The large-scale organization of the bacterial network of ecological co-occurrence interactions

Shiri Freilich[1,2,*], Anat Kreimer[3,4], Isacc Meilijson[1], Uri Gophna[5], Roded Sharan[1] and Eytan Ruppin[1,2]

[1]Blavatnik School of Computer Sciences, [2]School of Medicine, [3]School of Mathematical Science, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel, [4]Department of Biomedical Informatics, Columbia University, NY 10032, New York, USA and [5]Department of Molecular Microbiology and Biotechnology, Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel

ABSTRACT

In their natural environments, microorganisms form complex systems of interactions. Understating the structure and organization of bacterial communities is likely to have broad medical and ecological consequences, yet a comprehensive description of the network of environmental interactions is currently lacking. Here, we mine co-occurrences in the scientific literature to construct such a network and demonstrate an expected pattern of association between the species' lifestyle and the recorded number of co-occurring partners. We further focus on the well-annotated gut community and show that most co-occurrence interactions of typical gut bacteria occur within this community. The network is then clustered into species-groups that significantly correspond with natural occurring communities. The relationships between resource competition, metabolic yield and growth rate within the clusters correspond with the r/K selection theory. Overall, these results support the constructed clusters as a first approximation of a bacterial ecosystem model. This comprehensive collection of predicted communities forms a new data resource for further systematic characterization of the ecological design principals shaping communities. Here, we demonstrate its utility for predicting cooperation and inhibition within communities.

INTRODUCTION

In most natural environments, individual organisms do not live in isolation but rather form a complex system of inter-species interactions. These interactions shape the structure of the ecological community and play an important role in species' evolution (1). A rapidly expanding body of research indicates that complex social behaviors are commonly observed not only in animals but also in bacterial species that have been traditionally thought of as independent free-swimming organisms (2–5). Microbes are now known to form complex communities where, e.g. a metabolite produced by one organism can be used by another, which in return provides a different service (6). Within a community of species sharing limited resources, interactions can be described in terms of 'competition' and 'cooperation' where the structure of a community may be determined to a large extent by the type of relations between its members.

Considering the enormous cumulative mass of microorganisms in nature and their vast diversity, the structure of microbial communities has a considerable effect on the function of the ecosystem (7). The human body, e.g. hosts microorganisms estimated to outnumber human cells by a factor of 10, providing functions that humans did not evolve on their own (8,9). Variations in the identity and abundance of the microorganisms within the human body have important medical implications including obesity (10) and resistance (11) to pathogens. Other examples for ecosystems which are known to be affected by variations in the structure of their bacterial communities include bioreactors, agricultural fields and marine environments (7,12–14). Thus, understanding the organization of the network of bacterial interactions is likely to have broad medical and ecological consequences.

To date, a systematic description of the ecological communities formed by microorganisms is lacking (7,15,16). Although several resources are now available providing general annotations for the lifestyle of hundreds of fully sequenced microorganisms, no comprehensive data is currently on hand for sorting out the interrelationships among species. The accumulation of genomic and metagenomic data now enables a considerable progress: first, the availability of complete genomes from bacterial species exhibiting diverse lifestyles enables the development of new approaches towards predicting

ecological attributes (17–19). Such approaches include the construction of a genomic-driven ecological model for predicting the ability of microorganisms to inhabit natural-like environments (17,19) as well as the drawing of a microbial interaction network according to the co-occurrence of laterally transferred genomic elements (15). Second, the use of high-throughput sequencing and whole-community analysis techniques has led to a systematic characterization of species sharing the same ecological niche (7,12,20). However, this approach is to a large extent still limited to a few heavily sampled environments such as marine habitats and the human gut. Here, we put forward the use of a third source of data—the scientific literature (as reflected in PubMed)—in order to make use of this long-accumulated knowledge for achieving a systematic and comprehensive description of bacterial co-occurrence interactions. To this end, we apply a co-occurrence analysis—a technique previously shown to reflect true biological associations in genomic studies (21–23). We apply this approach for constructing a network of species connected according to their co-occurrence in PubMed entries. First, we provide evidence for the ecological plausibility of the network constructed as a model for studying ecological associations. Then, we analyze the patterns of interactions formed between species, the communities identified and the ecological design principles characterizing the different types of communities.

MATERIALS AND METHODS

Description of the data set

For a data set of 486 fully sequenced bacterial species (listed at Supplementary Table S1) we obtained annotations for their level of environmental diversity. Two types of annotations were used: fractions of regulatory genes were taken from (24), describing the fraction of transcription factors out of the total number of genes in the genome—an indicator of environmental variability (25). General environmental complexity estimates were also obtained from (25), where the natural environments of bacterial species were categorized based on the National Center for Biotechnology Information (NCBI) classification for bacterial lifestyle (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) and ranked according to the complexity of each category (1—obligatory symbionts; 2—specialized; 3—aquatic; 4—facultative host-associated; 5—multiple and 6—terrestrial species). Annotations for the fraction of regulatory genes were available for 133 species; annotations for the environmental complexity were available for 109 species. The taxonomic affiliation of each species at the phylum and class level is provided at Supplementary Table S1.

Constructing species interaction data according to species co-occurrence in PubMed

For each pair-wise combination of species, we automatically queried PubMed and counted the number of articles retrieved while asking for both species. Querying PubMed limits the search to the abstract and, therefore, reveals only these associations where species are highly relevant to the publication, making it more likely that the species are biologically associated. Web searches were done with the NCBI Entrez Programming utilities which are tools that provide access to Entrez data outside the regular web query interface and are helpful for automatically retrieving search results from multiple queries (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html). Queries were done using the corresponding species name and not the full description of the strain (i.e. *Escherichia coli* rather than *E. coli* K-12 MG1655; Supplementary Table S1) resulting in 336 distinct entries (termed 'species' entries').

The probability that two species co-occur together at a rate higher than chance expectation was determined by calculating a cumulative hypergeometric *P*-value. The corresponding size of the population is calculated as the sum of single-species entries (i.e. number of papers retrieved when querying for a single species, e.g. *E. coli*) minus the sum of papers retrieved when asking for pair-wise combination, yielding a value of 615268. Significance cut-off was determined by setting a False Discovery Rate threshold of 10%. Species pairs (the network of co-occurrence interactions) are listed at Supplementary Table S2. The network is shown in Supplementary Figure S1. To verify the robustness of our observations we also constructed alternative networks using false discovery rate (FDR) thresholds of 5% and 15%.

