Coming of age for COI metabarcoding of whole organism community DNA: towards bioinformatic harmonisation

Creedy T.J.[1], Andújar C.[2], Meramveliotakis E.[3], Noguerales V.[2,3], Overcast I.[4], Papadopoulou A.[3], Morlon H.[4], Vogler A.P.[1,5], Emerson B.C.[2] & Arribas P.[2]

[1] Department of Life Sciences, Natural History Museum, Cromwell Road, London SW5 7BD

[2] Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), 38206, S.C. La Laguna, Spain

[3] Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus

[4] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, Paris, France

[5] Department of Life Sciences, Imperial College London Silwood Park Campus, Ascot

Abstract

Metabarcoding of DNA extracted from community samples of whole organisms (whole organism community DNA, wocDNA) is increasingly being applied to terrestrial, marine and freshwater metazoan communities to provide rapid, accurate and high resolution data for novel molecular ecology research. The growth of this field has been accompanied by considerable development that builds on microbial metabarcoding methods to develop appropriate and efficient sampling and laboratory protocols for whole organism metazoan communities. However, considerably less attention has focused on ensuring bioinformatic methods are adapted and applied comprehensively in wocDNA metabarcoding. In this study we examined over 600 papers and identified 111 studies that performed COI metabarcoding of wocDNA. We then systematically reviewed the bioinformatic methods employed by these papers to identify the state-of-the-art. Our results show that the increasing use of wocDNA COI metabarcoding for metazoan diversity is characterised by a clear absence of bioinformatic harmonisation, and the temporal trends show little change in this situation. The reviewed literature showed (i) high heterogeneity across pipelines, tasks and tools used, (ii) limited or no adaptation of bioinformatic procedures to the nature of the COI fragment, and (iii) a worrying underreporting of tasks, software and parameters. Based upon these findings we propose a set of recommendations that we think the wocDNA metabarcoding community should consider to ensure that bioinformatic methods are appropriate, comprehensive and comparable. We believe that adhering to these recommendations will improve the long-term integrative potential of wocDNA COI metabarcoding for biodiversity science.

Keywords: *metabarcoding, COI barcode, animal communities, high-throughput sequencing, bioinformatics, community ecology*

Introduction

Metabarcoding of DNA extracted from community samples of whole organisms (whole organism community DNA, wocDNA) is a reliable and cost-efficient tool to study the biodiversity of metazoan communities (Bush et al., 2019; Ji et al., 2013; Porter & Hajibabaei, 2018). This approach, which has also been referred to as community DNA (e.g. Andújar et al., 2018b; Deiner et al., 2017) or bulk sample DNA (e.g. Braukmann et al., 2019; Yu et al., 2012) metabarcoding, primarily differs from other approaches such as eDNA (environmental or extra-organismal DNA; Taberlet et al., 2012) or iDNA (vertebrate DNA ingested by invertebrates; Schnell et al., 2012) in that the source material is a community of whole organisms collected through direct trapping or collection (e.g. malaise traps Ji et al., 2013, canopy fogging Creedy et al., 2019) or separated from an environmental sample (e.g. from soil Arribas et al. 2016 or water Suter et al., 2020). As a consequence, compared with eDNA and iDNA, wocDNA samples are characterised by (i) a comparatively low level of DNA degradation in the target species, (ii) a low proportion of non-target species, and (iii) the possibility for complementing, refining and/or validating metabarcoding-derived community data against other conventional morphological and molecular methods.

Metabarcoding of wocDNA samples is increasingly employed in community ecology, evolutionary ecology, biogeography, conservation biology, environmental management, and policy and decision-making (e.g. Bush et al., 2020; deWaard et al., 2019; Leese et al., 2018). Metazoan wocDNA metabarcoding has been adapted from pioneering approaches developed to inventory and characterise microbial diversity (e.g. Gilbert et al., 2010; Sogin et al., 2006). The majority of these adaptations have focused on sampling, and molecular laboratory steps, including adapted protocols to (i) sample, separate, enrich and/or clean animal wocDNA samples (Creedy et al., 2019; Fonseca et al., 2010, 2011), (ii) perform wocDNA extractions (Marquina et al., 2019; Nielsen et al., 2019), (iii) design and evaluate primers (Braukmann et al., 2019;

Elbrecht & Leese, 2017, Elbrecht et al. 2019), optimise amplification (Krehenwinkel et al., 2017) and prepare libraries (Yang et al., 2020). There is a growing consensus on the use of the mitochondrial cytochrome oxidase subunit I (COI) barcode, rather than other markers widely used for metabarcoding of non-metazoan communities, as the standard for wocDNA metabarcoding due to the range of COI primers with demonstrated efficiency (Braukmann et al., 2019; Elbrecht & Leese, 2017), and the potential of COI to improve the utility, resolution and reliability of wocDNA metabarcoding data (Andújar et al., 2018a; Turon et al., 2020).

However, in contrast to these advances in sampling and molecular processing, there has been limited effort to review and evaluate how bioinformatic processing has been adapted to metazoan wocDNA samples and the COI barcode, nor to examine consistency in bioinformatic approaches across the field. Broadly, bioinformatic tasks involve the computational cleaning, filtering and analysis of raw sequence data to produce biodiversity data comprising taxonomic units and their incidence across samples, implemented in a particular order (a ‘pipeline’). There are a wide array of software tools available for performing different bioinformatic tasks, from standalone tools to catch-all software packages (e.g. *OBItools* Boyer et al., 2016; *QIIME* Caporaso et al., 2010; *USEARCH/UPARSE* Edgar, 2013; and its open-source derivative *VSEARCH* Rognes et al., 2016). This software has been largely developed for metabarcode loci other than the COI region, with very few tools explicitly developed for protein coding metabarcodes (although see Andújar et al., 2021; Nugent et al., 2020; Ramirez-Gonzalez et al., 2013). To fully capitalise on the COI barcode for metabarcoding, bioinformatics should be specifically tailored to its evolutionary properties, such as the ability to interrogate the amino acid translation, and accounting for established patterns of sequence variation in protein coding genes for strict filtering. Additionally, metabarcoding employs a number of key bioinformatic tasks for which multiple alternative algorithms have been developed (e.g. denoising algorithms), with considerable variation in outcomes depending on parameters and thresholds applied.

The structure of a bioinformatic metabarcoding pipeline will depend strongly on the research aim, amplification and sequencing protocols, target locus, and target biodiversity fraction. The diversity of bioinformatic tasks and the software approaches to implement them is of course beneficial for designing appropriate pipelines, but such heterogeneity may also restrict integrated, standardised and synergistic growth in the field. As metazoan wocDNA metabarcoding becomes more accessible to researchers from a range of fields and backgrounds, harmonisation of bioinformatic approaches is important to ensure (i) high-quality, reproducible data amenable to qualitative or quantitative reviews and meta-analysis across studies, and (ii) a reliable, consistent methodology for wider implementation, development and expansion of wocDNA metabarcoding. We consider harmonisation not to mean strict prescription of the tasks and software to use, nor their order. Instead a harmonised field would recognise the diversity of approaches available, while recording key steps and establishing the effects of parameter choice on the outcome of metabarcoding studies. This approach could be enabled by the adoption of universal aligned standards for data generation and processing, while allowing for flexibility in implementation to adapt to varying research goals and take advantage of novel methodological development.

Harmonisation requires comprehensive examination of current practice to understand the aims and approaches of prior work, and a synthesis of the successes and failures in past implementations for the purposes of elaborating a framework to guide future research. Therefore it is our aim to summarise the state of the art for bioinformatic processing of metazoan wocDNA COI metabarcoding, and in doing so assess the potential for harmonisation. To this end, we performed a systematic review of peer-reviewed studies, collating information on the different bioinformatic pipelines, tasks and tools used in wocDNA COI metabarcoding in >100 recent studies (2011-2020). We use this data to (i) describe the diversity, heterogeneity and reproducibility of the bioinformatic procedures followed, (ii) identify the extent to which these

procedures are compatible with the evolutionary properties of the COI marker, and (iii) identify the key bioinformatic tasks, provide a framework for successful metabarcoding bioinformatics and make recommendations towards harmonised bioinformatic procedures for metazoan wocDNA COI metabarcoding.

Materials and Methods

Bibliographic search and screening

We focused this work on studies using whole organism community DNA (wocDNA) metabarcoding. In general, we define wocDNA samples as those where the target organisms were: (i) likely alive at the time of sampling, (ii) present as a largely complete specimen, and (iii) potentially identifiable using classical methods of morphological analysis. We exclude eDNA and iDNA metabarcoding due to the potentially different bioinformatic processing needs associated with these samples. In particular, eDNA and iDNA bioinformatic methods need to accommodate degraded DNA and a potentially high proportion of non-target reads. Furthermore, in many cases wocDNA metabarcoding is directly comparable to direct observation of specimens and conventional methods of taxonomic assignment not available for eDNA metabarcoding (Ji et al. 2013, Aylagas et al. 2016). This allows for more objective stringency thresholds in bioinformatic filtering and delimitation of operational taxonomic units (OTU).

We conducted a systematic search of peer-reviewed studies in the Web of Science (WOS) Core Collection (Science Citation Index Expanded, 1900-present) on 3rd November 2020, using the search “TS = (metabarcoding) NOT TS = (*micro* OR *bacteria* OR *myco* OR *archaea* OR fungi OR plant OR eDNA OR environmental DNA)”. These search parameters were selected in order to obtain a comprehensive set of wocDNA metabarcoding studies limited to Metazoa.

