



Microhaplotypes provide increased power from short-read DNA sequences for relationship inference

Diana S. Baetscher^{1,2} | Anthony J. Clemento^{2,3} | Thomas C. Ng^{2,4} | Eric C. Anderson^{2,3}  | John C. Garza^{1,2,3} 

¹Department of Ocean Sciences, University of California, Santa Cruz, CA, USA

²Southwest Fisheries Science Center, National Marine Fisheries Service, Santa Cruz, CA, USA

³Institute of Marine Sciences, University of California, Santa Cruz, CA, USA

⁴Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA

Correspondence

John C. Garza, Southwest Fisheries Science Center, National Marine Fisheries Service, Santa Cruz, CA, USA.

Email: carlos.garza@noaa.gov

Funding information

Division of Ocean Sciences, Grant/Award Number: 1260693

Abstract

The accelerating rate at which DNA sequence data are now generated by high-throughput sequencing instruments provides both opportunities and challenges for population genetic and ecological investigations of animals and plants. We show here how the common practice of calling genotypes from a single SNP per sequenced region ignores substantial additional information in the phased short-read sequences that are provided by these sequencing instruments. We target sequenced regions with multiple SNPs in kelp rockfish (*Sebastes atrovirens*) to determine “microhaplotypes” and then call these microhaplotypes as alleles at each locus. We then demonstrate how these multi-allelic marker data from such loci dramatically increase power for relationship inference. The microhaplotype approach decreases false-positive rates by several orders of magnitude, relative to calling bi-allelic SNPs, for two challenging analytical procedures, full-sibling and single parent–offspring pair identification. We also show how the identification of half-sibling pairs requires so much data that physical linkage becomes a consideration, and that most published studies that attempt to do so are dramatically underpowered. The advent of phased short-read DNA sequence data, in conjunction with emerging analytical tools for their analysis, promises to improve efficiency by reducing the number of loci necessary for a particular level of statistical confidence, thereby lowering the cost of data collection and reducing the degree of physical linkage amongst markers used for relationship estimation. Such advances will facilitate collaborative research and management for migratory and other widespread species.

KEYWORDS

high-throughput DNA sequencing, microhaplotype, parentage, population genetics, relationship inference

1 | INTRODUCTION

The proliferation of individual-based population genetic methods in ecology and evolution has led to a commensurate demand for increasing analytical power. The identification of first-order relatives, including parents and offspring, or full-siblings, is now commonplace in the study of animals and plants, with genotypes

serving both to identify relationships and as elements of larger data aggregations used in the estimation of population genetic parameter values. As the demands of such analyses grow, and extend to more difficult problems of relationship estimation, making optimal use of the data from high-throughput DNA sequencers is critical to achieving strong inference at low cost and with wide availability.

High-throughput sequencing technologies have dramatically increased the rate of data generation, making collection of data for genetic analysis cheaper and less time-consuming. Methodological advances in both generating and analysing these high-throughput sequencing data have made it more feasible to address difficult biological questions (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Kidd et al., 2014; McCormack, Hird, Zellmer, Carstens, & Brumfield, 2013; McKinney, Seeb, & Seeb, 2017). One such area of investigation that has benefited from these technological developments is the identification of family relationships and pedigree reconstruction. Since the inception of genetically informed relationship inference, a half-century ago, researchers have applied a number of different molecular markers to the problem of pedigree analysis, including allozymes, microsatellites and most recently single-nucleotide polymorphisms (SNPs). Key considerations for the utility of a molecular marker include (i) variability, (ii) ease of laboratory data generation and (iii) cost per individual.

Initial studies using single-locus protein-based markers, such as allozymes, had limited utility in species with low variability, with the added issue that data from these markers may not be consistent with neutral expectations (Parker, Snow, Schug, Booton, & Fuerst, 1998). Highly polymorphic microsatellite loci quickly became the molecular marker of choice for ecological studies with the widespread adoption of PCR in the early 1990s (Morin, Luikart, & Wayne, 2004). These DNA-based markers can have large numbers of alleles and, thus, high information content, and became the dominant marker for exclusion-based pedigree analysis (Parker et al., 1998). However, microsatellites also have many shortcomings, including substantial homoplasy and high genotyping error rates (Garza & Freimer, 1996; Hoffman & Amos, 2005; Morin et al., 2004; Pemberton, 2008). In addition, measurement error between genotyping platforms and laboratories makes reproducibility challenging (Pemberton, 2008; Seeb et al., 2007) and identifying sufficiently variable microsatellite loci, particularly in species with low diversity, has historically been difficult (Parker et al., 1998; Pastor, Garza, Allen, Amos, & Aguilar, 2004).

In contrast, single-nucleotide polymorphisms (SNPs) are the most abundant form of variation in the genome of most species (Brumfield, Beerli, Nickerson, & Edwards, 2003; Morin et al., 2004) and their characterization has become straightforward with the advent of high-throughput DNA sequencing. In addition, SNP genotypes can be called with much less human interaction, generally have low error rates, and facilitate data sharing and collaboration (Anderson & Garza, 2006; Clemento, Abadía-Cardoso, Starks, & Garza, 2011; Seeb, Pascal, Ramakrishnan, & Seeb, 2009). Despite the advantages of SNPs, the vast majority are bi-allelic and do not provide the same per-locus power as microsatellites. As such, many more SNPs than microsatellite loci are generally required to provide similar power for population genetic and molecular ecological studies (e.g., Hauser, Baird, Hilborn, Seeb, & Seeb, 2011; Kaiser et al., 2017; Narum et al., 2008; Weinman, Solomon, & Rubenstein, 2015).

The huge amounts of data generated by high-throughput DNA sequencers are transforming population biology, where they have

helped to elucidate species relationships, genetic connectivity and ecological processes (Andrews et al., 2016; Ekblom & Galindo, 2011; McCormack et al., 2013; Narum, Buerkle, Davey, Miller, & Hohenlohe, 2013). However, unlike traditional Sanger sequencing, precise control over instrument output is challenging, so most initial applications have involved the collection of large amounts of data from one or a small number of individuals, with sequencing reads either randomly sampling the genome or a reduced fraction of it. However, many questions in population biology do not require “whole genome” sequences or even the thousands of SNPs provided by most reduced representation methods, such as RADseq. As such, much effort has been expended to direct sequencing power to small numbers of genomic targets, allowing more individuals to be studied in a single instrument run.

