Kenneth Westerman directed study with Larry Parnell starting Jan. 2015

**Objective #1**

Keyword search of the GO database for all GOs relating to four families:

1. Protein kinases
2. GPCRs
3. Ion channels
4. Nuclear receptors

GO enrichment tool to search for GOs enriched in the list of genes from the CM GxE set

Used R to trim list of all GOs from the 4 groups down to only those enriched (p<0.05) in CM GxE set

* Source for CM GxE set: <http://www.biodatamining.org/content/7/1/21>

Repeat this process using a GxE set with exercise- and lipid-based interactions removed

* Keywords removed: physical, SFA, MUFA, PUFA, fat, pequi, olive, fish oil

**Objective #2**

Searched Pfam database for domains relevant to the 4 target families

**Objective #3**

Parsed GxE list for only those impacting triglycerides

Used Drug Gene Interaction database (dgidb) to generate list of 893 drugs targeting genes from the triglyceride-impacting GxE list

* Results included some chemical names that weren’t in the CHEMBL database or weren’t easily searchable and caused problems for downloading the dataset into R, so they were manually removed (\*this part of the workflow must be fixed for the general workflow\*)
* To attempt to remove “unsearchable” chemical names, rows with drug names including “-“ or “ “ (a.k.a. spaces) were removed from the list

Downloaded Chembl database into MySQL and queried to retrieve SMILES structures for each drug of interest

Downloaded FooDB database into MySQL and retrieved SMILES structures for all compounds

Converted SMILES structures (for drugs and for food compounds) into SDF format using OpenBabel

* According to ChemmineR feedback, a small number of SDF files from the OpenBabel conversion were not viable (23090 viable out of 23172)
* 49 compounds failed to return atom pairs

Manually combined drug and food SDFs into one file

R script using ChemmineR package:

* Input: file with all drug and food SDFs and file with drug names and structures
* Output: file with drug names, structures, chembl IDs, and all similar food compounds (at chosen Tanimoto threshold)

**Miscellaneous**

John Overington (from ChEMBL) talk w/ slide referencing the 4 target families of 70% of drugs

* <http://www.onhelix.com/Presentations/John_Overington.pdf>

Compiled list of proteins and associated diseases from UniProt database

CHEMnetBASE Dictionary of Food Compounds – database of compounds found in food

* <http://dfc.chemnetbase.com/dictionary-search.do?method=view&id=11030667&si>=

Spent a significant amount of time learning basic Python programming and how to perform queries in MySQL

Hypothetical Workflow #1

1. Parse GxE list for genes impacting phenotype of choice
2. Use Drug Gene Interaction database to list all drugs targeting genes of interest
3. Use ChEMBL (or drugbank?) to associate structures/CAS/SMILES to drugs in list
4. Use PubChem search (compounds, similarity (substructure?)) to associate chemical notation with similar compounds (search by similar vs. substructure?)
5. Compare this list of compounds with a database of food compounds using CHEMnetBASE DoFC
6. Link these food compounds with foods using PhytoHub, NutriChem, or other
7. Use FF data to perform GxE association tests

* Is there a way to combine steps 5&6?
* Hypothetical Workflow #2: Same, but start by using list of 620 FDA-approved drug targets (why is this different than simply searching chemical structure of all possible drugs?)