Data wrangling: (Shared)

For each counts table:

* Convert ENSG to HGNC names, ignoring ENSGs without a corresponding HGNC name
  + Sum up duplicate HGNCs such that each HGNC symbol is unique in the table
* Normalize counts to counts per million
* Final table should be patients (row) x genes (col)
  + As separate table or as a column add in True/False for AD status

Results in 3 tables representing AD data from different clinical sites (possibly consider additional datasets for further validation)

Machine learning (describe for Vineet’s class)

Choose a favorite classifier (random forrest? SVM? Bayes?), SciKitLearn has a bunch of off the shelf options, for simplicity lets just choose one

Approach 1: dump data into classifier, see how well it predicts AD

Approach 2: somehow engineer features (gotta be some python package that does this lol), then use the engineered features to predict AD

* Determine which features are most predictive of AD, compare genes contributing to a given feature with MSigDB to add “biological relevance” to classification
  + Exclude the annotated alzheimers gene sets (or at least ignore them here…)
* Maybe if there is a clear connection between a highly-predictive feature and a gene set we can discuss briefly the connection between that pathway and AD if known (adds to biological relevance of our ML)

Network comparison (describe for Trey’s class)

From Vineet’s project we will have a 3-layer network: genes->features->phenotype and we also have some idea of which features are predictive of AD phenotype.

Thought 1: if we somehow build a network representation for each dataset independently, maybe we can look for similar subgraph structures

Thought 2: within the 3-layer network there is some clustering of genes grouped together into highly predictive features (“highly predictive” being relative…), so use additional interaction data to assess if these genes interact with each other in a functional or structural manner.

* Trey seemed to lean towards identifying novel gene/protein complex relevant to AD so maybe use PPI networks?

Thought 3: if we generate networks to represent a whole dataset do we lose valuable patient-specific data? Like the smoothie metaphor for bulk RNAseq are per-patient networks more informative/useful?