Efficient near-telomere-to-telomere assembly of nanopore simplex reads

https://doi.org/10.1038/s41586-026-10105-6

Received: 14 April 2025

Accepted: 6 January 2026

Published online: 04 February 2026

Open access

Check for updates

Haoyu Cheng[1,5] ✉, Han Qu[2,5], Sean McKenzie[3], Katherine R. Lawrence[3], Rhydian Windsor[3], Mike Vella[3], Peter J. Park[2] & Heng Li[2,4] ✉

Telomere-to-telomere (T2T) assembly is the ultimate goal for de novo genome assembly. Existing algorithms[1,2] capable of near-T2T assembly all require Oxford Nanopore Technologies (ONT) ultra-long reads, which are costly and experimentally challenging to obtain and are thus often unavailable for samples without established cell lines[3]. Here we introduce hifiasm (ONT), an algorithm that can produce near-T2T assemblies from standard ONT simplex reads, eliminating the need for ultra-long sequencing. Compared with existing methods, hifiasm (ONT) reduces computational demands by an order of magnitude and reconstructs more chromosomes from telomere to telomere on the same datasets. This advance substantially broadens the feasibility of T2T assembly for applications previously limited by the high cost and experimental requirement of ultra-long reads.

With the advent of accurate long reads—notably, Pacific Biosciences high-fidelity (PacBio HiFi) reads[4]—the latest generation of assembly algorithms has revolutionized de novo assembly[5–7]. For diploid genomes, these tools routinely produce haplotype-resolved assemblies, accurately reconstructing both haplotypes[8], but they often struggle with centromeres or long segmental duplications, owing to the limited read lengths of PacBio HiFi at 10–20 kb. To achieve near-T2T human assemblies in which each chromosome is completely resolved from end to end[9], existing assemblers, including Verkko[2,10] and hifiasm (UL)[1], have to rely on ONT ultra-long reads[11,12] of at least 100 kb to assemble through regions that fail HiFi reads[13]. However, generating ultra-long reads is costly and demands a large amount of high-molecular-weight DNA at tens of micrograms per human sample[3]—around 40 times higher than the DNA input with the standard ONT protocol. As a result, ultra-long reads are rarely produced for clinical specimens or by biodiversity projects. This greatly limits the practicality of near-T2T assembly.

There is an urgent need to develop algorithms to improve the accessibility of T2T genome assembly. The longer read lengths and rapidly improving accuracy of ONT reads offer the potential for achieving T2T assembly using ONT as the sole long-read sequencing technology. Current ONT sequencing protocols produce two types of reads: duplex and simplex. Duplex reads are as accurate as PacBio HiFi reads and work as well in assembly[14], but they are expensive to generate and rarely available. ONT sequencing therefore focuses mainly on simplex reads[3], which are longer in length and cheaper to produce. Nevertheless, the de novo assembly of simplex reads remains challenging because of their higher non-random, recurrent sequencing error rates[15]. These errors conflict with the key assumption, used in haplotype-resolved assembly algorithms such as hifiasm[1,6,8], HiCanu[7], LJA[5] and Verkko[2,10], that sequencing errors are random. At present, ultra-long ONT sequencing can generate only simplex reads. Therefore, in this study, the term 'ultra-long' refers specifically to ultra-long simplex reads.

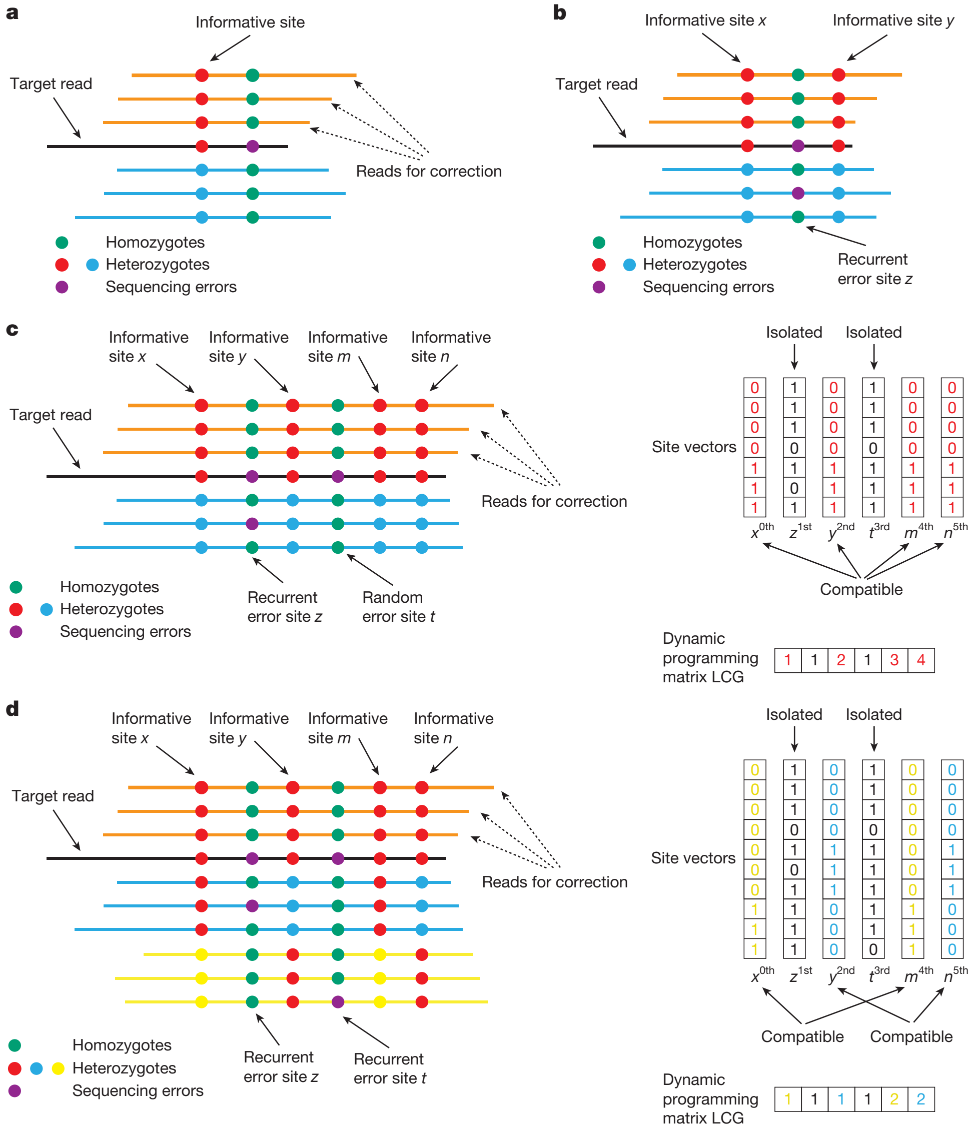
Several assembly workflows have been proposed to generate haplotype-resolved assemblies using ONT simplex reads[16–19]. They first build a consensus assembly by collapsing multiple haplotypes, and then reconstruct each haplotype through decompression from the consensus. This approach can fail in highly heterozygous regions and complex repetitive regions that cannot be accurately represented in the initial consensus assembly[20]. One error-correction tool, called HERRO[21], uses deep learning to correct ONT simplex reads before feeding them into existing assemblers such as hifiasm and Verkko. Although promising, HERRO is computationally intensive and demands high-end GPUs. Furthermore, although HERRO has been extensively validated using ultra-long simplex reads, it has not been shown to achieve near-T2T assembly with standard ONT simplex reads alone. HERRO is not an answer to near-T2T assembly at the population scale.

Here, to address the practical limitations of existing assembly methods, we developed hifiasm (ONT) to assemble the widely used ONT R10.4.1 standard simplex reads without ultra-long sequencing. It introduces a fast error-correction algorithm that uses read phasing to overcome the higher recurrent error rate of ONT simplex reads. On real data, hifiasm (ONT) often assembles multiple chromosomes from telomere to telomere with substantial reductions in time, labour and cost compared with existing methods.

The hifiasm (ONT) algorithm

Current assemblers optimized for PacBio HiFi reads all have an error-correction step to correct HiFi reads to nearly error-free. This step assumes that the errors are rare and random (Fig. 1a), which is approximately true for PacBio reads. However, the assumption does not hold for ONT simplex reads. In comparison with HiFi reads, ONT

[1]Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [3]Oxford Nanopore Technologies, Oxford, UK. [4]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA. [5]These authors contributed equally: Haoyu Cheng, Han Qu. ✉e-mail: haoyu.cheng@yale.edu; hli@ds.dfci.harvard.edu



x^{0th}	z^{1st}	y^{2nd}	t^{3rd}	m^{4th}	n^{5th}
0	1	0	1	0	0
0	1	0	1	0	0
0	1	0	1	0	0
0	0	0	0	0	0
1	1	1	1	1	1
1	0	1	1	1	1
1	1	1	1	1	1

1	1	2	1	3	4

x^{0th}	z^{1st}	y^{2nd}	t^{3rd}	m^{4th}	n^{5th}
0	1	0	1	0	0
0	1	0	1	0	0
0	1	0	1	0	0
0	0	0	0	0	0
0	1	1	1	0	1
0	0	1	1	0	1
0	1	1	1	0	1
1	1	0	1	1	0
1	1	0	1	1	0
1	1	0	0	1	0

1	1	1	1	2	2

Fig. 1 | Error correction of ONT simplex reads. a, Error correction in existing hifiasm for PacBio HiFi reads. Hifiasm identifies informative sites in which each allele (represented in red or blue) is supported by multiple reads. Sequencing errors are represented by purple dots. The algorithm then corrects the target read (black) using supporting reads (orange) that match the target read across all informative sites. **b**, Recurrent sequencing errors in ONT simplex reads. The existing error-correction approach in hifiasm incorrectly identifies recurrent sequencing errors as informative sites (illustrated as purple dots) because it is supported by two reads. In this case, all blue reads are correctly excluded owing to real informative sites (*x* and *y*), whereas all orange reads are mistakenly discarded because of the false-positive informative site *z*. With the existing error-correction approach, no reads remain available for correction. **c**, Error correction in hifiasm (ONT) with two haplotypes. The site vectors corresponding to informative sites *x*, *y*, *m* and *n* (highlighted in red) are identified as mutually compatible. These sites can be grouped into the same cluster using the dynamic programming matrix. Sites resulting from sequencing errors, such as *z* and *t*, are incompatible with other sites and remain unclustered. **d**, Error correction in hifiasm (ONT) with more than two haplotypes or repeat copies. The target read (black) and the orange reads originate from haplotype or repeat copy 1, the blue reads from haplotype or repeat copy 2 and the yellow reads from haplotype or repeat copy 3. Using the dynamic programming matrix, sites *x* and *m* can be grouped into one cluster (highlighted in yellow), whereas sites *y* and *n* form another cluster (highlighted in blue). Sites *x* and *m* exclude reads from haplotype or repeat copy 3, whereas sites *y* and *n* exclude reads from haplotype or repeat copy 2.

reads have higher error rates and the ONT sequencing errors tend to be recurrent—the same error can occur in multiple reads at the same genomic position (Fig. 1b). This makes it challenging to distinguish ONT sequencing errors from true heterozygous variants. As a result, error-correction algorithms in current HiFi assemblers do not work well with ONT simplex reads.

Hifiasm (ONT) overcomes the limitation of current methods by exploiting the phasing of long reads: a true heterozygous site is in phase with nearby heterozygotes sites, but a site loaded with recurrent sequencing errors is not (Fig. 1c,d). It used a dynamic programming-based algorithm for joint phasing and for identifying sequencing errors, and it considers base quality scores as well. With the new algorithm, hifiasm (ONT) can correct most ONT simplex reads to error-free. It also introduces other improvements to the assembly step, compared with earlier versions (Methods).

