Multiple instances of paraplythetic species and cryptic taxa revealed by mitochondrial and nuclear RAD data for Calandrella larks (Aves: Alaudidae)

Martin Stervander a,b,*, Per Alström c,d,e,*, Urban Olsson f, Ulf Ottosson b, Bengt Hansson a, Staffan Bensch a

a Molecular Ecology and Evolution Lab, Dept of Biology, Lund University, Ecology Building, SE-223 62 Lund, Sweden
b AP Leventis Ornithological Research Institute, Laminga, Jos East, Jos, Plateau, Nigeria
c Department of Animal Ecology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18 D, SE-752 36 Uppsala, Sweden
d Swedish Species Information Centre, Swedish University of Agricultural Sciences, Box 7007, SE-750 07 Uppsala, Sweden
e Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
f Department of Biological and Environmental Sciences, University of Göteborg, Box 463, SE-405 30 Göteborg, Sweden

A R T I C L E   I N F O
Article history:
Received 9 October 2015
Revised 7 May 2016
Accepted 24 May 2016
Available online 26 May 2016

Keywords:
Cryptic taxa
Phylogeny
RAD
SNAPP
Museum specimens

A B S T R A C T
The avian genus Calandrella (larks) was recently suggested to be non-monophyletic, and was divided into two genera, of which Calandrella sensu stricto comprises 4–5 species in Eurasia and Africa. We analysed mitochondrial cytochrome b (cytb) and nuclear Restriction-site Associated DNA (RAD) sequences from all species, and for cytb we studied 21 of the 22 recognised subspecies, with the aim to clarify the phylogenetic relationships within the genus and to compare large-scale nuclear sequence patterns with a widely used mitochondrial marker. Cytb indicated deep splits among the currently recognised species, although it failed to support the interrelationships among most of these. It also revealed unexpected deep divergences within C. brachydactyla, C. blanfordi/C. erlangeri, C. cinerea, and C. acuirostris. It also suggested that both C. brachydactyla and C. blanfordi, as presently circumscribed, are paraphyletic. In contrast, most of the many subspecies of C. brachydactyla and C. cinerea were unsupported by cytb, although two populations of C. cinerea were found to be genetically distinct. The RAD data corroborated the cytb tree (for the smaller number of taxa analysed) and recovered strongly supported interspecific relationships. However, coalescence analyses of the RAD data, analysed in SNAPP both with and without an outgroup, received equally strong support for two conflicting topologies. We suggest that the tree rooted with an outgroup – which is not recommended for SNAPP – is more trustworthy, and suggest that the reliability of analyses performed without any outgroup species should be thoroughly evaluated. We also demonstrate that degraded museum samples can be phylogenetically informative in RAD analyses following careful bioinformatic treatment. We note that the genus Calandrella is in need of taxonomic revision.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction
The avian family Alaudidae, larks, comprises 93–97 species in 21 genera (Dickinson and Christidis, 2014; Gill and Donsker, 2014). Larks are found on all continents except Antarctica. Although the majority of the species occur in Africa, followed by Eurasia, only one species each occur in Australia and the Americas (de Juana et al., 2004; Dickinson and Christidis, 2014; Gill and Donsker, 2014). Alström et al. (2013) presented the first comprehensive phylogeny of the family, which was based on two mitochondrial and three nuclear loci for >80% of all species and representatives from all recognised genera (though not all loci were available for all species), as well as mitochondrial data for multiple subspecies. They found several non-monophyletic genera and multiple cases of unpredicted deep divergences among taxa that were treated as conspecific, as well as shallow splits between some recognised species. They concluded that “few groups of birds show the same level of disagreement between taxonomy based on morphology and phylogenetic relationships as inferred from DNA sequences”, and proposed a revised generic classification.

Although the notion of cryptic species is several centuries old, reports of cryptic diversity have dramatically increased after the
introduction of PCR methods (Bickford et al., 2007). The detection of cryptic diversity is of great importance to science as well as conservation, since the lack thereof severely can misguide conservation policy (Bickford et al., 2007; Trontelj and Fiser, 2009). Whereas cryptic diversity has been unravelled across the animal kingdom, the frequency at which it occurs varies (Trontelj and Fiser, 2009). Birds are a class offering relatively few such discoveries, and the rate of discovery of cryptic diversity is, for example, twice as high in mammals and three times higher in amphibians (Trontelj and Fiser, 2009). Most larks are cryptically coloured and patterned, and many of the species are renowned for being difficult to distinguish (de Juana et al., 2004). Surprisingly few taxonomic revisions have been undertaken using molecular and/or vocal data. Nearly all of those have proposed taxonomic splits (Alström, 1998; Ryan et al., 1998; Ryan and Bloomer, 1999; Guillaumet et al., 2005, 2006, 2008; Alström et al., 2013), except one recent paper that advocated the lumping of two species (Spottriswoode et al., 2013). As remarked by Alström et al. (2013), it seems likely that the number of presently recognised lark species is underestimated.

One of the lark genera suggested by Alström et al. (2013) to be non-monomophetic was Calandrella. This genus was separated into two non-sister clades: (1) C. cinerea, C. brachydyactyla and C. acutirostris in a clade sister to the genus Eremophila ('horned larks'), and (2) C. rutila, C. rufescens, C. chelensis and C. athens in a sister position to a clade comprising the two monotypic genera Cherophysalus and Eremalauda. Based on morphology, all of these relationships were totally unexpected. By priority, the name Calandrella was restricted to the first of these clades, while the available name Alaudala was resurrected for the second Calandrella clade. Alström et al. (2013) also found C. brachydyactyla dukhunensis to be anciently separated from the other C. brachydyactyla subspecies, and actually sister to C. acutirostris – again completely unpredicted based on morphological assessments.

