Carl Zimmer Chromosome 3 SNP Burden Analysis

Jeff Mandell, Guannan Gong, Neng Wan, and Lukas Fuentes

Read in annotated SNP data.

This analysis will be run with two sets of SNP data. The first excludes intergenic and intronic SNPs, but keeps splice-site, but keeps UTR SNPs and SNPS within a neighborhood upstream/downstream of the exons, while the second set only includes SNPs annotated as exon/splice-site.

```
data = read.table("chr3_annotated_all.txt", sep = "\t", header = T, na.strings = ".",
    stringsAsFactors = F, quote = "")
# data = read.table('chr3_annotated_exon-splice.txt', sep = '\t', header=T,
# na.strings = '.', stringsAsFactors = F, quote = '')
```

Print gene lists

Next, print out the ten genes with highest mutational burden. Also print out lists of genes with the most non-synonymous mutations and with the most uncommon non-synonymous mutations ($< 5\%$ max popualtion frequency).

```
count_table = table(data$hgnc_symbol)
print(sort(count_table, decreasing = T)[1:10])
```

```
##
##   MUC4   MOBP ALS2CL  MUC20 CCDC50   CHL1 PHLDB2  GPD1L SLC9C1   VPS8
##     32     30     27     24     21     21     21     20     20     20
```

```
# Repeat, with synonymous variants excluded
nonsynonymous = data[is.na(data$ExonicFunc.ensGene) | data$ExonicFunc.ensGene !=
    "synonymous SNV", ]
nonsyn_table = table(nonsynonymous$hgnc_symbol)
print(sort(nonsyn_table, decreasing = T)[1:10])
```

```
##
##   MOBP   MUC4 ALS2CL  MUC20   CHL1 PHLDB2  GPD1L OR5H4P CCDC50 SLC9C1
##     29     26     24     21     20     20     19     19     18     18
```

```
# Repeat using uncommon variants only (max population frequency < 5% or
# unknown)
nonsynonymous_uncommon = nonsynonymous[is.na(nonsynonymous$PopFreqMax) | nonsynonymous$PopFreqMax <
    0.05, ]
nonsyn_uncommon_table = table(nonsynonymous_uncommon$hgnc_symbol)
print(sort(nonsyn_uncommon_table, decreasing = T)[1:10])
```

```
##
##     SENP2    CCDC66 EEF1A1P24 NPHP3-AS1     PLCL2    SLC6A1  ARMC10P1
##         5         4         3         3         3         3         2
##   B4GALT4      ECE2   ENPP7P3
##         2         2         2
```

Re-run with gene length normalization

Print the lists of genes with the highest variant burden (all variants, non-synonymous, and uncommon non-synonymous).

```
  ens_hcgna = data$Gene.ensGene
  names(ens_hcgna) = data$hgnc_symbol

  lengths = data$gene_length
  names(lengths) = data$Gene.ensGene
  normed = count_table / lengths[ens_hcgna[names(count_table)]]
  print(sort(normed, decreasing = T)[1:10])
```

```
##
##       OR5H6       OR5H8      OR5H15       ALG1L       PYDC2      SLC9C1
## 0.006683375 0.005741627 0.004068348 0.003558719 0.003401361 0.003089354
##        RTP2   LINC01100       EBLN2      CHST13
## 0.002971768 0.002702703 0.002382370 0.002235886
```

```
  nonsynonymous_normed = nonsyn_table / lengths[ens_hcgna[names(nonsyn_table)]]
  print(sort(nonsynonymous_normed, decreasing = T)[1:10])
```

```
##
##       OR5H8       OR5H6       PYDC2      OR5H15   LINC01100      SLC9C1
## 0.005741627 0.004177109 0.003401361 0.003254679 0.002702703 0.002614068
##       ALG1L        RTP2       EBLN2      MAGEF1
## 0.002372479 0.002228826 0.001786778 0.001222494
```

```
  nonsyn_uncommon_normed = nonsyn_uncommon_table / lengths[ens_hcgna[names(nonsyn_uncommon_table)]]
  print(sort(nonsyn_uncommon_normed, decreasing = T)[1:10])
```

```
##
##       PRSS45       ACTRT3        NRROS        PLCD1         GLB1
## 0.0009523810 0.0005984440 0.0003445899 0.0002318034 0.0002306805
##         UBA7        USP19         MFN1        XIRP1        RIOX2
## 0.0002197802 0.0001744896 0.0001682086 0.0001546312 0.0001453488
```