[Tutorial] Human Genome Annotation

Tutorial for Tidyverse(Chapter 4)

1. Introduction

1.1. What is gene annotation?

Over the past years, we have learnt that there are a number of chromosomes and genes in our genome. Counting the number of chromosomes is fairly easy but students might find difficult to say how many genes we have in our genome. If you can get an answer for this, could you tell how many genes encode protein and how many do not?

To answer this question, we need to access the database for gene annotation. Gene annotation is the process of making nucleotide sequence meaningful - where genes are located? whether it is protein-coding or noncoding. If you would like to get an overview of gene annotation, please find this link.

One of well-known collaborative efforts in gene annotation is the GENCODE consortium. It is a part of the Encyclopedia of DNA Elements (The ENCODE project consortium) and aims to identify all gene features in the human genome using a combination of computational analysis, manual annotation, and experimental validation (Harrow et al. 2012). You might find another database for gene annotation, like RefSeq, CCDS, and need to understand differences between them.

Figure 1. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction (Frankish et al. 2015). A) Mean number of alternatively spliced transcripts per multi-exon protein-coding locus B) Mean number of unique CDS per multi-exon protein-coding locus C) Mean number of unique (non-redundant) exons per multi-exon protein-coding locus D) Percentage genomic coverage of unique (non-redundant) exons at multi-exon protein-coding loci.

In this tutorial, we will access to gene annotation from the GENCODE consortium and explore genes and functional elements in our genome.

1.2. Aims

What we will do with this dataset:

-Be familiar with gene annotation modality.

-Tidy data and create a table for your analysis.

-Apply tidyverse functions for data munging.

Please note that there is better solution for getting gene annotation in R if you use a biomart. Our tutorial is only designed to have a practice on tidyverse exercise.

2. Explore your data

2.1. Unboxing your dataset

This tutorial will use a gene annotation file from the GENCODE. You will need to download the file from the GENCODE. If you are using terminal, please download file using wget:

```
# Run from your terminal, not R console
# wget ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_31/gencode.v31.basic.annotation.

# Once you downloaded the file, you won't need to download it again. So please comment out the command
```

Once you download the file, you can print out the first few lines using the following bash command (we will learn UNIX commands later):

```
# Run from your terminal, not R console
#gzcat gencode.v31.basic.annotation.gtf.gz | head -7
#zcat?
```

The file is the GFT file format, which you will find most commonly in gene annotation. Please read the file format thoroughly in the link above.

For the tutorial, we need to load two packages. If the package is not installed in your system, please install it.

-tidyverse, a package you have learnt from the chapter 5. -readr, a package provides a fast and friendly way to read. Since the file gencode.v31.basic.annotation.gtf.gz is pretty large, you will need some function to load data quickly into your workspace. readr in a part of tidyverse, so you can just load tidyverse to use readr functions.

Let's load the GTF file into your workspace. We will use read_delim function from the readr package. This is much faster loading than read.delim or read.csv from R base. However, please keep in mind that some parameters and output class for read_delim are slightly different from them.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
d = read_delim('gencode.v31.basic.annotation.gtf.gz', delim='\t', skip = 5, progress = F, col_names = F
```

```
## Rows: 1756502 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (7): X1, X2, X3, X6, X7, X8, X9
## dbl (2): X4, X5

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Can you find out what the parameters mean? Few things to note are:

-The GTF file contains the first few lines for comments (#). In general, the file contains description, provider, date, format.

-The GTF file does not have column names so you will need to assign 'FALSE for col_names.

This is sort of canonical way to load your dataset into R. However, we are using a GTF format, which is specific to gene annotation so we can use a package to specifically handle a GTF file.

Here I introduce the package rtracklayer. Let's install the package first.

```
#if (!requireNamespace("BiocManager", quietly = TRUE))
#    install.packages("BiocManager")

#::install("rtracklayer")
```

Then, now you can read the GTF file using this package. Then, you can check the class of the object d.

```
d_26 = rtracklayer::import('gencode.v31.basic.annotation.gtf.gz')
```

```
class(d_26)
```

```
## [1] "GRanges"
## attr(,"package")
## [1] "GenomicRanges"
```

You will find out that this is GRanges class. This is from the package Genomic Range, specifically dealing with genomic datasets but we are not heading into this in this tutorial. So please find this information if you are serious on this.

