Assignment 3

Theo-60985751, Hannah-88788427, Jingxuan-85741635

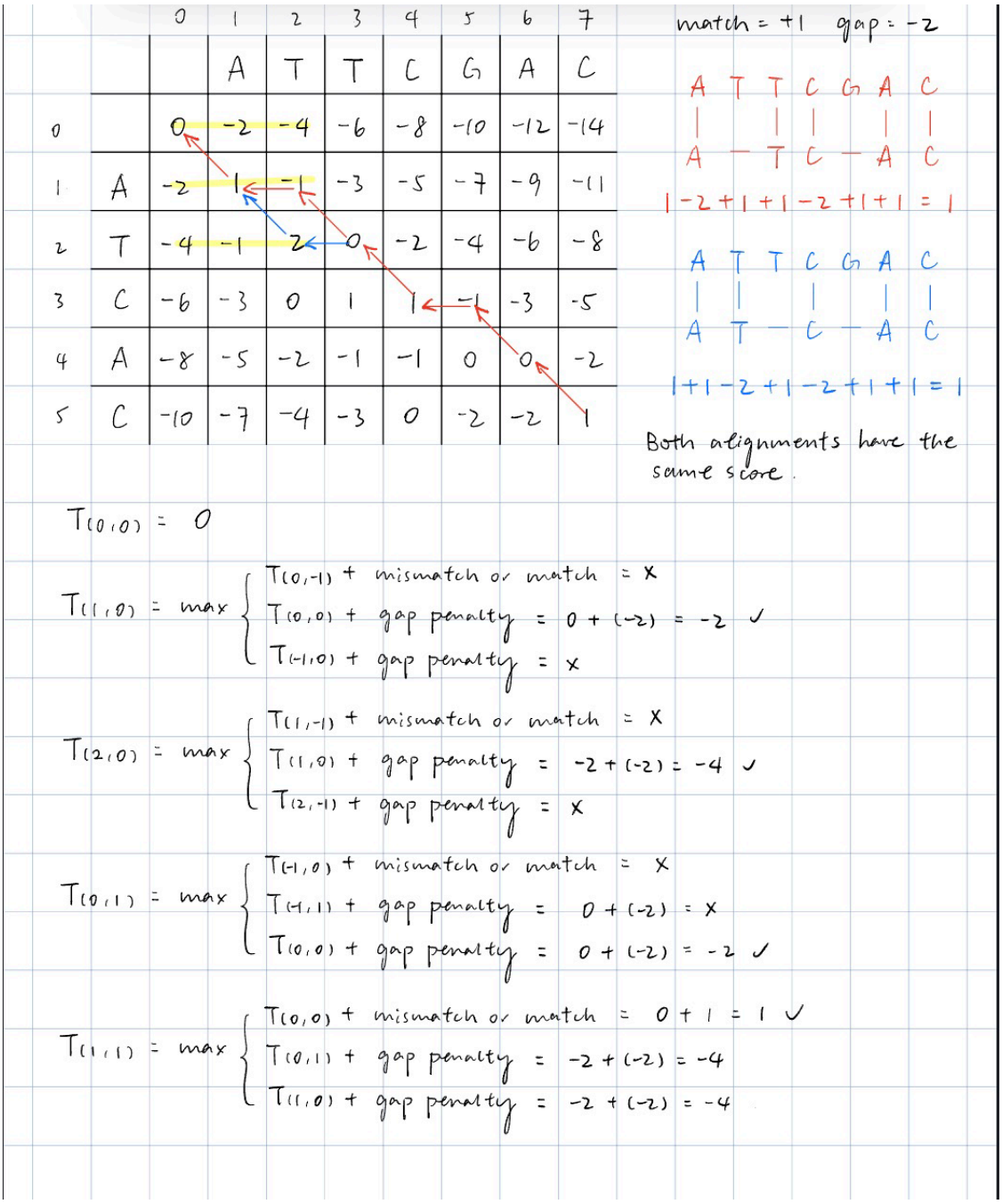
2023-11-18

Q1. Sequencing technologies

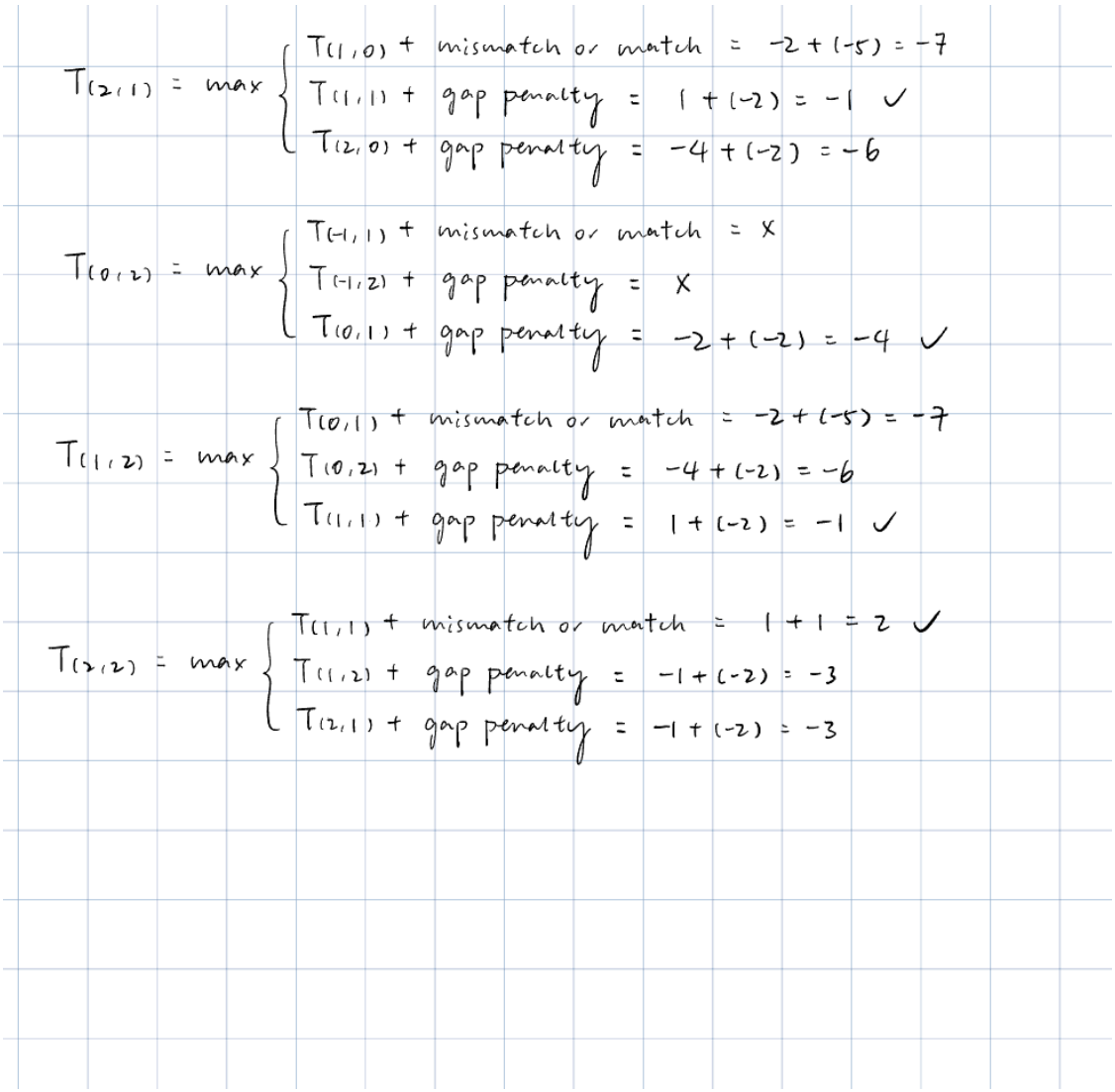
Areas of the genome with high GC content are harder to sequence because these regions coil up to themselves and require energy to separate them into straight strands that can finally be PCR-ed/sequenced. Source: https://www.neb.com/en/nebinspired-blog/four-tips-for-pcr-amplification -of-gc-rich-sequences#:~:text=Why%20can%20these%20regions%20be,break%20the%20three%20hydrogen%20bonds.

Q2. Global alignment exercise

```
knitr::include_graphics("image1.png")
```




```
knitr::include_graphics("image2.png")
```



Q3. Looking at the Metadata of an alignment (SAM) file

Q3.1

```
data_metadat_sam = read.csv("single_cell_RNA_seq_bam.sam", nrows=73,
sep="\t", header=FALSE,
fill=TRUE)
```

SN is reference sequence name, and LN is reference sequence length with range [1, 2^31 – 1]

Q3.2

```
