**HPPP Proteome: Combining GENCODE and UniProtKB datasets**

The goal of this pipeline is to combine the data from all protein coding genes in GENCODE with the information in UniProtKB

**Install and Run:**

git clone <https://github.com/HPPProteome/HPP_TargetList/>

cd HPP\_TargetList/build\_pipeline

python RUN\_ME.py

**All Scripts and their output:**

**Protein\_list\_builder.py:**

* Downloads the newest version GENCODE’s basic gene annotations GTF file and unzips it into gencode.v46.annotation.gtf if file is not already present.
* Outputs an excel file: coding\_protiens.xlsx, which contains all 19,411 protein coding genes in the GENCODE GTF file.

**Link\_to\_uniprot.py:**

* Downloads all human genes from UniProtKB in a tsv file and unzips it into uniprot.tsv if file is not already present.
* Takes in the coding\_protiens.xlsx file from Protein\_list\_builder.py to make connections with.
* Outputs an excel file: uniprot\_output.xlsx, which contains an updated table of the 19,411 genes with all relevant UniProtKB entries.

**Clean\_data.py:**

* Takes in excel file: uniprot\_output.xlsx from link\_to\_uniprot.py and removes any unnecessary UniProtKB entries.
* Outputs two files: look\_over.xlsx and full\_table.xlsx.
* look\_over.xlsx is a file that holds all genes where a single UniProtKB entry could not be isolates. (Currently 29)
* full\_table.xlsx is a table of all 19,411 genes, including those in look\_over.xlsx, where most UniProtKB entries have been removed to leave 1 GENCODE gene == 1 UniProtKB entry.

**Link\_to\_fasta:**

* Downloads the newest version of GENCODE’s Protein-coding transcript translation sequences FASTA file and unzips it into gencode.v46.pc\_translations.fa if file is not present.
* Takes in full\_table.xlsx from clean\_data.py and links the UniProtKB ENSP numbers with the GENCODE FASTA file to get GENCODE CDS length. Gene symbols are used when a ENSP is not present.
* Outputs final.xlsx, which is the most complete table.

RUN\_ME.py:

* Runs all scripts in order: protein\_list\_builder.py > link\_to\_uniprot.py > clean\_data.py > link\_to\_fasta.