Whole genome sequences from non-invasively collected samples

Rebecca Taylor[1], Micheline Manseau[2], Bridget Redquest[1], and Paul Wilson[1,3]

[1]Trent University Department of Biology
[2]Environment and Climate Change Canada National Wildlife Research Centre
[3]Trent University

April 28, 2020

Abstract

Conservation genomics is an important tool to manage threatened species under current biodiversity loss. Recent advances in sequencing technology mean that we can now use whole genomes to investigate demographic history, local adaptation, inbreeding, and more in unprecedented detail. However, for many rare and elusive species only non-invasive samples such as faeces can be obtained, making it difficult to take advantage of whole genome data. We present a method to extract DNA from the mucosal layer of faecal samples to reconstruct high coverage whole genomes using standard laboratory techniques, therefore in a cost-effective and efficient way. We use wild collected faecal pellets collected from wild caribou (Rangifer tarandus), a species undergoing declines in many parts of its range in Canada and subject to comprehensive conservation and population monitoring measures. We compare four faecal genomes to two tissue genomes sequenced in the same run. Quality metrics were similar between faecal and tissue samples with the main difference being the alignment success of raw reads to the reference genome likely due to differences in endogenous DNA content, affecting overall coverage. One of our faecal genomes was only reconstructed at low coverage (1.6X), however the other three obtained between 7 and 15X, compared to 19 and 25X for the tissue samples. We successfully reconstructed high-quality whole genomes from faecal DNA and, to our knowledge, are the first to obtain genome-wide data from wildlife faecal DNA in a non-primate species, representing an important advancement for non-invasive conservation genomics.

1 | INTRODUCTION

Human induced global biodiversity loss, for example due to habitat destruction and/or climate change, is accelerating (Brandies, Peel, Hogg, & Belov, 2019; Funk, Forester, Converse, & Darst, 2019; Harrison, Pavlova, Telonis-Scott, & Sunnucks, 2014; McMahon, Teeling, & Höglund, 2014; Shafer et al., 2015). Conservation genomics is one tool to help with the management of threatened taxa, particularly with recent advances in sequencing technologies and reducing costs (Brandies et al., 2019; Perry, Marioni, Melsted, & Gilad, 2010; Shafer et al., 2015). There are many articles outlining the advantages of genome-wide data for conservation including for estimating demographic histories, and for detecting genomic regions involved with local adaptation or inbreeding depression (e.g. Allendorf, Hohenloe, & Luikart, 2010; Harrison et al., 2014; McMahon et al., 2014; Shafer et al., 2015). Different genomic methods have been developed, including reduced-representation sequencing (RRS), however there are clear advantages of having whole genome information (Fuentes-Pardo & Ruzzante, 2017). Understanding adaptation is touted as one of the major advantages of genomics, however many adaptive traits are polygenic and may not be detected using RRS (Brandies et al., 2019; Fuentes-Pardo & Ruzzante, 2017; Funk, McKay, Hohenloe, & Allendorf, 2012; McMahon et al., 2014; Shafer et al., 2015). Similarly, whole genomes can be used to determine the genetic basis of phenotypic traits or diseases of interest to conservation (Brandies et al., 2019; Fuentes-Pardo & Ruzzante, 2017).

For many threatened taxa obtaining high-quality samples can be difficult, therefore advances in non-invasive

genetics have been important for conservation initiatives as they allow the study of rare or elusive species without needing to handle, or sometimes even see, the target species (Ozga et al., preprint, Smith & Wang, 2014; Snyder-Mackler et al., 2016). There are many types of non-invasive samples, including faeces, hair, urine, feathers, egg shells, and skin (Beja-Pereira, Oliveira, Alves, Schwartz, & Luikart, 2009; Russello, Waterhouse, Etter, & Johnson, 2015; Smith & Wang, 2014), however, faecal samples are commonly used since they are easy to obtain and can provide additional relevant information such as hormones, microbiome, and diet (Chiou & Bergey, 2018; Perry et al., 2010). However, obtaining genome-wide data from non-invasive samples is challenging due to low host (endogenous) DNA in extractions, fragmented DNA, the presence of PCR inhibitors, and high levels of allelic dropout, all of which are particularly true for faecal DNA (Chiou & Bergey, 2018; Perry et al., 2010; Smith & Wang, 2014; Snyder-Mackler et al., 2016).

New, promising approaches have been developed to sequence genomic data from faecal samples. Most have used sequence capture methodologies which use DNA or RNA baits to hybridise to target DNA (Chiou & Bergey, 2018). For example, Perry et al. (2010) used a DNA capture protocol with custom baits to enrich megabases of nuclear genomic regions and the mitochondrial genome from chimpanzees. Snyder-Mackler et al. (2016) were the first to use genome-wide enrichment capture from RNA baits to enrich faecal DNA, which resulted in low-coverage (a mean of 0.493) data for baboons. However, capture methodologies can be expensive and time consuming, have high PCR duplication rates, and bias the resulting datasets towards particular regions of the genome (Chiou & Bergey, 2018; Orkin et al., preprint). A recent study by Ozga et al. (preprint) tested different non-invasive samples from chimpanzees using both whole genome and exome capture methods, and found that urine had much higher success than faecal DNA, producing genome-wide data using the same extraction and sequencing methods as with high quality tissue samples (needing no extra methodological considerations). However, urine is not always easily collected for many taxa, and the capture method still does not give unbiased whole genome coverage.