Here, we describe how data from multiple SNPs that occur within the same small region, and which can be genotyped jointly from single reads from high-throughput DNA sequencers, can be used to much more efficiently derive accurate relationship inference. This method uses the phase information inherent in these short-read DNA sequences to derive multi-allelic “microhaplotype” markers from multiple, proximate SNPs (Kidd et al., 2013, 2014). We use data from a nearshore marine fish and simulation analysis to show how utilizing the additional information that comes from considering all variation in these short sequences provides large increases in inferential power for identifying kin relationships from the same amount of DNA sequence data.

As sequencing instruments are limited in the total number of sequencing reads produced in a single run, finding the optimal trade-off between the number of samples analysed and the number of genomic targets sequenced becomes critically important for population biological studies. For questions that are extremely data-intensive, or are focused specifically on genomic issues, whole genome sequencing or reduced representation methods may be necessary and appropriate, but they will be prohibitive when it is also necessary to analyse a large number of individuals. For projects that require analysis of thousands of samples, it is important to utilize data collection methods that make the most efficient use of sequencing technology, so that a modest number of loci, or genomic regions, are targeted, with these loci chosen to possess high information content. Multi-allelic microhaplotype markers meet this criterion and allow genotyping of many more individuals in a sequencing run, as many fewer such loci are necessary to achieve the same power than when just calling SNPs from such DNA sequence data.

Kidd et al. (2013, 2014) provided a proof of concept that microhaplotype markers exist in the human genome and are useful for forensic and pedigree-type questions. Gattepaille and Jakobsson (2012) showed analytically and empirically that such microhaplotypes increase the power for assignment of individuals to a population of origin, a result that was extended by McKinney et al. (2017) for natural populations of salmon.

We expand on this concept by describing a set of microhaplotype loci in an organism without a reference genome, kelp rockfish (*Sebastes atrovirens*), a Pacific Ocean nearshore species of ecological

and cultural importance. We then show how targeting gene regions with abundant natural variation allows development of a 96 locus microhaplotype panel with sufficient power for difficult relationship inference problems, including accurate identification of single parent–offspring and full-sibling pairs. We show how these microhaplotypes have significantly higher heterozygosity than 96 SNPs from the same data set and provide much more power for pedigree inference. While hundreds of SNP loci would be necessary to achieve similar accuracy, the panel of 96 microhaplotypes provides sufficiently low error rates for even the largest studies. We highlight how microhaplotypes will substantially increase the power for population genetic and ecological applications, and will be particularly useful for studies that require genetic markers that are easily genotyped and portable amongst laboratories that use benchtop sequencers to generate data. Microhaplotypes will substantially increase the efficiency of genotyping and provide greater analytical power, lowering costs and potentially enhancing collaboration and coordination in the study, management and conservation of animal and plant species.

2 | METHODS

2.1 | Samples

Tissue samples were obtained from field collections of rockfishes sampled at sites throughout Carmel and Monterey Bays, CA, USA. Adult kelp rockfish were sampled by hook-and-line capture followed by removal of a small sample from the caudal fin or by nonlethal, underwater pole-spear biopsy, and tissue samples were subsequently dried on blotting paper. Genomic DNA was extracted from the dried tissue of 144 unrelated adult fish using DNeasy 96 Blood and Tissue kits on a BioRobot 3000 (Qiagen, Inc.) with an elution volume of 200 μ l. DNA extracts were then stored at 4°C until analysis.

2.2 | SNP discovery and amplicon design

To identify sufficient nucleotide variation in kelp rockfish for the design of microhaplotype markers, we used reduced-representation genome sequencing to generate data from which we could identify variants and design small amplicons (100–130 bp) containing multiple SNPs. We performed double-digest restriction site-associated DNA sequencing (ddRADseq; Peterson, Weber, Kay, Fisher, & Hoekstra, 2012) on 20 adult kelp rockfish. DNA concentration was normalized across individuals, and samples were digested with two restriction enzymes, Sph1 and EcoR1, with all other details of the library preparation as in Peterson et al. (2012). We selected 350-bp genomic fragments using a Pippin Prep (Sage Science) and sequenced 12 samples in one run and eight samples in a second run on a MiSeq (Illumina, Inc.) using 600-cycle paired-end sequencing kits.

Additionally, several loci were identified from publicly available expressed sequence tags (ESTs) in an approach analogous to that used to discover and validate SNPs in other fish species (e.g., Abadía-Cardoso, Clemento, & Garza, 2011; Clemento et al., 2011). We selected 192 ESTs for screening by PCR to determine those that

effectively amplified. We then generated Sanger sequence data for each locus from two kelp rockfish individuals to identify variants.

Initial analysis of the ddRAD data was with STACKS v1.34 (Catchen, 2013). Raw reads were demultiplexed using the *process_radtags* module which assigns specific reads to individuals based on unique barcode sequences. Demultiplexed reads were then passed to Stacks for assembly, requiring a minimum stack depth of 4 (*m*), the distance allowed between stacks of 2 (*M*) and the distance allowed between catalog loci of 2 (*n*). Stacks identified 17,991 genomic regions in the 20 kelp rockfish samples, where each region should correspond to a unique DNA sequence. We then filtered the Stacks-assembled genomic regions according to two criteria: (i) the presence of at least one SNP and (ii) genotyping data present in at least eight samples. This filtering reduced the data set to 3,517 genomic regions. To ensure that amplicon design targeted unique genomic regions (e.g., no repetitive elements), we used BLAT—the BLAST-Like Alignment Tool (Kent, 2002)—to perform pairwise comparisons of each genomic region with every other region and removed 1,184 likely duplicates (those with greater than 95% similarity).

We then filtered the remaining sequences for (i) between two and six SNPs within 300 bp, with at least two of them within 100–130 bases, (ii) presence of multiple haplotypes observed across the 20 kelp rockfish sequenced and (iii) no obvious deviations from Hardy–Weinberg equilibrium (HWE). From the remaining Stacks loci, we randomly selected 192 small genomic regions (<200 bp) for amplicon design. We targeted regions <200 bp because such short regions appear to amplify more uniformly in multiplex reactions than larger DNA fragments (D. S. Baetscher & J. C. Garza, unpublished data). We then successfully designed PCR primers for 177 of these candidate microhaplotype markers using PRIMER3 software in GENEIOUS v7.1.7 (Kearse et al., 2012) and added 15 gene regions from the EST sequencing data.