The genus Calandrella (sensu Alström et al., 2013, subsequently followed by Dickinson and Christidis (2014) and Gill and Donsker (2014)) comprises four (Dickinson and Christidis, 2014) or five (Gill and Donsker, 2014) species. The taxonomy has been much debated over the years. Formerly, C. brachydyactyla was considered conspecific with C. cinerea under the latter name (e.g. Meinertzhagen, 1951; Vaurie, 1959; Mayr and Greenway, 1960). Based on widely allopatric distributions, different migratory behaviour and plumage differences, Vooos (1960) and Hall and Moreau (1970) proposed to treat these as separate species, and this has been followed by most subsequent authors (e.g. Glutz von Blotzheim and Bauer, 1985; Cramp, 1988; Sibley and Monroe, 1990; Keith et al., 1992). In addition, Mayr and Greenway (1960) separated C. blanfordi as a monotypic species; Sibley and Monroe (1990) agreed with this and also split off C. erlangeri (monotypic). In contrast, Dickinson and Christidis (2014) treated C. blanfordi as a polytypic species (with subspecies C. b. blanfordi in Eritrea; C. b. daarooodsiensis in northern Somalia; C. b. eremica in south-western Arabia; and C. b. erlangeri in the highlands of Ethiopia), whereas Gill and Donsker (2014) recognised C. blanfordi with three subspecies C. and C. erlangeri as a monotypic species. None of these authors gave any reasons for their treatments.

We here study all species and all except one of the subspecies in the genus Calandrella (sensu Alström et al., 2013) using full-length cytochrome b (cyt b) sequences for a sample of 46 individuals (all species but not all subspecies), a mix of full-length and short cyt b sequences for 114 individuals (all taxa) and Restriction-site Associated DNA (RAD) sequences for 12 individuals (all species, a few subspecies). We aim to clarify the phylogenetic relationships within the genus, to investigate the recently suggested paraphyly of C. brachydyactyla, and to compare analyses of large-scale nuclear sequence data with a mitochondrial marker.

2. Material and methods

2.1. Study group

We analysed all of the species in the genus Calandrella (sensu Alström et al., 2013) as well as all of the subspecies except C. brachydyactyla orientalis (Dickinson and Christidis, 2014; Gill and Donsker, 2014), in total 115 individuals (Table S1). Taxonomy follows Gill and Donsker (2014), except that we follow Ryan (2004) regarding the subspecies within C. cinerea. It should be noted that the taxonomic annotation of museum specimens is not always reliable, and we have classified the taxa according to current taxonomy (Ryan, 2004) and known distributions (Gombobaatar and Monks, 2011; BirdLife International and NatureServe, 2014; pers. obs.).

2.2. Laboratory procedures

Tissue samples were taken from toe pads of museum specimens at the Natural History Museum London (Tring; BMNH), the Royal Museum for Central Africa Tervuren (RMCA), the Natural History Museum of Denmark (ZMUC), the American Museum of Natural History (AMNH), and the University of Michigan Museum of Zoology (UMMZ).

Blood samples from live birds and tissue samples from dead embryos were extracted using standard phenol-chloroform protocol (Sambrook and Russell, 1989), whereas tissue samples from museum specimens were digested overnight at 55 °C in 100 μL lysis buffer (0.1 M Tris, 0.005 EDTA, 0.2% SDS, 0.2 M NaCl, pH 8.5) with 3–6 μL proteinase K (10 mg/ml) and then precipitated with ethanol.

We used Qiagen Multiplex PCR Kit (Qiagen Inc.), with amplification reactions containing 5 μL Qiagen Multiplex PCR Master Mix, 0.2 μl each of 10 μM forward and reverse primer (Table S2), 2 μL template DNA (5–10 ng/μl), and 2.6 μl water. We ran the PCR reactions for activation at 95 °C for 15 min; 30–45 three step cycles with denaturation at 94 °C for 30 s, annealing at varying temperatures for 90 s (Table S2), and extension at 72 °C for 90 s; final extension at 72 °C for 10 min. PCR products were checked on a 2% agarose gel, precipitated with NH₄Ac and ethanol, and then dissolved in 10–25 ml of water. We used 2 μl for sequencing of successful amplifications (BigDye sequencing kit; Applied Biosystems) in an ABI Prism 3100 capillary sequencer (Applied Biosystems).

The DNA sequence of the full cyt b gene was obtained from samples of wild-caught birds using amplification primers ND5-Syl (Stervander et al., 2015) and mtF-NP (Fregin et al., 2009), and internal sequencing primers Cytb_seq_H15541 (Stervander et al., 2015) and Cytb_seq_L15383 (this study; see Table S2). Owing to DNA degradation in the museum specimens, only shorter DNA fragments could be amplified, and we therefore designed specific primers based on the sequences of C. cinerea and C. brachydyactyla (Table S2). For some museum samples, only a 356 basepair (bp) fragment was sequenced, and specifically to type a number of C. acutirostris in Kashmir, a 107 bp diagnostic fragment was sequenced. PCR success was screened on 2% agarose gels.

We performed paired-end Restriction site-Associated DNA (RAD) sequencing on fresh samples representing C. a. acutirostris, C. a. tibetana, C. brachydyactyla dukhunensis, C. b. longipennis, C. b. rubiginosa, C. c. cinerea, C. c. williamsii, and C. c. saturator from Nigeria, as well as horned lark Eremophila alpestris as an outgroup. Library preparation followed Baird et al. (2008) and Etter et al. (2011) with the following modifications: We considered no RNase A treatment necessary; DNA fragments were sheared with a Bioruptor Standard UCD-200 (Diagenode) and selected to a size
range comprising 150–750 bp (length excluding adapters); PCR reactions were set up at 50 μl with 16 μl RAD library template running 16 cycles. Such protocol, with more library template and fewer cycles, leads to more P2–P2 fragments but fewer PCR clones. We used the restriction enzyme SfiI and the samples were multiplexed in a library with 16 samples (together with samples from other projects); the library was then sequenced on one Illumina HiSeq 2000 channel at BGI, Hong Kong.

