






## Article

# Assembling Quality Genomes of Flax Fungal Pathogens from Oxford Nanopore Technologies Data

Elizaveta A. Sigova <sup>1,2</sup>, Elena N. Pushkova <sup>1</sup>, Tatiana A. Rozhmina <sup>3</sup>, Ludmila P. Kudryavtseva <sup>3</sup>, Alexander A. Zhuchenko <sup>3,4</sup>, Roman O. Novakovskiy <sup>1</sup>, Daiana A. Zhernova <sup>1,5</sup>, Liubov V. Povkhova <sup>1,2</sup>, Anastasia A. Turba <sup>1</sup>, Elena V. Borkhert <sup>1</sup>, Nataliya V. Melnikova <sup>1</sup>, Alexey A. Dmitriev <sup>1,\*</sup> and Ekaterina M. Dvorianinova <sup>1,\*</sup>

<sup>1</sup> Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russia

<sup>2</sup> Moscow Institute of Physics and Technology, Moscow 141701, Russia

<sup>3</sup> Federal Research Center for Bast Fiber Crops, Torzhok 172002, Russia

<sup>4</sup> All-Russian Horticultural Institute for Breeding, Agrotechnology and Nursery, Moscow 115598, Russia

<sup>5</sup> Faculty of Biology, Lomonosov Moscow State University, Moscow 119234, Russia

\* Correspondence: alex\_245@mail.ru (A.A.D.); dvorianinova.em@phystech.edu (E.M.D.)

**Abstract:** Flax (*Linum usitatissimum* L.) is attacked by numerous devastating fungal pathogens, including *Colletotrichum lini*, *Aureobasidium pullulans*, and *Fusarium verticillioides* (*Fusarium moniliforme*). The effective control of flax diseases follows the paradigm of extensive molecular research on pathogenicity. However, such studies require quality genome sequences of the studied organisms. This article reports on the approaches to assembling a high-quality fungal genome from the Oxford Nanopore Technologies data. We sequenced the genomes of *C. lini*, *A. pullulans*, and *F. verticillioides* (*F. moniliforme*) and received different volumes of sequencing data: 1.7 Gb, 3.9 Gb, and 11.1 Gb, respectively. To obtain the optimal genome sequences, we studied the effect of input data quality and genome coverage on assembly statistics and tested the performance of different assembling and polishing software. For *C. lini*, the most contiguous and complete assembly was obtained by the Flye assembler and the Homopolish polisher. The genome coverage had more effect than data quality on assembly statistics, likely due to the relatively low amount of sequencing data obtained for *C. lini*. The final assembly was 53.4 Mb long and 96.4% complete (according to the glomerellales\_odb10 BUSCO dataset), consisted of 42 contigs, and had an N50 of 4.4 Mb. For *A. pullulans* and *F. verticillioides* (*F. moniliforme*), the best assemblies were produced by Canu–Medaka and Canu–Homopolish, respectively. The final assembly of *A. pullulans* had a length of 29.5 Mb, 99.4% completeness (dothideomycetes\_odb10), an N50 of 2.4 Mb and consisted of 32 contigs. *F. verticillioides* (*F. moniliforme*) assembly was 44.1 Mb long, 97.8% complete (hypocreales\_odb10), consisted of 54 contigs, and had an N50 of 4.4 Mb. The obtained results can serve as a guideline for assembling a de novo genome of a fungus. In addition, our data can be used in genomic studies of fungal pathogens or plant–pathogen interactions and assist in the management of flax diseases.

**Keywords:** *Aureobasidium pullulans*; *Colletotrichum lini*; *Fusarium verticillioides*; *Fusarium moniliforme*; pathogens; flax; nanopore sequencing; genome assembly



**Citation:** Sigova, E.A.; Pushkova, E.N.; Rozhmina, T.A.; Kudryavtseva, L.P.; Zhuchenko, A.A.; Novakovskiy, R.O.; Zhernova, D.A.; Povkhova, L.V.; Turba, A.A.; Borkhert, E.V.; et al. Assembling Quality Genomes of Flax Fungal Pathogens from Oxford Nanopore Technologies Data. *J. Fungi* **2023**, *9*, 301. <https://doi.org/10.3390/jof9030301>

Academic Editor: Zonghua Wang

Received: 31 January 2023

Revised: 22 February 2023

Accepted: 23 February 2023

Published: 26 February 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

*Colletotrichum lini* Manns et Bolley, *Aureobasidium pullulans* (de Bary) Arnaud, and *Fusarium verticillioides* (Sacc.) Nirenberg (*Fusarium moniliforme* Sheldon) are the fungal flax (*Linum usitatissimum* L.) pathogens, which cause diseases leading to significant crop losses. Flax is a highly valued cultivated plant because of its broad use. It is used for manufacturing food additives for people and animals because of the lignans, omega-3, and high fiber content in flaxseed [1–4]. Flax oil is widely used in industry as a component of coatings and paints [5]. Flax fiber is a popular product for manufacturing cloth and

paper [6]. Thus, economic profit from these valuable products depends on flax resistance to various pathogens, including *C. lini*, *A. pullulans*, and *F. verticillioides* (*F. moniliforme*) [7,8]. However, these types of pathogens demonstrate significant genetic diversity, which hinders the creation of universally resistant varieties [9–13].

To develop resistant varieties, it is essential to conduct detailed studies of phytopathogen genetics and determine genetic markers of pathogenicity. This knowledge is useful for phylogenetic studies and species resolution [14,15]. For initial identification, fungal barcodes can be used, e.g., the beta-tubulin (*TUB2*) gene or the ITS region [16–19]. However, molecular markers sometimes provide limited information on species. For example, in the *Aureobasidium* genus, using the ITS2 marker failed to distinguish between the species. In contrast, whole-genome phylogenetic analysis identified three separate species [20]. Thus, for the most complete representation of the pathogen at the genetic level, the sequence and structure of its genome should be revealed. For *C. lini*, the flax anthracnose causative agent, the genome sequence was unknown until this study.

For *A. pullulans*, full genome assemblies are available in the NCBI database. They have an average length of 28.5 Mb (23.8–31.0 Mb) and scaffold/contig level (<https://www.ncbi.nlm.nih.gov/assembly/?term=aureobasidium%20pullulans>, accessed on 10 January 2023). Nonetheless, obtaining complete genomes of this species is still reasonable due to the significant genetic diversity between the isolates. Sequencing 50 *A. pullulans* strains isolated from different sources revealed the absence of population structure. However, linkage disequilibrium analysis suggested high levels of recombination between *A. pullulans* strains [21]. This fact might explain the polyextremotolerance of the fungus and its adaptability to a variety of unfavorable conditions. Thus, the genome of a flax-isolated strain might become a useful source of information on markers of adaptation to *L. usitatissimum*.

For *F. verticillioides* (*F. moniliforme*), genome assemblies are also available in the NCBI database (<https://www.ncbi.nlm.nih.gov/assembly/?term=Fusarium+verticillioides>, accessed on 10 January 2023). The length of these assemblies is 42.7 Mb on average (41.8–44.7 Mb). Their levels vary from contig to chromosome. However, most of the genomes were assembled only from Illumina data and then scaffolded. As a result, they still have many gaps and low contig N50 values (<0.5 Mb). In addition, none of the sequenced *F. verticillioides* is specified as an isolate from flax. Meanwhile, *Fusarium* includes a broad range of remarkably diverse species [22–25]. Thus, an SSR marker analysis demonstrated that *F. verticillioides* isolates from maize are genetically diverse without correlation with a geographic region of isolation [26]. Therefore, sequencing the genome of an isolate from flax is beneficial for further research on the pathogenicity and evolution of the species.

In the current study, we performed sequencing on a third-generation sequencing platform. In comparison with second-generation sequencing, third-generation sequencing technologies (Pacific Biosciences (PacBio), Menlo Park, CA, USA and Oxford Nanopore Technologies (ONT), Oxford, UK) enable the construction of genomes with fewer gaps [27–29]. Notably, Oxford Nanopore Technologies allows the acquisition of super-long reads with a maximum length of 2.3 Mb (the maximum length obtained in a scientific laboratory [30]). However, the obtained volume of raw reads depends on many factors, such as the organism species, genome size and structure, and purity and mass of the extracted DNA [31–33]. Using an appropriate protocol for DNA extraction is a key factor in receiving a sufficient amount of raw data to obtain a high-quality whole-genome assembly [34–38]. Unfortunately, the accuracy of the obtained data is limited by the technology itself. Errors occur during the steps of sequencing and raw signal deciphering [39]. Nevertheless, basecalling parameters can still be varied to receive the optimal genome assembly.

Data accuracy and quantity are important factors in determining the quality of the resulting assembly [38,40,41]. Thus, our research aimed to study the effect of different types of ONT data (different coverage, accuracy, and species) on the quality of a genome assembly. We analyzed genome assemblers' performance in relation to raw ONT read volume and basecalling quality threshold. The effectiveness of different polishing tools for ONT data was also tested on the optimal raw assemblies of the sequenced flax pathogens.

## 2. Materials and Methods

### 2.1. Fungal Material

The following strains were used from the collection of the Institute for Flax (Torzhok, Russia): highly pathogenic to flax *Colletotrichum lini* #811, highly pathogenic to flax *Aureobasidium pullulans* #16, and *Fusarium verticillioides* (*Fusarium moniliforme*) #366 with low pathogenicity to flax. Mycelium was cultivated in test tubes with potato dextrose agar.

### 2.2. DNA Extraction and Purification

Pure, high-molecular-weight DNA was obtained according to the previously developed protocol for *Fusarium oxysporum* f. sp. *lini*, with several modifications [42]. After the step of incubation with RNase A, DNA samples were left at 8 °C overnight. Then, DNA purification was continued the next day, according to the protocol. For *C. lini* #811, all the precipitated DNA was taken for library preparation. This resulted in relatively low data output. To receive more sequencing data for *A. pullulans* #16 and *F. verticillioides* (*F. moniliforme*) #366, we took about 3 g of fungal material instead of 1 g, and visible DNA precipitate was removed from the samples. Only the remaining DNA was further used. The quality and quantity of the extracted DNA were evaluated with spectrophotometry (NanoDrop 2000C spectrophotometer, Thermo Fisher Scientific, Waltham, MA, USA) and fluorometry (Qubit 4.0 fluorometer, Thermo Fisher Scientific, Waltham, MA, USA). For *A. pullulans* #16 and *F. verticillioides* (*F. moniliforme*) #366, DNA was additionally purified with AMPure XP beads (Beckman Coulter, Brea, CA, USA) to achieve higher purity.

