Project 2: Alignment Analysis

Name: Syed Shahzaib Ali

Date: March 1st, 2021

*Course: BINF * 6110*

Introduction

Sequencing DNA is an important technique because it allows us to understand the various roles genes play in the physiological and biochemical functions of an organism. Over-time, many sequencing technologies have been developed; some that you might be familiar with are X-crystallography (1st generation), Sanger Sequencing (2nd generation), 454 machine by Roche (2nd generation), HT-NGS (3rd generation), Minion (3rd generation) etc. (Heather and Chain. 2016). One of the most important techniques is whole-genome sequencing that is conducted by high throughput techniques such as Illumina, but even though these technologies exist, it is not without errors (Pareek *et al*. 2011). The genomes they will produce will have a certain amount of coverage and if the technology is good and operated successfully, the genome will be complete, otherwise it will appear fragmented (Pareek *et al*. 2011).

There are many purposes of aligning sequence against a reference genome, some of them are: building phylogenetic trees to understand evolutionary connections and history, understand protein structures and function, data validation etc. (Kemena and Notredame. 2009). There are many challenges that come with the use of reference genomes. For example, if you use a single reference genome than there is probability of underestimating the variation that comes in the samples that represent a population (Gopalakrishnan et al. 2017). The reason why the under-estimation takes place is because of a concept called reference bias where the samples seem more and more similar to the reference genome (Gopalakrishnan et al. 2017). Same species fragmented genome can be helpful in understanding the variation within a population however the mapping that takes place might not be correct perhaps because they are mapped to the wrong position or remain unmapped (Lau, 2017). This is turn is a problem because it leads to a false analysis or, false negative or positive variant calls (Lau, 2017). Another advantage of using the same species genome as a reference though fragmented are that it can be useful in assisting the completion of another genome. Moving forward, the benefit of distantly related genomes is that they can allow for a better understanding of the evolutionary relationships and to what degree of conservation is seen between species (Galla et al. 2018). The problem with distant genome is that it doesn't give you an idea of how much variation there is within a population (Gopalakrishnan et al. 2017). More benefits of using distant genome is to understand the genetic structure of a specie that does not have a complete reference genome (Gopalakrishnan et al. 2017). Understanding the genetic structure and filling the gaps can be useful in characterizing the various biological roles the organism plays.

In conclusion, depending on the type of reference genome you have, comes with its unique challenges and mostly depends on the type of analysis you want to do. However, to make sure that there is very little error in your analysis the best approach would be to use a whole genome or a a genome with continuity regardless if it is the same species or a distantly related species, as it will induce less bias.

Discussion

If the goal of the project was to understand the evolutionary relationships than the preferred genome would be the cod genome. There are considerable advantages of using the cod genome of which the first one is that it is a complete genome. If the genome is complete then you can reach a proper conclusion to what degree, the burbot fish species is closest to the cod. The

other advantage is that the reference bias would be little to none meaning that your results would be reliable to go forward with more analysis on the sequences. On the other hand, if you were to understand how much variation happens in the species or to understand the population genetic structure than you should go ahead and use the burbot reference genome as you are trying to understand the variation within a population. If you are trying to understand the impact of environmental changes then the fragmented genome such as the burbot genome can be very useful.

From the data observed or results that came from the use of software for alignment specifically bwa and bowtie2, bowtie should be used as it is slightly better than bwa (if you specified the optimal arguments) for alignment. Reason being that bowtie did equal when the alignment required the burbot reference genome and did slightly better when the alignment required the cod reference genome in regard to bwa alignment. This tells us that bowtie software can be an important play in distant reference genome alignment. The conclusion of the results is supported by figure 1 which is a violin plot. If you compare the shapes of "bowtie_cod" with "bwa_cod", because the base is a bit wider and this can also be seen in table 1 the alignment rate is higher. However, when the mean was taken there is only 0.76% difference between these two. This suggests that even though bowtie is slightly better it doesn't cause a considerable amount of difference. When it comes the alignment towards cod the results show 30% alignment rate which is really low, and this suggests that the burbot reads do not really map well to the cod genome.

In the future, it would be better to compare alignments if both cod and burbot genomes were complete or fragmented, as comparing apples to oranges is a bit difficult and adds complexity to analysing how truly effective the bwa software in comparison to bowtie2.

