# Phasing or purging: tackling the genome assembly of a highly heterozygous animal species in the era of high-accuracy long reads

Nadège Guiglielmoni[1*] and Philipp H. Schiffer[1]

[1]Institut für Zoologie, Universität zu Köln, Zülpicher str. 47b, 50674 Cologne, Germany

[*]Corresponding author: nguiglie@uni-koeln.de

## Abstract

The revolution of high-accuracy long reads offers unprecedented quality and contiguity in genome assembly. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies have made significant strides in improving their sequencing technologies, yielding reads with error rates below 1% and lengths ranging from kilobases to megabases. These advancements have prompted the development of assembly tools tailored to leverage the enhanced accuracy of long reads. However, the challenge of collapsing haplotypes into high-quality haploid assemblies persists, especially for highly heterozygous genomes. This raises questions about the feasibility and desirability of phased assemblies versus collapsed haploid assemblies. To address these challenges, we benchmarked five assembly tools on ultra-low input PacBio HiFi and Nanopore R10.4 reads from the parthenogenetic nematode species *Plectus sambesii*. We propose a comprehensive methodology for assessing phased assemblies, repurposing existing evaluation programs to collect haplotype-relevant statistics. Our evaluation criteria include assembly size, contiguity, and completeness, with a focus on assessing the accuracy of phased assemblies by examining duplicated BUSCO orthologs and $k$-mer spectra. Additionally, we present strategies for generating collapsed assemblies by purging haplotigs. This study provides valuable insights and guidelines for generating high-quality phased and collapsed *de novo* genome assemblies from highly accurate long reads, particularly beneficial for non-model species genome assembly projects.

## Introduction

High-accuracy long reads have brought drastic improvements to the field of genome assembly, raising the bar for reference assemblies in terms of quality and contiguity. Pacific Biosciences (PacBio) introduced the first long reads with an error rate below 1% through Circular Consensus Sequencing. This technology is a development of the previous PacBio protocol to pass over a fragment multiple times and produce high-fidelity reads, then designated as PacBio HiFi reads [1], with accuracies approaching those of short Illumina reads and a length around 15 kilobases (kb). In parallel, Oxford Nanopore has also updated its chemistry in R10.4 flowcells to reduce errors and perform duplex sequencing to scan fragments twice [2]. The output reads are now deemed sufficient for *de novo* assembly without the need of additional data for polishing, which was crucial for R9.4 Nanopore reads, and their length can still reach values over 100 kb and even 1 Megabase (Mb) [3]. In response, assembly tools have been developed or adapted to benefit from the improved accuracy of these long reads [4].

These advancements have sparked new interest for phased assemblies. Genome assemblies typically provide a haploid representation of diploid and polyploid genomes; this entails collapsing homologous chromosomes (or haplotypes) into a single sequence, by representing homozygous sequences once and selecting one allele per heterozygous site. Collapsing haplotypes into a haploid assembly is increasingly difficult with higher levels of heterozygosity, and can result into suboptimal pseudohaploid assemblies with artefactual duplications [5]. The challenge of collapsing haplotypes for highly heterozygous genomes brings into question whether collapsed haploid assemblies should still be sought for rather than phased assemblies, which include all haplotypes and provide a more complete genome sequence. The error rate of low-accuracy long reads prohibited the discrimination of errors from alternative haplotypes, but high-accuracy long reads open new possibilities for haplotype-resolved assemblies. Some assemblers explicitly aim to yield phased assemblies, such as PECAT [6] and Verkko [7], but methods are missing to assess the correct and complete haplotype separation.

Assemblies need to be thoroughly evaluated to identify the ones that provide the best representation of the genome. The most basic assembly statistics consist in assembly size and contiguity. The assembly size should be close to the expected genome size, or the expected haploid genome size for a collapsed assembly. The contiguity is usually measured with the N50, which represents the largest fragment length in an assembly for which 50% of the assembly is represented in fragments of longer or equal length. This metric has also been declined as N60, N70, N80 and N90 for 60%, 70%, 80% and 90%. Subsequent evaluations look into the completeness of the assembly. One approach consists in searching for a set of orthologs based on the taxon that the species belongs to, notably with the Benchmarking Sets of Universal Single-Copy Orthologs (BUSCO) [8]. Another popular completeness assessment consists in comparing a set of $k$ length sequences (or $k$-mers) extracted from a high-accuracy sequencing dataset with $k$-mers in the assembly, as is done by the K-mer Analysis Toolkit [9].

More recently, Merqury [10] expanded on KAT by proposing a QV score to quantify the quality of an assembly, and an evaluation of phased assemblies based on parental data. By default, these tools do not provide statistics targeted for *de novo* phased assemblies.

In this study, we benchmarked five assemblers on their ability to yield high-quality phased and collapsed assemblies, namely Flye [11], hifiasm [12], NextDenovo [13], PECAT [6] and Verkko [7]. We tested these tools on ultra-low input PacBio HiFi and Nanopore R10.4 reads generated from the highly heterozygous parthenogenetic nematode species *Plectus sambesii*. A draft genome assembly of 187 Mb with a scaffold N50 of 23.4 kb was generated previously using short Illumina reads [14]. Assemblies were evaluated for their size, contiguity, and completeness, and candidate phased assemblies were selected for further analyses. We propose a comprehensive methodology to assess phased assemblies which repurposes state-of-the-art evaluation programs by collecting haplotype-relevant statistics. BUSCO searches for single-copy orthologs, which are expected only once in a haploid assembly. For a phased assembly, the number of duplicated BUSCO features would be increased, as a diploid assembly should have these orthologs in two copies (on each haplotype). In addition, KAT provides as output the proportion of $k$-mers represented once, twice, and more, in each peak. For a diploid genome, a $k$-mer spectrum should display two peaks: the first one for heterozygous $k$-mers; the second one for homozygous $k$-mers, at twice the multiplicity of the heterozygous peak, as these $k$-mers are found on the two haplotypes. Therefore, a complete diploid phased assembly would have heterozygous $k$-mers represented once, and homozygous $k$-mers represented twice. These expectations for orthologs and $k$-mers can also be extrapolated for polyploid assemblies. To evaluate the quality of diploid assemblies, we collected the number of two-copy BUSCO orthologs, the percentage of heterozygous $k$-mers represented once in the assembly, and the percentage of homozygous $k$-mers represented twice. Finally, we purged haplotigs in all assemblies to generate collapsed assemblies. Our analysis provides an overview of strategies to yield phased and collapsed *de novo* genome assemblies from highly accurate long reads and can serve as a guideline for genome assembly projects of non-model species.

