Second-generation environmental sequencing unmasks marine metazoan biodiversity

Citation for published version:
Fonseca, VG, Carvalho, GR, Sung, W, Johnson, HF, Power, DM, Neill, SP, Packer, M, Blaxter, ML, Lambshead, PJD, Thomas, WK & Creer, S 2010, 'Second-generation environmental sequencing unmasks marine metazoan biodiversity' Nature Communications, vol. 1, 98. DOI: 10.1038/ncomms1095

Digital Object Identifier (DOI):
10.1038/ncomms1095

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Nature Communications

Publisher Rights Statement:
This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Second-generation environmental sequencing unmasks marine metazoan biodiversity

Vera G. Fonseca, Gary R. Carvalho, Way Sung, Harriet F. Johnson, Deborah M. Power, Simon P. Neill, Margaret Packer, Mark L. Blaxter, P. John D. Lambshead, W. Kelley Thomas & Simon Creer

Biodiversity is of crucial importance for ecosystem functioning, sustainability and resilience, but the magnitude and organization of marine diversity at a range of spatial and taxonomic scales are undefined. In this paper, we use second-generation sequencing to unmask putatively diverse marine metazoan biodiversity in a Scottish temperate benthic ecosystem. We show that remarkable differences in diversity occurred at microgeographical scales and refute currently accepted ecological and taxonomic paradigms of meiofaunal identity, rank abundance and concomitant understanding of trophic dynamics. Richness estimates from the current benchmarked Operational Clustering of Taxonomic units from Parallel UltraSequencing analyses are broadly aligned with those derived from morphological assessments. However, the slope of taxon rarefaction curves for many phyla remains incomplete, suggesting that the true alpha diversity is likely to exceed current perceptions. The approaches provide a rapid, objective and cost-effective taxonomic framework for exploring links between ecosystem structure and function of all hitherto intractable, but ecologically important, communities.
Biodiversity is the product of millions of years of evolution and forms the basis of the earth’s life support system, but the magnitude and relative diversity of global species richness remain unknown. On earth, there may be over 100 million species1 but fewer than two million have been formally described1. There is also a pronounced bias towards the study of larger organisms, leaving the most speciose communities that are dominated by microscopic organisms understudied. To study diverse environments dominated by small taxa, second-generation sequencing has been used for the quantification of bacteria, archaea12 and viruses5–6, but large knowledge gaps still exist regarding the organization of diversity within several eukaryotic kingdoms, including the Metazoa7. The Kingdom Metazoa, also known as Animalia, consists of multicellular heterotrophic organisms ranging from Porifera to Chordata. Contemporary phylogenetic studies routinely recover a monophyletic Bilateria—Triploblasta clade (including deep-branaching Ecdysozoa, Lophotrochozoa and Deuterostomia)8, but no consensus view exists of the precise relationships between Bilateria, and the basal groups of Porifera, Ctenophora, Placozoa, Cnidaria and Coe-lenterata8. Marine benthic metazoan communities display some of the highest α-diversity on the planet and occupy one of the largest ecosystems on earth, where only 1% of species are estimated to be known9. Benthic meiofauna (small metazoans between 45 and 500µm in size) comprises members encompassing 60% of animal phyla and represents a major part of marine biodiversity10. Dominated by nematodes and characterized by high abundances (up to 109 individuals per 1 m²) and diversity11, meiofaunal assemblages perform essential roles in marine ecosystem processes, namely, nutrient cycling, secondary production, sediment transport and mineralization12.

In this paper, we apply a metagenetic approach (that is, the large-scale analysis of taxon richness through the analysis of homologous genes) using second-generation sequencing of the 18S nuclear small subunit (nSSU) ribosomal RNA (rRNA) gene to assess simultaneously the relative levels of richness and patterns of diversity of multiple metazoan phyla across an ecological gradient in a temperate benthic ecosystem. The heterogeneous levels of accumulating taxon richness derived from the benchmarked analyses were broadly congruent with those derived from intensive morphological assessments, but MEGABLAST annotation revealed a previously unidentified phylogenetic breadth of microbial metazoa. Moreover, the finding that the largely predacious turbellarian Platyhelminthes represent a substantial proportion of benthic diversity quantifies their hitherto unrecognized ecological importance in benthic food chains. Annotated metagenetic analyses enable the objective assessment of microbial biodiversity throughout all ecosystems, facilitating understanding of linkages between microbial biodiversity and ecosystem processes.

