MOLECULAR ECOLOGY
RESOURCES

Microhaplotypes provide increased power from short-read DNA sequences for relationship inference in fish and wildlife.

Journal:	*Molecular Ecology Resources*
Manuscript ID	Draft
Manuscript Type:	Resource Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Baetscher, Diana; University of California, Santa Cruz, Department of Ocean Sciences; Southwest Fisheries Science Center, Fisheries Ecology Division Clemento, Anthony; University of California, Santa Cruz, Department of Ocean Sciences; Southwest Fisheries Science Center, Fisheries Ecology Division Ng, Thomas; University of California, Santa Cruz, Department of Biomolecular Engineering; Southwest Fisheries Science Center, Fisheries Ecology Division Anderson, Eric; Southwest Fisheries Science Center, Fisheries Ecology Division; University of California, Santa Cruz, Department of Applied Math and Statistics Garza, John; Southwest Fisheries Science Center, Fisheries Ecology Division; University of California, Santa Cruz, Department of Ocean Sciences
Keywords:	Fish, Population Genetics - Empirical, Relationship Inference, Microhaplotype, Parentage, High throughput DNA sequencing

Microhaplotypes provide increased power from short-read DNA sequences for relationship inference in fish and wildlife

Diana S. Baetscher[1,2], Anthony J. Clemento[2,3], Thomas C. Ng[2,4], Eric C. Anderson[2,3], John Carlos Garza[1,2,3*],

[1] Department of Ocean Sciences, University of California, Santa Cruz, CA 95064, USA

[2] Southwest Fisheries Science Center, National Marine Fisheries Service, Santa Cruz, CA 95060, USA

[3] Institute of Marine Sciences, University of California, Santa Cruz, CA 95064, USA

[4] Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA

Correspondence: 110 McAllister Way Santa Cruz CA 95060, USA; Tel.: 831-420-3903; carlos.garza@noaa.gov

Keywords: Microhaplotype; population genetics; relationship inference; parentage; high-throughput DNA sequencing

Abstract

The accelerating rate at which DNA sequence data is now generated by high-throughput sequencing instruments provides both opportunities and challenges for population genetic and ecological investigations of fish and wildlife. 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 high-throughput 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 96 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, sibling and single parent-offspring pair identification. 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.

Introduction

The proliferation of individual-based population genetic methods in the study of 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 fish and wildlife, 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 analyzing these high-throughput sequencing data have made it more feasible to address difficult biological questions (McCormack et al. 2013, Kidd et al. 2014, Andrews et al. 2016, McKinney et al. 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 1) variability, 2) ease of laboratory data generation, and 3) 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 et al. 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 et al. 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 and Freimer 1996; Morin et al. 2004, Hoffman and Amos 2005, Pemberton 2008). In addition, measurement error between genotyping platforms and laboratories makes reproducibility challenging (Seeb et al. 2007; Pemberton 2008) and identifying sufficiently variable microsatellite loci, particularly in species with low diversity, has historically been difficult (Parker et al. 1998, Pastor et al. 2008).

In contrast, single-nucleotide polymorphisms (SNPs) are the most abundant form of variation in the genome of most species (Brumfield et al. 2003, Morin et al. 2004) and their identification has become simple 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 and Garza 2006, Seeb et al. 2009, Clemento et al. 2011). 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., Narum et al. 2008, Hauser et al. 2011, Weinman et al. 2015, Kaiser et al. 2016).

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 (Ekblom and Galindo 2011, McCormack et al. 2013, Narum et al. 2013, Andrews et al. 2016). However, unlike traditional Sanger sequencing, precise control over instrument output is challenging, so most initial applications have involved 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 analyzed 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 questions, whole genome sequencing or reduced representation methods may be necessary and appropriate, but they will be prohibitive when it is also necessary to analyze 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, since 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 (2015) showed analytically and empirically that such microhaplotypes increase the power for assignment of individuals to 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 accurately matching single parents and offspring, and identifying 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 among laboratories that use benchtop sequencers to generate data. Microhaplotypes will substantially increase the efficiency of genotyping, provide greater analytical power, lowering costs and potentially enhancing collaboration and coordination in the study, management and conservation of fish and wildlife species.

