Hybrid genome assembly and annotation of *Danionella translucida*, a transparent fish with the smallest known vertebrate brain

Mykola Kadobianskyi[1], Lisanne Schulze[1], Markus Schuelke[1,✉], and Benjamin Judkewitz[1,✉]

[1]Einstein Center for Neurosciences, NeuroCure Cluster of Excellence, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany.

Studying the activity of distributed neuronal circuits at a cellular resolution in vertebrates is very challenging due to the size and optical turbidity of their brains. We recently presented *Danionella translucida*, a close relative of zebrafish, as a model organism suited for studying large-scale 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 the smallest known vertebrate brain.

Hybrid genome assembly | Danionella translucida

Correspondence: ***benjamin.judkewitz@charite.de, markus.schuelke@charite.de***

Introduction

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, we recently developed a novel model organism for the optical investigation of neuronal circuit activity in vertebrates – *Danionella translucida* (DT), a transparent cyprinid fish (1, 2) with the smallest known vertebrate brain (3, 4). The majority of DT tissues remain transparent throughout its life (Fig. 1). DT displays a rich repertoire 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

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

(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%.

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 kit (Illumina). 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. Read library quality was controlled using FastQC v0.11.8 (10).

Illumina paired-end gDNA	
Number of reads	1.347×10^9
Total library size	136.047 Gb
Read length	2×101 bp
Estimated coverage	185×
Illumina mate-pair gDNA	
Number of reads	554.134×10^6
Total library size	55.968 Gb
Read length	2×101 bp
Estimated coverage	76×
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 mRNA	
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, mRNA for poly-A tailed RNA sequencing.

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 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 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 ~200 kb. Sequencing data statistics are summarised in Table 1.

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 750 Mb. This translates into 185-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, providing a large continuity boost even with low amount of input Nanopore reads (8, 9). 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 scaf-

