Results Section: Pipeline Design and Processing 43,000+ genomes

In this notebook will be generating statistics and plots related to processing 43,000+ genomes on Seven Bridges Cancer Genomics Cloud (CGC) platform.

Load Up Packages

```r
library(staphopia)
library(ggplot2)
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
```

Read In The Data

```r
results <- read.table("../data/cgc-runs.txt", header = TRUE, sep = "\t")
colnames(results)
```

```
## [1] "name"       "status"     "project"    "app"        "created_by"
## [6] "total_time" "run_time"   "queue_time" "price"
```

This leaves use with 9 columns:

1. name: Name of the job
2. status: Job's status
3. project: CGC project job was executed from.
4. app: CGC app used to execute the job.
5. created_by: User who submitted the job.
6. total_time: Total amount of time (in minutes) a job was queued and run
7. run_time: Total amount of time (in minutes) a job took to complete
8. queue_time: Total amount of time (in minutes) a job was queued
9. price: Total cost of the run

Clean Up The Data

Before we generate statistics and plots, we need to clean the data. There are jobs where the *run_time* and *price* were not properly reported from CGC. We will filter samples where the *run_time* is 0.

```r
results_clean <- results[results$run_time > 0, ]
nrow(results) - nrow(results_clean)
```

```
## [1] 11424
```

Job Summary

Run Time Summary

```
summary(results_clean$run_time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.75   47.26   51.23   52.39   56.26 1883.70
```

Number of Jobs With > 120 Minute Runtime

```
nrow(results_clean[results_clean$run_time > 120, ])
```

```
## [1] 160
```

Summary of Jobs With Run Time Between 10 and 120 Minutes

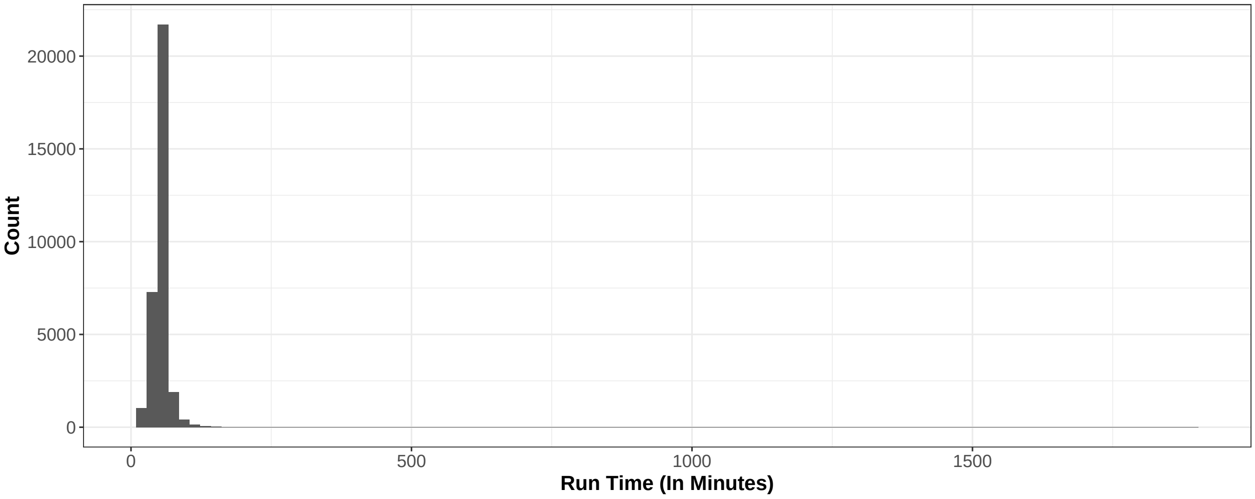
```
summary(results_clean[between(results_clean$run_time, 10, 120), ]$run_time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.83   47.24   51.19   51.76   56.11  119.79
```

