Check for updates

DATA NOTE

The genome sequence of the bootlace worm, *Lineus*

longissimus (Gunnerus, 1770) [version 1; peer review: 2

approved]

Dominic Kwiatkowski¹, Mark Blaxter¹, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective,

Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

¹Wellcome Sanger Institute, Cambridge, CB10 1SA, UK

v1	First pub https://do	olished: 14 Oct 2021, 6 :272 oi.org/10.12688/wellcomeopenres.17193.1	Open Peer Review Reviewer Status 🗹 🗸			
	Latest pu https://do	ublished: 14 Oct 2021, 6 :272 oi.org/10.12688/wellcomeopenres.17193.1				
0 h c				Invited F	Reviewers	
We p	oresent a	genome assembly from an individual Lineus longissimus		1	2	
(the Line majo pseu	bootlace idae). The ority of th udomolee	worm; Nemertea; Pilidiophora; Heteronemertea; e genome sequence is 391 megabases in span. The ne assembly is scaffolded into 19 chromosomal cules.	version 1 14 Oct 2021	report	report	
<mark>Keywords</mark> Lineus longissimus, bootlace worm, genome sequence, chromosomal			1. Jose Maria University o	1. Jose Maria Martin-Duran (D), Queen Mary University of London, London, UK		
		This article is included in the Tree of Life	2. Ferdinand London, Lo	2. Ferdinand Marletaz , University College London, London, UK Any reports and responses or comments on the		
2	a Marcaller	gateway.	Any reports an			

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Kwiatkowski D: Investigation, Resources; Blaxter M: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Supervision, Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (206194) and the Darwin Tree of Life Discretionary Award (218328).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Kwiatkowski D *et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Kwiatkowski D, Blaxter M, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of the bootlace worm**, *Lineus longissimus* (Gunnerus, 1770) [version 1; peer review: 2 approved] Wellcome Open Research 2021, 6:272 https://doi.org/10.12688/wellcomeopenres.17193.1

First published: 14 Oct 2021, 6:272 https://doi.org/10.12688/wellcomeopenres.17193.1

Species taxonomy

Eukaryota; Metazoa; Spiralia; Lophotrochozoa; Nemertea; Pilidiophora; Heteronemertea; Lineidae; Lineus; *Lineus longissimus* (Gunnerus, 1770) (NCBI:txid88925).

Introduction

Lineus longissimus (Lineidae, Heteronemertea, Nemertea) is a predatory ribbon worm, renowned as being the longest animal in the British and Irish biota, as individual specimens can exceed 25 m when fully extended. Their extensibility results in part from their unsegmented body form, where the coleom is limited to a rhynchocoel associated with the eversible proboscis. Phylum Nemertea is placed as sister to either Mollusca or Annelida within Eutrochozoa in the Spiralia (Struck & Fisse, 2008). Nemertea includes only 1200 described species worldwide that play important ecological roles in littoral and benthic communities (Gibson, 1972). Nemerteans have been studied for their ability to regenerate body parts (Zattara et al., 2019) and the potent venom neurotoxins secreted from the glandular epithelium of the proboscis (Stricker & Cloney, 1983). The nemertide alpha-1 toxin from L. longissimus shows promise as an insecticide (Bell et al., 2021).

Genome sequence report

The genome was sequenced from a single L. longissimus of unknown sex collected from White Bay, Great Cumbrae, North Avreshire, Scotland (latitude 55.790409, longitude -4.908826). A total of 79-fold coverage in Pacific Biosciences single-molecule long reads and 107-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 67 missing/misjoins and removed 5 haplotypic duplications, reducing the assembly length by 0.46% and the scaffold number by 63.10%, and increasing the scaffold N50 by 53.76%. The final assembly has a total length of 391 Mb in 32 sequence scaffolds with a scaffold N50 of 21 Mb (Table 1). Of the assembly sequence, 99.81% was assigned to 19 chromosomal-level scaffolds, representing 19 autosomes (numbered by sequence length) (Figure 1-Figure 4; Table 2). The assembly has a BUSCO (Simão et al., 2015) v5.1.2 completeness of 96.5% using the metazoa_odb10 reference set. While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

Methods

A single *L. longissimus* of unknown sex was collected from White Bay, Great Cumbrae, North Ayreshire, Scotland (latitude 55.790409, longitude -4.908826) by Dominic Kwiatkowski, Wellcome Sanger Institute (WSI) and preserved on dry ice prior to transfer to the WSI.

