A global metabarcoding analysis expands molecular diversity of Platyhelminthes and reveals novel early-branching clades

Konstantina Mitsi1, Alicia S. Arroyo1 and Iñaki Ruiz-Trillo1,2,3

1Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, Barcelona 08003, Catalonia, Spain
2Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona (UB), Barcelona 08028, Catalonia, Spain
3ICREA, Pg. Lluís Companys 23, Barcelona 08010, Catalonia, Spain

KM, 0000-0003-0228-2059; ASA, 0000-0001-7983-8583; IR-T, 0000-0001-6547-5304

Understanding biological diversity is crucial for ecological and evolutionary studies. Even though a great part of animal diversity has already been documented, both morphological surveys and metabarcoding analyses have previously shown that some animal groups, such as Platyhelminthes, may harbour hidden diversity. To better understand the molecular diversity of Platyhelminthes, one of the most diverse and biomedically important animal phyla, we here combined data from six marine and two freshwater metabarcoding expeditions that cover a broad variety of aquatic habitats and analysed the data by phylogenetic placement. Our results show that a great part of the hidden diversity is located in early-branching clades such as Catenulida and Macrostomorpha, as well as in late-diverging clades such as Proseriata and Rhabdocoela. We also report the first freshwater record of Gnosonesimida, a group previously thought to be exclusively marine. Finally, we identified two putative novel freshwater Platyhelminthes clades that branch between well-defined orders of the phylum. Thus, our analyses of several environmental datasets confirm that a large part of the diversity of Platyhelminthes remains undiscovered, point to groups with more potential novel species and identify freshwater environments as potential reservoirs for novel species of flatworms.

1. Introduction

To understand past and present biological processes and to make meaningful decisions for the future, it is of pivotal importance to decipher extant biodiversity [1]. Accurate biodiversity assessment is difficult because of sampling biases and the limitations of morphology-based taxonomic identification [2]. Sampling biases include restrictions in sampling site accessibility and preferential collection of specimens due to methodological constraints, both of which lead to a non-representative sample of the community under study. Traditional identification methods based on morphology are low throughput, time and resource consuming, require high taxonomic expertise that is particularly rare for most of the groups that are not well studied and fail to cope with cryptic diversity. As a consequence, it is estimated that real extant species diversity probably exceeds by 10-fold the current number of described species [3]. Although the unicellular eukaryotic lineages suffer from this bias more than their multicellular counterparts [4,5], there are groups of animals for which an assessment of diversity is incomplete [6,7].

Platyhelminthes (flatworms) is one of the most diverse [8] and relatively well-studied animal phyla. Initially considered to be an early-branching bilaterian...
potential novel, undescribed molecular diversity. The diversity of Platyhelminthes at the level of orders and detect the gap, we here analysed different environmental datasets from data on Platyhelminthes has never been done. To fill this gap, we here analysed different environmental datasets from various aquatic environments [20–24]. However, to our knowledge, a thorough analysis of metabarcoding data on Platyhelminthes has never been done. To fill this gap, we here analysed different environmental datasets from six marine and two freshwater habitats to both explore the diversity of Platyhelminthes at the level of orders and detect potential novel, undescribed molecular diversity.

2. Material and methods

We collected a total of 1380 representative sequences of clustered operational taxonomic units (OTUs) identified as Platyhelminthes in six marine and two freshwater environmental surveys. Each survey targeted a different hypervariable region of the 18S rRNA gene (table 1). We assigned sequences to either groups or taxonomic categories based on searches of the SILVA 128 SSU reference database. We constructed a reference tree of complete 18S rDNA sequences retrieved from the National Center for Biotechnology Information (NCBI) nucleotide database to use as a backbone for phylogenetic placement. We aimed to include all the known extant diversity of Platyhelminthes focusing on the freeliving representatives. To this end, we conducted a bibliographic search to select complete sequences of all the known families for each order of the phylum. We restricted our taxon sampling to 455 taxa so that the resulting reference tree could easily be visually inspected, while encompassing the diversity of the extant Platyhelminthes according to the latest complete phylogeny of the phylum.

