Various VDJdb statistics

N.B.

You have to first run `cd src/ && groovy -cp . BuildDatabase`, which will create `database/` folder with the most recent VDJdb build.

Record statistics by year

Load data

```
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(httr)
library(xml2)
library(ggplot2)
library(stringr)
select = dplyr::select

dt.vdjdb = fread("../database/vdjdb_full.txt", sep = "\t", fill=T)
```

```
## Warning in fread("../database/vdjdb_full.txt", sep = "\t", fill = T):
## Bumped column 31 to type character on data row 702, field contains '1ao7'.
## Coercing previously read values in this column from logical, integer
## or numeric back to character which may not be lossless; e.g., if '00'
## and '000' occurred before they will now be just '0', and there may be
## inconsistencies with treatment of ',,' and ',NA,' too (if they occurred
## in this column before the bump). If this matters please rerun and set
## 'colClasses' to 'character' for this column. Please note that column type
## detection uses a sample of 1,000 rows (100 rows at 10 points) so hopefully
## this message should be very rare. If reporting to datatable-help, please
## rerun and include the output from verbose=TRUE.
```

```
dt.vdjdb.s = dt.vdjdb %>%
  filter(species != "MacacaMulatta") %>%
  mutate(tcr_key = paste(v.alpha, j.alpha, cdr3.alpha, v.beta, j.beta, cdr3.beta),
         mhc_key = paste(mhc.a, mhc.b),
         paired = cdr3.alpha != "" & cdr3.beta != "") %>%
  select(reference.id, tcr_key, mhc_key, paired, antigen.epitope, species) %>%
  unique
```

Fetch pubmed info

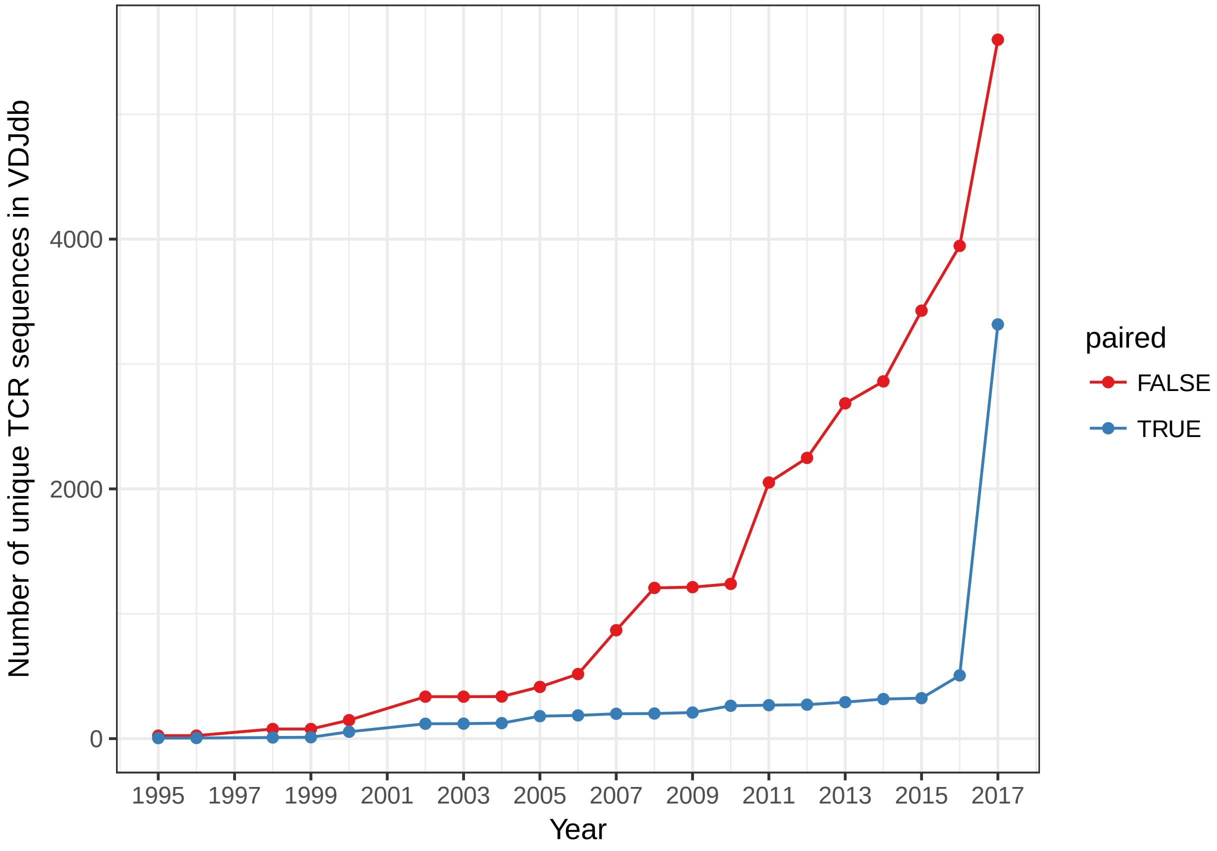
```
ids = unique((dt.vdjdb.s %>%
  filter(substr(reference.id,1,4)=="PMID") %>%
  mutate(id = str_split_fixed(reference.id, ":", 2)[,2]))$id)

pm_data = content(GET("https://eutils.ncbi.nlm.nih.gov/",
                      path = "entrez/eutils/esummary.fcgi",
                      query = list(db = "pubmed",
                                   id = paste0(ids,collapse = ","))))

write_xml(pm_data, "dates.xml")
pm_data_2 = readLines("dates.xml")
file.remove("dates.xml")
```

```
## [1] TRUE
```

