

A peer-reviewed version of this preprint was published in PeerJ on 21 April 2015.

[View the peer-reviewed version](https://doi.org/10.7717/peerj.895) (peerj.com/articles/895), which is the preferred citable publication unless you specifically need to cite this preprint.

Harvey MG, Judy CD, Seeholzer GF, Maley JM, Graves GR, Brumfield RT. 2015. Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. PeerJ 3:e895 <https://doi.org/10.7717/peerj.895>

Similarity thresholds used in short read assembly reduce the comparability of population histories across species

Michael G Harvey, Caroline Duffie Judy, Glenn F Seeholzer, James M Maley, Gary R Graves, Robb T Brumfield

Comparing inferences among datasets generated using short read sequencing may provide insight into the concerted effects of evolutionary processes across organisms, but comparisons are complicated by biases introduced during dataset assembly. Sequence similarity thresholds allow the *de novo* assembly of short reads into loci for analysis, but the resulting datasets are sensitive to both the similarity threshold used and to the variation naturally present in the organism under study. Stringent thresholds as well as highly variable species may result in datasets in which divergent alleles are lost or divided into separate loci ('over-splitting'), whereas liberal thresholds increase the risk of paralogous loci being combined into a single locus ('under-splitting'). Comparisons among datasets or species are therefore potentially biased if different similarity thresholds are applied or if the species differ in levels of genetic variation. We examine the impact of a range of similarity thresholds on assembly of empirical short read datasets from populations of four different non-model bird lineages (species or species pairs) with different levels of genetic divergence. We find that, in all species, stringent similarity thresholds result in fewer alleles per locus than more liberal thresholds, which appears to be the result of high levels of over-splitting at stringent thresholds. The frequency of putative under-splitting, conversely, is low at all thresholds. Inferred genetic distances between individuals, gene tree depths, and estimates of the ancestral mutation-scaled effective population size (θ) differ depending upon the similarity threshold applied. Relative differences in inferences across species differ even when the same threshold is applied, but may be dramatically different when datasets assembled under different thresholds are compared. We suggest some best practices for assembling short read data to maximize comparability, such as using more liberal thresholds and examining the impact of different thresholds on each dataset.

2 Michael G. Harvey

3
4 *Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA*
5 *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA*
6

7 Caroline Duffie Judy

8
9 *Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA*
10 *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA*
11 *Department of Vertebrate Zoology, MRC-116, National Museum of Natural History, Smithsonian*
12 *Institution, Washington, DC 20013, USA*
13

14 Glenn F. Seeholzer

15
16 *Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA*
17 *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA*
18

19 James M. Maley

20
21 *Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA*
22 *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA*
23 ⁵*Moore Laboratory of Zoology, Occidental College, Los Angeles, CA 90041 USA*
24

25 Gary R. Graves

26
27 ³*Department of Vertebrate Zoology, MRC-116, National Museum of Natural History, Smithsonian*
28 *Institution, Washington, DC 20013, USA*
29 ⁴*Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of*
30 *Copenhagen, DK-2100, Copenhagen Ø, Denmark*
31

32 Robb T. Brumfield

33
34 ¹*Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA*
35 ²*Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA*
36

37 Corresponding author:

38 Michael G. Harvey
39 Museum of Natural Science
40 119 Foster Hall
41 Louisiana State University
42 Baton Rouge, LA 70803, USA
43 225-578-2855
44 mharve9@lsu.edu
45

47 With the proliferation of population-level datasets obtained using massively parallel sequencing
48 technologies, there is increasing interest in studies that compare inferences from genomic datasets
49 obtained from different species (e.g., Leaché et al., 2013; Smith et al., 2013) or from different genomic
50 regions (e.g., Evans et al., 2014; Harvey et al., 2013; Leaché et al., in press). Assembly of short
51 sequence reads into orthologous loci is a key component of post-sequence processing, and commonly
52 used methods can lead to biases in population genetic parameter estimation (Ilut, Nydam & Hare,
53 2014). Here, we explore the effect of one major source of bias on the comparability of datasets and
54 inferences.

55 Sequence similarity provides the information necessary for assembling reads into orthologous
56 loci (Pop & Salzberg 2008; Chaisson, Brinza & Pevzner, 2009). By setting a sequence similarity
57 threshold, researchers attempt to assemble similar, presumably orthologous reads into loci while
58 separating or removing dissimilar, presumably non-orthologous reads (e.g. Etter et al., 2011; Catchen
59 et al., 2011). Selecting the most appropriate similarity threshold is challenging, primarily because the
60 amount of genetic (allelic) variation can vary greatly among orthologous loci within a species (Ilut,
61 Nydam & Hare, 2014). Because the amount of genetic variation also varies among species and
62 genomic regions, a particular similarity threshold may impact each dataset differently, potentially
63 impacting inferences in comparative studies.

64 Many methods default to a stringent similarity threshold, often requiring 98-99% sequence
65 similarity among reads for assembly (e.g., Catchen et al., 2011; Lu et al., 2013). However, stringent
66 similarity thresholds may split orthologous reads into multiple loci if the reads come from alleles that
67 are more different than the threshold permits (hereafter “over-splitting”; Fig. 1a). More liberal
68 similarity thresholds permit the assembly of more dissimilar orthologous reads into loci, but are more
69 susceptible to including paralogous reads in the assembly (hereafter “under-splitting”; Fig. 1b). Using
70 simulations, Rubin, Ree and Moreau (2012) found that under-splitting was frequent at more liberal

71 similarity thresholds in phylogenetic datasets, but did not strongly bias inference. Catchen et al. (2013)
72 examined RAD-Seq data from three-spined sticklebacks, and found that over-splitting was an issue
73 when datasets were processed with similarity thresholds more stringent than 96%. Ilut, Nydam and
74 Hare (2014) tested the impact of similarity threshold selection on both over- and under-splitting in
75 three simulated and one empirical RAD-Seq dataset. They found that under-splitting was minimal and
76 that affected loci were easily identified due to the presence of individuals with more alleles than
77 expected given their ploidy, but that over-splitting was significant at more stringent similarity
78 thresholds.

