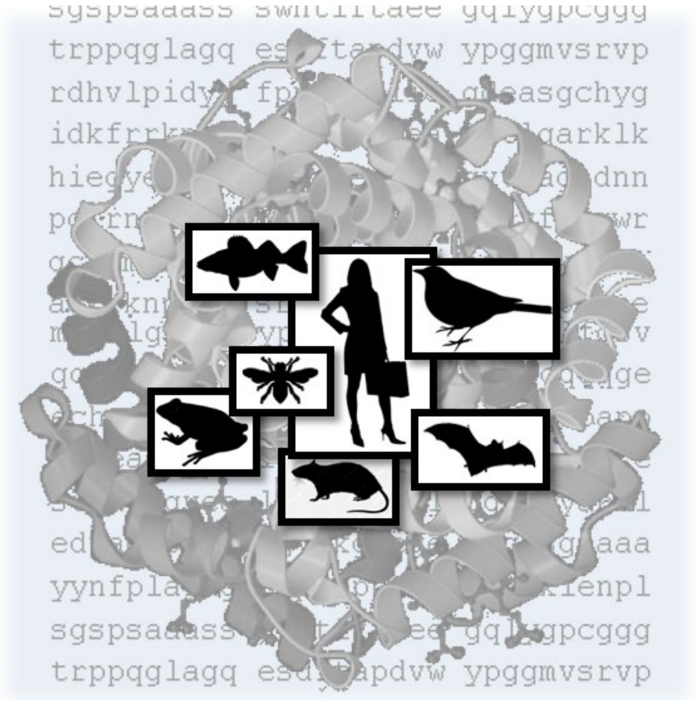
EPA/600/R-19/156

Sequence Alignment to Predict Across Species Susceptibility

(SeqAPASS)

VERSION 4.0



User Guide

Authors:

Carlie A. LaLone
Donovan J. Blatz
Colin P. Finnegan

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS): User Guide
Updated 09/10/19; Contact Carlie LaLone with Questions: LaLone.Carlie@epa.gov

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) User Guide

Quick Notes: Use Chrome for optimal performance and PLEASE DO NOT submit more than 10 Level 1 queries at a time. Wait until they run to completion prior to submitting more.

Table of Contents

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS): User Guide
Updated 09/10/19; Contact Carlie LaLone with Questions: LaLone.Carlie@epa.gov

Background

The SeqAPASS tool has been developed to predict across species relative intrinsic susceptibility to chemicals with known molecular targets (e.g., pharmaceuticals, pesticides) as well as evaluate conservation of molecular targets from high-throughput screening assays (i.e., U.S. Environmental Protection Agency ToxCast Program) and molecular initiating events (MIEs) and early key events in the adverse outcome pathway framework, as a means to extrapolate such knowledge across species. The term "relative" is used because it is recognized that molecular target similarity is one consideration, though an important one, for making predictions of susceptibility to a chemical. Other important considerations for susceptibility that are not evaluated using the SeqAPASS methodology include how well a chemical is absorbed, distributed, metabolized, and eliminated, life stage, and other life history traits. Also, "relative" indicates that the determination of sequence similarity between proteins is based on comparison to a single protein sequence for a specific species. Additionally, we describe "intrinsic susceptibility" as the vulnerability (or lack thereof) of an organism to chemical perturbation due to its inherent biological composition.

Cross-species comparisons of proteins can be conducted through examination of sequence and structural information, depending on how well the protein has been characterized and what is known about a chemical-protein interaction. SeqAPASS allows the user to assess various levels of protein sequence detail across species including comparisons of primary amino acid sequence (including ortholog detection), functional domain(s), and individual amino acid residue positions. Each level requires a greater understanding of the protein and its interaction with a chemical of interest (or similar ligand). Because human and veterinary drugs, as well as pesticides, are designed to act specifically on well characterized molecular targets, these chemical classes have proven useful for demonstrating the utility of the SeqAPASS tool and its application to various hazard assessment/research scenarios.

The pertinent information necessary to begin a SeqAPASS query includes: the identification of a single (or multiple) query species and a query protein, which would be the molecular target(s) of interest (e.g., receptor or enzyme).

The SeqAPASS algorithms mine, collect, and collate information from the National Center for Biotechnology Information (NCBI) protein database (http://www.ncbi.nlm.nih.gov/protein/), conserved domains database (http://www.ncbi.nlm.nih.gov/cdd/), taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy/), strategically utilizes the Stand-Alone Basic Local Alignment Search Tool for proteins (BLASTp; http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download and the Constraint-based Multiple Alignment Tool (COBALT; http://www.st-va.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi).

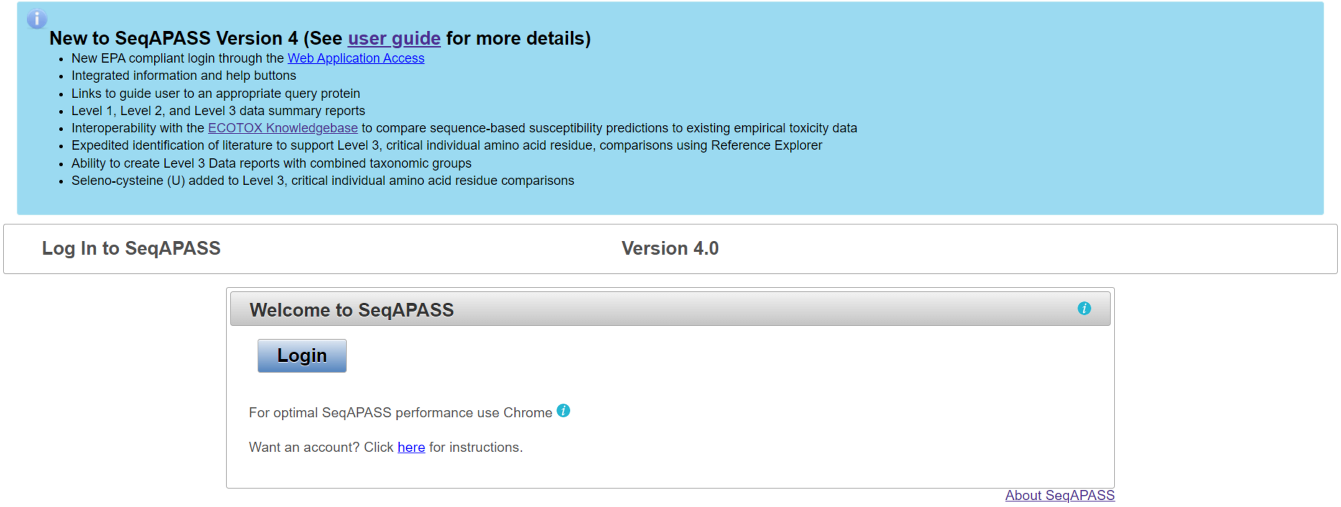
Accessing SeqAPASS

For optimal SeqAPASS performance use **Chrome**

Access SeqAPASS using the following URL: https://www.seqapass.epa.gov/seqapass/

Returning Users

Click "Login"



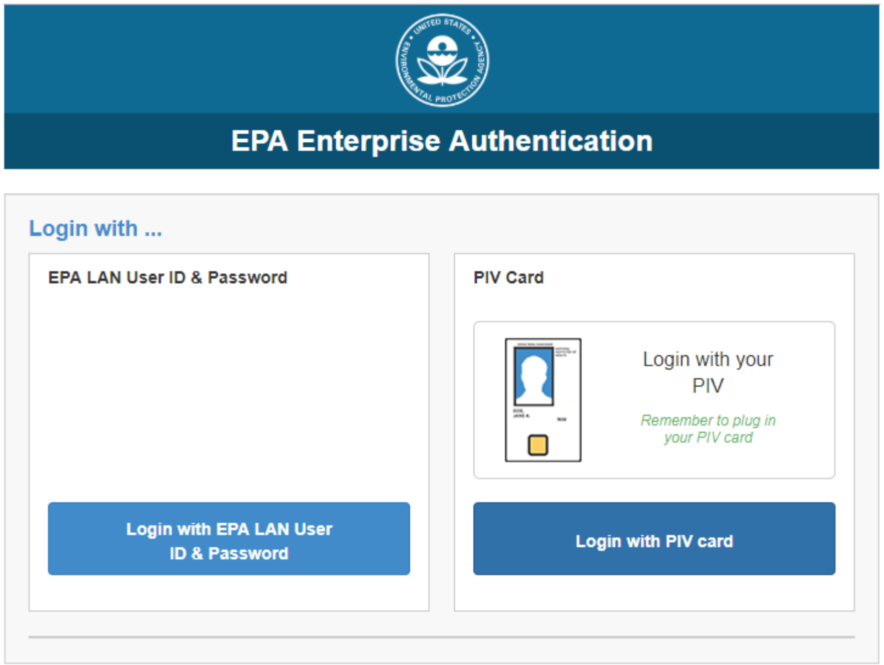
New to SeqAPASS Version 4 (See user guide for more details)

- New EPA compliant login through the Web Application Access
- Integrated information and help buttons
- Links to guide user to an appropriate query protein
- Level 1, Level 2, and Level 3 data summary reports
- Interoperability with the ECOTOX Knowledgebase to compare sequence-based susceptibility predictions to existing empirical toxicity data
- Expedited identification of literature to support Level 3, critical individual amino acid residue, comparisons using Reference Explorer
- Ability to create Level 3 Data reports with combined taxonomic groups
- Seleno-cysteine (U) added to Level 3, critical individual amino acid residue comparisons

For optimal SeqAPASS performance use Chrome

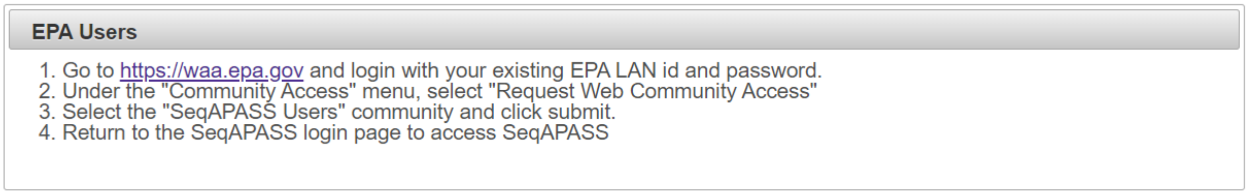
Want an account? Click here for instructions.

Select either "Login with EPA LAN User ID & Password" or "Login with PIV card" with two step verification to login.



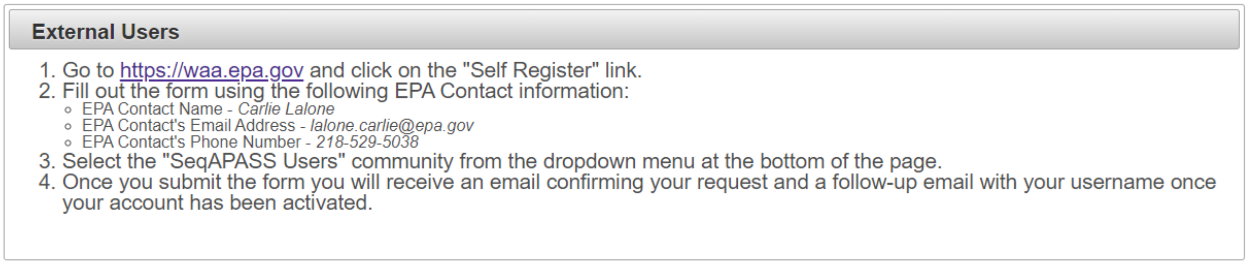
First time users

To request a username and password to access the SeqAPASS tool, select "here" below the login and follow the directions on the next page. The directions are different for the internal EPA user versus the external non-EPA user, however the user type does not limit access to the tool. Everyone that requests an account will be given one in a timely manner. Individual account allows users to store all previous SeqAPASS runs. Once the user has obtained their username, external users will select "Login with EPA LAN User ID and Password."



EPA Users

1. Go to https://waa.epa.gov and login with your existing EPA LAN id and password.
2. Under the "Community Access" menu, select "Request Web Community Access"
3. Select the "SeqAPASS Users" community and click submit.
4. Return to the SeqAPASS login page to access SeqAPASS



External Users

1. Go to https://waa.epa.gov and click on the "Self Register" link.
2. Fill out the form using the following EPA Contact information:
    - EPA Contact Name - *Carlie Lalone*
    - EPA Contact's Email Address - *lalone.carlie@epa.gov*
    - EPA Contact's Phone Number - *218-529-5038*
3. Select the "SeqAPASS Users" community from the dropdown menu at the bottom of the page.
4. Once you submit the form you will receive an email confirming your request and a follow-up email with your username once your account has been activated.

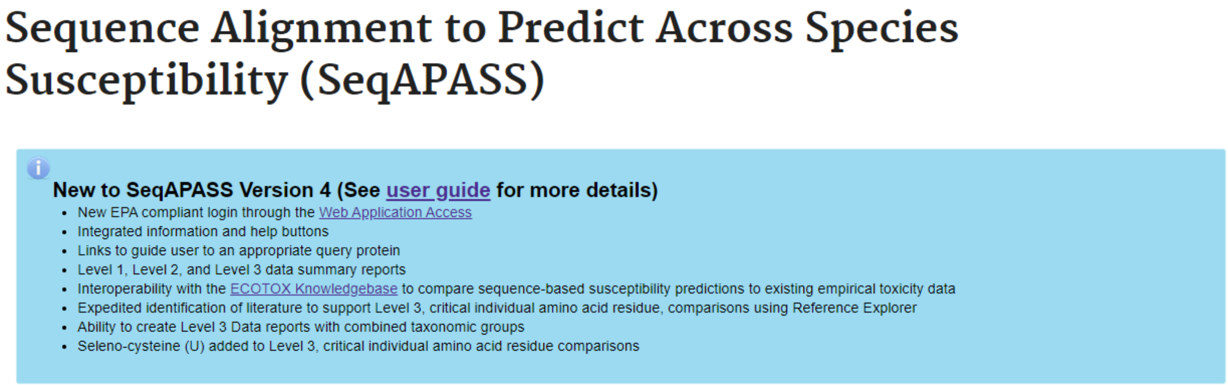
On the Log in screen the user will provide the necessary Login information:

EPA User: EPA LAN User ID & Password ***or*** PIV card with two step verification
External User: Username and Password

Upon creating your password, login to SeqAPASS as described above for *Returning Users*. To change a password at any time, go to waa.epa.gov and select "User Profile" to reset. The user will then use the new password to login.

Messages from the SeqAPASS development team

Look for messages about planned version releases, data updates, and/or fixes to the SeqAPASS tool. These will occasionally be displayed below the SeqAPASS banner when the development team has information to share with SeqAPASS users.



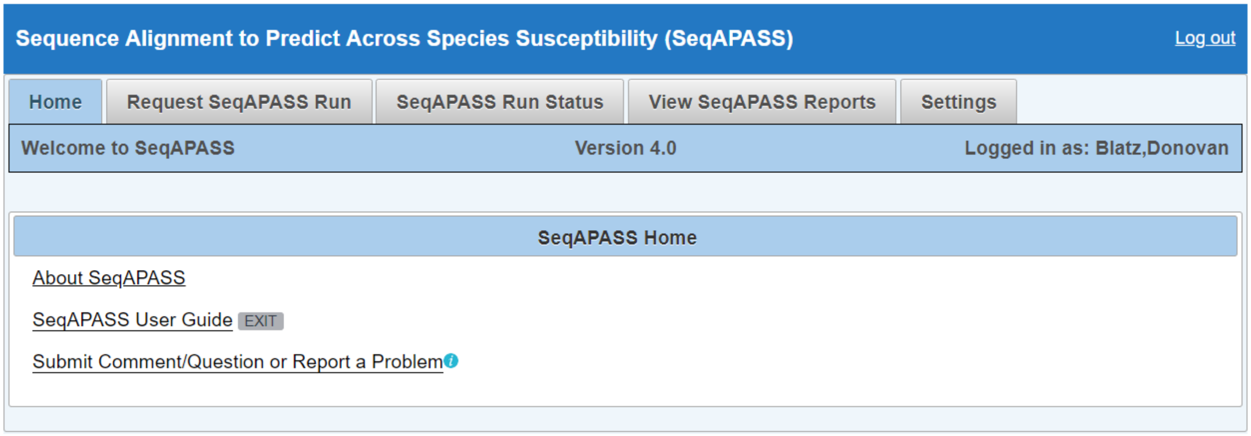
Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS)

New to SeqAPASS Version 4 (See user guide for more details)

- New EPA compliant login through the Web Application Access
- Integrated information and help buttons
- Links to guide user to an appropriate query protein
- Level 1, Level 2, and Level 3 data summary reports
- Interoperability with the ECOTOX Knowledgebase to compare sequence-based susceptibility predictions to existing empirical toxicity data
- Expedited identification of literature to support Level 3, critical individual amino acid residue, comparisons using Reference Explorer
- Ability to create Level 3 Data reports with combined taxonomic groups
- Seleno-cysteine (U) added to Level 3, critical individual amino acid residue comparisons

SeqAPASS Home Tab

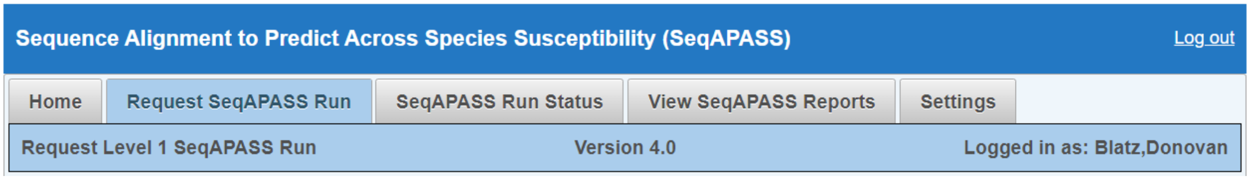
The “Home” tab indicates who is logged in to the tool (right-hand of the screen) and contains links to obtain information about the SeqAPASS tool (About SeqAPASS), including contact information for support and references to published articles describing the SeqAPASS tool and its applications. Other relevant references to databases and tools are also referenced. A link to the SeqAPASS User Guide can also be found on this page. To Submit a Comment/Question click on the “Submit Comment/Question” link to email the developer. “Log out” icon in upper right-hand corner of screen can be clicked at any time to log out. “Information” buttons are present throughout SeqAPASS to give the user additional information or instruction regarding features and functionality of the tool. “Exit” buttons are also present by each external (non-EPA) link that takes the user to a page ***NOT*** maintained by the EPA.



Welcome to SeqAPASS Version 4.0 Logged in as: Blatz,Donovan

Request SeqAPASS Run Tab

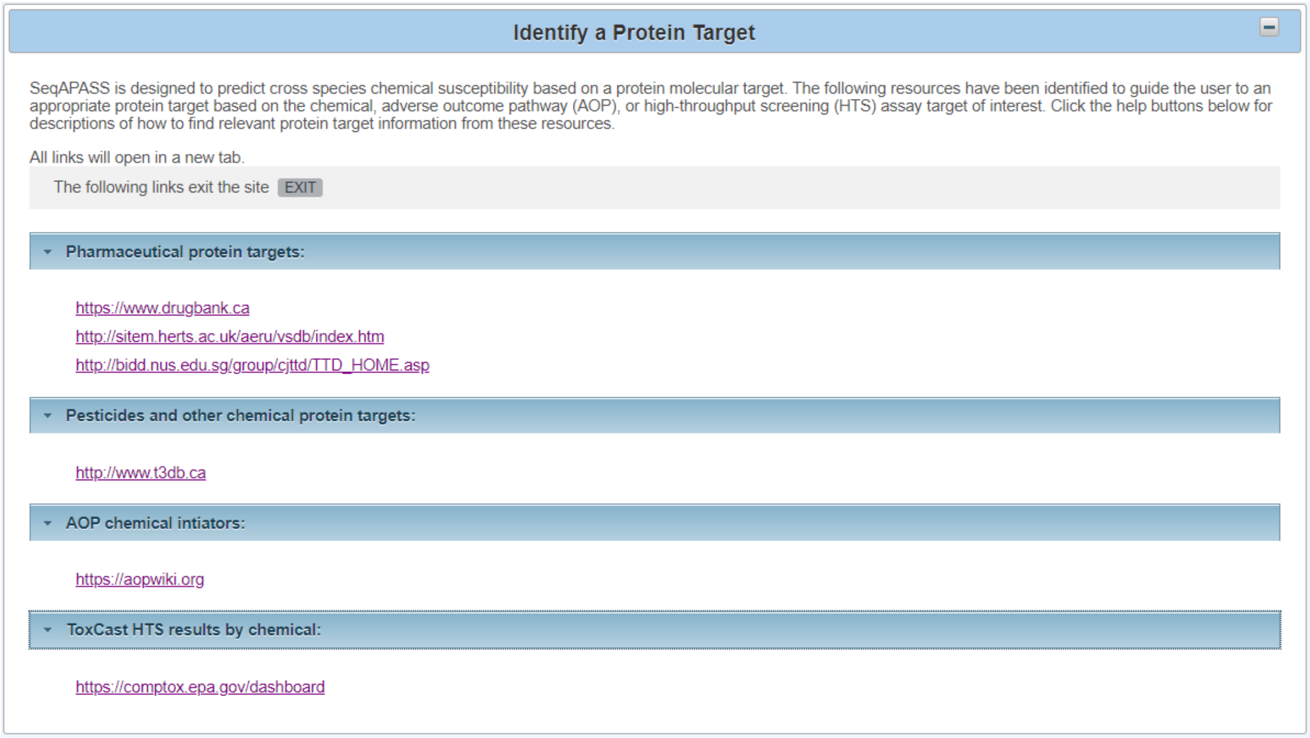
Clicking the “Request SeqAPASS Run” tab opens a page to enter the query information necessary for a SeqAPASS run. Each section of the “Request SeqAPASS Run” will be described below:



Request Level 1 SeqAPASS Run Version 4.0 Logged in as: Blatz,Donovan

Identify a Protein Target

SeqAPASS is designed to predict cross species chemical susceptibility. Protein targets are often decided based on chemical, adverse outcome pathway (AOP), or high-throughput screening (HTS) assay target. Resources have been provided, as links, to aid the user in searching for appropriate protein targets and can be accessed by selecting the drop-downs found in the "Identify a Protein Target" box.


Identify a Protein Target

SeqAPASS is designed to predict cross species chemical susceptibility based on a protein molecular target. The following resources have been identified to guide the user to an appropriate protein target based on the chemical, adverse outcome pathway (AOP), or high-throughput screening (HTS) assay target of interest. Click the help buttons below for descriptions of how to find relevant protein target information from these resources.

All links will open in a new tab.

The following links exit the site EXIT

Pharmaceutical protein targets:

https://www.drugbank.ca
http://sitem.herts.ac.uk/aeru/vsdb/index.htm
http://bidd.nus.edu.sg/group/cjttd/TTD_HOME.asp

Pesticides and other chemical protein targets:

http://www.t3db.ca

AOP chemical intiators:

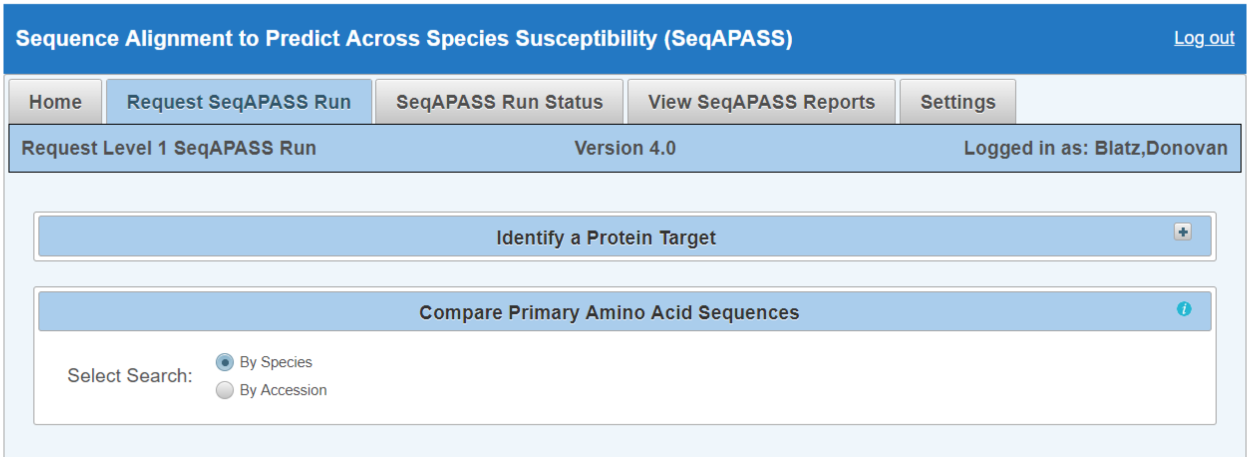
https://aopwiki.org

ToxCast HTS results by chemical:

https://comptox.epa.gov/dashboard

Select Search

There are two options for entering query information: "By Species" or "By Accession" (See radio buttons to the right of "Select Search"). Selecting "By Species" will allow the user to enter text and select from a dropdown list of species and then select a protein from any sequence available for that species in the NCBI protein database. Selecting "By Accession" allows the user to enter a NCBI protein accession.

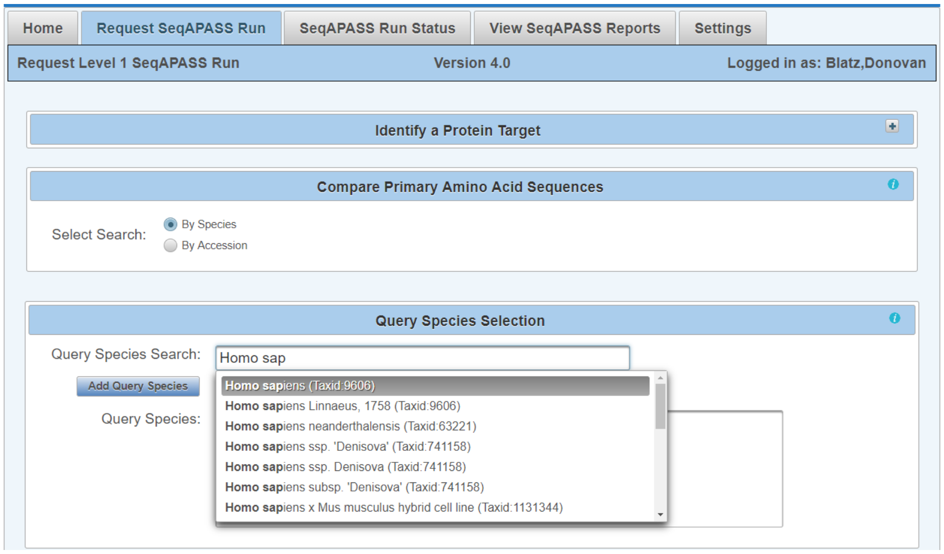


Request Level 1 SeqAPASS Run | Version 4.0 | Logged in as: Blatz,Donovan

Select Search:
● By Species
○ By Accession

Query "By Species"

Type the name of the query species of interest in the "Query Species Search" text box. The species common name, scientific name, or Taxid (ID number derived from the NCBI taxonomy database) may be typed into the search bar. This is the species you would like to compare all other species to. The search bar has an auto-complete function and will generate a list of species with corresponding Taxid. When text is typed into the search bar, the auto-complete function queries the database in the order of "starts with" then "contains." If an integer is typed in the search bar the auto-complete function queries the database in the order of "Taxid", "starts with", then "contains."



Logged in as: Blatz,Donovan

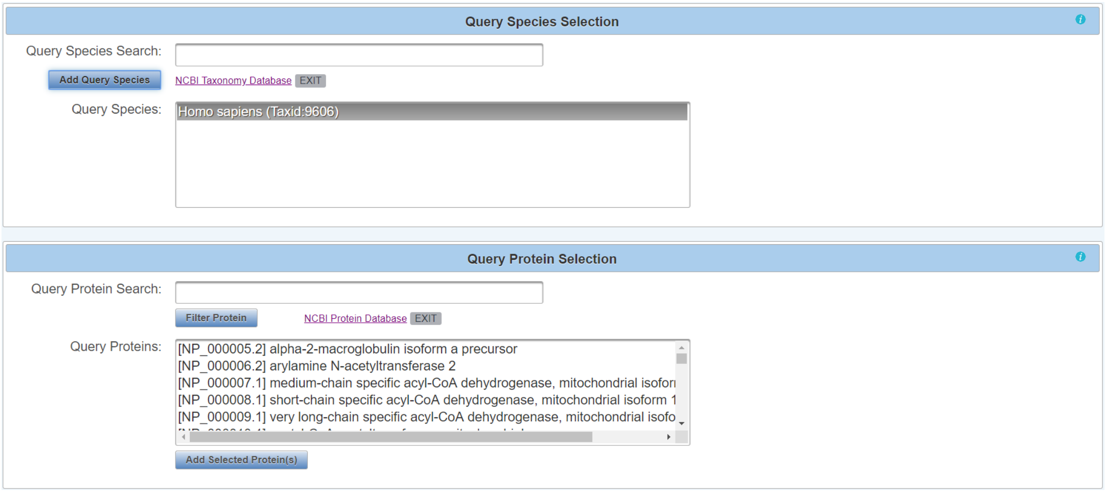
Select Search: ● By Species ○ By Accession

Query Species Search: Homo sap

Homo sapiens (Taxid:9606)
Homo sapiens Linnaeus, 1758 (Taxid:9606)
Homo sapiens neanderthalensis (Taxid:63221)
Homo sapiens ssp. 'Denisova' (Taxid:741158)
Homo sapiens ssp. Denisova (Taxid:741158)
Homo sapiens subsp. 'Denisova' (Taxid:741158)
Homo sapiens x Mus musculus hybrid cell line (Taxid:1131344)

Note: The user can also use the NCBI taxonomy database to identify query species using the NCBI link on the right-hand side of the "Add Query Species" button.

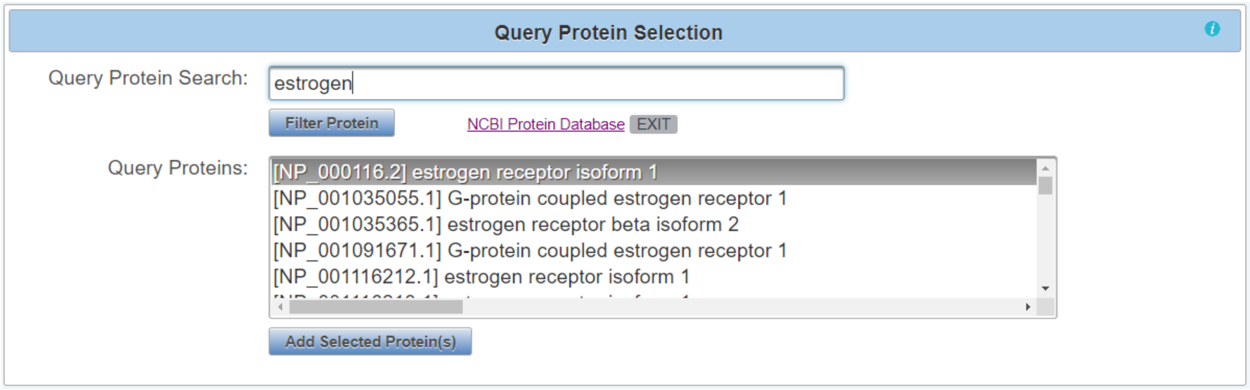
Select species of interest by clicking on the name in the drop-down box. Once species is selected, click "Add Query Species" button. This advances the species of interest to the "Query Species" box and fills the "Query Proteins" box with all available protein sequences for that species from the NCBI protein database (although the box only displays the initial 200 proteins/species based on lowest numerical accession number). The protein list includes the protein NCBI accession, protein name, and species scientific name.



Query Species: Homo sapiens (Taxid:9606)

Query Proteins:
[NP_000005.2] alpha-2-macroglobulin isoform a precursor
[NP_000006.2] arylamine N-acetyltransferase 2
[NP_000007.1] medium-chain specific acyl-CoA dehydrogenase, mitochondrial isofor
[NP_000008.1] short-chain specific acyl-CoA dehydrogenase, mitochondrial isoform 1
[NP_000009.1] very long-chain specific acyl-CoA dehydrogenase, mitochondrial isofo

To filter the query protein list, type the query protein name or partial name in the "Query Protein Search" box and click the "Filter Protein" button. This action will filter the protein list in the "Query Proteins" box to only display proteins that contain the user defined text (this search query does not contain an autofill feature due to the filter feature). Proteins will be listed in alphabetical order based on NCBI accession Example: typing "estrogen" retrieves all proteins that contain the word "estrogen" in the protein name (the user can scroll to identify proteins of interest).

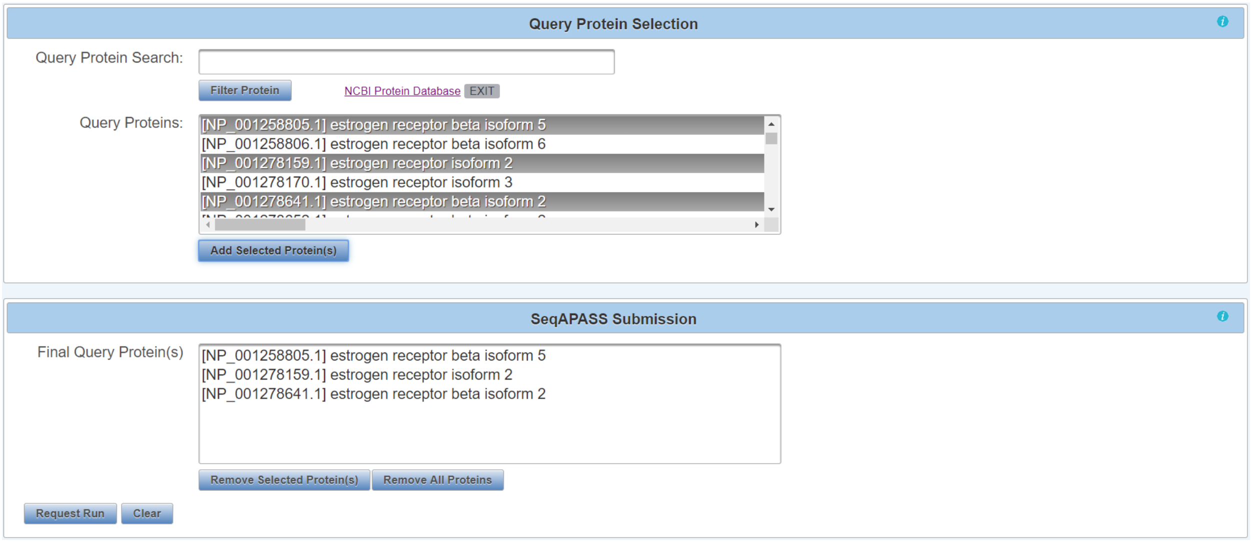


Query Protein Search: estrogen

Query Proteins:
[NP_000116.2] estrogen receptor isoform 1
[NP_001035055.1] G-protein coupled estrogen receptor 1
[NP_001035365.1] estrogen receptor beta isoform 2
[NP_001091671.1] G-protein coupled estrogen receptor 1
[NP_001116212.1] estrogen receptor isoform 1

Note: To explore details associated with a protein of interest, click the "NCBI Protein Database" link to the right of the "Filter Protein" button to open NCBI proteins database (See **SeqAPASS Documentation** section of user guide for details about searching for query proteins using NCBI database).

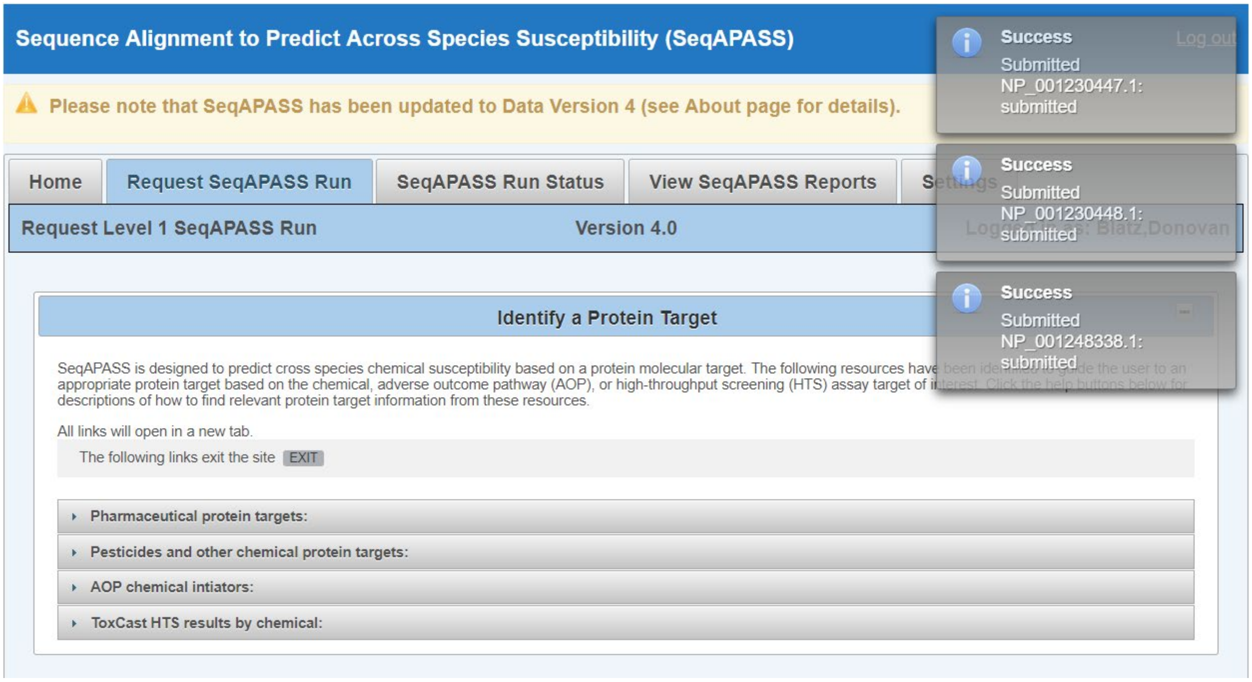
Highlight the protein or proteins of interest (Ctrl left click to select multiple proteins) in the "Query Proteins" box and click "Add Selected Protein(s)" button. This moves the protein(s) of interest to the "Final Query Protein(s)" box. To remove proteins from the "Final Query Protein(s)" box highlight those to be removed and click the "Remove Selected Protein(s)" button. Select "Remove All Proteins" to discard all proteins from "Final Query Protein(s)" box. The clear button removes all information previously entered on the "Request SeqAPASS Run" page.



Query Proteins:
[NP_001258805.1] estrogen receptor beta isoform 5
[NP_001258806.1] estrogen receptor beta isoform 6
[NP_001278159.1] estrogen receptor isoform 2
[NP_001278170.1] estrogen receptor isoform 3
[NP_001278641.1] estrogen receptor beta isoform 2

Final Query Protein(s)
[NP_001258805.1] estrogen receptor beta isoform 5
[NP_001278159.1] estrogen receptor isoform 2
[NP_001278641.1] estrogen receptor beta isoform 2

Once the user identifies the protein(s) to be queried, select “Request Run.” A message will briefly appear in upper right-hand corner of the screen for 10 seconds to alert the user of the request status.



Please note that SeqAPASS has been updated to Data Version 4 (see About page for details).

Success
Submitted NP_001230447.1: submitted

Success
Submitted NP_001230448.1: submitted

Success
Submitted NP_001248338.1: submitted

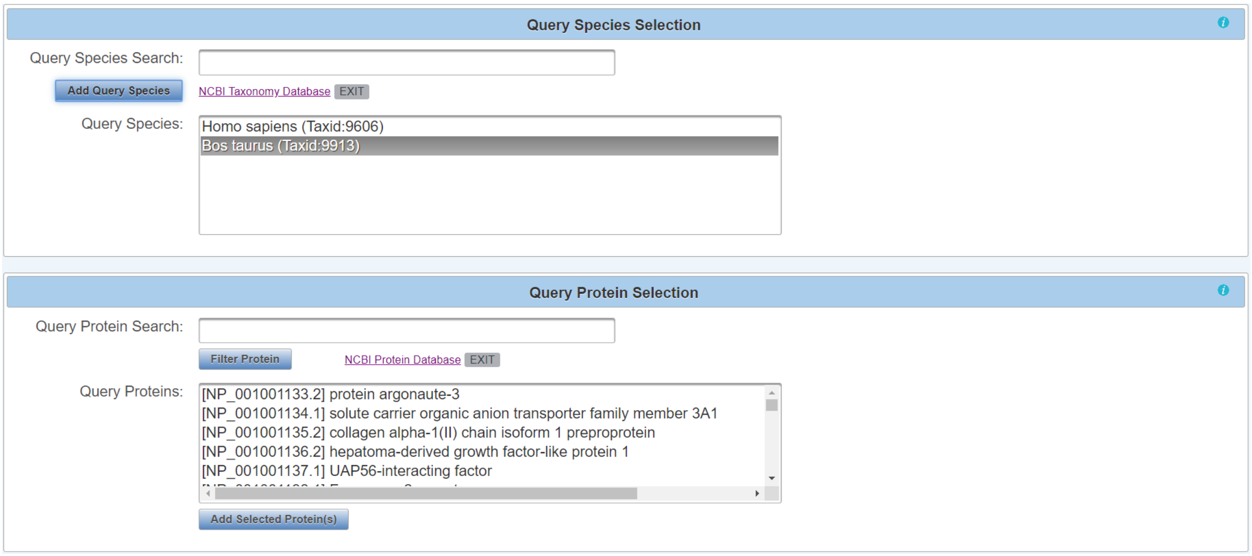
SeqAPASS is designed to predict cross species chemical susceptibility based on a protein molecular target. The following resources have been identified to guide the user to an appropriate protein target based on the chemical, adverse outcome pathway (AOP), or high-throughput screening (HTS) assay target of interest. Click the help buttons below for descriptions of how to find relevant protein target information from these resources.

All links will open in a new tab.

The following links exit the site

Multiple proteins can be added to the final list for multiple SeqAPASS runs. If another query species is desired, return to “Query Species Search” to select the next species. Follow the process described above for selecting the proteins associated with this species. The proteins populated in the “Query Proteins” box will always be associated with the species highlighted in the “Query Species” box.

Note: In the current version of SeqAPASS, ***PLEASE do not request more than 10 query proteins at a time*** to avoid longer wait times for the completion of a run.



Query Species Search:

Query Species:
Homo sapiens (Taxid:9606)
Bos taurus (Taxid:9913)

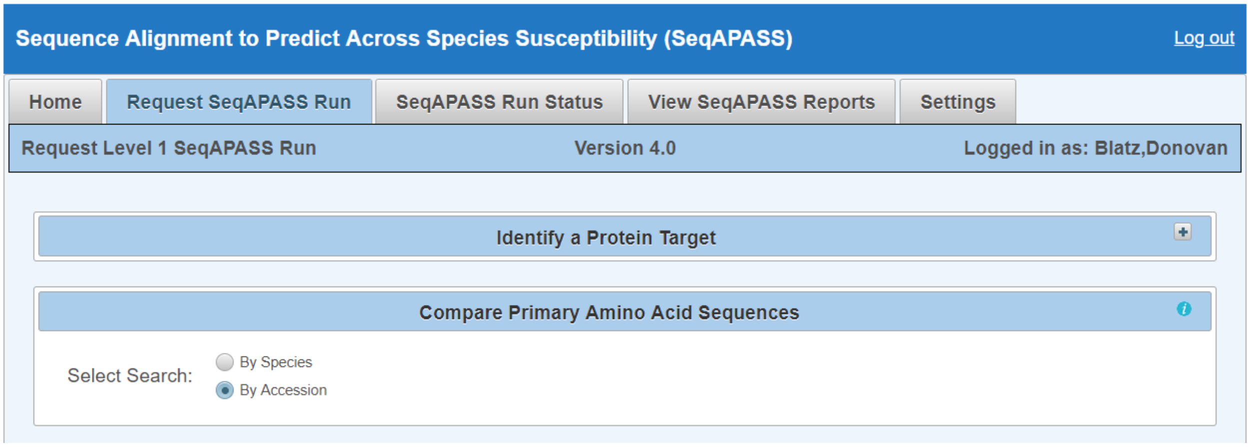
Query Protein Search:

Query Proteins:
[NP_001001133.2] protein argonaute-3
[NP_001001134.1] solute carrier organic anion transporter family member 3A1
[NP_001001135.2] collagen alpha-1(II) chain isoform 1 preproprotein
[NP_001001136.2] hepatoma-derived growth factor-like protein 1
[NP_001001137.1] UAP56-interacting factor

Note: A user may check the progress of the run by clicking on the “SeqAPASS Run Status” tab. (See **SeqAPASS Run Status** section of the user guide for more information)

Query "By Accession"

Users familiar with the NCBI database can utilize NCBI protein accessions (e.g., NP_000116.2) to query the SeqAPASS tool. This is done by selecting the "By Accession" radio button to the right of the "Select Search" text on the "Request SeqAPASS Run" page.

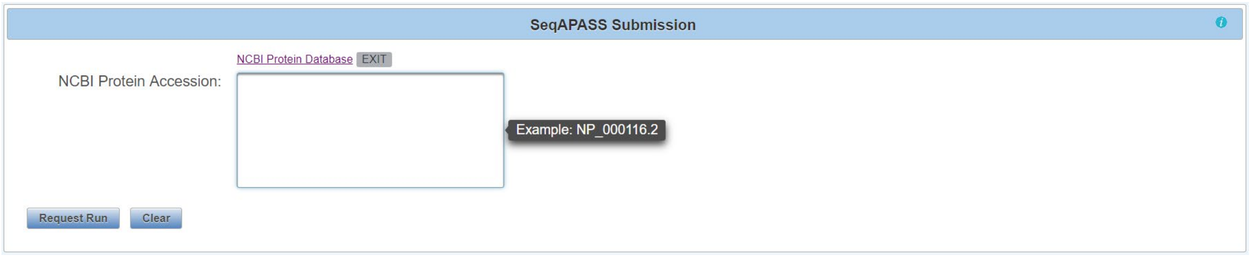


Logged in as: Blatz,Donovan

Select Search: ○ By Species ● By Accession

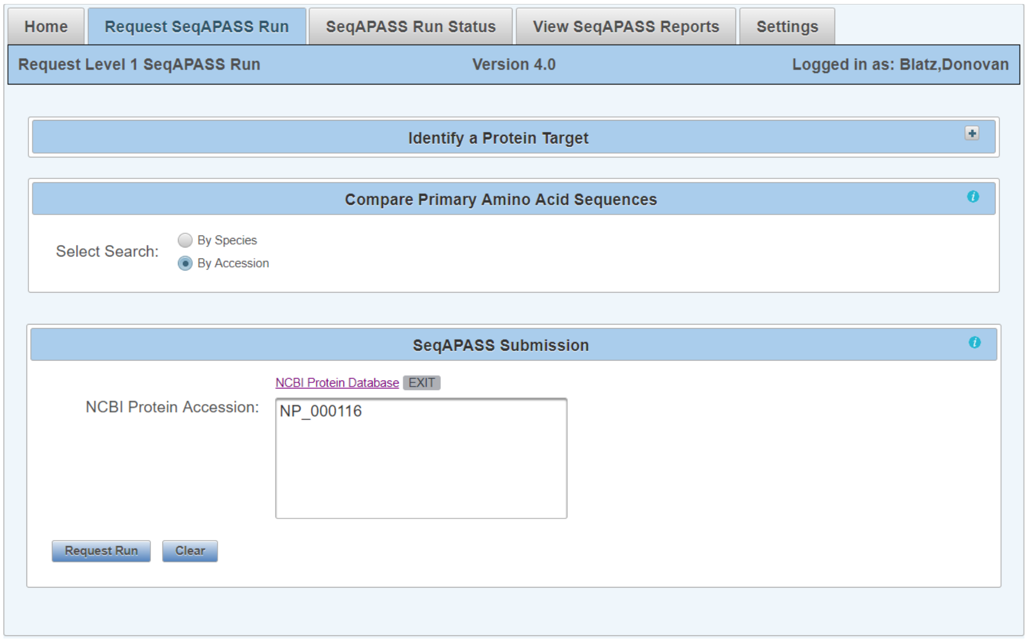
Upon selecting the "By Accession" radio button, a new query page will be displayed. Type the NCBI protein accession (e.g., NP_000116.2) for the protein of interest (this Accession comes from the NCBI protein database; See "**SeqAPASS Documentation**" for details) in the "NCBI Protein Accession" box. If desired, more than one NCBI Accession may be entered into the "NCBI Protein Accession" box by clicking the enter key after each additional NCBI Accession entry.

Upon clicking the "NCBI Protein Accession" text box, a pop-up message will appear in the middle of the text box, to provide an example for the proper format of Accessions to be entered.



Example: NP_000116.2

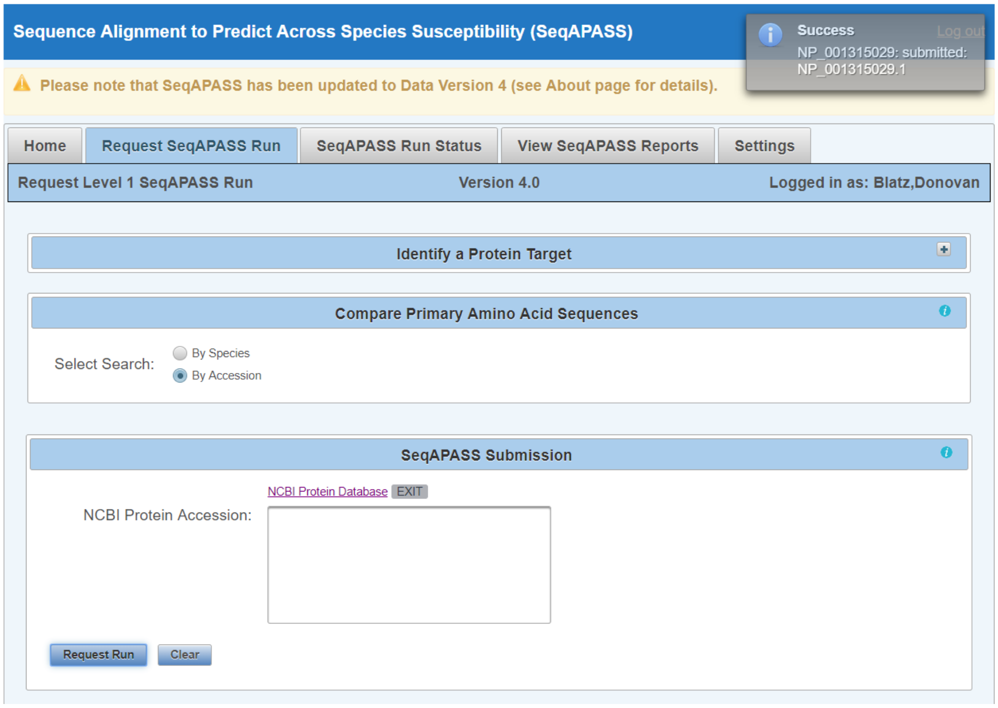
Note: To avoid longer wait times for the completion of a run, in the current version of SeqAPASS, ***please do not request more than 10 NCBI Accessions at a time***.