Methods

Samples

Tissue samples were obtained from field collections of rockfishes sampled at sites throughout Carmel and Monterey Bays, CA. Adult kelp rockfish were sampled by hook-and-line capture and removal of small caudal fin clip samples or non-lethal, underwater pole-spear biopsy (J. Smith, pers. comm.), and were subsequently dried on blotting paper. Genomic DNA was extracted from the dried tissue using DNeasy 96 Blood and Tissue kits on a BioRobot 3000 (Qiagen, Inc.) using an elution volume of 200 µl, with DNA extracts stored at 4°C until analysis.

SNP discovery and amplicon design

To identify sufficient nucleotide variation in kelp rockfish for design of microhaplotype markers, we used reduced-representation genome sequencing to generate data from which we could design small amplicons (100-130 bp) containing multiple SNPs. We performed double-digest restriction site-associated DNA sequencing (ddRADseq; Peterson et al. 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., Clemento et al. 2011, Abadía-Cardoso 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 Illumina sequencing data with Stacks v1.34 (Catchen et al. 2013) identified 17,991 gene regions in the 20 kelp rockfish samples, where each region should correspond to a unique DNA sequence. We then filtered the Stacksassembled gene regions according to two criteria: 1) the presence of at least one SNP, and 2) genotyping data present in at least eight samples. This filtering reduced the dataset to 3,517 gene regions. To ensure that amplicon design targeted unique gene regions (e.g., no repetitive elements), we used BLAT — the BLAST-Like Alignment Tool (Kent 2002) — to perform pairwise comparisons of each gene region with every other region and removed likely duplicates (those with greater than 95% similarity).

We then filtered the remaining sequences for 1) multiple SNPs within 100-130 bases and 2) presence of multiple haplotypes observed across the 20 kelp rockfish sequenced. From the remaining 2,333 gene regions, we selected 192 small gene 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 (unpublished data). We then designed PCR primers for candidate microhaplotype markers using Primer3 software in Geneious v7.1.7 (Kearse et al. 2012). Eight of these gene regions came from ESTs and 184 from our genomic sequencing data.

Amplicon sequencing

We used Genotyping-in-Thousands by Sequencing (GT-seq; Campbell et al. 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 sequencing instruments 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 150-cycle sequencing kit.

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.

Bioinformatic processing

Sequencing reads for each sample were grouped by index with the MiSeq Analysis Software (Illumina), paired-end reads were combined using the Fast Length Adjustment of SHort reads (FLASH; Magoc and Salzberg 2011) and then mapped to a reference file of consensus sequences using the Burrow-Wheeler Aligner (BWA-MEM; Li and 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 and Marth 2012) was used to call variants with settings that did not include an input set of variants, multi-nucleotide polymorphisms, or complex variation (composites of other types of variation). FreeBayes outputs a variant call format (vcf) file with cumulative 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 the large number of reads per locus. Accordingly, we developed MICROHAPLOT, a novel program that easily imports amplicon data containing microhaplotypes and allows filtering based on read depth and allelic ratio (the ratio of the most frequent haplotype to the second-most-frequent haplotype in an individual at a locus) before outputting individual haplotypes (Ng et al., DOI: 10.5281/zenodo.820110). MICROHAPLOT uses a reference vcf file containing target sites, a Sequence Alignment Map (SAM) file for each sample, and “population” information for each sample, if multiple populations or species are included. We filtered data to retain genotypes with a minimum of 20 reads per individual per locus and an allelic ratio of 0.1. We then excluded any locus that generated data for less than 75% of samples or produced more than two haplotypes per individual at a read depth threshold of 50. Loci with obvious deviations from Hardy Weinberg equilibrium, as determined by plots of observed and expected haplotype frequencies, were noted and 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 datasets to assess power in both marker types across all 165 gene regions and with sets of 96 gene regions. These datasets are as follows: microhaplotypes in all 165 loci (m165); the single SNP with the highest heterozygosity in each of the 165 gene regions (s165); the 96 microhaplotypes with the highest heterozygosities (m96); 96 SNPs, one each from amongst the 165 gene regions, having the highest heterozygosity (s96_top); and finally, the single SNP with the highest heterozygosity within the gene regions containing the best 96 microhaplotypes (s96_m). We then used Monte Carlo simulation to evaluate the power to accurately identify single parent-offspring pairs, and full- and half-sibling pairs using these five datasets.

