The ISME Journal (2014) 8, 1391–1402

www.nature.com/ismej

ORIGINAL ARTICLE

Classification and quantification of bacteriophage taxa in human gut metagenomes

This article has been corrected since Advance Online Publication and a corrigendum is also printed in this issue

Alison S Waller[1], Takuji Yamada[2], David M Kristensen[3], Jens Roat Kultima[1], Shinichi Sunagawa[1], Eugene V Koonin[3] and Peer Bork[1]

[1]*Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany;* [2]*Department of Biological Information, Tokyo Institute of Technology, Graduate School of Bioscience and Biotechnology, Yokohama, Japan and* [3]*National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD, USA*

Bacteriophages have key roles in microbial communities, to a large extent shaping the taxonomic and functional composition of the microbiome, but data on the connections between phage diversity and the composition of communities are scarce. Using taxon-specific marker genes, we identified and monitored 20 viral taxa in 252 human gut metagenomic samples, mostly at the level of genera. On average, five phage taxa were identified in each sample, with up to three of these being highly abundant. The abundances of most phage taxa vary by up to four orders of magnitude between the samples, and several taxa that are highly abundant in some samples are absent in others. Significant correlations exist between the abundances of some phage taxa and human host metadata: for example, '*Group 936 lactococcal phages*' are more prevalent and abundant in Danish samples than in samples from Spain or the United States of America. Quantification of phages that exist as integrated prophages revealed that the abundance profiles of prophages are highly individual-specific and remain unique to an individual over a 1-year time period, and prediction of prophage lysis across the samples identified hundreds of prophages that are apparently active in the gut and vary across the samples, in terms of presence and lytic state. Finally, a prophage–host network of the human gut was established and includes numerous novel host–phage associations.
The ISME Journal (2014) **8,** 1391–1402; doi:10.1038/ismej.2014.30; published online 13 March 2014
Subject Category: Microbial population and community ecology
Keywords: human gut; metagenomics; phage

Introduction

The crucial, multifaceted involvement of the gut microbiome in human health and diseases is being increasingly recognized through studies that reveal links between the intestinal prokaryotic communities and many conditions such as type 2 diabetes, obesity, Chrohn's disease, colitis, psoriasis, asthma, cardiovascular disease, colorectal cancer and HIV progression (Qin *et al.,* 2012; Cho and Blaser, 2012; De Vos and de Vos, 2012; Vujkovic-Cvijin *et al.,* 2013). Although these studies primarily focus on microbial communities, it is well known that viruses (primarily, bacteriophages or phages for brevity) are key components of any microbiome including that of the human gut (Breitbart *et al.,* 2003; Minot *et al.,* 2011; Barr *et al.,* 2013; Modi *et al.,* 2013). Bacteriophages have a major impact on the function and structure of bacterial communities, through horizontal gene transfer (Canchaya *et al.,* 2003; Kristensen *et al.,* 2010), impacting community composition (Duerkop *et al.,* 2012) and altering phenotypes such as virulence (Brüssow *et al.,* 2004; Busby *et al.,* 2013) or biofilm formation (Wang *et al.,* 2009; Carrolo *et al.,* 2010). Although these and other studies have demonstrated the effect of certain phages on the gut microbiome, a broader characterization of the phages and prophages is essential to fully characterize virus–host interactions in the human gut and metagenomics should provide sufficient data for this purpose.

Several metagenomic studies have been performed by enriching gut microbiome samples for viruses. In addition to early work that was performed on a limited scale (Breitbart *et al.,* 2003; Zhang *et al.,* 2006), three large-scale fecal metagenomic viral sequencing projects have been reported (Reyes *et al.,* 2010; Minot *et al.,* 2011; Kim *et al.,* 2011). Comparison of alpha-diversity profiles over time indicated that individual viromes

Correspondence: P Bork, Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, Heidelberg 69117, Germany.
E-mail: bork@embl.de
Received 14 August 2013; revised 17 January 2014; accepted 24 January 2014; published online 13 March 2014

were dominated by a few highly abundant temperate phages and that the virome composition remained stable over a year (Reyes *et al.,* 2010), although it could be altered through radical changes in the feeding regime (Minot *et al.,* 2011). However, these phage sequences have not been taxonomically classified beyond the family level (Reyes *et al.,* 2010; Kim *et al.,* 2011; Minot *et al.,* 2011) and due to the enrichment of the viruses, associations of phage taxa to bacterial taxa are difficult to establish.

An alternative to sequencing virus-enriched microbiomes (viromes) involves the identification of viral sequences in metagenomes whose sampling procedure was optimized for microbial communities. This has been performed on the Global Ocean Sampling (GOS) data set using a combination of BLAST and fragment recruitment to identify viral genes, which were classified at the family level (Williamson *et al.,* 2008) and viral scaffolds (Sharon *et al.,* 2011). An alternative approach was recently employed to investigate human gut metagenomes, using CRISPR spacer sequences to identify their cognate phage sequences. This analysis revealed a large common pool of phages that apparently are targeted by the CRISPR systems; these phages were also classified at the family level (Stern *et al.,* 2012).

The advances of viral ecology are hampered by the technical difficulties involved in experimental study of viruses as well as the paucity of bioinformatic tools to classify and quantify viral sequences in environmental samples (Duhaime and Sullivan, 2012; Solonenko *et al.,* 2013). Only a few software tools have been developed to assess the phylogenetic diversity within viral populations. Two independent techniques have been proposed to determine the overall alpha- and beta-diversity within viral communities (Angly *et al.,* 2005; Allen *et al.,* 2013). Two webservers have been developed for taxonomic classification of viruses: VIROME assigns viral open reading frames at the level of kingdom (that is, virus, bacteria, archaea, etc.; Wommack *et al.*, 2012), and metaVir assigns genes to viral families based on a list of 12 marker genes for broad viral families, with 4 markers being specific at the family level and 1 specific for a viral genus (Roux *et al.,* 2011). However, to our knowledge, no resource currently exists to systematically taxonomically classify and accurately quantify specific viral taxa. We recently identified a set of marker genes that are specific to certain phage taxa and can be used to quantify the abundance of each respective taxon (Kristensen *et al.,* 2013). Furthermore, these marker genes were chosen such that they are absent from 'non-prophage' regions of bacterial chromosomes, making them suitable to detect and quantify phage sequences in mixed metagenomic samples containing DNA from prokaryotes and phages.

Here, we use these marker genes to taxonomically classify and quantify phage taxa contained within 252 published metagenomes derived from fecal samples of 207 individuals. In addition to analyzing the phage taxa that are represented in the entire pool of metagenomic sequences, we further investigated the subset of phages that were identified as prophages integrated into bacterial chromosomes. We identified prophage regions within the metagenomic samples, and quantitatively predict patterns of prophage abundance and lysis. The derived taxonomic classification of the prophages was employed to infer an extensive network of prophage–host interactions within the gut microbiome.