DNA was extracted from midbody tissue at the Tree of Life laboratory, WSI. The tnLinLong1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Samples were cryogenically disrupted to a fine powder using a Covaris cryoPREP Automated Dry Pulveriser, receiving

Table 1. Genome data for Lineus longissimus, tnLinLong1.1.

Project accession data				
Assembly identifier	tnLinLong1.1			
Species	Lineus longissimus			
Specimen	tnLinLong1			
NCBI taxonomy ID	NCBI:txid88925			
BioProject	PRJEB45185			
BioSample ID	SAMEA7522833			
Isolate information	Unknown sex, anterior/mid/ posterior body			
Raw data accessions				
PacificBiosciences SEQUEL II	ERR6412039, ERR6436382, ERR6436383			
10X Genomics Illumina	ERR6054915-ERR6054918			
Hi-C Illumina	ERR6054914			
Genome assembly				
Assembly accession	GCA_910592395.1			
Accession of alternate haplotype	GCA_910592375.1			
Span (Mb)	391			
Number of contigs	109			
Contig N50 length (Mb)	10			
Number of scaffolds	32			
Scaffold N50 length (Mb)	21			
Longest scaffold (Mb)	29			
BUSCO* genome score	C:96.5%[S:96.1%,D:0.4%],F:2.2%,M: 1.3%,n:954			

*BUSCO scores based on the metazoa_odb10 BUSCO set using v5.1.2. C= complete [S= single copy, D=duplicated], F=fragmented, M=missing, n=number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/tnLinLong1.1/dataset/ CAJUZJ01/busco.

multiple impacts. Fragment size analysis of 0.01-0.5 ng then performed using an Agilent of DNA was FemtoPulse. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. Low molecular weight DNA was removed from a 200-ng aliquot of extracted DNA using 0.8X AMpure XP purification kit prior to 10X Chromium sequencing; a minimum of 50 ng DNA was submitted for 10X sequencing. HMW DNA was sheared into an average fragment size between 12-20 kb in a Megaruptor 3 system with speed setting 30. Sheared DNA was purified by solid-phase reversible immobilisation using AMPure PB beads with a 1.8X ratio of beads to sample to remove the shorter fragments and concentrate the DNA sample. The concentration of the sheared and purified DNA



Figure 1. Genome assembly of *Lineus longissimus*, tnLinLong1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/tnLinLong1.1/dataset/ CAJUZJ01/snail.



Figure 2. Genome assembly of Lineus longissimus, tnLinLong1.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/tnLinLong1.1/dataset/CAJUZJ01/blob. It should be noted that the tnLinLong1 genome is the first of its phylum (Nemertea) to be assembled, meaning there are few/no other related sequences for BlobToolKit to call in INSDC databases. The resultant identification of Mollusca, Chordata and Brachiopoda sequences here reflects the divergence of Nemertea from other phyla.



Figure 3. Genome assembly of Lineus longissimus, tnLinLong1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all chromosomes. Coloured lines show cumulative lengths of chromosomes assigned to each phylum using the buscogenes taxrule. Since the tnLinLong1 genome is the first of its phylum (Nemertea) to be assembled, there are few/no other related sequences for BlobToolKit to call in INSDC databases. The resultant identification of Mollusca, Chordata and Brachiopoda sequences here reflects the divergence of Nemertea from other phyla. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/tnLinLong1.1/dataset/CAJUZJ01/cumulative.



Figure 4. Genome assembly of Lineus longissimus, tnLinLong1.1: Hi-C contact map. Hi-C contact map of the tnLinLong1.1 assembly, visualised in HiGlass.

INSDC accession	Chromosome	Size (Mb)	GC%
OU342992.1	1	29.45	41.0
OU342993.1	2	29.19	41.1
OU342994.1	3	29.08	41.2
OU342995.1	4	25.60	41.2
OU342996.1	5	23.94	41.3
OU342997.1	6	21.66	41.8
OU342998.1	7	21.16	41.3
OU342999.1	8	20.96	41.3
OU343000.1	9	20.55	41.4
OU343001.1	10	19.41	41.4
OU343002.1	11	19.19	41.5
OU343003.1	12	17.98	41.6
OU343004.1	13	16.85	41.6
OU343005.1	14	16.76	41.5
OU343006.1	15	16.29	42.0
OU343007.1	16	16.53	41.5
OU343008.1	17	16.30	41.6
OU343009.1	18	15.89	41.4
OU343010.1	19	13.63	41.7
OU343011.1	MT	0.02	35.3
_	Unplaced	0.74	42.5

Table 2. Chromosomal pseudomolecules in the genome assembly of *Lineus longissimus*, tnLinLong1.1.

was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud sequencing libraries were constructed according to the manufacturers' instructions. Sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II and Illumina HiSeq X instruments. Hi-C data were generated using the Arima v2.0 kit and sequenced on HiSeq X.

