

1 Current geography masks dynamic history of gene flow during speciation in
2 northern Australian birds

3
4 Joshua V. Peñalba^{1,2,3}, Leo Joseph^{2,3}, Craig Moritz^{1,2}
5
6
7

8 Author affiliations
9

10 ¹ Ecology & Evolution, Australian National University, Acton, ACT 2601, Australia

11 ² Centre for Biodiversity Analysis, Acton, ACT 2601, Australia

12 ³ Australian National Wildlife Collection, CSIRO National Research Collections Australia,
13 Canberra, GPO BOX 1700, Canberra, ACT 2601, Australia
14

15 Corresponding author: Joshua Penalba (josh.penalba@gmail.com)
16

17 Key words: geographic mode, allopatry, parapatry, snowballing, tipping point
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

35 **Abstract**

36 During early stages of speciation, genome divergence is greatly influenced by gene flow.
37 As populations diverge, geography can allow or restrict gene flow in the form of barriers.
38 Current geography, e.g. whether sister species are allopatric or parapatric, is often used to predict
39 the potential for gene flow during the divergence process. We test the validity of this assumption
40 in eight meliphagoid bird species codistributed across four regions. These regions are separated
41 by known biogeographic barriers within and between northern Australia and Papua New Guinea.
42 We find that bird populations across the same barrier have a range of divergence levels and
43 probability of gene flow regardless of range connectivity. Geographic distance and maximum
44 range connectivity over time can better predict divergence and probability of gene flow than
45 whether populations are currently allopatric or parapatric. We also find support for a nonlinear
46 decrease of the probability of gene flow during the divergence process. This implies that
47 although gene flow influences divergence early in speciation, other factors associated with
48 higher divergence restrict gene flow later in speciation. Current geography may then mislead
49 inferences regarding potential for gene flow during speciation under a complex and dynamic
50 history of geographic and reproductive isolation.

51 **Background**

52 Gene flow, selection, and genetic drift shape divergence during speciation, while
53 geography sets the stage on which these forces act [1–3]. The geographic mode of speciation is
54 defined by the extent of spatial isolation during early stages of divergence. Population pairs can
55 have disjoint (allopatric), completely overlapping (sympatric), or separate yet partially adjoining
56 (parapatric) ranges [4]. This geographic context predicts potential gene flow between
57 populations. Variation in levels of gene flow affects genetic differentiation, in turn affecting the

58 strength of selection and drift that is necessary to drive population divergence. Alternatively,
59 under a purely population genetic framework the geographic mode of speciation is defined by
60 levels of gene flow; “allopatry” when the proportion of the population which are migrants (m)
61 equals zero, “sympatry” when $m = 0.5$, and “parapatry” when $0 < m < 0.5$ per generation [5].

62 Although geographic and genetic definitions are often assumed to correspond, this is not
63 always the case in nature [6]. Current range distributions between sister species are often used to
64 infer their geographic mode of speciation, although the dynamic nature of species ranges through
65 evolutionary time can lead to unreliable predictions of gene flow [7]. Range fluctuations during
66 the Pleistocene’s climatic cycling resulted in multiple periods of connectivity and discontinuity
67 before resulting in the distribution we see today [8]. In order to understand how gene flow has
68 influenced divergence during speciation, we must first understand how the geographic history
69 could have shaped the potential for gene flow through time. The discrepancy between the spatial
70 and population genetic definitions of the geographic mode begs the question: does current
71 geography adequately predict realized gene flow and, consequently, divergence during early
72 stages of speciation?

73 Although geographic connectivity would allow for gene flow early in speciation, later in
74 the process geographic connectivity can be insufficient for gene flow as populations diverge and
75 become reproductively isolated. Populations further along in the speciation process would have
76 reduced realized gene flow due to intrinsic incompatibilities or extrinsic selection [9]. Especially
77 in birds, prezygotic isolation from sexual selection on song or plumage could influence gene
78 flow in later stages of speciation. There has been growing support for the “snowball” model of
79 accumulating incompatibility loci, initially proposed by Orr [10]. Qualitatively, this model and
80 those derived from it specify a nonlinear accumulation of incompatibility loci resulting in a short

81 speciation duration [11]. The “snowballing” has also been modeled under scenarios with
82 moderate to no gene flow (parapatry to allopatry) meaning different geographic modes may yield
83 a similar short duration of speciation, varying only in how long it takes for speciation to initiate
84 [12–16]. This rapid accumulation of isolation has also been proposed under models of divergent
85 selection in speciation-with-gene flow and under neutral models [17–20]. Though the underlying
86 assumptions of these theories may differ, the trajectory converges to a rapid transition,
87 ‘snowballing’, or ‘tipping point’ during speciation (simply referred to ‘snowballing’ from this
88 point forward). There is increasing empirical support for this pattern from studies of individual
89 systems [21–23] and a taxonomically broad meta-analysis [24]. More broad, comparative studies
90 across multiple systems would help elucidate this trajectory to speciation.

91 Further studies of population divergence in various geographic contexts could further
92 clarify the role of gene flow, selection and genome architecture in influencing the landscape of
93 divergence [9,20,25,26]. During speciation, local genomic variation in mutation and
94 recombination rates influence the rate at which regions diverge [25,27,28]. The geographic mode
95 of speciation could influence this landscape of divergence. Theory predicts that populations
96 diverging in parapatry should have a skewed distribution of divergence across the genome with a
97 few loci resistant to gene flow due to selection while the rest are free to move between
98 populations [29]. Populations diverging in allopatry, on the other hand, are predicted to lack this
99 skew as drift would be the predominant force influencing divergence [30,31]. This landscape
100 should also change as populations move further along the speciation continuum [9,27].

101 Here we investigate multiple bird species with populations codistributed in the same
102 region with known biogeographic barriers. Empirical studies of gene flow during divergence in
103 relation to geography often survey closely related populations and species across different

104 geographic regions with varying biogeographic histories and selection pressures [30,31].
105 However, to understand how shifting geographic ranges can erode the correspondence between
106 current geography and realized gene flow, it is more relevant to compare a set of taxa across
107 common geography. Our study region comprises part of the monsoonal tropics of northern
108 Australia and southern Papua New Guinea containing congruent biogeographic barriers for many
109 taxa including birds [32–34]. Sea level rise since the last glacial maximum has formed a barrier
110 between northern Australia and Papua New Guinea [35]. Meanwhile, the aridification of
111 mainland Australia has resulted in multiple semipermeable terrestrial barriers with parapatrically
112 or allopatrically distributed populations. Multiple studies have shown congruent phenotypic and
113 genetic breaks for various plant and animal species in this region [34,36–39]. Here we use one
114 gerygone and seven honeyeater species co-distributed in four focal regions: Papua New Guinea
115 (PNG), Northern Territory (NT), Cape York Peninsula (CYP) and eastern Queensland (QLD)
116 that are separated by well-known barriers (figure 1). These species were chosen as they have
117 already been shown to have variation in divergence levels across known barriers between CYP
118 and QLD [40] and have varying degrees of range connectivity [41] setting the stage to compare
119 divergence in different biogeographical contexts. With this system we ask (1) how current
120 geography, or potential for gene flow, predicts realized gene flow and genome divergence during
121 early stages of speciation, (2) how genome divergence influences realized gene flow in later
122 stages of speciation, (3) and how divergence varies across different loci during the speciation
123 process. Focusing on the codistributed ranges of these species will allow us to understand how
124 the genome differentiates during speciation under different geographic contexts within and
125 between closely related systems.

