Metagenomics and the Units of Biological Organization

W. FORD DOOLITTLE AND OLGA ZHAXYBAYEVA

Metagenomics is a complex of research methodologies aimed at characterizing microbial communities and cataloging microbial diversity and distribution without isolating or culturing organisms. This approach will unavoidably engender new ways of thinking about microbial ecology that supplant the concept of "species." This concept—thanks to comparative genomics—has in any case become increasingly unsustainable, either as a way of binning diversity or as a biological reality. Communities will become the units of evolutionary and ecological study. Although metagenomic methods will increasingly find uses in protistology and mycology, the emphasis so far has been, and our focus here will be, on prokaryotes (bacteria and archaea).

Keywords: metagenomics, species, communities, ontology, microbial ecology

Technology and science coevolve. Knowledge informs the development of tools, while tools and the way we use them constrain and structure knowledge—not just how we acquire it (our epistemology) but also its content (our ontology). The history of genomics, the parent science to both metagenomics and systems biology, is most instructive here. Genomics began as a technology and was initially criticized for offering nothing more than a quantum acceleration in data accumulation. And yet it was soon apparent that genomics would revolutionize the science of life. Among its many effects on biology have been the acceptance that comparative methods and phylogenetics are essential for understanding function, a deeper recognition of the unity of life (Friedmann 2004), a relegitimization of exploration as science, the substitution of engineering for genetics as the main arena of intellectual competition, a regrettable new emphasis on commercial end points for basic inquiry, and the injection of computer science perspectives into areas of biology that had largely escaped their influence. Of the latter, Roger Brent wrote in 2000: "As the genomic inventories approach closure, the mass of this data will spur attempts to devise computational frameworks that integrate biological knowledge about cellular components and attempt to predict system behavior. During the early twenty-first century, this more predictive biology will have positive consequences for health and agriculture and will speed the development of a design-based biological engineering of cells and organisms to perform new functions" (p. 169).

This potentially predictive science is what we now call systems biology (and synthetic biology). It brings both new epistemology and new ontology, and, as O'Malley and Dupré (2005) pointed out, some systems biologists still embrace only the epistemology. These researchers, whom O'Malley and Dupré term pragmatic, are "united simply by an agreement that systems biology involves the study of interacting molecular phenomena through the integration of multilevel data and models." But others, arguably now in ascendance, are what O'Malley and Dupré call systems-theoretic biologists. For these biologists, "systems are taken to constitute a fundamental ontological category, and differences between biological and human-made (engineered) systems are considered less important than their similarities" (p. 1271). This is not a trivial shift in paradigms: Most molecular geneticists have understood the distinction between evolved and engineered systems to be a fundamental one, ever since the publication in 1977 of François Jacob's essay "Evolution and Tinkering."

As genomics was to genetics, so will metagenomics be to genomics

Metagenomics, a much newer science than genomics or systems biology, is already at a similar crossroads between pragmatic and systems theoretic. Metagenomics too began pragmatically. Environmental libraries of the sort we would now call metagenomic first began to appear in the early 1990s (Schmidt et al. 1991, Stein et al. 1996). The first use of the term "metagenome" (of which we aware) was in 1998 by Handelsman and colleagues, who viewed the "collective genomes" of a specific microflora (in this case soil) primarily as a repository of potentially useful genes to be mined by cloning and expression in *Escherichia coli*

BioScience 60: 102–112. ISSN 0006-3568, electronic ISSN 1525-3244. doi:10.1525/bio.2010.60.2.5

(a technique sometimes called functional metagenomics). The earliest exercises in sequence-driven metagenomics were undertaken simply for the purpose of assembling complete genomes of uncultivable prokaryotes, using environmental DNA fragments cloned in bacteria artificial chromosomes (Schleper et al. 1998). But now the recognition and "omic" characterization of biological entities more inclusive than genomes, organisms, or even species—loosely, communities—seem quite solidly integrated into the practice and developing theory of metagenomics as a discipline, its very ethos. Indeed, the terms "community genomics," "ecogenomics," or "environmental genomics" are sometimes used as synonyms for "metagenomics," although these former also accommodate whole-genome approaches.

