Results Section: Public Metadata

```
library(staphopia)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

Aggregating Data For Public Samples

First we'll get all publicly available *S. aureus* samples.

```
ps <- get_public_samples()
```

We now have 42949 samples to work with. Next we will acquire metadata associated with each sample.

We will also get information pertaining to submissions by year and how any publication links were made.

```
submissions <- get_submission_by_year()
publication_links <- get_publication_links()
```

Next we are going to pull down any metadata associated with the public samples.

```
metrics <- merge(
    ps,
    get_metadata(ps$sample_id),
    by='sample_id'
)
```

We are now going to add two columns `rank_name` and `year`.

```
metrics$year <- sapply(
    metrics$first_public,
    function(x) {
        strsplit(x, "-")[[1]][1]
    }
)

metrics$rank_name <- ifelse(
    metrics$rank.x == 3,
    'Gold',
    ifelse(
        metrics$rank.x == 2,
        'Silver',
```

```
        'Bronze'
    )
)
```

Publication Information

Summary

Here are details looking at total submissions and their publication status.

```
t(submissions[submissions$year == max(submissions$year),])
```

```
##                          8
## year                  2017
## published               17
## unpublished           6698
## count                 6715
## overall_published    11921
## overall_unpublished 31028
## overall              42949
```

Here is information on how publication links were made.

```
t(publication_links)
```

```
##                1
## elink       6712
## text        5656
## elink_pmid    48
## text_pmid     30
## total      11921
## total_pmid    78
```

There are 6 rows and their names are as follows:

1. elink: Number samples linked to a PubMed ID identified from eLink
2. text: Number samples linked to a PubMed ID identified from text mining (not through eLink)
3. elink_pmid: Number of PubMed IDs identified from eLink
4. text_pmid: Number of PubMed IDs identified from text mining (not through eLink)
5. total: Total number of samples associated with a PubMed ID
6. total_pmid: Total number of PubMed IDs associated with published samples

Percent of Samples Published

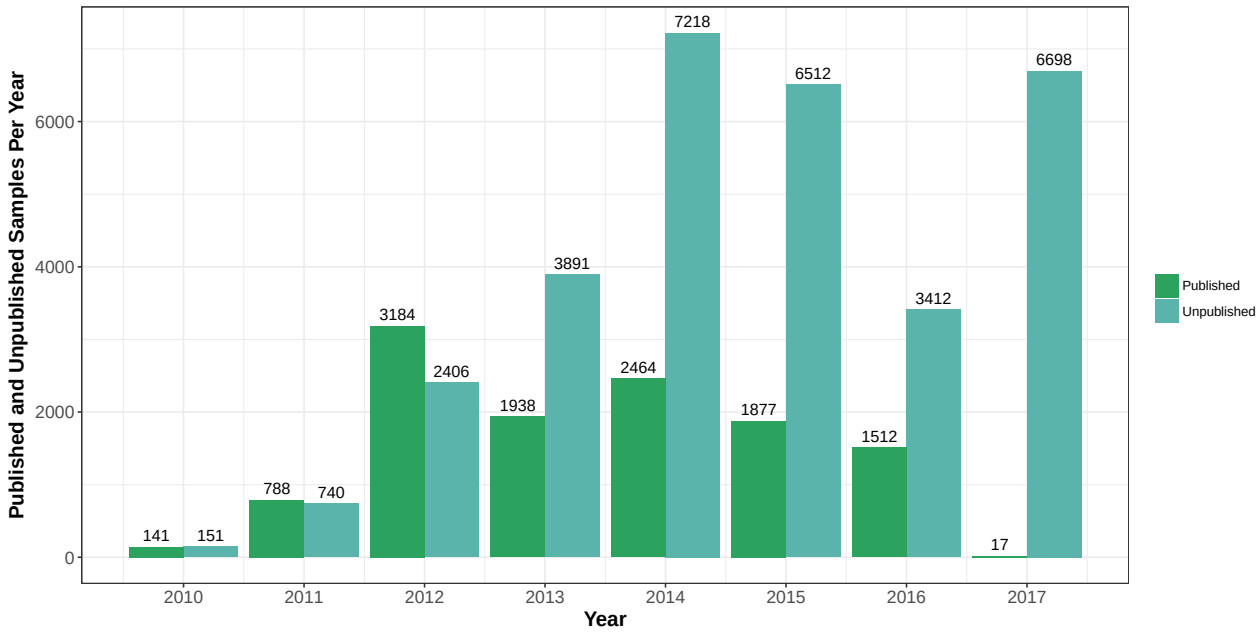
```
stats <- submissions[submissions$year == max(submissions$year),]
stats$overall_published / stats$overall * 100
```