# Material & Methods

## Long-read sequencing

High-molecular-weight DNA was extracted from *Plectus sambesii* as described in [15]: with a salting-out protocol on a few individuals for PacBio sequencing; and from plates of cultured worms with a CTAB-phenol-chloroform protocol for Nanopore sequencing. The PacBio HiFi library was prepared using the Express 2.0 Template kit (Pacific Biosciences, Menlo Park, CA, USA) and sequenced on a Sequel II/Sequel IIe instrument with 30 hours movie time. The Nanopore library was prepared with the LSK114 kit and sequenced on two MinION flowcells.

Reads were basecalled using Dorado v0.3.1 [16] in duplex mode with model dna_r10.4.1_e8.2_400bps_supv4.2.0 and the reads were converted to fastq using SAMtools v1.6 [17] with the module `samtools fastq`.

## Quality control

Quality and length of Nanopore reads and PacBio HiFi reads were plotted using NanoPlot v1.42.0 [18]. Genome size and heterozygosity were estimated using jellyfish v2.2.8 [19] and GenomeScope v2.0 [20] with $k$=21.

## *De novo* assembly

Nanopore reads were assembled using Flye v2.9 [11] with parameter `--nano-hq` or `--nano-corr`, NextDenovo v2.5 [13] with parameters `read_type=ont input_type=corrected genome_size=250m` and PECAT v0.0.3 [6] with the ONT config file and parameter `genome_size=250000000`. Nanopore Q20 reads were selected using chopper v0.5.0 with parameter `-q 20` and assembled using Flye v2.9 [11] with parameter `--nano-corr` or `--pacbio-hifi`, hifiasm v0.19 [12] with parameter `-l 0` or `-l 3` (for purging level), NextDenovo v2.5 [13] with parameter `genome_size=250m` and either `read_type=ont input_type=corrected` or `read_type=hifi` and PECAT v0.0.3 [6] with the ONT config file and parameter `genome_size=250000000`. PacBio HiFi reads were assembled using Flye v2.9 [11] with parameter `--pacbio-hifi`, hifiasm v0.19 [12] with parameter `-l 0` or `-l 3` and with and without Nanopore reads longer than 30 kb, NextDenovo v2.5 [13] with parameters `read_type=hifi genome_size=250m` and PECAT v0.0.3 [6] with the HiFi config file and parameter `genome_size=250000000`. For PECAT assemblies, the primary and alternate assemblies were concatenated into a draft phased assembly subsequently labelled as "combined". In addition, reads were filtered using Filtlong v0.2.1 [21] to reduce the dataset to 15 Gb and assembled using Verkko v2.0 [7] (as the assembler cannot handle very large sequencing depths), with the assistance of Nanopore reads longer than 30 kb.

## Decontamination

Bacterial contamination is expected for this species as its feeding source. Two assemblies, (1) Nanopore Flye `--nano-corr` and (2) PacBio HiFi and Nanopore hifiasm `-l 3`, were selected to identify contaminants stemming from the PacBio HiFi and Nanopore datasets. First, the assembly graphs were searched for circular contigs, which would be characteristic of bacterial genomes or the mitochondrion. Then, BUSCO v5.4.7 [8] was run with parameters `-m genome -l metazoa_odb10`, BLAST v2.10.0 [22] was run against the nt database with parameters `-outfmt "6 qseqid staxids bitscore std sscinames scomnames" -max_hsps 1 -evalue 1e-25` and minimap2 v2.24 [23] using the Nanopore reads against (1) with parameters `-ax map-ont` and the PacBio HiFi reads against (2) with parameters `-ax map-hifi`. The outputs were provided to BlobToolKit v4.3.2 [24] and

contigs flagged as Proteobacteria, Actinobacteria and Bacteroidetes were selected and combined with the circular contigs. BLAST v2.10.0 [22] was run for all assemblies against the contaminant dataset with parameters `-max_hsps 1 -evalue 1e-40 -perc_identity 90.0` and hits were removed from the assemblies.

### Assembly evaluation

Basic statistics were computed for all assemblies using assembly-stats v1.0.1 [25]. BUSCO v5.4.7 [8] was run as described previously. For candidate phased assemblies, KAT v2.4.2 was run with the module `kat comp` and the PacBio HiFi dataset as reference. Nanopore and PacBio HiFi reads mapped using minimap2 v2.24 [23] as described previously and the output was provided to purge_dups v1.2.5 [26] to identify haplotigs with thresholds `-l 30 -m 100 -u 800` when using Nanopore reads and `-l 40 -m 155 -u 800` when using PacBio HiFi reads. Mapped reads and haplotigs were then processed by Phased. This tool searches for reads mapping partially to one haplotig and to the alternative haplotig which would suggest a phase switch. Phased outputs the percentage of mapped reads supporting haplotype switches; ideally, the value should be 0% or 0 reads supporting haplotype switches to validate the correctness of a phased assembly.

### Haplotig purging

Haplotigs were purged from all assemblies using the output of minimap2 v2.24 [23] and purge_dups v1.2.5 [26], based on either the Nanopore or the PacBio dataset, to yield collapsed haploid assemblies. These assemblies were evaluated using assembly-stats v1.0.1 [25] and BUSCO v5.4.7 [8].

## Results

Sequencing yielded 35.8 Gb of Nanopore reads with 11.2 Gb over Q20, and 29.7 Gb of PacBio HiFi reads. Quality and length analyses highlight the advantage of Nanopore reads in terms of length, while PacBio HiFi reads have higher quality values (Figure S5). Preliminary analyses of the published Illumina dataset [14] and the newly sequenced PacBio HiFi dataset suggest a haploid genome size of 120-121 Mb with a heterozygosity of 3.80-3.99% (Figure S6). The $k$-mer spectra show two distinct peaks suggesting a diploid genome. The first peak includes heterozygous $k$-mers which are represented on one haplotype, and the second peak, at twice the coverage of the first peak, includes homozygous $k$-mers featured on the two haplotypes.

