

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## unpublished  6698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

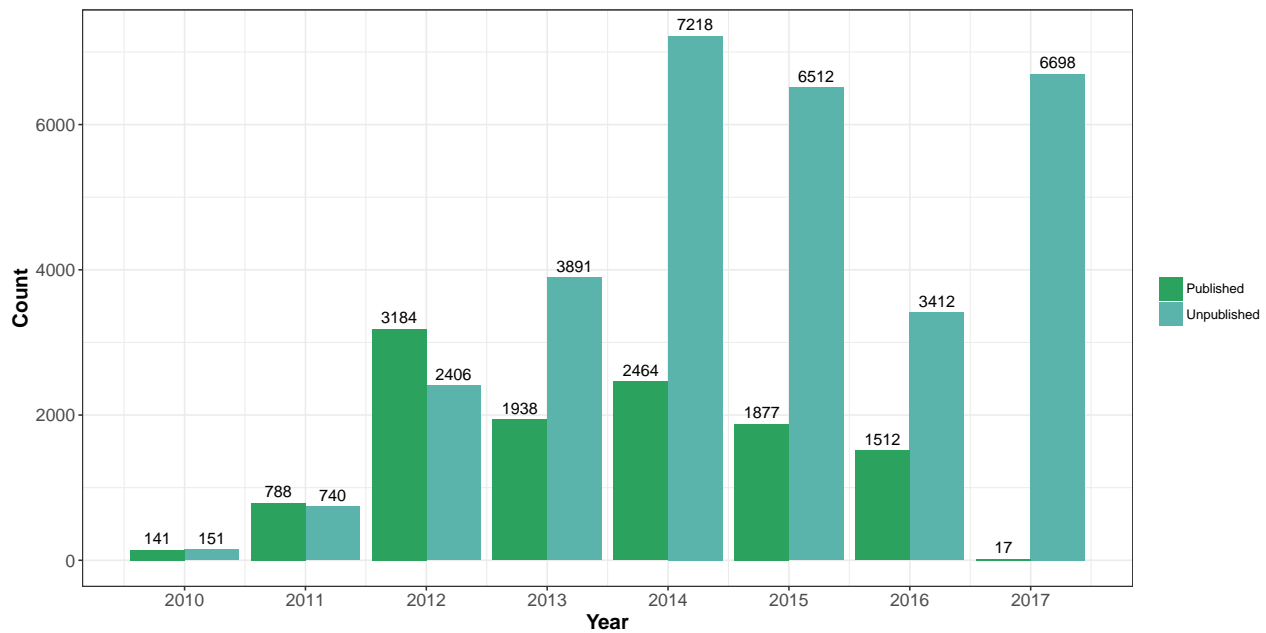
```
melted <- melt(submissions, id=c('year'),
               measure.vars = c('published', 'unpublished'))
melted$title <- ifelse(melted$variable == 'published', 'Published', 'Unpublished')
p <- ggplot(data=melted, aes(x=year, y=value, fill=title)) +
  xlab("Year") +
```

```

ylab("Count") +
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



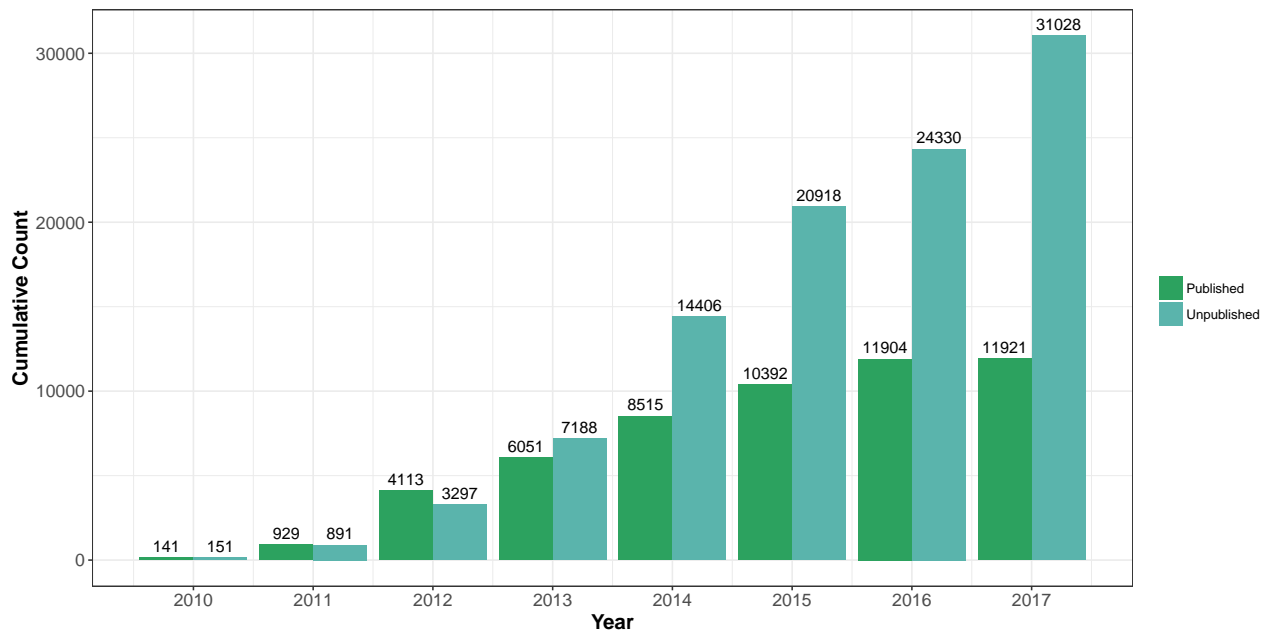
Overall Published vs Unpublished Submissions