126 **Methods**

127 *Sampling*

128 We sampled eight species with populations that occupy the four regions of interest
129 (figure 1). For each population, we sampled three to six individuals for a total of 157 individuals
130 (15-20 individuals per species; electronic supplementary material, table S1). Samples were
131 chosen from locations farther away from known contact zones to avoid recent hybrids [42].
132 Brown honeyeater and white-throated honeyeater had insufficient sampling for PNG so only the
133 Australian populations were used in the analyses. We extracted the DNA from all individuals
134 using a standard salting-out procedure.

135 *Sequencing*

136 We sanger sequenced NADH dehydrogenase-2 (ND2) using the primers L5204 (5'
137 TAACTAAGCTATCGGGCGCAT 3') and H6312 (5'CTTATTTAAGGCTTTGAAGGCC 3')
138 for measures of mitochondrial divergence and structure [43]. For the nuclear loci, we used a
139 slightly modified version of the ddRADseq protocol as described by [44]. In brief, we digested
140 the DNA using the restriction enzymes *PstI* and *EcoRI* and size-selected 345 - 407bp.
141 Approximately ten indexed individuals in ten pools were sequenced on a NextSeq500 for 150bp,
142 single-read, mid-output and the rest were sequenced on another NextSeq500 lane with similar
143 specification except with high-output. A more detailed description of lab methods is available in
144 the electronic supplementary material.

145 *Data processing and analyses*

146 For each species we generated a reference set of RAD loci via the pyrad pipeline [45].
147 We used individuals from a different species to serve as an outgroup to polarize the SNPs in
148 downstream analyses. Only RAD loci which have associated outgroup sequences were retained
149 for the reference (electronic supplementary material, table S2). The resulting set of sequences

150 were used as a reference for further processing. Individual reads were then mapped onto this
151 reference set using Bowtie2 (v. 2.2.2)[46]. In order to recover larger numbers of loci and avoid
152 biases from filtering for loci with higher coverage, the mapped reads were further processed
153 using ngsTools and ANGSD to incorporate genotype likelihood information for the various
154 population genetic measures (v. 0.911; <http://www.popgen.dk/angsd>) [47].

155 *Population structure*

156 It has been demonstrated that more loci, even of lower coverage, yields more accurate
157 estimates of population genetic statistics compared to fewer loci of higher coverage when using
158 genotype likelihoods [48]. We used ANGSD to further filter for SNPs to be used for downstream
159 analyses (electronic supplementary material, table S3; see electronic supplementary material for
160 detailed ANGSD commands). We used a minimum coverage cut-off of 2X and a maximum cut-
161 off of 40X per individual to optimize the number of loci to be used while reducing the likelihood
162 of recovering paralogous loci. The maximum coverage cut-off was determined by plotting a
163 histogram of the average coverage per RAD locus and finding the upper threshold where most
164 loci fell under. To determine population structure, we randomly chose a single SNP per RAD
165 locus and reran ANGSD only for that set. We then used used ngsDist [49] to generate a distance
166 matrix which we used for a principal coordinates analysis (PCoA) using ‘cmdscale’ from base R
167 (v3.2.2) and a population network using SplitsTree (figure 1, electronic supplementary material
168 figure S1) [50].

169 *Population divergence statistics*

170 To calculate the various population divergence statistics (F_{ST} , D_{XY} , and D_A) we used the
171 software within the ngsTools package. These statistics were used as they provide both relative
172 and absolute measures of genetic divergence and can be compared to previous studies [24,30,31].

173 We used all SNPs within each RAD locus for all per locus and global measures. To calculate
174 pairwise F_{ST} , we used realSFS on the ANGSD genotype likelihoods to estimate an unfolded 2D
175 site frequency spectrum (SFS) per population pair and used the SFS to derive the per locus F_{ST}
176 estimate [51,52]. We used the same outgroups as the *pyrad* filtering steps for the unfolded SFS.
177 To calculate global F_{ST} , we used the estimate of allele frequencies within each population and all
178 populations pooled together $F_{ST} = H_T - H_S / H_T$ where H_T is the heterozygosity of all populations
179 pooled and $H_S = \Sigma H_e / k$ and $H_e = 1 - \Sigma(p^2 + (1-p)^2)/m$ where k is the number of populations, m
180 is the number of loci, and p is the allele frequency [53]. To calculate D_{XY} , we used the estimate
181 of allele frequencies which incorporated the genotype likelihood from ANGSD [54]. To
182 calculate D_A , we used the π estimates from ANGSD for each population and the equation $D_A =$
183 $D_{XY} - (\pi_X + \pi_Y)/2$. Lastly, we used the package *ape* v. 4.1 to calculate the ND2 genetic p-distance
184 under the Jukes-Cantor model [55].

185 *Estimating the likelihood of migration*

186 We used an approximate Bayesian computation (ABC) model selection to estimate the
187 likelihood of migration (i.e. how strong the evidence is for some gene flow during the divergence
188 process) between each population pair . The ABC analyses and models followed that of Roux et
189 al. but using the unfolded 2DSFS as a summary statistic instead of the various population genetic
190 statistics used in their study [24]. We tested the models of isolation-with-migration (IM),
191 isolation-with-migration + heterogeneous N_e (IMhetN), isolation-with-migration +
192 heterogeneous migration (IMhetM), isolation-with-migration + heterogeneous N_e +
193 heterogeneous migration (IMhetNhetM), strict isolation (SI), and strict isolation + heterogeneous
194 N_e (SIhetN). Heterogeneous N_e is modeled to reflect variation in recombination rate throughout
195 the genome and heterogeneous migration is to reflect variation in gene flow across loci between

196 hybridizing populations. Models of SI preceding IM (secondary contact) and IM preceding SI
197 were excluded as they are typically difficult to distinguish from IM models [24]. We then used
198 the R package abc v. 2.1 and calculated the likelihood of each model using a neural network with
199 50 trained and 6 hidden networks [56]. We ran the abc analyses five times and used the average
200 model support of the replicates for further analyses. As in Roux et al., we used the sum of the
201 support for models containing a migration parameter as our probability for migration [24].
202 Additional details of the analyses can be found in the supplementary material.

203 To infer the relationship between divergence and realized gene flow, we correlated F_{ST} to
204 the probability for migration between each population pair. To negate possible circularity where
205 model support may be determined by F_{ST} itself, we calculated the F_{ST} of simulations under the SI
206 model and reran the ABC model selection on a range of F_{ST} values. If the ABC model selection
207 consistently recovers a low probability of migration, regardless of F_{ST} , high support for IM for
208 populations with lower F_{ST} in the empirical data would more likely be a biological phenomenon
209 rather than a model selection artifact.

