Published online 9 March 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 13* ***4207–4217***
doi:10.1093/nar/gkq140

SURVEY AND SUMMARY

Transposases are the most abundant, most ubiquitous genes in nature

Ramy K. Aziz[1,2,*], Mya Breitbart[3] and Robert A. Edwards[4,5]

[1]Computation Institute, University of Chicago, Chicago, IL 60637, USA, [2]Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, 11562 Cairo, Egypt, [3]College of Marine Science, University of South Florida, [4]Department of Computer Science, San Diego State University, San Diego, CA 92182 and [5]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

Received December 21, 2009; Revised and Accepted February 16, 2010

ABSTRACT

Genes, like organisms, struggle for existence, and the most successful genes persist and widely disseminate in nature. The unbiased determination of the most successful genes requires access to sequence data from a wide range of phylogenetic taxa and ecosystems, which has finally become achievable thanks to the deluge of genomic and metagenomic sequences. Here, we analyzed 10 million protein-encoding genes and gene tags in sequenced bacterial, archaeal, eukaryotic and viral genomes and metagenomes, and our analysis demonstrates that genes encoding transposases are the most prevalent genes in nature. The finding that these genes, classically considered as selfish genes, outnumber essential or housekeeping genes suggests that they offer selective advantage to the genomes and ecosystems they inhabit, a hypothesis in agreement with an emerging body of literature. Their mobile nature not only promotes dissemination of transposable elements within and between genomes but also leads to mutations and rearrangements that can accelerate biological diversification and—consequently—evolution. By securing their own replication and dissemination, transposases guarantee to thrive so long as nucleic acid-based life forms exist.

INTRODUCTION

Since life first emerged, organisms have been struggling for survival and competing over the finite resources within their ecosystems (1,2). This struggle for survival not only is confined to the organism level but it also applies to individual genes (3) and even non-coding DNA segments (4,5). As a corollary, a gene's success can be determined by its ability to persist in nature and to be spread throughout genomes and biomes (6). For this to take place, genes need some sequence plasticity to adapt to different environments while retaining enough sequence conservation to preserve the structure of their encoded proteins and the identity of their encoded biological functions (7,8).

Every time a new genome is sequenced, many genes are identified and annotated based on their homology to sequences available in databases, but new genes with novel functions are also identified, adding to the universal gene pool. To date, no study has systematically and directly surveyed the millions of protein-encoding genes (PEGs) deposited in sequence databases to identify their relative prevalence. There have been several challenges to such an endeavor: (i) the absence of numerical parameters to assess a gene's prevalence; (ii) the lack of fair representation of the tree of life within available sequence data (9,10) and (iii) the difficulty of defining what is meant by 'same gene' in different organisms and ecosystems.

To overcome these difficulties, (i) we calculated both the *abundance* and *ubiquity* of all known biological functions encoded in genomes and ecosystems to estimate their prevalence, with the assumption that these values will be correlated with gene fitness; (ii) we surveyed both genomic and metagenomic data sets to reduce bias caused by the uneven sampling of the tree of life in genomic data sets; and (iii) we defined similar genes as those encoding proteins with similar specific biological functions. In some instances, this definition could be regarded as an oversimplification, notably in cases of convergent evolution or homoplasy, where multiple genes of

*To whom correspondence should be addressed. Tel: +1 619 594 3137; Fax: +1 619 594 6746; Email: ramy.aziz@salmonella.org
Correspondence may also be addressed to Robert A. Edwards. Tel: +1 619 594 1672; Fax: +1 619 594 6746; Email: redwards@sciences.sdsu.edu

different origin evolve to perform similar biological functions. However, the majority of current gene annotations are specific enough to distinguish many instances of paralogous genes or different classes within gene/protein families. It is also understandable that different genes are under different selection pressures, as some are forced to endure mutations and tolerate sequence variability to escape pressure (e.g. bacterial genes encoding immunogenic proteins that are under pressure of the host immune system and genes encoding surface proteins that are easily recognized by predators) while others are under strict sequence conservation pressure (e.g. genes encoding housekeeping enzymes and essential biological functions).

Importantly, in determining gene prevalence we distinguished between ubiquity and abundance. Ubiquity is one of the indicators of essentiality, while abundance without ubiquity is an indicator of adaptive, organism-specific or habitat-specific functionality. In other words, ubiquitous genes are assumed to be those that carry essential functions and are thus indispensable in every genome (elements of core genomes) or every ecosystem (eco-essential genes). On the other hand, genes that are overly abundant in few ecosystems and absent in others are likely to play essential habitat-specific roles (e.g. photosynthesis, anaerobic metabolism, detoxification, etc.).

Contenders for the 'fittest gene' title include the gene encoding ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), an enzyme that plays a critical role in the fixation of carbon dioxide via the Calvin cycle and that has been touted as the single most successful, most abundant enzyme on the planet (11). Genes encoding ribosomal proteins are also plausible candidates. However, those are largely limited to cellular life forms, are not essential and almost absent in viruses (12), and are divergent between eukaryotes and prokaryotes. Additionally, DNA polymerase genes and other genes involved in DNA synthesis and nucleotide/nucleoside metabolism (e.g. ribonucleotide reductases, RNR) are essential for DNA-based life and are not restricted to cellular organisms, being found in viral genomes as well. Their essentiality favors them as strong competitors; yet, they are often present at one or few copies per genome. To our surprise, none of the previous candidate genes topped the list of the most abundant, most ubiquitous genes. Instead, our analysis singled out genes encoding transposases as the most abundant genes in genomes and metagenomes, and the most ubiquitous in metagenomes.

ANALYSIS OF GENOMES AND METAGENOMES

To determine the most abundant non-hypothetical PEG, we examined almost 10 million annotated genes or gene tags: over 3.2 million PEGs in fully sequenced viral, bacterial, archaeal and eukaryotic genomes (2137 genomes on 1 May 2009) and over 6.7 million environmental gene tags (EGT)—with significant matches to known proteins—in 187 random community genomes (metagenomes). For functional assignments, we mostly relied on the annotations available in the SEED database (13) because it uses subsystems-based controlled vocabulary curated by human experts and automatically propagated among genomes (14). For consistency, the same SEED subsystems were used for the annotation of all metagenomic data sets described in this study (15).

Analysis of complete genome sequences

We screened 2137 complete genomes (47 archaeal, 725 bacterial, 29 eukaryotic and 1336 viral genomes at the time this study was performed) available in the SEED database (URL: http://seed-viewer.theseed.org) and identified 37 258 PEGs (1.163% of all PEGs) annotated as transposase-related. Out of these, 26 625 (0.825% of all PEGs) were explicitly annotated as 'transposases', 360 were annotated as 'degenerate transposases', and then there were a variety of insertion sequence-related transposases, which may or may not be functional. Even when these ambiguous annotations were excluded from the final counts, transposases remained the most abundant PEGs in the completely sequenced genomes (Figure 1).

