Double-digest RAD sequencing outperforms microsatellite loci at assigning paternity and estimating relatedness: a proof of concept in a highly promiscuous bird

Derrick J. Thrasher³,⁴, Bronwyn G. Butcher¹, Leonardo Campagna¹,², Michael S. Webster³,⁴, Irby J. Lovette¹,²

¹Fuller Evolutionary Biology Program, Cornell Laboratory of Ornithology, Ithaca, NY 14850, USA
²Department of Ecology and Evolutionary Biology, Cornell University, E145 Corson Hall, Ithaca, NY 14853, USA
³Macaulay Library, Cornell Lab of Ornithology, 159 Sapsucker Woods Rd, Ithaca, NY 14850, USA
⁴Department of Neurobiology and Behavior, Cornell University, W361 Mudd Hall, 215 Tower Rd, Ithaca, NY 14853, USA

Keywords: double-digest restriction site-associated DNA sequencing (ddRAD-seq), microsatellite, single nucleotide polymorphism (SNP), parentage, relatedness, cooperative breeding

*Corresponding author: Derrick J. Thrasher, Macaulay Library, Cornell Lab of Ornithology, 159 Sapsucker Woods Rd, Ithaca, NY 14850, USA. Fax: (607) 254-2439. Email: djt224@cornell.edu

Running title: SNPs outperform microsatellite loci
Abstract

Information on genetic relationships among individuals is essential to many studies of the behavior and ecology of wild organisms. Parentage and relatedness assays based on large numbers of SNP loci hold substantial advantages over the microsatellite markers traditionally used for these purposes. We present a double-digest restriction site-associated DNA sequencing (ddRAD-seq) analysis pipeline that, as such, simultaneously achieves the SNP discovery and genotyping steps and which is optimized to return a statistically powerful set of SNP markers (typically 150-600 after stringent filtering) from large numbers of individuals (up to 240 per run). We explore the tradeoffs inherent in this approach through a set of experiments in a species with a complex social system, the variegated fairy-wren (Malurus lamberti), and further validate it in a phylogenetically broad set of other bird species. Through direct comparisons with a parallel dataset from a robust panel of highly variable microsatellite markers, we show that this ddRAD-seq approach results in substantially improved power to discriminate among potential relatives and substantially more precise estimates of relatedness coefficients. The pipeline is designed to be universally applicable to all bird species (and with minor modifications to many other taxa), to be cost- and time-efficient, and to be replicable across independent runs such that genotype data from different study periods can be combined and analyzed as field samples are accumulated.
Introduction

Advances in molecular techniques over the past several decades have substantially improved our ability to test questions about animal social behavior by providing reliable information on the genetic relationships among individuals (Westneat et al. 1990; Hughes 1998; Avise et al. 2002; Griffith et al. 2002; Solomon et al. 2004; Myers & Zamudio 2004). Microsatellites have been the molecular ‘tool-of-choice’ for this application since the 1990s, as microsatellite loci are often highly polymorphic, with up to dozens of co-segregating alleles at a single locus (Queller et al. 1993; Li et al. 2002; Selkoe & Toonen 2006; Guichoux et al. 2011). Accordingly, a small number of highly variable microsatellite loci can provide considerable power for discerning genetic relationships among individuals (Queller et al. 1993; Blouin 2003; Webster & Reichart 2005). However, microsatellite assays also have some practical drawbacks. Microsatellite laboratory protocols developed for one species are often not suitable for use in other species, especially in more distantly related taxa (Galbusera 2000; Decroocq et al. 2003; Hedgecock et al. 2004; Primmer et al. 2005). Traditional PCR-based microsatellite assays incur substantial financial and lab-bench time investments. The manual scoring of microsatellite alleles also requires substantial researcher time, and can involve various forms of error arising from alleles that have more than one clearly defined peak, allelic drop-out and null allele issues, and the various sources of human error that are inherent in any complicated workflow (Pemberton et al. 1995; Hedgecock et al. 2004; Hoffman & Amos 2005; Kalinowski et al. 2007).
Many of these limitations are less severe in assays based on single-nucleotide polymorphisms (SNPs), which require fewer steps and have greater automation (Gut 2001; Syvänen 2001; Seeb et al. 2011; Davey et al. 2011). SNPs are appropriate alternatives for studies of parentage and relatedness data because they are abundant in the genome, have low mutation rates (Brumfield et al. 2003; Morin et al. 2004), and can be scored semi-automatically (Garvin et al. 2010; Guichoux et al. 2011). In comparison to microsatellite-based relationship tests, the primary limitation of SNPs is that SNPs are typically biallelic, whereas microsatellite loci are often multiallelic, and hence the statistical power of SNP loci for discriminating parentage and relatedness is far lower on a per-locus basis (Ball et al. 2010). Compared to highly variable microsatellite loci, a substantially higher number of SNP markers is therefore required to achieve appropriate power in parentage and relatedness studies (Glaubitz et al. 2003; Morin et al. 2004; Coates et al. 2009).

