Comparing assembly rates to two different reference genomes

Rami Baghdan

2/14/2020

R Markdown

Compare percentage of reads assembled and quantify variation among individuals.

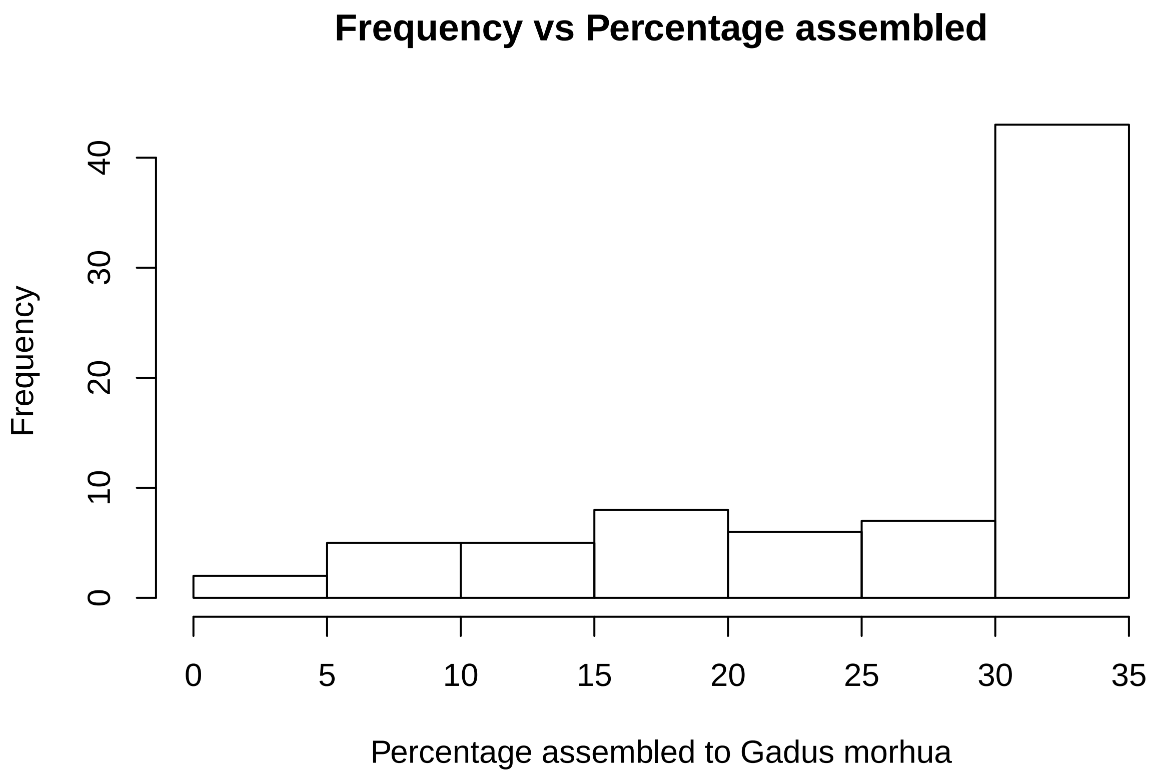
```r
library(stringr)
setwd("~/Desktop/genomicmeth")

# Read in Gadus morhua assembled files.

GMcounts <- read.delim("assembled_per_indGM.txt")

# Using stringr package, remove the path in the individual
# assembled read names.

# The word function is used to section string to obtain raw
# and assembled counts.

GMcounts$Rawcounts <- word(GMcounts$ind.raw.assembled, 2L)
GMcounts$Assembledcounts <- word(GMcounts$ind.raw.assembled,
    3L)

# Unecessary information such as the path to the fastq file
# is removed.
GMcounts$ind.raw.assembled <- str_remove(GMcounts$ind.raw.assembled,
    "/scratch/rbaghdan/genomicmethods/bwa_assem/bwa_assem2/")

GMcounts$ind.raw.assembled <- sub(" .*", "", GMcounts$ind.raw.assembled)

# The same is repeated for the new burbot genome alignment.
NBGcounts <- read.delim("assembled_per_indNBG.txt")
NBGcounts$Rawcounts <- word(NBGcounts$ind.raw.assembled, 2L)
NBGcounts$Assembledcounts <- word(NBGcounts$ind.raw.assembled,
    3L)
NBGcounts$ind.raw.assembled <- str_remove(NBGcounts$ind.raw.assembled,
    "/scratch/rbaghdan/genomicmethods/bwa_assem/bwa_assem3/")
NBGcounts$ind.raw.assembled <- sub(" .*", "", NBGcounts$ind.raw.assembled)

# Now, the percentage of reads that were assembled is found.
# Convert counts to a numeric class.
```

```
GMcounts$Rawcounts <- as.numeric(GMcounts$Rawcounts)
GMcounts$Assembledcounts <- as.numeric(GMcounts$Assembledcounts)
GMcounts$percentage <- GMcounts$Assembledcounts/GMcounts$Rawcounts *
    100
mean(GMcounts$percentage)
```

```
## [1] 25.41559
```

```
# The average percentage of assembled reads to the Gadus
# Morhua complete reference genome is 25.4%.
```

```
hist(GMcounts$percentage, main = "Frequency vs Percentage assembled",
    xlab = "Percentage assembled to Gadus morhua")
```