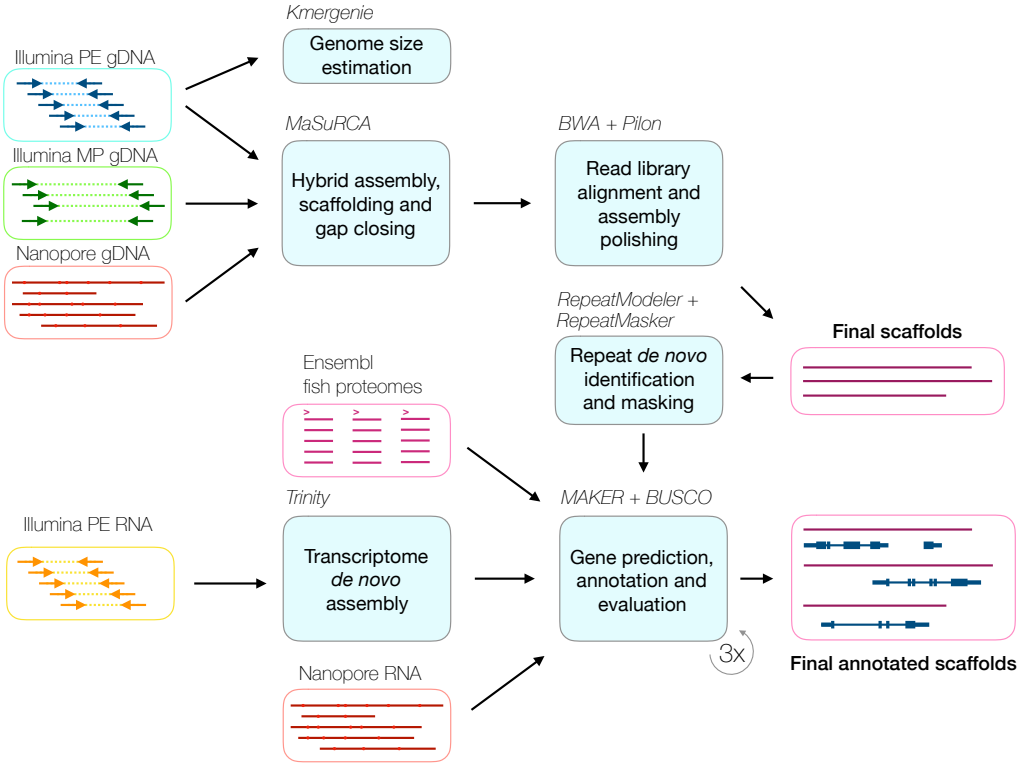

Fig. 2. DT genome assembly and annotation pipeline. **PE**, paired-end; **MP**, mate-pair.

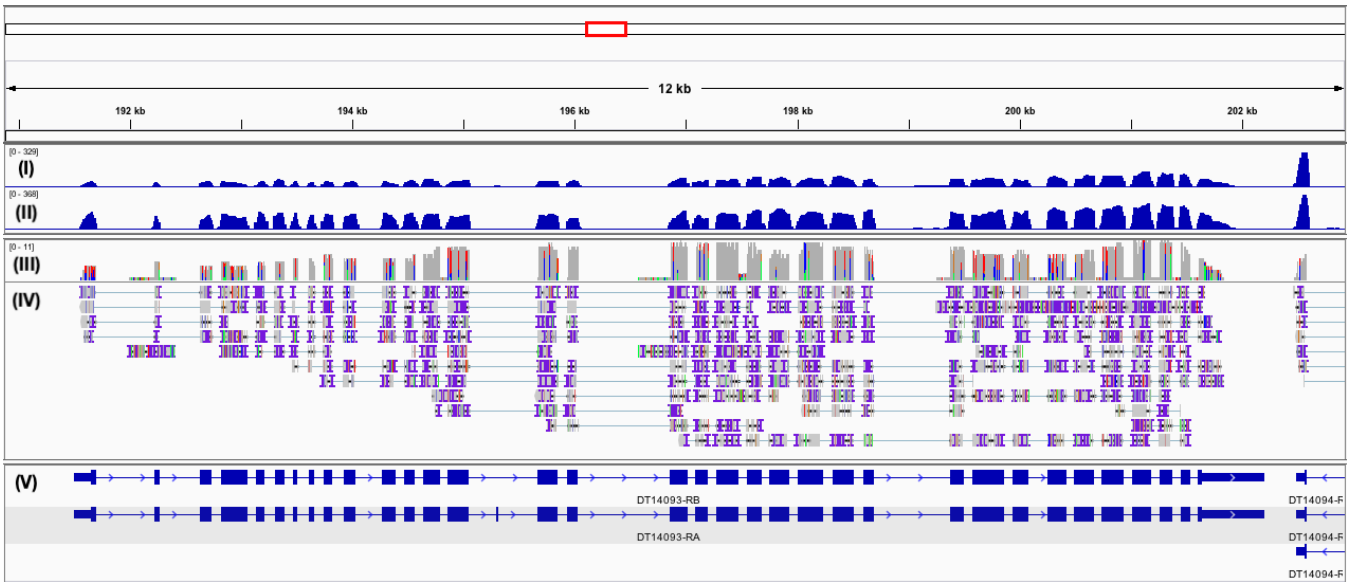

Fig. 3. IGV screenshot of the *dnmt1* locus in the DT genome assembly, with short-read RNA coverage, mapped Nanopore RNA-seq reads and alternative splicing annotation. Tracks from top to bottom: (I) adult RNA-seq coverage, (II) 3 dpf RNA-seq coverage, (III) Nanopore RNA-seq coverage, (IV) Nanopore RNA-seq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

folding and gap repair.

Following MaSuRCA author's recommendation (8), we have turned off the *frgcorr* module and provided raw read libraries for in-built read preprocessing. 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 (14). Statistics of the resulting assembly were generated using bbmap stats toolkit v37.32 (15) 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 (16) with the Actinopterygii ortholog dataset. In total, 91.5% of the orthologs were found in the assembly.

