MOLECULAR ECOLOGY

Collembola, the biological species concept, and the underestimation of global species richness.

Journal:	*Molecular Ecology*
Manuscript ID:	MEC-13-0466
Manuscript Type:	Original Article
Date Submitted by the Author:	03-May-2013
Complete List of Authors:	Cicconardi, Francesco; Sapienza - Univeristy of Rome, Department of physics Emerson, Brent; Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), Island Ecology and Evolution Research Group
Keywords:	Bioinfomatics/Phyloinfomatics, Insects, Phylogeography, Speciation, Systematics

COLLEMBOLA AND THE BIOLOGICAL SPECIES CONCEPT

Collembola, the biological species concept, and the underestimation of global species richness.

Francesco Cicconardi[1,2,*] and Brent C. Emerson[3,4]

[1]*Department of Physics, University of Rome "La Sapienza", P.le A. Moro 5, 00185 Rome, Italy*

[2]*Smithsonian Tropical Research Institute, Apartado 2072, Balboa, Republic of Panama*

[3]*Island Ecology and Evolution Research Group, IPNA-CSIC, 38206 La Laguna, Tenerife, Canary Islands, Spain*

[4]*School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.*

** Correspondence to be sent to: Francesco Cicconardi, Department of Physics, Sapienza, University of Rome, P.le A. Moro 5, 00185 Rome, Italy; E-mail: francicco@gmail.com*

Abstract. - Despite its ancient origin, global distribution, and abundance in nearly all habitats, the class Collembola is comprised of only 8,000 described species, and is estimated to number no more than 50,000. Many morphologically defined species have broad geographic ranges that span continents, and recent molecular work has revealed high genetic diversity within species. However, the evolutionary significance of this genetic diversity is unknown. In this study we sample five morphological species of the globally distributed genus Lepidocyrtus from 14 Panamanian sites to characterise genetic diversity and test morphospecies against the biological species concept. Mitochondrial and nuclear DNA sequence data was analysed and a total of 58 molecular lineages revealed. Very deep lineage diversification was recovered, with 30 evolutionary lineages estimated to have established more than 10 Mya, and almost all contemporary lineages preceding the Pleistocene (~2 Mya). Thirty-four lineages were sampled in sympatry revealing unambiguous co-segregation of mitochondrial and nuclear DNA sequence variation, consistent with biological species. Biological species richness within the class Collembola, and the geographic structure of this diversity, are substantially misrepresented components of terrestrial animal biodiversity. We speculate that global species richness could be at least an order of magnitude greater than a previous estimate of 50,000 species.

Keywords: *Lepidocyrtus*, molecular lineage, biodiversity, cryptic species, Panama

Soil and canopy rainforest communities comprise representatives of almost all higher order terrestrial taxonomic groups. They are tremendously diverse, both at local and global scales (Ødegaard 2000; Lavelle *et al.* 2006), and it has been estimated that as much as 23% and 40% of extant species reside in soil and canopy habitats respectively (Ozanne *et al.* 2003; Decaëns 2010). The quantification and spatial characterization of this biodiversity is a key goal for biologists, however, much of this diversity comprises animal species of very small body size, and as in any ecosystem the void of knowledge is negatively related to the size class of organisms. Within both soil and canopy communities the mesofauna (from 100 µm to 2 mm in size) of both habitats suffers from a conspicuous deficit of interest from the scientific community (Decaëns 2010). Collembola (springtails) are a class of wingless hexapods, and are often the most abundant mesofaunal arthropods collected in rainforest canopies (Rodgers & Kitching 2010) and soil ecosystems (Hopkin 1997). Despite their worldwide distribution they are only moderately diverse in terms of species number (approximately 8,000 described species) if compared with other hexapod groups (e.g. 360,000-400,000 species within the order Coleoptera alone (Scheffers *et al.* 2012)). Although significant progress that has been made regarding Collembola taxonomy over recent decades, with the identification of important new morphological characters and a significant growth of newly described species (Deharveng 2004), global species richness remains low. Estimates of global species richness within the hexapod class Insecta are in the order of 5,000,000 (Gaston 1991), while for Collembola, the other major hexapod class, global species richness is estimated to be only approximately 50,000 species when one takes into account the under-sampling of species with restricted distributions (Hopkin 1997). However, in addition to geographic under-sampling of local endemisms, recent molecular genetic studies present evidence that springtail biodiversity may be vastly underestimated because of pervasive cryptic species diversity within morphologically defined species of Collembola (e.g. Emerson et al. 2011; Porco et al. 2012a). Many morphologically defined species investigated to date have revealed remarkable levels of deeply divergent DNA sequence diversity at both broad and fine geographical scales (e.g. Timmermans et al. 2005; Stevens et al.

2006; Garrick et al. 2007; Cicconardi et al. 2010; Torricelli et al. 2010; Porco et al. 2012b; Ramirez-Gonzalez et al. 2013). More recently Emerson et al. (2011) have analysed geographically referenced barcode sequences from the mitochondrial DNA (mtDNA) *COI* gene belonging to 105 morphologically defined species. Their analyses revealed that highly divergent mtDNA sequence diversity within morphospecies is a taxonomically and geographically pervasive feature within Collembola, and most likely a signature of cryptic species diversity.

A hypothesis of pervasive cryptic species diversity is also consistent with the Collembola fossil record. From the geochronological point of view springtails have the most extensive record among all hexapod orders. They stretch from the Lower Devonian ~ 410 million years ago (Mya) (Hirst & Maulik 1926) to the Pleistocene ~ 1 Mya (Yosii 1974). There are two species reported from the Paleozoic: *Permobrya mirabilis* from the Upper Permian of South Africa (Riek 1976), and *Rhyniella praecursor* (397-391 Mya), originally described from the Rhynie Chert of England (Hirst & Maulik 1926). The latter fossil was redescribed several times and finally placed in the modern family Isotomidae because of the remarkably modern morphological characters (Whalley & Jarzembowski 1981; Greenslade & Whalley 1986). Baltic amber fossils from the Early Eocene (50-45 Mya) are dominated by species belonging to contemporary genera, and in some cases these have been assigned to extant species (Rapoport 1971). Morphological identity between fossil and modern assemblages is also reported in Early Miocene amber from Chiapas (23-16 Mya), Mexico, and from the Dominican Republic (23-20 Mya) where seven specimens could be placed into extant species (Christiansen 1971; Mari-Mutt 1983a). This propensity for morphological stasis over long periods of time fits well with the hypothesis that classically defined morphospecies are an underestimation of the true collembolan biodiversity.

Many Collembola genera are distributed globally, with the distributions of many species spanning continents. While extensive dispersal can be invoked to explain such distributions, it remains equally plausible that the broad distributions of many morphospecies could be representative of cryptic species diversity. The genus *Lepidocyrtus* Bourlet, 1839 (Entomobryidae:

Lepidocyrtinae) is one of the largest genera of Collembola (Hopkin 1997). It has a worldwide distribution and includes more than 250 described species, many of which have distributions spanning more than one continent. A recent molecular analysis of the seven *Lepidocyrtus* morphospecies inhabiting the North-Western Mediterranean basin, using both mitochondrial and nuclear DNA sequence data, has revealed lineage diversity within the seven species that is remarkable, both in terms of lineage number and age of origin (Cicconardi *et al.* 2010). Samples from 24 distinct geographic locations revealed 52 molecular lineages predating the onset of the Pleistocene 1.8 Mya, and 35 lineages originating more than 5.8 Mya. While the results of Cicconardi et al. (2010) are remarkable, the implications for cryptic diversity within the Collembola on a global scale are profound when one considers the broader geographical ranges of the seven morphospecies investigated. Their ranges extend far beyond the North-Western Mediterranean basin, with five of them spanning the Palaearctic and Nearctic regions. While evidence is mounting that morphologically defined species of Collembola are frequently comprised of divergent genetic lineages, the significance of these in term of species is unclear. Cicconardi et al. (2010) obtained some evidence that the genetic lineages they identified conform to species under Mayr's (1942) biological species concept (BSC), however their evidence was limited to an assessment of only two sympatric lineages.

The worldwide distribution of the genus *Lepidocyrtus*, provides ample opportunity for extending the sampling of Cicconardi et al. (2010) to investigate the temporal and spatial properties, and biological significance of cryptic lineage diversity within morphologically defined species of Collembola. The Isthmus of Panama is a narrow strip of land which links Central America with South America and it is among of the five most important global biodiversity hotspots (Myers *et al.* 2000; Briggs 2007; Joppa *et al.* 2011). For the purpose of our study the Isthmus represents a very distinct biogeographic region from that sampled by Cicconardi et al. (2010) within which to investigate morphospecies within the genus *Lepidocyrtus*. Contrasting with temperate species, that typically exhibit a more habitual association with soil and leaf litter, tropical collembolan species

are in fact ecologically distinct from their temperate relatives, exhibit different seasonality and a closer association with the sub-canopy. The combined processes of plate disassembly and redistribution, phases of global warming and cooling (Molnar 2008), together with well documented dating of volcanic arcs and geological formations (Coates *et al.* 2003; Wegner *et al.* 2010; Montes *et al.* 2012) also renders the Isthmus of Panama an informative region for biogeographic inference. In this study we sample five morphospecies within the genus *Lepidocyrtus* from 14 sites spanning the Isthmus of Panama for the quantification of molecular lineage diversity, its temporal and spatial structure, and the consistency of lineages with the BSC. To achieve this we analyse DNA sequences from the mtDNA cytochrome c oxidase subumit II (*COII*) gene and the nuclear elongation factor 1α (*EF1α*) gene. We reconstruct lineage origin in a temporal and spatial framework using molecular phylogenetic tools, a molecular clock for the *Lepidocyrtus* and geological data. We then evaluate divergent molecular lineages within morphospecies against the BSC, testing Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD) among lineages in sympatry.

