The Genome of the Poecilogonous Annelid Streblospio benedicti

Christina Zakas (1,*, Nathan D. Harry¹, Elizabeth H. Scholl², and Matthew V. Rockman³

¹Department of Biological Sciences, North Carolina State University, USA

²Bioinformatics Research Center, North Carolina State University, USA

³Department of Biology and Center for Genomics & Systems Biology, New York University, USA

*Corresponding author: E-mail: czakas@ncsu.edu.

Accepted: January 15, 2022

Abstract

Streblospio benedicti is a common marine annelid that has become an important model for developmental evolution. It is the only known example of *poecilogony* (where two distinct developmental modes occur within a single species) that is due to a heritable difference in egg size. The dimorphic developmental programs and life-histories exhibited in this species depend on differences within the genome, making it an optimal model for understanding the genomic basis of developmental divergence. Studies using *S. benedicti* have begun to uncover the genetic and genomic principles that underlie developmental uncoupling, but until now they have been limited by the lack of availability of genomic tools. Here, we present an annotated chromosomal-level genome assembly of *S. benedicti* generated from a combination of Illumina reads, Nanopore long reads, Chicago and Hi-C chromatin interaction sequencing, and a genetic map from experimental crosses. At 701.4 Mb, the *S. benedicti* genome is the largest annelid genomic analyses of development and evolution, as well as phylogenetic comparison within the annelida and the Lophotrochozoa. Despite having two developmental modes, there is no evidence of genome duplication or substantial gene number expansions. Instead, lineage-specific repeats account for much of the expansion of this genome compared with other annelids.

Key words: poecilogony, life-history evolution, developmental genomics, Lophotrochozoa.

Significance

The Lophotrochozoa is one of the largest clades of bilaterian diversity, but is an understudied group for genomic analyses. The *Streblospio benedicti* genome opens possibilities for understanding lineage-specific gene evolution in these animals. We present the chromosome-level genome assembly, with comparisons to other lophotrochozoan genomes.

Introduction

Genomic sequencing of lophotrochozoan animals has yielded important discoveries in genome evolution, adaptation, and novelty (Simakov et al. 2013; Albertin et al. 2015; Schiemann et al. 2017; Wang et al. 2017; Luo et al. 2018). Conserved genes and pathways across the bilateria have been investigated in other animal branches, however, many novel genes and gene functions within the Lophotrochozoa have yet to be explored (Tessmar-Raible and Arendt 2003; Paps et al. 2015). The Lophotrochozoa are a superb group for studying development, adaptation, and evolution on the genomic level due to the conserved patterns of ontogeny and a unique range of novel adaptations (Henry and Martindale 1999; Seaver 2014). However, they are severely lacking genomic resources: currently nine full genome data sets are listed in NCBI for annelids. Despite the biodiversity contained in this group, genome research has lagged relative to other bilaterian lineages, and little is known about genomic evolution and how that informs

© The Author(s) 2022. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com important process like convergence, gene regulatory network modification, and developmental systems drift. This gap in genomic resources impairs our understanding of basic developmental and evolutionary biology in a major animal lineage.