The Monte Carlo simulations were made using CKMRsim (Anderson, DOI: 10.5281/zenodo.820162), an R (R Core Development 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 versus 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 truly related pairs are deemed unrelated) and the false positive rates (the per-pair rate at which unrelated individuals are incorrectly inferred to be related) to be expected when any particular log-likelihood ratio threshold is used as a criterion for classifying a pair into a given relationship, versus 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 datasets (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 1536 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 realized 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 done 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).

Results

Three of the 192 loci were removed because they collectively accounted for nearly 73% of reads in one of the sequencing runs. For the 96 loci sequenced in replicate runs, the ordinal rank of loci by number of reads from the same 48 individuals was extremely strongly correlated (Spearman's coefficient = 0.99), demonstrating consistency of results for individual loci across runs. In addition, 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. 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 two and 13 haplotypes per locus (Figure 1).

Observed heterozygosities of the 165 microhaplotypes were substantially higher than those of the single most variable SNP in each of the 165 loci (Figure 2). Mean heterozygosity of the microhaplotypes was 0.41, versus 0.22 when just the SNP with the highest minor allele frequency (MAF) in each locus was called. The 96 most informative microhaplotype loci 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 it was 0.33 (range = 0.17-0.49). Mean heterozygosity of the most variable SNPs in each of the 96 best microhaplotype loci was very similar to that of the 96 best SNP loci and, as such, that set of polymorphisms was not evaluated further.

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

Even when the SNP dataset is expanded by a factor of four (for a total of 384 loci), the half sibling FPR for SNPs decreases to only 4.6 x 10^{-3} (Figure 4). In contrast, when the microhaplotype dataset is expanded by a factor of four, the resulting FPR at a FNR of 0.01 is 6.8 x 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 x 10^{-9} at 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.

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 inference analyses, require many fewer genetic markers than provided by popular reduced-representation genome sequencing approaches (e.g., RAD-seq, ddRADseq, etc.). In addition, such approaches usually yield different, albeit overlapping, sets of SNP markers in different sequencing runs for different investigators, complicating collaboration and replication. In contrast, directed 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 datasets that are predictable and easily replicated.

Here, we identified short gene 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 probabilistic phasing, 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 et al. 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 gene 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 microhaplotypes (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 microhaplotypes 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 (Anderson and Garza 2006; Abadía-Cardoso et al. 2013), 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 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 matching analysis involves a total of 2.5 x 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 x 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 x 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 by using microhaplotypes rather than SNPs. For example, searching for full siblings from amongst a dataset with 5,000 juveniles would involve nearly 1.25 x 10^{7} pairwise comparisons. False positive rates in a study with this number of pairwise comparisons between potential full-sibling pairs and using 96 microhaplotypes (FPR = 9.62 x 10^{-8} at FNR of 1%) is expected to produce less than one false positive. In contrast, with the best 96 SNPs (FPR = 2.54 x 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 full 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 of power over a pairwise approach. This situation occurs frequently when sampling large populations for the purpose of close-kin mark-recapture (Bravington et al. 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 et al. 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 (Pakstis et al. 2012; Kidd et al. 2014). Correspondingly, both initial ascertainment and microhaplotype locus screening and validation would need to include appropriate samples for this type of inference.

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 configuration. However, there is no inherent constraint on the number of microhaplotype loci that can be included in a particular study. 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 ecological and conservation studies that require genotyping a large number of samples, extracting more information per sequence is easily achieved using multi-allelic microhaplotypes rather than bi-allelic SNPs and will prove to be more efficient and cost effective. In addition, we analyze 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 (unpublished data).

