RESEARCH PAPER

Extensive gaps and biases in our knowledge of a well-known fauna: implications for integrating biological traits into macroecology

Elizabeth H. M. Tyler[1,2,3]*, Paul J. Somerfield[4], Edward Vanden Berghe[5], Julie Bremner[6], Emma Jackson[7], Olivia Langmead[7], Maria Lourdes D. Palomares[8] and Thomas J. Webb[1]*

[1]*Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK,* [2]*King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia,* [3]*Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK,* [4]*Plymouth Marine Laboratory, Prospect Place, Plymouth, Devon PL1 3DH, UK,* [5]*Ocean Biogeographic Information System, Rutgers University Institute of Marine and Coastal Sciences, 71 Dudley Road, New Brunswick, NJ 08901-8525, USA,* [6]*Centre for Environment, Fisheries and Aquaculture, Pakefield Road, Lowestoft, Suffolk NR33 0HT, UK,* [7]*Marine Life Information Network (MarLIN), Marine Biological Association, The Laboratory, Citadel Hill, Plymouth PL1 2PB, UK,* [8]*Fisheries Centre, University of British Columbia, 2202 Main Mall, Vancouver, Canada BC V6T 1Z4*

*Correspondence: Elizabeth H. M. Tyler and Thomas J. Webb, Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.
E-mail: elizabeth.tyler@kaust.edu.sa; t.j.webb@sheffield.ac.uk

ABSTRACT

Aim Ecologists seeking to describe patterns at ever larger scales require compilations of data on the global abundance and distribution of species. Comparable compilations of biological data are needed to elucidate the mechanisms behind these patterns, but have received far less attention. We assess the availability of biological data across an entire assemblage: the well-documented demersal marine fauna of the United Kingdom. We also test whether data availability for a species depends on its taxonomic group, maximum body size, the number of times it has been recorded in a global biogeographic database, or its commercial and conservation importance.

Location Seas of the United Kingdom.

Methods We defined a demersal marine fauna of 973 species from 15 phyla and 40 classes using five extensive surveys around the British Isles. We then quantified the availability of data on eight key biological traits (termed biological knowledge) for each species from online databases. Relationships between biological knowledge and our predictors were tested with generalized linear models.

Results Full data on eight fundamental biological traits exist for only 9% (*n* = 88) of the UK demersal marine fauna, and 20% of species completely lack data. Clear trends in our knowledge exist: fish (median biological knowledge score = six traits) are much better known than invertebrates (one trait). Biological knowledge increases with biogeographic knowledge and (to a lesser extent) with body size, and is greater in species that are commercially exploited or of conservation concern.

Main conclusions Our analysis reveals deep ignorance of the basic biology of a well-studied fauna, highlighting the need for far greater efforts to compile biological trait data. Clear biases in our knowledge, relating to how well sampled or 'important' species are suggests that caution is required in extrapolating small subsets of biologically well-known species to ecosystem-level studies.

Keywords
Benthic invertebrates, biological traits analysis, demersal fishes, ecosystem function, functional traits, marine biodiversity, marine macroecology, missing data, United Kingdom.

INTRODUCTION

Widespread appreciation of the fact that an organism's interactions with its environment and with other organisms are determined by its biological traits rather than its taxonomic nomenclature (McGill *et al.*, 2006; Raffaelli, 2007) has led to an explosion in the application of traits to community ecology (Naeem & Bunker, 2009), and macroecologists have gained an understanding of the mechanisms behind large-scale patterns in biodiversity by linking species traits to their abundance and distribution (Webb *et al.*, 2009; Buckley & Freckleton, 2010; Verberk *et al.*, 2010). Traits can predict the composition of ecological communities from their environment (Dolédec *et al.*, 1999; McGill *et al.*, 2006; Menezes *et al.*, 2010; Webb *et al.*, 2010), and when traits can be linked to ecosystem functioning (Bremner *et al.*, 2006; Petchey & Gaston, 2006; Lavorel *et al.*, 2010) it may prove possible to document the functional consequences of biodiversity loss (Solan *et al.*, 2004).

Traits are well-defined, measurable properties of an organism (McGill *et al.*, 2006), typically assigned at the species level. Those most relevant to macroecology and large-scale community ecology include life history (e.g. life span, growth rate, body size), reproduction (e.g. egg size, fecundity), feeding ecology (e.g. trophic level, diet breadth) and behaviour (e.g. dispersal ability) (Webb *et al.*, 2009). A consequence of this diversity of relevant traits is that whereas a 'taxonomic' macroecological study requires only information on the geographic distributions of species, traits-based analyses require an additional suite of trait measurements for each species (Schleuter *et al.*, 2010). Yet, as Naeem & Bunker (2009, 281) state, trait data are '. . . at best, dispersed throughout the literature, and at worst lacking altogether'. Missing data have a serious impact on statistical inference (Nakagawa & Freckleton, 2008), and so the patchiness of availability of trait data is a potential barrier to the trait-based approach to ecology; but there has been no effort to quantify the extent to which trait data are available for an entire, species-rich assemblage.

Here we assess the feasibility of applying traits-based approaches across an entire assemblage at a macroecological scale by quantifying the state of the biological knowledge of the marine fauna of the United Kingdom. This fauna is especially well documented in terms of its biogeography and taxonomy (Costello *et al.*, 2010), thanks to a long history of marine biological research (e.g. Hayward & Ryland, 1991a; Thurstan *et al.*, 2010). Traits-based descriptors of ecological composition and function, including applications to macroecology, have been attempted for subsets of the UK's marine fauna (e.g. Solan *et al.*, 2004; Bremner *et al.*, 2006; Tillin *et al.*, 2006; Webb *et al.*, 2009), but the extent to which such approaches can be generalized to the entire regional assemblage remains unknown.

We first derive a list of commonly observed demersal marine fish and benthic invertebrates from five spatially extensive surveys around the UK coast, resulting in 973 species from 15 phyla. We then collate information on eight broadly defined basic biological traits of particular relevance to macroecological analyses, to assess the extent of our biological knowledge of this fauna. Finally, we test for correlates of the state of this knowledge across our species, including taxonomic affiliation, biogeographic knowledge, body size, conservation status and commercial importance.

METHODS

Biogeographic records of the UK marine fauna in a global context

We used the Ocean Biogeographic Information System (OBIS; http://www.iobis.org), the largest holder of marine biogeographical datasets, to rank the exclusive economic zones of 229 countries in terms of the number and density of records (each record represents one observation of a particular species at a particular geographic location). Overseas territories, which are not contiguous with the main area of a country, were treated separately from their mother countries as they are geographically and biologically distinct (e.g. the British Indian Ocean Territory was separated from the United Kingdom). The UK ranked first in terms of number of records (4,901,017; the second ranked country is the USA with 3,137,478), and had the eighth highest density of records (3.5 km^{-2}), justifying our assertion that the marine fauna of the UK is among the best known in the world.

Defining the UK marine fauna

We focus on the marine fauna associated with the sea-bed (demersal fish and macrobenthic invertebrates) because this fauna is taxonomically well known and the focus of previous trait-based research (e.g. Jennings *et al.*, 1998; Bremner *et al.*, 2006; Tillin *et al.*, 2006) and (in the case of benthic invertebrates) largely sedentary, making them useful indicators of environmental change as they cannot easily avoid disturbance (Tillin *et al.*, 2006). To draw up as complete a list as possible of the UK demersal fish and macrobenthic invertebrate fauna, we chose five spatially extensive surveys of benthic habitats that were complementary in terms of geographic coverage, sampling methodology and habitats and taxa targeted (for full details of the locations surveyed and sampling methodologies employed see Figure S1 and Table S1 in Supporting Information). Surveys recorded (at least) the presence or absence of multiple species across multiple sites, and we focused on the most recent years available or the years with the best spatial coverage. These surveys combined to produce a list of 2641 unique taxa, fully identified to species level, and only parts of the north-western UK continental shelf were not well represented. We then excluded very rare species from each survey because our focus is on the 'core' members of the UK marine fauna, i.e. those that are regularly observed in biodiversity surveys. Species were excluded if they occurred in ≤ 1% of all samples, or in ≤ 10 individual samples, whichever was greater, but as long as a species exceeded this cut-off in at least one survey it was included in our final list. We also excluded taxa not identified to species level, and those from non-target groups (e.g. marine plants). However, we did

Table 1 The eight biological trait categories used in our analysis, with individual traits belonging to each category where relevant, and the importance of each trait category to macroecological analyses. A full definition of individual traits, together with additional information on their importance, is given in Table S2.