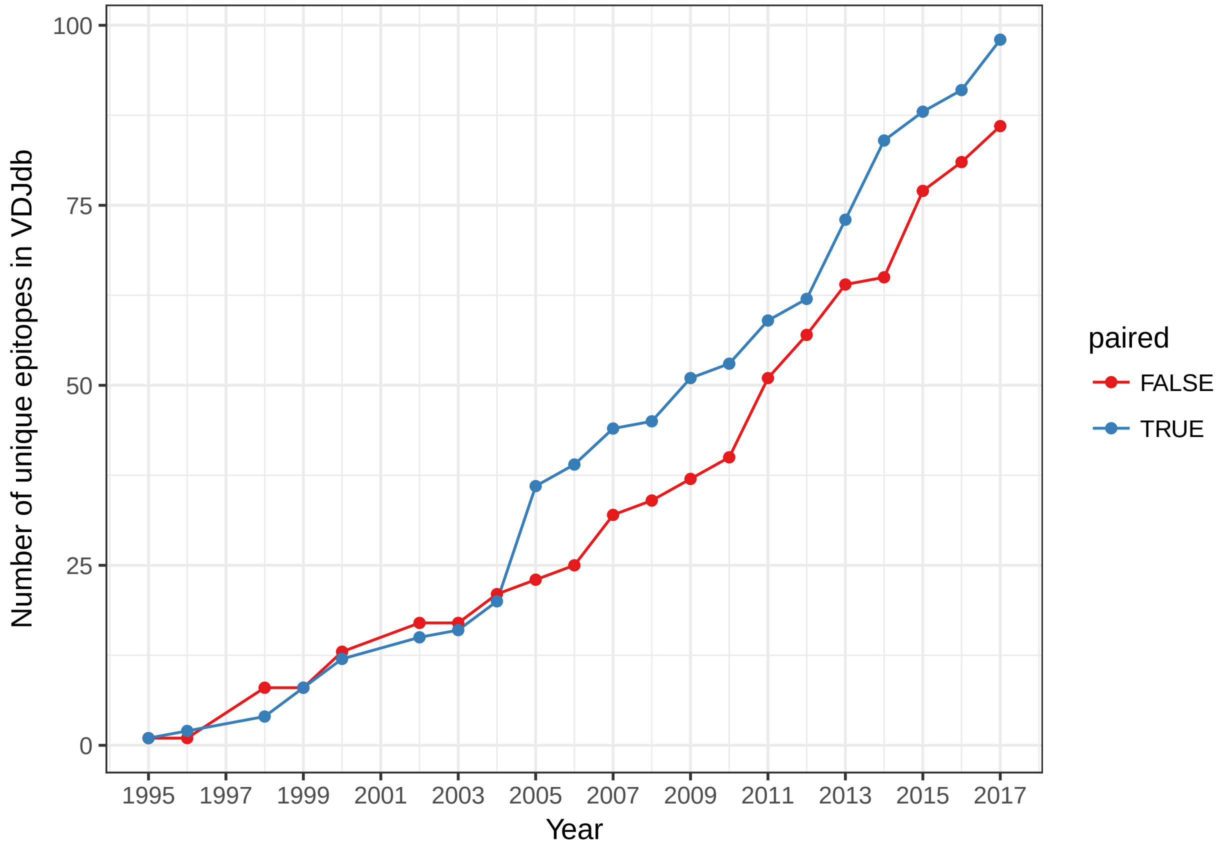
```
pm_data_id = str_split_fixed(pm_data_2[grepl("<Id>", pm_data_2)], "[<>]", n = Inf)[,3]
pm_data_date = str_split_fixed(pm_data_2[grepl('Name="PubDate"', pm_data_2)], "[<>]", n = Inf)[,3]

dt.pubdate = data.table(
  reference.id = paste0("PMID:", pm_data_id),
  pub_date = str_split_fixed(pm_data_date, " ", n = Inf)[,1]
)
```