Markers with higher information content intuitively reduce the amount of genotyping required for a set amount of inferential power (Rosenberg et al. 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 on 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 fishes and wildlife.

Acknowledgements

We thank Mark Carr, Dan Malone, Emily Saarman, Chris Edwards and Anna Lowe for assistance with sample collection and study design. We also thank Cassie Columbus, Laney Correa and Libby Gilbert-Horvath for assistance with sample processing and data generation. 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).

Data Accessibility

Consensus sequences and primer information for all 192 targeted amplicons, and genotype files for all 144 kelp rockfish are deposited in Dryad, DOI:XXXXXX, and all statistical analyses are documented in an R Notebook in the same repository.

References

Abadía-Cardoso A, Anderson EC, Pearse DE, Garza JC (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.

Abadía-Cardoso A, Clemento AJ, Garza JC (2011) Discovery and characterization of single-nucleotide polymorphisms in steelhead/rainbow trout, *Oncorhynchus mykiss*. *Molecular Ecology Resources*, **11 Suppl 1**, 31–49.

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, **172**, 2567–82.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLOS ONE*, **3**, e3376.

Bravington, MV, Skaug HJ, Anderson EC (2016) Close-kin mark-recapture. *Statistical Science*, **31**, 259–274.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, **18**, 249–256.

Campbell NR, Harmon SA, Narum SR (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.

Catchen JM (2013) Stacks: An analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Clemento AJ, Abadía-Cardoso A, Starks HA, Garza JC (2011) Discovery and characterization of single nucleotide polymorphisms in Chinook salmon, *Oncorhynchus tshawytscha*. *Molecular Ecology Resources*, **11 Suppl 1**, 50–66.

Ekblom R, Galindo J (2010) 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 JC, Freimer NB (1996) Homoplasy for size at microsatellite loci in humans and chimpanzees. *Genome Research*, **6**, 211–217.

Gattepaille LM, Jakobsson M (2012) Combining markers into haplotypes can improve population structure inference. *Genetics*, **190**, 159–174.

Gerber S, Mariette S, Streiff R, Bodénès C, Kremer A (2000) Comparison of microsatellites and amplified fragment markers for parentage analysis. *Molecular Ecology*, **9**, 1037–1048.

Gilbert-Horvath EA, Larson RJ, Garza JC (2006) Temporal recruitment patterns and gene flow in kelp rockfish (*Sebastes atrovirens*). *Molecular Ecology*, **15**, 3801–3815.

Hauser L, Baird M, Hilborn R, Seeb LW, Seeb JE (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.

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

Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. *Molecular Ecology*, **12**, 2511–2523.

Jones AG, Small CM, Paczolt KA, Ratterman NL (2010) A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, **10**, 6–30.

Kaiser SA, Taylor SA, Chen N *et al.* (2017) A comparative assessment of SNP and microsatellite markers for assigning parentage in a socially monogamous bird. *Molecular Ecology Resources*, **17**, 183–193.

Kearse M, Moir R, Wilson A *et al.* (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Kent WJ (2002) BLAT — The BLAST -Like Alignment Tool. *Genome Research*, **12**, 656–664.

Kidd KK, Pakstis AJ, Speed WC *et al.* (2013) Microhaplotype loci are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series*, **4**, e123–e124.

Kidd KK, Pakstis AJ, Speed WC *et al.* (2014) Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Forensic Science International: Genetics*, **12**, 215–224.

Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM (2013) Mendel: The Swiss army knife of genetic analysis programs. *Bioinformatics*, **29**, 1568–1570

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Magoč T, Salzberg SL (2011) FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

McKinney GJ, Seeb JE, Seeb LW (2017) Managing mixed-stock fisheries: genotyping multi-SNP haplotypes increases power for genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, doi:cjfas–2016–0443.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Morin PA, Luikart G, Wayne RK (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.

