MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 1909–1917 doi: 10.1111/j.1365-294X.2011.05383.x

Exploration of community traits as ecological markers in microbial metagenomes

ALBERT BARBERÁN,* ANTONI FERNÁNDEZ-GUERRA,* BRENDAN J. M. BOHANNAN† and EMILIO O. CASAMAYOR*

*Biogeodynamics & Biodiversity Group, Department of Continental Ecology, Centre d'Estudis Avançats de Blanes (CEAB-CSIC), Accés Cala St. Francesc 14, Blanes 17300, Spain, †Department of Biology, Center for Ecology and Evolutionary Biology, University of Oregon, 335 Pacific Hall, Eugene, OR 97403-5289, USA

Abstract

The rate of information collection generated by metagenomics is uncoupled with its meaningful ecological interpretation. New analytical approaches based on functional trait-based ecology may help to bridge this gap and extend the trait approach to the community level in vast and complex environmental genetic data sets. Here, we explored a set of community traits that range from nucleotidic to genomic properties in 53 metagenomic aquatic samples from the Global Ocean Sampling (GOS) expedition. We found significant differences between the community profile derived from the commonly used 16S rRNA gene and from the functional trait set. The traits proved to be valuable ecological markers by discriminating between marine ecosystems (coastal vs. open ocean) and between oceans (Atlantic vs. Indian vs. Pacific). Intertrait relationships were also assessed, and we propose some that could be further used as habitat descriptors or indicators of artefacts during sample processing. Overall, the approach presented here may help to interpret metagenomics data to gain a full understanding of microbial community patterns in a rigorous ecological framework.

Keywords: community ecology, functional traits, Global Ocean Sampling, metagenomics, microbial ecology

Received 22 July 2011; revision received 6 October 2011; accepted 25 October 2011

Introduction

In the field of community ecology, there is a resurging interest in understanding biogeographical patterns based on functional traits (i.e. biological characteristics linked to fitness; McGill *et al.* 2006; Kraft *et al.* 2008). The study of covarying traits in an environmental context is crucial to understand the ecological strategies that underlie community patterns (Green *et al.* 2008). In parallel, the new field of metagenomics is challenging the scientific community with an astonishing amount of complex data that intersect the disciplines of microbiology, genetics, ecology and bioinformatics (Handelsman 2004). Despite some computational advances, the analysis of community genomics data within a meaningful ecological framework remains an elusive goal (Raes *et al.* 2007a; Kunin *et al.* 2008). Metagenomics, however, produces data very suitable for extending traditional species-level functional-trait analyses (Wright *et al.* 2004) to the community level, resulting in an ecological approach to analysing metagenomic data that circumvent the confounding effects of horizontal gene transfer present at lower levels of organization (e.g. at the species or population level).

Recently, the Global Ocean Sampling (GOS) expedition (Rusch *et al.* 2007) has generated the largest marine metagenomic data set ever sampled along an environmental gradient, with approximately eight billion nucleotides present in more than 7 million DNA fragments. However, few attempts have been made to analyse this data within an ecological framework (Raes *et al.* 2011). A synergy between ecology and metagenomics may help bridge this gap, by providing theoretical and analytical tools that could unveil microbial community patterns and the processes that underlie them (Prosser *et al.* 2007).

Correspondence: Albert Barberán, Fax: +34 972 337806; E-mail: abarberan@ceab.csic.es

To achieve this objective, we have characterized a set of community traits in 53 GOS metagenomic samples taken from the near-surface marine planktonic environment. First, we tested the performance of each trait as a taxonomic, functional and habitat surrogate, respectively. Second, we compared the whole community profile derived from the commonly used 16S rRNA gene marker and from the functional trait set. Finally, we assessed intertrait relationships that could be further used as indicators of functional anomalies and/or for detection of artefacts during sample processing. Overall, the approach presented here is an important step towards developing taxonomic and functional analysis of metagenomic data in a rigorous ecological framework and to provide insights into community ecology beyond purely descriptive studies.

Materials and methods

Global Ocean Sampling metagenomic data

Unassembled genomic fragments (reads) from the GOS expedition (Rusch *et al.* 2007) were downloaded from the CAMERA database (Seshadri *et al.* 2007). We selected 53 surface water samples from picoplankton collected within the same size fraction (0.1–0.8 μm), and free of bacterial contamination during sample handling (DeLong 2005; see detailed information in Table S1, Supporting information). Based on current knowledge regarding the spatial and temporal scales of variation in marine microbial communities, a GOS sample represents approximately a week temporally, a few kilometres horizontally and a few metres vertically (Fuhrman 2009). The analysed metagenomic data set comprised approximately 8000 Mb contained in approximately 5 million reads.

Community traits calculation

Up to 15 traits were calculated with different level of complexity (see Table 1). First, three simple traits were calculated. Custom Perl scripts were used to calculate the GC content and its variance, whereas the odds ratio of dinucleotides was measured as previously described (Willner *et al.* 2009). Dinucleotides have been shown to perform better than tri- and tetranucleotides for explanation of habitat differences (Willner *et al.* 2009).

