SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Hybrid genome assembly and annotation of *Danionella translucida*

Received: 23 April 2019
Accepted: 26 July 2019
Published online: 26 August 2019

Mykola Kadobianskyi, Lisanne Schulze, Markus Schuelke & Benjamin Judkewitz

Studying neuronal circuits at cellular resolution is very challenging in vertebrates due to the size and optical turbidity of their brains. *Danionella translucida*, a close relative of zebrafish, was recently introduced as a model organism for investigating neural network interactions in adult individuals. *Danionella* remains transparent throughout its life, has the smallest known vertebrate brain and possesses a rich repertoire of complex behaviours. Here we sequenced, assembled and annotated the *Danionella translucida* genome employing a hybrid Illumina/Nanopore read library as well as RNA-seq of embryonic, larval and adult mRNA. We achieved high assembly continuity using low-coverage long-read data and annotated a large fraction of the transcriptome. This dataset will pave the way for molecular research and targeted genetic manipulation of this novel model organism.

Background & Summary

The size and opacity of vertebrate tissues limit optical access to the brain and hinder investigations of intact neuronal networks *in vivo*. As a result, many scientists focus on small, superficial brain areas, such as parts of the cerebral cortex in rodents, or on early developmental stages of small transparent organisms, like zebrafish larvae. In order to overcome these limitations, *Danionella translucida* (DT), a transparent cyprinid fish[1,2] with the smallest known vertebrate brain, was recently developed as a novel model organism for the optical investigation of neuronal circuit activity in vertebrates[3,4]. The majority of DT tissues remain transparent throughout its life (Fig. 1). DT displays a variety of social behaviours, such as schooling and vocal communication, and is amenable to genetic manipulation using genetic tools that are already established in zebrafish. As such, this species is a promising model organism for studying the function of neuronal circuits across the entire brain. Yet, a continuous annotated genome reference is still needed to enable targeted genetic and transgenic studies and facilitate the adoption of DT as a model organism.

Next-generation short-read sequencing advances steadily decreased the price of the whole-genome sequencing and enabled a variety of genomic and metagenomic studies. However, short-read-only assemblies often struggle with repetitive and intergenic regions, resulting in fragmented assembly and poor access to regulatory and promoter sequences[5,6]. Long-read techniques, such as PacBio and Nanopore, can generate reads up to 2 Mb[7], but they are prone to errors, including frequent indels, which can lead to artefacts in long-read-only assemblies[6]. Combining short- and long-read sequencing technologies in hybrid assemblies recently produced high-quality genomes in fish[8,9].

Here we report the hybrid Illumina/Nanopore-based assembly of the *Danionella translucida* genome. A combination of deep-coverage Illumina sequencing with a single Nanopore sequencing run produced an assembly with scaffold N50 of 340 kb and Benchmarking Universal Single-Copy Orthologs (BUSCO) genome completeness score of 92%. Short- and long-read RNA sequencing data used together with other fish species annotated proteomes produced an annotation dataset with BUSCO transcriptome completeness score of 86%.

Methods

Genomic sequencing libraries. For genomic DNA sequencing we generated paired-end and mate-pair Illumina sequencing libraries and one Nanopore library. We extracted DNA from fresh DT tissues with phenol-chloroform-isoamyl alcohol. For Illumina sequencing, we used 5 days post fertilisation (dpf) old larvae. A shotgun paired-end library with 500 bp insert size was prepared with TruSeq DNA PCR-Free kit (Illumina).

Einstein Center for Neurosciences, NeuroCure Cluster of Excellence, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117, Berlin, Germany. Correspondence and requests for materials should be addressed to M.S. (email: markus.schuelke@charite.de) or B.J. (email: benjamin.judkewitz@charite.de)

Fig. 1 Male adult *Danionella translucida* showing transparency.