Narum SR, Banks M, Beacham TD *et al.* (2008) Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology*, **17**, 3464–3477.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–7.

Pakstis AJ, Fang R, Furtado MR, Kidd JR, Kidd KK (2012) Mini-haplotypes as lineage informative SNPs and ancestry inference SNPs. *European Journal of Human Genetics*, **20**, 1148–1154.

Parker PG, Snow AA, Schug MD, Booton GC, Fuerst PA (1998) What molecules can tell us about populations: choosing and using a molecular marker. *Ecology*, **79**, 361–382.

Pastor T, Garza JC, Allen P, Amos W, Aguilar A (2004) Low genetic variability in the highly endangered Mediterranean monk seal. *Journal of Heredity*, **95**, 291-300.

Pemberton JM (2008) Wild pedigrees: the way forward. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 613–621.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLOS ONE*, **7**, e37135.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*, **73**, 1402–1422.

Seeb JE, Pascal CE, Ramakrishnan R, Seeb LW (2009) SNP genotyping by the 5’-nuclease reaction: advances in high-throughput genotyping with nonmodel organisms. In: *Single Nucleotide Polymorphisms: Methods and Protocols* (ed Komar AA), pp. 277–292. Humana Press, Totowa, NJ.

Seeb LW, Antonovich A, Banks MA *et al.* (2007) Development of a standardized DNA database for Chinook salmon. *Fisheries*, **32**, 540–552.

Wang J (2004) Sibship reconstruction from genetic data with typing errors. *Genetics*, **166**, 1963–1979.

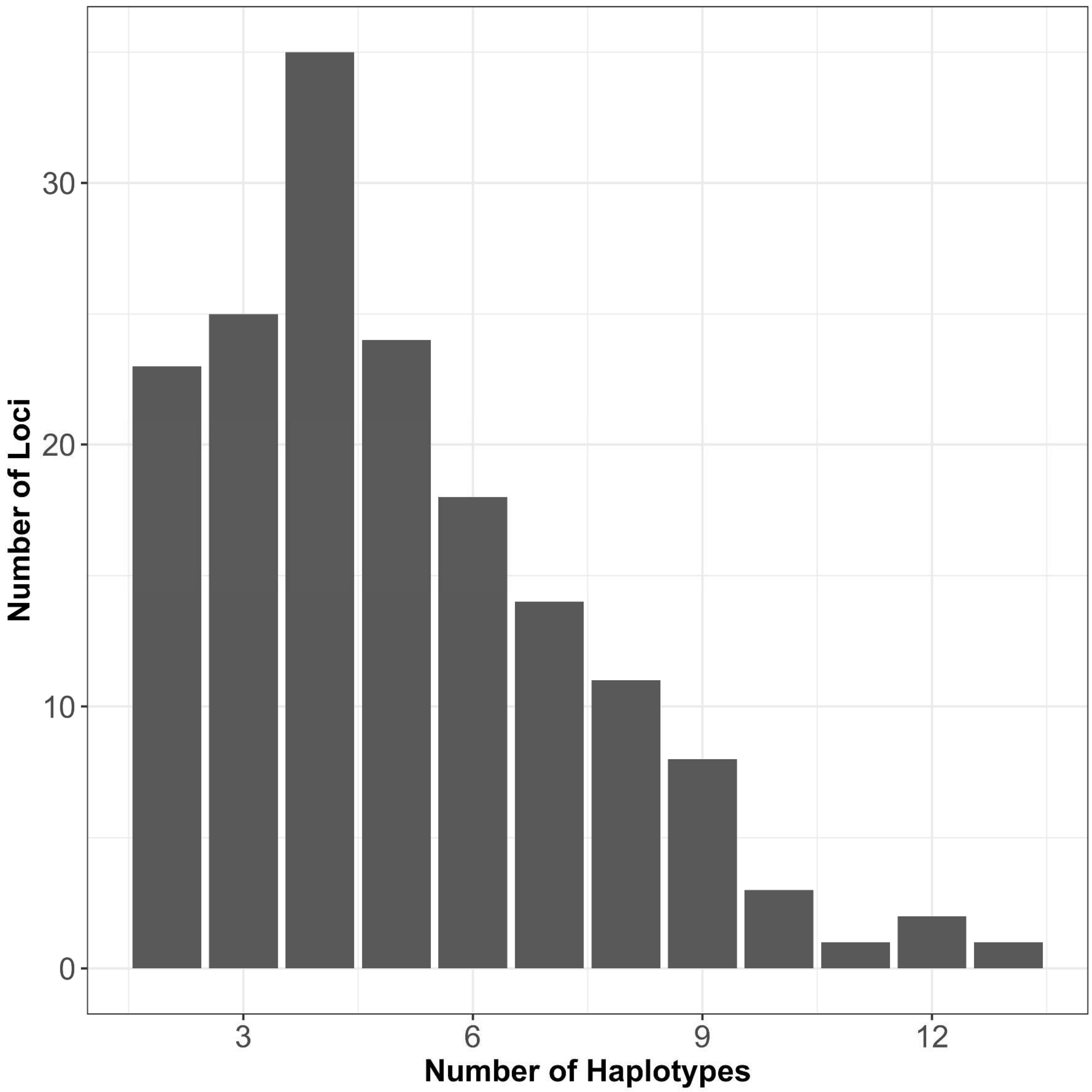
Weinman LR, Solomon JW, Rubenstein DR (2015) A comparison of single nucleotide polymorphism and microsatellite markers for analysis of parentage and kinship in a cooperatively breeding bird. *Molecular Ecology Resources*, **15**, 502–511.