2.3 | Amplicon sequencing

We used Genotyping-in-Thousands by Sequencing (GT-seq; Campbell, Harmon, & Narum, 2015) to generate sequence data for haplotype calling. Briefly, we used an initial multiplex PCR to select amplicon sequences from genomic DNA in each sample. We performed multiplex PCR with primers for 96 amplicons targeting DNA from 96 adult kelp rockfish in each reaction. The locus-specific primers were designed to include priming sites for the sequencing reactions, which allows the instrument to recognize start locations for sequencing. A second PCR added individual-specific indexes (DNA barcodes) that allow sequences to be identified to individual samples during bioinformatic analysis. After both PCRs, DNA concentration was normalized across samples to minimize variation in number of sequencing reads per individual. Post-normalization, indexed samples were combined and the sequencing library was quantified by Qubit Fluorometer (Thermo Fisher Scientific) and then by qPCR with the Illumina Library Quantification Kit (Kapa Biosystems). Finally, we sequenced the library on a MiSeq instrument using a paired-end approach and a 150-cycle sequencing kit. All other details of the

thermal cycling and library preparation are as in Campbell et al. (2015).

We tested 192 loci in two sets of 96 amplicons per sequencing run, with 96 DNA samples each. We replicated the first sequencing run with 48 of the same samples to evaluate consistency across sequencing runs and substituted half of the samples with 48 different individuals from the same collection to check for consistency of loci across samples. For the second set of 96 amplicons, we dropped three of the loci in the replicate run due to high read depth. These four sequencing runs provided variation information for a total of 144 individuals and each run produced 23.8–27.6 million reads that passed filter.

2.4 | Bioinformatic processing

Sequencing reads for each sample were grouped by index with the MISEQ Analysis Software (Illumina), and paired-end reads were combined using the Fast Length Adjustment of SHort reads (FLASH; Magoč & Salzberg, 2011). Only successfully paired reads were retained and then mapped to a reference file of consensus sequences using the Burrows-Wheeler Aligner (BWA-MEM; Li & Durbin, 2009). Mapped reads were converted from Sequence Alignment/Map (SAM) files to Binary Alignment/Map (BAM) files with SAMtools (Li et al., 2009), and then, FreeBayes (Garrison & Marth, 2012) was used to call variants with settings that did not include an input set of variants, multinucleotide polymorphisms or complex variation (composites of other types of variation). FreeBayes outputs a variant call format (VCF) file with information about the position of each SNP in each locus from the 144 rockfish evaluated.

Existing software was unable to reliably assemble haplotypes from specified variants, primarily due to a large number of reads per locus. Accordingly, we developed MICROHAPLOT, a novel program implemented as an R (R Core Development Team 2016) package and associated Shiny app (<http://shiny.rstudio.com/>), that easily imports amplicon data containing microhaplotypes and allows filtering based on various criteria before outputting individual haplotypes and their read depths in each individual (Ng et al., <https://doi.org/10.5281/zenodo.820110>). MICROHAPLOT uses a reference VCF file to specify the variant sites in each target region that are to be assembled into microhaplotypes, and then, it extracts those sites from SAM files of reads (one for each individual) aligned to the target regions. MICROHAPLOT properly accounts for indel variation within the target regions and uses the co-occurrence of variants on single reads to provide phase information to call microhaplotypes.

We filtered data to retain only those haplotypes with at least 20 reads at a locus within an individual. However, as some sequencing protocols can produce very high read depths at loci in some individuals, it is possible to find spurious microhaplotypes with read depth greater than 20, due to sequencing errors and “index switching” (Sinha et al., 2017). To filter these out, we also removed haplotypes with a read depth ratio of less than 0.2; that is, those haplotypes at a locus with less than 0.2 of the read depth of the haplotype with the highest read depth within an individual. For example, if an

individual has read depths 1,000, 800, 33 and 10 at a locus for haplotypes AGT, AAT, GGC and AGC, respectively, the AGC haplotype would be removed because it had fewer than 20 reads, and the GGC one because it has a read depth ratio of $33/1,000 = 0.033$ (<0.2). After this filtering, genotypes were called from the remaining haplotypes, and if two or more haplotypes remained, the individual was called as a heterozygote of the two haplotypes with highest read depth, and if only one haplotype remained, then the individual was called as a homozygote. Any locus that generated called genotypes for fewer than 75% of the samples was then excluded. Likewise, loci with obvious deviations from HWE and those that produced a third or fourth unfiltered haplotype with read depth >50 in more than 5% of individuals were removed. Finally, we removed monomorphic loci—those with only one haplotype present in the 144 test samples. Individual haplotypes from the 165 remaining loci were then exported from MICROHAPLOT for downstream analyses.

To determine the utility of microhaplotypes for pedigree analyses and compare their performance with bi-allelic SNPs, we generated five data sets to assess power in both marker types across all 165 genomic regions and with sets of 96 genomic regions. These data sets are as follows: microhaplotypes in all 165 genomic regions (m165); the single SNP with the highest heterozygosity in each of the 165 genomic regions (s165); the 96 microhaplotypes with the highest heterozygosities (m96); 96 SNPs with the highest heterozygosity, with no more than one per genomic region (s96_top); and finally, the single SNP with the highest heterozygosity within the genomic regions containing the best 96 microhaplotypes (s96_m). We then used Monte Carlo simulation to evaluate the power to accurately identify parent-offspring pairs, and full- and half-sibling pairs using these five data sets.

The Monte Carlo simulations were made using CKMRSIM (Anderson, <https://doi.org/10.5281/zenodo.820162>), an R (R Core Team, 2016) package that implements a variant of the importance-sampling algorithm of Anderson and Garza (2006) tailored to pairwise relationship inference and multi-allelic markers. Briefly, in CKMRSIM, the genotypes of related pairs of individuals are simulated from the estimated allele frequencies and the probabilities of those genotype pairs are calculated to compute a log-likelihood ratio of the true relationship vs. the hypothesis of no relationship. Similarly, genotypes of unrelated pairs are also simulated and their log-likelihood ratios computed. The simulated distributions of these log-likelihoods are used to compute the false-negative rates (the per-pair rate at which pairs that truly have the specified relationship are deemed unrelated) and the false-positive rates (the per-pair rate at which unrelated individuals are incorrectly inferred to be related with the specified relationship) to be expected when any particular log-likelihood ratio threshold is used as a criterion for classifying a pair into a given relationship, vs. unrelated. The importance-sampling algorithm permits accurate estimation of very small per-pair false-positive rates ($<10^{-10}$) which cannot be accurately estimated using conventional Monte Carlo.