210 *Speciation model fitting*

211 We fit our divergence versus probability of migration distribution to test support for
212 various theoretical trajectories of parapatric speciation presented by Yamaguchi and Iwasa [15].
213 The models include (1) ‘threshold’: where full incompatibility is reached after a certain
214 divergence level, (2) ‘constant’ rate of divergence increase, (3) ‘accelerated’ where increase in
215 divergence is small until a certain divergence threshold is reached and increase is accelerated, (4)
216 ‘decelerated’ increase where divergence accumulates quickly but slows down as it approaches
217 full incompatibility, and (5) ‘sigmoid’ where the rate of increase starts slow but accelerates after
218 a certain threshold before decelerating again prior to reaching complete incompatibility

219 suggesting a snowballing during the speciation process (electronic supplementary material, table
220 S4, figure S2). Our global F_{ST} parallels the ‘incompatibility genetic distance’ of Yamaguchi and
221 Iwasa [15]. The inverse of our probability of migration parallels their ‘incompatibility’ where no
222 support for migration ($P(m) = 0$) is equivalent to one or full support for incompatibility ($I(z) =$
223 1). Identical to F_{ST} and probability of migration, the measures Yamaguchi and Iwasa used are
224 bounded by zero and one. We used the resulting probability of migration as the summary statistic
225 and simulated 500k distributions under each speciation trajectory model and used ABC and
226 model rejection to estimate model support for our data. To test robustness of our ABC model
227 inference, we used 200 simulated datasets under each model to see if the correct model is
228 recovered (electronic supplementary material, table S5). Details regarding this method can be
229 found in the supplementary material.

230 *Species distribution modeling*

231 To infer how geographic distributions have changed through time we estimated
232 geographic connectivity over space through time using species distribution modeling and least
233 cost path analyses. We used vouchered specimen data from the Atlas of Living Australia
234 (<http://www.ala.org.au>) for occurrence points and environmental variables from WorldClim v.
235 1.4 (Community Climate System Model 4; <http://worldclim.org/paleo-climate1>) to predict
236 species ranges under past climates. To model species distributions in R v. 3.2.2, we followed
237 guidelines described in [57]. We used the R package *dismo* v. 1.1-1 for the maximum entropy
238 (MAXENT) analyses to predict species ranges [57]. We ran MAXENT using environmental
239 layers from the present, mid-Holocene and the LGM (electronic supplementary material, table
240 S6). Lastly, we selected a single coordinate (midpoint of the range) for each of the four
241 populations and calculated the geographic distance between those points to account for isolation-

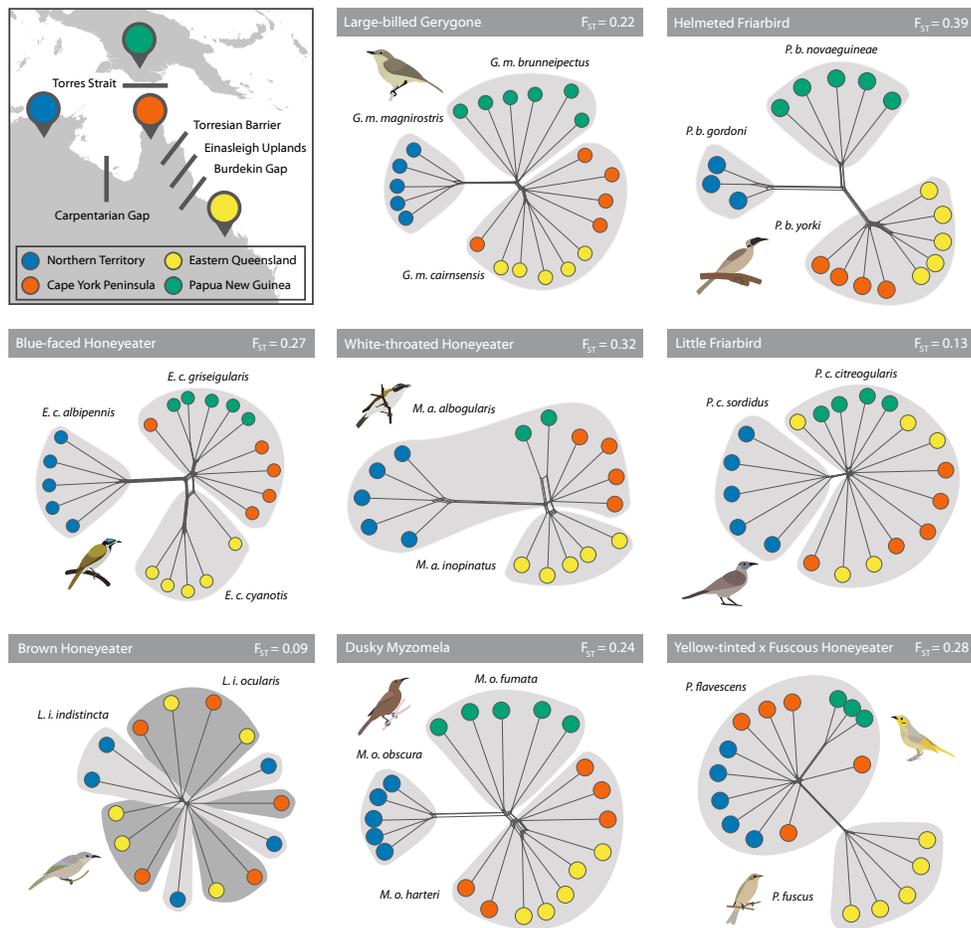
242 by-distance. Between population pairs, we also calculated the least cost path using the R package
243 gdistance (v. 1.1-9) based on modeled suitability [58] in the current range, mid-Holocene, and
244 LGM predictions. For each population pair, we chose the minimum resistance path between all
245 three time points to quantify the highest opportunity for migration through time. Habitat
246 resistance values can be found in the electronic supplementary material table S7 and species
247 distribution predictions can be found in figure S3.

248 **Results**

249 *Population structure and divergence*

250 Population structure varied between each taxon-pair though most had some form of
251 clustering across the four populations (figure 1). Of the eight species, five showed distinct
252 clusters for all four populations, one showed only three distinct clusters, one showed only two
253 distinct clusters, and the last showed no population clusters (F_{ST} 0.09 - 0.39). The degrees of
254 clustering and the population relationships are fairly variable, as shown by a PCoA and
255 population network of the genetic distances (electronic supplementary material, figure S1).
256 Currently allopatric populations are not more likely to be separate clusters, exemplified by
257 samples from the PNG population often being closer to CYP and QLD. Similarly, geographically
258 parapatric populations are not necessarily mixed as shown by NT often being the most
259 diverged and consistently separate cluster even for species with currently continuous geographic
260 ranges to CYP and QLD. Mitochondrial haplotype networks generally corroborated nuclear
261 SNP structure. Measures of genetic differentiation and divergence also varied. Autosomal and Z
262 chromosome divergences were consistently correlated with all divergence measures (electronic
263 supplementary material, figure S4, tables S8 and S9). Relative divergence (F_{ST}) and gene flow
264 scales with mitochondrial divergence with a few outliers. The transition to low probability of

265 migration is fairly rapid beyond 1% ND2 p-distance. D_{XY} , an absolute measure of divergence,
 266 also scales with F_{ST} but with more outliers compared to ND2 (electronic supplementary material,
 267 figure S5). Though different measures of divergence in different DNA classes generally
 268 correlated as expected, divergence and structure of populations varied between species with no
 269 immediate patterns corresponding to geography.