Next, we extended our approach to three additional traits that relied on the estimation of the number of genomes present in each metagenome. For the assessment of the effective genome size (EGS; Raes *et al.* 2007b), the number of rRNA genes per genome (Howard *et al.* 2008) and the number of genes per genome (Biers *et al.* 2009), we targeted 35 protein markers (see

Table 1 Summary of the metagenomic community traits explored

Trait	Mean ± SD PCA1*	Autocorrelation (Moran's *I*)	Taxonomic/functional (r_M)	Functional | taxonomic (r_M)	Coastal vs. Open ocean (ANOSIM *R*)	Atlantic vs. Indian vs. Pacific (ANOSIM *R*)
GC content	37.7 ± 4.2	0.08 (0.096)	0.09 (0.028)	**0.60 (0.001)**	**0.23 (0.001)**	**0.13 (0.004)**
Variance of GC content	82.17 ± 23.3	**0.18 (0.003)**	0.09 (0.055)	**0.30 (0.001)**	0.10 (0.026)	0.08 (0.022)
Dinucleotides	79.1%*	0.14 (0.018)	0.08 (0.048)	**0.68 (0.001)**	**0.23 (0.001)**	**0.14 (0.002)**
Effective genome size	1.8 ± 0.3	**0.16 (0.003)**	−0.08 (0.966)	**0.44 (0.001)**	0.06 (0.033)	**0.08 (0.009)**
Number of rRNA/genome	2.7 ± 0.6	**0.17 (0.006)**	−0.03 (0.748)	**0.20 (0.009)**	**0.15 (0.002)**	**0.18 (0.001)**
Number of genes/genome	1362 ± 260	**0.34 (0.001)**	**0.18 (0.002)**	−0.11 (0.991)	0.09 (0.036)	**0.42 (0.001)**
Codons	95.2%*	0.07 (0.134)	0.10 (0.032)	**0.69 (0.001)**	**0.21 (0.001)**	**0.13 (0.001)**
Amino acids	95.5%*	0.13 (0.020)	0.09 (0.026)	**0.69 (0.001)**	**0.24 (0.001)**	**0.17 (0.001)**
Acidic to basic amino acids ratio	0.86 ± 0.02	**0.27 (0.001)**	0.05 (0.120)	**0.36 (0.001)**	**0.32 (0.001)**	**0.24 (0.001)**
% of Transcriptional factors	$5{\cdot}10^{-3}$ ± $1{\cdot}10^{-3}$	−0.03 (0.805)	0.00 (0.522)	0.21 (0.011)	0.03 (0.123)	0.04 (0.083)
% of classified reads	65 ± 5	**0.18 (0.003)**	−0.02 (0.636)	**0.30 (0.003)**	0.01 (0.273)	0.06 (0.036)
Functional content	21.6%*	**0.49 (0.001)**	—	**1 (0.001)**	**0.25 (0.001)**	**0.19 (0.001)**
Functional diversity	5.5 ± 0.04	**0.21 (0.001)**	0.06 (0.097)	0.09 (0.094)	0.02 (0.709)	0.00 (0.459)
Taxonomic content	62.6%*	**0.49 (0.001)**	**1 (0.001)**	—	**0.31 (0.001)**	**0.52 (0.001)**
Taxonomic diversity	0.86 ± 0.13	**0.17 (0.007)**	**0.44 (0.001)**	−0.24 (1)	**0.17 (0.003)**	**0.30 (0.001)**
All community traits	40.1%*	**0.29 (0.001)**	**0.42 (0.001)**	**0.58 (0.001)**	**0.33 (0.001)**	**0.42 (0.001)**

Marked with an asterisk, the percentage of variation explained by the first principal component (PCA1) for the traits consisting of a multivariate matrix. Between parentheses, reported *P*-values. In bold, *P*-values < 0.01. *P*-values of partial Mantel correlations (r_M) and ANOSIM *R* statistic calculated after 999 permutations.

detailed information in Table S2, Supporting information) known to exist as single-copy genes, to be universally distributed along the tree of life, and that are likely recalcitrant to lateral gene transfer (Ciccarelli *et al.* 2006; Raes *et al.* 2007b; Wu & Eisen 2008).

Finally, a set of traits based on the functional annotation of the metagenomic reads was calculated. Automatic annotation of the reads and protein prediction were carried out using MG-RAST (Meyer *et al.* 2008), which removed strict duplicate reads. Codon and amino acid composition of the predicted proteins were calculated using the program *cusp* bundled in the EMBOSS package (Rice *et al.* 2000). The acidic (i.e. glutamic and aspartic acids) to basic (i.e. lysine, histidine and arginine) amino acids ratio (AB) was calculated following Rhodes *et al.* (2010). Functional content was based on the comparison against the SEED platform and reported at the subsystem level (Dinsdale *et al.* 2008). The SEED subsystems are manually curated collections of proteins with related functions (Overbeek *et al.* 2005). From the functional annotation, we used as traits the percentage of transcriptional factors (TF) and the percentage of SEED subsystems classified reads, both calculated over the reads predicted to be protein coding. Taxonomic content based on 16S rRNA genes was determined by comparing the reads against the Greengenes 16S rRNA gene database and reported at the order level (DeSantis *et al.* 2006). For each metagenome, the same parameters were used to ensure the congruity of subsequent analysis. Diversity of the taxonomic and functional contents was calculated using the Shannon index. To correct for unequal sample size, we report the mean of 1000 randomized subsamples.

