# WordCloud in Biology with tm and ggwordcloud

Camila Farias Amorim

04/06/2022

## 1) Load required packages, color palette and gg theme

```
library(tm) # for text  mining
library(ggwordcloud) # for plotting the wordcloud
library(tidyverse)

library(jcolors)
pal9 <- jcolors("pal9")
theme_set(theme_minimal())
```

## 2) AC1 - Import the members (genes) and text mining

```
term1 <- "Signaling by Interleukins"

## Load the text:
AC1 <- readLines("import_export/annotation_cluster1")
AC1
```

```
## [1] "IL11, ANXA1, CXCL8, CSF2, MMP1, CCL3L3, IL24, OSM, LILRA3, LILRA5, IL1A, IL1B, CCL4, CCL3"
## [2] "IL21, IL11, ANXA1, CXCL8, CSF2, MMP1, CCL3L3, IL24, OSM, IL1A, IFNG, IL1B, CCL4, CCL3, S100A12"
## [3] "IL1A, CXCL8, CSF2, IL1B, CCL3L3, CCL4, CCL3"
## [4] "IL21, IL11, TNFRSF6B, ANXA1, CXCL8, CSF2, MMP1, CCL3L3, IL24, OSM, ISG15, IL1A, IFNG, IL1B, CCL
## [5] "IL1A, CXCL8, ANXA1, MMP1, IL1B, OSM"
## [6] "IL1A, IL1B, OSM"
## [7] "IL1A, CXCL8, IL1B, CCL3L3, IL24, OSM"
## [8] "IL1A, IL11, IL1B, OSM"
```

```
AC1 <- Corpus(VectorSource(AC1)) # Load the data as a corpus (a modified list)
inspect(AC1)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 8
##
## [1] IL11, ANXA1, CXCL8, CSF2, MMP1, CCL3L3, IL24, OSM, LILRA3, LILRA5, IL1A, IL1B, CCL4, CCL3
## [2] IL21, IL11, ANXA1, CXCL8, CSF2, MMP1, CCL3L3, IL24, OSM, IL1A, IFNG, IL1B, CCL4, CCL3, S100A12
## [3] IL1A, CXCL8, CSF2, IL1B, CCL3L3, CCL4, CCL3
## [4] IL21, IL11, TNFRSF6B, ANXA1, CXCL8, CSF2, MMP1, CCL3L3, IL24, OSM, ISG15, IL1A, IFNG, IL1B, CCL4
## [5] IL1A, CXCL8, ANXA1, MMP1, IL1B, OSM
## [6] IL1A, IL1B, OSM
## [7] IL1A, CXCL8, IL1B, CCL3L3, IL24, OSM
```

```
## [8] IL1A, IL11, IL1B, OSM
```

```
## Text transformation:
#toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))

#AC1 <- tm_map(AC1, toSpace, "/")
AC1 <- tm_map(AC1, removePunctuation) # Remove punctuation
inspect(AC1)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 8
##
## [1] IL11 ANXA1 CXCL8 CSF2 MMP1 CCL3L3 IL24 OSM LILRA3 LILRA5 IL1A IL1B CCL4 CCL3
## [2] IL21 IL11 ANXA1 CXCL8 CSF2 MMP1 CCL3L3 IL24 OSM IL1A IFNG IL1B CCL4 CCL3 S100A12
## [3] IL1A CXCL8 CSF2 IL1B CCL3L3 CCL4 CCL3
## [4] IL21 IL11 TNFRSF6B ANXA1 CXCL8 CSF2 MMP1 CCL3L3 IL24 OSM ISG15 IL1A IFNG IL1B CCL4 CCL3 S100A12
## [5] IL1A CXCL8 ANXA1 MMP1 IL1B OSM
## [6] IL1A IL1B OSM
## [7] IL1A CXCL8 IL1B CCL3L3 IL24 OSM
## [8] IL1A IL11 IL1B OSM
```

```
## Build a term-document matrix and model frequency data
AC1 <- TermDocumentMatrix(AC1) # Calculates sparsity and word frequencies
inspect(AC1)
```

```
## <<TermDocumentMatrix (terms: 19, documents: 8)>>
## Non-/sparse entries: 72/80
## Sparsity           : 53%
## Maximal term length: 8
## Weighting          : term frequency (tf)
## Sample             :
##         Docs
## Terms    1 2 3 4 5 6 7 8
##   anxa1  1 1 0 1 1 0 0 0
##   ccl3   1 1 1 1 0 0 0 0
##   ccl3l3 1 1 1 1 0 0 1 0
##   ccl4   1 1 1 1 0 0 0 0
##   csf2   1 1 1 1 0 0 0 0
##   cxcl8  1 1 1 1 1 0 1 0
##   il11   1 1 0 1 0 0 0 1
##   il1a   1 1 1 1 1 1 1 1
##   il1b   1 1 1 1 1 1 1 1
##   osm    1 1 0 1 1 1 1 1
```

