

739A: Zebrafish Genome Resources at NCBI

Nuala O'Leary; Terence Murphy ; Francoise Thibaud-Nissen; Valerie Schneider; Eukaryotic Genome Annotation Group; RefSeq Curation Group

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA



U.S. National Library of Medicine
National Center for Biotechnology Information

Abstract

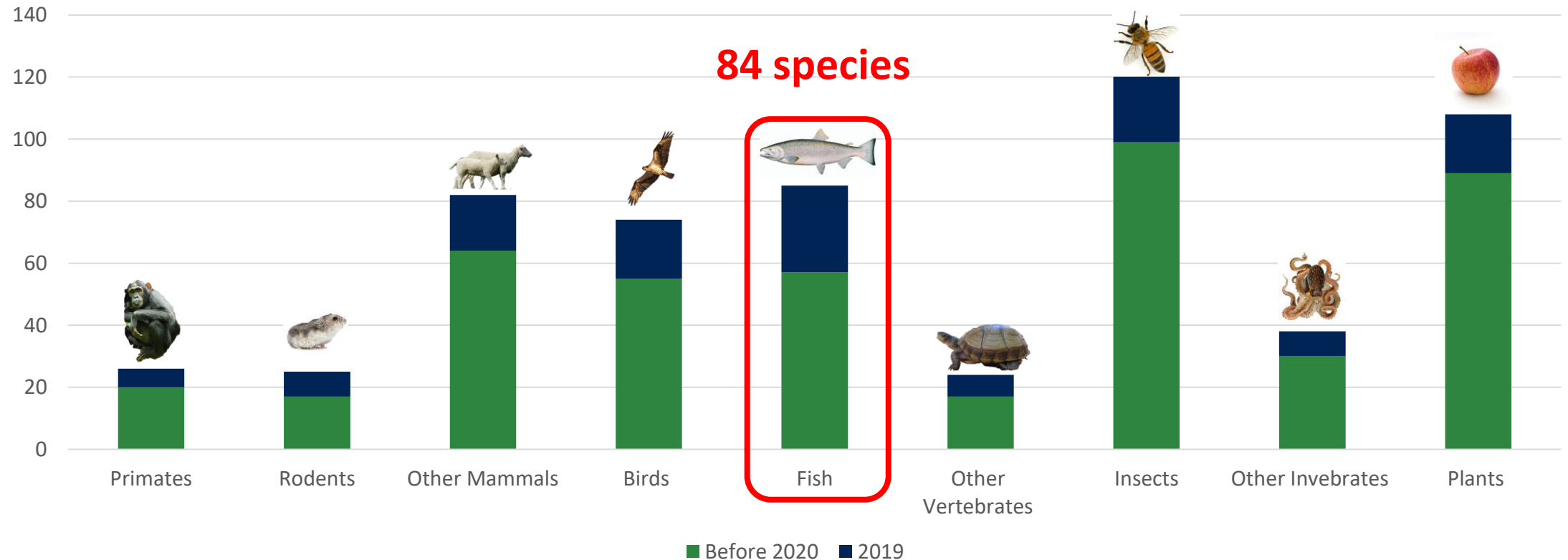


Genomic research on model organisms such as zebrafish (*Danio rerio*) relies on the availability of high-quality annotated reference genomes to facilitate consistent reporting and mapping of genetic data. The Reference Sequence (RefSeq) project (<https://www.ncbi.nlm.nih.gov/refseq/>) at the NCBI provides a comprehensive annotation of the zebrafish Tuebingen strain reference genome assembly (GRCz11) maintained by the Genome Reference Consortium (GRC). The zebrafish RefSeq dataset is generated by a combination of computational analysis and manual curation that results in an annotation that focuses on representation of all full-length, non-redundant transcripts.

The primary sources of data used in this annotation pipeline include mRNAs, expressed sequence tags (ESTs), protein data, RNA-seq datasets, and protein homology. Zebrafish is one of a select group of vertebrates that are the major focus of RefSeq's manual curation efforts, which involves the in-depth review of sequence data to define new transcript variants, resolve sequence errors, and remove inaccurate information. We also collaborate with expert groups, including the Zebrafish Information Network (ZFIN) and UniProtKB, to provide appropriate annotation and nomenclature for both genes and proteins.

