

TEACHING THE GENOME GENERATION

PolyPhen-2 Tutorials



PolyPhen-2 Tutorials

Use the following menu to link to a specific topic within this tutorial:

- A. [PolyPhen-2 Background and Home Page](#)
- B. [Submit a query for a variant in a human protein](#)
 - i. [Using a protein identifier](#)
 - ii. [Using a SNP identifier](#)
 - iii. [Using a protein sequence](#)
- C. [Accessing your results](#)
 - i. [Grid status and job status](#)
 - ii. [Viewing your results](#)
- D. [Interpreting a PolyPhen-2 Report](#)
 - i. [Query Section](#)
 - ii. [Results Section: Variant Effect Prediction](#)
 - iii. [Details Section](#)

Reference:

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, & Sunyaev SR. (2010). A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248-9. doi: 10.1038/nmeth0410-248.

A. PolyPhen-2 Background and Home Page

Polymorphism Phenotyping v2 (PolyPhen-2) is a bioinformatics tool that predicts the impact of a single amino acid substitution on the structure and function of human proteins. See also the [PolyPhen-2 Tutorial Series: Introduction](#) video.

1. Navigate to [PolyPhen-2](#). The top of the main page contains a short description of the tool (1) and menu tabs that direct to additional information about and documentation for the tool (2). At the bottom of the page is a form (3) into which you can enter details about your protein and variant of interest for submission to the tool.

PolyPhen-2 prediction of functional effects of human nsSNPs

Home About Help Downloads Batch query WHESS.db

1 PolyPhen-2 (Polymorphism Phenotyping v2) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. Please, use the form below to submit your query.

21-Jun-2021: Server has been migrated to new hardware. Note, all queries were terminated and user sessions data discarded in the process, hence you will need to resubmit your query if affected. We apologize for the inconvenience caused.

3 **Query Data**

Protein or SNP identifier

Protein sequence in FASTA format

Position

Substitution AA₁ A R N D C E Q G H I L K M F P S T W Y V
AA₂ A R N D C E Q G H I L K M F P S T W Y V

Query description

Submit Query Clear Check Status

Display advanced query options

Software & web support: wan adzhubey Web design & development: biobyte solutions

Figure 1. PolyPhen-2 Homepage. The PolyPhen-2 homepage includes (1) a description, (2) a menu, and (3) a query submission form.

2. For this tutorial, we will look at a variant in the human protein taste receptor type 2 member 38 (TAS2R38). We will investigate a variant that results in an amino acid change from isoleucine (I) to valine (V) at amino acid #296 in the TAS2R38 protein sequence.

B. Submit a query for a variant in a human protein

This section will demonstrate three different methods for how to submit a query for a variant in a human protein: using a protein identifier, a SNP identifier, or the protein sequence in FASTA format. This tutorial begins on the home page of [PolyPhen-2](#). See also the PolyPhen-2 Tutorial videos: [Submitting a Query using a Protein Identifier](#), [Submitting a Query using a SNP identifier](#), and [Submitting a Query using a Protein Sequence](#).

i. *Submit a query using a protein identifier*

PolyPhen-2 recognizes protein accession numbers and entry names from the [UniProtKB](#) database. It also recognizes RefSeq identifiers and standard gene symbols. However, use of UniProtKB identifiers is recommended. When using a UniProtKB identifier, you do not need to enter the protein FASTA sequence into the query submission form.

In this tutorial, we will be looking at human taste receptor type 2 member 38 (TAS2R38). The UniProtKB entry name TAS2R38 for is T2R38_HUMAN.

The screenshot shows the 'Query Data' section of the PolyPhen-2 interface. It includes a form with the following fields and actions:

- 1**: 'Protein or SNP identifier' field containing 'T2R38_HUMAN'.
- 2**: 'Position' field containing '296'.
- 3**: 'Substitution' table with two rows (AA₁ and AA₂) and columns for amino acids. The AA₁ row has 'I' selected under 'H'. The AA₂ row has 'V' selected under 'Y'.
- 4**: 'Query description' field containing 'TAS2R38 Variant'.
- 5**: 'Submit Query' button.
- 6**: 'Display advanced query options' link.

Figure 2. Query Submission with Protein Identifier. Submitting a query with (1) a protein identifier requires (2) the position of the variant amino acid, (3) the reference amino acid, and (4) the variant amino acid.

