read in each annotated.gd file; as a pd.df or as a dict

for each df:

parse out the freqs

parse out the position

if intergenic; throw it away

elif 3rd column is a .; throw it away

optional step: ignore synonymous mutations

if it’s a nonsynonymous or frameshift, keep it

parse out the locus tag

if there are multiple entries at one locus tag:

add up the frequencies

#now for each sample you should have a dict or df, with keys for locus tag; frequency; and a confirmation that the mutation is genic and not synonymous.

#now build a output dict or df

get the pooled number of locus\_tags that have mutations #this will be the number of columns in the df

get each locus\_tag’s name in a taglist

initiate df with rows = number of samples, columns = pooled\_length + 1 (sample name); make sure the columns are named for the values in the taglist

#add each row:

for each dict:

add a row where the pooled frequency values correspond to the column title