Environmental Microbiology (2009) **11**(7), 1752–1766 doi:10.1111/j.1462-2920.2009.01901.x

Metagenomic signatures of 86 microbial and viral metagenomes

Dana Willner,[1]* Rebecca Vega Thurber[1,2] and Forest Rohwer[1,3]

[1]*Department of Biology, LS301, and* [3]*Center for Microbial Sciences, San Diego State University, 5500 Campanile Dr., San Diego, CA 92182, USA.*

[2]*Department of Biological Sciences, Florida International University, 3000 NE 151 St., Miami, FL 33181, USA.*

Summary

Previous studies have shown that dinucleotide abundances capture the majority of variation in genome signatures and are useful for quantifying lateral gene transfer and building molecular phylogenies. Metagenomes contain a mixture of individual genomes, and might be expected to lack compositional signatures. In many metagenomic data sets the majority of sequences have no significant similarities to known sequences and are effectively excluded from subsequent analyses. To circumvent this limitation, di-, tri- and tetranucleotide abundances of 86 microbial and viral metagenomes consisting of short pyrosequencing reads were analysed to provide a method which includes all sequences that can be used in combination with other analysis to increase our knowledge about microbial and viral communities. Both principal component analysis and hierarchical clustering showed definitive groupings of metagenomes drawn from similar environments. Together these analyses showed that dinucleotide composition, as opposed to tri- and tetranucleotides, defines a metagenomic signature which can explain up to 80% of the variance between biomes, which is comparable to that obtained by functional genomics. Metagenomes with anomalous content were also identified using dinucleotide abundances. Subsequent analyses determined that these metagenomes were contaminated with exogenous DNA, suggesting that this approach is a useful metric for quality control. The predictive strength of the dinucleotide composition also opens the possibility of assigning ecological classifications to unknown fragments. Environmental selection may be responsible for this dinucleotide signature through direct selection of specific compositional signals; however, simulations suggest that the environment may select indirectly by promoting the increased abundance of a few dominant taxa.

Received 8 December 2008; accepted 31 January 2009. *For correspondence. E-mail willner9@aol.com; Tel. (+1) 619 594 1336; Fax (+1) 619 594 5676.

Introduction

Several studies have demonstrated sequence-based signatures in a wide variety of individual genomes (Burge *et al*., 1992; Karlin *et al*., 1997; 1998; Campbell *et al*., 1999; Gentles and Karlin, 2001), and genomic signatures have been both visualized and validated by chaos game representations (Deschavanne *et al*., 1999; Wang *et al*., 2005). Applications of genomic signature analysis include detection of lateral gene transfer in bacteria, molecular phylogeny, and binning of individual metagenomic fragments either for taxonomic assignment, or to infer possible host ranges for viruses (Teeling *et al*., 2004; Abe *et al*., 2005; Chapus *et al*., 2005; Dufraigne *et al*., 2005; Fertil *et al*., 2005; Woyke *et al*., 2006). Metagenomic data sets consist of DNA sequence fragments from consortia which contain both culturable and recalcitrant microbes and viruses. These data sets often contain a high percentage of fragments which show no significant similarity to known sequences, which has raised concerns among researchers about the validity of descriptions based on such a small subset of the data (Schloss and Handelsman, 2003; Teeling *et al*., 2004). Teeling and colleagues used tetranucleotide frequencies to assign taxonomic classifications to fosmid-sized fragments. The tetranucleotide abundances had high discriminatory power in metagenomes with low community diversity (Teeling *et al*., 2004). Environmental metagenomes, however, usually have high phylogenetic diversity and smaller sequences (less than 1 kb), many of which are not classifiable using database searches (Breitbart *et al*., 2002; Angly *et al*., 2006; Martin-Cuadrado *et al*., 2007; Wegley *et al*., 2007; Desnues *et al*., 2008; Dinsdale *et al*., 2008a). For example, in metagenomes derived from marine ecosystems, these 'unknown' sequences comprise up to 90% of the total sequence data, while in microbialites they account for more than 99% (Desnues *et al*., 2008). Therefore, similarity-based comparisons and characterizations, such as best BLAST hits, disregard an overwhelming proportion of metagenomic sequences, as they are incapable of classifying unknowns.

Analysis of the occurrence of oligonucleotide frequencies in eukaryotic, microbial (*Bacteria* and *Archaea*), and viral genomes has demonstrated that individual genomes possess sequence-based signatures, which reflect the specific patterns of dinucleotide abundances (Burge *et al*., 1992; Karlin and Ladunga, 1994; Karlin and Burge, 1995; Blaisdell *et al*., 1996; Karlin and Mrazek, 1997; Karlin *et al*., 1998; Campbell *et al*., 1999; Gentles and Karlin, 2001). This dinucleotide signature has more phylogenetic signal than genomic GC content, since the percentage of G and C nucleotides can vary widely across genomes (Teeling *et al*., 2004). The overabundance or relative absence of particular dinucleotides has been linked to DNA structural preferences as well as context-dependent cues, such as potential mutations and regulatory regions (Karlin *et al*., 1998). For example, vertebrate genomes, and especially the human genome, show a depression of the frequency of CG dinucleotides. This phenomenon has been hypothesized to be driven by the propensity for CG to TA mutations that arise following methylation and subsequent deamination (Burge *et al*., 1992; Karlin *et al*., 1998). Additionally, dinucleotide signatures have been shown to be influenced by amino acid preferences and codon biases, which in turn may be driven by the environment (Karlin *et al*., 1998; Singer and Hickey, 2003; Goodarzi *et al*., 2007; Paul *et al*., 2008). Nucleotide frequencies may also reflect environmental conditions such as temperature, pH, metal concentrations and other properties of an organism's habitat (Karlin *et al*., 1998). Finally, obligate intracellular bacterial parasites and viruses that require endogenous cellular machinery to replicate have genomic signatures that tend to mirror those of the host (Karlin *et al*., 1998; Campbell *et al*., 1999).

Dinsdale and colleagues (2008b) recently showed that metagenomes derived from similar environments exhibit similar metabolic profiles, based on functional annotations of component genes. However, as with previous metagenomes, these comparisons were based exclusively on sequences that contained identifiable protein-encoding genes, excluding a large proportion of the sequences from the analysis. Here we present an alternative approach for profiling a nearly identical set of metagenomes based on GC and di-, tri- and tetranucleotide frequencies to determine if related metagenomes show similar oligomeric composition. This was not expected to occur because metagenomes contain a mixture of sequences derived from a variety of individual genomes. Since GC content has been shown to be a poor discriminatory tool in binning of metagenomic sequences, it was used here as a comparison to evaluate the effectiveness of dinucleotides in characterizing metagenomes (Teeling *et al*., 2004). Both principal component analysis (PCA) and hierarchical clustering showed definitive groupings of metagenomes derived from similar environments based solely on dinucleotide abundances. We hypothesize that these groupings are driven by environmental selection, either for particular microbial and viral taxa with distinctive dinucleotide biases or by direct selection for DNA composition.

Results and discussion

Viral metagenomes have reduced GC content

GC content varies widely between individual genomes, as well as between metagenomes from different environments (Rocha and Danchin, 2002; Foerstner *et al*., 2005; Raes *et al*., 2007). Previous studies have used GC content to categorize both genomes and metagenomes and to bin sequences for taxonomic assignment, despite evidence that it performs poorly as a classification metric (Rocha and Danchin, 2002; Teeling *et al*., 2004; Foerstner *et al*., 2005; Raes *et al*., 2007). Here, GC content was used to characterize metagenomes to provide a standard for comparison with the performance of oligomeric abundances in classifying and describing both microbiomes and viromes.

To assess trends in GC content among the 86 metagenomes, average metagenomic GC content and standard deviations were calculated (Tables S1 and S2) and descriptive statistics were compiled by metagenome type (i.e. microbiomes or viromes; Fig. 1) and by biome (Fig. 2). GC content in both the microbial and viral data sets follows an approximately normal distribution (Fig. 1). Overall, the viromes have a lower average GC content, with a mean of 45.19% versus 49.56%. This mirrors the 4% average difference between GC content in phage and their bacterial hosts reported by Rocha and Danchin (2002). Similarly, the average GC content of all microbial and viral genomes available from NCBI (http://www.ncbi.nlm.nih.gov) are 49.70% and 44.32%, respectively, an approximately 5% difference. Although there is no clear consensus, it has been suggested that the increased AT content of viruses may be due either to shorter genome lengths or to an increased energetic cost associated with G and C nucleotides (Rocha and Danchin, 2002).

