

---

# **AquaMine**

***Release 1.2***

**Elsik Lab**

**Jul 26, 2023**



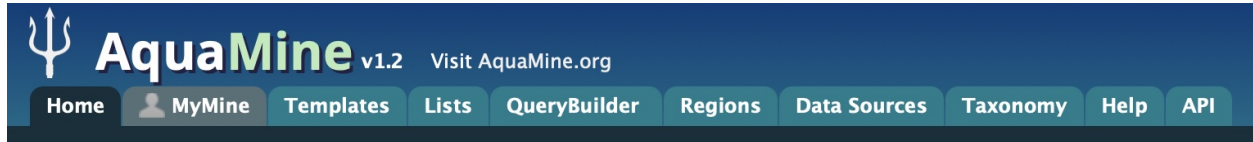
# TABLE OF CONTENTS

<b>1</b>	<b>Overview of AquaMine</b>	<b>3</b>
<b>2</b>	<b>Navigation and Searching in AquaMine</b>	<b>5</b>
2.1	Quick Search . . . . .	5
2.2	Templates . . . . .	7
2.2.1	Generate query code . . . . .	8
2.2.2	Download results . . . . .	9
2.2.3	Customize output and manage columns . . . . .	10
2.2.4	Optional filters . . . . .	10
2.3	QueryBuilder . . . . .	11
2.3.1	Model browser . . . . .	11
2.3.2	QueryBuilder Examples . . . . .	13
2.3.2.1	Example 1: Querying for protein coding genes . . . . .	13
2.3.2.2	Example 2: Querying for protein coding genes on a particular chromosome . . . . .	15
2.3.2.3	Example 3: Querying for Protein Coding genes on a particular chromosome and their exons . . . . .	15
<b>3</b>	<b>Report Pages</b>	<b>21</b>
3.1	Summary . . . . .	22
3.2	Transcripts . . . . .	22
3.3	Proteins . . . . .	23
3.4	Function . . . . .	24
3.5	Homology . . . . .	24
3.6	Publications . . . . .	26
3.7	Other . . . . .	26
<b>4</b>	<b>Genomic Regions Search</b>	<b>27</b>
<b>5</b>	<b>Lists</b>	<b>31</b>
5.1	Creating Lists . . . . .	31
5.2	Saving Lists . . . . .	34
<b>6</b>	<b>MyMine</b>	<b>37</b>
<b>7</b>	<b>API</b>	<b>39</b>
<b>8</b>	<b>Data Sources</b>	<b>41</b>
<b>9</b>	<b>How to cite</b>	<b>43</b>





**AquaMine** is a data mining resource that integrates genomic data of importance to U.S. aquaculture. It combines data generated by the aquaculture research community with data from external public databases. Many customized bioinformatics tools are available that researchers with or without bioinformatic programming skills can use to create their own crafted datasets for use in their research projects. AquaMine is powered by **InterMine**, an open-source data warehousing system that provides access to a wide variety of datasets.



Main site: <https://aquamine.rnet.missouri.edu/aquamine/begin.do>

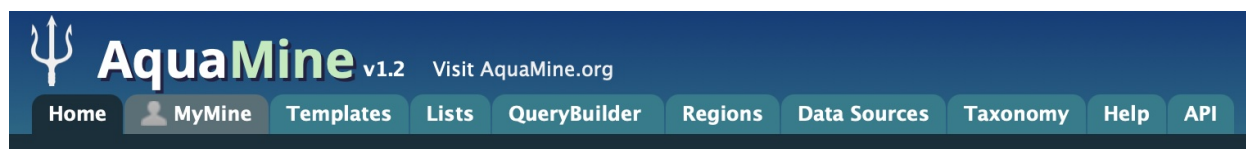
Link to the datasets used in AquaMine can be found here: <https://aquamine.rnet.missouri.edu/aquamine/dataCategories.do>

**AquaMine.org** is based upon work supported by Agriculture and Food Research Initiative Competitive grant no. 2021-67021-34254 from the USDA National Institute of Food and Agriculture. AquaMine is developed and hosted at the University of Missouri.



## OVERVIEW OF AQUAMINE

This section provides a brief overview of the layout for AquaMine.



The navigation panel highlights different functionalities of AquaMine.

**Home** - The home page for AquaMine

**MyMine** - The MyMine serves as a portal for account management. When logged in to AquaMine Users can access their saved templates, most recent queries and saved lists.

**Templates** - List of templates that users can select from based on the nature of their query. Each template is a predefined query with a simple form containing a description of what input is expected and the type of output that will be generated.

**Lists** - Allows users to upload lists of genes on which they can perform enrichment analyses and export the results. Users that log in to AquaMine can save their lists for future use.

**QueryBuilder** - A flexible interface that allows users to create their own custom query template while browsing the AquaMine data models. Queries can be exported in a variety of formats to share with other users.

**Regions** - The Genomic Region Search tool where users can enter a series of genomic coordinates, specify flanking regions and fetch all features that fall within the given interval. The result can be exported or saved as a list for further analyses.

**Data Sources** - Provides a summary of all the data loaded into AquaMine including their sources, associated publications and links to source sites.

**Taxonomy** - Taxonomic tree for all species in AquaMine (based on the NCBI taxonomy database), given as a reference for homologue queries.

**Help** - Links to the AquaMine help docs and tutorials

**API** - Describes the InterMine API that allows users to programmatically access AquaMine.

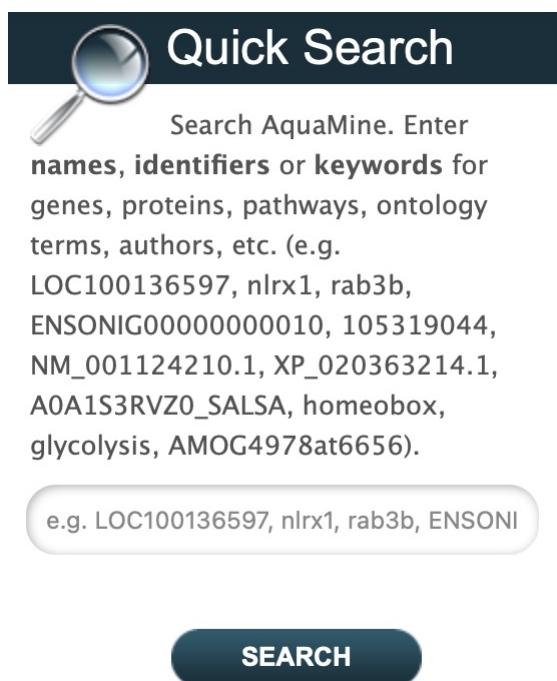


## NAVIGATION AND SEARCHING IN AQUAMINE

There are several ways for users to query AquaMine.

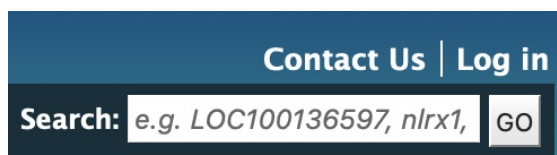
### 2.1 Quick Search

**Quick Search** allows users to search keywords from any of the AquaMine datasets. There is a Quick Search box on the AquaMine home page or a smaller search box in the upper right corner of all pages.



The image shows a 'Quick Search' interface. At the top, there is a dark blue header with a magnifying glass icon and the text 'Quick Search'. Below this, the text reads: 'Search AquaMine. Enter names, identifiers or keywords for genes, proteins, pathways, ontology terms, authors, etc. (e.g. LOC100136597, nlr1, rab3b, ENSONIG00000000010, 105319044, NM\_001124210.1, XP\_020363214.1, A0A1S3RVZ0\_SALSA, homeobox, glycolysis, AMOG4978at6656)'. Below this text is a light gray rounded rectangle containing the example text: 'e.g. LOC100136597, nlr1, rab3b, ENSONI'. At the bottom of the interface is a dark blue button with the word 'SEARCH' in white capital letters.

Fig. 1: Quick Search from home page



The image shows a smaller 'Quick Search' interface. It has a dark blue header with the text 'Contact Us | Log in' in white. Below this is a dark blue search bar with the text 'Search: e.g. LOC100136597, nlr1, GO' in white. The search bar is followed by a small white button with the word 'GO' in dark blue capital letters.

Fig. 2: Quick Search from any page

Quick Searches can be conducted with a number of identifiers including gene names, transcripts, pathways, gene identifiers or organisms. The wildcard character \* can be used to retrieve all results that match a particular search query.

For an example, we will use the Ensembl gene *ENSONIG00000000010*. Enter ENSONIG00000000010 in the search box and click **Search**. The results page is tabulated and displays a summary about your query, as shown below.

**Search our database by keyword**

ENSONIG00000000010

**Examples**

- Search this entire website. Enter identifiers, names or keywords for genes, pathways, authors, ontology terms, etc. (e.g. eve, embryo, zen, allele)
- Use OR to search for either of two terms (e.g. fly OR drosophila) or quotation marks to search for phrases (e.g. "dna binding").
- Boolean search syntax is supported: e.g. dros\* for partial matches or fly AND NOT embryo to exclude a term

**Search results 1 to 6 out of 6 for *ENSONIG00000000010***

**Categories**

**Hits by Category**

- Polypeptide: 4
- Gene: 2

**Hits by Organism**

- O. niloticus: 6

Type	Details	Score
Gene	<b>ENSONIG00000000010   hcls1</b> Source: Ensembl Biotype: protein_coding Description: src substrate cortactin [Source:NCBI gene;Acc:100700393] Length: 8449 <a href="#">FASTA</a> Chromosome: LG17: 14346614-14355062 Location: Organism: Oreochromis niloticus Assembly: O_niloticus_UMD_NMBU	*****
Gene	<b>100700393   LOC100700393</b> Source: RefSeq Biotype: protein_coding Description: src substrate cortactin Length: 9357 <a href="#">FASTA</a> Chromosome: LG17: 14346627-14355983 Location: Organism: Oreochromis niloticus Assembly: O_niloticus_UMD_NMBU	*****
Polypeptide	<b>ENSONIP00000000013</b> mRNA ID: ENSONIT00000000013 Gene ID: ENSONIG00000000010	*****
Polypeptide	<b>ENSONIP00000006465</b> mRNA ID: ENSONIT000000054600 Gene ID: ENSONIG00000000010	*****

Fig. 3: Results table for Quick Search with the gene *ENSONIG00000000010*

Hits are summarized in the box to the left of the results tables and can be filtered based on **Category** and **Organism**. Clicking on any of them will filter by the selected category. Note that for results with sequence data available, the sequence can be downloaded in FASTA format by clicking on the FASTA box within the hit Details box. The score column in the result table indicates the similarity of your query to each of the hits. The results page can also be converted to a list (and saved if users are logged in). To enable this feature click on **Gene** in the **Hits by Category**.

**Search our database by keyword**

ENSONIG00000000010

- or -

**Examples**

- Search this entire website. Enter identifiers, names or keywords for genes, pathways, authors, ontology terms, etc. (e.g. eve, embryo, zen, allele)
- Use OR to search for either of two terms (e.g. fly OR drosophila) or quotation marks to search for phrases (e.g. "dna binding").
- Boolean search syntax is supported: e.g. dros\* for partial matches or fly AND NOT embryo to exclude a term

**Search results 1 to 2 out of 2 for *ENSONIG00000000010***

Category restricted to Gene

**Categories**

Category: Gene

« show all

**Hits by Organism**

- O. niloticus: 2

Type	Details	Score
<input type="checkbox"/> Gene	<b>ENSONIG00000000010   hcls1</b> Source: Ensembl Biotype: protein_coding Description: src substrate cortactin [Source:NCBI gene;Acc:100700393] Length: 8449 <a href="#">FASTA</a> Chromosome: LG17: 14346614-14355062 Location: Organism: Oreochromis niloticus Assembly: O_niloticus_UMD_NMBU	*****
<input type="checkbox"/> Gene	<b>100700393   LOC100700393</b> Source: RefSeq Biotype: protein_coding Description: src substrate cortactin Length: 9357 <a href="#">FASTA</a> Chromosome: LG17: 14346627-14355983 Location: Organism: Oreochromis niloticus Assembly: O_niloticus_UMD_NMBU	*****

Fig. 4: Results table for Quick Search with the gene *ENSONIG00000000010* filtered by gene

After the table has been filtered for gene (and organism, if applicable), checkboxes will be available for users to select genes they would like to add to their list. Once the genes are selected, click on **CREATE LIST**. See the lists section

for more detail on creating and saving lists.