1. Locate the text box labeled “Protein or SNP Identifier” (1). Type the protein identifier into the box. For this example, type:
T2R38_HUMAN
Note: You could also use the UniProtKB accession number for human TAS2R38: P59533.
2. In the text box labeled “Position” (2), enter the position of the variant amino acid within the protein sequence. For this TAS2R38 variant, type:
296
3. Select the appropriate box in the AA₁ row (3) to represent the amino acid at the variant position in the reference sequence. For the TAS2R38 sequence, select I for isoleucine.
4. Select the appropriate box in the AA₂ row (4) to represent the substitution amino acid. For this TAS2R38 variant, select V for valine.
5. In the text box labeled “Query description” (5), enter a description that will enable you to remember the contents of this query. For this example, you might type:
TAS2R38 Variant

6. Leave all other settings as they are and press the “Submit Query” button (6).
7. Continue to section **C. Accessing your results**.

ii. *Submit a query using a SNP identifier*

PolyPhen-2 recognizes human reference SNPs from the [dbSNP](#) database. When using a reference ID from dbSNP, you do not need to enter the protein FASTA sequence, the variant position, or the amino acid substitution into the query submission form.

In this tutorial, we will be investigating a variant in TAS2R38 that results in an amino acid change from isoleucine (I) to valine (V) at amino acid #296 in the TAS2R38 protein sequence. The dbSNP reference ID for this variant is rs10246939.

The screenshot shows the PolyPhen-2 query submission form. The form is titled "Query Data" and contains several fields and buttons. A yellow box labeled "1" highlights the "Protein or SNP identifier" field, which contains the text "rs10246939". Below this field is a large text area for "Protein sequence in FASTA format", a "Position" field, and a "Substitution" dropdown menu. The "Substitution" menu is open, showing two rows of amino acid options: "AA₁ A R N D C E Q G H I L K M F P S T W Y V" and "AA₂ A R N D C E Q G H I L K M F P S T W Y V". A yellow box labeled "2" highlights the "Query description" field, which contains the text "TAS2R38 Variant". To the right of the "Query description" field are three buttons: "Submit Query", "Clear", and "Check Status". A yellow box labeled "3" highlights the "Submit Query" button. Below the "Submit Query" button is a link that says "Display advanced query options".

Figure 3. Query Submission with SNP Identifier. Submitting a query with a SNP identifier only requires the identifier.

1. Locate the text box labeled “Protein or SNP Identifier” (1). Type the SNP identifier into the box. For this example, type:
rs10246939
2. In the text box labeled “Query description” (2), enter a description that will enable you to remember the contents of this query. For this example, you might type:
TAS2R38 Variant
3. Leave all other settings as they are and press the “Submit Query” button (3).
4. Continue to section **C. Accessing your results**.

iii. *Submit a query using a protein sequence*

PolyPhen-2 also accepts protein sequences in FASTA format in place of a Protein or SNP identifier. You can find sequences in FASTA format by searching for your protein of interest on [UniProt](#) or the [NCBI Protein database](#).

Query Data																																										
Protein or SNP identifier	<input type="text"/>																																									
1 Protein sequence in FASTA format	>sp P59533 T2R38_HUMAN Taste receptor type 2 member 38 OS=Homo sapiens OX=9606 GN=TAS2R38 PE=2 SV=3 MLTLTRIRTVSYEVRSTFLFISVLEFAVGFLTNAFVFLVNFWDVVKRQALSNSDCVLLCL																																									
2 Position	296																																									
Substitution	<table border="0"> <tr> <td>AA₁</td> <td>A</td><td>R</td><td>N</td><td>D</td><td>C</td><td>E</td><td>Q</td><td>G</td><td>H</td><td>I</td><td>K</td><td>M</td><td>F</td><td>P</td><td>S</td><td>T</td><td>W</td><td>Y</td><td>V</td> </tr> <tr> <td>AA₂</td> <td>A</td><td>R</td><td>N</td><td>D</td><td>C</td><td>E</td><td>Q</td><td>G</td><td>H</td><td>I</td><td>L</td><td>K</td><td>M</td><td>F</td><td>P</td><td>S</td><td>T</td><td>W</td><td>Y</td><td>V</td> </tr> </table>	AA ₁	A	R	N	D	C	E	Q	G	H	I	K	M	F	P	S	T	W	Y	V	AA ₂	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
AA ₁	A	R	N	D	C	E	Q	G	H	I	K	M	F	P	S	T	W	Y	V																							
AA ₂	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V																						
5 Query description	TAS2R38 Variant																																									
6 <input type="button" value="Submit Query"/> <input type="button" value="Clear"/> <input type="button" value="Check Status"/>																																										

