Molecular Ecology Resources

WILEY

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

MhGeneS: An Analytical Pipeline to Allow for Robust Microhaplotype Genotyping

Julia C. Geue[1,2] | Peng Liu[2] | Sonesinh Keobouasone[2] | Paul Wilson[1] | Micheline Manseau[2,3]

[1]Biology Department, Trent University, Peterborough, Ontario, Canada | [2]Landscape Science and Technology, Environment and Climate Change Canada, Ottawa, Ontario, Canada | [3]Environmental and Life Sciences Graduate Program, Trent University, Peterborough, Ontario, Canada

Correspondence: Micheline Manseau (micheline.manseau@ec.gc.ca)

Received: 27 April 2024 | **Revised:** 12 September 2024 | **Accepted:** 20 September 2024

Handling Editor: Jeremy B Yoder

Funding: This work was supported by the Genomic Applications Partnership Program of Genome Canada [OGI-194] and Environment and Climate Change Canada.

Keywords: amplicon sequencing | microhaplotype genotyping | quality control | Seq2Sat

ABSTRACT

Microhaplotypes are small linked genomic regions comprising two or more single-nucleotide polymorphisms (SNPs) that are being applied in forensics and are emerging in wildlife monitoring studies and genomic epidemiology. Typically, targeted in non-coding regions, microhaplotypes in exonic regions can be designed with larger amplicons to capture functional non-synonymous sites and minimise insertion/deletion (indel) polymorphisms. Quality control is an important first step for high-confidence genotyping to counteract such false-positive variants. As genetic markers with higher polymorphism compared to biallelic SNPs, it is critical to ensure sequencing errors across the microhaplotype amplicon are filtered out to avoid introducing false-haplotypes. We developed the MhGeneS pipeline which works in tandem with Seq2Sat to help validate microhaplotype genotyping of the coding region of genes, with broader applicability to any microhaplotype profiling. We genotyped microhaplotype regions of the *Zfx* (≅160 bp) and *Zfy* (≅140 bp) genes, as well as an exon of the prion protein (*Prnp*) gene (≅370 bp) in caribou (*Rangifer tarandus*) using paired-end Illumina technology. As important quality metrics affecting microhaplotype calling, we identified the sequencing error rate profile related to the overlap or non-overlap of paired-end reads as well as the read depth as significant. In the case of *Prnp*, we achieved confident microhaplotype calling through MhGeneS by removing small sections of the 5′ and 3′ amplicons and using a minimum read depth of 20. Read depth and sequence trimming may be locus-specific, and validation of these parameters is recommended before the high-throughput profiling of samples.

1 | Introduction

Beyond generating whole-genome coverage, next-generation sequencing has facilitated characterising amplicons that can detect single-nucleotide polymorphisms (SNPs) through GT-Seq (e.g., Campbell, Harmon, and Narum 2015; Hayward et al. 2022; Natesh et al. 2019), microsatellites (e.g., Bradbury et al. 2018; Marcy-Quay et al. 2023; Pimentel et al. 2018; Vartia et al. 2016), and multiple SNPs as microhaplotypes (e.g., Baetscher et al. 2018; Delomas et al. 2023; Jones et al. 2009). These markers harness the depth of sequencing associated with genomic characterisation to target variable loci and increase sample throughput for a range of DNA quantity and quality (Acford-Palmer et al. 2023; Baetscher et al. 2018; Eriksson, Ruprecht, and Levi 2020; Meek and Larson 2019).

SNPs express a relatively low error rate; however, in diploid species, the expressed variation (four DNA bases per allele)

is lower than the per single-locus power demonstrated by microsatellite markers, resulting in less information per locus (Delomas et al. 2023; Jones et al. 2009). Alternatively, microhaplotype (MH) loci comprise two or more closely genetically linked SNPs with negligible recombination rates (Kidd et al. 2017) and are an emerging marker type in forensics. Typically, they are relatively small DNA fragments (< 200–300 bp) and combine the higher accuracy and repeatability of SNP markers with the higher per locus information of microsatellites (e.g., allelic diversity and heterozygosity; Baetscher et al. 2018; Delomas et al. 2023; Kidd et al. 2014; Oldoni, Kidd, and Podini 2019). Often, MH loci are screened encompassing two to four SNPs (Kidd and Pakstis 2022; Oldoni, Kidd, and Podini 2019). However, recent studies have screened MH loci with more SNPs, ranging anywhere from 3 to 12 linked SNPs (e.g., Fan et al. 2022; Gandotra et al. 2020). In human forensics, they have proven to be very efficient for individual assignment, mixture detection and pedigree reconstruction (Kidd et al. 2014; Kidd and Pakstis 2022; Oldoni, Kidd, and Podini 2019). In recent years, other research fields have discovered the benefits of microhaplotype markers for analyses such as genomic epidemiology (e.g., malaria (LaVerriere et al. 2022)), pedigree analyses in fisheries (Baetscher et al. 2018; Delomas and Campbell 2022; McKinney et al. 2020) and wildlife monitoring (Delomas et al. 2023).

Outside of standard operating procedures in human forensics, the uptake of microhaplotypes in a conservation genetic and molecular ecology context would significantly benefit from standardised approaches on processing raw sequencing reads from amplicons, filtering techniques, quality control measures and scoring. This is particularly true for MHs in genic regions where the objective is to capture informative coding regions for functional non-synonymous changes and associated synonymous changes to maximise haplotypic diversity. As a result, genic amplicon sizes can be beyond the standard for non-coding MHs (e.g., 75 bp in wolves (*Canis lupus*); Delomas et al. 2023).

A critical step of sequence-based amplicon typing requires that sequence artefacts be culled from the scoring (e.g., microsatellite sequencing; Liu et al. 2024), so understanding the baseline sequencing errors per locus in a range of species will facilitate more automated scoring of MHs through a software environment. As the result of the need for quality assurance and standardisation, we developed an analytical pipeline, MhGeneS (Microhaplotype Gene Screening), to optimise microhaplotype profiling of gene regions with broader applicability to non-coding microhaplotypes. As a case study, we applied this analytical pipeline to amplicon sequencing based on multiple genic regions of varying lengths in boreal caribou (*Rangifer tarandus caribou*). We used sex chromosome-specific amplicons, comprising a small fraction of *Zfx* and *Zfy* genes (161 and 142 bp, respectively; Ball et al. 2007) and a region of the prion protein (*Prnp*) gene (targeting a minimum of eight linked SNPs in 367 bp of exon 3). The *Prnp* gene in caribou (and other cervid species) is involved in mechanisms connected to the highly contagious and fatal prion disease, chronic wasting disease (CWD) (Arifin et al. 2020; Cheng et al. 2017; Escobar et al. 2020; Haley and Hoover 2015; Moazami-Goudarzi et al. 2021). Exon 3 contains a coding sequence associated with CWD susceptibility (Haley and Hoover 2015).

The DNA source for profiling these genes in caribou was non-invasively collected faecal samples. Faecal DNA can pose challenges for laboratory procedures affected by poor DNA quantity and quality. However, winter-collected caribou faecal pellets have been shown to produce high-quality extracts (Ball et al. 2007)—indeed full genomes have been sequenced from faecal DNA (Taylor et al. 2021). However, more PCR cycles during the profiling of non-invasive DNA sources may be required that can lead to relatively more errors within the read depth in the sequencing process (Eriksson, Ruprecht, and Levi 2020). These errors can cause problems with downstream analyses, and therefore evaluating those error rates (Pfeiffer et al. 2018) and thoughtful quality filtering are required (Bewicke-Copley et al. 2019; Bokulich et al. 2013; Puente-Sánchez, Aguirre, and Parro 2016). An important quality metric is the depth of sequencing reads. For samples with low sequencing depth, the information of one haplotype may be missing (allelic dropout), resulting in false homozygous genotype calling (Bilton et al. 2018; McKinney et al. 2020; O'Leary et al. 2018). With a higher depth of sequencing data, more reads can be used for calling a particular MH, which increases the certainty and decreases the genotyping errors in a dataset (O'Leary et al. 2018). Read filtering among different research fields often aims for a minimum of 10 reads per locus, including a recent wildlife study (Delomas et al. 2023) and genomic epidemiology (Kubik et al. 2021; LaVerriere et al. 2022). However, these assays are based on smaller MH regions that are more amenable to different tissue sources of DNA.