Results
Taxon richness estimates. nSSU rRNA amplicons were generated from eight benthic samples from the low-tide zone of an estuarine beach near Prestwick on the West coast of Scotland, and from one sample from a beach in Littlehampton in the South of England. The amplicons were processed for sequencing on a Roche 454 FLX platform generating a total of 353,896 sequences that were quality filtered to 305,702 for downstream analysis (Table 1). When performing metagenetic assessments of taxon richness, it is important to cluster taxonomic units in a biologically meaningful manner, as even slight differences in similarity cut-offs and using different algorithms can result in significantly different estimates of richness. The taxon clustering comparisons between ESPRIT11 and OCTUPUS (the Operational Clustering of Taxonomic Units from Parallel UltraSequencing, supplementary software available at http://octupus.sourceforge.net/) (Fig. 1) on the subsampled beach pyrosequencing data show that ESPRIT overestimates Operational Taxonomic Units (OTUs) richness compared with OCTUPUS (between 1.1x–4.4x, over the 90–99% cut-off range). Phylogenetic14 and barcoding15 studies based on the analysis of chain-termination nSSU gene sequences suggest that intraspecific divergence at least in nematode species is low (1–2%). However, the true level of intragenomic and intraspecific variation among rRNA gene repeats is largely unknown, and genome-wide analyses reveal a dominant set of conserved sequences accompanied by rare variant sequences6. Our benchmarking exercise, performed against a reference control data set comprising 41 species15, revealed that the 96% similarity OCTUPUS clustering algorithm with accompanying chimera screening estimated taxon richness that was most closely aligned with species richness. At 96% similarity, OCTUPUS resulted in 37 operational clustered taxonomic units (OTUs), whereas at 97%, OCTUPUS resulted in 51 OTUs from the control metagenetic analysis15. Thus, although we may be underestimating richness by at least 10%, we have opted for a more conservative approach of setting a 96% identity OCTU cut-off for all subsequent numerical comparisons. At this cut-off, an OCTU is likely to (at worst) correspond to a group of related species. Following the 96% similarity OCTUPUS clustering strategy, the total number of putatively non-chimeric tag reads and OCTUs was 217,221 and 428, respectively. Before chimera screening, 1013 OCTUs were clustered from the initial quality-screened 305,702 reads.

Community richness is closely linked to the environment. The peak of standardized OCTU richness for all phyla was within samples 6, 7 and 8 from Prestwick, and cluster analyses indicated clear and fine-scale hierarchical distinctions in OCTU composition within and between the two sites, with clear divergence of samples from Littlehampton (Fig. 2). OCTU richness and sediment grain size were clearly correlated (Spearman’s correlation coefficient, $n = 9, \rho = -0.82, P = 0.0108$).

Dominance of the Nematoda and rise of the Platyhelminthes. Plotting phylum richness rank for all nine independent samples (Fig. 3) shows that Nematoda are the most OCTU-rich in all nine samples, with Platyhelminthes and Arthropoda ranking second and third in all, respectively. Annotated metagenetic analyses can yield robust relative richness estimates and here it was possible to assign 374 OCTUs to phylum (Fig. 4a). The PCR primers used are not fully specific to
Metagenetics reveals phylogenetically distinct taxa. According to the comparisons of the OCTU sequences with the NCBI database, the majority (95%) of Nematoda OCTUs have never been sequenced before (Fig. 4a, Table 2). Similarly, only 6.5% of Platyhelminthes and 17.2% of Arthropoda OCTUs had 100% identity to previously sequenced specimens. The Annelida, Mollusca and Stramenopiles, however, had 50, 23 and 26.6%, respectively, of their OCTUs, with a 100% identity to previously sequenced taxa. None of the Gastrotricha OCTUs were identical to previously sequenced taxa. Overall, 54 OCTUs (representing 7,247 individual sequences) had identities below 90% to any reference nSSU sequence (OCTUs < 90% identity), and may represent previously unsampled diversity. Given the nature of the adopted clustering algorithm, the OCTUs with < 90% identity do not show any pattern of variation that would be expected of a chimeric assemblage of the already defined 428 OCTUs. Fifteen of the OCTUs with < 90% identity (300 sequences) were robustly placed within identified phyla (mainly Nematoda). A further 36 were either sisters to the sequences from known phyla, or represented deep, likely misplaced branches (Fig. 4b). Surprisingly, 11 OCTUs with < 90% identity, comprising 107 sequences in total distributed across the sampling area, were not affiliated to any phyla. The similarity of the latter sequences ranged between 83 and 89% to known nSSU genes, suggesting that they may represent unknown, distinct lineages. Furthermore, three OCTUs resulted in no MEGABLAST hit, but two of these exhibited signatures of relaxed selection, indicating that they may represent translocated ‘orphan’ nSSU genes that
Figure 4 | Percentage identity to known sequences and number of OCTUs found for the main phyla. (a) Number of different OCTUs for the main phyla recovered from the Prestwick and Littlehampton meiofaunal samples and their different levels of identity to nSSU in the GenBank/EMBL/DDBJ nucleotide database. (b) Putative taxonomic classification of 39 OCTUs that had less than 90% sequence identity to nSSU in GenBank/EMBL/DDBJ. We acknowledge and thank Richard Ling (Platyhelminthes), Frederico Batista (Mollusca), David Mann (Stramenopiles), Greg Rouse (Annelida), Charisa Wernick/MicroscopeWorld (Alveolata), Tim Ferrero (Nematoda), Kim Taylor/Warren Photographic (Arthropoda), Graham Matthews (Gastrotricha) and David Bass (Cercozoa), who gave permission to use the images in Figure 4a.