Display advanced query options

Figure 4. Query Submission with Protein Sequence. Submitting a query with (1) a FASTA sequence requires (2) the position of the variant amino acid, (3) the reference amino acid, and (4) the variant amino acid.

1. Locate the text box labeled “Protein sequence in FASTA format” (1). Copy and paste the protein sequence into the box. For this example, the TAS2R38 sequence from UniProt is:

```
>sp|P59533|T2R38_HUMAN Taste receptor type 2 member 38 OS=Homo
sapiens OX=9606 GN=TAS2R38 PE=2 SV=3
MLTLTRIRTVSYEVRSTFLFISVLEFAVGFLTNAFVFLVNFWDVVKRQALSNSDCVLLCL
SISRLFLHGLLFLSAIQQLTHFQKLSEPLNHSYQAIIMLWMIANQANLWLAACLSLLYCSK
LIRFSHTFLICLASWVSRKISQMLLGIILCSCICTVLCVWCFFSRPHFTVTTVLFMNNNT
RLNWQIKDLNLFYSFLFCYLWSVPPFLLFLVSSGMLTVSLGRHMRTMKVYTRNSRDPSE
AHIKALKSLVSFFCFVVISSCAAFISVPLLILWRDKIGVMVCVGIMAACPSGHAAIILISG
NAKLRRVMTILLWAQSSLKVRADHKADSRTL
```

Note: You must include a definition line (starting with “>”).

2. In the text box labeled “Position” (2), enter the position of the variant amino acid within the protein sequence. For this TAS2R38 variant, type:
296
3. Select the appropriate box in the AA1 row (3) to represent the amino acid at the variant position in the reference sequence. For the TAS2R38 sequence, select I for isoleucine.
4. Select the appropriate box in the AA2 row (4) to represent the substitution amino acid. For this TAS2R38 variant, select V for valine.

- In the text box labeled “Query description” (5), enter a description that will enable you to remember the contents of this query. For this example, you might type:

TAS2R38 Variant

- Leave all other settings as they are and press the “Submit Query” button (6).

C. Accessing your results

This section will focus on how to track your submission progress and access your results through the Grid Gateway Interface (GGI) page. See also the [PolyPhen-2 Tutorial Series: Accessing Your Results](#) video.

i. Grid status and job status

When you submit a query, your query, also called a *job*, will be placed in a queue. You will be directed to the Grid Gateway Interface (GGI) page to track your job status. The *grid* is a network of computers to which you can remotely submit queries or jobs. Bioinformatics applications like PolyPhen-2 use a remote computer network or grid because a network of computers can analyze queries more quickly than an individual computer can.

The screenshot shows the GGI interface with several key elements highlighted by numbered callouts:

- 1 Service Name:** PolyPhen-2
- Session ID:** 03f387de58cc7a212ce5aab887064071870d98b7 Overwrite default
- 2 Grid Status:**

Load	Health	Jobs:	Pending	Running
High	88%		549	51
- 3 Jobs (2 total):** Completed (1)

ID	Results	Errors	Date/Time	Delete	Description
8597706	View	-	2022-11-04 12:46:05	<input type="checkbox"/>	Family 2 Allele 1
- 4 Jobs (2 total):** Pending/Running (1/0)

ID	Pos.	State	Date/Time	Delete	Description
8622522	6	qw	2022-11-21 10:05:17	<input type="checkbox"/>	TAS2R38 Variant
- 5 Refresh** All items with **Delete** boxes checked will be removed!

Figure 5. Grid Gateway Interface (GGI). The GGI details (1) the application in use, (2) the status of the computer network, (3) completed jobs or queries, and (4) pending and running jobs or queries.

- The GGI page displays information about the application (1), the status of the grid (2), a list of your completed jobs (3), and a list of your pending and running jobs (4).