Methods (see all commands under supplementary code)

The analysis of the alignment was conducted in R but itself the generation of data was done in the Unix system by connecting to Compute Canada graham. Data of raw reads and mapped reads was obtained through Sam tools version 1.10 by using the stat package on (.bam) aligned files. The bam files were created through queueing system on graham using a shell script. The code can be seen in supplementary code. The directories of the files are mentioned below:

/home/sali12/new.percentages/bowtie_assem_cody

/home/sali12/new.percentages/bwa_assem_cody

/home/sali12/new.percentages/bowtie_assem_burbot

home/sali12/new.percentages/bwa_assem_burbot

The first step was to enter the "scratch/emandevi/genomic_methods/Project2" directory and copy the required files into my directory. Once that was done directories where the bam files would be saved were made. The files are mentioned above. Then the provided shell script was edited so that it will conduct the bwa and bowtie2 alignments. The 4 shell scripts have been provided in

the supplementary code in the R-mark down. The fourth step is to change the bash script so that it will conduct the BWA and bowtie2 alignments. The command that will conduct the alignment is mentioned below:

```
bowtie2 –very-sensitive-local -x
cod_reference_genome/GCF_902167405.1_gadMor3.0_genomic -U $fastq -S
bowtie_assem_cod/$basename.
bwa mem -t 16 cod_reference_genome/GCF_902167405.1_gadMor3.0_genomic.fna -U $fastq -
S bowtie_assem_cod/$basename.
```

The fifth step is to run the script through the que system. This can be done with the command below.

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

The sixth step taken was to use the package Sam tools to acquire the raw reads and the assembled reads and save into the .txt files.

```
for file in *sorted.bam; do Sam tools stats $file | grep "raw total sequences:" | sed 's/SN\t.*:\t//g';
done > actual_bwa_cody_raw.txt
```

```
for file in *sorted.bam; do Sam tools stats $file | grep "reads mapped:" | sed 's/SN\t.*:\t//g'; done
> actual_bwa_cody_raw.txt
```

Moving forward using scp command the .txt files with raw reads, assembled reads and .csv files with percentages were moved to the local directory, where R was used to create a data frame, and violin plot was made. The scp command is mentioned below but the R code is mentioned in R-mark down file.

```
scp sali12@graham.computecanada.ca:
/scratch/sali12/nearline/burbot_raw_data/bow_assem_cody/cod_bow_reads_mapped.txt
```

*****The code mentioned above was used over again but changed slightly depending on the files*****

References

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of applied genetics*, *52*(4), 413–435. https://doi.org/10.1007/s13353-011-0057-x

Kemena, C., & Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics (Oxford, England)*, *25*(19), 2455–2465. https://doi.org/10.1093/bioinformatics/btp452

Gopalakrishnan, S., Samaniego Castruita, J.A., Sinding, MH.S. *et al.* The wolf reference genome sequence (*Canis lupus lupus*) and its implications for *Canis spp.* population genomics. *BMC Genomics* **18,** 495 (2017). https://doi.org/10.1186/s12864-017-3883-3

Galla, S. J., Forsdick, N. J., Brown, L., Hoeppner, M., Knapp, M., Maloney, R. F., Moraga, R., Santure, A. W., & Steeves, T. E. (2018). Reference Genomes from Distantly Related Species Can Be Used for Discovery of Single Nucleotide Polymorphisms to Inform Conservation Management. *Genes,* 10(1), 9. https://doi.org/10.3390/genes10010009

Lau. (2017, May 22nd). References bias: Challenges and solutions. Retrieved from https://www.sevenbridges.com/reference-bias-challenges-and-solutions/

