# The systematics and evolution of the Sri Lankan rainforest land snail Corilla : new insights from RADseq-based phylogenetics

# The systematics and evolution of the Sri Lankan rainforest land snail *Corilla*: new insights from RADseq-based phylogenetics

D.C. Raheem[a,b,*], D.J. Gower[b], K. Breugelmans[c], K.B. Ranawana[d], T. Backeljau[c,e]

[a] Department of Biological Sciences, Faculty of Applied Sciences, Rajarata University of Sri Lanka, Mihintale 50300, Sri Lanka
[b] Department of Life Sciences, Natural History Museum, London SW7 5BD, UK
[c] Royal Belgian Institute of Natural Sciences, Vautierstraat 29, B-1000 Brussels, Belgium
[d] Department of Zoology, University of Peradeniya, Peradeniya, Sri Lanka
[e] Evolutionary Ecology Group, University of Antwerp, Universiteitsplein 1, B-2610 Antwerp, Belgium

* Corresponding author. E-mail address: draheem@as.rjt.ac.lk (D.C. Raheem)

Keywords:
*Corilla*
Land snail
RADseq
Phylogeny
Rainforest
Sri Lanka
Stylommatophora

## Abstract

The stylommatophoran land-snail genus *Corilla* is endemic to Sri Lanka and India's Western Ghats. On the basis of habitat distribution and shell morphology, the 10 extant Sri Lankan species fall into two distinct groups, lowland and montane. Here, we use phylogenetic analyses of restriction-site-associated DNA sequencing (RADseq) data and ancestral-state reconstructions of habitat association and shell morphology to clarify the systematics and evolution of Sri Lankan *Corilla*. Our dataset consists of 9 species of *Corilla*. Phylogenetic analyses were based on 88 assemblies (9,604–4,132,850 bp) generated by the RADseq assembler ipyrad, using four parameter combinations and different levels of missing data. Trees were inferred using a maximum likelihood (ML) approach. Ancestral states were reconstructed using maximum parsimony (MP) and ML approaches, with 1 binary state character analysed for habitat association (lowland *vs* montane) and 6 binary state characters analysed for shell morphology (shape, colour, lip width, length of upper palatal folds, orientation of upper palatal folds and collabral sculpture). Over a wide range of missing data (40–87% missing individuals per locus) and assembly sizes (62,279–4,132,850 bp), nearly all trees conformed to one of two topologies (A and B), most relationships were strongly supported and total branch support approached the maximal value. Apart from the position of *Corilla odontophora* 'south', topologies A and B showed similar, well-resolved relationships at and above the species level. Our study agrees with the shell-based taxonomy of *C. adamsi*, *C. beddomea*e, *C. carabinata*, *C. colletti* and *C. humberti* (all maximally supported as monophyletic species). It shows *C. erronea* and *C. fryae* constitute a single relatively widespread species (for which the valid name is *C. erronea*) and that the names *C. gudei* and *C. odontophora* each apply to at least two distinct, yet conchologically-cryptic species. The MP and ML ancestral-state reconstructions yielded broadly similar results and provide firm evidence that diversification in Sri Lankan *Corilla* has involved evolutionary convergence in the shell morphology of lowland lineages, with a pale shell and wide lip having evolved on at least two separate occasions (in *C. carabinata* and *C. colletti*) from montane ancestors having a dark, narrow-lipped shell.

## 1. Introduction

Sri Lanka and the Western Ghats of India harbour a highly threatened biota that is rich in taxa endemic to this region and is one of 36 global biodiversity hotspots(https://www.cepf.net/our-work/biodiversity-hotspots). Our understanding of the evolution and diversification of this biota is at a comparatively early stage and this is particularly the case for most invertebrate taxa, including land snails (but see Beenaerts et al., 2010; Joshi and Edgecombe, 2018; Jaiswara et al., 2019; Bhosale et al., 2021). The land-snail fauna of the Western Ghats–Sri Lanka hotspot is phylogenetically diverse and is dominated by species endemic to the region, many of them belonging to genera endemic or largely endemic to this region (Raheem et al., 2014). Among the genera endemic to the hotspot is the stylommatophoran genus *Corilla* Adams & Adams, 1858, the sole representative of the family Corillidae and a putative Gondwanan relict (Naggs and Raheem, 2005). The genus comprises 11 extant and currently recognized species, of which 10 are endemic to Sri Lanka and 1 to the Western Ghats (Gude, 1914; Barnacle, 1956).

Sri Lankan *Corilla* are highly restricted in their distribution, inhabiting tropical rainforest (lowland, submontane and montane rainforest) and moist monsoon forest (*sensu* Legg and Jewell, 1995), where they are usually found on the forest floor, among leaf litter and decaying wood. The current species-level taxonomy for *Corilla*, which largely follows Gude (1914), is based entirely on shell morphological characters: shell shape and size, the relative expansion of the lip, the arrangement of the folds inside the shell aperture (apertural dentition) and shell sculpture. On the basis of shell morphology and habitat distribution, Sri Lankan *Corilla* can be divided into two well-defined groups, lowland and montane (**Figs. 1, 2**). Lowland species have wide-lipped shells that are buff or rufous brown and are associated with lowland rainforest and/or moist monsoon forest, occurring up to elevations of *c.* 1000 m. Montane species, in contrast, have narrow-lipped shells that are dark purplish or blackish brown, and are restricted to submontane and/or montane rainforest, rarely occurring below 1000 m. The lowland species are *Corilla adamsi* (Gude, 1914), *C. carabinata* (Férussac, 1821), *C. colletti* Sykes, 1897 and *C. lesleyae* Barnacle, 1956, all of which are parapatric/allopatric. The montane species comprise three parapatric/allopatric taxa, *C. gudei* Sykes, 1897, *C. humberti* (Brot, 1864) and *C. odontophora* (Benson, 1865), and three partly sympatric ones, *C. beddomeae* (Hanley in Hanley & Theobald, 1876), *C. erronea* (Albers, 1853) and *C. fryae* Gude, 1896.

In the only molecular phylogenetic study carried out to date on Sri Lankan *Corilla* (Raheem et al., 2017), most deep-level relationships within the group were not resolved. Based on three mitochondrial (mt) genes 16S rRNA, cytochrome oxidase subunit I (COI) and NADH dehydrogenase 1 (ND1), this study showed that the current shell-based taxonomy of Sri Lankan *Corilla* was only partly supported by mtDNA sequence data; the morphologically defined limits of three species (*C. gudei*, *C.erronea* and *C. fryae*) were inconsistent with the mtDNA data and the taxonomic status of *C. odontophora* could not be resolved. This study also suggested that diversification may have involved convergence in shell morphology. Here, we harness the power of restriction-site-associated DNA sequencing (RADseq) to further clarify the systematics and evolution of Sri Lankan *Corilla*. RADseq and other reduced-representation approaches are high throughput (genomics) sequencing methods that sub-sample loci from across the genome (Baird et al., 2008; Andrews et al., 2016). Due to its affordability, flexibility and efficiency, RADseq is now used to tackle a diverse array of ecological and evolutionary questions (Andrews et al., 2016; Eaton and Overcast, 2020). An added advantage is that this approach does not require reference genomes, so is suitable for studies of nonmodel taxa (da Fonseca et al., 2016; Fitz-Gibbon et al., 2017; Lee et al., 2018). In systematics research, RADseq has been used at a range of phylogenetic scales (Eaton et al., 2017; Lecaudey et al., 2018) and to resolve phylogenetic questions that were intractable using Sanger-sequence-

based datasets (e.g. Near et al., 2018; Wagner et al., 2018). Using phylogenetic analysis of RADseq data and ancestral state reconstructions, the present study seeks to: (1) fully resolve phylogenetic relationships within Sri Lankan *Corilla* and (2) explore the history of habitat association and shell morphological change in this group.

## 2. Material and methods

### *2.1. Taxon sampling and RADseq library preparation*

The present study includes 9 of the 10 currently recognized species of Sri Lankan *Corilla* (Gude, 1914; Barnacle, 1956). The Sri Lankan species *Corilla lesleyae* (not seen since its discovery in the early 20$^{th}$ century) and the Indian species *C. anax* (Benson, 1865) were not sampled. We sampled 2 to 6 individuals from across the range of each species (**Fig. 1**), yielding a total of 30 individuals (**Table 1**; for registration details see **Supplementary Material Table S1**). Species identifications were based on Gude (1914) and Raheem et al. (2017).

Genomic DNA was extracted from *c.* 6–10 mm$^3$ pieces of foot tissue with the Nucleospin® Tissue kit (Macherey-Nagel, Düren, Germany), following the manufacturers' standard protocol, or using a CTAB (hexadecyltrimethylammonium bromide) protocol (Winnepenninckx et al., 1993). The quality of DNA extracts was assessed by quantification using a Qubit® 2.0 Fluorometer (Invitrogen, Life Technologies, Carlsbad, California, USA) and visualization on a 1% agarose gel.

Three RADseq libraries were prepared using an original RADseq (*sensu* Andrews et al., 2016) protocol based on Etter et al. (2011) and Baird et al. (2008), in which sequenced regions are adjacent to single restriction-enzyme cut sites. Individual 250-ng DNA samples were first digested with the restriction enzyme SbfI (SBfI-HF by NEB, USA) at 37 ºC for 60 min. Individually barcoded P1 adapters (IDT, Coralville, Iowa, USA), allowing discrimination of sequence reads from different individuals, were ligated to compatible ends of each digested DNA sample. The uniquely barcoded samples were pooled into multiplexed libraries and purified using DNA Clean & Concentrator™-5 (Zymo Research Corp, Irvine, California, USA). Each multiplexed library (50 µl volume) was sheared to an optimal size of ~300–400 bp in a M220 Focused-ultrasonicator™ (Covaris, Inc., Woburn, Massachusetts, USA), using a microTUBE-50 and the following instrument settings: peak incident power = 50W, duty factor = 20%, cycles per burst = 200, treatment time = 60 s. The sheared DNA was size-selected on an agarose gel and end-blunted using Quick Blunting™ Kit (NEB, USA). The libraries were then A-tailed and ligation of a second P2 adapter (IDT, Coralville, Iowa, USA) was carried out. Finally, using PCR amplification with P1 (IDT) and P2 (Sigma) primers and Phusion® High-Fidelity PCR Master Mix with HF Buffer (NEB, USA), the multiplexed libraries were enriched and tagged with a second, unique (library-specific) barcode. Each enriched library was quality checked (molar concentration ≥ 2 nM, size distribution of DNA fragments peaking at ~400 bp) using a Qubit® 2.0 Fluorometer and a High Sensitivity DNA assay on an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbron, Germany). Sequencing involved 250 bp, paired-end sequencing of each library on an Illumina MiSeq single-lane flow cell (MiSeq v 2 kit, 2 x 250 bp) and was done at Edinburgh Genomics, University of Edinburgh, Scotland and the Centre of Medical Genetics, University of Antwerp, Belgium.

### *2.2. Processing of RADseq data*

RADseq data were assembled using ipyrad (Eaton and Overcast, 2020) v. 0.6.11. The *de novo* assembly option was used, in which ipyrad uses the tool vsearch to cluster sequences and infer homology on the basis of sequence similarity. All ipyrad analyses were run on a server in the

molecular biology computing facility of the Natural History Museum (NHM), London, UK. *De novo* assembly of original RADseq paired-end reads in ipyrad is problematic so, as recommended (see: https://ipyrad.readthedocs.io/en/master/faq.html), only R1 reads were analysed.

Data assembly centred on three key parameters: (1) the clustering threshold, the percentage sequence similarity that is used to assess whether two sequences are homologous; (2) the minimum depth of read clusters for statistical base calling; and (3) the minimum number of individuals (= samples) per locus, that is the minimum number of individuals having data for a given locus for that locus to be retained in a final assembly. Four combinations of clustering threshold and minimum depth were used (**Table 2**); we hereafter refer to these as the four parameter combinations. We used a clustering threshold of 0.85 or 0.90 (the range 0.85–0.90 is generally reliable for most analyses because it balances the over-splitting of loci with over lumping; https://ipyrad.readthedocs.io/en/master/tutorial_intro_cli.html) and the minimum depth for statistical base calling was set at 7 or 9. For each of the four parameter combinations, the minimum individuals per locus varied from 4–30. Unless stated otherwise, all other parameter settings were held constant across the four parameter combinations.