We filtered our initial metabarcoding dataset both by alignment (using PaPaRa v. 2.5) to align the sequence queries to the reference alignments and by phylogenetic placement (using the EPA as implemented in RAxML). All trimming was done with trimAL. We removed all sequences that (i) did not align in the correct hypervariable region, (ii) were placed in the outgroup of the reference tree, (iii) had extremely long branches in the best-hit placement tree making the rate of nucleotide substitutions greater than 1, and (iv) had a placement hit both in Platyhelminthes branches but also in the outgroup branches. The filtered dataset (843 OTUs) was placed onto the reference tree using two different phylogenetic placement algorithms, RAxML-EPA and pplacer. We compared the resulting jplace files using compare_jplace_files.cpp as implemented in RAxML 

| Table 1. Sampling and filtering information. |
|-------------------------------------------|
| **dataset** | **habitat** | **18S rRNA variable region** | **no. OTUs before filtering** | **no. OTUs after filtering** | **reference** |
| TaraOceans | global marine water column | V9 | 185 | 117 | de Vargas et al. [25] |
| BioMarks | European coastal benthos and marine water column | V4 | 33 | 33 | Massana et al. [26] |
| Metabarpark | Atlantic and Mediterranean marine hard-bottom benthos | V7 | 232 | 1 | Wangensteen et al. [27] |
| DOSMARES | Mediterranean marine deep sea benthos (Blanes canyon) | V7 | 123 | 1 | Guardiola et al. [28] |
| INDEMARES | Mediterranean marine deep sea benthos | V7 | 29 | 0 | Guardiola et al. [29] |
| DeepSea | Atlantic and Pacific marine deep sea benthos | V8–V9 | 80 | 17 | Bik et al. [30] |
| Parana | river in Argentina | V4 | 14 | 7 | Arroyo et al. [31] |
| Sanabria | glacier lake in Iberian peninsula | V4 | 684 | 667 | in preparation |

We constructed maximum-likelihood trees for all the queries with top placement outside the known Platyhelminthes orders: (i) a tree in which we combined full-length reference sequences with short queries and (ii) a tree in which we manually trimmed the reference alignment to the length of the short queries. The trees were built (i) in RAxML under the GTR+Gamma substitution model with 1000 rapid bootstrap replicates and (ii) in IQTREE under the TN+F+R8 substitution model with 1000 ultrafast bootstraps and tested tree branches by SH-like LiTuck et al with 1000 replicates. All trees were visualized in iTOL.
identify potential diversity hotspots for future expeditions. To this end, we compiled the most comprehensive flatworm metabarcoding 18S rDNA dataset to date, comprising 1380 query sequences from six marine and two freshwater environmental surveys (table 1).

We first checked how many of those sequences corresponded to taxa already described and sequenced and how many represented novel taxa. We performed BLASTn searches against the SILVA 128 SSU database to confirm the identity of these potential Platyhelminthes. Among the 1245 query sequences that returned a flatworm sequence as the first hit, 60% had less than 97% sequence identity with the reference sequences (figure 1a). All groups, both parasitic and free living, showed high percentages of sequences with low BLAST identity (less than 97% sequence identity), Proseriata, Prolecithophora and Trematoda being the groups with the highest percentages. For example, 95% of the 248 Proseriata sequences had less than 97% sequence identity with reference sequences.

The quality of the reference database is of pivotal importance to evaluate the number of novel species inside a clade. In biodiversity assessments based on sequence similarity methods, a good reference database includes as many sequences as possible. By contrast, to evaluate the number of novel species using a phylogeny-driven approach, the number of taxa in a good reference tree should be small enough to allow the visual inspection of the results and broad enough to encompass all existing diversity. To this end, we inferred our Platyhelminthes 18S rDNA reference tree based on a broad taxon sampling of 455 complete 18S rDNA sequences including representatives from all major flatworm clades. Although our reference tree did not recover the same topology of orders as in multigene phylogenetic analyses [16,38], most orders were monophyletic. This is
important for the subsequent placement analysis as all the queries that fall into the known delimited orders expand the diversity inside these orders and all the queries that fall between well-defined orders represent completely novel molecular diversity.