Genome assembly statistics	
Total scaffolds	27,814
Total contigs	36,191
Total scaffold sequence	735.373 Mb
Total contig sequence	725.755 Mb
Gap sequences	1.308%
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 orthologs (Actinopterygii)	4,584

Table 2. DT genome assembly statistics and completeness.

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 RNA sequencing libraries, (ii) protein databases from sequenced and annotated fish species and (iii) trained gene prediction software. For Nanopore RNA sequencing we extracted total nucleic acids from 1-2 dpf embryos using phenol-chloroform-isoamyl alcohol extraction followed by DNA digestion with DNAse I. 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, 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 this resulted in 65.4 million read pairs for 3 dpf larvae and in 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 Trinity v2.8.4 assembler (17). Resulting RNA contigs, together with the Nanopore cDNA reads and proteomes of 11 fish species from Ensembl (18) were used as the transcript evidence in MAKER v2.31.10 annotation pipeline (19). Repetitive regions were masked using a *de novo* generated DT repeat library (RepeatModeler v1.0.11 (20)). The highest quality annotations with average annotation distance (AED) < 0.25 were used to train SNAP (21) and Augustus (22) gene predictors. Gene models were then polished over two addi-

tional rounds of re-training and re-annotation. The final set of annotations consisted of 24,099 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 (23) and protein domains from the interproscan v5.30-69.0 database (24). tRNAs were searched for and annotated using tRNAscan-SE v1.4 (25). The BUSCO transcriptome completeness search found 86% of complete *Actinopterygii* orthologs. An example Interactive Genomics Viewer (IGV) v2.4.3 (26) window with the *dnmt1* gene is shown on Fig. 3, demonstrating the annotation and RNA-seq coverage.

Total gene models	24,099
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 orthologs (Actinopterygii)	4,584

Table 3. DT transcriptome annotation statistics.

DT and zebrafish intron size distributions

The predicted genome size of DT is around one half of the zebrafish reference genome (27). *Danionella dracula*, a close relative of DT, possesses a unique developmentally truncated morphology (28) 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* (29), or an intron expansion in zebrafish (30). We therefore compared the distribution of total intron sizes from the combined Ensembl/Havana zebrafish annotation (18) 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. (29) 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

