Updating taxonomic biogeography in the light of new methods – examples from Collembola

Peter Shaw1,*, Christiana Faria2 and Brent Emerson2,3

1 University of Roehampton, Centre for Research in Ecology, Whitelands College, Holybourne Ave., London SW15 4JD, UK
2 School of Biological Sciences, Norwich Research Park, Norwich NR4 7TJ, UK
3 Present Address: Island Ecology and evolution group, IPNA-CSIC, C/Astrofisico Francisco Sánchez 3, 38206 La Laguna, Tenerife, Canary Islands, Spain
* Corresponding author, e-mail: p.shaw@roehampton.ac.uk

Received 10 January 2013 | Accepted 21 October 2013
Published online at www.soil-organisms.de 1 December 2013 | Printed version 15 December 2013

Abstract

The aim of this article is to review the consequences of advances in molecular genetics for researchers of Collembola biogeography. Gene sequence data have generally confirmed existing high-level taxonomy, but complicated species-level taxonomy by uncovering extensive cryptic diversity. Several commonly recorded European Collembola ‘species’ have proved to be complexes of closely related taxa, reducing the value of many older (pre-1990) records to near-zero. It seems likely that many more cryptic species of Collembola remain to be uncovered, even in well studied areas in Europe. The inevitable proliferation of genetically defined ‘species’ will be awkward to integrate into existing databases, which are based on morphospecies. Eventually the results will transform our understanding of Collembola biodiversity, since molecular data contain greatly more information, notably estimates of divergence times. In ancient and relatively pristine ecosystems (e.g. Antarctica, oceanic islands) genetic data can be used to show both extreme isolation of endemics (pre-dating ice ages) and the arrival of European/North American invasives. Collembola from the genus Lepidocyrtus in several very different areas show patterns of intra-specific divergence dated in the millions of years, a pattern likely to be repeated in other genera.

An exciting approach to speeding up these genetic studies comes from the advent of high throughput pyrosequencing technology, which already allows mass identification of mixed life forms from clustering then sequencing DNA amplicons (‘metabarcoding’), raising the possibility of automating the identification of bulked soil faunal extracts. A trial of this using soil extracts from Tenerife identified an apparent endemic Friesea species, along with mainland invasive species. These next generation techniques rely on Molecular Taxonomic units (MOTUs), identified purely by sequence analysis (usually of mtCOI), so vigilance will always be needed against the spurious identification of sequence artefacts such as nuclear copies of mitochondrial genes (NUMTS). A new risk of relying simply on high throughput sequence analysis is the possibility of inventing non-existent ‘species’ from irrelevant DNA.

Keywords phylobetageography | biodiversity | barcoding | mtCOI | endemism

1. Scope of review

One of the core questions needed for the management and conservation of biodiversity is to know what we have, and where it is (Vane-Wright et al. 1991, Crozier 1992). The recent increase in the availability of genetic data has been a great asset, for example in the unification of sexually dimorphic species and confirming species boundaries (Janzen et al. 2005, Bik et al. 2012). There is general agreement that genetic data are the pre-eminent tool to resolve the existence of cryptic species in many groups (Hogg & Hebert 2004, Chang et al. 2009, Porco
et al. 2013), but definitions of species boundaries remains contentious. Defining boundaries based on genetic distance for example involves choosing an essentially arbitrary point along a continuum to make a binary division, arguments about its location will inevitably persist. There is a move to a universal standard ‘barcode of life’ based on the mitochondrial Cytochrome c Oxidase 1 (Ratnasingham & Hebert 2007). There are however serious criticisms of the uncritical application of COI barcodes in taxonomy (Ebach & Holdrege 2005, Rubinoff et al. 2006), including the problem of ‘NUMTS’ (Nuclear Mitochondrial Sequences) (Song et al. 2008). This occurs when mitochondrial DNA sequences are copied into a nuclear chromosome, where they are copied down the generations but never transcribed, hence can accumulate mutations rapidly, unconstrained by selective pressure, and appear as the sequence of an unknown species. In the neotropical butterfly Astraptes fulgerator Walch, 1775, initial barcoding by Hebert et al. (2004) suggested that this clade comprises 10 cryptic species (Smith et al. 2006), but Brower (2006) reanalysed the COI data to show that it suggested no more than 7 clades, one of which probably an artefact caused by a NUMTS.