Trait categories	Individual traits	General importance
Body size	Maximum linear dimension (mm)	Key life history and macroecological trait (Gaston & Blackburn, 2000; Hildrew *et al.*, 2007), correlates with other traits, e.g. metabolic rate (Brown *et al.*, 2004), and response to disturbance, e.g. (Tillin *et al.*, 2006)
Diet	Major components of diet (e.g. plankton, detritus, invertebrates, fish)	Determines trophic structure, energy flow and nutrient cycling within communities
Feeding method	Mode of food capture (e.g. grazer, predator, filter feeder)	Influences inter-specific interactions, nutrient and energy cycling and predicts response to disturbance
Reproductive timing	Reproductive frequency (no. of breeding attempts per year); reproductive duration (length of breeding season); reproductive season (time of year of breeding season)	Determines annual productivity, as well as changes in phenology as a response to climate change (e.g. Edwards & Richardson, 2004)
Fecundity	Annual fecundity (number of eggs/young produced in a reproductive event); egg size (diameter of eggs/size of young)	Related to rate of population increase (e.g. Jennings *et al.*, 1998; McGill *et al.*, 2006) and thus to (meta)population dynamics and response to environmental change
Larval dispersal	Larval development mode (presence of a planktonic larval stage); larval duration (duration of the larval period; scored as 0 for live-bearers)	Influences geographic range size (Jablonski, 1986) and the relationship between distribution and abundance (Foggo *et al.*, 2007)
Adult dispersal	Adult movement method (mode of movement, e.g. crawling, swimming, burrowing, sessile); adult migration (existence of significant post-recruitment horizontal movements outside typical daily movements)	Important in determining large-scale colonization and extinction dynamics in response to changing environmental conditions, which is key to structuring macroecological relationships (Freckleton *et al.*, 2005)
Longevity	Adult life span (maximum recorded life span); age at maturity (minimum recorded age at maturity)	Influences population dynamics through effects on reproductive strategy and response to disturbance (Jennings *et al.*, 1998)

retain a few typically midwater species if they were commonly sampled by the survey methodology. Having derived this 'core' UK benthic and demersal community, all species names were checked against the standard World Register of Marine Species (WoRMS; http://www.marinespecies.org), enabling synonyms to be resolved and ensuring that all names, as well as higher taxonomic classifications, were consistent within and between all surveys. The final species list included 973 species from 15 phyla and 40 classes, with 148 fish and 825 invertebrates (see Appendix S1 for a full list).

The choice of which traits to use in traits-based analyses is important, and will depend on the specific questions to be addressed (Petchey & Gaston, 2006). Because our focus is on macroecology, we considered traits likely to influence the distribution, abundance and population dynamics of species at large spatial scales. We identified 14 such traits, which we amalgamated into eight broader trait categories (Table 1); further details on the 14 individual traits are available in Table S2. We limited our choice of traits to those most widely applicable across the range of taxa in our database, and used the eight broad trait categories to ensure cross-taxon relevance. For instance 'larval dispersal' includes developmental mode (most relevant for invertebrates) and planktonic larval duration (most frequently recorded for fish, Table 1). For each species, we scored the presence or absence of any information on each of the eight traits.

Information on any one of the component parts of a major trait was sufficient for that trait to be considered 'known'. For instance, 'reproductive timing' was considered known providing we had information on at least one of reproductive frequency, reproductive duration or reproductive season. By concentrating on simple data availability, we avoided the need to the assess accuracy of the specific trait values assigned to each species.

We sourced trait data for invertebrates from the Biological Traits Information Catalogue (BIOTIC), developed by the Marine Life Information Network (MarLIN; http://www.marlin.ac.uk/biotic) at the Marine Biological Association of the UK. Body size measures for invertebrates were taken from the standard reference for British marine species, namely Hayward & Ryland (1991a, 1991b). Fish data are from FishBase (Froese & Pauly, 2010). These databases have incorporated data from many sources, and we consider them to constitute the extent of the readily available data on the biology of an entire fauna. E.H.M.T., P.J.S. and T.J.W. contributed substantial data to BIOTIC from a previous project (Webb *et al.*, 2009).

Using these trait data, we define the 'biological knowledge' of each species as the total number of traits (zero to eight) for which we were able to find data. Only trait data resolved to the species level were considered (i.e. when information on a trait was missing for a species, we did not consider information from a related species as a surrogate). Data recorded for subspecies

were included within the parent species. In the case of multiple subspecies, a trait was scored as 'known' provided that information was available for at least one subspecies. For body size measures, the maximum value for any subspecies was taken, as this is the most commonly recorded statistic. We note that, in general, when data exist they do so as a single, species-level value for a given trait with little or no indication of the strength of evidence (e.g. number of individuals examined). Sometimes, attempts have been made to capture intra-specific variability in trait expression, for instance by using a fuzzy coding approach to score the affinities of a species to different trait modalities (Chevenet *et al.*, 1994); but for our purposes, any such intra-specific variation is subsumed by our reduction of knowledge of a trait to a simple binary variable.

Correlates of biological knowledge

We considered four potential correlates of biological knowledge: biogeographic knowledge, body size, conservation status and commercial importance. We also tested for differences in biological knowledge between taxonomic groups, for instance between fish (classes Actinopterygii, Elasmobranchii and Myxini) and invertebrates (a division that we term 'clade'), and between classes within each phylum. Details of how we included taxonomy in our analyses are provided in Analysis below.

We defined biogeographic knowledge for each species as the total number of times it occurs within OBIS (n_{OBIS}). Each occurrence corresponds to a record of that species at a particular geographic location. n_{OBIS} measures the extent to which a species has been sampled globally; itself a composite measure of global distribution, detectability by common sampling methods and spatial biases in global sampling.

Body size is the maximum recorded length of each species, as defined in Table 1. To avoid circularity, all models that included body size as a predictor excluded body size from the response, so that biological knowledge in these cases was knowledge of all traits *except* body size, and varied between 0 and 7.

We defined conservation status according to whether any of our species were included in at least one of several priority conservation lists for this region: the UK Biodiversity Action Plan's list of priority species (http://jncc.defra.gov.uk/page-5717), the OSPAR Commission's list of threatened and/or declining species or habitats (http://www.ospar.org/content/content.asp?menu=00180302000014_000000_000000), or the European Union Habitats Directive (ec.europa.eu/environment/nature/legislation/habitatsdirective/) Annex 2 (species of concern) or Annex 1 (habitats of concern; habitats are relevant because of the existence of biogenic reef-forming invertebrates in our dataset). Twenty-five fish and 15 invertebrate species from our list were categorized as being of 'conservation concern'.

Commercial importance of benthic invertebrates was rated by marine invertebrate experts (P.J.S. and, independently, three members of the shellfish team at the UK's central fisheries laboratory, http://www.cefas.defra.gov.uk/) and by reference to Barne *et al.* (1997), resulting in 37 commercially important invertebrate species. Fish species were scored as either being of no, minor, commercial or high commercial importance using the ratings given in FishBase (Froese & Pauly, 2010), which uses the Food and Agriculture Organization's (http://www.fao.org) list of commercial species, validated using the Sea Around Us database (http://www.seaaroundus.org). Minor, commercial and high commercial importance were collapsed in to a single 'commercially important' rating (containing 90 species) for all analyses that included invertebrates. Our full dataset, including data on all traits and potential correlates, is available in Appendix S1.

Analysis

We calculated the proportion of all species for which we had data for each of the eight traits. We compared biological knowledge between clades (fish versus invertebrates) using a generalized linear model (GLM). Because the response variable is bounded at both ends (between zero and eight traits), this and all other models were fitted with a binomial error structure and logit link. Next, we estimated the variance in biological knowledge explained by taxonomic class as a proportion of total variance in data availability (Gelman & Hill, 2007) by fitting a binomial generalized linear mixed effects model (GLMM) with no predictors and class as a random factor. We estimated 95% confidence intervals (CIs) of biological knowledge for each class from this model, identifying significant differences between classes on the basis of no overlap in CIs.