Simulations and likelihood calculations in CKMRSIM were made using a genotyping error model that includes allelic dropout and sequencing

errors. We set the rates of the errors so that, with both microhaplotypes and SNPs, the per-locus rate of calling an incorrect genotype was between 0.005 and 0.01. False-positive rates for parent-offspring and full- and half-sibling relationships were calculated for a range of false-negative rates from 0.01 to 0.3. In addition, to further evaluate the power to identify half-sibling pairs, we replicated two of the three 96-locus data sets (m96, s96_top) providing data sets that included 1, 2, 4, 8 and 16 times as many loci (i.e., providing allele frequencies for between 96 and 1,536 markers) and assessed power at a single FNR value of 0.01. Finally, as physical linkage between markers (even if they are not in linkage disequilibrium) results in a reduction in power for inference of siblings (relative to using entirely unlinked markers), and because close physical linkage becomes more likely with a larger number of markers, we evaluated the effects of physical linkage on the power of the replicated data sets for half-sibling inference. This was carried out by assuming a “typical vertebrate genome” (25 chromosomes of between 1 and 2 Morgans in recombinational length) into which loci were randomly positioned. Simulations in CKMRSIM were then performed assuming physical linkage using the package’s ability to call the software MENDEL (Lange et al., 2013).

3 | RESULTS

Three of the 192 loci were removed because they collectively accounted for nearly 73% of reads in one of the sequencing runs. After manually curating the remaining 189 loci in MICROHAPLOT using the criteria described above, 165 loci remained for analysis. These loci contained 825 unique haplotypes across 144 kelp rockfish, with between one and 11 SNPs per locus (mean 3.58) and two and 13 haplotypes per locus (Figure 1). For the 96 loci sequenced in replicate runs, the ordinal rank of loci by number of reads from the same 48 individuals was strongly correlated (Spearman’s coefficient = 0.99), demonstrating the consistency of results for individual loci across runs. In addition, read depths were consistently high (mean 1,483 reads per genotype) and genotype call rates were consistently high, ranging between 92% and 96% of all locus/individual combinations (at read depth of 20) in the four runs. Most of the missing data was concentrated in several individuals that appeared to have lower quality DNA extractions.

Observed heterozygosities of the 165 microhaplotype loci (m165) were substantially higher than those of the most variable single SNPs in each of the 165 loci (s165) (Figure 2). Mean heterozygosity of the microhaplotype loci was 0.41, vs. 0.22 when just the SNP with the highest minor allele frequency (MAF) in each locus was called. The 96 most informative microhaplotype loci (m96) had a mean of 5.64 alleles (haplotypes) per locus and mean heterozygosity of 0.54 (range = 0.37–0.82), whereas for the 96 most variable SNPs (s96_top) mean heterozygosity was 0.33 (range = 0.17–0.49). Mean heterozygosity of the most variable SNPs in each of the 96 best microhaplotype loci (s96_m) was very similar to that of the 96 best SNP loci (s96_top) and, as such, that set of polymorphisms was not evaluated further.

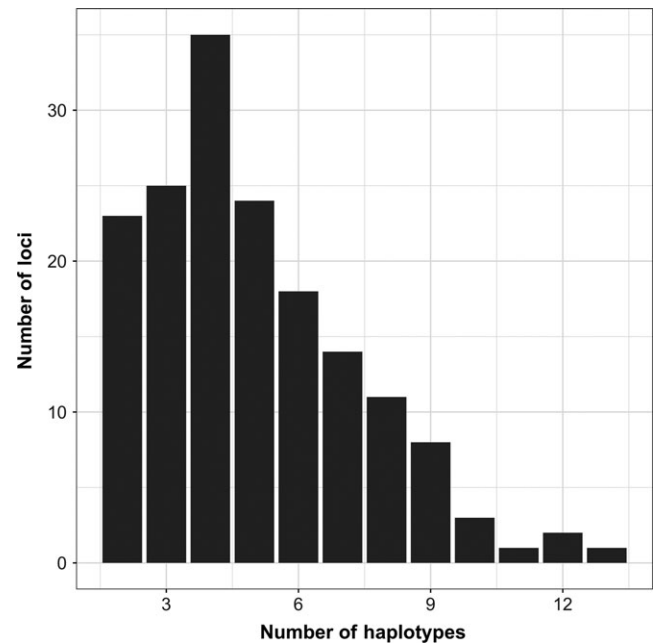


FIGURE 1 Distribution of microhaplotypes across 165 loci in 144 kelp rockfish samples. The number of haplotypes per locus ranged from two to 13

False-positive rates (FPR) for identifying parent-offspring pairs and full-sibling pairs, estimated using simulations, were much smaller with microhaplotypes than with SNPs (Figure 3). The FPR is inversely related to the false-negative rate (FNR), so that increasing FNR decreases FPR. At FNR = 0.01, matching single parents with offspring using 96 microhaplotype loci (i.e., m96 data set) resulted in an FPR of 8.43×10^{-11} , whereas with the top 96 SNPs (s96_top), it was 2×10^{-4} (Figure 3a). For identifying full-sibling pairs, also at FNR = 0.01, the FPR for m96 was 9.62×10^{-8} , and with s96_top, it was 2.54×10^{-3} (Figure 3b). In contrast, for identifying half-sibling pairs, considerably more power than provided by the set of either 96 microhaplotype loci or 96 SNPs is needed to achieve acceptable false-positive rates (Figure 4). With 96 microhaplotype loci, the FPR, again at FNR = 0.01, is 0.065, which means that more than one of twenty comparisons of nonsiblings would result in a false-positive identification. For the 96 SNPs, FPR = 0.44 at the same FNR of 0.01, indicating an almost complete lack of power to discriminate half-siblings from unrelated individuals.

Even when the SNP data set is expanded by a factor of four (for a total of 384 loci), the half-sibling FPR for SNPs decreases to only 4.6×10^{-3} (Figure 4). In contrast, when the microhaplotype data set is expanded by a factor of four, the resulting FPR at a FNR of 0.01 is 6.8×10^{-9} , which would be adequate for all but very large studies. Moreover, when taking into account physical linkage, which is unavoidable when the number of markers exceeds the number of chromosome arms and reduces the independence between markers for sibling inference, the apparent increase in power when adding markers is reduced, relative to unlinked markers, with the reduction increasing with the number of markers (Figure 4). Although the reduction is not extreme, to achieve an FPR of 1×10^{-9} at