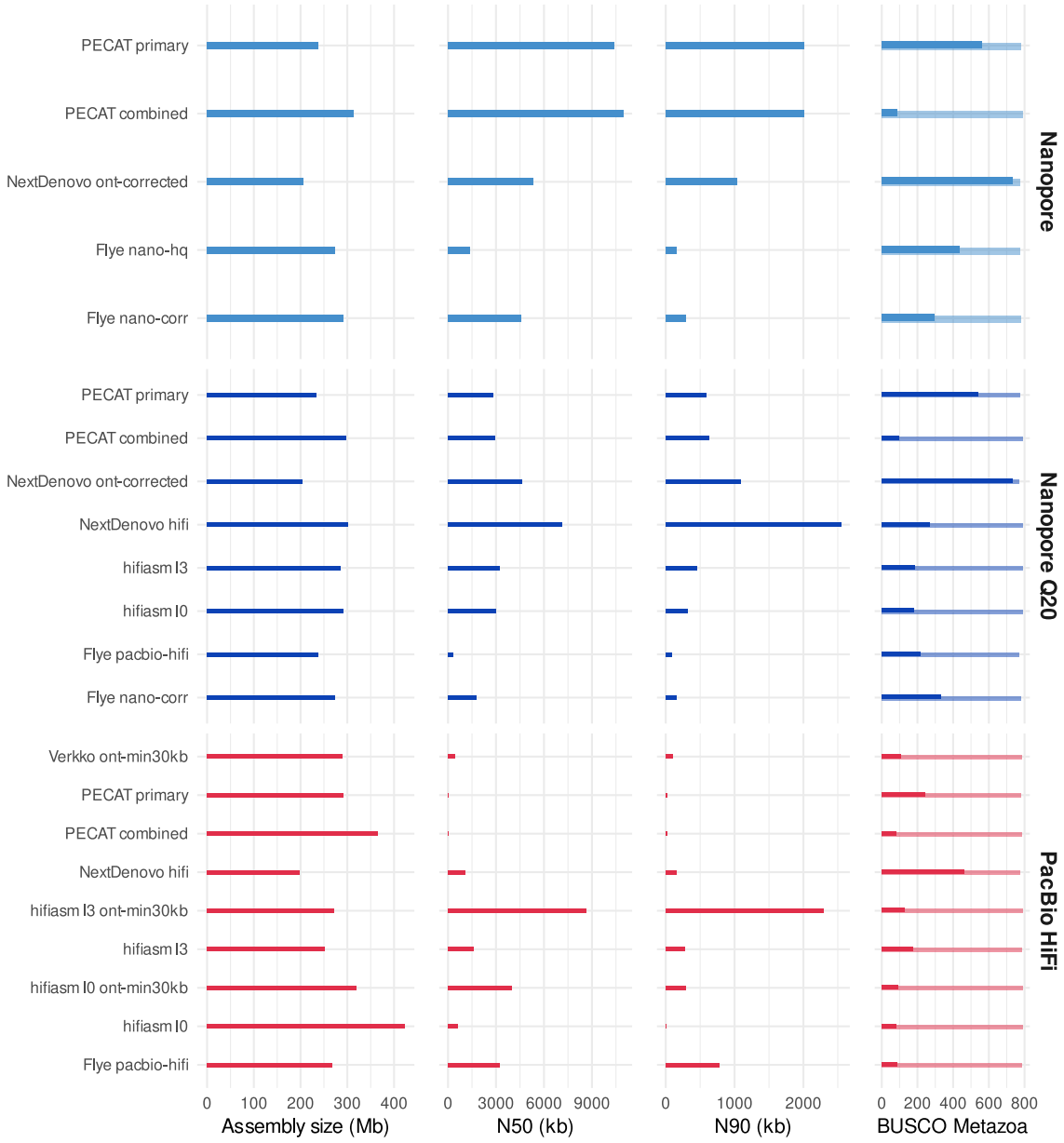



Figure 1: Initial assembly statistics, showing the assembly size, N50, N90 and the number of BUSCO orthologs against the Metazoa lineage.