Summarise

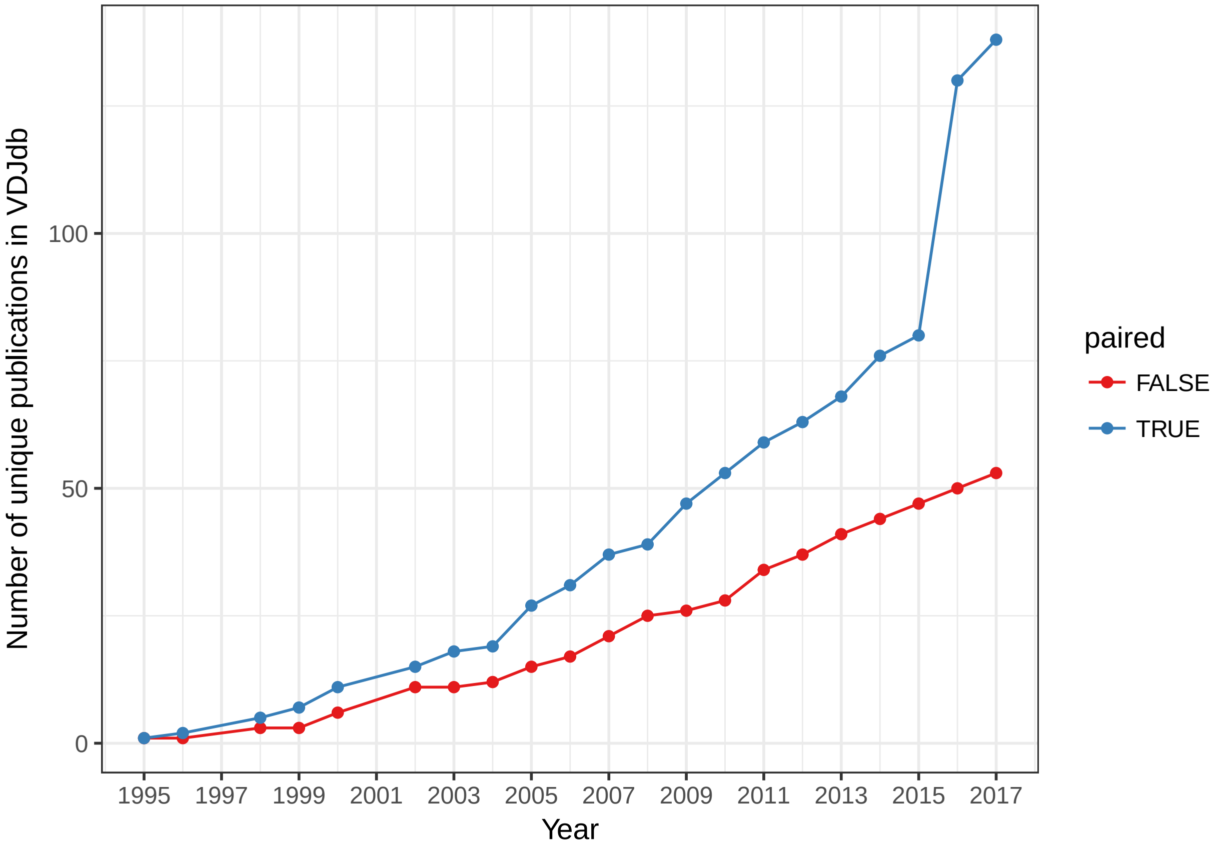
```
dt.vdjdb.s2 = dt.vdjdb.s %>%
  merge(dt.pubdate) %>%
  merge(data.table(pub_date2 = unique(dt.pubdate$pub_date)), allow.cartesian = T) %>%
  group_by(pub_date2, paired) %>%
  summarise(tcr_count = length(unique(tcr_key[which(pub_date <= pub_date2)])),
            epi_count = length(unique(antigen.epitope[which(pub_date <= pub_date2)])),
            ref_count = length(unique(reference.id[which(pub_date <= pub_date2)])),
            mhc_count = length(unique(mhc_key[which(pub_date <= pub_date2)])))

ggplot(dt.vdjdb.s2, aes(x = as.integer(pub_date2), y = tcr_count, color = paired)) +
  geom_line() +
  geom_point() +
  ylab("Number of unique TCR sequences in VDJdb") +
  scale_x_continuous("Year", breaks = seq(1995, 2017, by =2)) +
  scale_color_brewer(palette = "Set1") +
  theme_bw()
```



