CUBI literature in 2017-2022

library(readxl)  
library(tidyverse)  
library(ggplot2)  
pubs <- read\_excel("CUBI\_journal\_list\_with\_citations\_2022.xlsx") %>%  
 mutate(year=as.numeric(year)) %>% rename(Year=year) %>%  
 mutate(cited=as.numeric(cited)) %>%  
 mutate(IF=as.numeric(IF)) %>%  
 rename(Last=last, First=first) %>%  
 mutate(Last=ifelse(is.na(Last), "F", Last)) %>%  
 mutate(cubi= ifelse(First == "T" | Last == "T", "Yes", "No"))  
theme\_set(theme\_minimal(base\_size=18))  
pal <- c("#00B3C7", "#C12079", "#007296", "#00A4BD", "#005C7E", "#00B9CB")

df <- pubs %>%   
 group\_by(Year, cubi) %>%   
 summarise(n=n(), IF=sum(IF), cited=sum(cited)) %>%  
 ungroup() %>%  
 arrange(Year) %>%   
 group\_by(cubi) %>%  
 mutate(ifcum=cumsum(IF)) %>%  
 mutate(citcum=cumsum(cited)) %>%  
 mutate(ncum=cumsum(n))

ggplot(df, aes(x=Year, y=n, fill=cubi)) + geom\_bar(stat="identity") +   
 ylab("Number of publications") +  
 scale\_fill\_manual(values=c(No=pal[1], Yes=pal[2])) +  
 labs(fill="CUBI 1st/last") +  
 theme(legend.position=c(0.2, 0.85))

|  |
| --- |
| Figure 1: Number of publications per year |

ggplot(df, aes(x=Year, y=ncum, fill=cubi)) + geom\_bar(stat="identity") +   
 scale\_fill\_manual(values=c(No=pal[1], Yes=pal[2])) +  
 labs(fill="CUBI 1st/last") +  
 theme(legend.position=c(0.2, 0.85)) +  
 ylab("Cumulative number of publications")

|  |
| --- |
| Figure 2: Cumulative number of publications per year |

ggplot(df, aes(x=Year, y=IF, fill=cubi)) + geom\_bar(stat="identity") +   
 scale\_fill\_manual(values=c(No=pal[1], Yes=pal[2])) +  
 labs(fill="CUBI 1st/last") +  
 theme(legend.position=c(0.2, 0.85)) +  
 ylab("Sum of IF per year")

|  |
| --- |
| Figure 3: Sum of the journal impact factor (IF) per year |

ggplot(df, aes(x=Year, y=ifcum, fill=cubi)) + geom\_bar(stat="identity") +   
 scale\_fill\_manual(values=c(No=pal[1], Yes=pal[2])) +  
 labs(fill="CUBI 1st/last") +  
 theme(legend.position=c(0.2, 0.85)) +  
 ylab("Cumulative sum of IF")

|  |
| --- |
| Figure 4: Cumulative sum of the journal impact factor (IF) per year |

Note: citations are based on 2022; earlier papers will have more citations. This is why it makes little sense to make a year by year comparison.

ggplot(df, aes(x=Year, y=citcum, fill=cubi)) + geom\_bar(stat="identity") +   
 scale\_fill\_manual(values=c(No=pal[1], Yes=pal[2])) +  
 labs(fill="CUBI 1st/last") +  
 theme(legend.position=c(0.2, 0.85)) +  
 ylab("Cumulative sum of citations")

|  |
| --- |
| Figure 5: Cumulative sum of the number of citations per year |

ggplot(df, aes(x=Year, y=cited/n, fill=cubi)) + geom\_bar(stat="identity") +   
 scale\_fill\_manual(values=c(No=pal[1], Yes=pal[2])) +  
 labs(fill="CUBI 1st/last") +  
 theme(legend.position=c(0.2, 0.85)) +  
 ylim(0, 200) +  
 ylab("Cumulative sum of citations")

|  |
| --- |
| Figure 6: Per publication citation numbers |

dat\_c <- pubs %>%   
 filter(cubi == "Yes") %>%   
 mutate(lIF=log10(IF), lcited=log10(cited)) %>%  
 filter(!is.infinite(lIF) & !is.infinite(lcited))  
lm\_c <- lm(lcited ~ lIF, data=dat\_c)  
dat\_cp <- data.frame(IF=dat\_c$IF, cited=10^(predict(lm\_c, dat\_c)))  
  
dat\_s <- pubs %>%   
 filter(cubi == "No") %>%   
 mutate(lIF=log10(IF), lcited=log10(cited)) %>%  
 filter(!is.infinite(lIF) & !is.infinite(lcited))  
lm\_s <- lm(lcited ~ lIF, data=dat\_s)  
dat\_sp <- data.frame(IF=dat\_s$IF, cited=10^(predict(lm\_s, dat\_s)))  
  
  
ggplot(pubs, aes(x=IF, y=cited, color=cubi)) + geom\_point() +  
 scale\_color\_manual(values=c(No=pal[1], Yes=pal[2])) +  
 geom\_line(color=pal[1], data=dat\_sp, mapping = aes(x=IF, y=cited)) +  
 geom\_line(color=pal[2], data=dat\_cp, mapping = aes(x=IF, y=cited)) +  
 xlab("Journal Impact Factor") +  
 ylab("Number of citations") +  
 scale\_x\_log10() +  
 scale\_y\_log10()

|  |
| --- |
| Figure 7: Number of citations vs Impact Factor. Each dot is one paper. Lines show log-log linear regression model calculated separately for CUBI and sandwich papers. |