## 2.2 Templates

**Templates** or predefined queries are another search method within AquaMine. Popular templates are displayed on the home page, grouped by category (e.g., Genes, Proteins, Interactions) and the complete list can be seen by clicking the **Templates** menu tab.

GENES PROTEINS HOMOLOGY FUNCTION ENTIRE GENE SET ALIAS AND DBXREF

NCBI (RefSeq) genes are available for all AquaMine species. Check the Data Source page for the availability of Ensembl genes.

Query for genes:

- Gene → Transcripts + Exons
- Gene → Chromosomal location
- Chromosome → Genes
- Gene → Transcripts + Proteins
- Transcript → Protein
- Gene → Protein Sequences
- Transcript → Gene
- Gene → Publications

» [More queries](#)

popular templates

Fig. 5: Popular templates

**Templates**

Templates are predefined queries, each has a simple form and a description. Click on a template to run it, you can search for templates by keyword and filter them by category.

Filter:  Filter: -- all categories -- Reset

Actions: [Export selected](#) Options: ☒ Show descriptions ☐ Show Tags

You are not logged in. [Log in](#) to mark items as favourites .

- ☐ **Gene --> Transcripts + Exons**  
Given a gene id, retrieve transcript ids and locations and exon coordinates.
- ☐ **Organism --> All UniProt Proteins**  
Show all the UniProt proteins from a particular organism.
- ☐ **Orthologue Cluster ID --> Genes**  
Given an orthologue cluster ID from OrthoDB or AquaMine-Ortho, retrieve all genes in the cluster.
- ☐ **Gene --> Coding Sequences**  
Given a gene id, retrieve the coding sequences for report pages.
- ☐ **Gene --> Chromosomal location**  
Given a gene id, retrieve chromosomal location.
- ☐ **Gene Source and Organism --> Genes**  
Given a gene source (e.g. RefSeq or Ensembl), retrieve all gene ids. Use this to create a single-gene-set background list for enrichment.

Fig. 6: Full list of templates on Templates page

As an example, the **Gene Protein Sequences** template queries AquaMine to retrieve all homologues for a given gene. Here, we will do a search with the gene ID *110523811*.

**Gene → Protein Sequences**  
Given a gene id, retrieve protein sequences.

Gene > Gene ID

= 110523811

☐ constrain to be IN saved Gene list Crassostrea gigas RefSeq Genes

**Show Results** [Edit Query](#)

[web service URL](#) [Perl](#) | [Python](#) | [Ruby](#) | [Java \[help\]](#) [export XML](#)

Fig. 7: Example: Gene Protein Sequences

The results page displays all of the proteins for that query gene. When logged in to AquaMine, users can save their results as a list for further analyses by clicking on the **Save as List** button above the results table then choosing columns to save in their list. See the lists section for more detail on creating and saving lists. Note the “Trail: Query” text at the upper left of the results table. Clicking on the “Query” link will bring you back to the query that generated the table to allow for edits without having to start with a new template.

Trail: Query

**Gene → Protein Sequences**  
Given a gene id, retrieve protein sequences.

[Manage Columns](#) [Manage Filters](#) [Manage Relationships](#) [Save as List](#) [Generate Python code](#) [Export](#)

Showing 1 to 5 of 5 rows

Organism Name	Gene Gene ID	Polypeptides DB identifier	Polypeptides mRNA ID	Polypeptides Length	Polypeptides Is Representative	Sequence Residues
Oncorhynchus mykiss	110523811	XP_036832612.1	XM_036976717.1	1553	true	MPFAKRIVEPQLLCRHSPKESLVFEDLCTVNNVALSRTLRLQSLDLARHACSLFQELESDIVFTNQRV RGLQSKVGKLQQSISGLDAKQEAVPVSDLDVECKLSDHYVSPWLLQRNVFLPPTRPCLQELHRTAQ QGLRASHRDYQLRQQASSRRRKVTIAVSVAPPMPTYPSPHSARRRPQRGWSLTARQGKPFDDQQ...
Oncorhynchus mykiss	110523811	XP_036832666.1	XM_036976771.1	1538	false	MPFAKRIVEPQLLCRHSPKESLVFEDLCTVNNVALSRTLRLQSLDLARHACSLFQELESDIVFTNQRV RGLQSKVGKLQQSISGLDAKQEAVPVSDLDVECKLSDHYVSPWLLQRNVFLPPTRPCLQELHRTAQ QGLRASHRDYQLRQQASSRRRKVTIAVSVAPPMPTYPSPHSARRRPQRGWSLTARQGKPFDDQQ...
Oncorhynchus mykiss	110523811	XP_036832705.1	XM_036976810.1	1530	false	MPFAKRIVEPQLLCRHSPKESLVFEDLCTVNNVALSRTLRLQSLDLARHACSLFQELESDIVFTNQRV RGLQSKVGKLQQSISGLDAKQEAVPVSDLDVECKLSDHYVSPWLLQRNVFLPPTRPCLQELHRTAQ QGLRASHRDYQLRQQASSRRRKVTIAVSVAPPMPTYPSPHSARRRPQRGWSLTARQGKPFDDQQ...

Example: Gene Protein template search results.

## 2.2.1 Generate query code

The code for each template query can be retrieved by clicking on the arrow next to **Generate Python Code** and choosing the desired language from the pull-down menu. The language options are Python, Perl, Java, Ruby, JavaScript, and XML.



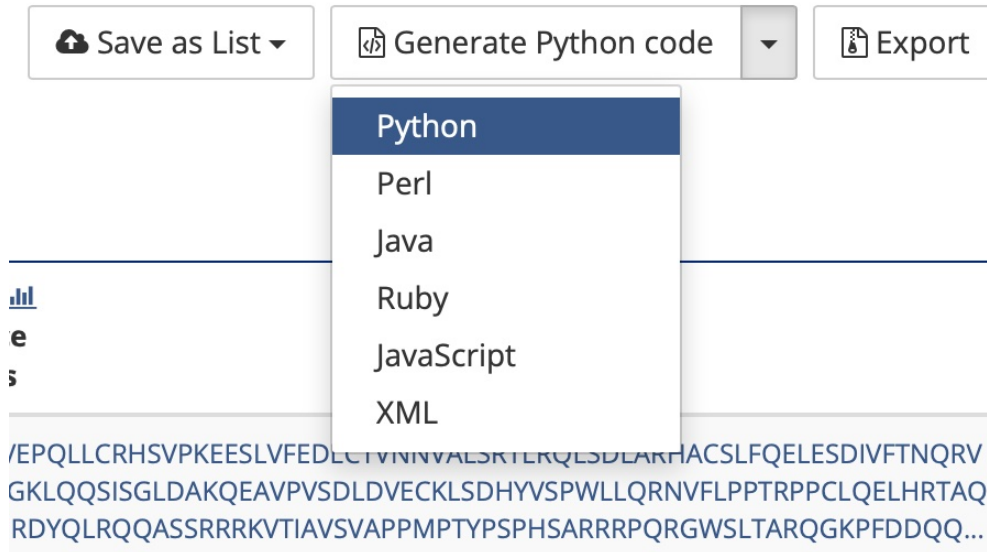


Fig. 8: Options for generating code from template query

## 2.2.2 Download results

The search results from a template query can be downloaded by clicking the **Export** button above the table and choosing the desired format from the pull-down menu to the right of the File name field. Available formats are tab-separated values, comma-separated values, XML, and JSON. When the results contain genomic features, they may also be downloaded in FASTA, GFF3, or BED format. Other options may be specified in the submenu to the left of the download box. By default, all rows and all columns are downloaded, but individual columns may be included or excluded by clicking on the toggles next to the column headers in the **All Columns** submenu. The number of rows and row offset are set in the **All Rows** submenu. Download the results as a compressed file by choosing GZIP or ZIP format in the **Compression** submenu (default is **No Compression**). Column headers are not added by default but may be included under the **Column Headers** submenu. Finally, the **Preview** submenu displays the first three rows of the file to be downloaded so that the desired format and options may be finalized before beginning the download. When ready, click the **Download file** button to download the results.

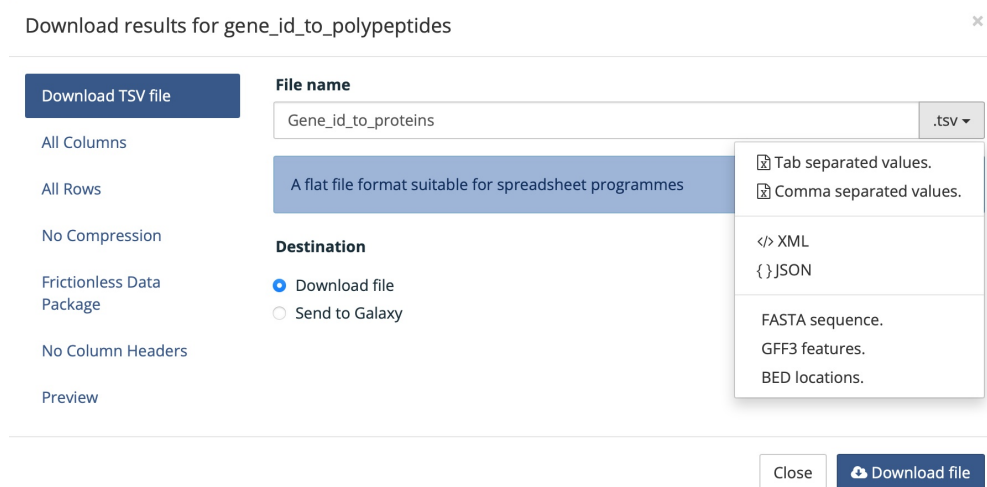


Fig. 9: Options for downloading results from template query

### 2.2.3 Customize output and manage columns

To customize the results table layout, click the **Manage Columns** button. This allows users to rearrange, remove or order columns. Filters can be edited by clicking the **Manage Filters** button. To specify the entity relationships within the query and change the way the results are presented in the table, click **Manage Relationships**. Clicking on the **Manage Relationships** option bring also brings up a blue information panel (“What does this do?”) that provides more detailed information.



Fig. 10: Options for customizing the results of a template query.

To further manage column data, each column has a set of icons in its header. Mousing over any icon will reveal what it does. The sort icon resembles two triangles and allows you to sort a column in either direction (ascending/descending; a->z, z->a). To delete any column from your table, click on the “x” icon. Alternatively, the three dots “...” can be selected to “Toggle column visibility” and hide the column. Once hidden, the icon turns into a double arrow that can be selected to expand the column into its original form. The next icon resembling a funnel can be selected to edit or remove any currently active filters. A summary of the data within a column can be viewed by clicking on the icon that resembles a graph. The data within each summary can be selected for further filtering or downloading.

Organism Name	Gene Gene ID	Polypeptides DB identifier	Polypeptides mRNA ID	Polypeptides Length	Polypeptides Is Representative	Sequence Residues

Fig. 11: Column header icons available to edit the results of a template query.

### 2.2.4 Optional filters

**\*\*If you are using a template with an optional filter and want to edit the template, the “Query” link at the top left of the results page will bring you back to the template that you used to create those results. Additionally, if you click on “Edit Query” in a template, the optional settings becomes non-optional and the default constraint is applied. To edit the template query, navigate back to the template by using the “Templates” tab on the home page menu bar.**

Some templates have optional filters that are disabled by default. For the **Gene ID Homologues** query, there is an optional filter to specify the Last Common Ancestor used in the query. To enable the filter, click **ON** next to the **Homologue > Last Common Ancestor** label to select an ancestor of choice.

**Gene ID → Homologues (OrthoDB or AquaMine-Ortho)**  
 Given a RefSeq gene id, retrieve homologues from OrthoDB or AquaMine-Ortho. Selection of a last common ancestor taxon is optional. You can turn it off if you do not know what to select, but making a selection will allow the query run faster.

**Gene > Gene ID**

=  ⓘ

☐ constrain to be  saved Gene list  ⓘ

**Homologue > Last Common Ancestor**

optional ☒ ON | OFF

=  ⓘ

[Show Results](#) [Edit Query](#)

[web service URL](#) [Perl](#) | [Python](#) | [Ruby](#) | [Java \[help\]](#) [export XML](#)

Fig. 12: Using the optional organism filter in the Gene ID Homologues query template.

## 2.3 QueryBuilder

The provided templates are suitable for many different types of searches, new queries may be built from scratch using the **QueryBuilder**. The possibilities of queries using the QueryBuilder are endless. The output may be formatted exactly as desired, and the query constraints may be chosen to perform complex search operations.

**QueryBuilder**  
 Advanced users can use a flexible query interface to construct their own data mining queries. The QueryBuilder lets you view the data model, apply constraints and select output. You can also export queries to share them with others.