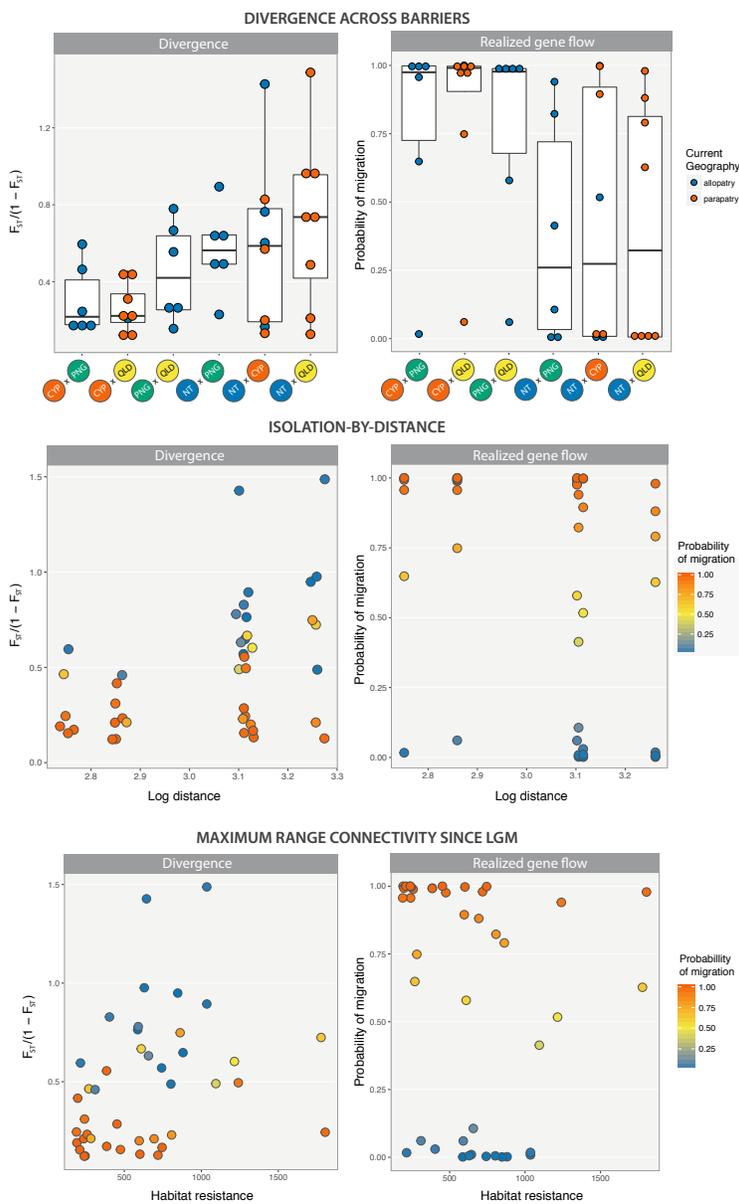


270
 271 **Figure 1.** Geographic regions that were sampled for this study with corresponding known
 272 biogeographical barriers. Sample networks were generated from a distance matrix estimated
 273 from the ddRAD genotype likelihoods. Within each system, the species and subspecies
 274 designation has been highlighted. F_{ST} represents global measures across all four populations.

275

276 *Geography and genome divergence*

277 Relative genome divergence (F_{ST}) and realized gene flow (probability of migration) is not
278 predicted by between the current geographic definitions of allopatry and parapatry: allopatry
279 between PNG and mainland Australia and either allopatry or parapatry within mainland
280 Australia, respectively (electronic supplementary material, figure 2, figure S3, top; Kruskal-
281 Wallis chi-squared=2.58e-03, df = 1, p = 0.9595). Population pairs exhibit a range of divergence
282 levels regardless of whether the current barrier is terrestrial or marine. Divergence due to
283 isolation-by-distance is well supported by a correlation between the adjusted F_{ST} and log of the
284 distance in km (Spearman's rank correlation: rho = 0.427936, p = 0.004698). Correspondingly,
285 the probability of migration also decreases with increasing distance (Spearman's rank
286 correlation: rho = -0.482106, p = 0.001225). Lastly, we compared the minimum landscape
287 resistance (ie. maximum range connectivity or maximum potential for gene flow corrected for
288 distance) between the three historic time points to F_{ST} for each population pair. All comparisons
289 with PNG had the LGM as the period where PNG was connected with north Australia (18 pairs).
290 The timing of maximum connectivity within mainland Australia varied: eight pairs in present-
291 day, seven during the mid-Holocene, and nine in the LGM (electronic supplementary material,
292 table S7, figure S3). There is a positive correlation between landscape resistance and F_{ST}
293 (Spearman's rank correlation test, S = 7048.6, rho = 0.4288, p = 0.004601), which translates to a
294 negative correlation between resistance and probability of gene flow (Spearman's rank
295 correlation test, S = 18102, rho = -0.466829, p = 0.00183). In sum, whereas classification of
296 allopatry and parapatry based on predictions of current distribution do not predict divergence or
297 probability of gene flow, other factors such as geographic distance or past connectivity proves to
298 be more appropriate predictors.



299
 300 **Figure 2.** *Top:* Relative divergence through the barriers colored by whether or not population
 301 pairs have disjunct ranges. Allopatry and parapatry are defined based on connectivity in the
 302 species distribution models. The barriers are ordered based on increasing geographic distance
 303 *Middle:* Adjusted F_{ST} values plotted against log of distance to show isolation-by-distance. Color
 304 gradient depicts the relative support for the models with gene flow. Points have been jittered by

305 0.03 to display the number of points. *Bottom*: Divergence plotted against the highest connectivity
306 across three time points (present, mid-Holocene, LGM).