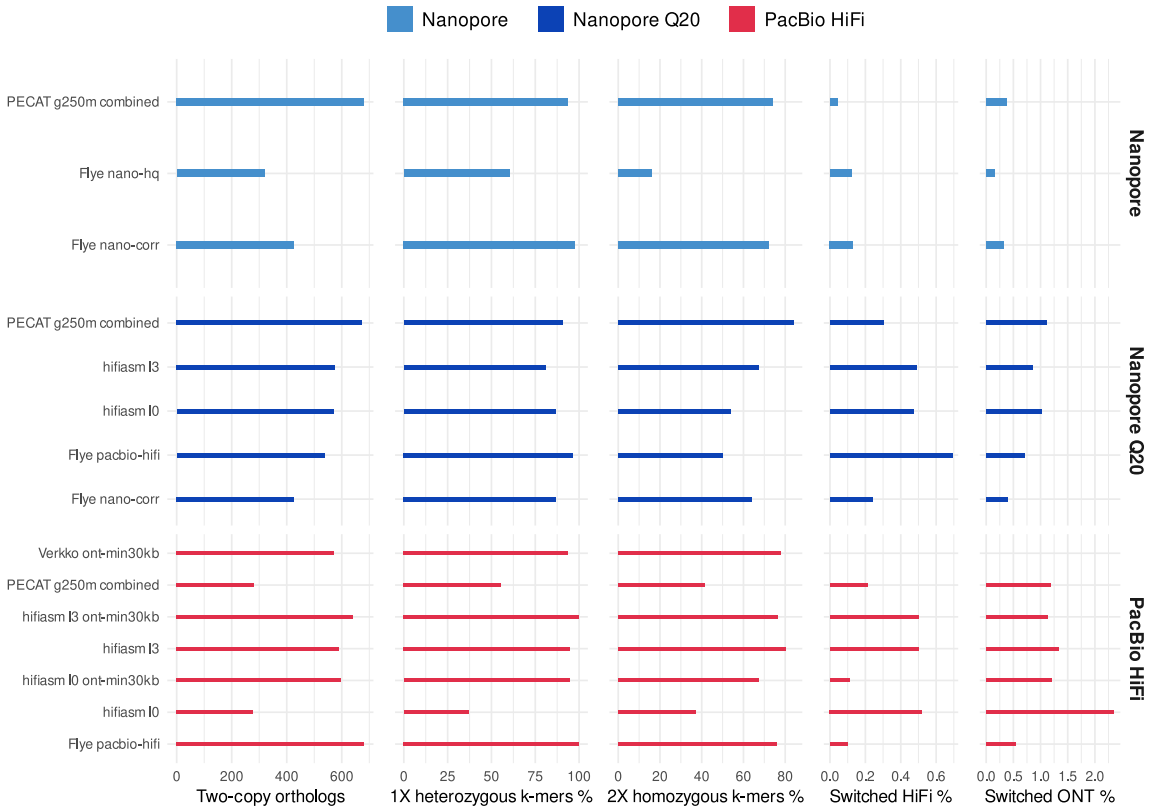



Figure 2: Phased assembly statistics, including the number of two-copy orthologs identified by BUSCO against the Metazoa lineage, the percentage of 1X heterozygous $k$-mers and 2X homozygous $k$-mers, and percentage of PacBio HiFi and Nanopore (ONT) reads supporting haplotype switches.

## Initial assemblies

All assemblies have a size over the expected haploid genome size of 120-121 Mb and close to the expected diploid genome size of 240-242 Mb (Figure 1). NextDenovo assemblies have the smallest genome sizes, as this assembler does not aim to phase assemblies but rather to collapse haplotypes. Flye and hifiasm of PacBio HiFi reads, all Nanopore assemblies, and almost all assemblies of Nanopore Q20 reads (with the exception of the Flye --pacbio-hifi assembly) have reached N50s over 1 Mb. The most contiguous assemblies were obtained using PECAT on Nanopore reads, NextDenovo hifi on Nanopore Q20 reads, and hifiasm on PacBio HiFi and Nanopore reads. Regarding the BUSCO score, all assemblies show similar numbers of recovered orthologs: from 770 to 790 out of 954 expected orthologs. The NextDenovo and primary PECAT assemblies have most of these orthologs in single copies, indicating that they would be partially or mostly collapsed, while the other assemblies have a majority of duplicated orthologs.

## Phased assemblies

Analyses of two-copy orthologs, $k$-mer completeness and haplotype switches provide a deeper evaluation of phasing (Figure 2). PECAT assemblies using Nanopore reads or Nanopore Q20 reads, the Flye assembly using PacBio HiFi reads, and the hifiasm -l 3 assembly using PacBio HiFi reads and Nanopore reads over 30 kb have the highest numbers of two-copy orthologs (681, 673, 681 and 641 respectively). Some assemblies provide a complete representation of heterozygous $k$-mers with 100.00% of 1X heterozygous $k$-mers for the Flye assembly of PacBio HiFi reads and the hifiasm -l 3 assembly of PacBio HiFi reads and Nanopore reads $>$ 30 kb. The percentage of 2X homozygous $k$-mers is overall lower, with up to 83.97% for the PECAT assembly of Nanopore Q20 reads and 80.49% for the hifiasm -l 3 of PacBio HiFi reads. $k$-mer spectra show that in general, Nanopore assemblies tend to have more missing heterozygous $k$-mers and 1X homozygous $k$-mers, while PacBio HiFi assemblies have more artefactual duplications, displayed as 2 to 5X heterozygous $k$-mers and 3-6+X homozygous $k$-mers (Figure S7-S9). Assemblies of Nanopore reads using Flye --nano-hq and PacBio HiFi reads using PECAT and hifiasm -l 0 achieved the lowest values of ortholog and $k$-mer completeness. Interestingly, the hifiasm -l 0 assembly underperfoms due to extensive artefactual duplications, yet it is greatly improved when combined with Nanopore reads over 30 kb. The PECAT assembly of PacBio HiFi reads also suffers from erroneous duplications. Regarding the percentage of switched long reads detected by Phased, the results do not indicate major errors in the assemblies, as the highest value reaches 2.35% of Nanopore reads supporting haplotype switches for the hifiasm -l 0 assembly of PacBio HiFi reads. The lowest values were obtained for all Nanopore assemblies, the Flye --nano-corr assembly of Nanopore Q20 reads, and the Flye --pacbio-hifi assemblies of PacBio HiFi reads. Identification of haplotigs with purge_dups failed for Verkko, thus haplotype switches could not be identified.

## Purged assemblies

Whether the assemblies were purged based on Nanopore or PacBio HiFi data does not affect the results overall in terms of N50, N90 and BUSCO score (Figures 3, 4). The main difference lies in the collapsed assembly sizes, which are consistently close to 150 Mb for assemblies purged based on PacBio HiFi reads, while assemblies purged based on Nanopore reads have more variability, up to a size of 198 Mb for NextDenovo ont-corrected. N50s and N90s are higher for assemblies purged with PacBio HiFi reads, but these values depend on the assembly size and the smaller values for assemblies purged using Nanopore reads may be owed to the larger assembly sizes. The Verkko assembly stands out as not collapsed, regardless of the dataset employed for purging, with unchanged size and number of BUSCO orthologs. Haplotig purging was suboptimal for PECAT PacBio HiFi assemblies, which have a larger amount of remaining duplicated orthologs (124 to 156 against 15 to 29 for other assemblies). The remaining assemblies all have low amounts of duplicated BUSCO orthologs and seem properly collapsed. Several