Select Search: ○ By Species ● By Accession

NCBI Protein Accession: NP_000116

After the NCBI accession(s) of interest have been typed in the “NCBI Protein Accession” box, click the “Request Run” button. To remove proteins from the “NCBI Protein Accession” box click the “Clear” button. A message will briefly appear in the upper right-hand corner of the screen to alert the user of their run request status.



Success
NP_001315029: submitted: NP_001315029.1

Please note that SeqAPASS has been updated to Data Version 4 (see About page for details).

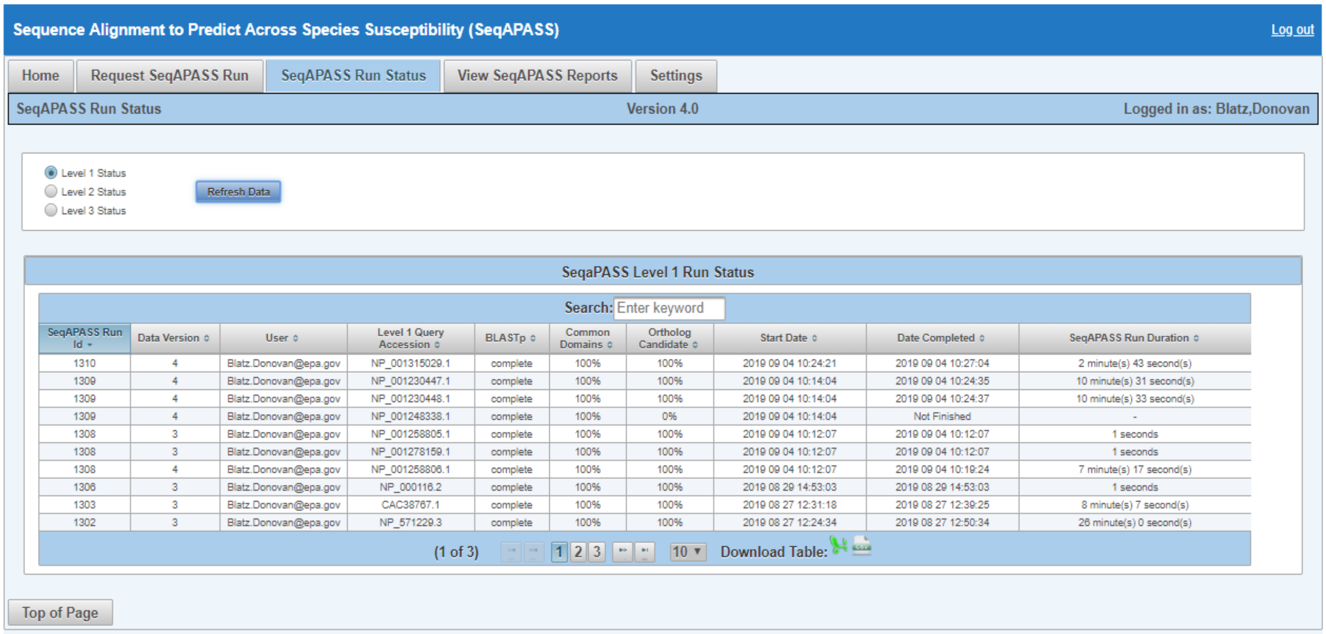
Select Search: ○ By Species ● By Accession

NCBI Protein Accession:

Note: All NCBI Accessions can include the version number (one digit after the decimal place, e.g., NP_000116.2). Otherwise, if the version is not included, the most recent version of the accession will be queried automatically.

SeqAPASS Run Status

Level 1 SeqAPASS (primary amino acid sequence comparisons) status is displayed as the default. The Accession in the column "Level 1 Query Accession" is that selected and queried by the user. For a query to finish it must display "complete" in the BLASTp column, 100% in the "Common Domains" column, and 100% in the "Ortholog Candidate" column. The "Common Domains" column displays the % completion for running Reverse Position Specific (RPS)-BLAST (Default E-value of ≤0.01) on the Accessions from the Level 1 Full Report. RPS-BLAST, and therefore "Common Domains" status, will take the longest to complete. The "Ortholog Candidate" column displays the % completion for running a reciprocal best hit BLAST evaluation for each hit sequence. The status for the "BLASTp" column is described as "started," "analyzing," or "complete." If the user's successfully submitted query has entered the run queue, the position of the submitted query in the queue will be indicated in the column (e.g., 2nd in queue). The "Common Domains" and "Ortholog Candidate" columns will also describe the position of the user's submitted query in the run queue. Once the run has begun processing, the % completed for RPS-BLAST or reciprocal best hit BLAST, respectively, will be displayed. Please see example below:



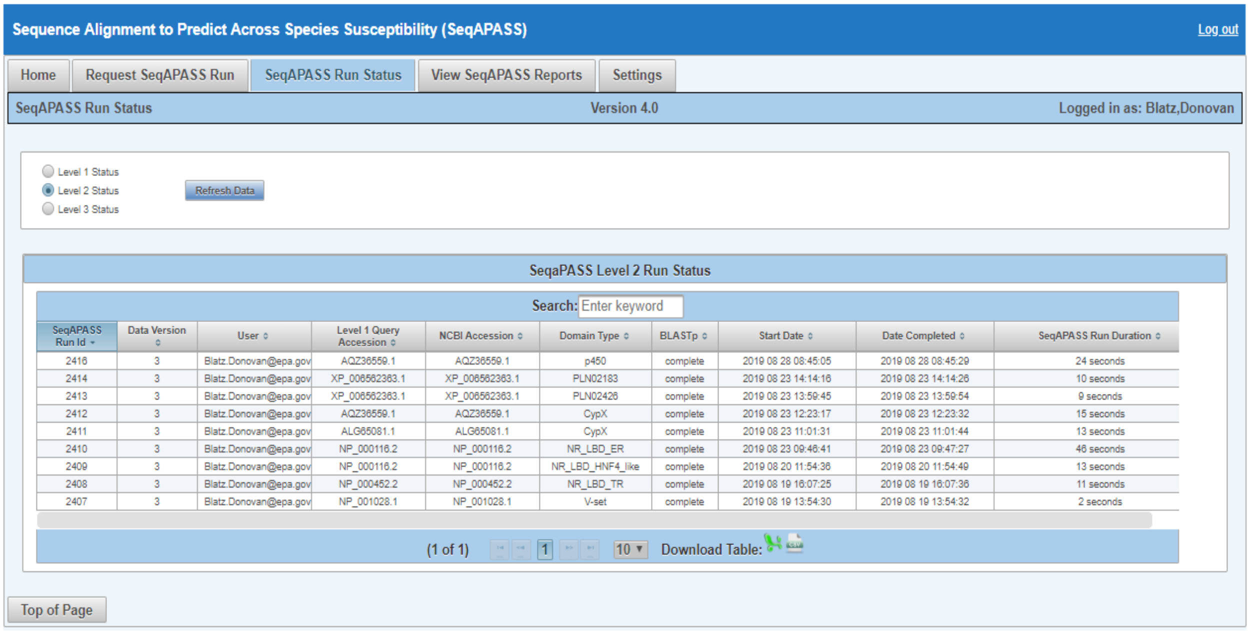
SeqaPASS Level 1 Run Status

SeqAPASS Run Id	Data Version	User	Level 1 Query Accession	BLASTp	Common Domains	Ortholog Candidate	Start Date	Date Completed	SeqAPASS Run Duration
1310	4	Blatz.Donovan@epa.gov	NP_001315029.1	complete	100%	100%	2019 09 04 10:24:21	2019 09 04 10:27:04	2 minute(s) 43 second(s)
1309	4	Blatz.Donovan@epa.gov	NP_001230447.1	complete	100%	100%	2019 09 04 10:14:04	2019 09 04 10:24:35	10 minute(s) 31 second(s)
1309	4	Blatz.Donovan@epa.gov	NP_001230448.1	complete	100%	100%	2019 09 04 10:14:04	2019 09 04 10:24:37	10 minute(s) 33 second(s)
1309	4	Blatz.Donovan@epa.gov	NP_001248338.1	complete	100%	0%	2019 09 04 10:14:04	Not Finished	-
1308	3	Blatz.Donovan@epa.gov	NP_001258805.1	complete	100%	100%	2019 09 04 10:12:07	2019 09 04 10:12:07	1 seconds
1308	3	Blatz.Donovan@epa.gov	NP_001278159.1	complete	100%	100%	2019 09 04 10:12:07	2019 09 04 10:12:07	1 seconds
1308	4	Blatz.Donovan@epa.gov	NP_001258806.1	complete	100%	100%	2019 09 04 10:12:07	2019 09 04 10:19:24	7 minute(s) 17 second(s)
1306	3	Blatz.Donovan@epa.gov	NP_000116.2	complete	100%	100%	2019 08 29 14:53:03	2019 08 29 14:53:03	1 seconds
1303	3	Blatz.Donovan@epa.gov	CAC38767.1	complete	100%	100%	2019 08 27 12:31:18	2019 08 27 12:39:25	8 minute(s) 7 second(s)
1302	3	Blatz.Donovan@epa.gov	NP_571229.3	complete	100%	100%	2019 08 27 12:24:34	2019 08 27 12:50:34	26 minute(s) 0 second(s)

The user can view the status of requested SeqAPASS runs. Each Run is assigned a unique "SeqAPASS Run Id." A ***Run*** is considered a query that was requested either individually or as a batch in the "Request SeqAPASS Run" tab. The user can view run start and end dates/times, and the duration of the run. (See **Search, View, and Download Data Tables** section of user guide for more information). The "Data Version" column indicates which version of NCBI data is being used (See "About" page for details on Data Versions)

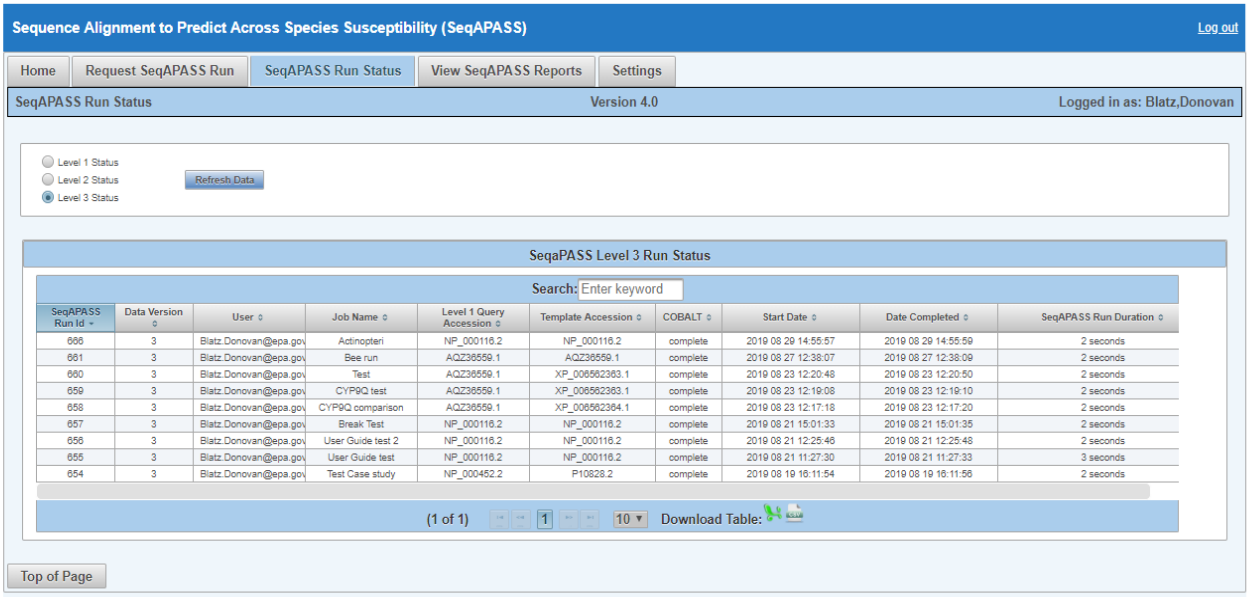
The user is also able to view the status of Level 2 (Functional domain(s)) and Level 3 (individual amino acid residue alignments).

View Level 2 Status by selecting the radio button. Also, while viewing the page, the user can click the "Refresh Data" button to refresh the data. "Level 1 Query Accession" column displays the NCBI accession selected and queried by the user. Please see below:



SeqAPASS Run Id	Data Version	User	Level 1 Query Accession	NCBI Accession	Domain Type	BLASTp	Start Date	Date Completed	SeqAPASS Run Duration
2416	3	Blatz.Donovan@epa.gov	AQZ36559.1	AQZ36559.1	p450	complete	2019 08 28 08:45:05	2019 08 28 08:45:29	24 seconds
2414	3	Blatz.Donovan@epa.gov	XP_006562363.1	XP_006562363.1	PLN02183	complete	2019 08 23 14:14:16	2019 08 23 14:14:26	10 seconds
2413	3	Blatz.Donovan@epa.gov	XP_006562363.1	XP_006562363.1	PLN02426	complete	2019 08 23 13:59:45	2019 08 23 13:59:54	9 seconds
2412	3	Blatz.Donovan@epa.gov	AQZ36559.1	AQZ36559.1	CypX	complete	2019 08 23 12:23:17	2019 08 23 12:23:32	15 seconds
2411	3	Blatz.Donovan@epa.gov	ALG65081.1	ALG65081.1	CypX	complete	2019 08 23 11:01:31	2019 08 23 11:01:44	13 seconds
2410	3	Blatz.Donovan@epa.gov	NP_000116.2	NP_000116.2	NR_LBD_ER	complete	2019 08 23 09:46:41	2019 08 23 09:47:27	46 seconds
2409	3	Blatz.Donovan@epa.gov	NP_000116.2	NP_000116.2	NR_LBD_HNF4_like	complete	2019 08 20 11:54:36	2019 08 20 11:54:49	13 seconds
2408	3	Blatz.Donovan@epa.gov	NP_000452.2	NP_000452.2	NR_LBD_TR	complete	2019 08 19 16:07:25	2019 08 19 16:07:36	11 seconds
2407	3	Blatz.Donovan@epa.gov	NP_001028.1	NP_001028.1	V-set	complete	2019 08 19 13:54:30	2019 08 19 13:54:32	2 seconds

View Level 3 Status by selecting the radio button. "Level 1 Query Accession" column displays the NCBI accession selected and queried by the user. The "Job Name" is the user defined name chosen to describe the Level 3 alignment. Also, while viewing the page, the user can click the "Refresh Data" button to refresh the data. Please see below:

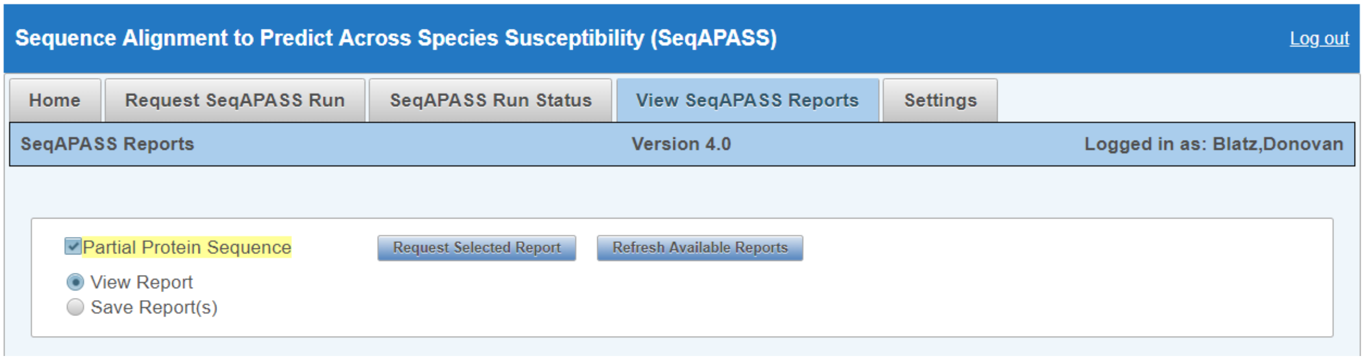


SeqAPASS Run Id	Data Version	User	Job Name	Level 1 Query Accession	Template Accession	COBALT	Start Date	Date Completed	SeqAPASS Run Duration
666	3	Blatz.Donovan@epa.gov	Actinopteri	NP_000116.2	NP_000116.2	complete	2019 08 29 14:55:57	2019 08 29 14:55:59	2 seconds
661	3	Blatz.Donovan@epa.gov	Bee run	AQZ36559.1	AQZ36559.1	complete	2019 08 27 12:38:07	2019 08 27 12:38:09	2 seconds
660	3	Blatz.Donovan@epa.gov	Test	AQZ36559.1	XP_006562363.1	complete	2019 08 23 12:20:48	2019 08 23 12:20:50	2 seconds
659	3	Blatz.Donovan@epa.gov	CYP9Q test	AQZ36559.1	XP_006562363.1	complete	2019 08 23 12:19:08	2019 08 23 12:19:10	2 seconds
658	3	Blatz.Donovan@epa.gov	CYP9Q comparison	AQZ36559.1	XP_006562364.1	complete	2019 08 23 12:17:18	2019 08 23 12:17:20	2 seconds
657	3	Blatz.Donovan@epa.gov	Break Test	NP_000116.2	NP_000116.2	complete	2019 08 21 15:01:33	2019 08 21 15:01:35	2 seconds
656	3	Blatz.Donovan@epa.gov	User Guide test 2	NP_000116.2	NP_000116.2	complete	2019 08 21 12:25:46	2019 08 21 12:25:48	2 seconds
655	3	Blatz.Donovan@epa.gov	User Guide test	NP_000116.2	NP_000116.2	complete	2019 08 21 11:27:30	2019 08 21 11:27:33	3 seconds
654	3	Blatz.Donovan@epa.gov	Test Case study	NP_000452.2	P10828.2	complete	2019 08 19 16:11:54	2019 08 19 16:11:56	2 seconds

To return to previous tabs click on "Home," "Request SeqAPASS Run," or "SeqAPASS Run Status" tabs.

View SeqAPASS Reports Tab

The “View SeqAPASS Reports” tab provides a table of completed SeqAPASS runs. From this page the user can choose to either “View Report” or “Save Report(s).”

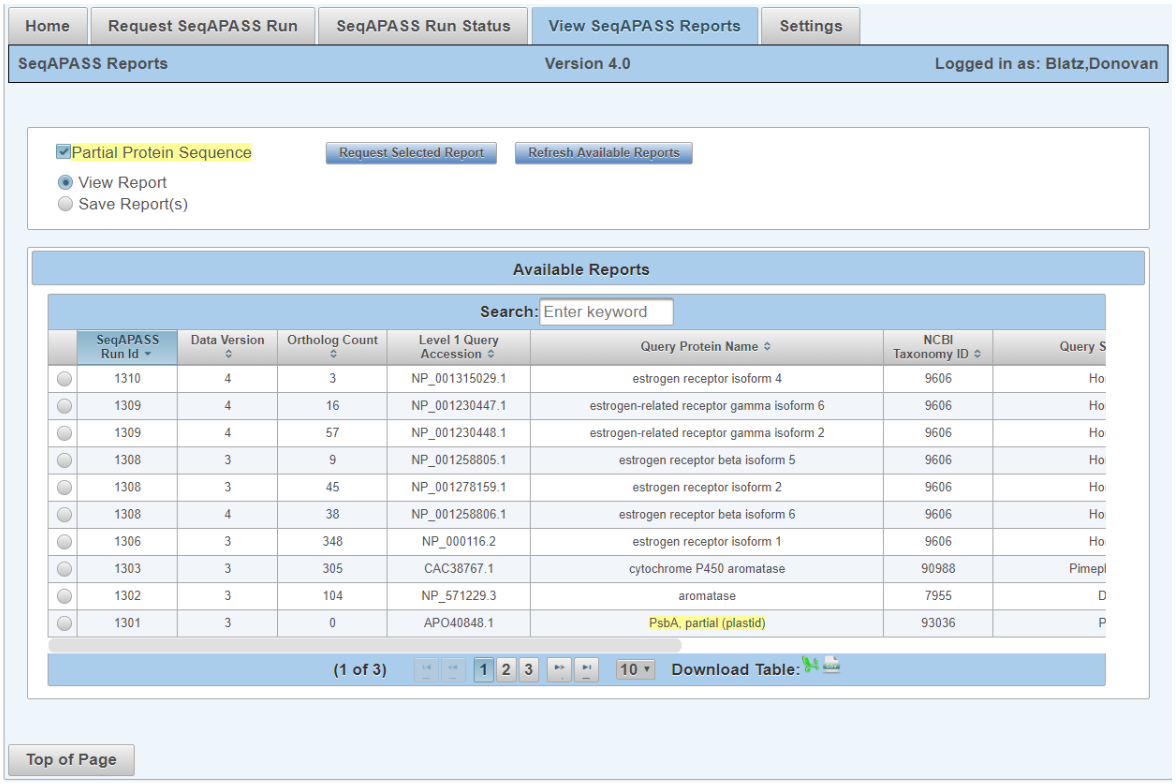


☑ Partial Protein Sequence

● View Report
○ Save Report(s)

The completed runs, by default, are listed in the order in which they were completed, with the most recent runs at the top. The table includes information for each run, such as SeqAPASS Run ID (unique for every run regardless of if it is the same protein/species combination ran twice), Data Version, Ortholog Count (number of orthologs detected from the aligned hit sequences in Level 1; see Detailed Documentation page 79), NCBI Accession, Query Protein Name, taxonomy information for the query species, and the date/time of run completion.

While viewing the page, the user can click the “Refresh Available Reports” button to refresh the table with additional completed runs. Partial protein sequences are highlighted in yellow as illustrated in the example below. (See **Search, View, and Download Data Tables** section of user guide for more information).



☑ Partial Protein Sequence

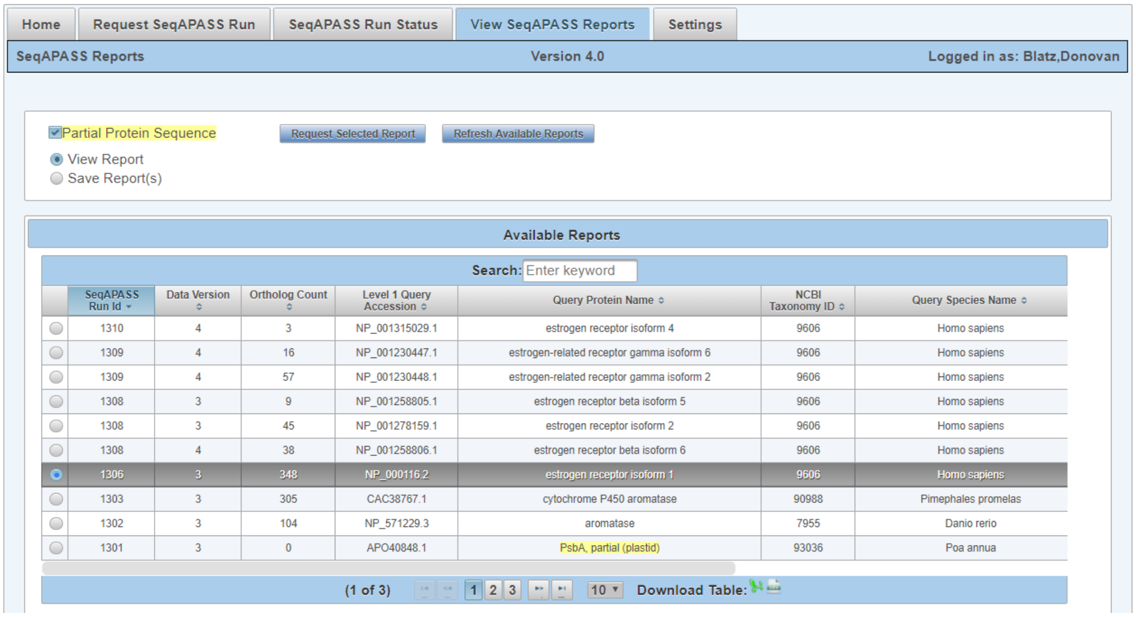
● View Report
○ Save Report(s)

	SeqAPASS Run Id	Data Version	Ortholog Count	Level 1 Query Accession	Query Protein Name	NCBI Taxonomy ID	Query S
○	1310	4	3	NP_001315029.1	estrogen receptor isoform 4	9606	Ho
○	1309	4	16	NP_001230447.1	estrogen-related receptor gamma isoform 6	9606	Ho
○	1309	4	57	NP_001230448.1	estrogen-related receptor gamma isoform 2	9606	Ho
○	1308	3	9	NP_001258805.1	estrogen receptor beta isoform 5	9606	Ho
○	1308	3	45	NP_001278159.1	estrogen receptor isoform 2	9606	Ho
○	1308	4	38	NP_001258806.1	estrogen receptor beta isoform 6	9606	Ho
○	1306	3	348	NP_000116.2	estrogen receptor isoform 1	9606	Ho
○	1303	3	305	CAC38767.1	cytochrome P450 aromatase	90988	Pimepl
○	1302	3	104	NP_571229.3	aromatase	7955	D
○	1301	3	0	APO40848.1	PsbA, partial (plastid)	93036	P

View Report

To select a completed run and view Level 1 data, select the corresponding radio button in the first column of the table and click “Request Selected Report.” This will open the Level 1 page to view the Level 1 data and to set up queries for Level 2 and Level 3.

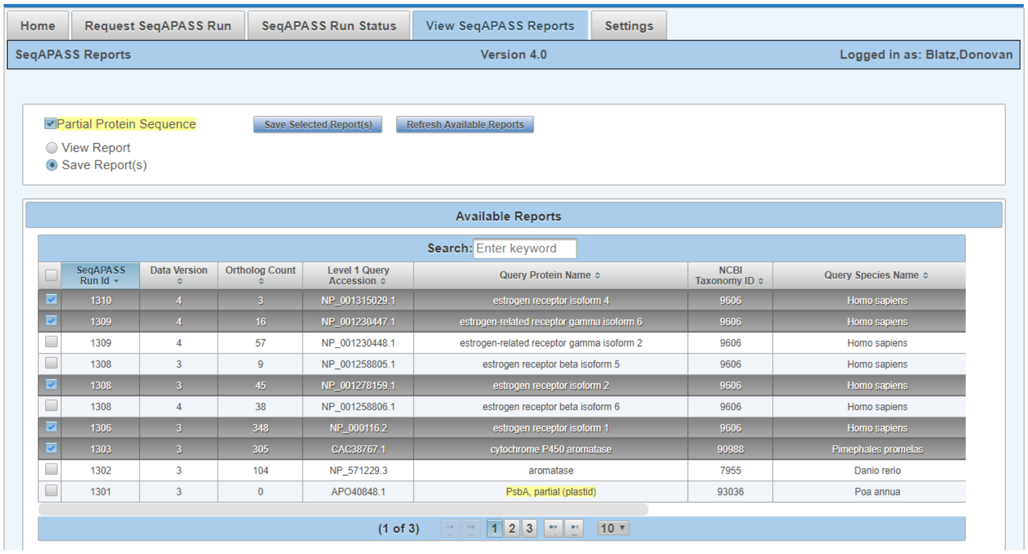
<u>Note:</u> The user ***MUST*** select a radio button ***PRIOR*** to clicking “Request Selected Report.” If the user fails to select a radio button and clicks “Request Selected Report” a Spinning Wheel will appear and disappear, and no completed run will be opened. Further, ***there is no pop-up message*** indicating that the user did not select a radio button.



	SeqAPASS Run Id	Data Version	Ortholog Count	Level 1 Query Accession	Query Protein Name	NCBI Taxonomy ID	Query Species Name
○	1310	4	3	NP_001315029.1	estrogen receptor isoform 4	9606	Homo sapiens
○	1309	4	16	NP_001230447.1	estrogen-related receptor gamma isoform 6	9606	Homo sapiens
○	1309	4	57	NP_001230448.1	estrogen-related receptor gamma isoform 2	9606	Homo sapiens
○	1308	3	9	NP_001258805.1	estrogen receptor beta isoform 5	9606	Homo sapiens
○	1308	3	45	NP_001278159.1	estrogen receptor isoform 2	9606	Homo sapiens
○	1308	4	38	NP_001258806.1	estrogen receptor beta isoform 6	9606	Homo sapiens
●	1306	3	348	NP_000116.2	estrogen receptor isoform 1	9606	Homo sapiens
○	1303	3	305	CAC38767.1	cytochrome P450 aromatase	90988	Pimephales promelas
○	1302	3	104	NP_571229.3	aromatase	7955	Danio rerio
○	1301	3	0	APO40848.1	PsbA, partial (plastid)	93036	Poa annua

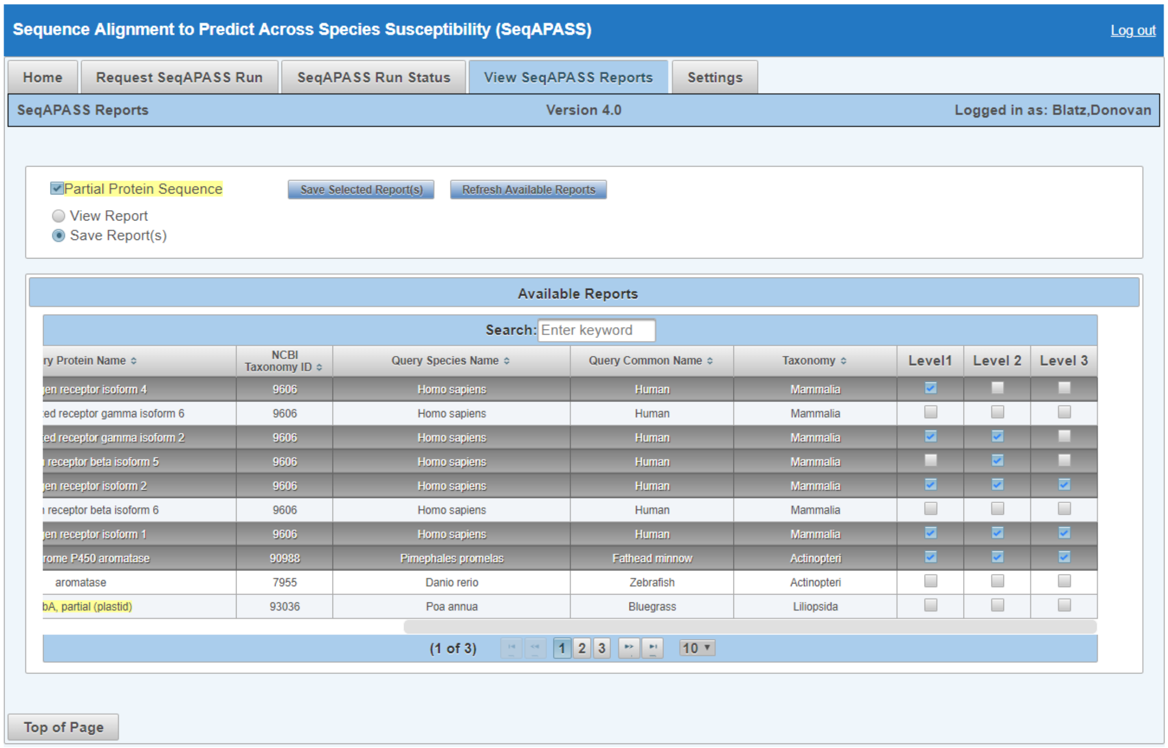
Save Report(s)

To download completed Level 1, 2, and/or 3 data, select the “Save Report(s)” radio button. Upon doing so the user can select which accession(s) to download by clicking the checkbox in the first column of the table associated with desired accession and click “Save Selected Report(s).”



☐	SeqAPASS Run Id	Data Version	Ortholog Count	Level 1 Query Accession	Query Protein Name	NCBI Taxonomy ID	Query Species Name
☑	1310	4	3	NP_001315029.1	estrogen receptor isoform 4	9606	Homo sapiens
☑	1309	4	16	NP_001230447.1	estrogen-related receptor gamma isoform 6	9606	Homo sapiens
☐	1309	4	57	NP_001230448.1	estrogen-related receptor gamma isoform 2	9606	Homo sapiens
☐	1308	3	9	NP_001258805.1	estrogen receptor beta isoform 5	9606	Homo sapiens
☑	1308	3	45	NP_001278159.1	estrogen receptor isoform 2	9606	Homo sapiens
☐	1308	4	38	NP_001258806.1	estrogen receptor beta isoform 6	9606	Homo sapiens
☑	1306	3	348	NP_000116.2	estrogen receptor isoform 1	9606	Homo sapiens
☑	1303	3	305	CAC38767.1	cytochrome P450 aromatase	90988	Pimephales promelas
☐	1302	3	104	NP_571229.3	aromatase	7955	Danio rerio
☐	1301	3	0	APO40848.1	PsbA, partial (plastid)	93036	Poa annua

The user can also deselect data that is not wanted in the download by scrolling to the far right of the table and deselecting the checkboxes for the different levels of the SeqAPASS analysis. By default, all available data for the selected accession will be downloaded in a zip file.



☑ Partial Protein Sequence

○ View Report
● Save Report(s)

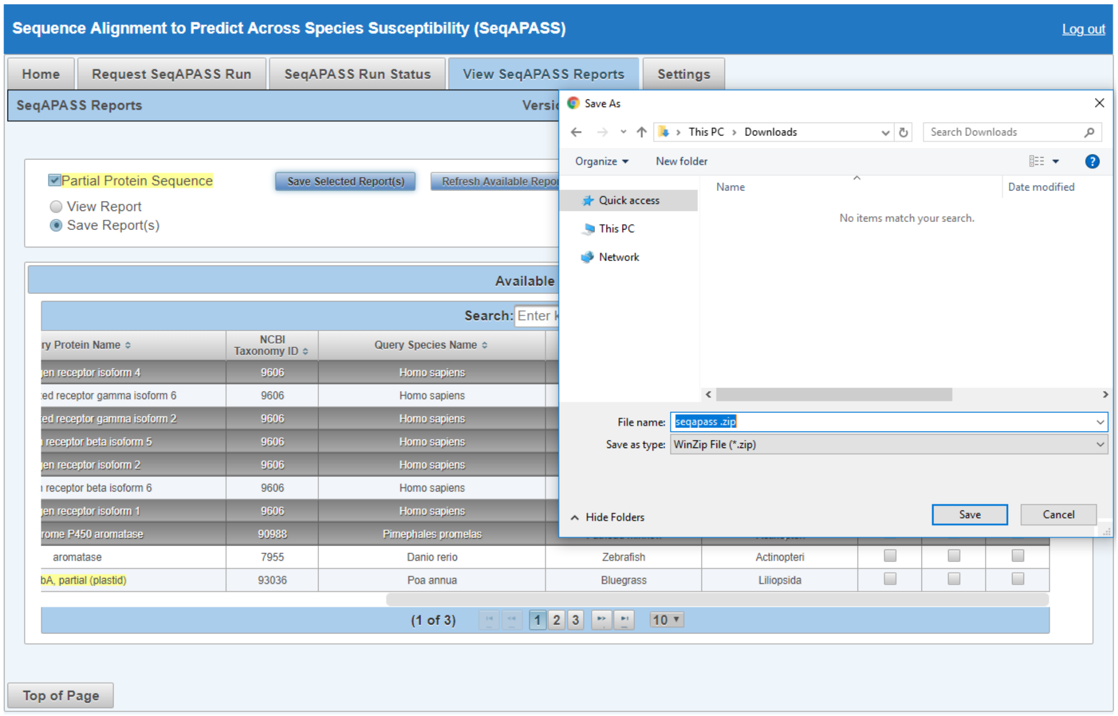
Available Reports

Search: Enter keyword

ry Protein Name	NCBI Taxonomy ID	Query Species Name	Query Common Name	Taxonomy	Level1	Level 2	Level 3
jen receptor isoform 4	9606	Homo sapiens	Human	Mammalia	☑	☐	☐
ted receptor gamma isoform 6	9606	Homo sapiens	Human	Mammalia	☐	☐	☐
ted receptor gamma isoform 2	9606	Homo sapiens	Human	Mammalia	☑	☑	☐
receptor beta isoform 5	9606	Homo sapiens	Human	Mammalia	☐	☑	☐
jen receptor isoform 2	9606	Homo sapiens	Human	Mammalia	☑	☑	☑
receptor beta isoform 6	9606	Homo sapiens	Human	Mammalia	☐	☐	☐
jen receptor isoform 1	9606	Homo sapiens	Human	Mammalia	☑	☑	☑
rome P450 aromatase	90988	Pimephales promelas	Fathead minnow	Actinopteri	☑	☑	☑
aromatase	7955	Danio rerio	Zebrafish	Actinopteri	☐	☐	☐
bA, partial (plastid)	93036	Poa annua	Bluegrass	Liliopsida	☐	☐	☐

(1 of 3) 1 2 3 10

A WinZip file will be created for all the selected Reports.

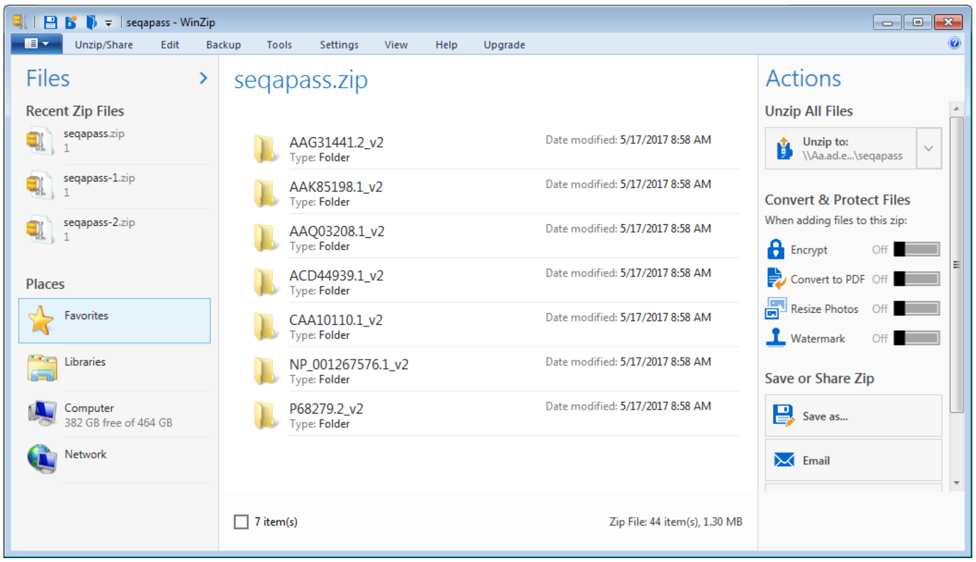


No items match your search.

File name: seqapass.zip

Save as type: WinZip File (*.zip)

A pop-up seqapass.zip file should appear with data files for each selected report. The naming convention is the NCBI Protein Accession and the Data Version (e.g., AAG31441.2_v2).

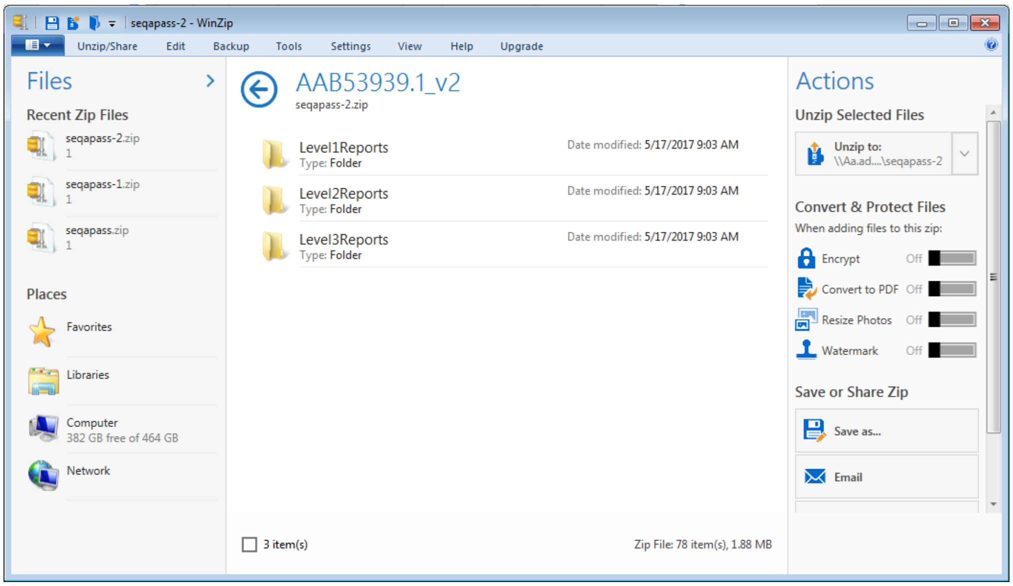


AAG31441.2_v2 Type: Folder Date modified: 5/17/2017 8:58 AM
AAK85198.1_v2 Type: Folder Date modified: 5/17/2017 8:58 AM
AAQ03208.1_v2 Type: Folder Date modified: 5/17/2017 8:58 AM
ACD44939.1_v2 Type: Folder Date modified: 5/17/2017 8:58 AM
CAA10110.1_v2 Type: Folder Date modified: 5/17/2017 8:58 AM
NP_001267576.1_v2 Type: Folder Date modified: 5/17/2017 8:58 AM
P68279.2_v2 Type: Folder Date modified: 5/17/2017 8:58 AM

☐ 7 item(s)
Zip File: 44 item(s), 1.30 MB

By clicking on one of the Reports for a Protein Accession_version, all available files for each Level of the SeqAPASS evaluation are available.

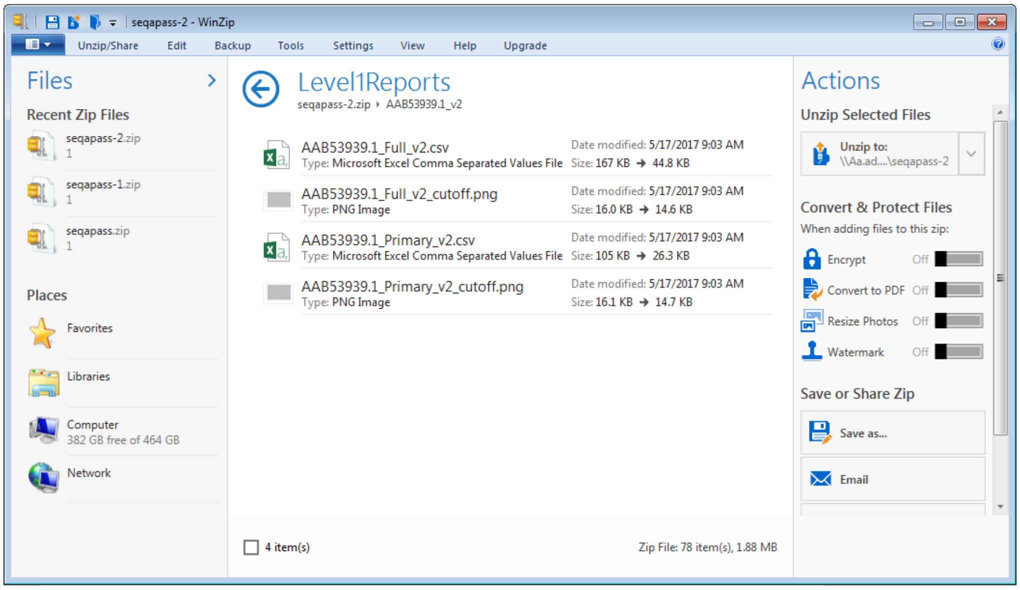
<u>Note:</u> This download includes default settings only. If susceptiblity cut-off or any defaults were manipulated on Level 1 or 2 pages they will ***NOT*** be downloaded here and can ***ONLY*** be downloaded directly from the Level 1 or Level 2 page where the setting was manipulated by the user. Also, data visualizations can ***ONLY*** be downloaded from the Level 1 and 2 pages. They ***DO NOT*** populate in the zip file folders.



Level1Reports Type: Folder Date modified: 5/17/2017 9:03 AM
Level2Reports Type: Folder Date modified: 5/17/2017 9:03 AM
Level3Reports Type: Folder Date modified: 5/17/2017 9:03 AM

☐ 3 item(s)
Zip File: 78 item(s), 1.88 MB

By selecting "Level1Reports", both full and primary reports are available as csv files as well as a graphic of the density plot for determining the susceptibility cut-off.

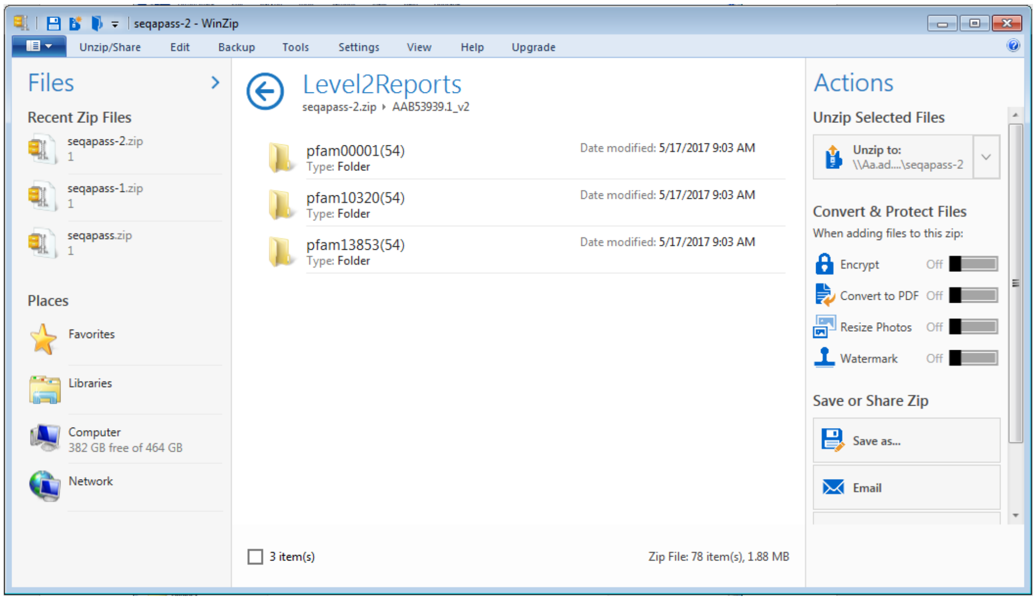


Level1Reports
seqapass-2.zip ▸ AAB53939.1_v2

Name	Type	Date modified	Size
AAB53939.1_Full_v2.csv	Microsoft Excel Comma Separated Values File	5/17/2017 9:03 AM	167 KB → 44.8 KB
AAB53939.1_Full_v2_cutoff.png	PNG Image	5/17/2017 9:03 AM	16.0 KB → 14.6 KB
AAB53939.1_Primary_v2.csv	Microsoft Excel Comma Separated Values File	5/17/2017 9:03 AM	105 KB → 26.3 KB
AAB53939.1_Primary_v2_cutoff.png	PNG Image	5/17/2017 9:03 AM	16.1 KB → 14.7 KB

☐ 4 item(s)
Zip File: 78 item(s), 1.88 MB

By selecting "Level2Reports", all completed domain comparisons will be available and named by NCBI domain accession with the starting amino acid residue position for the domain (e.g., pfam00001(54)).