Complex traits (i.e. taxonomic content, dinucleotides, codons, amino acids and functional content; see Table 1) were transformed by considering the projection on the first component of a principal component analysis (PCA).

Statistical analyses

To estimate the degree of spatial autocorrelation of the community traits, Moran's coefficient (*I*) was calculated. Partial Mantel tests were used to determine the correlation between the similarity of each trait and the taxonomic or functional community similarity. Additionally, analysis of similarities (ANOSIM) was used to test for significant differences within marine ecosystems (coastal vs. open ocean) and between oceans (Atlantic vs. Indian vs. Pacific). The ANOSIM *R* statistic is based on the difference of mean dissimilarity ranks between groups and within groups and ranges from 0 (no separation) to 1 (complete separation; Clarke 1993). To test for differences between habitats, PERmutational Multivariate ANOVA (PERMANOVA) was used (McArdle & Anderson 2001). To represent taxonomic and functional community similarity, we ran nonmetric multidimensional scaling using the Bray–Curtis distance metric after Hellinger standardization (Legendre & Gallagher 2001). All statistical analyses were carried out in the *R* environment (http://www.r-project.org) using the ape (Paradis *et al.* 2004) and vegan (Oksanen *et al.* 2008) packages.

Results and discussion

A more complete understanding of microbial processes and patterns is essential to understand ecosystem functions and to predict the Earth's response to global change (Fuhrman 2009). Community genomics is revealing an unprecedented level of microbial diversity and metabolic novelty in the world's oceans and is the most comprehensive approach currently used to reveal microbial processes and patterns in environmental samples (Handelsman 2004). To more completely understand these data in an ecological context, we analysed community-level functional traits in 53 selected metagenomic samples from the GOS expedition (Rusch *et al.* 2007; Table S1, Supporting information). The analyses produced (i) a defined set of community traits that serve as functional and ecological descriptors of the metagenomic samples; (ii) consistent relationships between traits that may be used for detection of irregularities and/or methodological artefacts and (iii) a different view on microbial communities based either on the taxonomic or on the functional content.

Community traits as functional descriptors of metagenomic samples

We estimated 15 community traits for each metagenomic sample (Table 1). To assess the performance of each of the selected traits as a community descriptor, we first tested their spatial autocorrelation. Most of the traits were positively spatially autocorrelated (i.e. closer communities tended to have more similar values), as expected for descriptors of ecological change (i.e. because the environment tends to be spatially autocorrelated). Taxonomic and functional composition showed the highest autocorrelation values (Table 1). Although both traits are subject to database scan biases, they summarize two key features of biological communities, that is, community identity and metabolic potential, respectively (Raes *et al.* 2007a).

The accuracy of the community traits used as descriptors of microbial metagenomes can be potentially related both to a truly functional cause (i.e. different metabolic potentials among different microbial

assemblages) or just an effect of community composition (i.e. different taxonomic/phylogenetic groups present in different samples). To distinguish these potential influences, we calculated the correlation between sample similarity for each trait with the taxonomic and the functional composition of each sample (separating the effects of possible intermatrix correlations with partial Mantel tests). Most of the traits showed a significant correlation with functional composition rather than taxonomic composition (Table 1), consistent with the hypothesis that they reflect functional differences, rather than just taxonomic differences, among samples. Specifically, GC content, dinucleotides and codon and amino acid compositions (all of them highly correlated) showed the strongest correlation (Table 1). Although nucleotidic signatures have been proven useful for the taxonomic assignment of individual genomic fragments (Teeling *et al.* 2004), some signatures have been successfully applied at the community level for ecological and environmental classification (Willner *et al.* 2009; Rhodes *et al.* 2010). For the taxonomic matrix, only taxonomic diversity and the number of genes per genome showed a significant (although weak) correlation (Table 1). Overall, the explored community traits were more correlated with the functional composition ($r_M = 0.58$) than with the taxonomic composition ($r_M = 0.42$).

We tested the ability of each trait to differentiate between coastal and pelagic communities and among communities from different oceans (Atlantic vs. Indian vs. Pacific). A subset of the community traits (GC content, dinucleotides, codon, amino acids, AB ratio and functional content) was able to clearly distinguish between coastal and open-ocean habitats (Table 1). A different subset of traits (number of rRNA per genome, number of genes per genome, taxonomic diversity and taxonomic composition) was effective at distinguishing oceanic origin (Table 1). In general, the community traits distinguished slightly better between coastal and pelagic samples (ANOSIM $R = 0.33$) than the taxonomic composition derived from the commonly used 16S rRNA gene marker (ANOSIM $R = 0.31$). Nevertheless, taxonomic composition was the best marker of oceanic origin (ANOSIM $R = 0.52$).