cat("Length of our X chromosome alignment: ", data_metadat_sam[22,3])

## Length of our X chromosome alignment:  LN:171031299
```

Q4. Looking at the Reads of an alignment (SAM) file

Q4.1

```
sam <- read.csv("single_cell_RNA_seq_bam.sam", sep="\t", header=FALSE,
comment.char="@", col.names = paste0("V",seq_len(30)), fill=TRUE)
sam <- sam[paste0("V",seq_len(11))]

cat("Number of reads in this BAM file:", nrow(sam))

## Number of reads in this BAM file: 146346
```

Q4.2

```
tenth_row <- as.character(sam[10, ])
print(tenth_row)

##  [1] "NS500668:199:HV73CBGX2:1:11203:20546:3351"
##  [2] "16"
##  [3] "1"
##  [4] "3365976"
##  [5] "255"
##  [6] "58M"
##  [7] "*"
##  [8] "0"
##  [9] "0"
## [10] "AATCAAAAAGGGGGCTGTCAGTAGGATGATATAAGATATAGATGTAGTTTATCTCCTA"
## [11] "EEEEEEEEEEA//AAAAEEEEEE/AEEAEAEEEEEEEEEEEEEAAEE///EEEAAA6A"
```

The chromosome to which the read was aligned is represented by the “RNAME” field, which corresponds to column V3 in the dataframe. The “QUAL” field in BAM corresponds to the column V11 in the dataframe. base.