A white paper, *The New Science of Metagenomics,* produced in 2007 by the US National Research Council, gives the following extended definition of the contemporary discipline, both its epistemology and its ontology (Handelsman et al. 2007):

> Like genomics itself, metagenomics is both a set of *research techniques*, comprising many related approaches and methods, and a *research field*. In Greek, *meta* means "transcendent." In its approaches and methods, metagenomics circumvents the unculturability and genomic diversity of most microbes, the biggest roadblocks to advances in clinical and environmental microbiology. *Meta* in the first context recognizes the need to develop computational methods that maximize understanding of the genetic composition and activities of communities so complex that they can only be sampled, never completely characterized. In the second sense, that of a research field, *meta* means that this new science seeks to understand biology at the aggregate level, transcending the individual organism to focus on the genes in the community and how genes might influence each other's activities in serving collective functions. (p. 13)

The rest of this article will be developed along just these lines. First, we assert that metagenomics as a set of research techniques came along just in time to rescue microbiology from one of its more intractable practical concerns, culturability, and also one of its more onerous concepts, the species. Second, we claim that data bearing on the key concerns of environmental microbiologists, which we consider to be the diversity, dispersal, and niche differentiation of microbial cells, can be collected and interpreted perfectly well without reference to the concept of species. And third, we note that as a research field, metagenomics is in the process of replacing the species as its fundamental unit with the community (or ecosystem). As with systems biology, more practitioners will accept this shift as epistemology rather than as ontology.

Culturability and "the species"

Microbiology's history is in many important ways divorced from those of zoology and botany, although it has imported from them principles of classification and underlying evolutionary concepts. It is worth remembering that even in Darwin's time, the status of bacteria as living organisms with evolutionary continuity (rather than as products of recurring events of spontaneous generation) was in doubt. And whether they had particulate heritable genes, like Mendel's peas or Dobzhansky's *Drosophila*, could be legitimately debated until the middle of the last century. Pure cultures were thus essential in proving the microbial causation of disease (Koch's postulates), and became the gold standard for environmental microbiology. Type specimens in culture collections became and are still a mainstay of formal classification (Stackebrandt et al. 2002), enshrining the notion that, as with animals and plants, bacteria have species and these species have "typical" members.

Applying these principles and the practice of laboratory microbiology to the environment (at large, but even in the more limited arena of the human microbiota) founders in a practical sense on the unculturability of most bacteria (and archaea), and in a conceptual sense on an astonishing genomic and phenomic variability within any grouping that we might want to call a species. As we discuss in the next section, the more unbiased culture-independent phylotyping methods that have largely replaced isolation and culturing do allow us to characterize environments by asking, "Who is there?" But they provide only partial answers to the question, "What are they doing?" and almost none to "What meaningful evolutionary and ecological units do they make up?"

Unculturability and the value of phylotyping

It is commonly asserted that anywhere from 90% to 99.9% of bacterial (and by extension archaeal) species cannot be grown under standard laboratory conditions. This disappointing fact is sometimes called "the great plate count anomaly," following Staley and Konopka (1985). It comes with the rueful corollary that those species that can be cultured from natural samples are likely to be neither the more numerous nor the more important members of the community, but are instead the "lab weeds." No doubt more nuanced and in particular more "omic" approaches will bring many intractable microbes to culture (Bollmann et al. 2007, Giovannoni and Stingl 2007), and the problem has most likely been hyped a bit in the interest of promoting culture-independent methods (including metagenomics). Still, many microbes will surely always be impossible to isolate in pure culture, either because they are obligate syntrophs or because their growth requirements are indefinable or unmeetable.

Thus for several decades, culture-independent molecular methods for studying microbial diversity, biogeography, and ecology have held sway. The field has been dominated by phylotyping, a powerful methodology introduced by

Norman Pace (1997) and based on biology's single and unarguably most broadly useful phylogenetic marker, small-subunit ribosomal RNA (SSU rRNA)—first made popular by Carl Woese (1987). This molecule is now represented in databases by more than three-quarters of a million individual gene sequences, and PCR (polymerase chain reaction) amplification from environmental DNA samples with universal, bacterial-specific, archaea-specific, or more selective primers allows phylogenetic identification (by precise positioning on the SSU rRNA tree) of the taxa present. There are obvious concerns about depth of sampling, within-genome SSU rRNA gene polymorphism, bias in priming and errors in sequencing, and, not least, the definition of taxa, especially "species" (see below). But by and large, SSU rRNA phylotyping has been the biggest boon to environmental microbiology since the Petri plate.

Unquestionably, this method has enormously enriched our appreciation of both microbial diversity at individual sites and differences between microbiota from different sites, while allowing us to rationalize what might be dominant biological and biogeochemical activities of site-specific microbial communities (e.g., Prosser and Nicol 2008). Much of the time, phylotype will be associated with and predictive of overall physiology. There can be little doubt, for instance, that the globally distributed SSU rRNA genes used to identify the marine cyanobacteria *Prochlococcus* and *Synechococcus* exist overwhelmingly only in oxygenic, photosynthetic unicells using chlorophyll a_2/b_2 and phycobilisomes, respectively (Ting et al. 2002).

