Review

Cell PRESS

Microbial systems biology

Combined phylogenetic and genomic approaches for the high-throughput study of microbial habitat adaptation

Jesse R.R. Zaneveld[1], Laura Wegener Parfrey[2], Will Van Treuren[2], Catherine Lozupone[2], Jose C. Clemente[2], Dan Knights[3], Jesse Stombaugh[2], Justin Kuczynski[1] and Rob Knight[2,4]

[1] Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309, USA
[2] Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309, USA
[3] Department of Computer Science, University of Colorado, Boulder, CO 80309, USA
[4] Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, MD 20815-6789, USA

High-throughput sequencing technologies provide new opportunities to address longstanding questions about habitat adaptation in microbial organisms. How have microbes managed to adapt to such a wide range of environments, and what genomic features allow for such adaptation? We review recent large-scale studies of habitat adaptation, with emphasis on those that utilize phylogenetic techniques. On the basis of current trends, we summarize methodological challenges faced by investigators, and the tools, techniques and analytical approaches available to overcome them. Phylogenetic approaches and detailed information about each environmental sample will be crucial as the ability to collect genome sequences continues to expand.

Setting the stage for high-throughput studies of habitat adaptation

We live in a world suffused with microbial life. Universal trees of life [1,2] constructed by a variety of methods unambiguously show that microbial bacteria, archaea and eukaryotes constitute the vast majority of life's diversity. These diverse organisms perform many important ecological functions across a wide range of natural and man-made environments: photosynthesis in the world's oceans [3]; nitrogen fixation and provision of carbohydrates in association with plant roots [4]; even modification of the chemistry of the upper atmosphere by communities in droplets of cloud-water [5]. The bodies of animals are also colonized internally and externally by microorganisms, which play crucial roles in the development [6], homeostasis [7] and behavior [8] of their hosts.

How have bacteria, archaea and microbial eukaryotes adapted to survive and thrive across such a range of lifestyles and habitats? Understanding the relationship between microbial genome sequence and fitness in a given environment is both a fundamental question in evolutionary biology and a matter of societal importance. As we seek to gain a predictive understanding of phenomena such as the emergence (or re-emergence) of pathogens, the impact of human activities from agriculture to the combustion of fossil fuels on ecosystems, or the effects of dietary or medical interventions on human health (e.g. administration of anti- or probiotics), accurate descriptions of the mechanisms by which microorganisms have adapted to environmental changes in the past will provide crucial guidance.

Traditionally, questions of microbial habitat adaptation have been addressed by experimental manipulation of microbes in pure culture, or by comparisons of genome sequences. More recently, however, large decreases in the cost of sequencing have allowed such approaches to be complemented by the collection of unprecedented quantities of 16S rRNA [9], metagenomic [10,11], transcriptomic [12] and whole-genome data [13]. The 'microbial data deluge' has spurred the development of new computational tools, and has also made possible systematic study of large-scale processes such as habitat adaptation in ways that would have been previously intractable. Here, we highlight how the increasing availability of sequence data from diverse environments is allowing researchers to systematically explore questions about the evolution of habitat adaptation in microbial genomes. We emphasize current trends in the use of tools and analytical approaches, highlighting those that have recently been applied to yield novel insights into this question (Table 1), as well as the outstanding methodological challenges that remain to be overcome.

High-throughput studies of microbial habitat adaptation

It is now well established that the distribution of microbial organisms across different habitat types is correlated with their phylogeny, both in terms of the β diversity of microbial communities [14,15] and the habitat range of individual lineages [16]. For example, 16S rRNA surveys clearly separate bacteria into host and free-living communities; planktonic saline and non-saline communities; and soil

Corresponding author: Knight, R. (rob.knight@colorado.edu).

 0966-842X/$ – see front matter doi:10.1016/j.tim.2011.07.006

Table 1. Links to software and resources discussed in the text

Category	Title	Description	Link	Refs
Ordination	QIIME	Tool for the analysis of community diversity in Python	http://qiime.sourceforge.net/	[70]
	Vegan	A package for the R statistical software tool containing several ordination methods, along with other tools	http://cc.oulu.fi/~jarioksa/softhelp/vegan.html	
	Mothur	Tools for community diversity analysis	http://www.mothur.org/	[71]
Ancestral state reconstruction	PAML	A multipurpose and widely used tool for evolutionary analysis after tree building, including ancestral state reconstruction	http://abacus.gene.ucl.ac.uk/software/paml.html	[72]
	EREM	A tool for ancestral state reconstruction of gene presence/absence	http://carmelab.huji.ac.il/software/EREM/erem.html	[73]
	Ape	A phylogeny package for R that includes functions for estimation of ancestral states	http://ape.mpl.ird.fr/ape_features.html	[74]
	MrBayes	A classic program for Bayesian phylogenetic inference	http://mrbayes.csit.fsu.edu/	[75]
	Mesquite	An extensive graphical suite for phylogenetic analysis	http://mesquiteproject.org/mesquite_folder/docs/mesquite/whyMesquite.html	
	BEAST	A tool for Bayesian phylogenetic inference	http://beast.bio.ed.ac.uk/Main_Page	[76]
Phylogenetic comparative measures	Ade4	Classical multivariate analysis R package, including methods for phylogenetic comparative measures	http://pbil.univ-lyon1.fr/ADE-4/home.php?lang=eng	[77]
	Adephylo	R package; describes phylogenetic signal present in data	http://cran.r-project.org/web/packages/adephylo/index.html	[63]
	Picante	R package containing phylogenetic comparative methods, as well as ordination techniques	http://picante.r-forge.r-project.org/	[64]
HGT detection	PhyloNet	A memory-efficient tool for phylogenetic HGT analysis	http://bioinfo.cs.rice.edu/phylonet/	[78]
	AnGST	Analyzer of gene and species trees	http://almlab.mit.edu/angst/	[33]
	DarkHorse	A distribution based HGT detection tool, with a database of precalculated results for many genomes	http://darkhorse.ucsd.edu/	[79]
	Phangorn	Package for the phylogenetic analysis of horizontal gene transfer	http://cran.r-project.org/web/packages/phangorn/index.html	[80]
Metadata curation	MG-RAST	Analysis, comparison and metadata curation for metagenomic sequences	http://metagenomics.anl.gov/	[43]
	QIIME-DB	A web server for running analysis of community diversity, backed by a large database of well-annotated samples	http://www.microbio.me/qiime	
	EMP submission portal	Submission portal for the earth microbiome project	http://www.microbio.me/emp	[44]
	GOLD	Manually curated metadata for genome and metagenomic sequences; accessible by HTML	www.genomesonline.org	[50]
Metadata standards	MIxS / MIMARKS	The minimal information about a MARKer gene (MIMARKS) or any gene (MIxS) standard	http://www.gensc.org/gc_wiki/index.php/MIMARKS	[41]

and sediment communities [14]. An association between habitat and phylogeny has also been detected in an analysis combining phylogenetically informative marker genes identified in metagenomic studies, comparison of the isolation environment for cultured organisms and 16S rRNA gene surveys [16]. These results suggest that microbial habitat preferences are fairly stable over evolutionary time. For example, we would not expect to see such patterns if horizontal gene transfer (HGT) was so rampant that all microorganisms were equally capable of adapting to a given environment (by rapid acquisition of the necessary genes from indigenous microbes). However, the observed correlation between phylogeny and habitat in microbial communities does not imply that habitat range for any individual organism can be perfectly predicted from phylogeny alone (nor does it contradict the observation of long tails of rare microbes in many samples [17]). Instead, this observation demonstrates that phylogenetic information can provide a useful first approximation for habitat range; accurate probabilistic models for determining how accurately phylogeny (or gene repertoire) can predict microbial habitat range remain a topic for future research.

