qc

Kate Weaver

12/20/2021

```
library(SummarizedExperiment)
```

```
load("~/mccoyLab/collabs/doubleseq_2021/summarized_experiment/create_summarized_experiment_allNov2021.R
```

```
counts <- as.matrix(assays(seAll)$counts)
```

```
library(rtracklayer)
library(tidyverse)

gtf <- rtracklayer::import("~/genomes/hg38_genome/gencode.v34.annotation.gtf") %>%
  as.data.frame() %>%
  filter(type == "gene") %>%
  select(gene_id, seqnames) %>%
  dplyr::rename(ensembl_gene_id = gene_id) %>%
  dplyr::rename(chromosome_name = seqnames)

gene_table <- gtf[match(rownames(counts), gtf$ensembl_gene_id),]
gene_table <- gene_table[gene_table$chromosome_name %in% paste0("chr", 1:22),]

counts_genes <- counts[gene_table$ensembl_gene_id,]
```

```
dim(counts)
```

```
## [1] 60669   143
```

```
dim(counts_genes)
```

```
## [1] 57640   143
```

```
gene_sum = colSums(counts_genes > 0)
length(gene_sum)
```

```
## [1] 143
```

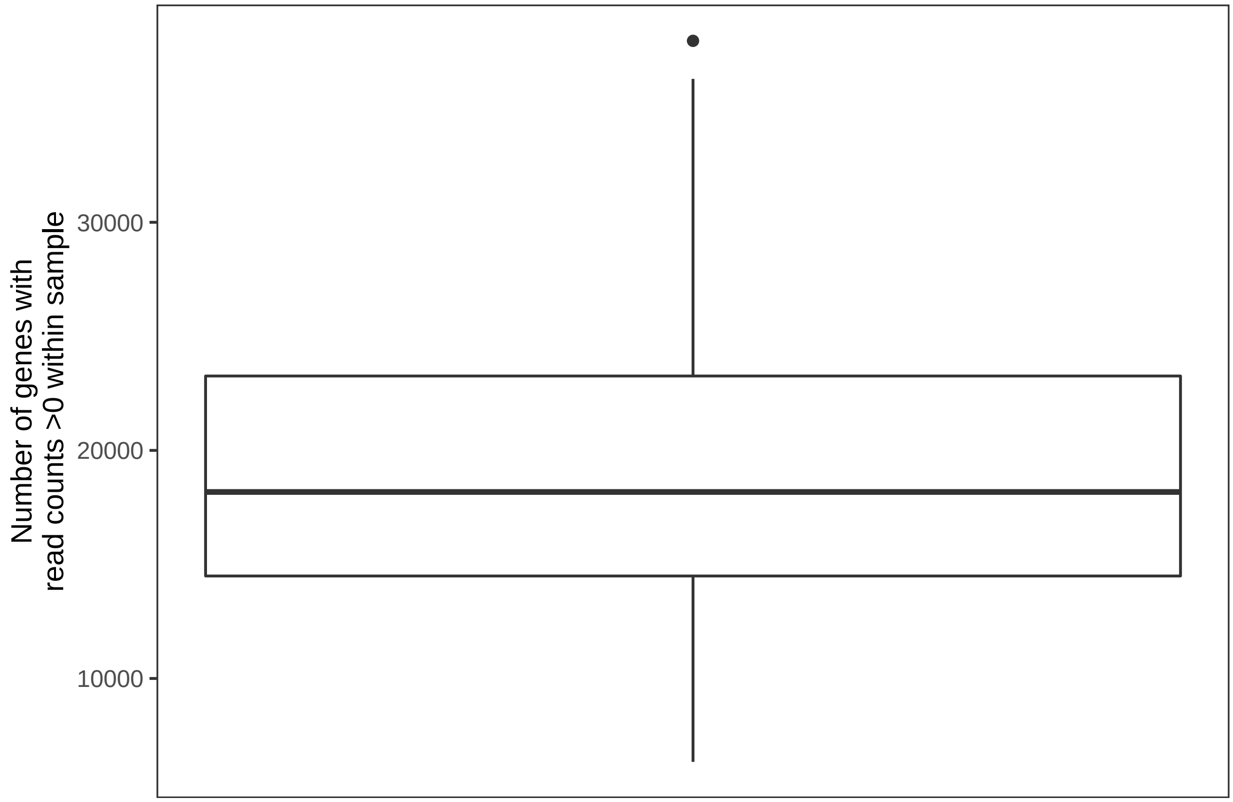
```
mean(gene_sum)
```

```
## [1] 19442.31
```

```
sd(gene_sum)
```

```
## [1] 6726.621
```

```
library(ggplot2)
gene_df <- data.frame(sumgenes = gene_sum)
ggplot(gene_df, aes(x=NULL, y=sumgenes)) + geom_boxplot() + ylab("Number of genes with\nread counts >0
```



```
sample_sum = rowSums(counts_genes > 0)
length(sample_sum)
```

```
## [1] 57640
```

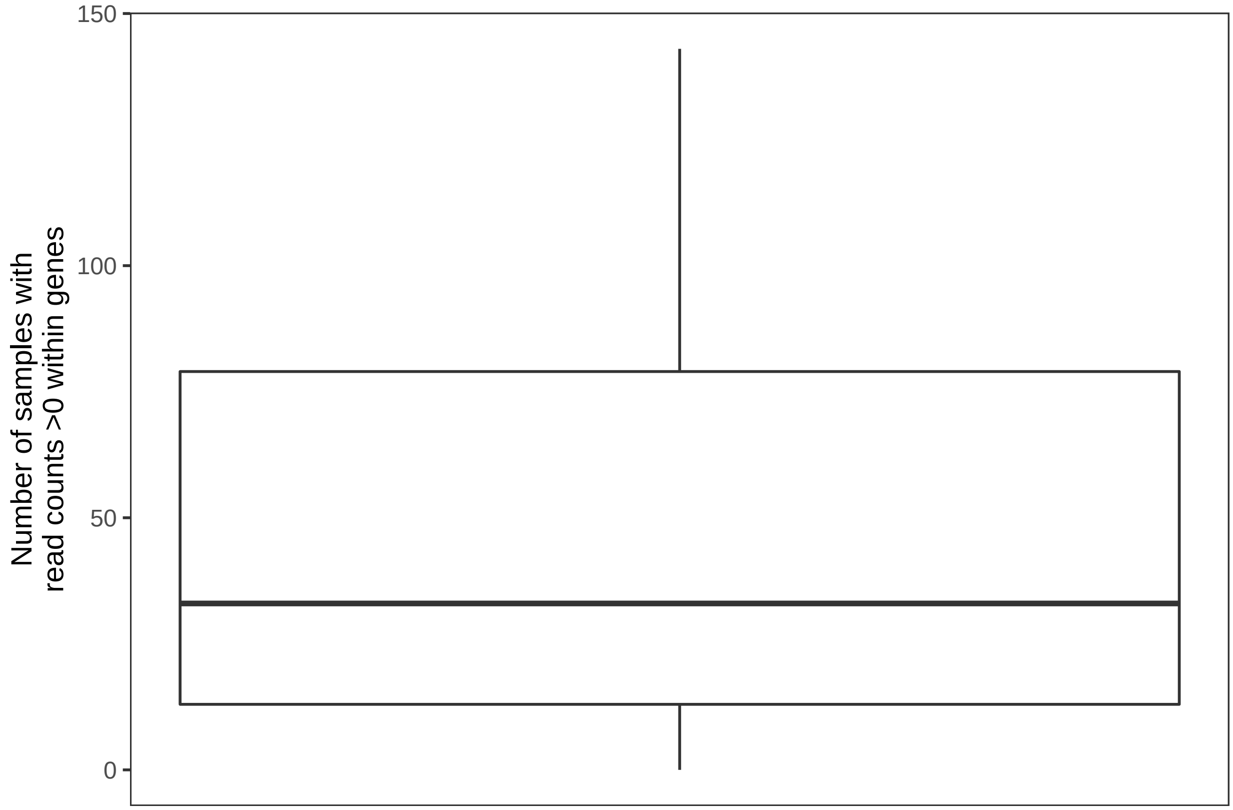
```
mean(sample_sum)
```

```
## [1] 48.23475
```

```
sd(sample_sum)
```

```
## [1] 42.69078
```

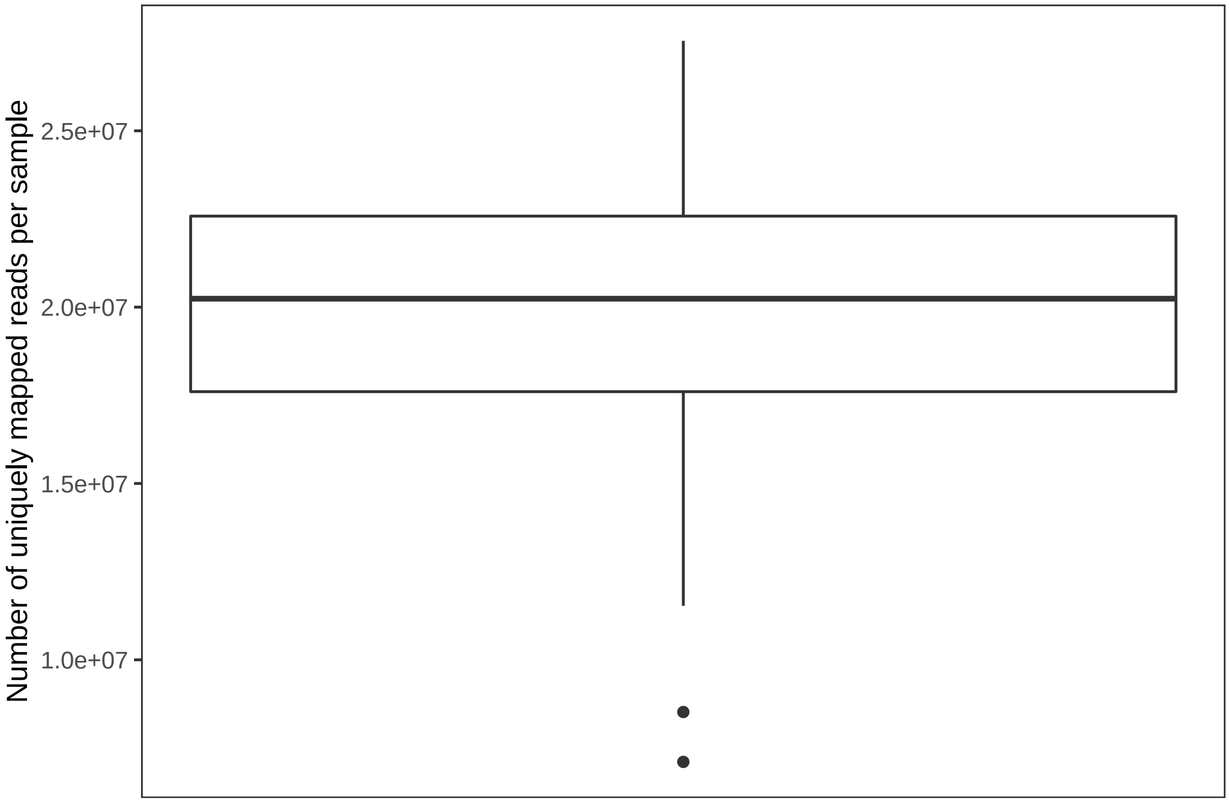
```
sample_df <- data.frame(sumsample = sample_sum)
ggplot(sample_df, aes(x=NULL, y=sumsample)) + geom_boxplot() + ylab("Number of samples with\nread counts
```



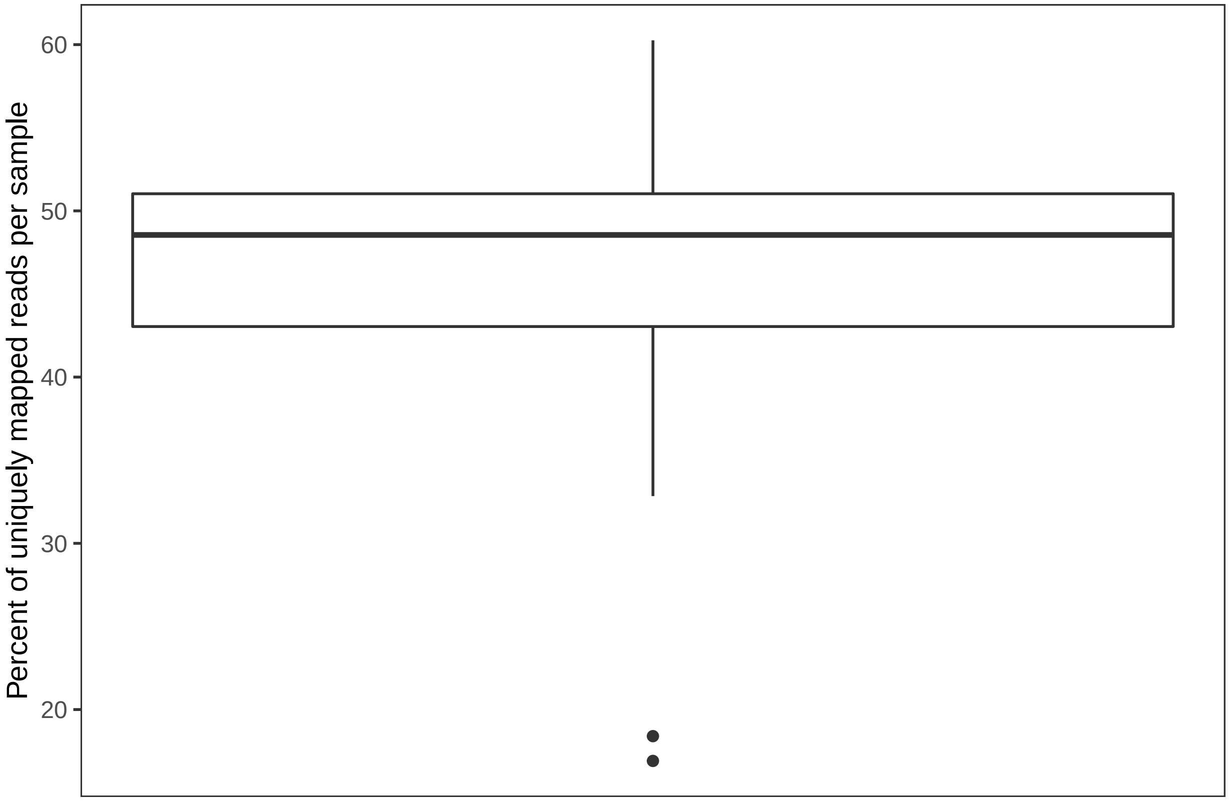
```
