Unlocking the potential of metagenomics through replicated experimental design

Rob Knight[1], Janet Jansson[2,3,4], Dawn Field[5], Noah Fierer[6], Narayan Desai[7], Jed A. Fuhrman[8], Phil Hugenholtz[9], Daniel van der Lelie[10], Folker Meyer[7,11], Rick Stevens[7,11], Mark J. Bailey[5], Jeffrey I. Gordon[12], George A. Kowalchuk[13,14], Jack A. Gilbert[7,15]

[1] Howard Hughes Medical Institute and Department of Chemistry & Biochemistry, University of Colorado at Boulder, Boulder, Colorado, USA.
[2] Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA.
[3] Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, California, USA.
[4] Joint Bioenergy Institute, Emeryville, California, USA.
[5] Centre for Ecology & Hydrology, Wallingford, Oxford, UK.
[6] Cooperative Institute for Research in Environmental Sciences and Department of Ecology and Evolutionary Biology, University of Colorado at Boulder, Boulder, Colorado, USA.
[7] Argonne National Laboratory, Argonne, Illinois, USA.
[8] Dept. of Biological Sciences, University of Southern California, Los Angeles, California, USA.
[9]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences & Institute for Molecular Bioscience, The University of Queensland, St Lucia, Australia.
[10] RTI, Research Triangle Park, North Carolina, USA
[11]The Computation Institute, University of Chicago, Chicago, Illinois, USA.
[12]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA.
[13]Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands.
[14] Institute of Ecological Science, VU University Amsterdam, Amsterdam, The Netherlands.
[15] Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA.

Abstract

Metagenomics holds enormous promise for discovering novel enzymes and organisms that are biomarkers or causes of processes relevant to disease, industry and the environment. In the last two years we have seen a paradigm shift in metagenomics to the application of broad cross-sectional and longitudinal studies enabled by advances in DNA sequencing and high-

performance computing. These technologies now make it possible to broadly assess microbial diversity and function, allowing systematic investigation of the largely unexplored frontier of microbial life. To achieve this aim, the global scientific community must collaborate and agree upon common objectives and data standards to enable comparative research across the Earth's microbiome. Improvements in comparability of data will facilitate the study of biotechnologically relevant processes such as bioprospecting for new glycoside hydrolases or identifying novel energy sources.

Introduction

The Earth hosts more than 10^{30} microbial cells[1], a figure that exceeds the number of known stars in the universe by nine orders of magnitude. This richness of single-celled life, the first life to evolve on the planet, still accounts for the vast majority of functional drivers of our planet's ecosystems[2]. Yet the diversity and interdependencies of these microscopic organisms remain largely unknown. Likewise, our understanding of the functional potential of most individual microbial taxa residing within any ecosystem is extremely limited and generally restricted to measurements of gross enzymatic processes of the community. Moreover, sequenced metagenomic datasets have, to date, only played a limited role in biotechnological knowledge discovery, with the majority of novel developments occurring through heterologous expression of enzymes.

Our knowledge of microbial diversity on Earth is poised to be revolutionized by the development of new technologies that will permit us to 'see' the 'who, what, when, where, why, and how?' of microbial communities. Most recently, next-generation sequencing methods have begun to rapidly improve our understanding of the functional and evolutionary processes necessary to advance the field of microbial ecology. Matching these technological strides are progress in scientific community cooperation, increases in interdisciplinary interaction, and the development of standards for experimental and sample contextual "metadata" acquisition, which are essential for downstream interpretation[3].

Here we discuss how advances in DNA sequencing, the handling of contextual data and improvements in study design can unlock the potential of metagenomics. We discuss the need for robust experimental design[4] (e.g., replication and improved ecosystem characterization) and

highlight the need for an Earth Microbiome Project that will rely on metagenomics to explore Earth's microbial dark matter across temporal and spatial scales and simultaneously facilitate novel gene discovery. Through standardized data generation approaches and metadata collection, we stand poised to make rapid progress toward advancing biotechnological goals.

Changing the paradigm in metagenomic experimental design

For more than 80 years, it has been recognized that the majority of microbial life cannot be easily cultured in the laboratory. This has constrained understanding of microbial ecosystems and impeded our ability to discover and utilize new beneficial functions derived from microorganisms (e.g., enzymes to drive biotechnological reactions, processes to enhance bioremediation, and biomarkers for disease diagnosis and therapeutic targets). Current biotechnology is still based on a small stable of "domesticated" species, yet technical improvements in molecular microbial ecology and synthetic biology offer the potential for novel enzyme discovery and exploitation from the previously inaccessible depths of the tree of life. However, in this age of exploration and discovery, as we test the capability and limits of these new tools, it is unsurprising that the majority of studies have failed to live up to expectations.

This has created a paradox, in that funding agencies are not providing the resources required to undertake metagenomic sequencing and analysis of the large and sufficiently replicated sample sets needed to produce scientifically valid investigations. Financial constraints should not compromise the need for scientific rigor. A genuine concern exists that such constraints have led some journals and reviewers to accept the argument that proper experimental design and true replication is logistically infeasible and therefore should not be required for publication of the observations made. Yet, as discovery moves from the description of apparent diversity to the genuine description of complexity and function, this is no longer acceptable or desirable.

Is it possible that metagenomics has failed to deliver what it promised—a fast, cheap, and comprehensive method to explore functional biochemistry in the natural world? It is too early to reach this conclusion, but several factors led to this perception, including underestimation of the complexity of microbial diversity, limited data concerning the source of each sample and the identity of many genes, difficulties in integrating and comparing results obtained with different technologies in different labs, mismatched expectations between researchers who sought to generate understanding of ecological patterns, with those who were excited to test the limits of

new technology, and the lack of agreed upon data standards. For the discovery of enzymes such as glycoside hydrolases[5] (important for biomass breakdown), information on the type of biomass, biological or physicochemical pretreatment (e.g. grinding of biomass by wood feeding insects), redox conditions, pH and temperature are important parameters to record. If we continue to develop these environmental data checklists for other types of sample sets, it will be feasible to search for relevant genes in databases created by metagenomic endeavors, which will greatly assist in finding genes relevant to a target biotechnology application.