Furthermore, we evaluated the inclusion of degraded museum samples of C. erlangeri and C. blanfordi daaroodensis into the library. While extractions had yielded reasonable DNA concentrations, an analysis with a DNA1000 chip for BioAnalyzer 2100 (Agilent) revealed that the C. erlangeri and C. blanfordi museum samples were heavily degraded, with few fragments exceeding 50 bp in length. Effectively, this made for a very low concentration of DNA that could be included in our RAD sequencing libraries together with DNA from fresh samples (with longer fragments), even without shearing. In order to be able to include the museum samples at all, we therefore modified the protocol as follows: The samples were first whole-genome amplified using a REPLi-g Mini Kit (Qiagen) and no shearing was performed.

2.3. Phylogenetic analyses of cytochrome b

All cyt b sequences were manually inspected, assembled, edited and aligned in Geneious v. 5–6 (Biomatters), and only acceptable sequences were retained (occasionally with ambiguity codes introduced for specifically problematic base calls). We created a full alignment (all 114 samples irrespective of sequence length; alignment FULL) and one alignment only containing sequences >900 bp (46 individuals; alignment LONG). As outgroup, we used either Horned Lark E. alpestris and Temincik’s Horned Lark E. bilopho, from the sister clade to Calandrella (Alström et al., 2013), or these two and three other closely related lark species pairs (Agulas Long-billed Lark Certhilauda brevirostris, Eastern Long-billed Lark C. semitorquata, Crested Lark Galerida cristata, Maghreb Lark G. macrorhyncha, Mongolian Lark Melanocorypha mongolica and Tibetan Lark M. tibetana), as well as the more distantly related Bearded Reedling Panurus biarmicus and Banded Prinia Prinia bairdii (cf. Alström et al., 2013).

Substitution models were evaluated with jModelTest v. 2.1.4 (Guindon and Gascuel, 2003; Darriba et al., 2012), selecting from 88 available models allowing for rate heterogeneity according to four gamma categories and for a proportion of invariable sites. Model selection was performed according to the Bayesian Information Criterion (BIC; Schwarz, 1978). The substitution model selected for both alignments was HKY (Hasegawa et al., 1985) with rate variation following a discrete gamma distribution with four rate categories (G; Yang, 1994) and with an estimated fraction of invariable sites (I; Gu et al., 1995).

Cyt b gene trees were computed within a Bayesian inference (BI) framework with BEAST v. 1.8.2 (Drummond et al., 2012), using a Birth-Death tree prior (Gernhard, 2008), and an uncorrelated relaxed molecular clock (Drummond et al., 2006) with a rate of 0.0105 ± 0.0005 substitution/site/lineage/my, based on overall cyt b substitution rates for a wide range of avian species (Weir and Schluter, 2008). We sampled trees every 1,000 generations, over 50 million generations, of which the first 25% were discarded as burn-in. Every analysis was repeated to ascertain congruence. The results were inspected using Tracer v. 1.6 (Rambaut et al., 2013), ensuring stationarity and effective sample sizes (ESS) of >300. Additionally, the cyt b gene was analysed by a maximum likelihood (ML) tree search (1000 replicates) and bootstrapping (1000 replicates) using RAxML version 8.2.4 (Stamatakis, 2014). GTRGAMMA was used for the bootstrapping phase and for the final tree inference.

2.4. Bioinformatic pipeline for nuclear RAD data

The paired raw reads were first filtered using the process_radtags and clone_filter components of Stacks v. 1.17 (Catchen et al., 2013). There is no genome assembly available for Calandrella larks, thus raw reads must either be assembled de novo, or mapped to a genome assembly of another species. The remarkably conserved syteny of the avian karyotype in most orders (Romanov et al., 2014; Zhang et al., 2014) allows for such interspecific mapping (e.g. Stervander et al., 2015). On the one hand, this results in excluding loci that are too diverged from the mapping target species. On the other hand, it effectively avoids problems of e.g. duplicated sequences, which can be much harder to detect when assembling raw RAD reads de novo. Furthermore, even though some genomic rearrangements are expected between different bird species, mapping towards a reference genome allows a coarse subset selection of SNPs based on genomic position. We thus mapped the filtered reads to the Zebra Finch Taeniopygia guttata genome (assembly taeGut3.2.4, downloaded from the UCSC genome website http://genome.ucsc.edu/) with Bowtie v. 2.2.2 (Langmead and Salzberg, 2012). Given the interspecific mapping we set maximum and minimum mismatch penalties (−mp) to 2 and 1, respectively. Further, we only accepted paired reads mapping in the expected relative direction within 150–1,000 bp from each other (−no-discordant, −no-mixed). From this point, two separate paths were followed for preparing data for concatenated sequence super-matrices and SNP-based analyses, respectively. For the sequence-based analyses, we removed read 2 from the resulting sam files and used read 1 as input to Stacks, following the ref_map track for phylogenetic and coalescent analyses. We used default parameters except for setting a higher minimum sequencing coverage to call a locus in a given sample (m = 4).

We filtered the output so that it only contained loci (i) in which all samples had been assigned 1–2 alleles; (ii) which contained 1–10 single nucleotide polymorphisms (SNPs); (iii) which comprised ≤12 alleles. The rationale for these filters was to eliminate any loci in which individuals had been assigned more than two alleles, and/or which were remarkably variable (cut-off values however assigned arbitrarily after inspecting data). By doing so, we may have traded off erroneous exclusion of highly variable loci in order not to include mis-assembled loci.

The Stacks output was then converted to biallelic sequences for all samples and loci, using custom scripts, and concatenated based on sample (but with random allele pairing) to long super-matrices. We included data from all high quality (fresh) samples in a super-matrix in which all individuals were included and no missing data at any locus was tolerated, comprising 38,038 loci with total length of 3.27 million bp (Mbp).

For SNP-based analyses, we merged all individual bam files (including museum samples) and performed variant calling with GATK. Index and dictionary files were prepared with PicardTools, then indels in the bam file were realigned using LeaAlignIndels. The UnifiedGenotyper was called and a limitation was set on maximum number of alternate alleles to 9 for the full dataset.

