nature biotechnology

Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes

H Bjørn Nielsen[1,2,32], Mathieu Almeida[3–5,32], Agnieszka Sierakowska Juncker[1,2], Simon Rasmussen[1], Junhua Li[6–8], Shinichi Sunagawa[9], Damian R Plichta[1], Laurent Gautier[1], Anders G Pedersen[1], Emmanuelle Le Chatelier[3,4], Eric Pelletier[10–12], Ida Bonde[1,2], Trine Nielsen[13], Chaysavanh Manichanh[14], Manimozhiyan Arumugam[7,9,13], Jean-Michel Batto[3,4], Marcelo B Quintanilha dos Santos[1], Nikolaj Blom[2], Natalia Borruel[14], Kristoffer S Burgdorf[13], Fouad Boumezbeur[3,4], Francesc Casellas[14], Joël Doré[3,4], Piotr Dworzynski[1], Francisco Guarner[14], Torben Hansen[13,15], Falk Hildebrand[16,17], Rolf S Kaas[18], Sean Kennedy[3,4], Karsten Kristiansen[7,19], Jens Roat Kultima[9], Pierre Léonard[3,4], Florence Levenez[3,4], Ole Lund[1], Bouziane Moumen[3,4], Denis Le Paslier[10–12], Nicolas Pons[3,4], Oluf Pedersen[13,20–22], Edi Prifti[3,4], Junjie Qin[6,7], Jeroen Raes[17,23,24], Søren Sørensen[25], Julien Tap[9], Sebastian Tims[26], David W Ussery[1], Takuji Yamada[9,27], MetaHIT Consortium[28], Pierre Renault[3], Thomas Sicheritz-Ponten[1,2], Peer Bork[9,29], Jun Wang[7,13,19,30], Søren Brunak[1,2] & S Dusko Ehrlich[3,4,31]

Most current approaches for analyzing metagenomic data rely on comparisons to reference genomes, but the microbial diversity of many environments extends far beyond what is covered by reference databases. *De novo* segregation of complex metagenomic data into specific biological entities, such as particular bacterial strains or viruses, remains a largely unsolved problem. Here we present a method, based on binning co-abundant genes across a series of metagenomic samples, that enables comprehensive discovery of new microbial organisms, viruses and co-inherited genetic entities and aids assembly of microbial genomes without the need for reference sequences. We demonstrate the method on data from 396 human gut microbiome samples and identify 7,381 co-abundance gene groups (CAGs), including 741 metagenomic species (MGS). We use these to assemble 238 high-quality microbial genomes and identify affiliations between MGS and hundreds of viruses or genetic entities. Our method provides the means for comprehensive profiling of the diversity within complex metagenomic samples.

Natural microbial communities typically contain a wide diversity of organisms, viruses, and other chromosomal and extra-chromosomal genetic elements. The microbiome of the human distal gut is among the most complex communities ever studied, with an estimated 1,000 different microbial species across human populations[1] and millions of different genes[2]. Current computational analysis strategies for metagenomic data rely largely on comparisons to reference genomes from cultivated microbes. However, these reference genomes represent only a fraction of the species and viruses present. Moreover, bacterial genomes from different isolates of the same species usually show considerable genetic heterogeneity when compared[3,4]. This variation may be the result of clonal differences, environmental

[1]Center for Biological Sequence Analysis, Technical University of Denmark, Kongens Lyngby, Denmark. [2]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark. [3]INRA, Institut National de la Recherche Agronomique, UMR 14121 MICALIS, Jouy en Josas, France. [4]INRA, Institut National de la Recherche Agronomique, US 1367 Metagenopolis, Jouy en Josas, France. [5]Department of Computer Science, Center for Bioinformatics and Computational Biology, University of Maryland, USA. [6]BGI Hong Kong Research Institute, Hong Kong, China. [7]BGI-Shenzhen, Shenzhen, China. [8]School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China. [9]European Molecular Biology Laboratory, Heidelberg, Germany. [10]Commissariat à l'Énergie Atomique et aux Énergies Alternatives, Institut de Génomique, Évry, France. [11]Centre National de la Recherche Scientifique, Évry, France. [12]Université d'Évry Val d'Essonne, Évry, France. [13]The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. [14]Digestive System Research Unit, University Hospital Vall d'Hebron, Ciberehd, Barcelona, Spain. [15]Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark. [16]Department of Structural Biology, VIB, Brussels, Belgium. [17]Department of Bioscience Engineering, Vrije Universiteit, Brussels, Belgium. [18]National Food Institute, Division for Epidemiology and Microbial Genomics, Technical University of Denmark, Kongens Lyngby, Denmark. [19]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [20]Hagedorn Research Institute, Gentofte, Denmark. [21]Institute of Biomedical Science, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [22]Faculty of Health, Aarhus University, Aarhus, Denmark. [23]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Belgium. [24]VIB Center for the Biology of Disease, Leuven, Belgium. [25]Section of Microbiology, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [26]Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands. [27]Department of Biological Information, Tokyo Institute of Technology, Yokohama, Japan. [28]A full list of members and affiliations appears at the end of the paper. [29]Max Delbrück Centre for Molecular Medicine, Berlin, Germany. [30]Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia. [31]King's College London, Centre for Host-Microbiome Interactions, Dental Institute Central Office, Guy's Hospital, United Kingdom. [32]These authors contributed equally to this work. Correspondence should be addressed to S.B. (brunak@cbs.dtu.dk) or S.D.E. (dusko.ehrlich@jouy.inra.fr).

Received 12 February; accepted 22 May; published online 6 July 2014; doi:10.1038/nbt.2939

adaptation or possibly artifacts from the cultivation process. Therefore, reference genomes represent only a small proportion of the biological diversity of microbial systems, and thus methods relying on them place limitations on structuring and analyzing metagenomic data. In particular, they limit our ability to segregate metagenomic data into coherent biological entities and fail to describe previously unknown species, phages and modules of genetic variation within microbial species.

De novo assembly of genomes from complex metagenomic data is inherently difficult due to the many sequence ambiguities that confuse the assembly process[5]. Hence, a typical metagenomic assembly will result in a large set of independent contigs that are not easily aggregated into biological entities[6]. Previous methods have addressed the assembly problem from different angles. Iverson *et al.*[7] used tetranucleotide-based binning to assist the assembly of species with a distinct and consistent skew in base composition. In complex communities this approach works only for a few organisms with extreme base compositions. Moreover, some known genomes have inconsistent tetranucleotide distributions, which would compromise any assembly using this approach. An alternative is to use differences in the abundance of genetic sequences measured in a single biological sample to separate organisms[8,9]. Such methods rely on the notion that abundance is constant across genetic entities (i.e., each gene on a specific bacterial chromosome will be found in a sample in the same abundance as any other gene on that chromosome). These abundance-based methods have been used to segregate the most abundant species in waste water[9], which is a community with limited complexity, or to further segregate a subset of sequence reads with similarity to known reference genomes[8]. However, both Albertsen *et al.*[9] and Wang *et al.*[8] acknowledge that their methods cannot segregate taxonomically related species or genera, respectively; also these methods cannot give comprehensive segregation of all entities in complex samples.

Proper structuring of the complete metagenomic composition is important not only for understanding the microbial communities[10], but also for making statistical associations between the metagenomic data and descriptors of the system. In the case of the human microbiome such descriptors include clinical parameters of the human host. For example, we have previously used co-abundance profiles to bin the 2% of a gene catalog with strongest correlation to the human type 2 diabetes phenotype[11]. This was manageable with 2% of the genes but such clustering using distance matrices is not possible for an entire microbiome (as calculating a 3.9 million × 3.9 million gene distance matrix is impractical even for large supercomputers).

The method we present here is based on segregating biological entities by co-abundance but overcomes the limited resolution of previous methods (e.g., not being able to segregate related organisms) and enables the complete segregation of complex metagenomic samples. The increased resolution is achieved by using co-abundance profiles across many samples. Here we use 396 samples, but we also show that species can be segregated accurately using as few as 18 samples. The computational problem of generating a distance matrix for a complete metagenome is overcome by using a method that extracts groups of genes that correlate (in terms of abundance) to randomly picked seed genes. This clustering approach can be done in hours on a powerful desktop computer.

Segregating a metagenome into groups of genes that have similar abundance (CAGs) allows identification of biological entities like species and phages, as well as small genetic entities representing co-inherited clonal heterogeneity. Phage sequences have previously been identified in complex communities by sequence similarity or size separation before sequencing[12,13], but a general method that can identify novel species, phages and genetic heterogeneity from generic metagenomics data has been lacking. In addition, small biological entities have, with a few exceptions[14,15], not been affiliated to specific microbial species within the community.

Here we assign genes from 396 human gut microbiome samples into 7,381 CAGs, and define subsets of these as MGS and phage-like CAGs. From the MGS, we assembled 238 unique genomes that meet the high-quality draft genome standard of the Human Microbiome Project[16]. We found that a large set of smaller CAGs could be affiliated to the MGS by dependency associations, and that the persistence of some MGS is related to the occurrence of specific dependency-associated CAGs.