### 2.3. DNA Library Preparation and Sequencing on the Oxford Nanopore Technologies Platform

SQK-LSK109 Ligation Sequencing Kit (Oxford Nanopore Technologies, Oxford, UK) was used to prepare libraries. Several modifications were introduced to the recommended manufacturer's protocol: the time of incubation at the step of DNA recovery at 20 °C was increased from 5 to 20 min, and the time of ligation was increased from 10 to 60 min. The prepared libraries were sequenced on a MinION instrument with the FLO-MIN-106 R9.4.1 flow cell.

### 2.4. Genome Assembly

The obtained reads were basecalled using Guppy 6.0.1 and the dna\_r9.4.1\_450bps\_sup.cfg config file with different quality filtration thresholds (min\_qscore). For *C. lini* strain #811, min\_qscore was taken in the range of 7 to 10. For *A. pullulans* #16 and *F. verticillioides* (*F. moniliforme*) #366, a default min\_qscore of 10 was chosen. Porechop 0.2.4 was used to remove adapters. For each min\_qscore value, draft assemblies were performed using Canu 2.2 (-nanopore-raw; -minInputCoverage = 5; -stopOnLowCoverage = 5; -genomeSize = 50 m (*C. lini* #811), -genomeSize = 30 m (*A. pullulans* #16), or -genomeSize = 45 m (*F. verticillioides* (*F. moniliforme*) #366)), Flye 2.8.1 (-genome-size 50000000 (*C. lini* #811), -genome-size 30000000 (*A. pullulans* #16), or -genome-size 45000000 (*F. verticillioides* (*F. moniliforme*) #366)), Miniasm 0.3-r179 (-x ava-ont), NextDenovo 2.5.0 (<https://github.com/Nextomics/NextDenovo>, accessed on 10 January 2023), Ra 0.2.1 (-x ont), Raven 1.5.1, Shasta 0.8.0 (Nanopore-Oct2021.conf), SmartDenovo, and Wtdbg-cns 1.1 (Wtdbg2 0.0) (-x ont; -g 50 m (*C. lini* #811), -g 30 m (*A. pullulans* #16), or -g 45 m (*F. verticillioides* (*F. moniliforme*) #366)) [43–49]. To analyze the quality of the obtained assemblies, BUSCO (Benchmarking Universal Single-Copy Orthologs) 5.3.2 and QUASt 5.0.2 were used [50,51]. The following datasets were used for the analysis with BUSCO: glomerellales\_odb10 (*C. lini* #811), dothideomycetes\_odb10 (*A. pullulans* #16), and hypocreales\_odb10 (*F. verticillioides* (*F. moniliforme*) #366). The following reference genomes were used for evaluating assembly contiguity: *C. lini* #811—*Colletotrichum higginsianum* IMI 349063 (GCA\_001672515.1, accessed on 10 January 2023, sequenced with PacBio, chromosome-level assembly); *A. pullulans* #16—*Aureobasidium pullulans* (GCA\_903819485.1, sequenced with ONT, contig-level assembly); *F. verticillioides* (*F. moniliforme*) #366—*Fusarium verticillioides* (GCA\_003316995.2, sequenced with Illumina, chromosome-level assembly). The obtained draft assemblies were

polished with ONT reads (the same basecalling quality threshold as was used to assemble the genomes) with Homopolish 0.3.4 (specified option: -m R9.4.pkl), MarginPolish 1.3.0 (allParams.np.json) (<https://github.com/UCSC-nanopore-cgl/MarginPolish>, accessed on 10 January 2023), Medaka 1.5.0 (<https://github.com/nanoporetech/medaka>, accessed on 10 January 2023), NextPolish 1.4.0, Pepper 0.0.6, Racon 1.4.20 [52–55]. For Homopolish, databases for polishing were created from assemblies available in NCBI: for *C. lini* #811—*Colletotrichum higginsianum* IMI 349063 (GCA\_001672515.1), *Colletotrichum fructicola* (GCA\_009771025.1), *Colletotrichum scovillei* (GCA\_011075155.1), *Colletotrichum australisnense* (GCA\_014706365.1), *Colletotrichum echinocloae* (GCA\_016618095.1), *Colletotrichum eleusines* (GCA\_016807845.1), *Colletotrichum horii* (GCA\_019693695.1), *Colletotrichum acutatum* (GCA\_001593745.1), *Colletotrichum sansevieriae* (GCA\_002749775.1), *Colletotrichum musae* (GCA\_002814275.1); for *A. pullulans* #16—*Aureobasidium pullulans* (GCA\_903819465.1, GCA\_003574545.1, GCA\_004917105.1, GCA\_004917135.1, GCA\_004917145.1, GCA\_004917155.1, GCA\_004917165.1, GCA\_004917185.1, GCA\_004917375.1, GCA\_004917415.1); for *F. verticillioides* (*F. moniliforme*) #366—*F. verticillioides* (GCA\_026119585.1, GCA\_020882315.1, GCA\_027571605.1, GCA\_013759275.1, GCF\_000149555.1, GCA\_003316975.2, GCA\_003316995.2, GCA\_003317015.2, GCA\_025503005.1, GCF\_000149555.1). If required, all prior alignments before polishing were produced with Minimap2 [56].

To compare the final assemblies of *C. lini*, *A. pullulans*, and *F. verticillioides* (*F. moniliforme*) with available genomes of *Colletotrichum*, *Aureobasidium*, and *Fusarium* species, the following assemblies were analyzed using BUSCO (glomerallales\_odb10) and QUAST: *Colletotrichum fructicola* (GCA\_009771025.1), *Colletotrichum scovillei* (GCA\_011075155.1), *Colletotrichum australisnense* (GCA\_014706365.1), *Colletotrichum echinocloae* (GCA\_016618095.1), *Colletotrichum eleusines* (GCA\_016807845.1), *Colletotrichum horii* (GCA\_019693695.1), *Colletotrichum acutatum* (GCA\_001593745.1), *Colletotrichum sansevieriae* (GCA\_002749775.1), *Colletotrichum musae* (GCA\_002814275.1), *Aureobasidium pullulans* (GCA\_903819465.1), *Aureobasidium pullulans* (GCA\_003574545.1), *Aureobasidium zeae* (GCA\_017580825.2), *Fusarium verticillioides* (GCA\_026119585.1), *Fusarium verticillioides* (GCA\_020882315.1), *Fusarium avenaceum* (GCA\_025948275.1), *Fusarium verticillioides* (GCA\_027571605.1), *Fusarium verticillioides* (GCA\_013759275.1).

### 3. Results

#### 3.1. Genome Assembly and Polishing

The purity of the sequenced DNA affects the volume and quality of raw nanopore reads. In this study, we used a previously developed protocol to extract pure high-molecular-weight DNA from the studied fungi [42]. For *C. lini* #811, the total DNA pool was used to prepare sequencing libraries, as no visible DNA precipitate could be isolated. However, this resulted in low data output. We received 1.65 Gb of raw ONT reads with an N50 of 15.7 kb (33× genome coverage for an expected genome length of 50 Mb). Most likely, long DNA fragments in the pool were insufficiently purified, resulting in a short lifetime of sequencing pores. We assumed that the isolation failure was due to a low amount of the input biological material. For *A. pullulans* #16 and *F. verticillioides* (*F. moniliforme*) #366, a greater mass of mycelium (2.5–3 g instead of 1 g) was taken for DNA isolation. Thus, visible DNA precipitate (likely presented by long but insufficiently pure DNA fragments) could be removed from the pool, and the remaining material was sequenced. We received the following volumes of raw ONT reads with the corresponding N50 parameters: 3.89 Gb (130× genome coverage for an expected genome length of 30 Mb) with N50 = 5.8 kb (*A. pullulans* #16) and 11.08 Gb (220× genome coverage for an expected genome length of 50 Mb) with N50 = 6.9 kb (*F. verticillioides* (*F. moniliforme*) #366).

For *C. lini* #811, the obtained sequencing data were basecalled with Guppy using the dna\_r9.4.1\_450bps\_sup.cfg config file and different quality filtration thresholds (min\_qscores of 7–10). The genome of the highly pathogenic *C. lini* #811 was assembled from these data with different tools (Figure 1, Table S1). We assessed the completeness and contiguity of the assembled sequences using BUSCO and QUAST. For each assembly,

QUAST statistics were evaluated with and without a reference sequence. The contiguity of the assemblies was estimated by the number of contigs and the N50 and L50 parameters: the higher the N50 and the lower the number of contigs and L50, the more contiguous the assembly. Reference-based parameters were used to evaluate the completeness (reference genome fraction, identified reference genomic features (CDS, exons, etc.); higher values indicate higher completeness), contiguity (NG50, LG50), and accuracy (mismatches/indels per 100 kb; the lower these statistics, the more accurate the assembly) of the obtained assemblies. In the case of reference-based assessment, the compared query and reference genomes can be significantly different. Thus, it was not the QUAST statistics of an assembly that were of interest but their ratio to those of other obtained assemblies. BUSCO was used for evaluating assembly completeness. A higher fraction of complete benchmarking universal single-copy orthologs inherent to an analyzed species group indicated higher completeness of an assembly.