Human genomes sequenced by ONT standard

To demonstrate the capabilities of hifiasm (ONT), we generated standard ONT simplex reads for seven human samples and thoroughly evaluated the performance of hifiasm (ONT). These samples (HG001, HG002, HG003, HG004, HG005, HG006 and HG007) have been well

Table 1 | Statistics of different assemblies using ONT standard simplex (SUP basecalling model) and PacBio HiFi reads

Phased	Sample	Dataset	Approach	Wall time (h)	T2T count		Multicopy genes retained (%)	N50 (Mb)		QV
					Scaffold	Contig		Scaffold	Contig	
Full (trio)	HG001	ONT (66×)	Hifiasm	15.0	16/17	11/11	92.2/92.2	154.8/134.8	127.1/109.3	49.1/49.2
			Verkko+HERRO	126.7	3/5	0/2	91.8/92.6	97.4/96.2	59.0/62.5	51.4/51.6
		HiFi (25×)	Hifiasm	2.0	0/0	0/0	93.6/92.3	48.4/52.1	21.2/21.5	50.2/50.0
	HG002	ONT (47×)	Hifiasm	8.4	15/17	7/15	92.1/95.0	146.9/143.8	131.5/143.8	47.5/47.1
			Verkko+HERRO	103.1	3/2	0/0	91.2/94.4	91.8/90.1	52.3/46.0	49.0/48.5
		HiFi (52×)	Hifiasm	5.4	1/1	0/0	92.0/95.0	96.9/97.6	78.7/63.1	53.8/53.0
	HG005	ONT (56×)	Hifiasm	10.8	11/12	8/4	92.7/94.0	134.4/134.8	107.3/99.5	48.3/48.4
			Verkko+HERRO	86.7	1/2	0/0	92.7/93.7	94.2/94.0	51.7/58.7	50.6/50.7
		HiFi (50×)	Hifiasm	5.6	2/2	1/0	93.3/94.6	104.4/104.4	83.7/92.5	53.6/53.4
Partial (dual)	HG001	ONT (66×)	Hifiasm	14.9	12/12	10/8	89.3/93.6	135.4/134.3	133.6/101.3	49.1/49.1
		HiFi (25×)	Hifiasm	1.9	0/0	0/0	90.8/92.2	55.0/54.6	27.5/22.9	50.1/50.0
	HG002	ONT (47×)	Hifiasm	8.2	15/15	7/10	94.5/91.4	141.6/133.5	103.7/110.7	47.3/47.3
		HiFi (52×)	Hifiasm	5.3	1/0	0/0	91.4/92.3	98.3/95.3	71.0/77.9	53.0/53.8
	HG003	ONT (71×)	Hifiasm	18.6	16/13	12/8	96.0/94.3	135.7/133.4	135.7/120.2	48.2/47.7
		HiFi (74×)	Hifiasm	9.3	3/2	0/0	93.3/94.5	94.9/104.5	91.4/71.8	51.8/52.4
	HG004	ONT (66×)	Hifiasm	14.2	13/12	9/6	93.5/92.5	111.1/133.6	104.0/97.4	48.5/48.6
		HiFi (61×)	Hifiasm	6.6	2/0	1/0	91.4/94.5	95.6/92.0	64.1/55.9	53.2/54.3
	HG005	ONT (56×)	Hifiasm	10.6	12/8	9/0	92.6/91.7	134.4/133.6	118.2/88.3	48.4/48.4
		HiFi (50×)	Hifiasm	5.5	2/3	0/2	94.7/90.3	97.5/100.7	85.7/85.7	53.5/53.5
	HG006	ONT (56×)	Hifiasm	10.4	15/13	9/6	93.6/92.1	134.4/134.1	107.6/100.7	47.9/47.6
		HiFi (73×)	Hifiasm	9.4	5/3	2/1	91.5/94.9	96.6/96.4	87.2/71.3	55.6/56.0
	HG007	ONT (66×)	Hifiasm	12.9	12/13	10/7	94.2/94.6	135.4/135.0	134.9/102.6	47.5/47.8
		HiFi (61×)	Hifiasm	7.2	1/2	0/2	93.1/93.2	95.2/93.4	85.4/87.5	53.4/54.5

Each assembly comprises two sets of sequences. These two sets represent either paternal/maternal sequences (for trio-binning assemblies) or haplotype 1/haplotype 2 (for hifiasm dual assemblies). The two numbers in each cell give the metrics for the two sets of sequences, respectively. The N50 of an assembly is defined as the sequence length of the shortest contig or scaffold at 50% of the total assembly size. 'Multicopy genes retained' is the percentage of multicopy genes in CHM13 (multiple mapping positions at ≥97% sequence identity) that are multicopy in the assembly. They were calculated by the asmgene method[38], with CHM13[36] as the reference genome. The sequencing read coverage of each dataset is provided in the 'Dataset' column. The number of T2T contigs represents how many chromosomes were reconstructed without assembly gaps (complete contigs), and the number of T2T scaffolds indicates how many chromosomes were reconstructed either with or without assembly gaps (complete scaffolds). QV scores, which reflect per-base assembly accuracy, were evaluated using the *k*-mer-based tool yak. Comprehensive *k*-mer-based evaluation results using both yak and Merqury[39] are provided in Extended Data Table 1.

characterized by the Genome in a Bottle (GIAB) Consortium for benchmarking[22]. Each sample was sequenced using ONT simplex sequencing with one or two R10.4 flow cells, targeting the production of non-ultra-long reads with approximately 50× or higher coverage per genome (Methods). The average read N50 value of these datasets is 30 kb (Extended Data Fig. 1). This value, defined as the length of the shortest read that cumulatively covers 50% of the total read size, is often used as an indicator of the average read length.

As shown in Extended Data Fig. 1b, ONT standard simplex reads are generally longer than the PacBio HiFi reads that we collected for the same samples. Moreover, ONT simplex sequencing yields a broader read length distribution than does PacBio HiFi, increasing the likelihood of obtaining reads that are substantially longer than average (Extended Data Fig. 1a). Such long reads are particularly valuable for resolving complex and repetitive regions during assembly[23]. Compared with ONT ultra-long reads (typically with N50 > 100 kb), ONT standard simplex reads are shorter; however, they offer several practical advantages, including higher throughput, lower cost and a reduction of up to 40-fold in the DNA input requirements[3,12]. These benefits make ONT standard sequencing more accessible, especially for sample types for which ultra-long protocols are not feasible.

Benchmark ONT standard human assemblies

We compared the hifiasm (ONT) algorithm from the hifiasm toolkit to another contemporary T2T assembler, Verkko[2,10]. For HG001, HG002 and HG005, with parental data for phasing, we performed trio-binning assembly[24] using both hifiasm (ONT) and Verkko for comparison. Because Verkko cannot directly assemble ONT simplex reads because of their higher error rates, we first preprocessed the reads with HERRO[21] for error correction, before assembling them with Verkko. To evaluate the remaining samples without parental data, we also evaluated assembly quality using hifiasm (ONT) in its dual-assembly mode[8], which can generate high-quality assemblies using ONT reads alone. Unlike trio-binning assemblies that are fully phased, ONT-only dual assemblies are partially phased, relying on homologous information between the two haplotypes, rather than on additional data such as Hi-C or parental information. Despite being only partially phased, a dual assembly still represents a complete diploid genome and is highly contiguous. Verkko was not used for samples without parental data because it does not support dual-assembly mode. All hifiasm assemblies were performed on a standard computational server with 64 CPUs, whereas Verkko + HERRO was run on a combination of servers optimized for performance, using more than 64 CPUs and multiple GPUs (see Supplementary Information section 1.4 for more details).

As shown in Table 1, hifiasm is approximately an order of magnitude faster than Verkko + HERRO when assembling ONT simplex reads generated using the super-accurate (SUP) basecalling model. In addition, unlike HERRO, which requires high-end GPUs for error correction, hifiasm is an all-in-one toolkit capable of assembling directly from raw reads using only CPUs. This further reduces computational costs and simplifies the deployment. Moreover, hifiasm assemblies are consistently higher quality than are those produced by Verkko + HERRO. Using ONT standard simplex reads, hifiasm successfully reconstructs 9–22 chromosomes from telomere to telomere across different samples, whereas Verkko + HERRO fails to produce any complete T2T contigs, except for HG001, with two T2T contigs. This advantage of hifiasm is also evident in assembly contiguity, particularly in the contig N50. We noted that Verkko + HERRO slightly outperforms hifiasm in terms of quality value (QV) scores, which reflect per-base assembly accuracy (Table 1 and Extended Data Table 1). To investigate this, we compared the HG002 assemblies generated by both tools with the HG002 Q100 ref. 25. We found that most per-base errors unique to the hifiasm assembly originated from long homopolymer regions (see Supplementary Fig. 1 and Extended Data Table 2). High error rates in long homopolymer regions are a known limitation of ONT reads[26], making these errors challenging to correct through assembly algorithms. We observed large numbers of sequencing errors remaining within long homopolymer regions in both hifiasm and Verkko + HERRO assemblies, although Verkko + HERRO performed slightly better. Fully resolving these errors would require an additional polishing step, either using PacBio HiFi reads with higher base accuracy, or reusing the ONT simplex reads with a polisher such as Medaka or Dorado Polish, which is more computationally intensive.

In addition to the assembly results shown in Table 1, which use ONT reads basecalled with the SUP model, we also tested hifiasm (ONT) assemblies using reads produced by the high-accuracy (HAC) model (Supplementary Table 1). Compared with the SUP model, the HAC model is about ten times faster in basecalling on an NVIDIA H100 GPU, but at the expense of lower base accuracy[27]. Nonetheless, for each sample, hifiasm (ONT) achieved a comparable assembly quality using the less accurate HAC reads, with QV scores approximately three points lower on the Phred scale. These QV scores can be further improved by applying a post-assembly polishing tool such as Dorado Polish, which is designed to work with both HAC and SUP reads and can substantially improve the base-level accuracy of the draft assembly.

Comparison of ONT standard and PacBio assemblies

We also compared assemblies generated from ONT standard simplex reads and PacBio HiFi reads. For fair comparison, we produced HiFi-based assemblies for each sample at similar coverage levels, except for HG001, for which there were insufficient publicly available HiFi data. As illustrated in Table 1, ONT assemblies exhibit substantially higher contiguity, as indicated by their N50 values and the number of T2T contigs and scaffolds. This superior contiguity is results mainly from the fact that ONT standard reads are tens of kilobases longer than PacBio HiFi reads. Evaluating assembly quality within repetitive regions by examining multicopy-gene retention rates, we found that ONT assemblies were comparable to PacBio HiFi assemblies in quality across all samples. This shows that despite the high sequencing error rates in ONT simplex reads, our algorithm effectively corrects these errors, while avoiding the common problem of collapsing highly similar repeats that is observed in existing ONT assemblers[20] such as Napu (Shasta)[19] (see Supplementary Table 2). We further compared the number of annotated immunological genes per human genome assembly[28] (see Supplementary Fig. 2). These genes, including the HLA and KIR loci, are typically repetitive and multi-allelic, indicating that ONT assemblies achieve a quality comparable to that of PacBio HiFi assemblies. As anticipated, ONT assemblies had lower QV scores than did PacBio HiFi assemblies, owing mainly to persistent sequencing errors in long homopolymer regions, which are challenging to fully correct. The lower per-base accuracy of ONT assemblies also results in slightly higher phasing switch and Hamming error rates, compared with PacBio HiFi assemblies (Extended Data Table 1). In terms of computational performance, ONT assemblies typically require 1.5 to 2 times longer than HiFi assemblies at similar coverage levels; however, most ONT assemblies can be completed within approximately half a day using 64 CPUs.

Assembly with ONT ultra-long reads

We then evaluated the performance of hifiasm (ONT) and Verkko + HERRO using ONT ultra-long reads (Fig. 2). To ensure a comprehensive comparison, we tested two diploid human genomes (HG002 and HG02818) from the Human Pangenome Reference Consortium (HPRC)[29] and three haploid non-human genomes: *Arabidopsis thaliana*[21], *Danio rerio* (zebrafish)[30] and *Solanum lycopersicum* (tomato)[14]. Consistent with previous results, hifiasm (ONT) remained an order of magnitude faster than Verkko + HERRO and did not require GPU resources. Although Verkko + HERRO assemblies improved substantially with ultra-long reads, hifiasm (ONT) still outperformed it in all quality metrics except QV (Fig. 2 and Extended Data Table 1). For most samples, hifiasm (ONT) successfully reconstructed most chromosomes from telomere to telomere: 41 out of 46 for HG002; 44 out of 46 for HG02818; 3 out of 5 for Arabidopsis; and 21 out of 25 for zebrafish. Tomato is an exception, probably owing to its lower sequencing coverage (33×, compared with 68× for HG002, 58× for HG02818, 153× for *Arabidopsis* and 145× for zebrafish). To further assess assembly quality in non-human genomes, we also tested the ONT assembly of *Linum usitatissimum* (flax)[31] (see Supplementary Table 3). Despite relying only on ONT standard reads, hifiasm (ONT) still outperformed Verkko + HERRO, reconstructing many chromosomes completely from telomere to telomere.

For HG002, using ultra-long reads instead of standard simplex reads further improved assembly quality, as evidenced by increased N50 values and a higher number of T2T contigs and scaffolds (see Fig. 2 and Table 1). For example, the number of T2T contigs produced by hifiasm (ONT) increased from 22 with standard reads to 33 with ultra-long reads. This ultra-long-based HG002 assembly of hifiasm (ONT) also reconstructs more T2T contigs and scaffolds than the recently reported Verkko HG002 assembly[10] requiring both HiFi and ultra-long reads. For the trio-binning assembly of HG002, hifiasm (ONT) produced 33 T2T chromosomes at the contig level and 44 at the scaffold level, whereas Verkko produced 22 T2T contigs and 32 T2T scaffolds combining both ONT ultra-long and PacBio HiFi reads.

Resolving challenging medically relevant genes

In clinical genomics, a key application of de novo assembly is the accurate reconstruction of challenging medically relevant genes that are often difficult to resolve using conventional alignment-based approaches. The GIAB Consortium has previously used hifiasm with PacBio HiFi reads to assemble a curated set of 273 medically important genes that are particularly difficult to reconstruct owing to their high repetitiveness[32]. Despite the success of this strategy, more than 100 other medically relevant genes remain unresolved in HiFi-based assemblies.