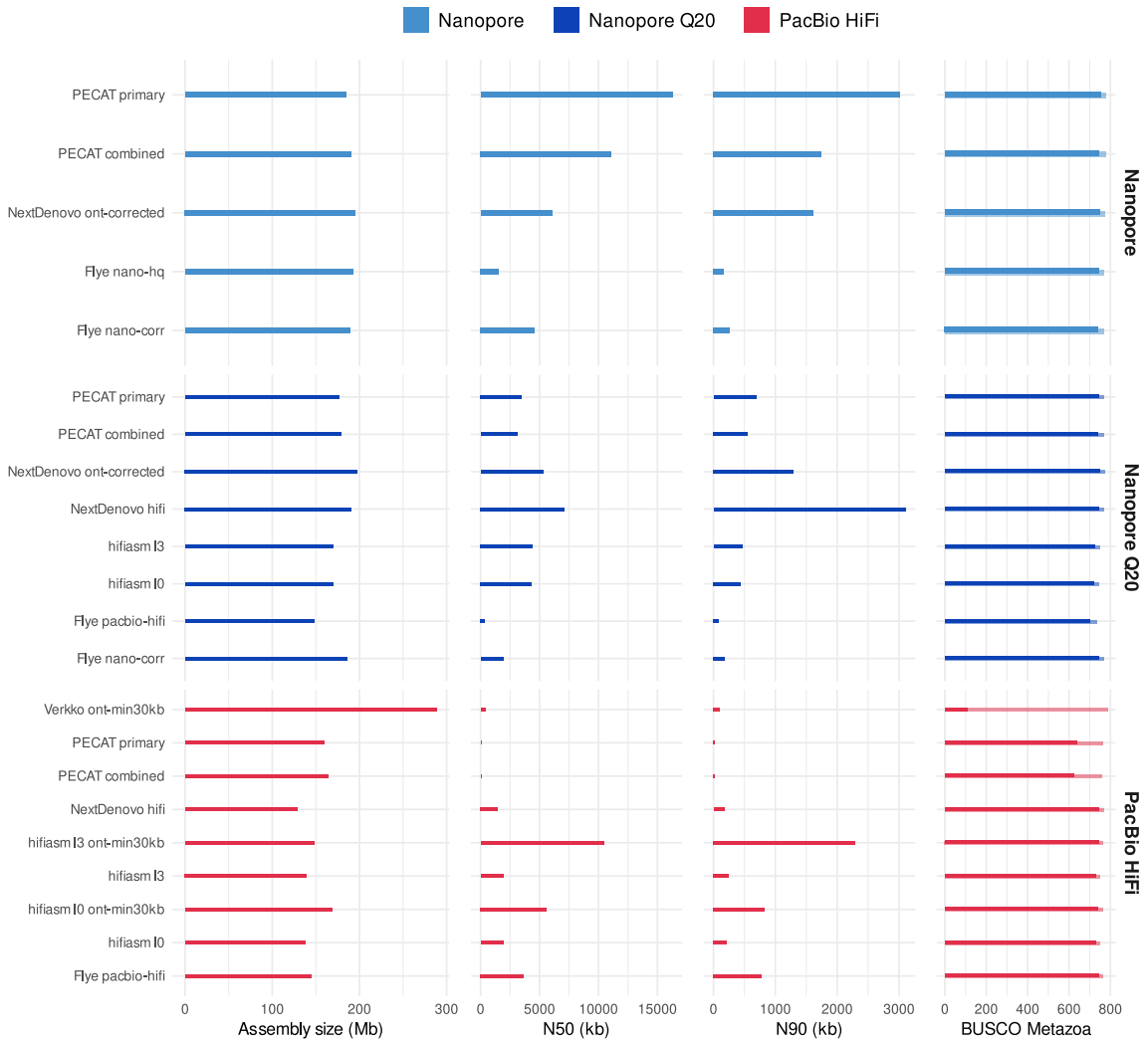



Figure 3: Assembly statistics after purging based on Nanopore reads, showing the assembly size, N50, N90 and the number of BUSCO orthologs against the Metazoa lineage.