Within our MhGeneS pipeline, we first summarised the sequencing error rate profile based on individual error rates for each position within the amplicon sequences. We then evaluated the dependency of the error rate profile on the length of the amplicon sequence and determined if trimming the amplicon sequences was an appropriate measure to improve the overall quality. We then assessed the depth-of-read for each sample with the number of microhaplotypes screened and identified potential artefacts in the form of rare microhaplotypes associated with low read numbers. We examined different levels of read depth filtering to demonstrate the importance of such quality filtering on the accuracy of the genotype scores. Finally, we recommend trimming and filtering strategies that can act as guidelines for automated profiling of genic microhaplotypes, which also apply to microhaplotypes within non-coding regions.

2 | Materials and Methods

2.1 | Caribou Sampling and Amplicon Sequencing

Between 2010 and 2023, 3871 faecal samples from boreal caribou (*Rangifer tarandus caribou*) were collected during winter surveys using a protocol developed by Hettinga et al. (2012). Faecal pellets were placed in sterile bags, labelled and kept frozen at −20°C during shipment for genetic analysis. DNA processing and analyses were conducted at Trent University.

DNA extraction was done with the Qiagen DNAeasy tissue kit, following the manufacturer's protocol. For comparison, we used relatively smaller sex chromosome-specific amplicons that comprised small regions within the *Zfx* (161 bp region) and *Zfy* (142 bp region) genes (Ball et al. 2007) and the

relatively large prion protein (*Prnp*) gene region in the caribou genome. For *Prnp*, we designed specific primers covering a 367 bp region in exon 3 of the prion protein gene. Both the forward 5′-CGCTGGAGCAGTGGTAGG-3′ and reverse 5′-TCCTACTATGAGAAAAATGAGGAA-3′ primers were designed to capture eight polymorphisms (Table S1), which have been identified as potential susceptibility factors for CWD in caribou (Arifin et al. 2020). Amplicon sequencing libraries were prepared following the protocol described by Liu et al. (2024) and were sequenced on the Illumina MiSeq platform. In summary, a multiplex polymerase chain reaction (PCR) was performed with the designed primers (*Zfx*, *Zfy* and *Prnp*-specific) using the QIAGEN Multiplex PCR Plus Kit (Qiagen, Hilden, German) and 1 ng of sample DNA. After an initial cleaning and purification step with 1:0.8 of AMPure XP beads (Beckman Coulter, Inc., United States), reactions were indexed with IDT for Illumina Nextera DNA Unique Dual Index Sets A–D (Integrated DNA Technologies, Inc., United States) and pooled (384 samples per sequencing library/run). Finalised libraries were quantified using Qubit and then diluted to 9.5 pM and 250 bp pair ends sequenced on an Illumina MiSeq machine.

2.2 | Genotyping and Microhaplotype Screening

Sequencing reads were demultiplexed with the 'generate FASTQ analysis module' by Illumina. Based on the indexes supplied in the library preparation, raw reads were assigned to each sample and FASTQ files were generated.

The MhGeneS pipeline (Figure 1) uses genotyping results from Seq2Sat, updated software to support accurate automated microsatellite (Liu et al. 2024) and microhaplotype screening (https://github.com/ecogenomicscanada/Seq2Sat). In brief, Seq2Sat (V.2.0.0.2) reads raw FASTQ reads of amplicon sequences into read packs (1000 reads/pack). Each read pack is processed independently and in parallel. An initial read quality check is performed on raw reads by removing too short and low-quality reads. Sequencing adaptors are also automatically detected and removed. Overlapping regions of paired-end (PE) reads are identified and analysed for sequencing error correction based on the base quality score. Non-overlapping regions remain with the original base quality score (uncorrected). A minimum of 30 bp overlapping between the forward and reverse strands is required to merge the two PE reads into a clean single-end (SE) read. The clean SE reads are then demultiplexed by assessing the edit distances between the primer pair and the reads (allowing two mismatches for each primer) using the Edlib library (Šošic and Šikic 2017). To reduce the possible inflation of a number of microhaplotype loci caused by sequencing errors in primer binding sites, primer information is given in a separate parameter file and primer sequences are subsequently trimmed from the clean reads (Figure 1 (1)).

Genotyping was also performed in Seq2Sat (V.2.0.0.2). To obtain accurate genotyping for the *Prnp* amplicon, we used algorithms to identify the primary main alleles and their proportion of reads. These are used to determine the microhaplotype locus' status as homozygous, heterozygous or inconclusive. First, each read variant is aligned against the *Prnp* gene reference (Table S1) to identify SNPs/artefacts using Edlib (Šošic and Šikic 2017). To determine the primary alleles, the two most abundant read variants are identified based on their read count, followed by the calculation of their relative proportion of haplotype reads. The sequence difference between the two most abundant read variants was then assessed by sequence alignment. A direct pairwise sequence alignment was done with Edlib (Šošic and Šikic 2017), to compare the two main read variants in their similarity in terms of mutations (SNPs). Then, the proportion of reads of the two sequences was used to determine the locus' zygosity based on threshold. If the two most common read variants have a proportional number of reads between 0.50 and 0.65, they are considered as heterozygous, whereas a value above 0.85 for one haplotype is considered homozygous and a read ratio between 0.65 and 0.85 is considered inconclusive requiring manual assessment. These thresholds constitute a relatively conservative approach (Nielsen et al. 2011) where a homozygous genotype is called when the proportion of the non-reference allele is above 80%. In the MhGeneS pipeline, this would translate in a sample being homozygous when the proportional number of reads between the two most abundant read variants is above 0.8. We opted for an even more conservative approach and added an inconclusive category which comprised samples with a rather uncertain genotype. These thresholds can be set with the htJetter, hmPerH and hmPerL parameters in Seq2Sat (V.2.0.0.2). We note here that the user can modify these parameters when running Seq2Sat.

For sex chromosome-specific amplicons (*Zfx* and *Zfy*; Ball et al. 2007), Seq2Sat (Liu et al. 2024) is similar to the microhaplotype identification for the *Prnp* amplicon and can identify the putative sex of each sample based on the proportional reads of the two most common MH variants. An initial depth-of-read filter of 10 reads was applied within Seq2Sat, and the *Zfx* marker was used to benchmark sex identification. If no read variant of the *Zfx* amplicon was determined or the two most common read variants had fewer than 10 reads, the sample was classified as inconclusive. If only the read variants of the *Zfx* amplicon could be determined or if the proportional reads between the most abundant *Zfy* and *Zfx* read variants were lower than 0.001, the samples were classified as female, and only one read variant was considered as the present microhaplotype. If the proportional reads of the *Zfy* to the *Zfx* marker were greater than 0.001, the sample was classified as male, and both read variants (of the *Zfy* and *Zfx* markers) were considered as the present microhaplotypes in this sample.