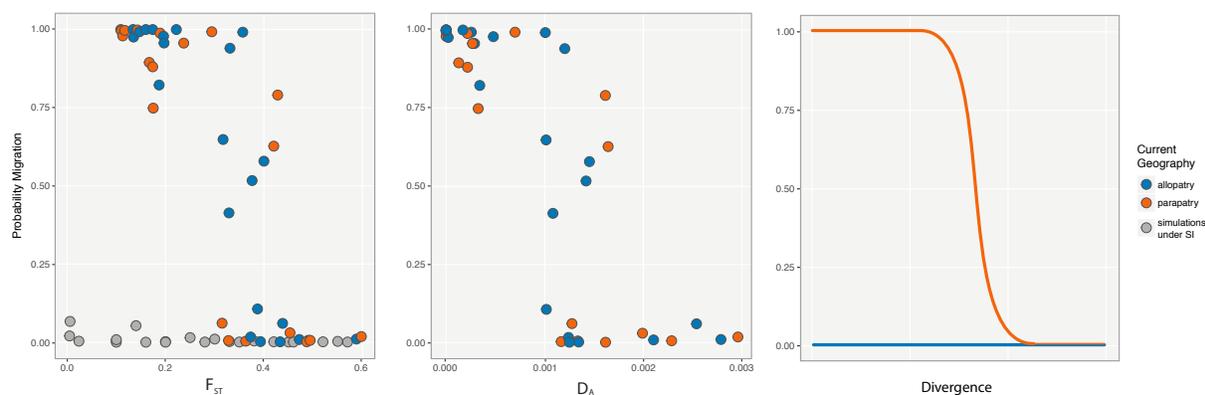
307

308 Although the patterns described above also apply to the Z chromosome, the Z
309 chromosome tends to have higher differentiation relative to the autosome. The difference
310 between the autosomal F_{ST} value and Z chromosome F_{ST} value increases as probability of gene
311 flow decreases (Spearman rank correlation: $\rho = -0.5504$, $p = 1.58e-4$) but does not correlate
312 with current population connectivity (Kruskal-Wallis rank sum test: $\chi^2 = 0.12665$, $df =$
313 1 , $p = 0.7219$). This means that the amount of difference between autosomal and Z chromosome
314 divergence is a function of overall divergence rather than spatial context. The difference between
315 autosomal and Z chromosome values increases with increasing divergence for both relative (F_{ST})
316 and absolute divergence (D_{XY}), though absolute divergence far less so (Spearman rank
317 correlation F_{ST} : $\rho = 0.715672$, $p = 1.001e-7$; Spearman rank correlation D_{XY} : $\rho = 0.285595$, p
318 $= 0.06674$; electronic supplementary material, figure S6). In this system, broad sampling of
319 population pairs reveal that difference between autosomal and Z chromosome divergence is
320 likely simply related to level of divergence rather than population connectivity.

321 *Genome divergence and speciation*

322 Designations of allopatry or parapatry in current distributions do not predict realized gene
323 flow, even for the less diverged populations. However, our data suggests that relative divergence
324 level influences realized gene flow throughout the entire range of divergence. We see a rapid
325 transition from high gene flow with low divergence to low gene flow with higher divergence
326 through a narrow range of divergence levels; similar to the model of snowballing during
327 speciation (figure 3). Support values for the different demographic models can be found in the

328 electronic supplementary material table S10. A subset of simulations under the strict isolation
329 model was always recovered to have low probability of migration regardless of F_{ST} value,
330 suggesting that the trajectory we see in our data is not likely due to artifacts in the model
331 selection (figure 3).

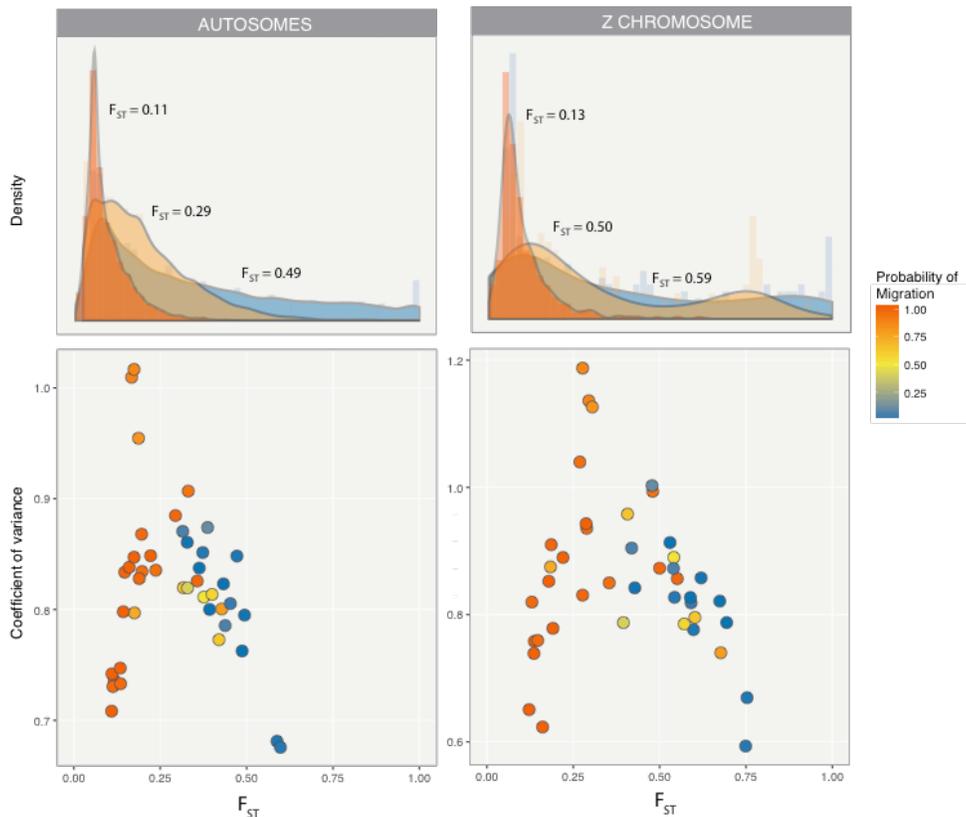


332
333 **Figure 3.** Probability of migration defined by the sum of ABC model supports for the 4 models
334 with a migration parameter (IM, IMhetM, IMhetN, IMhetNhetM) plotted against F_{ST} and D_A .
335 The third panel is a plot of our expectation of the change in probability of migration under
336 allopatry and parapatry.

337 After comparing our data to models proposed by Yamaguchi and Iwasa [15], all
338 simulations retained by the ABC model selection were those simulated under the sigmoidal
339 function representative of a snowballing effect (electronic supplementary material, figure S2).
340 There is no support that our data follow any other proposed trajectories. The F_{ST} range
341 corresponding to the tipping point spans $\sim 0.3 - 0.4$. The D_A range corresponding to the tipping
342 point is $\sim 0.1-0.17\%$. Species-specific points often span the entire range of the trajectory
343 (electronic supplementary material, figure S7). Plots for the Z chromosome, D_{XY} , and ND2
344 against probability of migration can be found in the electronic supplementary material, figure S8.
345 Our system provides additional empirical support for a snowballing pattern in speciation.

346 *Genome-wide divergence*

347 The coefficient of variance describes the distribution of the individual F_{ST} values across
348 the RAD loci. Higher coefficient of variance corresponds to a more skewed distribution - i.e. one
349 in which there is higher heterogeneity in levels of divergence across loci. At lower F_{ST} values
350 (and high gene flow) there is a lower coefficient of variance as most values are close to zero. The
351 coefficient of variance increases with increasing F_{ST} but peaks at intermediate levels of migration
352 and starts to decrease again when the support for migration decreases. The change in the
353 distribution of F_{ST} values follow a predictable pattern with increasing divergence where there is
354 an initial skew from low to moderate divergence levels followed by a more uniform distribution
355 from moderate to high divergence levels (figure 4). This may be due to similarities in genome
356 architecture rather than differences in current geographic classification.