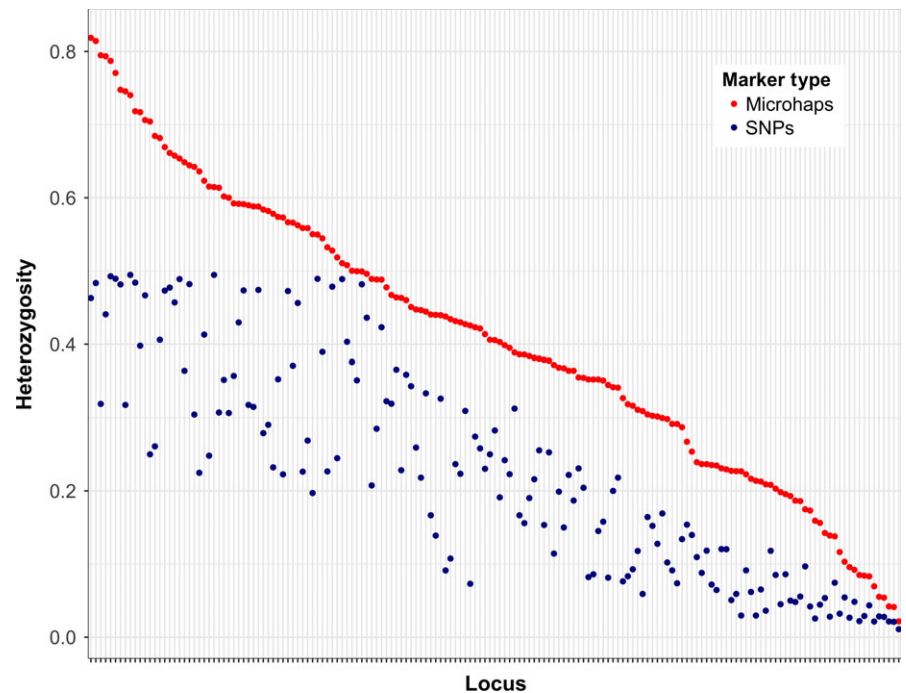


FIGURE 2 Heterozygosity of 165 microhaplotypes comprised of all SNPs in a locus compared to the single SNP with the highest minor allele frequency in that same locus. Bi-allelic SNPs have a maximum heterozygosity of 0.5

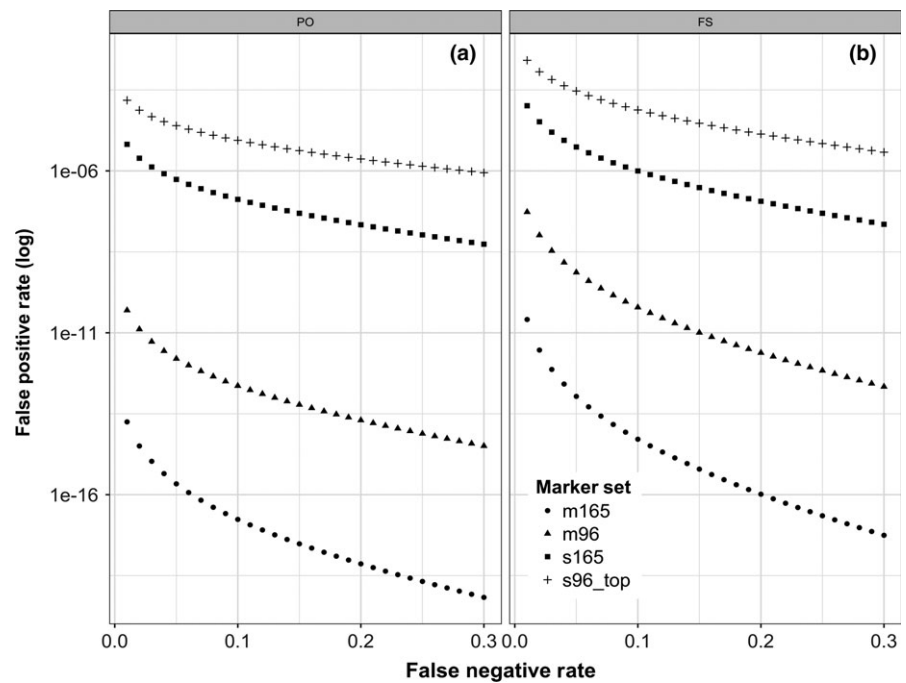


FIGURE 3 Simulated false-positive rates for matching (a) single parents with offspring and for (b) full-siblings at a given false-negative rate using the four sets of markers: 165 microhaplotypes (m165), 165 SNPs (s165), 96 microhaps (m96) and 96 SNPs with the highest heterozygosity (s96_top)

FNR = 0.01 in half-sibling analysis, about 50 more microhaplotypes are necessary than would be predicted without taking into account a typical pattern of linkage. In contrast, approximately 350 additional SNPs would be necessary to achieve such additional power in the face of physical linkage. Note that this analysis is intended to evaluate power for larger data sets with equivalent variation to the empirical data presented here, but will provide conservative estimates of the number of loci necessary if such larger data sets had lower mean heterozygosity.

4 | DISCUSSION

As population genetic and molecular ecology research transitions to use of data from high-throughput DNA sequencers, it is critical to determine which data collection methods provide the optimal balance between the necessary amount of data per individual and the maximum number of individuals that can be accommodated in each instrument run. Many population genetic questions, including elucidation of patterns of population structure and most relationship

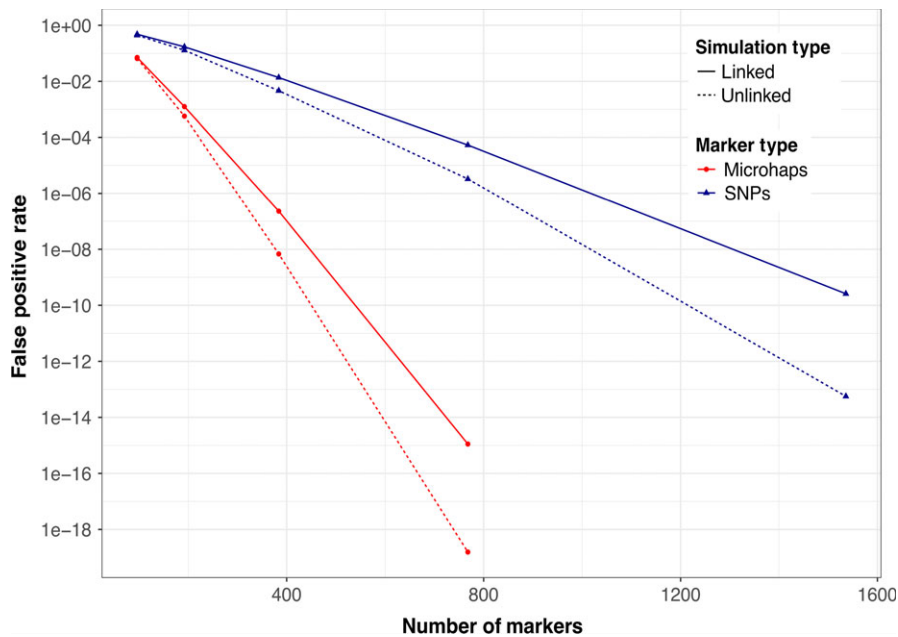


FIGURE 4 Simulated false-positive rates for identifying half-siblings with a given number of microhaplotype and SNP markers at a false-negative rate of 0.1. Data are simulated both including and excluding physical linkage

inference analyses, require many fewer genetic markers than provided by popular reduced-representation genome sequencing approaches (e.g., RADseq, ddRADseq). In addition, such approaches usually yield different, albeit overlapping, sets of genomic regions in different sequencing runs for different investigators, complicating collaboration and replication. In contrast, targeted methods, such as amplicon-sequencing and capture array approaches, offer the ability to optimize the trade-off between the number of samples and the amount of data acquired per sample, yielding data sets that are predictable and easily replicated.