```
## [1] 27.75618
```

Published vs Unpublished Submissions Per Year

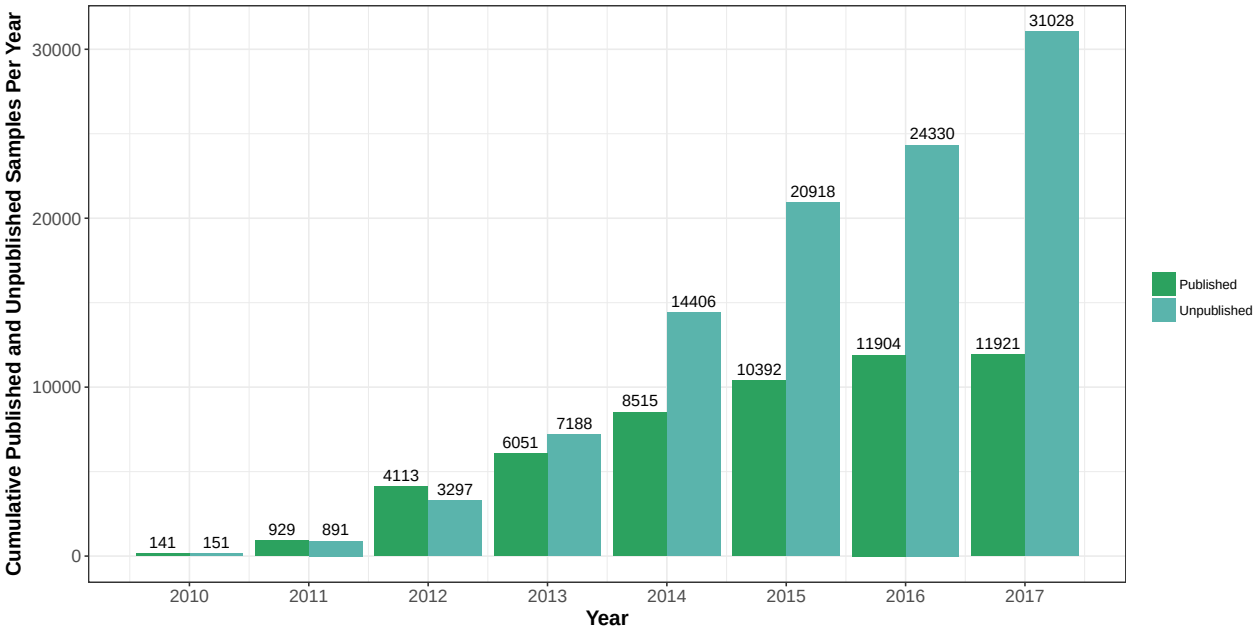
```
melted <- melt(submissions, id=c('year'),
               measure.vars = c('published', 'unpublished'))
melted$title <- ifelse(melted$variable == 'published', 'Published', 'Unpublished')
p <- ggplot(data=melted, aes(x=year, y=value, fill=title)) +
    xlab("Year") +
```

```
    ylab("Published and Unpublished Samples Per Year") +
    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(submissions$year), max(submissions$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
```



Overall Published vs Unpublished Submissions

```
melted <- melt(submissions, id=c('year'),
               measure.vars = c('overall_published', 'overall_unpublished'))
melted$title <- ifelse(melted$variable == 'overall_published', 'Published', 'Unpublished')
p <- ggplot(data=melted, aes(x=year, y=value, fill=title)) +
    xlab("Year") +
    ylab("Cumulative Published and Unpublished Samples Per Year") +
    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(submissions$year), max(submissions$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
```



Metadata Information

Number of Samples With A Collection Date

```
has_collection_date <- nrow(metrics[metrics$collection_date != "",])
paste0(has_collection_date," (", has_collection_date / nrow(metrics) * 100, " %)")
```