But phylotyping is nevertheless much less specifically predictive of phenotype than we would have imagined when the technique was introduced—both for taxonomically important but fugacious traits such as toxin production or the evasion of host defenses, and for seemingly basic features such as sugar metabolism or nitrogen fixation. As discussed below, comparing sequenced genomes even of supposedly conspecific isolates has revealed an astonishing variability in gene content, including variability in genes with crucial adaptive roles. Similarly for comparative environmental metagenomics, "Who is there?" may be a less-relevant question than "What is happening?" For instance, in their recent comparison of gut microbiota of obese and nonobese twins, Turnbaugh and colleagues (2009) found that the relative abundances of genes with similar functions are more highly conserved than are the relative abundances of phylotypes.

Gene content variability

In the mid-1990s, microbial genomicists looked forward to the expeditious completion—one species at a time—of a few dozen bacterial and archaeal genomes, chosen in the interests of phylogenetic breadth or medical, economic, or environmental importance. If we planned to examine multiple isolates from the same species, it would only be to look for the mutations responsible for diagnostic interstrain phenotypic differences, and for the very occasional gene picked up by lateral gene transfer (LGT), which were expected to be plasmid-encoded (figure 1).

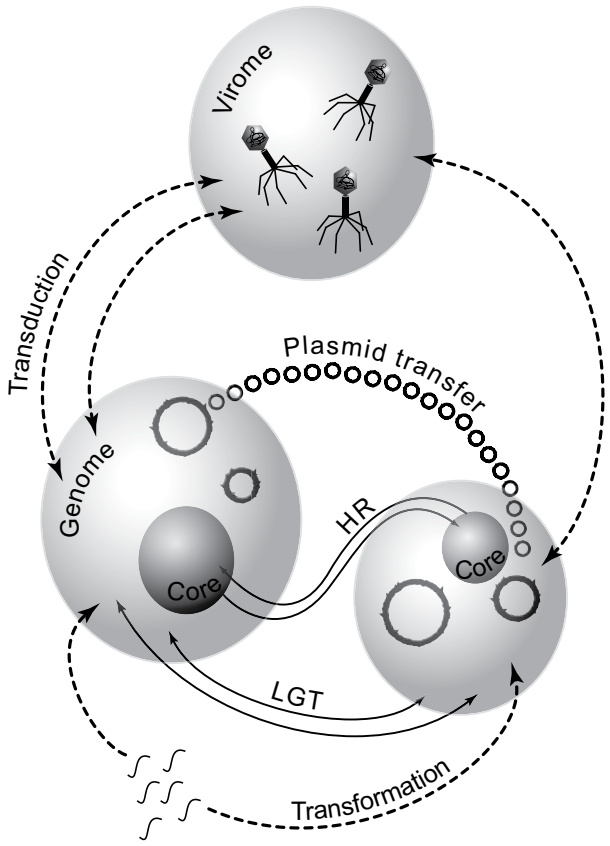

Figure 1. Flux of genes between two genomes and environment. It is not sufficient to consider genomes as isolated entities, as their content is continuously changed by acquisition of plasmids, naked DNA from the environment, and phage-carried genes. Dashed lines depict the mechanisms responsible for bringing foreign DNA into a genome. The inner "core" comprises those genes that are present in all isolates of a designated species—exchange between core genomes is often by homologous recombination. Each isolate will also bear many genes that are found only in some isolates—their acquisition will often be by lateral gene transfer (LGT). Evolutionary processes are shown in solid lines. Abbreviation: HR, homologous recombination.

A comparison study of the genomes of three strains of *E. coli* published six years ago made it clear that such limited sequencing efforts would not be sufficient (Welch et al. 2002). Welch and colleagues showed that most protein-coding genes in these strains were present in only one or two of them, with less than 40% being common to all three strains. Touchon and colleagues (2009) have recently updated this analysis to include 20 genomes of *E. coli* strains (some called *Shigella* but all at least 97% identical in sequences of shared genes). The total number of known genes from at least one strain is now between 11,000 and 18,000 (depending on how gene duplicates are counted), whereas the fraction of shared genes has fallen

