University of Zagreb
Faculty of Science
Department of Biology

Maša Roller Milošević

Analysis of metagenomes in extreme environments

GRADUATION THESIS

Zagreb, 2009

Ovaj rad izrađen je u Grupi za bioinformatiku, Zavod za molekularnu biologiju, Biološki odsjek, PMF, pod vodstvom prof. dr. sc. Kristana Vlahovičeka te je predan na ocjenu Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu radi stjecanja zvanja dipl. ing. biologije, smjer molekularna biologija.

Acknowledgments

I would like to thank the whole Bioinfo group (http://bioinfo.hr/) for companionship in work & play. Notably, Ivan Jurić &Tina Perica who started the work on metagenomics, Vedran Lucić for continuing it and Kristian Vlahoviček as the project supervisor & the boss. I thank Pero Jager for helping with Linux & Perl and recovering accidentally erased files.

Sunčana Špriovan (http://www.xvision.org/) is most gratefully acknowledged for beautifying introductory figures.

I am appreciative of the unfailing support of my family & friends.

BASIC DOCUMENTATION CARD

University of Zagreb

Faculty of Science

Department of Biology

GRADUATION THESIS

Analysis of metagenomes in extreme environments

Maša Roller Milošević

Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb

Not all prokaryotes are amenable to cultivation in laboratory conditions, and increasing amounts of knowledge about microbial diversity is gained from direct sampling of DNA from environments. A rapidly developing field that studies whole microbial communities is metagenomics, the culture-independent genomic study of organisms extracted directly from an ecological niche. Prokaryotic genomes show strong codon usage bias (CUB). CUB is directly correlated with expression levels of genes because codons used by genes expressed at high levels are encoded for by the most abundant tRNAs. Therefore, the CUB of a gene can be linked to its expression level through translational optimization. We show that metagenomes, much like genomes, also show CUB and that this phenomenon can be used to predict the expression level of genes at the level of the entire microbial community. We analyzed adaptation of organisms in metagenomes to their extreme environments through this approach.

(33 pages, 8 figures, 14 tables, 31 references, original in English)

Thesis is deposited in the Central Biological Library.

Keywords: metagenomics, codon usage bias, extreme environments

Supervisor: professor Kristian Vlahoviček, PhD

Reviewers: professor Mirjana Kalafatić, PhD and assistant professor Dijana Škorić, PhD

Thesis accepted: September 9th 2009.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek

DIPLOMSKI RAD

Analiza metagenoma ekstremnih okoliša

Maša Roller Milošević

Grupa za bioinformatiku, Zavod za molekularnu biologiju, Biološki odsjek, Prirodoslovno-matematički fakultet, Sveučilište u Zagrebu

Samo je malen broj bakterija prisutnih u biosferi moguće uzgojiti u laboratorijskim uvjetima, i stoga sve veća količina znanja o raznolikosti genoma prokariota dolaze iz direktne analize DNA iz okoliša pomoću visokoprotočnih metoda skupnog naziva Metagenomika. Metagenomom se naziva skup svih gena nekog mikrobnog okoliša. Poznato je da prokariotski genomi pokazuju značajnu nejednakost u korištenju sinonimnih kodona (CUB) (prema eng. *codon usage bias*). CUB je izravno povezan s razinom ekspresije gena stoga što su kodoni koje koriste visokoeksprimirani geni komplementarni s najrasprostranjenijim tRNA u organizmu. Dakle, CUB gena se može povezati s razinom ekspresije gena kroz translacijsku optimizaciju. Analizom različitih metagenomskih uzoraka pokazali smo da je moguće primijetiti nejednako iskorištenje kodona unutar pojedinog mikrobnog okoliša, kao i da je takvo svojstvo moguće iskoristiti za predviđanje razine ekspresije gena na razini cjelokupnog mikrobnog ekosustava. Primjena ovog pristupa omogućuje istraživanje genske adaptacije mikrobnih organizama na okoliš.

(33 stranica, 8 slika, 15 tablica, 31 literaturnih navoda, jezik izvornika: Engleski)

Rad je pohranjen u Središnjoj biološkoj knjižnici.

Ključne riječi: metagenomika, nejednakost u korištenju kodona, ekstremni okoliši

Voditelj: prof. dr. sc. Kristian Vlahoviček

Ocijenjitelji: prof. dr. sc. Mirjana Kalafatić i doc. dr. sc. Dijana Škorić

Rad prihvaćen: 9. rujan 2009.

TABLE OF CONTENTS

ABBREVIATIONS

BLAST	Basic Local Alignment Search Tool
COG	Clusters of Orthologous Groups of proteins
CUB	Codon Usage Bias
DNA	DeoxyriboNucleic Acid
EBPR	Enhanced Biological Phosphorus Removal
MELP	MILC-based Expression Level Predictor
MILC	Measure Independent of Length and Composition
NCBI	National Centre for Biotechnology Information
ORF	Open Reading Frame
RDP	Ribosomal Database Project
rRNA	ribosomal RiboNucleic Acid
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
tRNA	transfer RiboNucleic Acid

INTRODUCTION

Only a small portion of known prokaryotes are amenable to cultivation in laboratory conditions. For this reason culture-independent methods are introduced to study prokaryotic diversity. Metagenomics is a field that studies microbes through direct DNA sequencing from the environment, bypassing the need to cultivate them. The set of all individual genomes pertaining to each species present in one niche can be seen as one metagenome. Prokaryotes are not only diverse, but also show codon usage bias (CUB) at the level of genomes and between genes of the same genome. Codons are used unevenly with highly expressed genes showing codon usage patterns that are correlated with the most abundant tRNAs, in order to enable the most efficient translation of RNA. A metagenome can act as a genome and show codon usage bias that ultimately helps predict the expression level of genes.

NOT ALL PROKARYOTES CAN GROW IN CULTURE

Woese and Fox (1977) introduced a molecular phylogenetic approach to the classification of prokaryotes. On the basis of comparative analysis of the sequences of small subunit rRNAs (16S rRNAs), they identified three domains of life, one eukaryotic and two prokaryotic domains – Bacteria and Archea. Within the bacterial domain, they were able to identify only 11 distinct phyla from microbes that were amenable to cultivation. These were organisms that can grow on artificial media, under aerobic conditions and at moderate temperatures which makes them easy to isolate. Comparison of plate counts with direct microscopic counts has revealed that such easily isolated organisms constitute less than 1% of all species diversity in environmental samples (Staley and Konopka 1985), a phenomenon called the “great plate-count anomaly”.

Pace et al. (1986) introduced a novel method for sequencing small subunit rRNA genes that no longer required microorganisms to be cultivated. Nucleic acids can be directly isolated from environmental samples and the sequences of small subunit RNAs directly determined. Thus far, this method has provided evidence for 34 bacterial and 4 archeal phyla (Table 1), according to the classification of the Ribosomal Database Project (RDP)

(Cole et al. 2009). Six bacterial and one archeal phylum still have no cultivatable representatives. An additional six phyla of the prokaryotic domain have not a single representative with complete whole genome sequencing projects. Most of the knowledge gained from these poorly characterized organisms comes from environmental samples.