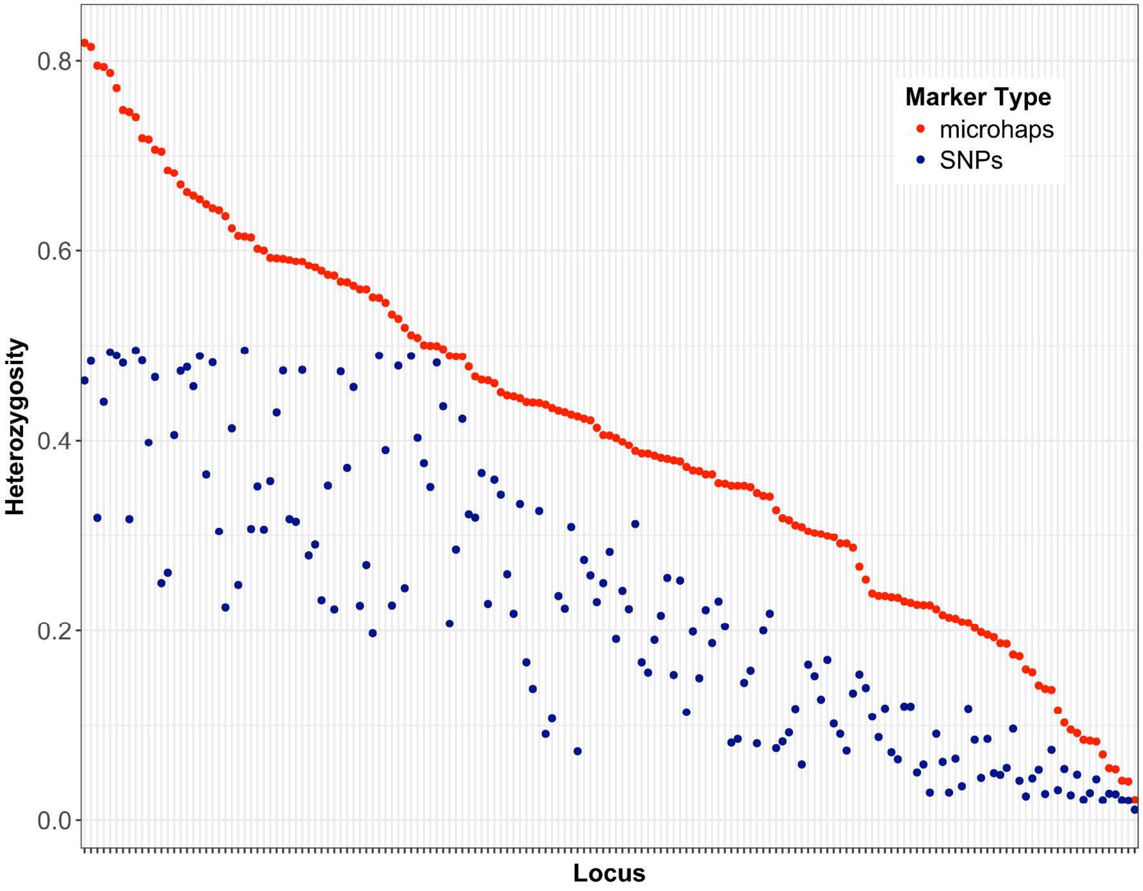
Figure 1. Distribution of microhaplotypes across 165 loci in 144 kelp rockfish samples. The number of haplotypes per locus ranged from two to 13.