samples <- read.delim("~/mccoyLab/collabs/doubleseq_2021/summarized_experiment/qc_files/sample_allNov202
uniq_read_num <- read.delim("~/mccoyLab/collabs/doubleseq_2021/summarized_experiment/qc_files/uniq_reads
uniq_read_per <- read.delim("~/mccoyLab/collabs/doubleseq_2021/summarized_experiment/qc_files/uniq_reads

uniq_read_df <- data.frame(sample = samples, num = uniq_read_num, percent = uniq_read_per)
two_samples <- uniq_read_df %>% dplyr::arrange(percent)
two_samples <- two_samples[c(1:2),]
two_samples$sample
```

```
## [1] "163-A-1-L_S19" "205-A-5-L_S40"
```

```
ggplot(uniq_read_df, aes(x=NULL, y=num)) + geom_boxplot() + ylab("Number of uniquely mapped reads per sa
```



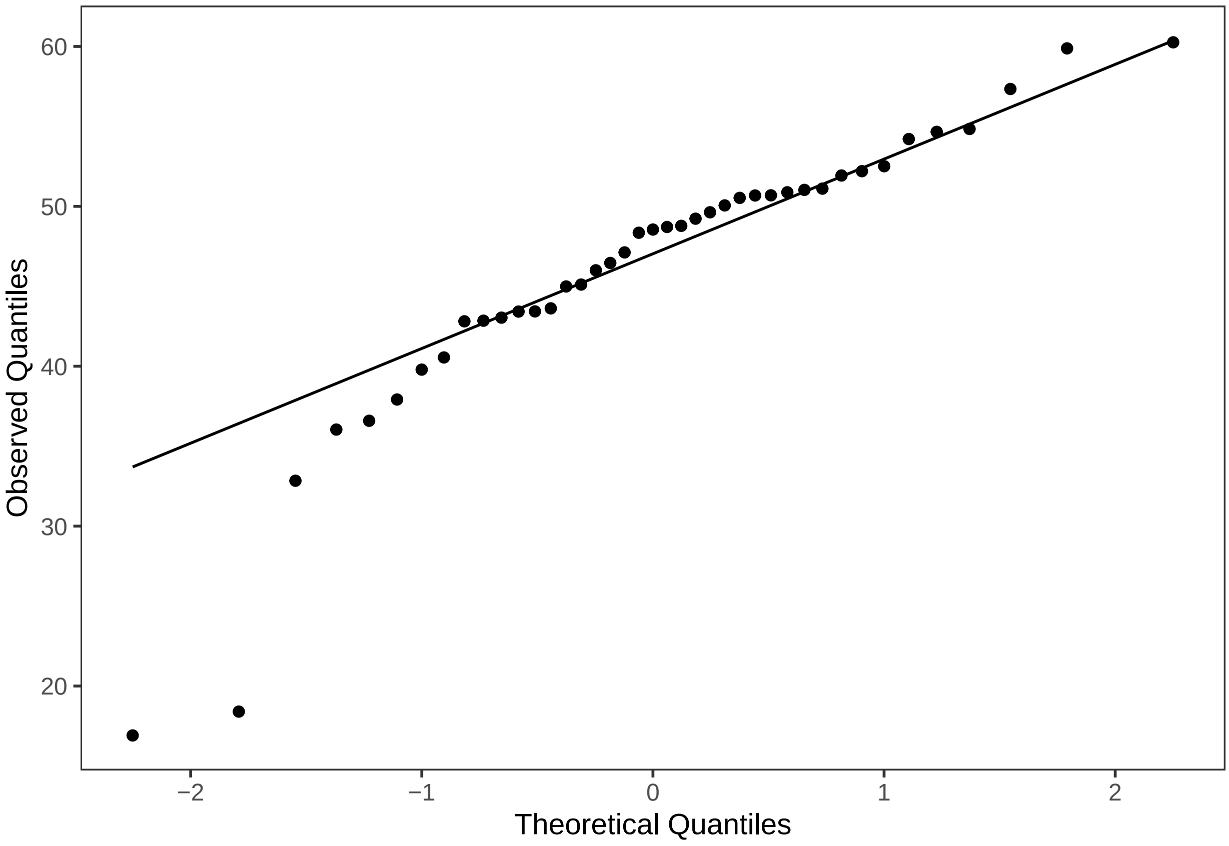
```
ggplot(uniq_read_df, aes(x=NULL, y=percent)) + geom_boxplot() + ylab("Percent of uniquely mapped reads p
```



```
