James Jedediah Smith

BIFX 550

Find-A-Gene Summary

5/4/2022

**Overview of MID1**

The gene I chose is known by the official name of Midline 1. Its official symbol is MID1. There does not appear to be any official alternative names, but NCBI says that it has also been known by the symbols OS, FXY, OSX, OGS1, XPRF, BBBG1, GBBB1, MIDIN, RNF59, ZNFXY, and TRIM18. The BBG1 and GBB1 names are likely references to the associated disease Opitz G/BBB Syndrome (NCBI, 2022).

MID1 has 667 amino acids, is located on Chromosome X at position Xp22.2, and can be considered negative stranded or anti-sense. There are 9 human transcripts for MID1 on NCBI, but up to 23 can be found by using Ensembl. Both have high quality transcripts available. NCBI has 327 orthologs listed for MID1, while Ensembl has 200. Further information about this gene is easily accessible through either NCBI Gene ID 4281 or Ensembl Gene ID ENSG00000101871 (NCBI, 2022).

**Role of MID1 in Opitz G/BBB Syndrome**

I chose to investigate MID1 for novel orthologs due to its association with Opitz G/BBB Syndrome. The MID1 protein forms homodimers that associate with microtubules (NCBI, 2022). This is thought to play an important role in organizing the cytoskeleton during development (OMIM, 2022). MID1 also interacts with a large array of other proteins, cells, and pathways involved in development. Perhaps then, it should be of no surprise that mutations to MID1 can have catastrophic developmental results.

Opitz G/BBB Syndrome is a rare developmental disease that comes in two flavors: Type I and Type II. Both types have similar symptoms, the distinction is in their inheritance mechanisms. Type I is the more studied. It is understood to be X-linked through the MID1 gene. Type II is thought to be autosomal dominant and presumably less prevalent. Symptoms of both usually include wide-spaced eyes, abnormal throat development, genital deformities, and slow growth. About half of cases also result in intellectual disability, cleft lip, cleft palate, and brain defects (OMIM, 2022).

MID1 mutations are certainly the culprit in Type I cases. Mutations that disrupt the C-terminus of the MID1 protein inhibit its ability to interact with other proteins and subunits (OMIM, 2022). The result is a buildup of other protein products that bind in the wrong places or disrupt the function of other pathways during development (OMIM, 2022). This cascading effect results in Opitz G/BBB Syndrome. Unfortunately, there is still much we do not know about MID1 and how its disruption effects development.

**Initial Search and False Positive**

The navigation through NCBI and BLAST tools can be daunting and time consuming, so I did everything I could to help narrow down my search. First, I took the NCBI transcript NM\_000381.4 and ran it through BLASTN against the EST database. I excluded humans and other major species that I knew were orthologs. From there, I sorted the results to display the best matches first, and tried to find species not listed under the NCBI or Ensembl MID1 orthologs.

This process took some trial and error, as at first I was only using the Ensembl ortholog list, and then after investigating further, realized that it was already listed on the NCBI ortholog list. After switching to the NCBI ortholog list, I was able to narrow things down more quickly, and I located a hit on ES557302.1 in the species *Eubalaena glacialis*. However, when I tried to confirm this via BLASTP and restricting my search to the species, it continued to give me a strange error.

After further investigation of the resources available for this species on NCBI, I discovered that there is almost no documentation of *Eubalaena glacialis* at all. I was getting an error for my species specific BLASTP because there were no proteins for this species identified in the NCBI databases! This extended to other NCBI databases as well, such as the gene database. It was rather shocking

Not yet deterred, with the help of Professor Ravi, I was able to translate ES557302.1 to a protein using the ORFfinder NCBI tool and compare it with NP\_000372.1, the protein product of the transcript I had been searching. Unfortunately, the novel protein product did not match under any of its reading frames. It was cut short by start or end codons in all cases. This was a rather discouraging false positive, especially given how far I had pursued it.

**Discovery of Novel Ortholog**

It stands to reason that after a false positive, one would move on to explore other species. But after doing some research on *Eubalaena glacialis*, I decided to continue with it. Known more commonly as the North Atlantic Right Whale, *Eubalaena glacialis* was hunted to the verge of extinction back in the 1890s, and has a current global population of approximately 300 (NOAA, 2022). This may be the reason behind their lacking presence on NCBI. It quickly became clear to me that my false positive did not necessarily exclude a true positive, given the lack of information on them.

And so I dived into the SRA databases. There were 5 massive SRA studies available for *Eubalaena glacialis*, some of the little information NCBI had on them. I searched all five databases, and got several tiny promising hits across the entire length of the MID1 search query. I ultimately chose the SRA study SRR10251454, read 3998192, backwards. The total ID was SRR10251454.3998192.2, but you cannot search it that way. You have to bring up the SRA dataset and search inside it.

Further analysis confirmed that this read was indeed novel. The novel and MID1 transcripts matched with 96.03% identity and an E-value of 7e-62. Although the coverage was only 2%, this was mainly due to the fact all the reads in this SRA study were raw fragments. The confidence was high enough to proceed.

Figure 1: BLASTN search identifies novel candidate.

Table

Description automatically generated

Figure 2: Side-by-side transcript comparison.

