

<https://helda.helsinki.fi>

De Novo Sequencing, Assembly, and Annotation of Four Threespine Stickleback Genomes Based on Microfluidic Partitioned DNA Libraries

Berner, Daniel

Multidisciplinary Digital Publishing Institute

2019-06-03

Berner, D.; Roesti, M.; Bilobram, S.; Chan, S.K.; Kirk, H.; Pandoh, P.; Taylor, G.A.; Zhao, Y.;
Jones, S.J.M.; DeFaveri, J. De Novo Sequencing, Assembly, and Annotation of Four
Threespine Stickleback Genomes Based on Microfluidic Partitioned DNA Libraries. *Genes*
2019, 10, 426.

<http://hdl.handle.net/10138/348525>

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Communication

De Novo Sequencing, Assembly, and Annotation of Four Threespine Stickleback Genomes Based on Microfluidic Partitioned DNA Libraries

Daniel Berner ^{1,*} , Marius Roesti ^{2,3}, Steven Bilobram ⁴, Simon K. Chan ⁴, Heather Kirk ⁴, Pawan Pandoh ⁴, Gregory A. Taylor ⁴, Yongjun Zhao ⁴, Steven J. M. Jones ^{4,5,6}  and Jacquelin DeFaveri ⁷

¹ Department of Environmental Sciences, Zoology, University of Basel, Vesalgasse 1, CH-4051 Basel, Switzerland

² Biodiversity Research Centre and Zoology Department, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; marius.roesti@iee.unibe.ch

³ Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland; sichan@bcgsc.ca

⁴ Canada's Michael Smith Genome Sciences Centre at BC Cancer, Vancouver, BC V5Z 4S6, Canada; sbilobram@bcgsc.ca (S.B.); sichan@bcgsc.ca (S.K.C.); hkirk@bcgsc.ca (H.K.); ppandoh@bcgsc.ca (P.P.); gtaylor@bcgsc.ca (G.A.T.); yzhao@bcgsc.ca (Y.Z.); sjones@bcgsc.ca (S.J.M.J.)

⁵ Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

⁶ Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

⁷ Faculty of Biological & Environmental Sciences, University of Helsinki, Viikinkaari 1, FI-00014 Helsinki, Finland; jacquelin.defaveri@helsinki.fi

* Correspondence: daniel.berner@unibas.ch; Tel.: +41-(0)-61-207-03-28; Fax: +41-(0)-61-207-03-01

Received: 13 March 2019; Accepted: 27 May 2019; Published: 3 June 2019



Abstract: The threespine stickleback is a geographically widespread and ecologically highly diverse fish that has emerged as a powerful model system for evolutionary genomics and developmental biology. Investigations in this species currently rely on a single high-quality reference genome, but would benefit from the availability of additional, independently sequenced and assembled genomes. We present here the assembly of four new stickleback genomes, based on the sequencing of microfluidic partitioned DNA libraries. The base pair lengths of the four genomes reach 92–101% of the standard reference genome length. Together with their *de novo* gene annotation, these assemblies offer a resource enhancing genomic investigations in stickleback. The genomes and their annotations are available from the Dryad Digital Repository (<https://doi.org/10.5061/dryad.113j3h7>).

Keywords: *Gasterosteus aculeatus*; genome assembly

1. Introduction

The threespine stickleback (*Gasterosteus aculeatus*) is a small teleost fish widely distributed in both marine and freshwater habitats across the northern hemisphere [1,2]. Because of its ability for colonizing and adapting to diverse types of habitats, the species represents an important system for investigating evolutionary diversification and the underlying genetics [3–5]. Such research has been facilitated by the release in 2006 of a high-quality reference genome assembled from Sanger-sequenced plasmids, fosmids, and BAC clones derived from a freshwater individual from Alaska [6]. This genome, hereafter called the “reference genome”, has a total ungapped size of 447 Mb, of which 95% has been anchored to the 21 chromosomes through three rounds of re-assembly of the original sequence scaffolds or contigs [7–9]. Despite the high quality of this resource, access to additional, independently sequenced

and assembled stickleback genomes appears desirable, given the wide geographic distribution of the species, and the extensive population structure and phenotypic and genetic diversity exhibited across its range. In this paper, we report the generation and annotation of four *de novo* stickleback genome assemblies.