- Find the line labeled “Service Name” (1) and confirm that it reads **PolyPhen-2**. If it does not, return to the [PolyPhen-2 homepage](#) and try your submission again.
- PolyPhen-2 is a real research tool. There may be researchers using the tool at the same time as you, which can affect how quickly you are able to obtain results. To get an idea of how long your query may take, check the Grid Status (2). The **Load** ranges from *Light* to *High* and the **Health** ranges from 0% to 100%. A *Light* load and a high percent Health indicate that you will likely get results quickly. A *High* load and/or low percent Health indicate that it may take a while for your results to be ready. You can also check the number of **Pending** jobs, which is the total number of queries currently in the queue to be analyzed. **Figure 6** shows an example of a grid status that would indicate a short wait.

Grid Status:				
Load	Health	Jobs:	Pending	Running
Light	100%		0	11

Figure 6. Light Grid Load.

- Check the status of your query under the Pending/Running Jobs section (4). When you submit your query, it will be given an identification (**ID**) number and placed in a queue to be processed. The **State** will likely read “qw” indicating that your query is in the queue. The Position (**Pos.**) column provides the position of your query in the queue. When the **Pos.** is 1, your query is next in line to be processed.
- Click the button labeled “Refresh” (5) to update the status of your query. If your query has completed, it will move to the Completed Jobs section (3). If your query is still listed under Pending/Running Jobs (4), wait a few minutes and then try to refresh again.

ii. *Viewing your results*

When your query has completed, it will be listed under Completed Jobs in the GGI.

Service Name: [PolyPhen-2](#)

Session ID: Overwrite default

Grid Status:

Load	Health	Jobs:	Pending	Running
High	88%		544	51

Jobs (2 total):

Completed (2)						
ID	Results	Errors	Date/Time	Delete	Description	
8597706	View	-	2022-11-04 12:46:05	<input type="checkbox"/>	Family 2 Allele 1	
8622522	View	-	2022-11-21 10:09:25	<input type="checkbox"/>	TAS2R38 Variant	

All items with **Delete** boxes checked will be removed!

Figure 7. Completed Job with Results. When PolyPhen-2 has finished analyzing your query, the results will be accessible via a link labeled View (2).

1. In the Completed Jobs section (1), look for a link labeled **View**. If the link is in the **Results** column, your query was processed successfully. Click the **View** link (2) in the row with your query to open the results.
2. If the **View** link is in the **Errors** column, as in the **Figure 8** below, PolyPhen-2 was not able to process your query. Click the **View** link to see a description of the error. You will need to resubmit your query, correcting any errors.

Jobs (1 total):

Completed (1)						
ID	Results	Errors	Date/Time	Delete	Description	
8638428	-	View	2022-12-02 16:46:17	<input type="checkbox"/>	TAS2R38 Variant	

Figure 8. Completed Job with Errors. If an error occurs during analysis, details of the error will be accessible via a link labeled View (2) in the Errors column.

D. Interpreting a PolyPhen-2 Report

This section will describe the information contained in a PolyPhen-2 results report. See also the [PolyPhen-2 Tutorial Series: Interpreting a PolyPhen-2 Report](#) video.

PolyPhen-2 reports contain three sections: Query (1), Results (2), and Details (3).

