

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

4
5 Derrick J. Thrasher^{3,4}, Bronwyn G. Butcher¹, Leonardo Campagna^{1,2}, Michael S.
6 Webster^{3,4}, Irby J. Lovette^{1,2}
7

8
9 ¹Fuller Evolutionary Biology Program, Cornell Laboratory of Ornithology, Ithaca, NY
10 14850, USA

11 ²Department of Ecology and Evolutionary Biology, Cornell University, E145 Corson Hall,
12 Ithaca, NY 14853, USA

13 ³Macaulay Library, Cornell Lab of Ornithology, 159 Sapsucker Woods Rd, Ithaca, NY
14 14850, USA

15 ⁴Department of Neurobiology and Behavior, Cornell University, W361 Mudd Hall, 215
16 Tower Rd, Ithaca, NY 14853, USA
17

18 **Keywords:** double-digest restriction site-associated DNA sequencing (ddRAD-seq),
19 microsatellite, single nucleotide polymorphism (SNP), parentage, relatedness,
20 cooperative breeding
21

22 *Corresponding author: Derrick J. Thrasher, Macaulay Library, Cornell Lab of
23 Ornithology, 159 Sapsucker Woods Rd, Ithaca, NY 14850, USA. Fax: (607) 254-2439.
24 Email: djt224@cornell.edu
25

26
27 Running title: ddRAD-seq SNPs outperform microsatellite loci
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43

44 **Abstract**

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

63

64

65

66 **Introduction**

67 Advances in molecular techniques over the past several decades have
68 substantially improved our ability to test questions about animal social behavior by
69 providing reliable information on the genetic relationships among individuals (Westneat
70 *et al.* 1990; Hughes 1998; Avise *et al.* 2002; Griffith *et al.* 2002; Solomon *et al.* 2004;
71 Myers & Zamudio 2004). Microsatellites have been the molecular ‘tool-of-choice’ for this
72 application since the 1990s, as microsatellite loci are often highly polymorphic, with up to
73 dozens of co-segregating alleles at a single locus (Queller *et al.* 1993; Li *et al.* 2002;
74 Selkoe & Toonen 2006; Guichoux *et al.* 2011). Accordingly, a small number of highly
75 variable microsatellite loci can provide considerable power for discerning genetic
76 relationships among individuals (Queller *et al.* 1993; Blouin 2003; Webster & Reichart
77 2005). However, microsatellite assays also have some practical drawbacks.
78 Microsatellite laboratory protocols developed for one species are often not suitable for
79 use in other species because the primers may not amplify well and targeted loci are
80 often not as polymorphic, especially in more distantly related taxa (Galbusera 2000;
81 Decroocq *et al.* 2003; Hedgecock *et al.* 2004; Primmer *et al.* 2005). Next-generation
82 sequencing has made the discovery of microsatellite loci for individual species more
83 attainable (Davey *et al.* 2011). However, discovering microsatellite loci can be very time
84 consuming and costly, largely due to protracted testing and optimization of candidate
85 primers after the initial sequencing. Additionally, traditional PCR-based microsatellite
86 assays also incur substantial financial and lab-bench time investments. The manual
87 scoring of microsatellite alleles also requires substantial researcher time, and can
88 involve various forms of error arising from alleles that have more than one clearly
89 defined peak, allelic drop-out and null allele issues, and the various sources of human

90 error that are inherent in any complicated workflow (Pemberton *et al.* 1995; Hedgecock
91 *et al.* 2004; Hoffman & Amos 2005; Kalinowski *et al.* 2007).

92 Many of these limitations are less severe in assays based on single-nucleotide
93 polymorphisms (SNPs), which require fewer steps and have greater automation (Gut
94 2001; Syvänen 2001; Seeb *et al.* 2011; Davey *et al.* 2011). SNPs are appropriate
95 alternatives for studies of parentage and relatedness data because they are abundant in
96 the genome, have low mutation rates (Brumfield *et al.* 2003; Morin *et al.* 2004), and can
97 be scored semi-automatically (Garvin *et al.* 2010; Guichoux *et al.* 2011). In comparison
98 to microsatellite-based relationship tests, the primary limitation of SNPs is that they are
99 typically biallelic, whereas microsatellite loci are often multiallelic, and hence the
100 statistical power of SNP loci for discriminating parentage and relatedness is far lower on
101 a per-locus basis (Ball *et al.* 2010). Compared to highly variable microsatellite loci, a
102 substantially higher number of SNP markers is therefore required to achieve appropriate
103 power in parentage and relatedness studies (Glaubitz *et al.* 2003; Morin *et al.* 2004;
104 Coates *et al.* 2009).

105 Recently, the application of SNPs for use in analyses of parentage, relatedness,
106 and overall population structure has received greater attention (Glaubitz *et al.* 2003;
107 Anderson & Garza 2006; Coates *et al.* 2009). Studies in birds (Cramer *et al.* 2011;
108 Weinman *et al.* 2015; Kaiser *et al.* 2017), fish (Hauser *et al.* 2011), and several
109 domesticated taxa (Tokarska *et al.* 2009; Fernández *et al.* 2013) have developed
110 sufficiently large SNP panels to attain a comparable, if not better, level of resolving
111 power as highly polymorphic microsatellite panels. While each of these studies manage
112 to identify powerful SNP panels, the SNP genotyping methods used are often labor

113 intensive, requiring a significant amount of preparatory work at the discovery stage prior
114 to genotyping of large numbers of individuals. Many of these methods also rely on
115 reference genomes (Anderson & Garza 2006; Heylar *et al.* 2011), or other genomic
116 resources (Fernández *et al.* 2013; Weinman *et al.* 2015; Kaiser *et al.* 2017) (e.g.
117 transcriptome, SNP microarray), for SNP identification. Ultimately, this has afforded
118 several beneficial examples of the utility of SNPs for parentage and relatedness
119 analyses, but without an efficient, universal method of SNP discovery and identification.

120 Restriction site-associated DNA sequencing (RAD-seq) is a reduced-
121 representation genomic technique that is widely used in molecular genetic studies
122 (Davey & Blaxter 2010; Etter *et al.* 2012; Puritz *et al.* 2014), particularly for linkage and
123 quantitative trait locus (QTL) mapping (Baird *et al.* 2008), genome wide association
124 studies (Davey *et al.* 2011), and phylogeography (Andrews *et al.* 2016). RAD-seq uses
125 restriction enzymes to fragment and sample a fraction of a genome; as it identifies SNPs
126 with no prior knowledge of the genome, it provides a more universal method of SNP
127 discovery (Willing *et al.* 2011). Double-digest restriction site-associated DNA sequencing
128 (ddRAD-seq) is a RAD-seq protocol that allows for selection of an even smaller fraction
129 of the genome through a size selection step, affording the ability to target a smaller total
130 number of SNPs in a greater number of individuals (Peterson *et al.* 2012; Puritz *et al.*
131 2014; Kess *et al.* 2016). This ability, in concert with the fact that no prior knowledge of
132 the genome is needed, makes ddRAD-seq an attractive method of simultaneous SNP
133 discovery and screening for use in discerning genetic relationships among individuals.

134 Here, we describe a ddRAD-based approach to the simultaneous discovery and
135 screening of high numbers of SNP loci with high power for testing questions about

136 parentage and relatedness. These protocols are optimized to generate an appropriately
137 robust set of SNP markers for 240 individuals per run, to be repeatable across runs to
138 allow the combination of SNP datasets generated at different times, and to be
139 universally applicable to birds (and with small modifications, to other organisms) without
140 requiring a species-specific marker discovery step. We validate these methods by
141 conducting a SNP-based parentage and relatedness study in the highly promiscuous,
142 and socially complex, variegated fairy-wren (*Malurus lamberti*). We compare the results
143 with previously generated paternity assignments and relatedness information, based on
144 microsatellite screens of the same fairy-wren individuals and social groups. To illustrate
145 the broad utility of this method we report the number of loci recovered for equivalent
146 studies of parentage that included different numbers of individuals (from less than 10 to
147 almost 500) of a variety of other species that collectively span much of the phylogenetic
148 diversity of living birds.

149

150 **Methods & Materials**

151 *Study population*

152 The variegated fairy-wren, endemic to Australia, is a cooperatively breeding bird
153 that lives in social groups composed of kin and non-kin (Schodde 1982; Rowley &
154 Russell 1997). Male dispersal is limited, and rates of extra-pair fertilizations (EPFs) are
155 high (~68% of all young, assessed with a panel of 12 species-specific microsatellites; DJ
156 Thrasher, unpublished data). We intensively monitored a color-banded population of the
157 nominate subspecies, *M. l. lamberti*, on Lake Samsonvale (27°16' S, 152° 41' E), 30 km
158 northwest of Brisbane, Queensland, Australia, from 2012 – 2016. The population

159 ranges from about 250-300 adults depending on year-to-year conditions. The study site
160 is bounded on most sides by Lake Samsonvale, and on its westernmost side by a major
161 highway, which increases our confidence in sampling most, if not all, of the adults in the
162 population. We also monitored all nesting attempts to measure, mark, and collect blood
163 samples from nestlings 6 days after hatching. Blood samples were immediately stored in
164 lysis buffer (White & Densmore 1992), and genomic DNA was later extracted using
165 Qiagen DNeasy Blood and Tissue kits. DNA concentration was determined using the
166 Qubit dsDNA BR Assay Kit and the Qubit® Fluorometer (Life Technologies) following
167 the manufacturers protocol.