METHODS

Data Collection

Samples were collected from 14 sites across the Isthmus of Panama (Fig. 1; Table 1). Specimens were collected from leaf litter and from plants to a height of two metres using Berlese/Tullgren funnel and an automatic aspirator and stored at -80 °C in 95% alcohol. Specimens belonging to the genus *Lepidocyrtus* Bourlet, 1839 were identified under a stereo microscope using characters as described in Gisin (1964a; b, 1965) and classified to morphospecies according to the literature available for Neotropical and Panama / Costa Rica springtails (Denis 1933; Mari-Mutt 1983b, 1986; Mari-Mutt & Bellinger 1990, 1996). Total genomic DNA was extracted using the DNeasy Tissue Kit (Qiagen Inc., Valencia, CA, USA) with the following variations. For all samples the entire specimen were soaked in lysis buffer at 50 °C for 12-24h, depending on the specimen size, in order

to allow for DNA extraction and the preservation of the thin cuticle. The cuticles were subsequently mounted on slides for further morphological analysis. The mtDNA gene *COII* was amplified via the polymerase chain reaction (PCR) using a combination of primers described in the literature (Cicconardi et al. 2010). The nuclear gene *EF1α* amplification was accomplished using a combination of primers designed specifically for *Lepidocyrtus* (EF1α_LepF: 5'-CAT GAT YAC KGG TAC CTC TCA-3'; EF1α_LepR1: 5'-GCA TCM CCM GAT TTG ATA GC-3'; EF1α_LepR2: 5'-GCC TCA ACG CAC ATG GGC TTA-3'). PCR products were gel purified (Wizard SV Gel and PCR Clean-Up kit, Promega) and sequenced in both directions at the John Innes Centre (Norwich, UK). Sequences were manually checked and corrected using FINCHTV v. 1.4.0 (Geospiza Inc., Seattle, WA, USA). Double peaks in the nuclear sequences, corresponding to allelic differences in heterozygous individuals, were scored and PHASE v. 2.1.1 (Stephens *et al.* 2001) was used to resolve the two alleles for heterozygous individuals, as in Cicconardi et al. (2010). All alleles that could not be phased were left unspecified and coded in accordance with the IUPAC code table. Uncorrected pairwise distances (π) (Nei 1987) were calculated with PAUP* v. 4.0b10 (Swofford 2003). The partitioning of genetic variance within both the *COII* and *EF1α* datasets among morphological species, among sampling sites, and within sampling sites was measured using an AMOVA strategy as implemented in ARLEQUIN v. 3.5.1.2 (Excoffier & Lischer 2010), and its statistical significance tested based on 10,000 permutations.

Phylogenetic analyses and divergence time estimation

Sequences from both markers were manually aligned with SEAVIEW v. 4.1 (Gouy *et al.* 2010) taking into account the translated amino acid sequence. For *COII* sequences all gaps were retained in the alignment. For *EF1α*, forward and reverse sequences were first assembled as contigs and the intron regions removed due to the inability to align them unambiguously. For maximum likelihood (ML) and Bayesian inference (BI) analyses the most appropriate substitution model was selected excluding invariant characters (I) to avoid over-fitting. For each of the two loci the best nucleotide

substitution model was inferred among 44 (for GARLI and BEAST) and 12 models (for MRBAYES for its more restricted model selection list) according to the Akaike Information Criterion as implemented in JMODELTEST v. 0.1.1 (Posada 2008). Three individuals from undescribed *Seira spp.* (Collembola: Entomobryidae: Seirinae), sampled in the same collection area, were used as outgroups in all analyses. For each dataset the ML searches, as implemented in GARLI v. 2 (Zwickl 2006), were performed, retaining the best tree out of 20+5 runs from random starting trees, using the best-fit model, allowing the algorithm to estimate parameters. Runs were continued until no improvement in log-likelihood > 0.01 was observed for 50,000 generations. After the best tree was found 1,000 ML non-parametric bootstrap replicates were performed and the result summarized using the SUMTREES v. 3.3.1 tool (Sukumaran & Holder 2010). The BI search was performed with the algorithm implemented in the MPI version of MRBAYES v. 3.2.1 (Huelsenbeck *et al.* 2001; Altekar *et al.* 2004; Ronquist 2004) using the best-fit model, estimating parameters. Fine-tuning of MCMC parameters such as number of runs, chains and temperature were conducted as in Cicconardi et al. (2010).

To build a "species" tree, where "species" should represent any group of individuals that, after some divergence time, are reproductively isolated respectively to individuals outside that group, a coalescence tree model for species or divergent populations within species needs to be identified (Heled & Drummond 2010). Because it has been demonstrated that mtDNA markers provide a better indication of true species diversity (Tang *et al.* 2012), we used the *COII* gene tree to define such coalescence units. A preliminary chronogram was built with BEAST v1.7.2 (Drummond *et al.* 2012) by performing two independent analyses (five independent runs of $5*10^6$ generations, a burn-in of $1*10^6$ generations, sampling every 100^{th} generation), using two different tree priors (coalescence: constant size; speciation: Yule process). The two parallel analyses were carried out to evaluate possible tree shape prior bias. All runs within each of the two analyses were combined to produce a maximum clade credibility trees to define "intra-species" branching pattern under the general mixed yule coalescent (GMYC) model (Pons *et al.* 2006) as implemented in the

R-script GMYC (Powell 2012) (http://www.r-project.org), using default settings for single and multiple threshold analysis. Coalescence units identified by the GMYC analysis were used as traits and both genetic markers were used to define phylogenetic relationships among coalescent units using the species tree method implemented in *BEAST (Heled & Drummond 2010), within BEAST v1.7.2. The multispecies coalescent model estimates a species tree by assuming gene trees are embedded inside it, and following the stochastic coalescent process back in time. The uncorrelated relaxed clock (Drummond *et al.* 2006) was used for both markers, setting a conservative *Lepidocyrtus COII* substitution rate of $2.45*10^{-2}$ *substitutions*/*site*/*myr* (described by Cicconardi et al. (2010)), and estimating the *EF1α* rate relative to mtDNA. The three specimens of *Seira spp.* were used as outgroups and the Yule process tree prior used. Five independent runs of $5*10^{8}$ MCMC generations were executed, sampling every 10,000th generation and samples combined using LogCombiner v. 1.7.2 (Drummond *et al.* 2012). Tracer v. 1.5 (Rambaut & Drummond 2007) was used to evaluate convergence among the independent runs and to define the appropriate burn-in. Posterior distributions and parameters were also evaluated by examining their effective sample size (ESS). TreeAnnotator v. 1.7.2 (Drummond *et al.* 2012) was subsequently used to summarize the obtained trees as a single consensus tree representing the posterior distribution. All phylogenetic analyses were performed independently and nodal support values extracted from the combined log file and presented on the tree topology extracted from the *BEAST analysis. We used the Wilcoxon rank-sum test (Wilcoxon–Mann–Whitney test), as implemented in R project (http://www.r-project.org), to evaluate hypotheses concerning the ages of lineages and their geographic distribution.

Molecular lineages and the biological species concept

We define a molecular lineage as a terminal monophyletic clade composed of the same set of individuals in both *COII* and *EF1α* gene trees. To graphically illustrate molecular lineages we constructed a visualization plot of terminal monophyletic clades for both the *COII* and *EF1α* gene

trees using the software package Circos (http://www.circos.ca/). Molecular lineages within morphospecies were formally evaluated for consistency with the biological species concept by testing for Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD) within sampling sites where two or more molecular lineages occur in sympatry. HWE and LD were calculated using ARLEQUIN v. 3.5.1.2 (Excoffier & Lischer 2010) and statistical significance tested using 100,000 dememorisation steps and 100,000 steps in Markov chain. After Bonferroni adjustments (Benjamini & Hochberg 1995), *p*-values were set to 0.05. We used the Wilcoxon rank-sum test to evaluate, for both markers, differences in the distributions of pairwise genetic distance of intra-sympatric lineage against intra-allopatric. We consider as null hypothesis that the distributions are equal (two-sided). Using coalescence units and morphospecies as unit of analysis, we performed individual-based rarefaction analyses, as implemented R package VEGAN v. 2.0-4 (Oksanen *et al.* 2007), to estimate expected species richness based on the two different measures, in order to evaluate the implications for biodiversity of the two different estimators of species richness.

