Preprocess_migmap_Mousedata

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
data <-fread('/home/sedreh/mus_musculus/My_results.csv', sep="\
t", header=TRUE)
data1 = read.csv
('/home/sedreh/university_files/ITMO/semester2/Bcellsproject/thir
d session/filterreddata.csv')
```

```
preprocess_data <- function(data){
separate(data,
         col = "read.header",
         into = c("read", "header"),
         sep = "_")
}
preprocess_data(data)
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in
9519 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,
20, ...].
```

```
##                       read header
##    1: >TCGAGGCCACGGCTAC-1 contig
##    2: >ATTGGACCACAGACTT-1 contig
##    3: >TGACTTTGTGCGGTAA-1 contig
##    4: >CCATTCGAGCTGAACG-1 contig
##    5: >TTCGAAGCAAACAACA-1 contig
```

```
##   ---
## 9515: >TCAGCAAAGTGAAGAG-1 contig
## 9516: >TGCCAAAAGACTTGAA-1 contig
## 9517: >TGCCAAAAGACTTGAA-1 contig
## 9518: >TGGTTAGAGCGAGAAA-1 contig
## 9519: >TGGTTAGAGCGAGAAA-1 contig
##                                                    cdr3nt
cdr3aa
##    1:                TGCCAGCAGTGGAGTAGTAACCCACCCACGTTC
CQQWSSNPPTF
##    2:                TGTCAACAGCATAATGAATACCCGCTCACGTTC
CQQHNEYPLTF
##    3:                TGTGCAAGAGGCTCAGGCCACTTTGACTACTGG
CARGSGHFDYW
##    4: TGTGCAAGAGGGCCTCTTCCCTATGATTACGACTGGTTTGCTTACTGG
CARGPLPYDYDWFAYW
##    5:                TGTCAACAGTGGAGTAGTTACCCATTCACGTTC
CQQWSSYPFTF
##   ---

## 9515:                TGCTGGCAAGGTACACATTTTCCGTACACGTTC
CWQGTHFPYTF
## 9516: TGTGCAAGAGGGAGTTACGCCCCTTATTACTATGTTATGGACTACTGG
CARGSYAPYYYVMDYW
## 9517:                TGTCAGCAGGATTATAGCTCTCCGCTCACGTTC
CQQDYSSPLTF
## 9518:          TGTGCCAGAGGTAGTAGCCCTTACTACTTTGACTACTGG
CARGSSPYYFDYW
## 9519:                TGTCAACAGGGTCAAAGTTATCCTCTCACGTTC
CQQGQSYPLTF
##       cdr.insert.qual mutations.qual     v.segment   d.segment
j.segment
##    1:                                    IGKV4-72*01           .
IGKJ4*01
##    2:                                  IGKV16-104*01           .
IGKJ5*01
##    3:         IIIIIII                    IGHV9-3*01  IGHD5-7*01
IGHJ2*01
##    4:     IIIIIIIIII                    IGHV1-72*01  IGHD2-4*01
IGHJ3*01
##    5:                                I   IGKV4-53*01           .
IGKJ4*01
##   ---

## 9515:                                   IGKV1-135*01           .
IGKJ2*01
## 9516:      IIIIIIIII              I   IGHV1-59*01 IGHD2-12*01
IGHJ4*01
## 9517:                                    IGKV6-32*01           .
IGKJ5*01
```

```
## 9518:            IIII                   IGHV2-2*01  IGHD1-1*01
IGHJ2*01
## 9519:                                IGKV15-103*01           .
IGKJ5*01
##       cdr1.start.in.read cdr1.end.in.read cdr2.start.in.read
##    1:                199              214                265
##    2:                172              190                241
##    3:                200              224                275
##    4:                152              176                227
##    5:                162              183                234
##   ---
## 9515:                175              208                259
## 9516:                192              216                267
## 9517:                193              211                262
## 9518:                164              188                239
## 9519:                184              202                253
##       cdr2.end.in.read cdr3.start.in.read cdr3.end.in.read
v.end.in.cdr3
##    1:              274                379              412
28
##    2:              250                355              388
24
##    3:              299                410              443
9
##    4:              251                362              410
9
##    5:              243                348              381
28
##   ---

## 9515:              268                373              406
23
## 9516:              291                402              450
9
## 9517:              271                376              409
23
## 9518:              260                371              410
9
## 9519:              262                367              400
26
##       d.start.in.cdr3 d.end.in.cdr3 j.start.in.cdr3 v.del
d.del.5 d.del.3
##    1:              -1            -1              26     0
-1      -1
##    2:              -1            -1              23     3
-1      -1
##    3:              16            21              19     0
20       4
##    4:              20            33              32     0
4       0
```

