PROCEEDINGS OF THE ROYAL SOCIETY B

rspb.royalsocietypublishing.org

Invited reply

Cite this article: Rossberg AG, Rogers T, McKane AJ. 2014 Current noise-removal methods can create false signals in ecogenomic data. *Proc. R. Soc. B* **281**: 20140191.
http://dx.doi.org/10.1098/rspb.2014.0191

Received: 24 January 2014
Accepted: 19 February 2014

Author for correspondence:
Axel G. Rossberg
e-mail: axel@rossberg.net

The accompanying comment can be viewed at http://dx.doi.org/10.1098/rspb.2013.3076.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2014.0191 or via http://rspb.royalsocietypublishing.org.

Royal Society Publishing

Current noise-removal methods can create false signals in ecogenomic data

Axel G. Rossberg[1], Tim Rogers[2] and Alan J. McKane[3]

[1]Centre for Environment, Fisheries and Aquaculture Science (Cefas), Pakefield Road, Lowestoft NR33 0HT, UK
[2]Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK
[3]Theoretical Physics Division, School of Physics and Astronomy, University of Manchester, Manchester M13 9PL, UK

In a recent article [1], we examined a simple and rather generic individual-based model consisting of a large number of organisms that undergo reproduction with mutation and death through competitive interaction. Our analysis revealed that the formation and coherence of species depend crucially on population size. Specifically, species are unlikely to form under high values of μK, the product of mutation rate (μ) with carrying capacity (K). The model contains only the two basic processes of competition and mutation. This simplicity allowed us to uncover the root cause of a phenomenon that we believe could be quite general.

To what extent do our theoretical findings manifest themselves in real ecological systems? We investigated this question by comparing the outputs of our model [1] with phylogenetic data derived from ecogenomic surveys in the literature [2,3]. We found that the reconstructed phylogenetic trees of organisms with body size around the millimetre scale or below have similar characteristics to those occurring in our model for parameters where species do not form. This finding led us to ask: 'are there species smaller than 1 mm?'

In their comment, Morgan *et al.* [4] propose that our theoretical findings, though correct, are not applicable to real ecological communities. They argue that the work reported in references [2,3] was flawed, specifically suggesting that the counts of operational taxonomic units (OTUs, interpretable as lineages) reported in those articles are highly inflated owing to errors in sequencing. If this were true, then the patterns observed in our figure 1 [1] would be artefacts, and their similarity to the results of our model mere coincidence. We believe that Morgan *et al.* are unjustified in dismissing these data and the conclusions we drew from them, as we now explain.

For the datasets in question, the number of OTUs found declines steadily with the maximal permitted genetic distance within OTUs. In the light of our theoretical findings, this fact suggests the absence of genetic species. Morgan *et al.* would like to demonstrate that species have in fact formed. To do this, they propose to 'clean' the underlying sequence data by removing large numbers of sequences, so as to reveal a pattern that they believe has been obscured by noise. The remarkable effect of this removal process can be seen in figure 2 of their comment [4], in which a plateau in the number of OTUs is recovered from data where OTUs previously declined smoothly. Morgan *et al.* claim that this plateau, which was absent from the untreated data, is the one predicted by our theory in the case when species have formed.

We would like to urge caution. Selectively removing parts of a dataset can profoundly alter it, and often imposes a new structure not present in the original data. Any noise removal requires some preconceptions about structure in the underlying data; one must have an extremely good understanding of both the system and the noise in order to attempt this. For ecogenomic pyrosequencing data, this understanding might still be insufficient at present. One can test for bias in a denoising algorithm such as the one used by Morgan *et al.* by inputting data known to have no structure, and seeing if the algorithm creates a structure where none previously existed (a false-positive).