Table 2 | OCTUs with 100% identity to nSSU sequences in the GenBank/EMBL/DDBJ database.

| OCTU number | Total sequences | Accession number | Phyla | Class | Species |
|-------------|----------------|-----------------|-------|-------|---------|
| 5           | 28,217         | AY854224.1      | Nematoda | Chromadorea | Anaplectus sp. |
| 43          | 1,773          | AY854198.1      | Nematoda | Enoplea | Viscocia viscosa |
| 153         | 133            | AY854217.1      | Nematoda | Chromadorea | Spinnia parasitifera |
| 320         | 2,239          | AY854224.1      | Nematoda | Chromadorea | Daptonea normandica |
| 378         | 13             | AY854209.1      | Nematoda | Chromadorea | Dichromadara sp. |
| 177         | 244            | AY854226.1      | Nematoda | Chromadorea | Sabatieria pulchra |
| 317         | 115            | FJ040466.1      | Nematoda | Chromadorea | Theristus agilis |
| 381         | 728            | AY284693.1      | Nematoda | Enoplea | Mesacanthion sp. |
| 5           | 6,604          | GU139755.1      | Nematoda | Chromadorea | Theristus agilis |
| 15          | 5,798          | AY284695.1      | Nematoda | Chromadorea | Theristus agilis |
| 233         | 543            | AY777546.1      | Platyhelminthes | Turbellaria | Schizochilus choriurus |
| 40          | 699            | AY777566.1      | Platyhelminthes | Turbellaria | Proxenetes quadrispinosus |
| 18          | 794            | AJ012508.1      | Platyhelminthes | Turbellaria | Diascorhynchus rubus |
| 25          | 15,111         | AJ012531.1      | Platyhelminthes | Turbellaria | Paromalostomum fusculum |
| 3           | 66,604         | Gu139755.1      | Nematoda | Enoplea | Mesacanthion sp. |
| 15          | 5,798          | AY284695.1      | Nematoda | Chromadorea | Theristus agilis |
| 233         | 543            | AY777546.1      | Platyhelminthes | Turbellaria | Schizochilus choriurus |
| 40          | 699            | AY777566.1      | Platyhelminthes | Turbellaria | Proxenetes quadrispinosus |
| 18          | 794            | AJ012508.1      | Platyhelminthes | Turbellaria | Diascorhynchus rubus |
| 25          | 15,111         | AJ012531.1      | Platyhelminthes | Turbellaria | Paromalostomum fusculum |
| 4           | 23,256         | AM774524.1      | Mollusca | Bivalvia | Angulus tenuis |
| 127         | 43             | DQ279943.1      | Mollusca | Bivalvia | Chamelea striatula |
| 146         | 337            | DQ640516.1      | Mollusca | Bivalvia | Mytilus tridens |
| 109         | 7              | EF526434.1      | Mollusca | Bivalvia | Modiolus modiolus |
| 6           | 6              | DQ093437.1      | Mollusca | Gastropoda | Littorina littorea |
| 415         | 28             | AF448164.1      | Annelida | Polychaeta | Scalpelis squamata |
| 253         | 12             | AF508122.1      | Annelida | Polychaeta | Ophelia bicornis |
| 176         | 90             | AF508125.1      | Annelida | Polychaeta | Prodradix sybisticus |
| 184         | 81             | AY446901.1      | Arthropoda | Maxillopoda | Centropages hamatus |
| 307         | 29             | EU868740.1      | Arthropoda | Maxillopoda | Crangon crangon |
| 86           | 27             | L81939.1        | Arthropoda | Maxillopoda | Calanus pacificus |
| 446         | 13             | GUS94643.1      | Arthropoda | Maxillopoda | Oithona sp. |
| 557         | 5              | AB087108.1      | Stramenopiles | Phaeophyceae | Dictyota linearis |
| 186         | 24             | AB178865.1      | Stramenopiles | Oomycetes | Halichondria sp. |
| 851         | 4              | EU818944.1      | Stramenopiles | Bacillariophyta | Oantella aurita |
| 515         | 4              | A519935.1       | Stramenopiles | Labyrinthulida | Aplanochytrium stochiophorus |
| 391         | 4              | AY620355.1      | Rhizaria | Cercozoa | Uncultured cercozoan |
| 964         | 1              | DQ073794.1      | Echinodermata | Echinodermata | Spatangus naschi |
are no longer constrained by function. Resampling to generate longer sequences beyond the 200 base pairs presented here will be required to confirm or refute the biological reality of these phylogenetically distinct sequences.