Basecalling quality	Basecalled data volume, Mb	N50 basecalled, kb	Assembler	No reference						With reference								
				Length, Mb	Number of contigs	N50, Mb	L50	BUSCO			Genome fraction, %	Genomic features		NG50, kb	LG50	Mismatches per 100 kb	Indels per 100 kb	
								C, %	D, %	F, %		Complete	Partial					
10	441	19.2	Canu	48.1	443	0.2	93	80.9	0.4	3.5	50.4	34,786	38,778	148	102	4383	213	
			Flye	52.9	111	1.1	15	89.7	0.3	2.1	54.8	41,724	37,484	1179	14	4515	202	
			Miniasm	27.9	290	0.1	84	20.2	0.0	2.4	14.1	2020	28,404	54	228	2936	11	
			NextDenovo															
			Ra	26.8	248	0.1	82	45.4	0.1	3.6	29.6	25,650	17,243	58	217	4097	195	
			Raven	32.2	351	0.1	104	50.8	0.1	3.1	31.4	19,422	29,298	69	-	4592	206	
			Shasta	27.1	691	0.1	142	41.3	0.1	3.5	27.4	15,008	27,144	16	503	4784	235	
			SmartDenovo	43.9	375	0.1	93	63.0	0.2	7.9	43.2	25,792	40,646	131	118	4691	197	
			Wtdbg2	50.3	178	0.6	26	67.4	0.2	5.5	46.6	24,676	48,039	577	27	4367	181	
9	558	19.1	Canu	51.7	260	0.3	46	86.6	0.6	3.6	53.8	37,660	40,334	330	45	4363	208	
			Flye	53.1	48	3.3	7	92.7	0.2	2.3	55.9	44,786	35,010	3313	7	4473	200	
			Miniasm	41.9	297	0.2	80	26.3	0.0	3.0	18.5	2640	38,872	151	106	2980	17	
			NextDenovo	0.1	1	0.1	1	0.3	0.0	0.0	0.1	148	71	-	-	4329	158	
			Ra	43.4	274	0.2	74	70.7	0.2	6.2	47.2	39,126	28,077	171	95	4044	182	
			Raven	45.5	285	0.2	75	71.6	0.2	4.1	45.1	29,104	38,891	188	89	4425	192	
			Shasta	42.1	762	0.1	165	55.5	0.2	4.5	39.2	17,926	44,131	67	222	5297	231	
			SmartDenovo	49.6	218	0.3	45	72.5	0.2	9.1	49.8	32,182	42,390	335	47	4454	190	
			Wtdbg2	51.6	98	1.3	12	73.6	0.2	5.3	50.3	27,748	49,101	1404	11	4423	198	
8	685	19.0	Canu	52.9	174	0.5	26	88.8	0.3	3.5	55.1	38,914	40,965	589	24	4278	206	
			Flye	53.4	37	3.4	6	93.7	0.3	2.1	56.4	45,910	34,160	3516	5	4634	214	
			Miniasm	48.5	191	0.4	40	27.0	0.0	3.1	18.6	2516	40,101	355	43	2970	14	
			NextDenovo	40.1	135	0.4	35	69.8	0.2	1.4	41.7	35,298	24,671	291	51	3981	172	
			Ra	50.5	177	0.4	37	82.2	0.1	7.4	54.3	45,136	31,914	429	38	4123	181	
			Raven	51.4	160	0.5	32	85.5	0.2	3.5	52.5	36,540	40,563	494	31	4263	184	
			Shasta	45.0	709	0.1	142	59.3	0.1	5.1	41.7	18,750	47,016	88	173	5300	218	
			SmartDenovo	52.2	119	0.7	22	77.0	0.2	9.5	53.1	36,420	41,643	735	21	4444	199	
			Wtdbg2	52.1	81	2.1	8	74.8	0.2	6.4	51.7	28,594	49,709	2101	8	4600	226	
7	830	19.0	Canu	53.4	134	0.8	21	88.8	0.5	3.8	55.3	38,292	41,980	827	19	4347	213	
			Flye	53.3	42	4.4	5	93.5	0.2	2.2	56.3	45,388	34,634	4437	5	4526	207	
			Miniasm	50.5	96	0.8	22	23.6	0.0	2.6	16.1	2170	35,553	842	22	2978	44,882	
			NextDenovo	52.3	69	1.2	13	91.6	0.3	2.1	55.3	45,430	32,813	1238	13	4339	194	
			Ra	52.5	98	0.9	22	85.7	0.1	6.8	56.1	46,478	32,886	879	21	4099	175	
			Raven	52.8	90	0.9	19	89.9	0.2	2.6	54.7	40,660	38,666	955	18	4348	196	
			Shasta	46.1	663	0.1	132	61.6	0.1	5.1	42.7	19,610	47,960	104	153	5260	219	
			SmartDenovo	52.7	85	1.1	17	76.6	0.2	9.7	53.3	36,114	42,255	1187	16	4578	209	
			Wtdbg2	51.8	40	3.2	6	71.0	0.1	6.4	51.2	27,012	50,902	3201	6	5122	270	

Figure 1. QUAST and BUSCO statistics of *C. lini* strain #811 draft genome assemblies. Basecalled data

volume, millions of bases—data volume obtained after basecalling with different quality thresholds (min\_qscore). N50 basecalled, kb—N50 of data obtained after basecalling with different quality thresholds (min\_qscore). BUSCO: C—complete, D—duplicated, F—fragmented (the glomerellales\_odb10 dataset). Reference—*Colletotrichum higginsianum* IMI 349063 (GCA\_001672515.1). Genomic features: Partial—partially covered features (an assembly contains at least 100 bp of a feature (CDS, exons, etc.) but not its whole sequence). The used colors indicate estimations of the value quality from dark green (best) to bright red (worst). The grey line indicates the absence of an assembly. Bold font indicates the highest quality assembly.

At all basecalling quality thresholds, the majority of the tools could assemble a genome of *C. lini* #811 with a completeness < 90% only. The assemblies with BUSCO completeness > 80% had an average length of 52.2 Mb (48.1–53.4 Mb), an average GC content of 54.04% (53.97–54.12%), and average duplication ratio values of 1.043 (1.037–1.048). For each min\_qscore, the highest assembly completeness was achieved by Flye (up to 93.7%) and the lowest by Miniasm (less than 30%). However, NextDenovo demonstrated even worse BUSCO completeness (0.3%) than Miniasm at min\_qscore = 9, as it failed to produce an assembly of a reasonable length. For min\_qscore = 10, the assembler was unable to produce a consensus sequence. At high min\_qscore values (9–10), all assemblers but Flye (at min\_qscore = 9) failed to construct genomes with a BUSCO completeness of more than 90%. At high quality thresholds, the contiguity of the obtained assemblies was poor for most assemblers. Only Flye, at min\_qscore = 9/10, and Wtdbg2, at min\_qscore = 9, could obtain assemblies with N50 values of the megabase order and numbers of contigs fewer than 100. Generally, all assemblers demonstrated better results at the lower min\_qscore values (7–8) than at the higher ones (9–10).

Notably, Flye outperformed other tools in terms of the key QUAST and BUSCO parameters. At each min\_qscore, Flye produced assemblies with the best reference-based parameters, e.g., genome fraction and genomic features (more than 50% genome fraction for all min\_qscore values). It obtained the most complete assembly from the basecalled data with min\_qscore = 8: BUSCO completeness of 93.7%, 45,910 complete and 34,160 partial genomic features, and a genome fraction of 56.4%. At a min\_qscore of 7, the tool produced the most contiguous assembly among all the raw genomes. It consisted of 42 contigs and had an N50 of 4.4 Mb. The second most continuous assembly was also performed by Flye at min\_qscore = 8 (37 contigs, N50 = 3.4 Mb). Although BUSCO completeness at min\_qscore = 8 (93.7%) was 0.2% higher than that at min\_qscore = 7, the assembly at min\_qscore = 7 had better QUAST statistics (N50, L50, and NG50). Therefore, we considered that the draft assembly at min\_qscore = 7 is optimal in terms of contiguity and completeness.

For *A. pullulans* #16 and *F. verticillioides* (*F. moniliforme*) #366, we obtained volumes of raw ONT data several times greater than those for *C. lini* #811. After basecalling with Guppy using the dna\_r9.4.1\_450bps\_sup.cfg config file and default min\_qscore = 10, the basecalled datasets were still several times bigger than those of *C. lini* #811 basecalled with min\_qscore = 7 (2.15 Gb and 5.06 Gb vs. 0.83 Gb). Thus, taking into account the high genome coverage with ONT reads, we decided to use the data basecalled only with min\_qscore = 10 for further genome assembly to reduce the number of low-quality reads. At min\_qscore = 10, assemblers performed better for the larger datasets. Thus, the assemblies of *A. pullulans* #16 and *F. verticillioides* (*F. moniliforme*) #366 generally had better QUAST and BUSCO statistics than those of *C. lini* #811 (Figure 2, Tables S2 and S3). For *A. pullulans* #16, the assemblies by Flye and Raven had a BUSCO completeness of 99.1%, which was 0.3% more than the parameter for the Canu assembly. However, the assembly by Canu had the highest N50, reference genome fraction, genomic features, and NG50. Thus, the assembly by Canu was considered optimal: 98.8% BUSCO completeness, length = 29.5 Mb, N50 = 2.4 Mb, and number of contigs = 32. For *F. verticillioides* (*F. moniliforme*) strain #366, the most contiguous assembly was also obtained using the Canu assembler: N50 = 4.4 Mb against 1.8 Mb for Raven and 1.2 Mb for Flye; L50 = 5 against 10 for Raven and 12 for Flye. Reference-based analysis showed that this genome assembly has the highest genome

fraction, genomic features, and NG50 parameters. Raven assembled the most complete genome according to BUSCO (96.1% completeness). However, N50 and other QUASt parameters of this assembly were substantially lower than those of the assembly by Canu. Therefore, the assembly by Canu was considered optimal for *F. verticillioides* (*F. moniliforme*) #366: 94.5% BUSCO completeness, length = 44.1 Mb, N50 = 4.4 Mb, and number of contigs = 54. Its BUSCO completeness can be further improved by polishing.

Species	Strain	Basecalled data volume, Mb	N50 basecalled, kb	Assembler	No reference					With reference									
					Length, Mb	Number of contigs	N50, Mb	L50	BUSCO			Genome fraction, %	Genomic features		NG50, kb	LG50	Mismatches per 100 kb	Indels per 100 kb	
									C, %	D, %	F, %		Complete	Partial					
<i>Colletotrichum lini</i>	811	441	19.2	Canu	48.1	443	0.2	93	80.9	0.4	3.5	50.4	34,786	38,778	148	102	4383	213	
				Flye	52.9	111	1.1	15	89.7	0.3	2.1	54.8	41,724	37,484	1179	14	4515	202	
				Miniasm	27.9	290	0.1	84	20.2	0.0	2.4	14.1	2020	28,404	54	228	2936	11	
				NextDenovo															
				Ra	26.8	248	0.1	82	45.4	0.1	3.6	29.6	25,650	17,243	58	217	4097	195	
				Raven	32.2	351	0.1	104	50.8	0.1	3.1	31.4	19,422	29,298	69	-	4592	206	
				Shasta	27.1	691	0.1	142	41.3	0.1	3.5	27.4	15,008	27,144	16	503	4784	235	
				SmartDenovo	43.9	375	0.1	93	63.0	0.2	7.9	43.2	25,792	40,646	131	118	4691	197	
				Wtdbg2	50.3	178	0.6	26	67.4	0.2	5.5	46.6	24,676	48,039	577	27	4367	181	
<i>Aureobasidium pullulans</i>	16	2145	6.3	Canu	29.5	32	2.4	5	98.8	0.2	0.3	84.2	86,676	3890	2385	5	2100	261	
				Flye	29.4	74	2.2	6	99.1	1.2	0.2	83.8	86,654	3904	2231	6	2120	252	
				Miniasm	26.7	142	0.3	33	37.8	0.1	12.7	49.3	15,812	48,879	251	41	7684	568	
				NextDenovo	29.0	61	1.1	9	94.9	0.2	1.9	83.6	86,166	4026	931	10	2104	262	
				Ra	28.9	87	0.5	21	95.0	0.4	1.7	82.9	85,450	4155	428	24	2102	252	
				Raven	29.1	20	2.3	5	99.1	0.2	0.2	83.9	86,445	4031	2296	5	2100	250	
				Shasta	25.0	639	0.1	137	72.2	0.1	5.3	71.1	70,902	7356	46	195	2186	410	
				SmartDenovo	31.1	637	0.1	174	71.0	2.2	7.1	70.1	68,402	9064	60	174	3109	584	
				Wtdbg2	32.9	859	0.5	20	94.3	0.3	1.3	82.1	84,327	4814	557	19	2314	334	
<i>Fusarium verticillioides</i> ( <i>Fusarium moniliforme</i> )	366	5060	6.6	Canu	44.1	54	4.4	5	94.5	0.2	2.9	96.0	2119	177	4356	5	440	47	
				Flye	43.5	337	1.2	12	95.2	1.3	2.3	95.7	2097	195	1175	12	442	39	
				Miniasm	36.7	364	0.2	78	36.1	0.2	10.1	61.0	500	908	136	97	5511	1213	
				NextDenovo	42.5	102	1.0	13	89.7	0.4	5.2	93.8	2112	168	997	13	456	80	
				Ra	41.7	160	0.4	35	93.4	0.7	1.9	93.2	1898	142	392	35	437	39	
				Raven	43.4	42	1.8	10	96.1	0.3	1.7	95.8	2106	183	1884	9	448	39	
				Shasta	39.3	940	0.1	195	74.4	0.2	10.0	86.9	1896	210	57	220	493	258	
				SmartDenovo	39.9	1123	0.0	398	51.5	4.2	6.2	58.6	924	233	36	431	1439	752	
				Wtdbg2	58.1	2852	0.1	184	76.9	0.8	7.2	87.5	1851	295	123	102	1159	446	