```
## [1] "17034 (39.660993271089 %)"
```

Number of Samples With A Location Information

```
has_location <- nrow(metrics[metrics$location != "unknown/missing",])
paste0(has_location," (", has_location / nrow(metrics) * 100, " %)")
```

```
## [1] "14983 (34.8855619455633 %)"
```

Number of Locations

```
nrow(as.data.frame(table(metrics[metrics$location != "unknown/missing",]$location)))
```

```
## [1] 123
```

Countries

```
country_data <- as.data.frame(table(metrics[(metrics$country != "unknown/missing" ) & (metrics$country
colnames(country_data) <- c("Country", "total")
country_data <- arrange(country_data, desc(total))
country_data
```

```
##                          Country total
## 1  United States of America (USA)  5823
## 2             United Kingdom (UK)  5177
## 3                         Germany   966
## 4                         Denmark   480
```

```
## 5                      Thailand   277
## 6                     Singapore   247
## 7                      Tanzania   153
## 8                   Netherlands   138
## 9                     Australia   131
## 10                   Luxembourg   122
## 11                      Ireland   111
## 12                       Gambia    88
## 13                  New Zealand    82
## 14                       Canada    59
## 15                     Colombia    59
## 16                        Gabon    59
## 17                       France    55
## 18                       Taiwan    54
## 19                      Belgium    53
## 20                    Argentina    50
## 21                        Spain    40
## 22                       Sweden    35
## 23                        Italy    29
## 24                     Portugal    28
## 25                       Russia    27
## 26                        Chile    25
## 27                  Switzerland    25
## 28                         Perú    24
## 29                       Poland    21
## 30                   Mozambique    17
## 31                     Malaysia    14
## 32                        Ghana    12
## 33                      Finland    10
## 34                       Norway     7
## 35                       Brazil     6
## 36                        China     6
## 37                       Greece     6
## 38                       Turkey     6
## 39                      Hungary     5
## 40                   Martinique     1
```

Number of Countries

```
paste0(nrow(country_data), " countries, represented by ", sum(country_data$total), " samples")
```

```
## [1] "40 countries, represented by 14528 samples"
```

Number of Samples With Isolation Source

```
has_source <- nrow(metrics[metrics$isolation_source != "",])
paste0(has_source," (", has_source / nrow(metrics) * 100, " %)")
```

```
## [1] "14768 (34.3849682181192 %)"
```

Isolation Sources

```
as.data.frame(table(metrics[metrics$isolation_source != "",]$isolation_source))
```

```
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21                                                                                                bakery envir
## 22                                                                  bakery environment - bottom metal shelf on tabl
## 23                                                                                      bakery environmen
## 24
## 25
## 26
## 27
## 28
## 29
## 30
## 31
## 32
## 33
## 34                                                                                                  bloodstrea
## 35
## 36
## 37
## 38
## 39
## 40
## 41
## 42
## 43
## 44
## 45
## 46
## 47
## 48
## 49
## 50
## 51
## 52
## 53                                                                                                          bu
```

```
## 54
## 55
## 56
## 57
## 58
## 59
## 60
## 61
## 62
## 63
## 64                                                                                                                  Ch
## 65
## 66
## 67
## 68
## 69
## 70
## 71
## 72
## 73
## 74
## 75
## 76
## 77
## 78
## 79
## 80
## 81
## 82
## 83
## 84
## 85
## 86
## 87
## 88
## 89
## 90
## 91
## 92
## 93
## 94
## 95
## 96
## 97
## 98  fatal septicaemia and septic arthritis in a 16-month-old American-Indian girl who had no risk fa
## 99
## 100
## 101
## 102
## 103
## 104
## 105
## 106
## 107
```

```
## 108
## 109
## 110
## 111
## 112
## 113
## 114
## 115
## 116
## 117
## 118
## 119
## 120
## 121
## 122
## 123
## 124
## 125                                                                                hexachl
## 126
## 127
## 128
## 129
## 130
## 131
## 132
## 133
## 134
## 135
## 136
## 137
## 138
## 139
## 140
## 141
## 142
## 143
## 144
## 145
## 146
## 147
## 148                                                     Isolated from pus and debrided tissu
## 149
## 150
## 151
## 152
## 153
## 154
## 155
## 156
## 157
## 158
## 159
## 160
## 161
```

