Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A practical framework for detecting and correcting Nanopore homopolymer errors in animal mitochondrial genome assembly

You ZHOU
City University of Hong Kong
Zhihao GUO
City University of Hong Kong
Xinyao Wu
City University of Hong Kong
Liesbeth Frias
City University of Hong Kong
Rebecca J. Traub
Murdoch University
Xiuqi WANG
South China Agricultural University
Runsheng LI
runsheng.li@cityu.edu.hk

City University of Hong Kong

Research Article

Keywords: Nanopore sequencing technology, Long homopolymer tracts (N > 10), simplex and duplex reads, taxonomic risk map

Posted Date: June 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9835208/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Mitochondrial genomes are fundamental to understanding genetic evolution and phylogenetic relationships. The development of novel long-read sequencing technologies enables the direct sequencing of the complete mitochondrial genome. For example, Oxford Nanopore Technologies (ONT) sequencing has streamlined the acquisition of mitogenomes and seen widespread application in model organisms such as humans and zebrafish. However, the current R10.4.1 flow cells still struggle to accurately resolve long homopolymer tracts (N > 10) within the mitochondrial genome.

Results

To evaluate the effect of accurate R10.4.1 flow cells, which caused long homopolymer tracts, we used six R10.4.1 mitochondrial genome datasets as curated case studies to benchmark homopolymer error rates, assembly outcomes, and correction strategies within a single, reproducible evaluation framework. Our results demonstrate that the accuracy of both simplex and duplex reads frequently falls below 50% in long homopolymer tracts (N > 10). Consequently, assemblies from both read types are prone to insertions and deletions (indels), especially deletions. Simplex assemblies exhibit a high frequency of indels. In contrast, duplex reads mitigate this issue by converting large indels into smaller ones. We recommend reference-guided polishing as an effective strategy for correcting indel errors in assemblies. Furthermore, we found that long homopolymer tracts (N > 10) are widespread across the animal kingdom and are particularly enriched in nematode mitochondrial genomes.

Conclusions

Collectively, this study provides a practical, reproducible framework and decision guide for diagnosing, quantifying, and correcting Nanopore homopolymer errors in mitochondrial genome assembly, illustrated with curated animal mitogenome case studies and a taxonomic risk map.

Introduction

Mitochondria play central roles in cellular metabolism, apoptosis, disease, and aging [1–3]. As the site of oxidative phosphorylation, they are responsible for ATP production and energy homeostasis. Unlike the nuclear genome, mitochondria possess their own genome (mtDNA), a small, circular molecule in animals that typically encodes 37 genes involved in oxidative phosphorylation and mitochondrial protein synthesis [4]. Owing to its maternal inheritance, absence of recombination, and relatively rapid rate of sequence evolution, mtDNA has long served as an important marker for phylogenetic and evolutionary studies [5]. While gene order is generally conserved within major lineages, variation between taxa makes mitochondrial genomes informative for exploring both phylogeny and genome evolution [6, 7].

The rapid development of DNA sequencing technologies has greatly expanded access to mitochondrial genomes, enabling comparative analyses across a wide range of animal groups. Among these technologies, Oxford Nanopore Technologies (ONT) sequencing has emerged as an effective approach for directly assembling complete mitochondrial genomes due to its long reads [8]. ONT sequencing infers nucleotide sequences by measuring changes in ionic current as DNA strands pass through nanoscale pores [9, 10]. Despite continuous improvements in sequencing chemistry and pore design, accurately resolving regions of low sequence complexity remains a major challenge. Homopolymer tracts composed of consecutive identical nucleotides generate limited variation in current signals, making it difficult to determine their exact length. This limitation frequently introduces insertion–deletion (indel) errors during basecalling and reduces assembly accuracy [11, 12].

DNA homopolymers are the simplest form of simple sequence repeats and are widespread in animal mitochondrial genomes with substantial variation among species [13]. The formation and elongation of homopolymer tracts are often attributed to replication slippage and imperfect repair processes, and their length can influence DNA structure, transcription, translation, and protein binding [14–17]. The abundance and distribution of homopolymers differ markedly among taxa, reflecting distinct evolutionary and functional constraints [17]. Long homopolymer tracts (N > 10 bp) are particularly common in AT-rich mitochondrial genomes, such as those of parasitic nematodes, where they may hinder accurate genome assembly [15, 18].

In this study, the impact of homopolymer tracts on read accuracy and mitochondrial genome assembly was systematically evaluated using six datasets generated with ONT R10.4.1 flow cells (Fig. 1A). Overall sequencing quality was assessed, followed by an analysis of the relationship between homopolymer length and read-level basecalling accuracy. Basecalling accuracy declined progressively with increasing homopolymer length, with a marked reduction observed when tract length exceeded approximately 10 bp. Although shorter homopolymers generally showed higher accuracy, occasional errors were still detected, while longer tracts did not invariably result in indels. A comparison of simplex and duplex reads was performed to assess improvements in sequencing accuracy within homopolymer regions, and the effectiveness of post-assembly polishing in correcting homopolymer-associated errors was further evaluated. To provide an evolutionary perspective, homopolymer distributions were examined across 7,583 mitochondrial genomes representing 23 animal phyla. Here, we provide a practical, reproducible workflow and decision guide for Nanopore mitogenome projects: diagnose homopolymer errors at the read level, quantify error propagation into assemblies, apply correction strategies, and prioritize validation using a taxonomic risk map.

Methods

Datasets retrieval and processing

In this study, a total of six datasets were used, of which five were retrieved from the NCBI database (https://www.ncbi.nlm.nih.gov) and one was obtained from our previous study (to be released soon).

Detailed information for the six datasets, including basecalling models, library preparation kits, and SRA accession numbers, is provided in Supplementary Table 1. Zebrafish (*Danio rerio*) and human (*Homo sapiens*) mitochondrial genomes were used as negative controls for long-homopolymer-associated errors, as their reference mitochondrial genomes contain no homopolymer tracts longer than 10 bp. In contrast, the remaining genomes contain varying numbers of homopolymer regions and were therefore included for comparative analysis. In addition, datasets from the common fruit fly (*Drosophila melanogaster*) and the subcutaneous nodule worm, *Dirofilaria asiatica*, included both simplex and duplex reads, enabling further comparison between these sequencing modes.

To isolate reads originating from mitochondrial genomes, all reads were aligned to the corresponding mitochondrial reference genomes using Minimap2 (2.22-r1101, https://github.com/lh3/minimap2) [19]. Properly mapped reads were subsequently extracted using Samtools (v1.13, https://github.com/samtools/samtools) [20]. For the two genomes that included both simplex and duplex datasets, these reads were analyzed separately to evaluate their relative performance in homopolymer accuracy assessment. Finally, NanoPlot (1.32.1, https://github.com/wdecoster/NanoPlot) was used to perform quality control on all datasets [21].

Mitochondrial genome assembly pipeline

For all mitochondrial genomes, a standardized bioinformatic pipeline was applied across all species. Briefly, raw Nanopore reads were first filtered using NanoFilt (v2.8.0, https://github.com/wdecoster/nanofilt) with stringent thresholds (quality score > 7 and read length > 1,000 bp) [21]. Subsequently, the mitochondrial genome of each species was assembled de novo using Flye (2.9.6, https://github.com/mikolmogorov/Flye) with the --meta parameter to improve the recovery of high-coverage sequences [22]. To reduce stochastic assembly variation, the Flye pipeline was executed 6 times independently for each dataset. We then statistically analyzed indels across the six replicate assemblies per dataset.

Sequencing coverage varied across datasets. However, all mitochondrial genomes included in this study exceeded a minimum coverage of 100×, which is sufficient for reliable homopolymer accuracy estimation. Therefore, differences in sequencing depth were not expected to substantially influence the comparative analyses.