Filtering with VCFTools (Danecek et al., 2011) was performed separately for fresh samples and museum samples. Both were filtered based on variant type (SNP), but minimum sequencing depth (8) per genotype (individual and site) was required only for fresh samples, since sequencing success was generally low for museum samples. For the fresh samples, hereafter called dataset fresh, sites were further filtered to require coverage in all individuals, followed by requiring no more or less than two alleles (i.e. variable, biallelic sites). We prepared one dataset including the outgroup E. alpestris, and one dataset in which only the ingroup was retained.

Museum samples (C. blanfordi daaroodensis and C. erlangeri) were processed separately, and preliminary phylogenetic analyses
placed them at a distance 2.6–3.2 times further from other *Calan- 
drella* samples compared to that of the distance from the latter to 
the outgroup (*Eremophilina*). This likely demonstrates that the 
sequenced DNA contained many changed nucleotides, probably 
either owing to errors introduced during whole-genome amplifica-
tion or by post mortem DNA damage (data not shown). The latter 
can be caused by, for example, hydrolytic deamination, strand 
crosslinks, or oxidative nucleotide damage, all of which can alter 
the sequenced DNA (cf. Pääbo et al., 2004; Willerslev and Cooper, 
2005; Stiller et al., 2006). In order to make use of museum spec-
imens for RAD sequence data, despite these being affected by 
DNA damage, we took the following approach to remove polymor-
phic positions likely to be erroneous. SNPs from the dataset FRESH 
(including outgroup) were intersected with nucleotide data from 
museum samples and only those sites polymorphic in the dataset 
FRESH were retained. In this way we retrieved SNPs covered by all 
taxa and variable among fresh samples, and excluded SNPs unique 
to the museum samples but monomorphic in other samples. This 
procedure was performed separately for *C. blanfordi daarooodensis* 
(dataset DAAARODENSIS) and *C. eremica* (dataset ERLANGERI), as well as 
for the intersection for both (dataset MUSEUM).

2.5. Phylogenomic analyses of nuclear RAD data

The RAD sequence super-matrix was analysed in RAxML 
v. 8.0.26 (Stamatakis, 2014), using the GTRGAMMA substitution model 
(general time-reversible model allowing for rate heterogeneity 
according to a gamma distribution). The best maximum likelihood 
(ML) tree from 1,000 searches was selected, and 1,000 bootstrap 
replicates were run. Convergence was verified running a posteriori 
bootstrap tests using the autoMRE function (Pattengale et al., 2009).

The datasets DAAARODENSIS, ERLANGERI, and MUSEUM were thinned to retain SNPs at least 10 Kbp apart. The dataset FRESH was first divided into three non-overlapping subsets spread over the whole genome, and then each genomic subset was thinned by 50 Kbp. Since phase is unknown, the pairing of alleles between sites was randomised, and two replicates of different allele pairing was analysed for each genomic subset of dataset FRESH, while three replicates were analysed 
for datasets DAAARODENSIS, ERLANGERI, and MUSEUM.

All datasets, including genomic subsets and replicated random 
allele pairing, were imported into BEAUti (Bouckaert et al., 2014), 
where the data were prepared for analyses with the SNAPP v. 
1.1.16 plugin (Bryant et al., 2012) in BEAST v. 2.1.3 (Bouckaert 
et al., 2014). The priors for forward (u) and reverse (v) mutation rates were set to be estimated and the remaining parameters were 
left at default values, e.g. species divergence rate λ = 0.00765, and θ 
deﬁned by a γ prior with shape parameter α = 11.750 and scale 
parameter β = 109.73. Runs were carried out for ≥1.5 million gener-
erations, sampling every 1,000 generations. Output was inspected 
with Tracer v. 1.6 (Rambaut et al., 2013), ensuring stationarity 
and ESS >200 for all parameters (with a few exceptions for single 
theta values of internal branches, in single replicates). We first 
alysed the data using each sample as a terminal taxon, in order 
to verify whether there were any deviations in topology between 
the nuclear and the mitochondrial trees, and then reran the analy-
ses with clades A (*C. brachydactyla*) and D1 (Nigerian *C. cinerea satu-
ration*; both estimated with cyt b being <600 ky old) collapsed into 
tree-tip taxa.

2.6. Within-lineage genetic variation

SNP heterozygosity was calculated, for individuals sampled 
from fresh material, as the proportion of heterozygous sites in 
dataset FRESH, comprising 215,996 high quality SNPs (full coverage
and a minimum sequence depth of ≥8 in each individual), which 
were variable among fresh samples (including outgroup). For the 
museum samples, this estimate was based on dataset MUSEUM, con-
taining 11,684 SNPs, after verifying that the heterozygosity levels 
for fresh samples were comparable to those obtained from the larger dataset FRESH.

Theta was co-estimated with the species tree in SNAPP runs. We 
combined multiple replicate runs for maximising ESS, all of which 
were ensured to be >200. Theta co-estimation was done for runs in 
which each individual was treated as a terminal taxon, as well as 
for runs in which clades D1 and A (see Section 3) were grouped.

3. Results

3.1. Phylogenetic analyses of cytochrome b

Both BI and ML analyses based on the full-length sequence data 
(Fig. 2) recovered a cyt b tree with seven strongly supported 
primary clades (labeled A–G). Sister relationships between clades 
B + C, D + E and F + G, respectively, were strongly supported, but 
the other, more ancient, relationships were unsupported. The tree 
shown in Fig. 2 resulted from a BI analysis where clade ABC was 
constrained based on the RAD data (see below); unconstrained 
analysis of these data resulted in a sister relationship between 
clades A and DE, with very poor support (BI posterior probability 
(PP)/ML bootstrap proportion (MLBS) <0.5; not shown).