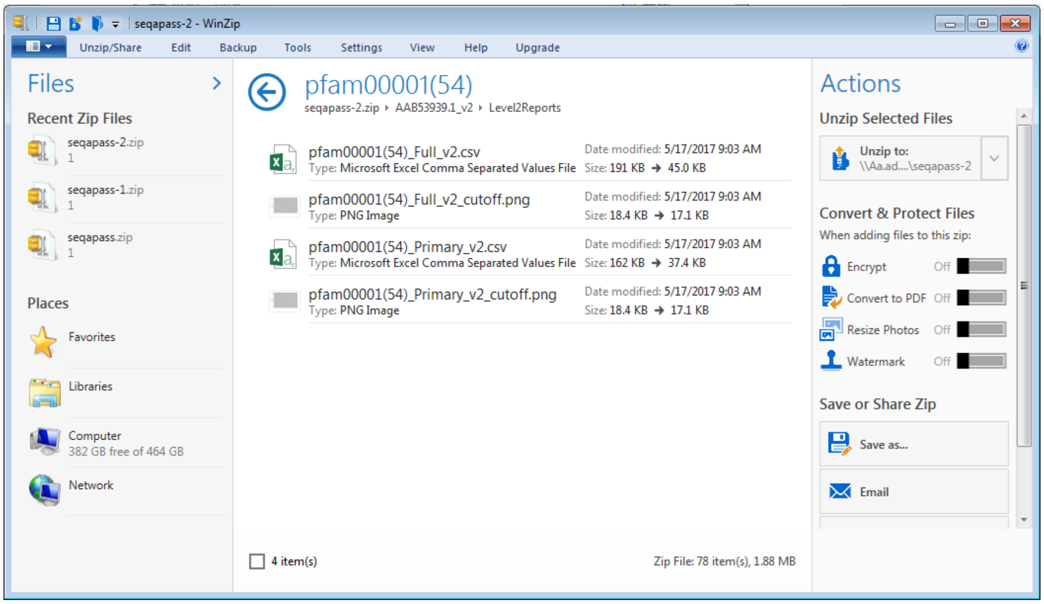


Level2Reports
seqapass-2.zip ▸ AAB53939.1_v2

Name	Type	Date modified
pfam00001(54)	Folder	5/17/2017 9:03 AM
pfam10320(54)	Folder	5/17/2017 9:03 AM
pfam13853(54)	Folder	5/17/2017 9:03 AM

☐ 3 item(s)
Zip File: 78 item(s), 1.88 MB

Upon selecting a domain file to view, both full and primary reports are available as csv files as well as a graphic of the density plot for determining the susceptibility cut-off.



pfam00001(54)
seqapass-2.zip ‣ AAB53939.1_v2 ‣ Level2Reports

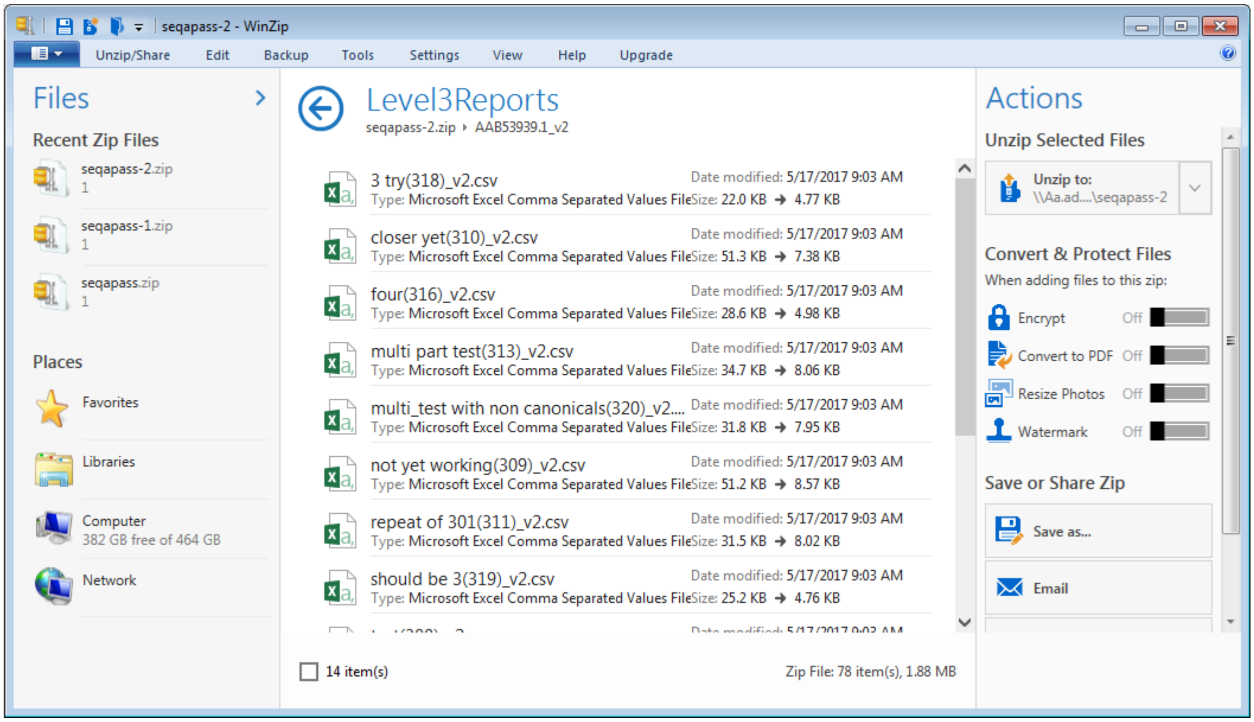
pfam00001(54)_Full_v2.csv — Type: Microsoft Excel Comma Separated Values File — Date modified: 5/17/2017 9:03 AM — Size: 191 KB → 45.0 KB
pfam00001(54)_Full_v2_cutoff.png — Type: PNG Image — Date modified: 5/17/2017 9:03 AM — Size: 18.4 KB → 17.1 KB
pfam00001(54)_Primary_v2.csv — Type: Microsoft Excel Comma Separated Values File — Date modified: 5/17/2017 9:03 AM — Size: 162 KB → 37.4 KB
pfam00001(54)_Primary_v2_cutoff.png — Type: PNG Image — Date modified: 5/17/2017 9:03 AM — Size: 18.4 KB → 17.1 KB

☐ 4 item(s)
Zip File: 78 item(s), 1.88 MB

Encrypt: Off
Convert to PDF: Off
Resize Photos: Off
Watermark: Off

By selecting “Level3Reports”, all user defined Level 3 alignments are available as csv.

Note: These csv files show the alignments across the entire sequence, not just those amino acid residues selected by the user.



Level3Reports
seqapass-2.zip ‣ AAB53939.1_v2

3 try(318)_v2.csv — Type: Microsoft Excel Comma Separated Values File — Date modified: 5/17/2017 9:03 AM — Size: 22.0 KB → 4.77 KB
closer yet(310)_v2.csv — Type: Microsoft Excel Comma Separated Values File — Date modified: 5/17/2017 9:03 AM — Size: 51.3 KB → 7.38 KB
four(316)_v2.csv — Type: Microsoft Excel Comma Separated Values File — Date modified: 5/17/2017 9:03 AM — Size: 28.6 KB → 4.98 KB
multi part test(313)_v2.csv — Type: Microsoft Excel Comma Separated Values File — Date modified: 5/17/2017 9:03 AM — Size: 34.7 KB → 8.06 KB
multi_test with non canonicals(320)_v2.... — Type: Microsoft Excel Comma Separated Values File — Date modified: 5/17/2017 9:03 AM — Size: 31.8 KB → 7.95 KB
not yet working(309)_v2.csv — Type: Microsoft Excel Comma Separated Values File — Date modified: 5/17/2017 9:03 AM — Size: 51.2 KB → 8.57 KB
repeat of 301(311)_v2.csv — Type: Microsoft Excel Comma Separated Values File — Date modified: 5/17/2017 9:03 AM — Size: 31.5 KB → 8.02 KB
should be 3(319)_v2.csv — Type: Microsoft Excel Comma Separated Values File — Date modified: 5/17/2017 9:03 AM — Size: 25.2 KB → 4.76 KB

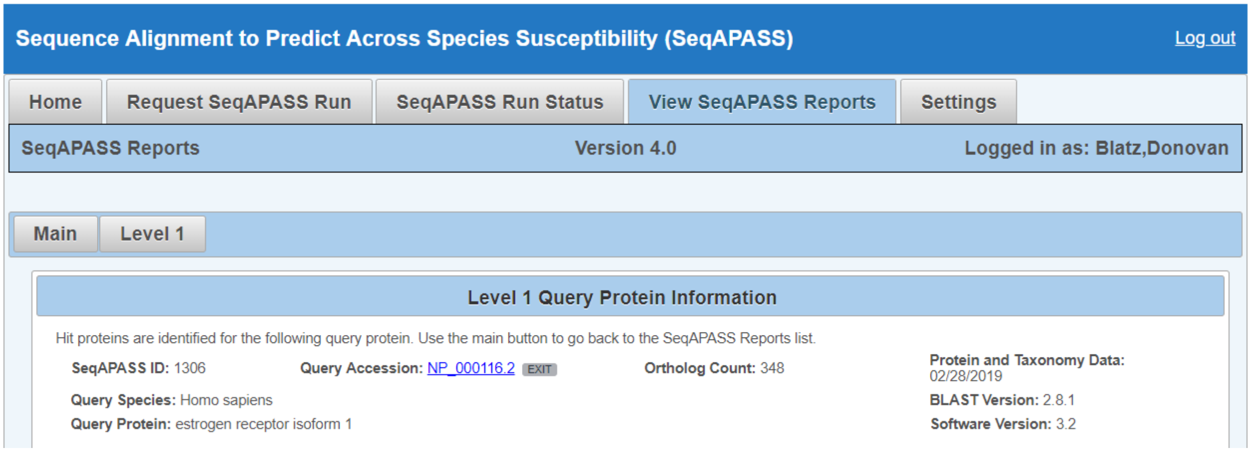
☐ 14 item(s)
Zip File: 78 item(s), 1.88 MB

Encrypt: Off
Convert to PDF: Off
Resize Photos: Off
Watermark: Off

Level 1: Primary Amino Acid Sequence Alignment

From the "View SeqAPASS Reports" tab, upon selecting a radio button and clicking "Request Selected Report" the Level 1 data will be displayed.

The "Level 1 Query Protein Information" box contains the SeqAPASS Run ID, Query Accession, Ortholog Count (# of hits identified as ortholog candidates to the query species protein sequence), NCBI Data updates ("Protein and Taxonomy Data:" displays the date that NCBI databases were downloaded and incorporated into the SeqAPASS database; BLAST Version: and Software Version: displays the version being used by the SeqAPASS tool for the selected data), Query Species, and Query Protein. Other information in this box will be described below.



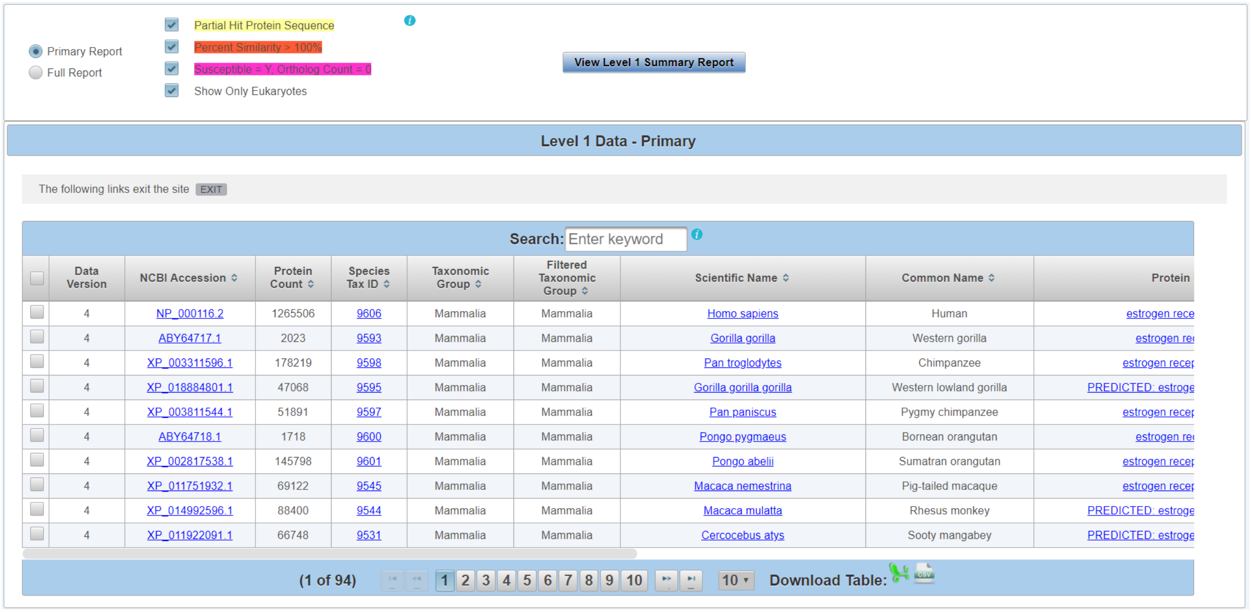
Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1306
Query Accession: NP_000116.2
Ortholog Count: 348
Protein and Taxonomy Data: 02/28/2019
Query Species: Homo sapiens
Query Protein: estrogen receptor isoform 1
BLAST Version: 2.8.1
Software Version: 3.2

The default table displayed at the bottom of the page is the "Primary Report", which includes query protein information in the first row below the column titles, followed by hit proteins whose sequences aligned with the query protein. The hit proteins are ordered from the highest to lowest percent similarity (Maximum percent similarity =100%). For each hit protein, Data version, NCBI Accession and species information is provided including the "Protein Count" which indicates the number of protein records per species in the NCBI protein database, taxonomic information (See **Primary Report Settings** section below in user guide for more detail on "Taxonomic Group" versus "Filtered Taxonomic Group" columns), and species names. Also included are the NCBI protein accession, protein name, BLASTp bitscore (describes overall quality of the alignment, See NCBI BLASTp tutorials), and percent similarity ([hit bitscore/query bitscore]*100). If the hit protein has been identified as an ortholog candidate (using reciprocal best hit blast method), it will be noted with a "Y" for yes or if not an ortholog candidate, a "N", for no. If the hit protein is predicted to be susceptible according to the susceptibility cut-off criteria, that will also be noted with a "Y" for yes or alternatively an "N" for no. The date the analysis was completed is also identified. The data also includes a column describing the number of ortholog candidates identified using the reciprocal best hit BLAST method. The susceptibility cut-off is also listed in a column. The cut-off is determined through identifying local minimums in the density plot of the percent similarity values for the primary report data set and evaluation of ortholog candidates. Additionally, there is a column that identifies if the species is a Eukaryote noted with a "Y" for yes or alternatively an "N" for no. Links out to the NCBI Protein Database, NCBI Taxonomy Database, and ECOTOX Knowledgebase (specific to the data row) are embedded in the Level 1 data table for "NCBI Accession," "Species Tax ID," "Scientific Name," "Protein Name", and "ECOTOX" columns. (See **Search, View, and Download Data Tables** section of user guide for more information).

Default highlights identify partial protein sequences, sequences with a bitscore higher than the query sequence and therefore percent similarity greater than 100% (commonly synthetic constructs), and when zero ortholog candidates are identified (in this case a user should consider a different query sequence or check the full report). Please see **Susceptibility Cutoff Box for Level 1** section of user guide for details when no orthologs are detected. Additionally, the default setting for the report shows only eukaryote data if a eukaryote is selected as the query protein, excluding prokaryote data from the table with the "Show Only Eukaryotes" checkbox checked. To view prokaryote data, deselect this checkbox. If a prokaryote is selected as the query protein, the default setting will include both eukaryote and prokaryote data and the "Show Only Eukaryotes" checkbox will not be selected. To limit the data to eukaryotes only, the user would check the "Show Only Eukaryotes" checkbox.

Columns in left side of table:



● Primary Report
○ Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

Level 1 Data - Primary

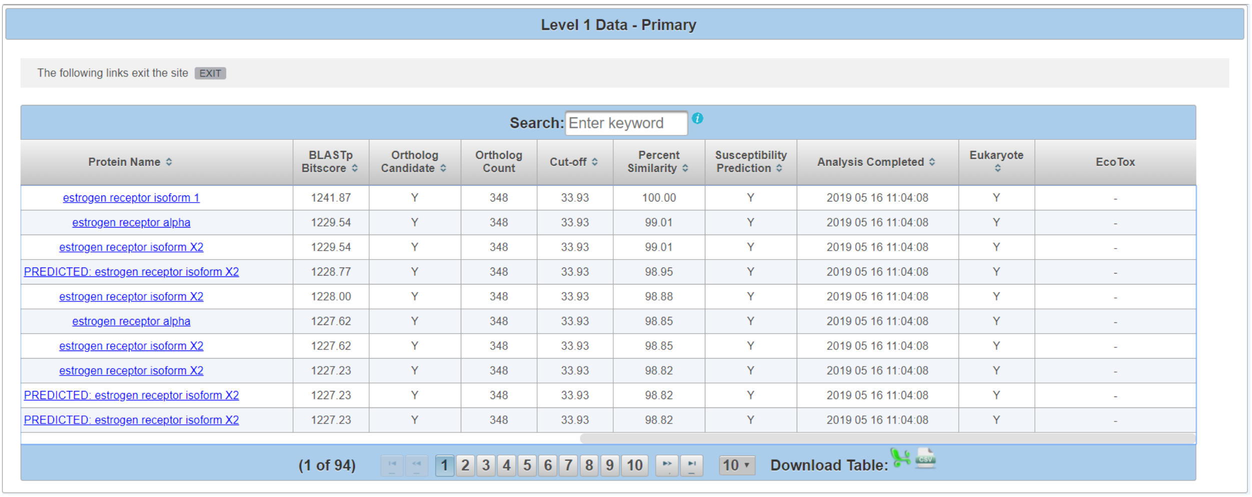
The following links exit the site EXIT

Search: Enter keyword

☐	Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name	Protein
☐	4	NP_000116.2	1265506	9606	Mammalia	Mammalia	Homo sapiens	Human	estrogen rece
☐	4	ABY64717.1	2023	9593	Mammalia	Mammalia	Gorilla gorilla	Western gorilla	estrogen re
☐	4	XP_003311596.1	178219	9598	Mammalia	Mammalia	Pan troglodytes	Chimpanzee	estrogen recep
☐	4	XP_018884801.1	47068	9595	Mammalia	Mammalia	Gorilla gorilla gorilla	Western lowland gorilla	PREDICTED: estroge
☐	4	XP_003811544.1	51891	9597	Mammalia	Mammalia	Pan paniscus	Pygmy chimpanzee	estrogen recep
☐	4	ABY64718.1	1718	9600	Mammalia	Mammalia	Pongo pygmaeus	Bornean orangutan	estrogen re
☐	4	XP_002817538.1	145798	9601	Mammalia	Mammalia	Pongo abelii	Sumatran orangutan	estrogen recep
☐	4	XP_011751932.1	69122	9545	Mammalia	Mammalia	Macaca nemestrina	Pig-tailed macaque	estrogen recep
☐	4	XP_014992596.1	88400	9544	Mammalia	Mammalia	Macaca mulatta	Rhesus monkey	PREDICTED: estroge
☐	4	XP_011922091.1	66748	9531	Mammalia	Mammalia	Cercocebus atys	Sooty mangabey	PREDICTED: estroge

(1 of 94) 1 2 3 4 5 6 7 8 9 10 | 10 | Download Table:

Columns in right side of table:



Level 1 Data - Primary

The following links exit the site EXIT

Search: Enter keyword

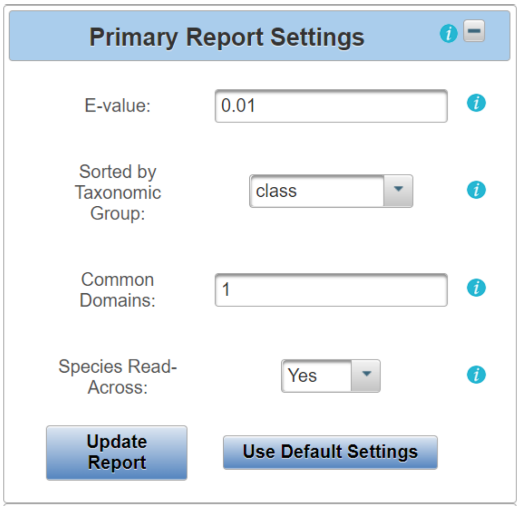
Protein Name	BLASTp Bitscore	Ortholog Candidate	Ortholog Count	Cut-off	Percent Similarity	Susceptibility Prediction	Analysis Completed	Eukaryote	EcoTox
estrogen receptor isoform 1	1241.87	Y	348	33.93	100.00	Y	2019 05 16 11:04:08	Y	-
estrogen receptor alpha	1229.54	Y	348	33.93	99.01	Y	2019 05 16 11:04:08	Y	-
estrogen receptor isoform X2	1229.54	Y	348	33.93	99.01	Y	2019 05 16 11:04:08	Y	-
PREDICTED: estrogen receptor isoform X2	1228.77	Y	348	33.93	98.95	Y	2019 05 16 11:04:08	Y	-
estrogen receptor isoform X2	1228.00	Y	348	33.93	98.88	Y	2019 05 16 11:04:08	Y	-
estrogen receptor alpha	1227.62	Y	348	33.93	98.85	Y	2019 05 16 11:04:08	Y	-
estrogen receptor isoform X2	1227.62	Y	348	33.93	98.85	Y	2019 05 16 11:04:08	Y	-
estrogen receptor isoform X2	1227.23	Y	348	33.93	98.82	Y	2019 05 16 11:04:08	Y	-
PREDICTED: estrogen receptor isoform X2	1227.23	Y	348	33.93	98.82	Y	2019 05 16 11:04:08	Y	-
PREDICTED: estrogen receptor isoform X2	1227.23	Y	348	33.93	98.82	Y	2019 05 16 11:04:08	Y	-

(1 of 94) 1 2 3 4 5 6 7 8 9 10 | 10 | Download Table:

Level 1: Primary Report Settings

Default settings

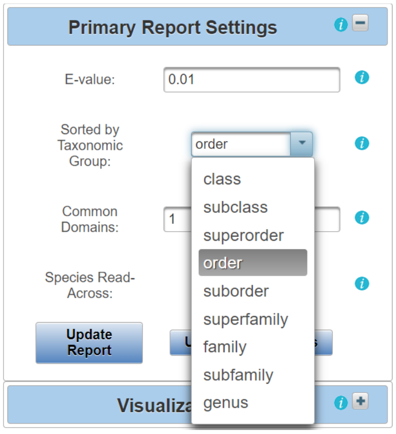
The “Primary Report Settings” drop down allows the user to view default settings on the table below and manipulate certain settings. “Primary Report Settings” are only available on the “Primary Report” display, not the “Full Report.” The default settings show data for hits whose E-value are ≤ 0.01 and have been identified to have ≥ 1 domain in common with the query sequence. The default setting for the “Sorted by Taxonomic Group” is “class,” therefore the “Filtered Taxonomic Group” column in the table is set to identify and report the taxonomic lineage of “class” from the NCBI Taxonomy Database. However, if class is not identified in the NCBI Taxonomic Hierarchy associated with the hit accession, then the algorithm will report the next available Taxonomic Group moving from class to subclass, to superorder, to order, to suborder, to superfamily, to family, to subfamily, to genus. Finally, the susceptibility predictions are set by using species read-across. (Please view **Documentation** Section of the User Guide for details on Read-Across settings). Briefly, Species Read-across is used to set the susceptibility prediction, where all ortholog candidates are Susceptible = Y; all species listed above the susceptibility cut-off are Susceptible = Y; all species below the cut-off from the same taxonomic group of one or more species above the cut-off are Susceptible = Y; and those below the cut-off that are not ortholog candidates and do not belong to a taxonomic group above the cut-off are Susceptible = N.



E-value: 0.01
Sorted by Taxonomic Group: class
Common Domains: 1
Species Read-Across: Yes

Changing Default Settings

The “E-value” and “Common Domains” settings can be manipulated by the user by entering the desired E-value or number of Common Domains in the respective text boxes and clicking “Update Report.” The table and data visualization will automatically be updated after a few seconds. The user may choose to change the level of the taxonomic hierarchy that is used for the susceptibility prediction. From the Sorted by Taxonomic Group” dropdown the user may choose to display a different taxonomic group in the “Filtered Taxonomic Group” column of the data table.



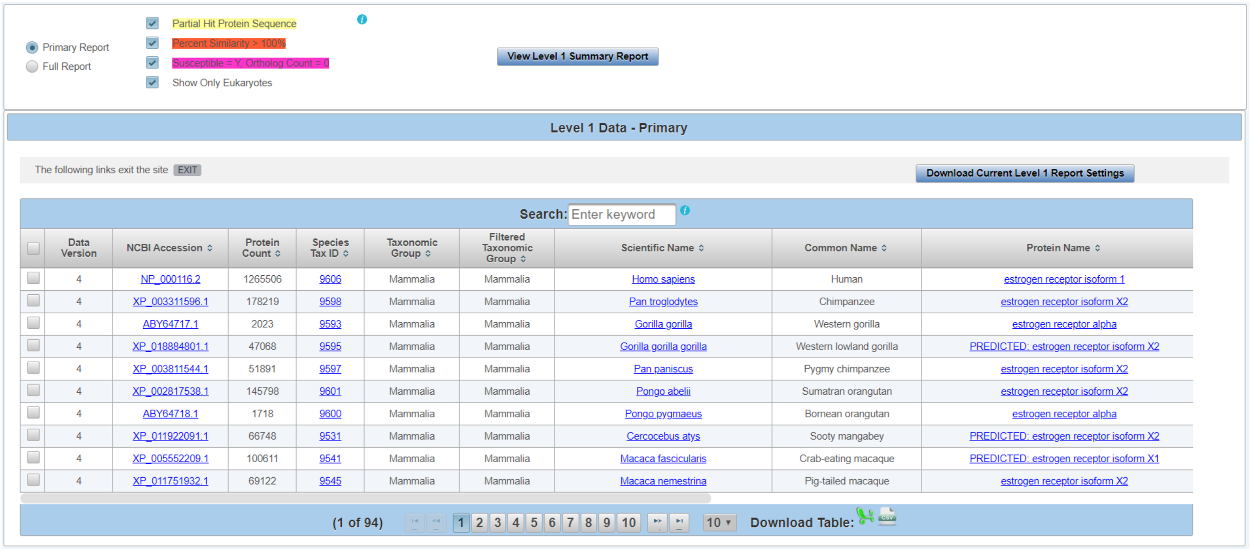
Primary Report Settings

E-value: 0.01

Sorted by Taxonomic Group: order

Common Domains: 1

If the user chooses "order" for example, the "Filtered Taxonomic Group" column in the data table will report the taxonomic lineage of "order" from the NCBI Taxonomy Database and all species read-across for the susceptibility prediction will be based on order instead of class. The data visualization will also update. As described previously, if order is not identified in the NCBI Taxonomic Hierarchy associated with the hit accession, then the algorithm will report the next available Taxonomic Group moving from suborder, to superfamily, to family, to subfamily, to genus. Upon selecting the Taxonomic Group from the dropdown and clicking "Update Report," the Level 1 Data for the Primary report will update to the selected taxonomic level.



● Primary Report
○ Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

Level 1 Data - Primary

The following links exit the site EXIT

Search: Enter keyword

☐	Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name	Protein Name
☐	4	NP_000116.2	1265506	9606	Mammalia	Mammalia	Homo sapiens	Human	estrogen receptor isoform 1
☐	4	XP_003311596.1	178219	9598	Mammalia	Mammalia	Pan troglodytes	Chimpanzee	estrogen receptor isoform X2
☐	4	ABY64717.1	2023	9593	Mammalia	Mammalia	Gorilla gorilla	Western gorilla	estrogen receptor alpha
☐	4	XP_018884801.1	47068	9595	Mammalia	Mammalia	Gorilla gorilla gorilla	Western lowland gorilla	PREDICTED: estrogen receptor isoform X2
☐	4	XP_003811544.1	51891	9597	Mammalia	Mammalia	Pan paniscus	Pygmy chimpanzee	estrogen receptor isoform X2
☐	4	XP_002817538.1	145798	9601	Mammalia	Mammalia	Pongo abelii	Sumatran orangutan	estrogen receptor isoform X2
☐	4	ABY64718.1	1718	9600	Mammalia	Mammalia	Pongo pygmaeus	Bornean orangutan	estrogen receptor alpha
☐	4	XP_011922091.1	66748	9531	Mammalia	Mammalia	Cercocebus atys	Sooty mangabey	PREDICTED: estrogen receptor isoform X2
☐	4	XP_005552209.1	100611	9541	Mammalia	Mammalia	Macaca fascicularis	Crab-eating macaque	PREDICTED: estrogen receptor isoform X1
☐	4	XP_011751932.1	69122	9545	Mammalia	Mammalia	Macaca nemestrina	Pig-tailed macaque	estrogen receptor isoform X2

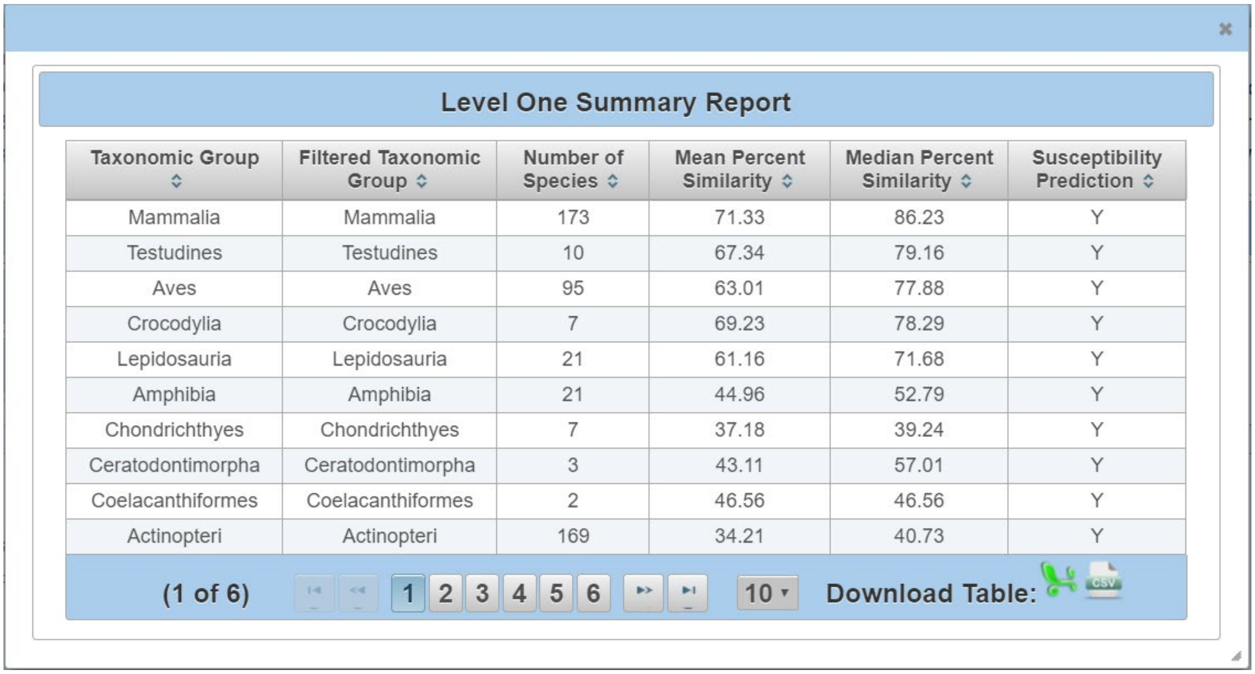
(1 of 94)

Level One Summary Report

The user can view a summary of the data for each taxonomic group by clicking on the “View Level 1 Summary Report” button. The data includes, number of species, mean percent similarity, median percent similarity and susceptibility prediction. This data can also be downloaded.

● Primary Report
○ Full Report

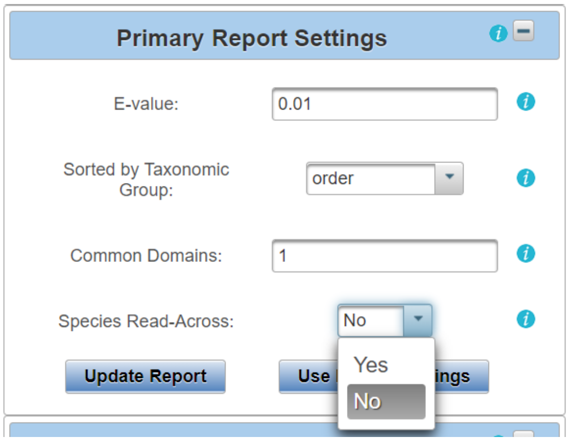
☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes



Level One Summary Report

Taxonomic Group	Filtered Taxonomic Group	Number of Species	Mean Percent Similarity	Median Percent Similarity	Susceptibility Prediction
Mammalia	Mammalia	173	71.33	86.23	Y
Testudines	Testudines	10	67.34	79.16	Y
Aves	Aves	95	63.01	77.88	Y
Crocodylia	Crocodylia	7	69.23	78.29	Y
Lepidosauria	Lepidosauria	21	61.16	71.68	Y
Amphibia	Amphibia	21	44.96	52.79	Y
Chondrichthyes	Chondrichthyes	7	37.18	39.24	Y
Ceratodontimorpha	Ceratodontimorpha	3	43.11	57.01	Y
Coelacanthiformes	Coelacanthiformes	2	46.56	46.56	Y
Actinopteri	Actinopteri	169	34.21	40.73	Y

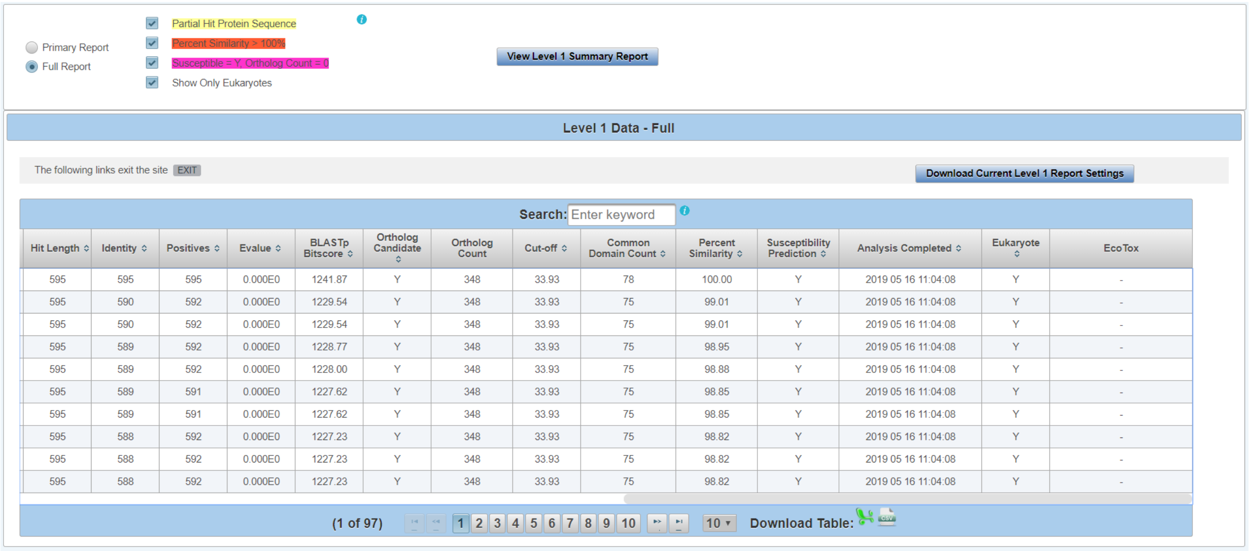
The user may also choose to turn species read-across off, by using the “Species Read-Across” drop-down and selecting “No” and clicking “Update Report.” When “No” is selected, the susceptibility predictions will only be “Y” in the table below if Percent Similarity is above the Cut-off or if the hit is identified as an Ortholog Candidate, yes or “Y.” Any hit below the cut-off will yield a susceptibility prediction of no or “N.”



Primary Report Settings

E-value: 0.01
Sorted by Taxonomic Group: order
Common Domains: 1
Species Read-Across: No

The user can select the “Full Report” on the “Level 1” page, which includes the same information as the “Primary Report” and additional information pertaining to the alignment of the protein sequence using BLASTp. Additional information includes the number of amino acid residues in the sequence (Hit Length), the number of exact matching amino acids between the hit and query sequence (Identity), the number of exact and similar matches in amino acids between the hit and the query sequence (Positives), the expect value (E-value) describing the number of different alignments expected to occur in the database search by chance, and the conserved domain count. The conserved domain count identifies all domains associated with the query protein in the NCBI conserved domains database (Specific hits, Non-specific hits, Superfamilies, and Multi-domains; See NCBI conserved domains database for details). SeqAPASS algorithms record the query sequence coverage of each curated domain and compares that coverage to that of the hit sequence. If the hit sequence covers the curated domain greater than or equal to the query sequence, then the domain is considered a common domain between the hit and query. The number of common domains comparing each hit sequence to the query sequence are summed and reported. This column displays “0” when the hit protein and query protein do not have any common domains. (See **Search, View, and Download Data Tables** section of user guide for more information). The user can also download the currently applied report settings by selecting the “Download Current Level 1 Report Settings.” This csv allows the user to track which settings were used or changed by the user when downloading a data table.



○ Primary Report
● Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

Level 1 Data - Full

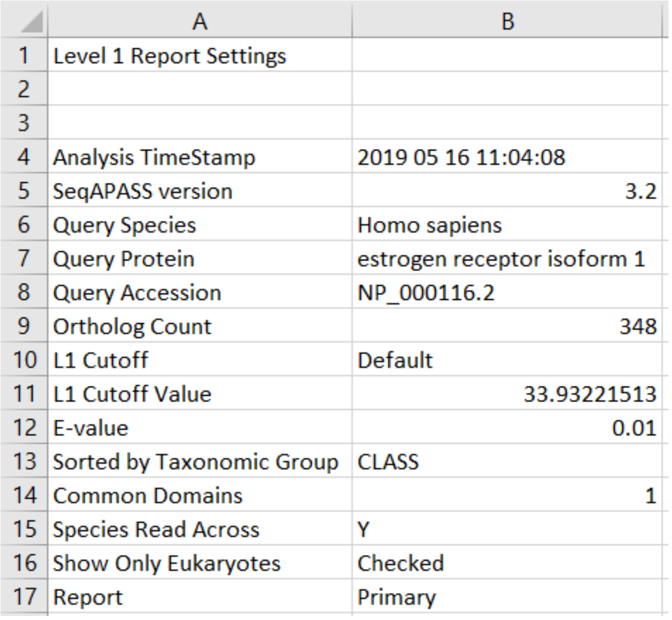
The following links exit the site EXIT

Search: Enter keyword

Hit Length	Identity	Positives	Evalue	BLASTp Bitscore	Ortholog Candidate	Ortholog Count	Cut-off	Common Domain Count	Percent Similarity	Susceptibility Prediction	Analysis Completed	Eukaryote	EcoTox
595	595	595	0.000E0	1241.87	Y	348	33.93	78	100.00	Y	2019 05 16 11:04:08	Y	-
595	590	592	0.000E0	1229.54	Y	348	33.93	75	99.01	Y	2019 05 16 11:04:08	Y	-
595	590	592	0.000E0	1229.54	Y	348	33.93	75	99.01	Y	2019 05 16 11:04:08	Y	-
595	589	592	0.000E0	1228.77	Y	348	33.93	75	98.95	Y	2019 05 16 11:04:08	Y	-
595	589	592	0.000E0	1228.00	Y	348	33.93	75	98.88	Y	2019 05 16 11:04:08	Y	-
595	589	591	0.000E0	1227.62	Y	348	33.93	75	98.85	Y	2019 05 16 11:04:08	Y	-
595	589	591	0.000E0	1227.62	Y	348	33.93	75	98.85	Y	2019 05 16 11:04:08	Y	-
595	588	592	0.000E0	1227.23	Y	348	33.93	75	98.82	Y	2019 05 16 11:04:08	Y	-
595	588	592	0.000E0	1227.23	Y	348	33.93	75	98.82	Y	2019 05 16 11:04:08	Y	-
595	588	592	0.000E0	1227.23	Y	348	33.93	75	98.82	Y	2019 05 16 11:04:08	Y	-

(1 of 97)

Note: SeqAPASS v2.0 and newer parse the BLASTp query and hit accessions to identify all the species/accessions from identical proteins. Therefore, if a hit sequence represents multiple species, all species with the identical sequence will be found in the data tables for Level 1. To determine which sequence/species was identified from BLASTp as a hit and which sequence/species was parsed from the identical sequence, view the “Full Report” for Level, column “Identical Protein,” where “N” is indicative of the original hit sequence and “Y” is the parsed sequence.

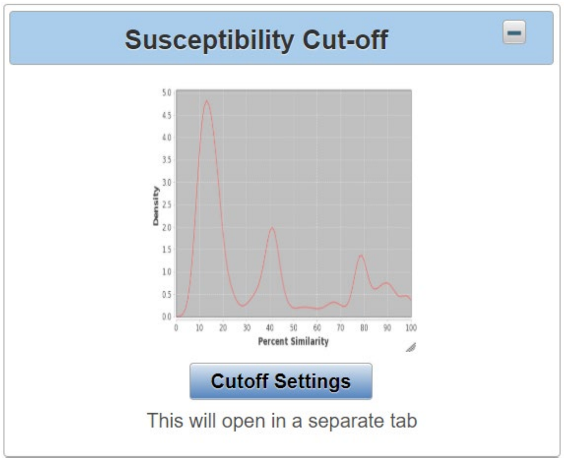


	A	B
1	Level 1 Report Settings	
2		
3		
4	Analysis TimeStamp	2019 05 16 11:04:08
5	SeqAPASS version	3.2
6	Query Species	Homo sapiens
7	Query Protein	estrogen receptor isoform 1
8	Query Accession	NP_000116.2
9	Ortholog Count	348
10	L1 Cutoff	Default
11	L1 Cutoff Value	33.93221513
12	E-value	0.01
13	Sorted by Taxonomic Group	CLASS
14	Common Domains	1
15	Species Read Across	Y
16	Show Only Eukaryotes	Checked
17	Report	Primary

When downloading the current level 1 report settings, the following information will be present in the csv. If the user decides to change the default settings, the csv can be utilized for quick information if the SeqAPASS page is no longer accessible.

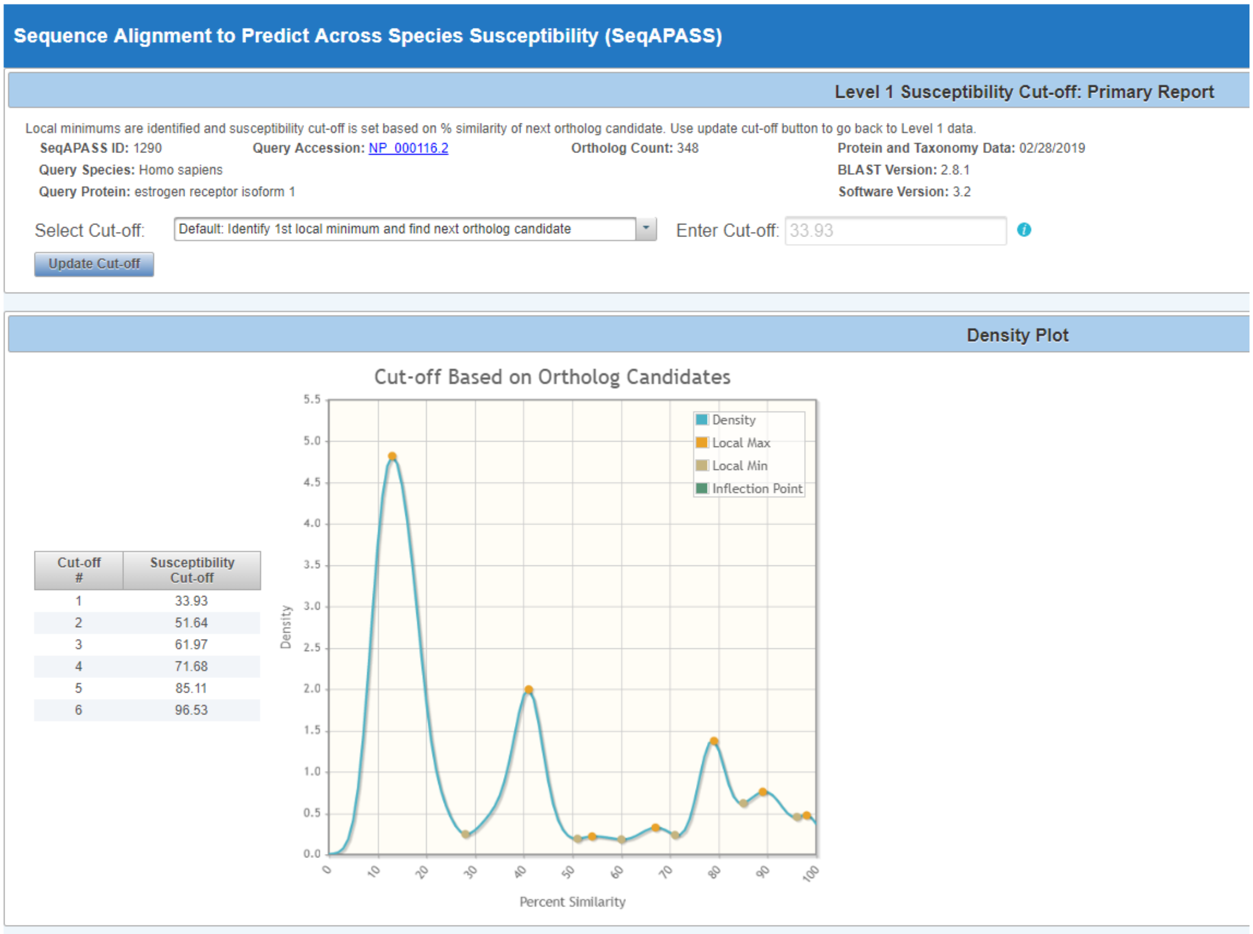
Susceptibility Cutoff Box for Level 1

The susceptibility prediction is determined by identifying ortholog candidates, sequences above a defined susceptibility cutoff, or by identifying those species below the susceptibility cut-off from an organism class above the susceptibility cutoff. The default susceptibility cut-off is set by plotting the distribution of percent similarities calculated for each hit protein. From this plot, the critical points are identified, and the local minimums and maximums reported. Using the ortholog candidate data, a susceptibility cut-off is automatically determined by identifying the first ortholog candidate at an equal or higher percent similarity than the first local minimum. The user can view this graph by clicking the “Cutoff Settings” button in the “Susceptibility Cut-off” box, which will open a new tab in the web browser. The “Select Cut-Off” drop-down can allow the user to select between the default cut-off, the 2nd local minimum or a user defined cut-off. The 2nd susceptibility cut-off is identified in the density plot by finding the 1st ortholog candidate at an equal or higher percent similarity to that of the 2nd local minimum. Upon selecting the User defined cut-off from the dropdown, the user can view and closely examine the density plot and manipulate the cut-off. The Enter Cut-off text box becomes active and the user can enter a number 1-100. To update the cut-off in the Level 1 data report and/or close the cutoff tab and return to the Level 1 page, click “Update Cut-off” button.



This will open in a separate tab

Note: The user should have a justification for changing the susceptibility cut-off, either based on evaluation of Ortholog cutoffs in the data visualization or from empirical evidence.



Local minimums are identified and susceptibility cut-off is set based on % similarity of next ortholog candidate. Use update cut-off button to go back to Level 1 data.

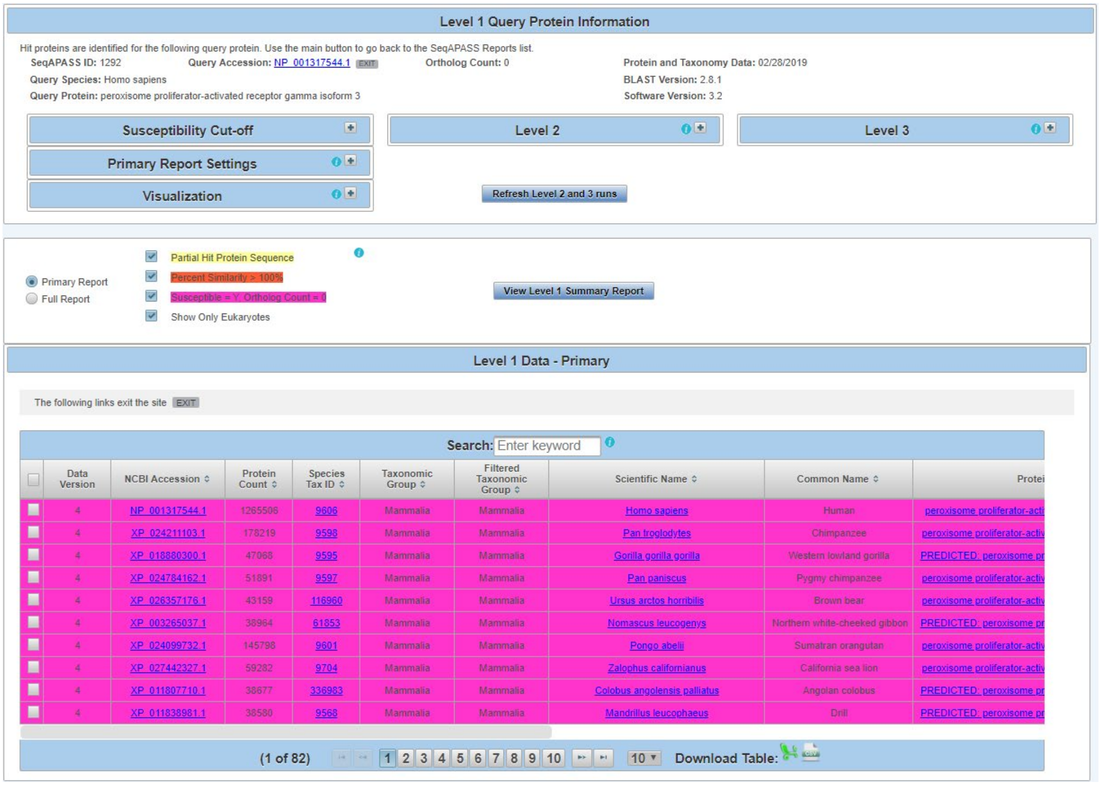
SeqAPASS ID: 1290
Query Accession: NP_000116.2
Ortholog Count: 348
Protein and Taxonomy Data: 02/28/2019
Query Species: Homo sapiens
BLAST Version: 2.8.1
Query Protein: estrogen receptor isoform 1
Software Version: 3.2

Select Cut-off: Default: Identify 1st local minimum and find next ortholog candidate
Enter Cut-off: 33.93

Cut-off #	Susceptibility Cut-off
1	33.93
2	51.64
3	61.97
4	71.68
5	85.11
6	96.53

All potential susceptibility cut-offs generated by the data distribution and ortholog candidate identification are reported in the table with columns “Cut-off #” and “Susceptibility Cut-off”. The user can use these numbers to define a cut-off if empirical evidence suggests that the “Default” or “2 minimum” are not supported.

No Orthologs Detected



Level 1 Query Protein Information

Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1292
Query Accession: NP_001317544.1 EXIT
Ortholog Count: 0
Query Species: Homo sapiens
Query Protein: peroxisome proliferator-activated receptor gamma isoform 3
Protein and Taxonomy Data: 02/28/2019
BLAST Version: 2.8.1
Software Version: 3.2

● Primary Report
○ Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

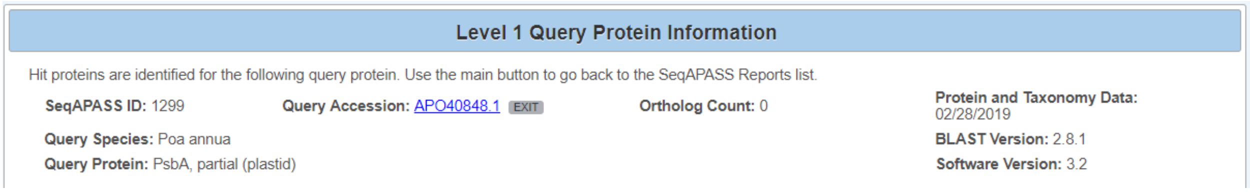
Level 1 Data - Primary

The following links exit the site EXIT

Search: Enter keyword

☐	Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name	Protei
☐	4	NP_001317544.1	1265506	9606	Mammalia	Mammalia	Homo sapiens	Human	peroxisome proliferator-acti
☐	4	XP_024211103.1	178219	9598	Mammalia	Mammalia	Pan troglodytes	Chimpanzee	peroxisome proliferator-activ
☐	4	XP_018880300.1	47068	9595	Mammalia	Mammalia	Gorilla gorilla gorilla	Western lowland gorilla	PREDICTED: peroxisome pr
☐	4	XP_024784162.1	51891	9597	Mammalia	Mammalia	Pan paniscus	Pygmy chimpanzee	peroxisome proliferator-activ
☐	4	XP_026357176.1	43159	116960	Mammalia	Mammalia	Ursus arctos horribilis	Brown bear	peroxisome proliferator-activ
☐	4	XP_003265037.1	38964	61853	Mammalia	Mammalia	Nomascus leucogenys	Northern white-cheeked gibbon	PREDICTED: peroxisome pr
☐	4	XP_024099732.1	145798	9601	Mammalia	Mammalia	Pongo abelii	Sumatran orangutan	peroxisome proliferator-activ
☐	4	XP_027442327.1	59282	9704	Mammalia	Mammalia	Zalophus californianus	California sea lion	peroxisome proliferator-activ
☐	4	XP_011807710.1	38677	336983	Mammalia	Mammalia	Colobus angolensis palliatus	Angolan colobus	PREDICTED: peroxisome pr
☐	4	XP_011838981.1	38580	9568	Mammalia	Mammalia	Mandrillus leucophaeus	Drill	PREDICTED: peroxisome pr

If no orthologs are detected from reciprocal best hit blast analysis, the "Ortholog Count" will be "0" at the top of the "Level 1 Query Protein Information" page. The cutoff will be set by the local minimums only, therefore the susceptibility prediction will NOT take into account ortholog candidates. ***It is recommended that the user checks the full report for Ortholog candidates or identifies a different query sequence for the susceptibility predictions***. Here, the susceptibility predictions will be highlighted in dark pink in the Level 1 data table to indicate that 0 orthologs were detected and the susceptibility cutoff was determined from plotting the distribution of percent similarities and identifying the local minimums.