Typically the use of genetic techniques to estimate the extent of evolutionary separation between taxa increases the number of species, or finds deep splits in taxa formerly considered as one species (e.g. Frati et al. 2000, Eggart et al. 2002, Fernando et al. 2003, Blaxter 2003, Timmermans et al. 2005, Janzen et al. 2005, Cicconardi et al. 2009, Cicconardi et al. 2013).

This review will contrast biodiversity of the Collembola patterns based on our current knowledge of morphospecies, and what is starting to be found using DNA sequence data, exemplified using results from three contrasting biomes.

2. Biodiversity within the Collembola

Collembola is an ancient group whose main families evolved before the split-up of Pangea around 200 million years before present (Hopkin 1997). Despite this, Collembola appear not to be very speciose, with recent estimates around 7,900 species globally (Janssens 2013) (up from Deharveng’s 2004 figure of c. 7000), compared to more than 300,000 Coleoptera. It is likely that the true species richness of Collembola is much higher: Cicconardi et al. (2013) suggest the true figure for Collembolan species richness is around 500,000, due to the widespread occurrence of cryptic species.

As with other life forms, long-isolated populations of Collembola can give rise to local endemic morphospecies; e.g. on remote islands (Deharveng 1980, Thibaud 1991, Thibaud & Najt 1993, Fjellberg 1995, Soto-Adames 2002), mountain tops (Dallas 1983, Deharveng & Bedos 1991, Cassagnau 1991) and ancient cave systems (Sendra & Reboreira 2012).

Although collembolan taxonomy at the higher taxonomic levels appears now to be stable (Rusek 2002, Deharveng 2004), at the intrageneric level modern biochemical/genetic techniques have repeatedly confirmed taxonomic splitting and uncovered cryptic species. Thus records hitherto attributed to Isotoma viridis (Bourlet, 1839) may now be called I. viridis, I. anglicana Lubbock, 1862, I. caerulea Bourlet, 1839 or I. riparia (Nicolet, 1842) (Simonsen et al. 1999, Burkhart & Filser 2005). The taxon hitherto called Isotomurus palustris (Müller, 1776) has been split into 14 species, including (for the UK alone) the original nominate species plus I. maculatus (Schäffer, 1896), I. prasinus (Reuter, 1891) and I. unifasciatus (Börner, 1901) (Carapelli et al. 2001, 2005). The morphological similarity between Dicyrtomina ornata (Nicolet, 1842) and D. saundersi (Lubbock, 1862) meant that authors from Lubbock (1873) to Hopkin (2007) doubted that they are validly separate species, but molecular data has confirmed the split, at least for mainland Europe (Fanciulli et al. 1995, 2001a). Porco et al. (2013a) attempted to determine Collembolan species richness at Churchill (Manitoba, Canada) by COI barcoding (using a 14% threshold for species separation), and found approximately twice as many genetic species (97) as morphospecies (45). Several common widespread, common Collembola (including Podura aquatica, Folsomia quadrioculata, Isotomurus plumosus and Sminthurides malmgreni) proved to conceal multiple Molecular Operational Taxonomic Units, effectively cryptic species.

To be confident that one is dealing with cryptic species one needs to cross-check mitochondrial phylogeny against a nuclear gene sequence, see, e.g., Porco et al. (2013b) who used mtCOI and nuclear 28S to show that the ubiquitous collembolan Parisolota notabilis (Schäffer, 1896) contains 4 genetic lineages each as distinct from each other as any of these ‘Parisolota notabilis’ are from other species in the genus.