We tested for relationships between biological knowledge and n_{OBIS}, conservation status and commercial importance using binomial GLMs. In such comparative analyses, species cannot be considered as independent data points, and so some means of controlling for the relatedness between species is required. Unfortunately, standard methods of phylogenetically weighted GLMs (e.g. Freckleton *et al.*, 2002) are not well developed for data with non-Gaussian error structures. We therefore applied the following modelling strategy. First, we note that the non-independence assumptions of a GLM fitted to comparative data are violated only if the residuals are not independent with respect to taxonomy. This can be assessed directly by estimating taxonomic signal in the residuals of a fitted model. To do this, we used the full taxonomic hierarchy of all species in our database (phylum, class, order, family, genus, species) to create a matrix of pairwise taxonomic distances between each species suitable for estimating the parameter λ, a measure of the strength of taxonomic signal which varies between 0 (no taxonomic signal) and 1 (variance increases with taxonomic distance according to a Brownian model; see Freckleton *et al.*, 2002, for details).

We fitted a GLM of the form

$$\text{Biological Knowledge} \sim \log(n_{OBIS}+1)+ \\ \text{Conservation Concern}+\text{Commercial Importance}$$

across all species (Model 1). Because of concerns about over-dispersion, this and all subsequent GLMs were fitted with quasibinomial errors and a logit link; this results in identical parameter estimates but wider confidence intervals when compared to a GLM with binomial errors (Faraway, 2006). We fitted

similar models separately for invertebrates and for fish. In each case, we estimated λ on the residuals. Because residuals retained taxonomic structure across all species, and at the clade level, we then fitted models at the class level. Class is an appropriate level for analysis as it effectively identifies major groups of marine species (e.g. gastropod molluscs, sea stars, polychaete worms, bony fish), and 15 classes in our database have ≥ 10 species. Taxonomic signal persisted in a few class-level models, in which case we ran models at the order level. However, we did not run models on groups below the order level, as sample sizes and taxonomic resolution are insufficient.

Note that our modelling approach is identical to that advocated by Bolker *et al.* (2009) as preliminary procedures to fitting GLMMs. Indeed, GLMMs with taxonomic groupings included as random factors, or 'taxonomic mixed models' (Hadfield & Nakagawa, 2010) would be an alternative way to incorporate taxonomic structure in our models. However, fitting such general models to our data is problematic, because most higher taxonomic groups contain very few species (e.g. a median of six species per taxonomic class, with nine monospecific classes), and many groups have limited variability in covariate states (e.g. there are 187 species in class Polychaeta, of which only three are of conservation concern and four are of commercial importance). Our approach means that we can use only sensible covariate combinations for each taxonomic grouping, and it avoids trying to estimate variances for groups with very few species.

In order to test the *a priori* prediction that large-bodied species may be better-studied, and thus better known, than small-bodied species, we fitted a second series of models of the form

$$\text{Biological Knowledge} \sim \log(n_{\text{OBIS}} +1) + \log(\text{maximum size}) + \text{Conservation Concern} + \text{Commercial Importance.}$$

In Model 2 we removed body size from the definition of biological knowledge (which thus varied between zero and seven) to enable us to include log-transformed maximum size as a predictor. Model fitting proceeded exactly as described for Model 1. Note that Model 2 included only those species for which we had an estimate of maximum size, whereas Model 1 was fitted to all species.

Our predictors are inter-related, meaning that models may suffer from problems associated with collinearity (Faraway, 2005). In particular, commercially exploited species tend to be large-bodied (linear model on log-transformed maximum length, positive effect of commercial importance, $t = 5.27$, $P < 0.001$), have a high n_{OBIS} (linear model on log-transformed n_{OBIS}, positive effect of commercial importance, $t = 8.26$, $P < 0.001$), and are likely to be of conservation concern (all 25 fish species of conservation concern are also commercially exploited). However, variance inflation in the models described above is within acceptable limits: standard errors for all predictors are inflated by a factor of ≤ 1.75 in Model 1 and ≤ 2.24 in Model 2, with values > 2 only observed in models of fish species where conservation concern and commercial importance are especially confounded. Individual parameter estimates may be somewhat sensitive to this level of variance inflation, but it is very unlikely to be sufficient to affect our conclusions.

We used standard GLM diagnostics (influence measures and distribution of residuals; Faraway, 2006) to inspect our models. Residuals were always satisfactory, but in a few cases we identified one or two influential species (high Cook's D), and so re-ran the relevant models excluding these species. All data manipulation and statistical analysis were carried out in R (R Development Core Team, 2009), using the *vegan* package to create distance matrices from taxonomies, the *ape* package to convert these to variance–covariance matrices, and the *CAIC* package to derive estimates of λ (Paradis *et al.*, 2004; Oksanen *et al.*, 2009; Orme *et al.*, 2009).

RESULTS

Across the 973 species in our dataset, body size is the best documented trait, with data available for 80% of all species, covering all fish and around 75% of invertebrates (Fig. 1). All other traits were known for fewer than a third of species, with data on fecundity being especially scarce (19%; Fig. 1).

Generally, fish are much better known than invertebrates: data on each trait were available for > 70% of fish species (except adult dispersal, 52%); but data were available for < 25% of invertebrate species for each trait apart from body size, with fecundity known for only 8% (Fig. 1). These differences are confirmed by a binomial GLM comparing biological knowledge between these two clades: mean biological knowledge for fish was 6.12 (95% CI 5.92–6.31), significantly more than for invertebrates (mean 1.98, 95% CI 1.76–2.20; $z = 30.9$, $P < 0.001$).

Nearly 60% of the variance in biological knowledge is explained by differences between classes (intra-class correlation estimated from a GLMM with taxonomic class fitted as a random effect = 0.59) and several classes stand out for having particularly low knowledge (e.g. Demospongiae) or high knowledge (e.g. Echinoidea and Actinopterygii) (Table 2). However, patterns in biological knowledge across the taxonomy are complex: median biological knowledge was 0 or 1 in 33 of 40 classes, there was wide overlap in CIs across classes and there was high variability between classes within most phyla (e.g. Mollusca, Annelida, Arthropoda; Table 2).

Model 1 revealed n_{OBIS} to be a consistently good predictor of biological knowledge across most taxonomic groups, with a general trend for biological knowledge to increase with biogeographic knowledge (Table 3, Fig. 2a; 22 of 29 values of $b_{n(\text{OBIS})}$ are significantly positive). In those groups that contained sufficient variance in the other predictors, there was also a very general pattern for increased biological knowledge in species of conservation concern and in species of commercial importance (Table 3; all values of $b_{\text{conservation}}$ and $b_{\text{commercial}}$ are positive, 9 of 14 significantly so; see Fig. 3). In the fish models, where commercial importance was scored on an ordinal scale, the significant positive coefficients indicate a monotonic increase in knowledge with increasing commercial exploitation. Models explained a moderate proportion of deviance (pseudo R^2 generally around 0.2–0.4; Table 4), and fitting models at the class level was in most cases sufficient to remove taxonomic signal from the residuals.
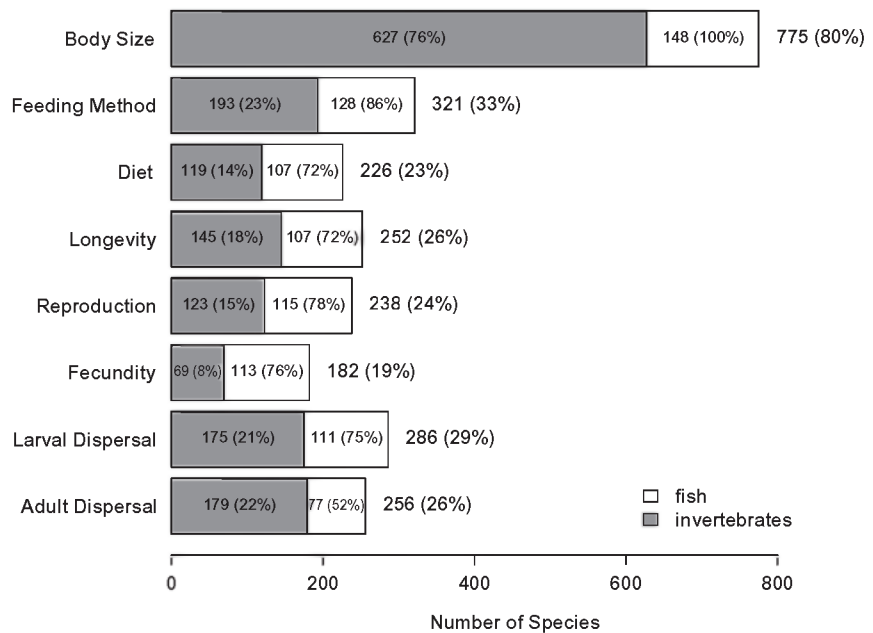
Figure 1 Availability of data on eight key biological traits across the UK marine fauna. For each trait, the total number of species for which we have data is given, as well as the subtotals for fish and invertebrates. We also give percentages, out of 973 species (total), 825 species (invertebrates) or 148 species (fish).