To change perspectives, national and global cooperation is needed to adopt minimum standards in experimental design and to convince funding agencies to make the appropriate levels of investment. Initial advances toward novel gene discovery using metagenomics relied on direct cloning and sequencing of DNA fragments extracted from uncultured microbial communities. Although an important step forward, these methods were also time consuming and expensive. For example, metagenomic data generation for the first leg of the Global Ocean Sampling expedition was estimated to cost > $10 million. Although costly, the dataset is comparatively limited by today's standards. Since the introduction of the first wave of 'next-generation' highly parallel DNA sequencers in 2006, there has been an explosion in gigabase- to terabase-scale metagenomic sequencing projects[6]. An illustrative, though not exhaustive, list includes the continued Global Ocean Survey (GOS), International Census of Marine Microbes, MetaHIT, the Human Microbiome Project (HMP), TARA Oceans, DeepSoil, MetaSoil, Genomic Observatories[7], the JGI Great Prairie pilot study, and the National Ecological Observatory Network (NEON).

Pioneering metagenomic studies of microbial community composition and function in different environments (e.g., acid mine drainage[8], soil/permafrost[9, 10], marine GOS[11], Hawaiian ocean time series[12], Western Channel Observatory L4[13], termite hindgut[14], cow rumen[15], human gastrointestinal tract[16], and mouse gastrointestinal tract[17]) provided a first glimpse into the potential of this approach to uncover previously unknown functional genes, phylogenetic types, and interactions among community members. Indeed, comparative metagenomic analyses have yielded considerable insight into the distribution of gene families across different ecosystems, and the role of specific functional attributes in adaptation to physical and chemical conditions[18-20]. However, these initial studies were limited by their status as pilot studies, often due to the need to develop and prove the technologies and the high cost of sequencing. Therefore, most of these studies were observational and were not able to adopt the normal scientific

methodological approach of well replicated coverage of the respective ecosystems for statistically relevant analyses[21] of the biological variation.

Now that sequencing costs have declined as throughput has dramatically increased, we expect, without any reasonable exceptions, rigorous experimental design to be applied to future metagenomics experiments. Further, we must take full advantage of this brave new world of rigorous metagenomic study design by thinking like cartographers, and creating a map that can be used to navigate the uncharted regions of the microbial universe. One example of this map could be a catalogue of all known proteins and the environments (including comprehensive metadata) in which they were found. To do this, it will be necessary to better characterize individual ecosystems with prolonged and in-depth investigations, comprehensive physical, chemical and biological contextual data, appropriate statistical design, and improved interpretation of functional and taxonomic characteristics (**Box 1**).

A metagenomic dataset is only as good as the contextual experimental and environmental data associated with it. Just as maps require a standard format to enable comparability, in-depth investigations also must be comparable, and be able to be linked, to uncover what features are common to multiple systems, or specific to each system. Standardization efforts enable further analyses, such as determining the distribution of these elements across time and space, thereby improving our understanding of microbial dynamics across planet Earth.

Defining the playing field through shallow and deep surveys

Ultra-deep sequencing of taxonomic or functional marker genes such as the small subunit ribosomal RNA gene (SSU-rRNA) or *nif*H has enabled comprehensive cataloging of the inhabitants of a variety of microbial ecosystems[22-26]. Deep sequencing of a few samples can provide information about rare taxa and rare genes, but without analyzing larger numbers of samples, limitations arise: the statistical significance of observed patterns cannot be determined, the patterns of co-occurrence between genes and taxa are difficult to assess, and the dominant biotic or abiotic factors structuring communities across time and space remain undetermined. As an analogy, if naturalists in the 19th century had only focused on plant and animal diversity in a few, isolated plots instead of exploring ecosystems across broad swaths of the globe, the fields of botany and zoology would have reached a standstill, and the global patterns of biogeography, which were crucial to forming our modern understanding of ecology

and evolution, would have remained unknown. Thus, for microbial biogeography, many samples from related or contrasting communities must be studied in parallel.

We recognize the recent advances that have been made by the deep sequencing of a few samples (e.g. generating billions or trillions of base pairs from a single sample). Indeed, broad, shallow sequencing from many thousands of samples can help to direct which samples should be analysed in more detail using deep sequencing, which enables additional data analyses that may lead to better interpretation of the biological information. For example, in order to obtain enough information to allow reliable assembly of specific genomic fragments (using currently available sequencing technologies), deep sequencing of random shotgun DNA is essential. Recent work on rumen samples obtained from two cows illustrates this point. Hess and colleagues[15] were able to assemble 15 near-complete bacterial genomes from short-read length shotgun sequencing data. However, Improved coverage is not the only answer, but can help to focus the question; for example, using a rough calculation of 4 Mbp (mega base pairs) per genome and a billion cells per gram, a single gram of soil could contain up to 3 Pbp (peta base pairs) of genetic data. Recently, Mackelprang et al.[9] used deep sequencing to successfully assemble a draft genome of a novel methanogen from highly diverse permafrost soil. Therefore, although soil is one of the most challenging ecosystems for metagenomics because of it's high diversity, advances in new assembly algorithms show great promise for genome reassembly from deep sequence studies [27].

The question of whether to sequence deeply or shallowly across many samples is dependant on the question you want to answer. Deep sequencing is required to observe rare members of microbial communities. Regardless of the habitat in question, rare members of the community can have key functional roles, such as nutrient cycling (e.g., methanogenesis[28], nitrogen fixation[26]), pathogenesis, stimulation of the immune system, and metabolite production (e.g., butyrate in the gut, or antibiotics). Moreover, microbes that are rare in one sample may be common in another. For example, in the European Meta-HIT project, metagenome sequences from fecal samples were obtained from 124 individuals, and the human gut microbes identified as being shared between individuals varied 8- to 1500-fold among different hosts[29].

