Received: 15 September 2017 | Revised: 18 May 2018 | Accepted: 24 May 2018

DOI: 10.1111/1755-0998.12910

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

A bioinformatic pipeline for identifying informative SNP panels for parentage assignment from RADseq data

Kimberly R. Andrews[1] | Jennifer R. Adams[1] | E. Frances Cassirer[2] | Raina K. Plowright[3] | Colby Gardner[4] | Maggie Dwire[4] | Paul A. Hohenlohe[5] | Lisette P. Waits[1]

[1]Department of Fish and Wildlife Sciences, University of Idaho, Moscow, Idaho

[2]Idaho Department of Fish and Game, Lewiston, Idaho

[3]Department of Microbiology and Immunology, Montana State University, Bozeman, Montana

[4]U.S. Fish and Wildlife Service, Albuquerque, New Mexico

[5]Department of Biological Sciences, Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho

Correspondence
Kimberly R. Andrews, Genetics and Genomics Group, NOAA Pacific Marine Environmental Lab, University of Washington JISAO, Seattle, WA 98115, USA.
Email: kimandrews@gmail.com

Present address
Kimberly R. Andrews, Genetics and Genomics Group, NOAA Pacific Marine Environmental Lab, University of Washington JISAO, Seattle, Washington.

Funding information
U.S. Fish and Wildlife Service; Oregon Department of Fish and Wildlife; Morris Animal Foundation, Grant/Award Number: D13ZO-081; Montana University System Research Initiative, Grant/Award Number: 51040-MUSRI2015-03; National Institutes of Health (NIH), Grant/Award Number: P20GM103474, P30GM110732, P30GM103324; University of Idaho College of Natural Resources; University of Idaho IBEST; National Science Foundation (NSF), Grant/Award Number: DEB-1316549; Mexican Wolf Interagency Field Team

Abstract

The development of high-throughput sequencing technologies is dramatically increasing the use of single nucleotide polymorphisms (SNPs) across the field of genetics, but most parentage studies of wild populations still rely on microsatellites. We developed a bioinformatic pipeline for identifying SNP panels that are informative for parentage analysis from restriction site-associated DNA sequencing (RADseq) data. This pipeline includes options for analysis with or without a reference genome, and provides methods to maximize genotyping accuracy and select sets of unlinked loci that have high statistical power. We test this pipeline on small populations of Mexican gray wolf and bighorn sheep, for which parentage analyses are expected to be challenging due to low genetic diversity and the presence of many closely related individuals. We compare the results of parentage analysis across SNP panels generated with or without the use of a reference genome, and between SNPs and microsatellites. For Mexican gray wolf, we conducted parentage analyses for 30 pups from a single cohort where samples were available from 64% of possible mothers and 53% of possible fathers, and the accuracy of parentage assignments could be estimated because true identities of parents were known a priori based on field data. For bighorn sheep, we conducted maternity analyses for 39 lambs from five cohorts where 77% of possible mothers were sampled, but true identities of parents were unknown. Analyses with and without a reference genome produced SNP panels with ≥95% parentage assignment accuracy for Mexican gray wolf, outperforming microsatellites at 78% accuracy. Maternity assignments were completely consistent across all SNP panels for the bighorn sheep, and were 74.4% consistent with assignments from microsatellites. Accuracy and consistency of parentage analysis were not reduced when using as few as 284 SNPs for Mexican gray wolf and 142 SNPs for bighorn sheep, indicating our pipeline can be used to develop SNP genotyping assays for parentage analysis with relatively small numbers of loci.

KEYWORDS
Canis lupus baileyi, CERVUS, maternity, *Ovis canadensis*, paternity, restriction site-associated DNA sequencing, single nucleotide polymorphism

1 | INTRODUCTION

The ability to identify the parents of individuals in wild populations can provide insight into a wide range of topics including inbreeding levels (Dunn, Clancey, Waits, & Byers, 2011; Pemberton, 2004), translocation success (Hogg, Forbes, Steele, & Luikart, 2006; Marker et al., 2008), hybridization (Adams, Kelly, & Waits, 2003; Steyer et al., 2016), demographic processes (D'Aloia et al., 2015; Douhard, Festa-Bianchet, & Pelletier, 2016), mating system (Dugdale, Macdonald, Pope, & Burke, 2007; Hogg & Forbes, 1997; Jones, Kvarnemo, Moore, Simmons, & Avise, 1998), disease transmission (Plowright et al., 2017) and quantitative genetics (DiBattista, Feldheim, Garant, Gruber, & Hendry, 2009; Janeiro, Coltman, Festa-Bianchet, Pelletier, & Morrissey, 2017; Nguyen, Hayes, & Ingram, 2014). Genetic data can provide a powerful tool for identifying parents, and currently the main type of genetic marker used for parentage analysis in wild populations is microsatellites (Jones, Small, Paczolt, & Ratterman, 2010; Pemberton, 2008). One of the greatest strengths of microsatellites for parentage analysis is high polymorphism levels, which lead to high statistical power. However, a major disadvantage of microsatellites is that genotyping involves subjective visual interpretation of images, which can lead to relatively high genotyping error rates and the inability to directly compare data across laboratories (Bonin et al., 2004; Pompanon, Bonin, Bellemain, & Taberlet, 2005). In addition, the discovery and genotyping of microsatellite loci can be expensive and time-consuming.

Because of these drawbacks, single nucleotide polymorphisms (SNPs) have long been hailed as advantageous over microsatellites for addressing many ecological and evolutionary questions, including parentage assignment (Anderson & Garza, 2006; Brumfield, Beerli, Nickerson, & Edwards, 2003; López-Herráez, Schafer, Mosner, Fries, & Wink, 2005; Morin, Luikart, Wayne, & Grp, 2004). Genotyping of SNPs is less subjective, and SNP data can often be directly compared across laboratories provided the same protocols have been used for laboratory and bioinformatic analyses. In addition, genotyping of SNPs is typically less time-consuming than microsatellites, and SNP loci are more abundant in the genome than microsatellites. However, each SNP locus has lower heterozygosity and therefore lower statistical power than each microsatellite locus, and therefore larger numbers of SNPs than microsatellites are required to achieve sufficient power for parentage analyses (Glaubitz, Rhodes, & Dewoody, 2003; Hauser, Baird, Hilborn, Seeb, & Seeb, 2011; Tokarska et al., 2009). Furthermore, until recently the process for discovering large numbers of SNPs was costly and time-consuming for nonmodel organisms (Morin et al., 2004), and therefore, SNPs were not practical for many applications (Glaubitz et al., 2003; Jones et al., 2010).

The development of new high-throughput sequencing (HTS) technologies over the last decade has dramatically increased the feasibility of SNP discovery and genotyping due to substantial decreases in time and cost for generating large quantities of sequence data. HTS facilitates several methods for discovering and genotyping large numbers of SNPs, including whole-genome sequencing (Ekblom & Wolf, 2014), transcriptome sequencing (De Wit, Pespeni, & Palumbi, 2015), DNA capture (Jones & Good, 2016) and restriction site-associated DNA sequencing (RADseq) (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016). Of these methods, RADseq has a number of advantages for identifying parentage-informative SNPs in nonmodel organisms. RADseq involves sequencing regions adjacent to restriction cut sites and therefore generates sequence data from across the genome, primarily in noncoding regions. RADseq is flexible in the number of loci it can target, but typically generates data from more than enough loci for parentage analyses. Compared to whole-genome and transcriptome sequencing, RADseq is much less expensive per sample, primarily because it generates data from a much smaller number of loci. In addition, RADseq is advantageous over transcriptome sequencing because it does not require high-quality tissue samples, which are often unavailable for wild populations. When compared to DNA capture, RADseq is advantageous because it requires no prior genomic knowledge, and is largely unaffected by ascertainment bias (Clark, Hubisz, Bustamante, Williamson, & Nielsen, 2005; Lachance & Tishkoff, 2013).

Although RADseq can be used directly for parentage analyses, projects with large sample sizes may save time and money by first using RADseq for SNP discovery with a subset of samples, and then using a different method for SNP genotyping of the remaining samples. RADseq typically generates data from at least several thousand SNPs, whereas parentage analyses usually require just tens or hundreds of SNPs (Anderson & Garza, 2006; Glaubitz et al., 2003; Kaiser et al., 2017; Tokarska et al., 2009). A number of approaches exist that are time- and cost-efficient for genotyping these smaller numbers of SNPs for large numbers of samples, including Fluidigm Dynamic Array (Fluidigm Corp, San Francisco, USA), MassARRAY (Agena Biosciences, San Digeo, USA), multiplex PCR amplicon sequencing (Campbell, Harmon, & Narum, 2015) and Rapture (Ali et al., 2016). For studies relying on noninvasively collected samples, such as faecal and hair samples (e.g., Constable, Ashley, Goodall, & Pusey, 2001; Rudnick, Katzner, Bragin, Rhodes, & Dewoody, 2005), some of these approaches will also likely outperform RADseq due to less stringent requirements for both quantity and quality of starting genomic DNA (Campbell et al., 2015; Kraus et al., 2015). However, each of these approaches requires primers and/or probes that have been custom-designed to target a predefined set of loci, which can be accomplished using data generated for a small subset of samples using HTS approaches like RADseq.

A small but growing number of studies are using HTS to discover and genotype SNPs for parentage analysis in wild populations. For example, Holman, de la Serrana, Onoufriou, Hillestad, and Johnston (2017) used RADseq data from 104 Atlantic salmon (*Salmo salar*) to identify 1,517 SNPs and then used these data to design a Fluidigm Dynamic Array assay to genotype 289 fish at 94 loci for parentage analysis. Weinman, Solomon, and Rubenstein (2015) used SNPs discovered from transcriptome sequencing of four superb starling (*Lamprotornis superbus*) samples to design a MassARRAY assay to

genotype 224 birds at 102 loci, and Kaiser et al. (2017) took a similar approach for black-throated blue warblers (*Setophaga caerulescens*). Nguyen et al. (2014) used whole-genome sequence data from one blue mussel (*Mytilus galloprovincialis*) to design a MassARRAY assay to genotype 227 SNPs for 3,711 samples. Each of these studies found comparable or improved performance for parentage analyses with SNPs when compared to microsatellites.

The optimal bioinformatic pipeline for discovering and genotyping informative SNPs will vary across HTS approaches and study systems. For example, locus assembly methods will differ for data generated from transcriptome sequencing, whole-genome sequencing, DNA capture or the wide variety of RADseq methods. In addition, the numbers of SNPs that must be identified to achieve strong statistical power for parentage analyses will vary across study systems based on genetic diversity, mating system and the number of individuals sampled. However, all bioinformatic pipelines will share some common goals; for example, all must take into account the relatively high genotyping error rate inherent in HTS data, identify and remove SNPs in paralogous and other repetitive genomic regions, and generate a set of unlinked loci.

Here, we develop a bioinformatic pipeline to identify a panel of informative SNPs for parentage analysis from RADseq data in two study systems: Mexican gray wolf (*Canis lupus baileyi*) and bighorn sheep (*Ovis canadensis*). High-quality reference genomes are available from closely related species for both of these taxa, and our pipeline includes the option of alignment to a reference genome. However, many nonmodel study systems will not have an available reference genome, and therefore, our pipeline also includes an option for assembling loci de novo. We compare the results of parentage analyses conducted with reference-based and de novo-assembled SNP panels, and with results generated using microsatellite markers. We expected large numbers of markers would be required to achieve sufficient statistical power for parentage analyses in both study systems, because both are small populations (about 80-110 individuals), and are therefore expected to have low genetic diversity. Another reason to expect low diversity in the Mexican gray wolf population is that its sole source is a captive population that was started by just seven individuals. Furthermore, the Mexican gray wolf has a mating system which results in large numbers of full siblings, as breeding typically occurs only between one breeding female and one breeding male within each pack across multiple years, and large numbers of markers may be required to distinguish between full siblings as potential parents. In contrast, the bighorn sheep has a polygynous mating system, with males competing for access to females for mating. Larger males tend to dominate matings, but most females breed and produce just one offspring every year, and therefore, the proportion of full siblings in the population should be lower than for the Mexican gray wolf.

