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)
```

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])
```

```
##
##  CCDC50  PLXND1 CRIP1P1    MUC4    MOBP  ALS2CL   MUC20   SHOX2    SNTN
##      42      38      32      32      30      27      24      22      22
##    CHL1
##      21
```

```
# 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])
```

```
##
##  CCDC50 CRIP1P1    MOBP    MUC4  ALS2CL  PLXND1   SHOX2    SNTN   MUC20
##      36      32      29      26      24      24      22      22      21
##    CHL1
##      20
```

```
# 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])
```

```
##
##     CRIP1P1   RNU6-822P   RNU6-557P   RN7SKP298     MIR4790      MIR563
##  0.17297297  0.08490566  0.08181818  0.06382979  0.06329114  0.06329114
##     ENPP7P3     MIR5186 RNU6ATAC15P  RNU6-1236P
##  0.06224066  0.05833333  0.05785124  0.05050505
```

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

```
##
##     CRIP1P1   RNU6-822P   RNU6-557P   RN7SKP298     MIR4790      MIR563
##  0.17297297  0.08490566  0.08181818  0.06382979  0.06329114  0.06329114
##     ENPP7P3     MIR5186 RNU6ATAC15P  RNU6-1236P
##  0.06224066  0.05833333  0.05785124  0.05050505
```

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

```
##
##    MIR548AB   RN7SKP298   RNU6-138P     ENPP7P3 RNU6ATAC26P     CRIP1P2
## 0.011904762 0.010638298 0.009345794 0.008298755 0.007936508 0.004385965
##   LINC02041   LINC02013    RN7SKP61       IQCF2
## 0.004065041 0.003802281 0.003389831 0.003184713
```