We are converting d into a data frame as following:

```
d = d %>% as.data.frame()
```

Let's overview few lines from the data frame, and explore what you get in this object.

```
head(d)
```

```
##     X1     X2         X3    X4    X5 X6 X7 X8
## 1 chr1 HAVANA       gene 11869 14409  .  +  .
## 2 chr1 HAVANA transcript 11869 14409  .  +  .
## 3 chr1 HAVANA       exon 11869 12227  .  +  .
## 4 chr1 HAVANA       exon 12613 12721  .  +  .
## 5 chr1 HAVANA       exon 13221 14409   .  +  .
## 6 chr1 HAVANA transcript 12010 13670   .  +  .
##
## 1
## 2                                                                     gene_id "ENSG00000223972.5"; tra
## 3                        gene_id "ENSG00000223972.5"; transcript_id "ENST00000456328.2"; gene_type "
## 4                        gene_id "ENSG00000223972.5"; transcript_id "ENST00000456328.2"; gene_type "
## 5                        gene_id "ENSG00000223972.5"; transcript_id "ENST00000456328.2"; gene_type "
## 6 gene_id "ENSG00000223972.5"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unprocessed
```

One thing you can find is that there is no columns in the data frame. Let's match which information is provided in columns. You can find the instruction page in the website (link).

Based on this, you can assign a name for 9 columns. One thing to remember is you should not use space for the column name. Spacing in the column name is actually working but not a good habit for your code. So please replace a space with underscore in the column name.

```
# Assign column names according to the GENCODE instruction.
cols = c('chrom', 'source', 'feature_type', 'start', 'end', 'score', 'strand', 'phase', 'info')
```

Now you can set up the column names into the col_names parameter, and load the file into a data frame.

```
d = read_delim('gencode.v31.basic.annotation.gtf.gz',
               delim='\t', skip = 5,
               progress = F,
               col_names = cols)
```

```
## Rows: 1756502 Columns: 9

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (7): chrom, source, feature_type, score, strand, phase, info
## dbl (2): start, end

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

You can find the column names are now all set.

```
head(d)
```

```
## # A tibble: 6 x 9
##   chrom source feature_type start   end score strand phase info
##   <chr> <chr>  <chr>        <dbl> <dbl> <chr> <chr>  <chr> <chr>
## 1 chr1  HAVANA gene         11869 14409 .     +      .     "gene_id \"ENSG00000~
## 2 chr1  HAVANA transcript   11869 14409 .     +      .     "gene_id \"ENSG00000~
## 3 chr1  HAVANA exon         11869 12227 .     +      .     "gene_id \"ENSG00000~
## 4 chr1  HAVANA exon         12613 12721 .     +      .     "gene_id \"ENSG00000~
## 5 chr1  HAVANA exon         13221 14409 .     +      .     "gene_id \"ENSG00000~
## 6 chr1  HAVANA transcript   12010 13670 .     +      .     "gene_id \"ENSG00000~
```

When you loaded the file, you see the message about the data class. You might want to overview this data.

```
summary(d)
```

```
##     chrom              source          feature_type           start
##  Length:1756502     Length:1756502     Length:1756502     Min.   :      577
##  Class :character   Class :character   Class :character   1st Qu.: 32101517
##  Mode  :character   Mode  :character   Mode  :character   Median : 61732754
##                                                           Mean   : 75288563
##                                                           3rd Qu.:111760181
##                                                           Max.   :248936581
##       end               score              strand              phase
##  Min.   :      647   Length:1756502     Length:1756502     Length:1756502
##  1st Qu.: 32107331   Class :character   Class :character   Class :character
##  Median : 61738373   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 75292632
##  3rd Qu.:111763007
##  Max.   :248937043
##      info
##  Length:1756502
##  Class :character
##  Mode  :character
##
##
##
```

2.2. How many feature types in the GENCODE dataset?

As instructed in the GENCODE website, the GENCODE dataset provides a range of annotations for the feature type. You can check feature types using ______ function.