For the Mexican gray wolf study system, we conducted both maternity and paternity analyses, and we knew a priori with high certainty the true identities of both parents for all sampled pups based on observational field data; therefore, this study system provided an excellent opportunity to test both the power and accuracy of our bioinformatic pipeline for parentage analyses. For our bighorn sheep study, we performed maternity analyses only, and for this study, we did not know the identities of mothers a priori. Therefore, we were not able to test the accuracy of the bighorn sheep maternity assignments, but were able to compare statistical power and consistency of maternity assignments across SNP panels and marker types. For both study systems, we determined the minimum number of SNPs that could be used without reducing the accuracy or consistency of parentage analyses, with the aim of designing a cost-effective assay for genotyping large numbers of samples in the future.

2 | METHODS

2.1 | Study systems and sample collection

Mexican gray wolves were originally distributed across the southwest United States and central Mexico, but were almost completely eradicated by the mid-1900s. In 1998, a wild population was re-established by releasing 11 captive individuals (Fish and Wildlife Service 2015). Through additional releases and natural reproduction, the wild population size had reached 113 at the end of 2016. Mexican gray wolves live in packs comprised of about four to eight individuals, including one breeding pair that is usually monogamous, and extended family members that typically do not breed. Mexican gray wolves begin breeding at 2 years of age, and litter size is usually four to six pups, and pups remain in the natal pack until at least 1.5 years of age (Fish and Wildlife Service 2010).

Mexican gray wolf adults, yearlings and pups are captured each year as part of a long-term monitoring project by the U.S. Fish and Wildlife Service. The age of each individual is estimated based on body size and tooth morphology, and blood samples are collected. For the study described here, we used blood samples from all individuals determined to have been born in 2014 (hereafter called "pups"), all individuals of the appropriate age to be potential parents of these pups, and eight additional individuals (pups and adults from other years) (Table 1a, Supporting Information Table S1). For each pup, we had high certainty of the true identities of both parents from field data: we knew the geographic location where each pup was sampled, the geographic range of each pack, and the identities of the male and female breeding pair for each pack.

Bighorn sheep were extirpated from much of their historical range in the western United States during the late 19th and early 20th century (Buechner, 1960). Extensive restoration efforts have successfully re-established this species in historical habitat, although these populations are often small and fragmented (Olson, Whittaker, & Rhodes, 2013). Bighorn sheep are polygynous, and males use various tactics to obtain matings, but not all males sire offspring (Hogg & Forbes, 1997). Females generally give birth to one lamb per year starting at age two, and lambs are weaned by about 4 months.

The Lostine bighorn sheep population in northeast Oregon was re-established in 1971 with the translocation of 20 individuals from Alberta, and the population grew and has remained around 80 sheep over the last 20 years (Cassirer et al., 2013; Coggins, 2006). Each

TABLE 1 Sample sizes and estimated per cent of all possible parents sampled in the full and reduced data sets for (a) Mexican gray wolf and (b) Bighorn sheep

	Full data set		Reduced data set	
Sample type	*n*	% parents	*n*	% parents
(a)				
Pups (2014)	30		30	
Potential mothers	21	95	14	64
Potential fathers	36	100	20	53
Other	8		8	
Total	95		72	
(b)				
Lambs (2011–2015)	42		31	
Lambs + potential mothers	8		8	
Potential mothers	47	90	41	77
Other	3		2	
Total	100		82	

winter from 2011 to 2016, we captured six- to nine-month-old lambs, yearlings and adults using baited corral traps and ground-darting; sampling of adults focused primarily on females (Table 1b, Supporting Information Table S2). We estimated the age of each individual based on morphology (for lambs), tooth eruption and capture history (Plowright et al., 2017), and collected blood and tissue samples (Table 1b, Supporting Information Table S2). We also collected faecal samples from unmarked individuals after they defecated. Lambs were sampled postweaning at 6–9 months of age, and most were no longer associating with their dam at this age; therefore, the identity of their dam was uncertain. Potential dams for each lamb were identified as adult females at least 2 years old and known to be alive at the time of the lamb's birth.

2.2 | Sample sizes

Generating RADseq data requires a greater quantity and higher quality of genomic DNA than does generating microsatellite data. Our RADseq protocol uses 50 ng of high-molecular-weight genomic DNA, whereas our microsatellite protocols are often successful even with nondetectable quantities of degraded DNA. Thus, our RADseq analyses used a subset of our total samples because some samples had insufficient quantity or quality of DNA. Hereafter, we refer to the sample set including all individuals as the "full data set" and the sample set including individuals with RADseq data as the "reduced data set" (Table 1, Supporting Information Tables S1 and S2).

For the Mexican gray wolf, we generated microsatellite data for 30 pups, 57 potential parents (21 females and 36 males), and eight additional individuals (pups and adults from other years), and we generated RADseq data for 30 pups, 34 potential parents (14 females and 20 males) and eight additional individuals (Table 1a, Supporting Information Table S1). These numbers do not include one potential mother that was later removed from the RADseq data set due to low numbers of genotyped RADseq loci (see below).

For the bighorn sheep, we generated microsatellite data from 42 lambs, 47 potential mothers, eight individuals that were both a lamb and a potential mother (e.g., a lamb born in 2011 could be the mother of a lamb born in 2015), and three additional males (Table 1b, Supporting Information Table S2). We generated RADseq data for 31 lambs, 41 potential mothers, eight individuals that were both a lamb and potential mother, and two additional males. These numbers do not include one lamb that was later removed from the RADseq data set due to low numbers of genotyped RADseq loci (see below).

2.3 | DNA extraction

DNA was extracted from blood and tissue samples using the DNeasy Blood and Tissue Kit (Qiagen, Inc.), and from faecal samples using the QIAamp Fast DNA Stool Mini Kit (Qiagen, Inc.). Faecal samples were extracted in a laboratory dedicated to low quality DNA samples. One negative control was included in each extraction to monitor for contamination of reagents.

2.4 | Microsatellite analyses

A total of 22 microsatellite loci were PCR amplified for each Mexican gray wolf sample (see Supporting Information Appendix S1 for PCR conditions), and a total of 15 microsatellite loci were PCR amplified in two multiplexes for each bighorn sheep sample (see Plowright et al., 2017 for PCR conditions). All PCRs were run with a negative control to test for reagent contamination. Each multiplex microsatellite PCR was performed twice for blood and tissue samples, and at least three times for faecal samples. PCR products were run on an Applied Biosystems 3130xl Genetic Analyzer and scored using GeneMapper v5.0 (Applied Biosystems, Inc). For faecal samples, we accepted a heterozygous genotype if it was observed in at least two PCRs, and a homozygous genotype if it was observed in at least three PCRs. We discarded samples with <50% amplification success or for which a consensus genotype was not obtained for at least 11 loci after five PCRs per multiplex. We used cervus 3.0 (Kalinowski, Taper, & Marshall, 2007) to estimate observed heterozygosity (H_o), expected heterozygosity (H_e) and combined nonexclusion probability across loci (the probability of not excluding a single unrelated candidate parent or parent pair from parentage assignment), and to test whether microsatellite loci deviated from Hardy–Weinberg equilibrium (HWE).

2.5 | RADseq library prep and sequencing

RADseq libraries were prepared following Ali et al. (2016), starting with 50 ng of high-molecular-weight genomic DNA per sample. Genomic DNA was digested using the restriction enzyme *SbfI*, and biotinylated RADseq adapters containing 8 bp barcodes were ligated to the restriction cut sites. The ligation products from all samples

were combined, and the multiplexed ligation product was then sheared to 400 bp using a Covaris M220 Focused-ultrasonicator. Genomic DNA fragments without ligated adapters (and therefore without restriction cut sites) were removed using Streptavidin bead washes, and the remaining DNA with ligated adapters was processed using the NEBNext Ultra DNA Library Prep Kit for Illumina, excluding the initial shearing step. The resulting libraries were sequenced using an Illumina HiSeq4000 at the University of California Berkeley QB3 Vincent J. Coates Genomics Sequencing Library with 150 bp paired end reads.

We performed RADseq library prep and Illumina sequencing twice independently for 18 of our bighorn sheep samples and nine of our Mexican gray wolf samples, for the purpose of choosing the best parameters for locus filtering (see below), or to increase the numbers of sequence reads for samples with low read counts after an initial Illumina run.

2.6 | RADseq de-multiplexing and quality control

Our bioinformatic pipeline is illustrated in Figure 1 and includes analytical methods for both reference genome-based and de novo RADseq analyses. For the RADseq method used here, the barcode and partial restriction site can occur on either the forward or reverse Illumina reads (Ali et al., 2016). Therefore, we used a custom perl script

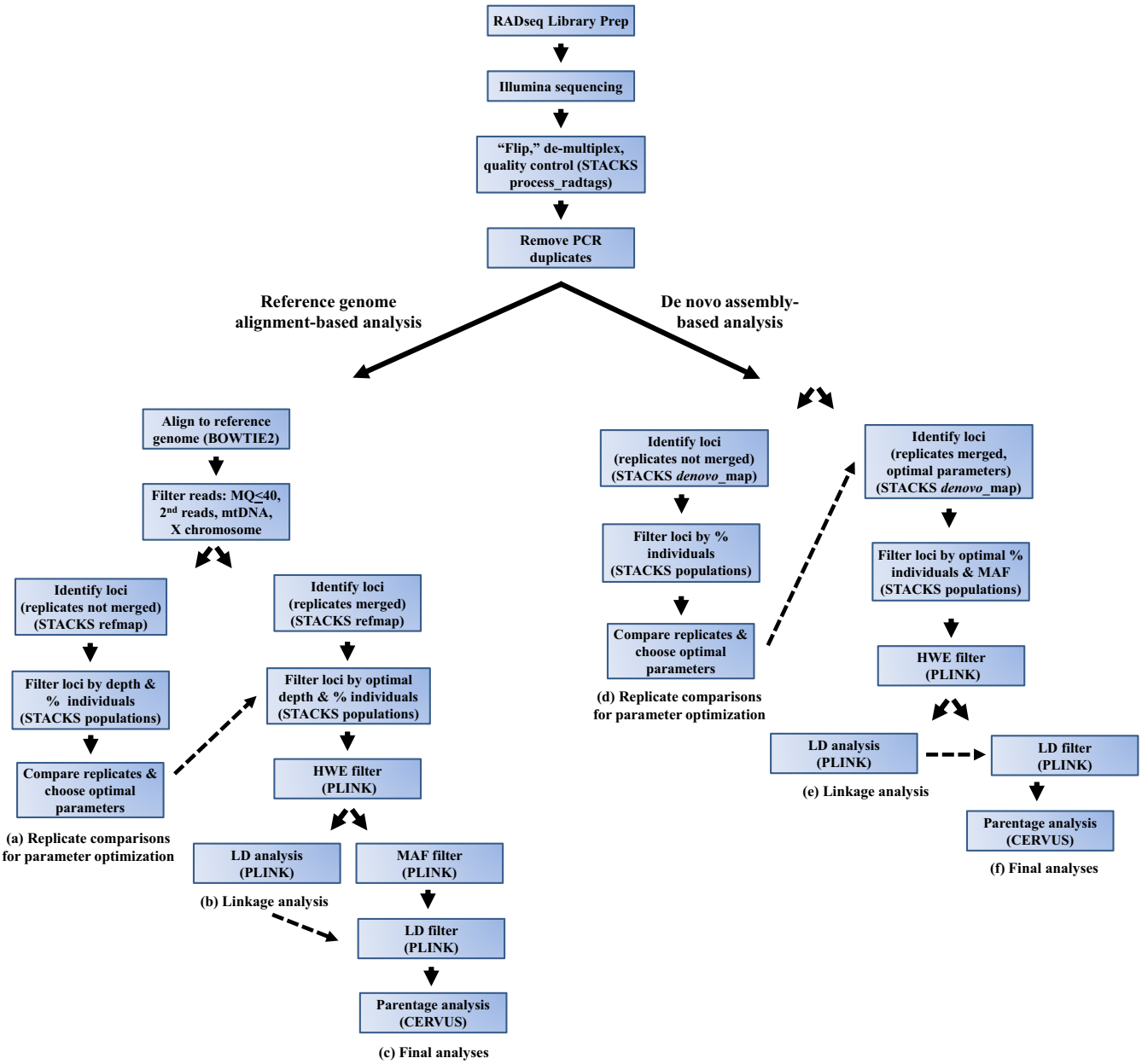

FIGURE 1 Bioinformatic pipeline for discovering and genotyping RADseq SNPs for parentage analysis with a reference genome ("Reference genome alignment-based analysis") and without a reference genome ("De novo assembly-based analysis"). Dashed lines indicate that parameter optimization analysis (a,d) and linkage disequilibrium analysis (b,e) are used to inform parameter choices for final analysis (c,f) [Colour figure can be viewed at wileyonlinelibrary.com]

to "flip" the raw sequence reads so that all reads starting at the restriction cut site were in one file, with the other reads ("second reads") in a second file (Supporting Information Appendix S2). The program PROCESS_RADTAGS in STACKS 1.42 (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) was used to demultiplex reads by barcode and to remove reads with poor sequence quality or uncalled bases. This program uses a sliding window approach to identify and remove reads with average phred scores less than 10 (90% probability of being correct) for each window. PCR duplicates were then removed using the CLONE_FILTER program in STACKS.

