Plotting for change: an analytical framework to aid decisions on which lineages are candidate species in phylogenomic species discovery

PETER J. UNMACK¹,²,³,* MARK ADAMS¹,⁴, MICHAEL P. HAMMER⁵, JERALD B. JOHNSON³,⁶, BERND GRUBER¹, ANDRÉ GILLES⁷, MATTHEW YOUNG¹ and ARTHUR GEORGES¹,*,

¹Institute for Applied Ecology, University of Canberra, Bruce, ACT 2617, Australia
²Centre for Applied Water Science, Institute for Applied Ecology, University of Canberra, Bruce, ACT 2617, Australia
³Department of Biology, Brigham Young University, Provo, UT 84602, USA
⁴Department of Biological Sciences, University of Adelaide, Adelaide, SA 5005, Australia
⁵Museum & Art Gallery of the Northern Territory, Darwin, NT 0801, Australia
⁶Monte L. Bean Life Science Museum, Brigham Young University, Provo, UT 84602, USA
⁷UMR 1467 RECOVER, Aix Marseille Univ, INRAE, Centre St Charles, 3 place Victor Hugo, 13331 Marseille, France

Received 27 March 2021; revised 27 May 2021; accepted for publication 28 May 2021

A recent study argued that coalescent-based models of species delimitation mostly delineate population structure, not species, and called for the validation of candidate species using biological information additional to the genetic information, such as phenotypic or ecological data. Here, we introduce a framework to interrogate genomic datasets and coalescent-based species trees for the presence of candidate species in situations where additional biological data are unavailable, unobtainable or uninformative. For de novo genomic studies of species boundaries, we propose six steps: (1) visualize genetic affinities among individuals to identify both discrete and admixed genetic groups from first principles and to hold aside individuals involved in contemporary admixture for independent consideration; (2) apply phylogenetic techniques to identify lineages; (3) assess diagnosability of those lineages as potential candidate species; (4) interpret the diagnosable lineages in a geographical context (sympatry, parapatry, allopatry); (5) assess significance of difference or trends in the context of sampling intensity; and (6) adopt a holistic approach to available evidence to inform decisions on species status in the difficult cases of allopatry. We adopt this approach to distinguish candidate species from within-species lineages for a widespread species complex of Australian freshwater fishes (Retropinna spp.). Our framework addresses two cornerstone issues in systematics that are often not discussed explicitly in genomic species discovery: diagnosability and how to determine it, and what criteria should be used to decide whether diagnosable lineages are conspecific or represent different species.


*Corresponding author. E-mail: georges@aerg.canberra.edu.au

Graphical Abstract

Schooling *Retropinna* from south-eastern Australia in an aquarium [Nerang River, Queensland, taxon SEQb]. Photo: Gunther Schmida

INTRODUCTION

The age of low-cost genomics is progressing at a rapid pace. Already it has delivered unparalleled genetic insights at all levels of the systematic hierarchy (Rokas & Abbot, 2009; Lemmon & Lemmon, 2013; Dohrmann & Wörheide, 2017), with expectations of much more to come (Andrew et al., 2013; Campbell et al., 2018; Lewin et al., 2018). Accompanying this exceptional potential are important discussions and inevitable disputes about how best to analyse the voluminous genomic datasets now being generated routinely by researchers world-wide.

Nowhere are these musings likely to be more contentious than those surrounding the use of genomic datasets to discover species (Carstens et al., 2013). In this pursuit, there have been debates and assertions over the choice of genomic markers (Camargo & Sites, 2013; Andrews et al., 2016; Edwards et al., 2016a, b), data filtering and missing data protocols (Hovmöller et al., 2013; Huang & Knowles, 2016; Molloy & Warnow, 2018), the merits of adding more data, specimens or taxa (Degnan & Rosenberg, 2009; Camargo et al., 2011; Blom et al., 2016), the involvement of phylogenetic tree-building approaches (Stenz et al., 2015; Edwards et al., 2016b; Morrison, 2016; Wen et al., 2016), the choice of algorithms and software (Carstens et al., 2013; Xu & Yang, 2016; Barley et al., 2017; Luo et al., 2018), the extent to which introgression, incomplete lineage sorting and recombination confound analyses (Mallet, 2005; Lanier & Knowles, 2012; Leaché & Oaks, 2017), and whether species should be diagnosed formally using molecular datasets alone (Solis-Lemus et al., 2015; Delić et al., 2017; Sukumaran & Knowles, 2017).

However, underpinning these separate debates is one fundamental and seemingly intractable problem: given the near ubiquity of within-species population structure (Avise, 2000; Dynesius & Jansson, 2014) and the power of genomic data to detect fine-scale heterogeneity (Benestan et al., 2015), how do we ensure that our genomic datasets can distinguish between distinct species and diagnosable allopatric lineages within a single geographically structured species?

The answer to this question depends heavily on one’s concept of species. For those who lean toward viewing lineages as species and cladogenesis as speciation (e.g. Fujita et al., 2012), the challenge is not great; it becomes one of refining the approaches to identifying lineages unambiguously using phylogenetic methods that account for and accommodate some level of gene flow between entities used in the phylogenetic analysis (Yang & Rannala, 2010; Ence & Carstens, 2011; Leaché et al., 2014; Chan et al., 2020). Those who adopt the biological species concept (Mayr, 1964) as a theoretical foundation upon which to base operational definitions admit the possibility of phylogenetic structure within species. For them, all species are lineages, but not all lineages, even substantially divergent lineages on independent evolutionary trajectories, are necessarily species. In this vein, a recent paper by Sukumaran & Knowles (2017) argued convincingly, in our view, that the most popular methodological approaches currently used to delineate species using genomic data, namely those based on the multispecies coalescent model (herein, MSC model), often diagnose genetic structure rather than species. Most importantly, this genetic structure will mirror underlying species boundaries fully only in cases for which speciation can be considered a singular event in time (i.e. comparatively rarely; Avise et al., 1998; Dynesius & Jansson, 2014). Sukumaran & Knowles (2017) make...
four key assertions: (1) speciation is usually a process, not an event; (2) as such, popular species delimitation approaches for genomic datasets mostly delineate lineages, not all of which are species; (3) given the unrivaled power of genomic data to identify fine-scale structure, the potential for taxonomic hyperinflation created by mistakenly designating lineages as species, without recourse to other criteria, will create major problems for many biological disciplines; and (4) until methods are formulated to decide which lineages are species, genome-based results should be validated with multiple data types, such as phenotypic or ecological information, before candidate taxa are regarded formally as species. Considering the anticipated rate of growth in the use of genomic datasets (Andrew et al., 2013), this seems an opportune time for the systematic community to consider deeply the merits of these assertions.