A two-sample *t*-test to compare the mean GC percentages between viral and microbial metagenomes revealed that GC content was significantly different with a *P*-value less than 0.0001, and the 95% confidence interval for the true mean difference is (2.52, 6.97). The relatively lower overall GC content of the viromes supports the idea that viruses, and especially phage, tend to be more AT-rich than their hosts (Rocha and Danchin, 2002). A single virome from the Arctic was an outlier with a GC content of 62.10%. This result directly contradicts the prevailing

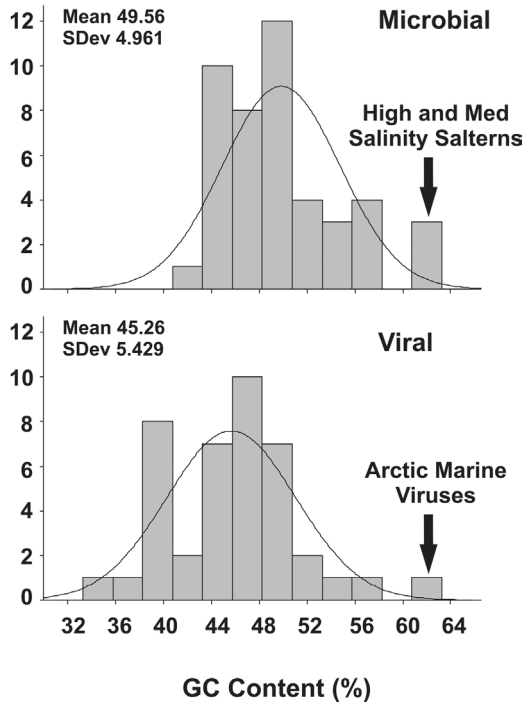

Fig. 1. Frequency histograms of per cent GC content for microbial and viral metagenomes with normal curve fitting and descriptive statistics. Extreme observations are indicated by arrows.

wisdom that genomes from colder environments would have higher AT content (Foerstner *et al*., 2005).

GC content is not a strong predictor of biomes

GC content among environments/biomes was also determined. GC content generally varied across the biomes with no overall trends. Despite some apparent clusters in Fig. 2, GC content only explained 34.9% and 13.9% of variation in microbial and viral metagenomes, respectively, across biomes (based on adjusted r^2 values from regression analysis). This corroborates previous work which demonstrated that GC content has little discriminatory power for binning of metagenomic fragments (Teeling *et al*., 2004). Although GC content has been shown to differ significantly between soil and marine metagenomes, no overall trends were observed among the 86 metagenomes in this study which encompassed a larger variety of environments (Foerstner *et al*., 2005; Raes *et al*., 2007).

Overview of dinucleotide relative abundances

Some microbial, viral and eukaryotic genomes show significant extremes in individual genomic dinucleotide abundances (Burge *et al*., 1992; Karlin *et al*., 1997; Campbell *et al*., 1999). To assess the over- and under-representation of dinucleotides in individual metagenomes, relative abundance odds ratios, ρ^*_{XY}, were calculated (see *Experimental procedures*). Standard deviations for each relative abundance odds ratio were also calculated to assess the degree of variation in dinucleotide usage between metagenomic sequences.

To compare relative abundance variations in metagenomes with those in individual genomes, 10 random subsets of genomic fragments with an average length of 100 bp were created for two microbial genomes (*Escherichia coli* K-12 and *Halobacterium salinarum* R1). For each fragment set, dinucleotide relative abundance ratios were calculated and averaged and then compared with the calculated dinucleotide usage profile for the entire genome, using the δ^* metric, explained in Karlin and colleagues (1997) and *Experimental procedures*. The same process was also conducted using random sets of 1000 sequences from a medium-salinity solar saltern microbiome and the Christmas Island marine microbiome. The distance between average dinucleotide relative abundance profiles of genomic fragments and microbial genomes was smaller than the distance between relative abundances for metagenomes and their subsets, regardless of coverage (Table S3). This indicates that as

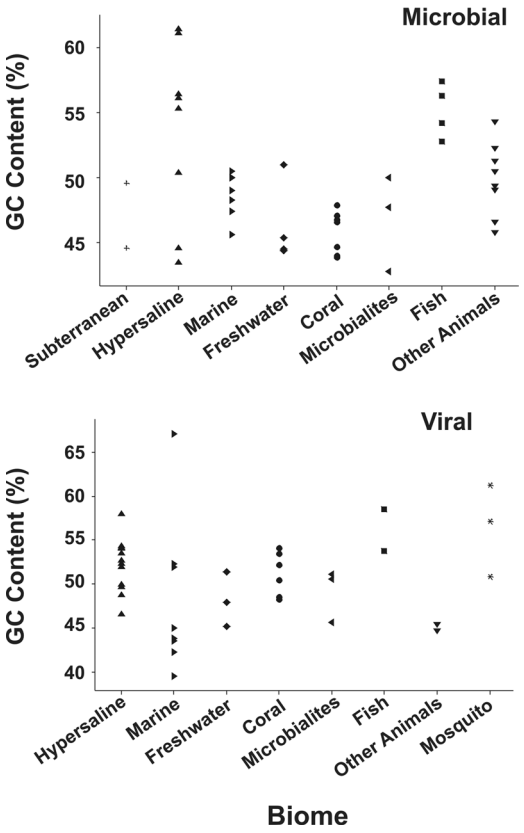

Fig. 2. Scatter plots of per cent GC content by biome for microbial and viral metagenomes.

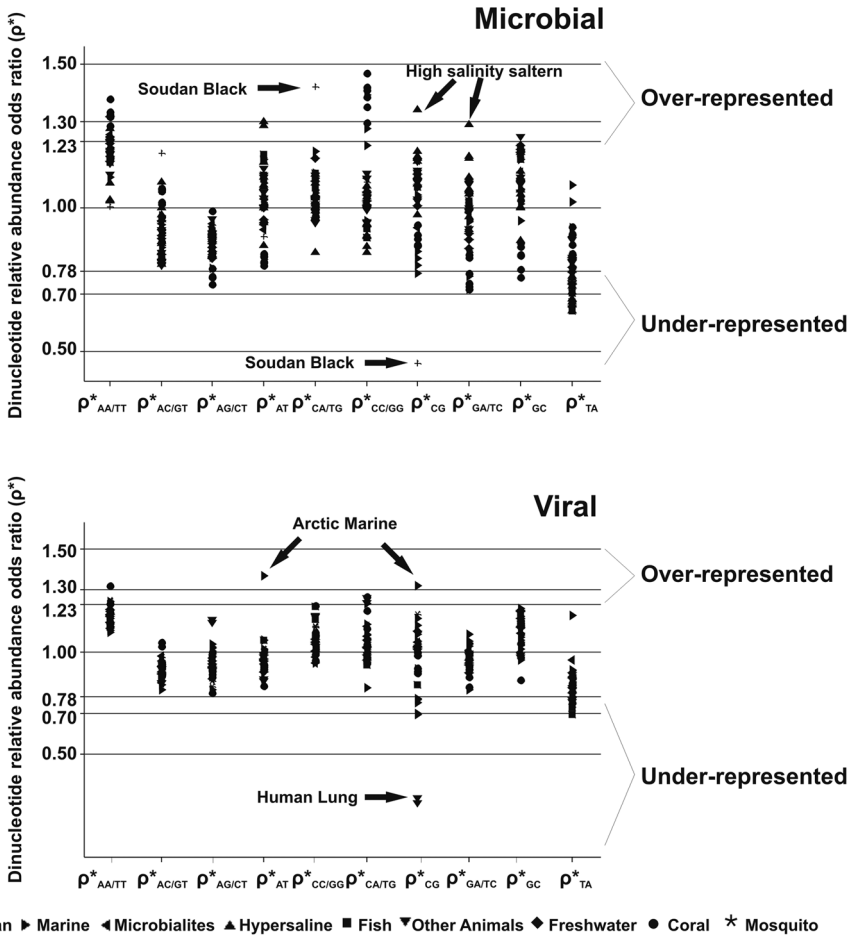

Fig. 3. Dinucleotide relative abundance odds ratios (ρ^*) for all microbial and viral metagenomes. Horizontal lines indicate cut-off points for dinucleotide abundance extremes as described by Karlin and colleagues (1997). Values outside of the normal range, $0.78 < \rho^* < 1.00$, indicate extreme dinucleotide abundances.

compared with individual genomes, metagenomes exhibit more variation in dinucleotide usage, which is expected, since metagenomes are comprised of sequence fragments from a variety of organisms.