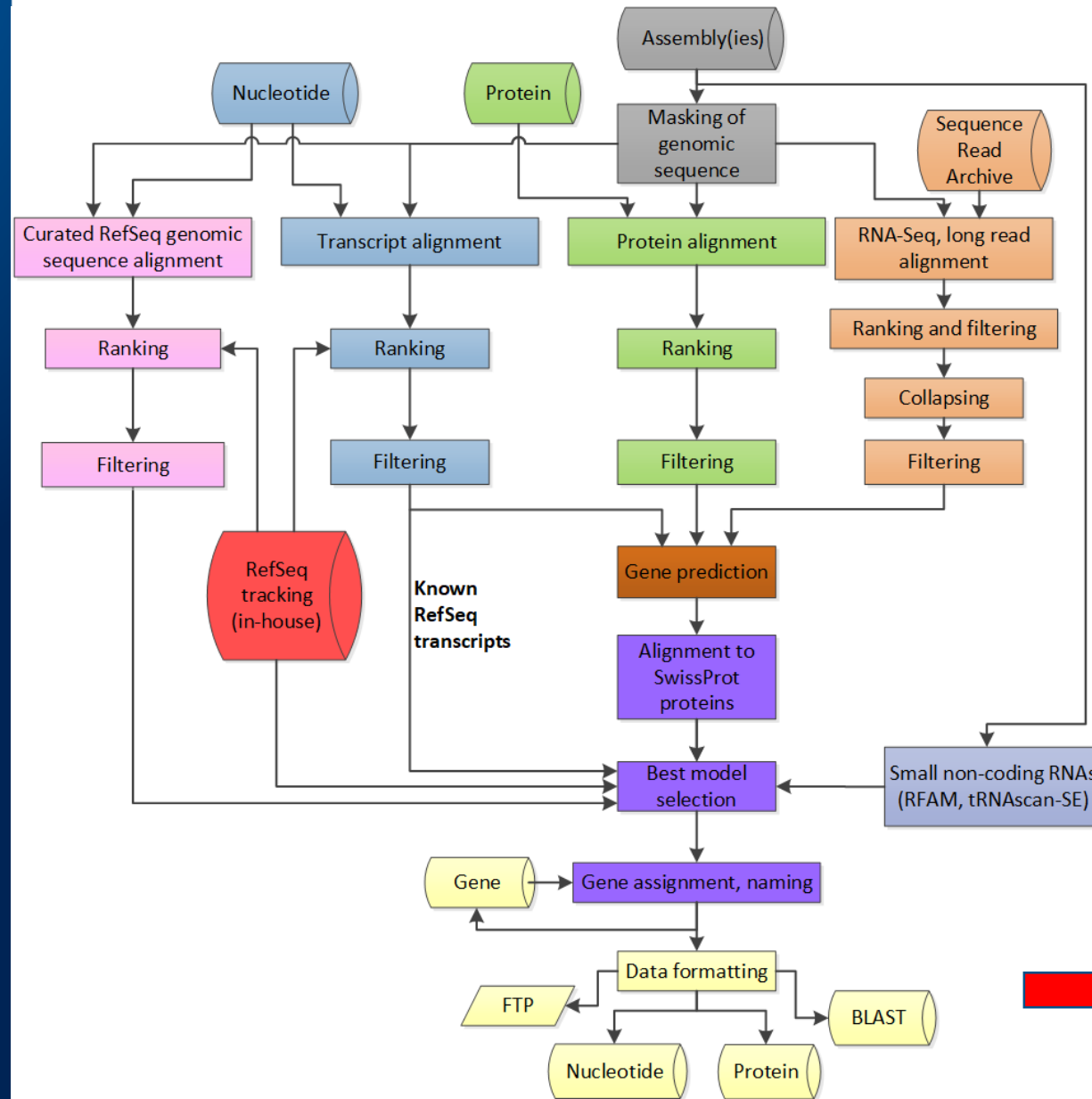
In addition to zebrafish, NCBI provides stable reference genome annotation for other fish species with high-quality genome assembly data submitted to NCBI's Assembly resource (<https://www.ncbi.nlm.nih.gov/assembly/>). To date, 80 fish species have RefSeq annotated genomes, providing a valuable resource for comparative genomic research. We have recently expanded our gene ortholog dataset for fish along with a new visualization tool for comparing homologous gene sets. In this poster presentation we will provide an overview of NCBI's zebrafish genome resources and highlight the utility of these resources to the zebrafish research community. We will also provide practical guidance on how to access RefSeq data and tools for analysis of individual genes as well as whole genome datasets.