```
##    5:               -1             -1               23     0
-1      -1
##   ---

## 9515:               -1             -1               23     3
-1      -1
## 9516:               12             18               25     0
8      12
## 9517:               -1             -1               23     3
-1      -1
## 9518:                9             18               22     2
11       3
## 9519:               -1             -1               24     0
-1      -1
##       j.del mutations.nt.FR1 mutations.nt.CDR1
mutations.nt.FR2
##    1:     3

##    2:     0

##    3:     3

##    4:     1

##    5:     0

##   ---

## 9515:     1

## 9516:     0

## 9517:     0

## 9518:     0

## 9519:     1

##       mutations.nt.CDR2 mutations.nt.FR3 mutations.nt.CDR3
##    1:
##    2:
##    3:
##    4:
##    5:                                                S288:C>T
##   ---
## 9515:
## 9516:                                                S319:C>T
## 9517:
## 9518:
```

```
## 9519:
##       mutations.nt.FR4    rc complete has.cdr3 in.frame
no.stop
##    1:                  FALSE     TRUE     TRUE     TRUE
TRUE
##    2:                  FALSE     TRUE     TRUE     TRUE
TRUE
##    3:                  FALSE     TRUE     TRUE     TRUE
TRUE
##    4:                  FALSE     TRUE     TRUE     TRUE
TRUE
##    5:                  FALSE     TRUE     TRUE     TRUE
TRUE
##   ---

## 9515:                  FALSE     TRUE     TRUE     TRUE
TRUE
## 9516:                  FALSE     TRUE     TRUE     TRUE
TRUE
## 9517:                  FALSE     TRUE     TRUE     TRUE
TRUE
## 9518:                  FALSE     TRUE     TRUE     TRUE
TRUE
## 9519:                  FALSE     TRUE     TRUE     TRUE
TRUE
##       mutations.aa.FR1 mutations.aa.CDR1 mutations.aa.FR2
##    1:
##    2:
##    3:
##    4:
##    5:
##   ---
## 9515:
## 9516:
## 9517:
## 9518:
## 9519:
##       mutations.aa.CDR2 mutations.aa.FR3 mutations.aa.CDR3
##    1:
##    2:
##    3:
##    4:
##    5:                                                S96:L>F
##   ---
## 9515:
## 9516:                                               S106:A>V
## 9517:
## 9518:
## 9519:
##       mutations.aa.FR4 pol.v pol.d.5 pol.d.3 pol.j canonical
```

