
















# 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@c228039](#) on September 30, 2021.

## Authors

---

Mark A. Arick<sup>1</sup> , Chuan-Yu Hsu<sup>1</sup> , Zenaida Magbanua<sup>1</sup> , Corrinne E. Grover<sup>2</sup> , Emma R. Miller<sup>2</sup> , Olga Pechanova<sup>1</sup> , Charles A. Thrash<sup>1</sup> , Ramey C. Youngblood<sup>1</sup> , Lauren Ezzell<sup>1</sup> , Samsul Alam<sup>3</sup> , John Benzie<sup>3</sup> , Matthew Hamilton<sup>3</sup> , Attila Karsi<sup>3</sup> , Mark L. Lawrence<sup>3</sup> , Daniel G. Peterson<sup>1</sup> 

## 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](http://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](http://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) [2] estimated the genome size using the Illumina paired reads digested into 50-mers.

Nanopore data was filtered to remove the control lambda-phage and sequences shorter than 1000 bases using the nanopack tool suite (v1.0.1) [3]. The filtered data was assembled into contigs using wtdbg2 (v2.4) [4]. The contigs were polished using the two iterations of racon (v1.4.0) [5] with minimap2 (v2.17) [6] mapping the nanopore reads. The contigs were further polished using pilon (v1.23) [7] with bwa (v0.7.10) [8] mapping the Illumina paired reads. The contigs were scaffolded using Bionano Solve (Solve3.4.1\_09262019) [9] and SALSA (v2.3) [10]. That scaffolds larger than 10Mb were linked and oriented based on the *Onychostoma macrolepis* genome [11], the most similar chromosome assembly available on NCBI, using ragtag (v1.1.1) [12].

RepeatModeler (v2.0.1) [13] and RepeatMasker (v4.1.1) [repeatmasker?] were used to create a species-specific repeat database, and mask those repeats in the genome. All available RNA-seq libraries for *L. rohita* [1] were downloaded from NCBI and mapped to the masked genome using hisat2 (v2.1.0) [14]. These alignments were used in the mikado (v2.0rc2) [15] and braker2 (v2.1.5) [16] pipelines. Mikado uses the putative transcripts assembled from the RNA-seq alignments using stringtie (v2.1.2) [17], cufflinks (v2.2.1) [18], and trinity (v2.11.0) [19]; along with the junction site prediction from portcullis (v1.2.2) [20], the alignments of the putative transcripts with UniprotKB Swiss-Prot (v2021.03) [uniprot?], and the ORFs from prodigal (v2.6.3) [21] to select the best representative for each locus. Braker2 uses the RNA-seq alignments and the gene prediction from GeneMark-ES (v4.61) [22] to train a species-specific Augustus (v3.3.3) [23] model. Maker2 (v2.31.10) [24] 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 (v5.47-82.0) [25] and BLAST+ (v2.9.0) [26] alignments against the UniprotKB Swiss-Prot database. Busco (v5.2.2) [27] was used to verify the completeness of both the genome and annotations against the actinopterygii\_odb10 database.

**Table 1:** List of SRA accessions used in annotation pipeline. A table of all metadata available for these accessions can be found [here](#).

BioProject	BioSample	Run	Instrument	sex	Tissue
PRJNA401304	SAMN07602342	SRR6003546	Illumina HiSeq 2000	female	Brain
PRJNA401304	SAMN07602341	SRR6003547	Illumina HiSeq 2000	female	Brain
PRJNA401304	SAMN07602344	SRR6003548	Illumina HiSeq 2000	female	Pituitary