Recently, the application of SNPs for use in analyses of parentage, relatedness, and overall population structure has received greater attention (Glaubitz et al. 2003; Anderson & Garza 2006; Coates et al. 2009). Studies in birds (Cramer et al. 2011; Weinman et al. 2014; Kaiser et al. 2016), fish (Hauser et al. 2011), and several domesticated taxa (Tokarska et al. 2009; Fernández et al. 2013) have developed SNP panels with a sufficient number of SNPs to attain a comparable, if not better, level of resolving power as highly polymorphic microsatellite panels. While each of these studies manage to identify powerful SNP panels, the SNP genotyping methods used are often labor intensive, requiring a significant amount of preparatory work at the discovery stage.
prior to genotyping of large numbers of individuals. Many of these methods also rely on reference genomes (Anderson & Garza 2006; Heylar et al. 2011), or other genomic resources (Fernández et al. 2013; Weinman et al. 2014; Kaiser et al. 2016) (e.g. transcriptome, SNP microarray), for SNP identification. Ultimately, this has afforded several beneficial examples of the utility of SNPs for parentage and relatedness analyses, but without an efficient, universal method of SNP discovery and identification.

Restriction site-associated DNA sequencing (RAD-seq) is widely used in molecular genetic studies (Davey & Blaxter 2010; Etter et al. 2012; Puritz et al. 2014), particularly for linkage and quantitative trait locus (QTL) mapping (Baird et al. 2008), genome wide association studies (Davey et al. 2011), and phylogeography (Andrews et al. 2016). RAD-seq uses a restriction enzyme to fragment and sample a fraction of a genome; as it identifies SNPs with no prior knowledge of the genome, it provides a more universal method of SNP discovery (Willing et al. 2011). Double-digest restriction site-associated DNA sequencing (ddRAD-seq) allows for selection of an even smaller fraction of the genome through the combined use of two restriction enzymes, affording the ability to target a smaller total number of SNPs in a greater number of individuals (Peterson et al. 2012; Puritz et al. 2014; Kess et al. 2016). This ability, in concert with the fact that no prior knowledge of the genome is needed, makes ddRAD-seq an attractive method of simultaneous SNP discovery and screening for use in discerning genetic relationships among individuals.

Here, we describe a ddRAD-based approach to the simultaneous discovery and screening of high numbers of SNP loci with high power for testing questions about
parentage and relatedness. These protocols are optimized to generate an appropriately robust set of SNP markers for 240 individuals per run, to be repeatable across runs to allow the combination of SNP datasets generated at different times, and to be universally applicable to birds (and with small modifications, to other organisms) without requiring a species-specific marker discovery step. We validate these methods by conducting a SNP-based parentage and relatedness study in the highly promiscuous, and socially complex, variegated fairy-wren (Malurus lamberti). We compare the results with previously generated paternity assignments and relatedness information, based on microsatellite screens of the same fairy-wren individuals and social groups. To illustrate the broad utility of this method we report summary statistics for equivalent studies of parentage in a variety of other species that collectively span much of the phylogenetic diversity of living birds.

Methods & Materials

Study Population

The variegated fairy-wren, endemic to Australia, is a cooperatively breeding bird that lives in social groups composed of kin and non-kin (Schodde 1982; Rowley & Russell 1997). Male dispersal is limited, and rates of extra-pair fertilizations (EPFs) are high (~68% of all young, assessed with a panel of 12 species-specific microsatellites; DJ Thrasher, unpublished data). We intensively monitored a color-banded population of the nominate subspecies, M. l. lamberti, on Lake Samsonvale (27°16’ S, 152° 41’ E), 30 km northwest of Brisbane, Queensland, Australia, from 2012 – 2016. The population
ranges from about 250-300 adults depending on year-to-year conditions. The study site is bounded on most sides by Lake Samsonvale, and on its westernmost side by a major highway, which increases our confidence in sampling most, if not all, of the adults in the population. We also monitored all nesting attempts to measure, mark, and collect blood samples from nestlings 6 days after hatching. Blood samples were immediately stored in lysis buffer (White & Densmore 1992), and genomic DNA was later extracted using Qiagen DNeasy Blood and Tissue kits. DNA concentration was determined using the Qubit dsDNA BR Assay Kit and the Qubit® Fluorometer (Life Technologies) following the manufacturers protocol.