We have undertaken such a test. We applied the procedure used by Morgan *et al.* to two synthetic datasets, each consisting of 5000 sequences of 200 base pair length. The first set was designed to mimic the low-diversity mock community used by Morgan *et al.*; it was obtained by repeatedly sampling from a set of 10 initial

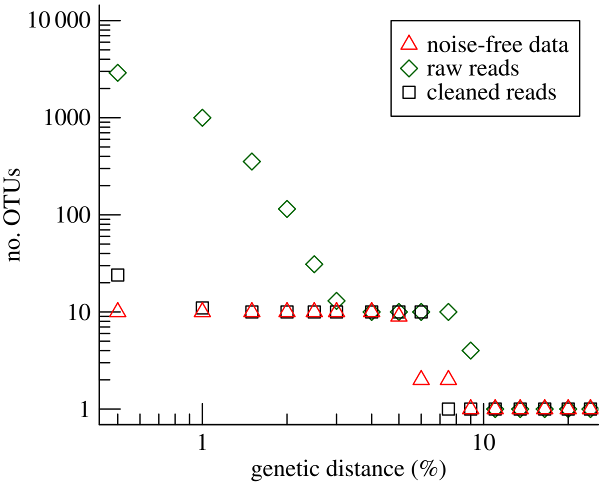

Figure 1. Relationship between genetic distance and observed number of OTUs in a sequence dataset derived from 10 unique and distinct sequences. (Online version in colour.)

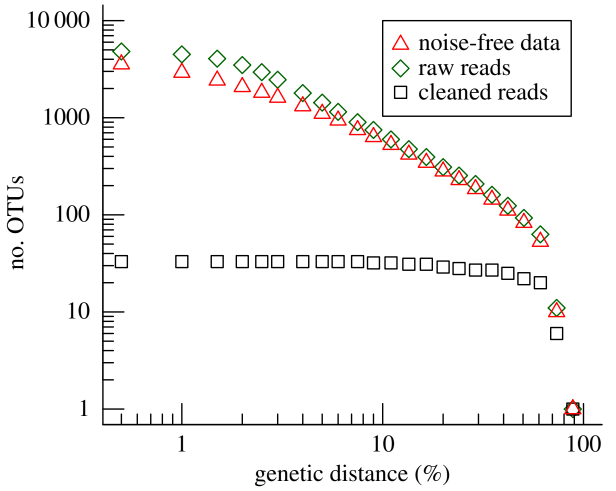

Figure 2. Relationship between genetic distance and observed number of OTUs in sequence data without species structure. (Online version in colour.)

sequences. The second was a high-diversity dataset generated by repeatedly replacing one randomly chosen sequence by a copy of another randomly chosen sequence, modified by random substitutions with probability 0.01 per base pair. This process simulates neutral evolution; after many iterations, it produces sequence data with no discernible species structure. Applying the fast clustering algorithm of OCTUPUS [3] to these datasets for a range of levels of genetic similarity leads to the expected [1,4] structures in figures 1 and 2 (triangles). We observe a plateau at low genetic distances for the low-diversity dataset, and a steady decline in the number of OTUs for the high-diversity set.

To model sequencing errors (the noise), sequences in both datasets were then subjected to random substitutions with a probability of 0.01 per base pair, simulating raw sequencer reads. In the output of the clustering algorithm (figures 1 and 2, diamonds), the addition of noise is observed to shift the original curves to the right. The low-diversity dataset exhibits highly inflated numbers of OTUs at small genetic distance, in line with concerns raised by Morgan *et al.* [4]. For the high-diversity dataset, however, the effect is weaker, suggesting that raw or slightly processed [3] high-diversity data can meaningfully be analysed in this format.

We then applied the APDP-SS algorithm [4,5] to delete some of the raw reads. The steps of the algorithm involving primer occurrences and comparison with GenBank were omitted as they are not relevant to synthetic data. For the low-diversity dataset, clustering after application of APDP (figure 1, squares) reveals a structure very similar to the original data, with a pronounced plateau at low genetic distances.

When applied to the high-diversity dataset, however, APDP again generates a plateau (figure 2, squares). This plateau is an artefact that would wrongly suggest the presence of only about 33 unique sequences in the original data; in fact, there were 4383. This result is important in the light of the similarity between our figure 2 and figure 2 of Morgan *et al.* [4]. In our case, the APDP algorithm has created a plateau from underlying data where this did not exist. In their case, Morgan *et al.* conclude that the algorithm has uncovered a true signal that was obscured by noise.

We have not analysed in detail exactly how APDP imposes the structure found in figures 1 and 2, although it appears to be mainly due to the blanket removal of all singleton sequences. This step was recognized by Morgan *et al.* as potentially problematic [5] but retained as 'a conservative approach', supported by its apparently successful inclusion in other recent algorithms [6]. Further analysis of this algorithm is clearly necessary. We have included as electronic supplementary material the R script used for the processing chain reported above, so that others may reproduce our test.