The histogram illustrates that the majority of the reads align to the Gadus morhua reference genome at a 30-35% rate.

```
summary(GMcounts$percentage)
```

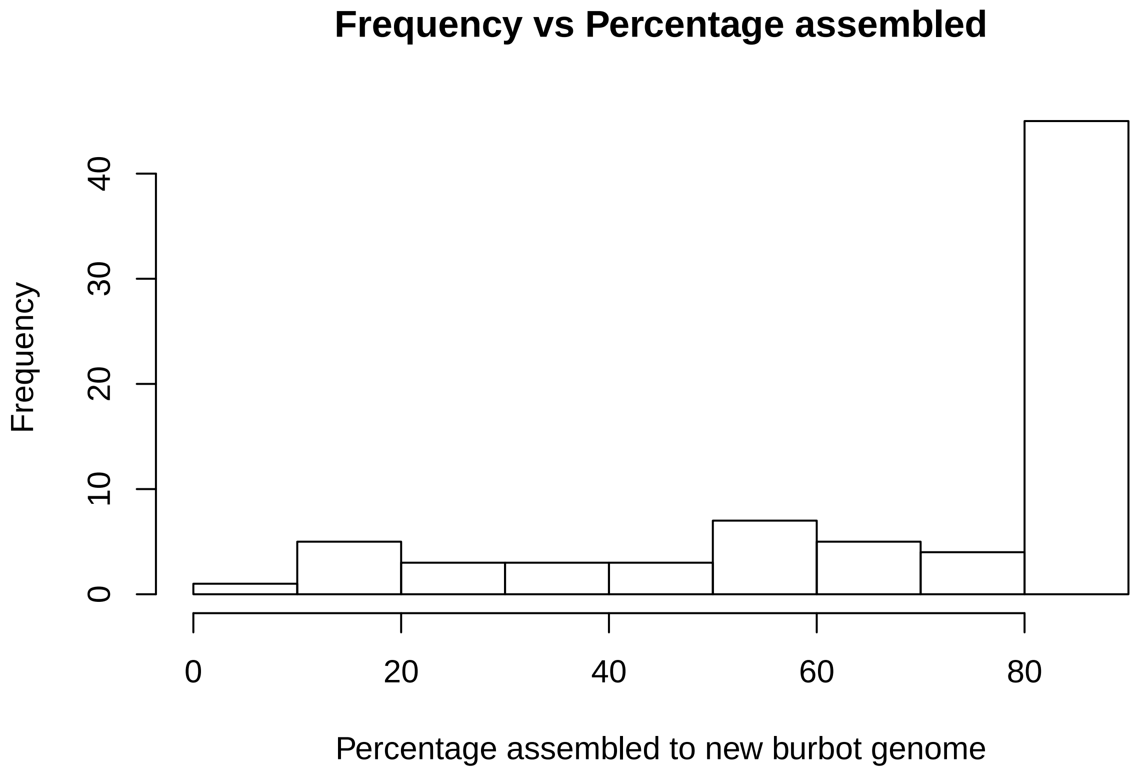
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.596  19.748  30.836  25.416  31.872  32.640
```

The minimum percentage of reads assembled for an individual was 3.6% and the maximum was 32.6% with an average of 25.4%. The median of 30.8% illustrates that there are some reads with very low alignment rates that are bringing down the overall average rate.