**Microsatellite development and genotyping**

We previously developed twelve polymorphic microsatellite loci for the variegated fairy-wren (Table S1, Supporting information) following methods described previously in Nali et al. (2014). Briefly, we extracted genomic DNA from blood in lysis buffer from eight adults in our study population, enriched the mix of DNA with repetitive sequences to develop an enriched microsatellite library, and conducted an Illumina MiSeq sequencing run. From this pool of sequences, we optimized twelve loci that amplified well using polymerase chain reaction (PCR), were polymorphic, and exhibited clearly defined peaks for genotyping. We designed three multiplexed PCRs for genotyping, and each amplification reaction contained 1ul of genomic DNA of varying concentrations (1 ng-40 ng per 1 ul). PCR products were combined with the GeneScan 500 base pair LIZ internal size standard for size-sorting using a 3730 DNA Analyzer. We used Geneious
version 8.0 (Kearse et al. 2012) to score alleles. The program automatically identifies alleles at each locus, and we manually inspected allele calls to minimize genotyping error. In total, we genotyped 287 adults and 482 nestlings from 226 nests sampled during the 2012-2016 breeding seasons.

**ddRAD sequencing**

We selected a subset of the individuals genotyped for microsatellite loci (120 adults and 40 nestlings) for use in our ddRAD-seq experiment and subsequent analyses. To assess the reliability of our SNP panel for parentage and relatedness analysis, we chose representative nestlings from all the years of our study. Typically, we selected one nestling from any individual nest. In a few cases, we selected two nestlings that prior microsatellite analysis had assigned to the same mother but different fathers. For each nestling, we included the mother, the social father, and the genetic father as assigned by previous microsatellite analysis. Our pool of candidate parents included 24 mothers and 78 putative fathers, and 18 randomly selected individuals of both sexes, for a total of 120 adults.

Our ddRAD-seq protocol is adapted from Peterson et al. (2012) (see Supporting information for a detailed protocol). Briefly, for each individual, 100ng - 500ng of DNA (20ul of DNA between concentrations of 5ng/ul - 25ng/ul) were digested with either SbfI and MspI, or SbfI and EcoRI (NEB), and ligated with one of 20 P1 adapters (each containing a unique inline barcode) and a P2 adapter (P2-MspI or P2-EcoRI). After digestion and ligation these samples were pooled in groups of 20 (each with a unique P1
adapter) and purified using 1.5X volumes of homemade MagNA made with Sera-Mag Magnetic Speed-beads (FisherSci) as described by Rohland & Reich (2012). Fragments of between 450 bp and 600 bp were selected using BluePippin (Sage Science) by the Cornell University Biotechnology Resource Center (BRC). Following size selection, index groups and Illumina sequencing adapters were added by performing 11 PCR cycles with Phusion DNA Polymerase (NEB). These reactions were cleaned up with 0.7x volumes of MagNA and pooled in equimolar ratios to create a single library for sequencing on one lane of Illumina HiSeq 2500 (100bp single end, performed by BRC).

By replicating 20 samples (run as a separate index group) with a wider BluePippin size selection range of 400-700 bp, we explored the inherent trade-off between the number of samples sequenced on a lane at a given coverage threshold and the number of SNPs recovered per sample. We similarly explored the effects of using a less frequent restriction enzyme digest (by substituting the 6 bp cutter EcoRI in place of the 4 bp cutter MspI) to generate a smaller total number of fragments in our size range, which in turn should increase the sequencing coverage of the loci screened.

To assess repeatability across index groups, we replicated two index groups, each comprised of 20 samples: one index group from the standard protocol, and the index group generated with the rare-cutter EcoRI enzyme (Table 1). We multiplexed the resulting 240 samples in 12 index groups (with 20 individuals each) which were pooled into a final library that was sequenced on one lane of an Illumina HiSeq 2500, producing 220,300,739 100 bp single end reads (see Table 1).
SNP data analysis

Quality filtering and demultiplexing

After the quality of the reads was assessed using FASTQC version 0.11.5 (www.bioinformatics.babraham.ac.uk/projects/fastqc), we trimmed all sequences to 97bp using fastX_trimmer (FASTX-Toolkit) to exclude low quality calls near the 3’ of the reads. We subsequently removed reads containing a single base with a Phred quality score of less than 10 (using fastq_quality_filter). We additionally removed sequences if more than 5% of the bases had a Phred quality score of less than 20. Using process_radtags module from the Stacks version 1.37 pipeline (Catchen et al. 2013), we demultiplexed the reads to obtain files with sequences that were specific to each individual.

De novo assembly of RAD loci

Because we do not have a sequenced genome for the variegated fairy wren or a close relative – which is likely to be the case for many non-model organisms involved in parentage studies – we assembled the sequences de novo using the Stacks pipeline (Catchen et al. 2013). First, we used denovo_map_pl to assemble the reads into a catalog allowing a minimum stack depth of 5 (m parameter), up to 5 mismatches per locus within an individual (M parameter), and 5 mismatches between loci of different individuals when building the catalog (n parameter). We ran the rxstacks module to filter loci with a log likelihood of less than -50 (lnl_lim -50) or that were confounded in at least
25% of the population (conf_lim 0.25). We then built a new catalog by rerunning cstacks and obtained individual genotype calls with sstacks.

