**BIFX-550  
S. Ravichandran, Ph.D.  
Biology, Hood College, Frederick, MD 21702**

Ongoing Assignment-bifx550a

Your Name: James Jedediah Smith

There will be three ongoing assignments, bifx550a-c. Assignments will be released as we progress through the class. These assignments are set up to help the students complete BIFX550’s find-a-gene project. You will use your find-a-gene query (gene/protein) to answer these questions. If you are working on bifx550a, please download the file, rename it as your firstinitiallastname\_bifx550a.docx (example, SRavichandran\_bifx550a.doc). Complete the assignment and upload the document to Hood College BIFX550 class BlackBoard site before the deadline

*Please note when I say, “this gene” or “this protein”, it means your query gene/protein respectively.*

**Assignment bifx550a (for due date lookup Blackboard announcement pages)**

1. What is the gene/protein that you have chosen for your final project? Please provide NCBI and Ensembl IDs.

Midline 1 is the official name. MID1 is the official symbol. The NCBI ID is 4281. The Ensembl ID is ENSG00000101871.

1. Are there any alternate names for this gene?

There doesn’t appear to be any alternative names, but NCBI says that it is also known as OS, FXY, OSX, OGS1, XPRF, BBBG1, GBBB1, MIDIN, RNF59, ZNFXY, and TRIM18.

1. On what chromosome is this gene located and is it a positive or negatively stranded gene? What NCBI database do you have to go to, to get this information?

MID1 is found on Chromosome X and it is negative or anti-sense. This can be found by going to the MID1 page of their gene database and looking under the genomic context tab. It can also be found on the GenBank and RefSeq entries of the gene.

1. Compare and comment about the transcripts between Ensembl and NCBI?

They have the same chromosome and location for MID1, but NCBI ony has 9 transcripts while Ensembl has 23.

1. Please provide the protein IDs (UniProt, NCBI, and Ensembl) of the query protein that you will be using for your find-a-gene search and provide its sequence length (i.e. how many amino acids?)

The protein name is E3 ubiquitin-protein ligase Midline-1. The NCBI ID is NP\_000372.1, the Ensembl ID is ENSP00000312678.4, and the UniProt ID is O15344-1. It has a length of 667 amino acids.

1. Briefly explain what the functions are for the gene. *For example, the Caspase-9 gene is an enzyme that is critical for the apoptotic pathway in many tissues*.

The MID1 protein forms homodimers that associate with microtubules. This is thought to play an important role in anchoring microtubules, making it an important part organizing the cytoskeleton microtubules.

1. Are there any known homologs for your gene/protein? *We will cover Homology in future classes.*

Yes, there are known homologs. MID1 is consereved in many animals, including Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog.

1. For a normal (healthy) individual, at what sites (lung, heart, etc.) is this protein expressed? Where in NCBI is this information stored? Do you know whether there is any association of the query protein expression with any disease (or disorder) states (i.e. overexpression for cancer patients (compared to a healthy individual)? If so, please provide them.

It is normally present in many tissues, including the colon, endometrium, gall bladder, urinary bladder, and prostate. NCBI stores this information under the expression tab for the gene. Mutations in this gene have been known to cause the X-linked form of Opitz syndrome. However, this does not appear to be associated with abnormal presence in certain tissues.

1. Is there a **TATA-[A/T]-A** motif in your corresponding transcript? If so, give the transcript ID (NCBI/Ensembl) and the position range (ex. 130,000-130,005) of this motif. Please indicate the Assembly (Example: GRCh38) that you are using for this question. For example, if my mRNA is NM\_001042594 then my corresponding will be NP\_001036059 and my position could be 130,000-130,005. Note the position number shown here is an example range and not the true motif location.

For ENST00000317552.9 with GRCh38 Assembly:

3474 – 3480, 3514 – 3520, 3645 – 3651, 3778 – 3784, 4201 – 4207, and 5917 – 5923 all contain a TATAAA motif.

4058 – 4064 contains a TATATA motif.

1. Is there a major pathway you can associate your gene to?

It is involved with the Ubl conjugation pathway, interferon signaling, cytokine signaling, and the immune system.

1. Do you have any information about what protein(s) your query protein might interact with? Where do you go to get this information (NCBI or Ensembl)?

Yes. It interacts with MID1IP1 and MID2, which are involved in similar processes, PPP2CB which is involved with cell growth, and IGBP1 which is involved wit B-cell differentiation. There are another ~100 associated things listed under the Interactions section of the NCBI gene page for MID1.

1. What conserved domains are known for your protein? Domains are conserved part of a protein sequence. Its 3D structure can independently exist, evolve, and function. For example, the kinase domain (function: phosphorylation) has been found in many protein families. There are roughly 53 million domains classified into some 2000 superfamilies. We covered this concept in class 3 and will go over it again in later classes.

NCBI lists approximately 210 conserved domains for the MID1 gene. Some select examples conserved among MID genes include BBC, PRY, zf-C3HC4, BBOX, FN3, SPRY, and RING. The examples are found on the NCBI HomoloGene page. But a more detailed list can be found at the second link.

<https://www.ncbi.nlm.nih.gov/homologene/7837>

<https://www.ncbi.nlm.nih.gov/cdd?LinkName=gene_cdd&from_uid=4281>

1. Are there any disease(s) associated with this gene/protein? Can you cite 1/2 relevant review/paper (PubMed links would be OK) that show associations of this gene with the disease(s)?

Yes. It is associated with Opitz GBBB Syndrome, Type 1.

<https://omim.org/entry/300552>

<https://pubmed.ncbi.nlm.nih.gov/18172692/>

<https://pubmed.ncbi.nlm.nih.gov/17221865/>

1. We will cover the NextGen or Micro-array technologies later in the class. In the genomic databases class, we briefly talked about where the NextGen or Microarray data is stored in NCBI? Please provide the database names and visit the databases to explore what information is available for your protein or gene. Please provide a one/two-line summary of the search results for your gene/protein from these databases.

The NextGen and Mircoarray data can be found under GEO DataSets. There are only 13 results for MID1, but only a handful of them have the gene name MID1 in their title or description. It seems most of the datasets are from humans and fruit flies.

1. Are there any SNPs known for your proteins? What database(s) do you go to get this information?

Yes. There are 103,852 SNPs for MID1. They can be found under the NCBI SNP database located at the link below. <https://www.ncbi.nlm.nih.gov/snp/?LinkName=gene_snp&from_uid=4281>

1. What organism(s) do you think you will avoid when you search for the novel gene? We haven’t learned about how to do the search but at this point, you can make an educated guess on what organisms (databases) you will avoid?

It is clear from the orthlog and homolog tables that MID1 is found in a variety of organisms from invertibrates to mammals. However, the mammals appear to make up a disproportionate amount of the discovered orthologs. As such, I think I will avoid searching for mammals since clearly MID1 has been studied there the most and I want to find something new.