```
d %>% group_by(feature_type) %>% count(feature_type)
```

```
## # A tibble: 8 x 2
## # Groups:   feature_type [8]
##   feature_type        n
##   <chr>           <int>
## 1 CDS            567862
## 2 exon           744835
## 3 gene            60603
## 4 Selenocysteine     96
## 5 start_codon     57886
## 6 stop_codon      57775
## 7 transcript     108243
## 8 UTR            159202
```

```
#table(d$feature_type)
```

> 8 feature types *Selenocysteine :Selenocysteine is produced when the UGA codon, which is normally recognized as a stop codon, is translated into a specific pathway. Selenocysteine, the 21st amino acid, has been found in 25 human selenoproteins and selenoenzymes important for fundamental cellular processes ranging from selenium homeostasis maintenance to the regulation of the overall metabolic rate.

2.3. How many genes we have?

Let's count the number of genes in our genome. Since we know that the column feature_type contains rows with gene, which contains obviously annotations for genes. We might want to subset those rows from the data frame.

```
d1 = filter(d, feature_type == 'gene')

#d1 = d[d$feature_type == 'gene', ]
```

2.4. Ensembl, Havana and CCDS.

Gene annotation for the human genome is provided by multiple organizations with different gene annotation methods and strategy. This means that information can be varying by resources, and users need to understand heterogeniety inherent in annotation databases.

The GENCODE project utlizes two sources of gene annotation.

1. Havana: Manual gene annotation (detailed strategy in here)
2. Ensembl: Automatic gene annotation (detailed strategy in here)

It provides the combination of Ensembl/HAVANA gene set as the default gene annotation for the human genome. In addition, they also guarantee that all transcripts from the Consensus Coding Sequence (CCDS) set are present in the GENCODE gene set. The CCDS project is a collaborative effort to identify a core set of protein coding regions that are consistently annotated and of high quality. Initial results from the Consensus CDS (CCDS) project are now available through the appropriate Ensembl gene pages and from the CCDS project page at NCBI. The CCDS set is built by consensus among Ensembl, the National Center for Biotechnology Information (NCBI), and the HUGO Gene Nomenclature Committee (HGNC) for human (link).

Right. Then now we count how many genes annotated with HAVANA and ENSEMBL.

```
d %>% group_by(source) %>% count(source)
```

```
## # A tibble: 2 x 2
## # Groups:   source [2]
##   source        n
##   <chr>     <int>
## 1 ENSEMBL  245185
## 2 HAVANA  1511317
```

2.5. do.call

Since the last column info contains a long string for multiple annotations, we will need to split it to extract each annotation. For example, the first line for transcript annotation looks like this:

```
#chr1    HAVANA    transcript    11869    14409    .    +    .    gene_id "ENSG00000223972.5"; transcri
```

If you would like to split transcript_support_level and create a new column, you can use strsplit function.

```
a = 'chr1    HAVANA    transcript    11869    14409    .    +    .    gene_id "ENSG00000223972.5"; trans

strsplit(a, 'transcript_support_level\\s+"')
```

```
## [[1]]
## [1] "chr1    HAVANA    transcript    11869    14409    .    +    .    gene_id \"ENSG00000223972.5\";
## [2] "1\"; hgnc_id \"HGNC:37102\"; tag \"basic\"; havana_gene \"OTTHUMG00000000961.2\"; havana_transc
```

After split the string, you can select the second item in the list ([[1]][2]).

```
strsplit(a, 'transcript_support_level\\s+"')[[1]][2]
```

```
## [1] "1\"; hgnc_id \"HGNC:37102\"; tag \"basic\"; havana_gene \"OTTHUMG00000000961.2\"; havana_transc
```

You can find the 1 in the first position, which you will need to split again.

```
b = strsplit(a, 'transcript_support_level\\s+"')[[1]][2]
strsplit(b, '\\"')
```

```
## [[1]]
##  [1] "1"                    "; hgnc_id "           "HGNC:37102"
##  [4] "; tag "               "basic"                "; havana_gene "
##  [7] "OTTHUMG00000000961.2" "; havana_transcript " "OTTHUMT00000362751.1"
## [10] ";"
```

From this, you will get the first item in the list ([[1]][1]).

Now you would like to apply strsplit function across vectors. For this, do.call function can be easily implemented to strsplit over the vectors from one column. Let's try this.