[Browse data model](#)  
[Import query from XML](#)  
[Login to view saved queries](#)

**Select a Data Type to Begin a Query**  
 Click on a class name for a description or double click on a class name to create a new query starting at that class

Gene  
 Protein  
 -----  
 Annotatable  
 Author  
 Bio-Entity  
 CDS  
 C Gene Segment  
 Chromosome  
 Coding Sequence

[Select](#)

To begin, select a **Data Type**. For example, select **Gene** as a Data Type then click the **Select** button to be taken to the Model browser.

### 2.3.1 Model browser

After selecting a data type, the **Model builder** appears displaying the attributes for the chosen feature class **Gene**.

First choose **Gene** as a Data Type in the QueryBuilder. Then click on **Select**. This will take you to a Model browser where you can select the attributes for the feature class “Gene”, which you would want to be shown in your results.

## Select a Data Type to Begin a Query


Click on a class name for a description or double click on a class name to create a new query starting at that class

Gene  
Protein  
-----  
Annotatable  
Author  
Bio-Entity  
CDS  
C Gene Segment  
Chromosome  
Coding Sequence

Select

Gene: "Gene" is an abstract term used to describe a collection of transcripts and related regulatory features.

Fig. 13: Selecting data type as **Gene** in QueryBuilder.

**Model browser**


Browse through the classes and attributes. Click on [SUMMARY](#) links to add summary of fields to the results table or on [SHOW](#) links to add individual fields to the results. Use [CONSTRAIN](#) links to constrain a value in the query.

Gene [SUMMARY](#) [CONSTRAIN](#)

- Biotype [SHOW](#) [CONSTRAIN](#)
- Description [SHOW](#) [CONSTRAIN](#)
- Exception [SHOW](#) [CONSTRAIN](#)
- Length [SHOW](#) [CONSTRAIN](#)
- Integer [SHOW](#) [CONSTRAIN](#)
- Name [SHOW](#) [CONSTRAIN](#)
- Note [SHOW](#) [CONSTRAIN](#)
- Gene ID [SHOW](#) [CONSTRAIN](#)
- Source [SHOW](#) [CONSTRAIN](#)
- Symbol [SHOW](#) [CONSTRAIN](#)
- CDSs CDS [SUMMARY](#) [CONSTRAIN](#)
- Chromosome Chromosome [SUMMARY](#) [CONSTRAIN](#)
- Chromosome Location Location [SUMMARY](#) [CONSTRAIN](#)
- Coding Sequences Coding Sequence [SUMMARY](#) [CONSTRAIN](#)
- Cross References Cross Reference [SUMMARY](#) [CONSTRAIN](#)
- Data Sets Data Set [SUMMARY](#) [CONSTRAIN](#)
- Exons Exon [SUMMARY](#) [CONSTRAIN](#)
- Expression Values Expression [SUMMARY](#) [CONSTRAIN](#)
- GO Annotation GO Annotation [SUMMARY](#) [CONSTRAIN](#)
- Homologues Homologue [SUMMARY](#) [CONSTRAIN](#)

☐ Show empty fields

**Fields selected for output**

**Query Overview**

no fields constrained

Constraint logic:

no constraints

Fig. 14: Model browser with “Gene” selected as the data type.

## 2.3.2 QueryBuilder Examples

The following three examples provide details as to how to use the QueryBuilder using “Gene” as the selected data type.

### 2.3.2.1 Example 1: Querying for protein coding genes

In the Model browser, click **Show** next to **Biotype**, **Gene ID** and **Symbol**, which will add these fields to the query. Notice that these two fields appear below the data type **Gene** in the Query Overview section.

The screenshot displays the QueryBuilder interface. On the left, the **Model browser** shows a tree structure of classes and attributes. The **Gene** class is expanded, showing attributes like **Biotype**, **Description**, **Exception**, **Length**, **Name**, **Note**, **Gene ID**, **Source**, and **Symbol**. Each attribute has a **SHOW** button and a **CONSTRAIN** button. Below the tree, there is a checkbox for **Show empty fields** and a section for **Fields selected for output**. On the right, the **Query Overview** section shows the **Gene** class with its selected attributes: **Biotype**, **Gene ID**, and **Symbol**. Below this, the **Constraint logic** section shows **no constraints**. At the bottom, the **Columns to Display** section shows the selected fields: **Gene > Biotype**, **Gene > Gene ID**, and **Gene > Symbol**, each with a red 'X' icon and a blue pencil icon.

Fig. 15: Example 1, Step 1: Select fields to be added to the query

Then click **Constrain** next to **Biotype**. The first drop-down menu defaults to “=” (equals sign). In the second drop-down menu, select **protein\_coding**, then click the **Add to query** button. This adds a constraint to the query to search only for protein coding genes. Notice that the Query Overview section now shows “Biotype = protein\_coding”. Also, two types of icons appear next to the attributes. Clicking on the red “X” icon next to an attribute will remove that field or constraint from the query. Clicking on the blue pencil icon next to a constraint brings up the constraint editing window where changes may be made to the query filters.

Lastly, click on **Show Results** above the Model Browser. The resulting table contains all protein coding genes in the database, with Gene ID, Gene Symbol and Gene Biotype as the table columns. Because the Biotypes should all be the same (protein coding), that column can be deleted by clicking the “x” above it. Alternatively, the three dots “...” can be selected to “Toggle column visibility” and hide the column. See the [Customize output and manage columns](#) section to review ways to manage column data. Clicking on “Query” in the “Trail: Query > Results” link at the upper left of the results table will bring you back to the query that generated the table to allow for edits without having to build a new query. For example, the column order can be changed by sliding the blue squares under the “Columns to Display” section at the bottom of the page.

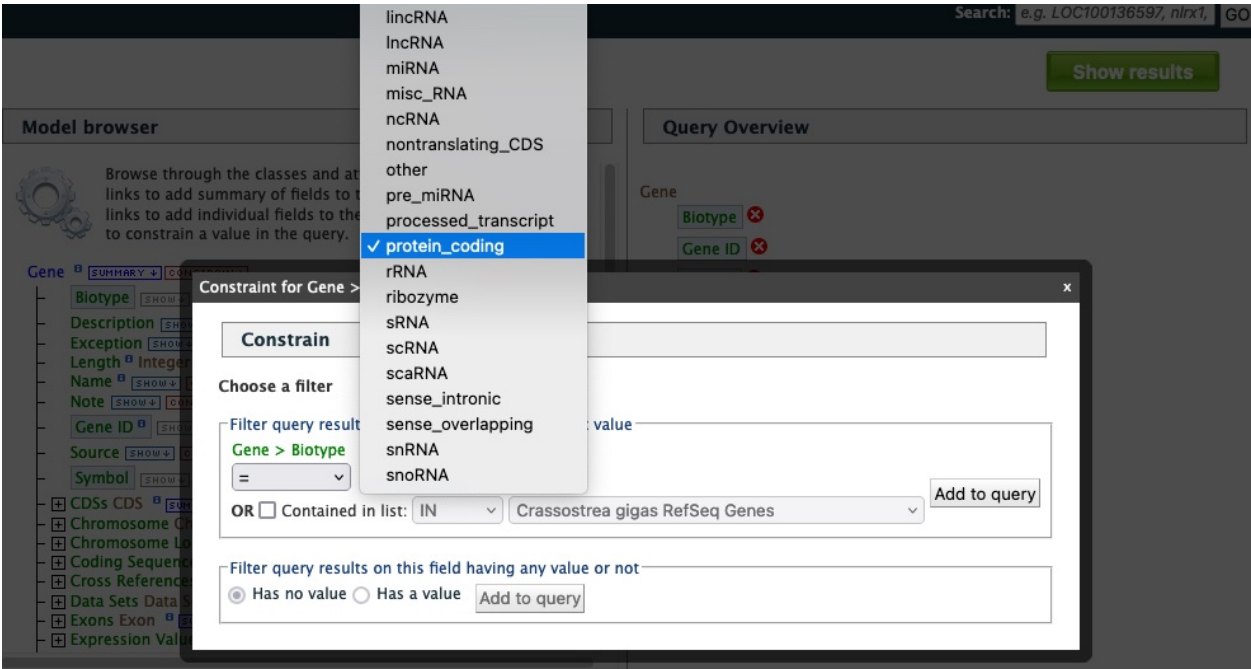


Fig. 16: Example 1, Step 2: Add a constraint to the query on Biotype.

Trail: Query > Results

☐ Manage Columns ☐ Manage Filters ☐ Manage Relationships

Showing 1 to 25 of 1,612,878 rows

Rows per page: 25

Gene Biotype	Gene Gene ID	Gene Symbol
protein_coding	1	A1BG
protein_coding	10	NAT2
protein_coding	100	ADA
protein_coding	1000	CDH2
protein_coding	10000	AKT3
protein_coding	100000006	slc12a8
protein_coding	100000009	sema5bb
protein_coding	100000024	sidkey-33c12.12
protein_coding	100000024-2	sidkey-33c12.12-2
protein_coding	100000026	aldh3a1
protein_coding	100000030	btr24
protein_coding	100000044	sidkey-100n10.2
protein_coding	100000058	zmp0000001102
protein_coding	100000061	slch211-133n4.9
protein_coding	100000073	onecut2
protein_coding	100000079	soul5

Fig. 17: Example 1, Step 3: Display query results.



### 2.3.2.2 Example 2: Querying for protein coding genes on a particular chromosome

Users can customize the previously run query by adding another constraint for **Chromosome**. On every query results page, note the “Trail: Query” text at the upper left of the results table. Clicking on the “Query” link will bring you back to the query that generated the table to allow for edits without having to build a new query. In the Model browser where you began Example 1, click on the “+” (plus sign) next to the Chromosome feature class to display its attributes.

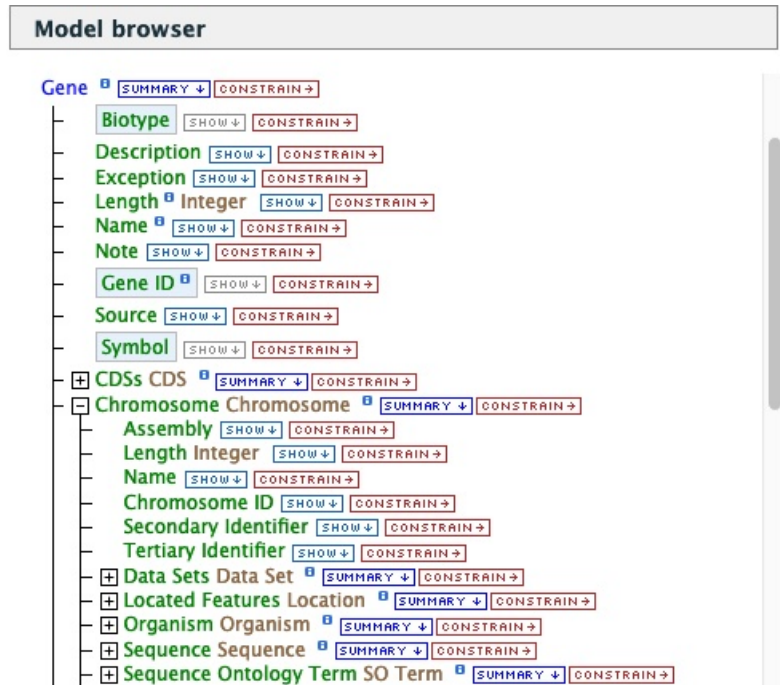


Fig. 18: Example 2, Step 1: View attributes of Chromosome feature class

Next click on **Constrain** next to the **Chromosome ID** attribute and in the text box of the pop-up window select “=” (equals). The chromosome ID format is specific to an organism. For this example, we will use the chromosome IDs for *Esox lucius*, the Northern pike. Enter **LG14** for chromosome number 14 then click on **Add to Query**, which adds the additional constraint to the query. We will add another constraint for the organism. Click on the **Organism** to reveal its attributes then click on **constrain** next to Name. In the resulting pop-up window, select “=” (equals) then “*Esox lucius*”.

Click on **Show results** and the query will result in all *Esox lucius* protein coding genes on the LG14 Chromosome. Note that the number of results has been reduced with the addition of the constraints.

### 2.3.2.3 Example 3: Querying for Protein Coding genes on a particular chromosome and their exons

This last example builds upon the previous queries to display all exons for each gene on a particular chromosome. Again, note the “Trail: Query” text at the upper left of the results table. Clicking on the “Query” link will bring you back to the query that generated the table to allow for edits without having to build a new query. Scroll down to **Exon** feature class and select the “+” (plus sign) next to Exons to display its attributes. Click **Show** next to **Length** and **DB Identifier**.

