Esther Olabisi-Adeniyi Student ID: 1238826
Project 2: Alignment of burbot sequences

Introduction

Genome assembly can be achieved in two different ways: de novo assembly and reference-based assembly. In de novo assembly, sequence reads are arranged without any reference template whereas reference-based assembly utilizes an already existing genome assembly as a template for aligning sequence reads. Having a reference genome is beneficial when it is from the same species as the sequence reads; however, when the only available option is that of a related species, genetic divergence among both groups could interfere with the mapping of conserved regions thereby interfering with the quality of alignment (Lou et al., 2021, p. 5969). In another instance, alignment to the reference genome of a different species can be useful for investigating evolutionary relationships between the two species (Gopalakrishnan et al., 2017). To investigate the impact of reference genomes on sequence alignment, I aligned raw reads of 10 burbot individuals to both a fragmented burbot reference genome and the high-quality reference genome of a related species, the cod fish.

Results

The following results were obtained from the alignment of raw sequence reads from 10 burbot individuals to burbot and cod reference genomes.

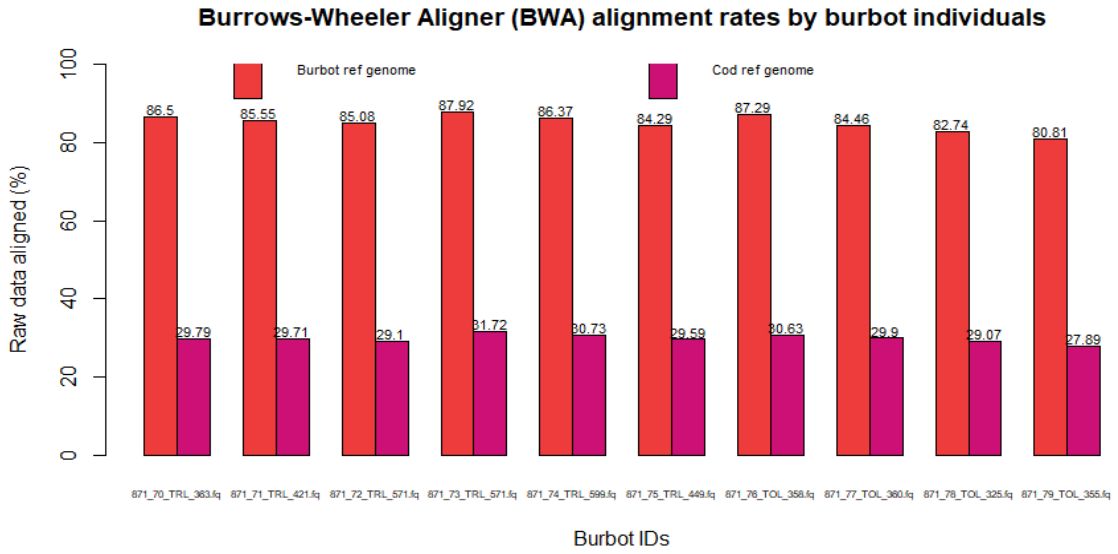

Fig. 1: BWA alignment success by burbot individuals.