Materials and methods

Marker genes for phage taxa

Phage Orthologous Groups (POGs) were constructed using the proteins contained in over 1000 phage genomes, including single-strand and double-strand DNA, single-strand and double-strand RNA phages and archaeal viruses (Kristensen *et al.,* 2013). Then, taxon-specific marker genes were identified that are never found in other viral taxa (that is, 100% precision), and not found in non-prophage regions of bacterial chromosomes (that is, viral quotient greater than 85% (see Kristensen *et al.* (2013) for details)). The presence of a phage in a given sample is determined by the detection of one of these marker genes. Those markers with recall $\geq 85\%$ (that is, present in $>85\%$ of the genomes in that taxa) and present in at most a single copy per virus are considered quantitative and were used for abundance calculations (Supplementary Table 1).

Analysis of published metagenomic data

Contigs within the metagenomic data from 81 samples from the GOS voyage were downloaded from CAMERA (Sun *et al.,* 2011); the samples are listed in Supplementary Table 4. Subsequently, genes were predicted using MetaGeneMark within the MOCAT pipeline (Kultima *et al.,* 2012). For the three published viromes, the reads were downloaded from the National Center for Biotechnology Information's Short Read Archive (Supplementary Table 5). Then using the SMASHCommunity pipeline, the reads were assembled into contigs and then genes were determined (Arumugam *et al.,* 2010).

Analysis of gut metagenomes

Altogether, 252 metagenomic samples, from 207 individuals, obtained from the MetaHIT project (71 Danish, 39 Spanish; all sampled once; Qin *et al.,* 2010), the NIH Human Microbiome Project (94 US individuals; 51 individuals sampled once, 41 sampled twice and 2 sampled three times; Peterson *et al.,* 2009) and Washington University (three US samples; all sampled once) were analyzed (Turnbaugh *et al.,* 2009). Sample collection and DNA extraction for the MetaHIT, and Human Microbiome Project samples followed their respective protocols (Manichanh *et al.,* 2006; McInnes and

Cutting, 2010). The Illumina sequence reads were processed using MOCAT where the reads were assembled into scaftigs and genes were detected (Kultima *et al.*, 2012). The genes were then clustered using CD-HIT-EST (-c 0.95 aS 0.9) to create a reference gene catalogue, herein called the 252refGene catalogue. Then, the abundance of each of these genes in each sample was determined by mapping the metagenomic reads from each sample to each reference gene using SoapAligner 2.21 (% nt id, minimum read length 45 nt), and then dividing the coverage per base pair by the gene length (bp).

Metagenomic phage taxa detection and abundance calculation

A database of the 252refGene catalogue was created and psiblast was run using the POG profiles to

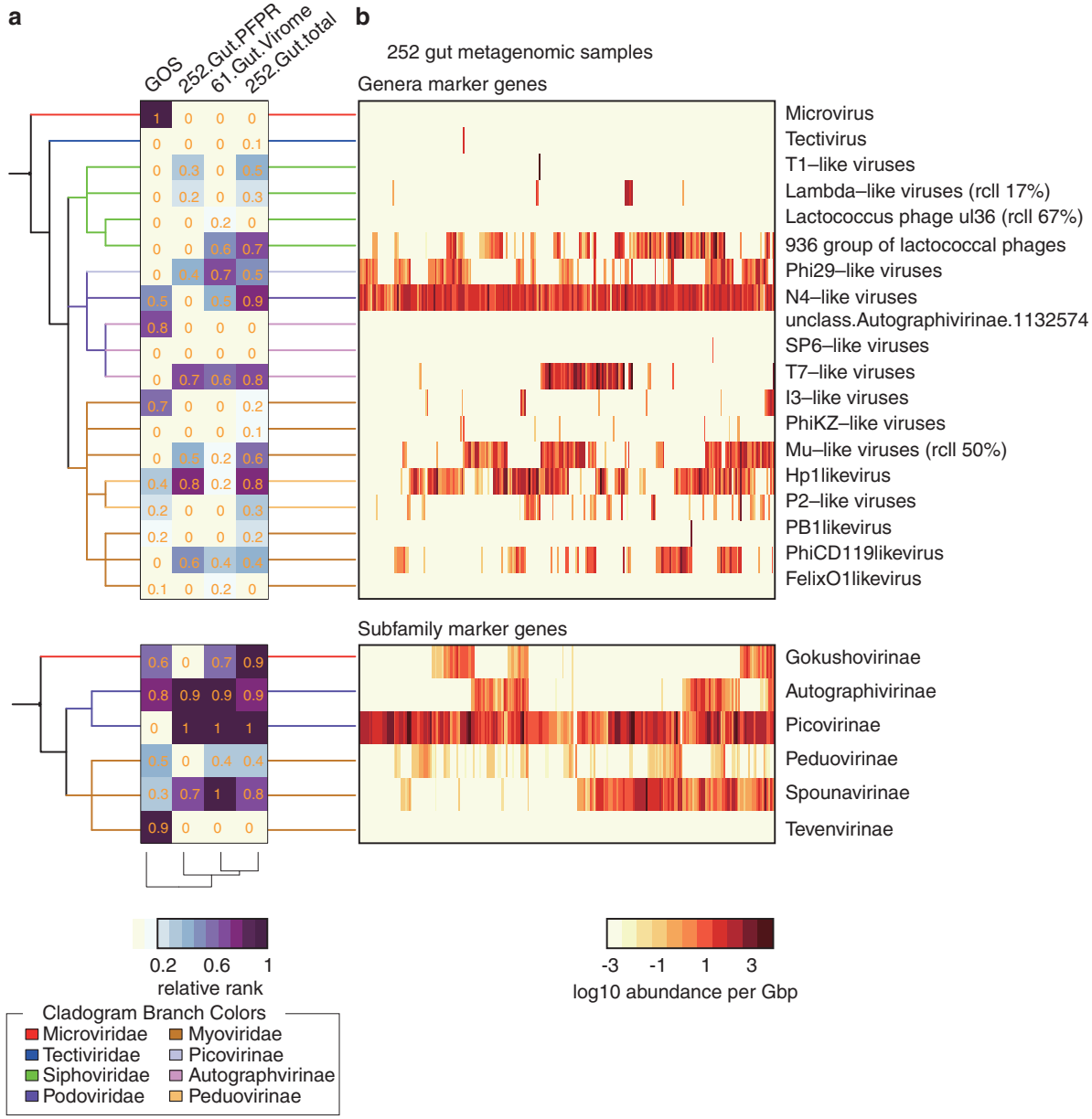

Figure 1 Taxonomic classification and quantification of phages in the human gut and the Oceans. (**a**) Relative rank of phage taxa in the human gut and the Ocean. The abundance of each taxon was used to determine its rank, and the relative rank is then the rank divided by the maximum rank in that sample set. Thereby, a relative rank of 1 is the most abundant taxon in that sample. Each column represents one of the four different sample sets: GOS (82 samples from the Global Ocean Sampling Expedition), 252.Gut.PFPR (only the subset of prophage-encoded genes in 252 gut metagenome samples), 61.Gut.Virome (all of the genes with 61 samples from three published studies of virus-enriched gut metagenomes), 252.Gut.total (all of the genes contained within 252 gut metagenome samples). The columns are clustered by similarity, as illustrated by the dendrogram below the columns. (**b**) Abundance of each phage taxon within the 252 human gut samples (total metagenomic genes). These abundances were derived by dividing the length-normalized base coverages by the total gene abundance in each sample, and multiplied by 10^9 to yield an abundance per Gbp. For those taxa for which marker genes are not quantitative (see Materials and methods), the recall is indicated in brackets.

