Análise exploratória de sequências CDR3

Primeira iteração

Matheus Cardoso

Jun 18, 2020

Contents

Introdução

Nesse documento será feita uma análise explotarória dos dados advindos do software `attila`.

Métodos

Processamento dos dados

Para começar a análise, eu carrego os dados de um arquivo binário que foi previamente salvo. Esse arquivo `rds` foi gerado por um script em `R` que está no meu fork do `attila`.

A baixo apresento um resumo do dataframe.

```r
library(tidyverse)
library(magrittr)

cdr <- read_rds("./data/binary/isaura_compressed.rds")
cdr %<>% mutate(type = factor(case_when(
                        str_detect(file, "Final") ~ "final",
                        str_detect(file, "Initial") ~ "initial"))) %>%
         ungroup() %>%
         group_by(file) %>%
         mutate(cdrp = quantity/sum(quantity)) %>%
         select(cdr3, type, cdrp, everything()) %>%
         arrange(-cdrp, -quantity) %>%
         ungroup()

dim(cdr)
```

```
## [1] 846376     40
```

```
names(cdr)
```

```
##  [1] "cdr3"      "type"      "cdrp"      "quantity"  "length"    "MW"       
##  [7] "AV"        "IP"        "flex"      "gravy"     "SSF_Helix" "SSF_Turn" 
## [13] "SSF_Sheet" "n_A"       "n_C"       "n_D"       "n_E"       "n_F"      
## [19] "n_G"       "n_H"       "n_I"       "n_K"       "n_L"       "n_M"      
## [25] "n_N"       "n_P"       "n_Q"       "n_R"       "n_S"       "n_T"      
## [31] "n_V"       "n_W"       "n_Y"       "aliphatic" "aromatic"  "neutral"  
## [37] "positive"  "negative"  "invalid"   "file"
```

```
knitr::kable(head(cdr))
```

cdr3	type	cdrp	quantity	length	MW	AV	IP	flex	gravy	SSF_Helix
GEESEIFGVVKY	initial	1.0000000	1	12	1356.4761	0.1667	4.2527	0.7529	-0.1333	0.4167
GEESEIFGVVKY	initial	1.0000000	1	12	1356.4761	0.1667	4.2527	0.7529	-0.1333	0.4167
FLVEVK	final	0.9629992	714166	6	733.8950	0.1667	6.0014	0.7018	1.2667	0.6667
FLVEVK	final	0.7156274	254505	6	733.8950	0.1667	6.0014	0.7018	1.2667	0.6667
DGVAVAGLDY	final	0.7025474	6481	10	979.0413	0.1000	4.0500	0.7231	0.6700	0.4000
DGVAVAGLDY	final	0.7025474	6481	10	979.0413	0.1000	4.0500	0.7231	0.6700	0.4000

```
summary(cdr)
```

```
##      cdr3                type             cdrp             quantity
##  Length:846376      final  : 82769   Min.   :1.30e-06   Min.   :     1.0
##  Class :character   initial:763607   1st Qu.:2.60e-06   1st Qu.:     1.0
##  Mode  :character                    Median :7.90e-06   Median :     1.0
##                                      Mean   :7.44e-05   Mean   :    10.8
##                                      3rd Qu.:3.07e-05   3rd Qu.:     4.0
##                                      Max.   :1.00e+00   Max.   :714166.0
##      length            MW                AV               IP
##  Min.   : 1.00   Min.   :  75.07   Min.   :0.0000   Min.   : 4.050
##  1st Qu.:10.00   1st Qu.:1078.18   1st Qu.:0.1364   1st Qu.: 4.050
##  Median :12.00   Median :1365.55   Median :0.2143   Median : 4.197
##  Mean   :12.06   Mean   :1391.10   Mean   :0.2200   Mean   : 4.983
##  3rd Qu.:14.00   3rd Qu.:1664.78   3rd Qu.:0.3000   3rd Qu.: 5.567
##  Max.   :32.00   Max.   :3702.13   Max.   :1.0000   Max.   :12.000
##       flex            gravy           SSF_Helix         SSF_Turn
##  Min.   :0.5670   Min.   :-4.5000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.7237   1st Qu.:-1.1583   1st Qu.:0.2667   1st Qu.:0.2000
##  Median :0.7437   Median :-0.6789   Median :0.3333   Median :0.2857
##  Mean   :0.7439   Mean   :-0.6256   Mean   :0.3464   Mean   :0.2961
##  3rd Qu.:0.7637   3rd Qu.:-0.1357   3rd Qu.:0.4286   3rd Qu.:0.3846
##  Max.   :0.9110   Max.   : 4.5000   Max.   :1.0000   Max.   :1.0000
##    SSF_Sheet           n_A             n_C               n_D
##  Min.   :0.0000   Min.   :0.000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:0.0714   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:1.000
##  Median :0.1333   Median :0.000   Median :0.00000   Median :1.000
##  Mean   :0.1501   Mean   :0.657   Mean   :0.02241   Mean   :1.516
##  3rd Qu.:0.2222   3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:2.000
##  Max.   :1.0000   Max.   :6.000   Max.   :2.00000   Max.   :8.000
##       n_E              n_F              n_G             n_H
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :2.000   Median :0.0000
##  Mean   :0.3984   Mean   :0.6725   Mean   :1.632   Mean   :0.2009
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :6.0000   Max.   :5.0000   Max.   :8.000   Max.   :4.0000
##       n_I              n_K              n_L             n_M
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.000   Median :0.0000
##  Mean   :0.3133   Mean   :0.1378   Mean   :0.549   Mean   :0.1766
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :5.0000   Max.   :4.0000   Max.   :6.000   Max.   :4.0000
##       n_N              n_P             n_Q              n_R
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.000   Median :0.0000   Median :0.000
##  Mean   :0.2924   Mean   :0.576   Mean   :0.1771   Mean   :0.533
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:1.000
##  Max.   :4.0000   Max.   :7.000   Max.   :5.0000   Max.   :6.000
##       n_S             n_T              n_V              n_W
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :1.068   Mean   :0.4342   Mean   :0.6591   Mean   :0.3986
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :8.000   Max.   :6.0000   Max.   :7.0000   Max.   :4.0000
##       n_Y           aliphatic         aromatic         neutral
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 1.000   1st Qu.: 3.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 1.000   Median : 4.000   Median : 2.000   Median : 2.000
##  Mean   : 1.645   Mean   : 4.563   Mean   : 2.716   Mean   : 1.994
##  3rd Qu.: 2.000   3rd Qu.: 6.000   3rd Qu.: 4.000   3rd Qu.: 3.000
##  Max.   :11.000   Max.   :17.000   Max.   :14.000   Max.   :10.000
##     positive         negative        invalid      file
##  Min.   :0.0000   Min.   :0.000   Min.   :0   Length:846376
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0   Class :character
##  Median :1.0000   Median :2.000   Median :0   Mode  :character
##  Mean   :0.8717   Mean   :1.914   Mean   :0
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:0
##  Max.   :7.0000   Max.   :9.000   Max.   :0
```

Isolando apenas as sequências CDR3 enriquecidas

Como é possível perceber pelos dados acima mostrados, temos muitas reads no dataframe. Entretanto, nosso interesse por agora é nas sequências que foram enriquecidas após várias etapas de seleção. Para isso, nós precisaremos criar um subset do dataframe inicial, contendo apenas CDR3s que apresentam alto percentual de predominância em seu respectivo arquivo de leitura.

Vou mostrar um exemplo do que quero dizer:

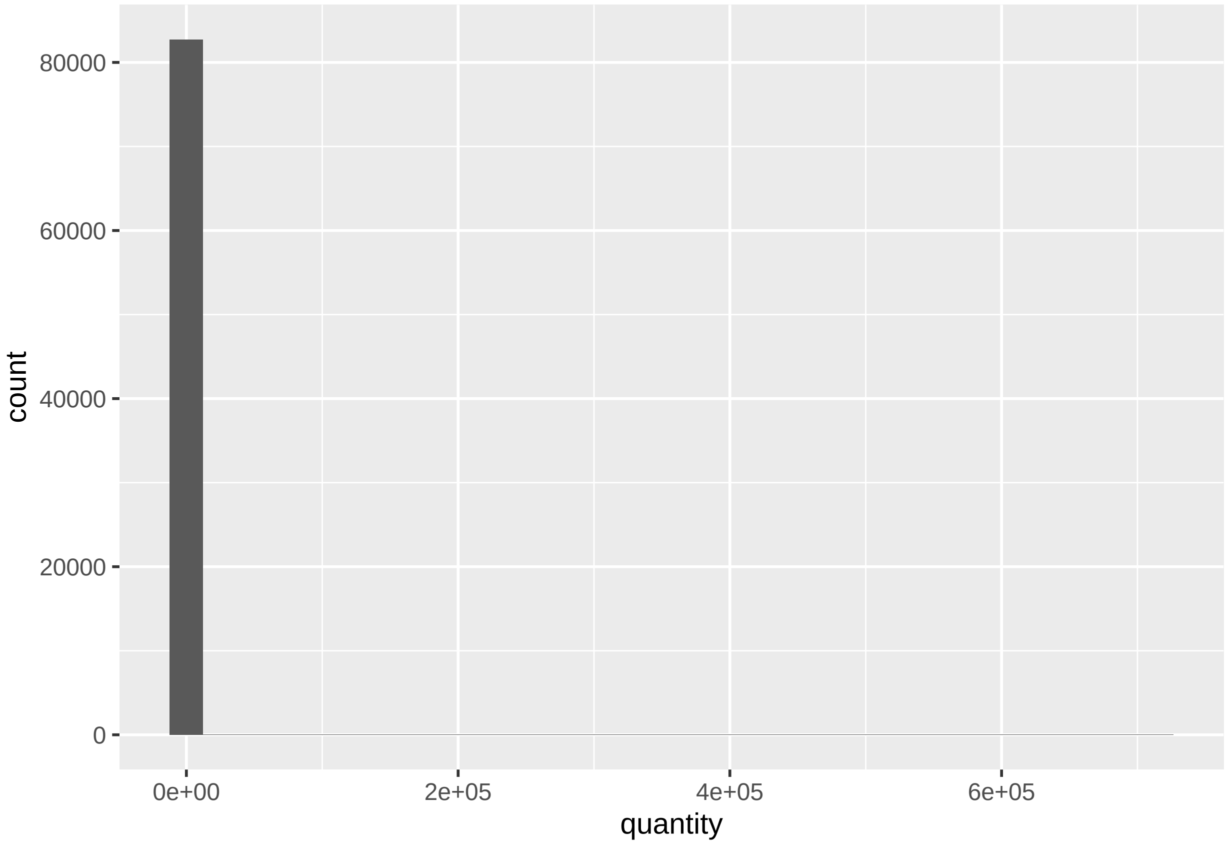
```r
cdr %>%
      select(cdr3, type, cdrp, quantity, file) %>%
      head() -> exemplo_unico_cdr

knitr::kable(exemplo_unico_cdr)
```

cdr3	type	cdrp	quantity	file
GEESEIFGVVKY	initial	1.0000000	1	mariajac_isaura_2_2_isaura_H_0eX4g_L_0bX4c_VH_InitialRoun
GEESEIFGVVKY	initial	1.0000000	1	mariajac_isaura_2_2_isaura_H_0eX4h_L_0bX4d_VH_InitialRour
FLVEVK	final	0.9629992	714166	mariajac_Isaura_Pd2_140819_R0xR5_b_VH_FinalRound_R5b_V
FLVEVK	final	0.7156274	254505	mariajac_Isaura_Pd2_140819_R0xR4_b_VH_FinalRound_R4b_V
DGVAVAGLDY	final	0.7025474	6481	mariajac_anteriores_isaura_1_isaura_HR01eXHR41h_LR01aXLR41
DGVAVAGLDY	final	0.7025474	6481	mariajac_anteriores_isaura_1_isaura_HR01eXHR41h_LR01aXLR41

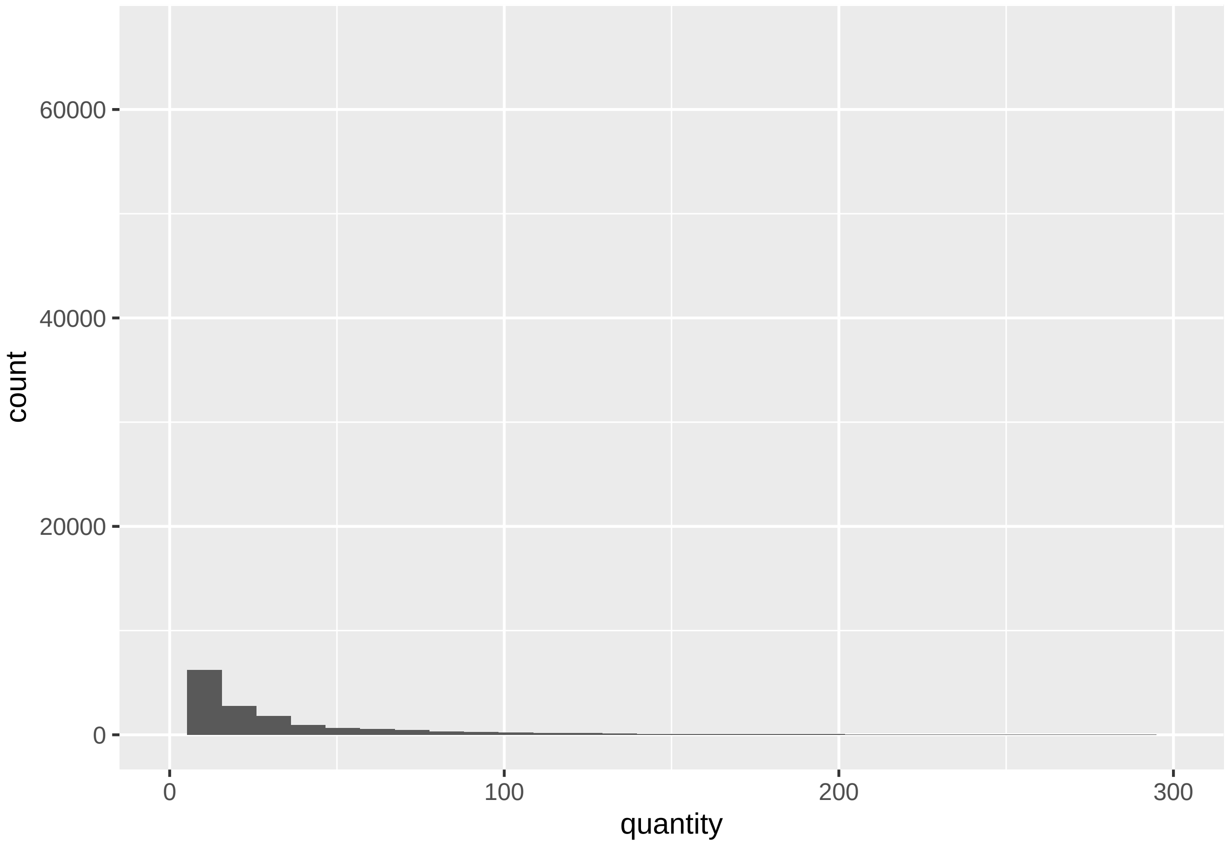
Como é possível observar, nas duas primeras linhas temos uma mesma sequência, que apresenta um percentual de 100% predôminancia em seu respectivo arquivo de leiura. (coluna `cdrp` - cdr percentage, variando de 0 a 1). Porém, observamos também que a mesma sequência aparece nesse arquivo somente uma vez. Ou seja, esses dois primeiros arquivos contém só uma leitura, e, portanto, seu percentual de predominância será de 100%. Isso, por outro lado, não reflete enriquecimento de CDR3, e, portanto, nós precisamos remover esses casos.