The systematic search resulted in 692 records, which were screened to to select only those studies that: (i) amplified some portion of the standard COI barcode “Folmer" region (Folmer et al., 1994), (ii) fit our definition of wocDNA samples, comprising mixtures of organisms extracted from the substrate, and (iii) provided a characterisation of metazoan communities. Studies targeting extra-organismal DNA (i.e. eDNA, iDNA) were excluded. We included studies of experimental mock communities composed of mixtures of DNA extracted from individual specimens or mixtures of specimens, and we also included studies where the target organisms remained partially or completely within an environmental substrate upon which DNA extraction was performed (e.g. parasites within a host, arthropods within soil), if the principal target was the whole organism community DNA. After reviewing the final set of filtered papers, 24 additional papers fitting the selection criteria but not present in the systematic WOS search were also included. A total of 111 articles constituted the set of core papers for subsequent assessment (see Table S2 for a complete list).

The core papers

All papers were systematically processed to record (i) the research aim and type of samples analysed; (ii) the bioinformatic tasks and pipelines implemented; and (iii) the software tools used and the reproducibility of the bioinformatic procedures employed. A detailed description of this process is provided in the Supplementary Methods and is summarised as follows.

The research aim was categorised according to whether the focus was (i) the comparison of molecular and/or bioinformatic procedures for metabarcoding, (ii) a proof-of-concept or feasibility study into the success of metabarcoding for uncovering accurate community data in the taxon/community/biome studied, or (iii) principally the study of ecological patterns and processes. We recorded whether the metabarcoded communities were sampled from marine,

7

freshwater, terrestrial biomes or from a host species, and finally if the targets were invertebrates or vertebrates.

Subsequently, the bioinformatic procedures for each paper were systematically parsed to identify the different tasks implemented, i.e. specific bioinformatic actions with a clearly-defined purpose and performed by a single tool. A total of 30 distinct bioinformatic tasks were identified starting from initial procedures on raw sequencing files through to the generation of community tables (see Table 1 for a description of each task). We focused solely on bioinformatic tasks that were presented as necessary for the generation of information about the occurrence or incidence of taxonomic units in the sampled communities (i.e. community data), and the taxonomic identity of these units. For example, we did not record any steps performing phylogenetics with a final OTU set, although we recorded steps where phylogeny-based methods were used as part of OTU delimitation and filtering. Similarly, we recorded tasks that performed filtering of community data for the purposes of removing OTUs or OTU records arising from erroneous sequences or from cross-talk/contamination (Edgar, 2018), but we did not record tasks that filtered community data for the purposes of statistical correction, such as normalisation or rarefaction.

Once the different tasks implemented by each article were identified, the pipeline used (i.e. the specific sequence of tasks in a particular order), was also recorded based on the order in which the different tasks were mentioned in the text, figures, supplementary material and/or cited papers. Where multiple mutually exclusive tasks were employed for the purposes of comparison of pipelines, we recorded that pipeline that the authors concluded to be empirically superior, or from which the authors used the output data for subsequent analysis.

For each of the bioinformatic tasks identified across the papers, we calculated (i) the number of papers implementing the task, (ii) the task's relative position within the pipelines, (iii) the

8

information reported on the software, version and parameters used, and (iv) the homogeneity in the software tools used to implement the task. We assessed homogeneity by calculating two indices, the *software homogeneity rate* and the *software dominance rate* (see Fig 5). Finally, we also summarised temporal trends in both the reporting and software heterogeneity of each task.

9

Results and Discussion

Diversity of bioinformatic methods

The 111 selected papers were published in 36 different journals with a broad focus on ecology and molecular ecology. There has been a steady increase in the number of papers published in this domain over time (Fig. 1). The earliest year of publication was 2011, but 77% of all papers were published in the last four years (2017-2020, n = 86, Fig. 1). Almost all papers studied invertebrate communities (n=108). Forty-five papers were focussed on terrestrial communities, 31 on freshwater, 30 on marine and five on parasite communities collected from a host vertebrate (see Table S2 for all the details on the core papers set).

Despite a clear trend for increased use of wocDNA COI metabarcoding, the field remains in a relatively early stage of implementation, reflected in the fact that in half of all papers (n=56, n=38 in the last four years) metabarcoding was undertaken as a proof-of-concept and the authors primarily discussed the feasibility of this method for the studied ecological system. Only 25 papers considered the sample sizes and metabarcoding procedures sufficiently rigorous to answer ecological questions. Thirty papers were primarily methodological, assessing the influences of primer choice, lab protocols and/or sequencing methods. However, within the methodological category, no paper solely studied the effect of bioinformatic pipeline choices. Indeed, only eight out of the 111 papers clearly stated that they compared different bioinformatic tools for the same task, despite the use of 116 discrete pieces of software or functions in our final count. These results illustrate the timely nature of this review, highlighting the inconsistent implementation of bioinformatic methods, in contrast to the relative maturity and harmonisation of field and laboratory methodologies.

10

205 ***High heterogeneity in tasks and pipelines***

206 The variety of bioinformatics pipelines reported across the 111 papers employed 108 unique
207 pipelines, i.e. sets of bioinformatics tasks carried out in a specific sequence. Three pipelines
208 were used twice; in two of these cases, a group of authors replicated their pipeline exactly, in the
209 other case the pipeline as reported consisted solely of a single step of searching raw reads against
210 a reference set. Although some of these pipelines were similar, with minor modifications to the
211 order, or the addition/removal of a few tasks, the heterogeneity of pipelines is remarkable. There
212 was also high heterogeneity in the number of tasks implemented within each pipeline, ranging
213 between 1 and 18 tasks, with half of the articles reporting fewer than 9 distinct bioinformatic
214 tasks (Fig. 2a). There was no particular trend in the number of tasks implemented over time
215 (Fig. 2b). The order in which these tasks were implemented also differed greatly (Fig. 2c),
216 although there was a tendency for certain tasks to be performed within similar general stages
217 within pipelines, that is, read preparation-based tasks tend to be implemented at the initial steps
218 of the pipelines, followed by filtering-based tasks and data generation tasks (Fig. 3).

219 Heterogeneity in the sequence of tasks may reflect the careful design and adaptation of
220 bioinformatic procedures within each study to the type and structure of sample and sequence
221 data and/or the specific research question, rather than the simple duplication of previously
222 published pipelines. However, high heterogeneity may equally result from the omission of
223 important tasks or their inappropriate implementation within the pipelines, and so result in low
224 comparability, integration and replication across studies. One clear example of this is associated
225 with the *Filtering* tasks of removal of erroneous sequence reads. Denoising (i.e. the removal of
226 sequencing errors based on models of error frequency parameterised by between-sequence
227 similarity, error sensitivity and/or relative frequency), was employed in just 18 studies and its
228 relative position within the pipelines was highly variable (see Table 1 and Fig. 3). While some
229 sequencing errors will be disregarded during OTU clustering, failure to incorporate denoising

can lead to false OTUs and thus OTU inflation (Shum & Palumbi, 2021) Furthermore, the trend towards examining haplotypic variation in metazoan wocDNA metabarcoding through use of amplicon sequence variants (ASVs, Callahan et al., 2017) requires minimising the number of spurious sequences, relying on stringent filtering such as denoising. Similarly, filtering to remove sequences with low copy number (that are often considered highly likely to be erroneous) was reported in only half (n=57) of the studies, despite being generally recommended (Calderón-Sanou et al., 2020; Ficetola et al., 2017) and a critical step for reducing spurious sequences surviving denoising including nuclear mitochondrial (NUMT, Lopez et al., 1994) copies (Andújar et al., 2021). It should be noted that while many task absences are cases of under-implementation, some may also be underreporting (see below).

Infrequent adaptation of pipelines to COI

The COI locus differs from many other metabarcoding loci (e.g. 18S, 16S, 12S, ITS) in that it is a protein coding gene, imparting strict expectations of amplicon sequence read properties that can be exploited in metabarcoding bioinformatics (Andújar et al., 2018b). However, the adaptation of pipelines to this fragment are in general rarely implemented in the papers of the core set. For example, only 22 papers (20%) used amino acid translations to identify erroneous sequences (“translation filtering”), using 11 different software tools for the task. The reason for low implementation of translation filtering is likely twofold; first, none of the major metabarcoding software packages include functions for translation filtering, and second, there is no standard straightforward command line software for undertaking this task. Those papers that carry out translation filtering do so by using one of three main approaches: (i) sequences are viewed and translated in a GUI application such as Geneious (https://www.geneious.com) or MEGA (Kumar et al., 2018), and those with stop codons manually removed, (ii) sequences are

12

processed through a custom script, some of which are available on github but none of which are used by research groups separate from the author, and (iii) sequences are aligned against references using MACSE (Ranwez et al., 2011) and those containing indels or stop codons are removed. The first option is time consuming and prone to human error, and custom scripts are challenging to document and maintain for a wider number of users. While MACSE is the most frequent single approach, it is computationally efficient only for small datasets. There may be some potential in the recent *coil* R package (Nugent et al., 2020) that uses Hidden Markov Models to identify and filter translation-based errors and appears to scale well to large datasets, although the R implementation presents a slight barrier to efficient inclusion in pipelines. Furthermore, the majority of translation filtering approaches are based solely on removing stop codons, while there may be other potential avenues for filtering based on amino acid translation. The extent to which expectations for protein structural properties can be applied to metabarcoding sequences for filtering other non-synonymous errors has been underexplored (but see Turon et al., 2020).