The Query Overview shows the query in progress with the selected fields. Also notice that a third type of icon, a blue square, appears next to some attributes. Clicking on a blue square icon brings up a window where the query Join Style may be modified. When adding a constraint, you can decide whether you want to show only those results with the information (genes with exons) or all results and the constrained feature if it exists (e.g., all genes and indicate exons if they exist). Click on the blue square icon next to **Exon collection** to bring up the Switch Join Style window. The

### Model browser

Browse through the classes and attributes. Click on [SUMMARY](#) links to add summary of fields to the results table or on [SHOW](#) links to add individual fields to the results. Use [CONSTRAIN](#) links to constrain a value in the query.

Gene [SUMMARY](#) [CONSTRAIN](#)

- Biotype [SHOW](#) [CONSTRAIN](#)
- Description [SHOW](#) [CONSTRAIN](#)
- Exception [SHOW](#) [CONSTRAIN](#)
- Length Integer [SHOW](#) [CONSTRAIN](#)
- Name [SHOW](#) [CONSTRAIN](#)
- Note [SHOW](#) [CONSTRAIN](#)
- Gene ID [SHOW](#) [CONSTRAIN](#)
- Source [SHOW](#) [CONSTRAIN](#)
- Symbol [SHOW](#) [CONSTRAIN](#)
- CDSs CDS [SUMMARY](#) [CONSTRAIN](#)
- Chromosome Chromosome [SUMMARY](#) [CONSTRAIN](#)
- Assembly [SHOW](#) [CONSTRAIN](#)
- Length Integer [SHOW](#) [CONSTRAIN](#)
- Name [SHOW](#) [CONSTRAIN](#)
- Chromosome ID [SHOW](#) [CONSTRAIN](#)
- Secondary Identifier [SHOW](#) [CONSTRAIN](#)
- Tertiary Identifier [SHOW](#) [CONSTRAIN](#)

☐ Show empty fields

### Query Overview

Gene

- Biotype [X](#)
- = protein\_coding [X](#) [C](#)
- Gene ID [X](#)
- Symbol [X](#)
- Chromosome Chromosome [X](#) [B](#)
- Chromosome ID [X](#)
- = LG14 [X](#) [A](#)
- Organism Organism [X](#) [B](#)
- Name [X](#)
- = Esox lucius [X](#) [B](#)

Constraint logic: A and B and C

A and B and C

### Fields selected for output

### Columns to Display

Use the [SHOW](#) or [SUMMARY](#) links to add fields to the results table. Click and drag the blue output boxes to choose the output column order. Click [\[1\]](#) to choose a column to sort results by, click again to select ascending [\[1\]](#) or descending [\[2\]](#). Use the [REMOVE ALL](#) link to remove all fields from the results table.

[REMOVE ALL](#)

Gene > Biotype [X](#)  
(no description) [\[1\]](#)

Gene > Gene ID [X](#)  
(no description) [\[2\]](#)

Gene > Symbol [X](#)  
(no description) [\[1\]](#)

Gene > Chromosome ID [X](#)  
(no description) [\[2\]](#)

Gene > Organism [X](#)  
(no description) [\[1\]](#)

[Show results](#)

Fig. 19: Example 2, Step 2: Add constraints to the Chromosome ID and Organism

Trail: [Query > Results](#)

☐ Manage Columns
☐ Manage Filters
☐ Manage Relationships

[Save as List](#)
[Generate Python code](#)
[Export](#)

Showing 1 to 25 of 2,178 rows

Rows per page: 25

H-
←
→
H+

page 1

←
→

Gene Biotype	Gene Gene ID	Gene Symbol	Gene Chromosome ID	Gene Organism
protein_coding	105005609	LOC105005609	LG14	Esox lucius
protein_coding	105005611	LOC105005611	LG14	Esox lucius
protein_coding	105005612	LOC105005612	LG14	Esox lucius
protein_coding	105005613	tmem178b	LG14	Esox lucius
protein_coding	105005663	sec31a	LG14	Esox lucius
protein_coding	105005665	lin54	LG14	Esox lucius
protein_coding	105005667	LOC105005667	LG14	Esox lucius
protein_coding	105005668	cops4	LG14	Esox lucius
protein_coding	105005669	plac8	LG14	Esox lucius
protein_coding	105005670	adams6	LG14	Esox lucius
protein_coding	105005691	LOC105005691	LG14	Esox lucius
protein_coding	105005693	LOC105005693	LG14	Esox lucius
protein_coding	105005694	LOC105005694	LG14	Esox lucius
protein_coding	105005695	LOC105005695	LG14	Esox lucius
protein_coding	105005697	LOC105005697	LG14	Esox lucius

Fig. 20: Example 2, Step 3: Result table after constraining by Chromosome ID and Organism



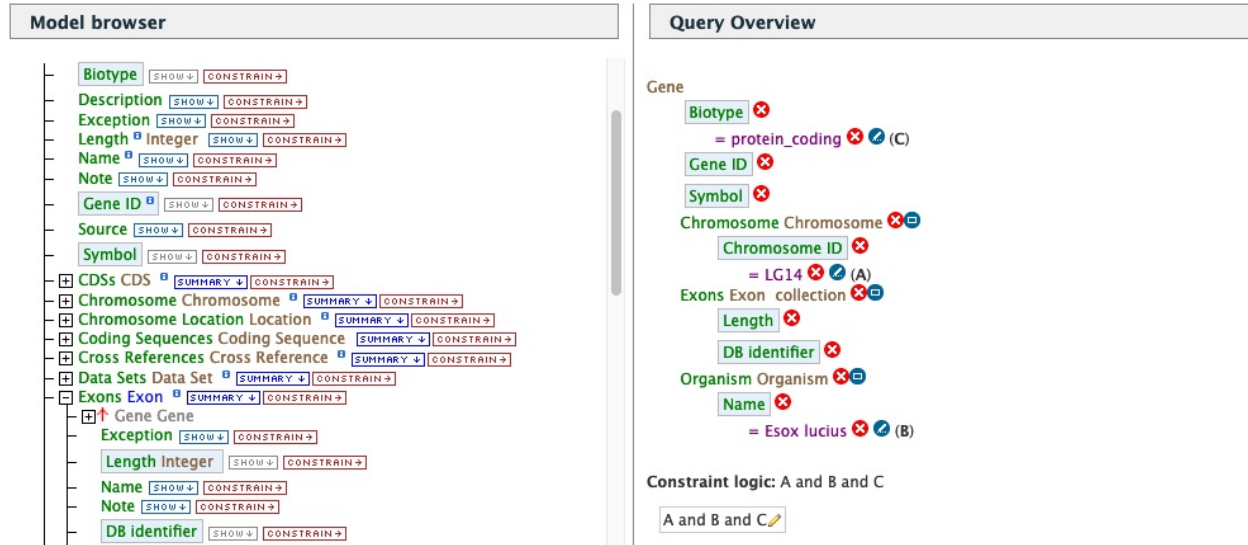


Fig. 21: Example 3, Step 1: Expand exon attributes and add fields to the query

default option is to show only Genes if they have a exon (inner join). Change this to **Show all Genes and show Exons if they are present** (outer join) then click **Add to query**.

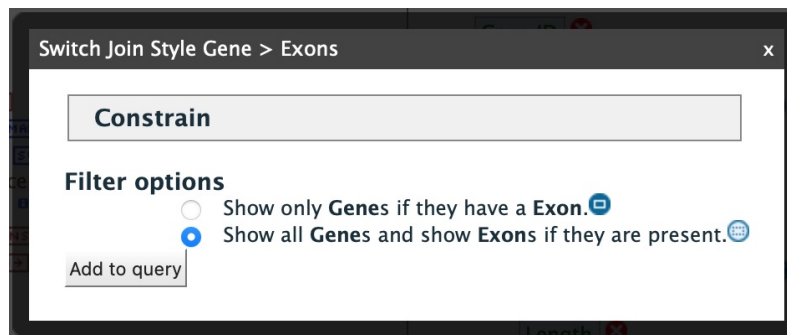


Fig. 22: Example 3, Step 2: Select join style for exons

Then click “Show results” to run the new query.

The results table now lists a new column **Gene Exons**, which we added to the query. If we look at the third row in the table, it lists 14 exons. Click on the **14 exons** link to expand that entry. That column now has additional rows containing the **Exon identifier** and **Length** for each of the 14 exons.

In changing the join style to an outer join, the exons have been grouped together by gene making it easier to determine how many exons are contained in each gene. If the same query is run with the default join (outer join) of **Show only Genes if they have an Exon**, the results table adds a new row for each new exon rather than grouping exons by gene. To change the join, click on the blue square icon next to **Exon collection** to bring up the Switch Join Style window as done above.

Trail: Query > Results

Manage Columns

Manage Filters

Manage Relationships

Save as List

Generate Python code

Export

Showing 1 to 25 of 2,178 rows

Rows per page: 25

Gene Biotype	Gene ID	Gene Symbol	Gene Chromosome ID	Gene Organism	Gene Exons
protein_coding	105005609	LOC105005609	LG14	Esos lucius	21 Exons
protein_coding	105005611	LOC105005611	LG14	Esos lucius	5 Exons
protein_coding	105005612	LOC105005612	LG14	Esos lucius	14 Exons
protein_coding	105005613	tmem178b	LG14	Esos lucius	5 Exons
protein_coding	105005663	sec31a	LG14	Esos lucius	346 Exons
protein_coding	105005665	lin54	LG14	Esos lucius	56 Exons
protein_coding	105005667	LOC105005667	LG14	Esos lucius	4 Exons
protein_coding	105005668	cops4	LG14	Esos lucius	19 Exons
protein_coding	105005669	plac8	LG14	Esos lucius	10 Exons
protein_coding	105005670	adams6	LG14	Esos lucius	77 Exons
protein_coding	105005691	LOC105005691	LG14	Esos lucius	34 Exons
protein_coding	105005693	LOC105005693	LG14	Esos lucius	204 Exons
protein_coding	105005694	LOC105005694	LG14	Esos lucius	13 Exons
protein_coding	105005695	LOC105005695	LG14	Esos lucius	8 Exons
protein_coding	105005697	LOC105005697	LG14	Esos lucius	58 Exons

Fig. 23: Example 3, Step 3: Query results with exon constraint

Trail: Query > Results

Manage Columns

Manage Filters

Manage Relationships

Save as List

Generate Python code

Export

Showing 1 to 25 of 2,178 rows

Rows per page: 25

Gene Biotype	Gene ID	Gene Symbol	Gene Chromosome ID	Gene Organism	Gene Exons
protein_coding	105005609	LOC105005609	LG14	Esos lucius	21 Exons
protein_coding	105005611	LOC105005611	LG14	Esos lucius	5 Exons
protein_coding	105005612	LOC105005612	LG14	Esos lucius	14 Exons
protein_coding	105005613	tmem178b	LG14	Esos lucius	<div><div>Length</div><div>DB identifier</div></div>
					256 exon-XM_013137303.3-1
					242 exon-XM_013137303.3-2
					84 exon-XM_013137303.3-3
					114 exon-XM_013137303.3-4
					87 exon-XM_013137303.3-5
					102 exon-XM_013137303.3-6
					126 exon-XM_013137303.3-7
					122 exon-XM_013137303.3-8
					213 exon-XM_013137303.3-9
					91 exon-XM_013137303.3-10
					265 exon-XM_013137303.3-11
					654 exon-XM_013137303.3-12
					248 exon-XM_013137303.3-13
					2738 exon-XM_013137303.3-14
protein_coding	105005613	tmem178b	LG14	Esos lucius	5 Exons

Fig. 24: Example 3, Step 4: Query results with exon column expanded

Trail: Query > Results

Manage Columns

Manage Filters

Manage Relationships

Save as List

Generate Python code

Export

Showing 1 to 25 of 83,546 rows

Rows per page: 25

Gene Biotype	Gene Gene ID	Gene Gene Symbol	Gene Chromosome ID	Gene Organism	Exons Length	Exons DB identifier
protein_coding	105005609	LOC105005609	LG14	Esox lucius	17	exon-XM_010863666.4-4
protein_coding	105005609	LOC105005609	LG14	Esox lucius	99	exon-XM_010863668.4-2
protein_coding	105005609	LOC105005609	LG14	Esox lucius	127	exon-XM_010863665.4-3
protein_coding	105005609	LOC105005609	LG14	Esox lucius	127	exon-XM_010863666.4-3
protein_coding	105005609	LOC105005609	LG14	Esox lucius	127	exon-XM_010863667.4-3
protein_coding	105005609	LOC105005609	LG14	Esox lucius	127	exon-XM_010863668.4-3
protein_coding	105005609	LOC105005609	LG14	Esox lucius	127	exon-XR_001198446.2-3
protein_coding	105005609	LOC105005609	LG14	Esox lucius	324	exon-XM_010863667.4-2
protein_coding	105005609	LOC105005609	LG14	Esox lucius	334	exon-XR_001198446.2-2
protein_coding	105005609	LOC105005609	LG14	Esox lucius	372	exon-XM_010863665.4-2
protein_coding	105005609	LOC105005609	LG14	Esox lucius	372	exon-XM_010863666.4-2
protein_coding	105005609	LOC105005609	LG14	Esox lucius	510	exon-XM_010863668.4-1

Fig. 25: Example 3, Step 5: Query results with default join style run for genes containing exons



## REPORT PAGES

All objects in AquaMine (e.g., gene, protein, transcript, publication) have report pages that can viewed after running a query. It allows users to view all available information for that object while providing links to related objects. As an example, we can revisit the **Gene Protein Sequences** template. In the list of templates under the **Templates** tab on the AquaMine home page, select **Gene Protein Sequences** to query AquaMine to retrieve all protein sequences for a given gene. Enter the Gene ID “100135786” into the the search box then click **Show Results**. In the results table, note that every entry is contains a link. You can mouse over any link to bring up a summary of that object. If we hover over the first Gene ID, we can see a summary box that includes information about that particular gene.