```
## 162
## 163
## 164
## 165
## 166
## 167
## 168
## 169
## 170
## 171
## 172
## 173                                                                lower respira
## 174
## 175
## 176                                                                          lu
## 177                                                                          lu
## 178                                                                          lu
## 179
## 180
## 181
## 182
## 183
## 184
## 185
## 186
## 187
## 188                                                                          MI
## 189                                                                           I
## 190
## 191
## 192
## 193
## 194
## 195
## 196
## 197
## 198
## 199
## 200
## 201
## 202
## 203
## 204
## 205
## 206
## 207
## 208
## 209
## 210
## 211
## 212
## 213
## 214
## 215
```

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230 peri
231
232
233
234
235
236
237
238
239
240
241
242
243
244 Purulen
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269

```
## 270
## 271
## 272
## 273
## 274
## 275
## 276
## 277
## 278
## 279
## 280
## 281
## 282
## 283
## 284
## 285                                                                    Stool of
## 286
## 287
## 288
## 289
## 290
## 291
## 292
## 293
## 294
## 295
## 296
## 297
## 298
## 299
## 300
## 301
## 302
## 303
## 304
## 305
## 306
## 307
## 308
## 309
## 310
## 311
## 312
## 313
## 314
## 315
## 316
## 317
## 318
## 319
## 320
## 321
## 322
## 323
```

```
## 324
## 325
## 326
## 327
## 328
## 329
## 330
## 331
## 332
## 333
## 334
## 335
## 336
## 337
## 338
## 339
## 340
## 341
## 342
## 343
## 344
## 345
## 346
## 347
## 348
## 349
## 350
## 351
## 352
## 353
## 354
## 355
## 356
## 357
## 358
## 359
## 360
## 361
## 362
## 363
## 364
## 365
## 366
## 367
## 368
## 369
## 370
## 371
## 372
## 373
## 374
## 375
## 376
## 377
```

```
## 378
## 379
## 380
## 381
## 382
## 383
## 384
## 385
## 386
## 387
## 388
## 389
## 390
## 391
## 392
## 393
## 394
## 395
## 396
##     Freq
## 1      4
## 2      3
## 3      2
## 4      4
## 5      4
## 6      8
## 7      7
## 8     54
## 9      2
## 10    10
## 11     1
## 12     1
## 13     4
## 14     5
## 15     2
## 16     1
## 17     2
## 18     2
## 19     2
## 20     1
## 21     2
## 22     1
## 23     1
## 24     1
## 25     1
## 26     3
## 27  1502
## 28   699
## 29     5
## 30     1
## 31     2
## 32     2
## 33    38
## 34     1
```

```
## 35     2
## 36   229
## 37     3
## 38     2
## 39    19
## 40     4
## 41     2
## 42     2
## 43    26
## 44     9
## 45    24
## 46     2
## 47     5
## 48     3
## 49     8
## 50    71
## 51    93
## 52     3
## 53     1
## 54     1
## 55     8
## 56     1
## 57     7
## 58     1
## 59     2
## 60     1
## 61     1
## 62     4
## 63     2
## 64     1
## 65     2
## 66     1
## 67    78
## 68    10
## 69     9
## 70     1
## 71     4
## 72     1
## 73     8
## 74   305
## 75   399
## 76     1
## 77     4
## 78    40
## 79     1
## 80     1
## 81     1
## 82    19
## 83     1
## 84     1
## 85     1
## 86     1
## 87     8
## 88   176
```

```
## 89     3
## 90     2
## 91     4
## 92     9
## 93     4
## 94     1
## 95     1
## 96     4
## 97     3
## 98     2
## 99     1
## 100    4
## 101    2
## 102   10
## 103   10
## 104    2
## 105   10
## 106    4
## 107    7
## 108    1
## 109    1
## 110    8
## 111   63
## 112    1
## 113    2
## 114    2
## 115   31
## 116    3
## 117    7
## 118   17
## 119   36
## 120    1
## 121    1
## 122    2
## 123    1
## 124    2
## 125    1
## 126    1
## 127    4
## 128    3
## 129    1
## 130    1
## 131  201
## 132   79
## 133    3
## 134   35
## 135    1
## 136  202
## 137  133
## 138   18
## 139   15
## 140   67
## 141    2
## 142    1
```