The marine annelid Streblospio benedicti (supplementary fig. S1. Supplementary Material online) is a particularly important model for understanding the genomic basis for developmental evolution. Streblospio benedicti is one of the rare cases of confirmed poecilogony-two distinct modes of development occurring within the species. Streblospio benedicti exhibits both indirect development, with a distinct larval phase, and abbreviated indirect development, with offspring that resemble small adult forms (Levin 1984; Levin and Bridges 1994). In S. benedicti, adults are gonochoristic (separate males and females) and females produce a fixed offspring type throughout their lives (Levin et al. 1987). There are two types of females in S. benedicti, which are essentially indistinguishable other than traits related to producing offspring (Gibson et al. 2010). The two types of females differ in the egg sizes they produce (\sim 100- vs. \sim 200- μ m-diameter eggs) and per-clutch fecundity (~200-400 vs. ~20-50 offspring per clutch; McCain [2008]). The resulting offspring are drastically different, with contrasting development modes (planktotrophy: indirect developing, obligately feeding larvae, vs. lecithotrophy: abbreviated indirect developing, nonobligately feeding larvae). These larval types differ in their planktonic development time (2-4 weeks swimming in the plankton vs. 0-2 days before settlement), larval ecologies (pelagic vs. benthic larvae), and overall life-history strategies. Importantly, these tradeoffs only occur during the embryonic and larval phases; by the time the worms become adults they are morphologically indistinct aside from some female reproductive anatomy (number of brood pouches and segments on which the brood pouches occur) and they occupy the same types of estuarine environments (Gibson et al. 2010). Interestingly, there is gene flow between adults of different types, but they usually do not directly co-occur (Zakas and Wares 2012). Furthermore, S. benedicti is the only known case of poecilogony where the developmental types are heritable with a strong additive genetic basis, as opposed to plasticity or polyphenism (Levin et al. 1991; Zakas and Rockman 2014; Zakas et al. 2018). Because these differences in development and life-history are contained within a single species, the ability to investigate the genomic basis of developmental variation is unparalleled. The genome of S. benedicti provides an opportunity to explore how a major transition in animal development happens on the genomic level.

High-quality genome assemblies are now enabling technical advances in using new models with naturally occurring traits of interest such as poecilogony. Here we present the genome of *S. benedicti*, which we constructed from a combination of Illumina short reads and Nanopore long reads, scaffolded with Hi-C and Chicago proximity-ligation data. We use the previously described genetic linkage map of *S. benedicti* to correct scaffolding arrangements and locate quantitative trait loci (QTL) markers in the genome (Zakas et al. 2018). This is the one of the few high-quality annelid genomes and opens opportunities for transformative research in this and related systems.

Results and Discussion

Genome Sequencing and Assembly

The genome assembly of *S. benedicti* is summarized in table 1. We recovered between 6 and 8.5 million reads per Nanopore flow cell, and 3–5 million reads were over 6.5 kb long. In total, we generated 30 million reads (with 15.6 million >6.5 kb). Illumina shotgun data alone yielded a draft assembly with scaffold N50 of 53 kb. The addition of the Nanopore, Chicago, and Hi-C data improved the assembly by an order of magnitude: increasing the scaffold N50 to 53.55 Mb (supplementary table S2 and fig. S2, Supplementary Material online).

To integrate the genetic map with the reference genome, we corrected the potential misassembly of scaffolds using Chromonomer (Catchen et al. 2020). Chromonomer breaks superscaffolds in regions of low confidence (stretches of "N" ambiguity) and rearranges them based on high-confidence markers in the genetic map. The *S. benedicti* genetic map was previously constructed from G₂ families with 702 markers in 11 linkage groups (Zakas et al. 2018). Chromonomer used 570 (81% of the total) informative markers to reconcile the reference to the genetic map. The resulting assembly increased the genome's N50 by 2.8 Mb, reduced the number of scaffolds in the assembly, and increased its Benchmarking Universal Single-Copy Orthologs (BUSCO) score by combining

Table 1

Genome Summary Statistics

| Number of scaffolds | 6 112 |
|-------------------------------------|------------------------------|
| 150/NI50 | 6/56 1 Mb |
| | |
| Total gene models predicted | 20,221 |
| BUSCO score [®] | 88.4(S) 3.4(D) 2.4(F) 5.8(M) |
| BUSCO total | 94.2 |
| GC% | 37.8 |
| %N | 0.3 |
| Average gene length | 3.54 kb |
| Median exon length | 157 bases |
| Median intron length | 410 bases |
| Transcribed regions | 17% |
| Mean number of exons per transcript | 3.1 |
| Protein-coding genes | 20,221 |
| mRNAs | 41,088 |
| Exons | 1,25,713 |
| CDSs | 1,24,795 |
| 5′ UTRs | 8,260 |
| tRNAs | 5,995 |
| 3' prime UTRs | 6,833 |

^aBUSCO: S, single; D, duplicated; F, fragmented; M, missing.

many smaller scaffolds with chromosomal scaffolds (supplementary table S2, Supplementary Material online). There are 11 chromosome-level scaffolds, from 38 to 65 Mb in length (supplementary table S1, Supplementary Material online). They correspond to the karyotype, which shows ten autosomes and one sex chromosome, and the 11 linkage groups constructed previously (Zakas et al. 2018).