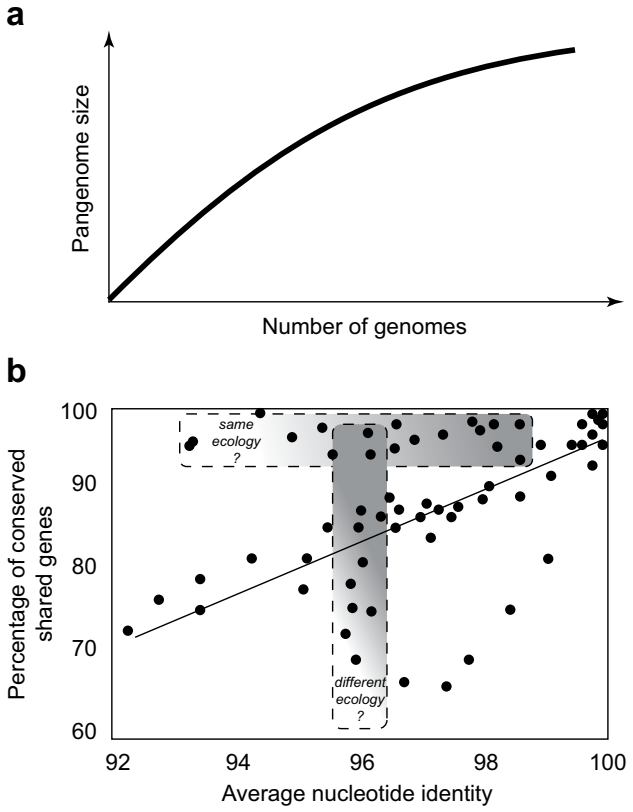

Figure 2. Examination of gene content of closely related genomes. (a) Illustration of pangenome growth as a function of the number of sequenced closely related genomes. Real data compiled for many groups of bacteria do not indicate that the curve is leveling off. (b) Illustration of relationship between gene content and genome sequence conservation. Each point represents a pairwise genome comparison. Although there is a positive correlation between gene content and genome sequence conservation, there are clear outliers, which are most likely to exhibit substantially different ecology. Source: Modified from figure 4 of Konstantinidis and Tiedje (2005). Original copyright by the National Academy of Sciences.

to about 20% (generously reckoned). The number of total genes in the so-called pangenome—the sum of all genes found in at least one strain—will only keep growing. Contrariwise, the fraction of genes in the core—genes found in all strains—will only shrink as more *E. coli* strain genomes are sequenced (figure 2a).

Lateral gene transfer by one mechanism or another must be invoked to explain such gene content variability, unless one is willing to entertain a scenario requiring more than 12,000 genes in the last common *E. coli* ancestor, with substantial differential gene loss in all its descendant lineages. Although phages, transposable elements, and functionally uncharacterized open-reading frames are concentrated among the noncore genes of the pangenome, these also include many genes that clearly serve strain-specific adaptive functions, including alternate substrate uptake and utilization, host specificity, and pathogenic potential (e.g., Schubert et al. 2009).

Such between-strain gene content variability is the rule, not the exception, in species for which multiple strains have been sequenced. Konstantinidis and Tiedje (2005) described a useful way of representing this variability, which is to plot the average nucleotide identity (ANI) between shared genes for any two strains versus the fraction of the average of their genome sizes that these shared genes represent. Figure 2b shows our visualization of such a representation. An ANI of 95% corresponds to a typical degree of similarity between strains for many recognized bacterial species, and this in turn generally corresponds to between 97% and 100% identity between SSU rRNA gene sequences (see below). Very similar or even identical SSU rRNA phylotypes can mask an enormous range of gene content and phenotypic variability, sometimes as much as 50%.

Two studies illustrate this point very well. The first is that of Martin Polz and his team (Thompson et al. 2005) at the Massachusetts Institute of Technology (MIT), working with coastal populations of *Vibrio splendidus,* a free-living bacterium that also can infect fish and shellfish. Using pulse-field gel electrophoresis, they documented astonishing between-strain genome size differences (up to 1 megabase) among isolates all differing by less than 1% in SSU rRNA sequence (Thompson et al. 2005). They concluded: "This group consists of at least a thousand distinct genotypes, each occurring at extremely low environmental concentrations (on average less than one cell per milliliter)" (p. 1311). In later work, using *hsp60* as a more sensitive marker (Hunt et al. 2008), these researchers showed that strains that were more similar to each other sorted more similarly when the water sample was divided into particle-associated and free-floating fractions. Therefore, gene content variation is most likely responsible for niche differentiation among the hundreds or thousands of subtypes that make up this "species," as defined by the SSU rRNA phylotype.

The second series of studies (Kettler et al. 2007, Martiny et al. 2008), from Sallie Chisholm's group (also at MIT), is an extensive one of the *Prochlorococcus* cluster, a collection of strains all within the traditional boundaries of a bacterial species (more than 97% rRNA identity), for which nearly a score of genome sequences and much environmental data are now available. These investigators used a phylogenetic tree (made from subsets of the 1273 "core" genes shared by 12 genomes examined) as a scaffold to tally gene gains and losses during diversification of this group into its various "ecotypes," for instance, by preferences for light intensity and nitrogen and phosphate utilization. In addition to the 140 genes lost from their common ancestor after its divergence from *Synechococcus* (the sister taxon), the high-light-adapted strains all share 257 genes (95 uniquely), several of which encode genes for DNA repair or protection from oxidative stress. Additionally, within high-light- and low-light-adapted groups, individual subclades and single strains differ in their

possession of genes that may correlate with the subdivisions of this species' niche, such as nitrate and phosphorous assimilation and light harvesting. Many of these hundreds of strain- or subclade-specific variable genes are found clustered in "genomic islands," but the mechanisms of acquisition are likely various: one source is surely phages (Lindell et al. 2004). Kettler and colleagues (2007) concluded: "We have barely begun to observe the extent of micro-diversity among *Prochlorococcus* in the ocean.... In particular, it will be enlightening to understand the complete genome diversity of the 10^5 cells in a milliliter of ocean water, and conversely, how widely separated in space two cells with identical genomes might be" (p. 2525).