Table 1: List of all archeal and bacterial phyla currently present in the Ribosomal Project Database and the number of members of each phyla that have known 16S rRNA sequences from whole genome sequencing projects. The highlighted phyla have no known cultivable representatives (Cole et al. 2009).

domain	phylum	whole genome sequencing projects
Achaea	Crenarchaeota	22
	Euryarchaeota	34
	Korarchaeota	1
	Nanoarchaeum	0
Bacteria	Aquificae	5
	Thermotogae	10
	Thermodesulfobacteria	0
	Deinococcus-Thermus	5
	Chrysiogenetes	0
	Chloroflexi	6
	Thermomicrobia	1
	Nitrospira	1
	Deferribacteres	0
	Cyanobacteria	34
	Chlorobi	9
	Proteobacteria	425
	Firmicutes	158
	Actinobacteria	55
	Planctomycetes	0
	Chlamydiae	12
	Spirochaetes	18
	Fibrobacteres	0
	Acidobacteria	3
	Bacteroidetes	15
	Fusobacteria	1
	Verrucomicrobia	3
	Dictyoglomi	2
	Gemmatimonadetes	1
	Lentisphaerae	0
	BRC1	0
	OP10	0
	OP11	0
	TM7	0
	WS3	0
	Dehalococcoides	3
	SR1	0
	OD1	0
	Tenericutes	22

GENOMES SHOW CODON USAGE BIAS

Not only are prokaryotes diverse, they also differ in the usage of codons. The genetic code is degenerate – 61 codons correspond to only 20 amino acids, therefore multiple (i.e synonymous) codons encode the same amino acid. Synonymous codons are not used equally in different organisms, a phenomenon called the codon usage bias (CUB) (Ikemura 1985). A study of 100 archeal and eubacterial genomes (Chen et al. 2004) proved a strong correlation between the GC content of a genome and preferred codons. Since this correlation can be calculated solely from intergenic sequences, i.e. non-coding sequences, these findings suggest a mutational pressure at the level of the entire genome on the selection of preferred codons. However, there is also natural selection acting on the choice of preferred codons in an organism. Preferred codons have been shown to be correlated with the abundance of tRNA molecules in diverse organisms from bacteria – *Escherichia coli* and *Mycoplasma capricolum* (Ikemura 1985) – to eukaryotic species – *Schizosacharomyces pombe*, *Drosophila melanogaster* and *Caenorhabditis elegans* (Kanaya et al. 2001). CUB is therefore also influenced genome-wide by natural selection. Natural selection also acts on differences in CUB between genes of the same genome. It has been shown that CUB correlates strongly with the level of gene expression in bacteria (Gouy and Gautier 1982) and metazoans (Duret 2002). Highly expressed genes have the most highly optimized codon usage. The most likely hypothesis is that genes encoded by codons recognized by the more abundant tRNA molecules can be translated more efficiently and that the translational selection acting on highly expressed genes is stronger.

The level of expression of genes can be quantified through CUB by previously described methods (Karlin and Mrazek 2000, Supek and Vlahovicek 2005). Briefly, MILC (Measure Independent of Length and Composition) measures the distance between the distribution of codon frequencies in a single gene (the *observed CU*) and a distribution of codon frequencies within a set of sequences, or even the whole genome (the *expected CU*). MELP (MILC-based Expression Level Predictor), a derivative of MILC, predicts expression through the ratio of a gene's MILC to a reference set, usually housekeeping genes that are optimized in CUB. MELP is a quantitative predictor of gene expression; the bulk of genes in an organism (i.e. the genomic average) have low MELP that

corresponds to low expression values, while housekeeping genes and genes responsible for environmental adaptation are marked by high MELP values.

METAGENOMICS HELP DESCRIBE MICROBIAL DIVERSITY

Metagenomics, the culture-independent study of microbial communities, is a rapidly growing field. There is an estimated 10^{30} microbes in the biosphere (Dinsdale et al. 2008), ranging from the microbiota of vertebrates to photothrophic organisms in oceans. Metagenomics have yielded important discoveries to elucidate the biology of these abundant and varied organisms. In a recent example, study of the endosymbionts of obese and lean mice has revealed that the gut microbiome of obese mice is enriched with genes for energy harvest and that the microbes are more invasive than those of lean mice (Turnbaugh et al. 2006). A metagenomic study of the nutrient rich environment of the Sargasso Sea revealed that genomic diversity of the genus *Prochlorococcus*, the most abundant photosynthetic organism in the seas, is concentrated in discrete regions of the genome and largely due to lateral transfer of genes by viruses (Venter et al. 2004). As shown in table 2, the development of technology now offers more sequences to be generated at a lower cost, a 2-fold and 3-fold decrease of the price per sequenced base with the pyrosequencing and Illumina approach, respectively, compared to traditional dye-termination (Hugenholtz and Tyson 2008). This promises an increase of data from environmental sequencing projects.

Table 2: Comparison of the cost and throughput of sequencing technologies. New technologies (454-Roche pyroseqeuncing and Illumina sequencing) generate far more sequence data per run, at a much lower cost than conventional dye-terminator sequencing, but the reads are shorter. (Table from (Hugenholtz and Tyson 2008)).

Sequencing technology	Million base pairs per run	Cost per base (US¢)	Average read length (base pairs)
Dye-terminator (ABI 3730xl)	0.07	0.1	700
454-Roche pyrosequencing (GS FLX titanium)	400	0.003	400
Illumina sequencing (GAii)	2,000	0.0007	35

Environmental sequencing projects most commonly use the whole genome shotgun sequencing approach, and are not only a culture independent method but also an approach that requires no prior knowledge about the genomes being sequenced (figure 1). Environmental samples are extracted directly from the habitat and usually prokaryotic organisms are extracted through size exclusion filtering. Their genomic DNA is liberated and randomly broken into smaller fragments that can be cloned into sequencing vectors. The clones are sequenced using random primers to finally give researches a set of fragments, or reads, each hundreds of nucleotides long. The most tedious step is the assembly of these fragments (Brown 2002).

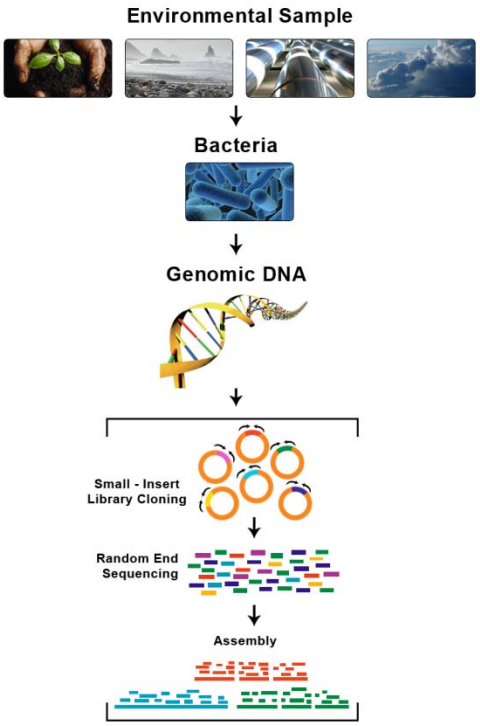

Figure 1: Overview of the metagenomics approach to the study of environmental samples. For a more detailed description, see text. (Adapted from http://camera.calit2.net)

To deduce the sequence of the whole original DNA molecule, reads must be computationally assembled (figure 2). The assembly process first finds overlaps between reads to create continuous stretches of sequence – fragments named contigs. The contigs are then assembled into scaffolds from information gained from mate pairs. A mate pair is a set of two reads, each from an alternate end of the same read in a vector,

with a known distance between the two reads. Therefore, scaffolds consist of a number of contigs with gaps of known length between them. To maximize the number of overlaps and ultimately achieve quality assembly each genome must be randomly broken and sequenced many times (Brown 2002).

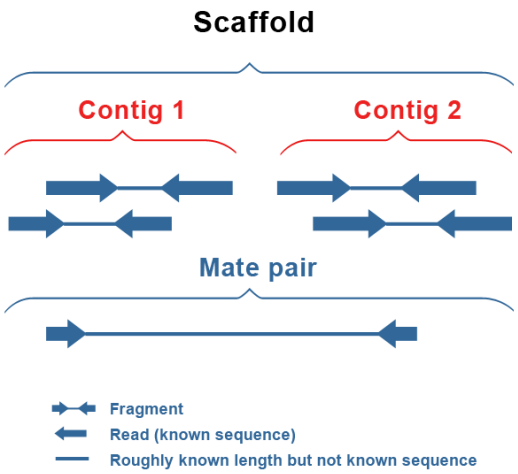

Figure 2: A schematic representation of the assembly process after whole genome shotgun sequencing. For a more detailed description, see text. (Adapted from http://genome.jgi-psf.org)

The resulting assembly of the whole set of microbial genomes present in the sampled space is the metagenome of that environment. Unfortunately, environmental sequencing is often done at small coverage rates, i.e. smaller length of sequenced DNA, so it rarely generates enough reads to assemble complete genomes of all the microbes present. This does not pose a huge impediment for the research of prokaryotic organisms because even the unassembled reads of shotgun sequencing typically contain at least one gene per read (Goo et al. 2004).