Clustering the PubMed-derived species co-occurrence network

The network of species connected by co-occurrence interactions was expressed as a network of interactions; this is known in graph theory as a line graph (26). This approach allows an overlapping partitioning of the co-occurrence network and hence allows species to be present in multiple modules. This one-to-many node's classification reflects the widely accepted view for the prevalence of species in nature—whereas some species are endemic and their dispersal is limited to a specific habitat, other species are ubiquitously spread in nature and are associated with different niches and communities. In this procedure the network of species (nodes) connected by interactions (edges) is expressed as a network of connected interactions. In the transformed graph, each node represents a co-occurrence interaction between two species and each edge represents pairs of interactions connected by a common species. Each binary interaction is condensed into a node that includes the two interacting species. These nodes are then linked by shared species content (27). We then use TribeMCL (26,28), an algorithm for clustering by flow simulation, at an inflation value of 2.5 to partition the interaction network and recover clusters of associated interactions. These clusters of interactions are then transformed back from an interaction–interaction representation to a species' representation for all subsequent validation and analysis. Partitioning this co-occurrence network yielded a total of 190 clusters, ranging in size from 2 to 32 species members ('Materials and Methods' section). The full list of clusters is provided

at the Supplementary Table S3. As a benchmarking, we constructed 1000 sets of random communities. Each set of random communities was constructed by shuffling between species in the original set of 190 communities (constructed using TribeMCL) where species A is always replaced by species B, hence maintaining the same size distribution of the original clusters. To verify the robustness of our observations we also clustered the data using alternative inflation values—2.3 and 2.7—yielding a total of 76 and 338 clusters with a maximal size of 33 and 27 species members, respectively. For each inflation value we also constructed 1000 sets of random communities that maintain the size distribution of the original clusters.

To repeat our analysis while using alternative clustering approaches, the original network of species' co-occurrences was also partitioned into clusters using NetworkBLAST (29) (beta = 0.8, true factor = 0.5), yielding 132 clusters ranging in size from 4 to 15 species members. A total of 1000 sets of random communities maintaining the size distribution of the original clusters were constructed.

Retrieving ecological co-occurrence data from environmental samples and NCBI annotations and identifying literature-driven clusters enriched in ecological groups

Occurrence of species from our dataset in environmental samples was inferred according to the results of the Basic Local Alignment Search Tool (BLAST) search (30) of the corresponding 16S RNA sequences against NCBI's env_nt (a comprehensive collection of sequences from environmental samples obtained in Whole Genome Shotgun sequencing projects, downloaded on May 2009 from ftp://ftp.ncbi.nih.gov/blast/db/FASTA/env_nt.gz), using the same parameters as in (20) (>98% sequence identity over at least 400 base pairs between query and hit). Overall, 77 species' entries were detected in 26 environmental samples including 17 samples from the gut, three marine samples, two samples from fresh water, two whale fall samples, one soil sample and one Enhanced Biological Phosphorus Removal (EBPR) Sludge sample. The occurrences of species within the environmental samples are provided at Supplementary Table S4.

Lifestyle annotations according to the NCBI classification scheme (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) were retrieved for all species analyzed and are listed in Supplementary Table S1. For each of the unsupervised, literature-driven clusters, we listed the number of species shared between the cluster and each ecological category (NCBI lifestyle categories and environmental samples). Significant enrichment in an ecological category was determined by calculating a cumulative hypergeometric *P*-value. Significance cutoff was determined by setting an FDR threshold of 10% (calculated independently for each ecological group). Significantly enriched clusters are listed in Supplementary Table S5.

Ecological characterization of clusters

As detailed in the following, we characterized each of the clusters with respect to three ecological parameters: respiration mode, maximal growth rate and the level of competition for natural resources among its members. The ecological characterization of the clusters is available at Supplementary Table S6.

Mode of respiration. Description of respiratory modes was retrieved from (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). From the species in the data set 53, 153, 187 and 16 are anaerobic, aerobic, facultative and microaerophilic, respectively (the respiratory mode of 77 species is described as 'unknown', Supplementary Table S1). In each cluster we looked at the distribution of anaerobic and aerobic bacteria—two categories where species exhibit a clear preference regarding the oxidative condition in their environment. Clusters that include aerobic bacteria (possibly together with facultative and microaerophilic species but not with anaerobic species) are considered aerobic clusters. Clusters that include anaerobic bacteria (possibly together with facultative and microaerophilic species but not with aerobic species) are considered anaerobic clusters. Out of 190 clusters, we found 150 clusters including at least a single aerobic or anaerobic species (Supplementary Table S6). The remaining 40 clusters contain only facultative and microaerophilic species, categories that do not provide sufficient information for determining the respiratory preferences of their community members and hence were not further analyzed. From these 150 clusters, 87 and 31 clusters include either aerobic or anaerobic, respectively, but never both simultaneously (all clusters may include facultative or microaerophilic species). From the remaining 32 clusters, five exhibit a preference toward anaerobic species (clusters 2, 5, 48, 56 and 59) and four exhibit a preference toward anaerobic species (clusters 14, 18, 19 and 51) although the small size of the clusters prevents conclusive statistical results (Supplementary Table S6, the distribution of clusters classified according to their mode of respiration is shown in Supplementary Figure S2).

Growth rate. Maximal growth rate information is available for 108 species in the data (17) (Supplementary Table S1). The median doubling time and the mean log doubling time of all its members for which doubling time is available was recorded for each community, as well as the actual distribution of doubling time within each community (Supplementary Table S6). To compare the intra- and inter-cluster similarity of the growth rate of species, we recorded the doubling-time distance for all possible species pairs according to the absolute value of the subtraction between the individual log doubling times, and compared the intra-cluster distance (recorded for pairs of species which are both members of the same community) with the inter-cluster distance outside communities (recorded for pairs of species which are not members of the same community). The inter- and intra-cluster distributions were compared in a

Wilcoxon rank sum test conducted by using the wilcox.test function at the R platform (providing a minimal *P*-value of 2.2e–16).

Metabolic overlap. Whereas the first two attributes are derived from experimental data, the level of competition between two species is estimated according to a qualitative score of the Effective Metabolic Overlap (EMO) in their nutritional demands. For each species, we have reconstructed its metabolic network, its optimal metabolic environment and a list of essential metabolites it has to produce in order to grow (17). Briefly, species-specific metabolic networks were constructed according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (release 46) (31). A list of metabolites that are likely to be essential for growth in most species was used for constructing species-specific target-metabolite lists according to the intersection between the generic target metabolites and the metabolites that each species produces. Metabolic growth environments were inferred for each species individually using the seed algorithm developed by (32), which similarly to other recently developed approaches (33–35), predicts the set of metabolites which are externally consumed by a species, given its metabolic network. To score the metabolic consequences of the presence of species B in the environment of species A on A's metabolic capacities, we removed from A's optimal environment all the metabolites consumed by B, simulated growth of species A on the media retrieved and then quantified the fraction of essential metabolites that A can still produce. Growth simulation was done by using the expansion method (36,37)—an approach where networks of increasing size are constructed starting from an initial set of substrates by stepwise addition of those reactions whose substrates are produced in the current core network. This approach requires the network topological backbone, as available here, and not a full blown stoichiometric model available for a limited number of species. EMO score denotes 1 minus the fraction of produced essential metabolites. Competition score (EMO score) 1 indicates that two species compete on the same resources; Competition score (EMO score) 0 indicates that two species utilize different metabolites for growth. Note that this asymmetric procedure may provide different competition relations between A $\rightarrow$ B than B $\rightarrow$ A, as one would intuitively expect. Level of competition was recorded between all species-pairs in the data. Notably, this approach only considers competition under optimal conditions, whereas, obviously, competition is also expected under less favorable conditions. For each cluster, we recorded the mean EMO score between all its members (Supplementary Table S6). To compare the intra- and inter-cluster level of competition we recorded the mean intra-cluster EMO (recorded for pairs of species which are both members of the same community) and the inter-cluster mean EMO (recorded for pairs of species which are not members of the same community). The inter- and intra-cluster distributions were compared in a Wilcoxon rank sum test conducted by using the wilcox.test function at the R platform.

