MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 502–511 doi: 10.1111/1755-0998.12330

A comparison of single nucleotide polymorphism and microsatellite markers for analysis of parentage and kinship in a cooperatively breeding bird

LUCIA R. WEINMAN,* JOSEPH W. SOLOMON* and DUSTIN R. RUBENSTEIN*,†
*Department of Ecology, Evolution and Environmental Biology, Columbia University, 1200 Amsterdam Avenue, New York, NY 10027, USA, †Department of Ornithology and Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA

Abstract

The development of genetic markers has revolutionized molecular studies within and among populations. Although poly-allelic microsatellites are the most commonly used genetic marker for within-population studies of free-living animals, biallelic single nucleotide polymorphisms, or SNPs, have also emerged as a viable option for use in non-model systems. We describe a robust method of SNP discovery from the transcriptome of a nonmodel organism that resulted in more than 99% of the markers working successfully during genotyping. We then compare the use of 102 novel SNPs with 15 previously developed microsatellites for studies of parentage and kinship in cooperatively breeding superb starlings (*Lamprotornis superbus*) that live in highly kin-structured groups. For 95% of the offspring surveyed, SNPs and microsatellites identified the same genetic father, but only when behavioural information about the likely parents at a nest was included to aid in assignment. Moreover, when such behavioural information was available, the number of SNPs necessary for successful parentage assignment was reduced by half. However, in a few cases where candidate fathers were highly related, SNPs did a better job at assigning fathers than microsatellites. Despite high variation between individual pairwise relatedness values, microsatellites and SNPs performed equally well in kinship analyses. This study is the first to compare SNPs and microsatellites for analyses of parentage and relatedness in a species that lives in groups with a complex social and kin structure. It should also prove informative for those interested in developing SNP loci from transcriptome data when published genomes are unavailable.

Keywords: cooperative breeding, kinship, microsatellite, parentage, relatedness, single nucleotide polymorphism, transcriptome

Received 13 February 2014; revision received 8 September 2014; accepted 10 September 2014

Introduction

The development of genetic markers and statistical methods for analysis of familial relationships in wild populations has enabled increasingly precise examination of genetic structure within and among populations of free-living organisms (Avise 1994; Hughes 1998; Sunnucks 2000; Jones & Ardren 2003; Garant & Kruuk 2005; Jones *et al.* 2010). Since the discovery that genetic monogamy is rare in birds (Gibbs *et al.* 1990; Westneat 1990), social insects (Boomsma & Ratnieks 1996) and other animals (Uller & Olsson 2008; Cohas & Allainé 2009), genetic markers have also allowed for detailed studies of parentage and mating systems. Microsatellites are the most commonly used codominant genetic marker for studies of parentage and kinship in nonmodel organisms (Glaubitz *et al.* 2003; Jones *et al.* 2010). These short DNA sequences are comprised of locus-specific, 2–5 nucleotide-long motifs repeated a variable number of times (Morin *et al.* 2004). Alleles of a microsatellite locus differ by the number of repeats in the sequence, and because loci can be highly polymorphic, large numbers of individuals within a population can be distinguished from one another using relatively few loci. Moreover, next-generation sequencing methods have made the identification of microsatellite loci easier and more affordable (Guichoux *et al.* 2011).

Another codominant marker that has seen increasing use in molecular studies of both model and nonmodel organisms is the single nucleotide polymorphism (SNP). SNPs are sequences of DNA for which two alleles of a

Correspondence: Dustin R. Rubenstein, Fax: 212-854-8188;
E-mail: dr2497@columbia.edu

particular locus differ by only one base pair in an otherwise identical sequence. The resolving power of a set of markers depends in large part on the number of independent alleles per locus and their frequency. Because SNPs generally only have two independent alleles per locus (but see Casci 2010; Zha *et al.* 2012), while microsatellites can have many more, microsatellites have so far dominated in studies of nonmodel, free-living organisms (Glaubitz *et al.* 2003; Jones *et al.* 2010). In contrast, SNPs have been more widely developed for genetic analyses in model organisms and domestic animals (Anderson & Garza 2006). SNPs have a variety of benefits over microsatellites in that they are more easily quantified on a per locus basis, have lower rates of genotyping error, and are cheaper to genotype per locus (Jones *et al.* 2010). SNPs are ideal for population genetics (Helyar *et al.* 2011), particularly in studies involving large numbers of individuals (Anderson & Garza 2006), because their patterns of evolution are well described by simple statistical models (Morin *et al.* 2004). They are also emerging as a viable option for parentage analysis of wild populations. For example, SNPs have been used in combination with microsatellites to examine rates of extra-pair paternity in banded wrens (Cramer *et al.* 2011). Simulations of parentage analysis showed that SNPs were more useful than microsatellites due to higher heterozygosity for studies of inbred European bison (Tokarska *et al.* 2009). In crustaceans, SNPs had similar or better rates of success to microsatellites in both parentage and pedigree assignments, despite having lower heterozygosity and lower numbers of independent alleles (Sellars *et al.* 2012). Additionally, increasing the number of SNPs improved assignment success. In studies of fish, similar rates of success between SNPs and microsatellites for parentage and kinship analysis were attributed to similar numbers of effective alleles (based on heterozygosity and number of independent alleles) (Hauser *et al.* 2011). As methods for SNP identification in nonmodel organisms become easier and cheaper (Helyar *et al.* 2011), it will be important to determine how well they compare to microsatellites for studies of parentage and kinship in a range of species, particularly those that live in societies with complex kin structure where determining paternity can be difficult because highly related individuals may complicate assignment.