```
head(do.call(rbind.data.frame, strsplit(a, 'transcript_support_level\\s+"'))[[2]])
```

```
## [1] "1\"; hgnc_id \"HGNC:37102\"; tag \"basic\"; havana_gene \"OTTHUMG00000000961.2\"; havana_transc
```

Now you can write two lines of codes to process two steps we discussed above.

```
# First filter transcripts and create a data frame.
d2 <- d %>% filter(feature_type == 'transcript')

# Now apply the functions.
d2$transcript_support_level <- as.character(do.call(rbind.data.frame,
                     strsplit(d2$info, 'transcript_support_level\\s+"'))[[2]])

d2$transcript_support_level <- as.character(do.call(rbind.data.frame,
                                                    strsplit( d2$transcript_support_level, '\\"'))[[1]]
```

Now you can check the strsplit works.

```
head(d2$transcript_support_level)
```

```
## [1] "1"  "NA" "NA" "NA" "5"  "5"
```

You can use the same method to extract other annotations, like gene_id, gene_name etc.

3. Exercises

Here I list the questions for your activity. Please note that it is an exercise for tidyverse functions, which you will need to use in your code. In addition, you will need to write an one-line code for each question using pipe %>%.

For questions, you should read some information thoroughly, including:

-Gene biotype.

-0 or 1 based annotation in GTF, BED format.

-Why some features have 1 bp length?

-What is the meaning of zero-length exons in GENCODE? Also fun to have a review for microexons

-Transcript support level (TSL)

3.1. Annotation of transcripts in our genome

1. Computes the number of transcripts per gene. What is the mean number of transcripts per gene? What is the quantile (25%, 50%, 75%) for these numbers? Which gene has the greatest number of transcript?

```
library(dplyr)
d_26 <- as.data.frame(d_26)
```

```
trans <- d_26 %>% group_by(gene_id) %>%
  filter(type == "transcript") %>%
  count() %>% ungroup()

trans
```

```
## # A tibble: 60,603 x 2
##    gene_id                n
##    <chr>              <int>
##  1 ENSG00000000003.14     3
##  2 ENSG00000000005.6      1
##  3 ENSG00000000419.12     2
##  4 ENSG00000000457.14     3
##  5 ENSG00000000460.17     5
##  6 ENSG00000000938.13     3
##  7 ENSG00000000971.15     3
##  8 ENSG00000001036.13     1
##  9 ENSG00000001084.13     5
## 10 ENSG00000001167.14     2
## # ... with 60,593 more rows
```

```r
trans %>% summarise(mean = mean(n))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  1.79
```

```r
trans %>% summarize(Quantile = quantile(n, c(0.25, 0.50, 0.75)), "%" = c(25, 50, 75))
```

```
## # A tibble: 3 x 2
##   Quantile   `%`
##      <dbl> <dbl>
## 1        1    25
## 2        1    50
## 3        2    75
```

```r
trans %>% summarise(max = gene_id[which.max(n)])
```

```
## # A tibble: 1 x 1
##   max
##   <chr>
## 1 ENSG00000109339.22
```

2. Compute the number of transcripts per gene among gene biotypes. For example, compare the number of transcript per gene between protein-coding genes, long noncoding genes, pseudogenes.

```r
s <- d_26 %>% group_by(gene_type, gene_id) %>%
  filter(type == "transcript") %>%
  count()

s %>% arrange(gene_type)
```

```
## # A tibble: 60,603 x 3
## # Groups:   gene_type, gene_id [60,603]
##    gene_type gene_id                n
##    <chr>     <chr>              <int>
##  1 IG_C_gene ENSG00000211592.8      1
##  2 IG_C_gene ENSG00000211675.2      1
##  3 IG_C_gene ENSG00000211677.2      1
##  4 IG_C_gene ENSG00000211679.2      1
##  5 IG_C_gene ENSG00000211685.3      1
##  6 IG_C_gene ENSG00000211890.4      1
##  7 IG_C_gene ENSG00000211891.6      1
##  8 IG_C_gene ENSG00000211892.4      1
##  9 IG_C_gene ENSG00000211893.4      1
## 10 IG_C_gene ENSG00000211895.5      1
## # ... with 60,593 more rows
```