It appears that a majority of practising molecular systematists accept that assertions 2 and 3 above follow logically from assertion 1, the almost unchallengeable reality that allopatric speciation, by far the most common means by which species are generated (Mayr, 1964), is usually a protracted process, often accompanied by sporadic genetic exchange between the incipient species, rather than a single point-in-time event (Turelli et al., 2001; Mallet, 2005; Georges et al., 2018). However, the same level of consensus is unlikely to apply to assertion 4, which concludes in effect that, in the short term, we require validation of delimited species using multiple types of data, such as traditional, non-molecular characters as the final arbiter of species integrity, despite these characters so often proving to be data deficient, unavailable, inadequate or misleading in many groups (Bickford et al., 2007; Hammer et al., 2013; Georges et al., 2018) and totally uninformative for genuinely cryptic species (Delić et al., 2017; Moritz et al., 2018). Furthermore, all modern species concepts agree that unequivocally diagnosable lineages in sympathy (i.e. via fixed differences at multiple, unlinked nuclear genes) represent distinct species for sexually reproducing organisms (Avise, 1994; Hammer et al., 2013; Mallet, 2013), regardless of whether the evidence is solely molecular. Sukumaran & Knowles (2017) restrict their attention to cases of allopatry, and this highly salient point is a counterbalance to their call for a need to validate conclusions on species boundaries using traditional morphological or ecological data. This raises an obvious question: can we devise some agreed protocols to mine genomic datasets better, such that they add value to MSC model tree-based analyses, and thereby provide a relatively objective framework for assessing which lineages merit recognition as species and which do not?

A concurrent and equally relevant concern with the use of any MSC model to delineate species is that, like all methods used to generate bifurcating phylogenetic trees, they assume that the data themselves are ‘tree-like’ (Bordewich & Tokac, 2016; Mallet et al., 2016). Unfortunately, there are a range of well-acknowledged reasons (e.g. hybridization/introgression, horizontal gene transfer, incomplete lineage sorting, gene conversion, intracistronic recombination and the misidentification of orthologues resulting from undetected gene duplication/loss) why genetic data may not be tree-like (Morrison, 2016), both among species and, most notably, for individuals and populations within a single species (Naciri & Linder, 2015; Xu & Yang, 2016). In particular, the genetic cohesion resulting from ongoing regular or sporadic gene flow among conspecific populations often ensures that any embedded phylogenetic relationships are, unavoidably, obscured by these high levels of reticulate evolution (Mallet et al., 2018). Summarizing these concerns, Morrison (2016) has contended that a tree model for phylogenetic history is not suitable for the study of genomes. This is especially relevant in studies of species delimitation, where the divergence of putative taxa under consideration is often shallow.

Although many researchers acknowledge that bifurcating trees are not a complete solution (Huson & Scornavacca, 2010; Edwards et al., 2016a; Morrison, 2016; Chan et al., 2020), it seems inevitable that they will continue to remain the cornerstone of future genome-based species discovery unless improved approaches become widely available. We have no quarrel with the use of bifurcating trees per se, only with them being the sole or primary focus in species delimitation. Our fundamental contention here is that species delimitation studies, whether traditional, genetic or genome based, should supplement any tree-based or network-based approach by cross-referencing with five additional tree-free analyses to yield a six-step process, namely: (1) visualize genetic affinities among individuals to identify both discrete and admixed genetic groups from first principles and to hold aside individuals involved in contemporary admixture for independent consideration; (2) apply phylogenetic techniques to identify lineages; (3) assess diagnosability of those lineages as potential candidate species; (4) interpret the diagnosable lineages in a geographical context (sympathy, parapatry, allopatry); (5) assess significance of difference or trends in the context of sampling intensity; and (6) adopt a holistic approach to available evidence (e.g. morphology, ecology, mode of reproduction, reproductive compatibility) to inform decisions on species status in the difficult cases of allopatry. Taken together, we suggest that these six steps provide a suitable framework to decide which lineages deserve recognition as valid candidate
species vs. those that should continue to be regarded as substantive lineages within species, subject to further evidence of speciation becoming available.

To illustrate our approach and demonstrate how it can complement the call for traditional taxonomic validation of the candidate species identified using genomic data (Sukumaran & Knowles, 2017), we consider a case study on abundant and widespread Australian fishes in the genus *Retropinna* (Teleostei: Retropinnidae) (Allen et al., 2002; Fig. 1). The current taxonomic framework (last reviewed by McDowall, 1979) formally recognizes two Australian species: *Retropinna tasmanica*, a possibly anadromous species found in estuaries and freshwaters of all larger, low-elevation streams in Tasmania, and *Retropinna semoni*, which is widespread and generally abundant in a wide variety of natural and human-made habitats throughout much of mainland south-eastern Australia (Fig. 2; Supporting Information, Table S1). A third species, *Retropinna retropinna*, is restricted to New Zealand.

The taxonomic nomenclature used herein is based on a comprehensive allozyme study by Hammer et al. (2007), which proposed five candidate species (CEQ, SEQ, SEC, MTV and COO; Fig. 2), each diagnosable by fixed differences at multiple allozyme loci. This includes two instances of sympatry (CEQ/SEQ and SEQ/SEC), one instance of parapatry (SEC/MTV) and six instances of allopatry (including the sister taxa MTV and COO). Hammer et al. (2007) also found one unresolved complex (MTV), where the allozyme data could not distinguish between a scenario of two allopatric taxa (MDB and TAS) that hybridize across a widespread but relatively narrow overlap zone (comprising sites designated herein as ADM; Fig. 2) vs. a scenario of one widespread taxon displaying strong clinal patterns across multiple loci. Thus, the Australian *Retropinna* provide an ideal case study to explore our proposed framework as a means of supplementing tree-based analyses of genomic and other molecular datasets. The availability of four additional molecular datasets [allozymes, mitochondrial DNA (mtDNA) and two different nuclear introns] allows us to confirm and buttress the primary genomic dataset [initially 89,870 single nucleotide polymorphisms (SNPs)] and assess the relative merits of different molecular datasets.

**MATERIAL AND METHODS**

**Taxonomic sampling**

We sampled 738 individuals from 91 sites across the Australian range of *Retropinna*, with six sites sampled on two occasions (Fig. 2; Supporting Information, Table S1). Five molecular datasets were generated from distributionally comprehensive subsets of these individuals and sites, namely: (1) our primary genomic dataset (*N* = 459 fish, 56 sites, 89,870 SNPs); (2) mtDNA sequence data for the genes *cytochrome b* (*cytb*) and *16S* (*N* = 229, 60 sites, 1600 bp); (3) nuclear DNA (nDNA) intron sequence data for the *alpha-tropomyosin* gene (*N* = 215, 60 sites, 314 bp for intron 5); (4) nDNA intron sequence data for the *S7* ribosomal protein gene (*N* = 200, 58 sites, 819 bp for intron 1); and (5) an updated regional allozyme study of the admixture zone in taxon MTV (*N* = 147 fish from 16 new sites for 29 polymorphic allozyme loci, integrated with an existing dataset [in the study by Hammer et al., 2007] for 250 fish from 35 sites). Full details of field collection procedures are presented in the Supporting Information (Appendix S1: Materials and Methods).