search for homologues to the POGs. Using all hits with an *E*-value <1E-5, metagenomic reads were assigned to POGs. Then to determine the abundance of each taxon, the abundances of all reference genes that were assigned to a specific POG were summed, and if a phage taxon was represented by multiple marker POGs the mean abundance was used. In addition, to compare taxa abundance across all the samples, the abundance of each taxon was normalized by the total gene abundance for each sample, resulting in a taxon abundance per Gbp total gene abundance. To compare virus abundances across diverse sample sets (Figure 1a), relative ranks were used. For the published datasets (GOS and gut-viromes), abundance was taken from the number of proteins with hits to each marker POG, and for the 252 gut metagenomes, abundance was calculated as described above, but for all markers (rather than just those that are quantitative), and then the rank abundance of each taxa was divided by the maximum rank in each sample, with a relative rank of 1 being the most abundant taxon.

Correlation analyses
To characterize correlations between the abundance of phage taxa and the human host metadata, the sample normalized taxon abundances in the 252 samples within each metadata category were compared using the Wilcoxon test, and *P*-values were adjusted for multiple testing using the fdr method as implemented in R-2.15.0 (R Development Core Team (R Foundation for Statistical Computing), 2011). To determine correlations between the abundance of phage taxa and bacterial taxa, Spearman correlation coefficients were calculated using R, and again the *P*-values were corrected using the fdr method.

Calculation of bacterial taxa abundances
The MOCAT pipeline was used to calculate the abundance of each bacterial taxon. Briefly, metagenomic reads were mapped to a set of 10 universal marker genes from 3496 reference prokaryotic genomes using 97% nucleotide identity. The abundance was calculated as the length-normalized coverage per base pair.

Detection of putative prophages

Prophages in metagenomic scaftigs (scaftig-prophages). To search for prophages that occur within the assembled scaftigs, we ran phage_finderv2.1 (Fouts, 2006) using all the scaftigs assembled in the 252 metagenomes. We then clustered the prophage regions by nucleotide similarity to remove redundancy, using CDhit (-c 0.95 –aS 0.85 –aL 0.7).

Prophages in reference bacterial genomes (refG-prophages). Prophage regions in a set of 2496 reference bacterial genomes downloaded from National Center for Biotechnology Information in November 2011 were determined by running phage_finder_v2.1. To determine which of these reference genome prophage regions occurs in the gut, we used the MOCAT pipeline to map the metagenomic reads to the prophage regions (minimum 95% id) and selected those prophages with a coverage of at least 0.75 bp/bp gene in at least one sample.

Calculation of prophage abundance and prophage to host ratio
To calculate the abundance of each predicted prophage in each sample, the metagenomic reads in each sample were mapped to each prophage region (start and stop coordinates determined by phage_finder) using SOAPaligner (95%), and the length-normalized base coverage was used. Then, to predict prophage lysis, this prophage abundance was divided by the host abundance to obtain a prophage to host (PtoH ratio). For the refG-prophages, the abundance of the host was calculated as a mean abundance of the universal marker genes (Ciccarelli *et al.,* 2006). For the scaftig-prophages, the abundance of the host was calculated as the mean base-pair coverage of the genomic regions, upstream and downstream of the prophage region.

Identification of antibiotic resistance and virulence factor genes on prophages
All prophage-encoded genes were blasted against the ARDB (Liu and Pop, 2009) and VFDB databases (Chen *et al.,* 2012). Sequence identity of >95% with at least 95% overlap was required for a match.

Identification of prophage–host interaction network
Prophage regions were identified in the scaftigs and reference genomes, as described above. Then, to taxonomically classify the prophages, the prophage-encoded proteins were scanned for homologues to the taxon-specific marker genes. For the reference genome prophages that occur in the gut, the taxonomic identity of the bacterial host is known. To determine the taxonomy of the bacterial host for the scaftig prophages, two separate techniques were employed using the scaftig sequences upstream and downstream of the prophage, namely a nucleotide classification method, Phylopithia (McHardy *et al.,* 2007), and BlastN (85% nt id) against reference bacterial genomes. The host identity was assigned to the most specific bacterial taxon at which the up- and downstream sequence agreed.

Results and discussion

Bacteriophage taxa in the human gut and in the ocean
To investigate the phage diversity in the human gut, we created a catalogue of all of the genes found in

252 metagenomic samples, from 207 individuals (Schloissnig *et al.,* 2013). We then searched this gene catalogue for homologues of the phage taxon-specific marker genes, thus identifying phage taxa represented within the given cohort. Using a set of marker genes for 33 taxa at the level of genus or lower, we detected 15 bacteriophage taxa (Figure 1a, column 252.Gut.total). In addition to the genus-specific marker genes, there are marker genes that are specific for six subfamilies; using these markers, five subfamilies were detected, two of which are not represented at the genera level (Figure 1a, column 252.Gut.total).

In addition, we sought to identify which of these phage taxa exist as prophages integrated in bacterial genomes within the metagenomic samples. To this end, we used Phage_Finder (Fouts, 2006) to identify putative prophage regions within the assembled scaftigs, and then searched the prophage-encoded genes for homologues of the taxon-specific marker genes. Within predicted prophage regions of the assembled scaftigs, we detected 7 of the total of 15 identified phage genera. Thus, representatives of these taxa appear to exist as temperate phages within the analyzed human gut samples (Figure 1a, column 252.Gut.PFPR). Conversely, those taxa detected in the 252 total metagenomic samples that were not detected in the prophage fraction may represent strictly virulent phages, such as the abundant *936 lactococcal phages* and *N4-like viruses*. However, for the less abundant taxa (such as *PhiKZ-like* and *I3-like viruses*), the possibility remains that they would be identified within prophage regions upon additional sequencing.

