8842–8852 *Nucleic Acids Research, 2013, Vol. 41, No. 19* *Published online 5 August 2013*
doi:10.1093/nar/gkt673

Environmental shaping of codon usage and functional adaptation across microbial communities

Maša Roller[1], Vedran Lucić[1], István Nagy[2], Tina Perica[3] and Kristian Vlahoviček[1,4,*]

[1]Bioinformatics Group, Department of Molecular Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia, [2]Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Temesvári körút 62, H-6726 Szeged, Hungary, [3]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK and [4]Department of Informatics, University of Oslo, PO Box 1080 Blindern, NO-0316 Oslo, Norway

Received May 28, 2013; Revised July 8, 2013; Accepted July 9, 2013

ABSTRACT

Microbial communities represent the largest portion of the Earth's biomass. Metagenomics projects use high-throughput sequencing to survey these communities and shed light on genetic capabilities that enable microbes to inhabit every corner of the biosphere. Metagenome studies are generally based on (i) classifying and ranking functions of identified genes; and (ii) estimating the phyletic distribution of constituent microbial species. To understand microbial communities at the systems level, it is necessary to extend these studies beyond the species' boundaries and capture higher levels of metabolic complexity. We evaluated 11 metagenome samples and demonstrated that microbes inhabiting the same ecological niche share common preferences for synonymous codons, regardless of their phylogeny. By exploring concepts of translational optimization through codon usage adaptation, we demonstrated that community-wide bias in codon usage can be used as a prediction tool for lifestyle-specific genes across the entire microbial community, effectively considering microbial communities as meta-genomes. These findings set up a 'functional metagenomics' platform for the identification of genes relevant for adaptations of entire microbial communities to environments. Our results provide valuable arguments in defining the concept of microbial species through the context of their interactions within the community.

INTRODUCTION

An estimated 10^{30} microbes (1), prokaryotes of both the archeal and bacterial domains of life, inhabit the biosphere with metabolic agility that makes them capable of populating almost every corner of our planet. For example, the genes of microbes living in the human gut outnumber the gene complement of their host by a factor of 10 and provide enhanced metabolic capabilities not available to the host alone (2).

Environmental diversity studies have by and large bypassed the obstacle of <1% of microbes being amenable to cultivation in laboratory conditions [i.e. *the great plate-count anomaly* (3)] by instead using high-throughput sequencing to extract genomic information directly from the environmental sample without prior culturing. This approach, termed metagenomics, provides a 'snapshot' of the genetic content of entire, usually microbial, communities. While a pure microbial culture contains the clones of a single organism, microbes in natural habitats live in communities of varied phyletic composition with high rates of horizontal gene transfer documented even between evolutionary distant phylogenies (4). This in turn contributes to a substantial amount of genetic (and metabolic) diversity even within a single microbial species (5,6) that cannot be captured by single-clone sequencing and prompts us to rethink the concept of species at the microbial level (7). One such example is the pan-genome paradigm (8).

The increased availability of sequencing methods has enabled numerous metagenomic projects. Some of the sampled environments include geological sites such as sea (9), soil (10) and various extreme habitats [e.g. acid drainage from a metal mine (11), as well as gastrointestinal tracts of diverse organisms—including human (2)

*To whom correspondence should be addressed. Tel: +385 1 4606306; Fax: +385 1 4606286; Email: kristian@bioinfo.hr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

and mouse (12)]. Analysis pipelines of sequence data originating from metagenomics projects are presently focused in two main directions. The first one classifies the functions of identified genes (open reading frames) according to annotation available through orthology databases such as COG/KOG (Clusters of Orthologous Groups of genes) (13) or KEGG-KO (Kyoto Encyclopaedia of Genes and Genomes—Orthology) (14) and subsequently ranking the relative 'importance' of a particular function according to its abundance in the environment. The second direction focuses on estimating the phyletic distribution of microbial species represented in the environment, based on similarity searches against known microbial species' sequences (15). In contrast to metagenomics, there are only a few metaproteomic studies (i.e. investigations of the total proteomes of microbial communities) (16–19) owing mostly to the complexity of experimental set-up and a relatively low throughput limited by high experimental costs (16).

Microbes in the same environment live within the same physical and chemical constraints, such as temperature, pH or ion concentration, and it was demonstrated that GC content is metagenome-specific (20). Furthermore, communities of microbes have been shown to share similar tRNA pools to facilitate horizontal gene transfer (21), which also implies a limited choice of preferred codons that are cognate to the shared community tRNA pool. It has also been shown that fast growth rates introduce stronger bias in synonymous codon usage at the level of whole metagenomes (22), much like the effect observed in single microbial species (23,24).

We hypothesize that under the same environmental constrains, under close contact and amenable to similar evolutionary pressures and horizontal gene transfer, the levels of integration within the community reach as deep as the constituent genes and that microbial communities effectively behave as meta-genomes. To validate our hypothesis and make a predictive model of communal gene expression, we explore the concept of translational optimization through synonymous codon usage bias.

Apart from the synonymous codon usage (CU) bias observed *between* genomes of different microbial species, synonymous codons are also used with unequal frequencies *within* a single microbial genome. The CU bias within a genome reflects the selection pressure for translational optimization of highly expressed genes—primarily the protein synthesis machinery such as ribosomal genes and elongation factors, but also genes with environmental adaptation functions (25). At the level of a single microbial genome, the effect of CU bias is routinely used to predict for functionally relevant and highly expressed genes (26–28). The choice of preferred codons in a single genome is most closely correlated with abundance of the cognate tRNA molecules (29–31) and further influenced by the genome's GC content (32,33).

By analysing 11 diverse microbial community sequencing samples, we first demonstrate that microbes living in the same ecological niche, regardless of their phyletic diversity, share a common preference for codon usage, i.e. that we can observe CU bias at the community level and that this bias is different between different communities. Second, we show that CU bias varies within the community, with distributions resembling that of single microbial species, i.e. that we can observe the inter-community CU bias. We then use the effects of inter-community CU bias and translational optimization concepts to identify genes with CU close to that of the meta-ribosomal sample and therefore with high predicted expression across the entire microbial community, defining its 'functional fingerprint'. We validate this functional fingerprint against metaproteomic samples and the literature, demonstrating the predictive potential for functional relevance. With this approach, we reveal higher level organizational patterns in metagenomes and propose a 'functional metagenomics' platform to predict functionally relevant genes at the level of the whole microbial community.

MATERIALS AND METHODS

Metagenomic data assembly and annotation

All metagenomic data used in this study are listed in Supplementary Tables S1 and S2. All data were downloaded from the National Centre for Biotechnology Information (NCBI), either preassembled or in trace format. Unassembled metagenomes were assembled in-house using the Celera Assembler (34).

We used the assembled data as query to the STRING/COG database version 8.0 (35) in a BLASTX (36) search with an e-value cut-off of 10^{-8}. Open reading frames (ORFs) in the query sequences were assigned with an in-house Perl script based on the 3-nearest neighbour consensus rule. Specifically, a COG category is assigned to a gene only if the three best hits (smallest E-values) are all from the same orthologous group.