2.7 | Reference alignment-based analyses

2.7.1 | Genome alignment and mapping quality filtering

Filtered Mexican gray wolf sequence reads were aligned to the domestic dog genome (*Canis lupus familiaris*, CanFam3.1), and filtered bighorn sheep sequence reads were aligned to the domestic sheep genome (*Ovis aries*, Oar_v3.1) using BOWTIE2 v.2.1.0 (Langmead & Salzberg, 2012) with the following parameters: –sensitive, –end-to-end, –X 900. Reads with mapping quality ≤40 and reverse reads were removed using PYSAM (https://github.com/pysam-developers/pysam, Li et al., 2009). We discarded reverse reads because these reads do not start at a restriction cut site when using the Ali et al. (2016) protocol, and instead start a variable distance away from the restriction cut site, and therefore, SNPs in the reverse reads are expected to have low depth of coverage. Reads aligning to the X chromosome and the mitochondrial DNA (mtDNA) were also removed.

2.7.2 | Parameter optimization: minimum depth and per cent individuals genotyped

We used the program REF_MAP in STACKS to identify SNPs for each individual sample from the reference-aligned sequence reads using a maximum likelihood approach. We required a minimum of three identical reads to create a stack and used an upper bound for the sequence error rate at 0.01. We then used the program POPULATIONS in STACKS to identify the best parameters for the "minimum depth of coverage to accept a locus" and the "minimum per cent of individuals genotyped to accept a locus" (hereafter "*r*"). To accomplish this, we chose five samples for which library prep and sequencing were conducted twice independently (hereafter called "replicate pairs") to compare the consistency of genotypes across replicates by calculating the genotype mismatch rate, or the proportion of loci for which genotypes were inconsistent between replicate pairs. We expected the genotype mismatch rate across samples to be related to the sequence read count, and therefore, we chose the five replicate pairs with the greatest range in sequence read count, to maximize our ability to predict error rates for samples across the range of read counts in our data set.

We tested 25 parameter sets in POPULATIONS, varying the minimum depth from five to nine, and varying *r* from 50% to 90%. These analyses were conducted using all samples in the data set; for each of the five chosen replicate pairs, the data from each of the two replicates were analysed individually, whereas sequence reads from all other replicate pairs were merged. For all parameter sets, we filtered out loci with minor allele frequency (MAF) <0.05 to minimize sequence errors present in the data set. We then estimated genotype mismatch rates between replicate samples using a modification of an R script developed by Mastretta-Yanes et al. (2015). The mismatch rate for each replicate pair was calculated as the number of loci for which the genotypes were different between replicates, divided by the total number of loci typed for both replicates.

2.7.3 | Filtering by HWE, MAF, and linkage

After identifying the parameter values that generated the lowest genotype mismatch rates, we merged the sequence data from the replicate samples and conducted the REF_MAP analysis using these merged samples, and then used the optimized parameter values in POPULATIONS to identify and genotype SNPs for all samples. We then used PLINK 1.90 (Purcell et al., 2007) to remove loci that were not in HWE ($p < 0.05$, using the mid-p adjustment recommended by the PLINK authors). To determine the best method for removing linked loci, we visualized linkage disequilibrium (LD) decay by plotting the squared allele count correlation (r^2) between all pairs of SNPs, except pairs more than 20 SNPs or 2,000 kb apart, calculated using PLINK. We then used PLINK to conduct a series of four MAF filters (retaining only loci with MAF>0.1, MAF>0.4, MAF>0.45, MAF>0.475) to determine the lowest number of the highest diversity loci that could provide sufficient statistical power for parentage analyses. We then used PLINK to filter out linked loci from each of the filtered locus sets. We also used PLINK to identify individual samples with >70% missing data. We then genotyped all individuals, excluding individuals with >70% missing data, for each of the filtered locus sets using "whitelists" (lists of the desired loci to be genotyped) in POPULATIONS.

2.8 | De novo assembly-based analyses

2.8.1 | Locus assembly and parameter optimization

We used the DENOVO_MAP pipeline in STACKS to assemble loci de novo. We only used the forward sequence reads for this analysis, for the same reasons described above for the reference-based analysis. Locus assembly in STACKS is determined by three main parameters: the minimum number of identical raw reads to create a stack within an individual (*m*), the maximum number of mismatches allowed between stacks to merge them into one locus within an individual (*M*) and the maximum number of mismatches allowed to merge stacks from different individuals into one locus (*n*) (Catchen et al., 2013). We tested 11 parameter combinations: $m = 2$–6 (with other parameters fixed at $M = 2$, $n = 1$), $M = 2$–5 (with $m = 3$, $n = 1$), and $n = 1$–4 (with $m = 3$, $M = 2$), using the same five replicate pairs as for the reference-based analysis described above. We

then used POPULATIONS to identify the optimal value for *r* (minimum per cent of individuals genotyped to accept a locus), testing four different values: 40%, 60%, 80%, 90%. We estimated genotype mismatch rates between replicate samples using the approach described above.

2.8.2 | Filtering by MAF, HWE and linkage

After identifying the de novo parameter values that generated the lowest genotype mismatch rates, we merged the sequence data from the replicate samples and conducted the DENOVO_MAP analysis using the optimal values for *m*, *M* and *n*, and then used the optimal value for *r* in POPULATIONS to identify and genotype SNPs for all samples. We also used POPULATIONS to filter SNPs by three MAF cut-offs (retaining SNPs with MAF>0.05, >0.3, >0.4); we used lower MAF cut-offs for de novo than reference-based analyses after initial analyses with optimal de novo parameters indicated that high MAF cut-offs resulted in fewer than 100 SNPs. We also used POPULATIONS to filter all except one SNP per RAD locus (parameter –write_single_snp) to reduce the number of physically linked loci retained. We then used PLINK to remove loci that were not in HWE ($p < 0.05$, using the mid-p adjustment). To determine the best method for removing physically linked loci under a hypothetical scenario in which we did not have a reference genome available, we calculated r^2 between all pairs of SNPs (filtered for HWE and MAF>0.05) using PLINK and plotted a histogram of these values. We then chose a maximum r^2 value cut-off that would conservatively exclude the tails of the distribution, and used PLINK to filter one of each SNP from pairs with r^2 values greater than this cutoff. We also used PLINK to identify individual samples with >70% missing data. We then genotyped all individuals, excluding individuals with >70% missing data, for each of the filtered locus sets using whitelists in POPULATIONS.

2.9 | Parentage analysis

For both of our study systems, we conducted parentage analyses with microsatellite and RADseq markers using CERVUS (Supporting Information Tables S3, S4, S5, S6, S7, Appendix S3). This program calculates the likelihood that each candidate female or male is the mother or father, taking into account population allele frequencies and genotyping errors. To determine whether a potential parent has a high enough likelihood score to assign parentage, CERVUS uses a simulation approach based on the observed allele frequencies to calculate the expected differences in likelihood between the true parent and other candidate parents.

As described above, we had lower sample sizes with RADseq data ("reduced data set") than with microsatellite data ("full data set") for both study systems. Therefore, we conducted microsatellite analyses using both the reduced and the full data sets (to compare the performance of microsatellites for these two different data set sizes), and we conducted RADseq analyses using just the reduced data set.

For all simulation analyses, we used 100,000 offspring (as recommended by the CERVUS authors), an estimated genotyping error rate of 1% for both the microsatellites and RADseq SNPs, and an estimated "per cent parents sampled" determined based on observational field data (see Results section). For the Mexican gray wolf, we accepted a parent-pair assignment at a 95% confidence level. If no parent pairs were assigned for a given pup, then we accepted a single-parent assignment at a 95% confidence level. For the bighorn sheep, we accepted assignments at lower stringency (80% confidence level) after we observed few assignments at the 95% confidence level for microsatellite data. We also used CERVUS to estimate H_o, H_e and combined nonexclusion probability for each SNP set.

2.10 | Delta scores

The statistical power of parentage analysis can be evaluated by examining delta scores calculated for each offspring by CERVUS. Delta scores are the difference in logarithm of odds (LOD, calculated as the natural log of the overall likelihood ratio) scores between the most likely candidate parent and the second most likely candidate parent, treating negative LOD scores as zero. Greater delta scores should indicate greater power to distinguish between candidate parents. We compared delta values for Mexican gray wolf maternal and paternal assignments, and for bighorn maternal assignments.

2.11 | Parent/offspring locus incompatibility rates

Locus incompatibility rates between assigned parents and offspring can provide insight into factors driving the performance of parentage analyses. Parent/offspring locus incompatibilities occur when the allelic composition of a locus is inconsistent with a parent/offspring relationship, for example if an offspring has no alleles in common with the assigned parent at the locus. When conducting parentage analyses in CERVUS, the number of incompatibilities allowed between offspring and assigned parents is dictated by the estimated genotyping error rate. We calculated the proportion of loci with incompatibilities for assigned parent/offspring pairs for all locus sets and sample sets for the Mexican gray wolf, for which we had prior knowledge of the correct parentage assignments.

3 | RESULTS

3.1 | Microsatellite diversity

For the Mexican gray wolf, no microsatellite loci significantly deviated from HWE, and an average of 99.8% of loci were typed across individuals. For the bighorn sheep, one of the 15 loci (BL4) deviated from HWE and was removed from subsequent analyses, and an average of 99.2% of the remaining loci were genotyped across individuals. Mean H_o and H_e across loci for the full data set were similar for the Mexican gray wolf ($H_o = 0.65$, $H_e = 0.62$) and bighorn sheep ($H_o = 0.59$, $H_e = 0.61$) (Table 2). Nonexclusion probability for parent pairs for the Mexican gray wolf (7.0×10^{-8}) was higher than nonexclusion probability for first parents for the bighorn sheep (0.026) (Table 2).

TABLE 2 Numbers of loci and diversity statistics for different microsatellite and RADseq marker sets and sample sets

	Mexican gray wolf									Bighorn sheep							
			H_o			H_e						H_o			H_e		
	# loci	NEP	Mean	Low	High	Mean	Low	High		# loci	NEP	Mean	Low	High	Mean	Low	High
Microsats																	
Full data set	22	7.0×10^{-8}	0.65	0.32	0.81	0.62	0.31	0.76		14	0.026	0.59	0.19	0.77	0.61	0.18	0.82
Reduced data set	22	6.0×10^{-8}	0.65	0.32	0.85	0.62	0.32	0.77		14	0.025	0.62	0.21	0.84	0.60	0.20	0.78
RADseq: Reference-aligned																	
MAF>0.1	1,478	7.3×10^{-180}	0.42	0.15	0.62	0.40	0.18	0.50		3,044	2.5×10^{-101}	0.37	0.15	0.62	0.37	0.18	0.50
MAF>0.4	480	4.4×10^{-69}	0.52	0.38	0.62	0.50	0.48	0.50		639	9.2×10^{-37}	0.49	0.38	0.62	0.50	0.48	0.50
MAF>0.45	292	1.6×10^{-42}	0.52	0.39	0.62	0.50	0.50	0.50		333	6.6×10^{-20}	0.50	0.38	0.62	0.50	0.50	0.50
MAF>0.475	159	1.6×10^{-23}	0.52	0.40	0.62	0.50	0.50	0.50		164	3.2×10^{-10}	0.50	0.39	0.62	0.50	0.50	0.50
RADseq: De novo																	
$m = 3$, MAF>0.05	363	3.2×10^{-45}	0.43	0.10	0.62	0.41	0.10	0.50	MAF>0.05	523	4.9×10^{-18}	0.10	0.36	0.61	0.10	0.36	0.50
$m = 4$, MAF>0.05	284	1.4×10^{-34}	0.41	0.08	0.63	0.40	0.10	0.50	MAF>0.3	240	2.2×10^{-13}	0.47	0.35	0.61	0.48	0.42	0.50
$m = 5$, MAF>0.05	223	3.8×10^{-26}	0.39	0.10	0.62	0.37	0.10	0.50	MAF>0.4	142	8.8×10^{-9}	0.48	0.38	0.61	0.50	0.48	0.50
$m = 6$, MAF>0.05	139	1.2×10^{-14}	0.33	0.10	0.58	0.31	0.10	0.50		NA	NA	NA	NA	NA	NA	NA	NA

Note. H_o: observed heterozygosity; H_e: expected heterozygosity; NEP: combined nonexclusion probability for parent pairs (Mexican gray wolf) and first parents (bighorn sheep). Parameter sets were the same for Mexican gray wolf and Bighorn sheep, except as indicated for the de novo analyses.