Clade A comprised all *C. brachydactyla* subspecies except *C. b. 
dukhunensis*. Although the divergences within this clade were very 
shallow, it was subdivided into a western clade (A1: NW Africa – 
*C. b. rubiginosa*; SW Europe – *C. b. brachydactyla*) and an eastern clade 
(A2: Kazakhstan, Uzbekistan, Somalia [winter] and India [winter] – 
*C. b. longipennis*), with one Greek sample (*C. b. brachydactyla*) in 
each of these two clades. The placement of one *C. b. longipennis* 
sample (64_KAZ) in the ML tree was inconclusive with regard to 
clades A1 and A2.

Clade B comprised the Asian *C. acutirostris*. There was a deep 
split, estimated to 1.5 million years ago (mya) [95% highest poste-
rior density (HPD) 0.9–2.2 mya], between the samples from the 
Tibetan plateau (China and India; B1) and Afghanistan (B2), respec-
tively. Clade C contained our single sample of *C. brachydactyla 
dukhunensis*, with an estimated divergence time from its sister 
clade, B, at 3.1 mya (95% HPD 2.1–4.2 mya).

Clades D and E contained the sub-Saharan *C. cinerea*, with a 
deep split, estimated at 2.7 mya (95% HPD 1.7–3.7 mya) that separ-
ated the Nigerian *C. c. saturatior* (D1) and Kenyan *C. c. williamsi 
(D2) from the South African *C. c. cinerea* and DR Congo *C. c. satura-
tior* (E). The divergence time between clades D1 and D2 was con-
siderably younger (0.8 mya, 95% HPD 0.4–1.3 mya).

Clade F included *C. b. blanfordi* (N Eritrea; F1) and *C. eremica* 
(Ethiopia; F2), separated from clade G, which contained *C. b. eremica* 
(Arabia; G1; one sample (44_SAU) inconclusively placed 
within clade G in ML analyses) and *C. b. daarooodensis* (N Somalia; 
G2), 4.3 mya (95% HPD 3.0–5.6 mya).

The cyt b tree including all sequences, i.e. a mix of full-length 
and short sequences (Fig. 3), recovered the same seven well-
supported clades as in the tree based on only the full-length 
sequences, as well as additional deep splits within clade B. Like 
in Fig. 2, the *C. brachydactyla* clade (A) was divided into a western 
(A1) and an eastern (A2) clade. In addition to the previously men-
tioned subspecies, clade A1 also included a sample from Hungary, 
presumably representing *C. b. hungarica*, and clade A2 also 
included *C. b. armemisiana* (Georgia, Iran), *C. b. hermonensis* (Leba-
non), a presumed *C. b. woltersi* (E Turkey), and a *C. b. brachydactyla* 
from W Turkey.

Clade B was divided into more clades than in Fig. 2. Clade B1 
contained three short sequence samples from India, whereas the
full-length Indian sample that fell in this clade in Fig. 2 (110_IND) was now placed in clade B4 (though lacking support). Clade B2 included six additional Afghan samples. Both the new clades B3 and B4 were comprised exclusively of samples from the Indian part of the Tibetan plateau. Clades B2 and B3 were sisters, with insignificant support (posterior probability [PP] 0.89), and had an estimated divergence at 1.5 mya (95% HPD 0.7–2.4 mya). The age of the divergence between these two pairs of clades was estimated to be 2.5 mya (95% HPD 0.5–2.5 mya). Clades B1 and B4 were sisters with strong support (PP 0.99) and an estimated split at 1.5 mya (95% HPD 0.7–2.4 mya). The age of the divergence between these two pairs of clades was estimated to be 2.5 mya (95% HPD 1.4–3.7 mya). Clade C included five samples of C. brachydactyla from the winter quarters (C India, Myanmar) in addition to the one from the breeding grounds included in Fig. 2. The estimated split between clades B and C was older than that estimated in Fig. 2 (3.8 mya, 95% HPD 2.3–5.5 mya).

The position of clade D was as in Fig. 2, with the addition of two samples of C. cinerea williamsi from Kenya in clade D2. Clade E contained, in addition to the samples in Fig. 2, C. alluvia and C. millardi from Botswana, C. f. fulvida from Zambia and Zimbabwe, C. c. niveni from South Africa, C. c. saturator from Angola, Malawi, Tanzania and Uganda, and C. c. spleniate from Namibia. Clades F and G agreed with Fig. 2 (same sequences).

3.2. Phylogenomic analyses of nuclear RAD data

Using single samples as terminal taxa, the SNAPP analyses did not deviate from clade assignment in the cyt analyses (Fig. S1), and we therefore continued the analyses by treating all C. brachydactyla samples except the deviating C. b. dukhunensis and both Nigerian C. cinerea saturatior samples as single taxa. The SNAPP analyses of the dataset FRESH, which included only the fresh Calandrella samples (i.e. all of the species except C. blanfordi and C. erlangeri) and with E. alpestris as outgroup (26,629 SNPs divided evenly over three genomic subsets, in turn run as two replicates with different random allele pairing), consistently recovered a tree (Fig. 4) that was largely in agreement with the cyt tree. However, unlike the cyt tree, it strongly (PP 1.0) supported a sister relationship between clade A (containing C. brachydactyla, excluding C. b. dukhunensis) and clade BC (comprising C. acutirostris and C. brachydactyla dukhunensis) (labeling as in Figs. 2 and 3). In contrast, the analyses of the same data but excluding the outgroup species (20,051 SNPs [after the exclusion of 6,578 sites that differed between E. alpestris and Calandrella but were monomorphic in Calandrella] over three genomic subsets) consistently recovered clades A and DE as sisters, with strong support (PP 1.0) (Fig. 4).

The SNAPP analyses of the dataset MUSEUM (1,590 SNPs), including all Calandrella species, i.e. also samples obtained from museum specimens of C. blanfordi and C. erlangeri, consistently recovered clade FG as sister to the other Calandrella species with strong support (PP 1.0) in two out of three replicates (Fig. 5a). Analyses of the larger dataset DAAROODENSIS (4,165 SNPs), containing C. blanfordi daaroodensis (G2), resulted in a tree with the same basic topology in one of the three replicates (Fig. 5b). However, the remaining MUSEUM and DAAROODENSIS replicates, as well as all replicates for dataset ERLANGERI (4,343 SNPs) containing C. erlangeri (F2), resulted in trees that placed the museum taxa as sisters to the outgroup...
E. alpestris (Fig. 5c). Thus, it should be noted, that the museum sequences still seem to suffer from erroneous bases due to DNA damage.

