Genome assembly of *Colletotrichum lini* from long Nanopore reads

Sigova E.A.[1, 2]*, Dvorianinova E.M.[1, 2], Rozhmina T.A.[1, 3], Kudryavtseva L.P.[3], Melnikova N.V.[1], Dmitriev A.A.[1]

[1] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

[2] Moscow Institute of Physics and Technology, Moscow, Russia

[3] Federal Research Center for Bast Fiber Crops, Torzhok, Russia

* sigova.ea@phystech.edu

Novosibirsk 2022

Motivation and Aim:

Colletotrichum lini is the malicious flax anthracnose causative agent. Studying the fungus at the genetic level is vital for the successful disease control. However, the lack of the *C. lini* whole genome sequence hinders extensive molecular research on the pathogen. Therefore, our aim was to obtain the first genome assembly of *C. lini* using the Oxford Nanopore Technologies (ONT) sequencing platform.

Methods and Algorithms:

C. lini highly pathogenic strain #811 was provided by the Institute for Flax (Torzhok, Russia).

1

Pure high-molecular DNA was obtained according to our previously developed protocol. The spectrophotometry (Nanodrop) and fluorometry (Qubit) methods were used to evaluate the quality and quantity of the extracted DNA

2

DNA libraries were prepared and sequenced on the ONT (MinION instrument, FLO-MIN-106 R9.4.1 flow-cell) platform according to the manufacturer's protocol

3

The obtained reads were bacecalled using Guppy 6.0.1 with different quality filtration thresholds (min_qscore, i.e. the minimum average read quality of a basecalled read, was taken in the range from 7 to 10)

4

Porechop 0.2.4 was used for adapter removing

5

Draft assemblies for each minimum quality score value were performed using
Canu 2.2, Flye 2.8.1, Raven 1.5.1, Shasta 0.8.0, Wtdbg-cns 1.1 (Wtdbg2 0.0), NextDenovo 2.5.0, Miniasm 0.3-r179, Ra 0.2.1, and SmartDenovo tools

6

BUSCO 5.3.2 and QUAST 5.0.2 were used to analyze the quality of the obtained assemblies

7

Draft assembly obtained with Flye (minimum quality score value = 7) was polished using
Medaka 1.5.0, Racon 1.4.20, Homopolish 0.3.4, MarginPolish 1.3.0, NextPolish 1.4.0, Nanopolish 0.13.3, Pepper 0.0.6

8

BUSCO 5.3.2 and QUAST 5.0.2 were used to analyze the quality of the obtained polished assemblies

Results: We obtained 1.7 Gb of raw ONT reads with an N50 of 15.7 kb.
Genomes assembled from this data with different tools and quality filtration thresholds have the statistics shown in the Table.

Q	Assembler	Assembly length, Mb	BUSCO, %	Number of contigs	N50, Mb	Q	Assembler	Assembly length, Mb	BUSCO, %	Number of contigs	N50, Mb
10	*Canu*	*48.1*	*80.9*	*443*	*0.15*	8	*Canu*	*52.9*	*88.8*	*174*	*0.53*
	Flye	*52.9*	*89.7*	*111*	*1.11*		*Flye*	*53.4*	*93.7*	*37*	*3.41*
	Raven	32.2	50.8	351	0.11		*Raven*	*51.4*	*85.5*	*160*	*0.49*
	Wtdbg2	50.3	67.4	178	0.59		Wtdbg2	52.1	74.8	81	2.10
	Shasta	27.1	41.3	691	0.06		Shasta	45.0	59.3	709	0.10
	NextDenovo	-	-	-	-		NextDenovo	40.1	69.8	135	0.37
	Ra	26.8	45.4	248	0.12		*Ra*	*50.5*	*82.2*	*177*	*0.43*
	SmartDenovo	401.6	75.8	24544	0.02		SmartDenovo	622.4	76.9	38180	0.02
	Miniasm	27.9	20.2	290	0.11		Miniasm	48.5	27.0	191	0.39
9	*Canu*	*51.7*	*86.6*	*260*	*0.33*	7	*Canu*	*53.5*	*88.8*	*134*	*0.81*
	Flye	*53.1*	*92.7*	*48*	*3.31*		***Flye***	***53.4***	***93.5***	***42***	***4.44***
	Raven	45.5	71.6	285	0.20		*Raven*	*52.8*	*89.9*	*90*	*0.95*
	Wtdbg2	51.6	73.6	98	1.30		Wtdbg2	51.8	71.0	40	3.20
	Shasta	42.1	55.5	762	0.08		Shasta	46.1	61.6	663	0.11
	NextDenovo	0.08	0.3	1	0.08		*NextDenovo*	*52.3*	*91.6*	*69*	*1.24*
	Ra	43.4	70.7	274	0.19		*Ra*	*52.5*	*85.7*	*98*	*0.87*
	SmartDenovo	507.7	76.6	31063	0.02		SmartDenovo	753.5	77.1	46314	0.02
	Miniasm	41.9	26.3	297	0.18		Miniasm	50.5	23.6	96	0.84