In addition to the potential of amino acid translation, the protein coding nature of COI leads to relatively stricter expectations of amplicon length. However, only half (n=54) of papers reported using length filtering, despite this being a relatively trivial procedure and with functions available in all metabarcoding software packages and as options in many more software tools. There may be some underreporting here; given the implementation of a length filtering parameter in many software tools that have a different primary purpose, authors may not have explicitly reported that length thresholds had been applied as part of a different procedure (note that we recorded when a single tool was reported to have fulfilled multiple tasks). Despite length filtering being widely available, and the relative algorithmic simplicity of implementation, there are no length filtering tools that allow for specification of thresholds outside of a simple minimum-maximum range, despite the internal barcode region of protein coding genes generally

13

being expected to vary in length only by multiples of 3 bases. While trivial to implement this programmatically for an experienced bioinformatician, this lack of straightforward user-friendly availability presents a barrier to appropriate threshold implementation by those with less experience.

Severe underreporting and increasing heterogeneity in the tools used for bioinformatic tasks

Of the 30 bioinformatic tasks identified (see Table 1 for a description of the tasks), only 11 were implemented in more than half of the papers (n<55) (Fig. 3). Quality filtering (n=92) and OTU delimitation (n=89) were the tasks most reported. Some of the less reported tasks were those associated with uncommon bioinformatic requirements of metabarcoding data, such as assembly or degapping; others have become redundant with modern computational power, such as preclustering. Low reporting of such tasks is likely an accurate reflection of rare implementation; however, there are many other tasks that are fundamental in metabarcoding bioinformatics but are poorly reported. For example, primer trimming was only reported by just over half of the papers (n=67), yet is a completely necessary step. Similarly, adapter trimming was underreported (n=21); while it is likely that in the majority of cases this is implemented by sequencing facilities prior to the authors receiving data, its reporting, including parameters and tools used, is fundamental to verify stringency of the read preparation procedures. The mapping of by-sample reads to OTUs was reported by only one third (n=30) of the papers that employed OTU delimitation, despite this being a necessary step for the production of ecological data for downstream analysis. Furthermore, OTU mapping is not a trivial step; the level of filtering/processing performed on the reads used for mapping (as opposed to filtering/processing performed on the sequences used for OTU delimitation), and the similarity threshold and tie-breaking algorithm employed to assign reads to OTU clusters could all substantially affect the

14

community data generated. The accurate reporting of this step is important to assess the validity of a pipeline, its comparability across studies, and/or its ability to be reproduced.

In addition to the clear underreporting of tasks within the pipelines as discussed above, the reporting of the bioinformatic tools and parameters used for those tasks cited in the papers was also very poor (Table 1). Only 21 of the 111 papers reported software name, version and parameters used for all of the bioinformatic tasks implemented, and 25 failed on all three counts (Fig. 4a). When considering the degree of underreporting by task (Fig. 4b), the most underreported software were used for some of the most perfunctory tasks (e.g. frequency filtering, length filtering, dereplication) that can be easily reproduced using many equivalent tools. Nonetheless, there remains relatively widespread underreporting, and this has remained unchanged over time (Fig. 5b).

Within the reported software, we identified 93 software tools used in metabarcoding bioinformatic pipelines (Table S3), of which 27% (25) were software ‘packages’. When taking into account distinct functions within packages, a total of 169 unique tools were recorded, however, this is likely an inaccurate picture given low reporting rates of functions used within software packages across all steps. There is a clear increase in the number of different software and software functions employed across all papers over time (Fig. 5a). Examining the diversity of software used within tasks over time, controlling for the number of papers published, there is limited improvement in homogeneity and a decrease in dominance of software (Figs 5c and 5d). Given that the number of metabarcoding publications is increasing year-on-year, there is thus a concomitant increase in the diversity of software used for a given task, and previously commonly used software are being used less (Figs 5c and 5d). These trends reflect that while new software tools are constantly being made available for metabarcoding, uptake is not consistent across the field and while some researchers use more recent tools, many researchers continue to use older methods, diversifying the field.

15

Toward a bioinformatic harmonisation of COI metabarcoding for metazoan wocDNA samples

Our results show that the increasing use of wocDNA COI metabarcoding for metazoan diversity is characterised by a clear absence of bioinformatic harmonisation, and the temporal trends show little change in this situation. The reviewed literature showed (i) high heterogeneity across pipelines, tasks and tools used, (ii) limited or no adaptation of bioinformatic procedures to the nature of the COI fragment, and (iii) a worrying underreporting of tasks, software and parameters.

The development of metabarcoding as a method for community ecology began with microbial studies over a decade ago, which have revealed the extensive diversity of bacteria and archaea on our planet and demonstrated the potential of metabarcoding for global biodiversity syntheses (Bates et al., 2013; Thompson et al., 2017). Although the integration and meta-analysis of microbial community data from independent studies is still challenging (e.g. Ramirez-Gonzalez et al., 2013), the success of international consortia such as the Earth Microbiome Project (EMP, Gilbert et al., 2010, 2014) has promoted the development of a harmonised framework for data generation and analyses within microbial eDNA research (see e.g. Tedersoo et al., 2015).

Through the adaptation of the microbial metabarcoding method to wocDNA samples, specific protocols to sample, sort and enrich community samples for wocDNA metabarcoding have been developed, targeting different taxonomic fractions and types of samples (e.g., Andújar et al., 2018a; Arribas et al., 2016; Creedy et al., 2019; Elbrecht & Leese, 2017; Fonseca et al., 2010; Yu et al., 2012). Additionally, recent efforts to adapt and optimise existing methods are increasing efficiency and versatility, for example through non-destructive DNA extraction techniques that retain specimens for morphological vouchering (Marquina et al., 2019; Nielsen

16

et al., 2019), or library preparation techniques tailored to metazoan samples (Yang et al., 2020). Although wocDNA COI metabarcoding remains in an expansive phase of development, standardisation in field and laboratory methods are emerging. This is in part boosted by collaborative initiatives such as the BIOSCAN initiative and its regional extensions (e.g. BIOALPHA), the Kruger Malaise Program, SITE-100, the Insect Biome Atlas Project, LIFEPLAN, and iBioGen (Arribas et al., 2021).

In contrast, there has been little advance in the development and validation of best practices associated with the bioinformatics processing of wocDNA COI metabarcoding data (but see Yang et al., 2020 for error reduction). Outside of taxonomic assignment, discussion of customising or parameterising tools for the purposes of working with wocDNA COI metabarcoding is very rare, with most papers simply reporting using tools with default settings. Our review has revealed heterogeneity in the number of tasks, the order of these within pipelines, and the tools used to implement them, along with a lack of even basic adaptations to the COI metabarcode for most of the papers. The majority of available software and resources for metabarcoding bioinformatics are still those that have been developed around the 16S rRNA gene (the primary target for microbiome metabarcoding), including the most popular software packages (e.g. USEARCH) and sets of wrapper scripts (e.g. QIIME, OBItools). While in many cases these methods may carry over to COI without issue, we observe very few studies that report consideration or analysis that assesses or validates the suitability of software choices for COI. These issues suggest that the expansion of wocDNA COI metabarcoding is proceeding at a pace and manner that could lose sight of or simply ignore the challenges inherent in producing high-quality data and reproducible methods (Baker et al., 2016; Zinger et al., 2019), and lose out on the potential for exploiting the benefits of the COI marker for wocDNA metabarcoding of Metazoa.

17

DNA metabarcoding has broad multidisciplinary potential, as demonstrated by the expansion in use of metazoan wocDNA COI metabarcoding among users from very diverse backgrounds. The diversity of applications of metabarcoding requires the concomitant bioinformatic techniques to be flexible and adaptable, and the field remains under active development. Thus it would not be productive to attempt to prescribe pipelines, tasks or even software tools in the name of standardisation, as there is no one-size-fits-all approach in metabarcoding. However, some degree of harmonisation is required to ensure quality, reproducibility and potential integration in metastudies (Tedersoo et al., 2015). Additionally, the absence of a harmonised framework of bioinformatic processing can act as a barrier for potential new users (Liu et al., 2020), hampering the growth of the field. To these ends, we thus propose a set of recommendations that we believe all researchers in the field should consider when designing and reporting their wocDNA COI metabarcoding bioinformatics pipeline, with the hope that they will catalyse harmonised implementation.

Fully report all tasks, software, software versions and parameters used, even if just the defaults. Our results show that underreporting is a recurrent problem. Comprehensive reporting of the tasks, pipelines and software used is essential for further integrating results in future reviews or meta-analyses (Tedersoo et al., 2015). Furthermore, care should be taken to report not just the name of the software package, but also the exact function, and if wrapper scripts are used then the underlying functions should be reported. Considering the trade-off with current constrictions for manuscript length, this could be achieved by the inclusion of a supporting table following the STAR-METHODs philosophy (Marcus, 2016), where task reference, order within the pipeline and software used are included. Note that the task lexicon and software lists compiled in this review (see Table 1) are a very useful resource for this purpose. This reporting effort for all the wocDNA COI metabarcoding will promote rigour and robustness with an

18

intuitive, consistent framework that makes reporting easier for the author and replication easier for the reader.