168

169 *Microsatellite development and genotyping*

170 We developed 12 polymorphic microsatellite loci for the variegated fairy-wren
171 (Table S1, Supporting Information) following methods described previously in Nali *et al.*
172 (2014). Briefly, we extracted genomic DNA from blood in lysis buffer from eight adults in
173 our study population, enriched the mix of DNA with repetitive sequences to develop an
174 enriched microsatellite library, and conducted an Illumina MiSeq sequencing run. From
175 this pool of sequences, we optimized 12 loci that amplified well using polymerase chain
176 reaction (PCR), were polymorphic, and exhibited clearly defined peaks for genotyping.
177 We designed three multiplexed PCRs for genotyping, and each amplification reaction
178 contained 1ul of genomic DNA of varying concentrations (1 ng/ul -40 ng/ul). PCR
179 products were combined with the GeneScan 500 base pair LIZ internal size standard for
180 size-sorting using a 3730 DNA Analyzer. We used Geneious version 8.0 (Kearse *et al.*
181 2012) to score alleles. The program automatically identifies alleles at each locus, and we

182 manually inspected allele calls to minimize genotyping error. In total, we genotyped 287
183 adults and 482 nestlings from 226 nests sampled during the 2012-2016 breeding
184 seasons.

185

186 *ddRAD sequencing*

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

197 Our ddRAD-seq protocol is adapted from Peterson et al. (2012) (see Supporting
198 Information for a detailed protocol). Briefly, for each individual, 100ng - 500ng of DNA
199 (20ul of DNA between concentrations of 5ng/ul - 25ng/ul) were digested with either SbfI
200 and MspI, or SbfI and EcoRI (NEB), and ligated with one of 20 P1 adapters (each
201 containing a unique inline barcode) and a P2 adapter (P2-MspI or P2-EcoRI). After
202 digestion and ligation these samples were pooled in groups of 20 (each with a unique P1
203 adapter) and purified using 1.5X volumes of homemade MagNA made with Sera-Mag
204 Magnetic Speed-beads (FisherSci) as described by Rohland & Reich (2012). Fragments

205 of between 450 bp and 600 bp were selected using BluePippin (Sage Science) by the
206 Cornell University Biotechnology Resource Center (BRC). Following size selection,
207 index groups and Illumina sequencing adapters were added by performing 11 PCR
208 cycles with Phusion DNA Polymerase (NEB). These reactions were cleaned up with 0.7x
209 volumes of MagNA and pooled in equimolar ratios to create a single library for
210 sequencing on one lane of Illumina HiSeq 2500 (100 bp single end, performed by BRC).
211 The sequencing was performed with a ~10% PhiX spike-in to introduce diversity to the
212 library.

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

220 To assess repeatability across index groups, we replicated two index groups,
221 each comprised of 20 samples: one index group from the standard protocol, and the
222 index group generated with the rare-cutter EcoRI enzyme (Table 1). We multiplexed the
223 resulting 240 samples in 12 index groups (with 20 individuals each) which were pooled
224 into a final library that was sequenced on one lane of an Illumina HiSeq 2500, producing
225 220,300,739 100 bp single end reads. Eight index groups included the samples used for
226 assigning paternity and estimating relatedness, and four additional index groups were
227 included to assess the variation in the number of loci recovered with changes in our

228 molecular protocol (choice of enzyme and size selection window), and to assess
229 repeatability (see Table 1). To avoid an additional source of variation these last four
230 index groups each included the same 20 individuals.

231

232 *SNP data analysis*

233

234 Quality filtering and demultiplexing

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

244

245 De novo assembly of RAD loci

246 Because we do not have a sequenced genome for the variegated fairy wren or a
247 close relative – which is likely to be the case for many non-model organisms involved in
248 parentage studies – we assembled the sequences de novo using the STACKS pipeline
249 (Catchen *et al.* 2013). If the genome of the species of interest (or a closely related
250 species) is available, a reference-based assembly of RAD loci is preferred (Shafer *et al.*

251 2017). First, we used *denovo_map.pl* to assemble the reads into a catalog allowing a
252 minimum stack depth of 5 (m parameter), up to 5 mismatches per locus within an
253 individual (M parameter), and 5 mismatches between loci of different individuals when
254 building the catalog (n parameter). This combination of parameters has been shown to
255 work well for other similarly polymorphic passerine birds (Campagna *et al.* 2015), but we
256 expect the optimal set of parameters to vary across datasets. For a detailed exploration
257 on how the different assembly parameters in STACKS impact the number and quality of
258 loci recovered, see Paris *et al.* (2017). We ran the *rxstacks* module to filter loci with a log
259 likelihood of less than -50 (lnl_lim -50) or that were confounded in at least 25% of the
260 population (conf_lim 0.25). We then built a new catalog by rerunning *cstacks* and
261 obtained individual genotype calls with *sstacks*.

262

263 SNP filtering

264 SNPs were exported using the *populations* module of the STACKS pipeline. All of
265 our samples were grouped in one population and a locus was exported if it was present
266 in 95% of the individuals in this population (r parameter) at a stack depth of at least 10
267 (m parameter). When a RAD locus had more than one SNP, the data were restricted to
268 the first one (*--write_single_snp*) to avoid including SNPs in high linkage disequilibrium
269 (LD). We required a minor allele frequency of at least 0.25 to process a nucleotide site (*-*
270 *-min_maf*).

271 We removed loci that were not in Hardy-Weinberg equilibrium using VCFtools
272 version 0.1.14 (Danecek *et al.* 2011; *---hwe*, $\alpha = 0.05$). VCFtools implements an exact

273 test which attempts to control for type I error in large datasets (Wigginton et al. 2005).
274 VCFtools was also used to test for LD by calculating r^2 values for every pairwise
275 combination of the 411 SNPs in the final dataset. This analysis confirmed that LD was
276 very low (average of 0.01; highest values: 0.43, 0.5 and 0.65). We decided to retain the
277 entire dataset for further analyses because of the overall low LD values across all
278 pairwise combinations. LD could be more pronounced when using enzymes that cut at
279 higher frequencies, therefore it is advisable that LD be assessed before completing
280 downstream parentage analyses. We obtained a variant call format (vcf) file that was
281 converted to GENEPOP format in PGD Spider version 2.0.5.0 (Lischer & Excoffier 2012)
282 and imported into CERVUS version 3.0.7 (Kalinowski *et al.* 2007).

283

284 Assessing repeatability and overlap in different sets of RAD loci

285 We conducted five independent de novo assemblies using STACKS, one for our
286 larger set of 160 samples, and one for each of the four index groups included in the
287 sequencing run to assess repeatability and understand the impact of variations in the
288 protocol on the number of loci recovered (Table 1). Once the assemblies were
289 completed, we filtered and exported loci independently as described above. We then
290 queried what the overlap among these sets of loci was (e.g., between the full set of 160
291 samples and a replicate of 20 samples – index 9). To better understand sources of
292 variation in our experimental protocol we also asked if the loci retained after filtering
293 were present in the catalogs of different assemblies. To assess overlap in sets of RAD
294 loci we first generated FASTA files with the sequence data from both the filtered loci
295 from each assembly and from the entire catalog. We then used BLAST version 2.3.0

296 (Altschul *et al.* 1990) with an E-value of 1e-10 to find matches among these FASTA files
297 and calculate the proportion of RAD tags shared by the different sets of sequences
298 (Tables 2 and 3).

299

300 *Parentage analysis*

301 We used CERVUS version 3.0.7 (Kalinowski *et al.* 2007) to assign paternity for all
302 nestlings using our microsatellite and SNP datasets separately. CERVUS uses a two-
303 step, likelihood-based approach to assign parentage. First, CERVUS compares each
304 offspring's genotype to that of a candidate parent and a random individual in the
305 population to calculate a likelihood ratio. This relationship is presented as an LOD score,
306 which is simply the natural logarithm of the calculated likelihood ratio. Positive LOD
307 scores indicate that a candidate parent is much more likely to be the true parent,
308 whereas negative LOD scores indicate that the candidate parent is highly unlikely to be
309 a true parent. Second, CERVUS conducts a simulation of parentage analysis based on
310 population allele frequencies and the proportion of potential parents included in the
311 analysis. The simulation accounts for the possibility of unsampled parents, missing data,
312 and genotyping errors. Considering these parameters, the simulation calculates critical
313 LOD scores by comparing the LOD distributions of the most likely parent and all other
314 candidate parents. The critical LOD score is used to determine the confidence (95% or
315 80%) of each parentage assignment.