3.2 | Locus identification and genotyping: reference-based methods

A large percentage of sequence reads aligned to the reference genomes for both species, including a mean of 96.9% of reads (range: 96.7%–97.0%) for Mexican gray wolf and a mean of 94.0% of reads (range: 78.0%–95.8%) for bighorn sheep. The Mexican gray wolf had a greater mean percentage of mapped reads retained after mapping quality filtering (mean: 84.0%, range 84.7%–85.8%) than bighorn sheep (mean: 72.5%, range: 45.9%–83.0%). After removing reads aligned to the X chromosome and mtDNA, and merging reads for individuals that were sequenced more than once, the total number of reads retained across samples was higher for Mexican gray wolf (mean: 2.88 million, range: 0.275 million–12.4 million) than bighorn sheep (mean 1.70 million, range: 0.299 million–4.91 million). For the five samples chosen as replicates for evaluating the best parameter values for minimum depth and minimum per cent individuals (see above), the mean number of sequence reads per replicate was 2.96 million (range: 0.686 million–6.86 million) for Mexican gray wolf and 1.16 million (range: 0.319 million–2.03 million) for bighorn sheep. Of the 25 parameter sets for which SNPs were identified and genotyped using POPULATIONS, the lowest mean genotype mismatch rate across replicates was obtained using the same parameters for both Mexican gray wolf and bighorn sheep: minimum depth of six, and minimum per cent individuals of 90% (Figure 2). The lowest mean genotype mismatch rate across replicates for SNPs identified with these parameters ranged from 0.0028 to 0.025 for Mexican gray wolf and 0.013 to 0.045 for bighorn sheep. As expected, mismatch rates were consistently lower across the 25 parameter sets for replicate pairs with higher numbers of sequence reads (Figure 2a,c).

Linkage decay analysis indicated overall stronger linkage for Mexican gray wolf than bighorn sheep (Figure 3). For Mexican gray wolf, r^2 decreased rapidly until 100 kb, and then decreased more slowly but remained above 0.2 (Figure 3a). For bighorn sheep, mean r^2 decreased rapidly until 200 kb and then remained around 0.1 (Figure 3b). Based on these results, we chose the following filter parameters to retain unlinked loci: For the Mexican gray wolf, we removed one locus from each pair of loci within a 1,000 kb sliding window with r^2 greater than 0.2, shifting windows by 1 bp steps; for bighorn sheep, we removed one locus from each pair of loci within a 500 kb

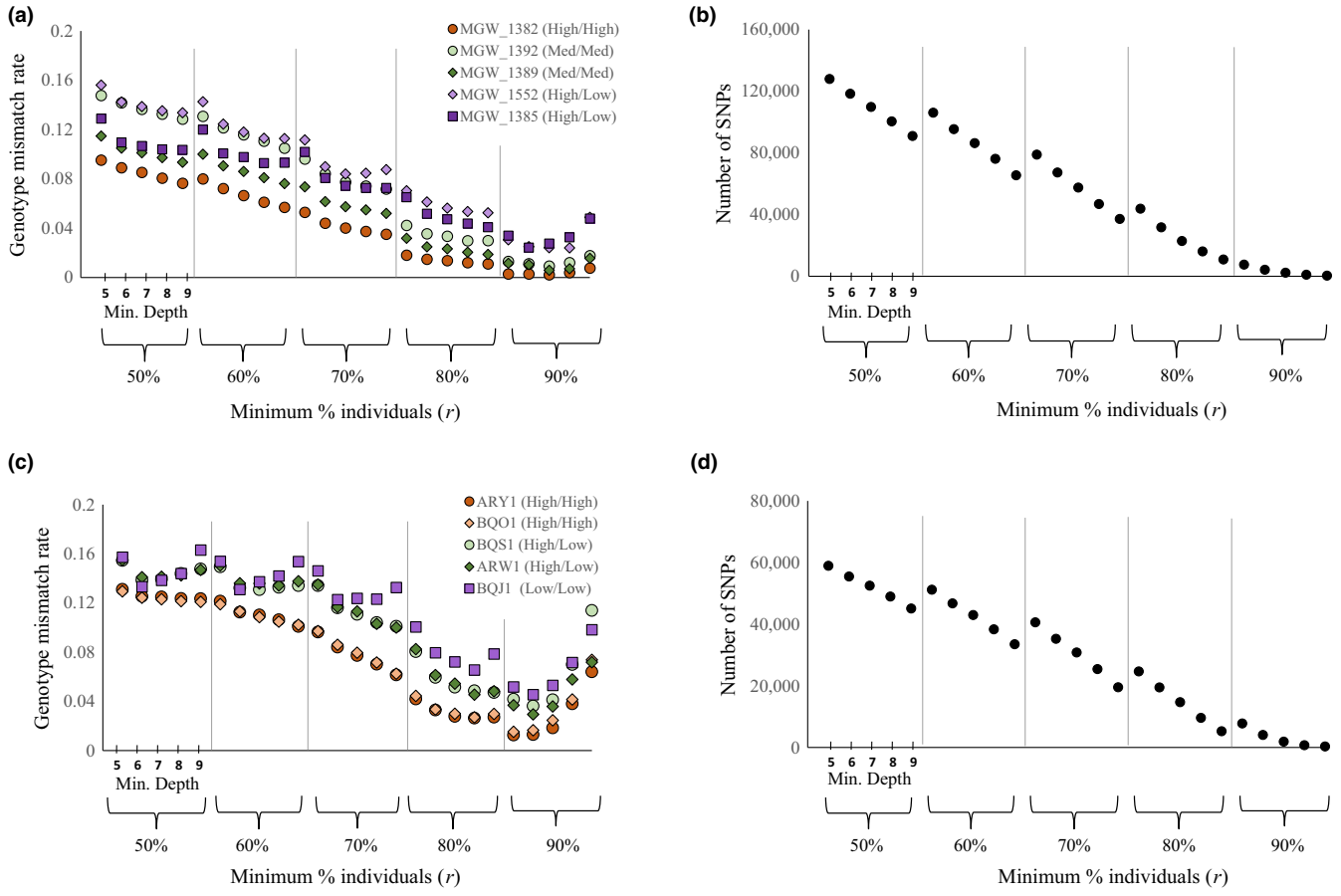

FIGURE 2 Genotype mismatch rates and number of SNPs for 25 parameter sets with reference genome-based RADseq analysis for five replicate samples each for Mexican gray wolf (a,b) and Bighorn sheep (c,d). Parameter sets vary in the minimum depth of coverage to accept a locus (ranging from 5 to 9) and the minimum per cent of individuals genotyped to accept a locus (*r*, ranging from 50% to 90%). High, medium and low refer to the relative number of sequence reads for each sample in the replicate pair. For example, "high/low" indicates one replicate had a relatively high number of sequence reads, and the second replicate had a relatively low number of reads

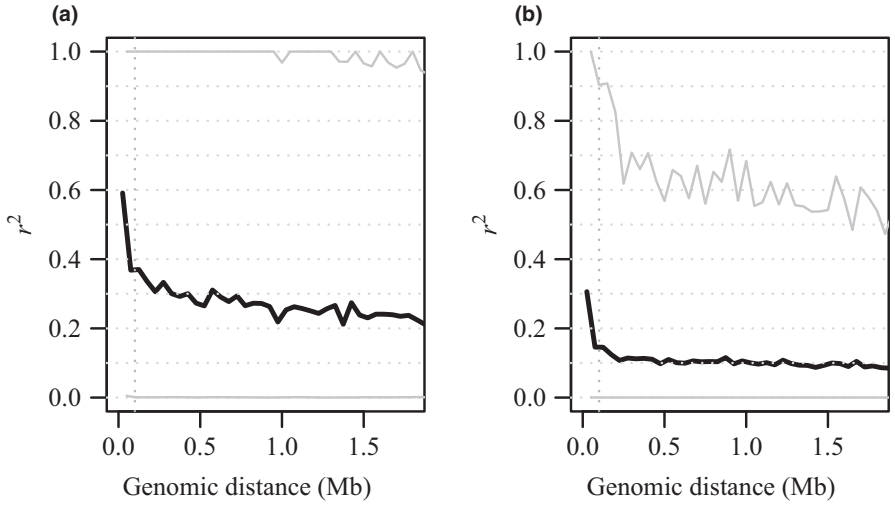

FIGURE 3 Linkage decay for Mexican gray wolf (a) and Bighorn sheep (b) for RADseq loci. Dotted vertical line is at 100,000 bp. Solid gray lines are the 2.5 and 97.5 percentiles

sliding window with r^2 greater than 0.1, shifting windows by 1 bp steps.

Despite a larger mean number of filtered sequence reads for Mexican gray wolf than bighorn sheep, fewer SNPs were retained after MAF filters (Mexican gray wolf: 159–1,478 SNPs; bighorn sheep: 164–3,044 SNPs), indicating lower diversity for this species (Table 2). For each species, one individual had <70% loci genotyped after MAF and LD filtering and was removed from subsequent analyses. Nonexclusion probabilities were lower for Mexican gray wolf (7.3×10^{-180}–1.6×10^{-23}) than bighorn sheep (2.5×10^{-101}–3.2×0^{-10}).

3.3 | Locus identification and genotyping: de novo assembly

Genotype mismatch rates from de novo analyses were most strongly influenced by the parameter r for both species, with the lowest mismatch rates occurring when r = 0.9 (Figure 4). Increasing values of m

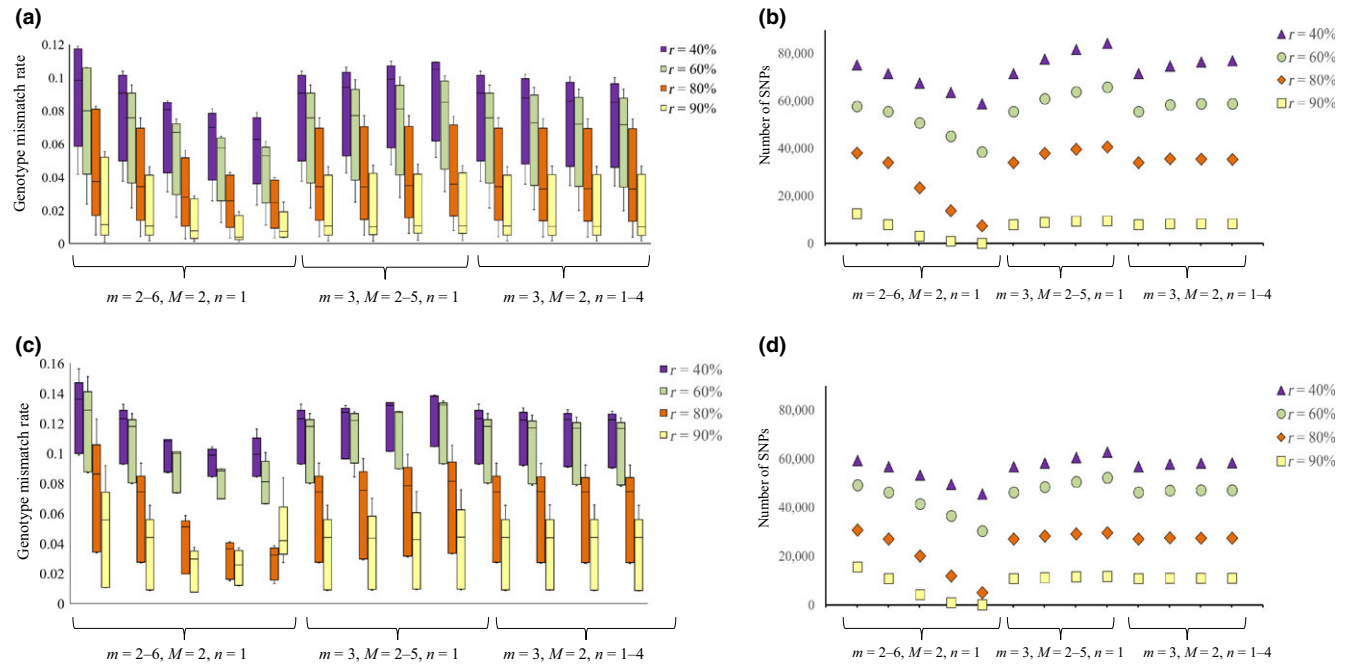

FIGURE 4 Genotype mismatch rates and number of SNPs for 44 parameter sets with de novo RADseq analysis for five replicate samples each for Mexican gray wolf (a,b) and Bighorn sheep (c,d). Parameter sets vary in the minimum number of identical raw reads to create a stack within an individual (m), the maximum number of mismatches allowed between stacks to merge them into one locus within an individual (M), the maximum number of mismatches allowed to merge stacks from different individuals into one locus (n) and the minimum per cent of individuals to accept a locus (r). Note that results from the parameter set with the default settings (m = 3, M = 2, n = 1) are shown three times for comparative purposes

