



Citation for published version:

Rossberg, AG, Rogers, T & McKane, AJ 2014, 'Current noise-removal methods can create false signals in ecogenomic data', *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1783, 20140191, pp. 1 - 3. <https://doi.org/10.1098/rspb.2014.0191>

DOI:

[10.1098/rspb.2014.0191](https://doi.org/10.1098/rspb.2014.0191)

Publication date:

2014

Document Version

Peer reviewed version

[Link to publication](#)

This is the Author's Accepted Manuscript of a paper published in Rossberg, AG, Rogers, T & McKane, AJ 2014, 'Current noise-removal methods can create false signals in ecogenomic data' *Proceedings of the Royal Society B: Biological Sciences*, vol 281, no. 1783, 20140191, pp. 1 - 3., and available online via: <http://dx.doi.org/10.1098/rspb.2014.0191>

University of Bath

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Current noise removal methods can create false signals in ecogenomic data

AXEL G. ROSSBERG¹, TIM ROGERS² & ALAN J. MCKANE³

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 & Astronomy, The University of 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 which undergo reproduction with mutation and death through competitive interaction. Our analysis revealed that the formation and coherence of species depends 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 which, we believe, could be quite general.

To what extent do our theoretical findings manifest themselves in real ecological systems? We investigated this question in [1] by comparing the outputs of our model with phylogenetic data derived from ecogenomic surveys in

22 the literature [2, 3]. We found that the reconstructed phylogenetic trees of
organisms with body size around the millimetre scale or below have similar
24 characteristics to those occurring in our model for parameters where species
do not form. This finding led us to ask the question: “Are there species
26 smaller than 1mm?”.

In their comment [4], Morgan *et al.* propose that our theoretical findings,
28 though correct, are not applicable to real ecological communities. They
argue that the work reported in references [2, 3] was flawed, specifically sug-
30 gesting that the counts of operational taxonomic units (OTUs, interpretable
as lineages) reported in those articles are highly inflated due to errors in se-
32 quencing. 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
34 coincidence. We believe that Morgan *et al.* are unjustified in dismissing this
data and the conclusions we drew from it, as we now explain.

36 For the datasets in question, the number of OTUs found declines steadily
with the maximal permitted genetic distance within OTUs. In light of our
38 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
40 they propose to “clean” the underlying sequence data by removing large
numbers of sequences, so as to reveal a pattern which they believe has been
42 obscured by noise. The dramatic effect of this removal process can be seen
in Figure 2 of their comment [4], in which a plateau in the number of OTUs
44 is recovered from data where OTUs previously declined smoothly. Morgan
et al. claim that this plateau, which was absent from the untreated data, is
46 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
48 profoundly alter it, and often imposes a new structure not present in the orig-

inal data. Any noise removal requires some preconceptions about structure
50 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 py-
52 rosequencing data, this understanding might still be insufficient at present.
One can test for bias in a denoising algorithm such as the one employed by
54 Morgan *et al.* by inputting data which is known to have no structure, and
seeing if the algorithm creates a structure where none previously existed (a
56 false positive).

We have undertaken such a test. We applied the procedure used by Morgan
58 *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
60 community used by Morgan *et al.*; it was obtained by repeatedly sampling
from a set of 10 initial sequences. The second was a high-diversity dataset
62 generated by repeatedly replacing one randomly chosen sequence by a copy
of another randomly chosen sequence, modified by random substitutions at
64 a rate 0.01. This process simulates neutral evolution; after many iterations
it produces sequence data with no discernible species structure. Applying
66 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 Figs. 1
68 and 2 (red 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
70 high-diversity set.

To model sequencing errors (the noise), sequences in both datasets were
72 then subjected to random substitutions with a probability of 0.01 per base
pair, simulating raw sequencer reads. In the output of the clustering algo-
74 rithm (Figs. 1,2, green diamonds), the addition of noise is observed to shift
the original curves to the right. The low-diversity dataset exhibits highly

76 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
78 effect is weaker, suggesting that raw or slightly processed [3] high-diversity
data can meaningfully be analysed in this format.

80 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 com-
82 parison with GeneBank were omitted as they are not relevant to synthetic
data. For the low-diversity dataset, clustering after application of APDP
84 (Fig 1, black squares) reveals a structure very similar to the original data,
with a pronounced plateau a low genetic distances.

