OXFORD

Briefings in Bioinformatics, 2022, **23(3)**, 1–13

https://doi.org/10.1093/bib/bbac146

Problem Solving Protocol

Benchmarking of long-read sequencing, assemblers and polishers for yeast genome

Xue Zhang , Chen-Guang Liu , Shi-Hui Yang, Xia Wang, Feng-Wu Bai and Zhuo Wang

Corresponding authors: Chen-Guang Liu, Zhuo Wang, State Key Laboratory of Microbial Metabolism, Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders of the Ministry of Education, Joint International Research Laboratory of Metabolic & Developmental Science of the Ministry of Education, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China. Tel: 86-21-34205125; Fax: 86-21-34208028; E-mail: cg.liu@sjtu.edu.cn (C.-G. Liu); Tel: 86-21-62933338; Fax: 86-21-62932059; E-mail: zhuowang@sjtu.edu.cn (Z. Wang)

Abstract

Background: The long reads of the third-generation sequencing significantly benefit the quality of the *de novo* genome assembly. However, its relatively high single-base error rate has been criticized. Currently, sequencing accuracy and throughput continue to improve, and many advanced tools are constantly emerging. PacBio HiFi sequencing and Oxford Nanopore Technologies (ONT) PromethION are two up-to-date platforms with low error rates and ultralong high-throughput reads. Therefore, it is urgently needed to select the appropriate sequencing platforms, depths and genome assembly tools for high-quality genomes in the era of explosive data production.

Methods: We performed 455 (7 assemblers with 4 polishing pipelines or without polishing on 13 subsets with different depths) and 88 (4 assemblers with or without polishing on 11 subsets with different depths) *de novo* assemblies of Yeast S288C on high-coverage ONT and HiFi datasets, respectively. The assembly quality was evaluated by Quality Assessment Tool (QUAST), Benchmarking Universal Single-Copy Orthologs (BUSCO) and the newly proposed Comprehensive_score (C_score). In addition, we applied four preferable pipelines to assemble the genome of nonreference yeast strains.

Results: The assembler plays an essential role in genome construction, especially for low-depth datasets. For ONT datasets, Flye is superior to other tools through C_score evaluation. Polishing by Pilon and Medaka improve accuracy and continuity of the preassemblies, respectively, and their combination pipeline worked well in most quality metrics. For HiFi datasets, Flye and NextDenovo performed better than other tools, and polishing is also necessary. Enough data depth is required for high-quality genome construction by ONT (>80X) and HiFi (>20X) datasets.

Keywords: de novo assembly, long-read sequencing, benchmarking, yeast, data depth, genome analysis

Introduction

The high-throughput and long-reads of next-generation sequencing technologies enabled the sequencing of entire genomes at an unprecedented speed, which has revolutionized biology in the past decades not only for laboratory research but for people's daily life [1]. The third-generation sequencing (TGS), Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are real-time, long-reads generated, single-molecule sequencing platforms, which can overcome the shortcomings of second-generation sequencing (SGS) technology such as relatively short reads, sequence-dependent biases, information loss [1], that greatly improved the continuity of *de novo* genome assembly [2].

Lately, high-throughput high-fidelity (HiFi) reads, obtained by PacBio sequel II system with the circular consensus sequencing mode, owned long reads (>10 kb) and high per-base accuracy (>99.9%). Unlike SGS and PacBio's optical monitoring systems that rely on DNA polymerase to read base sequences, ONT sequencing identifies DNA bases by measuring the changes in electrical conductivity generated as DNA strands pass through a biological pore that makes it generate ultralong reads. The most cutting-edge representation is PromethION platform that can produce 7 Tb reads per run with an average sequencing speed of ∼430 bases/s and N50 > 20 kb [3], which makes it possible to achieve data quickly and overcome the confusion of repetitive regions to

Xue Zhang is a Ph.D. candidate in the School of Life Sciences and Biotechnology at Shanghai Jiao Tong University. Her research interests are in comparative genomic analysis and integrative omics analysis.

Chen-Guang Liu is an Associate Professor in the School of Life Sciences and Biotechnology at Shanghai Jiao Tong University. His research interests are in biorefinery from lignocellulose and metabolic engineering.

Shi-Hui Yang is a Professor in the School of Life Sciences at Hubei University. His research interests are in metabolic engineering, synthetic biology, and renewable bioproducts.

Xia Wang is an Assistant Professor in the School of Life Sciences at Hubei University. Her research interests are in metabolic engineering, synthetic biology, and renewable bioproducts.

Feng-Wu Bai is a Professor in the School of Life Sciences and Biotechnology at Shanghai Jiao Tong University. His research interests are in biorefinery, bioprocess engineering, and metabolic engineering.

Zhuo Wang is an Associate Professor in Bio-X Institutes, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. Her research interests are in bioinformatics, network modeling and integrative omics analysis.

Received: November 27, 2021. **Revised:** March 26, 2022. **Accepted:** March 31, 2022

construct a continuous high-quality genome sequence, especially in the assembly of complex genomes such as high heterozygosity, high repetition and large genomes.

Although long reads do beneficial to genome construction, the single base error rate of ONT and HiFi sequencing is still higher than SGS. There are mainly two strategies for reducing errors in ONT genome assemblies; the first is to correct the random errors of long reads before genome assembly through high coverage, and the second is to polish the draft sequence after assembly which we called 'polishing' step.

Recently, a variety of new genome assembly and polishing methods have emerged. Canu is a successor of Celera assembler and introduces the adaptive overlapping strategy based on *tf-idf* weighted MinHash and the sparse assembly graph construction that avoids collapsing diverged repeats and haplotypes [4]. It was lately modified for HiFi reads and named HiCanu [5], which has homopolymer compression, overlap-based error correction and aggressive false overlap filtering steps. NECAT is an error correction and *de novo* assembly tool for ONT long noisy reads [6]. Similar to Canu, it first corrects the raw reads and then establishes the assembly, but Necat uses an adaptive read selection and two-step progressive method to quickly correct ONT reads to high accuracy. NextDenovo [7] is a string graph-based *de novo* assembler for multitype long reads such as Continuous Long-Read (CLR), HiFi and ONT. It uses the 'correct-then-assemble' strategy for ONT reads but no correction step for HiFi reads. Flye [8] is also compatible with ONT and HiFi reads and first generates the error-prone disjointings, then concatenates all disjointings into a single string in an arbitrary order and constructs the assembly graph and finally resolves the graph to obtain the accurate contigs. Hifiasm [9] is a recently published assembler that takes advantage of HiFi reads to faithfully resolve the haplotype information in a phased assembly graph. Miniasm constructs the string graph by overlapping a set of reads. It only performs the layout step of the overlap–layout-–consensus algorithm, which is different from Canu or Necat. Unlike the above long-read assemblers, Unicycler [10] can use short-read-only, long-read-only or hybrid reads for assembly. In the hybrid assembly process, it builds an initial assembly graph from short reads using SPAdes and then simplifies the graph using information from short and long reads. Pilon [11] and Racon are polishers that use short reads and long reads for correction, respectively. Racon is recommended to correct Miniasm's draft assembly [12], so we also design the ONT reads assembly pipeline named MiniRacon that combines the Miniasm and third round Racon iteration. Medaka is the first neural network-based polishing tool developed by ONT. NeuralPolish [13] is the latest polisher based on alignment matrix construction and orthogonal Bi-directional Gated Recurrent Unit (Bi-GRU) networks.