357

358 **Figure 4.** *Top:* Density of F_{ST} distributions of the ddRAD loci with increasing global F_{ST} and
359 decreasing probability of migration. Three representative population pairs were chosen to
360 represent different divergence levels. Low: Brown honeyeater NT & QLD, Medium: Blue-faced
361 honeyeater CYP & QLD, high: White-throated honeyeater NT & QLD. *Bottom:* Distribution of
362 the skew of F_{ST} distributions with increasing global F_{ST} and decreasing probability of migration.

363 **Discussion**

364 Geographic mode of speciation is expected to influence the roles of gene flow, selection,
365 and drift on divergence. However, our data suggests that although currently continuous
366 geographic ranges should have higher potential for gene flow relative to discontinuous ranges, it
367 does not necessarily translate to a higher probability of migration during divergence. The degree
368 of connectivity between populations is dynamic through their evolutionary history and this
369 results in reticulation of the genomes between those populations (figure 1, electronic
370 supplementary material, figure S3)[59]. Our system supports the idea that current classifications
371 of geography (allopatric, parapatric, and sympatric) are not reflective of gene flow during
372 population divergence [6,7]. It is possible that some currently allopatric populations have high
373 probability of gene flow reflecting past connectivity. Conversely, some currently parapatric
374 populations may have low probability of gene flow from incompatibilities accumulated during
375 past allopatry. Instead, there is support for geographic distance and minimum resistance being
376 negatively correlated with realized gene flow and positively correlated with divergence. Properly
377 classifying the geographic mode of speciation is particularly pertinent to a dynamic region like
378 northern Australia and PNG. Despite this correlation with distance and historical connectivity,
379 population divergence and realized gene flow of these birds across various barriers in northern
380 Australia and Papua New Guinea span the range of the speciation continuum potentially due to

381 lineage-specific biogeographic history or natural history (figure 3). Accumulating biogeographic
382 studies across suture zones or shared biogeographical barriers show similar dynamics of
383 population divergence in other avian systems as well as lizards and invertebrates [60–62].

384 Our data provides another empirical example in support for a snowballing pattern during
385 speciation. There exists a certain threshold of divergence where there is a rapid transition
386 between high and low gene flow likely resulting in speciation. Though populations at
387 intermediate stages are still observed, they persist at a narrow range of divergence levels. Our
388 data shows a similar rapid transition between the states of high and low likelihood for gene flow
389 (figure 3). This transition occurs at a narrow stage of nuclear and mitochondrial divergence; F_{ST}
390 $\sim 0.3-0.4$, $D_A \sim 0.1-0.17\%$, and ND2 p-distances of $\sim 1-1.5\%$. Using the substitution rate from
391 Pacheco et al. (2011), the ND2 p-distances would translate to 1.11 - 1.67 Mya. The range of D_A
392 where transition occurs is far lower to that found in Roux et al. ($\sim 0.5-2\%$) which may be a
393 characteristic specific to avian speciation. Additional studies in other avian systems would help
394 determine if these patterns in divergence are robust.

395 Unlike the nuclear loci, there are outliers in the mitochondrial data where populations
396 with high mtDNA divergence would still have high likelihood of gene flow and low F_{ST}
397 (electronic supplementary material, figure S4). These populations are from the dusky myzomela
398 (CYP & QLD, CYP & PNG, and PNG & QLD) and from the white-throated honeyeater (CYP &
399 QLD). The dusky myzomela population pairs are part of the small subset which are all allopatric
400 on mainland Australia. One possible explanation for the white-throated honeyeater is that gene
401 flow after secondary contact has homogenized the nuclear genome but maintained the local ND2
402 haplotypes in accordance with Haldane's rule [63].

403 Although neither ecological selection nor incompatibility loci were explicitly tested, this
404 sigmoidal trajectory of speciation is consistent with the parapatric models (divergence-with-gene
405 flow) proposed by Yamaguchi and Iwasa [15]. As suggested by the dynamic geographic history,
406 population pairs likely experienced varying rates of gene flow through time [8,59]. On the other
407 hand, it has also been shown that a nonlinear accumulation of divergence can occur under certain
408 neutral scenarios [20]. Ideally, estimates of population splitting time would inform us about the
409 timing and duration of speciation and therefore the rate of accumulating divergence; however,
410 we found that we could not use ddRAD to infer divergence times reliably [64].

411 Within the genome, there is variation of divergence across loci owing to various degrees
412 of standing genetic variation and recombination [25–27]. The change in the distribution of F_{ST}
413 values across loci is a coarse estimate of the change in landscape of divergence through time.
414 Although the L-shape or the skew of F_{ST} distributions between parapatrically diverging
415 populations have been attributed to a few loci under divergent selection with the rest
416 homogenized by gene flow [30,31], our broader sampling of population pairs show that the
417 change in F_{ST} distribution follow a predictable pattern with increasing F_{ST} or decreasing
418 probability of gene flow regardless of connectivity. It is also likely that this pattern of
419 accumulation of divergence is due to variation in nucleotide diversity across the loci [20,25]. The
420 change in skew with increasing F_{ST} could also be driven by linked selection instead of resistance
421 to gene flow as recent studies have shown [27,65]. It would be important to complement this
422 study with detailed studies across hybrid or suture zones to disentangle the role of gene flow and
423 linked selection on divergence between hybridizing populations.

424 Finally, we note some important points for the study of the geographic mode of
425 speciation. First, neither range overlap nor migration rate is static during speciation and it is

426 more likely that populations experienced various degrees of geographic connectivity and gene
427 flow through time [59]. This is particularly pertinent for highly vagile taxa, like birds, in highly
428 dynamic geographic regions, like the Australopapuan region. Additionally, populations that are
429 currently allopatric due to vicariance could have experienced periods of reduced but ongoing
430 gene flow during the formation of the barrier [66]. Second, population differentiation would also
431 vary if a population has accumulated isolating mechanisms during allopatry and accelerated
432 divergence during secondary contact (alloparapatry) [2]. Depending on the particular
433 demographic history, isolation with continuous migration (primary divergence) can be difficult
434 to distinguish from gene flow after secondary contact from genetic data alone [24]. Third,
435 population divergence in allopatry may not necessarily translate to speciation. Seeing that most
436 species definitions rely on degree of reproductive isolation there are no consistent criteria to
437 differentiate discontinuous populations with no gene flow from allopatric species [1,67,68]. The
438 combination of reduced gene flow, either completely or partially, and genetic drift may result in
439 population divergence but other factors likely play a more important role in speciation.

440 ***Conclusion***

441 Our comparative study of divergence of bird populations through the speciation process
442 highlights the dynamics of geographic history and their influence on divergence. Further, it
443 provides additional support for a snowballing pattern in speciation, and it characterizes broad
444 patterns of genomic divergence through time. The divergence that we discuss in this paper is
445 presumed to be neutral and therefore would benefit from a replicate study looking at divergence
446 in coding regions to observe whether different marker types have different trajectories in the
447 same populations. The results of this broad study also clearly describe the pattern of
448 accumulation of divergence which lend support to the emerging relevance of linked selection in

449 genome divergence. Comparative studies, particularly with multiple species in a shared
450 geographic region, help elucidate patterns of genome divergence during speciation. To fully
451 comprehend the patterns of neutral and adaptive genomic divergence, we need to sample broadly
452 both phylogenetically and geographically thus affecting shared patterns in speciation
453 differentiated from the exceptions.