Based on the reported mitochondrial genome lengths of the corresponding species available in GenBank, the contigs most closely matching the expected reference genome size were selected for each species. Assembly accuracy and structural features were further evaluated by visualizing the selected contigs using CGView (https://cgview.ca) [23].

Homopolymer analysis

To evaluate the accuracy of homopolymer sequencing across different datasets, we performed variant calling using Giraffe (https://github.com/lrslab/Giraffe_View) on the mitochondrial genome assemblies

[24]. Homopolymer regions were systematically categorized by both type (A/T/C/G). Statistical analyses and visualization were conducted using R (v4.4.2, https://www.r-project.org) with ggplot2, enabling comparative assessment of error patterns among different homopolymer classes [25, 26].

Homopolish correction analysis

Neither simplex nor duplex reads could accurately assemble the mitochondrial genome, as contigs contained multiple indels that fragmented CDS, tRNA, and rRNA regions. We therefore polished the contigs derived from duplex reads using an additional reference genome (NC_031365), compared pre- and post-polishing sequences with MUMmer (v4.0.1, https://github.com/mummer4/mummer), and visualized Homopolish-corrected sites, counts, and homopolymer lengths in R (v4.4.2, https://www.r-project.org) using ggplot2 [25, 27].

Animal mitochondrial genome selection and processing

To characterize the distribution of long homopolymer tracts in animal mitochondrial genomes, we downloaded all available complete animal mitochondrial genome sequences from the NCBI database. To minimize taxonomic redundancy, only one representative mitochondrial genome was retained for each genus. After filtering, 7,583 mitochondrial genomes representing distinct genera were included in the subsequent analyses.

To further examine the distribution of long homopolymers within the class Nematoda, a separate dataset was constructed at the order level. For this analysis, one representative mitochondrial genome was retained for each order. In total, 100 mitochondrial genomes representing different nematode orders were collected and used for downstream analyses.

Phylogenetic tree construction

A maximum-likelihood phylogenetic tree was reconstructed using concatenated sequences of the 12 mitochondrial protein-coding genes (ND1, ND2, ND3, ND4, ND4L, ND5, ND6, COX1, COX2, COX3, CYTB, and ATP6). Gene sequences were aligned using MAFFT (v7.525, https://github.com/GSLBiotech/mafft) with default parameters [28]. Subsequently, phylogenetic inference was performed using IQ-TREE (v3.0.1, https://github.com/iqtree/iqtree2) under the following parameters, with automated model selection (MFP), 1000 ultrafast bootstrap replicates, and 1000 SH-aLRT tests to assess nodal support [29].

Results

The datasets derived from R10.4.1 flow cells exhibited high-quality mitogenomic assembly and coverage

As the first module of our benchmarking framework, we evaluated overall read quality for six R10.4.1 mitogenome datasets (Fig. 1B). Four datasets (*D. rerio*, *Kyphosus azureus*, *H. sapiens*, and *Caenorhabditis elegans*) were selected to demonstrate the inherent accuracy limitations of simplex reads in homopolymer tracts. In addition, two datasets (*D. melanogaster* and *D. asiatica*) were included because they contained both simplex and duplex reads. These two datasets allowed for a direct performance comparison between the two basecalling modes. Across all datasets, median read accuracies remained high at 98–100%. This confirmed that the sequencing data were of sufficient quality for downstream analysis. In the datasets containing both read types, duplex reads consistently exhibited higher read accuracy than simplex reads (Fig. 1B). In *D. melanogaster*, duplex reads reached a median quality score of 29.1, compared with 17.9 for simplex reads. Similarly, in *D. asiatica*, the median read quality of duplex reads was 24.1, markedly higher than the 15.7 observed for simplex reads.

We further assessed overall homopolymer accuracy across the six datasets (Fig. 1B). In the simplex-only datasets (*D. rerio*, *K. azureus*, and *H. sapiens*), homopolymer accuracies for all four nucleotides exceeded 85%, whereas *C. elegans* showed lower overall homopolymer accuracy, with cytosine homopolymers dropping to 61.8%. Consistent with the read-level results, duplex reads in *D. melanogaster* and *D. asiatica* exhibited higher homopolymer accuracy than their simplex counterparts, confirming that duplex basecalling improves accuracy at both the read and homopolymer levels.

Nanopore mitochondrial genome assembly was affected by long homopolymers (N >10)

To evaluate whether homopolymer length further influences read accuracy in R10.4.1 flow cells, we examined the relationship between homopolymer length and basecalling accuracy across the four mitochondrial genomes. Accordingly, we further evaluated the relationship between homopolymer length and read accuracy for four-base homopolymers across the mitochondrial genomes of four species. Mitochondrial genome coverage across all four datasets exceeded 100×. In *D. rerio* and *H. sapiens*, the longest homopolymer observed was eight nucleotides, with read accuracy remaining above 65% for all bases. In contrast, in *K. azureus* and *C. elegans*, homopolymers of 11T and 11A, respectively, were identified, where read accuracy dropped below 50%. Overall, a consistent trend was observed across the four mitochondrial datasets, with read accuracy decreasing as homopolymer length increased (Fig. 2A). However, the accuracy curves were not smooth, reflecting the uneven distribution of homopolymer tracts of certain lengths within mitochondrial genomes.

Given substantial accuracy loss at homopolymer lengths of 11 nucleotides, we next evaluated whether these errors propagate into mitochondrial genome assembly. We therefore processed all four datasets through a unified pipeline and assembled the mitochondrial genomes, as shown in Fig. 2B. The mitochondrial genomes of *D. rerio* and *H. sapiens* contained no homopolymer regions exceeding 10 nucleotides, and the assembled contigs exhibited 100% average identity with their respective reference genomes (Fig. 2B). In contrast, the mitochondrial genome of *K. azureus* possessed a single 11T

homopolymer region that introduced an indel within the D-loop region, yielding an average identity of 99.98%. For *C. elegans*, one 11A homopolymer region was detected, which caused an indel in a tRNA gene and reduced the average identity to 99.15% (Fig. 2B-C). Overall, we defined the length of a homopolymer tract above 10 as a long homopolymer (N >10), in contrast is short homopolymer (N ≤10). And our results support a simple diagnostic rule within our framework: when read accuracy drops below 50%, errors are likely to occur in downstream assembly.

Duplex reads improved accuracy over simplex reads but remained susceptible to errors due to long homopolymer tracts

We used paired simplex and duplex datasets from *D. asiatica* and *D. melanogaster* to quantify duplex gains and remaining errors across homopolymer lengths (Fig. 3). In Nanopore sequencing, simplex reads are generated from a single DNA strand, whereas duplex reads combine signals from both complementary strands of the same DNA molecule, which generally improves basecalling accuracy. Recent advances in Nanopore chemistry and basecalling algorithms have substantially increased read accuracy, particularly with the introduction of duplex sequencing. Nevertheless, homopolymer tracts remain challenging for accurate basecalling. Consistent with previous studies, duplex reads in our datasets exhibited higher accuracy than simplex reads across most homopolymer lengths.

At the read level (Fig. 3A), accuracy remained high for short homopolymers but declined rapidly with increasing repeat length, regardless of whether duplex or simplex reads were used. In *D. asiatica*, read accuracy exceeded 80% for homopolymers shorter than 8 T, but dropped steeply thereafter, falling below 50% once the homopolymer length exceeded 10 and reaching the lowest level at 18. A similar pattern was observed in *D. melanogaster*, where both A- and T-homopolymers showed a sharp decrease in read accuracy beyond 10 bp. Across both species, the majority of homopolymers longer than 10 bp exhibited read accuracies below 50%, whereas only a small number of homopolymers shorter than 10 bp reached similarly low accuracy. Duplex reads consistently outperformed simplex reads, although this advantage diminished as homopolymer length increased.