Pensando em como fazer a seleção dessas sequências enriquecidas, fiz algumas análises:

```r
ggplot(filter(cdr, type == "final")) +
  geom_histogram(aes(quantity))
```



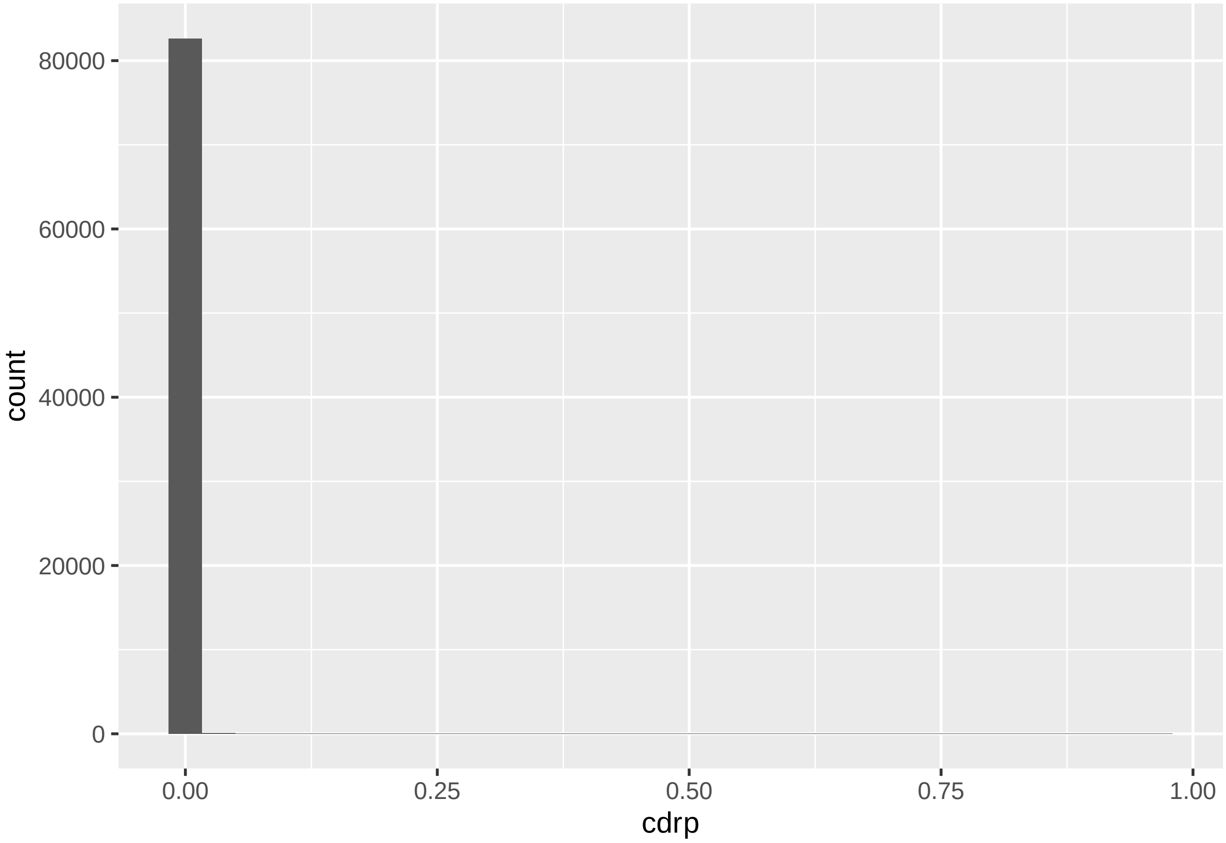
```
ggplot(filter(cdr, type == "final")) +
  geom_histogram(aes(quantity)) +
  xlim(0, 300)
```



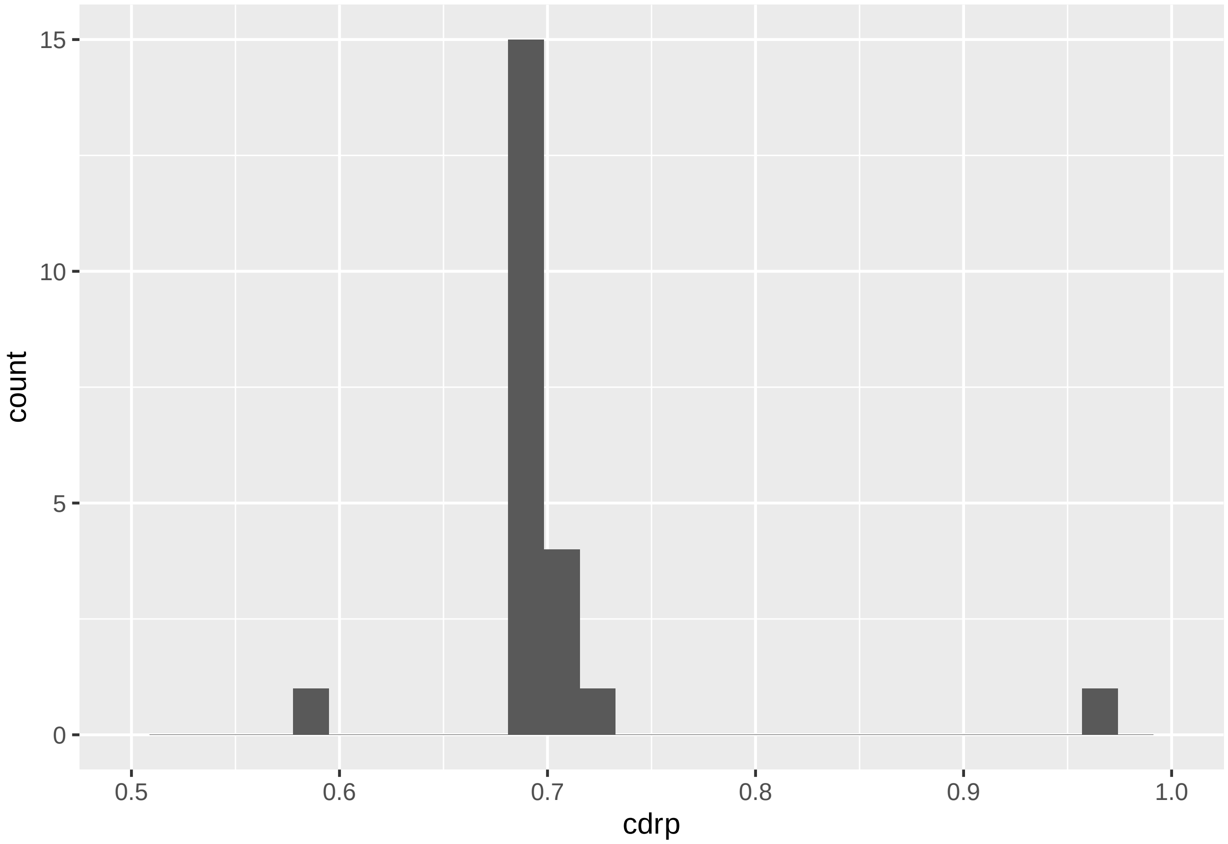
```r
cdr %>%
        filter(type == "final") %>%
        mutate(level = case_when(
                  quantity <= 300 ~ "quantity <= 300",
                  TRUE ~ "quantity > 300")) %>%
        group_by(level) %>%
        summarise("Number of CDR3 sequences" = n()) -> cdr_quantity_comparison_1

knitr::kable(cdr_quantity_comparison_1)
```

level	Number of CDR3 sequences
quantity <= 300	82055
quantity > 300	714

```r
ggplot(filter(cdr, type == "final")) +
  geom_histogram(aes(cdrp))
```



```
ggplot(filter(cdr, type == "final")) +
  geom_histogram(aes(cdrp)) +
  xlim(0.5, 1)
```



```
cdr %>%
        filter(type == "final") %>%
        mutate(level = case_when(
                  cdrp <= 0.3 ~ "cdrp <= 0.3",
                  TRUE ~ "cdrp > 0.3")) %>%
        group_by(level) %>%
        summarise("Percentage" = n()) -> cdr_cdrp_comparison_1

knitr::kable(cdr_cdrp_comparison_1, caption = "Percentage of prevalence of CDR3 sequence")
```

Table 4: Percentage of prevalence of CDR3 sequence

level	Percentage
cdrp <= 0.3	82746
cdrp > 0.3	23

```
cdr %>%
        filter(type == "final") %>%
        mutate(level = case_when(
                  cdrp < 0.5 ~ "cdrp < 0.5",
                  TRUE ~ "cdrp > 0.5")) %>%
        group_by(level) %>%
        summarise("Percentage" = n()) -> cdr_cdrp_comparison_2
```

```r
knitr::kable(cdr_cdrp_comparison_2, caption = "Percentage of prevalence of CDR3 sequence")
```

Table 5: Percentage of prevalence of CDR3 sequence

level	Percentage
cdrp < 0.5	82747
cdrp > 0.5	22

Como é possível notar, temos 23 sequências de CDR3 que apresentam prevalência maior que 30% em arquivos de leitura individual, e 22 se considerarmos 50% de prevalência.

Para termos noção do que isso significa, vejamos o seguinte:

```r
cdr$file %>% unique() %>% length() -> total_arquivos_leitura

filter(cdr, type == "final")$file %>% unique() %>% length() -> total_arquivos_leitura_final_read

tibble(
  "Arquivo de leitura" = c("Todos (Inicial + Final)", "Apenas Final", "Final com CDR3 prevalência >= 50
  "Quantidade de Arquivos" = c(total_arquivos_leitura, total_arquivos_leitura_final_read, cdr_cdrp_compa
) %>% knitr::kable()
```

Arquivo de leitura	Quantidade de Arquivos
Todos (Inicial + Final)	63
Apenas Final	31
Final com CDR3 prevalência >= 50%	22

E, para mostrar todos os arquivos com prevalência maior que 50%:

```r
cdr %>%
        filter(type == "final" & cdrp >= 0.5) %>%
        select(cdr3, cdrp, quantity, file) %>%
        knitr::kable()
```

cdr3	cdrp	quantity	file
FLVEVK	0.9629992	714166	mariajac_Isaura_Pd2_140819_R0xR5_b_VH_FinalRound_R5b_VH_S10_L0
FLVEVK	0.7156274	254505	mariajac_Isaura_Pd2_140819_R0xR4_b_VH_FinalRound_R4b_VH_S9_L00
DGVAVAGLDY	0.7025474	6481	mariajac_anteriores_isaura_1_isaura_HR01eXHR41h_LR01aXLR41c_VH_Fi
DGVAVAGLDY	0.7025474	6481	mariajac_anteriores_isaura_1_isaura_HR01eXHR41h_LR01aXLR41d_VH_Fi
DGVAVAGLDY	0.7025474	6481	mariajac_anteriores_isaura_1_isaura_HR01eXHR41h_LR01bXLR41c_VH_Fi
DGVAVAGLDY	0.7025474	6481	mariajac_anteriores_isaura_1_isaura_HR01eXHR41h_LR01bXLR41d_VH_Fi
DGVAVAGLDY	0.6955451	11866	mariajac_anteriores_isaura_1_isaura_HR01eXHR41g_LR01aXLR41c_VH_Fi
DGVAVAGLDY	0.6955451	11866	mariajac_anteriores_isaura_1_isaura_HR01eXHR41g_LR01aXLR41d_VH_Fi
DGVAVAGLDY	0.6955451	11866	mariajac_anteriores_isaura_1_isaura_HR01eXHR41g_LR01bXLR41c_VH_Fi
DGVAVAGLDY	0.6955451	11866	mariajac_anteriores_isaura_1_isaura_HR01eXHR41g_LR01bXLR41d_VH_Fi
DGVAVAGLDY	0.6911960	14540	mariajac_isaura_2_2_isaura_H_0eX4g_L_0aX4c_VH_FinalRound_VHR42g_
DGVAVAGLDY	0.6911960	14540	mariajac_isaura_2_2_isaura_H_0eX4g_L_0aX4d_VH_FinalRound_VHR42g
DGVAVAGLDY	0.6911960	14540	mariajac_isaura_2_2_isaura_H_0eX4g_L_0bX4c_2_VH_FinalRound_VHR4

cdr3	cdrp	quantity	file
DGVAVAGLDY	0.6911960	14540	mariajac_isaura_2_2_isaura_H_0eX4g_L_0bX4c_VH_FinalRound_VHR42g_
DGVAVAGLDY	0.6911960	14540	mariajac_isaura_2_2_isaura_H_0eX4g_L_0bX4d_VH_FinalRound_VHR42g
DGVAVAGLDY	0.6872554	11416	mariajac_isaura_2_2_isaura_H_0eX4h_L_0aX4c_2_VH_FinalRound_VHR4
DGVAVAGLDY	0.6872554	11416	mariajac_isaura_2_2_isaura_H_0eX4h_L_0aX4c_VH_FinalRound_VHR42h_
DGVAVAGLDY	0.6872554	11416	mariajac_isaura_2_2_isaura_H_0eX4h_L_0aX4d_VH_FinalRound_VHR42h
DGVAVAGLDY	0.6872554	11416	mariajac_isaura_2_2_isaura_H_0eX4h_L_0bX4c_VH_FinalRound_VHR42h
DGVAVAGLDY	0.6872554	11416	mariajac_isaura_2_2_isaura_H_0eX4h_L_0bX4d_2_VH_FinalRound_VHR4
DGVAVAGLDY	0.6872554	11416	mariajac_isaura_2_2_isaura_H_0eX4h_L_0bX4d_VH_FinalRound_VHR42h
GSHNSWDS	0.5791670	369801	mariajac_Isaura_Pd2_140819_R0xR4_a_VH_FinalRound_R4a_VH_S8_L00

Portanto, eu resolvi salvar esse dataframe como aquele contendo as sequências enriquecidas.