To compare the phage taxa found within these 252 gut samples optimized to enrich prokaryotic metagenomic DNA, to those detected in gut viromes, we searched the genes from three published virome studies (a total of 61 samples; Reyes *et al.,* 2010; Kim *et al.,* 2011; Minot *et al.,* 2011) for homologues of the taxon-specific marker genes. This analysis resulted in the identification of eight viral genera within the virome samples, seven of which were also identified in the 252 total metagenome catalogue, except for the FelixO1-like viruses (Figure 1a, column 61.Gut.Virome). This congruence of the results from the virome and metagenome analyses implies that the sampling methods involved in extracting total metagenomic DNA do not exclude phages in the lytic cycle that represent the bulk of the viromes, although under-representation of the lytic phages in the metagenome data cannot be ruled out. In addition, the fact that the *Gokushovirinae*, which belong to the *Microviridae* family of small viruses, are abundant in the viromes as well as in the 252 total metagenome catalogue indicates that sample extraction of total metagenomic DNA does not exclude these small viruses either contrary to previous suggestions (Stern *et al.,* 2012).

Last, to compare the taxonomic distribution of the phages identified in the gut to that in another well-studied environment, the Ocean, we used the taxon-specific marker genes to identify 13 phage taxa within 82 metagenomic samples from the GOS data set (Rusch *et al.,* 2007; Figure 1a, column GOS). A comparison of the relative ranks of phage taxa in the GOS samples with their relative ranks in the gut samples, as well as the fact that the GOS samples do not cluster with any of the gut samples based on hierarchical clustering of the relative rank profiles (dendrogram above the columns in Figure 1a), illustrates differences in the taxonomic composition between the gut and ocean viromes. Furthermore, the two most abundant taxa (*Microvirus* genus and *Tevenvirinae*) in the GOS samples are absent or rare in the gut. The abundance of the *Microvirus* and *Tevenvirinae* taxa in the Ocean is supported by several independent analyses (Williamson *et al.,* 2008; Tucker *et al.,* 2011; Holmfeldt *et al.,* 2012; Zhao *et al.,* 2013). Although the specific marker gene for the *T4-like virus* genus (a gene present in the *T4-like* phage genus but absent in the sister genus of *Schizot4-like virus*) that is considered abundant in the Ocean was not detected, the subfamily marker for *Tevenvirinae* (a gene shared by all members of both the *T4-like* and *Schizot4-like virus* genera) was detected.

Furthermore, the two most phage abundant taxa in the gut (*Picovirinae* and *Spounavirinae*) are absent or rare in the Ocean. Most of the *Picovirinae* and *Spounavirinae* phage genomes to date have been isolated from human-associated bacteria (for example, *Staphlococcus, Streptococcus, Clostridium, Mycoplasma, Listeria, Enterococcus, Bacillus*). Methodological differences (different sampling and DNA extraction techniques, as well as sequencing methods were used for the data sets) notwithstanding the differences in the relative ranks of the viral taxon between these data sets are pronounced, suggesting habitat preferences for at least some of the viral taxa such as *Picovirinae* and *Spounavirinae*.

Although identification and quantification of 33 phage taxa in these different data sets substantially expand our knowledge of viral ecology, it is essential to keep in mind that the marker gene-based approach relies on reference viral genomes, and thus we are only able to detect a small proportion of the phage taxa that actually exist in these environments. Recent estimates suggest that ~80% of the marine virome is unrelated to reference genome sequences (Hurwitz and Sullivan, 2013), and similarly high estimates of viral 'dark matter' (~70%) have been reported for gut viromes (Modi *et al.,* 2013). Furthermore, even among the known phages with completely sequenced genomes available, >60% remains unclassified at the genus level. Thus, marker genes cover less than half of those viral genera that are classified, mostly due to the extremely low number of virus genomes in the less well-studied taxa. Even when a marker is present, it should also be noted that, for the small and poorly characterized groups,

the possibility cannot be ruled that the marker gene has been shared with viruses belonging to a different as yet undiscovered taxa.

Quantification of phage taxa in 252 gut metagenomes
To quantify intra- and inter-personal variation within each of the 252 metagenomic samples, we then determined the sample-specific abundance of each taxon, by mapping the metagenomic reads from each sample to the taxon-specific marker genes that were identified in the 252 metagenomic gene catalogue. Between the samples, the abundances of most of the phage taxa varies by up to four orders of magnitude (Figure 1b). Overall, the *Picovirinae* subfamily is the most abundant taxon in the samples (Supplementary Figure 1). The *Picovirinae* and the *N4-like viruses* are ubiquitous, being present in 99% of the samples, whereas the majority of the phage taxa appear in only a portion of the samples (20–60%) and another seven taxa occur more rarely, in less than 10% of the samples (Supplementary Figure 2). Interestingly, some phages that were identified in a small fraction of the samples were extremely abundant when present. In particular, *T7-like viruses* that are present only in $<20\%$ of the samples are overall the fifth most abundant group of phages. Within each of the samples, an average of five phage taxa (ranging from 1 to 9) was detected (Supplementary Figure 3). Typically, one to three taxa are highly abundant and the others are several orders of magnitude less abundant, with *Picovirinae, N4-like viruses, Spounavirinae* and *Hp1-like viruses* most often represented among the high abundance taxa (Supplementary Figures 4 and 5). However, in a few samples, a rare taxon is the most abundant one, such as *I3-like, PB1-like, PhiCD119-like* or *T1-like*. It is not clear what leads to these apparent 'blooms' of otherwise rare viral genera.

Correlation of phage taxa abundance with host metadata
To test if any of the variation in the abundance of phage taxa can be explained by metadata of the human host, we used the Wilcoxon signed-rank test to compare the phage abundance profiles between samples with different associated metadata, such as country of origin, gender, age, and disease state (obesity, inflammatory bowel disease, Chron's or ulcerative colitis). We found a statistically significant association between the abundance of the *936 group of lactococcal phages* and the country of origin, with both higher abundance and higher prevalence in the Danish samples (Figure 2a). The *936 group of Lactococcus phages* is a group of virulent phages that infect strains of *Lactococcus lactis*, a bacterial strain used as a starter culture in the manufacturing of cheese and yogurt. Although the abundance of *L. lactis* is low in the final cheese product, group *936 phages* have been detected in the final products, and are known to withstand