Differences between the taxonomic and functional contents

Sequence identity of the 16S rRNA gene has been shown to be related to the overall genomic content in individual genomes (Zaneveld *et al.* 2010). However, we observed substantial differences between taxonomic (based on the 16S rRNA gene) and functional (based on SEED subsystems) contents in our samples (Mantel test: $r_M = 0.36$, P-value < 0.01; Fig. 1). Taxonomic content primarily separated the nonoceanic samples (i.e. hypersaline, mangrove, freshwater and estuaries) from the marine plankton (Fig. 1A), while functional content distinguished communities with different metabolic potentials (Fig. 1B). The single sample from a harbour (GS149) provides a striking example of how taxonomic community composition and functional content can provide different perspectives of the same complex microbial assemblage. While taxonomically the harbour metagenome was closer to other marine samples, in terms of its functional content it was a unique sample separated from the rest (Fig. 1A, B). This observation may suggest new research on genomic adaptation in polluted environments and on the dynamic processes that shape microbial communities. For microbial ecologists, it is still an unsolved question whether communities adapt more efficiently modifying the genomic repertoire of their members (as has been shown under laboratory conditions; Sniegowski *et al.* 1997) or changing the taxonomic composition by ecological processes such as immigration and dispersal.

Estuaries are productive habitats at the interface of terrestrial and oceanic ecosystems where a mixture of freshwater and marine-specific microorganisms is present (Bouvier & del Giorgio 2002; Crump *et al.* 2004). Estuarine samples GS11 (Delaware Bay) and GS12 (Chesapeake Bay) were intermediate between the single freshwater metagenomic sample (GS20) and the marine samples, both taxonomically and functionally (Fig. 1A, B). Although GS11 and GS12 samples differed in temperature (11 and 3.2 °C, respectively) and chlorophyll concentration (4.8 and 21 mg/m^{-3}, respectively; Table S1, Supporting information), they were more similar in composition to each other than to the other estuary sample (GS06 from the Bay of Fundy), which was more similar to other marine metagenomes (Fig. 1A, B). The temperature and chlorophyll concentration of sample GS06 were very similar to sample GS11. However, other relevant environmental data such as salinity, a key parameter known to greatly affect microbial community composition (Lozupone & Knight 2007; Auguet *et al.* 2010), have not been reported for the GOS samples. Thus, we cannot rule out the possible effects of other unmeasured parameters that may explain the observed patterns in community composition. Additionally, GS06 is an estuary sample that also consistently differed in other community traits (Fig. 2); for example, it had a very low GC content variance (Fig. 3B) compared with the other estuary samples.

We observed that both functional and taxonomic community patterns significantly varied between coastal and open ocean waters (ANOSIM $R = 0.25$ and 0.31, respectively). Coastal and open-ocean sites contain water masses with contrasting physicochemical