Chiou and Bergey (2018) developed a cost-effective method they called 'FaecalSeq', which takes advantage of the difference in CpG-methylation densities between bacterial and vertebrate genomes to enrich host faecal DNA. They validated their FaecalSeq approach using double-digest restriction-site associated DNA sequencing (ddRADseq) to obtain genome-wide SNP data in baboons (Chiou & Bergey, 2018). However, FaecalSeq still biases which genomic regions are captured based on methylation patterns, and can co-enrich non-target DNA such as from plant and animal food sources (Chiou & Bergey, 2018; Ozga et al., preprint). The only study to date to obtain unbiased 'uniform' high and low coverage whole genome sequences from faecal DNA is from Orkin et al. (preprint). They collected faecal samples from capuchin monkeys and used fluorescence activated cell sorting (FACS) to isolate mammalian cells from the feces, as an alternative to enriching host DNA. They successfully reconstructed one high coverage (12X) and 15 low coverage (0.1-4X) re-sequenced genomes, however the use of FACS adds additional expense ($40 per sample to isolate the cells) and assumes the availability of FACS resources (Orkin et al., preprint).

Using a protocol whereby we carefully extract DNA from the mucosal layer of faecal pellets collected from wild caribou (*Rangifer tarandus*), we attempted to reconstruct four high coverage re-sequenced whole genomes using only standard laboratory techniques, therefore in a very cost-effective and efficient way. Caribou (known as reindeer in Europe and Asia) occur across Canada in different ecozones from the High Arctic to the boreal forests (Banfield, 1967; COSEWIC, 2011). In Canada there are four subspecies and 12 conservation units, known as Designatable Units (DUs; Banfield, 1967; COSEWIC, 2011). All 11 of the extant DUs are listed as at risk of extinction (COSEWIC, 2014-2017), and many are threatened due to anthropogenic activities such as habitat destruction and climate change (Festa-Bianchet, Ray, Boutin, Côté, & Gunn, 2011; Vors & Boyce, 2009; Weckworth, Hebblewhite, Mariani, & Musiani, 2018). Caribou is a keystone species for the ecosystem and is of cultural and economic significance to indigenous communities (Festa-Bianchet et al., 2011; Polfus et al., 2016), highlighting the need for population monitoring and conservation initiatives.

Genetic analyses using microsatellites and mitochondrial DNA sequences from winter collected faecal samples have been fundamental in understanding population structure and evolutionary history of Canadian caribou (e.g. Horn et al., 2018; Klütsch, Manseau, Trim, Polfus, & Wilson, 2016; Klütsch, Manseau, & Wilson, 2012;

Polfus, Manseau, Klütsch, Simmons, & Wilson, 2017; Polfus et al., 2016), as well as for monitoring population trends (Hettinga et al., 2012), pedigree reconstruction and inbreeding estimations (e.g. McFarlane et al., 2018; Thompson, Klütsch, Manseau, & Wilson, 2019). To further aid with conservation efforts, understanding the genomic effects of inbreeding in small populations, as well as locally adapted genomic variation are important next steps. However, obtaining high quality samples such as tissue or blood for next-generation sequencing requires handling and some level of harm to the animals, which is less than ideal for a threatened species, or opportunistic sampling of dead individuals. We have a repository of ~40,000 winter collected faecal samples in the lab, and so a priority for us was to be able to obtain whole genome information from this vast non-invasively collected set of samples. Typically, the main issue with using faecal DNA for cost-effective genome-wide sequencing is low endogenous DNA (Chiou & Bergey, 2018; Perry et al., 2010; Snyder-Mackler et al., 2016), and so by extracting DNA from the mucosal layer on the outside of the faecal pellets we increased the likelihood of extracting host DNA from the epithelial cells of the intestines (Ball et al., 2007). We compare our four faecal genomes to two tissue genomes sequenced in the same run to assess performance and bias in the resulting data. As such, we are the first to assemble high coverage whole genomes from faecal DNA using standard laboratory and sequencing techniques alone. Similarly, to our knowledge all previous studies obtaining genome-wide data from faecal DNA have been from primate species, and so we are the first to do so in a non-primate non-model organism.

2 | MATERIALS AND METHODS

2.1 | Sample collection and DNA extraction

Faecal pellets were collected during winter aerial surveys between 2004 and 2018 (Table 1). Pellets were collected from the snow, bagged, and kept frozen for shipping to Trent University. Tissue samples were collected during harvesting activities in 1994 and 2002 (Table 1) and stored in 1X lysis buffer.

For the faecal genomes, three individuals were boreal caribou; two from Cold Lake, Alberta, and one from Wood Buffalo National Park, Northwest Territories. The fourth was a central mountain caribou from A La Peche. The samples were chosen to fill in sampling gaps for other conservation genomic analyses being undertaken in the lab for areas where we do not have tissue samples. Multiple faecal samples from these locations had been genotyped at microsatellite loci and to select which faecal samples to re-extract for whole genome sequencing, we surveyed the raw genotype files and looked for those with the cleanest, highest peaks, to select those most likely to have the highest amounts of high-quality endogenous DNA. Faecal pellets with the most amount of mucous layer visible were then chosen for the extractions.

