identical regions (⩾150 bp), comprising ~8.9% of the genome, accounted for 94 of 113 missed false negatives. Most false positives outside rDNA after excluding residual BEs were associated with homopolymer-related BEs, with the remainder located in identical regions.

We next evaluated the PWC pipeline. Of 19,887 true-positive sites detected by SAS, 99.66% were corrected to the YAO v2.0 sequence, with only 68 uncorrected sites confined to identical regions (**Table S5**). These results indicate that SAS provides a reliable basis for downstream correction, while homopolymers and identical regions remain limiting contexts for current sequencing technologies. Consistently, PWC achieved a high correction rate with minimal evidence of overcorrection.

Parallel simulations in HG002 v1.1 using identical spike-in errors and comparable sequencing depth showed similarly strong performance of SAS, but with increased false-positive rates and reduced precision and F1 scores relative to YAO. Accounting for residual errors in HG002 v1.1 improved precision from 94.56% to 97.73% for SEs and from 94.78% to 99.09% for BEs (**Figure. 6C, D**), indicating a larger contribution of pre-existing errors in HG002 than in YAO. Applying SAS to recent HG002 versions and comparing against curated corrections (v0.9→v1.0; v1.0→v1.1) showed overlap for most SEs (154/183, ⩾50 bp) but only a minority of BEs (1,359/10,398). Missed SEs showed no clear evidence of assembly defects upon manual inspection. Most SAS-missed BEs were located in homopolymer tracts (62.3%), with the remainder largely in identical regions or haplotype-switch contexts, suggesting that many curated BE corrections lie near current sequencing resolution limits. Finally, SAS identified residual signals in HG002 v1.1 outside rDNA that persist across recent versions, including misjoin signatures, abnormal-depth regions (>4 M b), and unsupported BE windows (**Figure. S8**). Independent analyses using Deeppolisher, DeepVariant, Sniffles2, and Flagger consistently reported more base-level and structural irregularities in HG002 v1.1 than in YAO v2.0, supporting the relative assessment provided by SAS (**Table 2**).

Improving the genome annotation

We generated the gene annotation of T2T-YAO v2.0 by transferring annotations from T2T-CHM13 v2.0 using Liftoff[28]. 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 3, Table S6**). To recover the missed protein-coding genes, we applied miniprot [29] which enabled the annotation of additional genes and extra gene copies based on MANE[30](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[31], 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 (**Table 3**). 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 S9**). 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 3**, **Table S6**). Notably, about one-third of the extra lncRNA copies are located on chromosome Y, with their biological significance largely unknown (**Figure S10**).

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 [32], encode protein sequences identical to their T2T-CHM13 counterparts (**Table**

S7). 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.

Without detectable errors in centromeres, T2T-YAO v2.0 provides unprecedented resolution of centromeric sequences. Refined annotation of centromeres, including higher-order repeats (HORs), and other satellite elements observed subtle but clear differences between T2T-YAO v2.0 and v1.1 (**Figure 4B, Supplement Information file 1**) [33,34], exemplified by updates to the paternal chromosome 1 assembly (**Figure S3B**). 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 S8**). 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[35], we constructed structural models for centromeric repeat clusters in T2T-YAO, confirming the correctness of the v2.0 centromere assemblies and revealed extensive variation in centromere size, structure, and repeat composition among chromosomes[34]. For instance, the HORs of maternal chromosomes 6 and 7 were substantially expanded compared to their paternal counterparts in T2T-YAO (**Figure S3A**).

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 [36] (**Figure 4B**). Motif composition and repeat length varied widely, without haplotype-, chromosome-, or arm-specific patterns, confirming earlier reports [37] but now with direct sequence evidence and strong read support, intriguing further investigation into telomere variations and their biological significance.

Implications in variant calling

Accumulating evidence indicates that high-quality, ancestry-matched reference genomes improve variant discovery by enhancing read mapping accuracy [6,38]. Compared with a uniform reference, ancestry-matched assemblies more effectively reduce mapping artifacts, resolve repetitive sequences, and sharpen functional inference. Leveraging the clearly resolved haplotypes in T2T-YAO v2.0, we estimated the impact of haplotype difference on variant detection. Across autosomes, small variants between T2T-YAO's paternal and maternal, 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 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 [34]. 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 S9**). 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 S9**). 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[39]. 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 [40], 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.

DISCUSSION

An error-free diploid human genome holds immense value for calibrating variant calling pipelines and enhancing the accuracy of clinical genetic diagnoses[7,41]. 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[42]. 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, providing not only more accurate baselines for variant discovery but also essential resources for precise human genomics.

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[9]. However, accurate evaluation of correctness and continuity remains more challenging and typically relies on alignment-based methods[43,44]. 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.

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, the simple MapQ often fails in diploid genomes, 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[45] and DeepVariant[12], 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[46]. (2) Phasing ONT and SPRQ reads with stringent criteria into haplotypes based on the hapmers they contain prior to mapping. 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.

Limitations of the study

Despite its sensitivity, SAS still misses some base-level errors in identical 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[20], 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[47]. 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.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Induced pluripotent stem cells (iPSCs) used for the T2T-YAO assembly were obtained from our Lab [13]. The collection of original cell used to establish iPSCs was approved by the Ethical Review Committee of Linfen Central Hospital, China (Approval No. 2022-20-1). The collection and storage of human samples were registered with and approved by the Human Genetic Resources Administration of China (HGRAC). Written informed consents were obtained from the participants.

METHOD DETAILS

Sample and sequencing

iPS cell line culture

Induced pluripotent stem cells (iPSCs) used for the T2T-YAO assembly were cultured in mTeSR™ PLUS medium (StemCell Technologies, Catalog No. 05825) in vitronectin (1:100 dilution, GIBCO, Catalog No. A14700) coated 6-well plates passaged at a 1:10 ratio using 0.02% EDTA (Versene, GIBCO, Catalog No. 15040-066).

ONT ultra-long library preparation and sequencing

ONT ultra-long DNA were extracted using the Grandomics BAC-long DNA Kit. The Blue Pippin system (Sage Science, Beverly, MA) was used to retrieve large DNA fragments by gel cutting. Approximately 8–10 µg of genomic DNA was selected (> 50 kb) with the SageHLS HMW library system (Sage Science), and then processed using the Ligation sequencing 1D Kit (Catalog No. SQK-LSK114, Oxford Nanopore Technologies, Oxford, UK) according the manufacturer's instructions. DNA libraries (approximately 400 ng) were constructed and sequenced on the PromethION (Oxford Nanopore Technologies) at the Genome Center of Grandomics (Wuhan, China). ONT sequencing data was generated in the pod5 format and underwent quality control using dorado in sup mode V5.0.0. This process filtered out low-quality fail reads (reads with a quality score lower than 7) to generate high-quality pass reads. These pass reads could be directly used for subsequent assembly.

Pacbio Revio® SPRQ library preparation and sequencing

High molecular weight genomic DNA was prepared by the CTAB method and followed by purification with Grandomics Genomic kit. According to the method description provided in the SMRTbell® prep kit 3.0 kit manual, the SPRQ libraries were prepared for sequencing. Sequencing data was subjected to quality control using SMRT Link V25.1 with default parameters.

Element AVITI library preparation and sequencing

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.

Pore-C library preparation and sequencing

Cells are fixed with formaldehyde (1%) to preserve 3D chromatin structure, quenched with glycine, and washed. Crosslinked chromatin is digested with DpnII (New England Biolabs, R0543L), adjacent DNA fragments are ligated by T4 DNA ligase (Vazyme, N103-01). After reverse crosslinking, DNA is purified and prepared using the ONT Ligation Kit (SQK-LSK114) for end-repair and adapter ligation. Libraries are loaded on a PromethION at the Genome Center of Grandomics (Wuhan, China).

