

Efficient near telomere-to-telomere assembly of Nanopore Simplex reads

Corresponding Author: Dr Haoyu Cheng

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)

Cheng and Qu et al. present a strategy for achieving near T2T genomes using standard Nanopore data. Their strategy is an update to hifiasm, a well-established tool for long read genome assemblies.

This work is a significant advance for de novo assemblies especially using data that are typically considered production-scale. Prior to this, T2T or alike genome assemblies have required PacBio HiFi data coupled with Nanopore Ultralong reads.

This is a successful attempt at generating high-quality assemblies using Nanopore unsheared DNA based simplex reads, and paved the path for de novo assemblies using standard sequencing that yields 30-60kb read N50s.

I am particularly impressed with the error correction strategy because it is leveraging long reads and accuracy improvements. I think this is a major improvement especially due to its speed and non-reliance on GPU-based compute.

I have a few minor comments:

1. It would be useful to include the ultralong assembly in Table 1 to serve as a contrast for the upper end of assembly contiguity and quality.

2. It might be worth diving into details for the major differences in assembly contiguity and structure for the GIAB samples when comparing the various data types. I think a general audience will appreciate the differences that are expected in performance as well as constraints to be mindful of when using them. For example, the SMN1 and SMN2 analysis was revealing in how UL reads helped resolve it.

3. Have the authors considered a mixed model with low coverage ultralong reads mixed in with standard simplex reads? I suspect this may be an interesting experiment since one could envision generating a mixture of data types in specific use cases.

4. I am curious if there is a function to speculate assembly contiguity based on read lengths? That is, how much of a T2T genome might be missed with a standard simplex based assembly vs a state-of-the-art assembly from HiFi + UL data?

5. Are 50 kb read N50s from unsheared Nanopore sequencing experiments significantly superior for assembly work? For example, the 1000G ONT long read consortium is generating data using unsheared DNA samples and Nanopore ligation-based sequencing. I am curious how much of an assembly boost could these read N50s provide (especially because 10-20% of the data exceed 100 kb read lengths).

I appreciated the analysis of medically-relevant genes, and think that it makes a good case for the success of such an approach as described in this study.

I think this is a major advance for de novo assemblies to be performed at scale for various production-level efforts, and will find adoption in the community.

(Remarks on code availability)

It is easy to install and run.

Referee #2

(Remarks to the Author)

Cheng et al present a new genome assembly pipeline hifiasm (ONT) that can generate T2T assemblies from ONT simplex reads, which are much easier to generate as compared to the standard requirement for T2T assemblies, ONT ultralong reads. The manuscript starts with an interesting introduction describing actual problems in the field of genome assembly from a user perspective.

The introduction of an assembly tool working only on ONT simplex reads is very timely as the generation of UL reads of 100+kb sizes is very complicated (or even impossible) in many non-animal/non-human species. To achieve similar results new algorithms that do not rely on UL reads are necessary.

Hifiasm (ONT) produces assemblies which are nearly as good as PacBio assemblies. The reduced quality (mainly due to false reconstructed homopolymers) might be acceptable considering the lower effort need to generate ONT simplex reads. With this hifiasm (ONT) enables T2T assembly on a population scale.

It is fair to assume that hifiasm (ONT) continues to be a widely used tool. The introduction of high quality assemblies from simplex reads will reduce the costs and efforts needed for T2T assemblies. No doubt this paper will be highly cited. The only main criticism I have, the actual advancement is limited to the error correction step.

Individual comments:

As a non-human researcher, I appreciate the analysis of the assembly qualities of other genomes than human as hifiasm is also the tool to use in our community. However, it is not clear, why these are done with Ultralong reads. Isn't the point of hifiasm (ONT) to be able to assemble genomes with simplex reads? Those or similar assemblies should be shown with simplex reads as well. Are the UL simplex reads, filtered simplex reads? It is not clear, to what read set the coverage values relate to. In case such a filtering was done, how did it affect the different samples? Was there different abundances of the UL simplex reads in the different samples? What would be the strategy to generate the required amount of UL simplex reads?

Is there a methodological difference in species with different ploidy? Specifically, I am wondering about inbred genomes, where only a single haplotype needs to be assembled, are those assemblies more error-prone due to the lack of linked variation that can help revealing (recurrent) sequencing errors?

Besides, during error correction, how are near-by errors distinguished from nearby polymorphisms for error correction?

The analysis of a representative duplicated gene pair is not sufficient to conclude on one algorithm outperforming the other ones. The analysis of more/all duplicated genes should be shown.

Minor:

Line 60. Coverage per haplotype or per genome?

Line 67 & 125 & 154. I am not on top of the terminology here. Are "Simplex Ultra long" reads and "Ultra long" reads the same? Or are they filtered simplex reads? Perhaps this should be clarified. Are "standard" reads equal to "simplex" reads?

(Remarks on code availability)

Referee #3

(Remarks to the Author)

This paper describes an extension of a very popular hifiasm assembler to simplex ONT data. The main novelty is the new algorithm for error correction of ONT reads that is robust to systematic errors. An alternative method for self error correction - HERRO - requires 10x more compute time and expensive GPU resources. Hifiasm ONT-only assemblies are similar to or better than the current best hybrid strategy employed by large assembly projects such as HPRC (using HiFi / Ultra-long ONT and/or Hi-C). The proposed hifiasm ONT version will likely have a great impact, and be adopted by many ongoing and future sequencing projects that rely on ONT simplex data. I am overall highly enthusiastic about this paper, but request some clarifications below.