```

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 Count") +
  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



```
# Output plot to PDF and PNG
staphopia::write_plot(p, paste0(getwd(), '/images/figure-x-published-per-year'))
```

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

```
df <- as.data.frame(table(substr(tolower(metrics[metrics$isolation_source != "",]$isolation_source), 1,
df[order(-df$Freq),]
```

##		Var1	Freq
## 23		blood	2201
## 184		nose	1548
## 174		nares	1236
## 333		wound	1196
## 188		not known	1116
## 65		culture	704
## 244		sputum	629
## 177		nasal	265
## 191		other	253
## 29		bodily fluid	229
## 171	mr	sa screen - nose/throat/perineum	228
## 221		respiratory	210
## 241		soft tissue	205
## 122		human body	202
## 118		host	201
## 262		throat	186
## 78		environment	176
## 169		mr	sa screen 164
## 123		human clinical	151
## 175		nares or umbillicus	141
## 326		urine	128
## 258	swabs, multiple swab locations		127
## 148		leg infection	125
## 180		nasal swab	120
## 131		invasive	119
## 143		laboratory	118
## 199		perineum	118
## 187		not collected	116
## 234		skin	97
## 43		bulk tank milk	93
## 237		skin or soft tissue	88
## 119		household surface	79
## 58		clinical specimen	78
## 265		tissue	74
## 42		bulk milk	71
## 165		milk	69
## 337		wound swab	69
## 125	human clinical specimen		67
## 99		foremilk	63
## 5		abscess	61
## 168		mr	sa [broth] 57
## 144		laboratory strain	52
## 136		joint fluid	46
## 190		osteomyelitis	45
## 223		respiratory sample	43
## 68		diabetic foot sample	40
## 26		blood for culture	38
## 120		human	38
## 227	sample from soft tissue		38
## 264		tip	38

## 106	ground turkey	36
## 103	groin	34
## 215	pus	34
## 213	prosthetic joint infection	30
## 178	nasal or rectal swab	29
## 35	bronchial alveolar lavage	26
## 259	swine facility	26
## 132	in vitro derived	25
## 37	bronchial washings	24
## 128	icu	24
## 31	bone	21
## 209	pork chop	21
## 91	fluid	20
## 72	drainage	19
## 105	ground beef	17
## 239	skin swab	17
## 266	tissues	17
## 277	ulcer swab	17
## 124	human clinical isolate	15
## 134	isolate from a human	15
## 176	nares/umbilicus/acilla	14
## 211	post surgical secretion	14
## 81	eye	13
## 335	wound infection	13
## 4	abdominal wound	12
## 151	leg wound	12
## 228	screen swab	12
## 7	abscess/pus swab	10
## 59	colonization	10
## 93	food	10
## 200	peritoneal fluid	10
## 206	pleural fluid	10
## 240	skin wound	10
## 36	bronchial secretions	9
## 60	commensal	9
## 129	infection	9
## 226	sample from bone or joint	9
## 255	surgical wound	9
## 24	blood culture	8
## 41	broncoscopy	8
## 47	catheter	8
## 64	csf	8
## 77	elbow wound	8
## 98	foot wound	8
## 183	non-icu	8
## 212	pressure sore	8
## 49	cellulitis of leg	7
## 95	footpad infection	7
## 104	groin swab	7
## 150	leg swab - left	7
## 170	mrsa screening swab	7
## 235	skin abscess	7
## 246	sputum from endotrachea	7
## 158	lung	6