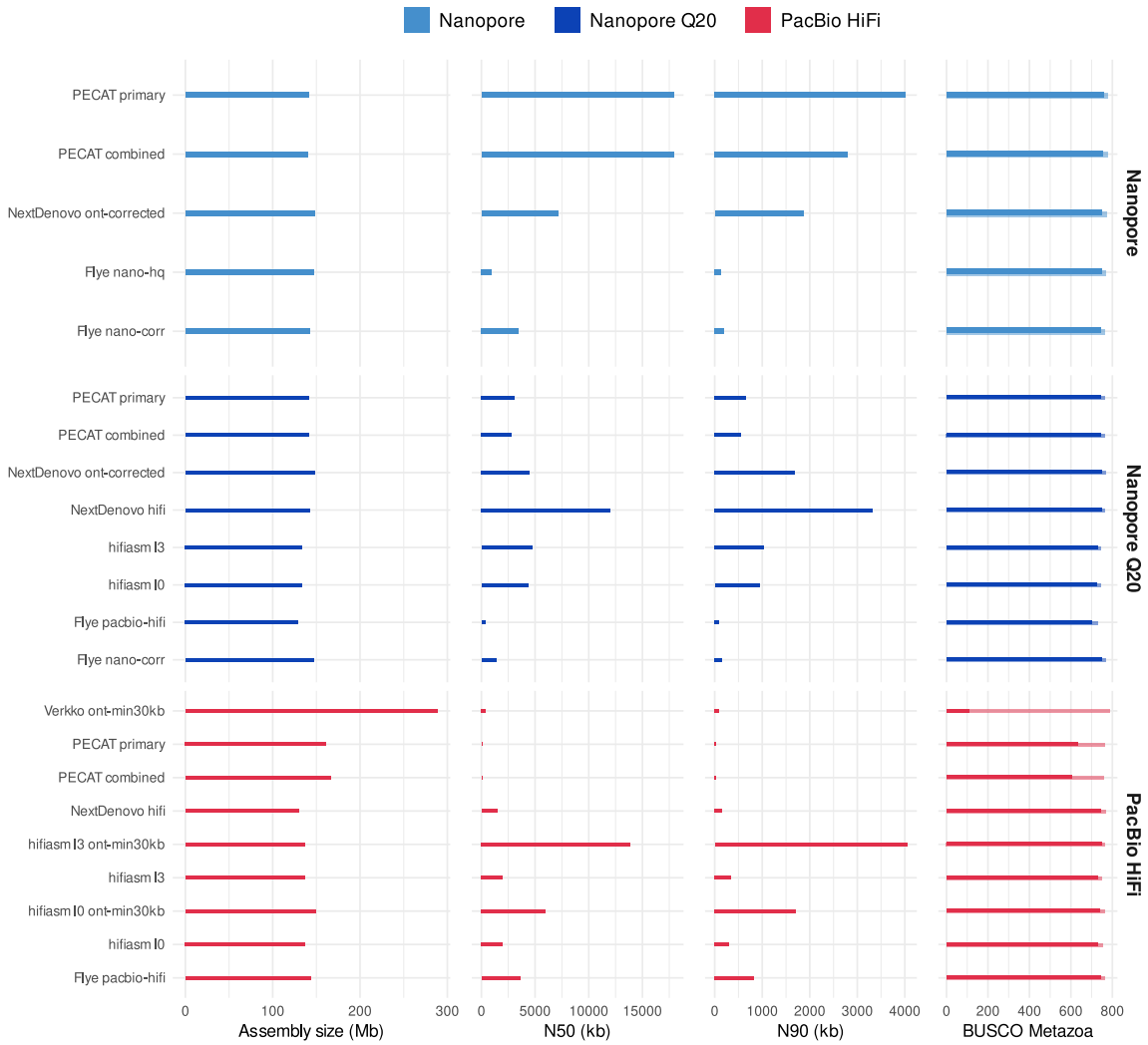



Figure 4: Assembly statistics after purging based on PacBio HiFi reads, showing the assembly size, N50, N90 and the number of BUSCO orthologs against the Metazoa lineage.

assemblies reach an N50 over 5 Mb and an N90 over 1 Mb, namely PECAT and NextDenovo on Nanopore reads, NextDenovo on Nanopore Q20 reads, hifiasm on PacBio HiFi reads assisted with Nanopore reads longer than 30 kb. The largest N50s are reached by: PECAT primary (11.1 Mb) and combined (16.3 Mb) using Nanopore reads and hifiasm -l 3 (10.5 Mb) using PacBio HiFi reads and Nanopore reads over 30 kb, after purging based on Nanopore reads; PECAT primary (17.9 Mb) and combined (17.9 Mb) using Nanopore reads, NextDenovo hifi (12.0 Mb) using Nanopore Q20 reads, hifiasm -l 3 (13.9 Mb) using PacBio HiFi and Nanopore reads over 30 kb, after purging based on PacBio HiFi reads.

## Discussion

Low-accuracy long reads were already commonly employed to yield high-quality collapsed assemblies for a large variety of species. High-accuracy long reads took genome assemblies one step further by enhancing collapsed assemblies and making phased assemblies an achievable goal. Our study shows that both PacBio HiFi and Nanopore R10.4 reads can yield complete haploid and diploid assemblies, and the two technologies can be used complementarily to improve contiguity and completeness. Although PacBio HiFi reads were generated using an ultra-low input protocol, while Nanopore reads were obtained from a large DNA input, no particular impact was observed on the results.

Initial assemblies resulted in draft sequences of variable length and contiguity, with similar overall BUSCO completeness. Assemblers generally performed similarly on the two datasets, with the exception of PECAT; this tool produced contigs of high contiguity and completeness when run on Nanopore reads, but led to a poor contiguity with PacBio HiFi reads.

Assemblies with larger numbers of duplicate BUSCO orthologs were selected as phased candidates and assessed for haplotype separation by examining the number of two-copy orthologs, the $k$-mer completeness with regards to heterozygous $k$-mers and homozygous $k$-mers, and the percentage of long reads supporting haplotype switches. PECAT generated promising diploid assemblies using Nanopore reads with high ortholog and $k$-mer completeness, and a low amount of long-read haplotype switches. hifiasm and Flye also yielded high-quality phased assemblies with adequate parameters, although haplotype switches were slightly more frequent for hifiasm assemblies. Verkko showed good completeness values, but could not be assessed for its correctness. Overall, incorporating Nanopore reads in assemblies led to lower percentages of long reads supporting haplotype switches, which could be owed to a better association of alleles using the longest Nanopore reads. PacBio HiFi reads resulted in assemblies with the best $k$-mer completeness, although this should be nuanced as these reads were used as the reference. The two long-read datasets seem to complement each other to yield more qualitative phased assemblies; for

instance, hifiasm was run using either the Nanopore Q20 reads, or the PacBio HiFi reads, or both PacBio HiFi and Nanopore reads, and the third case led to the best values for contiguity and completeness.

Draft assemblies were efficiently purged using either the Nanopore or the PacBio HiFi reads, with the exception of the Verkko assembly. Results were similar with both datasets, although assembly sizes were more consistent when PacBio HiFi reads were employed. The collapsed assemblies generally had a low amount of duplicated orthologs as expected. The most contiguous assemblies were obtained from the draft produced by PECAT using Nanopore reads, NextDenovo using Nanopore Q20 reads, and hifiasm using Nanopore and PacBio HiFi reads.

Both phased and collapsed assemblies highlight the benefit of combining Nanopore and PacBio HiFi reads to yield contigs and purged scaffolds of high quality. By incorporating Hi-C in the pipeline for scaffolding, we can expect to obtain chromosome-level haploid and haplotype-resolved assemblies.

## Data & Code availability

Datasets were deposited on the European Nucleotide Archive under the Bioproject PRJEB74285. Command lines are provided at github.com/nadegeguiglielmoni/phasing_or_purging Phased is available at github.com/nadegeguiglielmoni/Phase

## Funding

This project was supported through a DFG Emmy Noether Program (ENP) Projekt (434028868) to PHS. NG's position was first funded through a Deutsche Forschungsgemeinschaft (DFG) grant (458953049) to PHS and subsequently through the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101110569.

## Acknowledgements

We thank Laura I. Villegas for her contribution to long-read sequencing and the Genomics and Transcriptomics Laboratory for generation of PacBio HiFi sequencing data.

## References

[1] Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37**, 1155—-1162 (2019).

[2] Sereika, M. *et al.* Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature Methods* **19**, 823–826 (2022).

[3] Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**, 338–345 (2018).

[4] Yu, W. *et al.* Comprehensive assessment of 11 de novo HiFi assemblers on complex eukaryotic genomes and metagenomes. *Genome Research* (2024).