Sequence error rates for each position can be calculated to obtain an overview of the profile across the amplicon sequence. For each conclusive sample, every read variant identified that is independent of the number of reads for these variants is used in the calculation to assess the error distribution along the amplicons. The frequency of every position's base (A, C, T and G) from all identified read variants at a locus (*Zfx*, *Zfy* and *Prnp*) is recorded, and the error rate of each position is calculated by dividing the base frequency (different from the (two) main read variants at this position) by the total number of reads (see Table S2 for an example). The average error rate across the amplicon sequence for each sample is then recorded and can serve as a proxy for the sequencing performance for each sample.

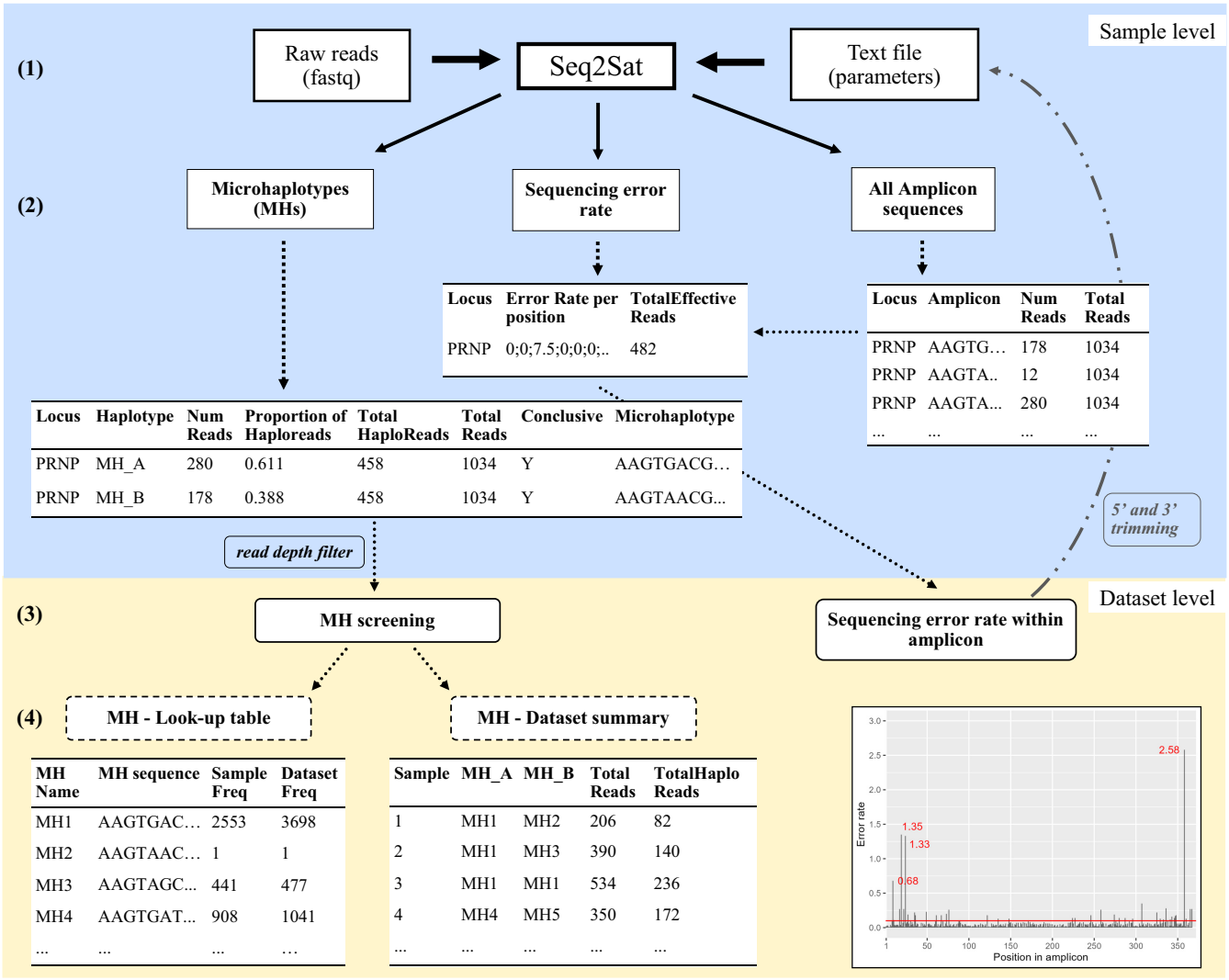

FIGURE 1 | Schematic of the MhGeneS pipeline (with the Prnp amplicon as an example). (1) Genotyping within Seq2Sat (Liu et al. 2024). (2) Output files from Seq2Sat used directly (microhaplotypes and sequencing error rate information) and indirectly (all amplicon sequences) in the MhGeneS pipeline and how information is shown. (3) The main steps within the MHGeneS pipeline: Microhaplotype screening and exploration of sequencing error rates within the amplicons. Quality control measures such as read depth filter or 5′ and 3′ trimming can be applied in Step 3. (4) Output from MhGeneS (GITHUB path to *R* script) comprising lookup and dataset summaries of microhaplotypes screened and visualised sequencing error rates.

Seq2Sat produces various output files (https://github.com/ecogenomicscanada/Seq2Sat), comprising genotype, microhaplotype and sex identification tables, sequencing error rate information as well as an html report (Figure S1 for an example), with graphs showing the distribution of microhaplotypes, error rates and plots on quality metrics of raw, as well as filtered, reads (see https://github.com/ecogenomicscanada/MhGeneS). Within the MhGeneS pipeline, the information on microhaplotypes and sequencing error rates produced by Seq2Sat are further processed using custom functions in *R* version 4.2.2, specifically developed for error rate exploration and microhaplotype screening (all *R* code and example files are available on https://github.com/ecogenomicscanada/MhGeneS) (Figure 1 (2)). We first summarised the mean error rates across the amplicon sequences for the entire dataset, based on the sequencing error rate per position and sample. We also reported on the average error rate across the entire amplicon sequence per sample and summarised this across the dataset (Figure S2). Finally, we applied an optional trimming step for the amplicon sequence (Figure 1 (3)). When longer amplicons (>250 bp) are sequenced paired-end on an Illumina platform, the reads do not fully overlap, leading to lower sequencing quality in the 5′ and 3′ end. Therefore, trimming those ends can minimise sequencing errors in the amplicon. When the trimming step is applied, the genotyping in Seq2Sat must be repeated with an edited parameter file with information on the number of base pairs to trim off the 5′ and 3′ ends of the sequence (Figure 1 (1)).

To screen for microhaplotypes (MHs) within the entire dataset (Figure 1 (3)), we first excluded samples with microhaplotypes considered inconclusive from previous analyses and testing. We identified the unique MHs in the dataset and generated a microhaplotype lookup table. This table summarises the present microhaplotypes and shows their frequency within the dataset (Figure 1 (4)). We generated a second dataset summary table, providing information on which MHs are assigned to each sample

and the number of reads assigned to each sample (Figure 1 (4)). For the two sex chromosome-specific markers (*Zfx* and *Zfy*), we screened for microhaplotypes separately and created two sets of tables: "lookup" and "data summary" (*Zfx*- and *Zfy*-specific). In this post-processing step, we applied different stringencies of depth-of-read filter: no filter, a minimum of 10, 20 and 50 reads. Filtering for a certain depth-of-read improves confidence in the called genotypes and controls for sequencing errors in a dataset.

To demonstrate the effect of trimming and depth-of-read filter on the quality of our data, we compared the sequencing error rate profile (Figure 1 (3)) among different trimming strategies, as well as the microhaplotype lookup table (Figure 1 (4)) for the different depth-of-read filters applied. We also explored the effect of those read filters on the number of microhaplotype loci and samples remaining in the dataset, as well as the frequency of rare microhaplotypes (here only present in 1%–5% of the samples) and ultra-rare microhaplotypes (here present in less than 1% of the samples). Rare microhaplotypes in a dataset can be considered spurious, since they are only present in few samples, most likely indicating false-positive genotyping, and were treated with caution.