**Single nucleotide polymorphism genotyping and data filtering**

Sequencing for the SNP dataset used DArTseq (Diversity Arrays Technology, Canberra, ACT, Australia), a variation of the double-digest restriction site associated DNA (ddRAD) technique that combines next generation sequencing, complexity reduction using restriction enzymes and implicit fragment size selection, as described by Kilian et al. (2012) and Georges et al. (2018).
The SNP dataset underwent two phases of filtering and error checking, one included automatically as part of DArTseq standard protocols (fully detailed by Georges et al., 2018), followed by various operator-defined choices implemented on the final ‘raw dataset’ using the R package DARTr v.1.0.5 (Gruber et al., 2018). These raw data were subjected to six sequential filtering procedures to generate two final SNP datasets, one for the phylogenetic trees and a more-stringent version for the multivariate plots among individuals and fixed difference analysis. For the latter, these filters removed: (1) loci that did not show close to 100% reproducibility (averaged over the two alleles, repAvg ≥ 0.995) for the ~30% of individuals that are randomly resequenced as a routine by DArT; (2) loci displaying > 10% missing genotypes (call rate by locus ≥ 0.9); (3) monomorphic loci; (4) tightly linked loci (filter secondaries, retaining one SNP at random); (5) individuals that display > 20%
missing genotypes (call rate by individual ≥ 0.8); and
(6) a final filter for any additional monomorphic loci created by removal of individuals. Given the low within-
population sample sizes (N ≤ 14), we did not filter loci for
departures from Hardy–Weinberg equilibrium or linkage
disequilibrium. Georges et al. (2018) discussed in detail
the rationale behind these filtering choices.

**STEP 1: ORDINATION OF THE GENETIC AFFINITIES
AMONG INDIVIDUALS**

As step 1, we advocate that all species delimitation
studies begin with some form of ordination to groupings
of individuals based on genetic similarity. This brings the
analyst close to the relevant attributes of the data and
provides an important level of expectation for what is likely
to emerge from the more complex analyses that are to follow.
Ordination can also detect putative hybrid individuals and sites
displaying substantial admixture among groups (Adams et al., 2014; Georges et al., 2018), both of which will
distort phylogenetic estimation and analyses designed to
establish diagnosability. Principal coordinates analysis
(PCoA) is a general technique for visualizing genetic
structure within a dataset that can be applied to all types
of data for which a sensible measure of genetic distance
can be applied. For the SNP dataset, we used a stepwise
implementation (Georges & Adams, 1992) of PCoA (DARTR
gl.pcoa and gl.pcoa.plot functions) to visualize the genetic
affinities among individuals with no priors. This allows
further examination of structure within clusters that are
well defined in the initial ordination. Ordinations were also
undertaken on the three DNA sequence datasets (mtDNA,
alpha-tropomyosin and S7) with a distance matrix of
p-distances (proportion of nucleotide sites differing) among
individuals, using MEGA 7.0.26 (Kumar et al., 2016).

**STEP 2: CONSTRUCTION OF GENE AND SPECIES TREES**

The second step in our interrogation framework is the
generation of species trees. We used single value
decomposition (SVD) quartets (Chifman & Kubatko, 2014,
2015) and maximum likelihood (RAxML; Stamatakis,
2014) applied to concatenated sequences for generating
genomic trees for Retropinna. Heterozygous SNP
positions were represented by standard ambiguity codes
(Felsenstein, 2004). We used SVDQUARTETS in PAUP*
(v.4.0a162; Swofford, 2003) with the following parameters:
evaQuartets = random, bootstrap = standard,
nreps = 10 000 and ambigs = distribute. Maximum
likelihood (ML) analyses were carried out with RAxML
v.8.2.10 on the CIPRES cluster (Miller et al., 2010) using
the model GTRCAT and searching for the best-scoring
ML tree using the model GTRGAMMA in a single
program run, with bootstrapping set to finish based on the
autoMRE majority rule criterion.

**STEP 3: ASSESSMENT OF LINEAGE DIAGNOSABILITY**

Although systematists accept that both lineages
and species ought to be diagnosable, there has been
comparatively little discussion in the literature about
what operational criteria should apply to this concept.
Continuing a long tradition from the pregenomic era,
studies based on the MSC model typically neglect to
provide an explicit operational definition, leaving the
reader uncertain regarding the taxonomic characters
under consideration (locus or nucleotide site) and
measures that are used to demonstrate that two
lineages are diagnosable (e.g. fixed differences or
large differences in allele frequency across multiple,
independent characters; large values for the fixation
index ($F_{ST}$) or other summary genetic distance; or
substantive lineages simply read off the tree).

Here, we have adopted two locus-based criteria for
our allozyme data, one absolute (a fixed difference
equates to no shared alleles or haplotypes at a locus;
see Georges et al., 2018) and one 'low tolerance'(two
populations are regarded as displaying an effective
fixed difference if the cumulative percentage of
shared alleles/haplotypes is between 0 and 10%;
see Chifman & Kubatko, 2015). Of these, the
first is likely to suit bi-allelic markers with low
expectations of homoplasy (i.e. SNPs), whereas the
second is likely to be more appropriate for higher-
homoplasic, multi-allelic markers (i.e. allozymes,
microsatellites and DNA sequences). For the SNP
data, we used only the first, absolute, approach. All
counts of the number of SNPs displaying fixed
differences were calculated in DARTR (function
gl.fixed.diff, criteria tloc = 0) on all pairwise
combinations of taxa (each reduced dataset having
first been refiltered for the more stringent, six-step
phase outlined previously). Many of the sample sizes
for our populations (sampling sites) were small; in
some cases, only one. This admits the possibility
that the true number of fixed differences between
two populations can be obscured by false positives
when comparing the samples drawn from those
populations; therefore, several operational decisions
needed to be incorporated into the analysis. First,
we combined populations with a sample size of one
manually with an adjacent population within the
same drainage (SEC.Mack, SEC.Glou, SEC.Timb
and SEQb.Twee; Supporting Information, Table S2D).
Second, we insisted that any two populations
had to have fixed differences that were corroborated
by more than five fixed differences in order to
regard them as distinct (by setting tpop = 5 in
gl.collapse of DARTR). This value was taken in
the context that the average number of fixed differences
between populations taken pairwise was 271; that
is, the decision was a very conservative measure to

© 2021 The Linnean Society of London, Biological Journal of the Linnean Society, 2021, XX, 1–21
control the influence of false positives arising from small sample sizes. Finally, for cases where the samples sizes were adequate (ideally, at least ten) we compared the observed count of fixed differences statistically with the expected rate of false positives using the test provided in gl.fixed.diff of DARTR.

Assessments of lineage diagnosability for the three DNA sequence datasets were undertaken within the R package SPIDER v.1.5.0 (Brown et al., 2012) to identify fully diagnostic nucleotide sites (using the nucDiag function, which lacks a low-tolerance option).

**Steps 4 and 5: Pairwise Comparisons of Candidate Species and Lineages**

Steps 4 and 5 in our interrogation framework require that all key lineages are compared pairwise for geographical distribution and sampling intensity (both geographical and genomic coverage). Of these, step 4 is of special relevance to the assessment of geographical context, because it builds upon a central but often-ignored tenet of taxonomy that, for sexually reproducing organisms in sympathy, absolute fixed differences at multiple, independent characters displaying codominant states provide unequivocal evidence for the existence of more than one species. Figure 3 presents our conceptual perspective on the diagnosability ‘burden of proof’ required to conclude that two lineages are candidate species for a range of sympatric, parapatric and allopatric scenarios. Figure 3 should be interpreted in the context that diagnosability is a necessary, but not sufficient, criterion for defining species.