RESULTS

Comprehensive co-abundance gene segregation

Our method for co-abundance segregation uses a metagenomic data set consisting of a number of samples of the same type. In this study we use a deeply sequenced data set of 396 human stool samples from Spanish and Danish individuals, including 124 samples from a previous study[2] (see Online Methods and **Supplementary Data 1** for details). Seventy-seven of the Spanish individuals were sampled twice, with an average of 6 months between the samplings. The first step is the assembly of the sequencing reads from each sample into genes, which are then pooled into a nonredundant gene catalog (**Fig. 1**). Our gene catalog contained 3,871,657 genes (**Supplementary Fig. 1**); only 10% of this gene catalog could be assigned to a taxonomic group at the species level (**Supplementary Note 1** and Online Methods). We picked a gene at random as a 'seed' and defined this and other genes with similar abundance profiles as a 'canopy', which was defined as those genes with Pearson correlation coefficient (PCC) > 0.9 to the seed gene profile. A canopy profile was then determined as the median abundance profile of the gene group and was used iteratively for recapturing the canopy until the canopy profile stabilizes (**Supplementary Fig. 2**). New random seed genes were picked until all genes were assigned to a canopy. Canopies that passed a canopy rejection criteria—canopies must contain more than two genes and 90% of the canopy profile signal must originate from more than three samples—were identified as CAGs. It should be noted that large canopies tend to be determined quickly, thereby reducing the computational cost of subsequent canopies. The method depends on two parameters: the gene inclusion criterion and the canopy rejection criteria.

Clustering of our data set binned 1.53 million genes (representing 68% of the mapped sequence reads; the remaining genes were in canopies that did not pass the rejection criteria) into 7,381 CAGs, which ranged in size from 3 to 6,319 genes. The size distribution of the groups, in terms of numbers of genes, was bimodal with peaks at ~50 genes and ~1,700 genes, respectively (**Fig. 2a**).

Most complete genomes of bacteria or archaea contain >700 genes and the 741 largest CAGs had >700 genes (**Supplementary Fig. 3** and **Supplementary Note 2**). The genes in these 741 CAGs were highly consistent in base composition, had highly correlated abundance profiles in an independent set of 115 samples[17] (**Supplementary Note 3**) and had consistent taxonomical annotation. For 115 of these CAGs, >95% of the taxonomically annotated genes were annotated to the same species (**Supplementary Figs. 4–6** and **Supplementary Data 2**) and most were similar to a reference sequence from a specific strain within a species. We refer to these CAGs with >700 genes as MGS. In nine cases we identified several distinct MGS from the same species (e.g., three from *Fecalibacterium prausnitzii*).

The individual gene-abundance profiles of all MGS were highly coherent and most were distinct from genes not included in the MGS.

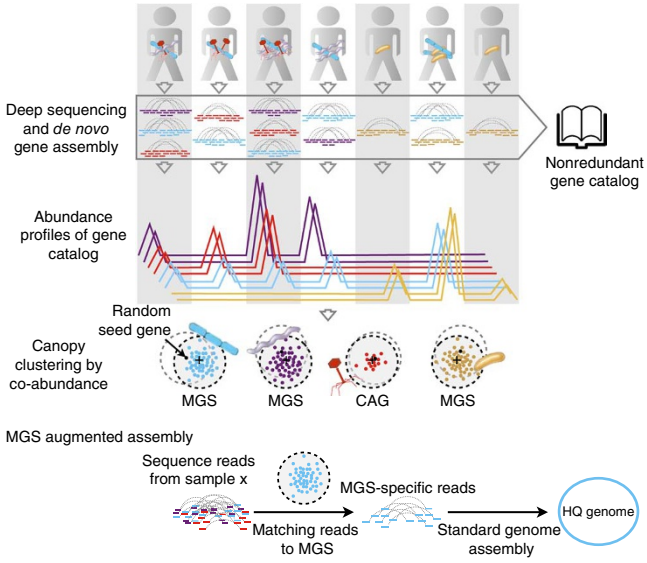


Figure 1 Overview of co-abundance clustering and the MGS-augmented assembly. DNA from a series of independent biological samples from microbial communities, here originating from the human gut microbiome, is extracted and shotgun sequenced. Genes assembled and identified in individual samples are then integrated to form a cross-sample, nonredundant gene catalog. The abundance profile of each gene in the catalog is assessed by counting the matching sequence reads in each sample. To facilitate co-abundance clustering of large gene catalogs, we used random seed genes as 'baits' for identifying groups of genes that correlate (PCC > 0.9, gray dashed circle) to the abundance profile of the bait genes. The fixed PCC distance threshold is called a canopy (dashed circles). To center the canopy on a co-abundance gene group (CAG), the median gene abundance profile of the genes within the original seed canopy (or subsequent canopies, symbolized as +) is used iteratively to recapture a new canopy until it settles on a particular profile (off-set circles). The gene content of a settled canopy (black dashed circles) is named a metagenomic species (MGS) if it contains 700 or more genes. The smaller groups remain referred to as CAGs. Sequence reads from individual samples that map to the MGS genes and their contigs are then extracted and used to assembly a draft genome sequence for an MGS; we refer to this process as MGS-augmented genome assembly. The use of sample-specific sequence reads in the assemblies helps discriminate between closely related strains.

Consequently, changing the gene inclusion criterion to PCC > 0.8 or PCC > 0.95 increased or reduced the size of the average MGS by only 5% and 17%, respectively (**Supplementary Figs. 7** and **8a** and **Supplementary Note 4**). In contrast, when we defined gene sets by sequence similarity to reference genomes, the abundance profiles were inconsistent (**Supplementary Fig. 8b**). Importantly, clustering of randomly permuted abundance profiles only resulted in very few canopies that pass the canopy rejection criterion and all of these were small (**Supplementary Fig. 9**).

Nineteen of the individuals sampled consumed a defined fermented milk product containing the previously sequenced *Bifidobacterium animalis subsp. lactis CNCM I-2494* (ref. 18), and we used this species as a benchmark to assess the ability of our method to identify and segregate a particular species. Although on average only 0.3% of the sequence reads in the 19 samples originated from *B. animalis*, we were able to capture 95% of the *B. animalis* reference genes into one MGS (MGS:337). Subsampling of the data showed that the *B. animalis* MGS can be segregated using as little as 700 K sequence reads per sample or from a much smaller sample set consisting of only 18 samples (**Fig. 3**).

Together the MGS and CAGs provide a detailed overview of the microbial community and precise estimates of the species richness that are in strong agreement with the observed gene richness (PCC = 0.96; **Supplementary Note 5** and **Supplementary Fig. 10**).

Genome assembly

Grouping genes into an MGS using our co-abundance method also assists *de novo* genome assembly by providing the basis for segregating sequence reads into those that derive from a distinct biological entity (**Fig. 1**). That is, for an individual sample, reads can be selected that map to a particular MGS; these are species-specific subsets of the full set of reads from that metagenomic sample. We used standard genome assembly tools to assemble these subsets of reads for all MGS. Of these 238 unique MGS genomes—including 181 new genomes from previously unsequenced species—met the Human Microbiome Project high-quality draft genome standard (**Supplementary Figs. 11** and **12** and **Supplementary Data 3**). We refer to this strategy as MGS-augmented assembly. The MGS-augmented assembly of MGS:337 covered 95% of the reference genome of the benchmark species *B. animalis subsp. lactis CNCM I-2494* with 99.9% identity (**Fig. 4**). In addition, 44 of the MGS-augmented assemblies were closely related to known reference genomes and covered an average of 78% of these genomes with an average sequence identity of 98.4% (**Supplementary Data 4**).

To reduce the possibility of making chimeric assemblies of closely related strains, we used only sequence reads from a single sample

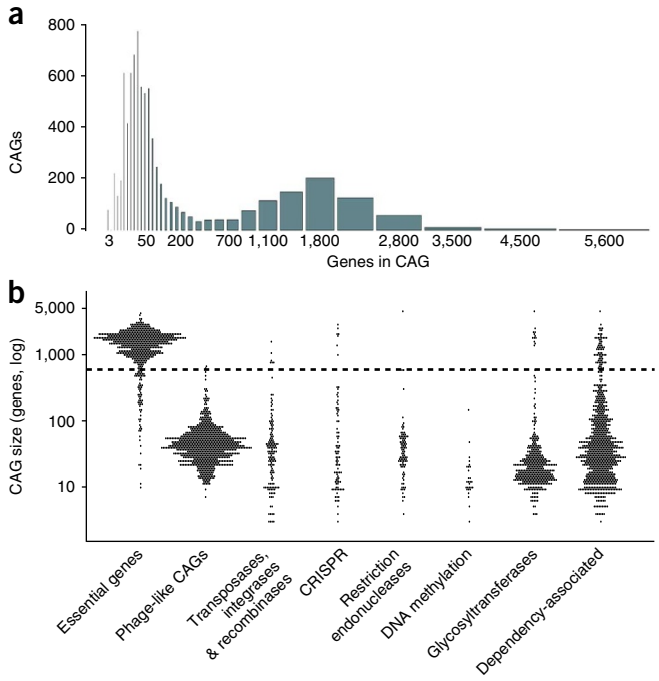


Figure 2 Size distributions of co-abundance gene groups (CAGs). (**a**) Histogram showing the CAG size distribution in terms of gene content. The scale is logarithmic as indicated by the bar widths. (**b**) Bee swarm plot showing CAGs that are significantly enriched (Fishers exact test, $P < 0.001$) for the indicated gene annotation, as well as phage-like CAGs and dependency-associated CAGs, plotted against the number of genes contained in the CAGs. Here every point represents an enriched CAG or MGS and the width of the swarms shows the distribution. The dashed line marks the 700-gene threshold separating small CAGs from MGS.