The ipyrad workflow consisted of 7 steps. In step 1, using the restriction-site overhang for SBfI (5′-TGCAGG-3′) and the barcodes specific to each individual, R1 reads were de-multiplexed separately for each library. Zero barcode mismatches were allowed. The demultiplexed reads for the 30 individuals were then assembled in a single directory and quality checked using the tool FastQC version 0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Step 2 involved filtering and editing the reads. Illumina adapters were removed using the strict filter option. Our FastQC analyses showed that some individuals had a biased sequence composition at *c.* the first 11 bases of the reads (i.e. the restriction-site overhang, comprising the first 6 bases, and the 5 bases immediately adjacent) and these bases were removed. FastQC analyses also showed that in many individuals quality scores dropped to < 20 towards the end of the reads (i.e. in the last 50 bases), so overall read lengths were trimmed to 200 bases. The trimmed reads were relatively long, so the threshold for the maximum number of low-quality bases was set at 10; bases with a quality score < 20 were treated as being ambiguous (denoted by N) and reads with > 10 Ns were removed.

In step 3, reads were clustered within individuals at a clustering threshold of either 0.85 or 0.90. The clustered reads were then aligned. Steps 4 involved joint estimation of error rates and heterozygosity and step 5 involved consensus base calling (minimum depth of 7 or 9) and filtering. Given the length of our trimmed reads, the maximum number of uncalled bases allowed in consensus sequences was set at 10 and the maximum number of heterozygous bases allowed in consensus sequences was set at 16. In step 6, reads were clustered across individuals at the same clustering threshold as step 3 and the read clusters were aligned. In step 7, loci were concatenated and data assemblies were generated on the basis of the minimum number of individuals per locus. Using a branched workflow (see: https://ipyrad.readthedocs.io/en/master/7-outline.html), we incrementally increased the setting of this parameter from 4–30 (all other parameters were held constant); 4 is the minimum number of individuals required to reconstruct a phylogeny. The maximum number of SNPs allowed in a final locus was 40, and the maximum number of indels allowed in a final locus was 10; again, these values reflect our relatively long read lengths. We also removed the last five bases from the edges of the final aligned loci; in *de novo* datasets, in particular, the 3′ edges of the final aligned loci tend to be less well aligned than the 5′ edges. Assemblies were output in phylip format for subsequent analysis.

### *2.3. Phylogeny reconstruction*

Phylogenies were inferred using a maximum likelihood (ML) approach. Analyses were carried out using RAxML-HPC2 on XSEDE (v. 8.2.10) on the CIPRES Science gateway (v.3.3; http://www.phylo.org/index.php, Miller et al., 2010). As recommended by Stamatakis (2016, p. 60), the GTR+G model of nucleotide substitution was used. Each RAxML analysis involved a single run and consisted of a rapid bootstrap analysis with the extended majority rule bootstopping criterion (maximum of 1,000 bootstrap replicates), followed by a search for the best scoring ML tree (Stamatakis, 2016). Raheem et al.'s (2017) analysis of mtDNA sequence data suggested that *C. beddomeae* is the sister of other Sri Lankan *Corilla*, so all trees were rooted on *C. beddomeae.* We considered branches with bootstrap values ≥70% to be strongly supported (Hillis and Bull, 1993) and those with bootstrap values <70% not to be strongly supported.

RADseq datasets with substantial amounts of missing data can be used to accurately reconstruct known phylogenies (Rubin et al., 2012; Razkin et al., 2016; Streicher et al., 2016). Following Streicher et al. (2016), we quantified missing data in two ways. First, as 'overall missing data', the proportion of empty cells in a data matrix/assembly/multiple sequence alignment. Second, as 'missing individuals per locus', the maximum proportion of individuals (or taxa when each tip in a phylogeny represents a different taxon) lacking data for any locus in a given assembly. The difference between these two measures and thus the effect of missing data on phylogenetic inference can vary from dataset to dataset (Streicher et al., 2016; Collins and Hrbek, 2018).

We compared branch support and phylogenetic relationships across a subset of 88 trees (i.e. those associated with 17–87% missing individuals per locus; for absolute numbers of missing individuals see **Fig. 3**); the remaining trees were not considered because as missing individuals per locus dropped below 17%, there was a sharp decline in resolution (most relationships were not resolved). In all 88 trees, the ingroup (clade 1 in **Fig. 4**) had 15 internal branches, so for each tree we calculated the sum of squared support (bootstrap values) for the 16 focal branches (the branch subtending clade 1 and the 15 internal branches within clade 1). We compared branch support using the sum of squared support rather than the sum of branch support or mean branch support (e.g. Streicher et al., 2016) because the first gives greater weight to strong support (Wilkinson and Crotti, 2017; M. Wilkinson, personal communication), thus making it easier to detect the impact of poorly/less strongly supported branches on total support.

We also assessed the relative distribution of assembled genomic data across different parts of the tree by quantifying coverage of loci per individual (i.e. the proportion of loci in an assembly for which data are available for a given individual) for 15 assemblies in each of the four ipyrad parameter combinations. The 15 assemblies are those associated with 40–87% missing individuals per locus and well-resolved trees (most or all branches strongly supported).

### *2.4. Ancestral-state reconstructions*

We traced the evolution of habitat association and six shell characters by reconstructing ancestral character states. This was done in Mesquite version 3.31 (Maddison and Maddison, 2017) using maximum parsimony (MP) and ML approaches. In both cases, we used the two best supported ML trees (topologies A and B in **Fig. 4**) from our RADseq analyses with individual nodes in each tree being assigned the same number as the associated branch (see **Fig. 4** and **Supplementary Material Figs S1 and S2**). All characters were coded as binary-state characters, with each character state being present in at least two of the species. The two states for habitat association were lowland and montane. The shell characters and states (in parentheses) analysed were: shape (discoid *vs* ovate), colour (dark *vs* pale), lip width (narrow *vs* wide), length of upper palatal folds (short or moderate *vs* long), orientation of upper palatal folds (parallel *vs* non-parallel) and collabral sculpture (bold *vs* fine or moderate) (see **Fig. 2** for

details). All shell characters were analysed together using the "trace all characters" option in Mesquite. All branches within three of the species-level clades (*C. beddomeae*, *C. carabinata* and *C. erronea–fryae*) were collapsed. In MP reconstructions, character transitions were treated as equal-cost and reversible (the unordered assumption). In ML-based reconstructions we used both the Mk1 (Markov k-state 1 parameter) and the AsymmMk (asymmetrical Markov k-state 2 parameter) model; rate parameters were estimated from the data and, in the case of the AsymmMk model, we used the option "root state frequencies same as equilibrium". Judgment of the 'best' ancestral character states was made using Mesquite's default likelihood decision threshold of 2 (i.e. if the negative log likelihoods of two states differ by ≥ 2, the one with lower likelihood was rejected). The fit of the Mk1 and AsymmMk models was compared using the second order information criterion (AICc; Sugiura, 1978) (for results in full see **Supplementary Material Tables S2–S5**). The AICc is appropriate for model selection in the context of small datasets ($n/K < 40$, where $n$ is the sample size and $K$ the number of estimable model parameters; Burnham and Anderson, 2002).

### *2.5. ZooBank registration*

This paper has been registered in ZooBank; the LSID for the article is: xxxxxxxxxxxxxxxxxxx.

## 3. Results

After demultiplexing, quality filtering and adapter filtering, a mean of 612,195 reads (R1) per individual (range = 241,864–1,323,833) was retained. After processing with ipyrad the mean number of consensus reads per individual ranged from 11,859–15,836 for the four ipyrad parameter combinations and the total number of loci ranged from 16,596–21,790. Across the four ipyrad parameter combinations, the length of the assemblies, the number of loci and the number of variable sites increased with increasing levels of missing individuals per locus (**Fig. 5A–C**). We also found that the two measures of missing data, the overall proportion of missing cells in an assembly and the proportion of missing individuals per locus, increased relative to each other (**Fig. 5D**). For all four parameter combinations, total branch support and topological resolution increased with increasing assembly size and declining levels of missing individuals per locus (**Fig. 3**). At 17–37% missing individuals per locus and assembly sizes of 9,604–198,413 bp, a range of topologies was observed, most relationships were strongly supported and total branch support ranged from 130,828–151,123. At 40–87% missing individuals per locus and assembly sizes of 162,279–4,132,850 bp, nearly all trees conformed to one of two topologies (here referred to as A and B; **Fig. 4**), most relationships were strongly supported and total branch support approached the maximal value (160,000); none of the trees had maximal support for all 16 focal branches. For assemblies with ≥50% missing individuals per locus, only topologies A and B were observed and support was always maximal for 13 of the 16 focal clades and always maximal or strongly supported for clade 3 (**Fig. 3; Table 3**). Topologies A and B differ in the composition of clade 4, that is specifically in the position of *Corilla odontophora* 'south' relative to other species (see **Fig. 4**). Clade 4 was not always strongly supported and maximal support for it was recovered only in a few topology-B trees. Clade 8 was also not always strongly supported but, in most cases, support was maximal or strong.

The largest assemblies (9,613–21,790 total loci, *c.* 1.8–4.1 million bp) in parameter combinations c85d7, c90d7 and c90d9 yielded topology A, with a shift to topology B as assembly sizes and missing data levels declined (**Fig. 3**). Across the four ipyrad parameter combinations, topology A was replaced by B when the clustering threshold was lowered (e.g. compare c90d7 and c85d7) or the minimum depth was raised (e.g. compare c85d7 and c85d9). Only topology B was observed for the largest assemblies in parameter combination c85d9. At

47–87% missing individuals per locus, topology B unlike topology A was associated with a wide range of assembly sizes (1,502–16,596 total loci, *c.* 35,000–3.2 million bp).

Coverage of loci was variable across individuals and its distribution across the tree was highly uneven (**Fig. 6)**. In the large to moderate-sized assemblies (87–40% missing individuals per locus) high-coverage individuals had approximately 40–95% of sampled loci and low-coverage individuals had *c.* 10–50% of loci. Clade 3 was dominated by high-coverage individuals, whereas other parts of the tree were dominated by low-coverage individuals. The pattern was similar for all four ipyrad parameter combinations.

The monophyly of species *C. adamsi*, *C. beddomeae*, *C. carabinata*, *C. colletti* and *C. humberti* was always maximally supported, but the remaining four species were consistently non-monophyletic (**Table 3**). Together, *C. erronea* and *C. fryae* formed a consistently maximally supported clade (*C. erronea–fryae* clade of **Fig. 4**), but most relationships within this clade were not strongly supported. However, at moderate to high levels of missing individuals per locus (47–87%; corresponding to a minimum of 5–27 individuals per locus) relationships within the clade comprising *C. erronea* 1, *C. erronea* 2 and *C. fryae* (three individuals) were almost always strongly supported, with both *C. erronea* and *C. fryae* being polyphyletic. In the case of *C. gudei*, while the single individual of *C. gudei* 'south' was always maximally supported as sister to *C. humberti*, the two individuals of *C. gudei* 'north' formed a maximally supported sister-group relationship with *C. carabinata*. We consistently recovered *C. odontophora* as polyphyletic. *Corilla odontophora* 'east' was usually strongly supported as sister to *C. colletti*, and *C. odontophora* 'north' was maximally supported as sister to the clade comprising *C. humberti* and *C. gudei* 'south'. The position of *C. odontophora* 'south' in clade 3 differed between topologies A and B. In topology A, *C. odontophora* 'south' was usually strongly supported as sister to clade 5 (the clade uniting *C. carabinata*, *C. gudei* 'north', *C. colletti* and *C. odontophore* 'east'), whereas in topology B it was usually strongly supported as sister to clade 6 (the clade comprising *C. odontophora* 'north', *C. gudei* 'south' and *C. humberti*).

