Steps for Data Preparation and ESKeDiT input

Simone Longo

The following shows the steps used in the command line to obtain regions that will be used to train the model. The goal is to capture a neutral mutation rate so the regions of interest will be free of all protein coding sequences, which are observed to be under a high degree of constraint.

The starting data is from the ENSEMBL Biomart using the following parameters:

- Using Ensembl Genes 100 dataset with GRCh38.p13 Human genes
- Restricted to include only autosomes (chr 1-22) AND only genes annotated as being "protein_coding" by ENSEMBL
- From the "Structures" option:
  - Chromosome/scaffold name
  - Exon region start (bp)
  - Exon region end (bp)
  - Exon stable ID
  - Exon rank in transcript
  - Strand
  - Gene name

In [1]:
```
!head -20 /Users/simonelongo/Documents/QuinlanLabFiles/ESKeDiT/notebooks/notebook_resources/ensembl_protein_coding_autosome_exons.txt
```

```
Chromosome/scaffold name        Exon region start (bp)  Exon region end (bp)    Exon stable ID  Exon rank in transcript  Strand  Gene name
1       201468341       201469188       ENSE00001444159 1       -1      PHLDA3
1       201464278       201466178       ENSE00001444157 2       -1      PHLDA3
1       201468345       201468581       ENSE00001824558 1       -1      PHLDA3
1       201467486       201467583       ENSE00002409826 2       -1      PHLDA3
1       201465853       201466178       ENSE00001901596 3       -1      PHLDA3
1       201466311       201466657       ENSE00001820952 1       -1      PHLDA3
1       201465944       201466178       ENSE00001936305 2       -1      PHLDA3
1       201469171       201469237       ENSE00001444155 1       -1      PHLDA3
1       201468341       201468948       ENSE00001444153 2       -1      PHLDA3
1       201465958       201466178       ENSE00001444152 3       -1      PHLDA3
1       207752054       207752309       ENSE00003898531 1       1       CD46
1       207757014       207757202       ENSE00003465327 2       1       CD46
1       207757540       207757642       ENSE00003504276 3       1       CD46
1       207759639       207759724       ENSE00003509765 4       1       CD46
1       207761249       207761446       ENSE00001167611 5       1       CD46
1       207767013       207767195       ENSE00001167601 6       1       CD46
1       207767779       207767823       ENSE00000792032 7       1       CD46
1       207770321       207770362       ENSE00003558309 8       1       CD46
1       207783292       207783330       ENSE00003582519 9       1       CD46
```

Convert to BED format

```
awk -v OFS='\t' \
'{ if ($6 > 0) { $6 = "+"} else { $6 = "-"} print "chr"$0 }' \
ensembl_protein_coding_autosome_exons.txt |\
tail -n +2 > ensembl_protein_coding_22june2020.bed



head -n 20 ensembl_protein_coding_22june2020.bed

chr1    201468341   201469188   ENSE00001444159 1   -   PHLDA3
chr1    201464278   201466178   ENSE00001444157 2   -   PHLDA3
chr1    201468345   201468581   ENSE00001824558 1   -   PHLDA3
chr1    201467486   201467583   ENSE00002409826 2   -   PHLDA3
chr1    201465853   201466178   ENSE00001901596 3   -   PHLDA3
chr1    201466311   201466657   ENSE00001820952 1   -   PHLDA3
chr1    201465944   201466178   ENSE00001936305 2   -   PHLDA3
chr1    201469171   201469237   ENSE00001444155 1   -   PHLDA3
chr1    201468341   201468948   ENSE00001444153 2   -   PHLDA3
chr1    201465958   201466178   ENSE00001444152 3   -   PHLDA3
chr1    207752054   207752309   ENSE00003898531 1   +   CD46
chr1    207757014   207757202   ENSE00003465327 2   +   CD46
chr1    207757540   207757642   ENSE00003504276 3   +   CD46
chr1    207759639   207759724   ENSE00003509765 4   +   CD46
chr1    207761249   207761446   ENSE00001167611 5   +   CD46
chr1    207767013   207767195   ENSE00001167601 6   +   CD46
chr1    207767779   207767823   ENSE00000792032 7   +   CD46
chr1    207770321   207770362   ENSE00003558309 8   +   CD46
chr1    207783292   207783330   ENSE00003582519 9   +   CD46
chr1    207785071   207785106   ENSE00003520315 10  +   CD46
```

Flatten BED file and Exclude Coding Regions

The BED must first be compressed and indexed by `tabix` for `bedtools` to function properly. This can be done with a function called `bedprep`.

Download `bedprep` here: https://github.com/SpacemanSpiff7/bedprep

The default options for `bedtools merge` for Version v2.29.2 are used.

```
bedprep ensembl_protein_coding_22june2020.bed
bedtools merge -i ensembl_protein_coding_22june2020_sorted.bed.gz > flat_ensembl_pc_22june2020.bed

head -20 flat_ensembl_pc_22june2020.bed

chr1    65419   65433
chr1    65520   65573
chr1    69037   71585
chr1    450703  451697
chr1    685679  686673
chr1    923928  924948
chr1    925150  925189
chr1    925731  925800
chr1    925922  926013
chr1    930155  930336
chr1    931039  931089
chr1    935772  935896
chr1    939040  939129
chr1    939272  939460
chr1    940346  940462
chr1    941076  941306
chr1    942103  943058
chr1    943253  943377
chr1    943698  944800
chr1    945042  945146
```

Bedtools requires a genome file to take the inverse of these regions. Here, I use `grch38.genome` and `-L` flag to limit the output to chromosomes contained in the input.

```
bedtools complement -L -i flat_ensembl_pc_22june2020.bed -g grch38.genome > pc_exon_complement_22june2020.bed

# Now we have the inverse
head -20 pc_exon_complement_22june2020.bed

chr1    0   65419
chr1    65433   65520
chr1    65573   69037
chr1    71585   450703
chr1    451697  685679
chr1    686673  923928
chr1    924948  925150
chr1    925189  925731
chr1    925800  925922
chr1    926013  930155
chr1    930336  931039
chr1    931089  935772
chr1    935896  939040
chr1    939129  939272
chr1    939460  940346
chr1    940462  941076
chr1    941306  942103
chr1    943058  943253
chr1    943377  943698
chr1    944800  945042
```

We can confirm the regions we have obtained by using IGV.

IGV shot of all genes (top track), protein coding genes (middle track), and protein coding complements (bottom track)

This BED file is now ready to be used as an input for model training. To reiterate, this BED file contains regions that don't explicitly code for proteins.

This removes exons and UTRs.

Using ESKeDiT 2.0.0

ESKeDiT is used to train the model.

Clone repository:

```
git clone https://github.com/SpacemanSpiff7/ESKeDiT
cd ESKeDiT
```

Declare variables to use and run.

```
nprocs=1
bed_path='/Users/simonelongo/Documents/QuinlanLabFiles/ESKeDiT/notebooks/notebook_resources/pc_exon_complement_22june2020.bed'
vcf_path='/Users/simonelongo/too_big_for_icloud/gnomAD_v3/gnomad.genomes.r3.0.sites.vcf.bgz'
fasta_path='/Users/simonelongo/too_big_for_icloud/ref_genome/hg38/hg38.fa'
meth_vcf_path='/Users/simonelongo/too_big_for_icloud/gnomAD_v3/gnomadv3_methylation_2.vcf.bgz'

python3 eskedit_main.py -f $fasta_path -v $vcf_path -b $bed_path -m $meth_vcf_path -@ $nprocs
```

In [ ]:
```
```