Implement filtering tasks such that spurious sequences are sufficiently removed to meet the assumptions of the research question. The quality of metabarcoding results is likely to depend most on the appropriate inclusion of filtering into a pipeline (Calderón-Sanou et al., 2020; Elbrecht et al., 2018; Zinger et al., 2019), so proper implementation of filtering tasks are critical for robust and harmonised use of COI metabarcoding. In metabarcoding, real amplicon sequence variants (ASVs, Callahan et al., 2017) amplified from target genes are inherently accompanied by spurious sequences, arising from multiple sources. Indeed, taxonomic inflation is a recurring issue demonstrated in communities with known haplotype composition (Creedy et al., 2020; Elbrecht et al., 2018). This can be exacerbated for mitochondrial markers like COI, due to the co-amplification of NUMTs and other non-authentic ASVs that are missed by denoising and require stringent, optimised filtering based on read abundances such as that implemented by the *metaMATE* software (Andújar et al., 2021). To ensure quality and reproducibility, metabarcoding studies should consider implementing the six most common filtering approaches, i.e Quality, Length, Chimera, Translation, and Frequency filtering, plus Denoising. For each of these tasks, appropriate thresholds should be considered, implemented and fully reported to a level that ensures reproducibility. Given the demonstrated importance of these tasks for most wocDNA metabarcoding studies, if any are not employed by a study the omission should be explained.

Adapt pipelines to the COI fragment. Suitable adaptations include read processing and filtering steps that leverage evolutionary properties of the protein coding nature of this fragment, or determining appropriate parameters for tools originally designed for other DNA regions. Some recent advances have been made in filtering tasks (*metaMATE*, Andújar et al., 2021; *coil*, Nugent et al., 2020; entropy-based denoising, Turon et al., 2020) but further development in

19

these promising areas is essential to fully exploit the potential of the COI gene for metabarcoding. As mentioned previously, there are no tools that enable simple length filtering variation that accounts for codon-level insertion or deletion. To our knowledge there is limited work exploring the extent to which protein structure inference might allow identification of erroneous sequences: for example the *SOAPbarcode* pipeline (Liu et al., 2013) includes a script that filters sequences based on translation hydrophilicity, but this is not comprehensively documented or discussed in the associated publications. Computation of protein structural properties is relatively trivial to perform, and seems like a fertile ground for novel development of filtering tools for protein coding markers.

For each task, consider all software available and try to select the most appropriate tool(s). This can only be approached with sufficient information about available software, and to this end we include a list of all software used for each task within Table S1, and Table S3 includes links to documentation and publications. The selection of the most appropriate tool is not always straightforward, but we suggest considering (i) the extent to which the tool was designed for the intended barcode region, purpose or dataset, (ii) the detail of available documentation and explanation to ensure a tool performs as expected, (iii) the availability and flexibility of options to appropriately apply the tool, (iv) the frequency of use of a tool in other studies with similar research aims, and (v) all else being equal, the simplest approach. Ideally, where multiple approaches exist, reasonable comparison between key methods should take place to fully understand the potential variation in conclusions that might arise from different bioinformatic choices, and the results of these comparisons should be reported. This is particularly the case when considering alternative, conceptually distinct algorithms for more bioinformatically complex tasks, such as denoising and OTU delimitation. The development of software packages and open access platforms integrating a catalogue of common bioinformatic tools, such mBRAVE (http://www.mbrave.net/), may play a fundamental role towards a proper

20

selection and harmonisation of the software used. However, software choices should be made on the basis of appropriateness and usefulness, rather than simply ease of availability and implementation due to inclusion in these packages/platforms.

Verify the compatibility of the tasks within a pipeline, especially with respect to task order. It is important to ensure that the assumptions of one task have not been violated by upstream processing; for example, UNOISE denoising employs a model of error rates in Illumina sequencing, and if errors have been removed by prior length or frequency filtering this model may not accurately fit to the data. Further, linked processes should be compatible: for instance, if OTU delimitation is based on a linkage algorithm such as swarm (Mahé et al., 2015), it is inappropriate to employ a simple similarity-based mapping method to assign reads to the resultant OTUs.

Aside from these recommendations, we also urge researchers to make data publicly available, both raw reads and final ASV and/or OTU sequences. Raw read datasets will become an invaluable resource for future work integrating many wocDNA metabarcoding studies across spatial and temporal scales, with continuing development and improvement of bioinformatic pipelines allowing for forward-compatibility of the data as analytical tools continue to evolve. Uploading ASV and/or OTU sequences, even with incomplete taxonomy, improves the capability of methods for taxonomic assignment that draw on these resources and provides fertile datasets for future development of bioinformatic methods.

Conclusions

The past decade has seen rapid growth in the development, testing and use of wocDNA COI metabarcoding. Much effort has been expended in the development of laboratory, sequencing and bioinformatic methodologies for wocDNA COI metabarcoding and for metabarcoding as a

21

whole. However, while much progress has been made towards harmonisation of lab and sequencing methods, bioinformatic processes have remained a tangle of varying software, pipelines and theoretical approaches, often suffering from underreported detail. This diversity allows for versatility, especially for those who are well-informed and experienced in bioinformatics and able to pick and choose the appropriate approach. However, choosing from the range of approaches could easily hinder new applications of metabarcoding for researchers coming from a limited bioinformatic background, and high heterogeneity can stymie the potential for future reviews and meta-analyses. Our review, which is the first evaluating the state of the art on this topic, highlights that this danger is clearly present in the field of metazoan wocDNA COI metabarcoding. The results of our assessment and the recommendations derived from it may help to improve bioinformatic harmonisation and thus the long-term integrative potential of wocDNA COI metabarcoding for biodiversity science.

Acknowledgements

This research was supported by the iBioGen project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 810729.

Conflict of interest

A.P.V. is a co-founder and scientific advisor of NatureMetrics, a private company providing commercial services in DNA-based monitoring. The authors declare that they have no other conflicts of interest.

Author contributions

22

T.J.C. and P.A, conceived the study. T.J.C. and P.A. assessed the initial paper set for inclusion, T.J.C. evaluated the methods of the core paper set and analysed the data. T.J.C. and P.A. wrote the initial draft and all co authors contributed to the final manuscript.

Data accessibility

Supporting Materials (methods, figures and tables) give the full details and methodological evaluation of the 111 publications making up the core papers.

References

Andújar, C., Arribas, P., Gray, C., Bruce, C., Woodward, G., Yu, D. W., & Vogler, A. P. (2018a). Metabarcoding of freshwater invertebrates to detect the effects of a pesticide spill. *Molecular Ecology*, *27*(1), 146–166. https://doi.org/10.1111/mec.14410

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018b). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, *27*(20), 3968–3975. https://doi.org/10.1111/mec.14844

Andújar, C., Creedy, T. J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A. J., Vogler, A. P., & Emerson, B. C. (2021). Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13337

Arribas, P., Andújar, C., Bidartondo, M. I., Bohmann, K., Coissac, É., Creer, S., deWaard, J. R., Elbrecht, V., Ficetola, G. F., Goberna, M., Kennedy, S., Krehenwinkel, H., Leese, F., Novotny, V., Ronquist, F., Yu, D. W., Zinger, L., Creedy, T. J., Meramveliotakis, E., … Emerson, B. C. (2021). Connecting high-throughput biodiversity inventories: Opportunities for a site-based genomic framework for global integration and synthesis. *Molecular Ecology*, *30*(5), 1120–1135. https://doi.org/10.1111/mec.15797

Arribas, P., Andújar, C., Hopkins, K., Shepherd, M., & Vogler, A. P. (2016). Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution*, *7*(9), 1071–1081. https://doi.org/10.1111/2041-210X.12557

Arribas, P., Andújar, C., Salces-Castellano, A., Emerson, B. C., & Vogler, A. P. (2021). The limited spatial scale of dispersal in soil arthropods revealed with whole-community haplotype-level metabarcoding. *Molecular Ecology*, *30*(1), 48–61. https://doi.org/10.1111/mec.15591

Baker, C. C. M., Bittleston, L. S., Sanders, J. G., & Pierce, N. E. (2016). Dissecting host-associated communities with DNA barcodes. *Philosophical Transactions of the Royal Society B: Biological*

23

Sciences, *371*(1702), 20150328. https://doi.org/10.1098/rstb.2015.0328

Bates, S. T., Clemente, J. C., Flores, G. E., Walters, W. A., Parfrey, L. W., Knight, R., & Fierer, N. (2013). Global biogeography of highly diverse protistan communities in soil. *The ISME Journal*, *7*(3), 652–659. https://doi.org/10.1038/ismej.2012.147

Boyer, F., Mercier, C., Bonin, A., Bras, Y. L., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., Waard, J. R. de, Sones, J. E., Zakharov, E. V., & Hebert, P. D. N. (2019). Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources*, *19*(3), 711–727. https://doi.org/10.1111/1755-0998.13008

Bush, A., Compson, Z. G., Monk, W. A., Porter, T. M., Steeves, R., Emilson, E., Gagne, N., Hajibabaei, M., Roy, M., & Baird, D. J. (2019). Studying Ecosystems With DNA Metabarcoding: Lessons From Biomonitoring of Aquatic Macroinvertebrates. *Frontiers in Ecology and Evolution*, *7*. https://doi.org/10.3389/fevo.2019.00434

