Steps (PW - 20190830):

Download results file (“2019-4-9\_NW1\_serum\_F002042\_cRAP filtered.csv”)

275 rows

Protein grouping has single sequence per family member

Narwhal serum?

(*https://github.com/pwilmart/fasta\_utilities*)

Download protein FASTA file (“GCF\_005190385.1\_NGI\_Narwhal\_1\_protein.faa.gz”)

Analyze FASTA file

45,047 sequences (count\_fasta.py)

7971 redundant sequences (check\_for\_duplicates.py)

84% of tryptic peptides are redundant (FASTA\_digester.py)

(*https://github.com/pwilmart/PAW\_BLAST*)

Open “2019-4-9\_NW1\_serum\_F002042\_cRAP filtered.csv” file with Excel

Save as Excel “.xlsx” format file

Save 275 accessions (with header) to text file (“identifications.txt”)

Make subset FASTA database (make\_subset\_DB\_from\_list\_3.py)

Needed to decompress file

File needed “.fasta” extension

Proteins are in the order they are in the FASTA file (not in list order)

Download a human canonical FASTA file (UniProt\_reference\_proteome\_manager.py)

Any version within a few months would also be fine

Make a BLAST subfolder

Move subset Narwhal FASTA file and human FASTA to subfolder

Run BLAST ortholog script (db\_to\_db\_blaster.py)

Select Narwhal DB first, then human

BLAST script makes a tab-delimited results file

Query protein are in the order that they are in the original DB

There is useful information before and after the mapping table

Open “2019-4-9\_NW1\_serum\_F002042\_cRAP filtered.xlsx” file

Turn on column filters for row 29

Add “Counter” and “Order” columns (Order is just a series fill column)

Sort ascending by “Accession”

Open BLAST script txt output file. With Excel:

“GCF\_005190385.1\_NGI\_Narwhal\_1\_protein\_subset\_vs\_2019.06\_UP000005640\_9606\_Homo\_sapiens\_canonical.txt”

Turn on column filters for row 6

Sort ascending by “query\_acc”

Select ortholog table (A6:O281) or subset and copy to clipboard

Paste orthologs to the right of the table in “2019-4-9\_NW1\_serum\_F002042\_cRAP filtered.xlsx”

Check that “Accession” and “query\_acc” columns are in the same row order

Sort to put table back in original order

Select hit\_acc column (without header) and copy to clipboard

*(https://github.com/pwilmart/annotations)*

Run annotation script (add\_protein\_annotations\_6.py)

Human species should be default radio button

Click “Get accessions” button

Check “Pathways” and “GO Terms” checkboxes (if wanted)

Click “Parse DAT file” button

Need to browse to a human Swiss-Prot flat text file

“keywlist.txt” file should also be present

Click “Add annotations” button

Have to select location for output files to be written

Annotations will be written to clipboard

Paste annotations to the right of table in “2019-4-9\_NW1\_serum\_F002042\_cRAP filtered.xlsx”

Check that “Index” column and “hit\_acc” column match row-by-row

Format table

Column widths

Format hyperlinks

Deleted any uninteresting columns

Add column key in a ReadMe tab?

Save annotated results