Previous work on the comparison of genome assembly tools of ONT reads mostly used prokaryotes, such as *Escherichia coli* [14], virus [15] and pathogenic bacteria [16, 17], and mostly used simulated datasets to construct low-quality sequences. Giordano *et al.* [18] used yeast as sequencing material for *de novo* assembly comparison in 2017, but the sequencing depth of the ONT datasets was very low, only ∼30X and the tools used were outdated, which could not guide the choice of advanced tools for the genome construction. Recently, the comparison of assembly on HiFi reads was carried out for *E. coli*, *Drosophila ananassae* [19] and rice [20]. However, HiFi has not been conducted and evaluated in yeast, not to mention the comparison of all available assemblers. Therefore, to select the advanced genome-build pipeline and appropriate sequencing depth for eukaryotes remains urgently to be investigated.

In this study, we comprehensively evaluated the influence of sequencing methods, assembly tools, polishing tools and sequencing depth on eukaryotic genome construction with high-coverage ONT, HiFi and BGISEQ paired-end datasets, using model organism yeast as the representative. This is not only the first time across the wide coverage as 800X on the ONT dataset but also the first HiFi dataset release of yeast. For the ONT dataset, the combination of 7 assembly methods (Canu, Flye, Necat, Miniasm, MiniRacon, NextDenovo and Unicycler) and 4 polishing methods (Medaka, Pilon, NeuralPolish and Medaka_Pilon) was carried out on 13 different depth subsets. While for the HiFi dataset, 4 latest tools, hifiasm, NextDenovo, Flye, and HiCanu, were applied on 11 different depth subsets with or without Pilon polish. The genome quality and computing performance of each pipeline was evaluated to determine the appropriate workflow and sequencing depth. In addition, we also sequenced and assembled two different genera of yeast strains, *Saccharomyces cerevisiae* CICC-1445 (SC) and *Schizosaccharomyces pombe* FLO-DUT (SP), to retrieve the high-quality genome by using the outstanding workflows. The advanced benchmark tests of updated tools based on high-coverage ONT and HiFi datasets provide valuable guidance for TGS genome construction of mold, microalgae and even complex genomes, such as human, to be exceptionally useful in understanding the fundamental mechanism of eukaryotic.

Materials and methods

Yeast strain and growth conditions

Saccharomyces cerevisiae strain S288C (ATCC 204508) and *S. cerevisiae* CICC-1445 were purchased from the American Type Culture Collection (ATCC) and China Center of Industrial Culture Collection (CICC), respectively; *S. pombe* FLO-DUT was preserved by our laboratory. The yeast strain was precultured on Yeast extract Peptone Dextrose medium (YPD) plates (10 g/l yeast extract, 20 g/l peptone, 20 g/l glucose and 20 g/l agar) at 30 °C for 24 h. A single colony was activated and then cultured in YPD media at 30 °C and 150 rpm and sampled when cells were growing in the log phase with Optical Density $(OD)_{600}$= 0.8∼0.9.

Genome sequencing and subsets generation from raw full datasets

Genomic DNA of three strains was extracted, tested and sequenced to generate ONT (PromethION) and BGISEQ reads and HiFi reads at BGI (Shenzhen, China) and Personalbio (Shanghai, China), respectively. *Saccharomyces cerevisiae* CICC-1445 and *S. pombe* FLO-DUT have not been sequenced before. Here, we generated the first complete genome sequences of these two strains. All raw data and three assembled genomes through optimal pipelines have been uploaded in the National Center for Biotechnology Information (NCBI) with BioProject accession PRJNA792930, PRJNA792931 and PRJNA792932 for *S. cerevisiae* S288C, *S. cerevisiae* CICC-1445 and *S. pombe* FLO-DUT, respectively.

Detailed statistic information about each sequenced dataset is summarized in Supplementary Tables S1–S3 (see Supplementary Data available online at https://academic.oup.com/bib). S288C has the deepest sequencing depth, about 800X ONT reads with N50 of ∼27 kb, 380X HiFi reads with N50 of ∼21 kb and average passes with 10.29 times and 240X depth of 2 × 150 bp BGISEQ paired reads. The raw ONT and HiFi reads of S288C were mapped to the latest gold-standard reference S288C genome (GCA_000146045.2) by pbmm2 (v1.3.0) [21] to calculate the mean-mapped concordance. Although sequencing depths are different for three strains, their ONT (Supplementary Figure S1A and B, see Supplementary Data available online at https://academic.oup.com/bib) and HiFi (Supplementary Figure S1C and D, see Supplementary Data available online at https://academic.oup.com/bib) datasets have similar read length distribution. To explore the dependence of the assembly pipelines on sequencing depth and the effect of that on the assembly quality, we randomly sampled 13 subsets with gradient depths of 10X, 20X, 40X, 60X, 80X, 100X, 120X, 140X, 160X, 320X, 480X, 640X and 800X for S288C ONT data, 11 subsets with gradient depths of 10X, 20X, 40X, 60X, 80X, 100X, 120X, 140X, 160X, 320X and 380X for S288C HiFi data and 6 subsets with depths of 10X, 20X, 40X, 80X, 160X and 240X for BGISEQ data by Seqtk v1.2. Each ONT or HiFi subset has a similar read length distribution (Supplementary Figure S1E and G, see Supplementary Data available online at https://academic.oup.com/bib) and coincident read length density (Supplementary Figure S1F and H, see Supplementary Data available online at https://academic.oup.com/bib).

De novo assembly and polishing pipelines

De novo assembly and polishing pipelines for our benchmark tests are shown in Figure 1. For 11 HiFi subsets of S288C, we performed 4 assemblers, HiCanu (v2.2), Flye (v2.8.3), NextDenovo (v2.5.0) and hifiasm (v0.16.1) by default settings. For 13 ONT subsets of S288C, we used 7 different tools for long-read assembly: Canu (v1.9), Flye (v2.8.3), Necat (v20200119), Miniasm (v0.3), Miniasm (v0.3)/Racon (v1.4.13), NextDenovo (v2.5.0) and hybrid-read assembler Unicycler (v0.4.8). Default parameters were used except Flye at super-high read depth, '–asm-coverage = 50' was set. The normal mode was used for the hybrid assembly by Unicycler. Then, each assembly was further polished by NeuralPolish, Medaka (v1.0.1) or Pilon (v1.23), respectively, with default parameters. Only ONT reads were used in NeuralPolish and Medaka, while only BGISEQ reads were input in Pilon. The model of 'r941_prom_high_g303' was used in Medaka. We also combined Medaka and Pilon pipelines (Medaka_Pilon) to obtain high-quality genome assembly. When lower than 160X, the depth of short reads was consistent with that of long reads during the assembly process. In other cases, all short reads (240X) were used.

Genome assembly assessment

We used QUAST (v5.0.2) to evaluate the quality of the assemblies generated by different assemblers and polishing tools [22]. All draft assemblies were compared to the latest reference S288C genome. From several metrics, we selected the number of contigs, N50, Mismatches and Indels per 100 kb to visualize in the main text. We also used BUSCO [23] to assess the genome annotation completeness of assemblies. In order to clarify the rate of different error types of NeuralPolish, we used Pomoxis v0.3.9 to report the insertion and deletion error rates separately. Furthermore, we proposed a new Comprehensive_score (C_score) to compositively evaluate the quality of genome assemblies from different pipelines in order to give some advice for needed users, which were described in detail in section 3.2. And we also attach the raw evaluating values from QUAST in the supplemental file for advanced users' information.

