Review of: "Shotgun metagenomics of soil invertebrate communities reflects taxonomy, biomass and reference genome properties"

Henrik Nilsson[1]

1 Göteborg University

Potential competing interests: The author(s) declared that no potential competing interests exist.

Review of Schmidt et al., ”Shotgun metagenomes of soil…”

Schmidt et al. present a timely and thought-provoking piece on metagenomics biases and potential pitfalls. I’m all for it.

Line 37 – “genomes, and” > “genomes and”

44 - “DNA sequencing, and” > “DNA sequencing and”

52 – “Biodiversity research, and” > “Biodiversity research and”

62 – Not all metabarcodes qualify as genes in the strict sense of the word, unlike what this sentence suggests. The formal fungal barcode – the nuclear ribosomal ITS region – is not a gene, for instance. I suggest “a genetic marker” instead of “a marker gene”.

66 - What about “: (i) you can effectively … , and (ii) small or rare…”. Or am I reading the sentence wrong?

66 – “, and avoid” > “ and avoid”

74 – is “circumvent” too strong a word here? I think so, and I would recommend “ameliorate” or something similar instead.

88 – the authors use several different solutions to enumerations. I would suggest the Oxford comma throughout, rather than variously using the Oxford comma, variously dropping it, and variously dropping the “and” altogether. On this line I would recommend “, and Oribatida”. Similarly on the next line, 156, and so on.

95 versus 96 – what is the difference between having “more biomass” and “larger biomass”? Or is no difference intended? Please clarify.

98 – I like the attention paid to replication in this study.

Table 1 – for Tardigrada and nearly all other groups of organisms, the author(s) of the species names are given within parentheses. However, this is not done for Enchytraeidae. What is this difference supposed to tell the reader? Different nomenclatural codes (but they’re not?)? Please clarify.

104 – I’d go for “a, b, and c”

112 – I’d go for “König 18, and”. After all, the authors use the Oxford comma just a few lines back, 105.

133 – “bp” - but an individual read is composed of bases (b) rather than base-pairs (bp), so why is “bp” use here?

139, 173, and elsewhere – I like the attention paid to open data and reproducibility. The use of FigShare is clever.

140 – please add a qualifier to “250 species”, such as “250 invertebrate species” or “250 eukaryotic species”. “250 species” could be interpreted to include prokaryotes, after all. And like the authors say, too many things in the life sciences are shaped by the needs of prokaryotic biology.

151 – I’d go for “from the data”

151 – I’d go for “Unclassified reads and”

158 – I’d go for “, and genome”

165, 167. In the interest of reproducibility, please provide the version number of these software packages.

171 - I’d go for “…in an average of 69 million paired-end reads per…”

188 – I’d go for “slightly – but statistically insignificantly – lower”

184-196 make for interesting reading.

197 – This is a very abrupt introduction of the abbreviation “GLM”, which is not defined, and which is not used elsewhere in the manuscript either. To be honest I don’t know what it stands for. Google suggested “Geostationary Lightning Mapper” – which does not strike me as the interpretation intended by the authors.

197 – I’d go for “completeness, and”

197 – “size on taxonomically assigned [PARAGRAPH STOP]”. At least one word is missing here. Some issue with the PDF conversion of the figure perhaps? Right now the sentence is unintelligible, at least to me, I’m afraid.

197 – The species names in the figure should be given in italics.

198 - I’d go for “, and genome size”

211 – I’d go for “thresholds but dropped”

222 – I’d go for “in low-complexity regions”

226 – I’d go for “in DNA-based analysis”

231 – I’d go for “, and biomass”

252 - Primer mismatch is a non-trivial problem, I agree. But so is amplicon length. PCR will favour shorter amplicons at the expense of longer ones, so species with longer barcodes won’t show up in metabarcoding, whereas those with shorter barcodes will dominate (https://academic.oup.com/femsec/article/82/3/666/492046 , https://mycokeys.pensoft.net/articles.php?id=4852&journal_name=mycokeys). Does the present wording give an overly simplistic view of the situation?

259 – I’d go for “of the number of reads”

267 – I’d go for “genome size influences”

269 – Shouldn’t this simply be “Beszteri et al., 2010”? (If not, then why not?) Similarly on 273.

273 – I’d go for “low-coverage”

276 – Hm. Up to this point the authors have been careful to point out that they work with invertebrates. But on this line they extrapolate their results to “metazoan mock communities”. Is it clear that the results of the authors can be extrapolated to all metazoans? Does this need discussion?

279 – I’d go for “, and developed”. That said, the authors are making a very good point on 278-279.

280 – I’d go for “genetic markers” instead of “marker genes”, since not all metabarcodes are genes in the strict sense of the word.

281 – I’d go for “, and repeat elements”

282 – “for some time” must be the understatement of the century. Have a look at https://link.springer.com/article/10.1007/s13225-021-00472-y for instance. We have something like 10 robust, complete fungal genomes right now (plus two truckloads of incomplete genomes, obviously). That took 30 years. When will we have the remaining 6.3 million ones? (And that’s just fungi. The tree of life is composed of more things than fungi, obviously.) I’d like to replace “for some time” with “for thousands of years” or something like that.

286 – Here the authors extrapolate their findings not only to metazoans, but to all “eukaryotic communities”. Is this justified? Does this need discussion?

290 and 292 – I’d use the Oxford comma here too. That said, 289-293 clearly pack important take-home messages. I’m all ears.

301, 302 – Oxford comma.

309 – *Acta Zoologica Hungarica* ?

314 and elsewhere – Why is the article title given in such a way that verbs and key nouns are given with leading uppercase letters here (“Metagenomics Reveals the Ancient Origin…”), but not on 320 (“amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples”)? Is a difference intended? If so, what? Please resolve this in a consistent way throughout the list of references.

352 – This reference is incompletely specified. And why is it important for the reader to know how the authors retrieved this reference?

363 versus 408 – Please homogenize the way books are specified.

417 – Species names should be written in italics.

I miss a legend for the supplementary item. Also, what is a “Countgroup”? That word is not found in the body of the manuscript.