Gene Annotations

An Iso-Seq transcriptome of pooled planktotrophic individuals of all developmental stages was used for annotations. Based on GMAP (version 2020-06-30; Wu and Watanabe [2005]) 24,117 of 24,317 high-quality Iso-seq transcripts mapped uniquely to the genome.

Using the Iso-Seq data as well as proteins from *Capitella teleta* as evidence, Maker v2.31.10 called 41,088 transcripts across the entire genome, including at least one gene on 1,899 of the 6,101 unplaced scaffolds. These transcripts represent 20,221 genes, indicating an average of two transcripts per locus throughout the genome. Additionally, 5,995 tRNA were identified (although 3,821 are noncoding/pseudo tRNAs).

Repeat Modeling

The genome of *S. benedicti* is 40.36% repetitive, which is greater than other annelids reported (table 1; supplementary table S7, Supplementary Material online), and it contains substantially more interspersed repeats. Most repeats are unclassified (comprising 30.19% of the genome) and are likely lineage-specific elements. This is not particularly unusual as emerging models tend to have novelity in repetitive elements and the families of repeats in the Lophotrochozoa remain understudied.

Comparison with Other Annelid Genomes

We used OrthoVenn2 to find gene cluster distributions across four annelid genomes (Helobdella robusta, Dimorphilus gyrociliatus, C. teleta, and S. benedicti). Of these four genomes, S. benedicti has the largest and least gene-dense genome. There are fewer genes per megabase and more interspersed repeats in S. benedicti (supplementary table S3, Supplementary Material online), based on our data set of 41,088 predicted transcripts. There are 15,259 clusters in this comparison and 8,511 of these contain S. benedicti proteins (supplementary fig. S3, Supplementary Material online). There are 1,948 (23% of all S. benedicti clusters) that are paralog clusters unique to S. benedicti, which is more than the other annelids. But unique proteins make up only 27% of the S. benedicti total proteins, which is comparable to the other species (supplementary table S3, Supplementary Material online). There are 11,281 transcripts found only in S. benedicti including those in the 1,948 clusters as well as 6,508 genes that are single-copy and single-isoform and are therefore not part of any clusters (supplementary fig. S4, Supplementary Material online). *Streblospio benedicti* has more predicted transcripts than the other annelids, but a similar number (20,211) of protein coding genes and a similar number of total gene clusters (supplementary fig. S3 and table S4, Supplementary Material online). *Streblospio benedicti* has more transcripts because there are, on average two transcripts per gene, although the distribution varies (supplementary fig. S5, Supplementary Material online). More robust comparisons about genomes and gene expansion can be addressed with the addition of new and updated annelid genomes available in the near future.

OrthoVen2 assigned GO terms to each of the gene clusters. Of the clusters unique to *H. robusta* (727) and *D. gyrociliatus* (759) there are no enriched GO terms relative to the other groups (supplementary table S3, Supplementary Material online). In *C. teleta's* unique gene clusters (1,794) there are eight GO terms that are enriched, whereas in *S. benedicti* there are 28 enriched GO terms listed in supplementary table S5, Supplementary Material online and figure 1. The proportion of unique clusters, singletons, and enriched GO categories suggest that there are more novel genes in the *S. benedicti* genome compared with the other annelids, although there is a range of novel genes reported across taxa within the Lophotrochozoa (Sun et al. 2020) and novel gene clusters are not necessarily correlated with functional novelty.