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021), haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). The assembly was polished with the 10X Genomics Illumina data by aligning to the assembly with longranger align, calling variants with freebayes (Garrison & Marth, 2012). One round of the Illumina polishing was applied. Scaffolding with Hi-C data

(Rao *et al.*, 2014) was carried out with SALSA2 (Ghurye *et al.*, 2019). The assembly was checked for contamination and corrected using the gEVAL system (Chow *et al.*, 2016) as described previously (Howe *et al.*, 2021). Manual curation was performed using gEVAL, HiGlass (Kerpedjiev *et al.*, 2018) and Pretext. The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2021). The genome was analysed and BUSCO scores generated within the BlobToolKit environment (Challis *et al.*, 2020). Table 3 contains a list of all software tool versions used, where appropriate.

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the Darwin Tree of Life Project Sampling Code of Practice. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project. Each transfer

Software tool	Version	Source
Hifiasm	0.12	Cheng <i>et al.,</i> 2021
purge_dups	1.2.3	Guan <i>et al.,</i> 2020
longranger	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/ other-pipelines
freebayes	v1.3.1-17- gaa2ace8	Garrison & Marth, 2012
MitoHiFi	2.0	Uliano-Silva et al., 2021
SALSA2	2.2	Ghurye <i>et al.</i> , 2019
gEVAL	N/A	Chow <i>et al.</i> , 2016
HiGlass	1.11.6	Kerpedjiev et al., 2018
PretextView	0.2.x	https://github.com/wtsi-hpag/PretextView
BlobToolKit	2.6.2	Challis <i>et al.</i> , 2020

Table 3. Software tools used.

of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances other Darwin Tree of Life collaborators.

Data availability

European Nucleotide Archive: Lineus longissimus (bootlace worm). Accession number PRJEB45185: https://identifiers.org/ena.embl:PRJEB45185

The genome sequence is released openly for reuse. The *L. longissimus* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1.

Acknowledgments

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893704.

Members of the Wellcome Sanger Institute Tree of Life programme collective are listed here: https://doi.org/10.5281/zenodo.5377053.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790456.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013542.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783559.

References

Bell J. Sukiran NA, Walsh S, et al.: The Insecticidal Activity of Recombinant Nemertide Toxin α-1 from Lineus Longissimus towards Pests and Beneficial Species. Toxicon. 2021; **197**: 79–86.

PubMed Abstract | Publisher Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit-Interactive Quality** Assessment of Genome Assemblies. *G3 (Bethesda)*. 2020; **10**(4): 1361–74. PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

Chow W, Brugger K, Caccamo M, et al.: gEVAL — a Web-Based Browser for

Evaluating Genome Assemblies. *Bioinformatics*. 2016; **32**(16): 2508–10. PubMed Abstract | Publisher Full Text | Free Full Text Garrison E, Marth G: Haplotype-Based Variant Detection from Short-Read Sequencing. arXiv: 1207.3907. 2012. Reference Source

Ghurye J, Rhie A, Walenz BP, et al.: Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly. PLoS Comput Biol. 2019; 15(8): e1007273.

PubMed Abstract | Publisher Full Text | Free Full Text Gibson R: Nemerteans. Hutchinson Radius. 1972.

Guan D, McCarthy SA, Wood J, et al.: Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies. *Bioinformatics*. 2020; 36(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): 170–75.

2896–98. PubMed Abstract | Publisher Full Text | Free Full Text

Howe K, Chow W, Collins J, et al.: Significantly Improving the Quality of Genome Assemblies through Curation. *Gigascience*. 2021; **10**(1): giaa153. PubMed Abstract | Publisher Full Text | Free Full Text

Kerpedjiev P, Abdennur N, Lekschas F, et al.: HiGlass: Web-Based Visual Exploration and Analysis of Genome Interaction Maps. Genome Biol. 2018; 19(1): 125. PubMed Abstract | Publisher Full Text | Free Full Text

Rao SSP, Huntley MH, Durand NC, et al.: A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014; **159**(7): 1665-80.

PubMed Abstract | Publisher Full Text | Free Full Text

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.