These data imply that out of a set of 2000 randomly sampled genes (the average number of genes in a typical bacterial genome), 22 genes are expected to encode transposases, at least 16 of which are likely functional. Obviously, genomes that have transposase genes tend to have them in multiple copies; this explains why although two-thirds of sequenced genomes (mostly viral) lack known functional transposases, the average number of transposases—when present—is 38.42 per genome (Table 1 and Supplementary Table S1). This observation is also in agreement with reports that transposases are unequally distributed among bacterial genomes, with higher abundance in facultative pathogens and free-living bacteria than in obligate pathogens and endosymbionts (16), and with extraordinarily high numbers in some species, e.g. *Crocosphaera watsonii* (17,18).

While the abundance of transposase genes in microbial genomes has been recognized for long time, only recently has it been exploited for inferring microbial cohabitation patterns and lateral gene transfer (19). Next to transposases, the most abundant functional roles in all sequenced genomes include ABC transporters, transcriptional regulators of different families, signal transduction kinases, chemotaxis proteins, acetyl- and glycosyl- transferases and cysteine desulfurase (Table 1 and Supplementary Table S1). On the other hand, the most ubiquitous functional roles in sequenced genomes are encoded by low-copy-number genes that consequently have a low overall abundance. Only four out of the 100 most ubiquitous functional genes have a mean copy number >2 per genome. These are genes encoding thioredoxin reductase; thioredoxin; cysteine desulfurase and the ABC transporter, ATP-binding protein (Supplementary Table S2). The list of most ubiquitous functional roles in genomes was topped by tRNA synthetases (Figure 1), and other genes associated with

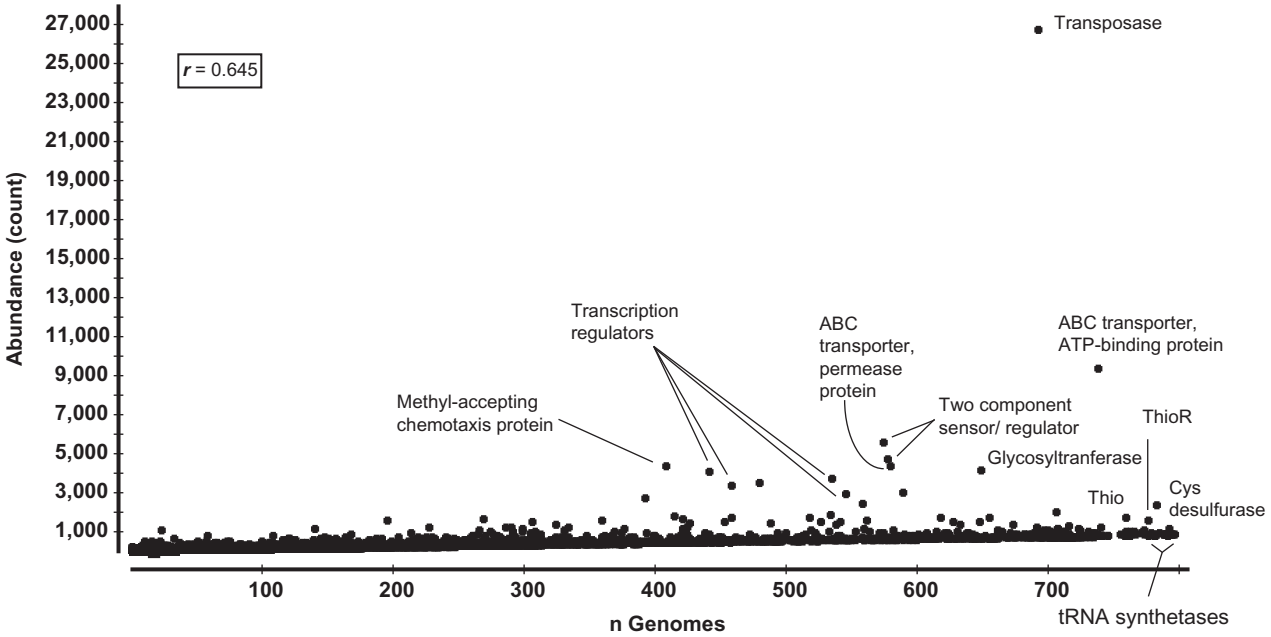

Figure 1. Abundance of different functional roles in 2137 genomes plotted against the ubiquity of these functional roles (defined as the number of genomes in which the functional role is represented at least once). *r*, Pearson's product moment correlation between abundance and ubiquity; *Cys*, cysteine; *Thio*, thioredoxin; *ThioR*, thioredoxin reductase. Proteins annotated solely based on their location or posttranslational modification but not their biological functions (e.g. membrane proteins, cytoplasmic proteins, secreted proteins, transmembrane proteins and generic lipoproteins) were excluded; an exception was the 'outer membrane protein' annotation as it describes specific bacterial proteins rather than protein localization.

protein synthesis and post-translational protein sorting (e.g. translation elongation factor and preprotein translocase, Supplementary Table S2).

Analysis of metagenomic sequences

In spite of the striking prevalence and high copy numbers of transposase genes in fully sequenced genomes, the use of those data sets is prone to biases. The available genomes unevenly represent the tree of life as they mostly correspond to cultured organisms from just four bacterial phyla (9). Moreover, there is an over-representation of microbes of interest to humans (20), such as bacterial pathogens and microbes used in agriculture or industry (21). Finally, while viruses are at least 10 times as abundant as bacteria in nature (22,23), sequenced viral genomes are lagging behind both in terms of numbers (~2:1 viral to bacterial genome ratio) and annotation quality (most encoded proteins are of unknown functions). In contrast, analysis of community genomes (metagenomes) offers a less-biased representation of life forms and biological functions in various habitats.

The term 'metagenome' describes the collective genomes found in a particular ecosystem (24,25). Since the first uncultured viral community genomic sequences were published in 2002 (26), metagenomics has emerged as a rapid and efficient method of identifying not only the species present in a given ecosystem but also the ecosystem-associated metabolic signatures or patterns (27–31). The emergence of low-cost, high-throughput next-generation sequencing technologies (32–37) has enabled the quick implementation of metagenomics in the analysis of different environments, allowing an unprecedented view of biodiversity (25,38–42). Over the past few years, metagenomic sequencing has been used to explore a wide range of environments, encompassing various marine ecosystems (28,43–47), hydrothermal vents (48,49), corals (50–52), salterns (53,54), soil (55–57), sludge (58), mines (59), human and animal guts (60–64) and lungs (65), microbialites (66,67) and even mosquitoes (31).

Metagenomic analysis is shifting the paradigm from organism/genome-centric to gene-centric and pathway-centric approaches to understanding biodiversity (68,69). Several bioinformatics and statistical tools allow the metabolic reconstruction of a particular ecosystem by enumerating EGTs in metagenomes and binning them either phylogenetically or biochemically (15,68,70–73), as well as the comparison of multiple metagenomes (59,74,75).