Single bacterial species' genomes and annotation

The complete genome sequence of the skin commensal *Propionibacterium acnes* was previously determined (37) and is publicly available (IDs: NC_006085.1). Genome sequencing of 12 *P. acnes* isolates (Supplementary Table S3) was performed on the SOLiD 3 System (Applied Biosystems, now part of Life Technologies) following the manufacturer's instructions. For this, DNA was extracted using the AquaGenomic kit (MultiTarget Pharmaceutical) and the preparation of the libraries and sequencing were performed using cycled ligation sequencing on a SOLiD 3 System; detailed procedure is described elsewhere (38–40). *P. acnes* strains were assigned STRING/COG categories and ORFs with the same method as metagenomes.

The complete genomes of six strains of *Rhodopseudomonas palustris* (41,42) from different environmental conditions were downloaded from the NCBI database including the complementary COG annotation (Supplementary Table S3). The *Escherichia coli* str. K-12 substr. DH10B, complete genome (ID: NC_010473) and complementary COG annotation were also retrieved from the NCBI.

Phylogenetic classification

We performed taxonomical analysis with the MEta-Genome ANalyzer (MEGAN) (15) from a BLASTX search of the all metagenomic samples against the NCBI non-redundant protein database (downloaded June 2010). Briefly, MEGAN classifies sequences to the last common ancestor of all significant BLAST hits and higher taxonomical levels do not include lower ones. For example, a sequence classified to the species level will not be counted again in the genus, family, order, etc.

Codon frequencies in metagenomes

The frequency of all codons whose cognate amino acids are encoded with more than one codon, i.e. synonymous codons, were normalized per amino acid for all metagenomic sets. The final set of codons includes all amino acids encoded with more than one codon, normalized so that the sum of frequencies of synonymous codons is equal to 1. According to MEGAN classification, the Sargasso Sea metagenome was divided into two groups (Supplementary Table S4): one containing only bacteria of the Alphaproteobacteria class and all its subgroups comprising ~36% of the whole sample and the other group all remaining phylogenies in the whole metagenome. The intraclass correlation coefficient (ICC) (43) quantifies the difference between measurements (codon frequencies) of data structured into groups (synonymous codons). Therefore, an ICC of 1 indicates no difference between datasets, while 0 indicates a large difference.

To examine differences in codon frequencies between a single species of the metagenome and an entire metagenome and also between same species in different metagenomes, we used sequences taxonomically annotated by MEGAN down to the species level. From all 11 metagenomes, we selected those species that had at least 2000 codons in each metagenome and were present in at least two metagenomes (Supplementary Table S4). For each species found, we computed the ICC between codon frequencies from the first and second metagenome, totalling 1029 comparisons. Also for each of those species, we computed the ICC between codon frequencies of the species and codon frequencies of the entire metagenome, totalling 2058 comparisons.

Distance in codon usage

For both the single microbial genomes and whole metagenomes, we used ribosomal protein genes from their cognate samples as the reference set, owing to their ubiquitous high level of expression. We analysed codon usage with another in-house Perl script that calculates the Measure Independent of Length and Composition (MILC) as previously described (44) for each ORF assigned through COG homology.

MILC quantifies *the distance in terms of codon usage* between a certain open reading frame and some expected distribution of codons. Mathematically, the measure is based on goodness of fit. Individual contribution of each amino acid to the MILC statistics is calculated as

$$Ma = 2\sum_{c} O_c ln\frac{O_c}{E_c} = 2\sum_{c} O_c ln\frac{f_c}{g_c}$$

where O_c is the actual observed count of codon c in a gene and E_c is the expected count of that codon. Observed counts can be replaced by frequencies, where f_c is the frequency of codon c in a gene and g_c is the expected frequency of that codon. The total difference in codon usage is then defined as

$$MILC = \frac{\sum_a M_a}{L} - C$$

The sum of all contributions (stop codons are excluded from the calculation) is divided by L, gene length in codons. C is the correction factor for overestimation of overall bias in shorter sequences. It is calculated as

$$C = \frac{\sum_a (r_a - 1)}{L} - 0.5$$

where r_a is the number of possible codons for the amino acid a, its degeneracy class.

For each gene in a genome, we derived the MILC distance to (i) the overall (meta)genome CU profile; and (ii) a reference set—a defined set of *optimally* encoded genes (i.e. ribosomal protein genes). We represented these two distances for each gene of a dataset by plotting them on a B-plot (26).

To test the variability in codon usage relative to phylogenetic distribution within ribosomal genes of a metagenome, we decomposed the Sargasso Sea ribosomal into subsets by species (Supplementary Table S5) using MEGAN. The MILC distance from the whole-metagenome ribosomal set of all subsets was measured and compared with the distance of all the genes of the metagenome and the whole ribosomal set.

Variability in codon usage for *P. acnes* and *R. palustris* isolates

The variability of codon usage within each COG was computed as the median MILC distance to the respective centroid for all genes from a given COG. The distances of two subsets, those within the 10% of smallest and 10% largest values, and the whole set were divided by STRING/COG supercategory. The occurrence of genes with the 10% smallest and largest values was tested against the whole set with the binomial test and the *P*-values corrected with the false discovery rate (FDR) method (Supplementary Table S6 and S7).

Codon usage distance between metagenomes

To measure the distance in codon usage between metagenomes, we used the MILC measure to derive the distance to (i) the overall metagenome CU profile of one metagenome; and (ii) the overall metagenome CU profile of another metagenome. By plotting the genes from both metagenomes in R (45), we can derive the distance in CU frequencies of the genes in each metagenome to both the genes in its own metagenome and those belonging to the other metagenome.

To simulate a situation in which a metagenome has no codon usage bias, we randomized the codons for each gene in the metagenome. The amino acid sequence was kept constant, but the synonymous codon used was randomly assigned using an in-house Perl script. The reference set of ribosomal genes was not randomized. We graphed the results in R in the same manner as the previous graph.

Enrichment of functions within highly expressed genes

We used an in-house Perl script to calculate the previously described MILC-Based Expression Level Predictor (MELP) (44) for every ORF, using ribosomal genes as the reference set. Briefly, the expression level of genes can be predicted through a ratio of a gene's distance in CU frequency from the overall usage in the genome and a reference set, i.e. the ratio of the two MILC distances.

MELP is defined as

$$\frac{MILC_{\text{genome average}}}{MILC_{\text{reference set}}}$$

$\text{MILC}_{\text{genome average}}$ is a measure of distance from the average codon usage of a microbial metagenome. $\text{MILC}_{\text{reference set}}$ is the distance from the average codon usage of a reference set, for which ribosomal genes were used.

Graphs produced in R were based on the number of ORFs in each STRING/COG supercategory with MELP values in the top 3% and low 3% (indicating high expression and low expression, respectively) normalized with the total occurrence of genes per supercategory regardless of MELP values. The occurrence of genes within the top and low 3% subsets for each STRING/COG supercategory was tested against the whole set, regardless of MELP values, with the binomial test and the *P*-values corrected with the FDR method.