```
# This was done for the new burbot reference genome alignment
# as well.
NBGcounts$Rawcounts <- as.numeric(NBGcounts$Rawcounts)
NBGcounts$Assembledcounts <- as.numeric(NBGcounts$Assembledcounts)
NBGcounts$percentage <- NBGcounts$Assembledcounts/NBGcounts$Rawcounts *
    100
mean(NBGcounts$percentage)
```

```
## [1] 70.34796
```

```
hist(NBGcounts$percentage, main = "Frequency vs Percentage assembled",
    xlab = "Percentage assembled to new burbot genome")
```



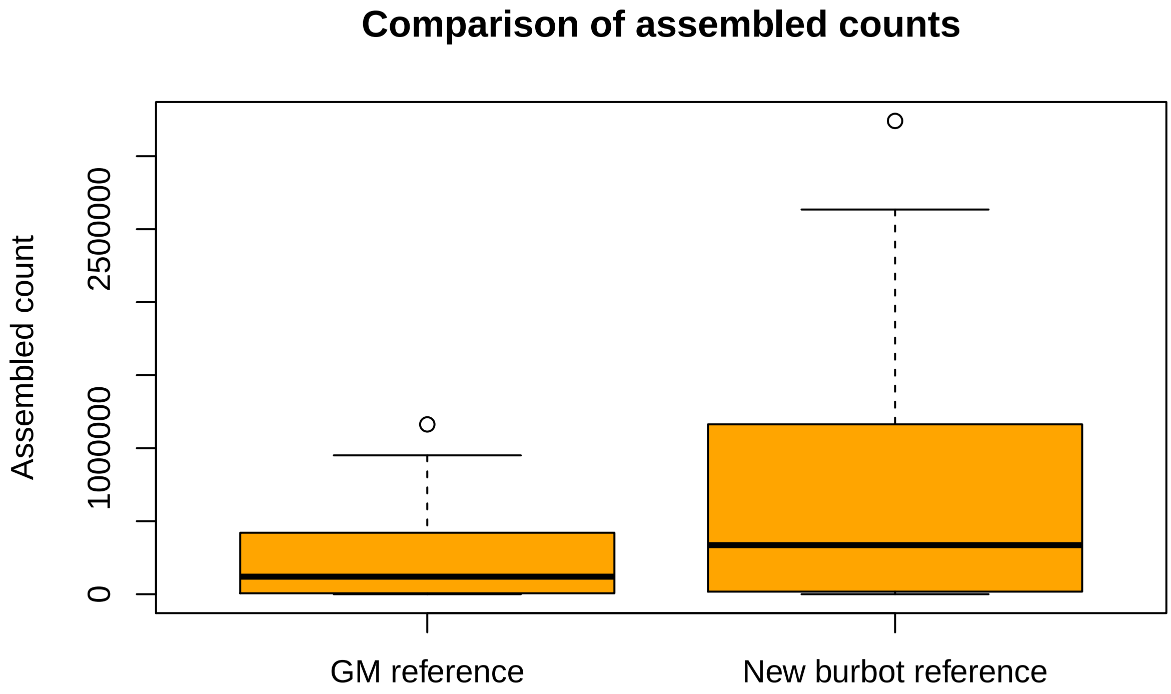
The histogram illustrates that the majority of the reads align to the reference at >80% rates.