BioProject	BioSample	Run	Instrument	sex	Tissue
PRJNA401304	SAMN07602343	SRR6003549	Illumina HiSeq 2000	female	Pituitary
PRJNA401304	SAMN07602346	SRR6003550	Illumina HiSeq 2000	female	Gonad
PRJNA401304	SAMN07602345	SRR6003551	Illumina HiSeq 2000	female	Gonad
PRJNA401304	SAMN07602348	SRR6003552	Illumina HiSeq 2000	female	Liver
PRJNA401304	SAMN07602347	SRR6003553	Illumina HiSeq 2000	female	Liver
PRJNA449818	SAMN08918388	SRR6987066	NextSeq 500	female	Pooled tissue
PRJNA449818	SAMN08918389	SRR6987067	NextSeq 500	male	Pooled tissue
PRJNA449818	SAMN08918390	SRR6987068	NextSeq 500	female	whole body
PRJNA528865	SAMN11246839	SRR8816555	Illumina HiSeq 2500	not applicable	Liver
PRJNA528865	SAMN11246841	SRR8816556	Illumina HiSeq 2500	not applicable	Liver
PRJNA528865	SAMN11246840	SRR8816557	Illumina HiSeq 2500	not applicable	Liver
PRJNA450719	SAMN08944450	SRR7027730	NextSeq 500	female	Pooled tissue
PRJNA450719	SAMN08944449	SRR7027731	NextSeq 500	male	Pooled tissue
PRJNA450719	SAMN08944451	SRR7027732	NextSeq 500	male	Whole body

## ddRAD-Seq & SNP Discovery

### WORLD FISH – 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 [28] or 1.5Gb using k-mer estimation [29]. However, the flow cytometry results (Table [2]) 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 2:** Flow cytometry results for 5 *L. rohita* blood samples, measured twice. 1) Trout erythrocyte 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 <sup>2</sup>	HAPLOID
Fish 1 Sample 1	16020	27350	2065	69247	2.049857756	1.024928878
Fish 2 Sample 1	13082	25929	6570	66671	2.018441451	1.0092207255
Fish 2 Sample 2	15402	25665	4354	67489	1.973674969	0.9868374845
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	0.9813804635
Fish 4 Sample 2	5624	26612	4097	68876	2.005288925	1.0026444625
Fish 5 Sample 1	6771	25761	3080	68825	1.942602107	0.9713010535
Fish 5 Sample 2	15926	26369	3352	68832	1.988248344	0.994124172
Standard only <sup>1</sup>	3	25258	3311	64331	NA	
				Average	1.982961434	0.9914807172
				Standard Deviation	0.03607582	0.01803790999

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. Table [3] contains a common assembly statistics for each step. The final genome size is 97.9% of the estimated genome size. The annotation pipeline produce 34,590 primary transcripts, 24,385 surviving the AED filter. BUSCO analysis show the genome completely contains 98.1% of the 3640 orthologs in the actinopterygii\_odb10 database with 36 (1%) duplicated; however, the filtered transcriptome only contained 80% of the totla orthologs complete with 70 duplicated. A complete comparison of the BUSCO analyses can be found in Table [4].

**Table 3:** Assembly statistics for each stage of the assembly

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
2872	2869	13	1479	2.88E+07	3.25E+07	3.61E+07	3.00E+07	4.53E+07	9.46E+08	ragtag