Here, we compare 102 novel SNPs with 15 microsatellites that have been used previously in analyses of parentage and kinship in a free-living population of cooperatively breeding superb starlings, *Lamprotornis superbus*. Superb starlings are plural cooperative breeders with helpers of both sexes that aid breeding pairs in feeding offspring and defending the nest from predators. Males are philopatric, resulting in higher genetic relatedness among males (i.e. potential fathers) than females (Rubenstein 2006). Extra-pair offspring sired by both subordinate males within the group as well as dominant males from outside the group occur in about 17–25% of nests (Rubenstein 2007a,b; D. R. Rubenstein unpublished data). To identify SNPs for this study, we first sequenced and assembled a transcriptome for the superb starling. We then compared the utility of microsatellites and SNPs in assigning parentage and in generating pairwise relatedness values. To do this, we genotyped more than 200 individual starlings with both sets of markers. We compared the results of parentage analyses using both marker systems and the resolving power of analyses employing different numbers of SNPs. Although a variety of recent studies have compared the usefulness of these two marker systems for parentage and relatedness (Hauser *et al.* 2011; Sellars *et al.* 2012; Steele *et al.* 2013), to our knowledge this is the first to do so in a species that lives in groups with a complex social and kin structure, where it is often difficult to accurately assign paternity because potential fathers can be highly related.

Materials and methods

Study system

Breeding activities of nine social groups of superb starlings were monitored through December 2012 (seven groups since April 2001 and two groups since January 2002) at the Mpala Research Centre, Laikipia, Kenya (0°17′ N, 37°52′ E). Superb starlings are endemic to East African savannas and live in extended family groups of up to 45 individuals (mean group size = 25.8). Groups contain up to six socially monogamous breeding pairs and a variable number of nonbreeding helpers that aid in the care of offspring. Birds defend territories year-round and breed during both the long (March–May) and short rains (October–November; Rubenstein 2007b, 2011). We observed groups continually throughout the year to identify the social parents and helpers at each active nest using intensive focal observations.

Sample collection

As of 2012, approximately 95% of the population had been captured (897 individuals). Birds were caught during the breeding season at active nests with mist nets or during the nonbreeding season using traps baited with papaya or corn meal powder (Rubenstein 2007a). Nestlings were banded and bled 7 days after hatching. Blood from superb starlings was collected in Queen's lysis buffer (Seutin *et al.* 1991), and genomic DNA was extracted using a DNeasy Blood & Tissue Kit (Qiagen). RNA for cDNA conversion and transcriptome assembly was extracted from the spleen using an RNeasy Tissue

Kit (Qiagen). All work was approved by Columbia University's Institutional Animal Use and Care Committee (AC-AAAB1128).

Sample size

During the course of our long-term study, we genotyped 897 individuals, including 549 offspring born into the nine social groups. CERVUS version 3.0 was used to assign parentage to the microsatellite-genotyped individuals (Rubenstein 2007a,b; D. R. Rubenstein unpublished data). By cross-referencing these parentage assignments with behavioural observations, we found that 37 of 208 (18%) nests from 2001 to 2012 contained extra-pair offspring. To compare the utility of SNP and microsatellite markers for assigning parentage, we selected a subset of 153 offspring for SNP genotyping, including all 93 offspring from 37 extra-pair nests and 60 offspring from 23 nests that did not contain extra-pair offspring (hereafter within-pair nests). Additionally, the mothers, social fathers (i.e. breeding partners) and extra-pair sires of each of these offspring as identified by microsatellite analysis of the entire population were also included, bringing the total number of individuals selected for SNP genotyping to 224. We note that some birds bred in multiple years (often with different mates), and some offspring went on to breed as adults. Although using only a subset of the adult males in the total population as candidate fathers in this study could bias paternity assignments, the focus of this study was not to assign fathers per se, but to determine how assignments made using SNPs differ from those made using microsatellites. Furthermore, all genotyped males were included as candidate fathers, so both types of markers ultimately had to distinguish among 105 candidates.

Microsatellite genotyping

Microsatellite markers were previously isolated using a modified enrichment protocol (Hamilton *et al.* 1999) that incorporated magnetic capture of streptavidin beads with biotinylated probes bound to microsatellite-containing genomic fragments (Kijas *et al.* 1994) isolated from blood (Rubenstein 2005). We chose 15 microsatellite markers from a set of 31 polymorphic loci developed for superb starlings (Rubenstein 2005) that have been used previously for studies of parentage and relatedness (Rubenstein 2007a,b). All of these markers were in Hardy–Weinberg equilibrium and did not contain null alleles. The forward primer of each pair was labelled using the fluorescent dyes 6-FAM, NED, PET or VIC (Applied Biosystems) and polymerase chain reaction (PCR) was conducted using methods and conditions published previously (Rubenstein 2005). Genotyping was performed on 3130xl, 3130 and 3100 Genetic Analyzers (Life Technologies). All alleles were scored automatically, but confirmed visually using GENEIOUS version 6.1 (Biomatters). Approximately 25% of the genotypes were rerun to ensure accuracy of the scoring, and for each locus, >100 individuals were screened twice to estimate genotyping errors; no such errors were found. Ultimately, there were no missing data in our sample, as every individual was scored at each locus.

Transcriptome assembly

Transcriptome assembly was conducted for the purpose of SNP discovery. Briefly, total RNA extractions from four females were submitted to the Genome Sequencing and Analysis Facility at the University of Texas at Austin. Samples were checked for quality on an Agilent BioAnalyzer 2100 and then treated using Ribo-Zero rRNA depletion [Epicentre Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat)], which removes rRNA through magnetic bead capture. The four samples underwent TruSeq library preparation, mRNA polyA selection and cDNA conversion prior to 2 × 101 bp paired-end barcoded sequencing over two shared lanes on an Illumina HiSeq 2000. The total read count across the four demultiplexed samples was 318 781 788 paired-end reads, with an additional 201 381 235 single-end reads obtained from an initial run that did not proceed past the paired-end turnaround chemistry step (Table S1, Supporting information).