We designed the improvement rate (IR) for each metric to evaluate the polishing pipeline through the Equation (1), where *i* means the order of the data depth and *n* means the number of subsets which is 13 for ONT data. The coefficient is +1 or −1 for metrics means that the higher the better (N50) or the smaller the better (Indels, Mismatches, contigs), respectively

$$\mathrm{IR} = (\pm 1) \times \frac{1}{n} \sum_{i=1}^{n} \frac{M_{\mathrm{i\_After}} - M_{\mathrm{i\_Before}}}{M_{\mathrm{i\_Before}}}. \tag{1}$$

Annotation of the assembled genome

De novo genome structure annotation was carried out by Augustus (v3.3.3) [24]. Then, annotated protein sequences were aligned to the S288C reference database by BLASP with a value = 1e−5 and max_target_seqs = 1. We calculated the percentage of annotated genes by Equation (2). The total gene number in reference was 6002. Venn diagram was drawn by VENNY (v2.1) [25] to compare the gene annotation ability of optimal pipelines.

$$\text{Percentage of annotated genes} = \frac{\text{Annotated gene numbers in assembly}}{\text{Total gene number in reference}}. \tag{2}$$

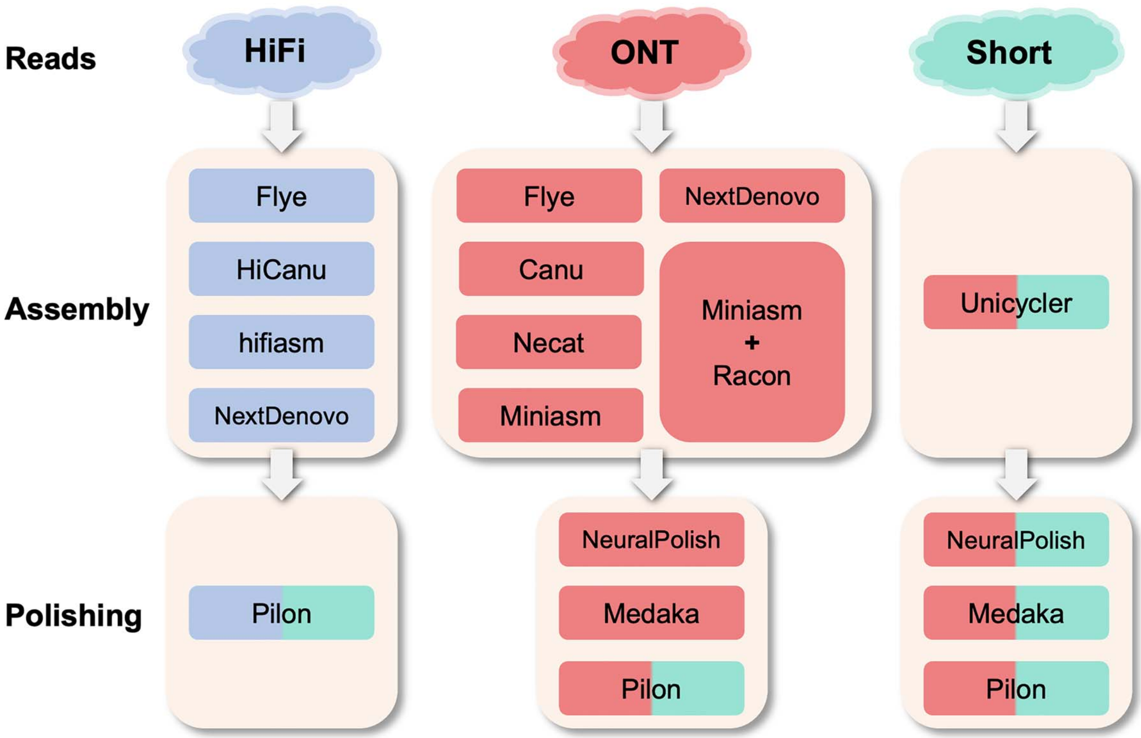

Figure 1. *De novo* assembly and polishing pipelines for benchmarking. The datasets used in each pipeline were represented by colors: Blue, HiFi; Red, ONT; Green, short reads from BGISEQ.

The sets of programs for assembly, polishing, evaluation and annotation were provided in supplemental file.

Results and discussion

De novo assemblers showed significant differences in assembly quality on ONT datasets

We used the reference strain of *S. cerevisiae*, S288C, to compare the capabilities of seven advanced assembly pipelines, Canu, Flye, Necat, Miniasm, MiniRacon, NextDenovo and Unicycler. These assembly pipelines showed significant differences in assembly quality. Genome accuracy especially the Indels per 100 kb was highly variable (Figure 2A). Unicycler has the most outstanding performance on Indels per 100 kb, with only a few errors, followed by MiniRacon, Necat and Flye. In addition to Miniasm, Canu has the most Indels. Most of the assemblers have relatively low Mismatches, while Flye and Necat presented the lowest mismatch number and have low variabilities in different read depths (Figure 2B). As for continuity, Necat and NextDenovo produced draft assemblies that had contig numbers closest to the chromosome numbers of the reference genome (16 chromosomes and 2 plasmids) and Flye has the robust N50 (Figure 2C and D). According to the result of BUSCO (Figure 2E and Supplementary Figure S2, see Supplementary Data available online at https://academic.oup.com/bib), Unicycler has the most complete gene number, followed by MiniRacon, Flye and Necat.

Canu has the least robust contig numbers, with fewer contigs when data depth was below 480X and dramatically increased when data depth was above 480X (Figure 2). Unicycler performed weakest in terms of contiguity but superior in terms of accuracy. Miniasm assembly presents the worst on both accuracy and completeness since it doesn’t include the base error correction and consensus steps. By BUSCO evaluation, it only has 2–11 complete genes, with a proportion of 0.1–0.5% (Supplementary Figure S2, see Supplementary Data available online at https://academic.oup.com/bib), which indicates the loss of correction and consensus during assembly has a great influence on the subsequent sequence characteristic analysis.

As the data depth of the subset increases from 10X to 800X, the computing resources required by all assemblers increase (Figure 2F). Miniasm and NextDenovo consumed the least CPU time, followed by MiniRacon and Necat. Unicycler consumed most computational resources and took 1536 CPU hours with 800X subsets, followed by Canu, while NextDenovo is a storage-consumed assembler and spends most memory (Supplementary Figure S3, see Supplementary Data available online at https://academic.oup.com/bib).

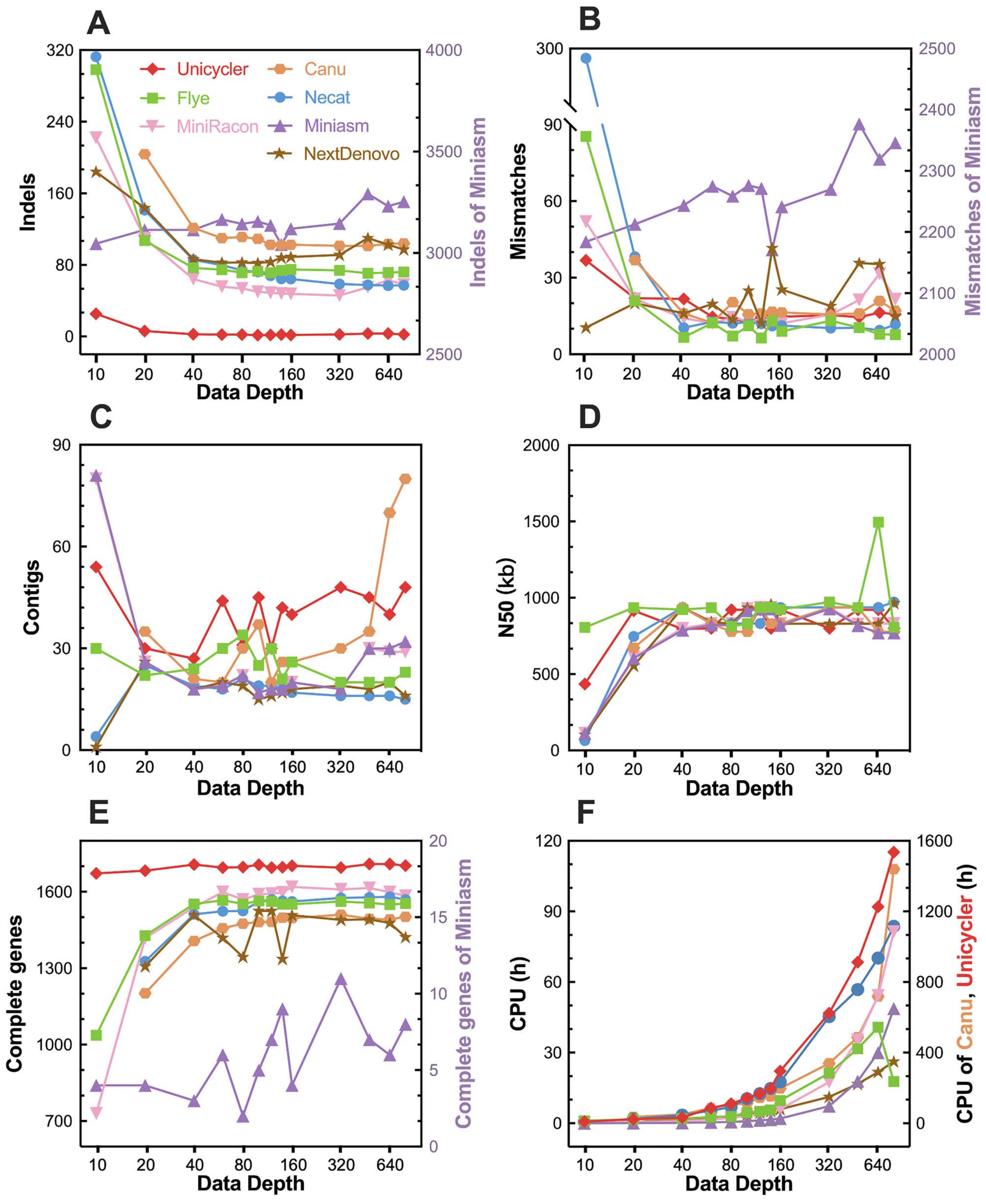

Figure 2. Main metrics of assemblies on 13 ONT subsets with different data depths. Number of Indels (**A**) and Mismatches (**B**) per 100 kb, contig numbers (**C**), length of N50 (**D**), complete genes' number from BUSCO (**E**) and computational time of different assemblers (**F**).

C_score evaluation of assemblers

Here, we integrate 5 metrics to comprehensively evaluate the genome qualities of assemblies from different assemblers, including contig numbers (contigs), the number of mismatches (Mismatches), Indels per 100 kb (Indels), N50 length in kb and the number of complete genes evaluated by BUSCO (completeness). For each metric, the value was scaled to [0, 1] by Min-Max Normalization across different pipelines and named Scaled_Metric (SM) through Equation (3), where M is the mean of all available subset

Table 1. The mean of metric values (M) and comprehensive scores (C_score) of different assemblers on ONT dataset.

	Contigs	N50 (kb)	Mismatches	Indels	Completeness	C_score
Flye	25.0	942.4	16.3	93.1	1506.7	0.904
Necat	17.2	825.4	33.3	91.8	1419.2	0.824
NextDenovo	17.2	798.2	22.3	101.8	1446.2	0.795
MiniRacon	26.5	788.1	19.7	70.9	1512.5	0.713
Canu	35.8	834.9	18.4	114.4	1458.1	0.678
Unicycler	40.3	836.6	17.6	4.4	1697.6	0.677
Miniasm	26.9	770.8	2265.1	3150.6	5.9	0.116

*Each M was calculated by the average of 13 ONT subset metrics (M_i) from 10X to 800X.

metrics (M_i). The M_{min} or M_{max} means the minimum or maximum M of pipelines that are to be compared.

$$SM = \frac{M - M_{min}}{M_{max} - M_{min}}. \tag{3}$$

For high-quality assembly, the metrics N50 and the number of complete genes should be as high as possible, and the metrics contigs, Mismatches, Indels should be as low as possible. Therefore, we define a raw_C_score in Equation (4) by summing up five SMs, whose coefficients were set as 1 for the first two metrics or −1 for the latter three metrics because of the positive and negative contribution of metrics. Then, we rescale the raw_C_score to [0, 1] also by Min-Max Normalization and obtained the C_score Equation (5), where the **raw_C_score**$_{t\_min}$ and **raw_C_score**$_{t\_max}$ mean the minimum and maximum theoretical value of the raw_C_score, which is −3 and 2, respectively

$$\begin{aligned} \text{raw\_C\_score} &= SM_{N50} + SM_{Completeness} - SM_{Contig} \\ &\quad - SM_{Mismatches} - SM_{Indels} \end{aligned} \tag{4}$$

$$\text{C\_score} = \frac{\text{raw\_C\_score} - \left(\text{raw\_C\_score}_{t_{min}}\right)}{\text{raw\_C\_score}_{t_{max}} - \text{raw\_C\_score}_{t_{min}}}. \tag{5}$$

The metric mean of the 13 subsets is used to calculate the C_score of 7 assemblers (Table 1). Flye has the best comprehensive performance, closely followed by Necat. Besides, Unicycler has excellent accuracy despite the fragmented assemblies and long CPU hours.

Influence of polishing process on assembly quality

For each obtained draft assembly of ONT subsets, we obtained a polished assembly by using each run of four polishing pipelines (NeuralPolish, Medaka, Pilon, Medaka_Pilon) and then assessed its performance through QUAST (Supplementary Figure S4, see Supplementary Data available online at https://academic.oup.com/bib) and calculated the main metrics IR of each polishing process (Figure 3).

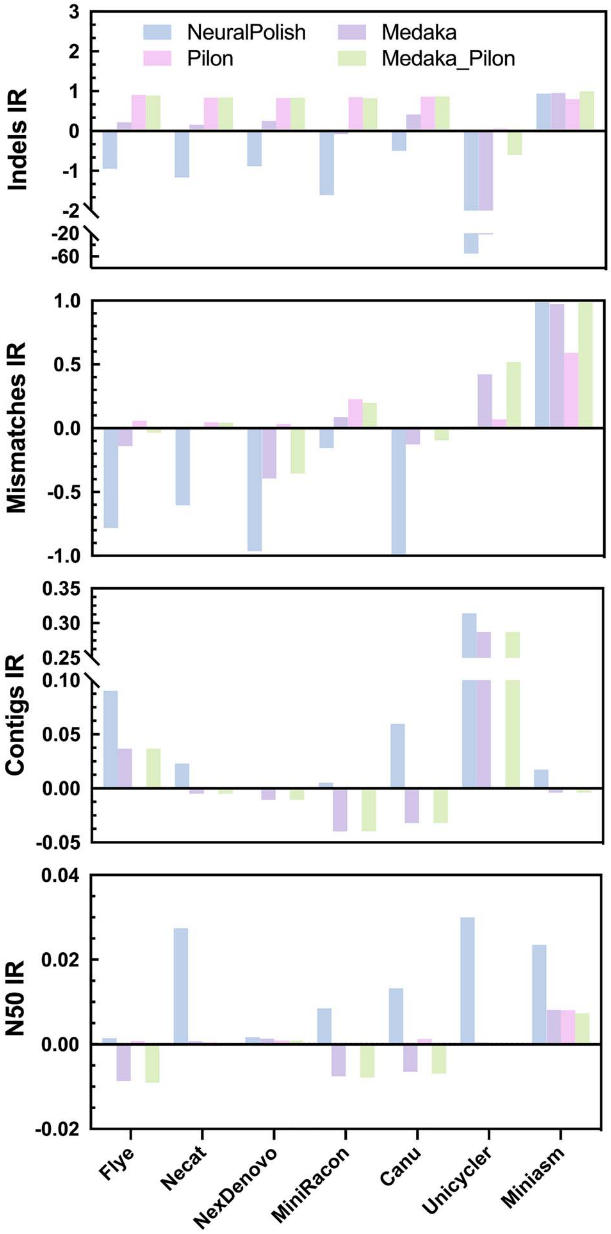

Figure 3. Metrics IR of assemblies after the polishing process.

