

Acquiring identifiers for autosomal genes

Warren Anderson
Joon Yuhl Soh
Mete Civelek

April 2, 2020

This guide provides code and documentation of analyses from Anderson et al., 2020, *Sex differences in human adipose tissue gene expression and genetic regulation involve adipogenesis*

Contents

1	Overview	2
2	Aquisition of GENCODE data	2

1 Overview

The following code was used to get a list of autosomal genes from the GENCODE annotation.

2 Aquisition of GENCODE data

The annotation data were downloaded from the GTEx website as described here.

```
# website
https://www.gtexportal.org/home/datasets

# download annotation, GTEx Analysis V8, Reference
gencode.v26.GRCh38.genes.gtf
mv gencode.v26.GRCh38.genes.gtf gencode0.txt

# isolate columns of interest and remove header
tail -n +7 gencode0.txt | cut -f1,3,4,5,9 | awk '($2 == "transcript") || ($2 == "gene")' > gencode.txt
```

The following R code was used to isolate the gene name – ensembl mappings and non-autosomal gene identifiers.

```
library(dplyr)

# get gene annotation information
gene_ann0 = read.table("gencode.txt",header=F,sep="\t",stringsAsFactors=F)

# get gene names
geneNames = sapply(gene_ann0[,5],function(x){
  out1 = strsplit(x," ")[[1]]
  ind = grep("gene_name",out1)
  out2 = out1[ind]
  out3 = strsplit(out1[ind+1], ";")[[1]][1]
  out = c(out2,out3)
  return(out)
}) %>% t
rownames(geneNames) = c(1:nrow(geneNames))
unique(geneNames[,1])
gene_info = cbind(gene_ann0[,c(1,3,4)], geneNames[,2]) %>% as.data.frame
names(gene_info) = c("chr","coor1","coord2","id")

# get list of non-autosomal genes
ind_X = which(gene_info$chr == "chrX")
ind_Y = which(gene_info$chr == "chrY")
ind_M = which(gene_info$chr == "chrM")
ind_non_auto = c(ind_X,ind_Y,ind_M)
non_auto_genes = gene_info$id[ind_non_auto] %>% unique %>% as.character

# get gene and/id annotation
geneID = sapply(gene_ann0[,5],function(x){
  out1 = strsplit(x," ")[[1]]
  ind = grep("gene_id",out1)
  out2 = out1[ind]
  out3 = strsplit(out1[ind+1], ";")[[1]][1]
  out = c(out2,out3)
  return(out)
}) %>% t
rownames(geneID) = c(1:nrow(geneID))
unique(geneID[,1])
gene_map = cbind(geneID[,2],geneNames[,2]) %>% as.data.frame
```

```
names(gene_map) = c("gene_id", "gene_name")

# write output
write.table(non_auto_genes, "genes_XYM.txt", col.names=F, row.names=F, quote=F, sep="\t")
write.table(unique(gene_map), "gencode_gene_map.txt", col.names=T, row.names=F, quote=F, sep="\t")
```