3 | Results and Discussion

We successfully genotyped 3487 and 3445 of the 3871 samples for the sex chromosome-specific amplicons (*Zfx*, *Zfy*) and the *Prnp* amplicon with Seq2Sat. For each successfully genotyped sample, Seq2Sat produced several output files (information on genotypes, microhaplotypes, sex-specific markers and sequencing error rate) and an html report (https://github.com/ecogenomicscanada/MhGeneS). Even though amplicon sequencing can achieve deep coverage, even in faecal samples (Eriksson, Ruprecht, and Levi 2020), we observed low sequencing depth in some of our samples for the targeted *Prnp* gene. The overall depth of read was higher for sex-specific amplicons (Figure 2).

3.1 | Trimming Based on the Sequencing Error Rate Profile

We observed varying patterns depending on the length of the amplicon sequence in the sequencing error rate profile (Figure 3). For the two relatively short sex-specific amplicons (*Zfx* and *Zfy*, 161 bp and 142 bp, respectively), sequencing errors were distributed randomly across their length, with slightly higher error rates for the *Zfy* marker around 20–35 bp within the 142 bp long amplicon sequence (Figure 3A). Due to the observed error rate profile of both sex-specific markers, no trimming step was applied. Using a 250 bp paired-end approach led to fully sequencing those two amplicons (142 bp and 161 bp) with a complete overlap. Trimming at the 5′ and 3′ ends would have decreased those smaller amplicons (142 bp and 161 bp) without significantly improving the overall error rate profile. No trimming strategies were assessed further. Ultimately, sex identification using *Zfx* and *Zfy* were based on the presence/absence of *Zfy* and not on sequence diagnostics.

In contrast, the *Prnp* amplicon was longer (367 bp) than required to be sequenced with full overlap, leading to a very different sequencing error rate profile. Due to its length, there was only a partial overlap which likely explains the elevated error rates at the non-overlapping 5′ and 3′ ends (Figure 3B); Illumina chemistry increased sequencing-derived errors at the end of the sequence chemistry (Schirmer et al. 2016). We tested different trimming strategies at the 5′ and 3′ ends to improve the error rate profile: (1) trimming 10 bp from each end (5′ and 3′) or (2) trimming 24 bp from the 5′ end and 10 bp from the 3′ end (tailored trimming). Applying both trimming strategies to the *Prnp* amplicon sequence improved the sequencing error rate profile (Figure 3C,D). Tailored trimming decreased the *Prnp* amplicon to some extent (from 367 bp to 333 bp); however, it was very efficient in removing the high error rates at the 5′ end. With testing different trimming approaches, we wanted to demonstrate the importance of investigating first the sequencing error rate profile of the amplicon. Trimming a sequence always leads to shortening it, and this decision should be case-specific. A really short amplicon with condensed information requires a different trimming strategy or even no trimming compared to a very long amplicon such as the *Prnp* here. By exploring the error rate profile as a quality control step, unnecessary trimming can also be prevented.

3.2 | Effect of Depth-of-Read Filtering on Microhaplotype Screening

Read depth filtering is often applied to different kinds of sequencing data as a first step of quality control (often with a threshold of 10 reads Delomas et al. 2023; Kubik et al. 2021; LaVerriere et al. 2022). Here, we tested different stringencies of depth-of-read filters: no filter, excluding samples with less than 10, 20 and 50 reads. We explored how these different thresholds affected the microhaplotype screening process (Figure 1 (3)) and especially the change in the number of unique and rare or ultra-rare microhaplotypes we identified. We found that the stringency of a depth-of-read filter affected the sample size as well as the number of identified microhaplotypes (Tables 1 and 2). An indicator of spurious microhaplotypes was the number of rare or ultra-rare microhaplotypes. Those were only present in less than 5% of the samples and therefore unlikely to be true microhaplotypes. By applying various depth-of-read filters, we managed to reduce the number of rare microhaplotypes, leading to a more confident list of unique microhaplotypes (Tables S3–S6).

The two sex-specific amplicons (*Zfx* and *Zfy*) were screened for microhaplotypes separately, and we created a lookup and dataset summary table for each marker (Figure 1 (4)). In the first step, inconclusive samples were excluded, and only those with a putative sex (with (two) common read variants) were included in the MH screening (Figure 1 (3)). A total of 3567 samples were screened for the *Zfx* marker and 1395 samples for the *Zfy* marker (Table 1). We found 13 unique *Zfx* microhaplotypes and 15 unique *Zfy* microhaplotypes, although only one MH for each marker was the most common in each dataset (Table S3). The number of samples did not change meaningfully with any of the depth-of-read filters applied (10, 20 or 50), but the number of microhaplotypes did decrease with the more stringent filters. In the *Zfx* dataset, the number of microhaplotypes was 13–10, 6 and 5 with no filters, a depth-of-read