The adaptation of microbial taxa to different habitats or lifestyles is reflected in their genome sequences. Some of the best-established examples of habitat adaptation identified in genomic studies include reduced genome size in intracellular endosymbionts [18] (Box 1), increases in genome size and the prevalence of two-component regulators

Box 1. Genome reduction

Genome reduction is one of the best-studied examples of genome evolution as a habitat adaptation in microbial organisms. Genomic minimalism is typically associated with organisms living in a host-associated environment, either as endosymbionts or obligate parasites (e.g. [18,82]), where increasing reliance on the host leads to loss of numerous pathways. The reduced genomes of the insect symbionts *Buchnera* (450 kb) and *Carsonella* (160 kb) have lost many biosynthetic pathways, but retain genes for amino acid biosynthesis, which forms the basis for their relationship with the host [18]. The extent of genomic reduction tends to increase as the length of the obligate relationship with the host increases, with the greatest reduction seen in the mitochondria and plastid organelles that have been stably incorporated in eukaryotic cells for more than one billion years and contain only a handful of genes [83]. Organelles also provide the most extreme example of eukaryotic genome reduction, in this case in the secondary plastids, which were acquired by acquisition of a eukaryotic alga. Two lineages with secondary plastids, cryptophytes and chlorarachniophytes, still retain a relict nucleus of the secondary red or green algal symbiont called a nucleomorph that has undergone extreme genome reduction and appears to be on a path to complete loss [84].

The genomic trajectory of obligate intracellular parasites has followed a similar reductive path, with extensive loss and/or reduction in biosynthetic pathways that corresponds to an increased reliance on the host [82]. Many eukaryotic lineages have undergone large-scale genomic streamlining when they become obligate parasites. The most extreme example is in microsporidia, a lineage of highly reduced fungi that are obligate intracellular parasites of diverse animals. The microsporidian *Enterocytozoon bieneusi*, an enteric pathogen in humans, has even lost the ability to synthesize its own ATP and instead has transporters to import ATP from its host [85]. Genomic reduction has also occurred in species of the highly abundant, free-living bacteria *Pelagibacter* and *Prochlorococcus*, where selection for efficient reproduction and/or reduced cell size are proposed to have selected for streamlining of genomic content [86,87]. In both *Pelagibacter ubique* and reduced strains of *Prochlorococcus*, loss of paralogous gene copies has been demonstrated to play a role in genome reduction [86,87]. In *Prochlorococcus* strains, loss of entire gene families has also played a role in genome reduction [87], whereas in *Pelagibacter* few ancestral pathways have been lost [86]. Genome reduction in *Pelagibacter* has instead been achieved by a reduction in the length of intergenic regions (these regions have a median length of only three nucleotides), the elimination of phage genes and pseudogenes and loss of recently duplicated paralogs [86].

in cosmopolitan organisms [19], increased acidic amino acids as a response to salinity ([20] and references contained therein) and increased rRNA copy number in fast-growing microorganisms ([21–24] and references contained therein). Additionally, numerous comparative genomic analyses have identified genomic changes associated with differences in habitat or lifestyle within specific taxa (see [25] and [26] for recent examples and [27] for a review).

Finally, metagenomic surveys have also shed light on many important aspects of habitat adaptation. These include changes in the aggregate functional profiles of microbial communities along gradients of depth [28], across diverse habitat categories [29] or between oligotrophic and copiotrophic communities [21].

Increasingly, research into microbial habitat adaptation is successfully leveraging publically available genome, marker gene and metagenome sequence data to contextualize new findings. Specifically, several recent studies of microbial co-occurrence [30], habitat adaptation [31,32], survival strategy [21] and genome evolution [33,34] have combined phylogenetic and genomic or metagenomic information to better understand microbial habitat adaptation. Such studies have converged on related strategies and faced common challenges. Based on these trends, we discuss a generalized workflow for comparative analysis (Figure 1), including the challenges involved in matching sequenced genomes to habitat assignments, determining which environmental parameters are most likely to be relevant for an analysis, separating the effects of habitat adaptation from those of shared evolutionary history and detecting HGT (Figure 1).

Challenges in defining microbial habitat range

To understand how microbial genomes change in response to environmental adaptation, comparative genomics approaches to habitat adaptation require an operational definition for environment, and a way to relate individual microbial genomes and the environments to which they are adapted (Figure 1). In the future, this problem may be resolved by single-cell genomics: careful selection of a range of environments, followed by the sequencing of large numbers of phylogenetically representative complete genomes directly from those environments would provide an unambiguous association between individual sequenced microorganisms and their habitat. In practice, however, this is not yet attainable on a large scale, although substantial progress is being made in techniques for obtaining genome sequences from single cells [35,36]. Direct assembly of genomes from deep metagenomic sequencing provides a similarly direct connection between genome and environment [37]. However, the assembly of complete genomes from metagenomic data is limited both because it can be difficult to obtain sufficient coverage for complete assembly in many complex communities and due to the potential for chimeric assemblies. Thus, many comparative genomics approaches currently rely on proxy information about the habitat range of an organism. Common proxy approaches for determining the habitat range of a sequenced microorganism include annotating environment based on the original isolation source (for cultured organisms), the reported collection site for environmental studies or database annotations based on one of these approaches. Annotating habitat from the source of the isolate is limited both by cultivation bias (the organisms that grow best in culture often represent a non-random subset of environmental diversity [38]) and because many organisms, especially those abundant in individual samples, are 'cosmopolitan' and can inhabit a variety of environments [39]. Careful surveys of the literature can be very useful in establishing a broader sense of the set of environments with which a sequenced organism must contend, but such surveys are laborious and are limited to the lineages actually discussed. An emerging alternative approach is to search community (marker gene and/or metagenomic) survey data for close relatives of sequenced genomes. Such an approach has the advantage that it can be conducted in a relatively unbiased manner and can associate sequenced organisms with the environmental samples in which their close relatives are found. As with annotations based on isolation source, however, care must be taken, as some organisms present in samples might simply be 'passing through' or might be contaminants