At the assembly level (Fig. 3B), homopolymer-associated errors were further amplified. Assembly accuracy remained close to 100% for homopolymers up to 4–6 bp but collapsed dramatically beyond this range. In *D. asiatica*, assembly accuracy showed minor fluctuations at 7–8 bp but remained above 90%, and recovered to 100% at 9–10 bp. However, when the homopolymer length reached 11 bp, assembly accuracy declined sharply, and for homopolymers longer than 12 T, assembly accuracy for both simplex and duplex data dropped to near zero. Although a partial recovery to ~33% was observed at 14 T for both read types, accuracy subsequently decreased to 0% at 15 T for simplex reads and at 18 T for duplex reads. In *D. melanogaster*, assembly accuracy for A- and T-homopolymers decreased to 100% once lengths reached 8 A and 7 T, respectively. Notably, for T-homopolymers, simplex and duplex reads exhibited identical assembly accuracy across all lengths. At long homopolymers (N > 10), assembly

accuracy abruptly failed, often resulting in complete loss of accuracy, indicating that long homopolymers represent a major bottleneck for correct assembly even when duplex data are used.

We then analyzed the types of errors observed at both the read and assembly levels (Fig. 3C). The error type analysis revealed a strong bias toward deletions over insertions. At the read level, deletion errors were more frequent than insertions in both species, with simplex reads showing substantially higher error counts than duplex reads. This trend was further pronounced at the assembly level, where deletion errors dominated the error spectrum, particularly 1 bp indels, followed by short 2–3 bp indels. Longer indels (4 bp or 5–7 bp, depending on species) were comparatively rare and occurred almost exclusively in assemblies derived from long homopolymers. Thus, compared with simplex reads, duplex reads substantially reduced the occurrence of longer indels, shifting the error profile toward shorter indels such as 1–3 bp events.

In summary, analysis of both read-level accuracy and assembly outcomes demonstrates that homopolymer length is a primary determinant of sequencing performance in R10.4.1 datasets. Duplex reads consistently achieved higher accuracy than simplex reads across homopolymer tracts of varying lengths, thereby reducing the number and size of homopolymer-associated indels in assembled mitochondrial genomes. However, once homopolymer length exceeded 10 nucleotides, read accuracy for both read types declined sharply to below 50%, resulting in persistent indel errors in the final assemblies.

Reference-guided correction effectively resolves mapping errors in *D. asiatica*

To illustrate how diagnostic read-level profiles translate to feature-specific assembly risk and how error correction redistributes errors, we analyzed T-homopolymer accuracy, tract-length composition, and Homopolish-corrected sites across genomic features in *D. asiatica* (Fig. 4).

At the genome-wide level (Fig. 4A), T-homopolymer accuracy was generally high for short homopolymers (N≤10), with most sites showing accuracies above 80% across the mitochondrial genome and only 4 sites exhibiting accuracies below 0.5 within this length range. In contrast, homopolymers longer than 10 bp exhibited markedly reduced accuracy, with most sites falling below the 0.5 threshold, while only a small number retained accuracies above 0.5. These low-accuracy long homopolymers were sparsely distributed across genomic positions but accounted for most severely error-prone sites, indicating a strong length-dependent effect rather than regional clustering.

Analysis of homopolymer length composition across mitochondrial features revealed that short homopolymers (N ≤10) dominated all functional categories (Fig. 4B). Coding sequences (CDS), rRNAs, tRNAs, and non-coding regions were all primarily composed of short homopolymer tracts; however, long homopolymers (N >10) were disproportionately enriched at specific loci, particularly within a subset of

CDS (e.g., ND genes and CYTB) and non-coding regions. rRNA genes also contained detectable fractions of long homopolymers, whereas tRNAs exhibited the lowest overall homopolymer burden.

We next examined the genomic positions and types of homopolymer corrections applied by Homopolish (Fig. 4C). Corrected sites were distributed across the mitochondrial genome, with the majority occurring in CDS regions, followed by rRNA and non-coding regions. Both insertion and deletion corrections were observed, with deletions being slightly more prevalent overall. Notably, correction events within CDS were broadly distributed along gene bodies, whereas corrections in rRNA and non-coding regions tended to cluster in regions containing longer homopolymer tracts.

Consistent with this pattern, the number of Homopolish-corrected sites differed markedly among functional categories (Fig. 4D). CDS accounted for the largest number of corrected sites, with multiple corrections observed in ND genes (*ND1–ND6* and *ND4L*), whereas rRNA genes (12S and 16S) and tRNAs showed fewer correction events. Non-coding regions exhibited a moderate number of corrections despite their shorter total length, suggesting a relatively high local density of error-prone homopolymers.

Finally, consistent with these observations, analysis of homopolymer length at corrected sites demonstrated that most reference-genome guided corrections occurred at long homopolymers (N > 10) (Fig. 4E). Corrected homopolymer lengths in CDS and rRNA genes frequently exceeded this threshold, whereas shorter homopolymers were rarely corrected. Non-coding regions contained both the longest homopolymers and the widest length range, including several extreme cases exceeding 15 bp.

Together, these results indicate that homopolymer length is the primary determinant of error occurrence and correction in mitochondrial genome assemblies, with long homopolymers (N >10) being rare but disproportionately responsible for low-accuracy sites and post-assembly corrections. These error-prone homopolymers are unevenly distributed across mitochondrial features, contributing most strongly to errors within coding and non-coding regions. This case study shows where reference-guided polishing is most informative, predominantly at N > 10 in CDS and rRNA, and how to visualize correction density alongside raw homopolymer risk.

Long homopolymer tracts are unevenly distributed across the animal phylum and nematode lineages

Finally, we constructed a taxonomic risk map by summarizing the presence of long homopolymer tracts (N >10) across 7,583 genus-representative animal mitogenomes (23 phyla; Supplementary file), then zoomed into Nematoda at the Order level (Fig. 5). To reduce sampling bias, phyla represented by fewer than 100 species were grouped into an “Others” category. As shown in Fig. 5A, species belonging to Chordata, Mollusca, and Cnidaria exhibited relatively low proportions of mitochondrial genomes containing homopolymer regions longer than 10 nucleotides. In contrast, Arthropoda, Nematoda, Annelida, Platyhelminthes, and the pooled “Others” category displayed markedly higher proportions of

species harboring such extended regions, indicating that long homopolymers are unevenly distributed across phyla.

To further examine this pattern within nematodes, we analyzed mitochondrial genomes across different orders. The order Rhabditida contained the largest number of species (90 in total), of which 65 harbored more than 10 homopolymer regions, while 25 contained fewer. By comparison, all four species within Trichinellida possessed fewer than 10 homopolymers, whereas Mermithida (two species) and Plectida (one species) exclusively harbored more than 10. Within Dorylaimida, three species were analyzed, with one species containing more than 10 homopolymers and two containing fewer (Fig. 5B). These findings highlight substantial variability among nematode lineages, with certain orders exhibiting consistent genomic patterns.

To provide phylogenetic context, we constructed a phylogenetic tree of 100 representative nematode species at the order level based on mitochondrial genome sequences. The topology resolved the major clades, with Rhabditida forming a large and heterogeneous cluster that reflects its diversity in homopolymer content. In contrast, Trichinellida and Mermithida formed compact and uniform branches, consistent with their conserved patterns of homopolymer representation.