Figures

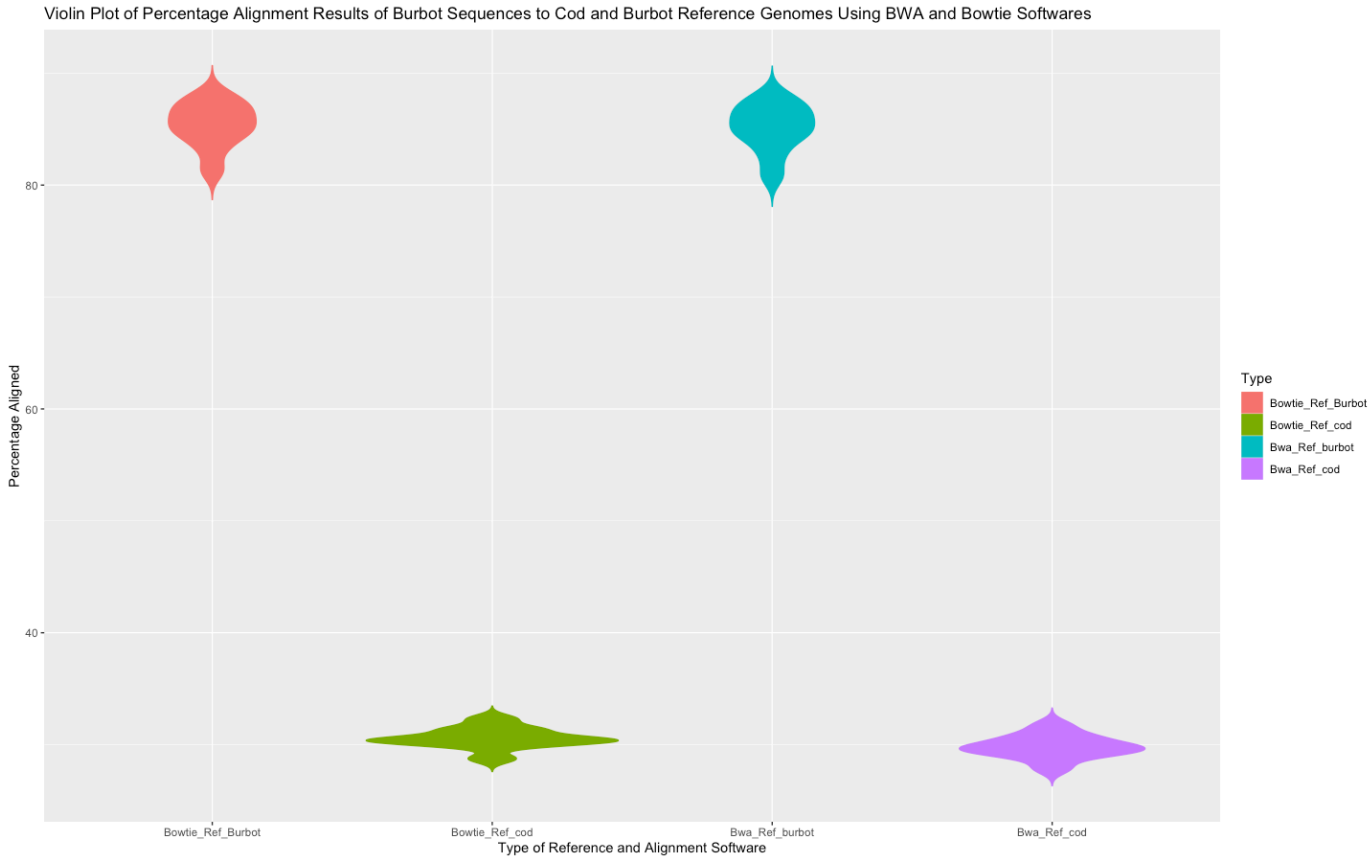

Figure 1: Show's a violin plot of the alignment rates determined for using burbot and cod reference genomes using two alignment softwares bwa and bowtie2 in graham. The plot shows that bowtie performs better when cod and burbot reference genome are used.