The majority of metagenomes did not show any extremes in dinucleotide relative abundances. All microbiomes showed abundances of AC/GT dinucleotides in the normal range ($0.78 < \rho^*_{XY} < 1.00$) under-represent AG/CT, and over-represent AA/TT, although not necessarily in the significant range (Fig. 3; Table S4). Microbial metagenomes also tended to under-represent TA, except for two microbiomes from marine environments. The coral-associated microbiomes derived from *Porites compressa* displayed an overabundance of AA/TT and CC/GG dinucleotides, while the microbiome derived from a second coral species, *Porites astreoides*, did not. This may be due to differences in isolation techniques as discussed below.

As with the microbial metagenomes, the majority of viromes showed no extreme dinucleotide abundances (Fig. 3; Table S5). In general, AC/GT dinucleotides tended to be under-represented and all viromes showed TA depression, although most were outside the significant range. Three coral viromes, two marine viromes, one freshwater virome and the mosquito viromes show elevation of AA/TT, which has been shown to be common in genomes from a wide variety of organisms (Burge *et al.*, 1992).

Dinucleotide biases as a tool for evaluating human contamination in metagenomes

Anomalies in dinucleotide relative abundance odds ratios can be used to quickly identify discrepancies in metagenomes such as human genomic DNA contamination. The Soudan Black microbiome showed an elevation of CA/TG dinucleotides and a severe depression of CG dinucleotides not observed in any other microbiome. This CG depression was indicative of contaminating human DNA sequences in the metagenome. BLASTN analysis was then

conducted on this metagenome and previously unidentified human sequences were found (data not shown) (Altschul *et al*., 1990). This human genomic contamination was introduced during sequencing, as 18S PCR indicated no human genomic DNA in the original samples (data not shown).

As indicated by the dinucleotide biases in Table S5 the two animal virome samples also showed a depression of CG nucleotides in ($\rho^*_{CG} = 0.29$ and $\rho^*_{CG} = 0.27$). These samples were viral particles collected from human lung sputum. As with the Soudan Black microbiome these viral metagenomes have abundances of CG nucleotides in ranges normally exhibited by vertebrate, and especially human DNA (Gentles and Karlin, 2001). Human genomic DNA contamination was confirmed by BLASTN analysis (Altschul *et al*., 1990). While BLAST analysis can require days to complete, the calculation of dinucleotide relative abundance odds ratios can be performed in a matter of minutes. We therefore suggest that this approach can be used as a quality control metric for human DNA contamination.

Dinucleotide relative abundance distances are large between unrelated metagenomes

To simultaneously compare abundance differences between metagenomes in all 16 dinucleotides, the δ^* statistic was calculated as previously described (Karlin *et al*., 1997; van Passel *et al*., 2006). A matrix of adjusted δ^* values for all pairwise comparisons of the 86 metagenomes was generated and scored using quartiles of the observed distribution of δ^* values to define similarity ranges (Fig. 4). The quartiles of the empirical distribution corresponded closely with the empirical cut-off values defined by Karlin and colleagues based on comparisons between reference genomes (Table S6), and therefore were given similar classifications (Karlin *et al*., 1998).

When compared with all other metagenomes using δ^*, the Soudan Black microbiome appeared very different from any of the other metagenomes, including the other subterranean sample (adjusted δ^* greater than 135 in all cases; Fig. 4; upper arrows), as confirmed by BLASTN analysis. The Arctic viral metagenome was distant from the other viromes, yet it bore weak similarity to many microbiomes from a variety of environments (Fig. 4, lower arrows). Prophage often develop genomic content similar to that of their bacterial hosts (Blaisdell *et al*., 1996). Consistent with this observation, Angly and colleagues (2006) showed that this virome contains a large prophage signal.

The six *P. compressa* coral microbiomes were more similar to each other than to all other metagenomes ($\delta^* < 90$ in all cases), and were different from the *P. astreoides* metagenome. These six *P. compressa* microbiomes where taken from corals treated with different stressors in aquaria (Vega Thurber *et al*., 2008). This experiment was performed in Hawaii. In contrast, the *P. astreoides* coral was taken directly from a marine environment near Panama (Wegley *et al*., 2007). Differences in overall dinucleotide frequencies may correspond to differences in the microbial consortia associated with different coral species and/or different coastal habitats. Alternatively, differences in isolation techniques may be responsible for the dissimilarity between the coral metagenomes, as the *P. astreoides* sample was known to contain mitochondrial DNA.

Three dimensions explain the majority of variance in metagenomes

For both viromes and microbiomes, PCA was conducted to reduce the dimensionality of the set of dinucleotide abundance predictors. While δ^* was used as a summary measure to simultaneously compare all differences between dinucleotide abundances, PCA combined the abundance variables into new variables which better explained dinucleotide differences between metagenomes. The eigenvalues for the first three principal components derived from ρ^*_{XY} values (Table 1) for microbial

Table 1. Eigenvalues and per cent of variance explained for the first three principal components derived from oligonucleotide composition of microbiomes and viromes.

	Dinucleotides			Trinucleotides			Tetranucleotides		
	PC1	PC2	PC3	PC1	PC2	PC3	PC1	PC2	PC3
Microbial									
Eigenvalue	4.77	1.95	1.35	23.86	10.24	7.99	94.37	58.94	26.94
Per cent variance explained	47.7%	19.5%	13.6%	37.3%	16.0%	12.5%	36.9%	23.0%	10.5%
Cumulative per cent variance explained	47.7%	67.2%	80.8%	37.3%	53.3%	65.7%	36.9%	59.9%	70.4%
Viral									
Eigenvalue	3.65	2.14	1.61	15.87	9.56	8.67	62.22	31.13	27.01
Per cent variance explained	36.5%	21.4%	16.1%	24.8%	14.9%	13.6%	24.3%	12.9%	10.6%
Cumulative per cent variance explained	36.5%	57.9%	74.0%	24.8%	39.7%	53.3%	24.3%	37.2%	47.8%