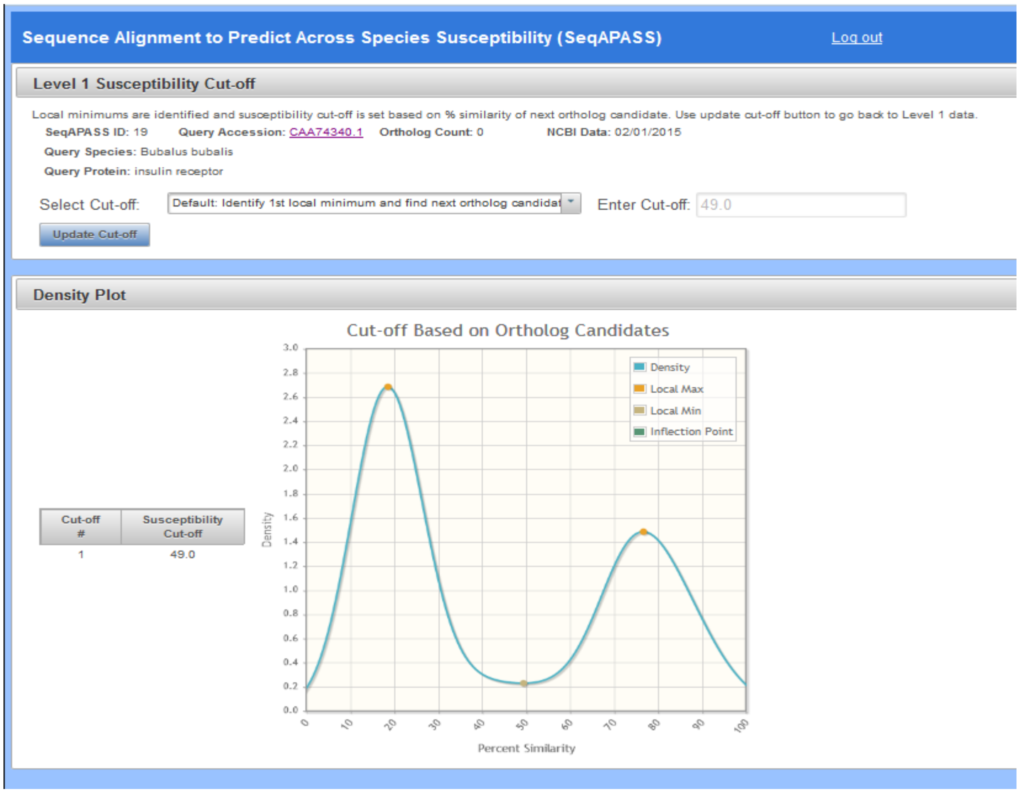
Level 1 Query Protein Information

Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1299
Query Accession: APO40848.1 EXIT
Ortholog Count: 0
Query Species: Poa annua
Query Protein: PsbA, partial (plastid)
Protein and Taxonomy Data: 02/28/2019
BLAST Version: 2.8.1
Software Version: 3.2

<u>Note:</u> De-select the "Show Only Eukaryotes" checkbox to see if prokaryotes were identified as orthologs.

By clicking on the “Cutoff Settings” button when no orthologs are detected, the “Cut-off #” and “Susceptibility Cut-off” columns will report only the local minimum values.



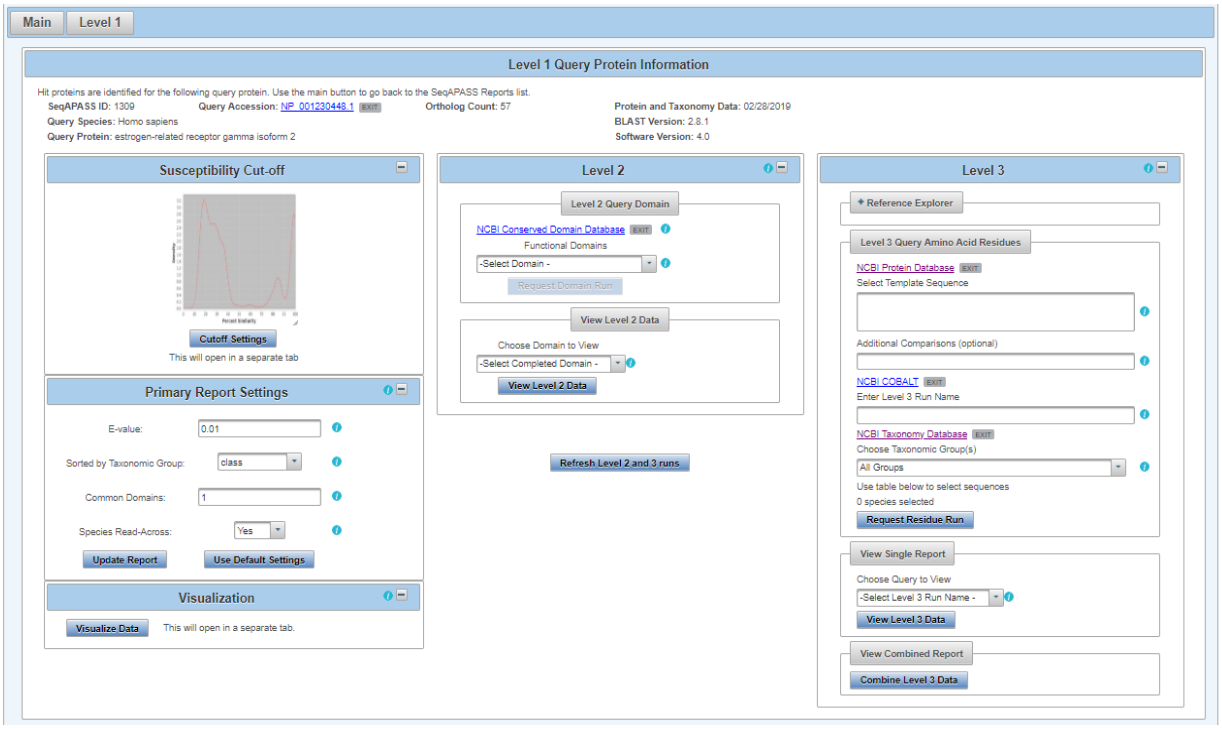
Local minimums are identified and susceptibility cut-off is set based on % similarity of next ortholog candidate. Use update cut-off button to go back to Level 1 data.

SeqAPASS ID: 19 **Query Accession:** CAA74340.1 **Ortholog Count:** 0 **NCBI Data:** 02/01/2015
Query Species: Bubalus bubalis
Query Protein: insulin receptor

Select Cut-off: Default: Identify 1st local minimum and find next ortholog candidat
Enter Cut-off: 49.0

Cut-off #	Susceptibility Cut-off
1	49.0

From the “Level 1” page the user can return to the list of completed SeqAPASS runs by clicking the “Main” button on the upper left-hand side of the “Level 1 Query Protein Information” page.



Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1309 **Query Accession:** NP_001230448.1 **Ortholog Count:** 57 **Protein and Taxonomy Data:** 02/28/2019
Query Species: Homo sapiens **BLAST Version:** 2.8.1
Query Protein: estrogen-related receptor gamma isoform 2 **Software Version:** 4.0

This will open in a separate tab

E-value: 0.01
Sorted by Taxonomic Group: class
Common Domains: 1
Species Read-Across: Yes

This will open in a separate tab.

Functional Domains
-Select Domain -

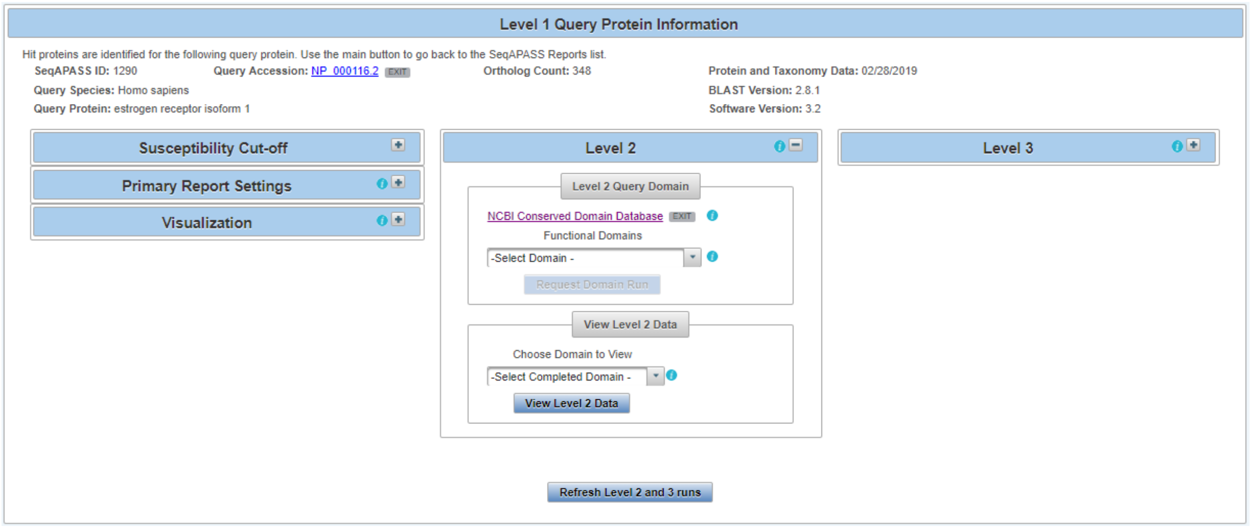
Choose Domain to View
-Select Completed Domain -

Select Template Sequence
Additional Comparisons (optional)
Enter Level 3 Run Name
Choose Taxonomic Group(s)
All Groups
Use table below to select sequences
0 species selected

Choose Query to View
-Select Level 3 Run Name -

Level 2: Functional Domain(s) Alignment

In the “View SeqAPASS Reports” tab, on the “Level 1 Query Protein Information” page, there is a “Level 2” box for comparing hit domains to the query domain. In the “Level 2” dropdown box, there is a link out to the “NCBI Conserved Domain Database” for the query protein of interest. Below this link the user will find a drop-down containing functional domains associated with the query sequence for comparison across species.

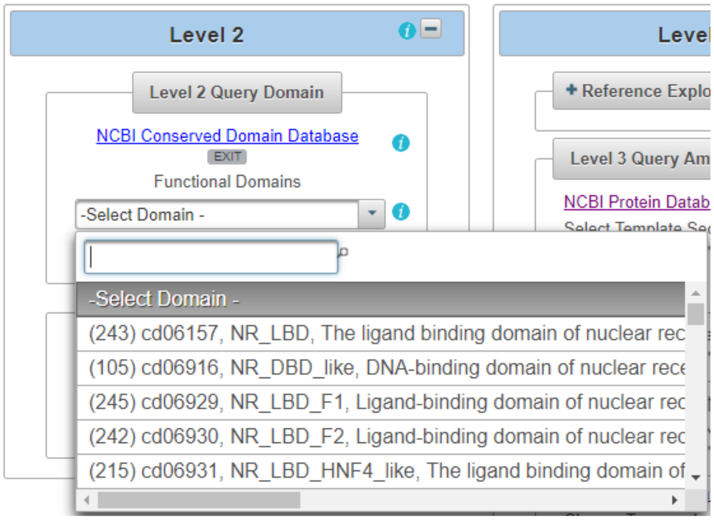


Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1290
Query Accession: NP_000116.2
Ortholog Count: 348
Protein and Taxonomy Data: 02/28/2019
Query Species: Homo sapiens
BLAST Version: 2.8.1
Query Protein: estrogen receptor isoform 1
Software Version: 3.2

In the drop-down box (below the words “Functional Domains”) the user will find all domains associated with the query protein listed in the NCBI Conserved Domains Database. To compare a domain from the query protein to domains of the hit proteins, the user will use the drop-down to highlight a domain and click the “Request Domain Run” button.

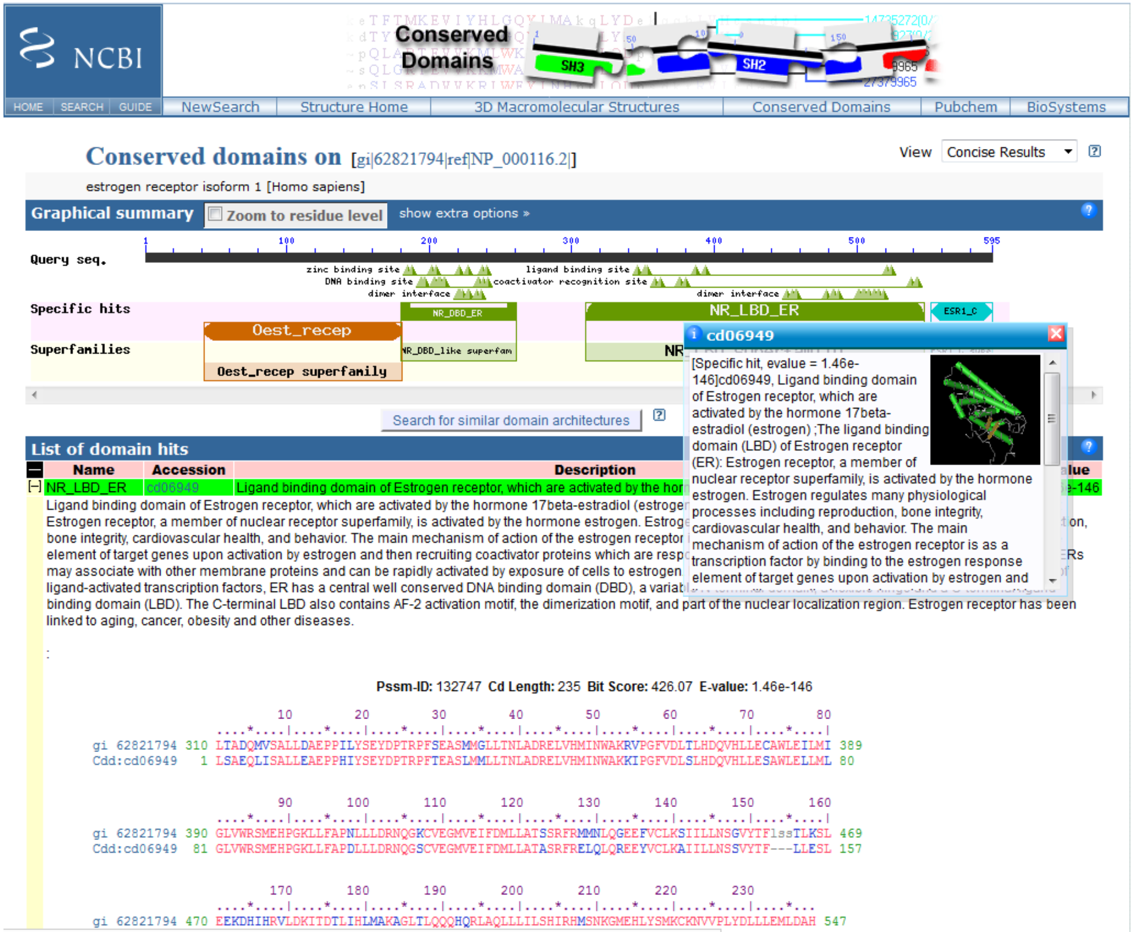
Note: Domains in the drop-down are listed with the first amino acid residue position that aligns with the NCBI curated domain in parenthesis, followed by the NCBI domain Accession, domain name, and description.



-Select Domain -
(243) cd06157, NR_LBD, The ligand binding domain of nuclear rec
(105) cd06916, NR_DBD_like, DNA-binding domain of nuclear rece
(245) cd06929, NR_LBD_F1, Ligand-binding domain of nuclear rec
(242) cd06930, NR_LBD_F2, Ligand-binding domain of nuclear rec
(215) cd06931, NR_LBD_HNF4_like, The ligand binding domain of

Note: The user can also use the text box on the top of the drop-down to search the “Functional Domain” list in the drop-down.

It is recommended that the user click on the "NCBI Conserved Domains Database" http://www.ncbi.nlm.nih.gov/cdd/ link to identify which domains are "Specific hits" in the NCBI Conserved Domains Database. On the NCBI page, the user can scroll over the graphical representation of the domains associated with the query sequence to highlight and identify the Accession associated with domain "Specific hits." The example below shows the user hovering over the NR_LBD_ER domain with the computer mouse.



Conserved domains on [gi|62821794|ref|NP_000116.2|]

estrogen receptor isoform 1 [Homo sapiens]

Ligand binding domain of Estrogen receptor, which are activated by the hor

[Specific hit, evalue = 1.46e-146]cd06949, Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estrogen) ;The ligand binding domain (LBD) of Estrogen receptor (ER): Estrogen receptor, a member of nuclear receptor superfamily, is activated by the hormone estrogen. Estrogen regulates many physiological processes including reproduction, bone integrity, cardiovascular health, and behavior. The main mechanism of action of the estrogen receptor is as a transcription factor by binding to the estrogen response element of target genes upon activation by estrogen and

Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estroge
Estrogen receptor, a member of nuclear receptor superfamily, is activated by the hormone estrogen. Estrog on,
bone integrity, cardiovascular health, and behavior. The main mechanism of action of the estrogen receptor
element of target genes upon activation by estrogen and then recruiting coactivator proteins which are resp ERs
may associate with other membrane proteins and can be rapidly activated by exposure of cells to estrogen
ligand-activated transcription factors, ER has a central well conserved DNA binding domain (DBD), a variab
binding domain (LBD). The C-terminal LBD also contains AF-2 activation motif, the dimerization motif, and part of the nuclear localization region. Estrogen receptor has been linked to aging, cancer, obesity and other diseases.

:

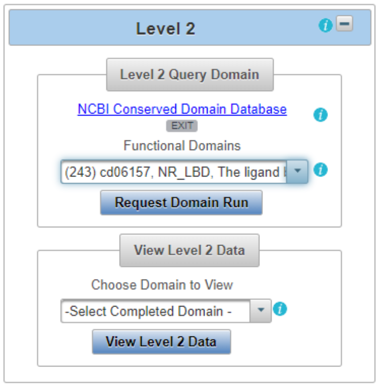
Pssm-ID: 132747 **Cd Length:** 235 **Bit Score:** 426.07 **E-value:** 1.46e-146

```
                     10        20        30        40        50        60        70        80
              ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 62821794 310 LTADQMVSALLDAEPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKRVPGFVDLTLHDQVHLLECAWLEILMI 389
Cdd:cd06949   1 LSAEQLISALLEAEPPHIYSEYDPTRPFTEASLMMLLTNLADRELVHMINWAKKIPGFVDLSLHDQVHLLESAWLELLML 80

                     90       100       110       120       130       140       150       160
              ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 62821794 390 GLVWRSMEHPGKLLFAPNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLNSGVYTFlssTLKSL 469
Cdd:cd06949  81 GLVWRSMEHPGKLLFAPDLLLDRNQGSCVEGMVEIFDMLLATASRFRELQLQREEYVCLKAIILLNSSVYTF---LLESL 157

                    170       180       190       200       210       220       230
              ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*...
gi 62821794 470 EEKDHIHRVLDKITDTLIHLMAKAGLTLQQQHQRLAQLLLILSHIRHMSNKGMEHLYSMKCKNVVPLYDLLLEMLDAH 547
```

After identifying the domain(s) of interest and the corresponding starting residue and domain Accession, the user can return to the SeqAPASS tool, scroll to the domain of interest in the drop-down. If that domain has not been previously run by the user, the "Request Domain Run" button will become active and the user can click it to submit the domain query.



Functional Domains: (243) cd06157, NR_LBD, The ligand l

Choose Domain to View: -Select Completed Domain -

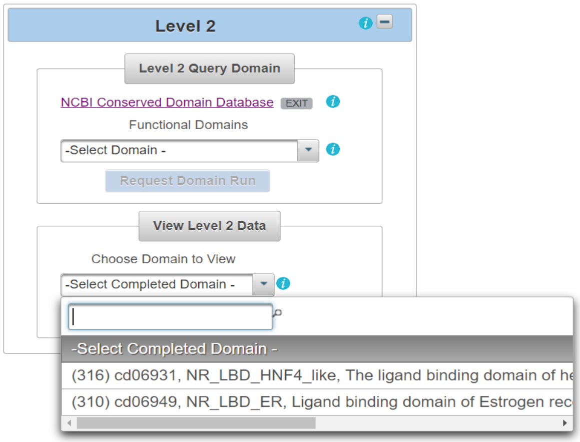
When user clicks the "Request Domain Run" button, the following message will appear if the runs has been submitted successfully.

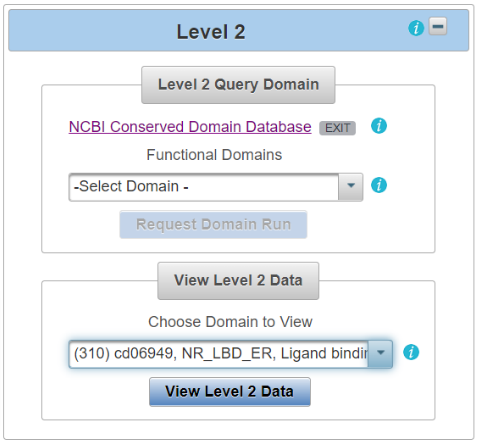
Level 2 Run Requested
Status queued

When sequence comparisons have completed for the selected functional domain, the domain will be present in the drop-down in the View Level 2 Data area. The ***drop-down is not automatically populated*** with the completed domain run. The ***user must click on the "Refresh Level 2 and 3 runs" button to update the page*** for the newly completed domain to present itself in the Choose Domain to View drop-down.

To view a completed Level 2 domain, highlight the domain of interest in the drop-down box and click the "View Level 2 Data" button. This will bring the user to the "Level 2" data page for the selected query protein/domain.

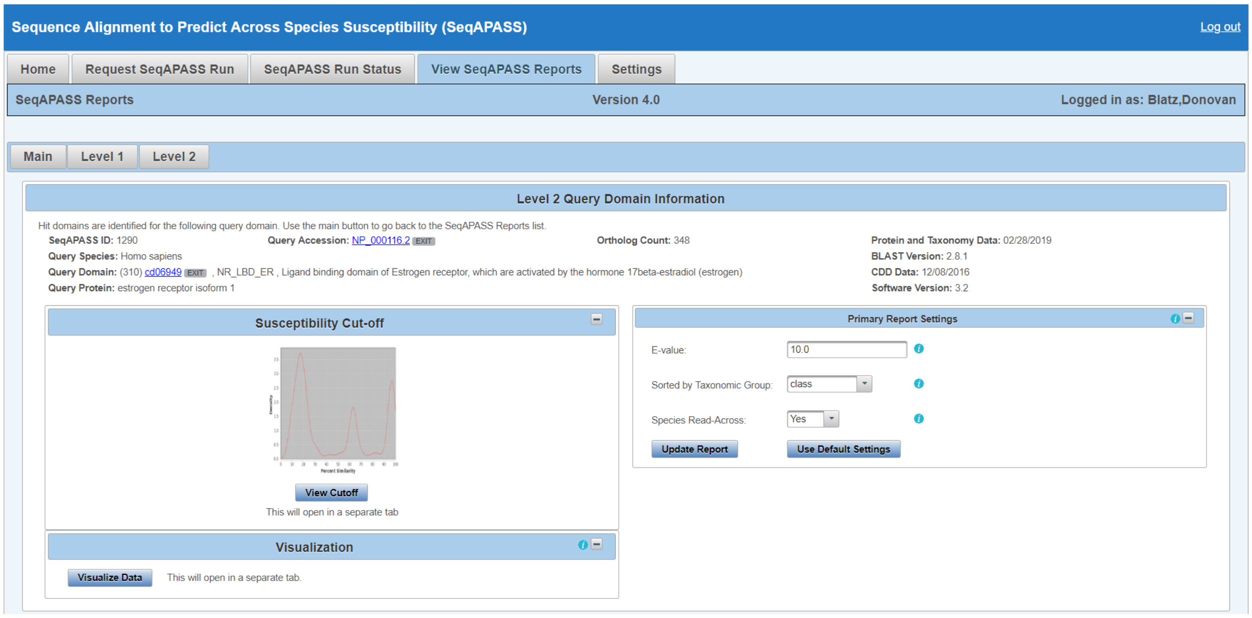
Note: The user can also use the text box on the top of the drop-down to search the "Completed Domain" list.





View Level 2 Data Page

The "Level 2 Query Domain Information" box contains the SeqAPASS Run ID, Query Accession, Ortholog Count (# of hits identified as ortholog candidates to the query species protein sequence), NCBI Data updates ("Protein and Taxonomy Data:" and "CDD Data:" display the dates that NCBI databases were downloaded and incorporated into the SeqAPASS database; "BLAST version:" and "Software Version:" displays the version being used by the SeqAPASS tool for the selected data), Query Species, Query Domain (with link out to NCBI domain page), Query Protein name.



Hit domains are identified for the following query domain. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1290
Query Accession: NP_000116.2 EXIT
Ortholog Count: 348
Protein and Taxonomy Data: 02/28/2019
Query Species: Homo sapiens
BLAST Version: 2.8.1
Query Domain: (310) cd06949 EXIT , NR_LBD_ER , Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estrogen)
CDD Data: 12/08/2016
Query Protein: estrogen receptor isoform 1
Software Version: 3.2

This will open in a separate tab

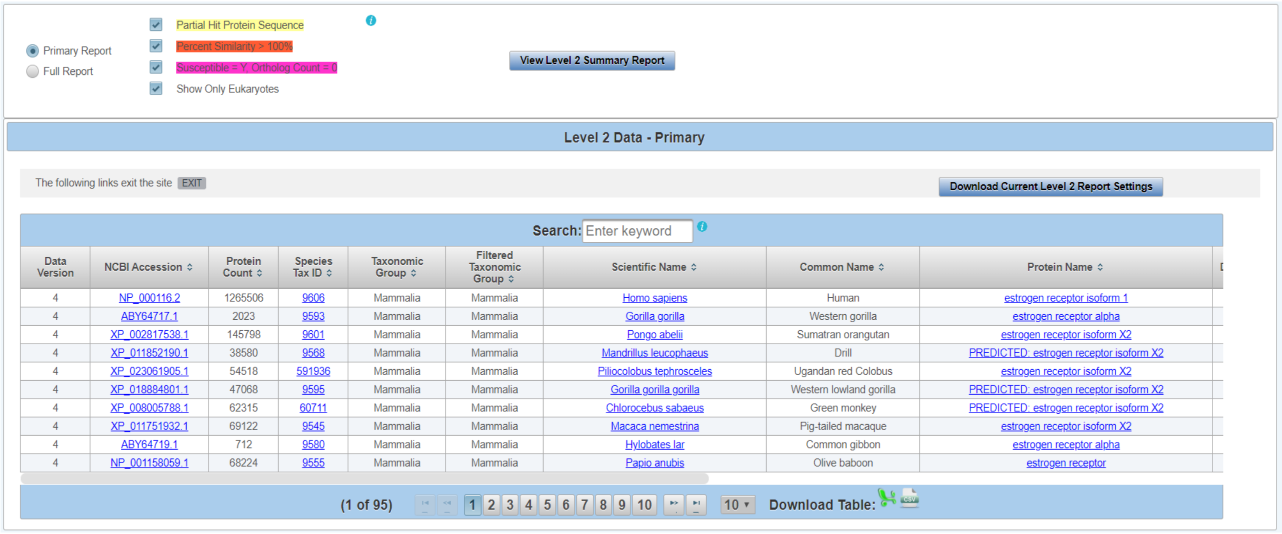
This will open in a separate tab.

E-value: 10.0
Sorted by Taxonomic Group: class
Species Read-Across: Yes

The default "Level 2" table is the "Primary Report", which includes query domain information in the first row below the column titles, followed by hit domains whose sequences aligned with the selected query domain. The hit domains are ordered from the highest to lowest percent similarity (Maximum percent similarity =100%). For each hit domain, Data Version, NCBI Accession and species information is provided, including the "Protein Count" which indicates the number of protein records per species in the NCBI protein database, taxonomic information, and species names. Also included are the NCBI accession for the query protein, query protein name, Domain Type, BLASTP bitscore (describes overall quality of the alignment, See NCBI BLASTp tutorials), and Domain percent similarity ([hit bitscore/query bitscore]*100). If the hit protein has been identified as an ortholog candidate (using reciprocal best hit BLAST method), it will be noted with a "Y" for yes or if not an ortholog candidate, a "N", for no.

A prediction of susceptibility is displayed based on the susceptibility cut-off, identified with a "Y" for yes or an "N" for no. The date/time the analysis was completed is also identified. (See **Search, View, and Download Data Tables** section of user guide for more information). There is a column that identifies if the species is a eukaryote, noted with a "Y" for yes or alternatively a "N" for no if the hit is a prokaryote. Additionally, a column with a link to the U.S. EPA ECOTOX Knowledgebase (https://cfpub.epa.gov/ecotox/help.cfm) is available when there are empirical toxicity data curated for the species identified in the row. This link allows the user to view available single chemical toxicity data from the literature for specific species.

Default highlights identify partial protein sequences, sequences with a bitscore higher than the query domain and therefore percent similarity greater than 100% (commonly synthetic constructs), and when zero ortholog candidates are identified (in this case a user should consider a different query sequence). Additionally, the default setting for the report shows only eukaryote data, excluding prokaryote data from the table with the "Shows Only Eukaryotes" checkbox checked. To view prokaryote data, deselect this checkbox.



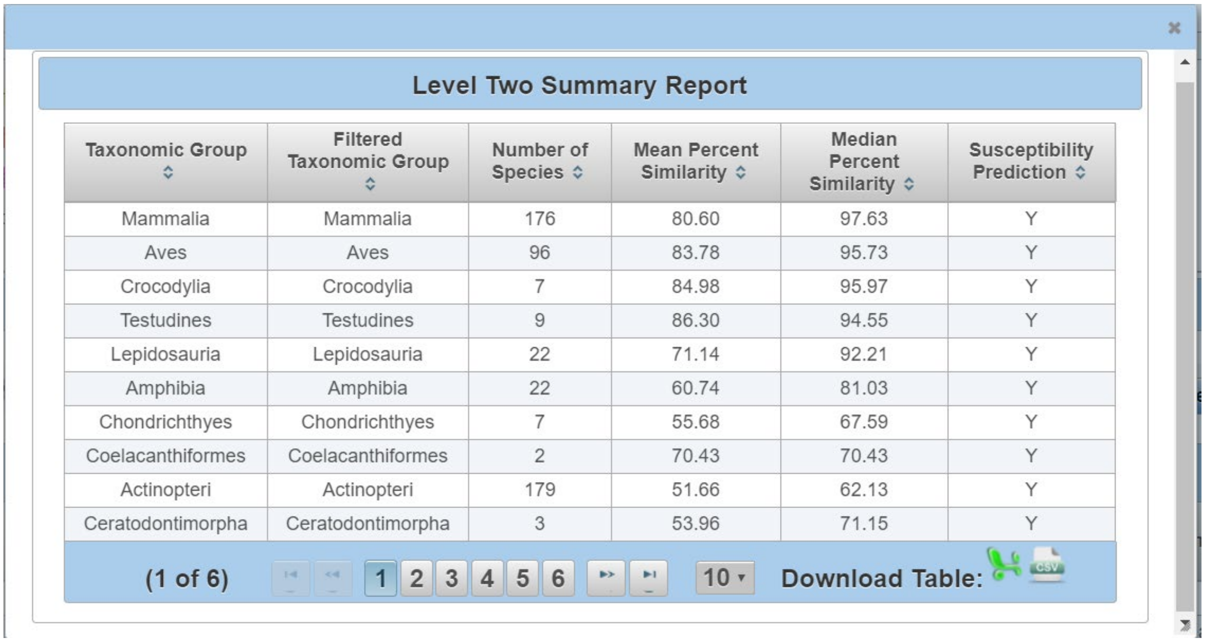
● Primary Report
○ Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name	Protein Name
4	NP_000116.2	1265506	9606	Mammalia	Mammalia	Homo sapiens	Human	estrogen receptor isoform 1
4	ABY64717.1	2023	9593	Mammalia	Mammalia	Gorilla gorilla	Western gorilla	estrogen receptor alpha
4	XP_002817538.1	145798	9601	Mammalia	Mammalia	Pongo abelii	Sumatran orangutan	estrogen receptor isoform X2
4	XP_011852190.1	38580	9568	Mammalia	Mammalia	Mandrillus leucophaeus	Drill	PREDICTED: estrogen receptor isoform X2
4	XP_023061905.1	54518	591936	Mammalia	Mammalia	Piliocolobus tephrosceles	Ugandan red Colobus	estrogen receptor isoform X2
4	XP_018884801.1	47068	9595	Mammalia	Mammalia	Gorilla gorilla gorilla	Western lowland gorilla	PREDICTED: estrogen receptor isoform X2
4	XP_008005788.1	62315	60711	Mammalia	Mammalia	Chlorocebus sabaeus	Green monkey	PREDICTED: estrogen receptor isoform X2
4	XP_011751932.1	69122	9545	Mammalia	Mammalia	Macaca nemestrina	Pig-tailed macaque	estrogen receptor isoform X2
4	ABY64719.1	712	9580	Mammalia	Mammalia	Hylobates lar	Common gibbon	estrogen receptor alpha
4	NP_001158059.1	68224	9555	Mammalia	Mammalia	Papio anubis	Olive baboon	estrogen receptor

Level Two Summary Report

The user can view a summary of the data for each taxonomic group by clicking on the “View Level 2 Summary Report”. The data includes, number of species, mean percent similarity, median percent similarity and susceptibility prediction. This data table can also be downloaded.

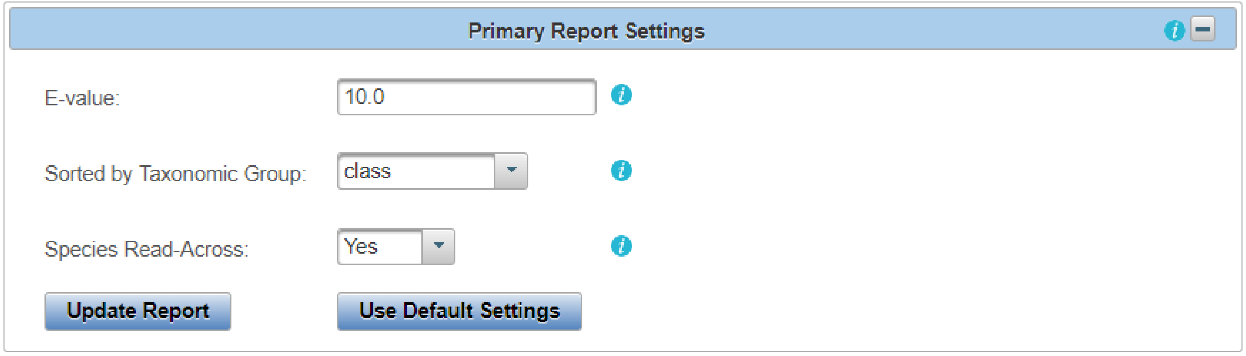


Taxonomic Group	Filtered Taxonomic Group	Number of Species	Mean Percent Similarity	Median Percent Similarity	Susceptibility Prediction
Mammalia	Mammalia	176	80.60	97.63	Y
Aves	Aves	96	83.78	95.73	Y
Crocodylia	Crocodylia	7	84.98	95.97	Y
Testudines	Testudines	9	86.30	94.55	Y
Lepidosauria	Lepidosauria	22	71.14	92.21	Y
Amphibia	Amphibia	22	60.74	81.03	Y
Chondrichthyes	Chondrichthyes	7	55.68	67.59	Y
Coelacanthiformes	Coelacanthiformes	2	70.43	70.43	Y
Actinopteri	Actinopteri	179	51.66	62.13	Y
Ceratodontimorpha	Ceratodontimorpha	3	53.96	71.15	Y

Level 2: Primary Report Settings

Default settings

The “Primary Report Settings” box allows the user to view default settings on the table below and manipulate certain settings. The “Primary Report Settings box is only available on the “Primary Report” display. The default settings show data for hits whose E-value are ≤ 10. The default setting for the “Sorted by Taxonomic Group” is “class,” therefore the “Filtered Taxonomic Group” column in the table is set to identify and report the taxonomic lineage of “class” from the NCBI Taxonomy Database. However, if class is not identified in the NCBI Taxonomic Hierarchy associated with the hit accession, then the algorithm will report the next available Taxonomic Group moving from class to subclass, to superorder, to order, to suborder, to superfamily, to family, to subfamily, to genus. Finally, the susceptibility predictions are set by using Species Read-Across. (Please view **SeqAPASS Documentation** Section of the User Guide for details on Read-Across settings). Briefly, Species Read-Across is used to set the susceptibility prediction, where all ortholog candidates are Susceptible = Y; all species listed above the susceptibility cut-off are Susceptible = Y; all species below the cut-off from the same taxonomic group of one or more species above the cut-off are Susceptible = Y; and those below the cut-off that are not ortholog candidates and do not belong to a taxonomic group above the cut-off are Susceptible = N.


Primary Report Settings

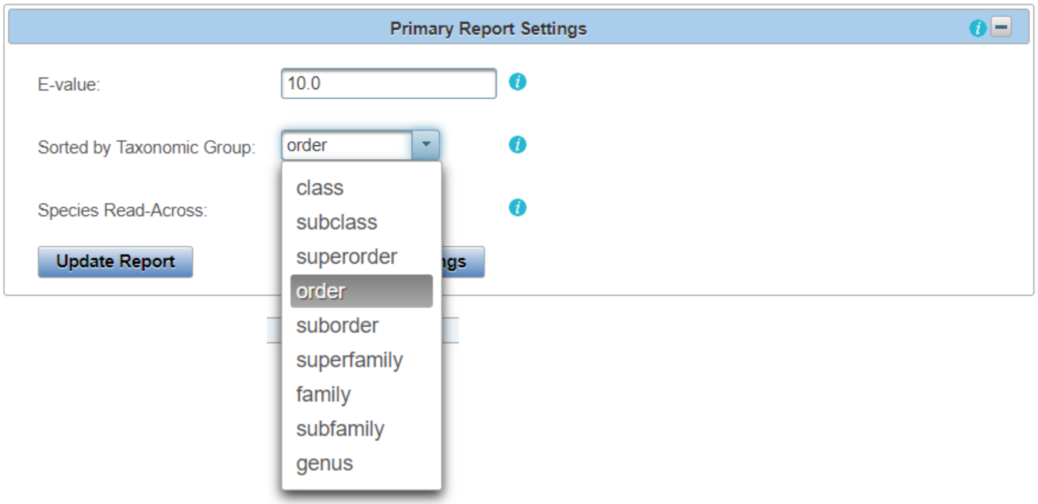
E-value: 10.0

Sorted by Taxonomic Group: class

Species Read-Across: Yes

Changing Default Settings

The user may choose to change the level of the taxonomic hierarchy that is used for the susceptibility prediction. From the Sorted by Taxonomic Group” dropdown the user may choose to display a different taxonomic group in the “Filtered Taxonomic Group” column of the data table.

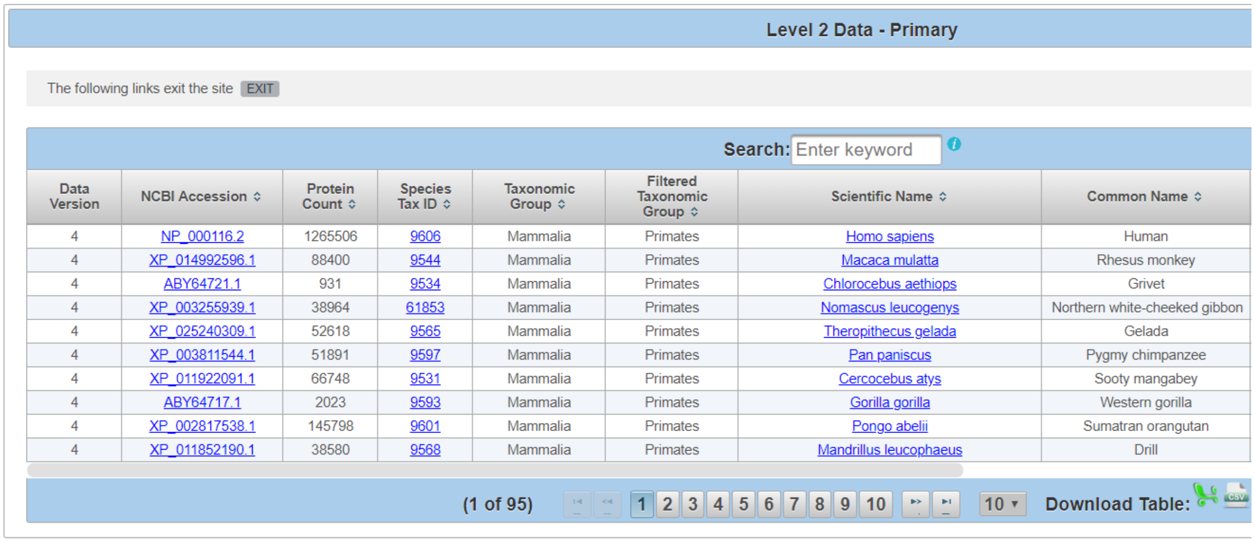

Primary Report Settings

E-value: 10.0

Sorted by Taxonomic Group: order

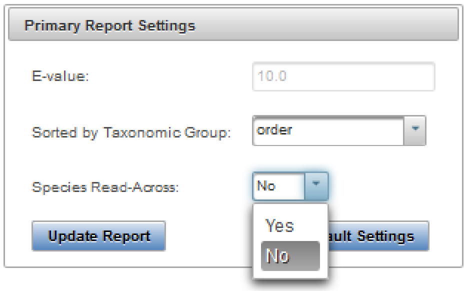
Species Read-Across:

If the user chooses "order" for example, the "Filtered Taxonomic Group" column in the data table will report the taxonomic lineage of "order" from the NCBI Taxonomy Database and all species read-across for the susceptibility prediction will be based on order instead of class. As described previously, if order is not identified in the NCBI Taxonomic Hierarchy associated with the hit accession, then the algorithm will report the next available Taxonomic Group moving from suborder, to superfamily, to family, to subfamily, to genus. Upon selecting the Taxonomic Group from the dropdown and clicking "Update Report," the Level 1 Data for the Primary report will update to the selected taxonomic level. The user can also download the currently applied report settings by selecting the "Download Current Level 2 Report Settings." This csv allows the user to track which settings were used or changed by the user when downloading a data table.



Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name
4	NP_000116.2	1265506	9606	Mammalia	Primates	Homo sapiens	Human
4	XP_014992596.1	88400	9544	Mammalia	Primates	Macaca mulatta	Rhesus monkey
4	ABY64721.1	931	9534	Mammalia	Primates	Chlorocebus aethiops	Grivet
4	XP_003255939.1	38964	61853	Mammalia	Primates	Nomascus leucogenys	Northern white-cheeked gibbon
4	XP_025240309.1	52618	9565	Mammalia	Primates	Theropithecus gelada	Gelada
4	XP_003811544.1	51891	9597	Mammalia	Primates	Pan paniscus	Pygmy chimpanzee
4	XP_011922091.1	66748	9531	Mammalia	Primates	Cercocebus atys	Sooty mangabey
4	ABY64717.1	2023	9593	Mammalia	Primates	Gorilla gorilla	Western gorilla
4	XP_002817538.1	145798	9601	Mammalia	Primates	Pongo abelii	Sumatran orangutan
4	XP_011852190.1	38580	9568	Mammalia	Primates	Mandrillus leucophaeus	Drill

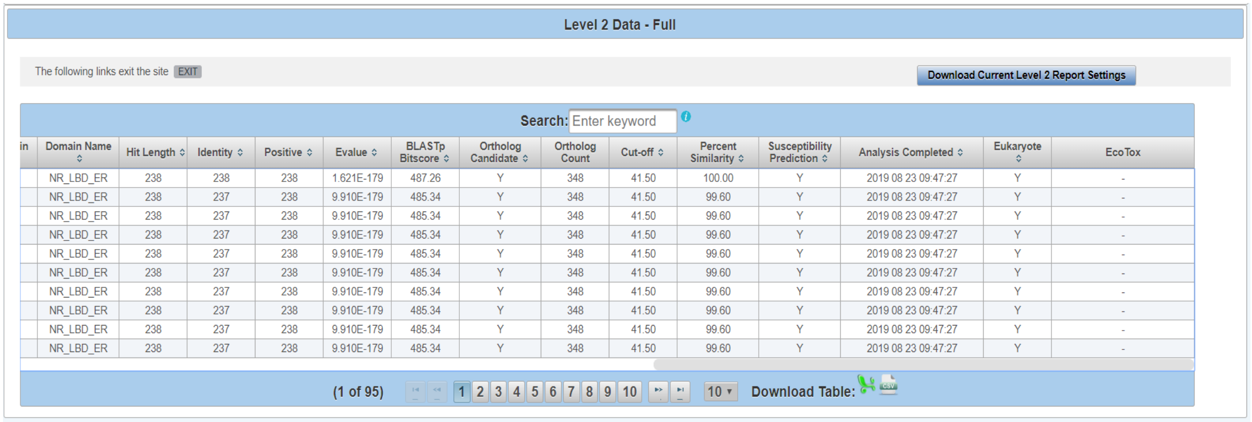
The user may also choose to turn species read across off, by using the "Species Read-Across" drop-down and selecting "No" and clicking "Update Report." When "No" is selected, the susceptibility predictions will only be "Y" in the table below if Percent Similarity is above the Cut-off or if the hit is identified as an Ortholog Candidate, yes or "Y." Any hit below the cut-off will yield a susceptibility prediction of no or "N."



E-value: 10.0
Sorted by Taxonomic Group: order
Species Read-Across: No

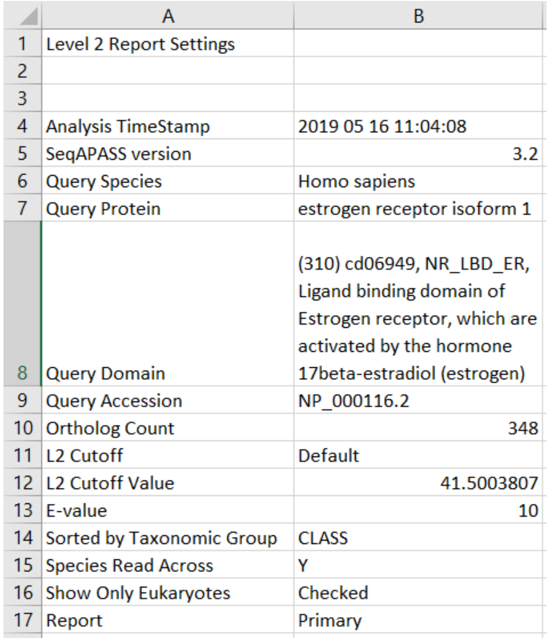
The user can select the "Full Report" on the "Level 2" data page, which includes the same information as the "Primary Report" and additional information pertaining to the alignment of the protein sequence using BLASTp and domain information. Additional information includes the NCBI PSSM ID, NCBI Domain ID, Domain Name, number of amino acid residues in the sequence (Hit Length), the number of exact matching amino acids between the hit and query sequence (Identity), the number of exact and similar

(similar side-chain substitutions) matches in amino acids between the hit and the query sequence (Positives), and the expect value (E-value) describing the number of different alignments expected to occur in the database search by chance. (See **Search, View, and Download Data Tables** section of user guide for more information).



in	Domain Name	Hit Length	Identity	Positive	Evalue	BLASTp Bitscore	Ortholog Candidate	Ortholog Count	Cut-off	Percent Similarity	Susceptibility Prediction	Analysis Completed	Eukaryote	EcoTox
	NR_LBD_ER	238	238	238	1.621E-179	487.26	Y	348	41.50	100.00	Y	2019 08 23 09:47:27	Y	-
	NR_LBD_ER	238	237	238	9.910E-179	485.34	Y	348	41.50	99.60	Y	2019 08 23 09:47:27	Y	-
	NR_LBD_ER	238	237	238	9.910E-179	485.34	Y	348	41.50	99.60	Y	2019 08 23 09:47:27	Y	-
	NR_LBD_ER	238	237	238	9.910E-179	485.34	Y	348	41.50	99.60	Y	2019 08 23 09:47:27	Y	-
	NR_LBD_ER	238	237	238	9.910E-179	485.34	Y	348	41.50	99.60	Y	2019 08 23 09:47:27	Y	-
	NR_LBD_ER	238	237	238	9.910E-179	485.34	Y	348	41.50	99.60	Y	2019 08 23 09:47:27	Y	-
	NR_LBD_ER	238	237	238	9.910E-179	485.34	Y	348	41.50	99.60	Y	2019 08 23 09:47:27	Y	-
	NR_LBD_ER	238	237	238	9.910E-179	485.34	Y	348	41.50	99.60	Y	2019 08 23 09:47:27	Y	-
	NR_LBD_ER	238	237	238	9.910E-179	485.34	Y	348	41.50	99.60	Y	2019 08 23 09:47:27	Y	-
	NR_LBD_ER	238	237	238	9.910E-179	485.34	Y	348	41.50	99.60	Y	2019 08 23 09:47:27	Y	-

Note: SeqAPASS v2.0 and newer parse the BLASTp query and hit accessions to identify all the species/accessions from identical proteins. Therefore, if a hit sequence represents multiple species, all species with the identical sequence will be found in the data tables for Level 2. To determine which sequence/species was identified from BLASTp as a hit and which sequence/species was parsed from the identical sequence, view the "Full Report" for Level, column "Identical Protein," where "N" is indicative of the original hit sequence and "Y" is the parsed sequence.