Plots

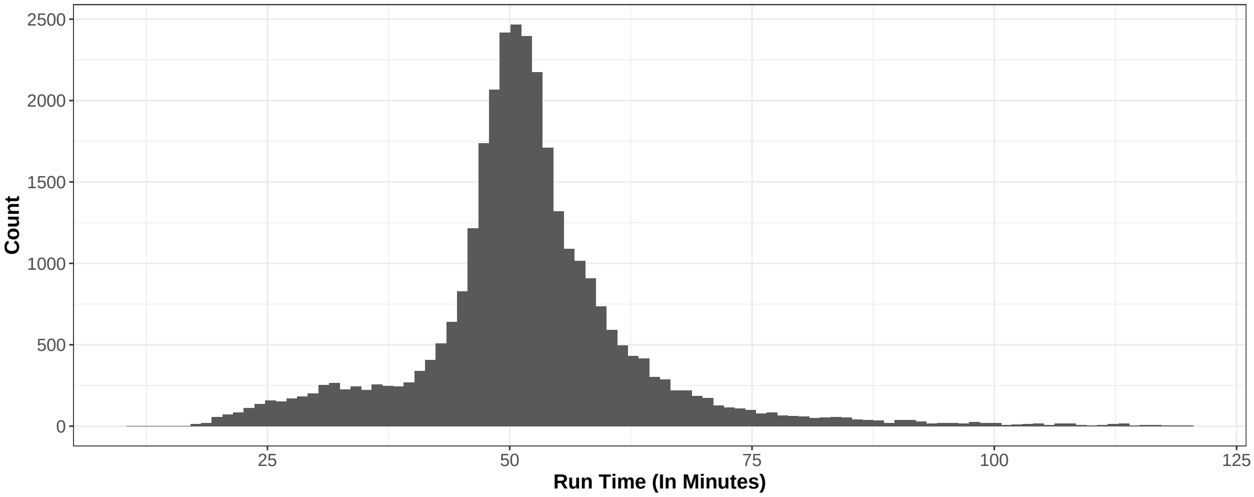
Run Time (Complete)

```
p <- ggplot(data=results_clean, aes(run_time)) +
    xlab("Run Time (In Minutes)") +
    ylab("Count") +
    geom_histogram(bins=100) +
    theme_bw() +
    theme(axis.text=element_text(size=12),
          axis.title=element_text(size=14,face="bold"))
p
```



Pipeline Run Time (Between 10-120 Minutes)

```r
p <- ggplot(data=results_clean[between(results_clean$run_time, 10, 120),], aes(run_time)) +
    xlab("Run Time (In Minutes)") +
    ylab("Count") +
    geom_histogram(bins=100) +
    theme_bw() +
    theme(axis.text=element_text(size=12),
          axis.title=element_text(size=14,face="bold"))
p
```



```r
# Output plot to PDF and PNG
staphopia::write_plot(p, paste0(getwd(), '/../figures/figure-03-pipeline-run-time'))
```

Session Info

```r
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] dplyr_0.7.4     ggplot2_2.2.1   staphopia_0.1.9
```

```
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.15     bindr_0.1.1      knitr_1.20       magrittr_1.5
##  [5] munsell_0.4.3    colorspace_1.3-2 R6_2.2.2         rlang_0.1.6
##  [9] stringr_1.2.0    plyr_1.8.4       tools_3.4.3      grid_3.4.3
## [13] gtable_0.2.0     htmltools_0.3.6  assertthat_0.2.0 yaml_2.1.18
## [17] lazyeval_0.2.1   rprojroot_1.3-2  digest_0.6.15    tibble_1.4.2
## [21] bindrcpp_0.2     glue_1.2.0       evaluate_0.10.1  rmarkdown_1.9
## [25] labeling_0.3     stringi_1.1.6    compiler_3.4.3   pillar_1.1.0
## [29] scales_0.5.0     backports_1.1.2  pkgconfig_2.0.1
```