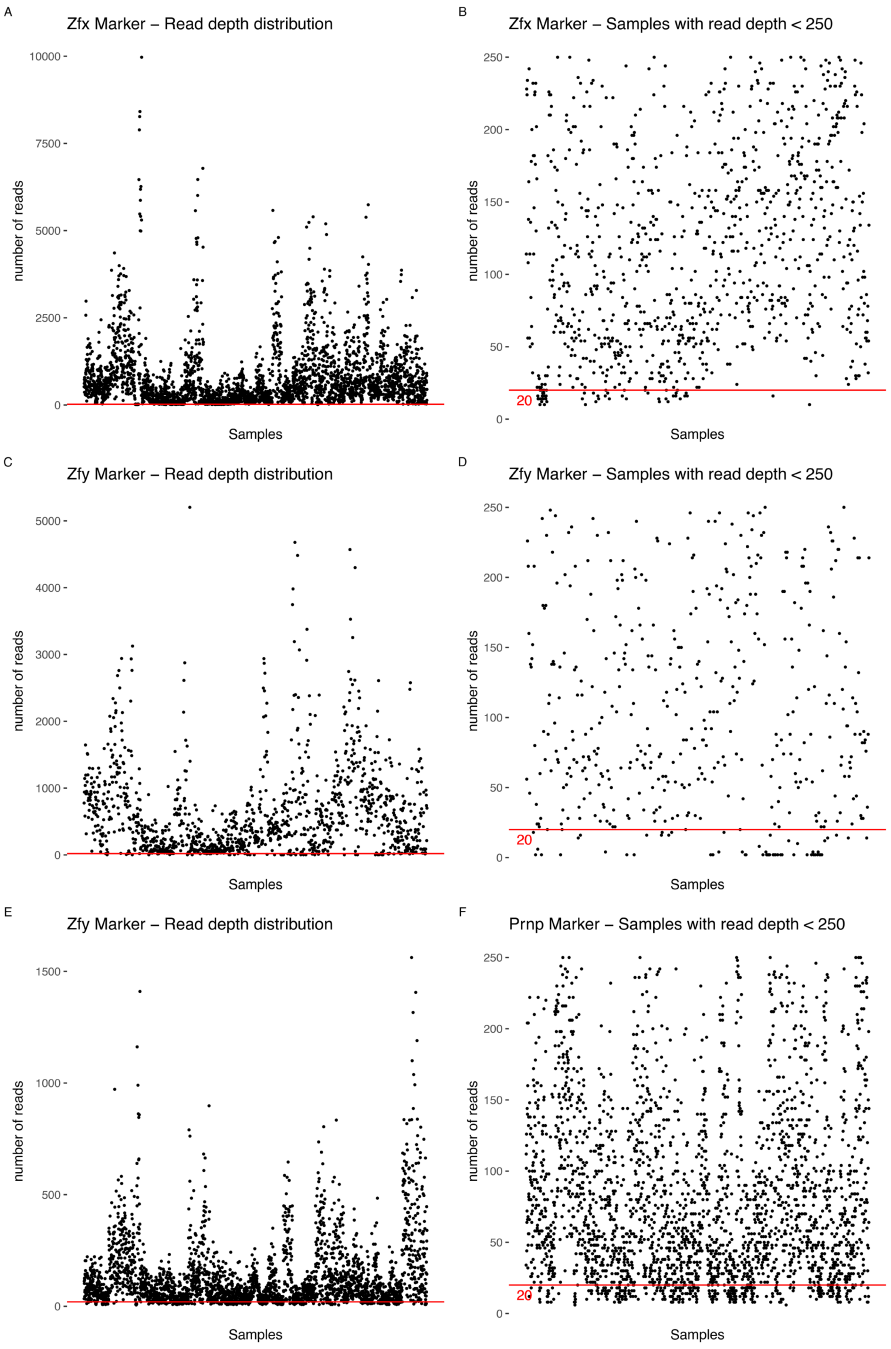

FIGURE 2 | Read depth distribution of the *Zfx* (A and B), *Zfy* (C and D) and *Prnp* (E and F) datasets. Plots on the left-hand side (A, C and E) show the read depth distribution within the entire datasets. Plots on the right-hand side only show samples with a read depth lower than 250 reads. The red line indicates the read depth filtering threshold of 20 reads.

filter of 10, 20, and 50, respectively. In the *Zfy* dataset, we only recorded a decline in unique microhaplotypes with depth-of-read filters for 20 and 50 reads to 10 and 3 unique MHs, respectively (Table 1; Tables S3–S6). In a dataset of thousands of samples, the likelihood of a microhaplotype detected in a small number of samples (< 5% of the samples) representing real biological variation is very small. To confirm if a rare or ultra-rare microhaplotype (here classified as spurious) is a true microhaplotype, re-sequencing at a deeper level is necessary. Another approach could be to add more samples from the sampled geographic origin to gain a potential bigger sample size of the rare or ultra-rare microhaplotype.

Therefore, the presence of rare MHs can indicate genotyping errors in a dataset. Reducing this number by applying filtering strategies is one way to control genotyping errors. For both sex-specific amplicons, we did not screen for MHs; however, all screened MHs, except for the most common, were considered ultra-rare and were present in less than 1% of samples. Applying depth-of-read filters of 10, 20 and 50 reduced the number of ultra-rare microhaplotypes to: 9, 5 and 4 for the *Zfx* marker and 14, 9 and 2 for the *Zfy* marker (Table 1; Tables S3–S6). After screening both markers for unique microhaplotypes, dataset summary tables (Figure 1 (4)) of *Zfx* and *Zfy* markers were merged to summarise the putative sex and the identified microhaplotypes for each sample (Tables S3–S6).

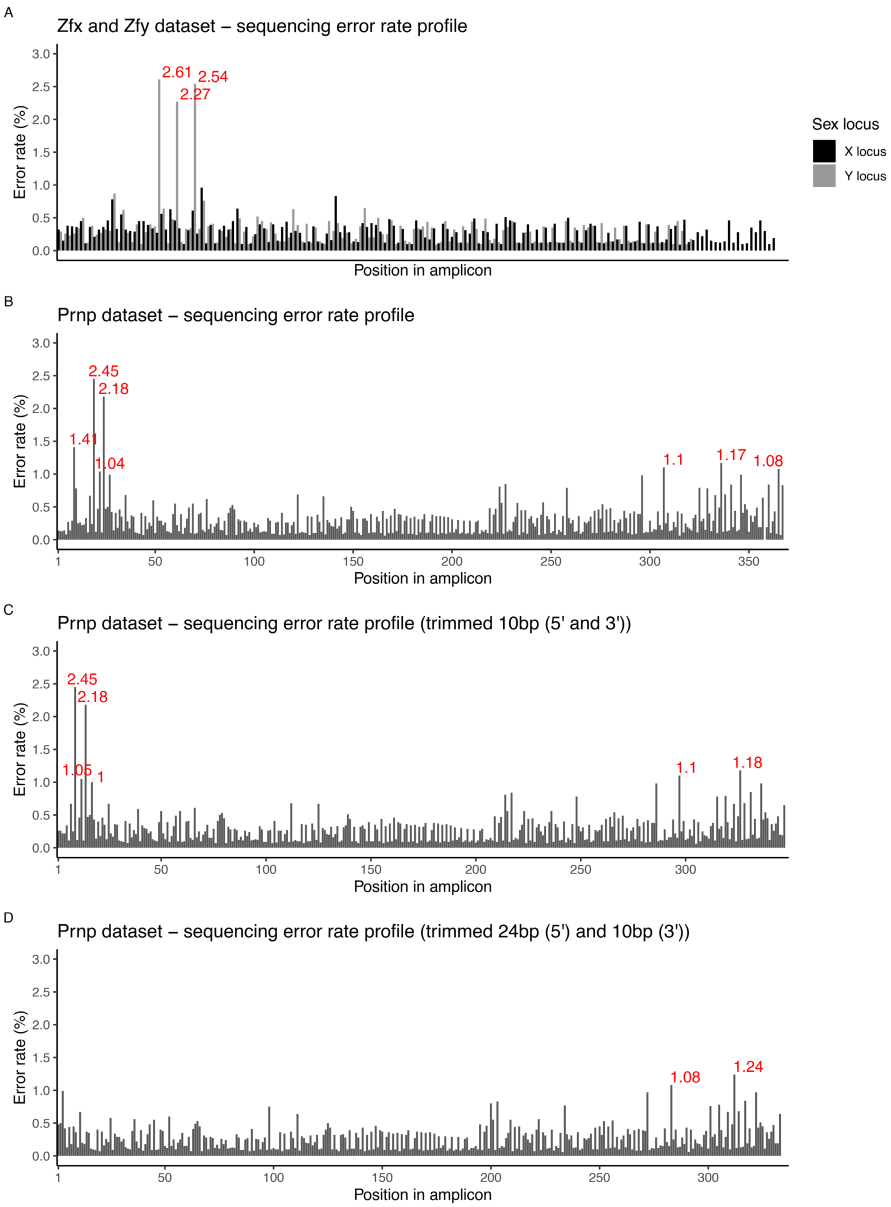

FIGURE 3 | Sequencing error rate profile for each dataset analysed: The Zfx and the Zfy marker (A), the untrimmed Prnp marker (B), the trimmed (10 bp off each 5′ and 3′ end) Prnp marker (C) and the trimmed (24 bp from the 5′ end and 10 bp from the 3′ end) Prnp marker (D).

After excluding inconclusive samples in the *Prnp* dataset, we processed 2821 samples during MH screening. We found 24 unique MHs when screening for microhaplotypes (Figure 1 (3)), 18 of which were only detected in a small number of samples (labelled as rare and ultra-rare microhaplotypes). The depth-of-read filters of 10 did not affect the number of samples or microhaplotypes in the dataset. When we applied more stringent depth-of-read filters (20 and 50), we reduced the number of samples from 2686 to 2101 and unique microhaplotypes from 21 to 11 (Table 1; Tables S7–S10). The number of ultra-rare microhaplotypes was also reduced to 14 and 3. Independent of which read depth filter was applied, the same six microhaplotypes (MH1–MH6; Tables S7–S10) were the most common in each dataset. Additionally, one microhaplotype (MH7; Tables S7–S10) was classified as rare since it occurred in 1%–2% of the samples, depending on the depth-of-read filter (e.g., 44 samples, when no depth-of-read filter was applied). The number of reads for samples with MH7 ranges between 44 and 898 reads, which may indicate that this microhaplotype represents real variation within the dataset (Table S11). A next step would be to re-sequence those samples to gain more certainty on the validity of their microhaplotype assignment.