Shallow sequencing, in contrast, enables the exploration of microbial community structure dynamics, which is fundamental to building a predictive understanding of an ecosystem[30]. Recent evidence suggests that some ecosystems maintain a temporally persistent but vast

microbial seed bank[31], suggesting that taxa identified by shallow surveys are merely indicative of the abundant taxa selected by the chemical, physical and biological processes leading up to and present at the time of sampling. However, one likely hypothesis states that "the dominant microorganisms in a sample are those that play the most important functional roles under normal conditions." Hence, if one is interested in the ecology of more abundant processes or taxa, ultra-deep metagenomic sequencing is not essential; relatively small fractions of the genetic diversity contained within samples can reveal ecological patterns that help define ecosystem structure[13]. The potential for reliance on shallow sequence data (either amplicon or shotgun) for some studies is supported by a study of gnotobiotic mice harbouring a defined consortium where the complete genome sequence of every community member was known. In that study it was possible to obtain accurate descriptions of the community's meta-transcriptome and meta-proteome based on short sequence reads[32].

Creating a highly detailed picture of an individual or environmental sample from one angle at one instant creates a static view of that sample that can be useful. However, it cannot capture temporal dynamics, or variability among individuals or habitats. Far more is gained from complementing such pictures with others, even if these others are taken at lower resolution, as it permits more accurate reconstruction of shape. Likewise, low-resolution pictures taken successively over time can provide a sense of motion and dynamics and low-resolution pictures of many different samples can provide a view of diversity and variability that cannot be obtained by a single sample. However, all these pictures or individual snap-shots must be well organized, as it is of little value to have them unsorted in a pile that prohibits retrieval of the series of the data sets, or images, necessary to reconstruct a view of a specific phenomenon under study.

To determine dynamic processes, it is necessary to apply broad sampling (both in time and space) at an appropriate resolution to determine the frequency of the dynamics. With most studies, an increase in the number of samples analyzed has a significant impact on analytical power (**Table 1**). Gilbert and colleagues[33] generated a 12-sample survey of the annual changes in the microbiota of surface waters in the English Channel, and found evidence for seasonal succession driven by temperature and nutrient availability. However, when they augmented this with 60 more samples, making a contiguous 72 sample time series over six years[22], the patterns were significantly refined, with the seasonality being extremely robust, and day-length being identified as the key driver of richness in the community (**Figure 1; Table 1**). Additionally, Arumugam and colleagues[34] used metagenomic sequencing from 22 individuals to show that

human gut microbiota could be classified into 3 enterotypes, which showed no correlation to diet or ethnicity. However, Wu et al.[35] performed the same analysis on 98 individuals and demonstrated that the increase analytical power found distinct correlations with diet (**Table 1**). Other examples of the power of sampling breadth can be routinely found in the literature (**Table 1**), and they demonstrate that using statistically relevant experimental design is vital to generating accurate analyses.

Defining the effect size and the power of a study is a particularly important challenge in the design of clinical microbiome-directed trials (e.g. probiotics, prebiotics, antibiotics and stool transplants) or the natural or man-made disturbance in any terrestrial or oceanic ecosystem. A recent attempt to define effect sizes in studies of the human microbiome[36] foundered due to the lack of comparability of different datasets and methodologies for taxon detection and assignment. Such effect sizes can only be determined with sufficiently large sample sizes of "normal" versus "altered" states, studied over sufficient temporal and spatial scales to reveal variation. The dilemma, especially for human studies, is that large samples are required to determine effect size, but such studies cannot gain Institutional Review Board approval because the effect size, and therefore the correct number of subjects required to achieve statistical power, is unknown.

Towards an Earth Microbiome Project

In recognition of the value of a multi-environmental survey of microbial diversity, we have instigated an initiative called the Earth Microbiome Project (EMP; www.earthmicrobiome.org). The EMP seeks to systematically characterize microbial taxonomic and functional biodiversity across global ecosystems, and to organize international environmental microbiology research by standardizing the protocols used to generate and analyze the data between studies. The Earth Microbiome Project (EMP) constitutes a restructuring and refocusing of microbial ecology. Individual projects are grouped (by single PI, or by consortium) into overarching science questions that can be used to define the fundamental purpose of a single project, or individual hypothesis-driven studies can be grouped under a larger question. While this framework provides a way to influence and globally organize environmental microbiology research, the novelty lies in the sheer scale of the endeavor and the standardization of the protocols used to generate and analyze the data between studies. The EMP standard operating procedures (SOPs) define a route to minimize bias between community analyses associated with different

material extraction techniques, analytical methods and core data quality control and analysis. However, currently, the EMP does not promote a standard physical sample acquisition protocol or preservation technique, but is working to explore the impact of these variables on ecological interpretation[37]. The EMP framework promotes open access research; hence all data is being made public, including to industry, and comparable within an open access forum, which creates a data resource capable of answering and asking fundamental questions about the function of microbes in different environmental habitats. However, it is not just data that must be open access. The scientists themselves also need to be more accessible through open science initiatives[38], ensuring that the right researchers are able to work on the most relevant topics, making the best use of reductionist expertise.

Additionally, the EMP framework enables multidisciplinary cooperation across funding agencies and scientific research areas. Stand-alone projects are mapped onto larger research themes, and these stack into overarching global questions, yielding multiple layers and scales of inquiry. This focus on multidisciplinary activity brings new dimensions to microbial investigation, through renewed interest in data processing, requirements for large-scale computational infrastructure, modeling community dynamics and functional capability, and linking the analyzed data and generated models to climate modeling informatics programs. It also merges aspects of biogeochemistry, microbiology, protein/enzyme interaction, and transcriptional feedback as we move from molecular scale processes to processes and dynamics on other scales. These range from cellular interaction, to community ecology, local, regional, national, continental and global scales. Such a broad knowledgebase will be critical for developing a predictive understanding of genes and organisms of biotechnological interest.

Of course, for large scale sequencing efforts such as the EMP to be focused and coordinated, the community must avoid the “sequence everything” approach, simply because “we can.” Hypotheses must guide our selection of the most appropriate samples to sequence. To a large extent these will be sample sets that have rich metadata, and samples that have the potential to provide fundamental new knowledge.