Reads mapping

Haplotype-specific long reads phasing and mapping

Before mapping, The ONT (>30kb) and SPRQ (>3k) reads were phased using splitHaplotype (https://github.com/KANGYUlab/genome-polish-pipeline, modified from the 'splitHaplotigs' script in Canu [48] with command: splitHaplotype -fastq -R long.fastq.gz -H mat.hapmer.meryl haplotype-Mat.fastq -H pat.hapmer.meryl haplotype-Pat.fastq -A unknown.fastq. Haplotype-specific reads were aligned to the corresponding haplotype assembly sequence using winnowmap[49] with parameters '-k 15 -W asm.meryl.repetit ive_k15.txt -ax map-pb/map-ont -Y'. Haplotype- unassigned reads were aligned to diploid assembly sequences.

NGS reads mapping

Element reads were aligned to diploid assembly sequences using BWA-MEM[50].

SAS assessment

Preprocessing of multi-platform sequencing data

To ensure the fidelity of all downstream analyses, we implemented a rigorous, technology-specific preprocessing and filtering pipeline for all input alignment (BAM) files.

For short-read data, alignments were subjected to stringent selection criteria, retaining only those reads that exhibited a perfect match to the reference genome (specified by the SAM tag NM:i:0) and a complete absence of soft or hard clipping events. This foundational step was critical for establishing a high-confidence baseline, essential for the subsequent high-precision identification of base-level errors.

For long-read data generated on PacBio HiFi (Revio) and Oxford Nanopore (ONT) platforms, our analyses were exclusively restricted to primary alignments. Secondary and supplementary

alignments were deliberately excluded to mitigate analytical ambiguities arising from reads mapping to multiple genomic loci. Furthermore, to manage artifacts associated with read fragmentation while preserving genuine structural information, we designed and applied an adaptive filter to regulate split alignments. The maximum permissible number of splits for a given read was calculated dynamically based on its length, following the equation:

$$\text{Max_splits} = 3.0 + (\text{Read_length_kb} \times 0.05)$$

This strategy effectively removes reads with excessive fragmentation, which often indicates low-quality sequencing or chimeric artifacts, while appropriately accommodating the natural correlation between read length and the incidence of split alignments in authentic long reads.

According to the aforementioned algorithm, read alignments were filtered with command "filter.sh long-read/short-reads -i in.bam -o filter.bam". All scrip file is available at https://github.com/KANGYUlab/genome-polish-pipeline). All filtered bam files from the same sequencing platform were merged.

Structural Error (SE) Detection

Framework for Structural Error Detection

Structural error (SE) detection was performed exclusively using ultra-long ONT sequencing data, leveraging its superior capacity to span large and complex genomic structures that are challenging for shorter reads. Our analytical framework integrates three independent modules, each targeting distinct structural features to provide comprehensive error identification.

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.

Insertion and deletion detection

Large insertions and deletions (INDEL, ≥50 bp) were systematically identified through CIGAR string analysis within 2-kb sliding windows across the genome. For each window, we quantified both the total count of insertion and deletion events and the number of unique reads supporting each event type. To ensure accurate representation of large deletions that may span multiple analysis windows, reads contributing deletion signals were allocated to all affected windows, thereby preventing underestimation of deletion events.

Anomalous depth profiling

Potential assembly collapses and structural misjoins were detected through systematic read depth analysis using mosdepth with 2-kb resolution windows. Genomic regions were flagged as potential structural errors if their local coverage deviated significantly from the genome-wide mean depth: regions with depth below 20% of the mean were classified as potential duplications or misjoins, while regions exceeding 200% of the mean depth were identified as potential collapses.

Clipping event analysis

Potential assembly misjoins were identified through comprehensive analysis of soft and hard clipping events within 2-kb sliding windows. A critical feature of this module is the deliberate exclusion of clipping events occurring at the precise chromosomal termini (i.e., the first and last base positions). This design choice specifically mitigates false positives arising from incomplete telomere extension in reference assemblies, thereby distinguishing genuine internal structural errors from assembly artifacts at chromosome ends.

Non-split Read Coverage (NScov%) Calculation

The Non-split Read Coverage (NScov%) was calculated to measure the genome-wide coverage provided by contiguous, non-chimeric reads. First, nanopore sequencing alignments (BAM file) were filtered with Samtools to retain only primary alignments, discarding all secondary and supplementary records. Next, any primary alignment containing an SA tag—indicative of a split-read or chimeric event—was also removed. The per-base depth of this filtered set was then

calculated using mosdepth. NScov% was defined as the proportion of the reference genome covered at a depth of at least 1× by these non-split reads.

Base-level Error Detection

Framework for base-level error detection

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[51] and Meryl[51] were used to assess k-mer-based quality metrics (QV), completeness, and phasing accuracy using 21-mers and 31-mers. 21-mers and 31-mer were counted in the child, maternal and paternal read sets, and haplotype-specific mers (hapmer) were created using Merqury scripts with the command 'hapmers.sh mat.meryl pat.meryl son.meryl'. The evaluation followed the methods described by McCartney et al[52]. The binary k-mer set comprised Element and SPRQ k-mers set with >1 occurrence, while the trinary k-mer set additionally included ONT read k-mers with >4 occurrences.

Variant Calling (SVs and SNVs)

Structural variants-like errors were called using Sniffles2[53] (ONT), Flagger[54] (ONT), and NucFlag (https://github.com/logsdon-lab/NucFlag) (SPRQ), with default parameters. Sniffiles2 SVs with

supportive reads less than 60% were removed. Base-level SNV-like errors were identified using DeepVariant [55] (ONT, --model_type=ONT_R104; Element, --model_type=WGS), Deeppolisher [56] (SPRQ), and GATK HaplotypeCaller [22] (Element). Variants with genotype quality (GQ) <20 were excluded. Heterozygous calls were also filtered out except for haplotype-aware calls from Deeppolisher.

Error Correction

Structural error correction

SE windows were first corrected using variant calls from Sniffles2[53] when available. In cases without confident calls, ONT reads spanning the SE and 5kb-10kb flanks were reassembled using Hifiasm[57] (--ont) or Flye[58] (--nano-raw). If required, flanking regions were extended to 50–100 kb for resolving large collapses or duplications. Reassembled contig was aligned back to reference, and the corrected sequences were substituted accordingly. Manual assessment and correction with Integrative Genomics Viewer (v2.6) [59] is often necessary for SE 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.

Base-level error polishing (PWC pipeline)

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

Candidate error sites were evaluated for reference sequence support across sequencing platforms. Sites were excluded as false positives if any platform showed >55% of supporting reads (⩾10 reads) supporting the reference sequence, with an additional window read-depth cutoff of ⩽300 for the Element platform.

nh-BEs were corrected using consensus sequences from supporting reads in the order of platform priority: Element > SPRQ > ONT. Consensus sequences were constructed using samtools consensus with adaptive thresholds based on sequence composition. For regions with AG or CT frequency <90%, normal mode was applied with consensus threshold 0.70. For regions with AG or CT frequency ⩾90%, relaxed mode was used with consensus threshold 0.50.

Consensus sequences were accepted only if minimum coverage (⩾10 reads) and consensus threshold requirements were met. Where consensus could not be reached from Element, SPRQ and ONT platforms were sequentially attempted. When all platforms failed, local ONT reassembly was performed for nh-BE regions.

h-BEs were polished using Element or SPRQ consensus only, due to ONT's known limitations in homopolymer resolution. The same consensus construction parameters and thresholds as for nh-BE were applied.

Polishing iterations continued until the number of detected BEs plateaued.