In this study, we followed a gene-centric approach by enumerating EGTs, and estimating the abundance and ubiquity of their different functional roles in 187 different metagenomic samples representing a broad range of environments. Assessing EGT abundance in metagenomic data is slightly different from determining PEG frequency in fully sequenced genomes. In genomes, a single, full-length copy of a gene reflects a single occurrence of that gene in one cell of an organism. In metagenomic data, multiple occurrence of an EGT can be attributed to either multiple copies of the same gene, multiple orthologs (from different genomes), multiple paralogs or just multiple sequences covering different parts of the exact same DNA segment. Moreover, the coding sequence length is a potential confounding factor: longer genes are more

Table 1. The 20 most abundant non-hypothetical protein-encoding genes in all sequenced genomes

Rank	Functional role	nG Count	V	A	B	E	C/n	%
1	Transposase	693 26 625	15 (1.1%) 21	31 (66%) 736	630 (86.9%) 25 226	17 (58.6%) 642	38.42	0.83
2	ABC transporter, ATP-binding protein	738 9382	1 (<1%) 1	39 (83%) 264	682 (94.1%) 8998	16 (55.2%) 119	12.71	0.29
3	Sensor histidine kinase	574 5575	–	22 (46.8%) 294	550 (75.9%) 5276	2 (6.9%) 5	9.71	0.17
4	DNA-binding response regulator	578 4708	–	13 (27.7%) 33	562 (77.5%) 4669	3 (10.3%) 6	8.20	0.15
5	Methyl-accepting chemotaxis protein	408 4389	1 (<1%) 1	15 (31.9%) 64	391 (53.9%) 4318	1 (3.4%) 6	10.76	0.14
6	ABC transporter, permease protein	580 4,377	–	33 (70.2%) 137	545 (75.2%) 4238	2 (6.9%) 2	7.55	0.14
7	Glycosyltransferase (EC 2.4.1.-)	649 4172	–	41 (87.2%) 287	598 (82.5%) 3863	10 (34.5%) 22	6.43	0.13
8	Transcriptional regulator, LysR family	441 4037	–	8 (17%) 10	430 (59.3%) 4017	3 (10.3%) 10	9.15	0.13
9	Transcriptional regulator, TetR family	535 3709	–	14 (29.8%) 45	521 (71.9%) 3664	–	6.93	0.12
10	Acetyltransferase, GNAT family	480 3516	–	19 (40.4%) 53	458 (63.2%) 3453	3 (10.3%) 10	7.33	0.11
11	Transcriptional regulator, AraC family	459 3382	1 (<1%) 1	7 (14.9%) 7	450 (62.1%) 3373	1 (3.4%) 1	7.37	0.11
12	Long-chain-fatty-acid–CoA ligase (EC 6.2.1.3)	589 2995	–	31 (66%) 68	533 (73.5%) 2728	25 (86.2%) 199	5.08	0.09
13	Transcriptional regulator, MarR family	546 2905	–	23 (48.9%) 71	522 (72%) 2831	1 (3.4%) 3	5.32	0.09
14	Permeases of the major facilitator superfamily	393 2733	1 (<1%) 1	12 (25.5%) 22	375 (51.7%) 2701	5 (17.2%) 9	6.95	0.09
15	Acetyltransferase (EC 2.3.1.-)	559 2436	4[a] (<1%) 4	22 (46.8%) 57	532 (73.4%) 2374	5 (17.2%) 5	4.36	0.08
16	Cysteine desulfurase (EC 2.8.1.7)	783 2362	–	36 (76.6%) 66	722 (99.6%) 2239	25 (86.2%) 57	3.02	0.07
17	3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100)	706 1975	–	27 (57.4%) 68	665 (91.7%) 1863	14 (48.3%) 44	2.80	0.06
18	Integrase	534 1829	70 (5.2%) 70	11 (23.4%) 19	448 (61.8%) 1729	5 (17.2%) 11	3.43	0.06
19	Outer membrane protein	415 1803	1 (<1%) 1	10 (21.3%) 12	402 (55.4%) 1786	2 (6.9%) 4	4.34	0.06
20	Permease of the drug/metabolite transporter (DMT) superfamily	518 1746	–	28 (59.6%) 53	486 (67%) 1688	4 (13.8%) 5	3.37	0.05

nG: number of genomes in which the functional role is present at least once; Count: number of genes in all sequenced genomes; V, A, B, E: viruses, archaea, bacteria, eukarya, respectively; C/n: average number of genes per positive genome; %: percentage of genes to the total number of genes in all genomes ($n = 3\,204\,918$).
[a]Acetyltransferase-like proteins that were missed in the automated analysis.

likely to be sampled by random sequencing (unless the sample is large enough to provide 100% coverage). For those reasons, the frequency of each EGT was normalized to the mean length of the most similar proteins [from BLASTX (76) results] to generate an abundance index, which was further divided by the number of informative sequence reads (those sequence tags matching annotated proteins in known databases) to generate a normalized abundance index (see the legend of Table 2 for more details).

The metagenomic data sets, which have been sequenced by different research groups, have been analyzed, consistently annotated and made publicly available through the metagenomics RAST server [http://metagenomics.theseed.org (15)]. They include both free-living and metazoan-associated viral, bacterial and eukaryotic sequences from autotrophic and heterotrophic communities from a wide variety of environments. In the analyzed metagenomes, the two most abundant functional genes are related to transposable elements [transposase and the retrotransposon-related p150 protein (77)]. Next to these, a set of photosynthesis-related genes; genes encoding viral structural, nonstructural, capsid and integrase proteins; genes associated with DNA replication; and genes involved in DNA synthesis are all among the most abundant biological functions in environmental metagenomes (Table 2 and Supplementary Table S3).

Since gene abundance in metagenomes is sensitive to sampling bias and sequencing depth, we also combined ubiquity with abundance data. The combined analysis confirmed the prevalence of transposases (abundant in 95% of the analyzed metagenomes) over the retrotransposon-related p150 genes (overly abundant in

only 36% of these metagenomes) and other replication and DNA metabolism-related genes that are equally ubiquitous but less abundant than transposases (Figure 2). The abundance of all analyzed non-hypothetical functions does not necessarily correlate with their ubiquity (Pearson correlation index = 0.524, Figure 2), i.e. many EGTs were pervasive in some ecosystems but absent in others (e.g. photosystem II proteins, p150 and viral structural genes; Table 2). Ubiquitous EGTs, on the other hand, include those matching transposases, DNA polymerases and enzymes involved in nucleotide metabolism (e.g. dTDP-glucose 4,6-dehydratase, UDP-glucose 4-epimerase and RNR; see Table 3 and Supplementary Table S4). Most of the ubiquitous EGTs are likely to be 'housekeeping' and essential for life, rather than habitat-specific (Figure 2). Additionally, many of these EGTs (e.g. DNA polymerases and RNRs) are found in all cellular and non-cellular biological entities, including viruses. As with genome sequence data, transposases are unequally distributed in ecosystems. This unequal distribution is in accordance with studies of ocean community genomics that showed a depth-dependent abundance of transposase genes (30) and a recent study that reported an unusually high abundance of transposase and retroviral integrase genes in a hydrothermal chimney biofilm (49).