**Figure 2.** QUASt and BUSCO statistics of *C. lini* #811, *A. pullulans* #16, and *F. verticillioides* (*F. moniliforme*) #366 draft genome assemblies at min\_qscore = 10. Basecalled data volume, millions of bases—data volume obtained after basecalling at min\_qscore = 10. N50 basecalled, kb—N50 of data obtained after basecalling at min\_qscore = 10. BUSCO: C—complete, D—duplicated, F—fragmented (the glomerellales\_odb10 (*C. lini* #811), dothideomycetes\_odb10 (*A. pullulans* #16), and hypocreales\_odb10 (*F. verticillioides* (*F. moniliforme*) #366) datasets). Genomic features: Partial—partially covered features (if an assembly contains at least 100 bp of a feature (CDS, exons, etc.) but not its whole sequence). The used colors indicate estimations of the value quality from dark green (best) to bright red (worst). The grey line indicates the absence of an assembly. Bold font indicates the highest quality assemblies.

The best assemblies of the three strains (*C. lini* #811: Flye at min\_qscore = 7, *A. pullulans* #16: Canu at min\_qscore = 10, and *F. verticillioides* (*F. moniliforme*) #366: Canu at

min\_qscore = 10) were polished with ONT reads basecalled with the same basecalling quality threshold. Polishing was performed with six various tools. Each polisher was used in three iterations (Figure 3: first round of polishing; Tables S4–S6: all three rounds of polishing). In comparison with raw assemblies, polished ones should have better BUSCO completeness and contain more reference genome sequences and genomic features due to the improvement in sequence accuracy and reduction in the number of erroneous mismatches and indels. For *C. lini* #811, only Homopolish increased BUSCO completeness (from 93.5% to 96.3%). Genome fraction and genomic features also significantly rose after correction with Homopolish: from 56.30% to 61.15% and from 41,724 to 64,784, respectively. The tool significantly decreased mismatches and indels per 100 kb: from 4526 to 4131 and from 207 to 124, respectively. After the second round of polishing with Homopolish, these values were slightly refined: from 4131 to 4110 and from 124 to 116, respectively (Table S4). BUSCO completeness and reference genome fraction were also improved a little after the second polishing round. They rose from 96.3% to 96.4% and from 61.2% to 61.7%, respectively. The third round failed to make significant changes. Therefore, the assembly of the *C. lini* #811 genome after the second round of Homopolish was the most complete (BUSCO completeness = 96.4%) and contiguous (assembly length = 53.4 Mb, N50 = 4.4 Mb, 42 contigs, L50 = 5).

Species	Strain	Polisher	No reference				With reference					
			Number of contigs	BUSCO			Genome fraction, %	Genomic features		Misassemblies	Mismatches per 100 kb	Indels per 100 kb
				C, %	D, %	F, %		Complete	Partial			
<i>Colletotrichum lini</i>	811	Flye assembly, min_qscore = 7	42	93.5	0.2	2.2	56.3	41,724	37,484	1520	4526	207
		Homopolish	42	96.3	0.2	0.9	61.2	64,784	16,926	1924	4131	124
		MarginPolish	37	85.4	0.2	7.2	56.4	46,334	33,712	1531	4515	199
		Medaka	42	89.4	0.2	5.1	56.8	49,422	30,408	1529	4449	191
		NextPolish	42	87.6	0.2	6.1	56.6	47,748	32,139	1517	4462	192
		Pepper	35	86.9	0.2	5.7	56.0	46,496	33,093	1531	4546	198
		Racon	36	86.1	0.2	7.1	56.5	46,356	33,566	1532	4483	199
<i>Aureobasidium pullulans</i>	16	Canu assembly, min_qscore = 10	32	98.8	0.2	0.3	84.2	86,676	3890	525	2100	261
		Homopolish	32	97.2	0.1	0.9	84.2	86,879	3694	528	2102	245
		MarginPolish	31	96.1	0.1	1.1	84.1	86,503	4138	523	2106	277
		Medaka	32	99.4	0.1	0.1	84.3	86,897	3687	527	2101	262
		NextPolish	32	96.2	0.1	1.6	84.2	86,785	3783	519	2102	261
		Pepper	31	94.9	0.1	2.5	84.3	86,621	3989	526	2105	266
		Racon	31	99.1	0.1	0.2	84.2	86,610	3923	529	2096	261
<i>Fusarium verticillioides</i> ( <i>Fusarium moniliforme</i> )	366	Canu assembly, min_qscore = 10	54	94.5	0.2	2.9	96.0	2119	177	564	440	47
		Homopolish	54	97.8	0.1	0.5	96.0	2113	179	561	440	25
		MarginPolish	54	86.0	0.2	9.0	95.9	2102	189	480	438	82
		Medaka	54	97.5	0.1	0.7	96.0	2119	174	562	440	30
		NextPolish	54	95.3	0.1	2.3	96.0	2118	177	559	440	40
		Pepper	54	97.5	0.1	0.8	96.0	2115	178	554	442	35
		Racon	53	97.5	0.2	2.1	96.0	2118	174	558	440	39

Figure 3. QUAST and BUSCO statistics of *C. lini* #811, *A. pullulans* #16, and *F. verticillioides* (*F. moniliforme*)



#366 polished genome assemblies. BUSCO: C—complete, D—duplicated, F—fragmented (the glomerellales\_odb10 (*C. lini* #811), dothideomycetes\_odb10 (*A. pullulans* #16), and hypocreales\_odb10 (*F. verticillioides* (*F. moniliforme*) #366) datasets). Genomic features: Partial—partially covered features (if an assembly contains at least 100 bp of a feature (CDS, exons, etc.) but not its whole sequence). The used colors indicate estimations of the value quality from dark green (best) to bright red (worst). Statistics in grey-blue are the ones of the best draft assembly that was further polished. All required prior alignments were made with Minimap2.

Polishing with Homopolish is based on the use of homologous sequences from the genomes of closely related species. Thus, a user needs to create a database of these sequences and pass it to the input of the polishing tool. We tried polishing the draft *C. lini* genome assembly using a single genome. To choose the closest sequence, we evaluated the QUAST parameters of the *C. lini* draft assembly using the genomes of *Colletotrichum* representatives: GCA\_001672515.1, GCA\_009771025.1, GCA\_011075155.1, GCA\_014706365.1, GCA\_016618095.1, GCA\_016807845.1, GCA\_019693695.1, GCA\_001593745.1, GCA\_002749775.1, and GCA\_002814275.1. *Colletotrichum higginsianum* GCA\_001672515.1 had the highest reference genome fraction in the assembly of *C. lini* #811 and was used for polishing. However, the use of a single genome provided nearly the same results as when it was included in the database of several genomes. After each of the three polishing rounds, BUSCO completeness values and relative numbers of indels were equal to those received with a multi-genome database. After the first round of Homopolish, the number of complete reference genomic features was 18 more than for the assembly polished with a wider database. However, the second round of polishing resulted in the same value as that after polishing with several genomes. The relative number of mismatches also remained virtually unchanged compared to those for each polishing round with the multi-genome database. Thus, our previously obtained polished assembly could still be considered optimal.

For *A. pullulans* strain #16, polishing with only two tools led to an increase in BUSCO completeness: from 98.8% to 99.4% for Medaka and from 98.8% to 99.1% for Racon. The other polishers decreased the parameter. In addition, all tools except Medaka, Homopolish, and NextPolish lowered the number of complete reference genomic features. For most tools, changes in mismatches and indels per 100 kb were insignificant. The second round of polishing with Racon left BUSCO completeness unchanged (Table S5). The second round of polishing with Medaka failed to improve any parameters significantly: the genome fraction changed from 84.27% to 84.29%, indels per 100 kb improved from 262 to 259, and mismatches per 100 kb improved from 2101 to 2099. However, BUSCO completeness decreased from 99.4% to 97.1%. The third round of polishing with Racon significantly decreased BUSCO completeness from 99.1% to 96.4%. After the third round of polishing with Medaka, QUAST statistics changed insignificantly again: genomic features changed from 86,761 to 86,766, mismatches per 100 kb did not change, indels per 100 kb improved from 259 to 258, and BUSCO completeness remained the same. The second and third rounds of polishing with other tools had almost no effect on BUSCO completeness. Thus, polishing the assembly of *A. pullulans* #16 with one round of Medaka was optimal. The resulting assembly had 99.4% BUSCO completeness. It had a total length of 29.5 Mb, N50 of 2.4 Mb, 32 contigs, and L50 of 5.

One iteration of polishing the assembly of *F. verticillioides* (*F. moniliforme*) strain #366 with Homopolish, Medaka, Pepper, or Racon significantly increased BUSCO completeness. For Homopolish, the increase was the greatest: from 94.5% to 97.8%. Reference genome fraction and complete genomic features were not improved substantially by any tool. Homopolish greatly decreased the relative number of indels from 47 to 25. The second and third rounds of polishing with any tool resulted in insignificant changes (Table S6). For Homopolish, the second round improved indels per 100 kb only from 25 to 24. Meanwhile, other parameters remained nearly the same (mismatches per 100 kb, genomic features, and BUSCO completeness). The third round of Homopolish also failed to change most of these parameters. Thus, indels per 100 kb remained at 24. The second and third rounds

of Medaka or Pepper made insignificant changes in the QUAST parameters. In contrast, BUSCO completeness decreased after the second and third rounds of Racon. Therefore, one round of polishing with Homopolish made the optimal assembly: BUSCO completeness = 97.8%, assembly length = 44.1 Mb, N50 = 4.4 Mb, 54 contigs, and L50 = 5.