79 Comparative phylogeographic and population genetics studies are particularly susceptible to
80 biases resulting from similarity thresholds, particularly over-splitting. Different species often exhibit
81 different levels of genetic diversity (Lewontin, 1974; Taberlet et al., 1998; Smith et al., 2014;
82 Romiguier et al. 2014), and this variation across species may interact with the application of similarity
83 thresholds to differentially bias datasets. Huang and Knowles (In press), for example, found that
84 mutational spectra of datasets simulated under deeper species trees were downward-biased relative to
85 those simulated under shallow species trees when processed with the same settings, including the same
86 similarity threshold. The impacts of similarity thresholds have not been examined, however, using
87 empirical data from species that vary in their levels of genetic diversity. Although diverse parameters
88 required for short read assembly are worthy of scrutiny, we focus on similarity thresholds as they are
89 particularly important for maintaining comparability across species with different levels of variation.

90 In this study, we examine the effect of similarity thresholds on dataset assembly and
91 phylogeographic inferences across four non-model bird lineages that vary in divergence. We sample
92 two populations or species within each lineage and assemble a RAD-Seq dataset for each species at a
93 series of similarity thresholds to assess the impact of different thresholds on the number of alleles
94 observed within assembled loci. We investigate the effect of different similarity thresholds on

95 estimates of standard population genetic and phylogeographic parameters within species and in
96 comparisons across species.

97

98

99 MATERIALS AND METHODS

100

101 *Study Species and Sampling*

102

103 We sampled four individuals from each of two populations in four lineages (Table S1). The
104 first lineage includes Clapper (*Rallus crepitans* J. F. Gmelin, 1788) and King (*R. elegans* J. J.
105 Audubon, 1834) rails, sister species of medium-sized water birds that interbreed in a narrow hybrid
106 zone centered on a salinity gradient (Maley 2012; Maley & Brumfield 2013). The Streamertail
107 (*Trochilus polytmus* C. Linnaeus, 1758) is a hummingbird endemic to Jamaica containing two
108 subspecies (*T. p. polytmus* and *T. p. scitulus*) that differ primarily in bill coloration, and which also
109 interbreed in a narrow hybrid zone (Gill et al., 1973; Coyne & Price, 2000). The Line-cheeked
110 Spinetail (*Cranioleuca antisiensis* P. L. Sclater, 1859) is a small insectivorous bird distributed along
111 the Andes Mountains (Remsen, 2003), from which we sampled two subspecies (*C. a. antisiensis* and *C.*
112 *a. baroni*) at either end of the distribution. Finally, we sampled two populations of Plain Xenops
113 (*Xenops minutus* A. E. Sparrman, 1788), a widespread insectivorous bird of lowland Neotropical
114 forests, that are separated by the Andes and differ in plumage, voice, and genetic markers (Remsen
115 2003; Burney 2009; Harvey & Brumfield, 2015).

116

117 *Laboratory Methods*

118

119 For each individual examined, we extracted total DNA from vouchered tissue samples using
120 DNeasy tissue kits (Qiagen, Valencia, CA, USA) following the manufacturer's protocol. We sent DNA
121 extracts to the Cornell Institute of Genomic Diversity (IGD) to collect data using Genotyping by
122 Sequencing, a RAD-Seq method (Elshire et al., 2011). Briefly, the IGD digested DNA using PstI
123 (CTGCAG) and ligated a sample-specific indexed adapter and common adapter to resulting fragments.
124 The IGD pooled and cleaned ligated samples using a QIAquick PCR purification kit (Qiagen),
125 amplified the pool using an 18-cycle PCR, purified the PCR product using QIAquick columns, and
126 quantified the amplified libraries using a PicoGreen assay (Molecular Probes, Carlsbad, CA, USA).
127 Based on the PicoGreen concentrations, the IGD then combined the samples for this project with
128 unrelated samples and ran plates of 96 samples on a 100-base pair, single-end Illumina HiSeq 2000
129 lane.

131 *Bioinformatics Processing*

132
133 We processed the raw GBS reads using the Stacks pipeline (Catchen et al., 2011; 2013) due to
134 its popularity in prior studies assembling RAD-Seq datasets within species. Although other dataset
135 assembly programs are available (e.g. Eaton, 2014; Sovic et al., in press), all should be subject to
136 similar artifacts. Datasets were assembled on compute nodes (2.93 GHz Quad Core Nehalem Xeon 64-
137 bt processors with 24GB 1333 MHz RAM or 96GB 1066MHz RAM) maintained by LSU High
138 Performance Computing. We demultiplexed raw reads, cleaned reads, and removed barcode and
139 adapter sequences using the program `process_radtags.pl`. We assembled alleles and loci *de novo* using
140 the program `denovo_map.pl`. We used custom Python (Python Software Foundation, 2007) scripts
141 (available at https://github.com/mgharvey/misc_Python) to obtain sequence alignments of both alleles
142 for each individual from the Stacks output files. Detailed settings are provided in the supplement.