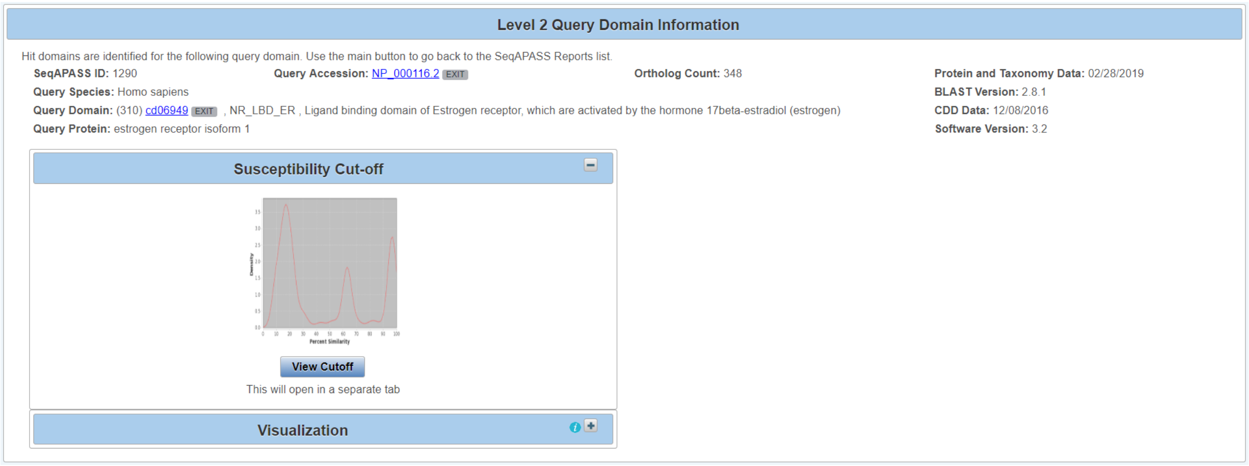


	A	B
1	Level 2 Report Settings	
2		
3		
4	Analysis TimeStamp	2019 05 16 11:04:08
5	SeqAPASS version	3.2
6	Query Species	Homo sapiens
7	Query Protein	estrogen receptor isoform 1
8	Query Domain	(310) cd06949, NR_LBD_ER, Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estrogen)
9	Query Accession	NP_000116.2
10	Ortholog Count	348
11	L2 Cutoff	Default
12	L2 Cutoff Value	41.5003807
13	E-value	10
14	Sorted by Taxonomic Group	CLASS
15	Species Read Across	Y
16	Show Only Eukaryotes	Checked
17	Report	Primary

When downloading the current level 2 report settings, the following information will be present in the csv. If the user decides to change the default settings, the csv can be utilized for quick information if the SeqAPASS page is no longer accessible.

Susceptibility Cutoff Box for Level 2

The susceptibility prediction is set by identifying ortholog candidates, sequences above a defined susceptibility cutoff, or by identifying those species below the susceptibility cut-off from an organism class above the susceptibility cutoff. The default susceptibility cut-off is set by plotting the distribution of percent similarities calculated for each hit protein. From this plot, the critical points are identified, and the local minimums and maximums reported. Using the ortholog candidate data, a susceptibility cut-off is automatically determined by identifying the first ortholog candidate at an equal or higher percent similarity than the first local minimum percent similarity. The user can view this graph by clicking the "View Cutoff" button in the "Susceptibility Cut-off" box. Radio buttons located to the right of the graphical display indicate which Cut-off has been applied for the evaluation of susceptibility in the report. These radio buttons can be selected to change the cut-off in the table to the 2nd local minimum, where the 2nd local minimum is identified in the density plot and the first ortholog candidate at an equal or higher percent similarity than the second local minimum percent similarity is used to set the cut-off. Or the user can define the local minimum by clicking on the "User Defined" radio button. Alternatively, the user can view the closely examine the density plot and manipulate the cut-off by clicking the "View Cutoff" button.



Hit domains are identified for the following query domain. Use the main button to go back to the SeqAPASS Reports list.

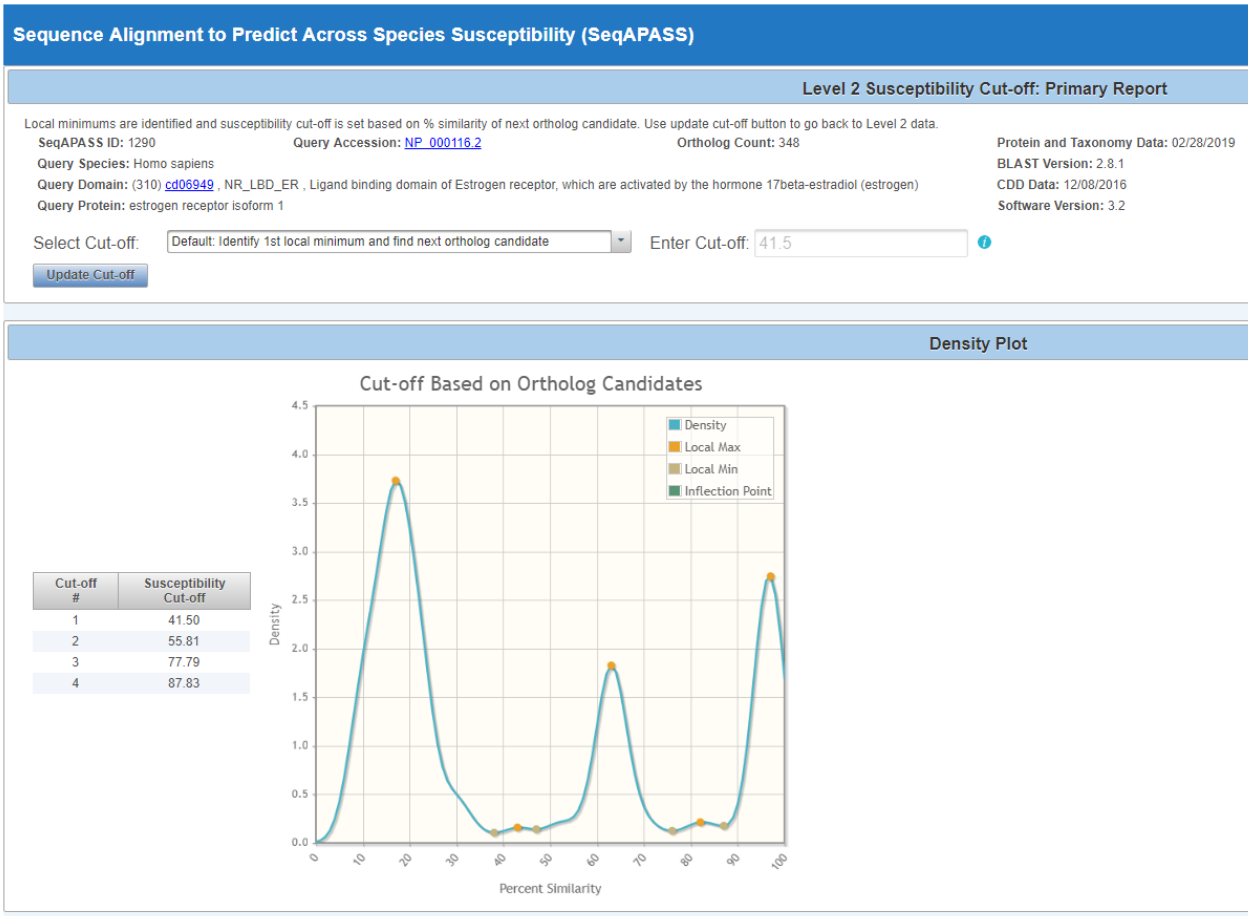
SeqAPASS ID: 1290
Query Accession: NP_000116.2
Ortholog Count: 348
Protein and Taxonomy Data: 02/28/2019
Query Species: Homo sapiens
BLAST Version: 2.8.1
Query Domain: (310) cd06949 , NR_LBD_ER , Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estrogen)
CDD Data: 12/08/2016
Query Protein: estrogen receptor isoform 1
Software Version: 3.2

This will open in a separate tab

Upon clicking "View Cutoff" button, a new page is displayed with a drop-down that allows the user to set the susceptibility cut-off using the first local minimum and the identified ortholog candidate, the second local minimum and the identified ortholog candidate, or by the "User defined cut-off" (where the user selects the cutoff). To update the cut-off in the Level 2 data report and/or return to the Level 2 page, click "Update Cut-off" button.

Note: The user should have direct empirical evidence that species above the user defined cutoff are susceptible via the protein of interest, or that the species below the user defined cutoff are not susceptible.

Upon selecting the User defined cut-off from the dropdown, the Enter Cut-off text box becomes active and the user can enter a number 1-100.



Local minimums are identified and susceptibility cut-off is set based on % similarity of next ortholog candidate. Use update cut-off button to go back to Level 2 data.

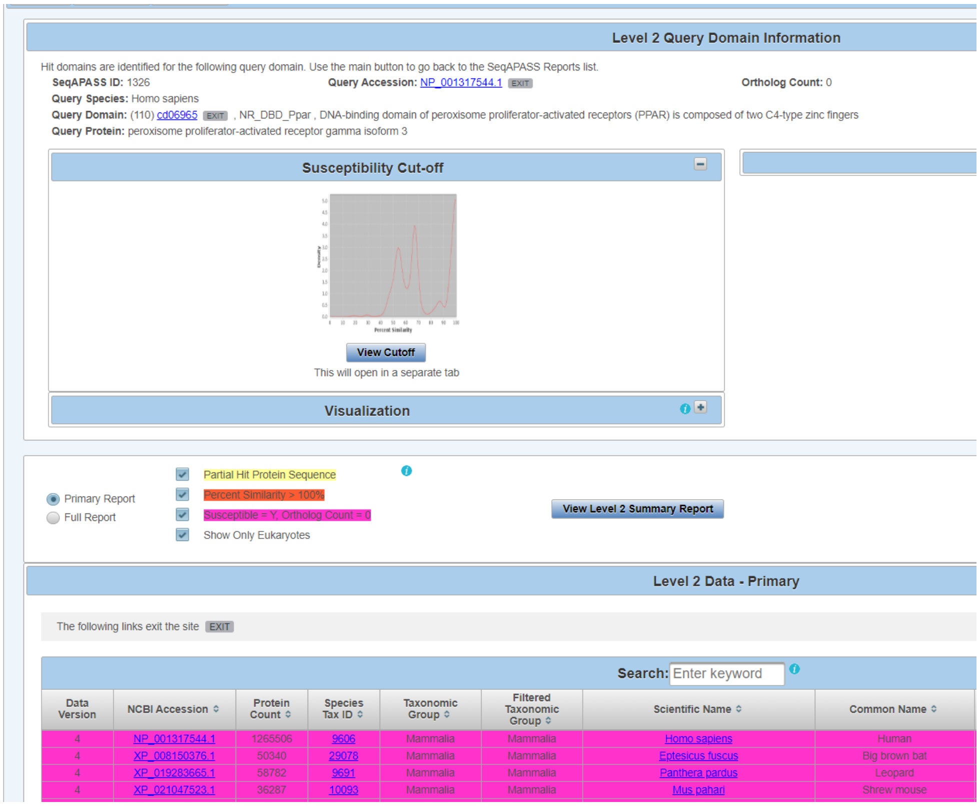
SeqAPASS ID: 1290
Query Accession: NP_000116.2
Ortholog Count: 348
Protein and Taxonomy Data: 02/28/2019
Query Species: Homo sapiens
BLAST Version: 2.8.1
Query Domain: (310) cd06949 , NR_LBD_ER , Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estrogen)
CDD Data: 12/08/2016
Query Protein: estrogen receptor isoform 1
Software Version: 3.2

Select Cut-off: Default: Identify 1st local minimum and find next ortholog candidate
Enter Cut-off: 41.5

Cut-off #	Susceptibility Cut-off
1	41.50
2	55.81
3	77.79
4	87.83

All potential susceptibility cut-offs generated by the data distribution and ortholog candidate identification are reported in the table with columns “Cut-off #” and “Susceptibility Cut-off”. The user can use these numbers to define a cut-off if empirical evidence suggests that the “Default” or “2 minimum” are not supported.

No Orthologs Detected



Level 2 Query Domain Information

Hit domains are identified for the following query domain. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1326 **Query Accession:** NP_001317544.1 EXIT **Ortholog Count:** 0
Query Species: Homo sapiens
Query Domain: (110) cd06965 EXIT , NR_DBD_Ppar , DNA-binding domain of peroxisome proliferator-activated receptors (PPAR) is composed of two C4-type zinc fingers
Query Protein: peroxisome proliferator-activated receptor gamma isoform 3

● Primary Report
○ Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

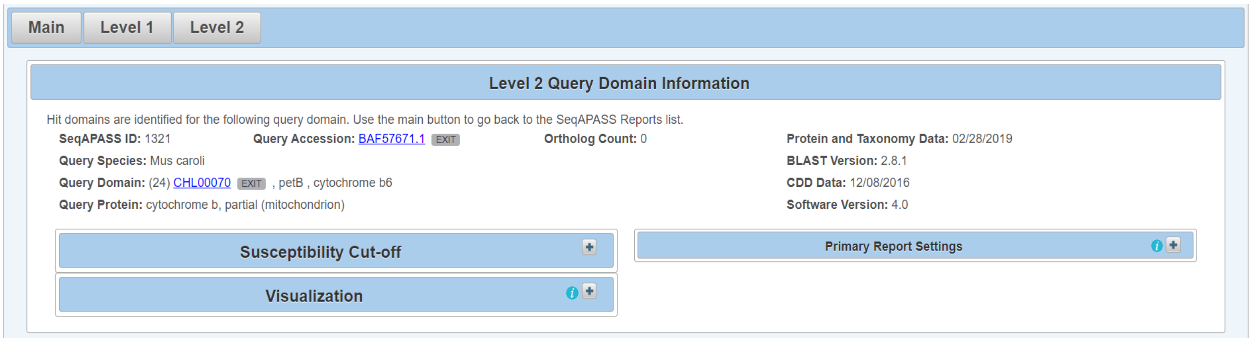
Level 2 Data - Primary

The following links exit the site EXIT

Search: Enter keyword

Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name
4	NP_001317544.1	1265506	9606	Mammalia	Mammalia	Homo sapiens	Human
4	XP_008150376.1	50340	29078	Mammalia	Mammalia	Eptesicus fuscus	Big brown bat
4	XP_019283665.1	58782	9691	Mammalia	Mammalia	Panthera pardus	Leopard
4	XP_021047523.1	36287	10093	Mammalia	Mammalia	Mus pahari	Shrew mouse

If no orthologs are detected from reciprocal best hit blast analysis, the "Ortholog Count" will be "0" at the top of the "Level 1 Query Protein Information" page. The cutoff will be set by the local minimums only, therefore the susceptibility prediction will NOT take into account ortholog candidates. ***It is recommended that the user checks the full report for Ortholog candidates or identifies a different query sequence for the susceptibility predictions***. Here, the susceptibility predictions will be highlighted in dark pink in the Level 1 data table to indicate that 0 orthologs were detected and the susceptibility cutoff was determined from plotting the distribution of percent similarities and identifying the local minimums.



Level 2 Query Domain Information

Hit domains are identified for the following query domain. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1321 **Query Accession:** BAF57671.1 EXIT **Ortholog Count:** 0 **Protein and Taxonomy Data:** 02/28/2019
Query Species: Mus caroli **BLAST Version:** 2.8.1
Query Domain: (24) CHL00070 EXIT , petB , cytochrome b6 **CDD Data:** 12/08/2016
Query Protein: cytochrome b, partial (mitochondrion) **Software Version:** 4.0

By clicking on the "View Cutoff" button when no orthologs are detected, the "Cut-off #" and "Susceptibility Cut-off" columns will report only the local minimum values.

Local minimums are identified and susceptibility cut-off is set based on % similarity of next ortholog candidate. Use update cut-off button to go back to Level 2 data.

SeqAPASS ID: 1326
Query Accession: NP_001317544.1
Ortholog Count: 0
Query Species: Homo sapiens
Query Domain: (110) cd06965 , NR_DBD_Ppar , DNA-binding domain of peroxisome proliferator-activated receptors (PPAR) is composed of two C4-type zinc fingers
Query Protein: peroxisome proliferator-activated receptor gamma isoform 3
Protein and Taxonomy Data: 02/28/2019
BLAST Version: 2.8.1
CDD Data: 12/08/2016
Software Version: 4.0

Select Cut-off: Default: Identify 1st local minimum and find next ortholog candidate
Enter Cut-off: 100.0

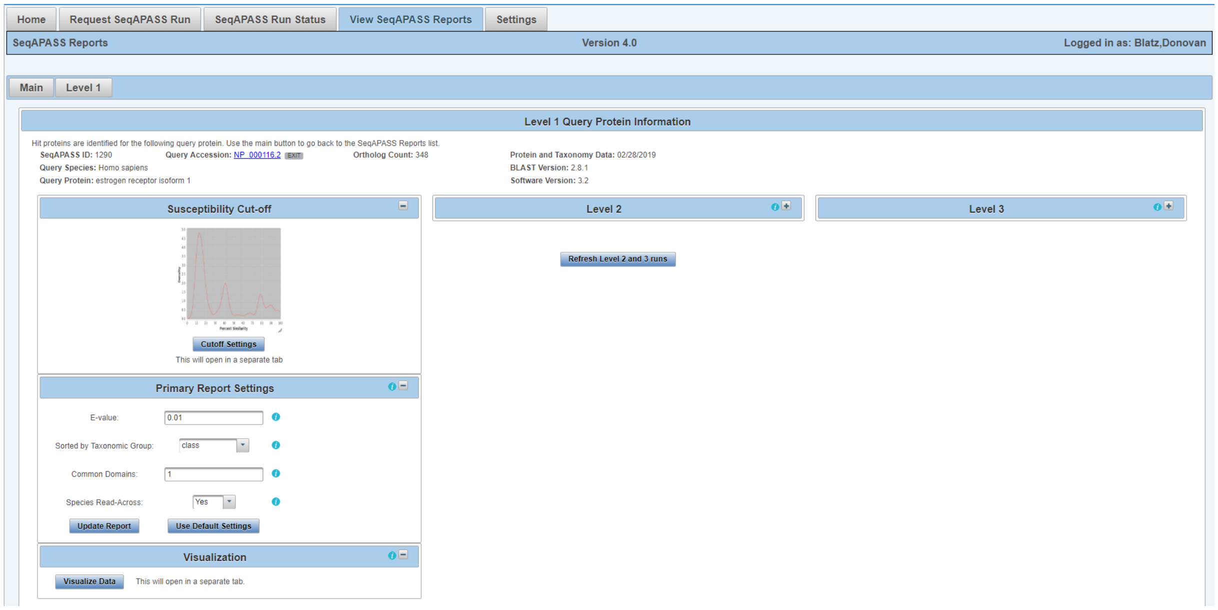
Cut-off #	Susceptibility Cut-off
1	100.00

The user can return to the “Level 2” data page by clicking the “Update Cut-off” button or exiting the tab.

Level 1 and Level 2: Data Visualization

From the Level 1 or Level 2-results page SeqAPASS users can access an interactive data visualization for both the “Primary Report” or “Full Report” by clicking on the “Visualize Data” button.

Example of Level 1 page:

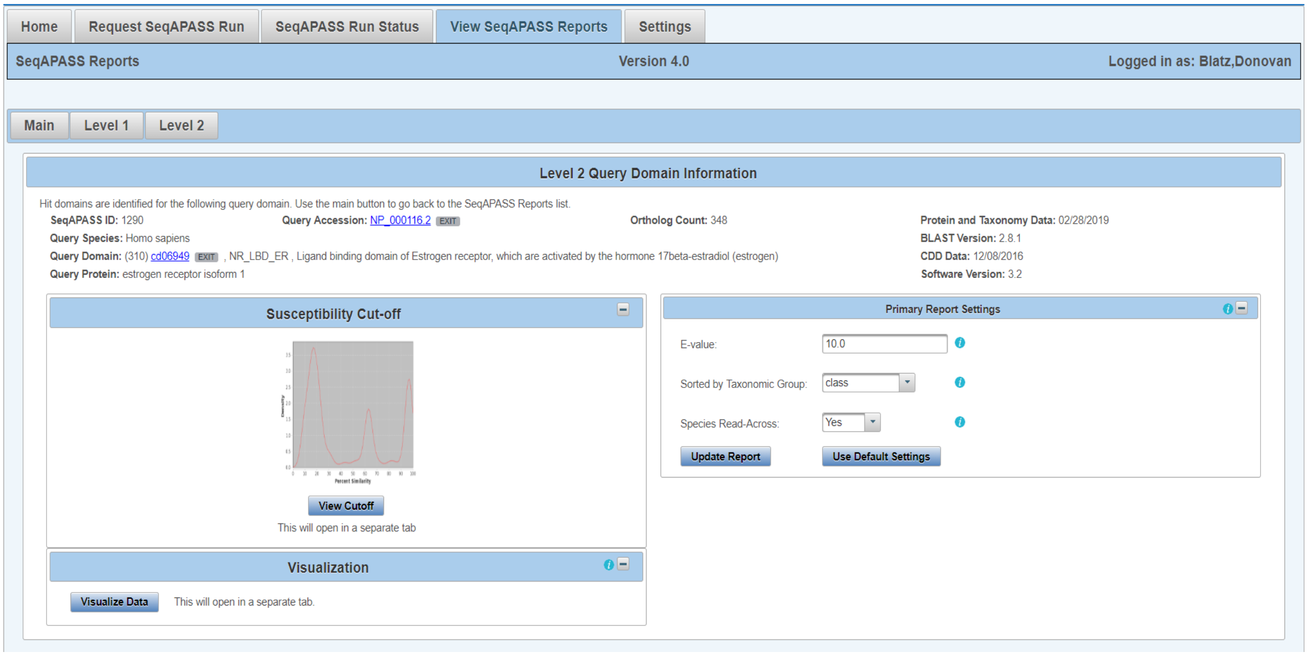


Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1290
Query Accession: NP_000116.2
Ortholog Count: 348
Query Species: Homo sapiens
Query Protein: estrogen receptor isoform 1
Protein and Taxonomy Data: 02/28/2019
BLAST Version: 2.8.1
Software Version: 3.2

E-value: 0.01
Sorted by Taxonomic Group: class
Common Domains: 1
Species Read-Across: Yes

Example of Level 2 page:



Hit domains are identified for the following query domain. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1290
Query Accession: NP_000116.2
Ortholog Count: 348
Protein and Taxonomy Data: 02/28/2019
Query Species: Homo sapiens
BLAST Version: 2.8.1
Query Domain: (310) cd06949 , NR_LBD_ER , Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estrogen)
CDD Data: 12/08/2016
Query Protein: estrogen receptor isoform 1
Software Version: 3.2

This will open in a separate tab

This will open in a separate tab.

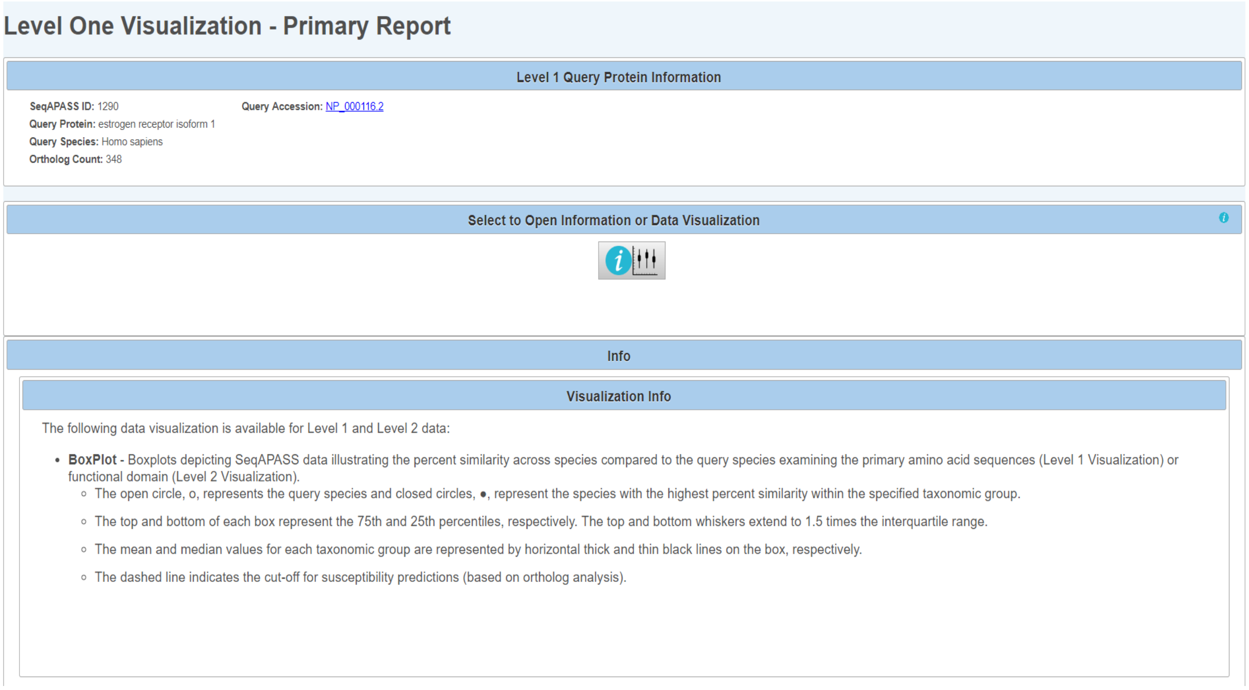
E-value: 10.0
Sorted by Taxonomic Group: class
Species Read-Across: Yes

The data visualization will then open in a new web browser tab, one for Level 1 and a different one for Level 2. The visualization will display for the report selected by the user on the Level 1 or Level 2 report page and be identified as "Level One Visualization – Primary Report" or "Level One Visualization – Full Report" and "Level Two Visualization – Primary Report" or "Level Two Visualization – Full Report."

Note: One report type at a time, either "Primary Report" or "Full Report," can be displayed in the visualization tab for Level 1 and Level 2. Therefore, if the user is viewing the "Level One Visualization – Primary Report" page and returns to the Level 1 results page and clicks the radio button for "Full Report," the data visualization tab will update to "Level One Visualization – Full Report."

Level 1 and 2 Information Page

The initial page that opens upon clicking the "Visualize Data" button provides the respective level query protein information, including SeqAPASS ID, query protein, query species, ortholog count, and query accession information. A link out to the NCBI protein database page corresponding to the queried accession is available by clicking the query accession. Information on the visualization is provided in the "Visualization Info" text box. To view the data visualization boxplots click the BoxPlot icon. The BoxPlot will then generate below the Visualization Info box.



Level One Visualization - Primary Report

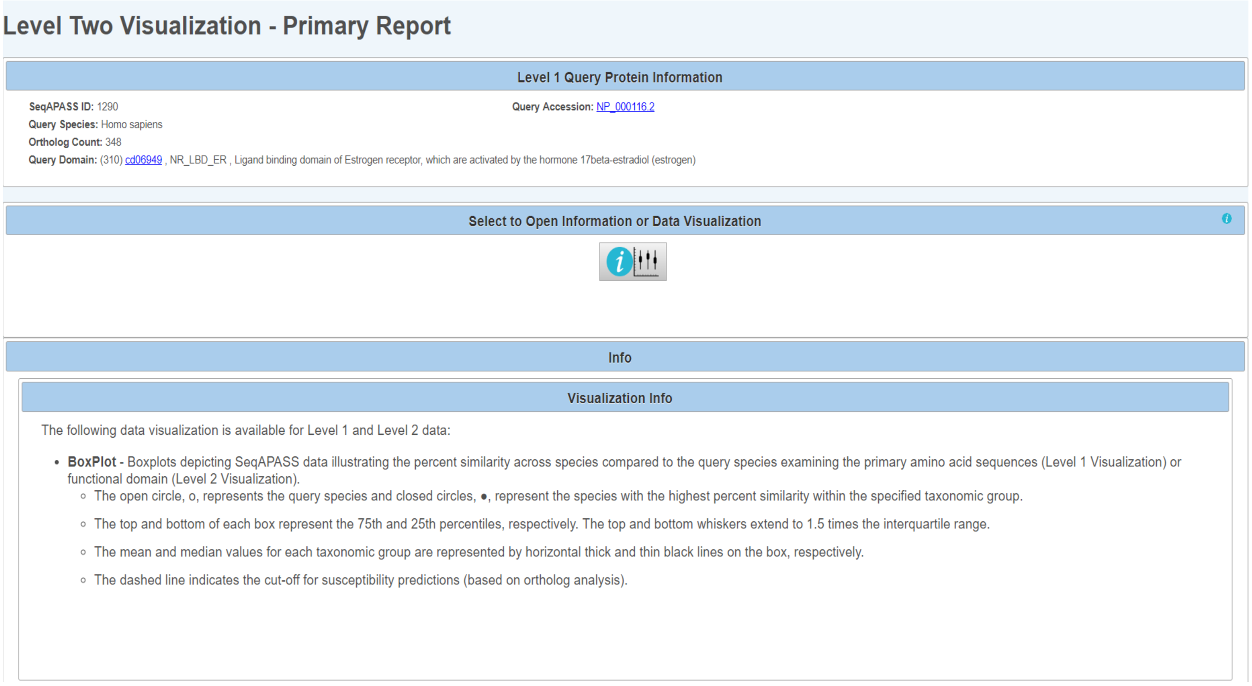
Level 1 Query Protein Information

SeqAPASS ID: 1290
Query Protein: estrogen receptor isoform 1
Query Species: Homo sapiens
Ortholog Count: 348
Query Accession: NP_000116.2

Visualization Info

The following data visualization is available for Level 1 and Level 2 data:

- **BoxPlot** - Boxplots depicting SeqAPASS data illustrating the percent similarity across species compared to the query species examining the primary amino acid sequences (Level 1 Visualization) or functional domain (Level 2 Visualization).
  - The open circle, o, represents the query species and closed circles, •, represent the species with the highest percent similarity within the specified taxonomic group.
  - The top and bottom of each box represent the 75th and 25th percentiles, respectively. The top and bottom whiskers extend to 1.5 times the interquartile range.
  - The mean and median values for each taxonomic group are represented by horizontal thick and thin black lines on the box, respectively.
  - The dashed line indicates the cut-off for susceptibility predictions (based on ortholog analysis).



Level Two Visualization - Primary Report

Level 1 Query Protein Information

SeqAPASS ID: 1290
Query Species: Homo sapiens
Ortholog Count: 348
Query Domain: (310) cd06949 , NR_LBD_ER , Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estrogen)
Query Accession: NP_000116.2

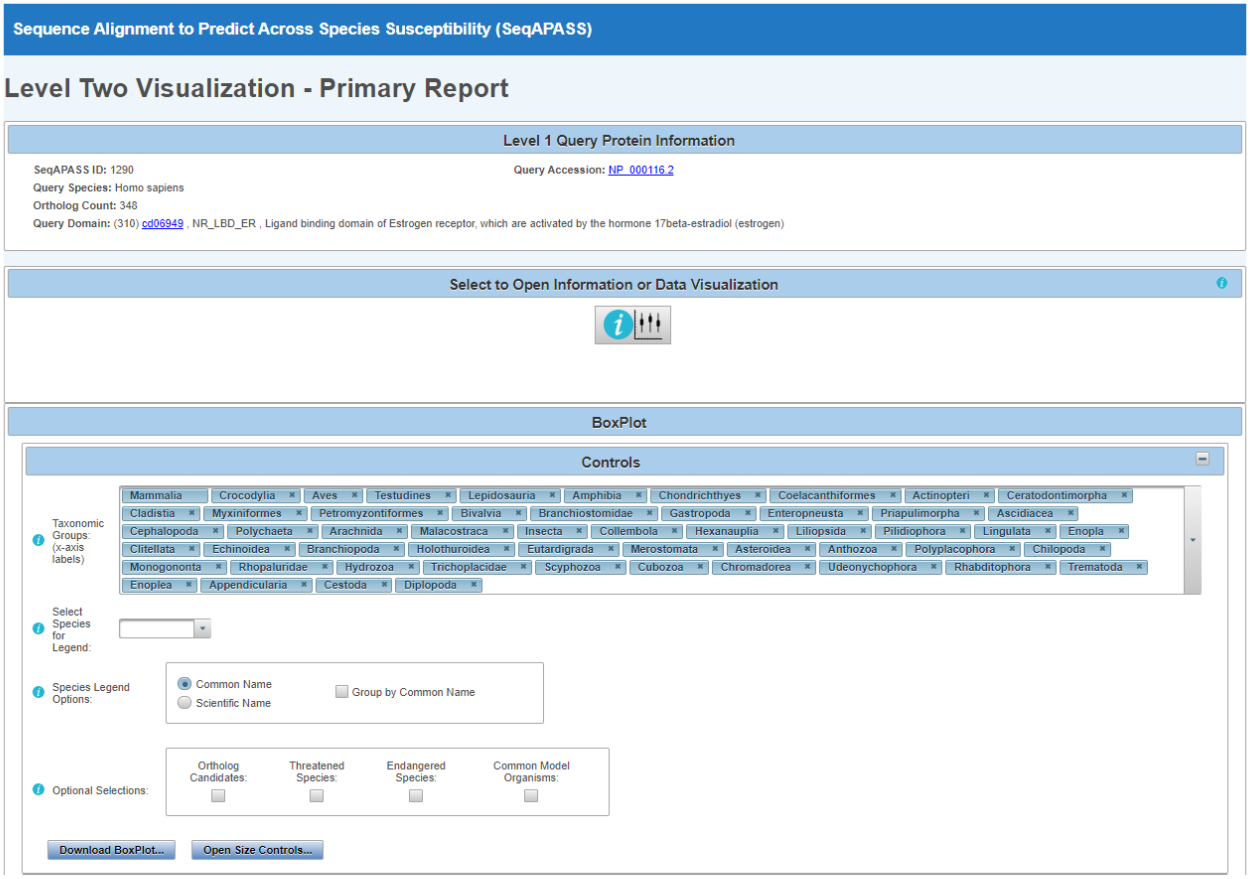
Visualization Info

The following data visualization is available for Level 1 and Level 2 data:

- **BoxPlot** - Boxplots depicting SeqAPASS data illustrating the percent similarity across species compared to the query species examining the primary amino acid sequences (Level 1 Visualization) or functional domain (Level 2 Visualization).
  - The open circle, o, represents the query species and closed circles, •, represent the species with the highest percent similarity within the specified taxonomic group.
  - The top and bottom of each box represent the 75th and 25th percentiles, respectively. The top and bottom whiskers extend to 1.5 times the interquartile range.
  - The mean and median values for each taxonomic group are represented by horizontal thick and thin black lines on the box, respectively.
  - The dashed line indicates the cut-off for susceptibility predictions (based on ortholog analysis).

Level 1 and 2 BoxPlot Page – Controls

Upon clicking the "BoxPlot" icon on either Level 1 or Level 2 Visualization Information pages, a box for the boxplot "Controls" and a box for the interactive boxplot will open, respectively.



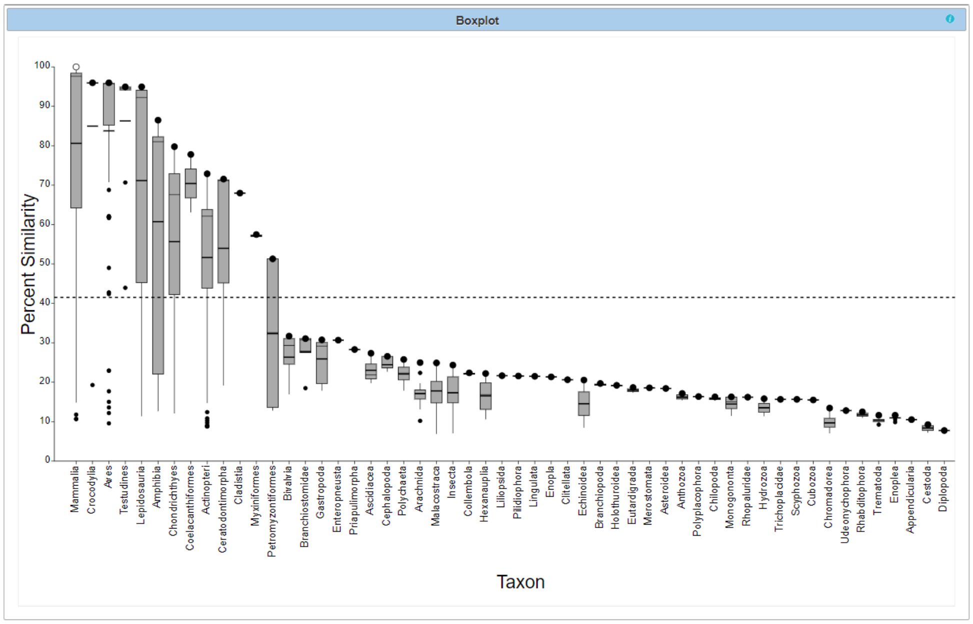
SeqAPASS ID: 1290
Query Species: Homo sapiens
Ortholog Count: 348
Query Domain: (310) cd06949 , NR_LBD_ER , Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estrogen)
Query Accession: NP_000116.2

Taxonomic Groups: (x-axis labels) Mammalia, Crocodylia, Aves, Testudines, Lepidosauria, Amphibia, Chondrichthyes, Coelacanthiformes, Actinopteri, Ceratodontimorpha, Cladistia, Myxiniformes, Petromyzontiformes, Bivalvia, Branchiostomidae, Gastropoda, Enteropneusta, Priapulimorpha, Ascidiacea, Cephalopoda, Polychaeta, Arachnida, Malacostraca, Insecta, Collembola, Hexanauplia, Liliopsida, Pilidiophora, Lingulata, Enopla, Clitellata, Echinoidea, Branchiopoda, Holothuroidea, Eutardigrada, Merostomata, Asteroidea, Anthozoa, Polyplacophora, Chilopoda, Monogononta, Rhopaluridae, Hydrozoa, Trichoplacidae, Scyphozoa, Cubozoa, Chromadorea, Udeonychophora, Rhabditophora, Trematoda, Enoplea, Appendicularia, Cestoda, Diplopoda

Select Species for Legend:

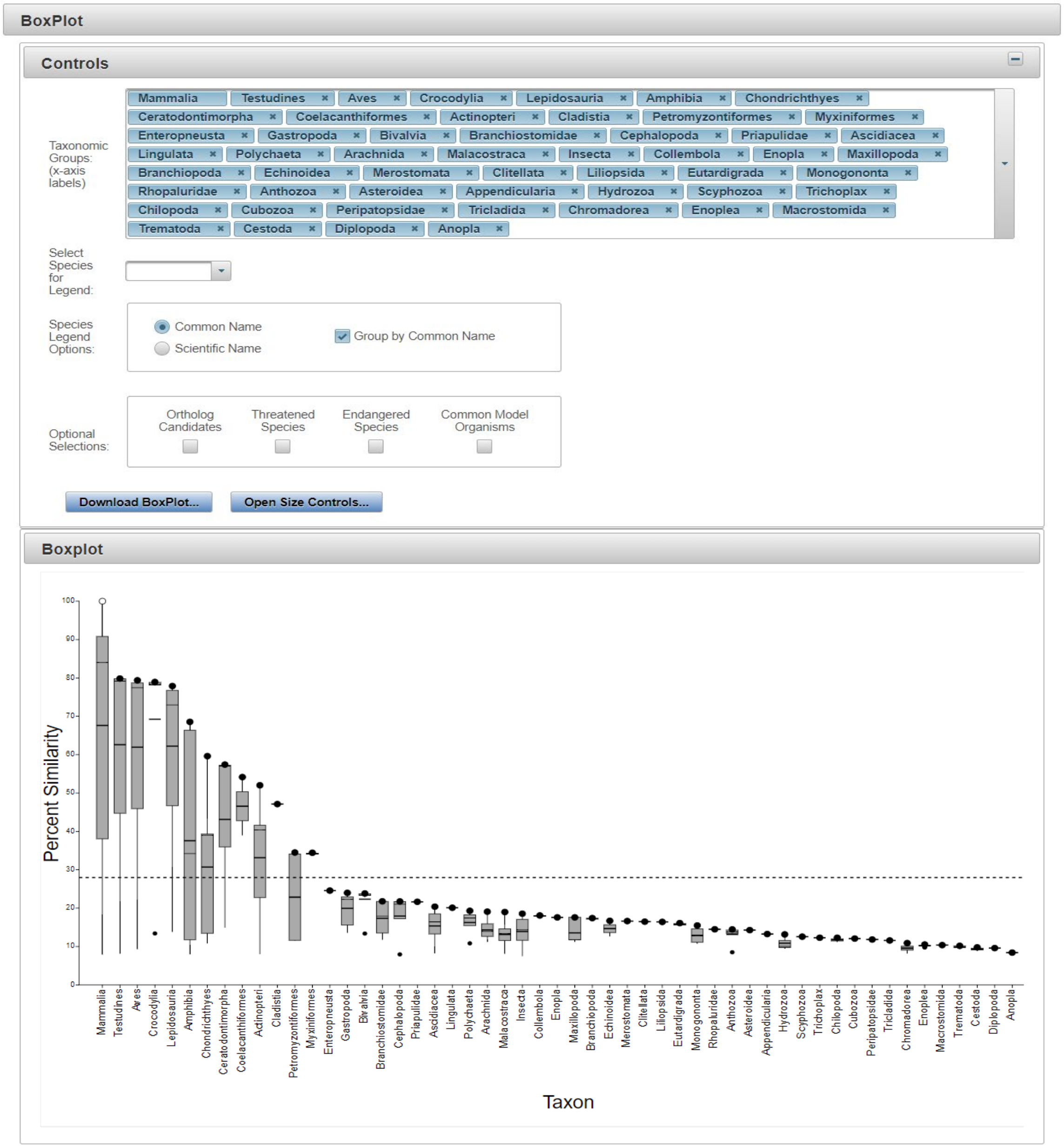
Species Legend Options: ● Common Name ○ Scientific Name ☐ Group by Common Name

Optional Selections: Ortholog Candidates: ☐ Threatened Species: ☐ Endangered Species: ☐ Common Model Organisms: ☐



Manipulating Taxonomic Groups on x-axis

The boxplot controls allow the user to edit the taxonomic groups that are displayed on the x-axis by clicking on the “X” for the Taxonomic Group name (e.g., Aves). This action removes the selected group from the x-axis. To the right of the “Taxonomic Groups” controls box is a drop-down that allows the user to remove or add back taxonomic groups to the x-axis of the boxplot graphic, by deselecting or selecting check-boxes in the dropdown. Similarly, unwanted taxonomic groups may be removed directly from the boxplot by hovering the cursor over the taxonomic groups listed along the x-axis. The user will notice that the selection arrow changes to a black arrow with a red ‘x’ next to it; clicking the taxonomic group will then remove it from the boxplot and the “Taxonomic Groups” controls box. The user can delete multiple species by pressing CTRL and either clicking individual species or slowly dragging across multiple species. Additionally, that taxonomic group will have the checkbox deselected in the “Taxonomic Groups” controls box drop-down list.



Taxonomic Groups: (x-axis labels) Mammalia, Testudines, Aves, Crocodylia, Lepidosauria, Amphibia, Chondrichthyes, Ceratodontimorpha, Coelacanthiformes, Actinopteri, Cladistia, Petromyzontiformes, Myxiniformes, Enteropneusta, Gastropoda, Bivalvia, Branchiostomidae, Cephalopoda, Priapulidae, Ascidiacea, Lingulata, Polychaeta, Arachnida, Malacostraca, Insecta, Collembola, Enopla, Maxillopoda, Branchiopoda, Echinoidea, Merostomata, Clitellata, Liliopsida, Eutardigrada, Monogononta, Rhopaluridae, Anthozoa, Asteroidea, Appendicularia, Hydrozoa, Scyphozoa, Trichoplax, Chilopoda, Cubozoa, Peripatopsidae, Tricladida, Chromadorea, Enoplea, Macrostomida, Trematoda, Cestoda, Diplopoda, Anopla

Select Species for Legend:

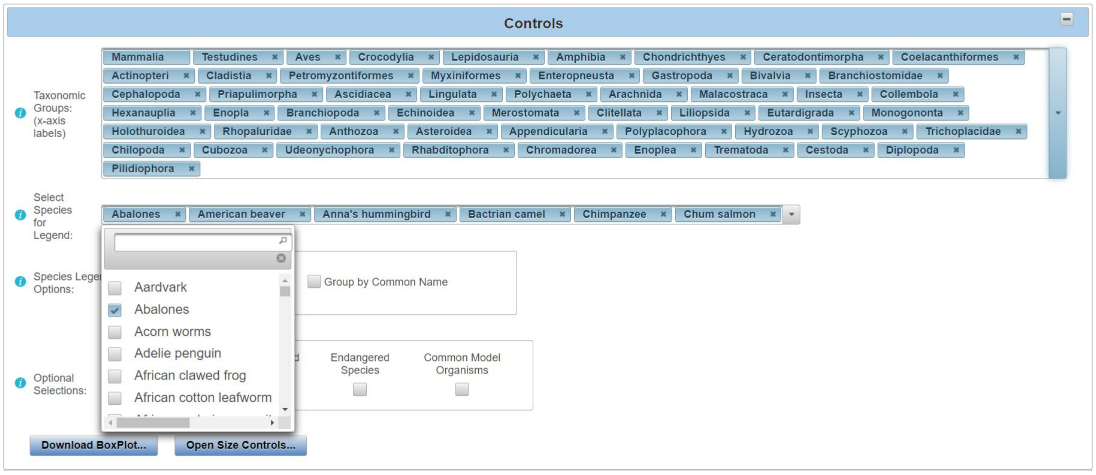
Species Legend Options: ● Common Name ○ Scientific Name ☑ Group by Common Name

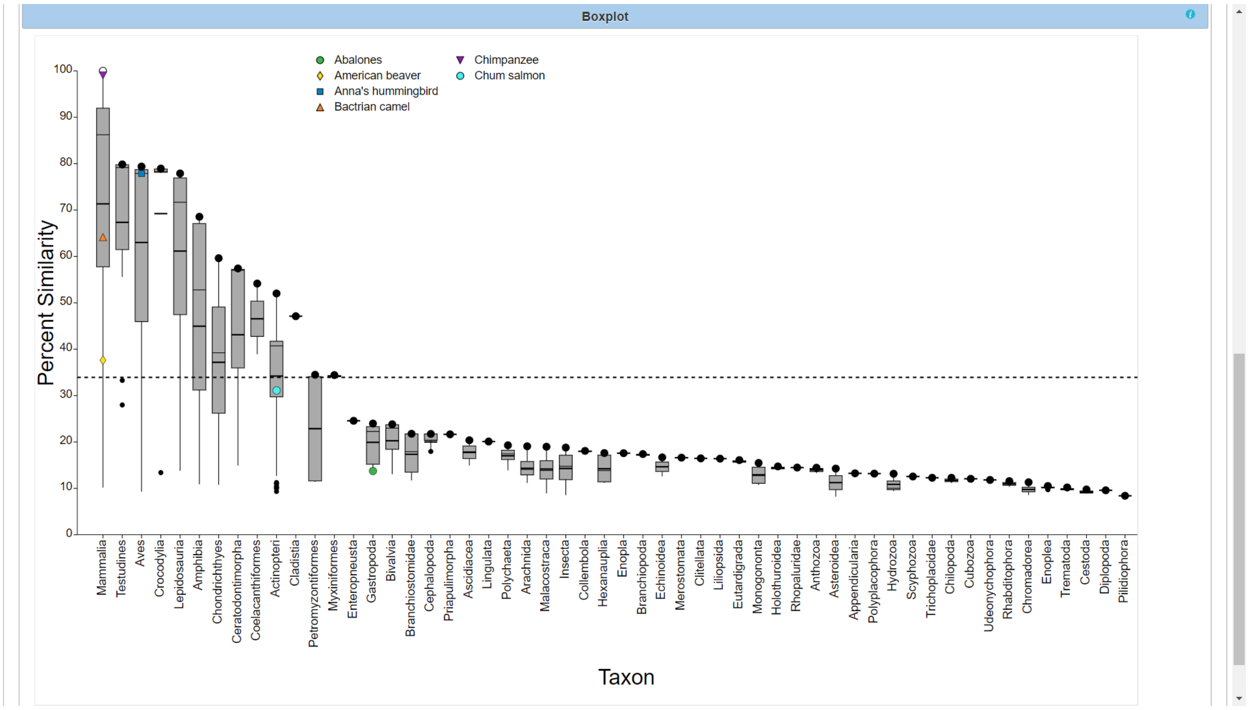
Optional Selections: ☐ Ortholog Candidates ☐ Threatened Species ☐ Endangered Species ☐ Common Model Organisms

Customize BoxPlot Legend

The user may customize the boxplot by adding a legend that will pinpoint species of interest on the boxplot. Upon clicking the drop-down for "Select Species for Legend" in the controls box the user may search in the text box for specific species to display in the boxplot legend. Upon identifying a species from the drop-down menu and selecting the checkbox the species name will be placed in the boxplot legend and a corresponding data point will be produced on the graph. The default settings display the species common name both in the "Select Species for Legend" dropdown and on the boxplot. However, if the species scientific name is desired, the user can select the radio button for "Scientific Name" in the controls box for "Species Legend Options." This action will change the drop-down menu and species in the legend to display the species scientific name.

Note: The database will take a brief moment to update the list upon changing between "Common Name" and "Scientific Name."

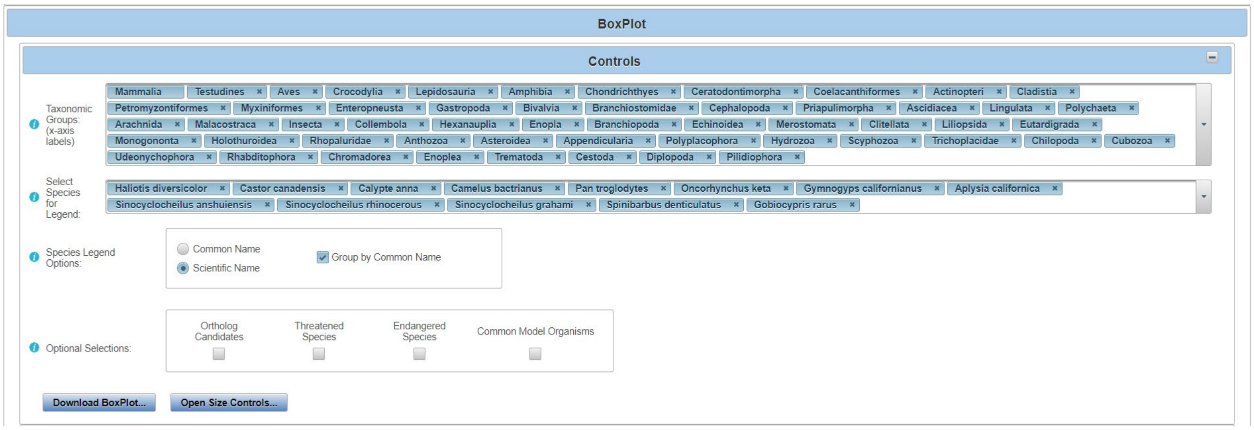




Change Species Display on Plot

Multiple scientific names can be represented by only one common name (e.g., Common name: Teleost fishes; corresponding scientific names: *Spinibarbus denticulatus*, *Sinocyclocheilus rhinocerous*, *Sinocyclocheilus grahami*, *Sinocyclocheilus anshuiensis*, *Gobiocypris rarus*, *Thamnaconus septentrionalis*). Therefore, if a species common name that represents multiple species was used to create the legend, and the user decides to instead select "Scientific Name," by default the boxplot legend will change to display multiple scientific names that representing the individual common name and each scientific name will be represented by a unique color/shape point on the plot. However, if the user selects the checkbox "Group by Common Name" in the "Species Legend Options" control box, then the scientific names that are represented by one common name will all display the same color/shape point on the plot.