316 CERVUS allows for different types of parentage analysis, including parent-pair
317 (sexes known or unknown), maternity (known father, but not mother), and paternity
318 (known mother, but not father). Variegated fairy-wrens at Lake Samsonvale are relatively

319 easy to observe, and we were able to assign known mothers behaviorally. We
320 subsequently confirmed this with microsatellite analyses: females that built and attended
321 a nest throughout incubation were always the mothers of the nestlings in that nest. In
322 many systems, a comparable level of demographic knowledge may not be available, so
323 a marker set must be powerful enough to assign parentage with minimal social
324 information. To investigate the broader utility of our ddRAD-seq method, we conducted
325 analyses that relied on the inclusion of the known mother, in addition to analyses
326 independent of the known mother, which were based only on the father-offspring
327 relationship. We simulated paternity assignments for 10,000 offspring to determine
328 critical LOD scores, using slightly different input parameters for each panel
329 (microsatellites and ddRAD sequencing derived SNPs). Simulations for both used the
330 following parameters: 78 candidate males, 95% of candidate males sampled, estimated
331 error rate of 0.01 for mistyped loci and likelihood scores. The proportion of loci typed
332 across all individuals was different for both panels: 0.997 for the microsatellite
333 simulation, and 0.961 for the SNP simulation.

334 For both paternity analyses, we used the trio LOD score and the father-offspring
335 LOD score from CERVUS to make assignments. The trio LOD score was calculated by
336 comparing the genotypes of the candidate male and offspring, relative to that of the
337 known mother. The father-offspring LOD score only accounts for the relationship of the
338 candidate male and the offspring, independent of the known mother. CERVUS ranked
339 candidate males by LOD scores in each category, and the highest-ranking males were
340 assigned as fathers. These rankings should be in agreement, but ambiguous

341 assignments (different top-ranking males assigned in each category) may occur when
342 multiple candidate male genotypes closely match an offspring's genotype.

343 We assessed each CERVUS assignment to determine whether it was plausible,
344 and whether the assigned male was the social father or an extra-pair sire. Our criteria for
345 accepting assignments differed slightly for microsatellites and SNPs. For microsatellites,
346 we automatically accepted the CERVUS assignment if the highest-ranking male was in
347 agreement for both the trio LOD and the father-offspring LOD, and if the number of
348 mismatches between the assigned male and the offspring was ≤ 1 (8% of 12 loci). For
349 SNPs, we also accepted the assignment if the highest-ranking males by LOD score type
350 were in agreement, and an allowable number of mismatches were not exceeded.
351 However, for SNPs, our allowable number of mismatches was based on the observed
352 maximum number of mismatches between a known mother and her known offspring
353 (max. = 7, mean = 3.4, 2% of 411 loci). For both panels, we accepted the social father
354 as the genetic sire if he met these respective criteria. If the social father mismatched the
355 offspring at higher numbers, or had negative LOD scores, the offspring was considered
356 sired by an extra-pair father. We accepted assignments of extra-pair fathers using the
357 same criteria outlined above. We did not observe cases in which an offspring could not
358 be assigned to either its social father, or an extra-pair sire.

359

360 *Relatedness analysis*

361 We used the package, 'related' (Pew *et al.* 2015), in R version 3.2.5 (R Core
362 Team 2016) to estimate pairwise relatedness (r) between all pairs of individuals in this
363 study. This package accounts for genotyping errors, missing data, and can estimate

364 relatedness using any of seven different estimators (4 non-likelihood-based, and 3
365 likelihood-based). 'Related' includes the function, *compareestimators*, which tests the
366 performance of different estimators on simulated data that share the same
367 characteristics as the real data. The program uses an allele frequency file to generate
368 simulated pairs of individuals of known relatedness, and automatically estimates
369 relatedness using four of the most commonly used estimators (all non-likelihood-based).
370 The function calculates a correlation coefficient between observed and expected values,
371 to evaluate which estimator performs best with the data set. Using *compareestimators* to
372 generate 200 simulated pairs of individuals for each degree of relatedness (i.e., half-sib,
373 full-sib, parent-offspring, unrelated), we determined that the Wang (2002) estimator
374 performed best for both our microsatellite and SNP datasets. However, SNP datasets
375 generated by ddRAD-seq are sometimes prone to genotyping error through allelic
376 dropout. Attard *et al.* (2018) found that this can cause relatedness estimators to produce
377 values that are very precise, but slightly downward-biased, especially for large datasets.
378 This should be considered when selecting an appropriate relatedness estimator for
379 standalone SNP data. For consistency across the comparisons of our microsatellite and
380 SNP datasets, we obtained point estimates of relatedness using the Wang (2002)
381 estimator, and evaluated all parent-offspring relationships that were previously
382 determined in our parentage analysis.

383 'Related' also evaluates how well different marker sets resolve degrees of
384 relatedness, given simulated genotypes based on allele frequency files. For both panels,
385 we used the *familysim* function to generate 200 pairs of individuals for each degree of
386 relatedness. We then used the *coancestry* function to analyze all pairwise relatedness

387 values with the Wang (2002) estimator. We created density plots representing
388 histograms of the relatedness values. These plots show the overlap in relatedness
389 values for degree of relatedness, and we used them to infer how well each panel
390 performed at discerning different relationships.

391

392 *Comparison with other avian ddRAD datasets*

393 We applied the same molecular protocol and bioinformatics pipeline described
394 above to other avian species, with the objective of assigning paternity and estimating
395 relatedness. These datasets ranged between 6 and 480 samples, and we calculated the
396 number of loci recovered for comparisons with the current data and to assess the utility
397 of our method with larger sample sizes. We also genotyped an additional 213 variegated
398 fairy-wren individuals and re-analyzed the data in combination with the 160 samples
399 included in this study.

400

401 **Results**

402 *SNP development and analysis*

403 After trimming, filtering and demultiplexing the data, we retained a total of
404 109,524,874 reads across all index groups, with an average of approximately 9,000,000
405 reads per index group. Two individuals failed (i.e., had less than 66,000 reads each; one
406 individual from each of two index groups) and were excluded from further analysis. The
407 number of reads per sample for the remaining individuals ranged from 213,544 to
408 810,966 (mean = 459,726 \pm 118,771 std. dev.).

409 Further analysis using the population program from STACKS identified loci with at
410 least 10X coverage, present in 95% of the individuals, and a minimum allele frequency
411 greater than 25%. We further retained only those loci that were in Hardy-Weinberg
412 equilibrium. When performing analyses on all 160 individuals from the primary
413 comparison runs (Table 1), we identified 411 loci that fulfilled these criteria and were
414 used for downstream analyses.

415 We varied two aspects of the ddRAD-seq protocol to assess the number of loci
416 recovered and the reproducibility of the method. As expected, both a reduction in the
417 range of fragment sizes selected during the construction of the library (compare groups
418 A (150 bp size selection) and C (300 bp size selection)), or the use of the less frequent
419 cutter, EcoRI (groups D and E), resulted in fewer loci recovered (Table 2). Those from
420 the EcoRI digest were mostly non-overlapping with those from the MspI index groups.
421 More importantly, between the replicated index groups in our standard protocol (A and
422 B) we recovered similar numbers of loci (797 and 645, respectively) with 549 (85.1%)
423 loci found in both datasets under our stringent filtering criteria. Moreover, when we
424 searched for the loci recovered in A in the catalog of B, and vice versa, the overlap was
425 above 99% (Table 3). We also found that 208 of the 248 non-overlapping loci recovered
426 in A were not recovered in B because they did not pass the stringent missing data ($r =$
427 0.95) and minimum depth of coverage ($m = 10$) filters (Table S2, Supporting
428 Information). This suggests that very similar sets of loci were recovered in these
429 independent assemblies, but that there was variation in the loci that passed our filtering
430 criteria. Although changing our filtering parameters led to a slight increase in the
431 proportion of overlap in the loci recovered between A and B, the total number of

432 overlapping loci increased substantially. When we accepted 20% missing data instead of
433 5%, the proportion of overlap was 86%, with 857 loci in total. When we changed the
434 minor allele frequency filter from 0.25 to 0.05 we obtained 88% overlap and a total of
435 1227 loci (Table S3, Supporting Information).

436

437 *Paternity assignments*

438 Both panels produced highly concordant results when assigning paternity, but the
439 SNP panel showed substantially higher power overall. Generally, the microsatellite loci
440 were more polymorphic, resulting in greater mean polymorphic information content (PIC)
441 for any given locus. Despite this, the SNP panel performed better because of the large
442 number of loci obtained through RAD sequencing. This greatly improved the non-
443 exclusion probabilities across different parentage assignment contexts (Table 4), and
444 reduced uncertainty in our assignments. Given the known mother, the microsatellite and
445 SNP panels assigned the same fathers to all 40 offspring with 95% confidence. When
446 paternity assignments were made without the known mother (no known candidate
447 parents), both panels again assigned fathers for all 40 offspring with 95% confidence.
448 However, 5 of these assignments were not in agreement between the two panels. For
449 these 5 cases, two candidate males had very similar LOD scores under the
450 microsatellite panel, and the assigned males did not match the males that were
451 assigned given the known mother. These (and all other) cases were resolved
452 unambiguously when the SNP panel was used, and the paternity assignments with and
453 without the known mother were in complete agreement for all offspring (Fig. 1).

