A critique of Rossberg et al.: noise obscures the genetic signal of meiobiotal ecospecies in ecogenomic datasets


High-throughput sequencing of DNA marker genes recovered from environmental samples (known as ecogenomics or metabarcoding) is an emerging tool for understanding patterns and processes in ecology and biodiversity [1]. The recent paper ‘Are there species smaller than 1 mm?’ [2] was inspired by a re-examination of published metabarcoding data from meiobiotic communities (including meiofauna and protists less than 1 mm) [3,4] which did not support the existence of well-defined genetic species. Rossberg et al. (hereafter referred to as RRM) noted that this observation is ‘at odds with much of the existing theoretical literature’ [2, p. 2]. Moreover, there are many empirical studies that demonstrate well-defined genetic species in meiobiotal organisms using phylogenetic, biological and morphological criteria [5–8]. Here, we offer a contrasting view highlighting a number of analytical and theoretical issues that cast doubt on their conclusion that available data are consistent with the hypothesis that ‘ecospecies form only for organisms with body sizes exceeding the millimetre scale’ [2, p. 6]. We provide new analyses to support our view that the cited observations for meiobiotic communities are affected by analytical artefacts generated by errors in the pyrosequencing reads that were not fully corrected in the original studies. We demonstrate that removing the noise generated by these errors results in small organisms exhibiting signals of species formation similar to those of larger species.

Models developed by RRM showed that ecospecies are unlikely to form under high values of $\mu K$, the product of mutation rate ($\mu$) and carrying capacity ($K$). Additionally, they showed that under conditions where ecospecies form, a lineage-through-time (LTT) plot on double-logarithmic axes exhibits a characteristic shape, including a plateau that separates the intra-specific and inter-specific diversification timescales. RRM compared their model predictions to published results using an analogue of LTT plots (the relationship between...
the published number of operational taxonomic units (OTUs) and increasing genetic clustering distance). Under these conditions, meiobiotic community data display a steady decline of OTU counts instead of the plateau displayed by DNA sequences from larger organisms (their fig. 1 [2]). RRM proposed three explanations why meiobiotic metabarcoding data do not exhibit the plateau: (i) the inflection points corresponding to ecospecies lie outside the range of genetic distances explored (1–10%); (ii) the plots represent methodological artefacts or (iii) species cannot form in meiofaunal assemblages, as μK is too large. RRM hypothesized that the higher carrying capacity of small organisms is potentially responsible for the inferred differences in species formation between large and small organisms [2], although they acknowledged that the underlying data may be affected by methodological artefacts, in particular that errors in pyrosequencing data can lead to inflated OTU counts.

We believe that the original meiobiotic results cited in RRM indeed are significantly influenced by methodological errors that obscure the true LTT relationships and alter their interpretation. It is now widely recognized that the metabarcoding protocol generates errors that lead to inflated diversity estimates in environmental samples [4,8]. Algorithms such as OCTUPUS [4] and ESPRIT [8], which were originally used to generate the results cited by RRM [3,4], attempt to account for combinations of these errors to provide more accurate estimates of taxonomic richness and composition.

Algorithms for detecting and removing errors are commonly tested by sequencing mock communities (pools of known DNA sequences) [9]. Mock communities are useful when addressing hypotheses about the signatures of species formation as they comprise unambiguous genetic species and should show LTT patterns characteristic of ecospecies formation. We therefore examined a widely used mock community comprising 21 genetically distinct species with high inter-species divergence, but no intra-specific variability at clustering distances above 1% [10]. LTT plots for the known reference sequences for this mock community (see the electronic supplementary material) display the expected plateau corresponding to the known number of genetic species at 1% clustering distance (figure 1—triangles). However, LTT plots for the raw pyrosequences from this community display a curve consistent with the lack of ecospecies over the same interval of genetic distances considered by RRM, and the number of OTUs is consistently overestimated (figure 1—diamonds). The result is similar using data processed with OCTUPUS (figure 1—circles). Now the plateau is more sharply defined with the inflection point close to the correct number of species, but it is not apparent within the genetic distance interval plotted by RRM (fig. 1 in [2]). Again, the number of OTUs is consistently overestimated at all clustering distances compared with the underlying reference sequences. This indicates that pyrosequencing noise, rather than the characteristics of the underlying community, is responsible for the pattern observed by RRM, and that the analytical methods used in the paper from which the data were taken were unable to account for this effect. Thus, even in cases where the underlying community comprises a known number of well-defined genetic species, failure to remove the noise generated by errors results in plots that lack a clustering threshold defining genetic species. When these errors are removed using the more effective method APDP [9] (figure 1—squares), the sequence data conform to the expected pattern characteristic of a community of well-defined genetic species, confirming that accurate error removal is possible, and vital to recovering the real signal in pyrosequenced DNA samples.