DNA was extracted from the mucosal coat on the faecal pellets from four individuals. To do this, four faecal pellets were put into a tube with 1ml of lysis buffer and gently rotated or washed for about 30 seconds. The faecal pellet and (after settling) any precipitate were discarded, following which 10µl of proteinase K was added to the lysis buffer and the sample incubated at 65°C for two hours. Another 10µl of proteinase K was then added and the sample was left at 37°C overnight. For each individual, we did this process twice. DNA extraction was then carried out using a DNAeasy Blood and Tissue Kit (Qiagen, Hilden, Germany). The DNA was eluted with 200µl of TE, and the two extractions for each individual combined for a total of 400µl. The samples were then run through a concentration column (Millipore 30K Device, Millipore Sigma, Burlington, MA, USA). To do this, the 400µl of Qiagen extracted sample was loaded into the column and spun at 14,000g for 5 minutes giving the final volume at ~50µl at a concentration factor of 12x.

Tissue samples from the Tay population in the Yukon (a northern mountain caribou) and the Fortymile caribou population (Grant's caribou) straddling the Yukon and Alaska border were also extracted using a DNAeasy Blood and Tissue Kit (Qiagen, Hilden, Germany).

2.2 | Quality control and sequencing

The DNA extractions were run on a 1.5% agarose gel, and run on a Qubit fluorometer (Thermo Fisher Scientific, MA, USA) using the High Sensitivity Assay Kit to ensure high DNA concentrations for sequencing. The samples were also run on a Nanodrop ND-8000 spectrophotometer (Nanodrop Technologies Inc.,

Wilmington, DE, USA) to assess purity. The DNA was normalized to 20ng/µl at a final volume of 50µl for the tissue samples and to 22ng/µl at a final volume of 50µl for the faecal samples and shipped to The Centre for Applied Genomics (TCAG) at the Hospital for Sick Children (Toronto, Ontario) for library preparation and sequencing. The samples were run alongside 10 other samples being used for another study (for a total of 16 samples) on 8 lanes of an Illumina HiSeq X (Illumina, San Diego, CA, USA). All raw reads will be made available on the NCBI by the time of publication

2.3 | Filtering reads and variant calling

We used Trimmomatic version 0.38 (Bolger, Lohse, & Usadel, 2014) to trim adaptors and other Illumina sequences from the reads which can result from sequencing very short DNA fragments, as may be expected from lower quality DNA. We used the sliding window approach (4 base pairs at a time) to trim reads once the phred score dropped below 15. Reads were aligned to the reference genome (Taylor et al., 2019) using Bowtie2 version 2.3.0 (Langmead & Salzberg, 2012), and the SAM file converted to a BAM file using Samtools version 1.5 (Li et al., 2009). We then sorted the BAM files using Samtools, and then removed duplicate reads and added correct read group information to each BAM file using Picard version 2.17.3 (Available: http://broadinstitute.github.io/picard/). We then re-sorted the BAM file and built an index using Picard.

We called variants using Haplotype Caller in GATK version 3.8 (McKenna et al., 2010) and produced a variant call format (VCF) file for each caribou. Individual VCF files were combined using the Combine GVCFs function, and then we performed joint genotyping using Genotype GVCFs, both in GATK. Due to one faecal genome being of poor quality compared to the others (see Results), we also produced a joint VCF file with this individual removed. We used VCFtools version 0.1.14 (Danecek et al., 2011) to filter the VCF files to ensure quality. We did two rounds of filtering, firstly to remove indels and any site with a depth of less than 5 or more than 40 (double the depth we were aiming for across the genome), and removed any low-quality genotype calls (minGQ) and low quality sites (minQ), with scores below 20, which in VCFtools are changed to missing data. Secondly, we filtered to remove all missing data.

2.4 | Quality assessments

The average depth for each BAM file was calculated both before and after duplicate removal using Samtools, and all BAM files were checked (after duplicate removal) using FastQC version 0.11.8 (Andrews, 2010). We also ran each BAM file through BUSCO version 3.0.2 (Benchmarking Universal Single-Copy Orthologs; Waterhouse et al., 2018) to reconstruct 4,104 conserved mammalian genes to assess the completeness of each genome. As our reference genome reconstructed 3,820 (93.1%; Taylor et al., 2019) complete mammalian BUSCO genes, this represents an upper limit for our re-sequenced individuals. We used Picard to run some quality checks on the BAM files, using 'CollectGcBiasMetrics' to assess GC content and produce statistics regarding GC bias in the genomes, 'CollectWgsMetrics' to assess the fraction of reads that pass quality filters for each of the genomes, and 'QualityScoreDistribution' to output quality scores of all bases.

We did two population genomic analyses with the genomes to assess how they performed. Using both combined VCF files, we performed a principle component analyses in R version 3.4.4 (R Development Core Team, 2006) using the packages vcfR (Knaus & Grünwald, 2017) and Adegenet (Jombart, 2008). We also used the populations module in Stacks version 2.4.1 (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) to convert our VCF files into input files for Treemix version 1.13 (Pickrell & Pritchard, 2012). We ran Treemix from 0-4 migration events, with three iterations of each. We grouped the SNPs in windows to account for possible linkage, using a block size of 100 for two of the iterations and 50 for one of the iterations (because to run the OptM package you must not have the same likelihood scores for each iteration). We plotted the resulting trees and residual plots in RStudio version 1.0.136 (RStudio Team, 2015), and used the R package OptM (Fitak. In Review) to calculate the second order rate of change in the log-likelihood of the different migration events (the ad hoc statistic delta M). We also tried running without accounting for linkage due to the smaller number of SNPs in the VCF file with all 6 caribou, however the trees were the same.