shapiro.test(uniq_read_df$percent)
```

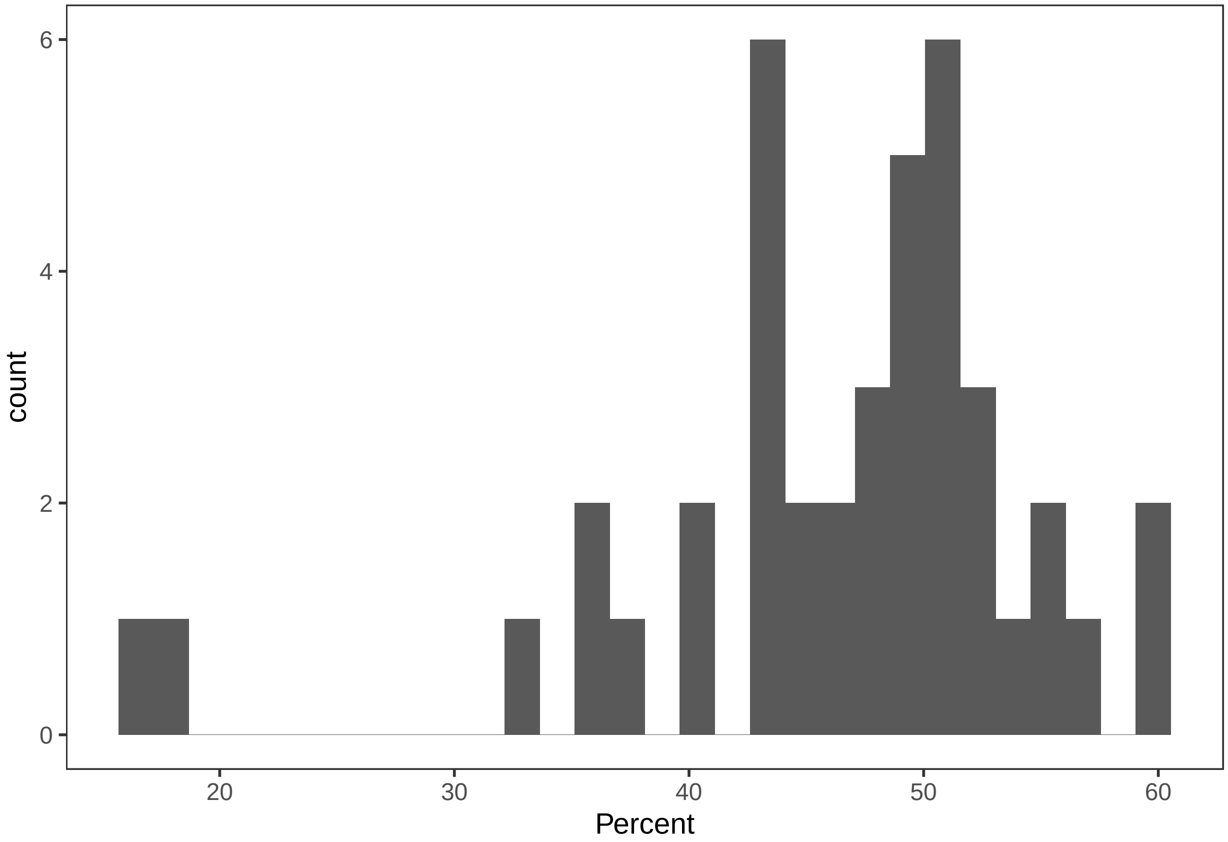
```
##
##  Shapiro-Wilk normality test
##
## data:  uniq_read_df$percent
## W = 0.87656, p-value = 0.0003616
```

```
ggplot(uniq_read_df, aes(sample=percent)) + geom_qq() + geom_qq_line() + theme_bw() + theme(panel.backg
```



```
ggplot(uniq_read_df, aes(x=percent)) + geom_histogram() + theme_bw() + theme(panel.background = element_
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
library(EnvStats)
```

```
rosnerTest(uniq_read_df$percent, k=3, warn=TRUE)
```

```
## $distribution
## [1] "Normal"
##
## $statistic
##      R.1      R.2      R.3
## 3.263874 3.681921 2.353409
##
## $sample.size
## [1] 41
##
## $parameters
## k
## 3
##
## $alpha
## [1] 0.05
##
## $crit.value
## lambda.1 lambda.2 lambda.3
## 3.046571 3.036097 3.025284
##
## $n.outliers
## [1] 2
##
```

```