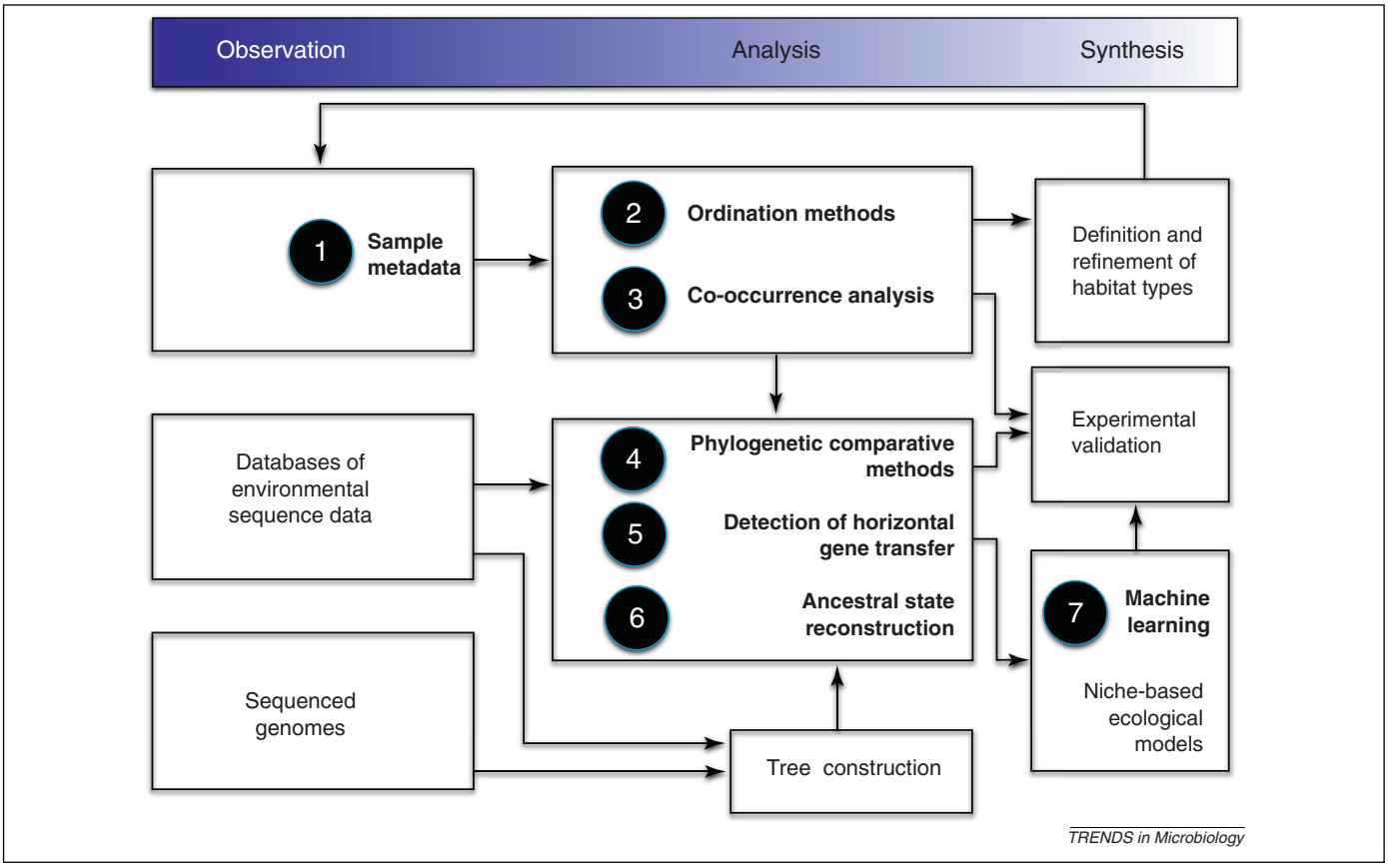

Figure 1. Recurrent themes in the analysis of microbial habitat adaptation. Numbered topics in bold correspond to sections in the text (see main text for additional detail). To compare microbes across habitats, it is first necessary to define the environmental factors that structure microbial communities. Insights into this question can be gained by combining sequence data from community surveys (e.g. 16S rRNA or other marker gene sequences) with rich metadata (Topic 1), using ordination techniques (Topic 2). These results can then help to define (and refine) important habitat categories. Interactions between organisms (such as competition or cooperation) can be characterized using co-occurrence analysis (Topic 3). When well-defined and annotated habitat categories (or data on environmental parameters) are available, surveys of microbial communities can be combined with genome sequence data and phylogenetic trees to allow more detailed study of habitat adaptation. Such studies include phylogenetic comparative measures (Topic 4), detection of horizontal gene transfer (Topic 5) and ancestral state reconstruction (Topic 6). Application of these techniques in combination allows for inference of traits involved in habitat adaptation: these traits/habitat associations can then be put into a predictive framework using machine learning techniques (Topic 7) or ecological modeling. Traits predicted to be important for habitat adaptation can be selected for detailed experimental study (e.g. by mutagenesis followed by competition in microcosms).

(Box 2). As databases of 16S rRNA and metagenomic community surveys accumulate, automated methods for surveying the habitat range of microbial taxa (see, e.g. [32,40]) using community surveys should become increasingly effective. However, this improvement is entirely dependent on the consistency and quality of the contextual information associated with marker gene and metagenomic community surveys.

Metadata annotation

The rapid accumulation of studies encompassing thousands of samples and billions of sequences has the potential to allow myriad new insights through comparative analysis. However, to maximize this potential, accurate contextual information about the samples (often called 'sequence metadata') is an increasingly important consideration (Figure 1, Topic 1). In addition to the existence of metadata, the form of that metadata is crucially important: if data are not consistently annotated in a standardized machine-readable format, large-scale comparative analyses become difficult or impossible. The utility of datasets for comparative analysis is thus frequently limited by the quality of metadata reported for the sampled environment. Such limitations can be introduced during data collection, data encoding or data reporting. During data collection, datasets are often limited by reporting only those physical, chemical or geographic parameters relevant to the particular hypothesis at hand (even if other parameters were collected). A lack of widely adopted standards for encoding the metadata that describes samples also presents significant challenges for comparative analyses. Differences in annotation can range from relatively simple (the use of different names or abbreviations to represent the same body site) to very challenging (differing definitions of environment types). Another limitation occurs during publication: although journals require that sequence data be made publically available, similar requirement have not been enforced for sample metadata.