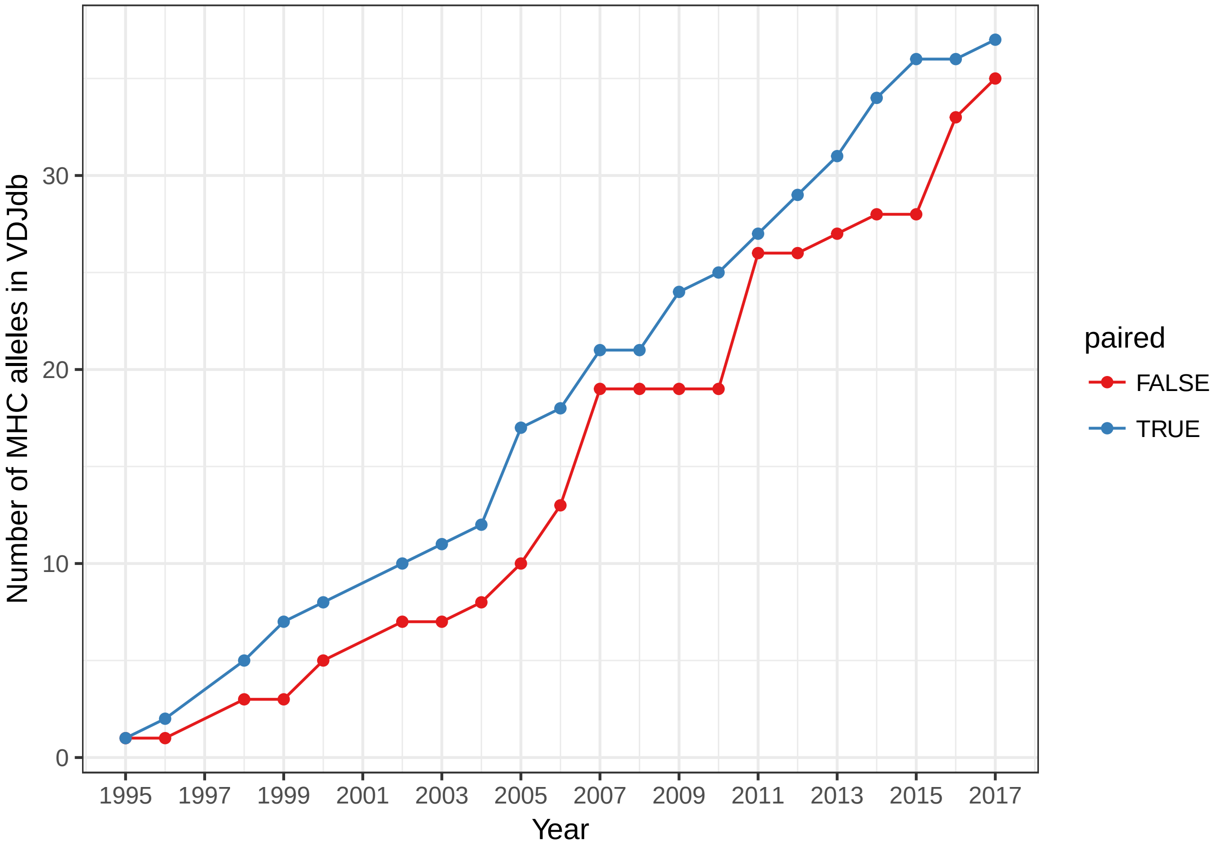
```
ggplot(dt.vdjdb.s2, aes(x = as.integer(pub_date2), y = epi_count, color = paired)) +
  geom_line() +
  geom_point() +
  ylab("Number of unique epitopes in VDJdb") +
  scale_x_continuous("Year", breaks = seq(1995, 2017, by =2)) +
  scale_color_brewer(palette = "Set1") +
  theme_bw()
```



```
ggplot(dt.vdjdb.s2, aes(x = as.integer(pub_date2), y = ref_count, color = paired)) +
  geom_line() +
  geom_point() +
  ylab("Number of unique publications in VDJdb") +
  scale_x_continuous("Year", breaks = seq(1995, 2017, by =2)) +
  scale_color_brewer(palette = "Set1") +
  theme_bw()
```



```r
ggplot(dt.vdjdb.s2, aes(x = as.integer(pub_date2), y = mhc_count, color = paired)) +
  geom_line() +
  geom_point() +
  ylab("Number of MHC alleles in VDJdb") +
  scale_x_continuous("Year", breaks = seq(1995, 2017, by =2)) +
  scale_color_brewer(palette = "Set1") +
  theme_bw()
```



```
fwrite(dt.vdjdb.s2, "vdjdb_stats_pubyear.txt", sep = "\t")
```