## $alternative
## [1] "Up to 3 observations are not\n                                 from the same Distribution."
##
## $method
## [1] "Rosner's Test for Outliers"
##
## $data
##  [1] 51.93 42.81 45.11 39.79 54.21 48.71 54.84 51.03 50.68 43.04 36.59 48.35
## [13] 16.91 37.92 42.85 49.23 46.46 32.84 43.43 46.00 44.99 50.06 49.63 36.04
## [25] 51.11 52.20 48.55 40.55 48.78 52.51 57.34 47.12 50.88 18.40 50.69 43.42
## [37] 50.53 60.26 59.88 54.66 43.62
##
## $data.name
## [1] "uniq_read_df$percent"
##
## $bad.obs
## [1] 0
##
## $all.stats
##   i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 46.19390 8.972130 16.91      13 3.263874   3.046571    TRUE
## 2 1 46.92600 7.747587 18.40      34 3.681921   3.036097    TRUE
## 3 2 47.65744 6.296159 32.84      18 2.353409   3.025284   FALSE
##
## attr(,"class")
## [1] "gofOutlier"
```

```
mean(uniq_read_df$percent) + (2 * sd(uniq_read_df$percent))
```

```
## [1] 64.13816
```

```
mean(uniq_read_df$percent) - (2 * sd(uniq_read_df$percent))
```

```
## [1] 28.24964
```

```
mean(uniq_read_df$percent) - (3 * sd(uniq_read_df$percent))
```

```
## [1] 19.27751
```

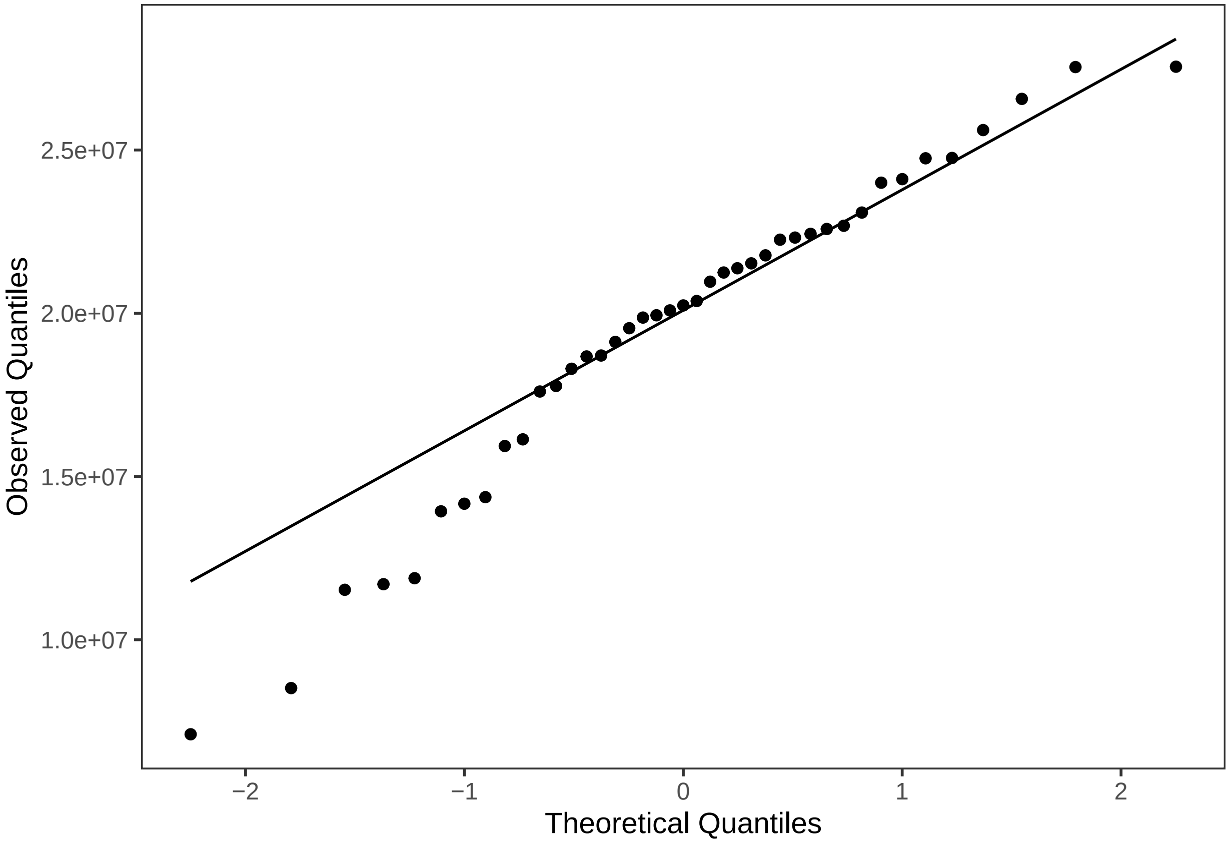
```
mean(uniq_read_df$percent) - (4 * sd(uniq_read_df$percent))
```

```
## [1] 10.30538
```

```
shapiro.test(uniq_read_df$num)
```

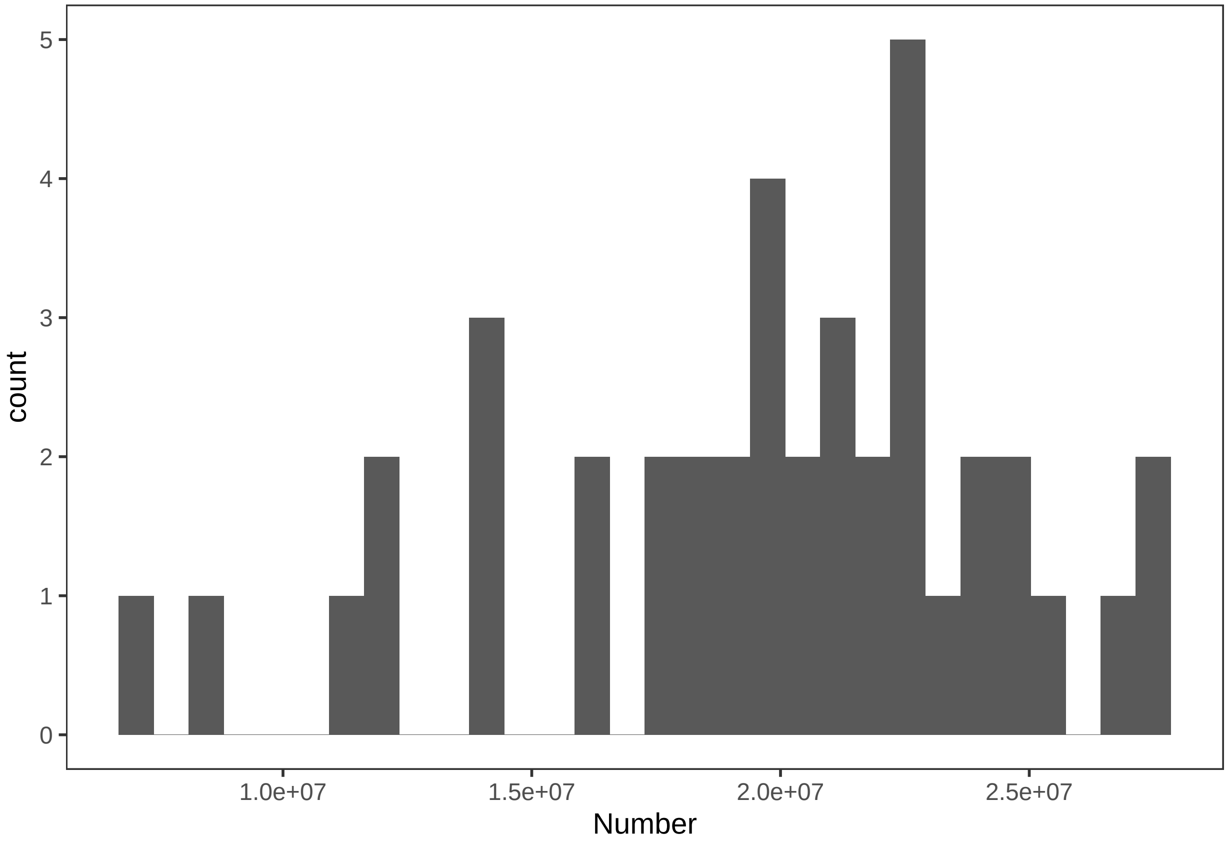
```
##
##  Shapiro-Wilk normality test
##
## data:  uniq_read_df$num
## W = 0.95586, p-value = 0.1126
```

```
ggplot(uniq_read_df, aes(sample=num)) + geom_qq() + geom_qq_line() + theme_bw() + theme(panel.backgroun
```



```
ggplot(uniq_read_df, aes(x=num)) + geom_histogram() + theme_bw() + theme(panel.background = element_blan
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