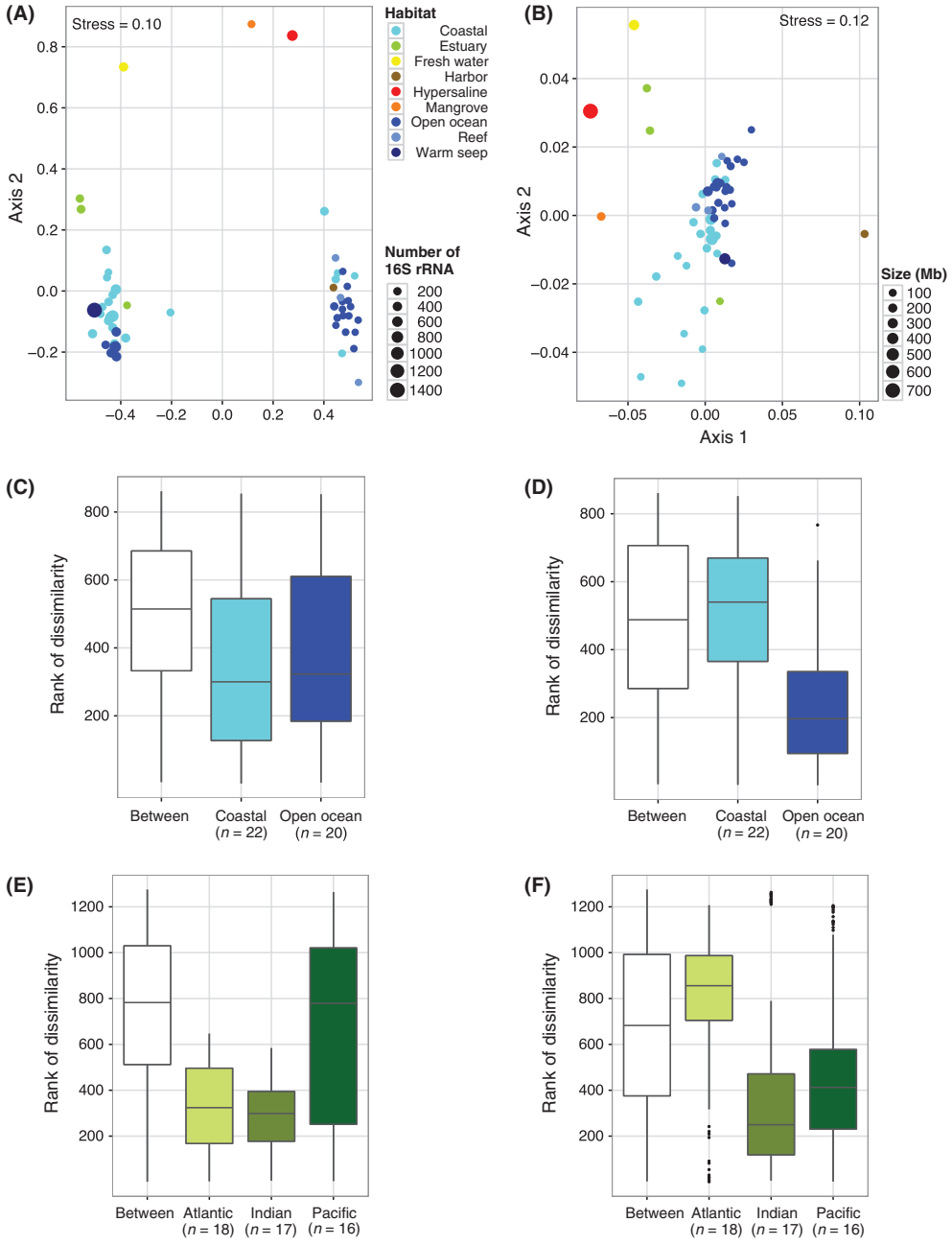

Fig. 1 Differences between taxonomic and functional community similarity matrices (β-diversity patterns). (A and B) Nonmetric multidimensional ordination plots for the taxonomic and functional matrices, respectively. Stress values are indicated. (C and D) Rank of dissimilarities between groups and within marine ecosystems (Coastal vs. Open ocean) for the taxonomic and functional matrices, respectively. (E and F) Rank of dissimilarities between groups and within oceans (Atlantic vs. Indian vs. Pacific) for the taxonomic and functional matrices, respectively.

characteristics, and several studies have shown different microbial composition (Acinas *et al.* 1997; Baltar *et al.* 2008). Although the ANOSIM R values were similar, functional content clustered all the open-ocean samples together (Fig. 1C, D). Taxonomic content differentiated better among oceans (ANOSIM $R = 0.52$) than the functional content ($R = 0.19$). Taxonomically, the samples from the Pacific Ocean were more heterogeneous (a few samples were more similar to the Atlantic and others to the Indian Ocean), while functionally, the Atlantic Ocean was the most heterogeneous (Fig. 1E, F; heterogeneous groups of samples show comparable mean rank of dissimilarities to the 'Between' category).

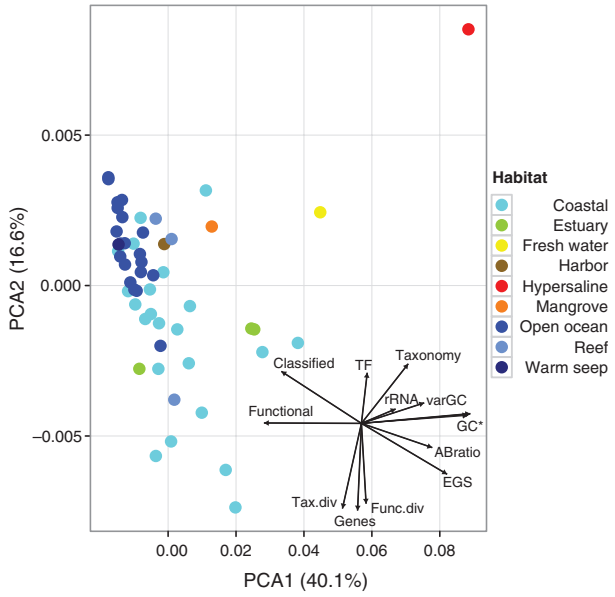

Fig. 2 Principal component analysis (PCA) ordination plot. In insert, variable loadings centred on (0,0). GC* refers to the highly correlated GC, dinucleotides, codon and amino acids traits. The variation explained by the two-first components is indicated on the axes.

Relationships among community traits

We used a PCA to determine the relationship among the community traits and to test how well they could discriminate among samples from different habitats (Fig. 2). Previous work on bacterial (Lozupone & Knight 2007) and archaeal (Auguet *et al.* 2010) community patterns based on the 16S rRNA gene indicated salinity as the major driving force at the global scale. In the ordination plot of community traits, the hypersaline (GS33) and freshwater (GS20) samples clustered away from the remaining samples (Fig. 2). The community traits distinguished better among samples from different habitats (PERMANOVA: $r^2 = 0.49$, P-value < 0.001) than taxonomic composition based on the 16S rRNA gene (PERMANOVA: $r^2 = 0.40$, P-value < 0.001).