Genomes annotated by NCBI



Eukaryotic Genome Annotation Pipeline

- Automated
- Evidence-based
- RNA-seq
- Long reads
- Proteins



Access Genome
sequence and
annotation through
NCBI Datasets

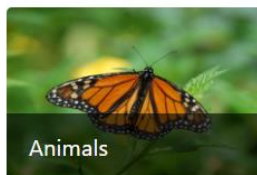
Introducing NCBI Datasets

A new way to
access genome
sequence and
annotation at NCBI





Welcome to NCBI Datasets BETA

NCBI Datasets is an experimental resource for finding and building datasets - and we're just getting started! Our web interface allows you to download genome sequence and annotation for eukaryotic organisms. For access to data for all organisms, including bacteria and viruses, use our command line tool and RESTful APIs.

[CONTINUE READING](#)



Well-studied species

	Homo sapiens human	121 assemblies
	Mus musculus house mouse	22 assemblies
	Arabidopsis thaliana thale cress	24 assemblies
	Danio rerio zebrafish	6 assemblies

Web access

Command line access

Command-line

Get data using our command-line program, available for Linux, Windows, and Mac.

[DOCUMENTATION](#)

API access

APIs

www.ncbi.nlm.nih.gov/datasets

Easy download of genome sequence and annotation

Choice of file types

Download

Data from 1 assembly

- ☒ Genomic sequence (FASTA)
- ☒ Annotated features (GFF3)
- ☐ Sequence and annotation (GBFF)
- ☐ Transcripts (FASTA)
- ☐ Protein (FASTA)

Selected files will be downloaded as a structure ZIP file

Estimated download size: 393.99 MB

Name your file

CANCEL **DOWNLOAD**

Genomes - NCBI Datasets BETA

NCBI Datasets is an experimental project for finding and building datasets

[NCBI Datasets](#) [Command-line tool](#) [API documentation](#)

DOWNLOAD 1

zebrafish All or RefSeq

Species	Assembly	Annotation	Level	Contig N50	Size	Year
<i>Danio rerio</i>						8 assemblies
<input checked="" type="checkbox"/> zebrafish Strain: Tuebingen	GRCz11 reference RefSeq: GCF_000002035.6	NCBI Release 106	Chromosome	1,422 kb	1.373 Gbp	2017
<input type="checkbox"/> zebrafish Strain: Tuebingen	GRCz11 reference GenBank: GCA_000002035.4	none submitted	Chromosome	1,422 kb	1.373 Gbp	2017
<input type="checkbox"/> zebrafish Strain: Tuebingen	GRCz10 RefSeq: GCF_000002035.5	none submitted	Chromosome	1,258 kb	1.372 Gbp	2014
<input type="checkbox"/> zebrafish Strain: Tuebingen	ASM869237v1	none submitted	Chromosome	417 kb	1.306 Gbp	2019
<input type="checkbox"/> zebrafish Strain: Tuebingen	Assembly WGS3...	none submitted	Scaffold	24 kb	1.443 Gbp	2015

Taxonomy filter

annotation

ncbi_datasets.zip

Data packaged in single folder

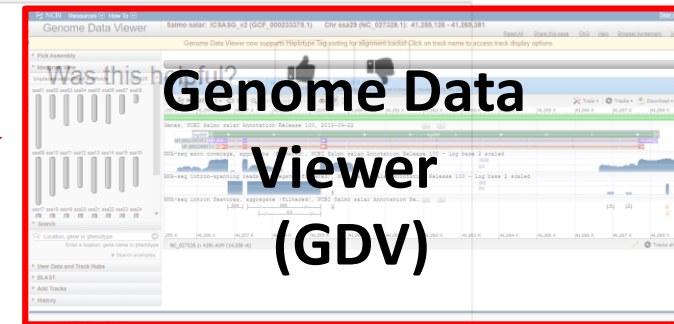
U.S. National Library of Medicine
National Center for Biotechnology Information

Finding genes

Search NCBI

zebrafish dmrt3a

Search



GENE

[dmrt3a – doublesex and mab-3 related transcription factor 3a](#)

[Danio rerio \(zebrafish\)](#)

Also known as: dmrt3a1, zDmrt3a, zgc:101766

GeneID: 450035

[RefSeq transcripts \(1\)](#) [RefSeq proteins \(1\)](#) [PubMed \(14\)](#)

Orthologs

Genome Browser

BLAST

Download

RefSeq Sequences

Showing 1 of 1 (by status, accession number)

Transcript	nt	Protein	aa	Isoform	Status
<u>NM_001005779.2</u>	1,920	<u>NP_001005779.2</u>	448		curated

Nucleotide/Protein

Protein

GenPept

myeloid differentiation primary response protein MyD88 [Salmo salar]

NCBI Reference Sequence: NP_001130017.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to:

LOCUS NP_001130017 283 aa linear VRT 22-DEC-2019

DEFINITION myeloid differentiation primary response protein MyD88 [Salmo salar].