### 3.2. Comparison with Available Genomes

To compare the obtained assemblies of *C. lini* #811, *A. pullulans* #16, and *F. verticillioides* (*F. moniliforme*) #366 with available assemblies from the NCBI database (accessed on 10 January 2023), genomes of the corresponding genus or species were downloaded and analyzed using BUSCO (the glomerellales\_odb10 (*Colletotrichum*), dothideomycetes\_odb10 (*Aureobasidium*), and hypocreales\_odb10 (*Fusarium*) datasets) and QUAST. We analyzed the following assemblies obtained from ONT data: *Colletotrichum australisense* (GCA\_014706365.1) sequenced with ONT GridION; *Colletotrichum horii* (GCA\_019693695.1) and *Colletotrichum scovillei* (GCA\_011075155.1) sequenced with ONT PromethION; *Aureobasidium pullulans* (GCA\_903819465.1) and *Fusarium avenaceum* (GCA\_025948275.1) sequenced with ONT MinION.

In addition, we calculated statistics for hybrid assemblies from ONT and Illumina data: *Colletotrichum fructicola* (GCA\_009771025.1), *Colletotrichum echinoclaoe* (GCA\_016618095.1), *Colletotrichum eleusines* (GCA\_016807845.1), *Fusarium verticillioides* (GCA\_026119585.1), and *Fusarium verticillioides* (GCA\_020882315.1).

To compare the assemblies with those constructed from data of sequencing technologies other than ONT, three genomes of *Colletotrichum* species (*Colletotrichum acutatum* (GCA\_001593745.1) sequenced with PacBio, *Colletotrichum sansevieriae* (GCA\_002749775.1) sequenced with IonTorrent, *Colletotrichum musae* (GCA\_002814275.1) sequenced with Illumina); two genomes of *Aureobasidium* species (*Aureobasidium pullulans* (GCA\_003574545.1) sequenced with Illumina, *Aureobasidium zeae* (GCA\_017580825.2) sequenced with PacBio HiFi); and two genomes of *Fusarium verticillioides* species (*Fusarium verticillioides* (GCA\_027571605.1) sequenced with PacBio HiFi, *Fusarium verticillioides* (GCA\_013759275.1) sequenced with Illumina) were downloaded from the NCBI database (<https://www.ncbi.nlm.nih.gov/>, accessed on 10 January 2023) from all available data types (Figure 4).

BUSCO completeness of the available genomes varied from 86.2% to 97.8%. Among the assemblies from different data types, the highest average completeness was observed for the assemblies produced completely or partly from long reads. On average, BUSCO completeness was 95.5% for the assemblies only from long-read data (ONT and PacBio), 96.9% for the hybrid assemblies, and 93.1% for the short-read assemblies. Except for the *C. fructicola* and *A. zeae* assemblies, the analyzed long-read and hybrid assemblies consisted of 9–42 contigs and had an N50 in the megabase range. Although *C. fructicola* was sequenced on both ONT and Illumina platforms, assembly statistics took an intermediate position between those of the other assemblies from long-read data and short-read data. *A. zeae* was sequenced on the PacBio platform. However, the statistics of the assembly were not close to those of the other assemblies from long-read data. The assemblies from short reads had thousands of contigs, and their N50 were in order of kilobases. Thus, the contiguity and completeness of the assemblies obtained in this study were superior to those of the analyzed assemblies from short-read data. However, the assemblies of the flax pathogens had comparable characteristics to most of the analyzed assemblies from long reads.

Species	Sequencing platform	No reference				
		Length, Mb	Number of contigs	N50, Mb	L50	BUSCO C, %
<i>Colletotrichum lini</i> #811	ONT	53.4	42	4.4	5	96.4
<i>Colletotrichum australisense</i>		55.3	28	5.7	4	87.9
<i>Colletotrichum horii</i>	ONT	74.3	42	3.1	8	97.1
<i>Colletotrichum scovillei</i>		52.0	16	4.8	5	97.4
<i>Colletotrichum fruticola</i>		58.1	447	0.9	23	97.5
<i>Colletotrichum echinochloae</i>	ONT and Illumina	62.2	21	5.3	5	96.1
<i>Colletotrichum eleusines</i>		53.5	15	5.1	5	95.9
<i>Colletotrichum acutatum</i>	PacBio	52.1	34	4.4	5	97.3
<i>Colletotrichum musae</i>	Illumina	49.1	10,618	0.007	2087	86.2
<i>Colletotrichum sansevieriae</i>	IonTorrent	51.2	8635	0.015	1008	91.5
<i>Aureobasidium pullulans</i> #16	ONT	29.5	32	2.4	5	99.4
<i>Aureobasidium pullulans</i> AWRI4230	ONT	29.8	15	2.5	5	96.7
<i>Aureobasidium pullulans</i> ASM357454v1	Illumina	29.5	456	0.2	35	97.0
<i>Aureobasidium zeae</i>	PacBio	23.6	94	0.7	12	95.4
<i>Fusarium verticillioides</i> ( <i>Fusarium moniliforme</i> ) #366	ONT	44.1	54	4.4	5	97.8
<i>Fusarium avenaceum</i>	ONT	42.0	9	4.9	4	94.1
<i>Fusarium verticillioides</i> ASM2611958v1		42.8	12	4.2	5	97.8
<i>Fusarium verticillioides</i> Fv10027_ITA	ONT and Illumina	44.7	21	2.9	5	97.2
<i>Fusarium verticillioides</i> ASM2757160v1	PacBio	41.9	11	4.2	5	97.8
<i>Fusarium verticillioides</i> ASM1375927v1	Illumina	41.9	857	0.1	118	97.7

**Figure 4.** QUASt and BUSCO statistics of available genome assemblies of *Aureobasidium*, *Colletotrichum*, and *Fusarium* species. BUSCO: the dothideomycetes\_odb10 (*Aureobasidium*), glomerellales\_odb10 (*Colletotrichum*), and hypocreales\_odb10 (*Fusarium*) datasets. The lines in grey-blue show statistics of the final assemblies of *C. lini* strain #811 (Flye, Homopolish x2), *A. pullulans* strain #16 (Canu, Medaka), and *F. verticillioides* (*F. moniliforme*) strain #366 (Canu, Homopolish).

#### 4. Discussion

In this study, we sequenced three fungal strains pathogenic to flax—*C. lini*, *A. pullulans*, and *F. verticillioides* (*F. moniliforme*)—on the ONT platform. This long-read technology allowed us to assemble genomes with high QUASt and BUSCO parameters: megabase-order N50, dozens of contigs, and BUSCO completeness of more than 95%. However, if the assemblies were based on short reads, the number of contigs would be in the order of thousands, and the N50 values would be in the order of kilobases. For instance, the *C. sansevieriae* assembly from IonTorrent data has nearly nine thousand contigs and an N50 of 150 kb (GCA\_002749775.1). Another example is the *C. musae* assembly from Illumina reads (GCA\_002814275.1), which consists of ten thousand contigs and has an N50 of 7 kb.

Meanwhile, whole-genome analysis has the potential to reveal key virulence factors and provides the direction for a thorough study of fungal pathogenicity [57]. However, the studied genomes must possess enough contiguity and accuracy to guarantee confidence in the results of genomic analysis. A contiguous and complete genome of a pathogen will advance further molecular research on the species' evolution, pathogenicity markers, genetic diversity, and plant–pathogen interactions [58–64]. This useful information can be retrieved from omics studies. Genome assemblies are used for mining genes, reconstructing phylogenetic relationships, studying recombination extents, revealing genetic determinants

of adaptations, etc. [65–67]. Thus, constructing a quality assembly is a primary task in studying the genomics of fungal pathogens.

Assembling the optimal genome of an organism usually implies benchmarking at least several bioinformatics tools [68]. Although the approach can be time- and resource-consuming for large genomes, testing different software for bacteria and fungi is a feasible task [69–71]. First, draft genome assemblies can be produced with a variety of instruments [72,73]. Then, a researcher can choose an optimal software for polishing with genomic reads [73]. However, the performance of bioinformatics instruments depends on the given amount and quality of data, as well as the complexity and length of the studied genome [74,75].

In this study, to construct the optimal genome of *C. lini*, we tested the performance of different assembly and polishing tools provided with different amounts of data. Sequencing reads were basecalled with mean quality thresholds from 7 to 10 (the `min_qscore` parameter in Guppy). Thus, the obtained datasets could be classified into two groups: the smaller ones of higher quality and the larger ones of lower quality. For the data basecalled at the strictest threshold (`min_qscore` = 10), most assemblers produced genomes of low completeness and poor contiguity (Figure 1). Notably, NextDenovo failed to output any genome sequence at all. At `min_qscore` = 9, assemblers showed better QUAST and BUSCO statistics. However, only Flye assembled a genome with a BUSCO completeness greater than 90%. At `min_qscore` = 8, the quality of the obtained assemblies improved again. The lowest basecalling threshold resulted in the highest genome coverage (~17× per a 50-Mb genome), while the N50 of the basecalled data remained the same (19 kb at each `min_qscore` value). Thus, genome coverage was critical for assembly statistics. Lowering the basecalling threshold and increasing genome coverage ~1.9 times gradually improved the QUAST and BUSCO parameters of the assemblies, except for those by Wtdbg2 and SmartDenovo. The lowest basecalling threshold allowed us to receive the most contiguous and complete assembly.

At each `min_qscore`, Flye outperformed other assemblers in the main QUAST (genome length, number of contigs, N50) and BUSCO parameters. Even at ~9× genome coverage (`min_qscore` = 10), the assembly by Flye had an N50 of the megabase order. Meanwhile, several other assemblers demonstrated consistently poor results. Miniasm provided the lowest assembly completeness. The assemblies by Raven and Ra could reach only a kilobase-order N50. Canu, one of the most widely used tools, neither reached a completeness of >90% nor an N50 of >0.8 Mb at all basecalling thresholds. This fact can be explained by the differences between the implemented algorithms. Flye is based on a graphing approach, while Canu employs an overlap–layout–consensus (OLC) paradigm [43,44]. Therefore, low genome coverage might be insufficient for Canu to find high-confidence overlaps and construct contigs. Probably, altering assembler parameters could slightly improve the results. Nonetheless, Flye showed quality assembly statistics at the default parameters.