143 To investigate the impact of similarity thresholds, we assembled seven datasets for each of the
144 four lineages under similarity thresholds (Stacks settings -M and -n) at all integer values from 93% (7
145 mismatches allowed) to 99% (1 mismatch allowed), reflecting the range of settings typically used for
146 assembling intraspecific datasets. Assembly with similarity thresholds less stringent than 93% failed
147 due to high computational demand in Stacks, but should not be necessary for the divergences examined
148 here. Reads with similarity values above the selected threshold clustered into assemblies, which we
149 treated as independently segregating loci in downstream analyses. We disabled the use of secondary,
150 more divergent reads for calling genotypes (Stacks setting -H) to prevent the assembly of reads that are
151 less similar than the similarity threshold used for primary stacks. We set minimum depth per allele
152 (Stacks setting -m) to seven, which provides a balance between the inclusion of singleton alleles
153 (potential errors) and the total size of the data matrix (Fig. S1). We set the maximum number of alleles
154 per individual (Stacks setting --max_locus_stacks) to three, one above the ploidy level of the study
155 organisms, in order to detect loci containing individuals with three or more alleles. We used custom
156 Python scripts (available at https://github.com/mgharvey/misc_Python) to format files and calculate
157 basic statistics and used COMPUTE (Thornton, 2003) to estimate standard population genetic
158 summary statistics.

159

160 *Number of Alleles*

161

162 We examined the number of alleles per locus across treatments to examine how different
163 similarity thresholds affected each dataset. As an index of the frequency of under-splitting in each
164 dataset, we calculated the number of loci containing individuals with more than two alleles. These loci
165 were presumed to contain paralogous reads and were removed from further analysis. To assess the
166 proportion of loci with putative over-split alleles, we mapped loci assembled under the more

167 conservative thresholds (94-99%) to the set of loci assembled under the most liberal threshold (93%).
168 This allowed us to detect instances in which multiple loci from the conservative threshold mapped to
169 the same locus from the stringent threshold. We used lastz (Harris, 2007) for mapping with minimum
170 identity set at 93% for all comparisons and no gaps permitted. We subtracted from each total the
171 number of loci from the liberal threshold (93%) that mapped to other loci assembled with the same
172 threshold using lastz.

173

174 *Genetic Distances and F_{st}*

175

176 Over-splitting may reduce estimates of genetic distance between individuals or populations by
177 splitting loci containing the most genetically dissimilar alleles. We calculated pairwise p-distances and
178 Jukes-Cantor corrected distances per unit sequence length at each locus. We measured distances
179 between individuals by measuring the average distance between both alleles. For loci containing
180 variable sites, we also estimated F_{st} between the two populations within each lineage using formula (3)
181 of Hudson, Slatkin & Maddison (1992).

182

183 *Gene Trees*

184

185 Over-splitting may also reduce average gene tree depth due to the loss of more variable loci. To
186 reduce computation, we selected a random subsample of 1,000 loci for each lineage at each threshold
187 for gene tree estimation. We selected the best-fit finite sites substitution model for each locus using
188 mrAIC.pl (Nylander, 2004) and conducted MrBayes (Ronquist and Huelsenbeck, 2003) runs with a
189 random starting tree, four Markov chains, and a 100,000-iteration burn-in followed by 1,000,000

190 sampling iterations. We measured the mean depth of gene trees in number of expected substitutions for
191 each sample using the R (R Core Team, 2014) package ape (Paradis, Claude & Strimmer, 2004).

192

193 *Demographic Parameter Estimation*

194

195 We used the 1,000 locus subsets from gene tree estimation to estimate the demographic history
196 of each lineage at each similarity threshold using the coalescent model implemented in BP&P (Yang
197 and Rannala, 2010). Although this method assumes no gene flow between populations, which may be
198 violated in some of our study lineages, simulations have demonstrated that BP&P performance is
199 robust to limited gene flow (Zhang et al., 2011). We used a speciation model containing two
200 populations and a divergence time parameter (τ) as well as population standardized mutation rate
201 parameters ($\theta = 4N_e\mu$, where N_e is the effective population size and μ is the substitution rate per site per
202 generation) for both daughter populations and an ancestral population. We set prior values using
203 gamma distributions determined by a shape parameter (α) and scale parameter (β). Priors for both
204 divergence time and population standardized mutation rate were set to $\alpha = 1$ and $\beta = 300$. We ran
205 analyses for a burn-in of 50,000 iterations and then sampled every other iteration for an additional
206 500,000 iterations.

207

208

209 RESULTS

210

211 After removing loci containing putative paralogous reads (see below), we recovered between
212 96,776 and 158,328 loci for the four lineages across the range of similarity thresholds (Table 1). The

213 similarity threshold used had an effect on the number of unique alleles per locus in all four lineages
214 (Kruskal Wallis test $p < 2.20 \cdot 10^{-16}$; Table S2). The number of alleles was low using the 99% similarity
215 threshold, but increased and plateaued as the threshold approached 93% (Fig. 2a). The number of
216 alleles was more similar across lineages at stringent thresholds than at liberal thresholds and this effect
217 impacted relative values between lineages. For example, *Xenops* contained, on average, 1.4 times as
218 many alleles as *Rallus* when processed with a 99% similarity threshold, but 1.66 times as many alleles
219 when processed with a 93% similarity threshold.

220 The proportion of loci containing putative paralogous reads increased slightly with increasing
221 similarity thresholds, but was less than 0.4% at all thresholds for all lineages (Fig. 2b). At all
222 thresholds, *Trochilus* exhibited roughly half the level of putative paralogy displayed in the other
223 lineages (Table S3). Depending on the lineage, 5 – 61% of loci represented putative over-split alleles
224 based on lastz mapping at the most stringent similarity threshold of 99%, but putative over-split alleles
225 decreased as thresholds became more liberal (Fig. 2b).

226 Genetic distances between individuals were reduced at more stringent similarity thresholds
227 (Fig. 3a). Variance across lineages in mean genetic distance increased as similarity thresholds became
228 more liberal (Fig. S2), although relative values between lineages were similar across thresholds. F_{st}
229 estimates between populations did not differ across thresholds (Fig. 3b).