Discussion

Read-level diagnostics in our framework recapitulate a consistent pattern, revealing that Nanopore is generally reliable for short homopolymers, whereas long homopolymer tracts (N > 10) remain the dominant source of systematic error. This pattern is evident in routine applications, such as human mitochondrial genome sequencing, where homopolymer lengths do not exceed 10 nucleotides, and R10.4.1 Nanopore chemistry can reliably resolve these tracts. The R10.4.1 chemistry achieves approximately 85% overall homopolymer accuracy [12, 30]. Beyond this 10 bp threshold, read basecalling accuracy drops sharply and continues to deteriorate with increasing homopolymer length. Duplex sequencing, which leverages signal redundancy from both DNA strands, improves basecalling fidelity [31]. Duplex reads, therefore, consistently outperform simplex reads. Nevertheless, both duplex and simplex reads systematically underestimate the lengths of extended homopolymers. Fundamental constraints of Nanopore technology cause these read-level errors [32]. The limited dynamic range of ionic current signals and the k-mer based decoding process drive these systematic biases. In summary, Nanopore technology fundamentally struggles to accurately basecall long homopolymer tracts at the raw read level.

At the whole mitochondrial genome assembly level, read-level basecalling biases directly cause structural inaccuracies in homopolymer-rich regions. During the assembly process, read-level underestimation manifests as large deletions within long homopolymer tracts. Applying duplex reads during whole mitochondrial genome assembly substantially reduces the frequency and size of insertion and deletion errors [33], and duplex-based assemblies therefore outperform simplex-based assemblies. Nevertheless, residual deletion errors persist in extended homopolymer regions of the assembled

genomes, even at high sequencing coverage. The systematic underestimation of homopolymer lengths prevents the accurate reconstruction of tracts exceeding 10 nucleotides. Recent improvements in Nanopore chemistry and duplex assembly strategies mitigate homopolymer-associated errors. These technological improvements do not eliminate the errors completely. Consequently, long homopolymers remain an intrinsic constraint for accurate whole mitochondrial genome reconstruction.

In the correction module of our workflow, reference-guided polishing can mitigate many homopolymer indels in *D. asiatica* when closely related references are available. Traditionally, short-read sequencing data are considered necessary for polishing long-read genome assemblies [34, 35]. However, generating additional next-generation sequencing (NGS) data increases project costs and delays turnaround times. Therefore, reference-guided polishing provides a highly efficient alternative. Reference-guided polishing strategies have been shown to correct various homopolymer errors in bacterial, viral, and eukaryotic genomes sequenced with both R9 and R10 Nanopore flow cells [36-38]. For instance, Homopolish effectively resolved 33 indels caused by Nanopore R9 errors in the African swine fever virus (ASFV) genome assembly [39]. However, this strategy requires careful consideration, as reliance on reference sequences may lead to overcorrection of true biological variation [27, 40]. The effectiveness of reference-guided polishing is highly dependent on the accuracy and completeness of the reference genome, which may itself contain unresolved errors, particularly in homopolymer-rich regions [34]. In the absence of short-read data, reference-guided polishing with closely related mitochondrial genomes provides a robust, practical solution for finalizing accurate whole mitochondrial genome assemblies.

Whole-animal mitochondrial genome assemblies reveal distinct distributions of long homopolymers across taxonomic phyla, with nematodes showing particularly extreme homopolymer patterns (Supplementary Table 2). Within Nematoda, homopolymers are predominantly characterized by long poly(T) tracts, and previous studies have also highlighted substantial differences in poly(T) tract abundance between free-living and parasitic nematodes [15, 41]. For example, the free-living model organism *C. elegans* (order Rhabditida) exhibits relatively low frequencies and mutation rates of poly(T) sequences, whereas several parasitic nematodes harbor more frequent poly(T) tracts longer than six bases, often exceeding ten consecutive nucleotides and associated with elevated AT content [42]. Consistent with this trend, the whole mitochondrial genome of *D. asiatica*, a recently described putative filarial parasite, strongly supports an extremely poly(T) distribution pattern. Notably, many of its long homopolymer tracts are primarily located within protein-coding regions in the whole genome, which contrasts with general biological observations that long homopolymers accumulate in non-coding regions to avoid frameshift mutations in protein-coding sequences [43, 44]. The role of host-associated selective pressures in driving these long homopolymer distributions remains unclear. Due to the extreme homopolymer content, Nanopore sequencing technology faces severe limitations in acquiring mitochondrial genomes from these parasitic nematodes. Therefore, we suggest incorporating NGS data or applying reference-guided polishing to correct Nanopore-induced errors in homopolymer-rich parasitic nematodes.

Several limitations of this study should be considered. First, homopolymer accuracy depends on the basecalling model used, and different model versions or training datasets may yield different absolute accuracy estimates. Second, assembly accuracy was assessed using available reference mitochondrial genomes, which themselves may contain unresolved errors, particularly in homopolymer-rich regions. Third, although short-read correction or polishing was discussed as a mitigation strategy, next-generation sequencing data are not error-free and may introduce additional biases in repetitive regions. Together, these factors may overestimate the homopolymer accuracy in homopolymer tracts.

In summary, our study demonstrates that long homopolymer tracts represent a critical and widespread limitation for Nanopore-based mitochondrial genome sequencing. Although R10.4.1 chemistry and duplex basecalling markedly improve overall read quality, they do not fully resolve errors associated with long homopolymer tracts (N >10). Addressing this limitation will likely require further improvements in basecalling algorithms, signal modeling, or hybrid sequencing strategies that integrate high-fidelity short reads. These considerations are particularly important for AT-rich mitochondrial genomes, such as those of parasitic nematodes, where long homopolymer tracts are both common. Together, our findings provide a practical, reproducible framework and decision guide for diagnosing, quantifying, and correcting Nanopore homopolymer errors in mitochondrial genome assembly, illustrated with curated animal mitogenome case studies and a taxonomic risk map.

Conclusion

In conclusion, our study shows that long homopolymer tracts remain a major limitation for Nanopore-based mitochondrial genome sequencing. Although R10.4.1 chemistry and duplex basecalling improve overall sequencing accuracy, systematic errors persist in long homopolymer regions, especially when the tract length exceeds 10 bp. These errors can lead to deletions and assembly inaccuracies in mitochondrial genomes, particularly in homopolymer-rich parasitic nematodes such as *D. asiatica*. Reference-guided polishing can reduce many of these errors when suitable reference genomes are available, although careful application is still required to avoid overcorrection. Overall, combining Nanopore sequencing with additional polishing strategies or short-read data remains important for improving the accuracy of mitochondrial genome assemblies with extreme homopolymer content.

Abbreviations

ONT: Oxford Nanopore Technologies

Indels: insertions and deletions

D. melanogaster: *Drosophila melanogaster*

D. asiatica: Dirofilaria asiatica

D. rerio: *Danio rerio*

H. sapiens: Homo sapiens

CDS: Coding sequences

NGS: next-generation sequencing

Declarations

Data availability

Detailed information for all datasets is available in Supplementary File 1. All code and workflow used in this study are available on GitHub (https://github.com/ZHOUYou-water/Homopolymer-analysis).

Acknowledgements

Not applicable.

Funding

This work was supported by the Early Career Scheme (CityU 21100521) and General Research Fund (11105524) from the Research Grants Council of the Hong Kong Special Administrative Region, China; the Hong Kong Health and Medical Research Fund (project number 08194126); and new Research Initiatives support from City University of Hong Kong (project number 9610497) to R.L.

Author contributions

Y.Z. participated in the design of the study, developed the code, and wrote the manuscript. Z.G. developed the code and helped revise the manuscript. X.W. participated in the data analysis of the study and helped revise the manuscript. L.F., RJT, X. W., and R.L. participated in the study design and helped revise the manuscript. All authors read and approved the final manuscript.

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Supplementary data

Supplementary data are attached to BMC Genomics.