A representative example is the pair of highly homologous genes *SMN1* and *SMN2*. Biallelic pathogenic variants in *SMN1* lead to spinal muscular atrophy (SMA), a progressive neurodegenerative disorder characterized by muscle weakness and atrophy resulting from a loss of motor neurons in the spinal cord[33]. Accurately determining the sequence of *SMN1* and its paralogue *SMN2* is crucial for diagnosing SMA and informing therapeutic decisions[32,34]. However, the HiFi-based assemblies used in the GIAB benchmarking framework are unable to fully resolve both *SMN1* and *SMN2*.

We evaluated the resolution of *SMN1* and *SMN2* in the haplotype-resolved assemblies generated by hifiasm (ONT) and Verkko + HERRO using both ONT standard and ONT ultra-long reads. To minimize potential inaccuracies caused by reference bias, we aligned the HG002 assemblies to the HG002 Q100 ref. 25, which was generated from the same

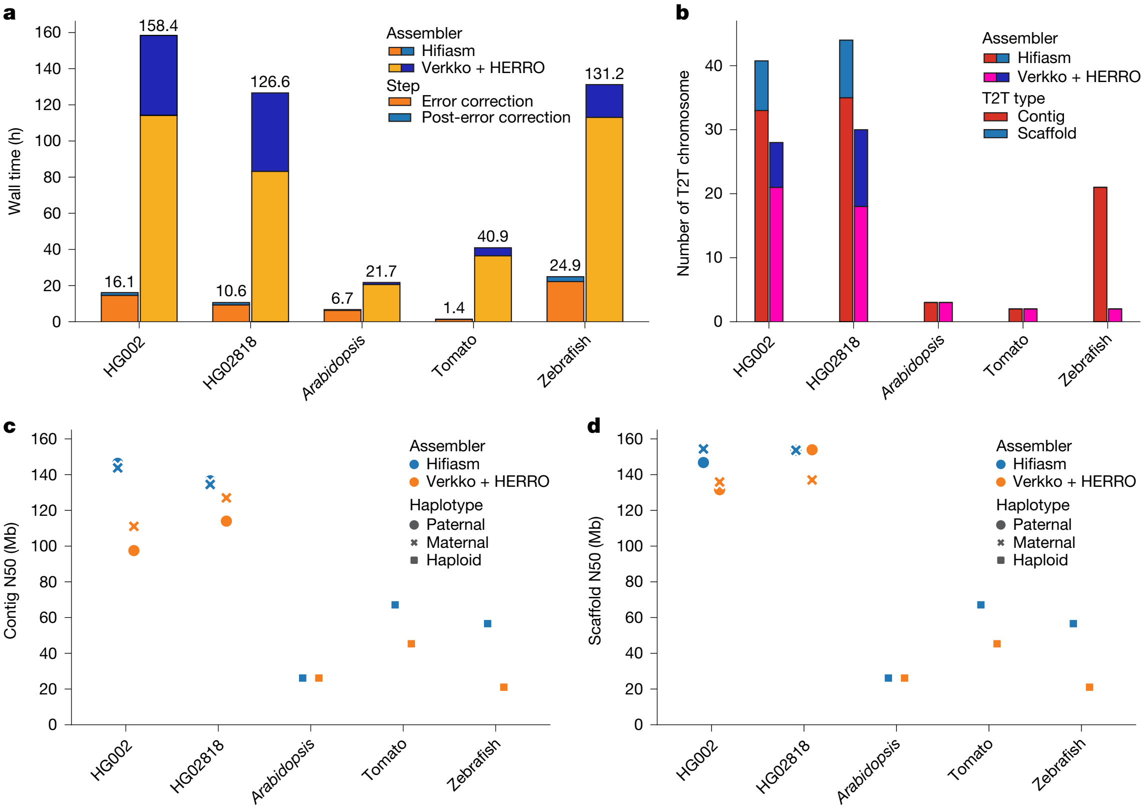


Fig. 2 | Results of assembly using ONT ultra-long reads. All assemblies generated by hifiasm (ONT) and Verkko + HERRO are used as is, except for the *Arabidopsis* assembly by hifiasm (ONT), in which low-coverage contigs were filtered out (Supplementary Information section 1.1). **a**, Running time of hifiasm (ONT) and Verkko + HERRO. Error-correction and post-error-correction times are shown separately for each sample. **b**, Number of T2T contigs and scaffolds. T2T contigs and scaffolds indicate entire chromosomes reconstructed without and with gaps, respectively. The genomes of human (diploid), *Arabidopsis* (haploid), tomato (haploid) and zebrafish (haploid) have 46, 5, 12 and 25 chromosomes, respectively. **c**,**d**, Assembly contiguity. Contig (**c**) and scaffold (**d**) N50 values. The assembly results for diploid human genomes (HG002 and HG02818) include two haplotypes, whereas the remaining three non-human genomes are haploid and have only one assembly.

individual. On the basis of the annotated coordinates of *SMN1* and *SMN2*, we selected the 77.5–78.5-Mb region on Chr. 5 (paternal) and the 71.5–73.0-Mb region on Chr. 5 (maternal) of the HG002 reference as the target loci. We then aligned these regions to the corresponding paternal and maternal assemblies produced by each assembler to assess the accuracy and completeness of *SMN1* and *SMN2* reconstruction.

As shown in Fig. 3, hifiasm (ONT) successfully reconstructs both *SMN1* and *SMN2* using either ONT standard or ONT ultra-long simplex reads. Notably, this shows that hifiasm (ONT) can resolve one of the most challenging medically relevant loci in the human genome using cost-effective and widely accessible ONT standard simplex reads. This capability broadens the feasibility of T2T assembly for clinical samples, for which ultra-long sequencing is often impractical. By contrast, Verkko + HERRO fails to faithfully reconstruct these regions. With ONT standard reads, it is unable to fully reconstruct either haplotype. When using ultra-long reads, Verkko + HERRO successfully assembles the paternal haplotype (Fig. 3c), but still fails to resolve the maternal haplotype (Fig. 3d). Hifiasm (ONT) outperforms Verkko + HERRO's ultra-long-read assemblies, even when using only ONT standard reads.

We then evaluated gene-level resolution in the fully phased HG002 assemblies. As in the multicopy-gene retention analysis in Table 1, genes were aligned to each assembly, and resolution was assessed by comparing the aligned genes to those in the reference genome. Instead of using a generic reference as in Table 1, here we used the HG002 Q100 reference from the same individual, ensuring an identical gene set across assemblies—which is crucial for accurately evaluating highly repetitive or multi-allelic genes. As shown in Extended Data Table 3, the HiFi assemblies contained hundreds of unresolved genes, many within medically relevant and difficult-to-assemble loci. By contrast, the ONT assemblies generated by hifiasm (ONT) reduced the number of unresolved genes by an order of magnitude and achieved gene-level resolution comparable with that of ONT ultra-long assemblies, even when using ONT standard reads. This is notable, because ONT standard reads are readily obtainable from clinical samples, whereas ONT ultra-long reads typically are not. With ONT standard reads, Verkko + HERRO achieved lower gene-level resolution than did hifiasm (ONT). Of note, the ONT standard read assembly using Verkko + HERRO resolved more challenging genes than the HiFi assembly did, despite a lower contig N50, highlighting the importance of read length in resolving challenging medically relevant genes.

Assembly correctness in human genomes

To evaluate different assembly approaches, we again used alignment to the HG002 Q100 reference. As shown in Fig. 4a and Supplementary Table 4, hifiasm (ONT) achieved the best results with ONT reads, outperforming the PacBio HiFi assembly and Verkko + HERRO assemblies. With hifiasm (ONT), ONT ultra-long reads not only yielded more contiguous assemblies than did ONT standard reads, but also resulted in the fewest misassemblies larger than 50 bp. Smaller misassemblies were ignored, because they can typically be resolved during post-assembly polishing. For samples lacking reference genomes, we used Flagger[29] and NucFlag[35] to perform reference-free evaluations by aligning reads back to their respective assemblies (Fig. 4b,c). The results again showed that hifiasm (ONT) generated the most accurate assemblies and that ONT assemblies were generally more accurate than PacBio

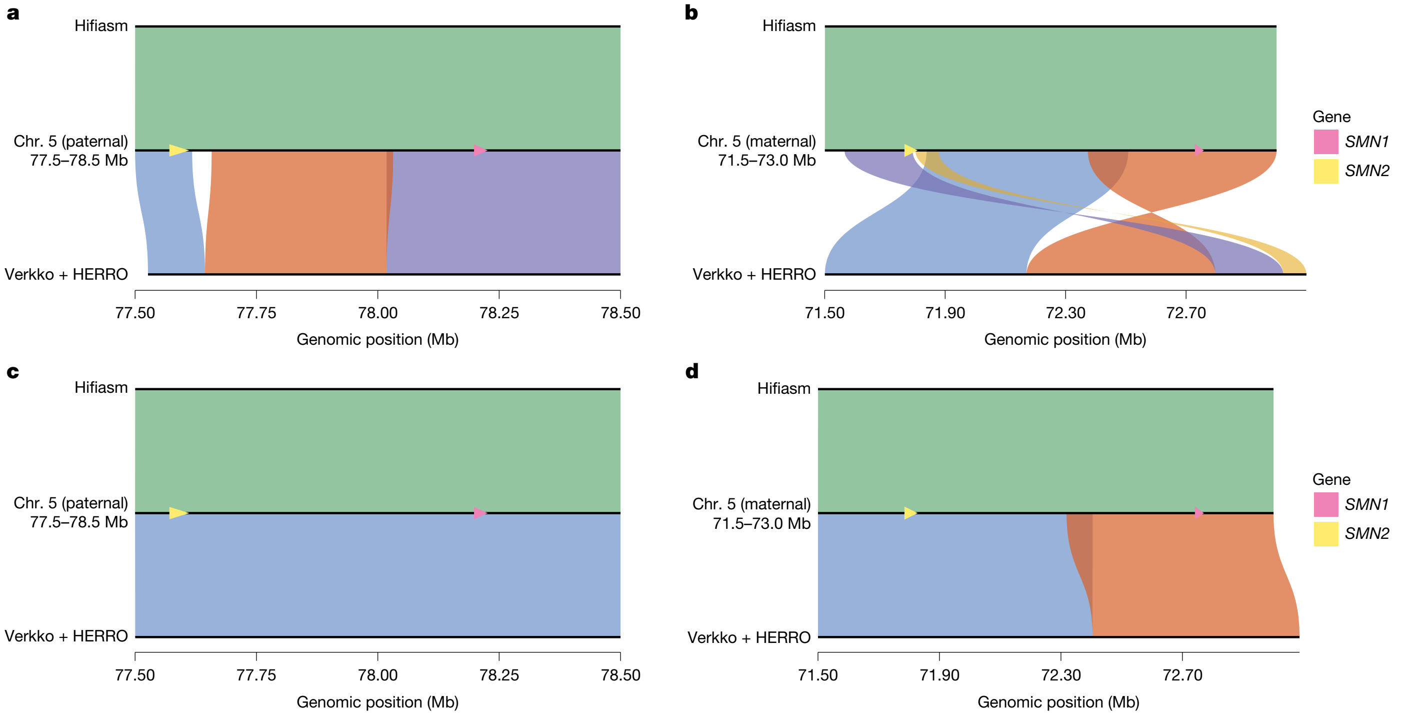


Fig. 3 | Comparison of HG002 assemblies with the HG002 T2T Q100 reference across the *SMN1* and *SMN2* region. Each plot, generated using SVbyEye[40], shows minimap2[38] alignment results of the assemblies to the annotated *SMN1* and *SMN2* regions within the HG002 Q100 reference. Alignments corresponding to the same contig are shown in the same colour. For consistency, several contigs aligned to the reverse-complement strand—such as the orange contig in Fig. 3b—are shown in their original orientations. Alignments of hifiasm (ONT) are shown above the reference, and alignments of Verkko + HERRO are shown below. The positions of *SMN1* and *SMN2* in the HG002 Q100 reference are highlighted in pink and yellow, respectively. **a**, HG002 paternal assemblies using ONT standard simplex reads. **b**, HG002 maternal assemblies using ONT standard simplex reads. **c**, HG002 paternal assemblies using ONT ultra-long simplex reads. **d**, HG002 maternal assemblies using ONT ultra-long simplex reads.

HiFi assemblies. We also observed that different misassembly evaluation methods yielded notably different results (Extended Data Fig. 2), with reference-free approaches tending to overestimate misassemblies relative to alignment-based ground truths. This suggests that reliable misassembly evaluation remains challenging. For HG002 fully phased assemblies, the genome-wide distribution of misassemblies was also assessed (Fig. 4d,e and Extended Data Fig. 3). The analysis revealed that most assembly errors occurred in highly repetitive regions such as centromeres.

Unresolved assembly gaps in human genomes

We next analysed the sequence composition of gaps that could not be resolved by the hifiasm toolkit. Assembly gaps were identified from alignments of contig ends that do not coincide with chromosome ends. For HG002, all contigs were mapped to the HG002 Q100 reference genome. As shown in Extended Data Fig. 4a–c, compared with the PacBio HiFi assembly, ONT assemblies resolved substantially more regions with high GA (TC) content, low complexity and extreme GC (AT) composition—areas that have been difficult to assemble using other sequencing technologies[9]. Ultra-long assemblies also resolved a substantially higher number of satellites, mainly from centromeric regions, indicating that ONT ultra-long reads remain essential for complete assembly of the human genome.