The role of metadata acquisition in improved experimental design

Initiatives like the EMP are saved from becoming simple natural history exercises in data collection by the requiring the acquisition and appropriate organization of the metadata that

accompany every sequence dataset generated. These environmental and experimental metadata are the primary data of many multidisciplinary research groups, who already work together to generate a comprehensive understanding of a particular environment, e.g. a marine sampling field expedition, or a temporal exploration of soil and ecosystem dynamics in one location. Such environmental parameters give context to the origin of the sequence data we rely upon to generate interpretative analyses about the microbial dynamics in that ecosystem. They include temperature, latitude and longitude, altitude, moisture content, nutrient concentrations, and standard ontologies for geolocators and ecosystem descriptors. But these must also be accompanied by experimental metadata that appropriately describe the methods used to create the sequence data, such as sample handling, nucleic acid extraction, PCR amplification method, sequence protocol, and bioinformatic analysis. Acquisition of these metadata are essential to the EMP, as they provide ecological grounding to analyses of the taxonomic and functional capacity of the sequenced microbial community. Hence, this robust framework for routine collection of metadata and reliable standards will enable comparison between studies. A suite of such standard languages is provided by the **M**inimum **I**nformation about any (**x**) **S**equence checklists (MIxS[39]). MIxS is an umbrella term to describe MIGS, MIMS and MIMARKS[3] and contains standard formats for recording environmental and experimental data.

The latest of these checklists, MIMARKS (Minimum Information about a MARKer Sequence) builds on the foundation of the MIGS (the Minimum Information about a Genome Sequence) and MIMS (the Minimum Information about a Metagenome Sequence) checklists[3], by including an expansion of the rich contextual information about each environmental sample. What is recorded depends on where the sample comes from. For example, human samples can be annotated with fields such as the age, weight, and health status of the subject, whereas seawater samples can be annotated with fields such as pH, salinity, depth and temperature. Additionally, detailed technical information such as the sequencing platform, and the genes and regions targeted are also required, making meta-analyses of many studies much easier to perform and interpret, because outliers can be traced back to technical differences or to biological differences automatically, rather than requiring the researcher to read scores of papers as is necessary for meta-analyses today[40]. This integration is especially important for finding enzymes that participate in processes that are potentially industrially useful but where the origin is irrelevant to the industrial application except for improving the possibility that the enzyme will work under the necessary conditions.

We believe that the MIxS standard will play a key role for three reasons: *First*, it will enable large-scale projects to collect massive datasets according to standard protocols at multiple sites, and to share these data to facilitate global understanding. *Second*, it will enable integration of each lab's individual projects into this universe of sequences, allowing community-level comparisons, unprecedented exploration of the diversity and distribution of life, easy detection and exclusion of contaminated samples, and the exploration of gene or taxon co-occurrence patterns. These features are especially crucial for accessing and integrating data from every clinic or every field site. *Third*, it will provide a framework for large-scale integration of efforts, especially predictive modeling. Stanislaw Ulam said, "Great scientists see analogies between theorems or theories. The very best ones see analogies between analogies." By providing a method of integrating both the systematically collected results of large-scale projects such as the EMP and the highly distributed efforts of smaller groups, standards such as MIxS will help enable a future in which analogies across spatial scales, temporal scales, and even theories are not only possible but routine.

As the cost of sequencing continues to decline, there has been a rapid adoption of the MIxS standard, and of sound sampling principles. For example, tools such as QIIME[41] and MG-RAST[42] are already MIxS-compliant and provide ways of viewing and analyzing MIxS-compliant data. INSDC has committed to incorporating a MIxS keyword as a standard, and large projects such as the HMP (https://commonfund.nih.gov/hmp/), NEON (http://www.neoninc.org/), the EMP (www.earthmicrobiome.org), the Bio Weather Map (http://bioweathermap.org/), and the Personal Genome Project (http://www.personalgenomes.org) have already pledged to support the standard. This rapid response is timely. As sequencing and computational methods co-evolve in a dynamic 'arms race' that spurs their mutual growth and progress, so too must data standards co-evolve.

International activities such as the EMP provide test beds to allow the community to agree on standards for exchange of data products that go well beyond the trading of consensus sequences and annotations (e.g., GenBank). Even given the expected advances in cloud computing and the predicted decrease in computation costs according to Moore's law, one main driver of innovation will be the need to provide analyses of datasets that are orders of magnitude larger without the corresponding need for vast increases in the bioinformatics budget. Investments in data reuse and usable data standards are critical. However, it is easier to create standards than it is to successfully promote their use. The Genomic Standards

Consortium (GSC) has conducted pioneering work on minimum information checklists that have enabled provenance standards, and it is now taking on the much more complicated task of defining standards for computed data products. In this regard, journals can play a role by universally adopting such standards as a requirement for accepting and publishing manuscripts.

The role of data generation in the discovery of novel enzymes and phylogenetic structure in microbial biodiversity must be complemented by improved functional and taxonomic databases that more appropriately represent the full breadth of microbial diversity. One critical aspect of this development will be mapping of metagenomic reads against reference genomes. The Earth Microbiome Project is partnered with the Genomic Encyclopedia of Bacteria and Archaea and Microbial Earth initiatives[43] that aim to improve the phylogenetic representation of sequenced genomes. These efforts combined with improved gene and protein database curation (e.g., IMG and IMG/M[44, 45]) will aid with metagenomic data interpretation, facilitating more efficient biodiscovery.