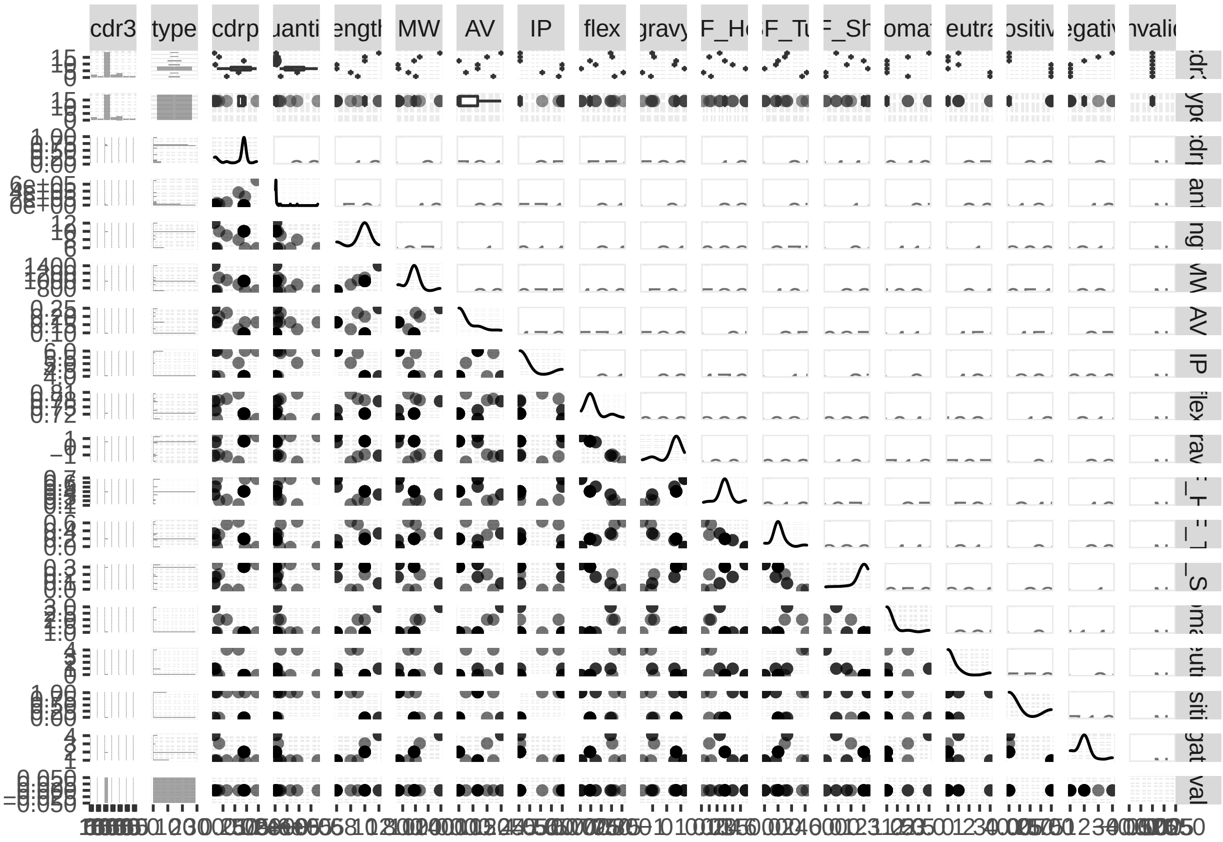
```r
cdr_rich <- cdr %>% filter(type == "final" & cdrp >= 0.5)
```

Todo o código feito a partir daqui é um rascunho

Peço perdão pela bagunça nos próximos blocos. Eu escrevi isso para me ajudar a entender os dados, sem a intenção de apresentar isso para ninguém.

Análise Exploratória

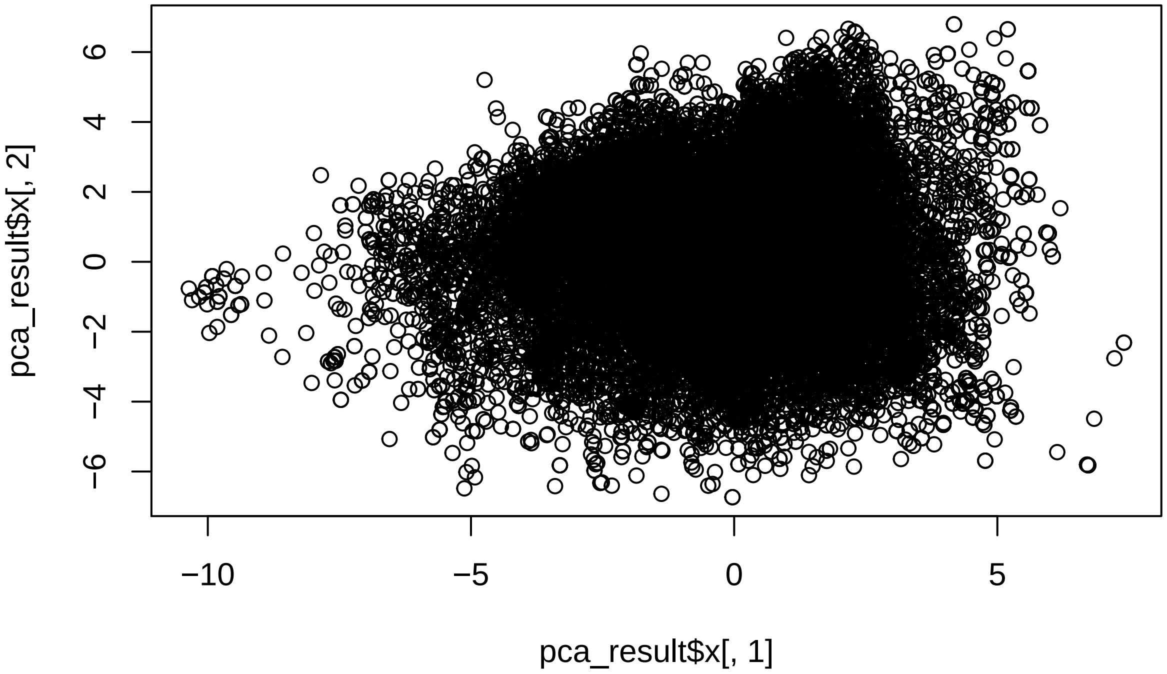
```r
cdr %>%
        ungroup() %>%
        arrange(-cdrp, type, file) %>%
        filter(quantity > 1) %>%
        filter(type == "final") -> cdr_final

cdr_final %>%
                filter(quantity > 1) %>%
                group_by(file) %>%
                slice_head(n = 1) -> cdr_enriched

library(GGally)
cdr_enriched %<>%
                select(cdr3:SSF_Sheet, aromatic:file)

cdr_enriched %>%
                ungroup() %>%
                select(-file) %>%
                ggpairs(aes(alpha = 0.4))
```



```
cdr_final %>%
            ungroup() %>%
            select(!c(cdr3, type, file, invalid)) -> cdr_final_pca

pca_result <- prcomp(cdr_final_pca, center = T, scale. = T)
summary(pca_result)
```

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5    PC6     PC7
## Standard deviation     2.2320 2.1424 1.82898 1.61396 1.52117 1.3145 1.22483
## Proportion of Variance 0.1384 0.1275 0.09292 0.07236 0.06428 0.0480 0.04167
## Cumulative Proportion  0.1384 0.2659 0.35881 0.43117 0.49544 0.5434 0.58511
##                            PC8     PC9   PC10    PC11   PC12    PC13    PC14
## Standard deviation     1.20982 1.12969 1.1144 1.05725 1.0462 0.98564 0.98080
## Proportion of Variance 0.04066 0.03545 0.0345 0.03105 0.0304 0.02699 0.02672
## Cumulative Proportion  0.62577 0.66122 0.6957 0.72677 0.7572 0.78416 0.81088
##                           PC15    PC16   PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.97047 0.95699 0.9431 0.90463 0.86672 0.86159 0.81767
## Proportion of Variance 0.02616 0.02544 0.0247 0.02273 0.02087 0.02062 0.01857
## Cumulative Proportion  0.83704 0.86248 0.8872 0.90991 0.93078 0.95140 0.96997
##                           PC22    PC23    PC24    PC25    PC26    PC27   PC28
## Standard deviation     0.72682 0.43550 0.34254 0.32517 0.26219 0.20040 0.1465
## Proportion of Variance 0.01467 0.00527 0.00326 0.00294 0.00191 0.00112 0.0006
## Cumulative Proportion  0.98465 0.98992 0.99318 0.99611 0.99802 0.99914 0.9997
##                           PC29     PC30      PC31      PC32      PC33      PC34
## Standard deviation     0.09783 1.62e-14 5.817e-15 5.189e-15 4.404e-15 3.802e-15
## Proportion of Variance 0.00027 0.00e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.00000 1.00e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
##                             PC35      PC36
## Standard deviation     3.541e-15 2.381e-15
## Proportion of Variance 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00
```

```
plot(pca_result$x[,1], pca_result$x[,2])
```



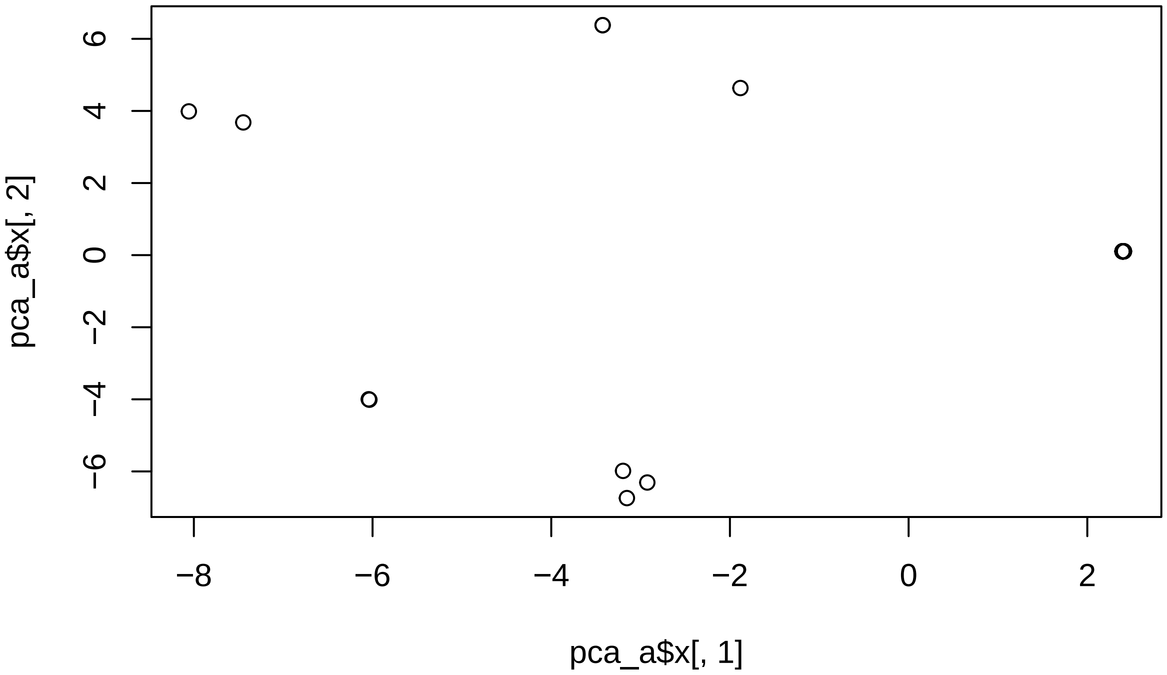
```
cdr_final_pca
```