The Collembolan genus Lepidocyrtus Bourlet, 1839 is a globally distributed genus, widespread in leaf litter (and to a lesser extent on arboreal surfaces) across the northern hemisphere. Current taxonomy of morphospecies within this genus depends heavily on the presence/absence of scales on the antennae and legs, the extent of hyper development of thorax 2 and on colour patterns. Taxonomists have expressed reservations about this, especially the problems of seeing scales on appendages after preservation (Hopkin 2007) and
recent genetic work has confirmed the presence of cryptic diversity within the genus. Cicconardi et al. (2009) collected seven morphospecies of Lepidocyrtus around one area of the western Mediterranean and found that two morphospecies [L. lignorum (Fabricius, 1775) and L. curvicollis Bourlet, 1839] were polyphyletic. ‘L. lignorum’ comprised four clades and ‘L. curvicollis’ formed two clades, all apparently separating over 10 million years ago. These clades were derived by a mitochondrial sequence (COII) and confirmed using the nuclear Elongation Factor 1 gene (EF1). Subsequently the same genus was studied in Panama, with a similar result: 5 morphospecies of Lepidocyrtus using COII and EF1 were found to contain 58 discrete lines with clade separations in excess of 10 million years ago, patterned by the palaeogeography of the region (Cicconardi et al. 2013). Soto-Adames (2002) used COI sequences to check the validity of existing Lepidocyrtus taxonomy in Puerto Rico. In most cases the validity of ‘colour-pattern species’ was upheld, with estimated separation dates in excess of eight million years. It seems an inescapable conclusion that we can expect to find further cryptic species within Lepidocyrtus, and that the taxonomy of the genus needs to be revised when sufficient data linking sequence data to morphospecies come available. There is no reason to think that Lepidocyrtus is an unusually diverse or ancient one, so it seems likely that many more cryptic species of Collembola remain undetected.

These deep divisions within the genus Lepidocyrtus can also be seen as an example of unexpectedly high phylobetadiversity (Emerson et al. 2011), an index of change in genetic composition across spatial gradients, either within species or of communities. The models underlying phylobetadiversity incorporate both history (since ancestral haplotypes may be inferred and their radiations mapped in space) and biotic interactions, making them a powerful new tool for the analysis of difficult communities such as soil fauna (Graham & Gillespie 2008).

Co-occurring with Lepidocyrtus spp in many habitats are members of the genus Entomobrya Rondani, 1861; both genera are routinely found in arboreal habitats as members of the bark and canopy faunas (Bowden et al. 1976, Shaw et al. 2007). The definitive review of the genus is given by Jordana (2012), though South (1961) is still a useful reference. Taxonomy is based on colour patterns and chaetotaxy, and for collections in the USA existing species boundaries are known to be unreliable, with ‘colour morphs’ generally being genetically isolated. However there are a minority of cases (e.g. Entomobrya clitellaria Guthrie, 1903) where valid species prove to have dimorphic colour patterns (Katz, in prep.).

### 3. Examples from contrasting biomes

It is instructive to compare the impact of genetic data on our understanding of biodiversity in Collembola from three contrasting biomes: the British Isles, remote oceanic islands and Antarctica. Our knowledge of UK Collembola suffers from the legacy of 150 years of study, since most old records are taxonomically unclear by modern species definitions, generating spurious endemics and unreliable species lists that need DNA sequence data to clarify true affiliations. By contrast remote oceanic islands are little-studied by Collembologists but do harbour endemic species of Collembola, unless displaced by invasive alien soil fauna. In Antarctica the few species of Collembola live in widely isolated populations that pre-date the Pleistocene ice ages, presumably relics of a far richer fauna that was exterminated by massive ice cover.

#### 3.1. The British Isles

From the viewpoint of evolutionary genetics, the British Isles are relatively young, having been repeatedly both connected to the mainland, and covered by ice or periglacial conditions until just ca. 10,000 years bp. (Bowen et al. 2002) They are also one of the best studied countries for many facets of biodiversity and were home to pioneers of Collembola taxonomy, so should have reliable lists and distribution data for the Collembola. In fact this is not the case.

Hopkin (2007) recorded 410 species of Collembola in the UK, but of these 101 names are suspect, defined as being based on a poorly preserved specimen or a probable misidentification. Within the especially problematic genus Protaphorura Absolon, 1901, 39 species names have been recorded for the UK (Hopkin 1997), of which only 24 were on the then world list maintained by Bellinger et al. (2009). By 2012 only 15 of these names remained on the world list, of which only 5 were considered to be ‘good’ species by Pomorski (1998). By contrast, it is probable that when molecular taxonomy is brought to bear on these 5 species of UK Protaphorura, they will be found to be hiding cryptic species that correspond poorly or not at all with the previously invalidated names. Hence the biogeography of Protaphorura in the UK is currently little known, despite there being many old records. Such bi-directional uncertainty undermines the credibility and accuracy of global species totals (May 1988, Mora et al. 2011). With all these caveats, the current list of accepted Collembola species in the UK is 322, of which 3 are invasives in specialised habitats and probably extinct.