```
summary(NBGcounts$percentage)
```

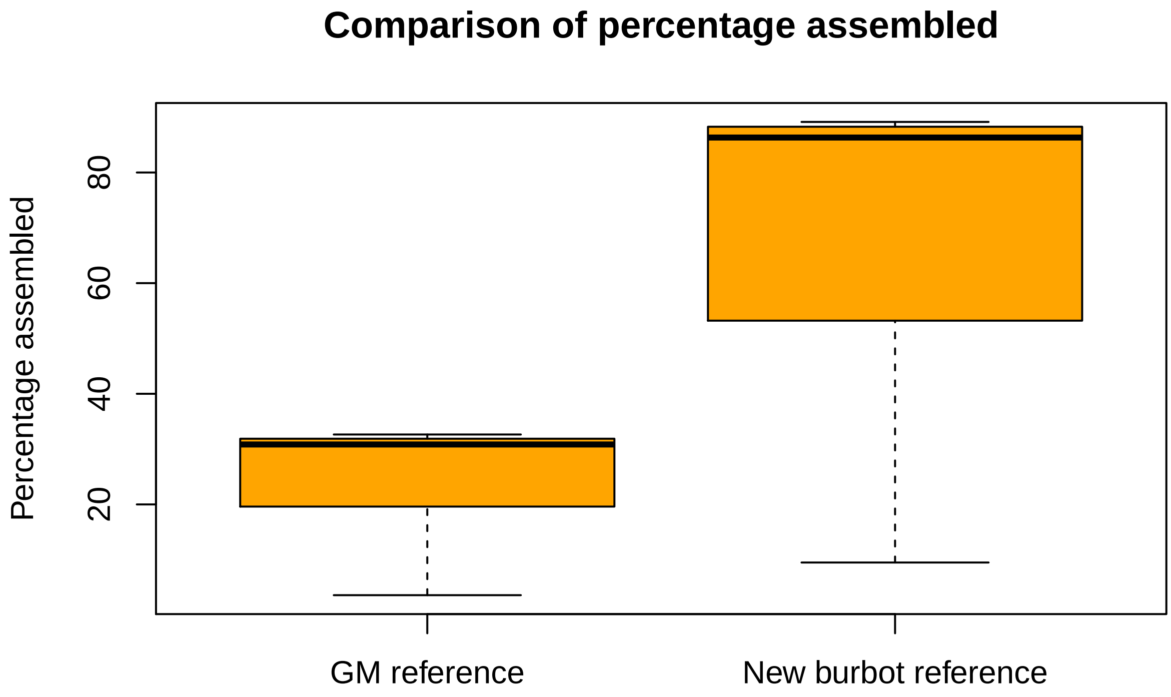
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   9.497  53.551  86.324  70.348  88.255  89.142
```

The minimum percentage of reads assembled for an individual was 9.5% and the maximum was 89.1% with an average of 70.3% of reads assembled. The median at 86.3% shows that there are a few reads with very low alignment rates bringing down the overall average alignment rate.

```r
# Visualize difference in counts and percentages between the
# two alignments using boxplots.

# Assembled counts.
boxplot(GMcounts$Assembledcounts, NBGcounts$Assembledcounts,
    main = "Comparison of assembled counts", names = c("GM reference",
        "New burbot reference"), ylab = "Assembled count", col = "orange")
```



```r
# Percentage assembled.

boxplot(GMcounts$percentage, NBGcounts$percentage, main = "Comparison of percentage assembled",
    names = c("GM reference", "New burbot reference"), ylab = "Percentage assembled",
    col = "orange")
```



The boxplots illustrate that reads among individuals aligned to the new burbot genome have higher assembled counts and higher alignment rates than those aligned to the Gadus morhua genome. I checked to see if there is a statistically significant difference between the two alignments regarding assembled percentage and assembled counts.

```
t.test(GMcounts$percentage, NBGcounts$percentage)
```

```
##
##  Welch Two Sample t-test
##
## data:  GMcounts$percentage and NBGcounts$percentage
## t = -14.968, df = 93.992, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -50.89280 -38.97193
## sample estimates:
## mean of x mean of y
##  25.41559  70.34796
```

```
t.test(GMcounts$Assembledcounts, NBGcounts$Assembledcounts)
```

```
##
##  Welch Two Sample t-test
##
## data:  GMcounts$Assembledcounts and NBGcounts$Assembledcounts
## t = -4.3851, df = 94.128, p-value = 3.019e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -624311.8 -235160.9
## sample estimates:
## mean of x mean of y
##  242444.6  672180.9
```

The burbot individuals have a statistically significant higher assembly rate to the new burbot reference genome than to the Gadus morhua complete genome.