# A high-quality chromosome-level genome assembly of rohu carp, Labeo rohita, and discovery of SNP markers

*This manuscript (permalink) was automatically generated from IGBB/rohu-genome@f366434 on September 29, 2021.*

## Authors


- **Mark A. Arick II**
  0000-0002-7207-3052 · maa146
  Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, MS, 39762 USA · Funded by Grant 7200AA18CA00030

- **Chuan-Yu Hsu**
  XXXX-XXXX-XXXX-XXXX
  Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, MS, 39762 USA

- **Zenaida Magbanua**
  XXXX-XXXX-XXXX-XXXX
  Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, MS, 39762 USA

- **Corrinne E. Grover**
  0000-0003-3878-5459
  Ecology, Evolution, and Organismal Biology Dept., Iowa State University, Ames, IA, 50010 USA

- **Emma R. Miller**
  0000-0001-9009-5303
  Ecology, Evolution, and Organismal Biology Dept., Iowa State University, Ames, IA, 50010 USA

- **Olga Pechanova**
  XXXX-XXXX-XXXX-XXXX
  Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, MS, 39762 USA

- **Charles A. Thrash**
  XXXX-XXXX-XXXX-XXXX
  Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, MS, 39762 USA

- **Ramey C. Youngblood**
  XXXX-XXXX-XXXX-XXXX
  Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, MS, 39762 USA

- **Lauren Ezzell**
  XXXX-XXXX-XXXX-XXXX
  Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, MS, 39762 USA

- **Samsul Alam**
  XXXX-XXXX-XXXX-XXXX
  placeholder

- **John Benzie**
  XXXX-XXXX-XXXX-XXXX
  placeholder

- **Matthew Hamilton**
  XXXX-XXXX-XXXX-XXXX
  placeholder

- **Attila Karsi**
  XXXX-XXXX-XXXX-XXXX
  placeholder

- **Mark L. Lawrence**
  XXXX-XXXX-XXXX-XXXX
  placeholder

- **Daniel G. Peterson**
  0000-0002-0274-5968
  Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, MS, 39762 USA

## Abstract

# Introduction

## Methods & Materials

### Genome Sequencing & Flowcytometry

ALAM – Rohu blood collection

#### Flow cytometry

The genome size of *Labeo rohita* was estimated for five samples via flow cytometry using trout erythrocyte nuclei (TENs; https://www.biosure.com/tens.html) as a standard (genome size=6.5pg). For each sample, nuclei were stabilized in 200 ul of LB01-propidium iodide (PI) buffer as per (Pellicer & Leitch, 2014), and two drops of TENs standard were used per 50ul of fish blood. Each sample was measured twice, totaling 10 runs overall. Only measurements with greater than 5,000 nuclei and a coefficient of variation (CV) of less than 3% were retained (Pellicer & Leitch, 2014).

#### Illumina short-read sequencing

A total of 2 µg of extracted genomic DNA was used for DNA-Seq library preparation using Illumina TruSeq DNA PCR-free Library Prep Kit (Illumina, San Diego, CA, USA). The final DNA-Seq library with the insert size range of 350 bp to 450 bp was submitted to Novogene company (www.en.novogene.com) for total of 2 lanes of PE150 Illumina HiSeq X-Ten (Illumina, San Diego, CA, USA) sequencing.

#### Oxford Nanopore sequencing

For each Nanopore sequencing run, 2 to 2.5 µg of genomic DNA was used in the library preparation with Nanopore Genomic DNA Ligation Sequencing Kit SQK-LSK 109 (Oxford Nanopore Technologies, Oxford, UK). The final library (about 700 to 750 ng) was loaded on a Nanopore Flow Cell R9.4.1 (Oxford Nanopore Technologies, Oxford, UK) and sequenced on GridION sequencer (Oxford Nanopore Technologies, Oxford, UK) for 48 hours. A total of 10 flow cell runs were conducted for the genome assembly.

#### Hi-C sequencing

One hundred µl of fish blood was subjected to the Hi-C library preparation using the Proximo Hi-C Animal Kit (Phase Genomics, Seattle, WA, USA). The final DNA-Seq library was submitted to Novogene company (www.en.novogene.com) for 1 lane of PE150 Illumina HiSeq X-Ten (Illumina, San Diego, CA, USA) sequencing run.