Conclusion

As it occurred with many other technologies such as computing, telecommunications and photography (which, like sequencing, began with scientific applications but rapidly transformed consumers' lives across the globe), metagenomics is in a time of transition. The community is moving from a situation in which technologies are first deployed centrally by large organizations, then by departments, by individual laboratories, and it is perhaps not unreasonable to speculate that sequencing devices will soon be owned by individuals, perhaps even in a handheld format. Standard protocols are necessary to integrate the information and to allow easy communication across studies—after all, the role played by the internet in today's world is only possible because computers everywhere can communicate with a set of standard, open protocols. While currently these initiatives are focused on DNA sequencing (amplicon sequencing and metagenomics), it will be necessary to determine integration of metabolomics, proteomics and single-cell genomics into these efforts to improve community characterization, and enable more appropriate ecological inferences. The 'omics ratio (ratio of applied techniques, e.g. genomics:transcriptomics:proteomics:metabolomics) should always be determined by the hypothesis. We believe and hope that MIxS and the EMP will enable the same type of functionality for ecologists, allowing us to construct not just a catalog of organisms on Earth but

also to understand and exploit the critical processes they perform in the environment over a vast range of spatial and temporal scales.

Acknowledgements: We wish to thank Jonathan Eisen for his constant support of the Earth Microbiome Project, and his help with writing this work. This work was in part supported by the U.S. Dept. of Energy under Contracts DE-AC02-06CH11357 and DE-AC02-05CH11231, the National Institutes of Health, the Natural Environment Research Council, UK the Crohns and Colitis Foundation of America, and the Howard Hughes Medical Institute. We thank Jens Reeder, Jesse Stombaugh, Cathy Lozupone, Daniel McDonald, Justin Kuczynski, and Jessica Metcalf for comments on drafts.

Tables 1A, B, C, D: Recent studies, number of samples, and reported results. Studies with more samples have a higher impact and clearer biological interpretations than studies with comparable amounts of sequencing but spread over fewer samples: the reason is ability to correlate information with biological or clinical parameters of the system. Three comparisons of successive studies are shown: *Table 1A - blue – marine; Table 1B - brown – human gut; Table 1C - green – human skin; Table 1D - orange – soil*.

Study	Number of samples	Sequencing target	Key results
Gilbert et al., Environmental Microbiology, 2009 [33]	12 monthly marine samples	16S RNA V6	Evidence of seasonally structured community diversity and for seasonal succession, significantly correlated to a combination of temperature, phosphate and silicate concentrations.
Gilbert et al., ISME J, 2011[22]	72 monthly marine samples	16S rRNA V6	Community had strong repeatable seasonal patterns, with winter peaks in diversity. Change in day length explained 65% of the diversity variance. The results suggested that seasonal changes in environmental variables are more important than trophic interactions. Relationships between Bacteria were stronger than with Eukaryotes or environment. The increase in temporal sampling over Gilbert et al., 2009, increased the capability to explore community relationships.
Zinger et al., PLoS ONE, 2011[46]	509 marine samples	16S rRNA	High variability of bacterial community composition specific to vent and coastal ecosystems. Both pelagic and benthic bacterial community distributions correlate with surface water productivity. Also, differences in physical mixing may play a fundamental role in the

			distribution patterns of marine bacteria, as benthic communities showed a higher dissimilarity with increasing distance than pelagic communities.

Study	Number of samples	Sequencing target	Key results
Arumugam et al., Nature, 2011[34]	22 human fecal samples	Metagenomes	Identification of three clusters (enterotypes) that are not nation or continent specific. Certain genes or functional groups do show correlation to certain host factors.
Muegge et al., Science, 2011[47]	51 mammalian fecal samples	16S rRNA Metagenomics	Fecal DNA from 33 mammalian species and 18 humans who kept detailed diet records, and we found that the adaptation of the microbiota to diet is similar across different mammalian lineages. Therefore, this study did not support the study of Arumugam et al., 2011.
Wu et al, Science, 2011[35]	98 human fecal samples	Metagenomes	Enterotypes were strongly associated with long-term diets, particularly protein and animal fat (*Bacteroides*) versus carbohydrates (*Prevotella*). Therefore, this study did not support the study of Arumugam et al., 2011; the increased breadth of the study improved the analytical capability.
Qin et al., Nature, 2010[29]	124 fecal samples	Metagenome	Over 99% of the genes are bacterial, most found in every sample, and indicate that the entire cohort harbors ~1,000 prevalent bacterial species. Each individual has at least 160 species, which are also largely shared.
Claesson et	170 fecal	16S rRNA	The fecal microbiota of the

al., PNAS, 2011[48]	samples		elderly shows temporal stability over limited time in the majority of subjects but is characterized by unusual phylum proportions and extreme variability.
Frank et al., PNAS, 2007[49]	190 human gut samples	16S rRNA	Statistically significant differences between the microbiotas of Crohn's Disease (CD) and ulcerative colitis (UC) patients and those of non-IBD controls. Significantly, our results indicate that a subset of CD and UC samples contained abnormal gut microbiotas.
Turnbaugh et al., Nature, 2009[50]	154 humans: fecal samples (twin pairs and mothers)	16S rRNA, shotgun	Identifies a core microbiome at the gene function but not the organismal lineage level; identifies systematic differences in diversity between lean and obese. Supported by Aruguman et al. 2011 on the obesity alpha diversity result.
Reyes et al. Nature 2010[51]	36 individuals: fecal samples (twin pairs and mothers, over 1 year)	16S rRNA, shotgun, viruses	Shows high levels of variability between individuals, magnitude of viral diversity, and absence of "kill-the-winner" dynamics.