3 | RESULTS

All DNA extractions had relatively high concentrations before normalising, however with three of the faecal extractions having slightly lower Nanodrop scores than the 1.8 "ideal" ratio (Table 1). All four faecal genomes and the two tissue genomes had comparable numbers of reads and percentage of reads retained after trimming (Table 1). However, the number of reads which successfully aligned to the reference genome was highly variable for the faecal genomes, ranging from 5 to 60%. In contrast, the two tissue genomes had very high alignment success at around 95-96% (Table 1). This resulted in varying depth of coverage overall for each individual. One faecal genome, from Wood Buffalo, had only an average depth of 1.63 after duplicate removal, the others achieving between 7 and 15X coverage. The tissue genomes were 19 and 25X coverage (Table 1). The depth did not drop significantly more for the faecal genomes after duplicate removal, however, indicating that they did not contain an inflated number of PCR or sequencing duplicates. All FastQC results from the BAM files looked good, with the per base sequencing quality not dropping below 28 even at the ends of the reads, high per sequence quality scores and no detected duplication levels, overrepresented sequences, or adaptor content.

BUSCO successfully reconstructed 92-93% of the conserved mammalian genes for all genomes apart from Wood Buffalo which reconstructed 44.6% (Table 2, Figure 1). The GC distribution was the same between all 6 genomes (Figures 2a-f). They all had similar mean base quality scores across regions of the genome with different percentage GC, although with windows of very high GC content dropping in quality score. Wood Buffalo, however, decreased dramatically in regions with high GC content (Figures 2a-f). Normalised coverage also seemed to be affected by GC content in all genomes, with Wood Buffalo again dropping dramatically compared to the others (Figures 2a-f). The BAM files all showed no adaptors or duplicates, as expected given prior filtering (Table 3). The percentage of bases with low quality scores and those in reads without a mapped read pair were all low and consistent between the genomes (Table 3). The percentage of bases with a low mapping score was higher, and slightly elevated in the Wood Buffalo genome (Table 3), showing the importance of quality filtering when producing the VCF file. Differences in coverage levels are also very apparent between the genomes (Table 3). The theoretical heterozygous SNP sensitivity scores, which is an estimate of the sensitivity to detect heterozygous sites (between 0 and 1), also varied. The score was low for the Wood Buffalo genome, but was high for all other genomes apart from the lowest quality Cold Lake individual which was intermediate (Table 3). The quality score distribution of the base pairs in the BAM files was consistent between all individuals, with the vast majority showing high quality scores (Figures 3a-f).

The VCF files with all 6 genomes before removing missing data contained 18,438,793 SNPs. However, the missing data was heavily skewed towards the faecal genome from Wood Buffalo which had 99% missing data (Table 2). This is potentially because we filtered for low quality genotype scores and sites, which are changed to missing data in VCFtools. After removing all missing data from the VCF file, only 25,390 SNPs remained. Additionally, the Wood Buffalo genome had an order of magnitude more private SNPs (Table 2). In the VCF file without the Wood Buffalo caribou, there were 18,261,032 SNPs before removing missing data. Missing data levels were quite high for one of the Cold Lake caribou (Table 2), however when removing all missing data 5,065,428 SNPs were still retained. The lower quality genome from Cold Lake had a slightly elevated number of private sites, indicating the potential for some errors due to quality affecting SNPs called for that individual (Table 2).

We used both VCF files with no missing data to do PCA's and Treemix analyses, to assess how well they would perform. The PCA using all six caribou showed a pattern that we expected (Taylor et al., In Review), although with the boreal caribou (both Cold Lake and the Wood Buffalo caribou) quite separated from one another. The Central mountain caribou (A La Peche) separated from all others, and the Northern Mountain (Tay) and Grant's caribou (Fortymile) sat closer together which matches the geography of the sampling sites (Figure 4a). The PCA without the Wood Buffalo caribou showed the two Cold Lake boreal caribou sitting closer together, and with the Northern mountain and Grant's caribou also sitting closer together (Figure 4b) which may be due to increased power from the greater number of SNPs used in the analysis. The Treemix analysis failed to build a tree when including the Wood Buffalo caribou, showing a large standard error bar (Figure 5a). In contrast, when removing Wood Buffalo the analysis could reconstruct a phylogeny which grouped the Grant's caribou (Fortymile) with Northern Mountain (Tay) as a separate clade to the other

three as expected (Figure 5b). When adding migration events, after 2 migration events no new migration events could be inferred. The OptM analysis gave 1 migration event as having the highest delta M, which showed a migration event from the ancestor of the Grant's (Fortymile) and Northern Mountain (Tay) caribou into a Cold Lake caribou (Figure 5c).

4 | DISCUSSION

We have successfully reconstructed both high and low coverage whole genomes from faecal DNA using only standard laboratory protocols and sequencing. We are also the first to obtain genome-wide data from faecal DNA in a non-primate species. This represents an important contribution for non-invasive conservation studies to move from genetics to genomics and investigate questions such as the local adaptation of populations. Further, our method to extract DNA from the mucosal layer of faecal pellets is cost effective. The cost of all lab processes and sequencing was $803.17 CAD (or about $567.92 USD) per sample ($6.60 for the extraction, $6.54 for the concentrator column, $0.83 for the Qubit HS kit, $135.00 for the library prep, and $657.50 for sequencing on half an Illumina lane on a HiSeq X). Our reference genome assembly is 2.205Gb (Taylor et al., 2019) and so sequencing effort for other taxa will need to be adjusted depending on genome size, however the costs of producing our faecal genomes was the same as for our tissue samples apart from the $6.54 for the concentrator column and half of the extraction cost ($3.30).