Trail: Query  
**Gene** ➔ **Protein Sequences**  
*Given a gene id, retrieve protein sequences.*

☐ Manage Columns ☐ Manage Filters ☐ Manage Relationships

Showing 1 to 3 of 3 rows

Gene Organism	Gene Gene ID
Oncorhynchus mykiss	<a href="#">100135786</a>
Oncorhynchus mykiss	100135786
Oncorhynchus mykiss	100135786

**Gene**

Biotype	protein_coding
Chromosome	USDA_OmykA_1.1
Description	nuclear pore complex glycoprotein p62
Gene ID	100135786
Length	11,774
Name	LOC100135786
Organism	Oncorhynchus mykiss
Source	RefSeq
Symbol	LOC100135786

Fig. 1: Summary for gene entry in query results table

Clicking on that same link will bring up its report page that includes a comprehensive overview for that gene. The report page header shows the Gene ID and its Biotype, for this example, protein coding. The tabs at the top of the page in the Quick Links menu bar quickly bring you to the data listed. The column on the right side of the report page displays external links to other Mines and databases.

The content of the report page is divided into categories based on the type of information provided for that particular object. Clicking on links within each category bring up more details about the objects of interest.

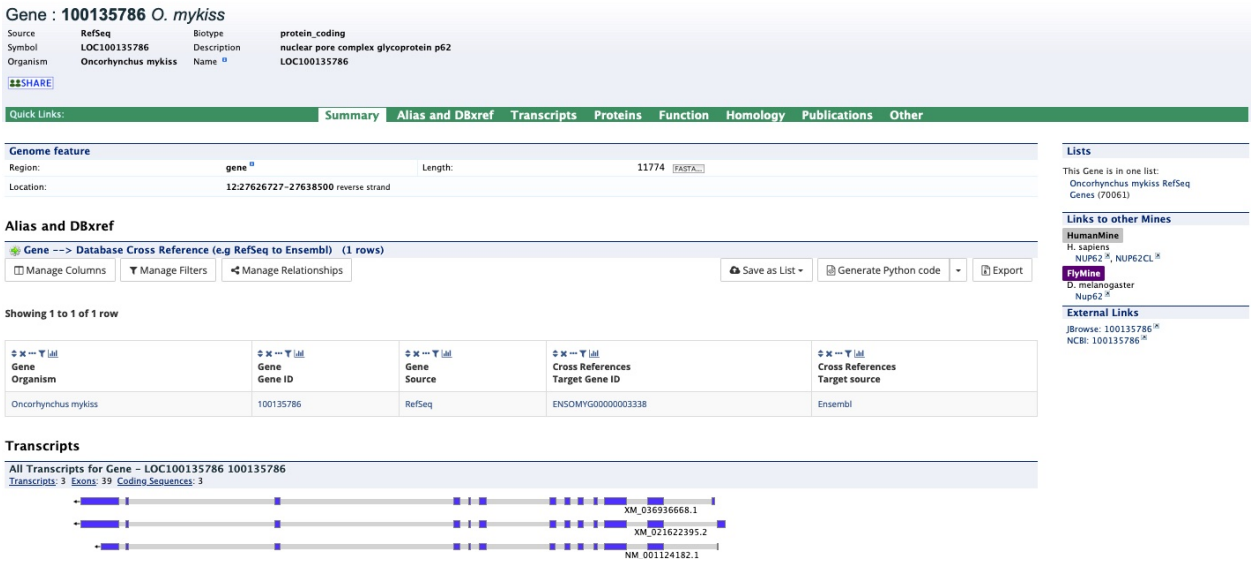


Fig. 2: Report page for protein coding gene

3.1 Summary

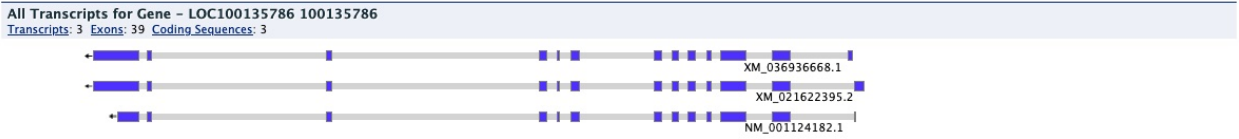
The **Summary** section near the top of the report provides information on the gene such as its length, chromosome location, and strand information. Users can also get the complete FASTA sequence of the gene by clicking on the FASTA tab.

Genome feature			
Region:	gene	Length:	11774 FASTA...
Location:	12:27626727-27638500 reverse strand		

3.2 Transcripts

The **Transcripts** section contains information about the gene model, such as transcripts and exons. Links to FASTA files are included where applicable.

Transcripts



Gene --> Coding Sequences (3 rows)

Manage Columns

Manage Filters

Manage Relationships

Save as List

Generate Python code

Export

Showing 1 to 3 of 3 rows

<div>Transcript DB identifier</div>	<div>Coding Sequences Is Representative</div>	<div>Sequence Residues</div>
NM_001124182.1	false	ATGAGTGGAGGATTCAACTTTGGACAAGCGTCCAGTACTGGATTGAGCTTTGGAGCTCCAAAACCAAGCTGCAACTGCCCCAGCCAGGCTTTGGGATGCCTAGTGCTGCAGCTCCAGGTGGTGGCTTCTCTTTCCGAACCCCAAGCACTCCCAAGGCCCAAGTAGCTGTCCCAAGAGCTCTGGGCTGTTGCTAT...
XM_021622395.2	true	ATGAGTGGAGGATTCAACTTTGGACAAGCGTCCAGTACTGGATTGAGCTTTGGAGCTCCAAAACCAAGCTGCAACTGCCCCAGCCAGGCTTTGGGATGCCTAGTGCTGCAGCTCCAGGTGGTGGCTTCTCTTTCCGAACCCCAAGCACTCCCAAGGCCCAAGTAGCTGTCCCAAGAGCTCTGGGCTGTTGCTAT...
XM_036936668.1	false	ATGAGTGGAGGATTCAACTTTGGACAAGCGTCCAGTACTGGATTGAGCTTTGGAGCTCCAAAACCAAGCTGCAACTGCCCCAGCCAGGCTTTGGGATGCCTAGTGCTGCAGCTCCAGGTGGTGGCTTCTCTTTCCGAACCCCAAGCACTCCCAAGGCCCAAGTAGCTGTCCCAAGAGCTCTGGGCTGTTGCTAT...

Gene --> Protein Sequences (1 rows)

Manage Columns

Manage Filters

Manage Relationships

Save as List

Generate Python code

Export

Showing 1 to 1 of 1 row

<div>Polypeptides DB identifier</div>	<div>Polypeptides mRNA ID</div>	<div>Polypeptides Length</div>	<div>Sequence Residues</div>
XP_021478070.2	XM_021622395.2	583	MSGGFNFGQASSTGFSFGAPKTTAATAPATGFGMPSAAPGGGFSFGTPTQPAPVAVPQSSGLFAMPTQGSTNPTGGFSFGTQAQSSPTAGGGFSFGTPIAKLNLGTPVASQPALTLTGMGATPSSGTGFLGGGLATQTAAAGGGFSFGTAGGLGAPVAQTQPQTQPAAGGLSLGTPAEATQIMGGGFSFGAAK...

3.3 Proteins

The **Proteins** section provides information about the protein product of the gene. The comments section gives a brief description about the protein along with the UniProt accession and links to any outside data sets.

Proteins

Curated comments from UniProt		Show proteins
Type	Comment	
similarity	Belongs to the nucleoporin NSP1/NUP62 family.	

1 Proteins

Manage Columns

Manage Filters

Manage Relationships

Save as List

Generate Python code

Export

Showing 1 to 1 of 1 row

<div>Proteins DB identifier</div>	<div>Proteins Primary Accession</div>	<div>Proteins Organism</div>	<div>Proteins Length</div>
O57397_ONCMY	O57397	Oncorhynchus mykiss	579

### 3.4 Function

The **Function** section displays Gene Ontology annotations for a gene. Annotations are divided into three categories:



- Cellular Component
- Molecular Function
- Biological Process



The GO terms are displayed along with the evidence code indicating how the annotations were derived. A results with Pathway information is also displayed if applicable.



**Function**



**Gene Ontology**



**cellular component**



nucleocytoplasmic transport  IEA 



nucleus  IEA 

nuclear envelope  IEA 



nuclear transport  IEA 



structural constituent of nuclear pore  IEA  ECO:0007669

endomembrane system  IEA 



nuclear pore  IEA  ECO:0007669



**molecular function**



structural constituent of nuclear pore  IEA  ECO:0007669



leucine zipper domain binding  ECO:0000315 IMP 

**biological process**


nucleocytoplasmic transport  IEA 


protein transport  IEA  ECO:0007669


nuclear transport  IEA 


structural constituent of nuclear pore  IEA  ECO:0007669


**1 Pathways**


 Manage Columns

 Manage Filters





 Manage Relationships

 Save as List

 Generate Python code

 Export

Showing 1 to 1 of 1 row

 Pathways Identifier	 Pathways Name	 Pathways URL
omy03013	Nucleocytoplasmic transport - Oncorhynchus mykiss (rainbow trout)	 <a href="https://www.genome.jp/pathway/omy03013">https://www.genome.jp/pathway/omy03013</a>

### 3.5 Homology

The **Homology** section provides information for all homologues and displays a summarized view of the homologues reported in different organisms.



Homology

Homologues	
A. mexicanus	nup62l
C. clupearformis	LOC121579955 LOC121573705
C. gigas	LOC105325291
C. harengus	nup62l
C. virginica	LOC111127431
D. melanogaster	Nup62
D. rerio	nup62l
E. cragini	nup62l
E. lucius	LOC105026973
G. aculeatus aculeatus	nup62l
G. morhua	nup62
H. americanus	LOC121878137

Fig. 3: Homologues

Orthologue Clusters

Manage Columns

Manage Filters

Manage Relationships

Save as List

Generate Python code

Export

Showing 1 to 10 of 10 rows

<div><div>Orthologue Clusters</div><div>Cluster ID</div></div>	<div><div>Orthologue Clusters</div><div>Last Common Ancestor</div></div>
1588739at7742	Vertebrata
37232at7898	Actinopterygii
46333at33208	Metazoa
AMOG20556at1489388	Euteleosteiomorpha
AMOG22019at41705	Protacanthopterygii
AMOG22374at8015	Salmonidae
AMOG26644at33511	Deuterostomia
AMOG26719at7898	Actinopterygii
AMOG27151at32443	Teleostei
AMOG33412at33213	Bilateria

Gene --> Reciprocal Best Hits (8 rows)

Manage Columns

Manage Filters

Manage Relationships

Save as List

Generate Python code

Export

Showing 1 to 8 of 8 rows

<div><div>Gene</div><div>Gene ID</div></div>	<div><div>Gene</div><div>Organism</div></div>	<div><div>Reciprocal Best Hit</div><div>Gene ID</div></div>	<div><div>Reciprocal Best Hit</div><div>Organism</div></div>
100135786	Oncorhynchus mykiss	100196770	Salmo salar
100135786	Oncorhynchus mykiss	109892610	Oncorhynchus kisutch
100135786	Oncorhynchus mykiss	111981177	Salvelinus sp. IW2-2015
100135786	Oncorhynchus mykiss	112257891	Oncorhynchus tshawytscha
100135786	Oncorhynchus mykiss	115105716	Oncorhynchus nerka
100135786	Oncorhynchus mykiss	118391303	Oncorhynchus keta
100135786	Oncorhynchus mykiss	121573705	Coregonus clupeaformis
100135786	Oncorhynchus mykiss	123999511	Oncorhynchus gorbuscha

Fig. 4: Orthologue clusters (from OrthoDB/AquaMine-Ortho) and reciprocal best hits

### 3.6 Publications

The **Publications** section displays a table of publications related to the gene with links to full citations.