From these and the dozens of other bacterial and archaeal species currently under study, it is not clear that there will be any clustering of cellular lineages so fine-grained (aside from identity) that it will not be matched by an equally fine-grained parceling out of the environment into micro-, nano-, or pico-niches. It is clear from these same observations, however, that if the limit to our ecological characterization of environmental microbes were not culturability (as it is not, e.g., for *V. splendidus*), it would be money, time, and person-power. Even with the cheapest new technologies, there are far too many genome sequences to complete, and far too many pure culture studies to carry out in developing anything approaching an adequate understanding of the ecology of even one currently recognized prokaryotic species, or even the simplest natural community. We need a new way to conduct microbial ecology independent of pure cultures but more information-rich than phylotype-based molecular methods, and in particular directed at "What is being done?" not "Who is doing it?" And we must be content with samples, not complete inventories. Necessity is indeed one of the mothers of metagenomics.

What is a "species," anyway?

We also clearly need a new way of thinking about species. This topic is a perennially vexatious one for microbiologists, usually argued in terms of the difficulty in reconciling a species definition with a species concept. The former would comprise a set of criteria by which different isolates could be judged to be members of the same species, and the latter a theory according to which such groupings might be more than arbitrary, more than mere matters of convenience.

The current bacterial (and by extension archaeal) species definition embodies a scientific community consensus: "A species is a category that circumscribes a (preferably) genomically coherent group of individual isolates/strains sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardized conditions" (Stackebrandt et al. 2002, p. 1044). This vague invocation is underwritten by the criterion that conspecific strains must show at least a value of 70% (which does not, by the way, translate to 70% shared gene content) in a standardized DNA-to-DNA hybridization test, and of 97% identity between SSU rRNA sequences. In all the species studies above, and in most under way, it would be difficult to claim either genomic coherence or a "high degree of similarity" in phenotype—or at least to claim anything like the degree of coherence and similarity exhibited by a typical animal species (or even perhaps a typical animal family, order or phylum, when considering only gene content). Konstantinidis and Tiedje (2005) suggested that using ANI as a measure and setting the cutoff at 99% would come closer to nonmicrobial usage. (An ANI of 95% corresponds to the current consensus definition.) However, they do not actually recommend this, for fear of a chaotic expansion in the number of bacterial "species."

Moreover, when the microbial sequence data are viewed in their entirety, it is unclear how often prokaryotes comprise "genomically coherent groups," as opposed to occupying (albeit with vastly differing abundances) a continuum of potential "genome space." For example, Konstantinidis and colleagues (2006) admitted that "an important issue that remains unresolved is whether bacteria exhibit a genetic continuum in nature," (p. 1935), and Gevers and colleagues (2005), in a group effort aimed at a consensus concept, confessed that "it might not be possible to delineate groups within a continuous spectrum of genotypic variation" (p. 733).

The situation may be somewhat more resolved when limited sampling sites are considered. In a sample from 4000-meters deep in the mid-Pacific, Konstantinidis and DeLong (2008) recently found reasonably discretely clustered metagenomic sequences, about as divergent as expected for typical "species" (5% to 6%). But they go on to note that "comparisons to surface-dwelling relatives of the Sargasso Sea revealed that distinct sequence-based clusters were not always detectable, presumably due to environmental variations, further underscoring the important relationship between environmental contexts and genetic mechanisms, which together shape and sustain microbial population structure" (p. 1052).

Assuming that clustering does in fact occur, at least under some environmental regimens such as the deep Pacific, do we have a genetic and ecological process theory about why it should—do we have a species concept? Nonmicrobes are generally considered to conform to Ernst Mayr's dominant biological species concept, which holds that "species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups" (Mayr 1942). Some nonmicrobes do not fit this definition, notably asexual lineages, and were considered by Mayr not to have species—along with bacteria and archaea, for which reproduction never entails mating (Mayr 1996). But many bacterial groups are now known to engage in frequent interlineage recombination. In fact, for some, such as *Neisseria* or *Helicobacter* (Hanage et al. 2006a) or the archaeon *Halorubrum* (Papke et al. 2007), diverging lineages accumulate many more sequence differences from recombination with other lineages than they do through mutations accumulating within the lineage, and it can be argued that such recombination-acquired differences are more likely

to be adaptively significant (Townsend et al. 2003). Thus, some microbiologists currently entertain the idea of a species concept applicable to such recombinogenic groups and analogous to the biological species concept (Hanage et al. 2006a, Fraser et al. 2007, Achtman and Wagner 2008).