Across those species for which we had an estimate of body size, there is a general trend for biological knowledge to increase with body size (Model 2; Table 4, Fig. 2b). However, although most (16/21) body size coefficients are positive, only nine are significantly so (Table 4); there is rather more variation between taxonomic groups (Table 4, Fig. 2b), and including body size typically adds rather little to the explanatory power of the models (Table 4).

DISCUSSION

We provide the first assessment of the state of biological knowledge of an entire fauna at a national scale. Given that we consider broad definitions of basic aspects of biology across the most commonly occurring species from two major faunal groups, the lack of information is startling, with full knowledge for just 9% ($n = 88$) of the 973 species in our database. For invertebrates, including familiar and commercially important species such as mussels, oysters, crabs and lobsters, we have no data at all on the biological traits of nearly a quarter (192) of 825 species; the median biological knowledge score across invertebrates is one, and exceeds zero only because of the relatively wide availability of estimates of body size (available for 76% of invertebrates). Some traits, such as fecundity, are virtually unknown for British marine invertebrates. Fish are generally much better known: body size is known for all 148 species, and over a quarter of species attain the maximum score of eight. Nonetheless, data on most basic life-history traits are lacking for around a quarter of British demersal fish species, with no information other than body size for one species (Stephen's goby, *Gobius gasteveni*).

The wide availability of size information is reassuring, as body size has a defining role in aquatic ecosystems at a range of scales (Jennings *et al.*, 1998; Tillin *et al.*, 2006; Hildrew *et al.*, 2007; Olden *et al.*, 2007; Blanchard *et al.*, 2008; Webb *et al.*, 2009).

Thus, broad-scale analyses using this single life-history trait are both feasible with existing data, and potentially revealing of the structure and function of marine ecosystems. However, until we have a greater understanding of how size is related to other traits important in determining the large-scale distribution of species, or their ecological function, consideration of body size alone will not suffice. Moreover, examination of links between species diversity, trait diversity and ecosystem function (Webb *et al.*, 2010) require information on multiple traits, data that appear to be lacking at present.

Additional data on some traits for some species doubtless exist in the literature, or as expert knowledge not yet incorporated in existing databases; but we believe our results to reflect a genuine lack of data on the biological traits of UK marine species, not least because in recent decades basic studies of the biological traits of individual species have seldom been a high research priority. Furthermore, we have made no attempt to assess the accuracy of published trait data. This ignores inconsistencies of observation, inaccurate propagation of knowledge, misidentifications and what is likely to be very significant intraspecific (both ontogenetic and geographic) variation in trait expression. One consequence of this is that even if a strong link were made between a particular combination of biological traits and a specific ecosystem function (Solan *et al.*, 2004) our ability to predict the functional consequences of environmental change in any given spatio-temporal setting would be limited by a lack of knowledge of the actual expression of traits by individuals of different species present at that time.

Some taxonomic groups emerge as particularly well (or poorly) known, but taxonomy was generally not a good predictor of knowledge, which is very variable within most classes. However, well-known species do not form a random subset of the entire fauna. For example, the number of occasions that a species is recorded in OBIS was the strongest predictor of bio-

Table 2 Biological knowledge across the 40 taxonomic classes in our dataset. Biological knowledge is defined as the number of traits for which we have data, and varies between zero and eight. It is shown here as the minimum, median (in bold) and maximum value across the n species in each class. Also shown is the number of species with a biological knowledge score of 0 (n_{NONE}) and the number of species with the maximum biological knowledge score (n_{FULL}). The vertical lines on the left of the table join those classes with overlapping 95% confidence intervals of biological knowledge, according to a generalized linear mixed model. Classes are ordered from lowest to highest mean biological knowledge, and the three classes of fish are shown in italics.

Class	Phylum	Biological knowledge	n	n_{NONE}	n_{FULL}
Scyphozoa	Cnidaria	**0**	4	4 (100%)	0
Demospongiae	Porifera	0**1**...7	42	13 (30.1%)	0
Clitellata	Annelida	0**1**	6	3 (50.0%)	0
Calcarea	Porifera	0**1**	7	3 (42.3%)	0
Cephalopoda	Mollusca	0**1**	11	3 (27.3%)	0
Stenolaemata	Bryozoa	0**1**	8	2 (25.0%)	0
Gymnolaemata	Bryozoa	**0**...8	45	24 (53.3%)	1 (2.2%)
Caudofoveata	Mollusca	**0**	2	2 (100%)	0
Solenogastres	Mollusca	**0**	2	2 (100%)	0
Tentaculata	Ctenophora	**0**	2	2 (100%)	0
Holothuroidea	Echinodermata	0**1**.4	9	1 (11.1%)	0
Anopla	Nemertina	0**1**	3	1 (33.3%)	0
Polyplacophora	Mollusca	**1**	4	0	0
Sipunculidea	Sipuncula	**1**	3	0	0
Rhabditophora	Platyhelminthes	**0**	1	1 (100%)	0
Craniata	Brachiopoda	**1**	1	1 (100%)	0
Entoprocta	Entoprocta	**1**	1	0	0
Insecta	Arthropoda	**1**	1	0	0
Phascolosomatidea	Sipuncula	**1**	1	0	0
Rhynchonellata	Brachiopoda	**1**	1	0	0
Scaphopoda	Mollusca	**1**	1	0	0
Staurozoa	Cnidaria	**1**	1	0	0
Hydrozoa	Cnidaria	0**1**...8	38	3 (7.9%)	1 (2.6%)
Crinoidea	Echinodermata	0**1**...7	3	1 (33.3%)	0
Phoronida	Phoronida	2**3**4	2	0	0
Polychaeta	Annelida	0**1**...8	187	86 (46.0%)	6 (3.2%)
Pycnogonida	Arthropoda	**1**...5	5	0	0
Ascidiacea	Chordata	0**1**...8	33	4 (12.1%)	4 (12.1%)
Maxillopoda	Arthropoda	0**1**...8	10	1 (10.0%)	2 (20.0%)
Gastropoda	Mollusca	0**1**...8	85	11 (12.9%)	5 (5.9%)
Leptocardii	Chordata	**5**	1	0	0
Myxini	*Chordata*	*5*	*1*	*0*	*0*
Malacostraca	Arthropoda	0**1**...8	154	10 (6.5%)	8 (5.2%)
Anthozoa	Cnidaria	0**1**...8	40	6 (15.0%)	1 (2.5%)
Bivalvia	Mollusca	0**1**...8	75	8 (10.7%)	10 (13.3%)
Ophiuroidea	Echinodermata	0**1**...8	13	1 (7.7%)	1 (7.7%)
Asteroidea	Echinodermata	**0**...6.8	15	1 (6.7%)	5 (33.3%)
Echinoidea	Echinodermata	**1**...8	8	0	5 (62.5%)
Elasmobranchii	*Chordata*	*3.6.8*	*15*	*0*	*1 (6.7%)*
Actinopterygii	*Chordata*	*1...78*	*132*	*0*	*38 (28.8%)*
Invertebrates		0**1**...8	825	192 (23.3%)	49 (5.9%)
Fish		**1**...6.8	148	0	39 (26.4%)
Total		0**1**...8	973	192 (20.0%)	88 (9.0%)

logical knowledge, suggesting that, for the most part, we tend to study those species that we most often encounter. n_{OBIS} is an inexact measure of geographic extent, but this result does suggest that knowledge is likely to be especially scarce for rare species (see also Gaston & Kunin, 1997b). The consequence of this for our understanding of ecosystems may depend on the extent to which common species dominate community properties and ecosystem functioning (e.g. Gaston & Fuller, 2008;