The user has the option of removing selected species from the legend either by removing them directly from the "Select Species for Legend" drop-down box or by hovering the mouse directly over the species name in the legend. The mouse will change to a black arrow with a red 'x' next to it. Clicking the name while this arrow is displayed will remove the species from the legend and from the control box.

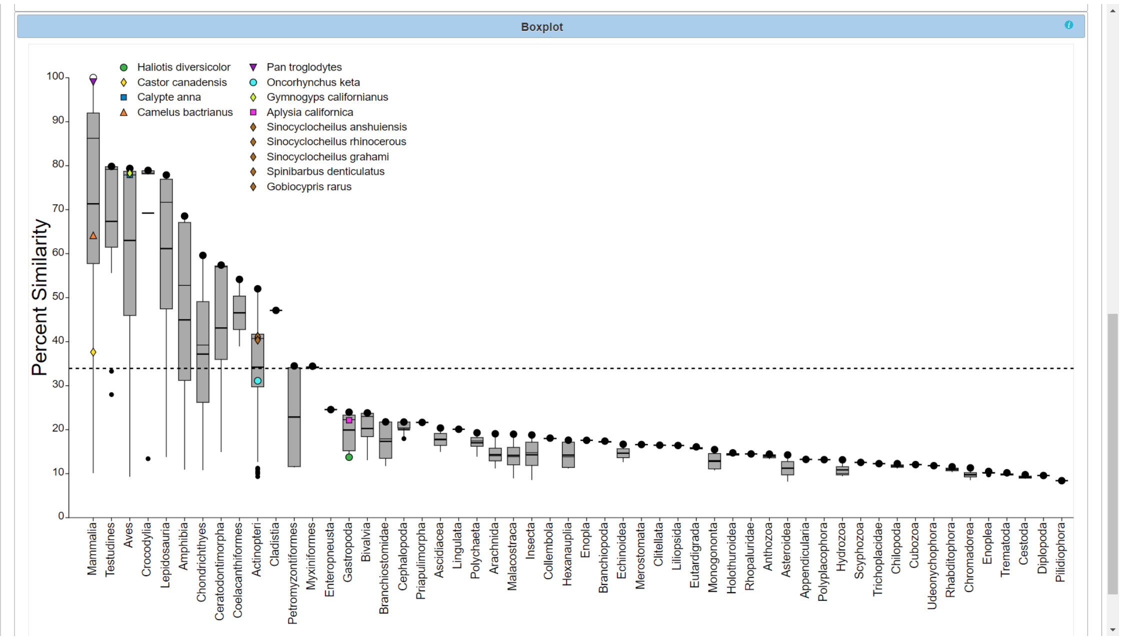


Taxonomic Groups: (x-axis labels) Mammalia, Testudines, Aves, Crocodylia, Lepidosauria, Amphibia, Chondrichthyes, Ceratodontimorpha, Coelacanthiformes, Actinopteri, Cladistia, Petromyzontiformes, Myxiniformes, Enteropneusta, Gastropoda, Bivalvia, Branchiostomidae, Cephalopoda, Priapulimorpha, Ascidiacea, Lingulata, Polychaeta, Arachnida, Malacostraca, Insecta, Collembola, Hexanauplia, Enopla, Branchiopoda, Echinoidea, Merostomata, Clitellata, Liliopsida, Eutardigrada, Monogononta, Holothuroidea, Rhopaluridae, Anthozoa, Asteroidea, Appendicularia, Polyplacophora, Hydrozoa, Scyphozoa, Trichoplacidae, Chilopoda, Cubozoa, Udeonychophora, Rhabditophora, Chromadorea, Enoplea, Trematoda, Cestoda, Diplopoda, Pilidiophora

Select Species for Legend: Haliotis diversicolor, Castor canadensis, Calypte anna, Camelus bactrianus, Pan troglodytes, Oncorhynchus keta, Gymnogyps californianus, Aplysia californica, Sinocyclocheilus anshuiensis, Sinocyclocheilus rhinocerous, Sinocyclocheilus grahami, Spinibarbus denticulatus, Gobiocypris rarus

Species Legend Options: ○ Common Name ● Scientific Name ☑ Group by Common Name

Optional Selections: Ortholog Candidates ☐ Threatened Species ☐ Endangered Species ☐ Common Model Organisms ☐

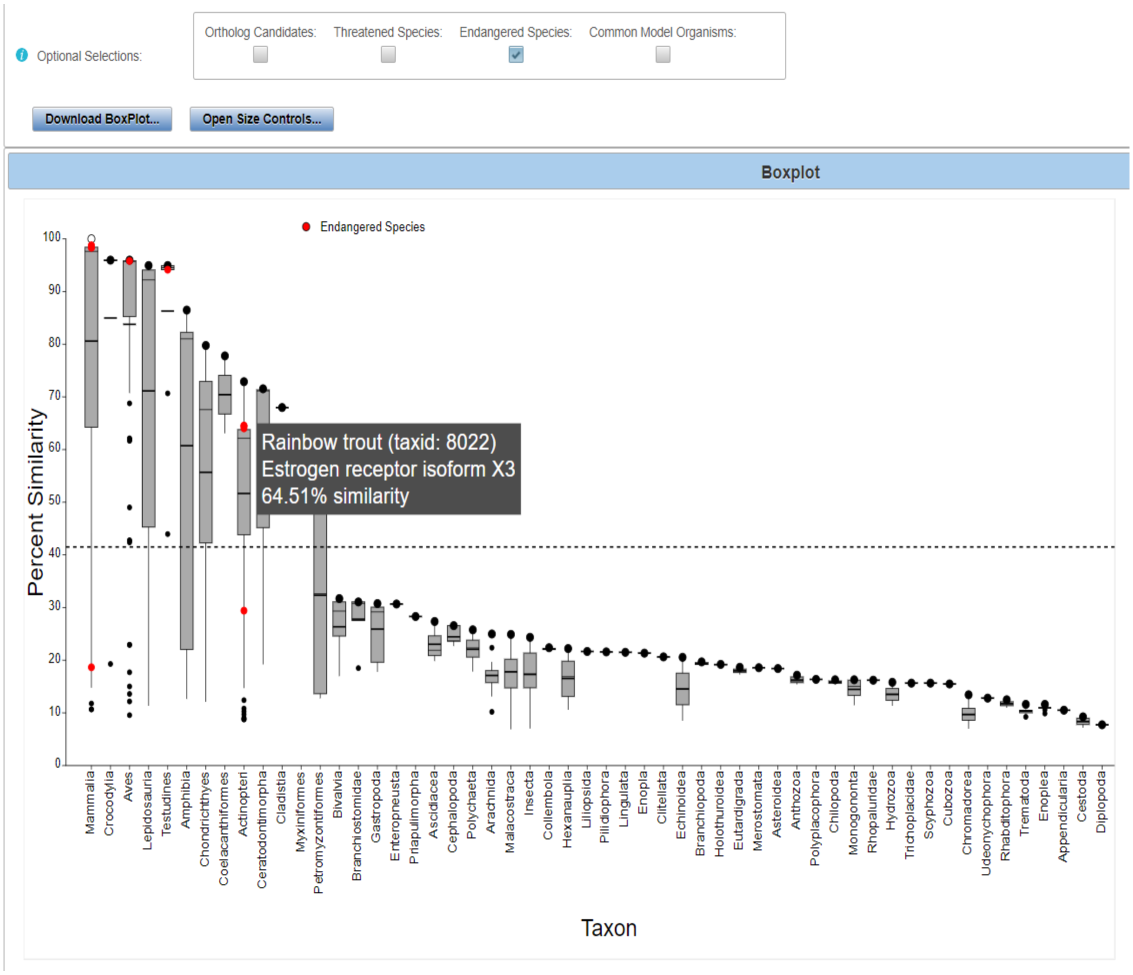


Customize the Legend to Display Species Groups of Interest

In the "Optional Selections" controls box, the user has the option of displaying "Ortholog Candidates," "Threatened Species," "Endangered Species," or "Common Model Organisms." Upon selecting one of the checkboxes, red data points corresponding to species will be displayed on the boxplot. By hovering the mouse over a single red point, a pop-up box will appear with the corresponding species name, taxonomic ID, query protein, and percent similarity.

Note: The user can select to display either species common name or scientific name in the hover over information box by selecting from the "Species Legend Options."

If the user selects either "Threatened Species" or "Endangered Species," clicking on an individual red dot will open a new web browser tab and link to the corresponding species page on th US Fish and Wildlife Service's Environmental Conservation Online System (USFWS, ECOS; e.g.,) (https://ecos.fws.gov/ecp0/profile/speciesProfile?sId=1506).



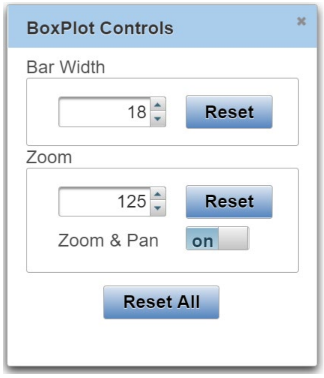
Optional Selections:
Ortholog Candidates: ☐ Threatened Species: ☐ Endangered Species: ☑ Common Model Organisms: ☐

Rainbow trout (taxid: 8022)
Estrogen receptor isoform X3
64.51% similarity

BoxPlot Controls Widget for Bar Width, Zoom and Pan

By clicking the “Open Size Controls” button, a “BoxPlot Controls” widget opens that allows the user to adjust the size of the bars on the boxplot by increasing or decreasing the “Bar Width” using the up and down arrows. The minimum and maximum size for bars are 6 and 60, respectively. To reset the bar width on the boxplot to default size, click the “Reset” button to the right of the “Bar Width” adjustment box in the “BoxPlot Controls” box. The user can also Zoom and Pan the boxplot by toggling the on /off button under the “Zoom” heading. The user can then zoom in or out by clicking the up or down arrows or entering a number in the text box and clicking enter. To reset the zoom on the boxplot to default size, click the “Reset” button to the right of the “Zoom” adjustment box in the “BoxPlot Controls” widget. The pan option is available when the “Zoom and Pan” option is toggled to the “on” position, which allows the user to click on the boxplot and drag the plot around the screen to reposition. To reset all BoxPlot Controls to default settings click the “Reset All” button.

<u>Note:</u> Upon exiting out of the BoxPlot Controls widget, the Zoom and Pan options are automatically turned off.



Bar Width: 18

Zoom: 125

Zoom & Pan: on

Download BoxPlot Widget

To download the boxplot, click “Download BoxPlot” button in the controls box. A “Download Boxplot” Widget will pop up. It will be necessary to specify which type of file (SVG, PNG, or JPG,) to downloaded by clicking on the desired radio button for “Image Type.” The user may customize the resolution of the boxplot for PNG and JPG files prior to download by altering the “Width” and “Height” of the BoxPlot. To change “Width” or “Height,” enter the desired number in the text boxes. Click “Download Image” button to download the file. To close the “Download Boxplot” widget, click the “x” on the top right of the widget.

Image Type: ○ SVG ● PNG ○ JPG

Width: 1,236

Height: 755

Hover-over Features in the BoxPlot

By hovering over a taxonomic group name on the x-axis of the box plot, an information box will pop-up listing the top three species in order by highest percent similarity. If only one or two species are represented in the taxonomic group, then only those species will be displayed. Hovering the mouse over any of the species in the boxplot, that is present in the legend, will generate a pop-up box with the corresponding species name, taxonomic ID, query protein, and percent similarity. The susceptiblity cut-off is displayed in a pop-up text box upon hovering over the dashed horizontal cut-off line.

Summary Table for Species in a Specific Taxonomic Group

By clicking on a box representing a taxonomic group in the boxplot a table will pop-up providing summary information for that particular group. The table header will provide summary statistics (i.e., mean and median percent similarity), including the Taxonomic Group name, number of species represented in the box, the overall susceptiblity prediciton for the selected taxonomic group. Data table includes protein and species information along with metrics for evaluated protein similarity and predicting suseptiblity. Also inlcuded in the table are columns indicating if a species belongs to a certain group of interest (e.g., Threatened Species; Endangered Species, Model Organism). Table can be downloaded by clicking on the icon for excel or csv file.

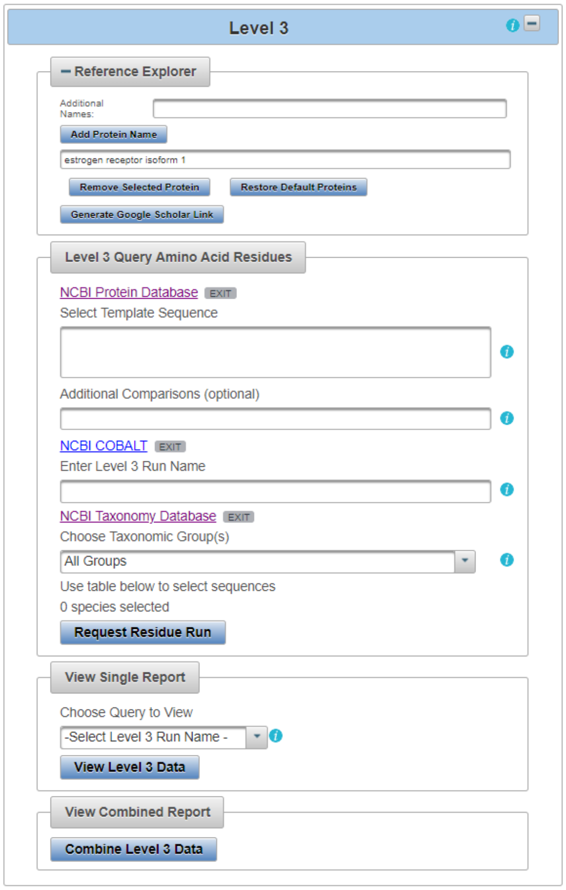
Interactive Visualization with Level 1 Data Page and Level 2 Data Page

The data visualization is programmed to update with changes made to the Level 1 Data page and Level 1 Data page, respectively. Therefore, if the user updates the Susceptibility Cut-off (See user guide section **Susceptibility Cutoff Box for Level 1** and **Susceptibility Cutoff Box for Level 2)** to the “Second Local Minimum” or “User Defined Cut-off,” the previously opened data visualization boxplot tab will update the cut-off accordingly. Similarly, the user modifies the Primary Report Settings (See user guide section **Level 1: Primary Report Settings** and **Level 2: Primary Report Settings)**, the data visualization will update accordingly.

Note: If the user updates the “Primary Report Settings” for “Sorted by Taxonomic Group” the boxplot will update to display the new taxonomic group selection that is present in the “Filtered Taxonomic Group” column in the data table. The user should be aware that manipulating the “Sorted by Taxonomic Group” to a different level in the taxonomic lineage (e.g., from class to order; from class to genus) adds a larger number of taxonomic groups to the x-axis. Therefore, the plot may require greater user manipulation using the BoxPlot Controls to view the data.

Level 3: Individual Amino Acid Residue Alignment

In the “View SeqAPASS Reports” tab, on the “Level 1 Query Protein Information” page, there is a “Level 3” dropdown for setting up the query for comparing individual amino acid residues to a template sequence. It is anticipated that the choice of template sequence and residues that are selected to align will be derived from the published literature in most cases. Publications evaluating homology models, protein crystal structures, pesticide field resistance, or utilizing site-directed mutagenesis are a few examples of the types of studies that may contain such information to guide a Level 3 SeqAPASS evaluation.



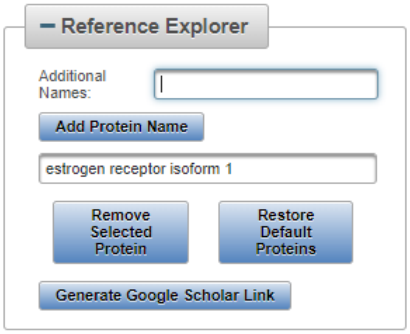
estrogen receptor isoform 1

Choose Taxonomic Group(s): All Groups

Use table below to select sequences

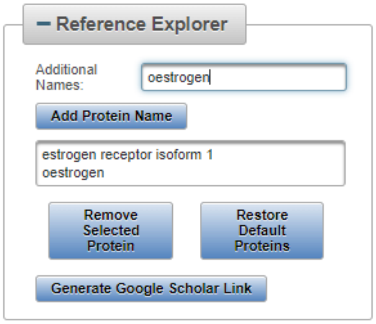
0 species selected

Relevant literature containing these data can be identified using the SeqAPASS “Reference Explorer.” The user can search for literature with the protein(s) of interest with an auto-populated search term that is integrated into a predefined Boolean string and generate a Google Scholar link that will take them to scientific articles containing their protein(s).



estrogen receptor isoform 1

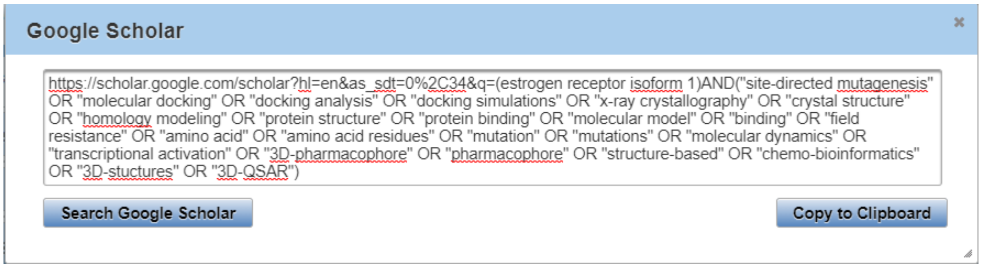
The user can modify the Boolean search string by adding text to the "Additional Names" text box and clicking the "Add Protein Name" button. By selecting a name that is currently in the text box and clicking the "Remove Selected Protein" button, the user can delete names from the text box and therefore these names will not be included in the Boolean string for the Google Scholar search.



Additional Names: oestrogen

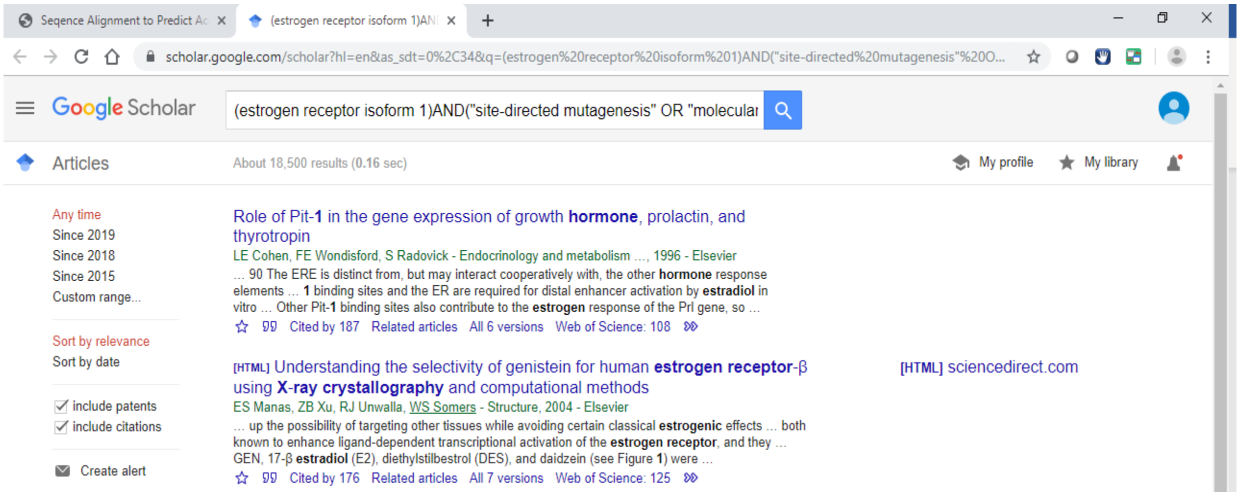
estrogen receptor isoform 1
oestrogen

When satisfied with the protein names to be included in the Boolean search string, the user will select the "Generate Google Scholar Link" button. A pop-up will appear displaying the Boolean sting to be searched in Google Scholar. The user can continue to modify the Boolean string by clicking in the text and adding additional information. The Boolean string can be copied and pasted elsewhere by the user by clicking the "Copy to Clipboard" button. The user can also choose to use the generated Boolean string to search Google Scholar. To do so the user will select the "Search Google Scholar" button.



```
https://scholar.google.com/scholar?hl=en&as_sdt=0%2C34&q=(estrogen receptor isoform 1)AND("site-directed mutagenesis"
OR "molecular docking" OR "docking analysis" OR "docking simulations" OR "x-ray crystallography" OR "crystal structure"
OR "homology modeling" OR "protein structure" OR "protein binding" OR "molecular model" OR "binding" OR "field
resistance" OR "amino acid" OR "amino acid residues" OR "mutation" OR "mutations" OR "molecular dynamics" OR
"transcriptional activation" OR "3D-pharmacophore" OR "pharmacophore" OR "structure-based" OR "chemo-bioinformatics"
OR "3D-stuctures" OR "3D-QSAR")
```

Upon selecting the "Search Google Scholar" button, a new tab will be generated in the browser for Google Scholar that contains the Boolean string in the search with publications and articles that matched the SeqAPASS generated Boolean sting. The literature displayed by Google Scholar for the user should be evaluated to identify appropriate articles for determining Level 3 template sequences and critical individual amino acids for comparisons across species.



scholar.google.com/scholar?hl=en&as_sdt=0%2C34&q=(estrogen%20receptor%20isoform%201)AND("site-directed%20mutagenesis"%20O...

(estrogen receptor isoform 1)AND("site-directed mutagenesis" OR "molecular

About 18,500 results (**0.16** sec)

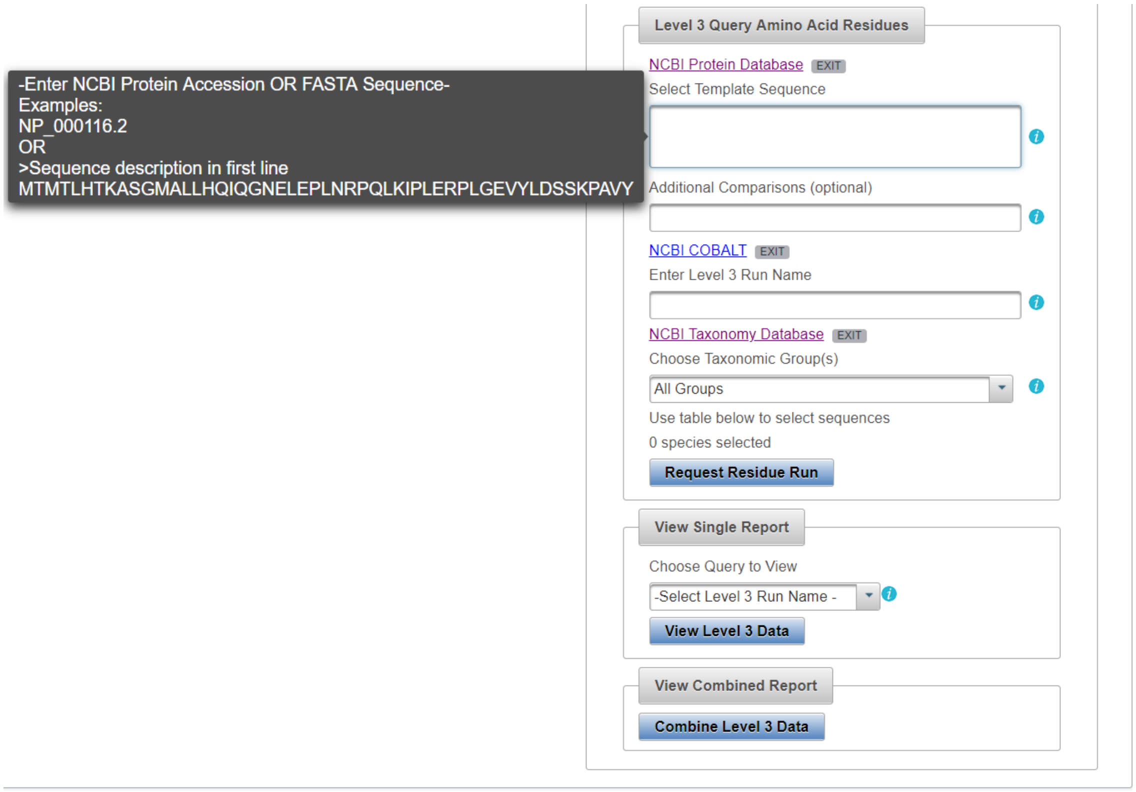
Role of Pit-**1** in the gene expression of growth **hormone**, prolactin, and thyrotropin
LE Cohen, FE Wondisford, S Radovick - Endocrinology and metabolism ..., 1996 - Elsevier
... 90 The ERE is distinct from, but may interact cooperatively with, the other **hormone** response elements ... **1** binding sites and the ER are required for distal enhancer activation by **estradiol** in vitro ... Other Pit-**1** binding sites also contribute to the **estrogen** response of the Prl gene, so ...
Cited by 187 Related articles All 6 versions Web of Science: 108

[HTML] Understanding the selectivity of genistein for human **estrogen receptor**-β using **X-ray crystallography** and computational methods
ES Manas, ZB Xu, RJ Unwalla, WS Somers - Structure, 2004 - Elsevier
... up the possibility of targeting other tissues while avoiding certain classical **estrogenic** effects ... both known to enhance ligand-dependent transcriptional activation of the **estrogen receptor**, and they ... GEN, 17-β **estradiol** (E2), diethylstilbestrol (DES), and daidzein (see Figure **1**) were ...
Cited by 176 Related articles All 7 versions Web of Science: 125

[HTML] sciencedirect.com

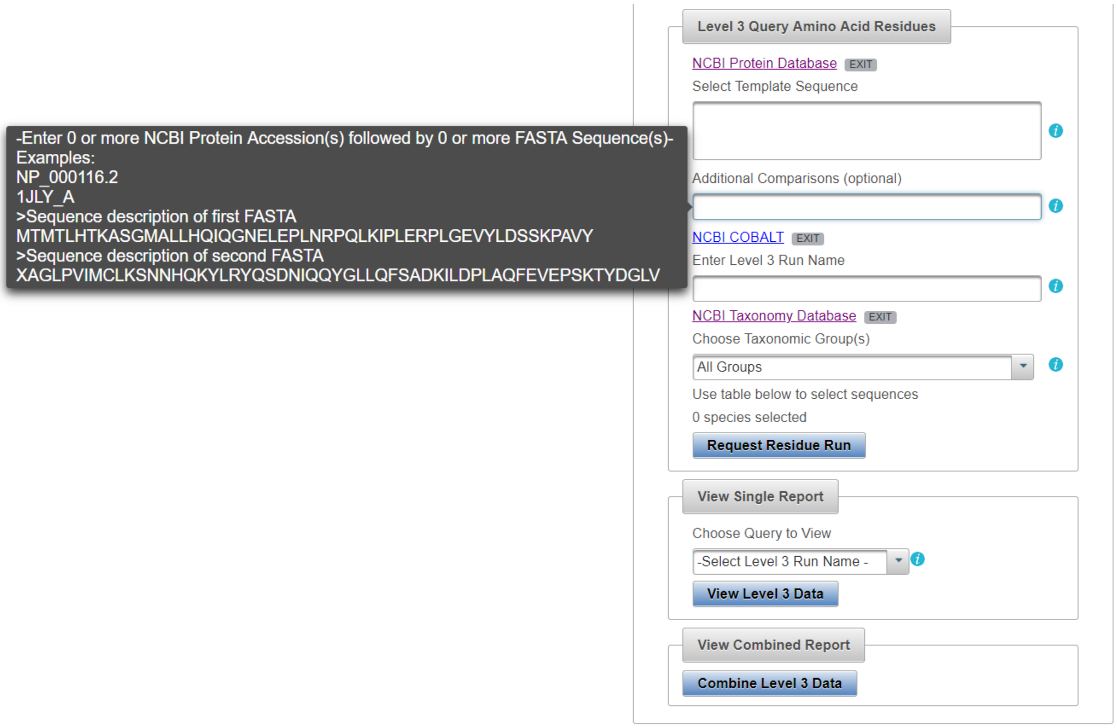
In the "Level 3" box, there is a link out to the "NCBI Protein Database" for identifying the template sequence of interest. Below this link the user will find a text box where the user can enter an NCBI Protein Accession with the version number (e.g., NP_000116.2) or a FASTA formatted sequence (e.g., <
>gi|62821794|ref|NP_000116.2| estrogen receptor isoform 1 [Homo sapiens]
MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGEVYLDSSKPAVYNYPEGAAYEFNA
AAAANA
QVYGQTGLPYGPGSEAAAFGSNGLGGFPPLNSVSPSPLMLLHPPPQLSPFLQPHGQQVPYYLENE
PSGYT
VREAGPPAFYRPNSDNRRQGGRERLASTNDKGSMAMESAKETRYCAVCNDYASGYHYGVWSC
EGCKAFFK
RSIQGHNDYMCPATNQCTIDKNRRKSCQACRLRKCYEVGMMKGGIRKDRRGGRMLKHKRQRD
DGEGRGEV
GSAGDMRAANLWPSPLMIKRSKKNSLALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASMMG
LLTNLA
DRELVHMINWAKRVPGFVDLTLHDQV).

Upon clicking on in the "Select Template Sequence" text box, a pop-up message will appear to provide examples for the proper format of Accessions or FASTA files to be entered. A link out to the NCBI Protein Database is available for the user and found above the template entry text box.


-Enter NCBI Protein Accession OR FASTA Sequence-
Examples:
NP_000116.2
OR
>Sequence description in first line
MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGEVYLDSSKPAVY

Additional sequences can (this is an optional field the user can choose to fill in) also be incorporated into the Level 3 alignment using the "Additional Comparisons (optional)" text box. Upon clicking on the "Additional Comparisons (optional)" text box, a pop-up message will appear to provide examples for the proper format of Accessions or FASTA files to be entered.

Note: In the "Additional Comparisons (optional)" text box, zero or more NCBI Protein Accession must be entered prior to FASTA sequence(s) if they are to be included in the Level 3 alignment.



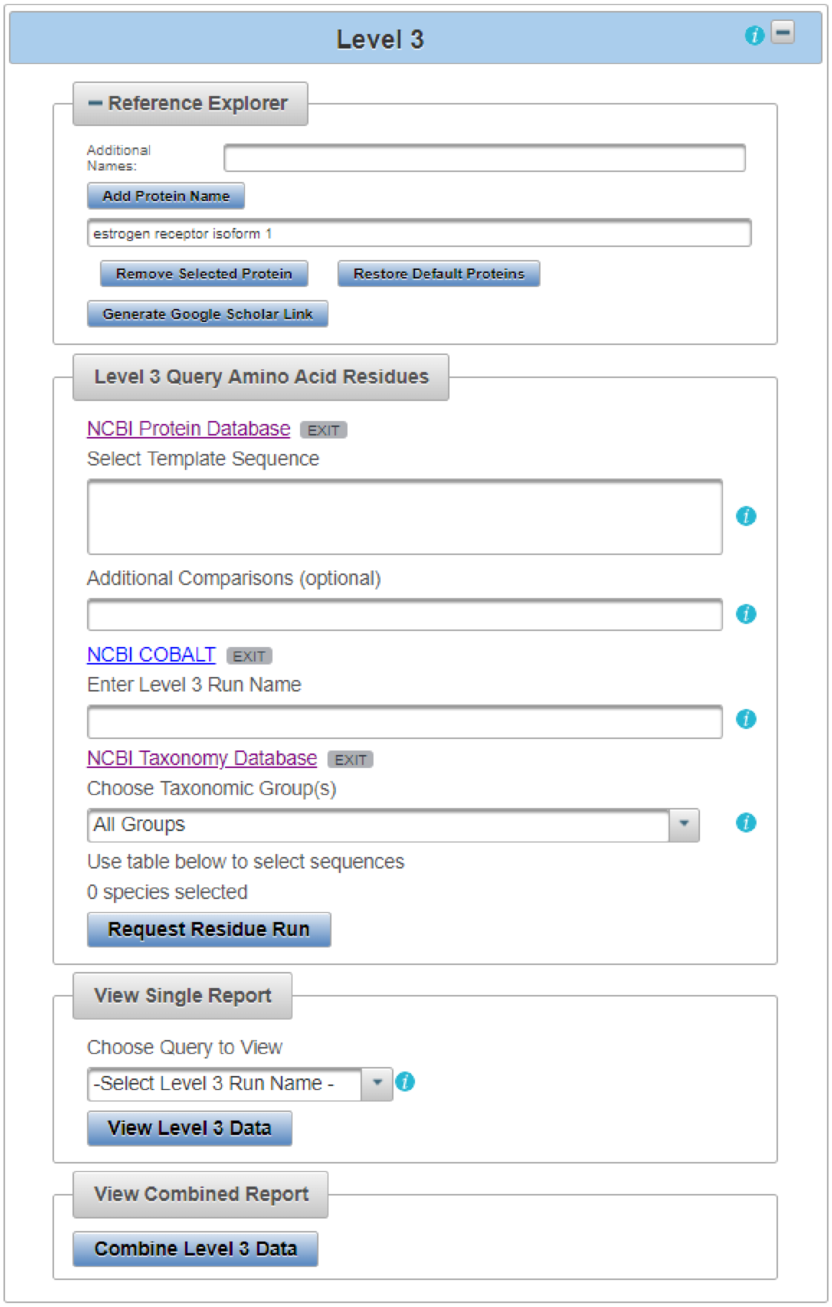
-Enter 0 or more NCBI Protein Accession(s) followed by 0 or more FASTA Sequence(s)-
Examples:
NP_000116.2
1JLY_A
>Sequence description of first FASTA
MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGEVYLDSSKPAVY
>Sequence description of second FASTA
XAGLPVIMCLKSNNHQKYLRYQSDNIQQYGLLQFSADKILDPLAQFEVEPSKTYDGLV

All Groups
0 species selected
-Select Level 3 Run Name -

Below the text box where the user can choose to add additional sequences for comparison, is a link to NCBI COBALT (Constraint-based Multiple Protein Alignment Tool). The NCBI COBALT allows the user to align multiple sequences and is the alignment tool that SeqAPASS algorithms utilize to set up the query of individual amino acid residues across species.

Note: The user does not need to use the COBALT link to run a Level 3 evaluation, however the link is available in case the user chooses to further evaluate or compare multiple potential template sequences.

Under the text "Enter Level 3 Run Name," there is a text box where the user can enter a user defined name for the run. The user may only enter letters or integers as text for the name. The user defined name will appear in the "View Level 3 Data" dropdown upon completion of the Level 3 sequence alignment.

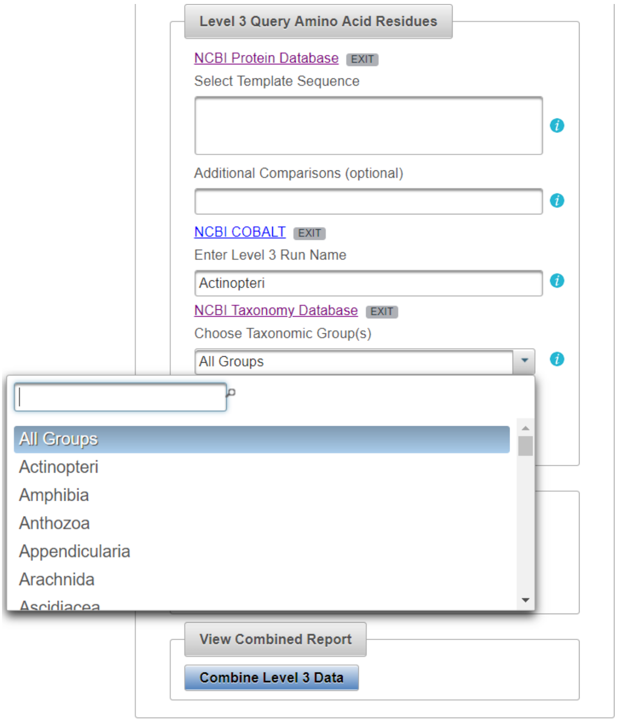


estrogen receptor isoform 1

Choose Taxonomic Group(s): All Groups

0 species selected

To complete the set-up for a Level 3 query the user must select which sequences to compare to the identified template sequence. Listed in the Choose Taxonomic Group(s) drop-down are all Taxonomic Groups that were identified as hits in the "Level 1" primary amino acid sequence alignment data. Because COBALT is used to align all sequences that are selected, it is recommended that the user selectively identify sequences from the hit table below to align. For example, selecting sequences with low similarity to the template sequence along with sequences sharing high similarity to the template sequence can skew the alignment because COBALT is trying to align all the sequences together. It is recommended that the user select sequences by first selecting a taxonomic group from the "Choose Taxonomic Group(s) drop-down. The user can also use the NCBI taxonomy link to type in the name of the "Taxonomic Groups" found in the drop-down to look up which species fall in that group.

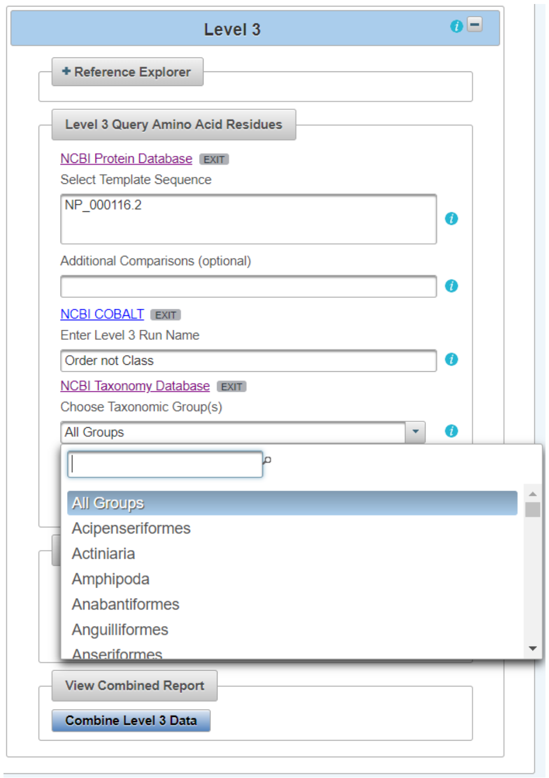


Select Template Sequence:

Enter Level 3 Run Name: Actinopteri

Choose Taxonomic Group(s): All Groups

Note: The “Choose Taxonomic Group(s):” drop-down will display the level of the taxonomic hierarchy being displayed in the “Filtered Taxonomic Group” column of the Level 1 Data table. For example, if the user changes the default option from “class” to “order,” then “order will be displayed in the dropdown.



Select Template Sequence: NP_000116.2

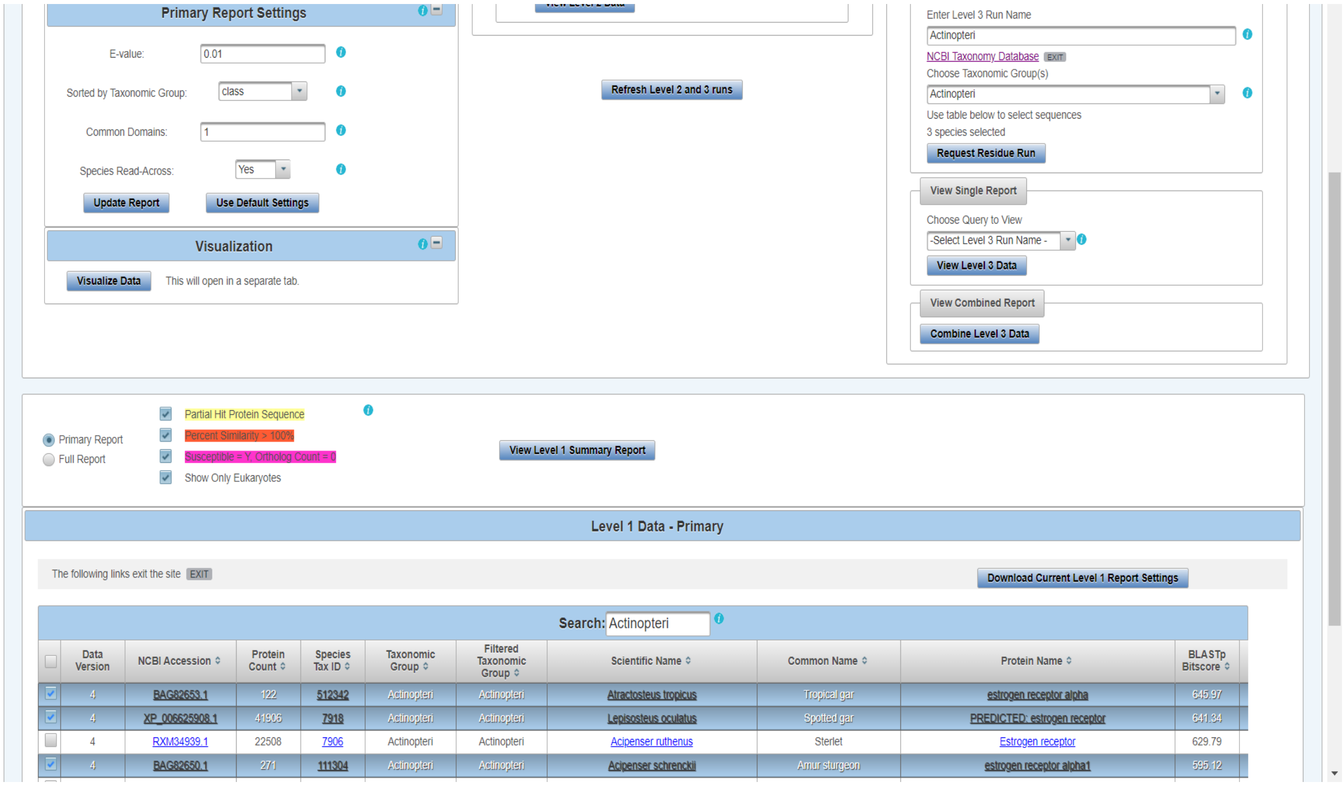
Enter Level 3 Run Name: Order not Class

Choose Taxonomic Group(s): All Groups

By choosing a group from the drop-down menu, the "Level 1" table below will be filtered by the selected Taxonomic Group (see column "Taxonomic Group" in Level 1 data table). When a "Taxonomic group is selected from the drop-down, it can take up to a few seconds for the Level 1 data table to filter completely, depending on the size of the table. The user can then examine each hit protein in the Level 1 table and select those that they would like to compare to the template sequence. To select sequences/species from the filtered Level 1 data table, the user will select the check boxes in the first column of the table. Although it is not typically recommended, the user may also select the header check box in the first column to select all sequences/species in the filtered table.

Note: The user can also type the "Taxonomic Group" of interest in the text search box at the top of the drop-down for quick filtering.

Below is an example where the user selected the "Taxonomic Group" Actinopteri from the drop-down and then selected individual sequences/species to align with the template sequence. The number of selected species will be shown in the text above the "Request Residue Run" button.



E-value: 0.01
Sorted by Taxonomic Group: class
Common Domains: 1
Species Read-Across: Yes

This will open in a separate tab.

Enter Level 3 Run Name: Actinopteri
Choose Taxonomic Group(s): Actinopteri
Use table below to select sequences
3 species selected

Choose Query to View: -Select Level 3 Run Name -

● Primary Report
○ Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

The following links exit the site

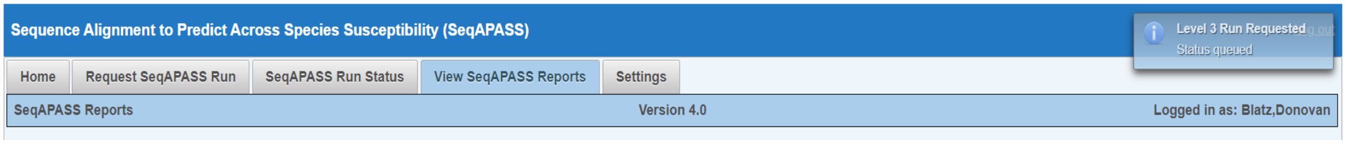
Search: Actinopteri

☐	Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name	Protein Name	BLASTp Bitscore
☑	4	BAG82653.1	122	512342	Actinopteri	Actinopteri	Atractosteus tropicus	Tropical gar	estrogen receptor alpha	645.97
☑	4	XP_006625908.1	41906	7918	Actinopteri	Actinopteri	Lepisosteus oculatus	Spotted gar	PREDICTED: estrogen receptor	641.34
☐	4	RXM34939.1	22508	7906	Actinopteri	Actinopteri	Acipenser ruthenus	Sterlet	Estrogen receptor	629.79
☑	4	BAG82650.1	271	111304	Actinopteri	Actinopteri	Acipenser schrenckii	Amur sturgeon	estrogen receptor alpha1	595.12

(See **Search, View, and Download Data Tables** section of user guide for more information)

The user can choose to align sequences/species from multiple taxonomic groups with the template sequence, by going back to the "Choose Taxonomic Group" drop-down and selecting another group, which filters the Level 1 table based on the group selected, and then the user can select additional species from the newly filtered table. As before, the number of selected species can be tracked in the text above the "Request Residue Run" button that reads "X species selected."

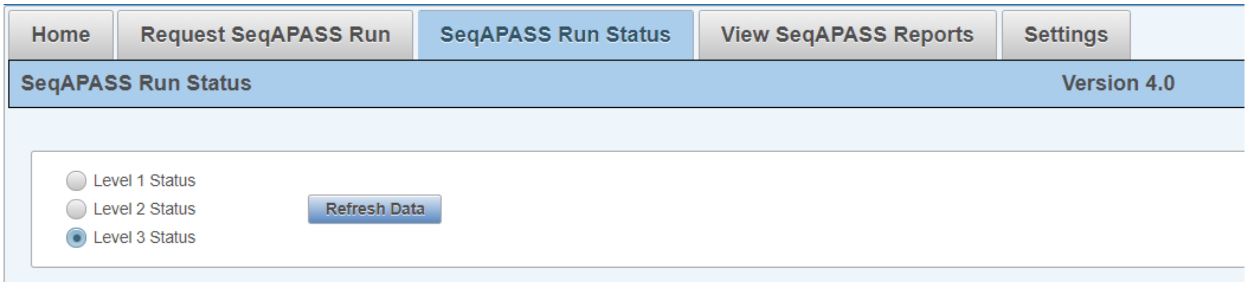
When the user has selected all sequences they want to align, then click the "Request Residue Run" button. Upon successful submission of a Level 3 query the user will see the following pop-up message. If submission is unsuccessful, a message will appear describing the reason for the unsuccessful submission.



Level 3 Run Requested
Status queued

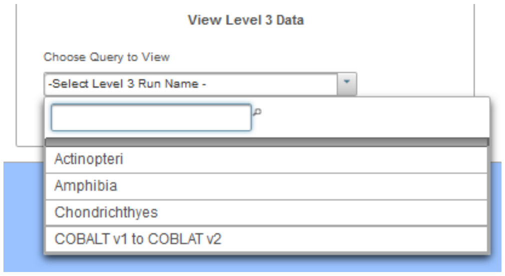
Logged in as: Blatz,Donovan

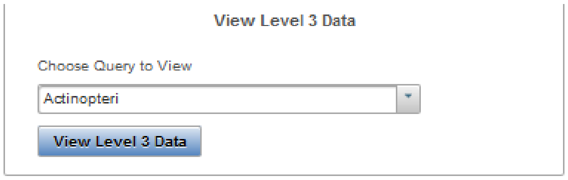
To update the “Choose Query to View” drop-down menu with the completed Level 3 alignments, the user can click on the “Refresh Level 2 and 3 runs” button.



○ Level 1 Status
○ Level 2 Status
● Level 3 Status

Additionally, the user can check the status of the Level 3 run by clicking the “SeqAPASS Run Status” tab and the radio button for “Level 3 Status.” Typically, Level 3 alignments complete in a few seconds. When the Level 3 query completes and the Level 1 page has been updated, the user defined Level 3 Run Name will be available in the “Choose Query to View” drop-down menu. After selecting the desired Run Name from the drop-down, click “View Level 3 Data” button to view the aligned sequences and set up the individual amino acid residue alignments with the selected sequences/species.