```
#mean
s %>% group_by(gene_type) %>%
  summarise(m_transcripts = mean(n))
```

```
## # A tibble: 40 x 2
##    gene_type       m_transcripts
##    <chr>                   <dbl>
##  1 IG_C_gene                1
##  2 IG_C_pseudogene          1
##  3 IG_D_gene                1
##  4 IG_J_gene                1
##  5 IG_J_pseudogene          1
##  6 IG_pseudogene            1
##  7 IG_V_gene                1
##  8 IG_V_pseudogene          1
##  9 lncRNA                   1.48
## 10 miRNA                    1
## # ... with 30 more rows
```

```
#quantile(0%, 25%, 50%, 75%, 100%)
s %>% group_by(gene_type) %>%
  summarise(Quantile = quantile(n, c(0.00, 0.25, 0.50, 0.75, 1.00)), "%" = c(0, 25, 50, 75, 100), .grou
```

```
## # A tibble: 200 x 3
## # Groups:   gene_type [40]
##    gene_type       Quantile   `%`
##    <chr>              <dbl> <dbl>
##  1 IG_C_gene              1     0
##  2 IG_C_gene              1    25
##  3 IG_C_gene              1    50
##  4 IG_C_gene              1    75
##  5 IG_C_gene              1   100
##  6 IG_C_pseudogene        1     0
##  7 IG_C_pseudogene        1    25
##  8 IG_C_pseudogene        1    50
##  9 IG_C_pseudogene        1    75
## 10 IG_C_pseudogene        1   100
## # ... with 190 more rows
```

```
#max
s %>% ungroup() %>%
  top_n(1, n)
```

```
## # A tibble: 1 x 3
##   gene_type      gene_id                 n
##   <chr>          <chr>               <int>
## 1 protein_coding ENSG00000109339.22    87
```

3. Final task is to compute the number of transcripts per gene per chromosome.

```
l <- d_26 %>% group_by(seqnames, gene_type, gene_id) %>%
  filter(type == "transcript") %>%
  count()

l
```

```
## # A tibble: 60,603 x 4
## # Groups:   seqnames, gene_type, gene_id [60,603]
##    seqnames gene_type       gene_id                n
##    <fct>    <chr>           <chr>              <int>
##  1 chr1     IG_V_pseudogene ENSG00000276674.1      1
##  2 chr1     lncRNA          ENSG00000116652.6      1
##  3 chr1     lncRNA          ENSG00000116883.8      1
##  4 chr1     lncRNA          ENSG00000117242.7      1
##  5 chr1     lncRNA          ENSG00000153363.13     9
##  6 chr1     lncRNA          ENSG00000162888.4      1
##  7 chr1     lncRNA          ENSG00000162913.10     2
##  8 chr1     lncRNA          ENSG00000175147.13     4
##  9 chr1     lncRNA          ENSG00000176320.2      1
## 10 chr1     lncRNA          ENSG00000176754.13     2
## # ... with 60,593 more rows
```

```
#mean
l %>% group_by(seqnames) %>%
  summarise(Mean = mean(n))
```

```
## # A tibble: 25 x 2
##    seqnames  Mean
##    <fct>    <dbl>
##  1 chr1      1.80
##  2 chr2      1.77
##  3 chr3      1.93
##  4 chr4      1.76
##  5 chr5      1.74
##  6 chr6      1.78
##  7 chr7      1.76
##  8 chr8      1.75
##  9 chr9      1.70
## 10 chr10     1.78
## # ... with 15 more rows
```

```
#quantile(0%, 25%, 50%, 75%, 100%)
l %>% group_by(seqnames) %>%
  summarise(Quantile = quantile(n, c(0.00, 0.25, 0.50, 0.75, 1.00)), "%" =c(0, 25, 50, 75, 100) ,.groups
```