Illumina paired-end gDNA	
Number of reads	1.347×10^9
Total library size	136.047 Gb
Insert size	500 bp
Read length	2 × 101 bp
Estimated coverage	186×
Illumina mate-pair gDNA	
Number of reads	554.134×10^6
Total library size	55.968 Gb
Insert size	10 kb
Read length	2 × 101 bp
Estimated coverage	77×
Nanopore gDNA	
Number of reads	824.880×10^3
Total library size	4.288 Gb
Read length N50	11.653 kb
Estimated coverage	5.8×
Nanopore cDNA	
Number of reads	208.822×10^3
Total library size	279.584 Mb
Read length N50	1.812 kb
BGI 3 dpf larvae mRNA	
Number of reads	130.768×10^6
Total library size	13.077 Gb
Read length	2 × 100 bp
BGI adult mRNA	
Number of reads	128.546×10^6
Total library size	12.855 Gb
Read length	2 × 100 bp

Table 1. Sequencing library statistics. gDNA stands for genomic DNA sequencing, cDNA for reverse-transcribed complementary DNA, mRNA for poly-A tailed RNA sequencing.

Sequencing on HiSeq 4000 generated 1.347 billion paired-end reads. A long ~10 kb mate-pair library was prepared using the Nextera Mate Pair Sample Prep Kit and sequenced on HiSeq 4000, resulting in 554 million paired-end reads. Raw read library quality was assessed using FastQC v0.11.8[10].

A Nanopore sequencing high-molecular-weight gDNA library was prepared from 3 months post fertilisation (mpf) DT tails. We used 400 ng of DNA with the 1D Rapid Sequencing Kit (SQK-RAD004) according to manufacturer's instructions to produce the longest possible reads. This library was sequenced with the MinION sequencer on a single R9.4 flowcell using MinKNOW v1.11.5 software for sequencing and base-calling, producing a total of 4.3 Gb sequence over 825k reads. The read library N50 was 11.6 kb with the longest read being approximately 200 kb. Sequencing data statistics are summarised in Table 1.

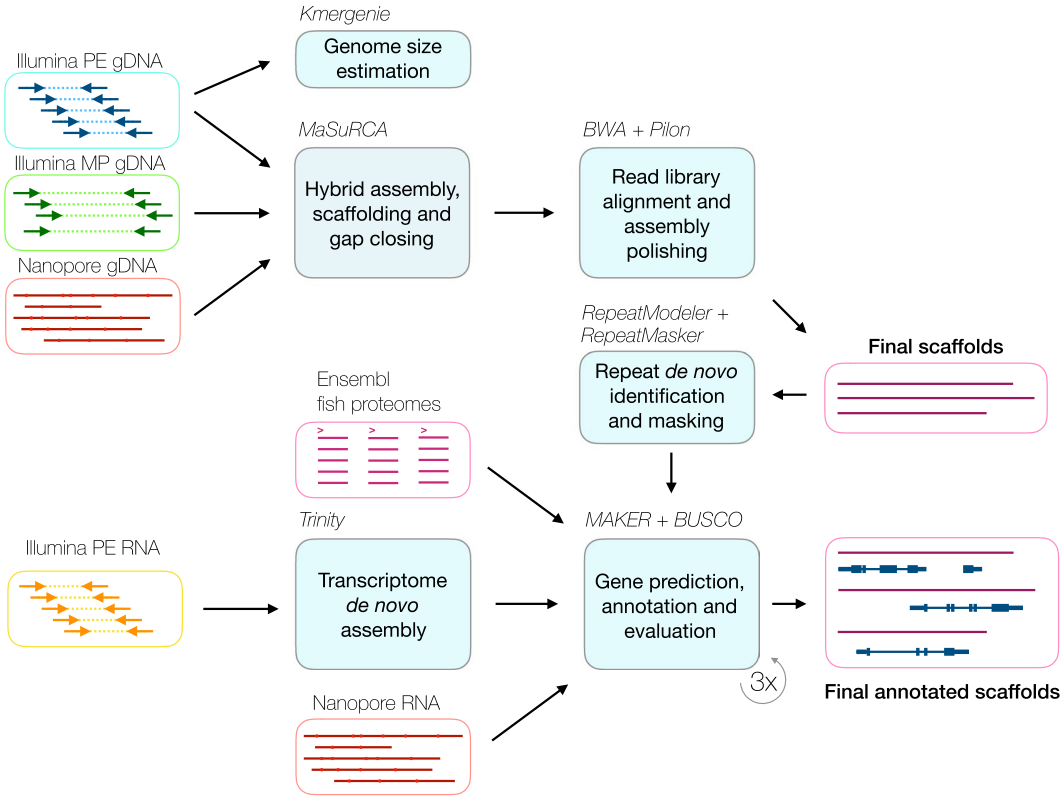

Fig. 2 DT genome assembly and annotation pipeline. **PE**, paired-end; **MP**, mate-pair.