Results:

1. Although assembly contiguity and QV suggest high quality of hifiasm contigs, direct analysis of misassemblies is needed,

particularly for the comparison with PacBio-based assemblies. In particular, quality of the repetitive sequence (centromeres and SegDups) need to be benchmarked - not necessarily for all assemblies, but at least for a subset where high-quality orthogonal benchmarks are available.

2. De novo diploid assemblies are expected to improve SV detection. Perhaps it makes sense to generate assembly-based SV calls and evaluate against the HG002 SV benchmark, to confirm that all heterozygous SVs are preserved in the corrected reads?

3. Did the authors directly measure corrected read error rates? Are systematic errors also corrected or masked? Do the authors anticipate that error-corrected reads could be used for other applications (such as reference variant calling)?

4. HERRO-based assemblies seem to have slightly higher QV scores, but could this be because of regions that were uniquely assembled by hifiasm? Maybe restricting the comparison to the same regions that were assembled by both methods would help answer this?

Methods:

5. Does the algorithm implicitly assume that all haplotype copies are balanced (in either diploid or repetitive regions)? Can error correction be potentially used for metagenome or cancer assembly? If the answer is “no”, a simple clarification is sufficient.

6. Does all-vs-all read alignment remain a computational bottleneck, compared to the error correction and the rest of the assembly? Did the authors make any improvements to take advantage of more accurate iterations of ONT data?

7. Will error correction work for R9 or earlier R10 data with previous basecaller versions? Can the authors provide minimum recommendations? Is the algorithm compatible with data from other large-scale ONT sequencing projects, such as G1k resequencing efforts or long-read CARD initiative?

8. I understand the main principle of the DP approach to compute the maximum sequence of compatible sites. However I was unable to find information on how given two sites, their pairwise compatibility is decided ($S[j] \leftrightarrow S[i]$)? This seems to be an important practical detail.

Misc:

9. It would be helpful to have a supplementary table that lists all data accession IDs.

(Remarks on code availability)

The core repository seems to be in a good shape, and I previously used hifiasm in other modes.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

Dear authors,

Thank you for addressing my comments. I don't have anything further.

All the best

(Remarks on code availability)

Easy to use, as is the case with other tools developed by Heng Li

Referee #2

(Remarks to the Author)

I am satisfied with all responses the authors gave to my questions. Congratulations to the authors.

(Remarks on code availability)

Referee #3

(Remarks to the Author)

I thank the authors for answering all my requests. I am highly enthusiastic about this work.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Editor:

Q1.0.1: We would be interested in seeing whether hifiasm (ONT) can be applied to more complex contexts (cf. comments made by reviewer #3) such as metagenomics or cancer genomes.

A1.0.1: The error correction algorithm in hifiasm (ONT) does not have the diploid genome assumption, making it work well for resolving highly repetitive repeats like centromeres including large numbers of highly similar repeat copies. For assembly algorithms, each repeat copy is essentially similar to a haplotype in a polyploid genome. We have already demonstrated the power of hifiasm (ONT) in resolving highly repetitive regions of human and non-human genomes in our manuscript.

Assembling cancer genomes presents challenges similar to assembling highly repetitive regions due to the uncertainty of ploidy levels. We generated de novo assemblies for two tumor cell lines—HG008 (pancreatic cancer) and COLO829 (metastatic melanoma)—using ONT standard Simplex reads (see Table r1). hifiasm (ONT) achieved significantly better contiguity than Verkko, as reflected by higher assembly N50 values and more T2T contigs. Because the ploidy and chromosome structures of cancer cell lines are often uncertain, direct quality evaluation remains difficult. Nevertheless, compared with Verkko, hifiasm (ONT) produced larger assemblies with substantially higher contiguity and far fewer contigs, suggesting that it handles polyploid or multi-haplotype genomes more effectively. These results indicate that hifiasm (ONT) currently represents the most capable assembler for cancer genome reconstruction.

Table r1. Statistics of cancer cell line (tumor) assemblies using ONT standard Simplex reads (SUP basecalling model)

Sample	Read N50 (kb)	Approach	Assembly size (Gb)	# Contigs	T2T count		Assembly N50 (Mb)	
					Scaffold	Contig	Scaffold	Contig
HG008 (45x)	36.63	hifiasm	2.99/1.68	82/80	6/3	5/2	100.82/89.88	93.52/55.11
		Verkko	1.49/2.83	1,749/3,920	1/2	0/0	39.32/78.35	10.34/22.78
COLO829 (49x)	26.52	hifiasm	1.95/3.09	148/115	1/9	0/7	97.85/132.26	62.47/96.08
		Verkko	2.12/2.4	1,564/1,113	1/4	0/0	60.03/82.78	20.50/22.39

Each assembly consists of two sets of sequences representing the paternal and maternal haplotypes. The two numbers in each cell give the metrics for the two sets of sequences, respectively. In addition to ONT reads, Hi-C data were also incorporated for phasing.

Compared with single-sample genome assembly, metagenomic assembly has distinct requirements. Although metagenomic assemblers must handle large numbers of microbial

genomes, these genomes are typically small and contain relatively few repetitive sequences. Therefore, the error-correction algorithms of hifiasm (ONT) remain applicable, but specialized downstream assembly algorithms are needed to fully resolve metagenomic data. To demonstrate this, we used hifiasm (ONT) to correct raw metagenomic ONT reads and then input the corrected reads into hifiasm-meta, which was originally developed for PacBio HiFi data. As shown in Figure r1, hifiasm achieved performance comparable to metaFlye, even though it is not optimized for metagenomic assembly, but remained less effective than myloasm and nanoMDBG, which are specifically designed for ONT R10 metagenomic data. These results indicate that while hifiasm (ONT) performs robust error correction on metagenomic reads, achieving optimal assembly quality requires metagenome-specific downstream algorithms.

