Results Section: Public Metadata

```
library(staphopia)
library(ggplot2)
library(reshape2)
```

Aggregating Data For Public Samples

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

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

Variation From *S. aureus* N315

In Staphopia all samples had variants (SNPs and InDels) called using *S. aureus* N315 as the reference genome. In this section we'll visualize the total number of variants each sample has. This will give us an idea of the sequenced genitic diversity with respect to N315.

Gather Variant Counts

We will use `get_variant_counts()` to get the variant counts for each sample. We will also order the counts by the total.

```
variant_counts <- get_variant_counts(ps$sample_id)
variant_counts <- variant_counts[order(total),]
```

Summary of Variant Counts

Total Variants (SNPs and InDels)

```
summary(variant_counts$total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10   19457   23891   26505   37343  146962
```

SNPs

```
summary(variant_counts$snp_count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       6   18712   23162   25560   36062  141893
```

InDels

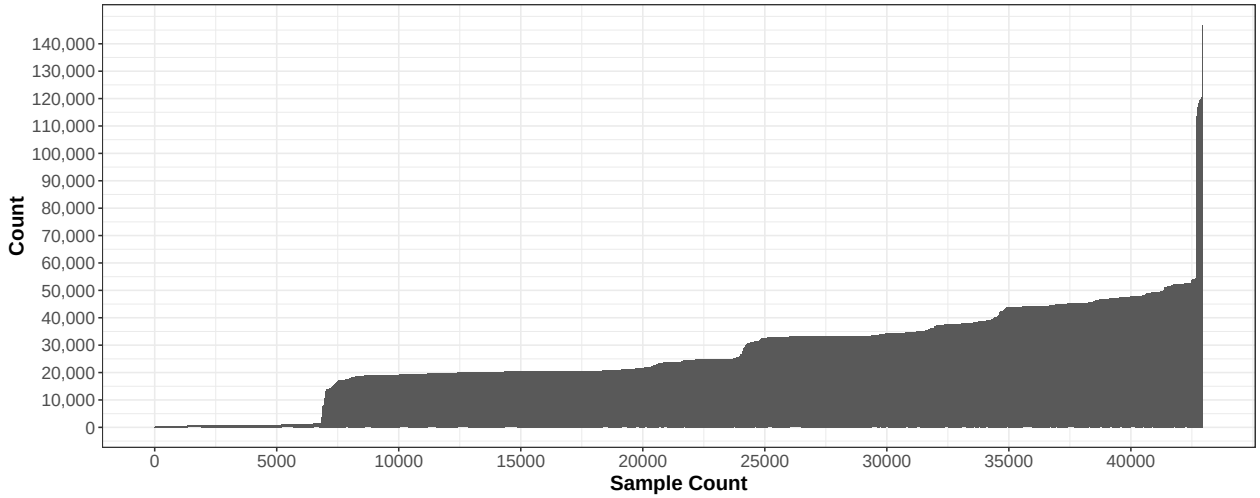
```
summary(variant_counts$indel_count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   709.0   901.0   944.4  1293.0  5125.0
```

Visualizing Variant Counts

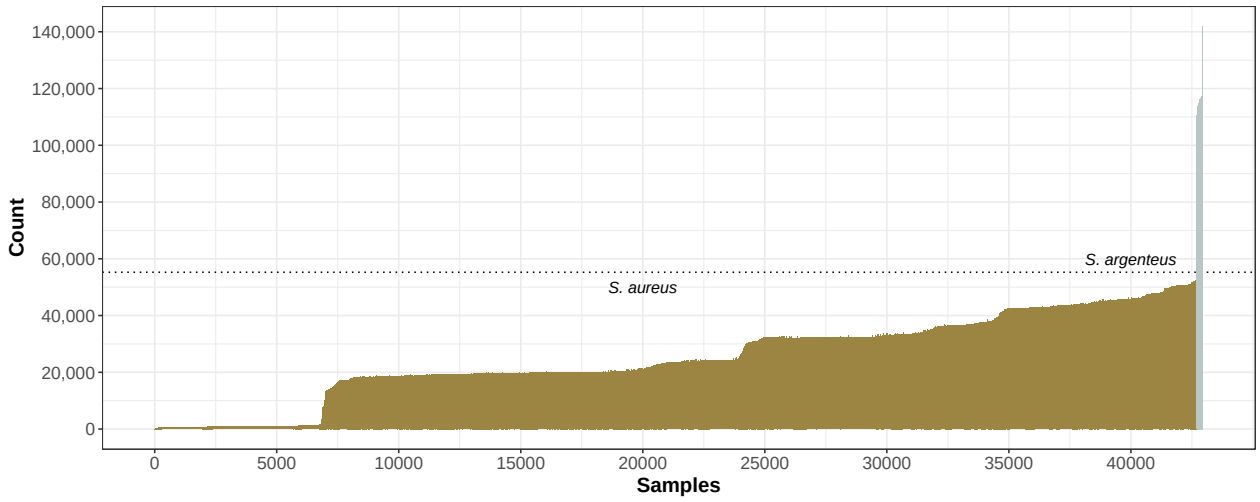
Total Variants (SNPs and InDels)

```
p <- ggplot(data=variant_counts, aes(x=seq(1,nrow(variant_counts)), y=total)) +
    xlab("Sample Count") +
    ylab("Count") +
    geom_bar(stat='identity') +
    scale_x_continuous(breaks = seq(0, nrow(variant_counts), by = 5000)) +
    scale_y_continuous(breaks = seq(0, max(variant_counts$total), by=10000), labels = scales::comma) +
    theme_bw() +
    theme(axis.text=element_text(size=12),
          axis.title=element_text(size=14,face="bold"))
p
```