Medaka can reduce the contig number to a certain degree and NeuralPolish can improve the N50 metric since they use the long reads to improve the continuity. Pilon has no obvious effect on the improvement of continuity because it uses the short reads from SGS for fine polishing of bases. However, in terms of accuracy, Pilon is the most robust polisher to reduce Mismatches and Indels among three single polishers, closely followed by Medaka. Medaka can reduce the Mismatches of draft sequences, especially for Unicycler. After Unicycler's assembly and Medaka's polishing, we obtained the

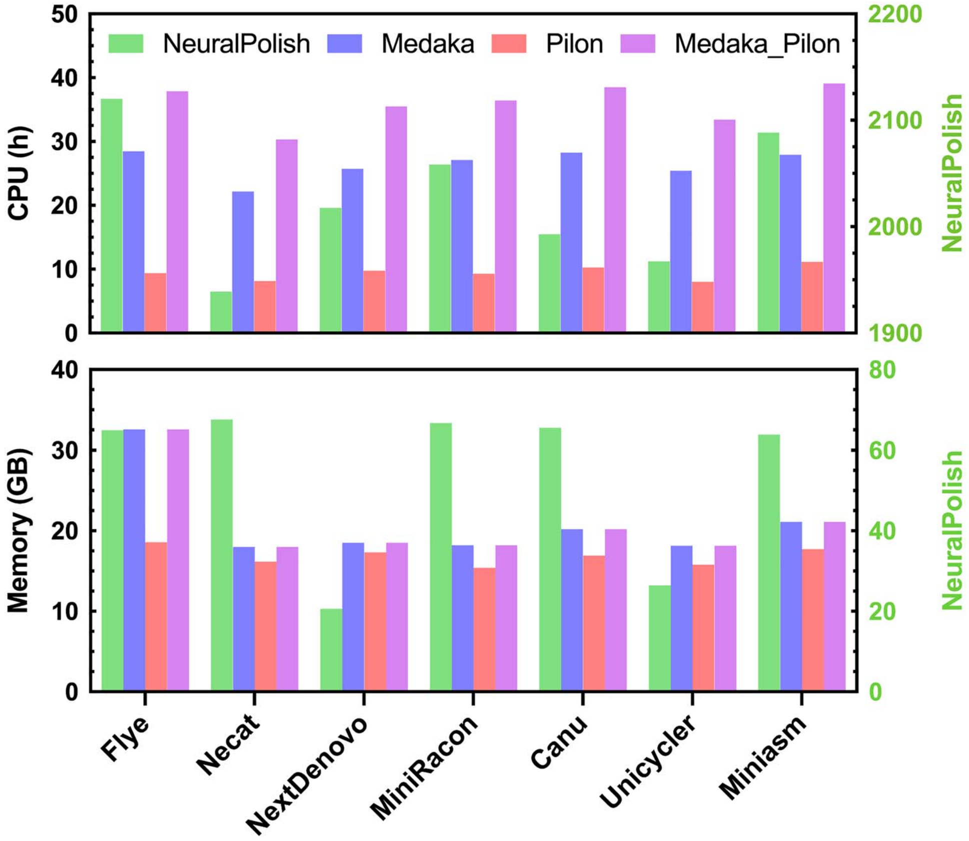

Figure 4. Computational resources required for polishers. The value is the sum of 13 subsets with data depths.

least mismatched sequences (Supplementary Figure S4, see Supplementary Data available online at https://academic.oup.com/bib). NeuralPolish showed variable performance on accuracy. On one hand, it can reduce Mismatches and Indels of assemblies from Miniasm, on the other hand, it really increases that of other assemblies, which is due to the introduction of insertion error rather than deletion error (Supplementary Figure S5, see Supplementary Data available online at https://academic.oup.com/bib). It is suspected that in the polishing process of NeuralPolish, in order to correct the deletion error more greedily, some correct positions were misjudged as the deletion sites [13], and then NeuralPolish modified these positions to introduce new insertion errors.

The comparison results of the three single polishers demonstrated that Pilon and Medaka perform better, in which Medaka is more effective for the increase of continuity and the decrease of Mismatches, and Pilon is very competitive in the improvement of sequence accuracy. Consequently, we combined the two polishers and demonstrated Medaka_Pilon performed better in most of the quality metrics (Figure 3).

In terms of the computational performance, we calculated the sum of CPU time and memory utilized for 13 subsets (Figure 4). Pilon took the least time and memory followed by Medaka. Both of them consumed fewer computing resources compared with assemble process. However, NeuralPolish took a very long time, especially in the neural network prediction step independent of the input data depth (Supplementary Figure S6, see Supplementary Data available online at https://academic.oup.com/bib), which is about 160 CPU hours for each

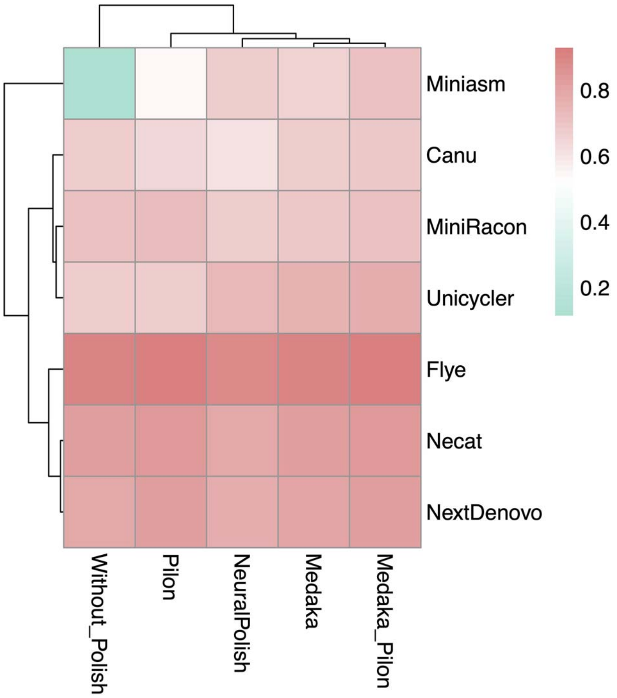

Figure 5. The C_scores heatmap of four polishing pipelines by seven assemblers independently.