The diversity of species in a metagenome depends on the nature of the environmental niche it inhabits. Prokaryotic abundance can be estimated through fluorescence microscopy and total genomic diversity calculated from the reassociation rate of DNA isolated from a particular niche. Table 3 presents an overview of one such analysis that provided an estimate of the species diversity of seven different environments (Torsvik, Ovreas and Thingstad 2002). The DNA-based species definition defines a species as

prokaryotic organisms with more than 70% sequence identity. Through this estimate, forest soil, a nutrient rich environment, has an estimated 6,000 species while a salt-crystallizing pond, an extreme environment due to 22% salinity, is estimated to support a mere 7 distinct species. Nutrient rich environments harbour more diverse species, while those with extreme conditions are inhabited by a much smaller number of species.

Table 3: Prokaryotic abundance is determined by fluorescence microscopy while the total community genome complexity is calculated from reassociation rate of DNA. Distinct genomes are calculated assuming a species has >90% sequence similarity and with genome equivalents of the *Escherichia coli* genome (4.1 x 106 bp). (Table adapted from (Torsvik et al. 2002))

DNA source	Abundance (cells cm^{-3})	Community genome complexity (bp)	Distinct genomes
Forest soil	4.8×10^{9}	2.5×10^{10}	6000
Pasture soil	1.4×10^{7}	(1.5×10^{10})-(3.5×10^{10})	3500-8800
Pristine marine sediment	3.1×10^{9}	4.8×10^{10}	11,400
Marine fish-farm sediment	7.7×10^{9}	2.0×10^{8}	50
Salt-crystalizing pond	6.0×10^{7}	2.9×10^{7}	7

GOALS OF THE PROJECT

The primary aim of this project is to test whether there is translational optimization (through the existence of codon usage bias) at the level of entire metagenomes and whether this optimization acts on environment-specific genes. Tringe and collaborators (2005) found that roughly half of the predicted samples in metagenomes from three distinct environments showed homology to the COG (clusters of orthologous groups of proteins) database. Thus, we will explore whether our environmental samples have sufficient evolutionary links to better researched organisms to functionally annotate their organisms through homology searches.

Codon usage bias exists at the level of genomes. On the basis of three observations we hypothesize that CUB exists at the level of metagenomes as well. First, organisms living in the same environment live in similar conditions – i.e. temperature, pH or ion compositions – that influence the composition of DNA, most importantly GC bias. There are significant differences between the GC content of sequenced metagenomes (Foerstner et al. 2005), a factor that strongly influences CUB (Chen et al. 2004). Second, horizontal gene transfer is an important means of adaptation between organisms in microbial communities (Gogarten, Doolittle and Lawrence 2002). Third, organisms in the same environment often share the same essential functions, for example phosphorus removal in enhanced biological phosphorus removal (EBPR) sludge communities (Martin et al. 2006).

SOFTWARE AND BIOLOGICAL DATABASES

NCBI Trace Archive

The National Center for Biotechnology Information (NCBI) Trace Archive is a public repository of reads from various large-scale sequencing projects, including metagenomic projects. The data is deposited in the format of DNA sequence chromatograms (traces), as base calls and quality estimates.

We used the ftp site of the NCBI Trace archive (ftp://ftp.ncbi.nih.gov/pub/TraceDB/) for retrieving datasets of traces from metagenomics projects. The ftp makes the following files available:

- fasta.organism.XXX.gz: FASTA format files.
- qual.organism.XXX.gz: quality scores in FASTA format.
- clip.organism.XXX.gz: quality clip values for a read as provided by the sequencing center.
- anc.organism.XXX.gz: tab delimited files containing the ancillary information for each read
- xml.organism.XXX.gz: ancillary data in xml format

Celera Assembler

I used the Celera Assembler WGS version 5.2 in this project for the assembly of traces from metagenomic projects. It was first used for assembly of the *Drosophila melanogaster* genome from whole genome shotgun data (Myers et al. 2000). This is an open source software, available through SourceForge (http://sourceforge.net/), designed for a Linux operating system.

The STRING/COG database

The Clusters of Orthologous Groups of proteins (COGs) database is a tool for classification of proteins on the basis of the concept of orthologs– homologous proteins derived by vertical descent (Tatusov et al. 2003). The identification of orthologs in a genome proceeds through comparison of protein sequences and relies on the premise

that orthologs are more similar to each other than to any other protein in the genome. The construction protocol includes automatic detection of orthologs as well as manual curation and annotation.

The STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database (http://string.embl.de/) includes a more comprehensive COG database extended through a protocol similar to the original COG database. I used version 8.0 in this project (Jensen et al. 2009), hence it is referred to as the STRING/COG database.

BLAST

The Basic Local Alignment Search Tool (BLAST) is a freely available tool (http://www.ncbi.nlm.nih.gov/BLAST/download.shtml) for searching a database of sequences with a query sequence (Altschul et al. 1997). I used BLAST version 2.2.10 on a Linux system.

Perl

The Perl scripting language (http://www.perl.com) is a widely used dynamic programming language due to its flexibility and adaptability. It has become one of the most popular tools for bioinformatics because of the facility with which it can manipulate text files, the most commonly used format in biological databases. The usability of Perl is enhanced with an existing freely available set of Perl modules designed for bioinformatic solutions in the life sciences – Bioperl (Stajich et al. 2002).

All the Perl scripts used in this project are available upon request.

R

R is a freely available program for statistical computation and graphics. It is available through local CRAN mirrors (http://cran.r-project.org/), websites that host the R program distributions on its domains and make it available for download. The program's ability to manipulate large datasets with ease makes it a good choice for statistical analysis of metagenomes.

MILC and MELP

MILC (Measure Independent of Length and Composition) quantifies the distance in codon usage between a certain open reading frame and some expected distribution of codons (Supek and Vlahovicek 2005). 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.

MELP (MILC-based Expression Level Predictor) is a statistic that predicts the expression level of genes. It is defined as

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

$MILC_{genome\ average}$ is a measure of distance from the average codon usage of a microbial metagenome. $MILC_{reference\ set}$ is the distance from the average codon usage of a reference set, for which housekeeping genes are used.

Even though measurements for MILC are corrected for length we used a threshold of 100 codons for all our open reading frames, as is recommended by many researchers when using codon usage measures.

GC CONTENT MEASURE

The metagenomic GC content was calculated through the measure

$$GC_{measure} = \frac{(GC_{\max} - GC_{actual})}{(GC_{max} - GC_{min})}$$

where GC_{max} and GC_{min} are the theoretical maximum and minimum GC content, respectively, of a nucleotide sequence encoding a certain protein sequence. GC_{actual} is the true GC content of the sequence.