## 2	abdominal fluid	5
## 11	armpit	5
## 39	bronchoalveolar lavage	5
## 53	chicken breast	5
## 155	liver infection	5
## 157	lower respiratory tract specimens of patients	5
## 172	mrsa screen - other site/specimen	5
## 182	neck wound	5
## 189	oral	5
## 323	urethra	5
## 1	abdominal abscess	4
## 3	abdominal swab	4
## 10	arm	4
## 15	aspirate	4
## 32	bone marrow infection	4
## 62	corneal ulcer	4
## 67	decubitus ulcer	4
## 82	eye drainage	4
## 85	faeces	4
## 89	finger wound	4
## 94	foot	4
## 114	hip infection	4
## 142	knee wound	4
## 153	liver	4
## 193	peg tube drainage	4
## 208	pooled	4
## 232	septic arthritis	4
## 236	skin infection	4
## 250	stool	4
## 269	toe wound	4
## 270	tonsillar abscess	4
## 320	unknown7	4
## 22	bile	3
## 30	body fluid	3
## 40	bronchoscopy	3
## 44	burn	3
## 79	environmental	3
## 86	farm	3
## 115	hip joint fluid	3
## 130	intra-abdominal abscess	3
## 160	lungs of cystic fibrosis patient a	3
## 194	penile swab	3
## 238	skin or soft tissue infection	3
## 242	spinal fluid	3
## 249	sternal wound	3
## 257	swab	3
## 284	umbilicus	3
## 298	unknown22	3
## 301	unknown26	3
## 6	abscess/pus collection	2
## 12	arm swab - left	2
## 14	ascitic fluid	2
## 17	bakery environment - assembly production room	2
## 25	blood - culture	2

## 28	blops	2
## 33	brain abscess	2
## 34	bronch	2
## 38	bronchoalveolar aspirate	2
## 51	chest	2
## 54	child - hospital pneumology ward	2
## 56	clinical	2
## 80	excreted bodily substance	2
## 87	fatal septicaemia and septic arthritis in a 16-mon	2
## 90	fish drying yard	2
## 92	fluid left elbow	2
## 101	graft	2
## 102	granuloma	2
## 109	heart valve	2
## 111	hematoma	2
## 126	human samples	2
## 133	isolated from pus and debrided tissue at surgical	2
## 135	joint aspirate	2
## 137	jp drainage	2
## 138	jugular catheter	2
## 140	knee	2
## 152	lesion	2
## 163	mass	2
## 192	p.e.g site swab	2
## 198	perineal	2
## 203	peritoneum infection	2
## 205	pin tract	2
## 216	pus swab	2
## 222	respiratory culture	2
## 256	suture	2
## 263	throat swab	2
## 272	tracheal aspirate	2
## 274	tracheostomy site swab	2
## 280	ulcer swab - left leg	2
## 282	ulcer swab - right leg	2
## 285	unknown1	2
## 287	unknown11	2
## 288	unknown12	2
## 289	unknown13	2
## 291	unknown15	2
## 292	unknown16	2
## 303	unknown3	2
## 309	unknown35	2
## 314	unknown4	2
## 315	unknown41	2
## 318	unknown5	2
## 321	unknown8	2
## 322	unknown9	2
## 329	urine (nephrostomy)	2
## 334	wound from outpatient	2
## 336	wound site	2
## 338	wound swab (site unspecified)	2
## 8	abscess swab	1
## 9	ankle swab - right	1