Overall, the quality statistics were very similar between our tissue and faecal genomes, apart from the Wood Buffalo individual, with no sign of lowered per base sequencing quality, a skewed GC content, inflated numbers of PCR duplicates, percentage of bases in reads with lower mapping quality, skewed quality score distributions of base pairs, or the number of genes reconstructed in a BUSCO analysis, for example (Table 3, Figures 1, 2a-f, and 3a-f). As expected, the main difference was in the alignment success of reads to the reference genome likely reflecting endogenous DNA content. We aimed for high coverage whole genomes, and the Wood Buffalo individual was only reconstructed at low coverage. Another of our genomes, from the Cold Lake population, was also a bit low at around 7X coverage which may have affected the quality of genotyping as it has a slightly elevated number of private SNPs and a lower theoretical heterozygous SNP sensitivity. With a slight refinement in laboratory techniques, it may be possible to increase the likelihood of selecting samples with higher endogenous DNA content to ensure results like our more successful faecal genomes. To select which samples to sequence, we looked at raw genotype peaks from microsatellite scores to assess quality and endogenous DNA content. However, a better method (or to use in combination) might be to use qPCR to screen for proportion of host DNA within extractions (Chiou & Bergey, 2018; Hayward et al., 2020; Snyder-Mackler et al., 2016), or the PCR method developed by Ball et al. (2007). If used in combination with our DNA extraction technique, it is likely that genomes to the standard of our high quality faecal genomes will be more consistently produced, further increasing the cost effectiveness of our method. Unfortunately, we have no more DNA left from the extractions used for our faecal genomes so we cannot screen them post-hoc to test for a correlation, but we plan to use PCR quantification technique moving forward for the next batch of faecal genomes (Ball et al., 2007; Hayward et al., 2020). This should likely be standard practice for researchers choosing samples for sequencing.

Other important and standard checks we completed included measuring DNA concentration using a Qubit and purity using a Nanodrop. As we only did four samples we cannot do a quantitative analysis, but it is interesting that our best faecal genome, which attained almost the same coverage as our tissue samples, was also the only faecal sample to reach a Nanodrop reading of 1.8 (Table 1). Also, our faecal sample which could only be used as a low coverage genome, Wood Buffalo, had a very high Qubit reading compared with our 'average' faecal samples. At 80ng/µl, it has reached the same concentrations as we see with our tissue extractions (Table 1). We wonder if a spuriously high DNA concentration may indicate high levels of bacterial DNA, something which would be worth testing in future.

With any genomic data produced from non-invasive samples, strict filtering and careful monitoring of data quality is vital. We performed extensive data quality assessments with our genomes to assess potential areas of bias (Tables 1-3, Figures 1-3). Filtering for low quality sites is standard practice with any whole genome data but is even more important with faecal samples as we expect higher genotyping errors with

poorer quality DNA. We filtered low quality sites (both base and mapping quality) in VCFtools which is changed to missing data. Missing data filtering before further analyses is therefore crucial due to missing data levels being heavily skewed towards lower quality samples (Table 2). After removing the missing data, if we included the low quality faecal genome we ended up with an order of magnitude fewer SNPs in the VCF file than when we excluded it (25,390 vs 5,065,428). We also tested the performance of the genomes with two standard analyses, PCA and Treemix. Clearly the inclusion of the lower quality Wood Buffalo genome affected the results, especially the Treemix which completely failed to reconstruct a phylogeny (Figure 5). As we only included one individual per population we did not use a minor allele frequency filter, although with multiple individuals per population in a larger dataset this could also be an important filtering step.

Our comparisons between faecal and tissue DNA would have been more useful if we had used samples from the same individual, however obtaining tissue samples can only be done opportunistically or through regulated harvesting activities. As such we do not have any samples with both faecal and tissue. We could potentially have sequenced both faecal and tissue samples from the same area, however we needed to fill in key gaps in our sampling for evolutionary and conservation genomic analyses being undertaken in our lab. As such, we chose to sequence four faecal samples where we have no available tissue given tight resources as is common for conservation genomics studies.

One potential drawback of our method, as well as many other methods being developed for producing genome-wide data from faecal samples (Orkin et al. preprint; Perry et al., 2010; Snyder-Mackler et al., 2016), is the need for a reference genome. The technique developed by Chiou and Bergey (2018) does not require a reference for production of genome-wide ddRADseq SNP data, however to account for possible co-enrichment of food or contaminant sources, alignment to a reference would be highly beneficial. Additionally, for any whole-genome resequencing project, a reference genome is essential (Fuentes-Pardo et al., 2017). However, with the costs coming down and increased availability of bioinformatics pipelines for non-model species (Brandies et al. 2019; Fuentes-Pardo et al., 2017), the availability of a reference genome is becoming less of an issue, especially with initiatives such as the CanSeq150 (www.cgen.ca/canseq150) and the Genome 10K project (Koepfli, Paten, The Genome 10K Community of Scientists, O'Brien, 2015). Another advantage we had, in addition to a high-quality reference genome, is the collection of faecal samples in winter from the snow. The fact that the samples are collected while frozen will mean lower degradation of DNA than if they had been collected, for example, in the tropics (Smith & Wang, 2014). As such, for many taxa the collection of fresh faecal matter which is immediately frozen or appropriately stored would be highly beneficial.