**Metazoan richness curves do not approach saturation.** Along the Prestwick transect, the slope of OCTU rarefaction curves at 96% cut-off for the Nematoda, Platyhelminthes, Arthropoda, Mollusca, Stramenopiles and Annelida phyla did not reach an asymptote (Fig. 5). However, 217,221 samples and 374 OCTU-defined taxa fail to achieve saturation, even for low abundance phyla in which rarefaction curves tend to converge.

**Discussion**

Both ESPRIT\(^{13}\) and OCTUPUS can be used to cluster OTUs from the size of the environmental metagenetic data set presented here, without the need of a computing cluster. Nevertheless, ESPRIT may overestimate OTU richness, presumably because it counts both substitutional and independent insertion/deletion (indel) events in OCTU clustering. One of the most widely publicized artefacts of 454 Roche pyrosequencing is its inherent inaccuracy in calling homopolymer runs, including extensions (insertions), incomplete extensions (deletions) and carry forward errors (insertions and substitutions)\(^{16-21}\). Therefore, clustering algorithms that use indel data informatively, especially as independent events at homologous base positions\(^{22}\), are likely to be susceptible to higher OTU estimates. Analyses herein, and those of Sun et al.\(^{13}\), suggest that OCTUPUS generates a conservative number of OTUs compared with both ESPRIT and combinations of MUSCLE\(^{23}\) and DOTUR\(^{24}\). Nevertheless, it is imperative, when possible, to benchmark the analytical framework against real or simulated data, in order to estimate taxon richness as realistically as possible\(^{25}\).

The patterns of taxon richness along the Prestwick transect exhibited remarkable differences even at a microgeographical scale. Although numerous environmental factors influence meiofaunal distribution and assemblage, grain size, examined here, is known to be the predominant driver of meiofaunal community structure and diversity\(^{12}\). Nevertheless, the fine-scale community structuring also indicates that there are likely to be a host of additional biotic (for example, prokaryote communities and organic matter) and abiotic (sediment grain shape, surface composition) microgeographical factors responsible for community structuring within the benthos\(^{12}\).

Nematodes account for approximately 80% of all individual animals on earth, but quantifying the levels of relative richness between major Metazoan phyla has not been practical hitherto, because of the taxonomic magnitude of the challenge. Remarkably, along only an 800 m transect, we detected 182 Nematoda OCTUs, compared with 450 species of Nematode that have been described from around the whole British Isles\(^{25}\). From a geographical perspective, these data represent the discovery of 40% of the previously known phylum richness from a transect that represents 0.004% of the length of the British coastline (~17,820 km, Ordnance Survey). Molecular definition of substantial meiofaunal taxon richness at microgeographical scales is in alignment with intensive morphological studies\(^{35}\), yet recent metagenetic studies from the microbial biosphere have suggested that taxon richness may be orders of magnitude more than previously expected\(^{14-16}\). Reed and Knight\(^{17}\), Quince et al.\(^{18}\) and Huse et al.\(^{19}\) have recently provided alternative interpretations, based on pyrosequencing error, DNA recombination and clustering approaches, for earlier studies reporting extensive richness and rare biospheres derived from Roche 454 pyrosequencing. The data here have been quality/size filtered; only substitutional variance has contributed to taxon identification and a particularly aggressive chimeras screen has been applied to featured OCTUs. In combination with the benchmarking exercise performed against known taxa, it is therefore highly unlikely that the delineated taxa are based on sequencing error, and taxon estimates are likely to underestimate richness by at least 10%. Although global nematode diversity estimates range from 1 to 100 million species, only 24,000 species have been described\(^{20}\), with ~2,000 marine species catalogued in Europe\(^{30}\). Such relative paucity of taxonomic knowledge of nematode biodiversity is a product of their numerical dominance, the number of taxa and the high level of cryptic species: this taxon is thus a prime candidate for taxonomic exploration using metagenetic analyses. Why are Nematoda so dominant, in both abundance and diversity, in meiofaunal communities? Benthic nematodes are small in size and possess rapid generation times, and these attributes may have facilitated rapid adaptation to local conditions and habitats, especially in interstitial niches. The taxonomic richness of nematodes is, in addition, likely promoted by ancient ancestry\(^{11}\) and the absence of a dispersal phase, promoting speciation across structurally and spatially heterogeneous environments during glacial and interglacial periods. Nematodes also have cryptobiotic adaptations, including resting eggs and highly resistant larvae\(^{22}\), and they are considered to be especially resistant to environmental challenges, being among the last taxa to disappear in an environmental catastrophe\(^{20}\). As no other meiofaunal phylum shares all such characteristics simultaneously, nematodes may have had the time and potential to diversify, endowing them with collectively optimal traits for domination of the marine benthos. Considering all metazoan taxa, the detection of 428 OCTUs on only two sites is likely a gross underestimate of actual metazoan richness. In all, 70% of Nematoda OCTUs were unique to Prestwick and 58–100% of the OCTUs for the other phyla were only present in Prestwick. In the absence of large-scale dispersal events sustaining ubiquitous meiofaunal communities\(^{12}\), levels of marine meiofaunal α-diversity, at the very least, in other parts of the world are likely to exceed expectations.