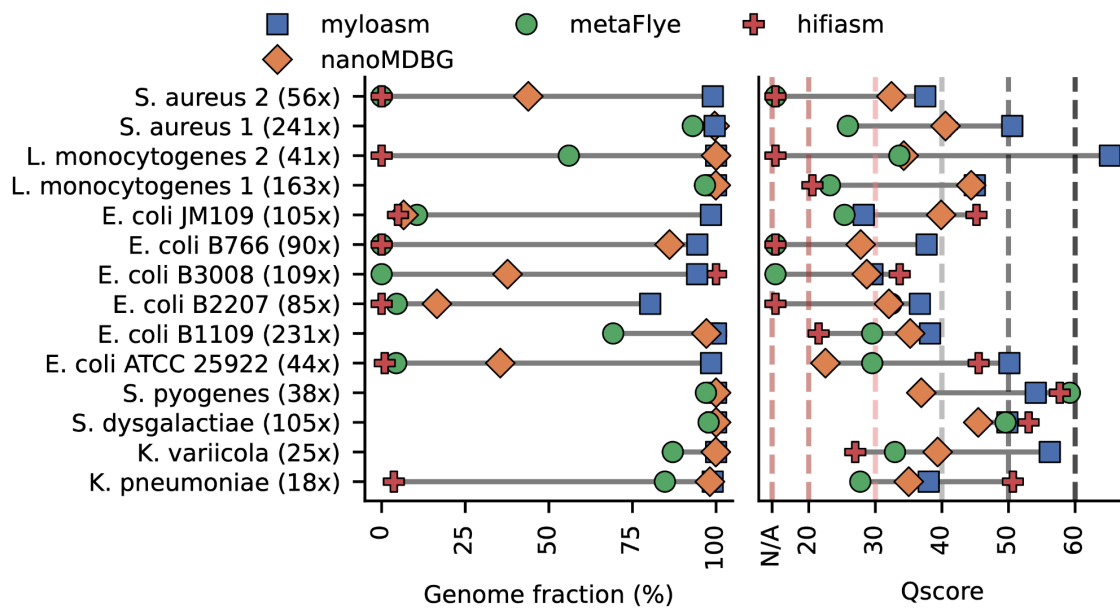


Figure r1. Metagenomic assembly results using ONT R10 Simplex reads. Hifiasm represents assemblies generated by hifiasm-meta using corrected reads produced by hifiasm (ONT). metaFlye is a general-purpose metagenomic assembler without specific optimization for ONT R10 data, while myloasm and nanoMDBG are metagenomic assemblers specifically designed for ONT R10 reads.

In summary, because the error-correction module in hifiasm (ONT) does not rely on a diploid genome assumption, it can effectively handle complex genomes such as cancer and metagenomic samples. However, achieving optimal performance for these challenging assemblies requires further optimization in the downstream assembly stage. We thus did not present these results in the manuscript. In future work, we will focus on these improvements, particularly for cancer genome assembly.

Referees' comments:

Referee #1 (Remarks to the Author):

Cheng and Qu et al. present a strategy for achieving near T2T genomes using standard Nanopore data. Their strategy is an update to hifiasm, a well-established tool for long read genome assemblies.

This work is a significant advance for de novo assemblies especially using data that are typically considered production-scale. Prior to this, T2T or alike genome assemblies have required PacBio HiFi data coupled with Nanopore Ultralong reads.

This is a successful attempt at generating high-quality assemblies using Nanopore unshread DNA based simplex reads, and paved the path for de novo assemblies using standard sequencing that yields 30-60kb read N50s.

I am particularly impressed with the error correction strategy because it is leveraging long reads and accuracy improvements. I think this is a major improvement especially due to its speed and non-reliance on GPU-based compute.

I have a few minor comments:

Q1.1.1: It would be useful to include the ultralong assembly in Table 1 to serve as a contrast for the upper end of assembly contiguity and quality.

A1.1.1: We thank the reviewer for this valuable suggestion and agree that it highlights an important point. To make this clearer for the readers, we compared assembly contiguity and error rates in Figure 4a using different data types. The results clearly show the trend that ultra-long > ONT standard > PacBio HiFi in both contiguity and misassembly metrics. We believe this figure provides a clearer and more intuitive comparison for the audience.

Q1.1.2: It might be worth diving into details for the major differences in assembly contiguity and structure for the GIAB samples when comparing the various data types. I think a general audience will appreciate the differences that are expected in performance as well as constraints to be mindful of when using them. For example, the SMN1 and SMN2 analysis was revealing in how UL reads helped resolve it.

A1.1.2: As noted by the reviewer, providing a more in-depth analysis of assembly structure would be valuable for the audience. Accordingly, we have incorporated the following results into the revised manuscript:

- a) Figure 4: Distribution of misassemblies across all GIAB samples.
- b) Figure 5: Genomic regions that remain unassembled (contig ends).

