**Clarification Question About HNH Endonuclease Function Determination in view of hits to the Ref Sequences**

In view phagesDB & HHPred data, we are seeking clarification of the HNH function status of the following five draft Glaske16 genes at positions: 44853-45341 bp (gp 70), 51656-52198 bp (gp 83), 54100-54426 bp (gp 91), 56773-57150 bp (gp 98), and 60940-61320 bp (gp 117). There respective sequences are provided, along with background information.

>Glaske16\_gp70\_(44853-45341 bp)

MPDGNQPACKYGACNDPVLARGFCKLHYYRNRDGKPMDGPRRSYSTGPRAWTYERLASVPITSTGAHQRVRRLWGSASLYPCATCGGPAKDWAYDGTDPTHYYEQGRKAWSHFSRWPEFYMPMCKPCHSNHDRRAAADELREYRQWKMRNPGKTLEDLEGVAZ

>Glaske16\_gp83\_(51656-52198 bp)

MDTIWKPIPQDPTGLYLASQDGRILRKEYVIEKLQSHGHLYRRVMPEKIVKQCIKDRAPSHGVHPIIQMRSSTQYASTVERRVSSLIAAAWHGLPYEAGDRTAQNDWRIGFIDGDPSNVHADNLEWVSNQGVNTHHSHDFYYENLKAYRAQAAVETAESFLARYYSPDEIDWSTAERIAAZ

>Glaske16\_gp91\_( 54100-54426 bp)

MPTNSKNGPRSRGRTGGKFERAKWRVLKANQICAHPDCRQLIDLDLKWPDPMSPTVNHIIPVKDLAWDDPLTYSVENLEPMHLVCNQRLGAGPRKKKPKHPQSRNWREZ

>Glaske16\_gp98\_(56773-57150 bp)

MALAGEAKREYQRQWRANRRAAWFAGKACVRCGSDEDLELDHVDPTLKVTNAVWSWSQERRDVELAKCQVLCNACHKAKTISQTVITIGLKAYRHGTCSMYEHHRCRCGLCRLWARNKKRRQRAAZ

>Glaske16\_gp117\_(60940-61320 bp)

MQREYMRRWVANRRSAFFASKQCAMCGAGEELELDHIDPTKKVDHRIWSWTDARRSEELAKCQVLCASCHKKKTGEQWYANRSVSENAHHGTSRRYRKMKCRCGLCRLGNTNRSRALRQRHRVPVEZ

HNH is expected to have a typical ββα-metal fold and Zn-finger motif (which would need protein modeling software to decipher; DOI: <https://doi.org/10.1038/srep42542>), and the Official Function List simply states that it “Must have H-N-H over a 30 aa span.” It would help students if there was an easy way to make this determination since it may not always be obvious in HHPred. Besides just considering the percent probability, should we also consider the e-values (and probably have an e-value cut-off)? Additionally, must it always hit chain A as well as the Zn-finger motif, could it hit other chains such as chain D, with non-zinc motifs such as for Manganese or strontium ions?

Reference sequences for HNH endonuclease are Sisi gp 99 and Arianna gp 54. Both match *Geobacillus virus* E2 hit 5H0M\_A in PDB, with Sisi having a 93.5% alignment, 98.7% probability, and E-value: 1e-7, while Arianna has 67.3% alignment, 98.7% probability, and E-value: 2.2e-7.

>Sisi\_gp99 MPRAPKVCRHAGCTTLTTTGTCPQHTTHRWGNHQGRKVPHRLQQATFRRDNWTCQSCGHTATPGSGQLHADHIQPRSRGGADTLDNMRTLCKACHAPKSRAEARGSNT

> Arianna\_gp54

MAWSNGSSRTSSKHWQALRASAKKQLGYYCCAVCGITPAGGARLELDHIIPVAEGGSDEMANLQWLCARHHAIKTRAESRRGAQRRAARRRLPQRPHPGLR

HHPred for Arianna is: PDB, *Geobacillus virus* E2, hit # 5H0M\_A, 67.3% alignment, Probability: 98.67%, E-value: 2.2e-7.

A screenshot of a computer

Description automatically generated with low confidence

The PDB hit for Arianna (<https://www.rcsb.org/structure/5H0M>) is below

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, font, number

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

In view of the above, we can now specifically ask about the following five draft Glaske16 genes at positions: 44853-45341 bp (gp 70), 51656-52198 bp (gp 83), 54100-54426 bp (gp 91), 56773-57150 bp (gp 98), and 60940-61320 bp (gp 117).