Study	**Number of samples**	**Sequencing target**	**Key results**
Costello et al., Science, 2009[52]	27 body sites in 9 individuals	16S rRNA	Community composition was determined primarily by body habitat. Within habitats, interpersonal variability was high, whereas individuals exhibited minimal temporal variability. Several skin locations harbored more diverse

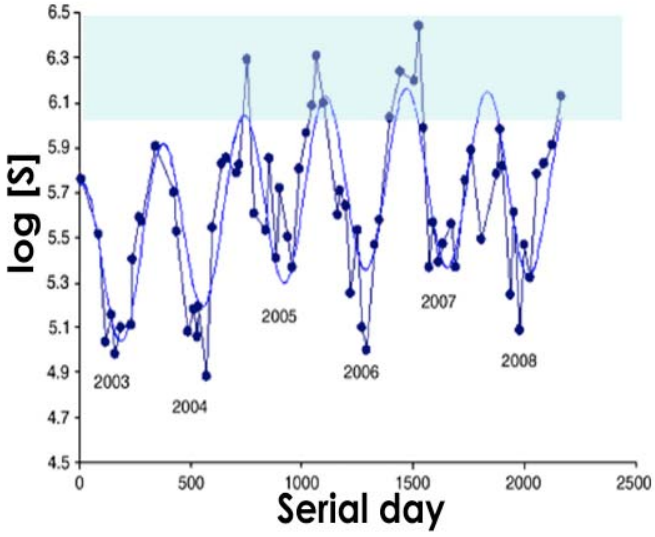
			communities than the gut and mouth, and skin locations differed in their community assembly patterns.
Fierer et al., PNAS, 2010[53]	90 keyboard keys 30 phalange skin	16S rRNA	Structure of microbial communities can be used to differentiate objects handled by different individuals, even if those objects have been left untouched for up to 2 weeks at room temperature.
Caporaso et al., Genome Biology, 2011[24]	396 time points for four body sites	16S rRNA	Despite stable differences between body sites and individuals, there is variability in an individual's microbiota across time. Only a small fraction of taxa are temporally persistent, hence no core temporal microbiome exists at high abundance. Strikingly, this study confirmed the results of a previous study (Costello et al., 2009) with a massive increase in data.
Ravel et al., PNAS, 2011[54]	396 vaginal swabs	16S rRNA	Patterns were associated with the diagnosis of bacterial vaginosis. The inherent differences within and between women in different ethnic groups strongly argues for a more refined definition of bacterial communities normally found in healthy women.

Study	**Number of samples**	**Sequencing target**	**Key results**
Rasche et al., ISMEJ, 2010[55]	72 soil samples	16S rRNA tRFLP	Seasonal dynamics displayed by key phylogenetic and nitrogen (N) cycling functional groups were found to be tightly coupled with seasonal alterations in labile C and N pools as well as with variation in soil temperature and soil moisture.

Mackelprang et al., Nature, 2011[9]	12 soil samples (permafrost & active layer - before & after thaw)	Metagenomes	Permafrost thaw caused a rapid shift in several phylogenetic and functional genes and C and N cycling pathways. A draft genome of a novel methanogen was assembled from the metagenome data.

Figure 1: Conceptual diagram of why replicated samples, especially across a gradient or along a time series, are critical for interpretation of results. Structure that is externally imposed via study design greatly improves our ability to recover biologically meaningful relationships rather than simply finding statistical differences between samples (especially important because every pair of biological samples will be different if sequenced deeply enough). In this case, we show the L4 Western English Channel ocean time series samples [22]: Sampling only during the summer, highlighted in blue, would only reveal the tip of the iceberg of variability in this ecosystem, which is driven by seasonal change (the graph shows day on the x-axis; log of the observed number of species on the y-axis). Similar principles apply in other ecosystems that have other major drivers of variation that, when overlooked, can influence the results in ways that are puzzling, or give a misleading picture of variation.

Figure 2: Importance of metadata-enabled studies. Matched-pair diagrams showing visualizations from recently published, high-impact studies with and without metadata, showing the importance of metadata. Examples taken from Costello et al. 2009[52] (PCoA plot of UniFrac distances between human body habitat associated communities reveals clustering by human body habitat type), Ley et al. 2008[56] (where a bipartite network diagram shows that the main clustering of mammalian fecal communities is by diet), and Fierer et al. 2010[53] (where an NMDS plot of UniFrac distances between soil communities shows that the main factor driving variation in these communities is pH). These relationships are immediately and intuitively obvious when the right metadata are applied, but would be essentially impossible to see otherwise.


log [S]
6.5
6.3
6.1
5.9
5.7
5.5
5.3
5.1
4.9
4.7
4.5
0
500
1000
1500
2000
2500
Serial day
2003
2004
2005
2006
2007
2008