Apart from the position of *C. odontophora* 'south', topologies A and B showed similar, well-resolved relationships at and above the species level. The two deepest nodes have given rise to three major lineages: the montane *C. beddomeae*, the lowland *C. adamsi* and clade 1, which comprises all the other taxa. The sister-group relationship between the lowland *C. adamsi* and clade 1 was always maximally supported. Clade 1 consisted of two maximally supported sister clades: the montane *C. erronea–fryae* clade and clade 3. In both topologies A and B, clade 3 contained clades 5 and 6. Clade 6, a wholly montane clade, comprised a maximally supported sister group-relationship between *C. odontophora* 'north' and the clade of *C. gudei* 'north' and *C. humberti*. Clade 5 comprised a maximally supported sister-group relationship between two clades, each clade containing one lowland and one montane species: the lowland *C. colletti* and its sister, the montane *C. odontophora* 'east', and the lowland *C. carabinata* and its sister the montane *C. gudei* 'north'.

The results of the ancestral-state reconstructions for habitat association and shell morphology were similar for topologies A and B, with the states inferred for the two deepest nodes and for the common ancestors of clade 1, of the *C. odontophora* 'east' + *C. colletti* clade and of the *C. gudei* 'north' + *C. carabinata* clade being the same for both topologies (**Figs 7–9; Supplementary Material Tables S2–S5, Figs S1, S2**). Topology B was associated with a wide range of assembly sizes including some of the largest ones, so we focus here on the results for this topology. MP ancestral-state reconstructions of habitat association (**Fig. 7A**) unambiguously inferred all ancestral states, indicating that the state at the two deepest nodes was montane and that species diversification in Sri Lankan *Corilla* has involved transitions from a montane ancestor to a lowland descendant in three distinct lineages, *C. adamsi*, *C. carabinata* and *C. colletti*. In the two ML reconstructions (**Fig. 7B, Supplementary Material Tables S4 and S5, Fig. S2**), the state at the two deepest splits could not be resolved: a montane state was not significantly more likely than a lowland one. However, for both MP and

ML reconstructions, the states inferred for clade 1 were similar: the ancestor of clade 1 was montane and there have been transitions from montane to lowland states in two lineages, *C. carabinata* and *C. colletti*. Although the AICc score for the Mk1 model was lower than that for the AsymmMk1 model, the difference was small and indicated a negligible difference in fit.

MP reconstructions unambiguously inferred ancestral states for five of six shell characters at all 20 nodes and for one shell character (orientation of upper palatal folds) at 17 nodes (**Figs 8A, 9A**); the state of this last character at the two deepest nodes and the node subtending the *C. adamsi* clade was equivocal. The ML reconstructions (**Figs 8B, 9B; Supplementary Material Tables S4, S5 and Fig. S2**) were broadly similar to the MP ones, but uncertainty was greater. The ML reconstructions unambiguously inferred states for three characters (shape, colour and lip width) at 18 nodes. For the other three characters (length of palatal folds, orientation of palatal folds and collabral sculpture), states were inferred at 12–16 nodes; the states at the remaining nodes were unreconstructable (4 nodes) or unresolved (1 or 2 nodes) for the first two of these characters and were all unresolved (4 or 8 nodes) for the third character. The Mk1 model generally yielded a lower AICc score than the AsymmMk1 model for the six shell characters, but the difference between the two models was negligible.

In the MP reconstructions, the two deepest nodes were associated with a discoid, dark and narrow lipped shell with short upper palatal folds and bold collabral sculpture. In contrast, in the ML reconstructions the state at the deepest node could be inferred only for the length of the palatal folds (short), with the states of the other five characters being unresolved. At the next-to-deepest node, states were unresolved for shell shape, colour, lip width and collabral sculpture, and were unreconstructable for length and orientation of upper palatal folds. The MP and ML analyses yielded similar states for four of the shell characters at 14 of the 16 nodes in clade 1. Two of the 14 nodes in clade 1 (the ancestors of *C. carabinata* and of *C. colletti*) were associated with an ovate, pale and wide-lipped shell with parallel upper palatal folds; the 12 remaining nodes in clade 1, including all the deepest nodes, the ancestor of the *C. odontophora* 'east' + *C. colletti* clade and the ancestor of the *C. gudei* 'north' + *C. carabinata* clade, were associated with an ovate, dark and narrow lipped shell with non-parallel upper palatal folds.

## 4. Discussion

### *4.1. Phylogenetic relationships and taxonomic implications*

Apart from the position of *Corilla odontophora* 'south', all deep nodes in the RADseq-based trees were well resolved, with the two common RADseq-based trees (topologies A and B) being congruent with each other and consistent with Raheem et al.'s (2017) mtDNA-based tree in which most of the deeper nodes were not well resolved. Our study thus agrees with the shell-based taxonomy of *C. adamsi*, *C. beddomeae*, *C. colletti*, *C. carabinata* and *C. humberti*, all of which were maximally supported as monophyletic species. However, our results indicate that the current conchologically defined species limits of *C. gudei*, *C. erronea*, *C. fryae* and *C. odontophora* are inconsistent with the RADseq data. *Corilla gudei* 'north' and 'south' were inferred to be polyphyletic, with *C. gudei* 'north' sister to the lowland species *C. carabinata* and *C. gudei* 'south' sister to *C. humberti* (**Fig. 4**). This confirms Raheem et al.'s (2017) finding that *C. gudei* 'north' and 'south' are distinct species. The polyphyly of the nominal taxon *C. odontophora* was also strongly supported in the RADseq trees, whereas in Raheem et al.'s (2017) mtDNA-based study, only *C. odontophora* 'east' and 'north' were included and their relationships were not well resolved. This finding suggests that *C. odontophora* 'east', 'north' and 'south' are distinct species, with geographically adjacent but isolated ranges on the hills bounding the Uva Basin.

Although *C. erronea* and *C. fryae* constitute a strongly supported clade, the topology within the clade does not correspond to any obvious taxonomic, morphological, elevational or geographical patterns. These findings agree with those of Raheem et al. (2017). *Corilla erronea* and *C. fryae* are confusingly similar in shell morphology, but the name *C. fryae* has traditionally been applied to specimens from the southwestern part of the Central Highlands, whereas *C. erronea* has been used for individuals from other montane areas. Taken together, the RADseq, mtDNA, morphological and distributional data suggest that *C. erronea* and *C. fryae* constitute a single relatively widespread species, for which the valid name is *C. erronea* (Albers, 1853) with *C. fryae* Gude, 1896 as a junior synonym (ICZN, 1999, Article 23.1). This *C. erronea–fryae* lineage is readily separable from other Sri Lankan *Corilla* by the faint, but distinctive striae that intersect the collabral sculpture (see Raheem et al. 2017: fig. 1E). Some relationships within this clade were poorly resolved, which may partly reflect the low coverage of loci per individual relative to other parts of the tree (**Fig. 6**).

Excluding *C. lesleyae* (not sampled in this study), the current shell-based taxonomy recognizes 9 species of Sri Lankan *Corilla*, whereas the RADseq data indicate at least 11. Interestingly, the more extensive geographic sampling of Raheem et al.'s (2017) mtDNA study revealed deep phylogenetic divergences within *C. adamsi*, *C. beddomeae* and *C. colletti*, suggesting that these species may contain additional unrecognized species. On the basis of currently used shell characters, *C. odontophora* 'east', 'north' and 'south' are conchologically cryptic and can be differentiated from each other only if precise locality (or DNA-sequence) data are available. *Corilla gudei* 'north' and 'south', which are restricted to adjacent but non-overlapping ranges in the Knuckle Massif (**Fig. 1**), pose a similar problem, the subtle conchological differences between the two (Raheem et al., 2017) not being always obvious. While the findings of our RADseq-based study broadly agree with those of the earlier mtDNA-based work by Raheem et al. (2017), they clearly demonstrate that the species-level diversity of *Corilla* has been underestimated. The obvious next step is a thorough taxonomic revision of the group.

### *4.2. Topological incongruence*

We found that branch support and phylogenetic resolution increased with increasing assembly size, despite increasing levels of missing data. This agrees with other RADseq-based studies (Rubin et al., 2012; Razkin et al., 2016; Tripp et al., 2017; Wang et al., 2017; Crotti et al., 2019). Overall, our results were strongly supported and consistent for moderate to large data assemblies, with missing individuals per locus ranging from ≈50–87%. Topology A was recovered for only the largest assemblies in three of the four parameter combinations and was recovered less often or not at all when the clustering threshold was low and/or the minimum depth of read clusters was high. These results underline the impact of assembly parameters on the reliable estimation of phylogenetic patterns and suggest a possible effect of assembly size on topology. Topological incongruence could be due to two, mutually non-exclusive factors: biological processes, such as gene duplication, incomplete lineage sorting or historical hybridization (Som, 2015; Schrempf and Szöllösi, 2020), all of which are known to occur in stylommatophoran land snails (Saur and Hausdorf, 2010; Böckers et al., 2016; Richards et al., 2017; Koch et al., 2017; Noda et al., 2019), and methodological artefacts including biased taxon sampling (Som, 2015) and differences in library preparation (Lambert et al., 2019).

The most unstable part of the tree (clade 3) has noticeably high loci coverage per individual and thus relatively little missing data in comparison to other parts of the tree. Nonetheless, several members of clade 3 have relatively low loci coverage, so insufficient sequence data may be an issue. This may also explain the variable support recovered for 3 of the 16 focal branches (branches 3, 4 and 8) at 50–87% missing individuals per locus (support for all the other focal branches was always maximal). We note the non-random distribution of loci coverage across the tree, with greater coverage per individual in clade 3. This may indicate

a genealogical sampling bias (relatively fewer loci being detected between divergent taxa), a particular issue with the RADseq approach—this arises because of mutations that disrupt enzyme recognition sites (Rubin et al., 2012; Arnold et al., 2013; Eaton et al., 2017; Collins and Hrbek, 2018) or may reflect biases arising from batch effects (e.g. differences between libraries, sequencing runs) or the quality and concentration of DNA extracts (da Fonseca et al., 2016; Leigh et al., 2018). Potential signal conflict among loci could also arise through the biased sampling of RADseq loci; simulations have shown that as the proportion of missing data declines and the size of data assemblies shrink, the mutational spectrum of the constituent loci becomes biased such that loci with the highest mutation rates are disproportionately excluded (Huang and Knowles, 2016; see also Crotti et al., 2019). An important avenue for future work on the systematics of *Corilla* is to investigate the topological incongruence observed in our RADseq-based analyses, to try and establish the underlying cause/s (e.g. see Richards et al., 2017).

### *4.3. History of habitat association and shell morphological change*

The habitats associated with the 16 nodes of clade 1 were unambiguously reconstructed by both MP and ML approaches. They indicate that the ancestor of clade 1 occurred in montane habitats and that transitions from a montane to a lowland association have occurred on at least two separate occasions (*C. carabinata* and *C. colletti*). This is consistent with Raheem et al.'s (2017) preliminary phylogeny. Outside clade 1, there may have been one more transition from a montane to a lowland habitat (i.e. in *C. adamsi*), but that ancestral-state reconstruction was uncertain because it was supported by only the MP and not by the ML analysis. This MP *vs* ML uncertainty also applies to the habitat associated with the deepest node in the tree.