Bush, A., Monk, W. A., Compson, Z. G., Peters, D. L., Porter, T. M., Shokralla, S., Wright, M. T. G., Hajibabaei, M., & Baird, D. J. (2020). DNA metabarcoding reveals metacommunity dynamics in a threatened boreal wetland wilderness. *Proceedings of the National Academy of Sciences*, *117*(15), 8539–8545. https://doi.org/10.1073/pnas.1918741117

Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L., & Thuiller, W. (2020). From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices? *Journal of Biogeography*, *47*(1), 193–206. https://doi.org/10.1111/jbi.13681

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Creedy, T. J., Ng, W. S., & Vogler, A. P. (2019). Toward accurate species-level metabarcoding of arthropod communities from the tropical forest canopy. *Ecology and Evolution*, *9*(6), 3105–3116. https://doi.org/10.1002/ece3.4839

Creedy, T. J., Norman, H., Tang, C. Q., Chin, K. Q., Andujar, C., Arribas, P., O'Connor, R. S., Carvell, C., Notton, D. G., & Vogler, A. P. (2020). A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput DNA barcoding. *Molecular Ecology Resources*, *20*(1), 40–53. https://doi.org/10.1111/1755-0998.13056

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., Vere, N. de, Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

deWaard, J. R., Levesque-Beaudin, V., deWaard, S. L., Ivanova, N. V., McKeown, J. T. A., Miskie, R., Naik, S., Perez, K. H. J., Ratnasingham, S., Sobel, C. N., Sones, J. E., Steinke, C., Telfer, A. C., Young, A. D., Young, M. R., Zakharov, E. V., & Hebert, P. D. N. (2019). Expedited assessment of terrestrial arthropod diversity by coupling Malaise traps with DNA barcoding. *Genome*, *62*(3), 85–95. https://doi.org/10.1139/gen-2018-0093

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998. https://doi.org/10.1038/nmeth.2604

Edgar, R. C. (2018). UNCROSS2: Identification of cross-talk in 16S rRNA OTU tables. *BioRxiv*, 400762. https://doi.org/10.1101/400762

Elbrecht, V., & Leese, F. (2017). Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science*, *5*. https://doi.org/10.3389/fenvs.2017.00011

Elbrecht, V., Vamos, E. E., Steinke, D., & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*, *6*, e4644. https://doi.org/10.7717/peerj.4644

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M., Zakharov, E. V., Hebert, P. D. N., & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, *7*, e7745. https://doi.org/10.7717/peerj.7745

Ficetola, G. F., Mazel, F., & Thuiller, W. (2017). Global determinants of zoogeographical boundaries. *Nature Ecology & Evolution*, *1*(4), 1–7. https://doi.org/10.1038/s41559-017-0089

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*(5), 294–299.

Fonseca, V. G., Carvalho, G. R., Sung, W., Johnson, H. F., Power, D. M., Neill, S. P., Packer, M., Blaxter, M. L., Lambshead, P. J. D., Thomas, W. K., & Creer, S. (2010). Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, *1*(1), 98. https://doi.org/10.1038/ncomms1095

Fonseca, V. G., Packer, M., Carvalho, G. R., Power, D. M., Lambshead, P. J. D., & Creer, S. (2011). Isolation of marine meiofauna from sandy sediments: From decanting to DNA extraction. *Protocol Exchange*. https://doi.org/10.1038/nprot.2010.157

Gilbert, J. A., Jansson, J. K., & Knight, R. (2014). The Earth Microbiome project: Successes and aspirations. *BMC Biology*, *12*(1), 69. https://doi.org/10.1186/s12915-014-0069-1

Gilbert, J. A., Meyer, F., Jansson, J., Gordon, J., Pace, N., Tiedje, J., Ley, R., Fierer, N., Field, D., Kyrpides, N., Glöckner, F.-O., Klenk, H.-P., Wommack, K. E., Glass, E., Docherty, K., Gallery,

R., Stevens, R., & Knight, R. (2010). The Earth Microbiome Project: Meeting report of the "1st EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6th 2010. *Standards in Genomic Sciences*, *3*(3), 249. https://doi.org/10.4056/aigs.1443528

Hajibabaei, M., Spall, J. L., Shokralla, S., & van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, *12*(1), 28. https://doi.org/10.1186/1472-6785-12-28

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P. M., Woodcock, P., Edwards, F. A., Larsen, T. H., Hsu, W. W., Benedick, S., Hamer, K. C., Wilcove, D. S., Bruce, C., Wang, X., Levi, T., Lott, M., … Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, *16*(10), 1245–1257. https://doi.org/10.1111/ele.12162

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, *7*(1), 17668. https://doi.org/10.1038/s41598-017-17333-x

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., Ekrem, T., Čiampor, F., Čiamporová-Zaťovičová, Z., Costa, F. O., Duarte, S., Elbrecht, V., Fontaneto, D., Franc, A., Geiger, M. F., Hering, D., Kahlert, M., Kalamujić Stroil, B., Kelly, M., … Weigand, A. M. (2018). Chapter Two - Why We Need Sustainable Networks Bridging Countries, Disciplines, Cultures and Generations for Aquatic Biomonitoring 2.0: A Perspective Derived From the DNAqua-Net COST Action. In D. A. Bohan, A. J. Dumbrell, G. Woodward, & M. Jackson (Eds.), *Advances in Ecological Research* (Vol. 58, pp. 63–99). Academic Press. https://doi.org/10.1016/bs.aecr.2018.01.001

Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2020). A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology*, *45*(3), 373–385. https://doi.org/10.1111/een.12831

Liu, S., Li, Y., Lu, J., Su, X., Tang, M., Zhang, R., Zhou, L., Zhou, C., Yang, Q., Ji, Y., Yu, D. W., & Zhou, X. (2013). SOAPBarcode: Revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution*, *4*(12), 1142–1150. https://doi.org/10.1111/2041-210X.12120

Lopez, J. V., Yuhki, N., Masuda, R., Modi, W., & O'Brien, S. J. (1994). Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution*, *39*(2), 174–190. https://doi.org/10.1007/BF00163806

Mahé, F., Rognes, T., Quince, C., Vargas, C. de, & Dunthorn, M. (2015). Swarm v2: Highly-scalable and

high-resolution amplicon clustering. *PeerJ*, *3*, e1420. https://doi.org/10.7717/peerj.1420

Marcus, E. (2016). A STAR Is Born. *Cell*, *166*(5), 1059–1060. https://doi.org/10.1016/j.cell.2016.08.021

Marquina, D., Esparza-Salas, R., Roslin, T., & Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. *Molecular Ecology Resources*, *19*(6), 1516–1530. https://doi.org/10.1111/1755-0998.13071

Nielsen, M., Gilbert, M. T. P., Pape, T., & Bohmann, K. (2019). A simplified DNA extraction protocol for unsorted bulk arthropod samples that maintains exoskeletal integrity. *Environmental DNA*, *1*(2), 144–154. https://doi.org/10.1002/edn3.16

Nugent, C. M., Elliott, T. A., Ratnasingham, S., & Adamowicz, S. J. (2020). coil: An R package for cytochrome *c* oxidase I (COI) DNA barcode data cleaning, translation, and error evaluation. *Genome*, *63*(6), 291–305. https://doi.org/10.1139/gen-2019-0206

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, *27*(2), 313–338. https://doi.org/10.1111/mec.14478

Ramirez-Gonzalez, R., Yu, D. W., Bruce, C., Heavens, D., Caccamo, M., & Emerson, B. C. (2013). PyroClean: Denoising Pyrosequences from Protein-Coding Amplicons for the Recovery of Interspecific and Intraspecific Genetic Variation. *PLOS ONE*, *8*(3), e57615. https://doi.org/10.1371/journal.pone.0057615

Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE*, *6*(9), e22594. https://doi.org/10.1371/journal.pone.0022594

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Shum, P., & Palumbi, S. R. (2021). Testing small-scale ecological gradients and intraspecific differentiation for hundreds of kelp forest species using haplotypes from metabarcoding. *Molecular Ecology*. https://doi.org/10.1111/mec.15851

Schnell, I. B., Thomsen, P. F., Wilkinson, N., Rasmussen, M., Jensen, L. R. D., Willerslev, E., Bertelsen, M. F., & Gilbert, M. T. P. (2012). Screening mammal biodiversity using DNA from leeches. *Current Biology*, *22*(8), R262–R263. https://doi.org/10.1016/j.cub.2012.02.058

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., & Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences*, *103*(32), 12115–12120. https://doi.org/10.1073/pnas.0605127103

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., & Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences*, *103*(32), 12115–12120. https://doi.org/10.1073/pnas.0605127103

27

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Tedersoo, L., Anslan, S., Bahram, M., Põlme, S., Riit, T., Liiv, I., Kõljalg, U., Kisand, V., Nilsson, H., Hildebrand, F., Bork, P., & Abarenkov, K. (2015). Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys*, *10*, 1–43. https://doi.org/10.3897/mycokeys.10.4852

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., Navas-Molina, J. A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J. T., Mirarab, S., Zech Xu, Z., Jiang, L., … Knight, R. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, *551*(7681), 457–463. https://doi.org/10.1038/nature24621

Turon, X., Antich, A., Palacín, C., Præbel, K., & Wangensteen, O. S. (2020). From metabarcoding to metaphylogeography: Separating the wheat from the chaff. *Ecological Applications*, *30*(2), e02036. https://doi.org/10.1002/eap.2036

Yang, C., Bohmann, K., Wang, X., Cai, W., Wales, N., Ding, Z., Gopalakrishnan, S., & Yu, D. W. (2020). Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *BioRxiv*, 2020.07.07.187666. https://doi.org/10.1101/2020.07.07.187666

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*(4), 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., Barba, M. D., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*(8), 1857–1862. https://doi.org/10.1111/mec.15060

28

Tables

Table 1: Table of all bioinformatic tasks performed across the core papers set. Tasks are grouped into four groups by broad purposes, and a detailed definition of each task is given along with summary statistics of the implementation of each task across the 111 papers. For a list of the software used for each task, Table S1 is an expanded version of this table.