References

1. Boore JL: **Animal mitochondrial genomes**. *Nucleic acids research* 1999, **27**(8):1767-1780.
2. Chen XJ, Butow RA: **The organization and inheritance of the mitochondrial genome**. *Nature Reviews Genetics* 2005, **6**(11):815-825.
3. Rackham O, Filipovska A: **Organization and expression of the mammalian mitochondrial genome**. *Nature Reviews Genetics* 2022, **23**(10):606-623.
4. De Giorgi C, Saccone C: **Mitochondrial genome in animal cells: Structure, organization, and evolution**. *Cell biophysics* 1989, **14**(1):67-78.
5. Rokas A, Ladoukakis E, Zouros E: **Animal mitochondrial DNA recombination revisited**. *Trends in Ecology & Evolution* 2003, **18**(8):411-417.
6. Pereira SL: **Mitochondrial genome organization and vertebrate phylogenetics**. *Genetics and Molecular biology* 2000, **23**:745-752.
7. Dowton M, Castro L, Austin A: **Mitochondrial gene rearrangements as phylogenetic characters in the invertebrates: the examination of genome'morphology'**. *Invertebrate Systematics* 2002, **16**(3):345-356.
8. Zascavage RR, Thorson K, Planz JV: **Nanopore sequencing: An enrichment-free alternative to mitochondrial DNA sequencing**. *Electrophoresis* 2019, **40**(2):272-280.
9. Deamer D, Akeson M, Branton D: **Three decades of nanopore sequencing**. *Nature biotechnology* 2016, **34**(5):518-524.
10. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X: **The potential and challenges of nanopore sequencing**. *Nature biotechnology* 2008, **26**(10):1146-1153.
11. Agah S, Zheng M, Pasquali M, Kolomeisky AB: **DNA sequencing by nanopores: advances and challenges**. *Journal of Physics D: Applied Physics* 2016, **49**(41):413001.
12. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q: **Opportunities and challenges in long-read sequencing data analysis**. *Genome biology* 2020, **21**(1):30.
13. Sung W, Tucker A, Bergeron RD, Lynch M, Thomas WK: **Simple sequence repeat variation in the Daphnia pulex genome**. *BMC genomics* 2010, **11**(1):691.
14. Tóth G, Gáspári Z, Jurka J: **Microsatellites in different eukaryotic genomes: survey and analysis**. *Genome research* 2000, **10**(7):967-981.
15. Riepsamen AH, Gibson T, Rowe J, Chitwood DJ, Subbotin SA, Dowton M: **Poly (T) variation in heteroderid nematode mitochondrial genomes is predominantly an artefact of amplification**. *Journal of molecular evolution* 2011, **72**(2):182-192.
16. Levinson G, Gutman GA: **Slipped-strand mispairing: a major mechanism for DNA sequence evolution**. *Molecular biology and evolution* 1987, **4**(3):203-221.

17. Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M: **Frameshift mutations and the genetic code**. In: *Cold Spring Harbor symposia on quantitative biology: 1966*: Cold Spring Harbor Laboratory Press; 1966: 77-84.
18. Hu M, Chilton NB, Gasser RB: **The mitochondrial genome of Strongyloides stercoralis (Nematoda)–idiosyncratic gene order and evolutionary implications**. *International journal for parasitology* 2003, **33**(12):1393-1408.
19. Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics* 2018, **34**(18):3094-3100.
20. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM: **Twelve years of SAMtools and BCFtools**. *Gigascience* 2021, **10**(2):giab008.
21. De Coster W, Rademakers R: **NanoPack2: population-scale evaluation of long-read sequencing data**. *Bioinformatics* 2023, **39**(5):btad311.
22. Kolmogorov M, Yuan J, Lin Y, Pevzner PA: **Assembly of long, error-prone reads using repeat graphs**. *Nature biotechnology* 2019, **37**(5):540-546.
23. Stothard P, Wishart DS: **Circular genome visualization and exploration using CGView**. *Bioinformatics* 2005, **21**(4):537-539.
24. Liu X, Shao Y, Guo Z, Ni Y, Sun X, Leung AYH, Li R: **Giraffe: a tool for comprehensive processing and visualization of multiple long-read sequencing data**. *Computational and Structural Biotechnology Journal* 2024, **23**:3241-3246.
25. Ihaka R, Gentleman R: **R: a language for data analysis and graphics**. *Journal of computational and graphical statistics* 1996, **5**(3):299-314.
26. Wickham H: **ggplot2**. *Wiley interdisciplinary reviews: computational statistics* 2011, **3**(2):180-185.
27. Huang Y-T, Liu P-Y, Shih P-W: **Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing**. *Genome biology* 2021, **22**(1):95.
28. Katoh K, Misawa K, Kuma Ki, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic acids research* 2002, **30**(14):3059-3066.
29. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R: **IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era**. *Molecular biology and evolution* 2020, **37**(5):1530-1534.
30. Ni Y, Liu X, Simeneh ZM, Yang M, Li R: **Benchmarking of Nanopore R10. 4 and R9. 4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing**. *Computational and structural biotechnology journal* 2023, **21**:2352-2364.
31. Hall MB, Wick RR, Judd LM, Nguyen AN, Steinig EJ, Xie O, Davies M, Seemann T, Stinear TP, Coin L: **Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data**. *Elife* 2024, **13**:RP98300.
32. Cheng H, Qu H, McKenzie S, Lawrence KR, Windsor R, Vella M, Park PJ, Li H: **Efficient near-telomere-to-telomere assembly of nanopore simplex reads**. *Nature* 2026:1-8.

33. Sanderson ND, Kapel N, Rodger G, Webster H, Lipworth S, Street TL, Peto T, Crook D, Stoesser N: **Comparison of R9. 4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction**. *Microbial genomics* 2023, **9**(1):000910.
34. Wick RR, Holt KE: **Polypolish: short-read polishing of long-read bacterial genome assemblies**. *PLoS computational biology* 2022, **18**(1):e1009802.
35. Wick RR, Judd LM, Holt KE: **Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing**. *PLOS Computational Biology* 2023, **19**(3):e1010905.
36. Zakerzade R, Chang C-H, Chatla K, Krishnapura A, Appiah SP, Zhang J, Unckless RL, Blumenstiel JP, Bachtrog D, Wei KH: **Diversification and recurrent adaptation of the synaptonemal complex in Drosophila**. *PLoS genetics* 2025, **21**(1):e1011549.
37. Schiffer AM, Rahman A, Sutton W, Putnam ML, Weisberg AJ: **A comparison of short-and long-read whole-genome sequencing for microbial pathogen epidemiology**. *Msystems* 2025, **10**(12):e01426-01425.
38. Lule DB, Ssemwanga D, Kaleebu P, Tully DC: **The utility of integrating nanopore sequencing into routine HIV-1 drug resistance surveillance**. *Microbial genomics* 2025, **11**(3):001375.
39. Li K, Han X, Shao Y, Zhang Y, Zhao X, Johnson EW, Li R: **ANASFV: a workflow for ASFV whole genome sequencing, assembly, and evaluation**. *bioRxiv* 2024:2024.2007. 2008.602471.
40. Huang Y-T, Liu P-Y, Shih P-W: **High-quality genomes of nanopore sequencing by homologous polishing**. *BioRxiv* 2020:2020.2009. 2019.304949.
41. Hu M, Jex AR, Campbell BE, Gasser RB: **Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing**. *Nature protocols* 2007, **2**(10):2339-2344.
42. Jacob JE, Vanholme B, Van Leeuwen T, Gheysen G: **A unique genetic code change in the mitochondrial genome of the parasitic nematode Radopholus similis**. *BMC Research Notes* 2009, **2**(1):192.
43. Hu M, Gasser R, El-Osta YA, Chilton N: **Structure and organization of the mitochondrial genome of the canine heartworm, Dirofilaria immitis**. *Parasitology* 2003, **127**(1):37-51.
44. Yilmaz E, Fritzenwanker M, Pantchev N, Lendner M, Wongkamchai S, Otranto D, Kroidl I, Dennebaum M, Le TH, Anh Le T: **The mitochondrial genomes of the zoonotic canine filarial parasites Dirofilaria (Nochtiella) repens and Candidatus Dirofilaria (Nochtiella) honkongensis provide evidence for presence of cryptic species**. *PLoS neglected tropical diseases* 2016, **10**(10):e0005028.