Contrary to most morphological assessments of marine meiofaunal richness\(^{34,35}\), Platyhelminthes were consistently ranked as the second richest phylum in all the investigated samples. Platyhelminthes normally have to be identified from living or well-preserved material and their low ranking in other surveys perhaps arises from a global lack of taxonomic expertise, from extensive cryptis and from delicate body structures that do not survive harsh
The meiofaunal size fraction was mechanically separated from the sand and concentrated by filtering five times with filtered tap water through a 450 µm filter. Subsequent separation from fine silt was achieved by repetitive centrifugation in 1.16 specific gravity LUDOX-TM solution. Following centrifugation, each sample was retained on a distinct mesh sieve, which was then folded, sliced and placed in a 15 ml falcon tube and kept at −80 °C until DNA extraction. Samples were lysed overnight at 55 °C in lysis buffer (100 mM Tris–HCl, pH 7.5; 100 mM NaCl; 100 mM EDTA, 500 µg ml−1 proteinase K, assisted by spinning wheel mixing, and DNA extracted with the QIAamp DNA Blood Maxi Kit (Qiagen) following the manufacturer’s protocol (http://www.nature.com/naturecommunications).

**Primer design and PCR strategy.** The genes coding for rRNA have been used for decades for the identification of microbial species46 and will, henceforward, mainly be used because they are ubiquitous in cellular organisms. They are of relatively large size, contain highly conserved and variable regions that facilitate primer design and are present in tandemly repeated, multiple copies enabling efficient PCR amplification47,48. For environmental metagenetic discovery, we chose the nSSU, as considerably more reference sequences are available in public databases for nSSU (1,246,462) than for the large subunit (LSU) (180,344) (ref. 32). Moreover, ‘universal’ nSSU primers have been shown to amplify more taxa from more communities than from those designed for the LSU7. The 5’ region of the nSSU exhibits more segregating sites than the 3’ region in Metazoa9 and so was selected as the target area for OCTU discrimination. MEGA version 4.1 (ref. 35) was used to align and compare a wide range of metazoon nSSU sequences with respect to existing representative and degenerate newly priming sites. A size fraction between 250 and 500 bp in length (a combination of average mean length and maximum permitted amplicon size of a 454 Life Sciences, Roche Applied Science GSFLX amplicon sequencing run). Of all primer permutations, SSU_FO4 (5′-GCTT GTCYCAAGATTAACC-3′) and SSU_R22 (5′-GCTTGCTGCTCTTCTTTT GGATAC-3′) primers were selected, as they exhibited pronounced similarity across a wide phylogenetic phylogeny, but also flanked a highly variable dinucleotide of the nSSU where these were used for subsequent amplicon generation. Fusion primers were then developed in which a proprietary primer sequence (Adaptor A or B) of the Roche 454 GSFLX sequencing technology and a sample-specific five nucleotide key tag (to differentiate between multiple samples) were included46. All fusion primers were designed using PRIMER PREMIER 5.0 (Premier Biosoft International), considering physical and structural properties of the oligonucleotides (such as annealing temperature, G + C percentage, hairpins and false priming). The forward fusion primers were −45 bases in length and designed such that the 454 A-adaptor is followed by the tag (each of which differed by at least two bases)32, and then by the experimental forward primer (SSU_FO4). The reverse primers were designed similarly, to optimize thermal compatibility and yield amplification. The nSSU region was performed using 1 µl of genomic DNA template (1:500 dilutions) in 3 × 40 µl reactions using Pfu DNA polymerase (Promega). PCR conditions involved a 2 min denaturation at 95 °C followed by 35 cycles of 1 min at 95 °C, 45 s at 57 °C, 3 min at 72 °C and a final extension of 10 min at 72 °C. Negative controls (ultrapure water instead of DNA template) were included for all amplification reactions. Electrophoresis of triplicate PCR products was undertaken on a 2% gel with Top Vision LM GQ Agarose (Fermentas), and the expected 450 bp fragment was purified using the QIAquick Gel Extraction Kit (Qiagen), following the manufacturer’s instructions, before pooling identical samples. All purified PCR products were then quantified with an Agilent Bioanalyzer 2100, diluted to the same concentration, pooled together and sequenced in single sample and single direction (A) on half a plate of a Roche 454 GSFLX (454 Life Sciences, Roche Applied Science) sequencing platform at Liverpool University’s Centre for Genomic Research, UK. Further details regarding the rationale, theory and execution of environmental metagenetic analyses of the meiofaunal biosphere can be found in Creer et al.12.