RESULTS

Morphospecies delineation

Given the very limited number of studies on the taxonomy of Central American springtails, collected specimens (Table 1, Fig. 1) were named according to the closest literature description available. Morphological characters such as labial, thoracic and abdominal chaetotaxy (especially number and position of macrochaete and chaetotaxy of trichobothrial complex of abdomen II-IV), body pigmentation, and shape of the foot complex were taken into consideration for a broad assessment of morphological differences and characterisation of morphospecies groups. Five different morphospecies were identified. Two of them appear closely related to the species *Lepidocyrtus balteatus* Mari Mutt 1983. They differ from each other only by their body pigmentation pattern and are named as *L.* cf. *balteatus* with one and two pigmented bands

249 respectively. A third species, well differentiated from the previous two, may be attributed to the
250 species *L. vexans* Denis, 1933. Compared with the previous species this latter shows higher
251 variability, specifically in the pigmentation pattern and macrochaetae distribution; nonetheless in
252 line with degree of morphological variability commonly considered among Entomobryidae (e.g.
253 Dallai 1967, 1969; Jordana & Baquero 2005; Zhang et al. 2009). These three morphospecies were
254 the most abundant and more geographically widespread. A further two species, one also referable to
255 the species *L. balteatus* Mutt, 1983 - with no pigmentation; and another completely undescribed (*L.*
256 *sp.*), whose habitus differs from the others by a clear and distinct distribution and abundance of
257 thoracic macrochaetotaxy, were less abundant and were recovered in only a few sampling sites
258 (Table 1).
259
260 *DNA sequence variation*
261 Sequences from the mitochondrial gene *COII* were obtained from 96 samples from across the 14
262 sampling sites, yielding 77 unique haplotypes from across 30 morphospecies populations. The
263 number of sequences per morphospecies ranges from 19 to 51 for the three more abundant species
264 and three sequences each for the remaining two species. MtDNA *COII* sequences range in length
265 from 665 to 677 nucleotides (nt). Length variation is explained by an indel of 12 nucleotides
266 corresponding to four contiguous amino acids, and another indel of a single triplet. The Hasegawa-
267 Kishino-Yano model plus Gamma (HKY+Γ) was selected as the most appropriate model for GARLI
268 (ML) and BEAST (BI), while the general time-reversible model plus gamma (GTR+Γ) was selected
269 for MRBAYES (BI). Individuals with identical haplotypes belong to the same morphological species
270 from the same site, with the exception of a single haplotype sampled from two specimens of the
271 same morphospecies in two different sites. Genetic variation within the *COII* locus is remarkably
272 high. Considering *COII* pairwise *p*-distances (π), the distribution has a median of 0.23π, a mean
273 0.21π and the highest value of 0.31π. Furthermore, almost all pairwise comparisons (97%) have
274 values above 0.10π (Fig. S1). The ANOVA analysis reveals that 28.21% of the genetic variation

within the *COII* gene is explained by differences among morphospecies, 27.53% occurs among populations within morphospecies, while the highest proportion, 44.26%, occurs within populations (Table 2, Fig. S2).

Sequences from the nuclear gene *EF1α* were obtained from 80 specimens from 14 sampling sites, representing a total of 29 morphospecies populations. Sixteen samples could not be amplified due to limited DNA template. The number of sequenced individuals per morphospecies ranges from 15 to 45 in the three more common species, and two and three in the remaining two species. The *Lepidocyrtus EF1α* has three exons of differing length (exon I: 396bp; exon II: 219bp; exon III: 236bp), and four intronic sequences with high nucleotide polymorphism and length variation (intron I: 46-102bp; intron II: 84-108bp; intron III: 108-130bp; intron IV: 52-131bp). High variation within the *EF1α* locus is also exemplified by the 3rd intron that is found in only three individuals, representing three different morphospecies from three different sites. The model of molecular evolution that best fits the data for both GARLI (ML) and BEAST (BI) is the equal-frequency Tamura-Nei model plus Gamma (TrNef+Γ) model while the general time-reversible model plus gamma (GTR+Γ) model was selected for MRBAYES (BI). A total of 77 unique alleles are recovered after identical sequences are collapsed. Considering only the coding region of the gene, the locus shows an observed heterozygosity (H_o) of 0.28, which increases to 0.53 when introns are also taken into account. Fourteen alleles were sampled from more than one specimen, of which 13 were sampled from different specimens from the same site, with the remaining allele sampled from two geographically very close populations (SF and CM, Fig. 1). As expected the nuclear locus exhibits less genetic divergence with respect to mtDNA loci, but divergences within and among morphospecies are still substantial for a nuclear locus. Considering the distribution of nucleotide diversity, the median value is 0.11π, with a mean value of 0.08π and a maximum value of 0.14π (Fig. S1). An AMOVA analysis reveals that 70.82% of the observed variation is explained by differences among morphospecies, while 16.04% and 13.14% of variation is explained by differences among populations within morphospecies and within populations respectively (Table 2,

Fig. S2).

Phylogenetic analysis and divergent time estimation

Phylogenetic analyses using both ML (Garli) and BI (MRBAYES) with the same molecular marker resulted in the same topology, with the exception of only a few nodes characterized by very low support (Fig. S3-S4). For both markers and methods the topologies are composed of two very well characterized clades: one composed by the morphospecies *L.* cf. *vexans*, and another with the remaining samples, where morphospecies are highly polyphyletic. The sample compositions of terminal clades are highly consistent between the two gene trees, with very few exceptions. In order to estimate a species tree with *BEAST, coalescence units were first defined by two BEAST analyses for the mtDNA partition, the first with a Yule tree prior, and the second with a coalescent tree prior (see Methods). Both analyses yielded high effective sampling size values (ESS > 400) after sampling $2*10^7$ states. GMYC analyses (threshold probability of 0.90) gave very consistent values for input trees derived from both BEAST analyses, yielding 65 coalescent units for the coalescence tree prior and 62 units for the Yule prior. The more conservative estimate of 62 units was subsequently used for the *BEAST analysis. Five independent runs for the species tree analysis converged and reached stability after $25*10^7$ generations. After applying a burn-in of $5*10^7$ generations log files were merged, yielding a total of $1*10^9$ states giving high ESS values, higher than 5,000 for 85% of the priors.

Using a conservative *Lepidocyrtus COII* evolutionary rate of $2.45*10^{-2}$ *substitutions*/*site*/*myr* (see Cicconardi et al. 2010), BEAST estimated the *EF1α* rate to be $1.38*10^{-3}$ *substitutions*/*site*/*myr*, approximately 18 times slower than *COII*. The use of a conservative (fast) substitution rate means that our estimations of lineage ages are more likely to be underestimates than overestimates (Cicconardi et al. 2010). Adopting this conservative approach to date the species tree (Fig. 2) reveals very deep lineage diversification, with the origin of almost all lineages estimated to precede the Pleistocene (~2 Mya), and basal divergence within the ingroup to have initiated at least 40 Mya

(upper HPD: 51.2; lower HPD: 33.3). The restricted geographic ranges of Collembola lineages, combined with the geologically old branching pattern among lineages, reflect long-term persistence of evolutionary lineages within restricted geographic limits. Very similar findings have been observed in other studies of Collembola, where geologically old and geographically discrete lineages have also been recovered (e.g. Garrick et al. 2007; Cicconardi et al. 2010; Torricelli et al. 2010; Ramirez-Gonzalez et al. 2013).

The number of lineages through time plot (Fig. 2) indicates two inflation points, one between 20 and 15 Mya (T_1) and another more recent, approximately 2.5 Mya (T_2). Although we consider our rate calibration to potentially underestimate the timing of divergence events, these inflation points exhibit good agreement with major paleogeographic and paleoclimatic phenomena. The Panama region went through a general and rapid shallowing (19-12 Mya), from a deep-marine (lower-bathyal) environment to a coastal/fluvial environment, with continual emergence and erosion of volcanic arcs in the early-late Miocene (Coates *et al.* 2003; Wegner *et al.* 2010; Montes *et al.* 2012). At the beginning of the Pleistocene paleogeographic and paleoclimatic events influenced the area, with a global climate change towards cooler temperatures (~2.8 Mya), and the final closure of the Isthmus of Panama (2.7-2.6 Mya) triggering a significant biogeographical event, the great biotic American interchange (see Molnar 2008 for a review). Given the apparent fit of phylogenetic events with paleogeographic and paleoclimatic events, we further evaluated the hypothesis that the geographical persistence of lineages and their diversification may correlate with the regional geological history of land formation. To do this we used available dates for the magmatic evolution of Isthmus of Panama to test whether sequentially formed terrains were randomly colonized by lineages (H_0), or if the temporal emergence of lineages correlates with the temporal sequence of magmatic evolution and the formation of new colonisable territories (H_1). Based on their geographic distributions we assigned the most recent common ancestor (MRCA) of each lineage to one of the three main volcanic arcs designated by Wegner et al. (2010): Adakite suite (Western Panama, < 2 Mya), Cordilleran arc (Midwestern Panama, 22–7 Mya) and Chagres-Bayano arc

(Central Panama, 66–42 Mya). We then tested the hypothesis that the three MRCA distributions are consistent with the progressive volcanic arc formation of the three geographical regions against a null hypothesis of random association. Wilcoxon rank-sum tests reveal that Chagres-Bayano arc MRCAs are significantly older than both Cordilleran arc MRCAs (p-value < 0.0005) and Adakite suite MRCAs (p-value < 0.0004), and that Cordilleran arc MRCAs are older than Adakite suite MRCAs (p-value < 0.01), in perfect agreement with the volcanic arc series (Fig. S5).