Genome assembly. The genome assembly and annotation pipeline is shown in Fig. 2. We estimated the genome size using the k-mer histogram method with Kmergenie v1.7016 on the paired-end Illumina library preprocessed with fast-mcf v1.04.807[11,12], which produced a putative assembly size of approximately 744 Mb. This translates into 186-fold Illumina and 5.8-fold Nanopore sequencing depths.

Multiple published assembly pipelines utilise a combination of short- and long-read sequencing. Our assembler of choice was MaSuRCA v3.2.6[13], since it has already been used to generate high-quality assemblies of fish genomes[8,9], providing a large continuity boost even with low amount of input long reads[14]. Briefly, Illumina paired-end shotgun reads were non-ambiguously extended into the superreads, which were mapped to Nanopore reads for error correction, resulting in megareads. These megareads were then fed to the modified CABOG assembler that assembles them into contigs and, ultimately, mate-pair reads were used to do scaffolding and gap repair.

Following MaSuRCA author's recommendation[8], we have turned off the *frgcorr* module and provided raw paired-end and mate-pair read libraries for in-built preprocessing with the QuorUM error corrector[13,15]. The initial genome assembly size estimated with the Jellyfish assembler module was 938 Mb. After the MaSuRCA pipeline processing we have polished the assembly with one round of Pilon v1.22, which attempts to resolve assembly errors and fill scaffold gaps using preprocessed reads mapped to the assembly[16]. Leftover contaminants were filtered during the processing of the genome submission to the NCBI database. Statistics of the resulting assembly were generated using bbmap stats toolkit v37.32[17] and are presented in Table 2.

The resulting 735 Mb assembly had a scaffold N50 of 341 kb, the longest scaffold being more than 3 Mb. To assess the completeness of the assembly we used BUSCO v3[18] with the Actinopterygii ortholog dataset. In total, 91.5% of the orthologs were found in the assembly.

Transcriptome sequencing and annotation. We used three sources of transcriptome evidence for the DT genome annotation: (i) assembled poly-A-tailed short-read and raw Nanopore cDNA sequencing libraries, (ii) protein databases from sequenced and annotated fish species and (iii) trained gene prediction software. For Nanopore cDNA sequencing we extracted total nucleic acids from 1–2 dpf embryos using phenol-chloroform-isoamyl alcohol extraction followed by DNA digestion with DNAse I. The resulting total RNA was converted to double-stranded cDNA using poly-A selection at the reverse transcription step with the Maxima H Minus Double-Stranded cDNA Synthesis Kit (ThermoFisher). The double-stranded cDNA sequencing library was prepared and sequenced in the same way as the genomic DNA with MinKNOW v1.13.1, resulting in 190 Mb sequence data distributed over 209k reads. These reads were filtered to remove 10% of the shortest ones. For short-read RNA-sequencing, we have extracted total RNA with the TRIzol reagent (Invitrogen) from 3 dpf larvae and from adult fish. RNA was poly-A enriched and sequenced as 100 bp paired-end reads on the BGISEQ-500 platform. After preprocessing the library sizes were 65.4 million read pairs for 3 dpf larvae and 64.3 million read pairs for adult fish specimens (Table 1). We first assembled the 100 bp paired-end RNA-seq reads *de novo* using

Genome assembly statistics	
Total scaffolds	27,639
Total contigs	36,005
Total scaffold sequence	735.303 Mb
Total contig sequence	725.703 Mb
Gap sequences	1.306%
Scaffold N50	340.819 kb
Contig N50	133.131 kb
Longest scaffold	3.085 Mb
Longest contig	995.155 kb
Fraction of genome in >50 kb scaffolds	88.3%
BUSCO genome completeness score	
Complete	91.5%
Single	87.0%
Duplicated	4.5%
Fragmented	3.6%
Missing	4.9%
Total number of Actinopterygii orthologs	4,584

Table 2. DT genome assembly statistics and completeness.