Other than the predominance of transposases, ABC transporter ATP-binding proteins and phage integrases (Table 2), there is little agreement in the gene abundance data between genomes and metagenomes (Tables 1 and 2).

Table 2. The 20 most abundant functional roles in metagenomes

Rank	Functional role	nMG	nCAI
1	Transposase	178	4026.17
2	Retrotransposon-related p150 protein	69	3412.12
3	Viral structural protein	126	1909.75
4	ABC transporter, ATP-binding protein	170	1528.03
5	Replication-associated protein	32	1481.67
6	Photosystem II CP43 protein (PsbC)	47	1429.44
7	Photosystem II protein D2 (PsbD)	71	1224.89
8	Replication protein Rep	66	1213.18
9	Photosystem II protein D1 (PsbA)	83	930.2
10	Cytochrome b6-f complex subunit, cytochrome b6	51	925.32
11	Viral nonstructural protein	39	847.57
12	ATP synthase alpha chain (EC 3.6.3.14)	157	804.47
13	Ribonucleotide reductase of class Ia (aerobic), alpha subunit (EC 1.17.4.1)	165	776.57
14	Thymidylate synthase thyX (EC 2.1.1.-)	140	771.16
15	Single-stranded DNA-binding protein	151	769.41
16	Major capsid protein	100	745.51
17	ATP synthase beta chain (EC 3.6.3.14)	156	661.21
18	UDP-glucose 4-epimerase (EC 5.1.3.2)	169	657.36
19	Ribonucleotide reductase of class Ia (aerobic), beta subunit (EC 1.17.4.1)	150	652.32
20	Integrase	164	633.18

nMG: number of metagenomes in which the functional role is present at least once; nCAI: normalized cumulative abundance index. For each metagenome, a normalized abundance index (nAI) was calculated as the relative, length-normalized number of functional roles per million EGTs, and the nAI values for each functional role were added up to generate the normalized cumulative abundance index (nCAI).

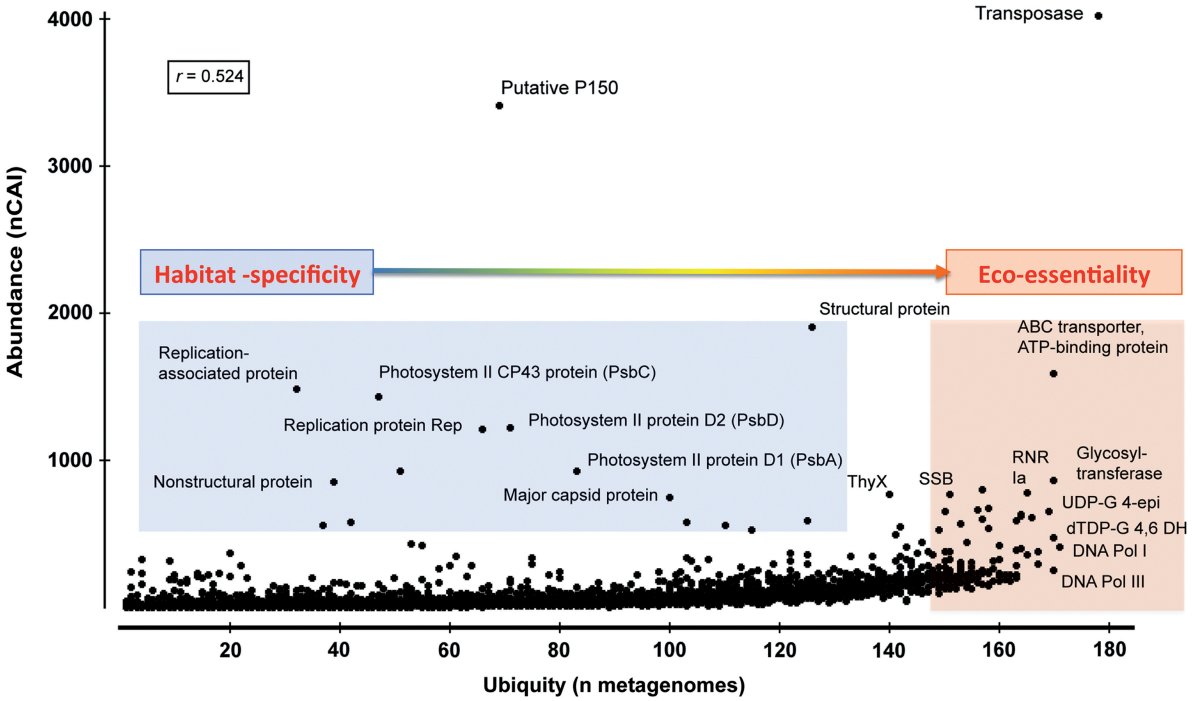

Figure 2. The normalized cumulative abundance indices (nCAI) of different functional roles in 187 metagenomes plotted against the ubiquity of these functional roles (defined as the number of metagenomes in which the functional role is represented at least once). *r*, Pearson's product moment correlation between abundance and ubiquity; DNA Pol, DNA polymerase; dTDP-G 4,6 DH, dTDP-glucose 4,6 dehydratase; Rep, replication-associated protein; RNR, ribonuleotide reductase; SSB, single-stranded DNA-binding protein; ThyX, thymidylate synthase thyX (EC 2.1.1.-); UDP-G 4-epi, UDP-glucose 4-epimerase.

Table 3. The 20 most ubiquitous functional roles in metagenomes

Rank	Functional role	nMG	%
1	Transposase	178	95.19
2	DNA polymerase I (EC 2.7.7.7)	171	91.44
3	dTDP-glucose 4,6-dehydratase (EC 4.2.1.46)	170	90.91
4	DNA polymerase III alpha subunit (EC 2.7.7.7)	170	90.91
5	ABC transporter, ATP-binding protein	170	90.91
6	UDP-glucose 4-epimerase (EC 5.1.3.2)	169	90.37
7	Heat shock protein 60 family chaperone GroEL	167	89.30
8	Chaperone protein DnaK	167	89.30
9	Ribonucleotide reductase of class II (coenzyme B12-dependent) (EC 1.17.4.1)	166	88.77
10	Ribonucleotide reductase of class Ia (aerobic), alpha subunit (EC 1.17.4.1)	165	88.24
11	Replicative DNA helicase (EC 3.6.1.-)	165	88.24
12	Integrase	164	87.70
13	Long-chain-fatty-acid–CoA ligase (EC 6.2.1.3)	164	87.70
14	Phosphate starvation-inducible protein PhoH, predicted ATPase	163	87.17
15	Carbamoyl-phosphate synthase large chain (EC 6.3.5.5)	163	87.17
16	DNA primase (EC 2.7.7.-)	163	87.17
17	Glycosyltransferase	163	87.17
18	Valyl-tRNA synthetase (EC 6.1.1.9)	163	87.17
19	Thymidylate synthase (EC 2.1.1.45)	163	87.17
20	ATP-dependent Clp protease ATP-binding subunit clpX	162	86.63

nMG: number of metagenomes in which the functional role is present at least once; %: percentage of nMG to the total number of metagenomes analyzed (187).