rosnerTest(uniq_read_df$num, k=3, warn=TRUE)
```

```
## $distribution
## [1] "Normal"
##
## $statistic
##      R.1      R.2      R.3
## 2.538399 2.500707 2.054527
##
## $sample.size
## [1] 41
##
## $parameters
## k
## 3
##
## $alpha
## [1] 0.05
##
## $crit.value
## lambda.1 lambda.2 lambda.3
## 3.046571 3.036097 3.025284
##
## $n.outliers
## [1] 0
##
## $alternative
## [1] "Up to 3 observations are not\n                                 from the same Distribution."
##
## $method
## [1] "Rosner's Test for Outliers"
##
## $data
##  [1] 24745475 19937382 22580639 20375163 25609866 17771933 26564955 19541303
##  [9] 19121735 14367199 11701856 13932671  7105408 11885209 20966915 22431906
## [17] 22679341 14166180 20239139 21528015 17602642 22253050 23998516 15933764
## [25] 18705612 22317550 21377754 18300084 16137505 19867051 11530093 24106916
## [33] 21247315  8519313 23084070 18676659 21773229 27550907 27539621 24755541
## [41] 20087601
##
## $data.name
## [1] "uniq_read_df$num"
##
## $bad.obs
## [1] 0
##
## $all.stats
##   i   Mean.i    SD.i    Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 19576026 4912790  7105408      13 2.538399   3.046571   FALSE
## 2 1 19887792 4546105  8519313      34 2.500707   3.036097   FALSE
## 3 2 20179291 4209825 11530093      31 2.054527   3.025284   FALSE
##
## attr(,"class")
## [1] "gofOutlier"
```

```
mean(uniq_read_df$num) + (2 * sd(uniq_read_df$num))
```

```
## [1] 29401606
```

```
mean(uniq_read_df$num) - (2 * sd(uniq_read_df$num))
```

```
## [1] 9750447
```

```
mean(uniq_read_df$num) - (3 * sd(uniq_read_df$num))
```

```
## [1] 4837657
```

```
mean(uniq_read_df$num) - (4 * sd(uniq_read_df$num))
```

```
## [1] -75132.36
```