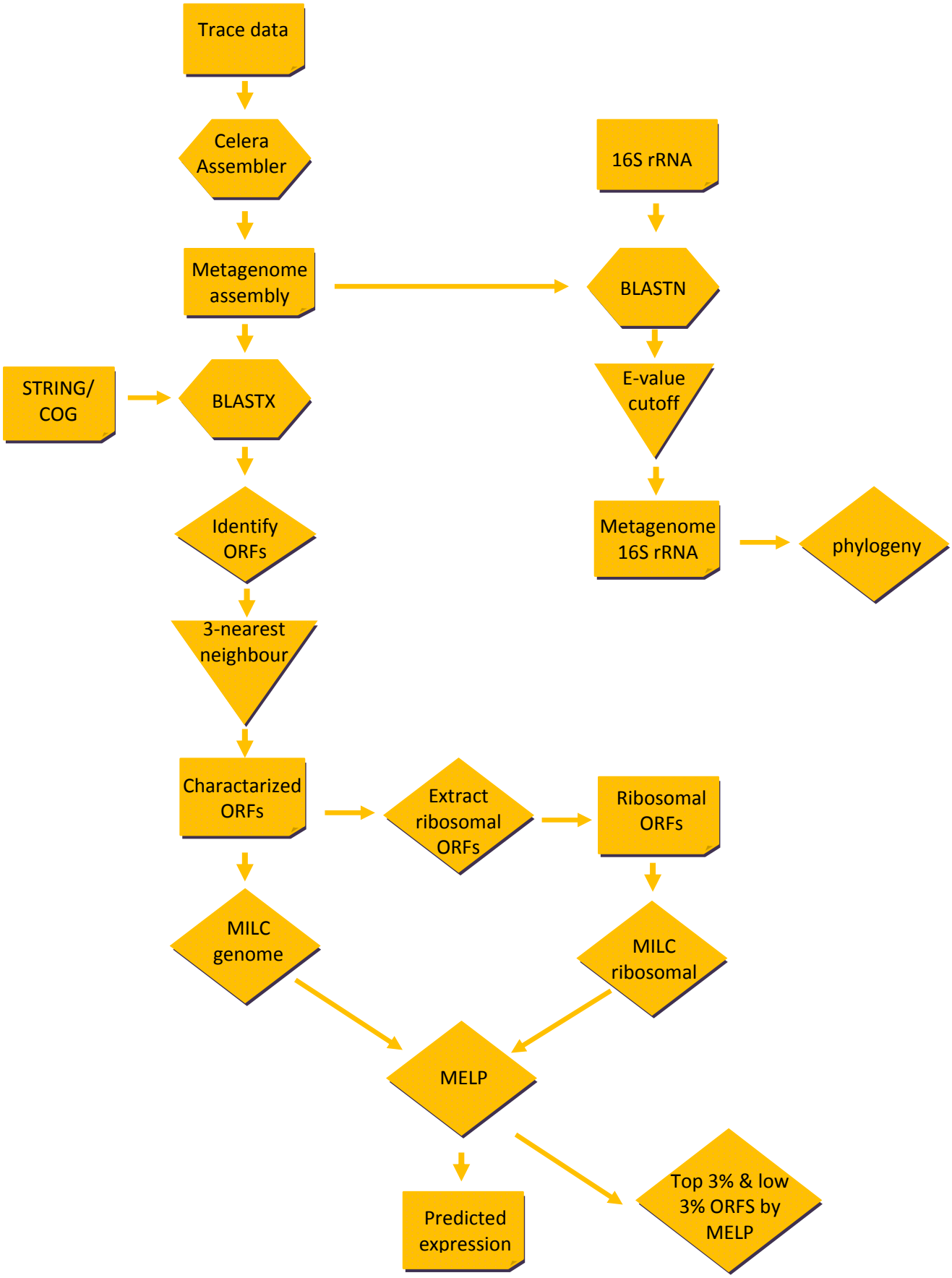

Figure 3: overview of the analysis pipeline. For a more detailed description see text.

Assembly

I downloaded the trace data, including fasta, quality, clip, ancillary information and mate pairs (where available) from the NCBI trace archive.

Table 4: Names of metagenomes used in this project, their NCBI Project IDs and references for the original sequencing projects.

Metagenome	NCBI Project ID	Reference
5-Way (CG) Acid Mine Drainage Biofilm Metagenome	13696	(Tyson et al. 2004)
Human Distal Gut Biome	16729	(Gill et al. 2006)
Lean Mouse 1 Gut Metagenome	17391	(Turnbaugh et al. 2006)
Obese Mouse 1 Gut Metagenome	17397	
US EBPR Sludge Metagenome	17657	(Martin et al. 2006)
OZ EBPR Sludge Metagenome	17659	

I performed the trace data assembly using the Celera Assembler, a program specifically developed for assembly of the human genome, using modified options to accommodate the inherent differences of metagenomic data to that of a single eukaryote. To circumvent the smaller coverage rate, I increased the error rates for building contigs to 14%, reduced the overlap length size to 14 and disabled fragment correction. Bubble popping was also disabled, a feature designed to identify haplotype differences in whole genome assembly projects.

Assigning ORFs and COG categories

I used all of the assembled data – singletons, contigs and scaffolds – as query to the STRING/COG database in a BLASTX search (compares a nucleotide query to a protein database) with an e-value cutoff of 10^{-8}. The position of matches to proteins in the COG database helped us identify open reading frames (ORFs) in the query sequences of metagenomic data.

An in-house script assigns COG categories to genes based on the 3-nearest neighbour consensus rule. More exactly, a COG category is assigned to a gene only if the three best hits (smallest E-values) are all from the same orthologous group.

Analyzing codon usage

An in-house script extracted ribosomal open reading frames using STRING/COG characterization. I extracted ORFs that matched COGs of ribosomal proteins in the translation, ribosomal structure and biogenesis (J) super category.

We analyzed codon usage through an in-house Perl script that calculates the Measure Independent of Length and Composition (MILC) and GC content through calculations previously described. $MILC_{(genome\ average)}$ calculates the distance of an ORF from the genome average, while $MILC_{(ribosomal)}$ shows the distance of each ORF from the reference set – ribosomal genes. We graphed the results for every metagenome in R, showing $MILC_{(genome\ average)}$ and $MILC_{(ribosomal)}$ each on an axis, colouring each ORF according to GC content and showing a regression line through the graph where $MILC_{(genome\ average)}$ equals $MILC_{(ribosomal)}$.

Randomizing codons

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

Predicting expression values

We used and in-house Perl script to calculate the MILC-Based Expression Level Predictor (MELP) for every ORF, using ribosomal genes as the reference set due to their ubiquitous high level of expression. The graphical representation shows the number of ORFs that have MELP values in the top 3% and low 3% (indicating high expression and low expression, respectively) separated according to their STRING/COG super category.

Species diversity

The taxonomic annotation of assembled sequences was derived from a BLAST search against all 16S rRNA sequences available in the Ribosomal Database Project (RDP) version release 10.13 (Cole et al. 2009). We performed a BLASTN search of all the metagenomic sequences against the RDP database with an E value cutoff of 10^{-10} and word size of 20. The top hit from the query metagenome (i.e. smallest E-value) for each query was statisticaly processed in R and provided us with an estimate of species diversity.

RESULTS

ASSEMBLY OF METAGENOMES

Table 5: Statistics of the Celera Assembler for extreme metagenomes assemblies. Contigs in scaffolds, contigs not in scaffolds (degenerate contigs) and singleton reads are highlighted and were used for further analysis. Singleton reads are reads that could not be placed in a contig, but are neither ambiguous nor repetitive. The length of all scaffolds (TotalBasesInScaffolds) and degenerate contigs (DegenContigLength) are shown in DNA bases.

Metagenome	Acid mine biofilm	US EBPR sludge	OZ EBPR sludge	Human gut	Lean mouse gut	Obese mouse gut
Total Scaffolds	32551	3493	5306	7247	419	440
TotalContigsInScaffolds	32551	4762	6485	7488	419	440
TotalBasesInScaffolds	68181719	12323873	21292466	22377089	547050	577912
TotalDegenContigs	17834	4612	5740	9129	50191	29592
DegenContigLength	30052223	1054	6563029	10978589	515839	5182865
SingletonReads	104351	26848	51535	62996	772190	543487

Table 6: Coverage statistics of metagenome assemblies after assembly with the Celera Assembler. The measure ContigsOnly is the length of all contigs in scaffolds and their exact repeats divided by scaffold length. ContigsAndDegens adds degenerate contigs divided by scaffold length. The AllReads measure includes all the reads used for the assembly divided by scaffold length.

Metagenome	Acid mine biofilm	US EBPR sludge	OZ EBPR sludge	Human gut	Lean mouse gut	Obese mouse gut
ContigsOnly	1.44	4.44	2.92	2.14	1.30	2.02
ContigsAndDegens	2.74	6.15	3.81	3.35	3.60	19.30
AllReads	3.79	7.25	5.13	5.48	16.17	40.62

BLAST

Table 7: Summary of BLASTX results for each metagenome (query) against the COG database at E value cutoff of 10^{-8}. The number of sequences used as query (scaffolds, contigs and singletons) and the number and percent of the query sequences that had a match in the database are given.

Metagenome	Acid mine biofilm	US EBPR sludge	OZ EBPR sludge	Human gut	Lean mouse gut	Obese mouse gut
Number of sequences used as query	154736	36222	63760	79613	822800	573519
Number of sequences with match in database	110591	32367	49785	59553	33177	18561
% of sequences with match in database	71.47	89.36	78.08	74.80	4.03	3.24

Table 8: Number of open reading frames (ORFs) found in metagenomes and the number and percent of found ORFs used for further study. ORFs that are degenerate, malformed or shorter than 100 codons were eliminated.