The addition of more contiguous annelid genomes in the future will reveal the extent of chromosomal rearrangement that has occurred in the annelids, but initial investigation revealed little evidence of macrosynteny across *S. benedicti* and the three other annelid genomes. The genome's Hox gene cluster was found on chromosome 7 by tBLASTn with queries from other annelids including *C. teleta*, *Platynereis dumerilii*, and *Myzostoma cirriferum*. This genome contains the full 11 *Antp* set of Hox genes on this chromosome.

The S. benedicti genome assembly and annotation provide a critical tool for understanding the genetic basis of phenotypic diversity, including genomic modifications that ultimately lead to evolutionary changes in ontogeny. The S. benedicti genome adds to the growing collection of assembled and annotated lophotrochozoan genomes. There are limited full annelid genome assemblies, but S. benedicti is one of the most complete and contiguous genomes for the annelida to date (supplementary table S3, Supplementary Material online). This assembly provides a methodology for assembling and annotating other lophotrochozoan genomes, which often have limited tissue availability, high heterozygosity, and a low-representation of lineage-specific genes in major ontogeny databases. The heterozygosity estimated from the Illumina reads is 0.29%, after nine generations of sibmating, lower than most lophotrochozoans (Kocot et al. 2020; Varney et al. 2021). Heterozygosity in outbred S. benedicti has been estimated at 0.5–1% (Rockman 2012) and the modest reduction after inbreeding may reflect the effects of inbreeding depression.

There are some notable assembly caveats: We used females for the Illumina reads because males have a long Y chromosome that is likely repetitive (Zakas et al. 2018), and we wanted to minimize assembly issues. The contents of the Y chromosome, and the sex chromosomes in general, warrants further investigation, especially as a QTL maps to the X chromosome and has contrasting directional parental contributions to offspring size (Zakas and Rockman 2021). This genome and transcriptome are generated from planktotrophic animals only, leaving the possibility that structural rearrangements or duplications may have happened between the two types. Previous work has indicated that major genomic rearrangements are unlikely to be a major source of genome divergence (Zakas et al. 2018). Future genomic sequencing and mapping of the lecithotrophic morph should reveal regions of genomic divergence between the types.

This high-quality, chromosomal-scale genome will aid in future population and functional genomics analysis in *S. benedicti*. Genomic comparisons with other lophotrochozoan animals, as they become increasingly available, will also provide insight into lineage-specific novelty with respect to *poecilogony* and the evolution of marine larval forms. We find that despite having a larger genome than other annelids, and a unique ability to produce different larval types, the *S. benedicti* genome does not contain significant genomic duplications. There are lineage-specific repetitive regions and longer introns and intergenic regions in *S. benedicti* compared with other current annelid genomes.

Materials and Methods

Genome Sequencing

We assembled the chromosomal-level genome of *S. benedicti* using individuals of the planktotrophic morph from Bayonne, New Jersey. The genome is assembled from five classes of data:

- 1. PCR-free Illumina shotgun sequence data generated from a pool of 13 females from a single F₉ inbred line.
- 2. Nanopore long-read data generated from multiple pools including 71 wild-caught females
- 3. Chicago proximity-ligation data generated from a pool of three males from the inbred line.
- 4. Hi-C data generated from a pool of two females from the inbred line.
- 5. Genetic map data from an experimental cross.

The initial assembly was generated using wtdbg2 (Ruan and Li 2020) an Pilon (Walker et al. 2014). This physical assembly was polished using two rounds of polishing with Racon (Vaser et al. 2017) with the Illumina shotgun data. The Chicago and the Hi-C data were then incorporated using the HiRise software pipeline (Putnam et al. 2016).

Genome Correction with Genetic Map (Chromonomer)

Chromonomer is a tool to combine a scaffolded, unfinished genome sequence with a genetic map by aligning the markers in the QTL map to the genome. It rearranges scaffolds in the unfinished assembly at contig breaks or low-confidence alignment regions such that they match the order of the markers in the map according to their recombination distance. Out of 702 markers in the genetic map, 598 total markers mapped to the final genome assembly with alignment scores over 10 (fig. 1). In our analysis, Chromonomer makes 394 rearrangements, all of which occur at contig breaks where separate contigs were scaffolded together denoted by "N" ambiguity symbols.