86 When applied to the high-diversity dataset, however, APDP again gener-
ates a plateau (Fig 2, black squares). This plateau is an artefact which
88 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 light
90 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
92 where this did not exist. In the other case, Morgan *et al.* conclude that the
algorithm has uncovered a true signal which was obscured by noise.

94 We have not analysed in detail exactly how APDP imposes the structure
found in Figs. 1 and 2, although it appears to be mainly due to the blanket
96 removal of all singleton sequences. This step was recognised as potentially
problematic in [5] but retained as “a conservative approach”, supported
98 by its apparent successful inclusion in other recent algorithms [6]. Further
analysis of this algorithm is clearly necessary. We have included as supple-
100 mentary material the R script used for the processing chain reported above,
so that others may reproduce our test.

102 In our original article [1], we began a theoretical investigation of the basic

mechanisms leading to genetic clustering. As well as challenging the result
104 of Refs. [2, 3], Morgan *et al.* have speculated about some aspects of our
model which they believe are too simple, for example, asexual reproduction.
106 Our experience suggests that the mechanism of cluster formation is generic
and will hold in more realistic models. Crucially, we have already demon-
108 strated that the same phenomenon occurs in both the phenotypic [7] and
genotypic [8] versions of the model, which appear very different *a priori*.
110 We are currently studying other variants of the model, incorporating sex-
ual reproduction, and hope that other researchers will also investigate this
112 question.

Although the simulated organisms in our models do not form species when
114 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
116 phenotypically (or genetically) differentiated and adapted to their niches.
We expect that future theoretical work will establish that many population-
118 level features (including biogeographic structure, ecological differentiation,
etc. [4]) are not dependent on the existence of coherent species. Indeed,
120 even reproductive isolation of two sub-populations [9] does not conclusively
demonstrate the separation of species; the same would be observed if speci-
122 mens were taken from opposite ends of a ring species.

Further work is needed to accurately assess the extent of species forma-
124 tion in the meiofaunal biosphere. As we have seen, the handling of errors
produced in current high-throughput sequencing technologies poses a major
126 challenge. Possible areas for improvement include: more extensive genetic
and phylogenetic analyses of selected meiofaunal taxa, potential for syn-
128 thesising population-level surveys with selective whole-genome sequencing
and the development of more sophisticated mathematical models incorpo-

130 rating the effects of sequencing errors. The question of species formation is
closely related to the problem of identifying so-called barcoding gaps [10, 11],
132 however, in the present literature the existence of species is often assumed
a priori. Re-analysis of existing data without this assumption could well
134 provide new insights. As the quantity and quality of ecogenomic data im-
proves, we may find that the concept of ‘species’ is no longer central to our
136 understanding of many aspects of ecology and biodiversity.

References

- 138 [1] Rossberg, A. G., Rogers, T. & McKane, A. J., 2013 Are there species
smaller than 1mm? *Proc. R. Soc. B* **280**, 1767.
- 140 [2] Creer, S., Fonseca, V. G., Porazinska, D. L., Giblin-Davis, R. M.,
Sung, W., Power, D. M., Packer, M., Carvalho, G. R., Blaxter, M. L.,
142 Lamshead, P. J. D. *et al.*, 2010 Ultrasequencing of the meiofaunal
biosphere: practice, pitfalls and promises. *Mol. Ecol.* **19**, 4–20. (doi:
144 10.1111/j.1365-294X.2009.04473. x).
- [3] Fonseca, V. G., Carvalho, G. R., Sung, W., Johnson, H. F., Power,
146 D. M., Neill, S. P., Packer, M., Blaxter, M. L., Lamshead, P. J. D.
& Thomas, W. K., 2010 Second-generation environmental sequencing
148 unmasks marine metazoan biodiversity. *Nature communications* **1**, 98.
- [4] Morgan, M. J., Bass, D., Bik, H., Birky, C. W., Blaxter, M., Crisp,
150 M. D., Derycke, S., Fitch, D., Fontaneto, D., Hardy, C. M. *et al.*,
2014 A critique of Rossberg et al.: noise obscures the genetic signal
152 of meiobiotal ecospecies in ecogenomic datasets. *Proc. R. Soc. B (to
appear)* .
- 154 [5] Morgan, M. J., Chariton, A. A., Hartley, D. M. & Hardy, C. M., 2013