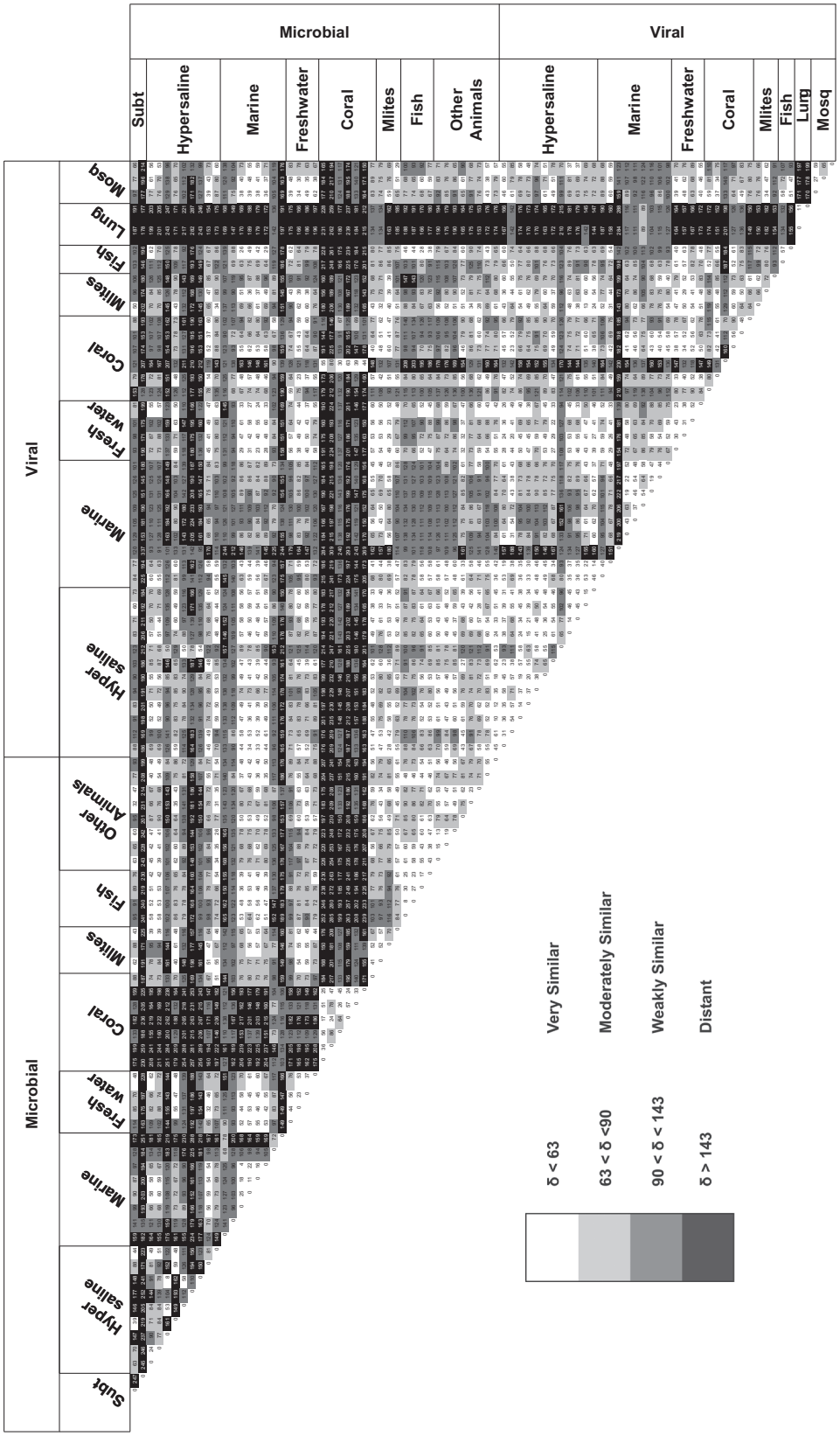

Fig. 4. Dinucleotide relative abundance distance (δ^*) values for all pairwise comparisons of metagenomes. Values are shaded according to the degree of similarity between metagenomes based on relative dinucleotide abundances. Subt indicates subterranean microbiomes, Milites indicates microbialites and mosq indicates mosquito viromes.

and viral metagenomes were all greater than one and thus were retained (Quinn and Keough, 2002). For microbiomes, the first three principal components explained 80.8% of the variance between metagenomes. For the viromes, the fourth principal component also had an eigenvalue greater than one; however, this component was excluded, since the first three components explained nearly 74% of the variance. This is comparable to what was obtained by functional genomic analyses using two principal components (~70%; Dinsdale *et al*., 2008b).

Principal component analysis was also conducted for tri- and tetranucleotide relative abundance odds ratios to determine whether differences in sequence composition of metagenomes could be better explained by higher-order oligonucleotides. Dinucleotides explained the highest proportion of variance between microbial and viral metagenomes (Table 1) as compared with other oligonucleotides. These results support the notion originally posited by Karlin that dinucleotides are sufficient to capture the majority of variations in sequence composition (Karlin *et al*., 1997).

Previous work by Sandberg and colleagues (2003) demonstrated that longer oligonucleotides are more useful for classification of genomic fragments; however, it is important to note that Sandberg and colleagues utilized raw oligonucleotide frequencies, while this analysis uses a corrected measure, the relative abundance odds ratio (Burge *et al*., 1992; Karlin *et al*., 1997). This measure controls for underlying prevalences of lower-order terms, and allows for true determination of oligonucleotide biases (Burge *et al*., 1992; Karlin *et al*., 1997). To illustrate the overestimation of discriminatory power which occurs when unadjusted measures are used, PCA was conducted using di-, tri- and tetranucleotide raw frequences (Table S7). In all cases, the per cent of variability explained was inflated when raw frequences were used. However, regardless of whether raw or adjusted frequency measures were used, dinucleotides always explained more of the variability between metagenomes than tri- and tetranucleotides. These results differ from the work of Teeling and colleagues which showed the utility of tetranucleotides for binning of genomic fragments, and also that of Pride and colleagues which demonstrated strong tetranucleotide biases across microbial and phage genomes (Pride *et al*., 2003; Teeling *et al*., 2004). While both of these studies did use corrected abundance measures to control for lower-order terms, they considered much longer sequences (from 40 kb to entire genomes) than the metagenomic fragments used in this study, which averaged 100 bp in length (Pride *et al*., 2003; Teeling *et al*., 2004). Therefore, while for genomes and longer sequence fragments, higher-order oligonucleotides may provide more discriminatory power, dinucleotides perform best for the description of naturally occurring metagenomic signatures based on short sequence fragments.

Metagenomic clustering based on oligonucleotide relative abundances

Three-dimensional scatter plots of the first three principal components derived from dinucleotide relative abundance odds ratios showed that metagenomes derived from similar biomes cluster together (Fig. 5). All four fish microbiomes (top left, yellow circles) clustered tightly, as did two metagenomes derived from mice (top left, red circles enclosed in black ellipse). The Soudan Black subterranean microbiome fell distinctly outside any distinguishable cluster, reflecting its high level of compositional discordance with other metagenomes, as also demonstrated in frequency analysis using the δ^* statistic. The sample used to generate this metagenome was taken from the reduced mine sediments in Minnesota, an extreme, anoxic environment unlike any others included in this study (Edwards *et al*., 2006) and it is highly likely that the composition of the microbial community greatly differs from those of the other microbiomes. Additionally, viromes from similar marine environments grouped together (bottom left, blue circles), while the Arctic marine viral metagenome was very distant from all other viromes, due to the high abundance of prophage sequences.

The scatter plots also revealed several interesting trends among metagenomes from related environments. Three of the four metagenomes from the Northern Line Islands (top left, blue circles enclosed in black ellipse) clustered, while the fourth, taken from Christmas island, fell separately. This fourth microbiome represents Christmas Island, which is the mostly highly inhabited of all four islands (Dinsdale *et al*., 2008a). Not only does Christmas Island have distinctly different water chemistry from the other three islands, but also the reef-associated microbes are more heterotrophic and pathogenic, as exhibited by the structural and functional 'metabolic profiles' (Dinsdale *et al*., 2008a,b). Within the coral samples (green circles), all of the *P. compressa* microbial metagenomes clustered tightly near the PC3 axis, while the *P. astreoides* metagenome fell more centrally in the plot. These data again support the hypothesis that there are inherent differences in these samples despite the fact that they were classified *a priori* as belonging to the same biome. Two of the three mosquito viromes (bottom left, black circles) clustered separately from the third. BLAST analyses demonstrated that these two metagenomes were overwhelmingly dominated by sequences from a single-stranded virus of mosquitoes, *Aedes albopictus densovirus* (data not shown). The third mosquito sample was subjected to single-stranded DNA digestion prior to sequencing and containing no sequences with BLAST similarities to this virus.