Table 3 Generalized linear models of biological knowledge (number of biological traits for which we have data) of the UK marine fauna as a function of biogeographic knowledge (n_{OBIS}), conservation concern, and commercial importance (Model 1). n is the number of species in each group. Model coefficients (b_x) for each predictor x are given with their 95% CIs, as is the deviance explained by the model (Pseudo R^2), and an estimate of taxonomic signal in the residuals (λ) which equals 0 when residuals are independent of taxonomy, and has a maximum value of 1 (see text for details). Models were run across all taxa, separately for all invertebrates, and all fish, and at the level of taxonomic class for those classes with sufficient (≥ 10) species. When λ indicated taxonomic signal in the class-level models, additional models were run at the order level (in italics in the Class column). Conservation concern and commercial importance were only included as predictors for groups in which variance in these parameters was sufficient for meaningful estimation.

Clade	Class	n	$b_{n(OBIS)}$	$b_{conservation}$	$b_{commercial}$	Pseudo R^2	λ
All taxa		973	**0.49 (0.42–0.56)*****	**1.43 (0.84–2.05)*****	**1.24 (0.94–1.54)*****	0.36	0.31
All invertebrates		825	**0.62 (0.53–0.71)*****	**1.79 (1.05–2.56)*****	**0.52 (0.06–0.98)***	0.26	0.22
	Polychaeta	187	**0.86 (0.65–1.08)*****			0.31	0.19
	Phyllodicida	74	**0.80 (0.52–1.10)*****			0.35	< 0.0001
	Terebellida	31	**1.25 (0.55–2.17)*****			0.44	< 0.0001
	Spionida	25	**1.34 (0.55–2.40)****			0.34	< 0.0001
	Sabellida	20	**1.11 (0.33–2.22)***			0.31	0.91†
	Eunicida	12	0.66 (−0.200–1.89)			0.21	< 0.0001
	Malacostraca	154	**0.49 (0.31–0.67)*****		0.44 (−0.30–1.17)	0.23	0.28
	Amphipoda	55	**0.51 (0.21–0.85)****			0.18	0.81†
	Decapoda	78	**0.44 (0.22–0.68)*****		0.38 (−0.46–1.21)	0.25	< 0.0001
	Maxillopoda	10	0.75 (−0.25–2.21)			0.19	< 0.0001
	Gymnolaemata	45	**0.81 (0.19–1.55)****			0.16	< 0.0001
	Ascidiacea	33	**0.81 (0.21–1.51)****			0.22	< 0.0001
	Anthozoa	40	**0.70 (0.35–1.10)*****	1.44 (−0.01–2.91)		0.30	< 0.0001
	Hydrozoa	38	0.35 (−0.03–0.77)			0.09	< 0.0001
	Asteroidea	15	0.49 (−0.20–1.36)			0.11	< 0.0001
	Ophiuroidea	13	**1.37 (0.56–2.58)****			0.55	< 0.0001
	Gastropoda	85	**0.58 (0.36–0.82)*****			0.28	0.49
	Nudibranchia	25	−0.04 (−0.17–0.09)			0.01	< 0.0001
	Littorinimorpha	17	0.28 (−0.20–0.92)			0.07	0.89†
	Neogastropoda	12	**1.00 (0.51–1.75)*****			0.70	< 0.0001
	Bivalvia	75	**0.79 (0.49–1.12)*****		0.34 (−0.54–1.22)	0.33	0.49
	Veneroida	34	**0.67 (0.28–1.12)****			0.25	0.95†
	Cephalopoda	11	−0.06 (−0.19–0.08)			0.04	< 0.0001‡
	Demospongiae	42	**0.59 (0.21–1.03)****			0.22	0.306§
All fish		148	**0.16 (0.08–0.24)*****	0.63 (0.03–1.31)	**0.40 (0.21–0.59)*****	0.35	0.154
	Actinopterygii	132	**0.16 (0.07–0.25)****	**1.06 (0.18–2.16)***	**0.41 (0.20–0.62)*****	0.36	< 0.0001
	Elasmobranchii	15	**0.22 (0.12–0.33)****	**0.72 (0.28–1.16)***	**0.46 (0.19–0.74)***	0.81	< 0.0001

***$P < 0.001$; **$P < 0.01$; *$P < 0.05$. Statistically significant results are shown in bold. Note that because models were fitted with quasibinomial errors, we used F-tests to test for the significance of parameter estimates (Faraway, 2006) which may not always agree entirely with the profile likelihood-based confidence intervals.

†High values of λ in order-level analyses are indicative of interesting patterns at lower taxonomic levels, but sample sizes and taxonomic resolution are insufficient to examine these statistically.

‡One species was identified as influential (high Cook's D), but results are unaffected by removal of this species.

§λ reduces to < 0.0001 on removal of a single influential (high Cook's D) species, but other results are unaffected.

Gaston, 2010). If certain combinations of traits were considered to be key to ecosystem integrity, they could perhaps best be maintained by directing management efforts at common species which possess these traits (Gaston & Fuller, 2008). However, it is worth noting that rare species may possess unique (possibly functionally important) traits or combinations of traits (Gaston & Kunin, 1997a); the poor state of biological knowledge for rare species that we document here means we simply do not know if this is the case for UK marine species.