ACCESSION NP_001130017

VERSION NP_001130017.1

ORGANISM Salmo salar

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Protacanthopterygii; Salmoniformes; Salmonidae; Salmoninae; Salmo.

REFERENCE 1 (residues 1 to 283)

AUTHORS Iliev DB, Sophtehz H, Fremmerlid K and Jorgensen JB.

TITLE MyD88 interacts with interferon regulatory factor (IRF) 3 and IRF7 in Atlantic salmon (Salmo salar): transgenic SsMyD88 modulates the INF-induced type I interferon response and accumulates in B6B-resomes

JOURNAL J. Biol. Chem. 286 (49), 42715-42724 (2011)

REMARK GenBank: MyD88 interacts with interferon regulatory factor (IRF) 3 and IRF7 in Atlantic salmon (Salmo salar)

REFERENCE 2 (residues 1 to 283)

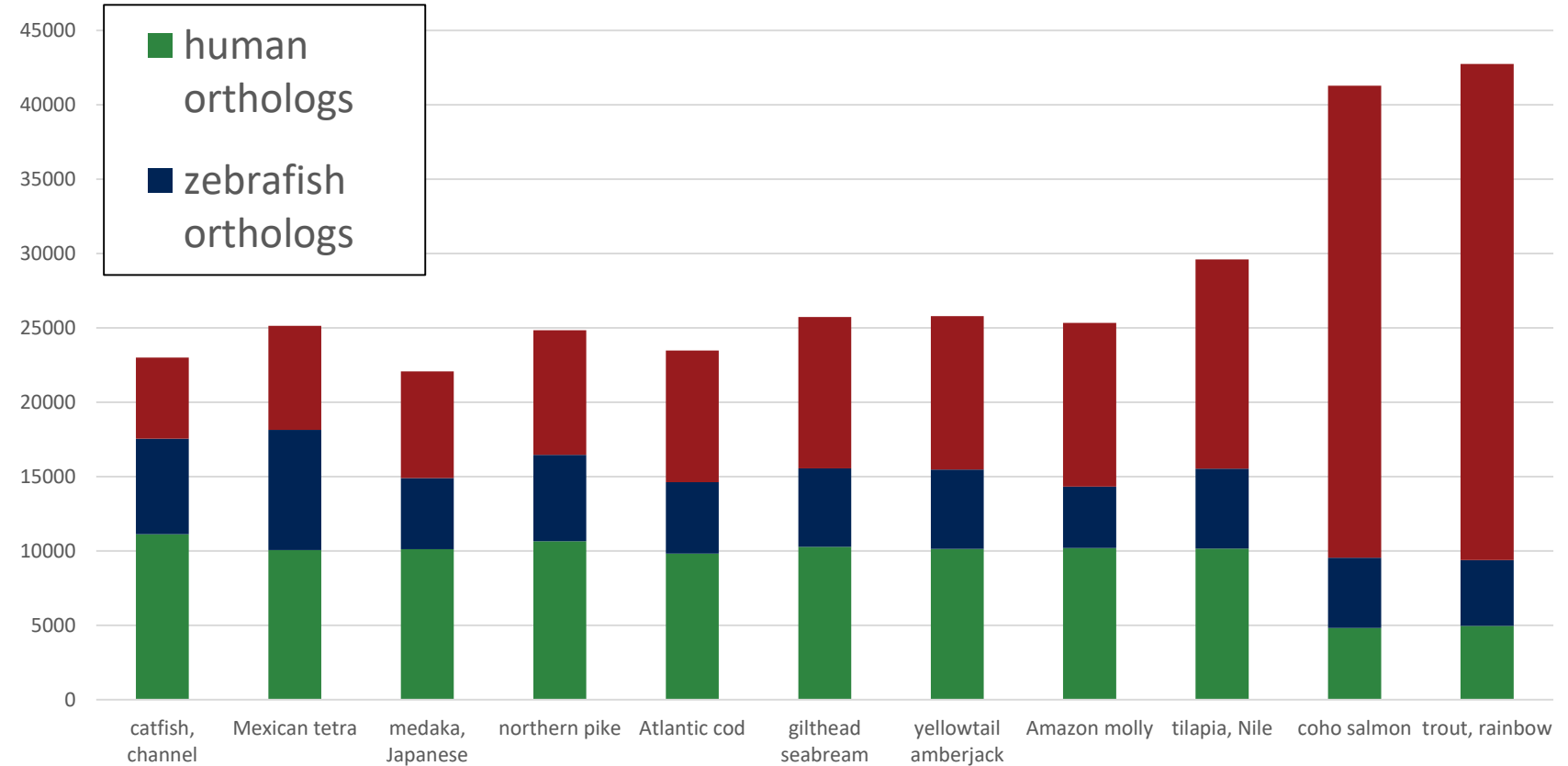
AUTHORS Skjæveland I, Iliev DB, Strandskog G and Jorgensen JB.