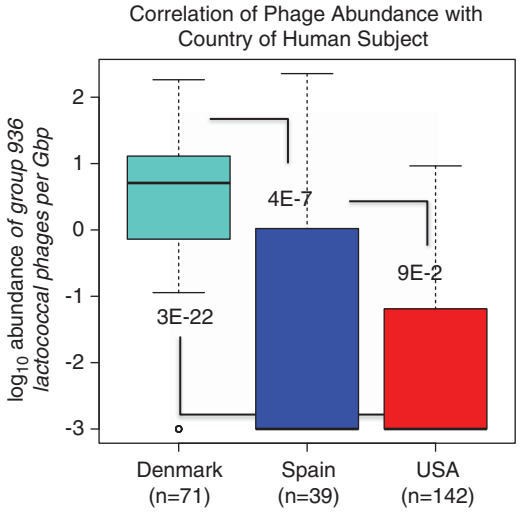

Figure 2 Correlation between the abundance of the group 936 lactococcal phages and the country of origin of the human subject in the study. Boxplots show the abundance of *group 936 lactococcal phages* in samples from individuals from different countries. Fdr-corrected *P*-values between the groups are indicated on the boxplot. Fdr-corrected *P*-values between the groups are indicated on the boxplot.

pasteurization (Mahony *et al.,* 2012). Thus, the high abundance of 936 *lactococcal phages* in the Danish samples might stem from the higher level of consumption of fermented milk products (such as cheese and yougurt) in Denmark (Wielicka and Gorynska-Goldmann, 2005; Agriculture and Horticulture Development Board, 2011; US Department of Agriculture, 2012; Carlucci *et al.,* 2013) and/or from differences between fermentation practices among these countries. As additional large-scale disease-based metagenomic studies with additional metadata become available, the ability to quantify these phage taxa will most likely allow for the identification of correlations between phage abundance and/or phage lysis and disease states.

Individuality and temporal stability of prophage abundance profiles
To investigate the prevalence of temperate phages, which can integrate into bacterial chromosomes as prophages, in the fecal metagenomes, we employed two independent approaches to identify prophages in the metagenomic sequences. First, we searched the assembled scaftigs for prophages using phage_finder (7% false positives and 9% false negatives; Fouts, 2006), resulting in 2518 unique predicted scaftig-prophages. Second, we used phage_finder to identify prophages in a set of reference bacterial genomes that are present in the 252 gut samples, resulting in 463 predicted reference genome prophages (refG-prophages), from 230 unique genomes (detailed information on these prophages and related sequences is available at: http://www.bork.embl.de/Docu/metaG_phage_supp/). Clustering of the scaftig- and refG-prophages by nucleotide

identity (>95%id) revealed that most of the scaftig-prophages are not identically represented in the reference genome prophages as only 13 out of the 2518 clustered with a refG-prophage, suggesting novel prophages in the gut. We then determined the abundance of each prophage region in each sample and defined the prophage abundance profile for each sample as the vector of the proportional abundance of each prophage region.

We then used the prophage abundance profiles to determine if the set of prophages detected within a metagenomic sample can be a 'personalized' indicator unique to the human host. To do this, we examined data from a subset of 43 individuals for which multiple samples (94 in total) were taken over the span of 1 year. We calculated the similarity between prophage abundance profiles for all samples, using the Euclidean distance between the $\log_{10}$ abundance profiles. Different samples from the same individual at two different times were found to be more similar to each other than to any of the other samples, even over a 1-year time period (Figure 3). These findings demonstrate the individual specificity and stability of the prophage abundance profiles.

Quantification of the bacterial lysis by temperate phages

To investigate which of the prophages are active and enter the lytic cycle, we determined the 'lytic potential' of the prophages by calculating the abundance of the prophage sequences relative to the abundance of the host bacterial chromosome PtoH ratio. In theory, if the PtoH ratio is equal to 1, then the prophage is stably integrated (that is, every instance of the phage is integrated within the host chromosome in each genome of the respective bacterial host), whereas if the PtoH ratio is greater than 1, then the phage is at least partially in the lytic phase (that is, viruses are present both within the host chromosome and in virions), and if the PtoH ratio is less than 1, then that prophage region is absent in some of the bacterial host genomes. We compared the PtoH ratios of all prophages, across the 252 samples and found that some of the prophages prefer lysogeny, some are absent in a large fraction of the samples and lysogenized in others, and rare occasions of lysis were detected (Figure 4a).

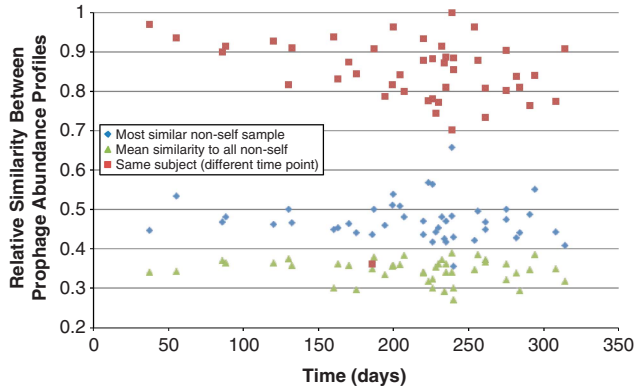

Figure 3 Similarity between prophage abundance profiles of human gut metagenomic samples. For each sample, a prophage abundance distribution was determined (that is, the relative abundance of each prophage region detected in the sample). The Euclidean distance between each sample's prophage abundance profile was then determined. The red dots indicate the similarity between the prophage abundance profiles from the same individual sampled at different time points, the time between the sampling points is indicated on the *x* axis. The blue diamonds represent the similarity between each of the time-point samples and the next most similar sample. Over the whole year sampling period, two samples from the same individual are always most similar to each other than to any other sample.

Some of the detected prophages encode genes that might affect human health, such as genes for antibiotic resistance or virulence factors. The implications of the variation in the prophage patterns is that, although some individuals may harbour near-identical bacterial genomes, presence or absence of a prophage can have phenotypic effects on the bacterial host, as well as ultimately the human host. Trends in the antibiotic resistance capacity of the prophages were similar to those identified in the analysis of the bulk bacterial sequences (Forslund *et al.,* 2013). However, some classes of resistance genes appear to be either enriched or depleted in the prophages, such as Streptomycin-resistant genes, which are enriched, or Erythromycin-resistant genes, which are depleted. In addition, by analyzing the samples for which we have time series for the same person, it can be seen that the PtoH ratio of an individual changes over time and thus represents temporal changes in the extent of lysogenization of the given prophage (Figure 4b).

By conservatively defining an 'active prophage' as one with a PtoH ratio greater than 10, we identified 625 predicted scaftig-prophages that are active in at least two samples, and a core of about 50 active prophages that are present in at least half of the samples (Supplementary Figure 6). For most of the scaftig-prophages, we do not know the bacterial host but for the refG-prophages the hosts are known. Altogether, we identified 200 active predicted refG-prophages (137 unique genomes) present in at least two samples, and a core of about 50 active prophages (43 genomes) that are present in at least half of the samples (Supplementary Figure 7, Supplementary Table 2). These findings greatly expand the list of known prophages that are active in the gut.