```
## # A tibble: 125 x 3
## # Groups:   seqnames [25]
##    seqnames Quantile   `%`
##    <fct>       <dbl> <dbl>
##  1 chr1            1     0
##  2 chr1            1    25
##  3 chr1            1    50
##   4 chr1             2    75
##   5 chr1            33   100
##   6 chr2             1     0
##   7 chr2             1    25
##   8 chr2             1    50
##   9 chr2             2    75
##  10 chr2            31   100
## # ... with 115 more rows
```

```
l %>% ungroup() %>%
  top_n(1,n)
```

```
## # A tibble: 1 x 4
##   seqnames gene_type      gene_id                  n
##   <fct>    <chr>          <chr>                <int>
## 1 chr4     protein_coding ENSG00000109339.22    87
```

3.2. Gene length in the GENCODE

1. What is the average length of human genes?

```
d_26 %>%
  summarise(mean = mean(width))
```

```
##       mean
## 1 4069.518
```

2. Is the distribution of gene length differed by autosomal and sex chromosomes? Please calculate the quantiles (0%, 25%, 50%, 75%, 100%) of the gene length for each group.

```
d_chrtype <- d_26 %>%
  filter(seqnames != "chrM") %>%
  mutate(chr_type = case_when(
    seqnames %in% c("chrX", "chrY") ~ "sex",
    TRUE ~ "autosomal"
  ))

d_chrtype %>%
  group_by(chr_type) %>%
  summarise(Quantile = quantile(width, c(0.00, 0.25, 0.50, 0.75, 1.00)), "%" = c(0, 25, 50, 75, 100), .g
```

```
## # A tibble: 10 x 3
## # Groups:   chr_type [2]
##    chr_type  Quantile   `%`
##    <chr>        <dbl> <dbl>
##  1 autosomal        1     0
##  2 autosomal       80    25
##  3 autosomal      129    50
##  4 autosomal      222    75
##  5 autosomal  2473537   100
##  6 sex              1     0
##  7 sex             78    25
##  8 sex            127    50
##  9 sex            230    75
## 10 sex        2241765   100
```

3. Is the distribution of gene length differed by gene biotype? Please calculate the quantiles (0%, 25%, 50%, 75%, 100%) of the gene length for each group.

```
d_26 %>%
  group_by(gene_type) %>%
  summarise(Quantile = quantile(width, c(0.00, 0.25, 0.50, 0.75, 1.00)), "%" = c(0, 25, 50, 75, 100), .g
```

```
## # A tibble: 200 x 3
## # Groups:   gene_type [40]
##    gene_type       Quantile   `%`
##    <chr>              <dbl> <dbl>
##  1 IG_C_gene              3     0
##  2 IG_C_gene             92    25
##  3 IG_C_gene            312.    50
##  4 IG_C_gene            336    75
##  5 IG_C_gene           8914   100
##  6 IG_C_pseudogene       34     0
##  7 IG_C_pseudogene      293    25
##  8 IG_C_pseudogene      316    50
##  9 IG_C_pseudogene      424    75
## 10 IG_C_pseudogene     5211   100
## # ... with 190 more rows
```

3.3. Transcript support levels (TSL)

The GENCODE TSL provides a consistent method of evaluating the level of support that a GENCODE transcript annotation is actually expressed in humans.

1. With transcript, how many transcripts are categorized for each TSL?

```
d_26 %>%
  group_by(transcript_support_level) %>%
  filter(type == "transcript") %>%
  count()
```

```
## # A tibble: 7 x 2
## # Groups:   transcript_support_level [7]
##   transcript_support_level     n
##   <chr>                    <int>
## 1 1                        31801
## 2 2                        13372
## 3 3                         7228
## 4 4                         2245
## 5 5                        13674
## 6 NA                       27843
## 7 <NA>                     12080
```

2. From the first question, please count the number of transcript for each TSL by gene biotype.

