Análise exploratória

Jean Resende

Contexto

Agrupamento das tabelas

Iniciei o agrupamento das sequências setando os diretorios que se encontravam as tabelas e gerando as listas que conterão as tabelas.

```
library(data.table)

# diretorios contendo os arquivos
diretorio_tcrbcr_sequencias <- "../../../Projetos/Bigdata/BigData/BigData/repertorio_tcrbc

diretorio_tcrbcr_clones <- "../04_anlytics/00_clones/data/"

# listas para armazenar as tabelas
lista_tcrbcr_sequencias <- list()
lista_tcrbcr_clones <- list()
```

Em seguida, aplico um loop importando as tabelas para o ambiente R, gero uma variável com o nome das amostras e junto as tabelas da lista para um única tabela.

```
# loop para importar as tabelas
arquivos_sequencias <- list.files(path = diretorio_tcrbcr_sequencias,
                                  pattern = "\\.tsv$", full.names = TRUE)

for (arquivo in arquivos_sequencias) {
  tabela <- fread(arquivo, sep = "\t")
  tabela$sample_id <- gsub("_report.tsv", "", basename(arquivo))
  lista_tcrbcr_sequencias[[basename(arquivo)]] <- tabela
}
```

```r
# combina todas as tabelas em uma unica tabela
tcrbcr_sequencias <- rbindlist(lista_tcrbcr_sequencias, idcol = "fonte_tabelas")
```

```r
# loop para importar as tabelas
arquivos_clones <- list.files(path = diretorio_tcrbcr_clones,
                                  pattern = "\\.tsv$", full.names = TRUE)

for (arquivo in arquivos_clones) {
  tabela <- fread(arquivo, sep = "\t")
  tabela$sample_id <- gsub("_report.tsv", "", basename(arquivo))
  lista_tcrbcr_clones[[basename(arquivo)]] <- tabela
}

# combina todas as tabelas em uma unica tabela
tcrbcr_clones <- rbindlist(lista_tcrbcr_clones, idcol = "fonte_tabelas")
```

Após ter construído a tabela para as sequencias e para os clones, salvei estes objetos e tabelas e limpei o ambiente de trabalho.

```r
save(tcrbcr_sequencias, file = "tcrbcr_sequencias.RData")
save(tcrbcr_clones, file = "tcrbcr_clones.RData")

write.csv(tcrbcr_sequencias, file = "tcrbcr_sequencias.csv")
write.csv(tcrbcr_clones, file = "tcrbcr_clones.csv")
```

Após salvar os objetos e tabelas, limpo o ambiente de trabalho, para então iniciar a análise exploratória dos dados.

```r
objetos <- ls()

for (objeto  in objetos) {
  if(objeto %in% c("tcrbcr_clones", "tcrbcr_sequencias") == FALSE){
    rm(list = objeto)
  }
}
```