## 13	arthritis aspirates	1
## 16	aspiration	1
## 18	bakery environment - bottom metal shelf on table u	1
## 19	bakery environment - concentrated whipped topping	1
## 20	bakery environment - hallway	1
## 21	bal	1
## 27	bloodstream of an adult female icu patient	1
## 45	buttock abscess; community aquired	1
## 46	buttock swab - left	1
## 48	catheter specimen urine	1
## 50	cerebrospinal fluid	1
## 52	chest cavity abscess	1
## 55	child in a hospital pneumology ward	1
## 57	clinical sample	1
## 61	community aquired	1
## 63	cough swab	1
## 66	darccocystitis	1
## 69	diced chicken	1
## 70	doctor's hands	1
## 71	drain	1
## 73	drain site swab	1
## 74	ear swab	1
## 75	ear swab - left	1
## 76	elbow swab - right	1
## 83	eye swab - right	1
## 84	face swab	1
## 88	fatting pig at farm	1
## 96	foot swab - right	1
## 97	foot ulcer of a diabetic patient	1
## 100	gastrostomy site swab	1
## 107	hand swab - right	1
## 108	hardware	1
## 110	heel swab - left	1
## 112	hexachlorocyclohexane-contaminated soil	1
## 113	high vaginal swab	1
## 116	hip - left	1
## 117	hospital environment	1
## 121	human abscess	1
## 127	human urine	1
## 139	kidney infection	1
## 141	knee swab - left	1
## 145	lab strain	1
## 146	lean turkey	1
## 147	leg abcess	1
## 149	leg swab	1
## 154	liver cyst	1
## 156	lower jaw abscess	1
## 159	lung infection	1
## 161	lungs of cystic fibrosis patient b	1
## 162	lungs of cystic fibrosis patient c	1
## 164	mid-stream urine	1
## 166	minced pork	1
## 167	mouth swab	1
## 173	mrna screen - throat	1

## 179	nasal sample	1
## 181	neck swab	1
## 185	nose swab	1
## 186	nostril	1
## 195	penis	1
## 196	pericardic fluid	1
## 197	pericardium infection infection	1
## 201	peritoneal fluid inpatient/outpatient	1
## 202	peritoneum fluid	1
## 204	pernasal swab	1
## 207	pool	1
## 210	pork valentine	1
## 214	purulent sputum, cardio thoracic surgery	1
## 217	raw chicken	1
## 218	raw pork mince	1
## 219	raw turkey	1
## 220	rectal swab	1
## 224	respiratory; pharyngeal smear	1
## 225	right ankle	1
## 229	secretion left hip	1
## 230	secretion surgical	1
## 231	sepsis patient	1
## 233	sheep abscess	1
## 243	spleen infection	1
## 245	sputum - cf	1
## 247	sputum induced	1
## 248	staphylococcus aureus usa 300	1
## 251	stool of child with non-specific diarrhea	1
## 252	sub-cutaneous abscess	1
## 253	surface	1
## 254	surgical ward	1
## 260	thigh	1
## 261	thigh swab -right	1
## 267	toe swab	1
## 268	toe swab - left	1
## 271	trachael aspirate	1
## 273	tracheal secretion	1
## 275	transtracheal aspirate fluid	1
## 276	ulcerated maxilla	1
## 278	ulcer swab - key-in	1
## 279	ulcer swab - left foot	1
## 281	ulcer swab - right heel	1
## 283	umbilical swab	1
## 286	unknown10	1
## 290	unknown14	1
## 293	unknown17	1
## 294	unknown19	1
## 295	unknown2	1
## 296	unknown20	1
## 297	unknown21	1
## 299	unknown23	1
## 300	unknown24	1
## 302	unknown28	1
## 304	unknown30	1

```
## 305                unknown31      1
## 306                unknown32      1
## 307                unknown33      1
## 308                unknown34      1
## 310                unknown36      1
## 311                unknown37      1
## 312                unknown38      1
## 313                unknown39      1
## 316                unknown42      1
## 317                unknown43      1
## 319                unknown6       1
## 324                urethral meatus 1
## 325                urinary catheter site swab 1
## 327                urine collection bag 1
## 328                urine from long term care facility 1
## 330                vaginal tampon 1
## 331                ventral vulva abscess 1
## 332                veterinary school 1
```

Number of Isolation Sources

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

## [1] 338
```