The raw fastQ files were evaluated using FASTQC version 0.10.1 and trimmed for quality and adapter contamination using TRIMMOMATIC version 0.27 with the parameters: 2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 HEADCROP:12 MINLEN:36, corresponding to MacManes (2014). After quality trimming, the remaining 271 021 514 paired-end and 190 112 938 single-end reads were de novo assembled using the Trinity platform version 2013-02-25 (Grabherr *et al.* 2011; Haas *et al.* 2013). For more efficient processing, reads were in silico normalized prior to assembly with a targeted maximum coverage of 30. The resulting assembly consisted of 939 687 total transcripts and 470 614 components (inclusive of lowly-expressed isoforms) with an N50 of 4422. Paired-end reads from all samples were then aligned back to the Trinity assembly as a reference using BWA version 0.7.4-r385 (Li & Durbin 2009). The extra single-end reads were excluded from alignments so as to more reliably identify sequence duplicates and reduce potential false positives in variant calling. To ensure the highest quality results and downstream compatibility with the SAMTOOLS version 0.1.19 mpileup variant calling function (Li *et al.* 2009) and the Genome Analysis Toolkit version 2.5-2 (GATK; McKenna *et al.*

2010), two separate alignments were performed using BWA-MEM (Li 2013a) and BWA-backtrack/sampe (Li & Durbin 2009).

SNP marker development

Alignments were processed and validated, and sequence duplicates were identified with PICARD-TOOLS version 1.92 and MarkDuplicates (Wysoker *et al.* 2013). To be conservative, we used two different pipeline methods to identify SNPs and only considered loci recognized by both methods. BWA-MEM alignments were used for variant calling in the mpileup pipeline, whereas BWA sampe alignments were used in the GATK pipeline. SAMTOOLS variant calling was conducted on only properly paired alignments using mpileup recommended parameters of -D -S -C50 (Li 2013b). Variants were then filtered using vcfutils.pl varFilter function for mapping quality, coverage and proximity to other variants or gaps, with a minimum SNP RMS mapping quality (*-Q*) of 30, minimum read depth (*-d*) of 30, maximum read depth (*-D*) of 100, minimum indel score for nearby SNP filtering (*-G*) of 10, gap-filtering of SNPs within bp range (*-w*) of 60 and a window size (*W*) for filtering dense SNPs of 500 (gap parameters are those suggested in Fass 2010).

Within the GATK version 2.5-2 pipeline (McKenna *et al.* 2010), and following GATK Best Practices, we performed indel realignment on only properly paired alignments and ran HaplotypeCaller for variant discovery with -minPruning set to 5 for efficiency given the size of our de novo transcriptome data set. Variant calls were then stringently hard-filtered using GATK recommendations and standard exome analysis practices for mapping quality, coverage, strand-bias, MQRankSum, ReadPosRankSum and masking for called indels (see Appendix S1, Supporting information). Base quality score recalibration was completed using the highest confidence SNPs as known SNPs, and HaplotypeCaller was then rerun with another round of variant filtration (DePristo *et al.* 2011; Van der Auwera *et al.* 2013).

The GATK HaplotypeCaller pipeline, with conservative and repeated filtering, resulted in 64 536 SNP VCF entries versus 208 642 produced by SAMTOOLS mpileup. Employing GATK SelectVariants, we created a concordance file of 40 458 SNPs from the two VCF outputs for all four individual samples. With VCFTOOLS version 0.1.10 (Danecek *et al.* 2011), we calculated a transition/transversion ratio of 2.29 for the concordance set, and found that of these 40 458 SNPs, 40 428 (99.9%) passed a Hardy–Weinberg equilibrium test with an alpha threshold of 0.001. Because we were not interested in mapping low-frequency variants within the superb starling population, we derived a subset of high heterozygosity SNPs setting the minor allele threshold to 0.3. We then selected sequences with >300 bp flanking sequence on both sides to design primers for Sequenom MassArray SNP genotyping. This process resulted in 9376 SNPs that met all of these criteria. Of these 9376 SNPs, we derived a sample of 305 SNPs that were discovered at intervals evenly spread across the Trinity assembly putative 'gene' components and that passed GATK Validation Amplicons (which we used to convert SNP VCF entries to FASTA format for Sequenom Assay Design).

SNP Primer design

Primers for Sequenom's iPLEX Gold reaction were designed for the 305 chosen SNPs using the Sequenom online Assay Design Suite version 1.0, with Extend Primer Assay Design version 4.1.0.17. The Design Suite was set to multiplex wells up to 36-plex, the maximum currently allowed (see Appendix S1, Supporting information for design parameters). Because our a priori target for this study was 100 SNP loci (*sensu* Hauser *et al.* 2011), we chose the three wells with the highest plex for genotyping; the first 36-plex had a multiplex SNP capture confidence score of 70.6%, the second 35-plex had a score of 65.9%, and the third 32-plex had a score of 74.2%. This resulted in 103 SNP loci spread across the three plexes (see Table S2, Supporting information for primer information).