```
##    1:                      -1      -1      -1    -1      TRUE
##    2:                      -1      -1      -1    -1      TRUE
##    3:                      -1      -1      -1    -1      TRUE
##    4:                      -1      -1      -1    -1      TRUE
##    5:                      -1      -1      -1    -1      TRUE
##   ---
## 9515:                      -1      -1      -1    -1      TRUE
## 9516:                      -1      -1      -1    -1      TRUE
## 9517:                      -1      -1      -1    -1      TRUE
## 9518:                      -1      -1      -1    -1      TRUE
## 9519:                      -1      -1      -1    -1      TRUE
##
contignt
##    1:
CAAATTGTTCTCTCCCAGTCTCCAGCAATCCTGTCTGCATCTCCAGGGGAGAAGGTCACAATGAC
TTGCAGGGCCAGCTCAAGTGTAAGTTACATGCACTGGTACCAGCAGAAGCCAGGATCCTCCCCCA
AACCCTGGATTTATGCCACATCCAACCTGGCTTCTGGAGTCCCTGCTCGCTTCAGTGGCAGTGGG
TCTGGGACCTCTTACTCTCTCACAATCAGCAGAGTGGAGGCTGAAGATGCTGCCACTTATTACTG
CCAGCAGTGGAGTAGTAACCCACCCACGTTCGGCTCGGGGACAAAGTTGGAAATAAAACGGGCTG
ATGCTGCACCAACTGTATCCATCTTCCCACCATCCAGTGAGCAGTTAACATCTGGAGGTGCCTCA
GTCGTGTGCTTC
##    2:
GATGTCCAGATAACCCAGTCTCCATCTTATCTTGCTGCATCTCCTGGAGAAACCATTACTATTAA
TTGCAGGGCAAGTAAGAGCATTAGCAAATATTTAGCCTGGTATCAAGAGAAACCTGGGAAAACTA
ATAAGCTTCTTATCTACTCTGGATCCACTTTGCAATCTGGAATTCCATCAAGGTTCAGTGGCAGT
GGATCTGGTACAGATTTCACTCTCACCATCAGTAGCCTGGAGCCTGAAGATTTTGCAATGTATTA
CTGTCAACAGCATAATGAATACCCGCTCACGTTCGGTGCTGGGACCAAGCTGGAGCTGAAACGGG
CTGATGCTGCACCAACTGTATCCATCTTCCCACCATCCAGTGAGCAGTTAACATCTGGAGGTGCC
TCAGTCGTGTGCTTCC
##    3:
CAGATCCAGTTGGTACAGTCTGGACCTGAGCTGAAGAAGCCTGGAGAGACAGTCAAGATCTCCTG
CAAGGCTTCTGGGTATACCTTCACAACCTATGGAATGAGCTGGGTGAAACAGGCTCCAGGAAAGG
GTTTAAAGTGGATGGGCTGGATAAACACCTACTCTGGAGTGCCAACATATGCTGATGACTTCAAG
GGACGGTTTGCCTTCTCTTTGGAAACCTCTGCCAGCACTGCCTATTTGCAGATCAACAACCTCAA
AAATGAGGACACGGCTACATATTTCTGTGCAAGAGGCTCAGGCCACTTTGACTACTGGGGCCAAG
GCACCACTCTCACAGTCTCCTCAGGTAATGAAAAGGGACCTGACATGTTCCTCCTCTCAGAGTGC
AAAGCCCCAGAGGAAAATGAAAAGATAAACCTGGGCTGTTTAGTAATTGGAAGTCAGCCACTGAA
AATCAGCTGGGAGCCAAAGAAGTCAAGTATAGTTGAACATGTCTTCCCCTCTGAAATGAGAAATG
GCAATTATACAATGGTCCTCCAGGTCACTGTGCTGGCCTCAGAACTGAACC
##    4:
CAGGTCCAACTGCAGCAGCCTGGGGCTGAGCTTGTGAAGCCTGGGGCTTCAGTGAAGCTGTCCTG
CAAGGCTTCTGGCTACACCTTCACCAGCTACTGGATGCACTGGGTGAAGCAGAGGCCTGGACGAG
GCCTTGAGTGGATTGGAAGGATTGATCCTAATAGTGGTGGTACTAAGTACAATGAGAAGTTCAAG
AGCAAGGCCACACTGACTGTAGACAAACCCTCCAGCACAGCCTACATGCAGCTCAGCAGCCTGAC
ATCTGAGGACTCTGCGGTCTATTATTGTGCAAGAGGGCCTCTTCCCTATGATTACGACTGGTTTG
CTTACTGGGGCCAAGGGACTCTGGTCACTGTCTCTGCAGAGAGTCAGTCCTTCCCAAATGTCTTC
CCCCTCGTCTCCTGCGAGAGCCCCCTGTCTGATAAGAATCTGGTGGCCATGGGCTGCCTGGCCCG
GGACTTCCTGCCCAGCACCATTTCCTTCACCTGGAACTACCAGAACAACACTGAAGTCATCCAGG
GTATCAGAACCTTCCCAACACTGAGGACAGGGGGCAAGTACCTAGCCACCTCGCA
##    5:
```

GAAATTGTGCTCACCCAGTCTCCAGCACTCATGGCTGCATCTCCAGGGGAGAAGGTCACCATCAC
CTGCAGTGTCAGCTCAAGTATAAGTTCCAGCAACTTGCACTGGTACCAGCAGAAGTCAGAAACCT
CCCCCAAACCCTGGATTTATGGCACATCCAACCTGGCTTCTGGAGTCCCTGTTCGCTTCAGTGGC
AGTGGATCTGGGACCTCTTATTCTCTCACAATCAGCAGCATGGAGGCTGAAGATGCTGCCACTTA
TTACTGTCAACAGTGGAGTAGTTACCCATTCACGTTCGGCTCGGGGACAAAGTTGGAAATAAAAC
GGGCTGATGCTGCACCAACTGTATCCATCTTCCCACCATCCAGTGAGCAGTTAACATCTGGAGGT
GCCTCAGTCGTGTGCTTC

9515:
GATGTTGTGATGACCCAGACTCCACTCACTTTGTCGGTTACCATTGGACAACCAGCCTCCATCTC
TTGCAAGTCAAGTCAGAGCCTCTTAGATAGTGATGGAAAGACATATTTGAATTGGTTGTTACAGA
GGCCAGGCCAGTCTCCAAAGCGCCTAATCTATCTGGTGTCTAAACTGGACTCTGGAGTCCCTGAC
AGGTTCACTGGCAGTGGATCAGGGACAGATTTCACACTGAAAATCAGCAGAGTGGAGGCTGAGGA
TTTGGGAGTTTATTATTGCTGGCAAGGTACACATTTTCCGTACACGTTCGGAGGGGGGACCAAGC
TGGAAATAAAACGGGCTGATGCTGCACCAACTGTATCCATCTTCCCACCATCCAGTGAGCAGTTA
ACATCTGGAGGTGCCTCAGTCGTGTGCTTC
9516:
CAGGTCCAACTGCAGCAGCCTGGGGCTGAGCTGGTGAGGCCTGGGACTTCAGTGAAGTTGTCCTG
CAAGGCTTCTGGCTACACCTTCACCAGCTACTGGATGCACTGGGTAAAGCAGAGGCCTGGACAAG
GCCTTGAGTGGATCGGAGTGATTGATCCTTCTGATAGTTATACTAACTACAATCAAAAGTTCAAG
GGCAAGGCCACATTGACTGTAGACACATCCTCCAGCACAGCCTACATGCAGCTCAGCAGCCTGAC
ATCTGAGGACTCTGCGGTCTATTACTGTGCAAGAGGGAGTTACGCCCCTTATTACTATGTTATGG
ACTACTGGGGTCAAGGAACCTCAGTCACCGTCTCCTCAGGTAATGAAAAGGGACCTGACATGTTC
CTCCTCTCAGAGTGCAAAGCCCCAGAGGAAAATGAAAAGATAAACCTGGGCTGTTTAGTAATTGG
AAGTCAGCCACTGAAAATCAGCTGGGAGCCAAAGAAGTCAAGTATAGTTGAACATGTCTTCCCCT
CTGAAATGAGAAATGGCAATTATACAATGGTCCTCCAGGTCACTGTGCTGGCCTC
9517:
AGTATTGTGATGACCCAGACTCCCAAATTCCTGCTTGTATCAGCAGGAGACAGGGTTACCATAAC
CTGCAAGGCCAGTCAGAGTGTGAGTAATGATGTAGCTTGGTACCAACAGAAGCCAGGGCAGTCTC
CTAAACTGCTGATATACTATGCATCCAATCGCTACACTGGAGTCCCTGATCGCTTCACTGGCAGT
GGATATGGGACGGATTTCACTTTCACCATCAGCACTGTGCAGGCTGAAGACCTGGCAGTTTATTT
CTGTCAGCAGGATTATAGCTCTCCGCTCACGTTCGGTGCTGGGACCAAGCTGGAGCTGAAACGGG
CTGATGCTGCACCAACTGTATCCATCTTCCCACCATCCAGTGAGCAGTTAACATCTGGAGGTGCC
TCAGTCGTGTGCTTC
9518:
CAGGTGCAGCTGAAGCAGTCAGGACCTGGCCTAGTGCAGCCCTCACAGAGCCTGTCCATCACCTG
CACAGTCTCTGGTTTCTCATTAACTAGCTATGGTGTACACTGGGTTCGCCAGTCTCCAGGAAAGG
GTCTGGAGTGGCTGGGAGTGATATGGAGTGGTGGAAGCACAGACTATAATGCAGCTTTCATATCC
AGACTGAGCATCAGCAAGGACAATTCCAAGAGCCAAGTTTTCTTTAAAATGAACAGTCTGCAAGC
TGATGACACAGCCATATATTACTGTGCCAGAGGTAGTAGCCCTTACTACTTTGACTACTGGGGCC
AAGGCACCACTCTCACAGTCTCCTCAGAGAGTCAGTCCTTCCCAAATGTCTTCCCCCTCGTCTCC
TGCGAGAGCCCCCTGTCTGATAAGAATCTGGTGGCCATGGGCTGCCTGGCCCGGGACTTCCTGCC
CAGCACCATTTCCTTCACCTGGAACTACCAGAACAACACTGAAGTCATCCAGGGTATCAGAACCT
TCCCAACACTGAGGACAGGGGGCAAGTACCTAGCCACCTCGCAGGTGTTGCTGTCTCCCAAGAGC
ATCCTTGAAGG
9519:
GACATCCAGATGAACCAGTCTCCATCCAGTCTGTCTGCATCCCTTGGAGACACAATTACCATCAC
TTGCCATGCCAGTCAGAACATTAATGTTTGGTTAAGCTGGTACCAGCAGAAACCAGGAAATATTC
CTAAACTATTGATCTATAAGGCTTCCAACTTGCACACAGGCGTCCCATCAAGGTTTAGTGGCAGT