To address these issues, many new sequencing efforts are now adopting the minimal information about any (x) sequence (MIxS) standards (http://www.gensc.org/gc_wiki/index.php/MIxS) which was proposed by the Genomic Standards Consortium [41]. The MIxS standard encapsulates three metadata compliant data types, which are the minimal information about a (meta)genome sequence [42] and the minimal information about a marker gene sequence [41]. These standards require researchers to supply their metadata using controlled vocabulary terminology and ontological values, which will greatly

Box 2. Source/sink dynamics

Attempts to map the habitat range of an organism using (metagenomic or marker gene) community surveys is that the presence of a microbe in an assemblage is not proof that the organism is adapted for life there. If a productive (source) and an unproductive (sink) environment are linked by high rates of migration, even relatively abundant organisms in the unproductive environment can be maintained primarily by migration from the source, rather than reproduction in the sink [88]. Such source/sink dynamics have been extensively documented in the ecology of micro- and macroscopic organisms [88,89] and are likely to play important roles in many microbial communities. For example, microbial assemblages from the human gut may contain transient populations of microorganisms associated with ingested food or the mouth community, in addition to the indigenous community. The complexities presented by source/sink dynamics are compounded by the prevalence of dormancy in microbial populations [90], which can increase the ability of microbes to emigrate to, and persist in, marginal habitats. Currently available techniques for minimizing the effect of source/sink dynamics when annotating habitat range from community surveys include requiring the presence of an OTU across multiple samples, considering the relative presence of an organism in a habitat as a proportion of its total abundance across all environments and experimental comparison of rRNA and rDNA ratios to test for metabolism in the sample can indicate the presence of alive and actively transcribing organisms as opposed to just their DNA. One additional recent approach to this problem involves new algorithms for tracking recent migration from a source environment [91]. This approach can also detect laboratory contamination, which can lead to inappropriate conclusions about cosmopolitanism (see [39] and references contained therein). However, accurate techniques for inferring microbial habitat adaptation (fitness in a particular habitat, rather than merely presence) from community surveys remain a topic where further development is needed.

benefit cross-study comparisons. Owing to the adoption of such standards, some databases are also starting to require MIxS compliance during metadata submission. These include the Metagenomics RAST Server (MG-RAST; http://metagenomics.anl.gov/) [43], the Human Microbiome Project (HMP; http://www.hmpdacc.org/), the Earth Microbiome Project (EMP; http://www.earthmicrobiome.org/) [44] and the QIIME Database (http://www.microbio.me/qiime).

Ordination methods

When investigating habitat adaptation in microbes, it is crucial to first have a baseline understanding of how microbial communities vary across environmental samples (microbial β diversity) and the main factors that drive such variation (Figure 1, Topic 2). Ordination methods have been widely and fruitfully applied to address these questions. By assessing the microbial composition of each microbial community, ordination methods allow an assessment of the extent to which communities are partitioned into distinct clusters or arrayed along a continuous gradient based on environmental factors (see [45] for a survey of ordination methods).

Ordination analyses performed on microbial community composition data acquired via sequencing of the gene encoding the small subunit ribosomal RNA have been used to distinguish microbial communities, and to identify environmental factors that contribute to both large and small-scale differences between communities. For example, Lozupone *et al.* [46] found a clear split between saline and non-saline environments among non-host associated microbial communities. King *et al.* combined ordination techniques with biogeography to demonstrate the dominant role of pH, plant abundance and snow depth in shaping the microbial communities found in alpine soil and to build global distribution models for microorganisms in this habitat [47], and Fierer *et al.* used 16S rRNA composition data to show that microbial communities on individuals' hands were far more similar to the communities on their computer keyboards than they were to communities from other individuals' hands [48].

A community-wide perspective on the factors structuring microbial diversity can also be obtained by shotgun metagenomic data. DNA or RNA sequences from random locations on the genomes of many microbes in a community can be assigned to functional (or other) categories, and again ordination methods can be applied to the resulting data. The (dis)agreement between 16S rRNA data and metagenomic data can then be visualized and quantified via Procrustes analysis, which compares the similarity of pairs of ordinations (see [9] for an example of applying this technique to the 5′ and 3′ paired-end reads of the same rRNA molecules in environmental samples). Such comparisons are one method of determining, at the community level, the degree to which the pool of functional genes in a microbial assemblage is predictable from phylogeny (relative to other reference communities). An unusual degree of difference between phylogeny and gene content may be a biologically interesting signature of competition or functional convergence [32].

Finally, ordination methods can help to inform high-throughput studies of microbial habitat adaptation by determining which environmental parameters are most important in structuring community diversity. Objective methods for defining relevant metadata parameters and defining working habitat categories are crucial, because many studies rely heavily on the lifestyle or habitat categories defined in a small number of online databases (primarily NCBI [49] and GOLD [50]) to test comparative genomic hypotheses. Careful refinement of these categories and addition of more detailed subcategories (based in part on the results of ordination techniques) would yield rapid dividends in comparative analysis.

Application of machine learning techniques

Machine learning techniques hold promise for relating gene functions to habitat distributions (Figure 1, Topic 7). These techniques have been used in taxonomic classification of metagenomic data and many other problems in bioinformatics. Although their application to classification and clustering of microbial communities by habitat is relatively new [51], machine learning techniques have been applied extensively to habitat classification in microarray data [52]. This emerging approach has been successful for classifying microbial communities across several different habitat types. For example, Muegge *et al.* [53] used a nearest-neighbor approach to demonstrate that phylogenetic characterizations of microbial communities can be used to predict metagenomic profiles of those communities. Werner *et al.* used supervised classifiers to identify a small subset of operational taxonomic units (OTUs) that were highly predictive of the type of bioreactor in

brewery wastewater-treatment systems [54]. Supervised classifiers have also recently been applied to source tracking of fecal contamination in water supplies [55].

The primary purpose of supervised machine learning in the context of microbial habitat adaptation is to build predictive models of the differences between habitats. A supervised classifier takes as its input a set of biological samples (training data) characterized by, for example, observations of OTUs, or counts of gene categories, along with metadata identifying the source habitats of those communities. The output is a model designed to predict the source habitat for novel biological samples not included in the training data and an estimate of the expected future accuracy of the model. In many cases the classifier will also report a measure of the predictive capability of each of the dependent variables (e.g. gene categories). One of the main advantages of machine learning techniques is that they are designed to discover general trends present in the training data even when the number of dependent variables is much larger than the number of samples, while avoiding overfitting. This is a challenging task, however, in data as sparse and high-dimensional as microbial community surveys or metagenomic analyses, and exceptional caution must be taken to avoid overestimating the future accuracy of supervised classifiers with small sets of training data [56]. Novel techniques might also need to account for the compositional nature of metagenomics data; for example, changes in a dominant community member could introduce spurious correlations between minor members [57]. Nonetheless, one exciting direction is that once sufficient genomes linked to environmental samples have been collected, machine learning techniques will be ideal for understanding which genes, regulatory structures or other properties of the genome are specifically associated with presence in an environment, especially when combined with the phylogenetic methods discussed in the next section.