For the shell characters, as with habitat association, there was no agreement between the MP- and ML-inferred states at the two deepest nodes: all MP-inferred states were unequivocal, while nearly all ML-inferred states were unresolved. The state at the deepest node could only be inferred with certainty for the length of the upper palatal folds. Thus, all we can infer is that the ancestor of Sri Lankan *Corilla* had a shell with short/medium upper palatal folds. In contrast, character-state reconstructions for clade 1 unambiguously inferred ancestral states for shell shape, colour, lip width and the orientation of the upper palatal folds. The ancestor of clade 1, thus, had an ovate, dark and narrow-lipped shell with non-parallel upper palatal folds. This combination of character states has been retained by all species in this clade, apart from *C. humberti*, *C. carabinata* and *C. colletti*. *Corilla humberti* has retained the ancestral state of clade 1 for shell shape, colour and lip width but lacks palatal folds. *Corilla carabinata* and *C. colletti* have retained the ancestral ovate shape of the shell and *C. carabinata* has also retained the ancestral non-parallel orientation of the upper palatal folds. Conversely, these two species exhibit a pale shell colour and wide lip, both derived states; *C. colletti* also has the derived parallel orientation of the upper palatal folds.

Our ancestral-state reconstructions for habitat association and shell morphology provide strong evidence that diversification in Sri Lankan *Corilla* has involved evolutionary convergence (*sensu* Losos, 2011) in the shell morphology of lowland lineages, with a pale shell and wide lip having evolved on at least two separate occasions (in *C. carabinata* and *C. colletti*) from montane ancestors having a dark, narrow-lipped shell. A number of studies of land snails have explained morphological and ecological similarities between non-sister species as being the result of convergent adaptation (e.g. Emberton, 1995; Dowle et al., 2015; Köhler and Criscione, 2015; Hyman et al., 2017). This explanation has, however, not been rigorously investigated for land snails (Davison, 2002) as it has been for other taxa, such as *Anolis* lizards (Losos et al., 1998). Very little is known about the biology of *Corilla*, so we can only speculate on the causes and processes underlying habitat-associated differences and convergence in shell morphology. In a few well-studied land snails, shell colour has been linked to factors, such as heat absorption or protection from predators (Saenko and Schilthuizen, 2021); predation has also

been invoked to explain the structure of the apertural dentition (Wada and Chiba, 2013). The applicability of these and other explanations for the shell morphological variation of *Corilla* remains to be investigated.

### *4.4. Conclusions*

By largely resolving the phylogenetic relationships of Sri Lankan *Corilla*, this RADseq-based study substantially extends our understanding of the diversity and diversification of the group. The unstable position of *C. odontophora* 'south', however, does require further investigation. Our study also shows that the current, shell-based taxonomy clearly underestimates the species diversity of Sri Lankan *Corilla*: the names *C. gudei* and *C. odontophora*, as currently used, each apply to at least two distinct, yet conchologically-cryptic species. This finding points to the need for detailed revisionary studies and hints at the possibility that further undescribed species remain to be discovered—there are significant extents of habitat within the range of the genus that have yet to be sampled. We found firm evidence for evolutionary convergence in the shell morphology of lowland lineages, with a pale shelled, wide lipped lowland morphology having evolved from a dark shelled, narrow-lipped montane ancestor on at least two separate occasions. The underlying factors, however, are unclear and our understanding of them is hindered by the sparsity of data on the natural history of *Corilla* (e.g. life cycle, diet and natural enemies). This lack of data may also have serious implications for the conservation of these threatened, forest-living snails, and field- and lab-based studies on the biology of *Corilla* are urgently needed.

## Acknowledgements

This study was funded by the Royal Belgian Institute of Natural Sciences (RBINS) and a Post-doctoral Research Fellowship awarded to DCR by the Belgian Federal Science Policy Office (BELSPO) and the Marie Curie Actions of the European Commission (FP7-PEOPLECOFUND-2008). The work was done in the context of the FWO research community W0.009.11N 'Belgian network for DNA barcoding' and with support from the 'Joint Experimental Molecular Unit' (JEMU) at RBINS. Additional funding was provided by the British Ecological Society, the Malacological Society of London, the NHM and Synthesys BE-TAF. The study is based on extensive field surveys and collections-based study and analysis, which were made possible through permits issued by the Forest Department and Department of Wildlife Conservation, Government of Sri Lanka. Much of the fieldwork was done during the course of a collaborative Darwin Initiative Project; separate surveys carried out by DCR generated substantial additional data. The DI project was led by F. Naggs and involved the NHM and two Sri Lankan institutions, the Department of National Museums (DNM), Colombo and the University of Peradeniya. We would like to thank Y. Mapatuna, Director of the DNM during the Darwin Project, for her wholehearted support; L. Kariyawasam, the curator then in charge of the DNM's mollusc collections, who played a key role in running the project; and R. Pethiyagoda for his role in making the project a success. Among the present staff of the DNM, we are grateful to L. Somaratne for arranging access to the collections and to T. Dasanayake, T. Gamage, C. Munasinghe and M. De Silva for their assistance with the curation and registration of material. At RBINS, C. De Busschere, S. Derycke, S. Gomber, F. Hendrickx, M. Quinzin, G. Sonet and C. Vangestel helped with RADseq library preparation, data processing and/or analyses; O. Razkin gave support with RADseq data processing; and Y. Barette provided expertise on software and IT-related issues. At the NHM, P. Forster, J. Streicher and M. Wilkinson gave valuable feedback on the analyses and H. Taylor provided high-quality images. DCR thanks D. Abeyewardene and the late A. Abeyewardene for generously supporting her in the UK. We would also like to thank the following people: L. W. Perera, H.W.V. Pushpamal, S.

Wickramanayake, A. Palihawadana, G. Weerakoon, N. Perera, M.J.B. Green, L.B. Karunaratne, the late S.U. Deraniyagala and the late E. de Silva.

## References

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17, 81–92. https://doi.org/10.1038/nrg.2015.28.

Arnold, B., Corbett-Detig, R.B., Hartl, D., Bomblies, K., 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol. Ecol. 22, 3179–3190. https://doi.org/10.1111/mec.12276.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3, e3376. https://doi.org/10.1371/journal.pone.0003376.

Barnacle, G.A.S. 1956. A new species of *Corilla* from Ceylon. J. Conchol. 24, 95–96.

Beenaerts, N., Pethiyagoda, R., Ng, P.K.L., Yeo, D.C.J., Bex, G.J., Bahir, M.M., Artois, T., 2010. Phylogenetic diversity of Sri Lankan freshwater crabs and its implications for conservation. Mol. Ecol. 19, 183–196. https://doi.org/10.1111/j.1365-294X.2009.04439.x.

Bhosale, A.R., Saadi, A.J., Wade, C.M., Thackeray, T.U., Tamboli, A.S., Kadam, S.K., Muley, D. V., Raheem, D.C., 2021. *Varadia*, a new helicarionoidean semi-slug genus from India's Western ghats (Stylommatophora: Helicarionoidea). Eur. J. Taxon. 757, 50–79. https://doi.org/10.5852/ejt.2021.757.1413.

Böckers, A., Greve, C., Hutterer, R., Misof, B., Haase, M., 2016. Testing heterogeneous base composition as potential cause for conflicting phylogenetic signal between mitochondrial and nuclear DNA in the land snail genus *Theba* Risso 1826 (Gastropoda: Stylommatophora: Helicoidea). Org. Divers. Evol. 16, 835–846. https://doi.org/10.1007/s13127-016-0288-0.

Burnham, K., Anderson, D. 2002. Model Selection and Multimodel Inference: A Practical Information-theoretic Approach, second edn. Springer-Verlag, New York.

Collins, R.A., Hrbek, T., 2018. An *in silico* comparison of protocols for dated phylogenomics. Syst. Biol. 67, 633–650. https://doi.org/10.1093/sysbio/syx089.

Cooray, P.G. 1984. An Introduction to the Geology of Sri Lanka (Ceylon), second (revised) edn. National Museums of Sri Lanka, Colombo.

Cox, L.R. 1960. General characteristics of Gastropoda, in: Moore, C. (Ed.), Treatise on Invertebrate. Palaeontology Part I, Mollusca 1. Geological Society of America and University of Kansas Press, USA, pp. I84–I169.

Crotti, M., Barratt, C.D., Loader, S.P., Gower, D.J., Streicher, J.W., 2019. Causes and analytical impacts of missing data in RADseq phylogenetics: insights from an African frog (*Afrixalus*). Zool. Scr. 48, 157–167. https://doi.org/10.1111/zsc.12335.

da Fonseca, R.R., Albrechtsen, A., Themudo, G.E., Ramos-Madrigal, J., Sibbesen, J.A., Maretty, L., Zepeda-Mendoza, M.L., Campos, P.F., Heller, R., Pereira, R.J., 2016. Next-generation biology: sequencing and data analysis approaches for non-model organisms. Mar. Genomics 30, 3–13. https://doi.org/10.1016/j.margen.2016.04.012.

Davison, A., 2002. Land snails as a model to understand the role of history and selection in the origins of biodiversity. Popul. Ecol. 44, 129–136. https://doi.org/10.1007/s101440200016.

Dowle, E.J., Morgan-Richards, M., Brescia, F., Trewick, S.A., 2015. Correlation between shell phenotype and local environment suggests a role for natural selection in the evolution of *Placostylus* snails. Mol. Ecol. 24, 4205–4221. https://doi.org/10.1111/mec.13302.

Eaton, D.A.R., Overcast, I., 2020. Ipyrad: interactive assembly and analysis of RADseq datasets. Bioinformatics 36, 2592–2594. https://doi.org/10.1093/bioinformatics/btz966.

Eaton, D.A.R., Spriggs, E.L., Park, B., Donoghue, M.J., 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst. Biol. 66, 399–412. https://doi.org/10.1093/sysbio/syw092.
Emberton, K.C., 1995. Sympatric convergence and environmental correlation between two land-snail species. Evolution 49, 469–475. https://doi.org/10.1111/j.1558-5646.1995.tb02279.x.
Etter, P.D., Preston, J.L., Bassham, S., Cresko, W.A., Johnson, E.A., 2011. Local *de novo* assembly of rad paired-end contigs using short sequencing reads. PLoS One 6, e18561. https://doi.org/10.1371/journal.pone.0018561.
Fitz-Gibbon, S., Hipp, A.L., Pham, K.K., Manos, P.S., Sork, V.L., 2017. Phylogenomic inferences from reference-mapped and *de novo* assembled short-read sequence data using RADseq sequencing of California white oaks (*Quercus* section *Quercus*). Genome 60, 743–755. https://doi.org/10.1139/gen-2016-0202.
Gaston, K.J., 1991. How large is a species' geographic range? Oikos 61, 434. https://doi.org/10.2307/3545251.
Gude, G.K. 1914. The Fauna of British India including Ceylon and Burma. Mollusca – II. Trochomorphidae–Janellidae. Taylor and Francis, London.
Hillis, D.M., Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42, 182–192. https://doi.org/10.1093/sysbio/42.2.182.
Huang, H., Lacey Knowles, L., 2016. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of rad sequences. Syst. Biol. 65, 357–365. https://doi.org/10.1093/sysbio/syu046.
Hyman, I.T., Lamborena, I.D.L.I., Köhler, F., 2017. Molecular phylogenetics and systematic revision of the south-eastern Australian Helicarionidae (Gastropoda, Stylommatophora). Contrib. to Zool. 86, 51–95. https://doi.org/10.1163/18759866-08601004.
ICZN. (1999). International Code of Zoological Nomenclature, fourth edn. The International Trust for Zoological Nomenclature, London. https://www.iczn.org/the-code/the-international-code-of-zoological-nomenclature (accessed 9 July 2022).
Jaiswara, R., Dong, J., Robillard, T., 2019. Phylogenetic relationships in the cricket tribe Xenogryllini (Orthoptera, Gryllidae, Eneopterinae) and description of the Indian genus *Indigryllus* gen. nov. J. Zool. Syst. Evol. Res. 57, 789–805. https://doi.org/10.1111/jzs.12298.
Joshi, J., Edgecombe, G.D., 2018. Molecular phylogeny and systematics of the centipede genus *Ethmostigmus* Pocock (Chilopoda: Scolopendromorpha) from peninsular India. Invertebr. Syst. 32, 1316–1335. https://doi.org/10.1071/IS18030.
Koch, E.L., Neiber, M.T., Walther, F., Hausdorf, B., 2017. High gene flow despite opposite chirality in hybrid zones between enantiomorphic door snails. Mol. Ecol. 26, 3998–4012. https://doi.org/10.1111/mec.14159.
Köhler, F., Criscione, F., 2015. A molecular phylogeny of camaenid land snails from north-western Australia unravels widespread homoplasy in morphological characters (Gastropoda, Helicoidea). Mol. Phylogenet. Evol. 83, 44–55. https://doi.org/10.1016/j.ympev.2014.11.009.
Lambert, S.M., Streicher, J.W., Fisher-Reid, M.C., Méndez de la Cruz, F.R., Martínez-Méndez, N., García-Vázquez, U.O., Nieto Montes de Oca, A., Wiens, J.J., 2019. Inferring introgression using RADseq and DFOIL: power and pitfalls revealed in a case study of spiny lizards (*Sceloporus*). Mol. Ecol. Resour. 19, 818–837. https://doi.org/10.1111/1755-0998.12972.
Lecaudey, L.A., Schliewen, U.K., Osinov, A.G., Taylor, E.B., Bernatchez, L., Weiss, S.J., 2018. Inferring phylogenetic structure, hybridization and divergence times within Salmoninae (Teleostei: Salmonidae) using RAD-sequencing. Mol. Phylogenet. Evol. 124, 82–99. https://doi.org/10.1016/j.ympev.2018.02.022.
Lee, K.M., Kivelä, S.M., Ivanov, V., Hausmann, A., Kaila, L., Wahlberg, N., Mutanen, M., 2018. Information dropout patterns in restriction site associated DNA phylogenomics and a