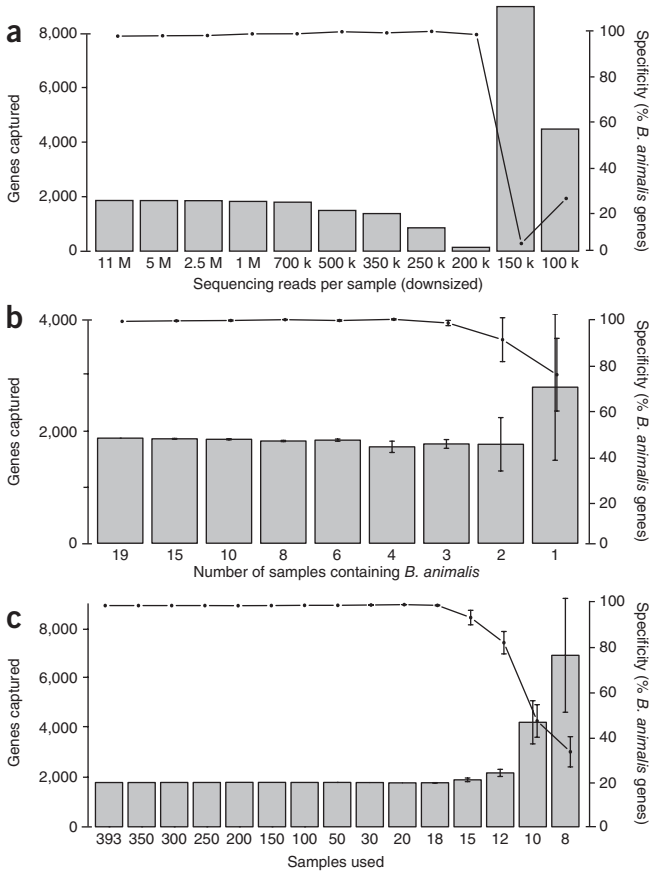


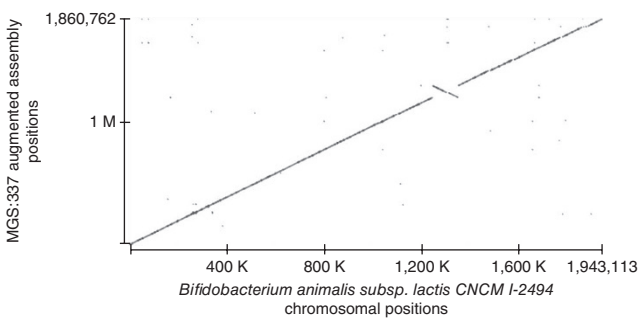
Figure 3 Benchmarking sensitivity and specificity of the co-abundance clustering across a range of sequencing depths or sample numbers. *B. animalis subsp. lactis CNCM I-2494* was used as a benchmark species because 19 samples originated from individuals who had consumed a defined fermented milk product containing this strain. (For each clustering, the size (number of genes captured) is shown as bars (left axis); and the specificity (percentage of genes matching the *B. animalis* reference genome with > 95% sequence identity over 100 bp or better captured in the MGS that is most similar to *B. animalis*) is shown as a line (right axis).) (**a**) Co-abundance clustering using reduced data sets to simulate the sequencing depths indicated (*x* axis). At a sequence depth of 700 K reads, 97% of the *B. animalis* genes were captured, and at a depth of 200 K reads 98.6% of the captured genes were from *B. animalis*. (**b**) Co-abundance clustering of random sample subsets containing the indicated number of samples (*x*-axis) from individuals that consumed the DFMP. Here the total sample size was kept constant at 375 samples. (**c**) Co-abundance clusterings on a series of random sample subsets of the indicated size (*x* axis). These sample subsets included 19 samples from individuals who had consumed the DFMP, except when they contained <19 samples (i.e., 19-8 DFMP individuals per subset). In **b** and **c**, samples were downsized to 11 million sequence reads per sample. Error bars, ±1 s.d. from the mean ($n = 5$).

for every such assembly. However, >100 MGS could be assembled from multiple samples, hence, the total number of high-quality draft genome assemblies was 360.

Functional characterization of small CAGs

The 6,640 CAGs with <700 genes (an average of 44 genes) showed abundance correlations similar to those of the MGS and several lines of evidence suggest that many of these CAGs represent biological entities or clonal differences within species. Of the small CAGs, 848 were enriched for proteins characteristic of phages[13] or for genes with similarity to specific known phage taxa (**Supplementary Data 5**). These phage-like CAGs have a median coding sequence length of 28kb, which is substantially longer than previous phage assemblies from complex communities (in which ~60% are <1 kb)[13]. An average of 113 (±37, s.d.) phage-like CAGs were identified per sample. Although bacteriophage taxonomy is fairly limited, we observed consistent species- or family-level taxonomical annotation in 35 and 172 phage-like CAGs, respectively. As expected from the presence of phage-like CAGs, transposase, integrase and recombinase encoding genes were enriched in the smaller CAGs (**Fig. 2b**).

Another class of functions that were enriched in smaller CAGs are functions that are important for biotic interactions. These include clustered, regularly interspaced, short palindromic repeats (CRISPR)-associated genes, which function in bacteria and archaea as a sequence-dependent adaptive immune system against foreign nucleic acids[19]. In addition to core CRISPR-associated genes, several CAGs were enriched for specific subtypes of these genes (**Supplementary Fig. 13**).



Similarly, restriction endonucleases and DNA methylases, which are part of the nonadaptive defense system, were enriched in 120 small CAGs. Also, genes involved in modification of bacterial exterior surfaces, which are important for bacterial identification and masking, were enriched in more than 400 small CAGs. These included genes involved in modifications of the cell wall and, in particular, glycosyltransferases.

Dependency associations affiliate small CAGs to MGS

The existence of small CAGs that represent phages and clonal differences implies that such CAGs depend on cellular organisms for their proliferation. In relationships that are nonpromiscuous, a dependent CAG should therefore never occur independently of the hosting microorganism. Indeed, we identified significant dependency associations by comparing the absence-presence profiles throughout all samples for all pairs of CAGs (including the MGS) using Fisher's exact test and excluding relationships where a potential dependent CAG was observed independently of the potential hosting CAG (**Fig. 5a**). In this way the network becomes directional from the dependency-associated CAG to the hosting CAG (which may be observed independently).

Figure 4 Comparison of the MGS:337 augmented assembly and the *B. animalis* reference genome. BLAST dot-plot comparing the MGS:337 augmented assembly (*y* axis) to the *B. animalis subsp. lactis CNCM I-2494* reference genome (*x* axis). The dot-blot shows the relative chromosomal positions of matching sequence on the MGS-augmented assembly and the *B. animalis* reference genome. The MGS-augmented assembly covers 95% of the reference genome with 99.9% identity. The plot shows an inversion in the assembly relative to the reference genome around position 1,300 K.

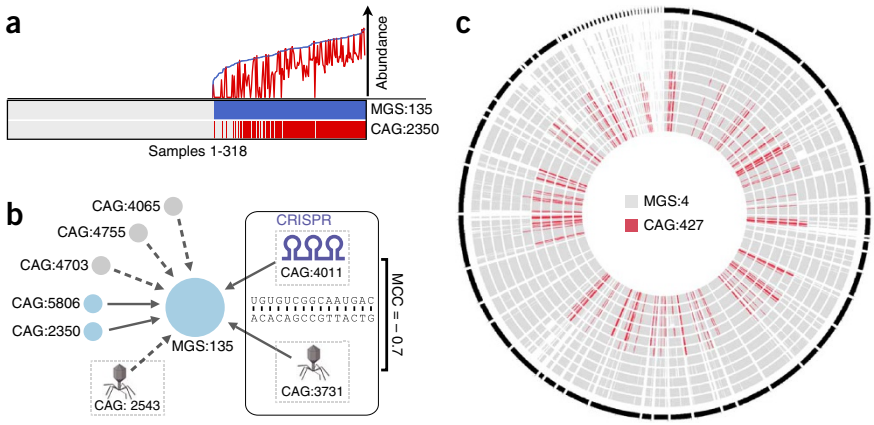


Figure 5 Dependency associations among MGS and CAGs. (**a**) A typical example of a significant dependency association. The abundance of the MGS:135 (*S. wadsworthensis*) and the small CAG:2350 across 318 fecal samples are shown as blue and red curves, respectively (upper panel, logarithmic scale). Below the sample-wise presence of the two CAGs is shown as bars. CAG:2350 is significantly co-occurring with MGS:135 and never detected independently (Fishers exact test, $P = 9 \times 10^{-74}$). The samples were sorted according to the abundance of MGS:135. (**b**) The dependency-association subnetwork of CAGs associated to *S. wadsworthensis* (MGS:135). Arrows show dependency associations and solid arrows indicate that co-assembly of the MGS and the CAG in one or more samples supported the association. Blue coloring indicates CAGs dominated by genes with species level similarity to *S. wadsworthensis*. CAG:2543 and CAG:3731 are enriched for phage genes, and CAG:4011 contains a series of CRISPR-associated genes and a CRISPR cluster. The CRISPR complex containing CAG:4011 and one of the phages-like CAG:3731 anti-correlate (Matthews correlation coefficient = –0.7) and spacers of the CRISPR show sequence complementarity to the phage. (**c**) The *E. coli* (MGS:4) and its nine dependency-associated CAGs were co-assembled to high-quality draft genomes in each of 11 samples. The outer black circle represents the consensus assembly of the *E. coli*–centered agglomerate and each of the gray circles represents alignment of the assembly from a particular sample. The positions and sequence coverage of CAG:427 are marked in red, across the assemblies.