Obligatory versus facultative communities. To describe communities as 'obligatory' or 'facultative'—i.e. whether species within these communities can be found in a limited or a vast range of alternative communities, respectively—we calculated the mean number of clusters to which their species' members are classified (Supplementary Table S6). 'Obligatory' and 'facultative' communities are these where the mean number of alternative clusters per community members is lower and higher than the median values over all communities, respectively.

RESULTS

Literature-driven data allow the reconstruction of ecologically plausible data set of pair-wise co-occurrence interactions

As a first step we constructed a matrix describing all pair-wise associations between 486 fully sequenced bacterial species. For every pair of species, we recorded the number of PubMed entries retrieved when searching for both species together ('Materials and Methods' section). Two species are then considered to be associated if their observed co-occurrence rate is significantly higher than the chance expectation ('Materials and Methods' section), leading to the overall detection of 1086 associations involving 466 species. To study the ecological plausibility of the recorded co-occurrence interactions we recorded the number of associated partners of species exhibiting different lifestyles (Figure 1). In accordance with ecological data we observed that the lowest number of interacting partners is recorded for obligatory symbionts and specialized bacteria while the highest number of co-occurrences is recorded for soil bacteria (25,38,39). Similar results are obtained when setting alternative FDR values (5 and 15%; an FDR threshold of 10% is otherwise reported) for determining the threshold for the significance of co-occurrence rate (Supplementary Figure S3). Notably, our network of co-occurrence interactions is focused on
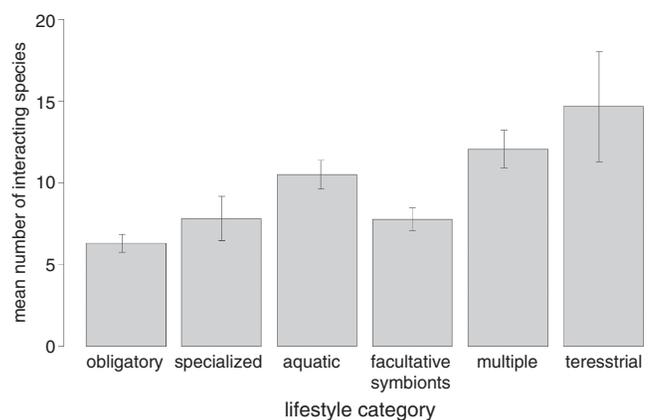

Figure 1. Mean number of co-occurring partners identified for species of different lifestyle. Annotations of lifestyle are according to (25), lifestyle categories are ordered according to their complexity from the most simple communities of obligatory host-associated species to the most complex terrestrial communities ('Materials and Methods' section). Number of species in each category are (ordered as in the figure): 27, 5, 4, 42, 28, 3. Bars show the standard error.

describing the interactions between the 486 species studied, possibly introducing biases toward more widely covered categories. It is hence re-assuring that a low number of interactions is observed for a relatively well-represented category such as the obligatory symbionts (Figure 1). On the other hand it can be expected that a higher coverage of species from other lifestyle categories (such as soil bacteria) will reveal additional interacting partners within this category and possibly lead to even more drastic differences between the numbers of interacting partners of species from these lifestyle categories. The specialized bacterium *Psychrobacter arcticum* is an example of a species with no interacting partners. *Psychrobacter arcticum* is commonly isolated from cold and hostile environments and hence provides a model for studying the growth of organisms in potentially lifeless space environments (40). The highest number of co-occurring partners—26—is recorded for *Enterococcus faecalis*, a mostly commensal organism inhabiting the highly populated human gastrointestinal tract. Beyond these specific examples we observe a significant positive correlation between the number of co-occurring partners a species has and its environmental diversity. Environmental diversity was estimated using two species-specific measures ('Materials and Methods' section): fraction of regulatory genes (Spearman correlation: 0.24, *P*-value: 3*e*–4) and the variability of species' habitats (environmental complexity estimate, Spearman correlation: 0.38, *P*-value: 6*e*–5).

Since co-occurrence in the literature potentially reflects association types other than ecological overlap, we ruled out the effect of another type of association—taxonomic proximity—on our observations. To this end, co-occurrence interactions between species of close phylogenetic proximity (class members, the phylum affiliation was considered for these species lacking a class affiliation) were excluded, leaving 495 inter-family co-occurrence interactions. For this data set we still observed a significant positive correlation between the number of co-occurring partners and the species' environmental diversity (estimated according to the fraction of regulatory genes: Spearman correlation: 0.25, *P*-value: 4*e*–3; environmental complexity estimate: 0.36, *P*-value: 1*e*–4).

Pair-wise co-occurrence interactions in the environmental samples

Although the estimates of ecological diversity allow a systematic annotation of species' lifestyle characteristics, they only provide a crude estimate of the true environmental complexity and do not provide direct information in support of the predicted pair-wise co-occurrence interactions. The ecological plausibility of the literature-based pair-wise co-occurrences was further assessed by benchmarking against co-occurrence of species in environment-specific samples. To this end, we focused on the two niches where comprehensive data is available (gut and marine; 'Materials and Methods' section). Considering any species identified in the same sample to be ecologically associated, we identify 1159 co-occurrence

interactions between members of the same sample or 1433 co-occurrence interactions between members of the same sample type (i.e. different gut samples). Of these, 109 and 131 interactions, respectively, are also detected by the literature-driven data. Overall, the experimental data support more than 10% of the literature-driven co-occurrences, where the overlap between these sets is higher than expected by chance (hypergeometric *P*-value = 0) hence providing support to the ecological plausibility of our predictions. The overlap between the sets remains significant when considering only intra- and or inter-family co-occurrences (hypergeometric *P*-value < 0.05, Text S1 in the Supplementary Data). As an illustration of how well the literature-driven co-occurrences allow the reconstruction of naturally occurring communities we looked at the co-occurring partners identified for five species in our data set— *Bacteroides fragilis, Bifidobacterium adolescentis, B. longum, B. thetaiotaomicron* and *E. faecalis*; these species can be considered obligatory gut bacteria as the gut is their main natural environment. Overall, the five species form 69 co-occurrence interactions (edges in the network; Figure 2). As expected, the large majority of these co-occurrence interactions (78%) occur within the gut community, i.e. between the obligatory gut bacteria and themselves or between the obligatory gut bacteria and facultative gut bacteria (i.e. gut species that also occupy alternative environments, e.g. *Propionibacterium acnes* (41); 'Materials and Methods' section).