Although one of our faecal genomes could not be used as a high coverage genome, all of them could be used a low-coverage genomes which are typically between 1-4X per individual (Fuentes-Pardo et al., 2017). More individuals would be needed for genotype likelihood calls which may be cost prohibitive, although improved lab screening for samples with higher amounts of endogenous DNA will improve sample selection and therefore the number of samples which could be run on one sequencing lane if low coverage genomes are the desired outcome. Sequencing whole genomes from few individuals per population for in-depth analyses (e.g. investigations of local adaptation or runs of homozygosity) to supplement traditional genetic methods may be most cost-effective for non-invasive monitoring of threatened taxa.

Overall, our method to extract high-quality DNA for whole genome sequencing from non-invasively collected faecal samples is an important step forward in our ability to study and monitor caribou using our already existing sample collection. We will now be able to sequence genomes from populations for which we had no existing tissue samples for comprehensive investigations of adaptation, inbreeding, and demographic histories of caribou across North America which will be invaluable knowledge to inform the conservation of this declining species. Further, our protocol for extracting DNA from the mucosal layer of faecal matter could be used in other taxa, especially if they have access to winter or freshly collected samples. Together with our thorough considerations of data quality and bias, we hope other research groups will be able to produce high-quality whole genome data for other rare or elusive species.

ACKNOWLEDGEMENTS

Funding was provided through an NSERC Collaborative Research & Development (CRD) grant, NSERC grant RGPIN-2015-04477, Manitoba Hydro, Saskatchewan Power, and Weyerhaeuser Inc. We thank Austin Thompson for help with lab work, Sonesinh Keobouasone for help with data management, Melinda-Lee Baker for help with literature searches, and The Centre for Applied Genomics (TCAG) at the Hospital for Sick Children (Toronto, Ontario) for library preparation and sequencing, to the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute Canada/ Calcul Canada gme-665-ab.Compute Canada (RRG gme-665-ab), and Amazon Cloud Computing for high-performance computing services. We thank sample collectors – Mary Gamberg who collected tissue samples during harvesting activities for the Environment and Climate Change Canada Northern Contaminants Program, biologists with the Alberta government, Jasper and Wood Buffalo National Parks.

AUTHOR CONTRIBUTIONS

R.S.T. helped with study design, did the bioinformatics, and wrote the manuscript, M.M. conceived the study, acquired funding, and edited the manuscript, B.R. helped with study design, did the laboratory work, and edited the manuscript, and P.J.W. conceived the study, acquired funding, and edited the manuscript.

DATA AVAILABILITY STATEMENT

All raw data will be available on the NCBI database upon acceptance.

REFERENCES

Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics* ,*11* , 697–709. http://doi.org/10.1038/nrg2844

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Ball, M. C., Pither, R., Manseau, M., Clark, J., Petersen, S. D., Kingston, S., ... Wilson, P. (2007). Characterization of target nuclear DNA from faeces reduces technical issues associated with the assumptions of low-quality and quantity template. *Conservation Genetics, 8* , 577–586. http://doi.org/10.1007/s10592-006-9193-y

Banfield, A.W.F. (1967). *A Revision of the Reindeer and Caribou, Genus Rangifer* . National Museum of Canada, Bulletin No. 177, Queen's Printer: Ottawa, ON, Canada.

Beja-Pereira, A., Oliveira, R., Alves, P. C., Schwartz, M. K., & Luikart, G. (2009). Advancing ecological understandings through technological transformations in noninvasive genetics. *Molecular Ecology Resources* , *9* , 1279–1301. http://doi.org/10.1111/j.1755-0998.2009.02699.x

Bolger, A. M., Lohse, M., Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30,*2114–2120. http://doi.org/10.1093/bioinformatics/btu170

Brandies, P., Peel, E., Hogg, C. J., & Belov, K. (2019). The value of reference genomes in the conservation of threatened species.*Genes, 10,* 846. http://doi.org/10.3390/genes10110846

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics.*Molecular Ecology* , *22* , 3124-3140. https://doi.org/10.1111/mec.12354

Chiou, K. L., & Bergey, C. M. (2018). Methylation-based enrichment facilitates low-cost , noninvasive genomic scale sequencing of populations from feces. *Scientific Reports, 8* , 1975. http://doi.org/10.1038/s41598-018-20427-9

COSEWIC (2016). *COSEWIC assessment and status report on the Caribou Rangifer tarandus, Barrenground population, in Canada.* Ottawa, ON: Committee on the Status of Endangered Wildlife in Canada.

COSEWIC (2017). *COSEWIC assessment and status report on the Caribou Dolphin and Union population (Rangifer tarandus), in Canada.*Ottawa, ON: Committee on the Status of Endangered Wildlife in Canada.

COSEWIC (2017). *COSEWIC assessment and status report on the Caribou Rangifer tarandus, Eastern Migratory population and Torngat Mountains population, in Canada.* Ottawa, ON: Committee on the Status of Endangered Wildlife in Canada.

COSEWIC (2014). *COSEWIC assessment and status report on the Caribou Rangifer tarandus, Newfoundland population, Atlantic-Gaspésie population and Boreal population, in Canada.* Ottawa, ON: Committee on the Status of Endangered Wildlife in Canada.

COSEWIC (2014). *COSEWIC assessment and status report on the Caribou Rangifer tarandus, Northern Mountain population, Central Mountain population and Southern Mountain population in Canada.* Ottawa, ON: Committee on the Status of Endangered Wildlife in Canada.