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

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

470

471 *Relatedness analysis*

472 The SNP panel produced simulated data that closely matched the observed allele
473 and genotype frequencies (Pearson's correlation coefficient = 0.975). The microsatellite
474 panel also matched well, but was not as reliable as the SNP panel (Pearson's
475 correlation coefficient = 0.877). This resulted in better estimates of pairwise relatedness
476 for parent-offspring using the SNP panel (Fig. 3). Overall, the SNP panel produced

477 better simulated estimates for each degree of relatedness (Fig. 4), greatly reducing the
478 variance around expected relatedness values (unrelated = 0, half-sib = 0.25, full-sib =
479 0.5, and parent-offspring = 0.5). This bolsters the confidence with which actual
480 relationships can be discerned when calculating pairwise relatedness of a population for
481 which there is little prior knowledge of social relationships.

482

483 **Discussion**

484 Several recent studies have rigorously investigated the use of SNPs in population
485 genetic studies for several non-model organisms (Morin *et al.* 2004; Slate *et al.* 2010;
486 Garvin *et al.* 2010; Heylar *et al.* 2011; Seeb *et al.* 2011), with growing support for the use
487 of SNPs in studies of parentage (Anderson & Garza 2006; e.g. Hauser *et al.* 2011; e.g.
488 Kaiser *et al.* 2017; e.g. Kess *et al.* 2016) and relatedness (e.g. Glaubitz *et al.* 2003;
489 Wang 2007). SNPs have proven to perform as well, if not better than microsatellites in
490 these types of studies. To our knowledge, this is the first study to describe a ddRAD-seq
491 method for use in parentage and relatedness analyses of wild populations, and to test it
492 in a diverse set of avian taxa. Our study is also the first to compare the efficiency of
493 microsatellites versus SNPs for determining genetic relationships in a bird species that is
494 both socially complex and highly promiscuous. We show that SNPs developed from our
495 modified ddRAD-seq method are substantially more powerful than a moderate number
496 of species-specific microsatellite loci at assigning paternity and estimating relatedness
497 among individuals. Our method is highly attractive as an alternative to traditional
498 microsatellite genotyping, especially for systems where no microsatellites have been
499 developed. This is largely due to the combination of its cost and researcher time

500 efficiency, the ease of this non-species-specific method that combines the SNP
501 discovery and screening steps, and the large number of SNPs reliably recovered.

502 The total approximate materials cost for our ddRAD-seq analyses, including DNA
503 extraction, normalization of the DNA concentrations, library preparation, sequencing and
504 computational time was US \$3,270.00 for 240 samples, or approximately \$13.6 per
505 sample. The initial investment in oligonucleotides (i.e., primers and adaptors; see the
506 Supplementary ddRAD-seq Protocol in the Supporting Information) was US \$2000 and
507 is sufficient for the analysis of thousands of samples, making the per sample cost
508 negligible. The use of a homemade MagNA in place of commercial SPRI beads provides
509 significant savings. This cost is similar to that for genotyping 240 individuals at 12
510 microsatellite loci (in 3 multiplexed PCR mixes), in a situation where the labelled primers
511 have already been designed, purchased, and tested. However, a substantial additional
512 benefit of this ddRAD-seq method is that it does not require any locus discovery or
513 development before starting. The time required for library preparation, once DNA has
514 been extracted, is modest, and once the sequence data have been obtained, SNP
515 calling for the entire dataset can be performed in less than a day through a largely
516 automated bioinformatics pipeline (see Supporting Information). Unlike manually scoring
517 peaks in traditional microsatellite genotyping analyses, the identification of SNPs is less
518 subjective and takes far fewer hours of hands-on analysis (as most is performed
519 computationally). The tools for analyzing these ddRAD data are freely available and
520 widely used (e.g., STACKS, VCFtools). Nevertheless, we note that assembling RAD loci
521 can still be challenging, and the choice of bioinformatics pipelines and specific
522 combinations of assembly parameters can influence the quality and quantity of loci

523 recovered (for detailed discussions on these issues see Eaton 2014, Mastretta-Yanes *et*
524 *al.* 2015, Shafer *et al.* 2017 and Paris *et al.* 2017).

525 For this study, our conditions and protocol allowed us to recover 411 high quality
526 SNP loci for 160 individual samples (although 240 samples were multiplexed together on
527 one lane of sequencing). However, we show that through simple variations in the size
528 selection window or the specificity of the restriction enzyme, more or fewer loci can be
529 obtained. For some applications, it could be advantageous to multiplex a greater number
530 of individuals and achieve similar coverage by aiming to recover fewer loci (e.g., using
531 EcoRI rather than MspI). Alternatively, for applications where more loci are required, the
532 size selection window could be widened and concordantly the number of individuals
533 would have to be lowered. It is also possible to vary the number of loci retained by
534 applying different bioinformatics filters. With strict filtering parameters (5% missing data,
535 minor allele frequency of 0.25 and minimum depth of coverage of 10x) we recovered
536 411 loci, which contained sufficient information to accurately assign paternity and
537 estimate relatedness among the individuals in our study. However, with only slight
538 modification of these parameters it is possible to greatly increase the number of loci
539 recovered. A total of 506 loci were retained when we allowed a minimum coverage of 5x,
540 742 with up to 20% missing data, and 910 with a minor allele frequency of 0.05 (we
541 varied one filter at a time). With respect to the entire catalog from 160 individual samples
542 (49662 loci from the ALL group), when we applied one filter at a time, we found that
543 approximately 47% of loci remained when a minimum coverage of 10x was required,
544 while only 5% remained when we allowed up to 5% missing data (Table S4, Supporting

545 Information). This suggests that the greatest loss of loci across all individuals can be
546 attributed to missing data, and to a lesser extent, depth of coverage. These missing loci
547 may be the result of DNA quality, conservation of restriction sites, size selection, or per
548 sample coverage during sequencing. Despite these losses under our stringent filtering
549 criteria, we still recovered more than a sufficient number of informative SNPs to perform
550 highly robust parentage and relatedness analyses.

551 The number of SNPs needed to perform robust parentage and relatedness
552 analyses depends on characteristics of the study population. Populations with reduced
553 genetic diversity will likely require a greater number of loci than those that are more
554 genetically diverse (Saunders *et al.* 2007; Strucken *et al.* 2016; Tortereau *et al.* 2017).
555 Obtaining more loci from the outset would aid in overcoming any issues relating to
556 population genetic diversity. Additionally, when studying species with complex social
557 systems, including for example both variable levels of genetic relatedness among
558 individuals and high rates of extra-pair fertilizations, it is imperative to obtain a sufficient
559 number of markers to discern genetic relationships robustly (Hughes 1998; Ross 2001;
560 Weinman *et al.* 2015). Our case study, using the variegated fairy-wren, shows that our
561 modified ddRAD-seq method recovers more than enough SNP loci to confidently discern
562 relationships in a species with a complex social system. Most parentage and
563 relatedness analysis programs are well equipped to handle large numbers of loci, so a
564 greater number of loci would not hinder analyses. Once an appropriate number of SNPs
565 are identified for performing robust analyses, conditions can be varied to maximize the
566 number of individuals to be genotyped. For our purposes, we conducted parentage and
567 relatedness analyses in CERVUS and the R package 'related,' respectively, to reliably

568 compare the performance of our microsatellite and SNP panels. Several other pedigree
569 reconstruction programs are readily available (e.g. COLONY, MasterBayes, and
570 Sequoia) and researchers can easily input SNP data into their preferred program (for
571 detailed comparisons of some of these programs see Karaket *et al.* 2012 and Weinman
572 *et al.* 2015). The R package, 'Sequoia' (Huisman 2017), is specifically tailored for SNP
573 data, and can reconstruct multi-generational pedigrees with as few as 100 SNPs and
574 many non-genotyped individuals. Given these considerations, 'Sequoia' may be
575 particularly useful for studies with limited social information or incomplete population
576 sampling.

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

591 the SNP panel, ambiguous assignments were nonexistent, and these cases were clearly
592 resolved (Fig. 1).

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

603 This protocol was designed to be universally applicable across bird species, and
604 we have successfully applied it in a range of other avian study systems (Table 5). While
605 different numbers of individuals were used in each study, and therefore different
606 numbers of loci were recovered, in all cases paternity was confidently assigned to
607 nestlings using CERVUS (unpublished results). We note that when the number of
608 individuals genotyped was larger (we have so far tested up to 480 individuals), the
609 number of loci recovered after filtering was smaller (as low as 135 SNPs with stringent
610 filtering parameters). Our controls designed to assess repeatability (e.g., compare group
611 A and B or D and E in Tables 2 and 3) suggest that there is a high degree of overlap
612 among replicates, but also variation in which loci pass our filters. In cases in which a
613 larger number of loci are needed to accurately assess paternity or estimate relatedness,

614 some of the filtering parameters may be relaxed. However, further testing would be
615 required to assess the informativeness of loci obtained under less stringent filtering
616 criteria. For example, for a set of 373 Variegated fairy-wren individuals (Table 5) the
617 number of loci retained increased from 157 to 410 when 20% missing data was allowed
618 instead of 5%. Accordingly, our protocol is suitable for long-term studies in which
619 samples are accumulated across several years, especially if not all individuals need to
620 be compared simultaneously (e.g. non-overlapping generations or individuals from
621 different years). It is also likely that this protocol can be applied successfully for studies
622 (short or long-term) where thousands of individuals need to be compared at one time.
623 This remains to be shown, though, and in such cases it may be more appropriate to use
624 techniques based on microarrays (see Fernández *et al.* 2013, Liu *et al.* 2016, and
625 Tortereau *et al.* 2017).