Assessing bivariate relationships may also help to define ecological strategies across community axes of variation (Wright *et al.* 2004). All bivariate Spearman's rank correlations between community traits are detailed in Table S3, Supporting information. A few noteworthy outliers deserve further attention (Fig. 3). Although a positive relationship (Spearman's $\rho = 0.78$, P-value < 0.01) between the GC content and its variance was detected as a general trend (Fig. 3B), the hypersaline sample (GS33) clearly deviated, showing a high GC content with low variance (Fig. 3B). This may reflect a constraining effect of extreme environments at the community level captured in the nucleotide composition, in agreement with the content in the individual genomes reported for hypersaline inhabitants such as the Sphingobacteria *Salinibacter ruber* (% GC = 66.1) and the Euryarchaeota *Haloquadratum walsbyi* (% GC = 47.9). Further investigations are needed to confirm whether or not this is specifically related to salinity or it can be extrapolated to other extreme environments such as hydrothermal vents or hot springs. Initial analyses using statistical physics methods point to a significant effect of the ecological lifestyle and the composition of functional genes on long-range correlation structure in microbial genomes (Garcia *et al.* 2008, 2011). Overall, the GC content appeared as a very convenient parameter for initial exploration of metagenomic samples (Foerstner *et al.* 2005) owing to its straightforward calculation and higher correlation with other properties at higher organizational levels such as dinucleotides, codons, amino acids and functional content (as already known for bacterial genomes) because of the highly dependence on nucleotidic composition (Binnewies *et al.* 2006; see Table S3, Supporting information, and Fig. 3A for an example of correlation with the amino acid composition).

A general negative trend (Spearman's $\rho = -0.60$, P-value < 0.01) between the per cent of TF and EGS (Raes *et al.* 2007b) was observed (Fig. 3C). It has been shown that the number of genes in functional categories scales as a power law of the genome size (van Nimwegen 2003). The freshwater sample (GS20) appeared as an outlier, with a large proportion of TF relative to the remaining metagenomic samples (Fig. 3C). It has been proposed that TF could be an indicator of environmental variability because transcription factors are more strongly selected in variable than in constant environments (Parter *et al.* 2007). Lakes are small closed systems, highly diverse and more sensitive to environmental changes than the ocean, and thus promoters of higher microbial diversity (Auguet *et al.* 2010; Barberán & Casamayor 2010, 2011; Barberán *et al.* 2011). More metagenomic samples from different environments and particularly, freshwater samples are certainly needed to confirm this observation.

Another interesting relationship that still needs to be fully explained was the lower percentage of classified reads observed in metagenomic samples with bigger EGS (Fig. 3D; Spearman's $\rho = -0.86$, P-value < 0.01). This may result from a truly functional relationship or to sampling bias (the genomes available in public databases may under-represent microorganisms with bigger genomes and larger genomes contain more orphan genes; Skovgaard *et al.* 2001) or to larger percentage of picoeukaryotes or phages, which are much less characterized.

Finally, rRNA copy number is a trait that had previously attracted considerable attention because it reflects