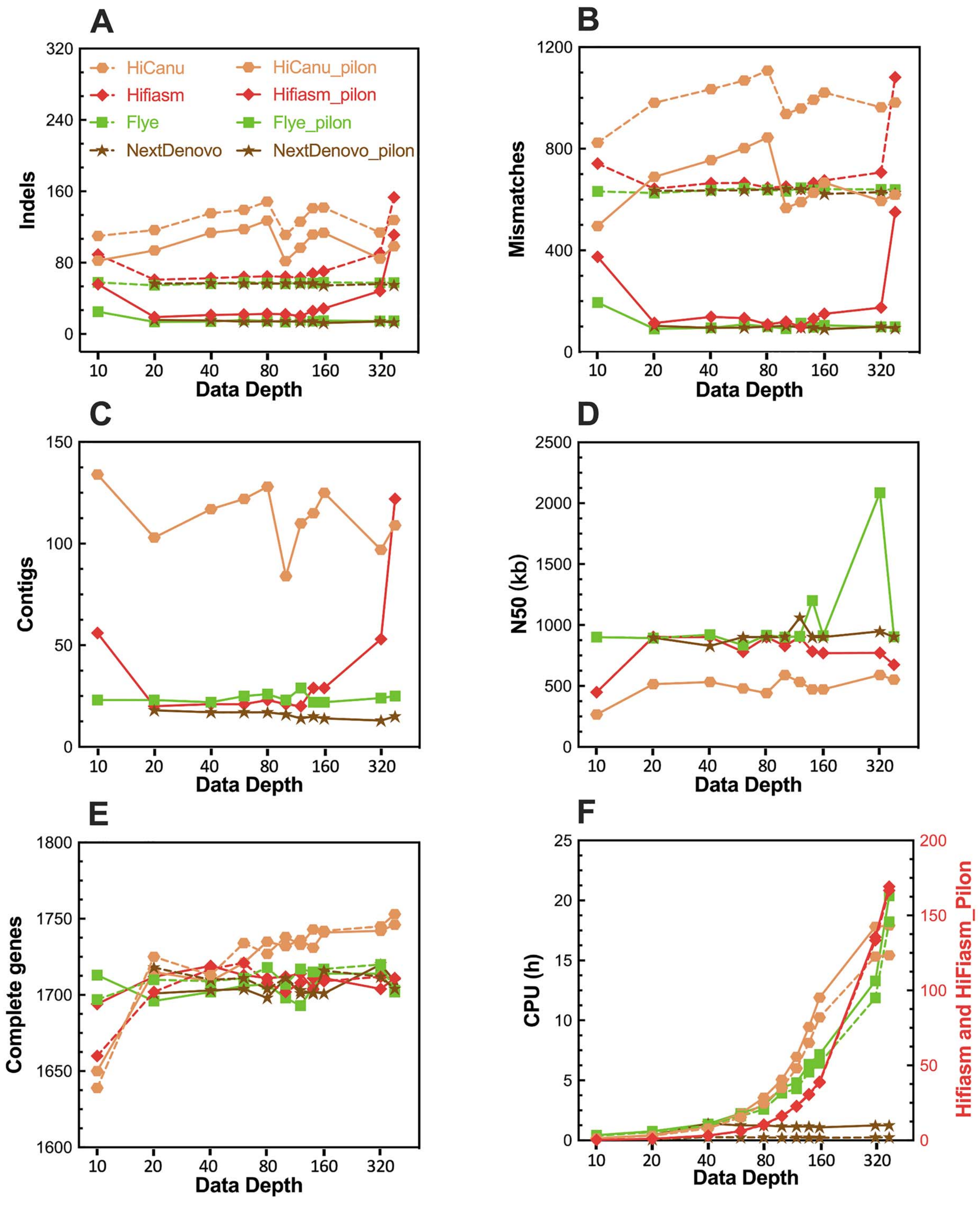

Figure 6. Main metrics of assemblies on 11 HiFi subsets with different data depths. Number of Indels (**A**) and Mismatches (**B**) per 100 kb, contig numbers (**C**), length of N50 (**D**), complete genes number from BUSCO (**E**) and computational time (**F**).

assembly, with more intensive resources than most of their assemble process.

C_score evaluation of polishing pipelines after ONT dataset assembly

We also calculated the C_score of all pipelines of ONT datasets and displayed by the heatmap in Figure 5. The results showed that the Medaka_Pilon polishing performed well with most assemblers. Medaka has similar performance and is clustered in the nearest branch with Medaka_pilon, followed by another neural network-based polisher NeuralPolish. Pilon also has a good performance except with Miniasm. In terms of assemblers, similar polishing features were observed in the draft assemblies from Necat and NextDenovo. Flye is a superior tool that achieves the highest C_scores

Table 2. The mean of metric values (M) and comprehensive scores (C_score) of different pipelines on HiFi dataset.

	Contigs	N50 (kb)	Mismatches	Indels	Completeness	C_score
Flye_Pilon	24.0	1034.4	108.8	15.5	1704.6	0.778
NextDenovo_Pilon	15.6	914.5	97.4	14.1	1705.0	0.760
hifiasm_Pilon	37.7	787.8	190.3	36.1	1709.6	0.649
Flye	24.0	1034.6	637.8	57.2	1711.3	0.647
NextDenovo	15.6	914.6	634.7	56.2	1709.0	0.602
hifiasm	37.7	787.9	707.6	77.4	1705.6	0.424
HiCanu_Pilon	113.1	495.3	659.4	101.9	1723.5	0.292
HiCanu	113.1	495.3	988.3	128.3	1726.6	0.200

*Each M was calculated by the average of 11 HiFi subset metrics (Mi) from 10X to 380X.

Table 3. Optimal pipelines with C_score>0.9 by comparison of all pipelines on ONT and HiFi datasets.

Pipelines	Flye_Pilon_HiFi	ND_Pilon_HiFi	Flye_HiFi	Flye_Pilon_ONT	Flye_MP_ONT
C_score	0.971	0.942	0.922	0.921	0.919

*All C_scores of 43 pipelines can be found in Supplementary Table S4, see Supplementary Data available online at https://academic.oup.com/bib

with three polishers, Pilon, Medaka_Pilon and Medaka. Miniasm is the most distinct assembler due to its absence of raw reads correction step.

Evaluation of assemblers on HiFi datasets

We compared four assembly tools designed to leverage the full potential of HiFi reads, HiCanu, hifiasm, Flye and NextDenovo (Figure 6). HiCanu, showed significant differences in assembly quality. Assemblies by HiCanu have the lowest quality in both continuity and accuracy, with the lowest N50 and the highest Mismatches and Indels at any depth, except 380x. However, HiCanu is good at improving genomic integrity, and it obtained the best BUSCO evaluation results with the highest number of complete genes in most subsets (Figure 6E).

The other three tools performed similarly, except that hifiasm was more sensitive to the data depth. The quality of the genome constructed by hifiasm decreased at very low (10x) or high (>300x) depths, while the number of contigs, Mismatches and Indels increased dramatically at 320x and 380x.

The effect of polishing process on HiFi datasets was also tested. Each assembly obtained from HiFi datasets was polished by Pilon. Pilon correction significantly improves genome accuracy and genome integrity regardless of the assembly tool. Therefore, the conclusion that the post-assembly correction process is necessary to improve genome quality is not only applicable to the ONT dataset but also to the HiFi dataset.

C_scores of these eight pipelines (four assemblers with or without Pilon polishing) showed that Flye_Pilon and NextDenovo_Pilon are two superior pipelines with C_score of 0.778 and 0.760 (Table 2), which indicates that these two pipelines are the most stable choice for assembling Hifi data.

Comparison of optimal pipelines from ONT and HiFi datasets

To investigate the impact of the two sequencing methods on genome construction, we compared all 43 pipelines (35 ONT-based and 8 HiFi-based) through C_score. Mean metric values of 10 subsets from 10X to 320X were used and scaled across these 43 pipelines and further calculated the C_scores (Table S4). Pipelines with C_scores>0.9 are shown in Table 3. Both pipelines, Flye_Pilon and NextDenovo_Pilon using HiFi reads, constructed the best assemblies, closely followed by the Flye pipeline using HiFi reads and Flye_Pilon and Flye_MP pipelines using ONT reads.