```
## # A tibble: 30,818 x 36
##     cdrp quantity length    MW    AV    IP  flex gravy SSF_Helix SSF_Turn
##    <dbl>    <int>  <int> <dbl> <dbl> <dbl> <dbl> <dbl>     <dbl>    <dbl>
##  1 0.963   714166      6  734. 0.167  6.00 0.702  1.27     0.667      0
##  2 0.716   254505      6  734. 0.167  6.00 0.702  1.27     0.667      0
##  3 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  4 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  5 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  6 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  7 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  8 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  9 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
## 10 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
## # ... with 30,808 more rows, and 26 more variables: SSF_Sheet <dbl>, n_A <int>,
## #   n_C <int>, n_D <int>, n_E <int>, n_F <int>, n_G <int>, n_H <int>,
## #   n_I <int>, n_K <int>, n_L <int>, n_M <int>, n_N <int>, n_P <int>,
## #   n_Q <int>, n_R <int>, n_S <int>, n_T <int>, n_V <int>, n_W <int>,
## #   n_Y <int>, aliphatic <int>, aromatic <int>, neutral <int>, positive <int>,
## #   negative <int>
```

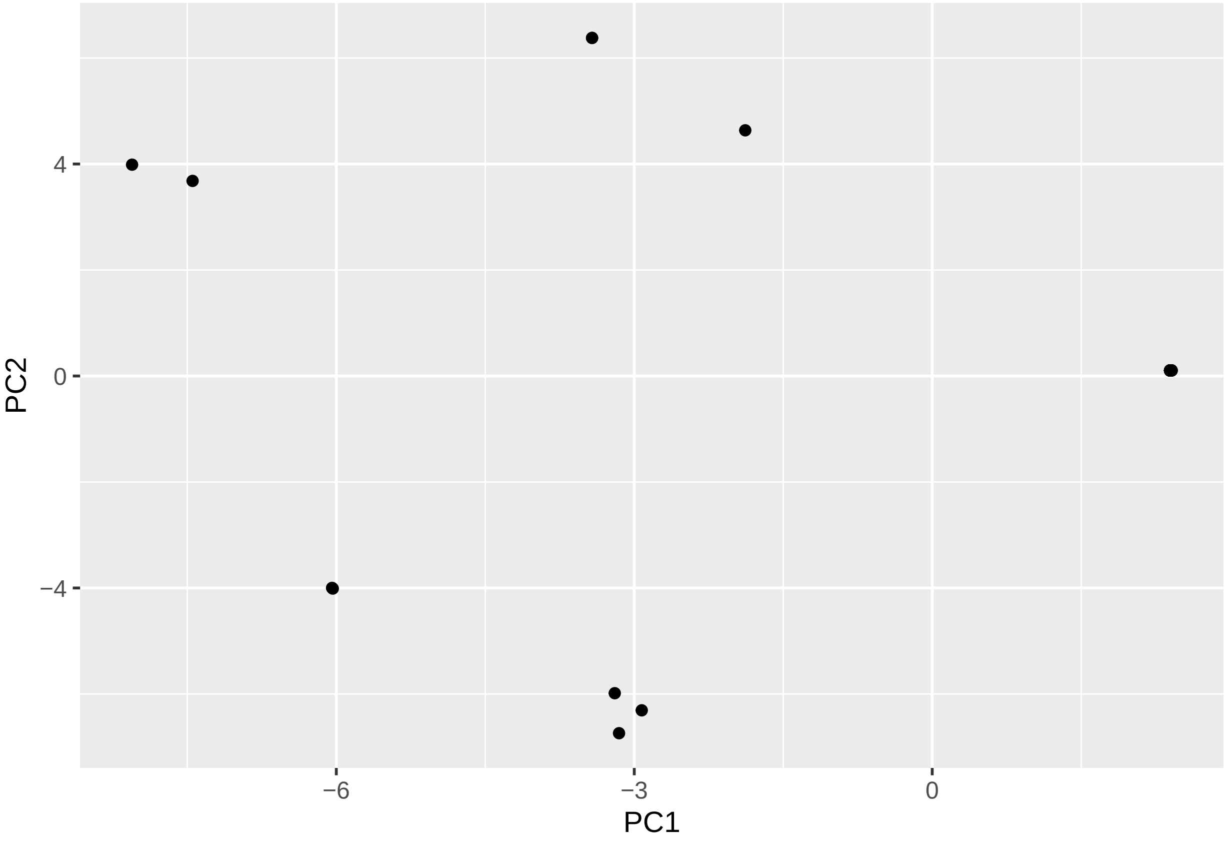
```r
cdr_final %>%
              group_by(file) %>%
              arrange(-cdrp) %>%
              slice_head(n = 1) %>%
              ungroup() %>%
              select(!c(cdr3, type, file, invalid)) %>%
              arrange(-cdrp) -> a

# in this line we remove all collumns that have variance equal to 0
# Doing this, we can apply a pca to the dataframe without erros
# credit goes to: https://stackoverflow.com/a/40317343
a <- select(a, !c(which(apply(a, 2, var)==0)))
pca_a <- prcomp(a, center = T, scale. = T)
summary(pca_a)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5    PC6     PC7
## Standard deviation     3.578 3.1940 2.0814 1.42383 1.19178 0.9347 0.56375
## Proportion of Variance 0.400 0.3188 0.1354 0.06335 0.04439 0.0273 0.00993
## Cumulative Proportion  0.400 0.7188 0.8542 0.91753 0.96192 0.9892 0.99915
##                            PC8      PC9      PC10      PC11      PC12      PC13
## Standard deviation     0.16506 1.82e-15 2.891e-16 2.891e-16 2.891e-16 2.891e-16
## Proportion of Variance 0.00085 0.00e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.00000 1.00e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
##                             PC14      PC15      PC16      PC17      PC18
## Standard deviation     2.891e-16 2.891e-16 2.891e-16 2.891e-16 2.891e-16
## Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
##                             PC19      PC20      PC21      PC22      PC23
## Standard deviation     2.891e-16 2.891e-16 2.891e-16 2.891e-16 2.891e-16
## Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
##                             PC24      PC25      PC26      PC27      PC28
## Standard deviation     2.891e-16 2.891e-16 2.891e-16 2.891e-16 2.891e-16
## Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
##                             PC29
## Standard deviation     2.891e-16
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```

```r
plot(pca_a$x[,1], pca_a$x[,2])
```



```
ggplot(as_tibble(pca_a$x)) +
  geom_point(aes(PC1, PC2))
```



```r
pca_a$x
```

```
##              PC1        PC2        PC3         PC4         PC5         PC6
##  [1,] -3.153187 -6.7396691  0.5423173  2.02558876 -3.16509933  1.12651448
##  [2,] -2.924818 -6.3081964  1.1150958  1.00898751 -1.56987430  0.22477835
##  [3,]  2.410307  0.1050559 -0.4026319 -0.12398379  0.07212154 -0.09425090
##  [4,]  2.410307  0.1050559 -0.4026319 -0.12398379  0.07212154 -0.09425090
##  [5,]  2.410307  0.1050559 -0.4026319 -0.12398379  0.07212154 -0.09425090
##  [6,]  2.410307  0.1050559 -0.4026319 -0.12398379  0.07212154 -0.09425090
##  [7,]  2.400501  0.1030796 -0.4000082 -0.11578167  0.06524087 -0.08679969
##  [8,]  2.400501  0.1030796 -0.4000082 -0.11578167  0.06524087 -0.08679969
##  [9,]  2.400501  0.1030796 -0.4000082 -0.11578167  0.06524087 -0.08679969
## [10,]  2.400501  0.1030796 -0.4000082 -0.11578167  0.06524087 -0.08679969
## [11,]  2.395003  0.1023694 -0.3978832 -0.11203533  0.06286417 -0.08337386
## [12,]  2.395003  0.1023694 -0.3978832 -0.11203533  0.06286417 -0.08337386
## [13,]  2.395003  0.1023694 -0.3978832 -0.11203533  0.06286417 -0.08337386
## [14,]  2.395003  0.1023694 -0.3978832 -0.11203533  0.06286417 -0.08337386
## [15,]  2.395003  0.1023694 -0.3978832 -0.11203533  0.06286417 -0.08337386
## [16,]  2.394928  0.1060044 -0.3918603 -0.11979058  0.07640050 -0.09021272
## [17,]  2.394928  0.1060044 -0.3918603 -0.11979058  0.07640050 -0.09021272
## [18,]  2.394928  0.1060044 -0.3918603 -0.11979058  0.07640050 -0.09021272
## [19,]  2.394928  0.1060044 -0.3918603 -0.11979058  0.07640050 -0.09021272
## [20,]  2.394928  0.1060044 -0.3918603 -0.11979058  0.07640050 -0.09021272
## [21,]  2.394928  0.1060044 -0.3918603 -0.11979058  0.07640050 -0.09021272
## [22,] -8.056297  3.9872538 -7.6565091  3.56935244  0.68136201  0.04397812
## [23,] -7.447150  3.6813651 -2.2018788 -5.56131636 -2.65614682 -0.03862617
## [24,] -1.882441  4.6356466  2.8223272 -0.01769929  1.19699610  4.33679216
## [25,] -3.196508 -5.9850602  1.8092377  0.42683642 -0.35366382 -0.28270965
## [26,] -6.035579 -4.0109123  1.0990276 -1.14920572  2.91225518 -0.52596797
## [27,] -6.042641 -3.9984501  1.1237524 -1.17303982  2.95899415 -0.54683492
## [28,] -3.424598  6.3788035  4.4488843  1.55423903 -0.66349835 -1.32778820
## [29,] -3.424598  6.3788035  4.4488843  1.55423903 -0.66349835 -1.32778820
##                PC7          PC8           PC9          PC10          PC11
##  [1,]  1.57620348  0.301175021 -1.767683e-15 -1.639314e-16 -1.665335e-16
##  [2,] -0.56361237 -0.711391980 -1.323594e-15  5.811324e-17 -3.885781e-16
##  [3,]  0.01986187 -0.044071866 -1.769581e-15  2.034505e-16  2.602085e-17
##  [4,]  0.01986187 -0.044071866 -1.769581e-15  2.034505e-16  2.602085e-17
##  [5,]  0.01986187 -0.044071866 -1.769581e-15  2.034505e-16  2.602085e-17
##  [6,]  0.01986187 -0.044071866 -1.769581e-15  2.034505e-16  2.602085e-17
##  [7,]  0.02793611 -0.002614030 -1.825092e-15  2.034505e-16  2.602085e-17
##  [8,]  0.02793611 -0.002614030 -1.825092e-15  2.034505e-16  2.602085e-17
##  [9,]  0.02793611 -0.002614030 -1.825092e-15  2.034505e-16  2.602085e-17
## [10,]  0.02793611 -0.002614030 -1.825092e-15  2.034505e-16  2.602085e-17
## [11,]  0.03044869  0.020579178 -1.825092e-15  2.034505e-16  2.602085e-17
## [12,]  0.03044869  0.020579178 -1.825092e-15  2.034505e-16  2.602085e-17
## [13,]  0.03044869  0.020579178 -1.825092e-15  2.034505e-16  2.602085e-17
## [14,]  0.03044869  0.020579178 -1.825092e-15  2.034505e-16  2.602085e-17
## [15,]  0.03044869  0.020579178 -1.825092e-15  2.034505e-16  2.602085e-17
## [16,]  0.01202754  0.020451584 -1.825092e-15  2.034505e-16  2.602085e-17
## [17,]  0.01202754  0.020451584 -1.825092e-15  2.034505e-16  2.602085e-17
## [18,]  0.01202754  0.020451584 -1.825092e-15  2.034505e-16  2.602085e-17
## [19,]  0.01202754  0.020451584 -1.825092e-15  2.034505e-16  2.602085e-17
## [20,]  0.01202754  0.020451584 -1.825092e-15  2.034505e-16  2.602085e-17
## [21,]  0.01202754  0.020451584 -1.825092e-15  2.034505e-16  2.602085e-17
## [22,] -0.28752985 -0.003268379 -7.546047e-16 -3.647256e-15  9.992007e-16
## [23,] -0.09436330  0.005718434 -1.448494e-15 -1.315788e-15  2.775558e-16
## [24,] -0.28995939 -0.017090684 -7.515689e-16 -6.019490e-16 -5.551115e-17
## [25,] -2.25341359  0.391067521 -1.989728e-15  9.462917e-16 -3.885781e-16
## [26,]  0.66659000 -0.013224714 -1.684416e-15 -1.426810e-15  6.106227e-16
## [27,]  0.60209347  0.014936875 -1.684416e-15 -1.426810e-15  6.106227e-16
## [28,]  0.11419546 -0.003391954 -1.936819e-15 -8.951173e-16  1.110223e-16
## [29,]  0.11419546 -0.003391954 -1.936819e-15 -8.951173e-16  1.110223e-16
##                PC12          PC13          PC14          PC15          PC16
##  [1,]  1.110223e-16 -5.551115e-17 -1.643650e-16 -1.665335e-16 -1.665335e-16
##  [2,]  5.551115e-16  1.665335e-16 -1.643650e-16  9.436896e-16 -6.106227e-16
##  [3,]  1.110223e-16  3.469447e-18 -6.559423e-18  1.058181e-16  6.245005e-17
##  [4,]  1.110223e-16  3.469447e-18 -6.559423e-18  1.058181e-16  6.245005e-17
##  [5,]  1.110223e-16  3.469447e-18 -6.559423e-18  1.058181e-16  6.245005e-17
##  [6,]  1.110223e-16  3.469447e-18 -6.559423e-18  1.058181e-16  6.245005e-17
##  [7,]  8.326673e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
##  [8,]  8.326673e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
##  [9,]  8.326673e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [10,]  8.326673e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [11,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [12,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [13,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [14,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [15,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [16,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [17,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [18,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [19,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [20,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [21,]  5.551115e-17  3.469447e-18 -6.559423e-18 -5.204170e-18  6.245005e-17
## [22,] -1.110223e-16  3.330669e-16 -4.419208e-16  1.110223e-16 -1.665335e-16
## [23,] -9.992007e-16 -5.551115e-16  1.131907e-16  8.326673e-16 -6.106227e-16
## [24,] -2.498002e-16 -1.665335e-16  4.566660e-16 -9.714451e-16 -2.220446e-16
## [25,]  1.110223e-16 -5.551115e-17  2.797242e-16 -3.885781e-16 -3.885781e-16
## [26,] -9.992007e-16  4.440892e-16 -3.308985e-16 -1.665335e-16 -3.885781e-16
## [27,] -9.992007e-16  3.330669e-16 -3.308985e-16 -1.665335e-16 -3.885781e-16
## [28,]  1.665335e-16  6.661338e-16 -3.712308e-16  4.440892e-16  0.000000e+00
## [29,]  1.665335e-16  6.661338e-16 -3.712308e-16  4.440892e-16  0.000000e+00
##                PC17          PC18          PC19          PC20          PC21
##  [1,] -5.551115e-17  6.106227e-16  2.220446e-16  5.551115e-17  4.163336e-17
##  [2,]  1.665335e-16  1.665335e-16  2.220446e-16 -3.885781e-16  9.298118e-16
##  [3,] -1.006140e-16  2.428613e-16 -2.151057e-16 -3.469447e-17 -5.074066e-17
##  [4,] -1.006140e-16  2.428613e-16 -2.151057e-16 -3.469447e-17 -5.074066e-17
##  [5,] -1.006140e-16  2.428613e-16 -2.151057e-16 -3.469447e-17 -5.074066e-17
##  [6,] -1.006140e-16  2.428613e-16 -2.151057e-16 -3.469447e-17 -5.074066e-17
##  [7,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
##  [8,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
##  [9,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [10,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [11,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [12,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [13,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [14,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [15,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [16,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [17,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [18,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [19,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [20,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [21,] -1.006140e-16  2.428613e-16 -1.040834e-16 -3.469447e-17 -1.062518e-16
## [22,]  5.273559e-16  3.330669e-16  2.220446e-16  1.665335e-16  3.747003e-16
## [23,]  5.273559e-16 -3.330669e-16  4.440892e-16  3.885781e-16  4.163336e-17
## [24,]  3.053113e-16 -1.387779e-16  7.771561e-16  2.359224e-16 -2.289835e-16
## [25,] -5.551115e-17 -7.216450e-16  4.440892e-16  5.551115e-17  2.636780e-16
## [26,]  1.942890e-16 -1.942890e-16  8.881784e-16  4.996004e-16  1.249001e-16
## [27,]  1.942890e-16 -1.942890e-16  8.881784e-16  4.996004e-16  1.249001e-16
## [28,]  1.665335e-16 -8.326673e-16  6.661338e-16 -5.551115e-16  4.857226e-16
## [29,]  1.665335e-16 -8.326673e-16  6.661338e-16 -5.551115e-16  4.857226e-16
##                PC22          PC23          PC24          PC25          PC26
##  [1,]  1.665335e-16  2.775558e-16  1.665335e-16 -2.220446e-16 -8.326673e-17
##  [2,]  3.885781e-16  1.054712e-15 -2.775558e-16  4.440892e-16  1.387779e-16
##  [3,]  7.979728e-17 -9.020562e-17  4.510281e-17  4.597017e-17  1.283695e-16
##  [4,]  7.979728e-17 -9.020562e-17  4.510281e-17  4.597017e-17  1.283695e-16
##  [5,]  7.979728e-17 -9.020562e-17  4.510281e-17  4.597017e-17  1.283695e-16
##  [6,]  7.979728e-17 -9.020562e-17  4.510281e-17  4.597017e-17  1.283695e-16
##  [7,]  7.979728e-17 -1.179612e-16  4.510281e-17 -9.540979e-18  1.283695e-16
##  [8,]  7.979728e-17 -1.179612e-16  4.510281e-17 -9.540979e-18  1.283695e-16
##  [9,]  7.979728e-17 -1.179612e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [10,]  7.979728e-17 -1.179612e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [11,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [12,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [13,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [14,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [15,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [16,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [17,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [18,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [19,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [20,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [21,]  7.979728e-17 -1.457168e-16  4.510281e-17 -9.540979e-18  1.283695e-16
## [22,] -5.551115e-16  3.885781e-16  5.551115e-16 -4.996004e-16 -4.718448e-16
## [23,]  1.110223e-16  1.665335e-16 -2.775558e-16  3.885781e-16 -6.938894e-16
## [24,] -2.220446e-16  1.110223e-16 -2.775558e-16 -3.885781e-16 -8.326673e-17
## [25,]  6.106227e-16  3.885781e-16  1.665335e-16 -4.440892e-16 -8.326673e-17
## [26,]  5.551115e-16  4.996004e-16  3.330669e-16 -3.885781e-16 -2.775558e-17
## [27,]  5.551115e-16  4.996004e-16  3.330669e-16 -3.885781e-16 -2.775558e-17
## [28,] -1.110223e-16  8.881784e-16  0.000000e+00  1.387779e-16 -1.665335e-16
## [29,] -1.110223e-16  8.881784e-16  0.000000e+00  1.387779e-16 -1.665335e-16
##                PC27          PC28          PC29
##  [1,]  1.110223e-16  2.498002e-16 -6.938894e-17
##  [2,]  5.551115e-16 -1.942890e-16 -5.134781e-16
##  [3,] -3.295975e-17 -5.898060e-17  1.474515e-17
##  [4,] -3.295975e-17 -5.898060e-17  1.474515e-17
##  [5,] -3.295975e-17 -5.898060e-17  1.474515e-17
##  [6,] -3.295975e-17 -5.898060e-17  1.474515e-17
##  [7,] -3.295975e-17 -5.898060e-17  7.025630e-17
##  [8,] -3.295975e-17 -5.898060e-17  7.025630e-17
##  [9,] -3.295975e-17 -5.898060e-17  7.025630e-17
## [10,] -3.295975e-17 -5.898060e-17  7.025630e-17
## [11,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [12,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [13,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [14,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [15,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [16,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [17,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [18,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [19,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [20,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [21,] -1.439820e-16 -5.898060e-17  7.025630e-17
## [22,]  0.000000e+00  2.498002e-16  3.191891e-16
## [23,] -2.775558e-16  3.608225e-16  2.636780e-16
## [24,] -4.718448e-16  1.387779e-16  2.498002e-16
## [25,]  3.330669e-16  2.498002e-16  1.526557e-16
## [26,]  2.220446e-16 -1.942890e-16 -1.804112e-16
## [27,]  2.220446e-16 -1.942890e-16 -1.804112e-16
## [28,] -2.775558e-16  3.885781e-16 -8.326673e-17
## [29,] -2.775558e-16  3.885781e-16 -8.326673e-17
```

```r
str(pca_a)
```

```
## List of 5
##  $ sdev    : num [1:29] 3.58 3.19 2.08 1.42 1.19 ...
##  $ rotation: num [1:32, 1:29] 0.186 -0.131 0.157 0.054 -0.215 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:32] "cdrp" "quantity" "length" "MW" ...
##   .. ..$ : chr [1:29] "PC1" "PC2" "PC3" "PC4" ...
##  $ center  : Named num [1:32] 5.64e-01 6.42e+04 9.34 9.65e+02 1.30e-01 ...
##   ..- attr(*, "names")= chr [1:32] "cdrp" "quantity" "length" "MW" ...
##  $ scale   : Named num [1:32] 2.58e-01 1.48e+05 1.70 1.53e+02 4.86e-02 ...
##   ..- attr(*, "names")= chr [1:32] "cdrp" "quantity" "length" "MW" ...
##  $ x       : num [1:29, 1:29] -3.15 -2.92 2.41 2.41 2.41 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:29] "PC1" "PC2" "PC3" "PC4" ...
##  - attr(*, "class")= chr "prcomp"
```

```r
pca_cdr_result <- cdr %>%
                        select(!c(cdr3, type, file, invalid)) %>%
                        prcomp(center = T, scale. = T)
summary(pca_cdr_result)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.2070 2.0517 1.83300 1.61745 1.57126 1.26983 1.25221
## Proportion of Variance 0.1353 0.1169 0.09333 0.07267 0.06858 0.04479 0.04356
## Cumulative Proportion  0.1353 0.2522 0.34556 0.41823 0.48681 0.53160 0.57516
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.1940 1.17812 1.07633 1.03880 1.02983 1.01582 1.00061
## Proportion of Variance 0.0396 0.03855 0.03218 0.02997 0.02946 0.02866 0.02781
## Cumulative Proportion   0.6148 0.65331 0.68549 0.71547 0.74493 0.77359 0.80140
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.99209 0.98029 0.95983 0.95470 0.90047 0.87291 0.85869
## Proportion of Variance 0.02734 0.02669 0.02559 0.02532 0.02252 0.02117 0.02048
## Cumulative Proportion  0.82874 0.85544 0.88103 0.90635 0.92887 0.95003 0.97052
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.75771 0.44052 0.30147 0.29571 0.24285 0.18264 0.12589
## Proportion of Variance 0.01595 0.00539 0.00252 0.00243 0.00164 0.00093 0.00044
## Cumulative Proportion  0.98646 0.99185 0.99438 0.99681 0.99845 0.99937 0.99981
##                           PC29      PC30      PC31     PC32      PC33     PC34
## Standard deviation     0.08195 1.326e-12 8.635e-14 6.31e-14 6.009e-14 3.98e-14
## Proportion of Variance 0.00019 0.000e+00 0.000e+00 0.00e+00 0.000e+00 0.00e+00
## Cumulative Proportion  1.00000 1.000e+00 1.000e+00 1.00e+00 1.000e+00 1.00e+00
##                             PC35      PC36
## Standard deviation     2.841e-14 1.485e-14
## Proportion of Variance 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00
```

```
cdr_final %>%
              group_by(file) %>%
              arrange(-cdrp) %>%
              slice_head(n = 10) %>%
              ungroup() %>%
              select(!c(cdr3, type, file, invalid)) %>%
              arrange(-cdrp) -> b

b <- select(b, !c(which(apply(b, 2, var)==0)))
b
```