Molecular lineages, species boundaries and species richness

An assessment of terminal monophyletic clades, comprised of the same individuals in both the *COII* and *EF1α* trees (Fig. S3-S4), results in a total of 58 molecular lineages (Fig. 3). Across the 14 sampling sites 34 molecular lineages are found in sympatry with at least one other molecular lineage belonging to the same morphospecies. Perhaps not surprisingly, sympatric lineages (s.l.) are found only within the more densely sampled morphospecies (*L.ve*: 22 sympatric lineages; *L.b1b*: 7; *L.b2b*: 5). To evaluate the significance of apparent mitochondrial and nuclear DNA sequence co-segregation patterns among sympatric molecular lineages, we tested for linkage disequilibrium (LD) and Hardy-Weinberg equilibrium (HWE), considering as a null hypothesis that individuals of the same morphospecies within a sampling site belong to a panmictic population. Tests were performed for morphospecies populations in which sympatric molecular lineages are represented by four or more individuals. This resulted in the analysis of six morphospecies populations comprising a total of 19 molecular lineages (*L.b1b-SanFelix* 3 sympatric lineages; *L.b2b-ElValle* 3; *L.ve-Darien* 4; *L.ve-Fortuna* 3; *L.ve-P.I.L.A.* 3; *L.ve-SanFelix* 3). Within all the six morphospecies populations both null hypotheses of linkage equilibrium and HWE were rejected (p-adj < 0.04). As a more rigorous test, we performed pairwise analyses between sympatric molecular lineages sampled for two or more specimens (*L.b1b-SanFelix-2* vs *-3*; *L.ve-Fortuna-2* vs *-3*; *L.ve-P.I.L.A.-1* vs *-2*; *L.ve-SanFelix-1* vs *-2*). All pairwise comparisons resulted in the rejection of both random association of alleles between the two loci and HWE (p-adj < 0.03). The two results are therefore consistent with

379 reproductive isolation among cryptic species. We applied the Wilcoxon rank-sum test to evaluate
380 whether the distribution of pairwise genetic *p*-distances (π) within allopatric molecular lineages
381 differs from that of sympatric molecular lineages. We found no significant difference between the
382 two distributions (*COII p*-value = 0.06; *EF1α p*-value = 0.95) indicating that allopatric molecular
383 lineages exhibit genetic divergences consistent with reproductive isolation from other such lineages
384 (Fig. S6).

385 A rarefaction analysis was performed for both morphospecies and molecular lineages (Fig.
386 4). While the rarefaction analysis suggests it is unlikely that there are additional morphospecies that
387 we have not sampled, the implications for molecular lineages are quite different. The steepness of
388 the molecular lineage curve suggests that regional diversity is much greater than the 58 molecular
389 lineages sampled. A rarefaction analysis of data generated by Cicconardi et al. (2010) yields a
390 similar result for morphospecies and molecular lineages within the North-Western Mediterranean
391 basin (Fig. 4).

392

393

394 ## DISCUSSION

395 *Temporal and geographical lineage diversification*

396 The five morphologically defined species of *Lepidocyrtus* sampled within the region of Panama
397 reveal a striking level of molecular genetic differentiation, with both mitochondrial and nuclear
398 markers exhibiting divergent DNA sequences with restricted geographic ranges. The few identical
399 sequences recovered in our analyses were always sampled from the same morphospecies within the
400 same sampling site, with only two exceptions: a single mtDNA haplotype and a single nuclear allele
401 are shared between two geographically proximate sampling sites. Diversification within the focal
402 group is estimated to have initiated at least 40 Mya, resulting in 62 coalescent units defined from a
403 joint Bayesian analysis of mitochondrial and nuclear DNA sequence data. Thirty molecular lineages
404 were already distinct more than 10 Mya, while the Pleistocene is characterised by 46 molecular

lineages prior to its onset that survived through this 1.8 Mya period. An estimated increase in the rate of diversification approximately 16 Mya coincides with the emergence of the first colonisable lands within the Isthmus, and our analyses indicate that the geographic distribution of molecular lineages within Collembola has been shaped by paleogeographic events within the region. The temporal pattern of establishment of molecular lineages follows a westward progression, with more ancient eastern populations subsequently and progressively colonising newly emergent terrains to the west. This pattern is in general accordance with that observed within freshwater fishes of the region, where two dispersal events occurred in the area, both with a westward direction, before (~7 Mya) and after (~3 Mya) the closure of the Isthmus of Panama (Bermingham & Martin 1998).

The pattern of deeply divergent but geographically localised molecular lineages observed within Panamanian morphospecies of *Lepidocyrtus* is consistent with results obtained by Cicconardi et al. (2010), who examined seven traditionally described morphospecies from the genus *Lepidocyrtus* in the North-Western Mediterranean basin. Their combined analysis of highly concordant mtDNA and nuclear *EF1α* gene sequences revealed an Oligocenic or pre-Oligocenic origin of more than 23 Mya for lineage diversity within and among morphospecies. Cicconardi et al. (2010) identified 35 evolutionary lineages in the North-Western Mediterranean basin that were already distinct more than 5.8 Mya, indicating their survival and persistence through the dramatic environmental perturbations of the Messinian Salinity Crisis. The Pleistocene is characterized by 52 evolutionary lineages prior to its onset that survived through the climatic oscillations during this 1.8 Mya period. Deeply divergent DNA sequence lineages are frequently recovered within morphologically defined species of Collembola, with lineages often exhibiting narrow geographic ranges (e.g. Timmermans et al. 2005; Stevens et al. 2006; Garrick et al. 2007; Cicconardi et al. 2010; Torricelli et al. 2010; Ramirez-Gonzalez et al. 2013). While this diversity has been suggested to be indicative of cryptic species diversity (Cicconardi et al. 2010; Torricelli et al. 2010; Emerson et al. 2011), until now only limited evidence has been available to evaluate this hypothesis (Cicconardi et al. 2010). The fortuitous sampling of divergent molecular lineages in sympatry

within the Panama has allowed us to formally evaluate sequence diversity within morphospecies against the BSC.

Genetic diversity and its biological significance

The class Collembola has a worldwide distribution, but it is only moderately diverse in terms of species number, with only approximately 8,000 species described on the basis of morphological variation. While few in number, morphologically defined species of Collembola frequently exhibit unusually broad geographic ranges for species incapable of flight. While 8,000 species is recognised to be an underestimation of true species diversity, a projected estimate of approximately 50,000 species after taking into account the probable under sampling of locally endemic species (Hopkin 1997) remains a small number of species when compared to estimates of up to 5,000,000 species within the class Insecta (Gaston, 1991). Recent molecular studies raise the possibility that the under-sampling hypothesis is not the only plausible explanation to interpret low species richness within the Collembola, and Emerson et al. (2011) presented evidence supporting a hypothesis of taxonomically and geographically pervasive cryptic species diversity within morphologically defined species. While Emerson et al. (2011) were able to document the widespread occurrence of divergent genetic lineages within morphologically defined species of Collembola, the significance of this diversity in terms of biological species can only be assumed. Cicconardi et al. (2010) were able to directly assess two molecular lineages within one of their seven morphospecies of *Lepidocyrtus* against the BSC, finding support for their status as biological species, but the remainder of the lineages they described could not be assessed as they were only sampled in allopatry. Our sampling of five morphologically defined species of *Lepidocyrtus* in the Isthmus of Panama recovered 58 well-defined molecular lineages (Fig. 2-3) in only 14 sampling sites. Among these 58 lineages, 34 were sampled in sympatry, among which 19 were sufficiently sampled to allow us to formally test expectations from the BSC concerning LD and HWE. Our analyses revealed unambiguous co-segregation between *COII* and *EF1α* markers in sympatry and the

457 disruption of HWE. Four sympatric pairs of molecular lineages were sufficiently sampled to
458 perform pairwise analyses that were in each case consistent with reproductive isolation among
459 molecular lineages. The pairwise genetic distances among molecular lineages in sympatry were not
460 significantly different from pairwise distances among the remaining allopatric lineages, suggesting
461 all 58 molecular lineages are likely to exhibit reproductive isolation.
462
463 *Morphospecies diversity vs biological species diversity*
464 Molecular sampling of *Lepidocyrtus* within only 39 sites across Panama and the North-Western
465 Mediterranean yields a species richness estimate of more than half that of a morphologically based
466 global species richness estimate for the genus. The genus *Lepidocyrtus* is globally distributed with
467 266 described morphospeices (www.collembola.org). We have ascertained that molecular lineages
468 within morphospecies conform to biological species, and we have characterised 58 such lineages
469 across 14 sites in Panama. A previous investigation of *Lepidocyrtus* in the North-Western
470 Mediterranean (Cicconardi et al. 2010) recovered 87 molecular lineages across 25 sampling sites,
471 yielding an estimated 145 biological species. Rarefaction analysis (Fig 4) reveals that
472 morphospecies diversity has been exhaustively sampled while, in contrast, it is clear that biological
473 species diversity, as estimated by molecular lineages, has only been partially sampled within each
474 region. These results have significant implications for global biological species diversity, both
475 within the genus *Lepidocyrtus* and within the many other genera of Collembola with molecular
476 signals of cryptic species diversity (Emerson et al. 2011).
477