Table 1: Data-Frame of Alignment Results

Column1	Raw_Read	Mapped_Read	Type	Percentage.Alignment
1	1494524	1292764	Bwa_Ref_burbot	86.50004951
2	820654	702096	Bwa_Ref_burbot	85.5532295
3	969375	824748	Bwa_Ref_burbot	85.08038685
4	1468330	1290920	Bwa_Ref_burbot	87.91756621
5	1461492	1262243	Bwa_Ref_burbot	86.36674029
6	708367	597061	Bwa_Ref_burbot	84.2869586
7	734576	641224	Bwa_Ref_burbot	87.29171658
8	1116625	943105	Bwa_Ref_burbot	84.46031568
9	914486	756624	Bwa_Ref_burbot	82.73762529
10	984048	795237	Bwa_Ref_burbot	80.8128262
11	1494524	445165	Bwa_Ref_cod	29.78640691
12	820654	243821	Bwa_Ref_cod	29.7105723
13	969375	282112	Bwa_Ref_cod	29.10246293
14	1468330	465748	Bwa_Ref_cod	31.71957258
15	1461492	449056	Bwa_Ref_cod	30.72586097
16	708367	209586	Bwa_Ref_cod	29.5872055
17	734576	225019	Bwa_Ref_cod	30.63250093
18	1116625	333832	Bwa_Ref_cod	29.89651853
19	914486	265862	Bwa_Ref_cod	29.0722876
20	984048	274423	Bwa_Ref_cod	27.88715591
21	1494524	454723	Bowtie_Ref_cod	30.42594164
22	820654	249780	Bowtie_Ref_cod	30.43670049
23	969375	290107	Bowtie_Ref_cod	29.92722115
24	1468330	474380	Bowtie_Ref_cod	32.30745132
25	1461492	458933	Bowtie_Ref_cod	31.40167719
26	708367	215304	Bowtie_Ref_cod	30.39441419
27	734576	230085	Bowtie_Ref_cod	31.32215047
28	1116625	341842	Bowtie_Ref_cod	30.61385873
29	914486	276058	Bowtie_Ref_cod	30.18723086
30	984048	282897	Bowtie_Ref_cod	28.74829277
31	1494524	1295237	Bowtie_Ref_Burbot	86.66552026
32	820654	703530	Bowtie_Ref_Burbot	85.72796818
33	969375	827290	Bowtie_Ref_Burbot	85.34261767
34	1468330	1293125	Bowtie_Ref_Burbot	88.06773682
35	1461492	1265273	Bowtie_Ref_Burbot	86.57406267
36	708367	599516	Bowtie_Ref_Burbot	84.63353036
37	734576	642343	Bowtie_Ref_Burbot	87.44404936
38	1116625	944426	Bowtie_Ref_Burbot	84.57861861
39	914486	761529	Bowtie_Ref_Burbot	83.27399217
40	984048	800088	Bowtie_Ref_Burbot	81.30578996

Table 1 shows a data-frame that contains the information of percentage alignments of the burbot sequences to reference genomes of Cod and Burbot fish using two software's BWA and Bowtie2.