The assemblers gave better results at lower min_qscore values (7-8), since at high min_qscore values (9-10) the coverage was insufficient to obtain a quality assembly

The average length of the assemblies with BUSCO > 80% was 52.2 Mb (48.1-53.5 Mb)

The highest assembly completeness for each min_qscore was achieved by Flye (up to 93.7%)

For the basecalled data with a min_qscore of 7, Flye produced the most contiguous assembly: N50 of 4.4 Mb for a total length of 53.4 Mb, 42 contigs

The most contiguous and complete assembly obtained with Flye (min_qscore = 7) was polished with various polishing tools.

Polisher	no reference				with reference (*C. higginsianum* GCA_001672515.1)						
	Length, Mb	Contigs	BUSCO		Genome fraction %	Genomic features	Misassemblies	Misassembled contigs	Misassembled contigs length, Mb	Mismatches per 100 kbp	Indels per 100 kbp
			Complete, %	Fragmented, %							
Assembly Flye, min_qscore = 7	53.35	42	93.5	2.2	56.3	45388	1520	13	41.9	4526	207
Medaka	53.39	42	89.4	5.1	56.8	49422	1529	14	42.3	4449	191
Medaka2	53.39	42	89.4	5.1	56.7	49414	1523	14	42.3	4448	190
Medaka3	53.39	42	89.4	5.1	56.7	49408	1529	14	42.3	4446	190
Racon	53.43	36	86.1	7.1	56.5	46356	1532	12	42.0	4483	199
Racon2	53.40	34	86.0	7.2	56.5	46482	1534	12	42.0	4487	200
Racon3	53.39	33	85.7	7.3	56.4	46224	1520	12	42.0	4479	198
Homopolish	53.36	42	96.3	0.9	61.1	64802	1925	17	50.0	4131	124
Homopolish2	**53.37**	**42**	**96.4**	**0.9**	**61.7**	**66188**	**1923**	**17**	**50.1**	**4108**	**116**
Homopolish3	53.37	42	96.4	0.8	61.8	66274	1982	18	50.9	4111	117
MarginPolish	53.42	37	85.4	7.2	56.4	46334	1531	12	42.1	4515	199
MarginPolish2	53.40	33	85.3	7.2	56.5	46402	1533	12	42.1	4513	200
MarginPolish3	53.40	33	85.3	7.2	56.4	46356	1531	12	42.1	4518	200
NextPolish	53.37	42	87.6	6.1	56.6	47748	1517	13	41.9	4462	192
NextPolish2	53.37	42	87.5	6.2	56.6	47776	1512	13	41.9	4459	192
NextPolish3	53.37	42	87.6	6.1	56.6	47844	1514	13	41.9	4458	193
Nanopolish	53.35	42	85.0	7.4	56.3	45532	1519	13	41.9	4521	204
Nanopolish2	53.35	42	85.0	7.4	56.3	45526	1516	13	41.9	4521	204
Nanopolish3	53.35	42	85.0	7.4	56.3	45526	1517	13	41.9	4523	204
Pepper	53.26	35	86.9	5.7	56.0	46496	1531	12	41.9	4546	198
Pepper2	53.23	30	89.1	4.5	56.2	48284	1525	12	41.9	4490	190
Pepper3	53.21	29	87.3	5.5	56.0	46876	1534	12	41.9	4539	196

The highest completeness of polished assembly (96.4%) and the highest percent of covered reference genome fraction (61.8%) were achieved by Homopolish. The smallest amount of mismatches per 100 kbp and indels per 100 kbp was also achieved by Homopolish (second iteration).

The genome assembly of *C. lini* strain #811 obtained with Flye from the ONT data basecalled with min_qscore = 7 and polished with Homopolish twice can be considered the most complete and contiguous: N50 of 4.4 Mb for a total length of 53.4 Mb, 42 contigs, completeness 96.4%.

Conclusion: We obtained the first *C. lini* genome assembly from long ONT reads. This knowledge is a starting point for further detailed research on *C. lini* and the flax-pathogen interaction.

Acknowledgements: This work was financially supported by the Russian Science Foundation, grant 22-16-00169.