3.3 | Combined Effect of Read Trimming and Depth-of-Read Filtering on the *Prnp* Amplicon

Trimming positions with significantly higher sequencing error rates of the *Prnp* amplicon improved the overall genotyping performance of Seq2Sat. More samples were retrieved and were thus genotyped as heterozygous or homozygous rather than inconclusive. Within the untrimmed *Prnp* dataset, we screened 2821 conclusive samples. With the two trimming

TABLE 1 | Read depth filter comparison for all tested datasets (Zfx, Zfy and Prnp).

	Zfx dataset				*Zfy* dataset				*Prnp* dataset			
Depth-of-read filter	**Samples**	**Unique MHs**	**Rare MHs**	**Ultra-rare MHs**	**Samples**	**Unique MHs**	**Rare MHs**	**Ultra-rare MHs**	**Samples**	**Unique MHs**	**Rare MHs**	**Ultra-rare MHs**
No filter	3567	13	0	12	1395	15	0	14	2821	24	1	17
> 10 reads	3564	10	0	9	1395	15	0	14	2821	24	1	17
> 20 reads	3460	6	0	5	1346	10	0	9	2686	21	1	14
> 50 reads	3280	5	0	4	1257	3	0	2	2101	11	2	3

Note: Various read depth filters were tested: No filter, 10, 20 and 50 read depth filters. Number of samples, unique microhaplotypes (MHs), rare microhaplotypes (occurring in 1%–5% of the samples) and ultra-rare microhaplotypes (occurring in less than 1% of the samples) are shown.

strategies, however, trimming 10 bp from each end (5′ and 3′) and trimming 24 bp from the 5′ end and 10 bp from the 3′ end (tailored trimming), we screened 2883 and 3048 samples, respectively (Table 2). When removing highly erroneous positions from sequenced amplicon reads, read variants differed in fewer sites. This resulted in a reduction of rare and ultra-rare microhaplotypes (Table 2; no filter) that are only present in a handful of samples due to sequencing errors in non-overlapping parts of the amplicon sequence (5′ and 3′ ends). Therefore, read variants with differing sites are less frequent in a sample, leading to more favourable proportional reads. As described above, if the number of proportional reads is less than 0.65 or greater than 0.85, a sample is assigned a heterozygous or homozygous genotype, respectively.

After trimming 10 bp from the 5′ and 3′ ends, we again excluded the inconclusive samples in the *Prnp* dataset and applied 10, 20 and 50 depth-of-read filters on the dataset of 2883 samples (Table 2; *Prnp-10 bp at 5′ and 3′* ends). As with the untrimmed *Prnp* dataset, a depth-of-read filter of 10 did not change the number of samples nor microhaplotypes. A filter of 20 or 50, however, led to a reduction in sample size (2768 and 2242) and unique MHs (15 and 9). We detected 10 and 4 rare or ultra-rare microhaplotypes after filtering out reads below 20 or 50. Independent of the depth-of-read filter applied, the same common five microhaplotypes were found in each dataset trimmed by 10 bp at the 5′ and 3′ ends (MH1–MH5; Tables S12–S15).

After applying a tailored trimming step (24 bp at 5′ and 10 bp at 3′ ends), we processed 3048 samples during MH screening (Figure 1 (3)). While screening these samples for microhaplotypes, we found 16 unique MHs, 12 of which were only detected in a small number of samples (labelled as rare and ultra-rare microhaplotypes). A depth-of-read filter of 10 did not decrease the number of samples or microhaplotypes in the dataset. When we applied more stringent depth-of-read filters (20 and 50), however, we reduced the number of samples and unique microhaplotypes (2945 and 2415; and 14 and 9, respectively; Table 2; *Prnp-24 bp at 5′ and 10 bp at 3′* ends). The number of rare and ultra-rare microhaplotypes was reduced to 9 and 4 MHs. In the tailored trimmed dataset, we also found the same five common microhaplotypes among the screened samples (MH1–MH5; Tables S16–S19).

3.4 | Implementations of the MhGeneS Pipeline for Microhaplotype Screening

Sequencing errors are inevitable, and stringent quality control measures need to be applied to ensure accurate haplotype calling (Bewicke-Copley et al. 2019; Bokulich et al. 2013; Puente-Sánchez, Aguirre, and Parro 2016). Sequencing errors, as well as low read depth, can lead to insufficient data and cause problems in the genotyping process, i.e., false-positive variant calling and allelic dropout (Bilton et al. 2018; McKinney et al. 2020; O'Leary et al. 2018). Sequencing errors can also lead to more spurious microhaplotypes, resulting in higher uncertainty when calling genotypes (Baetscher et al. 2018; Sinha et al. 2017). An initial evaluation of the error rate within the amplicon is therefore recommended (Pfeiffer et al. 2018) before applying any trimming to a sequence or depth-of-read filtering.

TABLE 2 | Trimming strategy and read depth filter comparison for the Prnp dataset.

	Prnp—not trimmed				*Prnp*—10 bp at 5′ and 3′				*Prnp*—24 bp at 5′ and 10 bp at 3′			
Depth-of-read filter	Samples	Unique MHs	Rare MHs	Ultra-rare MHs	Samples	Unique MHs	Rare MHs	Ultra-rare MHs	Samples	Unique MHs	Rare MHs	Ultra-rare MHs
No filter	2821	24	1	17	2883	19	1	13	3048	16	1	11
> 10 reads	2821	24	1	17	2883	19	1	13	3048	16	1	11
> 20 reads	2686	21	1	14	2768	15	1	9	2945	14	1	8
> 50 reads	2101	11	2	3	2242	9	1	3	2415	9	1	3

Note: Three different trimming strategies were tested: No trimming, 10 bp at the 5′ and 3′ ends and tailored trimming (24 bp at the 5′ and 10 bp at the 3′ ends). Various read depth filters were tested: No filter, 10, 20 and 50 read depth filters. Number of samples, unique microhaplotypes (MHs), rare microhaplotypes (occurring in 1%–5% of the samples) and ultra-rare microhaplotypes (occurring in less than 1% of the samples) are shown.

Applying our new MhGeneS pipeline to three different amplicon sequences (sex-specific genes *Zfx* and *Zfy*, as well as the *Prnp* gene) varying in length, we demonstrate how exploring the sequencing error rate profile and potential read trimming (Figure 3), as well as the application of different depth-of-read filters, affect the number of unique and rare microhaplotypes identified (Tables 1 and 2). Reporting the number of spurious microhaplotypes (rare and ultra-rare) across the different filtering strategies is important to better understand the dataset structure. However, the frequency of those microhaplotypes does not indicate if they are real or artefacts. Within a dataset of thousands of samples, however, it is very unlikely that a microhaplotype only present in one sample is presenting real biological variation (e.g., Table S11; *MH_Y2*). If a microhaplotype is present in more than one sample, even with an acceptable read depth (e.g., Table S11; *MH7 and MH17 for Prnp*), depending on the depth-of-read filter applied, a microhaplotype can be flagged as rare. Ultimately, re-sequencing samples expressing those rare MHs (present in < 5% of the samples) at a higher depth for re-analysis will elucidate whether these MHs represent real biological variation.