The resulting network of the most significant (Fishers exact test: $P < 10^{-10}$, corrected) dependency associations is shown in **Supplementary Figure 14** and contains 882 relationships between 1,205 CAGs (**Supplementary Data 6**). As expected, the network is significantly over-represented for small CAGs that associate to an MGS (odds ratio 12.7, Fisher's exact test: $P << 1 \times 10^{-100}$) and many of the subnetworks are centered on an MGS, but there are also nine small CAGs that connect MGS pairs of the same genus. Biologically these may be genetic elements or phages that are shared between related species.

For 413 of the associations, sequence contigs were found in individual samples that bridged the dependent and hosting CAGs (enrichment relative to all CAGs pairs: odds ratio 2,513, Fisher's exact test $P << 1 \times 10^{-100}$). This indicates that many dependency-associated CAGs are genomically integrated in the hosting CAG in some samples.

The dependency associations connect CAGs into subnetworks, which can be used to guide further study of the components. For example, the subnetwork centered on *Sutterella wadsworthensis* (MGS:135, **Fig. 5b**) contains eight dependency associations including the phage-like CAG:3731 and a CAG containing CRISPR-associated genes and a repeat region (CAG:4011). The sample-wise detection of the CRISPR and phage-like CAGs were anti-correlated (Matthew's correlation coefficient –0.7) and one of the CRISPR spacers had a 15-bp sequence match to the phage, suggesting that the CRISPR prevents the phage from infecting the bacterium[19].

Sample-specific, MGS-augmented assemblies of the *Escherichia coli* MGS:4 and its dependency-associated CAGs (**Supplementary Fig. 15** and **Supplementary Data 7**) demonstrated strong sequence similarities throughout the majority of the chromosome, but had some differences. In particular the largest of the *E. coli*–associated CAGs (CAG:427, containing 345 genes) was absent in some sample-specific assemblies (**Fig. 5c**). In the MGS:4 genome the integration of CAG:427 was spread across several locations, suggesting that it represents several independent insertion and/or rearrangement events.

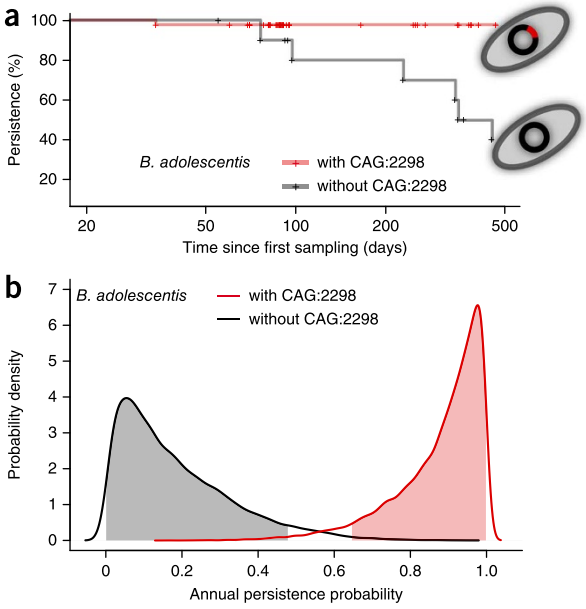


Figure 6 Gut persistence probability for *B. adolescentis*. The gut persistence of *B. adolescentis* (MGS:119) populations stratified by the presence (red curves) or absence (black curves) of the dependency-associated CAG:2298 observed across 54 human individuals who had the bacterium in the first of two fecal samples. *B. adolescentis* had substantially higher persistence probability with CAG:2298 present. (**a**) Interval-censored Kaplan-Meier curves showing the cumulative loss of populations of *B. adolescentis* over time across the cohort of human individuals. Points (+) indicate time (in days) of the second of two samplings from a human individual. The curve shows the "losses" when they are registered at the second time point and not when the loss actually happened (i.e., the data are interval-censored). (**b**) Model-based estimates of annual gut persistence probability for *B. adolescentis* with or without the dependency-associated CAG. Note that annual persistence probability with the CAG (mean estimate = 88%) is much larger than without (mean estimate = 18%). In the Bayesian logistic regression framework used here, estimates are expressed as probability distributions over the possible values for parameters of interest. We therefore obtain both an estimate of a parameter, and quantification of how certain we are of the estimate. This figure shows the posterior probability distribution over possible values for the annual persistence probabilities; with the shaded areas indicating the 95% highest-density intervals (i.e., the parameter values with the most support).

MGS persistence is influenced by associated CAGs

To investigate the effect that dependency-associated CAGs may have on their host MGS, we analyzed 73 of the individuals that were sampled at two different time points (four of the original 77 sample pairs were discarded). From each of these sample pairs we assessed if a given MGS was present at the first time point and whether it was still present at the second time point. Based on this information, it was possible to estimate the persistence of an MGS across the cohort of human individuals, and whether this persistence was influenced by dependency-associated CAGs.

We found that there was a greater probability of *B. adolescentis* (MGS:119) persisting at the second time point when its dependency-associated CAG:2298 was also present (**Fig. 6a**). To further analyze this phenomenon, we used logistic regression to infer the annual persistence probabilities for MGS with or without their dependency-associated CAGs (**Supplementary Data 6**). The credibility of these estimates was quantified using Bayesian statistical methods[20], which, in brief, output a posterior probability distribution over the possible annual persistence probabilities. From this analysis we identified 26 cases in which the presence of a specific dependency-associated CAG correlated with a substantially altered annual persistence probability of its host MGS. For example, the annual persistence probability of *B. adolescentis* (MGS:119) was estimated to be 88% in individuals where it was observed in association with CAG:2298, but 18% in individuals where CAG:2298 was absent (**Fig. 6b**; posterior probability that the CAG:2298 effect is larger than zero = 99.94%). We observed similar, positive effects for CAGs dependency-associated with *Prevotella copri*, *E. coli*, *F. prausnitzii* and 12 other MGS (**Supplementary Fig. 15** and **Supplementary Data 6**). In addition, ten dependency-associated CAGs had a substantial negative effect on the persistence probability of their hosting MGS.

The dependency-associated CAGs that increased the MGS persistence probability contained a range of gene sets, including CRISPR-associated genes (CAG:2720), collagen adhesion protein and gram-positive anchor proteins (CAG:2888), and thioredoxin family proteins that might be important for the tolerance of reactive oxygen species (ROS). This is in line with our observation that the most common species in the human gut microbiome have genes that mediate ROS tolerance (**Supplementary Note 6** and **Supplementary Data 8**). Among the dependency-associated CAGs that contributed negatively to the MGS persistence probability, we observed three phage-like CAGs.

DISCUSSION

The method presented here allows complete co-abundance clustering of microbiomes. The resulting CAGs provide insight into the microbial species present in metagenomic samples and their genetic makeup, and thereby provide details important for understanding the content of a microbial community. Clustering is purely data driven and therefore circumvents the need for reference genomes, presequencing filtering and cultivation of microbial species. The ability of our method to discriminate between strains of the same species indicates the power of co-abundance to segregate closely related biological entities, in contrast to the findings from gene sets that are defined by sequence similarity to known reference genomes (**Supplementary Fig. 8**). Inaccurate discrimination among closely related species potentially leads to false associations between clinical conditions and putative species. Although we identify a few cases of chimeric assemblies (**Supplementary Note 7**), we have no indication of CAGs constituting multiple species; however, such entities could in principle exist in very close co-abundance.

The method should be generally applicable to sets of deep-sequenced shotgun metagenomics samples. The exact number of samples and sequencing depth needed depends on the complexity of the microbial community and the abundance of the microbes. However, our benchmarking using *B. animalis* suggests that the number of samples used is critical (here 18 samples), whereas the necessary sequencing depth (here 0.7 M reads) is easily reached with current sequencing technologies (**Fig. 3**). This emphasis on sample number over sequencing depth is likely to hold across different types of microbial communities.

Interestingly, we found that most genes involved in resistance to antibiotics, except vancomycin resistance genes, were not found as members of any CAG. This is in line with the fact that most antibiotic genes, except vancomycin resistance genes, are known to act alone to provide antibiotic resistance (**Supplementary Note 8**).