- c) Supplementary Table 7: Gene-level resolution in HG002 assemblies, using the HG002 Q100 T2T reference as the ground truth.
 - d) Supplementary Table 5, Supplementary Figure 4 and 5: HG002-specific misassemblies
- Together with Table 1, these additions provide a comprehensive overview of the strengths and limitations of hifiasm (ONT).

Q1.1.3: Have the authors considered a mixed model with low coverage ultralong reads mixed in with standard simplex reads? I suspect this may be an interesting experiment since one could envision generating a mixture of data types in specific use cases.

A1.1.3: As shown in Table r2, we compared HG002 assemblies generated by mixing ONT standard and ultra-long reads in different proportions. Incorporating ultra-long reads consistently improved assembly quality. Notably, the Mix 4 dataset (47x standard + 40x ultra-long) produced a higher-quality assembly than the ultra-long-only dataset, even though the latter had a much higher read N50 (115.99 kb vs. 45.74 kb). In contrast, adding very low coverage (10x) ultra-long reads (Mix 1) resulted in reduced assembly quality. This decline is likely due to the contained read problem in the assembly algorithm, which often arises from uneven read length distributions. In future work, we aim to address this problem to further reduce the ultra-long read coverage required for optimal assembly.

Table r2. Statistics of HG002 trio-binning assemblies combining ONT standard and ultra-long Simplex reads using hifiasm

Dataset	Coverage			Read N50 (kb)	T2T count		Assembly N50 (Mb)	
	Standard	Ultra-long	Total		Scaffold	Contig	Scaffold	Contig
Mix 1	47x	10x	57x	32.33	4/3	3/1	52.69/60.48	33.53/35.16
Mix 2	47x	20x	67x	35.61	18/16	12/9	147.02/154.38	123.95/135.23
Mix 3	47x	30x	77x	39.68	17/20	16/12	146.78/154.38	135.71/135.23
Mix 4	47x	40x	87x	45.74	20/21	16/16	146.78/154.38	133.57/135.23
Standard-only	47x	0x	47x	29.40	15/17	7/15	146.87/143.80	131.52/143.80
Ultra-long-only	0x	68x	68x	115.99	22/19	19/14	146.77/154.38	146.34/143.79

Each assembly consists of two sets of sequences representing the paternal and maternal haplotypes. The two numbers in each cell give the metrics for the two sets of sequences, respectively. All mixed datasets include the full set of ONT standard reads together with the longest 10x, 20x, 30x, and 40x ultra-long reads for Mix 1, Mix 2, Mix 3, and Mix 4, respectively. For Mix 3 and Mix 4, the parameter `--chem-c` was set to 3 and 4, respectively, while the other assemblies were generated using default settings. This adjustment was made because Mix 3 and Mix 4 have higher coverage, and hifiasm automatically increases the chimeric read detection threshold at higher coverages to reduce assembly errors. All reads were basecalled using the SUP basecalling model.

Q1.1.4: I am curious if there is a function to speculate assembly contiguity based on read lengths? That is, how much of a T2T genome might be missed with a standard simplex based assembly vs a state-of-the-art assembly from HiFi + UL data?

A1.1.4: As shown in Table r2, incorporating more ultra-long reads improves read N50 and generally enhances assembly contiguity. However, higher read N50 does not always result in better contiguity, as seen when comparing Mix 4 with the ultra-long-only dataset in Table r2. We believe that final assembly quality is influenced by multiple factors, including base accuracy, read length, coverage and sequencing biases (e.g. PacBio HiFi is biased against GA-rich regions). In addition to read length, our observations indicate that read coverage plays a critical role in assembly contiguity. The current version of hifiasm (ONT) requires approximately 40–50× ONT Simplex standard reads to achieve near T2T assemblies of human genomes. In future work, we plan to further optimize hifiasm (ONT) to reduce this coverage requirement.

Q1.1.5: Are 50 kb read N50s from unsheared Nanopore sequencing experiments significantly superior for assembly work? For example, the 1000G ONT long read consortium is generating data using unsheared DNA samples and Nanopore ligation-based sequencing. I am curious how much of an assembly boost could these read N50s provide (especially because 10-20% of the data exceed 100 kb read lengths).

A1.1.5: It is difficult to obtain ONT standard Simplex reads with a 50 kb read N50. While the 1000 Genomes ONT Consortium has generated data of this scale, the currently published datasets are based on ONT R9 chemistry. However, next-generation ONT assemblers such as hifiasm (ONT) and HERRO require ONT R10 data to achieve high-quality assemblies. To address this, we simulated such data by mixing ONT standard and ultra-long reads. As shown in Table r2, adding ultra-long reads significantly improved assembly quality. Notably, the Mix 4 dataset (47x standard + 40x ultra-long) achieved higher assembly quality than the ultra-long-only dataset, even though the latter had a much greater read N50 (115.99 kb vs. 45.74 kb). This improvement is likely due to the higher overall coverage in Mix 4, indicating that read coverage plays a key role in the current version of hifiasm (ONT). In future work, we aim to further optimize the algorithm to reduce the read coverage requirement.

Q1.1.6: I appreciated the analysis of medically-relevant genes, and think that it makes a good case for the success of such an approach as described in this study.

I think this is a major advance for de novo assemblies to be performed at scale for various production-level efforts, and will find adoption in the community.

A1.1.6: We thank the reviewer for identifying our strength. In the revised manuscript, we have also added Supplementary Table 7, which evaluates all genes in the HG002 assemblies using the HG002 Q100 T2T reference as the ground truth. This clearly shows that hifiasm (ONT) reduces the number of unresolved genes by an order of magnitude compared to the HiFi assembly. Most of them are challenging medically relevant loci such as *SMN1* and *SMN2*.