Supplementary Code

```
---
title: 'Project 2: Alignment Analysis'
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

The first step was to get enter the directory to get the required files in the graham.

cd /scratch/emandevi/genomic_methods_w2021/Project2

Then the files had to be copied to my directory on graham.

cp -R burbot_raw_data /scratch/sali12/nearline/

cp -R burbot_reference_genome /scratch/sali12/nearline/burbot_raw_data

cp -R cod_reference_genome /scratch/sali12/nearline/burbot_raw_data

In order to get the bam files, through the shell script some directories need to be made because the files will appear in the direcotry specified in the shell script.

mkdir bow_assem_cod

mkdir bow_assem_burbot

mkdir bwa_assem_cod

mkdir bwa_assem_burbot

Then the shell script code used below is mentioned below. In order to get the files for each alignment software used and reference the directory name was changed and the command for aignment as well. The rest was kept the same.

#!/bin/sh

This script uses bwa to map reads (.fastq) to reference genome
usage (for testing with just one individual):
sbatch run_bwa_queuesub.sh $fastq

#SBATCH --account=def-emandevi
#SBATCH --time=0-00:15:00 ## days-hours:minutes:seconds
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=4 # number of threads
#SBATCH --mem=16000 # requested memory (in MB)
#SBATCH --mail-type=END

module load bwa/0.7.17
module load samtools/1.9
```

```
fastq=$1
basename=`echo $fastq | sed 's/\.f.*\.gz.*//'`

echo "Starting alignment of $fastq to reference genome"
bwa mem -t 16 burbot_reference_genome/
GCA_900302385.1_ASM90030238v1_genomic.fna $fastq >  bwa_assem_burbot/
$basename.sam
echo "Converting sam to bam for $basename"
samtools view -b -S -o bwa_assem_burbot/$basename.bam bwa_assem_burbot/
$basename.sam

echo "Sorting and indexing bam files for $basename"
samtools sort bwa_assem_burbot/$basename.bam -o bwa_assem_burbot/
$basename.sorted.bam
samtools index bwa_assem_burbot/$basename.sorted.bam


###### The following code represents the script for bowtie2 alignment
software that through the sbatch will be used through queueing system. The
script is designed to run for 15 minutes.

#!/bin/sh

## This script uses bwa to map reads (.fastq) to reference genome
## usage (for testing with just one individual):
## sbatch run_bwa_queuesub.sh $fastq

#SBATCH --account=def-emandevi
#SBATCH --time=0-00:15:00 ## days-hours:minutes:seconds
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=4 # number of threads
#SBATCH --mem=16000 # requested memory (in MB)
#SBATCH --mail-type=END

module load bowtie2
module load samtools/1.9

fastq=$1
basename=`echo $fastq | sed 's/\.f.*\.gz.*//'`

echo "Starting alignment of $fastq to reference genome"
bowtie2 --very-sensitive-local -x cod_reference_genome/
GCF_902167405.1_gadMor3.0_genomic -U  $fastq -S bowtie_assem_cod/
$basename.sam
echo "Converting sam to bam for $basename"
samtools view -b -S -o bowtie_assem_cod/$basename.bam bowtie_assem_cod/
$basename.sam

echo "Sorting and indexing bam files for $basename"
samtools sort bowtie_assem_cod/$basename.bam -o bowtie_assem_cod/
$basename.sorted.bam
samtools index bowtie_assem_cod/$basename.sorted.bam
```

```
# The script is for running bowtie2 using the burbot genome

#!/bin/sh

## This script uses bwa to map reads (.fastq) to reference genome
## usage (for testing with just one individual):
## sbatch run_bwa_queuesub.sh $fastq

#SBATCH --account=def-emandevi
#SBATCH --time=0-00:15:00 ## days-hours:minutes:seconds
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=4 # number of threads
#SBATCH --mem=16000 # requested memory (in MB)
#SBATCH --mail-type=END

module load bowtie2
module load samtools/1.9

fastq=$1
basename=`echo $fastq | sed 's/\.f.*\.gz.*//'`

echo "Starting alignment of $fastq to reference genome"
bowtie2 --very-sensitive-local -x  burbot_reference_genome/
GCA_900302385.1_ASM90030238v1_genomic -U $fastq -S bowtie_assem_burbot/
$basename.sam
echo "Converting sam to bam for $basename"
samtools view -b -S -o bowtie_assem_burbot/$basename.bam
bowtie_assem_burbot/$basename.sam

echo "Sorting and indexing bam files for $basename"
samtools sort bowtie_assem_burbot/$basename.bam -o bowtie_assem_burbot/
$basename.sorted.bam
samtools index bowtie_assem_burbot/$basename.sorted.bam


###### The script for running BWA alignment software using the cod genome
#########

#!/bin/sh

## This script uses bwa to map reads (.fastq) to reference genome
## usage (for testing with just one individual):
## sbatch run_bwa_queuesub.sh $fastq

#SBATCH --account=def-emandevi
#SBATCH --time=0-00:15:00 ## days-hours:minutes:seconds
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=4 # number of threads
#SBATCH --mem=16000 # requested memory (in MB)
#SBATCH --mail-type=END

module load bwa/0.7.17
module load samtools/1.9
```

```
fastq=$1
basename=`echo $fastq | sed 's/\.f.*\.gz.*//'`

echo "Starting alignment of $fastq to reference genome"
bwa mem -t 16 cod_reference_genome/GCF_902167405.1_gadMor3.0_genomic.fna
$fastq >  bwa_assem_cod/$basename.sam
echo "Converting sam to bam for $basename"
samtools view -b -S -o bwa_assem_cod/$basename.bam bwa_assem_cod/
$basename.sam

echo "Sorting and indexing bam files for $basename"
samtools sort bwa_assem_cod/$basename.bam -o bwa_assem_cod/
$basename.sorted.bam
samtools index bwa_assem_cod/$basename.sorted.bam

# The command to use bowtie2 and bwa is mentioned below and that was used
in the script. The --very-sensitive local argument allows for better
alignment and gap penalty. The -x referes to the file that will be used for
reference. -U is for the files to be aligned and -S is for the output of
the file. The format indicated is .sam which will be converted .bam files.

bowtie2 --very-sensitive-local -x cod_reference_genome/
GCF_902167405.1_gadMor3.0_genomic -U $fastq -S bowtie_assem_cod/$basename.

bwa mem -t 16 cod_reference_genome/GCF_902167405.1_gadMor3.0_genomic.fna -U
$fastq -S bowtie_assem_cod/$basename.

# Once the files were obtained then samtools stats was used to get the
information of raw reads and assembled reads. The raw reads give the total
length of the reads and the assembled reads tell you how much of the burbot
sequneces aligned or matched.

for file in *sorted.bam; do samtools stats $file | grep "raw total
sequences:" | sed 's/SN\t.*:\t//g'; done > actual_bwa_cody_raw.txt

for file in *sorted.bam; do samtools stats $file | grep "reads mapped:" |
sed 's/SN\t.*:\t//g'; done > actual_bwa_cody_raw.txt

for file in *sorted.bam; do samtools stats $file | grep "raw total
sequences:" | sed 's/SN\t.*:\t//g'; done > actual_bwa_cody_raw.txt

for file in *sorted.bam; do samtools stats $file | grep "reads mapped:" |
sed 's/SN\t.*:\t//g'; done > actual_bwa_cody_raw.txt

for file in *sorted.bam; do samtools stats $file | grep "raw total
sequences:" | sed 's/SN\t.*:\t//g'; done > actual_bwa_cody_raw.txt

for file in *sorted.bam; do samtools stats $file | grep "reads mapped:" |
sed 's/SN\t.*:\t//g'; done > actual_bwa_cody_raw.txt

for file in *sorted.bam; do samtools stats $file | grep "raw total
sequences:" | sed 's/SN\t.*:\t//g'; done > actual_bwa_cody_raw.txt
```

```
for file in *sorted.bam; do samtools stats $file | grep "reads mapped:" |
sed 's/SN\t.*:\t//g'; done > actual_bwa_cody_raw.txt


##### The files can be seen here. The directores have been made sure that
they are accesible by everyone other than me and for files to be viewed.
#######

/home/sali12/new.percentages/bowtie_assem_cody

/home/sali12/new.percentages/bwa_assem_cody

/home/sali12/new.percentages/bowtie_assem_burbot

home/sali12/new.percentages/bwa_assem_burbot

# The percentage calculated were done in R in the graham and were saved as
csv files. The Module load r command loads the r but does not initiate it
until r is written in another line. The files were read using the
read.table function because read.csv or read.delim causes the first number
to be indicated as the column. Finally simple analysis was done where
assembled reads were divided by raw read numbers and multiplied with a 100.
The file dataa was saved as csv and transported back to the local computer.
The following is a demonstration with one of the directories which only had
the bam alignment files that resulted from bowtie2 using the burbot
fragmented genome as a reference.

Module load r

R

Actual = read.table(“ bow_bur_raw.txt")

Assem_read = read.table(“bow_percentages_bur.csv")

Final = (Assem_read / Actual)*100

Write.csv(Final, Percentage.Burbot.Bowtie)


##### copying the files of raw reads, percentages and assembled reads to
the local computer. The path is now broken and inaccessible.

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/bow_bur_raw.txt

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/bow_percentages_bur.csv

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/bur_bow_reads_mapped.txt

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/bwa_cod_raw.txt
```