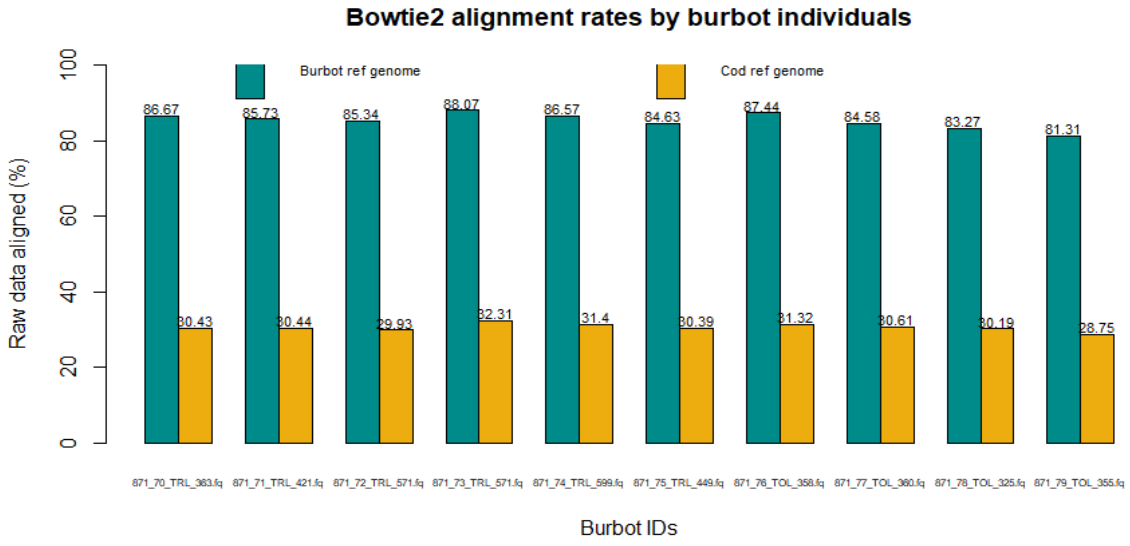

Fig. 2: Bowtie2 alignment success by burbot individuals.

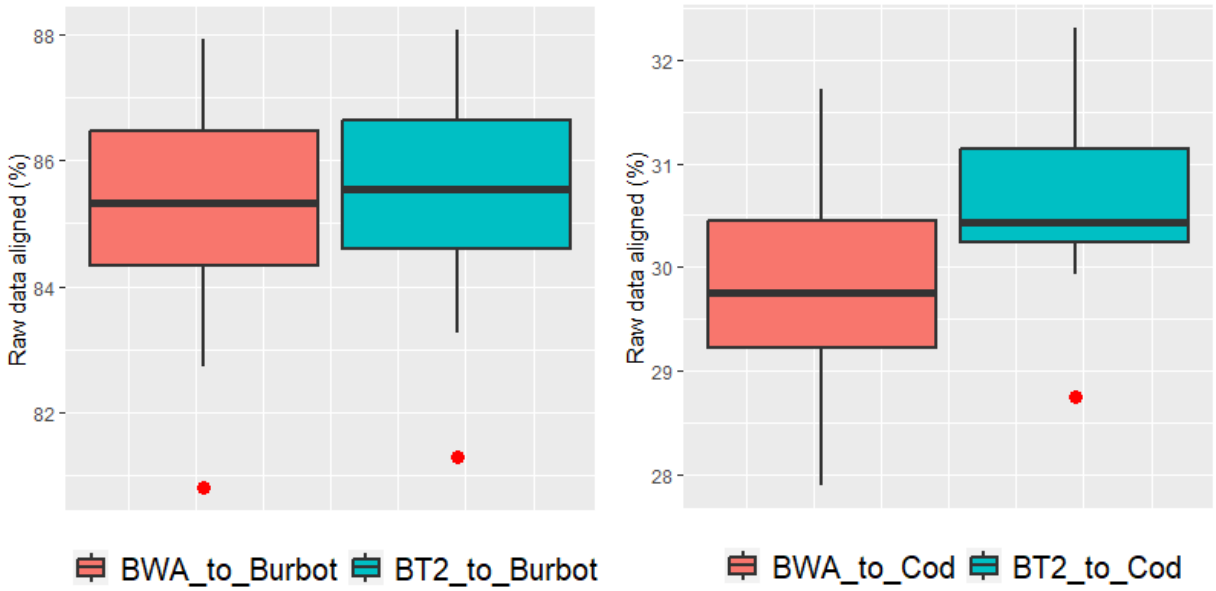

Fig. 3: Boxplots summarizing alignment success by reference genome. Outliers are displayed as red circles.