Species of conservation concern were better known across both invertebrates and fish. Given that conservation listing is generally of rare species (e.g. Purvis *et al.*, 2005), about which our results have suggested we tend to know least, this result may seem counter-intuitive. The likely explanation is that a certain amount of knowledge is required before a species can be listed. Thus, while only 2% of invertebrates in our dataset are currently of conservation concern, this increases to 9% when we consider only those species that are biologically well known (i.e. have data available for seven or eight traits). This suggests that many of those species without conservation listing – especially those rarely sampled – should be considered unknown, rather than safe. It is noteworthy, too, that all 25 species of fish listed as being
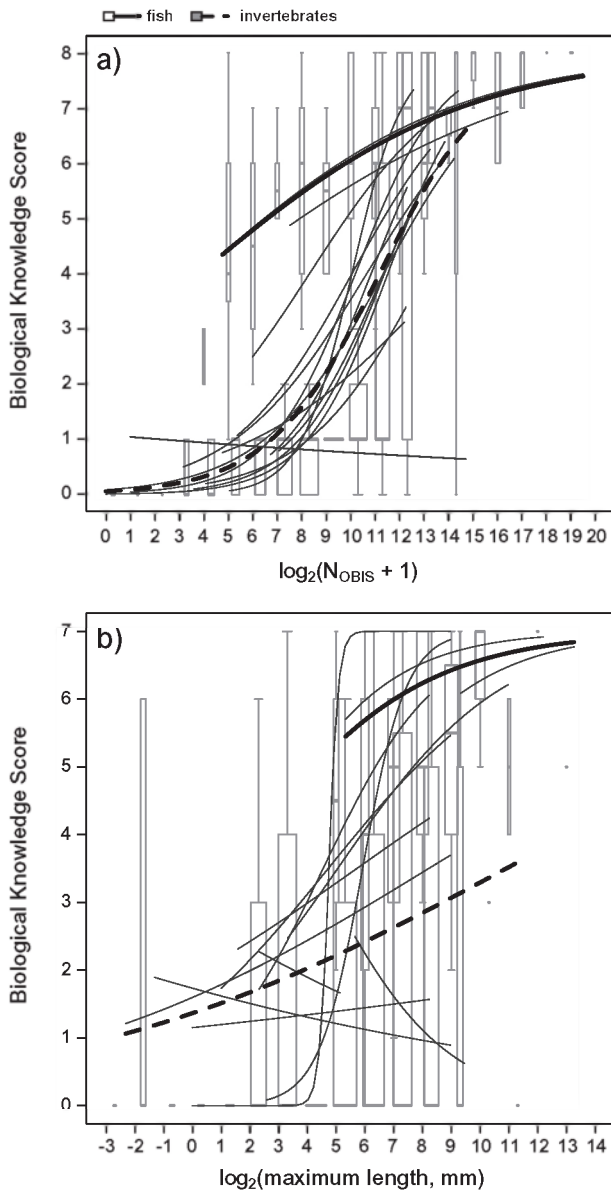
Figure 2 Biological knowledge increases with biogeographic knowledge in both fish and invertebrates (a) whereas body size is only significant for a small proportion of invertebrate groups (b). The number of records in OBIS (n_{OBIS}) and body size are placed in $\log_2$ bins for clarity of presentation. Boxes extend to the inter-quartile range around the median value, the whiskers extend a further 1.5× the inter-quartile range from the boxes, and the width of the boxes is proportional to the number of species in each category. Lines are fitted generalized linear models of biological knowledge as a function of (a) $\log(n_{OBIS} + 1)$ and (b) $\log$(body size), showing results across all fish (thick solid lines), all invertebrates (thick dashed lines) and separately for each sufficiently species-rich class (fine solid lines). Note that data availability varies from zero to eight traits in (a), but only from zero to seven traits in (b), because body size is used as a predictor so cannot contribute to the response. For the same reason, there are 973 species in (a) and 775 species in (b). Results of detailed statistical analyses of these patterns are presented in Tables 3 & 4.
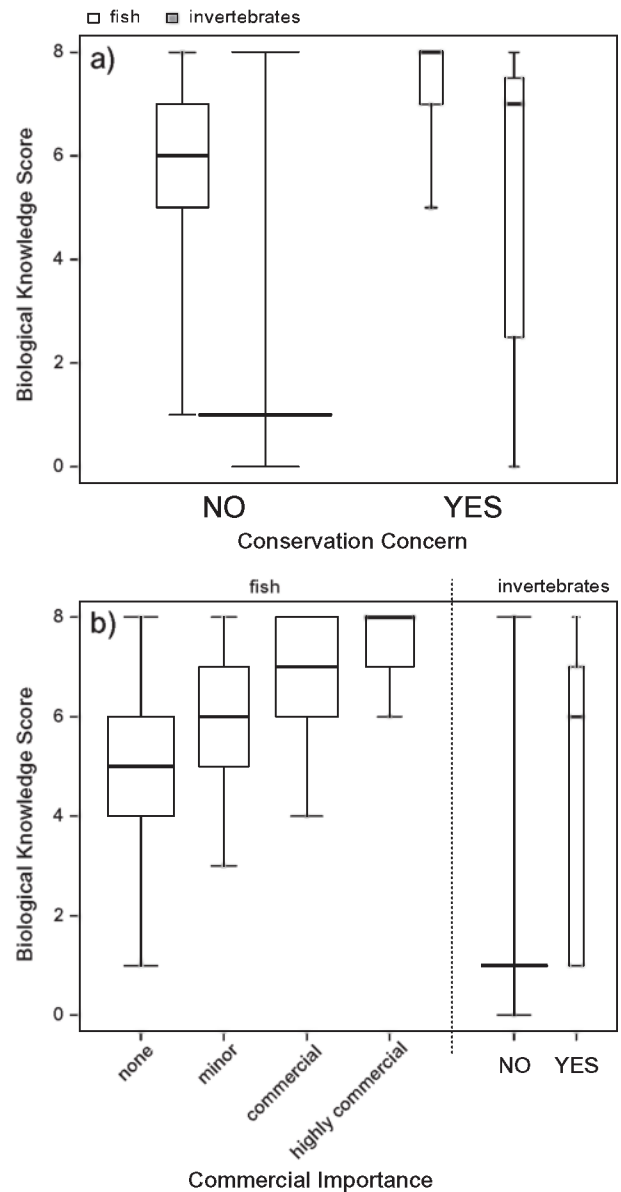
Figure 3 Biological knowledge is higher in species of conservation concern (a) and in commercially exploited species (b) in both fish and invertebrates. The number of records in OBIS (n_{OBIS}) and body size are placed in $\log_2$ bins for clarity of presentation. Boxes extend to the inter-quartile range around the median value, the whiskers extend a further 1.5× the inter-quartile range from the boxes, and the width of the boxes is proportional to the number of species in each category. Commercial importance was scored on a four-point scale for fish, but only on a binary scale for invertebrates. Results of detailed statistical analyses of these patterns are presented in Tables 3 & 4.

of conservation concern are also fished commercially, 13 of them to a major extent. In most cases, fishing is the cause of the conservation listing; it is striking that species that would be afforded protection on land, remain heavily exploited in the sea.

Dealing with the large knowledge gaps identified here is not straightforward. As Dolédec *et al.* (1999) point out, filling these

Table 4 Generalized linear models of biological knowledge of the UK marine fauna as a function of biogeographic knowledge (n_{OBIS}), maximum body size, conservation concern and commercial importance (Model 2). n is the number of species in each group. Model coefficients (b_x) for each predictor x are given with their 95% CIs, as is the deviance explained by the model (Pseudo R^2), and an estimate of taxonomic signal in the residuals (λ) which equals 0 when residuals are independent of taxonomy, and has a maximum value of 1 (see text for details). Models were run across all taxa, separately for all invertebrates, and all fish, and at the level of taxonomic class for those classes with sufficient (≥ 10) species. When λ indicated taxonomic signal in the class-level models, additional models were run at the order level (in italics in the Class column). Conservation concern and commercial importance were only included as predictors for groups in which variance in these parameters was sufficient for meaningful estimation. Methodology is similar to that of Model 1 (Table 2), but in Model 2 body size was included as a predictor, and removed from calculations of the biological knowledge of each species, and so models were fitted using only those species for which body size is known.

Clade	Class	n	$b_{n(OBIS)}$	$b_{body\ size}$	$b_{conservation}$	$b_{commercial}$	Pseudo R^2	λ
All taxa		775	**0.55 (0.44–0.65)***	**0.18 (0.08–0.28)***	**1.56 (0.78–2.42)***	**0.84 (0.43–1.25)***	0.31	0.33
All invertebrates		627	**0.86 (0.70–1.04)***	0.05 (–0.07–0.18)	**2.42 (1.24–3.74)***	0.21 (–0.47–0.88)	0.24	0.18
	Polychaeta	99	**0.78 (0.45–1.15)***	0.11 (–0.15–0.40)			0.21	<0.0001
	Malacostraca	141	**0.83 (0.49–1.22)***	**0.40 (0.01–0.81)***		–0.69 (–2.24–0.78)	0.24	<0.0001
	Amphipoda	51	**0.85 (0.24–1.62)***	**1.10 (0.08–2.22)***			0.24	0.69†
	Decapoda	77	**0.71 (0.30–1.20)***	**0.76 (0.06–1.51)***		–0.84 (–2.64–0.82)	0.30	<0.0001
	Gymnolaemata	20	0.71 (–0.21–1.98)	–0.19 (–0.60–0.19)			0.13	<0.0001
	Ascidiacea	29	**1.64 (–0.17–5.06)***	–0.27 (–1.68–1.27)			0.27	0.231
	Enterogona‡	21	**5.27 (2.64–10.13)***	–0.17 (–1.46–0.99)			0.63	0.84
	Anthozoa	34	0.54 (–0.10–1.29)	0.51 (–0.15–1.26)			0.20	<0.0001
	Hydrozoa	35	–0.46 (–1.44–0.46)	**9.11 (6.78–12.03)***			0.90	<0.0001
	Asteroidea	14	0.51 (–0.47–1.69)	–0.95 (–3.07–0.69)			0.11	<0.0001
	Ophiuroidea	12	**2.34 (0.71–5.08)***	0.05 (–2.06–2.08)			0.51	<0.0001
	Gastropoda	74	**1.10 (0.47–1.92)***	**0.90 (0.08–1.87)***			0.35	0.71
	Nudibranchia‡	24						
	Littorinimorpha	17	0.60 (–0.20–1.91)	**1.97 (0.05–4.87)***			0.32	0.95†
	Bivalvia	67	**1.05 (0.58–1.59)***	0.21 (–0.36–0.81)		0.05 (–1.35–1.41)	0.28	0.415
	Veneroida	34		**1.07 (0.42–1.90)***	**0.96 (0.10–2.02)***			0.32†
	Demospongiae‡	29	0.64 (–0.46–1.88)	**1.90 (–0.11–5.00)***			0.24	0.36
All fish		148	**0.19 (0.10–0.28)***	–0.15 (–0.38–0.08)	**0.74 (0.09–1.46)***	**0.49 (0.26–0.72)***	0.35	0.08
	Actinopterygii	132	**0.17 (0.07–0.27)***	0.01 (–0.28–0.30)	**1.08 (0.16–2.22)***	**0.42 (0.17–0.39)***	0.36	<0.0001
	Elasmobranchii	15	**0.25 (0.14–0.36)***	0.16 (–0.22–0.56)	**0.71 (0.23–1.20)***	**0.46 (0.14–0.79)***	0.83	<0.0001§