2. Methods and Materials

2.1. Stickleback Samples, DNA Library Preparation, Sequencing

Populations were chosen to mirror the species' wide ecological diversity and geographic distribution. Specifically, we selected a marine (anadromous) and a freshwater-resident population from both the Atlantic (SYL, NID) and the Pacific region (BAM, BOT) (Table 1). From each site, a single individual was sampled with unbaited minnow traps on breeding grounds during the spring of 2016, except for the SYL individual sampled from a laboratory line in the same year. After euthanasia according to standard protocols, the focal individuals were immediately frozen at $-20\text{ }^{\circ}\text{C}$ or $-80\text{ }^{\circ}\text{C}$ to minimize DNA degradation. To facilitate the assembly of chromosome 19, which is the sex chromosome heterogametic (XY) in threespine stickleback males [10], we considered only female individuals.

To obtain high molecular weight DNA, we slowly thawed the specimens, immediately sampled 50 mg of liver (SYL, BAM, BOT) or muscle (NID) tissue, and performed extractions using the QIAGEN MagAttract high molecular weight DNA Kit (Qiagen, Germantown, MD, USA), following the manufacturer's protocol. DNA integrity was assessed using pulsed-field gel electrophoresis (PFGE; the resulting gel images are provided as Figure S1 in the Supplementary Materials). The DNA obtained (predominantly $<50\text{ kb}$) was then used without size selection to generate microfluidic partitioned libraries using the Chromium System (10x Genomics Inc., Pleasanton, CA, USA). Details on DNA quality assessment and the Chromium library preparation protocol are specified in [11]. For the individuals SYL and NID, only a single Chromium library was generated, while for BAM and BOT, two replicate libraries were produced. These six total DNA libraries were then barcoded individually, pooled to equal total molarity among individuals, and paired-end sequenced to 150 base pairs (bp) in three lanes of an Illumina HiSeq X instrument. This produced between 504 and 542 million raw sequence reads per individual in total. The raw sequence data are available at the NCBI Short Read Archive (SRA) under the BioProject accession number PRJNA525775 (the genome assemblies and their annotations are, in addition to Dryad, also available under the same NCBI BioProject number).

Table 1. Characterization of the four stickleback individuals and their genome assemblies.

Assembly	SYL	NID	BAM	BOT
Region	Atlantic	Atlantic	Pacific	Pacific
Habitat type	Marine	Freshwater	Marine	Freshwater
Locality [Reference]	List, Sylt, Germany	Aach stream, Switzerland [12]	Bamfield Inlet, Vancouver Island, Canada [13]	Boot Lake, Vancouver Island, Canada [14]
Geographic coordinates	55°01'49.04" N, 8°25'37" E	47°33'29.25" N, 9°16'42.38" E	48°49'12.69" N, 125°8'57.9" W	50°03'00.2" N, 125°32'27.4" W
Number of scaffolds	15,853	10,246	25,430	18,433
N50 (Mb)	0.396	3.636	0.446	0.307
Longest scaffold (Mb)	3.12	16.02	4.33	3.81
Total assembly length (Mb) (gapped length in parentheses)	417.5 (431.8)	452.5 (467.5)	411.7 (445.7)	414.9 (427.3)
Number of annotated genes with experimental evidence	18,513	19,928	17,789	18,413

2.2. Genome Assembly and Annotation

As a first step, we filtered reagent sequences (approximately 1%) from the raw paired-end reads. Then we performed genomic assemblies by using the proprietary Supernova assembler (10x Genomics, San Francisco, CA, USA) with default parameters on a 750 Gb server, running the jobs serially to avoid resource conflict. To maximize assembly quality, we explored different combinations of assembler versions (1.20 and 2.01) and sequence coverages (from approximately 30× to 145×). For NID, the highest quality, as judged by N50 values and total assembly length, was achieved with Supernova version 2.01, while for the other individuals, version 1.20 performed best. Moreover, for all four stickleback individuals, optimal sequence coverage was above the maximum of 56× recommended for Supernova assembly of human genomes (89×, 142×, 145×, and 77× for SYL, NID, BAM, and BOT, respectively). An effort to further increase contiguity by using ARCS [15] did not improve the quality of the assemblies; hence, our genomes represent the assemblies obtained by Supernova, with scaffolds smaller than 1 kb excluded. No attempt was made to arrange our assembled scaffolds to the level of physical chromosomes.