Artificial metagenomes

For constructing artificial metagenomes, we downloaded all whole genome bacterial sequences from the NCBI Genbank database (downloaded June 2013). NCBI bacterial COG annotated sequences were used to create artificial metagenomes with similar COG supercategory composition to actual metagenomes (<1% difference). For each COG supercategory, the same number of sequences as in real metagenomes was randomly selected from the entire NCBI bacterial COG annotated set to make the corresponding artificial metagenome.

Metaproteome comparisons

The human gut (18) and Sargasso Sea metaproteomic studies (17) were chosen for benchmarking MELP values because both include >100 proteins and report relative abundances of proteins. The human gut study classified peptides into corresponding proteins, and we used the COG annotations for comparison with metagenomic COGs. The expression levels as predicted from metagenomes for each COG were computed as medians of MELP while the abundance of proteins in the metaproteome were computed as medians of normalized spectral abundance factors (NSAFs) per COG. The Sargasso Sea metaproteomic study provided sequences of proteins most closely matching peptides and the number of spectra they appeared in which we used as quantification of their abundance. We found the closest metagenomic genes for these proteins using a BLASTX search of the Sargasso Sea metagenome and show each gene's MELP value and the corresponding protein's spectral count. For each comparison, we computed the correlation of metaproteomic and metagenomic values and Spearman's correlation coefficient in R.

RESULTS

We examine genes in 11 metagenomes from eight distinct environments: the Sargasso Sea (9), three whale fall carcass samples (10), Waseca farm soil (10), human gut microbiome (2), lean and obese mouse gut microbiomes (12), an acid mine drainage (11) and two geographically distant enhanced phosphorous removal sludges (46).

Metagenome-centric bias in codon usage is phylogeny-independent

To investigate whether microbes living in the same ecological niche share a bias in CU, we compared the distance of each gene's CU in a metagenome from overall metagenome CU in the same metagenome of origin and all other metagenomes. Genes originating from one metagenome form a distinct cluster (as shown in one example in Figure 1A, and other examples in Supplementary Figure S1), and have CU predominantly closer to that metagenome overall CU than genes from the other metagenome. If the amino acid sequence and composition of each gene is preserved but the synonymous codons are randomly chosen (Figure 1B), the genes' CU become equidistant to both metagenomes regardless of their metagenome of origin.

To demonstrate the independence of metagenome codon usage (CU) to phyletic composition, we tested two parameters (i) the variability of single species' CU across metagenomes; and (ii) the variability of CU in a metagenome on removal of dominant phyla.

The variability of single species' codon usage across metagenomes

We de-composed all metagenome genes into their respective phylogenies (see Materials and Methods section and Supplementary Table S4). Then, genes pertaining to each species identified in two metagenomes were compared in terms of CU distance with (i) their respective metagenome overall CU; and (ii) CU of genes from the same species in a different metagenome. The resulting distance distributions, quantified with the intraclass correlation coefficient measure (ICC, described in Materials and Methods section) show a statistically significant difference in CU patterns of compared phylogenies—the within-species' CU pattern is more variable between metagenomes than

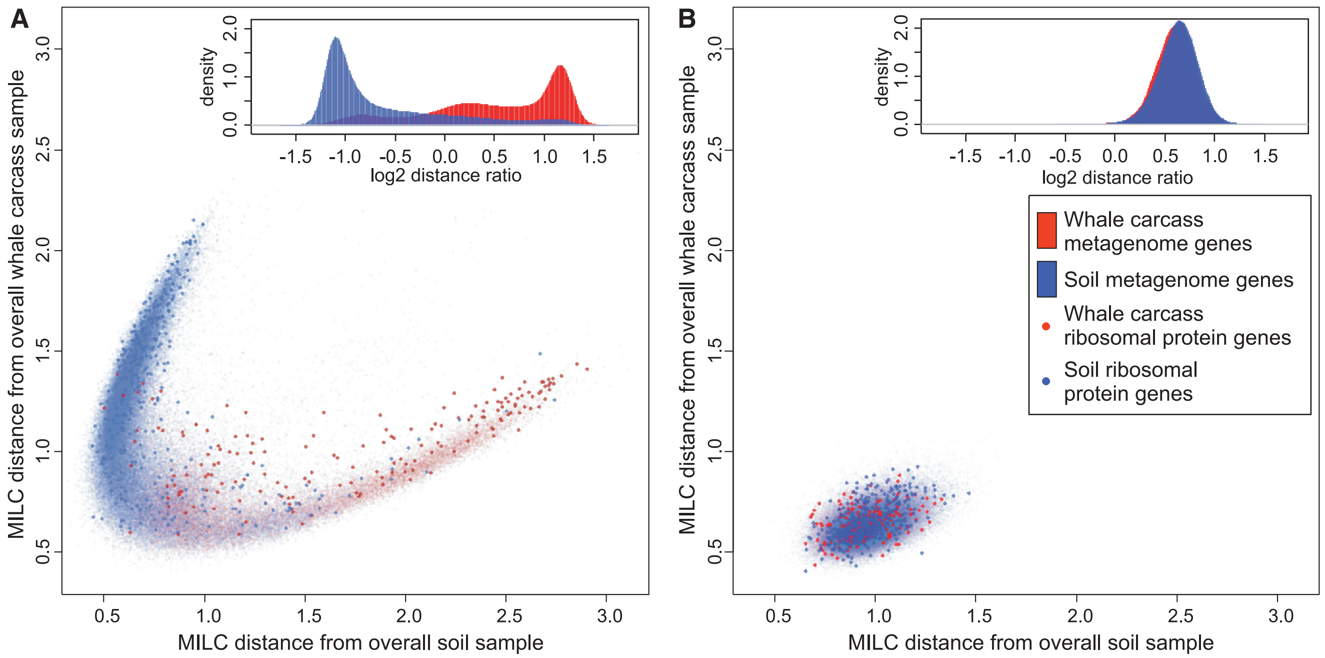

Figure 1. Codon usage is metagenome-specific. Soil versus Santa Cruz Whale fall Santa Cruz Bone codon usage (CU) frequencies. (**A**) The distance (MILC, outlined in 'Materials and Methods' section) of each gene's CU frequency to overall CU frequencies of two microbial communities. Genes [red in whale carcass (N = 33 422) and blue in Waseca soil (N = 88 696) metagenome] are predominantly closer to their respective metagenome of origin therefore forming two distinct groups (the distribution of log2 ratio of the two distances for each gene are shown in the inset). If the amino acid composition of metagenomes is kept constant and the codons randomly chosen, CU bias of each metagenome would be eliminated resulting in uniform distribution of CU distances and overlap of two colours, as shown in (**B**).