In genomic data, the most abundant functional roles reflect the over-representation of bacterial proteins in currently available fully sequenced genomes (2.5 million bacterial proteins versus 560 000 eukaryotic, 100 000 archaeal and 40 000 viral proteins). This bias may decrease when more viral genomes are sequenced and better annotated to reflect their actual distribution in nature. In metagenomic data, abundance indices reflect an overrepresentation of bacterial, archaeal and viral over eukaryotic sequences in currently available data sets; however, this overrepresentation is in agreement with reports that bacteria and archaea dominate the cellular world (78) while viruses are the most abundant biological entities (22,23).

DISCUSSION

The main assumption of this study is that the most successful genes are likely to be prevalent in genomes and ecosystems. We defined the most prevalent gene as the one 'spreading its DNA around' and not the one expressing the most protein molecules. Thus, while RuBisCO, for example, is claimed as the most abundant enzyme on Earth (11) based on the estimated number of its protein molecules, its genes are neither the most abundant nor most widely distributed (Supplementary Table S5). In addition, we focused on PEGs and did not include genes encoding ribosomal RNA in the analysis; those are absent in viruses and usually present in multiple copies in cellular genomes [1–15, mean = 4, (79)], which would place them at the 12th rank in gene abundance in all sequenced genomes (compare with Table 1).

This study demonstrates that transposases are the most abundant genes in both completely sequenced genomes and environmental metagenomes, and are also the most ubiquitous in metagenomes. Transposase genes encode DNA-binding enzymes, members of the polynucleotidyl transferase superfamily, that catalyze 'cut-and-paste' or 'copy-and-paste' reactions promoting the movement of DNA segments to new sites (80). The term transposase is often used to describe what are classically known as DNA transposases or type II transposases. These move double-stranded DNA directly by excision and insertion, and are sometimes associated with insertion sequences, but often just catalyze their own mobilization (81,82). The major group of dsDNA transposases is known as DDE transposases due to their possession of a non-contiguous, highly conserved catalytic triad of two aspartate (D) and one glutamate (E) residues (83). Other protein families that essentially use transposition but lack the DDE motif include tyrosine and serine recombinases, and rolling-circle transposases (82). In addition, within these transposase subclasses, several protein family domains [PFam domains (84)] have been described (49,83), yet a large fraction of transposases identified in genomes and metagenomes remain unclassified.

There are two other classes of transposable elements (Types I and III) that are distinguished as separate categories and were not as abundant or ubiquitous as Type II transposases in our analyzed data sets. Type I includes retrotransposons, which use the enzyme retrotransposase to move DNA by reverse transcription of an RNA intermediate (85). Retrotransposases (Type I transposases) are suggested to be responsible for the majority of 'junk' repeats, which make up >40% of the human genome and seem to code for no other genes (86–88). Type III transposable elements are associated with miniature inverted-repeat transposable elements (MITEs) (89,90). Transposases, in general, and Type II transposases, in particular, constitute a highly diverse group of enzymes. It is difficult to provide a robust, consistent scheme for classifying transposase sequences in ecosystems; however, structure-based classification schemes are being developed (83).

The prevalence of transposons (Type II) and retrotransposons (Type I) in eukaryotic genomes has been well documented, but in these genomes they are mostly associated with non-coding, repetitive DNA (91–93). Moreover, Type II transposases are continuously being detected in bacterial, archaeal and, to a lesser extent, bacteriophage genomes. In this work, we demonstrate that these jumping genes are also almost omnipresent in every ecosystem that contains nucleic acid-based life forms.

OUTLOOK

Transposase genes have been classically considered as 'selfish genes' with no other purpose than spreading themselves and are thus expected to be universal DNA parasites (6,85). If this were their only raison-d'être, they