SNP filtering

SNPs were exported using the populations module of the Stacks pipeline. All of our samples were grouped in one population and a locus was exported if it was present in 95% of the individuals in this population (r parameter) at a stack depth of at least 10 (m parameter). The data were restricted to the first SNP per locus (--write_single_snp), and a minor allele frequency of at least 0.25 was required to process a nucleotide site (--min_maf).

Using vcftools version 0.1.14 (Danecek et al. 2011), we removed loci that were not in Hardy-Weinberg equilibrium (--hwe). We obtained a variant call format file that was converted to structure format in PGD Spider version 2.0.5.0 (Lischer & Excoffier 2012) and further modified to a format compatible with CERVUS version 3.0.7 (Kalinowski et al. 2007) using a custom perl script.

Parentage Analysis

We used CERVUS version 3.0.7 (Kalinowski et al. 2007) to assign paternity for all nestlings using our microsatellite and SNP datasets separately. CERVUS uses a two-step, likelihood-based approach to assign parentage. First, CERVUS compares each offspring’s genotype to that of a candidate parent and a random individual in the population to calculate a likelihood ratio. This relationship is presented as an LOD score,
which is simply the natural logarithm of the calculated likelihood ratio. Positive LOD scores indicate that a candidate parent is much more likely to be the true parent, whereas negative LOD scores indicate that the candidate parent is highly unlikely to be a true parent. Second, CERVUS conducts a simulation of parentage analysis based on population allele frequencies and the proportion of potential parents included in the analysis. The simulation accounts for the possibility of unsampled parents, missing data, and genotyping errors. Considering these parameters, the simulation calculates critical LOD scores by comparing the LOD distributions of the most likely parent and all other candidate parents. The critical LOD score is used to determine the confidence (95% or 80%) of each parentage assignment.

CERVUS allows for different types of parentage analysis, including parent-pair (sexes known or unknown), maternity (known father, but not mother), and paternity (known mother, but not father). Variegated fairy-wrens at Lake Samsonvale are relatively easy to observe, and we were able to assign known mothers behaviorally. We subsequently confirmed this with microsatellite analyses: females that built and attended a nest throughout incubation were always the mothers of the nestlings in that nest. In many systems, a comparable level of demographic knowledge may not be available, so a marker set must be powerful enough to assign parentage with minimal social information. To investigate the broader utility of our ddRAD-seq method, we conducted analyses that relied on the inclusion of the known mother, in addition to analyses independent of the known mother, which were based only on the father-offspring relationship. We simulated paternity assignments for 10,000 offspring to determine
critical LOD scores, using slightly different input parameters for each panel (microsatellites and ddRAD sequencing derived SNPs). Simulations for both used the following parameters: 78 candidate males, 95% of candidate males sampled, estimated error rate of 0.01 for mistyped loci and likelihood scores. The proportion of loci typed across all individuals was different for both panels: 0.997 for the microsatellite simulation, and 0.961 for the SNP simulation.

For both paternity analyses, we used the trio LOD score and the father-offspring LOD score from CERVUS to make assignments. The trio LOD score was calculated by comparing the genotypes of the candidate male and offspring, relative to that of the known mother. The father-offspring LOD score only accounts for the relationship of the candidate male and the offspring, independent of the known mother. CERVUS ranked candidate males by LOD scores in each category, and the highest-ranking males were assigned as fathers. These rankings should be in agreement, but ambiguous assignments (different top-ranking males assigned in each category) may occur when multiple candidate male genotypes closely match an offspring’s genotype.

We assessed each CERVUS assignment to determine whether it was plausible, and whether the assigned male was the social father or an extra-pair sire. Our criteria for accepting assignments differed slightly for microsatellites and SNPs. For microsatellites, we automatically accepted the CERVUS assignment if the highest-ranking male was in agreement for both the trio LOD and the father-offspring LOD, and if the number of mismatches between the assigned male and the offspring was ≤ 1 (8% of 12 loci). For SNPs, we also accepted the assignment if the highest-ranking males by LOD score type
were in agreement, and an allowable number of mismatches were not exceeded.

However, for SNPs, our allowable number of mismatches was based on the observed maximum number of mismatches between a known mother and her known offspring (max. = 7, mean = 3.4, 2% of 411 loci). For both panels, we accepted the social father as the genetic sire if he met these respective criteria. If the social father mismatched the offspring at higher numbers, or had negative LOD scores, the offspring was considered sired by an extra-pair father. We accepted assignments of extra-pair fathers using the same criteria outlined above. We did not observe cases in which an offspring could not be assigned to either its social father, or an extra-pair sire.