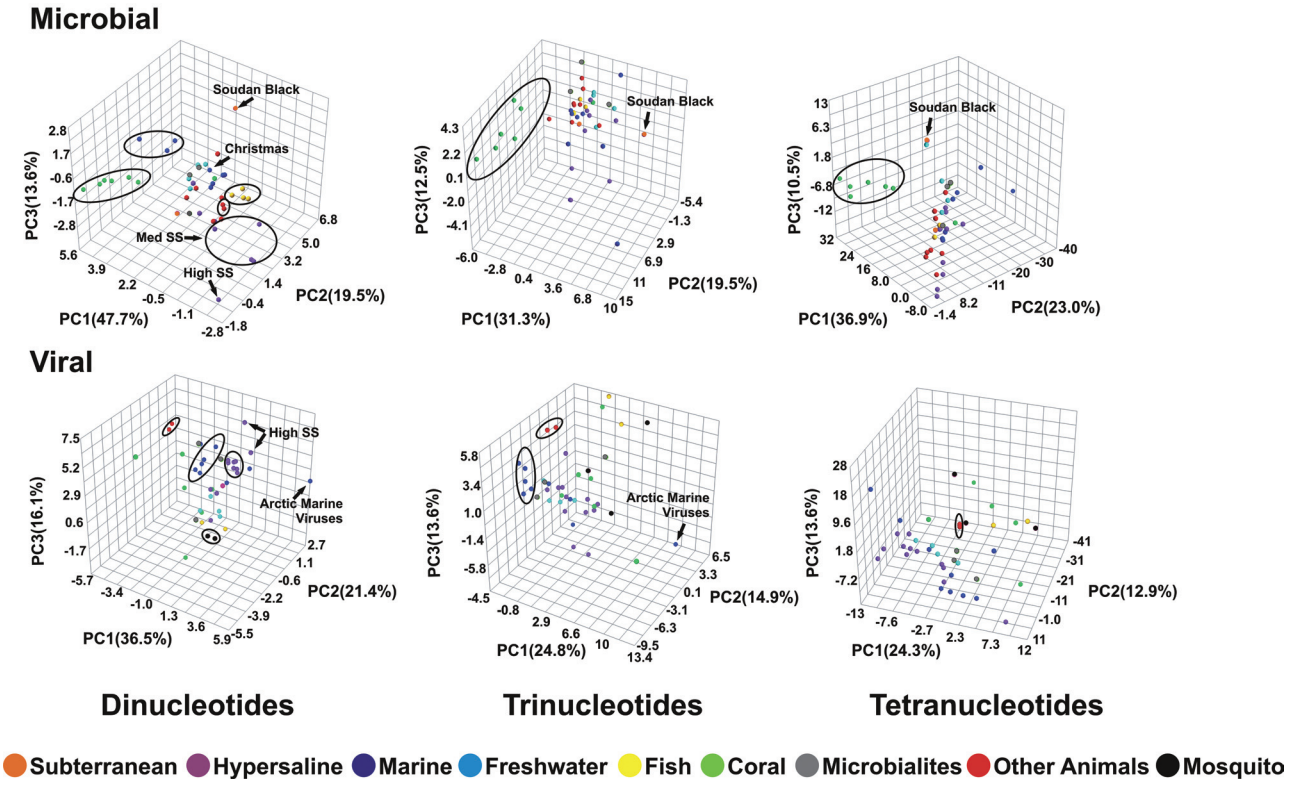

Fig. 5. Three-dimensional scatter plots of principal components from PCA of microbial and viral oligonucleotide frequencies. The per cent of variation each principal component explains is indicated in parentheses adjacent to the component axis.

Interestingly, for both microbiomes and viromes, hypersaline metagenomes (top and bottom left, purple circles) clustered according to a salinity gradient with the high-salinity saltern lying farthest, followed by the medium-salinity salterns, with the low-salinity samples lying more centrally. This corroborates results from previous studies in solar salterns, which show that diversity decreases as saltern ponds become more saline (Benlloch *et al*., 2002; Casamayor *et al*., 2002). The decrease in diversity leads to the marked dominance of extreme halophiles at high salinity, thus creating a metagenomic dinucleotide signature dominated by the inputs of relatively few species (Benlloch *et al*., 2002; Casamayor *et al*., 2002). This supports the hypothesis that dominant taxa in the environment may drive metagenomic dinucleotide signals.

To further test this dominant taxa hypothesis, we compared the genomic signatures of abundant taxa in medium- and high-salinity salterns to the metagenomic signatures. For each genome identified by BLASTN, dinucleotide relative abundance odds ratios were calculated. Weighted averages of odds ratios were calculated using subsets of the most abundant taxa as described in *Experimental procedures*. For each subset, the distance between the weighted average dinucleotide signature and the metagenomic signature was expressed using the δ^* metric of Karlin and colleagues (1997). As shown by the rank abundance curves, the high-salinity saltern metagenome had lower diversity and was less even than the medium-salinity saltern, with BLASTN hits to fewer known genomes (Fig. 6A and B, Table S8). At high salinity, over 70% of BLAST hits are attributable to only two microbial genomes (*Salinibacter ruber* and *Haloquadratum walsaybi*), while at medium salinity nearly 80 taxa must be included to account for 70% of BLAST hits. When the number of genomes used to calculate the weighted average abundance ratios versus δ^* was plotted, the curves mirrored the BLASTN-based rank–abundance curve for each metagenome (Fig. 6C and D). At high salinity, the distance between the estimated dinucleotide abundances and the true abundances changed little as more genomes were added, while the distance continuously decreased with the addition of genomes at medium salinity. This supports the hypothesis that dominant taxa are driving dinucleotide signatures, since each taxon seems to contribute to the metagenomic dinucleotide signature according to its relative abundance. It should be noted that even when all genomes identified by BLASTN hits are

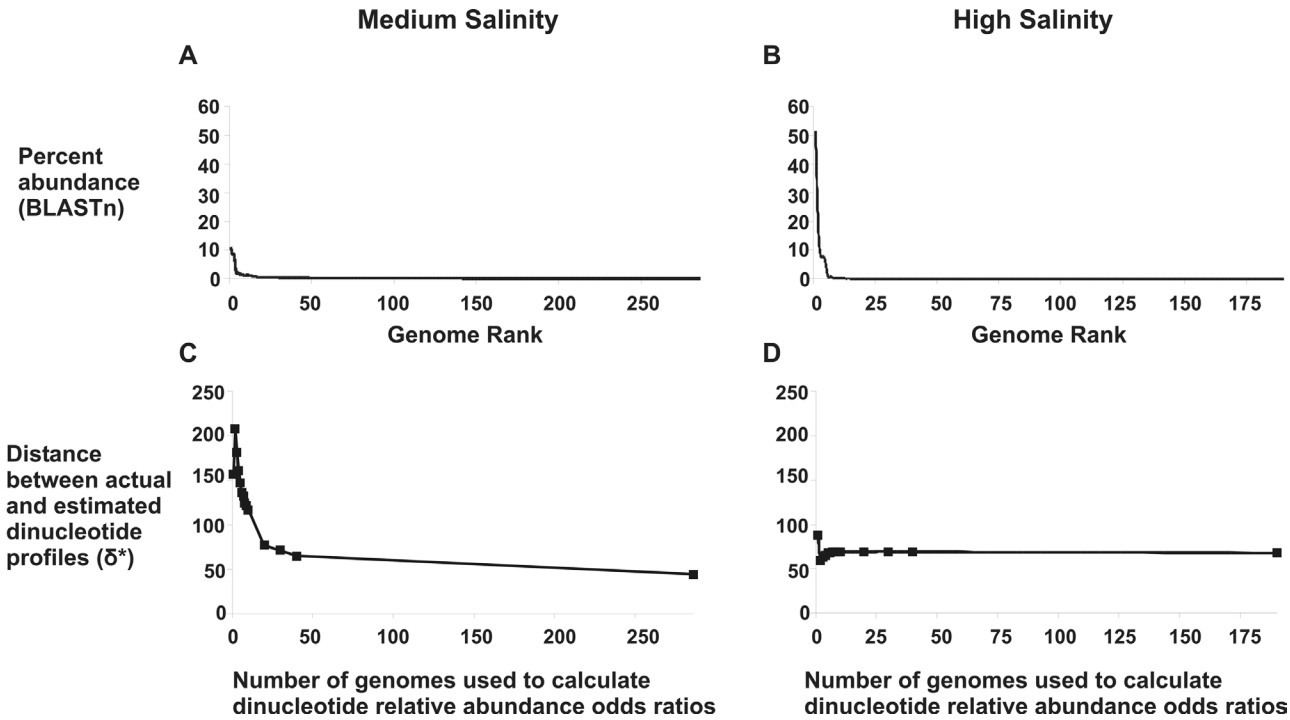

Fig. 6. Rank abundance curves (A and B) and dinucleotide relative abundance distances (C and D) for medium- and high-salinity solar saltern microbiomes.

considered, there is still considerable distance between dinucleotide relative abundance estimates and the true metagenomic signature ($\delta^* = 44$ for medium salinity, and $\delta^* = 68$ for low salinity). This is attributable to the large percentage of sequences in each metagenome with no similarity to known microbial genomes (88% at medium salinity and 68% at high salinity), which may have large contributions to the metagenomic signatures.

Scatter plots created using the first three principal components from PCA with tri- and tetranucleotide relative abundance odds ratios (Fig. 5, centre and right) did not provide any additional clustering of metagenomes. In fact, clusters and overall trends appeared to decline and/or disappear when longer oligonucleotides were used. Dinucleotide relative abundances exhibited substantial discriminatory power to cluster metagenomes by environment and also to provide biologically relevant information about similarities and differences between metagenomic libraries.

Dinucleotide clustering is robust to the addition of metagenomes containing longer sequences

The set of 86 microbial and viral metagenomes used in this study were selected because all samples were processed and sequenced analogously, producing approximately 100 base pair pyrosequencing reads. Since the calculated dinucleotide relative abundance ratios reflect average abundances over the entire metagenome, it might be assumed that the introduction of longer sequence reads would have little effect, although as previously stated, longer sequences may have stronger signatures using higher-order oligonucleotides. Other approaches to metagenomics, such as the use of fosmids and BACs, may be subject to differences in cloning efficiencies which could introduce bias. Additionally, the inclusion of further metagenomes could increase the discriminatory power of dinucleotides if more distinct environments were represented, but if additional data were to come from similar environments, it could also potentially decrease resolution. To determine the effects of the addition of metagenomic libraries on clustering behaviour, 11 microbial and 3 viral metagenomes were added to the analysis. GC content, dinucleotide relative abundance odds ratios and other characteristics of these additional metagenomes are provided in Table S9.