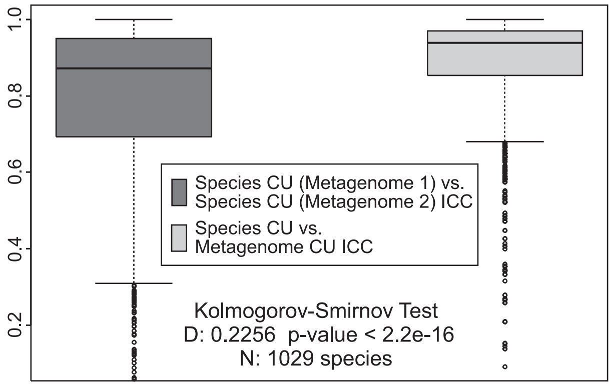

Figure 2. Codon usage variability between same species in different metagenomes is larger than within a metagenome. ORFs from each identified species (using MEGAN) were compared against their originating metagenome (dark grey, total comparisons N = 2058) and against same-species ORFs in a different metagenome (light grey, total comparisons N = 1029 comparisons). ICC measures were calculated, representing how 'close' the CU profiles match, with ICC = 1 denoting the perfect match. The light grey distribution shows less variability and is shifted towards higher ICC values, denoting the closer overall match of species' CU to their metagenome of origin.

in different species within the same metagenome (Figure 2).

To further test whether CU is a dynamic property that changes with different environmental constraints at the level of single bacterial species, we compared the CU variability of independently sequenced strains of microbes living in distinct niches. The commensal Gram-positive bacteria *P. acnes* live in consistent environmental conditions, predominantly on the human skin. Nagy *et al.* (37–40) have sequenced the genomes of 12 *P. acnes* isolates and analysed them along with the publicly available genomes. When compared with the metabolically versatile and a cosmopolitan bacterium *R. palustris* with six publicly available genome projects on isolates from diverse environments (41,42), the *P. acnes* strains show less variation in CU per orthologous group than do the *R. palustris* strains (Figure 3). Despite the fact that the sampling includes more than twice as many strains from constrained environmental conditions (*P. acnes*) than variable conditions (*R. palustris*), the variability in CU is smaller in the constrained environmental conditions.

The variability of codon usage in metagenomes on removal of dominant phyla

We analysed CU frequencies of the Sargasso Sea metagenome, the largest dataset in this study, in comparison with (i) other investigated metagenomes; (ii) itself, with dominant phyla removed; and (iii) by taking the equal sequence sample from the phyla where ≥30 sequences are present (32 phyla total with 32 sequences, for a total of 1024 sequences). The comparisons (Supplementary Figure S2) between Sargasso Sea CU frequencies and other metagenomes all show ICC < 0.75, while the same Sargasso sample with dominant phyla of the Alphaproteobacteria class removed (~36% of the whole set, see Supplementary Table S4) and the Alphaproteobacteria class itself show virtually no deviation (ICC > 0.98 and 0.95, respectively) from the

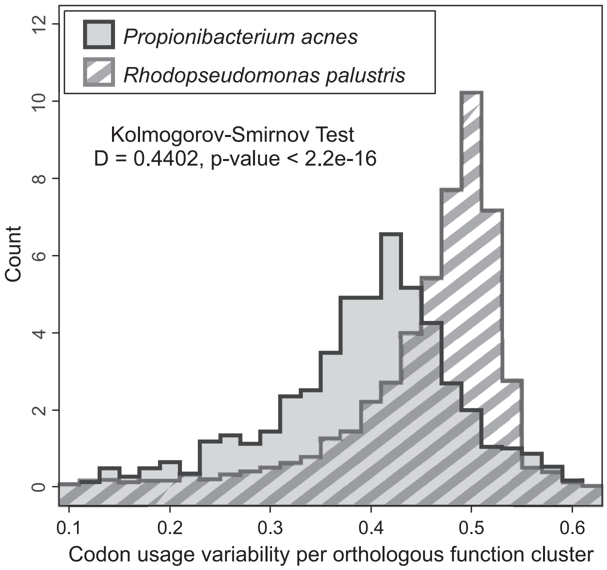

Figure 3. Environmental variability of codon usage. Variability of codon usage per COG Category in six strains of *R. palustris* and in 12 strains of *P. acnes*. The codon usage variability (calculated as median CU distance from the ribosomal set within an orthologous group to its centroid CU) for the strains of *P. acnes* ($N = 15436$), living in consistent environmental conditions, is shifted to the left, i.e. shows smaller variation and higher bias, than for the *R. palustris* strains ($N = 24071$) living in diverse environmental conditions.

original metagenome CU (Supplementary Figures S2A and B). The comparison with a small subset of equally represented phyla (Supplementary Figure S2C), even with a 800× reduction in sample, still shows ICC closer to its original metagenome. This demonstrates that the community-level codon usage bias is not an effect caused by the most abundant species.

Codon usage within metagenomes follows similar patterns as in single microbial genomes

Next, we tested whether the existence of CU bias within whole metagenomes mirrors the bias observed within single microbial genomes that facilitates detection of translationally optimized genes. As has been established at the level of single microbial genomes (29,30), the distance of each gene's CU frequency to the overall CU of the whole genome and to that of a 'reference set' of highly expressed genes (ribosomal protein genes) gives a characteristic crescent-shaped plot [Supplementary Figure S3A, introduced by (26)], separating the genes in two distinct groups based on the ratio of two distances. The group similar in CU to genes with high potential for expression (i.e. ribosomal proteins), are considered 'optimized for translation'.

In comparison, the distribution of genes in metagenomes according to the distance of their CU frequencies to their respective meta-ribosomal genes (i.e. the metagenomic reference set) and the overall CU frequencies of the whole metagenome is shown in Supplementary Figure S3B for the Santa Cruz whale carcass bone microbial sample (distributions of other metagenomes are shown in Supplementary Figures S5A–D, left panels). Metagenomes exhibit similar CU distance distributions to those observed in single bacterial genomes, despite the fact that they are composed of genes that originate from diverse phylogenies. To test the significance of the CU distribution shapes, we performed three different procedures: (i) testing for CU distance within a meta-ribosomal reference set (Supplementary Figure S4); (ii) randomization of synonymous codons while preserving the overall amino-acid content within a metagenome (Supplementary Figure S5); and (iii) construction of environment-independent artificial metagenomes by randomly selecting the whole bacterial genome sequences from the NCBI while maintaining the original metagenome functional profile (COG category composition; Supplementary Figures S6–S8). When the Sargasso Sea ribosomal genes are de-composed into subsets by species (Supplementary Table S5) the six most abundant species cluster close to the whole metagenomic set indicating similar CU patterns (see 'Materials and Methods' section and Supplementary Figure S4). If the amino acid sequence composition of genes in a metagenome is kept constant but the synonymous codons are randomly chosen (Supplementary Figure S5), the crescent plot shape analogous to single bacterial genomes and CU bias is lost. Artificial metagenomes with the same size and functional properties (COG category composition) of the original metagenomes show emergence of a similar crescent shape in CU frequency distance for all artificial metagenomes (Supplementary Figure S7). The ratio of CU distances from the cognate genome average and the (meta)ribosomal reference set, represented as MELP values in Supplementary Figures S7 and S8, is uniform in all artificial metagenomes but distinct between real metagenomes. The crescent shape is therefore a property of both the amino-acid bias in various functional groups and the bias in the choice of synonymous codons.