**Publications**

2 Publications

Manage Columns

Manage Filters

Manage Relationships

Save as List

Generate Python code

Export

Showing 1 to 2 of 2 rows

<div>Publications</div> <div>First Author</div>	<div>Publications</div> <div>Title</div>	<div>Publications</div> <div>Year</div>	<div>Publications</div> <div>Journal</div>	<div>Publications</div> <div>Volume</div>	<div>Publications</div> <div>Pages</div>	<div>Publications</div> <div>PubMed ID</div>
Pasquier Jeremy	Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database.	2016	BMC Genomics	17	368	27189481
Yamashita A	cDNA cloning of a novel rainbow trout SRY-type HMG box protein, rtSox23, and its functional analysis.	1998	Gene	209	193-200	9524264

### 3.7 Other

This last section provides miscellaneous information that do not fit into any of the above categories. This example lists other data sets that the gene belongs to.

**Other**

11 Data Sets

Oncorhynchus mykiss NCBI RefSeq gene set for USDA\_OmyKA\_1.1, AquaMine-Ortho data set, PubMed to gene mapping, AquaMine GO annotation data set, Gene ID Cross References (Ensembl ↔ RefSeq) data set, TrEMBL data set, KEGG pathways data set, OrthoDB data set, Oncorhynchus mykiss RefSeq Coding Sequences, Oncorhynchus mykiss RefSeq Protein Sequences, AquaMine reciprocal best hits data set

1 Synonyms

LOC100135786

## GENOMIC REGIONS SEARCH

The **Genomic Regions Search** is a tool to fetch features that are within a given set of genomic coordinates or are within a given number of bases flanking the coordinates.

To begin this type of search, click the **Regions** tab on the menu bar. A form will appear asking for the search parameters (organism, feature types, genomic coordinates, etc.)

The coordinates must have one of three formats:

1. `chromosome_number:start..end`
2. `chromosome_number:start-end`
3. `chromosome_number start end` (tab delimited)

Click on the input examples above the text input box (number 4) to view a representative set of coordinates in each format. Click the **More genome coordinates help** link near the top of the form for more detailed information on the input format requirements.

During a search, regions may be extended on either side of the genomic coordinates using the slider or by entering text in the field to the left of the slide bar. There is also the option to perform a strand-specific region search using the checkbox at the bottom of the form (number 6).

As an example, select *Ictalurus punctatus* from the **Select Organism** drop-down, and notice that the associated Assembly appears in the next drop-down menu, *IpCoco\_1.2*. Under the **Select Feature Types** section, check the box next to **Gene**, and enter the following coordinates into the genomic regions search text field:

```
14:60000000..68000000
```

Click the **Search** box to conduct the genomic regions search. If there are no overlaps within your search coordinates, the search can be done again with the search region extended using the slide bar or entering text into the search box (e.g., 10k).

The search results page presents a list of features present within the genomic interval that was searched. In this case, the feature type was limited to Gene. The results may be exported as tab-separated or comma-separated values. If they contain genomic features, there is also the option to save the results in GFF3 or BED format. The FASTA sequences of the features may also be downloaded. Links within the features provide detailed reports. If users are interested in creating a list of particular features from the result page then they can filter based on feature type (if applicable), shown in red box, and click on **Go**.

## Search for features within Genomic Regions

Search for features that overlap a list of genome coordinates you enter or upload, e.g. 6:50000..100000. The primary Chromosome IDs used in AquaMine are the "Assigned-Molecule" from the NCBI assembly report. You can retrieve all cross-reference identifiers using the template query **Organism Name → Chromosome IDs** found on the AquaMine home page under "ENTIRE GENE SET". Most species use numbers for autosomes. The mitochondrial, X and Y chromosomes are MT, X and Y (when available). Scaffolds that are not assigned to chromosomes use RefSeq ids ("NW\_...") for all species. The following species do not have full chromosomes, and use only RefSeq ids: *H. americanus*, *M. salmoides*, *M. saxatilis*, *P. vannamei*, *S. dumerili* and *S. lalandi dorsalis*. The following species use alternative identifiers (eg. LG1, LG01 or ssa01): *E. lucius*, *O. kisutch*, *O. nerka*, *O. niloticus*, *O. tshawytscha*, *S. salar* and *Salvelinus* sp. You should use the "Organism Name → Chromosome IDs" template query to view ids for those species.

**Warning:** After running your search, using your browser back key to revise the search may cause unexpected changes to the parameters. It would be better to start a new search.

[More genome coordinates help](#)

1. Select Organism:

2. Select Assembly:

3. ☐ Select Feature Types:

☐ CDS

☐ D Loop

☐ Guide RNA

☐ mRNA

☐ Origin Of Replication

☐ Pseudogenic Transcript

☐ scRNA

☐ tRNA

☐ Y RNA

☐ C Gene Segment

☐ Exon

☐ J Gene Segment

☐ miRNA

☐ Pseudogene

☐ rRNA

☐ snRNA

☐ Transcript

☐ D Gene Segment

☐ Gene

☐ lncRNA

☐ ncRNA

☐ Pseudogenic Exon

☐ SNV

☐ snoRNA

☐ V Gene Segment

4. Type/Paste in genomic regions in ☒ base coordinate ☐ interbase coordinate

(example for input format chr:1..1000)▼

(example for input format chr:1-1000)▼

(example for tab delimited input format)▼

or Upload genomic regions from a .txt file...

No file selected.

5. Extend your regions at both sides:



6. ☐ Check this box to perform a strand-specific region search (search + strand if region start<end; search - strand if region end<start)

Fig. 1: Genomic Regions search form

## Search for features within Genomic Regions

Search for features that overlap a list of genome coordinates you enter or upload, e.g. 6:50000..100000. The primary Chromosome IDs used in AquaMine are the "Assigned-Molecule" from the NCBI assembly report. You can retrieve all cross-reference identifiers using the template query [Organism Name → Chromosome IDs](#) found on the AquaMine home page under "ENTIRE GENE SET". Most species use numbers for autosomes. The mitochondrial, X and Y chromosomes are MT, X and Y (when available). Scaffolds that are not assigned to chromosomes use RefSeq ids ("NW\_...") for all species. The following species do not have full chromosomes, and use only RefSeq ids: *H. americanus*, *M. salmoides*, *M. saxatilis*, *P. vannamei*, *S. dumerili* and *S. lalandi dorsalis*. The following species use alternative identifiers (eg. LG1, LG01 or ssa01): *E. lucius*, *O. kisutch*, *O. nerka*, *O. niloticus*, *O. tshawytscha*, *S. salar* and *Salvelinus* sp. You should use the "Organism Name → Chromosome IDs" template query to view ids for those species.

**Warning:** After running your search, using your browser back key to revise the search may cause unexpected changes to the parameters. It would be better to start a new search.

[More genome coordinates help](#)

- Select Organism:
- Select Assembly:
- Select Feature Types:
 

<input type="checkbox"/> CDS	<input type="checkbox"/> C Gene Segment	<input type="checkbox"/> D Loop
<input type="checkbox"/> Exon	<input checked="" type="checkbox"/> Gene	<input type="checkbox"/> J Gene Segment
<input type="checkbox"/> lncRNA	<input type="checkbox"/> mRNA	<input type="checkbox"/> miRNA
<input type="checkbox"/> ncRNA	<input type="checkbox"/> Pseudogene	<input type="checkbox"/> Pseudogenic Exon
<input type="checkbox"/> Pseudogenic Transcript	<input type="checkbox"/> rRNA	<input type="checkbox"/> scRNA
<input type="checkbox"/> snRNA	<input type="checkbox"/> snoRNA	<input type="checkbox"/> tRNA
<input type="checkbox"/> Transcript	<input type="checkbox"/> V Gene Segment	
- Type/Paste in genomic regions in ☒ base coordinate ☐ interbase coordinate  
 (example for input format chr:1..1000)▼  
 (example for input format chr:1-1000)▼  
 (example for tab delimited input format)▼  

14:6000000..6800000

 or Upload genomic regions from a .txt file...  
 No file selected.
- Extend your regions at both sides:  

1k

10k

100k

1M

10M
- ☐ Check this box to perform a strand-specific region search (search + strand if region start<end; search - strand if region end<start)

Fig. 2: Genomic Regions search example with *Ictalurus punctatus*

Selected organism: *I. punctatus*

Selected assembly: *IpCoco\_1.2*

Selected feature types: Gene

Hide

Export data for all features within all regions: 

TABCSVGFF3BEDFASTA

Export entire sequences for all regions: 

FASTA

Create list by feature type: 

Gene

Go

Page size 10 

<< First < Prev | Next > Last >>

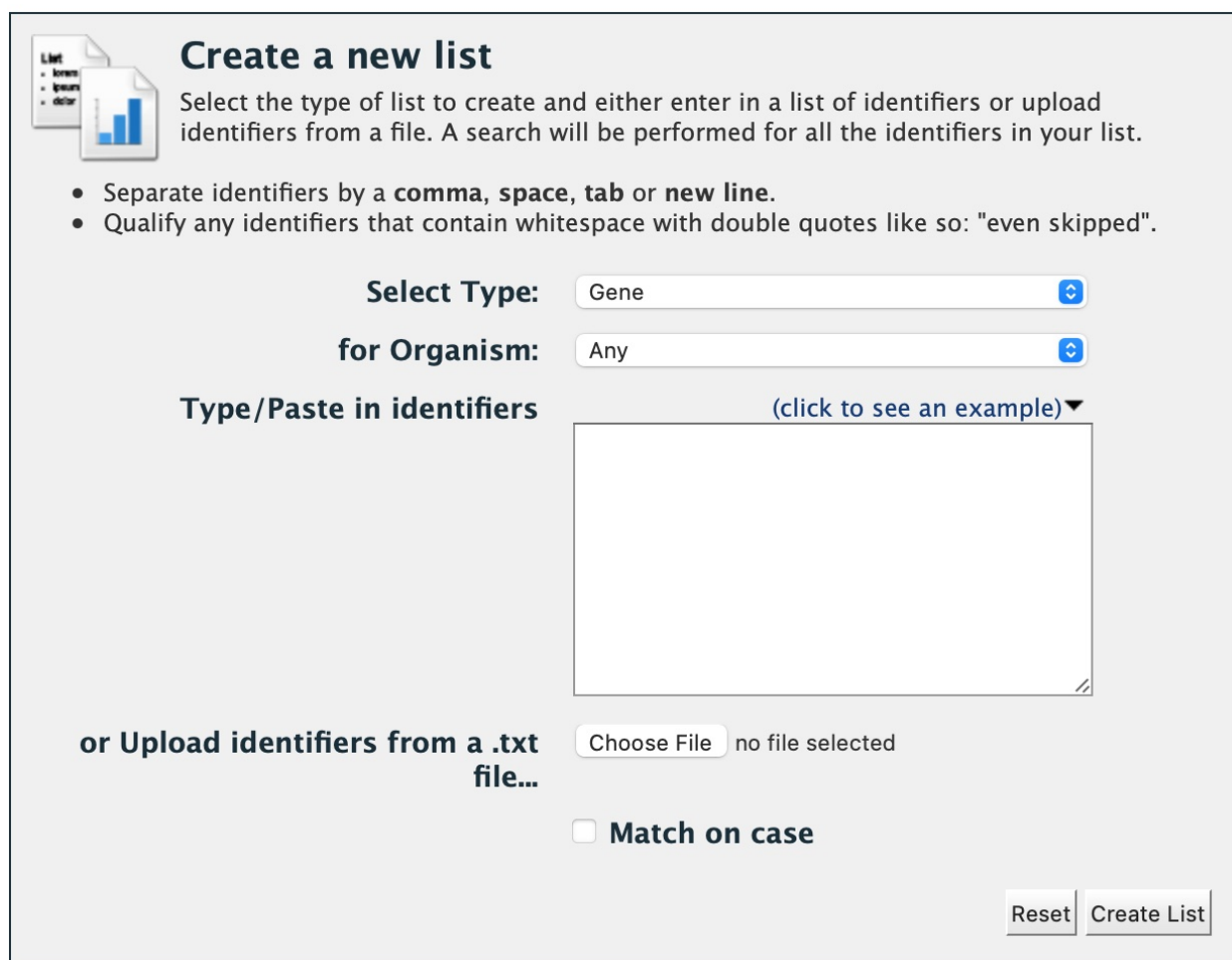
GENOME REGION	FEATURE	FEATURE TYPE	LOCATION
14:60000000..68000000 Export sequence for entire region: <div>FASTA</div> <div>TABCSVGFF3BEDFASTAGALAXY</div> <div>Create List by <div>Gene</div> <div>Go</div></div>	syt7 108275255	Gene	14:5884617..6022151
	syt7a ENSIPUG000000008858	Gene	14:5895585..6022361
	slc25a22 108274700	Gene	14:6031983..6059790
	slc25a22a ENSIPUG000000008927	Gene	14:6032990..6049147
	5S_rRNA ENSIPUG000000008991	Gene	14:6041279..6041372
	pidd1 ENSIPUG000000009018	Gene	14:6058149..6072046
	pidd1 108274699	Gene	14:6063462..6072020
	pnpla2 ENSIPUG000000009037	Gene	14:6076653..6092627
	pnpla2 108274701	Gene	14:6076804..6087777
LOC108274702 108274702	Gene	14:6093022..6128197	