Metagenome	Acid mine	US EBPR sludge	OZ EBPR sludge	Human gut	Lean mouse gut	Obese mouse gut
number of ORFs found	112232	30244	46758	57058	27744	17267
number of ORFs used	79257	20175	29754	47765	4955	4058
% of ORFs used	70.62	66.71	63.63	83.71	17.86	23.50

SPECIES DIVERSITY

Table 9: Top 5 species in the acid mine drainage biofilm metagenome, according to their RDP classification.

Species	% hits
Shewanella algae	52.01
uncultured Chloroflexi bacterium TK-SH14	31.50
Prochlorococcus marinus str. MIT 9313	4.59
uncultured bacterium MES_rTCB88	0.66
Xanthomonas axonopodis	0.66

Table 10: Top 5 species in the US EBPR sludge metagenome, according to their RDP classification.

Species	% hits
Shewanella algae	49.27
uncultured Chloroflexi bacterium TK-SH14	34.10
Zymomonas sp. S5	2.79
uncultured bacterium MES_rTCB88	2.18
Prochlorococcus marinus str. MIT 9313	2.06

Table 11: Top 5 species in the OZ EBPR sludge metagenome, according to their RDP classification.

Species	% hits
uncultured Chloroflexi bacterium TK-SH14	39.61
Shewanella algae	39.57
Zymomonas sp. S5	10.72
Mannheimia granulomatis	6.62
Prochlorococcus marinus str. MIT 9313	1.64

Table 12: Top 5 species in the human distal gut metagenome, according to their RDP classification.

Species	% hits
uncultured *Catenibacterium sp*	1.80
Prochlorococcus marinus str. MIT 9313	1.68
uncultured bacterium p-922-s962-5	1.28
uncultured bacterium JPL1_61	1.26
uncultured bacterium orang1_aai55d06	1.18

Table 13: Top 5 species in the lean mouse gut metagenome, according to their RDP classification.

Species	% hits
Prochlorococcus marinus str. MIT 9313	17.50
Mycoplasma arthritidis	2.27
uncultured bacterium HD5--2	1.94
uncultured bacterium myd5_aaa04h09	1.72
uncultured bacterium C14_g02_2	1.60

Table 14: Top 5 species in the obese mouse gut metagenome, according to their RDP classification.

Species	% hits
Prochlorococcus marinus str. MIT 9313	19.10
uncultured bacterium mcbc118	3.33
uncultured bacterium mcbc51	3.26
uncultured bacterium mcbc51	2.72
uncultured bacterium MD27_aaa04g10	2.18

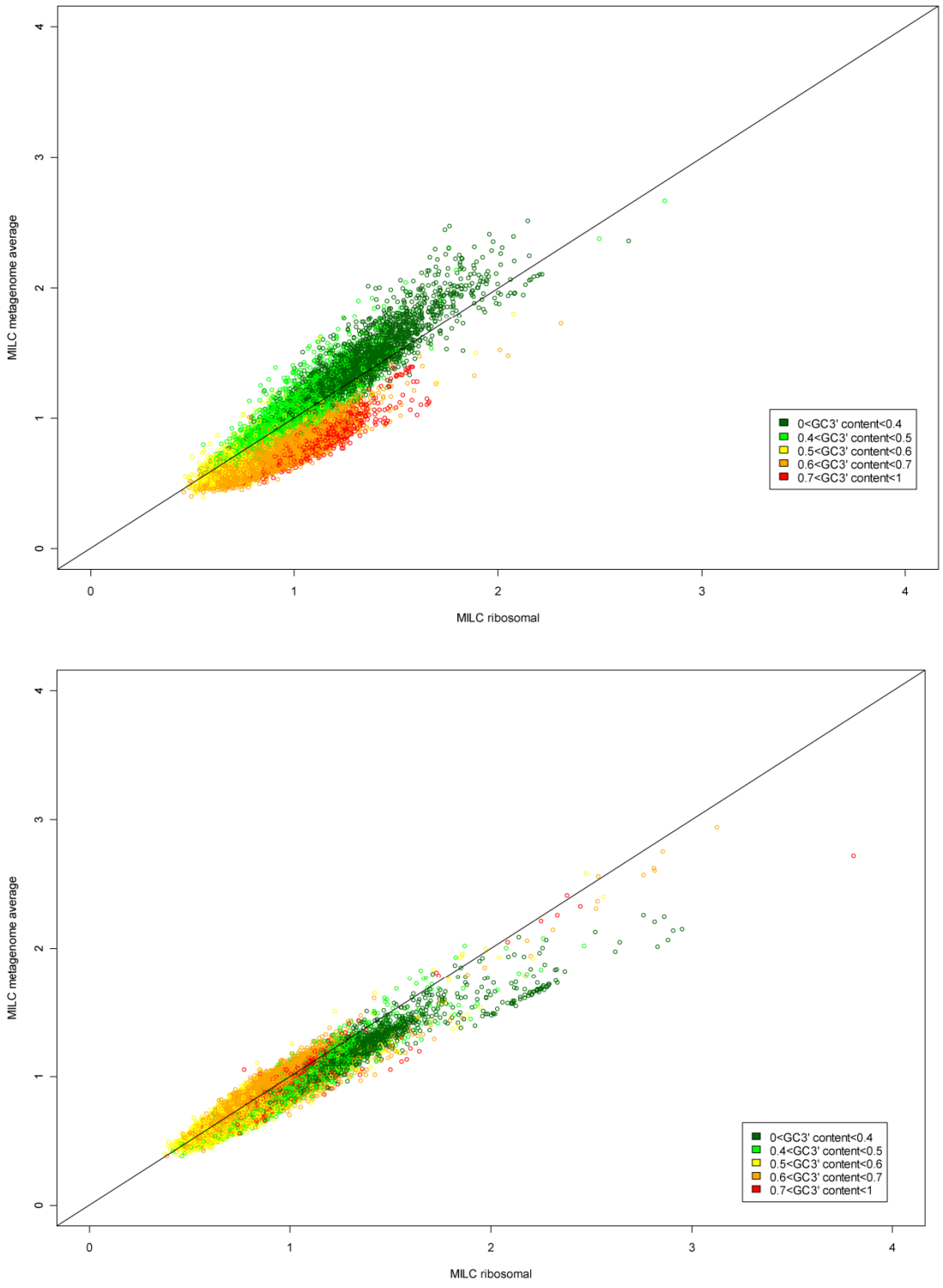

Figure 3: MILC metagenome average vs. MILC ribosomal for the US EBPR sludge metagenome (top) and OZ EBPR sludge metagenome (bottom). In figures 3-7 each ORF is plotted according to its distance from the metagenome average (MILC metagenome average) and distance from the reference ribosomal set (MILC ribosomal). ORFs are coloured according to the GC content. Additionally, a regression line is plotted where the distance from the metagenome average equals the distance from the ribosomal set.