To search for correlations between the lytic potential (PtoH ratio) of the predicted prophages and abundances of bacterial taxa, we compared the abundances of the bacterial taxa in each of the 252 samples with all the PtoH ratios in the same samples. There were some statistically significant correlations between the lytic potential of some prophages, and the mean abundance of the non-prophage region of the host bacterial chromosomes.

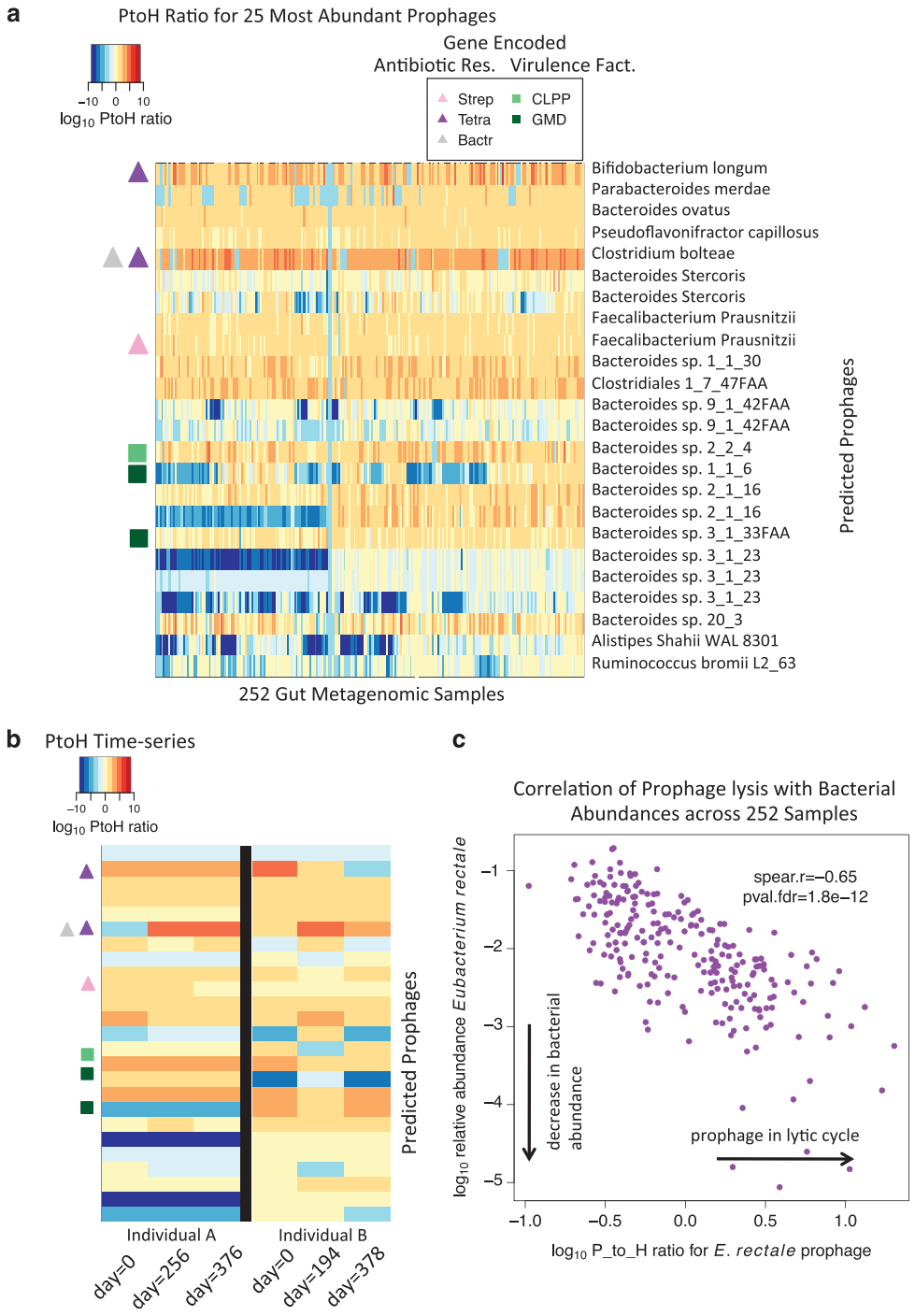

Figure 4 Analysis of prophage lysis. (**a**) A comparison of the PtoH ratio for the 25 most abundant refG-predicted prophages across the 252 samples. The PtoH ratio for a prophage is the abundance of the prophage over the mean abundance of the host chromosome. Each row represents a predicted prophage, the bacterial host of the prophage is indicated on the right of each row, and a triangle or square on the left of the row indicates if the prophage contains an antibiotic resistance (Res.) gene or virulence factor (Fact.) gene: Strep (steptomycin resitance), Tetra (tetracyclin resistance), Bact (bactracin resistance), CLPP (casinolytic protease), GMD (GDP-mannose 4,6-dehydratase). Each column represents one of the 252 gut metagenomic samples, and the colour of the cell indicates the $\log_{10}$PtoH ratio, with yellow indicating that the prophage is lysogenized, blue indicating that the prophage is absent and red indicating that the prophage is in lysis. (**b**) Trends of the PtoH ratio over time for the 25 most abundant prophages. The PtoH ratio was plotted for the two individuals with samples from three different time points. Similar temporal variability was observed in the individuals that were only sampled twice. (**c**) Correlation of prophage lysis with bacterial abundances across 252 samples. The *x* axis shows the $\log_{10}$ PtoH ratio for the *Eubacterium rectale* prophage in the 252 samples. The *y* axis shows the $\log_{10}$ relative abundance of the host bacterium in the samples. The spearman correlation coefficient and fdr-corrected *P*-value are shown in the upper right hand corner. The negative correlation indicates that in samples where the prophage is in the lytic phase, the relative abundance of the bacterial host is lower, presumably due to bacterial lysis by the phage, indicating that the PtoH ratio reflects the lytic state of the phage.

Most of these are negative correlations between a prophage and its known bacterial host, such as *Eubacterium rectale* or *Clostridium leptum* (Figure 4c). Thus, in samples with a high lytic potential of a given prophage, the abundance of the respective host bacterium was lower than in samples with a low-lytic potential. These correlations lend credence to the use of the PtoH ratio as an indicator of the lytic or lysogenic state of prophages.