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.

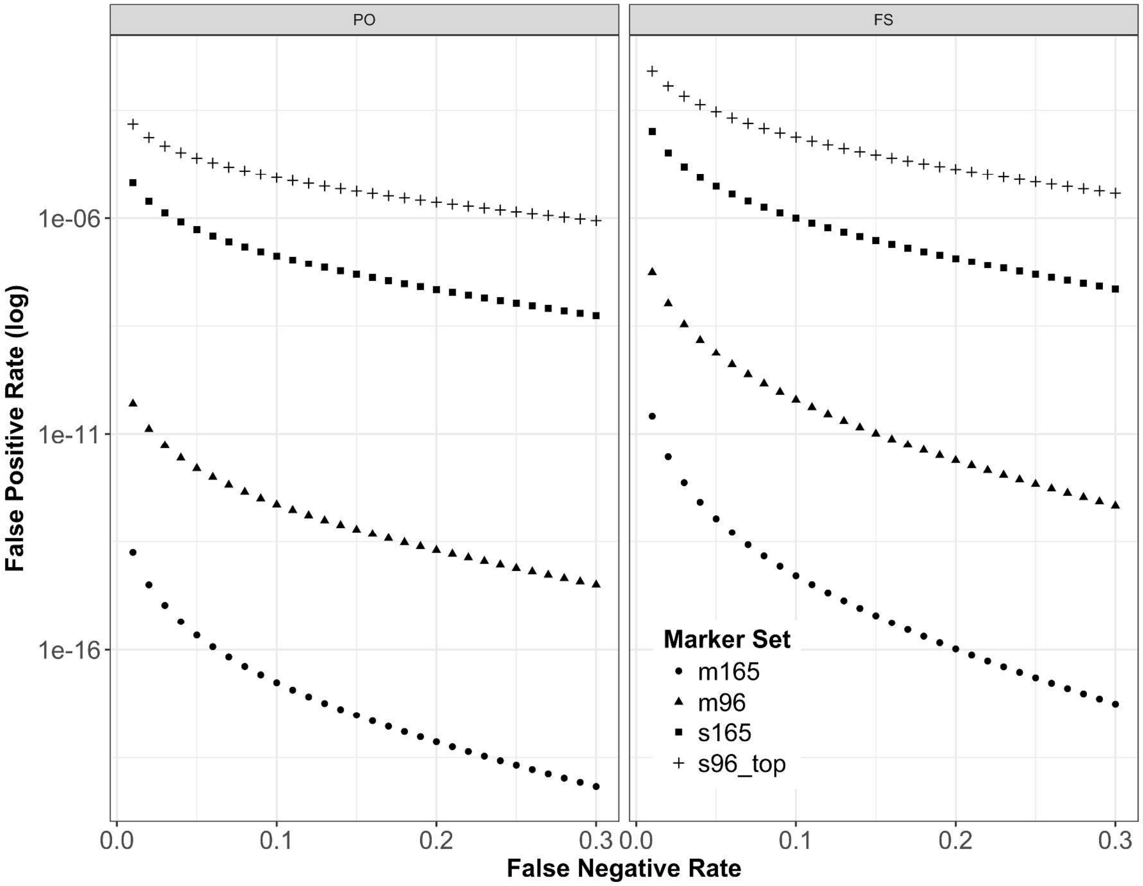
Figure 3. Simulated false positive rates for matching (a) single parents with offspring and for (b) full-siblings at a given false negative rate using the four sets of markers: 165 microhaplotypes (m165), 165 SNPs (s165), 96 microhaps (m96) and 96 SNPs with the highest heterozygosity (s96_top).

