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

```
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

```
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.

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

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

Job Summary

Run Time

```
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

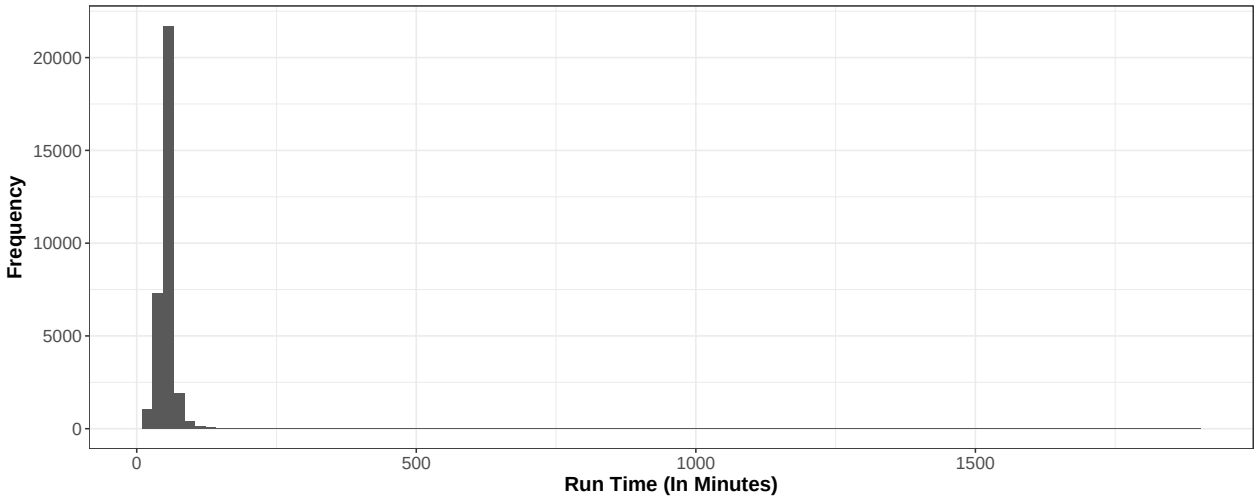
```
summary(results_clean[between(results_clean$run_time, 10, 120), ]$run_time)
```

```
##    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("Frequency") +
    geom_histogram(bins=100) +
    theme_bw() +
    theme(axis.text=element_text(size=12),
          axis.title=element_text(size=14,face="bold"))
p
```



Run Time (Between 10-120 Minutes)

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