478 ## CONCLUSIONS

479 Collembola are one of the more ancient groups of arthropods to have colonized the terrestrial
480 environment, with broad environmental tolerances among species appearing to have facilitated their
481 ubiquitous occurrence across a wide range of ecosystems, from dry arctic environments through
482 alpine tundra and deserts to humid tropical rain forests. Their low species diversity and typically

483 broad geographic ranges have been an enigma that has recently been challenged, with the
484 suggestion that morphology may be an inappropriate tool for the delineation of species (Emerson et
485 al. 2011). This is supported by recent molecular studies reporting high levels of DNA sequence
486 divergence within morphologically defined species, but without a demonstrated relationship
487 between this variation and biological species boundaries. We have assessed 58 intra-morphospecies
488 molecular lineages, sampled from within five morphologically defined species of Collembola
489 sampled in the Isthmus of Panama, against the biological species concept, finding that molecular
490 lineages are consistent with biological species. Biological species richness within the class
491 Collembola is clearly underestimated by morphological approaches, with local species richness
492 estimates from molecular lineage sampling indicating the need for intensive sampling to estimate
493 species richness, even at a small local scale. Biological species richness within the class
494 Collembola, and the geographic structure of this diversity, are substantially misrepresented
495 components of terrestrial animal biodiversity. We speculate that global species richness could be at
496 least an order of magnitude greater than a previous estimate of 50,000 species (Hopkin 1997).
497
498

499 ## ACKNOWLEDGMENTS

500 F.C. gratefully acknowledges Prof. Pietro Paolo Fanciulli from the University of Siena for his help
501 in the morphological identification of specimens, the Smithsonian Institution for Tropical Research
502 and thanks the National Environmental Authority of Panama's (ANAM) for granting collecting and
503 export permits. He also wishes to thank Dr. Oris Sanjur for his extensive help with various aspects
504 of the sampling campaign, Dr. Donald M. Windsor, Vanessa Sanchez and all the members of the
505 Windsor Lab at STRI for their kind support and outstanding assistance throughout all phases of the
506 stay in Panamà. F.C. wants also to thank Prof. Alessandro Nardone from University of Tuscia, for
507 letting use the molecular biology resource facility. This research was financial supported by the
508 Smithsonian Institution Short-fellowship Program Partial, the Italian Ministry of Education,

509 Universities and Research and by a Research Fellowship awarded to Brent Emerson from The
510 Leverhulme Trust.

511

512

513

514 ## REFERENCES

515

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, **20**, 407–415.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B Methodological*, **57**, 289–300.

Bermingham E, Martin AP (1998) Comparative mtDNA phylogeography of neotropical freshwater fishes: testing shared history to infer the evolutionary landscape of lower Central America. *Molecular Ecology*, **7**, 499–517.

Briggs JC (2007) The Genesis of Central America: Biology versus Geophysics. *Global Ecology and Biogeography*, **4**, 169–172.

Christiansen K (1971) Notes on Miocene Amber Collembola from Chiapas. *University of California Publications in Entomology*, **63**, 45–48.

Cicconardi F, Nardi F, Emerson BC, Frati F, Fanciulli PP (2010) Deep phylogeographic divisions and long-term persistence of forest invertebrates (Hexapoda: Collembola) in the North-Western Mediterranean basin. *Molecular ecology*, **19**, 386–400.

Coates AG, Aubry M, Berggren WA *et al.* (2003) Early Neogene history of the Central American arc from Bocas del Toro , western Panama.

Dallai R (1967) Ricerche sui collemboli - II. *Archivio Botanico e Biogeografico italiano*, **12**, 424–449.

Dallai R (1969) Ricerche sui Collemboli. VI. Le isole di Capraia e di Pianosa. *Redia*, **51**, 277–304.

Decaëns T (2010) Macroecological patterns in soil communities. *Global Ecology and Biogeography*, **19**, 287–302.

Deharveng L (2004) Recent advances in Collembola systematics. *Pedobiologia*, **48**, 415–433.

Denis JR (1933) Contributo alla conoscenza del Microgenton di Costa Rica. III. Collemboles de Costa Rica avec une contribution au species de l ordre. Deuxiene note. *Boll. Lab. Zool. Portici*, **27**, 222–322.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. (D Penny, Ed,). *PLoS biology*, **4**, e88.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*, **29**, 1969–73.

Emerson BC, Cicconardi F, Fanciulli PP, Shaw PJA (2011) Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, **366**, 2391–2402.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, **10**, 564–567.

Garrick RC, Sands CJ, Rowell DM, Hillis DM, Sunnucks P (2007) Catchments catch all: long-term population history of a giant springtail from the southeast Australian highlands--a multigene approach. *Molecular Ecology*, **16**, 1865–1882.

Gaston KJ (1991) The Magnitude of Global Insect Species Richness. *Conservation Biology*, **5**, 283–296.

Gisin H (1964a) Collemboles d'Europe. VI. *Revue Suisse de Zoologie*, **71**, 383–400.

Gisin H (1964b) Collemboles d'Europe. VII. *Revue Suisse de Zoologie*, **71**, 649–678.

Gisin H (1965) Nouvelle notes taxonomiques sur les Lepidocyrtus. *Revue d'Écologie et de Biologie du Sol*, **2**, 519–524.

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, **27**, 221–224.

Greenslade P, Whalley P (1986) The systematic position of Rhyniella praecursor Hirst and Maulik (Collembola). The earliest known hexapod. In: *Second International Seminar on Apterygota, Siena, Italy* (ed Dallai R), pp. 319–323. University of Siena, Siena, Italy.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Molecular biology and evolution*, **27**, 570–80.

Hirst S, Maulik S (1926) On some Arthropod Remains from the Rhynie Chert (Old Red Sandstone). *Geological Magazine*, **63**, 69–71.

Hopkin SP (1997) Biology of the springtails (Insecta: Collembola). *Journal of Fluid Mechanics*, **491**, 285–300.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Joppa LN, Roberts DL, Myers N, Pimm SL (2011) Biodiversity hotspots house most undiscovered plant species. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 13171–6.

Jordana R, Baquero E (2005) A proposal of characters for taxonomic identification of Entomobrya species (Collembola , Entomobryomorpha), with description of a new species. *Abh. Ber. Naturkundemus. Görlitz*, **76**, 117–134.

Lavelle P, Decaëns T, Aubert M *et al.* (2006) Soil invertebrates and ecosystem services. *European Journal of Soil Biology*, **42**, S3–S15.

Mari-Mutt JA (1983a) Collembola in amber from the Dominican Republic. *Proceedings of the Entomological Society of Washington*, **85**, 575–587.

Mari-Mutt JA (1983b) Two new species of Lepidocyrtus from Paramo de Mucubaji, Merida, Venezuela (Collembola: Entomobryidae). *Carribean Journal of Science*, **19**, 53–60.

Mari-Mutt JA (1986) Puerto Rican Lepidocyrtus and species of Pseudosinella (COLLEMBOLA: ENTOMOBRYIDAE). *Carribean Journal of Science*, **22**, 1–48.

Mari-Mutt JA, Bellinger PF (1990) *A catalog of the Neotropical Collembola, including Nearctic areas of Mexico*. Sandhill Crane Press, Gainesville, Fla. (USA).

Mari-Mutt JA, Bellinger PF (1996) Supplement to the Catalog of the Neotropical Collembola—August 1989 to April 1996. *Caribbean Journal of Science*, **32**, 166–175.

Mayr E (1942) *Systematics and the origin of species from the viewpoint of a zoologist*. Columbia University Press.

Molnar P (2008) Closing of the Central American Seaway and the Ice Age: A critical review. *Paleoceanography*, **23**, 1–15.

Montes C, Cardona a., McFadden R *et al.* (2012) Evidence for middle Eocene and younger land emergence in central Panama: Implications for Isthmus closure. *Geological Society of America Bulletin*, **124**, 780–799.

Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca GA, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press.

Ødegaard F (2000) How many species of arthropods? Erwin's estimate revised. *Biol J Linnean Soc*, **71**, 583–597.

Oksanen J, Blanchet FG, Kindt R *et al.* (2007) vegan: Community Ecology Package. *R package version*, **1**, R package version 1.17–9.

Ozanne CMP, Anhuf D, Boulter SL *et al.* (2003) Biodiversity meets the atmosphere: a global view of forest canopies. *Science*, **301**, 183–186.

Pons J, Barraclough TG, Gomez-Zurita J *et al.* (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**, 595–609.

Porco D, Bedos A, Greenslade P *et al.* (2012) Challenging species delimitation in Collembola: cryptic diversity among common springtails unveiled by DNA barcoding. *Invertebrate Systematics*, **26**, 470.