**Table 4:** BUSCO analysis for the genome and transcriptome, before and after AED filtering.

Type	Genome	Unfiltered Transcriptome	Filtered Transcriptome
Complete BUSCOs (C)	3571	3036	2915
Complete and single-copy BUSCOs (S)	3535	2962	2845
Complete and duplicated BUSCOs (D)	36	74	70
Fragmented BUSCOs (F)	23	225	200
Missing BUSCOs (M)	46	379	525
Total BUSCO groups searched	3640	3640	3640

```

=====
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 elements*	length occupied	percentage 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, P-element, Transib)	900	209002 bp	0.02 %
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](https://doi.org/10.1093/bioinformatics/btr011) · PMID: [21217122](https://pubmed.ncbi.nlm.nih.gov/21217122/) · PMCID: [PMC3051319](https://pubmed.ncbi.nlm.nih.gov/PMC3051319/)
2. **GenomeScope: fast reference-free genome profiling from short reads**  
Gregory W Vulture, Fritz J Sedlazeck, Maria Nattestad, Charles J Underwood, Han Fang, James Gurtowski, Michael C Schatz  
*Bioinformatics (Oxford, England)* (2017-07-15)  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5870704/>  
DOI: [10.1093/bioinformatics/btx153](https://doi.org/10.1093/bioinformatics/btx153) · PMID: [28369201](https://pubmed.ncbi.nlm.nih.gov/28369201/) · PMCID: [PMC5870704](https://pubmed.ncbi.nlm.nih.gov/PMC5870704/)
3. **NanoPack: visualizing and processing long-read sequencing data**  
Wouter De Coster, Sven D'Hert, Darrin T Schultz, Marc Cruts, Christine Van Broeckhoven  
*Bioinformatics* (2018-08-01) <https://doi.org/gf38cx>  
DOI: [10.1093/bioinformatics/bty149](https://doi.org/10.1093/bioinformatics/bty149) · PMID: [29547981](https://pubmed.ncbi.nlm.nih.gov/29547981/) · PMCID: [PMC6061794](https://pubmed.ncbi.nlm.nih.gov/PMC6061794/)
4. **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](https://doi.org/10.1038/s41592-019-0669-3) · PMID: [31819265](https://pubmed.ncbi.nlm.nih.gov/31819265/) · PMCID: [PMC7004874](https://pubmed.ncbi.nlm.nih.gov/PMC7004874/)
5. **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](https://doi.org/10.1101/gr.214270.116) · PMID: [28100585](https://pubmed.ncbi.nlm.nih.gov/28100585/) · PMCID: [PMC5411768](https://pubmed.ncbi.nlm.nih.gov/PMC5411768/)
6. **Minimap2: pairwise alignment for nucleotide sequences**  
Heng Li  
*Bioinformatics* (2018-09-15) <https://doi.org/gdhbqt>  
DOI: [10.1093/bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191) · PMID: [29750242](https://pubmed.ncbi.nlm.nih.gov/29750242/) · PMCID: [PMC6137996](https://pubmed.ncbi.nlm.nih.gov/PMC6137996/)
7. **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](https://doi.org/10.1371/journal.pone.0112963) · PMID: [25409509](https://pubmed.ncbi.nlm.nih.gov/25409509/) · PMCID: [PMC4237348](https://pubmed.ncbi.nlm.nih.gov/PMC4237348/)
8. **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](https://doi.org/10.1093/bioinformatics/btp324) · PMID: [19451168](https://pubmed.ncbi.nlm.nih.gov/19451168/) · PMCID: [PMC2705234](https://pubmed.ncbi.nlm.nih.gov/PMC2705234/)
9. **Bionano Access Software**  
Bionano Genomics  
<https://bionanogenomics.com/support-page/bionano-access-software/>
10. **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](https://doi.org/10.1371/journal.pcbi.1007273) · PMID: [31433799](https://pubmed.ncbi.nlm.nih.gov/31433799/) · PMCID: [PMC6719893](https://pubmed.ncbi.nlm.nih.gov/PMC6719893/)

11. **Chromosome-level genome assembly of a cyprinid fish <i>Onychostoma macrolepis</i> 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](https://doi.org/10.1111/1755-0998.13190) · PMID: [32419357](https://pubmed.ncbi.nlm.nih.gov/32419357/)