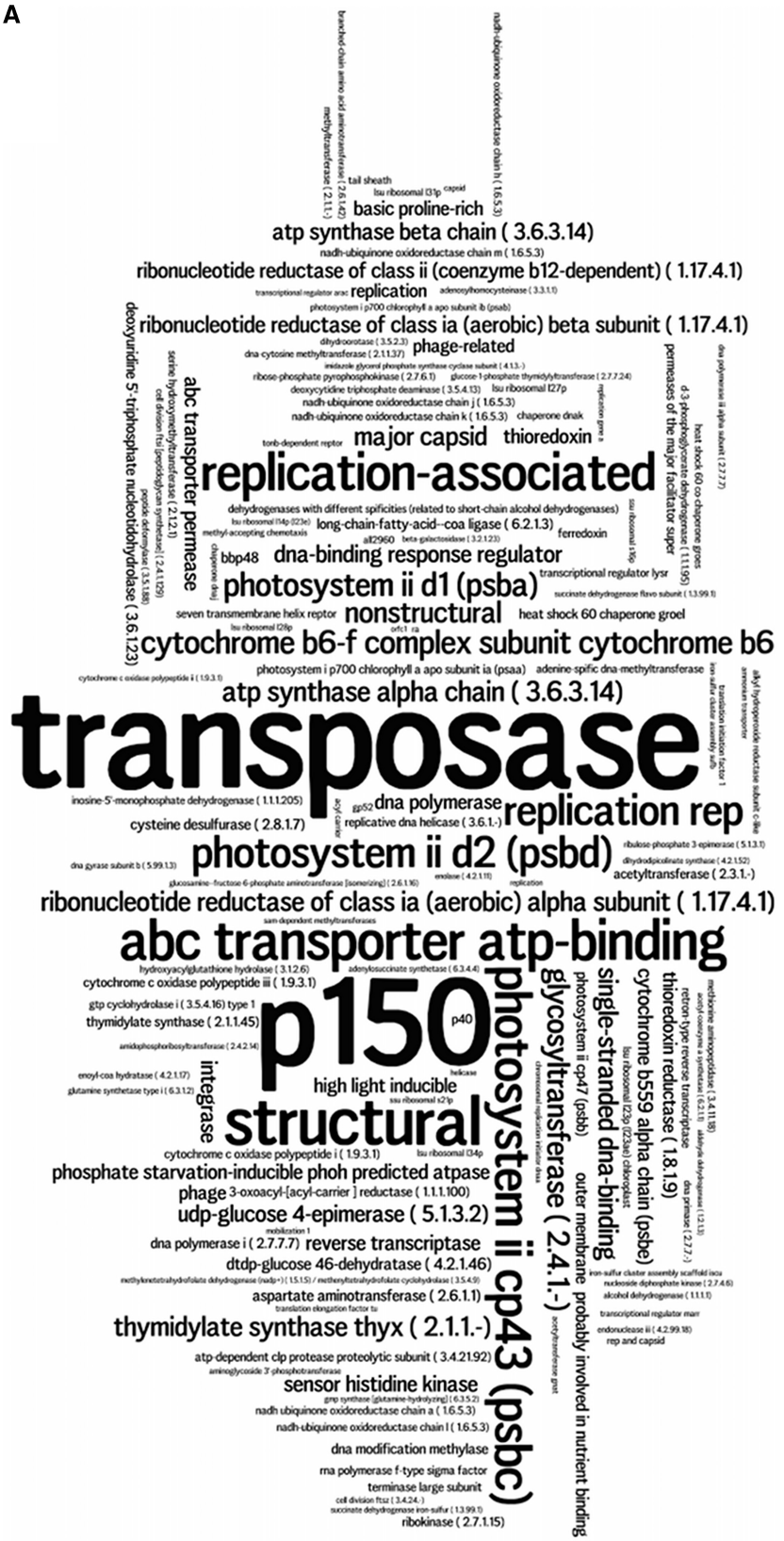

Figure 3. Word clouds (created on http://www.wordle.net) representing (**A**) the 100 most abundant functional roles (Supplementary Table S3) and (**B**) the 100 most ubiquitous functional roles (Supplementary Table S4) in metagenomes. The font size of each functional role is proportional to its (A) abundance index or (B) number of metagenomes in which it is present.

Figure 3. Continued.

have certainly fulfilled it by surviving, persisting and prevailing in all ecosystems. An open question is whether their ubiquity is also an indication of eco-essentiality. The finding that transposases are as ubiquitous as housekeeping DNA-processing enzymes but that they outnumber all essential genes (Figure 3) supports the idea that these mobile, self-replicating genes strive to inhabit and multiply in as many genomes as possible.

Besides the obvious detrimental effect that transposition can cause to host genomes—by inactivating housekeeping genes or impairing the chromosome's structure—transposases also play beneficial roles (92). For example, transposases may mobilize or activate genes that enhance their hosts' fitness (94,95), induce advantageous rearrangements (96) or enrich the host's gene pool (97–100). There are accruing documented examples of transposase genes co-opted by the host to encode transcription factors (99), centromere-binding proteins (100) or generators of diversity in the immune system (97,98), a process described as exaptation [or domestication, from a host-centric view (94)]. Such cases can involve one or a few transposases per genome or, as more recently shown, thousands of transposases (95).

Despite their ubiquity and abundance, there is neither evidence nor reason to believe that transposases encode

conserved essential cellular functions. In our opinion, the role of transposases as diversifying agents (94,101) is beneficial enough to be selected for; however, the cost of transposon-induced mutations also puts pressure on the cells to inactivate or delete their transposases (16,91,93,101).

In conclusion, the prevalence of transposases in metagenomes and completely sequenced genomes from bacteria, archaea, eukaryotes and viruses is in accordance with suggestions that they may offer a selective advantage to the genomes and ecosystems that they 'parasitize' (17,94,101). The diversification they induce in these genomes and ecosystems is arguably an essential way of maintaining, diversifying and evolving life on our planet.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Anca Segall, Elizabeth Dinsdale, Forest Rohwer, Peter Salamon, Jim Nulton and Ben Felts for stimulating discussions and helpful suggestions, and Moselio Schaechter and Stanley Maloy for valuable suggestions to improve the manuscript.

FUNDING

National Science Foundation, Division of Biological Infrastructure (DBI-0850356 to R.A.E. and DBI-0850206 to M.B.); the NMPDR project was supported by National Institutes of Health (HHSN266200400042C). Funding for open access charge: National Science Foundation, Division of Biological Infrastructure (DBI-0850356 to R.A.E.).

Conflict of interest statement. None declared.

REFERENCES

1. Darwin,C. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.
2. Huxley,J.S. (1942) *Evolution: The Modern Synthesis*, 1st edn. Harper, London.
3. Dawkins,R. (1976) *The Selfish Gene*. Oxford University Press, Oxford.
4. Edgell,D.R., Fast,N.M. and Doolittle,W.F. (1996) Selfish DNA: the best defense is a good offense. *Curr. Biol.*, **6**, 385–388.
5. Doolittle,W.F. and Sapienza,C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*, **284**, 601–603.
6. Orgel,L.E. and Crick,F.H. (1980) Selfish DNA: the ultimate parasite. *Nature*, **284**, 604–607.
7. Drummond,D.A. and Wilke,C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134**, 341–352.
8. Koonin,E.V. (2009) Darwinian evolution in the light of genomics. *Nucleic Acids Res.*, **37**, 1011–1034.
9. Hugenholtz,P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol.*, **3**, REVIEWS0003.
10. Wu,D., Hugenholtz,P., Mavromatis,K., Pukall,R., Dalin,E., Ivanova,N.N., Kunin,V., Goodwin,L., Wu,M., Tindall,B.J. *et al.* (2009) A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*, **462**, 1056–1060.
11. Dhingra,A., Portis,A.R. Jr and Daniell,H. (2004) Enhanced translation of a chloroplast-expressed RbcS gene restores small subunit levels and photosynthesis in nuclear RbcS antisense plants. *Proc. Natl Acad. Sci. USA*, **101**, 6315–6320.
12. Kristensen,D.M., Mushegian,A.R., Dolja,V.V. and Koonin,E.V. (2010) New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.*, **18**, 11–19.
13. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V., Chuang,H.Y., Cohoon,M., de Crecy-Lagard,V., Diaz,N., Disz,T., Edwards,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.
14. Aziz,R.K., Bartels,D., Best,A.A., DeJongh,M., Disz,T., Edwards,R.A., Formsma,K., Gerdes,S., Glass,E.M., Kubal,M. *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.
15. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
16. Ochman,H. and Davalos,L.M. (2006) The nature and dynamics of bacterial genomes. *Science*, **311**, 1730–1733.
17. Mes,T.H. and Doeleman,M. (2006) Positive selection on transposase genes of insertion sequences in the *Crocosphaera watsonii* genome. *J. Bacteriol.*, **188**, 7176–7185.
18. Zehr,J.P., Bench,S.R., Mondragon,E.A., McCarren,J. and DeLong,E.F. (2007) Low genomic diversity in tropical oceanic N2-fixing cyanobacteria. *Proc. Natl Acad. Sci. USA*, **104**, 17807–17812.
19. Hooper,S.D., Mavromatis,K. and Kyrpides,N.C. (2009) Microbial co-habitation and lateral gene transfer: what transposases can tell us. *Genome Biol.*, **10**, R45.
20. Aziz,R.K. (2009) The case for biocentric microbiology. *Gut. Pathogen.*, **1**, 16.
21. Ahmed,N. (2009) A flood of microbial genomes-do we need more? *PLoS ONE*, **4**, e5831.
22. Furuse,K., Osawa,S., Kawashiro,J., Tanaka,R., Ozawa,A., Sawamura,S., Yanagawa,Y., Nagao,T. and Watanabe,I. (1983) Bacteriophage distribution in human faeces: continuous survey of healthy subjects and patients with internal and leukaemic diseases. *J. Gen. Virol.*, **64(Pt 9)**, 2039–2043.
23. Breitbart,M. and Rohwer,F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.*, **13**, 278–284.
24. Handelsman,J., Rondon,M.R., Brady,S.F., Clardy,J. and Goodman,R.M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.*, **5**, R245–R249.
25. Riesenfeld,C.S., Schloss,P.D. and Handelsman,J. (2004) Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, **38**, 525–552.
26. Breitbart,M., Salamon,P., Andresen,B., Mahaffy,J.M., Segall,A.M., Mead,D., Azam,F. and Rohwer,F. (2002) Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA*, **99**, 14250–14255.
27. 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.
28. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
29. 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.
30. DeLong,E.F., Preston,C.M., Mincer,T., Rich,V., Hallam,S.J., Frigaard,N.U., Martinez,A., Sullivan,M.B., Edwards,R., Brito,B.R. *et al.* (2006) Community genomics among stratified