Porco D, Potapov M, Bedos A *et al.* (2012) Cryptic diversity in the ubiquist species Parisotoma notabilis (Collembola, Isotomidae): a long-used chimeric species? (D Steinke, Ed,). *PloS one*, **7**, e46056.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–1256.

Powell JR (2012) Accounting for uncertainty in species delineation during the analysis of environmental DNA sequence data. *Methods in Ecology and Evolution*, **3**, 1–11.

Rambaut A, Drummond AJ (2007) Tracer v14, Available from http://beast.bio.ed.ac.uk/Tracer.

Ramirez-Gonzalez R, Yu DW, Bruce C *et al.* (2013) PyroClean: Denoising Pyrosequences from Protein-Coding Amplicons for the Recovery of Interspecific and Intraspecific Genetic Variation. (A Torkamani, Ed,). *PloS one*, **8**, e57615.

Rapoport E (1971) The geographical distribution of neotropical and antartic Collembola. *Pacific Insects Monograph*, **25**, 99–118.

Riek EF (1976) New Upper Permian insects from Natal, South Africa. *Annals of the Natal Museum*, **22**, 755–789.

Rodgers DJ, Kitching RL (2010) Rainforest Collembola (Hexapoda: Collembola) and the insularity of epiphyte microhabitats. *Insect Conservation and Diversity*, **4**, 99–106.

Ronquist F (2004) Bayesian inference of character evolution. *Trends in Ecology & Evolution*, **19**, 475–481.

Scheffers BR, Joppa LN, Pimm SL, Laurance WF (2012) What we know and don't know about Earth's missing biodiversity. *Trends in ecology & evolution*, **27**, 501–10.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, **68**, 978–989.

Stevens MI, Greenslade P, Hogg ID, Sunnucks P (2006) Southern hemisphere springtails: Could any have survived glaciation of Antarctica? *Molecular Biology and Evolution*, **23**, 874–882.

Sukumaran J, Holder MT (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.

Swofford DL (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.* Sinauer, Sunderland, MA.

Tang CQ, Leasi F, Obertegger U *et al.* (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 16208–12.

Timmermans MJTN, Ellers J, Mariën J *et al.* (2005) Genetic structure in Orchesella cincta (Collembola): strong subdivision of European populations inferred from mtDNA and AFLP markers. *Molecular Ecology*, **14**, 2017–2024.

Torricelli G, Carapelli A, Convey P *et al.* (2010) High divergence across the whole mitochondrial genome in the "pan-Antarctic" springtail Friesea grisea: evidence for cryptic species? *Gene*, **449**, 30–40.

Wegner W, Worner G, Harmon RS, Jicha BR (2010) Magmatic history and evolution of the Central American Land Bridge in Panama since Cretaceous times. *Geological Society of America Bulletin*, **123**, 703–724.

Whalley P, Jarzembowski EA (1981) A new assessment of Rhyniella, the earliest known insect, from the Devonian of Rhynie, Scotland. *Nature*, **291**, 317–317.

Yosii R (1974) Fossil Collembola contained in the Mizunami amber (Insecta: Collembola). *Bulletin of the Mizunami Fossil Museum*, **1**, 409–411.

Zhang F, Deharveng L, Greenslade P, Chen J-X (2009) Revision of Acanthocyrtus (Collembola: Entomobryidae), with description of a new genus from eastern Asia. *Zoological Journal of the Linnean Society*, **157**, 495–514.

Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin.

F.C. and B.C.E designed the study and wrote the manuscript, F.C. performed sampling, sequencing and analysis.

Data Accessibility:

- DNA sequences: Genbank accessions XXXXXXXXXX

- Final DNA sequence assembly uploaded as online supplemental material

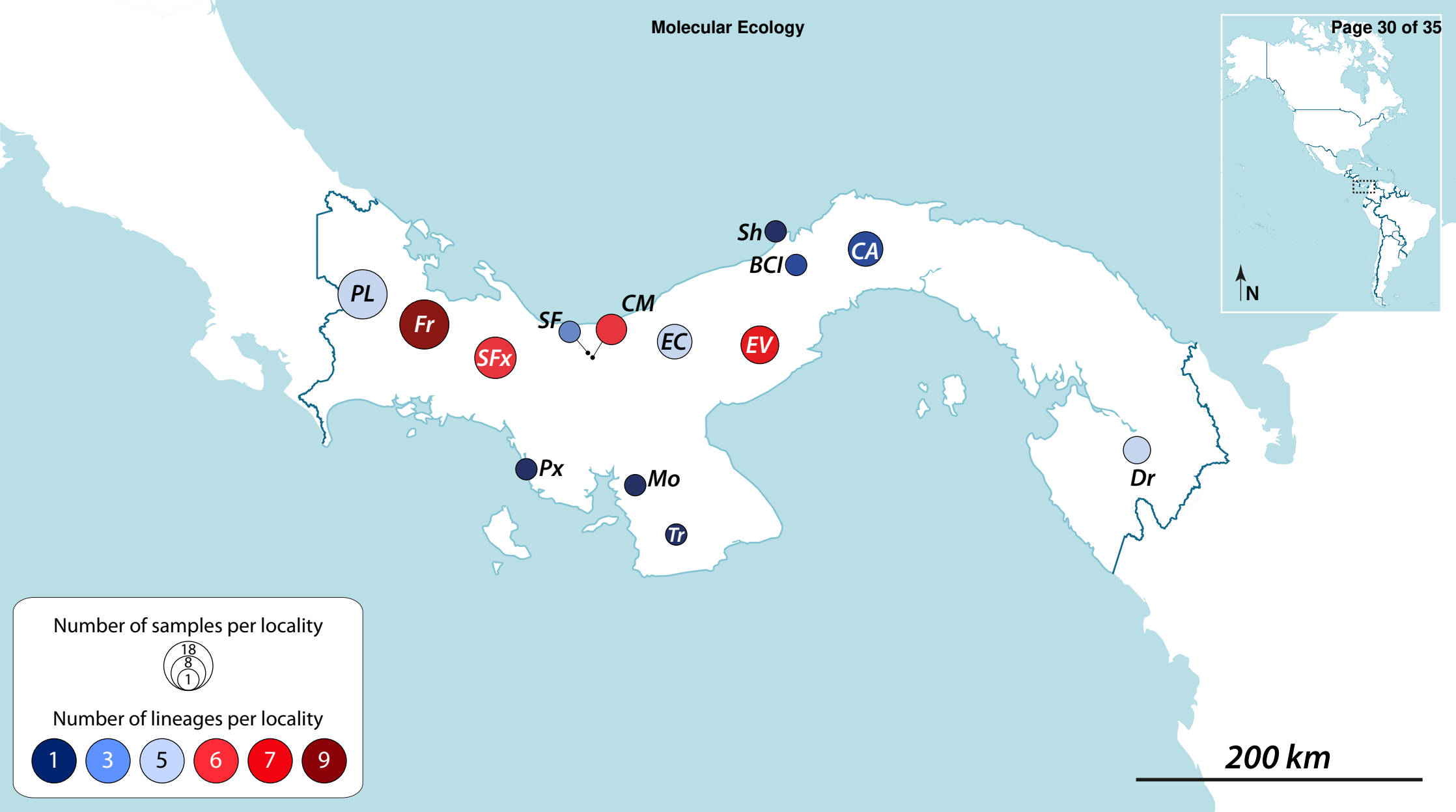
Figure legends

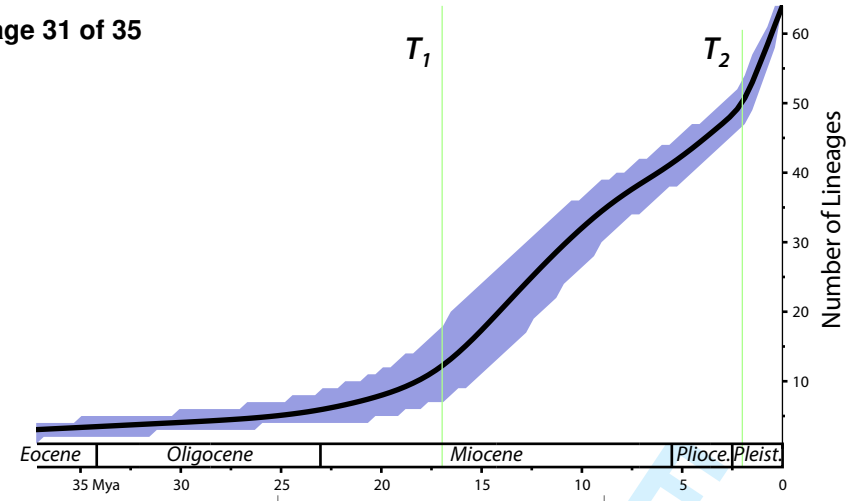
FIGURE 1. Map of sampling sites within the Isthmus of Panama for morphospecies within the genus *Lepidocyrtus*, as coded as in Table 1. Circle size represents the number of sampled specimens, colours indicate the number of mtDNA lineages within each site.