Unlike Protaphorura, the genus Orchesella Templeton, 1835 is taxonomically uncomplicated and species may
be determined by eye using colour patterns. Distribution data for *Orchesella* species in the UK should therefore be reliable. The commonest and most distinctive is *Orchesella cincta* (Linnaeus, 1758), with 230 UK records of this species dating back to 1834 (Templeton 1835). The UK distribution map (Fig. 1) shows a strong bias towards the London area, with very few records in Cornwall, Wales or northern Scotland. Since the first author of the present review has collected this species with approximately equal ease in all parts of England, Wales and the Scottish lowlands, this distributional bias in fact surely reflects recorder bias (Prendergast et al. 1993), though its apparent absence from high mountains may be accurate, as it is said to be replaced by *Orchesella alticola* Uzel, 1891 (Hopkin 2007). In summary, the main influence visible on the biogeography of this common, large and distinctive species in the UK is recorder bias. For less common or distinctive species this bias can only be worse.

This well-known observer bias problem greatly reduces the usefulness of national records as tools to map distributions, but does not preclude a limited study of temporal change in the distribution of colour-pattern species. Thus the UK has two species of *Dicyrtomina*, *D. ornata* and *D. saundersi*. These can easily be distinguished by the pattern on abd 5 (a solid splodge in *D. ornata* but a multi-lobed pattern in *D. saundersi*). Older sources suggest *D. ornata* is the commoner species, contradicting modern experience, and the distribution of dates of records (Fig. 2) strongly suggests that *D. ornata* has undergone range contraction in the UK while *D. saundersi* has expanded. (Median year of record between the two species are highly significantly different; \( p < 0.001 \) by a U test). There is no obvious explanation for this, which may be a case of un-noticed competitive displacement, or even the recent displacement of a native line of *D. saundersi* with a more invasive foreign (continental) clade, as seems to have happened with the common reed *Phragmites australis* Trin. ex Streudel in the USA since 1910 (Saltonstall 2002).

The UK apparently supports nine endemic species of Collembola (Table 1), but most of these are *Incertae sedis* (based on inadequate descriptions and ruined type specimens). The only credible of these seems to be *Sminthurinus lawrencei* Gisin, 1963 (with two records and a distinctive mucro). By contrast, *Entomobrya lawrencei* Baquero & Jordana, 2008 (just one collection) is separated from *Entomobrya handschini* Stach, 1922 only by the chaetotaxy of the slide-mounted type specimen (Baquero & Jordana 2008), and the remaining endemics lack a useable type specimen. No UK endemic cave dwelling Collembola have yet been discovered, although ancient cave systems in unglaciated areas may...
Updating taxonomic biogeography in the light of new methods

contain some (e.g. an undescribed entomobryid has been collected from Lidstone Cave in Devon). In summary, the best explanation for the endemic Collembola in Table 1 is that they are merely taxonomic artefacts, though it is likely that there are UK-endemic Collembola waiting to be discovered and described. The impact of genetic data on our understanding of UK endemicity can be guessed as leading to a great increase in the number of UK-endemic clades, though whether these deserve specific status will depend on their distinctiveness.

3.2. Island faunas

Ancient oceanic islands can support high proportions of endemic species, including Collembola. Christiansen & Bellinger (1992) reported 92 Collembola species of the 154 species from Hawaiian Islands to be endemic. They also observed that there was good reason to think that both endemic and established immigrant Collembola had become extinct in the Hawaiian islands in the past 50 years, as had happened with birds and plants (e.g. Loope et al. 1988). There is no proof yet of island-endemic Collembola species being lost to invasive predators, unlike other invertebrates (Gillespie & Reimer 1993, Ashmole & Ashmole 2000, Coote & Loeve 2003), but there are examples suggesting that losses of endemic Collembola have occurred by ecological displacement. The island of St Helena, being ancient and isolated with many endemic taxa, would have been expected to harbour endemic (or at least exotic) Collembola. However, a systematic collection of Collembola from throughout the island in 2006 found no endemics at all, just a low diversity set of common European species dominated by 2 large, ubiquitous species Orchesella cincta (Linnaeus, 1758) and Tomocerus minor (Lubbock, 1867) by chaetotaxy of dens and dental pattern on mucro.