Predicting metagenomic expression and functional profiles through synonymous codon usage

Given that CU can vary in different environmental constraints in single bacterial species and that metagenomes share CU, we can use the CU bias in metagenomes to predict the expression levels of genes in the same manner as is routinely used to predict genes optimized for high levels of expression in single microbial genomes (26,28,44). The resulting predictions at the level of whole metagenomes, using the meta-ribosomal protein reference set, are shown in Figure 4 with additional examples in Supplementary Figures S9 and S10. Briefly, the most significantly enriched functions in the high expression level set are (i) amino-acid transport and metabolism (COG supercategory E) for Sargasso sea; (ii) energy production and conservation (COG supercategory C) for the whale fall metagenomes; and (iii) inorganic ion transport and metabolism (COG supercategory P) for the acid mine biofilm metagenome. Furthermore, the most striking difference we found was the lack of enrichment in energy production and carbohydrate metabolism (COG supercategories C and G) in the obese mice microbiota

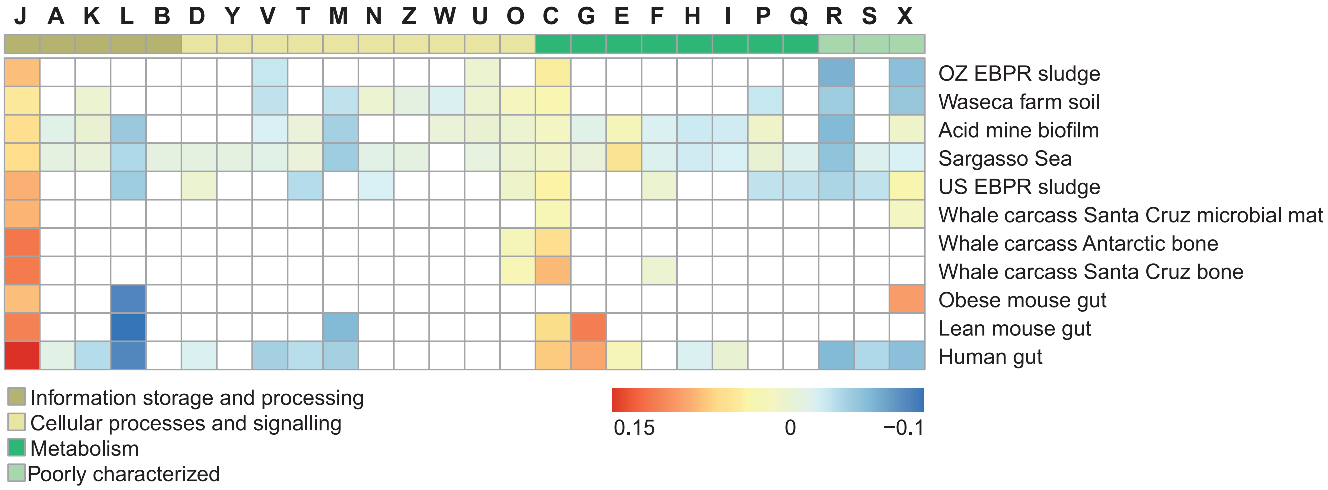

Figure 4. Enrichment of functions within highly expressed genes in metagenomes. Enrichment or depletion of functional annotations in the 3% genes with highest predicted expression (highest MELP measure) relative to the abundance of each COG supercategory in the whole metagenome for the OZ EBPR sludge (*N* = 29 754), Waseca farm soil (*N* = 88 696), acid mine biofilm (*N* = 79 257), Sargasso Sea (*N* = 688 539), US EBPR sludge (*N* = 20 175), Whale fall Santa Cruz microbial mat (*N* = 40 916), Whale fall Antarctic bone (*N* = 30 503), Whale fall Santa Cruz bone (*N* = 33 422), obese mouse gut (*N* = 4058), lean mouse gut (*N* = 4955) and human gut (*N* = 47 765), Santa Cruz whale fall bone (*N* = 33 422) and acid mine (*N* = 79 257). Metagenomes show different functional enrichment patterns that are consistent with environmental requirements (e.g. metabolite transport functions [E] in the Sargasso Sea or energy conversion [C] in the whale carcass metagenome). Non-significant enrichments are shown in white. Letters at the top represent COG supercategories: [J] Translation, ribosomal structure and biogenesis; [A] RNA processing and modification; [K] Transcription; [L] Replication, recombination and repair; [B] Chromatin structure and dynamics; [D] Cell cycle control, cell division, chromosome partitioning; [Y] Nuclear structure; [V] Defence mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [N] Cell motility; [Z] Cytoskeleton; [W] Extracellular structures; [U] Intracellular trafficking, secretion and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones; [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E]Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism; [R] General function prediction only; [S] Function unknown; [X] Uncharacterized.

sample, in contrast to both lean human and mouse microbiota samples.

To test whether the functional composition of metagenomes influences expression profiles, we again tested the effect on artificially constructed metagenomes. Supplementary Figure S11 shows the expression profiles for these artificial sets calculated in the same manner as for the original metagenomes. The resulting expression profiles, despite having the same functional composition, lost environment-specific signal, indicating that there is an environment-wide translational optimization for specific functional categories.

Validation with metaproteomics data

We compared our predictions of gene expression with two available metaproteomic studies that provided abundances of a comparably large number of proteins, the Sargasso Sea metaproteomic study (17) and a functionally (COG) classified subset of the human gut metaproteomic study (18). The predicted expression values based on CU optimization positively correlate with abundance in metaproteomic studies (Figure 5), both for the comparison of each gene with the protein most similar in sequence (Sargasso Sea *rho* = 0.34) and when median values per gene and protein COG are compared (Human Gut *rho* = 0.34). The comparison of the human gut metaproteome with our predictions in the Sargasso Sea metagenome, results in a correlation coefficient of *rho* = −0.03.

DISCUSSION

Microbial communities have properties of single bacterial genomes

The phenomenon of translational optimization through CU has been a well-established method of gene regulation for single bacterial genomes, and provides a predictive link between genomic and expression data (25,27). We have adapted this approach to the analysis of genomic information from entire microbial communities. Analysis of 11 distinct metagenomes shows that microbial communities exhibit codon usage bias similar to that already described for single microbial species (Supplementary Figure S3B and left panels of Supplementary Figure S5). It suggests that microbial communities sharing an environment are likely to have synchronized regulation mechanisms of translational optimization for expression of environment-specific genes. In the opposite scenario, where each microbial constituent of the community would use independent translational optimization mechanisms, we would expect as a net community-wide effect a uniform codon usage, without a particular preference for any subset of synonymous codons. In fact, this is exactly the scenario we demonstrated with our randomized samples (Supplementary Figure S5) and artificial metagenomes (Supplementary Figures S6–S8).