When the 11 microbial metagenomes were added, PCA produced comparable results to the analysis which included only pyrosequenced microbiomes. Previously, dinucleotide relative abundances accounted for 80.8%, with the addition of the other microbiomes; 76.4% of the variation was explained. Despite this reduction, metagenomes still exhibited nearly the same clustering behaviour as in the previous analysis, with the Soudan Black metagenome failing to associate with other microbiomes (Fig. S1, left, orange circle), and six of seven coral

metagenomes (green circles) clustering together. The clustering of hypersaline metagenomes (purple circles) once again occurred along a salinity gradient, and the newly added metagenome from a Spanish saltern (purple square) clustered with pre-existing high-salinity solar saltern samples (Legault *et al*., 2006). Two of the whale fall metagenomes (yellow circles) which were derived from whale bones clustered together, while the third sample, which was taken from a microbial mat, did not (Tringe *et al*., 2005).

For viromes, the total per cent of variance explained by the first three principal components was 73.8%, a decrease from 78.4% in the previous analysis. Again, however, similar clustering behaviour was observed, with the Arctic viral metagenome and the two known contaminated lung samples (Fig. S1, right, red circles) appearing anomalous. The two hot springs metagenomes (green squares), which were derived from an environment unlike any other in the data set, lay distinctly outside the rest of the points on the scatter plot, displaying a close association with each other but a dramatic difference with the rest of the viromes (Schoenfeld *et al*., 2008).

Hierarchical clustering by dinucleotide relative abundance odds ratios

Hierarchical clustering was used to quantify the grouping behaviour and trends of the 45 microbial and 41 viral metagenomes demonstrated visually by the three-dimensional scatter plots. Using dinucleotides, both the microbiomes (Fig. 7) and viromes (Fig. S2) formed many clusters containing metagenomes exclusively from identical or similar biomes. Consistent with the PCA results, the fish microbiomes clustered together, as did the mice microbiomes, similar marine viromes and the two contaminated human lung viromes. Additional associations which were not apparent in the scatter plots were clearly delineated in the dendrograms, such as the grouping of chicken and cow rumen microbiomes. Trends in metagenomic clustering, such as the appearance of a salinity gradient, were also consistent with the scatter plot results. Hierarchical clustering of viromes did not segregate metagenomes by environment as well as clustering for microbiomes. This reflects the higher total percentage of variance explained by dinucleotide relative abundances

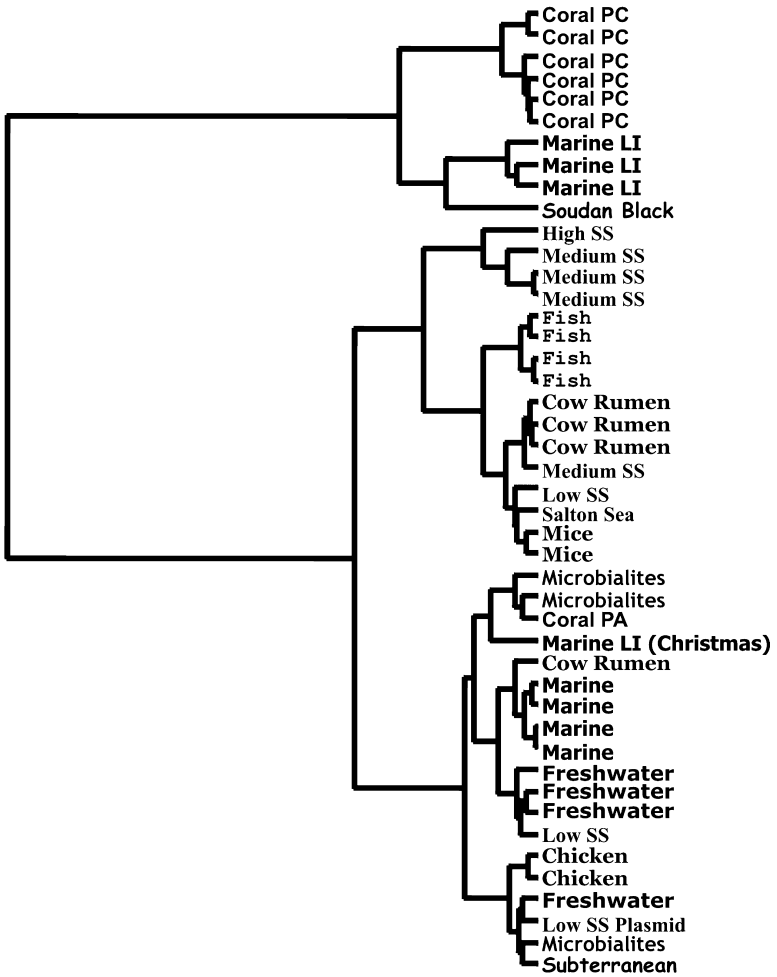

Fig. 7. Hierarchical clustering of microbial metagenomes by the first three principal components from dinucleotide relative abundances. Metagenomes are labelled according to biome. Coral PC indicates *P. compressa* coral microbiomes, Marine LI indicates microbiomes from the four Line Islands: Kiribati, Taburean, Palmyra and Christmas, and SS indicates solar salterns.

for microbiomes (80.8%) versus viromes (78.4%). Additionally, there may be unknown similarities between environments which are driving clustering behaviour, creating associations between biomes which have been classified *a priori* as different, and challenging traditional notions of what constitutes a biome.

Caveats

The majority of the metagenomic DNAs were amplified using multiple displacement amplification with Phi29 polymerase prior to sequencing, which could artificially inflate the occurrence of sequences from small circular as well as large linear genomes, and potentially exclude small linear viral genomes, thus biasing dinucleotide frequencies (Pinard *et al*., 2006; Spits *et al*., 2006). However, multiple displacement amplification generally provides an even representation of genomes except at the ends, and bias created by amplification would be away from dinucleotide extremes (Dean *et al*., 2002). Additionally, all of the pyrosequenced metagenomes used in this study were collected and processed in an identical manner, thus equally exposing them to any potential biases due to sampling or amplification. Additionally, it should be noted that the sequences used here were generated from pyrosequencing using the GS20 platform, which has been reported to have an error rate as high as 4% (Huse *et al*., 2007). However, in practice this error rate has been determined to be much lower, on the order of 0.25% (Huse *et al*., 2007).

Conclusions

Previous work has demonstrated the presence of distinctive oligonucleotide signatures in a variety of prokaryotic, eukaryotic and viral genomes, as well as marked differences in dinucleotide abundances between the genomes of distantly related organisms (Burge *et al*., 1992; Karlin and Ladunga, 1994; Blaisdell *et al*., 1996; Karlin *et al*., 1997; Gentles and Karlin, 2001; Teeling *et al*., 2004). Metagenomes represent a diverse cross-section of a particular environmental community, and therefore are not composed of DNA from a single type of organism, but a mixture of DNA from a variety of organisms (Tringe and Rubin, 2005). Initially, we hypothesized that an individual metagenome would not have a characteristic signature, since it essentially represents an averaging of genomes from multiple species. Instead, dinucleotide compositional analysis showed that despite this high level of diversity, metagenomes do have distinct sequence-based signatures. These dinucleotide signatures are driven by environmental selection, in that environments may be dominated by a group of highly abundant taxa whose sequence composition accounts for trends in dinucleotide abundances. Alternatively, the environment itself might be selecting for particular patterns of dinucleotides, irrespective of taxonomy. With the current data set, the metagenomic dinucleotide profiling performs better than profiling using higher-order oligonucleotides, and explains approximately the same proportion of variance between metagenomes as functional genomic analyses. This approach also challenges preconceived notions regarding what constitutes a biome, indicating that the data may carry information that undermines *a priori* biome classifications. The predictive power of this approach suggests that it may be possible to identify anomalous sequences in a manner analogous to that used for individual genomes. The dinucleotide composition was also useful for determining subtle signals in sequence data, such as the presence of contamination, and should be used as a rapid quality check for metagenomes. Together these results show that using dinucleotide abundances allows for more complete characterization of metagenomic content and for rapid comparisons between metagenomes. These analyses could also be used in combination with functional analyses such as those presented in Dinsdale and colleagues (2008b). Since functional annotations rely on similarities to known sequences while compositional analyses do not, environmental clustering of metagenomes by function could be corroborated using dinucleotide signatures, thus providing greater power to discriminate between environments.