Discussion

According to the results of this project (Fig. 1), the fragmented burbot genome is a better choice than the cod genome for aligning the raw burbot data, and this observation is consistent across both aligners (Fig 1. and 2). The low alignment rates associated with the cod reference genome are possibly because burbots and cods diverged from their common ancestor about 44.4 million years ago (Han et al., 2020). Both species have thereby developed unique genetic variations that significantly interfere with conserved regions, resulting in lower alignment quality. Despite being fragmented, the fragmented burbot genome still offers better alignment rates in this case.

Looking at alignment success by individual, 871_79 consistently had the lowest alignment rates regardless of the reference genome or aligner used (Fig. 1 and 2). In addition, the individual is an outlier in three of four alignment groups (Fig 3) suggesting that there may have been issues during DNA sequencing for this particular individual. Conversely, individual 871_73 maintains the highest alignment rates across all groups. Overall, alignment rates are generally consistent among aligners and reference genomes, with the exception of the BT2_to_Cod alignment, which appears to be skewed (Fig. 3).

Whilst the burbot reference genome proves to be better for aligning these burbot datasets, it is important to note that choosing the most appropriate reference genome depends on the specific research inquiry. For instance, if a study aims to estimate admixture among burbots, cods, and related species, the preferred reference genome may change based on the group of species being considered. In another scenario, if the aim is to investigate the relationship between burbots and cods, mapping burbots against cod may be appropriate. Through this project, I have gained a better understanding of the impact that reference genomes have in biological studies.

Methods

Burrows-Wheeler aligner (BWA) and Bowtie2 aligner were used to align raw reads of 10 burbot fish to a fragmented burbot reference genome and a high-quality cod reference genome. Data alignment and processing were done on Graham HPC cluster while plots were prepared on RStudio 2022.12.0. This is the Graham directory path to all files: /scratch/eolabisi/MyProject2.

To begin, I unzipped all raw burbot data with a “for” loop. I also made a copy of the bwa script to customize parts of it, such as the host's username and reference genome's file path, for my use.

```
for file in *.gz; do gunzip $file; done
scp run_bwa_queuesub.sh burbot_raw_data/run_bwa_queuesub2.sh
```

In the raw burbot directory, I created a bwa_assembly for the output files that will be produced.

```
mkdir bwa_assem
```

Next, I used a loop to carry out bwa alignment and samtools statistics against the burbot reference genome. The script aligns the raw reads before sorting and indexing the .bam files.

```
for file in *.fq; do sbatch run_bwa_queuesub2.sh $file; done
```

I copied the content of bwa_assem to the designated folder for “bwa alignment to burbot reference genome”: Burbot_Ref_BWA. I emptied bwa_assem and repeated the previous step against the cod genome ensuring that I changed the file path in the script accordingly. Output files were copied to Cod_Ref_BWA.

```
rsync -a bwa_assem/ ../Aligned_data/Burbot_Ref_BWA/    #copy
rsync -a bwa_assem/ ../Aligned_data/Cod_Ref_BWA/
```

I made a copy of the bwa script to customize it for bowtie2 alignments.

```
  scp run_bwa_queuesub2.sh run_bowtie_queuesub.sh
```

The following changed were made to the bwa script for bowtie2 alignments. I added the command line for bowtie2 alignment and changed my assembly folder name throughout.

```
module load bowtie2/2.4.4  ##

bowtie2 -x
../burbot_reference_genome/GCA_900302385.1_ASM90030238v1_genomic
-U $fastq -S bowtie2_assem/$basename.sam --very-sensitive-local
     ##

samtools view -b -S -o bowtie2_assem/$basename.bam
bowtie2_assem/$basename.sam  ##

samtools sort bowtie2_assem/$basename.bam -o
bowtie2_assem/$basename.sorted.bam   ##

samtools index bowtie2_assem/$basename.sorted.bam   ##

if [[ -s bowtie2_assem/$basename.bam ]]   ##
  then
    rm bowtie2_assem/$basename.sam  ##
    echo "removed $basename.sam"

else
  echo "$basename.sam is empty! Something's fishy..."
fi

if [[ -s bowtie2_assem/$basename.bam ]]   ##
  then
    rm bowtie2_assem/$basename.bam  ##
    echo "removed $basename.bam"
```