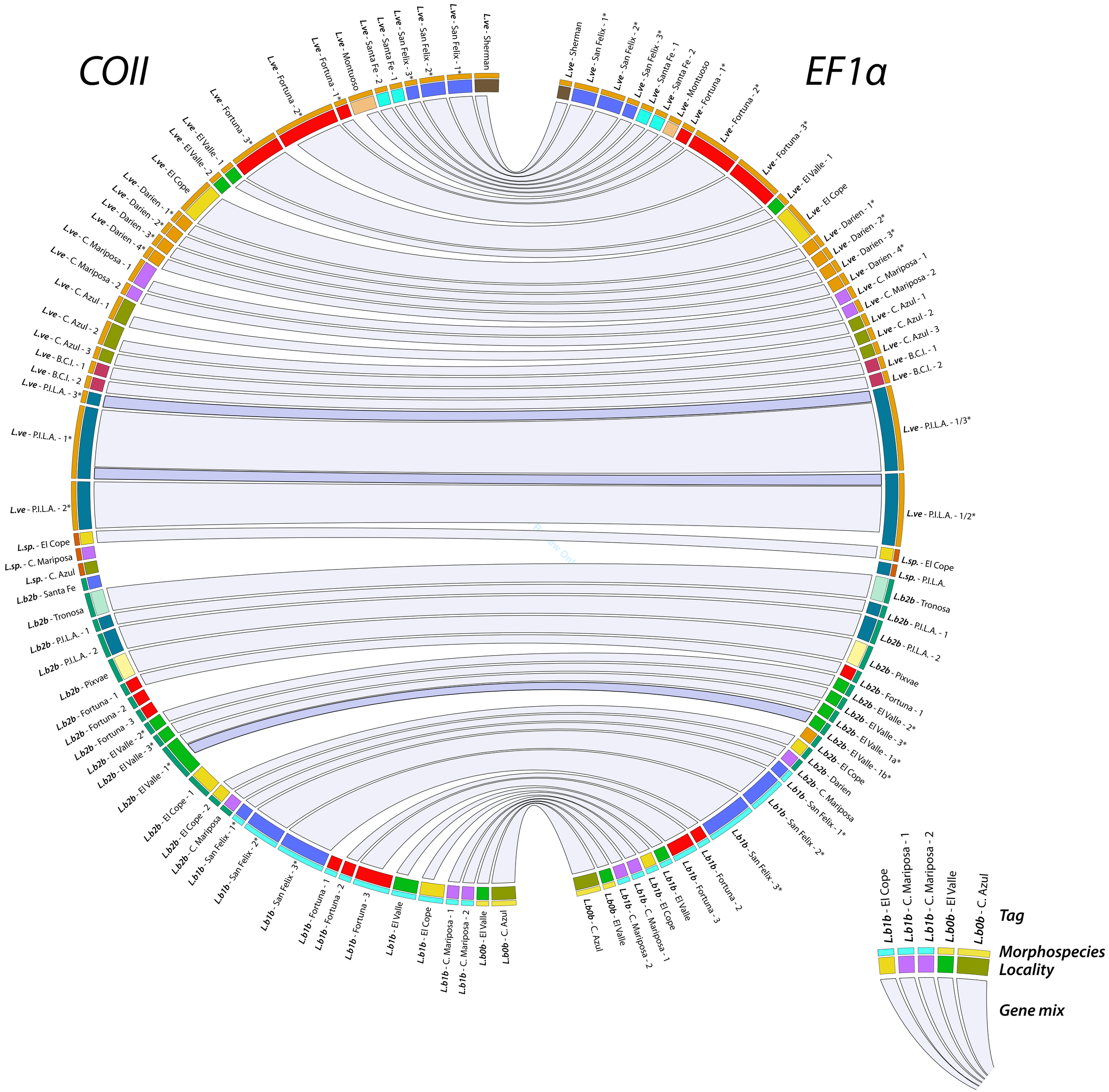
FIGURE 2. Species tree derived from the Bayesian analysis of *COII* and *EF1α* DNA sequence data from morphospecies within the genus *Lepidocyrtus* using *BEAST. Posterior probabilities are indicated above nodes only if higher than 0.70. Branch colors represent morphospecies, while colored boxes represent molecular lineages within each sampling site. The plot in the insert represents lineages through time (LTT). The black line indicates the mean, while the blue area represents the 95% HPD.

FIGURE 3. Segregation of molecular loci within morphospecies of the genus *Lepidocyrtus*. The correspondence between mitochondrial (left) and nuclear (right) molecular lineages is plotted with Circos. Differently coloured outer rectangles represent different morphospecies, while differently coloured inner rectangles represent different sampling sites. Rectangle width reflects sample size. Internally, light-gray ribbons represent co-segregation of markers, while darker ribbons represent mixed segregation. Tags show species name and sampling site, with numbers differentiating sympatric lineage of the same morphospecies.

FIGURE 4. Rarefaction curves from sampling at the regional scale within the Isthmus of Panama and the northwestern Mediterranean basin for species within the genus *Lepidocyrtus* defined upon the basis of morphology or molecular lineages.


Sh
BCI
CA
PL
Fr
SF
CM
EC
EV
SFx
Px
Mo
Tr
Dr
N
Number of samples per locality
18
8
1
Number of lineages per locality
1
3
5
6
7
9
200 km




COII
EF1α
L.ve - Sherman
L.ve - San Felix - 1*
L.ve - San Felix - 2*
L.ve - San Felix - 3*
L.ve - Santa Fe - 1
L.ve - Santa Fe - 2
L.ve - Montuoso
L.ve - Fortuna - 1*
L.ve - Fortuna - 2*
L.ve - Fortuna - 3*
L.ve - El Valle - 1
L.ve - El Valle - 2
L.ve - El Cope
L.ve - Darien - 1*
L.ve - Darien - 2*
L.ve - Darien - 3*
L.ve - Darien - 4*
L.ve - C. Mariposa - 1
L.ve - C. Mariposa - 2
L.ve - C. Azul - 1
L.ve - C. Azul - 2
L.ve - C. Azul - 3
L.ve - B.C.I. - 1
L.ve - B.C.I. - 2
L.ve - P.I.L.A. - 3*
L.ve - P.I.L.A. - 1*
L.ve - P.I.L.A. - 2*
L.sp. - El Cope
L.sp. - C. Mariposa
L.sp. - C. Azul
L.b2b - Santa Fe
L.b2b - Tronosa
L.b2b - P.I.L.A. - 1
L.b2b - P.I.L.A. - 2
L.b2b - Pixvae
L.b2b - Fortuna - 1
L.b2b - Fortuna - 2
L.b2b - Fortuna - 3
L.b2b - El Valle - 2*
L.b2b - El Valle - 3*
L.b2b - El Valle - 1*
L.b2b - El Cope - 1
L.b2b - El Cope - 2
L.b2b - C. Mariposa
L.b1b - San Felix - 1*
L.b1b - San Felix - 2*
L.b1b - San Felix - 3*
L.b1b - Fortuna - 1
L.b1b - Fortuna - 2
L.b1b - Fortuna - 3
L.b1b - El Valle
L.b1b - El Cope
L.b1b - C. Mariposa - 1
L.b1b - C. Mariposa - 2
L.b0b - El Valle
L.b0b - C. Azul
L.b0b - C. Azul
L.b0b - El Valle
L.b1b - C. Mariposa - 2
L.b1b - C. Mariposa - 1
L.b1b - El Cope
L.b1b - El Valle
L.b1b - Fortuna - 3
L.b1b - Fortuna - 2
L.b1b - San Felix - 3*
L.b1b - San Felix - 2*
L.b1b - San Felix - 1*
L.b2b - C. Mariposa
L.b2b - El Cope - 2
L.b2b - El Cope - 1
L.b2b - El Valle - 1*
L.b2b - El Valle - 3*
L.b2b - El Valle - 2*
L.b2b - Fortuna - 3
L.b2b - Fortuna - 2
L.b2b - Fortuna - 1
L.b2b - Pixvae
L.b2b - P.I.L.A. - 2
L.b2b - P.I.L.A. - 1
L.b2b - Tronosa
L.b2b - Santa Fe
L.sp. - C. Azul
L.sp. - C. Mariposa
L.sp. - El Cope
L.ve - P.I.L.A. - 2*
L.ve - P.I.L.A. - 1*
L.ve - P.I.L.A. - 3*
L.ve - B.C.I. - 2
L.ve - B.C.I. - 1
L.ve - C. Azul - 3
L.ve - C. Azul - 2
L.ve - C. Azul - 1
L.ve - C. Mariposa - 2
L.ve - C. Mariposa - 1
L.ve - Darien - 4*
L.ve - Darien - 3*
L.ve - Darien - 2*
L.ve - Darien - 1*
L.ve - El Cope
L.ve - El Valle - 2
L.ve - El Valle - 1
L.ve - Fortuna - 3*
L.ve - Fortuna - 2*
L.ve - Fortuna - 1*
L.ve - Montuoso
L.ve - Santa Fe - 2
L.ve - Santa Fe - 1
L.ve - San Felix - 3*
L.ve - San Felix - 2*
L.ve - San Felix - 1*
L.ve - Sherman
L.ve - P.I.L.A. - 1/3*
L.ve - P.I.L.A. - 1/2*
L.sp. - El Cope
L.sp. - P.I.L.A.
L.b2b - Tronosa
L.b2b - P.I.L.A. - 1
L.b2b - P.I.L.A. - 2
L.b2b - Pixvae
L.b2b - Fortuna - 1
L.b2b - El Valle - 2*
L.b2b - El Valle - 3*
L.b2b - El Valle - 1a*
L.b2b - El Valle - 1b*
L.b2b - El Cope
L.b2b - Darien
L.b2b - C. Mariposa
L.b1b - San Felix - 1*
L.b1b - San Felix - 2*
L.b1b - San Felix - 3*
L.b1b - Fortuna - 2
L.b1b - Fortuna - 3
L.b1b - El Valle
L.b1b - El Cope
L.b1b - C. Mariposa - 1
L.b1b - C. Mariposa - 2
L.b0b - El Valle
L.b0b - C. Azul
L.b1b - El Cope
L.b1b - C. Mariposa - 1
L.b1b - C. Mariposa - 2
L.b0b - El Valle
L.b0b - C. Azul
Tag
Morphospecies
Locality
Gene mix