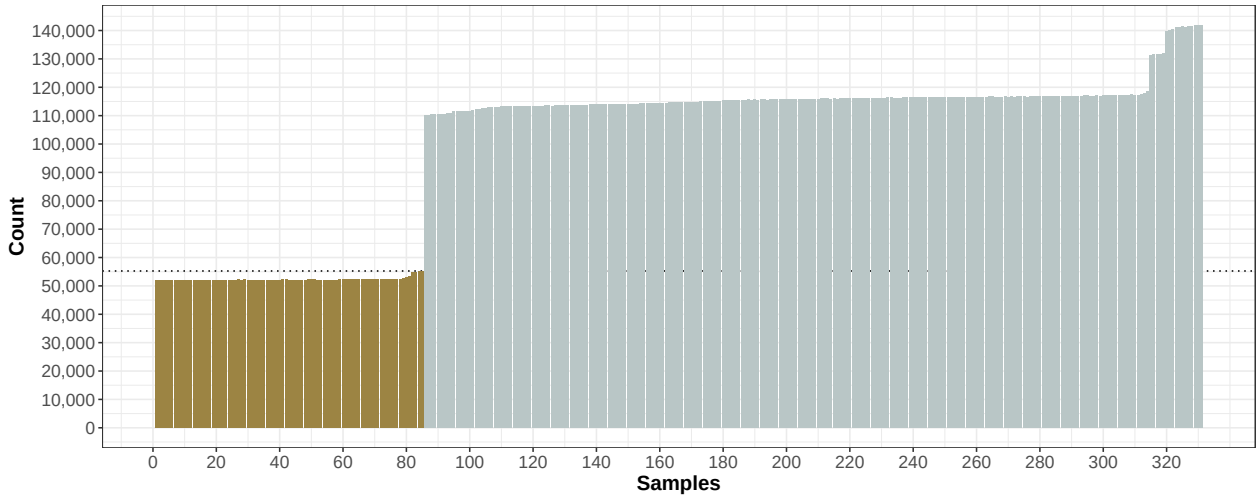
SNPs Only

```
cutoff <- max(variant_counts[variant_counts$snp_count < 60000,]$snp_count)
variant_counts$fill <- ifelse(variant_counts$snp_count > cutoff, TRUE, FALSE)
p <- ggplot(data=variant_counts, aes(x=seq(1,nrow(variant_counts)), y=snp_count, fill=fill)) +
    xlab("Samples") +
    ylab("Count") +
    geom_hline(yintercept = cutoff, linetype="dotted") +
    geom_bar(stat='identity') +
    annotate("text", x = 40000, y = 60000, label = "S. argenteus", fontface=3) +
    annotate("text", x = 20000, y = 50000, label = "S. aureus", fontface=3) +
    scale_x_continuous(breaks = seq(0, nrow(variant_counts), by = 5000)) +
    scale_y_continuous(breaks = seq(0, max(variant_counts$snp_count), by=20000), labels = scales::comma
    scale_fill_manual(values=c("#9C8443", "#B9C6C6")) +
    theme_bw() +
    theme(axis.text=element_text(size=12),
          axis.title=element_text(size=14,face="bold"),
          legend.position="none")
p
```




```
# Output plot to PDF and PNG
staphopia::write_plot(p, paste0(getwd(), '/../figures/figure-09-snp-accumulation'))
```

```
cutoff <- max(variant_counts[variant_counts$snp_count < 60000,]$snp_count)
variant_counts$fill <- ifelse(variant_counts$snp_count > cutoff, TRUE, FALSE)
p <- ggplot(data=variant_counts[variant_counts$snp_count > 52000,], aes(
        x=seq(1,nrow(variant_counts[variant_counts$snp_count > 52000,])),
        y=snp_count,
        fill=fill)
    ) +
    xlab("Samples") +
    ylab("Count") +
    geom_hline(yintercept = cutoff, linetype="dotted") +
    geom_bar(stat='identity') +
    scale_x_continuous(breaks = seq(0, nrow(variant_counts[variant_counts$snp_count > 52000,]), by = 20
    scale_y_continuous(breaks = seq(0, max(variant_counts$snp_count), by=10000), labels = scales::comma
    scale_fill_manual(values=c("#9C8443", "#B9C6C6")) +
    theme_bw() +
    theme(axis.text=element_text(size=12),
          axis.title=element_text(size=14,face="bold"),
          legend.position="none")
p
```



InDels Only

```
p <- ggplot(data=variant_counts, aes(x=seq(1,nrow(variant_counts)), y=indel_count)) +
    xlab("Sample Count") +
    ylab("Count") +
    geom_bar(stat='identity') +
    scale_x_continuous(breaks = seq(0, nrow(variant_counts), by = 5000)) +
    scale_y_continuous(breaks = seq(0, max(variant_counts$indel_count), by=500), labels = scales::comma
    theme_bw() +
    theme(axis.text=element_text(size=12),
          axis.title=element_text(size=14,face="bold"))
p
```