Total protein-coding gene models	24,097
Total functionally annotated gene models	21,491
Gene models with AED <0.5	95%
Mean AED	0.18
BUSCO annotation completeness score	
Complete	86.3%
Single	80.6%
Duplicated	5.7%
Fragmented	7.1%
Missing	6.6%
Total number of Actinopterygii orthologs	4,584

Table 3. DT transcriptome annotation statistics.

Trinity v2.8.4 assembler[19]. This produced 222448 contigs with an N50 length of 3586 bp, clustered into 146103 "genes". BUSCO transcriptome analysis revealed 96% of complete Actinopterygii orthologs in the Trinity assembly. These contigs, together with the Nanopore cDNA reads and proteomes of 11 fish species from Ensembl[20] were used as the transcript evidence in MAKER v2.31.10 annotation pipeline[21]. Repetitive regions were masked using a *de novo* generated DT repeat library (RepeatModeler v1.0.11)[22]. The highest quality annotations with average annotation distance (AED) < 0.25 were used to train SNAP[23] and Augustus[24] gene predictors. Gene models were then polished over two additional rounds of re-training and re-annotation. The final set of annotations consisted of 24,097 protein-coding gene models with an average length of 13.4 kb and an average AED of 0.18 (Table 3). We added putative protein functions using MAKER from the UniProt database[25] and protein domains from the interproscan v5.30–69.0 database[26]. tRNAs were searched for and annotated using tRNAscan-SE v1.4[27]. The BUSCO transcriptome completeness search found 86% of complete *Actinopterygii* orthologs in the annotation set. An example Interactive Genomics Viewer (IGV) v2.4.3[28] window with the *dnmt1* gene is shown on Fig. 3, demonstrating the annotation and RNA-seq coverage.

Data Records

Raw sequencing libraries and genome and transcriptome assemblies are deposited to NCBI SRA as part of the BioProject SRP136594[29].

The genome assembly with gene and transcript annotations has been deposited at GenBank under the accession number SRMA00000000[30] (the version described in this paper is SRMA01000000), as well as on figshare in FASTA/GFF3 format[31]. The Trinity transcriptome assembly has been deposited at NCBI TSA under accession number GHNV00000000[32] (the version described in this paper is GHNV01000000), as well as on figshare[31].

Kmergenie-generated kmer abundance histograms and a summary report together with the genome size estimation are deposited at figshare[31].

Fig. 3 IGV screenshot of the *dnmt1* locus in the DT genome assembly, with short-read RNA coverage, mapped Nanopore cDNA-seq reads and alternative splicing annotation. Tracks from top to bottom: (I) adult RNA-seq coverage, (II) 3 dpf RNA-seq coverage, (III) Nanopore cDNA-seq coverage, (IV) Nanopore cDNA-seq read mapping and (V) annotation with alternative splicing isoforms.