Figures

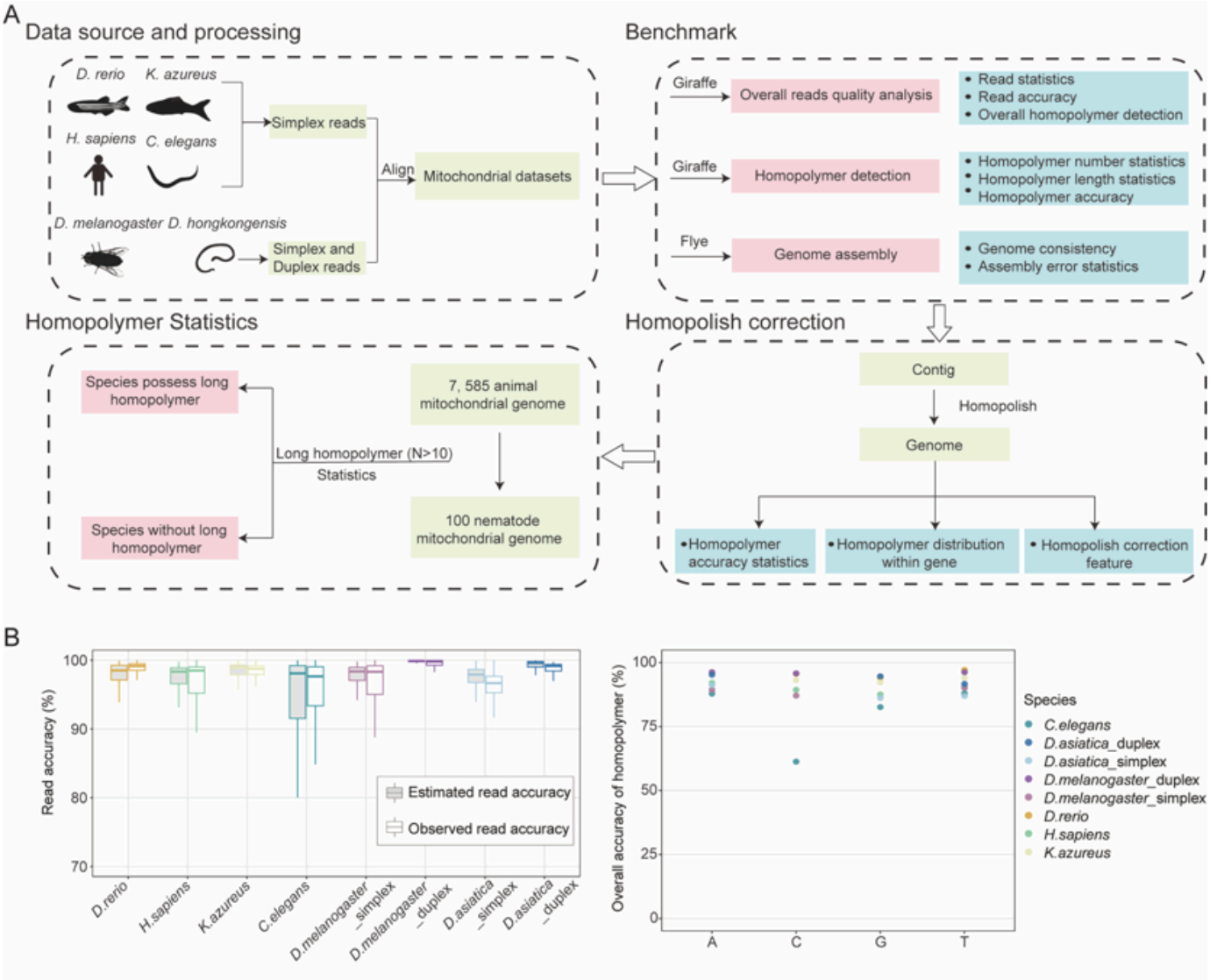

Figure 1

Overview of the analytical workflow and homopolymer-associated accuracy assessment in mitochondrial genomes using R10.4.1 Nanopore sequencing. **A.** Schematic overview of the data sources, processing pipeline, and analytical framework used to evaluate homopolymer-associated errors in animal mitochondrial genomes. Mitochondrial reads generated from simplex and duplex sequencing were subjected to read quality assessment, homopolymer detection, genome assembly, and homopolymer error correction, followed by large-scale homopolymer statistics across animal and nematode mitochondrial genomes. **B.** Read-level accuracy and homopolymer basecalling performance across multiple species. The left panel shows the distribution of estimated and observed read accuracies, while the right panel summarizes overall homopolymer accuracy for each nucleotide (A, C, G, and T), highlighting species-specific differences and the improved performance of duplex reads relative to simplex reads.