comparison with multilocus Sanger data in a species-rich moth genus. Syst. Biol. 67, 925–939. https://doi.org/10.1093/sysbio/syy029.
Legg, C., Jewell, N. 1995. A 1:50,000-scale forest map of Sri Lanka: the basis for a national forest geographic system. Sri Lanka Forester, Special Issue (Remote Sensing), 3–24.
Leigh, D.M., Lischer, H.E.L., Grossen, C., Keller, L.F., 2018. Batch effects in a multiyear sequencing study: false biological trends due to changes in read lengths. Mol. Ecol. Resour. 18, 778–788. https://doi.org/10.1111/1755-0998.12779.
Losos, J.B., 2011. Convergence, adaptation, and constraint. Evolution 65, 1827–1840. https://doi.org/10.1111/j.1558-5646.2011.01289.x.
Losos, J.B., Jackman, T.R., Larson, A., De Queiroz, K., Rodríguez-Schettino, L., 1998. Contingency and determinism in replicated adaptive radiations of island lizards. Science 279 (5359), 2115–2118. https://doi.org/10.1126/science.279.5359.2115.
Maddison, W. P., D.R. Maddison. 2017. Mesquite: A Modular System for Evolutionary Analysis (Version 3.31). http://mesquiteproject.org.
Miller, M.A., Pfeiffer, W., Schwartz, T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees, in: Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010. New Orleans, LA, pp.1–8.
Naggs, F., Raheem, D., 2005. Sri Lankan snail diversity: faunal origins and future prospects. Rec. West. Aust. Museum, Suppl. 68, 11–29. https://doi.org/10.18195/issn.0313-122x.68.2005.011-029.
Near, T.J., MacGuigan, D.J., Parker, E., Struthers, C.D., Jones, C.D., Dornburg, A., 2018. Phylogenetic analysis of Antarctic notothenioids illuminates the utility of RADseq for resolving Cenozoic adaptive radiations. Mol. Phylogenet. Evol. 129, 268–279. https://doi.org/10.1016/j.ympev.2018.09.001.
Noda, T., Satoh, N., Asami, T., 2019. Heterochirality results from reduction of maternal diaph expression in a terrestrial pulmonate snail. Zool. Lett. 5, 1–13. https://doi.org/10.1186/s40851-018-0120-0.
Raheem, D.C., Breugelmans, K., Wade, C.M., Naggs, F.C., Backeljau, T., 2017. Exploring the shell-based taxonomy of the Sri Lankan land snail *Corilla* H. and A. Adams, 1855 (Pulmonata: Corillidae) using mitochondrial DNA. Mol. Phylogenet. Evol. 107, 609–618. https://doi.org/10.1016/j.ympev.2016.12.020.
Raheem, D. C., Taylor, H., Ablett, J., Preece, R. C., Aravind, N. A., Naggs, F. 2014. A systematic revision of the land snails of the Western Ghats of India. Tropical Natural History, Supplement 4, 1–294. https://li01.tci-thaijo.org/index.php/tnh/article/view/103091.
Razkin, O., Sonet, G., Breugelmans, K., Madeira, M.J., Gómez-Moliner, B.J., Backeljau, T., 2016. Species limits, interspecific hybridization and phylogeny in the cryptic land snail complex *Pyramidula*: the power of RADseq data. Mol. Phylogenet. Evol. 101, 267–278. https://doi.org/10.1016/j.ympev.2016.05.002.
Richards, P.M., Morii, Y., Kimura, K., Hirano, T., Chiba, S., Davison, A., 2017. Single-gene speciation: mating and gene flow between mirror-image snails. Evol. Lett. 1, 282–291. https://doi.org/10.1002/evl3.31.
Rubin, B.E.R., Ree, R.H., Moreau, C.S., 2012. Inferring phylogenies from RAD sequence data. PLoS One 7, e33394. https://doi.org/10.1371/journal.pone.0033394.
Saenko, S. V., Schilthuizen, M., 2021. Evo-devo of shell colour in gastropods and bivalves. Curr. Opin. Genet. Dev. 69, 1–5. https://doi.org/10.1016/j.gde.2020.11.009.
Sauer, J., Hausdorf, B. 2010. Reconstructing the evolutionary history of the radiation of the land snail genus *Xerocrassa* on Crete based on mitochondrial sequences and AFLP markers. BMC Evolutionary Biology, 10, 299. https://doi.org/10.1186/1471-2148-10-299.
Schrempf, D., Szöllösi, G. 2020. The sources of phylogenetic conflicts, in: Scornavacca, C., Delsuc, F., Galtier, N. (Eds), Phylogenetics in the Genomic Era. No commercial publisher, authors open access book, pp. 3.1:1–3.1:23. https://hal.archives-ouvertes.fr/hal-02535482/document (accessed 9 July 2022).
Som, A., 2014. Causes, consequences and solutions of phylogenetic incongruence. Brief. Bioinform. 16, 536–548. https://doi.org/10.1093/bib/bbu015.

Stamatakis, A., 2016. The RAxML v8.2.X manual. Heidelberg Institute for Theoretical Studies. https://cme.h-its.org/exelixis/resource/download/NewManual.pdf (accessed 9 July 2022).
Streicher, J.W., Schulte, J.A., Wiens, J.J., 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. 65, 128–145. https://doi.org/10.1093/sysbio/syv058.
Sugiura, N., 1978. Further analysis of the data by Anaike's information criterion and the finite corrections. Commun. Stat. - Theory Methods 7, 13–26. https://doi.org/10.1080/03610927808827599.
Tripp, E.A., Tsai, Y.H.E., Zhuang, Y., Dexter, K.G., 2017. RADseq dataset with 90% missing data fully resolves recent radiation of *Petalidium* (Acanthaceae) in the ultra-arid deserts of Namibia. Ecol. Evol. 7, 7920–7936. https://doi.org/10.1002/ece3.3274.
Wada, S., Chiba, S., 2013. The dual protection of a micro land snail against a micro predatory snail. PLoS One 8, e54123–e54123. https://doi.org/10.1371/journal.pone.0054123.
Wagner, N.D., Gramlich, S., Hörandl, E., 2018. RAD sequencing resolved phylogenetic relationships in European shrub willows (*Salix* L. subg. *Chamaetia* and subg. *Vetrix*) and revealed multiple evolution of dwarf shrubs. Int. J. Bus. Innov. Res. 17, 8243–8255. https://doi.org/10.1002/ece3.4360.
Wang, X., Ye, X., Zhao, L., Li, D., Guo, Z., Zhuang, H., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (Poaceae: Bambusoideae). Sci. Rep. 7, 11546. https://doi.org/10.1038/s41598-017-11367-x.
Wilkinson, M., Crotti, M., 2017. Comments on detecting rogue taxa using RogueNaRok. Syst. Biodivers. 15, 291–295. https://doi.org/10.1080/14772000.2016.1252440
Winnepenninckx, B., Backeljau, T., De Wachter, R., 1993. Extraction of high molecular weight DNA from molluscs. Trends Genet. 9, 407. https://doi.org/10.1016/0168-9525(93)90102-N.

## Figure Captions

**Figure 1.** Distribution of Sri Lankan *Corilla*, showing range limits (dotted lines) of species and the 30 localities sampled for this study (red dots; sample codes follow **Table 1**, with asterisks indicating sampled type localities). The range limits (approximating to extent of occurrence *sensu* Gaston, 1991) are based on unpublished distributional data from field surveys and museum collections (primarily two Sri Lankan institutions, the Colombo Museum and the University of Peradeniya, and the NHM, London, UK). **A**. The lowland species *Corilla adamsi*, *C. carabinata* and *C. colletti*, and the montane species *C. beddomeae*, *C. gudei* and *C. humberti*. **B.** The montane species *C. erronea*, *C. fryae* and *C. odontophora*. For clarity, we show the range limits of *C. gudei* 'north' and 'south', and *C. odontophora* 'east', 'north' and 'south'. The Central Highlands, Knuckles Massif and Rakwana Massif are the three main mountain masses of Sri Lanka; the Uva Basin lies in the eastern half of the Central Highlands (Cooray, 1984).

**Figure 2.** Shell and external morphology of *Corilla*. **A–F.** Shells of the lowland species *Corilla adamsi* (**A**, **E**) and *C. carabinata* (**D**), and the montane species *C. erronea* (**B**, **F**) and *C. beddomeae* (**C**); white arrows indicate the lip, which is wide and strongly reflected in the lowland species and narrow in montane species. **G, H.** The shell sculpture of *Corilla* consists of lines that correspond with the direction of growth of the shell and run parallel to the edge of the lip (i.e. collabral sculpture *sensu* Cox, 1960). The collabral sculpture (indicated by black arrows) may be bold, consisting of well-defined raised lines, as in *C. adamsi* (**G**). Or, as in the case of *C. humberti* (**H**), it may be faint, being composed of shallow, incised lines (striae). **I–K.** The upper palatal folds (shown by white arrows), as viewed through the outer wall of the shell. These may vary in length, being long in *C. carabinata* and short in *C. gudei* and *C. beddomeae*. The upper palatal folds also vary in orientation: they are parallel in *C. beddomeae* (**K**) and non-parallel (converge posteriorly) in *C. carabinata* (**I**) and *C. gudei* (**J**). **L, M.** Live examples of the lowland *C. adamsi* (**L**) and montane *C. humberti* (**M**). Dark mottling on the part of the body within the shell, as in **L**, may sometimes be visible through the shell wall of lowland species; the shell itself lacks such markings. The very dark shell and body colour in **M** is typical of all montane species in their living state (the colour of empty shells found in the field are usually faded, resembling **B**, **C**, **F** and **K**). Image credits: **A–K**, H. Taylor; **L**, **M**, S. Nanayakkara.

**Figure 3.** Variation in total branch support, topology and assembly size across the four ipyrad parameter combinations for 17–87% missing individuals per locus (88 assemblies in total). In the heatmap, the columns represent the four ipyrad parameter combinations, with 'c' denoting the clustering threshold (0.85 or 0.90) and 'd' the minimum depth (7 or 9). Each cell in the heatmap represents a single assembly. The rows are levels of missing data, which are shown as percentages with absolute values within parentheses. The columns are the four parameter combinations, with each column comprising 22 assemblies. The topology of the ML consensus tree obtained by RAxML for a given assembly is denoted by one of the letters 'A/a' to 'k'; 16 distinct ML consensus trees were obtained in total for the 88 assemblies, with topologies A and B being the most common. Uppercase lettering indicates that all 16 focal branches are strongly supported, lowercase lettering that at least one of these branches is not strongly supported. The approximate size (x $10^6$ bp) of each assembly is given at the bottom left of the corresponding cell.