230 Mean gene tree depth, based on the depth of the deepest node, increased as more liberal
231 similarity thresholds were applied in each lineage (Fig. 3c). Variance in mean gene tree depths across
232 lineages was inversely related to threshold stringency (Fig. S2) and relative values across lineages were
233 contingent on the threshold applied. For example, the mean gene tree depth for *Xenops* was $1.48 \times$
234 greater than for *Rallus* at 99% similarity, but $1.91 \times$ greater at 93% similarity.

235 Ancestral θ estimates were higher at more liberal similarity thresholds for all four lineages (Fig
236 3d), but contemporary θ estimates and population divergence times (τ) showed no association with
237 similarity thresholds (Figs. S3, S4). Ancestral θ estimates, as with genetic distance and gene tree depth,
238 displayed lower variance across lineages at stringent relative to liberal thresholds (Fig. S2). Relative
239 values across lineages also differed across thresholds. The ancestral θ for *Xenops* was 1.89 \times greater
240 than for *Rallus* at 99% similarity, for example, but 2.95 \times greater at 93% similarity.

241

242

243 DISCUSSION

244

245 Comparability of parameter estimates is essential for comparative studies of phylogeographic
246 structure and genetic diversity across species or among genomic regions (Nybom 2004). Our results
247 reveal, however, that inferences differ not only among lineages with different population histories, but
248 also according to the similarity threshold applied during dataset assembly. Differences in the impact of
249 similarity thresholds across datasets not only reduce the utility of those datasets for comparative
250 studies, but also preclude the application of standardized mutation rate estimates that would allow
251 demographic parameters in non-model species to be converted to real values (DaCosta & Sorenson,
252 2014). The issues discussed here are not restricted to RAD-Seq datasets, but are of concern for all short
253 read datasets requiring similarity-based *de novo* assembly, including those from sequence capture and
254 transcriptomics. Mapping reads to existing reference sequences also requires the application of
255 similarity thresholds and, although identifying under-splitting is more straightforward with a reference
256 genome, divergent alleles may still be lost to over-splitting if the threshold used for mapping is too
257 stringent (Trapnell & Salzberg, 2009; Lunter & Goodson, 2011). Careful selection of similarity

258 thresholds for assembly is an important issue for diverse sequencing projects, particularly if
259 comparisons are to be made across datasets.

260 We found that datasets assembled under stringent similarity thresholds included fewer unique
261 alleles per locus than those assembled under more liberal thresholds. Similarly, Ilut, Nydam and Hare
262 (2014) found heterozygosity was reduced when stringent similarity thresholds were applied, but
263 increased with more liberal thresholds across three simulated and one empirical dataset. The reduced
264 number of alleles per locus in datasets assembled with stringent thresholds is likely due to the higher
265 frequency of putative over-splitting in those datasets. Prior studies also demonstrated that over-splitting
266 is frequent when datasets are processed at stringent similarity thresholds, and that this leads to allele
267 loss (Catchen et al., 2013; Ilut, Nydam & Hare, 2014). Our results suggest that under-splitting occurs at
268 low frequencies across similarity thresholds and has little impact on datasets. The impact of under-
269 splitting may be more severe in species with highly repetitive genomes or in studies across deep,
270 phylogenetic timescales that require more liberal similarity thresholds for assembly (e.g., Rubin, Ree &
271 Moreau, 2012; Eaton & Ree, 2013).

272 Variation in datasets resulting from the similarity threshold applied has important effects on
273 downstream parameter estimation. In addition to the biases in population genetic and phylogeographic
274 estimates that we found, Huang and Knowles (In press) found that mutational spectra are downward-
275 biased as a result of the loss of the most divergent loci and phylogenetic estimates are more accurate
276 when more liberal similarity thresholds are applied to simulated data (Rubin, Ree & Moreau, 2012;
277 Huang & Knowles, in press). Unlike other parameters, our F_{st} estimates were not strongly impacted by
278 variation in similarity thresholds, perhaps because F_{st} is calculated using the ratio of between- and
279 within-population divergence, both of which are impacted by allele loss. In addition, θ values from
280 contemporary populations were similar across thresholds, while ancestral θ values were lower at more

281 stringent thresholds. This may result if stringent thresholds result in the loss of alleles that are fixed
282 between the two divergent populations at a higher rate than those that are variable within populations.
283 Despite these exceptions, it seems likely that observed biases in datasets across similarity thresholds
284 would impact diverse population genetic and phylogeographic parameter estimates.

285 Stringent similarity thresholds (98-99%) are widely applied currently to population-level
286 studies (e.g. Emerson et al., 2010; Reitzel et al., 2013; Chu et al., 2014), perhaps under the supposition
287 that they are more conservative and less likely to permit the assembly of non-orthologous reads or as
288 an attempt to reduce dataset size and computation times (Ilut, Nydam & Hare, 2014). We concur with
289 Ilut, Nydam and Hare (2014) and Huang and Knowles (In press) that defaulting to stringent thresholds
290 is generally not appropriate. Over-splitting decreases at more liberal similarity thresholds and the
291 number of alleles per locus asymptotes near the 96% threshold, suggesting that datasets assembled
292 under similarity thresholds of 96% or less stringency are relatively less biased by over-splitting.
293 Although this asymptote will vary depending on the divergence within a dataset, other studies have
294 found asymptotes at similar threshold values, for example at roughly 95-96% in empirical data from
295 sticklebacks (Catchen et al., 2013) or between roughly 88% and 96% in simulated tunicate,
296 stickleback, and soybean datasets and an empirical tunicate dataset (Ilut, Nydam & Hare, 2014). The
297 approach suggested by Ilut, Nydam and Hare (2014) in which datasets are assembled at a series of
298 similarity thresholds, the location of the asymptote in over-splitting is identified, and that threshold is
299 used for final assembly is preferable to defaulting to stringent thresholds.