The island of Tenerife also showed a mixture of endemics and invasives, with records from two pan-European invasive species of Collembola (Parisotoma notabilis (Schäffer, 1896) and Ceratophysella gibbosa (Bagnall, 1940) but also two geographically localised surviving endemic lineages (Ramirez-Gonzalez et al. 2013).

3.3. Antarctica

The biogeography of Antarctic Collembola has received a disproportionate amount of study despite (or because) of their isolation, low biodiversity and high percentage
endemism (said to be 70% by Block 1984, but this figure will have changed as new species have been added to the Antarctic list). The consistent finding is that ice sheets are the geographical constraints acting as barriers to dispersal over geological time periods, with populations staying genetically isolated for millions of years. Eastern and western Antarctica contain almost entirely different Collembola, with only one species (*Friesea grisea* Schäffer, 1891) in common (Greenslade 1995). The division between maritime and coastal Antarctica is known as the Gressitt line and is such a sharp biogeographical discontinuity that Chown & Convey (2007) likened it to the Wallace line of south-east Asia. Within localised regions of Antarctica, the same pattern recurs that geographic isolation predicts biogeographical patterns (McGaughran et al. 2010). Thus, a major ice tongue (the Drygalski ice tongue) splits Victoria Land (VL) into two biogeographical zones (Northern VL and Southern VL), with different species of Collembola in the north and south (Sinclair & Stevens 2006). Similarly, the boundary between *Desoria klovstadi* (Carpenter, 1902) and *Gressittacantha terranova* Wise, 1967 is the Mariner glacier (Adams et al. 2006, Stevens et al. 2007). *Gressittacantha terranova* has its population trisected into genetically isolated groups by the Aviator and Campbell glaciers (Fanciulli et al. 2001b, Hawes et al. 2010). In summary the patterns of endemism in Antarctica are dominated by isolation imposed by physical geography.

These findings have been amplified by genetic data, which confirm isolation between remote populations, e.g. Stevens & Hogg (2003) found that *Gomphiocephalus hodgsoni* Carpenter, 1908 (Hypogastruridae) populations on Ross island and Beaufort islands were long-separated from a mainland (Taylor Valley) population (about 2% separation in COI). Antarctic Collembola populations consistently show long periods of divergence, often pre-dating the pleistocene. Mitochondrial sequence data show that the two populations of *Friesea grisea* separated by the Gressitt line show extreme genetic divergence (> 20% in protein coding genes), suggesting prolonged isolation (10–20 million years), pre-dating the development of the Antarctic ice sheet (Torricelli et al. 2010). Similarly, some endemic clades of *Cryptopygus* sp. on the Antarctic mainland were separated by 10–20 million years, suggested to have survived in ice-free refugia since the miocene (Stephens et al. 2006)

There is no evidence yet from either morphological or genetic studies that natural habitats in continental Antarctica have experienced invasions of alien Collembola, presumably due to the inaccessibility and harshness of conditions (Sinclair & Stevens 2006). However colonisation by non-native Collembola has already started in the maritime Antarctic (Slabber et al. 2007, Greenslade et al. 2012), and at least two non-Antarctic Collembola have been found in McMurdo and Scott Base research stations, in both cases associated with indoor horticultural facilities (Sinclair & Stevens 2006).