Q4.3

```
sam$X_allign = sam$V3 == 'X'
cat("There are", sum(sam$X_allign), "reads that alligns to chromosome X")

## There are 5999 reads that alligns to chromosome X
```

Q4.4

```
phred33toQ = function(ascii){
  return( as.numeric(charToRaw(ascii)) - 33)
}

sam_xchromo = sam[which(sam$V3 == "X"),]



total_mean_quality_xchromo = 0
for (quality in sam_xchromo$V11){
  # print(mean(phred33toQ(quality)))
  total_mean_quality_xchromo = total_mean_quality_xchromo +
```

```
mean(phred33toQ(quality))
}

mean_quality_xchromo = total_mean_quality_xchromo/nrow(sam_xchromo)

cat("The mean base quality for reads aligning to chromosome X is",
mean_quality_xchromo)

## The mean base quality for reads aligning to chromosome X is 32.72349
```

Q4.5

including libraries

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ lubridate 1.9.3     ✓ tibble    3.2.1
## ✓ purrr     1.0.2     ✓ tidyr     1.3.0
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.2.3
```

separating the sam file into smaller ones

mini_sam1

```
mini_sam1 = sam[1:50000,]

df_quality_1 = matrix(nrow = 58, ncol = 1)
read_number = 1
for (read in mini_sam1$V11){
  quality = phred33toQ(read)
  col_name = paste0("read", read_number)
  df_quality_1 = cbind(df_quality_1, quality)
  read_number = read_number + 1
}

df_quality_1 = df_quality_1[,-c(1)]
# df_quality_1 = t(df_quality_1)
# colnames(df_quality_1) = c(1:58)
# df_quality_1 = stack(as.data.frame(df_quality_1))
#
# ggplot(data = df_quality_1)+
#   geom_boxplot(aes(x=ind, y = values))
```

mini_sam2

```
mini_sam2 = sam[50001:100000,]

df_quality_2 = matrix(nrow = 58, ncol = 1)
read_number = 1001
for (read in mini_sam2$V11){
  quality = phred33toQ(read)
  col_name = paste0("read", read_number)
  df_quality_2 = cbind(df_quality_2, quality)
  read_number = read_number + 1
}
df_quality_2 = df_quality_2[,-c(1)]
```

mini_sam3

```
mini_sam3 = sam[100001:146346,]

df_quality_3 = matrix(nrow = 58, ncol = 1)
read_number = 2001
for (read in mini_sam3$V11){
  quality = phred33toQ(read)
  col_name = paste0("read", read_number)
  df_quality_3 = cbind(df_quality_3, quality)
  read_number = read_number + 1
```

```
}
df_quality_3 = df_quality_3[,-c(1)]
```

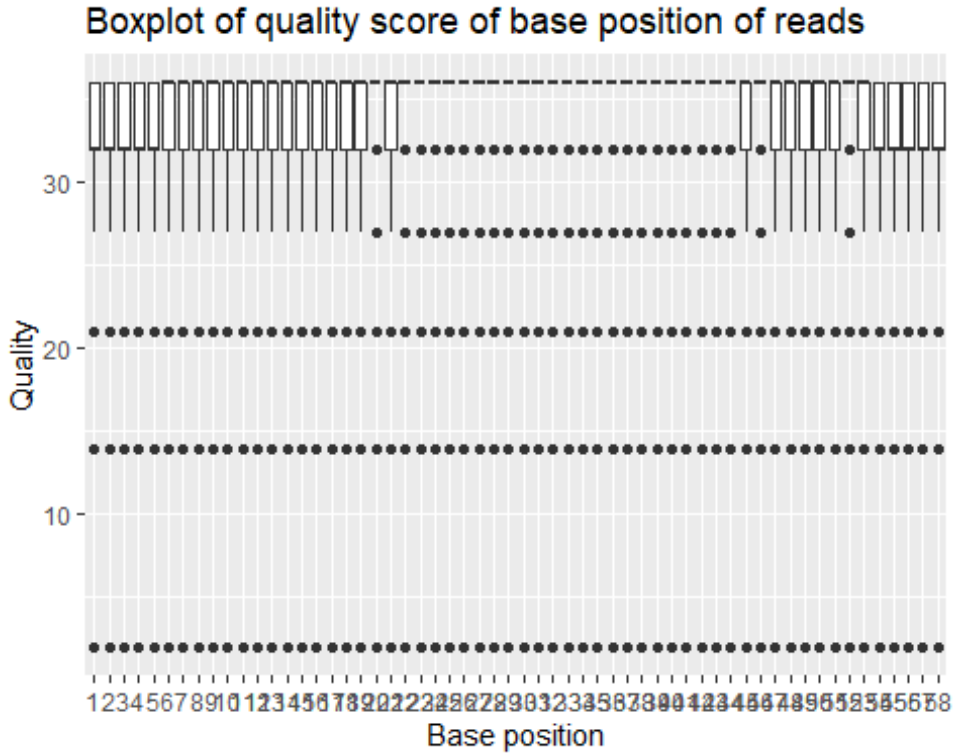
combining smaller sam quality scores into one big file