Exploração dos dados

Começo importando para o R os metadados para auxiliar na exploração dos dados de TCR e BCR.

```
load("../../../Projetos/Bigdata/BigData/BigData/repertorio_tcrbcr_acc/data/metadata.RData"

load("../../../Projetos/Bigdata/BigData/BigData/repertorio_tcrbcr_acc/data/coldataACC.RDat

#load("tcgaACC_pre_processed.RData")
```

```
library(TCGAbiolinks)

ACC_clinical <- GDCquery_clinic("TCGA-ACC")
```

```
idx_barcode <- ACC_clinical$submitter_id %in%
                              substr(metadata$barcode, 1,12)
# generos
table(ACC_clinical$gender[idx_barcode])
```

```
female   male
    47     29
```

```
# idade media
mean(ACC_clinical$age_at_index[idx_barcode])
```

```
[1] 46.60526
```

```
range(ACC_clinical$age_at_index[idx_barcode])
```

```
[1] 14 77
```

```
# steroid
table(metadata$steroid)
```

```
HSP LSP
 45  30
```

```
sum(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,2) == "IG"])
```

```
[1] 159663
```

```
sum(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,2) == "TR"])
```

```
[1] 1485
```

```
sum(tcrbcr_clones$`count`[substr(tcrbcr_clones$V, 1,2) == "IG"])
```

```
[1] 159663
```

```
sum(tcrbcr_clones$`count`[substr(tcrbcr_clones$V, 1,2) == "TR"])
```

```
[1] 1485
```

```
sum(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "IGH"])
```

```
[1] 43684
```

```
range(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "IGH"])
```

```
[1]    1 7131
```

```
sum(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "IGK"])
```

```
[1] 50139
```

```
  range(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "IGK"])
```

```
[1]    1 7924
```

```
  sum(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "IGL"])
```

```
[1] 65840
```

```
  range(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "IGL"])
```

```
[1]     1 15022
```

```
  sum(tcrbcr_clones$count[substr(tcrbcr_clones$V, 1,3) == "IGH"])
```

```
[1] 43684
```

```
  range(tcrbcr_clones$count[substr(tcrbcr_clones$V, 1,3) == "IGH"])
```

```
[1]    1 7861
```

```
  sum(tcrbcr_clones$count[substr(tcrbcr_clones$V, 1,3) == "IGK"])
```

```
[1] 50139
```

```
  range(tcrbcr_clones$count[substr(tcrbcr_clones$V, 1,3) == "IGK"])
```

```
[1]    1 8953
```

```
  sum(tcrbcr_clones$count[substr(tcrbcr_clones$V, 1,3) == "IGL"])
```

```
[1] 65840
```

```
  range(tcrbcr_clones$count[substr(tcrbcr_clones$V, 1,3) == "IGL"])
```

```
[1]     1 15719
```

```
  sum(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "TRA"])
```

```
[1] 400
```

```
  range(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "TRA"])
```

```
[1]  1 67
```

```
  sum(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "TRB"])
```

```
[1] 1015
```

```
  range(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "TRB"])
```

```
[1]   1 139
```

```
  sum(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "TRD"])
```

```
[1] 18
```

```
  range(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "TRD"])
```

```
[1]  1 14
```

```
  sum(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "TRG"])
```

```
[1] 52
```

```
  range(tcrbcr_sequencias$`#count`[substr(tcrbcr_sequencias$V, 1,3) == "TRG"])
```

```
[1] 1 7
```

```
  sum(tcrbcr_clones$`count`[substr(tcrbcr_clones$V, 1,3) == "TRA"])
```

```
[1] 400
```

```
  range(tcrbcr_clones$`count`[substr(tcrbcr_clones$V, 1,3) == "TRA"])
```

```
[1]  1 67
```

```
  sum(tcrbcr_clones$`count`[substr(tcrbcr_clones$V, 1,3) == "TRB"])
```

```
[1] 1015
```

```
  range(tcrbcr_clones$`count`[substr(tcrbcr_clones$V, 1,3) == "TRB"])
```

```
[1]   1 147
```

```
  sum(tcrbcr_clones$`count`[substr(tcrbcr_clones$V, 1,3) == "TRD"])
```

```
[1] 18
```

```
  range(tcrbcr_clones$`count`[substr(tcrbcr_clones$V, 1,3) == "TRD"])
```

```
[1]  1 15
```

```
  sum(tcrbcr_clones$`count`[substr(tcrbcr_clones$V, 1,3) == "TRG"])
```

```
[1] 52
```

```
range(tcrbcr_clones$`count`[substr(tcrbcr_clones$V, 1,3) == "TRG"])
```

```
[1] 1 7
```

```
data <- tcrbcr_sequencias
data$type <- ifelse(substr(data$V, 1,2) == "IG", "BCR",
                    ifelse(substr(data$V, 1,2) == "TR","TCR",NA))

head(data)
```

```
                                                fonte_tabelas #count frequency
                                                        <char>  <int>     <num>
1: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003_report.tsv      7      0.28
2: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003_report.tsv      3      0.12
3: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003_report.tsv      3      0.12
4: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003_report.tsv      3      0.60
5: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003_report.tsv      2      0.08
6: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003_report.tsv      2      0.08
                                              CDR3nt           CDR3aa
                                              <char>           <char>
1:          TGCTGCTCATATGCAGGTAGTACCACTTTCGCGGTATTT    CCSYAGSTTFAVF
2:               TGTCAACAGGCTTACAGTCCCCCTGAGACGTTC      CQQAYSPPETF
3:               TGTCAGCAGTATGGTACCTCACCTGAAATGTTC      CQQYGTSPEMF
4: TGTGCCAGCAGCGTAGACCGGACAGGAGGGGACTATGGCTACACCTTC CASSVDRTGGDYGYTF
5:               TGTCAAAAGTATGACAGTGTCCCGCTCACTTTC      CQKYDSVPLTF
6:                  TGCATGCAAACTCTACAAAGGGAGACGTTC       CMQTLQRETF
            V        D          J      C         cid cid_full_length
       <char>   <char>     <char> <char>      <char>           <int>
1: IGLV2-23*02        .   IGLJ2*01   IGLC   assemble6               0
2: IGKV1-12*01        .   IGKJ1*01   IGKC  assemble46               0
3: IGKV3-20*01        .   IGKJ1*01   IGKC  assemble49               0
4:    TRBV9*01 TRBD1*01 TRBJ1-2*01   TRBC  assemble38               0
5: IGKV1-37*01        .   IGKJ4*01   IGKC  assemble59               0
6: IGKV2-28*01        .   IGKJ1*01   IGKC assemble102               0
                                      sample_id   type
                                         <char> <char>
1: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003    BCR
2: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003    BCR
3: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003    BCR
4: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003    TCR
```

```
5: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003    BCR
6: 130723_UNC9-SN296_0386_BC2E4WACXX_ACTTGA_L003    BCR
```