**Step 6: Bringing in Other Evidence**

Although not strictly necessary for this study, we present four other comprehensive molecular datasets on *Retropinna*. Not only does their inclusion minimize any concern that a stand-alone genomic study might lack the certainty framework available when buttressed by older genetic technologies, but these additional datasets can also serve as part of our ‘other biological differences’ assessment. All details of the procedures used to generate and analyse the mtDNA, nDNA (two introns) and allozyme datasets are

![Figure 3](https://academic.oup.com/biolinnean/advance-article-doi/10.1093/biolinnean/blab095/6384958)

**Figure 3.** Diagrammatic representation of the conceptual framework used in this study to classify all pairwise comparisons of lineages and candidate species for their comparative geographical distributions. The codes (highlighted in yellow) assigned to each scenario match those used in Table 1. The terms shallow, moderate and deep relate to how the gap distance compares with the combined geographical areas occupied by the two taxa.
presented in the Supporting Information (Appendix S1: Materials and Methods).

RESULTS

STEP 1: VISUALIZE GENETIC AFFINITIES AMONG INDIVIDUALS

Stringent filtering of the SNP raw data (459 fish; 89 870 loci; 34.4% missing data) resulted in a final dataset comprising the genotypes of 457 Retropinna spp. (Supporting Information, Table S1) for 3954 loci (4.6% missing data). Principal co-ordinates analysis, with individuals as entities and loci as attributes, revealed five clearly defined clusters in the first two dimensions (Fig. 4A), four of which correspond to the following putative taxa or combinations of taxa: CEQ, SEQ, ADM/COO/MDB/TAS and SEC. The distinctive fifth grouping comprised a single individual from site 15, an SEQ × SEC F₁ hybrid that: (1) was heterozygous at all diagnostic loci for the two putative parent taxa; (2) co-occurred at site 15 with 19 other individuals that were from one or the other parental taxon (SEC, N = 14 or SEQ, N = 5 respectively; Supporting Information, Table S1); (3) was highly heterozygous (observed heterozygosity Hₒ = 0.193) compared with all other taxa (Hₒ range 0.011–0.020); and (4) occupied an appropriate intermediate position in the ordination.

Ordinations applied separately to each of the four clearly defined clusters (sensu Georges & Adams, 1992; Fig. 4A) revealed further structure. There were two or more clearly defined, mostly allopatric groups within each cluster: three for CEQ (Fig. 4B), three for SEC (Fig. 4C), four within SEQ, with a primary subdivision into northern (SEQa) and southern (SEQb) clusters (Fig. 4D), and either two (COO vs. MDB/ADM/TAS) or three (COO, MDB and TASN) clusters within the ‘MTV/COO’ complex, depending on whether the ADM sites were included (Fig. 4E shows the ADM sites as seamlessly bridging the gap between the otherwise distinctive MDB and TASN clusters; Fig. 4F shows ADM sites excluded). In summary, ordination of the genomic data fully supported four of the five primary candidate species (CEQ, SEQ, SEC and MTV) proposed by Hammer et al. (2007), provided some support for the distinctiveness of taxon COO, identified a number of secondary, geographically based clusters within four of these candidate species, highlighted a single F₁ hybrid for exclusion from the species tree analyses and confirmed that there was no obvious cryptic genetic heterogeneity at individual sites other than the already-proposed MDB/TASN historical admixture zone (ADM sites). These results also closely mirrored those obtained for the original allozyme study (summary tree in Supporting Information, Fig. S1A) and for our expanded regional allozyme study of the admixture zone in MTV (Supporting Information, Fig. S1B).

STEP 2: IDENTIFY LINEAGES

The SNP dataset for our two phylogenetic analyses was less stringently filtered than for the PCoA, because the phylogenetic algorithms are more tolerant of missing values. The data comprised 448 individuals (excluding the F₁ individual and ten individuals dropped with a call rate < 70%) from 58 populations scored for 11 980 loci (10.6% missing data). The SVD quartets tree (Chifman & Kubatko, 2014, 2015; Fig. 5A) has a topology fully consistent with the PCoA, in that the major clades revealed in the tree correspond to the major groupings in the PCoA (minus the F₁ hybrid). There is clear bootstrap support for the five candidate species of Hammer et al. (2007) except the MTV vs. COO split; COO appears as an early-branching lineage within a well-supported MTV/COO clade. The SVD quartets tree and ordination also agree with Hammer et al. (2007) in recognizing SEQa and SEQb as distinct lineages but are concordant only in part on how most other secondary PCoA groupings are delineated. Importantly, the SVD quartets phylogram struggles to handle the varying degrees of MDB × TAS historical admixture present among the ADM sites and, in common with all bifurcating trees, is constrained to placing these reticulated populations in somewhat chaotic and often early-branching positions between the two parental lineages.

A similar topology to the SVD quartets tree is evident in the RAxML tree of the 446 individuals (Supporting Information, Fig. S2), a condensed summary of which is presented in Figure 5B. Bootstrapping ceased automatically at 592 replications. Here, strong support is evident for all five candidate species of Hammer et al. (2007) and for the presence of two well-differentiated lineages within SEQ. Once again, the ADM sites are scattered in a superficially random manner throughout the MTV-defining clade.

STEP 3: ASSESS DIAGNOSABILITY

A fixed difference analysis, with no prior assignment of populations to the candidate taxa of Hammer et al. (2007), was used to identify diagnosable aggregations of populations [operational taxonomic units (OTUs)]. Four populations with sample sizes of one were amalgamated manually with their closest neighbouring population within the same catchment (Supporting Information, Table S2D). The diagnosable OTUs arising from the fixed difference analysis are provided in the Supporting Information (Table S2H) and depicted in Figure 5A. The largest OTU combined the nine populations from Tasmania (TAS), the 13
populations from the Murray–Darling Basin (MDB) and the ten intervening populations in Victoria (ADM) as a single diagnosable taxon corresponding to the candidate species MTV of Hammer et al. (2007). This aggregation appears clinal (Fig. 4E), with considerable internal genetic variability. The populations of SEC were characterized by six diagnosable OTUs, and each of CEQ, SEQa and SEQb by three (Supporting Information, Table S2H). The two COO populations were not diagnostically different from other OTUs but could not be assigned unambiguously (Supporting Information, Table S2G).

All five candidate species of Hammer et al. (2007) were fully diagnosable by absolute fixed differences at 12–110 loci (Table 1), although the differences between COO and MTV ($p = 1.00$), COO and CEQ ($p = 1.00$) and COO and SEQ ($p = 0.48$) were not statistically significant, suggesting that further sampling is needed to resolve the taxonomic status of COO. In no case did the definition of diagnosable OTUs challenge

---

**Figure 4.** Ordination plots in the first two dimensions for the initial principal coordinates analysis (PCoA; A) and follow-up PCoAs for all four primary groups (B–F) identified in panel A. Axes are scaled to reflect their relative contribution to total variance (as shown in parentheses). Colours denote candidate species or lineages of Hammer et al. (2007). Secondary clusters in A–D are numbered according to site. Panel F has ADM removed.
the concept of the candidate species identified by Hammer et al. (2007), and the fixed difference analysis strengthened the conclusion that the MTV is a single diagnosable taxon.

**Step 4: Interpret in geographical context**

We now need to decide whether diagnosable lineages merit recognition as candidate species. Underpinning this process is a fundamental taxonomic principle that the presence of unequivocally diagnosable lineages (i.e. fixed differences at multiple, unlinked loci) in sympatry for sexually reproducing organisms provides compelling evidence for the presence of two species, regardless of which modern species concept is adopted. Any candidate species thus delineated then furnishes an informal ‘genomic yardstick’ that might help to assess the status of other parapatric and allopatric lineage pairs, being mindful that the number of diagnostic genomic characters required to nominate candidate species increases across the full spectrum from widespread parapatry to deep allopatry with an unsampled gap (Fig. 3).