**Relatedness Analysis**

We used the package, “related” (Pew et al. 2015), in R version 3.2.5 (R Core Team 2016) to estimate pairwise relatedness (r) between all pairs of individuals in this study. This package accounts for genotyping errors, missing data, and can estimate relatedness using any of seven different estimators (4 non-likelihood-based, and 3 likelihood-based). “related” includes the function, `compareestimators`, which tests the performance of different estimators on simulated data that share the same characteristics as the real data. The program uses an allele frequency file to generate simulated pairs of individuals of known relatedness, and automatically estimates relatedness using four of the most commonly used estimators (all non-likelihood-based). The function calculates a correlation coefficient between observed and expected values, to evaluate which estimator performs best with the data set. Using `compareestimators`
to generate 200 simulated pairs of individuals for each degree of relatedness (i.e., half-
sib, full-sib, parent-offspring, unrelated), we determined that the Wang (2002) estimator
performed best for both our microsatellite and SNP datasets. We obtained point
estimates of relatedness using the Wang (2002) estimator, and evaluated all parent-
offspring relationships that were previously determined in our parentage analysis.

“Related” also evaluates how well different marker sets resolve degrees of
relatedness, given simulated genotypes based on allele frequency files. For both panels,
we used the familiesim function to generate 200 pairs of individuals for each degree of
relatedness. We then used the coancestry function to analyze all pairwise relatedness
values with the Wang (2002) estimator. We created density plots representing
histograms of the relatedness values. These plots show the overlap in relatedness
values for degree of relatedness, and we used them to infer how well each panel
performed at discerning different relationships.

Results

SNP development and analysis

After trimming, filtering and demultiplexing the data, we retained a total of
109,524,874 reads across all index groups, with an average of approximately 9,000,000
reads per index group. Two individuals failed (i.e., had less than 66,000 reads each; one
individual from each of two index groups) and were excluded from further analysis. The
number of reads per sample for the remaining individuals ranged from 213,544 to
810,966 (mean = 459,726 ± 118,771 std. dev.).
Further analysis using the population program from stacks identified loci with at least 10X coverage, present in 95% of the individuals, and a minimum allele frequency greater than 25%. We further retained only those loci that were in Hardy-Weinberg equilibrium. When performing analyses on all 160 individuals from the primary comparison runs (Table 1), we identified 411 loci that fulfilled these criteria and were used for downstream analyses.

We varied two aspects of the ddRAD-seq protocol to assess the number of loci recovered and the reproducibility of the method. As expected, both a reduction in the range of fragment sizes selected during the construction of the library (compare index 1 (150 bp size selection) and index 10 (300 bp size selection)) or the use of the less frequent cutter, EcoRI (index 11 and 12), resulted in fewer loci recovered (Table 3). Those from the EcoRI digest were non-overlapping with those from the MspI index groups. More importantly, between the replicated index groups in our standard protocol (index 1 and index 9) we recovered similar numbers of loci (797 and 645, respectively) with 549 (85.1%) loci found in both datasets (under our stringent filtering criteria).

Paternity assignments

Both panels produced highly concordant results when assigning paternity, but the SNP panel showed substantially higher power overall. Generally, the microsatellite loci were more polymorphic, resulting in greater mean polymorphic information content (PIC) for any given locus. Despite this, the SNP panel performed better because of the large number of loci obtained through RAD sequencing. This greatly improved the non-
exclusion probabilities across different parentage assignment contexts (Table 2), and reduced uncertainty in our assignments. Given the known mother, the microsatellite and SNP panels assigned the same fathers to all 40 offspring with 95% confidence. When paternity assignments were made without the known mother (no known candidate parents), both panels again assigned fathers for all 40 offspring with 95% confidence. However, 5 of these assignments were not in agreement between the two panels. For these 5 cases, two candidate males had very similar LOD scores under the microsatellite panel, and the assigned males did not match the males that were assigned given the known mother. These (and all other) cases were resolved unambiguously when the SNP panel was used, and the paternity assignments with and without the known mother were in complete agreement for all offspring (Fig. 1).

Overall, both panels assigned 23 out of 40 nestlings (57.5%) to males that were not their social father. Due to the nature of our non-random sampling of individuals for this experiment, and the overall smaller sample size, this value is slightly lower than the overall rate of 67.6% extra-pair young observed for all years of the study (unpublished data).

As a measure of certainty for our assignments, we calculated the difference between LOD scores for the two top-ranked males assigned to each nestling, under each panel (Fig. 2). Typically, this difference was $8 - 10x$ higher for the SNP panel ($n=40$, mean $= 165.0$) than for the microsatellite panel ($n=40$, mean 19.1), a reflection of the much higher discriminatory power of the SNP dataset. For the SNP panel, many of the second-ranked males had a strongly negative LOD score, making them extremely
unlikely to be the true father. This was less often true for the microsatellite panel, as the second-ranked males often had positive, or just slightly negative, LOD scores. Overall this result illustrates the increased discrimination power achieved by the SNP panel compared to the microsatellites, which allowed us to assign paternity in cases in which the microsatellite assignments remained ambiguous or (albeit rarely) misleading.