300 We were unable to directly investigate the frequency of under-splitting and over-splitting in our
301 datasets because we lack genome sequences for the non-model organisms examined. Our indirect
302 measure of over-splitting may detect not just over-split loci, but also loci that are under-split in the
303 assembly from the most liberal threshold but correctly separated in the assembly from the more
304 stringent threshold. The frequency of under-splitting appears to be low enough, however, that this

305 effect would be minimal. Broad concordance between our results and prior investigations into over-
306 splitting in systems with a genome for reference (Catchen et al., 2013; Ilut, Nydam & Hare, 2014)
307 suggest that our test for over-split alleles is a reasonable proxy for use in non-model organisms.

308 Results from our indirect measure of under-splitting are also broadly consistent with the low
309 levels of under-splitting observed in prior work using reference genomes (Ilut, Nydam & Hare, 2014)
310 and were expected given the low level of paralogy in avian genomes (e.g. chicken; Hillier et al., 2004).
311 Our measure of under-splitting, the number of loci containing individuals with more alleles than
312 expected, has been used previously to filter out loci with paralogous data from RAD-Seq datasets
313 (Parchman et al., 2012; Peterson et al., 2012). Some loci may contain reads from paralogous loci but
314 may not contain sufficient numbers of alleles to trip this filter, potentially inflating estimates of
315 variation. Prior work, however, suggests that paralogous reads lack strong signal conflicting with that
316 from entirely orthologous loci and have relatively minor effects on inferences (Rubin, Ree & Moreau,
317 2011). Other indicators, such as extreme heterozygosity (White et al., 2013) or gene tree topologies
318 suggesting a history of duplication, might also be used to detect additional loci containing paralogous
319 reads in situations where under-splitting is a concern.

320 We uncovered differences in allelic diversity and population history inferences across the four
321 study lineages examined. *Xenops minutus* generally displayed the greatest allelic diversity and also the
322 largest genetic distances between individuals, deepest gene trees, and highest θ values, which was
323 perhaps not surprising given prior evidence of deep genetic divergences within this species (Smith et
324 al., 2014; Harvey & Brumfield, 2015). The other lineages were more similar by most measures,
325 although *Trochilus polytmus* was slightly higher than *Cranioleuca* and *Rallus* in allelic diversity,
326 genetic distance, and gene tree depths. Interestingly, *Trochilus polytmus* also exhibited roughly half the

327 amount of under-splitting, or putative paralogous loci, of the other three species, which may be related
328 to the small genome size of hummingbirds (Gregory et al., 2009).

329 Our results suggest that the similarity threshold used for assembly impacts the level of variation
330 in a dataset as well as downstream population genetic and phylogeographic estimates. Comparisons
331 across datasets are also biased by the impact of similarity thresholds, appearing more similar across
332 datasets when stringent thresholds are used or in some cases more different if species are assembled
333 with different thresholds. Methods for threshold selection exist that limit these biases (Ilut, Nydam &
334 Hare, 2014), but they need to be further developed and applied more widely across studies if we are to
335 be able to compare inferences and integrate inferences across studies, genomic regions, and organisms.

336

337

338 ACKNOWLEDGEMENTS

339

340 We thank the many collectors and museum curators and staff involved in obtaining and maintaining the
341 samples used for this study, in particular John M. Bates and David E. Willard (FMNH), Mark B.
342 Robbins (KUMNH), Donna L. Dittmann (LSUMNS), and James Dean (USNM). Jeremy M. Brown,
343 Michael E. Hellberg, Prosanta Chakrabarty, Jacob A. Esselstyn, and the LSU Vert Lunch group
344 provided helpful comments on the issue addressed in this paper.

345

346

347 REFERENCES

348

349 Burney CW. 2009. *Comparative phylogeography of Neotropical birds*. D. Phil. Dissertation, Louisiana
350 State University.

351

352 Catchen JM, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH. 2011. Stacks: building and
353 genotyping loci *de novo* from short-read sequences. *G3 Genes Genomes Genetics* 1:171-182.

354

355 Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for
356 population genomics. *Molecular Ecology* 22:3124-3140.

357

358 Chaisson MJ, Brinza D, Pevzner PA. 2009. *De novo* fragment assembly with short mate-paired reads:
359 does the read length matter? *Genome Research* 19:336-346.

360

361 Chu ND, Kaluziak ST, Trussell GC, Vollmer SV. 2014. Phylogenomic analyses reveal latitudinal
362 population structure and polymorphisms in heat stress genes in the North Atlantic snail *Nucella*
363 *lapillus*. *Molecular Ecology* 23:1863-1873.

364

365 Coyne JA, Price TD. 2000. Little evidence for sympatric speciation in island birds. *Evolution* 54:2166-
366 2171.

367

368 Eaton DA. 2014. PyRAD: Assembly of *de novo* RADseq loci for phylogenetic analyses.
369 *Bioinformatics* 30:1844-1849.

370

371 Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: An example
372 from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology* 62:689-706.

373

- 374 Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust,
375 simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One*
376 **6**:e19379.
- 377
- 378 Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. 2010.
379 Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the*
380 *National Academy of Sciences* **107**:16196-16200.
- 381
- 382 Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA. 2011. Local *de novo* assembly of RAD
383 paired-end contigs using short sequencing reads. *PLoS One* **6**:e18561.
- 384
- 385 Evans BJ, Zeng K, Esselstyn JA, Charlesworth B, Melnick DJ. 2014. Reduced representation genome
386 sequencing suggests low diversity on the sex chromosomes of Tonkean macaque monkeys.
387 *Molecular Biology and Evolution* **31**:2425-2440.
- 388
- 389 Gill F, Stokes C. 1973. Contact zones and hybridization in the Jamaican hummingbird, *Trochilus*
390 *polytmus* (L.). *Condor* **75**:170-176.
- 391
- 392 Gregory TR, Andrews CB, McGuire JA, Witt CC. 2009. The smallest avian genomes are found in
393 hummingbirds. *Proceedings of the Royal Society B* **276**:3753-3757.
- 394
- 395 Harris RS. 2007. *Improved pairwise alignment of genomic DNA*. D. Phil. Dissertation, The
396 Pennsylvania State University.
- 397

398 Harvey MG, Brumfield RT. 2015. Genomic variation in a widespread Neotropical bird (*Xenops*
399 *minutus*) reveals divergence, population expansion, and gene flow. *Molecular Phylogenetics*
400 *and Evolution* 83:305-316.