The relative geographical distributions of all candidate species and selected conspecific lineages are summarized in Table 1. For allopatric lineages, we have reviewed current knowledge of whether the gap is unsampled (populations likely to exist but not sampled, e.g. SEQa vs. SEQb; Islam et al., 2018) or genuine (a probable real absence of intermediate populations, Figure 5. Phylogenomic trees for 58 populations of Retropinna based on the single nucleotide polymorphism dataset. *Nodes receiving strong bootstrap support (> 97%). A, SVD quartets species tree, with populations labelled by the candidate taxon plus site and sympatric populations additionally identified by the # symbol. Text is coloured on the terminals of the SVD tree according to our final determination of species (refer to main text). Left-hand vertical bars indicate diagnosable operational taxonomic units based on the fixed difference analysis; taxa without a vertical bar are diagnostic as a single population. Right-hand vertical boxes represent the candidate taxa of Hammer et al. (2007) plus ADM. B, the maximum likelihood tree showing the primary lineages for all 448 individuals (full tree in Supporting Information, Fig. S2).
<table>
<thead>
<tr>
<th>Pairwise comparison</th>
<th>FD</th>
<th>No_loci</th>
<th>%FD</th>
<th>Comparative distribution</th>
<th>Geographical sampling intensity</th>
<th>mtDNA</th>
<th>Alpha-tropomyosin</th>
<th>S7</th>
<th>Other biological data</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEQ vs. SEQ</td>
<td>52</td>
<td>3935</td>
<td>1.32</td>
<td>SP (site 4)</td>
<td>Adequate: adequate</td>
<td>33</td>
<td>2</td>
<td>37</td>
<td>6</td>
</tr>
<tr>
<td>SEQ vs. SEC</td>
<td>101</td>
<td>3944</td>
<td>2.56</td>
<td>SP (sites 14 and 15)</td>
<td>Adequate: strong</td>
<td>114</td>
<td>0</td>
<td>26</td>
<td>13</td>
</tr>
<tr>
<td>SEC vs. MTV</td>
<td>16</td>
<td>3944</td>
<td>0.41</td>
<td>PP/AS</td>
<td>Strong:-strong</td>
<td>79</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>CEQ vs. SEC</td>
<td>106</td>
<td>3935</td>
<td>2.69</td>
<td>AMg</td>
<td>Adequate: strong</td>
<td>142</td>
<td>2</td>
<td>42</td>
<td>11</td>
</tr>
<tr>
<td>CEQ vs. MTV</td>
<td>40</td>
<td>3935</td>
<td>1.02</td>
<td>AS</td>
<td>Adequate: strong</td>
<td>132</td>
<td>8</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>CEQ vs. COO</td>
<td>110</td>
<td>3916</td>
<td>2.81</td>
<td>AMg</td>
<td>Adequate: adequate</td>
<td>184</td>
<td>12</td>
<td>39</td>
<td>14</td>
</tr>
<tr>
<td>SEC vs. MTV</td>
<td>63</td>
<td>3944</td>
<td>1.60</td>
<td>ASb</td>
<td>Adequate: strong</td>
<td>108</td>
<td>1</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>SEC vs. COO</td>
<td>136</td>
<td>3925</td>
<td>3.46</td>
<td>ADg</td>
<td>Adequate: adequate</td>
<td>154</td>
<td>4</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>SEC vs. COO</td>
<td>67</td>
<td>3925</td>
<td>1.71</td>
<td>ADg</td>
<td>Strong:adequate</td>
<td>127</td>
<td>4</td>
<td>35</td>
<td>13</td>
</tr>
<tr>
<td>MTV vs. COO</td>
<td>12</td>
<td>3925</td>
<td>0.31</td>
<td>AS/AMg</td>
<td>Strong:adequate</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>MDB vs. COO</td>
<td>32</td>
<td>3936</td>
<td>0.81</td>
<td>AS/AMg</td>
<td>Strong:adequate</td>
<td>22</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>SEQa vs. SEQb</td>
<td>30</td>
<td>3905</td>
<td>0.77</td>
<td>AMu/AS</td>
<td>Deficient:deficient</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TAS vs. MDB</td>
<td>2</td>
<td>3954</td>
<td>0.05</td>
<td>(gap = ADM)</td>
<td>Strong:strong</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>ADM vs. MDB</td>
<td>0</td>
<td>3954</td>
<td>0.00</td>
<td>PW</td>
<td>Strong:strong</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ADM vs. TAS</td>
<td>0</td>
<td>3954</td>
<td>0.00</td>
<td>PW</td>
<td>Strong:strong</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Codes for diagnostic single nucleotide polymorphisms (SNPs; step 3): FD, number of absolute fixed differences; No_loci, total number of SNPs after removal of monomorphs; %FD, FD/No_loci as a percentage. Codes for comparative distributions (step 4) use the conceptual terminology outlined in Figure 3. Codes for geographical sampling intensity (step 5): adequate, fewer sites or lower coverage when compared with 'strong' rating; deficient, lacking in either the number of sites/specimens or in matching sampling coverage with putative distribution; poor, only range-restricted site(s), with real distribution of putative lineage unknown; strong, multiple sites/individuals, and geographical coverage matches putative distribution. Step 6 (other non-SNP data) includes an 'other molecular diagnoses' component, which presents the number of fixed nucleotide differences between the nominated lineages for the three DNA sequence datasets [mitochondrial DNA (mtDNA), alpha-tropomyosin and S7] together with the number of near-fixed (tolerance 5%) allozyme loci; raw data are in the paper by Hammer et al. (2007).
e.g. MTV vs. COO (Hammer et al., 2007) or the gap is occupied by another taxon, e.g. CEQ vs. SEC. As indicated, there are two instances of taxa where their ranges overlap, bringing them into sympatry. These cases validate the three candidate species CEQ, SEQ, and SEC (CEQ and SEC are allopatric but display levels of divergence in fixed alleles that are comparable to those found between each and its sympatric congener, SEQ). Indeed, all pairwise values among candidate species are generally comparable to or exceed the CEQ/SEQ yardstick (52 fixed differences), except for the MTV vs. COO comparison (12 fixed differences; Table 1).

**Step 5: Assessment of Sampling Intensity**

Table 2 summarizes the extent of genomic and geographical coverage and sampling intensity for all candidate species and selected conspecific lineages. Here, we have rated all genomic sampling as strong, given that SNPs provide intensive and random, genome-wide coverage for mostly codominant genetic markers. As shown, the geographical coverage and sampling intensity for all candidate species is either strong (and arguably very strong for MTV; Fig. 2) or adequate. As such, and given the very extensive geographical distributions of both SEC and MTV taken together with their partial parapatry status (Table 1), we contend that our genomic data demonstrate MTV to be a valid candidate species, despite its somewhat lower levels of diagnosability with SEC (16 fixed differences). In contrast, other conspecific lineages or populations either lack sufficient diagnostic loci (within CEQ and MTV) or, although clearly diagnosable, lack the geographical sampling intensity required to conclude with any confidence that unsampled intermediate populations will not also be intermediate genetically (as found in the TAS/ADM/MDB admixture complex in MTV). Thus, although the SEQa vs. SEQb dichotomy is intriguing, we affirm that these lineages should continue to be regarded as conspecific, pending further genomic characterization of known but unsampled, geographically intermediate populations (Islam et al., 2018).

