This Rscript was uses BiomaRt to create an annotation\_tsv for S. scrofa (pig)

library(plyr)

library(biomaRt)

mart = useMart("ENSEMBL\_MART\_ENSEMBL", host="www.ensembl.org") #Change host here to use older version

pigMart<- useDataset("sscrofa\_gene\_ensembl",mart=mart)

#Keeping track of current version used

currentMarts <- listMarts()

currentMarts$version[currentMarts$biomart == "ENSEMBL\_MART\_ENSEMBL"]

#Ensembl Genes 97

#Gets attributes using transcript id from biomaRt

attr<- getBM(attributes=c("ensembl\_gene\_id", "entrezgene\_id", "external\_gene\_name", "description"), mart=pigMart)

rename\_col\_by\_name <- function (df, old, new) {

#https://stackoverflow.com/a/16490387

names(df)[names(df) == old] <- new

return(df)

}

#Standardize column name of unique ID to 'gene\_id' - This corresponds to the unique ID used in GTF

attr <- rename\_col\_by\_name(attr, 'ensembl\_gene\_id', 'gene\_id')

#For queries with multiple entrez gene IDs, return these IDs in a single cell, as a string with commas separating the individual IDs

collapsed.entrezIDs <- ddply(attr, .(gene\_id), summarize, entrezgene\_id=paste(entrezgene\_id, collapse=","))

#Sanity check. Gene symbols should be consistent, without conflicting results

test1 <- ddply(attr, .(gene\_id), summarize, external\_gene\_name=paste(external\_gene\_name, collapse=","))

View(test1[grepl(",", test1$entrezgene\_id),])

#Another way of testing automagically

test1.split <- strsplit(test1$external\_gene\_name[grepl(",", test1$external\_gene\_name)], ",")

unique(sapply(test1.split, function(x) {length(unique(x))})) == c(1)

#Sanity check. Descriptions are also consistent, no conflicting results

test2 <- ddply(attr, .(gene\_id), summarize, description=paste(description, collapse="@"))

test2.split <- strsplit(test2$description[grepl("@", test2$description)], "@")

unique(sapply(test2.split, function(x) {length(unique(x))})) == c(1)

#Since entrez gene IDs are the only ones with conflicting entries, can merge our collapsed entrezIDs from before with the attr table

mapping\_table <- merge(collapsed.entrezIDs, attr, on="gene\_id", all.x=T)

write.table(mapping\_table, "Sscrofa11.1\_annotation.tsv", quote=F, sep="\t", row.names=F)