Figures

Figure 1: Year of publication of the articles in the core papers set. Bar fills and numbers refer to the number of articles within each research aim category. Note that only articles indexed by Web of Science by 3rd November 2020 were included.

Figure 2: Bioinformatic pipelines implemented by the core papers set. A) Frequency distribution of the number of tasks by study, B) Number of tasks by study against the year of publication, with best fit regression line in blue with shaded 95% confidence intervals around the line. Slight horizontal jitter added to points to better show density. C) Network diagram of tasks and different pipeline routes through these tasks. All pipelines start and end on the respective orange nodes. All other nodes are coloured according to the four main categories of bioinformatic tasks; red for read preparation tasks, blue for sequence processing, green for filtering and purple for data generation tasks. Arrows link tasks performed consecutively, with direction of arrow showing order of tasks. Thickness of arrows shows relative frequency of pairs of consecutive tasks. Arrows coloured orange are the top 10% of consecutive task pairs by relative frequency. Note that while this illustrates a possible complete pipeline from Start to End, this “average” pipeline is not in fact performed by any of the papers assessed by this review.

Figure 3: Violin plot of standardised task position within pipelines. Increasing x-axis position denotes later placement of task within pipelines, vertical dashed lines denote 25%, 50% and 75% of the way through the pipeline respectively. Tasks are separated into task groups and ordered within task group by mean standardised pipeline position. Points denote task positions where tasks occurred too infrequently to compute density profile for violin plots. Values report the total number of papers implementing each task.

Figure 4: Plots summarising the reporting of three key aspects of bioinformatic tools (software name, version and parameters) by the core papers. A). Venn diagram shows the number of papers fully reporting each detail, i.e. giving the software used for every task reported, and giving the parameters and version for each task where software is given; 86 papers reported at

29

least one of the three details for all steps, 25 further papers failed to fully report all three details in all steps. B) Bar chart details the proportion of papers employing a specific task that failed to report the software used for that task, with longer bars denoting a greater proportion of papers not reporting software for that specific task

Figure 5: Consistency in software reporting and use over time. A) The total number of unique software functions reported across all papers for each year of publication. B) For each paper, the proportion of the total number of bioinformatic tasks for which the software used for a task was not reported. C) The software homogeneity rate is one minus the number of different software tools used for a given task in a given year, divided by the number of papers employing the task in that year, calculated only when more than one paper reported a task in a given year. A value of 1 means all papers used the same tool for a given task in a given year. D) The software dominance rate is the proportion of papers that use the most common software tool for a given task in a given year, calculated only when more than one paper reported a task in a given year. A value of 1 means all papers used the same tool for a given task in a given year. B-D) Best fit regression lines are shown in blue with shaded 95% confidence intervals around the line. Horizontal jitter added to points to illustrate density within years; C & D) colours denote different tasks, see Figure S1.

30

Task Group	Task	Description	Number papers reporting task	Number papers not reporting software	Total number of software tools	Total number of software functions	Number of papers performing manually
Read preparation	***quality control***	*Generating a report of sequence quality information from a sample or set of samples - no modification is done to data*	**19**	**0**	**4**	**4**	**0**
	adapter trimming	*Trimming of sequencing adapters*	**9**	**1**	**6**	**6**	**0**
	demultiplexing	*Separation of sequences from a mixed pool into separate pools based on the occurence of a unique set of bases (index or tag)*	**55**	**17**	**16**	**19**	**0**
	pair merging	*The assembly of mate pair reads into a single contig*	**63**	**1**	**10**	**18**	**0**
	quality trimming	*The removal of bases from either or both ends of sequences in a pool based on quality scores*	**20**	**1**	**8**	**10**	**0**
	mate pairing	*The identification and syncronisation of mate pair reads between two samples, often involving arranging reads in identical orders and/or removal of reads without a mate pair*	**3**	**0**	**3**	**3**	**0**
	primer trimming	*Trimming of PCR primers*	**66**	**8**	**15**	**17**	**0**
	reverse complementation	*Reverse complementing the sequences in a pool*	**7**	**3**	**2**	**2**	**0**
	sequence conversion	*Converting sequences from fastq to fasta*	**3**	**0**	**2**	**3**	**0**
	length trimming	*The removal of bases from either or both ends of sequences in a pool, either the removal of a fixed number of bases or the removal of a variable number of bases to reduce sequences to a standard length*	**10**	**3**	**6**	**7**	**0**
	pair concatenation	*Concatenating mate pair reads into a single contig (where reads don't overlap)*	**8**	**4**	**4**	**4**	**0**
	assembly	*The assembly of reads into contigs, applied when more than one pair of overlapping fragments have been metabarcoded*	**6**	**0**	**4**	**4**	**0**
	degapping	*Removal of gaps from sequences*	**1**	**0**	**1**	**1**	**0**
Sequence processing	***dereplication***	*The removal of duplicate reads to retain only unique sequences in a pool; often the total number of copies of a sequence is recorded in the header of the retained sequence*	**58**	**10**	**11**	**19**	**0**
	size sorting	*The sorting of a fasta file according to a size annotation in the header*	**10**	**2**	**3**	**4**	**0**
Filtering	***quality filtering***	*Removal and/or trimming of sequences from a pool based on quality information. Also often converts from fastq to fasta.*	**81**	**11**	**20**	**27**	**0**
	similarity filtering	*Removal of sequences based on similarity to an alignment, either based on sequence identity or alignment position*	**9**	**1**	**4**	**4**	**0**

	length filtering	*The removal of sequences from a pool that are less than, more than, or fall within or outside of a specified length threshold or thresholds*	**54**	**21**	**17**	**23**	**0**
	preclustering	*Reduction of sequence variation in a dataset prior to further processing - a form of denoising*	**12**	**1**	**3**	**6**	**0**
	denoising	*The removal of reads containing putative PCR or sequencing errors based on statistical assessment*	**18**	**1**	**8**	**8**	**0**
	normalisation	*A process by which the number of sequences for each of a set of samples is reduced where necessary such that the output set of samples all have the same number of sequences while maintaining the relative frequencies of OTUs*	**2**	**0**	**1**	**1**	**1**
	chimera filtering	*The filtering of putative chimeric assemblies from a pool of mate paired reads*	**63**	**4**	**6**	**16**	**1**
	translation filtering	*Removal of sequences from a set of sequence based on their translation, usually removing sequences with inframe stop codons or frameshifts due to erroneous indels or substitutions caused by sequencing errors*	**22**	**3**	**11**	**12**	**0**
	frequency filtering	*Removal of sequences based on their frequency in a pool*	**51**	**37**	**11**	**15**	**1**
	taxonomy filtering	*Removal of sequences based on an assigned taxonomy or a taxonomic classification*	**9**	**5**	**1**	**1**	**1**
	mistag filtering	*Removal of sequences based on putative tagging errors*	**3**	**1**	**1**	**1**	**0**
Data generation	***OTU delimitation***	*The grouping of a set of sequences into OTUs by some method*	**84**	**5**	**12**	**22**	**0**
	OTU mapping	*The mapping of sequences to OTUs to provide read counts for each OTU*	**30**	**3**	**7**	**11**	**0**
	uncurated taxonomic assignment	*The assignment (identification or classification) of taxonomy to OTUs using a global uncorated reference database (e.g. GenBank, BOLD)*	**55**	**2**	**11**	**13**	**0**
	reference taxonomic assignment	*The assignment (identification or classification) of taxonomy to OTUs using a purpose-built and/or specially curated reference set of sequences*	**60**	**9**	**18**	**23**	**1**

Table 1: *Table of all bioinformatic tasks performed across the core papers set. Tasks are grouped into four groups by broad purposes, and a detailed definition of each task is given along with summary statistics of the implementation of each task across the 111 papers. For a list of the software used for each task, Table S1 is an expanded version of this table.*