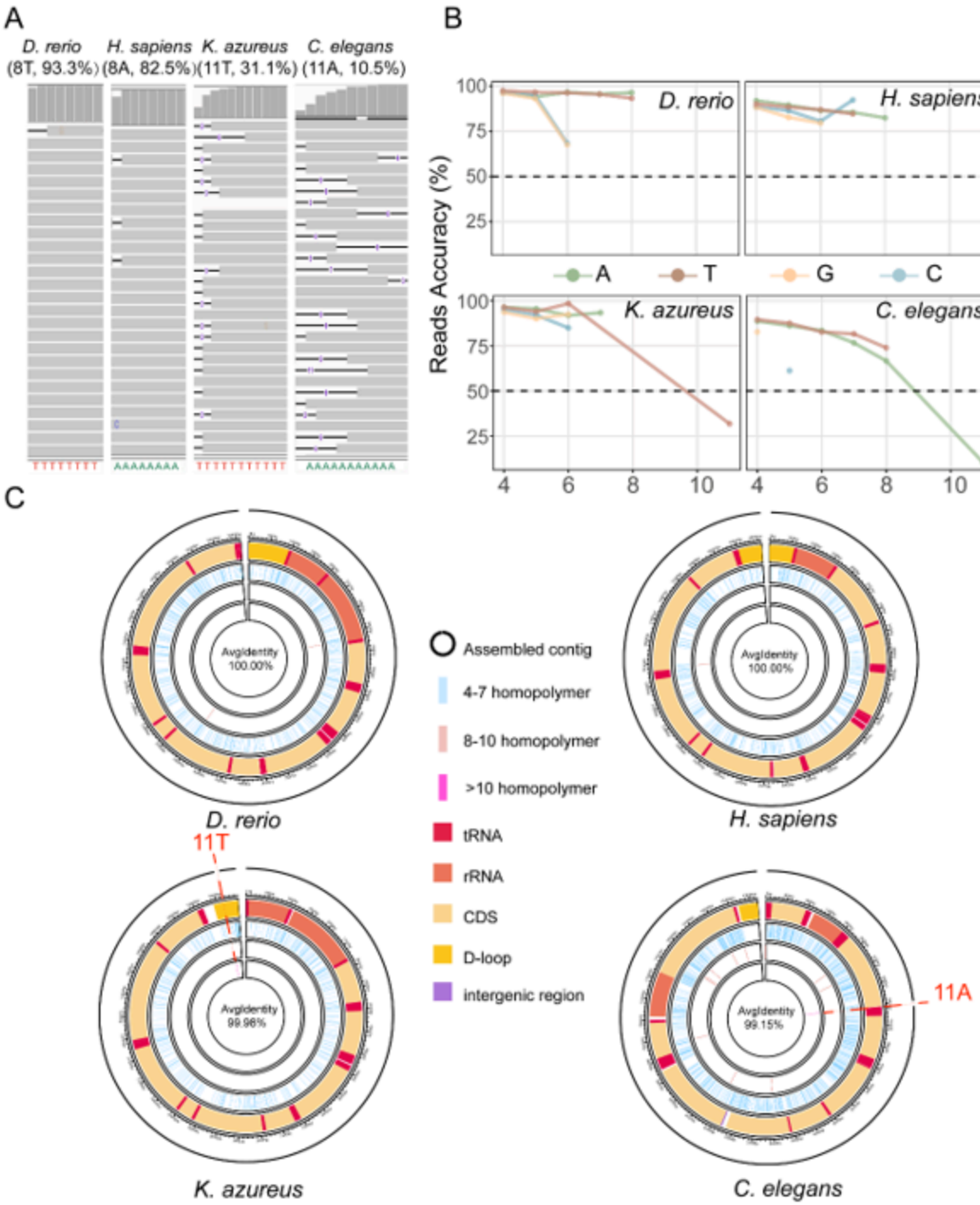

Figure 2

The impact of homopolymer length on read accuracy and mitochondrial genome assembly. **A.** IGV snapshot for the longest homopolymer region with reads in four species. The long homopolymer tracts (N > 10) exhibited low coverage at this position and varied numbers of indels in reads. **B**. The overall trend is that as the homopolymer length increases, the read accuracy gradually decreases and drops below 50% at 11 nucleotides. **C**. A homopolymer length of 11 nucleotides consistently induced an indel error at the corresponding position in the assembled contig.

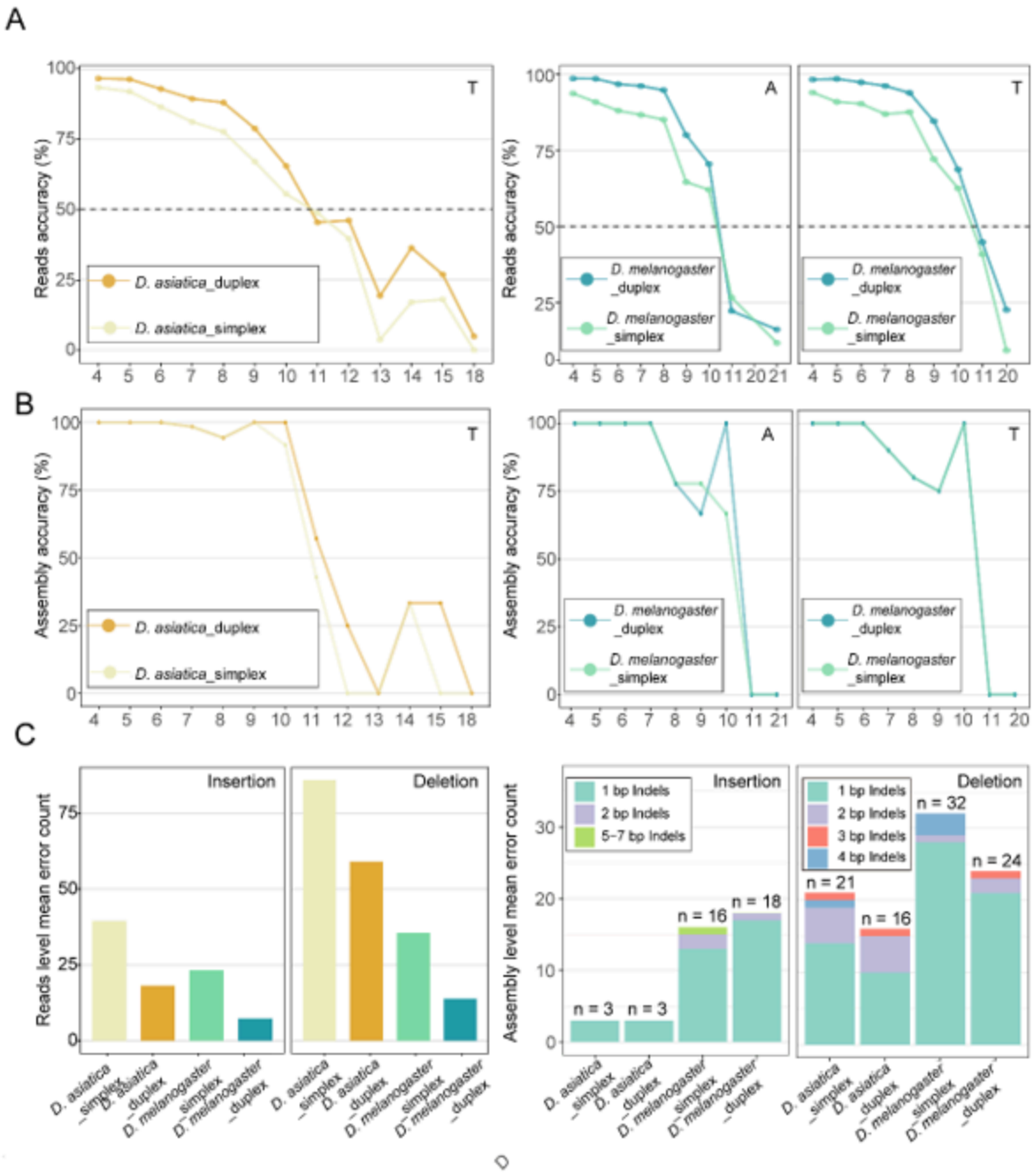

Figure 3

Duplex reads exhibited higher overall read accuracy and improved assembly accuracy. **A.** When the homopolymer length reached 11 nucleotides, read accuracy dropped below 50% for both simplex and duplex data, despite duplex reads being consistently more accurate overall. In *D. asiatica* mitochondrial genome, the numbers of homopolymers across length classes were 265, 159, 95, 63, 35, 24, 12, 7, 4, 6, 3, 3, and 1. In *D. melanogaster mitochondrial genome*, the corresponding counts were 185, 87, 54, 17, 18, 9, 3, 1, and 1 for A homopolymers, and 197, 83, 48, 10, 10, 16, 3, 2, and 1 for T homopolymers. **B.** In genomes containing abundant homopolymers longer than 10 nucleotides, assemblies generated from duplex reads contained fewer indel errors within contigs than those generated from simplex reads. **C.**

Homopolymer tracts predominantly caused deletion errors, and the use of duplex reads substantially reduced the number of indels associated with long homopolymers (N > 10).