Q1.1.7: Referee #1 (Remarks on code availability):

It is easy to install and run.

A1.1.7: We thank the reviewer for identifying our strength.

Referee #2 (Remarks to the Author):

Cheng et al present a new genome assembly pipeline hifiasm (ONT) that can generate T2T assemblies from ONT simplex reads, which are much easier to generate as compared to the standard requirement for T2T assemblies, ONT ultralong reads. The manuscript starts with an interesting introduction describing actual problems in the field of genome assembly from a user perspective.

The introduction of an assembly tool working only on ONT simplex reads is very timely as the generation of UL reads of 100+kb sizes is very complicated (or even impossible) in many non-animal/non-human species. To achieve similar results new algorithms that do not rely on UL reads are necessary.

Hifiasm (ONT) produces assemblies which are nearly as good as PacBio assemblies. The reduced quality (mainly due to false reconstructed homopolymers) might be acceptable considering the lower effort needed to generate ONT simplex reads. With this hifiasm (ONT) enables T2T assembly on a population scale.

It is fair to assume that hifiasm (ONT) continues to be a widely used tool. The introduction of high quality assemblies from simplex reads will reduce the costs and efforts needed for T2T assemblies. No doubt this paper will be highly cited. The only main criticism I have, the actual advancement is limited to the error correction step.

Individual comments:

Q1.2.1: As a non-human researcher, I appreciate the analysis of the assembly qualities of other genomes than human as hifiasm is also the tool to use in our community. However, it is not clear, why these are done with Ultralong reads. Isn't the point of hifiasm (ONT) to be able to assemble genomes with simplex reads? Those or similar assemblies should be shown with simplex reads as well. Are the UL simplex reads, filtered simplex reads? It is not clear, to what read set the coverage values relate to. In case such a filtering was done, how did it affect the different samples? Was there different abundances of the UL simplex reads in the different samples? What would be the strategy to generate the required amount of UL simplex reads?

A1.2.1: For non-human genomes, hifiasm (ONT) performs well with ONT standard (non-ultra-long) Simplex reads. However, few publicly available ONT R10 Simplex datasets that are not ultra-long currently exist, and both hifiasm (ONT) and HERRO require ONT R10 data.

This scarcity largely reflects recent trends—over the past few years, most high-quality, haplotype-resolved assemblers have been developed for PacBio HiFi data supplemented with ONT ultra-long reads. Thus few projects have generated standard ONT R10 Simplex datasets.

In the revised manuscript, to further demonstrate the effectiveness of hifiasm (ONT), we assembled a flax genome using ONT standard Simplex R10 reads. As shown in Supplementary Table 4, hifiasm (ONT) produced a more contiguous assembly than Verkko + HERRO on this dataset.

The Darwin Tree of Life (DToL) project also rapidly adopted hifiasm (ONT) for assembling samples sequenced with ONT standard Simplex data. For example, DToL internally used hifiasm (ONT) to assemble three cichlid fish genomes. These genomes are challenging to assemble with PacBio HiFi reads, as HiFi sequencing often shows coverage dropouts in fish genomes, resulting in low contig N50 values (~1 Mb). In contrast, using hifiasm (ONT) with ONT standard Simplex reads, the assemblies achieved contig N50s of ~40 Mb, with 4, 8, and 6 chromosomes assembled to T2T level, respectively. As these data are unpublished, the results are not included in this manuscript.

Q1.2.2: Is there a methodological difference in species with different ploidy? Specifically, I am wondering about inbred genomes, where only a single haplotype needs to be assembled, are those assemblies more error-prone due to the lack of linked variation that can help revealing (recurrent) sequencing errors?

A1.2.2: hifiasm (ONT) utilizes long-range information across variants for phasing and error correction without assuming a diploid genome. Therefore, no specific adjustments are required for species with different ploidy levels (see Figure 1d in the main manuscript). As shown in Section A1.0.1 and Table r1, hifiasm (ONT) can effectively assemble cancer genomes with unknown or variable ploidy.

For inbred or nearly homozygous genomes, the error-correction algorithm within hifiasm (ONT) also performs well, as illustrated by the following scenarios:

- a) Non-repetitive regions: Only sequencing errors are present, and these errors are generally incompatible with one another (see Figure 1 in the main manuscript). hifiasm (ONT) correctly identifies such inconsistencies as sequencing errors rather than true variants and corrects them accordingly.
- b) Repetitive regions: From an assembly perspective, each repeat copy behaves similarly to a haplotype in a polyploid genome. Figure 1d in the main manuscript provides an example where hifiasm (ONT) successfully resolves such complex regions.

We have demonstrated the capability of hifiasm (ONT) in resolving inbred or nearly homozygous genomes, including *Arabidopsis thaliana* (Arabidopsis), *Danio rerio* (zebrafish), and *Solanum lycopersicum* (tomato). In the revised manuscript, we also incorporated an additional nearly homozygous sample, *Linum usitatissimum* L. (flax), for evaluation.

Q1.2.3: Besides, during error correction, how are near-by errors distinguished from nearby polymorphisms for error correction?

A1.2.3: hifiasm (ONT) leverages the consistency among multiple nearby variants to distinguish true variants from sequencing errors, regardless of how closely they occur (see Figure 1 in the main manuscript). As long as sequencing errors do not exactly overlap with true variants, hifiasm (ONT) can correctly identify and use the variant information.