Each of the four assemblies was then annotated for protein coding genes by using the MAKER platform (version 2.31.9; Yandell Lab, Salt Lake City, UT, USA). MAKER produces a single set of annotated genes by combining the ab initio gene prediction from three programs (AUGUSTUS [16], Snap [17], GeneMark [18]), informed by experimental gene evidence. As gene evidence, we used the threespine stickleback cDNA underlying the annotation of the reference genome (27,628 transcripts; available on Ensembl), and 538,010 protein sequences included in Swiss-Prot. AUGUSTUS predictions were trained on zebrafish genes, SNAP was trained using the 2586 highly conserved vertebrate genes predicted by BUSCO (version 2.0.1) [19,20], while GeneMark was self-trained.

2.3. Comparative Sequence Alignment

To illustrate the value of the new assemblies as resources for sequence alignment, we used a sample of 5 million 150 bp paired-end reads produced by Illumina HiSeq2500 whole-genome sequencing of pooled DNA from 72 field-caught stickleback from the NID population (the individuals derive from an experimental study [21]; sequence data: Laurentino and Berner, manuscript in preparation). These reads were paired-end aligned with Novoalign (version 3.00; <http://www.novocraft.com/products/novoalign>) to the reference genome and to each of the four new assemblies, each time using identical alignment parameters (main settings: $-t540$, $-g40$, $-x12$). The observed proportion of unique alignment to each assembly was standardized by the alignment proportion observed when using the reference genome. The reason for this standardization was that we were not primarily interested in the absolute alignment success, but in how alignment success compared among the assemblies to which the same set of reads was matched.

Since the NID population is located in the Atlantic part of the species' range, we repeated the above alignment protocol by using an analogous sample of whole-genome sequence reads generated in a similar way from a pool of DNA from 62 stickleback individuals from a Pacific freshwater population (Misty Lake, Vancouver Island, Canada [22,23]; sequence data: Haenel and Berner, manuscript in preparation).

3. Results and Discussion

The four new stickleback genome assemblies varied in length from 412 to 453 Mb (Table 1), thus ranging from approximately 8% smaller to 1% larger than the reference genome (Figure 1A). Contiguity was modest and highly variable, with scaffold numbers varying 2.5-fold and N50 values varying 12-fold among the individuals (Table 1). Interestingly, the assemblies based on two replicate Chromium libraries (BAM, BOT) were not superior in completeness or contiguity to those derived from a single library. This leads us to speculate that molecular weight of the extracted DNA—which was highest in the NID individual (details not presented)—may be a more critical determinant of assembly quality

than the number of libraries when using the Chromium system. We also found no indication of a correlation among individuals between total read number and total assembly length.

The average number of genes annotated was around 18,000 across the assemblies (genes with evidence; Table 1). Annotated gene number was strongly correlated with total assembly length ($r = 0.97$), indicating that the sequences missing in the less complete assemblies were not unusual with respect to gene content.

Aligning whole-genome sequence reads from the NID population with identical alignment parameters to all genomes revealed highest success when matched to SYL and especially NID—the two Atlantic assemblies (Figure 1B, left). This result cannot be attributed to genome completeness alone because the SYL genome is relatively incomplete, but must reflect substantial overall sequence divergence between Atlantic and Pacific stickleback. Consistent with this view, when aligning whole-genome reads from a Pacific population (Misty Lake) to all genomes, the success of alignment to the Atlantic assemblies dropped below the success observed for the reference genome (Figure 1B, right). Our comparative alignment analysis thus highlights the potential of our new assemblies, particularly the Atlantic ones, to complement population genomic analyses based on the reference genome. We further anticipate that our new resources will facilitate primer design for applications like targeted sequencing and genome editing, and the identification of structural variation within the stickleback genome.