Glaske16 gp 70 (44853-45341 bp has the top PhagesDb hit as Skinny gp 71 which is called Hypothetical Protein, yet it is 100% identical, q1:s1, but has >10 hits to HNH endonuclease). I am inclined to call this an HNH endonuclease, except if the forum suggests otherwise. Again, below is its aa sequence:

MPDGNQPACKYGACNDPVLARGFCKLHYYRNRDGKPMDGPRRSYSTGPRAWTYERLASVPITSTGAHQRVRRLWGSASLYPCATCGGPAKDWAYDGTDPTHYYEQGRKAWSHFSRWPEFYMPMCKPCHSNHDRRAAADELREYRQWKMRNPGKTLEDLEGVAZ

However, this gene, like the two reference sequences, hits HNH chain A of the same *Geobacillus virus* E2 hit 5H0M\_A in PDB, with 75.5% alignment, 99.19% probability, and E-value: 2.5e-11, with everything exactly as seen above for the two reference sequences, including the HNH endonuclease at position 76-124 (<https://www.rcsb.org/structure/5H0M>).

A screenshot of a computer

Description automatically generated with medium confidence

Notably, Skinny gp 93 which is called HNH has got poor e-values

Next is Glaske16 gene at 51656-52198 bp (draft gp 83); its sequence is below:

MDTIWKPIPQDPTGLYLASQDGRILRKEYVIEKLQSHGHLYRRVMPEKIVKQCIKDRAPSHGVHPIIQMRSSTQYASTVERRVSSLIAAAWHGLPYEAGDRTAQNDWRIGFIDGDPSNVHADNLEWVSNQGVNTHHSHDFYYENLKAYRAQAAVETAESFLARYYSPDEIDWSTAERIAAZ

This one too hits HNH endonuclease in phagesDB. HHPred shows it in PDB with 54.1% alignment, Probability: 99.76%, E-value: 6.2e-18, but notably, it does not hit the same chain as the ref chain (it hits 1U3E\_M; <https://www.rcsb.org/structure/1U3E>) and no Zn+2 motif, but instead Mn+2 and Sr+2, but it also has the βα. See below:

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, line, font, plot

Description automatically generated

What is your verdict on this gene in Glaske16 at 51656-52198 bp in view of the above? I am inclined to call it HNH endonuclease, except if the forum suggests otherwise.

A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a phone

Description automatically generated with low confidence

Next is Glaske16 gp 91 at position 54100-54426 bp. Has several hits to HNH in phagesDB.

MPTNSKNGPRSRGRTGGKFERAKWRVLKANQICAHPDCRQLIDLDLKWPDPMSPTVNHIIPVKDLAWDDPLTYSVENLEPMHLVCNQRLGAGPRKKKPKHPQSRNWREZ

This has a low e-value but hits the same chain as the ref sequence, and the zinc motif (<https://www.rcsb.org/structure/5H0M> ), and is called HNH endonuclease, and another hit at 4H9D\_A (<https://www.rcsb.org/structure/4H9D>).

A picture containing text, font, screenshot

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

What is your verdict on this gene in Glaske16 gp91 at 54100-54426 bp in view of the above? I am inclined but wary to call it HNH endonuclease because of the e-values, but again, it hits are the same as the Ref sequences; any suggestions?

The next question is about the Glaske16 gp98 at position 56773-57150 bp. It has more than 60 hits to HNH endonuclease in phagesDB. Its sequence is below:

MALAGEAKREYQRQWRANRRAAWFAGKACVRCGSDEDLELDHVDPTLKVTNAVWSWSQERRDVELAKCQVLCNACHKAKTISQTVITIGLKAYRHGTCSMYEHHRCRCGLCRLWARNKKRRQRAAZ

It also hits the same hit 5H0M\_A in PDB with the same everything as the reference sequences, and high probability (98%), alignment 52.4%, but with not as great an e value (0.000029). What is your verdict on this one? See data below:

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated with low confidence

Finally, the Glaske16 gp117 at 60940-61320 bp. This gene has more than 70 hits to HNH endonuclease in phagesDB. What is your verdict on this one? Its sequence is:

MQREYMRRWVANRRSAFFASKQCAMCGAGEELELDHIDPTKKVDHRIWSWTDARRSEELAKCQVLCASCHKKKTGEQWYANRSVSENAHHGTSRRYRKMKCRCGLCRLGNTNRSRALRQRHRVPVEZ

It also hits the same hit 5H0M\_A in PDB (<https://www.rcsb.org/structure/5H0M>) with the same everything as the reference sequences Sisi gp 99 and Arianna gp 54, and high probability (98.05%), alignment 49.6%, but with not as great an e-value (0.000017). What is your verdict on this one? See data below.

A screenshot of a computer screen

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with low confidence

A picture containing text, electronics, screenshot, display

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence