Hm04_final

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

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

```
library(tidyr)
```

```
# read the data and look at the structure
data <- read.csv("/home/sedreh/ITMO/semester2/Statistic-R/4/data-cleaning.csv", heade
r=TRUE)
summary(data)
```

```
##                barcode       is_cell
##  AGGGTGACATGAAGTA-1:   16   False:25753
##  TGAGCCGGTGACAAAT-1:   16   True :23337
##  CCTTCCCAGCCTTGAT-1:   15
##  GTCAAGTTCTCCAGGG-1:   15
##  AGAGCGAGTTTCCACC-1:   14
##  ATGAGGGAGTACGACG-1:   14
##  (Other)           :49000
##                       contig_id     high_confidence     length
##  AAACCTGAGAAAGTGG-1_contig_1:    1   False:10113     Min.   : 250.0
##  AAACCTGAGACCTTTG-1_contig_1:    1   True :38977     1st Qu.: 492.0
##  AAACCTGAGACCTTTG-1_contig_2:    1                   Median : 500.0
##  AAACCTGAGACCTTTG-1_contig_3:    1                   Mean   : 521.4
##  AAACCTGAGACCTTTG-1_contig_4:    1                   3rd Qu.: 560.0
##  AAACCTGAGACCTTTG-1_contig_5:    1                   Max.   :1161.0
##  (Other)                    :49084
##      chain             v_gene          d_gene          j_gene
##  IGK    :29043   None     :11412   None   :40503   IGKJ2  :12314
##  IGH    :12091   IGKV6-17 : 5443   IGHD1-1: 2793   IGKJ1  : 5924
##  Multi  : 4356   IGKV4-51 : 4282   IGHD2-4: 1418   IGKJ4  : 4554
##  IGL    : 3021   IGKV6-20 : 2957   IGHD2-6: 1373   None   : 4322
##  TRA    :  384   IGKV10-96: 2557   IGHD2-3:  857   IGKJ5  : 3332
##  TRG    :  110   (Other)  :22437   (Other): 2141   IGHJ3  : 3206
##  (Other):   85   NA's     :    2   NA's   :    5   (Other):15438
##      c_gene      full_length   productive               cdr3
##  IGKC   :27520   False:13547   False: 1241   None          :16235
##  IGHM   :10495   True :35543   None :16297   CQQHYSTPYTF   : 4791
##  IGHD   : 3491                 True :31552   CQQWSGYPYTF   : 3694
##  None   : 2654                               CGQSYSYPFTF   : 2712
##  IGLC2  : 1949                               CQQGNTLPPTF   : 1975
##  IGLC1  : 1604                               CVRPYSNYWYFDVW:  720
##  (Other): 1377                               (Other)       :18963
##                                     cdr3_nt          reads
##  None                                   :16235   Min.   :    12
##  TGTCAGCAACATTATAGTACTCCGTACACGTTC       : 4742   1st Qu.:   166
##  TGCCAGCAGTGGAGTGGTTACCCATACACGTTC       : 3685   Median :   250
##  TGTGGACAGAGTTACAGCTATCCATTCACGTTC       : 2711   Mean   :  1045
##  TGCCAACAGGGTAATACGCTTCCTCCGACGTTC       : 1964   3rd Qu.:   599
##  TGTGTGAGACCTTATAGTAACTACTGGTACTTCGATGTCTGG:  720   Max.   :180397
##  (Other)                                :19033
##       umis              raw_clonotype_id              raw_consensus_id
##  Min.   :    1.0   None         :25763   None                  :39600
##  1st Qu.:    1.0   clonotype3   :   17   clonotype1_consensus_1:    5
##  Median :    1.0   clonotype3887:   16   clonotype2_consensus_1:    4
##  Mean   :   10.5   clonotype751 :   16   clonotype3_consensus_1:    4
##  3rd Qu.:    4.0   clonotype1596:   15   clonotype3_consensus_2:    4
##  Max.   :13527.0   clonotype3160:   15   clonotype4_consensus_1:    4
##                    (Other)      :23248   (Other)               : 9469
```

```
#looking at data
head(data)
```

barcode	is_cell	contig_id	high_confidence	length	c
<fctr>	<fctr>	<fctr>	<fctr>	<int>	<

	barcode <fctr>	is_cell <fctr>	contig_id <fctr>	high_confidence <fctr>	length <int>	c <</fctr>
1	AAACCTGAGACCTTTG-1	True	AAACCTGAGACCTTTG-1_contig_1	True	577	I
2	AAACCTGAGACCTTTG-1	True	AAACCTGAGACCTTTG-1_contig_2	True	503	I
3	AAACCTGAGACCTTTG-1	True	AAACCTGAGACCTTTG-1_contig_3	False	348	I
4	AAACCTGAGACCTTTG-1	True	AAACCTGAGACCTTTG-1_contig_4	False	373	I
5	AAACCTGAGACCTTTG-1	True	AAACCTGAGACCTTTG-1_contig_5	True	711	I
6	AAACCTGAGACCTTTG-1	True	AAACCTGAGACCTTTG-1_contig_6	True	655	I

6 rows | 1-8 of 19 columns

```
#for our project we need just 5 column of this data
chosen_column <- data %>%
  select ("barcode","contig_id", "v_gene", "d_gene", "j_gene", "productive")
chosen_column
```

barcode <fctr>	contig_id <fctr>	v_gene <fctr>	d_gene <fctr>	j_ge… <fctr>	product <fctr>
AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1_contig_1	IGKV4-79	None	IGKJ4	True
AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1_contig_2	None	None	None	None
AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1_contig_3	None	None	IGKJ1	None
AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1_contig_4	None	*NA*	IGKJ5	None
AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1_contig_5	IGHV5-17	IGHD2-8	IGHJ4	True
AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1_contig_6	IGHV2-5	IGHD2-4	IGHJ4	False
AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1_contig_7	None	None	IGLJ1	None
AAACCTGAGCAACGGT-1	AAACCTGAGCAACGGT-1_contig_1	None	None	IGHJ1	None
AAACCTGAGCAACGGT-1	AAACCTGAGCAACGGT-1_contig_2	IGKV4-70	None	IGKJ4	True

barcode <fctr>	contig_id <fctr>	v_gene <fctr>	d_gene <fctr>	j_ge… <fctr>	product <fctr>
AAACCTGAGCAACGGT-1	AAACCTGAGCAACGGT-1_contig_3	IGKV4-70	None	IGKJ4	True

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 … 1000 Next

```
#In contig_id column we need separate cell(AAATTTGGGCCAAATTGGG) from contig
split_contig <- separate (chosen_column, contig_id, c("contig", "id"))
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 49090 rows [1,
## 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
head(split_contig)
```

	barcode <fctr>	contig <chr>	id <chr>	v_gene <fctr>	d_gene <fctr>	j_gene <fctr>	productive <fctr>
1	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG	1	IGKV4-79	None	IGKJ4	True
2	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG	1	None	None	None	None
3	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG	1	None	None	IGKJ1	None
4	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG	1	None	*NA*	IGKJ5	None
5	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG	1	IGHV5-17	IGHD2-8	IGHJ4	True
6	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG	1	IGHV2-5	IGHD2-4	IGHJ4	False

6 rows

```
#My gole was spliting "contig_id" column like this: "AAACCTGAGACCTTTG-1_contig_1" =
 "AAACCTGAGACCTTTG" and "-1_contig_1" so i will do it by separate function
separate_contig <- separate(chosen_column,
         col = "contig_id",
         into = c("contig", "id"),
         sep = "_")
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 49090 rows [1,
## 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
head(separate_contig)
```

	barcode <fctr>	contig <chr>	id <chr>	v_gene <fctr>	d_gene <fctr>	j_ge… <fctr>	productive <fctr>
1	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	IGKV4-79	None	IGKJ4	True
2	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	None	None	None	None
3	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	None	None	IGKJ1	None
4	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	None	*NA*	IGKJ5	None
5	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	IGHV5-17	IGHD2-8	IGHJ4	True
6	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	IGHV2-5	IGHD2-4	IGHJ4	False

6 rows

```
#In data we have "NA" values for genes' column that shows non_productive chains!! so
 we should delete them
sum(is.na(separate_contig))
```

```
## [1] 7
```

```
na <-na.omit(separate_contig)
head(na)
```

	barcode <fctr>	contig <chr>	id <chr>	v_gene <fctr>	d_gene <fctr>	j_ge… <fctr>	productive <fctr>
1	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	IGKV4-79	None	IGKJ4	True
2	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	None	None	None	None
3	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	None	None	IGKJ1	None
5	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	IGHV5-17	IGHD2-8	IGHJ4	True
6	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	IGHV2-5	IGHD2-4	IGHJ4	False
7	AAACCTGAGACCTTTG-1	AAACCTGAGACCTTTG-1	contig	None	None	IGLJ1	None

6 rows

```
#Also we need to know number of cells that shows with unique barcodes! here we have m
any duplicate! so should extract just unique ones!
number_of_cells <- separate_contig %>%
    distinct(barcode) %>%
    count()
  number_of_cells$n
```

```
## [1] 25396
```