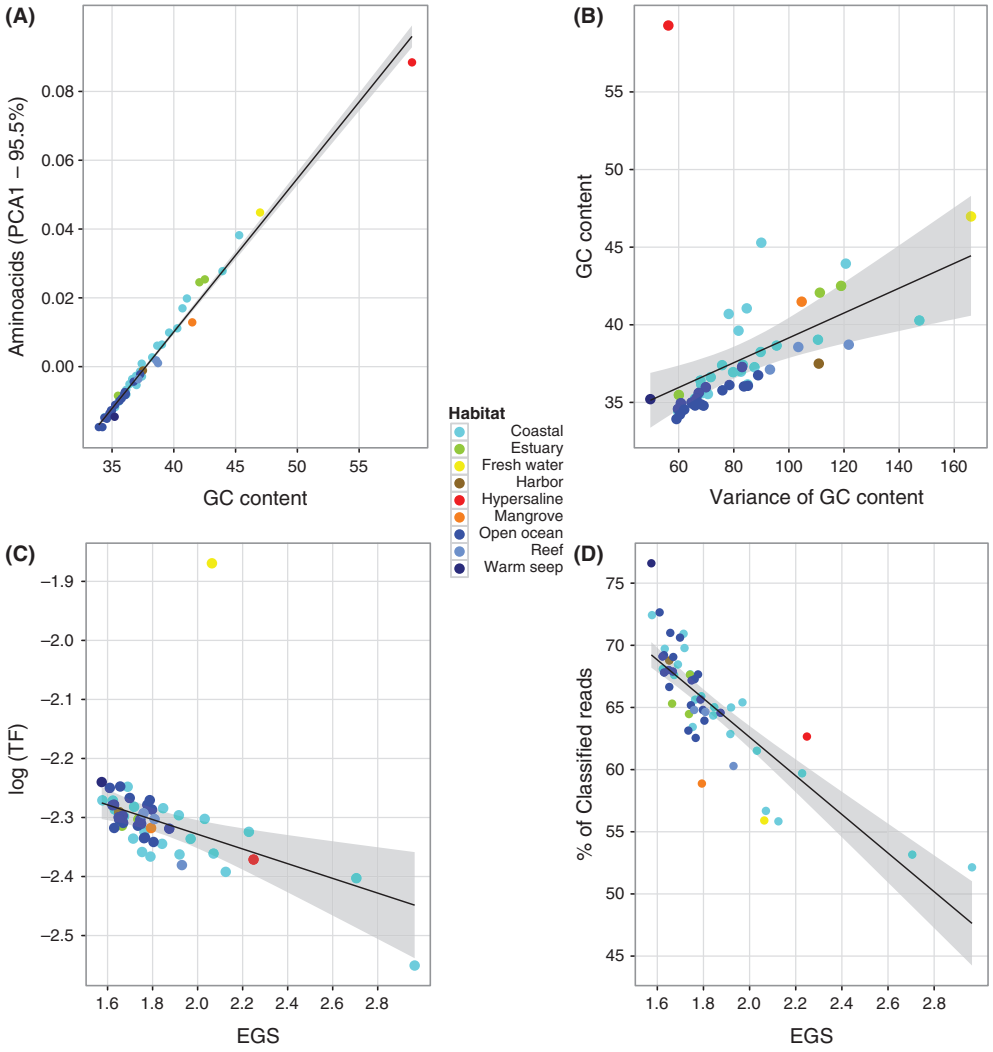

Fig. 3 Bivariate relationships between some community traits. (A) GC content was highly correlated with the aminoacid composition (also with dinucleotide and codon composition). (B) Positive relationship between the GC content and its variance. (C) Negative relationship between the percentage of transcriptional factors (TF) and the effective genome size (EGS). (D) Negative relationship between the percentage of classified reads and the EGS. The general trend is illustrated using a linear regression. See all Spearman's rank correlations in Table S3 (Supporting information).

ecological strategies directly related to succession (Klappenbach *et al.* 2000; Fierer *et al.* 2007). At the community level, however, the trait with the highest correlation with rRNA copy number was the ratio of AB amino acids (Spearman's $\rho = 0.50$, P-value < 0.01). New experimental studies should explore how rRNA copy number scales from the population to the community level and how is affected by the environment.

Final conclusions

Overall, the novel approach presented here may help to bridge the gap that exists between the disciplines of general ecology and microbial ecology. The recently developed methodology of metagenomics and trait-based community ecology seems totally compatible and useful for the ecological analysis of complex communities of microorganisms. Although trait-based approaches to microorganisms are largely unexplored (but see Litchman 2008; Green *et al.* 2008; for recent reviews), conserved properties at the molecular level (i.e. single-copy genes; Ciccarelli *et al.* 2006; Wu & Eisen 2008) and gene length (Xu *et al.* 2006) serve as anchors to extend the trait approach to the community level in complex environmental genetic data sets.

Acknowledgements

We thank members of the Green laboratory in UO for helpful discussions. AB is supported by the Spanish

FPU predoctoral scholarship program and EOC laboratory by grants PIRENA CGL2009-13318 from the Spanish Ministerio de Ciencia e Innovación (MICINN) and the EU-COST Action number ES1103: Microbial Ecology & The Earth System: Collaborating for Insight and Success with the new generation of sequencing tools (CISME).

References

Acinas SG, Rodriguez-Valera F, Pedrós-Alió C (1997) Spatial and temporal variation in marine bacterioplankton diversity as shown by RFLP fingerprinting of PCR amplified 16S rDNA. *FEMS Microbiology Ecology*, **24**, 27–40.

Auguet JC, Barberán A, Casamayor EO (2010) Global ecological patterns in uncultured Archaea. *ISME Journal*, **4**, 182–190.

Baltar F, Arístegui J, Gasol JM, Hernández-León S, Herndl GJ (2008) Strong coast-ocean and surface-depth gradients in prokaryotic assemblage structure and activity in a coastal transition zone region. *Aquatic Microbial Ecology*, **50**, 63–74.

Barberán A, Casamayor EO (2010) Global phylogenetic community structure and β-diversity patterns in surface bacterioplankton metacommunities. *Aquatic Microbial Ecology*, **59**, 1–10.

Barberán A, Casamayor EO (2011) Euxinic freshwater hypolimnia promote bacterial endemicity in continental areas. *Microbial Ecology*, **61**, 465–472.

Barberán A, Fernández-Guerra A, Auguet JC, Galand PE, Casamayor EO (2011) Phylogenetic ecology of widespread uncultured clades of the Kingdom Euryarchaeota. *Molecular Ecology*, **20**, 1988–1996.

Biers EJ, Sun S, Howard EC (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Applied and Environmental Microbiology*, **75**, 2221–2229.

Binnewies TT, Motro Y, Hallin PF *et al.* (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional and Integrative Genomics*, **6**, 165–185.

Bouvier TC, del Giorgio PA (2002) Compositional changes in free-living bacterial communities along a salinity gradient in two temperate estuaries. *Limnology and Oceanography*, **47**, 453–470.

Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.

Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. *Austral Ecology*, **18**, 117–143.

Crump BC, Hopkinson CS, Sogin ML, Hobbie JE (2004) Microbial biogeography along an estuarine salinity gradient: combined influences of bacterial growth and residence time. *Applied and Environmental Microbiology*, **70**, 1494–1505.

DeLong EF (2005) Microbial community genomics in the ocean. *Nature Reviews Microbiology*, **3**, 459–469.

DeSantis TZ, Hugenholtz P, Larsen N *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, **72**, 5069–5072.