Phylogenetic comparative methods

Once habitats have been assigned to organisms, relating genome properties to habitats is still challenging. Because all organisms share a common ancestry, each genome sequence cannot be counted as an independent observation when conducting statistical analyses, including machine learning techniques. Instead, the evolutionary history that relates organisms must be taken into account [58] (Figure 1, Topic 4). The importance of this well-established, but often ignored, principle is illustrated in Figure 2. Phylogenetic comparative methods are of particular relevance to microbial ecologists because the organisms selected for genome sequencing are not distributed across the tree of life evenly (although efforts are underway to ameliorate this problem [13]). This sequencing bias exacerbates the problems of interpretation introduced when traits are correlated with phylogeny.

Recent investigations of microbial adaptation to the human gut [32], global co-occurrence patterns [30] and genomic changes associated with growth rate [21] have investigated phylogenetic patterns by plotting relevant traits against phylogenetic distance, and found useful information in both trends that can largely be explained by phylogeny (e.g. similarity in GC content [21,30]) and those can only be partially explained by phylogeny (e.g. gene content during adaptation to life in the gut [32]; and gene content and genome size in co-occurring organisms [30]). Other studies have employed rarefaction, in which data are evened out across categories by discarding members of overrepresented taxa. Rarefaction can provide a useful check on the effects of oversampled taxa but suffers from the obvious drawback that it frequently discards a large portion of the data, and is limited by the least sampled taxon. Nonetheless, the utility of relatively unsophisticated methods such as rarefaction and regression against phylogenetic distance suggests that inclusion of more formal analyses of phylogenetic signal (e.g. phylogenetic independent contrasts [59] and phylogenetic generalized least squares), along with reconstructions of ancestral states (Box 3) could play an important role in future studies of microbial habitat adaptation. The development [59,60] and testing [61,62] of phylogenetic comparative methods for quantitative traits, as well as software packages [63,64] to make such methods easily accessible, are active areas of research, but many tools exist for estimating these traits without phylogenetic bias (Table 1) and should be applied in microbial studies.

Relating co-occurrence patterns to bacterial genomes

One way to understand potential interactions between organisms that might impact environmental distribution is through the application of co-occurrence analysis (Figure 1, Topic 3). For instance, species that support each other's growth, such as in syntrophic relationships where one organism produces metabolites that are consumed by the other, would be expected to positively co-occur across samples. By contrast, species that competitively exclude each other (e.g. because of similar metabolic requirements) might negatively co-occur. Co-occurrence patterns, however, are confounded because both positive and negative associations can also be driven by environmental preferences [30,65]. Additionally, differences in the depth of sampling between environmental isolates could obscure co-occurrence patterns, especially for rare taxa.

Combining co-occurrence studies with comparative genomics can clarify the biological properties that drive associations among microbes [30]. As an example, Chaffron *et al.* performed a global analysis of co-occurrence patterns using 16S rRNA surveys representing 3000 distinct sampling events for which sequence data were deposited in GenBank [30]. They then assessed the genomic properties of the subset of OTUs for which close relatives had genome sequences. Although some of the positive associations in the 16S rRNA OTU network reflected known or suspected syntrophic associations, such as a consortium involved in the anaerobic oxidation of methane, the general trends suggested that the major factor driving positive associations was shared environmental preference. Positively co-occurring OTUs were more phylogenetically related than random OTU pairs, extending to lineages that diverged up to 10% at the 16S rRNA level (these would typically be placed in different taxonomic families). Interestingly, positively co-occurring OTUs had more similar genome size, GC content (the proportion of nucleotides that are guanine

(a)

Raw habitat data

Node name	Habitat	Metabolic genes	Other genes
A	Habitat 1	2,042	2,957
B	Habitat 1	2,018	2,981
C	Habitat 1	2,021	2,978
D	Habitat 1	2,043	2,956
E	Habitat 1	2,046	2,953
F	Habitat 1	2,016	2,983
G	Habitat 1	2,050	2,949
H	Habitat 1	2,062	2,937
I	Habitat 2	2,102	2,897
J	Habitat 2	2,111	2,888
K	Habitat 2	2,077	2,922
L	Habitat 2	2,080	2,919
M	Habitat 2	2,067	2,932
N	Habitat 2	2,083	2,916
O	Habitat 1	2,032	2,967
P	Habitat 1	2,032	2,967

Summary

Habitat	Metabolic genes	Other genes	Average % metabolic
Habitat 1	20,362	29,628	49.7%
Habitat 2	17,474	12,520	58.3%

Result appears significant by G-test for independence:

G = 7.886759, p = 0.0024899

(b)