12. **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](https://doi.org/10.1186/s13059-019-1829-6) · PMID: [31661016](https://pubmed.ncbi.nlm.nih.gov/31661016/) · PMCID: [PMC6816165](https://pubmed.ncbi.nlm.nih.gov/PMC6816165/)
13. **RepeatModeler2 for automated genomic discovery of transposable element families**  
Jullien M Flynn, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G Clark, Cédric Feschotte, Arian F Smit  
*Proceedings of the National Academy of Sciences* (2020-04-28) <https://doi.org/ggsnv2>  
DOI: [10.1073/pnas.1921046117](https://doi.org/10.1073/pnas.1921046117) · PMID: [32300014](https://pubmed.ncbi.nlm.nih.gov/32300014/) · PMCID: [PMC7196820](https://pubmed.ncbi.nlm.nih.gov/PMC7196820/)
14. **Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype**  
Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, Steven L Salzberg  
*Nature Biotechnology* (2019-08-02) <https://doi.org/gf5395>  
DOI: [10.1038/s41587-019-0201-4](https://doi.org/10.1038/s41587-019-0201-4) · PMID: [31375807](https://pubmed.ncbi.nlm.nih.gov/31375807/) · PMCID: [PMC7605509](https://pubmed.ncbi.nlm.nih.gov/PMC7605509/)
15. **Leveraging multiple transcriptome assembly methods for improved gene structure annotation**  
Luca Venturini, Shabhonam Caim, Gemy George Kaithakottil, Daniel Lee Mapleson, David Swarbreck  
*GigaScience* (2018-08) <https://doi.org/gmx7k4>  
DOI: [10.1093/gigascience/giy093](https://doi.org/10.1093/gigascience/giy093) · PMID: [30052957](https://pubmed.ncbi.nlm.nih.gov/30052957/) · PMCID: [PMC6105091](https://pubmed.ncbi.nlm.nih.gov/PMC6105091/)
16. **BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database**  
Tomáš Brůna, Katharina J Hoff, Alexandre Lomsadze, Mario Stanke, Mark Borodovsky  
*NAR Genomics and Bioinformatics* (2021-03) <https://doi.org/gmx7k6>  
DOI: [10.1093/nargab/lqaa108](https://doi.org/10.1093/nargab/lqaa108) · PMID: [33575650](https://pubmed.ncbi.nlm.nih.gov/33575650/) · PMCID: [PMC7787252](https://pubmed.ncbi.nlm.nih.gov/PMC7787252/)
17. **Transcriptome assembly from long-read RNA-seq alignments with StringTie2**  
Sam Kovaka, Aleksey V Zimin, Geo M Pertea, Roham Razaghi, Steven L Salzberg, Mihaela Pertea  
*Genome Biology* (2019-12-16) <https://doi.org/ghrk2k>  
DOI: [10.1186/s13059-019-1910-1](https://doi.org/10.1186/s13059-019-1910-1) · PMID: [31842956](https://pubmed.ncbi.nlm.nih.gov/31842956/) · PMCID: [PMC6912988](https://pubmed.ncbi.nlm.nih.gov/PMC6912988/)
18. **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**  
Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, Lior Pachter  
*Nature Protocols* (2012-03-01) <https://doi.org/f4pbzd>  
DOI: [10.1038/nprot.2012.016](https://doi.org/10.1038/nprot.2012.016) · PMID: [22383036](https://pubmed.ncbi.nlm.nih.gov/22383036/) · PMCID: [PMC3334321](https://pubmed.ncbi.nlm.nih.gov/PMC3334321/)
19. **Full-length transcriptome assembly from RNA-Seq data without a reference genome**  
Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, ... Aviv Regev

*Nature Biotechnology* (2011-05-15) <https://doi.org/b2bctj>  
DOI: [10.1038/nbt.1883](https://doi.org/10.1038/nbt.1883) · PMID: [21572440](https://pubmed.ncbi.nlm.nih.gov/21572440/) · PMCID: [PMC3571712](https://pubmed.ncbi.nlm.nih.gov/PMC3571712/)

20. **Efficient and accurate detection of splice junctions from RNA-seq with Portcullis**  
Daniel Mapleson, Luca Venturini, Gemy Kaithakottil, David Swarbreck  