Unsupervised partitioning of the co-occurrence network produces clusters which correspond with naturally occurring communities

Next, we examined whether unsupervised approaches can be applied for automatically partitioning the co-occurrence network into modules which correspond to naturally occurring communities. In correspondence with true environments, each species-node in the network was allowed to be classified to more than a single cluster ('Materials and Methods' section). For example, the ecologically versatile *Pseudomonas aeruginosa* (42) is classified into nine clusters. The typical gut bacteria *E. faecalis*— although having the highest number of co-occurring partners (26 in comparison to 15 partners recorded for *P. aeruginosa*)—is mapped into a single cluster. Notably, from the 27 members of this cluster, *E. faecalis* is clustered together with 21 other members of the gut community (manually curated according to the scientific literature as obligatory and facultative gut bacteria, Text S1 in the Supplementary Data), thus creating a highly specialized cluster corresponding to the natural occurring gut community. These two highly connected species correspond to the two types of hubs described for protein–protein interaction networks (43). The single-community *E. faecalis* can be considered as a 'party' hub, that is a node which interacts with most of its partners simultaneously, whereas *P. aeruginosa* can be considered as a 'date' hub, that is a node that is linked with different partners at different locations. In analogy to the proteome network, such 'date' nodes have the potential to link ecological communities,
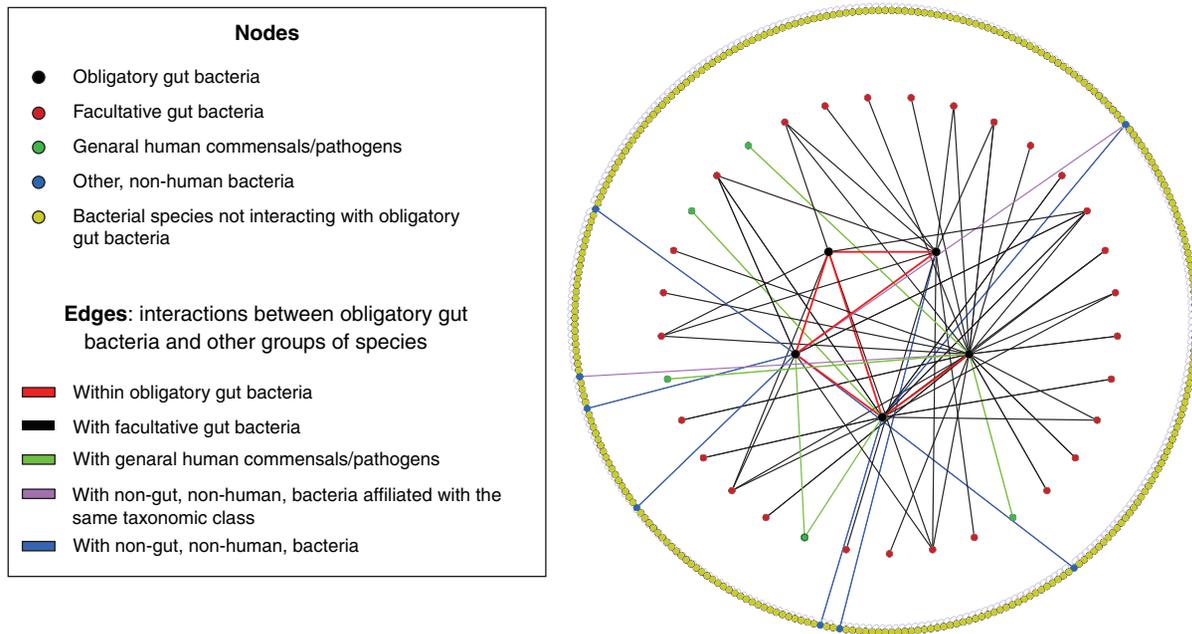
Figure 2. Co-occurrence interactions between obligatory gut bacteria and other bacterial species. Typical gut bacteria (*B. fragilis, B. adolescentis, B. longum, B. thetaiotaomicron* and *E. faecalis*—represented by black dots) interact with 41 species. Manual survey of these species identified 28 of them as facultative gut bacteria (represented by red dots), five as non-gut human-associated bacteria (green dots) and eight as non-human associated species (blue dots). Overall, the five typical gut bacteria form 69 co-occurrence interactions, 54 of them (78%) are among themselves (six co-occurrence interactions, red lines) and with facultative gut species (48 co-occurrence interactions, black lines). Only 15 co-occurrence interactions are with non-gut species (red, purple and blue lines). The full list of co-occurrence interactions, species, and ecological annotations is provided at Text S1 in the Supplementary Data.

e.g. by transferring genetic elements between otherwise ecologically disconnected species.

Beyond the gut example (which relates to a well-documented population), the general correspondence between the unsupervised, literature-driven clusters and naturally occurring communities was tested by looking for the enrichment of cluster-members in groups of ecologically associated species, as inferred from external, independent data sources. First, experimental data derived from environmental samples were used for this validation ('Materials and Methods' section), where species that are detected within the same sample types (e.g. marine sample) are considered to be ecologically associated. In total, 52, 92 and 100% of the species detected in marine, gut and EBPR sludge samples, respectively, were classified into clusters significantly enriched in one of these sample types (Figure 3; 'Materials and Methods' section), with a mean cluster's specificity (the fraction of species in the cluster which are found in the corresponding ecological sample) of 0.7. Totally, 32 of the clusters (rows) were significantly enriched in one of the five ecological groups (columns). In comparison, in 1000 random sets of communities maintaining the size and connectivity of the original clusters ('Materials and Methods' section), the maximal number of enriched clusters in a random set of 190 clusters was six (with a mean of 0.5 enriched clusters for each set of 190 clusters). Similarly, the number of enriched clusters revealed by using alternative inflation values and an alternative clustering approach (NetworkBLAST) is much higher than the number of clusters revealed in the corresponding

random sets of communities (1000 sets of random communities were constructed for each alternative clustering approach maintaining the same size distribution of the original clusters, results are shown at Supplementary Table S7). The identification of several clusters that highly correspond with independently defined groups of co-occurring species further validates our co-occurrence analysis as an efficient large-scale procedure for stratifying environmental associations and communities.

We further examined whether the 'guilt by association' principle, allowing transferring functional annotations between co-clustered proteins (27,44) can be applied for predicting the environmental distribution of un-annotated species in an ecologically characterized group. To this end, we looked in detail at few of the ecologically enriched clusters, shown in Figure 3, and searched the literature to obtain a description of the ecological distribution of cluster members that are not detected at the corresponding environmental sample (Text S1 in the Supplementary Data). We first looked at the clusters enriched with species found in marine samples. From the six marine clusters, the lowest specificity (<50%) was recorded in cluster 11—the biggest marine cluster. It is composed of 21 species-members, out of which nine species were identified in marine samples. Manual curation of the remaining 12 species revealed that nine of them can indeed be found in aquatic samples (Figure 3), implying that the co-occurrence signal can provide an indication for the ecological distribution of un-annotated species.