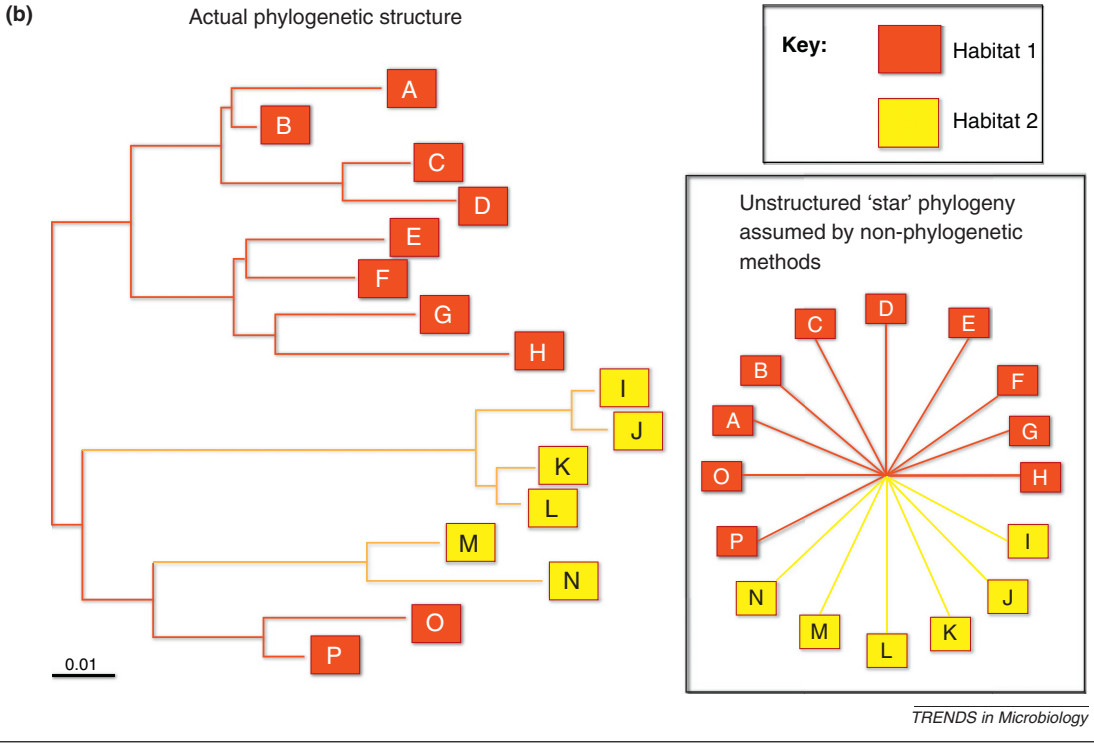

Figure 2. The importance of phylogenetic correction in comparing traits across habitats. Consider the problem faced by an investigator seeking to test whether adaptation to a copiotrophic environment (Habitat 2) is correlated with acquisition of additional metabolic genes relative to an oligotrophic environment (Habitat 1). To illustrate how phylogenetic structure can complicate such analyses, panels **(a)** and **(b)** summarize hypothetical (simulated) data representing a case in which habitat adaptation and metabolic gene evolution are purely independent. Given gene presence/absence data derived from whole-genome sequences **(a)**, it may be tempting to use traditional statistical methods without phylogenetic correction to test whether the habitat of the organism influences the number of metabolic genes in its genome. Naïve assessment of the effect of habitat on gene content using a non-phylogenetic test may lead an investigator to conclude that the increase in representation of metabolic genes between organisms found in Habitat 2 over Habitat 1 (58.53% vs 49.7%) is statistically significant. In the example, illustrated in **(a)**, the G-test for independence yields a highly significant *P*-value ($P = 0.00249$), despite no actual connection between habitat and gene content. Examination of the phylogeny relating the genomes **(b)** reveals a great deal of phylogenetic structure that is ignored by any statistical test that does not incorporate evolutionary relatedness (including, but not limited to, the G-test). Instead, such non-phylogenetic statistical tests implicitly assume the unstructured 'star' phylogeny (panel b, inset). Ignoring the hidden patterns of correlation caused by shared evolutionary history in this manner frequently produces false positive results such as that in **(a)**. For example, non-phylogenetic tests would ignore the correlations caused by the close phylogenetic relationship between lineages I and J as well as K and L (thus overcounting genes from the lineages). In this simple hypothetical example, we can readily observe that phylogenetically unaware statistical methods can generate false positive results. This phenomenon is widely recognized in the literature on phylogenetic comparative methods (see [59–64,81] for more detailed discussion). To illustrate that the false positive result obtained in this hypothetical example is not

Box 3. Ancestral state reconstruction

Reconstruction of ancestral states is a powerful tool to understand molecular and genomic evolution, which is increasingly being applied to the study of microbial habitat adaptation. Ancestral traits for a group of species can be inferred based on a phylogenetic tree, an alignment of the observed states and a model of evolution of the character under study. By analyzing a character in a group of extant species, the most probable state the character had in the common ancestor of these species can be determined, thus identifying changes that have occurred since divergence. The ancestral sequence can be estimated by one of several methods, such as parsimony [92], maximum likelihood ([93] and references therein) or Bayesian inference ([94] and references therein). Selected tools for performing ancestral state reconstruction are listed in Table 1. For relatively recent evolutionary events, it is sometimes possible to infer probable gene sequences at ancestral nodes. The estimated sequence can then be synthesized, cloned into a vector that is transfected into a cell and the expressed protein can subsequently be purified to study its properties. Based on this process, new insights into the evolution of dim-light vision [95] and steroid receptors [96] have been obtained. In addition to inferring ancestral gene sequences, ancestral state reconstruction has also been applied to infer other traits, such as mitochondrial metabolism [97] and the content of genomes [98]. In the future, it seems probable that integrated studies of genomic evolution including both ancestral state reconstruction of genome contents, sequence-based analyses of selective pressure (e.g. via ratios of synonymous to non-synonymous nucleotide substitutions [99]), tests of the order of trait divergence [100] and detection of horizontal gene transfer could yield new insights into the evolution of microbial habitat adaptation.

or cytosine) and relative coverage of KEGG functional pathways than random OTU pairs. Phylogeny could largely explain the high similarity in GC content but not similarities in genome size and KEGG functional pathway coverage. Thus, inhabiting the same environment could drive convergence of genome size and metabolic potential in divergent microbes [30].

Horizontal gene transfer (HGT)

Ongoing studies have continued to document the important roles played by HGT in microbial habitat adaptation (Figure 1, Topic 5). HGT can be detected by several methods: phylogenetic methods, which typically compare gene trees with a 'species tree'; compositional methods, which analyze deviations in nucleotide, codon or amino acid composition; or mobile-element methods, which search for specific genes or sequences associated with DNA mobility (see [66] for a review). Although there is ongoing controversy [67,68] about the total extent of HGT, and the implications of HGT for microbial (especially bacterial and archaeal) phylogeny [69], it is increasingly clear that both (i) HGT has played a major role in bacterial evolution and (ii) trees of the universal or nearly universal genes give the same overall phylogenetic pattern on average [68], implying that the extent of HGT is not so great that measures of vertical inheritance, such as 16S rRNA phylogenies, are meaningless. Several recent studies of HGT have therefore focused on separating the relative contribution of HGT (by conjugation, phage transduction, transformation, etc. [66]) and vertical descent (including gene loss, duplication, evolution of new gene families and sequence divergence) to the evolution of gene content.

Schliep *et al.* [34] used information embedded in the set (or 'forest') of gene trees from 100 bacteria and archaea to identify sections of gene trees that were not consistent with vertical descent, but did correspond to lifestyle (e.g. 'anaerobe') or habitat (e.g. 'soil') features as derived from NCBI annotations. This analysis yielded sets of gene families that could be better explained by lifestyle or habitat annotations than by taxonomy (~19% of gene families analyzed for hyperthermophiles) as well as networks of gene exchange among taxa and clusters of genes that were gained or lost in association with lifestyle.

David and Alm [33] used AnGST, a model that tests for gene duplication, gene loss and HGT within a single framework, to reconstruct the evolutionary history of 3983 gene families. The results implied an 'archaeal expansion' 3.33–2.65 billion years ago in which the number of gene families expanded by ~26% during a period of rapid diversification. By examining the timing of the expansion, and finding that the gene categories increasing during this event were primarily associated with redox and electron transfer (O_2 binding, Fe binding and Fe-S binding were the most enriched categories), David and Alm were able to connect this expansion to the 'great oxygenation event': a dramatic biotically mediated event in Earth's history, in which the production of oxygen by photosynthesis began to exceed buffering capacity and thus raise O_2 levels in the atmosphere and ocean.