```
GGATCTGGAACAGGTTTCACATTAACCATCAGCAGCCTGCAGCCTGAAGACATTGCCACTTACTA
CTGTCAACAGGGTCAAAGTTATCCTCTCACGTTCGGTGCTGGGACCAAGCTGGAGCTGAAACGGG
CTGATGCTGCACCAACTGTATCCATCTTCCCACCATCCAGTGAGCAGTTAACATCTGGAGGTGCC
TCAGTCGTGTGCTTC

New_data <-function(preprocess_data){
  new_data <- preprocess_data %>% select(read, v.segment,
d.segment, j.segment)
 new_data
}
New_data(preprocess_data(data))

## Warning: Expected 2 pieces. Additional pieces discarded in
9519 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,
20, ...].

##                        read     v.segment   d.segment j.segment
##    1: >TCGAGGCCACGGCTAC-1   IGKV4-72*01           .  IGKJ4*01
##    2: >ATTGGACCACAGACTT-1 IGKV16-104*01           .  IGKJ5*01
##    3: >TGACTTTGTGCGGTAA-1    IGHV9-3*01  IGHD5-7*01  IGHJ2*01
##    4: >CCATTCGAGCTGAACG-1   IGHV1-72*01  IGHD2-4*01  IGHJ3*01
##    5: >TTCGAAGCAAACAACA-1   IGKV4-53*01           .  IGKJ4*01
##   ---
## 9515: >TCAGCAAAGTGAAGAG-1  IGKV1-135*01           .  IGKJ2*01
## 9516: >TGCCAAAAGACTTGAA-1   IGHV1-59*01 IGHD2-12*01  IGHJ4*01
## 9517: >TGCCAAAAGACTTGAA-1   IGKV6-32*01           .  IGKJ5*01
## 9518: >TGGTTAGAGCGAGAAA-1    IGHV2-2*01  IGHD1-1*01  IGHJ2*01
## 9519: >TGGTTAGAGCGAGAAA-1 IGKV15-103*01           .  IGKJ5*01

Final_preprocess_data <-function(preprocess_data){
preprocess_data %>%
  rename(
    barcode = read,
    v_gene= v.segment,
    d_gene = d.segment,
    j_gene= j.segment
    )
}
Final_data =
Final_preprocess_data(New_data(preprocess_data(data)))

## Warning: Expected 2 pieces. Additional pieces discarded in
9519 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,
20, ...].

Final_data

##                    barcode        v_gene      d_gene   j_gene
##    1: >TCGAGGCCACGGCTAC-1   IGKV4-72*01           . IGKJ4*01
##    2: >ATTGGACCACAGACTT-1 IGKV16-104*01           . IGKJ5*01
```

```
##    3: >TGACTTTGTGCGGTAA-1    IGHV9-3*01  IGHD5-7*01 IGHJ2*01
##    4: >CCATTCGAGCTGAACG-1   IGHV1-72*01  IGHD2-4*01 IGHJ3*01
##    5: >TTCGAAGCAAACAACA-1   IGKV4-53*01           . IGKJ4*01
##   ---
## 9515: >TCAGCAAAGTGAAGAG-1  IGKV1-135*01           . IGKJ2*01
## 9516: >TGCCAAAAGACTTGAA-1   IGHV1-59*01 IGHD2-12*01 IGHJ4*01
## 9517: >TGCCAAAAGACTTGAA-1   IGKV6-32*01           . IGKJ5*01
## 9518: >TGGTTAGAGCGAGAAA-1    IGHV2-2*01  IGHD1-1*01 IGHJ2*01
## 9519: >TGGTTAGAGCGAGAAA-1 IGKV15-103*01           . IGKJ5*01
```

```
#Number of cells (As we have several copies of each barcode,we
need to count just one copy of unique barcoe)

barcode_summary <- function(Final_data)
{
  number_of_cells <- Final_data %>%
    distinct(barcode) %>%
    count()
  number_of_cells$n
}
barcode_summary(Final_data)
```

```
## [1] 4420
```

```
apply(Final_data[1:3,], 1, function(x) {
  substr(x[2:3], 1, 4)
})
```

```
##        [,1]   [,2]   [,3]
## v_gene "IGKV" "IGKV" "IGHV"
## d_gene "."    "."    "IGHD"
```

```
 matrix_gene_data <- as.matrix(Final_data[,2:4])
 matrix_gene_data <- substr(matrix_gene_data, 1, 3) # get only
first three characters
 Final_data$chain <- apply(matrix_gene_data, 1, function(x) {
   x <- x[!(x %in% ".")] # removeing .
   x <- unique(x) # get unique value from row
   if(length(x) == 0) { # if all are . then return none
     "None"
   } else if (length(x) > 1) { # if more than 1 unique value then
it's multi
     "Multi"
   } else { # otherwise just single chain value
     x
   }
 })
 Final_data
```