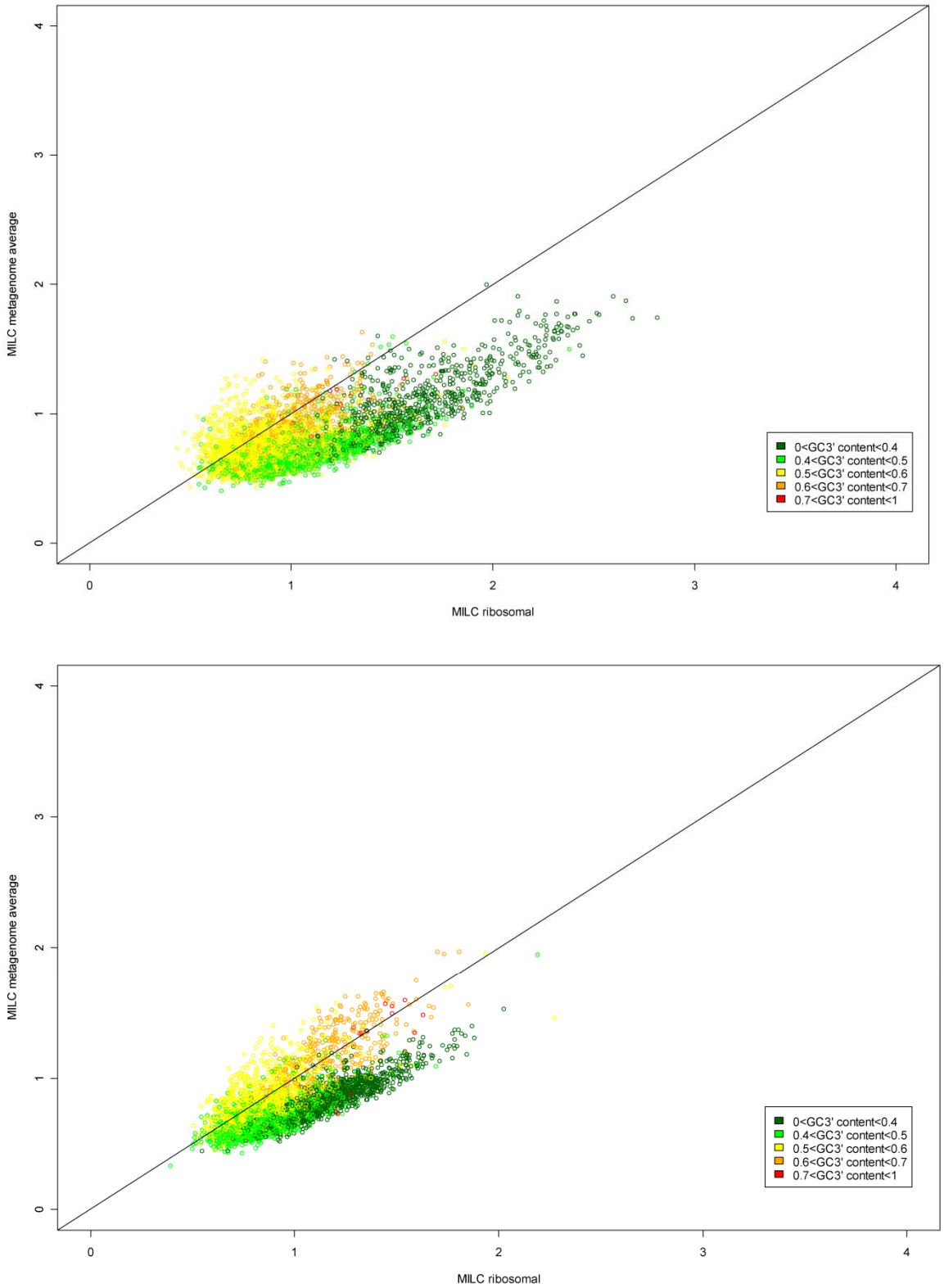

Figure 4: MILC metagenome average vs. MILC ribosomal for the lean mouse gut metagenome (top) and obese mouse gut metagenome (bottom). For a detailed explanation of the plot see figure 3.

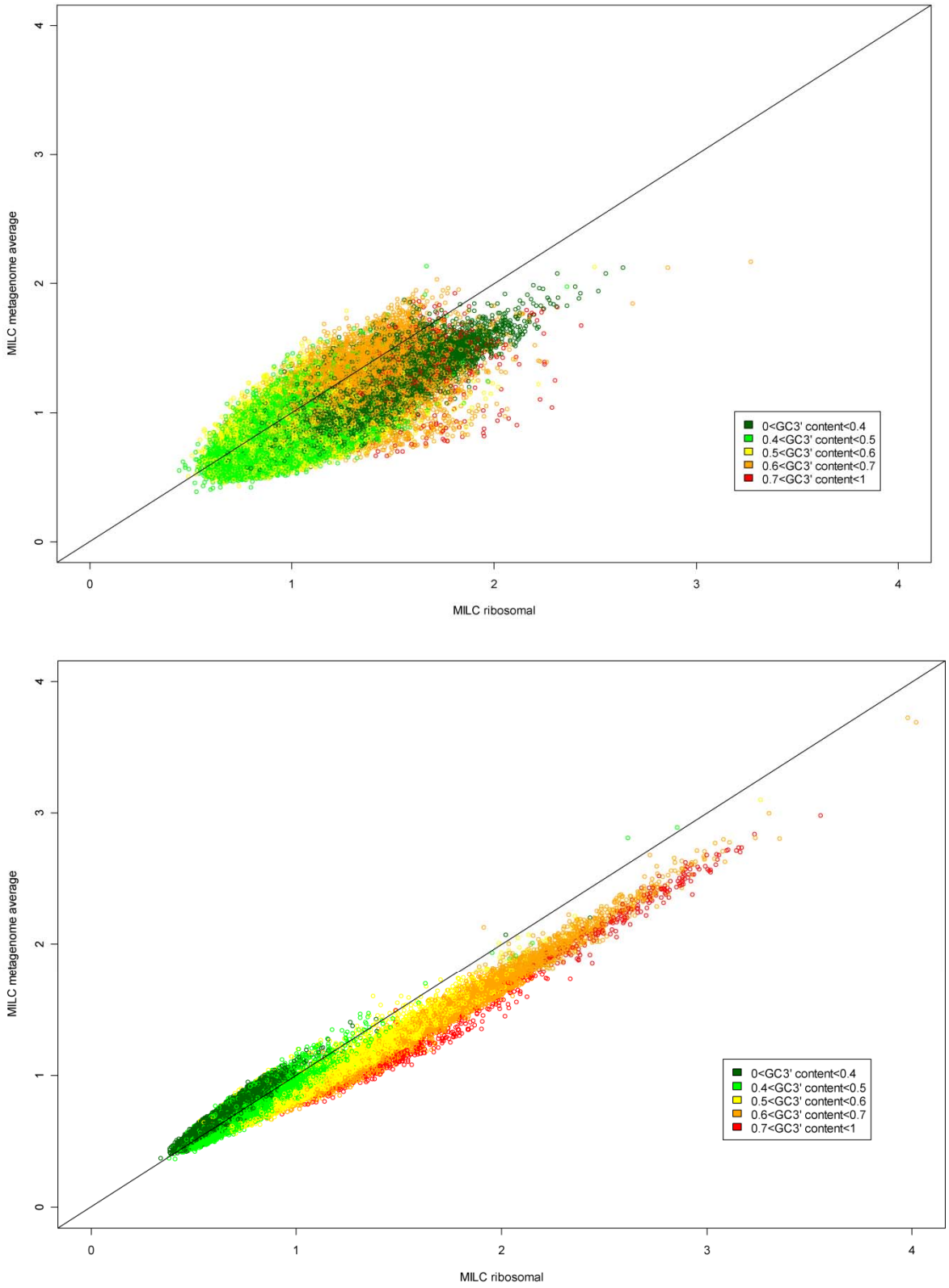

Figure 5: MILC metagenome average vs MILC ribosomal for the human distal gut metagenome (top) and acid mine drainage biofilm metagenome (bottom). For a detailed description of the plot see figure 3.