Here, we identified short genomic regions containing multiple SNPs segregating as haplotypes and designed amplicons that can be easily multiplexed and sequenced using such targeted protocols. These microhaplotypes contain more information than single bi-allelic SNPs and offer the benefit of providing much more inferential power per locus than the SNP data typically derived from high-throughput DNA sequencers. The microhaplotype information is provided directly in such data, without the need for statistical phasing (Stephens & Donnelly, 2003), because the sequences are replicated from single molecules and therefore preserve phase information for variants located in the same sequencing read. This phase information allows much higher inferential and statistical power for examining population biology questions, including data-intensive inference of pedigree relationships, by calling multi-allelic microhaplotypes from the same sequence data typically used to call bi-allelic SNPs.

The value of utilizing incomplete linkage disequilibrium (LD) between proximate SNPs (Pakstis, Fang, Furtado, Kidd, & Kidd, 2012) and leveraging the phase information that comes from high-throughput DNA sequencing instruments (Kidd et al., 2013, 2014) has been previously recognized. Kidd et al. (2013) demonstrated that areas of the human genome where two or more SNPs occur within ~200 bp are common and that the SNPs were generally not in

complete LD, with recombination, genetic drift and/or selection creating population ancestry-informative alleles (Kidd et al., 2014).

Here, we extend the documentation of microhaplotype utility, by showing how selecting genomic regions with multiple SNPs in close proximity for use with targeted sequencing approaches, including the amplicon-sequencing approach we employ, allows much more power for relationship inference to be derived from the same amount of high-throughput sequencing data. In the example rockfish data set, simulations demonstrated that 96 microhaplotype loci generate false-positive rates for single parent-offspring identification on the order of 10^{-11} , at FNR = 0.01, whereas the most informative single SNPs from each of 96 loci provided false-positive rates of 10^{-4} (Figure 3a). Similarly, power for the more challenging problem of full-sibling identification was substantially higher with the 96 microhaplotype loci (FPR on the order of 10^{-8}) than with the 96 best SNPs (FPR on the order of 10^{-3}).

The much higher mean heterozygosity of the microhaplotype loci compared with the SNPs is indicative of their greater information content for population genetic analyses, particularly relationship inference. Indeed, the simulations demonstrated that the microhaplotype markers substantially outperformed the corresponding SNP loci in all cases (Figure 3). While 96 SNP loci with modest mean MAF have been shown to be sufficient to identify parent pair-offspring trios (Abadía-Cardoso, Anderson, Pearse, & Garza, 2013; Anderson & Garza, 2006), single parent-offspring pair identification is considerably more challenging—there is greater separation between the likelihood ratio distributions of parent pair-offspring trios and unrelated trios than there is between single parent-offspring pairs and unrelated pairs. While the false-positive rates estimated for the 96 SNPs might seem low, even exceedingly small rates can lead to a large number of false-positive errors, because these are *per-pair* rates. The expected number of false-positive errors is found by multiplying the FPR by the total number of pairwise comparisons necessary. Many

studies, particularly in natural populations, involve very large numbers of pairwise comparisons. For example, with samples from 5,000 adults and 5,000 juveniles, a single parent–offspring identification analysis involves a total of 2.5×10^7 pairwise comparisons. Thus, even false-positive rates of 10^{-6} could result in dozens of incorrectly inferred pedigree relationships. With the best 96 microhaplotype loci from kelp rockfish, the false-positive rate for single parent–offspring pair analysis is 8.43×10^{-11} at a false-negative rate of 1%. This means that even with 100,000 parents and 100,000 offspring genotyped (for a total of 10^{10} pairwise comparisons), less than one falsely inferred parent–offspring relationship between unrelated individuals would be expected. In contrast, with this same sampling scheme, and a dataset with the 96 best SNPs (FPR = 2×10^{-4} at FNR of 1%), we would expect thousands of false positives (Figure 3a).

Similarly, Monte Carlo evaluation of the false-positive rate for full-siblings demonstrates the substantial increase in power obtained using microhaplotypes rather than SNPs. For example, searching for full-siblings from amongst a data set with 5,000 juveniles would involve nearly 1.25×10^7 pairwise comparisons. False-positive rates in a study with this number of pairwise comparisons between potential full-sibling pairs and using 96 microhaplotype loci (FPR = 9.62×10^{-8} at FNR of 1%) are expected to produce less than one false positive. In contrast, with the best 96 SNPs (FPR = 2.54×10^{-3} at FNR of 1%) such an analysis would potentially result in thousands of false positives, highlighting how most published studies that attempt to identify pairs of siblings, particularly half-siblings, have been underpowered. While 96 SNPs can be sufficient for accurately identifying large full-sibling groups in a joint analysis (e.g., with COLONY [Wang, 2004]), if only a small number of sibling pairs are present in the sample, the joint analysis offers no increase in power over a pairwise approach. This situation occurs frequently when sampling large populations for the purpose of close-kin mark–recapture (Bravington, Skaug, & Anderson, 2016).

Another analytical application that will benefit from increased power with microhaplotypes is genetic stock identification (GSI) or individual assignment. The microhaplotype panel described here is for a species without significant population structure (Gilbert-Horvath, Larson, & Garza, 2006), but utilizing the phased data from short-read sequences for haplotype determination has recently been shown to increase power for GSI as well, although by a much smaller margin than for relationship inference applications (McKinney et al., 2017). However, with targeted ascertainment, it is feasible to over-represent loci with haplotypes that have highly diagnostic frequencies across different populations and with high power for identifying particular ancestry (Kidd et al., 2014; Pakstis et al., 2012). In addition, Willis, Hollenbeck, Puritz, Gold, and Portnoy (2017) have shown how microhaplotypes can be used to identify paralogs and genotyping error in RADseq data sets. The use of microhaplotypes might also increase analytical power for other population genetic applications, such as estimation of effective population size or phylogeography, but such applications would require appropriately ascertained loci, and not ones chosen, as here, to include as many SNPs as

possible in 100- to 150-bp fragments. As such, both initial ascertainment and microhaplotype locus screening and validation would need to involve appropriate discovery strategies and samples for the type of inference that will be pursued.