library(stringr)  
library(tagcloud)  
words <- unlist(strsplit(pubs$title, split="[^a-zA-Z0-9]+")) %>% tolower()  
words <- table(words)  
wdf <- data.frame(N=as.vector(words), words=names(words)) %>%  
 filter(N > 2) %>%  
 filter(nchar(words) > 2) %>% arrange(-N) %>%   
 filter(!words %in% c("and", "the", "for", "with", "after", "from", "identifies")) %>%  
 mutate(words=gsub("dna", "DNA", words)) %>%  
 mutate(words=gsub("hla", "HLA", words)) %>%  
 mutate(words=gsub("sars", "SARS", words)) %>%  
 mutate(color=sample(pal, n(), replace=TRUE))  
tagcloud(tags=wdf$words, weights=wdf$N, col=wdf$color)

|  |
| --- |
| Figure 8: Tag cloud showing frequency of words used in titles of CUBI publications (size corresponds to number of publications). Colors are random. |

library(stringr)  
library(tagcloud)  
aut <- gsub("[^a-zA-Z]+", " ", pubs$authors)  
words <- unlist(strsplit(aut, split="[^a-zA-Z0-9]+")) %>% tolower() %>% str\_to\_title()  
words <- table(words)  
wdf <- data.frame(N=as.vector(words), words=names(words)) %>%  
 filter(N > 2) %>%  
 filter(nchar(words) > 2) %>% arrange(-N) %>%   
 filter(!words %in% c("Sch", "and", "the", "for", "with", "after", "from", "identifies")) %>%  
 mutate(words=gsub("dna", "DNA", words)) %>%  
 mutate(words=gsub("hla", "HLA", words)) %>%  
 mutate(words=gsub("sars", "SARS", words)) %>%  
 mutate(color=sample(pal, n(), replace=TRUE))  
tagcloud(tags=wdf$words, weights=wdf$N, col=wdf$color)

|  |
| --- |
| Figure 9: Tag cloud showing frequency of authors and co-authors of CUBI publications (size corresponds to number of publications). Colors are random. |

# Number of citations per publication

library(knitr)  
pubs$dYear <- 2023 - pubs$Year  
tab <- "Category,  
Total publications,  
,sandwich  
,CUBI  
Mean citations per publication,  
,sandwich  
,CUBI  
Median citations per publication,  
,sandwich  
,CUBI  
Mean citations per publication per year,  
,sandwich  
,CUBI  
Median citations per publication per year,  
,sandwich  
,CUBI  
Total IF,  
,sandwich  
,cubi  
Mean IF per publication,  
,sandwich  
,CUBI  
Median IF per publication,  
,sandwich  
,CUBI  
" %>% {read.table(text = ., sep=",", header=T)}  
  
pubs\_cubi <- pubs %>% filter(cubi == "Yes")  
pubs\_sand <- pubs %>% filter(cubi == "No")  
  
tab$Value <- c(  
nrow(pubs),  
nrow(pubs\_sand),  
nrow(pubs\_cubi),  
  
format(mean(pubs$cited), digits=2),  
format(mean(pubs\_sand$cited), digits=2),  
format(mean(pubs\_cubi$cited), digits=2),  
  
format(median(pubs$cited), digits=2),  
format(median(pubs\_sand$cited), digits=2),  
format(median(pubs\_cubi$cited), digits=2),  
  
format(mean(pubs$cited/pubs$dYear), digits=2),  
format(mean(pubs\_sand$cited/pubs\_sand$dYear), digits=2),  
format(mean(pubs\_cubi$cited/pubs\_cubi$dYear), digits=2),  
  
format(median(pubs$cited/pubs$dYear), digits=2),  
format(median(pubs\_sand$cited/pubs\_sand$dYear), digits=2),  
format(median(pubs\_cubi$cited/pubs\_cubi$dYear), digits=2),  
  
format(sum(pubs$IF), digits=2),  
format(sum(pubs\_sand$IF), digits=2),  
format(sum(pubs\_cubi$IF), digits=2),  
format(mean(pubs$IF), digits=2),  
format(mean(pubs\_sand$IF), digits=2),  
format(mean(pubs\_cubi$IF), digits=2),  
format(median(pubs$IF), digits=2),  
format(median(pubs\_sand$IF), digits=2),  
format(median(pubs\_cubi$IF), digits=2)  
  
)  
  
tab %>% kable()

| Category | X | Value |
| --- | --- | --- |
| Total publications |  | 85 |
|  | sandwich | 60 |
|  | CUBI | 25 |
| Mean citations per publication |  | 25 |
|  | sandwich | 24 |
|  | CUBI | 26 |
| Median citations per publication |  | 9 |
|  | sandwich | 8.5 |
|  | CUBI | 10 |
| Mean citations per publication per year |  | 7.8 |
|  | sandwich | 8.2 |
|  | CUBI | 6.7 |
| Median citations per publication per year |  | 3 |
|  | sandwich | 3 |
|  | CUBI | 2.5 |
| Total IF |  | 776 |
|  | sandwich | 655 |
|  | cubi | 121 |
| Mean IF per publication |  | 9.1 |
|  | sandwich | 11 |
|  | CUBI | 4.8 |
| Median IF per publication |  | 6.9 |
|  | sandwich | 7.4 |
|  | CUBI | 4.9 |

authors <- c("Beule", "Ivanov", "Holtgrewe", "Blanc", "Messerschmidt",  
 "Obermayer", "Weiner")  
names(authors) <- authors  
  
dfa <- map\_dfr(authors, ~ {  
 aut <- .x  
 pubs %>% filter(grepl(aut, authors, ignore.case = T)) %>%  
 summarise(Author=aut, N=n(), IF=sum(IF), Cited=sum(cited))  
 }) %>% arrange(-N)