626 Applying this general protocol to many non-avian taxa will simply require ensuring
627 that specific restriction enzymes and fragment size windows are chosen appropriately.
628 The size of the genome and the number of individuals multiplexed will have to be taken
629 into consideration to achieve the desired coverage.

630 In summary, our ddRAD-seq method provides a cost effective and robust way to
631 identify SNPs for use in studies utilizing parentage and relatedness analyses. Our
632 experiment shows that a majority of the same SNPs can be obtained across groups,
633 using the same size selection windows and restriction enzymes. Future individuals can
634 be genotyped and incorporated to the analysis by re-running the STACKS pipeline.
635 Using a bird exhibiting great social complexity, and high promiscuity, we have shown

636 that SNPs identified by ddRAD-seq are more effective at assigning paternity and
637 estimating relatedness than highly polymorphic, species-specific microsatellite loci.

638

639 **Acknowledgements**

640 We thank D. Baldassarre, K. Gielow, J. Welklin, and field technicians at Lake
641 Samsonvale who assisted with field efforts for this study. We are grateful to L. Stenzler
642 and S. Bogdanowicz for help with microsatellite discovery, development, and
643 genotyping. D. Baldassarre, E. Greig, & A. Dalziel provided valuable discussion on the
644 study design. We thank J. LaPergola, J. Welklin, B. Van Doren, and M. Kinnaird for
645 providing the unpublished data shown in Table 5. We are grateful to Y. Bourgeois and
646 two anonymous reviewers for valuable comments on previous versions of this
647 manuscript. This research was supported by the US National Science Foundation (IOS-
648 1353681 and DEB-1721662).

649

650

651

652

653

654

655 **References**

- 656
657 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment
658 search tool. *Journal of Molecular Biology*, **215**, 403-410.
659 Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-
660 scale parentage inference. *Genetics*, **172**, 2567–2582.
661 Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the

- 662 power of RADseq for ecological and evolutionary genomics. *Nature Reviews*
663 *Genetics*, **17**, 81–92.
- 664 Attard, RMC, Beheregaray, LB, Möller, LM (2018) Genotyping-by-sequencing for
665 estimating relatedness in nonmodel organisms: Avoiding the trap of precise bias.
666 *Molecular Ecology Resources*, **00**, 1–10.
- 667 Avise JC, Jones AG, Walker D, Dewoody JA (2002) Genetic mating systems and
668 reproductive natural histories of fishes: lessons for ecology and evolution. *Annual*
669 *Review of Genetics*, **36**, 19–45.
- 670 Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping
671 using sequenced RAD markers. *PLoS ONE*, **3**, e3376.
- 672 Ball AD, Stapley J, Dawson DA *et al.* (2010) A comparison of SNPs and microsatellites
673 as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*).
674 *BMC Genomics*, **11**, 218.
- 675 Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis
676 in natural populations. *Trends in Ecology & Evolution*, **18**, 503–511.
- 677 Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide
678 polymorphisms in inferences of population history. *Trends in Ecology & Evolution*,
679 **18**, 249–256.
- 680 Campagna L, Gronau I, Silveira LF, Siepel A, Lovette IJ (2015) Distinguishing noise
681 from signal in patterns of genomic divergence in a highly polymorphic avian
682 radiation. *Molecular Ecology*, **24**, 4238–4251.
- 683 Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an
684 analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.
- 685 Coates BS, Sumerford DV, Miller NJ *et al.* (2009) Comparative Performance of Single
686 Nucleotide Polymorphism and Microsatellite Markers for Population Genetic
687 Analysis. *Journal of Heredity*, **100**, 556–564.
- 688 Cramer ERA, Hall ML, De Kort SR, Lovette IJ, Vehrencamp SL (2011) Infrequent Extra-
689 Pair Paternity in the Banded Wren, a Synchronously Breeding Tropical Passerine.
690 *The Condor*, **113**, 637–645.
- 691 Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools.
692 *Bioinformatics*, **27**, 2156–2158.
- 693 Davey JW, Blaxter ML (2010) RADSeq: next-generation population genetics. *Briefings in*
694 *Functional Genomics*, **9**, 416–423.
- 695 Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker
696 discovery and genotyping using next-generation sequencing. *Nature Reviews*
697 *Genetics*, **12**, 499–510.
- 698 Decroocq V, Fave MG, Hagen L, Bordenave L, Decroocq S (2003) Development and
699 transferability of apricot and grape EST microsatellite markers across taxa.
700 *Theoretical and Applied Genetics*, **106**, 912–922.
- 701 Eaton DA (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses.
702 *Bioinformatics*, **30**, 1844–1849.
- 703 Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2012) SNP discovery
704 and genotyping for evolutionary genetics using RAD sequencing. In: *Molecular*
705 *Methods for Evolutionary Genetics*. Methods in Molecular Biology. pp. 157–178.
706 Humana Press, Totowa, NJ.

- 707 Fernández ME, Goszczynski DE, Lirón JP *et al.* (2013) Comparison of the effectiveness
708 of microsatellites and SNP panels for genetic identification, traceability and
709 assessment of parentage in an inbred Angus herd. *Genetics and Molecular Biology*,
710 **36**, 185–191.
- 711 Galbusera P (2000) Cross-species amplification of microsatellite primers in passerine
712 birds. *Conservation Genetics*, **1**, 163–168.
- 713 Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms
714 to non-model species: a technical review. *Molecular Ecology Resources*, **10**, 915–
715 934.
- 716 Glaubitz JC, Rhodes OE, Dewoody JA (2003) Prospects for inferring pairwise
717 relationships with single nucleotide polymorphisms. *Molecular Ecology*, **12**, 1039–
718 1047.
- 719 Griffith SC, Owens IPF, Thuman KA (2002) Extra pair paternity in birds: a review of
720 interspecific variation and adaptive function. *Molecular Ecology*, **11**, 2195–2212.
- 721 Guichoux E, Lagache L, Wagner S *et al.* (2011) Current trends in microsatellite
722 genotyping. *Molecular Ecology Resources*, **11**, 591–611.
- 723 Gut IG (2001) Automation in genotyping of single nucleotide polymorphisms. *Human*
724 *Mutation*, **17**, 475–492.
- 725 Hadfield JD, Richardson DS, Burke T (2006) Towards unbiased parentage assignment:
726 combining genetic, behavioural and spatial data in a Bayesian framework. *Molecular*
727 *Ecology*, **15**, 3715–3730.
- 728 Hauser L, Baird M, Hilborn R, Seeb LW, Seeb JE (2011) An empirical comparison of
729 SNPs and microsatellites for parentage and kinship assignment in a wild sockeye
730 salmon (*Oncorhynchus nerka*) population. *Molecular Ecology Resources*, **11**, 150–
731 161.
- 732 Hedgecock D, Li G, Hubert S, Bucklin K (2004) Widespread null alleles and poor cross-
733 species amplification of microsatellite DNA loci cloned from the Pacific oyster,
734 *Crassostrea gigas*. *Journal of Shellfish Research*, **23**, 379–385.
- 735 Heylar SJ, Hemmer Hansen J, Bekkevold D *et al.* (2011) Application of SNPs for
736 population genetics of nonmodel organisms: new opportunities and challenges.
737 *Molecular Ecology Resources*, **11**, 123–136.
- 738 Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches,
739 common sources and consequences for paternal exclusion. *Molecular Ecology*, **14**,
740 599–612.
- 741 Hughes C (1998) Integrating molecular techniques with field methods in studies of social
742 behavior: a revolution results. *Ecology*, **79**, 383–399.
- 743 Huisman, J (2017) Pedigree reconstruction from SNP data: parentage assignment,
744 sibship clustering and beyond. *Molecular Ecology Resources*, **17**, 1009-1024.
- 745 Kaiser SA, Taylor SA, Chen N *et al.* (2017) A comparative assessment of SNP and
746 microsatellite markers for assigning parentage in a socially monogamous bird.
747 *Molecular Ecology Resources*, **17**, 183–193.
- 748 Karaket T, Poompuang S (2012) CERVUS vs. COLONY for successful parentage and
749 sibship determinations in freshwater prawn *Macrobrachium rosenbergii* de Man.
750 *Aquaculture*, **324**, 307-311.
- 751 Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program