*Relatedness analysis*

The SNP panel produced simulated data that closely matched the observed allele and genotype frequencies (Pearson’s correlation coefficient = 0.975). The microsatellite panel also matched well, but was not as reliable as the SNP panel (Pearson’s correlation coefficient = 0.877). This resulted in better estimates of pairwise relatedness for parent-offspring using the SNP panel (Fig. 3). Overall, the SNP panel produced better simulated estimates for each degree of relatedness (Fig. 4), greatly reducing the variance around expected relatedness values (unrelated = 0, half-sib = 0.25, full-sib = 0.5, and parent-offspring = 0.5). This bolsters the confidence with which actual relationships can be discerned when calculating pairwise relatedness of a population for which there is little prior knowledge of social relationships.

**Discussion**

Several recent studies have rigorously investigated the use of SNPs in population genetic studies for several non-model organisms (Morin *et al.* 2004; Slate *et al.* 2010; Garvin *et al.* 2010; Heylar *et al.* 2011; Seeb *et al.* 2011), with growing support for the use
of SNPs in studies of parentage (Anderson & Garza 2006; e.g. Hauser et al. 2011; e.g. Kaiser et al. 2016; e.g. Kess et al. 2016) and relatedness (e.g. Glaubitz et al. 2003; Wang 2007). SNPs have proven to perform as well, if not better, than microsatellites in these types of studies. To our knowledge this is the first study to describe a universal ddRAD-seq method for use in parentage and relatedness analyses of wild populations of birds. Our study is also the first to compare the efficiency of microsatellites versus SNPs for determining genetic relationships in a species that is both socially complex, and highly promiscuous. We show that SNPs developed from our modified ddRAD-seq method are substantially more powerful than a moderate number of species-specific microsatellite loci at assigning paternity and estimating relatedness among individuals. Our method is highly attractive as an alternative to traditional microsatellite genotyping, especially for systems where no microsatellites have been developed. This is largely due to the combination of its cost and researcher time efficiency, the ease of this non-species-specific method that combines the SNP discovery and screening steps, and the large number of SNPs reliably recovered.

The total approximate materials cost for our ddRAD-seq analyses, including DNA extraction, normalization of the DNA concentrations, library preparation, sequencing and computational time was US$3,270.00 for 240 samples, or approximately $13.63 per sample. The use of a homemade MagNA in place of commercial SPRI beads provides significant savings. This cost is similar to that for genotyping 240 individuals at 12 microsatellite loci (in 3 multiplexed PCR mixes), in a situation where the labelled primers have already been designed, purchased, and tested. However, a substantial additional
benefit of this ddRAD-seq method is that it does not require any locus discovery or development before starting. The time required for library preparation, once DNA has been extracted, is modest, and once the sequence data have been obtained, SNP calling for the entire dataset can be performed in less than a day through a largely automated bioinformatics pipeline. Unlike manually scoring peaks as in a traditional microsatellite genotyping analysis, the identification of SNPs is less subjective and takes far fewer hours of hands-on analysis (as most is performed computationally). The tools for analyzing these ddRAD data are freely available and widely used (e.g., Stacks, VCFtools).

For this study, our conditions and protocol allowed us to recover 411 high quality SNP loci for 240 individual samples. However, we show that through simple variations in the size selection window or the specificity of the restriction enzyme, more or fewer loci can be obtained. For some applications, it could be advantageous to multiplex a greater number of individuals and achieve similar coverage by aiming to recover fewer loci (e.g., using EcoRI rather than MspI). Alternatively, for applications where more loci are required, the size selection window could be widened and concordantly the number of individuals would have to be lowered.

The number of SNPs needed to perform robust parentage and relatedness analyses depends on characteristics of the study population. Populations with reduced genetic diversity will likely require a greater number of loci than those that are more genetically diverse (Saunders et al. 2007; Strucken et al. 2016; Tortereau et al. 2017). Obtaining more loci from the outset would aid in overcoming any issues relating to
population genetic diversity. Additionally, when studying species with complex social systems, including for example both variable levels of genetic relatedness among individuals and high rates of extra-pair fertilizations, it is imperative to obtain a sufficient number of markers to discern genetic relationships robustly (Hughes 1998; Ross 2001; Weinman et al. 2014). Our case study, using the variegated fairy-wren, shows that our modified ddRAD-seq method recovers more than enough SNP loci to confidently discern relationships in a species with a complex social system. Most paternity and relatedness analysis programs are well equipped to handle large numbers of loci, so a greater number of loci would not hinder analyses. Once an appropriate number of SNPs are identified for performing robust analyses, conditions can be varied to maximize the number of individuals to be genotyped.

For both paternity and relatedness analyses, our SNP panel far outperformed our microsatellite panel by providing much more power and improving the overall confidence for assignments. Variegated fairy-wrens are relatively easy to observe, and every nest found can be assigned to a known mother by watching the female that builds the nest and/or incubates the eggs. This level of knowledge may not be the norm for most study systems, so we also investigated the CERVUS output for male-offspring relationships, independent of known mothers. In doing so, the reliability of the SNP panel became even more evident. In CERVUS, the higher the LOD score, the more likely that a given male is the true father. Using SNPs, CERVUS typically output only a single male with a positive LOD score, and the difference in LOD scores between the top-two ranked males was dramatically different for SNP assignments (Fig. 2). When social information about
the known mother was excluded from the paternity analysis, the microsatellite panel
sometimes produced assignments that were ambiguous (two males had similar LOD
scores), and occasionally the wrong male was assigned paternity of the offspring. Under
the SNP panel, ambiguous assignments were nonexistent, and these cases were clearly
resolved (Fig. 1).