[5] Guiglielmoni, N., Houtain, A., Derzelle, A., Van Doninck, K. & Flot, J.-F. Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics* **22**, 303 (2021).

[6] Nie, F. *et al.* *De novo* diploid genome assembly using long noisy reads via haplotype-aware error correction and inconsistent overlap identification. *bioRxiv* (2023).

[7] Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnology* 1–9 (2023).

[8] Manni, Mosè and Berkeley, Matthew R and Seppey, Mathieu and Simão, Felipe A and Zdobnov, Evgeny M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* **38**, 4647–4654 (2021).

[9] Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2016).

[10] Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 1–27 (2020).

[11] Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540–546 (2019).

[12] Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 1–6 (2021).

[13] NextOmics. NextDenovo, `https://github.com/Nextomics/NextDenovo` (2019).

[14] Rošić, S. *et al.* Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity. *Nature Genetics* **50**, 452–459 (2018).

[15] Guiglielmoni, N., Villegas, L. I., Kirangwa, J. & Schiffer, P. H. Revisiting genomes of non-model species with long reads yields new insights into their biology and evolution. *Frontiers in Genetics* **15**, 1308527 (2024).

[16] Oxford Nanopore Technologies. Dorado, `https://github.com/nanoporetech/dorado` (2022).

[17] Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10** (2021). Giab008.

[18] De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).

[19] Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

[20] Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432 (2020).

[21] Wick, R. R. Filtlong, `https://github.com/rrwick/Filtlong` (2017).

[22] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

[23] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

[24] Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. Blobtoolkit–interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics* **10**, 1361–1374 (2020).

[25] Pathogen Informatics, Wellcome Sanger Institute. assembly-stats, `https://github.com/sanger-pathogens` (2014).

[26] Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

# Supplementary material

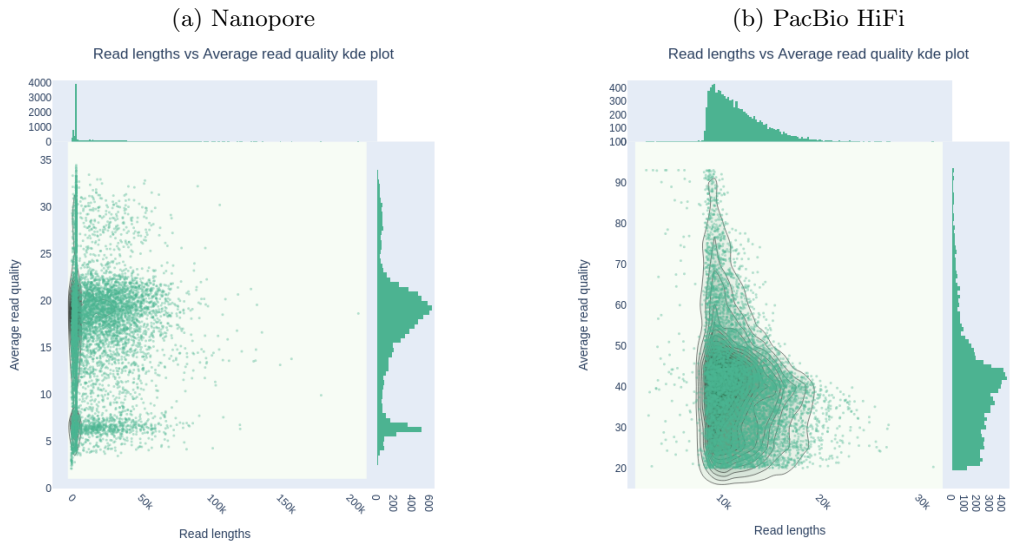



Figure 5: NanoPlot analysis of the Nanopore and PacBio HiFi reads.

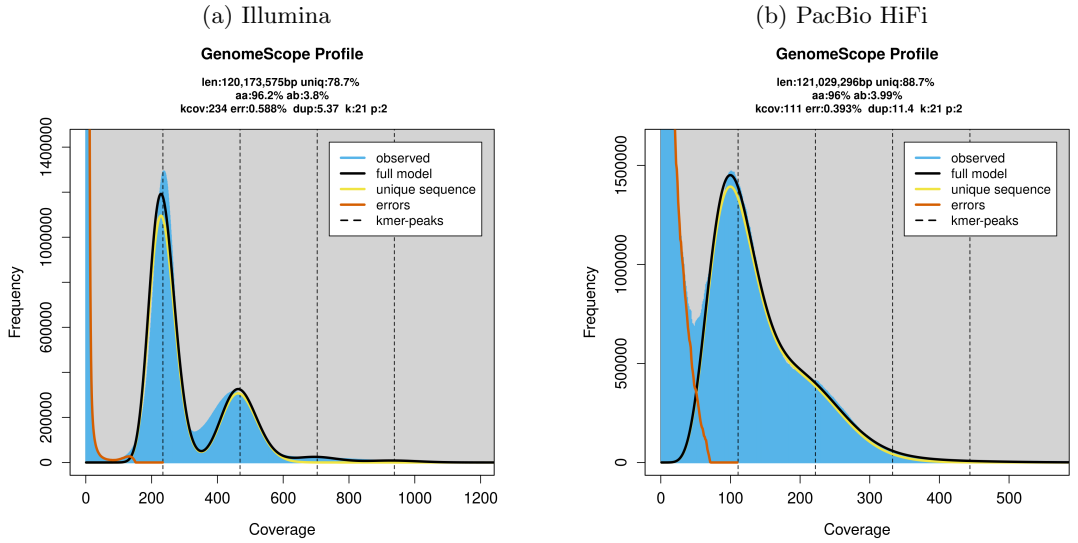
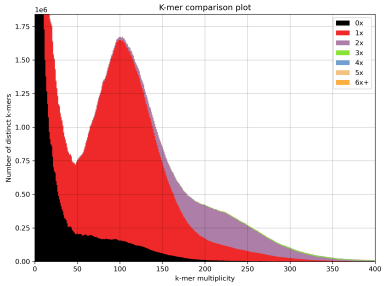
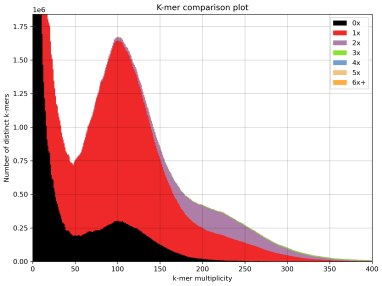



Figure 6: GenomeScope analysis of the $k$-mers in the Illumina and PacBio HiFi datasets with $k$=21.