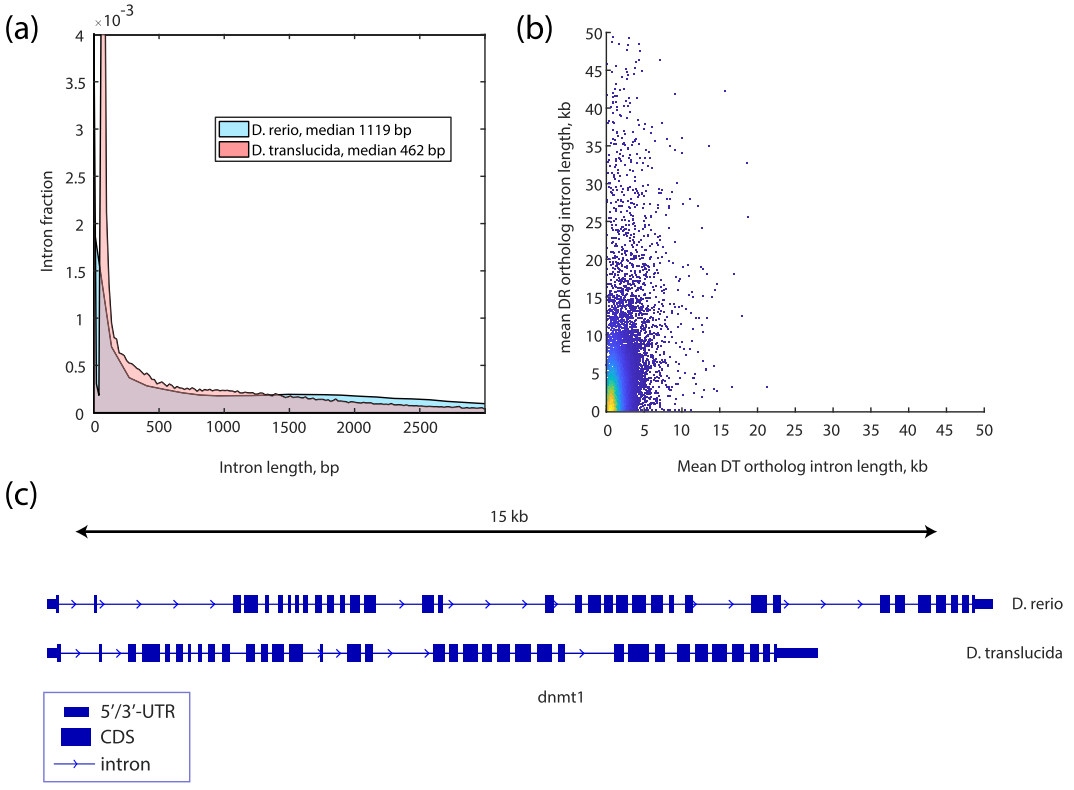

Fig. 4 Intron size distribution in DT (red) in comparison to zebrafish (DR, blue). (**a**) Intron size distribution of all transcripts in DR and DT. (**b**) Intron size relationship for identified DR-DT orthologous proteins. (**c**) A comparison of *dnmt1* orthologous loci in both fish. **5′/3′ UTR**, untranslated regions; **CDS**, coding sequence.

MAKER pipeline annotation output GFF3 file containing evidence mapping, identified repetitive elements and gene models, MAKER-predicted transcripts and proteins, IGV-compatible short-read and long-read RNA-seq coverage, raw sequencing read library FASTQC quality analysis report and intron orthology data together with their custom analysis code are available on figshare[31].

Technical Validation

DT and zebrafish intron size distributions. The predicted genome size of DT is around one half of the zebrafish reference genome[33]. *Danionella dracula*, a close relative of DT, possesses a unique developmentally truncated morphology[34] and has a genome of a similar size (ENA Accession Number GCA_900490495.1). In order to validate our genome assembly, we set out to compare the compact genome of DT to the zebrafish reference genome.

Changes in the intron lengths have been shown to be a significant part of genomic truncations and expansions, such as a severe intron shortening in another miniature fish species, *Paedocypris*[35], or an intron expansion in zebrafish[36]. We therefore compared the distribution of total intron sizes from the combined Ensembl/Havana zebrafish annotation[20] to the MAKER-produced DT annotation (Fig. 4a). We found that the DT intron size distribution is similar to other fish species investigated in ref.[35] which stands in stark contrast to the large tail of long introns in zebrafish. Median intron length values are in the range of the observed genome size difference (462 bp in DT as compared to 1,119 bp in zebrafish).

To investigate the difference in intron sizes on the transcript level, we compared average intron sizes for orthologous protein-coding transcripts in DT and zebrafish. We have identified orthologs in DT and zebrafish protein databases with the help of the conditional reciprocal best BLAST hit algorithm (CRB-BLAST)[37]. In total, we have identified 19,192 unique orthologous protein pairs. For 16,751 of those orthologs with complete protein-coding transcript exon annotation in both fish we calculated their respective average intron lengths (Fig. 4b). The distribution was again skewed towards long zebrafish introns in comparison to DT. As an example, Fig. 4c shows *dnmt1* locus for the zebrafish and DT orthologs.

Code Availability

Software used for read preprocessing, genome and transcriptome assembly and annotation is described in the Methods section together with the versions used. Custom MATLAB code used for orthology analysis is deposited on figshare[31].