Although, many of the CAGs and their dependency association are not understood at present, our findings suggest that even small CAGs represent biologically meaningful entities, either in the form of phages or clonal differences of microbial species. This is consistent with previous findings that the genetic differences that make the *E. coli* O104:H4 strain a cause of severe food poisoning are just a few virulence factors, including a Shiga toxin 2–encoding prophage[21]. Therefore, thorough descriptions of genetic heterogeneity, which are enabled by our co-abundance method, are likely to be important for disease association studies. Moreover, although the relationships we observed between specific smaller CAGs and their host MGS are only associations, they do suggest functionally and/or evolutionarily important relationships that may prove critical for future understanding and engineering of microbial communities. Furthermore the approach that allowed us to determine conditional persistence probabilities for microbial species in a community is likely to be important for revealing relationships or conditions that are critical for the persistence or elimination of specific microbes in future studies. Here the functions of genes within the CAGs associated with differential persistence suggest that tolerating ROS and anchoring to the intestinal epithelia are important for microbial persistence in the gut. Our findings also suggest specific phage-microbe relationships that reduce persistence. Therefore, our co-abundance–based method should facilitate advances in understanding microbial biology as well as enabling *de novo* genome assembly and comprehensive characterization of complex metagenomic samples.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data were deposited at EBI with the accession code ERP002061; and MGS-augmented assemblies were deposited at EBI under (PRJEB674 to PRJEB1046). 454 sequencing reads were added to the NCBI BioProjectID 32811. The 3.9 M gene catalog and the CAGs are available for download from https://www.cbs.dtu.dk/projects/CAG/. Source code for the MGS canopy algorithm is available as **Supplementary Software** and from: http://git.dworzynski.eu/mgs-canopy-algorithm.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7- HEALTH-F4-2007-201052: Metagenomics of the Human Intestinal Tract (MetaHIT) and FP7-HEALTH-2010-261376: International Human Microbiome Standards, as well as the Novo Nordisk Foundation Center for Biosustainability. Work on the clustering concept has been supported by the OpenGPU FUI collaborative research projects, with funding

from DGCIS. Researchers on the project were granted access to the HPC resources of CCRT under the allocation 2011-036707 made by GENCI (Grand Equipement National de Calcul Intensif). The company Alliance Services Plus (AS+) has provided help to scale up the process, especially, V. Arslan, D. Tello, V. Ducrot, T. Saidani and S. Monot. The authors affiliated with MGP are funded, in part, by the Metagenopolis ANR-11-DPBS-0001 grant. Ciberehd is funded by the Instituto de Salud Carlos III (Spain). M.A. was supported by a grant from the Ministère de la Recherche et de l'Education Nationale (France).

AUTHOR CONTRIBUTIONS

All authors are members of the Metagenomics of the Human Intestinal Tract (MetaHIT) Consortium. S.D.E. and S.B. managed the project. F.C., N.B., F.G., T.H., K.S.B. and T.N. performed clinical sampling. F.L. and C.M. performed DNA extraction. J.L., E.P. and D.L.P. performed sequencing. S.D.E., H.B.N., M.A., A.S.J., S.R., P.R. and P.B. designed the analyses. H.B.N., A.S.J., S.R., M.A., A.G.P., D.R.P., L.G., I.B., M.B., M.B.Q.d.S., M.A., J.L., J.T., S.S., T.Y., E.P., D.L.P. and R.S.K. performed the data analyses. H.B.N., S.B., A.S.J., S.R., A.G.P. and M.A. wrote the manuscript. H.B.N., S.B., S.D.E., D.R.P., I.B., P.B., E.P., O.P. and D.W.U. revised the manuscript. The MetaHIT Consortium members contributed to the design and execution of the study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Fodor, A.A. *et al.* The "most wanted" taxa from the human microbiome for whole genome sequencing. *PLoS ONE* **7**, e41294 (2012).
2. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
3. Lukjancenko, O., Wassenaar, T.M. & Ussery, D.W. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.* **60**, 708–720 (2010).
4. Fitzsimons, M.S. *et al.* Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. *Genome Res.* **23**, 878–888 (2013).
5. Pop, M. Genome assembly reborn: recent computational challenges. *Brief. Bioinform.* **10**, 354–366 (2009).
6. Wooley, J.C., Godzik, A. & Friedberg, I. A primer on metagenomics. *PLOS Comput. Biol.* **6**, e1000667 (2010).
7. Iverson, V. *et al.* Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* **335**, 587–590 (2012).
8. Wang, Y., Leung, H.C.M., Yiu, S.M. & Chin, F.Y.L. MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics* **28**, i356–i362 (2012).
9. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
10. Raes, J. & Bork, P. Molecular eco-systems biology: towards an understanding of community function. *Nat. Rev. Microbiol.* **6**, 693–699 (2008).
11. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
12. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
13. Minot, S. *et al.* The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625 (2011).
14. Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* **22**, 1985–1994 (2012).
15. Zhang, Q., Rho, M., Tang, H., Doak, T.G. & Ye, Y. CRISPR-Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. *Genome Biol.* **14**, R40 (2013).
16. Chain, P.S.G. *et al.* Genomics. Genome project standards in a new era of sequencing. *Science* **326**, 236–237 (2009).
17. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
18. Chervaux, C. *et al.* Genome sequence of the probiotic strain Bifidobacterium animalis subsp. lactis CNCM I-2494. *J. Bacteriol.* **193**, 5560–5561 (2011).
19. Terns, M.P. & Terns, R.M. CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol.* **14**, 321–327 (2011).
20. Kruschke, J.K. Bayesian data analysis. *Wiley Interdiscip. Rev. Cogn. Sci.* **1**, 658–676 (2010).
21. Karch, H. *et al.* The enemy within us: lessons from the 2011 European *Escherichia coli* O104:H4 outbreak. *EMBO Mol. Med.* **4**, 841–848 (2012).

MetaHIT Consortium:

H Bjørn Nielsen[1,2,32], Mathieu Almeida[3–5,32], Agnieszka S Juncker[1,2], Simon Rasmussen[1], Junhua Li[6–8], Shinichi Sunagawa[9], Damian R Plichta[1], Laurent Gautier[1], Anders G Pedersen[1], Emmanuelle Le Chatelier[3,4], Eric Pelletier[10–12], Ida Bonde[1,2], Trine Nielsen[13], Chaysavanh Manichanh[14], Manimozhiyan Arumugam[9], Jean-Michel Batto[3,4], Marcelo B. Quintanilha dos Santos[1], Nikolaj Blom[2], Natalia Borruel[14], Kristoffer S Burgdorf[13], Fouad Boumezbeur[3,4], Francesc Casellas[14], Joël Doré[3,4], Piotr Dworzynski[1], Francisco Guarner[14], Torben Hansen[13,15], Falk Hildebrand[16,17], Rolf S Kaas[18], Sean Kennedy[3,4], Karsten Kristiansen[19], Jens Roat Kultima[9], Pierre Leonard[3], Florence Levenez[3,4], Ole Lund[1], Bouziane Moumen[3,4], Denis Le Paslier[10–12], Nicolas Pons[3,4], Oluf Pedersen[13,20–22], Edi Prifti[3,4], Junjie Qin[6,7], Jeroen Raes[17,23,24], Søren Sørensen[25], Julien Tap[9], Sebastian Tims[26], David W Ussery[1], Takuji Yamada[9,27], Pierre Renault[3], Thomas Sicheritz-Ponten[1,2], Peer Bork[9,29], Jun Wang[7,13,19,30], Søren Brunak[1,2], S Dusko Ehrlich[3,4,31], Alexandre Jamet[3], Alexandre Mérieux[33], Antonella Cultrone[3], Antonio Torrejon[14], Benoit Quinquis[4], Christian Brechot[33], Christine Delorme[3], Christine M'Rini[33], Willem M de Vos[26], Emmanuelle Maguin[3], Encarna Varela[14], Eric Guedon[3], Falony Gwen[16], Florence Haimet[4], François Artiguenave[10], Gaetana Vandemeulebrouck[3], Gérard Denariaz[34], Ghalia Khaci[3], Hervé Blottière[4], Jan Knol[35], Jean Weissenbach[10], Johan E T van Hylckama Vlieg[34], Jørgensen Torben[26], Julian Parkhill[36], Keith Turner[36], Maarten van de Guchte[3], Maria Antolin[14], Maria Rescigno[37], Michiel Kleerebezem[26], Muriel Derrien[34], Nathalie Galleron[4], Nicolas Sanchez[3], Niels Grarup[21], Patrick Veiga[34], Raish Oozeer[35], Rozenn Dervyn[3], Séverine Layec[3], Thomas Bruls[10], Yohanan Winogradski[3] & Zoetendal Erwin G[26]

[33]Institut Mérieux, Lyon, France. [34]Danone Research, Palaiseau, France. [35]Gut Biology & Microbiology, Danone Research, Center for Specialized Nutrition, Wageningen, the Netherlands. [36]The Wellcome Trust Sanger Institute, Hinxton, Cambridge, U.K. [37]Istituto Europeo di Oncologia, Milan, Italy.

ONLINE METHODS

Sample description. 396 stool samples from 177 Danish and 141 Spanish human individuals were collected (**Supplementary Data 1**). 124 of the samples were sequenced and used previously[2]. The Spanish samples included 13 individuals with Crohn's disease and 69 with ulcerative colitis. 77 of the Spanish individuals were sampled twice with, on average, 6 months between the samplings. The Danish samples include healthy individuals ranging in body mass index from 18 to 42. All were subjected to Illumina deep sequencing resulting in 4.5 Gb sequence per sample on average, and a total of 23.2 billion high-quality sequencing reads with an average length of 77 bp. The study was approved by the local Ethical Committees of the Capital Region of Denmark (HC-2008-017) and Clinical Research Ethics Committees (Comités de Ética en Investigación Clínica, CEIC, Spain) and informed consent was obtained.