We also observed that ONT ultra-long assemblies contained more contig ends aligned to chromosome ends and segmental duplication regions. Most of these originated from small, erroneous contigs shorter than 500 kb, constructed from a limited number of reads with high sequencing error rates. By filtering out these small contigs, the number of contig ends within chromosome ends and segmental duplications was substantially reduced without compromising overall assembly quality, because these genomic regions were already represented in longer, higher-quality contigs. We hypothesize that the ONT standard assemblies contained fewer such erroneous contigs because ONT standard reads were processed using a newer version of the Dorado basecaller (v.0.7.2), whereas the ONT ultra-long assemblies were generated from reads processed using an older version (v.0.4.0). As Dorado continues to improve, base-level read accuracy will increase, reducing the frequency of erroneous short contigs.

Finally, we compared assembly gaps observed in PacBio HiFi and ONT simplex standard human genome assemblies (Extended Data Fig. 4d,e; 20 haplotypes in total for different assembly types, including both fully and partially phased assemblies). Because no complete reference genome is available for most samples, contigs were aligned to the CHM13 T2T reference. Although accurate alignment of contig ends can be challenging because of their repetitive and multi-allelic nature across individuals, we observed strong concordance: most assembly gaps clustered within repetitive regions. This observation is consistent with previous findings[9] and with the distribution of misassemblies in Fig. 4d,e and Extended Data Fig. 3. Notably, ONT standard assemblies now resolve many highly challenging genomic regions that remain unresolved in PacBio HiFi assemblies (Extended Data Fig. 4d,e). On the basis of haplotype-level density (red chromosomal bars), PacBio HiFi assemblies exhibit substantially more assembly gaps near chromosome ends (for example, chr. 16, chr. 19 and chr. X). However, because assembly gaps were measured by aligning assemblies to the CHM13 reference, there is alignment noise and uncertainty, making the haplotype-level density less accurate than the HG002-specific evaluation shown in Extended Data Fig. 4a–c. This is mostly because more than 5% of each assembly could not be aligned to the CHM13 reference, and vice versa, with most assembly gaps located around these regions. Such regions are not only difficult

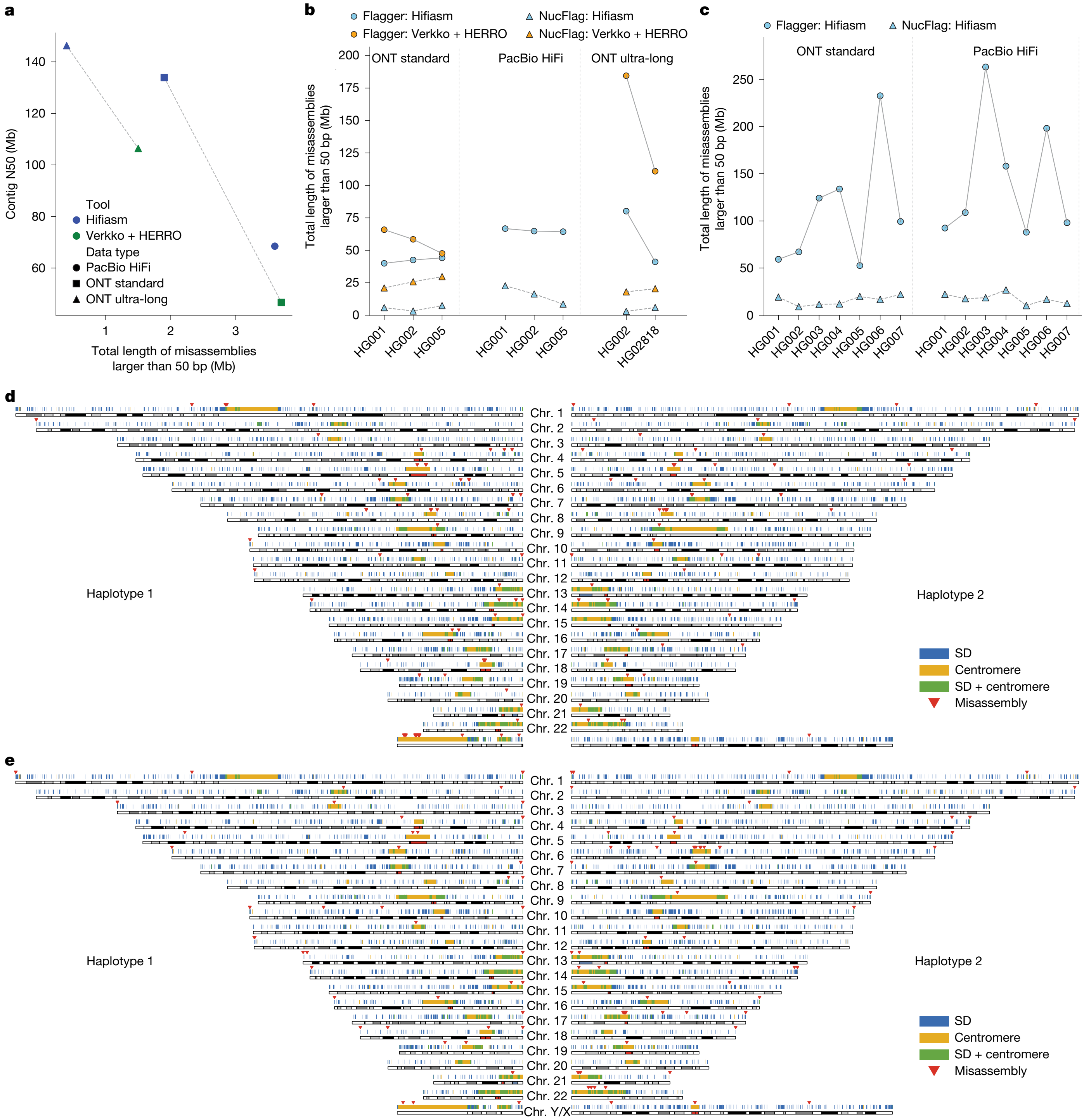


Fig. 4 | Evaluation of misassemblies in human genome assemblies. Only misassemblies larger than 50 bp are counted, because smaller errors can typically be corrected by polishing. For the HG002-specific results in **a**,**d**,**e**, misassemblies are accurately assessed against the HG002 Q100 ground truth. By contrast, other assemblies are approximately evaluated using Flagger[29] and NucFlag[35], because no ground truth is available. In **d**,**e**, segmental duplication (SD) and centromere annotations are shown in blue and in yellow, respectively, with overlapping regions highlighted in green and misassembly sites marked with red triangles. **a**, Relationship between misassembly length and assembly contiguity in fully phased HG002 assemblies. **b**, Misassemblies in all trio-binning human assemblies. **c**, Misassemblies in all non-trio-binning human assemblies. **d**, Genome-wide distribution of misassemblies in the HG002 ONT standard fully phased assembly produced by hifiasm (ONT). **e**, Genome-wide distribution of misassemblies in the HG002 ONT ultra-long fully phased assembly produced by hifiasm (ONT).

to assemble[36], often containing unresolved gaps, but also difficult to align[37], owing to their high sequence divergence among individuals. To mitigate this, we also provide the diploid-genome-level density (black bars above the chromosomal tracks), which represents the number of diploid genomes (out of ten total) with at least one haplotype exhibiting contig ends at each position. Both haplotypes of each diploid genome were merged and counted as a single observation to prevent double counting. This analysis shows more clearly that ONT assemblies outperform PacBio HiFi assemblies across the genome.

Discussion

Since the release of the first T2T human genome[36], reconstructing entire genomes from telomere to telomere has become feasible and is attracting increasing attention. Nevertheless, achieving this goal remains challenging for both sequencing technologies and computational methods. Here we present hifiasm (ONT), an efficient de novo assembly algorithm that is designed to use the long read lengths of ONT simplex data to achieve T2T assemblies without relying on complex hybrid assembly strategies. It introduces an improved error-correction approach that addresses the recurrent sequencing errors within ONT simplex reads. This enables high-accuracy de novo assembly methods, originally developed for PacBio HiFi data, to be applied directly to ONT R10.4.1 simplex reads. By contrast, existing tools—including current T2T hybrid assemblers—remain unable to perform de novo assembly of ONT simplex reads as effectively as PacBio HiFi data. By using the potential of ONT simplex sequencing fully, hifiasm (ONT) achieves superior performance over other assemblers that do not make effective use of this data type.

In comparison to HERRO, which was developed in parallel for error correction of ONT simplex reads, hifiasm (ONT) is substantially faster, owing to its use of an efficient dynamic programming method instead of a time-consuming deep-learning approach. This reduces the total assembly time by an order of magnitude and eliminates the need for multiple high-end GPUs (for example, four A100, L40S or RTX 8000 GPUs for assembling a human genome), effectively removing a major computational barrier for ONT simplex assembly. Furthermore, by tightly integrating and optimizing error correction with its ONT-specific assembly strategy, hifiasm (ONT) outperforms Verkko + HERRO by a wide margin across nearly all evaluation metrics.

With ONT ultra-long simplex reads, hifiasm (ONT) reconstructs more human chromosomes from telomere to telomere than any other approach. It not only outperforms Verkko + HERRO using the same ultra-long data, but also exceeds the performance of hybrid assemblies that combine ONT ultra-long and PacBio HiFi reads. Hifiasm (ONT) also shows that T2T genome assembly can be achieved using cost-effective and easily accessible ONT standard simplex reads. This advance enables population-scale T2T assembly and makes it feasible for clinical samples, in which ultra-long sequencing is often impractical. For example, hifiasm (ONT) is able to fully resolve the highly homologous and medically important gene pair *SMN1* and *SMN2*, which has remained unresolved in previous PacBio HiFi-based assemblies. We anticipate that hifiasm (ONT) will soon make routine near-T2T genome assembly possible across various research and clinical applications.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10105-6.

1. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat. Methods* **21**, 967–970 (2024).
2. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
3. Li, H. & Durbin, R. Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet.* **25**, 658–670 (2024).
4. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
5. Bankevich, A., Bzikadze, A. V., Kolmogorov, M., Antipov, D. & Pevzner, P. A. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat. Biotechnol.* **40**, 1075–1081 (2022).
6. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
7. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
8. Cheng, H. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
9. Porubsky, D. et al. Gaps and complex structurally variant loci in phased genome assemblies. *Genome Res.* **33**, 496–510 (2023).
10. 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).
11. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
12. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
13. Logsdon, G. A. et al. Complex genetic variation in nearly complete human genomes. *Nature* **644**, 430–441 (2025).
14. Koren, S. et al. Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *Genome Res.* **34**, 1919–1930 (2024).
15. Liu, Y. et al. Repeat and haplotype aware error correction in nanopore sequencing reads with DeChat. *Commun. Biol.* **7**, 1678 (2024).
16. Nie, F. et al. De novo diploid genome assembly using long noisy reads. *Nat. Commun.* **15**, 2964 (2024).
17. Lorig-Roach, R. et al. Phased nanopore assembly with Shasta and modular graph phasing with GFAse. *Genome Res.* **34**, 454–468 (2024).
18. Darian, J. C., Kundu, R., Rajaby, R. & Sung, W. K. Constructing telomere-to-telomere diploid genome by polishing haploid nanopore-based assembly. *Nat. Methods* **21**, 574–583 (2024).
19. Kolmogorov, M. et al. Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat. Methods* **20**, 1483–1492 (2023).
20. Jarvis, E. D. et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).
21. Stanojević, D., Lin, D., Nurk, S., Florez de Sessions, P. & Šikić, M. Telomere-to-telomere phased genome assembly using HERRO-corrected simplex Nanopore reads. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.18.594796 (2024).
22. Dwarshuis, N. et al. The GIAB genomic stratifications resource for human reference genomes. *Nat. Commun.* **15**, 9029 (2024).
23. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
24. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
25. Hansen, N. F. et al. A complete diploid human genome benchmark for personalized genomics. Preprint at *bioRxiv* https://doi.org/10.1101/2025.09.21.677443 (2025).
26. Delahaye, C. & Nicolas, J. Sequencing DNA with nanopores: troubles and biases. *PLoS One* **16**, e0257521 (2021).
27. Hall, M. B. et al. Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data. *eLife* **13**, RP98300 (2024).
28. Zhou, Y., Song, L. & Li, H. Full-resolution HLA and KIR gene annotations for human genome assemblies. *Genome Res.* **34**, 1931–1941 (2024).
29. Liao, W. W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
30. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
31. Arkhipov, A. A. et al. Nanopore data-driven chromosome-level assembly of flax genome. *Plants* **13**, 3465 (2024).
32. Wagner, J. et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).
33. Lunn, M. R. & Wang, C. H. Spinal muscular atrophy. *Lancet* **371**, 2120–2133 (2008).
34. Biros, I. & Forrest, S. Spinal muscular atrophy: untangling the knot? *J. Med. Genet.* **36**, 1–8 (1999).
35. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
36. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
37. Qin, Q. & Li, H. Challenges in structural variant calling in low-complexity regions. *Gigascience* **14**, giaf154 (2025).
38. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
39. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
40. Porubsky, D. et al. SVbyEye: a visual tool to characterize structural variation among whole-genome assemblies. *Bioinformatics* **41**, btaf332 (2025).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Overview of hifiasm (ONT)

The existing hifiasm assembly toolkit consists of three approaches: the original hifiasm[6], hifiasm (Hi-C)[8] and hifiasm (UL)[1], each designed for specific purposes. These methods are proposed, respectively, for trio-binning haplotype-resolved assembly using parental data, single-sample haplotype-resolved assembly using Hi-C reads and T2T hybrid assembly. The core component shared by these methods is constructing a high-quality assembly graph through de novo assembly of PacBio HiFi reads. However, owing to the limited length of PacBio HiFi reads, long, repetitive genomic regions often cannot be fully resolved within this core assembly graph.