In essence, hifiasm corrects sequencing errors using reads from the same haplotype. Most errors are not phased with real polymorphisms and can therefore be detected and removed. Even in the rare instance where an error overlaps a SNP, such an error typically does not recur across reads and can still be corrected. We have now added a discussion of this point in the Methods section of the main manuscript.

Q1.2.4: The analysis of a representative duplicated gene pair is not sufficient to conclude one algorithm outperformed the other ones. The analysis of more/all duplicated genes should be shown.

A1.2.4: We thank the reviewer for the suggestion. In the revised manuscript, we further evaluated the completeness of all genes across the HG002 assemblies (Supplementary Table 7). Because HG002 has a Q100 T2T reference genome, we used the aligned genes in the Q100 reference as the ground truth. The gene completeness of each HG002 assembly was assessed by comparing its aligned genes to those in the reference. As shown in Supplementary Table 7, the ONT assemblies generated by hifiasm (ONT) contain only 20–40 genes that remain unresolved when using either ONT standard or ultra-long reads. In comparison, the HiFi assemblies contain approximately an order of magnitude more unresolved genes. This observation is consistent with Figure 3 in our manuscript.

In addition, to better evaluate the assemblies' ability to resolve complex genomic regions, we have incorporated the following new results in the revised manuscript:

- a) Figure 4 (main manuscript): Distribution of misassemblies across all GIAB samples.
- b) Figure 5 (main manuscript): Genomic regions that remain unassembled (contig ends).
- c) Supplementary Fig. 3 (supplementary material): Number of annotated immunological genes per human genome assembly. These genes, including HLA and KIR loci, are typically repetitive and multi-allelic, making them difficult to assemble. For a given assembly, a gene may be unreported if its sequence diverges substantially from the reference.

Minor:

Q1.2.5: Line 60. Coverage per haplotype or per genome?

A1.2.5: We now clarify in the text that the coverage is reported per genome rather than per haplotype.

Q1.2.6: Line 67 & 125 & 154. I am not on top of the terminology here. Are “Simplex Ultra long” reads and “Ultra long” reads the same? Or are they filtered simplex reads? Perhaps this should be clarified. Are “standard” reads equal to “simplex” reads?

A1.2.6: We thank the reviewer for pointing out this lack of clarity. Compared with PacBio, ONT offers a wider range of sequencing protocols. Based on read length, ONT libraries can be classified as:

- a) Standard: shorter read length, low DNA input requirement.
- b) Ultra-long: longer read length, high DNA input requirement, typically work with cell lines.

In terms of base accuracy, ONT distinguishes between:

- a) Simplex: lower base accuracy with recurrent sequencing errors but high yield.
- b) Duplex: higher base accuracy with minimal recurrent errors but lower scalability; currently discontinued by ONT.

In theory, these categories can be combined to produce four sequencing types: Simplex standard, Simplex ultra-long, Duplex standard, and Duplex ultra-long. However, in practice, Duplex ultra-long libraries are nearly impossible to produce. Moreover, ONT has recently decided to discontinue support for Duplex sequencing; thus, only Simplex standard and Simplex ultra-long are currently available.

In our study, we originally used the term “Simplex ultra-long” to distinguish these reads from Duplex reads. However, since Duplex sequencing is no longer supported by ONT, we have replaced “Simplex ultra-long” with “ultra-long” to avoid confusion. In practice, “Simplex” reads are often referred to simply as “standard” ONT reads. This clarification has been added to the revised text.

Referee #3 (Remarks to the Author):

This paper describes an extension of a very popular hifiasm assembler to simplex ONT data. The main novelty is the new algorithm for error correction of ONT reads that is robust to systematic errors. An alternative method for self error correction - HERRO - requires 10x more compute time and expensive GPU resources. Hifiasm ONT-only assemblies are similar to or better than the current best hybrid strategy employed by large assembly projects such as HPRC (using HiFi / Ultra-long ONT and/or Hi-C). The proposed hifiasm ONT version will likely have a great impact, and be adopted by many ongoing and future sequencing projects that rely on ONT simplex data. I am overall highly enthusiastic about this paper, but request some clarifications below.

Results:

Q1.3.1: Although assembly contiguity and QV suggest high quality of hifiasm contigs, direct analysis of misassemblies is needed, particularly for the comparison with PacBio-based

assemblies. In particular, quality of the repetitive sequence (centromeres and SegDups) need to be benchmarked - not necessarily for all assemblies, but at least for a subset where high-quality orthogonal benchmarks are available.

A1.3.1: We thank the reviewer for this valuable suggestion. In the revised manuscript, we have added Figure 4 (main manuscript), which illustrates the distribution of misassemblies across all GIAB human samples. For HG002, where a high-quality ground truth is available, we directly compared each assembly to the HG002 reference genome. For other samples lacking ground truth, we evaluated misassemblies using the reference-free methods Flagger and NucFlag by aligning sequencing reads back to the assemblies. The results demonstrate that hifiasm outperforms Verkko, and that ONT-based assemblies generally achieve higher overall quality than HiFi-based assemblies. Most misassemblies are located around centromeric and SegDup regions.

Q1.3.2: De novo diploid assemblies are expected to improve SV detection. Perhaps it makes sense to generate assembly-based SV calls and evaluate against the HG002 SV benchmark, to confirm that all heterozygous SVs are preserved in the corrected reads?