SAS assessment of HG002 assembly versions

All sequencing data and annotations of HG002 used in this study are publicly available. No new data were generated in this study. ONT UL reads were obtained from https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/scratch/HG002/sequencing/ont/12_1_22_R1041_ULCIR_HG002_dorado0.4.0/, HiFi reads were obtained from https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/HG002/assemblies/polishing/HG002/v1.0/mapping/hifi_revio_pbmay24/hg002v1.0.1_hifi_revio_pbmay24.bam convert to fastq, Element reads were obtained from https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/scratch/HG002/sequencing/element/trio/HG002/). CenSat region was obtained from https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/HG002/assemblies/annotation/centromere/hg002v1.1_v2.0/hg002v1.1.cenSatv2.0.noheader.bb. The correction patches VCF files between recent versions were obtained from https://github.com/marbl/hg002.

According to the part of “Reads mapping” and “SAS assessment” described earlier in the Methods, sequencing reads of HG002 sample from three platforms were aligned to each assembly version of HG002 (v0.9, v1.0.1, v1.1), followed by BAM filtering and SAS evaluation.

Benchmarking with simulated SVs and SNVs

For structural errors (SEs), we randomly introduced 4,900 large indels (52–9,998 bp) and 50 inversions (1,067–4,852 bp) into the hg002v1.1 and YAO v2.0 assemblies using SURVIVOR (v1.0.7), excluding acrocentric chromosomes. For base-level errors (BEs), we introduced 18,000 SNP

s and 2,000 small indels into both assemblies using SimuG (v1.0.1) with parameters -snp_count 18000 -indel_count 2000 -ins_del_ratio 0.5. Haplotype-specific long reads (ONT and HiFi), together with element reads, were aligned to the spike-in genomes according to the previous parameter process. We applied SAS for both SE and BE detection, and used Sniffles2 and DeepPolisher as comparison methods for SE and BE detection, respectively. We used BEDTools intersect to obtain the counts of true positives (TP), false negatives (FN), and false positives (FP), and subsequently calculated recall, precision, and the F1 score. The coordinates of residual error regions were lifted over from the original assemblies to the spike-in genomes using minimap2 (v2.27) and transanno (v0.4.4). For evaluation of simulated error detection, to avoid potential bias caused by positional shifts between the simulated SNV coordinates and the actual mismatch or misjoin sites based on reads alignment, we expanded the SAS-SE and SAS-BE regions by one detection window on both flanks (50 bp for BE detection and 2 kb for SE detection). All downstream performance assessments were conducted based on these expanded regions.

Identification of identical regions (IRs): IRs were identified using genmap (K=150, E=0) to generate a genome-wide k-mer mapping frequency bedgraph. Regions with non-unique k-mer coverage were identified, and their end positions were extended to 150 bp downstream from the start position, constrained by chromosome boundaries. Overlapping intervals were merged using bedtools merge to generate the final IR.

Annotation

Genome annotation

We generated annotations for T2T-YAO v2.0 maternal and paternal haplotypes separately by mapping genes and transcripts from the latest annotation of T2T-CHM13 (GCF_009914755.1-RS_2025_08) to T2T-YAO v2.0 each haplotype.

Before global annotation, we masked some complex genomics regions include the V(D)J gene segments and the ribosomal DNA (rDNA) arrays to minimize mapping errors in these regions. We lifted V(D)J gene annotations from the RefSeq GFF of the T2T-CHM13 (GCF_009914755.1-RS_2025_08) onto T2T-YAO v2.0 each haplotype using Liftoff [60], and curated manually to ensure accurate regions of V(D)J gene. Total rDNA arrays were identified by aligning reference rDNA

sequence KY962518.fasta to the T2T-YAO v2.0 using nucmer (--maxmatch -l 31 -c 100; delta-filter -i 96 -l 1000).

Then, we used Liftoff (-chroms chroms.txt -copies -sc 0.95 -exclude_partial -polish) to map all features from T2T-CHM13 (GCF_009914755.1-RS_2025_08) onto T2T-YAO v2.0 each haplotype, excluding V(D)J gene segments and rDNA arrays. We then manually removed the previously identified complex genomic regions (V(D)J gene segments and rDNA arrays) from the generated annotations. This process yielded 56919 genes on MAT and 56546 genes on PAT.

To obtain comprehensive annotation of T2T-YAO v2.0, particularly for protein-coding genes, we supplemented the annotation with GRCh38 MANE (v1.4) annotations. Following the extra-copy search submodule strategy in LiftOn[32], we aligned MANE protein sequences to T2T-YAO each haplotype using miniport[29] with default parameters. Comparing the miniprot gene feature to the previously annotations lifted from T2T-CHM13, we filtered out any miniprot features that overlapped ≥10% of their own length with an existing gene feature or spanned more than two existing adjacent gene loci to avoid two genes were annotated in the same genomic location. This procedure added 15 genes to MAT and 29 to PAT.

We further refined the annotation coordinates using the protein-maximization algorithm in LiftOn, generating a comprehensive and accurate T2T-YAO v2.0 gene annotation dataset for downstream analyses. Synteny of the genes between T2T-CHM13 and T2T-YAO were plotted using LiftoffTools[60] synteny module (liftofftools synteny -r <reference.fa> -t <target.fa> -rg <reference.gff3> -tg <target.gff3>).

Centromere and Telomere Annotation

Centromere regions were annotated using RepeatMasker [61](based on Dfam39 library) with default parameters, and only hits with scores >50 were retained. The α-satellite arrays were identified via HumAS-HMMER[62] with default parameters. The sequence composition of human centromeric α-satellite HOR arrays was visualized using CenMAP (v0.5.1) (https://github.com/logsdon-lab/CenMAP). The orientation, and organization of centromere protein B (CENP-B) binding motifs were annotated using Genomic Centromere Profiling (GCP) pipeline[63]. Telomeric repeats were detected and quantified using Teloscope with default parameters[64] [https://doi.org/10.1093/bioinformatics/btac460].

Analysis of Genomic Variations

T2T-YAO v2.0 was separated into paternal (pat) and maternal (mat) haplotypes. In addition to an internal comparison between the two haplotypes, each haplotype was independently aligned against both the T2T-CHM13v2.0 and GRCh38.p14.

For each pairwise comparison, whole-genome alignment was performed using Minimap2 (v2.28) [65](-cx asm5 --cs, --eqx). Syntenic one-to-one alignments were extracted from the resulting PAF files, sorted, and then processed with paftools.js call (-q 60 -L 1000) to generate a comprehensive VCF file. The VCF file was then filtered to retain only SNPs and short indels (1-3bp) for each chromosome comparison pair. Ti/Tv ratios, SNP and short indels density were calculated from these one-to-one aligned regions.

Functional consequences of all identified variants were annotated using the Ensembl Variant Effect Predictor (VEP, v114, https://github.com/Ensembl/ensembl-vep) with respect to the corresponding reference annotation for each pair. Putative loss-of-function (pLoF) variants were identified by the Loss-Of-Function Transcript Effect Estimator [66] plugin and flagged as 'High-Confidence' (HC).

Pore-C chromosomal configuration data analysis

Haplotype phasing were validated with Pore-C contact maps using wf-pore-C v1.3.0 (https://github.com/epi2me-labs/wf-pore-c) with parameters '--cutter 'DpnII' –ref asm.diploid.fasta'. Contact maps were visualized in Juicebox v2.20 (https://github.com/aidenlab/Juicebox).

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 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. 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.

