

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, RAS, 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

Key words: flax pathogens, *Colletotrichum lini*, genome assembly, Oxford Nanopore

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). Pure high-molecular DNA was obtained according to our previously developed protocol. The spectrophotometry (Nanodrop) and fluorometry (Qubit 4) methods were used to evaluate the quality and quantity of the extracted DNA. 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. The obtained reads were basecalled 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). Porechop 0.2.4 was used for adapter removing. 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. BUSCO 5.3.2 and QUAST 5.0.2 were used to analyze the quality of the obtained assemblies.

Results: We obtained 1.7 Gbases (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 Table 1.

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 highest assembly completeness for each min_qscore was achieved by Flye (up to 93.7 %), the lowest – by Miniasm (less than 30 %). The average length of the assemblies with BUSCO > 80 % was 52.2 Mb (48.1–53.5 Mb). For the basecalled data with a min_qscore of 7, Flye also produced the most contiguous assembly: N50 of 4.4 Mb for a total length of 53.4 Mb, 42 contigs. Thus, the genome assembly of *C. lini* strain #811 obtained with Flye from the ONT data basecalled with min_qscore = 7 can be considered the most complete and contiguous.

Table 1. Statistics on the obtained genome assemblies

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

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 No. 22-16-00169.