The evidence of distinct functional profiles of microbial communities regardless of the actual phylogenetic composition (47), emphasizes the community-wide

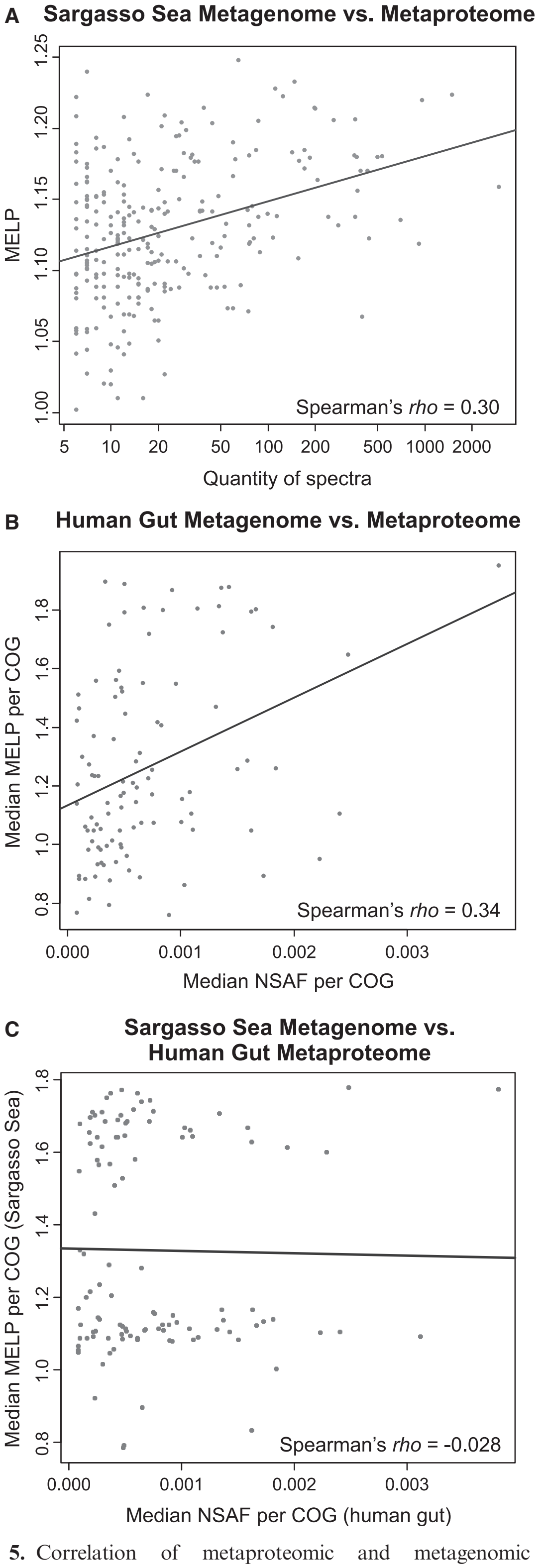

Figure 5. Correlation of metaproteomic and metagenomic data. Spearman's correlation coefficient between metaproteomic and metagenomic data is shown for (**A**) the Sargasso Sea (*N* = 257)—each proteins' number of spectra in the metaproteome versus the corresponding gene's MELP value and (**B**) the human gut—the median NSAF value per protein COG versus the median MELP value per gene COG (*N* = 116) and (**C**) the Human gut median NSAF value per protein COG versus the median MELP value per Sargasso Sea gene COG. Spearman's rho correlation is positive and greater for comparisons of metaproteomes with their own metagenomes then with foreign metagenomes.

translational optimization effect as an important metagenomic feature with relevant predictive power. This helps us rank metabolic functions and orthologous groups at the systems level that are more likely to be important for the adaptation of the entire metagenome to its particular environment. More importantly, our predictions can be used to identify and characterize genes of unknown functions with codon usage patterns similar to the community-wide lifestyle genes. By using the approach described here, we can effectively screen metagenomes for dominant functional properties that are not dependent on gene abundance data, but rather reflect higher organizational levels and give us essentially a 'characteristic functional fingerprint' of a microbial ecosystem, as shown in Figure 4.

Tracking functional adaptations across metagenomes

Two marine metagenomes, the Sargasso Sea and Santa Cruz whale carcass bone, have distinct profiles of gene enrichment within COG supercategories in environment-specific functions. The Sargasso Sea is a nutrient poor environment (9), and its genes for ABC-type transporters in the supercategory for amino acid transport and metabolism (COG supercategory E) are translationally optimized. On the other hand, the whale fall carcass metagenome, where microbes live on an abundant food source, shows optimization in the category for energy production and conversion (COG supercategory C). This difference in gene optimization can reflect functional adaptation of microbes to different environmental conditions and is lost in artificial metagenomes of the same functional composition (Supplementary Figure S11). The metaproteomic study of the Sargasso Sea environment (17) reported a dominance of the ABC-type transporters, consistent with our prediction of overrepresentation of these proteins in the set of genes with high translational optimization. Additionally, acid-mine drainage biofilms have been documented to frequently use inorganic ions as a source of energy (48), which is also validated in our analysis.

Alternative hypothesis for the influence of gut microbiota to host metabolism

The lifestyles of both the lean human and lean mouse gut metagenomes (Figure 4; for emphasis shown separately in Supplementary Figure S10) have similar profiles despite spatial isolation and different hosts. They show optimization enrichment in genes responsible for energy production and conversion (COG supercategory C) and carbohydrate transport and metabolism (COG supercategory G), i.e. the main functions of the biota inhabiting the intestinal track that aid the host in digestion of food. We show that these energy harvesting genes in the lean mouse's microbiome, shown previously to be less abundant than those in the obese gut (12), are in fact optimized for translational efficiency. The same category of genes in the obese mouse's gut lacks translational optimization, implying that these microbes might be less efficient at processing food therefore leaving more nutrients to be absorbed and metabolized by the obese host.

A metaproteomic study of a healthy (lean) human gut microbiome (18) showed a greater fraction of proteins responsible for carbohydrate transport and metabolism (COG supercategory G) than is the fraction of coding sequences identified in the corresponding metagenome, consistent with our predictions of high expression in the human gut metagenome, and provides an alternative hypothesis on the association of intestinal fauna with obesity.

Environmental effect on codon usage stability in microbial species

We address how variable environmental conditions influence CU by examining the variability in CU in 12 strains of bacteria living in the same environmental conditions (*P. acnes*) and six strains of bacteria living in different environmental conditions (*R. palustris)*. Consistent with our claims and the findings of Botzman and Margalit (49), the *R. palustris* samples show on overall higher variability in CU, suggesting plasticity of translational optimization that adopts to each specific environment. Even though the *R. palustris* strains generally show more variation in CU (Figure 3) both species, regardless of environmental constraints, show the least relative variation of CU within the COG categories (i.e. orthologous genes) for housekeeping, including ribosomal protein genes (Supplementary Table S6 and S7). The *R. palustris* strains show enrichment of genes with the highest variation in CU, predominantly in functional categories responsible for environment-specific metabolism, i.e. multidrug efflux pumps (supercategory V) and dehydrogenases (supercategories I, Q and R). This is consistent with previous findings that the optimal set of codons is consistent in varied bacterial phylogenies (23), and suggests that adaptation to environmental constraints occurs mainly through translational optimization of environment-specific functions, rather than equally for the entire bacterial genome(50). The observed effects can explain scenarios such as the intestinal fauna communities (51) where the genetic complement of a community can remain fairly constant, while they adapt to different environments through translational optimization, effectively changing the baseline expression levels of relevant genes to reflect the availability of a particular nutrient profile.