SNP genotyping

SNP genotyping was conducted at the Taub Genomics Core Facility at the Columbia University Medical Center, using the iPLEX Gold assay on a Sequenom MassArray. SNP loci were amplified from genomic DNA using standard PCR. A second specialized PCR reaction yielded an allele-specific difference in mass for each SNP locus. Mass spectrometry of the amplified samples assigned a genotype at each locus for each individual. We ran 11% of samples in duplicate as a control; duplicate samples were checked for concordance and no errors were found. Of the 103 SNPs selected for genotyping, one failed across all 224 individuals. Thus, >99% of our SNP loci were successful for genotyping. The mean call rate for the remaining 102 loci was 97%; 72 of these had a 100% call rate and the rest ranged from 94% to 99.6%.

Parentage analysis

We used both CERVUS version 3.0 (Kalinowski *et al.* 2007) and COLONY2 (Jones & Wang 2010) to assign parentage with microsatellite and SNP genotype data separately and together. CERVUS uses maximum-likelihood methods to predict parent-offspring relationships. For each offspring, a metric of the likelihood that each potential

father is the actual father is calculated as a log-likelihood (LOD) score. CERVUS uses the range of LOD score distributions to calculate a critical LOD score against which LOD scores from the actual population can be evaluated. LOD scores are then computed for each potential father using their actual genotype for each individual offspring. The critical LOD score calculated from the simulation step is used to assign fathers to each offspring (Kalinowski *et al.* 2007). In contrast, COLONY2 utilizes a full-pedigree likelihood method that jointly infers parentage and sibship. The program considers the likelihoods of entire pedigree configurations, rather than individual pairs of offspring and parents (Jones & Wang 2010).

We divided the parentage analyses into three groups: (i) those using microsatellite genotypes only (15 loci); (ii) those using SNP genotypes only (102 loci); and (iii) those using a combined genotype of microsatellites and SNPs (117 loci; Table 1). We also ran 18 separate analyses comprising the 10, 20, 30, 40, 50, 60, 70, 80 and 90 most and least heterozygous loci to compare the power of different combinations of SNP markers. In CERVUS, separate parentage analyses were also conducted with SNPs and microsatellites that incorporated an estimation of relatedness of candidate males to each other. Relatedness of candidate males was set to 0.12, and the proportion of candidate fathers that are related was estimated to be 0.5 (L. Pollack & D. R. Rubenstein unpublished data). The use of this information in assignment did not qualitatively change the results. In all CERVUS analyses, parentage was assigned at the default levels of 80% (relaxed) and 95% (strict) confidence. Known mothers (based on focal observations at the nest, but also confirmed through microsatellite-based parentage analysis of the entire population) were included in the analysis for all offspring. SNPs identified these mothers as the genetic mothers, confirming previous results from microsatellites. Importantly, in 14 years of studying this population, no evidence of intraspecific brood parasitism, or egg-dumping by females, exists; molecular analysis from microsatellites and SNPs confirmed this observation. All CERVUS analyses assumed a 1% genotype error (the program default) and an estimate of 10 candidate fathers for any given offspring in any given year (*sensu* Rubenstein 2007a,b). In COLONY2, assignments were made using one run with medium run lengths and likelihood precision, as well as all other program defaults. The mating system was assumed to be polygamous for both males and females, and to be without inbreeding. A list of known maternal sibships was provided, as well as a list of males to exclude as candidate fathers for each offspring, based on birth date. For both programs, all 105 genotyped males were considered as candidates for each offspring in each analysis.

Assignments made in COLONY2 using the full-pedigree method agreed with those made in CERVUS for both microsatellites and SNPs for 149 of 153 (97%) offspring. In the remaining four cases, COLONY2 was unable to assign a father using SNPs with a probability >0.001, but it assigned the same fathers as CERVUS did when using microsatellites. When COLONY2 used the pairwise likelihood method, all assignments agreed with those made in CERVUS when using both SNPs and microsatellites. Because the programs largely agreed on all paternity assignments, we present the results from CERVUS only.

Relatedness analysis

KINGROUP version 2 (Konovalov *et al.* 2004) was used to calculate pairwise relatedness values with likelihood ratio tests based on Queller and Goodnight's *r* (Queller & Goodnight 1989; Goodnight & Queller 1999). Matrices

Table 1 Comparison of marker characteristics for microsatellites, single nucleotide polymorphisms (SNPs) and the combined marker system, including the number of individuals genotyped, the number of loci used, the mean number of alleles per locus, the mean proportion of individuals genotyped, mean expected (H_E) and observed heterozygosities (H_O), mean polymorphic information content (PIC) and the combined nonexclusion probabilities for SNPs, microsatellites and the combined marker system of SNPs and microsatellites

Parameter	SNPs	Microsatellites	SNPs and Microsatellites
Individuals genotyped	224	224	224
Number loci used	102	15	117
Mean alleles per locus	2.00	6.20	2.54
Range of alleles per locus	—	4–16	2–16
Mean proportion individuals typed	0.99	1.00	0.99
Mean H_E	0.40	0.51	0.42
Mean H_O	0.38	0.54	0.40
Mean PIC	0.32	0.48	0.34
Nonexclusion probability (first parent)	8.4×10^{-5}	2.7×10^{-2}	2.3×10^{-6}
Nonexclusion probability (second parent)	2×10^{-8}	1.2×10^{-3}	2.68×10^{-11}
Nonexclusion probability (parent pair)	3.53×10^{-13}	9.4×10^{-6}	3.31×10^{-18}

of pairwise *r* values were calculated separately using microsatellites and SNPs. These matrices were then compared with a Mantel test based on a Pearson's product-moment correlation in the R package 'vegan' (Oksanen *et al.* 2013). Significance was tested using 999 permutations.