In addition to *C. lini*, we sequenced two more flax pathogen genomes. We obtained a greater data volume for *A. pullulans* and *F. verticillioides* (*F. moniliforme*) than for *C. lini*; the *A. pullulans* and *F. verticillioides* (*F. moniliforme*) genomes were covered with raw data ~130 and ~250 times, respectively. In comparison with the dataset for the *C. lini* #811 assembly, the datasets for *A. pullulans* #16 and *F. verticillioides* (*F. moniliforme*) #366 were larger. We assumed that data quality might not be critical for assembly accuracy at high data volumes. Therefore, we basecalled raw data with default parameters (`min_qscore` = 10). This reduced the genome coverage to ~70× and ~110×. Using the basecalled data, we tested the performance of different assembly software to choose the optimal draft genomes. For *A. pullulans* #16, Canu, Flye, NextDenovo, and Raven produced assemblies with the lowest number of contigs and highest N50. Although Flye and Raven assembled 0.3% more complete genome sequences than Canu, Canu still assembled a genome with a higher N50. In addition, the assembly by Canu had fewer contigs than that by Flye. For *F. verticillioides* (*F. moniliforme*) #366, assemblies by Canu and Raven had the best statistics. However, the assembly by Canu had a significantly higher N50 than the assemblies by other tools.

Thus, for both *A. pullulans* and *F. verticillioides* (*F. moniliforme*), Canu constructed the most contiguous assemblies. At high genome coverage, this OLC assembler performed the best of all the tools.

In addition to high contiguity and completeness, accuracy is another important characteristic of a quality genome assembly. This parameter can be assessed using QUASt reference-based statistics and BUSCO completeness. Most representative QUASt statistics include the number of the identified reference genomic features (CDS, exons, etc.), mismatches, and indels per a certain number of base pairs. For *C. lini* #811, we calculated these statistics for the assemblies from basecalled data of different qualities. Despite different quality of genomic reads, we observed fluctuations in the relative number of indels for each assembler instead of a single tendency. For example, for Flye, the parameter changed in the following manner: 202–200–214–207 (min\_qscores from 10 to 7). The relative number of mismatches showed the same trend for these assemblies. However, the number of the complete reference genomic features gradually increased with lowering data quality to min\_qscore = 8. At min\_qscore = 7, the statistic decreased, probably due to the poorer data quality. For the other assemblers, the parameters changed in varying ways. Therefore, the quality of the input data had an inconsistent effect. For *A. pullulans* #16 and *F. verticillioides* (*F. moniliforme*) #366, we applied a single basecalling threshold. Assemblies with the best BUSCO and basic QUASt statistics had the highest number of detected genomic features and one of the lowest numbers of mismatches and indels.

To improve the accuracy of the raw genomes, we performed polishing with ONT data of the same min\_qscore values that were chosen as optimal for the high-quality draft assemblies. For *C. lini* #811 and *F. verticillioides* (*F. moniliforme*) #366, Homopolish polished raw assemblies to the smallest number of indels and the highest completeness and reference genome fraction. For *A. pullulans* #16, the polisher reduced BUSCO completeness. Since Homopolish corrects systematic errors using homologous sequences from the sequences provided by a user (database of 10 available *A. pullulans* assemblies was used for strain #16), it might have adjusted the *A. pullulans* genome to the provided sequences [53]. However, a large part of the *A. pullulans* genome can differ from those available in NCBI [21,76]. The highest assembly completeness of the fungal genome was achieved by Medaka. The parameter is crucial for an assembly, as it indicates the presence of the universal sequences of a taxon. Therefore, we regarded Medaka as the optimal polisher for *A. pullulans* strain #16.

Using Flye and Homopolish (two iterations), we received the first *C. lini* genome assembly from 1.65 Gb of raw ONT reads (N50 = 15.7 kb) basecalled at min\_qscore = 7. The assembly is 53.4 Mb-long, consists of 42 contigs with an N50 of 4.4 Mb, and has a completeness of 96.4%. We compared the assemblies of the three sequenced phytopathogens with the available genomes of the corresponding genus or species (Figure 4). The completeness of the *C. lini* assembly is close to the median value of 96.6% for the deposited assemblies produced from ONT data, either completely or partially. The obtained assembly has an N50 higher than that of the genomes of *C. fructicola* (N50 = 0.9 Mb) and *C. horii* (N50 = 3.1 Mb). However, the other four genomes from ONT data (*C. scovillei*, *C. australisnense* (*nom. inval.*), *C. echinoclaoe*, and *C. eleusines*) have slightly higher N50 values (4.8–5.7 Mb) and a lower or the same number of contigs (15–42). Nonetheless, only the *C. australisnense* assembly has an L50 lower than that of the obtained *C. lini* assembly. The *C. lini* assembly has higher completeness and contiguity than assemblies from Illumina and IonTorrent data. However, these parameters are close to those of the *C. acutatum* assembly from PacBio reads. Therefore, the analyzed assemblies demonstrated comparable contiguity. Most likely, it was the relatively high N50 of the received sequencing reads that positively influenced the reached genome contiguity.

For *A. pullulans* and *F. verticillioides* (*F. moniliforme*), the optimal assemblies were obtained by Canu–Medaka (total length of 29.5 Mb, 32 contigs, an N50 of 2.4 Mb, 99.4% completeness) and Canu–Homopolish (total length of 44.1 Mb, 54 contigs, an N50 of 4.4 Mb, 97.8% completeness), respectively. The assembly of *A. pullulans* #16 has the highest completeness (99.4%) among all the analyzed assemblies of *Auereobasidium* representatives.

This indicates that high coverage (more than 70×) with error-prone ONT reads results in improved accuracy of the resulting assembly. The N50 value of *A. pullulans* #16 is close to that of the *A. pullulans* AWRI4230 assembly from ONT data and higher than those of the assemblies from Illumina and PacBio reads. For *F. verticillioides* (*F. moniliforme*), the final assembly has a completeness (97.8%) equal to that of the *F. verticillioides* ASM2611958v1 genome from ONT reads and the *F. verticillioides* ASM2757160v1 genome from PacBio data. The assembly of *F. verticillioides* (*F. moniliforme*) #366 has an N50 slightly higher than the median value of 4.2 Mb for all the analyzed *Fusarium* genomes. Thus, the obtained amount of sequencing data for *A. pullulans* and *F. verticillioides* (*F. moniliforme*) (more than 70× and 110× genome coverage after basecalling, respectively) allowed us to construct assemblies with the main QCAST and BUSCO statistics close to those of the long-read genome assemblies from NCBI.

In this study, we assembled the genomes of the three flax pathogens (*C. lini*, *A. pullulans*, and *F. verticillioides* (*F. moniliforme*)) using ONT data and analyzed the influence of the basecalling read quality threshold and choice of assemblers and polishers on the assembly statistics. We defined the best approaches to obtain a genome assembly with the highest completeness and contiguity for each of the studied pathogens. As a result, high-quality assemblies of *C. lini* (53.37 Mb, N50 of 4.4 Mb, 96.4% complete), *A. pullulans* (29.5 Mb, N50 of 2.4 Mb, 99.4% complete), and *F. verticillioides* (*F. moniliforme*) (44.1 Mb long, N50 of 4.4 Mb, 97.8% complete) strains with known pathogenicity to flax were obtained for the first time.

## 5. Conclusions

Our results can guide the choice of a fungal de novo genome assembly strategy based on the use of ONT sequencing data. If low amounts of sequencing data were obtained (as in the case of *C. lini*), the genome coverage had more effect on assembly statistics than the quality of ONT reads. Therefore, using lower filtration threshold (`min_qscores`) values (7–8) for basecalling could be more effective than using the higher ones (9–10). Constructing the assemblies of *C. lini* demonstrated that Flye provided the best results at low genome coverage. Testing the assemblers' performance at high genome coverage (the datasets for *A. pullulans* and *F. verticillioides* (*F. moniliforme*)) showed that Canu achieved the best results. Polishing with Homopolish yields better results on assemblies with low initial (before polishing) BUSCO completeness values (as in the case of *C. lini*). When the initial value of BUSCO completeness was already high (as for *A. pullulans*), Medaka and Racon were the most useful tools for increasing it.

The assembled genomes of the flax pathogens—*C. lini*, *A. pullulans*, and *F. verticillioides* (*F. moniliforme*)—with high completeness and contiguity can be included in comparative genomic studies of plant pathogens. For differently virulent strains of *C. lini*, *A. pullulans*, and *F. verticillioides* (*F. moniliforme*), such an analysis could be useful in determining pathogenicity mechanisms. Thus, our study contributes to future screening for genetic markers of pathogenicity and diagnosing or controlling fungal diseases of crops. However, in the present study, we obtained genome assemblies only for single representatives of the three examined species. Moreover, only few high-quality genome assemblies of these species with known virulence are available in public databases. Therefore, for studying the association of pathogenicity with genome features, it is necessary to obtain high-quality assemblies for larger sets of *C. lini*, *A. pullulans*, and *F. verticillioides* (*F. moniliforme*) with known virulence and, ideally, to create pan-genomes of these species.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/jof9030301/s1>; Table S1: QCAST and BUSCO statistics of *Colletotrichum lini* strain #811 draft genome assemblies; Table S2: QCAST and BUSCO statistics of *Aureobasidium pullulans* strain #16 draft genome assemblies; Table S3: QCAST and BUSCO statistics of *Fusarium verticillioides* (*Fusarium moniliforme*) strain #366 draft genome assemblies; Table S4: QCAST and BUSCO statistics of *Colletotrichum lini* #811 polished genome assemblies; Table S5: QCAST and BUSCO statistics of *Aureobasidium pullulans* #16 polished genome assemblies; Table S6: QCAST and BUSCO statistics of *Fusarium verticillioides* (*Fusarium moniliforme*) #366 polished genome assemblies.

**Author Contributions:** Conceptualization, N.V.M., A.A.D. and E.M.D.; Methodology, E.A.S., E.N.P., T.A.R., L.P.K., R.O.N., D.A.Z., L.V.P., E.V.B. and E.M.D.; Investigation, E.A.S., E.N.P., T.A.R., L.P.K., A.A.Z., R.O.N., D.A.Z., L.V.P., A.A.T., E.V.B., N.V.M., A.A.D. and E.M.D.; Formal analysis, E.A.S., A.A.Z., A.A.T., N.V.M., A.A.D. and E.M.D.; Writing—original draft preparation, E.A.S. and E.M.D.; Writing—review and editing, E.A.S., N.V.M., A.A.D. and E.M.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** The work was financially supported by the Russian Science Foundation, grant number 22-16-00169.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The generated dataset for this study can be found in the NCBI database under the BioProject accession numbers PRJNA929545, PRJNA929546, and PRJNA929547.