```
##                  barcode        v_gene      d_gene   j_gene
chain
```

```
##    1: >TCGAGGCCACGGCTAC-1   IGKV4-72*01           . IGKJ4*01
IGK
##    2: >ATTGGACCACAGACTT-1 IGKV16-104*01           . IGKJ5*01
IGK
##    3: >TGACTTTGTGCGGTAA-1    IGHV9-3*01  IGHD5-7*01 IGHJ2*01
IGH
##    4: >CCATTCGAGCTGAACG-1   IGHV1-72*01  IGHD2-4*01 IGHJ3*01
IGH
##    5: >TTCGAAGCAAACAACA-1   IGKV4-53*01           . IGKJ4*01
IGK
##   ---

## 9515: >TCAGCAAAGTGAAGAG-1  IGKV1-135*01           . IGKJ2*01
IGK
## 9516: >TGCCAAAAGACTTGAA-1   IGHV1-59*01 IGHD2-12*01 IGHJ4*01
IGH
## 9517: >TGCCAAAAGACTTGAA-1   IGKV6-32*01           . IGKJ5*01
IGK
## 9518: >TGGTTAGAGCGAGAAA-1    IGHV2-2*01  IGHD1-1*01 IGHJ2*01
IGH
## 9519: >TGGTTAGAGCGAGAAA-1 IGKV15-103*01           . IGKJ5*01
IGK

#show the number of occurance of each copy

barcode_summary <- function(Final_data)
  {
  barcode_summary <- group_by(Final_data, barcode)
  barcode_summary <- summarize(barcode_summary, count=n())
  barcode_summary <- group_by(barcode_summary, count)
  barcode_summary
}
result <- barcode_summary(Final_data)
result

## # A tibble: 4,420 x 2
## # Groups:   count [7]
##    barcode             count
##    <chr>               <int>
##  1 >AAACCTGAGACCTTTG-1     2
##  2 >AAACCTGAGCAACGGT-1     2
##  3 >AAACCTGAGTAGCGGT-1     2
##  4 >AAACCTGCAAACCCAT-1     2
##  5 >AAACCTGCAATCACAC-1     2
##  6 >AAACCTGCACTAGTAC-1     2
##  7 >AAACCTGCAGCCTTGG-1     2
##  8 >AAACCTGCATCTACGA-1     2
##  9 >AAACCTGGTTATCCGA-1     2
## 10 >AAACCTGGTTGATTCG-1     2
## # … with 4,410 more rows
```

```r
write.csv(Final_data,'Final_preprocess_data.csv')

#How many IGK,IGH, IGL and Multi we have in the data?
Summary <- function(Final_data)
{
  chain_summary <- group_by(Final_data, chain)
  chain_summary <- summarize(chain_summary, count=n())
  chain_summary
}
Summary (Final_data)
```

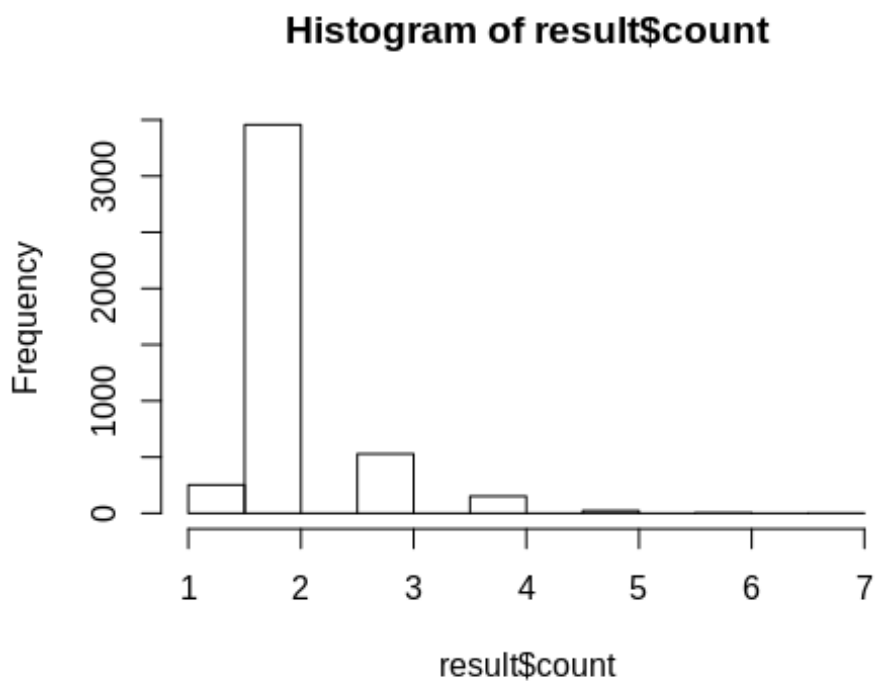
```
## # A tibble: 3 x 2
##   chain count
##   <chr> <int>
## 1 IGH    4387
## 2 IGK    4691
## 3 IGL     441
```