In all of the SNAPP analyses of the RAD data the relative branch lengths differed markedly from the cyt b trees (Figs 2 and 3), with clades DE and FG being considerably shallower. Divergences within...
Fig. 3. Relationships within the genus *Calandrella* based on a dataset comprising both full-length and short mitochondrial cytochrome *b* sequences, analysed under a lognormal relaxed clock with a rate of 2.1%/million years. Posterior probabilities (PPs) are indicated at the nodes; * means PP 1.00, and – means PP < 0.50. Clade labels and colours as in Figs. 2, 4 and 5 for simplified comparison. # indicates full length sample (also used in analysis in Fig. 2).
clade A, when individuals were treated as terminal taxa, were markedly deeper compared to clade BC (Fig. S1). Moreover, within clade A the relationship among *C. b. brachydactyla*, *C. b. rubiginosa* and *C. b. longipennis* varied among different analyses, with short internode distances and long terminal branches (Fig. S1).

The RAxML bootstrap analysis of the RAD sequence super-matrix based on the fresh samples (including *E. alpestris* as outgroup; 3.27 Mbp) recovered the same topology as the SNAPP tree with 100% support for all clades (see Fig. 4).

3.3. Within-lineage genetic variation

Calculated for the individual samples, the SNP heterozygosity was almost twice as high (0.13–0.14) in *C. brachydactyla* (excluding *C. b. dukhunensis*) as in the samples from all the other taxa (0.03–0.09) (Table 1). There was considerably greater variation in theta: lowest in the African clades DE (0.012–0.033) and FG (0.015–0.016), intermediate in the Asian clade BC (0.025–0.041), and highest in *C. brachydactyla* (clade A; 0.067–0.087) (Table 1). When collapsed to a single terminal taxon, clade A had a particularly high estimated theta (0.331; Table 1).

4. Discussion

4.1. Interspecific relationships and dating

*Cytb* supported deep splits among the currently recognised species, although it failed to support the interrelationships among most of these. It also inferred deep divergences within *C. cinerea* and *C. acutirostris*, and two non-monophyletic species, namely *C. brachydactyla* and *C. blanfordi*. The RAD data strongly supported sister relationships between *C. acutirostris* and *C. brachydactyla dukhunensis* (clade BC), between this clade and the other subspecies of *C. brachydactyla* (clade ABC), and between clade FG (including only representatives for clades F2 and G2) and the rest of the species (clade DE).

The divergence times inferred from the *cytb* data had large confidence intervals, and are based on a molecular clock, and should therefore be interpreted with caution. This is particularly true for the deepest nodes that were associated with low posterior probabilities, as well as the divergence times in Fig. 3, which were based on an analysis in which the majority of the sequences were incomplete. However, as the divergence times were estimated under the best-fit model, HKY + G + I, rather than the GTR + G model used by Weir and Schluter (2008) in their analysis that confirmed a mean *cytb* rate of 2.1%/my across 74 avian calibrations, our time estimates are likely to be underestimates rather than overestimates. The overall somewhat younger divergences compared to those estimated by Alström et al. (2013) are also likely the result of the use of different models (GTR + G in Alström et al. (2013)).

4.2. Intraspecific variation and divergence

In the *cytb* tree, the *C. brachydactyla* clade (A) showed very shallow divergence across its large range, here sampled from Morocco to eastern Kazakhstan. Still, there was evidence of geographical
structuring, with western and eastern birds in separate clades (A1 and A2, respectively). The only signs of geographical overlap between these two clades were from Crete, where two birds representing each of these two clades had been collected in the latter half of May. Although we have no proof that these two birds were breeding there, the chance of one or both of them being migrants on their way to breeding grounds further northeast and passing through Crete that late in spring seems slight (cf. Cramp, 1988; de Juana et al., 2004). In contrast to the cytb data, the RAD data failed to support any geographic structuring within *C. brachydactyla* (clade A), likely due to incomplete lineage sorting, which is more common in nuclear markers than in mitochondrial as a result of the four times larger effective population size in nuclear compared to mitochondrial DNA (see further Section 4.3, below). *C. brachydactyla* species, and by far exhibits the largest genetic diversity (Table 1).

The deep divergence between *C. b. dukhunensis* (clade C) and the other *C. brachydactyla* subspecies (clade A), and the sister
relationship between the former and *C. acutirostris* (clade B), was first found by Alström et al. (2013). They used the same *C. b. dukhunensis* sample from the Mongolian breeding grounds as analysed here (73_MNG). Their finding was corroborated by additional samples of *C. b. dukhunensis* from the winter quarters in Myanmar and eastern India and, more importantly, by RAD data from the same breeding site bird (73_MNG). The RAD data refute that the pattern might be caused by mitochondrial introgression. The taxonomic status of *C. b. dukhunensis* is in obvious need of revision.

The deep splits inferred within *Calandrella acutirostris* in the cytB tree were unexpected. The estimated age of clade B differed between the analyses in Figs. 2 and 3 (mean 1.5 mya and 2.6 mya, respectively). The more ancient separation in the latter tree is likely due to the addition of samples that suggest deep splits within Indian birds. This may also be the main reason for the older age of clade BC in Fig. 3 than in Fig. 2 (3.9 mya and 3.1 mya, respectively), perhaps in combination with the addition of more *dukhunensis* samples in the former analysis. Further studies are warranted to clarify the relationship between the clades and the timing of divergence.

