Public Submissions By Year

This is a quick example on generating plots of the publicly available *Staphylococcus aureus* sequencing data by year.

Import Packages

```
library(staphopia)
library(ggplot2)
library(reshape2)
USE_DEV = TRUE
```

Get Our Data

We'll use the **get_submission_by_year()** function to do exactly that, retrieve submission counts by year.

```
results <- get_submission_by_year()
results
```

```
##   year published unpublished count overall_published overall_unpublished
## 1 2010       169         159   328               169                 159
## 2 2011       830         790  1620               999                 949
## 3 2012      3195        2464  5659              4194                3413
## 4 2013      1976        3918  5894              6170                7331
## 5 2014      2534        7264  9798              8704               14595
## 6 2015      1930        6588  8518             10634               21183
## 7 2016      1543        3438  4981             12177               24621
## 8 2017        17        7157  7174             12194               31778
##   overall
## 1     328
## 2    1948
## 3    7607
## 4   13501
## 5   23299
## 6   31817
## 7   36798
## 8   43972
```

In the table above, there are seven columns:

1. year: The year in which an experiment was made public in ENA/SRA
2. published: The number of submissions associated with a publication for a given year
3. unpublished: The number of submissions **not** associated with a publication for a given year
4. count: The number of submissions for a given year
5. overall_published:The sum of published samples of the previous years
6. overall_unpublished: The sum of unpublished samples of the previous years
7. overall: The sum of each of the previous years

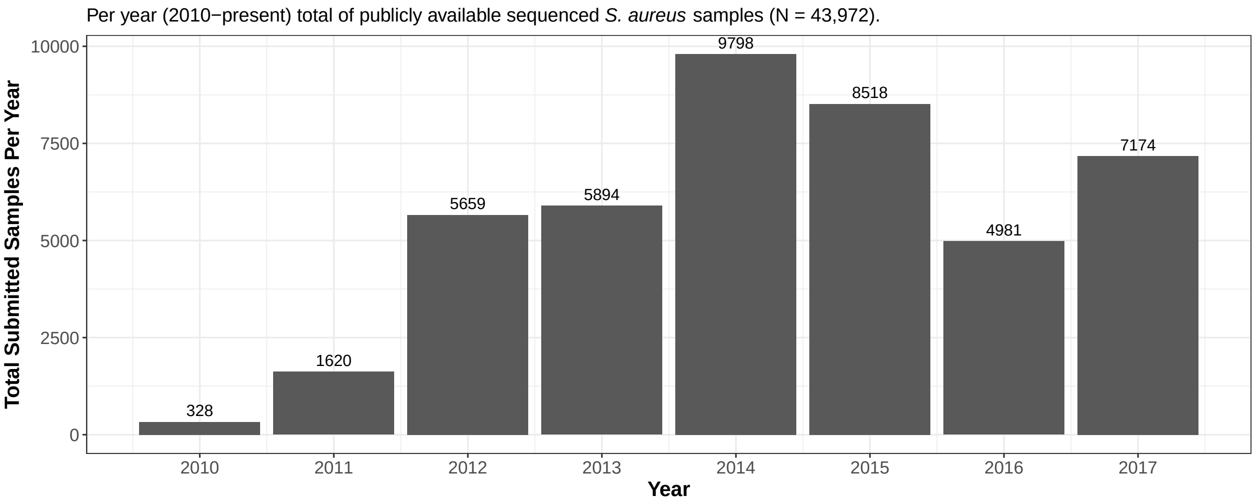
Plotting Our Data

We'll use *ggplot2* to make two simple plots to visualize our data.

Submissions Per Year

Here we are just going to look at the per year submissions.