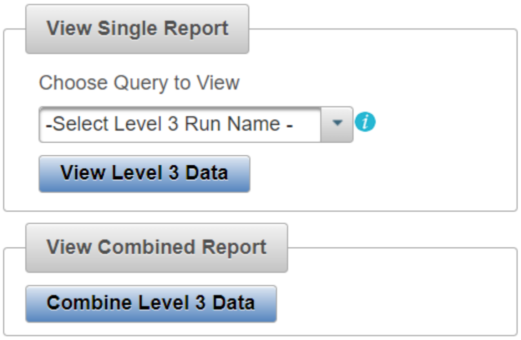




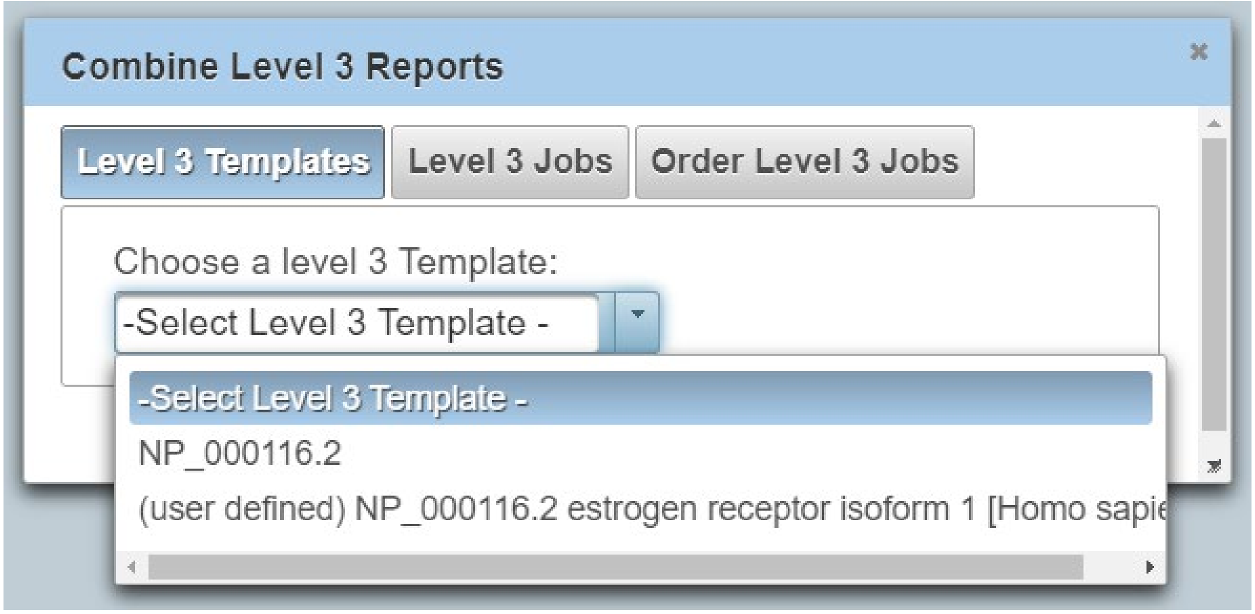
Choose Query to View: Actinopteri

Upon a successful Level 3 query submission a pop-up message will be displayed as follows in the upper right-hand side of the screen:

Level 3 Run Requested
Status queued



Once the Level 3 run has completed, the user can select the "Select Level 3 Run Name" drop down in the View Single Report box to view an individual user defined level 3 run. If the user has completed multiple Level 3 alignments, between a template sequence and more than one taxonomic groups, the user can combine Level 3 reports by selecting the "Combine Level 3 Data" button. A pop-up will appear for the Combine Level 3 Reports. There are a series of three steps to combine Level 3 reports. First the user will Choose a Level 3 Template from the dropdown that contains a list of all templates used to generate alignments in Level 3 by the user. The template sequence must be in-common to the Level 3 runs that will be combined.

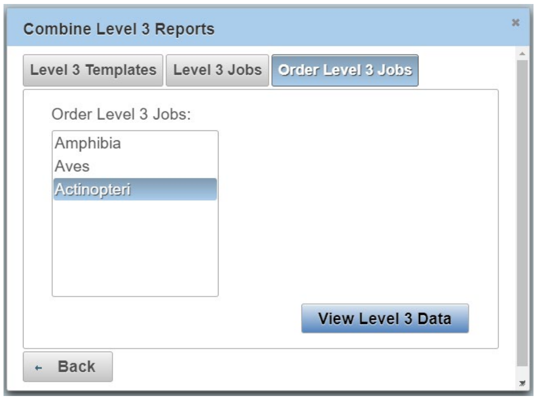


After selecting the template, the user will click the "Next" button. At this point the user will select all Level 3 Jobs that are to be combined by selecting the check box in the "Level 3 Jobs" dropdown next to the user defined names. After all jobs that are to be combined are selected the user will click the "Next" button. Note that as the user moves through each step of the Combine level 3 Reports feature, the step the user is currently on is indicated by highlighting the button in blue coloring (example "Level 3 Jobs" button is highlighted when working on selecting Jobs to combine).

☐

☑ Amphibia
☑ Aves
☑ Actinopteri

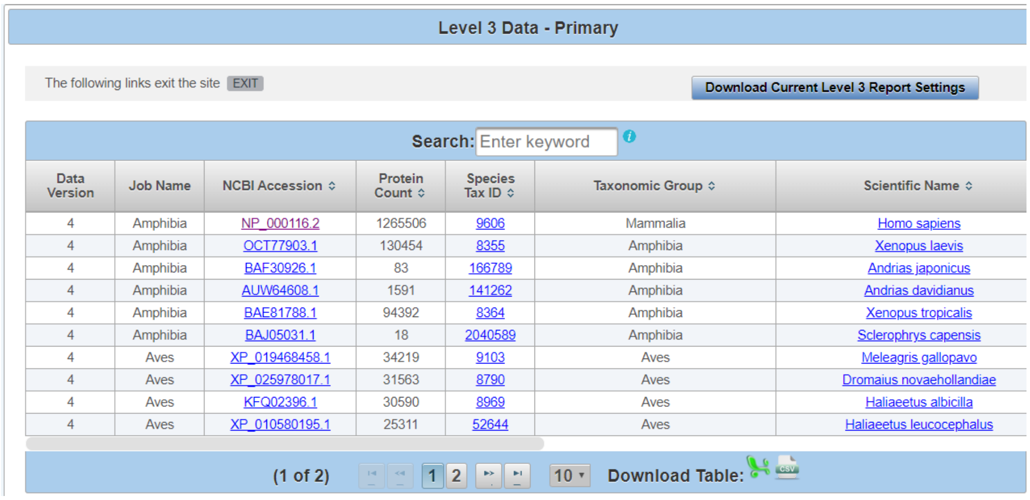
The next step in the Combine Level 3 Reports feature is to put the jobs in order as to how they should be displayed in the output. Typically, sequences from an individual taxonomic group are aligned to a template sequence and named accordingly (e.g., Actinopteri, Amphibia, Aves, etc.). It may be useful to order the combined report similarly to how the taxonomic groups are displayed on the x-axis of the Level 1 or Level 2 data visualization. Therefore, the user can select the user defined name from the *Order Level 3 Jobs:* text box and drag and drop the name to the desired order from top to bottom. To move on to select individual amino acids for sequence comparisons the user will select the “View Level 3 Data” button.



Order Level 3 Jobs:

Amphibia
Aves
Actinopteri

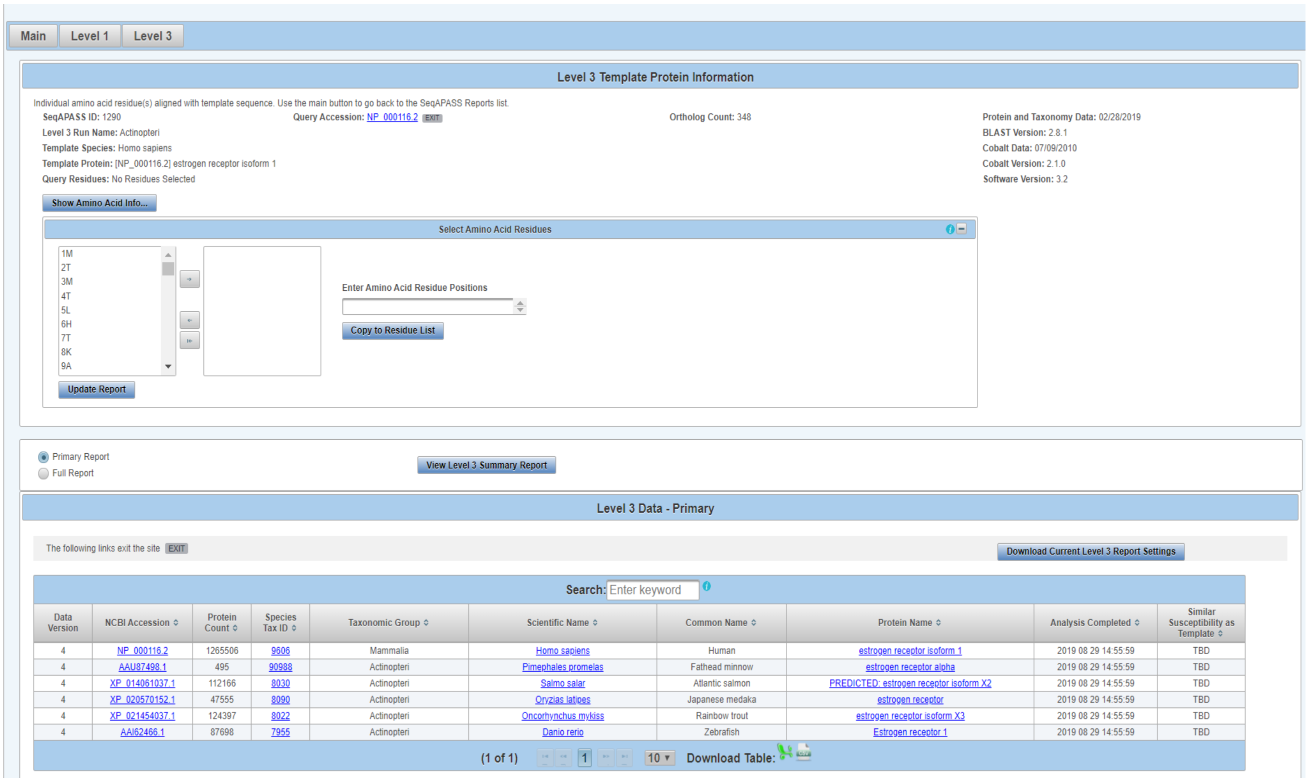
The order selected will translate to the top to bottom order displayed in the data table, with the template sequence only displayed once in the first row and all selected jobs below.



Data Version	Job Name	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Scientific Name
4	Amphibia	NP_000116.2	1265506	9606	Mammalia	Homo sapiens
4	Amphibia	OCT77903.1	130454	8355	Amphibia	Xenopus laevis
4	Amphibia	BAF30926.1	83	166789	Amphibia	Andrias japonicus
4	Amphibia	AUW64608.1	1591	141262	Amphibia	Andrias davidianus
4	Amphibia	BAE81788.1	94392	8364	Amphibia	Xenopus tropicalis
4	Amphibia	BAJ05031.1	18	2040589	Amphibia	Sclerophrys capensis
4	Aves	XP_019468458.1	34219	9103	Aves	Meleagris gallopavo
4	Aves	XP_025978017.1	31563	8790	Aves	Dromaius novaehollandiae
4	Aves	KFQ02396.1	30590	8969	Aves	Haliaeetus albicilla
4	Aves	XP_010580195.1	25311	52644	Aves	Haliaeetus leucocephalus

View Level 3 Individual Amino Acid Query and Data Page

Clicking the "View Level 3 Data" button, the Level 3 data page opens. The "Level 3 Template Protein Information" box contains the SeqAPASS Run ID, Query Accession (with link out to NCBI), Ortholog Count (# of hits identified as ortholog candidates to the query species protein sequence), NCBI Data (displays the date that NCBI databases and executables were downloaded and incorporated into SeqAPASS), Level 3 Run Name (defined by user), Template Species (Entered by user in Level 3 query), Template Protein, and Query Residues (this field is populated with residues upon selection and successful table update).



Individual amino acid residue(s) aligned with template sequence. Use the main button to go back to the SeqAPASS Reports list.

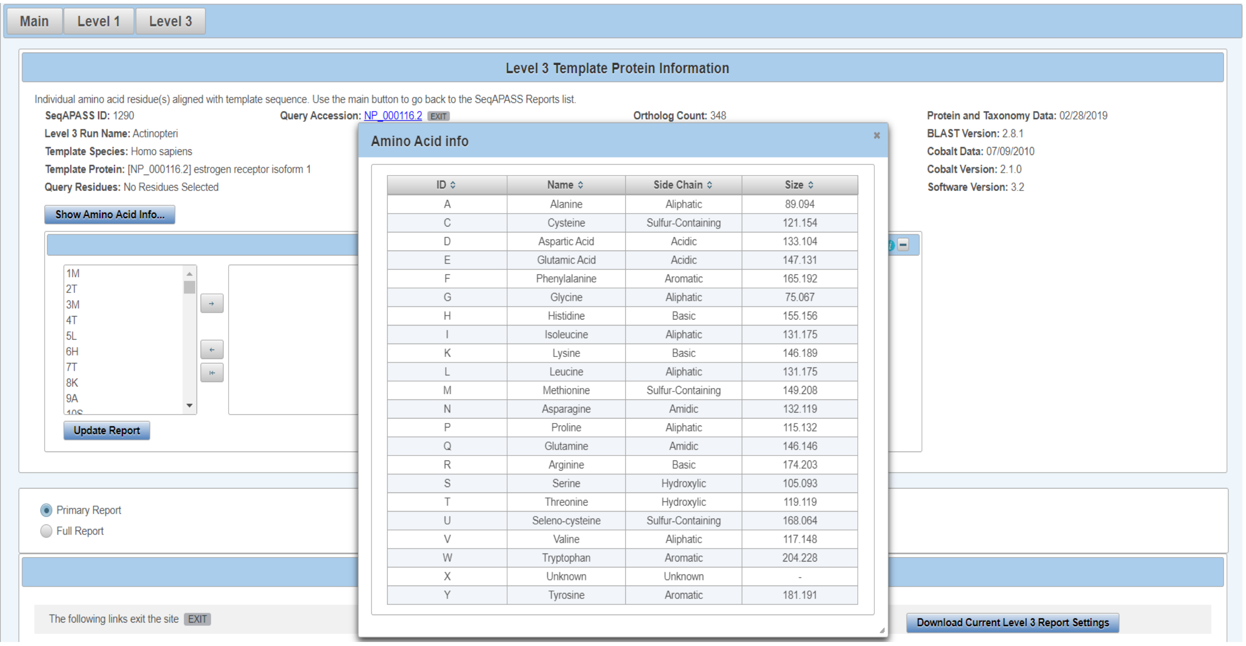
SeqAPASS ID: 1290
Query Accession: NP_000116.2 EXIT
Ortholog Count: 348
Protein and Taxonomy Data: 02/28/2019
Level 3 Run Name: Actinopteri
BLAST Version: 2.8.1
Template Species: Homo sapiens
Cobalt Data: 07/09/2010
Template Protein: [NP_000116.2] estrogen receptor isoform 1
Cobalt Version: 2.1.0
Query Residues: No Residues Selected
Software Version: 3.2

● Primary Report
○ Full Report

Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Scientific Name	Common Name	Protein Name	Analysis Completed	Similar Susceptibility as Template
4	NP_000116.2	1265506	9606	Mammalia	Homo sapiens	Human	estrogen receptor isoform 1	2019 08 29 14:55:59	TBD
4	AAU87498.1	495	90988	Actinopteri	Pimephales promelas	Fathead minnow	estrogen receptor alpha	2019 08 29 14:55:59	TBD
4	XP_014061037.1	112166	8030	Actinopteri	Salmo salar	Atlantic salmon	PREDICTED: estrogen receptor isoform X2	2019 08 29 14:55:59	TBD
4	XP_020570152.1	47555	8090	Actinopteri	Oryzias latipes	Japanese medaka	estrogen receptor	2019 08 29 14:55:59	TBD
4	XP_021454037.1	124397	8022	Actinopteri	Oncorhynchus mykiss	Rainbow trout	estrogen receptor isoform X3	2019 08 29 14:55:59	TBD
4	AAI62466.1	87698	7955	Actinopteri	Danio rerio	Zebrafish	Estrogen receptor 1	2019 08 29 14:55:59	TBD

The user can view the "Level 3" data page, which includes the Data Version, NCBI Accession, Protein Count, Taxonomic information, Protein Name, and date/time the Level 3 run completed. The data table remains in order of percent similarity, with those sequences having the highest percent similarity to the template sequence, on the top, to those with the lowest percent similarity on the bottom. (See **Search, View, and Download Data Tables** section of user guide for more information).

For additional information on Amino Acid Residues, including definition of the acronym, the amino acid residue name, the classification for the amino acid side chain and the size of the amino acid residue based on molecular weight, the user can click the “Show Amino Acid Info…” button. A pop-up table, “Amino Acid info,” will be displayed providing this information.



Individual amino acid residue(s) aligned with template sequence. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 1290
Query Accession: NP_000116.2
Ortholog Count: 348
Protein and Taxonomy Data: 02/28/2019
Level 3 Run Name: Actinopteri
BLAST Version: 2.8.1
Template Species: Homo sapiens
Cobalt Data: 07/09/2010
Template Protein: [NP_000116.2] estrogen receptor isoform 1
Cobalt Version: 2.1.0
Query Residues: No Residues Selected
Software Version: 3.2

● Primary Report
○ Full Report

The following links exit the site

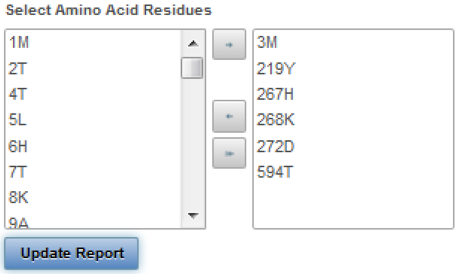
Amino Acid info

ID	Name	Side Chain	Size
A	Alanine	Aliphatic	89.094
C	Cysteine	Sulfur-Containing	121.154
D	Aspartic Acid	Acidic	133.104
E	Glutamic Acid	Acidic	147.131
F	Phenylalanine	Aromatic	165.192
G	Glycine	Aliphatic	75.067
H	Histidine	Basic	155.156
I	Isoleucine	Aliphatic	131.175
K	Lysine	Basic	146.189
L	Leucine	Aliphatic	131.175
M	Methionine	Sulfur-Containing	149.208
N	Asparagine	Amidic	132.119
P	Proline	Aliphatic	115.132
Q	Glutamine	Amidic	146.146
R	Arginine	Basic	174.203
S	Serine	Hydroxylic	105.093
T	Threonine	Hydroxylic	119.119
U	Seleno-cysteine	Sulfur-Containing	168.064
V	Valine	Aliphatic	117.148
W	Tryptophan	Aromatic	204.228
X	Unknown	Unknown	-
Y	Tyrosine	Aromatic	181.191

To obtain individual amino acid residue alignment data in the Level 3 data table, the user must use the shuttle in the “Level 3 Template Protein Information box to select positions and amino acid residues from the chosen template sequence to align with the sequences/species that were selected by taxonomic group. Single letter abbreviations are used for the amino acid sequences.

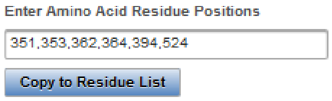
G: Glycine A: Alanine S: Serine T: Threonine C: Cysteine V: Valine
L: Leucine I: Isoleucine M: Methionine P: Proline F Phenylalanine U: Seleno-cysteine
Y: Tyrosine W: Tryptophan D: Aspartic Acid E: Glutamic Acid
N: Asparagine Q: Glutamine H: Histidine K: Lysine R: Arginine

The user can select one residue at a time by clicking and highlighting the residue of interest and then clicking the top right arrow shuttle button to move the residue to the right-hand box for inclusion in the alignment. Each time a residue is added to the right-hand box, the left-hand box resets itself to the 1st residue. Or the user can select multiple residues at the same time by holding the Ctrl button, clicking on residues, and then clicking the top right arrow shuttle button to move the residues to the right-hand box. The user can choose to remove selected residues by using the left arrow button to clear one at a time or the double left arrow button to remove all selected residues at once. When residues of interest (likely defined from the literature as described above) have been selected, click the "Update Report" button, which then updates the Level 3 Data table with the individual residue alignment data.



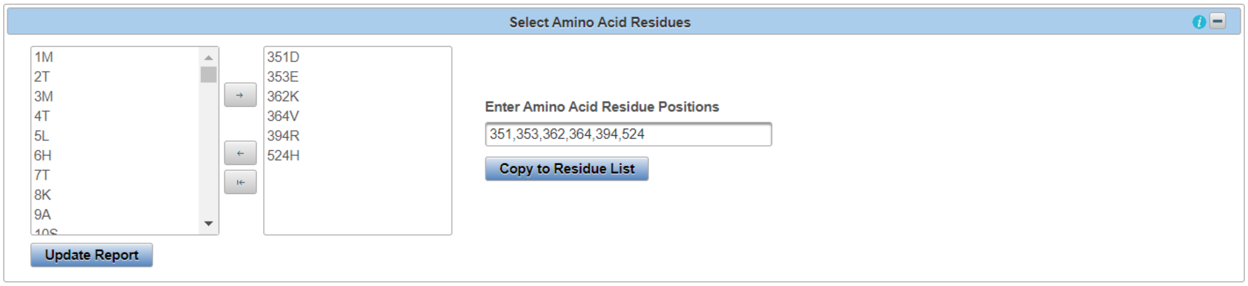
3M
219Y
267H
268K
272D
594T

Alternatively, the user can enter the amino acid positions in the "Enter Amino Acid Residue Positions" text box (e.g., 351,353,362) and click the "Copy to Residue List" button.



Enter Amino Acid Residue Positions: 351,353,362,364,394,524

Upon clicking "Copy to Residue List" the "Select Amino Acid Residues" shuttle box is populated with the position and residues typed. The user can then click the update Report button to produce Level 3 results in the table below.



351D
353E
362K
364V
394R
524H

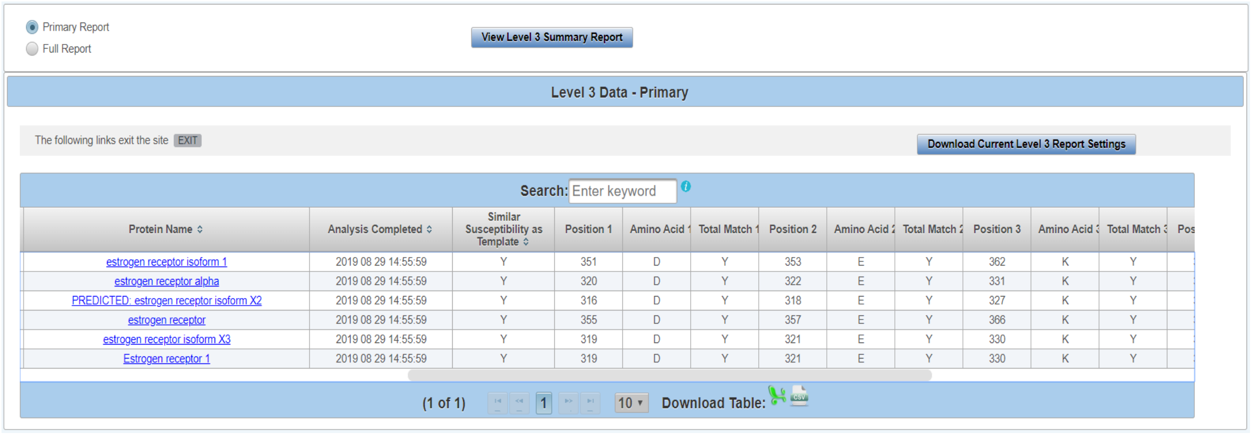
Enter Amino Acid Residue Positions: 351,353,362,364,394,524

The individual amino acid residue alignment data will then be updated on the right most columns of the Level 3 Data table. The user ***can submit a maximum of 50 individual amino acid residues*** from the template sequence to compare to the other selected sequences. The individual amino acid residues will be listed in numerical order starting with the 1st position in the template sequence to the last position in the template sequence.

Level 3 Data – Primary Report

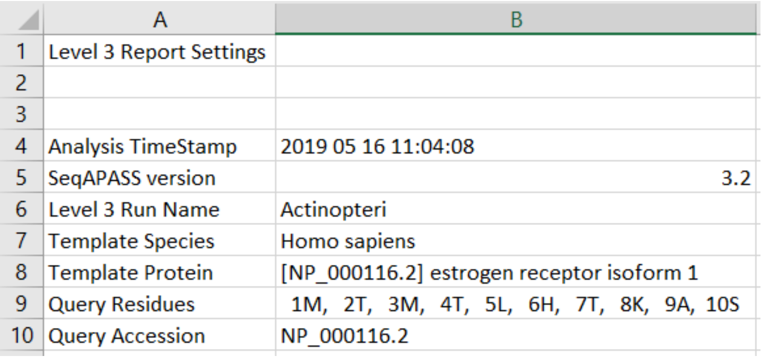
The default report is the "Primary Report" and can be recognized as such because the radio button for "Primary Report" above the "Level 3 Data" table is selected.

The "Primary Report" columns for the alignment will be titled "Similar Susceptibility as Template" ("Y" or "N" for yes or no, respectively), followed by Position 1, Amino Acid 1, Total Match 1, Position 2 Amino Acid 2, Total Match 2, Position 3, Amino Acid 3, Total Match 3…. The template sequence will always be in the top row of the Level 3 Data table followed by the previously selected sequences. Further, the residues selected in the shuttle will also be displayed in the top row corresponding to the template sequence. Each Position and Amino Acid in the following rows are those corresponding to the Protein Accession identified in that row and aligning with the template sequence. The Total Match X describes whether the amino acid residue matches the template based on side-chain classification and molecular weight, "Y," for yes, or "N," for not a match to the template. The user can evaluate this data to understand how well conserved an amino acid residue is across species or in a species of interest to add an additional line of evidence to support (or question) susceptibility predictions. The user can also download the current report settings by selecting the "Download Current Level 3 Report Settings." This csv allows the user to track which settings were used or changed by the user when downloading a data table.



Protein Name	Analysis Completed	Similar Susceptibility as Template	Position 1	Amino Acid 1	Total Match 1	Position 2	Amino Acid 2	Total Match 2	Position 3	Amino Acid 3	Total Match 3	Pos
estrogen receptor isoform 1	2019 08 29 14:55:59	Y	351	D	Y	353	E	Y	362	K	Y	
estrogen receptor alpha	2019 08 29 14:55:59	Y	320	D	Y	322	E	Y	331	K	Y	
PREDICTED: estrogen receptor isoform X2	2019 08 29 14:55:59	Y	316	D	Y	318	E	Y	327	K	Y	
estrogen receptor	2019 08 29 14:55:59	Y	355	D	Y	357	E	Y	366	K	Y	
estrogen receptor isoform X3	2019 08 29 14:55:59	Y	319	D	Y	321	E	Y	330	K	Y	
Estrogen receptor 1	2019 08 29 14:55:59	Y	319	D	Y	321	E	Y	330	K	Y	

When downloading the current level 2 report settings, the following information will be present in the csv. If the user decides to change the default settings, the csv can be utilized for quick information if the SeqAPASS page is no longer accessible.

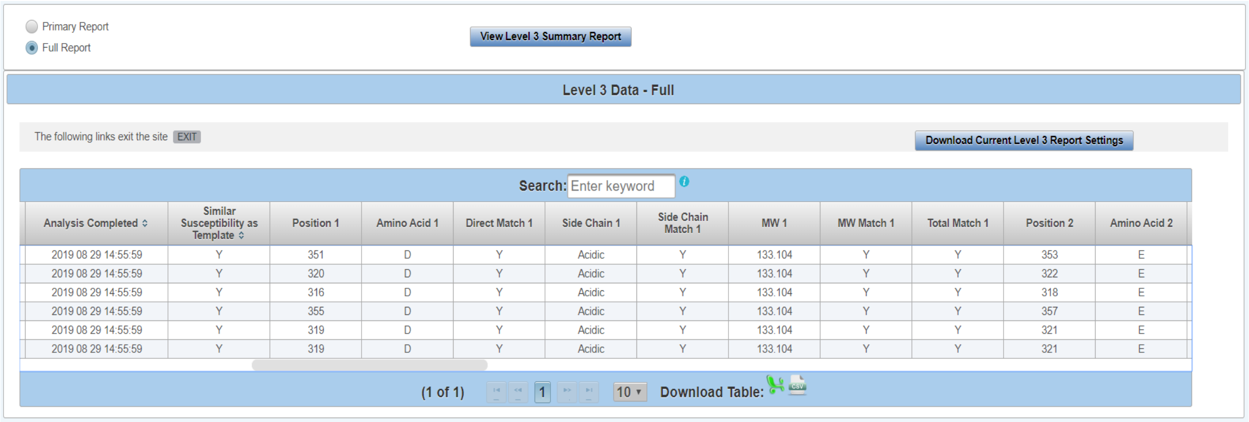


	A	B
1	Level 3 Report Settings	
2		
3		
4	Analysis TimeStamp	2019 05 16 11:04:08
5	SeqAPASS version	3.2
6	Level 3 Run Name	Actinopteri
7	Template Species	Homo sapiens
8	Template Protein	[NP_000116.2] estrogen receptor isoform 1
9	Query Residues	1M, 2T, 3M, 4T, 5L, 6H, 7T, 8K, 9A, 10S
10	Query Accession	NP_000116.2

Level 3 Data – Full Report

The user may choose to view the Full Report for Level 3 data by selecting the radio button above the "Level 3 Data" table for "Full Report." The table below will automatically update to display all of the alignment details.

The "Full Report" columns for the alignment will be titled "Similar Susceptibility as Template" ("Y" or "N" for yes or no), followed by Position 1, Amino Acid 1, Direct Match 1, Side Chain 1, MW1, MW Match 1Total Match 1, Total Match 1, Position 2, Amino Acid 2, Direct Match 2, Side Chain 2, MW2, MW Match Total Match 2, Total Match 2……… The template sequence will always be in the top row of the Level 3 Data table followed by the previously selected sequences. Further, the residues selected in the shuttle will also be displayed in the top row corresponding to the template sequence. Each Position and Amino Acid in the following rows are those corresponding to the Protein Accession identified in that row align with the template sequence. The Total Match X describes whether the amino acid residue matches the template based on side-chain classification and molecular weight, "Y," for yes, or "N," for not a match to the template. The user can evaluate this data to understand how well conserved an amino acid residue is across species or in a species of interest to add an additional line of evidence to support (or question) susceptibility predictions.

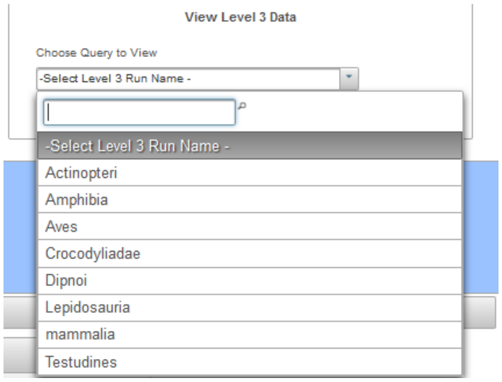


Analysis Completed	Similar Susceptibility as Template	Position 1	Amino Acid 1	Direct Match 1	Side Chain 1	Side Chain Match 1	MW 1	MW Match 1	Total Match 1	Position 2	Amino Acid 2
2019 08 29 14:55:59	Y	351	D	Y	Acidic	Y	133.104	Y	Y	353	E
2019 08 29 14:55:59	Y	320	D	Y	Acidic	Y	133.104	Y	Y	322	E
2019 08 29 14:55:59	Y	316	D	Y	Acidic	Y	133.104	Y	Y	318	E
2019 08 29 14:55:59	Y	355	D	Y	Acidic	Y	133.104	Y	Y	357	E
2019 08 29 14:55:59	Y	319	D	Y	Acidic	Y	133.104	Y	Y	321	E
2019 08 29 14:55:59	Y	319	D	Y	Acidic	Y	133.104	Y	Y	321	E

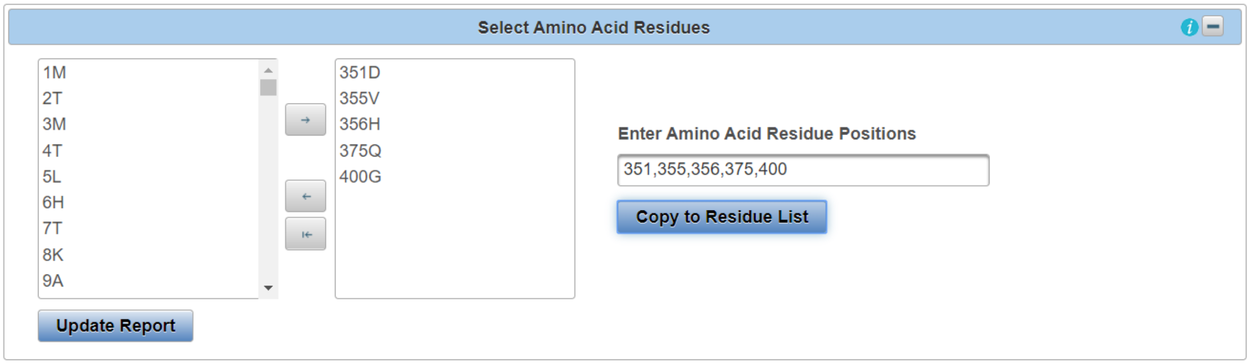
The "Direct Match X" column describes whether the hit amino acid is an exact match to the template amino acid, providing a "Y" or "N" for yes or no, respectively. The "Side Chain X" column indicates the side chain classification for the amino acid residue (click on "Show Amino Acid Info…for more information on classifications). The "Side Chain Match X" column indicates whether the hit side chain has the same classification as the template amino acid, providing a "Y" or "N" for yes or no, respectively. The "MW X" column indicates the molecular weight (g/mol) of the amino acid residue and the "MW Match X" column indicates whether the hit molecular weight has a difference in molecular weight greater than or equal to 30 g/mol compared to the template amino acid, providing a "Y" or "N" for yes or no, respectively. For the "Total Match X" to be "Y," both "Side Chain Match X" and "MW Match X" should be either "Y" and Y" or one "Y" and one "N," respectively. Only if both "Side Chain Match X" and "MW Match X" are "N" and "N," then the "Total Match X" is "N" for no. Ultimately, the Total Match 1, 2, 3, 4…. are used to inform the "Similar Susceptibility as Template" column. If there is one or more "N" for Total Match comparing any amino acid residue to the template across a row for a given species, then the "Similar Susceptibility as Template" is "N" for no, indicating that the hit species is predicted NOT to have the same susceptibly prediction as the template sequence. However, if all "Total Match X" are "Y" for yes, then the "Similar Susceptibility as Template" is "Y" indicating that the hit species is predicted to have the same susceptibly prediction as the template sequence.

Multiple Level 3 Runs Requiring the Same Amino Acid Residue Comparisons

Typically, Level 3 individual amino acid residue alignments are submitted repetitively, comparing species from one taxonomic group at a time to the template amino acid residue(s).



Therefore, to increase efficiency in submitting the same alignments in Level 3 over and over again, the user can take advantage of the "Copy to Residue List" button. For the first alignment of amino acid residues, the user would select the amino acid residues to align and click the "Update Report" button.



Enter Amino Acid Residue Positions: 351,355,356,375,400

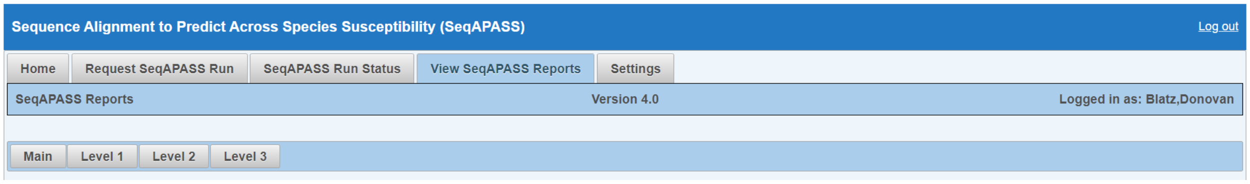
By clicking "Update Report" the residues that were selected will be copied into the "Enter Amino Acid Residue Positions" text box. When the user selects a new Level 3 Run Name from the same Level 1 query accession) to view by using the "View Level 3 Data" dropdown and clicking the "View Level 3 Data" button on the Level 1 Query Protein Information page, the "Enter Amino Acid Residue Positions" text box will be populated with the amino acid residues selected from the previous run.

Enter Amino Acid Residue Positions: 351,353,362,364,394,524

The user can keep, add, or delete, residue positions in this box and click "Copy to Residue List" button. The amino acid residues will then be moved to the "Select Amino Acid Residues Shuttle" and the user can then click "Update Report" to view the data in the table below.

Moving Between Level 1, Level 2, and Level 3 Data Pages

As a user chooses to view Level 1, Level 2, or Level 3 data in the "View SeqAPASS Reports" tab, new buttons become available for allowing the user to move between Levels of an analysis. Please see snapshot below.



The user can use the "Main" button to return to the list of completed Level 1 runs and select a different query accession to view. The "Level 1" button brings the user to the Level 1 data page, where the user can set up queries for Level 2 and Level 3, as well as select the button to view Level 2 and Level 3 data pages. Open Level 1, Level 2, and Level 3 pages remain open until the user selects a different run to view on the "Main" page. Moving between tabs, such as "Home," Request SeqAPASS Run," and "SeqAPASS Run Status", does not close the Level 1, Level 2, or Level 3 pages that have been opened.

Note: If the user logs out of the SeqAPASS tool, upon logging back in, the data will reset to default settings. Therefore, the View SeqAPASS Reports tab will not display the "Main," "Level 1," "Level 2," or "Level 3" buttons, until a query is chosen and Level 2 and Level 3 pages are opened.

Search, View, and Download Data Tables

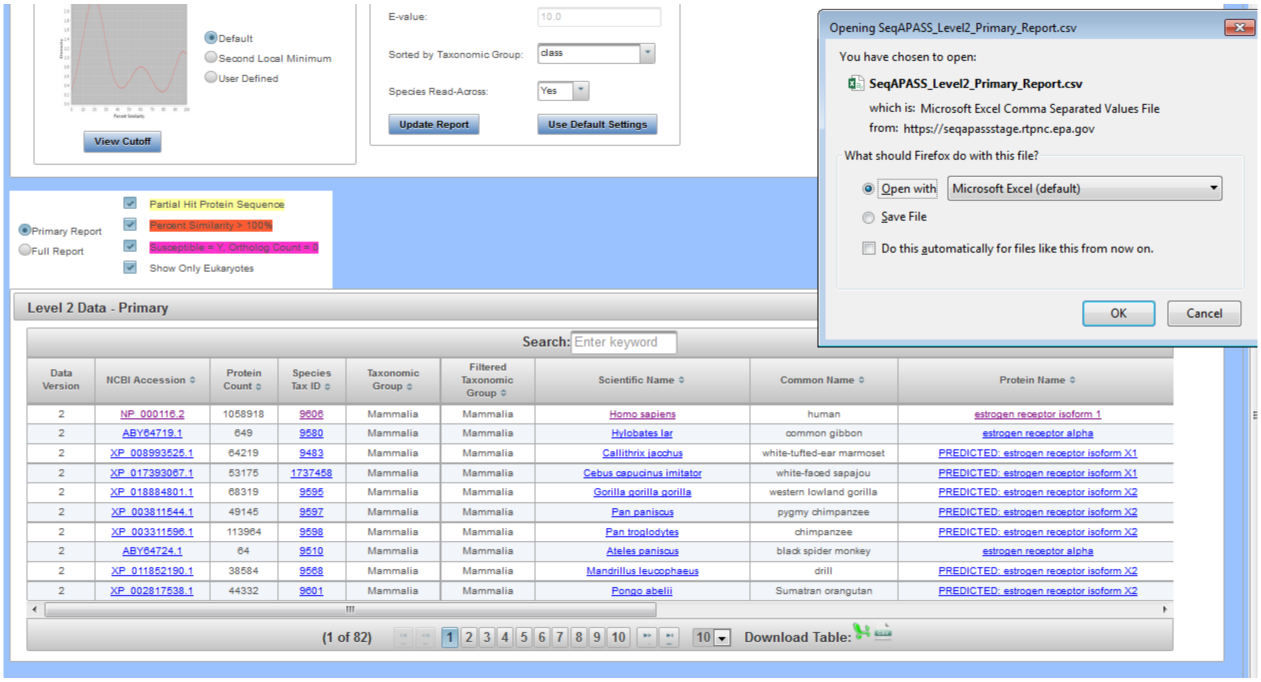
The user can use the "Search" box to enter text to search the table. Further, the user can use the arrow buttons and page numbers on the bottom of the screen to view all data and the drop-down to expand the table to 10, 20, or 50 rows. There are also left and right scroll bars at the bottom of the tables to allow the user to view all columns of the table.

Search using text box on top of tables:

Options for viewing data:

All data tables in the SeqAPASS tool can be downloaded as Excel or csv files. The icons for downloading the files are present on the bottom right-hand side of all tables. Click the icon to down load data.

Upon selecting a csv file, the user can choose to save or open the file. Each file is appropriately named by Level of the SeqAPASS evaluation and report type.



● Default
○ Second Local Minimum
○ User Defined

E-value: 10.0
Sorted by Taxonomic Group: class
Species Read-Across: Yes

● Primary Report
○ Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

Level 2 Data - Primary

Search: Enter keyword

Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name	Protein Name
2	NP_000116.2	1058918	9606	Mammalia	Mammalia	Homo sapiens	human	estrogen receptor isoform 1
2	ABY64719.1	649	9580	Mammalia	Mammalia	Hylobates lar	common gibbon	estrogen receptor alpha
2	XP_008993525.1	64219	9483	Mammalia	Mammalia	Callithrix jacchus	white-tufted-ear marmoset	PREDICTED: estrogen receptor isoform X1
2	XP_017393067.1	53175	1737458	Mammalia	Mammalia	Cebus capucinus imitator	white-faced sapajou	PREDICTED: estrogen receptor isoform X1
2	XP_018884801.1	68319	9595	Mammalia	Mammalia	Gorilla gorilla gorilla	western lowland gorilla	PREDICTED: estrogen receptor isoform X2
2	XP_003811544.1	49145	9597	Mammalia	Mammalia	Pan paniscus	pygmy chimpanzee	PREDICTED: estrogen receptor isoform X2
2	XP_003311596.1	113964	9598	Mammalia	Mammalia	Pan troglodytes	chimpanzee	PREDICTED: estrogen receptor isoform X2
2	ABY64724.1	64	9510	Mammalia	Mammalia	Ateles paniscus	black spider monkey	estrogen receptor alpha
2	XP_011852190.1	38584	9568	Mammalia	Mammalia	Mandrillus leucophaeus	drill	PREDICTED: estrogen receptor isoform X2
2	XP_002817538.1	44332	9601	Mammalia	Mammalia	Pongo abelii	Sumatran orangutan	PREDICTED: estrogen receptor isoform X2

(1 of 82) 1 2 3 4 5 6 7 8 9 10 10

Opening SeqAPASS_Level2_Primary_Report.csv

You have chosen to open:

SeqAPASS_Level2_Primary_Report.csv

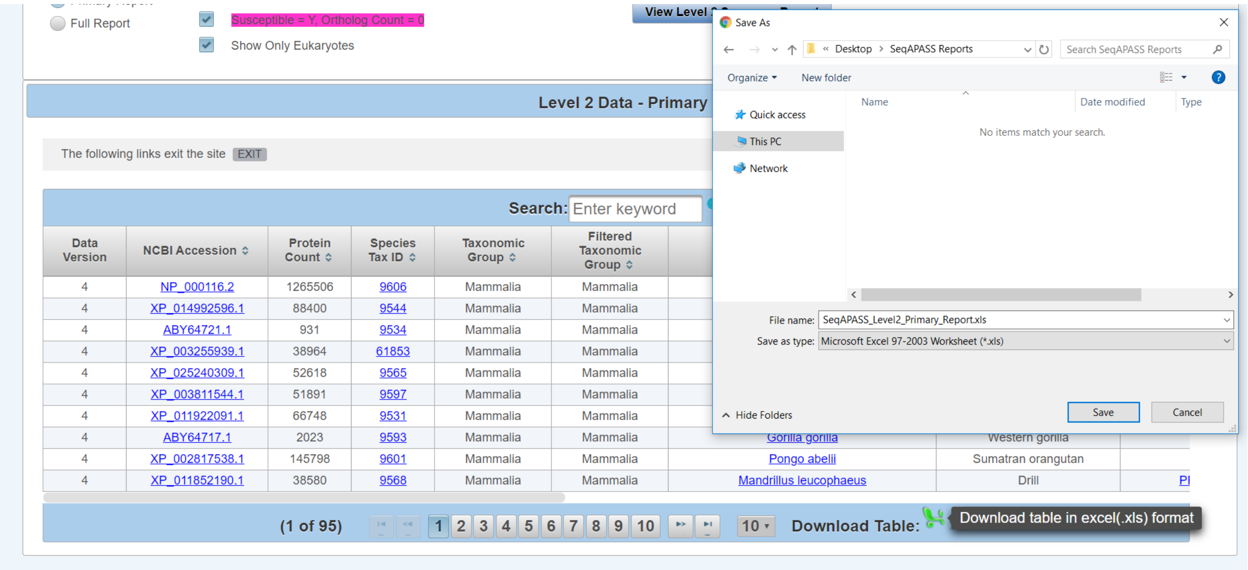
which is: Microsoft Excel Comma Separated Values File
from: https://seqapassstage.rtpnc.epa.gov

What should Firefox do with this file?

● Open with Microsoft Excel (default)
○ Save File

☐ Do this automatically for files like this from now on.

Upon selecting a .xls file, the user can save the report to their desired location. Each file is appropriately named by Level of the SeqAPASS evaluation and report type.



○ Full Report

☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

Level 2 Data - Primary

The following links exit the site EXIT

Search: Enter keyword

Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group			
4	NP_000116.2	1265506	9606	Mammalia	Mammalia			
4	XP_014992596.1	88400	9544	Mammalia	Mammalia			
4	ABY64721.1	931	9534	Mammalia	Mammalia			
4	XP_003255939.1	38964	61853	Mammalia	Mammalia			
4	XP_025240309.1	52618	9565	Mammalia	Mammalia			
4	XP_003811544.1	51891	9597	Mammalia	Mammalia			
4	XP_011922091.1	66748	9531	Mammalia	Mammalia			
4	ABY64717.1	2023	9593	Mammalia	Mammalia	Gorilla gorilla	Western gorilla	
4	XP_002817538.1	145798	9601	Mammalia	Mammalia	Pongo abelii	Sumatran orangutan	
4	XP_011852190.1	38580	9568	Mammalia	Mammalia	Mandrillus leucophaeus	Drill	Pl

(1 of 95) 1 2 3 4 5 6 7 8 9 10 10

Download table in excel(.xls) format

Save As

Desktop › SeqAPASS Reports

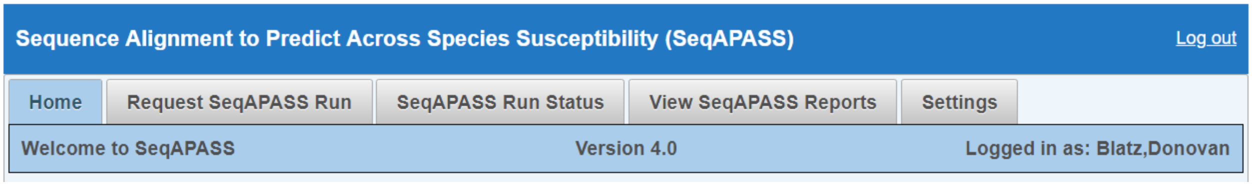
Search SeqAPASS Reports

No items match your search.

File name: SeqAPASS_Level2_Primary_Report.xls
Save as type: Microsoft Excel 97-2003 Worksheet (*.xls)

Log out

The user can log out from any page in SeqAPASS, by clicking the “Log out” link on the upper right-hand side of the page. If a user clicks Log out and then Logs back in, all settings will be set back to default. User can log out at any time by clicking the “Log out” link on the upper right-hand side. Any successfully submitted queries that were requested prior to logging out will continue running and when completed, will be available to the user in the “View SeqAPASS Reports” tab.



Pop-up Messages

The Spinning Wheel pop-up is used as an indicator to alert the user that an action is taking place, where the interface of the SeqAPASS tool is contacting the backend database. For example, upon clicking the “SeqAPASS Run Status” tab, “Refresh Data” button, “View Level 2 Data” button, or “View Level 3 Data” button the Spinning Wheel will pop-up and disappear from the screen. There are multiple other instances where the spinning wheel is used as an indicator to the user that an action is occurring.

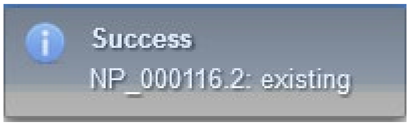
Querying database ... Please wait

Pop-up messages are meant to guide the user to submit the correct information for a query, inform the user of a successful or failed query submission, or otherwise inform the user of an error. All pop-up messages will appear for 10 seconds on the upper right-hand side of the screen, and then disappear. If the user would like to close the message before the 10 seconds is up, click on the message and an “x” will appear of the upper right-hand corner of the message box. Click the x to close the message.

In the “Request SeqAPASS Run” tab, Compare Primary Amino Acid Sequences “By Species” page, a successful Level 1 query submission will display a pop-up message indicating that the query has been submitted to the run queue or if “existing’ message appears indicating that the accession has been ran previously either by a user and is available to view.

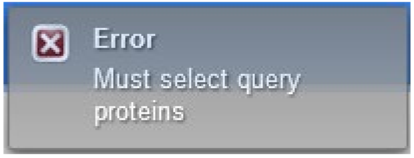
Success
Submitted NP_064393.2: submitted

OR



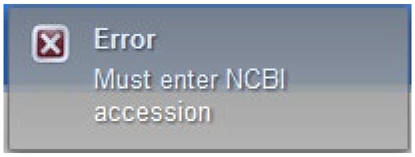
Success
NP_000116.2: existing

User did not select any query proteins from the "Request SeqAPASS Run" tab, Compare Primary Amino Acid Sequences "By Species" or "By Accession" page, and clicked "Request Run" button.



Error
Must select query proteins

OR



Error
Must enter NCBI accession

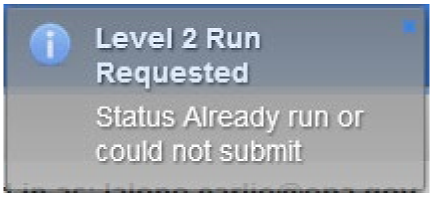
If the user enters non-sense text (or any text that is not an NCBI accession) into the "NCBI Protein Accession" text box for submitting a Level 1 query in the "Request SeqAPASS Run" tab, in the Compare Primary Amino Acid Sequences "By Accession" page, and clicked "Request Run" button, the message below will pop-up indicating that the Accession entered is not in the SeqAPASS database.

Success
fgafgaf: not in database

In the "View SeqAPASS Reports" tab, Level 1 page, if a user clicks "View Level 2 Data," a successful Level 2 query submission will display a pop-up message indicating that the query has entered the run queue.

Level 2 Run Requested
Status queued

In the "View SeqAPASS Reports" tab, Level 1 page, if a user selects a domain that has already been submitted (but not completed) and clicks "Request Domain Run" a message for successful Level 2 query submission will display a pop-up message indicating that the query has entered the run queue



Level 2 Run Requested
Status Already run or could not submit

In the “View SeqAPASS Reports” tab, Level 1 page, if a user clicks “View Level 2 Data” without selecting a domain to view from the drop-down, the message below will pop-up to indicate that the user must select a domain.

Error
Must select domain from drop-down

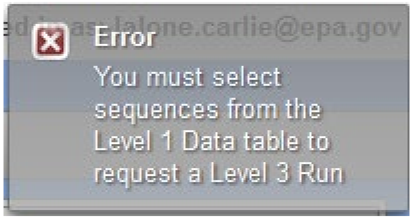
In the “View SeqAPASS Reports” tab, Level 1 page, a successful Level 3 query submission will display a pop-up message indicating that the query has entered the run queue.