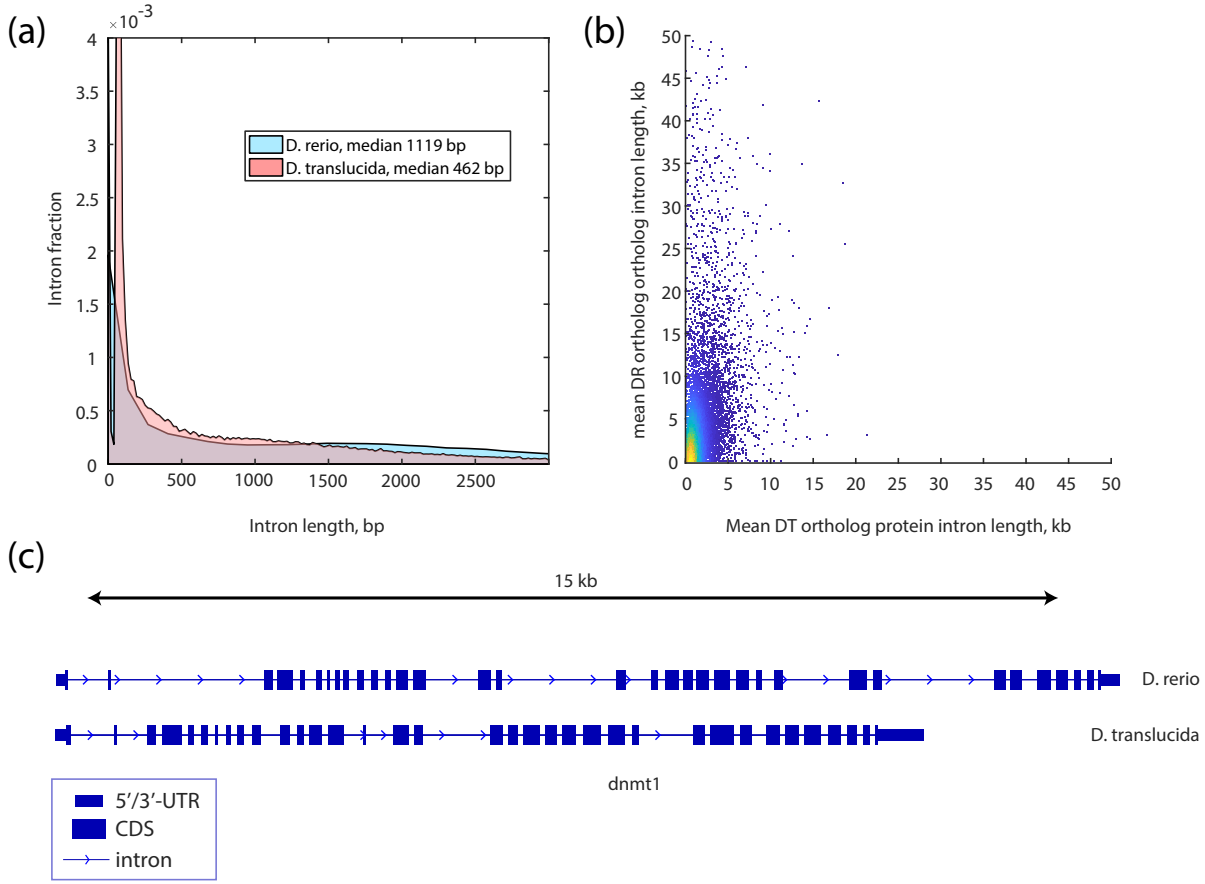

Fig. 4. Intron size distribution in DT in comparison to zebrafish (DR). (a) Intron size distribution of all transcripts in DR and DT. (b) Intron size relationship for identified DR-DT orthologous proteins. (c) *dnmt1* ortholog locus in both fish.

with the help of the conditional reciprocal best BLAST hit algorithm (CRB-BLAST) (31). 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.

Conclusions

In this work we describe whole-genome sequencing, assembly and annotation of the *Danionella translucida* genome. Using deep-coverage short reads and low-coverage long Nanopore reads, we achieved a high level of assembly continuity and completeness. We have functionally annotated the assembly with both long- and short-read RNA-seq, which allowed us to quantify the intron size distribution in DT in comparison to zebrafish. We expect that this work will provide an important resource for the use of *Danionella translucida* in biomedical research.

Data availability

The genome assembly and annotation files and data analysis codes will be made available on g-node.

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 (Exc. 257) to MS and BJ. BJ 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.