401

402 Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT. 2013. Sequence capture versus
403 restriction site associated DNA sequencing for phylogeography. *arXiv*:1312.6439.

404

405 Huang H, Knowles LL. In press. Unforeseen consequences of excluding missing data from next-
406 generation sequences: simulation study of RAD sequences. *Systematic Biology* doi:10.1093.

407

408 Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence
409 data. *Genetics* **132**:583-589.

410

411 Hillier WH et al. 2004. Sequence and comparative analysis of the chicken genome provide unique
412 perspectives on vertebrate evolution. *Nature*, **432**, 695-716.

413

414 Ilut DC, Nydam ML, Hare MP. 2014. Defining loci in restriction-based reduced representation
415 genomic data from nonmodel species: sources of bias and diagnostics for optimal clustering.
416 *BioMed Research International* **2014**:675158.

417

418 Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW. In press. Phylogenomics
419 of Phrynosomatid lizards: Conflicting signals from sequence capture versus restriction site
420 associated DNA sequencing. *Genome Biology and Evolution* doi:10.1093/gbe/evv026.

421

- 422 Leaché AD, Harris RB, Maliska ME, Linkem CW. 2013. Comparative species divergence across eight
423 triplets of spiny lizards (*Sceloporus*) using genomic sequence data. *Genome Biology and*
424 *Evolution* **5**:2410-2419.
- 425
- 426 Lewontin RC. 1974. *The genetic basis of evolutionary change*. New York: Columbia University Press.
- 427
- 428 Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES, Costich DE. 2013.
429 Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based
430 SNP discovery protocol. *PLoS Genetics* **9**:e1003215.
- 431
- 432 Lunter G, Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina
433 sequence reads. *Genome Research* **21**:936-939.
- 434
- 435 Maley JM. 2012. *Ecological speciation of King Rails (Rallus elegans) and Clapper Rails (Rallus*
436 *longirostris)*. D. Phil. Dissertation, Louisiana State University.
- 437
- 438 Maley JM, Brumfield RT. 2013. Mitochondrial and next-generation sequence data used to infer
439 phylogenetic relationships and species limits in the Clapper/King rail complex. *The Condor*
440 **115**:316-329.
- 441
- 442 Nybom H. 2004. Comparison of different nuclear DNA markers for estimating intraspecific genetic
443 diversity in plants. *Molecular Ecology* **13**:1143-1155.
- 444
- 445 Nylander JAA. 2004. MrAIC.pl. Available at: <http://www.abc.se/~nylander> (Accessed 12/20/2013)

446

447 Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language.

448 *Bioinformatics* **20**:289-290.

449

450 Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA. 2012. Genome-wide

451 association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology* **21**:2991-3005.

452

453 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an

454 inexpensive method for de novo SNP discovery and genotyping in model and non-model

455 species. *PloS One* **7**:e37135.

456

457 Pop M, Salzberg SL. 2008. Bioinformatics challenges of new sequencing technologies. *Trends in*

458 *Genetics*, **24**:142-149.

459

460 Python Software Foundation. 2007. Python version 2.7. Available at: <http://www.python.org> (Accessed

461 12/20/2013).

462

463 R Core Team. 2014. R: A language and environment for statistical computing. Available at:

464 <http://www.R-project.org/> (Accessed 12/20/2013)

465

466 Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM. 2013. Going where traditional

467 markers have not gone before: Utility of an promise for RAD sequencing in marine invertebrate

468 phylogeography and population genomics. *Molecular Ecology* **22**:2953-2970.

469

- 470 Remsen, JV Jr. 2003. Family Furnariidae (Ovenbirds). In: del Hoyo J, et al., eds. *Handbook of the*
471 *Birds of the World*. Barcelona: Lynx Edicions, 162-357.
- 472
- 473 Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L,
474 Faivre N, Loire E, Lourenco JM, Nabholz B, Roux C, Tsagkogeorga G, Weber AAT, Weinert
475 LA, Belkhir K, Bierne N, Glémin S, Galtier N. 2014. Comparative population genomics in
476 animals uncovers the determinants of genetic diversity. *Nature* **515**:261-263.
- 477
- 478 Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models.
479 *Bioinformatics* **19**:1572-1574.
- 480
- 481 Rubin BER, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PLoS One*
482 **7**:e33394.
- 483
- 484 Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2013. Target capture and massively
485 parallel sequencing of ultraconserved elements (UCEs) for comparative studies at shallow
486 evolutionary time scales. *Systematic Biology* **63**:83-95.
- 487
- 488 Smith BT, McCormack JE, Cuervo AM, Hickerson MJ, Aleixo A, Cadena CD, Pérez Eman JE, Burney
489 CW, Xie X, Harvey MG, Faircloth BC, Glenn TC, Derryberry EP, Prejean J, Fields S,
490 Brumfield RT. 2014. The drivers of tropical speciation. *Nature*, **515**, 406-409.
- 491
- 492 Sovic MG, Fries AC, Lisle Gibbs H. In press. AfrRAD: A pipeline for accurate and efficient de novo
493 assembly of RADseq data. *Molecular Ecology Resources* doi:10.1111/1755-0998.12378.

494

495 Taberlet P , Fumagalli L, Wust-Saucy AG, Cosson JF. 1998. Comparative phylogeography and
496 postglacial colonization routes in Europe. *Molecular Ecology*, **7**, 453-464.