Next, we aimed to identify clusters that correspond to ecological environments that are more specialized than the
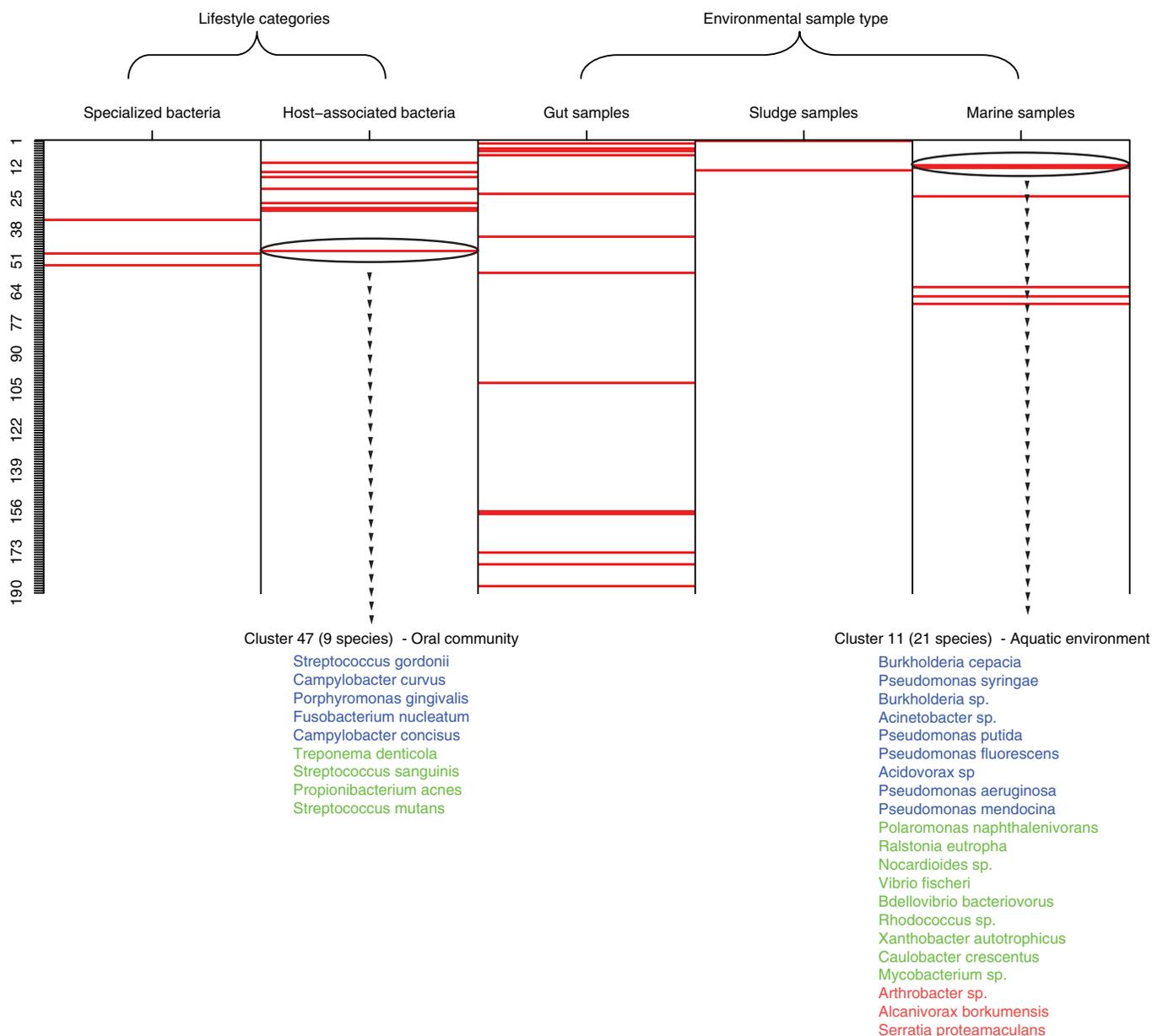
Figure 3. Correspondence between unsupervised, literature-driven clusters and ecological groups, delineating three clusters which are enriched in specialized bacteria, eight in host-associated bacteria, 13 in gut bacteria, two in bacteria from sludge samples and six in marine bacteria. Rows denote clusters (ordered according to their size), columns denote ecological groups. Enriched clusters (detected as described at the 'Materials and Methods' section) are shown in red. Only ecological groups (lifestyle categories and environmental sample types) for which at least a single cluster is enriched are shown (data is provided in Supplementary Table S5 in the Supplementary Data). The full list of species classified to cluster 11 and 47 discussed in the main text is detailed in the figure. Species in blue are classified into a specific ecological group (marine species according to detection of species in environmental sample; oral species according to manually curated list composed of five typical oral species). The remaining species within these clusters were manually examined for their ecological distribution; species colored in green are those that can inhabit the corresponding ecological cluster (Text S1 in the Supplementary Data), species colored in red are those whose ecological habitat is different.

yet generic aquatic environment (which is composed of many micro-environments). To this end, we used the general annotations for species' lifestyle ('Materials and methods' section), considering species sharing the same annotation (e.g. 'specialized' or 'host-associated' species) to be ecologically associated. Three clusters were found to be enriched in specialized bacteria species (Figure 3); a manual survey yielded a more specialized ecological description of each of these clusters. Cluster 48 is

enriched in species occupying anaerobic environments with typical high levels of sulfate and nitrate, grouping together sulfate-reducing bacteria and autotrophic denitrifiers and exhibiting similar species distribution to the one observed in granular sludge-bed reactor communities (45). The two other specialized clusters are also found to be enriched in hyper-thermophilic and radiation-resistant bacteria (Text S1 in the Supplementary Data). Similarly, we examined whether the clusters

enriched with host-associated bacteria correspond with more narrowly defined, naturally occurring communities. To this end we first defined a group of five species typical of the oral community—*Campylobacter curvus*, *Streptococcus gordonii*, *Fusobacterium nucleatum*, *Porphyromonas gingivalis* and *C. concisus*. We identified a single host-associated cluster which exhibits a significant enrichment in the typical oral bacteria (including all five species) and then examined the ecological distribution of the remaining species in the cluster—remarkably, all species were found to be inhabitants of the oral cavity (Figure 3).

Community members exhibit similarity in their ecological properties

The automated approach, together with the manual surveys, supports the ecological plausibility of this high-throughput classification procedure and indicates that indeed the literature-driven clusters correspond, at least in part, to naturally occurring communities. Hence, the ecological network of co-occurrences and its unsupervised partitioning into ecological groups represents an innovative type of data: a comprehensive collection of bacterial ecological clusters. As one expects that community members will exhibit a similarity in their ecological properties, we examined whether communities show a typical respiration mode, growth rate and metabolic preferences. For all three attributes we indeed observe an intra-cluster similarity. First, the large majority of clusters (80%) can be characterized by a distinct respiratory mode containing either aerobic or anerobic bacteria. Second, within a community, species exhibit a significantly higher similarity in their maximal growth rates in comparison to species outside the community (P-value $< 2.2e$–16 in a Wilcoxon test; Methods). The significance of the difference is kept when limiting the analysis to pairs of species which are not members of the same taxonomic family (P-value $< 2.2e$–16 in a Wilcoxon test). Third, within a community, species show a higher similarity in their metabolic demands (higher EMO) and hence possibly fiercer competition on the available natural resources, in comparison to species outside the community (P-value $< 2.2e$–16 in a Wilcoxon test; Methods). The significance of the difference is kept when limiting the analysis to pairs of species which are not members of the same taxonomic family (P-value $= 1e$–6 in a Wilcoxon test).