```
df_quality = cbind(df_quality_1, df_quality_2)
df_quality = cbind(df_quality, df_quality_3)

df_quality = t(df_quality)
colnames(df_quality) = c(1:58)
df_quality = stack(as.data.frame(df_quality))
```

plotting the barplot

```
ggplot(data = df_quality)+
  geom_boxplot(aes(x=ind, y = values)) +
  ggtitle("Boxplot of quality score of base position of reads") +
  xlab("Base position") +
  ylab("Quality")
```



The Base Quality varies at the beginning and end of each read, with the mean being toward the lower end (30). For the middle part, the base Quality stays consistently high.

Q4.6

Column 4 contains the leftmost mapping position of the read

Q4.7

```
sam$hspa8_allign = sam$V4 > 40801273 & sam$V4 < 40805199

cat("There are", sum(sam$hspa8_allign), "reads that have their leftmost
mapping
    position aligned with the cordinate for Hspa8 protein")

## There are 134 reads that have their leftmost mapping
##     position aligned with the cordinate for Hspa8 protein
```

Q4.8

```
sam$MAPQ_50less = sam$V5 < 50

cat("There are",sum(sam$MAPQ_50less), "reads that have mapping quality score
less than 50")

## There are 61527 reads that have mapping quality score less than 50
```

Q4.9

```
mean_MAPQ_50less = mean(sam[sam$MAPQ_50less == TRUE,]$V5)
cat("Mean mapping quality of the reads that has MAPQ less than 50:",
mean_MAPQ_50less)

## Mean mapping quality of the reads that has MAPQ less than 50: 0.2418125
```

Q4.10

Q5. Investigating the Variants

Q5.1

```
vcf_con <- file("RNA_seq_annotated_variants.vcf", open="r")
vcf_file <- readLines(vcf_con)
close(vcf_con)
vcf <- data.frame(vcf_file)
header <- vcf[grepl("##", vcf$vcf_file), ]
# factor(header)
variants <- read.csv("RNA_seq_annotated_variants.vcf", skip=length(header),
header=TRUE, sep="\t")
```

Reference and alternative allele of the first variant

```
ref_1 = variants$REF[1]
alt_1 = variants$ALT[1]

cat("Reference allele base of the first variant: ", ref_1,"\n")

## Reference allele base of the first variant:  G

cat("Alternative allele base of the first variant calledby STrelka: ", alt_1)
```

```
## Alternative allele base of the first variant calledby STrelka:  A
```

Q5.2

```
info_1 = as.character(variants$INFO[1])

splitted_1 = strsplit(info_1,";")
ANN_1 = splitted_1[[1]][3]
cat("The ANN info is:", ANN_1)

## The ANN info is:
ANN=A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000
0088585.9|protein_coding|2/21|c.-
133+17418G>A||||||,A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcr
ipt|ENSMUST00000177608.7|protein_coding|2/21|c.-
133+17418G>A||||||,A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcr
ipt|ENSMUST00000180062.7|protein_coding|1/20|c.-132-
39973G>A||||||,A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|
ENSMUST00000186051.6|protein_coding|2/21|c.-
133+17418G>A||||||,A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcr
ipt|ENSMUST00000187376.6|retained_intron|2/11|n.420+17418G>A||||||,A|intron_v
ariant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST00000186405.6|prot
ein_coding|1/2|c.-
133+17418G>A||||||WARNING_TRANSCRIPT_INCOMPLETE,A|intron_variant|MODIFIER|Sul
f1|ENSMUSG00000016918|transcript|ENSMUST00000189541.6|protein_coding|2/3|c.-
133+17418G>A|||||WARNING_TRANSCRIPT_INCOMPLETE
```

Q5.3