(a) Flye –nano-corr




(b) Flye –nano-hq




(c) PECAT

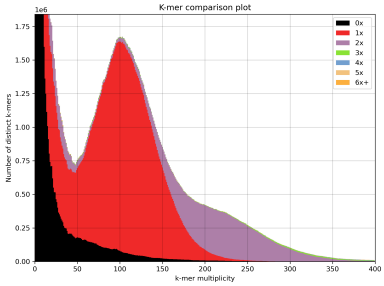
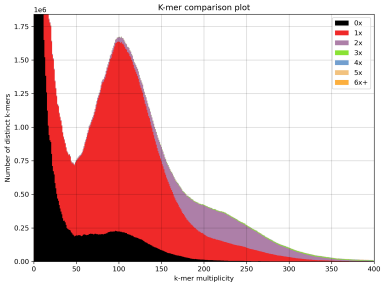
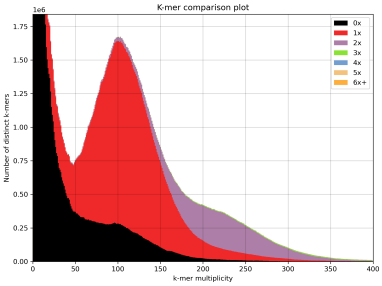



Figure 7: KAT analysis of Nanopore assemblies.

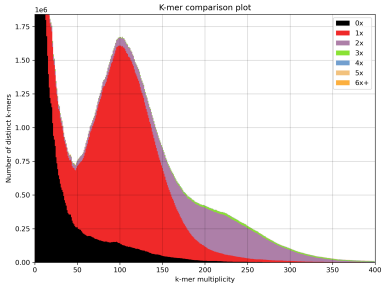
(a) Flye –nano-corr



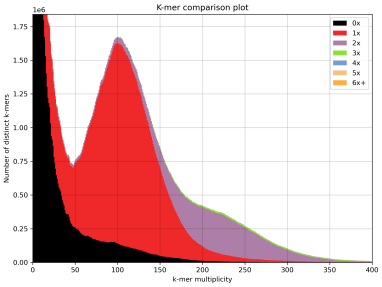

(b) Flye –pacbio-hifi




(c) hifiasm -l 0




(d) hifiasm -l 0 Nanopore 30+ kb




(e) PECAT

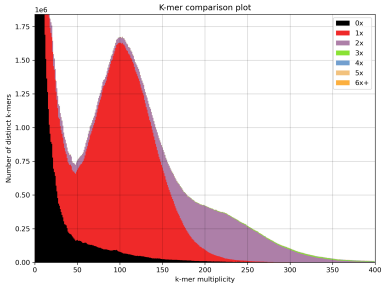
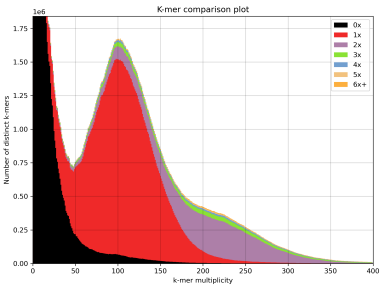
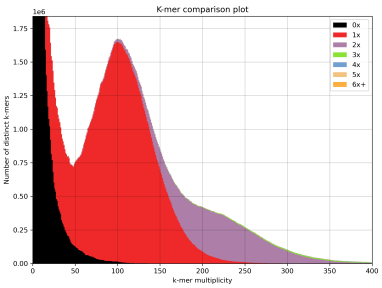



Figure 8: KAT analysis of Nanopore Q20+ assemblies.

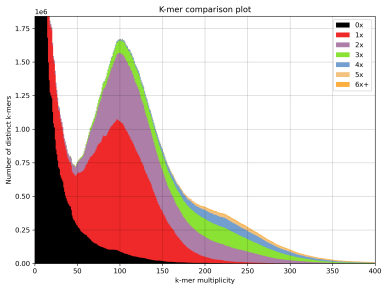
(a) Verkko Nanopore 30+ kb



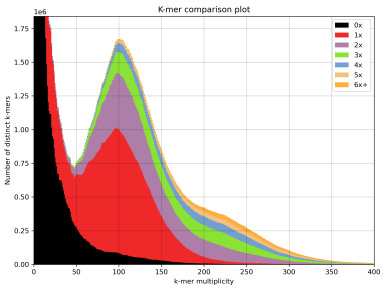

(b) Flye



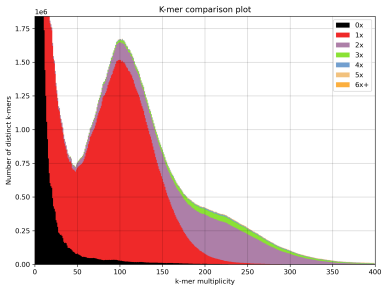

(c) PECAT



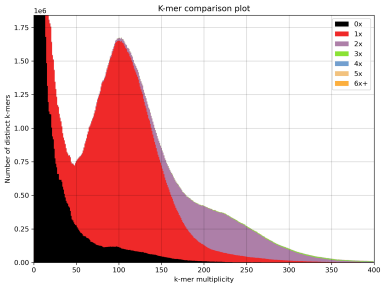

(d) hifiasm -l 0




(e) hifiasm -l 0 Nanopore 30+ kb




(f) hifiasm -l 3




(g) hifiasm -l 3 Nanopore 30+ kb

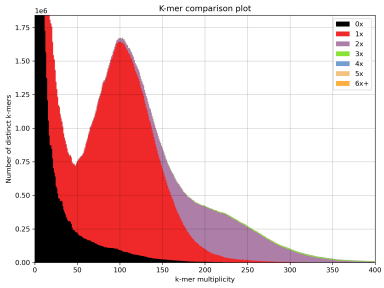



Figure 9: KAT analysis of PacBio HiFi assemblies.