References

1. Roberts, T. R. Danionella translucida, a new genus and species of cyprinid fish from Burma, one of the smallest living vertebrates. *Environ. Biol. Fishes* **16**, 231–241 (1986).
2. Britz, R., Conway, K. W. & Rüber, L. Spectacular morphological novelty in a miniature cyprinid fish, danionella dracula n. sp. *Proc. Biol. Sci.* **276**, 2179–2186 (2009).
3. Schulze, L. *et al.* Transparent danionella translucida as a genetically tractable vertebrate brain model. *Nat. Methods* **15**, 977–983 (2018).
4. Penalva, A. *et al.* Establishment of the miniature fish species Danionella translucida as a genetically and optically tractable neuroscience model. Preprint at https://doi.org/10.1101/444026v1.full (2018).
5. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
6. Watson, M. Mind the gaps - ignoring errors in long read assemblies critically affects protein prediction. Preprint at https://doi.org/10.1101/285049v1 (2018).
7. Payne, A., Holmes, N., Rakyan, V. & Loose, M. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast 5 files. Preprint at https://doi.org/10.1101/312256v1.full (2018).
8. Tan, M. H. *et al.* Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the Clownfish (Amphiprion ocellaris) genome assembly. *GigaScience* **7**, 1–6 (2018).
9. Tørrensen, O. K. *et al.* An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics* **18**, 1–23 (2017).
10. Andrews, S. *FastQC: a quality control tool for high throughput sequence data*, http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010).
11. Aronesty, E. Comparison of Sequencing Utility Programs. *Open Bioinforma J* **7**, 1–8 (2013).
12. Chikhi, R. & Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**, 31–37 (2014).
13. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
14. Tan, M. H. *et al.* A hybrid de novo assembly of the sea pansy (Renilla muelleri) genome. *GigaScience* **8**, 1–7 (2019).
15. Marçais, G., Yorke, J. A. & Zimin, A. QuorUM: An Error Corrector for Illumina Reads. *PLoS One* **10**, 1–13 (2015).
16. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, 1–14 (2014).
17. Bushnell, B. *BBmap short-read aligner, and other bioinformatics tools*, http://sourceforge.net/projects/bbmap/ (2016).
18. 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* **31**, 3210–3212 (2015).
19. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
20. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
21. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
22. Smit, A. F. A. & Hubley, R. *Repeat Modeler Open-1.0*, http://www.repeatmasker.org (2008).
23. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 1–9 (2004).
24. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
25. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
26. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
27. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
28. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics* **14**, 178–192 (2013).
29. *NCBI Sequence Read Archive*, http://identifiers.org/ncbi/insdc.sra:SRP136594 (2019).
30. *GenBank*, http://identifiers.org/ncbi/insdc:SRMA00000000 (2019).
31. Kadobianskyi, M., Schulze, L., Schuelke, M. & Judkewitz, B. Hybrid genome assembly and annotation of Danionella translucida. *figshare*. https://doi.org/10.6084/m9.figshare.c.4437488 (2019).
32. *GenBank* http://identifiers.org/ncbi/insdc:GHNV00000000 (2019).

33. Howe, K. *et al*. The zebrafish reference genome sequence and its relationship to the human genome. *Nat. Commun.* **496**, 498–503 (2013).
34. Britz, R. & Conway, K. W. Danionella dracula, an escape from the cypriniform Bauplan via developmental truncation? *J. Morphol.* **277**, 147–166 (2016).
35. Malmstrøm, M. *et al*. The most developmentally truncated fishes show extensive hox gene loss and miniaturized genomes. *Genome Biol. Evol.* **10**, 1088–1103 (2018).
36. Moss, S. P., Joyce, D. A., Humphries, S., Tindall, K. J. & Lunt, D. H. Comparative analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage. *Genome Biol. Evol.* **3**, 1187–1196 (2011).
37. Aubry, S., Kelly, S., Kümpers, B. M. C., Smith-Unna, R. D. & Hibberd, J. M. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis. *PLoS Genet.* **10**, 1–16 (2014).

Acknowledgements

We would like to thank Jörg Henninger for helpful discussions and critical reading of this manuscript. This work was funded by the NeuroCure Cluster of Excellence (DFG, project EXC-2049-390688087) to M.S. and B.J. B.J. is a recipient of a Starting Grant by the European Research Council (ERC-2016-StG-714560) and the Alfried Krupp Prize for Young University Teachers, awarded by the Alfried Krupp von Bohlen und Halbach-Stiftung.

Author Contributions

M.K. collected samples, conducted sequencing, genome assembly and annotation and conducted data analysis. L.S. collected samples and participated in data analysis. M.S. and B.J. conceived and supervised the study.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.