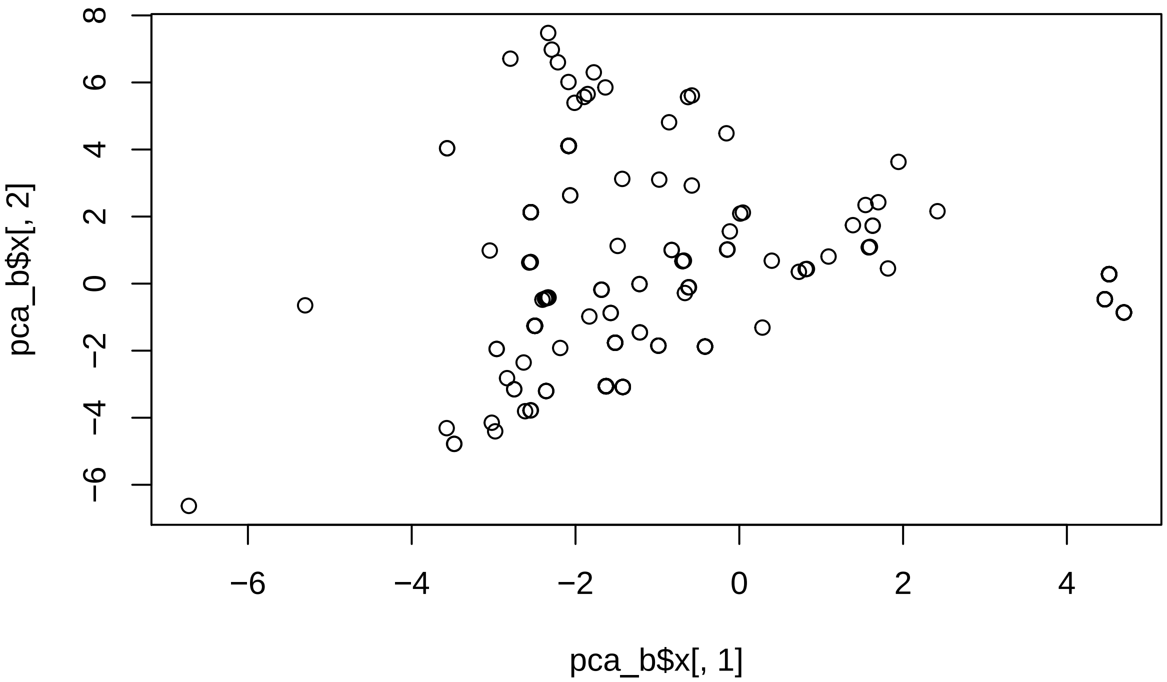
```
## # A tibble: 290 x 36
##     cdrp quantity length    MW    AV    IP  flex gravy SSF_Helix SSF_Turn
##    <dbl>    <int>  <int> <dbl> <dbl> <dbl> <dbl> <dbl>     <dbl>    <dbl>
##  1 0.963   714166      6  734. 0.167  6.00 0.702  1.27     0.667      0
##  2 0.716   254505      6  734. 0.167  6.00 0.702  1.27     0.667      0
##  3 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  4 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  5 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  6 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  7 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  8 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
##  9 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
## 10 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4        0.2
## # ... with 280 more rows, and 26 more variables: SSF_Sheet <dbl>, n_A <int>,
## #   n_C <int>, n_D <int>, n_E <int>, n_F <int>, n_G <int>, n_H <int>,
## #   n_I <int>, n_K <int>, n_L <int>, n_M <int>, n_N <int>, n_P <int>,
## #   n_Q <int>, n_R <int>, n_S <int>, n_T <int>, n_V <int>, n_W <int>,
## #   n_Y <int>, aliphatic <int>, aromatic <int>, neutral <int>, positive <int>,
## #   negative <int>
```

```
pca_b <- prcomp(b, center = T, scale. = T)
summary(pca_b)
```