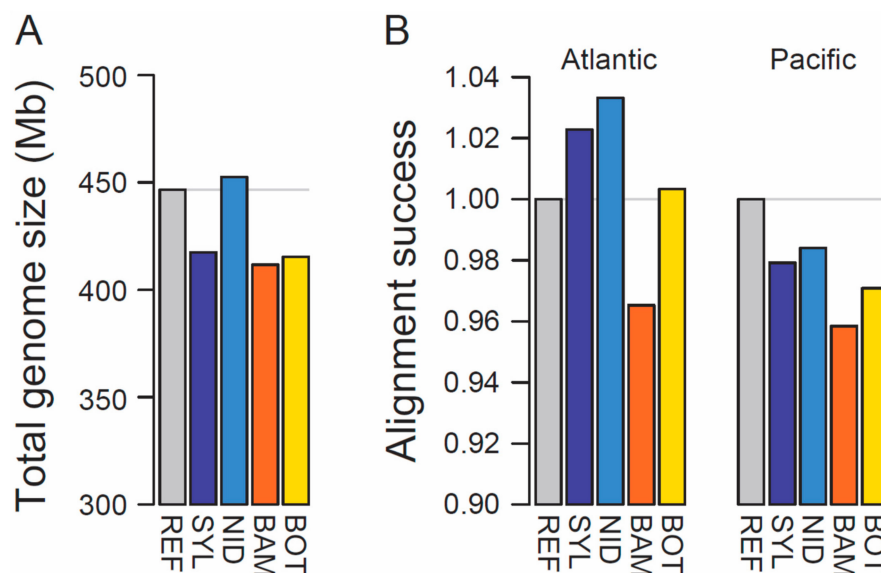


Figure 1. (A) Total ungapped length of the reference (REF) stickleback genome and the four new assemblies (SYL, Atlantic marine; NID, Atlantic freshwater; BAM, Pacific marine; BOT, Pacific freshwater). (B) Alignment success of whole-genome sequence reads from an Atlantic (left, NID) and a Pacific (right, Misty Lake) freshwater population when matched to each of the five assemblies, standardized by the alignment success achieved with the reference genome. Note that in both (A) and (B), the y -axis is strongly truncated to increase visual resolution in the upper range of the scale. The gray horizontal lines indicate the values for the reference genome.

Supplementary Materials: The following is available online at <http://www.mdpi.com/2073-4425/10/6/426/s1>, Figure S1: DNA electrophoresis gel image.

Author Contributions: Conceptualization, M.R., S.J.M.J. and J.D.; Methodology and laboratory work, M.R., H.K., P.P., G.A.T., Y.Z. and J.D.; Data analysis, D.B., S.B., S.K.C. and G.T.; Writing, D.B.; Project Administration, D.B., M.R., S.J.M.J. and J.D.; Funding Acquisition, D.B. and J.D.

Funding: This research was funded by the Swiss National Science Foundation SNF (grants 31003A_165826 to DB and P300PA_174344 to M.R.), and the Academy of Finland (grant 1286889 to J.D.).