For the three amplicons, we found the depth-of-read filter of 20 to be a reasonable compromise to reduce the number or rare microhaplotypes without excluding too many samples, with the above caveat that confirmatory re-sequencing is recommended. This was also reported in a recent microhaplotype study using tissue samples of kelp rockfish (Baetscher et al. 2018). We, however, highlight that the trade-off between the number of unique and rare or ultra-rare microhaplotypes and decisions on the reduction in sample size need to be evaluated carefully for each dataset.

Targeted amplicon sequencing can improve the cost and time efficiency of large-scale projects requiring processing and analysis of thousands of samples (Baetscher et al. 2018). With the new MhGeneS pipeline, we demonstrated how assessing data quality in the form of sequencing error rates and read depth filter can increase the power and confidence in downstream genotype calling based on genic regions. We emphasise that MhGeneS can also be applied for genotype calling within non-coding regions.

Author Contributions

M.M. and P.W. initiated and secured funding for the genotyping project. J.C.G., P.L. and S.K. developed the pipeline. P.L. modified Seq2Sat to fit the pipeline, and J.C.G. and S.K. wrote code in *R* for the analytical part of the pipeline. J.C.G., P.L., S.K., M.M. and P.W. helped improve the workflow of the pipeline. J.C.G. wrote the manuscript, and P.L., M.M. and P.W. helped to edit the manuscript. All authors approved the final version of this manuscript for publication.

Acknowledgements

We thank Austin Thompson and Bridget Redquest for extracting DNA and sequencing the caribou samples. We thank Broderick Crosby and Rebecca Taylor for helping to design the *Prnp* primers. We thank members of the Ecogenomics team (https://www.ecogenomicscanada.ca/) for insightful conversations. Samples were collected by the governments of British Columbia, Saskatchewan, Manitoba and Ontario, the Sahtu Renewable Resource Board, Parks Canada, the Canadian Wildlife

Service and Manitoba Hydro as part of ongoing monitoring work and research activities. Open Access funding provided by the Environment and Climate Change Canada library.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Metadata and raw sequencing files are all deposited under the BioProject accession number PRJNA1132949. The read sequences for the caribou dataset were deposited in the NCBI Sequence Read Archive (SRA) (split into four separate projects due to the larger number of files). R code and example files can be found on the MhGeneS GitHub page: https://github.com/ecogenomicscanada/MhGeneS.

Benefit Sharing Statement

Benefits from this research accrue from sharing the pipeline, relevant code and data as described above.

References

Acford-Palmer, H., J. E. Phelan, F. G. Tadesse, et al. 2023. "Identification of Two Insecticide Resistance Markers in Ethiopian Anopheles Stephensi Mosquitoes Using a Multiplex Amplicon Sequencing Assay." *Scientific Reports* 13, no. 1: 5612. https://doi.org/10.1038/s41598-023-32336-7.

Arifin, M. I., A. Staskevicius, S. Y. Shim, et al. 2020. "Large-Scale Prion Protein Genotyping in Canadian Caribou Populations and Potential Impact on Chronic Wasting Disease Susceptibility." *Molecular Ecology* 29, no. 20: 3830–3840. https://doi.org/10.1111/mec.15602.

Baetscher, D. S., A. J. Clemento, T. C. Ng, E. C. Anderson, and J. C. Garza. 2018. "Microhaplotypes Provide Increased Power From Short-Read DNA Sequences for Relationship Inference." *Molecular Ecology Resources* 18, no. 2: 296–305. https://doi.org/10.1111/1755-0998.12737.

Ball, M. C., R. Pither, M. Manseau, et al. 2007. "Characterization of Target Nuclear DNA From Faeces Reduces Technical Issues Associated With the Assumptions of Low-Quality and Quantity Template." *Conservation Genetics* 8, no. 3: 577–586. https://doi.org/10.1007/s10592-006-9193-y.

Bewicke-Copley, F., E. Arjun Kumar, G. Palladino, K. Korfi, and J. Wang. 2019. "Applications and Analysis of Targeted Genomic Sequencing in Cancer Studies." *Computational and Structural Biotechnology Journal* 17: 1348–1359. https://doi.org/10.1016/j.csbj.2019.10.004.

Bilton, T. P., J. C. McEwan, S. M. Clarke, et al. 2018. "Linkage Disequilibrium Estimation in Low Coverage High-Throughput Sequencing Data." *Genetics* 209, no. 2: 389–400. https://doi.org/10.1534/genetics.118.300831.

Bokulich, N. A., S. Subramanian, J. J. Faith, et al. 2013. "Quality-Filtering Vastly Improves Diversity Estimates From Illumina Amplicon Sequencing." *Nature Methods* 10, no. 1: 57–59. https://doi.org/10.1038/nmeth.2276.

Bradbury, I. R., B. F. Wringe, B. Watson, et al. 2018. "Genotyping-By-Sequencing of Genome-Wide Microsatellite Loci Reveals Fine-Scale Harvest Composition in a Coastal Atlantic Salmon Fishery." *Evolutionary Applications* 11, no. 6: 918–930. https://doi.org/10.1111/eva.12606.

Campbell, N. R., S. A. Harmon, and S. R. Narum. 2015. "Genotyping-In-Thousands by Sequencing (GT-Seq): A Cost Effective SNP Genotyping Method Based on Custom Amplicon Sequencing." *Molecular Ecology Resources* 15, no. 4: 855–867. https://doi.org/10.1111/1755-0998.12357.

Cheng, Y. C., M. Musiani, M. Cavedon, and S. Gilch. 2017. "High Prevalence of Prion Protein Genotype Associated With Resistance to Chronic Wasting Disease in One Alberta Woodland Caribou Population." *Prion* 11, no. 2: 136–142. https://doi.org/10.1080/19336896.2017.1300741.

Delomas, T. A., and M. R. Campbell. 2022. "Grandparent Inference From Genetic Data: The Potential for Parentage-Based Tagging Programs to Identify Offspring of Hatchery Strays." *North American Journal of Fisheries Management* 42, no. 1: 85–95. https://doi.org/10.1002/nafm.10714.

Delomas, T. A., J. Struthers, T. Hebdon, and M. R. Campbell. 2023. "Development of a Microhaplotype Panel to Inform Management of Gray Wolves." *Conservation Genetics Resources* 15, no. 1: 49–57. https://doi.org/10.1007/s12686-023-01301-x.

Eriksson, C. E., J. Ruprecht, and T. Levi. 2020. "More Affordable and Effective Noninvasive Single Nucleotide Polymorphism Genotyping Using High-Throughput Amplicon Sequencing." *Molecular Ecology Resources* 20, no. 6: 1505–1516. https://doi.org/10.1111/1755-0998.13208.

Escobar, L. E., S. Pritzkow, S. N. Winter, et al. 2020. "The Ecology of Chronic Wasting Disease in Wildlife." *Biological Reviews of the Cambridge Philosophical Society* 95, no. 2: 393–408. https://doi.org/10.1111/brv.12568.

Fan, H., Q. Xie, L. Wang, et al. 2022. "Microhaplotype and Y-SNP/STR (MY): A Novel MPS-Based System for Genotype Pattern Recognition in Two-Person DNA Mixtures." *Forensic Science International* 59: 102705. https://doi.org/10.1016/j.fsigen.2022.102705.

Gandotra, N., W. C. Speed, W. Qin, et al. 2020. "Validation of Novel Forensic DNA Markers Using Multiplex Microhaplotype Sequencing." *Forensic Science International: Genetics* 47: 102275. https://doi.org/10.1016/j.fsigen.2020.102275.

Haley, N. J., and E. A. Hoover. 2015. "Chronic Wasting Disease of Cervids: Current Knowledge and Future Perspectives." *Annual Review of Animal Biosciences* 3: 305–325. https://doi.org/10.1146/annurev-animal-022114-111001.