```
detailANN_1 = strsplit(ANN_1,',')
detailANN_1

## [[1]]
## [1]
"ANN=A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST000
00088585.9|protein_coding|2/21|c.-133+17418G>A||||||"
## [2]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000017
7608.7|protein_coding|2/21|c.-133+17418G>A||||||"
## [3]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000018
0062.7|protein_coding|1/20|c.-132-39973G>A||||||"
## [4]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000018
6051.6|protein_coding|2/21|c.-133+17418G>A||||||"
## [5]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000018
7376.6|retained_intron|2/11|n.420+17418G>A||||||"
## [6]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000018
6405.6|protein_coding|1/2|c.-133+17418G>A||||||WARNING_TRANSCRIPT_INCOMPLETE"
## [7]
```

```
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000018
9541.6|protein_coding|2/3|c.-133+17418G>A||||||WARNING_TRANSCRIPT_INCOMPLETE"
```

We know that the variant is most likely a modifier variant, located on the Sulf1 gene, transcripted, and within the protein coding region. The annotation also has other information about the position on the cDNA, feature ID, and more.

Q5.4

Repeating with varient on line 683:

```
ref_683 = variants$REF[683]
alt_683 = variants$ALT[683]

cat("Reference allele base of the first variant: ", ref_683,"\n")

## Reference allele base of the first variant:  ACAGGGG

cat("Alternative allele base of the first variant calledby STrelka: ",
alt_683, "\n")

## Alternative allele base of the first variant calledby STrelka:  A

info_683 = as.character(variants$INFO[683])

splitted_683 = strsplit(info_683,";")
ANN_683 = splitted_1[[1]][3]
cat("The ANN info is:", ANN_683)

## The ANN info is:
ANN=A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000
0088585.9|protein_coding|2/21|c.-
133+17418G>A||||||,A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcr
ipt|ENSMUST00000177608.7|protein_coding|2/21|c.-
133+17418G>A||||||,A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcr
ipt|ENSMUST00000180062.7|protein_coding|1/20|c.-132-
39973G>A||||||,A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|
ENSMUST00000186051.6|protein_coding|2/21|c.-
133+17418G>A||||||,A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcr
ipt|ENSMUST00000187376.6|retained_intron|2/11|n.420+17418G>A||||||,A|intron_v
ariant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST00000186405.6|prot
ein_coding|1/2|c.-
133+17418G>A||||||WARNING_TRANSCRIPT_INCOMPLETE,A|intron_variant|MODIFIER|Sul
f1|ENSMUSG00000016918|transcript|ENSMUST00000189541.6|protein_coding|2/3|c.-
133+17418G>A||||||WARNING_TRANSCRIPT_INCOMPLETE

detailANN_683 = strsplit(ANN_683,',')
detailANN_683

## [[1]]
## [1]
"ANN=A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST000
```

```
00088585.9|protein_coding|2/21|c.-133+17418G>A||||||"
## [2]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000017
7608.7|protein_coding|2/21|c.-133+17418G>A||||||"
## [3]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000018
0062.7|protein_coding|1/20|c.-132-39973G>A||||||"
## [4]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000018
6051.6|protein_coding|2/21|c.-133+17418G>A||||||"
## [5]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000018
7376.6|retained_intron|2/11|n.420+17418G>A||||||"
## [6]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000018
6405.6|protein_coding|1/2|c.-133+17418G>A||||||WARNING_TRANSCRIPT_INCOMPLETE"
## [7]
"A|intron_variant|MODIFIER|Sulf1|ENSMUSG00000016918|transcript|ENSMUST0000018
9541.6|protein_coding|2/3|c.-133+17418G>A||||||WARNING_TRANSCRIPT_INCOMPLETE"
```

The variant would be affecting the Sulf1 gene

Q5.5

```
for (index in 1:nrow(variants)){
  variants$HIGH[index] = grepl("HIGH",strsplit((variants$INFO[index]),";"))
  variants$MODERATE[index] =
grepl("MODERATE",strsplit(as.character(variants$INFO[index]),";"))
  variants$LOW[index] =
grepl("LOW",strsplit(as.character(variants$INFO[index]),";"))
  variants$MODIFIER[index] =
grepl("MODIFIER",strsplit(as.character(variants$INFO[index]),";"))
}

cat("Possible frameshift indels:",sum(variants$HIGH), "\n")

## Possible frameshift indels: 4

cat("Possible nonsynonymous SNVs:",sum(variants$MODERATE) +
sum(variants$MODIFIER) + sum(variants$HIGH), "\n")

## Possible nonsynonymous SNVs: 897

cat("Possible synonymous SNVs:",sum(variants$HIGH), "\n")

## Possible synonymous SNVs: 4
```

Q5.6

A frameshift variant is when there's an indel that affects 3 or multiple of 3 base pairs, thus moving the sequence after the indel up or down a whole codon/codons.

It has a greater effect on the resultant protein compared to a missense varriant, since missense only affects that single amino acid, while frameshift affects everything that comes after it.

Q5.7