Experimental procedures

Data sets

The primary data used for this study consist of pyrosequencing (Roche/454 Life Sciences) reads for a total of 86 metagenomes (45 microbial and 41 viral) derived from nine different biomes, classified as in Dinsdale and colleagues (2008b). Table 2 shows the biome classifications along with how many metagenomes from each biome were used in the analysis. The metagenomic sequences are freely available from both the SEED platform and NCBI, and accession numbers as well as descriptions of the metagenomes are provided in

Table 2. The 86 microbial and viral metagenomes used in the study classified by biome.

Biome	Microbial metagenomes	Viral metagenomes
Subterranean	2	–
Hypersaline	9	12
Marine	8	9
Freshwater	4	4
Coral	7	6
Microbialites	3	3
Fish	4	2
Other animals	8	2
Mosquito	–	3
Total	45	41

Tables S10A and S10B. Data sets are classified as microbial or viral depending on whether sample DNA was extracted for sequencing from a whole microbial fraction or viral fraction derived from caesium chloride density gradient ultracentrifugation as previously described (Angly *et al*., 2006; Wegley *et al*., 2007; Desnues *et al*., 2008; Dinsdale *et al*., 2008b). Metagenomic sequences have an average length of 100.2 bp, and metagenomic sizes range from 4645 sequences to 688 590 sequences. Additional metagenomic data sets used in PCA were obtained from CAMERA (http://camera.calit2.net).

Frequency tabulation

GC content and mono- and di-, tri-, tetranucleotide counts over each entire metagenome were calculated using a self-written Perl script. This script and all other programs used in this analysis are available at http://sourceforge.net/projects/dinucleotidesig. Counts of N nucleotides (representing a poor sequencing read) as well as oligonucleotides containing N nucleotides were tabulated and in all cases comprised less than 1% of the total frequency data. All N mono- and di- and higher-order oligonucleotides were removed from the data sets prior to further data processing.

The odds ratio measure of dinucleotide bias ρ^*_{XY} was used to evaluate dinucleotides for over- and under-representation in the metagenomes (Burge *et al*., 1992; Karlin *et al*., 1997). This measure accounts both for the underlying frequencies of individual mononucleotides and for the complementary nature of double-stranded DNA (Burge *et al*., 1992; Karlin *et al*., 1997). Raw frequency values were corrected by averaging the frequency of each mono- or dinucleotide with the frequency of its reverse complement, and assigning the average frequency, *f**, to both (Burge *et al*., 1992). Following frequency correction, ρ^*_{XY} was calculated for all possible dinucleotides over all metagenomes as $\rho^*_{XY} = \frac{f^*_{XY}}{f^*_X f^*_Y}$ which corrects the observed dinucleotide frequency for the lower-order mononucleotide frequency terms (Karlin *et al*., 1997; 1998). As shown in *Results*, ρ^*_{XY} values were then classified as normal or extreme according to the given criteria (Karlin *et al*., 1997; 1998). Standard deviations for dinucleotide relative abundance odds ratios were calculated by determining ρ^*_{XY} values for each individual sequence and then calculating the adjusted average difference between each sequence and ρ^*_{XY} for the whole metagenomes. Relative abundance odds ratios were calculated for trinucleotides using the third-order measure $\gamma^*_{XYZ} = \frac{f^*_{XYZ} f^*_X f^*_Y f^*_Z}{f^*_{XY} f^*_{YZ} f^*_{XNZ}}$, and for tetranucleotides using the fourth-order metric $\tau^*_{XYZW} = \frac{f^*_{XYZW} f^*_{XY} f^*_{XNZ} f^*_{XNMW} f^*_{YZ} f^*_{YNW} f^*_{ZW}}{f^*_{XYZ} f^*_{XYNW} f^*_{YZW} f^*_X f^*_Y f^*_Z f^*_W}$, where *N* and *M* represent any nucleotide (Karlin *et al*., 1997).

Calculation of relative abundance differences

Calculation of ρ^*_{XY} allows for comparison of dinucleotide frequencies across an individual metagenome, and identifies potential dinucleotide bias. To compare the overall frequencies between metagenomes, the average dinucleotide relative abundance difference, δ^*, was calculated for all pairwise combinations of the 86 metagenomes (Karlin *et al*., 1997) using a self-written Java program. The value of δ^* was calculated as $\delta^*(f, g) = \frac{1}{16}\sum_{XY} |\rho^*_{XY}(f) - \rho^*_{XY}(g)|$ where *f* and *g* represent two different metagenomes (Karlin *et al*., 1997). Descriptive statistics and a graphical summary of the distribution of δ^* values were generated using Minitab Version 15 software (Minitab State College, PA, USA), and values were classified by quartile. The average relative abundance differences reported in *Results* are multiplied by 1000 for easier comparison.

Comparison of dinucleotide relative abundance variation in genomes versus metagenomes

Ten randomly selected sets of 1000 sequences each were selected from a medium-salinity saltern microbiome and the Christmas Island microbiome using a self-written Perl script, to give 0.125× and 0.005× coverage of the metagenomes respectively. The genomes of *E. coli* K-12 substrain MG1655 (NC_000913) and *H. salinarum* R1 (NC_010364) were downloaded from NCBI (http://www.ncbi.nlm.nih.gov), and dinucleotide relative abundance odds ratios were calculated as described above. Random sets of genomic fragments were generated using a self-written Perl script which allows the user to specify the desired genomic coverage as well as an average fragment length. Ten set of fragments with an average length of 100 bp were generated for both 0.125× and 0.005× coverage. For each set of genomic and metagenomic sequences, dinucleotide relative abundance odds ratios were calculated as described above. Relative abundance odds ratios were averaged over 10 repetitions and compared with the dinucleotide profiles for the original genome or metagenome using the δ^* metric of Karlin and colleagues (1997) described above.

BLAST analysis of metagenomes and rank–abundance calculations

Metagenomes were compared with the database of all microbial genomes available from NCBI (http://www.ncbi.nlm.nih.gov) using BLASTN with an *e*-value cut-off of 10^{-5} (Altschul *et al*., 1990). Complete genomes for each organism detected by BLAST were downloaded from NCBI and dinucleotide relative abundance odds ratios were calculated as described above. To determine the relative contribution of each genome to the metagenomic signature, weighted averages of dinucleotide relative abundances were calculated by multiplying each dinucleotide relative abundance odds ratio for a genome by the percentage of total BLAST hits to that genome in the subset of genomes considered, summing the weighted odds ratios, and then dividing by the number of genomes in the subset.

Statistical analysis

All statistical analysis was performed using Minitab Version 15 software (Minitab, State College, PA, USA). Simple descriptive statistics as well as histograms and box plots of

GC content for microbial and viral genomes were created using the Minitab Descriptive Statistics option. GC content was also explored by biome using a scatter plot generated using the 2D Plot routine. Trends in di-, tri- and tetranucleotide frequencies were examined in Minitab using PCA as well as hierarchical clustering. For all analyses, relative abundance odds ratios were used instead of raw frequency values, because the raw values were too highly correlated to provide meaningful results. Principal component analysis takes a set of correlated variables and reduces them to a smaller set of uncorrelated variables, which can then be used for further analysis (Quinn and Keough, 2002). Principal component analysis was performed in this study separately on the microbial and viral metagenomes, using the correlation association matrix to compensate for unequal variances of predictor variables (Quinn and Keough, 2002). Predictor sets were reduced to three principal components based on eigenvalues in both cases, and these principal components were used to generate three-dimensional scatter plots with the software package Graphis (KyleBank Software, Ayr, UK). Eigenvalues greater than one indicate that a principal component explains more of the variance than would be expected by chance. Since PCA conducted using a correlation matrix standardizes each original variable to have a mean of zero and a variance of 1, the total variance equals the total number of original predictors. Thus, the larger the eigenvalue, the larger proportion of the variance a principal component is explaining (Quinn and Keough, 2002).

Hierarchical clustering using Euclidean distances with Ward linkage was performed on standardized data generated from both the raw values of ρ^*_{XY}, γ^*_{XYZ} and τ^*_{XYZW} and the principal component values generated from the PCA. Both methods assigned the same number of optimal clusters of identical composition.

Acknowledgements

We thank Steve Rayhawk for helpful discussion, and Katie Barott, Elizabeth Dinsdale, Matt Haynes and Linda Wegley for critical readings of the manuscript. This work was supported by the Marine Microbiology Initiative at the Gordon and Betty Moore Foundation (to F.R.) and the Canadian Institute for Advanced Research (CIFAR) Integrated Microbial Biodiversity Program (to F.R.).