generally resulted in decreasing mismatch rates, but varying the *M* and *n* parameters had little impact on genotype mismatch rates. For both species, the optimal de novo parameter set was $m = 5$, $M = 2$, $n = 1$, $r = 0.9$. The Mexican gray wolf data set had lower mismatch rates and greater numbers of SNPs than the bighorn sheep data set across parameter sets, likely due to the higher average depth of coverage for the Mexican gray wolf. Mean mismatch rates were slightly lower for de novo than reference-based analyses (Mexican gray wolf de novo range 0.0009–0.019 compared to reference-based range 0.0028–0.025; bighorn sheep de novo range 0.012–0.037 compared to reference-based range 0.013–0.045). However, de novo analysis resulted in fewer SNPs than reference-based analysis for both species when using optimal parameters. For example, the genotype mismatch analysis with optimal parameters for Mexican gray wolf resulted in 4,222 SNPs for reference-based analysis compared to 1,082 SNPs for de novo analysis, and the analysis for bighorn sheep resulted in 4,094 SNPs for reference-based analysis compared to 1,141 SNPs for de novo analysis.

We report r^2 values for SNPs identified using the optimal de novo and *r* parameter values and filtered for HWE and MAF>0.05 (Figure 5). Most r^2 values were <0.1, but the tails of the distribution extended to $r^2 = 1.0$ for both species. We chose r^2 cut-offs of 0.25 for Mexican gray wolf and 0.20 for bighorn sheep to exclude SNP pairs with r^2 values in the tails of the distribution. For the Mexican gray wolf, MAF filtering of SNPs identified using the best de novo parameter set retained fewer than 100 SNPs when using MAF>0.3 and MAF>0.4 filters. To retain a larger number of SNPs, we used a MAF>0.05 filter for the four de novo parameter sets with the lowest genotype mismatch rates (Figure 4; $M = 2$, $n = 1$, $m = 3$–6, $r = 0.9$). In contrast, for bighorn sheep, the filtering steps retained >100 SNPs for all MAF filters for the optimal de novo parameter set.

More loci were lost to the LD filter in de novo analyses than reference-based analyses. For Mexican gray wolf, the LD filter for de novo analyses removed 63.3%–95.5% of SNPs across panels, compared to a loss of 32.3%–67.7% of SNPs across panels for reference-based analysis. For bighorn sheep, the LD filter for de novo analysis removed 37.4–83.9% of SNPs across panels, compared to a loss of 12.8%–40.9% for reference-based analysis. For both species, one individual had <70% loci genotyped after MAF and LD filters and was removed from subsequent analyses for each parameter set, except one parameter set for Mexican gray wolf (i.e., $m = 6$, $M = 2$, $n = 1$) for which two individuals had <70% loci genotyped and were removed.

Final SNP sets included between 139 and 363 loci for Mexican gray wolf and between 142 and 523 loci for bighorn sheep (Table 2). Mean H_o, mean H_e and nonexclusion probabilities were generally lower for de novo SNP sets than reference-aligned SNP sets for both species (Table 2). This likely results from less stringent MAF filtering for de novo than reference-based SNPs, and therefore lower diversity for de novo SNP panels, as well as the retention of fewer loci for a given MAF cut-off for de novo SNPs.

3.4 | Parentage analysis

For the Mexican gray wolf, field data indicated the full data set included 95% of potential mothers and all potential fathers in the population. However, we used a parameter value of 90% potential parents sampled (for each sex) for CERVUS parentage simulations with the full data set to allow for the possibility of unsampled parents. For the reduced data set, we used parameter values of 64% potential mothers sampled and 53% potential fathers sampled, after taking into account the lower sample size for this data set.

As we knew with high certainty the identities of the true parents of each Mexican gray wolf pup, we could estimate assignment error rates for our genetic parentage analyses. The most common type of error across all locus sets was the misassignment of a parent when the true parent was absent from the data set (Table 3). Microsatellite analysis of the full data set resulted in fewer assignment errors (5.0% errors) than microsatellite analysis of the reduced data set (22.0% errors, Table 3). When comparing the performance of RADseq and microsatellites for the reduced data set, both reference-based and de novo RADseq had consistently lower error rates than microsatellites. For the three least-stringent MAF-filtered SNP sets of the reference-based analyses (MAF > 0.1, MAF > 0.4, MAF > 0.45), the error rate was 5.0%, and for the MAF>0.475 locus set, the error rate was 10.0% (Table 3). For the de novo-assembled

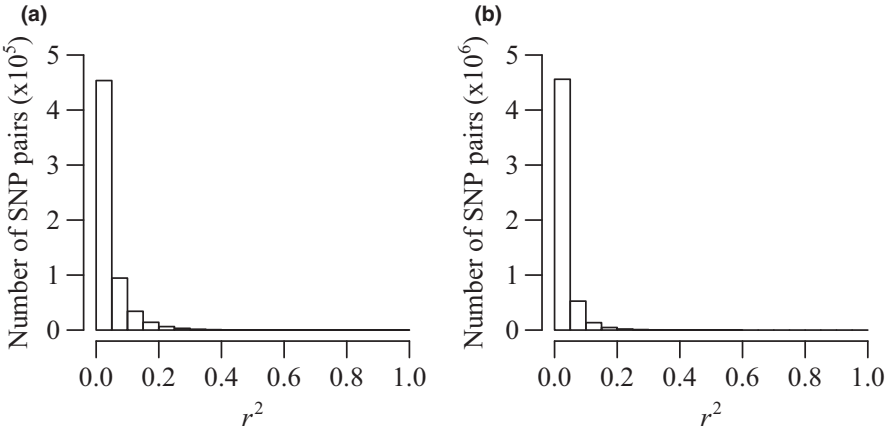

FIGURE 5 Squared allele count correlation (r^2) between all pairs of SNPs generated using de novo RADseq analysis and filtered for HWE and MAF > 0.05 for Mexican gray wolf (a) and bighorn sheep (b)

TABLE 3 Per cent incorrect Mexican gray wolf parentage assignments for different marker sets and sample sets

	# loci	Absent, misassigned (%)	Present, misassigned (%)	Present, unassigned (%)	Total (%)
Microsatellites					
Reduced	22	20.0	2.0	0.0	22.0
Full	22	3.3	1.7	0.0	5.0
RADseq: reference-aligned					
MAF>0.1	1,478	5.0	0.0	0.0	5.0
MAF>0.4	480	5.0	0.0	0.0	5.0
MAF>0.45	292	5.0	0.0	0.0	5.0
MAF>0.475	159	8.3	1.7	0.0	10.0
RADseq: De novo					
m = 3, MAF>0.05	363	3.0	0.0	2.0	5.0
m = 4, MAF>0.05	284	3.3	0.0	0.0	3.3
m = 5, MAF>0.05	223	3.0	3.0	2.0	8.0
m = 6, MAF>0.05	139	8.0	0.0	3.0	11.0

Note. "Absent, misassigned": the true parent was absent from the sample set, and another parent was misassigned. "Present, misassigned": the true parent was present in the sample set, but a different parent was misassigned. "Present, unassigned": the true parent was present in the sample set, but no parent was assigned. Results are shown for the 30 pups present in both the full and reduced sample sets.

TABLE 4 Per cent of bighorn sheep maternity assignments at 80% and 95% confidence levels for different marker sets and sample sets

	Microsatellites		RADseq
Confidence	Reduced (22 loci)	Full (22 loci)	All SNP sets (142–3,044 loci)
95%	28.2%	41.0%	84.6%
80%	64.1%	56.4%	0%
Not assigned	7.7%	2.6%	15.4%

Note. Results are shown for the 39 lambs present in both the full and reduced sample sets. Results were identical for all RADseq SNP sets, including all de novo and reference alignment-based analyses.

SNP sets, the error rate was lowest at m = 3 (5.0%) and m = 4 (3.3%), and then increased as m increased and the number of loci decreased (up to 11.0%).

For bighorn sheep, the number of possible mothers in our full data set for each lamb ranged from 36 to 46 (mean 40.2), and we estimated we had sampled 90% of all potential mothers based on observational field data. For the reduced data set, the number of possible mothers ranged from 30 to 40 (mean 34.7), and we estimated this data set included 77% of all potential mothers. We did not have prior knowledge regarding the true mothers and therefore could not calculate error rates. When comparing the maternity assignment results across markers and sample sets for all lambs that were present in both the full and reduced data sets (n = 39 lambs, after removing one lamb with >70% missing RADseq data), microsatellite analyses resulted in more assignments at the 95% confidence level for the full data set (41.0% lambs) than the reduced data set (28.2% lambs), and fewer lambs were unassigned for the full data set (2.6% for full data set, 7.7% for reduced data set, Table 4). The RADseq analysis resulted in more assignments at the 95% confidence level (84.6% lambs) and more unassignments (15.4%) than either of the microsatellite analyses (Table 4). In fact, RADseq analysis resulted in no assignments at the 80% confidence level. As described above, we did not have samples from all potential mothers in the population, and therefore, unassignments may be correct.

The identities of the mothers that were assigned to each lamb were identical across analyses for all reference-based and de novo RADseq SNP sets, as were the assignment confidence levels. When comparing the RADseq results with the full-data set microsatellite results, there were five (12.8%) assignment disagreements for the 39 offspring that were present in both the full and reduced data sets. In all five cases, different mothers were assigned for the analyses with the two marker types, despite the fact that both mothers were present in both data sets. When comparing the RADseq results with the reduced data set microsatellite results, there were 10 (25.6%) assignment disagreements. These included the same disagreements as in the full data set comparison, as well as four (10.3%) disagreements in which RADseq did not assign a mother, but the microsatellites did (although these were all assigned at 80% confidence), and one (2.6%) disagreement in which the microsatellites did not assign a mother, but RADseq did.