A network of temperate phages and their bacterial hosts in the gut

Finally, to detect association of specific phage taxa incorporated within certain bacterial taxa as prophages, we combined the taxonomic classification of the gut prophages, with taxonomic classification of the surrounding host bacterial chromosomes, to produce a network of temperate phages and their bacterial hosts in the gut (Figure 5). Seven viral genera were detected (*P22-like, P2-like, Hp1-like, Mu-like, T7-like, Phi29-like* and *PhiCD119-like*), all of which yielded novel interactions with bacterial taxa. Some phage taxon appear to have a more specific host range, such as the *P22-like* genus, which is only connected to genera of the *Enterobacteraceae* family (*Escherichia, Salmonella, Klebsiella*), or *Spounavirinae,* which only interacts with genera of the *Bacteroidales* order; whereas others have a broader host range such as the *Hp1-like* and *Mu-like viruses,* which each connect with bacterial taxa from four different phyla. Some of the unclassified subfamilies that are represented by National Center for Biotechnology Information taxids (for example, 196894) are also highly connected, apparently infecting a number of different bacterial taxa. Although previous work in marine virology identified phage strains as having either specific or broad host ranges (even infecting different bacterial phyla; Sullivan *et al.,* 2003), recent analysis of large-scale marine infection networks revealed many more specialists than generalists among viruses (Flores *et al.,* 2013). These studies also have shown that geographic diversity patterns of phage and host impacted the infection network (Flores *et al.,* 2013). Given the diversity of bacterial taxa present in the confined human intestines, broader host ranges for gut-associated phage seem plausible.

Multiple novel associations were identified for some of the most abundant members of the gut microbiome, such as *Bacteroides, Eubacterium, Faecalibacterium* and *Prevotella*. The most highly connected bacterial taxa are the genus *Escherichia*, the orders *Selenomonadales* and *Clostridiales,* which are infected by three, five and five different phage taxa. All of the interactions for *Escherichia* are derived from the analysis of reference genomes, reflecting its dominance in the genome databases. Conversely, all of the interactions with the *Selenomonadales* order were derived from the analysis of the scaftigs, which probably reflects the lack of gut-specific *Selenomondales* genomes in the reference genome database. Analysis of marine

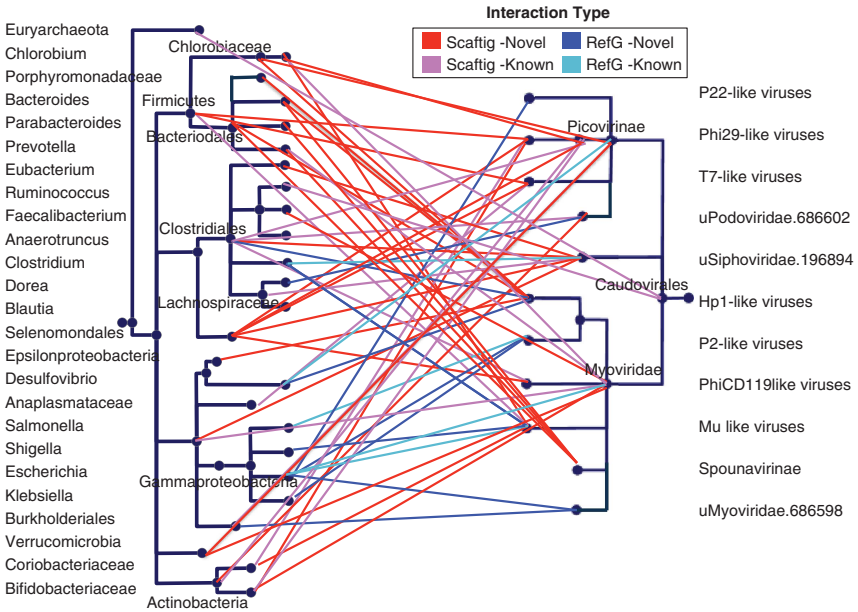

Figure 5 Network of temperate phages and their bacterial hosts in the gut. The cladogram on the right indicates the taxonomy of the phage taxa that were found as integrated prophages. A line connects the phage taxa to its bacterial host on the left side. The taxon labels besides the cladograms indicate the leaves of the branches, internal nodes involved in interactions are labelled within the cladograms. The colour of the line indicates if the interaction was determined from analysis of assembled metagenomic scaftigs (scaft-prophages), red and pink, or from analysis of reference bacterial genomes that occur in the gut (refG-prophages), blue and turquoise. In addition, red and dark blue indicate that these interactions have not been reported in the literature, whereas pink and turquoise represent known interactions.

phage–bacterial infection networks also shows that many marine bacteria are infected by multiple phages (Flores *et al.,* 2013); the present results indicate that this trend holds across diverse environments.

This first network of gut phage–bacteria associations, albeit harbouring lots of novelty, can only be seen as a lower limit and is likely to be considerably larger. This is due to biases in metagenomic sampling that do not capture all lytic phages, and the currently short read lengths that prevent a higher fraction of bacterial hosts to be associated to prophages as well as strict thresholds for taxonomic identification of bacterial hosts that had to be implemented to avoid false positives. However, novel experimental techniques now allow one to follow-up on these identified interactions (Tadmor *et al.,* 2011; Deng *et al.,* 2012) and even determine the dynamics of some of the phage infections (Allers *et al.,* 2013).

Conclusions

Through the application of taxon-specific marker genes to a large collection of metagenomes, we identified and quantified phage taxa that are present in the human gut. To our knowledge, this is the first large-scale study on taxonomic classification and quantification of phages in the human gut down to the genus level, and the results substantially expand our knowledge of the phage diversity as well as intra- and inter-personal variation in the abundances of specific phage taxa. A comparison of the gut phage repertoire to that of the Ocean revealed substantial differences. The gut virome seems to be more diverse than the ocean virome, although we cannot rule out that this difference was due to a poorer recognition of marine viruses with our set of markers. By quantifying the phage genera, we identified a significant enrichment of a phage taxon according to the country of the host and found correlations between the abundances of phage and bacterial taxa. Then, through systematic detection of prophage regions within this large data set, we identified hundreds of putative novel prophages. Furthermore, the analysis of prophage-abundance profiles from the same individual over the course of a year shows that these profiles are significantly individual-specific. By calculating the lytic potential of the identified prophages, we delineated a core of over 50 prophages that are capable of lysis and are present in at least half of the samples, and identified correlations between the lytic state of certain prophage regions and the abundances of the host bacteria. By combining the taxonomic classifications of the prophages and the bacterial hosts, we derived a network of temperate phages that infect bacteria in the human gut. These phage–bacteria associations together with the quantification of phage taxa and their lytic potential pave the way for a better understanding of the role of phages in the ecology of microbial communities. Although the analysis reported here most likely identifies only a small fraction of the gut virome, it nevertheless substantially expands the knowledge on virus–host interactions in the gut.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

ASW was supported by a grant from the International Human Microbiome Consortium (IHMS). DMK and EVK are supported by intramural funds of the US Department of Health and Human Services (to the National Library of Medicine, NIH).

References

Agriculture and Horticulture Development Board (2011). Dairy statistics. An insider's guide 2011, http://www.dairyco.org.uk.

Allen HK, Bunge J, Foster JA, Bayles DO, Stanton TB. (2013). Estimation of viral richness from shotgun metagenomes using a frequency count approach. *Microbiome* **1**: 5.

