Public Sequencing Rank 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_rank_by_year()** function to do exactly that, retrieve submission counts by year.

```
results <- get_rank_by_year()
results
```

```
##   year bronze silver gold count overall_bronze overall_silver overall_gold
## 1 2010    292      0    0   292            292              0            0
## 2 2011   1452     21   55  1528           1744             21           55
## 3 2012   1598   1097 2895  5590           3342           1118         2950
## 4 2013    426    475 4928  5829           3768           1593         7878
## 5 2014    431   1138 8113  9682           4199           2731        15991
## 6 2015    519    588 7282  8389           4718           3319        23273
## 7 2016    454    990 3480  4924           5172           4309        26753
## 8 2017    645   1809 4261  6715           5817           6118        31014
##   overall
## 1     292
## 2    1820
## 3    7410
## 4   13239
## 5   22921
## 6   31310
## 7   36234
## 8   42949
```

In the table above, there are seven columns:

1. year: The year in which an experiment was made public in ENA/SRA
2. bronze: The number of bronze ranked samples for a given year
3. silver: The number of silver ranked samples for a given year
4. gold: The number of gold ranked samples for a given year
5. count: The number of submissions for a given year
6. overall_bronze: The sum of bronze ranked samples of the previous years
7. overall_silver: The sum of silver ranked samples of the previous years
8. overall_gold: The sum of gold ranked samples of the previous years
9. overall: The sum of each of the previous years

Plotting Our Data

We'll use *ggplot2* to visualize our data. We'll look at the number of Bronze, Silver and Gold ranked samples of

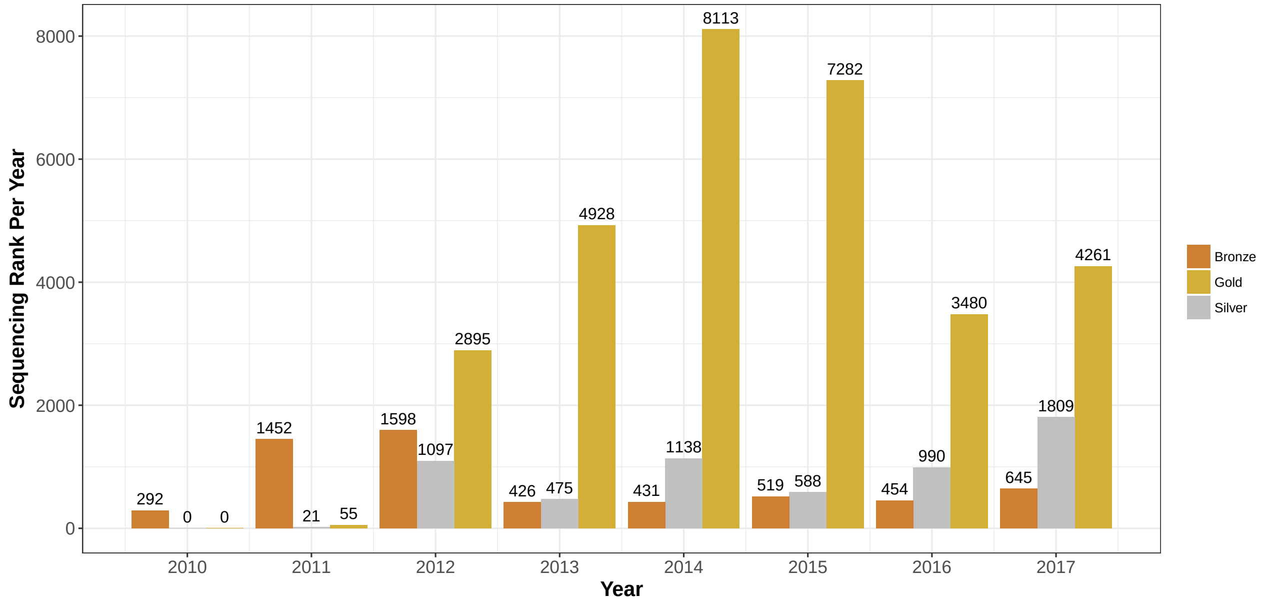
Published vs Unpublished By Year

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.

```r
melted <- melt(results, id=c('year'),
               measure.vars = c('bronze', 'silver', 'gold'))
melted$title <- ifelse(melted$variable == 'gold', 'Gold',
                       ifelse(melted$variable == 'silver', 'Silver', 'Bronze'))
melted$rank <- ifelse(melted$variable == 'gold', 3,
                      ifelse(melted$variable == 'silver', 2, 1))

title <- substitute(paste("Sequencing ranks (Bronze = ", b, ", Silver = ", s,
                          ", Gold = ", g, ") of publicly available ",
                          italic('S. aureus')," samples between ", min_year,
                          " and ", max_year, "."), list(
    b=format(max(results$overall_bronze), big.mark=',', scientific=FALSE),
    s=format(max(results$overall_silver), big.mark=',', scientific=FALSE),
    g=format(max(results$overall_gold), big.mark=',', scientific=FALSE),
    min_year=min(results$year),
    max_year=max(results$year)
))
p <- ggplot(data=melted, aes(x=year, y=value, fill=title, group=rank, label=title)) +
    xlab("Year") +
    ylab("Sequencing Rank Per Year") +
    ggtitle(title) +
    geom_bar(stat='identity', position='dodge') +
    geom_text(aes(label=value), vjust = -0.5, position = position_dodge(.9)) +
    scale_fill_manual(values=c("#CD7F32", "#D4AF37", "#C0C0C0")) +
    scale_x_continuous(breaks = round(seq(min(results$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
```



Sequencing ranks (Bronze = 5,817, Silver = 6,118, Gold = 31,014) of 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
```