To address this limitation, longer but less accurate ONT simplex reads, especially ultra-long reads, have been used by existing hybrid T2T assemblers such as hifiasm (UL)[1] and Verkko[2,10]. However, these assemblers do not make full use of ONT ultra-long reads because they cannot perform de novo assembly with them directly. The higher error rate and recurrent sequencing errors of ONT reads pose challenges for distinguishing genomic variations from sequencing errors. As a result, both hifiasm (UL) and Verkko use a two-stage strategy: first, constructing an assembly graph from PacBio HiFi reads with de novo assembly; and second, aligning ONT ultra-long reads to resolve regions that are unresolved by HiFi reads alone. For ONT reads, this alignment-based approach remains inherently limited by inaccuracies and biases introduced in the HiFi-based assembly graph.

Our hifiasm (ONT) approach enhances this core capability by enabling effective de novo assembly of ONT simplex reads, making full use of their longer length. Compared with HiFi-based assembly graphs, ONT-based assembly graphs tend to be more accurate and cleaner, and to resolve more repetitive regions. This improved ONT-based assembly graph substantially improves existing methods in the hifiasm toolkit, such as trio-binning or Hi-C phased assembly. The specific strategies of hifiasm (ONT) for using ONT simplex reads in T2T assembly are detailed below.

Error correction of ONT simplex reads

Error correction generates near-error-free reads by fixing sequencing errors in raw data, which is a crucial step for genome assembly. To correct a given target read *R*, assembly algorithms need to first collect all related reads originating from the same genomic region. Conventional algorithms assume that any overlapping read belongs to the same genomic region as *R* if they share high sequence similarity. However, this approach fails to distinguish highly similar repeat copies or haplotypes. As a result, many false-positive reads from other repeats or haplotypes might be incorrectly used to correct *R*, leading to an overcorrection issue that might collapse repeats and haplotypes in the final assembly.

To address this problem when assembling PacBio HiFi reads, hifiasm uses a key assumption that most sequencing errors in PacBio HiFi data occur randomly and typically appear in only a single read. Figure 1a shows an example. When correcting the target read *R*, hifiasm identifies mismatches through pairwise alignments between *R* and all overlapping reads. Mismatches that are supported by multiple overlaps are considered as informative sites representing true genomic variants, whereas those that appear in only one read are treated as sequencing errors and ignored. This strategy is essentially similar to widely used variant-calling methods. Subsequently, only overlapping reads showing no differences at these informative sites compared to *R* are used for correction.

However, this existing hifiasm error-correction strategy is unsuitable for ONT simplex reads (Fig. 1b). Although the overall sequencing error rate of ONT simplex reads is not significantly higher than that of PacBio HiFi reads, ONT simplex reads exhibit a higher frequency of recurrent, non-random errors. Therefore, the current hifiasm method would mistakenly identify these recurrent errors as informative sites. For a target simplex read *R*, overlapping reads originating from the same genomic region would thus be incorrectly discarded if they differed from *R* at these recurrent error sites. As a result, most ONT simplex reads would remain uncorrected, and could not be used for de novo genome assembly.

In hifiasm (ONT), we introduce an approach that makes use of the long-range phasing information of ONT reads to improve error correction. The basic idea is that true informative sites representing real genomic variants usually appear together and are mutually compatible with other informative sites. In practice, hifiasm (ONT) clusters potential informative sites on the basis of their compatibility. As shown in Fig. 1c, given a target ONT simplex read awaiting correction, overlapping reads at a potential informative site (x, y, m, n, z or t) can be classified into two phases: phase 0 (matching the target read) and phase 1 (differing from the target read). For two sites, if both consistently classify overlapping reads into identical phases, they are considered compatible and grouped together. True variants are expected to be compatible with multiple other real variant sites. By contrast, isolated sites that lack compatibility with others (such as z and t) have a higher likelihood of representing sequencing errors. This method is conceptually similar to using haplotype phasing to improve the accuracy of variant calling. To further enhance reliability, hifiasm (ONT) adopts the following criteria: (i) an isolated site is considered informative only if it is supported by a sufficiently high number of reads; and (ii) any grouped site that is supported by more than one read is regarded as an informative site, because such sites are inherently more reliable.

In practice, it is necessary to develop an efficient algorithm for grouping potential informative sites, because this operation must be performed for each read during error correction. To achieve this, we propose a dynamic programming method designed to identify the largest compatible group for each site. Specifically, let R be the target read awaiting correction, and S be the list of N potential informative sites within R, sorted by their positions. Here, $S[i]$ denotes the i-th site in the list, and $S[i][k]$ represents the phase assignment of the k-th read at site $S[i]$. The value of $S[i][k]$ can take one of the following states:

$$S[i][k] \in \{0, 1, *\},$$

where 0 and 1 indicate that the k-th read is assigned to phase 0 or phase 1, respectively, and * indicates that the k-th read does not cover site $S[i]$. The details of the dynamic programming method are described as follows:

1. Subproblem. Let LCG[i] be the size of the largest compatible group in S that ends at index i and is compatible with $S[i]$. The goal is to compute LCG[0] to LCG[$N-1$] for all sites and identify those with values greater than 1, which indicate a compatible group rather than an isolated site.
2. Recurrence relation. Formally, the recurrence relation is defined as follows:

   $$\mathrm{LCG}(i) = \max_{\substack{j<i \\ S[j] \leftrightarrow S[i]}} \{\mathrm{LCG}(j)\} + 1,$$

   where $S[j] \leftrightarrow S[i]$ indicates that $S[j]$ is compatible with $S[i]$. Two sites $S[j]$ and $S[i]$ are considered compatible if and only if

   $$S[i][k] = S[j][k] \text{ for all } k, \text{such that } S[i][k] \in \{0, 1\} \text{ and } S[j][k] \in \{0, 1\}.$$

   Figure 1c provides an example of the dynamic programming matrix LCG.
3. Traceback for grouping sites. Hifiasm (ONT) identifies any entry where LCG[i] > 1, starting from the highest value and proceeding downward. For each site $S[i]$ with $LCG[i] > 1$ that has not yet been assigned to a cluster, the algorithm traces back through its compatible prefix sites $S[j]$, following the path used to compute $LCG(i)$. For example, in Fig. 1c, hifiasm (ONT) starts from LCG(5), which

holds the highest score, and groups the corresponding sites $S[5]$, $S[4]$, $S[2]$ and $S[0]$ (that is, n, m, y and x) during the traceback process.

The time and space complexity of this dynamic programming method are $O(n^2)$ and $O(n)$, respectively, making it efficient for error correction in de novo assembly. An additional advantage of this approach is that it does not rely on the diploid-genome assumption, enabling it to handle polyploid genomes or highly similar repeats with more than two repeat copies. As shown in Fig. 1d, hifiasm (ONT) successfully identifies x and m as one group and n and y as another group when there are three haplotypes available.

We further improve the error correction by filtering out low-quality base pairs as follows:

1. Potential homopolymer sequencing errors. ONT reads are known to exhibit a high sequencing error rate within homopolymer regions. If a potential informative site is located in a homopolymer region, hifiasm (ONT) discards it, because it is more likely to result from homopolymer-induced sequencing errors.
2. Strand bias. Given the target read R, hifiasm (ONT) excludes an informative site if all reads that support R originate from one strand, whereas all other reads that differ from R at this site originate from the opposite strand. Strand bias is a common sequencing error observed in ONT reads.
3. Low base quality score. In addition to sequence data, hifiasm (ONT) also loads base quality scores into memory. Any base pair with a quality score lower than 10 is considered a potential sequencing error and excluded from the calculation of informative sites.

One potential challenge for hifiasm (ONT) arises when sequencing errors occur at the exact position of a true variant, making that variant incompatible with others. However, in practice, such cases are rare. Even if a true variant in one read is affected by sequencing errors, other nearby variants that remain unaffected can still be accurately detected, allowing hifiasm (ONT) to effectively separate haplotypes and resolve repeat copies.

Improved strategies for T2T assembly

With the error-correction approach in hifiasm (ONT), most sequencing errors within ONT simplex reads can be corrected effectively. These nearly error-free reads are then used with the existing assembly strategies in hifiasm to construct a high-quality assembly graph. For haploid genomes, a graph cleaning strategy is applied to produce linear assembly results. For diploid genomes, further data—such as parental or Hi-C reads—are required to generate haplotype-resolved assemblies using hifiasm's existing trio-binning or Hi-C phasing approaches.

To further improve T2T assembly, we developed a strategy to retain telomere sequences in the final assembly. A common issue in hifiasm is that although it can reconstruct entire chromosomes, it can still miss telomeric sequences at chromosome ends. This occurs because in the assembly graph—particularly for diploid genomes—telomere ends often appear as tips. During graph cleaning, hifiasm typically discards these tips, because most are caused by assembly errors. As a result, telomere sequences might be inadvertently removed from the final assembly. To address this, the improved T2T assembly strategy in hifiasm (ONT) first checks whether any reads contain telomeric sequences before the assembly. If such reads are detected, hifiasm preserves the corresponding graph tips during graph cleaning. This approach helps to retain more telomere ends and results in an increased number of T2T contigs and scaffolds.

We also developed a dual-scaffold approach to assemble more chromosomes from telomere to telomere at the scaffold level. The goal is to scaffold gapless contigs into longer, gapped scaffolds by using information from both haplotypes. The basic idea is that, for an assembly gap in haplotype 1, the dual-scaffold approach examines the corresponding homologous regions in haplotype 2. If the region in haplotype 2 is completely assembled without gaps, the dual-scaffold method fills the gap in haplotype 1 with ambiguous nucleotides (Ns), using the estimated length inferred from the complete sequence in haplotype 2. In essence, this approach performs reference-guided scaffolding for each haplotype[41], using the other haplotype as a reference.

ONT sequencing and basecalling

ONT standard simplex sequencing data for the GIAB samples HG001–HG007 have been deposited in the official ONT open data repository (s3://ont-open-data/giab_2025.01/). Cell lines for these samples were obtained from the Human Genetic Cell Repository at the Coriell Institute for Medical Research and cultured according to the supplier's recommended protocols. High-molecular-weight DNA was extracted using the QIAGEN Puregene cell extraction kit, followed by library preparation with the SQK-LSK114 kit according to ONT protocols, and sequencing was performed on PromethION flow cells using P48 instruments. Basecalling was done using Dorado v.0.7.2 with both HAC v.5.0.0 and SUP v.5.0.0 models. For HG001, HG003, HG004, HG005, HG006 and HG007, reads from two flow cells were basecalled and used for assembly. For HG002, only data from a single flow cell were used, because one flow cell was sufficient to produce ONT simplex reads at approximately 50× coverage. In addition, we re-basecalled the existing *D. rerio* (zebrafish) dataset using Dorado v.0.8.3 with the SUP v.5.0.0 model to improve read-level base accuracy.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Human reference genome: GRCh38, CHM13v2 and HG002 Q100; ONT standard reads of HG001 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG001; ONT standard reads of HG001 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG001; PacBio HiFi reads of HG001: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/HudsonAlpha_PacBio_CCS/; Illumina short reads of HG001 and parents: ERR194147 (HG001), ERR194160 (paternal), ERR194161 (maternal); ONT standard reads of HG002 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG002/PAW70337; ONT standard reads of HG002 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG002/PAW70337; ONT ultra-long reads of HG002: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/5b73fa0e-658a-4248-b2b8-cd16155bc157–UCSC_GIAB_R1041_nanopore/HG002_R1041_UL/dorado/v0.4.0_wMods/*ULCIR*.bam and https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/5b73fa0e-658a-4248-b2b8-cd16155bc157–UCSC_GIAB_R1041_nanopore/HG002_R1041_UL/dorado/v0.4.0_wMods/*ULNEB*.bam; PacBio HiFi reads of HG002: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_HiFi-Revio_20231031/; Illumina short reads of HG002 and parents: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/raw_data/Illumina/; ONT standard reads of HG003 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG003; ONT standard reads of HG003 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG003; PacBio HiFi reads of HG003: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/PacBio_HiFi-Revio_20231031/ and https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/PacBio_CCS_Google_15kb/; Illumina short reads of HG003: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/raw_data/Illumina/parents/HG003/; ONT standard reads of HG004 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG004; ONT standard reads of HG004 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG004;