**Acknowledgments:** We thank the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, EIMB RAS for providing the computing power and techniques for the data analysis. This work was performed using the equipment of EIMB RAS “Genome” center ([http://www.eimb.ru/ru1/ckp/ccu\\_genome\\_ce.php](http://www.eimb.ru/ru1/ckp/ccu_genome_ce.php), accessed on 10 January 2023).

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Lorenc, F.; Jarosova, M.; Bedrnicek, J.; Smetana, P.; Barta, J. Structural Characterization and Functional Properties of Flaxseed Hydrocolloids and Their Application. *Foods* **2022**, *11*, 2304. [[CrossRef](#)] [[PubMed](#)]
2. Ivanov, S.; Rashevskaya, T.; Makhonina, M. Flaxseed additive application in dairy products production. *Procedia Food Sci.* **2011**, *1*, 275–280. [[CrossRef](#)]
3. Saka, İ.; Baumgartner, B.; Özkaya, B. Usability of microfluidized flaxseed as a functional additive in bread. *J. Sci. Food Agric.* **2022**, *102*, 505–513. [[CrossRef](#)] [[PubMed](#)]
4. Kezimana, P.; Dmitriev, A.A.; Kudryavtseva, A.V.; Romanova, E.V.; Melnikova, N.V. Secoisolariciresinol Diglucoside of Flaxseed and Its Metabolites: Biosynthesis and Potential for Nutraceuticals. *Front. Genet.* **2018**, *9*, 641. [[CrossRef](#)] [[PubMed](#)]
5. Chaudhary, B.; Tripathi, M.; Pandey, S.; Bhandari, H.R.; Meena, D.; Prajapati, S. Uses of flax (*Linum usitatissimum*) after harvest. *Int. J. Trop. Agric.* **2016**, *34*, 159–164.
6. Hall, L.M.; Booker, H.; Siloto, R.M.P.; Jhala, A.J.; Weselake, R.J. Chapter 6—Flax (*Linum usitatissimum* L.). In *Industrial Oil Crops*; McKeon, T.A., Hayes, D.G., Hildebrand, D.F., Weselake, R.J., Eds.; AOCS Press: Urbana, IL, USA, 2016; pp. 157–194. [[CrossRef](#)]
7. da Silva, L.L.; Moreno, H.L.A.; Correia, H.L.N.; Santana, M.F.; de Queiroz, M.V. Colletotrichum: Species complexes, lifestyle, and peculiarities of some sources of genetic variability. *Appl. Microbiol. Biotechnol.* **2020**, *104*, 1891–1904. [[CrossRef](#)]
8. Leslie, J.F. Introductory biology of *Fusarium moniliforme*. *Adv. Exp. Med. Biol.* **1996**, *392*, 153–164. [[CrossRef](#)]
9. Alkemade, J.; Messmer, M.; Voegelé, R.; Finckh, M.; Hohmann, P. Genetic diversity of *Colletotrichum lupini* and its virulence on white and Andean lupin. *Sci. Rep.* **2021**, *11*, 13547. [[CrossRef](#)]
10. Wang, J.-H.; Ndoye, M.; Zhang, J.-B.; Li, H.-P.; Liao, Y.-C. Population structure and genetic diversity of the *Fusarium graminearum* species complex. *Toxins* **2011**, *3*, 1020–1037. [[CrossRef](#)]
11. Bashyal, B.; Aggarwal, R.; Banerjee, S.; Gupta, S.; Sharma, S. Pathogenicity, Ecology and Genetic Diversity of the *Fusarium* spp. Associated with an Emerging Bakanae Disease of Rice (*Oryza sativa* L.) in India. In *Microbial Diversity and Biotechnology in Food Security*; Springer: New Delhi, India, 2014; p. 307.
12. Schena, L.; Ippolito, A.; Zahavi, T.; Cohen, L.; Nigro, F.; Droby, S. Genetic diversity and biocontrol activity of *Aureobasidium pullulans* isolates against postharvest rots. *Postharvest Biol. Technol.* **1999**, *17*, 189–199. [[CrossRef](#)]
13. Russell, G.E. *Plant Breeding for Pest and Disease Resistance: Studies in the Agricultural and Food Sciences*; Butterworth-Heinemann: Oxford, UK, 2013.
14. Schoch, C.L.; Seifert, K.A.; Huhndorf, S.; Robert, V.; Spouge, J.L.; Levesque, C.A.; Chen, W.; Consortium, F.B.; List, F.B.C.A.; Bolchacova, E. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 6241–6246. [[CrossRef](#)]
15. Capella-Gutierrez, S.; Kauff, F.; Gabaldon, T. A phylogenomics approach for selecting robust sets of phylogenetic markers. *Nucleic Acids Res.* **2014**, *42*, e54. [[CrossRef](#)]
16. Damm, U.; O’Connell, R.J.; Groenewald, J.Z.; Crous, P.W. The *Colletotrichum destructivum* species complex—*Hemibiotrophic* pathogens of forage and field crops. *Stud. Mycol.* **2014**, *79*, 49–84. [[CrossRef](#)]
17. Xu, J. Fungal DNA barcoding. *Genome* **2016**, *59*, 913–932. [[CrossRef](#)]

18. Lücking, R.; Aime, M.C.; Robbertse, B.; Miller, A.N.; Ariyawansa, H.A.; Aoki, T.; Cardinali, G.; Crous, P.W.; Druzhinina, I.S.; Geiser, D.M.; et al. Unambiguous identification of fungi: Where do we stand and how accurate and precise is fungal DNA barcoding? *IMA Fungus* **2020**, *11*, 14. [[CrossRef](#)]
19. Novakovskiy, R.O.; Dvorianinova, E.M.; Rozhmina, T.A.; Kudryavtseva, L.P.; Gryzunov, A.A.; Pushkova, E.N.; Povkhova, L.V.; Snezhkina, A.V.; Krasnov, G.S.; Kudryavtseva, A.V.; et al. Data on genetic polymorphism of flax (*Linum usitatissimum* L.) pathogenic fungi of Fusarium, Colletotrichum, Aureobasidium, Septoria, and Melampsora genera. *Data Brief* **2020**, *31*, 105710. [[CrossRef](#)]
20. Onetto, C.A.; Schmidt, S.A.; Roach, M.J.; Borneman, A.R. Comparative genome analysis proposes three new Aureobasidium species isolated from grape juice. *FEMS Yeast Res.* **2020**, *20*, foaa052. [[CrossRef](#)]
21. Gostinčar, C.; Turk, M.; Zajc, J.; Gunde-Cimerman, N. Fifty Aureobasidium pullulans genomes reveal a recombining polyextremotolerant generalist. *Environ. Microbiol.* **2019**, *21*, 3638–3652. [[CrossRef](#)]
22. Liu, S.; Wu, B.; Lv, S.; Shen, Z.; Li, R.; Yi, G.; Li, C.; Guo, X. Genetic Diversity in FUB Genes of Fusarium oxysporum f. sp. cubense Suggests Horizontal Gene Transfer. *Front. Plant Sci.* **2019**, *10*, 1069. [[CrossRef](#)]
23. Laraba, I.; McCormick, S.P.; Vaughan, M.M.; Geiser, D.M.; O'Donnell, K. Phylogenetic diversity, trichothecene potential, and pathogenicity within Fusarium sambucinum species complex. *PLoS ONE* **2021**, *16*, e0245037. [[CrossRef](#)]
24. Akbar, A.; Hussain, S.; Ullah, K.; Fahim, M.; Ali, G.S. Detection, virulence and genetic diversity of Fusarium species infecting tomato in Northern Pakistan. *PLoS ONE* **2018**, *13*, e0203613. [[CrossRef](#)] [[PubMed](#)]
25. Dvorianinova, E.M.; Pushkova, E.N.; Novakovskiy, R.O.; Povkhova, L.V.; Bolsheva, N.L.; Kudryavtseva, L.P.; Rozhmina, T.A.; Melnikova, N.V.; Dmitriev, A.A. Nanopore and Illumina Genome Sequencing of Fusarium oxysporum f. sp. lini Strains of Different Virulence. *Front. Genet.* **2021**, *12*, 662928. [[CrossRef](#)] [[PubMed](#)]
26. Ren, X.; Zhu, Z.; Li, H.; Duan, C.; Wang, X. SSR marker development and analysis of genetic diversity of Fusarium verticillioides isolated from maize in China. *Sci. Agric. Sin.* **2012**, *45*, 52–66.
27. Ravi, R.K.; Walton, K.; Khosroheidari, M. MiSeq: A Next Generation Sequencing Platform for Genomic Analysis. *Methods Mol. Biol.* **2018**, *1706*, 223–232. [[CrossRef](#)]
28. Rhoads, A.; Au, K.F. PacBio Sequencing and Its Applications. *Genom. Proteom. Bioinform.* **2015**, *13*, 278–289. [[CrossRef](#)]
29. Lu, H.; Giordano, F.; Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genom. Proteom. Bioinform.* **2016**, *14*, 265–279. [[CrossRef](#)]
30. Payne, A.; Holmes, N.; Rakyar, V.; Loose, M. BulkVis: A graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **2019**, *35*, 2193–2198. [[CrossRef](#)]
31. Treangen, T.J.; Salzberg, S.L. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat. Rev. Genet.* **2011**, *13*, 36–46. [[CrossRef](#)]
32. Laver, T.; Harrison, J.; O'Neill, P.A.; Moore, K.; Farbos, A.; Paszkiewicz, K.; Studholme, D.J. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **2015**, *3*, 1–8. [[CrossRef](#)]
33. Bouso, J.M.; Planet, P.J. Complete nontuberculous mycobacteria whole genomes using an optimized DNA extraction protocol for long-read sequencing. *BMC Genom.* **2019**, *20*, 793. [[CrossRef](#)]
34. Helmersen, K.; Aamot, H.V. DNA extraction of microbial DNA directly from infected tissue: An optimized protocol for use in nanopore sequencing. *Sci. Rep.* **2020**, *10*, 2985. [[CrossRef](#)]
35. Russo, A.; Mayjonade, B.; Frei, D.; Potente, G.; Kellenberger, R.T.; Frachon, L.; Copetti, D.; Studer, B.; Frey, J.E.; Grossniklaus, U.; et al. Low-Input High-Molecular-Weight DNA Extraction for Long-Read Sequencing From Plants of Diverse Families. *Front. Plant Sci.* **2022**, *13*, 883897. [[CrossRef](#)]
36. Dvorianinova, E.M.; Bolsheva, N.L.; Pushkova, E.N.; Rozhmina, T.A.; Zhuchenko, A.A.; Novakovskiy, R.O.; Povkhova, L.V.; Sigova, E.A.; Zhernova, D.A.; Borkhert, E.V. Isolating *Linum usitatissimum* L. Nuclear DNA Enabled Assembling High-Quality Genome. *Int. J. Mol. Sci.* **2022**, *23*, 13244. [[CrossRef](#)]
37. Melnikova, N.V.; Pushkova, E.N.; Dvorianinova, E.M.; Beniaminov, A.D.; Novakovskiy, R.O.; Povkhova, L.V.; Bolsheva, N.L.; Snezhkina, A.V.; Kudryavtseva, A.V.; Krasnov, G.S.; et al. Genome Assembly and Sex-Determining Region of Male and Female Populus × sibirica. *Front. Plant Sci.* **2021**, *12*, 625416. [[CrossRef](#)]
38. Dmitriev, A.A.; Pushkova, E.N.; Novakovskiy, R.O.; Beniaminov, A.D.; Rozhmina, T.A.; Zhuchenko, A.A.; Bolsheva, N.L.; Muravenko, O.V.; Povkhova, L.V.; Dvorianinova, E.M.; et al. Genome Sequencing of Fiber Flax Cultivar Atlant Using Oxford Nanopore and Illumina Platforms. *Front. Genet.* **2021**, *11*, 590282. [[CrossRef](#)]
39. Wick, R.R.; Judd, L.M.; Holt, K.E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **2019**, *20*, 129. [[CrossRef](#)]
40. Dmitriev, A.; Pushkova, E.; Melnikova, N. Plant Genome Sequencing: Modern Technologies and Novel Opportunities for Breeding. *Mol. Biol.* **2022**, *56*, 495–507. [[CrossRef](#)]
41. Sun, Y.; Shang, L.; Zhu, Q.-H.; Fan, L.; Guo, L. Twenty years of plant genome sequencing: Achievements and challenges. *Trends Plant Sci.* **2022**, *27*, 391–401. [[CrossRef](#)]
42. Krasnov, G.S.; Pushkova, E.N.; Novakovskiy, R.O.; Kudryavtseva, L.P.; Rozhmina, T.A.; Dvorianinova, E.M.; Povkhova, L.V.; Kudryavtseva, A.V.; Dmitriev, A.A.; Melnikova, N.V. High-Quality Genome Assembly of Fusarium oxysporum f. sp. lini. *Front. Genet.* **2020**, *11*, 959. [[CrossRef](#)]



43. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722–736. [[CrossRef](#)]
44. Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P.A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **2019**, *37*, 540–546. [[CrossRef](#)] [[PubMed](#)]
45. Vaser, R.; Šikić, M. Time- and memory-efficient genome assembly with Raven. *Nat. Comput. Sci.* **2021**, *1*, 332–336. [[CrossRef](#)]
46. Ruan, J.; Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **2020**, *17*, 155–158. [[CrossRef](#)] [[PubMed](#)]
47. Vaser, R.; Šikić, M. Yet another de novo genome assembler. In Proceedings of the 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), Dubrovnik, Croatia, 23–25 September 2019; pp. 147–151.
48. Liu, H.; Wu, S.; Li, A.; Ruan, J. SMARTdenovo: A de novo assembler using long noisy reads. *Gigabyte* **2021**, 1–9. [[CrossRef](#)]
49. Li, H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **2016**, *32*, 2103–2110. [[CrossRef](#)]
50. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [[CrossRef](#)]
51. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [[CrossRef](#)]
52. Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **2017**, *27*, 737–746. [[CrossRef](#)]
53. 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 Biol.* **2021**, *22*, 95. [[CrossRef](#)]
54. Hu, J.; Fan, J.; Sun, Z.; Liu, S. NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **2020**, *36*, 2253–2255. [[CrossRef](#)]
55. Shafin, K.; Pesout, T.; Chang, P.C. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **2021**, *18*, 1322–1332. [[CrossRef](#)]
56. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **2021**, *37*, 4572–4574. [[CrossRef](#)]
57. O’Connell, R.J.; Thon, M.R.; Hacquard, S.; Amyotte, S.G.; Kleemann, J.; Torres, M.F.; Damm, U.; Buiate, E.A.; Epstein, L.; Alkan, N.; et al. Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. *Nat. Genet.* **2012**, *44*, 1060–1065. [[CrossRef](#)]
58. Frantzeskakis, L.; Kusch, S.; Panstruga, R. The need for speed: Compartmentalized genome evolution in filamentous phytopathogens. *Mol. Plant Pathol.* **2019**, *20*, 3–7. [[CrossRef](#)]
59. Qian, W.; Jia, Y.; Ren, S.-X.; He, Y.-Q.; Feng, J.-X.; Lu, L.-F.; Sun, Q.; Ying, G.; Tang, D.-J.; Tang, H. Comparative and functional genomic analyses of the pathogenicity of phytopathogen *Xanthomonas campestris* pv. *campestris*. *Genome Res.* **2005**, *15*, 757–767. [[CrossRef](#)]
60. Mat Razali, N.; Cheah, B.H.; Nadarajah, K. Transposable Elements Adaptive Role in Genome Plasticity, Pathogenicity and Evolution in Fungal Phytopathogens. *Int. J. Mol. Sci.* **2019**, *20*, 3597. [[CrossRef](#)]
61. Lefebvre, F.; Joly, D.L.; Labbé, C.; Teichmann, B.; Linning, R.; Belzile, F.; Bakkeren, G.; Bélanger, R.R. The Transition from a Phytopathogenic Smut Ancestor to an Anamorphic Biocontrol Agent Deciphered by Comparative Whole-Genome Analysis. *Plant Cell* **2013**, *25*, 1946–1959. [[CrossRef](#)]
62. Thynne, E.; Saur, I.M.L.; Simbaqueba, J.; Ogilvie, H.A.; Gonzalez-Cendales, Y.; Mead, O.; Taranto, A.; Catanzariti, A.-M.; McDonald, M.C.; Schwessinger, B.; et al. Fungal phytopathogens encode functional homologues of plant rapid alkalization factor (RALF) peptides. *Mol. Plant Pathol.* **2017**, *18*, 811–824. [[CrossRef](#)]
63. Potgieter, L.; Feurtey, A.; Dutheil, J.Y.; Stukenbrock, E.H. On Variant Discovery in Genomes of Fungal Plant Pathogens. *Front. Microbiol.* **2020**, *11*, 626. [[CrossRef](#)]
64. Ye, Z.; Pan, Y.; Zhang, Y.; Cui, H.; Jin, G.; McHardy, A.C.; Fan, L.; Yu, X. Comparative whole-genome analysis reveals artificial selection effects on *Ustilago esculenta* genome. *DNA Res.* **2017**, *24*, 635–648. [[CrossRef](#)]
65. Batista, T.M.; Hilario, H.O.; de Brito, G.A.M.; Moreira, R.G.; Furtado, C.; de Menezes, G.C.A.; Rosa, C.A.; Rosa, L.H.; Franco, G.R. Whole-genome sequencing of the endemic Antarctic fungus *Antarctomyces pellizariae* reveals an ice-binding protein, a scarce set of secondary metabolites gene clusters and provides insights on Thelebolales phylogeny. *Genomics* **2020**, *112*, 2915–2921. [[CrossRef](#)] [[PubMed](#)]
66. Drott, M.T.; Satterlee, T.R.; Skerker, J.M.; Pfannenstiel, B.T.; Glass, N.L.; Keller, N.P.; Milgroom, M.G. The Frequency of Sex: Population Genomics Reveals Differences in Recombination and Population Structure of the Aflatoxin-Producing Fungus *Aspergillus flavus*. *mBio* **2020**, *11*, e00963-20. [[CrossRef](#)] [[PubMed](#)]
67. Wyka, S.A.; Mondo, S.J.; Liu, M.; Dettman, J.; Nalam, V.; Broders, K.D. Whole-Genome Comparisons of Ergot Fungi Reveals the Divergence and Evolution of Species within the Genus *Claviceps* Are the Result of Varying Mechanisms Driving Genome Evolution and Host Range Expansion. *Genome Biol. Evol.* **2021**, *13*, evaa267. [[CrossRef](#)] [[PubMed](#)]
68. Meleshko, D.; Korobeynikov, A. Benchmarking state-of-the-art approaches for norovirus genome assembly in metagenome sample. *bioRxiv* **2022**. [[CrossRef](#)]
69. Gavrielatos, M.; Kyriakidis, K.; Spandidos, D.A.; Michalopoulos, I. Benchmarking of next and third generation sequencing technologies and their associated algorithms for *de novo* genome assembly. *Mol. Med. Rep.* **2021**, *23*, 251. [[CrossRef](#)]

70. Breckell, G.L.; Silander, O.K. Do You Want to Build a Genome? Benchmarking Hybrid Bacterial Genome Assembly Methods. *bioRxiv* **2021**. [[CrossRef](#)]
71. Zhang, X.; Liu, C.-G.; Yang, S.-H.; Wang, X.; Bai, F.-W.; Wang, Z. Benchmarking of long-read sequencing, assemblers and polishers for yeast genome. *Brief. Bioinform.* **2022**, *23*, bbac146. [[CrossRef](#)]
72. Deng, Z.-L.; Dhingra, A.; Fritz, A.; Götting, J.; Münch, P.C.; Steinbrück, L.; Schulz, T.F.; Ganzenmüller, T.; McHardy, A.C. Evaluating assembly and variant calling software for strain-resolved analysis of large DNA viruses. *Brief. Bioinform.* **2020**, *22*, bbaa123. [[CrossRef](#)]
73. Gettle, N.; Gallone, B.; Verstrepen, K.J.; Stelkens, R. Harnessing the power of technical and natural variation in 116 yeast datasets to benchmark long read assembly pipelines. *bioRxiv* **2022**. [[CrossRef](#)]
74. Jung, H.; Jeon, M.-S.; Hodgett, M.; Waterhouse, P.; Eyun, S.-I. Comparative Evaluation of Genome Assemblers from Long-Read Sequencing for Plants and Crops. *J. Agric. Food Chem.* **2020**, *68*, 7670–7677. [[CrossRef](#)]
75. Chen, Z.; Erickson, D.L.; Meng, J. Benchmarking Long-Read Assemblers for Genomic Analyses of Bacterial Pathogens Using Oxford Nanopore Sequencing. *Int. J. Mol. Sci.* **2020**, *21*, 9161. [[CrossRef](#)]
76. Urzì, C.; De Leo, F.; Passo, C.L.; Criseo, G. Intra-specific diversity of *Aureobasidium pullulans* strains isolated from rocks and other habitats assessed by physiological methods and by random amplified polymorphic DNA (RAPD). *J. Microbiol. Methods* **1999**, *36*, 95–105. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.