Level 3 Run Requested
Status queued

In the “View SeqAPASS Reports” tab, Level 1 page, if a user fails to type a user defined Level 3 Run Name, the message below will pop-up to indicate that the user must do so.

Error
You must specify a Template Sequence and Level 3 Run Name

In the “View SeqAPASS Reports” tab, Level 1 page, if a user fails to select species from the Level 1 Data table to be compared with the template sequence, the message below will pop-up.



Error
You must select sequences from the Level 1 Data table to request a Level 3 Run

In the “View SeqAPASS Reports” tab, Level 1 page, if a user fails to select a Level 3 Run Name from the Choose Query to View drop-down and clicks the “View Level 3 Date” button, the message below will pop-up.

Error
Must select level 3 run from drop-down

In the "View SeqAPASS Reports" tab, "Level 3 Template Protein Information" data page, if a user fails to select amino acid residues using the "Select Amino Acid Residues" shuttle and clicks the "View Level 3 Date" button, the message below will pop-up.

No Residues Selected
User must select residues

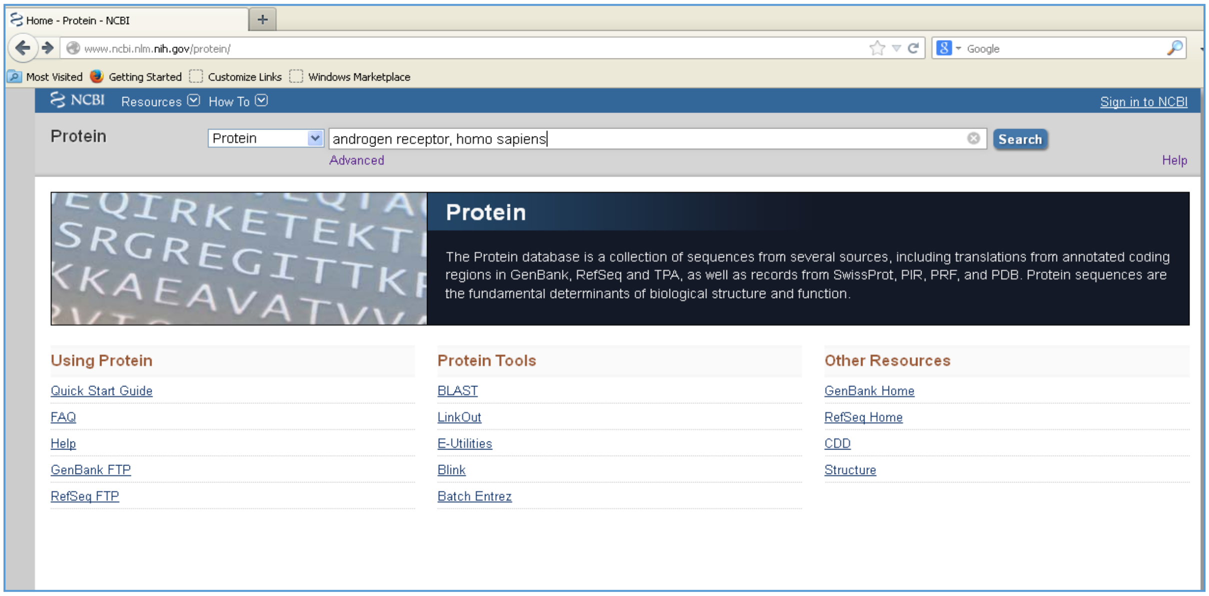
Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) Documentation

Query Species: The selection of the query species for a SeqAPASS analysis is dependent upon the question the user is addressing. For example, the query species can be the target species (i.e., human or companion animal in the case of drugs; or insect, plant, fungus, or pest in the case of pesticides) or, depending on the application of the susceptibility prediction, the query species may be a species known or hypothesized to be sensitive to a chemical acting on the protein molecular target of interest. There may be instances where a protein for the species of interest has not been sequenced, in this case it may serve the users purpose to identify another taxonomically related species from the same organism Class, Order, Family, or Genus as a surrogate query species. In certain cases, when there is interest in the susceptibility of a particular species (e.g., honey bee) and in the case that there are numerous potential target species (e.g., neonicotinoids are intended to cause mortality in a number of pest insects) the species of particular concern may serve as the query species.

Query Protein: SeqAPASS can be queried with any protein sequence available in the NCBI protein GenBank database, by protein name, or NCBI Accession. It is suggested that the user of SeqAPASS examines their query protein and species in the NCBI protein database prior to submitting a run to SeqAPASS (use NCBI link on query page). It is not uncommon for a protein of a specific species to be represented by more than one sequence. In such cases there are some guiding principles for identification of the best sequence available for the SeqAPASS run.

General guidelines: These guidelines describe best practices for identifying the most useful sequence for a species susceptibility prediction in SeqAPASS, however, in some cases, limited sequence information is available and therefore less desirable sequences may be used. It is up to the user of SeqAPASS to recognize the quality and limitations of the sequence chosen for the SeqAPASS query. The information about a particular protein can be found on the Protein page in the NCBI database (http://www.ncbi.nlm.nih.gov/protein/).

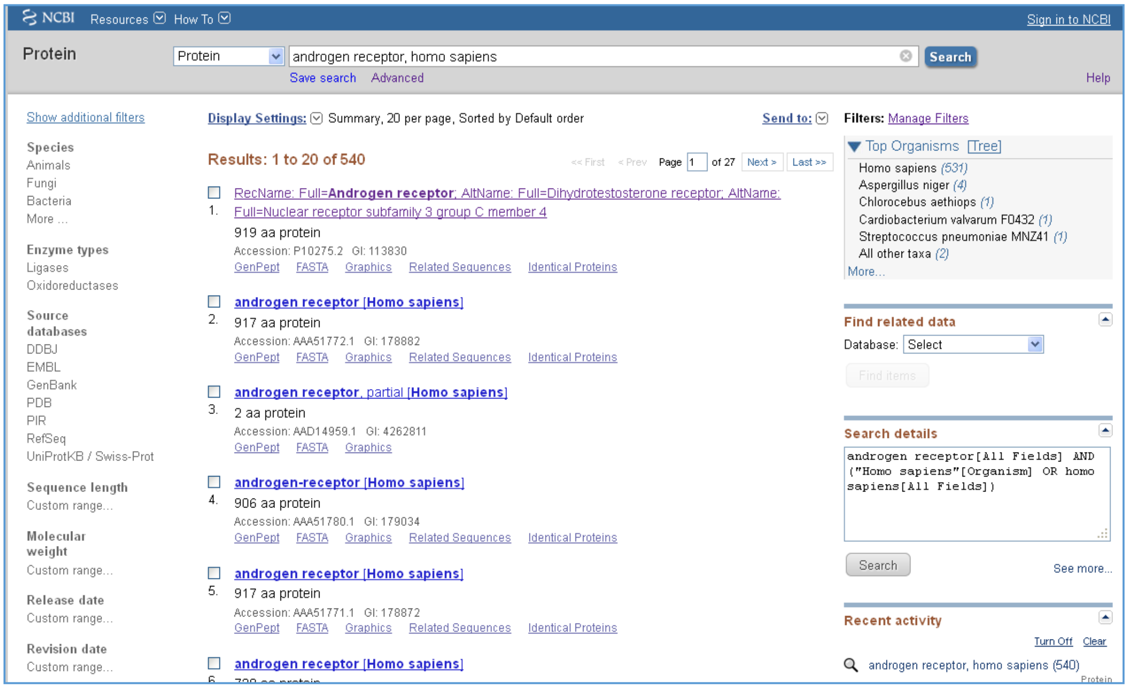
http://www.ncbi.nlm.nih.gov/protein/



Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

Search for a protein of interest using protein name and/or species of interest: For the example above, multiple hit proteins were identified.



Display Settings: Summary, 20 per page, Sorted by Default order

Results: 1 to 20 of 540

1. ☐ RecName: Full=Androgen receptor; AltName: Full=Dihydrotestosterone receptor; AltName: Full=Nuclear receptor subfamily 3 group C member 4
919 aa protein
Accession: P10275.2 GI: 113830
GenPept FASTA Graphics Related Sequences Identical Proteins

2. ☐ androgen receptor [Homo sapiens]
917 aa protein
Accession: AAA51772.1 GI: 178882
GenPept FASTA Graphics Related Sequences Identical Proteins

3. ☐ androgen receptor, partial [Homo sapiens]
2 aa protein
Accession: AAD14959.1 GI: 4262811
GenPept FASTA Graphics

4. ☐ androgen-receptor [Homo sapiens]
906 aa protein
Accession: AAA51780.1 GI: 179034
GenPept FASTA Graphics Related Sequences Identical Proteins

5. ☐ androgen receptor [Homo sapiens]
917 aa protein
Accession: AAA51771.1 GI: 178872
GenPept FASTA Graphics Related Sequences Identical Proteins

6. ☐ androgen receptor [Homo sapiens]

Filters: Manage Filters

Top Organisms [Tree]

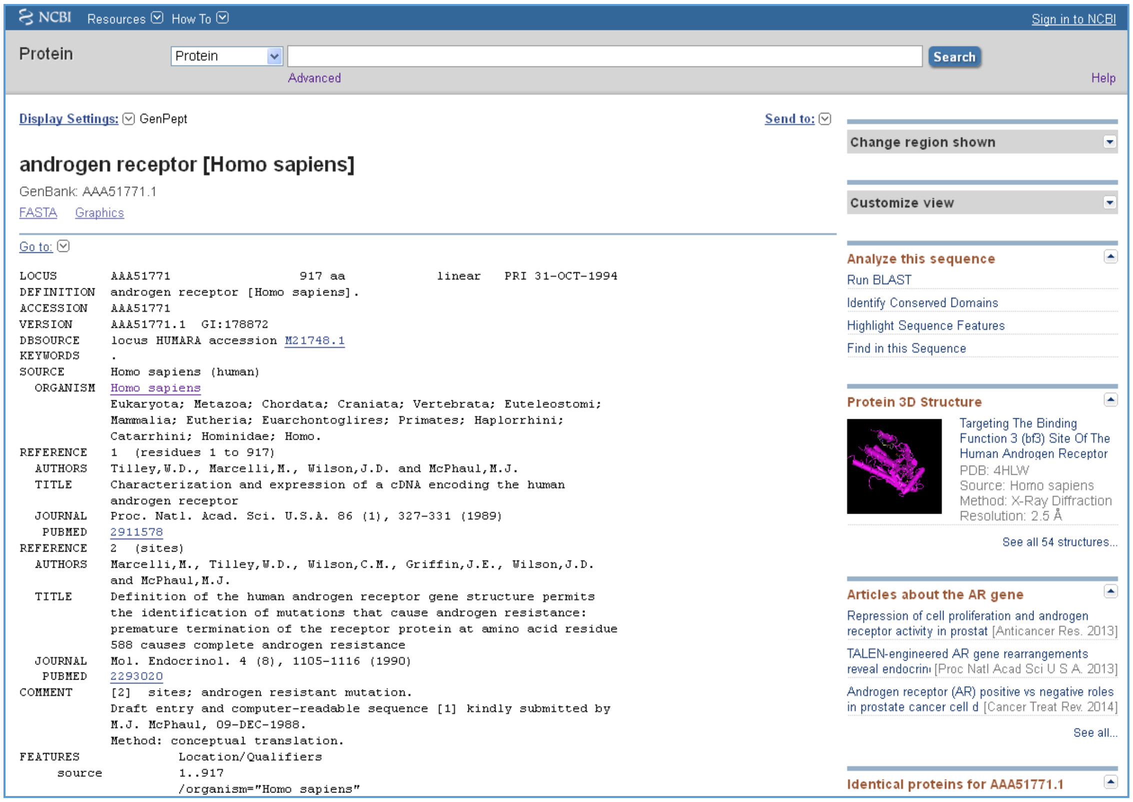
Homo sapiens	(531)
Aspergillus niger	(4)
Chlorocebus aethiops	(1)
Cardiobacterium valvarum F0432	(1)
Streptococcus pneumoniae MNZ41	(1)
All other taxa	(2)

Search details

```
androgen receptor[All Fields] AND
("Homo sapiens"[Organism] OR homo
sapiens[All Fields])
```

androgen receptor, homo sapiens (540)

Select one of the proteins by clicking on the link shown above to see detailed information about the protein



androgen receptor [Homo sapiens]

GenBank: AAA51771.1

```
LOCUS       AAA51771                 917 aa            linear   PRI 31-OCT-1994
DEFINITION  androgen receptor [Homo sapiens].
ACCESSION   AAA51771
VERSION     AAA51771.1  GI:178872
DBSOURCE    locus HUMARA accession M21748.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 917)
  AUTHORS   Tilley,W.D., Marcelli,M., Wilson,J.D. and McPhaul,M.J.
  TITLE     Characterization and expression of a cDNA encoding the human
            androgen receptor
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 86 (1), 327-331 (1989)
   PUBMED   2911578
REFERENCE   2  (sites)
  AUTHORS   Marcelli,M., Tilley,W.D., Wilson,C.M., Griffin,J.E., Wilson,J.D.
            and McPhaul,M.J.
  TITLE     Definition of the human androgen receptor gene structure permits
            the identification of mutations that cause androgen resistance:
            premature termination of the receptor protein at amino acid residue
            588 causes complete androgen resistance
  JOURNAL   Mol. Endocrinol. 4 (8), 1105-1116 (1990)
   PUBMED   2293020
COMMENT     [2]  sites; androgen resistant mutation.
            Draft entry and computer-readable sequence [1] kindly submitted by
            M.J. McPhaul, 09-DEC-1988.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..917
                     /organism="Homo sapiens"
```

<u>Guiding principles:</u> On the NCBI protein page, rows to examine include: “DEFINITION,” “REFERENCES,” COMMENTS,” and “FEATURES.” The information provided in these rows can aid a SeqAPASS user in the identification of an ideal query sequence for SeqAPASS.

It is desirable to:

a. Use accessions with the following prefix: NP_

b. Avoid use of protein sequences labeled “partial,” “PREDICTED,” “PROVISIONAL,” “INFERRED,” or “hypothetical”

c. Avoid using those labeled “TPA” (Third Party Annotation), however if TPA is all that is available “TPA: experimental” would be preferred over “TPA: inferential”

d. Look at the date associated with the protein in the “LOCUS” row of the detailed protein page. A more recent date can have the most up-to-date annotation of the protein. Under the “DBSOURCE” row of the detailed protein page other accessions associated with past protein sequences can be viewed. Many times, if the “xrefs” row is heavily populated and has the most recent annotation update date, it is likely to be the best sequence to use as a query sequence in SeqAPASS.

d. Short sequences should be avoided when possible as query sequences. Many times, if one selects the protein from the protein output derived from the NCBI protein database query, they will find that the short sequence is actually a partial sequence described in the “DEFINITION” row of the Protein page.

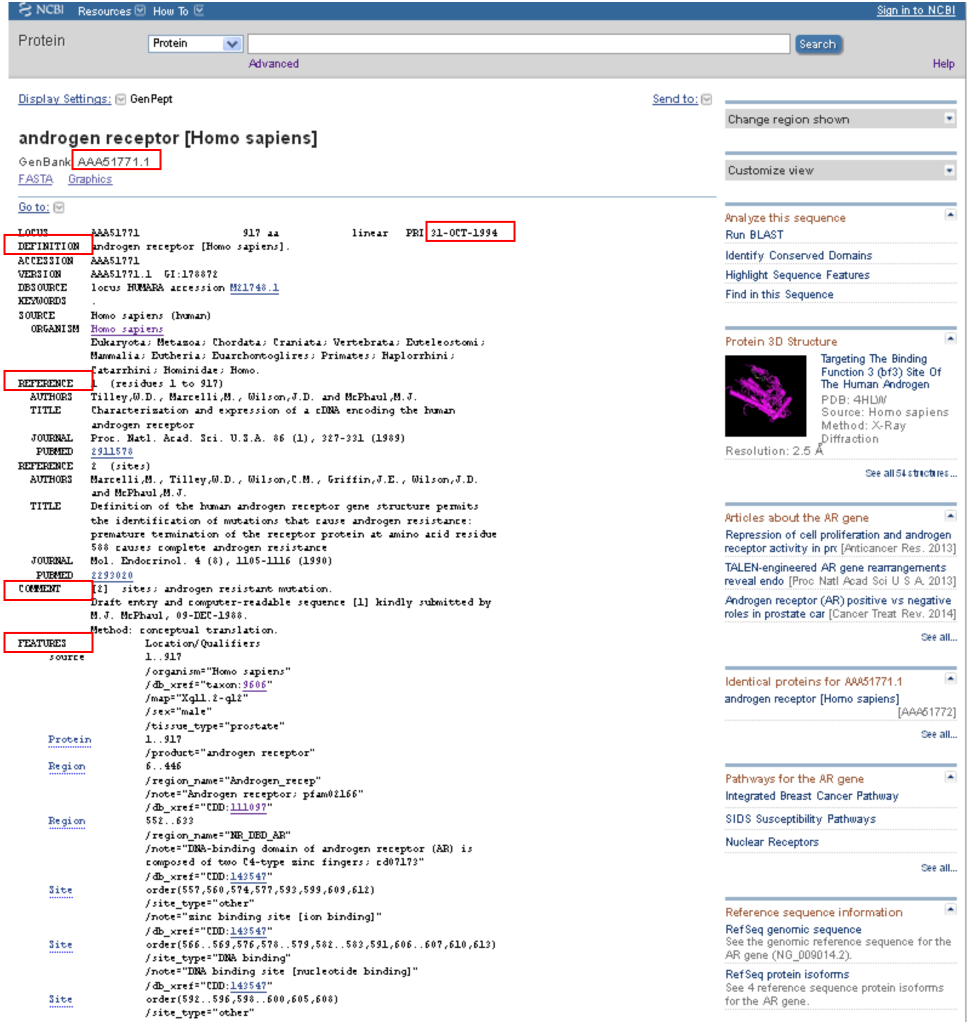
e. Unless there is reason for doing so (based on the question the user is trying to address), splice-variants labeled in “FEATURES” rows of the Protein page as “alternatively spliced” would be less desirable

f. It is important to check the references associated with the selected query protein. In some cases, certain sequences are associated with sensitivity to a given chemical. This can be particularly useful when predicting susceptibility to pesticides, where certain strains of insects are produced to be readily sensitive or insensitive to a chemical.

g. A secondary check of the sequence used in the SeqAPASS run would be to look at the output derived and see whether ortholog candidates were detected. Ideally a preferential sequence would have more ortholog candidates identified.

Important Note: To identify which query protein has the greatest number of Ortholog Candidates the user can choose to submit multiple proteins with the same species and protein. Upon the Level 1 runs completing for those similar proteins, the user can then select the "View SeqAPASS Reports" tab and look at the table for "Ortholog Count" the protein with the highest number is likely to be the most appropriate query species for a SeqAPASS evaluation.

Example: Androgen receptor, *Homo sapiens*



androgen receptor [Homo sapiens]

GenBank AAA51771.1

```
LOCUS       AAA51771                 917 aa            linear   PRI 31-OCT-1994
DEFINITION  androgen receptor [Homo sapiens].
ACCESSION   AAA51771
VERSION     AAA51771.1  GI:178872
DBSOURCE    locus HUMARA accession M21748.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 917)
  AUTHORS   Tilley,W.D., Marcelli,M., Wilson,J.D. and McPhaul,M.J.
  TITLE     Characterization and expression of a cDNA encoding the human
            androgen receptor
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 86 (1), 327-331 (1989)
   PUBMED   2911578
REFERENCE   2  (sites)
  AUTHORS   Marcelli,M., Tilley,W.D., Wilson,C.M., Griffin,J.E., Wilson,J.D.
            and McPhaul,M.J.
  TITLE     Definition of the human androgen receptor gene structure permits
            the identification of mutations that cause androgen resistance:
            premature termination of the receptor protein at amino acid residue
            588 causes complete androgen resistance
  JOURNAL   Mol. Endocrinol. 4 (8), 1105-1116 (1990)
   PUBMED   2293020
COMMENT     [2]  sites; androgen resistant mutation.
            Draft entry and computer-readable sequence [1] kindly submitted by
            M.J. McPhaul, 09-DEC-1988.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..917
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /map="Xq11.2-q12"
                     /sex="male"
                     /tissue_type="prostate"
     Protein         1..917
                     /product="androgen receptor"
     Region          6..446
                     /region_name="Androgen_recep"
                     /note="Androgen receptor; pfam02166"
                     /db_xref="CDD:111097"
     Region          552..633
                     /region_name="NR_DBD_AR"
                     /note="DNA-binding domain of androgen receptor (AR) is
                     composed of two C4-type zinc fingers; cd07173"
                     /db_xref="CDD:143547"
     Site            order(557,560,574,577,593,599,609,612)
                     /site_type="other"
                     /note="zinc binding site [ion binding]"
                     /db_xref="CDD:143547"
     Site            order(566..569,576,578..579,582..583,591,606..607,610,613)
                     /site_type="DNA binding"
                     /note="DNA binding site [nucleotide binding]"
                     /db_xref="CDD:143547"
     Site            order(592..596,598..600,605,608)
                     /site_type="other"
```

Targeting The Binding Function 3 (bf3) Site Of The Human Androgen
PDB: 4HLW
Source: Homo sapiens
Method: X-Ray Diffraction
Resolution: 2.5 Å
See all 54 structures...

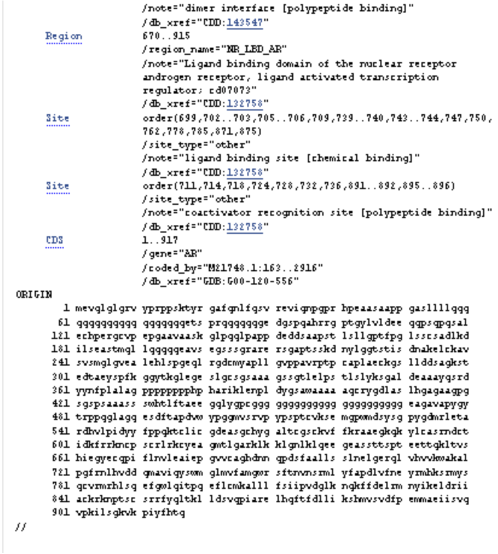
Articles about the AR gene
Repression of cell proliferation and androgen receptor activity in pro [Anticancer Res. 2013]
TALEN-engineered AR gene rearrangements reveal endo [Proc Natl Acad Sci U S A. 2013]
Androgen receptor (AR) positive vs negative roles in prostate car [Cancer Treat Rev. 2014]
See all...

Identical proteins for AAA51771.1
androgen receptor [Homo sapiens] [AAA51772]
See all...

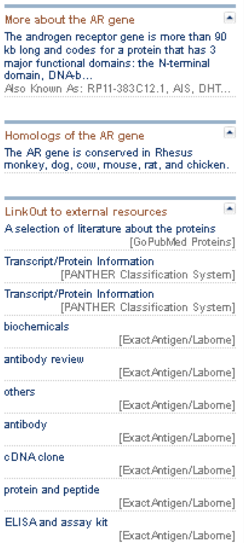
Pathways for the AR gene
Integrated Breast Cancer Pathway
SIDS Susceptibility Pathways
Nuclear Receptors
See all...

Reference sequence information
RefSeq genomic sequence
See the genomic reference sequence for the AR gene (NG_009014.2).
RefSeq protein isoforms
See 4 reference sequence protein isoforms for the AR gene.

Example cont:



```
                     /note="dimer interface [polypeptide binding]"
                     /db_xref="CDD:143547"
     Region          670..915
                     /region_name="NR_LBD_AR"
                     /note="Ligand binding domain of the nuclear receptor
                     androgen receptor, ligand activated transcription
                     regulator; cd07073"
                     /db_xref="CDD:132758"
     Site            order(699,702..703,705..706,709,739..740,743..744,747,750,
                     762,778,785,871,875)
                     /site_type="other"
                     /note="ligand binding site [chemical binding]"
                     /db_xref="CDD:132758"
     Site            order(711,714,718,724,728,732,736,891..892,895..896)
                     /site_type="other"
                     /note="coactivator recognition site [polypeptide binding]"
                     /db_xref="CDD:132758"
     CDS             1..917
                     /gene="AR"
                     /coded_by="M21748.1:163..2916"
                     /db_xref="GDB:G00-120-556"
ORIGIN
        1 mevqlglgrv yprppsktyr gafqnlfqsv reviqnpgpr hpeaasaapp gaslllllqqq
       61 qqqqqqqqqq qqqqqqqets prqqqqqqge dgspqahrrg ptgylvldee qqpsqpqsal
      121 echpergcvp epgaavaask glpqqlpapp deddsaapst lsllgptfpg lsscsadlkd
      181 ilseastmql lqqqqqeavs egsssgrare rsgaptsskd nylggtstis dnakelckav
      241 svsmglgvea lehlspgeql rgdcmyapll gvppavrptp caplaeckgs llddsagkst
      301 edtaeyspfk ggytkglege slgcsgsaaa gssgtlelps tlslyksgal deaaayqsrd
      361 yynfplalag pppppppphp hariklenpl dygsawaaaa aqcrygdlas lhgagaagpg
      421 sgspsaaass swhtlftaee gqlygpcggg gggggggggg gggggggggg eagavapygy
      481 trppqglagq esdftapdvw ypggmvsrvp ypsptcvkse mgpwmdsysg pygdmrleta
      541 rdhvlpidyy fppqktclic gdeasgchyg altcgsckvf fkraaegkqk ylcasrndct
      601 idkfrrkncp scrlrkcyea gmtlgarklk klgnlklqee geassttspt eettqkltvs
      661 hiegyecqpi flnvleaiep gvvcaghdnn qpdsfaalls slnelgerql vhvvkwakal
      721 pgfrnlhvdd qmaviqyswm glmvfamgwr sftnvnsrml yfapdlvfne yrmhksrmys
      781 qcvrmrhlsq efgwlqitpq eflcmkalll fsiipvdglk nqkffdelrm nyikeldrii
      841 ackrknptsc srrfyqltkl ldsvqpiare lhqftfdlli kshmvsvdfp emmaeiisvq
      901 vpkilsgkvk piyfhtq
//
```



More about the AR gene

The androgen receptor gene is more than 90 kb long and codes for a protein that has 3 major functional domains: the N-terminal domain, DNA-b...
Also Known As: RP11-383C12.1, AIS, DHT...

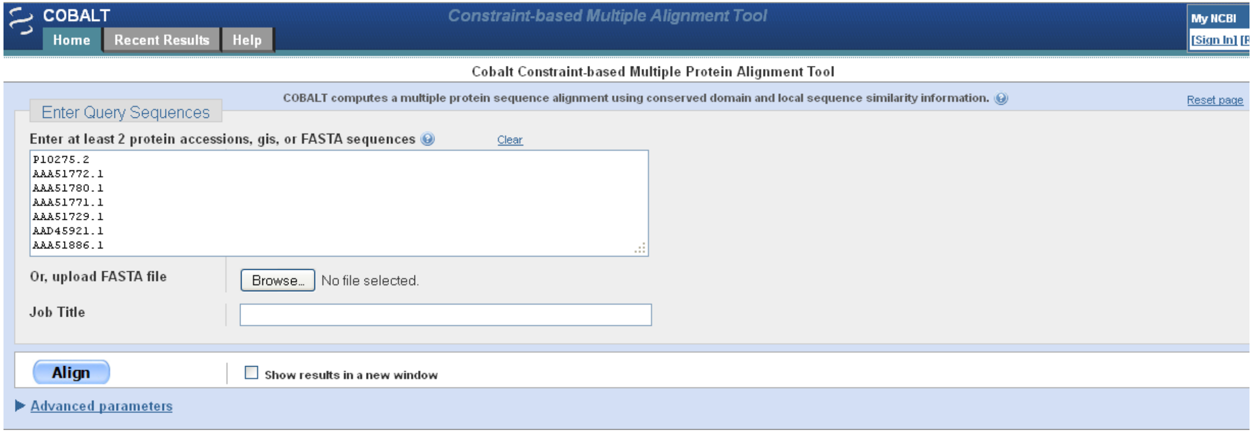
Homologs of the AR gene

The AR gene is conserved in Rhesus monkey, dog, cow, mouse, rat, and chicken.

LinkOut to external resources

A selection of literature about the proteins [GoPubMed Proteins]
Transcript/Protein Information [PANTHER Classification System]
Transcript/Protein Information [PANTHER Classification System]
biochemicals [ExactAntigen/Labome]
antibody review [ExactAntigen/Labome]
others [ExactAntigen/Labome]
antibody [ExactAntigen/Labome]
cDNA clone [ExactAntigen/Labome]
protein and peptide [ExactAntigen/Labome]
ELISA and assay kit [ExactAntigen/Labome]

h. If multiple proteins appear to be the best query protein for SeqAPASS, the sequences can be aligned using NCBI's COBALT. Enter (copy and paste from NCBI protein search list) accessions and align.



Enter at least 2 protein accessions, gis, or FASTA sequences

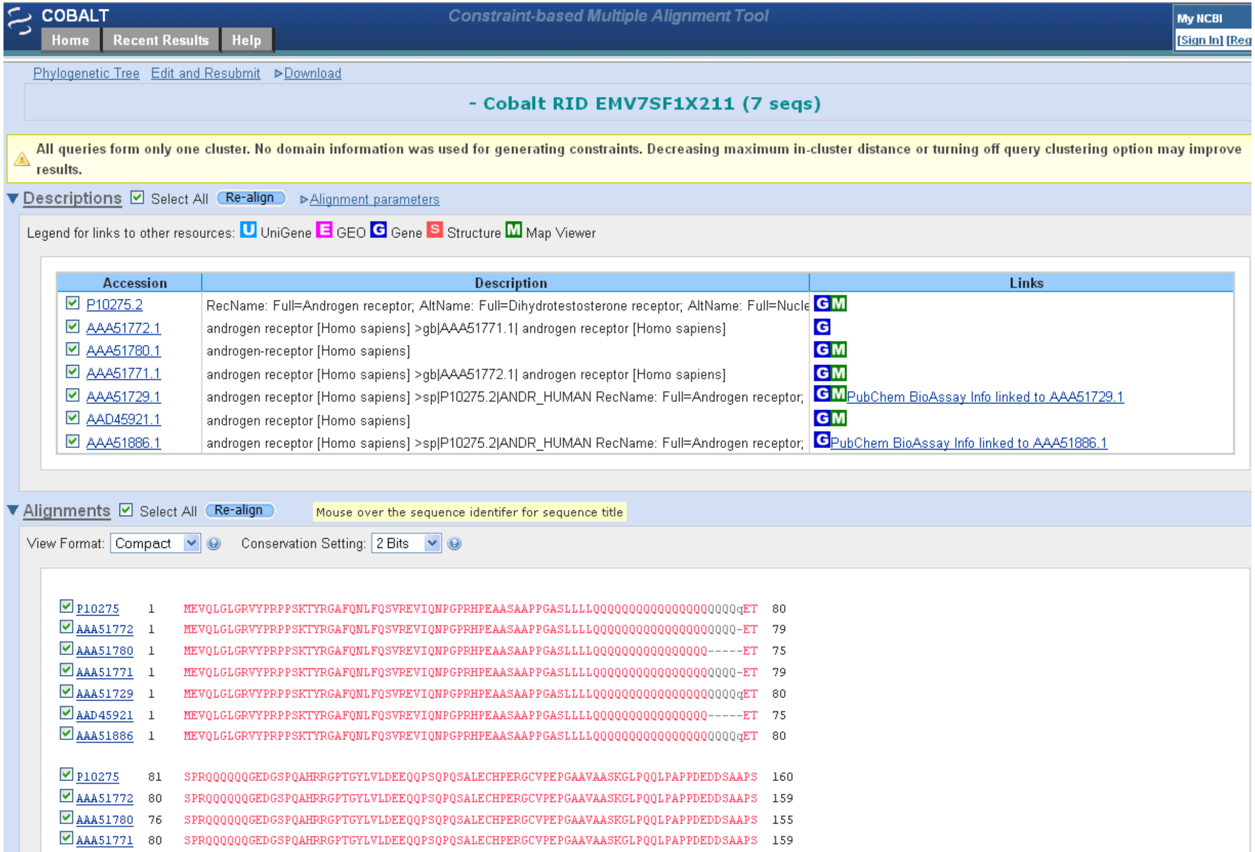
```
P10275.2
AAA51772.1
AAA51780.1
AAA51771.1
AAA51729.1
AAD45921.1
AAA51886.1
```

Or, upload FASTA file: No file selected.

Job Title:

☐ Show results in a new window

Alignment page will be generated



- Cobalt RID EMV7SF1X211 (7 seqs)

All queries form only one cluster. No domain information was used for generating constraints. Decreasing maximum in-cluster distance or turning off query clustering option may improve results.

Descriptions ☑ Select All Re-align Alignment parameters

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Accession	Description	Links
☑ P10275.2	RecName: Full=Androgen receptor; AltName: Full=Dihydrotestosterone receptor; AltName: Full=Nucle	G M
☑ AAA51772.1	androgen receptor [Homo sapiens] >gb|AAA51771.1| androgen receptor [Homo sapiens]	G
☑ AAA51780.1	androgen-receptor [Homo sapiens]	G M
☑ AAA51771.1	androgen receptor [Homo sapiens] >gb|AAA51772.1| androgen receptor [Homo sapiens]	G M
☑ AAA51729.1	androgen receptor [Homo sapiens] >sp|P10275.2|ANDR_HUMAN RecName: Full=Androgen receptor;	G M PubChem BioAssay Info linked to AAA51729.1
☑ AAD45921.1	androgen receptor [Homo sapiens]	G M
☑ AAA51886.1	androgen receptor [Homo sapiens] >sp|P10275.2|ANDR_HUMAN RecName: Full=Androgen receptor;	G PubChem BioAssay Info linked to AAA51886.1

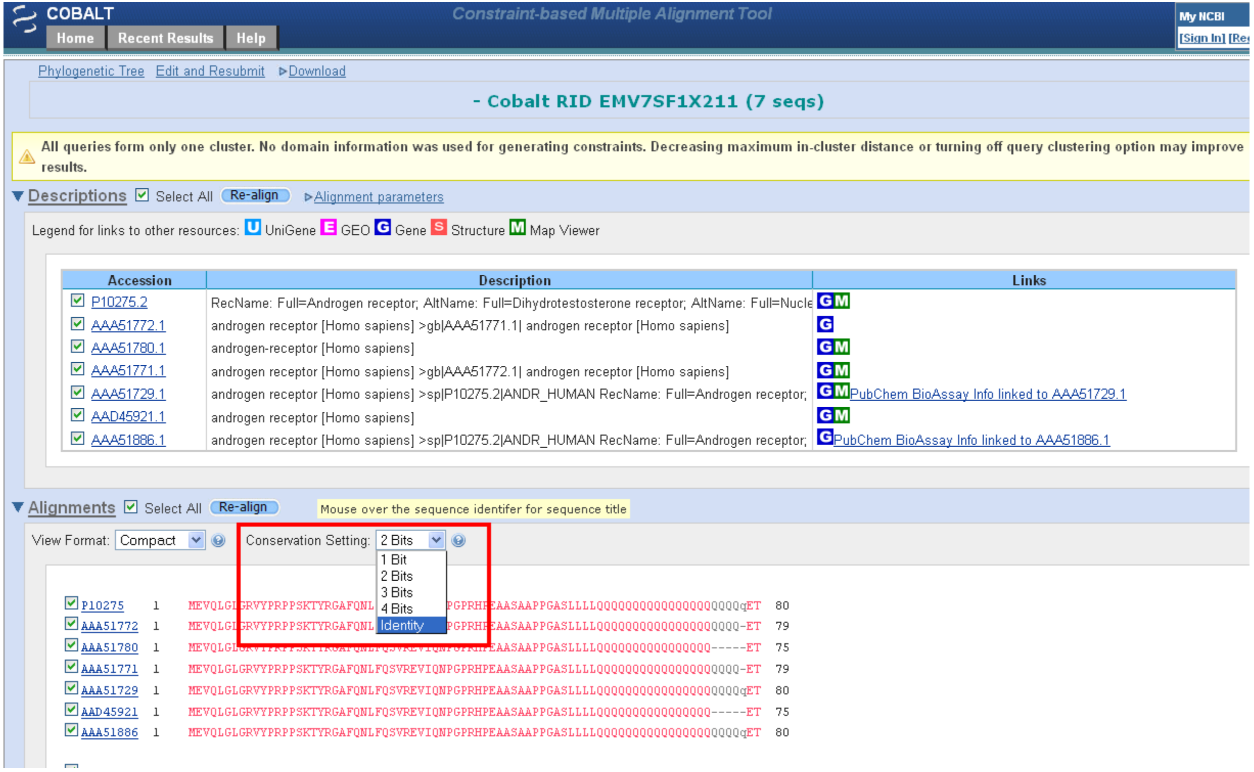
Alignments ☑ Select All Re-align Mouse over the sequence identifer for sequence title

View Format: Compact Conservation Setting: 2 Bits

```
☑ P10275     1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQqET   80
☑ AAA51772   1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQ-ET   79
☑ AAA51780   1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQ-----ET   75
☑ AAA51771   1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQ-ET   79
☑ AAA51729   1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQqET   80
☑ AAD45921   1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQ-----ET   75
☑ AAA51886   1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQqET   80

☑ P10275     81   SPRQQQQQQGEDGSPQAHRRGPTGYLVLDEEQQPSQPQSALECHPERGCVPEPGAAVAASKGLPQQLPAPPDEDDSAAPS   160
☑ AAA51772   80   SPRQQQQQQGEDGSPQAHRRGPTGYLVLDEEQQPSQPQSALECHPERGCVPEPGAAVAASKGLPQQLPAPPDEDDSAAPS   159
☑ AAA51780   76   SPRQQQQQQGEDGSPQAHRRGPTGYLVLDEEQQPSQPQSALECHPERGCVPEPGAAVAASKGLPQQLPAPPDEDDSAAPS   155
☑ AAA51771   80   SPRQQQQQQGEDGSPQAHRRGPTGYLVLDEEQQPSQPQSALECHPERGCVPEPGAAVAASKGLPQQLPAPPDEDDSAAPS   159
```

To evaluate sequences, change settings for “Conservation Setting” from “2 Bits” to “Identity”



- Cobalt RID EMV7SF1X211 (7 seqs)

All queries form only one cluster. No domain information was used for generating constraints. Decreasing maximum in-cluster distance or turning off query clustering option may improve results.

Descriptions ☑ Select All Re-align Alignment parameters

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Accession	Description	Links
☑ P10275.2	RecName: Full=Androgen receptor; AltName: Full=Dihydrotestosterone receptor; AltName: Full=Nucle	G M
☑ AAA51772.1	androgen receptor [Homo sapiens] >gb|AAA51771.1| androgen receptor [Homo sapiens]	G
☑ AAA51780.1	androgen-receptor [Homo sapiens]	G M
☑ AAA51771.1	androgen receptor [Homo sapiens] >gb|AAA51772.1| androgen receptor [Homo sapiens]	G M
☑ AAA51729.1	androgen receptor [Homo sapiens] >sp|P10275.2|ANDR_HUMAN RecName: Full=Androgen receptor;	G M PubChem BioAssay Info linked to AAA51729.1
☑ AAD45921.1	androgen receptor [Homo sapiens]	G M
☑ AAA51886.1	androgen receptor [Homo sapiens] >sp|P10275.2|ANDR_HUMAN RecName: Full=Androgen receptor;	G PubChem BioAssay Info linked to AAA51886.1

Alignments ☑ Select All Re-align Mouse over the sequence identifer for sequence title

View Format: Compact Conservation Setting: 2 Bits

```
☑ P10275     1    MEVQLGLGRVYPRPPSKTYRGAFQNL                  PGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQqET   80
☑ AAA51772   1    MEVQLGLGRVYPRPPSKTYRGAFQNL                  PGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQ-ET   79
☑ AAA51780   1    MEVQLGL                                      AASAAPPGASLLLLQQQQQQQQQQQQQQQQ-----ET   75
☑ AAA51771   1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQ-ET   79
☑ AAA51729   1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQqET   80
☑ AAD45921   1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQ-----ET   75
☑ AAA51886   1    MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQqET   80
```

Look for differences in the sequence (e.g., conserved residues, gaps) and start by eliminating sequences that have gaps.
i. If, after the suggested evaluations of the proteins are performed, questions remain as to which sequence would be best to run in SeqAPASS, run all relevant sequences in SeqAPASS for the evaluation. The individual residue differences between commonly named sequences will become most important when evaluating residues known to be important for binding the chemical or activating the protein (Level 3 SeqAPASS analysis). After completing the SeqAPASS run, select the data that has the greatest number of ortholog candidates for your evaluation of conservation and further predictions of cross species susceptibility. Depending on the protein of interest, multiple subunits may be associated with a protein. In this case, all relevant subunits can be queried using SeqAPASS.

Level 1 Calculated Percent Similarity

The SeqAPASS algorithms submit the query to NCBI's standalone BLASTp (using default settings, including BLOSUM-62 matrix), which aligns the query protein with all proteins available in the NCBI protein database and provides a variety of metrics associated with each pairwise alignment between the query and hit sequences. SeqAPASS selectively captures output from BLASTp, including one sequence per species with the highest bit score. Detailed descriptions of metrics derived from BLASTp (e.g., BLASTp Bitscore, E-Value, Positives, Identity, Hit length) can be found in:
The NCBI Handbook: (http://www.ncbi.nlm.nih.gov/books/NBK21106/);
BLAST® Help: (http://www.ncbi.nlm.nih.gov/books/NBK62051/) and the
NCBI Glossary Field Guide: (http://www.ncbi.nlm.nih.gov/Class/FieldGuide/glossary.html)

The top row of the Level 1 data corresponds to the queried protein selected by the user. For each sequence queried, the Level 1, top row query sequence is used to determine the maximum bitscore for the analysis, which is derived from aligning the query sequence to itself using BLASTp. To calculate percent similarity, the bitscore for each hit sequence is normalized to the maximum bit score and then multiplied by 100.

<u>Note:</u> SeqAPASS v2.0 and newer parse the BLASTp query and hit accessions to identify all the species/accessions from the identical proteins. Therefore, if a hit sequence represents multiple species, all species with the identical sequence will be found in the data table for Level 1 and Level 2. To determine which sequence/species was identified from BLASTp as a hit and which sequence/species was parsed from the identical sequence, view the "Full Report" for Level 1 or Level 2, column "Identical Protein," Where "N" is indicative of the original hit sequence and "Y" is the parsed sequence.

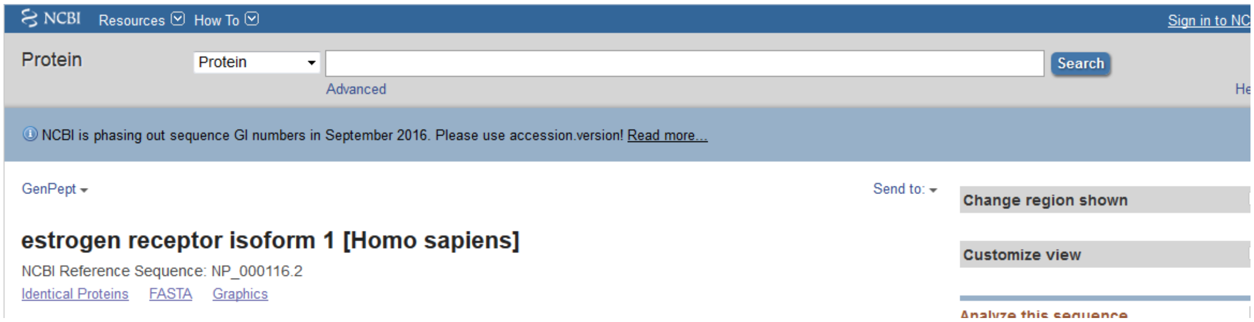
Common Domain Count

Reversed Position Specific BLAST (RPS BLAST) is used to compare each query and hit sequence to conserved domains defined in NCBIs Conserved Domain Database. A hit domain is considered in common with the query domain if it contains the same domain accession as the query and it aligns with the NCBI curated domain with the same or greater amino acid residue coverage than the query sequence.

Ortholog Candidate Identification

Ortholog sequences are those that have diverged from a speciation event and therefore are more likely to maintain similar function. SeqAPASS uses reciprocal best hit (RBH) BLAST for ortholog detection by automatically comparing each hit protein to all protein sequences available for the query species and if the original query protein or one of its identical protein matches is identified to by the best match to the hit or maintain the same bitscore, then the hit sequence would be considered an ortholog candidate. The sequence is indicated an Ortholog Candidate or not with a yes (Y) or no (N) in the column.

Note: Many NCBI protein accessions represent multiple identical protein sequences in the BLASTp output. This is due to BLASTp querying and presenting data from the non-redundant protein database. Sometimes the identical sequences are from different species. This can be checked by following the link for the top row "NCBI Accession" in the table to the NCBI protein page. Below the protein name [species] title will be a link to "Identical Proteins."
Click the "Identical Proteins" link and look for a sequence in the list from the user defined query species.



NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! Read more...

estrogen receptor isoform 1 [Homo sapiens]

NCBI Reference Sequence: NP_000116.2

Identical Proteins FASTA Graphics

Note: If the top hit is a Protein DataBank (PDB) code (e.g., 1AHR_A) from RBH BLAST there will be no ortholog candidates identified. BLASTp when ran against all accessions for a given species does not return PDB codes. It is recommended that the user identify a similar/identical sequence to the PDB code and use that sequence as the query sequence.

Susceptibility cut-off

The susceptibility cut-off values listed on the "Level 1 (and Level 2) Susceptibility Cut-off" page are determined by plotting the % similarity data from the "Primary Report" or "Full Report" and identifying the local minimums in the data. The default cut-off is determined by taking the 1st local minimum and moving up in percent similarity until the next ortholog candidate is found. The susceptibility cut-off displayed in the list is the percent similarity of the identified ortholog candidate.

Criteria for Susceptibility Prediction (when "Primary Report Settings" is set to "Species Read-Across:" Yes)

All sequences identified above the susceptibility cut-off are predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes"

If the hit sequence is below the susceptibility cut-off, but identified as an Ortholog Candidate = Y, for "yes," then the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes"

If the hit sequence is below the susceptibility cut-off but belongs to any organism class found above the susceptibility cut-off, the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes". This criterion allows susceptibility predictions to be made across taxonomic groups based on the likelihood that the sequences above the cut-off are better matches to the query.

If the hit sequence is below the susceptibility cut-off and not identified as an ortholog candidate (Ortholog Candidate = N, for "no,") and does not belong to any organism class found above the susceptibility cut-off, the hit is predicted to not be susceptible; therefore, Susceptibility Prediction = N for "no"

Note that the "Primary Report" may yield different Susceptibility Predictions than the "Full Report," as the predictions are based on the data in the different reports. The Primary Report is filtered to only display E-value ≤ 0.01 and Common Domain Count ≥ 1.

Criteria for Susceptibility Prediction (when "Primary Report Settings" is set to "Species Read-Across:" No)

All sequences identified above the susceptibility cut-off are predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes"

If the hit sequence is below the susceptibility cut-off, but identified as an Ortholog Candidate = Y, for "yes," then the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes"

If the hit sequence is below the susceptibility cut-off and not identified as an ortholog candidate (Ortholog Candidate = N, for "no,"), the hit is predicted to not be susceptible; therefore, Susceptibility Prediction = N for "no"

Level 2 Calculated Percent Similarity

Data obtained from the Level 1 RPS BLAST evaluation is used to assign sequence ranges that aligned with a user selected domain (from the NCBI CDD database) to each accession from the Level 1 Full report. BLASTp is then used to align the query domain range to each hit domain range. The percent similarity is calculated based on the bit scores from the BLASTp alignment of the domain regions. For each sequence queried, the Level 2, top row query species is used to determine the maximum bitscore for the analysis, which is derived from aligning the query sequence to itself using BLASTp. To calculate percent similarity, the bitscore for each hit sequence is normalized to the maximum bit score and then multiplied by 100.

Susceptibility cut-off (same method as used in Level 1)

The susceptibility cut-offs listed on the "Level 2 Susceptibility Cut-off" page are determined by plotting the % similarity data from the "Primary Report" or "Full Report" and identifying the local minimums in the data. The default cut-off is determined by taking the 1st local minimum and moving up in percent similarity until the next ortholog candidate is found. The susceptibility cut-off displayed in the list is the percent similarity of the identified ortholog candidate.

Level 2 Criteria for Susceptibility Prediction (when "Primary Report Settings" is set to "Species Read-Across:" Yes)

All sequences identified above the susceptibility cut-off are predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes"

If the hit sequence is below the susceptibility cut-off, but identified as an Ortholog Candidate = Y, for "yes," then the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes"

If the hit sequence is below the susceptibility cut-off but belongs to any organism class found above the susceptibility cut-off, the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes". This criterion allows susceptibility predictions to be made across taxonomic groups based on the likelihood that the sequences above the cut-off are better matches to the query.

If the hit sequence is below the susceptibility cut-off and not identified as an ortholog candidate (Ortholog Candidate = N, for "no,") and does not belong to any organism class found above the susceptibility cut-off, the hit is predicted to not be susceptible; therefore, Susceptibility Prediction = N for "no"

Note that the "Primary Report" may yield different Susceptibility Predictions than the "Full Report," as the predictions are based on the data in the different reports. The Primary Report is filtered to only display E-value ≤ 0.01 and Common Domain Count ≥ 1.

Level 2 Criteria for Susceptibility Prediction (when "Primary Report Settings" is set to "Species Read-Across:" No)

All sequences identified above the susceptibility cut-off are predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes"

If the hit sequence is below the susceptibility cut-off, but identified as an Ortholog Candidate = Y, for "yes," then the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes"

If the hit sequence is below the susceptibility cut-off and not identified as an ortholog candidate (Ortholog Candidate = N, for "no,"), the hit is predicted to not be susceptible; therefore, Susceptibility Prediction = N for "no"

Level 3 Sequence Alignments

COBALT is used to align all user selected sequences (from Level 1 hits) with a user defined template sequence. Because COBALT algorithms align all sequences, it is recommended that the user align the template sequence with sequences that are most similar to one another. As a means to capture the most similar sequences from the SeqAPASS data it is recommended that the user filter the Level 1 data by taxonomic group and step through the Level 1 data pages one by one while selecting sequences. It is recommended that the user look at the name of the sequence and exclude 'partial' sequences when possible. Requesting a query from one taxonomic group at a time, breaks the data down in manageable alignments.

Selecting Amino Acid Residues to Align

The user may select up to 50 amino acid residues to compare across selected species in Level 3.