The phylogeographic pattern within clade DE in the cytB tree was unexpected, with a total lack of geographical structure throughout almost the entire distribution of *C. cinerea*, from South Africa to DR Congo, Uganda, and southern Tanzania (clade E) and comprising seven different subspecies. In contrast, clades D (Nigeria and Kenya) and E were estimated to have separated around 2.5 mya, and there was a deep split between the geographically isolated Nigerian and Kenyan populations (clade D1 and D2, respectively). The ancient divergence between the samples from Uganda (C. c. saturatior) and Kenya (C. c. williamsi), on either side of Lake Victoria, are particularly noteworthy. It should be noted that the Nigerian population (clade D1) is treated as belonging to *C. c. saturatior*, as are the samples from Angola, DR Congo, Malawi, Tanzania, and Uganda (clade E) (Ryan, 2004; Dickinson and Christidis, 2014). This calls for a taxonomic revision. The RAD data supported the topology of the cytB tree for the four samples included in the RAD analysis. It is noteworthy that the Nigerian samples of *C. c. saturatior* displays intermediate genetic diversity (Table 1), despite seemingly very restricted current range and population size (pers. obs.).

The old age of clade FG, and the approximately 4 my old divergence between clades F (C. blanfordi eremica and C. b. darrowensis) and G (C. erlangeri and C. blanfordi blanfordi) were totally unexpected. These deep splits among the Horn of Africa/Arabian taxa stand in strong contrast to the lack of divergence within the much larger geographical area covered by the taxa in clade E, although it resembles (though is much older than) the separation of the geographically close populations from opposite sides of Lake Victoria (in clades D2 and E, respectively). The taxonomy of the two species in clade FG is in need of revision. The genetic diversity measured in *C. b. darrowensis* and *C. erlangeri* was low (Table 1). This is compatible with their rather small geographical distributions, but our estimates may be somewhat deflated owing to less power to detect heterozygote positions due to low sequencing coverage.

Trontelj and Fiser (2009) pointed out that meaningful investigation of cryptic diversity should be undertaken at genus level, as we have done with extensive sampling, and that information about any resulting cryptic lineages are of value for conservation policy. We have demonstrated the distinctiveness, both of previously unknown lineages (within *C. cinerea* and *C. acutirostris*) and between deep lineages that have previously been classified as subspecies. Pending a taxonomic revision, this information may prove very important for the management and conservation of e.g. the small and restricted Nigerian population of what is presently treated as *C. cinerea saturatior*.

### 4.3. SNAPP analyses of RAD data

Individual gene trees frequently differ among each other as well as from the phylogeny they are contained within (reviews in e.g. Avise, 1994, 2000; Maddison, 1997; Page and Charleston, 1998; Funk and Omland, 2003; Degnan and Rosenberg, 2009). Moreover, mitochondrial genes are particularly prone to introgress across species boundaries, at least during early stages of divergence (e.g. Funk and Omland, 2003; Chan and Levin, 2005; cf. e.g. Alström et al., 2008; Irwin et al., 2009; Wang et al., 2014). Hence, any single-locus tree needs to be corroborated by analyses of independent loci. In the present study, the RAD data provided independent testing of the cytB hypothesis, although RAD data were only available for a small number of individuals. In particular, the non-monophyly of *C. brachyactyla* first reported by Alström et al. (2013), and the deep divergence between clades F and G and between clade FG and the others found in our cytB trees, were corroborated by the RAD data. Moreover, as expected, the RAD dataset, which was vastly larger in terms of number of nucleotides than the cytB matrix, resolved the deeper nodes much better than cytB.

Concatenation, as in the present RAxML analysis of the RAD data, has been suggested to be statistically inconsistent under some circumstances (e.g. Kolaczkowski and Thornton, 2004; Kubatko and Degnan, 2007; Degnan and Rosenberg, 2006), but is still used without evaluation against other methods (e.g. Eaton and Ree, 2013; Wagner et al., 2013; Cruaud et al., 2014). To deal with the inconsistency problem, multiple “species tree methods” have been proposed. Some of these (e.g. BEST, Liu and Pearl, 2007; Liu, 2008; BEAST, Heled and Drummond, 2010; MP-EST, Liu et al., 2010) are based on the multispecies coalescent model (e.g. Nielsen, 1998; Rannala and Yang, 2003; Degnan and Salter, 2005; Liu and Pearl, 2007), which assumes that gene trees evolve under coalescent processes in every branch of a species phylogeny. Unlike previous algorithms integrating the species tree in parallel with marker trees (e.g. Liu, 2008; Heled and Drummond, 2010), SNAPP (Bryant et al., 2012) is tailored for SNP datasets and integrates a coalescent species tree without inferring marker trees. Unlike multi-marker sequence data SNPs cannot easily be partitioned and indeed SNAPP assumes unlinked markers. Consequently, to decrease linkage SNPs should first be thinned based on their physical position, as we did by 50 Kbp (10 Kbp for the small datasets including museum samples). As the results from the SNAPP analyses agreed with the RAxML of the concatenated SNPs, concatenation is probably not compromised for these data (but compare with the inconsistencies found by Stervander et al. (2015)).

SNAPP will sample the root position along with the rest of the nodes in the tree and therefore, in theory, does not require an outgroup to root the tree (Bryant et al., 2012). This view is also presented in the instructions for sequence-based analyses in BEAST v. 2 (http://beast2.org/faq/). However, our analyses of the dataset FRESH, with and without an outgroup, respectively, received equally strong support (PP 1.0) for two conflicting topologies [(outgroup, ((A, BC), DE)) vs. ((A, DE), BC)]. From both a biogeographical and morphological viewpoint, the former seems more likely (cf. Fig. 1 and de Juana et al., 2004)), and this best reflects the general sequence (SNP) similarity: clade A–BC 70.5–72.6% vs. clade A–DE 67.4–68.4%. We suggest that the use of an outgroup produces a more reliable result, at least for datasets which contain short internode distances close to the root. Coalescence-based analyses of SNP datasets are often very computationally demanding and time-consuming, and it may therefore be compelling to follow the recommendation by Bryant et al. (2012), not to include an outgroup. Indeed, adding an outgroup to the analyses does not only
increase the taxon matrix by one, but it will perhaps double the number of SNPs (owing to the inherent longer evolutionary distance between the outgroup and the ingroup, than between any of the ingroup taxa). In practice, this will often incur a trade-off, where the inclusion of an outgroup may limit the number of SNPs that can be included in analyses. Based on our results, we suggest that the reliability of analyses without outgroup be thoroughly evaluated.