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/bow_cod_ref.csv

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/cod_bow_reads_mapped.txt

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/actual_bwa_cody_raw.csv

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/bwa_cod_percentages.csv

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/actual_cody_bwa_reads_mapped.txt

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/bwa_bur_raw.txt

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/bur_bwa_reads_mapped.txt

scp sali12@graham.computecanada.ca: /scratch/sali12/nearline/
burbot_raw_data/bow_assem_cody/bur_ref_bur.csv

The ggplot2 library is required to make the violin plot.

```{r cars}
library(ggplot2)
library(tidyverse)
library(dplyr)
```

Here the files that were produced in graham were uploaded to make a
dataframe. It's a bit to rename the columns because 3/4 had the same name
thus 4 individual dataframes were created. This particular dataframe has
the information of raw reads, assemebled reads and alignment rate acquired
using bowtie2 software using the burbot reference genome.

```{r pressure, include=TRUE}
# This creates the data frame for burbot sequences using the burbot
reference genome
getwd()

Bowtie.burbot.raw = read.table("bow_bur_raw.txt")

Bowtie.burbot = read.csv("bow_percentages_bur.csv")

Bowtie.burbot.mapped = read.table("bur_bow_reads_mapped.txt")

Bowtie.burbot.raw = Bowtie.burbot.raw %>% rename(Raw_Read = V1 )

Bowtie.burbot.mapped = Bowtie.burbot.mapped %>% rename(Mapped_Read = V1 )
```

```
Bowtie.burbot$X <- "Bowtie_Ref_Burbot"

Bowtie.burbot = Bowtie.burbot %>% rename(Type = X, Percentage.Alignment =
V1 )

Bowtie.burbot.data <- cbind(Bowtie.burbot.raw,Bowtie.burbot.mapped,
Bowtie.burbot )

View(Bowtie.burbot.data)
```