COSEWIC (2015). *COSEWIC assessment and status report on the Peary Caribou Rangifer tarandus pearyi in Canada.* Ottawa, ON: Committee on the Status of Endangered Wildlife in Canada.

COSEWIC (2015). *COSEWIC Assessment Process, Categories and Guidelines* . Ottawa, ON: Committee on the Status of Endangered Wildlife in Canada.

COSEWIC (2011). *Designatable Units for Caribou (Rangifer tarandus) in Canada* . Ottawa, ON: Committee on the Status of Endangered Wildlife in Canada.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics, 27,*2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Festa-Bianchet, M., Ray, J. C., Boutin, S., Côté, S. D., & Gunn, A. (2011). Conservation of caribou (*Rangifer tarandus*) in Canada: An uncertain future. *Canadian Journal of Zoology* , *89* , 419-434. https://doi.org/10.1139/z11-025

Fitak, R. R. (submitted). optM: an R package to optimize the number of migration edges using threshold models. *Journal of Heredity.*

Fuentes-pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology, 26,* 5369–5406. http://doi.org/10.1111/mec.14264

Funk, W. C., Forester, B. R., Converse, S. J., Darst, C., & Morey, S. (2019). Improving conservation policy with genomics: a guide to integrating adaptive potential into U.S. Endangered Species Act decisions for conservation practitioners and geneticists.*Conservation Genetics* , *20* , 115–134. http://doi.org/10.1007/s10592-018-1096-1

Funk, W. C., Mckay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology and Evolution* , *27* , 489–496. http://dx.doi.org/10.1016/j.tree.2012.05.012

Harrisson, K. A., Pavlova, A., Telonis-Scott, M., & Sunnucks, P. (2014). Using genomics to characterize evolutionary potential for conservation of wild populations. *Evolutionary Applications, 7* , 1008–1025. http://doi.org/10.1111/eva.12149

Hayward, K. M., Harwood, M. P., Lougheed, S. C., Sun, Z., de Groot, P. V. C., & Jensen, E.L., (2020). A real-time PCR assay to accurately quantify polar bear DNA in fecal extracts. *PeerJ, 8* , e8884. http://doi.org/10.7717/peerj.8884

Hettinga, P. N., Arnason, A. N., Manseau, M., Cross, D., Whaley, K., & Wilson, P. J. (2012). Estimating size and trend of the North Interlake woodland caribou population using fecal-DNA and capture-recapture models. *Journal of Wildlife Management, 76* , 1153–1164. https://doi.org/10.1002/jwmg.380

Horn, R. L., Marques, A. J. D., Manseau, M., Golding, G. B., Klütsch, C. F. C., Abraham, K., & Wilson, P. J. (2018). Parallel evolution of site-specific changes in divergent caribou lineages. *Ecology and Evolution, 8* , 6053–6064. http://dx.doi.org/10.1002/ece3.4154

Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics, 24,* 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Klütsch, C. F. C., Manseau, M., Trim, V., Polfus, J. L., & Wilson, P. J. (2016). The eastern migratory caribou: the role of genetic introgression in ecotype evolution. *Royal Soceity Open Science ,3,* 150469. https://doi.org/10.1098/rsos.150469

Klütsch, C. F. C., Manseau, M., & Wilson, P. J. (2012). Phylogeographical analysis of mtDNA data indicates postglacial expansion from multiple glacial refugia in woodland caribou (*Rangifer tarandus caribou*). *PLOS ONE, 7,* e52661. https://doi.org/10.1371/journal.pone.0052661

Knaus, B. J., & Grüwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources, 17,* 44–53. https://doi.org/10.1111/1755-0998.12549

Koepfli, K., Paten, B., The Genome 10K Community of Scientists, & O'Brien, S. J. (2015). The Genome 10K Project: A way forward.*Annual Review of Animal Biosciences, 3* , 57–111. http://doi.org/10.1146/annurev-animal-090414-014900

Langmead, B., & Salzberg. S. (2012). Fast gapped-read alignment with Bowtie2. *Nature Methods, 9,* 357-359. https://doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... 1000 Genome Profect Data Processing Subgroup. (2009). The sequence alignment / map format and SAMtools. *Bioinformatics ,25* , 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

McFarlane, S., Manseau, M., Flasko, A., Horn, R. L., Arnason, N., Neufeld, L., ... Wilson, P. J. (2018). Genetic influences on male and female variance in reproductive success and implications for the recovery of severely endangered mountain caribou. *Global Ecology and Conservation, 16* , e00451. https://doi.org/10.1016/j.gecco.2018.e00451

McKenna, A., Hanna, M., Banks, E., Sivachenki, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* , *20* , 1297–1303. https://doi.org/10.1101/gr.107524.110

McMahon, B. J., Teeling, E. C., & Höglund, J. (2014). How and why should we implement genomics into conservation? *Evolutionary Applications, 7* , 999–1007. http://doi.org/10.1111/eva.12193

Orkin, J. D., de Manuel, M., Krawetz, R., del Campo, J., Fontsere, C., Kuderna, L. F. K., ... Melin, A. D. Unbiased whole genomes from mammalian feces using fluorescence-activated cell sorting.*BioRxiv* . https://doi.org/10.1101/366112

Ozga, A. T., Webster, T. H., Gilby, I. C., Wilson, M. A., Nockerts, R. S., Wilson, M. L., ... Stone, A. C. Urine as a high-quality source of host genomic DNA from wild populations. *BioRxiv* . https://doi.org/10.1101/2020.02.18.955377