Table

Description automatically generated

After confirming the transcripts had a match, I needed to check the proteins. This step had me anxious, as it is where I had been stopped last time. Using the ORFfinder NCBI tool, I identified ORF 5 as the frame to use for my novel gene. It runs the whole length of the protein and matched to MID1 in several different species via SmartBLAST. The region we used is part of the FN3 domain.

Figure 3: ORFfinder determines ORF 5 to be ideal frame.Graphical user interface, application

Description automatically generated

Figure 4: SmartBLAST links identified ORF to MID1.

Graphical user interface, text, application

Description automatically generated

Although I was convinced at this point, I double checked that the protein was novel using a BLASTP search of the refseq database. It was already established that there are no proteins identified for *Eubalaena glacialis*, so the results of this were predictable: no hits for that species. I also aligned the novel protein with other MID1 proteins using BLASTP and then COBALT. This backed up all the previous data and confirmed it was novel.

Figure 5: BLASTP confirms candidate is novel.

Graphical user interface, application

Description automatically generated

Figure 6: Novel protein aligned with human MID1 protein.

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generatedFigure 7: Novel protein aligned with various MID1 proteins.

**Further Analysis**

After confirming my novel ortholog was actually novel, I created phylogenetic trees using MEGA X to further compare it with other similar sequences. Initially I started with a parsimony tree, but the sequences were too similar for the program to distinguish between. I labeled most of them as “ambiguous” and would shuffle them around randomly upon creating a new tree. So I moved on to a maximum likelihood tree with bootstrapping, but the likelihood values appeared to be small.

Figure 8: Maximum likelihood tree with bootstrapping.

Graphical user interface, text, application

Description automatically generated

The last step of this journey was to generate a model for my novel protein. I did this using Phyre2. However, it was a struggle to find a MID1 model to compare it to for reference. I initially looked at the Protein Data Bank. Although there did not appear to be a full model available for MID1, there some models for specific regions or fragments of MID1. But unfortunately, I was unable to locate a fragment that contained the region aligning to my novel sequence. So instead, I looked on another site called AlphaFold, which did have a full protein model for MID1.

Upon comparing the models and bases, it becomes clear that my novel sequence fragment is not at either terminus of the protein. It is towards one side, but definitely in the middle. About half of it forms the bridge to a large alpha helix spiral, while the other half forms into the crisscrossing beta sheets.

Figure 9: Phyre2 predicted protein model for novel MID1 protein fragment.

Map

Description automatically generated

Figure 10: AlphaFold model for MID1 with novel region in the center.

Graphical user interface, application, Word

Description automatically generated

**Conclusions**

Due to the location of my novel protein in MID1, we can conclude that mutations in the region are not likely to cause of Optiz G/BBB Syndrome, as it is not at the end of the MID1 protein, and therefore not part of the C-terminus. However, the similarity between orthologs hints to the fact that this gene is essential to proper development, as its structure is highly conserved across species.

The identified novel region appears to be part of an FN3 domain, which are responsible for RGD cell recognition (NCBI, 2020). FN3 domains can found in a wide variety of animal proteins, and are estimated to be present in nearly 2% of all animal proteins (NCBI, 2020). Though it is worth noting, the novel protein we identified is clearly unique to MID1 type proteins, as evidenced by my BLASTP searches.

Over all, this novel identification of a MID1 ortholog in *Eubalaena glacialis* will likely do little to save the species from extinction or develop cures for humans with Optiz G/BBB Syndrome. However, it certainly speaks to how important MID1 is to the development of complex organisms. It would be interesting to see what type of organisms lack MID1 entirely or whether Opitz G/BBB Syndrome is a uniquely human disease. Perhaps we will learn more in the future.

**Important Sequences**

Here is a quick list of the nucleotide and protein sequences associated with my novel MID1 ortholog. They are in FASTA format for convenience.

>gnl|SRA|SRR10251454.3998192.2

CTGTTTGTCTTCAACTTCCCAGGCTCACTGCTGCGGCTGCCCGCCTGGTTGATGGCCTTGACAATGAAGATGTACTTGGTGCCGCTCTGCAGACCGTGGACGGTGTAGTGGTTCTGCTTGATGTTGGGCACGATCATCCAGCTATCGGCCG

>lcl|ORF5|Novel\_MID1\_Eubalaena\_glacialis

ADSWMIVPNIKQNHYTVHGLQSGTKYIFIVKAINQAGSRSSEPGKLKTN

**References**

NCBI, National Center for Biotechnology Information. (Updated 2022). “MID1 Midline 1 [ Homo sapiens (human) ].” Accessed May 4, 2022. <https://www.ncbi.nlm.nih.gov/gene/4281>

NCBI, National Center for Biotechnology Information. (Updated 2020). “Conserved Protein Domain Family: FN3.” Accessed May 4, 2022. <https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=238020>

NOAA, National Oceanic and Atmospheric Agency. (Updated 2022). “Species Directory: North Atlantic Right Whale.” Accessed May 4, 2022. <https://www.fisheries.noaa.gov/species/north-atlantic-right-whale>

OMIM, Online Mendelian Inheritance in Man. (Edited 2022). “#300000 OPITZ GBBB SYNDROME, TYPE I.” Accessed May 4, 2022. <https://omim.org/entry/300000>