A1.3.2: We agree that for the HG002 assembly, it is important to compare the results against the high-quality ground truth. Instead of comparing assembly-based SV calls with the benchmark SV set, we chose to directly align the assemblies to the HG002 ground truth reference and identify structural variants (≥ 50 bp) that represent misassemblies. This approach provides a more direct and unbiased assessment, avoiding potential artifacts introduced by specific SV callers. Moreover, it offers a more comprehensive evaluation than comparing SV callsets, since even in the current HG002 benchmark, defining and representing SVs within highly repetitive regions (i.e., non-confident regions) such as centromeres remains challenging. Direct alignment of assemblies allows more accurate evaluation in these difficult regions. In addition to large SVs (≥ 50 bp), smaller variants may also reflect local misassemblies, but such errors can typically be corrected using post-processing tools such as assembly polishers. As such we only focus on SVs.

As shown in Table r3 (and incorporated into Figure 4 and Supplementary Table 5 in the revised manuscript), only hundreds of SVs (i.e., misassemblies ≥ 50 bp) were detected in each assembly. For instance, hifiasm with ONT ultra-long reads identified only 184 SVs, which were further reduced to 135 after merging nearby SVs within 10 kb. In contrast, the latest HG002 SV benchmark reports 28,188 SVs across 2.76 Gb of confident regions. Notably, most SVs or misassemblies identified in hifiasm and Verkko assemblies are located around centromeric regions that fall outside these 2.76Gb confident regions. Thus, we believe that hifiasm (ONT) successfully preserves nearly all true SVs while maintaining high assembly accuracy.

For HG002 specifically, we have also added gene-level resolution analysis, which clearly shows that hifiasm (ONT) performs well in preserving gene content (Supplementary Table 7).

Table r3. Structural variant statistics from HG002 phased assemblies compared to the HG002 Q100 reference genome

Data	Approach	Contig N50 (Mb)	SV total length (MB)	# SV (no merge)	# SV (merge by distance)		
					1,000bp	5,000bp	10,000bp
ONT ultra-long	hifiasm	146.34/143.79	0.4	184	160	146	135
	Verkko	97.49/111.04	1.5	751	679	399	269
ONT standard	hifiasm	131.52/143.80	1.9	402	353	246	191
	Verkko	52.29/45.99	3.7	934	837	727	661
PacBio HiFi	hifiasm	78.7/63.1	3.6	772	657	524	450

Each assembly consists of two sets of sequences representing the paternal and maternal haplotypes. The two numbers in each cell give the metrics for the two sets of sequences, respectively.

Q1.3.3: Did the authors directly measure corrected read error rates? Are systematic errors also corrected or masked? Do the authors anticipate that error-corrected reads could be used for other applications (such as reference variant calling)?

A1.3.3: There is currently no mature tool specifically designed for assessing ONT read accuracy. To address the first question, we aligned the HERRO- and hifiasm (ONT)-corrected ONT standard reads of HG002 to the T2T-HG002 reference and calculated both the total number of aligned bases (matches, mismatches, and gaps; column 11 in the PAF file) and the total number of matched bases (column 10). We then derived a BLAST-like alignment identity as the ratio of matched bases to aligned bases, as follows:

- a) HERRO: produced 124.5 Gb of corrected bases with 99.7 % identity
- b) hifiasm (ONT): produced 141.3 Gb corrected bases of 99.1% identity.

We found that HERRO and hifiasm (ONT) put different emphasis on accuracy versus recall. HERRO discards a larger fraction of reads with high error rates, whereas hifiasm (ONT) retains them. Even if these reads cannot be fully corrected, they are still valuable for the assembly step, where they contribute to improved completeness and contiguity. It is worth noting that the “identity” reported here is not equivalent to base-level accuracy, as problematic reads that are not fully corrected often originate from highly repetitive regions and may map to multiple positions in the T2T-HG002 reference. This leads to an overestimation of the error rate.

To address the second question, we also performed variant calling using a well-established HG002 truth set. Both the raw and hifiasm-corrected ONT standard reads were aligned to GRCh38, and variant calling was performed to assess read accuracy (see Table r4). As shown in Table r4, the variant-calling results of the corrected reads are generally comparable to those of the raw reads. DeepVariant tends to perform better on raw ONT reads, as it was originally trained on raw HG002 ONT error profiles. After error correction, both the read quality values and error spectrum change, leading to a mismatch with the model’s learned error characteristics. We

do not recommend variant calling with error-corrected reads because 1) error correction requires much more computing resources and variant calling, and 2) deep-learning-based variant callers such as DeepVariant and Clair3 are trained on raw reads and are suboptimal for corrected reads.

In future work, we plan to further optimize hifiasm (ONT) for assembly-based variant calling, particularly for detecting small-scale variants where current assembly-based approaches remain less effective.

Table r4. Read-based variant calling benchmark of HG002 ONT standard Simplex reads before and after hifiasm (ONT) error correction

Data	INDEL			SNP			ALL		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Raw	0.691102	0.658042	0.674167	0.995107	0.987573	0.991326	0.936363	0.92449	0.930389
Corrected	0.650316	0.726876	0.686468	0.994734	0.987643	0.991176	0.918942	0.937724	0.928238

Raw and corrected reads were aligned to GRCh38 and called using deepvariant. Variant calling accuracy was evaluated using hap.py.

Q1.3.4: HERRO-based assemblies seem to have slightly higher QV scores, but could this be because of regions that were uniquely assembled by hifiasm? Maybe restricting the comparison to the same regions that were assembled by both methods would help answer this?