497

498 Thornton K. 2003. libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*
499 **19**: 2325-2327.

500

501 Trapnell C, Salzberg SL. 2010. How to map billions of short reads onto genomes. *Nature*
502 *Biotechnology* **27**:455-457.

503

504 Yang Z, Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings*
505 *of the National Academy of Sciences* **107**:9264-9269.

506

507 White TA, Perkins SE, Heckel G, Searle JB. 2013. Adaptive evolution during an ongoing range
508 expansion: the invasive bank vole (*Myodes glareolus*) in Ireland. *Molecular Ecology* **22**:2971-
509 2985.

510

511 Zhang RM, Zhang DX, Zhu T, Yang Z. 2011. Evaluation of a Bayesian coalescent method of species
512 delimitation. *Systematic Biology* **60**:747-761.

513

514

515

516

517

518

519

520

521 **Table 1.** Attributes and summary statistics (SD) of datasets assembled under the similarity thresholds
 522 examined.
 523

	Similarity Threshold	Loci	Samples per Locus	Segregating Sites per Locus
<i>Cranioleuca</i>	99	147,123	8.28 (3.64)	0.20 (0.44)
	98	145,423	8.39 (3.64)	0.30 (0.63)
	97	144,475	8.42 (3.64)	0.34 (0.74)
	96	143,780	8.43 (3.63)	0.38 (0.86)
	95	142,897	8.45 (3.63)	0.41 (0.98)
	94	141,880	8.46 (3.63)	0.44 (1.11)
	93	140,801	8.47 (3.62)	0.48 (1.26)
<i>Rallus</i>	99	100,086	6.59 (2.62)	0.17 (0.41)
	98	99,300	6.61 (2.61)	0.24 (0.60)
	97	98,680	6.62 (2.61)	0.28 (0.73)
	96	98,206	6.62 (2.60)	0.30 (0.83)
	95	97,808	6.62 (2.60)	0.33 (0.93)
	94	97,321	6.62 (2.60)	0.36 (1.07)
	93	96,776	6.63 (2.59)	0.40 (1.22)
<i>Trochilus</i>	99	125,594	7.65 (3.34)	0.32 (0.56)
	98	125,966	7.74 (3.39)	0.46 (0.77)
	97	125,697	7.76 (3.39)	0.51 (0.87)
	96	125,437	7.76 (3.39)	0.54 (0.95)
	95	125,118	7.77 (3.39)	0.56 (1.02)
	94	124,669	7.78 (3.39)	0.59 (1.13)
	93	123,926	7.79 (3.39)	0.62 (1.25)
<i>Xenops</i>	99	155,933	7.54 (3.42)	0.65 (0.79)
	98	158,496	7.87 (3.47)	1.05 (1.17)
	97	158,281	7.99 (3.48)	1.25 (1.41)
	96	158,328	8.02 (3.48)	1.35 (1.56)
	95	158,078	8.03 (3.48)	1.40 (1.66)
	94	157,534	8.03 (3.48)	1.45 (1.76)
	93	156,640	8.05 (3.48)	1.50 (1.87)

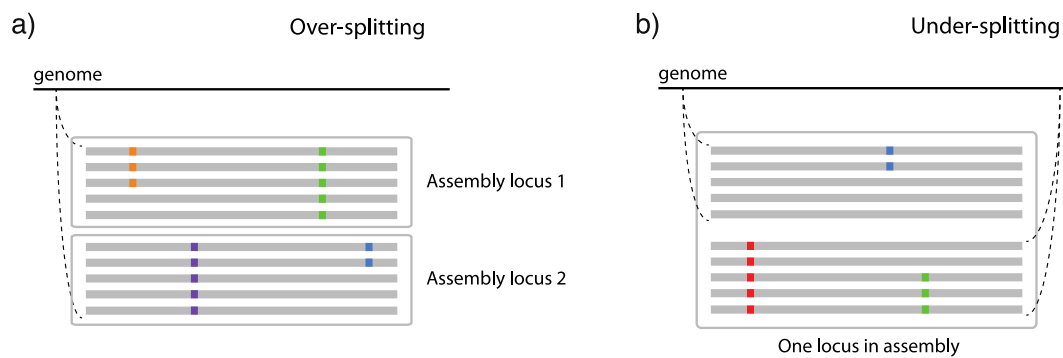
524

525

526

527
528
529
530
531
532

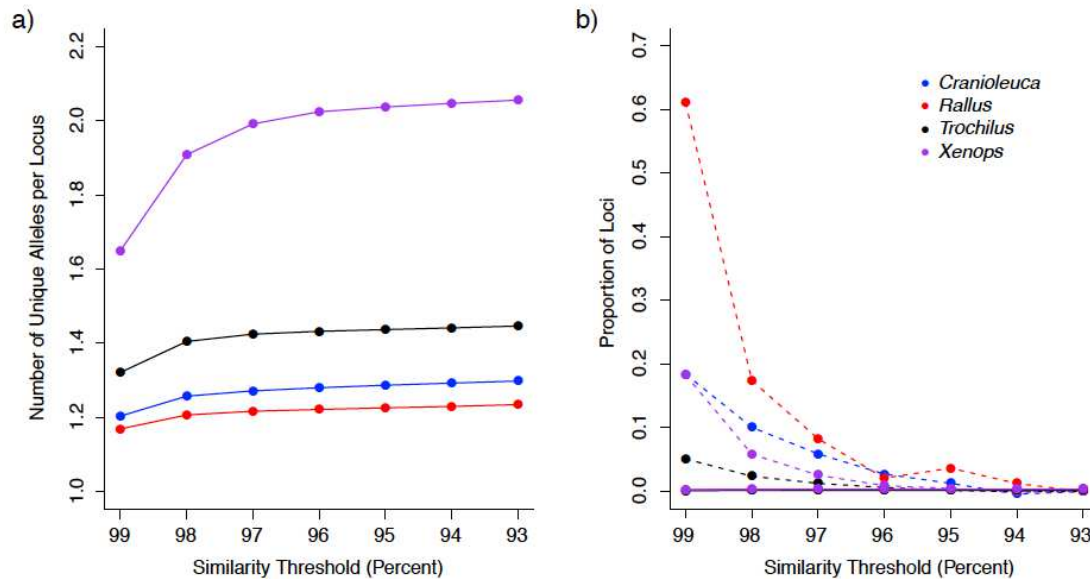
533 **Fig. 1.** Two ways in which similarity thresholds can result in spurious assemblies: (a) over-splitting
534 occurs when reads from different alleles from the same genomic position are spuriously split into
535 multiple loci due to lower similarity than the similarity threshold parameter, and (b) under-splitting
536 occurs when reads from different genomic positions are clustered into a single locus due to higher
537 similarity than the similarity threshold parameter. Gray bars represent identical sequence across reads,
538 whereas colored squares represent alternate alleles at SNPs.