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**
	abnormal depth	1806	1151	2957	423	1085	1508
	Indel/clipping	176	211	387	13	70	83
	not spanned	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**
	abnormal depth	1373	172	1545	6	0	6
	Indel/clipping	137	133	270	0	0	0
	not spanned	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
	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
	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

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
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. Residual issues in T2T-YAO v2.0 and HG002 v1.1*

	T2T-YAO v2.0		HG002 v1.1	
	Count	Length**	Count	Length**
SAS-detected SE (merged)				
Not spanned(internal)	0	0	5	522,000
Not spanned (telomeric)	0	0	27	112,756
Indel/clipping	0	0	70	179,127
Abnormal depth	1	12,000	160	4,562,000
SAS-detected BE (merged)				
nhBE	0	0	615	71,617
hBE	342	17,400	580	33,300
Sniffles2	18	6,017	126	902,692
DeepVariant (Element, homozygous)				
GQ>20	102	-	364	-
GQ>10	533	-	911	-
Deeppolisher				
GQ>20 (non-homopolymer)	53	-	556	-
GQ>20 (homopolymer)	1,325	-	423	-
GQ>10 (non-homopolymer)	608	-	5,223	-
GQ>10 (homopolymer)	10,863	-	4,096	-
Flagger(>50bp)				
Erroneous region	26	44,401	75	619,529
Duplicated region	91	241,364	0	0
Collapsed region	740	155,227	31,517	15,191,801

* All statistics are counted for issues outside rDNA cluster

**Lengths are reported for merged regions where applicable

Table 3. 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

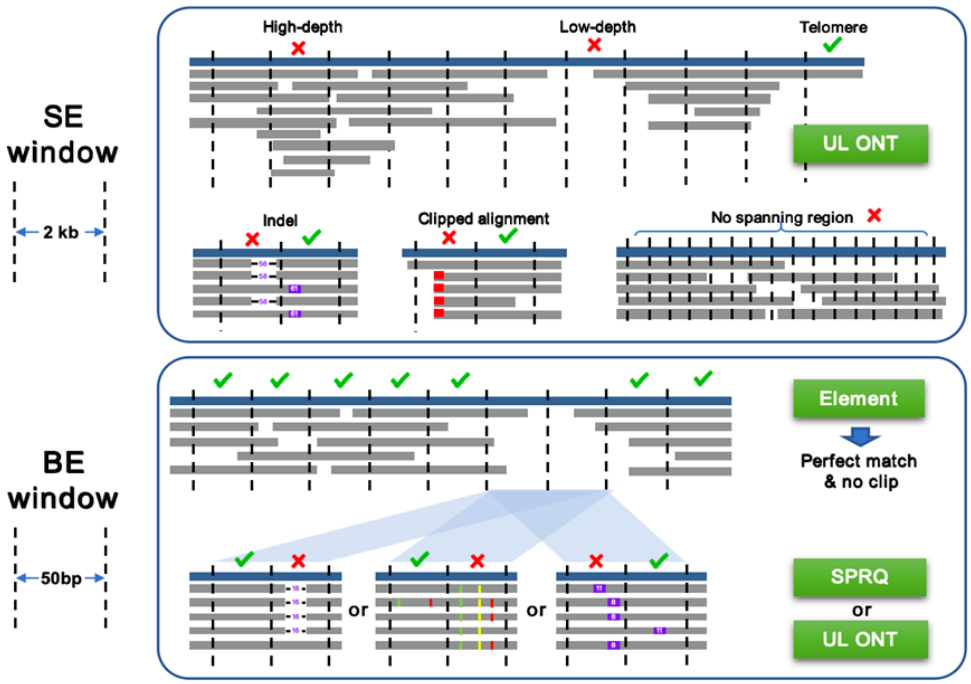
6. **Homopolymer sites lacking sufficient read support**: <0.02% of total homopolymer sites

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

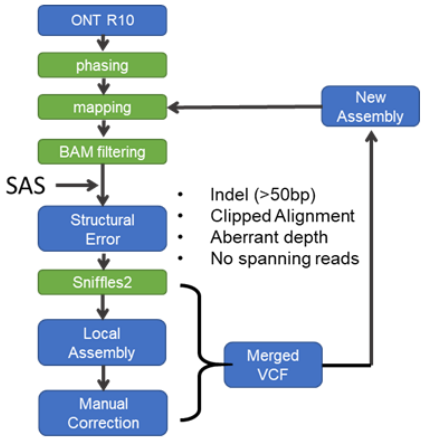
FIGURE TITLES AND LEGENDS

Figure 1

A



B



C

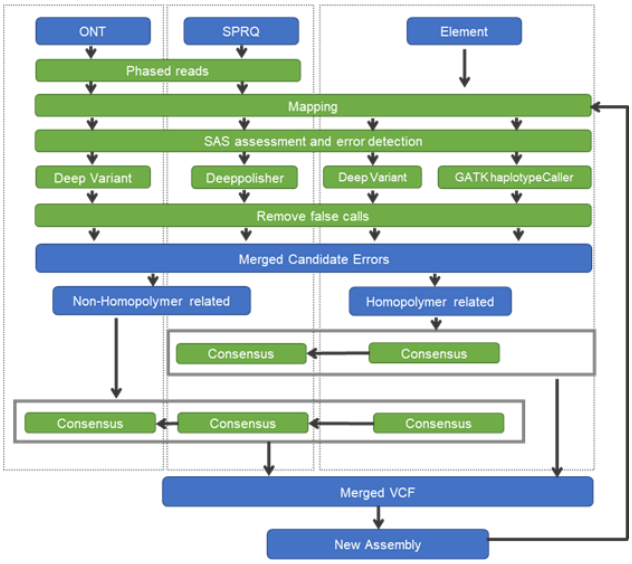
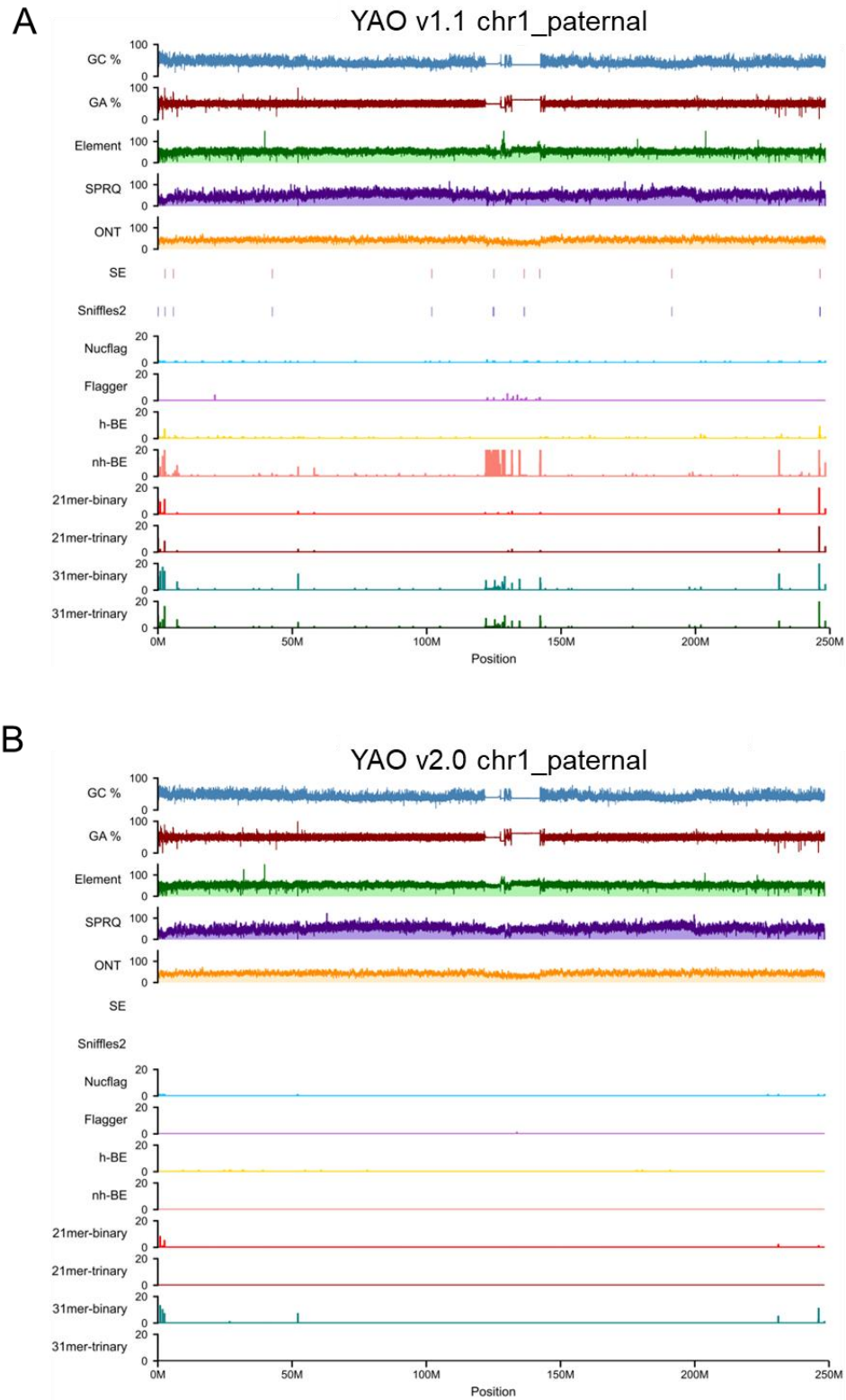