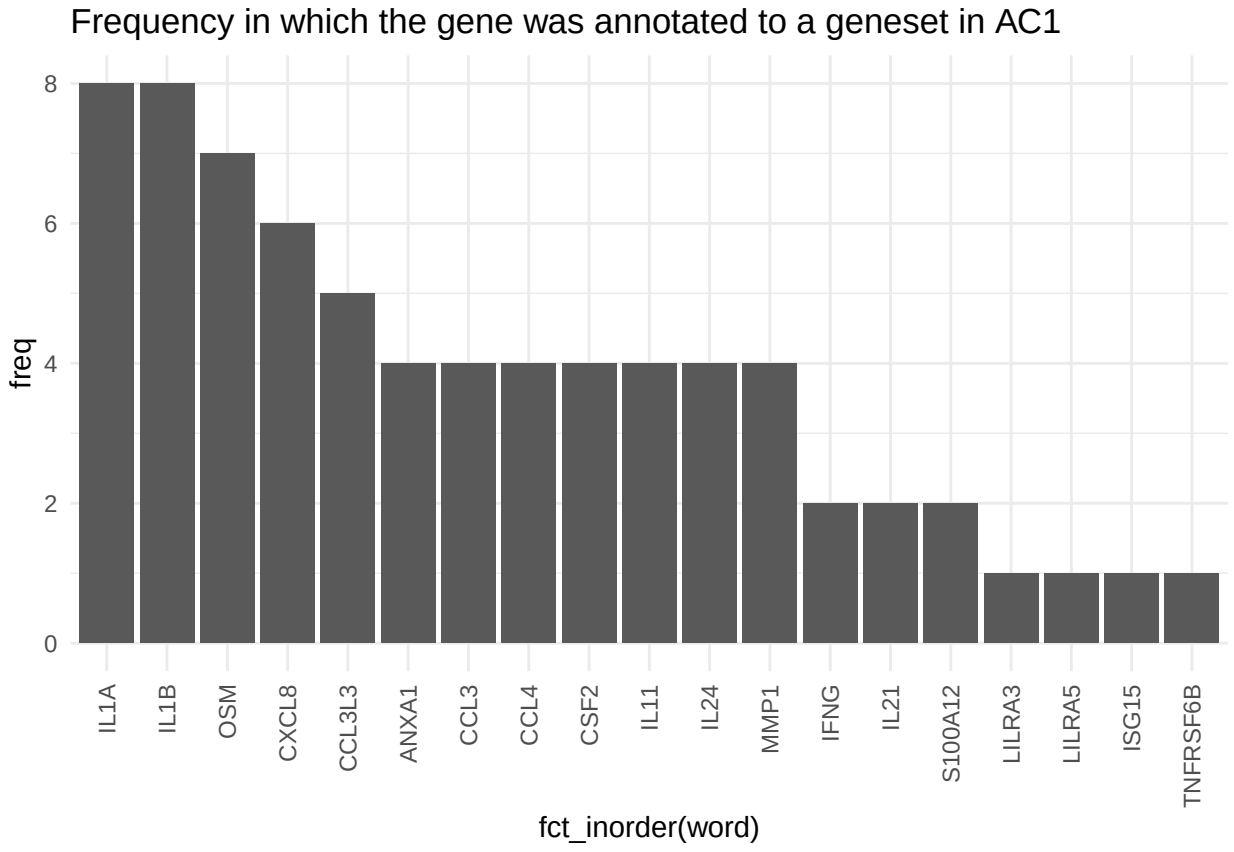
```
AC1 <- as.matrix(AC1)
rownames(AC1) <- str_to_upper(rownames(AC1)) # convert back to capital
AC1 <- sort(rowSums(AC1),decreasing=TRUE)
AC1 <- data.frame(word = names(AC1),freq=AC1) # new modeled data
AC1$freq <- as.numeric(AC1$freq)
AC1 <- arrange(AC1, desc(freq))
AC1$word <- as.factor(AC1$word)
AC1
```

```
##              word freq
## IL1A         IL1A    8
```

```
## IL1B          IL1B    8
## OSM            OSM    7
## CXCL8        CXCL8    6
## CCL3L3     CCL3L3    5
## ANXA1        ANXA1    4
## CCL3          CCL3    4
## CCL4          CCL4    4
## CSF2          CSF2    4
## IL11          IL11    4
## IL24          IL24    4
## MMP1          MMP1    4
## IFNG          IFNG    2
## IL21          IL21    2
## S100A12   S100A12    2
## LILRA3     LILRA3    1
## LILRA5     LILRA5    1
## ISG15        ISG15    1
## TNFRSF6B TNFRSF6B    1
```