Given that each study targeted different hypervariable regions of the 18S rRNA gene, our initial dataset was a mixture of V4, V7, V8–V9 and V9 18S rDNA queries (table 1). Queries from a hypervariable region must map to a full-length 18S rDNA with minimal ambiguity to serve as a reliable phylogenetic marker. However, in many cases, queries do not align unambiguously because of the fast-evolving nucleotide sites resulting in unreliable trees. To overcome this pitfall, we refined the unfiltered dataset of 1245 query sequences by alignment. More than one-quarter of the initial sequences were removed because of misalignment (table 1). The majority of V7 queries were removed during this filtering step, indicating that the V7 hypervariable region was too short and variable to be useful as a molecular marker for Platyhelminthes. By contrast, all V4 and V9 queries were retained, showing that these variable regions can serve as quality molecular markers for Platyhelminthes.

Our phylogenetic placement analyses showed that the majority of OTUs grouped with free-living taxa (figure 1b). We detected phylogenetic placements in the internal nodes of early-branching clades such as Catenulida, Macrostomorpha, Prorhynchida and Polycladida that potentially indicate novel groups yet to be described. Many OTUs were placed within Polycladida, a well-described clade with more than 800 described species, of which only 30 have representative sequences. Thus, these placements probably reflect a lack of molecular data in the reference database rather than real novel diversity. Many other OTUs were placed within Proseriata and Rhabdocoela, probably representing novel diversity.
given that these two clades are well sampled for the 18S rRNA gene.

We then inquired whether we could detect marine OTUs in groups considered freshwater and vice versa (figure 1b,c). As expected, marine groups were only detected in marine samples and freshwater groups in freshwater samples, except for one clade, Gnosonesimida. We recovered the first freshwater record of Gnosonesimida, a group formed by only six described species thought to be exclusively marine [16,18]. Polycladida and Prolecithophora have both freshwater and marine representatives, but we could only detect them in our marine datasets. Even though Neodermata is formed by obligate parasites, we recovered OTUs in the marine water column. Those OTUs were placed inside the three orders of Neodermata, suggesting that free-living stages may have been sampled.

We then analysed potential novel clades within Platyhelminthes. In our best-hit placement tree (figure 1c), we localized OTUs that were placed outside the limits of known flatworm orders (figure 2a) with high likelihood–weight scores and characterized them as ‘interesting placements’. We inferred a maximum-likelihood tree from the alignment of the full-length reference sequences and those OTUs with phylogenetically interesting placements (figure 2b). The phylogenetic tree revealed two novel freshwater clades, clade 1 and clade 2, branching as monophyletic clades in between major Platyhelminthes orders. Clade 1 was formed by three sequences that had 92% sequence identity with sequences of the genus Castrada (Typhloplanidae) and clade 2 by 19 sequences with sequence identity to Otomosostoma auditivum and Invenusta aestas (Coelogynoporidae) that ranged between 91% and 94%. We also inferred a maximum-likelihood tree using only the V4 hypervariable region of the reference alignment and the short queries; this tree was not informative because of the weak phylogenetic signal. Even though the exact phylogenetic position of the new clades within Platyhelminthes remains unclear, they certainly form two separate, well-defined groups, probably in early-branching positions.