We used 96 loci to compare the power of microhaplotypes with SNPs, primarily because of the standard 96-well microplate configuration, and the associated standardization of laboratory equipment, including many traditional genotyping and sequencing platforms, around this 96-well microplate configuration. However, there is no inherent constraint on the number of microhaplotype loci that can be included in a particular study. While compiling data from microhaplotypes can be implemented with any type of short-read sequencing protocol that preserves phase information, including RAD approaches (Willis et al., 2017) and capture arrays, one benefit of the targeted sequencing approaches, such as the GT-seq amplicon protocol used here, is the ability to include any number of loci (Campbell et al., 2015), so that panels of microhaplotype loci can be tailored to the study-specific requirements for analytical power in relationship inference (Anderson & Garza, 2006). In addition, the GT-seq protocol that we employ here interrogates fragments of 100–150 bp in length that contain multiple variable sites. In principal, longer fragments could contain more variable sites on average and provide more power per locus than the relatively short fragments we analyse. We note, however, that longer fragments involve trade-offs of sequencing cost and instrument run time, and that the highly multiplexed PCR involved in our approach is less reliable with longer fragments (D. S. Baetscher & J. C. Garza, unpublished data).

Similarly, although we used FreeBayes to call SNPs and create the VCF files for input to MICROHAPLOT, in which haplotypic genotypes were then called, there are several alternative analytical pathways, including FreeBayes and Stacks, that can be used to call SNPs in a haplotype-aware manner that preserves phase information, and then call genotypes from the resulting haplotype data. A full treatment of the advantages and drawbacks of different analysis software is beyond the scope of the current study, but we note that the filtering and data handling capabilities of these other analytical approaches are not always well suited to the high read depths and other aspects of amplicon-sequencing data sets such as the one we present.

Our criteria for calling genotypes after filtering of haplotypes using read depth and read depth ratio in MICROHAPLOT were empirically derived. It is likely that a small number of additional genotypes could be called (or a small number of genotypes could be called with less error) by applying different filtering criteria. For example, our approach would lead to an individual with 21 reads of haplotype A and 18 reads of haplotype B being called as a homozygote of haplotype B. We note, however, that only 1.8% of genotype calls in our experiments had sufficiently low read depth (homozygotes with <40 reads) to even have the chance of such an error. Although a better approach that utilizes a probabilistic model to call haplotypes and genotypes from the sequencing information could undoubtedly be developed, it is beyond the scope of the present work.

In ecological and conservation studies that require genotyping a large number of samples, extracting more information per sequence than the typically called bi-allelic SNPs is easily achieved with multi-allelic microhaplotypes and will prove to be more efficient and cost-effective. In addition, we analyse 96 samples here for convenience, but in subsequent work have shown that we can reliably generate genotypes for 384 fish at 96 of these microhaplotype loci in a single such sequencing run of an Illumina MiSeq—a medium throughput benchtop sequencer—achieving call rates above 99%, at a minimum read depth of 20 for all individual/locus combinations (D. S. Baetscher & J. C. Garza, unpublished data).

Markers with higher information content intuitively reduce the amount of genotyping required for a set amount of inferential power (Rosenberg, Li, Ward, & Pritchard, 2003). Moreover, it is particularly important to minimize the number of genetic markers used in the identification of close kin, because of the challenge of physical linkage in some such analyses. Because recombination does not occur between many loci that are on the same chromosome during any single segregation event, such loci on the same chromosome do not provide independent observations of relatedness for sibling relationship categories. We show that the effect of physical linkage on relationship inference with a small number of markers is minimal, and it is thus unlikely to greatly affect parent–offspring and full-sibling identification. However, with the much larger number of markers necessary for half-sibling analysis, linkage increases the false-positive rate substantially and the discrepancy becomes greater as the number of markers increases, so that hundreds of additional SNP markers are necessary to account for this linkage and achieve FPR values that might be necessary for studying natural populations. Furthermore, the reduced cost per individual of genotyping a panel with a modest number (e.g., 96) of microhaplotype loci compared to methods that target larger proportions of the genome will allow enhanced monitoring and evaluation of lower-profile species, benefiting management and conservation of many different animal and plant species.

ACKNOWLEDGEMENTS

We thank M. Carr, D. Malone, E. Saarman, C. Edwards and A. Lowe for assistance with sample collection and study design. We also thank C. Columbus, E. Correa and E. Gilbert-Horvath for assistance with sample processing and data generation. D. Pearse and three anonymous reviewers provided feedback that significantly improved the manuscript. This work was supported by a grant from the National Science Foundation (Award number 1260693; PIs M. Carr, E.C. Anderson, C. Edwards and J.C. Garza).

AUTHOR CONTRIBUTIONS

DSB, ECA and JCG designed the study. DSB conducted laboratory work. DSB, AJC, TCN, ECA and JCG analysed data. DSB and ECA performed simulations. DSB, ECA and JCG wrote the manuscript with input from all authors.

DATA ACCESSIBILITY

Consensus sequences and primer information for all 192 targeted amplicons, the Binary Alignment/Map (BAM) files and variant call format (VCF) files for all retained regions genotype files for all 144 kelp rockfish and an R Notebook documenting all statistical analyses are deposited in Dryad, <https://doi.org/10.5061/dryad.5863d>.