#### Bionano

### Assembly & Annotation

Jellyfish (v2.2.10) [1] and GenomeScope (v1.0) [**pcmid:PMC5870704?**] estimated the genome size using the Illumina paried reads digested into 50-mers.

Nanopore data was filtered to remove the control lambda-phage and sequences shorter the 1000 bases using the nanopack tool suite (v1.0.1) [2]. The filtered data was assembled into contigs using wtdbg2 (v2.4) [3]. The contigs were polished using the two iterations of racon (v1.4.0) [4] with minimap2 (v2.17) ([5]) mapping the nanopore reads. The contigs were further polished using pilon (v1.23) [6] with bwa (v0.7.10) [7] mapping the Illumina paired reads. The contigs were scaffolded using Bionano Solve ( Solve3.4.1_09262019) [[8]/) and SALSA (v2.3) [9]. That scaffolds larger than 10Mb were linked and oriented based on the Onychostoma macrolepis genome [10], the most similar chromosome assembly available on NCBI, using ragtag (v1.1.1) [11].

RepeatModeler [VERSION] (CITE) and RepeatMasker [VERSION] (CITE) were used to create a species-specific repeat database, and mask those repeats in the genome. All available RNA-seq libraries for *L. rohita* (ADD TABLE) were downloaded from NCBI and mapped to the masked genome using hisat2 [VERSION] (CITE). These alignments were used in the mikado [VERSION] (CITE) and braker2 [VERSION] (CITE) pipelines. Mikado uses the putative transcripts assembled from the RNA-seq alignments using stringtie [VERSION] (CITE), cufflinks [VERSION] (CITE), and trinity [VERSION] (CITE); along with the junction site prediction from portcullis [VERSION] (CITE), the alignments of the putative transcripts with uniport swiss-prot [VERSION] (CITE), and the ORFs from prodigal [VERSION] (CITE) to select the best representative for each locus. Braker2 [VERSION] (CITE) uses the RNA-seq alignments and the gene prediction from GeneMark [VERSION] (CITE) to train a species-specific Augustus model [VERSION] (CITE). Maker2 [VERSION] (CITE) predicts genes based on the new Augustus, GeneMark, and SNAP models; modifying the predictions based on the available RNA and protein evidence from the *Cyprinidae* family in the NCBI RefSeq database. Any predicted genes with an AED above 0.47 were removed from further analysis. The surviving genes were functionally annotated using InterProScan [VERSION] (CITE) and Blast [VERSION] (CITE) alignments against the uniprot swiss-prot database [VERSION] (CITE). Busco [VERSION] (CITE) using the actinopterygii_odb10 database was used to verify the completeness of both the genome and annotations.

### ddRAD-Seq & SNP Discovery

### WORLDFISH – Fin clipping collection

### ZENAIDA – ddRAD-Seq method

### CORRINNE – SNP Discovery, population analyses, interested in using the gender, which contigs associated with sex

### Data Availability

The data used for the *L. rohita* genome and annotation is available at NCBI under the BioProject PRJNA650519. The assembled genome sequence and annotations are available at GenBank under accessions JACTAM000000000. The raw data is available at the SRA (Sequence Read Archive) under accessions SRR12580210 – SRR12580221.

## Results & Discussion

### Sequencing & Assembly

The C-value of *L. rohita* was previously reported as 1.99 pg (~1.95Gb) using Feulgen densitometry [12] or 1.5Gb using k-mer estimation [13]. However, the flow cytometry results [1] show a C-value of 0.99 pg (~0.97Gb) with a standard deviation of 0.02 across all measurements. The smaller C-value is also

closer to the genome estimate produced by GenomeScope (0.97Gb) and the final genome assembly size of 0.95 Gb.