This particular dataframe has the information of raw reads, assemebled reads and alignment rate acquired using bowtie2 software using the cod reference genome.

```{r pressure, echo=FALSE}
# This creates the data frame for Bowtie alignment of burbot sequences
using the cod reference genome

Bowtie.cod.raw = read.table("bwa_cod_raw.txt")

Bowtie.cod  = read.csv("bow_cod_ref.csv")

Bowtie.cod.mapped = read.table("cod_bow_reads_mapped.txt")

Bowtie.cod.raw = Bowtie.cod.raw %>% rename(Raw_Read = V1 )

Bowtie.cod.mapped = Bowtie.cod.mapped %>% rename(Mapped_Read = V1 )

Bowtie.cod$X <- "Bowtie_Ref_cod"

Bowtie.cod = Bowtie.cod %>% rename(Type = X, Percentage.Alignment = V1 )

Bowtie.cod.data <- cbind(Bowtie.cod.raw,Bowtie.cod.mapped, Bowtie.cod )

View(Bowtie.cod.data)

```

This particular dataframe has the information of raw reads, assemebled reads and alignment rate acquired using b software using the cod reference genome.

```{r pressure, echo=FALSE}
# This creates the data frame for Bowtie alignment of burbot sequences
using the cod reference genome

Bwa.cod.raw = read.table("actual_bwa_cody_raw.csv")

Bwa.cod  = read.csv("bwa_cod_percentages.csv")
```

```
Bwa.cod.mapped = read.table("actual_cody_bwa_reads_mapped.txt")

Bwa.cod.raw = Bwa.cod.raw %>% rename(Raw_Read = V1 )

Bwa.cod.mapped = Bwa.cod.mapped %>% rename(Mapped_Read = V1 )

Bwa.cod$X <- "Bwa_Ref_cod"

Bwa.cod = Bwa.cod %>% rename(Type = X, Percentage.Alignment = V1 )

Bwa.cod.data <- cbind(Bwa.cod.raw,Bwa.cod.mapped, Bwa.cod )

View(Bwa.cod.data)
```

This particular dataframe has the information of raw reads, assemebled reads and alignment rate acquired using bwa software using the burbot reference genome.

```{r pressure, echo=FALSE}
# This creates the data frame for Bowtie alignment of burbot sequences using the cod reference genome

Bwa.burbot.raw = read.table("bwa_bur_raw.txt")

Bwa.burbot = read.csv("bur_ref_bur.csv")

Bwa.burbot.mapped = read.table("bur_bwa_reads_mapped.txt")

Bwa.burbot.raw = Bwa.burbot.raw %>% rename(Raw_Read = V1 )

Bwa.burbot.mapped = Bwa.burbot.mapped %>% rename(Mapped_Read = V1 )

Bwa.burbot$X <- "Bwa_Ref_burbot"

Bwa.burbot = Bwa.burbot %>% rename(Type = X, Percentage.Alignment = V1 )

Bwa.burbot.data <- cbind(Bwa.burbot.raw,Bwa.burbot.mapped, Bwa.burbot )

View(Bwa.burbot.data)

Final_data = do.call("rbind", list(Bwa.burbot.data,
Bwa.cod.data,Bowtie.cod.data, Bowtie.burbot.data  ))

View(Final_data)
```

The below cod allows for the creation of violin plot to visualize the alignment rate. The data frame that was generated was written as a .csv file and has been posted in the write up.

```{r pressure, echo=FALSE}
ggplot(Final_data, aes(x = Type, y = Percentage.Alignment, color = Type,
fill = Type)) + geom_violin(trim = FALSE) + labs(y = "Percentage Aligned",
x = "Type of Reference and Alignment Software", title =  "Violin Plot of
Percentage Alignment Results of Burbot Sequences to Cod and Burbot
Reference Genomes Using BWA and Bowtie Softwares")

write.csv(Final_data, "Final_data_alignment.csv")
```