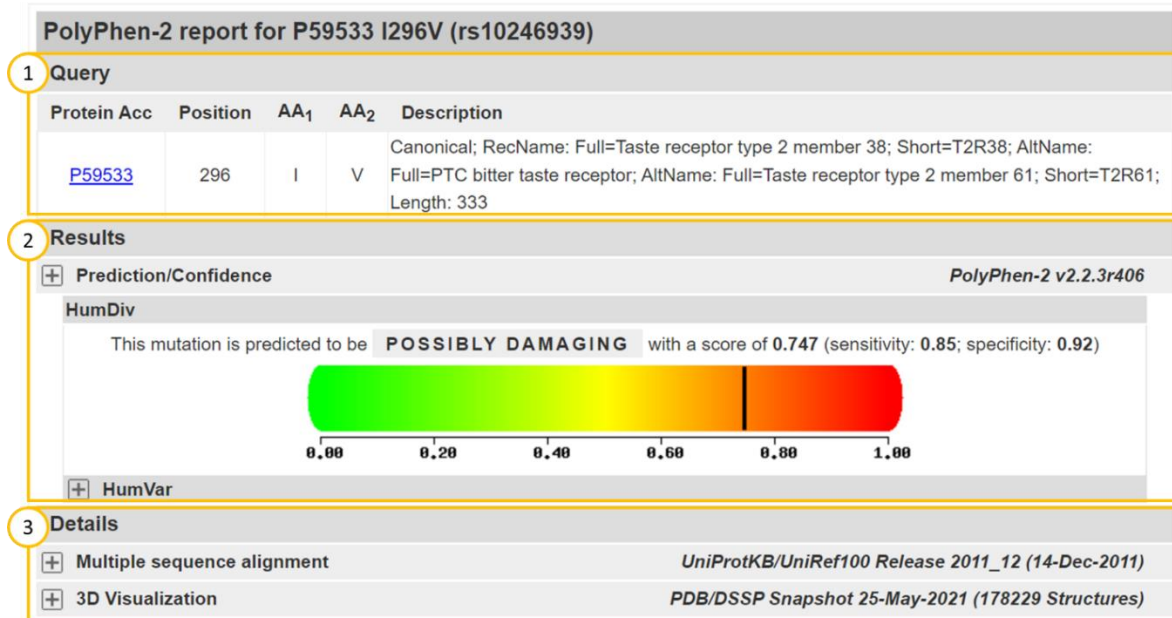


Figure 9. PolyPhen-2 Results Report. A results report includes (1) query information, (2) prediction results, and (3) data about the protein of interest.


i. Query Section

1. The Query section displays information about the protein from your query submission, including the Protein Accession Number (**Protein Acc**), the **Position** of the variant, the original (**AA₁**) and variant (**AA₂**) amino acids, and a **Description** of the protein.

ii. Results Section: Variant Effect Prediction

1. The Results section displays the results of the variant effect prediction models.
2. PolyPhen-2 uses two different models to make variant effect predictions: HumDiv and HumVar. These two models were trained using different datasets, and each have slightly different applications. In general, HumDiv is best used to evaluate rare variants involved in complex disease, whereas HumVar is best used for diagnostics of Mendelian diseases. For more information about the differences between these two models and applications for each, refer to the **Prediction** section of the [PolyPhen-2 overview documentation](#).

- By default, only the HumDiv prediction is displayed. To also view the HumVar prediction, click the plus [+] button next to the HumVar label.

 HumVar

- Both HumDiv and HumVar produce predictions as a score which falls between 0 and 1. High scores represent a high probability that the variant will affect protein function. Low scores represent a low probability that the variant will affect protein function. If there is insufficient data about your protein of interest, PolyPhen-2 will not provide a prediction score and will report the HumDiv and HumVar results as **unknown**.
- The results for both models display as a color gradient bar that represents the possible prediction scores. A vertical black bar represents the score for the variant in your query.

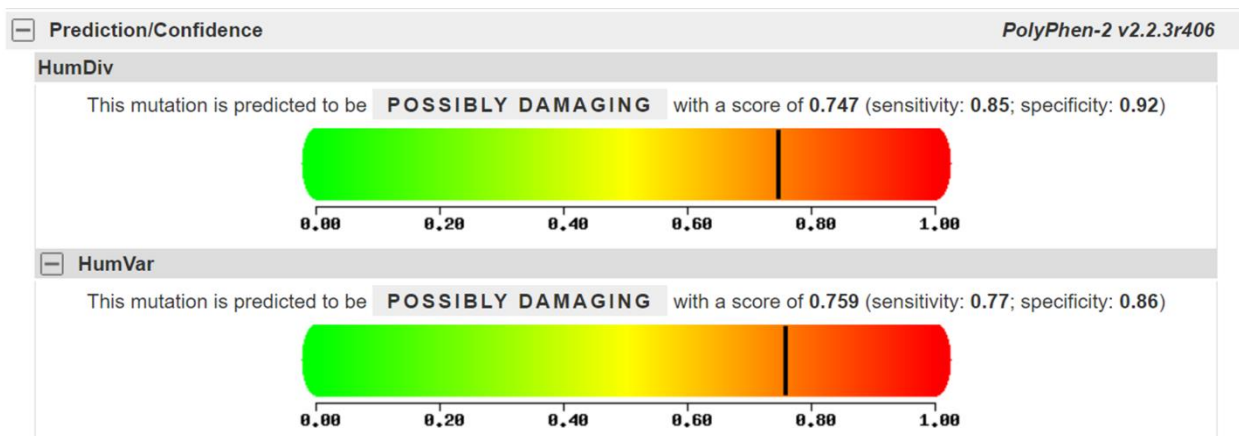


Figure 10. Prediction/Confidence Report. PolyPhen-2 provides variant effect predictions using two different models: HumDiv and HumVar.