454 **Author contributions**

455 JVP conceived ideas and designed the project, collected the data, performed the analyses, and led
456 the writing of the manuscript. LJ and CM provided ideas during the analysis and writing of the
457 manuscript

458 **Acknowledgments**

459 We would like to thank Ian J. Mason and Alex Drew for greatly appreciated help with collecting
460 samples in the field, Alexander Xue for assistance with demographic analyses, Matteo Fumigalli
461 for assistance with ngsTools, and Daniel Rosauer for assistance with species distribution
462 modelling. We would also like to thank Sonal Singhal, Sally Potter, and Daniel R. Wait for
463 helpful discussions and comments on the manuscript. Lastly all data collection was carried out in
464 Australian National University's Biomolecular Research Facility and most analyses were carried
465 out in the ABC Bioinformatics Development Cluster.

466 **Funding**

467 This research was funded by BirdLife Stuart Leslie Bird Research Award 2015.

468 **References**

- 469 1. Mayr E. 1942 *Systematics and the Origin of Species, from the Viewpoint of a Zoologist*.
470 Harvard University Press.
- 471 2. Coyne JA, Orr HA, Others. 2004 *Speciation*. Sinauer Associates Sunderland, MA.
- 472 3. White MJD. 1978 *Modes of speciation*. San Francisco: WH Freeman 455p.-Illus., maps,

- 473 chrom. nos.. General (KR, 197800185).
- 474 4. Mayr E. 1963 Animal species and evolution. *Animal species and evolution*.
- 475 5. Gavrilets S. 2003 Perspective: models of speciation: what have we learned in 40 years?
476 *Evolution* **57**, 2197–2215.
- 477 6. Butlin RK, Galindo J, Grahame JW. 2008 Sympatric, parapatric or allopatric: the most
478 important way to classify speciation? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 2997–
479 3007.
- 480 7. Losos JB, Glor RE. 2003 Phylogenetic comparative methods and the geography of
481 speciation. *Trends Ecol. Evol.* **18**, 220–227.
- 482 8. Hofreiter M, Stewart J. 2009 Ecological change, range fluctuations and population
483 dynamics during the Pleistocene. *Curr. Biol.* **19**, R584–94.
- 484 9. Seehausen O *et al.* 2014 Genomics and the origin of species. *Nat. Rev. Genet.* **15**, 176–192.
- 485 10. Orr HA. 1995 The population genetics of speciation: the evolution of hybrid
486 incompatibilities. *Genetics*
- 487 11. Welch JJ. 2004 Accumulating Dobzhansky-Muller incompatibilities: reconciling theory and
488 data. *Evolution* **58**, 1145–1156.
- 489 12. Orr HA, Turelli M. 2001 The evolution of postzygotic isolation: accumulating Dobzhansky-
490 Muller incompatibilities. *Evolution* **55**, 1085–1094.
- 491 13. Gavrilets S. 2000 Waiting time to parapatric speciation. *Proc. Biol. Sci.* **267**, 2483–2492.
- 492 14. Gavrilets S. 2014 Models of Speciation: Where Are We Now? *J. Hered.* **105**, 743–755.
- 493 15. Yamaguchi R, Iwasa Y. 2017 A tipping point in parapatric speciation. *J. Theor. Biol.* **421**,
494 81–92.
- 495 16. Yamaguchi R, Iwasa Y. 2013 First passage time to allopatric speciation. *Interface Focus* **3**,
496 20130026.
- 497 17. Via S. 2012 Divergence hitchhiking and the spread of genomic isolation during ecological
498 speciation-with-gene-flow. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 451–460.
- 499 18. Feder JL, Egan SP, Nosil P. 2012 The genomics of speciation-with-gene-flow. *Trends*
500 *Genet.* **28**, 342–350.
- 501 19. Feder JL, Nosil P, Wacholder AC, Egan SP, Berlocher SH, Flaxman SM. 2014 Genome-
502 Wide Congealing and Rapid Transitions across the Speciation Continuum during Speciation
503 with Gene Flow. *J. Hered.* **105**, 810–820.
- 504 20. Southcott L, Kronforst MR. In press. A neutral view of the evolving genomic architecture

- 505 of speciation. *Ecol. Evol.* (doi:10.1002/ece3.3190)
- 506 21. Matute DR, Butler IA, Turissini DA, Coyne JA. 2010 A Test of the Snowball Theory for
507 the Rate of Evolution of Hybrid Incompatibilities. *Science* **329**, 1518–1521.
- 508 22. Moyle LC, Nakazato T. 2010 Hybrid incompatibility ‘snowballs’ between *Solanum* species.
509 *Science* **329**, 1521–1523.
- 510 23. Riesch R *et al.* 2017 Transitions between phases of genomic differentiation during stick-
511 insect speciation. *Nature Ecology & Evolution* **1**, 0082.
- 512 24. Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. 2016 Shedding Light on
513 the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLoS Biol.* **14**,
514 e2000234.
- 515 25. Cruickshank TE, Hahn MW. 2014 Reanalysis suggests that genomic islands of speciation
516 are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157.
- 517 26. Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, Noor MAF, Mehlig B,
518 Westram AM. 2017 Interpreting the genomic landscape of speciation: a road map for
519 finding barriers to gene flow. *J. Evol. Biol.* **30**, 1450–1477.
- 520 27. Burri R *et al.* 2015 Linked selection and recombination rate variation drive the evolution of
521 the genomic landscape of differentiation across the speciation continuum of *Ficedula*
522 flycatchers. *Genome Res.* **25**, 1656–1665.
- 523 28. Ellegren H *et al.* 2012 The genomic landscape of species divergence in *Ficedula*
524 flycatchers. *Nature* **491**, 756–760.
- 525 29. Martinsen GD, Whitham TG, Turek RJ, Keim P. 2001 Hybrid populations selectively filter
526 gene introgression between species. *Evolution* **55**, 1325–1335.
- 527 30. Martin SH *et al.* 2013 Genome-wide evidence for speciation with gene flow in *Heliconius*
528 butterflies. *Genome Res.* (doi:10.1101/gr.159426.113)
- 529 31. Nosil P, Gompert Z, Farkas TE, Comeault AA, Feder JL, Buerkle CA, Parchman TL. 2012
530 Genomic consequences of multiple speciation processes in a stick insect. *Proc. Biol. Sci.*
531 **279**, 5058–5065.
- 532 32. Edwards RD, Crisp MD, Cook DH, Cook LG. 2017 Congruent biogeographical
533 disjunctions at a continent-wide scale: Quantifying and clarifying the role of biogeographic
534 barriers in the Australian tropics. *PLoS One* **12**, e0174812.
- 535 33. Cracraft J. 1986 Origin and evolution of continental biotas: speciation and historical
536 congruence within the Australian avifauna. *Evolution* **40**, 977–996.
- 537 34. Bowman DMJS *et al.* 2010 Biogeography of the Australian monsoon tropics. *J. Biogeogr.*
538 **37**, 201–216.