**Sediment analysis.** Particle size analysis was carried out using a Malvern Mastersizer 2000, which uses laser diffraction to calculate the particle size distribution for individual sediment grains in the range 0.02–2000 µm. To prevent flocculation, before testing, the samples were immersed for 24 h in distilled water with added detergent (sodium hexametaphosphate). The Mastersizer determines the particle size distribution by volume and the results of the particle size analysis are reported as the cumulative median grain size.

**Data analysis.** Sequences generated from the Roche 454 GSFLX pyrosequencing were analysed using the OCTUPUS pipeline (supplementary software available at http://octopus.sourceforge.net/). Briefly, OCTUPUS comprises a number of Perl scripts that concatenate quality trimming, tag matching and size culling, before the assignment of user-defined substitutional difference-based cutoff clustering. The clustering module of OCTUPUS involves three steps. Initially, sequences are compared successively with each other by MEGALAST to define different OCTU groups, separated by a user-defined genetic distance. If an unsigned sequence matches an existing OCTU sequence (for example, 97% similarity or more), the companion sequences are aligned using MUSCLE21, and if the resulting consensus sequence differs from the original alignment, the consensus OCTU sequence is changed to reflect the diversity of sequences within the OCTU cluster. If the consensus OCTU does not change following a preset number of novel comparisons, further consensus OCTU matching

**Methods**

**Sample collection.** A total of 24 benthic samples were collected from the low-tide mark along an 800 m transect, using a standard corer methodology, from marine sandy beach substrata in Prestwick (three every 100 m between 55°30.481′N, 4°37.489′W and 55°30.194′N, 4°37.368′W) and a further three from Littlehampton (50°48.021′N, 00°32.530′W), UK, during the summer of 2007. The latter sampling site was used as a geographically disparate out-group comparison. For the sequencing analysis, each biological sample comprised three pooled 44 mm diameter × 100 mm benthic samples taken approximately 10 m apart. An additional core was taken for sediment analysis. All samples were immediately fixed in 500 ml storage pots containing 300 ml of DESS (20% DMSO and 0.25 M disodium EDTA, saturated with NaCl, pH 8.0).
sequences will be placed within the OCTU cluster, bypassing a computation- ally intractable number of multiple sequence alignments that would otherwise be required in the analysis of large metagenetic data sets. Benchmarking trials have shown that 10 additions to the OCTU cluster without consensus amend- ment provide a stable estimate of OCTU numbers and were used throughout in this study. OCTUs were annotated using MEGALAST (megalast -d database path -D 2 -p 90 -a 2 -b 1 -v 1 -i infile -F outf ile) against the downloaded GenBank/EMBL/DBJ nucleotide database and taxonomic annotation was restricted to matches of 90% and higher.