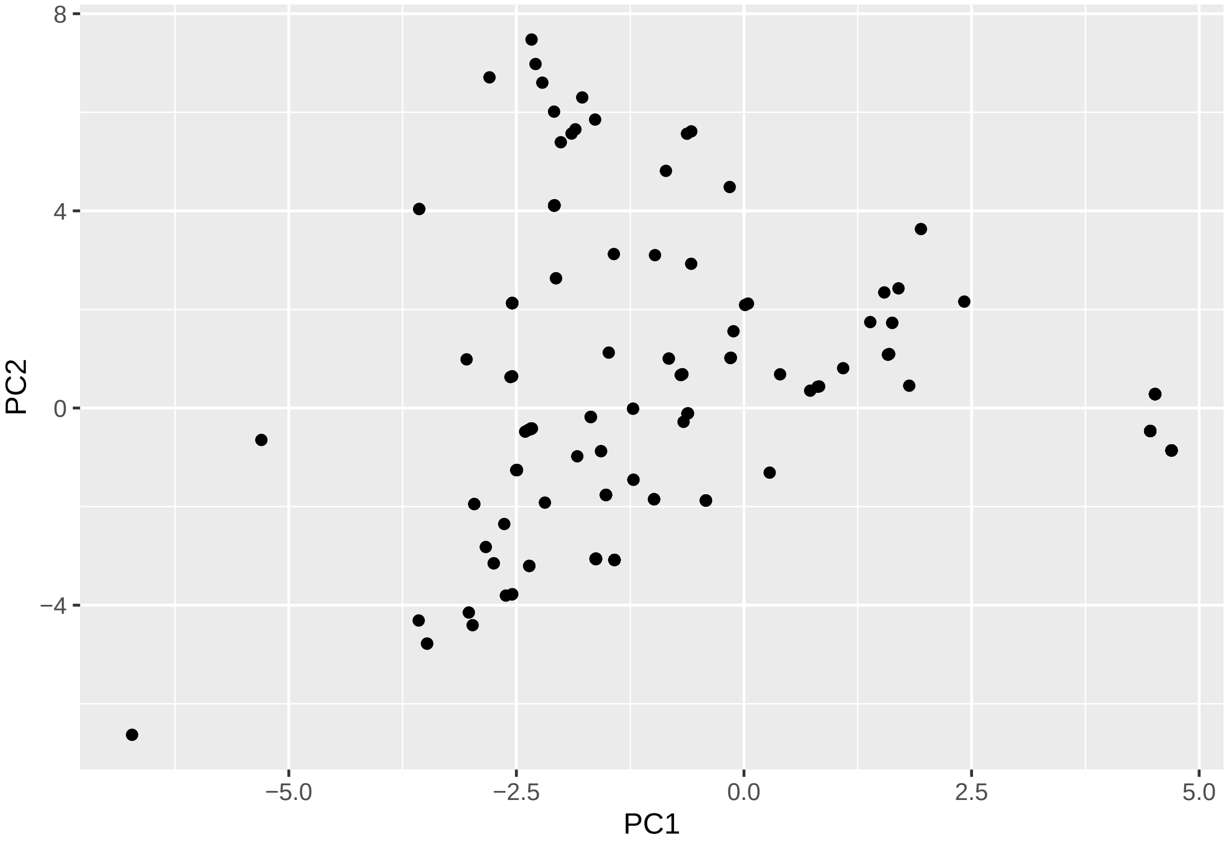
```
## Importance of components:
```

```
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     2.9161 2.4665 1.9498 1.7410 1.52362 1.35730 1.26191
## Proportion of Variance 0.2362 0.1690 0.1056 0.0842 0.06448 0.05117 0.04423
## Cumulative Proportion  0.2362 0.4052 0.5108 0.5950 0.65949 0.71066 0.75490
##                            PC8     PC9    PC10    PC11    PC12    PC13   PC14
## Standard deviation     1.14684 1.07537 1.04875 0.97586 0.82782 0.79248 0.7074
## Proportion of Variance 0.03653 0.03212 0.03055 0.02645 0.01904 0.01745 0.0139
## Cumulative Proportion  0.79143 0.82355 0.85411 0.88056 0.89959 0.91704 0.9309
##                           PC15    PC16    PC17   PC18    PC19    PC20   PC21
## Standard deviation     0.66953 0.63727 0.57337 0.5629 0.50397 0.45910 0.3981
## Proportion of Variance 0.01245 0.01128 0.00913 0.0088 0.00706 0.00585 0.0044
## Cumulative Proportion  0.94339 0.95467 0.96381 0.9726 0.97966 0.98552 0.9899
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.38050 0.30999 0.23902 0.17910 0.12711 0.09736 0.07041
## Proportion of Variance 0.00402 0.00267 0.00159 0.00089 0.00045 0.00026 0.00014
## Cumulative Proportion  0.99394 0.99661 0.99820 0.99909 0.99954 0.99980 0.99994
##                           PC29      PC30      PC31      PC32      PC33
## Standard deviation     0.04712 1.566e-15 9.467e-16 8.825e-16 6.685e-16
## Proportion of Variance 0.00006 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.00000 1.000e+00 1.000e+00 1.000e+00 1.000e+00
##                             PC34      PC35      PC36
## Standard deviation     4.817e-16 4.617e-16 2.517e-16
## Proportion of Variance 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00
```

```
plot(pca_b$x[,1], pca_b$x[,2])
```



```
ggplot(as_tibble(pca_b$x)) +
  geom_point(aes(PC1, PC2))
```



```r
summary(pca_b)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     2.9161 2.4665 1.9498 1.7410 1.52362 1.35730 1.26191
## Proportion of Variance 0.2362 0.1690 0.1056 0.0842 0.06448 0.05117 0.04423
## Cumulative Proportion  0.2362 0.4052 0.5108 0.5950 0.65949 0.71066 0.75490
##                            PC8     PC9    PC10    PC11    PC12    PC13   PC14
## Standard deviation     1.14684 1.07537 1.04875 0.97586 0.82782 0.79248 0.7074
## Proportion of Variance 0.03653 0.03212 0.03055 0.02645 0.01904 0.01745 0.0139
## Cumulative Proportion  0.79143 0.82355 0.85411 0.88056 0.89959 0.91704 0.9309
##                           PC15    PC16    PC17   PC18    PC19    PC20   PC21
## Standard deviation     0.66953 0.63727 0.57337 0.5629 0.50397 0.45910 0.3981
## Proportion of Variance 0.01245 0.01128 0.00913 0.0088 0.00706 0.00585 0.0044
## Cumulative Proportion  0.94339 0.95467 0.96381 0.9726 0.97966 0.98552 0.9899
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.38050 0.30999 0.23902 0.17910 0.12711 0.09736 0.07041
## Proportion of Variance 0.00402 0.00267 0.00159 0.00089 0.00045 0.00026 0.00014
## Cumulative Proportion  0.99394 0.99661 0.99820 0.99909 0.99954 0.99980 0.99994
##                           PC29      PC30      PC31      PC32      PC33
## Standard deviation     0.04712 1.566e-15 9.467e-16 8.825e-16 6.685e-16
## Proportion of Variance 0.00006 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.00000 1.000e+00 1.000e+00 1.000e+00 1.000e+00
##                             PC34      PC35      PC36
## Standard deviation     4.817e-16 4.617e-16 2.517e-16
## Proportion of Variance 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00
```

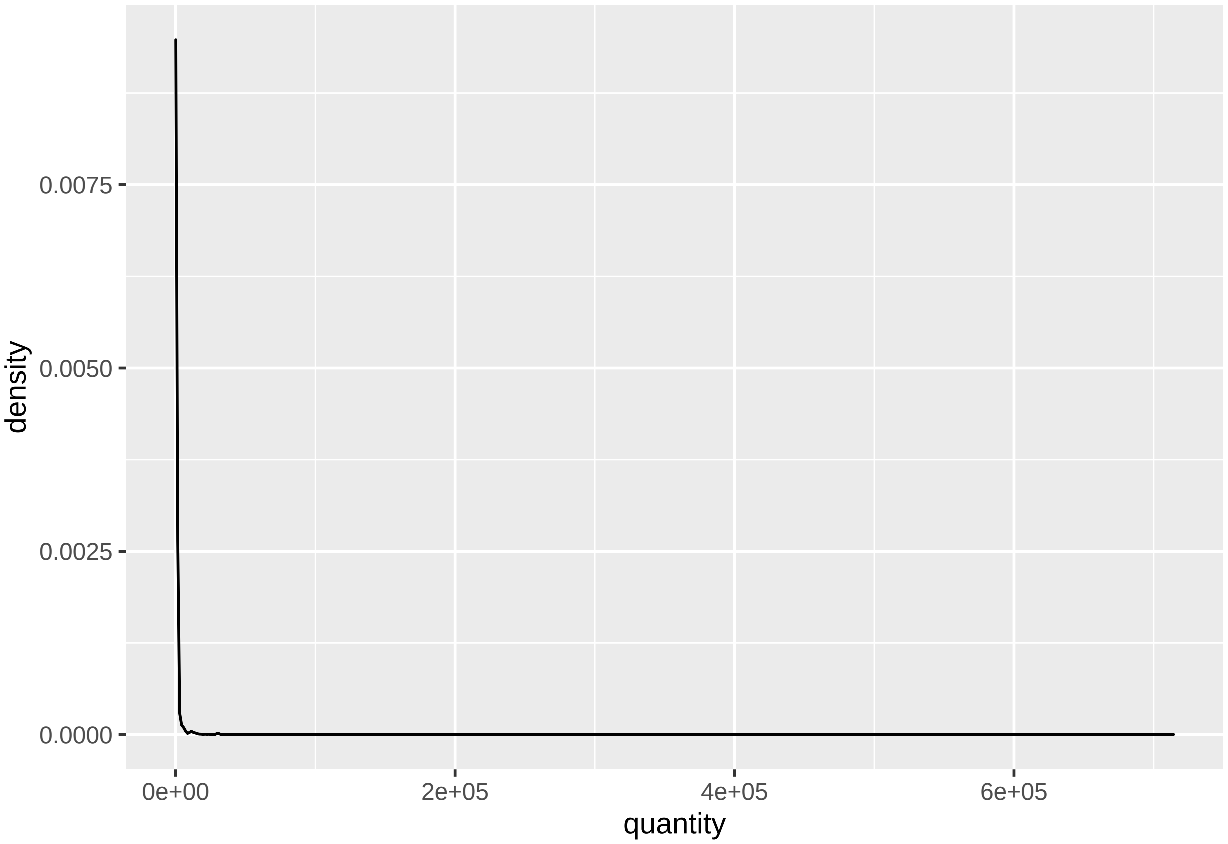
```
summary(cdr$quantity)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      1.0      1.0      1.0     10.8      4.0 714166.0
```

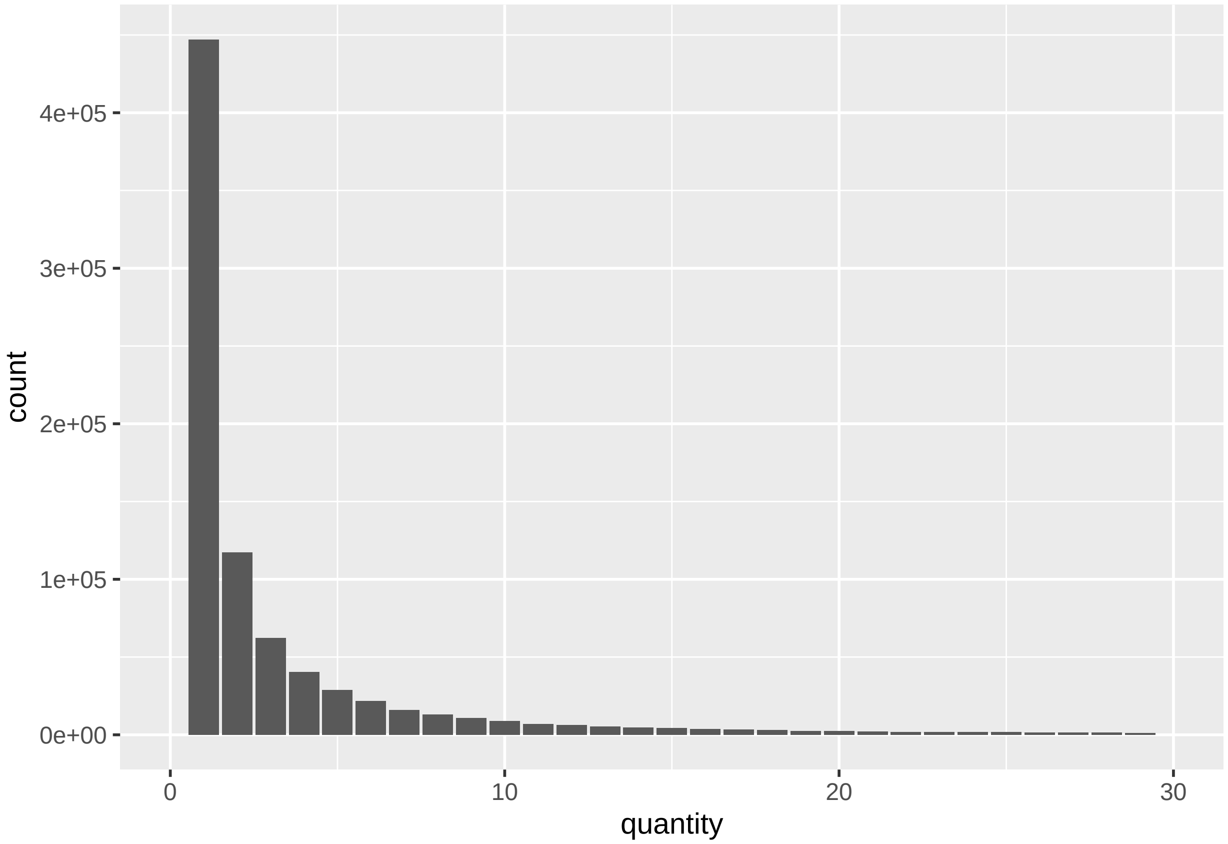
```
cdr %>%
        filter(quantity >= 100) -> a
a
```

```
## # A tibble: 5,027 x 40
##    cdr3  type   cdrp quantity length    MW    AV    IP  flex gravy SSF_Helix
##    <chr> <fct> <dbl>    <int>  <int> <dbl> <dbl> <dbl> <dbl> <dbl>     <dbl>
##  1 FLVE~ final 0.963   714166      6  734. 0.167  6.00 0.702  1.27     0.667
##  2 FLVE~ final 0.716   254505      6  734. 0.167  6.00 0.702  1.27     0.667
##  3 DGVA~ final 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4
##  4 DGVA~ final 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4
##  5 DGVA~ final 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4
##  6 DGVA~ final 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4
##  7 DGVA~ final 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4
##  8 DGVA~ final 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4
##  9 DGVA~ final 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4
## 10 DGVA~ final 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4
## # ... with 5,017 more rows, and 29 more variables: SSF_Turn <dbl>,
## #   SSF_Sheet <dbl>, n_A <int>, n_C <int>, n_D <int>, n_E <int>, n_F <int>,
## #   n_G <int>, n_H <int>, n_I <int>, n_K <int>, n_L <int>, n_M <int>,
## #   n_N <int>, n_P <int>, n_Q <int>, n_R <int>, n_S <int>, n_T <int>,
## #   n_V <int>, n_W <int>, n_Y <int>, aliphatic <int>, aromatic <int>,
## #   neutral <int>, positive <int>, negative <int>, invalid <int>, file <chr>
```