Fig. 3: Genomic Regions Search results

## LISTS

### 5.1 Creating Lists

Users may create and save lists of features, such as gene IDs, transcript IDs, gene symbols, etc. The list tool searches the database for the list items and attempts to convert each identifier to the selected type. Click on the Lists tab from the menu to access the full list upload form. A short version of the form is also in the Quick List box on the home page.



**Create a new list**

Select the type of list to create and either enter in a list of identifiers or upload identifiers from a file. A search will be performed for all the identifiers in your list.

- Separate identifiers by a **comma, space, tab or new line**.
- Qualify any identifiers that contain whitespace with double quotes like so: "even skipped".

**Select Type:** Gene

**for Organism:** Any

**Type/Paste in identifiers** [\(click to see an example\)](#)

**or Upload identifiers from a .txt file...** Choose File no file selected

☐ **Match on case**

Reset Create List

Fig. 1: List upload form

Fig. 2: Quick list from AquaMine home page

As an example, enter the following comma-separated identifiers into the Lists upload form under the **Lists** tab. Notice that they do not have to be in the same format. A Summary table is displayed with the results of searching for each of the five identifiers in the list.

LOC105319044, 114545253, wisp1, ENSONEG000000000134, gstk1

Leave the **Select Type** drop-down menu to **Gene** and the **Organism** drop-down to **Any**. Click on **Create List**. Note that you can also upload a list from a .txt file.

Duplicates found - which one(s) do you want? [?](#) Add all Remove all

Page 1 of 6 1 2 3 4 5 6 ... 5 rows per page

Identifier you provided	Matches <a href="#">?</a>									Action <a href="#">?</a>
	biotype	symbol	chromosome assembly	length	description	organism name	source	primary identifier	class	
gstk1	protein_coding	gstk1	AstMex3_surface	30231	glutathione S-transferase kappa 1	Astyanax mexicanus	RefSeq	103043665	Gene	<span>Add</span>
	protein_coding	gstk1	Ch_v2.0.2	6241	glutathione S-transferase kappa 1	Clupea harengus	RefSeq	105901652	Gene	<span>Add</span>
	protein_coding	gstk1	GRCz11	4725	glutathione S-transferase kappa 1	Danio rerio	RefSeq	436833	Gene	<span>Add</span>
	protein_coding	gstk1	Eluc_v4	6376	glutathione S-transferase kappa 1	Esox lucius	RefSeq	105019054	Gene	<span>Add</span>
	protein_coding	gstk1	gadMor3.0	2038	glutathione S-transferase kappa 1	Gadus morhua	RefSeq	115554426	Gene	<span>Add</span>

Fig. 3: List Example: Search results for list of identifiers

The summary table provides information regarding those identifiers that had a direct hit without any duplicates. If there are any duplicates, users can decide to add the relevant entries individually by clicking on the **Add** button under the **Action** column or choosing the **Add all** tab. Here we will click **Add all**. Once the selections have been added, the list can be saved by clicking the **Save a list of 31 Genes** button on the top of the summary table. Name the list by entering text into the **Choose a name for the list** box at the top of the results page.

After the list is saved, users are presented with a **List Analysis** page. This page provides users with widgets to perform analyses on gene lists that they have created.

The selection of widgets provided on the List Analysis page depend on the contents of the list. The available widgets for this list example include:



Before we show you the results ...

Choose a name for the list

All\_organism\_gene search

(e.g. Smith 2013)

Add additional matches

You entered: 5 identifiers  
We found: 3 Genes

Save a list of 31 Genes

Why are the numbers different? See below.

Duplicates found - which one(s) do you want?

Add all Remove all

Page 1 of 6 1 2 3 4 5 6

5 rows per page

Identifier you provided	Matches									Action
	biotype	symbol	chromosome assembly	length	description	organism name	source	primary identifier	class	
gstk1	protein_coding	gstk1	AstMex3_surface	30231	glutathione S-transferase kappa 1	Astyanax mexicanus	RefSeq	103043665	Gene	Remove
	protein_coding	gstk1	Ch_v2.0.2	6241	glutathione S-transferase kappa 1	Clupea harengus	RefSeq	105901652	Gene	Remove
	protein_coding	gstk1	GRCz11	4725	glutathione S-transferase kappa 1	Danio rerio	RefSeq	436833	Gene	Remove
	protein_coding	gstk1	Eluc_v4	6376	glutathione S-transferase kappa 1	Esox lucius	RefSeq	105019054	Gene	Remove
	protein_coding	gstk1	gadMor3.0	2038	glutathione S-transferase kappa 1	Gadus morhua	RefSeq	115554426	Gene	Remove

Fig. 4: List Example: Saving list of identifiers

List Analysis for All\_organism\_gene search (31 Genes)

Manage Columns

Manage Filters

Manage Relationships

Save as List

Generate Python code

Export

Showing 1 to 25 of 28 rows Rows per page: 25

Gene ID	Gene Source	Gene Biotype	Gene Symbol	Gene Description	Gene Length	Gene Chromosome ID	Gene Chromosome Location Start
100305229	RefSeq	protein_coding	gstk1	Glutathione S-transferase kappa 1	9820	2	19987402
100528071	RefSeq	protein_coding	wisp1	wnt1-inducible-signaling pathway protein 1	4434	24	10474060
100528083	RefSeq	protein_coding	gstk1	glutathione S-transferase kappa 1	3712	1	2665889
100694768	RefSeq	protein_coding	wisp1	WNT1 inducible signaling pathway protein 1	7803	LG22	23164664
100709083	RefSeq	protein_coding	gstk1	glutathione S-transferase kappa 1	4264	LG11	22928341
102683192	RefSeq	protein_coding	wisp1	WNT1 inducible signaling pathway protein 1	8614	LG11	22373380
103043665	RefSeq	protein_coding	gstk1	glutathione S-transferase kappa 1	30231	6	26605535
105019054	RefSeq	protein_coding	gstk1	glutathione S-transferase kappa 1	6376	LG20	19404149
105319044	RefSeq	protein_coding	LOC105319044	putative transcription factor p65 homolog	9348	LG3	56616354
105901652	RefSeq	protein_coding	gstk1	glutathione S-transferase kappa 1	6241	11	9482505

Orthologues

A. mexicanus (19) C. clupeaformis (35) C. gigas (4)  
C. harengus (27) C. virginica (8) D. melanogaster (3)  
D. rerio (27) E. cragini (20) E. lucius (20)  
G. aculeatus aculeatus (16) G. morhua (18) H. americanus (8)  
H. hippoglossus (19) H. rufescens (4) H. sapiens (14)  
H. stenolepis (19) I. punctatus (20) L. gigantea (4)  
L. oculatus (15) M. cephalus (23) M. mercenaria (5)  
M. salmoides (84) M. saxatilis (19) O. gorbuscha (31)  
O. keta (28) O. kisutch (40) O. mykiss (37) O. nerka (33)  
O. niloticus (205) O. tshawytscha (32) P. clarkii (8)  
P. flavescens (17) P. monodon (4) P. vannamei (5)  
S. dumerili (24) S. lalandi dorsalis (28) S. salar (37)  
S. sp. IW2-2015 (4) X. gladius (19)

External Links

No external links.

Fig. 5: List Example: Analysis for gene list

5.1. Creating Lists

33

1. Gene Ontology Enrichment
2. Publication Enrichment
3. Pathway Enrichment

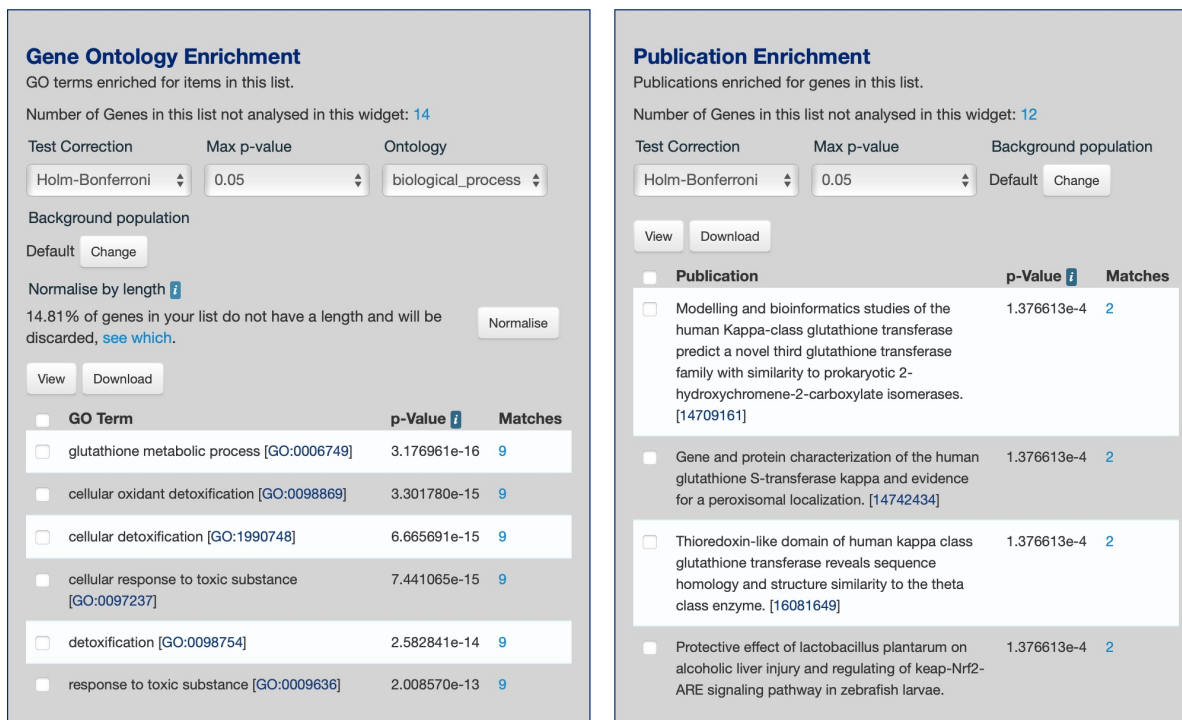
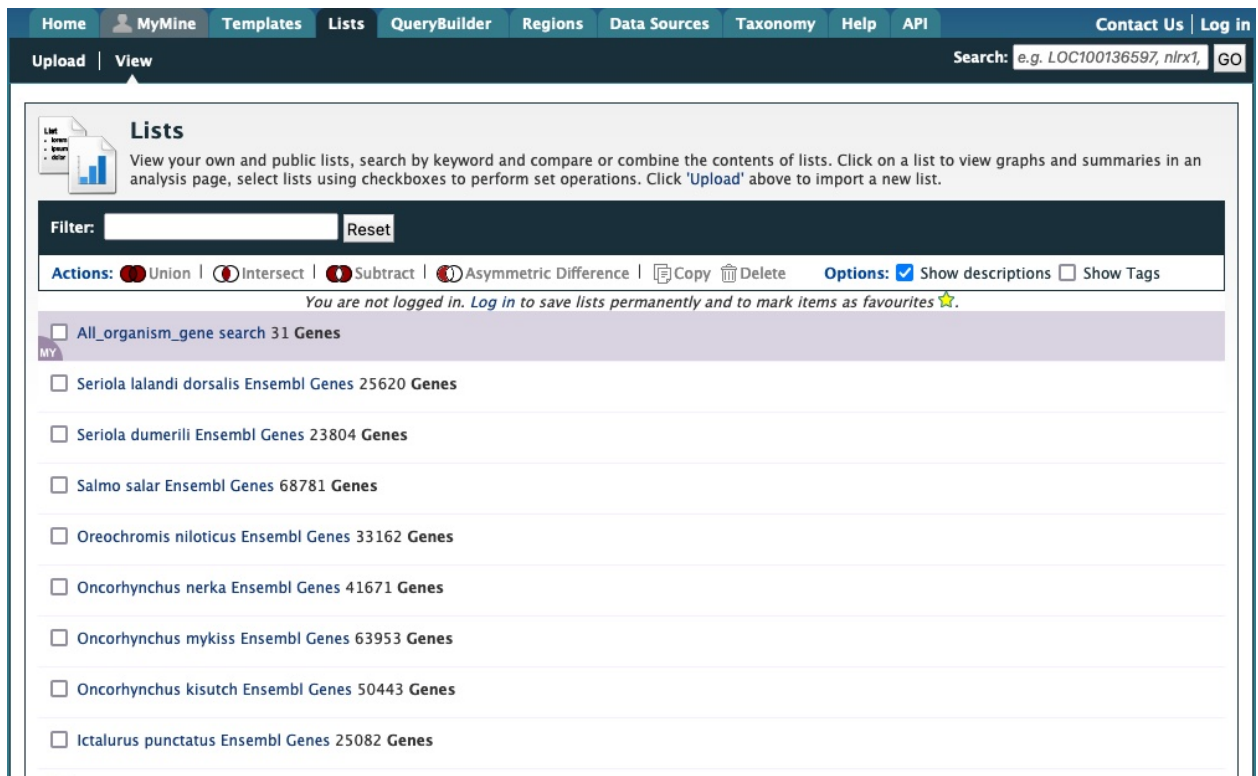