Hayward, K. M., R. B. G. Clemente-Carvalho, E. L. Jensen, et al. 2022. "Genotyping-In-Thousands by Sequencing (GT-Seq) of Noninvasive Faecal and Degraded Samples: A New Panel to Enable Ongoing Monitoring of Canadian Polar Bear Populations." *Molecular Ecology Resources* 22, no. 5: 1906–1918. https://doi.org/10.1111/1755-0998.13583.

Hettinga, P. N., A. N. Arnason, M. Manseau, D. Cross, K. Whaley, and P. J. Wilson. 2012. "Estimating Size and Trend of the North Interlake Woodland Caribou Population Using Fecal-DNA and Capture-Recapture Models." *Journal of Wildlife Management* 76, no. 6: 1153–1164. https://doi.org/10.1002/jwmg.380.

Jones, B., D. Walsh, L. Werner, and A. Fiumera. 2009. "Using Blocks of Linked Single Nucleotide Polymorphisms as Highly Polymorphic Genetic Markers for Parentage Analysis." *Molecular Ecology Resources* 9, no. 2: 487–497. https://doi.org/10.1111/j.1755-0998.2008.02444.x.

Kidd, K. K., and A. J. Pakstis. 2022. "State of the Art for Microhaplotypes." *Genes* 13, no. 8: 1322. https://doi.org/10.3390/genes13081322.

Kidd, K. K., A. J. Pakstis, W. C. Speed, et al. 2014. "Current Sequencing Technology Makes Microhaplotypes a Powerful New Type of Genetic Marker for Forensics." *Forensic Science International. Genetics* 12: 215–224. https://doi.org/10.1016/j.fsigen.2014.06.014.

Kidd, K. K., W. C. Speed, A. J. Pakstis, et al. 2017. "Evaluating 130 Microhaplotypes Across a Global Set of 83 Populations." *Forensic Science International. Genetics* 29: 29–37. https://doi.org/10.1016/j.fsigen.2017.03.014.

Kubik, S., A. C. Marques, X. Xing, et al. 2021. "Recommendations for Accurate Genotyping of SARS-CoV-2 Using Amplicon-Based Sequencing of Clinical Samples." *Clinical Microbiology and Infection* 27, no. 7: 1036.e1–1036.e8. https://doi.org/10.1016/j.cmi.2021.03.029.

LaVerriere, E., P. Schwabl, M. Carrasquilla, et al. 2022. "Design and Implementation of Multiplexed Amplicon Sequencing Panels to Serve Genomic Epidemiology of Infectious Disease: A Malaria Case Study." *Molecular Ecology Resources* 22, no. 6: 2285–2303. https://doi.org/10.1111/1755-0998.13622.

Liu, P., P. Wilson, B. Redquest, S. Keobouasone, and M. Manseau. 2024. "Seq2Sat and SatAnalyzer Toolkit: Towards Comprehensive Microsatellite Genotyping From Sequencing Data." *Molecular Ecology Resources* 24, no. 3: e13929. https://doi.org/10.1111/1755-0998.13929.

Marcy-Quay, B., C. C. Wilson, C. A. Osborne, and J. E. Marsden. 2023. "Optimization of an Amplicon Sequencing-Based Microsatellite Panel and Protocol for Stock Identification and Kinship Inference of Lake Trout (*Salvelinus namaycush*)." *Ecology and Evolution* 13, no. 4: e10020. https://doi.org/10.1002/ece3.10020.

McKinney, G. J., C. E. Pascal, W. D. Templin, et al. 2020. "Dense SNP Panels Resolve Closely Related Chinook Salmon Populations." *Canadian Journal of Fisheries and Aquatic Sciences* 77, no. 3: 451–461. https://doi.org/10.1139/cjfas-2019-0067.

Meek, M. H., and W. A. Larson. 2019. "The Future is Now: Amplicon Sequencing and Sequence Capture Usher in the Conservation Genomics Era." *Molecular Ecology Resources* 19, no. 4: 795–803. https://doi.org/10.1111/1755-0998.12998.

Moazami-Goudarzi, K., O. Andréoletti, J.-L. Vilotte, and V. Béringue. 2021. "Review on PRNP Genetics and Susceptibility to Chronic Wasting Disease of Cervidae." *Veterinary Research* 52, no. 1: 128. https://doi.org/10.1186/s13567-021-00993-z.

Natesh, M., R. W. Taylor, N. K. Truelove, et al. 2019. "Empowering Conservation Practice With Efficient and Economical Genotyping From Poor Quality Samples." *Methods in Ecology and Evolution* 10, no. 6: 853–859. https://doi.org/10.1111/2041-210X.13173.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. "Genotype and SNP Calling From Next-Generation Sequencing Data." *Nature Reviews. Genetics* 12, no. 6: 443–451. https://doi.org/10.1038/nrg2986.

Oldoni, F., K. K. Kidd, and D. Podini. 2019. "Microhaplotypes in Forensic Genetics." *Forensic Science International. Genetics* 38: 54–69. https://doi.org/10.1016/j.fsigen.2018.09.009.

O'Leary, S. J., J. B. Puritz, S. C. Willis, C. M. Hollenbeck, and D. S. Portnoy. 2018. "These Aren't the Loci You'e Looking for: Principles of Effective SNP Filtering for Molecular Ecologists." *Molecular Ecology* 27, no. 16: 3193–3206. https://doi.org/10.1111/mec.14792.

Pfeiffer, F., C. Gröber, M. Blank, et al. 2018. "Systematic Evaluation of Error Rates and Causes in Short Samples in Next-Generation Sequencing." *Scientific Reports* 8, no. 1: 10950. https://doi.org/10.1038/s41598-018-29325-6.

Pimentel, J. S. M., A. O. Carmo, I. C. Rosse, et al. 2018. "High-Throughput Sequencing Strategy for Microsatellite Genotyping Using Neotropical Fish as a Model." *Frontiers in Genetics* 9: 73. https://doi.org/10.3389/fgene.2018.00073.

Puente-Sánchez, F., J. Aguirre, and V. Parro. 2016. "A Novel Conceptual Approach to Read-Filtering in High-Throughput Amplicon Sequencing Studies." *Nucleic Acids Research* 44, no. 4: e40. https://doi.org/10.1093/nar/gkv1113.

Schirmer, M., R. D'Amore, U. Z. Ijaz, N. Hall, and C. Quince. 2016. "Illumina Error Profiles: Resolving Fine-Scale Variation in Metagenomic Sequencing Data." *BMC Bioinformatics* 17, no. 1: 125. https://doi.org/10.1186/s12859-016-0976-y.

Sinha, R., G. Stanley, G. Gulati, et al. 2017. "Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples in Illumina HiSeq 4000 DNA Sequencin." https://doi.org/10.1101/125724.

Šošic, M., and M. Šikic. 2017. "Edlib: A C/C++ Library for Fast, Exact Sequence Alignment Using Edit Distance." *Bioinformatics* 33, no. 9: 1394–1395. https://doi.org/10.1093/bioinformatics/btw753.

Taylor, R. S., M. Manseau, B. Redquest, et al. 2021. "Whole Genome Sequences From Non-Invasively Collected Caribou Faecal Samples." *Conservation Genetics Resources* 14, no. 1: 53–68. https://doi.org/10.1007/s12686-021-01235-2.

Vartia, S., J. L. Villanueva-Cañas, J. Finarelli, et al. 2016. "A Novel Method of Microsatellite Genotyping-By-Sequencing Using Individual Combinatorial Barcoding." *Royal Society Open Science* 3, no. 1: 150565. https://doi.org/10.1098/rsos.150565.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.