```
d_26 %>%
  group_by(gene_type, transcript_support_level) %>%
  filter(type == "transcript") %>%
  count()
```

```
## # A tibble: 91 x 3
## # Groups:   gene_type, transcript_support_level [91]
##    gene_type       transcript_support_level     n
##    <chr>           <chr>                    <int>
##  1 IG_C_gene       1                            1
##  2 IG_C_gene       5                            1
##  3 IG_C_gene       NA                           7
##  4 IG_C_gene       <NA>                         5
##  5 IG_C_pseudogene NA                           9
##  6 IG_D_gene       NA                          37
##  7 IG_J_gene       NA                          18
##  8 IG_J_pseudogene NA                           3
##  9 IG_pseudogene   NA                           1
## 10 IG_V_gene       5                            3
## # ... with 81 more rows
```

3. From the first question, please count the number of transcript for each TSL by source.

```
d_26 %>%
  group_by(source, transcript_support_level) %>%
  filter(type == "transcript") %>%
  count()
```

```
## # A tibble: 14 x 3
## # Groups:   source, transcript_support_level [14]
##    source  transcript_support_level     n
##    <fct>   <chr>                    <int>
##  1 HAVANA  1                        29434
##  2 HAVANA  2                        12052
##  3 HAVANA  3                         6964
##  4 HAVANA  4                         2116
##  5 HAVANA  5                        10157
##  6 HAVANA  NA                       19962
##  7 HAVANA  <NA>                     11901
##  8 ENSEMBL 1                         2367
##  9 ENSEMBL 2                         1320
## 10 ENSEMBL 3                          264
## 11 ENSEMBL 4                          129
## 12 ENSEMBL 5                         3517
## 13 ENSEMBL NA                        7881
## 14 ENSEMBL <NA>                       179
```

3.4. CCDS in the GENCODE

1. With gene, please create a data frame with the columns - gene_id, gene_name, hgnc_id, gene_type, chromosome, start, end, and strand. Then, please create new columns for presence of hgnc and ccds.

For example, you can put 1 in the column isHgnc, if hgnc annotation is avaiable, or 0 if not. Then, you can put 1 in the column isCCDS, if ccds annotation is avaiable, or 0 if not.

```
d_3.4 <- d_26 %>%
  select(gene_id, gene_name, hgnc_id, gene_type, seqnames, start, end, strand) %>%
  mutate(isHgnc = case_when(
         is.na(hgnc_id) ~ "0",
         TRUE ~ "1")
         ) %>%
  mutate(isCCDS = case_when(
    is.na(d_26$ccdsid) ~ "0",
    TRUE ~ "1")
    )
```

2. Please count the number of hgnc by gene biotypes.

```
d_3.4 %>%
  filter(isHgnc == "1") %>%
  group_by(gene_id, gene_type) %>%
  count() %>% ungroup() %>% group_by(gene_type) %>%
  count()
```

```
## # A tibble: 36 x 2
## # Groups:   gene_type [36]
##    gene_type           n
##    <chr>           <int>
##  1 IG_C_gene          14
##  2 IG_C_pseudogene     9
##  3 IG_D_gene          37
##  4 IG_J_gene          18
##  5 IG_J_pseudogene     3
##  6 IG_V_gene         142
##  7 IG_V_pseudogene   185
##  8 lncRNA           3970
##  9 miRNA            1856
## 10 misc_RNA         1033
## # ... with 26 more rows
```

3. Please count the number of hgnc by level. Please note that level in this question is not TSL. Please find information in this link: 1 (verified loci), 2 (manually annotated loci), 3 (automatically annotated loci).

```
d_3.4.3 <- d_26  %>%
  mutate(isHgnc = case_when(
         is.na(hgnc_id) ~ "0",
         TRUE ~ "1")
         )

d_3.4.3 %>%
  filter(isHgnc == "1")%>%
  group_by(level) %>%
  count()
```

```
## # A tibble: 3 x 2
## # Groups:   level [3]
##   level       n
##   <chr>   <int>
## 1 1      107054
## 2 2     1279964
## 3 3      237265
```

```
#without creating "isHgnc" column, using grepl()

d_26 %>%
  filter(grepl("HGNC", d_26$hgnc_id) == TRUE) %>%
  group_by(level) %>%
  count()
```

```
## # A tibble: 3 x 2
## # Groups:   level [3]
##   level       n
##   <chr>   <int>
## 1 1      107054
## 2 2     1279964
## 3 3      237265
```

3.5. Transcripts in the GENCODE

1. Which gene has the largest number of transcripts?