4. High throughput techniques

A current standard high-throughput platform, e.g. the Roche 454, allows for highly parallel sample processing. This means that researchers may work with pooled DNA samples (such as from a collection of arthropods whose DNA is co-extracted) and reconstruct the community from first generating clusters of similar amplicons, then identifying them by sequence comparison to known species (Yu et al. 2012). Unlike standard barcode sequences (e.g. the 648 base pairs of cytochrome oxidase 1), 454 amplicons tend to be shorter (100–300 base pairs) for reasons of technical stability. This is a transformative new technology, and consistently finds substantial numbers of unrecognised taxa in environmental collections (e.g. Sogin et al. 2006 for deep sea bacteria, Buée et al. 2009 for fungi, Nolte et al. 2010 for protozoa in lake water, Larsen et al. 2010 for human gut flora, Creer et al. 2010 for meiofauna, Fonseca et al. 2010 for general marine metazoan biodiversity, Porazinska et al. 2010 for rainforest nematodes). The disadvantage of this pooled DNA method is the lack of a collectable specimen of the unrecognised potentially new taxon, whereas DNA sequencing of single specimens can (in principle) be done using a non-destructive extraction method in order to retain the cuticle for morphological analyses (Porco et al. 2010).

Preliminary work applying 454 high-throughput technology to community samples of Collembola on Tenerife successfully reconstructed COI barcodes from the pooled amplicons, confirming the presence of *Parisotoma notabilis* and *Ceratophysella gibbosa* (ubiquitous in synanthropic habitats and found as identical sequences in mainland Europe), along with new sequences close to *Friesea truncata* Cassagnau, 1958 that appeared to be endemic to specific regions of the island (Ramirez-Gonzalez et al. 2013).

Although the field is in its infancy, a logical extension of using high throughput techniques for mapping the genetic diversity of Collembola would be applying the same principle to the amplification and sequencing of environmental DNA (eDNA). Andersen et al. (2012) have largely reconstructed the vertebrate fauna of a safari park by high throughput analyses of eDNA recovered from
the soil. Taberlet et al. (2012) showed that eDNA could be extracted from bulk soil and used to identify both oligochaetes and fungi. qPCR amplification of eDNA in a range of natural ponds streams and lakes in Europe found amplicons from the crustacean Lepidurus apus Leach, 1816 and the insect Leucorrhinia pectoralis Charpentier, 1825 (Thomsen et al. 2012), along with sequences from fish, mammals and riverside birds. Thus eDNA profiles might prove useful to define aquatic Collembola in the diverse but low-density communities found in freshwater systems such as upwellings (Bketschko & Christian 1989, Shaw et al. 2012) or caves (Deharveng et al. 2008, Sendra & Riboleira 2012). If eDNA extraction from soils and litters can be standardised, it may be possible to replace the expertise-demanding steps of high-gradient extraction, sorting then visual identification of Collembola (and all other soil fauna) by a wholly automated procedure that defines the community by the barcode sequences of its dominant eDNA amplicons. Under this scenario, worries about mis-identification of damaged or juvenile animals will be replaced by worries that unfamiliar sequences are merely NUMTS or amplification artefacts.

5. Conclusions

Our current understanding of Collembolan biogeography, based on morphospecies, is set to be transformed by recent increases in the availability of DNA sequencing. The first effect has been to complicate species distribution maps of the Collembola by revealing an unexpected degree of intra-specific complexity. Generally, though not inevitably, colour pattern ‘species’ have been confirmed to have been genetically isolated, and many literature records for Collembola will become progressively downgraded in value as their morphospecies are found to contain cryptic lineages. The initial consequences of widespread genetic data on our understanding of Collembolan biogeography will be disruptive as old taxa are split, but as data accumulate we will start to see the true biogeography of clades, unhampered by taxonomic uncertainty. In addition to refining our detection of species boundaries, genetic data allow dating of population splits. These have consistently given surprisingly large ages (> 10 MYBP), which correlate with local palaeogeography. Barcoding also allows identification of modern range expansions by aggressive/invasive clades. High throughput techniques raise the possibility of automating the collection of taxonomic barcodes for Collembola (or any other soil arthropod), although this has yet to be demonstrated in practice.

7. Acknowledgements

We are grateful to the Systematics Association and the Linnean Society for SYNTAX funding to undertake barcoding of UK Entomobrya and Lepidocyrtus. We are also grateful to the anonymous referees who made many helpful comments on the text.

8. References


Fratii, F., E. Dell’Ampio, S. Casasanta, A. Carapelli & P. P. Fanciulli (2000): Large Amounts of Genetic Divergence among Italian Species of the Genus Orchesella (Insecta, Collembola) and the
Relationships of Two New Species. – Molecular Phylogenetics and Evolution 17: 456–461.