PacBio HiFi reads of HG004: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/PacBio_HiFi-Revio_20231031/ and https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/PacBio_CCS_Google_15kb/; Illumina short reads of HG004: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/raw_data/Illumina/parents/HG004/; ONT standard reads of HG005 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG005; ONT standard reads of HG005 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG005; PacBio HiFi reads of HG005: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG005_NA24631_son/PacBio_CCS_15kb_20kb_chemistry2/uBAMs/; Illumina short reads of HG005: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG005/raw_data/Illumina/; ONT standard reads of HG006 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG006; ONT standard reads of HG006 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG006; PacBio HiFi reads of HG006: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG006_NA24694-huCA017E_father/PacBio_CCS_15kb_20kb_chemistry2/uBAMs/ and https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG006_NA24694-huCA017E_father/PacBio_HiFi_Google/; Illumina short reads of HG006: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG005/raw_data/Illumina/parents/HG006/; ONT standard reads of HG007 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG007; ONT standard reads of HG007 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG007; PacBio HiFi reads of HG007: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG007_NA24695-hu38168_mother/PacBio_CCS_15kb_20kb_chemistry2/uBAMs/ and https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG007_NA24695-hu38168_mother/PacBio_HiFi_Google/; Illumina short reads of HG007: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG005/raw_data/Illumina/parents/HG007/; ONT ultra-long reads of HG02818: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG02818/raw_data/nanopore/dorado0.7.2_sup4.1.0_5mCG_5hmCG/; Illumina short reads of HG02818 and parents: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG02818/raw_data/Illumina/; ONT ultra-long reads of *A. thaliana*: SRR29061597; Illumina short reads of *Arabidopsis thaliana*: https://ngdc.cncb.ac.cn/gsa/browse/CRA005350; ONT ultra-long data of *Danio rerio*: https://genomeark.s3.amazonaws.com/index.html?prefix=species/Danio_rerio/fDanRer17/genomic_data/ont/pod5/ and https://genomeark.s3.amazonaws.com/index.html?prefix=species/Danio_rerio/fDanRer17/genomic_data/ont/fast5/; Illumina short reads of *D. rerio*: https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=1029986; ONT ultra-long data of *S. lycopersicum*: https://obj.umiacs.umd.edu/marbl_publications/duplex/Solanum_lycopersicum_heinz1706/UL/R10.4_40x.noduplex.fastq.gz; Illumina short reads of *S. lycopersicum*: https://ngdc.cncb.ac.cn/gsa/browse/CRA003995/CRX232533; ONT reads of *L. usitatissimum* (flax): SRR31124331; ONT reads used for comparison with Napu (Shasta): https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/Napu_paper_ONT_Coriell_SingleFC_2023/HG0*_R10/reads/*.bam; and Napu (Shasta) assemblies: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/Napu_paper_ONT_Coriell_SingleFC_2023/HG0*_R10/assembly/*contigs*fasta. All evaluated assemblies generated by hifiasm and Verkko + HERRO are available at https://zenodo.org/records/15203417 (human genome assemblies from ONT reads) (ref. 42), https://zenodo.org/records/17613526 (non-human genome assemblies from ONT reads) (ref. 43) and https://zenodo.org/records/15205178 (assemblies from PacBio HiFi reads) (ref. 44). All links and accession identifiers for the sequencing data are also listed in Supplementary Table 5.

Code availability

Hifiasm (ONT) is released under the MIT licence and is freely available at https://github.com/chhylp123/hifiasm. The specific version used in this study has been archived at https://doi.org/10.5281/zenodo.18079611 (ref. 45).

41. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
42. Cheng, H. Efficient near telomere-to-telomere assembly of nanopore simplex reads (human genome assemblies from ONT reads). *Zenodo* https://doi.org/10.5281/zenodo.15203417 (2025).
43. Cheng, H. Efficient near telomere-to-telomere assembly of nanopore simplex reads (non-human genome assemblies from ONT reads). *Zenodo* https://doi.org/10.5281/zenodo.17613526 (2025).
44. Cheng, H. Efficient near telomere-to-telomere assembly of Nanopore simplex reads (HiFi assemblies). *Zenodo* https://doi.org/10.5281/zenodo.15205178 (2025).
45. Cheng, H. Efficient near telomere-to-telomere assembly of Nanopore simplex reads (hifiasm-0.25.0-r726). *Zenodo* https://doi.org/10.5281/zenodo.18079611 (2025).
46. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).

Acknowledgements This study was supported by the US National Institutes of Health (grants R00HG012798 and RM1HG011558 to H.C., R01HG010040, U01HG010971 and U41HG010972 to H.L. and UM1DA058230 and R01CA269805 to P.J.P.). We thank the HPRC for making ultra-long datasets publicly available.

Author contributions H.C. designed the algorithm, implemented hifiasm (ONT), drafted the manuscript and benchmarked hifiasm (ONT). H.Q. designed the evaluation of genome assemblies and led the comprehensive benchmarking of all approaches, supervised by P.J.P. S.M., K.R.L., R.W. and M.V. collected nanopore sequencing data and contributed to the assembly evaluation. H.L. directed the entire project. All authors contributed to writing the manuscript.

Competing interests S.M., K.R.L., R.W. and M.V. are employees of ONT. The remaining authors declare no competing interests.

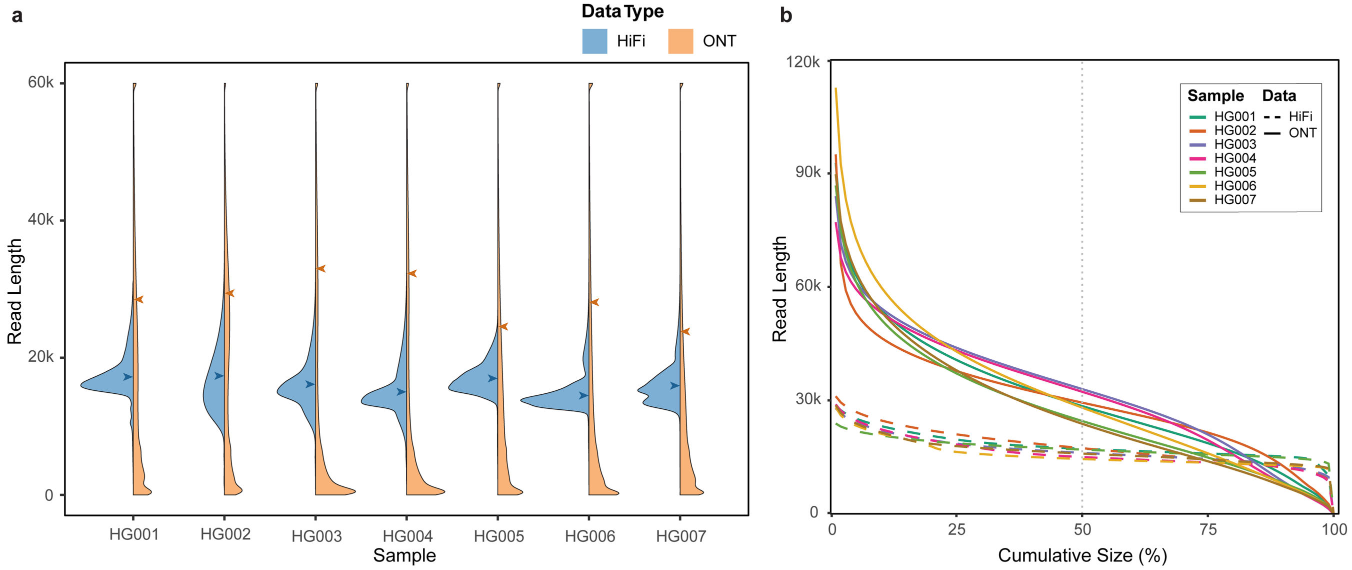
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10105-6.

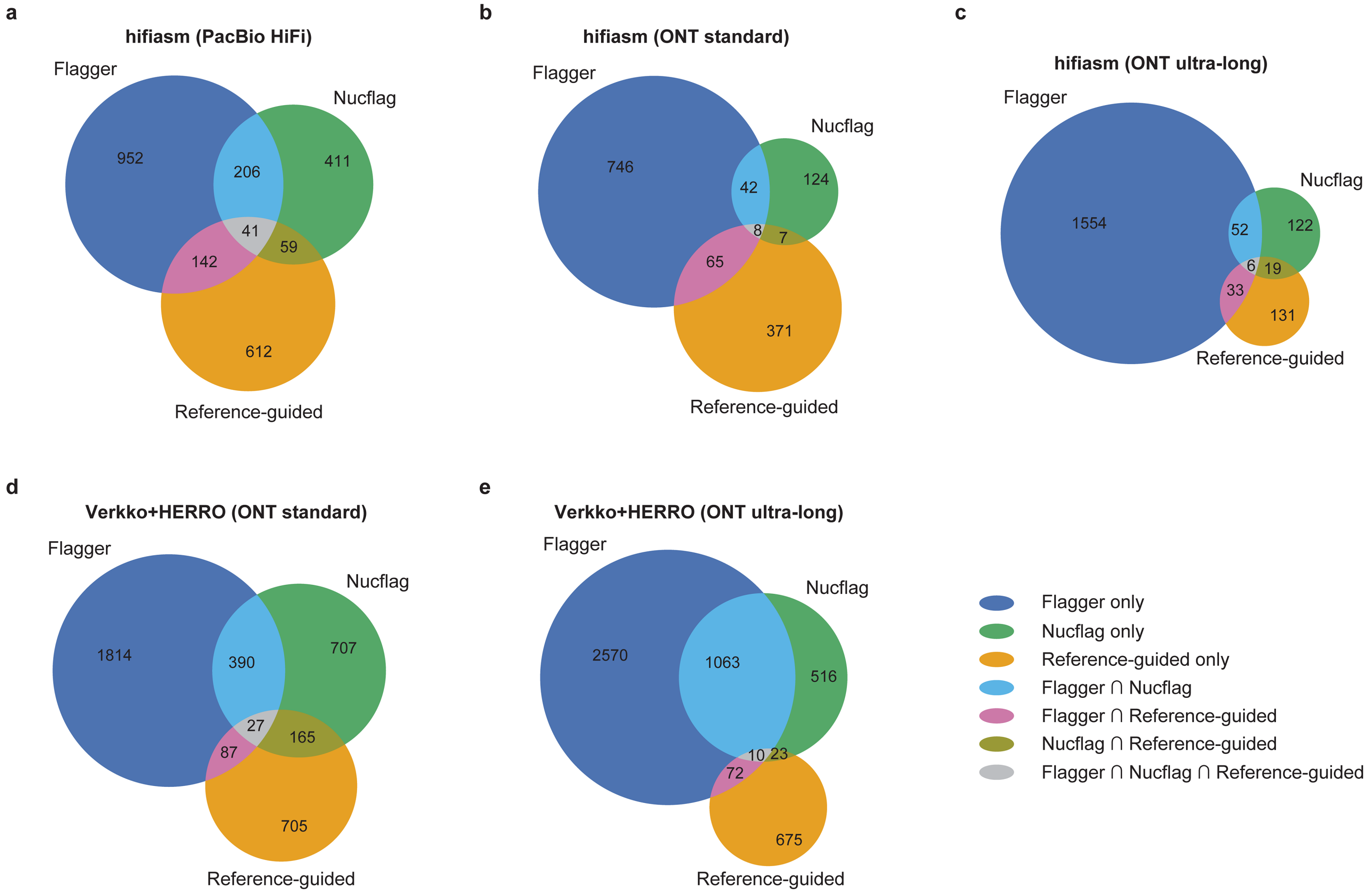
Correspondence and requests for materials should be addressed to Haoyu Cheng or Heng Li.

Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

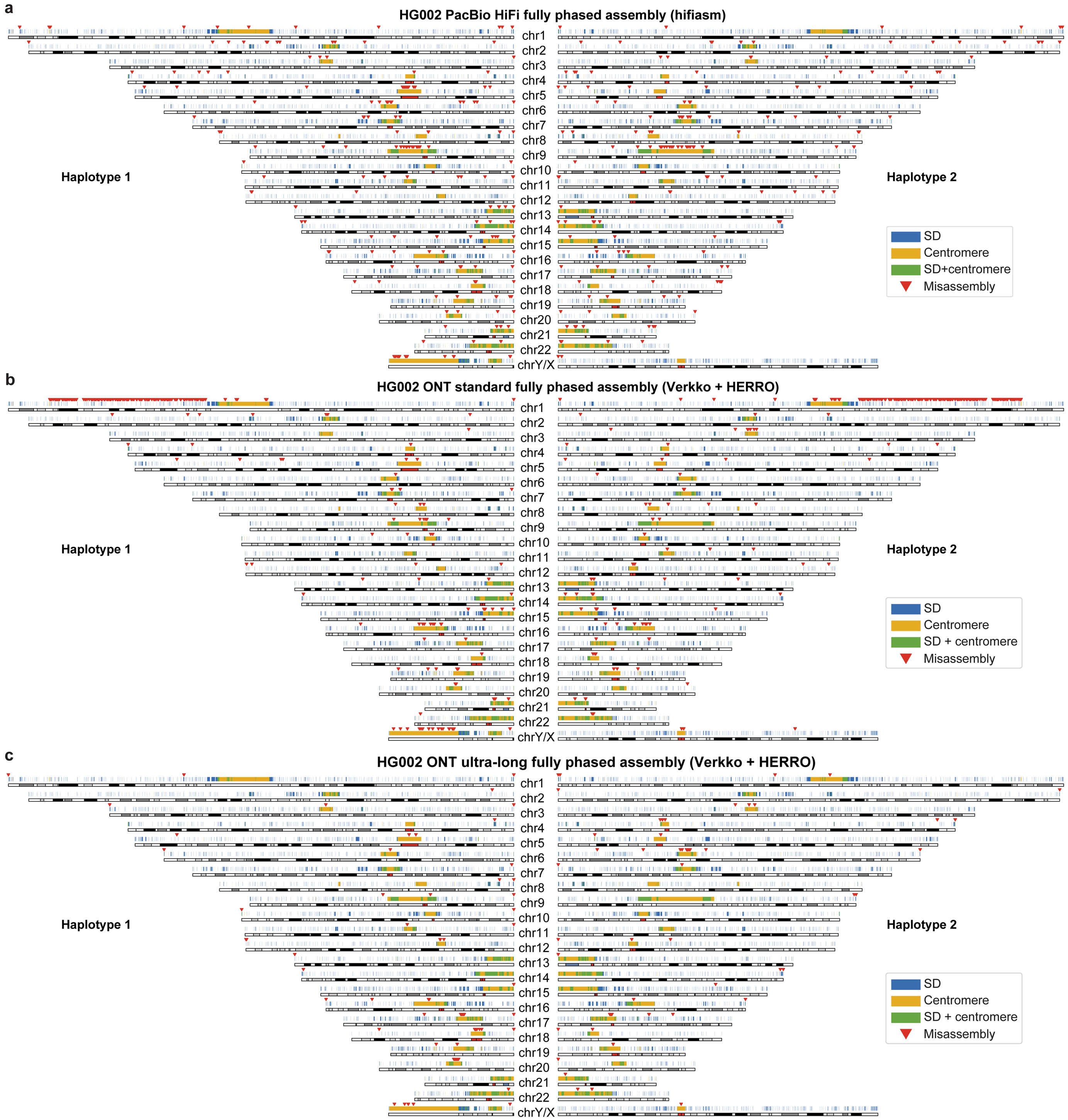
Reprints and permissions information is available at http://www.nature.com/reprints.



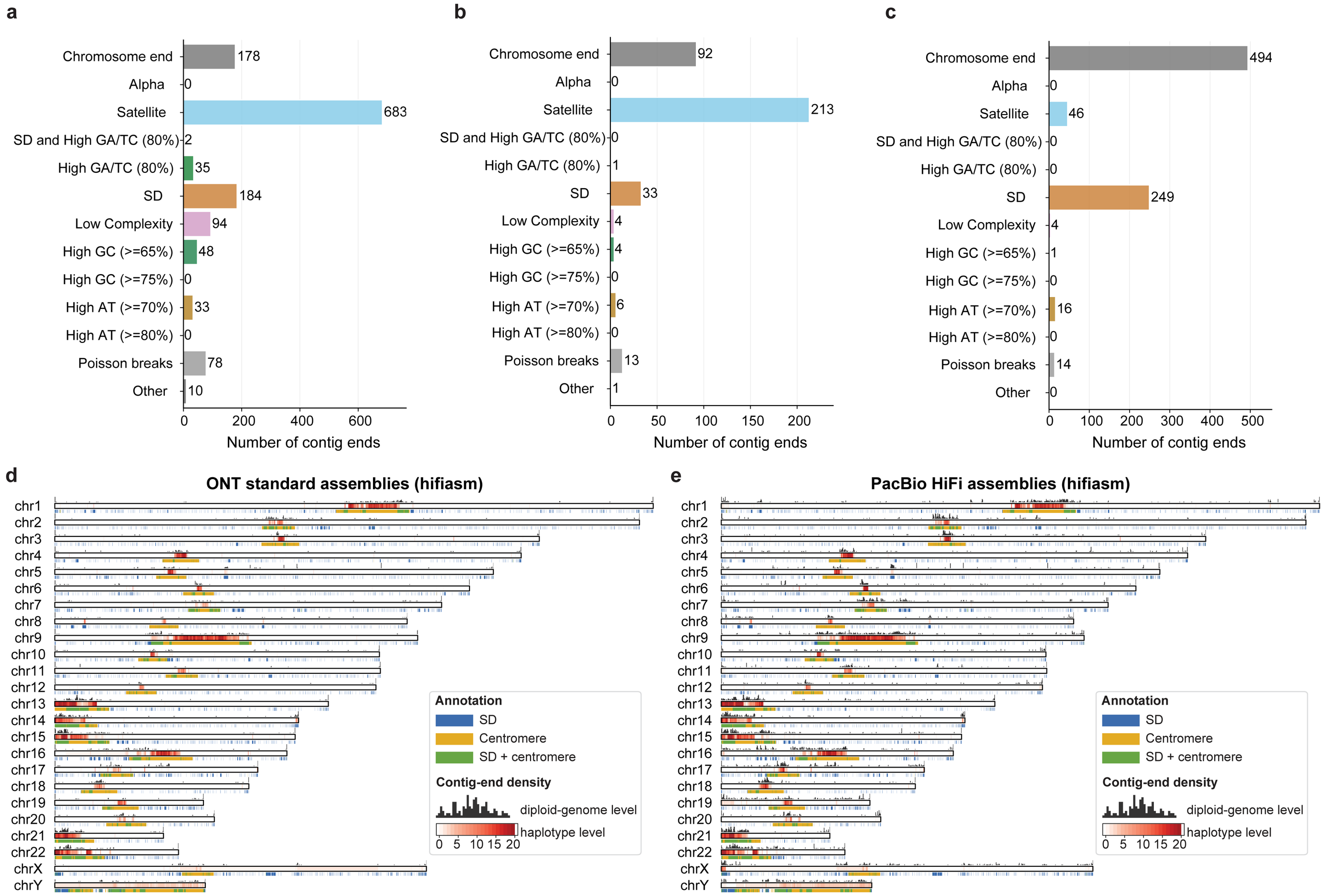
Extended Data Fig. 1 | Comparison of ONT standard simplex and PacBio HiFi data for HG001–HG007. a, Read length distribution for each sample. Blue and orange arrows indicate the N50 values of ONT standard simplex and PacBio HiFi reads, respectively. Read lengths greater than 60 kb are grouped into a single bin. **b**, Nx plot illustrating read length distributions. For each dataset, the Nx plot shows read lengths sorted from longest to shortest, relative to cumulative read length as a percentage of the total yield.



Extended Data Fig. 2 | Number of overlapping misassemblies identified by Flagger, NucFlag and reference-guided analyses in HG002 assemblies. Only misassemblies larger than 50 bp were included. Reference-guided misassemblies were evaluated against the HG002 Q100 ground truth, whereas Flagger and NucFlag represent reference-free methods that align reads back to assemblies. **a**, HG002 PacBio HiFi fully phased assembly generated by hifiasm. **b**, HG002 ONT standard fully phased assembly produced by hifiasm (ONT). **c**, HG002 ONT ultra-long fully phased assembly produced by hifiasm (ONT). **d**, HG002 ONT standard fully phased assembly produced by Verkko + HERRO. **e**, HG002 ONT ultra-long fully phased assembly produced by Verkko + HERRO.



Extended Data Fig. 3 | Genome-wide misassembly distribution in HG002 PacBio HiFi and Verkko + HERRO assemblies. Only misassemblies larger than 50 bp were counted, and all misassemblies were accurately assessed against the HG002 Q100 reference genome. Segmental duplication and centromere annotations are shown in blue and yellow, with overlapping regions highlighted in green and misassembly sites marked with red triangles. **a**, Misassemblies in the HG002 PacBio HiFi fully phased assembly generated by hifiasm. **b**, Misassemblies in the HG002 ONT standard fully phased assembly produced by Verkko + HERRO. **c**, Misassemblies in the HG002 ONT ultra-long fully phased assembly produced by Verkko + HERRO.



Extended Data Fig. 4 | Genome-wide distribution of contig ends representing potential assembly gaps in human genome assemblies generated by hifiasm. Contig ends correspond either to chromosome ends or to unresolved assembly gaps. We aligned contig ends from each assembly to either the HG002 or CHM13 T2T reference genome. For the HG002-specific results (**a**–**c**), evaluations were performed using the HG002 Q100 ground truth. For the results in **d**,**e**, assemblies were aligned to the CHM13 T2T reference genome. **a**–**c**, Number of contig ends in the HG002 phased assemblies produced by hifiasm using PacBio HiFi (**a**), ONT standard (**b**) and ONT ultra-long (**c**) reads. Each contig end was hierarchically classified according to its overlap with annotated sequence features. Poisson breaks refer to isolated contig ends (≤2 ends within a ± 100 kb window), which are likely to represent random assembly breaks. **d**,**e**, Genome-wide distribution of assembly gaps across 10 diploid human genome assemblies (20 haplotypes) generated using ONT standard simplex (**d**) and PacBio HiFi (**e**) reads, including both trio and non-trio assemblies. Chromosomal bars show haplotype-level gap coverage as a red heat map, indicating the number of haplotypes with gaps at each position (max: 20 for autosomes, sex-aware for chrX/Y). Black bars above represent the number of diploid genomes with assembly gaps at each position (max: 10, after diploid-genome-level deduplication). Annotations below indicate segmental duplications (blue), centromeric satellites (yellow), and overlaps (green). More details are provided in Supplementary Information sections 1.13 and 1.14.

Extended Data Table 1 | *k*-mer-based assembly evaluation using yak and Merqury

Data type	Dataset	Assembler	yak			Merqury		
			QV	switch (%)	hamming (%)	QV	switch (%)	hamming (%)
ONT (SUP) standard	HG001	hifiasm (trio)	49.08/49.16	3.38/1.83	4.41/1.21	57.21/57.64	0.33/0.08	0.60/0.10
		Verkko+HERRO (trio)	51.38/51.61	3.46/1.86	4.35/1.27	57.43/57.64	0.29/0.03	0.48/0.10
		hifiasm (dual)	49.12/49.13	/	/	57.24/57.69	/	/
	HG002	hifiasm (trio)	47.54/47.12	1.59/2.52	1.22/1.69	51.50/51.50	0.08/0.11	0.11/0.16
		Verkko+HERRO (trio)	48.97/48.48	1.62/2.61	1.23/1.73	51.67/51.57	0.06/0.07	0.08/0.10
		hifiasm (dual)	47.33/47.29	/	/	51.49/51.64	/	/
	HG003	hifiasm (dual)	48.20/47.69	/	/	64.69/64.87	/	/
	HG004	hifiasm (dual)	48.53/48.61	/	/	64.46/63.13	/	/
	HG005	hifiasm (trio)	48.28/48.42	0.88/2.08	0.57/1.46	49.78/49.89	0.05/0.09	0.06/0.16
		Verkko+HERRO (trio)	50.64/50.68	0.91/2.10	0.56/1.43	49.87/49.93	0.01/0.03	0.02/0.04
		hifiasm (dual)	48.45/48.35	/	/	49.80/49.89	/	/
	HG006	hifiasm (dual)	47.88/47.64	/	/	63.42/61.87	/	/
	HG007	hifiasm (dual)	47.49/47.83	/	/	59.57/61.45	/	/
ONT (SUP) ultra-long	HG002	hifiasm (trio)	46.12/45.66	1.69/2.71	1.26/1.75	50.76/50.44	0.06/0.06	0.07/0.09
		Verkko+HERRO (trio)	47.79/47.37	1.68/2.44	1.25/1.59	51.08/51.03	0.07/0.06	0.07/0.09
	HG02818	hifiasm (trio)	45.57/45.59	3.18/3.16	1.94/2.20	54.42/54.17	0.04/0.05	0.04/0.25
		Verkko+HERRO (trio)	47.46/47.48	3.21/3.12	1.98/2.04	55.10/55.14	0.05/0.06	0.06/0.06
	Arabidopsis	hifiasm	43.93	/	/	41.73	/	/
		Verkko+HERRO	45.30	/	/	33.35	/	/
	Tomato	hifiasm	42.48	/	/	45.54	/	/
		Verkko+HERRO	43.18	/	/	45.40	/	/
	Zebrafish	hifiasm	43.06	/	/	56.95	/	/
		Verkko+HERRO	43.59	/	/	51.05	/	/
HiFi	HG001	hifiasm (trio)	50.21/49.97	2.65/1.14	4.04/0.86	51.86/51.75	0.07/0.08	0.14/0.11
		hifiasm (dual)	50.12/50.02	/	/	51.90/51.78	/	/
	HG002	hifiasm (trio)	53.78/53.00	0.80/0.96	0.77/0.70	51.86/51.75	0.07/0.08	0.14/0.11
		hifiasm (dual)	53.03/53.81	/	/	51.67/51.88	/	/
	HG003	hifiasm (dual)	51.75/52.39	/	/	56.61/62.83	/	/
	HG004	hifiasm (dual)	53.22/54.29	/	/	60.30/62.12	/	/
	HG005	hifiasm (trio)	53.59/53.39	0.29/0.73	0.21/0.62	49.71/49.74	0.03/0.06	0.04/0.10
		hifiasm (dual)	53.48/53.52	/	/	49.63/49.81	/	/
	HG006	hifiasm (dual)	55.56/55.97	/	/	54.53/63.54	/	/
	HG007	hifiasm (dual)	53.42/54.55	/	/	57.44/58.45	/	/
ONT (HAC) standard	HG001	hifiasm (trio)	45.68/45.66	3.94/2.39	4.86/1.62	53.19/53.10	0.41/0.15	0.74/0.23
		hifiasm (dual)	45.62/45.71	/	/	52.98/53.18	/	/
	HG002	hifiasm (trio)	44.43/44.01	1.98/3.36	1.50/2.25	49.21/49.06	0.12/0.18	0.16/0.25
		hifiasm (dual)	44.07/44.35	/	/	49.11/49.13	/	/
	HG003	hifiasm (dual)	45.28/44.92	/	/	56.86/57.68	/	/
	HG004	hifiasm (dual)	45.73/45.78	/	/	57.68/57.68	/	/
	HG005	hifiasm (trio)	45.06/45.15	1.22/2.82	0.83/1.98	48.67/48.68	0.09/0.17	0.15/0.24
		hifiasm (dual)	45.13/45.10	/	/	48.71/48.63	/	/
	HG006	hifiasm (dual)	44.27/44.13	/	/	55.13/55.48	/	/
	HG007	hifiasm (dual)	44.38/44.66	/	/	53.69/55.35	/	/

The phasing switch error rate refers to the proportion of adjacent haplotype-specific marker pairs originating from different haplotypes, and the phasing Hamming error rate represents the percentage of haplotype-specific markers that are incorrectly phased. Only fully phased trio-binning assemblies were evaluated for phasing accuracy. All assemblies were assessed for QV. For *Linum usitatissimum* (flax), we did not perform *k*-mer-based assembly evaluation because short-read data from the same individual are not available.