We applied the same error-removal approach (APDP) to one of the marine littoral benthos environmental datasets [4] (referred to as FO in [2]) to test whether errors similarly influence the observed relationships in RRM. We see a similar relationship for raw reads and error-cleaned sequences to that observed for the mock community sequences (figure 2), and the cleaned sequences now display the plateau that was absent from RRM’s results [2]. Here, the initial steep gradient representing intra-specific variation is absent, likely because this 18S region is highly conserved even between species of the same genus and intra-specific variation is expected to be below the range plotted by RRM [6,11–13].
The models developed by RRM suggest that ecotypes are unlikely to form under high values of $\mu K$, and RRM hypothesized that the higher carrying capacity $K$ of small organisms could be responsible for their observed differences in species formation [2]. Alternatively, we propose that the incomplete removal of minor sequence variants, generated errors from the real gene sequences in the underlying community, will mimic high mutation rates ($\mu$). That is, the patterns for pyrosequenced meiobiotal communities look similar to those generated by the model for high $\mu K$ values because $\mu$ is artificially elevated, not because small organisms have higher carrying capacities than large organisms.

Further to the interpretation of the cited empirical data, we have concerns about the application of the model to meiobiota. Firstly, the model describes ecotypes formation in a single panmictic population, and it is unclear whether the application of such a model to meiobiota community ecology is valid. The macroinvertebrate data analysed in RRM featured a limited number of species from one beetle genus and a complex of neotropical butterflies [13,14], whereas the meiobiota data included taxa from approximately 14 phyla of meiobiotic and protist lineages [3,4]. Given the variability in rates and patterns of molecular evolution, life histories, taxonomic complexity and population sizes represented in such communities, a continuum of lineages at different levels of sequence similarity should be predicted a priori. Secondly, the model used in RRM proposes a constant rate of asexual reproduction in all individuals and constant carrying capacity in a unidimensional ecological niche, with no opportunities for allopatic or parapatric speciation. This model reflects the mode of evolution of parthenogenetic species and of mtDNA in sexually reproducing species, but the meiobiota biosphere violates the assumptions of the model in many ways. While many meiobiota reproduce asexually, the majority are sexual. Importantly, meiobiota species also differ markedly in size (44 $\mu$m–1 mm [3,4]), consequent reproductive rate (e.g. between 1 and 55 generations per year in nematodes [15]) and carrying capacities are strongly affected by nutrient inputs [16]. Furthermore, interstitial taxa are notoriously patchy and often possess life histories lacking a dispersal phase [16]. Asynchronous reproductive rates, variable carrying capacity and heterogeneous ecological distributions will introduce temporal and population genetic variability in levels of gene flow, hence enhanced opportunity for drift and natural selection to act on temporally and spatially disjunct populations. Although all models are simplifications of real-world processes, the present simulations of RRM do not take into account these potentially significant deviations from the assumptions of their model. Finally, there is ample independent empirical evidence for species below 1 mm in size. Briefly, many taxon exist as populations that are reproductively isolated [5] and display concordant genetic variation at nuclear and mitochondrial loci [17], and even very closely related meiobiotic species display consistent morphological [6,18] and behavioural differences [19] that also coincide with ecological differentiation [19]. There is also clear evidence for biogeographic structuring of microscopic eukaryotes [20,21]. While the proposal that small organisms cannot form species might be supported if all organisms existed in conditions defined by the model, the existence of clearly defined genetic and ecological species supports the proposal that these models are not appropriate for organisms that compose meiobiotal communities.

Microscopic organisms are a vital component of the biosphere and underpin the majority of ecosystem processes. Given the recent advances in sequencing technology, we are now in a position to explore microscopic biodiversity, associated ecosystem function and reaction to environmental change. However, accurate interpretation of the taxonomic diversity of these data will be vital in forming and testing hypotheses about ecological and evolutionary patterns and processes. The extraordinary claim that species cannot form for small organisms is clearly at odds with much of the existing observational and theoretical literature, and it is far from clear that the currently available data provided by RRM support it. On the basis of our re-analyses that account for the noise in metabarcoding datasets, the available data are not consistent with the hypothesis that ecotypes form only for larger organisms. In conjunction with the existing literature, which provides strong evidence that meiobiotal species have been, and continue to be, observed experimentally, there is little empirical evidence to support a distinction between the abilities of large and small organisms to form genetic species on the basis of size alone.

References

11. Tang CQ, Leasi F, Obertegger U, Kieneke A, Baracough TG, Fontaneto D. 2012 The widely used small subunit 18S rDNA molecule greatly