- 539 35. Lambeck K, Chappell J. 2001 Sea Level Change Through the Last Glacial Cycle. *Science*
540 **292**, 679–686.
- 541 36. Baldassarre DT, White TA, Karubian J, Webster MS. 2014 Genomic and morphological
542 analysis of a semipermeable avian hybrid zone suggests asymmetrical introgression of a
543 sexual signal. *Evolution* **68**, 2644–2657.
- 544 37. Catullo RA, Lanfear R, Doughty P, Keogh JS. 2014 The biogeographical boundaries of
545 northern Australia: evidence from ecological niche models and a multi-locus phylogeny of
546 *Uperoleia* toadlets (Anura: Myobatrachidae). *J. Biogeogr.* **41**, 659–672.
- 547 38. Edwards SV, Potter S, Schmitt CJ. 2016 Reticulation, divergence, and the phylogeography–
548 phylogenetics continuum. *Proceedings of the*
- 549 39. Kearns AM, Joseph L, Omland KE, Cook LG. 2011 Testing the effect of transient Plio-
550 Pleistocene barriers in monsoonal Australo-Papua: did mangrove habitats maintain genetic
551 connectivity in the Black Butcherbird? *Mol. Ecol.* **20**, 5042–5059.
- 552 40. Peñalba JV, Mason IJ, Schodde R, Moritz C, Joseph L. In press. Characterizing divergence
553 through three adjacent Australian avian transition zones. *J. Biogeogr.*
554 (doi:10.1111/jbi.13048)
- 555 41. R. Schodde, Mason IJ. 1999 The Directory of Australian Birds. Passerines. CSIRO
556 Publishing: Melbourne
- 557 42. Ford J. 1987 Hybrid Zones in Australian Birds. *Emu* **87**, 158–178.
- 558 43. Sorenson MD, Ast JC, Dimcheff DE, Yuri T, Mindell DP. 1999 Primers for a PCR-based
559 approach to mitochondrial genome sequencing in birds and other vertebrates. *Mol.*
560 *Phylogenet. Evol.* **12**, 105–114.
- 561 44. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012 Double digest RADseq:
562 an inexpensive method for de novo SNP discovery and genotyping in model and non-model
563 species. *PLoS One* **7**, e37135.
- 564 45. Eaton DAR. 2014 PyRAD: assembly of de novo RADseq loci for phylogenetic analyses.
565 *Bioinformatics* **30**, 1844–1849.
- 566 46. Langmead B, Salzberg SL. 2012 Fast gapped-read alignment with Bowtie 2. *Nat. Methods*
567 **9**, 357–359.
- 568 47. Korneliussen TS, Albrechtsen A, Nielsen R. 2014 ANGSD: Analysis of Next Generation
569 Sequencing Data. *BMC Bioinformatics* **15**, 356.
- 570 48. Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A,
571 Nielsen R. 2013 Quantifying population genetic differentiation from next-generation
572 sequencing data. *Genetics* **195**, 979–992.

- 573 49. Vieira FG, Lassalle F, Korneliussen TS, Fumagalli M. 2016 Improving the estimation of
574 genetic distances from Next-Generation Sequencing data. *Biol. J. Linn. Soc. Lond.* **117**,
575 139–149.
- 576 50. Huson DH. 1998 SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*
577 **14**, 68–73.
- 578 51. Reynolds J, Weir BS, Cockerham CC. 1983 Estimation of the coancestry coefficient: basis
579 for a short-term genetic distance. *Genetics* **105**, 767–779.
- 580 52. Weir BS, Cockerham CC. 1984 ESTIMATING F-STATISTICS FOR THE ANALYSIS OF
581 POPULATION STRUCTURE. *Evolution* **38**, 1358–1370.
- 582 53. Hartl DL, Clark AG, Clark AG. 1997 *Principles of population genetics*. Sinauer associates
583 Sunderland.
- 584 54. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. 2013 Calculation of Tajima’s D and
585 other neutrality test statistics from low depth next-generation sequencing data. *BMC*
586 *Bioinformatics* **14**, 289.
- 587 55. Paradis E, Claude J, Strimmer K. 2004 APE: Analyses of Phylogenetics and Evolution in R
588 language. *Bioinformatics* **20**, 289–290.
- 589 56. Csilléry K, François O, Blum MGB. 2012 abc: an R package for approximate Bayesian
590 computation (ABC). *Methods Ecol. Evol.* **3**, 475–479.
- 591 57. Hijmans RJ, Elith J. 2013 Species distribution modeling with R. *R package version 0. 8-11*
- 592 58. van Etten J. 2017 R Package gdistance: Distances and Routes on Geographical Grids.
593 *Journal of Statistical Software, Articles* **76**, 1–21.
- 594 59. Rheindt FE, Edwards SV. 2011 Genetic introgression: an integral but neglected component
595 of speciation in birds. *Auk* **128**, 620–632
- 596 60. Singhal S, Bi K. 2017 History cleans up messes: the impact of time in driving divergence
597 and introgression in a tropical suture zone. *Evolution* (doi:10.1111/evo.13278)
- 598 61. Winger BM, Bates JM. 2015 The tempo of trait divergence in geographic isolation: avian
599 speciation across the Marañón Valley of Peru. *Evolution* **69**, 772–787.
- 600 62. Whinnett A, Zimmermann M, Willmott KR, Herrera N, Mallarino R, Simpson F, Joron M,
601 Lamas G, Mallet J. 2005 Strikingly variable divergence times inferred across an Amazonian
602 butterfly ‘suture zone’. *Proc. Biol. Sci.* **272**, 2525–2533.
- 603 63. Price TD, Bouvier MM. 2002 The evolution of F1 postzygotic incompatibilities in birds.
604 *Evolution* **56**, 2083–2089.
- 605 64. Shafer ABA, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW, Wolf JBW. 2016

- 606 Bioinformatic processing of RAD-seq data dramatically impacts downstream population
607 genetic inference. *Methods Ecol. Evol.* (doi:10.1111/2041-210X.12700)
- 608 65. Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016 Variation in Linked Selection and
609 Recombination Drive Genomic Divergence during Allopatric Speciation of European and
610 American Aspens. *Mol. Biol. Evol.* **33**, 1754–1767.
- 611 66. Yang M, He Z, Shi S, Wu C-I. 2017 Can genomic data alone tell us whether speciation
612 happened with gene flow? *Mol. Ecol.* **26**, 2845–2849.
- 613 67. Harvey MG, Seeholzer GF, Smith BT, Rabosky DL, Cuervo AM, Brumfield RT. 2017
614 Positive association between population genetic differentiation and speciation rates in New
615 World birds. *Proc. Natl. Acad. Sci. U. S. A.* (doi:10.1073/pnas.1617397114)
- 616 68. Sukumaran J, Knowles LL. 2017 Multispecies coalescent delimits structure, not species.
617 *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1607–1612.