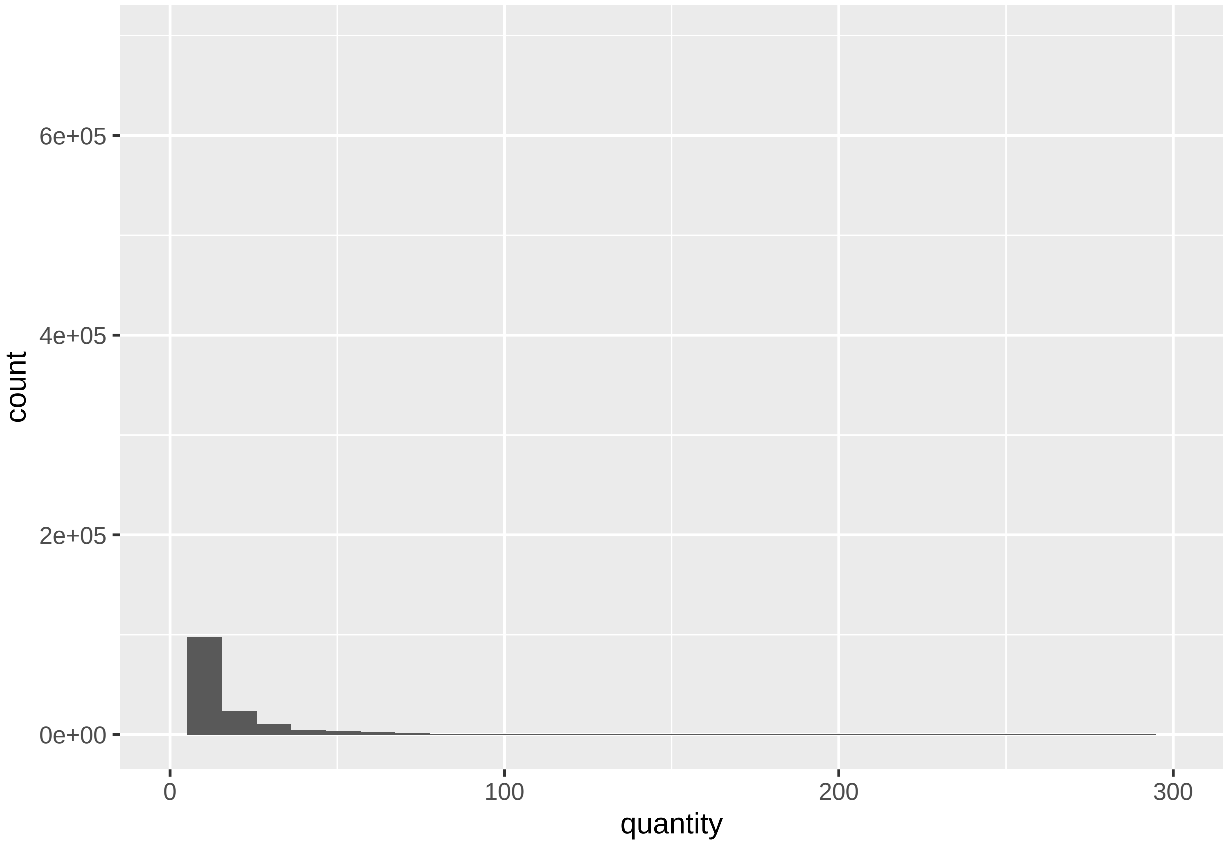
```
ggplot(a) +
  geom_density(aes(quantity))
```



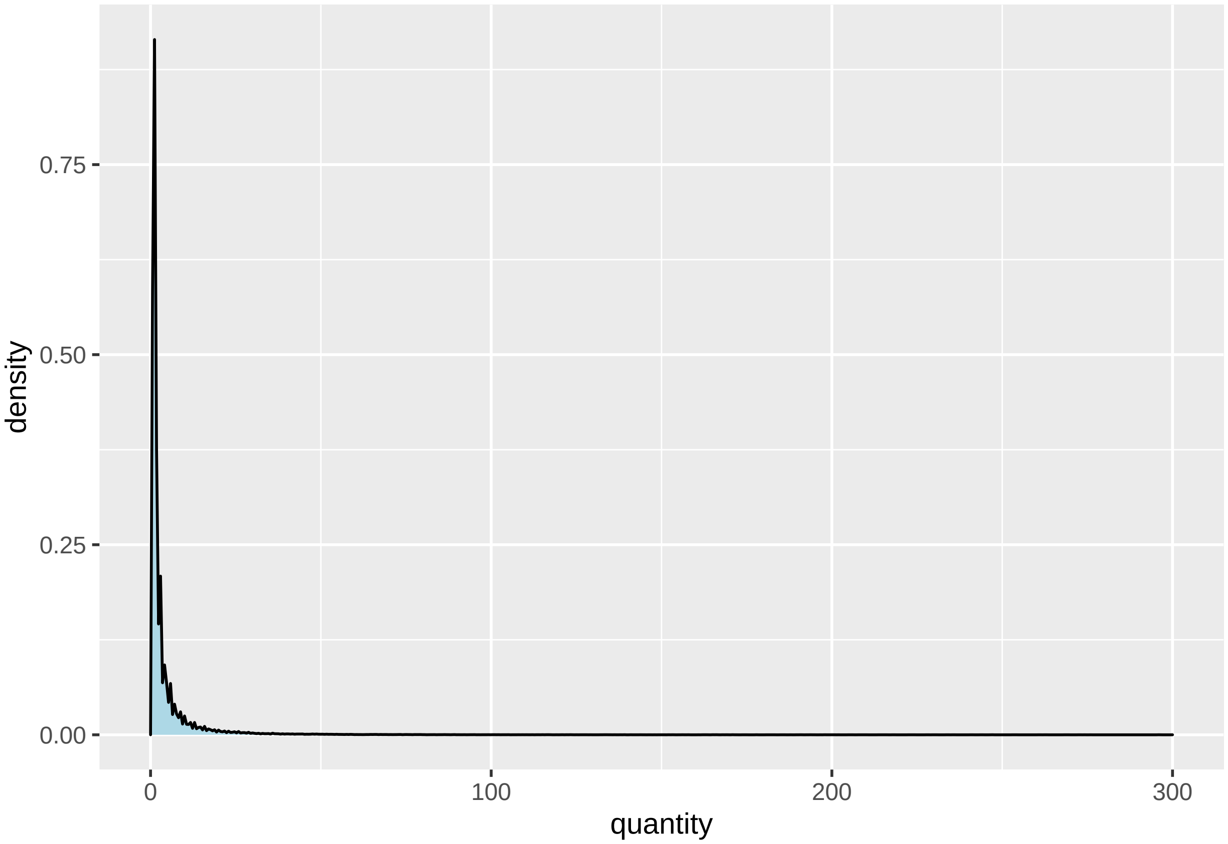
```
ggplot(cdr) +
  geom_bar(aes(quantity)) +
  xlim(0, 30)
```



```
ggplot(cdr) +
  geom_histogram(aes(quantity)) +
  xlim(0, 300)
```



```
ggplot(cdr) +
  geom_density(aes(quantity), fill = "lightblue") +
  xlim(0, 300)
```



```
quantile(cdr$quantity)
```

```
##     0%    25%    50%    75%   100%
##      1      1      1      4 714166
```

```
dim(cdr)
```

```
## [1] 846376     40
```

```
cdr %>% filter(quantity >= 1E3) %>% dim()
```

```
## [1] 555  40
```

```
cdr %>% filter(quantity >= 1E4) %>% dim()
```

```
## [1] 87 40
```

```
cdr %>% filter(quantity >= 1E5) %>% dim()
```

```
## [1]  5 40
```

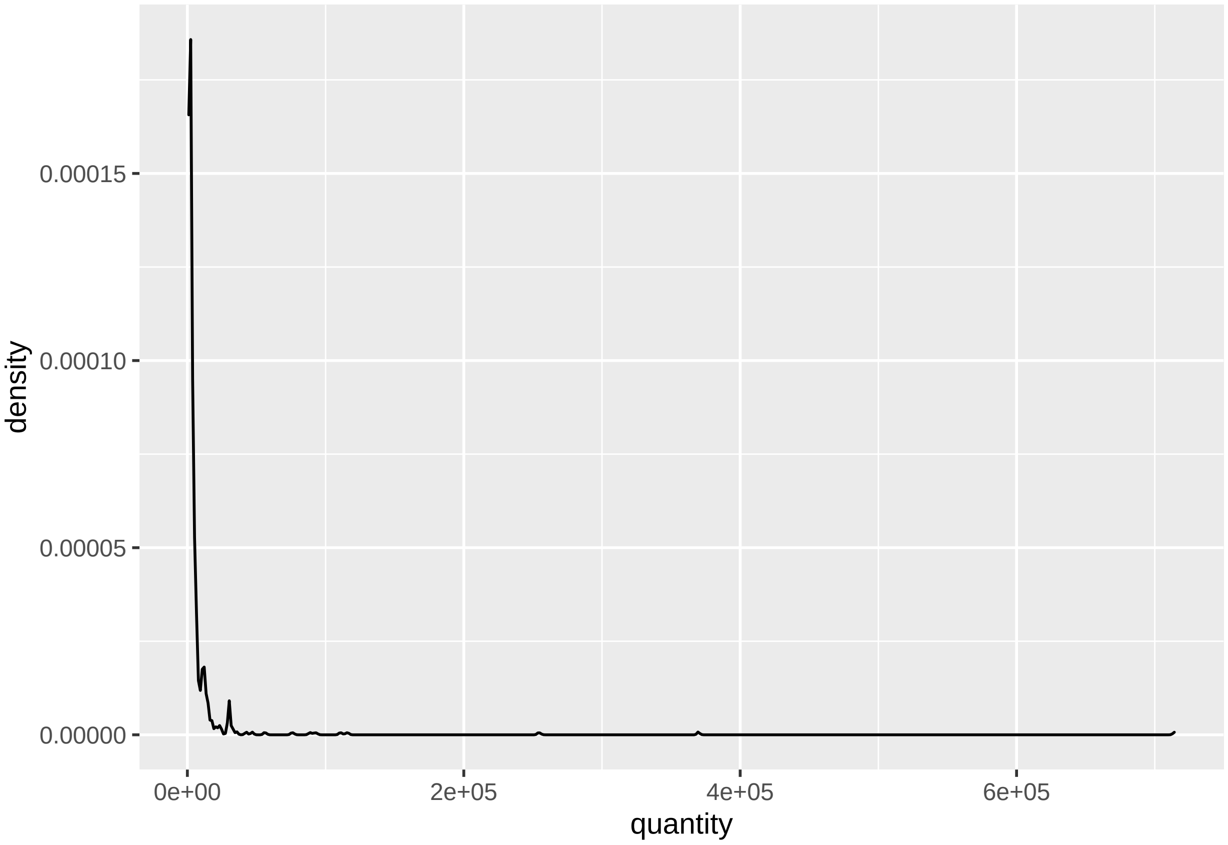
```
cdr %>% filter(quantity >= 1E3) -> b
b %>% group_by(type) %>% summarise(total = n())
```

```
## # A tibble: 2 x 2
##   type    total
##   <fct>   <int>
## 1 final     299
## 2 initial   256
```

```
b %>% group_by(type) %>% summarise(quantile = quantile(cdrp)) -> b_quantiles
b_quantiles <- add_column(b_quantiles, quantiles = rep(attr(quantile(b$quantity), "names"), 2))
knitr::kable(b_quantiles)
```

type	quantile	quantiles
final	0.0013875	0%
final	0.0044044	25%
final	0.0093961	50%
final	0.0336626	75%
final	0.9629992	100%
initial	0.0028332	0%
initial	0.0038802	25%
initial	0.0074028	50%
initial	0.0224001	75%
initial	0.0972894	100%

```
ggplot(b) +
  geom_density(aes(quantity))
```



```r
b %>%
      group_by(cdr3, type) %>%
      arrange(-cdrp)
```

```
## # A tibble: 555 x 40
## # Groups:   cdr3, type [193]
##    cdr3  type   cdrp quantity length    MW    AV    IP  flex gravy SSF_Helix
##    <chr> <fct> <dbl>    <int>  <int> <dbl> <dbl> <dbl> <dbl> <dbl>     <dbl>
##  1 FLVE~ final 0.963   714166      6  734. 0.167  6.00 0.702  1.27     0.667
##  2 FLVE~ final 0.716   254505      6  734. 0.167  6.00 0.702  1.27     0.667
##  3 DGVA~ final 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4
##  4 DGVA~ final 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4
##  5 DGVA~ final 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4
##  6 DGVA~ final 0.703     6481     10  979. 0.1    4.05 0.723  0.67     0.4
##  7 DGVA~ final 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4
##  8 DGVA~ final 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4
##  9 DGVA~ final 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4
## 10 DGVA~ final 0.696    11866     10  979. 0.1    4.05 0.723  0.67     0.4
## # ... with 545 more rows, and 29 more variables: SSF_Turn <dbl>,
## #   SSF_Sheet <dbl>, n_A <int>, n_C <int>, n_D <int>, n_E <int>, n_F <int>,
## #   n_G <int>, n_H <int>, n_I <int>, n_K <int>, n_L <int>, n_M <int>,
## #   n_N <int>, n_P <int>, n_Q <int>, n_R <int>, n_S <int>, n_T <int>,
## #   n_V <int>, n_W <int>, n_Y <int>, aliphatic <int>, aromatic <int>,
## #   neutral <int>, positive <int>, negative <int>, invalid <int>, file <chr>
```

```r
b %>%
      group_by(cdr3, type) %>%
      select(cdr3, type, cdrp, quantity) %>%
      arrange(-cdrp, -quantity) %>%
      slice_head(n = 1) %>%
      arrange(-cdrp, -quantity)
```

```
## # A tibble: 193 x 4
## # Groups:   cdr3, type [193]
##    cdr3              type   cdrp quantity
##    <chr>             <fct> <dbl>    <int>
##  1 FLVEVK            final 0.963   714166
##  2 DGVAVAGLDY        final 0.703     6481
##  3 GSHNSWDS          final 0.579   369801
##  4 RGSSSSFDY         final 0.329    92824
##  5 ELVGATYY          final 0.250    88917
##  6 DPTWRMATIGSLGTY   final 0.181   115672
##  7 DDYGPAAFDP        final 0.167    46831
##  8 FIVESK            final 0.152    42538
##  9 DRSYYDSSGYYSD     final 0.108    30233
## 10 GNDYVWGSYIEPNYFDY final 0.106    29756
## # ... with 183 more rows
```