Acknowledging concerns regarding the misinterpretation of levels of richness due to the formation of recombinant DNA molecules in environmental DNA sequencing15,16, we adopted an aggressive putative chimera culling regime embed- ded within the OCTUPUS pipeline for the primary data set. The OCTUPUS chimera culling regime was approved by manual removal of putative false positive OCTUs (that is, those that exhibited more than 10 base pair mismatches with the MEGALAST reference sequence) and retrieval of clear false-negative chimeras (that is, those exhibited 100% length matches with already sequenced taxa). To compare independent estimates of taxonomic richness, the OCTUPUS clustering- algorithm was also tested against the HCluster algorithm of ESPRIT on a subsample of the larger data set. Finally, OCTU richness generated from the above OCTUPUS algorithm was benchmarked at a range of percentage similarity cutoff against a reference data set comprising the metagenetic analysis of a phylogeneti- cally diverse combination of 41 nematode5 species. Scripts for preprocessing the above data are available from the Natural Environment Research Council’s Environ- mental Bioinformatics Centre Script Repository (http://nebc.nerc.ac.uk/tools/ scripts/general-bioinformatics).

For the 54 OCTUs exhibiting BLAST hits below 90% identity, manual MEGALAST searches were conducted and phylogenetic affinities investigated by means of the NCBI taxonomy browser. Sequence data and all associated fusion primer codes have been deposited in the Entrez short read archive under accession number SRA009394.

For direct ecological comparisons of between-sample OCTU richness, the original data set was reanalysed using 15,000 randomly picked sequences (over 200 bases in length, 8.135,000) from each sample, before OCTUPUS clustering and annotation. The total number of OCTUs generated from the original and stand- ardized data sets was significantly correlated (Spearman’s coefficient: ρ = 0.783, P = 0.0132); nevertheless, interpretations derived from direct comparisons of richness between samples refer to the standardized data set.

Sample cluster analyses (UPGMA) were performed using the Multivariate Statistical Package10 using Sorensen’s Coefficient on a binary (presence/absence of OCTU) data matrix. Phylum-specific rarefaction curves for the Prestwick transect were generated using the microbiodiversity program (Version 8.0, R.K. Colwell) using a range of estimators (for example, ACE, Chao1, Jackknife1 and Bootstrap) that yielded very similar results. The ACE abundance-based coverage estimator was used because it represents a consensus view of the analyses and has proven to work well for the analysis of metagenetic data sets.