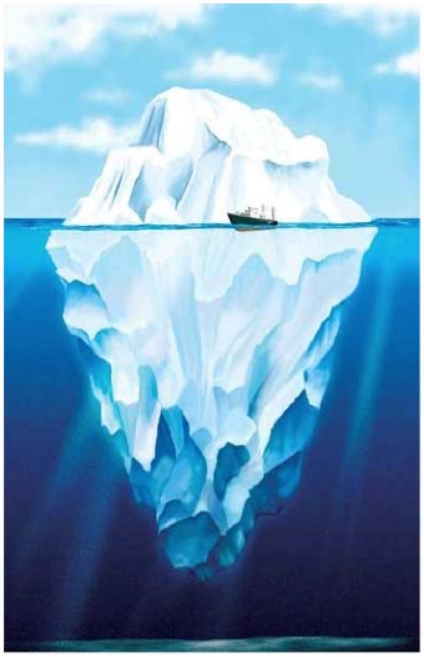
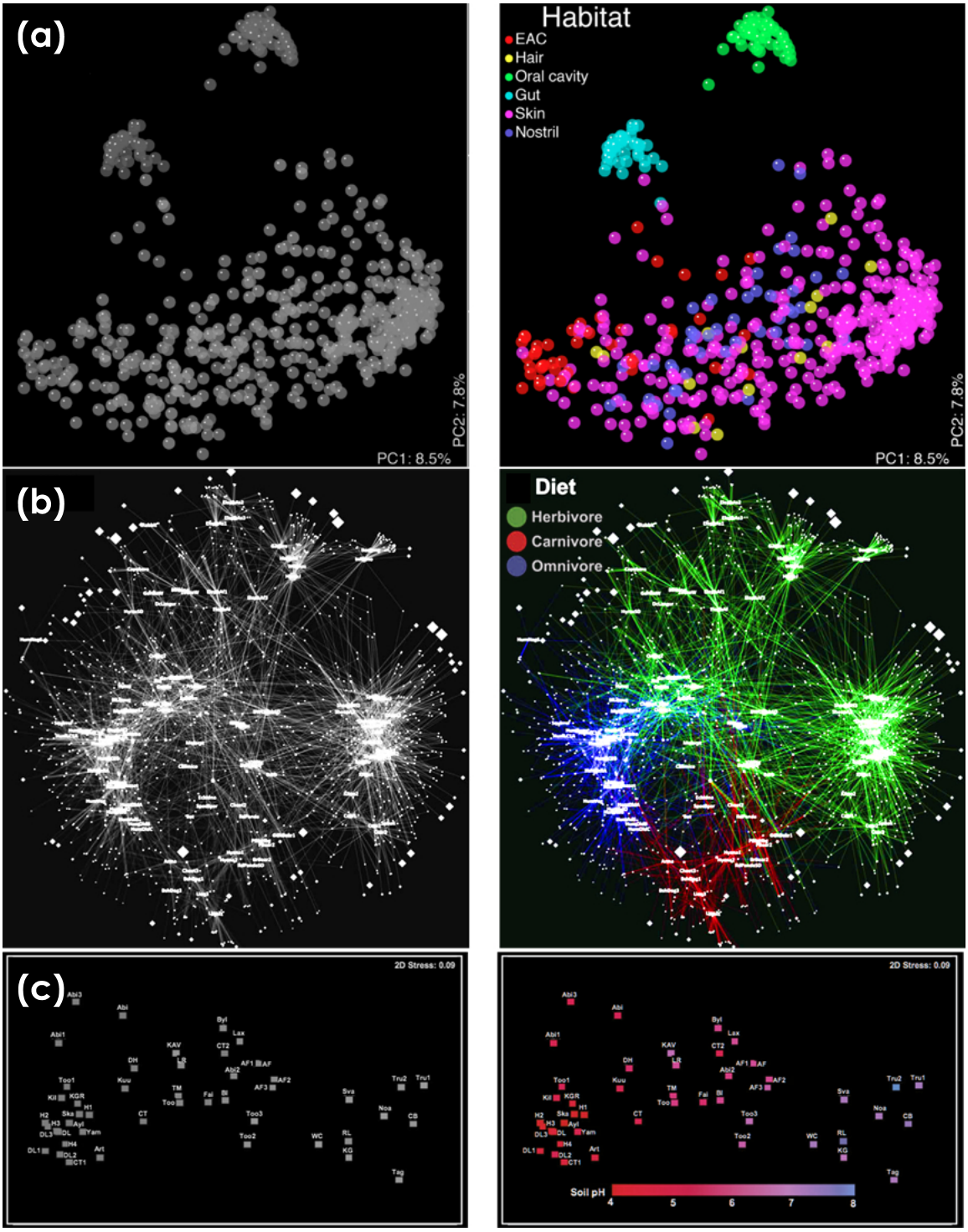


Figure 1

Figure 2


(a)
PC1: 8.5%
PC2: 7.8%
Habitat
EAC
Hair
Oral cavity
Gut
Skin
Nostril
PC1: 8.5%
PC2: 7.8%
(b)
Diet
Herbivore
Carnivore
Omnivore
(c)
2D Stress: 0.09
2D Stress: 0.09
Soil pH
4
5
6
7
8

Box 1: Key decisions in the metagenomic pipeline that affect the utility of the data and ability to leverage existing and future studies in its interpretation.

Challenge	Decision	Pitfall	Consequence
Biological and technical replicates are expensive and time-consuming	Whether to perform replication, or gamble that a single sample in each group is informative with sufficiently well-described ecosystem parameters	Often non-replicated designs are not interpretable, or are over interpreted (e.g. attributing differences in a single healthy versus diseased person to the disease)	Conclusions cannot be replicated by other researchers, and may not be generalizable beyond the specific samples analyzed
A fixed sequencing budget must be divided among some number of samples (e.g. by multiplexing at some level)	Whether to sequence few samples deeply, or many samples more shallowly	The appropriate number of samples and sequencing depth are unknown	Few samples may be uninformative and may preclude informative analysis of variation in the system and/or replication; shallow sequencing may miss rare but important taxa or functions
Experimental challenges due to low yield of DNA and/or high community diversity	Whether to adopt new protocols for improved DNA extraction, amplification and/or assembly	DNA extraction and manipulation steps all introduce biases that may make it difficult to compare between studies	For unique or rare samples that require special treatment it is essential that all steps in the treatment are considered if comparing results to

			those from other studies.
Defining the dimensions of variation that matter in a given system is challenging, and often is the purpose of the study itself	Which scales and parameters to select, and how much variation to cover	"Extremes" of variation in the system being studied are expensive and difficult to obtain (tail of distribution) and may not even be extreme from the microbes' perspective; relevant variation often unknown	Conclusions from one population or study site inappropriately generalized to other populations or study sites; relevant variation in system undiscovered; extreme efforts to obtain exotic samples are unrewarded
Must choose a sequencing platform	Trade-off between read length and number of sequences; must decide when to adopt new technology	All sequencing technologies and processing pipelines have drawbacks, not all of which are widely advertised; technology changes rapidly	Sequences may be too short, too few too error-prone to interpret, or too passé to publish
Interpretation of sequence data	Must decide whether to use reference-based or de novo methods for assembly, taxonomy and functional assignment, and if so which reference to use	Different reference databases give different results; de novo is unbiased but far less powerful when appropriate references exist; analyses differ as reference databases update rapidly, limiting comparisons	Incorrect and/or hard-to-reconcile functional and taxonomic assignments

		between studies. Current assembly algorithms are insufficient for highly complex metagenome data.	
Metadata collection	Must decide what metadata (i.e. sample or site data) to collect and associate with sample	Too complex to be implemented; fields inconsistent with previous studies due to lack of standards-compliance; data model can't accommodate	Chaos!
Centralization	Whether to centralize sample collection, metadata curation, DNA extraction, sequencing, data storage, and data analysis	Decentralization can lead to inconsistencies that make data difficult to interpret; centralization can lead to delays while funding is acquired or capacity is built, and can limit creativity	Either the dataset may be vast but too inconsistent to interpret, or it may be extremely consistent but limited in scope and/or interpretation. Specific considerations apply to each stage; the EMP currently favors decentralized sample collection and centralization of other steps on a case-by-case basis