Bibliography

1. Tyson R Roberts. Danionella translucida, a new genus and species of cyprinid fish from burma, one of the smallest living vertebrates. *Environmental Biology of Fishes*, 16(4):231–241, 1986.
2. Ralf Britz, Kevin W. Conway, and Lukas Rüber. Spectacular morphological novelty in a miniature cyprinid fish, danionella dracula n. sp. *Proceedings of the Royal Society B: Biological Sciences*, 276(1665):2179–2186, 2009.
3. Lisanne Schulze, Jörg Henninger, Mykola Kadobianskyi, Thomas Chaigne, Ana Isabel Faustino, Nahid Hakiy, Shahad Albadri, Markus Schuelke, Leonard Maler, Filippo Del Bene, and Benjamin Judkewitz. Transparent Danionella translucida as a genetically tractable vertebrate brain model. *Nature Methods*, 15:977–983, 2018.
4. Ariadne Penalva, Jacob Bedke, Elizabeth S.B. Cook, Joshua P. Barrios, Erin P.L. Bertram, and Adam D Douglass. Establishment of the miniature fish species danionella translucida as a genetically and optically tractable neuroscience model. *bioRxiv*, 2018.
5. Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26 (10):1135–1145, 2008.
6. Mick Watson. Mind the gaps - ignoring errors in long read assemblies critically affects protein prediction. *bioRxiv*, 2018.
7. Alex Payne, Nadine Holmes, Vardhman Rakyan, and Matthew Loose. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv*, 2018.
8. Mun Hua Tan, Christopher M Austin, Michael P Hammer, Yin Peng Lee, Laurence J Croft, and Han Ming Gan. Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (Amphiprion ocellaris) genome assembly. *GigaScience*, 7(3):1–6, 2018.
9. Justin Jiang, Andrea M Quattrini, Warren R Francis, Joseph F Ryan, Estefania Rodriguez, and Catherine S McFadden. A Hybrid de novo Assembly of the Sea Pansy (Renilla muelleri) Genome. *bioRxiv*, 2018.
10. Simon Andrews. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. 2010.
11. Erik Aronesty. Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal*, 2013.
12. Rayan Chikhi and Paul Medvedev. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1):31–37, 2014.
13. Aleksey V Zimin, Guillaume Marçais, Daniela Puiu, Michael Roberts, Steven L Salzberg, and James A Yorke. The MaSuRCA genome assembler. *Bioinformatics*, 29(21):2669–2677, 2013.
14. Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, and Ashlee M Earl. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, 9(11):1–14, 2014.
15. B Bushnell. BBMap short-read aligner, and other bioinformatics tools. Available online at: http://sourceforge.net/projects/bbmap/. 2016.
16. Felipe A Simao, Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, and Evgeny M Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015.
17. Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29 (7):644–652, 2011.
18. Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R Laird, Ilias Lavidas, Zhicheng Liu, Jane E Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M Staines, Stephen J Trevanion, Bronwen L Aken, Fiona Cunningham, Andrew Yates, and Paul Flicek. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018.
19. Brandi L Cantarel, Ian Korf, Sofia MC Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. Maker: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18(1):188–196, 2008.
20. Arian FA Smit and Robert Hubley. RepeatModeler Open-1.0. Available online at: http://www.repeatmasker.org. 2008.
21. Ian Korf. Gene finding in novel genomes. *BMC Bioinformatics*, 5(1):59, 2004.
22. Mario Stanke, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. Augustus: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, 34:W435–W439, 2006.
23. The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.
24. Philip Jones, David Binns, Hsin-Yu Y Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Y Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, 30(9):1236–1240, 2014.
25. Todd M Lowe and Sean R Eddy. trnascan-se: A program for improved detection of transfer rna genes in genomic sequence. *Nucleic Acids Research*, 25(5):955–964, 1997.
26. Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2013.
27. Kerstin Howe, Matthew D Clark, Carlos F Torroja, James Torrance, Camille Berthelot, Matthieu Muffato, John E Collins, Sean Humphray, Karen McLaren, Lucy Matthews, Stuart McLaren, Ian Sealy, Mario Caccamo, Carol Churcher, Carol Scott, Jeffrey C Barrett, Romke Koch, Gerd-Jörg Rauch, Simon White, William Chow, Britt Kilian, Leonor T Quintais, José A Guerra-Assunção, Yi Zhou, Yong Gu, Jennifer Yen, Jan-Hinnerk Vogel, Tina Eyre, Seth Redmond, Ruby Banerjee, Jianxiang Chi, Beiyuan Fu, Elizabeth Langley, Sean F Maguire, Gavin K Laird, David Lloyd, Emma Kenyon, Sarah Donaldson, Harminder Sehra, Jeff Almeida-King, Jane Loveland, Stephen Trevanion, Matt Jones, Mike Quail, Dave Willey, Adrienne Hunt, John Burton, Sarah Sims, Kirsten McLay, Bob Plumb, Joy Davis, Chris Clee, Karen Oliver, Richard Clark, Clare Riddle, David Elliot, David Eliott, Glen Threadgold, Glenn Harden, Darren Ware, Sharmin Begum, Beverley Mortimore, Beverly Mortimer, Giselle Kerry, Paul Heath, Benjamin Phillimore, Alan Tracey, Nicole Corby, Matthew Dunn, Christopher Johnson, Jonathan Wood, Susan Clark, Sarah Pelan, Guy Griffiths, Michelle Smith, Rebecca Glithero, Philip Howden, Nicholas Barker, Christine Lloyd, Christopher Stevens, Joanna Harley, Karen Holt, Georgios Panagiotidis, Jamieson Lovell, Helen Beasley, Carl Henderson, Daria Gordon, Katherine Auger, Deborah Wright, Joanna Collins, Claire Raisen, Lauren Dyer, Kenric Leung, Lauren Robertson, Kirsty Ambridge, Daniel Leongamornlert, Sarah McGuire, Ruth Gilderthorp, Coline Griffiths, Deepa Manthravadi, Sarah Nichol, Gary Barker, Siobhan Whitehead, Michael Kay, Jacqueline Brown, Clare Murnane, Emma Gray, Matthew Humphries, Neil Sycamore, Darren Barker, David Saunders, Justene Wallis, Anne Babbage, Sian Hammond, Maryam Mashreghi-Mohammadi, Lucy Barr, Sancha Martin, Paul Wray, Andrew Ellington, Nicholas Matthews, Matthew Ellwood, Rebecca Woodmansey, Graham Clark, James D Cooper, James Cooper, Anthony Tromans, Darren Grafham, Carl Skuce, Richard Pandian, Robert Andrews, Elliot Harrison, Andrew Kimberley, Jane Garnett, Nigel Fosker, Rebekah Hall, Patrick Garner, Daniel Kelly, Christine Bird, Sophie Palmer, Ines Gehring, Andrea Berger, Christopher M Dooley, Zübeyde Ersan-Ürün, Cigdem Eser, Horst Geiger, Maria Geisler, Lena Karotki, Anette Kirn, Judith Konantz, Martina Konantz, Martina Oberländer, Silke Rudolph-Geiger, Mathias Teucke, Christa Lanz, Günter Raddatz, Kazutoyo Osoegawa, Baoli Zhu, Amanda Rapp, Sara Widaa, Cordelia Langford, Fengtang Yang, Stephan C Schuster, Nigel P Carter, Jennifer Harrow, Zemin Ning, Javier Herrero, Steve M J Searle, Anton Enright, Robert Geisler, Ronald H A Plasterk, Charles Lee, Monte Westerfield, Pieter J de Jong, Leonard I Zon, John H Postlethwait, Christiane Nüsslein-Volhard, Tim J P Hubbard, Hugues Roest Crollius, Jane Rogers, and Derek L Stemple. The zebrafish reference genome sequence and its relationship to the human genome. *Nature Communications*, 496(7446):498–503, 2013.
28. Ralf Britz and Kevin W Conway. Danionella dracula, an escape from the cypriniform bauplan via developmental truncation? *Journal of Morphology*, 277(2):147–166, 2016.

29. Martin Malmstrøm, Ralf Britz, Michael Matschiner, Ole K Tørresen, Renny Kurnia Hadiaty, Norsham Yaakob, Heok Hui Tan, Kjetill Sigurd Jakobsen, Walter Salzburger, and Lukas Rüber. The most developmentally truncated fishes show extensive hox gene loss and miniaturized genomes. *Genome Biology and Evolution*, 10(4):1088–1103, 2018.
30. Stephen P Moss, Domino A Joyce, Stuart Humphries, Katherine J Tindall, and David H Lunt. Comparative analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage. *Genome Biology and Evolution*, 3:1187–1196, 2011.
31. Sylvain Aubry, Steven Kelly, Britta MC Kümpers, Richard D Smith-Unna, and Julian M Hibberd. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of c4 photosynthesis. *PLOS Genetics*, 10(6):1–16, 2014.