```r
#Number of copies of each cell

barcode_summary <- function(Final_data)
  {
  barcode_summary <- group_by(Final_data, barcode)
  barcode_summary <- summarize(barcode_summary, count=n())
  barcode_summary
}
barcode_summary(Final_data)
```

```
## # A tibble: 4,420 x 2
##    barcode             count
##    <chr>               <int>
##  1 >AAACCTGAGACCTTTG-1     2
##  2 >AAACCTGAGCAACGGT-1     2
##  3 >AAACCTGAGTAGCGGT-1     2
##  4 >AAACCTGCAAACCCAT-1     2
##  5 >AAACCTGCAATCACAC-1     2
##  6 >AAACCTGCACTAGTAC-1     2
##  7 >AAACCTGCAGCCTTGG-1     2
##  8 >AAACCTGCATCTACGA-1     2
##  9 >AAACCTGGTTATCCGA-1     2
## 10 >AAACCTGGTTGATTCG-1     2
## # … with 4,410 more rows
```

```r
#show the number of occurance of each copy

barcode_summary <- function(Final_data)
  {
  barcode_summary <- group_by(Final_data, barcode)
  barcode_summary <- summarize(barcode_summary, count=n())
  barcode_summary <- group_by(barcode_summary, count)
  barcode_summary
```

```
}
chian_Migmap <- barcode_summary(Final_data)
hist(result$count)
```

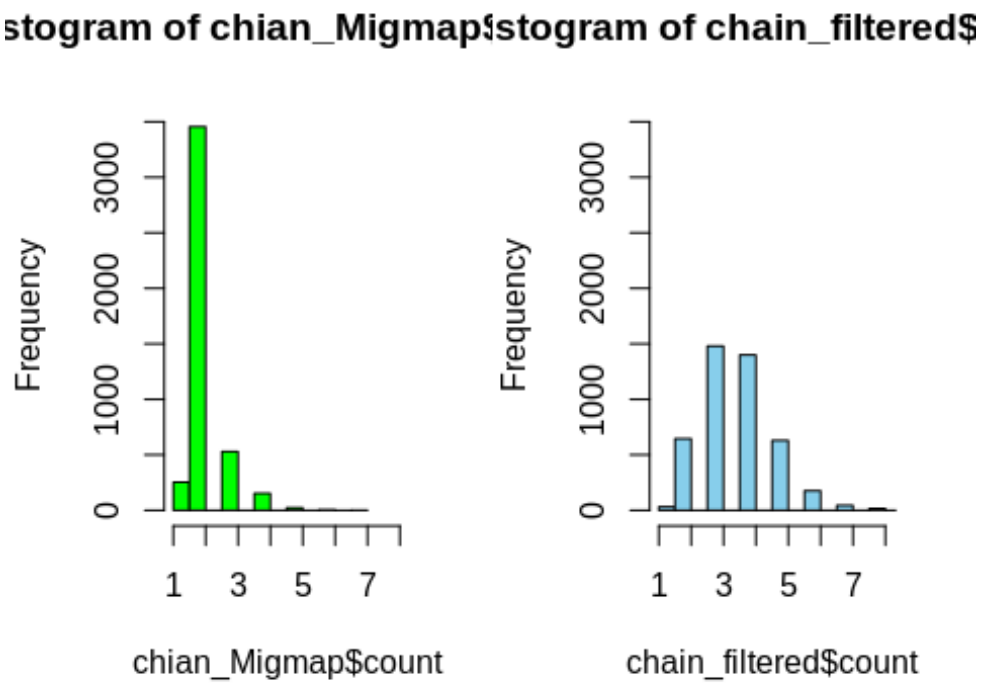



```
#show the number of occurance of each copy

barcode_summary <- function(data1)
  {
  barcode_summary <- group_by(data1, barcode)
  barcode_summary <- summarize(barcode_summary, count=n())
  barcode_summary <- group_by(barcode_summary, count)
  #barcode_summary <- summarize(barcode_summary, count_total=n())
  barcode_summary
}
chain_filtered <- barcode_summary(data1)
result
```