Figure 1. Overview of the sufficient alignment support (SAS) metric for detecting assembly errors and error correction platform-integrated window consensus (**PWC) pipeline.** (**A**). SAS identifies structural errors (SEs) and base-level errors (BEs) based on read alignment patterns across defined windows. 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 PWC (Platform-integrated Window Consensus) pipeline.

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
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
0M
50M
100M
150M
200M
250M
Position

976

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

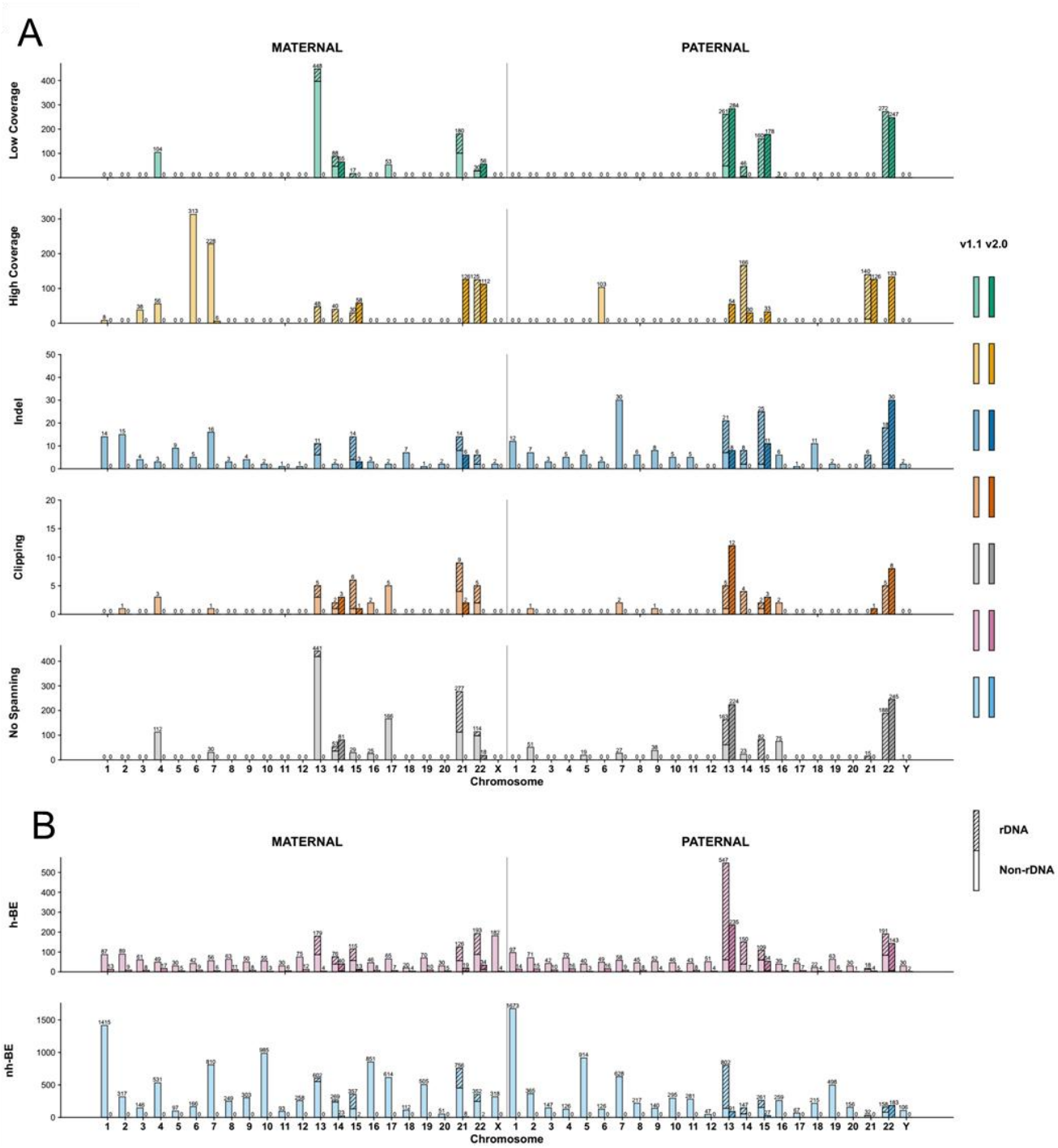
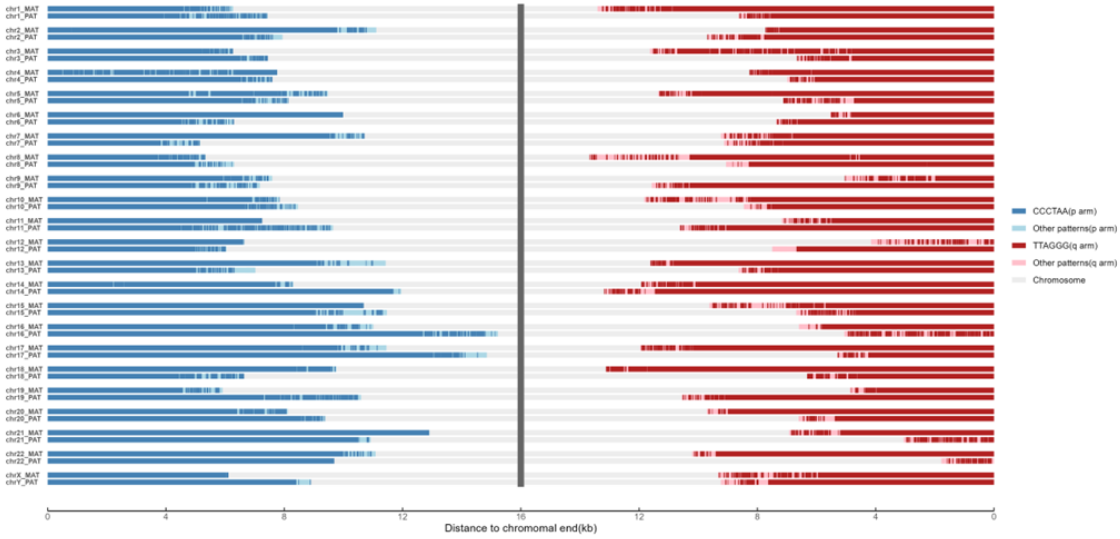


Figure 3. Counts of base-level and structural errors in T2T-YAO before and after iterative correction. Distribution of structural error (SE) windows (**A**) and base-level error (BE) windows (**B**) across chromosomes in maternal and paternal haplotypes of T2T-YAO v1.1 compared to v2.0. SEs