## 3) AC1 - plotting

```
AC1 %>%
  ggplot(., aes(x=fct_inorder(word), y=freq)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5)) +
  ggtitle("Frequency in which the gene was annotated to a geneset in AC1")
```



```
AC1_plot <- AC1 %>%
  add_column(Gene = "yes") %>%
  ggplot(., aes(label = word, size = freq, color=Gene)) +
  scale_radius(range = c(1, 10), limits = c(0, NA)) +
  geom_text_wordcloud(area_corr = TRUE) +
  scale_size_area(max_size = 24) +
  scale_color_manual(values = pal9[1]) +
  ggtitle(term1) +theme(plot.title = element_text(hjust = 0.5, color = pal9[1]))
AC1_plot
```


## 4) AC2 and AC3

```
# AC2
term2 <- "Innate immune responses and Neutrophil degranulation"
AC2 <- readLines("import_export/annotation_cluster2")
AC2 <- Corpus(VectorSource(AC2))
AC2 <- tm_map(AC2, removePunctuation)
AC2 <- TermDocumentMatrix(AC2) #from tm
AC2 <- as.matrix(AC2)
rownames(AC2) <- str_to_upper(rownames(AC2))
AC2 <- sort(rowSums(AC2),decreasing=TRUE)
AC2 <- data.frame(word = names(AC2),freq=AC2)
AC2$freq <- as.numeric(AC2$freq)
AC2 <- arrange(AC2, desc(freq))
AC2$word <- as.factor(AC2$word)
AC2_plot <- AC2 %>%
  add_column(Gene = "yes") %>%
  ggplot(., aes(label = word, size = freq, color=Gene)) +
  scale_radius(range = c(1, 10), limits = c(0, NA)) +
  geom_text_wordcloud(area_corr = TRUE) +
  scale_size_area(max_size = 24) +
  scale_color_manual(values = pal9[2]) +
  ggtitle(term2) +theme(plot.title = element_text(hjust = 0.5, color = pal9[2]))
```

```
# AC3:
term3 <- "Chemotaxis"
AC3 <- readLines("import_export/annotation_cluster3")
AC3 <- Corpus(VectorSource(AC3))
AC3 <- tm_map(AC3, removePunctuation)
AC3 <- TermDocumentMatrix(AC3) #from tm
AC3 <- as.matrix(AC3)
rownames(AC3) <- str_to_upper(rownames(AC3))
AC3 <- sort(rowSums(AC3),decreasing=TRUE)
AC3 <- data.frame(word = names(AC3),freq=AC3)
AC3$freq <- as.numeric(AC3$freq)
AC3 <- arrange(AC3, desc(freq))
AC3$word <- as.factor(AC3$word)
AC3_plot <- AC3 %>%
  add_column(Gene = "yes") %>%
  ggplot(., aes(label = word, size = freq, color=Gene)) +
  scale_radius(range = c(1, 10), limits = c(0, NA)) +
  geom_text_wordcloud(area_corr = TRUE) +
  scale_size_area(max_size = 24) +
  scale_color_manual(values = pal9[3]) +
  ggtitle(term3) +theme(plot.title = element_text(hjust = 0.5, color = pal9[3]))
```