*GigaScience* (2018-12) <https://doi.org/gkzg5r>  
DOI: [10.1093/gigascience/giy131](https://doi.org/10.1093/gigascience/giy131) · PMID: [30418570](https://pubmed.ncbi.nlm.nih.gov/30418570/) · PMCID: [PMC6302956](https://pubmed.ncbi.nlm.nih.gov/PMC6302956/)
21. **Prodigal: prokaryotic gene recognition and translation initiation site identification**  
Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, Loren J Hauser  
*BMC Bioinformatics* (2010-03-08) <https://doi.org/cktxnm>  
DOI: [10.1186/1471-2105-11-119](https://doi.org/10.1186/1471-2105-11-119) · PMID: [20211023](https://pubmed.ncbi.nlm.nih.gov/20211023/) · PMCID: [PMC2848648](https://pubmed.ncbi.nlm.nih.gov/PMC2848648/)
22. **Eukaryotic Gene Prediction Using GeneMark.hmm-E and GeneMark-ES**  
Mark Borodovsky, Alex Lomsadze  
*Current Protocols in Bioinformatics* (2011-09-07) <https://doi.org/b75r2g>  
DOI: [10.1002/0471250953.bi0406s35](https://doi.org/10.1002/0471250953.bi0406s35) · PMID: [21901742](https://pubmed.ncbi.nlm.nih.gov/21901742/) · PMCID: [PMC3204378](https://pubmed.ncbi.nlm.nih.gov/PMC3204378/)
23. **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources**  
Mario Stanke, Oliver Schöffmann, Burkhard Morgenstern, Stephan Waack  
*BMC Bioinformatics* (2006-02-09) <https://doi.org/cv8xsn>  
DOI: [10.1186/1471-2105-7-62](https://doi.org/10.1186/1471-2105-7-62) · PMID: [16469098](https://pubmed.ncbi.nlm.nih.gov/16469098/) · PMCID: [PMC1409804](https://pubmed.ncbi.nlm.nih.gov/PMC1409804/)
24. **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects**  
Carson Holt, Mark Yandell  
*BMC Bioinformatics* (2011-12-22) <https://doi.org/fz39nj>  
DOI: [10.1186/1471-2105-12-491](https://doi.org/10.1186/1471-2105-12-491) · PMID: [22192575](https://pubmed.ncbi.nlm.nih.gov/22192575/) · PMCID: [PMC3280279](https://pubmed.ncbi.nlm.nih.gov/PMC3280279/)
25. **InterProScan 5: genome-scale protein function classification**  
P Jones, D Binns, H-Y Chang, M Fraser, W Li, C McAnulla, H McWilliam, J Maslen, A Mitchell, G Nuka, ... S Hunter  
*Bioinformatics* (2014-01-21) <https://doi.org/f53532>  
DOI: [10.1093/bioinformatics/btu031](https://doi.org/10.1093/bioinformatics/btu031) · PMID: [24451626](https://pubmed.ncbi.nlm.nih.gov/24451626/) · PMCID: [PMC3998142](https://pubmed.ncbi.nlm.nih.gov/PMC3998142/)
26. **BLAST+: architecture and applications**  
Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, Thomas L Madden  
*BMC Bioinformatics* (2009-12-15) <https://doi.org/cnjxgz>  
DOI: [10.1186/1471-2105-10-421](https://doi.org/10.1186/1471-2105-10-421) · PMID: [20003500](https://pubmed.ncbi.nlm.nih.gov/20003500/) · PMCID: [PMC2803857](https://pubmed.ncbi.nlm.nih.gov/PMC2803857/)
27. **BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes**  
Mosè Manni, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, Evgeny M Zdobnov  
*Molecular Biology and Evolution* (2021-10-01) <https://doi.org/gmgv52>  
DOI: [10.1093/molbev/msab199](https://doi.org/10.1093/molbev/msab199) · PMID: [34320186](https://pubmed.ncbi.nlm.nih.gov/34320186/) · PMCID: [PMC8476166](https://pubmed.ncbi.nlm.nih.gov/PMC8476166/)
28. <http://zotero.org/users/local/gVY292am/items/GYK6YE2I>
29. **<i>De novo</i> Assembly and Genome-Wide SNP Discovery in Rohu Carp, <i>Labeo rohita</i>**  
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](https://doi.org/10.3389/fgene.2020.00386) · PMID: [32373166](https://pubmed.ncbi.nlm.nih.gov/32373166/) · PMCID: [PMC7186481](https://pubmed.ncbi.nlm.nih.gov/PMC7186481/)