We then compare the assemblies from 2 optimal pipelines of HiFi data and ONT data in detail (Figure S7), the contig numbers of assemblies obtained by the HiFi data are lower than that of the ONT data, the N50 and complete gene number from BUSCO of assemblies obtained by HiFi were slightly higher than that of ONT (Supplementary Figure S7, see Supplementary Data available online at https://academic.oup.com/bib), indicating that the genome quality from HiFi reads has an overall improvement compared to ONT, which is consistent with the C_score comparison. However, HiFi pipelines have introduced more Mismatches or Indels than that of ONT pipelines when mapped to the reference genome from the NCBI database. Considering the high quality of HiFi raw reads (Supplementary Figure S8, see Supplementary Data available online at https://academic.oup.com/bib), there may be a possibility that the advanced HiFi method detected some details that were ignored in the previous genome construction and further improved the quality of the reference genome. For computational performance, genome construction based on HiFi data consumed less computational resources compared to ONT. Here we apply three tools, Canu, Flye and NextDenovo, to both ONT and HiFi datasets,

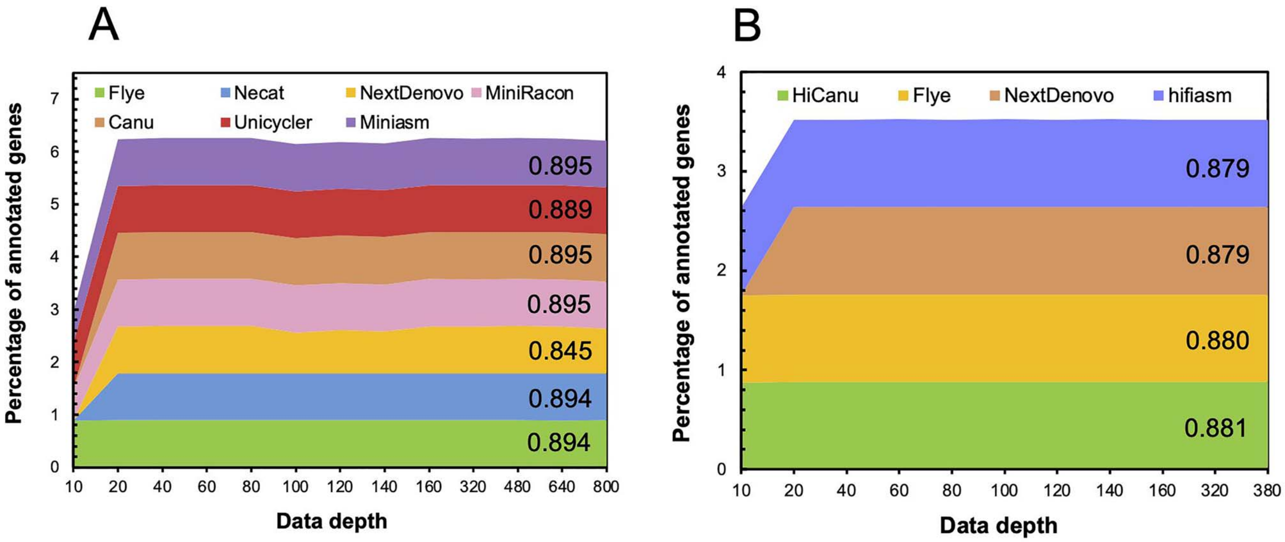

Figure 7. The percentage of annotated genes built by assemble_Medaka_Pilon pipeline on ONT (**A**) and assemble_Pilon pipeline on HiFi (**B**) datasets. Numbers on the graph's right side are the percentage of annotated genes at 800X (**A**) or 380X (**B**).

the CPU time and memory usage of HiFi datasets are smaller than that of ONT datasets at the same depth (Figures 2F and 6F).

Notably, when the data depth is extremely low such as 10X, Flye performs outstanding in both HiFi and ONT data. However, the characteristics of different tools should be considered, for example, the powerful phased assemble ability of hifiasm will open the door for the haplotype resolution of polyploid genomes.

The effect of data depth on assemblers and polishers

The assembly quality was highly variable across the different assemblers on low-depth subsets (10X, 20X, 40X in Figure 2 and 10X in Figure 6) and kept robust on relative high-depth subsets. Different from the assembly process, the polishing process is not sensitive to data depth (Supplementary Figure S4, see Supplementary Data available online at https://academic.oup.com/bib), Notably, it is not the deeper the better in the choice of sequencing depth. On one hand, the subsets with higher coverage will hardly further improve the quality of assembly over a certain threshold. On the other hand, the higher the depth, the more computing resources will be required, and ultrahigh-depth data may even confuse the assembler to crash so that cannot obtain the assembly. Therefore, it is necessary to determine a suitable depth. In our results, the dataset with about 20X can build most genomes, but for further improvement on genome quality, an increase in sequencing depth is necessary. For high-quality genome construction of fungi such as yeast, the sequencing depth should not be less than 80X for ONT (Figure 2) and 20X for HiFi data (Figure 6), but it is also not recommended to exceed 320X if assembled by Canu or Miniasm on ONT dataset.

Yeast genome annotation

The percentage of genes that have been built by ONT and HiFi datasets with Medaka_Pilon polishing and Pilon polishing process are shown in Figure 7. For the full ONT or HiFi datasets, the genomes obtained by most pipelines can consist of over 87% genes. Flye has the best gene-built ability that is independent of sequencing method and data depth even at 10X. The reduction of the total annotated genes percentage in other assemblers such as Necat and Canu on ONT data or NextDenovo on HiFi data was mainly due to the sensitivity to low-depth subsets, especially at 10X.

To investigate the ability of genome annotation by different pipelines, we compared the genes identified from four optimal pipelines with 320X subset. Most of the genes (5259 genes) can be successfully constructed by all pipelines and there are 79 and 6 genes that can only be built by ONT and HiFi datasets, respectively (Supplementary Figure S9, see Supplementary Data available online at https://academic.oup.com/bib), which indicates that the choice of sequencing method may have an effect on genome annotation.

Case study

According to the comparison results above, four optimal pipelines were selected, Flye_Pilon and Flye_MP for ONT data and Flye_Pilon and NextDenovo_Pilon for HiFi data, to build the draft genome of other two industrial yeast strains of different genera, *S. pombe* FLO-DUT (SP) and *S. cerevisiae* CICC-1445 (SC). Although the accuracy of the assembly such as Mismatches or Indels cannot be assessed since the reference is not available for these strains, we still evaluated other metrics on SP and SC assemblies and demonstrated in Tables 4 and 5.

Table 4. Statistic information on the *de novo* assemblies from the best four pipelines for the *S. pombe* FLO-DUT (SP)

Assembly		**SP**			
		Flye_Pilon_ONT	**Flye_MP_ONT**	**ND_Pilon_HiFi**	**Flye_Pilon_HiFi**
Number of contigs		8	6	6	4
Largest contig (Mb)		5.554	5.555	5.624	5.595
Total length (Mb)		12.73	12.73	12.81	12.65
GC (%)		36.05	36.05	36.05	36.05
N50 (Mb)		4.495	4.495	4.555	4.554
BUSCO	**Complete**	811	797	804	801
	Fragmented	110	108	111	111

Table 5. Statistic information on the *de novo* assemblies from the best four pipelines for the *S. cerevisiae* CICC-1445 (SC)

Assembly		**SC**			
		Flye_Pilon_ONT	**Flye_MP_ONT**	**ND_Pilon_HiFi**	**Flye_Pilon_HiFi**
Number of contigs		31	32	15	61
Largest contig (Mb)		1.475	1.479	2.417	0.978
Total length (Mb)		11.84	11.87	12.05	12.21
GC (%)		38.37	38.38	38.40	38.27
N50 (Mb)		0.811	0.818	0.946	0.580
BUSCO	**Complete**	1622	1613	1712	1683
	Fragmented	149	148	144	149

Both assemblers can build the high-continuity genome with N50 about 4500 kb in SP and 800 kb in SC. NextDenovo_Pilon pipeline with HiFi data obtained genome with the highest N50 in both strains. And both HiFi pipelines have significantly higher complete gene numbers on SC assemblies than that of ONT pipelines.