In-silico metaproteomics

Metaproteomics studies are crucial in comprehensive functional studies of protein families responsible for adaptation of entire microbial communities to an environment—many genes in metagenomes that have no identifiable homologue in the databases but show high expression levels or propensity for horizontal transfer. Even though there is an abundance of available metagenomic data, analogous studies of the whole proteome complement of an environment (17–19) are in their infancy and their throughput cannot rival that of metagenomics. Therefore, the most accessible approach to define communities at the level of proteome content will in the near future remain (meta)genome-based. We present a method that can link genomic data from easily accessible metagenomic sequences to the abundances of proteins in the corresponding metaproteome and provide insight into the functions of proteins important for survival in an environment. Despite the present scarcity of available metaproteome datasets resulting in the limited ability to perform quantitative comparisons, there is a general trend for proteins of higher abundance to have higher predicted metagenomic expression values and *vice versa* (Figure 5).

Possible mechanisms for the cause of CU bias in metagenomes

There is increasing evidence of novel mechanisms (52) that suggest an even higher amount of horizontal gene transfer events occurring within microbial communities than previously expected (53–56). We speculate that the *inter-community* CU bias we observe may be a consequence of an extensive exchange of genetic information in an environment, up to and including the level that is the strongest selection force for CU optimization—the shared tRNA pool (57,58). This in turn can be considered beneficial in particular for lifestyle-specific genes. With the common tRNA pool, these genes have an additional advantage for fixation within a community: better *a priori* translational optimization for more rapid protein production and therefore a faster penetrance. Drastic examples include the rapid propagation of antibiotic resistance genes (59). As the net effect of having a common community-wide optimization target, we observe the same signature as in single bacterial genomes—uneven *intra-community* CU bias regardless of its phylogenetic composition. Another contributing mechanism leading to the observed effect of CU bias at the community level would include parallel evolution of each microbial species' CU in such a way to support GC content required by the environmental constraints. However, for each of the analysed data samples, we observe a statistically significant deviation of the ribosomal protein genes' GC content with respect to that of the whole environmental sample (Supplementary Figure S12), indicating that the genes encoded with optimal codons are not under exclusive selection for GC content. In fact, both mechanisms combined seem to produce a detectable bias in CU and, as the net effect, allow us to build a predictive model based on the concept of translational optimization.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Ivan Jurić, Marija Buljan and Nenad Bartoniček for their contribution to the early stages of the analysis, Petar Jager for providing high performance computing support, Ozren Polašek with the discussions on statistical methods and Vedran Franke for discussion on analysis methods, manuscript versions and help with the figures. We also thank all members of the bioinfo.hr group and numerous colleagues for discussions on how to improve the manuscript.

FUNDING

European Molecular Biology Organization Young Investigator Programme Installation grant [1431/2006 to K.V.]; International Centre for Genetic Engineering and Biotechnology Collaborative Research Programme [CRP/CRO07-03 to K.V.]; Integra-Life Seventh Framework Programme [315997 to K.V.]; and the Croatian Ministry of Science Education and Sports [119-0982913-1211 to K.V.]. I.N. was supported by the Hungarian National Office for Research and Technology Teller program [OMFB-00441/2007] and by the French-Hungarian Associated European Laboratory SkinChroma [OMFB-00272/2009]. Funding for open access charge: EU Seventh Framework Programme [INTEGRA-LIFE RegPot 315997]; EMBO Young Investigator Installation grant [1431/2006]; Croatian Ministry of Science Education and Sports [119-0982913-1211].

Conflict of interest statement. None declared.

REFERENCES

1. Whitman,W.B., Coleman,D.C. and Wiebe,W.J. (1998) Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA*, **95**, 6578–6583.
2. Gill,S.R., Pop,M., DeBoy,R.T., Eckburg,P.B., Turnbaugh,P.J., Samuel,B.S., Gordon,J.I., Relman,D.A., Fraser-Liggett,C.M. and Nelson,K.E. (2006) Metagenomic analysis of the human distal gut microbiome. *Science*, **312**, 1355–1359.
3. Staley,J.T. and Konopka,A. (1985) Measurement of insitu activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.*, **39**, 321–346.
4. Keeling,P.J. and Palmer,J.D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.*, **9**, 605–618.
5. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl Acad. Sci. USA*, **102**, 13950–13955.
6. Willenbrock,H., Hallin,P., Wassenaar,T. and Ussery,D. (2007) Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. *Genome Biol.*, **8**, R267.
7. Achtman,M. and Wagner,M. (2008) Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.*, **6**, 431–440.
8. Mira,A., Martin-Cuadrado,A.B., D'Auria,G. and Rodriguez-Valera,F. (2010) The bacterial pan-genome: a new paradigm in microbiology. *Int. Microbiol.*, **13**, 45–57.
9. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D.Y., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
10. Tringe,S.G., von Mering,C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W., Podar,M., Short,J.M., Mathur,E.J., Detter,J.C. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.
11. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. and Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
12. Turnbaugh,P.J., Ley,R.E., Mahowald,M.A., Magrini,V., Mardis,E.R. and Gordon,J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.
13. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 14.
14. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
15. Huson,D.H., Auch,A.F., Qi,J. and Schuster,S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.
16. Keller,M. and Hettich,R. (2009) Environmental proteomics: a paradigm shift in characterizing microbial activities at the molecular level. *Microbiol. Mol. Biol. Rev.*, **73**, 62–70.
17. Sowell,S.M., Wilhelm,L.J., Norbeck,A.D., Lipton,M.S., Nicora,C.D., Barofsky,D.F., Carlson,C.A., Smith,R.D. and Giovanonni,S.J. (2008) Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J.*, **3**, 93–105.
18. Verberkmoes,N.C., Russell,A.L., Shah,M., Godzik,A., Rosenquist,M., Halfvarson,J., Lefsrud,M.G., Apajalahti,J., Tysk,C., Hettich,R.L. *et al.* (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME J.*, **3**, 179–189.
19. Wilmes,P., Wexler,M. and Bond,P.L. (2008) Metaproteomics provides functional insight into activated sludge wastewater treatment. *PLoS One*, **3**, e1778.
20. Foerstner,K.U., von Mering,C., Hooper,S.D. and Bork,P. (2005) Environments shape the nucleotide composition of genomes. *EMBO Rep.*, **6**, 1208–1213.
21. Tuller,T., Girshovich,Y., Sella,Y., Kreimer,A., Freilich,S., Kupiec,M., Gophna,U. and Ruppin,E. (2011) Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Res.*, **39**, 4743–4755.
22. Vieira-Silva,S. and Rocha,E.P. (2010) The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.*, **6**, e1000808.
23. Rocha,E.P. (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.*, **14**, 2279–2286.
24. Sharp,P.M., Bailes,E., Grocock,R.J., Peden,J.F. and Sockett,R.E. (2005) Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.*, **33**, 1141–1153.
25. Supek,F., Škunca,N., Repar,J., Vlahoviček,K. and Šmuc,T. (2010) Translational selection is ubiquitous in prokaryotes. *PLoS Genet.*, **6**, e1001004.
26. Karlin,S. and Mrazek,J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.*, **182**, 5238–5250.
27. Plotkin,J.B. and Kudla,G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.*, **12**, 32–42.
28. Sharp,P. and Li,W. (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
29. Ikemura,T. (1985) Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13–34.
30. Kanaya,S., Yamada,Y., Kinouchi,M., Kudo,Y. and Ikemura,T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.*, **53**, 290–298.
31. Tuller,T., Carmi,A., Vestsigian,K., Navon,S., Dorfan,Y., Zaborske,J., Pan,T., Dahan,O., Furman,I. and Pilpel,Y. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**, 344–354.
32. Chen,S.L., Lee,W., Hottes,A.K., Shapiro,L. and McAdams,H.H. (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl Acad. Sci. USA.*, **101**, 3480–3485.
33. Hershberg,R. and Petrov,D.A. (2009) General rules for optimal codon choice. *PLoS Genet.*, **5**, e1000556.
34. Myers,E.W., Sutton,G.G., Delcher,A.L., Dew,I.M., Fasulo,D.P., Flanigan,M.J., Kravitz,S.A., Mobarry,C.M., Reinert,K.H.J., Remington,K.A. *et al.* (2000) A whole-genome assembly of *Drosophila. Science*, **287**, 2196–2204.
35. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.*