**Step 6: Final Assessment, Bringing in Other Evidence**

There remains one candidate taxon, COO, for which the genomic data are somewhat equivocal regarding whether it represents an allopatric sister species to species MTV or a distinct lineage of species MTV. Both taxa are diagnosable but not significantly so, and COO could not be assigned unambiguously to another aggregation based on fixed differences (Supporting Information, Table S2). This is consistent with the low bootstrap support for the node uniting COO with MTV in the SDV quartets tree. This is where step 6 can help to inform the final decision. Table 1 summarizes all the other molecular diagnoses and other biological differences among pairwise taxa. Both mtDNA sequence data and allozymes clearly diagnose all five candidate species from one another, including COO from MTV. Despite this, COO haplotypes are nested within the MTV complex in the mtDNA gene tree (summary tree in Fig. 6A , with full tree in Supporting Information, Fig. S3), providing a further demonstration that diagnosability cannot be assessed automatically using tree-based analyses. As commonly

<table>
<thead>
<tr>
<th>Taxon/lineage</th>
<th>Genome coverage</th>
<th>Geographical coverage</th>
<th>Geographical sampling intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total sites</td>
<td>Total N</td>
</tr>
<tr>
<td>CEQ</td>
<td>Strong</td>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>SEQ</td>
<td>Strong</td>
<td>13</td>
<td>69</td>
</tr>
<tr>
<td>SEC</td>
<td>Strong</td>
<td>17</td>
<td>96</td>
</tr>
<tr>
<td>MTV</td>
<td>Strong</td>
<td>62</td>
<td>531</td>
</tr>
<tr>
<td>COO</td>
<td>Strong</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>SEQa</td>
<td>Strong</td>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td>SEQb</td>
<td>Strong</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td>ADM</td>
<td>Strong</td>
<td>23</td>
<td>171</td>
</tr>
<tr>
<td>MDB</td>
<td>Strong</td>
<td>28</td>
<td>267</td>
</tr>
<tr>
<td>TAS</td>
<td>Strong</td>
<td>11</td>
<td>93</td>
</tr>
</tbody>
</table>

Terminology for geographical sampling intensity (step 5) matches that used in Table 1. ‘Total sites’ includes all genetic datasets.
found for single nDNA genes, alpha-tropomyosin and S7 are able to diagnose some, but not all, pairwise combinations of candidate species (Table 1), although both show fixed differences between COO and MTV at two or three nucleotide sites. This lack of full diagnosability across all species is clearly evident in the two gene trees (alpha-tropomyosin summary tree in Fig. 5B, with full tree in Supporting Information, Fig. S4; S7 summary tree in Fig. 5C, with full tree in Supporting Information, Fig. S5).

Not only is COO consistently diagnosable from MTV across four of our molecular datasets, but several other biological differences also bolster our contention that both are valid candidate species (Table 1). Species COO displays several unique morphological/ecological differences from other populations in the complex, most notably smaller overall size but with larger eyes (Wager & Unmack, 2000), pronounced sexual dimorphism at a smaller size, higher modal counts for dorsal fin rays and lower modal counts for vertebrae (McDowall, 1979). This conclusion is also consistent with an earlier taxonomic decision by Lake (1971), before the revision by McDowall (1979).

DISCUSSION

One perhaps unanticipated outcome of the genomics age is that vastly increased volumes of detailed molecular genetic data have not equated automatically to greater ease in delineating species (Coates et al., 2018; Leaché et al., 2019). On the contrary, this task is now arguably harder, because a greater number of subjective taxonomic decisions are required to accommodate the corresponding increase in diagnosable allopatric lineages detectable using genomic datasets (Georges et al., 2018; Singhal et al., 2018). This difficulty is exacerbated by recent observations that species trees based on the MSC model delineate population structure rather than species per se (Sukumaran & Knowles, 2017; Leaché et al., 2019), a finding that is likely to apply, in principle, to all methods that generate strictly bifurcating trees (Pickrell et al., 2012; Morrison, 2016; Georges et al., 2018).

Among those who acknowledge these limitations, responses have varied from attempting to develop or refine ‘better’ approaches to the estimation of species trees (e.g. Bouckaert, 2010; Pickrell et al., 2012; Leaché et al., 2019), supplementing MSC model trees with additional, delineation-oriented statistical analyses.
(Miralles & Vences, 2013; Edwards & Knowles, 2014; Hime et al., 2016; Barrow et al., 2018; McCartney-Melsted et al., 2018), arguing for conservatism in deciding which allopatric lineages merit recognition as distinct species (Sukumaran & Knowles, 2017; Coates et al., 2018; Singhal et al., 2018) or advocating that all ‘well-supported’ lineages (particularly, high-profile taxa under threat of extinction) be considered to be valid evolutionary species (Freudenstein et al., 2016; Groves et al., 2017).

In this study, we have outlined a more comprehensive approach, which draws on many of the points of view mentioned above. Our conservative six-step framework for sexually reproducing organisms attempts to combine the fundamental taxonomic principles that underpin species delineation in the pre-genomics era (the need to identify fully diagnostic characters, assess comparative geographical distribution and consider sampling intensity explicitly) (Mayr, 1964; Helbig et al., 2002; Hammer et al., 2013) with the power of MSC model analyses to identify evolutionary lineages (Coates et al., 2018) and the ability of ordination to detect admixture and assess the genetic distinctiveness of such lineages independently (Georges et al., 2018). Moreover, it can also be applied retrospectively to any existing genomic or molecular study of species boundaries, regardless of the type of data analysed or the tree-building method chosen. In the case of an MSC model species tree, one can simply apply step 1, followed by steps 3–5 for each pairwise comparison of candidate taxa, to assess whether there is robust evidence to reject the proposition that two lineages are conspecific (step 6). Importantly, our approach is broadly compatible with most modern species concepts (except a rigid application of the phylogenetic species concept; Groves et al., 2017) and will help to counter taxonomic inflation (Isaac et al., 2004).

Species delineation in Australian Retropinna
Applying this framework to Australian Retropinna has allowed us unequivocally to delineate four candidates that warrant recognition at the level of species (CEQ, SEQ, SEC and MTV), provide justification for recognizing a fifth species (COO), reveal the presence of two distinctive lineages in species SEQ that require further genomic sampling in the intervening region before any follow-up re-evaluation of their taxonomic status, identify an F1 hybrid and thus prevent it from contaminating key analyses, and demonstrate a broad zone of historical admixture between two diagnosable allopatric lineages (TAS and MDB) within species MTV, a scenario that was not clearly evident on our species trees alone. It is worth noting here that these MDB and TAS populations, if considered without the intervening ADM populations, are morphologically distinctive (Table 1), to the point where they have previously been considered separate species (McDowall, 1979; Hammer et al., 2007). This highlights the importance of comprehensive sampling, in order that individuals examined at isolated sampling locations are not misinterpreted as distinct taxa (Chambers & Hillis, 2020). Despite this, all our genetic/genomic datasets show the lineages representing populations at the extremes of the cline (TAS and MDB) to be diagnosable only marginally in shallow allopatry (Table 1), suggesting that a genomic study of species MTV based on only a handful of sites would still have reached the same conclusions under our framework. This was the case for the two diagnosable lineages within SEQ that require further genomic sampling in the unsampled gap. Having been alerted to this need already by Hammer et al. (2007), the present study has been able to demonstrate conclusively that the putative ADM taxon is genetically intermediate but undiagnosable from either the MDB or the TAS population (Table 1).