Fig. 6: List Example: Displayed widgets for list analysis

Before using the widgets, be sure to read the **Important Notes for Enrichment Widgets to avoid false positives**.

## 5.2 Saving Lists

To see your saved lists, click the **View** tab on the **Lists** page. Note the saved list is highlighted in light purple and has a “MY” label in the corner. The remaining lists are gene lists for the organisms in AquaMine that can be used for future analyses. If not logged in, lists will be saved temporarily during your current session. However, you must be logged in to save your lists permanently. Further analyses of lists can be done with the **Actions** links at the top of the list. The links become active once lists are selected for analyses. Saved lists may also be accessed from the **MyMine** menu tab.



The screenshot shows the AquaMine web interface. The top navigation bar includes links for Home, MyMine, Templates, Lists, QueryBuilder, Regions, Data Sources, Taxonomy, Help, and API. On the right, there are links for Contact Us and Log in. Below the navigation bar, there is a search bar with the text "Search: e.g. LOC100136597, nlr1," and a "GO" button. The main content area is titled "Lists" and contains a description: "View your own and public lists, search by keyword and compare or combine the contents of lists. Click on a list to view graphs and summaries in an analysis page, select lists using checkboxes to perform set operations. Click 'Upload' above to import a new list." Below the description, there is a "Filter:" input field and a "Reset" button. A row of action buttons includes "Union", "Intersect", "Subtract", "Asymmetric Difference", "Copy", and "Delete". To the right of these buttons are "Options:" with checkboxes for "Show descriptions" (checked) and "Show Tags" (unchecked). A message states: "You are not logged in. Log in to save lists permanently and to mark items as favourites." Below this, a list of saved user lists is shown, each with a checkbox and a label: "All\_organism\_gene search 31 Genes", "Seriola lalandi dorsalis Ensembl Genes 25620 Genes", "Seriola dumerili Ensembl Genes 23804 Genes", "Salmo salar Ensembl Genes 68781 Genes", "Oreochromis niloticus Ensembl Genes 33162 Genes", "Oncorhynchus nerka Ensembl Genes 41671 Genes", "Oncorhynchus mykiss Ensembl Genes 63953 Genes", "Oncorhynchus kisutch Ensembl Genes 50443 Genes", and "Ictalurus punctatus Ensembl Genes 25082 Genes".

Fig. 7: List Example: Saved user lists



## MYMINE

MyMine serves as a portal where logged-in users may manage their lists, queries, templates, and account details.

To access MyMine, click on the MyMine menu tab. A submenu appears with six options:

**Lists** - Lists saved by the user when logged in.

**History** - List of most recently run queries.

**Queries** - List of saved queries.





**Templates** - Templates created or marked as “favorite” by the user.

**Password** - Password reset form.

**Account Details** - User preferences form.



### Your Lists

<input type="checkbox"/>	LIST NAME	DESCRIPTION	TYPE	NUMBER OF OBJECTS	DATE CREATED
<input type="checkbox"/>	All_organism_gene search   <a href="#">Add tags</a> <a href="#">Share with users</a>		Gene	27 values	2022-05-20 16:04
<input type="checkbox"/>	S_salar_ontology_terms   <a href="#">Add tags</a> <a href="#">Share with users</a>		Ontology Term	1055 values	2022-05-20 16:14

New list name:

Fig. 1: Saved lists found under MyMine. Note that currently saved lists can be selected for analyses to contribute to new lists.



An API is available for users who would like to programmatically access FAANGMine.

### Perl Web Service Client

The Perl web service client library makes it easy to run queries in AquaMine directly from Perl programs. You can use these modules to construct any query you could run from web interface and fetch the results in a number of formats, including native Perl data structures and objects, and TSV/CSV strings.

Like all our code, it is open-source, coming licensed under the LGPL. For information on our API, visit our [wiki pages](#).

- **Prerequisites**

You should install the [Perl webservice client library module](#) to get started. You can install it directly from CPAN (the comprehensive Perl archive network). This is a public repository of thousands of modules. Using CPAN will ensure you get the latest version, and that your dependencies are automatically managed for you.

To install the InterMine Perl client library type the following command into a shell:

```
> sudo cpan Webservice::InterMine
```

- **Examples of using the downloaded script**

On each Template Query page and the QueryBuilder there is a link to get Perl code to run that particular query using the web service API. Just click the link, save the generated Perl script in a file and execute it. You can use the generated code as a starting point for your own programs.

You can run the downloaded script by running the following command in a shell:

```
> perl path/to/downloaded/script.pl
```

If you get an error saying `Cannot find Webservice::InterMine in @INC`, or similar then see 'Prerequisites' above.

Feel free to edit the script – these are designed to be spring-boards to help you get where you want to. See the [Cookbook](#), which contains a set of short tutorial 'recipes' that demonstrate particular features of the Perl API, for ways to edit the scripts, and even extend their functionality.

For a good reference to writing programs in Perl, see [here](#).

Perl, Python, Ruby, and Java are the languages supported by the InterMine API.

For more detailed information, view the [InterMine documentation](#).





## DATA SOURCES

The Data Sources table provides a description of the datasets that are integrated into AquaMine, along with their download location, version or release, citations wherever applicable, and any additional comments. It is a useful place to begin to find out what data has been loaded for your organism of interest.

Data Category	Description	Organism (* US Aquaculture Species)	Source	PubMed	Link
Assembly	Genome Assembly	<i>Astyanax mexicanus</i> (Mexican tetra/Mexican cavefish)	AstMex3_surface	Warren et al. – PubMed 33664263	NCBI FTP
		<i>Clupea harengus</i> (Atlantic herring)	Ch_v2.0.2	Pettersson et al. – PubMed 31649060	NCBI FTP
		<i>Coregonus clupeaformis</i> (Lake whitefish)	ASM2061545v1	Mérot et al. – PubMed 35416336	NCBI FTP
		<i>Crassostrea gigas</i> (Pacific oyster)*	cgigas_uk_roslin_v1	Peñaloza et al. – PubMed 33764468	NCBI FTP
		<i>Crassostrea virginica</i> (Eastern oyster)*	C_virginica-3.0	Gómez-Chiari et al. – PubMed 25982405	NCBI FTP
		<i>Danio rerio</i> (Zebrafish)	GRCz11	Howe et al. – PubMed 23594743	NCBI FTP
		<i>Drosophila melanogaster</i> (Fruit fly)	Release_6_plus_ISO1_MT	Hoskins et al. – PubMed 25589440	NCBI FTP
		<i>Esox lucius</i> (Northern pike)	Eluc_v4	Rondeau et al. – PubMed 25069045	NCBI FTP
		<i>Etheostoma cragini</i> (Arkansas darter)	CSU_Ecrag_1.0	Reid et al. – PubMed 33058399	NCBI FTP
		<i>Gadus morhua</i> (Atlantic cod)	gadMor3.0	Tørrisen et al. – PubMed 28100185	NCBI FTP
		<i>Gasterosteus aculeatus aculeatus</i> (Three-spined stickleback)	GAculeatus_UGA_version5	Nath et al. – PubMed 33598708	NCBI FTP
		<i>Haliotis rufescens</i> (Red abalone)*	xgHaliRuf1.0.p	Griffiths et al. – PubMed 36190478	NCBI FTP
		<i>Hippoglossus hippoglossus</i> (Atlantic halibut)	fHipHip1.pri	Einfelt et al. – PubMed 33655659	NCBI FTP
		<i>Hippoglossus stenolepis</i> (Pacific halibut)	HSTE1.2	Jasonowicz et al. – PubMed 35569134	NCBI FTP
		<i>Homarus americanus</i> (American lobster)	GMCI_Hamer_2.0	Polinski et al. – PubMed 34162536	NCBI FTP
		<i>Homo sapiens</i> (Human)	T2T-CHM13v2.0	Nurk et al. – PubMed 35357919	NCBI FTP
		<i>Ictalurus punctatus</i> (Channel catfish)*	IpCoco_1.2	Liu et al. – PubMed 27249958	NCBI FTP
		<i>Lepisosteus oculatus</i> (spotted gar)	LepOcu1	Braasch et al. – PubMed 26950095	NCBI FTP
		<i>Lottia gigantea</i> (Owl limpet)	Heiro1	Simakov et al. – PubMed 23254933	NCBI FTP
		<i>Mercenaria mercenaria</i> (Northern quahog)	ASM1480567v1.1	Song et al. – PubMed 33487168	NCBI FTP
		<i>Micropterus salmoides</i> (Largemouth bass)	ASM1485139v1	Sun et al. – PubMed 32985096	NCBI FTP
		<i>Morone saxatilis</i> (Striped sea bass)*	NCSU_SB_2.0	Not published. North Carolina State University	NCBI FTP
		<i>Mugil cephalus</i> (Flathead mullet/Striped mullet)	CIBA_Mcephalus_1.1	Shekhar et al. – PubMed 35783261	NCBI FTP
		<i>Oncorhynchus gorbuscha</i> (Pink salmon)	OgorEven_v1.0	Christensen et al. – PubMed 34919547	NCBI FTP
		<i>Oncorhynchus keta</i> (Chum salmon)	Oket_V1	Rondeau EB et al. – bioRxiv 2021.12.27.474290	NCBI FTP
		<i>Oncorhynchus kisutch</i> (Coho salmon)*	Okis_V2	Kim et al. – PubMed 26614614	NCBI FTP
		<i>Oncorhynchus mykiss</i> (Rainbow trout)*	USDA_OmyKA_1.1	Gao et al. – PubMed 33616628	NCBI FTP
		<i>Oncorhynchus nerka</i> (Sockeye salmon)	Oner_1.0	Christensen et al. – PubMed 33119641	NCBI FTP
		<i>Oncorhynchus tshawytscha</i> (Chinook salmon)	Otrsh_v2.0	Christensen et al. – PubMed 29621340	NCBI FTP
		<i>Oreochromis niloticus</i> (Nile tilapia)*	O_niloticus_UMD_NMBU	Conte et al. – PubMed 30942871	NCBI FTP

Fig. 1: AquaMine Data Sources table



## HOW TO CITE

AquaMine is a project supported by the National Science Foundation to address the need for a high performance data mining resource that enables fine-grained querying and integrating of the heterogeneous aquatic species data with existing information, such as functions of known genes and research datasets.

For more generic examples on how to use InterMine, click [here](#). These are tutorials created by FlyMine that showcase the different features of InterMine.