It is sometimes difficult to obtain appropriate demographic data to use in a formal
parentage analysis, and for many studies, this level of detail may not be necessary.
Population allele frequencies can be used to estimate pairwise relatedness for
individuals, and to reconstruct pedigrees using maximum likelihood-based methods.
Variance in estimates of pairwise relatedness (r) for known parent-offspring pairs was
dramatically reduced when using SNPs (Fig. 3). For our simulations, SNPs greatly
improved the differentiation between distributions for individuals of known degrees of
relatedness (Fig. 4). This is particularly important for systems with minimal demographic
and observational data, where these distributions can be used to determine familial
relationships between individuals, in conjunction with actual estimated r-values.

In summary, our ddRAD-seq method provides a cost effective and robust way to
identify SNPs for use in studies utilizing parentage and relatedness analyses. Our
experiment shows that a majority of the same SNPs can be obtained across groups,
using the same size selection windows and restriction enzymes. Future individuals can
be genotyped and incorporated to the analysis by re-running the Stacks pipeline. Using
a bird exhibiting great social complexity, and high promiscuity, we have shown that
SNPs identified by ddRAD-seq are more effective at assigning paternity and estimating
relatedness than highly polymorphic, species-specific microsatellite loci. This protocol was designed to be universally applicable across bird species, and we have successfully applied it in a range of other avian study systems (Table 4). While different numbers of individuals were used in each study and therefore different numbers of loci were recovered, in all cases paternity was confidently assigned to nestlings using CERVUS (unpublished results). Applying this general protocol to many non-avian taxa may simply require ensuring that the specific restriction enzymes and fragment size windows are chosen appropriately.

Acknowledgements
We thank D. Baldassarre, K. Gielow, J. Welklin, and field technicians at Lake Samsonvale who assisted with field efforts for this study. We are grateful to L. Stenzler and S. Bogdanowicz for help with microsatellite discovery, development, and genotyping. D. Baldassarre, E. Greig, & A. Dalziel provided valuable discussion on the study design. This research was supported by the US National Science Foundation (IOS-1353681 and DEB-1721662).
References


Fernández ME, Goszczynski DE, Lirón JP et al. (2013) Comparison of the effectiveness...


Molluscan Studies, 82, 104–109.


Australia.


Data Accessibility
The raw data used in this manuscript will be stored in the Dryad Digital Repository upon acceptance.

Author Contributions
D.J.T designed the study, collected field data, performed microsatellite development and analysis, conducted parentage and relatedness analyses, and drafted the manuscript with help from all co-authors. B.G.B and L.C. designed the study, and performed SNP discovery and analysis. M.S.W and I.J.L. helped design the study, and secured funding.
## Tables

### Table 1: Experimental design.

<table>
<thead>
<tr>
<th># of samples (index groups)</th>
<th>enzymes</th>
<th>size selection interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>160 samples (8 index groups, index 1-8)</td>
<td>SbfI - MspI</td>
<td>450 - 600 bp</td>
</tr>
<tr>
<td>20 samples (Index 9)</td>
<td>SbfI - MspI</td>
<td>450 - 600 bp (replicate of above)</td>
</tr>
<tr>
<td>20 samples (index 10)</td>
<td>SbfI - MspI</td>
<td>400 - 700 bp (wide size selection)</td>
</tr>
<tr>
<td>20 samples (index 11)</td>
<td>SbfI - EcoRI</td>
<td>450 - 600 bp (infrequent 3’ cutter)</td>
</tr>
<tr>
<td>20 samples (index 12)</td>
<td>SbfI - EcoRI</td>
<td>450 - 600 bp (replicate of above)</td>
</tr>
</tbody>
</table>

### Table 2. Marker characteristics

<table>
<thead>
<tr>
<th>Marker Panel</th>
<th>Number of loci</th>
<th>Mean proportion loci typed</th>
<th>Mean alleles per locus</th>
<th>Mean $H_e$</th>
<th>Mean $H_o$</th>
<th>Mean PIC</th>
<th>Nonexclusion probability (first parent)</th>
<th>Nonexclusion probability (second parent)</th>
<th>Nonexclusion probability (parent pair)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsatellites</td>
<td>12</td>
<td>0.99</td>
<td>14.17</td>
<td>0.77</td>
<td>0.76</td>
<td>0.74</td>
<td>$1.9 \times 10^{-4}$</td>
<td>$1.9 \times 10^{-8}$</td>
<td>$1.5 \times 10^{-10}$</td>
</tr>
<tr>
<td>SNPs</td>
<td>411</td>
<td>0.98</td>
<td>2.00*</td>
<td>0.45</td>
<td>0.45</td>
<td>0.35</td>
<td>$5.2 \times 10^{-20}$</td>
<td>$6.9 \times 10^{-35}$</td>
<td>$1.0 \times 10^{-55}$</td>
</tr>
</tbody>
</table>