3.5 | Mechanisms underlying parentage analysis performance

Delta scores were consistently higher for all RADseq SNP sets than microsatellites for both species (Figure 6), indicating SNPs had greater statistical power to distinguish correct from incorrect parents. These results were supported by a substantial decrease in

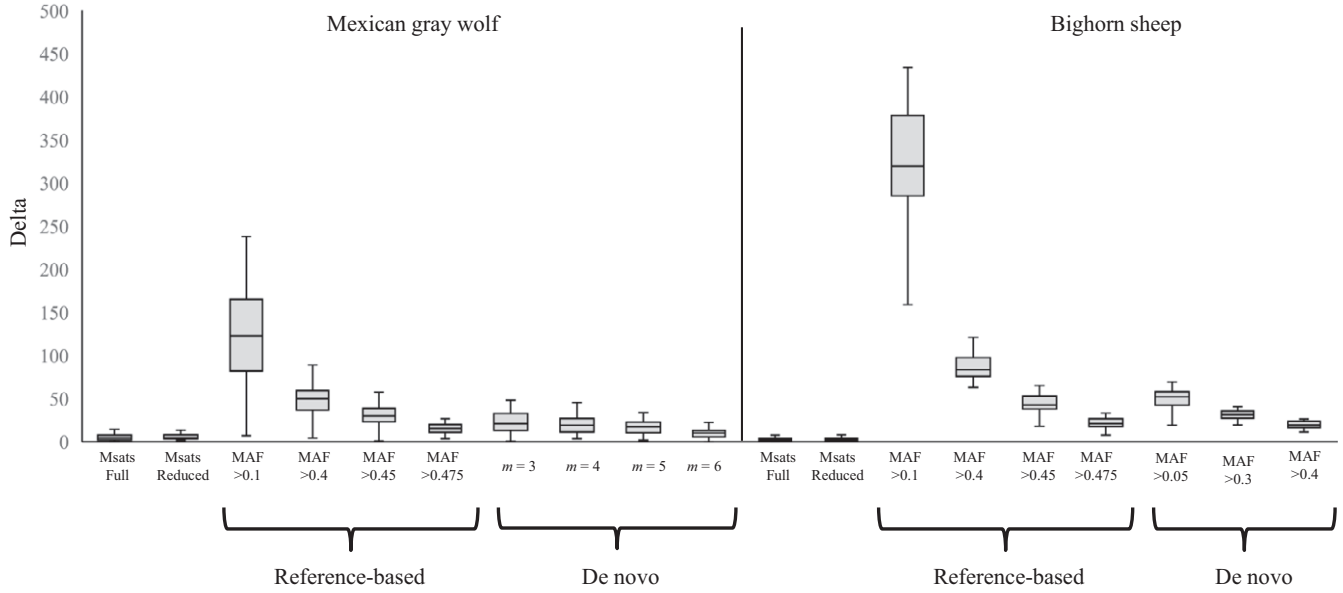

FIGURE 6 Delta values for Mexican gray wolf combined maternal and paternal assignments accepted at 95% confidence, and bighorn sheep maternal assignments accepted at 80% confidence. "Msats full" = microsatellite analysis of full data set; "Msats reduced" = microsatellite analysis of reduced data set. "MAF" refers to minor allele frequency filtered RADseq SNP subsets, and "*m*" refers to the minimum number of identical raw reads to create a stack within an individual. See Table 2 for the numbers of loci for each analysis

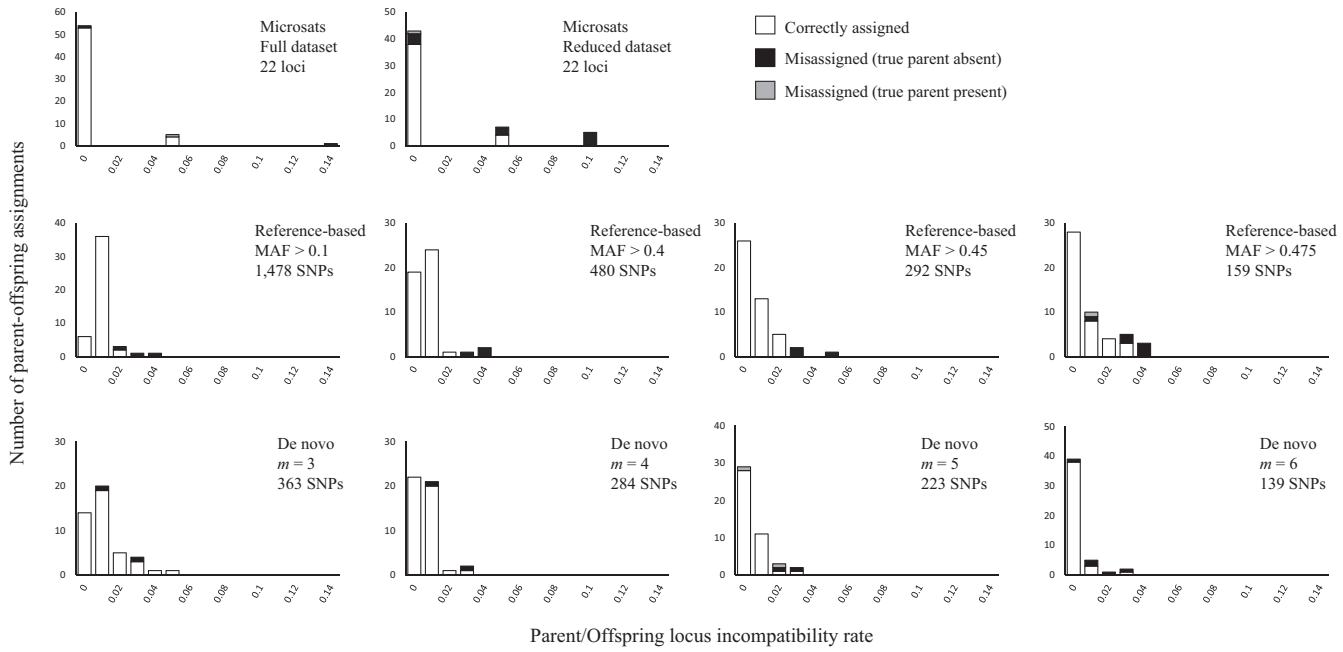

FIGURE 7 Locus incompatibility rates for all parent/offspring assignments accepted at 95% confidence for Mexican gray wolf using microsatellites for full and reduced data sets, and RADseq markers for the reduced data set. Incompatibility rates for correct assignments are in white, misassignments when the true parent was absent are in black, and misassignments when the true parent was present are in gray

nonexclusion probabilities for SNPs compared to microsatellites (Table 2). Delta values decreased as the number of RADseq SNPs decreased, and delta values for bighorn sheep were consistently higher than for Mexican gray wolf. Notably, however, the number of loci retained at each MAF filter was also higher for bighorn sheep, which likely leads to increased statistical power (Table 2).

For Mexican gray wolf, prior knowledge of the true parents allowed a comparison of locus incompatibility rates for correct and incorrect parentage assignments (Figure 7). For microsatellite analyses and the two de novo SNP panels with the lowest numbers of loci ($m = 5$ and $m = 6$), incorrect assignments sometimes had incompatibility rates equal to zero (Figure 7). When incorrect assignments

have zero incompatibilities, CERVUS can only successfully exclude the incorrect parent using population allele frequency data. Therefore, the success of this analysis will depend on the accuracy of population allele frequency estimates, as well as the utility of the allele frequencies for distinguishing between potential parents; increased accuracy of allele frequencies likely explains the substantially improved performance of microsatellites for the full data set compared to the reduced data set. In contrast, incorrect assignments in the rest of the RADseq analyses never had locus incompatibility rates of zero, and in most cases, the incorrect assignments had more incompatibilities than any correct assignments, especially for the reference-based analyses (Figure 7). Overall, these results indicate that high assignment error rates for the reduced data set microsatellites and the two de novo SNP panels with the lowest numbers of loci are likely influenced by a lack of diversity to distinguish correct from incorrect assignments.

4 | DISCUSSION

4.1 | RADseq bioinformatic pipeline identifies informative SNP panels for parentage analysis

Our RADseq bioinformatic pipeline identified SNP panels with high power and accuracy for parentage assignment when using both the reference-based and de novo options. For all Mexican gray wolf and bighorn sheep SNP panels, nonexclusion probabilities (Table 2) and delta values (Figure 6) indicated high statistical power, and higher statistical power than for microsatellites. For the Mexican gray wolf study system, we had prior knowledge of the true identities of the parents and therefore could directly evaluate the accuracy of the SNP panels for parentage analysis. Assignment accuracy was 95% or higher across most SNP panels, despite the presence of many full siblings and low diversity for this study system. This was much higher than the accuracy for 22 microsatellites (78% for the "reduced data set" which includes the same samples as the RADseq data set). The performance of the SNP panels only decreased when the number of SNPs was lower than 284.

For the bighorn sheep study system, we did not know the true mothers a priori and therefore could not directly assess accuracy. However, maternity assignments were identical across all RADseq SNP panels, indicating no reduction in power or accuracy for as few as 142 SNPs. Furthermore, the proportion of assigned lambs (84.6%) was consistent with our expectation based on the estimated proportion of potential dams sampled (77%). We found evidence that the SNP panels had higher power than microsatellites for this species; for microsatellites, only 28.2% of lambs assigned a mother with 95% confidence, and an additional 64.1% of lambs assigned a mother with 80% confidence (Table 4). In contrast, RADseq analyses either assigned mothers at 95% confidence or did not assign a mother at all. The identities of the assigned mothers were largely consistent across the two marker types (74.4% agreement for RADseq and reduced data set microsatellites, and 87.2% agreement for RADseq and full-data set microsatellite results; the microsatellite full-data set had a larger sample size than the RADseq data set).

Comparison of delta scores, nonexclusion probabilities, and locus incompatibility rates across locus and sample sets provided evidence that the reduced power and accuracy for microsatellites compared to SNPs was driven by low overall genetic diversity for the locus set. This low diversity likely constrained the abilities to distinguish between potential parents and to exclude an incorrect parent when the true parent was absent from the data set.

4.2 | Comparing performance of de novo and reference-based RADseq pipelines

Genotype mismatch rates were slightly lower for de novo than reference-based analyses, indicating high genotyping accuracy for this approach. However, de novo analysis with optimal parameters resulted in almost four times fewer SNPs than reference-based analysis for both species. In addition, a greater proportion of loci was lost to the LD filter in de novo analysis than reference-based analysis. This is not surprising, given that our de novo LD filter cannot incorporate information regarding physical positions of SNPs along the genome (aside from information as to whether SNPs are located on the same RAD locus), and thus, many physically unlinked loci were likely lost in the de novo LD filter. Because few SNPs remained after the LD filter for both species, and especially for the Mexican gray wolf, we used low MAF cut-offs to retain sufficient numbers of SNPs. Thus, the de novo SNP panels had lower diversity and lower statistical power than the reference-based panels for both species. However, this reduced power had no impact on the accuracy of parentage assignment for the Mexican gray wolf panels with ≥284 SNPs or any of the bighorn sheep panels, and minimal impact on accuracy for Mexican gray wolf panels with ≤223 SNPs (described further in the next section).

4.3 | Comparing performance of different RADseq SNP sets

To identify a subset of informative RADseq SNPs that could be used to develop a time- and cost-effective assay for large sample sizes, we targeted SNPs with the highest diversity, as markers with higher diversity should have higher power for parentage analysis. To accomplish this, we conducted a series of MAF filters to determine the lowest number of high-diversity loci that would provide sufficient power and accuracy for parentage analyses (reference-based analysis: MAF>0.1, >0.4, >0.45, >0.475; de novo analysis for bighorn sheep: MAF>0.05, >0.3, >0.4; for Mexican gray wolf de novo analysis, we only used MAF>0.05 due to low overall numbers of SNPs, as described above). These filters retained 139–1,478 SNPs for Mexican gray wolf and 142–3,044 SNPs for bighorn sheep. For bighorn sheep, all SNP subsets produced identical results in terms of the identities of the mothers assigned and the confidence levels. For Mexican gray wolf, results for all SNP subsets had ≥95% assignment accuracy except for the SNP panels with the lowest numbers of loci

(reference-based analysis: 159 SNPs, 90% accuracy; de novo analysis: 139–223 SNPs, 89%–92% accuracy) (Table 3). Overall, these results indicate that the performance of parentage analysis was not compromised using as few as 284 SNPs for Mexican gray wolf and 142 SNPs for bighorn sheep.

4.4 | Choosing a bioinformatic pipeline

The optimal bioinformatic pipeline for discovering and genotyping RADseq SNPs for parentage analysis will depend on many factors, including the RADseq protocol chosen (see Andrews et al., 2016 for a review of the many RADseq protocols), depth of sequencing, availability of a reference genome and various characteristics of the study system, including the size, diversity, complexity and linkage patterns of the genome. For both our study species, reference genomes were available from closely related species, and thus, our bioinformatic pipeline incorporates the option of reference alignment. Performing a strict mapping quality filter with BOWTIE2 (MQ≥40) allowed us to filter out loci that mapped poorly to the genome, or mapped to multiple locations in the genome, which could indicate paralogs. Furthermore, alignment to a reference genome helped us to identify and filter linked loci. If a reference genome were not available, a genome from a more distantly related species could also be used, although a less stringent mapping quality filter may be necessary. Alternatively, we demonstrated here that loci can be assembled de novo without a reference genome, and linked loci can be identified based on allele correlations.