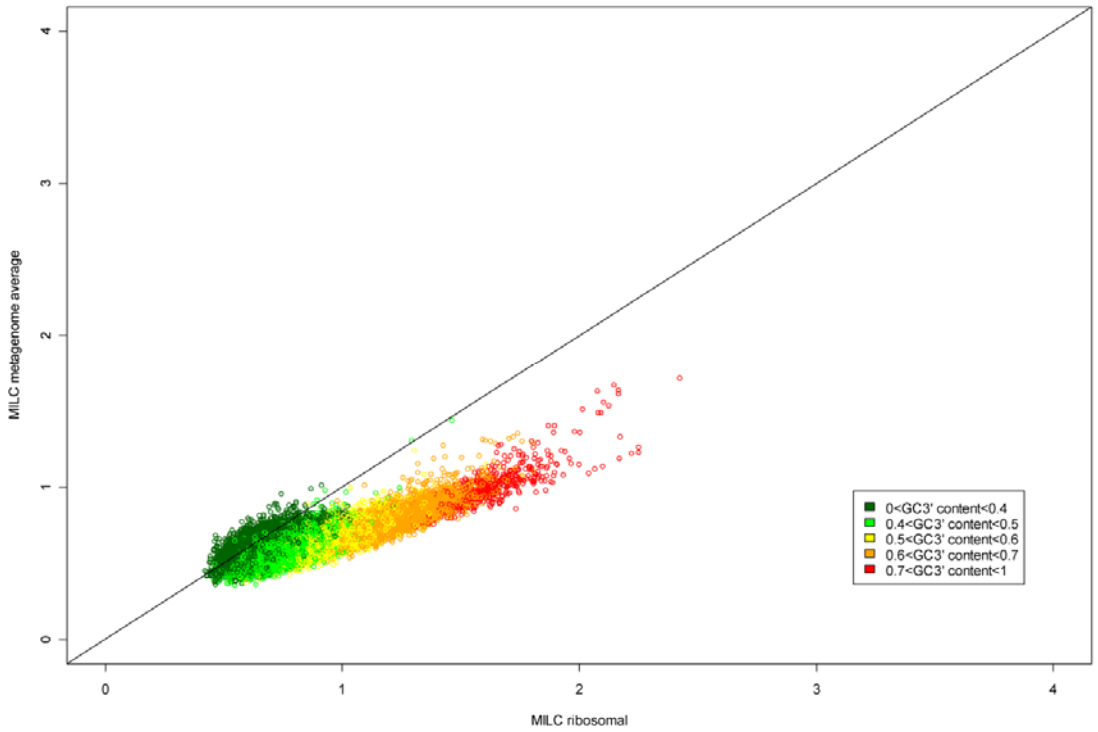

Figure 6: MILC metagenome average vs. MILC genome average for the acid mine sludge metagenome with randomized codons. The amino acid sequence of each ORF is preserved, while the codons are chosen randomly. The reference ribosomal set is not randomized. For a detailed description of the plot see figure 3.

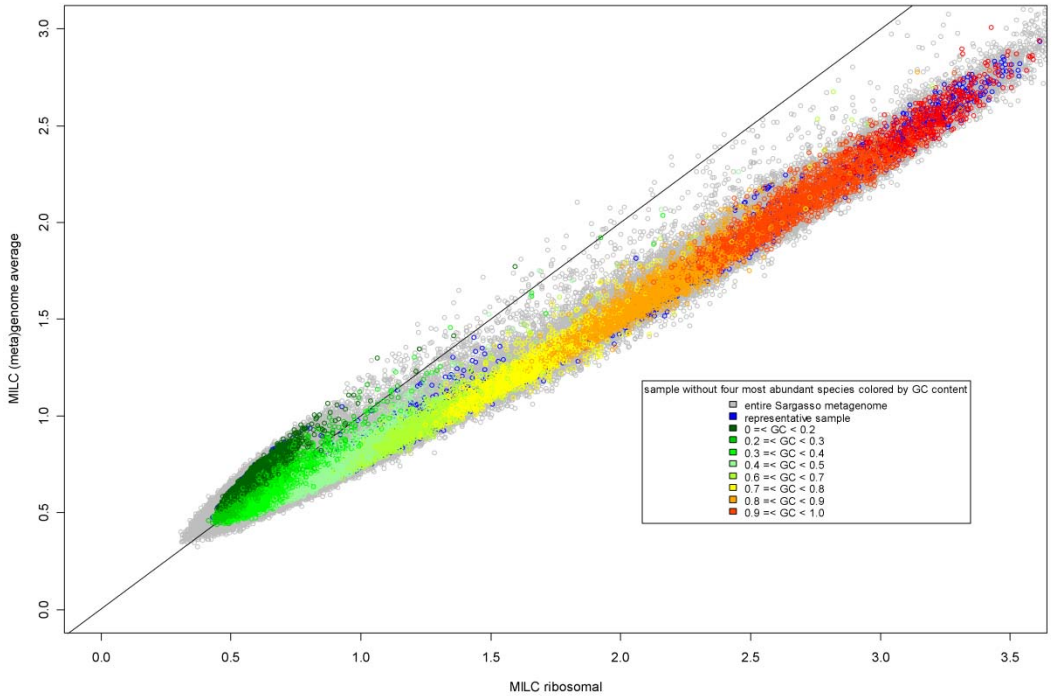

Figure 7: MILC metagenome average vs. MILC genome average for the Saragasso Sea metagenomic sample with the four most abundant species coloured blue. Only part of the sample is coloured according to GC content. For a detailed description of the rest of the plot see figure 3. (Perica and Lucic 2008).

Figure 8: Representation of the top 3% and low 3% of ORFs by MELP (red bars and green bars, respectively) within STRING/COG super families. Yellow bars show the representation of a super category in the entire metagenome. High and low MELP values are indicative of high and low expression of ORFs, respectively. For explanations of the STRING/COG super categories see supplementary table 1.

DISCUSSION

INSIGHTS INTO PROKARYOTIC DIVERSITY

The phylogenetic classification of organisms in metagenome samples through the 16S rRNA sequences present (tables 9-14) provides an estimate of prokaryotic diversity in the metagenomic samples. Extreme environments are expected to harbour a small number of species (Torsvik et al. 2002). The acid mine (table 9), United States (table 10) and Australian EBPR sludge (table 11) show the immense prevalence of the top 2 species, each present in more than 30% of the sample. On the other hand, the microbiome the human gut (table 12) has no species present in more than 2% of the sample despite its unfavourable environment. The gut is not a classic extreme metagenome because the organisms living there exist in a symbiosis with the host organism.

Before exploring the species diversity of the mouse's guts, an important note is that only 4% and 3% of assembled sequences of the lean and obese mouse, respectively, were functionally annotated (table 7). A homology search with databases like STRING/COG is only as successful with a comprehensive database. The mouse gut metagenomes obviously contain genes that still need to be more closely investigated. We assigned STRING/COG categories according to the 3-neighbour rule, which might have failed to annotate some of the genes that could have been assigned functions with a less stringent process. It was of higher importance to avoid false positives, i.e. genes assigned wrong functions, than to annotate all genes. The presence of close to 20% of the top species in mouse microbiomes (tables 13 and 14) might be biased by the lack of homologous sequences in the RDP database, similarly to the STRING/COG database.

METAGENOMES SHOW CODON USAGE BIAS AT THE LEVEL OF GENOMES

The distance in CUB from genomic average versus the distance from ribosomal reference set was plotted for each metagenome (figures 3-5). Figure 6 shows the plot of $MILC_{(metagenome)}$ vs. $MILC_{(ribosomal)}$ for the acid mine sludge biofilm metagenome where the amino acid sequence of ORFs from the metagenome was retained but the codons used

randomized, while the reference set was not randomized. Comparison of the CUB plot of the acid mine in figure 5 and the same metagenome randomized in figure 6 shows a clear loss in the shape of the graph. In figure 6, the distance from the genomic average is more or less constant for all ORFs. This clearly shows that a selection on the choice of codons exists at the level of metagenomes.

CUB is known to exist at the level of genomes of single organisms (Ikemura 1985). The influence of the most abundant prokaryotic species in a sample on the codon usage of a whole metagenome has been previously studied (Perica and Lucic 2008). Figure 7 demonstrates that the four most abundant species (coloured blue), comprising approximatly 25% of the whole sample, follow the same trend as the rest of the metagenome.

What is the biological purpose of maintaining a codon usage bias at the level of metagenomes? Horizontal gene transfer is known to be frequent between unicellular organisms living in the same ecological niche (Gogarten et al. 2002). The transfer of a gene is not in itself sufficient for the gene to function in a different organism, the expression of the gene needs to be regulated equivalently as well. We suggest that CUB is maintained in metagenomes to, at least at a transcriptional level, facilitate expression of horizontally transferred genes.

EXPRESSION LEVELS WITHIN COG FUNCTIONAL CATEGORIES REVEAL CLUES ON ENVIRONMENTAL ADAPTATION

MELP can help predict expression values of genes in genomes, and therefore metagenomes, which show codon usage bias. Homology through tools such as the STRING/COG database can help functionally classify genes. A marriage of these two computational methods reveals insights into the adaptations organisms make to their environment and is shown in figure 8. The counts of ORFs in the top 3% and low 3% by MELP values are shown within each STRING/COG super category.