Phylogenetic placement outperforms the conceptually problematic but often used practice of reconstructing de novo phylogenies from short reads that do not contain sufficient phylogenetic signal to reproduce a reasonable tree. It is a reliable method to classify short DNA sequences, a common output of metabarcoding and metagenomic studies, and has been extensively used for taxonomic assignment in diversity studies. Overall, our analyses show a high diversity of Platyhelminthes in both marine and freshwater environments, with the latter habitat likely containing as yet unnamed taxa. We found that Proseriata and Rhabdocoela are the two flatworm groups with more potential novel species. Our data also show a high novelty of molecular data in Catenulida and Macrostromorpho that may correspond either to unsequenced data or to new taxa. Moreover, we identified, in freshwater environments, two novel clades that group outside the well-known Platyhelminthes orders. While our data demonstrate the utility of metabarcoding analyses in search of novel diversity, we emphasize the need for more traditional taxonomic efforts to have a good understanding of animal diversity.

Data accessibility. Alignments, trees and jplace files are available as electronic supplementary material.

Authors’ contributions. K.M. collected the data, conducted the analyses, interpreted the data, designed the figures and wrote the draft and the revised manuscript. A.S.A. collected the data and revised the manuscript. I.R.-T. conceived and designed the study, supervised the work and revised the manuscript. All authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by grants (BFU2014-57779-P and BPF2017-90114-P) from Ministerio de Economía y Competitividad (MINECO), Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarrollo Regional (FEDER) to I.R.-T. It has also received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. H2020-MSCA-ITN-2015-675752 (SINGEK Project, http://www.singek.eu/).

Acknowledgements. We are thankful to Marta Álvarez, David López-Escardó, Michelle Leger, Meritxell Antó and Daniel Marquina for their comments and suggestions.

References

1. Pawlowski J, Lejzerowicz F, Apotheloz-Perret-Gentil L, Visco J, Eiling P. 2016 Protist metabarcoding and environmental biomonitoring: time for change. Eur. J. Protistol. 55, 12–25. (doi:10.1016/j.ejop.2016.02.003)
2. Wangensteen OS, Turon X. 2016 Metabarcoding and of the genus Castrada (Typhloplanidae) and clade 2 by 19
3. Del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I. 2014 The others: our biassed perspective of eukaryotic genomes. Trends Ecol. Evol. 29, 252–259. (doi:10.1016/j.tree.2014.03.006)
4. Arroyo AS, López-Escardó D, de Vargas C, Ruiz-Trillo I. 2016 Hidden diversity of Acoelomorpha revealed through metabarcoding. Biol. Lett. 12, 20160674. (doi:10.1098/rsbl.2016.0674)
5. Baguña J. 1999 Acoel flatworms: earliest extant animal, plant, and fungal kingdoms. PlaSt Biol. 10, e1001419. (doi:10.1371/journal.pbio.1001419)
6. Arroyo AS, López-Escardó, Michelle Leger, Meritxell Antó and Daniel Marquina for their comments and suggestions.
7. Dunn CW. 2012 What is regeneration, and improves resolution of the animal tree of life. Nature 452, 745–749. (doi:10.1038/nature06614)
8. Zhang Z-Q. 2011 Animal biodiversity: an introduction to higher-level classification and taxonomic richness. Zootaxa 3148, 7–12. (doi:10.11646/zootaxa.3148.1.3)
9. Carznaza S, Baguiti J, Riutort M. 1997 Are the Platyhelminthes a monophyletic primitive group? An assessment using 18s rDNA sequences. Mol. Biol. Evol. 14, 485–497. (doi:10.1093/oxfordjournals.molbev.a025785)
10. Ruiz-Trillo I, Riutort M, Littlewood DT, Herniou EA, Baguña J. 1999 Acrol flatworms: earliest extant bilaterian Metazoa, not members of Platyhelminthes. Science 283, 1919–1923. (doi:10.1126/science.283.5409.1919)
11. Dunn CW et al. 2008 Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749. (doi:10.1038/nature06614)
12. Sánchez Alvarado A. 2012 What is regeneration, and why look to planarians for answers? BMC Biol. 10, 88. (doi:10.1186/1741-7007-10-88)
13. Saló E, Baguñà J. 2002 Regeneration in planarians and other worms: new findings, new tools, and