Extended Data Table 2 | HG002 fully phased assembly variant calling compared with the HG002 benchmark (SUP basecalling model)

Data type	Assembler	Region	SNP (%)			INDEL (%)			ALL (%)		
			Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
ONT standard	hifiasm	homopolymer	96.56	96.00	96.28	47.80	64.55	54.93	74.47	84.46	79.15
		non-homopolymer	97.72	97.04	97.38	94.77	93.57	94.17	97.39	96.69	97.04
	Verkko + HERRO	homopolymer	96.83	96.23	96.53	55.71	65.54	60.23	79.68	84.97	82.24
		non-homopolymer	97.75	97.12	97.43	95.65	94.68	95.16	97.51	96.87	97.19
ONT ultra-long	hifiasm	homopolymer	96.22	95.80	96.01	41.67	61.31	49.61	70.41	83.15	76.25
		non-homopolymer	97.72	97.08	97.40	93.44	93.35	93.39	97.25	96.70	96.98
	Verkko + HERRO	homopolymer	96.59	96.06	96.32	50.96	63.24	56.44	77.01	84.02	80.36
		non-homopolymer	97.70	97.13	97.41	93.80	93.44	93.62	97.27	96.76	97.02

Assemblies were aligned to GRCh38, and variants were called using dipcall[46]. Variant-calling results were evaluated using hap.py.

Extended Data Table 3 | Summary of gene resolution in HG002 fully phased assemblies relative to the HG002 Q100 reference genome

Data	Approach	Contig N50 (Mb)	# Single-copy resolved	# Multicopy resolved	# False duplicated	# Partially resolved	# Unresolved		
							> 50%	10–50%	≤ 10%
HG002 ref	mannually curated	146.8 / 154.3	34114/35754	1203/1276	NA	NA	NA	NA	NA
PacBio HiFi	hifiasm	78.7 / 63.1	33876/35403	1176/1265	32/23	7/5	10/26	5/8	184/289
ONT (SUP)	hifiasm	131.5 / 143.8	34092/35731	1199/1272	8/7	0/0	2/1	0/0	12/15
standard	Verkko + HERRO	52.3 / 46.0	34001/35631	1188/1260	70/74	3/16	6/4	3/1	31/28
ONT (SUP)	hifiasm	146.3 / 143.8	34078/35698	1198/1274	13/34	0/0	3/2	1/0	19/20
ultra-long	Verkko + HERRO	97.5 / 111.0	34090/35733	1197/1274	11/9	0/0	0/2	0/0	13/10

Each assembly consists of two sets of sequences representing the paternal and maternal haplotypes. The two numbers in each cell correspond to the metrics for the two haplotypes, respectively. For each assembly or genome, genes were identified by aligning cDNA sequences with a sequence identity of at least 99%. 'HG002 ref' refers to the T2T Q100 reference genome for HG002, whose aligned genes were used as the ground truth. Gene completeness was evaluated by comparing the genes identified in each assembly with those in the HG002 reference. 'Single-copy resolved' denotes the number of single-copy genes in the HG002 reference that remain single copy in the assembly. 'Multicopy resolved' denotes the number of multicopy genes in the HG002 reference that remain multicopy in the assembly. 'False duplicated' refers to single-copy genes in the HG002 reference that appear as multicopy in the assembly. 'Partially resolved' refers to genes that are fully present but fragmented across multiple pieces in the assembly. 'Unresolved' (>50%, 10–50%, ≤10%) indicates genes that cannot be completely identified in the assembly, in which the aligned length covers more than 50%, between 10% and 50% or less than 10% of the full gene length, respectively.

nature portfolio

Corresponding author(s): Haoyu Cheng

Last updated by author(s): Jan 2, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	□	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software were used for data collection.

Data analysis

The manuscript describes the hifiasm (ONT) assembly algorithm, which is freely available at: https://github.com/chhylp123/hifiasm. The specific version used in this study has been archived in a DOI-minted repository at https://doi.org/10.5281/zenodo.18079611. Assemblies using ONT Simplex and PacBio HiFi reads were generated with hifiasm (version 0.25.0-r726). For comparison, assemblies from ONT Simplex reads were also generated using HERRO (R10.4.1 model, https://github.com/lbcb-sci/herro) and Verkko (version v2.2.1). Seqtk (version 1.4-r130-dirty) was used to downsample HERRO-corrected reads to the target coverage for Verkko input. To identify telomere-to-telomere (T2T) contigs and scaffolds, we used a script provided by the Human Pangenome Reference Consortium (HPRC): https://github.com/biomonika/HPP/blob/main/assembly/wdl/workflows/assessAsemblyCompletness.wdl. K-mer-based assembly quality was evaluated with yak (version 0.1-r62-dirty) and Merqury (version 1.3). Meryl (version 1.4.1) was used together with Merqury to build the k-mer index. For human genome assemblies, gene completeness was assessed by aligning cDNAs to both the reference genome and the assembled contigs using minimap2 (version 2.28-r1209). Gene-level completeness was then evaluated using paftools.js, a utility included in the minimap2 package. The alignments used for SVbyEye visualization (commit fdee406) were also generated with minimap2 (version 2.28-r1209). Immuannot (commit 31362c3) was used to annotate immunological genes. Assembly-based variant calling was performed using Dipcall v0.3 and benchmarked against the GIAB HG002 T2TQ100-v1.0 truth set and its corresponding confident regions on GRCh38. Benchmarking of the resulting variant calls was then carried out using hap.py v0.3.15. VCF files were processed using bcftools (version 1.21) and tabix from SAMtools (version 1.22.1). Flagger v1.1.0 and NucFlag v0.3.6 were used to detect misassemblies in a reference-free manner. To improve alignment accuracy in highly repetitive regions—particularly for reference-based disassembly detection—we employed a hybrid strategy combining MashMap v3.1.3 and Minimap2 v2.30-r1287. Nucleotide composition at each contig end was calculated using bedtools nuc v2.31.0. bedtools was also used to identify assembly gaps. IGV v2.16.0 was used to visualize read and assembly alignments.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Human reference genome: GRCh38, CHM13v2, and HG002 Q100; ONT standard reads of HG001 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG001; ONT standard reads of HG001 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG001; PacBio HiFi reads of HG001: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/HudsonAlpha_PacBio_CCS/; Illumina short reads of HG001 and parents: ERR194147 (HG001), ERR194160 (paternal), ERR194161 (maternal); ONT standard reads of HG002 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG002/PAW70337; ONT standard reads of HG002 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG002/PAW70337; ONT ultra-long reads of HG002: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/5b73fa0e-658a-4248-b2b8-cd16155bc157--UCSC_GIAB_R1041_nanopore/HG002_R1041_UL/dorado/v0.4.0_wMods/*ULCIR*.bam and https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/5b73fa0e-658a-4248-b2b8-cd16155bc157--UCSC_GIAB_R1041_nanopore/HG002_R1041_UL/dorado/v0.4.0_wMods/*ULNEB*.bam; PacBio HiFi reads of HG002: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_HiFi-Revio_20231031/; Illumina short reads of HG002 and parents: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/raw_data/Illumina/; ONT standard reads of HG003 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG003; ONT standard reads of HG003 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG003; PacBio HiFi reads of HG003: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/PacBio_HiFi-Revio_20231031/ and https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/PacBio_CCS_Google_15kb/; Illumina short reads of HG003: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/raw_data/Illumina/parents/HG003/; ONT standard reads of HG004 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG004; ONT standard reads of HG004 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG004; PacBio HiFi reads of HG004: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/PacBio_HiFi-Revio_20231031/ and https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/PacBio_CCS_Google_15kb/; Illumina short reads of HG004: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/raw_data/Illumina/parents/HG004/; ONT standard reads of HG005 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG005; ONT standard reads of HG005 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG005; PacBio HiFi reads of HG005: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG005_NA24631_son/PacBio_CCS_15kb_20kb_chemistry2/uBAMs/; Illumina short reads of HG005: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG005/raw_data/Illumina/; ONT standard reads of HG006 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG006; ONT standard reads of HG006 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG006; PacBio HiFi reads of HG006: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG006_NA24694-huCA017E_father/PacBio_CCS_15kb_20kb_chemistry2/uBAMs/ and https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG006_NA24694-huCA017E_father/PacBio_HiFi_Google/; Illumina short reads of HG006: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG005/raw_data/Illumina/parents/HG006/; ONT standard reads of HG007 (SUP): s3://ont-open-data/giab_2025.01/basecalling/sup/HG007; ONT standard reads of HG007 (HAC): s3://ont-open-data/giab_2025.01/basecalling/hac/HG007; PacBio HiFi reads of HG007: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG007_NA24695-hu38168_mother/PacBio_CCS_15kb_20kb_chemistry2/uBAMs/ and https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG007_NA24695-hu38168_mother/PacBio_HiFi_Google/; Illumina short reads of HG007: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG005/raw_data/Illumina/parents/HG007/; ONT ultra-long reads of HG02818: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG02818/raw_data/nanopore/dorado0.7.2_sup4.1.0_5mCG_5hmCG/; Illumina short reads of HG02818 and parents: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG02818/raw_data/Illumina/; ONT ultra-long reads of Arabidopsis thaliana: SRR29061597; Illumina short reads of Arabidopsis thaliana: https://ngdc.cncb.ac.cn/gsa/browse/CRA005350; ONT ultra-long data of Danio rerio: https://genomeark.s3.amazonaws.com/index.html?prefix=species/Danio_rerio/fDanRer17/genomic_data/ont/pod5/ and https://genomeark.s3.amazonaws.com/index.html?prefix=species/Danio_rerio/fDanRer17/genomic_data/ont/fast5/; Illumina short reads of Danio rerio: https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=1029986; ONT ultra-long data of Solanum lycopersicum: https://obj.umiacs.umd.edu/marbl_publications/duplex/Solanum_lycopersicum_heinz1706/UL/R10.4_40x.noduplex.fastq.gz; Illumina short reads of Solanum lycopersicum: https://ngdc.cncb.ac.cn/gsa/browse/CRA003995/CRX232533; ONT reads of Linum usitatissimum (Flax): SRR31124331; ONT reads used for comparison with Napu (Shasta): https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/Napu_paper_ONT_Coriell_SingleFC_2023/HG0*_R10/reads/*.bam; Napu (Shasta) assemblies: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/Napu_paper_ONT_Coriell_SingleFC_2023/HG0*_R10/assembly/*contigs*fasta; All evaluated assemblies generated by hifiasm and Verkko+HERRO are available at: https://zenodo.org/records/15203417 (human genome assemblies from ONT reads), https://zenodo.org/records/17613526 (non-human genome assemblies from ONT reads), and https://zenodo.org/records/15205178 (assemblies from PacBio HiFi reads). All links and accession identifiers for the sequencing data are also listed in Supplementary Table 5.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not Applicable
Reporting on race, ethnicity, or other socially relevant groupings	Not Applicable
Population characteristics	Not Applicable
Recruitment	Not Applicable
Ethics oversight	Not Applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Does not apply since the study does not include statistical analysis of any hypotheses. All assembly algorithms are deterministic.
Data exclusions	No data were excluded from analysis.
Replication	Not applicable since this study only describes deterministic algorithms without statistic analysis.
Randomization	Does not apply, since this study introduces a method and does not include biological hypotheses analysis. All assembly algorithms are deterministic.
Blinding	Not applicable since this study does not involve statistic analysis and data acquisition.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	Not Applicable
Novel plant genotypes	Not Applicable
Authentication	Not Applicable