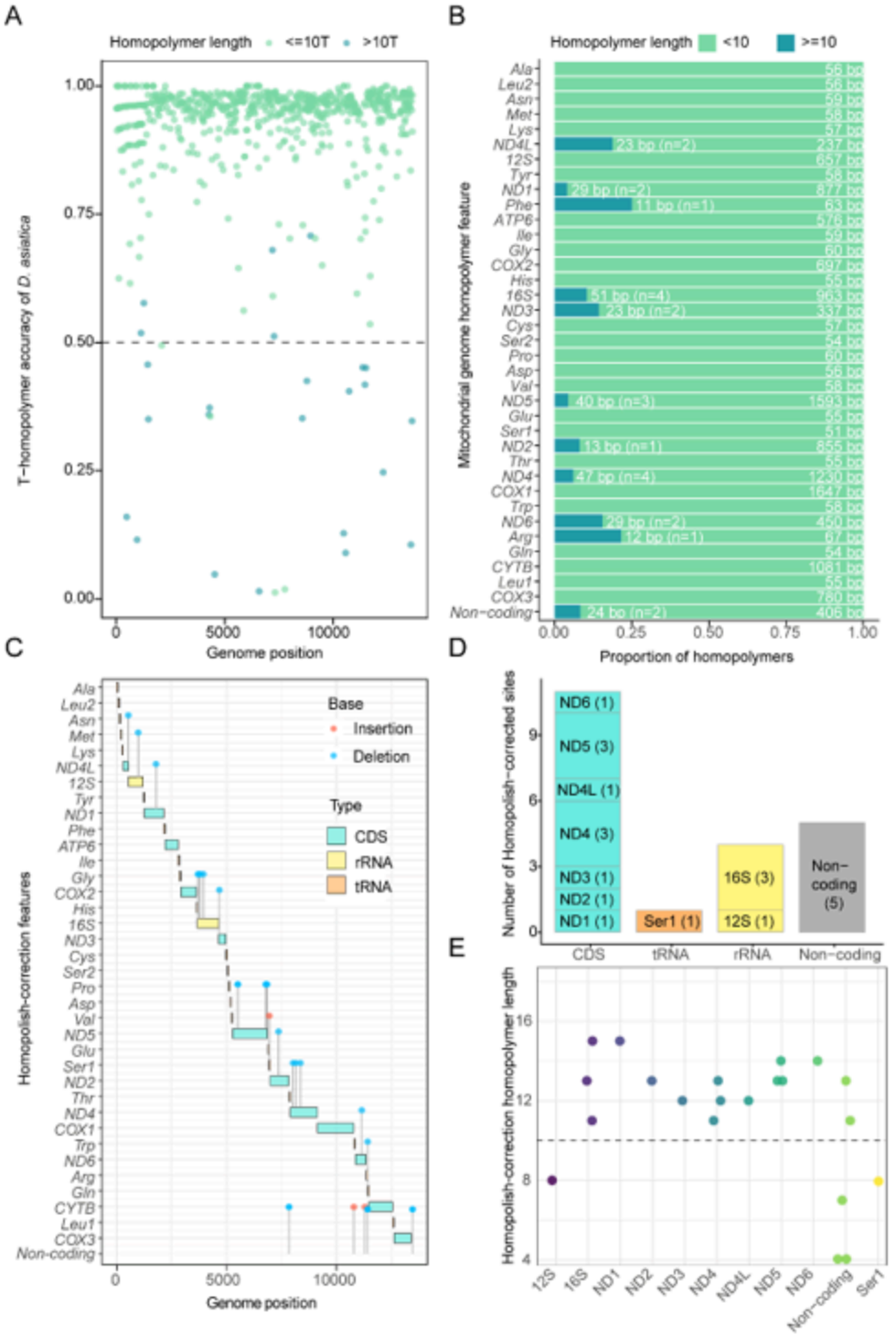

Figure 4

Reference guide polishes predominantly correct long homopolymers in mitochondrial coding regions of the *D. asiatica* mitochondrial genome. **A**. In most genomic loci, T-homopolymers longer than 10 nucleotides show accuracy below 0.5, whereas only four loci with T-homopolymers of 10 nucleotides or

shorter fall below this threshold. Specifically, 708 loci were in the ≤10 group and 24 loci were in the >10 group. **B**. Long homopolymer tracts were predominantly located in ND genes and the 16S rRNA, with a small number also present in tRNA genes such as Phe, and Gln. **C.** Homopolish primarily corrected deletions caused by homopolymer tracts. **D.** Consistent with long homopolymer distribution, homopolish predominantly corrected errors within CDS. **E.** Homopolish corrected indels caused by homopolymers ranging from 4 to 18 nucleotides in length. Length distribution of homopolymers corrected by homopolish across different mitochondrial genomic features, indicating that corrections mainly target homopolymer tracts longer than 10 nucleotides.

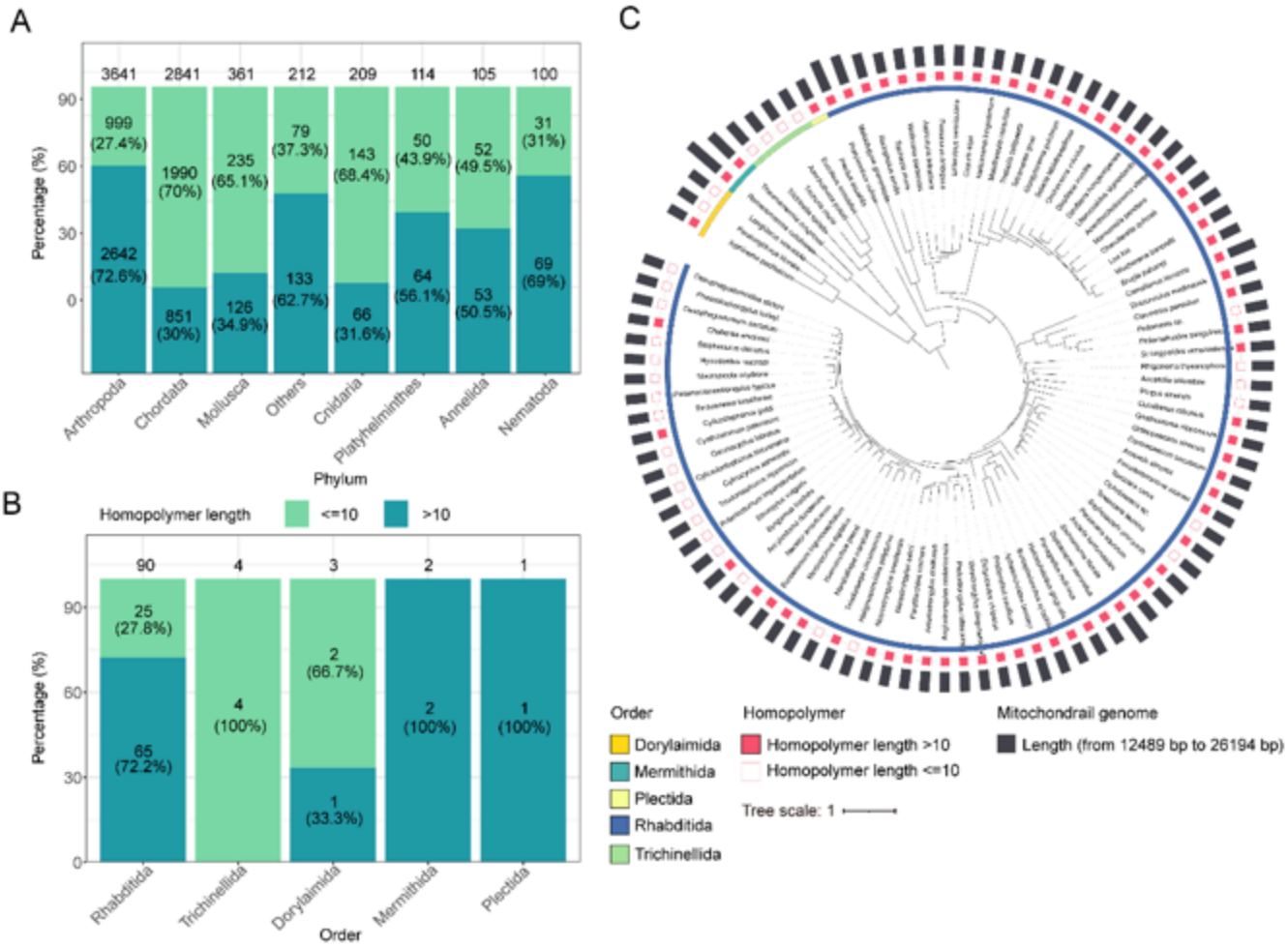

Figure 5

Distribution of homopolymers in phylum and order and phylogenetic tree constructed based on Nematoda. **A.** Proportion of species with mitochondrial genomes containing homopolymer regions longer than 10 nucleotides versus those with shorter homopolymers, summarized at the phylum level. **B.** Distribution of species with or without homopolymer regions longer than 10 nucleotides across different orders within the phylum Nematoda. **C.** Phylogenetic tree constructed from mitochondrial genomes of

representative nematode species, highlighting variations in homopolymer distribution across the nematode.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryFigure.docx