Acknowledgments: Sean Rogers kindly contributed the BAM stickleback individual, Lauren Coombe and Rene Warren aided the production of the genomes, and two reviewers provided comments on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Bell, M.A.; Foster, S.A. *The Evolutionary Biology of the Threespine Stickleback*; Oxford University: Oxford, UK, 1994.
- Biology of the Three-Spined Stickleback*; Östlund-Nilsson, S., Mayer, I., Huntingford, F.A., Eds.; CRC: Boca Raton, FL, USA, 2007.
- McKinnon, J.S.; Rundle, H.D. Speciation in nature: The threespine stickleback model systems. *Trends Ecol. Evol.* **2002**, *17*, 480–488. [[CrossRef](#)]
- Peichel, C.L. Fishing for the secrets of vertebrate evolution in threespine sticklebacks. *Dev. Dyn.* **2005**, *234*, 815–823. [[CrossRef](#)] [[PubMed](#)]
- Cresko, W.A.; McGuigan, K.L.; Phillips, P.C.; Postlethwait, J.H. Studies of threespine stickleback developmental evolution: Progress and promise. *Genetica* **2007**, *129*, 105–126. [[CrossRef](#)] [[PubMed](#)]
- Jones, F.C.; Grabherr, M.G.; Chan, Y.F.; Russell, P.; Mauceli, E.; Johnson, J.; Swofford, R.; Pirun, M.; Zody, M.C.; White, S.; et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **2012**, *484*, 55–61. [[CrossRef](#)] [[PubMed](#)]
- Roesti, M.; Moser, D.; Berner, D. Recombination in the threespine stickleback genome - patterns and consequences. *Mol. Ecol.* **2013**, *22*, 3014–3027. [[CrossRef](#)] [[PubMed](#)]
- Glazer, A.M.; Killingbeck, E.E.; Mitros, T.; Rokhsar, D.S.; Miller, C.T. Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. *G3* **2015**, *5*, 1463–1472. [[CrossRef](#)]
- Peichel, C.L.; Sullivan, S.T.; Liachko, I.; White, M.A. Improvement of the threespine stickleback genome using a Hi-C-based proximity-guided assembly. *J. Hered.* **2017**, *108*, 693–700. [[CrossRef](#)]
- Peichel, C.L.; Ross, J.A.; Matson, C.K.; Dickson, M.; Grimwood, J.; Schmutz, J.; Myers, R.M.; Mori, S.; Schluter, D.; Kingsley, D.M. The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. *Curr. Biol.* **2004**, *14*, 1416–1424. [[CrossRef](#)]
- Jones, S.J.M.; Taylor, G.A.; Chan, S.; Warren, R.L.; Hammond, S.A.; Bilobram, S.; Mordecai, G.; Suttle, C.A.; Miller, K.M.; Schulze, A.; et al. The genome of the beluga whale (*Delphinapterus leucas*). *Genes* **2017**, *8*, 378. [[CrossRef](#)]
- Berner, D.; Roesti, M.; Hendry, A.P.; Salzburger, W. Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Mol. Ecol.* **2010**, *19*, 4963–4978. [[CrossRef](#)]
- Higham, T.E.; Jamniczky, H.A.; Jagnandan, K.; Smith, S.J.; Barry, T.N.; Rogers, S.M. Comparative dynamics of suction feeding in marine and freshwater three-spined stickleback, *Gasterosteus aculeatus*: Kinematics and geometric morphometrics. *Biol. J. Linn. Soc.* **2017**, *122*, 400–410. [[CrossRef](#)]
- Hendry, A.P.; Taylor, E.B. How much of the variation in adaptive divergence can be explained by gene flow? An evaluation using lake-stream stickleback pairs. *Evolution* **2004**, *58*, 2319–2331. [[CrossRef](#)] [[PubMed](#)]
- Yeo, S.; Coombe, L.; Warren, R.L.; Chu, J.; Birol, I. ARCS: Scaffolding genome drafts with linked reads. *Bioinformatics* **2017**, *34*, 725–731. [[CrossRef](#)] [[PubMed](#)]
- Stanke, M.; Tzvetkova, A.; Morgenstern, B. AUGUSTUS at EGASP: Using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.* **2006**, *7*, S11. [[CrossRef](#)] [[PubMed](#)]
- Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **2004**, *5*. [[CrossRef](#)] [[PubMed](#)]
- Lukashin, A.V.; Borodovsky, M. GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Res.* **1998**, *26*, 1107–1115. [[CrossRef](#)]
- Simao, 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](#)]
- Waterhouse, R.M.; Seppey, M.; Simao, F.A.; Manni, M.; Ioannidis, P.; Klioutchnikov, G.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **2018**, *35*, 543–548. [[CrossRef](#)]

21. Berner, D.; Ammann, M.; Spencer, E.; Rüegg, A.; Lüscher, D.; Moser, D. Sexual isolation promotes divergence between parapatric lake and stream stickleback. *J. Evol. Biol.* **2017**, *30*, 401–411. [[CrossRef](#)]
22. Lavin, P.A.; McPhail, J.D. Parapatric lake and stream sticklebacks on northern Vancouver Island: Disjunct distribution or parallel evolution? *Can. J. Zool.* **1993**, *71*, 11–17. [[CrossRef](#)]
23. Hendry, A.P.; Taylor, E.B.; McPhail, J.D. Adaptive divergence and the balance between selection and gene flow: Lake and stream stickleback in the Misty system. *Evolution* **2002**, *56*, 1199–1216. [[CrossRef](#)] [[PubMed](#)]



© 2019 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 (<http://creativecommons.org/licenses/by/4.0/>).