- Improved inference of taxonomic richness from environmental DNA.
156 *PloS one* **8**, e71974.
- [6] Bragg, L., Stone, G., Imelfort, M., Hugenholtz, P. & Tyson, G. W.,
158 2012 Fast, accurate error-correction of amplicon pyrosequences using
Acacia. *Nature Methods* **9**, 425–426.
- [7] Rogers, T., McKane, A. J. & Rossberg, A. G., 2012 Demographic noise
160 can lead to the spontaneous formation of species. *Europhys. Lett.* **97**,
162 40008. (doi:10.1143/JPSJ.77.044002).
- [8] Rogers, T., McKane, A. J. & Rossberg, A. G., 2012 Spontaneous genetic
164 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
166 two free-living nematode species complexes. *Biol. J. Linn. Soc.* **94**,
168 737–753.
- [10] Wiemers, M. & Fiedler, K., 2007 Does the DNA barcoding gap exist?—a
170 case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in
Zoology* **4**, 1–16.
- [11] Meier, R., Zhang, G. & Ali, F., 2008 The use of mean instead of smallest
172 interspecific distances exaggerates the size of the barcoding gap and
174 leads to misidentification. *Systematic Biology* **57**, 809–813.

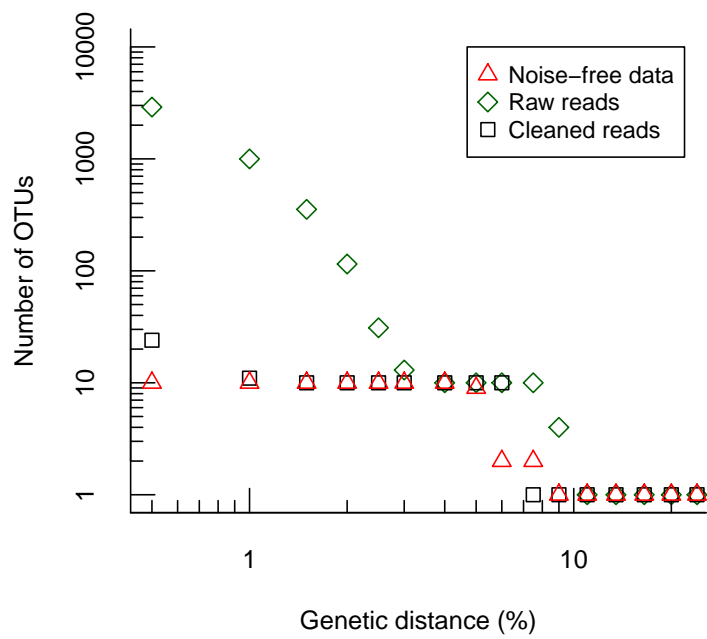


Figure 1: Relationship between genetic distance and observed number of OTUs in a sequence dataset derived from 10 unique and distinct sequences.

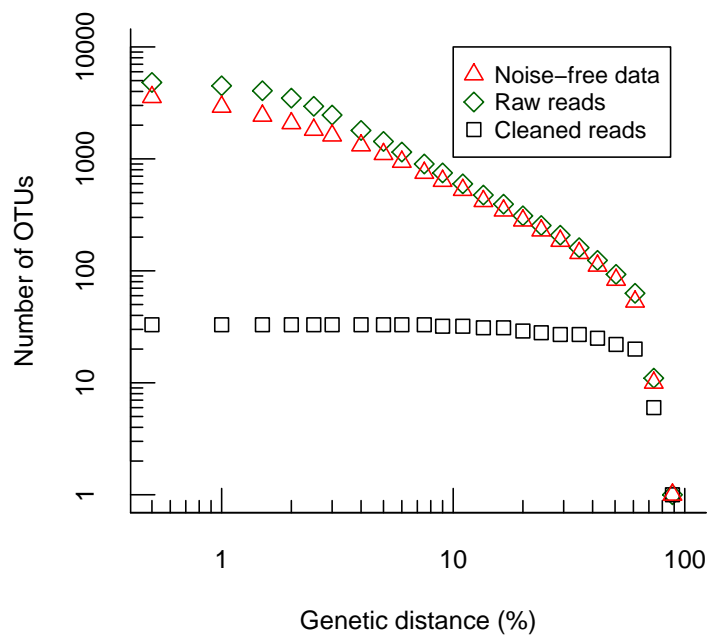


Figure 2: Relationship between genetic distance and observed number of OTUs in sequence data without species structure.