At issue is not whether within-group recombination by transduction, conjugation, or transformation is analogous to interbreeding (in its genetic consequences, it is), but whether such bacterial groups are equivalently "isolated from other such groups." Barriers against interspecific "mating" (incorporation of foreign genes) are not likely to be selected for directly as they can be in animal species, because the acquisition of any genes—detrimental or beneficial—is extremely infrequent in the lives of bacteria. But selection for resistance to agents effecting transfer (phages in particular) may be common, as would the loss of competence in transformation. As well, sequence divergence between lineages will reduce the frequency of recombination between their DNAs, and we can construct models in which "sympatric speciation" (without physical separation) occurs spontaneously as a result. But such modeling indicates that with experimentally measured rates of recombination as a function of sequence divergence, we can only expect "distance-scaled recombination to reinforce and maintain genetic separations which are initially created by allopatry or niche differentiation, but not to generate them" (Fraser et al. 2007, p. 2043). And there are many means by which LGT can abrogate such barriers.

Fred Cohan has long defended a recombination-independent model for bacterial speciation based on periodic selection (see Cohan and Perry 2007 for a recent formulation). He imagines that within ecologically defined populations, "genomic coherence" is maintained by selective sweeps driven by favorable fitter-type mutations—as in the original periodic selection chemostat experiments of Atwood and colleagues (1951). Since (in Cohan's model) there is little recombination, such sweeps purge diversity at unselected loci, maintaining coherence within—while driving divergence between—ecologically defined populations. He considers these populations, which he calls ecotypes and which comprise much tighter clusters than most existing named bacterial "species," to be the microbial analog of animal species.

Real bacterial evolutionary behavior will be variable and complex, sometimes approaching the strictures of the biological species concept, sometimes Cohan's ecotype model, but in both cases, constrained by the vagaries of phage availability, the specificity and avidity of conjugation and DNA uptake machineries, and the wild card that is LGT (Doolittle and Zhaxybayeva 2009). There is no reason to suppose that either concept (or any model with a finite number of parameters) will apply to all prokaryote cells, or that when different mechanisms do effect some degree of genomic or phenomic cohesion, this will correspond to any generally accepted and uniform cutoffs in terms of similarity in gene sequence, gene content, or breadth of niche. Indeed, frequent transitions between clonal- and recombination-based behavior have long been part of a sophisticated understanding of prokaryotic population genetics (Feil and Spratt 2001). Thus the species "category" (comprising all entities we might call species) can have no uniform or conserved properties. There is no reason not to consider everything—from a single cell possessing the only copy of its genome within the milliliter of seawater it inhabits, up to all the cells we might call *Prochlorococcus* in all the world's oceans—a significant player in the ecological drama, each at its own level. Such questions as how many species there are in some particular environment or on this planet, or whether bacterial species are cosmopolitan or endemic, have no answers that do not depend on an arbitrary framing of the question. However, the interests in diversity, differentiation, and dispersal that impel us to ask these questions can be addressed through "species-independent" metagenomics.

Metagenomics as culture- and "species"-free biology

Although some metagenomic investigators present results in the context of species ontologies based on genomic cohesion or ecotype differentiation (Whitaker and Banfield 2006, Cohan and Perry 2007, Hunt et al. 2008, Konstantinidis and DeLong 2008), we can always understand and make predictions from such data without reference to species as a category. In particular, questions about diversity, dispersal, and differentiation of niches can be recast as questions about the level of clustering at which clades become uniform with respect to some particular property (one simple representation of which is figure 3), or the extent to which clusters based on one character (for instance ANI) correspond to those based on another (for instance shared phenome, as in figure 2b). Answers will vary with parameters (marker genes and traits clustered), yet this is all we can expect or need from the data to understand diversity, dispersal, or differentiation. If such measures converge, they can identify potential units of biological organization—but this convergence could occur with entities as lowly as the single cell or as complex as a multilineage ecosystem, and it is up to us what degree of convergence is required to recognize such units as natural.

Diversity. The way we quantify diversity depends on what we choose as a unit (Shaw et al. 2008). In recognition of the problematic nature of the word "species," most diversity studies to date speak of operational taxonomic units (OTUs), usually defined—perhaps not coincidentally—by at least 97% identity of SSU rRNA sequences. Since it is impossible to sequence everything, overall diversity is usually inferred from a limited sample of sequences (see Bohannan and Hughes 2003 or Lozupone and Knight 2008), by either parametric or nonparametric methods. In parametric methods, sample data are fitted into an OTU-abundance distribution. The latter can be either set a priori or chosen through Bayesian methods (Quince et al. 2008). The total diversity will be the area under the OTU-abundance distribution curve. In nonparametric methods, the relative abundance of sampled OTUs is used to extrapolate the total abundance. For example, analyses

of 16S rRNA sequences from the Global Ocean Survey by these methods predict up to 3000 OTUs (Quince et al. 2008). Examination of a hypervariable region of 16S rRNA from deep-sea hydrothermal vents revealed an enormous diversity of OTUs (a "rare biosphere"), most at very low abundance levels (Sogin et al. 2006, Huber et al. 2007), with predictions of 50,000 to 300,000 OTUs (Quince et al. 2008).