References

1. Blaxter, M. Molecular systematics—counting angels with DNA. Nature 421, 122–124 (2003).
2. May, R. M. How many species are there on earth. Science 241, 1441–1449 (1988).
3. Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. Ironing out the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity. Nature Methods 6, 636–639 (2009).
4. Quince, C. et al. Accurate determination of microbial diversity from 454 pyrosequencing data. Nat. Methods 6, 639–641 (2009).
5. Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. 12, 1889–1898 (2010).
6. Platt, H. M. & Warwick, R. M. Free-living marine nematodes. Part III. British Monhysterids (Field Studies Council, 1998).
7. Reeder, J. & Knight, R. The ‘rare biosphere’: a reality check. Nat. Methods 6, 636–639 (2009).
8. Widmann, M. et al. Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. Mar. Ecol-Prog. Ser. 320, 1–9 (2006).
9. Stage, D. E. & Eickbush, T. H. Sequence variation within the rRNA gene loci of vertebrates. Prog. Zoo1. 171–180 (1988).
10. Markmann, M. & Tautz, D. Reverse taxonomy: an approach towards determining the integral structure of a metazoan. Mol. Ecol. 6, 764–770 (2005).
11. Boucher, G. & Lambshead, P. J. D. Ecological biodiversity of marine nematodes in samples from temperate, tropical, and deep sea regions. Conserv. Biol. 19, 1594–1604 (1995).
12. Heip, C. H. et al. Community analysis of the benthic meiofauna of Erford/Langesundfjord. Environ. Microbiol. 5, 12115–12120 (2003).
13. Boucher, G. & Lambshead, P. J. D. Phylogenetic character analysis of nematode identification. Mol. Ecol. 11, 839–850 (2002).
14. Lawton, J. H. et al. Biodiversity inventories, indicator taxa and the effects of habitat modification in tropical forest. Nature 391, 72–76 (1998).
15. Lambshead, P. J. D. Sub-catastrophic sewage and industrial-waste contamination as revealed by marine nematode faunal analysis. Mar. Ecol-Prog. Ser. 29, 247–260 (1986).
16. Markmann, M. & Tautz, D. Reverse taxonomy: an approach towards determining the diversity of meioebenthic organisms based on ribosomal RNA signature sequences. Philos. Trans. R Soc. Lond B Biol. Sci. 360, 1917–1924 (2005).
17. Danovaro, R. et al. Early-phase functional decline of deep-sea ecosystem functioning linked to benthic biodiversity loss. Curr. Biol. 18, 1–8 (2008).
18. Worm, B. et al. Biodiversity loss in the ocean: how bad is it? Response. Science 316, 1282–1284 (2007).
19. Soltan, M. et al. Extinction and ecosystem function in the marine benthos. Science 306, 1177–1180 (2004).
20. Palumbi, S. R., McLeod, K. L. & Grunbaum, D. Ecosystems in action: lessons from marine ecology about recovery, resistance, and reversibility. BioScience 58, 33–42 (2008).
21. Platt, H. M. & Warwick, R. M. Free-Living Marine Nematodes (Cambridge Univ Press, 1983).
22. Blaxter, M. L. et al. A molecular evolutionary framework for the phylum Nematoda. Nature 392, 71–75 (1998).
23. Bhadury, P. et al. Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. Mar. Ecol-Prog. Ser. 320, 1–9 (2006).
24. Sogin, M. L. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbiol. 71, 1507–1516 (2005).
25. Sogin, M. L. et al. Microbial population structures in the deep marine environment. Proc. Natl Acad. Sci. USA 102, 1439–1450 (2005).
26. Platt, H. M. & Warwick, R. M. Free-living marine nematodes. Part I. British Enchytraeids (Cambridge Univ Press, 1983).
27. Warwick, R. M. & Somerfield, P. J. The integral structure of a benthic faunal assemblage. J. Exp. Mar. Biol. Ecol. 330, 12–18 (2006).
28. Schloss, P. & Handelsman, J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbiol. 71, 1507–1516 (2005).
29. Margulis, L. et al. Eukaryotic phylogeny: internal transcribed spacer data and a new hypothesis for the origin of eukaryotic cells. Proc. Natl Acad. Sci. USA 91, 376–380 (1994).
30. Kusum, V., Engelbrektsen, A., Ochman, H. & Hugenholtz, P. Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. Environ. Microbiol. 12, 118–123 (2010).
31. Creer, S., Poon, C. E., Malhotra, A. & Thorpe, R. S. Optimal intron analyses in the trimerserus radiation of Asian pitvipers. Syst. Biol. 55, 57–72 (2006).
32. De Jonge, V. & Bouwman, L. A simple density separation technique for the rapid identification of nematodes. Mar. Ecol-Prog. Ser. 4, 169–172 (1986).
33. Boucher, G. & Lambshead, P. J. D. Phylogenetic character analysis of nematode identification. Mol. Ecol. 11, 839–850 (2002).
34. Heip, C. H. et al. Community analysis of the benthic meiofauna of Erford/Langesundfjord. Mar. Ecol-Prog. Ser. 46, 171–180 (1998).
35. Heip, C. H. et al. Community analysis of the benthic meiofauna of Erford/Langesundfjord. Mar. Ecol-Prog. Ser. 46, 171–180 (1998).
36. Bhadury, P. et al. Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. Mar. Ecol-Prog. Ser. 320, 1–9 (2006).
37. Stage, D. E. & Eickbush, T. H. Sequence variation within the rRNA gene loci of vertebrates. Prog. Zoo1. 171–180 (1988).
38. Widmann, M. et al. Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. Mar. Ecol-Prog. Ser. 320, 1–9 (2006).
39. Stage, D. E. & Eickbush, T. H. Sequence variation within the rRNA gene loci of vertebrates. Prog. Zoo1. 171–180 (1988).
Acknowledgments

We thank Neil Hall, Andrew Cossins and Margaret Hughes of the Liverpool Advanced Molecular Genetic Facility for Roche 454 sequencing support, in addition to the staff of Edinburgh University’s Gene-Pool facility and Holly Bik for field support. We also thank Emlyn Roberts, Wendy Grail, Simeon Walker and technical support staff at Bangor University. This work was supported by a Natural Environment Research Council (NERC) New Investigator Grant (NE/E001505/1) and a Post Genomic and Proteomics Grant (NE/F001266/1) to S.C. and a Portuguese Foundation for Science and Technology (FCT) grant (SFRH/BD/27413/2006) to V.G.F.

Author contributions

S.C., V.G.F., G.R.C. and W.K.T. conceived and designed the experiments. V.G.F., H.F.J., S.P.N., M.P. and S.C. performed the experiments. V.G.F and S.C. analysed the data. S.C., G.R.C., M.L.B., M.P., P.J.D.L., W.K.T., S.P.N. and D.M.P. were involved in the critical review of the manuscript. V.G.F., G.R.C., D.M.P., M.L.B. and S.C. wrote the paper.

Additional information

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Fonseca, V.G. et al. Second-generation environmental sequencing unmasks marine metazoan biodiversity. Nat. Commun. 1:98 doi: 10.1038/ncomms1095 (2010).

License: This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/