References:

1. **Whitman, W.B., Coleman, D.C. & Wiebe, W.J. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* 95, 6578-6583 (1998).**
2. **Falkowski, P.G., Fenchel, T. & Delong, E.F. The microbial engines that drive Earth's biogeochemical cycles. *Science* 320, 1034-1039 (2008).**
3. **Field, D. et al. The Genomic Standards Consortium. *Plos Biology* 9 (2011).**
4. **Fisher, R.A. The Design of Experiments. (1935).**
5. **Gilbert, J.A. et al. in Bioprospecting using metgaenomics. (ed. M. Himmel) (Springer, 2011).**
6. **Jansson, J. Towards "Tera-Terra": Terabase Sequencing of Terrestrial Metagenomes. *ASM MIcrobe Magazine* (2011).**
7. **Davies, N., Field, D. & Genomic Observatories, N. Sequencing data: A genomic network to monitor Earth. *Nature* 481, 145 (2012).**
8. **Tyson, G.W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37-43 (2004).**
9. **Mackelprang, R. et al. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480, 368-371 (2011).**
10. **Delmont, T.O. et al. Structure, fluctuation and magnitude of a natural grassland soil metagenome. *Isme J* (2012).**
11. **Rusch, D.B. et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5, e77 (2007).**
12. **DeLong, E.F. et al. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311, 496-503 (2006).**
13. **Gilbert, J.A. et al. The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS One* 5, e15545 (2010).**
14. **Warnecke, F. et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450, 560-565 (2007).**
15. **Hess, M. et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 331, 463-467 (2011).**
16. **Gill, S.R. et al. Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355-1359 (2006).**
17. **Turnbaugh, P.J. et al. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027-1031 (2006).**
18. **Delmont, T.O. et al. Metagenomic mining for microbiologists. *Isme J* 5, 1837-1843 (2011).**
19. **Dinsdale, E.A. et al. Functional metagenomic profiling of nine biomes. *Nature* 452, 629-632 (2008).**
20. **Tringe, S.G. et al. Comparative metagenomics of microbial communities. *Science* 308, 554-557 (2005).**
21. **Prosser, J.I. Replicate or lie. *Environ Microbiol* 12, 1806-1810 (2010).**
22. **Gilbert, J.A. et al. Defining seasonal marine microbial community dynamics. *Isme J* 6, 298-308 (2012).**

23. Caporaso, J.G. et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 108 Suppl 1, 4516-4522 (2011).
24. Caporaso, J.G. et al. Moving pictures of the human microbiome. *Genome Biol* 12, R50 (2011).
25. Parsons, R.J., Breitbart, M., Lomas, M.W. & Carlson, C.A. Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *Isme J* (2011).
26. Farnelid, H. et al. Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLoS One* 6, e19223 (2011).
27. Desai, N., Antonopoulos, D., Gilbert, J.A., Glass, E.M. & Meyer, F. From genomics to metagenomics. *Current Opinion in Biotechnology* 23, 72-76 (2012).
28. Thauer, R.K., Kaster, A.K., Seedorf, H., Buckel, W. & Hedderich, R. Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat Rev Microbiol* 6, 579-591 (2008).
29. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59-65 (2010).
30. Larsen, P., Field, D. & Gilbert, J.A. Predicting bacterial community assemblages using an artificial neural network approach. *Nat Methods* Published Online 15th April 2012 (2012).
31. Caporaso, J.G., Paszkiewicz, K., Field, D., Knight, R. & Gilbert, J.A. The Western English Channel contains a persistent microbial seed bank. *Isme J* (2011).
32. Mahowald, M.A. et al. Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. *Proc Natl Acad Sci U S A* 106, 5859-5864 (2009).
33. Gilbert, J.A. et al. The seasonal structure of microbial communities in the Western English Channel. *Environ Microbiol* 11, 3132-3139 (2009).
34. Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* 473, 174-180 (2011).
35. Wu, G.D. et al. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science* (2011).
36. Kuczynski, J. et al. Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol* 11, 210 (2010).
37. Gilbert, J.A. et al. The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010. *Stand Genomic Sci* 3, 249-253 (2010).
38. Woelfle, M., Olliaro, P. & Todd, M.H. Open science is a research accelerator. *Nat Chem* 3, 745-748 (2011).
39. Yilmaz, P. et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotech* 29, 415-420 (2011).
40. Lozupone, C.A. & Knight, R. Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A* 104, 11436-11440 (2007).
41. Caporaso, J.G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7, 335-336 (2010).
42. Meyer, F. et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9, 386 (2008).
43. Wu, D. et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462, 1056-1060 (2009).
44. Markowitz, V.M. et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* 40, D115-122 (2012).

45. **Markowitz, V.M. et al. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* 40, D123-129 (2012).**
46. **Zinger, L. et al. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS One* 6, e24570 (2011).**
47. **Muegge, B.D. et al. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 332, 970-974 (2011).**
48. **Claesson, M.J. et al. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc Natl Acad Sci U S A* 108 Suppl 1, 4586-4591 (2011).**
49. **Frank, D.N. et al. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A* 104, 13780-13785 (2007).**
50. **Turnbaugh, P.J. et al. A core gut microbiome in obese and lean twins. *Nature* 457, 480-U487 (2009).**
51. **Reyes, A. et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466, 334-338 (2010).**
52. **Costello, E.K. et al. Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science* 326, 1694-1697 (2009).**
53. **Fierer, N. et al. Forensic identification using skin bacterial communities. *Proc Natl Acad Sci U S A* 107, 6477-6481 (2010).**
54. **Ravel, J. et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A* 108 Suppl 1, 4680-4687 (2011).**
55. **Rasche, F. et al. Seasonality and resource availability control bacterial and archaeal communities in soils of a temperate beech forest. *Isme J* 5, 389-402 (2011).**
56. **Ley, R.E. et al. Evolution of mammals and their gut microbes. *Science* 320, 1647-1651 (2008).**