Algorithms that include a unified model of gene evolution hold great promise for the study of habitat adaptation in microbial genomes (Table 1). The separation of genome evolution into specific vertical or horizontal components, and relating patterns in each to changes in habitat or lifestyle are also promising avenues for future research.

Concluding remarks

The increasing availability of 16S rRNA and metagenomic community surveys, in combination with new genome sequences, provides novel opportunities to conduct large-scale studies relating the survival strategies of microbial organisms to their genomic features (Box 4). Using the structure of the tree of life will be essential in establishing baseline predictions for trait conservation given phylogeny and thereby distinguishing novel adaptations to a particular habitat from traits preserved solely due to shared evolutionary history. Given this phylogenetic baseline, large collections of community surveys with backing metadata can be used to detect genomic variations associated

specific to the details of the tree, nor the small number of genes depicted, we repeated the procedure depicted in **(a)** and **(b)** across 1000 simulated 256 taxon trees. In each case, 5000 binary characters (representing gene presence/absence), plus one habitat character were simulated in a purely neutral manner. Because there was no genuine correlation between habitat and gene content, we would expect no more than a 5% false positive rate from a valid statistical test. However, in 38.4% (384/1000) of these hypothetical situations, a G-test of gene content versus habitat would falsely reveal a statistically significant result ($P < 0.05$). This simple example illustrates that the application of phylogenetic comparative measures in studies of microbial habitat adaptation should be considered essential (see Table 1 for available software and references [59–64,81] for studies that address this issue).

Box 4. Outstanding questions

- What is the relative role of sequence change in existing genes versus transfer of new genes in microbial habitat adaptation?
- Given that microbes may be detected as present in an assemblage, but not genuinely adapted for life there (due to source/sink dynamics or laboratory contamination), how can we best tell which organisms are adapted?
- How can we determine the order of adaptive events that permits colonization of a new environment?
- Do adaptations to environments with similar features (e.g. the guts of various mammals) share similar features?
- Can adaptation to other members of the community (e.g. syntrophy) be distinguished from shared adaptation to common abiotic factors present in the environment?
- Given source/sink dynamics, what is the best (experimental or bioinformatic) measure of habitat adaptation?
- If HGT has played an important role in microbial adaptation to a variety of environments, then what is the timing of specific genomic changes that allow for adaptation to a habitat relative to dispersal into a new habitat?
- To what extent has HGT affected microbial eukaryotes?
- What are the relative contributions of different gene transfer mechanisms to adaptive evolution across habitats?
- Do adaptation to habitat (e.g. the human gut) and adaptation to other organisms co-occurring in a specific assemblage (e.g. syntrophy) show interchangeable genomic signals, or are these patterns distinct?

with life in a range of environmental conditions. Statistical tools are now available for investigating adaptation along ecological gradients, detecting HGT, reconstructing the evolutionary history of genes involved in environmental adaptation and inferring correlations in species abundance. A major challenge for future studies will be designing accessible, high-throughput pipelines that combine these tools to gain biological insight and generate testable hypotheses from the large-scale sequence collection efforts currently underway.

Acknowledgments

The authors would like to thank Mike Robeson for useful comments on the draft. The work from our laboratory described in this review was supported in part by the National Institutes of Health, the Crohn's and Colitis Foundation of America, the Bill and Melinda Gates Foundation, and the Howard Hughes Medical Institute.

References

1 Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science* 276, 734–740
2 Ciccarelli, F.D. *et al.* (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283–1287
3 Johnson, Z.I. *et al.* (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* 311, 1737–1740
4 Lugtenberg, B. and Kamilova, F. (2009) Plant-growth-promoting rhizobacteria. *Annu. Rev. Microbiol.* 63, 541–556
5 Womack, A.M. *et al.* (2010) Biodiversity and biogeography of the atmosphere. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 365, 3645–3653
6 Cheesman, S.E. *et al.* (2011) Epithelial cell proliferation in the developing zebrafish intestine is regulated by the Wnt pathway and microbial signaling via Myd88. *Proc. Natl. Acad. Sci. U.S.A.* 108 (Suppl. 1), 4570–4577
7 Samuel, B.S. *et al.* (2008) Effects of the gut microbiota on host adiposity are modulated by the short-chain fatty-acid binding G protein-coupled receptor, Gpr41. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16767–16772
8 Sharon, G. *et al.* (2010) Commensal bacteria play a role in mating preference of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 20051–20056
9 Caporaso, J.G. *et al.* (2011) 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
10 Qin, J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65
11 Peterson, J. *et al.* (2009) The NIH Human Microbiome Project. *Genome Res.* 19, 2317–2323
12 Stewart, F.J. *et al.* (2011) Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome Biol.* 12, R26
13 Wu, D. *et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462, 1056–1060
14 Lozupone, C.A. and Knight, R. (2007) Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11436–11440
15 Ley, R.E. *et al.* (2008) Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.* 6, 776–788
16 von Mering, C. *et al.* (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 315, 1126–1130
17 Pedros-Alio, C. (2007) Ecology. Dipping into the rare biosphere. *Science* 315, 192–193
18 Moran, N.A. *et al.* (2008) Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* 42, 165–190
19 Konstantinidis, K.T. and Tiedje, J.M. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 3160–3165
20 Rhodes, M.E. *et al.* (2010) Amino acid signatures of salinity on an environmental scale with a focus on the Dead Sea. *Environ. Microbiol.* 12, 2613–2623
21 Vieira-Silva, S. and Rocha, E.P. (2010) The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* 6, e1000808
22 Klappenbach, J.A. *et al.* (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* 66, 1328–1333
23 Klappenbach, J.A. *et al.* (2001) rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res.* 29, 181–184
24 Stevenson, B.S. and Schmidt, T.M. (2004) Life history implications of rRNA gene copy number in *Escherichia coli*. *Appl. Environ. Microbiol.* 70, 6670–6677
25 Cho, Y.J. *et al.* (2010) Genomic evolution of *Vibrio cholerae*. *Curr. Opin. Microbiol.* 13, 646–651
26 Deng, X. *et al.* (2010) Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. *BMC Genomics* 11, 500
27 Binnewies, T.T. *et al.* (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct. Integr. Genomics* 6, 165–185
28 DeLong, E.F. *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311, 496–503
29 Dinsdale, E.A. *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* 452, 629–632
30 Chaffron, S. *et al.* (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* 20, 947–959
31 Merhej, V. *et al.* (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol. Direct.* 4, 13
32 Zaneveld, J.R. *et al.* (2010) Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res.* 38, 3869–3879
33 David, L.A. and Alm, E.J. (2011) Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 469, 93–96
34 Schliep, K. *et al.* (2011) Harvesting evolutionary signals in a forest of prokaryotic gene trees. *Mol. Biol. Evol.* 28, 1393–1405
35 Ishoey, T. *et al.* (2008) Genomic sequencing of single microbial cells from environmental samples. *Curr. Opin. Microbiol.* 11, 198–204
36 Coleman, M.L. and Chisholm, S.W. (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18634–18639
37 Narasingarao, P. *et al.* (2011) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J.* DOI: 10.1038/ismej.2011.78