Results

Marker characteristics

The mean number of independent alleles per locus was 6.2 (range 4–16 per locus) for microsatellites and 2 for SNPs (Table 1). The mean observed and expected heterozygosities, as well as mean polymorphic information content (PIC), were greater for microsatellites, but nonexclusion probabilities were lower for SNPs (Table 1). CERVUS identified six SNPs that were not in Hardy–Weinberg equilibrium. However, because this was a highly kin-structured population and our samples represent a nonrandom subset of the group, we chose to include these markers in the analyses. Importantly, removal of these six loci did not qualitatively change the results, and no consistent patterns were observed between loci not found to be in Hardy–Weinberg equilibrium and locus mismatches between parents and offspring.

Comparison of SNP and microsatellite assignments

Of the parentage assignments that CERVUS made using only SNPs, 141 of 153 (92.2%) offspring were assigned with 95% confidence, 8 of 153 (5.2%) offspring were assigned with 80% confidence and the remaining 4 of 153 (2.5%) offspring were not assigned a father. Similar probabilities were made using only microsatellites, with 142 of 153 (92.8%) offspring assigned with 95% confidence, 8 of 153 (5.2%) offspring with 80% confidence and 3 of 153 (2.0%) offspring not assigned a father. Combining marker types improved assignments slightly, as 147 of 153 (96.0%) offspring were assigned with 95% confidence, 3 of 153 (2.0%) offspring with 80% confidence and 3 of 153 (2.0%) offspring were not assigned a father.

CERVUS generally assigned the same fathers whether SNPs or microsatellites were used. For 130 of 153 (85%) offspring, both markers assigned the same fathers (Table 2). On average, there were 2.28 mismatches for the SNP assignments and 0 mismatches for the microsatellite assignments (Table 2). In 112 of 130 (86%) cases, SNPs and microsatellites assigned the same father based on LOD score alone (which for 90 of these cases could be confirmed where behavioural information was available; Table 2). For the remaining 18 of 130 (14%) cases, SNPs and microsatellites assigned different fathers based on LOD scores alone (Table 2). For 11 of these 18 cases, behavioural data confirmed the father top ranked by SNPs as the social father and ruled out those ranked above him by microsatellites as other close relatives (10 cases), or confirmed the father top ranked by microsatellites as the social father and ruled out those ranked above him by SNPs (1 case). In each of these cases, the male in question still had the same number of mismatches (<1 mismatches for microsatellites, <5 mismatches for SNPs) as the top-ranked male.

Table 2 Comparison of CERVUS parentage assignments for single nucleotide polymorphism (SNP) and microsatellite markers. Assignments are separated into the number of cases in which the analyses using each marker system agree and disagree on the same father using log-likelihood (LOD) scores alone. Cases in which they disagree based on LOD scores alone are broken down into (i) those that were resolved when behavioural information was considered, (ii) those that remained unresolved when behavioural information was considered and (iii) those for which no relevant behavioural observations were available. The mean number of locus mismatches for each marker system is detailed

		Mean number of locus mismatches		
LOD assignments	*n*	Microsatellites	SNPs	Combined
Agreements (assignment of same father)	112	0.01	2.20	2.21
Disagreements (assignment of different father)				
Resolved by behavioural data	11	0.09	1.18	1.57
Unresolved by behavioural data	4	0	2	2
No behavioural data available	3	0	0.33	0.33

The remaining 7 of 18 cases were more difficult to assign parentage because the two potential fathers for all seven offspring were themselves a father and son pair with an unusually high genetic relatedness ($r = 0.8$). These seven offspring belonged to three broods, two broods with three offspring and one brood with a single offspring. For the first brood with three offspring, behavioural data revealed that SNPs had assigned an extra-pair male while giving the social father a low or negative LOD score and a high number of mismatches. However, microsatellites gave positive LOD scores and identified a low number of mismatches to both the social father and the putative extra-pair male, suggesting that microsatellites could not accurately discriminate among the potential fathers. For the brood with one offspring, behavioural data showed that SNPs had assigned the social father while giving the other candidate male a negative LOD score and a high number of mismatches.

However, for this same offspring, microsatellites generated positive LOD scores and low numbers of mismatches to both candidate males. For the brood with three offspring, no relevant behavioural information about the social father was available (Table 2). In these three cases, SNPs assigned one male to all three offspring while giving the other candidate male a low or negative LOD score and a high number of mismatches. In contrast, microsatellites gave positive LOD scores and low numbers of mismatches to both candidate males. Interestingly, in COLONY2, there was no disagreement in assignments for the brood with one offspring or the second brood with three offspring. For the first brood with three offspring, SNPs again assigned the extra-pair father over the social father for all three offspring, while microsatellites assigned the extra-pair father for one of the offspring, and the social father for the other two.

For 15% of the offspring (n = 23 individuals), no consensus between assignments made with microsatellites and assignments made with SNPs could be found, and no male was a good genetic match by either marker system. Assignments made with SNPs had ≥6 mismatches (SNP mean = 9.3), whereas assignments made with microsatellites had ≥2 mismatches (mean = 1.8). We concluded that the genetic fathers of these offspring have not yet been captured and are extra-group extra-pair males (Rubenstein 2007a).

In summary, for 146 of 153 offspring genotyped (95%), microsatellites and SNPs either identified the same father (when behavioural data were included) or suggested that the genetic father had not been captured. For the remaining 7 of 153 (5%) of offspring, SNPs could clearly distinguish between two closely related candidate fathers, whereas microsatellites could not. Agreement between marker systems improved to 150 of 153 offspring (98%) when assignments were made in COLONY2. COLONY2 likely had greater success than CERVUS because of the program's greater ability to distinguish between the two closely related candidate fathers in those seven cases. Importantly, the addition of behavioural information greatly improved assignment results in all cases. In the three cases where no behavioural data were available, SNPs were better able to distinguish between two closely related candidate fathers than microsatellites.