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

A



B

997

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

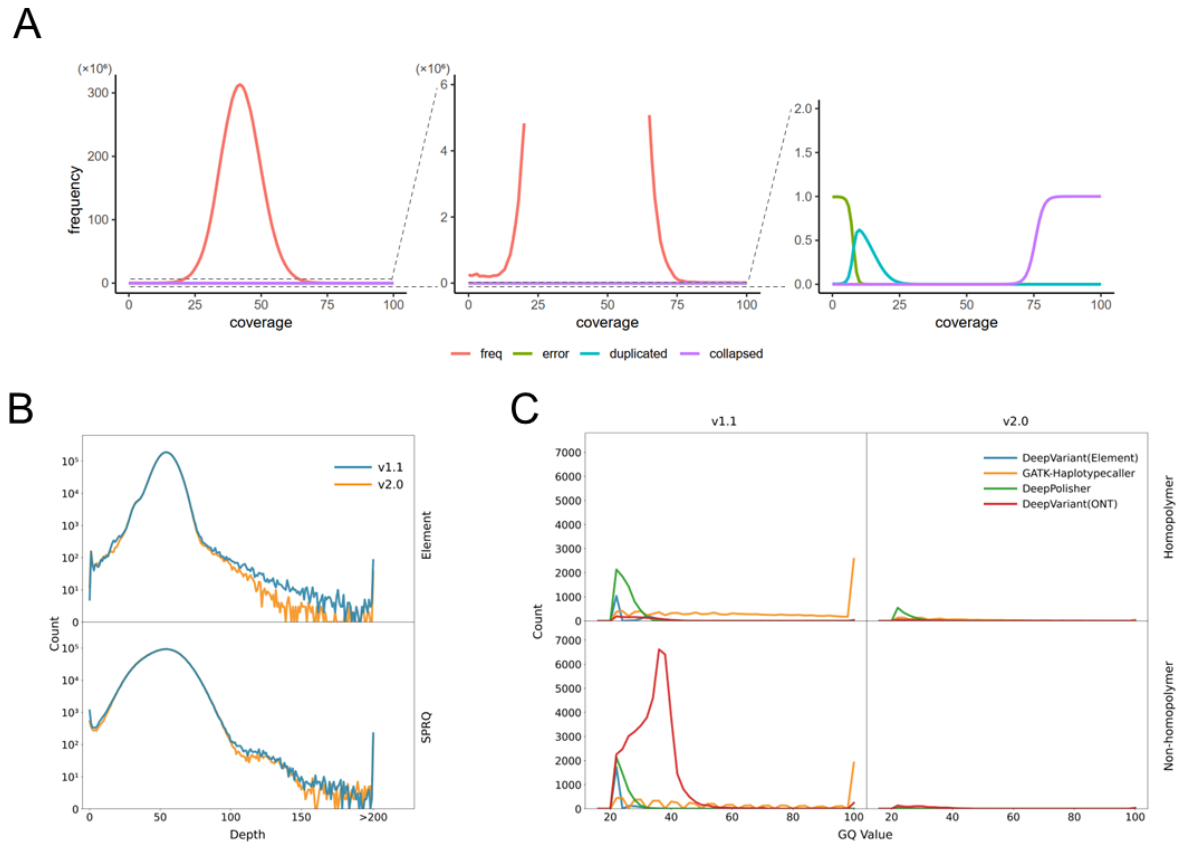


Figure 5. Read depth distribution and SNV-like errors in T2T-YAO v1.1 and v2.0

(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

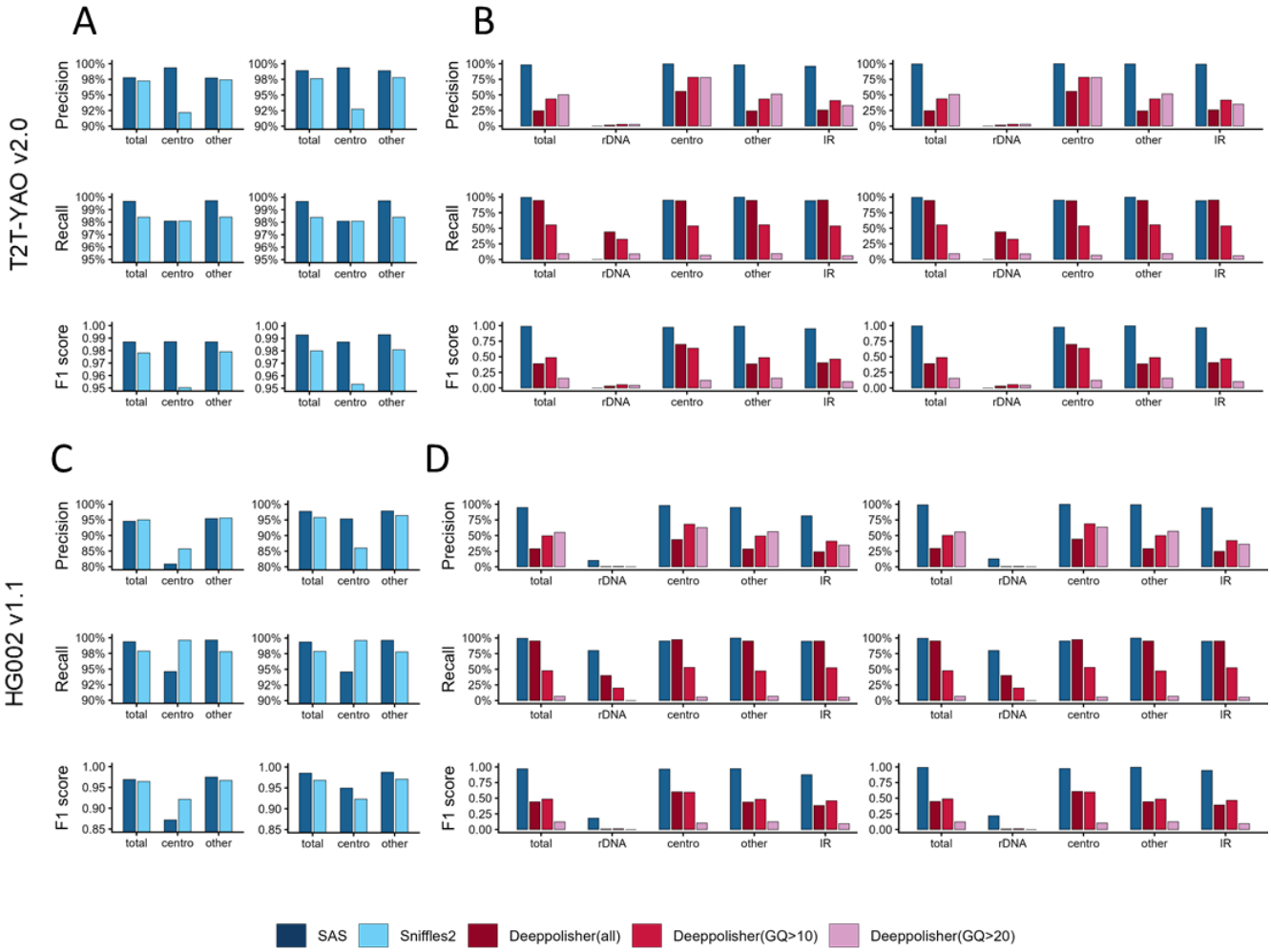
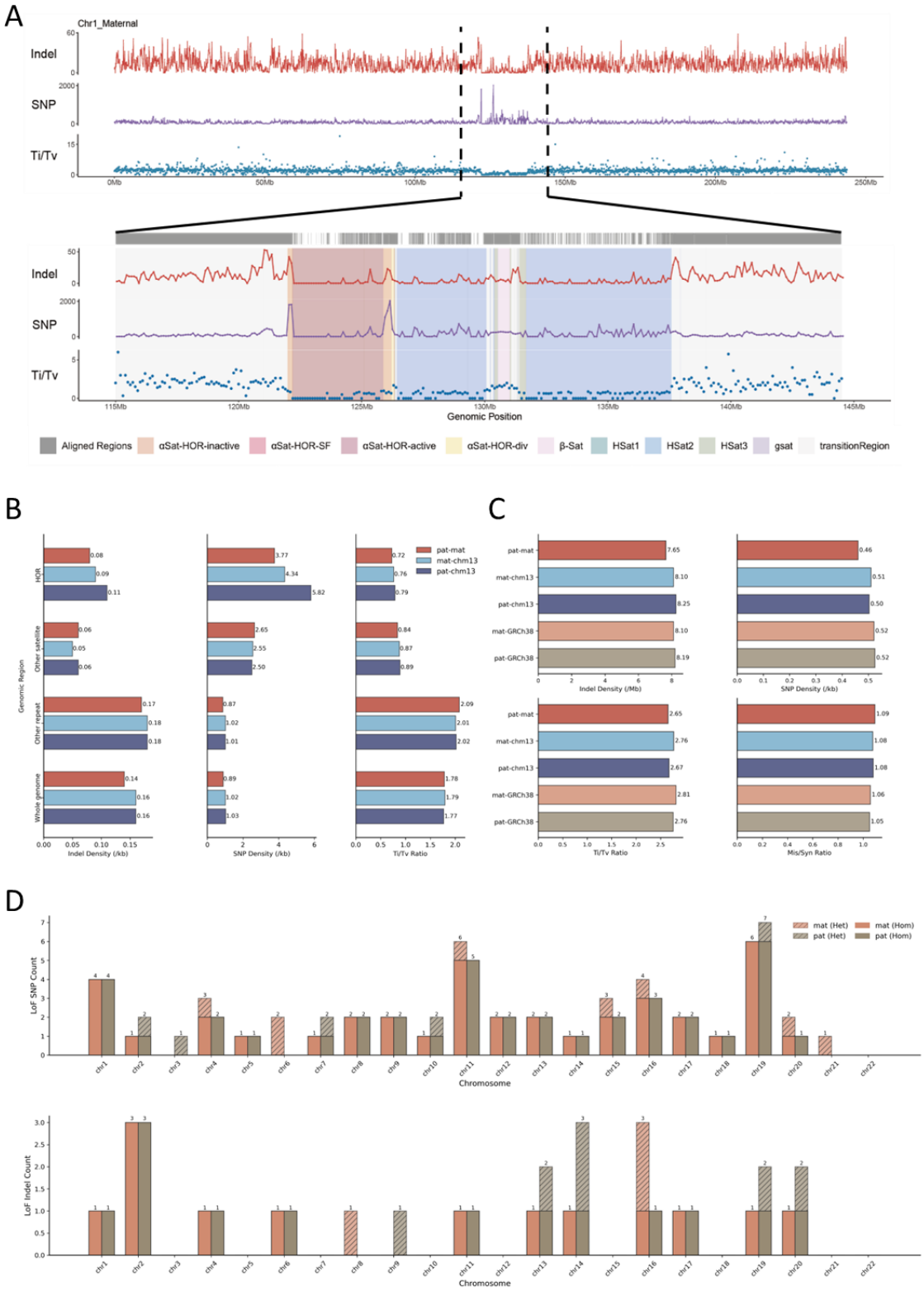