Bioinformatic pipelines should also be tailored to maximize genotyping accuracy, which is particularly important for parentage analyses, especially in study systems with low diversity and large numbers of close relatives (Hoffman & Amos, 2005). For reference-based analyses, two parameters that strongly impact genotyping error rates for RADseq are the minimum depth of coverage to accept a locus, and minimum per cent individuals genotyped to accept a locus. For de novo analyses, additional relevant parameters include the maximum numbers of genotype mismatches allowed when merging reads into one locus within and between individuals (Catchen et al., 2013). The optimal values for these parameters are likely to vary across RADseq protocols and study systems, but these values are rarely chosen based on empirical evidence (but see Catchen et al., 2013; Mastretta-Yanes et al., 2015; Fountain, Pauli, Reid, Palsboll, & Peery, 2016; Paris, Stevens, & Catchen, 2017). Replicate comparison has been used to choose parameters for de novo assembly of RADseq loci (Mastretta-Yanes et al., 2015), and here, we show that this strategy can also be used to choose parameters for reference-based RADseq analysis. Notably, alternative methods have been described for choosing optimal parameters for de novo assembly, including the comparison of sequence data from known parents and offspring (Fountain et al., 2016) and comparison of the numbers of loci obtained across a variety of parameters (Paris et al., 2017).

Bioinformatic analyses to identify informative loci for parentage analysis should ideally ensure that the loci chosen are not linked, because parentage analyses typically assume loci are independent (Jones & Ardren, 2003), although some methods are relatively robust to linkage or do not assume linkage (e.g., Staples et al., 2014; Wang & Santure, 2009). Different populations and species will have different levels of LD depending on a number of factors including demographic history, mating system and recombination rate (Gaut & Long, 2003; Gray et al., 2009; Miller, Poissant, Malenfant, Hogg, & Coltman, 2015; Pritchard & Przeworski, 2001). Therefore, optimal methods for identifying a set of unlinked loci will vary across study systems. Here, we demonstrate that RADseq data can be used to characterize LD patterns both with and without a reference genome, and this information can then be used to tailor bioinformatic analyses to identify a subset of unlinked loci for the study system.

4.5 | Designing SNP assays

RADseq data can be used to design SNP assays that rely on technologies like amplicon sequencing, DNA capture, Fluidigm Dynamic Array and MassARRAY. These approaches are more time- and cost-effective than RADseq for genotyping small numbers of loci for large numbers of samples, but require prior genomic information to design primers and/or probes. Under some circumstances, RADseq data can be used directly to design these primers and probes. RADseq data are typically generated with Illumina HiSeq technology, which generates sequence reads up to 150 bp long. This length is sufficient for designing probes for a DNA capture approach (Ali et al., 2016), but not for designing primers for PCR-based approaches like amplicon sequencing, Fluidigm Dynamic Array, and MassARRAY. However, if a reference genome is available, as for Mexican gray wolf and bighorn sheep, RADseq reads can be aligned to the reference, and then primers and probes can be designed directly from the reference sequence. If a reference genome is not available, probes can be developed from the forward and/or reverse reads, or primers may be developed by assembling the forward and reverse reads, provided these reads have substantial overlap. This approach would be most tractable for RADseq methods that allow assembly of forward and reverse reads into long contigs several hundred bases long, as in Hohenlohe et al. (2013). This long contig-assembly approach is only possible when using RADseq methods that use mechanical shearing to generate fragments of variable length for each RAD locus (i.e., Baird et al. 2008, Ali et al., 2016), and is not possible for double digest RAD (ddRAD, Peterson, Weber, Kay, Fisher, & Hoekstra, 2012), Genotyping by Sequencing (GBS, Elshire et al., 2011), or many other RADseq techniques (see Andrews et al., 2016). However, some of these RADseq approaches may allow the assembly of small contigs that can be used for primer development, if there is considerable overlap between the forward and reverse sequence reads (e.g., Jacobsen et al., 2017).

Another consideration when designing SNP assays is the number of loci needed. Several empirical studies have evaluated the numbers of SNPs required for parentage analysis, demonstrating that the minimum requirement varies across study systems (e.g., Holman et al., 2017; Kaiser et al., 2017; Tokarska et al., 2009; Weinman et al., 2015). One important factor influencing the minimum required loci is the mating system; for example, mating systems with high

proportions of close relatives, such as the Mexican gray wolf, should require more loci for sufficient power to distinguish between putative parents that are close relatives. Another factor influencing the number of required loci is the diversity of the loci; the more diverse the loci, the higher the statistical power, and therefore, the fewer loci are needed (Anderson & Garza, 2006; Morin et al., 2004). For species and populations with high genetic diversity, it may be easier to discover high-diversity SNPs and therefore easier to design a smaller SNP panel with high statistical power. The thoroughness of sampling of candidate parents is also important, as shown here by the much lower accuracy of microsatellite parentage analysis for the Mexican gray wolf data set that had many unsampled candidate parents (i.e., comparing performance of the full vs. reduced data set). Thus far, most studies evaluating the number of SNPs required for parentage analysis have used data sets with few missing parents (but see Trong, van Bers, Crooijmans, Dibbits, & Komen, 2013), and therefore, our study provides a unique perspective on the performance of SNPs when 23%–47% of putative parents are missing from the data set. Finally, the amount of prior knowledge of parentage is another important factor influencing the minimum loci needed. For example, in many parentage studies the mothers of offspring are known a priori, and incorporating that information into the parentage analysis will increase statistical power and decrease the number of loci required (e.g., Kaiser et al., 2017; Weinman et al., 2015).

4.6 | Choosing markers for parentage analysis

We demonstrate here that RADseq is an effective tool for parentage analysis, and provides greater power and accuracy than 14–22 microsatellites for our two study systems. The cost of supplies for our RADseq analysis was approximately US$5.00 per sample for library prep, and US$32.00 per sample for sequencing (about 80 samples per HiSeq lane). In contrast, the cost of supplies for our microsatellite analysis was approximately US$8.00 per sample for the multiplex PCR and fragment length analysis; this does not include the cost of supplies for marker development and protocol optimization. However, after the bioinformatic pipeline had been optimized, RADseq laboratory and bioinformatic analysis were substantially less time-consuming than microsatellite genotyping. Furthermore, RADseq analysis requires no extra time or expense for locus discovery, as RADseq simultaneously discovers and genotypes loci. However, for projects with large sample sizes, an approach of using RADseq to identify a small panel of loci for SNP assay development will likely be more cost-effective than using RADseq directly for parentage analysis. RADseq typically generates data from many more genetic markers than is necessary for parentage analysis, and requires substantially more hard drive space and computational power for analysis than microsatellites or SNP assays. In addition, if very large sample sizes are used for RADseq studies, higher depth of coverage (and therefore greater sequencing costs and hard drive space) may be required than was used in this study to ensure that sufficient numbers of loci pass the "per cent individuals genotyped" filter. The laboratory cost of SNP genotyping assays varies widely across approaches, but is generally comparable to the costs of RADseq. The laboratory and bioinformatic analyses for SNP genotyping assays are considerably less time-consuming than for microsatellites or RADseq, and the genotypes generated using the same assay method across laboratories should be highly comparable.

4.7 | Conclusions

Here, we describe a bioinformatic pipeline for identifying informative RADseq SNP panels for parentage analysis with or without a reference genome, and test the performance of these panels for small populations with high proportions of close relatives for two different species. Our pipeline identified SNP panels with higher power and accuracy for parentage analysis than 14–22 microsatellite loci for both species. Subsets of 284 SNPs for Mexican gray wolf and 142 SNPs for bighorn sheep provided parentage analysis results consistent with results generated using more than >1,000 SNPs, indicating that RADseq can be used to discover SNPs for designing time- and cost-effective assays to genotype small numbers of loci for large numbers of samples. More RADseq SNPs were needed for parentage analysis of Mexican gray wolf than bighorn sheep, likely due to lower diversity in the Mexican gray wolf population resulting from a small founding population size, as well as a mating system that results in a greater proportion of close relatives. Our pipeline incorporates methods for optimizing bioinformatic parameters to maximize genotyping accuracy, remove linked loci and select loci with high statistical power and therefore can be used across a wide range of study systems.

ACKNOWLEDGEMENTS

We thank Meaghan Clark, Brendan Epstein, Digpal Gour, Sarah Hendricks, Michelle Keyes, Mitch Kissler and Amanda Stahlke for assistance with laboratory work and data analysis. We thank the Mexican Wolf Interagency Field Team and Oregon Department of Fish and Wildlife for assistance with sample collection. This work was funded by the U.S. Fish and Wildlife Service, Oregon Department of Fish and Wildlife, Morris Animal Foundation (grant D13ZO-081), Montana University System Research Initiative (51040-MUSRI2015-03), National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) under Award Number P20GM103474, P30GM110732 and P30GM103324, NSF grant DEB-1316549, University of Idaho College of Natural Resources, and University of Idaho IBEST. The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the U.S. Fish and Wildlife Service. This is PMEL contribution number 4785, and Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement NA15OAR4320063 contribution number 2018-0147.

DATA ACCESSIBILITY

RADseq sequence data are available at the NCBI Short Read Archive (BioProject ID PRJNA454718, SRA accession SRP144608).

These sequences were demultiplexed (separated by barcode into individuals) and quality filtered using process_radtags in STACKS. For each individual, forward and reverse reads are provided, as well as remainder reads that were removed from further analysis. For individuals that were sequenced more than once, sequence data from each replicate are provided separately (i.e., replicates are not merged).

AUTHOR CONTRIBUTIONS

All authors contributed to study design. K.R.A. generated and analysed RADseq data. J.R.A. generated microsatellite data; J.R.A. and K.R.A. analysed microsatellite data. E.F.C., R.K.P., C.G. and M.D. collected samples. P.A.H. assisted with RADseq data analysis. K.R.A. led the writing of the article, and all authors contributed to the writing.

ORCID


Kimberly R. Andrews http://orcid.org/0000-0003-4721-1924

REFERENCES

Adams, J. R., Kelly, B. T., & Waits, L. P. (2003). Using faecal DNA sampling and GIS to monitor hybridization between red wolves (*Canis rufus*) and coyotes (*Canis latrans*). *Molecular Ecology*, *12*, 2175–2186. https://doi.org/10.1046/j.1365-294X.2003.01895.x

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD capture (rapture): Flexible and efficient sequence-based genotyping. *Genetics*, *202*, 389–400. https://doi.org/10.1534/genetics.115.183665

Anderson, E. C., & Garza, J. C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, *172*, 2567–2582.

Andrews, K., Good, J., Miller, M., Luikart, G., & Hohenlohe, P. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*, 81–92. https://doi.org/10.1038/nrg.2015.28

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A. … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *Plos One*, *3*, e3376.

Bonin, A., Bellemain, E., Eidesen, P. B., Pompanon, F., Brochmann, C., & Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, *13*, 3261–3273. https://doi.org/10.1111/j.1365-294X.2004.02346.x

Brumfield, R. T., Beerli, P., Nickerson, D. A., & Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, *18*, 249–256. https://doi.org/10.1016/S0169-5347(03)00018-1

Buechner, H. K. (1960). The bighorn sheep in the United States, its past, present, and future. *Wildlife Monographs*, *4*, 1–174.

Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, *15*, 855–867. https://doi.org/10.1111/1755-0998.12357

Cassirer, E. F., Plowright, R. K., Manlove, K. R., Cross, P. C., Dobson, A. P., Potter, K. A., & Hudson, P. J. (2013). Spatio-temporal dynamics of pneumonia in bighorn sheep. *Journal of Animal Ecology*, *82*, 518–528. https://doi.org/10.1111/1365-2656.12031

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

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., & Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, *15*, 1496–1502. https://doi.org/10.1101/gr.4107905

Coggins, V. L. (2006) Selenium supplementation, parasite treatment, and management of bighorn sheep at Lostine River, Oregon. *Bienn. Symp. North. Wild Sheep 15*, 98–106.