PROTEIN STABILITY NEEDS TO BE RETAINED IN ACIDIC ENVIRONMENTS

The acid mine drainage metagenome is extracted from a pink biofilm that grew on the surface of a sulphuric acidic (pH ~0.8) solution at ~42 C. Previous proteomic studies have shown enrichment of the metagenome in ribosomal and chaperon proteins (Ram et al. 2005). In concordance, we found that ORFs with the functions of information storage and processing, i.e. category J – translation, ribosomal structure and biogenesis, are overrepresented in the categories with high MELP (figure 8). This indicates that the high expression of genes involved in translation and related activities is necessary to ensure protein stability at a very low pH at least partly by refolding. RNA processing and modification (super category A) shows very low MELP values, indicating that these processes might be altogether absent in such an acidic environment. The human distal gut metagenome also exists at a very low pH, hence it shows high expression values of the J category responsible for protein stability.

METAGENOMES IN DIGESTIVE TRACKS ARE ADAPTED FOR ENERGY HARVEST

Microorganisms living in the human gut synthesize essential amino acids and vitamins and process otherwise indigestible substances such as certain plant polysaccharides (Gill et al. 2006). In agreement with these functions is the overall high expression levels of metabolic categories energy production and conversion (C) and carbohydrate transport and metabolism (G) shown in figure 8.

The diet of a mouse is similar to that of humans. Consequently, the metagenomes of both the lean and obese mouse have high MILC values in metabolic categories C and G (figure 8). Interestingly, the lean mouse shows much higher MILC values in these categories than the obese mouse. Previous studies have shown that the microbiome in an obese mouse's gut has an increase in the number of genes involved in energy harvest (Turnbaugh et al. 2006). We propose that due to the smaller number of genes in the lean mouse microbiome organisms, it must optimize the expression of genes for energy harvest to ensure enough nutrition for itself. The microbiome of the obese mouse has a greater choice of genes for energy harvest, and therefore they need not all be optimized for translation.

METABOLISM OF METAGENOMES DEPENDS ON ENVIRONMENTAL CONDITIONS

Enhanced biological phosphorous removal (EBPR) sludge communities are maintained in reactors for biological removal of excess inorganic phosphate from wastewater. Two geographically distant metagenomes were used in this project, one located in the United States of America (US) and the other in Australia (OZ) (Martin et al. 2006). Regardless of the distance between them, both metagenomes show a very similar distribution of expression within COG categories (figure 8) due to very similar environmental conditions and metabolic processes. The same set of genes needs to be highly expressed in both metagenomes and, similarly, the same set of genes is rarely used. Genes in the category for replication, recombination and repair (L) have very low MELP values (figure 8) suggesting very high mutation rates in the metagenomes of EBPR sludge communities.

CONCLUSIONS

- Metagenomes show codon usage bias.
- Levels of expression of genes can be predicted through codon usage.
- Genes in a metagenome can be functionally characterised through computational homology methods.
- Important conclusions about the metabolism of metagenomes can be gained from the marriage of gene expression prediction and functional characterisation.
- Extreme metagenomes have adaptations to their environments that are evident in the control of gene expression.

REFERENCES

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller & D. J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25**,** 3389-3402.

Brown, T. A. 2002. *Genomes - Second Edition*. BIOS Scientific Publishers Ltd.

Chen, S. L., W. Lee, A. K. Hottes, L. Shapiro & H. H. McAdams (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 3480-3485.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity & J. M. Tiedje (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research,* 37**,** D141-D145.

Dinsdale, E. A., R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. L. Li, L. McDaniel, M. A. Moran, K. E. Nelson, C. Nilsson, R. Olson, J. Paul, B. R. Brito, Y. J. Ruan, B. K. Swan, R. Stevens, D. L. Valentine, R. V. Thurber, L. Wegley, B. A. White & F. Rohwer (2008) Functional metagenomic profiling of nine biomes. *Nature,* 452**,** 629-U8.

Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development,* 12**,** 640-649.

Foerstner, K. U., C. von Mering, S. D. Hooper & P. Bork (2005) Environments shape the nucleotide composition of genomes *EMBO reports,* 6**,** 1208–1213.

Gill, S. R., M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett & K. E. Nelson (2006) Metagenomic analysis of the human distal gut microbiome. *Science,* 312**,** 1355-1359.

Gogarten, J. P., W. F. Doolittle & J. G. Lawrence (2002) Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution,* 19**,** 2226-2238.

Goo, Y. A., J. Roach, G. Glusman, N. S. Baliga, K. Deutsch, M. Pan, S. Kennedy, S. DasSarma, W. V. Ng & L. Hood (2004) Low-pass sequencing for microbial comparative genomics. *Bmc Genomics,* 5**,** 19.

Gouy, M. & C. Gautier (1982) Codon Usage in Bacteria - Correlation with Gene Expressivity. *Nucleic Acids Research,* 10**,** 7055-7074.

Hugenholtz, P. & G. W. Tyson (2008) Microbiology - Metagenomics. *Nature,* 455**,** 481-483.

Ikemura, T. (1985) Codon Usage and Transfer-RNA Content in Unicellular and Multicellular Organisms. *Molecular Biology and Evolution,* 2**,** 13-34.

Jensen, L. J., M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork & C. von Mering (2009) STRING 8-a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research,* 37**,** D412-D416.

Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo & T. Ikemura (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. *Journal of Molecular Evolution,* 53**,** 290-298.

Karlin, S. & J. Mrazek (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *Journal of Bacteriology,* 182**,** 5238-5250.

Martin, H. G., N. Ivanova, V. Kunin, F. Warnecke, K. W. Barry, A. C. McHardy, C. Yeates, S. M. He, A. A. Salamov, E. Szeto, E. Dalin, N. H. Putnam, H. J. Shapiro, J. L. Pangilinan, I. Rigoutsos, N. C. Kyrpides, L. L. Blackall, K. D. McMahon & P. Hugenholtz (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotechnology,* 24**,** 1263-1269.

Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. W. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Q. Zheng, G. M. Rubin, M. D.

Adams & J. C. Venter (2000) A whole-genome assembly of Drosophila. *Science,* 287**,** 2196-2204.

Pace, N. R., D. A. Stahl, D. J. Lane & G. J. Olsen (1986) The Analysis of Natural Microbial-Populations by Ribosomal-RNA Sequences. *Advances in Microbial Ecology,* 9**,** 1-55.

Perica, T., & V. Lucic (2008) Personal communication.

Ram, R. J., N. C. VerBerkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. Blake, M. Shah, R. L. Hettich & J. F. Banfield (2005) Community proteomics of a natural microbial biofilm. *Science,* 308**,** 1915-1920.

Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson & E. Birney (2002) The bioperl toolkit: Perl modules for the life sciences. *Genome Research,* 12**,** 1611-1618.

Staley, J. T. & A. Konopka (1985) Measurement of Insitu Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annual Review of Microbiology,* 39**,** 321-346.

Supek, F. & K. Vlahovicek (2005) Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *Bmc Bioinformatics,* 6**,** 15.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin & D. A. Natale (2003) The COG database: an updated version includes eukaryotes. *Bmc Bioinformatics,* 4**,** 14.

Torsvik, V., L. Ovreas & T. F. Thingstad (2002) Prokaryotic diversity - Magnitude, dynamics, and controlling factors. *Science,* 296**,** 1064-1066.

Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz & E. M. Rubin (2005) Comparative metagenomics of microbial communities. *Science,* 308**,** 554-557.

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis & J. I. Gordon (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature,* 444**,** 1027-1031.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar & J. F. Banfield (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature,* 428**,** 37-43.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Y. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers & H. O. Smith (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science,* 304**,** 66-74.

Woese, C. R. & G. E. Fox (1977) Phylogenetic Structure of Prokaryotic Domain - Primary Kingdoms. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 5088-5090.

SUPPLEMENTARY

Supplementary table 1: Codes and descriptions of COG functional categories.

Code	Description
Information storage and processing	
J	Translation, ribosomal structure and biogenesis
A	RNA processing and modification
K	Transcription
L	Replication, recombination and repair
B	Chromatin structure and dynamics
Cellular processes and signalling	
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
Metabolism	
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
Poorly characterized	
R	General function prediction only
S	Function unknown
X	Uncharacterised