ORCID

Eric C. Anderson  <http://orcid.org/0000-0003-1326-0840>

John C. Garza  <http://orcid.org/0000-0002-7325-6803>

REFERENCES

- Abadía-Cardoso, A., Anderson, E. C., Pearse, D. E., & Garza, J. C. (2013). Large-scale parentage analysis reveals reproductive patterns and heritability of spawn timing in a hatchery population of steelhead (*Oncorhynchus mykiss*). *Molecular Ecology*, 22, 4733–4746. <https://doi.org/10.1111/mec.12426>
- Abadía-Cardoso, A., Clemento, A. J., & Garza, J. C. (2011). Discovery and characterization of single-nucleotide polymorphisms in steelhead/rainbow trout, *Oncorhynchus mykiss*. *Molecular Ecology Resources*, 11 (Suppl 1), 31–49. <https://doi.org/10.1111/j.1755-0998.2010.02971.x>
- Anderson, E. C., & Garza, J. C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, 172, 2567–2582.
- Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (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>
- Bravington, M. V., Skaug, H. J., & Anderson, E. C. (2016). Close-kin mark-recapture. *Statistical Science*, 31, 259–274. <https://doi.org/10.1214/16-STS552>
- 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](https://doi.org/10.1016/S0169-5347(03)00018-1)
- 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>
- Catchen, J. M. (2013). STACKS: An analysis tool set for population genomics. *Molecular Ecology*, 22, 3124–3140. <https://doi.org/10.1111/mec.12354>
- Clemento, A. J., Abadía-Cardoso, A., Starks, H. A., & Garza, J. C. (2011). Discovery and characterization of single nucleotide polymorphisms in Chinook salmon, *Oncorhynchus tshawytscha*. *Molecular Ecology Resources*, 11(Suppl 1), 50–66. <https://doi.org/10.1111/j.1755-0998.2010.02972.x>
- Eklom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107, 1–15.
- Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907v2*, 9.
- Garza, J. C., & Freimer, N. B. (1996). Homoplasmy for size at microsatellite loci in humans and chimpanzees. *Genome Research*, 6, 211–217. <https://doi.org/10.1101/gr.6.3.211>
- Gattepaille, L. M., & Jakobsson, M. (2012). Combining markers into haplotypes can improve population structure inference. *Genetics*, 190, 159–174. <https://doi.org/10.1534/genetics.111.131136>

- Gilbert-Horvath, E. A., Larson, R. J., & Garza, J. C. (2006). Temporal recruitment patterns and gene flow in kelp rockfish (*Sebastes atrovirens*). *Molecular Ecology*, 15, 3801–3815. <https://doi.org/10.1111/j.1365-294X.2006.03033.x>
- 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.
- 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>
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... Drummond, A. (2012). GENEIOUS BASIC: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28, 1647–1649. <https://doi.org/10.1093/bioinformatics/bts199>
- Kent, W. J. (2002). BLAT – The BLAST – Like Alignment Tool. *Genome Research*, 12, 656–664. <https://doi.org/10.1101/gr.229202>
- Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., & Ihuegbu, N. (2013). Microhaplotype loci are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series*, 4, e123–e124.
- Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagacé, R., Chang, J., Wootton, S., ... Kidd, J. R. (2014). Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Forensic Science International: Genetics*, 12, 215–224. <https://doi.org/10.1016/j.fsigen.2014.06.014>
- Lange, K., Papp, J. C., Sinsheimer, J. S., Sripracha, R., Zhou, H., & Sobel, E. M. (2013). MENDEL: The Swiss army knife of genetic analysis programs. *Bioinformatics*, 29, 1568–1570. <https://doi.org/10.1093/bioinformatics/btt187>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- 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 SAMtools. *Bioinformatics*, 25, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27, 2957–2963.
- McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, 66, 526–538. <https://doi.org/10.1016/j.ympev.2011.12.007>
- McKinney, G. J., Seeb, J. E., & Seeb, L. W. (2017). Managing mixed-stock fisheries: Genotyping multi-SNP haplotypes increases power for genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, 74, 429–434. <https://doi.org/10.1139/cjfas-2016-0443>
- Morin, P. A., Luikart, G., & Wayne, R. K. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, 19, 208–216. <https://doi.org/10.1016/j.tree.2004.01.009>
- Narum, S. R., Banks, M., Beacham, T. D., Bellinger, M. R., Campbell, M. R., Dekoning, J., ... Garza, J. C. (2008). Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology*, 17, 3464–3477.
- Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, 22, 2841–2847. <https://doi.org/10.1111/mec.12350>
- Pakstis, A. J., Fang, R., Furtado, M. R., Kidd, J. R., & Kidd, K. K. (2012). Mini-haplotypes as lineage informative SNPs and ancestry inference SNPs. *European Journal of Human Genetics*, 20, 1148–1154. <https://doi.org/10.1038/ejhg.2012.69>
- Parker, P. G., Snow, A. A., Schug, M. D., Booton, G. C., & Fuerst, P. A. (1998). What molecules can tell us about populations: Choosing and using a molecular marker. *Ecology*, 79, 361–382.
- Pastor, T., Garza, J. C., Allen, P., Amos, W., & Aguilar, A. (2004). Low genetic variability in the highly endangered Mediterranean monk seal. *Journal of Heredity*, 95, 291–300. <https://doi.org/10.1093/jhered/esh055>
- 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>
- R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rosenberg, N. A., Li, L. M., Ward, R., & Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics*, 73, 1402–1422. <https://doi.org/10.1086/380416>
- Seeb, L. W., Antonovich, A., Banks, M. A., Beacham, T. D., Bellinger, M. R., Blankenship, S. M., ... Smith, C. T. (2007). Development of a standardized DNA database for Chinook salmon. *Fisheries*, 32, 540–552. [https://doi.org/10.1577/1548-8446\(2007\)32\[540:DOASDD\]2.0.CO;2](https://doi.org/10.1577/1548-8446(2007)32[540:DOASDD]2.0.CO;2)
- Seeb, J. E., Pascal, C. E., Ramakrishnan, R., & Seeb, L. W. (2009). SNP genotyping by the 5′-nuclease reaction: Advances in high-throughput genotyping with nonmodel organisms. In A. A. Komar (Ed.), *Single nucleotide polymorphisms: Methods and protocols* (pp. 277–292). Totowa, NJ: Humana Press.
- Sinha, R., Stanley, G., Gulati, G. S., Ezran, C., Travaglini, K. J., Wei, E., ... Weissman, I. L. (2017). Index switching causes “spreading-of-signal” among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv*, 125724. <https://doi.org/10.1101/125724>
- Stephens, M., & Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, 73, 1162–1169. <https://doi.org/10.1086/379378>
- Wang, J. (2004). Sibship reconstruction from genetic data with typing errors. *Genetics*, 166, 1963–1979. <https://doi.org/10.1534/genetics.166.4.1963>
- 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>
- Willis, S. C., Hollenbeck, C. M., Puritz, J. B., Gold, J. R., & Portnoy, D. S. (2017). Haplotyping RAD loci: An efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources*, 17, 955–965. <https://doi.org/10.1111/1755-0998.12647>

How to cite this article: Baetscher DS, Clemento AJ, Ng TC, Anderson EC, Garza JC. Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. *Mol Ecol Resour*. 2018;18:296–305. <https://doi.org/10.1111/1755-0998.12737>