# 7) Proposed final image

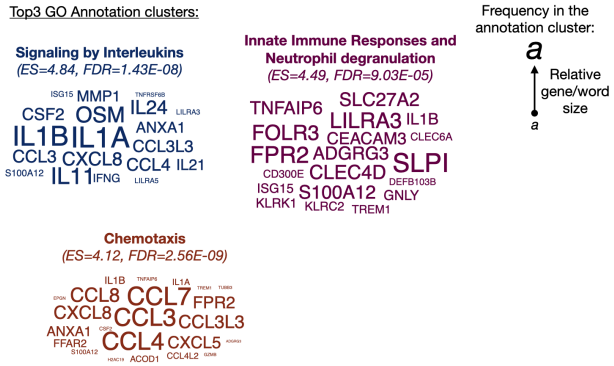



Figure 1: .

# 8) Highlight specific genes of interest

```
AC1 %>%
  add_column(Gene = "yes") %>%
  mutate(IL1_cytokines = case_when(
    word %in% c("IL1A", "IL1B") ~ "biomarker",
    TRUE ~ "not_biomarker")) %>%
  ggplot(., aes(label = word, size = freq, color=IL1_cytokines)) +
  scale_radius(range = c(1, 10), limits = c(0, NA)) +
  geom_text_wordcloud(area_corr = TRUE) +
  scale_size_area(max_size = 24) +
```

```
  scale_color_manual(values = c(pal9[4],pal9[6])) +
  ggtitle(term1) + theme(plot.title = element_text(hjust = 0.5))
```


## Session Info

**Session Info:** R version 4.1.2 (2021-11-01) Platform: x86_64-apple-darwin17.0 (64-bit) Running under: macOS Big Sur 10.16

Matrix products: default BLAS: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib

locale: [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages: [1] stats graphics grDevices utils datasets methods base

other attached packages: [1] jcolors_0.0.4 forcats_0.5.1 stringr_1.4.0
[4] dplyr_1.0.8 purrr_0.3.4 readr_2.1.2
[7] tidyr_1.2.0 tibble_3.1.6 tidyverse_1.3.1
[10] ggwordcloud_0.5.0.9000 ggplot2_3.3.5 tm_0.7-8
[13] NLP_0.2-1

loaded via a namespace (and not attached): [1] Rcpp_1.0.8.3 lubridate_1.8.0 png_0.1-7 assertthat_0.2.1 [5] digest_0.6.29 utf8_1.2.2 slam_0.1-50 R6_2.5.1
[9] cellranger_1.1.0 backports_1.4.1 reprex_2.0.1 evaluate_0.15
[13] highr_0.9 httr_1.4.2 pillar_1.7.0 rlang_1.0.2
[17] readxl_1.3.1 rstudioapi_0.13 rmarkdown_2.13 labeling_0.4.2
[21] munsell_0.5.0 gridtext_0.1.4 broom_0.7.12 compiler_4.1.2
[25] modelr_0.1.8 xfun_0.30 pkgconfig_2.0.3 htmltools_0.5.2 [29] tidyselect_1.1.2 fansi_1.0.2 crayon_1.5.0 tzdb_0.2.0
[33] dbplyr_2.1.1 withr_2.5.0 grid_4.1.2 jsonlite_1.8.0
[37] gtable_0.3.0 lifecycle_1.0.1 DBI_1.1.2 magrittr_2.0.2
[41] scales_1.1.1 cli_3.2.0 stringi_1.7.6 farver_2.1.0
[45] fs_1.5.2 xml2_1.3.3 ellipsis_0.3.2 generics_0.1.2
[49] vctrs_0.3.8 tools_4.1.2 glue_1.6.2 markdown_1.1
[53] hms_1.1.1 parallel_4.1.2 fastmap_1.1.0 yaml_2.3.5
[57] colorspace_2.0-3 rvest_1.0.2 knitr_1.37 haven_2.4.3