Figure 6 | Benchmarking the accuracy of SAS using simulated errors in T2T-YAO v2.0 and HG002 v1.1.

Precision–recall (PR) analysis of structural error (SE) and base-level error (BE) detection using simulated spike-in errors. **(A, B)** SE (A) and BE (B) detection in T2T-YAO v2.0 simulations. **(C, D)** SE (C) and BE (D) detection in HG002 v1.1 simulations. For each dataset, left panels show results obtained without accounting for residual errors in the reference assembly, whereas right panels show results after residual errors were removed.

Figure 7



1031

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 INFORMATION

Document S1.

Table S1-S9

Supplemental Information file 1: S1_Figures 1-45.

Supplemental Information file 2: S2_Figures 1-24.

Figure S1 Statistics of read depth and homopolymer length deviation.

(A) Genome-wide distribution of ONT read depth. Green solid lines indicate the upper and lower outlier cutoffs calculated using adjusted boxplot (based on MedCouple), while red dashed lines denote the median and thresholds for abnormal depth. **(B)** Median absolute deviation (MAD) of the number of repeat units along homopolymer tracks in SPRQ (left panel) and ONT (right panel) reads.

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. 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 S6. Distribution of three major types of homopolymers in each chromosome of T2T-YAO v2.0.

Figure S7. Distribution of k-mer frequencies across sequencing datasets. (A) 21mer. **(B)** 31mer.

Figure S8. Comparison of corrections and SAS-detected residual errors in recent versions of HG002.

Analysis of the corrections between recent versions of HG002, and their difference between SAS calls in SE **(A)** and BE **(B)**.

Figure S9. 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 S10. **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. Benchmarking the accuracy of SAS and PWC using simulated spike-in errors.

Table S6. The copy number and location of each gene in T2T-CHM13, maternal and paternal haplotype of T2T-YAO v2.0.

Table S7. Transcript variants and sequence identity between T2T-YAO v2.0 and T2T-CHM13

Table S8. Specific HOR units in T2T-YAO v2.0

Table S9. 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.

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 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

4 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

5 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

6 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

7 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

8 Formenti, G. & Howe, K. An assembly line for an improved human reference genome. *Nature* **611**, 519–+ (2022). https://doi.org/10.1038/d41586-022-03151-3

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

11 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

12 Shafin, K. *et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature methods* **18**, 1322–1332 (2021). https://doi.org/10.1038/s41592-021-01299-w

13 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

14 Merot, C., Oomen, R. A., Tigano, A. & Wellenreuther, M. A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends Ecol Evol* **35**, 561–572 (2020). https://doi.org/10.1016/j.tree.2020.03.002

15 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

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

17 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

18 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

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

20 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

21 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

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 Res* **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 Res* **30**, 415–426 (2020). https://doi.org/10.1101/gr.255349.119

28 Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics (Oxford, England)* **37**, 1639–1643 (2021). https://doi.org/10.1093/bioinformatics/btaa1016

29 Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics (Oxford, England)* **39** (2023). https://doi.org/10.1093/bioinformatics/btad014

30 Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022). https://doi.org/10.1038/s41586-022-04558-8

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

32 Chao, K. H. *et al.* Combining DNA and protein alignments to improve genome annotation with LiftOn. *Genome research* **35**, 311–325 (2025). https://doi.org/10.1101/gr.279620.124

33 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

34 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

35 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

36 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

37 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

38 Guo, S. *et al.* Enhancing Variant Calling in Whole-exome Sequencing Data Using Population-matched Reference Genomes. *Genomics Proteomics Bioinformatics* **22** (2024). https://doi.org/10.1093/gpbjnl/qzae070

39 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome biology* **17**, 122 (2016). https://doi.org/10.1186/s13059-016-0974-4

40 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). https://doi.org/10.1038/s41586-020-2308-7

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

42 Telomere-to-telomere-consortium. *HG002 Q100 project*, 2023).

43 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

44 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

45 Chen, Q., Yang, C., Zhang, G. & Wu, D. GCI: a continuity inspector for complete genome assembly. *bioRxiv*, 2024.2004.2006.588431 (2024). https://doi.org/10.1101/2024.04.06.588431

46 Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020). https://doi.org/10.1038/s41586-020-2547-7

47 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

48 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. *Genome Research* **27**, 722–736 (2017). https://doi.org/10.1101/gr.215087.116

49 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–+ (2022). https://doi.org/10.1038/s41592-022-01457-8

50 Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *Int Parall Distrib P*, 314–324 (2019). https://doi.org/10.1109/Ipdps.2019.00041

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

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

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

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

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

56 Mastoras, M. A.-O. *et al.* Highly accurate assembly polishing with DeepPolisher. (2024). https://doi.org/10.1101/2024.09.17.613505

57 Cheng, H. Y., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nature methods* **21** (2024). https://doi.org/10.1038/s41592-024-02269-8

58 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540–+ (2019). https://doi.org/10.1038/s41587-019-0072-8