- 752 CERVUS accommodates genotyping error increases success in paternity
753 assignment. *Molecular Ecology*, **16**, 1099–1106.
- 754 Kearse M, Moir R, Wilson A *et al.* (2012) Geneious Basic: An integrated and extendable
755 desktop software platform for the organization and analysis of sequence data.
756 *Bioinformatics*, **28**, 1647–1649.
- 757 Kess T, Gross J, Harper, F, Boulding EG (2016) Low-cost ddRAD method of SNP
758 discovery and genotyping applied to the periwinkle *Littorina saxatilis*. *Journal of*
759 *Molluscan Studies*, **82**, 104–109.
- 760 Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic
761 distribution, putative functions and mutational mechanisms: a review. *Molecular*
762 *Ecology*, **11**, 2453–2465.
- 763 Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for
764 connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–
765 299.
- 766 Liu S, Palti Y, Gao G, Rexroad CE (2016) Development and validation of a SNP panel
767 for parentage assignment in rainbow trout. *Aquaculture*, **452**, 178–182.
- 768 Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC (2015)
769 Restriction site-associated DNA sequencing, genotyping error estimation and de
770 novo assembly optimization for population genetic inference. *Molecular Ecology*
771 *Resources*, **15**, 28–41.
- 772 Morin PA, Luikart G, Wayne RK (2004) SNPs in ecology, evolution and conservation.
773 *Trends in Ecology & Evolution*, **19**, 208–216.
- 774 Myers EM, Zamudio KR (2004) Multiple paternity in an aggregate breeding amphibian:
775 the effect of reproductive skew on estimates of male reproductive success.
776 *Molecular Ecology*, **13**, 1951–1963.
- 777 Nali RC, Zamudio KR, Prado CPA (2014) Microsatellite markers for Bokermannohyla
778 species (Anura, Hylidae) from the Brazilian Cerrado and Atlantic Forest domains.
779 *Amphibia-Reptilia*, **35**, 355–360.
- 780 Paris JR, Stevens JR, Catchen JM (2017) Lost in parameter space: a road map for
781 Stacks. *Methods in Ecology and Evolution*. **8**, 1360–1373.
- 782 Pemberton JM, Slate J, Bancroft DR, Barrett JA (1995) Nonamplifying alleles at
783 microsatellite loci: a caution for parentage and population studies. *Molecular*
784 *Ecology*, **4**, 249–252.
- 785 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest
786 RADseq: An inexpensive method for de novo SNP discovery and genotyping in
787 model and non-model species (L Orlando, Ed.). *PLoS ONE*, **7**, e37135.
- 788 Pew J, Muir PH, Wang J, Frasier TR (2015) related: an R package for analysing pairwise
789 relatedness from codominant molecular markers. *Molecular Ecology Resources*, **15**,
790 557–561.
- 791 Primmer CR, N Painter J, T Koskinen M, U Palo J, Merilä J (2005) Factors affecting
792 avian cross-species microsatellite amplification. *Journal of Avian Biology*, **36**, 348–
793 360.
- 794 Puritz JB, Matz MV, Toonen RJ *et al.* (2014) Demystifying the RAD fad. *Molecular*
795 *Ecology*, **23**, 5937–5942.
- 796 Queller DC, Strassmann JE, Hughes CR (1993) Microsatellites and kinship. *Trends in*

- 797 *Ecology & Evolution*, **8**, 285–288.
- 798 R Core Team (2016) R: A Language and Environment for Statistical Computing. R
799 Foundation for Statistical Computing, Vienna, Austria. [http:// www.R-project.org/](http://www.R-project.org/).
- 800 Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for
801 multiplexed target capture. *Genome Research*, **22**, 939–946.
- 802 Ross KG (2001) Molecular ecology of social behaviour: analyses of breeding systems
803 and genetic structure. *Molecular Ecology*, **10**, 265–284.
- 804 Rowley I, Russell EM (1997) *Fairy-wrens and Grasswrens: Maluridae*. Oxford University
805 Press.
- 806 Saunders IW, Brohede J, Hannan GN (2007) Estimating genotyping error rates from
807 Mendelian errors in SNP array genotypes and their impact on inference. *Genomics*,
808 **90**, 291–296.
- 809 Schodde, R (1982). *The fairy-wrens* (ed. Bass T). The Craftsman Press. Victoria,
810 Australia.
- 811 Seeb JE, Carvalho G, Hauser L *et al.* (2011) Single nucleotide polymorphism (SNP)
812 discovery and applications of SNP genotyping in nonmodel organisms. *Molecular*
813 *Ecology Resources*, **11**, 1–8.
- 814 Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using
815 and evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.
- 816 Shafer A, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW, Wolf JB (2016)
817 Bioinformatic processing of RAD-seq data dramatically impacts downstream
818 population genetic inference. *Methods in Ecology and Evolution*, **8**, 907–917.
- 819 Slate J, Gratten J, Beraldi D *et al.* (2010) Gene mapping in the wild with SNPs:
820 guidelines and future directions. *Genetica*, **136**, 97–107.
- 821 Solomon NG, Keane B, Knoch LR (2004) Multiple paternity in socially monogamous
822 prairie voles (*Microtus ochrogaster*). *Canadian Journal of Zoology*, **82**, 1667–1671.
- 823 Strucken EM, Lee SH, Lee HK *et al.* (2016) How many markers are enough? Factors
824 influencing parentage testing in different livestock populations. *Journal of Animal*
825 *Breeding and Genetics*, **133**, 13–23.
- 826 Syvänen A-C (2001) Accessing genetic variation: genotyping single nucleotide
827 polymorphisms. *Nature Reviews Genetics*, **2**, 930–942.
- 828 Tokarska M, Marshall T, Kowalczyk R *et al.* (2009) Effectiveness of microsatellite and
829 SNP markers for parentage and identity analysis in species with low genetic
830 diversity: the case of European bison. *Heredity*, **103**, 326–332.
- 831 Tortereau F, Moreno CR, Tosser-Klopp G, Servin B, Raoul J (2017) Development of a
832 SNP panel dedicated to parentage assignment in French sheep populations. *BMC*
833 *Genetics*, **18**, 50.
- 834 Wang J (2002) An estimator for pairwise relatedness using molecular markers. *Genetics*,
835 **160**, 1203–1215.
- 836 Wang J (2007) Triadic IBD coefficients and applications to estimating pairwise
837 relatedness. *Genetical Research*, **89**, 135–153.
- 838 Webster MS, Reichart L (2005) Use of microsatellites for parentage and kinship
839 analyses in animals. *Methods in Enzymology*, **395**, 222–238.
- 840 Weinman LR, Solomon JW, Rubenstein DR (2015) A comparison of single nucleotide
841 polymorphism and microsatellite markers for analysis of parentage and kinship in a

- 842 cooperatively breeding bird. *Molecular Ecology Resources*, **15**, 502–511.
843 Westneat DF, Sherman PW, Morton ML (1990) The ecology and evolution of extra-pair
844 copulations in birds. *Current Ornithology*, **7**, 331–370.
845 White PS, Densmore LD (1992) Mitochondrial DNA isolation. In: *Molecular Genetic*
846 *Analysis of Populations: A Practical Approach* (ed. Hoelzel AR), pp. 50–51. Oxford
847 University Press, New York, NY.
848 Wigginton, JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg
849 equilibrium. *American Journal of Human Genetics*, **76**, 887-893.
850 Willing E-M, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for
851 de novo assembly and marker design without available reference. *Bioinformatics*,
852 **27**, 2187–2193.

853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869

870 **Data Accessibility**

871 RAD loci from de novo assembly in STACKS (Groups All and A-E), and Microsatellite
872 and SNP genotypes: Dryad Digital Repository: doi:10.5061/dryad.c76pf34.

873

874 **Author Contributions**

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

879
880
881
882
883
884
885
886

887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919

Tables

Table 1: Experimental design. Index groups 9-12 were included in our sequencing run to assess how changes in our molecular protocol impacted the number of loci recovered. Therefore, we selected the same 20 individuals for these four index groups to reduce the possible sources of variation.

Number of samples (index groups)	Enzymes	Size selection interval	Group
160 samples (8 index groups, index 1-8)	Sbfl - MspI	450 - 600 bp	All
20 samples (Index 1)	Sbfl - MspI	450 - 600 bp (replicate of above)	A
20 samples (Index 9)	Sbfl - MspI	450 - 600 bp (replicate of above)	B
20 samples (index 10)	Sbfl - MspI	400 - 700 bp (wide size selection)	C
20 samples (index 11)	Sbfl - EcoRI	450 - 600 bp (infrequent 3' cutter)	D
20 samples (index 12)	Sbfl - EcoRI	450 - 600 bp (replicate of above)	E

920
921

922 Table 2. Overlap in the RAD loci that were obtained while varying different steps of the
 923 protocol (size selection and restriction enzymes). The diagonal indicates the total
 924 number of loci recovered for each treatment. Values above the diagonal represent the
 925 percent overlapping loci between groups (relative to the group with the smallest number
 926 of loci), while values below the diagonal list the number of loci that were overlapping
 927 between groups.
 928

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

929

All: 160 samples; Sbf1/Msp1; 450-600 bp.
 A: 20 samples; Sbf1/Msp1; 450-600 bp (Index 1).
 B: 20 samples; Sbf1/Msp1; 450-600 bp (Index 9).
 C: 20 samples; Sbf1/Msp1; 400-700 bp (Index 10).
 D: 20 samples; Sbf1/EcoRI; 450-600 bp (Index 11).
 E: 20 samples; Sbf1/EcoRI; 450-600 bp (Index 12).

930
 931 Table 3. Overlap in filtered RAD loci and those present in the catalogs from the different
 932 assemblies. The filtered loci from the groups in the different rows were aligned against
 933 databases generated from the catalogs of the groups in the columns. The numbers in
 934 the table are the percent of filtered loci from each group that match loci that are in the
 935 catalog of the target database. The diagonal contains the total number of loci recovered
 936 for each group and the total number of loci in each catalog are 49662 (All), 27085 (A),
 937 25521 (B), 29554 (C), 11823 (D), 16724 (E).
 938

	All	A	B	C	D	E
All	411	99.5	99.3	99.0	5.6	9.2
A	99.2	797	99.1	99.1	7.9	11.9
B	99.2	99.1	645	99.1	7.9	12.6
C	98.5	88.7	89.8	1440	7.4	11.4
D	37.1	22.8	21.8	22.5	596	99.7
E	42.8	21.3	19.0	20.2	96.3	516

939

940

941 Table 4. Marker characteristics. He: Expected heterozygosity; Ho: Observed
 942 heterozygosity; PIC: Polymorphic information content.