**Table 1:** Flow cytometry results for 5 *L. rohita* blood samples, measured twice. 1) (Trout erthrocyte nuclei) Genome size = 5.19pg 2) Genome estimate calculated as (average sample fluorescence/ average standard fluorescence * standard genome size ) in picograms DIPLOID (i.e., 2C)

| Specimen Name | Number of Sample nuclei | Average sample fluorescence | Number of standard nuclei | Average standard fluorescence | Estimated Genome size [2] | HAPLOID |
|---|---|---|---|---|---|---|
| Fish 1 Sample 1 | 16020 | 27350 | 2065 | 69247 | 2.049857756 | 1.024928878 |
| Fish 2 Sampe 1 | 13082 | 25929 | 6570 | 66671 | 2.018441451 | 1.009220725 5 |
| Fish 2 Sample 2 | 15402 | 25665 | 4354 | 67489 | 1.973674969 | 0.986837484 5 |
| Fish 3 Sample 1 | 15124 | 25798 | 4442 | 68195 | 1.963364176 | 0.981682088 |
| Fish 3 Sample 2 | 14923 | 25763 | 4823 | 68837 | 1.942414254 | 0.971207127 |
| Fish 4 Sample 1 | 13320 | 26346 | 5913 | 69665 | 1.962760927 | 0.981380463 5 |
| Fish 4 Sample 2 | 5624 | 26612 | 4097 | 68876 | 2.005288925 | 1.002644462 5 |
| Fish 5 Sample 1 | 6771 | 25761 | 3080 | 68825 | 1.942602107 | 0.971301053 5 |
| Fish 5 Sample 2 | 15926 | 26369 | 3352 | 68832 | 1.988248344 | 0.994124172 |
| Standard only [1] | 3 | 25258 | 3311 | 64331 | NA | |
| | | | | Average | 1.982961434 | 0.991480717 2 |
| | | | | Standard Deviation | 0.03607582 | 0.018037909 99 |

A total of 130.5 Gb of Nanopore long reads from 44.7 million read, and 261 Gb of Illumina short reads from 870 million pairs were produced, along with 382 million pairs (114 Gb) for the Hi-C library. The initial *de novo* assembly consisted of 4999 contigs with an N50 of 1.28 Mb. After the Bionano and HiC data was incorporated, the total number of sequences dropped to 2899 and the N50 increased to 29.9 Mb. The final assembly consisted of 25 chromosome length scaffolds and 2844 unplaced scaffolds, ranging in size from 1,479bp to 7.18 Mb. [2] contains a list of assembly statistics for each step. The final genome size is 97.9% of the estimated genome size. BUSCO analysis show the genome completely contains 98.1% of the 3640 genes in the actinopterygii_odb10 database with a 36 (1%) genes duplicated, 23 (0.6%) fragmented, and 46 (1.3%) missing.

**Table 2:** Assembly statistics

| n | n:500 | L50 | min | N75 | N50 | N25 | E-size | max | sum | name |
|---|---|---|---|---|---|---|---|---|---|---|
| 4999 | 4999 | 202 | 1348 | 514919 | 1281850 | 2395030 | 1727184 | 7832582 | 9.43E+08 | wtdbg2 |
| 3709 | 3706 | 15 | 1479 | 1.13E+07 | 2.65E+07 | 3.08E+07 | 2.20E+07 | 3.79E+07 | 9.46E+08 | bionano |
| 2899 | 2896 | 14 | 1479 | 2.64E+07 | 2.99E+07 | 3.43E+07 | 2.69E+07 | 4.45E+07 | 9.46E+08 | hic |

| n | n:500 | L50 | min | N75 | N50 | N25 | E-size | max | sum | name |
|---|---|---|---|---|---|---|---|---|---|---|
| 2872 | 2869 | 13 | 1479 | 2.88E+07 | 3.25E+07 | 3.61E+07 | 3.00E+07 | 4.53E+07 | 9.46E+08 | ragtag |