(2009) STRING 8-a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
36. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
37. Bruggemann,H., Henne,A., Hoster,F., Liesegang,H., Wiezer,A., Strittmatter,A., Hujer,S., Durre,P. and Gottschalk,G. (2004) The complete genome sequence of Propionibacterium acnes, a commensal of human skin. *Science*, **305**, 671–673.
38. Horvath,B., Hunyadkurti,J., Voros,A., Fekete,C., Urban,E., Kemeny,L. and Nagy,I. (2012) Genome sequence of Propionibacterium acnes type II strain ATCC 11828. *J. Bacteriol.*, **194**, 202–203.
39. Hunyadkurti,J., Feltoti,Z., Horvath,B., Nagymihaly,M., Voros,A., McDowell,A., Patrick,S., Urban,E. and Nagy,I. (2011) Complete genome sequence of Propionibacterium acnes type IB strain 6609. *J. Bacteriol.*, **193**, 4561–4562.
40. McDowell,A., Hunyadkurti,J., Horvath,B., Voros,A., Barnard,E., Patrick,S. and Nagy,I. (2012) Draft genome sequence of an antibiotic-resistant Propionibacterium acnes strain, PRP-38, from the novel type IC cluster. *J. Bacteriol.*, **194**, 3260–3261.
41. Oda,Y., Larimer,F.W., Chain,P.S.G., Malfatti,S., Shin,M.V., Vergez,L.M., Hauser,L., Land,M.L., Braatsch,S., Beatty,J.T. *et al.* (2008) Multiple genome sequences reveal adaptations of a phototrophic bacterium to sediment microenvironments. *Proc. Natl Acad. Sci. USA*, **105**, 18543–18548.
42. Larimer,F.W., Chain,P., Hauser,L., Lamerdin,J., Malfatti,S., Do,L., Land,M.L., Pelletier,D.A., Beatty,J.T., Lang,A.S. *et al.* (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris. *Nat. Biotechnol.*, **22**, 55–61.
43. Donner,A. and Koval,J.J. (1980) The estimation of intraclass correlation in the analysis of family data. *Biometrics*, **36**, 19–25.
44. Supek,F. and Vlahovicek,K. (2005) Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics*, **6**, 15.
45. R Development Core Team. (2009) R Foundation for Statistical Computing.
46. Martin,H.G., Ivanova,N., Kunin,V., Warnecke,F., Barry,K.W., McHardy,A.C., Yeates,C., He,S.M., Salamov,A.A., Szeto,E. *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.*, **24**, 1263–1269.
47. Burke,C., Steinberg,P., Rusch,D., Kjelleberg,S. and Thomas,T. (2011) Bacterial community assembly based on functional genes rather than species. *Proc. Natl Acad. Sci.*, **108**, 14288–14293.
48. Johnson,D.B. (1998) Biodiversity and ecology of acidophilic microorganisms. *Fems. Microbiol. Ecol.*, **27**, 307–317.
49. Botzman,M. and Margalit,H. (2011) Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol.*, **12**, R109.
50. Retchless,A.C. and Lawrence,J.G. (2012) Ecological adaptation in bacteria: speciation driven by codon selection. *Mol. Biol. Evol.*, **29**, 3669–3683.
51. Arumugam,M., Raes,J., Pelletier,E., Le Paslier,D., Yamada,T., Mende,D.R., Fernandes,G.R., Tap,J., Bruls,T., Batto,J.M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.
52. McDaniel,L.D., Young,E., Delaney,J., Ruhnau,F., Ritchie,K.B. and Paul,J.H. (2010) High frequency of horizontal gene transfer in the oceans. *Science*, **330**, 50.
53. Boto,L. (2010) Horizontal gene transfer in evolution: facts and challenges. *Proc. Biol. Sci.*, **277**, 819–827.
54. Caro-Quintero,A., Deng,J., Auchtung,J., Brettar,I., Hofle,M.G., Klappenbach,J. and Konstantinidis,K.T. (2011) Unprecedented levels of horizontal gene transfer among spatially co-occurring *Shewanella* bacteria from the Baltic Sea. *ISME J.*, **5**, 131–140.
55. Kuo,C.H. and Ochman,H. (2009) The fate of new bacterial genes. *FEMS Microbiol. Rev.*, **33**, 38–43.
56. Malmstrom,R.R., Rodrigue,S., Huang,K.H., Kelly,L., Kern,S.E., Thompson,A., Roggensack,S., Berube,P.M., Henn,M.R. and Chisholm,S.W. (2013) Ecology of uncultured Prochlorococcus clades revealed through single-cell genomics and biogeographic analysis. *ISME J.*, **7**, 184–198.
57. Tuller,T. (2011) Codon bias, tRNA pools and horizontal gene transfer. *Mob. Genet. Elements*, **1**, 75–77.
58. Diene,S.M., Merhej,V., Henry,M., El Filali,A., Roux,V., Robert,C., Azza,S., Gavory,F., Barbe,V., La Scola,B. *et al.* (2013) The rhizome of the multidrug-resistant *Enterobacter aerogenes* genome reveals how new "killer bugs" are created because of a sympatric lifestyle. *Mol. Biol. Evol.*, **30**, 369–383.
59. Forsberg,K.J., Reyes,A., Wang,B., Selleck,E.M., Sommer,M.O. and Dantas,G. (2012) The shared antibiotic resistome of soil bacteria and human pathogens. *Science*, **337**, 1107–1111.