Using intra-cluster ecological properties for studying the design principles of ecological communities

The similarity between cluster members allows the characterization of communities in terms of competition—growth rate—respiration mode and a systematic study of the inter-relationships between these ecological attributes. We studied these inter-relationships in 38 clusters where information on maximal growth rate is available for at least half of the cluster members. For each cluster we calculated its typical growth rate, its typical level of intra-cluster competition and the ratio between aerobic to anaerobic species within the cluster ('Materials and

Methods' section). The anerobic to aerobic ratio in a cluster is used for estimating the typical growth yield—the efficiency of the conversion of substrate into biomass, whereas low yield is characteristic of anaerobic species and high yield is characteristic of aerobic species (46). We observe a general trend of lower competition (lower EMO score) in clusters with a typical fast growth rate (Spearman correlation: 0.6; P-value $< 1e$–4). Overall, the large majority of clusters exhibit either fast growth rate associated with low competition or slow growth potential associated with intense competition (Figure 4). Higher anaerobic/aerobic ratio (low yield) is observed in the fast-growing communities (Figure 4). Notably, a lower mean correlation between growth rate and competition (0.26) is observed across the 1000 random-communities sets. A correlation equal or higher than the correlation observed in the natural-like set is observed in $<5\%$ of the communities. Similarly, the correlation between growth rate and level of competition observed in clusters produced by NetworkBLAST—an alternative clustering approach—is much higher than the mean correlation observed in corresponding sets of random communities (Spearman correlation: 0.6 is observed in the true communities in comparison to mean correlation of 0.1 in the corresponding sets of random communities).

Notably, this division of clusters into low competition–fast growth–low yield versus high competition–slow growth–high yield category is in agreement with the r/K selection theory. The r/K selection theory, originally suggested for animals and plants (47), aims to explain the choice between slow and fast growth, given the environmental conditions and the level of competition encountered by a species. In general, r-strategies are adapted to maximize the rate of growth while K-strategies are adapted to compete and survive when resources are limited (48). When applied to bacterial species, r-selection is suggested to be typical of communities occupying rich metabolic environments where species can exhibit high growth rate and low yield; K-selection is suggested to be typical of species occupying less abundant environments, these species typically exhibit a lower growth potential but a higher capability to compete for substrates (49,50). Hence, the low competition–fast growth–low yield clusters corresponds with the characteristics of r-selection whereas high competition–slow growth–high yield clusters corresponds with K-selection. Clusters enriched with gut bacteria—a rich and highly populated metabolic environment—fall into the r-selection category (Figure 4). Clusters which fall into the K-selection are typically composed of host-associated pathogens and symbionts; e.g. in cluster 13 the majority of species populate plant's tissues (9 out of 13 are plant's symbionts and pathogens; Figure 4). The typical slow growth observed in these clusters possibly reflects the basic strategy of many host-associated bacteria that do not scavenge too much of the basic nutrients essential for the host metabolism since, otherwise, their host will soon starve to death and the bacteria will lose their protective niche (51). Possibly, K-selection is also typical of free-living communities and the lack of such clusters in our data set is the outcome of
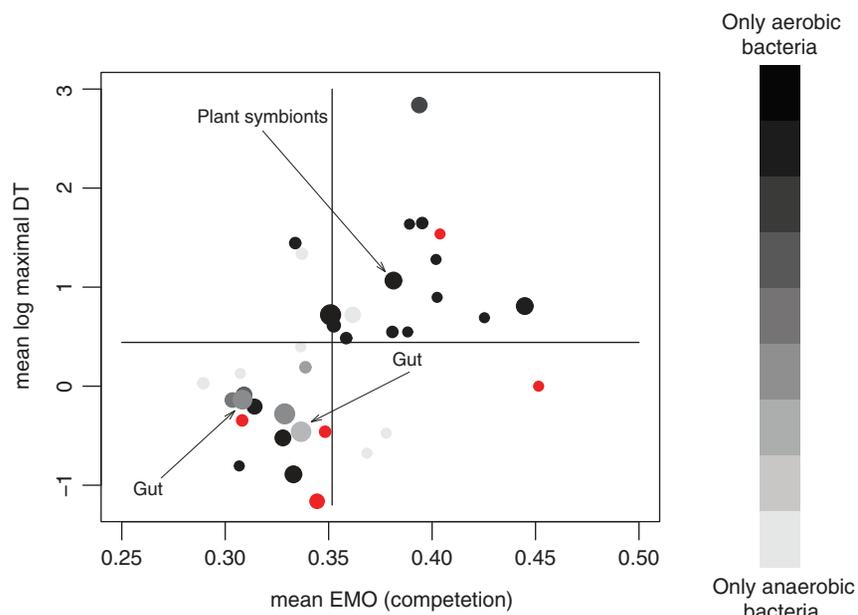
Figure 4. Level of metabolic competition versus growth rate in 38 clusters where growth rate information is available for a least half of the cluster members. Sizes of dots correspond to the sizes of the clusters. Colors of dots correspond to the anaerobic/aerobic ratio. Red dots correspond to cluster which do not contain any aerobic or anaerobic bacteria (alternative annotations include facultative, microaerophilic and unknown species). The plot is divided according to median values of the axes. The clusters indicated by arrows are described at Text S1 in the Supplementary Data; the ecological attributes of all clusters are provided in Table S6 in the Supplementary Data. Abbreviations: EMO—effective metabolic overlap; DT—duplication time.

biases in the collection of sequenced bacteria toward easily cultured, fast-grower species. For example, >99% of the soil bacteria (mostly fastidious growers) have not been cultivated and characterized and hence the soil ecosystems are, to a large extent, uncharted (50). The inclusion of uncultivated species in our data set can possibly lead to the identification of K-selection in communities dominated by free-living species. It is interesting to note that, in support of the ecological plausibility of our findings, 'obligatory' communities (i.e. communities whose members are classified into a limited number of alternative clusters) fit better to the r/K selection theory than 'facultative' communities (i.e. communities whose members are classified into a relatively large number of clusters), where the inverse relationship between growth rate and level of competition are more prominent is this subset of communities (Spearman correlation: 0.7; P-value = $7e$–4, observed in 21 'obligatory' communities, versus Spearman correlation: 0.6; P-value = $2e$–5, observed in the overall data and no significant correlation when only considering the 'facultative' communities; 'Materials and Methods' section).