38 Lozupone, C. and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235
39 Nemergut, D.R. *et al.* (2011) Global patterns in the biogeography of bacterial taxa. *Environ. Microbiol.* 13, 135–144
40 Lozupone, C.A. *et al.* (2008) The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc. Natl. Acad. Sci. U.S.A.* 105, 15076–15081
41 Yilmaz, P. *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* 29, 415–420
42 Kottmann, R. *et al.* (2008) A standard MIGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* 12, 115–121
43 Meyer, F. *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* 9, 386
44 Gilbert, J.A. *et al.* (2010) 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
45 Ramette, A. (2007) Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* 62, 142–160
46 Knight, R. *et al.* (2007) PyCogent: a toolkit for making sense from sequence. *Genome Biol.* 8, R171
47 Rousk, J. *et al.* (2010) Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.* 4, 1340–1351
48 Fierer, N. *et al.* (2010) Forensic identification using skin bacterial communities. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6477–6481
49 Benson, D.A. *et al.* (2011) GenBank. *Nucleic Acids Res.* 39, D32–D37
50 Liolios, K. *et al.* (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 38, D346–D354
51 Knights, D. *et al.* (2011) Supervised classification of human microbiota. *FEMS Microbiol. Rev.* 35, 343–359
52 Lee, J.W. *et al.* (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. Data Anal.* 48, 869–885
53 Muegge, B.D. *et al.* (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 332, 970–974
54 Werner, J.J. *et al.* (2011) Bacterial community structures are unique and resilient in full-scale bioenergy systems. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4158–4163
55 Smith, A. *et al.* (2010) Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. *Water Res.* 44, 4067–4076
56 Efron, B. and Tibshirani, R. (1997) Improvements on cross-validation: the .632+ bootstrap method. *J. Am. Stat. Assoc.* 92, 548–560
57 Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*, Chapman & Hall
58 Harvey, P.H. and Pagel, M.D. (1991) *The Comparative Method in Evolutionary Biology*, Oxford University Press
59 Felsenstein, J. (1985) Phylogenies and the comparative method. *Am. Nat.* 125, 1–15
60 Blomberg, S.P. *et al.* (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57, 717–745
61 Laurin, M. (2010) Assessment of the relative merits of a few methods to detect evolutionary trends. *Syst. Biol.* 59, 689–704
62 Freckleton, R.P. *et al.* (2002) Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* 160, 712–726
63 Jombart, T. *et al.* (2010) adephylo: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics* 26, 1907–1909
64 Kembel, S.W. *et al.* (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26, 1463–1464
65 Horner-Devine, M.C. *et al.* (2007) A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology* 88, 1345–1353
66 Zaneveld, J.R. *et al.* (2008) Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. *Microbiology* 154, 1–15
67 Galtier, N. and Daubin, V. (2008) Dealing with incongruence in phylogenomic analyses. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 363, 4023–4029
68 Koonin, E.V. *et al.* (2011) Comparison of phylogenetic trees and search for a central trend in the "forest of life". *J. Comput. Biol.* 18, 917–924
69 Andam, C.P. *et al.* (2010) Biased gene transfer mimics patterns created through shared ancestry. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10679–10684
70 Caporaso, J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336
71 Schloss, P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541
72 Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591
73 Carmel, L. *et al.* (2010) EREM: parameter estimation and ancestral reconstruction by expectation-maximization algorithm for a probabilistic model of genomic binary characters evolution. *Adv. Bioinform.* 2010, 167408
74 Paradis, E. *et al.* (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290
75 Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574
76 Drummond, A.J. and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7, 214
77 Dray, S. and Dufour, A.B. (2007) The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Software* 22, 1–20
78 Than, C. *et al.* (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinform.* 9, 322
79 Podell, S. *et al.* (2008) A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. *BMC Bioinform.* 9, 419
80 Schliep, K.P. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics* 27, 592–593
81 Jombart, T. *et al.* (2010) Putting phylogeny into the analysis of biological traits: a methodological approach. *J. Theor. Biol.* 264, 693–701
82 Pallen, M.J. and Wren, B.W. (2007) Bacterial pathogenomics. *Nature* 449, 835–842
83 Gray, M.W. (1999) Evolution of organellar genomes. *Curr. Opin. Genet. Dev.* 9, 678–687
84 Archibald, J.M. and Lane, C.E. (2009) Going, going, not quite gone: nucleomorphs as a case study in nuclear genome reduction. *J. Hered.* 100, 582–590
85 Keeling, P.J. *et al.* (2010) The reduced genome of the parasitic microsporidian *Enterocytozoon bieneusi* lacks genes for core carbon metabolism. *Genome Biol. Evol.* 2, 304–309
86 Giovannoni, S.J. *et al.* (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309, 1242–1245
87 Luo, H. *et al.* (2011) Genome reduction by deletion of paralogs in the marine cyanobacterium *Prochlorococcus*. *Mol. Biol. Evol.* DOI: 10.1093/molbev/msr081
88 Kawecki, T.J. (2000) Adaptation to marginal habitats: contrasting influence of the dispersal rate on the fate of alleles with small and large effects. *Proc. Biol. Sci.* 267, 1315–1320
89 Sokurenko, E.V. *et al.* (2006) Source-sink dynamics of virulence evolution. *Nat. Rev. Microbiol.* 4, 548–555
90 Jones, S.E. and Lennon, J.T. (2010) Dormancy contributes to the maintenance of microbial diversity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5881–5886
91 Knights, D. *et al.* (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* DOI: 10.1038/nmeth.1650
92 Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113
93 Koshi, J.M. and Goldstein, R.A. (1996) Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* 42, 313–320
94 Pagel, M. *et al.* (2004) Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* 53, 673–684

95 Yokoyama, S. *et al.* (2008) Elucidation of phenotypic adaptations: molecular analyses of dim-light vision proteins in vertebrates. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13480–13485
96 Thornton, J.W. *et al.* (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* 301, 1714–1717
97 Gabaldon, T. and Huynen, M.A. (2003) Reconstruction of the proto-mitochondrial metabolism. *Science* 301, 609
98 Paten, B. *et al.* (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* 18, 1829–1843
99 Marri, P.R. *et al.* (2006) Gene gain and gene loss in *Streptococcus*: is it driven by habitat? *Mol. Biol. Evol.* 23, 2379–2391
100 Ackerly, D.D. *et al.* (2006) Niche evolution and adaptive radiation: testing the order of trait divergence. *Ecology* 87, S50–S61