After preparing the bowtie2 script, I made a bowtie2_assem directory and ran the alignment and samtools statistics using a loop as with bwa.

```
mkdir bowtie2_assem
for file in *.fq; do sbatch run_bowtie_queuesub.sh $file; done
```

I copied bowtie2_assem's content to its designated folder, Cod_Ref_BOWTIE2. Finally, I did the last bowtie2 alignment to burbot reference genome using the above "for" loop command and the correct file path. Output files were copied to Burbot_Ref_BOWTIE2.

```
rsync -a bowtie2_assem/ ../Aligned_data/Cod_Ref_BOWTIE2/
rsync -a bowtie2_assem/ ../Aligned_data/Burbot_Ref_BOWTIE2/
```

Now in the class's module2_assem directory, I copied example_bash_script to my directory and customized it for my use.

```
  scp example_bash_script.sh
~/Genomic_class/MyProject2/Aligned_data
```

The following changes were made to the bash script to obtain both number of raw reads and mapped reads for bwa alignment to burbot genome. Apart from customizing the output file name, I also added a sed command to replace the descriptive phrase on each line with a space character.

```
echo "ind raw assembled" > b_bwa_assembled_per_ind.txt ##

for file in Burbot_Ref_BWA/*sorted.bam  ##
do
indname=`echo $file | sed 's/.*\///g' | sed
's/\.sorted\.bam//g'`
raw=`samtools stats $file | grep "raw total sequences:" | sed
's/SN\t.*:\t//g' | sed 's/\t# excluding supplementary and
secondary reads/ /g'`    ##
echo "$indname $raw $assembled" >> b_bwa_assembled_per_ind.txt
```

Next, I requested an srun session to run the bash script. The following steps were repeated for the other (3) alignment groups ensuring that file paths and names were changed as required.

```
srun --pty --account def-lukens -t 0-00:15 /bin/bash

./example_bash_script.sh
```

All text files were moved to my PC for analysis on R software:

```
scp -r
eolabisi@graham.computecanada.ca:~/Genomic_class/MyProject2/Alig
ned_data/*.txt .
```

######R CODE#######

```
library(tidyverse)
library(reshape2)
library(ggplot2)
```

I created a function that reads a space-delimited text file as a data frame. The function also adds a column containing calculated percentage of raw data aligned.

```
create.table <- function(txt_file) {
  txt_table <- read.table(txt_file, header = T)
  txt_table$percent_aligned <-
(txt_table$assembled/txt_table$raw) *100
  txt_table$percent_aligned <- round(txt_table$percent_aligned,
digits = 2)
  return(txt_table)
}
```

Next, I read in each text file to produce data frames.

```
BWA_to_Bur <- create.table("b_bwa_assembled_per_ind.txt")#bwa
alignment to burbot genome
BWA_to_Cod <- create.table("c_bwa_assembled_per_ind.txt") #bwa
alignment to cod genome
BT2_to_Bur <- create.table("b_bt2_assembled_per_ind.txt")
#bowtie2 alignment to burbot genome
BT2_to_Cod <- create.table("c_bt2_assembled_per_ind.txt")
##bowtie2 alignment to cod genome
```

To put all my percentage values in one place, I created another dataframe merging all the needed columns. Starting with one of the tables, I selected the ID and percent-aligned columns to be kept and added percent-aligned columns from other tables, each with their respective column titles.

```
Merged_table <- BWA_to_Bur %>%
  select(ind, BWA_to_Burbot=percent_aligned) %>%
  add_column(BWA_to_Cod= BWA_to_Cod$percent_aligned)  %>%
  add_column(BT2_to_Burbot = BT2_to_Bur$percent_aligned) %>%
  add_column(BT2_to_Cod = BT2_to_Cod$percent_aligned)

#view
Merged_table

#remove first copies
rm(BWA_to_Bur, BWA_to_Cod, BT2_to_Bur, BT2_to_Cod)
```