Predicting various types of interactions between community members

Mapping clusters along the r/K-selection continuum axis might be an indicator for prevalence of different interaction types within different communities. It is reasonable to assume that some communities are more cooperative and less competitive than others. Revealing the intra-cluster interaction types can provide an indication to the real composition of a community at a given time

point: cooperating species are likely to coexist; competing species, on the other hand, are likely to exhibit a complementary distribution whereas only one of them will be dominant under a specific environmental setting. To further examine competition versus cooperation relations, we focused on pairs of species that coexist in at least three communities, assuming this testifies to a close environmental proximity between pair members. Within this group we looked for two different types of metabolic relationships: reduced versus elevated competition between pair members, assuming reduced competition might provide an indication for metabolic dependence and cooperation whereas elevated competition between species which typically populate the same niche provides an indication that presence of one of the species will inhibit the growth of the other. From the 193 pairs sharing at least three communities we identified 32 and 62 pairs with reduced or enhanced metabolic competition (pair-wise EMO is at least one SD lower/higher than the mean competition recorded for a species with other co-occurring partners, respectively). *Arthrobacter* sp. and *Acinetobacter* sp. provide an example for a pair where the resource competition of *Acinetobacter* sp. with *Arthrobacter* sp. is lower than the competition it exhibits with other co-occurring species (pair-wise EMO score 0.16 versus mean EMO score of 0.32, SD: 0.09; Supplementary Table S8). Notably, these two species form a consortium allowing the utilization of butyl benzyl phthalate (BBP) as a sole carbon source in municipal waste-contaminated soil in a pathway which requires the participation of genes from both species (52). Thus, the two species are distinct and complementary to each other in their metabolic activities

for the degradation of BBP (52), hence providing an example for metabolic dependence and cooperative interaction. *Listeria monocytogenes* and *Lactococcus lactis* provide an example for an ecologically associated pair with increased resource competition, where the resource competition between *L. monocytogenes* with *L. lactis* is higher than the competition it exhibits with other co-occurring species (pair-wise EMO score: 0.36 versus mean EMO score: 0.3, SD: 0.05; Supplementary Table S8). Interestingly, the inhibitory effect of bacteriocin produced by *L. lactis* on the growth of *L. monocytogenes* (53) supports the competitive nature of the interaction between the two species and their non-overlapping existence in a community. These two examples suggest that the integration of various types of systematic ecological data (in this case, literature-driven co-occurrences data together with genomic-driven metabolic data) can be used for the prediction and characterization of interaction types within a community.

DISCUSSION

In this analysis we show that by applying systematic literature-mining approaches, such as those previously shown to be very useful for the study of genomic and proteomic networks, we can chart the first large-scale network of bacterial ecological co-occurrences. Several lines of evidence—examining the ecological plausibility of the analysis from a species perspective, a community perspective and an inter-community perspective—support the constructed ecological clusters as a first approximation of a bacterial ecosystem model. First, from a species perspective, the degree of a species-node in the co-occurrence network (number of co-occurring species) corresponds with other species-specific ecological features; when focusing on the gut community as a case study, we observe that most of the co-occurrences occur within this community. Second, from a community perspective, we demonstrate that communities constructed in an unsupervised manner significantly correspond with natural occurring communities. Moreover, these communities exhibit coherent characteristic ecological behaviors regarding the respiration mode, growth rate and metabolic demands of their members. Finally, the inter-relationship between growth rate, metabolic yield and level of competition observed in these communities correspond with patterns expected according to the r/K-selection theory, hence providing the first large-scale account of this fundamental ecological theory. The availability of such a model raises several intriguing new challenges and allows a wide and diverse source of knowledge for exploring both environmental models, which were previously suggested based on the study of specific communities, and patterns of genome evolution where ecological communities data allow to systematically explore the transfer of genomic data between co-occurring species. For example, one can explore the potential role of 'dynamic' species—i.e. species classified to many ecological communities, in transferring genomic elements between otherwise disconnected ecological species.

Focusing on the possible ecological applications, the current dataset can serve as a basis for the future investigation of several interesting questions. Here, we provided two examples where ecological patterns delineated from the network of ecological data correspond with expected ecological behavior. However, up to now, our approach is limited to provide only direct predictions for the level of competition. Notably, alternative approaches such as 'metabolic synergy' provide direct predictions for the level of cooperation between species (54). Hence, the use of alternative approaches can allow a wider perspective for describing the broad range of interactions between species. Predicting cooperative interaction types between species, e.g. where the products of one interacting partner provide essential substrates for its partner, hence allowing it to survive in an otherwise unviable environment, can be applied for developing enrichment techniques toward successful cultivation. An example is the successful cultivation of *Leptospirillum* spp. following the characterization of functional partitioning among community members which allowed a better understanding of its metabolic requirements (2). Similarly, predicting inhibitory interactions (in which the presence of one species inhibits the growth of its competitor) can be used for inducing controlled shifts in bacterial populations. Considering the implications of gaining a systematic understanding of the structure and dynamics within bacterial communities, the further development of large-scale approaches for constructing and characterizing ecological communities is a major, innovative challenge in systems biology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Losos,J.B., Leal,M., Glor,R.E., De Queiroz,K., Hertz,P.E., Rodriguez Schettino,L., Lara,A.C., Jackman,T.R. and Larson,A.

(2003) Niche lability in the evolution of a Caribbean lizard community. *Nature*, **424**, 542–545.

2. Allen,E.E. and Banfield,J.F. (2005) Community genomics in microbial ecology and evolution. *Nat. Rev. Microbiol.*, **3**, 489–498.

3. Kolter,R. and Greenberg,E.P. (2006) Microbial sciences: the superficial life of microbes. *Nature*, **441**, 300–302.

4. Nadell,C.D., Xavier,J.B. and Foster,K.R. (2009) The sociobiology of biofilms. *FEMS Microbiol. Rev.*, **33**, 206–224.

5. West,S.A., Diggle,P.D., Buckling,A., Gardner,A. and Griffin,A.S. (2007) The Social lives of microbes. *Annu. Rev. Ecol. Evol. Systemtatics*, **38**, 53–77.

6. Marx,C.J. (2009) Microbiology. Getting in touch with your friends. *Science (New York, N.Y.)*, **324**, 1150–1151.

7. Fuhrman,J.A. (2009) Microbial community structure and its functional implications. *Nature*, **459**, 193–199.

8. Backhed,F., Ley,R.E., Sonnenburg,J.L., Peterson,D.A. and Gordon,J.I. (2005) Host-bacterial mutualism in the human intestine. *Science (New York, N.Y.)*, **307**, 1915–1920.

9. Turnbaugh,P.J., Ley,R.E., Hamady,M., Fraser-Liggett,C.M., Knight,R. and Gordon,J.I. (2007) The human microbiome project. *Nature*, **449**, 804–810.

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

11. Follows,M.J., Dutkiewicz,S., Grant,S. and Chisholm,S.W. (2007) Emergent biogeography of microbial communities in a model ocean. *Science (New York, N.Y.)*, **315**, 1843–1846.

12. DeLong,E.F. (2009) The microbial ocean from genomes to biomes. *Nature*, **459**, 200–206.

13. Lenski,R.E., Mongold,J.A., Sniegowski,P.D., Travisano,M., Vasi,F., Gerrish,P.J. and Schmidt,T.M. (1998) Evolution of competitive fitness in experimental populations of *E. coli*: what makes one genotype a better competitor than another? *Antonie van Leeuwenhoek*, **73**, 35–47.

14. Sloan,W.T., Woodcock,S., Lunn,M., Head,I.M. and Curtis,T.P. (2007) Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microb. Ecol.*, **53**, 443–455.