**Figure 4.** The two main ML topologies recovered for *Corilla* (30 individuals). The key difference between topologies A and B is in the composition of clade 3 (the arrow shows the branch subtending this clade). Branches are colour-coded as follows: black, montane taxa; green, lowland taxa; and grey, *C. odontophora* 'south'. Where relevant, taxon labels include the abbreviations 'n' for north, 's' for south or 'e' for east (e.g. *C. gudei* 'north' is shown as *C. gudei* 'n'). The 16 focal branches are indicated by the larger, bold numbers (above the relevant

branches). The smaller numbers are bootstrap support values (only values ≥70% are shown); maximal support is indicated by an asterisk. The scale bar indicates substitutions per site.

**Figure 5.** Variation in assembly length (**A**), number of loci (**B**), number of variable sites (**C**) and overall amount of missing data (**D**) relative to % missing individuals per locus for the four parameter combinations. The combinations are coded as 'c' (clustering threshold, 0.85 or 0.90) and 'd' (minimum depth, 7 or 9).

**Figure 6.** Coverage of loci per individual for assemblies generated with a clustering threshold of 0.85, minimum depth of 7 (ipyrad parameter combination c85d7) and missing individuals per locus ranging from 40–87%. In the heatmap, columns represent different assembly sizes and levels of missing data, and rows are different individuals; a single cell represents the proportion of loci in an assembly for which data are available for a given individual (% coverage of loci per individual). Sample codes follow **Figure 4**. The arrow indicates clade 3, for which relationships differ between topologies A and B. Colour-coding of branches: black, montane taxa; green, lowland taxa; grey, *Corilla odontophora* 'south'.

**Figure 7.** Reconstructions of ancestral character states for habitat association (topology B). Habitats are colour-coded black for montane and grey for lowland. Circles at terminals represent current (observed) character states. **A.** MP reconstructions. Circles at nodes denote reconstructed states. **B.** ML-based reconstructions. Pie charts above and below nodes denote reconstructed state probabilities for the Mk1 and AsymmMk models, respectively. Numerical probabilities for reconstructed states are shown in **Supplementary Material Tables S4** (Mk1 model) and **S5** (AsymmMk model).

**Figure 8.** Reconstructions of ancestral character states for the three shell characters, shape, colour and lip width (topology B). At each node, shape, colour and lip width are shown, respectively, from left to right as separate, differently colour-coded circles/pie charts. Circles at terminals represent current character states. **A.** MP reconstructions. Circles at each node denote reconstructed states. **B.** ML reconstructions. Pie charts above and below internal branches denote state probabilities reconstructed under the Mk1 and AsymmMk models, respectively. Numerical probabilities for reconstructed states are shown in **Supplementary Material Tables S4** (Mk1 model) and **S5** (AsymmMk model).

**Figure 9.** Reconstructions of ancestral character states for the three shell characters, length of upper palatal folds, orientation of upper palatal folds and collabral sculpture (topology B). At each node, length of upper palatal folds, orientation of upper palatal folds and collabral sculpture are shown, respectively, from left to right as separate, differently colour-coded circles/pie charts. Circles at terminals represent current character states. **A.** MP reconstructions. Circles at each node denote reconstructed states. **B.** ML reconstructions. Pie charts above and below internal branches denote state probabilities reconstructed under the Mk1 and AsymmMk models, respectively. Numerical probabilities for reconstructed states are shown in **Supplementary Material Tables S4** (Mk1 model) and **S5** (AsymmMk model).

**Table 1**
Taxa used in this study with sample codes, locality data and selected read statistics of the ipyrad assembly process. Sampled localities with associated sample codes are shown in **Figure 1**. The use of 'east'. 'north' or 'south' for *Corilla gudei* and *C. odontophora* indicate if individuals originate from the eastern ('e'), northern ('n') or southern ('s') parts, respectively, of the ranges of these taxa. Precise and accurate type locality data are only available for *C. colletti*, *C. fryae* and *C. odontophora* (see Raheem et al., 2017); all of these type localities (shown by asterisk after sample code) have been sampled. District names are abbreviated as: BAD, Badulla; GAL, Galle; KAL, Kalutara; KAN, Kandy; KEG, Kegalle; MAT, Matara; NUW, Nuwara Eliya; RAT, Ratnapura. The ipyrad statistics are reads available after demultiplexing, filtering and editing, and mean cluster depth after *de novo* clustering of reads within samples.

| | Taxon | Locality | No. of filtered reads | Mean depth of clusters (no. of reads) | |
|---|---|---|---|---|---|
| | | | | Clustering threshold 0.85 | Clustering threshold 0.90 |
| 1 | *Corilla adamsi* | 1.6 km ne Ingiriya, KAL | 741,260 | 18.76 | 17.31 |
| 2 | *C. adamsi* | Homadola, near Udugama, GAL | 241,864 | 8.95 | 8.34 |
| 3 | *C. beddomeae* | Adam's Peak (eastern face), NUW | 586,716 | 11.76 | 12.72 |
| 4 | *C. beddomeae* | 7.5 km se Maskeliya, NUW | 1,224,730 | 18.88 | 18.16 |
| 5 | *C. beddomeae* | 4.5 km se Pupuressa, KAN | 1,327,308 | 13.99 | 13.59 |
| 6 | *C. carabinata* | 1.6 km ne Dedugala, KEG | 368,295 | 21.79 | 19.37 |
| 7 | *C. carabinata* | Padavigampola, near Rambukkana, KEG | 741,374 | 12.79 | 11.69 |
| 8 | *C. carabinata* | Matale, MAT | 623,622 | 14.07 | 13.38 |
| 9* | *C. colletti* | Balangoda, RAT | 728,201 | 5.29 | 4.85 |
| 10 | *C. colletti* | Rajawaka, RAT | 462,071 | 8.39 | 7.82 |
| 11 | *C. erronea* | above Northcove Estate, Bogowantalawa, NUW | 455,757 | 21.89 | 19.67 |
| 12 | *C. erronea* | 8 km sw Maskeliya, NUW | 406,006 | 27.14 | 24.17 |
| 13 | *C. erronea* | 4.8 km nw Ambewela, NUW | 398,850 | 13.82 | 12.56 |
| 14 | *C. erronea* | Kikilimana, NUW | 287,332 | 24.58 | 21.7 |
| 15 | *C. erronea* | 5.1 km sw Aranayaka, KEG | 338,813 | 23.07 | 20.75 |
| 16 | *C. erronea* | Galheeria Estate (1360 m), above Elkaduwa, MAT | 733,087 | 27.75 | 25.52 |
| 17* | *C. fryae* | Albion Estate, 7km sw Lindula, NUW | 527,745 | 25.11 | 22.5 |
| 18 | *C. fryae* | Adam's Peak (eastern face), NUW | 699,571 | 15.63 | 14.26 |
| 19 | *C. fryae* | above Queensberry Estate, Dimbula, NUW | 450,606 | 17.68 | 15.98 |
| 20 | *C. gudei* 'north' | Laggala, above Illukkumbura, MAT | 641,166 | 19.1 | 18.04 |
| 21 | *C. gudei* 'north' | Galheeria Estate (1280 m), above Elkaduwa, MAT | 786,469 | 22.57 | 20.77 |
| 22 | *C. gudei* 'south' | Nawanagala, KAN | 523,744 | 24.48 | 22.41 |
| 23 | *C. humberti* | Between Nuwara Eliya and Hakgala, NUW | 833,545 | 17.08 | 15.81 |
| 24 | *C. humberti* | Nuwara Eliya, NUW | 840,897 | 14.39 | 13.4 |
| 25* | *C. odontophora* 'east' | Namunukula, above Bibilegama, BAD | 746,590 | 16.63 | 15.52 |
| 26 | *C. odontophora* 'east' | 5.7 km nw Lunugala, BAD | 631,138 | 12.76 | 11.53 |
| 27* | *C. odontophora* 'north' | 2.4 km ne Paranagama, NUW | 284,815 | 21.88 | 19.59 |
| 28* | *C. odontophora* 'north' | 5.2 km ne Paranagama, NUW | 786,806 | 14.49 | 13.55 |
| 29* | *C. odontophora* 'south' | Obadaella, near Bandarawela, BAD | 310,754 | 12.74 | 11.87 |
| 30 | *C. odontophora* 'south' | Thangamalai, Haputale, BAD | 640,195 | 16.85 | 15.73 |

**Table 2**
Key parameter settings for ipyrad data assembly and the total number of assemblies (and ML consensus trees) generated for each of the four parameter combinations. A few assemblies had sequence data for only 29 of the 30 individuals, hence the differing totals for assemblies and/or trees.

| | **Parameter combination** | | | |
|---|---|---|---|---|
| | **1. c85d7** | **2. c85d9** | **3. c90d7** | **4. c90d9** |
| **Parameter** | | | | |
| Clustering threshold | 0.85 | 0.85 | 0.90 | 0.90 |
| Minimum depth | 7 | 9 | 7 | 9 |
| Minimum individuals per locus | 4–30 | 4–30 | 4–30 | 4–30 |
| **Number of assemblies (and ML trees)** | 27 (27) | 26 (26) | 27 (27) | 26 (25) |

**Table 3**
Support for clades in the well resolved ML consensus trees obtained for the four ipyrad parameter combinations. The trees are those associated with the 15 largest assemblies (levels of missing individuals per locus range from 50–87%). Clades are numbered as in **Figure 4**; 'A' and 'B' indicate clades unique to topologies A and B, respectively. Bootstrap supports are given as means, and for clades with non-maximal support the range in absolute values is also shown (i.e. within parentheses). Maximal support is indicated by an asterisk.

| **Clades** | **Bootstrap support (%)** | | | |
|---|---|---|---|---|
| | **1. c85d7** | **2. c85d9** | **3. c90d7** | **4. c90d9** |
| 1: Clade 1 | * | * | * | * |
| 2: *C. erronea–fryae* | * | * | * | * |
| 3: Clade 3 | 98.9 (95–100) | 99.9 (99,100) | 99.8 (99,100) | 99.8 (98,100) |
| 4A: *C. odontophora* 'south' + ((*C. odontophora* 'east' + *C. colletti*) + (*C. gudei* 'north' + *C. carabinata*)) | 76.3 (65–85) | - | 89 (73–97) | 85 (61–95) |
| 4B: *C. odontophora* 'south' + (*C. odontophora* 'north' + (*C. gudei* 'south' + *C. humberti*)) | 83.5 (60–99) | 89.4 (61–100) | 85 (63–97) | 76 (42–97) |
| 5: (*C. odontophora* 'east' + *C. colletti*) + (*C. gudei* 'north' + *C. carabinata*) | * | * | * | * |
| 6: *C. odontophora* 'north' + (*C. gudei* 'south' + *C. humberti*) | * | * | * | * |
| 7: *C. gudei* 'south' + *C. humberti* | * | * | * | * |
| 8 : *C. odontophora* 'east' + *C. colletti* | 91.1 (49–100) | 90.2 (55–100) | 99.3 (95–100) | 94.9 (85–100) |
| 9 : *C. gudei* 'north' + *C. carabinata* | * | * | * | * |
| 10: *C. odontophora* 'south' | * | * | * | * |
| 11: *C. odontophora* 'north' | * | * | * | * |
| 12: *C. humberti* | * | * | * | * |
| 13: *C. odontophora* 'east' | * | * | * | * |
| 14: *C. colletti* | * | * | * | * |
| 15: *C. gudei* 'north' | * | * | * | * |
| 16: *C. carabinata* | * | * | * | * |
| *C. adamsi* + clade 1 (ingroup) | * | * | * | * |
| *C. adamsi* | * | * | * | * |
| *C. beddomeae* | * | * | * | * |