Bioinformatics. 2015; **31**(19): 3210–12. **PubMed Abstract | Publisher Full Text** Stricker SA, Cloney RA: **The Ultrastructure of Venom-Producing Cells in** Parametres Peregrina (Nemertea, Hoplonemertea). J Morphol. 1983; **177**(1): 89–107.

PubMed Abstract | Publisher Full Text

Struck TH, Fisse F: **Phylogenetic Position of Nemertea Derived from Phylogenomic Data.** *Mol Biol Evol.* 2008; **25**(4): 728–36. **PubMed Abstract | Publisher Full Text**

Uliano-Silva M, Nunes JGF, Krasheninnikova K, et al.: marcelauliano/MitoHiFi: mitohifi_v2.0. 2021. **Publisher Full Text**

Zattara EE, Fernández-Álvarez FA, Hiebert TC, et al.: A Phylum-Wide Survey Reveals Multiple Independent Gains of Head Regeneration in Nemertea. Proc Biol Sci. 2019; **286**(1898): 20182524.

PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 15 November 2021

https://doi.org/10.21956/wellcomeopenres.18996.r46958

© **2021 Marletaz F.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Ferdinand Marletaz 匝

Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London, UK

I think this provides an accurate report of the genome assembly of the nemertean *Lineus longissimus*.

"Their extensibility results in part from their unsegmented body form, where the coleom is limited to a rhynchocoel associated with the eversible proboscis "

-> I think it is not so opportune to link extensibility with reduced coelome, a lot of animals with reduced or absent coelom and they are not so extensibly (planarians, etc...), so I would suggest to rephrase

"Phylum Nemertea is placed as sister to either Mollusca or Annelida within Eutrochozoa in the Spiralia (Struck & Fisse, 2008). "

-> I think you should cite some more recent references and discuss the alternative phylogenetic positions as suggested by Dr. Martin-Duran.

It would be interesting to mention the estimated polymorphism and the estimated genome size.

I would also recommend to cite Luo et al. (2018)¹ in the introduction as it is the first nermetean genome characterised.

References

1. Luo YJ, Kanda M, Koyanagi R, Hisata K, et al.: Nemertean and phoronid genomes reveal lophotrochozoan evolution and the origin of bilaterian heads.*Nat Ecol Evol*. 2018; **2** (1): 141-151 PubMed Abstract | Publisher Full Text

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others? $\ensuremath{\mathsf{Yes}}$

Are the datasets clearly presented in a useable and accessible format? Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: animal phylogeny, regulatory and comparative genomics, spiralians, evodevo

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 19 October 2021

https://doi.org/10.21956/wellcomeopenres.18996.r46445

© **2021 Martin-Duran J.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Jose Maria Martin-Duran 匝

School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

This is a clear and concise report on the sequencing of Lineus longissiums genome, an iconic species in the European biota. Methods and the resulting assembly are gold-standard. I only have three minor comments:

- L. longissimus might actually be the longest animal species ever recorded (see https://www.guinnessworldrecords.com/world-records/longest-animal)
- The phylogenetic position of Nemerteans is contentious. Recent phylogenomic analyses place them either as sister to Platyhelminthes (https://doi.org/10.1016/j.cub.2018.11.042¹) or Lophophorata (https://doi.org/10.1016/j.cub.2015.06.068²), in addition to the previous phylogenetic associations with Annelida and Mollusca.
- Importantly, this is not the first Nemertean genome to be sequenced (https://www.nature.com/articles/s41559-017-0389-y³). In any case, it is not unusual that nemertean sequences do not appear in public databases.

References

1. Marlétaz F, Peijnenburg K, Goto T, Satoh N, et al.: A New Spiralian Phylogeny Places the

Enigmatic Arrow Worms among Gnathiferans. *Current Biology*. 2019; **29** (2): 312-318.e3 Publisher Full Text

2. Laumer CE, Bekkouche N, Kerbl A, Goetz F, et al.: Spiralian phylogeny informs the evolution of microscopic lineages.*Curr Biol*. 2015; **25** (15): 2000-6 PubMed Abstract | Publisher Full Text 3. Luo YJ, Kanda M, Koyanagi R, Hisata K, et al.: Nemertean and phoronid genomes reveal lophotrochozoan evolution and the origin of bilaterian heads.*Nat Ecol Evol*. 2018; **2** (1): 141-151 PubMed Abstract | Publisher Full Text

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others? Yes

Are the datasets clearly presented in a useable and accessible format? γ_{PS}

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: evolutionary developmental biology, evolutionary genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.