Construction of a nonredundant metagenomic gene catalog. Illumina raw sequencing reads from 396 metagenomic samples (**Supplementary Data 1**) were processed using the MOCAT software package[22]. In brief, >23.2 billion raw sequencing reads were filtered using the FastX software (http://hannonlab.cshl.edu/fastx_toolkit) with a quality cutoff of 20 and reads shorter than 30 bp discarded. High-quality reads (92% of raw reads) were assembled into scaftigs (i.e., continuous sequences within scaffolds) using SOAPdenovo (version 1.05)[23]. Genes were predicted on 18.5 M scaftigs longer than 500 bp (35 Gbp in total) using MetaGeneMark[24]. Predicted genes from all samples (45.4 M in total) were clustered using BLAT by single linkage. Any two genes with greater than 95% identity and covering more than 90% of the shorter gene were clustered together. Finally, cluster representatives shorter than 100 bp were discarded resulting in a set of 4,201,877 nonredundant genes. From this set, we removed genes that were considered spurious or likely originated from human, animals or plants to yield a final set of 3,871,657 genes that formed the reference gene catalog. For a comparison to our previous gene catalog[12], see **Supplementary Data 9**.

Quantification of reference gene abundances. High-quality reads were mapped to the reference gene catalog using the screen function in MOCAT[22]. Briefly, reads were mapped with SOAPaligner (version 2.21)[25] with options: –M 4 (find best hits), –l 30 (seed length), –r 1 (random assignment of multiple hits) and –v 5 (maximum number of mismatches). Mapped reads were subsequently filtered using a 30-bp length and 95% identity cutoff and gene-length normalized base counts were calculated using the soap.coverage script (available at: http://soap.genomics.org.cn/down/soap.coverage.tar.gz). For samples where 11 M or more sequence reads were obtained (n = 393), 11 M sequence reads were drawn randomly (without replacement). These randomly drawn reads were mapped to the gene catalog and the number of reads counted to form a downsized depth or abundance matrix. The 11 M downsized depth matrix was used to estimate CAG abundances, gene and MGS richness. Similar downsizings were done for the reduced sampling depths (**Fig. 3**).

Taxonomical annotation. Catalog genes were assigned taxonomical annotation by sequence similarity to a database of 3,048 reference genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/ and ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria_DRAFT/, July 2012), using BLASTN, only accepting alignments with 100 bp or longer. Sequence similarity of 95%, 85% and 75% or better was used for species, genus and phylum level taxonomical annotation, respectively. MGS were assigned a species level annotation if more than 50% of the genes comprising the CAG were assigned a given species level taxonomy (including genes with no match). MGS were described to have 'clear and unambiguous similarity to a known species' when 90% or more of the genes were annotated to the same species. Selected CAGs that appear in figures and could not be assigned to a genus or species by DNA similarity (MGS:11, MGS:17, MGS:124 and MGS:225) were in addition taxonomically annotated by similarity to the UniProt database (BLASTP, best hit, E < 0.001) to get an approximate taxonomical annotation.

Phage definition and taxonomy annotation. A CAG was called phage-like if it passed one of two criteria. (i) If a CAG contained a minimum of ten phage-taxonomy annotated genes and 80% of these were consistent at the species, genus or family level. Here phage-taxonomy annotated genes were defined as genes with a top-3 blastp hit (E < 0.001, against the combined NCBI nr Sept. 2013 and ACLAME[26] 0.4 database) to a viral organism listed in the International Committee on Taxonomy of Viruses (ICTV) master species list (release 2012). (ii) If a CAG-encoded five or more distinct characteristic phage functions and ≥40% of the CAG genes were most similar to known phage genes. Phage-functional classes were defined as proteins with a best-hit (hmmscan[27], domE < 0.001, against Pfam-A[28] 27.0) to one of 16 phage-specific Pfam functions defined by Minot *et al.*[13] or as proteins matching the corresponding set of functions identified among phage orthologous groups (blastp, E < 0.001, against POG VQ[29]). A characteristic phage function was counted only once per CAG. Furthermore, a gene most similar to known phage genes was defined as a gene with a best hit (blastp, E < 0.001, against the combined NCBI nr and the ACLAME 0.4 database) to a viral organism. All phage-like CAGs were taxonomically annotated to species, genus or family level using a 50% consistency criteria across ICTV annotated genes (top-3 blastp hits, E < 0.001, against the combined NCBI nr and ACLAME[26] database). Interestingly, the functions "tail," "portal," "terminase" and "capsid" were each found in ≥70% of all phage-like CAGs and on average in only 5% of other small CAGs.

Gene annotations and enrichment analysis. Functional annotation (including CRISPR-associated genes) of the gene catalog was obtained by aligning predicted proteins to the UniProt database using BLASTP (best hit with E < 0.001) and proteins from the eggNOG (v3) database[30] using BLASTP (WU-BLAST 2.0, default parameters except $E = 1 \times 10^{-5}$, B = 10,000) and were assigned to an orthologous group as described elsewhere[31].

Genes of MGS:11, CAG:4957, MGS:17 and MGS:124 (appearing in **Supplementary Fig. 16c**) was aligned to proteins listed by Roessner *et al.*[32] as experimentally verified and strictly anaerobe corrin ring biosynthesis proteins (60 coverage, 40% identity). CRISPR repeat-spacer segments were identified with CRISPR-recognition tool (CRT, ver. 1.2)[33] in selected CAG assemblies. Genes were annotated as virulence or antibiotic resistance genes when BLASTP alignments exceeded 80% identity over 80% of the length of protein in the Virulence Factor Database (VFDB, February 2012 version) or ResFinder[34] (version 1.2) database, respectively.

From 271 essential genes from the genome of *Bacillus subtilis* strain 168 (ref. 35), 252 clusters of orthologous genes (COGs) were deduced (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Bacillus_subtilis_168_uid57675/NC_000964.ptt manually curated, see **Supplementary Data 10**). Genes aligning to these COGs were termed essential genes.

CAGs significantly enriched for a specific annotation were identified using Fisher's exact test (P < 0.001 for **Fig. 2b**). Significant biases in eggNOG[30] annotation, as a function of the MGS observation frequency across the samples, were identified using Wilcoxon rank sum test ($P < 1 \times 10^{-15}$; **Supplementary Data 8**).

Co-abundance clustering. The canopy-based clustering of the gene catalog was performed by iteratively picking a seed gene among the not yet clustered genes and aggregate genes with abundance profiles within a fixed distance from the seed gene abundance profile (Pearson correlation coefficient > 0.9 and Spearman's rank correlation coefficient > 0.6) into the seed canopy. Canopies with median abundance profiles within a distance of 0.97 PCC from one another were merged. Canopies with 2 or less genes (such canopies included a total of 1.7 M genes), or for which the canopy abundance signal from any three samples constituted 90% or more of the total signal across all samples, for which the median profile was detected in less than four samples, or for which one sample made up 90% of the total signal (1.1 M genes) were discarded for having insufficient supporting evidence (based on Monte Carlo simulation; **Supplementary Fig. 9**). Canopies that passed these criteria were called CAGs. CAGs with more than 700 genes are also referred to as MGS or just species. Note that the number of clusters was not predefined for the canopy-based clustering. CAG abundance profiles were calculated as the sample-wise median gene depth signal (downsized). A CAG was considered observed in a sample when its abundance profile exceeded zero in that sample.

Source code for the canopy algorithm is freely available as **Supplementary Software** and at http://git.dworzynski.eu/mgs-canopy-algorithm.

MGS-augmented assembly. For each of the 741 MGS we performed a *de novo* MGS-augmented assembly, using the subset of sequence reads that mapped to the contigs from which the MGS genes originated. For each MGS, we performed independent and sample-specific augmented assemblies on the two samples from which most sequence reads mapped to the MGS and the sample from which most of the MGS gene containing *de novo* contigs were derived. For a given sample, the reads were aligned using Burrows-Wheeler Aligner[36] (bwa-0.5.9) to the MGS specific scaffolds and the mapped reads, including unmapped mates, were extracted. These reads were then corrected by Quake[37] using $k = 15$. The reads were then *de novo* assembled with Velvet (1.2.01) using *k*-mers from 21 to 45 and the parameters '-cov_cutoff auto' and '-exp_cov auto'. As several samples were used for assembly of each MGS, the best assembly was selected based on ranking of contig N50 and the number of contigs in the assemblies[38]. Contigs with read depth of less than half the average depth of all contigs were removed from the assemblies[38,39]. The contigs and scaffolds were then filtered to 100- and 500-bp minimum lengths, respectively, and gaps in scaffolds were filled using SOAPdenovo GapCloser (1.10).