Perry, G. H., Marioni, J. C., Melsted, P., & Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces.*Molecular Ecology* , *19,* 5332–5344. http://doi.org/10.1111/j.1365-294X.2010.04888.x

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genetics, 8,* e1002967. https://doi.org/10.1371/journal.pgen.1002967

Polfus, J. L., Manseau, M., Klütsch, C. F. C., Simmons, D., & Wilson, P. J. (2017). Ancient diversification in glacial refugia leads to intraspecific diversity in a Holarctic mammal. *Journal of Biogeography* , *44* , 386–396. https://doi.org/10.1111/jbi.12918

Polfus, J. L., Manseau, M., Simmons, D., Neyelle, M., Bayha, W., Andrew, F., ... Wilson, P. J. (2016). Łeghagots'enete (learning together) the importance of indigenous perspectives in the identification of biological variation. *Ecology and Society, 21* , 18. http://dx.doi.org/10.5751/ES-08284-210218

R Core Team. (2015). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available at https://www.R-project.org/

RStudio Team. (2015). RStudio: Integrated Development for R. Boston, MA: RStudio, Inc. Available at http://www.rstudio.com/.

Russello, M. A., Waterhouse, M. D., Etter, P. D., & Johnson, E. A. (2015). From promise to practice: pairing non-invasive sampling with genomics in conservation. *PeerJ, 3* , e1106. http://doi.org/10.7717/peerj.1106

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., ... Zielinski, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution* , *30* , 78–87. http://dx.doi.org/10.1016/j.tree.2014.11.009

Smith, O., & Wang, J. (2014). When can noninvasive samples provide sufficient information in conservation genetics studies? *Molecular Ecology Resources, 14* , 1011–1023. http://doi.org/10.1111/1755-0998.12250

Snyder-mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., ... Tung, J. (2016). Efficient genome-wide sequencing and noninvasively collected samples. *Genetics* ,*203* , 699–714. http://doi.org/10.1534/genetics.116.187492

Taylor, R. S., Horn, R. L., Zhang, X., Golding, G. B., Manseau, M., & Wilson, P. J. (2019). The caribou (*Rangifer tarandus*) genome.*Genes, 10* , 540. https://doi.org/10.3390/genes10070540

Taylor, R. S., Manseau, M., Horn, R. L., Keobouasone, S., Golding, G. B., & Wilson, P. J. (In Review). The role of introgression and ecotypic parallelism in delineating intra-specific conservation units.*Molecular Ecology* .

[dataset] Taylor, R. S., Manseau, M., Redquest, B., & Wilson, P. J; (2020); Whole genome sequences for 'Whole genome sequences from non-invasively collected samples'; Will be archived into the NCBI database by the time of acceptance.

Thompson, L. M., Klütsch, C. F. C., Manseau, M., & Wilson, P. J. (2019). Spatial differences in genetic diversity and northward migration suggest genetic erosion along the boreal caribou southern range limit and continued range retraction. *Ecology and Evolution, 9* , 7030–7046. https://doi.org/10.1002/ece3.5269

Vors, L. S., & Boyce, M. S. (2009). Global declines of caribou and reindeer. *Global Change Biology* , *15* , 2626-2633. https://doi.org/10.1111/j.1365-2486.2009.01974.x

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ionnidis, P., Klioutchnikov, G., ... Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics.*Molecular Biology and Evolution* , *35* **,** 543–548. https://doi.org/10.1093/molbev/msx319

Weckworth, B. V., Hebblewhite, M., Mariani, S., & Musiani, M. (2018). Lines on a map: conservation units, meta-population dynamics, and recovery of woodland caribou in Canada. *Ecosphere, 9* , e02323. https://doi.org/10.1002/ecs2.2323

Hosted file

`image1.emf` available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

FIGURE 1 BUSCO results for each genome showing the number of conserved mammalian genes successfully reconstructed and in single copy (light blue), duplicate (dark blue), fragmented (yellow) or missing (red). All successfully reconstruct within 1% of all possible genes (given the reference genome), aside from the Wood Buffalo genome which only reconstructed 1,829 (44.6%) complete and single copy genes.

Hosted file

`image2.emf` available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

Hosted file

image3.emf available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

Hosted file

image4.emf available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

FIGURE 2 GC bias plots for each caribou showing the distribution of GC content in red bars along the bottom. For each window, the normalised coverage is shown in blue circles and the mean base quality (phred score) is shown by the green line.

Hosted file

image5.emf available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

Hosted file

image6.emf available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

Hosted file

image7.emf available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

FIGURE 3 Quality score distributions for all base pairs in the BAM file for each genome.

Hosted file

image8.emf available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

FIGURE 4 PCA with all six caribou included (a) and without the Wood Buffalo caribou (b).

Hosted file

image9.emf available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

Hosted file

image10.emf available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

FIGURE 5 Treemix analysis using all six genomes failed to show any topology and had a large standard error bar (a), however when removing the lowest quality genome a maximum Likelihood tree could be produced (b). Adding one migration event inferred migration from the ancestor of the clade containing the Grant's and northern mountain caribou into a Cold Lake individual (c).

Hosted file

Taylor_etal_TABLE1.docx available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

Hosted file

Taylor_etal_TABLE2.docx available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples

Hosted file

Taylor_etal_TABLE3.docx available at https://authorea.com/users/304235/articles/446329-whole-genome-sequences-from-non-invasively-collected-samples