microbial assemblages in the ocean's interior. *Science*, **311**, 496–503.
31. Dinsdale,E.A., Edwards,R.A., Hall,D., Angly,F., Breitbart,M., Brulc,J.M., Furlan,M., Desnues,C., Haynes,M., Li,L. *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.
32. Ronaghi,M., Karamohamed,S., Pettersson,B., Uhlen,M. and Nyren,P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, **242**, 84–89.
33. Ronaghi,M. (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Res.*, **11**, 3–11.
34. Bennett,S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433–438.
35. Kartalov,E.P. and Quake,S.R. (2004) Microfluidic device reads up to four consecutive base pairs in DNA sequencing-by-synthesis. *Nucleic Acids Res.*, **32**, 2873–2879.
36. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
37. Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**, 16–18.
38. Schloss,P.D. and Handelsman,J. (2003) Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.*, **14**, 303–310.
39. Handelsman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**, 669–685.
40. Edwards,R.A. and Rohwer,F. (2005) Viral metagenomics. *Nat. Rev. Microbiol.*, **3**, 504–510.
41. Xu,J. (2006) Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol. Ecol.*, **15**, 1713–1731.
42. Casas,V. and Rohwer,F. (2007) Phage metagenomics. *Methods Enzymol.*, **421**, 259–268.
43. Angly,F.E., Felts,B., Breitbart,M., Salamon,P., Edwards,R.A., Carlson,C., Chan,A.M., Haynes,M., Kelley,S., Liu,H. *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol.*, **4**, e368.
44. Yooseph,S., Sutton,G., Rusch,D.B., Halpern,A.L., Williamson,S.J., Remington,K., Eisen,J.A., Heidelberg,K.B., Manning,G., Li,W. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.*, **5**, e16.
45. Rusch,D.B., Halpern,A.L., Sutton,G., Heidelberg,K.B., Williamson,S., Yooseph,S., Wu,D., Eisen,J.A., Hoffman,J.M., Remington,K. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, e77.
46. McDaniel,L., Breitbart,M., Mobberley,J., Long,A., Haynes,M., Rohwer,F. and Paul,J.H. (2008) Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. *PLoS ONE*, **3**, e3263.
47. Persson,O.P., Pinhassi,J., Riemann,L., Marklund,B.I., Rhen,M., Normark,S., Gonzalez,J.M. and Hagstrom,A. (2009) High abundance of virulence gene homologues in marine bacteria. *Environ. Microbiol.*, **11**, 1348–1357.
48. Grzymski,J.J., Murray,A.E., Campbell,B.J., Kaplarevic,M., Gao,G.R., Lee,C., Daniel,R., Ghadiri,A., Feldman,R.A. and Cary,S.C. (2008) Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility. *Proc. Natl Acad. Sci. USA*, **105**, 17516–17521.
49. Brazelton,W.J. and Baross,J.A. (2009) Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. *ISME J.*, **3**, 1420–1424.
50. Dinsdale,E.A., Pantos,O., Smriga,S., Edwards,R.A., Angly,F., Wegley,L., Hatay,M., Hall,D., Brown,E., Haynes,M. *et al.* (2008) Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE*, **3**, e1584.
51. Vega Thurber,R.L., Barott,K.L., Hall,D., Liu,H., Rodriguez-Mueller,B., Desnues,C., Edwards,R.A., Haynes,M., Angly,F.E., Wegley,L. *et al.* (2008) Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. *Proc. Natl Acad. Sci. USA*, **105**, 18413–18418.
52. Vega Thurber,R., Willner-Hall,D., Rodriguez-Mueller,B., Desnues,C., Edwards,R.A., Angly,F., Dinsdale,E., Kelly,L. and Rohwer,F. (2009) Metagenomic analysis of stressed coral holobionts. *Environ. Microbiol.*, **11**, 2148–2163.
53. Santos,F., Meyerdierks,A., Pena,A., Rossello-Mora,R., Amann,R. and Anton,J. (2007) Metagenomic approach to the study of halophages: the environmental halophage 1. *Environ. Microbiol.*, **9**, 1711–1723.
54. Rodriguez-Brito,B., Li,L., Wegley,L., Furlan,M., Angly,F., Breitbart,M., Buchanan,J., Desnues,C., Dinsdale,E., Edwards,R. *et al.* (2010) Viral and microbial community dynamics in four aquatic environments. *ISME J.* [12 February 2010, Epub ahead of print].
55. Kim,K.H., Chang,H.W., Nam,Y.D., Roh,S.W., Kim,M.S., Sung,Y., Jeon,C.O., Oh,H.M. and Bae,J.W. (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl. Environ. Microbiol.*, **74**, 5975–5985.
56. Pang,H., Zhang,P., Duan,C.J., Mo,X.C., Tang,J.L. and Feng,J.X. (2009) Identification of cellulase genes from the metagenomes of compost soils and functional characterization of one novel endoglucanase. *Curr. Microbiol.*, **58**, 404–408.
57. Zhang,K., He,J., Yang,M., Yen,M. and Yin,J. (2009) Identifying natural product biosynthetic genes from a soil metagenome by using T7 phage selection. *Chembiochem*, **10**, 2599–2606.
58. Kunin,V., He,S., Warnecke,F., Peterson,S.B., Garcia Martin,H., Haynes,M., Ivanova,N., Blackall,L.L., Breitbart,M., Rohwer,F. *et al.* (2008) A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res.*, **18**, 293–297.
59. Edwards,R.A., Rodriguez-Brito,B., Wegley,L., Haynes,M., Breitbart,M., Peterson,D.M., Saar,M.O., Alexander,S., Alexander,E.C. Jr and Rohwer,F. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, **7**, 57.
60. Breitbart,M., Hewson,I., Felts,B., Mahaffy,J.M., Nulton,J., Salamon,P. and Rohwer,F. (2003) Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.*, **185**, 6220–6223.
61. 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.
62. Frank,D.N. and Pace,N.R. (2008) Gastrointestinal microbiology enters the metagenomics era. *Curr. Opin. Gastroenterol.*, **24**, 4–10.
63. Turnbaugh,P.J., Hamady,M., Yatsunenko,T., Cantarel,B.L., Duncan,A., Ley,R.E., Sogin,M.L., Jones,W.J., Roe,B.A., Affourtit,J.P. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.
64. Tuohy,K.M., Gougoulias,C., Shen,Q., Walton,G., Fava,F. and Ramnani,P. (2009) Studying the human gut microbiota in the trans-omics era–focus on metagenomics and metabonomics. *Curr. Pharm. Des.*, **15**, 1415–1427.
65. Willner,D., Furlan,M., Haynes,M., Schmieder,R., Angly,F.E., Silva,J., Tammadoni,S., Nosrat,B., Conrad,D. and Rohwer,F. (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE*, **4**, e7370.
66. Desnues,C., Rodriguez-Brito,B., Rayhawk,S., Kelley,S., Tran,T., Haynes,M., Liu,H., Furlan,M., Wegley,L., Chau,B. *et al.* (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature*, **452**, 340–343.
67. Breitbart,M., Hoare,A., Nitti,A., Siefert,J., Haynes,M., Dinsdale,E., Edwards,R., Souza,V., Rohwer,F. and Hollander,D. (2009) Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Cienegas, Mexico. *Environ. Microbiol.*, **11**, 16–34.
68. Eisen,J.A. (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol.*, **5**, e82.
69. Hugenholtz,P. and Tyson,G.W. (2008) Microbiology: metagenomics. *Nature*, **455**, 481–483.
70. Angly,F., Rodriguez-Brito,B., Bangor,D., McNairnie,P., Breitbart,M., Salamon,P., Felts,B., Nulton,J., Mahaffy,J. and Rohwer,F. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics*, **6**, 41.