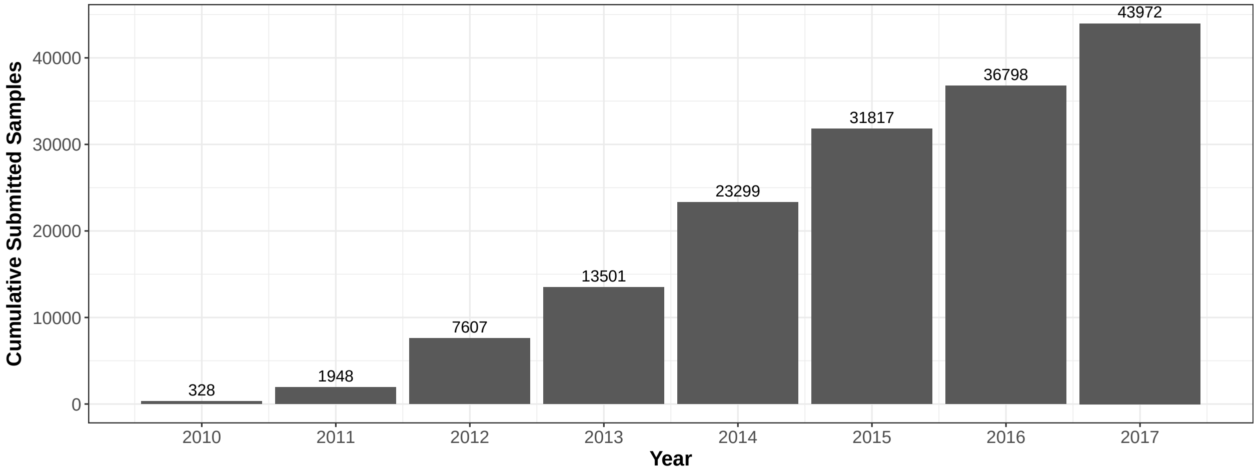
```
title <- substitute(paste("Per year (2010-present) total of publicly available sequenced ",
                          italic('S. aureus')," samples (N = ", x,")."),
                    list(x=format(max(results$overall), big.mark=',', scientific=FALSE)))
p <- ggplot(data=results, aes(x=year, y=count)) +
    xlab("Year") +
    ylab("Total Submitted Samples Per Year") +
    ggtitle(title) +
    geom_bar(stat='identity') +
    geom_text(aes(label=count), vjust = -0.5) +
    scale_x_continuous(breaks = round(seq(min(results$year), max(results$year), by = 1),1)) +
    theme_bw() +
    theme(axis.text=element_text(size=12),
          axis.title=element_text(size=14,face="bold"))
p
```



Overall Submissions

Changing gears a litte, here we are going to look at the overall growth of *S. aureus* sequencing data over the years.

```
title <- substitute(paste("Cumulative total of publicly available sequenced ",
                          italic('S. aureus')," samples (N = ", x,") by year (2010-present)."),
                    list(x=format(max(results$overall), big.mark=',', scientific=FALSE)))
p <- ggplot(data=results, aes(x=year, y=overall)) +
    xlab("Year") +
    ylab("Cumulative Submitted Samples") +
    ggtitle(title) +
    geom_bar(stat='identity') +
    geom_text(aes(label=overall), vjust = -0.5) +
    scale_x_continuous(breaks = round(seq(min(results$year), max(results$year), by = 1),1)) +
    theme_bw() +
    theme(axis.text=element_text(size=12),
          axis.title=element_text(size=14,face="bold"))
p
```


Cumulative total of publicly available sequenced *S. aureus* samples (N = 43,972) by year (2010−present).