Genome Comparisons

We compared the *S. benedicti* genome with three other annelid reference genomes using OrthoVenn2, which is a graphbased method of identifying orthologous clusters (Xu et al. 2019). OrthoVenn2 identifies gene clusters that contain sets of orthologs or paralogs from these different species. We used the genomes of *C. teleta*, which is the most closely related species with a reference genome, *H. robusta* (leech), and *D. gyrociliatus*, which is the most compact and complete sequenced annelid genome (Simakov et al. 2013; Martín-Durán et al. 2020).

Gene Annotations

For the annotations we generated a PacBio Iso-Seq RNA transcriptome: RNA was extracted (Qiagen) and pooled from a mix of males and females from all stages including embryos. Total RNA was frozen at -80 °C and sequenced by the Duke Sequencing core, with no size selection on two SMRT cells; 264,600 reads were generated. Trimmed reads were clustered and polished with PacBio SMRT Link version 8.0 Isoseq3 tools. There are 24,317 Iso-Seq transcripts. High quality Iso-Seq reads were mapped to the chromosome sequences with GMAP (Wu and Watanabe 2005) version 2020-06-30. Alignment format was set to "gene" and failed alignments were suppressed from the gff3-formatted output.

Gene predictions were generated using maker (Cantarel et al. 2008) version 2.31.10. The 11 chromosomes as well as the 6,101 unplaced scaffolds were used for the predictions. A fasta file of high-quality Iso-Seq sequences (n = 24,317) were provided as same-species EST evidence. Protein homology evidence came from 31,978 *C. teleta* proteins downloaded from NCBI. Softmasking was selected for repeat masking.

The full set of 41,088 predicted proteins from transcripts were used in a BlastP search of NCBI's RefSeq protein database with a significance cutoff of 1.0e–05. The GI number for the top match for each of the proteins was extracted and *blastdbcmd* used to assign a descriptor associated with that GI number. We created a lookup table which associated a



Fig. 1.—Plot of genome assembly for 11 chromosomes. Red histogram shows the placement of gene transcripts identified in the *S. benedicti* GO enrichment categories. In total, 595 of 702 linkage markers are mapped on each chromosome. Three additional markers were mapped to unplaced scaffolds (supplementary table S6, Supplementary Material online).

protein name with the protein ID and gene name for each top match. These were added to the annotation file generated by maker. 37,822 (92%) proteins have a putative annotation.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We would like to thank Mohammed Khalfan for assistance with scripts to basecall on the NYU HPC cluster, Julian Catchen for feedback on Chromonomer, and the NCSU bioinformatics consulting core for help with hosting Jbrowser. We also thank Luke Noble for assistance with early shortread assemblies. We also thank John Yuen and Arielle Martel for help maintaining laboratory cultures. This work was supported by the Zegar Family Foundation, National Science Foundation (CAREER Grant No. IOS-1350926 to M.V.R.), National Institute of Health (Grant No. GM108396 to C.Z.), and startup funds from NYU and NCSU.

Data Availability

Results from the genome assembly and annotation are available: http://zakas.statgen.ncsu.edu/jbrowse/?data=sbenedicti (last accessed January 25, 2022).

The genome assembly is available at NCBI (BioProject: PRJNA721660).

Literature Cited

- Albertin CB, et al. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature 524(7564):220–224.
- Cantarel BL, et al. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18(1):188–196.
- Catchen J, Amores A, Bassham S. 2020. Chromonomer: a tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny. G3 (Bethesda). 10(11):4115–4128.
- Gibson G, MacDonald K, Dufton M. 2010. Morphogenesis and phenotypic divergence in two developmental morphs of *Streblospio benedicti* (Annelida, Spionidae). Invertebr Biol. 129(4):328–343.
- Henry JJ, Martindale MQ. 1999. Conservation and innovation in Spiralian development. Hydrobiologia 402:255–265.
- Kocot KM, Poustka AJ, Stöger I, Halanych KM, Schrödl M. 2020. New data from Monoplacophora and a carefully-curated dataset resolve molluscan relationships. Sci Rep. 10:1–8.
- Levin LA. 1984. Multiple patterns of development in *Streblospio benedicti* webster (spionidae) from three coasts of North America. Biol Bull. 166(3):494–508.
- Levin LA, Bridges TS. 1994. Control and consequences of alternative developmental modes in a poecilogonous polychaete. Am Zool. 34(3):323–332.
- Levin LA, Caswell H, DePatra KD, Creed EL. 1987. Demographic consequences of larval development mode: planktotrophy vs. lecithotrophy in *Streblospio benedicti*. Ecology. 68(6):1877–1886.