Constable, J. L., Ashley, M. V., Goodall, J., & Pusey, A. E. (2001). Noninvasive paternity assignment in Gombe chimpanzees. *Molecular Ecology*, *10*, 1279–1300. https://doi.org/10.1046/j.1365-294X.2001.01262.x

D'Aloia, C. C., Bogdanowicz, S. M., Francis, R. K., Majoris, J. E., Harrison, R. G., & Buston, P. M. (2015). Patterns, causes, and consequences of marine larval dispersal. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 13940–13945. https://doi.org/10.1073/pnas.1513754112

De Wit, P., Pespeni, M. H., & Palumbi, S. R. (2015). SNP genotyping and population genomics from expressed sequences - Current advances and future possibilities. *Molecular Ecology*, *24*, 2310–2323. https://doi.org/10.1111/mec.13165

DiBattista, J. D., Feldheim, K. A., Garant, D., Gruber, S. H., & Hendry, A. P. (2009). Evolutionary potential of a large marine vertebrate: Quantitative genetic parameters in a wild population. *Evolution*, *63*, 1051–1067. https://doi.org/10.1111/j.1558-5646.2008.00605.x

Douhard, M., Festa-Bianchet, M., & Pelletier, F. (2016). Maternal condition and previous reproduction interact to affect offspring sex in a wild mammal. *Biology Letters*, *12*, pii: 20160510.

Dugdale, H. L., Macdonald, D. W., Pope, L. C., & Burke, T. (2007). Polygynandry, extra-group paternity and multiple-paternity litters in European badger (*Meles meles*) social groups. *Molecular Ecology*, *16*, 5294–5306. https://doi.org/10.1111/j.1365-294X.2007.03571.x

Dunn, S. J., Clancey, E., Waits, L. P., & Byers, J. A. (2011). Inbreeding depression in pronghorn (*Antilocapra americana*) fawns. *Molecular Ecology*, *20*, 4889–4898. https://doi.org/10.1111/j.1365-294X.2011.05327.x

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*, 1026–1042. https://doi.org/10.1111/eva.12178

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*, e19379. https://doi.org/10.1371/journal.pone.0019379

Fish and Wildlife Service. (2010). Mexican Wolf Conservation Assessment. Region 2, Albuquerque, New Mexico, USA.

Fish and Wildlife Service, Interior (2015). Endangered and threatened wildlife and plants; endangered status for the Mexican wolf. *Federal Register*, *80*(11), 2488–2512.

Fountain, E. D., Pauli, J. N., Reid, B. N., Palsboll, P. J., & Peery, M. Z. (2016). Finding the right coverage: The impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Molecular Ecology Resources*, *16*, 966–978. https://doi.org/10.1111/1755-0998.12519

Gaut, B. S., & Long, A. D. (2003). The lowdown on linkage disequilibrium. *Plant Cell*, *15*, 1502–1506. https://doi.org/10.1105/tpc.150730

Glaubitz, J. C., Rhodes, O. E., & Dewoody, J. A. (2003). Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology*, *12*, 1039–1047. https://doi.org/10.1046/j.1365-294X.2003.01790.x

Gray, M. M., Granka, J. M., Bustamante, C. D., Sutter, N. B., Boyko, A. R., Zhu, L., … Wayne, R. K. (2009). Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics*, *181*, 1493–1505. https://doi.org/10.1534/genetics.108.098830

Hauser, L., Baird, M., Hilborn, R., Seeb, L. W., & Seeb, J. E. (2011). An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. *Molecular Ecology Resources*, *11*, 150–161. https://doi.org/10.1111/j.1755-0998.2010.02961.x

Hoffman, J. I., & Amos, W. (2005). Microsatellite genotyping errors: Detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, *14*, 599–612.

Hogg, J. T., & Forbes, S. H. (1997). Mating in bighorn sheep: Frequent male reproduction via a high-risk "unconventional" tactic. *Behavioral Ecology and Sociobiology*, *41*, 33–48. https://doi.org/10.1007/s002650050361

Hogg, J. T., Forbes, S. H., Steele, B. M., & Luikart, G. (2006). Genetic rescue of an insular population of large mammals. *Proceedings of the Royal Society B-Biological Sciences*, *273*, 1491–1499. https://doi.org/10.1098/rspb.2006.3477

Hohenlohe, P. A., Day, M. D., Amish, S. J., Miller, M. R., Kamps-Hughes, N., Boyer, M. C., … Luikart, G. (2013). Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, *22*, 3002–3013. https://doi.org/10.1111/mec.12239

Holman, L. E., de la Serrana, D. G., Onoufriou, A., Hillestad, B., & Johnston, I. A. (2017). A workflow used to design low density SNP panels for parentage assignment and traceability in aquaculture species and its validation in Atlantic salmon. *Aquaculture*, *476*, 59–64. https://doi.org/10.1016/j.aquaculture.2017.04.001

Jacobsen, M. W., Christensen, C., Madsen, R., Nygaard, R., Jónsson, B., Præbel, K., & Hansen, M. M. (2017). Single nucleotide polymorphism markers for analysis of historical and contemporary samples of Arctic char (*Salvelinus alpinus*). *Conservation Genetics Resources*, *9*, 587–589. https://doi.org/10.1007/s12686-017-0728-y

Janeiro, M. J., Coltman, D. W., Festa-Bianchet, M., Pelletier, F., & Morrissey, M. B. (2017). Towards robust evolutionary inference with integral projection models. *Journal of Evolutionary Biology*, *30*, 270–288. https://doi.org/10.1111/jeb.13000

Jones, A. G., & Ardren, W. R. (2003). Methods of parentage analysis in natural populations. *Molecular Ecology*, *12*, 2511–2523. https://doi.org/10.1046/j.1365-294X.2003.01928.x

Jones, M., & Good, J. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*, 185–202. https://doi.org/10.1111/mec.13304

Jones, A. G., Kvarnemo, C., Moore, G. I., Simmons, L. W., & Avise, J. C. (1998). Microsatellite evidence for monogamy and sex-biased recombination in the Western Australian seahorse *Hippocampus angustus*. *Molecular Ecology*, *7*, 1497–1505. https://doi.org/10.1046/j.1365-294x.1998.00481.x

Jones, A. G., Small, C. M., Paczolt, K. A., & Ratterman, N. L. (2010). A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, *10*, 6–30. https://doi.org/10.1111/j.1755-0998.2009.02778.x

Kaiser, S. A., Taylor, S. A., Chen, N., Sillett, T. S., Bondra, E. R., & Webster, M. S. (2017). A comparative assessment of SNP and microsatellite markers for assigning parentage in a socially monogamous bird. *Molecular Ecology Resources*, *17*, 183–193. https://doi.org/10.1111/1755-0998.12589

Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, *16*, 1099–1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x

Kraus, R. H. S., vonHoldt, B., Cocchiararo, B., Harms, V., Bayerl, H., Kühn, R., … Nowak, C. (2015). A single-nucleotide polymorphism-based approach for rapid and cost-effective genetic wolf monitoring in Europe based on noninvasively collected samples. *Molecular Ecology Resources*, *15*, 295–305. https://doi.org/10.1111/1755-0998.12307

Lachance, J., & Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays*, *35*, 780–786. https://doi.org/10.1002/bies.201300014

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–354. https://doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAM tools. *Bioinformatics*, *25*, 2078–2079.

López-Herráez, D., Schafer, H., Mosner, J., Fries, H. R., & Wink, M. (2005). Comparison of microsatellite and single nucleotide polymorphism markers for the genetic analysis of a Galloway cattle population. *Journal of Biosciences*, *60C*, 637–643.

Marker, L. L., Wilkerson, A. J. P., Sarno, R. J., Martenson, J., Breitenmoser-Würsten, C., O'Brien, S. J., & Johnson, W. E. (2008). Molecular genetic insights on cheetah (*Acinonyx jubatus*) ecology and conservation in Namibia. *Journal of Heredity*, *99*, 2–13. https://doi.org/10.1093/jhered/esm081

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, *15*, 28–41. https://doi.org/10.1111/1755-0998.12291

Miller, J. M., Poissant, J., Malenfant, R. M., Hogg, J. T., & Coltman, D. W. (2015). Temporal dynamics of linkage disequilibrium in two populations of bighorn sheep. *Ecology and Evolution*, *5*, 3401–3412. https://doi.org/10.1002/ece3.1612

Morin, P. A., Luikart, G., Wayne, R. K., & Grp, S. N. P. W. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, *19*, 208–216. https://doi.org/10.1016/j.tree.2004.01.009

Nguyen, T. T. T., Hayes, B. J., & Ingram, B. A. (2014). Genetic parameters and response to selection in blue mussel (*Mytilus galloprovincialis*) using a SNP-based pedigree. *Aquaculture*, *420*, 295–301. https://doi.org/10.1016/j.aquaculture.2013.11.021

Olson, Z. H., Whittaker, D. G., & Rhodes, O. E. (2013). Translocation history and genetic diversity in reintroduced bighorn sheep. *Journal of Wildlife Management*, *77*, 1553–1563. https://doi.org/10.1002/jwmg.624

Paris, J. R., Stevens, J. R., & Catchen, J. (2017). Lost in parameter space: A road map for STACKS. *Methods in Ecology and Evolution*, *8*, 1360–1373. https://doi.org/10.1111/2041-210X.12775

Pemberton, J. (2004). Measuring inbreeding depression in the wild: The old ways are the best. *Trends in Ecology & Evolution*, *19*, 613–615. https://doi.org/10.1016/j.tree.2004.09.010

Pemberton, J. M. (2008). Wild pedigrees: The way forward. *Proceedings of the Royal Society B-Biological Sciences*, *275*, 613–621. https://doi.org/10.1098/rspb.2007.1531

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*, e37135. https://doi.org/10.1371/journal.pone.0037135

Plowright, R. K., Manlove, K. R., Besser, T. E., Paez, D. J., Andrews, K. R., Matthews, P. E., … Cassirer, E. F. (2017). Age-specific infectious period shapes dynamics of pneumonia in bighorn sheep. *Ecology Letters*, *20*, 1325–1336.

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: Causes, consequences and solutions. *Nature Reviews Genetics*, *6*, 847–859. https://doi.org/10.1038/nrg1707

Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics*, *69*, 1–14. https://doi.org/10.1086/321275

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575. https://doi.org/10.1086/519795

Rudnick, J. A., Katzner, T. E., Bragin, E. A., Rhodes, O. E., & Dewoody, J. A. (2005). Using naturally shed feathers for individual identification, genetic parentage analyses, and population monitoring in an endangered Eastern imperial eagle (*Aquila heliaca*) population from Kazakhstan. *Molecular Ecology*, *14*, 2959–2967. https://doi.org/10.1111/j.1365-294X.2005.02641.x

Staples, J., Qiao, D., Cho, M. H., Silverman, E. K., University of Washington Center for Mendelian Genomics, Nickerson, D. A., & Below, J. E. (2014). PRIMUS: Rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *American Journal of Human Genetics*, *95*, 553–564. https://doi.org/10.1016/j.ajhg.2014.10.005

Steyer, K., Kraus, R. H. S., Mölich, T., Anders, O., Cocchiararo, B., Frosch, C., … Nowak, C. (2016). Large-scale genetic census of an elusive carnivore, the European wildcat (Felis s. silvestris). *Conservation Genetics*, *17*, 1183–1199. https://doi.org/10.1007/s10592-016-0853-2

Tokarska, M., Marshall, T., Kowalczyk, R., Wójcik, J. M., Pertoldi, C., Kristensen, T. N., … Bendixen, C. (2009). Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: The case of European bison. *Heredity*, *103*, 326–332. https://doi.org/10.1038/hdy.2009.73

Trong, T. Q., van Bers, N., Crooijmans, R., Dibbits, B., & Komen, H. (2013). A comparison of microsatellites and SNPs in parental assignment in the GIFT strain of Nile tilapia (Oreochromis niloticus): The power of exclusion. *Aquaculture*, *388*, 14–23. https://doi.org/10.1016/j.aquaculture.2013.01.004

Wang, J., & Santure, A. W. (2009). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics*, *181*, 1579–1594. https://doi.org/10.1534/genetics.108.100214

Weinman, L. R., Solomon, J. W., & Rubenstein, D. R. (2015). A comparison of single nucleotide polymorphism and microsatellite markers for analysis of parentage and kinship in a cooperatively breeding bird. *Molecular Ecology Resources*, *15*, 502–511. https://doi.org/10.1111/1755-0998.12330

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Andrews KR, Adams JR, Cassirer EF, et al. A bioinformatic pipeline for identifying informative SNP panels for parentage assignment from RADseq data. *Mol Ecol Resour*. 2018;18:1263–1281. https://doi.org/10.1111/1755-0998.12910