Dinsdale EA, Edwards RA, Hall D *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.

Fierer N, Bradford MA, Jackson RB (2007) Toward an ecological classification of soil bacteria. *Ecology*, **88**, 1354–1364.

Foerstner KU, von Mering C, Hooper SD, Bork P (2005) Environments shape the nucleotide composition of genomes. *EMBO Reports*, **6**, 1208–1213.

Fuhrman JA (2009) Microbial community structure and its functional implications. *Nature*, **459**, 193–199.

Garcia JAL, Bartumeus F, Roche D, Giraldo J, Stanley HE, Casamayor EO (2008) Ecophysiological significance of scale-dependent patterns in prokaryotic genomes unveiled by a combination of statistic and genometric analyses. *Genomics*, **91**, 538–543.

Garcia JAL, Fernández-Guerra A, Casamayor EO (2011) A close relationship between primary nucleotides sequence structure and the composition of functional genes in the genome of prokaryotes. *Molecular Phylogenetics and Evolution*, **61**, 650–658.

Green JL, Bohannan BJM, Whitaker RJ (2008) Microbial biogeography: from taxonomy to traits. *Science*, **320**, 1039–1043.

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, **68**, 669–685.

Howard EC, Sun S, Biers EJ, Moran MA (2008) Abundant and diverse bacteria involved in DMSP degradation in marine surface waters. *Environmental Microbiology*, **10**, 2397–2410.

Klappenbach JA, Dunbar JM, Schmidt TM (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology*, **66**, 1328–1333.

Kraft NJ, Valencia R, Ackerly DD (2008) Functional traits and niche-based tree community assembly in an Amazonian forest. *Science*, **322**, 580–582.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews*, **72**, 557–578.

Legendre P, Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, **129**, 271–280.

Litchman E (2008) Trait-based community ecology of phytoplankton. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 615–639.

Lozupone CA, Knight R (2007) Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 11436–11440.

McArdle BH, Anderson MJ (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, **82**, 290–297.

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

Meyer F, Paarmann D, D'Souza M *et al.* (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

van Nimwegen E (2003) Scaling laws in the functional content of genomes. *Trends in Genetics*, **19**, 479–484.

Oksanen J, Kindt R, Legendre P *et al.* (2008). *Vegan: Community Ecology Package. R Package Version 1, 15.11*. Available from http://vegan.r-forge.r-project.org/.

Overbeek R, Begley T, Butler RM *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, **33**, 5691–5702.

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

Parter M, Kashtan N, Alon U (2007) Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biology*, **7**, 169.

Prosser JI, Bohannan BJM, Curtis TP *et al.* (2007) The role of ecological theory in microbial ecology. *Nature Reviews Microbiology*, **5**, 384–392.

Raes J, Foerstner KU, Bork P (2007a) Get the most out of your metagenome: computational analysis of environmental sequence data. *Current Opinion in Microbiology*, **10**, 490–498.

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007b) Prediction of effective genome size in metagenomic samples. *Genome Biology*, **8**, R10.

Raes J, Letunic I, Yamada T, Jensen LJ, Bork P (2011) Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Molecular Systems Biology*, **7**, 473.

Rhodes ME, Fitz-Gibbon ST, Oren A, House CH (2010) Amino acid signatures of salinity on an environmental scale with a focus on the Dead Sea. *Environmental Microbiology*, **12**, 2613–2623.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, **16**, 276–277.

Rusch DB, Halpern AL, Sutton G *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, **5**, e77.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. *PLoS Biology*, **5**, e75.

Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends in Genetics*, **17**, 425–428.

Sniegowski PD, Gerrish PJ, Lenski RE (1997) Evolution of high mutation rates in experimental populations of *E. coli*. *Nature*, **387**, 703–705.

Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, **6**, 938–947.

Willner D, Thurber RV, Rohwer F (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environmental Microbiology*, **11**, 1752–1766.

Wright IJ, Reich PB, Westoby M *et al.* (2004) The worldwide leaf economics spectrum. *Nature*, **428**, 821–827.

Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biology*, **9**, R151.

Xu L, Chen H, Hu X, Zhang R, Zhang Z, Luo ZW (2006) Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Molecular Biology and Evolution*, **23**, 1107–1108.

Zaneveld JR, Lozupone C, Gordon JI, Knight R (2010) Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Research*, **38**, 3869–3879.

A.B. is completing a PhD on ecological and phylogenetic patterns in microbial communities. A.F.G. is a PhD student focused on bioinformatics and phylogenetic and evolutionary processes of microbial functioning. E.O.C.'s and B.J.M.B.'s interests rely on applying ecological and evolutionary thinking on the biology of microorganisms.

Data accessibility

Data have been downloaded from the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) database at http://camera.calit2.net.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Summary of the metagenomic samples used and the community traits calculated.

Table S2 List of the 35 marker genes and their Clusters of Orthologous Groups (COG) annotation.

Table S3 All bivariate Spearman's Rank correlations between community traits. Lower triangle values correspond to Spearman's ρ. Upper triangle values are *P*-values. In bold, strong ($\rho > 0.6$) and significant (*P*-value > 0.01) correlations.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.