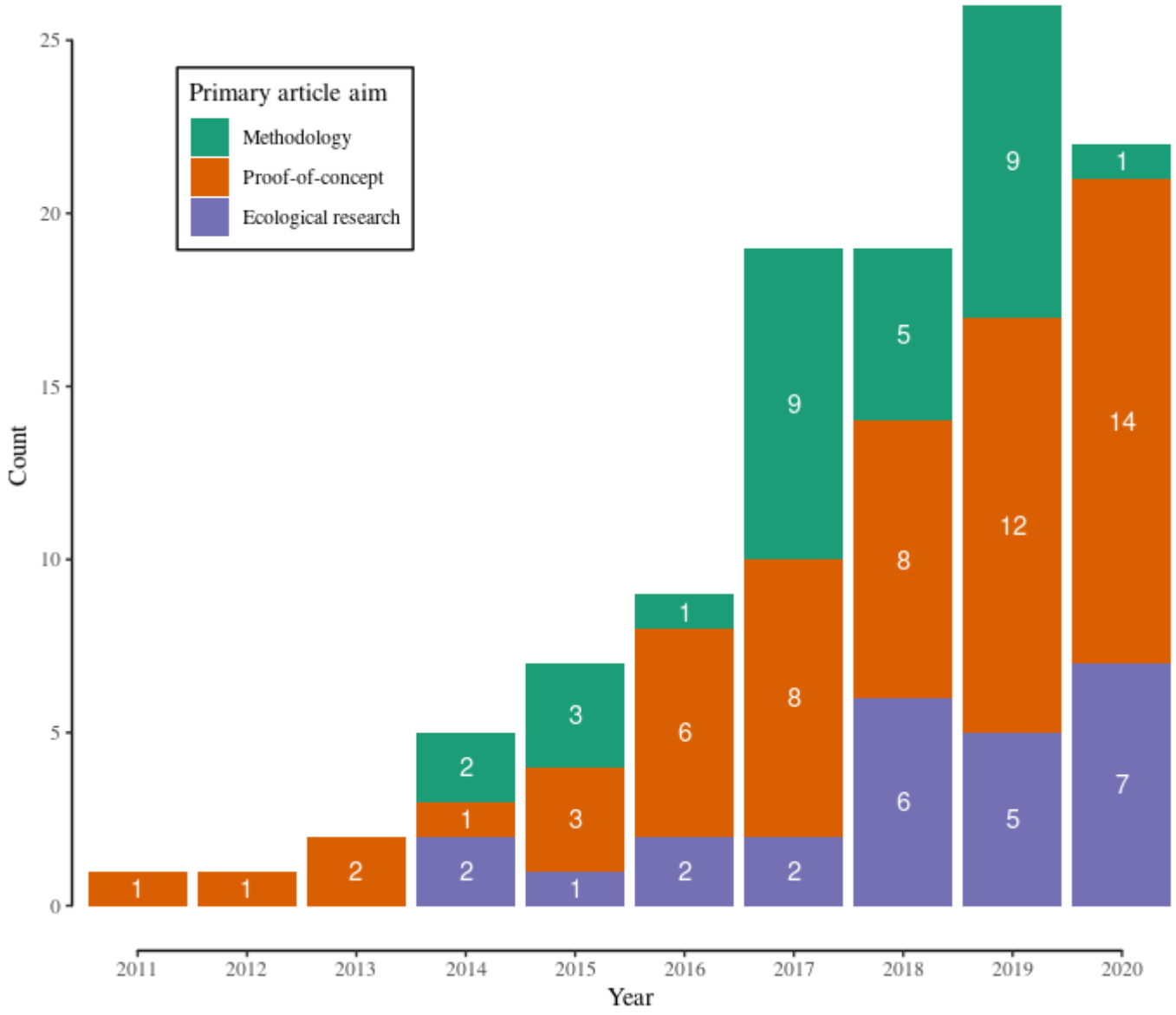

Figure 1: *Year of publication of the articles in the core papers set. Bar fills and numbers refer to the number of articles within each research aim category. Note that only articles indexed by Web of Science by 3rd November 2020 were included.*

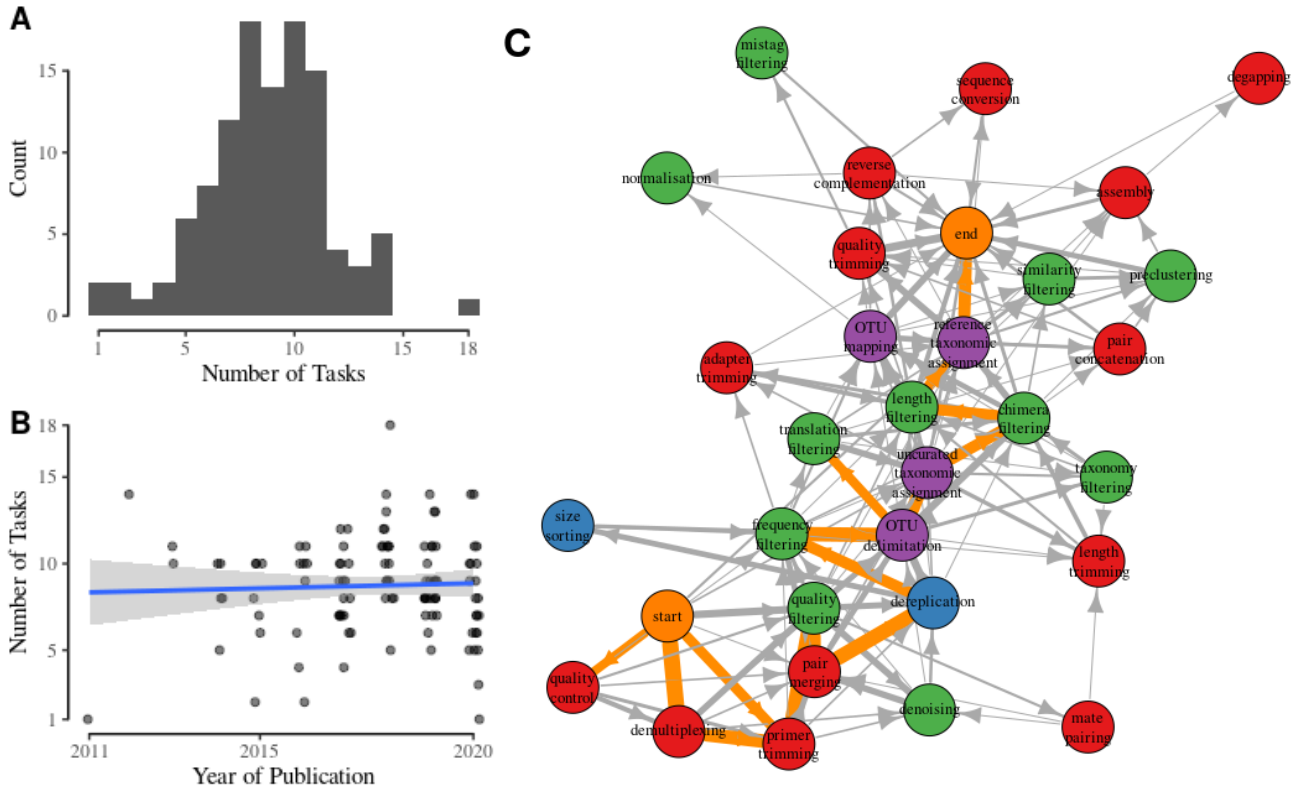

Figure 2: *Bioinformatic pipelines implemented by the core papers set. Left: A) Frequency distribution of the number of tasks by study, B) Number of tasks by study against the year of publication, with best fit regression line in blue with shaded 95% confidence intervals around the line. Slight horizontal jitter added to points to better show density. Right: C) Network diagram of tasks and different pipeline routes through these tasks. All pipelines start and end on the respective orange nodes. All other nodes are coloured according to the four main categories of bioinformatic tasks; red for read preparation tasks, blue for sequence processing, green for filtering and purple for data generation tasks. Arrows link tasks performed consecutively, with direction of arrow showing order of tasks. Thickness of arrows shows relative frequency of pairs of consecutive tasks. Arrows coloured orange are the top 10% of consecutive task pairs by relative frequency. Note that while this illustrates a possible complete pipeline from Start to End, this "average" pipeline is not in fact performed by any of the papers assessed by this review.*