```
for (index in 1:nrow(variants)){
  variants$intronic[index] =
grepl("intron_variant",strsplit(as.character(variants$INFO[index]),";"))
  variants$intergenic[index] =
grepl("intergenic_region",strsplit(as.character(variants$INFO[index]),";"))
}


cat("Number of intronic variants: ", sum(variants$intronic) +
sum(variants$intergenic),"\n")

## Number of intronic variants:  606

cat("Intronic/intergeneic variants make up", (sum(variants$intronic) +
sum(variants$intergenic))/nrow(variants), "of all of the variants found in
the VCF file" )

## Intronic/intergeneic variants make up 0.7248804 of all of the variants
found in the VCF file
```

Q5.8

```
variants[variants$HIGH == TRUE,]$INFO

## [1]
"SNVHPOL=3;MQ=255;ANN=G|splice_acceptor_variant&intron_variant|HIGH|Ddx1|ENSM
USG00000037149|transcript|ENSMUST00000071103.8|protein_coding|25/25|c.2093-
2A>C||||||;LOF=(Ddx1|ENSMUSG00000037149|1|1.00)"
## [2]
"SNVHPOL=3;MQ=255;ANN=T|stop_gained&splice_region_variant|HIGH|Rps14|ENSMUSG0
0000024608|transcript|ENSMUST00000025511.9|protein_coding|4/5|c.388G>T|p.Glu1
30*|567/683|388/456|130/151||,T|stop_gained&splice_region_variant|HIGH|Rps14|
ENSMUSG00000024608|transcript|ENSMUST00000122279.1|protein_coding|3/4|c.388G>
T|p.Glu130*|552/667|388/456|130/151||,T|stop_gained&splice_region_variant|HIG
H|Rps14|ENSMUSG00000024608|transcript|ENSMUST00000118551.7|protein_coding|4/5
|c.388G>T|p.Glu130*|491/602|388/456|130/151||,T|stop_gained&splice_region_var
iant|HIGH|Rps14|ENSMUSG00000024608|transcript|ENSMUST00000137400.7|protein_co
ding|4/5|c.388G>T|p.Glu130*|417/444|388/415|130/137||WARNING_TRANSCRIPT_INCOM
PLETE,T|downstream_gene_variant|MODIFIER|Rps14|ENSMUSG00000024608|transcript|
ENSMUST00000142980.1|retained_intron||n.*470G>T|||||470|,T|downstream_gene_va
riant|MODIFIER|Rps14|ENSMUSG00000024608|transcript|ENSMUST00000127568.7|prote
in_coding||c.*30G>T|||||30|WARNING_TRANSCRIPT_INCOMPLETE,T|downstream_gene_va
riant|MODIFIER|Gm8731|ENSMUSG00000080779|transcript|ENSMUST00000118088.1|proc
essed_pseudogene||n.*1094C>A|||||1094|"
## [3]
"CIGAR=1M6D;RU=CAGGGG;REFREP=1;IDREP=0;MQ=0;ANN=A|frameshift_variant&splice_a
```