A1.3.4: Compared with Verkko + HERRO, hifiasm (ONT) uniquely assembles through highly repetitive genomic regions. These regions cannot be accurately evaluated by current QV assessment tools, which estimate QV based on k-mer profiling. Such methods are unreliable in highly repetitive regions, as they contain large numbers of abnormal high-frequency k-mers that distort the score.

To address this limitation, we performed an assembly-based variant-calling comparison using the HG002 ground truth and incorporated the results into the revised manuscript (Supplementary Table 6 in the Supplementary Material). We found that the main difference in QV between hifiasm (ONT) and HERRO arises from indels within homopolymer regions. In contrast, SNP accuracy is highly comparable between the two, both within and outside of homopolymer regions. These results confirm our earlier conclusion that QV differences are primarily driven by homopolymer-associated indels, which typically occur outside of functionally important regions. It is also worth noting that HERRO was trained specifically on HG002 data, which may provide it with an inherent advantage in this comparison.

Methods:

Q1.3.5: Does the algorithm implicitly assume that all haplotype copies are balanced (in either diploid or repetitive regions)? Can error correction be potentially used for metagenome or cancer assembly? If the answer is “no”, a simple clarification is sufficient.

A1.3.5: The error correction algorithm in hifiasm (ONT) does not rely on a diploid genome assumption, which is critical for resolving highly repetitive regions such as centromeres, where large numbers of highly similar repeat copies occur. As shown in Table r1, Figure r1, and Section **A1.0.1**, the error correction module of hifiasm (ONT) performs effectively for both metagenomic and cancer genome assemblies. However, achieving optimal performance for these complex assemblies requires further optimization in the downstream assembly stage. We thus did not present these results in the manuscript. In future work, we will focus on these improvements, with particular emphasis on cancer genome assembly.

Q1.3.6: Does all-vs-all read alignment remain a computational bottleneck, compared to the error correction and the rest of the assembly? Did the authors make any improvements to take advantage of more accurate iterations of ONT data?

A1.3.6: As shown in Figure 2 of the manuscript, error correction is the primary bottleneck in the entire assembly process, with the all-vs-all read alignment step being the most computationally intensive component. Unlike methods that exploit the improved accuracy of ONT data to simplify correction, hifiasm (ONT) reimplements the all-vs-all alignment algorithm to better tolerate the higher error rates of ONT reads compared to PacBio HiFi data. In the original hifiasm, alignments are performed in windows of 375 bp, assuming relatively low sequencing error rates. In contrast, hifiasm (ONT) performs full-length alignments between each pair of reads to accommodate the greater error rate inherent to ONT sequencing.

Q1.3.7: Will error correction work for R9 or earlier R10 data with previous basecaller versions? Can the authors provide minimum recommendations? Is the algorithm compatible with data from other large-scale ONT sequencing projects, such as G1k resequencing efforts or long-read CARD initiative?

A1.3.7: We have now clarified in the revised manuscript that hifiasm (ONT) is compatible only with R10 data. ONT currently supports only the R10 protocol, meaning all newly generated ONT reads are R10, with no R9 data being produced. There is no specific requirement regarding the basecaller version. The assemblies presented in Supplementary Table 3 (Supplementary Material) were generated using reads basecalled with an earlier version of Guppy (v6.1.2) rather than the latest Dorado basecaller, and hifiasm (ONT) performed well on these reads.

Hifiasm (ONT) has also been tested and adopted in several large-scale ONT sequencing projects using R10 data. As shown in Supplementary Table 3, it successfully assembled R10 datasets from the CARD initiative. Additionally, based on our personal communication, hifiasm (ONT) has been used by the 1000 Genomes Project ONT Sequencing Consortium for assembling R10 data.

Q1.3.8: I understand the main principle of the DP approach to compute the maximum sequence of compatible sites. However I was unable to find information on how given two sites, their pairwise compatibility is decided ($S[j] \leftrightarrow S[i]$)? This seems to be an important practical detail.

A1.3.8: We thank the suggestion of the reviewer and have now clarified this in the revised manuscript. Two sites $S[j]$ and $S[i]$ are considered compatible if and only if $S[i][k] = S[j][k]$, for all k such that $S[i][k] \in \{0, 1\}$ and $S[j][k] \in \{0, 1\}$. 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 $S[i]$.

Misc:

Q1.3.9: It would be helpful to have a supplementary table that lists all data accession IDs.

A1.3.9: We have now incorporated this in the revised manuscript (Supplementary Table 8).

Referee #3 (Remarks on code availability):

Q1.3.10: The core repository seems to be in a good shape, and I previously used hifiasm in other modes.

A1.3.10: We thank the reviewer for identifying our strength.

Referees' comments:

Referee #1 (Remarks to the Author):

Q1.1: Thank you for addressing my comments. I don't have anything further. All the best

A1.1: We thank the reviewer for the constructive comments.

Referee #1 (Remarks on code availability):

Q1.2 Easy to use, as is the case with other tools developed by Heng Li.

A1.2: We thank the referee for the positive comment on the usability of the software.

Referee #2 (Remarks to the Author):

Q2.1: I am satisfied with all responses the authors gave to my questions. Congratulations to the authors.

A2.1: We thank the reviewer for the constructive comments.

Referee #3 (Remarks to the Author):

Q3.1: I thank the authors for answering all my requests. I am highly enthusiastic about this work.

A3.1: We thank the reviewer for the constructive comments.