Assembly statistics. General assembly statistics were calculated using assemblathon_stats.pl[40] and coverage was calculated by aligning reads to the contigs using bwa (0.5.9)[36] and BEDtools. To assess the quality of the assemblies, we adopted the six high-quality draft assembly criteria from the Human Microbiome Project (HMP)[16]. Five of these criteria address the contiguity of the assembly, and one criterion, genome completeness, by counting core genes contained in the assembly. The criteria are (i) 90% of the genome assembly must be included in contigs >500 bp, (ii) 90% of the assembled bases must be at > 5× read coverage, (iii) The contig N50 must be >5 kb, (iv) scaffold N50 must be >20 kb, (v) average contig length must be >5 kb and (vi) >90% of the core genes must be present in the assembly. The core gene ratios were determined using HMP standard operating procedure for both bacteria and archaea. In short, blastx was used to identify core genes from the scaffolds and proteins with at least 30% identity and 30% coverage for Bacteria and 50% identity and 70% coverage for Archaea were considered a core gene hit. The ratio of core genes identified was then calculated using get_coregroups_coverage.pl (HMP tools and protocols). In total 360 sample-specific MGS-augmented assemblies from 247 unique MGS passed all six criteria (**Supplementary Data 3**). In addition, 139 unique assemblies passed five criteria.

We determined the number of novel species by aligning all proteins to Uniprot using blastp and converted taxids from strain to species level using NCBI taxonomy. An assembly was considered previously unsequenced if less than 10% of the genes could be aligned with a minimum of 95% identity over 33aa to genes from a species. 181 of the 238 high-quality assembled draft genomes plus 83 assemblies passing 5 criteria were identified as novel species.

Screening for chimeric assemblies. Because the HMP criteria were created for single genome assembly, we applied three additional metrics to account for putative chimeric assemblies arising from metagenomic data, (i) uniformity of the contig read depth distribution, (ii) identification of multiple copies of conserved 40 COGs[30] and (iii) inter-assembly tetranucleotide frequency (TNF) consistency.

Because assemblies consisting of genomic regions from different organisms are likely to have multimodal coverage distributions, we performed peak detection on the contig read coverage distributions for all assemblies passing 4-6 HMP criteria and assemblies with more than 1 peak were manually inspected. From the presence of multiple copies of COGs, we were able to identify three assemblies as chimeric. Of the 247 unique high-quality draft assemblies, 9 (3.6%) were identified as potentially chimeric, and for the additional 139 assemblies that passed five criteria, we identified 3 potential chimeric assemblies (2.3%) and 1 without any core genes (MGS:3246). The remaining assemblies have been deposited at the European Nucleotide Archive (ENA).

Furthermore, tetranucleotide frequencies *z*-scores were calculated for all assemblies and HMP reference assemblies as described by Teeling *et al.*[41]. For each assembly the frequencies were calculated in windows of 5 kb to avoid biases introduced by different scaffold lengths. If a scaffold was shorter than the window size it was still included in the calculations. Within each assembly a median tetranucleotide frequency *z*-profile was created and the tetranucleotide frequency *z*-scores of each 5-kb window were correlated to this median profile using PCC. The resulting high-quality draft genomes showed comparable TNF correlations to the single organism HMP reference genomes indicating a low rate of chimeric assemblies (**Supplementary Fig. 12**).

Comparison of MGS-augmented assemblies and reference genomes. To estimate the completion level of the MGS-augmented assemblies, 299 draft reference genomes from the human intestinal tract HMP DACC database and the NCBI collection of complete reference genomes (both version updated from 2012/04) were used as a reference set for a blast comparison procedure. 44 of the assemblies that passed 5 or more of the 6 HMP criteria (including the bacteria/archaea core ratio criteria) were similar to a reference genome. The contigs and scaffolds of these assemblies were projected on their closest reference genomes using the GAGE pipeline for assembly quality evaluation[42]. First *nucmer* (default parameters) was used to align the contigs/scaffolds to the reference genome. Then *delta-filter* was used to remove low identity match (parameters: -I 95, -o 80). Finally *dnadiff* was used to compare the assemblies and the closest reference genome and estimate the mean identity and coverage of each contigs and scaffolds (**Supplementary Data 4**). Additionally, the MGS:337 assembly, which did not meet the six criteria, was 99.9% identical to *B. animalis subsp. lactis CNCM I-2494* (ref. 18) and covered 95% of this reference genome (**Fig. 4**).

To search for potential contaminants, unaligned scaffold fragments were blasted to the complete reference genome set, and the best hit (with identity and coverage threshold of ≥95% and ≥80%, respectively) was extracted. Scaffolds that matched to a different genus were considered potential contaminants. Of the 44 MGS-augmented assemblies, only 16 contained any scaffolds with similarity to an alternative genus. In general these scaffolds were small with an average size of only 2,721 bp. If we consider unaligned scaffold with similarity to an alternative genus as a potential contaminant, the mean contamination rate was estimated to 1.00 scaffold per HQ assembly.

MGS-augmented assembly gap closure using Sanger sequence data. To further experimentally validate the coherence of the sample-specific MGS-augmented assemblies, we used Sanger sequence data from eight samples[2]. Fecal microbial DNA from those individuals was used to construct plasmid-based (pCNS) clone libraries of 3 kb long inserts, containing 250,000 clones each. Clone insert ends were sequenced using the Sanger technology (ABI3730XL). Sequences were subsequently subjected to vector cleaning and quality trimming, generating on average 230,468 (±5,145) reads per sample. The same DNA was used for pyrosequencing (454GSFlx-Titanium), resulting in 2,362,978 reads on average per sample (±3,245,603). For each reference subject, Sanger and 454 reads as well as Velvet contigs generated from Illumina sequencing of the same DNA were combined for assembly using the 454-Newbler software (v2.3). CAGs detected in a given reference subject were compared with Sanger reads from that individual using blastn. High-scoring segment pair (HSPs) covering at least 90% of the length of the smallest read or velvet contig with at least 90% identity were selected, and corresponding reads extracted. Scaffolding of the CAG contigs with paired Sanger reads was then achieved using the bambus software[43]. Only assemblies with >1× coverage were kept, and used to assess the rate of gap closure (**Supplementary Data 3**). On average 64% of the assembly gaps were closed, and in particular, the MGS:710 assembly was closed to only 3 scaffolds from an initial 32 scaffolds.

Phylogeny of the MGS assemblies. We used all nonchimeric assemblies passing 5 and 6 HMP criteria (139 and 247 assemblies, respectively) and 296 HMP gut microbiome reference genomes (HMPDACC) and 1,506 reference genomes to construct a phylogeny based on 40 phylogenetic marker proteins (COGs)[44]. For each assembly, proteins were predicted using Prodigal and aligned using blastp to the individual COG proteins, and the best hits were selected requiring at least 50% id over 50% of the COG sequence. For each COG the MGS assembly and reference proteins were aligned using *muscle* and here joined to a single alignment file for each COG using *muscle-profile*. The 40 individual protein alignments were concatenated to a single alignment for each reference

doi:10.1038/nbt.2939

genome/MGS assembly and alignments containing less than 35 COGs were removed from further analysis, resulting in 337 MGS assemblies for the final tree. The phylogenetic tree was constructed with FastTree using the JTT substitution matrix with the parameters "-gamma -pseudo -spr 4 -mlacc 3 -slownni" and visualized using ITOL[45].

Co-assembly of *E. coli* and dependency-associated CAGs. A pool of the main *E. coli* (MGS:4) and its nine dependency-associated CAGs (CAG:427, CAG:1345, CAG:2136, CAG:2318, CAG:2530, CAG:2610, CAG:3070, CAG:3196 and CAG:5108) were used for recruiting 1,708 contigs for a pooled assembly, across 247 selected samples. Subsequently, *de novo* assembly (as described above) from 13 of these samples passed five or more HMP criteria (**Supplementary Data 7**). A consensus assembly was generated from the contigs of these assemblies using minimus2, where each assembly was joined to the consensus in separate steps[46]. The consensus assembly contained 4.3 Mb sequences in 45 contigs with a contig N50 of 129 kb. Subsequently all the individual assemblies were aligned with blastn to the consensus assembly, and contigs without a significant hit were pooled and clustered using cd-hit-est with the parameters "-c 0.8 -n 7"[47]. To further reduce redundancy of the extra contigs they were cut into 500-bp 'reads' with 250-bp overlap and reassembled using Newbler 2.6. The resulting 157 contigs were then added to the consensus assembly obtained from minimus2 to a final assembly of 4.91 Mb in 202 contigs.

Dependency associations. A CAG was considered dependency associated on another CAG when the sample-wise overlapping detections of the CAG pair were statistically significantly over-represented (Fisher's exact test, upper tail, Bonferroni corrected $P < 1 \times 10^{-10}$) and the dependency-associated CAG was not detected independently.

Smaller CAGs (<700 genes) were considered 'co-existence associated' when their detections were significantly enriched (Fisher's exact test, Bonferroni corrected $P < 0.05$) in samples where an *MGS* pair was co-observed, and never occurred independently of one of the two MGS (the host). Here an MGS pair consisted of a host MGS and a companion MGS. An MGS was considered a potential companion if it co-existed with a potential host species in samples from ten or more individuals and was found independently of the host species in samples from ten or more human individuals. For the co-existence-associated relationships where the small CAG was not observed independently of any of the two MGS, the host species were determined as the MGS with the strongest abundance correlation to the small CAG across samples where both were detected, and by the sample specific co-assembly. No inconsistency between these measures was found.