943

Marker Panel	Number of loci	Mean proportion loci typed	Mean alleles per locus	Mean H _e	Mean H _o	Mean PIC	Nonexclusion probability (first parent)	Nonexclusion probability (second parent)	Nonexclusion probability (parent pair)
Microsatellites	12	0.99	14.17	0.77	0.76	0.74	1.9 x 10 ⁻⁴	1.9 x 10 ⁻⁶	1.5 x 10 ⁻¹⁰
SNPs	411	0.98	2.00*	0.45	0.45	0.35	5.2 x 10 ⁻²⁰	6.9 x 10 ⁻³⁵	1.0 x 10 ⁻⁵⁵

*Only biallelic SNPs were retained. If a locus had 3 alleles across the population, it was filtered from the dataset.

944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965

Table 5: Summary information for ddRAD-seq studies performed to investigate parentage in other bird species. Only SNPs with up to 5% missing data, with a minimum coverage of 10x and a minor allele frequency of 0.25 were retained. HWE: Hardy-Weinberg equilibrium.

Species	Scientific name	Number of individuals	Number of loci	Number of loci in HWE
Variegated fairy-wren*	<i>Malurus lamberti</i>	373	234	157
Hispaniolan woodpecker**	<i>Melanerpes striatus</i>	288	179	135
Northern Red-billed hornbill	<i>Tockus erythrorhynchus</i>	40	475	414
Von der Decken's hornbill	<i>Tockus deckeni</i>	112	490	410
Sapayoa	<i>Sapayoa aenigma</i>	6	672	671
Red-backed fairy-wren***	<i>Malurus melanocephalus</i>	480	329	167

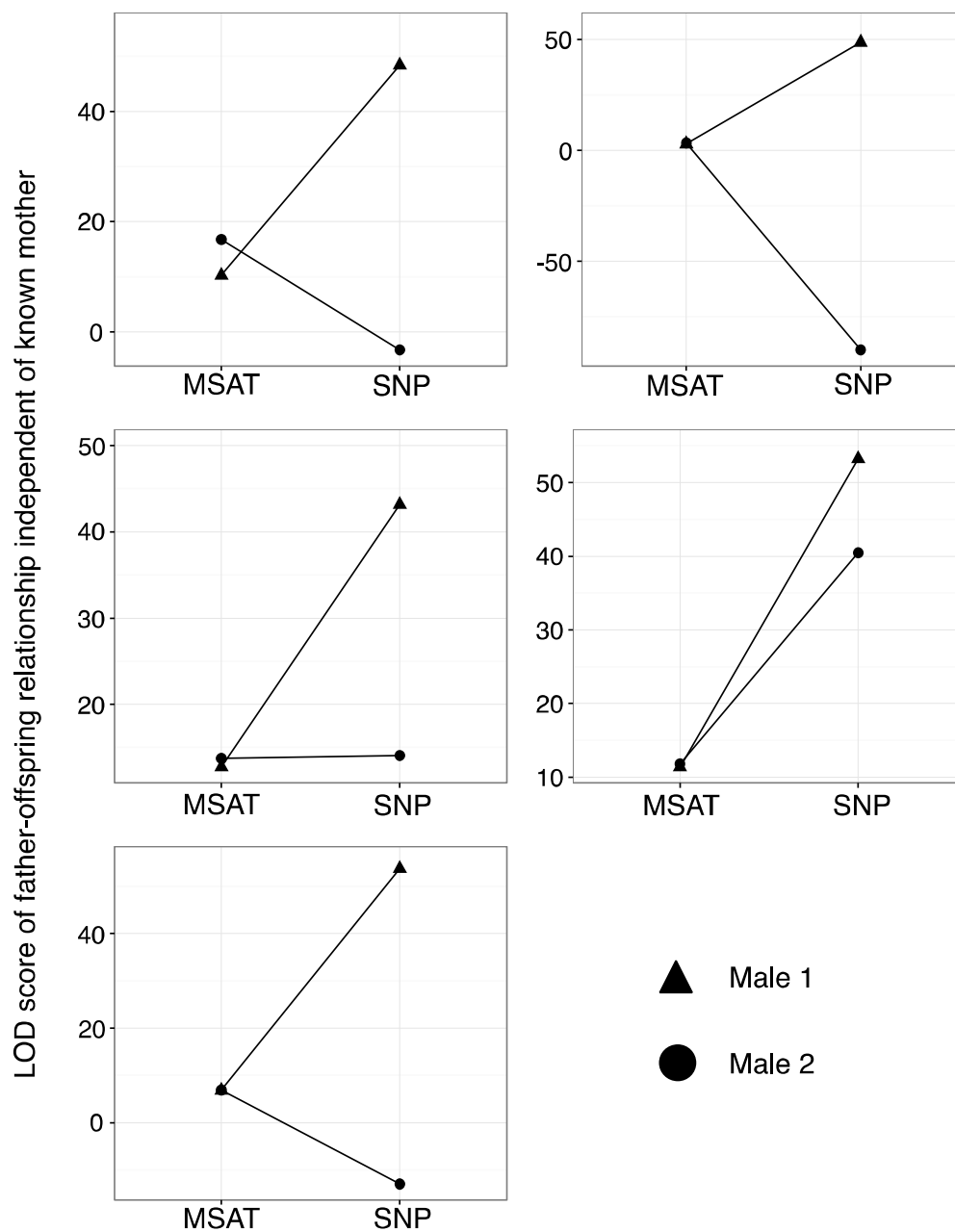
966
967 * Two independent ddRAD-seq experiments were performed – one with 160 samples
968 (those from the current study) and a second with 213 samples. After filtering and
969 demultiplexing, the data from 373 samples were combined for denovo assembly and
970 SNP identification.
971 ** Samples were run in two experiments and combined, one with 240 samples and the
972 other with 48.
973 *** Two independent ddRAD-seq experiments were run on each set of 240 samples.

974
975

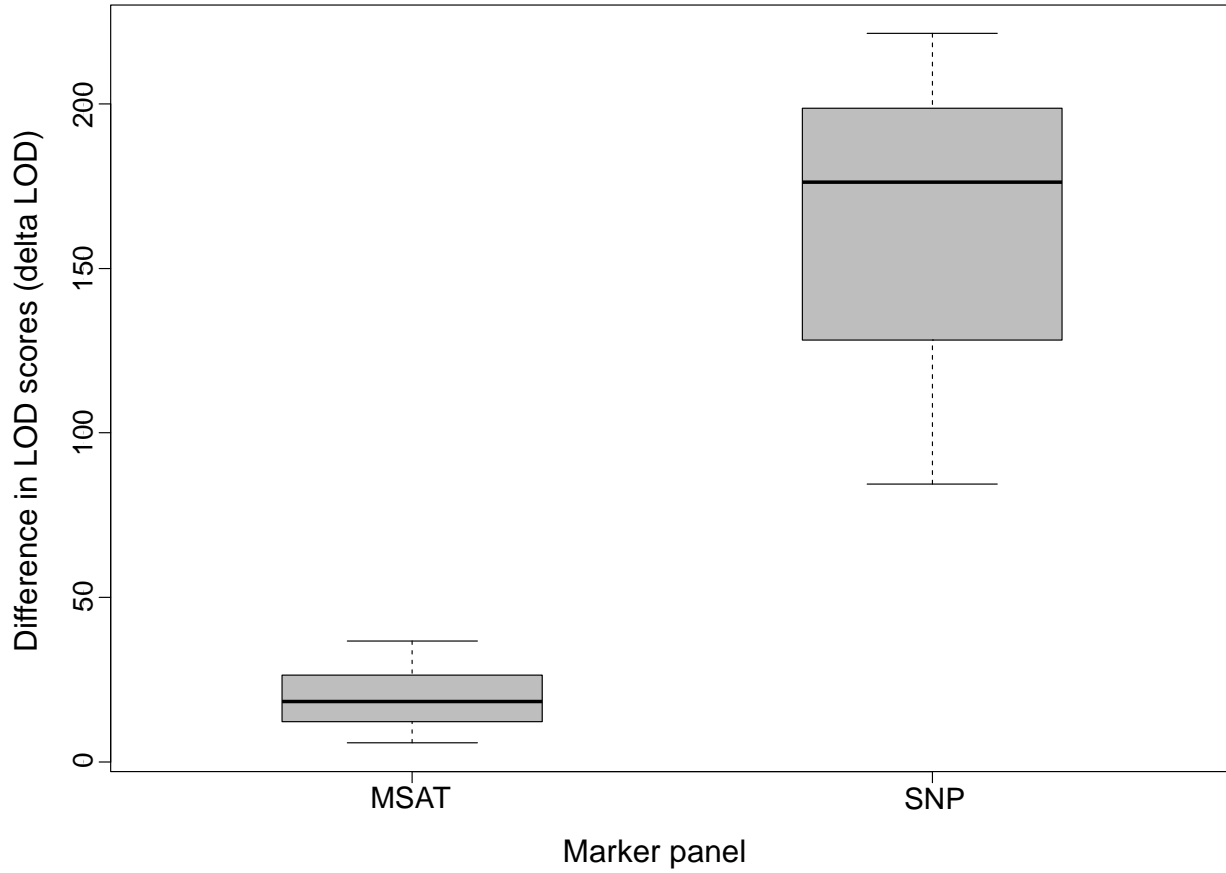
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994

995 **Figures**

996 Figure 1. Resolved paternity assignments for 5 nestlings with ambiguous assignments
997 under the microsatellite panel, but not with the SNP panel. Each panel in the graph
998 represents an individual offspring, and the two top-ranked males are depicted as a
999 triangle and a circle, respectively. Lines connecting like shapes show the change in LOD
1000 score for each male, using each marker type (microsatellites versus SNPs). Note that
1001 the y-axis scale varies among panels in the graph.