Comparison of assignments from different numbers of SNPs

For the comparison of varying numbers of SNPs, we excluded the 23 offspring whose fathers were not captured, resulting in a sample size of 130 individuals. In each comparison, we quantified the number of assignments from CERVUS that were not the same as those made with all 102 SNPs (Fig. 1). Comparisons of the groups of increasing numbers of the most heterozygous and least heterozygous SNPs were also made both with and without reference to behavioural information (Fig. 1). Importantly, the inclusion of behavioural information greatly improved the performance of the markers; >60 of the most or least heterozygous SNPs had no unresolved disagreements when behavioural information was taken into account, and the group of 50 SNPs had only one disagreement. Thus, a reduced set of ~60 SNP loci with a mean heterozygosity of 0.41 spread over two wells is sufficient for assigning paternity in a kin-structured population, but only when behavioural information is available to aid in assignment. When behavioural data are unavailable, >60 SNPs are recommended (Fig. 1).

Comparison of microsatellite and SNP relatedness values

A comparison of relatedness values generated for the microsatellites and SNPs revealed that despite high variation in individual pairwise r values (Fig. S1, Supporting information), the two matrices were highly correlated

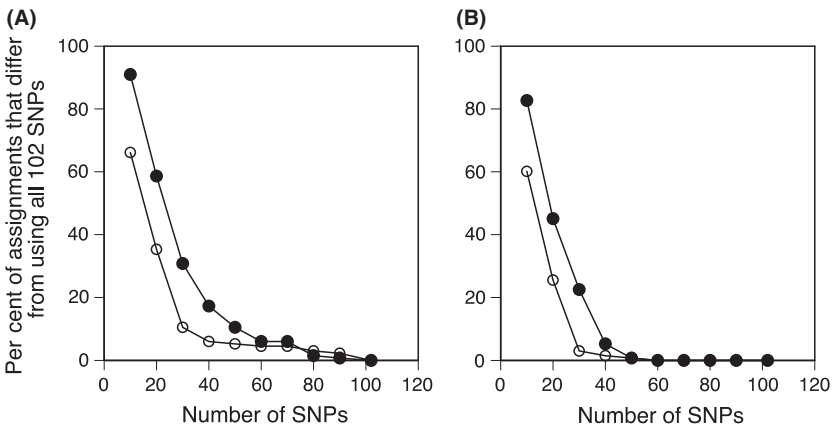

Fig. 1 Percentage of CERVUS parentage assignments made using different numbers of the most heterozygous (open circles) and least heterozygous (closed circles) single nucleotide polymorphisms (SNPs) that differ from those made by using all 102 SNPs for the 130 offspring where SNPs and microsatellites identified the same fathers. Assignments were made (A) based only on SNP loci or (B) based on SNP loci and behavioural observations.

(Mantel $r = -0.26$, $P = 0.001$). Upper quantiles of null model permutations ranged from 0.013 (90%) to 0.023 (99%). The observed cluster of high relatedness in Fig. S1 (Supporting information) is unsurprising considering that starlings live in kin-structured social groups; high r values are most likely a mixture of mother or father offspring pairs, or sibling relationships.

Discussion

SNPs and microsatellites have recently been compared in free-living populations for studies of population structure, genetic diversity or parentage in whales (Morin *et al.* 2012), fish (Hauser *et al.* 2011; DeFaveri *et al.* 2013; Trọng *et al.* 2013) and crustaceans (Sellars *et al.* 2012). However, our study is the first to compare SNPs and microsatellites in a cooperatively breeding species living in social groups with high kin structure. We show that 102 SNPs (mean $H_O = 0.38$, 2 alleles per locus) can be as, if not more, useful than 15 microsatellites (mean $H_O = 0.54$, 4–16 alleles per locus, mean of 6.2) for parentage analysis in the superb starling. Based on LOD scores alone, SNPs and microsatellites ultimately agreed on parentage in 88% of offspring genotyped. Moreover, when behavioural information was included in the assignments, the two marker systems agreed for 95% of the offspring. SNPs did, however, perform slightly better than microsatellites in the few cases where candidate fathers were highly related. Finally, we found that despite high variation in pairwise r values, the high correlation of the SNP and microsatellite relatedness matrices indicates comparable performance of the two markers for kinship studies.

The number of SNP loci required for parentage analysis is related to the average heterozygosity of the set of SNP loci under consideration, where the more heterozygous the markers, the fewer loci are required (Morin *et al.* 2004). Previous studies have shown that 80–100 SNP loci can be effective for parentage analysis (Hauser *et al.* 2011), whereas simulations in an inbred population of European bison predict that 50–60 SNPs with a heterozygosity of 0.5 should be adequate for paternity assignment (Tokarska *et al.* 2009). Moreover, a comparison of microsatellites and SNPs for parentage in an inbred herd of Angus cattle found that 29–34 SNPs performed as well as 17–18 microsatellite loci (Fernández *et al.* 2013). We confirmed that <80 SNP loci can be useful for parentage if the mean observed heterozygosity of the set is above 0.4. Comparison of assignments made with increasing numbers of the most and least heterozygous SNPs gave a similar result; groups of >50 SNPs performed comparably to 102 SNPs, but only when including behavioural information in the assignments. Overall, this suggests that ~60 SNP loci, with a mean heterozygosity in the 0.4–0.5 range, are adequate for parentage analysis of a free-living population with high kin structure when behavioural information about the parents is included in the assignments. Without behavioural data, groups of SNPs <102 never performed as well as having a complete panel. Thus, for kin-structured populations like those of many cooperative breeders, having observations of the social parents at a nest is critical for accurate assignment when fewer numbers of SNP loci are used.