All such estimates rely on a human-set cutoff to define the OTU. It is tempting and common to believe that some OTU cutoff values will prove more "natural" than others by corresponding to discrete genotypic clusters. But the rare biosphere discovery raises the possibility that there are few real gaps in diversity, only local variations in abundances.

In some instances, phylogenetic information is used as an alternative to arbitrarily delimited OTUs. In this case, the clustering takes into account variable phylogenetic distances between sequences, and does not rely on a preset divergence cutoff. Diversity is assessed by examining the shapes of phylogenetic trees; for example, by lineage-through-time plots (Martin 2002). Questions to be addressed with such trees are (a) Are there nonrandom patterns of clustering? (b) How do we demarcate such patterns? and (c) Is there an underlying process that would lead to their formation?

Inspection of phylogenetic trees of various prokaryotic groups reveals some clustering, although sometimes the groups are fuzzy and require outside knowledge, such as the assigned names of the examined taxa (Hanage et al. 2006a, Papke et al. 2007). Quantitative assessment of OTU clustering at different cutoff rates in a bacterioplankton rRNA sample revealed a discontinuity (gap) at 99% similarity cutoff (see figure 1 in Acinas et al. 2004). A similar diversity pattern was observed in *Bacillus* populations isolated from Israel's Evolution Canyon (Koeppel et al. 2008). Stephen Giovannoni (2004) and Koeppel and colleagues (2008) noted that these data fit the ecotype model of Cohan and Perry (2007). However, this does not preclude a fit to a simpler null model of random birth and death.

Theoretically, the divergence times in populations governed only by genetic drift follow an exponential distribution (Martin 2002); that is, lineage-through-time plots should look linear if the number of lineages is plotted on a logarithmic scale (figure 1 in Martin 2002). If, however, we consider only extant lineages (all that we have, in the absence of sequence data from prokaryotic fossils), the number of lineages grows faster than exponentially under equal birth and death rates (Zhaxybayeva and Gogarten 2004). It has yet to be shown that the distribution under this simple neutral model is significantly different from observed OTU clustering from microbial sampling. Effects of recombination (LGT) and sampling, as well as relative rates and constancy of birth and deaths on lineage-through-time plots, can also be substantial (Harvey et al. 1994, Martin et al. 2004, Zhaxybayeva and Gogarten 2004, Pagel et al. 2006). Simulations of bacterial populations under such birth-death models with mutation and recombination reveal that clusters form (and go extinct) at many combinations of mutation and recombination parameters (Hanage et al. 2006b).

Discussions of clustering in environmental sequence data are reminiscent of the debate that occurred in the field of paleobiology in the 1970s and 1980s on whether observed clades are formed by stochastic processes or not (Gould et al. 1977, Stanley et al. 1981). Since it is not yet clear if the observed clusters are "natural," we should ask whether we actually need them to analyze the metagenomic data sets in an environmental microbiology context. Perhaps new methods could be devised that "integrate out" cluster as a variable, and concentrate efforts on only relative differences in community composition (so called beta-diversity). Shaw and colleagues (2008) showed that community samples could be meaningfully compared without invoking total diversity, and noted that correlation between environmental variables depends primarily on how OTUs are defined.

Differentiation (niche partitioning). Adam Martiny and colleagues (2008) recently described a mapping of selected phenotypes to divergence of internal transcribed spacers between rRNA genes in various oceanic *Prochlorococcus* populations. With this more sensitive (rapidly evolving) phylogenetic marker, and logic as shown in figure 3, they found that "light correlates with broad-scale diversity (90% cutoff), temperature with intermediate scale (95%), whereas no correlation with phosphate was observed" (p. 823). From a similarly motivated metagenomic study of resource partitioning among marine vibrios, Hunt and colleagues (2008) concluded that these comprise "numerous ecologically distinct populations at different levels of phylogenetic differentiation," but caution that "using marker genes to assess community-wide diversity may not capture some ecological specialization," since this can be so fine-grained (as suggested from figure 2b). Again, metagenomic analysis may support a more nearly continuously graded mapping of ecotypic properties to genomic divergence than would be consistent with a robust notion of species. The AdaptML analysis developed by Polz's team (Hunt et al. 2008) holds great promise for such work, we think. As they describe it, AdaptML aims "to identify populations as groups of related strains sharing a common projected habitat, which reflects their relative abundance in the measured environmental categories (size fractions and seasons)" (p. 1083). In practice, the model inputs are the phylogeny, season, and size fraction of the strains. AdaptML then maps changes in environmental preference onto the tree by predicting projected habitats for each extant and ancestral strain in the phylogeny.