59 Robinson, J. T., Thorvaldsdottir, H., Turner, D. & Mesirov, J. P. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics (Oxford, England)* **39** (2023). https://doi.org/ARTN btac830 10.1093/bioinformatics/btac830

60 Shumate, A. & Salzberg, S. LiftoffTools: a toolkit for comparing gene annotations mapped between genome assemblies. *F1000Res* **11**, 1230 (2022). https://doi.org/10.12688/f1000research.124059.2

61 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
62 Logsdon, G. A. *et al.* The variation and evolution of complete human centromeres. *Nature* **629** (2024). https://doi.org/10.1038/s41586-024-07278-3
63 Corda, L. & Giunta, S. Chromosome-specific centromeric patterns define the centeny map of the human genome. *Science (New York, N.Y.)* **389** (2025). https://doi.org/ARTN eads3484 10.1126/science.ads3484
64 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
65 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**, 3094–3100 (2018). https://doi.org/10.1093/bioinformatics/bty191
66 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). https://doi.org/10.1038/s41586-020-2308-7
67 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10** (2021). https://doi.org/10.1093/gigascience/giab008
68 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–842 (2010). https://doi.org/10.1093/bioinformatics/btq033
69 Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics (Oxford, England)* **34**, 867–868 (2018). https://doi.org/10.1093/bioinformatics/btx699
70 Marcais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, e1005944 (2018). https://doi.org/10.1371/journal.pcbi.1005944

STAR★METHODS

KEY RESOURCES TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
vitronectin	GIBCO	Catalog No. A14700
Critical commercial assays		
mTeSR™ PLUS medium	StemCell Technologies	Catalog No. 05825
Versene	GIBCO	Catalog No. 15040-066
Grandomics BAC-long DNA Kit	GrandOmics	Catalog No. XJZZ-BDE-003-1
Ligation sequencing 1D Kit	Oxford Nanopore Technologies	Catalog No. SQK-LSK114
Grandomics Genomic kit	GrandOmics	Cat. No. XJZZ-HMW-001-003
SMRTbell® prep kit 3.0	Pacific Biosciences	Cat. No.102-141-700
Elevate Enzymatic Library Prep Kit	Element AVITI	Cat. No. 830-00009
Long UDI Adapter Kit Set A	Element AVITI	Cat. No. 830-00010
AVITI 2×150 Sequencing Kit CB UltraQ	Element AVITI	Cat. No. 860-00018
DpnII	New England Biolabs	Cat. No. R0543L
T4 DNA ligase	Vazyme	Cat. No. N103-01

ONT Ligation Kit	Oxford Nanopore Technologies	Cat. No. SQK-LSK114
Deposited data		
raw sequencing data for T2T-YAO v2.0	This study	https://ngdc.cncb.ac.cn/gsa-human/browse/HRA011075
raw sequencing data for T2T-YAO v1.1	This study	https://ngdc.cncb.ac.cn/gsa-human/browse/HRA004987
T2T-YAO v2.0 genome	This study	material, https://download.cncb.ac.cn/gwh/Animals/Homo_sapiens_T2T_YAO_v2_mat_GWHGEYC00000000.1/GWHGEYC00000000.1.genome.fasta.gz; paterial, https://download.cncb.ac.cn/gwh/Animals/Homo_sapiens_T2T_YAO_v2_pat_GWHGEYB00000000.1/GWHGEYB00000000.1.genome.fasta.gz
T2T-YAO v1.1 genome	This study	material, https://download.cncb.ac.cn/gwh/Animals/Homo_sapiens_v1.1_GWHDQZJ00000000/GWHDQZJ00000000.genome.fasta.gz; paterial, https://download.cncb.ac.cn/gwh/Animals/Homo_sapiens_ChTY001.v1.1_pat_GWHDOOG00000000/GWHDOOG00000000.genome.fasta.gz
BAM alignment files for T2T-YAO v2.0 genome	This study	https://download.cncb.ac.cn/gsa-human/HRA011075/HRR2695327/HRR2695327.bam; https://download.cncb.ac.cn/gsa-human/HRA011075/HRR2695328/HRR2695328.bam; https://download.cncb.ac.cn/gsa-human/HRA011075/HRR2695329/HRR2695329.bam
Experimental models: Cell lines		
genomic DNA and iPSCs	This study	
Software and algorithms		
SAS	This study	https://github.com/KANGYUlab/sas-pipeline
genome-polish-pipeline	This study	https://github.com/KANGYUlab/genome-polish-pipeline

Meryl v1.3	[51]	https://github.com/marbl/meryl
Merqury v1.1	[51]	https://github.com/marbl/merqury
Winnowmap v2.03	[49]	https://github.com/marbl/Winnowmap
minimap2 v2.27	[65]	https://github.com/lh3/minimap2
BWA-MEM v0.7.17	[50]	https://github.com/lh3/bwa
SAMtools v1.19.2	[67]	https://github.com/samtools/samtools/
BEDtools v2.25.0	[68]	https://github.com/arq5x/bedtools2
IGV v2.17.3	[59]	https://github.com/igvteam/igv.js#igvjs
mosdepth	[69]	https://github.com/brentp/mosdepth
NucFlag v0.3.6	https://github.com/logsdon-lab/NucFlag	https://github.com/logsdon-lab/NucFlag
Flagger v0.3.2	[54]	https://github.com/mobinasri/flagger/
Sniffles2 v2.4	[53]	https://github.com/fritzsedlazeck/Sniffles
GATK v4.3.0.0	[22]	https://github.com/broadinstitute/gatk
Deepvariant v1.9.0	[55]	https://github.com/google/deepvariant
Deepploisher r0.1.0	[10]	https://github.com/google/deeppolisher
Flye v2.9.5	[58]	https://github.com/mikolmogorov/Flye
Hifiasm v0.25.0	[57]	https://github.com/chhylp123/hifiasm
Bcftools v1.20	[67]	https://github.com/samtools/bcftools/
RepeatMasker v4.1.9	[61]	https://www.repeatmasker.org/RepeatMasker
HumAS-HMMER	[62]	https://github.com/enigene/HumAS-HMMER
Genomic Centromere Profiling (GCP) pipeline	[63]	https://github.com/GiuntaLab/GCP-Centeny
CenMAP v0.5.1	[62]	https://github.com/logsdon-lab/CenMAP
LOFTEE v1.0.4	[66]	https://github.com/konradjk/loftee/tree/grch38
VEP	https://github.com/Ensembl/ensembl-vep	https://github.com/Ensembl/ensembl-vep
mummer	[70]	https://github.com/mummer4/mummer
Liftoff 1.6.3	[28]	https://github.com/agshumate/Liftoff

Miniprot-0.18 (r281)	[29]	https://github.com/lh3/miniprot
Lifton v1.0.5	[32]	https://github.com/Kuanhao-Chao/LiftOn
Liftofftools 0.4.4	[60]	https://github.com/agshumate/LiftoffTools
Gffutils	https://github.com/seandavi/GFFutils	https://github.com/seandavi/GFFutils
Teloscope v0.1.1	[64]	https://github.com/vgl-hub/teloscope
wf-pore-C v1.3.0	https://github.com/aidenlab/Juicebox	https://github.com/aidenlab/Juicebox
Genmap v1.3.0	https://github.com/cpockrandt/genmap	https://github.com/cpockrandt/genmap
ggplot2 v3.5.1	https://cran.r-project.org/web/packages/ggplot2/index.html	https://cran.r-project.org/web/packages/ggplot2/index.html
R v4.4.0	https://cran.r-project.org/	https://cran.r-project.org/
Matplotlib v3.7.1	https://matplotlib.org/	https://matplotlib.org/
NumPy v1.24.3	https://pypi.org/project/numpy/	https://pypi.org/project/numpy/
Pandas v2.0.3	https://pandas.pydata.org/	https://pandas.pydata.org/
Python v3.9.16	https://www.python.org/	https://www.python.org/