Dependency-associated small CAGs were considered significantly absent in samples where a specific companion species was found, when it was significantly enriched in samples where the companion species was absent compared to samples where the host MGS was found (Fisher's exact test, Bonferroni corrected $P < 0.05$). Furthermore, the small CAG could never be observed independent of the hosting species. Again, an MGS was considered as a potential companion species if it co-existed with a host species in samples from ten or more individuals.

For all types of dependency associations a CAG was considered detected in a sample if the CAG abundance profile exceeded zero. Furthermore, only CAGs detected in ≥10 and ≤308 samples were considered. To ensure independence between the observations only one sample per individual was used ($n = 318$). Dependency-associated, 'co-existence–associated' and 'co-existence–absent' CAGs showed correlation to the species richness comparable to that of all CAGs.

Estimation of CAG persistence probabilities. Data from 73 human individuals, which were sampled twice, were used to estimate the annual persistence probabilities of MGS with or without dependency-associated CAGs (**Fig. 6**, **Supplementary Fig. 17** and **Supplementary Data 6**). All of the 2 × 73 stool samples included in this analysis resulted in at least 11 M sequence reads, and samples yielding more than this were downsized to 11 M reads. Furthermore, all included sample pairs exhibited strong stability between the samplings, in that they were more similar to each other than to 99% of the other samples in the cohort (using the Spearman correlation coefficient of the MGS abundances as similarity measure). Four of the original 77 sample pairs did not pass these criteria.

The main idea in this analysis was the following: for a fraction of the 73 sample pairs, a given MGS is present in the sample obtained at time point 1. For a subset of these sample pairs, the same MGS was also present at time point 2. Based on these, data logistic regression can be used to estimate an annual persistence probability for the MGS. The predictor variable (time between two consecutive samples) is continuous, whereas the outcome variable (presence or absence of an MGS) is binary. Logistic regression is used to estimate how the probability of an MGS still being present depends on the amount of time passed.

This computation is based on the assumption that an MGS has a typical probability per time unit of persisting in the gut of an individual. Thus the likelihood of observing an MGS at time point 2 is expected to be smaller the more time that has passed between the two samplings. Specifically, this decline is assumed to be exponential; thus if the probability that a given MGS will persist for a year is $P(1) = 0.7$, then the probability that it will persist for two years is $P(2) = 0.7^2 = 0.49$, etc. This assumption seems to fit well with Kaplan-Meier curves constructed from these data (**Fig. 6a**). Of course, the persistence of a given MGS in any individual is likely to depend on the specific conditions in that individual. We simply assume that there is a typical overall annual persistence probability associated with the MGS (on average, a given MGS has a typical tendency to persist in the gut of any individual), and real data will be scattered around this average according to unidentified covariates and stochastic effects.

Annual persistence probabilities were estimated in a probabilistic (Bayesian) model–based framework that explicitly accounts for time dependence. In this approach we assume that the annual persistence probability for an MGS is determined by the inherent resilience of the MGS itself, in combination with possible additional effects (positive or negative) caused by a set of dependency-associated CAGs. Specifically, we assume that the annual persistence probability, P, for a given MGS, depends on the effects of a set of dependency-associated CAGs in the form of a logistic regression model: $\ln(P/[1\text{-}P]) = \mathrm{logit}(P) = b_0 + \mathrm{Sum}[b_i X_i]$ or, equivalently: $P = \mathrm{expit}(b_0 + \mathrm{Sum}[b_i X_i])$. Here, the regression coefficient b_0 corresponds to the inherent persistence tendency of the MGS itself, b_i corresponds to the effect of dependency-associated CAG number i and X_i is a binary variable indicating whether dependency-associated CAG number i is present or absent for a given sample. "Expit" is the sigmoidal, logistic function (the inverse of the logit function). The index, i, runs over all the dependency-associated CAGs for a given MGS.

The probability that a CAG will survive for t days, $P(t)$, can be found from its annual persistence probability, P, in the following way: $P(t) = P^{[t/365]}$. The likelihood for a data point where the MGS survives (i.e., where it is still present at the second sample, after t days have elapsed) is therefore given by the following expression: $L = P^{[t/365]} = [\mathrm{expit}(b_0 + \mathrm{Sum}[b_i X_i])]^{[t/365]}$. For data points where a CAG does not survive, the likelihood is simply: $L = 1 - P^{[t/365]} = 1 - [\mathrm{expit}(b_0 + \mathrm{Sum}[b_i X_i])]^{[t/365]}$. As recommended in Gelman *et al.*[48] the priors for all b_0 regression coefficients (which correspond to the inherent persistence of all MGS) are t-distributions with $\mu = 0$, d.f. = 1, and rate = 0.1 (corresponding to scale = 10). The priors for all b_i regression coefficients (corresponding to the effects on persistence of the dependency-associated CAGs) are t-distributions with $\mu = 0$, d.f. = 1 and rate = 0.4 (corresponding to scale = 2.5). These are conservative priors that help keep the correlation coefficients close to zero. Given these expressions for priors and likelihoods, it is possible to perform a Bayesian analysis of the model, resulting in estimates of the above-mentioned regression coefficients. However, since the regression coefficients themselves can be difficult to interpret, we instead report the following derived measures: (i) the annual persistence probability for each MGS. This can be computed as: $P = \mathrm{expit}(b_0)$. (ii) The annual persistence probability for a specific MGS when together with a given dependency-associated CAG. This can be computed as: $P = \mathrm{expit}(b_0 + b_j)$, where j refers to the specific dependency-associated CAG. (iii) The effect of the dependency-associated CAG. We have chosen to simply express this as the absolute difference between the above two measures. (For instance, if the annual persistence probability of an MGS, together with a specific dependency-associated CAG, is 0.75, and the annual persistence probability of the MGS alone is 0.5, then the effect of the dependency-associated CAG is reported as 0.75 – 0.5 = 0.25).

For the analysis of coexistence between a pair of MGS and an associated CAG (**Supplementary Note 9**, **Supplementary Fig. 16** and **Supplementary Data 11**), there were insufficient data to obtain estimates for each individual CAG. We therefore pooled all data points for CAGs having a positive effect on the persistence of their MGS host, and estimated an overall effect for these.

Note that, in the Bayesian framework, estimates are expressed as probability distributions over the possible values for parameters of interest[20]. We therefore obtain both an estimate of a parameter, and quantification of how certain we are of the estimate. To declare an effect to be substantially different from zero, we require that its 95% highest posterior density interval (the "95% HDI") should be located entirely outside of a "region of practical equivalence" to 0 (a "ROPE"). In this analysis the ROPE was defined to be [−0.02, 0.02]. The 95% HDI is the narrowest interval that includes 95% of the probability. By design, all parameter values inside a 95% HDI will be more likely than all values outside. In this work we have identified 26 dependency-associated CAGs where we are more than 95% certain that they have a nonzero effect on the persistence probability of an MGS (**Supplementary Data 6**).

The model was implemented and analyzed in a Bayesian framework by Markov chain Monte Carlo (MCMC) using the JAGS package[49]. Convergence of MCMC was checked by running two independent chains and verifying that they arrived at similar posterior distributions. In particular it was checked that the potential scale reduction factor ("R-hat") for each estimated parameter was <1.1 (ref. 50).

22. Kultima, J.R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLOS ONE* **7**, e47656 (2012).
23. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
24. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132 (2010).
25. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
26. Leplae, R., Lima-Mendez, G. & Toussaint, A. ACLAME: a classification of mobile genetic elements, update 2010. *Nucleic Acids Res.* **38**, D57–D61 (2010).
27. Finn, R.D., Clements, J. & Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–37 (2011).
28. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
29. Kristensen, D.M., Cai, X. & Mushegian, A. Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *J. Bacteriol.* **193**, 1806–1814 (2011).
30. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40**, D284–D289 (2012).
31. Tringe, S.G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).
32. Roessner, C.A. & Scott, A.I. Fine-tuning our knowledge of the anaerobic route to cobalamin (vitamin B12). *J. Bacteriol.* **188**, 7331–7334 (2006).
33. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
34. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).
35. Kobayashi, K. *et al.* Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. USA* **100**, 4678–4683 (2003).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
37. Kelley, D.R., Schatz, M.C. & Salzberg, S.L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* **11**, R116 (2010).
38. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
39. Mavromatis, K. *et al.* Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods* **4**, 495–500 (2007).
40. Earl, D. *et al.* Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* **21**, 2224–2241 (2011).
41. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F.O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6**, 938–947 (2004).
42. Salzberg, S.L. *et al.* GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22**, 557–567 (2012).
43. Koren, S., Treangen, T.J. & Pop, M. Bambus 2: scaffolding metagenomes. *Bioinformatics* **27**, 2964–2971 (2011).
44. Ciccarelli, F.D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
45. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).
46. Treangen, T.J., Sommer, D.D., Angly, F.E., Koren, S. & Pop, M. Next generation sequence assembly with AMOS. *Curr. Protoc. Bioinformatics* Chapter 11, Unit 11.8 (2011).
47. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
48. Gelman, A., Jakulin, A., Pittau, M.G. & Su, Y. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2**, 1360–1383 (2008).
49. Plummer, M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. in *Proc. 3rd Int. Work. Distrib. Stat. Comput.* March, 20–22 (2003).
50. Gelman, A. & Rubin, D. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–511 (1992).

doi:10.1038/nbt.2939