71. Krause,L., Diaz,N.N., Goesmann,A., Kelley,S., Nattkemper,T.W., Rohwer,F., Edwards,R.A. and Stoye,J. (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.*, **36**, 2230–2239.
72. Kunin,V., Copeland,A., Lapidus,A., Mavromatis,K. and Hugenholtz,P. (2008) A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.*, **72**, 557–578.
73. Schloss,P.D. and Handelsman,J. (2008) A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics*, **9**, 34.
74. Rodriguez-Brito,B., Rohwer,F. and Edwards,R.A. (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics*, **7**, 162.
75. Huson,D.H., Richter,D.C., Mitra,S., Auch,A.F. and Schuster,S.C. (2009) Methods for comparative metagenomics. *BMC Bioinformatics*, **10(Suppl. 1)**, S12.
76. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., 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.
77. Sassaman,D.M., Dombroski,B.A., Moran,J.V., Kimberland,M.L., Naas,T.P., DeBerardinis,R.J., Gabriel,A., Swergold,G.D. and Kazazian,H.H. Jr (1997) Many human L1 elements are capable of retrotransposition. *Nat. Genet.*, **16**, 37–43.
78. Whitman,W.B., Coleman,D.C. and Wiebe,W.J. (1998) Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA*, **95**, 6578–6583.
79. Lee,Z.M., Bussema,C., 3rd. and Schmidt,T.M. (2009) rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Res.*, **37**, D489–D493.
80. Rice,P.A. and Baker,T.A. (2001) Comparative architecture of transposase and integrase complexes. *Nat. Struct. Biol.*, **8**, 302–307.
81. Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (2002) *Mobile DNA II*. ASM press, Washington, DC.
82. Curcio,M.J. and Derbyshire,K.M. (2003) The outs and ins of transposition: from mu to kangaroo. *Nat. Rev. Mol. Cell. Biol.*, **4**, 865–877.
83. Hickman,A.B., Chandler,M. and Dyda,F. (2010) Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. *Crit. Rev. Biochem. Mol. Biol.*, **45**, 50–69.
84. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
85. Wright,S. and Finnegan,D. (2001) Genome evolution: sex and the transposable element. *Curr. Biol.*, **11**, R296–R299.
86. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
87. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
88. Cordaux,R. and Batzer,M.A. (2009) The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.*, **10**, 691–703.
89. Wessler,S.R., Bureau,T.E. and White,S.E. (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.*, **5**, 814–821.
90. Feschotte,C. and Mouches,C. (2000) Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Mol. Biol. Evol.*, **17**, 730–737.
91. Ohshima,K., Hattori,M., Yada,T., Gojobori,T., Sakaki,Y. and Okada,N. (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol.*, **4**, R74.
92. Mills,R.E., Bennett,E.A., Iskow,R.C. and Devine,S.E. (2007) Which transposable elements are active in the human genome? *Trends Genet.*, **23**, 183–191.
93. Pace,J.K. 2nd and Feschotte,C. (2007) The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res.*, **17**, 422–432.
94. Benjak,A., Forneck,A. and Casacuberta,J.M. (2008) Genome-wide analysis of the "cut-and-paste" transposons of grapevine. *PLoS ONE*, **3**, e3107.
95. Nowacki,M., Higgins,B.P., Maquilan,G.M., Swart,E.C., Doak,T.G. and Landweber,L.F. (2009) A functional role for transposases in a large eukaryotic genome. *Science*, **324**, 935–938.
96. Mendiola,M.V., Bernales,I. and de la Cruz,F. (1994) Differential roles of the transposon termini in IS91 transposition. *Proc. Natl Acad. Sci. USA*, **91**, 1922–1926.
97. Agrawal,A., Eastman,Q.M. and Schatz,D.G. (1998) Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*, **394**, 744–751.
98. Hiom,K., Melek,M. and Gellert,M. (1998) DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. *Cell*, **94**, 463–470.
99. Lin,R., Ding,L., Casola,C., Ripoll,D.R., Feschotte,C. and Wang,H. (2007) Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. *Science*, **318**, 1302–1305.
100. Casola,C., Hucks,D. and Feschotte,C. (2008) Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. *Mol. Biol. Evol.*, **25**, 29–41.
101. Condit,R., Stewart,F.M. and Levin,B.R. (1988) The population biology of bacterial transposons: A priori conditions for maintenance as parasitic DNA. *Am. Nat.*, **132**, 129–147.