***P < 0.001; **P < 0.01; *P < 0.05. Statistically significant results are shown in bold. Note that because models were fitted with quasibinomial errors, we used F tests to test for the significance of parameter estimates (Faraway, 2006) which may not always agree entirely with the profile likelihood-based confidence intervals.

†High values of λ in order-level analyses are indicative of interesting patterns at lower taxonomic levels, but sample sizes and taxonomic resolution are insufficient to examine these statistically.

‡In these groups, limited variation in the response variable (biological knowledge) suggests caution in interpretation of models, even when they could be fitted (biological knowledge is > 0 in 3/21 species of Enterogona, 0/24 species of Nudibranchia, 3/29 species of Demospongiae).

§One species is identified as influential (high Cook's D; this is the largest species in our database (the basking shark *Cetorhinus maximus*). Excluding it causes $b_{body\ size}$ to approach significance ($P = 0.0593$), commercial importance to lose significance ($P = 0.0932$), and pseudo R^2 to increase (to 0.88), but has little effect on parameter estimates.

gaps, even for a relatively small and taxonomically uniform assemblage such as the freshwater invertebrates of the Upper Rhone river, is a major undertaking (in that example, around 60 person-years). Previous approaches have therefore attempted to deal with gaps by predicting values based on the degree of variability within a taxonomic group (Foggo *et al.*, 2007; Matthew *et al.*, 2009), inputting zero values (Dolédec *et al.*, 1999; Bremner *et al.*, 2006) or deleting whole species or traits (Baird & van den Brink, 2007). None of these is ideal, and all may require the development of novel statistical methods. The validity of predicting traits from congeners or confamilials has not formally been tested, but requires accurate information on the evolutionary relationships between species, and sufficient trait data to establish general patterns of phylogenetic conservatism in different traits. Using taxonomic relationships as a surrogate for phylogeny, we estimated a λ value of 0.98 for maximum body size. This suggests that estimating the size of a species based on that of its close relatives is probably justified, but it is unclear whether this result would generalize to other traits. Replacing missing values with zeros may break up co-evolved trait complexes, and cause well-known traits (e.g. body size) to appear to be the defining ones, simply because analyses on other traits lack statistical power. Deleting cases (species or traits which include missing values) reduces degrees of freedom and thus statistical power. Trait-based analyses should consider adopting standard methods for dealing with missing data such as multiple imputation (e.g. Nakagawa & Freckleton, 2008), especially because missing data are unlikely to be randomly distributed across the assemblage. In our data, for example, there is significant taxonomic signal in both biogeographic knowledge (n_{OBIS}, $\lambda = 0.43$) and biological knowledge ($\lambda = 0.37$). It remains unclear, however, how robust techniques for dealing with missing data are in situations in which available data are substantially outnumbered by gaps.

Incorporating biological traits into macroecological analyses holds great promise for revealing the mechanistic basis behind emergent patterns and relationships (e.g. Webb *et al.*, 2009; Buckley & Freckleton, 2010; Verberk *et al.*, 2010) and can provide alternative indicators of environmental change (Bremner, 2008; Baiser & Lockwood, 2011). In this first attempt to catalogue the extent of biological knowledge of an entire fauna, however, we have documented severe shortages of data for most biological traits. We expect even greater knowledge gaps in other, less-studied, marine faunas. In terrestrial and freshwater systems, trait databases tend to concentrate on specific taxonomic groups (Poff *et al.*, 2006; Naeem & Bunker, 2009) we therefore expect that multiphylum studies in most systems will suffer from similar shortcomings to those we have documented here for the UK demersal marine fauna.

Filling gaps in our knowledge will require both additional data collection and development of the statistical techniques for estimating missing trait values. In addition we agree with Naeem & Bunker (2009) that a concerted effort is required to collate existing trait data and to make them available in a standard format – possibly using existing portals, such as Sealifebase (http://www.sealifebase.org) for marine species. Recent international initiatives such as OBIS (see also Somerfield *et al.*, 2009) have vastly expanded our knowledge of where marine species occur. Finding out what they actually *do* remains a major challenge.

REFERENCES

Baird, D.J. & van den Brink, P.J. (2007) Using biological traits to predict species sensitivity to toxic substances. *Ecotoxicology and Environmental Safety*, **67**, 296–301.

Baiser, B. & Lockwood, J.L. (2011) The relationship between functional and taxonomic homogenization. *Global Ecology and Biogeography*, **20**, 134–144.

Barne, J.H., Robson, C.F., Kaznowska, S.S., Doody, J.P., Davidson, N.C. & Buck, A.L. (1997) *Coasts and seas of the United Kingdom* [series]. Joint Nature Conservation Committee, Peterborough.

Blanchard, J.L., Jennings, S., Law, R., Castle, M.D., Mccloghrie, P., Rochet, M.-J. & Benoît, E. (2008) How does abundance scale with body size in coupled size-structured food webs? *Journal of Animal Ecology*, **78**, 270–280.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.-S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, **24**, 127–135.

Bremner, J. (2008) Species' traits and ecological functioning in marine conservation and management. *Journal of Experimental Marine Biology and Ecology*, **366**, 37–47.

Bremner, J., Rogers, S.I. & Frid, C.L.J. (2006) Methods for describing ecological functioning of marine benthic assemblages using biological traits analysis (BTA). *Ecological Indicators*, **6**, 609–622.

Brown, J.H., Gillooly, J.F., Allen, A.P., Savage, V.M. & West, G.B. (2004) Toward a metabolic theory of ecology. *Ecology*, **85**, 1771–1789.

Buckley, H.L. & Freckleton, R.P. (2010) Understanding the role of species dynamics in abundance–occupancy relationships. *Journal of Ecology*, **98**, 645–658.

Chevenet, F., Dolédec, S. & Chessel, D. (1994) A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology*, **31**, 295–309.

Costello, M.J., Coll, M., Danovaro, R., Halpin, P., Ojaveer, H. & Miloslavich, P. (2010) A census of marine biodiversity knowledge, resources, and future challenges. *PLoS ONE*, **5**, e12110.

Dolédec, S., Statzner, B. & Bournard, M. (1999) Species traits for future biomonitoring across ecoregions: patterns along a human-impacted river. *Freshwater Biology*, **42**, 737–758.

Edwards, M. & Richardson, A.J. (2004) Impact of climate change on marine pelagic phenology and trophic mismatch. *Nature*, **430**, 881–884.

Faraway, J.J. (2005) *Linear models with R*. Chapman and Hall/CRC, Boca Raton, FL.

Faraway, J.J. (2006) *Extending the linear model with R*. Chapman and Hall/CRC, Boca Raton, FL.

Foggo, A., Bilton, D.T. & Rundle, S.D. (2007) Do developmental mode and dispersal shape abundance–occupancy relationships in marine macroinvertebrates? *Journal of Animal Ecology*, **76**, 695–702.

Freckleton, R.P., Harvey, P.H. & Pagel, M. (2002) Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist*, **160**, 712–726.