539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562

563
564
565
566
567
568
569
570
571
572

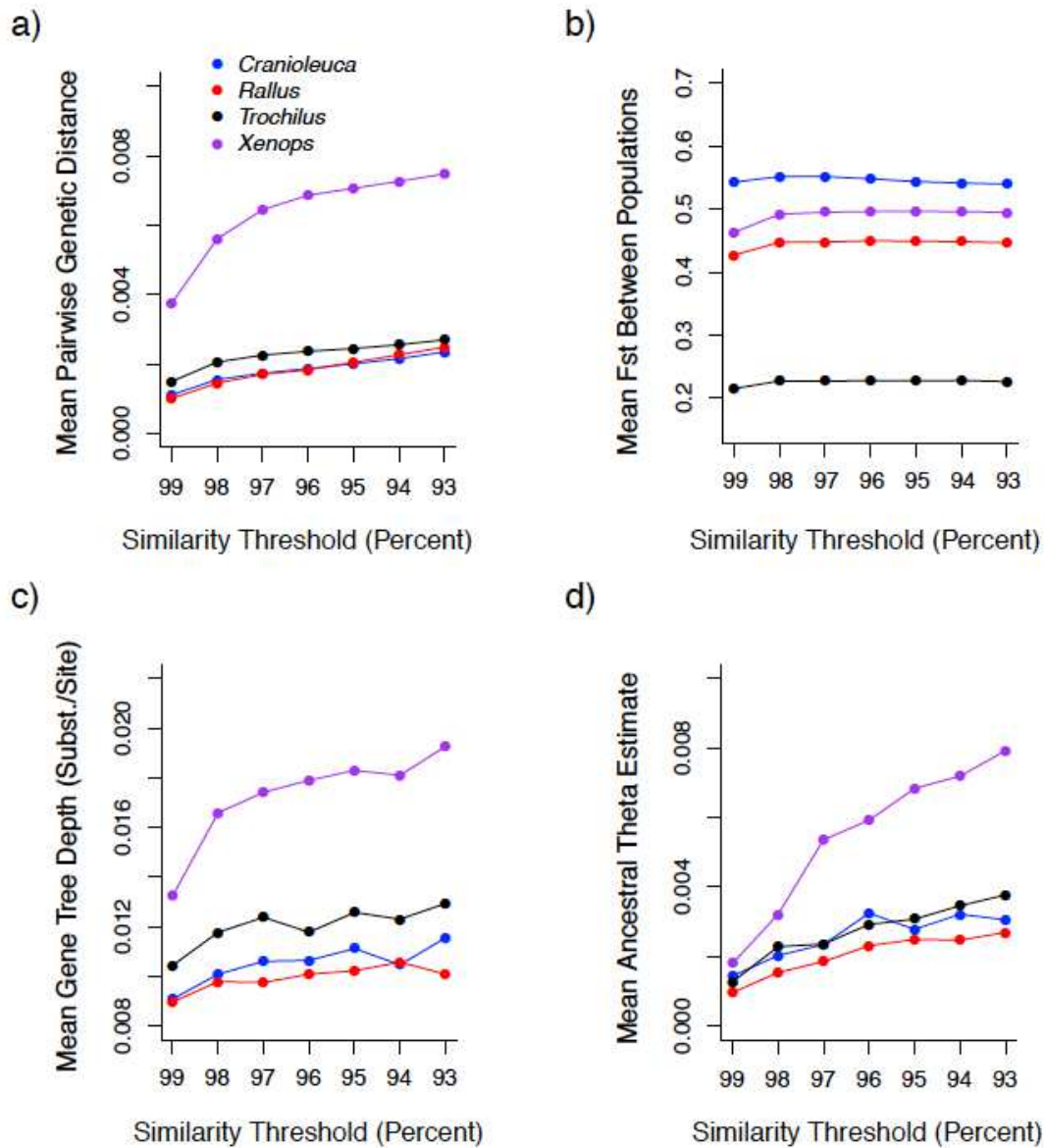
Fig. 2. The impact of similarity thresholds on empirical datasets from four bird lineages. (a) Stringent similarity thresholds resulted in fewer unique alleles per locus relative to more liberal thresholds. (b) Putative over-split loci (connected by dashed lines) were more frequent in datasets assembled at stringent similarity thresholds, whereas loci containing under-split reads (solid lines) occurred at low frequency across all similarity thresholds.



573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592

593
594
595
596
597
598
599

Fig. 3. The impact of similarity thresholds on population genetic parameter estimates of (a) mean pairwise genetic distance between individuals, (b) mean F_{ST} between populations, (c) mean gene tree depth and (d) ancestral theta based on a coalescent model.



600
601