Project 2: Reference-guided assembly of short reads from burbot (*Lota lota*): impact of reference genome and alignment tool selection

J Moggridge

23/02/2021

Introduction

Reference-guided assembly involves mapping new sequencing reads to an existing genome; this is an important processing step prior to genotyping for analysis of genome-wide associations or phylogenetic inference, for example. Reference-guided assembly is simpler and faster than *de-novo* assembly but choosing a suitable reference genome to align reads to is crucial. Depending on the species of interest, the reference genome may be of low quality (*i.e.* many gaps) or unavailable entirely, as there are very few reference genomes compared to the enormous number of known species.

If there is no reference for the species of interest, it is common practice to use a high quality reference genome of a closely-related species instead. However, if there is not a high degree of homology between the subject and reference genomes, many reads will fail to align to the assembly regardless of quality. If the reference genome used is of low quality, many reads may fail to be aligned due to gaps in the assembly, even if the individual and the reference are closely-related. In either of these cases, incorrectly mapped or unmapped reads leads to errors in genotyping that can bias later analyses (Bohling 2020). Even when considering the genomes of closely-related species, the choice of reference can subtly impact findings, for example the number of variants found or heterozygosity estimates (Gopalakrishnan et al. 2017).

In this work, the trade-off in these concerns (relatedness and assembly quality) for read mapping is examined through an analysis of data from a genotyping by sequencing study of burbot (*Lota lota*). A burbot reference genome exists but is of low quality, while a high quality reference is available for the Atlantic cod (*Gadus morhua*; Tørresen et al. 2017). Illumina single-end reads from 10 individuals were mapped to both cod and burbot reference genomes using both `bwa` and `bowtie2` (very sensitive option); reads mapped were recorded for comparison of reference genomes and aligners.

Methods

All the scripts and output files for this analysis can be found on graham in `/scratch/jmoggrid/Project2/`. I performed the alignments using the `bash` commands below. In practice, each aligner/reference combination was executed separately with smaller time allocation (scripts named: `run_<aligner>_<reference>.sh`). Lines that are too long (with \) have been wrapped for readability.

```
#!/bin/sh
#SBATCH --account=def-nricker
#SBATCH --time=0-03:00:00
#SBATCH --nodes=1
```

```
#SBATCH --ntasks-per-node=4
#SBATCH --mem=16000
#SBATCH --mail-type=END
module load bwa/0.7.17
module load bowtie2

## align each fq file to COD ref using BWA
for file in *.fq.gz; do
  basename=`echo $file | sed 's/\.fq\.gz//'`;
  bwa mem -t 16 cod_reference_genome/GCF_902167405.1_gadMor3.0_genomic.fna $file \
    > $basename.bwa.cod.sam
done
## align each fq file to BURBOT ref using BWA
for file in *.fq.gz; do
  basename=`echo $file | sed 's/\.fq\.gz//'`;
  bwa mem -t 16 burbot_reference_genome/GCA_900302385.1_ASM90030238v1_genomic.fna $file \
    > $basename.bwa.burbot.sam
done
## align each fq file to COD ref using BOWTIE2
for file in *.fq.gz; do
  basename=`echo $file | sed 's/\.fq\.gz//'`;
  bowtie2 -x cod_reference_genome/GCF_902167405.1_gadMor3.0_genomic -U $file \
    -S $basename.bowtie.cod.sam --very-sensitive-local;
done
## align each fq file to BURBOT ref using BOWTIE2
for file in *.fq.gz; do
  basename=`echo $file | sed 's/\.fq\.gz//'`;
  bowtie2 -x burbot_reference_genome/GCA_900302385.1_ASM90030238v1_genomic -U $file \
    -S $basename.bowtie.burbot.sam --very-sensitive-local;
done
```