```
## 143   24
## 144    9
## 145    3
## 146  119
## 147   25
## 148    2
## 149   15
## 150    2
## 151    3
## 152   43
## 153    2
## 154    2
## 155    1
## 156    2
## 157    1
## 158    4
## 159  118
## 160   52
## 161    1
## 162    1
## 163    1
## 164  125
## 165    1
## 166    7
## 167   12
## 168    2
## 169    4
## 170    1
## 171    5
## 172    1
## 173    5
## 174    6
## 175    1
## 176    3
## 177    1
## 178    1
## 179    2
## 180    1
## 181   69
## 182    1
## 183    1
## 184   57
## 185  153
## 186   11
## 187    7
## 188  228
## 189    5
## 190    1
## 191  641
## 192  595
## 193  141
## 194   14
## 195  225
## 196   40
```

```
## 197   29
## 198    1
## 199    6
## 200  112
## 201    2
## 202    1
## 203    4
## 204    1
## 205    8
## 206 1311
## 207  234
## 208    3
## 209    1
## 210    1
## 211  116
## 212  865
## 213  238
## 214   13
## 215    5
## 216   45
## 217  253
## 218    2
## 219    4
## 220    1
## 221    2
## 222    1
## 223    1
## 224    1
## 225    2
## 226   88
## 227   30
## 228    1
## 229    9
## 230    1
## 231    1
## 232    2
## 233    1
## 234    2
## 235    7
## 236    3
## 237    1
## 238    4
## 239   21
## 240    1
## 241   14
## 242    8
## 243   30
## 244    1
## 245   34
## 246    2
## 247    1
## 248    1
## 249    1
## 250    1
```

```
## 251  210
## 252    2
## 253   43
## 254    1
## 255    1
## 256    9
## 257   38
## 258   12
## 259    1
## 260    1
## 261    1
## 262    1
## 263    3
## 264    1
## 265    9
## 266   87
## 267    1
## 268    7
## 269    4
## 270   88
## 271    3
## 272   17
## 273   10
## 274  205
## 275    3
## 276    1
## 277  363
## 278  266
## 279    1
## 280    7
## 281    1
## 282    1
## 283    3
## 284    4
## 285    1
## 286    1
## 287    1
## 288    1
## 289    9
## 290    2
## 291    1
## 292    2
## 293  127
## 294   26
## 295    1
## 296    1
## 297  171
## 298   12
## 299    3
## 300    2
## 301   38
## 302   10
## 303   64
## 304   17
```

```
## 305    1
## 306    1
## 307    4
## 308    4
## 309    1
## 310    1
## 311    1
## 312    1
## 313    2
## 314    1
## 315    1
## 316   15
## 317    2
## 318    1
## 319    1
## 320    2
## 321    1
## 322    2
## 323    1
## 324    3
## 325    1
## 326    1
## 327    1
## 328    1
## 329    1
## 330    1
## 331    1
## 332    1
## 333    1
## 334    1
## 335    1
## 336    1
## 337    1
## 338    1
## 339    1
## 340    1
## 341    1
## 342    1
## 343    1
## 344    3
## 345    1
## 346    1
## 347    3
## 348    1
## 349    1
## 350    1
## 351    1
## 352    1
## 353    1
## 354    1
## 355    1
## 356    2
## 357    1
## 358    1
```

```
## 359    1
## 360    1
## 361    1
## 362    1
## 363    2
## 364    1
## 365    1
## 366    1
## 367    1
## 368    1
## 369    1
## 370    3
## 371    1
## 372    1
## 373    1
## 374    1
## 375    1
## 376    4
## 377    1
## 378    1
## 379   36
## 380   91
## 381    1
## 382    1
## 383    1
## 384    2
## 385    1
## 386    1
## 387    1
## 388  199
## 389  997
## 390    2
## 391   13
## 392    1
## 393    1
## 394   51
## 395   18
## 396    2
```

Number of Isolation Sources

```
nrow(as.data.frame(table(metrics[metrics$isolation_source != "",]$isolation_source)))
```

```
## [1] 396
```