```
#which.max
trans %>%
  summarise(max = gene_id[which.max(n)])
```

```
## # A tibble: 1 x 1
##   max
##   <chr>
## 1 ENSG00000109339.22
```

```
#top_n
trans %>% ungroup() %>%
  top_n(1,n)
```

```
## # A tibble: 1 x 2
##   gene_id                n
##   <chr>              <int>
## 1 ENSG00000109339.22    87
```

2. Please calculate the quantiles (0%, 25%, 50%, 75%, 100%) of the gene length for protein coding genes and long noncoding genes.

```
d_26 %>%
  filter(gene_type %in% c("protein_coding", "lncRNA")) %>%
  group_by(gene_type) %>%
  summarise(Quantile = quantile(width, c(0.00, 0.25, 0.50, 0.75, 1.00)), "%" = c(0, 25, 50, 75, 100), .g
## # A tibble: 10 x 3
## # Groups:   gene_type [2]
##    gene_type      Quantile   `%`
##    <chr>             <dbl> <dbl>
##  1 lncRNA                1     0
##  2 lncRNA              126    25
##  3 lncRNA              339    50
##  4 lncRNA             2658    75
##  5 lncRNA          1375317   100
##  6 protein_coding        1     0
##  7 protein_coding       76    25
##  8 protein_coding      123    50
##  9 protein_coding      193    75
## 10 protein_coding  2473537   100
```

3. Please count the number of transcripts by chromosomes.

```
d_26 %>%
  filter(type == "transcript") %>%
  group_by(seqnames) %>%
  count()
```

```
## # A tibble: 25 x 2
## # Groups:   seqnames [25]
##    seqnames     n
##    <fct>    <int>
##  1 chr1      9827
##  2 chr2      7432
##  3 chr3      6157
##  4 chr4      4662
##  5 chr5      5203
##  6 chr6      5455
##  7 chr7      5292
##  8 chr8      4350
##  9 chr9      3949
## 10 chr10     4157
## # ... with 15 more rows
```

3.6. Autosomal vs. Sex chromosomes.

1. Please calculate the number of genes per chromosome.

```
d_26 %>% filter(type == "gene") %>%
  group_by(seqnames) %>%
  count()
```

```
## # A tibble: 25 x 2
## # Groups:   seqnames [25]
##    seqnames     n
##    <fct>    <int>
##  1 chr1      5471
##  2 chr2      4196
##   3 chr3       3185
##   4 chr4       2651
##   5 chr5       2983
##   6 chr6       3059
##   7 chr7       3014
##   8 chr8       2482
##   9 chr9       2327
##  10 chr10      2332
## # ... with 15 more rows
```

2. Please compare the number of genes between autosomal and sex chromosome (Mean, Median).

```
d_chrtype %>% filter(type == "gene") %>%
  group_by(seqnames, chr_type) %>%
  count() %>% ungroup() %>%
  group_by(chr_type) %>%
  summarise(Mean = mean(n), Median = median(n))
```

```
## # A tibble: 2 x 3
##   chr_type   Mean Median
##   <chr>     <dbl>  <dbl>
## 1 autosomal 2617.  2604.
## 2 sex       1494.  1494.
```

3. Please divide the genes into groups ‘protein coding’ and ‘long noncoding’, and then compare the number of genes in each chromosomes within groups.

```
d_26 %>%
  filter(type == "gene" & gene_type %in% c("protein_coding", "lncRNA")) %>%
  group_by(seqnames, gene_type) %>%
  count()
```

```
## # A tibble: 49 x 3
## # Groups:   seqnames, gene_type [49]
##    seqnames gene_type          n
##    <fct>    <chr>          <int>
##  1 chr1     lncRNA          1416
##  2 chr1     protein_coding  2048
##  3 chr2     lncRNA          1241
##  4 chr2     protein_coding  1247
##  5 chr3     lncRNA           861
##  6 chr3     protein_coding  1075
##  7 chr4     lncRNA           790
##  8 chr4     protein_coding   751
##  9 chr5     lncRNA           950
## 10 chr5     protein_coding   886
## # ... with 39 more rows
```