Then, I collected the percentage of reads mapped for each individual in each of the four treatments using `samtools stats` and `flagstats` (script: `run_collect_stats.sh`).

```
#!/bin/sh

## Create a csv file with header line
touch alignments_data.csv
echo "Sample,Aligner,Reference,Total,Mapped,Unmapped,mQ0, \
      PercentMapped,Secondary,Supplementary" > alignments_data.csv

## Wrangle data of interest from samtools stats and flagstats
module load samtools
for file in *.sam; do
  echo $file;
  sample=`echo $file | sed 's/\..*//'`;
  if echo $file | grep -q 'bwa'; then
    aligner='bwa'
  elif echo $file | grep -q 'bowtie'; then
    aligner='bowtie2'
  fi;
  if echo $file | grep -q 'cod'; then
    reference='cod'
  elif echo $file | grep -q 'burbot'; then
```

```
    reference='burbot'
  fi;
  touch temp;
  samtools stats $file > temp;
  total=`grep 'raw total sequences' temp | cut -f3`;
  mapped=`grep 'reads mapped:' temp | cut -f3`;
  unmapped=`grep 'reads unmapped:' temp | cut -f3`;
  mQ0=`grep 'reads MQ0:' temp | cut -f3`;
  samtools flagstats $file > temp;
  percentMapped=`grep % temp | sed -r 's/.*\(|% : .*\)//g'`;
  secondary=`grep 'secondary' | sed 's/ + 0 secondary//'`
  supplementary= `grep 'supplementary' | sed 's/ + 0 supplementary//'`
  echo "$sample,$aligner,$reference,$total,$mapped,$unmapped, \
    $mQ0,$percentMapped,$secondary,$supplementary" >> \
    alignments_data.csv;
done
rm temp
```

R code for statistical analysis:

```
library(tidyverse)
library(rstatix)
# compute group means and sd of % mapped
burbot <- read_csv('alignments_data.csv') %>%
  group_by(Aligner, Reference) %>%
  mutate(`% mapped mean` = mean(PercentMapped),
         `% mapped sd` = sd(PercentMapped)) %>%
  ungroup()
# repeated measures ANOVA
burbot.anova <- rstatix::anova_test(
  data = burbot, dv = PercentMapped, wid = Sample, within = c(Reference, Aligner))
```

Results

Statistical analysis of the number of aligned reads was conducted in the R language. I computed group means and standard deviations for the percentage of reads mapped by each treatment (table 1). Because of the within-subjects design of this two-factor experiment, I performed a repeated measures ANOVA using the **rstatix** package to discern whether the choice of alignment tool or reference genome has an effect on the proportion of reads mapped (table 2).

Table 1: Summary table for percentage of reads mapped in fourconditions: with cod or burbot reference genome, with bwa or bowtie2 alignment.

Aligner	Reference	% mapped (sd)
bowtie2	burbot	85.4 (2)
bowtie2	cod	30.6 (1)
bwa	burbot	85.2 (2.1)
bwa	cod	30 (1.1)

Of the two reference genomes, a significantly greater proportion of burbot reads are unambiguously aligned when using the highly fragmented burbot reference (~85 % mapped) as opposed to the higher-quality but more genetically-distant cod reference (~30 % mapped; tables 1,2, fig. 2R). The effect of alignment tool is rather minimal in comparison, but this was still significant according to repeated measures ANOVA (with each individual taken into account as a random effect). Overall, there is relatively little variation in the percentage of reads mapped among individuals for a given treatment. Interestingly, a similar trend in individuals' values is seen across all treatments (fig. 2L).