```
## # A tibble: 4,420 x 2
## # Groups:   count [7]
##    barcode             count
##    <chr>               <int>
##  1 >AAACCTGAGACCTTTG-1     2
##  2 >AAACCTGAGCAACGGT-1     2
##  3 >AAACCTGAGTAGCGGT-1     2
##  4 >AAACCTGCAAACCCAT-1     2
##  5 >AAACCTGCAATCACAC-1     2
##  6 >AAACCTGCACTAGTAC-1     2
```

```
##  7 >AAACCTGCAGCCTTGG-1     2
##  8 >AAACCTGCATCTACGA-1     2
##  9 >AAACCTGGTTATCCGA-1     2
## 10 >AAACCTGGTTGATTCG-1     2
## # … with 4,410 more rows
```

```
par(mfrow = c(1,2))
hist(chian_Migmap$count, col="green", xlim=c(1,8),
ylim=c(0,3500))
hist(chain_filtered$count,col="skyblue", xlim=c(1,8),
ylim=c(0,3500))
```




```
#counts number of occurance of each chain type for every unique
barcodehow

occurance_of_each_chain <- function(Final_data)
{
  result <- Final_data %>%
    group_by(barcode, chain) %>%
    filter(chain %in% c('IGK','IGH','IGL')) %>%
    summarize(count=n())
  result
}
occurance_of_each_chain(Final_data)
```

```
## # A tibble: 8,752 x 3
## # Groups:   barcode [4,420]
```

```
##    barcode             chain count
##    <chr>               <chr> <int>
##  1 >AAACCTGAGACCTTTG-1 IGH       1
##  2 >AAACCTGAGACCTTTG-1 IGK       1
##  3 >AAACCTGAGCAACGGT-1 IGH       1
##  4 >AAACCTGAGCAACGGT-1 IGK       1
##  5 >AAACCTGAGTAGCGGT-1 IGH       1
##  6 >AAACCTGAGTAGCGGT-1 IGK       1
##  7 >AAACCTGCAAACCCAT-1 IGH       1
##  8 >AAACCTGCAAACCCAT-1 IGK       1
##  9 >AAACCTGCAATCACAC-1 IGH       1
## 10 >AAACCTGCAATCACAC-1 IGK       1
## # … with 8,742 more rows

# I have all conditions here

occurance_of_each_chain <- function(Final_data)
  {
  result <- Final_data %>%
    group_by(barcode, chain) %>%
    filter(chain %in% c('IGK','IGH','IGL')) %>%
    summarize(count=n())
  result <- unite(result, 'result_chain', count, chain, remove=F,
sep='')
  result <- summarize(result, type=paste(result_chain,
collapse='_'), count=sum(count))
  result
}
occurance_of_each_chain(Final_data)

## # A tibble: 4,420 x 3
##    barcode             type      count
##    <chr>               <chr>     <int>
##  1 >AAACCTGAGACCTTTG-1 1IGH_1IGK     2
##  2 >AAACCTGAGCAACGGT-1 1IGH_1IGK     2
##  3 >AAACCTGAGTAGCGGT-1 1IGH_1IGK     2
##  4 >AAACCTGCAAACCCAT-1 1IGH_1IGK     2
##  5 >AAACCTGCAATCACAC-1 1IGH_1IGK     2
##  6 >AAACCTGCACTAGTAC-1 1IGH_1IGL     2
##  7 >AAACCTGCAGCCTTGG-1 1IGH_1IGK     2
##  8 >AAACCTGCATCTACGA-1 1IGH_1IGK     2
##  9 >AAACCTGGTTATCCGA-1 1IGH_1IGK     2
## 10 >AAACCTGGTTGATTCG-1 1IGH_1IGK     2
## # … with 4,410 more rows

# I added whichever condition that I want from data

Condition <- function (result)
  {

  many_conditions = c(
```

```
  '1IGL', '1IGH', '1IGK', '1IGH_1IGK_1IGL' , '1IGH_1IGL' ,
'1IGH_1IGK', '1IGH_2IGK', 'IGH_2IGL')
  result %>%
  filter(type %in% many_conditions) %>%
  group_by(type) %>%
  summarise(total=n())
}
Condition(occurance_of_each_chain(Final_data))
```

```
## # A tibble: 7 x 2
##   type           total
##   <chr>          <int>
## 1 1IGH              30
## 2 1IGH_1IGK       3208
## 3 1IGH_1IGK_1IGL   141
## 4 1IGH_1IGL        220
## 5 1IGH_2IGK        326
## 6 1IGK             209
## 7 1IGL              14
```