- Locate the short description above the graph for each model. The description includes both the numerical probability score and a category. For the TAS2R38 variant, HumDiv calculated a score of 0.747 and HumVar calculated a score of 0.759.
- PolyPhen-2 divides all variant effect prediction scores into three categories: **benign**, **possibly damaging**, and **probably damaging**. A result of **benign** indicates that the variant is not likely to affect protein function. A result of **possibly damaging** indicates that the variant may affect protein function. A result of **probably damaging** indicates that the variant is likely to affect protein function. For the TAS2R38 variant, both models predicted that the variant is **possibly damaging**.
- The results also include sensitivity and specificity scores, which are measures of the confidence associated with these predictions. Generally, a higher score represents higher confidence. For more information about the sensitivity and specificity scores, refer to the **Prediction** section of the [PolyPhen-2 overview documentation](#).

iii. Details Section

1. The Details section of the PolyPhen-2 report contains additional information about the protein. Click the plus [+] button next to the **Multiple sequence alignment**.



2. The **Multiple sequence alignment** tab shows the alignment of 75 amino acids near the variant position in your protein of interest with similar proteins from other species. The top sequence is from your protein of interest, labeled QUERY. The variant amino acid is highlighted with a black box.

Multiple sequence alignment		UniProtKB/UniRef100 Release 2011_12 (14-Dec-2011)	
QUERY	SSCAAFISVPLLLIWRDKIGVMVCGIMAACPSG-AA	I	ISGNAK
sp F7BJT1#1	SFCAALISMPLLFLLWRNKIGMMVCGIMAACPSG-AA	I	ISSNAK
sp D2HHN1#1	SFCAALISVPLLLMLWLNKIGVMICVGIACPSI-AA	I	ISSNAK
sp G1Q9F1#1	SLCAALLSVPLLLVWLNKIGAMVCGIMAACPSG-AA	I	ISGNAK
sp G1M985#1	SFCAALISVPLLLMLWLNKIGVMICVGIACPSI-AA	I	ISSNAK
sp UPI000210723C#1	SFCAAVISVPLLLMLWLNKIGVMVCGILAACPSG-AV	I	IAGNAK
sp F1PAP8#1	SFCVALISVPLTMVWLNKIGVMICVGIACPSI-AA	I	ISGNAK
sp F6QSW4#1	SFCAVLISVPLLLML-HSKIVVMVSAWIMAVCPSG-AA	I	ISGNVAK
sp Q2ABD2#1	SFCVALISVPLTMVWLNKIGVMICVGIACPSI-AA	I	ISGNAK
sp F1PQS7#1	SFCVALISVPLTMVWLNKIGVMICVGIACPSI-AA	I	ISGNAK
sp F1PQS9#1	SFCVALISVPLTMVWLNKIGVMICVGIACPSI-AA	I	ISGNAK
sp Q7TQA6#1	SFCAALISVPLLLMLWLNKIGVMICVGIACPSI-AA	I	ISGNAK
sp Q4VHE7#1	SFCAALISVPLLLVWLNKIGVMVCGIMMAACPSG-AA	I	ISGNAK
sp UPI00022F5975#1	SFCAALASMPLLVWLNKIGVMVCGIMMAACPSG-AV	I	ISGNAK
sp F1MZD2#1	SLCAALFISVPLLLMLWLNKIGVMVCGIMAACPSG-AV	I	ISGNAK

Shown are 73 amino acids surrounding the mutation position (marked with a black box)

Figure 11. Multiple Sequence Alignment. PolyPhen-2 provides an alignment of the protein sequence near the variant site with similar proteins from other species.

3. The aligned sequences are listed underneath your query sequence. To find out more about any of the aligned sequences, click the sequence name to the left of the sequence. This will bring you to the relevant UniProt entry for that sequence.