In our original article [1], we began a theoretical investigation of the basic mechanisms leading to genetic clustering. As well as challenging the result of references [2,3], Morgan *et al.* have speculated about some aspects of our model that they believe are too simple; for example, asexual reproduction. Our experience suggests that the mechanism of cluster formation is generic and will hold in more realistic models. Crucially, we have already demonstrated that the same phenomenon occurs in both the phenotypic [7] and genotypic [8] versions of the model, which appear very different *a priori*. We are currently studying other variants of the model incorporating sexual reproduction, and hope that other researchers will also investigate this question.

Although the simulated organisms in our models do not form species when μK is large, it is important to note that the populations do still exhibit a certain structure. In particular, while not forming species, individuals are phenotypically (or genetically) differentiated and adapted to their niches. We expect that future theoretical work will establish that many population-level features (including biogeographic structure, ecological differentiation, etc. [4]) are not dependent on the existence of coherent species. Indeed, even reproductive isolation of two subpopulations [9] does not conclusively demonstrate the separation of species; the same would be observed if specimens were taken from opposite ends of a ring species.

Further work is needed to accurately assess the extent of species formation in the meiofaunal biosphere. As we have seen, the handling of errors produced in current high-throughput sequencing technologies poses a major challenge. Possible areas for improvement include more extensive genetic and phylogenetic analyses of selected meiofaunal taxa, potential for synthesizing population-level surveys with selective whole-genome sequencing, and the development of more sophisticated mathematical models incorporating the effects of sequencing errors. The question of species formation is closely related to the problem of identifying so-called barcoding gaps [10,11]; however, in the present literature, the existence of species is often assumed *a priori*. Reanalysis of existing data without this assumption could well provide new insights. As the quantity and quality of ecogemonic data improves, we may find that the concept of ‘species’ is no longer central to our understanding of many aspects of ecology and biodiversity.

References

1. Rossberg AG, Rogers T, McKane AJ. 2013 Are there species smaller than 1 mm? *Proc. R. Soc. B* **280**, 20131248. (doi:10.1098/rspb.2013.1248)
2. Creer S *et al*. 2010 Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Mol. Ecol.* **19**, 4–20. (doi:10.1111/j.1365-294X.2009.04473. x)
3. Fonseca VG *et al*. 2010 Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nat. Commun.* **1**, 98. (doi:10.1038/ncomms1095)
4. Morgan MJ *et al*. 2014 A critique of Rossberg *et al*.: noise obscures the genetic signal of meiobiotal ecospecies in ecogenomic datasets. *Proc. R. Soc. B* **281**, 20133076. (doi:10.1098/rspb.2013.3076)
5. Morgan MJ, Chariton AA, Hartley DM, Hardy CM. 2013 Improved inference of taxonomic richness from environmental DNA. *PLoS ONE* **8**, e71974. (doi:10.1371/journal.pone.0071974)
6. Bragg L, Stone G, Imelfort M, Hugenholtz P, Tyson GW. 2012 Fast, accurate error-correction of amplicon pyrosequences using *Acacia*. *Nat. Methods* **9**, 425–426. (doi:10.1038/nmeth.1990)
7. Rogers T, McKane AJ, Rossberg AG. 2012 Demographic noise can lead to the spontaneous formation of species. *Europhys. Lett.* **97**, 40008. (doi:10.1143/JPSJ.77.044002)
8. Rogers T, McKane AJ, Rossberg AG. 2012 Spontaneous genetic clustering in populations of competing organisms. *Phys. Biol.* **9**, 066002. (doi:10.1088/1478-3975/9/6/066002)
9. Fonseca G, Derycke S, Moens T. 2008 Integrative taxonomy in two free-living nematode species complexes. *Biol. J. Linn. Soc.* **94**, 737–753. (doi:10.1111/j.1095-8312.2008.01015.x)
10. Wiemers M, Fiedler K. 2007 Does the DNA barcoding gap exist? A case study in blue butterflies (Lepidoptera: Lycaenidae). *Front. Zool.* **4**, 1–16. (doi:10.1186/1742-9994-4-8)
11. Meier R, Zhang G, Ali F. 2008 The use of mean instead of smallest interspecific distances exaggerates the size of the barcoding gap and leads to misidentification. *Syst. Biol.* **57**, 809–813. (doi:10.1080/10635150802406343)