Freckleton, R.P., Gill, J.A., Noble, D. & Watkinson, R. (2005) Large-scale population dynamics, abundance–occupancy relationships and the scaling from local to regional population size. *Journal of Animal Ecology*, **74**, 353–364.

Froese, R. & Pauly, D. (eds) (2010) *Fishbase*. Available at: http://www.fishbase.org (accessed 12 August 2010).

Gaston, K.J. (2010) Valuing common species. *Science*, **327**, 154–155.

Gaston, K.J. & Blackburn, T.M. (2000) *Patterns and processes in macroecology*. Blackwell Science, Oxford.

Gaston, K.J. & Fuller, R.A. (2008) Commonness, population depletion and conservation biology. *Trends in Ecology and Evolution*, **23**, 14–19.

Gaston, K.J. & Kunin, W.E. (1997a) Rare–common differences: an overview. *The biology of rarity* (ed. by W.E. Kunin and K.J. Gaston), pp. 12–29. Chapman and Hall, London.

Gaston, K.J. & Kunin, W.E. (1997b) Concluding comments. *The biology of rarity* (ed. by W.E. Kunin and K.J. Gaston), pp. 262–272. Chapman and Hall, London.

Gelman, A. & Hill, J. (2007) *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge.

Hadfield, J.D. & Nakagawa, S. (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, **23**, 494–508.

Hayward, P.J. & Ryland, J.S. (1991a) *The marine fauna of the British Isles and north-west Europe. Volume I: introduction and protozoans to arthropods*. Oxford University Press, Oxford.

Hayward, P.J. & Ryland, J.S. (1991b) *The marine fauna of the British Isles and north-west Europe. Volume II: molluscs to chordates*. Oxford University Press, Oxford.

Hildrew, A., Raffaelli, D. & Edmonds-Brown, R. (2007) *Body size: the structure and function of aquatic ecosystems*. Cambridge University Press, Cambridge.

Jablonski, D. (1986) Larval ecology and macroevolution in marine invertebrates. *Bulletin of Marine Science*, **39**, 565–587.

Jennings, S., Reynolds, J.D. & Mills, S.C. (1998) Life history correlates of responses to fisheries exploitation. *Proceedings of the Royal Society B: Biological Sciences*, **265**, 333–339.

Lavorel, S., Grigulis, K., Lamarque, P., Colace, M.-P., Garden, D., Girel, J., Pellet, G. & Douzet, R. (2010) Using plant functional traits to understand the landscape distribution of multiple ecosystem services. *Journal of Ecology*, **99**, 135–147.

McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. (2006) Rebuilding community ecology from functional traits. *Trends in Ecology and Evolution*, **21**, 178–185.

Matthew, R.H., Wendel, K., Paterson, M.J., Yan, N.D., Cannon, C.H. & Rusak, J.A. (2009) Communities contain closely related species during ecosystem disturbance. *Ecology Letters*, **13**, 162–174.

Menezes, S., Baird, D.J. & Soares, A.M.V.M. (2010) Beyond taxonomy: a review of macroinvertebrate trait-based community descriptors as tools for freshwater biomonitoring. *Journal of Applied Ecology*, **47**, 711–719.

Naeem, S. & Bunker, D.E. (2009) TraitNet: furthering biodiversity research through the curation, discovery, and sharing of species trait data. *Biodiversity, ecosystem functioning and human wellbeing* (ed. by D.E.B. Shahid Naeem, D. Bunker, A. Hector, M. Loreau and C. Perrings), pp. 281–289. Oxford University Press, Oxford.

Nakagawa, S. & Freckleton, R.P. (2008) Missing inaction: the dangers of ignoring missing data. *Trends in Ecology and Evolution*, **23**, 592–596.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Simpson, G.L., Solymos, P., Stevens, M.H.H. & Wagner, H. (2009) *Vegan: community ecology package. R package version 1.15-4*. Available at: http://CRAN.R-project.org/package=vegan.

Olden, J.D., Hogan, Z.S. & Vander Zanden, M.J. (2007) Small fish, big fish, red fish, blue fish: size-biased extinction risk of the world's freshwater and marine fishes. *Global Ecology and Biogeography*, **16**, 694–701.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T. & Fritz, S. (2009) *CAIC: comparative analyses using independent contrasts. R package version 1.0.4-94/r94*. Available at: http://R-Forge.R-project.org/projects/caic/.

Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Petchey, O.L. & Gaston, K.J. (2006) Functional diversity: back to basics and looking forward. *Ecology Letters*, **9**, 741–758.

Poff, N.L., Olden, J.D., Vieira, N.K.M., Finn, D.S., Simmons, M.P. & Kondratieff, B.C. (2006) Functional trait niches of North American lotic insects: traits-based ecological applications in light of phylogenetic relationships. *Journal of the North American Benthological Society*, **25**, 730–755.

Purvis, A., Cardillo, M., Grenyer, R. & Collen, B. (2005) Correlates of extinction risk: phylogeny, biology, threat and scale. *Phylogeny and conservation* (ed. by A. Purvis, J.L. Gittleman and T. Brooks), pp. 295–316. Cambridge University Press, Cambridge.

R Development Core Team (2009) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Raffaelli, D. (2007) Food webs, body size and the curse of the Latin binomial. *From energetics to ecosystems: the dynamics and structure of ecological systems* (ed. by N. Rooney, K. Mccann and D. Noakes), pp. 53–64. Springer, Dordrecht.

Schleuter, D., Daufresne, M., Massol, F. & Argillier, C. (2010) A user's guide to functional diversity indices. *Ecological Monographs*, **80**, 469–484.

Solan, M., Cardinale, B.J., Downing, A.L., Engelhardt, K.A.M., Ruesink, J.L. & Srivastava, D.S. (2004) Extinction and ecosystem function in the marine benthos. *Science*, **306**, 1177–1180.

Somerfield, P.J., Arvanitidis, C. & Vanden-Berghe, E. (2009) Large-scale studies of the European benthos: the MacroBen database. *Marine Ecology Progress Series*, **382**, 221–311.

Thurstan, R.H., Brockington, S. & Roberts, C.M. (2010) The effects of 118 years of industrial fishing on UK bottom trawl fisheries. *Nature Communications*, **1**, 1–6.

Tillin, H.M., Hiddink, J.G., Jennings, S. & Kaiser, M.J. (2006) Chronic bottom trawling alters the functional composition of benthic invertebrate communities on a sea-basin scale. *Marine Ecology Progress Series*, **318**, 31–45.

Verberk, W.C.E.P., van der Velde, G. & Esselink, H. (2010) Explaining abundance–occupancy relationships in specialists and generalists: a case study on aquatic macroinvertebrates in standing waters. *Journal of Animal Ecology*, **79**, 589–601.

Webb, T.J., Tyler, E.H.M. & Somerfield, P.J. (2009) Life history mediates large-scale population ecology in marine benthic taxa. *Marine Ecology Progress Series*, **396**, 293–306.

Webb, C.T., Hoeting, J.A., Ames, G.M., Pyne, M.I. & Poff, N.L. (2010) A structured and dynamic framework to advance traits-based theory and prediction in ecology. *Ecology Letters*, **13**, 267–283.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1 Spatial extent and sampling intensity of the surveys used to define the UK marine fauna.

Table S1 Summary of large-scale biodiversity surveys used to compile the UK marine species list.

Table S2 The 14 individual traits included in the study, with their definitions and reasons for inclusion.

Appendix S1 Data used in Tyler et al., *Biological Knowledge of the UK Marine Fauna*.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCH

This work forms part of ongoing efforts to incorporate information on the biological traits of marine species into macroecological investigations, led by T.J.W. at the University of Sheffield, UK and P.J.S. at the Plymouth Marine Laboratory, UK. It also builds on general interests in the development and applications of biological traits analysis led by J.B. Author contributions: E.H.M.T., P.J.S. and T.J.W. conceived the ideas; E.H.M.T., J.B., E.J., O.L. and M.L.D.P. collected or contributed traits data; E.V.B. contributed the OBIS analyses; T.J.W. and E.H.M.T. analysed the data; E.H.M.T. and T.J.W. led the writing, to which all co-authors also contributed.

Editor: Janne Soininen