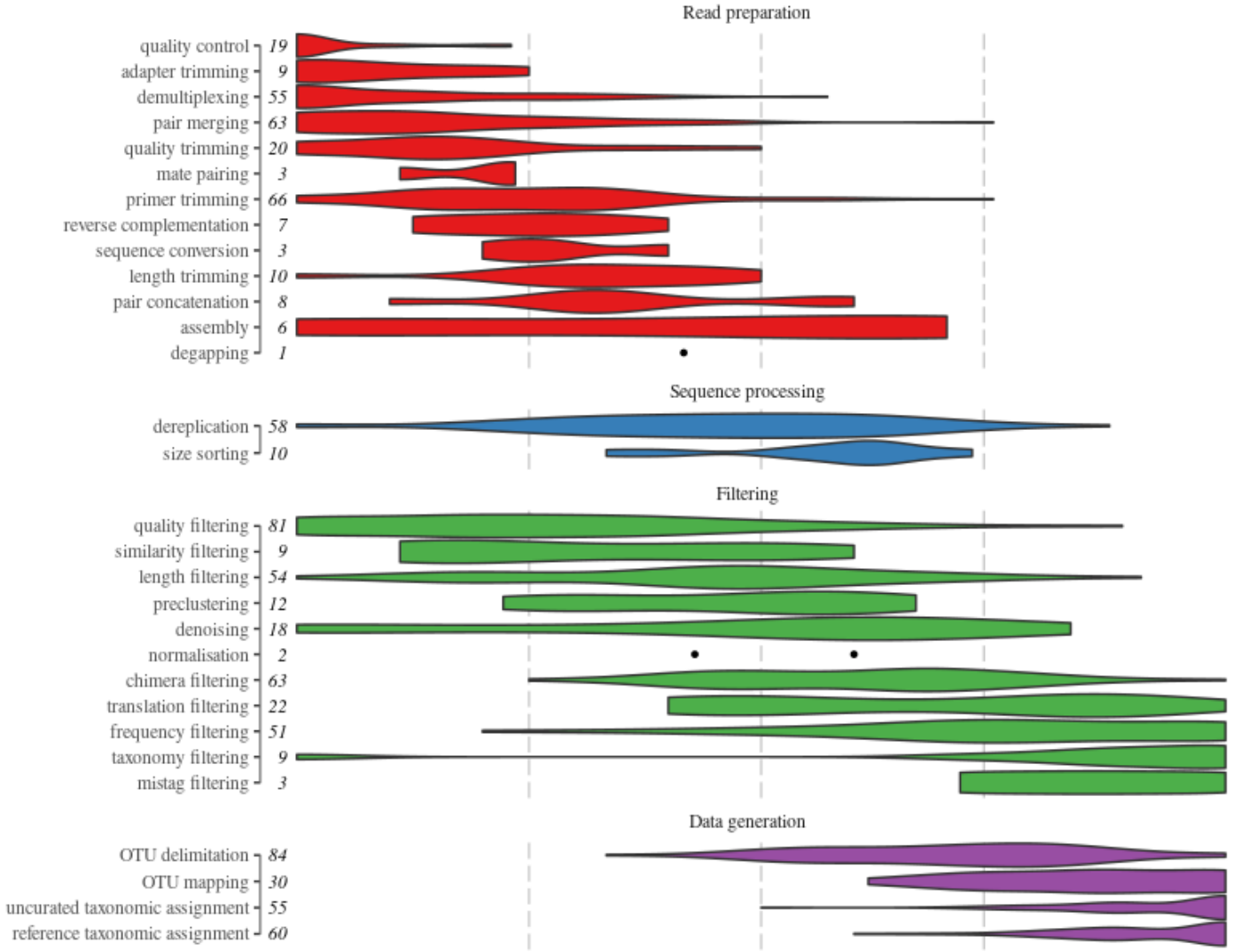

Figure 3: *Violin plot of standardised task position within pipelines. Increasing x-axis position denotes later placement of task within pipelines, vertical dashed lines denote 25%, 50% and 75% of the way through the pipeline respectively. Tasks are separated into task groups and ordered within task group by mean standardised pipeline position. Points denote task positions where tasks occurred too infrequently to compute density profile for violin plots. Values report the total number of papers implementing each task.*

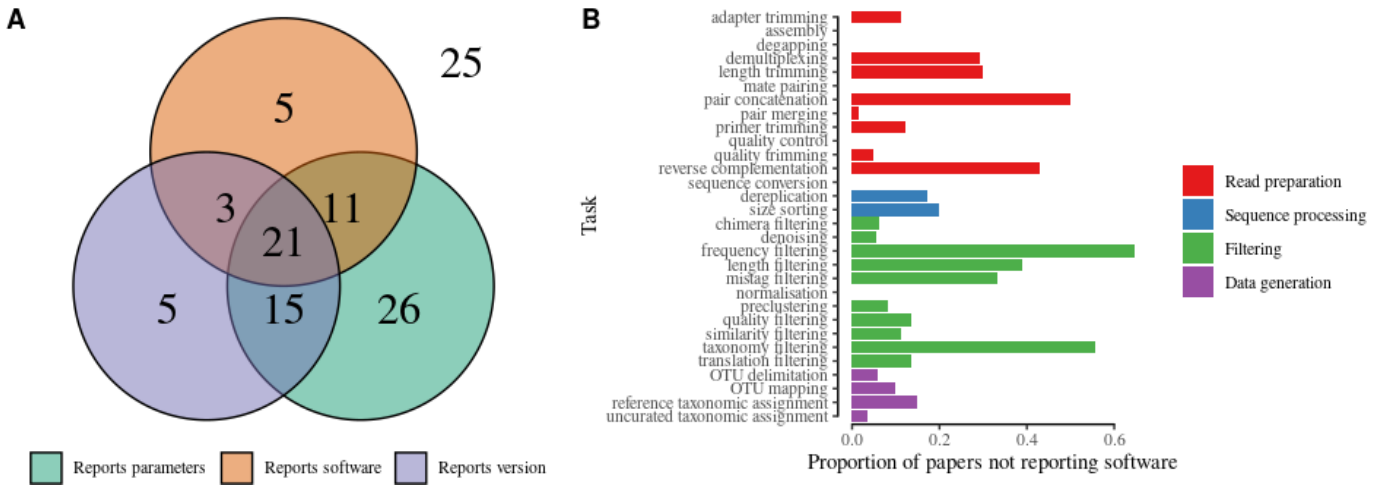

Figure 4: *Plots summarising the reporting of 3 key methodological details by papers. A) Venn diagram shows the number of papers fully reporting each detail, i.e. giving the software used for every task reported, and giving the parameters and version for each task where software is given; 86 papers reported at least one of the three details for all steps, 25 further papers failed to fully report all three details in all steps. B) Bar chart details the proportion of papers employing a specific task that failed to report the software used for that task, with longer bars denoting a greater proportion of papers not reporting software for that specific task.*

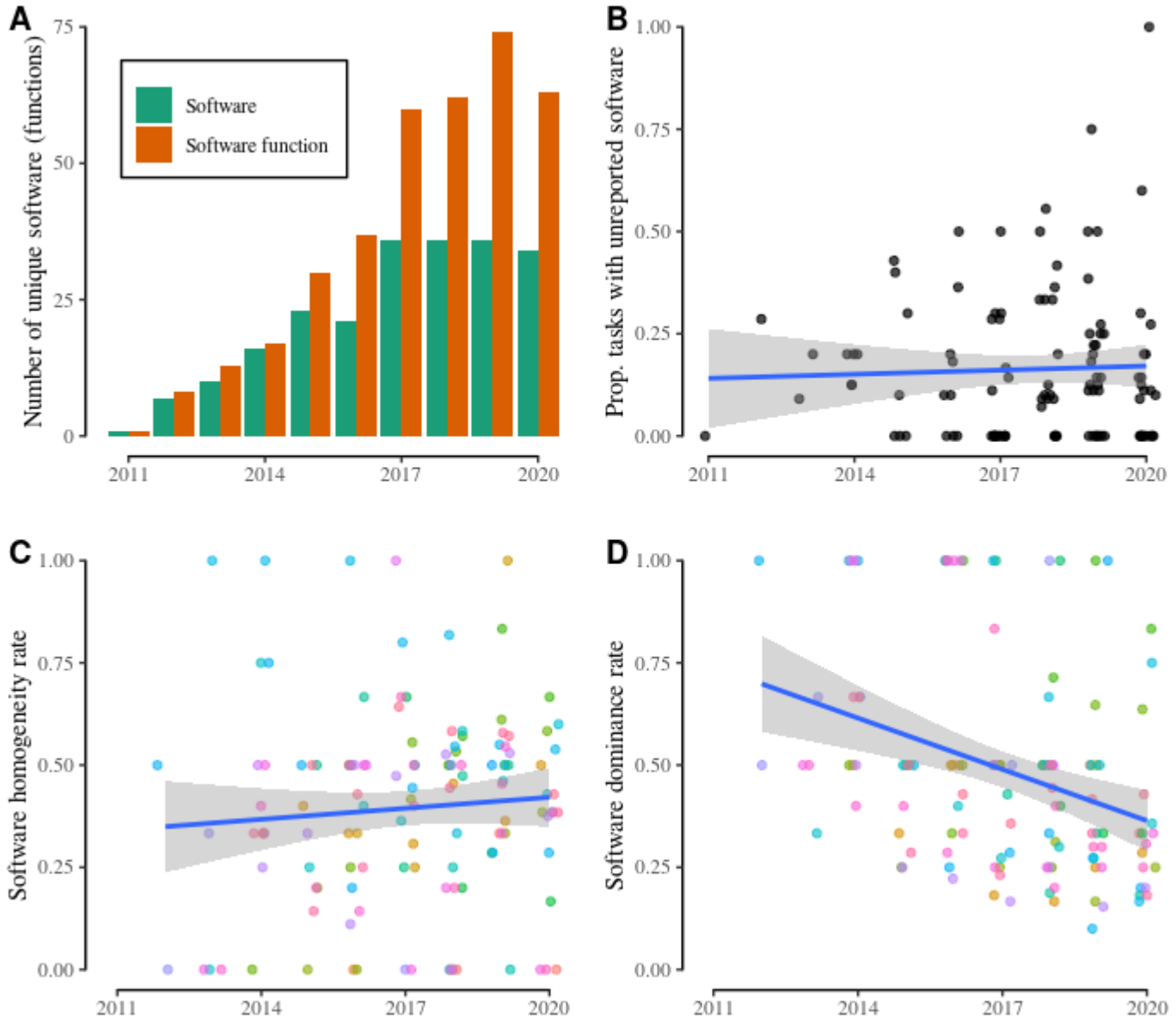

***Figure 5**: Consistency in software reporting and use over time. A) The total number of unique software functions reported across all papers for each year of publication. B) For each paper, the proportion of the total number of bioinformatic tasks for which the software used for a task was not reported. C) The software homogeneity rate is one minus the number of different software tools used for a given task in a given year, divided by the number of papers employing the task in that year, calculated only when more than one paper reported a task in a given year. A value of 1 means all papers used the same tool for a given task in a given year. D) The software dominance rate is the proportion of papers that use the most common software tool for a given task in a given year, calculated only when more than one paper reported a task in a given year. A value of 1 means all papers used the same tool for a given task in a given year. B-D) Best fit regression lines are shown in blue with shaded 95% confidence intervals around the line. Horizontal jitter added to points to illustrate density within years; C & D) colours denote different tasks, see Figure S1.*