- Levin LA, Zhu J, Creed E. 1991. The genetic basis of life-history characters in a polychaete exhibiting planktotrophy and lecithotrophy. Evolution 45(2):380–397.
- Luo Y-J, et al. 2018. Nemertean and phoronid genomes reveal lophotrochozoan evolution and the origin of bilaterian heads. Nat Ecol Evol. 2(1):141–151.
- Martín-Durán JM, et al. 2020. Conservative route to genome compaction in a miniature annelid. Nat Ecol Evol. 5(2):231–242.
- McCain ER. 2008. Poecilogony as a tool for understanding speciation: early development of *Streblospio benedicti* and *Streblospio gynobranchiata* (Polychaeta: spionidae). Invertebr Reprod Dev. 51(2):91–101.
- Paps J, Xu F, Zhang G, Holland PW. 2015. Reinforcing the egg-timer: recruitment of novel lophotrochozoa homeobox genes to early and late development in the pacific oyster. Genome Biol Evol. 7(3):677–688.
- Putnam NH, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26(3):342–350.
- Rockman MV. 2012. Patterns of nuclear genetic variation in the poecilogonous polychaete Streblospio benedicti. Integr Comp Biol. 52(1):173–180.
- Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. Nat Methods 17(2):155–158.
- Schiemann SM, et al. 2017. Clustered brachiopod Hox genes are not expressed collinearly and are associated with lophotrochozoan novelties. Proc Natl Acad Sci U S A. 114(10):E1913–E1922.
- Seaver EC. 2014. Variation in spiralian development: insights from polychaetes. Int J Dev Biol. 58(6–8):457–467.
- Simakov O, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. Nature 493(7433):526–531.
- Sun J, Chen C, et al. 2020. The Scaly-foot Snail genome and implications for the origins of biomineralised armour. Nat Commun. 11:1–12.
- Tessmar-Raible K, Arendt D. 2003. Emerging systems: between vertebrates and arthropods, the Lophotrochozoa. Curr Opin Genet Dev. 13(4):331–340.
- Varney RM, Speiser DI, McDougall C, Degnan BM, Kocot KM. 2021. The iron-responsive genome of the chiton *Acanthopleura granulata*. Genome Biol Evol. 13:evaa263.
- Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27(5):737–746.
- Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9(11):e112963.
- Wang S, et al. 2017. Scallop genome provides insights into evolution of bilaterian karyotype and development. Nat Ecol Evol. 1:1–12.
- Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21(9):1859–1875.
- Xu L, et al. 2019. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 47(W1):W52–W58.
- Zakas C, Deutscher JM, Kay AD, Rockman MV. 2018. Decoupled maternal and zygotic genetic effects shape the evolution of development. eLife 7:e37143.
- Zakas C, Rockman MV. 2014. Dimorphic development in *Streblospio ben-edicti*: genetic analysis of morphological differences between larval types. Int J Dev Biol. 58(6–8):593–599.
- Zakas C, Rockman MV. 2021. Baby makes three: maternal, paternal, and zygotic genetic effects shape larval phenotypic evolution. Evolution 75(7):1607–1618.
- Zakas C, Wares JP. 2012. Consequences of a poecilogonous life history for genetic structure in coastal populations of the polychaete *Streblospio benedicti*. Mol Ecol. 21(22):5447–5460.

Associate editor: Dennis Lavrov