238509 | CDS |
239799 | exon |
9824 | five_prime_UTR |
24385 | gene |
24385 | mRNA |
1509 | three_prime_UTR |

```
==> 99-final/Rohu.busco.genome.txt <==
# BUSCO version is: 5.2.2
# The lineage dataset is: actinopterygii_odb10 (Creation date: 2020-08-05,
number of genomes: 3640, number of BUSCOs: 26)
# Summarized benchmarking in BUSCO notation for file /work/maa146/rohu-
genome/annotations/0-ref/Rohu.genome.fa
# BUSCO was run in mode: genome
# Gene predictor used: metaeuk

        ***** Results: *****

        C:98.1%[S:97.1%,D:1.0%],F:0.6%,M:1.3%,n:3640
        3571    Complete BUSCOs (C)
        3535    Complete and single-copy BUSCOs (S)
        36      Complete and duplicated BUSCOs (D)
        23      Fragmented BUSCOs (F)
        46      Missing BUSCOs (M)
        3640    Total BUSCO groups searched
```

```
==> 99-final/Rohu.busco.transcript.txt <==
# BUSCO version is: 5.2.2
# The lineage dataset is: actinopterygii_odb10 (Creation date: 2020-08-05,
number of genomes: 3640, number of BUSCOs: 26)
# Summarized benchmarking in BUSCO notation for file /work/maa146/rohu-
genome/annotations/10-rename/Rohu.transcripts.fa
# BUSCO was run in mode: transcriptome

        ***** Results: *****

        C:80.1%[S:78.2%,D:1.9%],F:5.5%,M:14.4%,n:3640
        2915    Complete BUSCOs (C)
        2845    Complete and single-copy BUSCOs (S)
        70      Complete and duplicated BUSCOs (D)
        200     Fragmented BUSCOs (F)
        525     Missing BUSCOs (M)
        3640    Total BUSCO groups searched
```

```
==> 99-final/Rohu.busco.unfiltered_transcripts.txt <==
# BUSCO version is: 5.2.2
# The lineage dataset is: actinopterygii_odb10 (Creation date: 2020-08-05,
number of genomes: 3640, number of BUSCOs: 26)
# Summarized benchmarking in BUSCO notation for file /work/maa146/rohu-
genome/annotations/5-maker/Rohu/all.transcripts.fasta
# BUSCO was run in mode: transcriptome

        ***** Results: *****

        C:83.4%[S:81.4%,D:2.0%],F:6.2%,M:10.4%,n:3640
        3036    Complete BUSCOs (C)
        2962    Complete and single-copy BUSCOs (S)
        74      Complete and duplicated BUSCOs (D)
        225     Fragmented BUSCOs (F)
        379     Missing BUSCOs (M)
        3640    Total BUSCO groups searched
```

```
==================================================
file name: Rohu.genome.fa
sequences:          2872
total length: 1128029156 bp  (945637473 bp excl N/X-runs)
GC level:         36.05 %
bases masked:  465289941 bp ( 41.25 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
Retroelements       123482     53424977 bp    4.74 %
   SINEs:                0            0 bp    0.00 %
   Penelope              0            0 bp    0.00 %
   LINEs:            29838     13842585 bp    1.23 %
    CRE/SLACS            0            0 bp    0.00 %
     L2/CR1/Rex      15537      8265099 bp    0.73 %
     R1/LOA/Jockey     981       643697 bp    0.06 %
     R2/R4/NeSL        936       732043 bp    0.06 %
     RTE/Bov-B         271       181676 bp    0.02 %
     L1/CIN4          7978      1801119 bp    0.16 %
   LTR elements:     93644     39582392 bp    3.51 %
     BEL/Pao          1287      1138420 bp    0.10 %
     Ty1/Copia         221       110739 bp    0.01 %
     Gypsy/DIRS1     32714     19073624 bp    1.69 %
       Retroviral     2443      1815904 bp    0.16 %

DNA transposons      79815     22485920 bp    1.99 %
   hobo-Activator     2242       639067 bp    0.06 %
   Tc1-IS630-Pogo    49209     16661924 bp    1.48 %
   En-Spm                0            0 bp    0.00 %
   MuDR-IS905            0            0 bp    0.00 %
   PiggyBac              0            0 bp    0.00 %
   Tourist/Harbinger  1794       259934 bp    0.02 %
   Other (Mirage,      900       209002 bp    0.02 %
    P-element, Transib)

Rolling-circles        196        71055 bp    0.01 %

Unclassified:       1774260    361869982 bp   32.08 %

Total interspersed repeats:   437780879 bp   38.81 %


Small RNA:                0            0 bp    0.00 %

Satellites:               1          267 bp    0.00 %
Simple repeats:      504151     23839806 bp    2.11 %
Low complexity:       58354      3597934 bp    0.32 %
==================================================
```

```
* most repeats fragmented by insertions or deletions
  have been counted as one element


RepeatMasker version 4.1.1 , default mode

run with rmblastn version 2.10.0+
The query was compared to classified sequences in "Rohu-families.fa"
```