Figure 1

Click here to access/download;Figure;Raheem-et-al_fig1.jpg

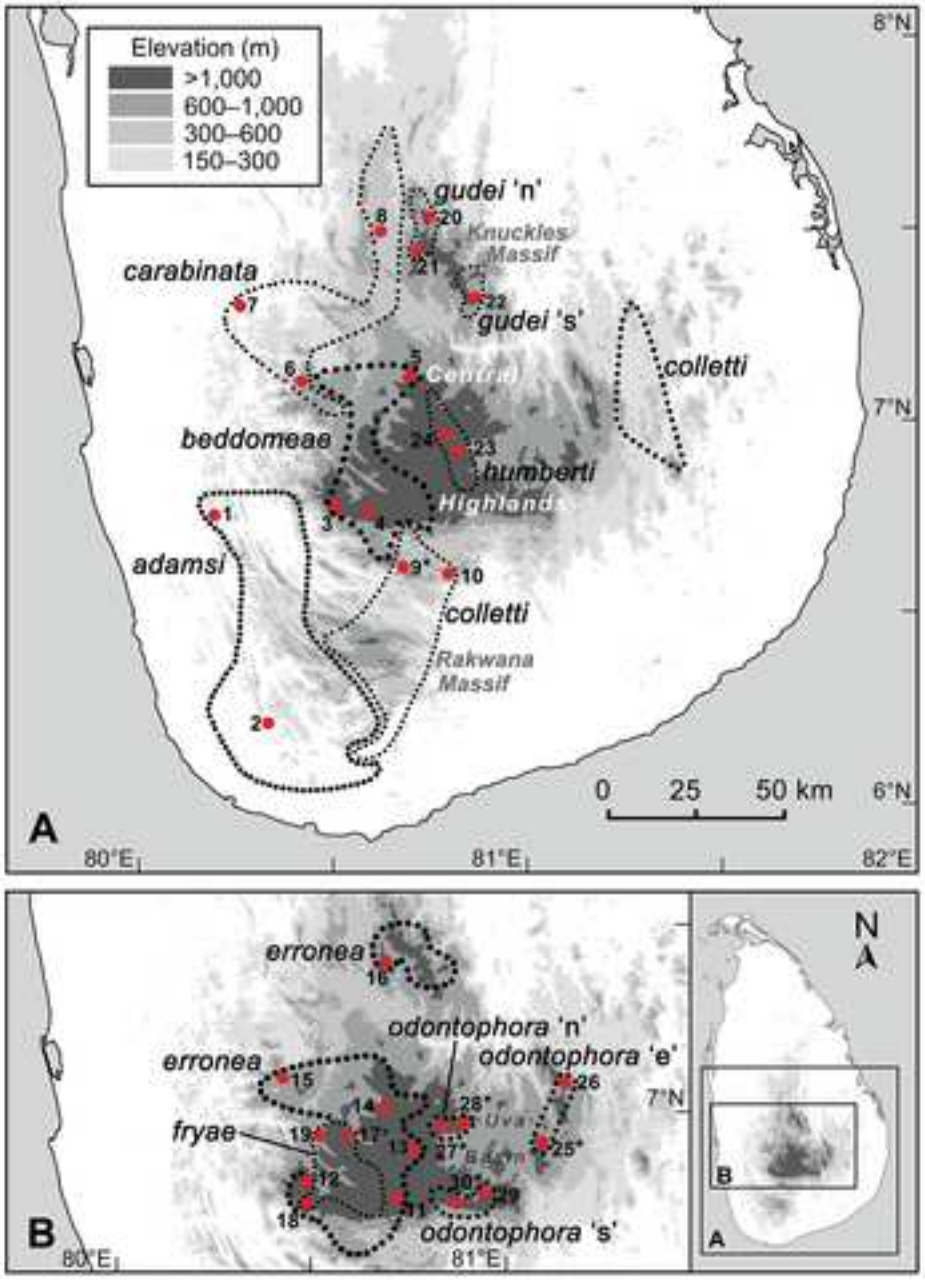

Figure 2



Figure 3



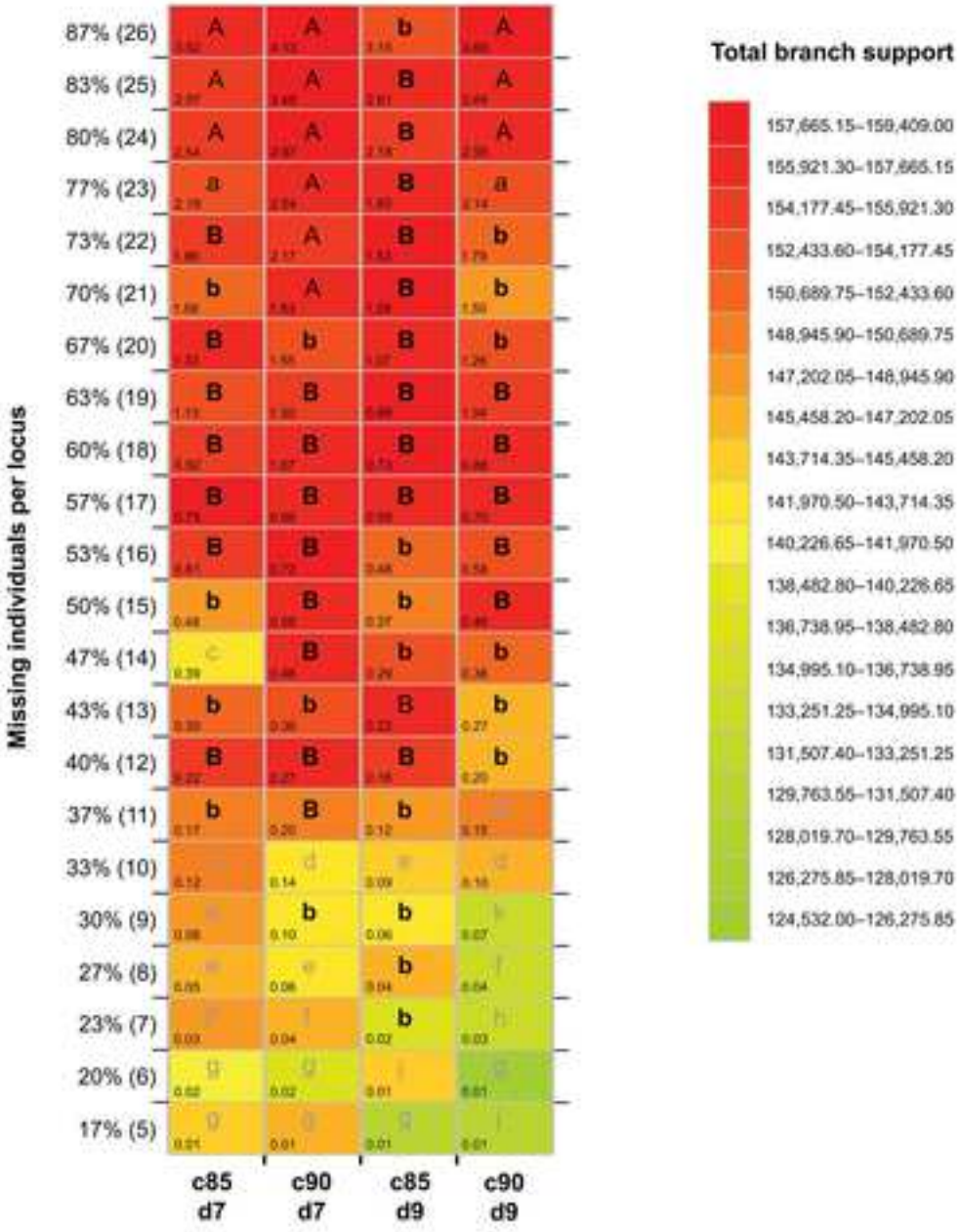

Figure 4



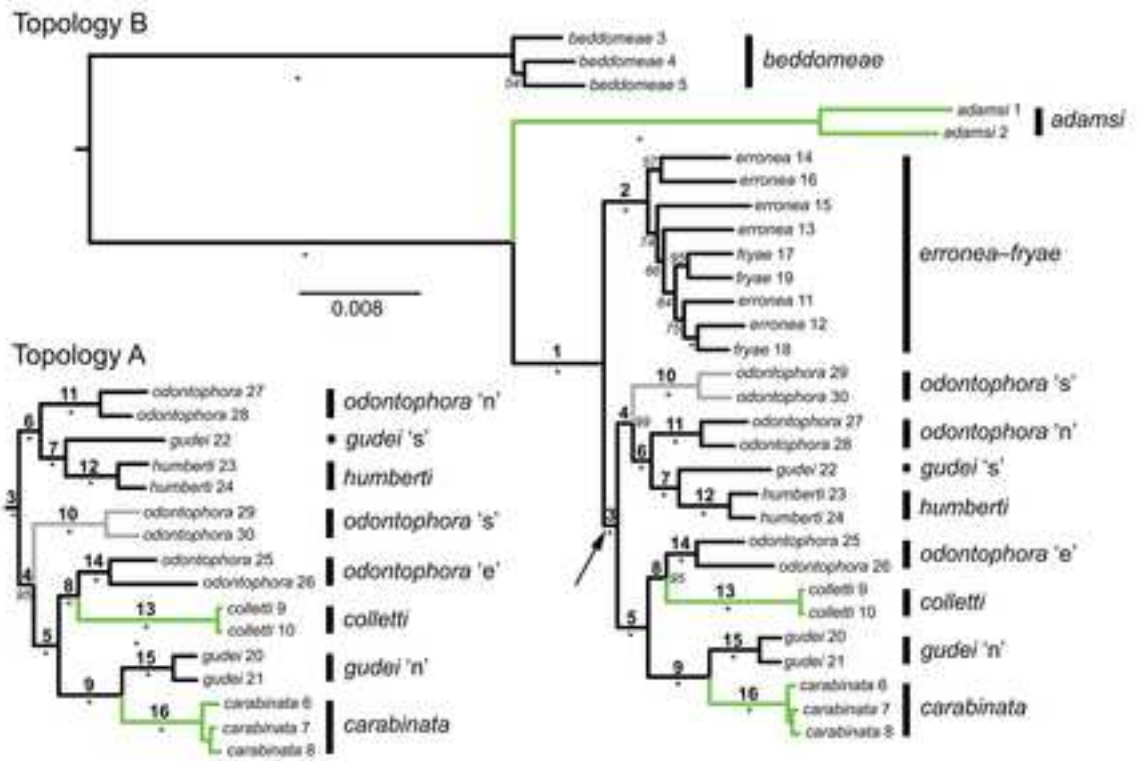

Figure 5