```
cceptor_variant&splice_region_variant&intron_variant|HIGH|Rps19|ENSMUSG000000
40952|transcript|ENSMUST00000108428.7|protein_coding|5/5|c.357-
2_360delAGGGGC|p.Gly120fs||357/639|119/212||INFO_REALIGN_3_PRIME,A|frameshift
_variant&splice_acceptor_variant&splice_region_variant&intron_variant|HIGH|Rp
s19|ENSMUSG00000040952|transcript|ENSMUST00000108430.9|protein_coding|5/6|c.3
57-
2_360delAGGGGC|p.Gly120fs||357/438|119/145||INFO_REALIGN_3_PRIME,A|frameshift
_variant&splice_acceptor_variant&splice_region_variant&intron_variant|HIGH|Rp
s19|ENSMUSG00000040952|transcript|ENSMUST00000108429.7|protein_coding|5/6|c.3
57-
2_360delAGGGGC|p.Gly120fs||357/438|119/145||INFO_REALIGN_3_PRIME,A|frameshift
_variant&splice_acceptor_variant&splice_region_variant&intron_variant|HIGH|Rp
s19|ENSMUSG00000040952|transcript|ENSMUST00000156372.7|protein_coding|5/5|c.3
90-
2_393delAGGGGC|p.Gly131fs||390/413|130/136||WARNING_TRANSCRIPT_INCOMPLETE&INF
O_REALIGN_3_PRIME,A|frameshift_variant&splice_acceptor_variant&splice_region_
variant&intron_variant|HIGH|Rps19|ENSMUSG00000040952|transcript|ENSMUST000001
24035.1|protein_coding|4/4|c.465-
2_468delAGGGGC|p.Gly156fs||465/530|155/175||WARNING_TRANSCRIPT_INCOMPLETE&INF
O_REALIGN_3_PRIME,A|frameshift_variant&splice_acceptor_variant&splice_region_
variant&intron_variant|HIGH|Rps19|ENSMUSG00000040952|transcript|ENSMUST000001
53451.8|protein_coding|5/5|c.357-
2_360delAGGGGC|p.Gly120fs||357/400|119/132||WARNING_TRANSCRIPT_INCOMPLETE&INF
O_REALIGN_3_PRIME,A|splice_acceptor_variant&3_prime_UTR_variant&intron_varian
t|HIGH|Rps19|ENSMUSG00000040952|transcript|ENSMUST00000129847.7|nonsense_medi
ated_decay|6/7|c.*219-
2_*222delAGGGGC|||||2752|INFO_REALIGN_3_PRIME,A|splice_acceptor_variant&splic
e_region_variant&intron_variant&non_coding_transcript_exon_variant|HIGH|Rps19
|ENSMUSG00000040952|transcript|ENSMUST00000130335.7|retained_intron|2/3|n.277
-
2_280delAGGGGC|||||INFO_REALIGN_3_PRIME,A|splice_region_variant|LOW|Rps19|EN
SMUSG00000040952|transcript|ENSMUST00000129847.7|nonsense_mediated_decay|6/7|
c.*219-
2_*222delAGGGGC||||||INFO_REALIGN_3_PRIME,A|downstream_gene_variant|MODIFIER|
Rps19|ENSMUSG00000040952|transcript|ENSMUST00000138662.1|retained_intron||n.*
1145_*1150delCAGGGG|||||1145|,A|non_coding_transcript_exon_variant|MODIFIER|R
ps19|ENSMUSG00000040952|transcript|ENSMUST00000146004.1|processed_transcript|
1/2|n.166_171delAGGGGC||||||INFO_REALIGN_3_PRIME,A|non_coding_transcript_vari
ant|MODIFIER|Rps19|ENSMUSG00000040952|transcript|ENSMUST00000129847.7|nonsens
e_mediated_decay|6/7|c.*219-
2_*222delAGGGGC||||||INFO_REALIGN_3_PRIME;LOF=(Rps19|ENSMUSG00000040952|10|0.
60)"
## [4]
"SNVHPOL=3;MQ=255;ANN=T|stop_gained&splice_region_variant|HIGH|Hnrnpl|ENSMUSG
00000015165|transcript|ENSMUST00000174477.7|protein_coding|5/12|c.769A>T|p.Ly
s257*|770/2180|769/1845|257/614||WARNING_TRANSCRIPT_NO_START_CODON,T|stop_gai
ned&splice_region_variant|HIGH|Hnrnpl|ENSMUSG00000015165|transcript|ENSMUST00
000038572.14|protein_coding|5/13|c.796A>T|p.Lys266*|817/2142|796/1761|266/586
||,T|stop_gained&splice_region_variant|HIGH|Hnrnpl|ENSMUSG00000015165|transcr
ipt|ENSMUST00000174548.7|protein_coding|6/14|c.796A>T|p.Lys266*|1354/2679|796
```