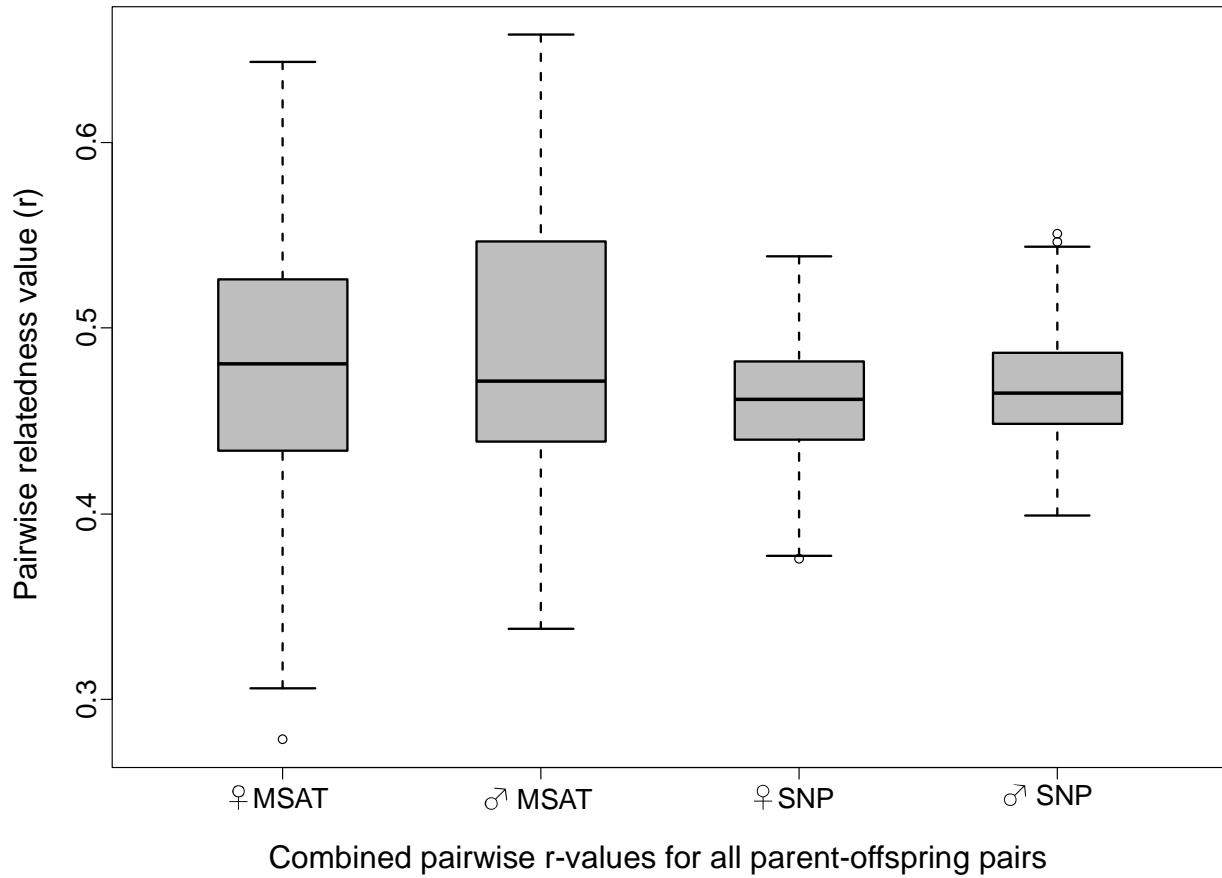


1003 Figure 2. Difference in CERVUS LOD scores (delta LOD) between the most likely father
1004 of a nestling and the second possible father in the population, for both marker panels.



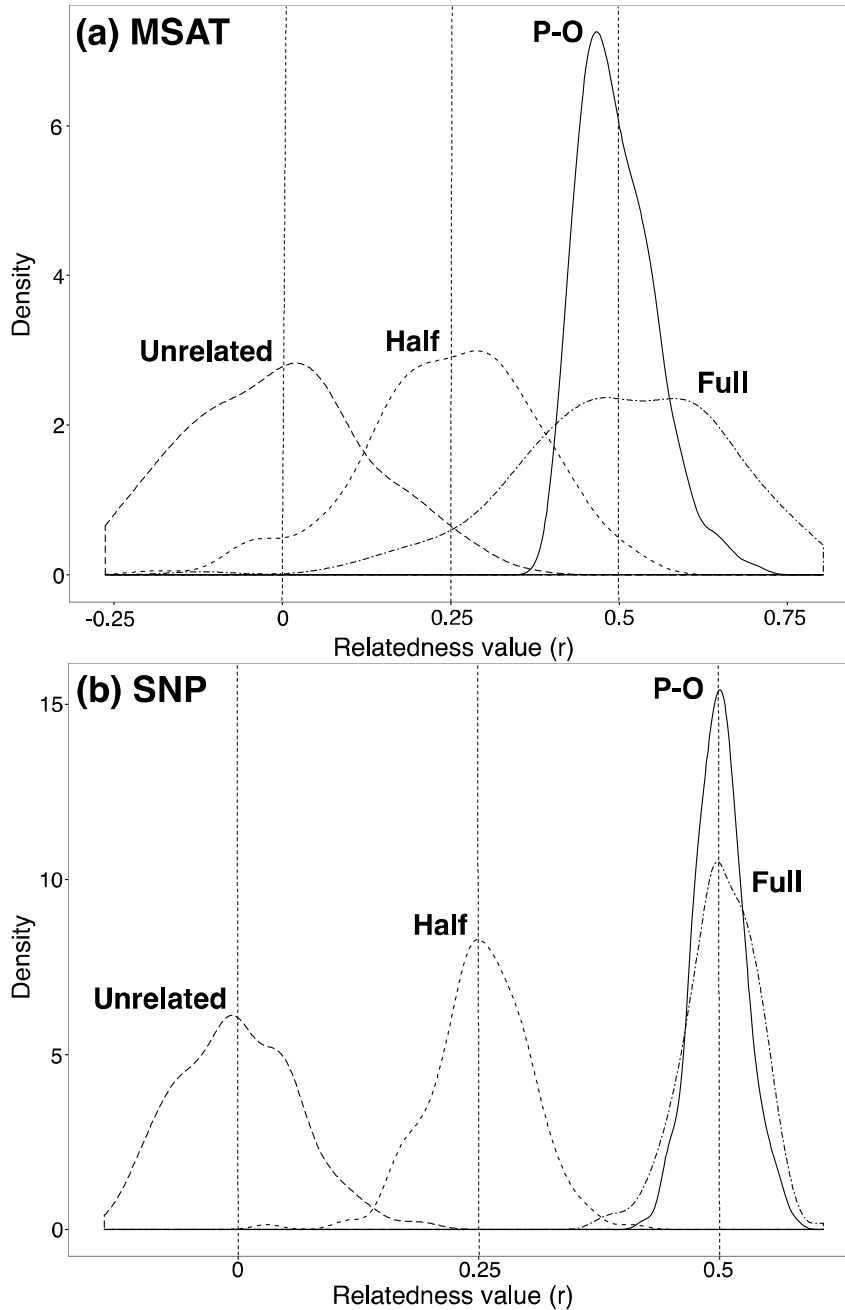
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021

1022 Figure 3. Box plot of pairwise relatedness values for all parent-offspring (40 mother-
1023 offspring and 40 father-offspring) relationships, using population allele frequencies from
1024 each marker panel.
1025



1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040

1041 Figure 4. Density plots of relatedness values for simulated pairs of known relatedness
1042 (unrelated, half-sibling, full-sibling, and parent-offspring) using population allele
1043 frequencies from each marker panel (a. MSAT; b. SNP). Overlap in distributions
1044 indicates the overlap between relatedness value estimators for pairs of individuals of
1045 different relationships. The spread of each distribution indicates the reliability of
1046 observed relatedness values based on their deviation from expected relatedness values
1047 (Unrelated = 0, Half-sib = 0.25, Full-sib = 0.5, and Parent-offspring (P-O) = 0.5, denoted
1048 by vertical dashed lines).
1049



1050