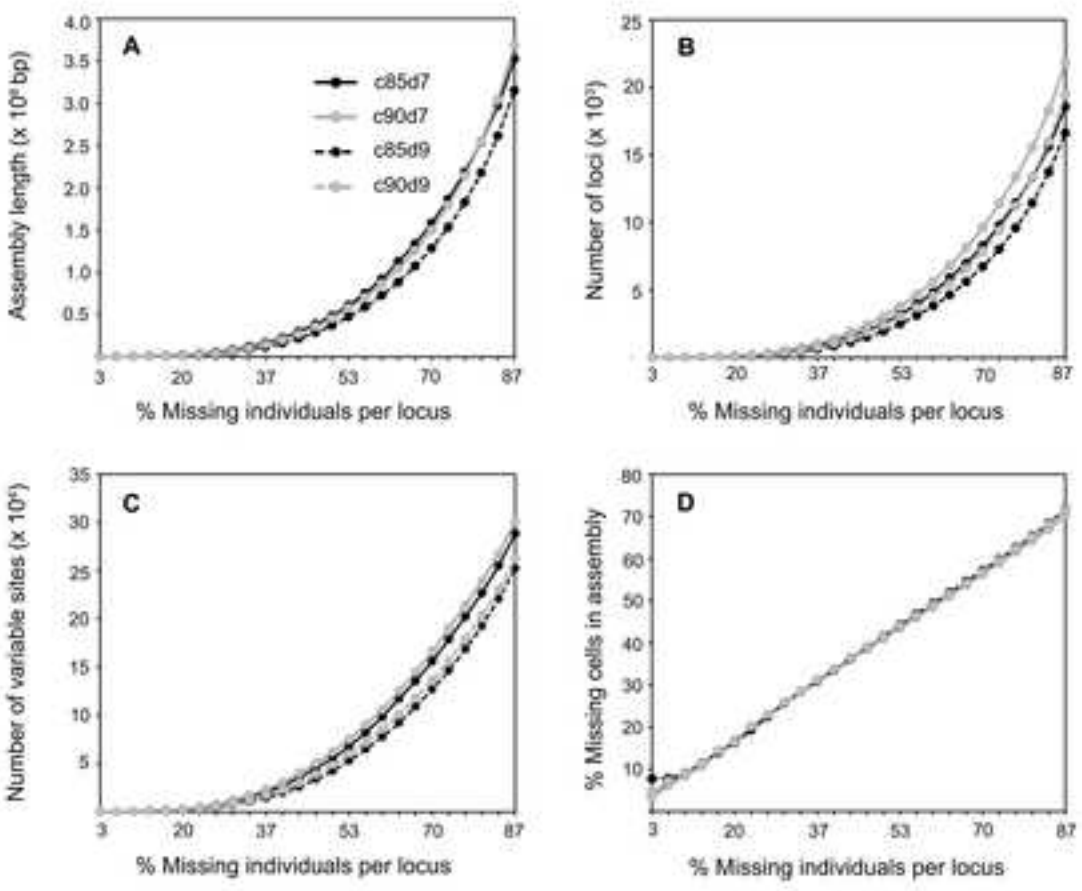

Figure 6



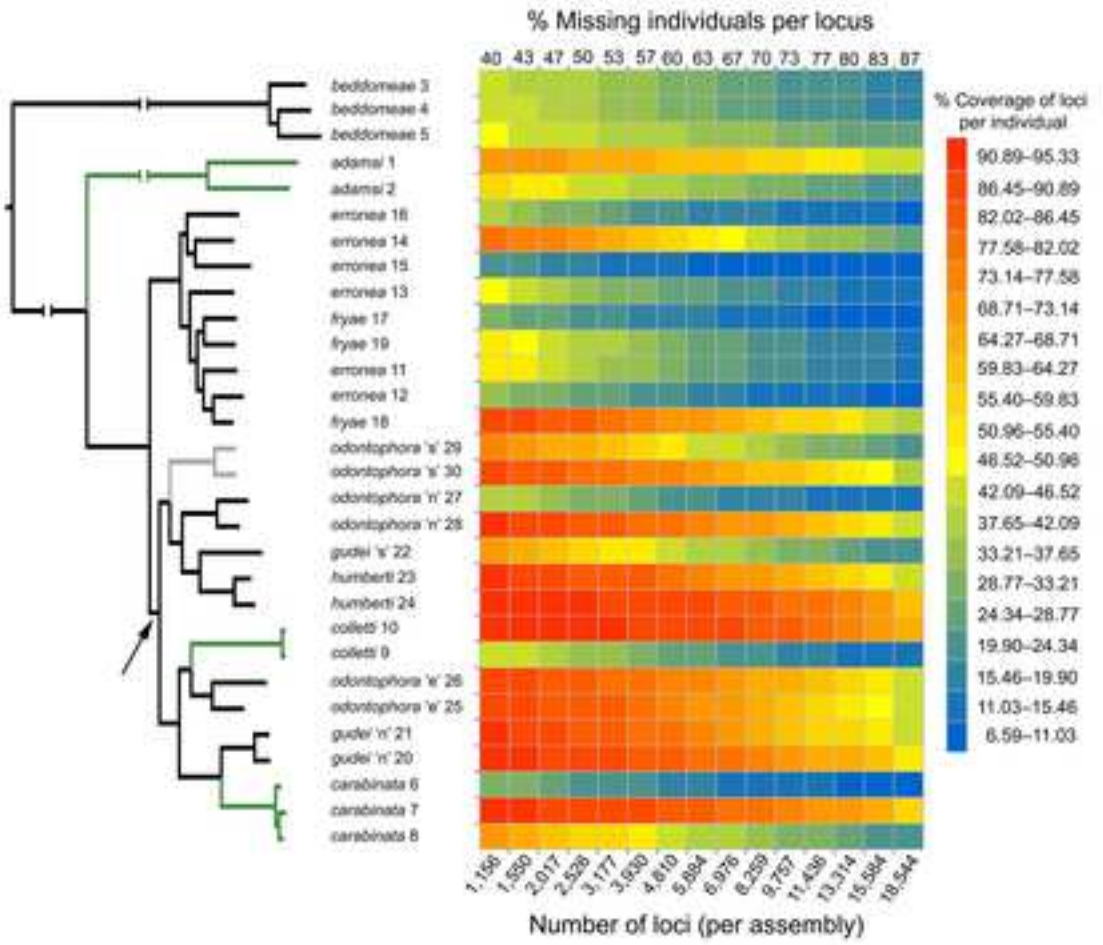

Figure 7



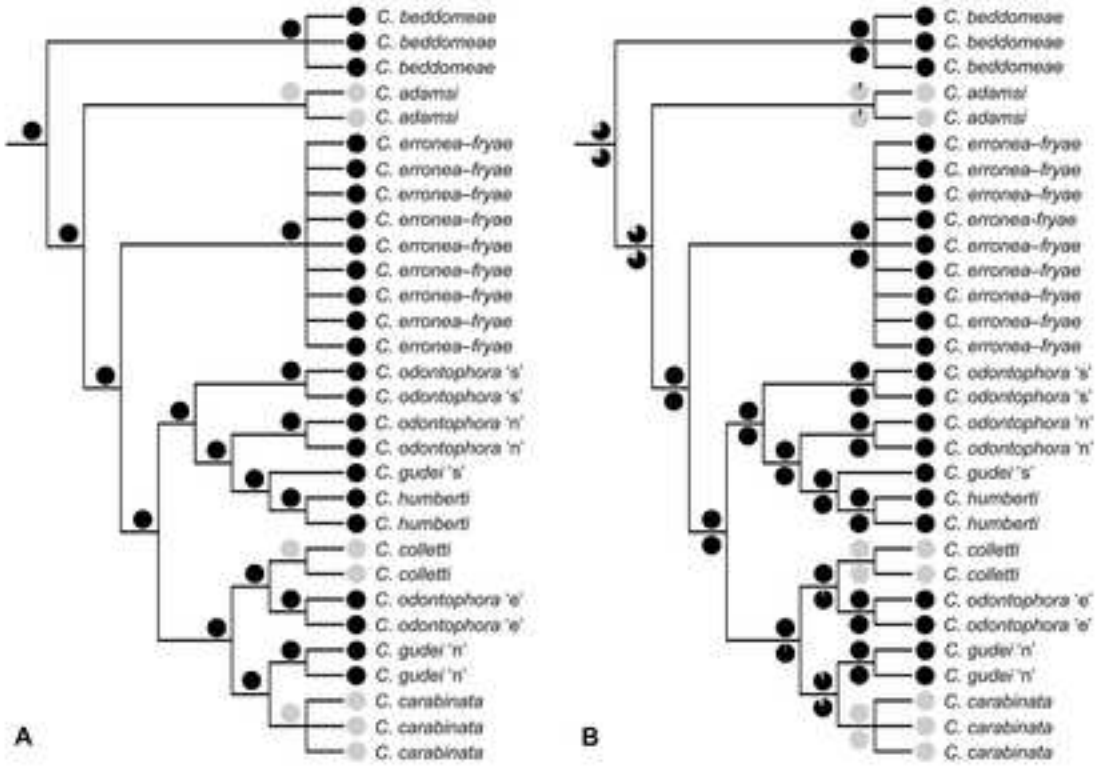

Figure 8



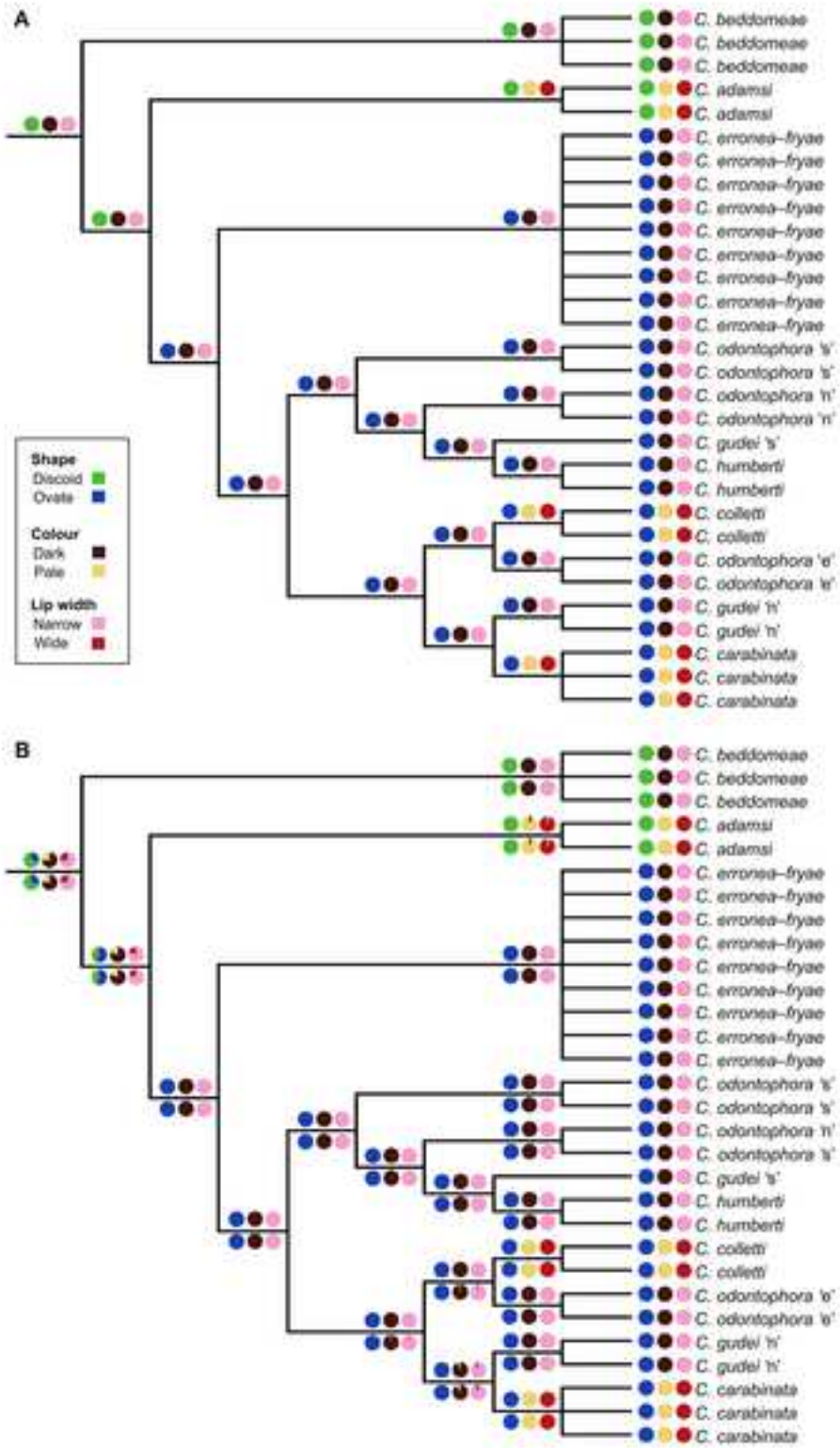

Figure 9

Click here to access/download;Figure;Raheem-et-al_fig9.jpg