```
/1761|266/586||,T|stop_gained&splice_region_variant|HIGH|Hnrnpl|ENSMUSG000000
15165|transcript|ENSMUST00000172529.7|protein_coding|5/13|c.406A>T|p.Lys136*|
590/1898|406/1371|136/456||,T|stop_gained&splice_region_variant|HIGH|Hnrnpl|E
NSMUSG00000015165|transcript|ENSMUST00000174882.7|nonsense_mediated_decay|4/1
3|c.472A>T|p.Lys158*|472/1850|472/606|158/201||WARNING_TRANSCRIPT_NO_START_CO
DON,T|splice_region_variant&non_coding_transcript_exon_variant|LOW|Hnrnpl|ENS
MUSG00000015165|transcript|ENSMUST00000173750.7|retained_intron|5/6|n.592A>T|
|||||,T|splice_region_variant&non_coding_transcript_exon_variant|LOW|Hnrnpl|E
NSMUSG00000015165|transcript|ENSMUST00000174755.7|retained_intron|5/6|n.767A>
T||||||,T|splice_region_variant&non_coding_transcript_exon_variant|LOW|Hnrnpl
|ENSMUSG00000015165|transcript|ENSMUST00000173818.7|retained_intron|1/5|n.370
A>T||||||,T|splice_region_variant&non_coding_transcript_exon_variant|LOW|Hnrn
pl|ENSMUSG00000015165|transcript|ENSMUST00000172841.1|retained_intron|1/5|n.8
8A>T||||||,T|upstream_gene_variant|MODIFIER|Hnrnpl|ENSMUSG00000015165|transcr
ipt|ENSMUST00000173578.1|retained_intron||n.-
2641A>T|||||2641|,T|upstream_gene_variant|MODIFIER|Gm44702|ENSMUSG00000109420
|transcript|ENSMUST00000209194.1|antisense||n.-
4696T>A|||||4696|,T|downstream_gene_variant|MODIFIER|Hnrnpl|ENSMUSG0000001516
5|transcript|ENSMUST00000174396.1|retained_intron||n.*746A>T|||||746|,T|downs
tream_gene_variant|MODIFIER|Hnrnpl|ENSMUSG00000015165|transcript|ENSMUST00000
172884.7|protein_coding||c.*68A>T|||||68|WARNING_TRANSCRIPT_INCOMPLETE;LOF=(H
nrnpl|ENSMUSG00000015165|13|0.38);NMD=(Hnrnpl|ENSMUSG00000015165|13|0.38)"
```

All of them has the potential to affect the final transcripted protein

Q5.10

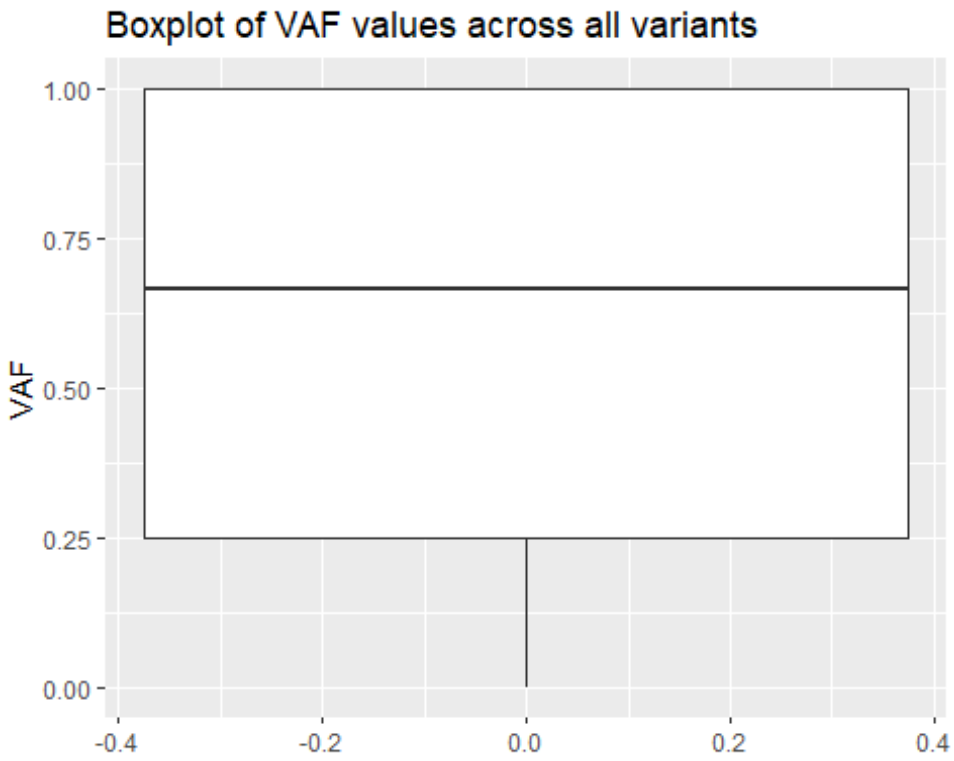
```
library(dplyr)
for (index in 1:nrow(variants)){
  variants$AD[index] = strsplit(variants[index,10],":")[[1]][6]
  variants$ref_count[index] =
strsplit(variants[index,17][[1]][1],",")[[1]][1]
  variants$alt_count[index] =
strsplit(variants[index,17][[1]][1],",")[[1]][2]

  alt_count_index = as.integer(variants$alt_count[index][[1]][1])
  ref_count_index =  as.integer(variants$ref_count[index][[1]][1])
  variants$VAF[index] = alt_count_index / (alt_count_index + ref_count_index)

}

ggplot(data = variants) +
  geom_boxplot(aes(y = VAF))+
  ggtitle("Boxplot of VAF values across all variants")

## Warning: Removed 8 rows containing non-finite values (`stat_boxplot()`).
```




```
variants$VAFgreater5 = variants$VAF > 0.05
cat("Number of variants with VAF > 5%:", sum(variants$VAFgreater5, na.rm =
TRUE), "\n")

## Number of variants with VAF > 5%: 816

for (index in 1:nrow(variants)){
  variants$VAFgreat_codingregion[index] = variants$VAF[index] > 0.05 &
grepl("protein_coding",variants$INFO[index])
}

cat("Number of variants with VAF > 5% and in protein coding region:",
    sum(variants$VAFgreat_codingregion, na.rm = TRUE))

## Number of variants with VAF > 5% and in protein coding region: 681
```

Contributions

Team members all worked individually and compared results. Hannah added bonus question 4.10. Hannah and Jingxuan reviewed and editted Q2. Hannah editted Q4.2. Theo rendered and knitted the submission file