15. Hooper,S.D., Mavromatis,K. and Kyrpides,N.C. (2009) Microbial co-habitation and lateral gene transfer: what transposases can tell us. *Genome Biol.*, **10**, R45.

16. Pignatelli,M., Moya,A. and Tamames,J. (2009) EnvDB, a database for describing the environmental distribution of prokaryotic taxa. *Environ. Microbiol. Rep.*, **1**, 191–197.

17. Freilich,S., Kreimer,A., Borenstein,E., Yosef,N., Sharan,R., Gophna,U. and Ruppin,E. (2009) Metabolic-network-driven analysis of bacterial ecological strategies. *Genome Biol.*, **10**, R61.

18. Janga,S.C. and Babu,M.M. (2008) Network-based approaches for linking metabolism with environment. *Genome Biol.*, **9**, 239.

19. Freilich,S., Kreimer,A., Borenstein,E., Gophna,U., Sharan,R. and Ruppin,E. (2010) Decoupling environment-dependent and independent genetic robustness across bacterial species. *PLoS Comp. Biol.*, **6**, 1–10.

20. Lozupone,C.A., Hamady,M., Cantarel,B.L., Coutinho,P.M., Henrissat,B., Gordon,J.I. and Knight,R. (2008) The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc. Natl Acad. Sci. USA*, **105**, 15076–15081.

21. Muller,H. and Mancuso,F. (2008) Identification and analysis of co-occurrence networks with NetCutter. *PLoS ONE*, **3**, e3178.

22. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.

23. Stapley,B.J. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pacific Symp. Biocomput.*, 529–540.

24. Madan Babu,M., Teichmann,S.A. and Aravind,L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.*, **358**, 614–633.

25. Parter,M., Kashtan,N. and Alon,U. (2007) Environmental variability and modularity of bacterial metabolic networks. *BMC Evol. Biol.*, **7**, 169.

26. Dongen,S.v. (2000) Graph clustering by flow simulation. *Ph.D. Thesis*. University of Utrecht.

27. Pereira-Leal,J.B., Enright,A.J. and Ouzounis,C.A. (2004) Detection of functional modules from protein interaction networks. *Proteins*, **54**, 49–57.

28. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

29. Sharan,R., Suthram,S., Kelley,R.M., Kuhn,T., McCuine,S., Uetz,P., Sittler,T., Karp,R.M. and Ideker,T. (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*, **102**, 1974–1979.

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

31. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

32. Borenstein,E., Kupiec,M., Feldman,M.W. and Ruppin,E. (2008) Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proc. Natl Acad. Sci. USA*, **105**, 14482–14487.

33. Aguilar,D., Aviles,F.X., Querol,E. and Sternberg,M.J. (2004) Analysis of phenetic trees based on metabolic capabilites across the three domains of life. *J. Mol. Biol.*, **340**, 491–512.

34. Ebenhoh,O. and Handorf,T. (2009) Functional classification of genome-scale metabolic networks. *EURASIP. J. Bioinformat. Sys. Biol.*, **2009**.

35. Handorf,T., Christian,N., Ebenhoh,O. and Kahn,D. (2008) An environmental perspective on metabolism. *J. Theoret. Biol.*, **252**, 530–537.

36. Ebenhoh,O., Handorf,T. and Heinrich,R. (2004) Structural analysis of expanding metabolic networks. *Genome Informat.*, **15**, 35–45.

37. Handorf,T., Ebenhoh,O. and Heinrich,R. (2005) Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J. Mol. Evol.*, **61**, 498–512.

38. Mavromatis,K., Ivanova,N., Barry,K., Shapiro,H., Goltsman,E., McHardy,A.C., Rigoutsos,I., Salamov,A., Korzeniewski,F., Land,M. *et al.* (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods*, **4**, 495–500.

39. Tringe,S.G., von Mering,C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W., Podar,M., Short,J.M., Mathur,E.J., Detter,J.C. *et al.* (2005) Comparative metagenomics of microbial communities. *Science (New York, N.Y.)*, **308**, 554–557.

40. Bowman,J.P., McCammon,S.A., Brown,M.V., Nichols,D.S. and McMeekin,T.A. (1997) Diversity and association of psychrophilic bacteria in Antarctic sea ice. *Appl. Environ. Microbiol.*, **63**, 3068–3078.

41. Allison,C. and Macfarlane,G.T. (1989) Dissimilatory nitrate reduction by *Propionibacterium acnes*. *Appl. Environ. Microbiol.*, **55**, 2899–2903.

42. Stover,C.K., Pham,X.Q., Erwin,A.L., Mizoguchi,S.D., Warrener,P., Hickey,M.J., Brinkman,F.S., Hufnagle,W.O., Kowalik,D.J., Lagrou,M. *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, **406**, 959–964.

43. Han,J.D., Bertin,N., Hao,T., Goldberg,D.S., Berriz,G.F., Zhang,L.V., Dupuy,D., Walhout,A.J., Cusick,M.E., Roth,F.P. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

44. Schwikowski,B., Uetz,P. and Fields,S. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.

45. Chen,C., Ren,N., Wang,A., Yu,Z. and Lee,D.J. (2008) Microbial community of granules in expanded granular sludge bed reactor for simultaneous biological removal of sulfate, nitrate and lactate. *Appl. Microbiol. Biotech.*, **79**, 1071–1077.

46. Pfeiffer,T., Schuster,S. and Bonhoeffer,S. (2001) Cooperation and competition in the evolution of ATP-producing pathways. *Science (New York, N.Y.)*, **292**, 504–507.

47. MacArthur,R.H. and Wilson,E.O. (1967) *The Theory of Island Biogeography*. Princeton University Press, Princeton.

48. Pianka,E. (1970) On r- and K-selection. *Am. Nat.*, **104**, 592–597.

49. Fierer,N. and Jackson,R.B. (2006) The diversity and biogeography of soil bacterial communities. *Proc. Natl Acad. Sci. USA*, **103**, 626–631.
50. Torsvik,V. and Ovreas,L. (2002) Microbial diversity and function in soil: from genes to ecosystems. *Curr. Opin. Microbiol.*, **5**, 240–245.
51. Joseph,B. and Goebel,W. (2007) Life of *Listeria monocytogenes* in the host cells' cytosol. *Microbes Infect./Institut. Pasteur*, **9**, 1188–1195.
52. Chatterjee,S. and Dutta,T.K. (2008) Complete degradation of butyl benzyl phthalate by a defined bacterial consortium: role of individual isolates in the assimilation pathway. *Chemosphere*, **70**, 933–941.
53. Benkerroum,N., Oubel,H., Zahar,M., Dlia,S. and Filali-Maltouf,A. (2000) Isolation of a bacteriocin-producing *Lactococcus lactis* subsp. *lactis* and application to control *Listeria monocytogenes* in Moroccan jben. *J. Appl. Microbiol.*, **89**, 960–968.
54. Christian,N., Handorf,T. and Ebenhoh,O. (2007) Metabolic synergy: increasing biosynthetic capabilities by network cooperation. *Genome Informat.*, **18**, 320–329.