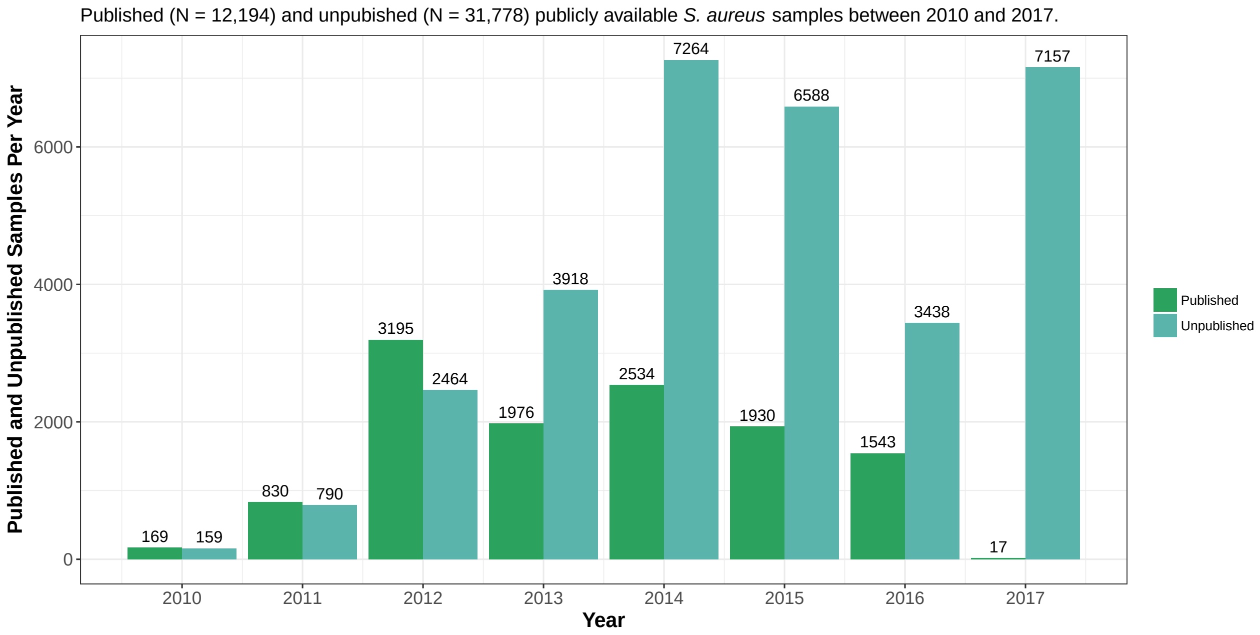
Published vs Unpublished By Year

For our final plot, we'll look at the number of samples that were referenced in a publication along side those that weren't. We'll need to melt the data in order to plot our groups.

```
melted <- melt(results, id=c('year'),
               measure.vars = c('published', 'unpublished'))
melted$title <- ifelse(melted$variable == 'published', 'Published', 'Unpublished')

title <- substitute(paste("Published (N = ", p, ") and unpubished (N = ", u, ") publicly available ",
                          italic('S. aureus')," samples between ", min_year, " and ", max_year, "."),
    list(
        p=format(max(results$overall_published), big.mark=',', scientific=FALSE),
        u=format(max(results$overall_unpublished), big.mark=',', scientific=FALSE),
        min_year=min(results$year),
        max_year=max(results$year)
))
p <- ggplot(data=melted, aes(x=year, y=value, fill=title)) +
    xlab("Year") +
    ylab("Published and Unpublished Samples Per Year") +
    ggtitle(title) +
    geom_bar(stat='identity', position='dodge') +
    geom_text(aes(label=value), vjust = -0.5, position = position_dodge(.9)) +
    scale_fill_manual(values=c("#2ca25f", "#5ab4ac")) +
    scale_x_continuous(breaks = round(seq(min(results$year), max(results$year), by = 1),1)) +
    theme_bw() +
    theme(axis.text=element_text(size=12),
          axis.title=element_text(size=14,face="bold"),
          legend.title = element_blank())


p
```



Published (N = 12,194) and unpubished (N = 31,778) publicly available *S. aureus* samples between 2010 and 2017.

```
sessionInfo()
```

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.2 LTS
##
## Matrix products: default
## BLAS: /usr/lib/libblas/libblas.so.3.6.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] reshape2_1.4.3  ggplot2_2.2.1   staphopia_0.1.9
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.15     knitr_1.20       magrittr_1.5     munsell_0.4.3
##  [5] colorspace_1.3-2 R6_2.2.2         rlang_0.1.6      stringr_1.2.0
##  [9] httr_1.3.1       plyr_1.8.4       tools_3.4.3      grid_3.4.3
## [13] gtable_0.2.0     htmltools_0.3.6  yaml_2.1.18      lazyeval_0.2.1
## [17] rprojroot_1.3-2  digest_0.6.15    tibble_1.4.2     curl_3.1
## [21] evaluate_0.10.1  rmarkdown_1.9    labeling_0.3     stringi_1.1.6
## [25] compiler_3.4.3   pillar_1.1.0     scales_0.5.0     backports_1.1.2
## [29] jsonlite_1.5
```