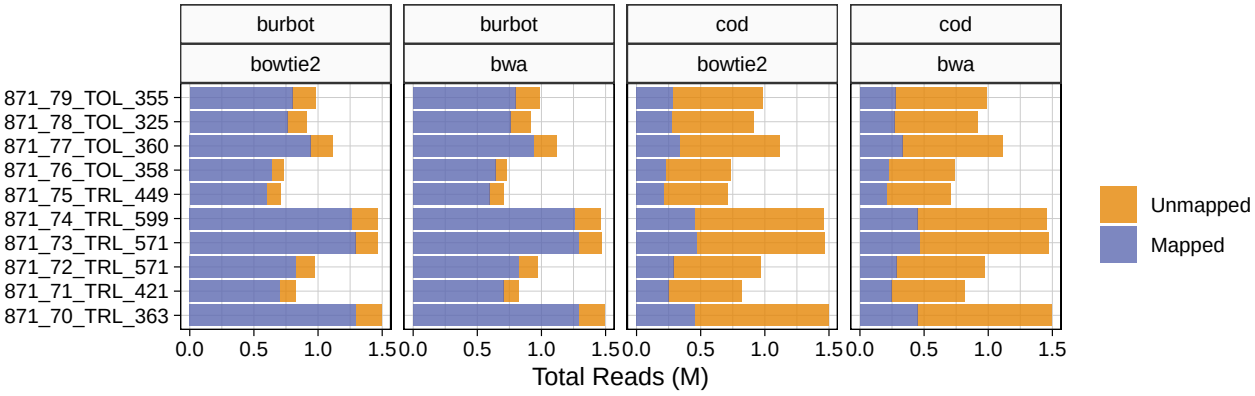

Figure 1: Total and mapped reads for each indivdual to either burbot or cod reference, with either 'bwa' or 'bowtie2' aligners.

Table 2: Summary table for repeated measures ANOVA for effects of reference genome and alignment tool choices on the percentage of burbot reads mapped.

Effect	DFn	DFd	F	p	p<.05	ges
Reference	1	9	18742	3.01e-16	*	0.997
Aligner	1	9	54.25	4.26e-05	*	0.013
Reference:Aligner	1	9	368.2	1.31e-08	*	0.006

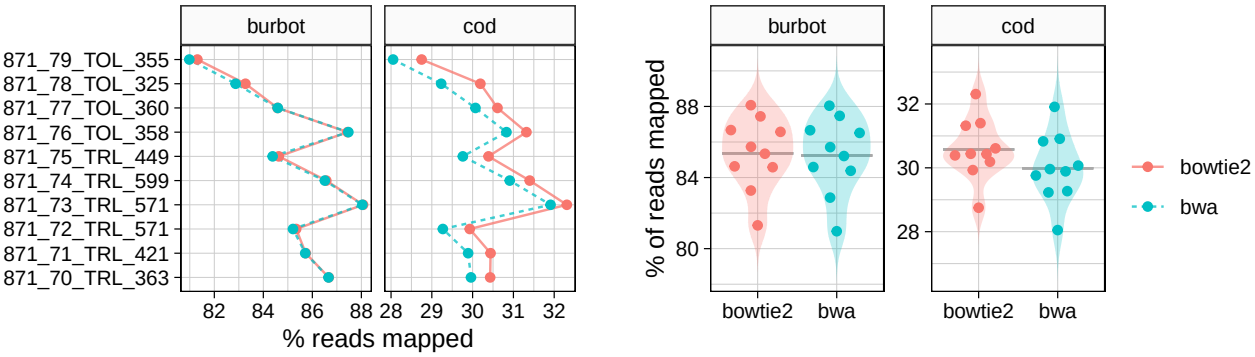

Figure 2. Percentage of burbot reads mapped against either burbot or cod reference, using either bwa or bowtie2. Each point is an individual, lines indicate the group mean. Note that vastly different scales are used for the % of reads mapped each reference genome.

Discussion

Reference-guided assembly is a critical processing step prior to identification of variants for genomics analyses. However, reference choice may be a source of bias in read-mapping and genotyping, influencing later analyses. Reference-bias occurs because read mapping has greater success for sequences that are similar to the reference specimen, leading to an underestimation of non-reference alleles (Brandt et al. 2015). In this work, I investigated whether the choice of cod or burbot reference genome would impact the proportion of reads mapped for ten burbots. These reference genomes have vast differences in assembly quality, where the cod genome (length 670 Mb) is comprised of 227 scaffolds and the burbot genome (length 397 Mb) has 106,616 scaffolds.