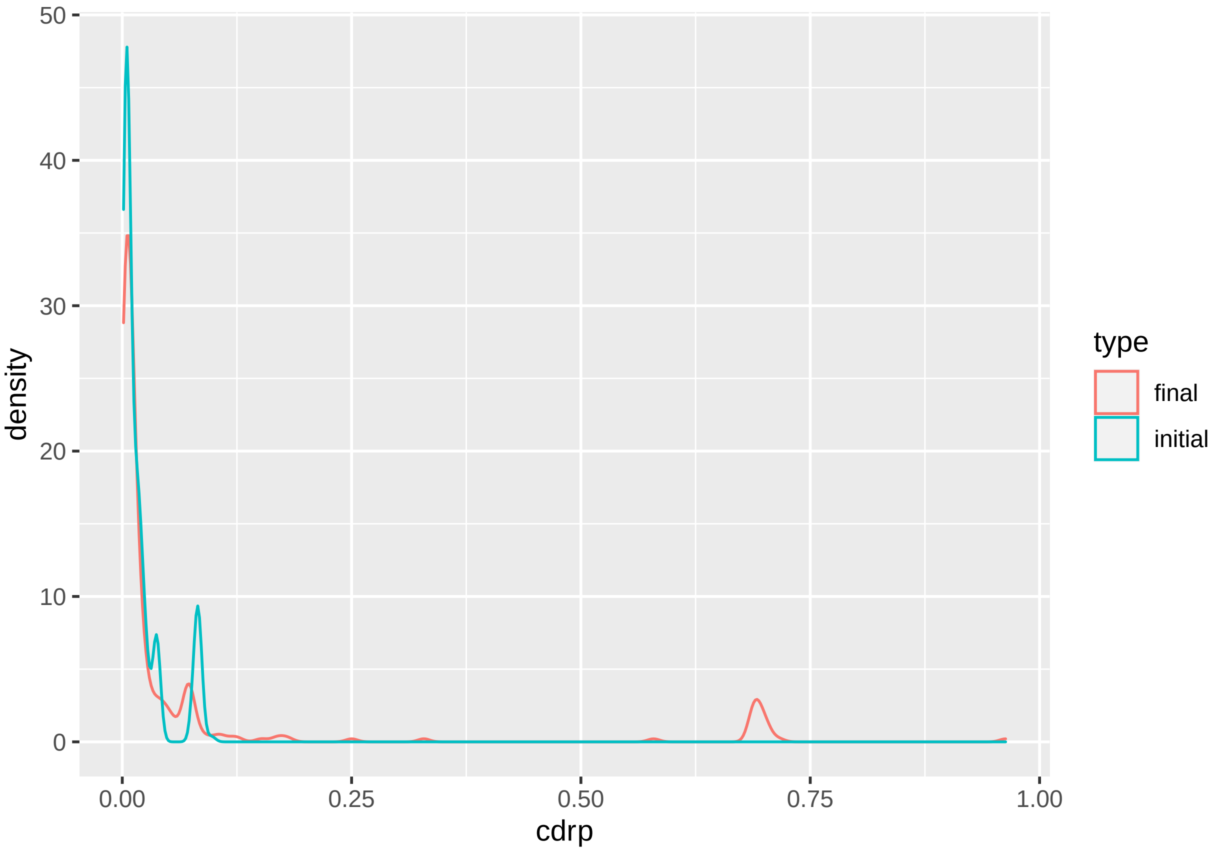
```r
b %>%
      group_by(type, cdr3) %>%
      summarise(total = n()) %>%
      arrange(-total)
```

```
## # A tibble: 193 x 3
## # Groups:   type [2]
##    type    cdr3       total
##    <fct>   <chr>      <int>
##  1 initial FIVESK        29
##  2 initial DLGIPDDY      21
##  3 initial EMWGPEY       21
##  4 initial FLVESK        21
##  5 final   DGVAVAGLDY    19
##  6 initial DGVAVAGLDY    19
##  7 final   GRWGSY        15
##  8 initial DLHWGAADY     10
##  9 initial EMWGPDY       10
## 10 initial ETWGPEY       10
## # ... with 183 more rows
```

```r
b %>%
      group_by(cdr3, type) %>%
      select(cdr3, type, cdrp, quantity) %>%
      arrange(-cdrp, -quantity) -> c

c %>% filter(type == "initial") %>% slice_head(n = 1)
```

```
## # A tibble: 37 x 4
## # Groups:   cdr3, type [37]
##    cdr3             type       cdrp quantity
##    <chr>            <fct>     <dbl>    <int>
##  1 DGVAVAGLDY       initial 0.0378      2463
##  2 DIAAAGDFDY       initial 0.00298     1001
##  3 DISPVGYWFDP      initial 0.00405     1362
##  4 DLGIPDDY         initial 0.0338     11372
##  5 DLHWGAADY        initial 0.00720     2737
##  6 DLYLGYYYDSSGHSY  initial 0.00333     1119
##  7 DPIVVVPAASNWFDP  initial 0.00513     1726
##  8 DPYDSSGYSELTRFDP initial 0.00802     2699
##  9 DQNY             initial 0.00435     1465
## 10 DRTIVGASFDY      initial 0.0138      4628
## # ... with 27 more rows
```

```
ggplot(c) +
  geom_density(aes(cdrp, color = type), alpha = .4)
```



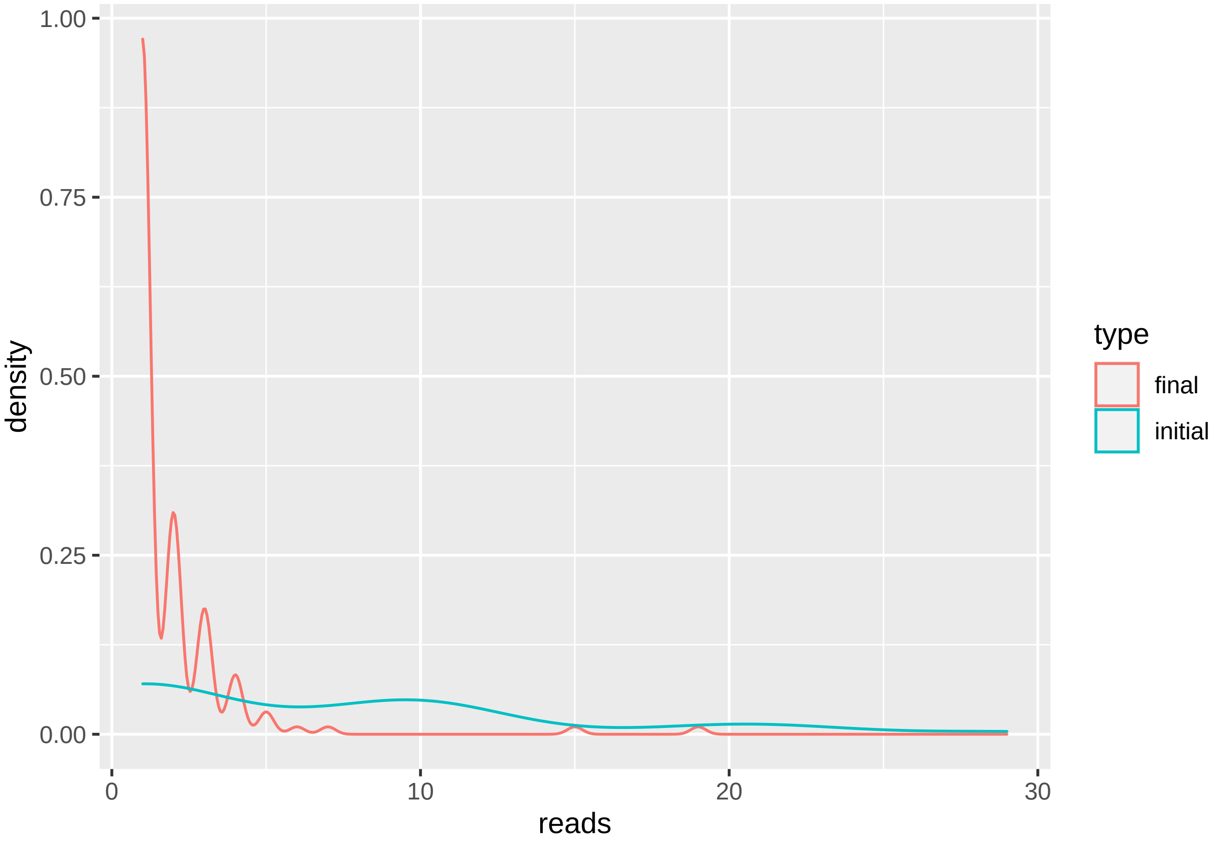
```
b %>%
    group_by(cdr3, type) %>%
    summarise(
      quantity = sum(quantity),
      reads    = n()) %>%
    arrange(-quantity, -reads) -> d

d
```

```
## # A tibble: 193 x 4
## # Groups:   cdr3 [159]
##    cdr3            type    quantity reads
##    <chr>           <fct>      <int> <int>
##  1 FLVEVK          final    1091572     5
##  2 GSHNSWDS        final     413702     4
##  3 FIVESK          initial   384738    29
##  4 DGVAVAGLDY      final     214584    19
##  5 RGSSSSFDY       final     204747     3
##  6 FIVESK          final     161781     7
##  7 DPTWRMATIGSLGTY final     127813     2
##  8 DLGIPDDY        initial   118382    21
##  9 ELVGATYY        final     100078     4
## 10 EMWGPEY         initial    79364    21
## # ... with 183 more rows
```

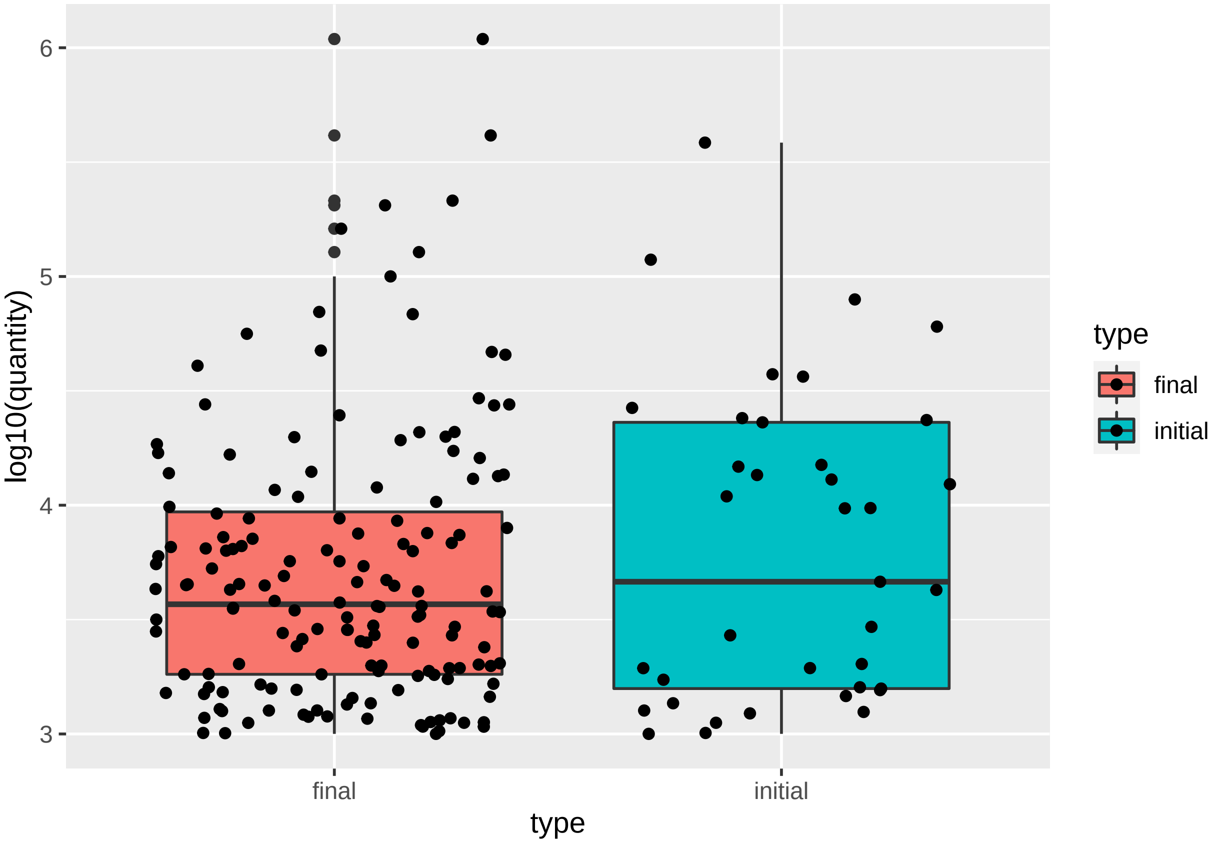
```
d %>% group_by(type) %>% summarise(n = n())
```

```
## # A tibble: 2 x 2
##   type        n
##   <fct>   <int>
## 1 final     156
## 2 initial    37
```

```
ggplot(d) +
  geom_density(aes(reads, color = type))
```



```
ggplot(d) +
  geom_boxplot(aes(type, log10(quantity), fill = type)) +
  geom_jitter(aes(type, log10(quantity), fill = type))
```



```
d
```

```
## # A tibble: 193 x 4
## # Groups:   cdr3 [159]
##    cdr3            type    quantity reads
##    <chr>           <fct>      <int> <int>
##  1 FLVEVK          final    1091572     5
##  2 GSHNSWDS        final     413702     4
##  3 FIVESK          initial   384738    29
##  4 DGVAVAGLDY      final     214584    19
##  5 RGSSSSFDY       final     204747     3
##  6 FIVESK          final     161781     7
##  7 DPTWRMATIGSLGTY final     127813     2
##  8 DLGIPDDY        initial   118382    21
##  9 ELVGATYY        final     100078     4
## 10 EMWGPEY         initial    79364    21
## # ... with 183 more rows
```

Resultados

Conclusão