Given that we found similar results using microsatellites and SNPs for both parentage and kinship analyses, the decision of which marker to use may ultimately come down to a question of cost and development time. SNPs are generally cheaper to use than microsatellites, and the price is decreasing rapidly (Helyar *et al.* 2011). Moreover, development time for the SNP loci was considerably shorter than for microsatellites for this species. One primary difference is that microsatellite marker development requires significant laboratory work for not only the isolation and characterization of loci, but also their optimization and variability screening. In contrast, the development of SNP loci requires significant bioinformatics processing. We found that the simultaneous use of two discovery methods—the GATK HaplotypeCaller pipeline (here confirming its value beyond genome and exome analysis to that of RNAseq and de novo transcriptome) and the SAMTOOLS mpileup—and the application of variant filters for mapping quality, read coverage, depth, potential errors and false positives resulted in the identification of more than 40 000 high-confidence loci that can be further refined for use in studies like these. Filtration of the loci for minor allele frequency and sufficient flanking sequence reduced this set to 9376 loci. We arbitrarily chose 300 of these nearly 10 000 loci and were able to design primers for >97% of them. More than 99% of these primers worked successfully for genotyping, suggesting that using such a conservative approach to SNP discovery can result in a robust marker set. We ultimately chose 103 rather than 300 or even all 9376 loci because of cost issues and suggestions from previously published studies (Tokarska *et al.* 2009; Hauser *et al.* 2011). However, as the iPLEX Gold assay on a Sequenom MassArray is able to analyse more than 36 loci per plex, the cost of studies like this one will decrease greatly. Although other methods of SNP identification and/or analysis (e.g. ddRADseq, Peterson *et al.* 2012) may be cheaper and utilize more loci, they also have drawbacks. For example, even with careful size selection methods, such techniques may identify different loci in each experiment. Adding more individual genotypes to a data set overtime may therefore be challenging. With our approach, new individuals can be genotyped continually using the same loci, much the way they can using microsatellite markers. Thus, our approach will be useful for long-term studies where

new individuals to the population need to be continually genotyped.

We have demonstrated an affordable and powerful way to develop SNP loci from the transcriptome of a nonmodel organism. We further showed that these SNP markers are as, if not more, useful that microsatellite markers for studies of parentage and kinship in a cooperatively breeding bird that lives in social groups with complex kin structure. However, parentage assignments are greatly enhanced when behavioural observations of social parents are available. Thus, for studies of group-living species where field data exist, surprisingly small numbers of SNP loci can be useful for parentage assignment. We encourage further comparisons of different molecular markers in species with a range of social structures.

Acknowledgements

We thank K. Gaynor, L. Pollack and the staff of the Sackler Institute for Comparative Genomics at the American Museum of Natural History for help with microsatellite genotyping. S. Kisselev at the Taub Genomics Core Facility at the Columbia University Medical Center helped with SNP genotyping. S. Guindre-Parker and K. Apakupakul helped run parentage analyses and compile behavioural data. R. Maia helped with statistical analyses. U. Basu Roy helped with RNA extractions. K. Apakupakul, M. MacManes, L. Stenzler and N. Boehlman provided comments on previous versions of this manuscript. This research was supported by the US National Science Foundation (IOS-1121435 and IOS-1257530).

References

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

Avise JC (1994) *Molecular Markers, Natural History and Evolution*. Chapman Hall, Sunderland, MA.

Boomsma JJ, Ratnieks FLW (1996) Paternity in eusocial Hymenoptera. *Philosophical Transactions of the Royal Society of London Series B*, **351**, 947–975.

Casci T (2010) Population genetics: SNPs that come in threes. *Nature Reviews Genetics*, **11**, 8.

Cohas A, Allainé D (2009) Social structure influences extra-pair paternity in socially monogamous mammals. *Biology Letters*, **5**, 313–316.

Cramer ER, Hall ML, de Kort SR, Lovette IJ, Vehrencamp SL (2011) Infrequent extra-pair paternity in the banded wren, a synchronously breeding tropical passerine. *The Condor*, **113**, 637–645.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFTOOLS. *Bioinformatics*, **27**, 2156–2158.

DeFaveri J, Viitaniemi H, Leder E, Merilä J (2013) Characterizing genic and nongenic molecular markers: comparison of microsatellites and SNPs. *Molecular Ecology Resources*, **13**, 377–392.

DePristo M, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Fass J (2010) "SNP Discovery in Novel Genomes". Available from: http://bsc2010.bioinformatics.ucdavis.edu/wp-content/uploads/2010/10/Fass-Joseph-SNP-Discovery.pdf.

Fernández ME, Goszczynski DE, Lirón JP *et al.* (2013) Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd. *Genetics and Molecular Biology*, **36**, 185–191.

Garant D, Kruuk LEB (2005) How to use molecular marker data to measure evolutionary parameters in wild populations. *Molecular Ecology*, **14**, 1843–1859.

Gibbs HL, Weatherhead PJ, Boag PT, White BN, Tabak LM, Hoysak DJ (1990) Realized reproductive success of polygynous red-winged blackbirds revealed by DNA markers. *Science*, **250**, 1394–1397.

Glaubitz JC, Rhodes OE, Dewoody JA (2003) Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology*, **12**, 1039–1047.

Goodnight KF, Queller DC (1999) Computer software for performing likelihood rests of pedigree relationships using genetic markers. *Molecular Ecology*, **8**, 1231–1234.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, **1529**, 644–652.

Guichoux E, Lagache L, Wagner S *et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources*, **11**, 591–611.

Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.

Hamilton MB, Pincus EL, Di Fiore A, Fleischer RC (1999) Universal linker and ligation procedures for construction of genomic DNA libraries enriched for microsatellites. *BioTechniques*, **27**, 500–502.

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.

Helyar SJ, Hemmer-Hansen J, Bekkevold D *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, **11**, 123–136.

Hughes CR (1998) Integrating molecular techniques with field methods in studies of social behavior: a revolution results. *Ecology*, **79**, 383–399.

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

Jones OR, Wang J (2010) COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, **10**, 551–555.

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

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, **16**, 1099–1106.

Kijas JMH, Fowler JCS, Garbett CA, Thomas MR (1994) Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. *BioTechniques*, **16**, 656.

Konovalov DA, Manning C, Henshaw MT (2004) KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Molecular Ecology Notes*, **4**, 779–782.

Li H (2013a) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2.

Li H (2013b) SAMTOOLS website. Available from: http://samtools.sourceforge.net/, accessed 11 June 2013.

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.

MacManes MD (2014) On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, **5**, 13.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

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

Morin PA, Archer FI, Pease VL *et al*. (2012) An empirical comparison of SNPs and microsatellites for population structure, assignment, and demographic analyses of bowhead whale populations. *Endangered Species Research*, **19**, 129–147.

Oksanen J, Blanchet FG, Kindt R *et al*. (2013) vegan: Community Ecology Package. R package version 2.0-7. Available from: http://CRAN.R-project.org/package=vegan.

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.

Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. *Evolution*, **43**, 258–275.

Rubenstein DR (2005) Isolation and characterization of polymorphic microsatellite loci in the plural cooperatively breeding superb starling, *Lamprotornis superbus*. *Molecular Ecology Notes*, **5**, 739–744.

Rubenstein DR (2006) *The evolution of the social and mating systems of the plural cooperatively breeding superb starling, Lamprotornis superbus*. PhD dissertation, Cornell University, Ithaca, NY.

Rubenstein DR (2007a) Territory quality drives intraspecific patterns in extrapair paternity. *Behavioral Ecology*, **18**, 1058–1064.

Rubenstein DR (2007b) Female extrapair mate choice in a cooperative breeder: trading sex for help and increasing offspring heterozygosity. *Proceedings of the Royal Society of London Series B*, **274**, 1895–1903.

Rubenstein DR (2011) Spatiotemporal environmental variation, risk aversion and the evolution of cooperative breeding as a bet-hedging strategy. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 10816–10822.

Sellars MJ, Dierens L, McWilliam S *et al*. (2012) Comparison of microsatellite and SNP DNA markers for pedigree assignment in Black Tiger shrimp, *Penaeus monodon*. *Aquaculture Research*, **3**, 417–426.

Seutin G, White BN, Boag PT (1991) Preservation of avian blood and tissue samples for DNA analysis. *Canadian Journal of Zoology*, **69**, 82–90.

Steele CA, Anderson EC, Ackerman MW *et al*. (2013) A validation of parentage-based tagging using hatchery steelhead in the Snake River basin. *Canadian Journal of Fisheries and Aquatic Sciences*, **70**, 1046–1054.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology and Evolution*, **15**, 199–203.

Tokarska M, Marshall T, Kowalczyk R *et al*. (2009) Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: the case of European bison. *Heredity*, **103**, 326–332.

Trọng TQ, van Bers N, Crooijmans R, Dibbits B, Komen H (2013) A comparison of microsatellites and SNPs in parental assignment in the GIFT strain of Nile tilapia (*Oreochromis niloticus*): the power of exclusion. *Aquaculture*, **388–391**, 14–23.

Uller T, Olsson M (2008) Multiple paternity in reptiles: patterns and processes. *Molecular Ecology*, **17**, 2566–2580.

Van der Auwera GA, Carneiro M, Hartl C *et al*. (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, **43**, 11.10.1–11.10.33.

Westneat DF (1990) Genetic parentage analysis in the indigo bunting: a study using DNA fingerprinting. *Behavioral Ecology and Sociobiology*, **27**, 67–76.

Wysoker A, Tibbetts K, Fennell T (2013) Picard-Tools 1.9.2. Available from: http://picard.sourceforge.net/, accessed 28 May 2013.

Zha L, Yun L, Chen P, Luo H, Yan J, Hou Y (2012) Exploring of tri-allelic SNPs using pyrosequencing and the SNaPshot methods for forensic application. *Electrophoresis*, **33**, 841–848.

L.R.W. collected data and helped write the manuscript. J.W.S. assembled the transcriptome, performed SNP discovery and helped write the manuscript. D.R.R. designed the study, performed statistical analyses and helped write the manuscript.

Data accessibility

Microsatellite-containing DNA sequences: GenBank accessions AY906870–AY906917.

SNP-containing DNA sequences: dbSNP/GenBank accessions ss947850405–ss947850507.

Raw (adapter and quality trimmed) Illumina transcriptome sequences: SRA accessions SRR1565487, SRR1566294, SRR1567314, SRR1567346, SRR1565489, SRR1567380, SRR1565490, SRR1567393.

Microsatellite and SNP genotypes: DRYAD entry doi:10.5061/dryad.jc2pj.

Summary of behavioural focal observations of social parents: DRYAD entry doi:10.5061/dryad.jc2pj.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Supplementary materials.

Figure S1 Correlation between pairwise relatedness values generated from microsatellite and SNP markers.

Table S1 Single- and paired-end read counts from two Illumina HiSeq 2000 lanes for each superb starling sample.

Table S2 List of forward, reverse, and extend primer sequences for all 103 SNPs.