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.

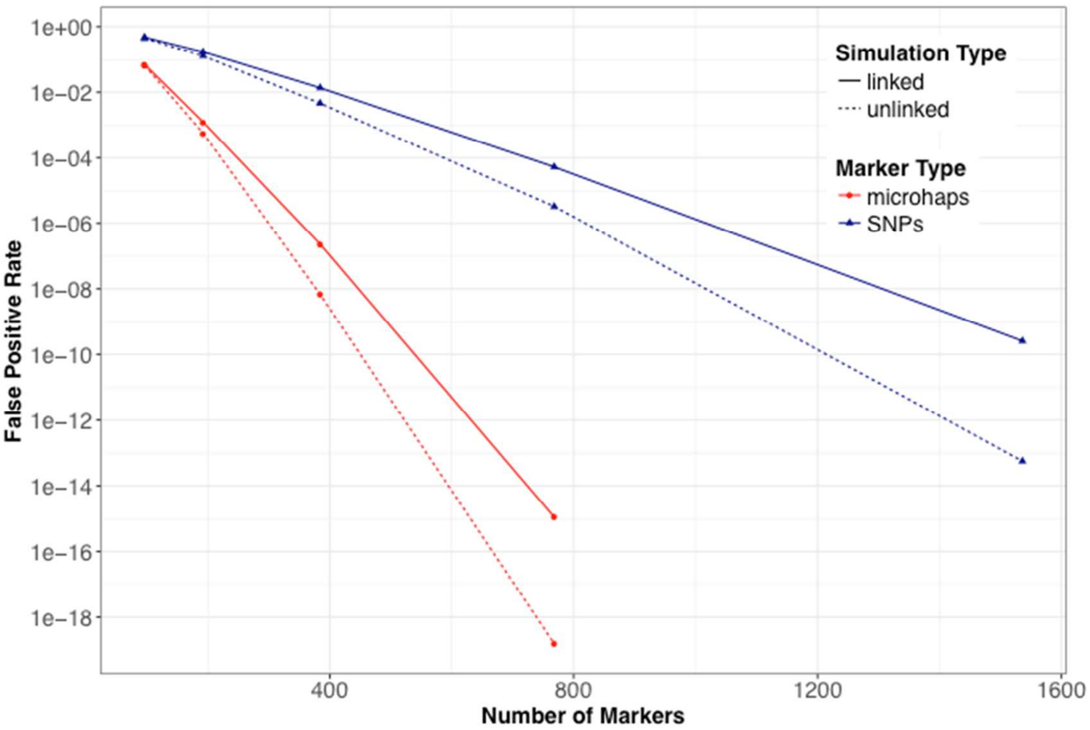

Number of Loci
30
20
10
0
3
6
9
12
Number of Haplotypes


Marker Type
● microhaps
● SNPs
Heterozygosity
0.0
0.2
0.4
0.6
0.8
Locus

658


PO
FS
1e-06
1e-11
1e-16
False Positive Rate (log)
0.0
0.1
0.2
0.3
False Negative Rate
Marker Set
● m165
▲ m96
■ s165
+ s96_top

659
660


False Positive Rate
1e+00
1e−02
1e−04
1e−06
1e−08
1e−10
1e−12
1e−14
1e−16
1e−18
400
800
1200
1600
Number of Markers
Simulation Type
linked
unlinked
Marker Type
microhaps
SNPs

661