The branch lengths differed strongly between the cyt b and RAD trees. This might represent true differences between these datasets. As has already been remarked, different loci may have different histories. However, it seems unlikely that C. brachydactyla, which is the most northerly distributed and most highly migratory species and which shows an extremely shallow population structure in the cyt b tree (excluding C. b. dukhunensis), actually has the deepest intraspecific divergence of all species in nuclear DNA (Fig. S1). It seems more likely that this pattern results from random effects due to small sample sizes for all taxa, which prevents reliable estimates of effective population sizes and hence branch lengths. The higher heterozygosity and estimated theta within C. brachydactyla than in any of the other taxa in the RAD data (Table 1) might be the main reason for the anomalous pattern, but larger samples for all taxa are needed to evaluate whether this is consistent or not.

4.4. Use of degraded DNA from museum samples

We tested including degraded DNA from museum samples in our RAD libraries, by using whole-genome amplification and applying no shearing. This seems to have introduced errors in the DNA sequences. Posterior to our library construction, other studies have successfully amplified RAD tags from museum material, with one important difference – they made separate libraries with museum material only, which allowed much shorter fragment lengths (e.g. Tin et al., 2014) than our target, which was optimized for the fresh samples.

However, despite these problems, we attempted to make use of the data through a process of careful bioinformatic scrutiny. After all putative erroneous nucleotides had been excluded from the degraded museum specimens representing clade FG from the Horn of Africa, we could place that clade basal within the genus (Fig. 5). In the attempt to exclude erroneous nucleotides, we acknowledge the risk to also exclude true variants unique to clade FG, however that risk may be balanced by the risk of included erroneous nucleotides among the sites variable among the fresh samples.

Lastly, while traditional RAD sequencing is an option, the new hyRAD method (Suchan et al., 2016) and other NGS approaches such as whole-genome shotgun sequencing and sequence capture approaches are available and may be preferable for the degraded DNA from museum samples (reviewed by Burrell et al. (2015)).

5. Conclusions

Although the mitochondrial cyt b and nuclear RAD data differed strongly in number of nucleotides (≤1,143 bp vs. ≤3,270,000 bp) and taxa (22 vs. 10), they agreed well. As expected, the RAD data provided better resolution and support for deeper nodes. Moreover, representing a vastly larger proportion of the genome, which was analysed using a coalescent based method (SNAPP), the RAD data should provide a more robust estimate of the phylogeny of the genus Calandrella than cyt b. The pronounced differences in relative branch lengths between cyt b and RAD trees that we noted need to be evaluated by more densely sampled RAD data. Suggestions that outgroup rooting is superfluous in SNAPP (and related methods, e.g. “BEAST) need to be evaluated. For our data, we suggest that analyses without an outgroup resulted in a strongly supported but incorrect tree. We showed that precious museum specimens can be successfully used in RAD sequencing analyses, although they should preferably be prepared in separate libraries. Nevertheless, we managed to include them in our analyses despite heavy degradation and DNA damage. In the genus Calandrella, we discovered several cryptic lineages within C. acutirostris and C. cinerea, whereas many described subspecies were not genetically distinct; the family is thus in need of taxonomic revision, which we will address elsewhere.

Author contributions

The study was conceived in two separate processes by M.S., U.Ot., B.H., S.B., and P.A. and U.Ol. respectively. The parts were merged, and its final design was determined with M.S. from input from all other authors; P.A., U.Ol., and U.Ot. carried out the field work, whereas museum sample collection was done by M.S., U.Ol., and U.Ot.; M.S. performed the major parts of the lab work, with additions from U.Ol.; M.S. and P.A. performed all analyses with input from B.H. and S.B.; P.A. and M.S. wrote the paper with input from the other authors, all of which have approved of the final version.

Acknowledgements

We would like to thank Michel Louette, Alain Reygel and Gael Carin (RMCA), Mark Adams, Robert Prys-Jones, and Hein van Grouw (BMNH), Paul Sweet, Peter Capainolo, and Thomas J. Trombone (AMNH), Janet Hinshaw (UMMZ), Jon Fjeldså and Jan Bolding Kristensen (ZMUk), and Sharon Birks (UWBM) for providing data and museum samples for genetic analyses. Yahkat Barshep, Mark Hulme, Nick Horrocks, Jesus Garcia, Pierre-André Crochet, Claire Spottiswoode, Jan Lifjeld, Paul J. Leader, and Stephane Ostrowski provided samples from the field, while Keith Barnes and Paulette Bloomer provided unpublished sequences. Naomi Keenhen and Peter Pruisscher were instrumental in creating the RAD sequencing library, and are warmly thanked. Robert Prys-Jones and Pamela C. Rasmussen kindly shared their expertise in scrutinizing specimens from R. Meierzhagen’s collections, in order to validate that these were likely genuine. Martin Trestedt gave advice on DNA extraction from ancient tissues, and Martin Larsson and Linus Hedh assisted in some Sanger sequencing. We thank Dennis Hasselquist and one anonymous reviewer for comments on the manuscript. M.S. has received grants from the Royal Physiographic Society and BirdLife Sweden, P.A. and U.O were supported by the Sound Approach and the Jornwall Foundation; B.H. by the Swedish Research Council, the Crafoord Foundation, and EU-FP7 (project Avian Genomics); and S.B. by the Swedish Research Council. This is contribution no. 109 from A. P. Leventis Ornithological Research Institute.

Appendix A. Supplementary material

Sanger sequences >200 bp are deposited at GenBank with accession numbers KX379897–KX379995. Sanger and RAD datasets are available at DataDryad, see http://dx.doi.org/10.5061/dryad.16mg6. Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2016.05.032.

References