I created a function to transform the merged data frame and create a bar plot of the values. The transformation allows me to display bars of two different datasets on one plot. Various graphical modifications were also included.

```
create_bar_plot <- function(column1, column2, column_for_x_axis,
colour_vector, plot_title){
  joined_data <- t(cbind(column1, column2))
  y <- as.matrix(joined_data)
  x <- barplot(joined_data, ylim=c(0,100), beside=T,
               names.arg=column_for_x_axis, cex.names=0.5,
col=colour_vector,
               xlab="Burbot IDs", ylab="Raw data aligned (%)",
               main=plot_title)
  legend(x="top", legend = c("Burbot ref genome", "Cod ref
genome"),
         fill = colour_vector,  cex = 0.7,
         horiz=T, bty="n", , inset=-0.3)
  text(x, y+2, labels=y, cex=0.7)
}
```

Bar plot 1 compares Burrows-Wheeler Aligner (BWA) results.

```
create_bar_plot(Merged_table$BWA_to_Bur,
Merged_table$BWA_to_Cod, Merged_table$ind, c("brown2",
"deeppink3"), plot_title = "Burrows-Wheeler Aligner (BWA)
alignment rates by burbot individuals")
```

Bar plot 2 compares Bowtie2 alignment results.

```
create_bar_plot(Merged_table$BT2_to_Bur,
Merged_table$BT2_to_Cod, Merged_table$ind, c("cyan4",
"darkgoldenrod2"), plot_title = "Bowtie2 alignment rates by
burbot individuals")
```

Finally, I created another function to make a box plot for each column of a data frame. This allows me to compare summaries of alignment percentages between columns. The melt() function modifies the data frame for use with ggplot.

```
Box_plot <- function(df,column_names_vector){
  modified <- melt(df, id.vars=NULL, measure.vars =
column_names_vector)
  colnames(modified) <- c("Key" ,"Raw data aligned (%)")

  #creating a plot
  ggplot(modified, aes(x=NULL, y=`Raw data aligned (%)`,
fill=Key)) +
```

```
    geom_boxplot(linewidth=1, outlier.size = 3,
                 outlier.colour="red") +

    theme(axis.title.x=element_blank(),
          axis.text.x=element_blank(),
          axis.ticks.x=element_blank(),
          legend.position = "bottom",
legend.title=element_blank(),
          legend.text = element_text(size=15))
}
```

The first graph shows boxplots for burbot reference genome while the second shows boxplots for cod reference genome. Both graphs will be compared.

```
Box_plot(Merged_table, c('BWA_to_Burbot','BT2_to_Burbot'))
Box_plot(Merged_table, c('BWA_to_Cod','BT2_to_Cod'))
```

References

Gopalakrishnan, S., Samaniego Castruita, J. A., Sinding, M.-H. S., Kuderna, L. F. K., Räikkönen, J., Petersen, B., Sicheritz-Ponten, T., Larson, G., Orlando, L., Marques-Bonet, T., Hansen, A. J., Dalén, L., & Gilbert, M. T. P. (2017). The wolf reference genome sequence (Canis lupus lupus) and its implications for Canis spp. population genomics. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3883-3

Han, Z., Liu, M., Liu, Q., Zhai, H., xiao, shijun, & Gao, T. (2020). *Chromosome-level genome assembly of burbot (Lota lota) provides insights into the evolutionary adaptations in freshwater*. https://doi.org/10.22541/au.160218269.98008776/v1

Hooley-underwood Z, Mandeville EG, Gerrity P, Deromedi J, Johnson K (2018) Combining Genetic , Isotopic, and Field Data to Better Describe the Influence of Dams and Diversions on Burbot Movement in the Wind River Drainage, Wyoming. pp. 606–620.

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, *30*(23), 5966–5993. https://doi.org/10.1111/mec.16077