The results of the present study will help to stabilize the systematic framework for Australian smelt without the need to await the discovery of concordant phenotypic or ecological differences to validate these species, as recommended recently by Sukumaran & Knowles (2017). Although valuable, and recognizing the need to bring all available evidence to bear on the decisions, we suggest that phenotypic or ecological evidence is not necessary and is no longer realistic given the high prevalence of genuinely or apparently cryptic species (Adams et al., 2014; Struck et al., 2018), a steady decline in rates of species description and, in some groups (including the Australian fishes), the number of taxonomists (Pearson et al., 2011; Adams et al., 2013; Sangster & Luksenburg, 2015), the impracticality of integrating historical museum vouchers into genomics-led taxonomic revisions owing to their lack of data for one or more of the three required datasets (genetic/genomic, phenotypic/morphological and ecological), and the decline in and/or impossibility of collecting new voucher specimens in many groups (Dakota et al., 2017; Hope et al., 2018). The new data, with a refined analytical approach, will assist the reduced pool of taxonomists to validate candidate species formally, providing a level of qualified confidence to undertake taxonomic decisions. The imperative is high, given the current and projected increases in rates of species extinction (Toukhsati, 2018) and that the average time between species discovery and formal description is currently >20 years (Fontaine et al., 2012).

Genomic sequence data
Assessments of the number of diagnostic genetic markers among lineages are ideally suited to genomic datasets composed of unlinked, mostly codominant
loci, such as SNPs, but their application to genomic sequence data is clearly complicated by several issues. First, researchers will need to define the unlinked genetic markers under consideration. Second, they will need to decide how to define diagnosability at these individual loci. Finally, they will require access to software capable of undertaking pairwise comparisons of user-defined lineages to count the number of diagnosable loci detected, often across hundreds or thousands of these loci.

Given that our study has generated gene sequence data for two linked mitochondrial genes and two unlinked nuclear introns, we have presented some simple proof-of-concept analyses for *Retropinna* here (PCoA plots in Fig. 7; Supporting Information, Fig. S6; diagnosability measures in Table 1). As shown, the PCoA plots for the three unlinked loci (mtDNA, alpha-tropomysin and S7; separate plots in Fig. 7) recover the same primary and secondary aggregations as identified in the individual gene trees (Fig. 6), and the PCoA plot for the concatenated sequences (Supporting Information, Fig. S6) also produces clusters that inform the delineation of diagnosable lineages. Hopefully, this demonstration will encourage those relying on genomic

**Figure 7.** Scatterplots of ordination scores in the first two dimensions for the principal coordinates analyses (PCoAs) undertaken for the DNA sequence data. A, initial PCoA for mitochondrial DNA plus follow-up PCoAs of three composite clusters (red or blue boxes/envelopes/arrows). B, initial PCoA for alpha-tropomyosin. C, initial PCoA for S7. Axes are scaled to reflect their relative importance (as shown in parentheses) in explaining total multivariate variability. Symbols denote candidate species or lineages.
sequence data for species delineation to explore these issues more fully and/or enhance their MSC model analytical software to generate some of the key analyses currently not provided, such as measures of pairwise diagnosability among user-defined lineages across all user-defined genetic markers, and the ability to vary the tolerance levels for shared alleles/haplotypes when defining diagnosability.

**Diagnosability vs. Distinctiveness**

One of the core differences between our use of individual-based clustering methods (step 1, stepwise PCoA) and other studies adopting such approaches is that we require the process of identifying primary genetic clusters to be linked explicitly with assessments of whether the clusters thus identified are also fully diagnosable (step 3) rather than simply distinguishable on allele frequency profiles. Our approach places the concept of lineage diagnosability, based here on the presence of multiple fixed differences at unlinked genetic markers, at the heart of species delineation (Wiens & Servedio, 2000; Helbig et al., 2002; Adams et al., 2014; Georges et al., 2018; Unmack et al., 2019). This contrasts with other studies that have also incorporated multivariate ordination or a variety of other approaches, such as Bayesian assignment tests (e.g. using computer programs such as STRUCTURE and STRUCTURAMA), where the clusters thus defined are typically only matched qualitatively to those identified on the primary species tree (Carstens et al., 2013; Baumsteiger et al., 2017; Posso-Terranova & Andrés, 2018). Our focus on quantitative assessments of taxon diagnosability (summarized in Table 1) also argues for the use of assumption-free clustering approaches, such as ordination, rather than analyses, such as STRUCTURE, that delineate groups based on Hardy–Weinberg expectations. These latter analyses can delineate clusters even where they display only modest differences in allele frequency across a small percentage of loci (Georges et al., 2018) and are thus not diagnosable at the level required for species delineation (Wiens & Servedio, 2000; Helbig et al., 2002).

There are two reasons why we believe that even sympatric lineages ought to be diagnosable fully, rather than simply distinctive based on allele frequency profiles, before being considered different species in the absence of other compelling information (e.g. known reproductive incompatibility). First, all sound taxonomies require the existence of unlinked, highly heritable characters that not only diagnose the two species involved unequivocally but are also capable of identifying unsampled individuals as pure species A, pure species B, pure F₁ hybrid, or determining the approximate level of admixture (F₂, backcross, multigenerational hybrid, etc.). This expectation that a species diagnosis will also work on individuals, whether part of the original diagnosis or to be identified in the future, remains a key component of best practice in systematics for both sympatric and allopatric species (Edwards & Knowles, 2014). Second, these sympatric comparisons are central to the comparative or ‘yardstick’ approach that systematists often rely on as proxy measures of reproductive isolation when assessing the taxonomic status of allopatric lineages (Shelley et al., 2018; Singhal et al., 2018). Thus, it makes sense to use the same standards for diagnosability for all taxon comparisons (step 4), regardless of their comparative geographical distribution (step 5).

Although we advocate strongly that diagnosability ought to centre around the number of fully fixed or near-fixed differences as part of a conservative perspective on genomic species delineation, we acknowledge that some systematists favour less stringent criteria, particularly where distinctive lineages are sympatric or parapatric (Helbig et al., 2002; Hammer et al., 2013). Clearly, this is a topic that merits future debate. Nevertheless, by explicitly defining diagnosability in their genomic studies, researchers will allow others to assess objectively the strength of any assertion that two lineages are not conspecific based on genomic evidence.

**Sampling intensity and distributional gaps**

Step 4 in our framework involves explicit consideration of the sampling intensity used for geographical and genomic coverage. Although it is obviously desirable that both are sampled as intensively as possible, low geographical coverage (e.g. few individuals, small number of sample sites, poor representation of the distribution of a putative lineage) is generally a far more serious concern than is low coverage of the genome (Chambers & Hillis, 2020). This is because random sampling of the genome for neutral markers is unlikely to produce spurious primary lineages, whereas it will not be possible to sample a lineage genuinely using only a few sites unless that lineage itself is genetically homogeneous across its range (i.e. unlikely for the majority of species; Avise, 2000; Dynesius & Jansson, 2014). In addition, measures of both diagnosability and distinctiveness are particularly inaccurate, and often inflated, when only small numbers of individuals and sites are surveyed (Richardson et al., 1986; Baeverstock & Moritz, 1996; Georges et al., 2018). This point is illustrated in the Supporting Information (Fig. S7), which provides an empirical demonstration of how the percentages of fixed SNP differences between any two Retropinna species (here, CEQ and SEC, although the same result is obtained for other combinations) increase...
as sampling is reduced and escalate considerably when only single sites are compared.