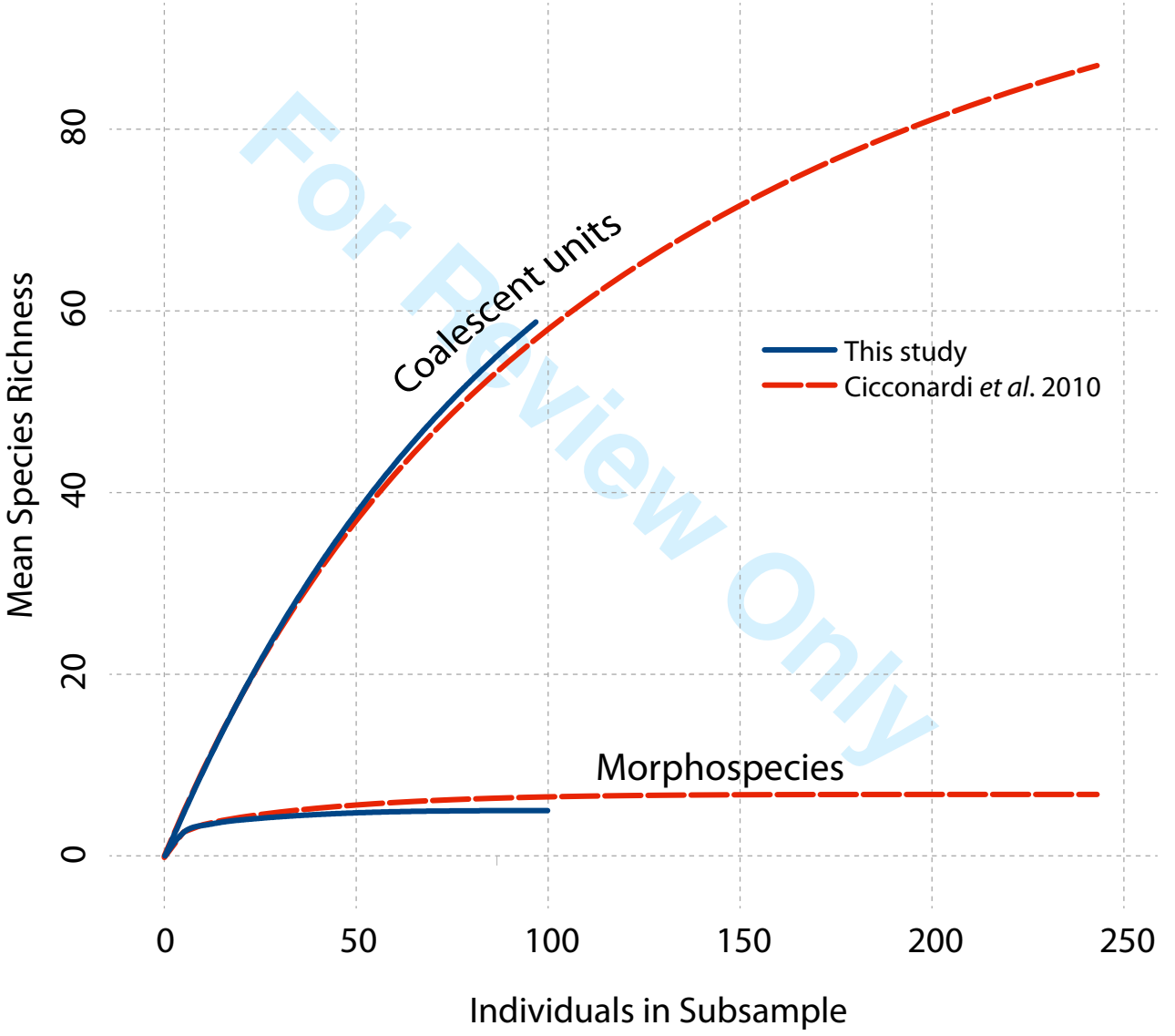


Table 1 Sampling sites

Table 1. Sampling sites within the Isthmus of Panama for morphospecies within the genus *Lepidocyrtus*

Code	Location	Province	Latitude	Longitude	Altitude a.s.l.	Morphotype (# of specimen)	# Lineages	Vegetation
BCI	*Barro Colorado Island*	Panamà	9° 9'5.33"N	79°51'32.47"W	130m	*L.ve* (2)	2	1-2
CA	*Cerro Azul*	Panamà	9°14'16"N	79°24'38"W	890m	*L.b0b* (2); *L.sp.* (1); *L.ve* (5)	4	2-1
CM	*Cerro Mariposa*	Veraguas	8°29'59.80"N	81° 6'53.20"W	1200m	*L.b1b* (2); *L.b2b* (1); *L.sp.* (1); *L.ve* (3	6	3-1
Dr	*Darién*	Darién	8° 0'1.86"N	77°42'23.72"W	480m	*L.b2b* (1); *L.ve* (4)	5	1-1
EC	*El Copé*	Coclé	8°39'15"N	80°37'30"W	1180m	L.b1b (2); L.b2b (3); *L.sp.* (1); *L.ve* (3)	5	2-1
EV	*El Valle*	Coclé	8°37'27"N	80° 7'6"W	1000	L.b0b (1); L.b1b (2); L.b2b (5); *L.ve* (2	7	2-1
Sh	*Fort Sherman*	Colón	9°21'11.98"N	79°57'45.17"W	40m	*L.ve* (2)	1	1-1
Fo	*Fortuna station*	Bocas del Toro	8°47'12.50"N	82°13'17.80"W	1230m	*L.b1b* (5); *L.b2b* (3); *L.ve* (10)	9	2-1
Tr	*La Tronosa Forest Res.*	Los Santos	7°23'52"N	80°38'15"W	560m	*L.b2b* (2)	1	1-2
Mo	*Montuoso Forest Res.*	Herrera	7°43'36"N	80°50'54"W	930m	*L.ve* (2)	1	1-2
PL	*P.I.L.A.*	Chiriqui	8°53'58.60"N	82°37'11.60"W	2430m	L.b2b (3); *L.sp.* (1); *L.ve* (12)	5	4-1
Px	*Pixvae*	Veraguas	7°52'7"N	81°31'55"W	430m	*L.b2b* (2)	1	1-2
Sfx	*San Felix*	Ngöbe-Buglé	8°29'57.80"N	81°46'21.10"W	1690m	*L.b1b* (8); *L.ve* (5)	6	3-1
SF	*Santa Fe*	Veraguas	8°33'36"N	81°10'29"W	570m	*L.b2b* (1); *L.ve* (2)	3	2-1

Vegetation Legend

Code	Eng
1-1	Low-land tropical evergreen broadleaf ombrophilous rainforest
1-2	Low-land tropical semideciduous rainforest
2-1	Sub-mountain tropical evergreen broadleaf ombrophilous rainforest
3-1	Mountain tropical evergreen broadleaf ombrophilous rainforest
4-1	High mountain tropical evergreen broadleaf ombrophilous rainforest

Table 2. Amova analyses of mtDNA COII and nuclear EF1α sequence data from morphospecies within the genus *Lepidocyrtus*.

AMOVA

Source of variation COII	Variance components	Percentage of variation
Among morphospecies	23.83	28 %
Among population within morphospecies	23.25	28 %
Within morphospecies populations	37.38	44 %
total	84.46	100 %
Among T_1 lineages	26.80	32 %
Among coalescent units within T_1 lineages	53.56	64 %
Within coalescent units	3.78	4 %
total	84.14	100 %
Among T_2 lineages	14.37	44 %
Among coalescent units within T_2 lineages	17.46	54 %
Within coalescent units	0.73	2 %
total	32.55	100 %

AMOVA

Source of variation EF1α	Variance components	Percentage of variation
Among morphospecies	31.14	71 %
Among population within morphospecies	7.05	16 %
Within morphospecies populations	5.78	13 %
total	43.97	100 %
Among T_1 lineages	29.23	70 %
Among coalescent units within T_1 lineages	12.08	29 %
Within coalescent units	0.73	2 %
total	42.04	100 %
Among T_2 lineages	14.37	44 %
Among coalescent units within T_2 lineages	17.46	54 %
Within coalescent units	0.73	2 %
total	32.55	100 %