*Only biallelic SNPs were retained. If a locus had 3 alleles across the population, it was filtered from the dataset.*
Table 3. Overlap in the RAD loci that were obtained while varying different steps of the protocol (size selection and restriction enzymes). The diagonal indicates the total number of loci recovered for each treatment. Values above the diagonal represent the percent overlapping loci between groups (relative to the group with the smallest number of loci), while values below the diagonal list the number of loci that were overlapping between groups.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>411</td>
<td>85.4</td>
<td>78.1</td>
<td>79.1</td>
<td>1.2</td>
<td>1.0</td>
</tr>
<tr>
<td>A</td>
<td>351</td>
<td>797</td>
<td>85.1</td>
<td>75.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>B</td>
<td>321</td>
<td>549</td>
<td>645</td>
<td>83.4</td>
<td>2.0</td>
<td>2.3</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>604</td>
<td>538</td>
<td>1440</td>
<td>4.2</td>
<td>3.9</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>15</td>
<td>12</td>
<td>25</td>
<td>596</td>
<td>68.4</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>13</td>
<td>12</td>
<td>20</td>
<td>353</td>
<td>516</td>
</tr>
</tbody>
</table>

All: 160 samples; Sbf1/Msp1; 450-600 bp.
A: 20 samples; Sbf1/Msp1; 450-600 bp.
B: 20 samples; Sbf1/Msp1; 450-600 bp.
C: 20 samples; Sbf1/Msp1; 400-700 bp.
D: 20 samples; SbfI/EcoRI; 450-600 bp.
E: 20 samples; SbfI/EcoRI; 450-600 bp.

Table 4: Summary information for ddRAD-seq studies performed to investigate parentage in other bird species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Scientific name</th>
<th>Number of individuals</th>
<th>Number of loci</th>
<th>Number of loci in HWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variegated fairy-wren</td>
<td>Malurus lamberti</td>
<td>160</td>
<td>552</td>
<td>411</td>
</tr>
<tr>
<td>Hispaniolan woodpecker*</td>
<td>Melanerpes striatus</td>
<td>288</td>
<td>179</td>
<td>135</td>
</tr>
<tr>
<td>Northern Red-billed hornbill</td>
<td>Tockus erythrorhynchus</td>
<td>40</td>
<td>475</td>
<td>414</td>
</tr>
<tr>
<td>Von der Decken's hornbill</td>
<td>Tockus deckeni</td>
<td>112</td>
<td>490</td>
<td>410</td>
</tr>
<tr>
<td>Sapayoa</td>
<td>Sapayoa aenigma</td>
<td>6</td>
<td>672</td>
<td>671</td>
</tr>
<tr>
<td>Red-backed fairy-wren**</td>
<td>Malurus melanocephalus</td>
<td>240</td>
<td>483</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td></td>
<td>240</td>
<td>291</td>
<td>174</td>
</tr>
</tbody>
</table>

* Two independent ddRAD-seq experiments were performed - one with 240 samples and the other with 48. After quality filtering and demultiplexing the data from 288 samples was combined for denovo assembly and SNP identification.

** Two independent ddRAD-seq experiments were run on each set of 240 samples. There is 74% overlap in the loci identified in each experiment (of the loci in HWE).
Figures

Figure 1. Resolved paternity assignments for 5 nestlings with ambiguous assignments under the microsatellite panel, but not with the SNP panel. Each panel in the graph represents an individual offspring, and the two top-ranked males are depicted as a triangle and a circle, respectively. Lines connecting like shapes show the change in LOD score for each male, using each marker type (microsatellites versus SNPs). Note that y-axis scale varies among panels in the graph.
Figure 2. Difference in CERVUS LOD scores (delta LOD) between the most likely father of a nestling and the second possible father in the population, for both marker panels.
Figure 3. Box plot of pairwise relatedness values for all parent-offspring (40 mother-offspring and 40 father-offspring) relationships, using population allele frequencies from each marker panel.
Figure 4. Density plots of relatedness values for simulated pairs of known relatedness (unrelated, half-sibling, full-sibling, and parent-offspring) using population allele frequencies from each marker panel (a. MSAT; b. SNP). Overlap in distributions indicates the overlap between relatedness value estimators for pairs of individuals of different relationships. The spread of each distribution indicates the reliability of observed relatedness values based on their deviation from expected relatedness values (Unrelated = 0, Half-sib = 0.25, Full-sib = 0.5, and Parent-offspring (P-O) = 0.5, denoted by vertical dashed lines).