## Acknowledgements

ISU Office of Biotechnology Flow Cytometry Facility

ResearchIT Iowa State University

# References


1. **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers**
Guillaume Marçais, Carl Kingsford
*Bioinformatics* (2011-03-15) https://doi.org/b7gkd6
DOI: 10.1093/bioinformatics/btr011 · PMID: 21217122 · PMCID: PMC3051319

2. **NanoPack: visualizing and processing long-read sequencing data**
Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, Christine Van Broeckhoven
*Bioinformatics* (2018-08-01) https://doi.org/gf38cx
DOI: 10.1093/bioinformatics/bty149 · PMID: 29547981 · PMCID: PMC6061794

3. **Fast and accurate long-read assembly with wtdbg2**
Jue Ruan, Heng Li
*Nature Methods* (2019-12-09) https://doi.org/ggd8j9
DOI: 10.1038/s41592-019-0669-3 · PMID: 31819265 · PMCID: PMC7004874

4. **Fast and accurate de novo genome assembly from long uncorrected reads**
Robert Vaser, Ivan Sović, Niranjan Nagarajan, Mile Šikić
*Genome Research* (2017-05) https://doi.org/gfkmzr
DOI: 10.1101/gr.214270.116 · PMID: 28100585 · PMCID: PMC5411768

5. **Minimap2: pairwise alignment for nucleotide sequences**
Heng Li
*Bioinformatics* (2018-09-15) https://doi.org/gdhbqt
DOI: 10.1093/bioinformatics/bty191 · PMID: 29750242 · PMCID: PMC6137996

6. **Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement**
Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, Ashlee M Earl
*PLoS ONE* (2014-11-19) https://doi.org/gfkmz5
DOI: 10.1371/journal.pone.0112963 · PMID: 25409509 · PMCID: PMC4237348

7. **Fast and accurate short read alignment with Burrows-Wheeler transform**
H Li, R Durbin
*Bioinformatics* (2009-05-18) https://doi.org/dqt59j
DOI: 10.1093/bioinformatics/btp324 · PMID: 19451168 · PMCID: PMC2705234

8. **Bionano Access Software**
Bionano Genomics
https://bionanogenomics.com/support-page/bionano-access-software/

9. **Integrating Hi-C links with assembly graphs for chromosome-scale assembly**
Jay Ghurye, Arang Rhie, Brian P Walenz, Anthony Schmitt, Siddarth Selvaraj, Mihai Pop, Adam M Phillippy, Sergey Koren
*PLOS Computational Biology* (2019-08-21) https://doi.org/gf74qw
DOI: 10.1371/journal.pcbi.1007273 · PMID: 31433799 · PMCID: PMC6719893

10. **Chromosome-level genome assembly of a cyprinid fish Onychostoma macrolepis by integration of nanopore sequencing, Bionano and Hi-C technology**
Lina Sun, Tian Gao, Feilong Wang, Zuliang Qin, Longxia Yan, Wenjing Tao, Minghui Li, Canbiao Jin, Li Ma, Thomas D Kocher, Deshou Wang

*Molecular Ecology Resources* (2020-07-20) https://doi.org/gmx33w
DOI: 10.1111/1755-0998.13190 · PMID: 32419357

11. **RaGOO: fast and accurate reference-guided scaffolding of draft genomes**
Michael Alonge, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J Sedlazeck, Zachary B Lippman, Michael C Schatz
*Genome Biology* (2019-10-28) https://doi.org/ggctnf
DOI: 10.1186/s13059-019-1829-6 · PMID: 31661016 · PMCID: PMC6816165

12. http://zotero.org/users/local/gVY292am/items/GYK6YE2I

13. **De novo Assembly and Genome-Wide SNP Discovery in Rohu Carp, Labeo rohita**
Paramananda Das, Lakshman Sahoo, Sofia P Das, Amrita Bit, Chaitanya G Joshi, Basdeo Kushwaha, Dinesh Kumar, Tejas M Shah, Ankit T Hinsu, Namrata Patel, … Joykrushna Jena
*Frontiers in genetics* (2020-04-21) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7186481/
DOI: 10.3389/fgene.2020.00386 · PMID: 32373166 · PMCID: PMC7186481