References

Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S., and Ikemura, T. (2005) Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res* **12:** 281–290.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4:** e368.

Benlloch, S., Lopez-Lopez, A., Casamayor, E.O., Ovreas, L., Goddard, V., Daae, F.L., *et al.* (2002) Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern. *Environ Microbiol* **4:** 349–360.

Blaisdell, B.E., Campbell, A.M., and Karlin, S. (1996) Similarities and dissimilarities of phage genomes. *Proc Natl Acad Sci USA* **93:** 5854–5859.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., *et al.* (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* **99:** 14250–14255.

Burge, C., Campbell, A.M., and Karlin, S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA* **89:** 1358–1362.

Campbell, A., Mrazek, J., and Karlin, S. (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci USA* **96:** 9184–9189.

Casamayor, E.O., Massana, R., Benlloch, S., Ovreas, L., Diez, B., Goddard, V.J., *et al.* (2002) Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. *Environ Microbiol* **4:** 338–348.

Chapus, C., Dufraigne, C., Edwards, S., Giron, A., Fertil, B., and Deschavanne, P. (2005) Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol Biol* **5:** 63.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., *et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA* **99:** 5261–5266.

Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., and Fertil, B. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* **16:** 1391–1399.

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., *et al.* (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452:** 340–343.

Dinsdale, E.A., Pantos, O., Smriga, S., Edwards, R.A., Angly, F., Wegley, L., *et al.* (2008a) Microbial ecology of four coral atolls in the northern line islands. *PLoS ONE* **3:** e1584.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., *et al.* (2008b) Functional metagenomic profiling of nine biomes. *Nature* **452:** 629–632.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., and Deschavanne, P. (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res* **33:** e6.

Edwards, R.A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D.M., *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7:** 57.

Fertil, B., Massin, M., Lespinats, S., Devic, C., Dumee, P., and Giron, A. (2005) GENSTYLE: exploration and analysis of DNA sequences with genomic signature. *Nucleic Acids Res* **33:** W512–W515.

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

Gentles, A.J., and Karlin, S. (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res* **11:** 540–546.

Goodarzi, H., Torabi, N., Najafabadi, H.S., and Archetti, M. (2007) Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons. *Gene* **407:** 30–41.

Huse, S., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8:** R:143.

Karlin, S., and Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11:** 283–290.

Karlin, S., and Ladunga, I. (1994) Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci USA* **91:** 12832–12836.

Karlin, S., and Mrazek, J. (1997) Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci USA* **94:** 10227–10232.

Karlin, S., Mrazek, J., and Campbell, A.M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179:** 3899–3913.

Karlin, S., Campbell, A.M., and Mrazek, J. (1998) Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32:** 185–225.

Legault, B., Lopez-Lopez, A., Alba-Casado, J.C., Doolittle, W.F., Bolhuis, H., Rodriguez-Valera, F., and Papke, R.T. (2006) Environmental genomics of '*Haloquadratum walsbyi*' in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics* **7:** 171.

Martin-Cuadrado, A.B., Lopez-Garcia, P., Alba, J.C., Moreira, D., Monticelli, L., Strittmatter, A., *et al.* (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE* **2:** e914.

van Passel, M.W., Bart, A., Luyf, A.C., van Kampen, A.H., and van der Ende, A. (2006) Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics* **7:** 26.

Paul, S., Bag, S.K., Das, S., Harvill, E.T., and Dutta, C. (2008) Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol* **9:** R70.

Pinard, R., de Winter, A., Sarkis, G.J., Gerstein, M.B., Tartaro, K.R., Plant, R.N., *et al.* (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7:** 216.

Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., and Blaser, M.J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13:** 145–158.

Quinn, G.P., and Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge, UK: Cambridge University Press.

Raes, J., Foerstner, K.U., and Bork, P. (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* **10:** 490–498.

Rocha, E.P., and Danchin, A. (2002) Base composition bias might result from competition for metabolic resources. *Trends Genet* **18:** 291–294.

Sandberg, R., Bränden, C.I., Ernberg, I., and Cöster, J. (2003) Quantifying the species-specificity in genomic signatures, synonymous codon choice, animo acid usage and G+C content. *Gene* **11:** 35–42.

Schloss, P.D., and Handelsman, J. (2003) Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* **14:** 303–310.

Schoenfeld, T., Patterson, M., Richardson, P.M., Wommack, K.E., Young, M., and Mead, D. (2008) Assembly of viral metagenomes from Yellowstone Hot Springs. *Appl Environ Microbiol* **74:** 4164–4174.

Singer, G.A.C., and Hickey, D.A. (2003) Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317:** 39–47.

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., and Sermon, K. (2006) Whole-genome multiple displacement amplification from single cells. *Nat Protoc* **1:** 1965–1970.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., and Glockner, F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6:** 938–947.

Tringe, S.G., and Rubin, E.M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* **6:** 805–814.

Tringe, S., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., *et al.* (2005) Comparative metagenomics of microbial communities. *Science* **308:** 554–557.

Vega Thurber, R., Barott, K.L., Hall, D., Liu, H., Rodriguez-Mueller, B., Desnues, C., *et al.* (2008) Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. *PNAS* **105:** 18413–18418.

Wang, Y., Hill, K., Singh, S., and Kari, L. (2005) The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene* **346:** 173–185.

Wegley, L., Edwards, R., Rodriguez-Brito, B., Liu, H., and Rohwer, F. (2007) Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*. *Environ Microbiol* **9:** 2707–2719.

Woyke, T., Teeling, H., Ivanova, N.N., Huntemann, M., Richter, M., Gloeckner, F.O., *et al.* (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443:** 950–955.

Supporting information

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

Fig. S1. Three-dimensional scatter plot of principal components from PCA of microbial and viral dinucleotide frequencies for original and additional metagenomes. The per cent of variation each principal component explains is indicated in parentheses.

Fig. S2. Hierarchical clustering of viromes by the first three principal components from dinucleotide relative abundances. Metagenomes are labelled according to biome. Coral PC indicates *P. compressa* coral viromes, Marine LI indicates viromes from the Line Islands and SS indicates solar salterns.

Table S1. GC content with standard deviations for all microbiomes.

Table S2. GC content with standard deviations for all viromes.
Table S3. Results of simulation to evaluate the degree of variation in dinucleotide relative abundance profiles of genomes versus metagenomes. Metagenomes included are a medium-salinity solar saltern microbiome with SEED ID 4440416.4 and the Christmas Island microbiome with SEED ID 4440041.3.
Table S4. Over- and under-represented dinucleotides in microbiomes. Values outside the normal range are shaded and standard deviations are in parentheses. Values less than 0.78 correspond to under-representations, ρ^*_{XY} greater than 1.23 indicates over-representation. Light grey indicates ρ^*_{XY} less than 0.50, medium grey between 0.50 and 0.70, dark grey between 0.70 and 0.78, charcoal grey between 1.23 and 1.30, and black between 1.30 and 1.50.
Table S5. Over- and under-represented dinucleotides in viral metagenomes. Values of ρ^*_{XY} outside the normal range are shaded and standard deviations are given in parentheses. Values less than 0.78 correspond to under-representations while ρ^*_{XY} greater than 1.23 corresponds to over-representation. Light grey shading indicates ρ^*_{XY} less than 0.50, medium grey between 0.50 and 0.70, dark grey between 0.70 and 0.78, charcoal grey between 1.23 and 1.30, and black between 1.30 and 1.50.
Table S6. Ranges and classifications for the four quartiles of δ^* values (δ^* given as multiplied by 1000) compared with classifications given by Karlin and colleagues (1998).
Table S7. Eigenvalues and per cent of variance explained for the first three principal components derived from raw oligonucleotide frequencies of microbiomes and viromes.
Table S8. Results of BLASTN analysis (*e*-value < 0.00001) comparing a medium-salinity solar saltern microbiome (SEED ID: 444.0416.4) and a high-salinity solar saltern microbiome (SEED ID: 4440419.3) to all microbial genomes in NCBI.
Table S9. Characteristics of additional metagenomes containing long sequence reads used in PCA. All metagenomic sequences were obtained from CAMERA (http://camera.calit2.net). M indicates microbial metagenomes while V indicates viral metagenomes.
Table S10A. Description of microbial metagenomes including SEED and NCBI accession numbers and references.
Table S10B. Description of viral metagenomes including SEED and NCBI accession numbers and references.

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