Of the two reference genomes, a greater proportion of burbot reads are unambiguously aligned when using the burbot reference than the cod reference (fig. 2R). Interestingly, the choice of reference appears to scale the proportion of reads mapped in a nearly identical fashion for each individual (fig. 2L). This suggests that variation in mapping rates among individuals is not due to differences in genetic distance to either reference. We would take this as evidence that read-mapping success is not differentially-biased among individuals by the choice of reference for this data.

In contrast to reference selection, the choice of `bowtie2` or `bwa` as the alignment tool has only a small impact on mapping rates (fig. 2), which is not surprising given that they use similar approaches based on the Burrows-Wheeler transform (Li and Durbin 2009). The two aligners had equal mapping rates when the burbot reference is used, while `bowtie2` had slightly better performance across all individuals when mapping to the cod reference (fig. 2L). Despite the small effect size relative to variation among individuals, the choice of alignment and the interaction effect of reference genome and alignment tool were both significant according to repeated measures ANOVA (table 2). Nevertheless, it seems that either tool should work equally well on a given reference genome, with `bowtie2` potentially having slightly better performance when reads are mapped to a more distant reference.

Generally, we would choose the burbot genome for reference-guided assembly of burbot reads, since this yields a far greater proportion of mapped reads from a given individual and will provide us with more data for genotyping and later analysis. However, if conducting a study comparing burbot and other species rather than only burbot populations, it may be beneficial to use the more complete but genetically-distant cod reference to reduce bias in read mapping (Gopalakrishnan et al. 2017). Depending on the goals of the experiment, the completeness and functional annotation of the reference genomes may also be important factors to consider. If we were wanting to compare results to others in the literature, it might be beneficial to use the same reference (Lloret-Villas et al. 2021). A caveat of this analysis is that reference genome quality and content were not investigated, only the proportions of reads mapped were compared among treatments. From this data alone, I cannot speculate on whether choice of reference genome or alignment tool would lead to biases in variant calling for these individuals.

References

Bohling, Justin. 2020. "Evaluating the Effect of Reference Genome Divergence on the Analysis of Empirical RADseq Datasets." *Ecology and Evolution* 10 (14): 7585–7601. https://doi.org/https://doi.org/10.1002/ece3.6483.

Brandt, Débora Y. C., Vitor R. C. Aguiar, Bárbara D. Bitarello, Kelly Nunes, Jérôme Goudet, and Diogo Meyer. 2015. "Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data." *G3: GenesGenomesGenetics* 5 (5): 931–41. https://doi.org/10.1534/g3.114.015784.

Gopalakrishnan, Shyam, Jose A. Samaniego Castruita, Mikkel-Holger S. Sinding, Lukas F. K. Kuderna, Jannikke Räikkönen, Bent Petersen, Thomas Sicheritz-Ponten, et al. 2017. "The Wolf Reference Genome

Sequence (Canis Lupus Lupus) and Its Implications for Canis Spp. Population Genomics." *BMC Genomics* 18 (1): 495. https://doi.org/10.1186/s12864-017-3883-3.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics (Oxford, England)* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Lloret-Villas, Audald, Meenu Bhati, Naveen Kumar Kadri, Ruedi Fries, and Hubert Pausch. 2021. "Investigating the Impact of Reference Assembly Choice on Genomic Analyses in a Cattle Breed." *bioRxiv*, January, 2021.01.15.426838. https://doi.org/10.1101/2021.01.15.426838.

Tørresen, Ole K., Bastiaan Star, Sissel Jentoft, William B. Reinar, Harald Grove, Jason R. Miller, Brian P. Walenz, et al. 2017. "An Improved Genome Assembly Uncovers Prolific Tandem Repeats in Atlantic Cod." *BMC Genomics* 18 (1): 95. https://doi.org/10.1186/s12864-016-3448-x.