Allers E, Moraru C, Duhaime MB, Beneze E, Solonenko N, Barrero-Canosa J *et al.* (2013). Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. Environ. *Microbiol* **15**: 2306–2318.

Angly FE, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P *et al.* (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**: 41.

Arumugam M, Harrington ED, Foerstner KU, Raes J, Bork P. (2010). SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* **26**: 2977–2978.

Barr JJ, Auro R, Furlan M, Whiteson KL, Erb ML, Pogliano J *et al.* (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc Natl Acad Sci USA* **110**: 10771–10776.

Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P *et al.* (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**: 6220–6223.

Brüssow H, Canchaya C, Hardt W-D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* **68**: 560–602.

Busby B, Kristensen DM, Koonin EV. (2013). Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens. *Environ Microbiol* **15**: 307–312.

Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann M-L, Brüssow H. (2003). Phage as agents of lateral gene transfer. *Curr Opin Microbiol* **6**: 417–424.

Carlucci D, Stasi A, Nardone G, Seccia A. (2013). Explaining price variability in the italian yogurt market: a hedonic analysis. *Agirbusiness* **0**: 1–13.

Carrolo M, Frias MJ, Pinto FR, Melo-Cristino J, Ramirez M. (2010). Prophage spontaneous activation promotes DNA release enhancing biofilm formation in *Streptococcus pneumoniae*. *PLoS One* **5**: e15678.

Chen L, Xiong Z, Sun L, Yang J, Jin Q. (2012). VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res* **40**: D641–D645.

Cho I, Blaser MJ. (2012). The human microbiome: at the interface of health and disease. *Nat Rev Genet* **13**: 260–270.

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

De Vos WM, de Vos EAJ. (2012). Role of the intestinal microbiome in health and disease: from correlation to causation. *Nutr Rev* **70**(Suppl 1): S45–S56.

Deng L, Gregory A, Yilmaz S, Poulos BT, Hugenholtz P, Sullivan MB. (2012). Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging. *MBio* **3**: pii e00373–12.

Duerkop BA, Clements CV, Rollins D, Rodrigues JLM, Hooper LV. (2012). A composite bacteriophage alters colonization by an intestinal commensal bacterium. *Proc Natl Acad Sci USA* **109**: 17621–17626.

Duhaime MB, Sullivan MB. (2012). Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **434**: 181–186.

Flores CO, Valverde S, Weitz JS. (2013). Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *ISME J* **7**: 520–532.

Forslund K, Sunagawa S, Kultima JR, Mende DR, Arumugam M, Typas A *et al.* (2013). Country-specific antibiotic use practices impact the human gut resistome. *Genome Res* **23**: 1163–1169.

Fouts DE. (2006). Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* **34**: 5839–5851.

Holmfeldt K, Odić D, Sullivan MB, Middelboe M, Riemann L. (2012). Cultivated single-stranded DNA phages that infect marine Bacteroidetes prove difficult to detect with DNA-binding stains. *Appl Environ. Microbiol* **78**: 892–894.

Hurwitz BL, Sullivan MB. (2013). The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* **8**: e57355.

Kim M-S, Park E-J, Roh SW, Bae J-W. (2011). Diversity and abundance of single-stranded DNA viruses in human faeces. *Appl Environ Microbiol* **77**: 8062–8070.

Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. (2010). New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* **18**: 11–19.

Kristensen DM, Waller AS, Yamada T, Bork P, Mushegian AR, Koonin EV. (2013). Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J Bacteriol* **195**: 941–950.

Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR *et al.* (2012). MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One* **7**: e47656.

Liu B, Pop M. (2009). ARDB–Antibiotic Resistance Genes Database. *Nucleic Acids Res* **37**: D443–D447.

Mahony J, Murphy J, van Sinderen D. (2012). Lactococcal 936-type phages and dairy fermentation problems: from detection to evolution and prevention. *Front Microbiol* **3**: 335.

Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L *et al.* (2006). Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**: 205–211.

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4**: 63–72.

McInnes P, Cutting M. (2010). Core Microbiome Sampling, Protocol A—HMP Protocol #07-001.

Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD *et al.* (2011). The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res* **21**: 1616–1625.

Modi SR, Lee HH, Spina CS, Collins JJ. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**: 219–222.

Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA *et al.* (2009). The NIH Human Microbiome Project. *Genome Res* **19**: 2317–2323.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**: 59–65.

Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F *et al.* (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**: 55–60.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria.

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F *et al.* (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**: 334–338.

Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D *et al.* (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**: 3074–3075.

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

Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A *et al.* (2013). Genomic variation landscape of the human gut microbiome. *Nature* **493**: 45–50.

Sharon I, Battchikova N, Aro E-M, Giglione C, Meinnel T, Glaser F *et al.* (2011). Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J* **5**: 1178–1190.

Solonenko SA, Ignacio-Espinoza JC, Alberti A, Cruaud C, Hallam S, Konstantinidis K *et al.* (2013). Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics* **14**: 320.

Stern A, Mick E, Tirosh I, Sagy O, Sorek R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res* **22**: 1985–1994.

Sullivan MB, Waterbury JB, Chisholm SW. (2003). Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. *Nature* **424**: 1047–1051.

Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S *et al.* (2011). Community cyberinfrastructure for advanced

microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res* **39**: D546–D551.

Tadmor AD, Ottesen EA, Leadbetter JR, Phillips R. (2011). Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science* **333**: 58–62.

Tucker KP, Parsons R, Symonds EM, Breitbart M. (2011). Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J* **5**: 822–830.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.

US Department of Agriculture. (2012). Food Intakes Converted to Retail Commodities Data: Table 3—Dairy Consumption, http://www.ers.usda.gov/data-products/commodity-consumption-by-population-characteristics.aspx#.UgJp2FNKnfY.

Vujkovic-Cvijin I, Dunham RM, Iwai S, Maher MC, Albright RG, Broadhurst MJ *et al.* (2013). Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Sci Transl Med* **5**: 193ra91.

Wang X, Kim Y, Wood TK. (2009). Control and benefits of CP4-57 prophage excision in *Escherichia coli* biofilms. *ISME J* **3**: 1164–1179.

Wielicka A, Gorynska-Goldmann E. (2005). World and Poland per capita cheese consumption. *Rocz Akad Rol w Pozaniu* **4**: 157–166. http://www.jard.edu.pl/pub/17_4_2005.pdf (Accessed on 3 June 2013).

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI *et al.* (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. Hall, N, ed. *PLoS One* **3**: e1456.

Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S *et al.* (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci* **6**: 427–439.

Zhang T, Breitbart M, Lee WH, Run J-Q, Wei CL, Soh SWL *et al.* (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* **4**: e3.

Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC *et al.* (2013). Abundant SAR11 viruses in the ocean. *Nature* **494**: 357–360.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)