The complete genes of SC evaluated from BUSCO were more than that of SP, this is not due to the difference in assembly quality but the difference in reference annotated species datasets. The closest reference dataset of SC is saccharomycetes_odb10 (class level) while that of SP is ascomycota_odb10 (phylum level), phylum has the higher taxonomic level so that contains more other species-specific genes.

Discussion

In this study, we performed 455 and 88 *de novo* assemblies of *S. cerevisiae* S288C on high-coverage ONT and HiFi data, respectively, to comprehensively evaluate the influence of assembly tools, polishing tools and sequencing depth on eukaryotic genome construction. According to our C_score, the pipelines based on Flye assembler perform best on both ONT and HiFi datasets, and NextDenovo is another recommended choice for HiFi data. The polishing process is necessary to improve the qualities of assemblies and Medaka_Pilon performs best for ONT datasets. In the case application of two nonreference strains, SP and SC, the genome obtained from HiFi data are more continuous and complete than that of ONT data. For the selection of these two sequencing platforms, other factors should also be considered such as price, throughput and convenience. Unlike Flye, which can recover some short sequences such as plasmids therefore result in more contigs, NextDenovo has fewer contig numbers, even lower than the number of chromosomes in high-depth datasets, suggesting that there may be excessive overlap problems.

Among all tested assemblers, only Unicycler takes the hybrid assembly pipeline and performs inferior in contiguity but superior in accuracy. On one hand, the addition of short reads increases the fragmentation of the assembly, while on the other hand, it really reduces the Indels through the integration of short reads [26]. In addition, since Unicycler was intended to build the genome of prokaryotes which has a relatively short length and low complexity, it doesn't need to balance the assembly quality and computing performance. While in *S. cerevisiae*, computational consumption is its bottleneck, which is ~20 times that of Flye.

Different from the previous evaluation work of ~30X ONT data in *S. cerevisiae* [18], we used the ultrahigh-depth dataset for the first time to comprehensively assess the impact of sequencing depth on the construction of genomes with different tools. Generally speaking, the quality of the genome improves with the sequencing depth and tends to be stable after 80X and 20X on ONT and HiFi datasets, respectively, but the computing consumption will continue increasing. For most assemblers, 20X data is the lower limit for genome construction, while an increase in sequencing depth is necessary for further improvement. Qualities of *de novo* genome assembly have an important impact on downstream annotation and comparative genomics applications [27]. This benchmark study not only contributes to the high-quality genome construction of yeast, but also provides insights for other eukaryote genomes such as mold, microalgae and even complex human genomes.

Key Points

- Comprehensive benchmarking on the latest assembly and polishing tools for two advanced TGS datasets (ONT and HiFi) for eukaryotic model organism.
- Comparison of different assemblers across a wide range of sequencing depths for the first time.
- Flye is the most robust assembler on both ONT and HiFi datasets and NextDenovo also performs outstandingly on HiFi datasets.
- High-quality genome construction of two unsequenced industrial yeast strains.

Data Availability Statement

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary material. Besides, all raw sequencing data are available from the NCBI database with BioProject accession PRJNA792930, PRJNA792931, and PRJNA792932 for *S. cerevisiae* S288C, *S. cerevisiae* CICC 1445, and *S. pombe* FLO-DUT respectively. All the generated assemblies of this study are available from the corresponding author on request.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

The computations in this paper were run on the π 2.0 cluster supported by the Center for High-Performance Computing at Shanghai Jiao Tong University. We also thank the Core Facilitiy and Service Center (CFSC) in School of Life Sciences and Biotechnology, SJTU for strain culture and DNA sample preparation.

Funding

National Key Research and Development Program of China (2018YFB1501401, 2021YFC2100600, 2019YFA-0905400); National Natural Science Foundation of China (21978167, 31970026, 32070679); Open Project Funding of the State Key Laboratory of Biocatalysis and Enzyme Engineering (SKLBEE2018016).

References

1. Shendure J, Balasubramanian S, Church GM, *et al.* DNA sequencing at 40: past, present and future. *Nature* 2017;**550**(7676):345–53.
2. Amarasinghe SL, Su S, Dong X, *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020;**21**(1):30.
3. Wang Y, Zhao Y, Bollas A, *et al.* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* 2021;**39**(11): 1348–65.
4. Koren S, Walenz BP, Berlin K, *et al.* Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* 2017;**27**(5):722–36.
5. Nurk S, Walenz BP, Rhie A, *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* 2020;**30**(9):1291–305.
6. Chen Y, Nie F, Xie SQ, *et al.* Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun* 2021;**12**(1):60.
7. https://github.com/Nextomics/NextDenovo.
8. Kolmogorov M, Yuan J, Lin Y, *et al.* Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;**37**(5):540–6.
9. Cheng H, Concepcion GT, Feng X, *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 2021;**18**(2):170–5.
10. Wick RR, Judd LM, Gorrie CL, *et al.* Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;**13**(6):e1005595.
11. Walker BJ, Abeel T, Shea T, *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;**9**(11):e112963.
12. Vaser R, Sovic I, Nagarajan N, *et al.* Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 2017;**27**(5):737–46.
13. Huang N, Nie F, Ni P, *et al.* NeuralPolish: a novel Nanopore polishing method based on alignment matrix construction and orthogonal Bi-GRU Networks. *Bioinformatics* 2021;**37**(19): 3120–7.
14. Senol Cali D, Kim JS, Ghose S, *et al.* Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Brief Bioinform* 2019;**20**(4):1542–59.
15. Islam R, Raju RS, Tasnim N, *et al.* Choice of assemblers has a critical impact on de novo assembly of SARS-CoV-2 genome and characterizing variants. *Brief Bioinform* 2021;**22**(5):bbab102. https://doi.org/10.1093/bib/bbab102.
16. Chen Z, Erickson DL, Meng J. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC Genomics* 2020;**21**(1):631.
17. Zhang P, Jiang D, Wang Y, *et al.* Comparison of de novo assembly strategies for bacterial genomes. *Int J Mol Sci* 2021;**22**(14): 7668.
18. Giordano F, Aigrain L, Quail MA, *et al.* De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci Rep* 2017;**7**(1):3935.
19. Tvedte ES, Gasser M, Sparklin BC, *et al.* Comparison of long read sequencing technologies in interrogating bacteria and fly genomes. *G3 (Bethesda)* 2021;**11**(6):jkab083. https://doi.org/10.1093/g3journal/jkab083.
20. Lang D, Zhang S, Ren P, *et al.* Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi

reads of Pacific Biosciences Sequel II system and ultra-long reads of Oxford Nanopore. *Gigascience* 2020;**9**(12):giaa123. https://doi.org/10.1093/gigascience/giaa123.

21. https://github.com/PacificBiosciences/pbmm2.
22. Gurevich A, Saveliev V, Vyahhi N, *et al.* QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;**29**(8): 1072–5.
23. Manni M, Berkeley MR, Seppey M, *et al.* BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* 2021;**38**(10):4647–54.
24. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 2003;**19**(Suppl 2):ii215–25.
25. https://bioinfogp.cnb.csic.es/tools/venny/index.html.
26. Sovic I, Krizanovic K, Skala K, *et al.* Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. *Bioinformatics* 2016;**32**(17):2582–9.
27. Giani AM, Gallo GR, Gianfranceschi L, *et al.* Long walk to genomics: history and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J* 2020;**18**: 9–19.