TITLE Identification and characterization of TLR8 and MyD88 homologs in Atlantic salmon (Salmo salar)

NCBI

New! Fish ortholog sets

- 1:1 orthologs vs zebrafish
- Report human gene if identified
- Improve nomenclature through ortholog connections



NCBI Orthologs

www.ncbi.nlm.nih.gov/gene/58524/ortholog/?scope=7898

DMRT3 - doublesex and mab-3 related transcription factor 3

Genes similar to DMRT3

NCBI Orthologs [How was this calculated?](#)

0 items

SEARCH THE TAXONOMY TREE

Enter taxonomic name

Craniata

- vertebrates
 - birds
 - alligators and others
 - turtles
 - lizards
 - mammals
 - amphibians
 - coelacanth
- bony fishes**
 - gars
 - teleost fishes
 - reedfish
- lampreys
- cartilaginous fishes
 - chimaeras
 - Callorhynchus milii
 - sharks and rays
 - Rhincodon typus
 - Amblyraja radiata

Taxonomy navigation

Genes Literature

71 genes for: bony fishes (*Actinopterygii*)

0 selected.

Add to cart

Protein alignment

Download

Download sequence and tabular data

Transcript and protein info

Species	Gene	Architecture	aa				
<input type="checkbox"/> <i>Danio rerio</i> zebrafish	dmrt3a doublesex and mab-3 related transcription factor 3a		448				
<div><div>RefSeq transcripts (1)</div><div>RefSeq proteins (1)</div><div>Architecture</div><div>aa</div></div> <table><tbody><tr><td>NM_001005779.2</td><td>NP_001005779.2</td><td></td><td>448</td></tr></tbody></table> <div><div>Genome Browser</div><div>InterPro ↗</div></div>				NM_001005779.2	NP_001005779.2		448
NM_001005779.2	NP_001005779.2		448				
<input type="checkbox"/> <i>Takifugu rubripes</i> torafugu	dmrt3a doublesex and mab-3 related transcription factor 3a		468				
<input type="checkbox"/> <i>Astyanax mexicanus</i> Mexican tetra	dmrt3a doublesex and mab-3 related transcription factor 3a		436				
<input type="checkbox"/> <i>Lepidosteus oculatus</i>	dmrt3a		452				

Search genes and regions

View NCBI gene annotation and navigate through exons

Download Track Data

Add UCSC-style
track hubs and
custom data

Genome Data Viewer

Danio rerio
(zebrafish)

Assembly: **GRCz11**

Search assembly
Location, gene or phenotype

Examples ▶

► Pick Assembly

Ideogram View

▼ User Data and Track Hubs ?

- no added tracks or track hubs - Options ▼

Supported File/Data types

► BLAST

► Add Tracks

► Assembly Region Details

► History



Acknowledgements

- This work was done as part of the Authors' official duties as NIH employees and is a Work of the United States Government. Therefore, copyright may not be established in the United States. 17 U.S.C. § 105

Send Feedback

<https://www.ncbi.nlm.nih.gov/projects/RefSeq/update.cgi>

Watch NCBI News for updates!

<https://ncbiinsights.ncbi.nlm.nih.gov/>

<https://www.youtube.com/user/NCBINLM>