As shown above, poor geographical coverage can inflate the apparent genetic distinctiveness of allopatric lineages considerably. Of more significance is the reality that the genetic and taxonomic affinities of any unsampled individuals that might occur in the gap between two allopatric lineages are usually some of the most crucial pieces of evidence required to determine conclusively whether these two lineages are valid species or distinctive phylogeographical groups that admix fully where they abut. For these two reasons, it is in the interests of everyone that researchers undertaking genomic species discovery discuss explicitly whether their gaps are real, unsampled or unable to be sampled (as per Fig. 3 and our step 4). As an example, species COO and MTV are believed to be fully allopatric, because no smelt has ever been collected from the intervening river basin (the Bulloo; Wager & Unmack, 2000). However, should they ever turn up in this gap, our study has provided 19 unlinked diagnostic markers (12 fixed SNPs, four allozyme loci, diagnostic nucleotides at each of our three DNA sequence datasets) that can, in combination, determine their comparative genetic and taxonomic status unequivocally.

CONCLUDING REMARKS

This study highlights how the inclusion of individual-based ordination plots, combined with assessments of lineage diagnosability, comparative lineage distribution and overall sampling intensity, together provide a framework that enhances our ability to assess independently which of the often many lineages identifiable on any species tree are well-supported candidate species vs. those do not meet a conservative standard of proof (Coates et al., 2018). Our proposals can be used to mine genomic datasets more thoroughly, given that they contain many more taxonomic insights than only those revealed in all the possible bifurcating trees generated using any method.

Adoption of our framework in no way stifles debate on any other theoretical and operational front. Thus, we support the efforts by others to improve tree-building software and to explore other tree-based approaches that have been proposed to detect and account for reticulate evolution and admixture/introgression, e.g. phylogenetic networks (Than et al., 2008; Solis-Lemus et al., 2017), fuzzy trees and TREEMIX (Bouckaert, 2010; Huson & Scornavacca, 2010; Pickrell et al., 2012). Instead, we hope that by encouraging future species discovery studies to adopt a common framework that explicitly presents a standard set of key biological parameters for others to review, this will both stimulate debate in general and allow group-specialist researchers to focus on any specific taxonomic controversies.

ACKNOWLEDGEMENTS

We thank the many people who assisted with sample collection and the South Australian Museum for access to their tissue collection. We are particularly indebted to the late Richard Norris, who led the bid for the Australian Collaborative Research Network for Murray-Darling Basin Futures that made this work possible. We also thank John A. Allen and an anonymous reviewer for their helpful comments on the penultimate version of this paper. This research was undertaken with the approval of the Animal Ethics Committee of the University of Canberra. This work was supported by the Australian Collaborative Research Network for Murray-Darling Basin Futures (A.G., B.G. and P.J.U.), the Cooperative Research Centre for Freshwater Ecology (A.G.), the Australian Research Council (LP140100521: P.J.U., B.G. and A.G.) and the Institute for Applied Ecology. The authors confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

DATA AVAILABILITY

The genomic data have been deposited in the Dryad repository (Georges et al., 2021), and all DNA sequences have been deposited in GenBank (accession numbers BankIt2464486: MZ301446–MZ301689; BankIt2464674: MZ301690–MZ301933; BankIt2464675: MZ301934–MZ302157; BankIt2464683: MZ302158–MZ302368).

REFERENCES


SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site:

Table S1. Sample size details and locality information, plus tissue and tree codes for all *Retropinna* sites surveyed. Site codes follow those used in Figure 2. Sites indicated by an asterisk are where two species were found in sympathy; sites labelled a and b indicate two samples from the same locality taken at different times. The superscript + for the single nucleotide polymorphism sample sizes indicates where individuals were discarded owing to a low call rate. Abbreviations: Alloz, sample sizes for the allozyme component of the present study; Hammer, sites and sample sizes as screened for allozyme variation by Hammer et al. (2007); Nmax, maximum number of individuals screened for at least one genetic/genomic technique; Tissue code, SA Museum freezer location; Tree code, locality code used in Supporting Information, Figs S2–S5).

Table S2. Outcomes of the fixed difference analysis. A, initial populations and sample sizes. B, count of fixed allelic differences between sympatric populations. C, putative contemporary hybrids identified visually in the principal coordinates analysis. D, populations with sample sizes of one were amalgamated manually with a second population in the same drainage. E, amalgamations of populations with no corroborated (tpop = 1) fixed allelic differences. F, further amalgamation of populations with five or fewer fixed differences. G, test of significance of aggregations. Note that COO could not be amalgamated reliably on lack of statistical significance because of ambiguity arising from non-transitivity. SECMacl+ and SECMackTimM were amalgamated. H, final set of diagnosable operational taxonomic units. See Figure 4.

Figure S1. Allozyme trees. A, summary unweighted pair group method with arithmetic mean tree for the original dataset of Hammer et al. (2007). B, neighbour-joining tree among the 51 sites surveyed within candidate species MTV. Sites are labelled and colour-highlighted by taxon (as in Fig. 2) plus site code (Supporting Information, Table S1). Sites not included in the regional allozyme study by Hammer et al. (2007) are also labelled with the # symbol.

Figure S2. Mid-point rooted RAxML tree for the concatenated sequences for 448 *Retropinna* based on the genomic dataset (11 980 single nucleotide polymorphisms). An asterisk indicates that the nodes received > 97% support.

Figure S3. Maximum likelihood tree for *Retropinna* based on analysis of mitochondrial cytochrome b gene. An asterisk indicates that the nodes received > 97% support. Each operational taxonomic unit (OTU) code is based on the population number and locality described in the Supporting Information (Table S1) and Figure 2.

Figure S4. Maximum likelihood tree for *Retropinna* based on analysis of the fifth intron of the alpha-tropomyosin nuclear gene. An asterisk indicates that the nodes received > 97% support. Each operational taxonomic unit (OTU) code is based on the population number and locality described in the Supporting Information (Table S1) and Figure 2.

Figure S5. Maximum likelihood tree for *Retropinna* based on analysis of the first intron of the S7 nuclear gene. An asterisk indicates that the nodes received > 97% support. Each operational taxonomic unit (OTU) code is based on the population number and locality described in the Supporting Information (Table S1) and Figure 2.

Figure S6. Scatterplots of ordination scores in the first two dimensions for the principal coordinates analyses (PCoAs) undertaken on the concatenated DNA sequence data for two mitochondrial DNA genes, alpha-tropomyosin and S7. Follow-up PCoAs for two composite clusters are shown in red boxes/envelopes/arrows. Dimensions are scaled to reflect their relative importance (as shown in parentheses) in explaining total multivariate variability. Symbols denote candidate species or lineages.

Figure S7. Relationship between diagnosability estimates for candidate species CEQ and SEC vs. the number of sites surveyed.

Figure S8. Schooling *Retropinna*. 