When microbial censuses are taken from separated but otherwise "identical" habitats, the level of concordance among metagenomic-sequence-defined functional and phylogenetic parameters will be the proof of the pudding for the utility of metagenomics. The most extensive of such characterizations will come from comparisons of the gut microbiomes of humans and other animals, since there will be special emphasis on establishing strong statistical correlations between "who is there" (in terms of phylotype, at whatever level of refinement); "what are they doing" (the functions of genes present);

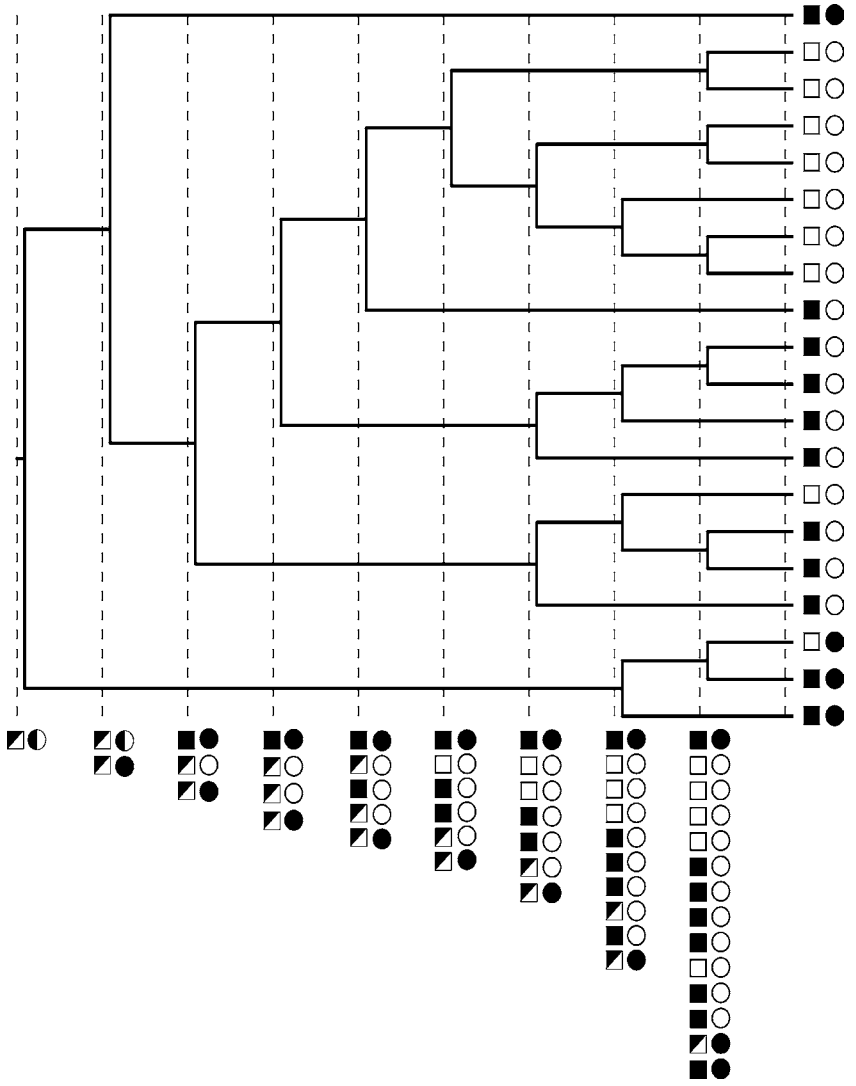

Figure 3. Establishment of correlation between a trait (functional or spatial) and an evolutionary history of a gene (tree backbone). Two traits, denoted as circles and squares, are considered to be of two types (black and white). Assessment of clade traits is performed at different cutoffs of clustering (vertical dashed lines). At each cutoff level the clade composition with respect to the trait is summarized as pure (black or white shapes) or as mixed (half-shaded shapes). At each level a test for significance of the covariation between gene's evolutionary history and a trait can be performed (e.g., phylogenetic test of Martin 2002). This example also illustrates that a different trait (circle versus square) exhibits best correlation to gene history at different clustering cutoffs.

and the nutritional, health, and genetic status of the hosts. A recent review of SSU rRNA data by Ley and colleagues (2008) reported that vertebrate gut microbiota are reproducibly correlated with host phylogeny, host morphology, and diet, and that we humans are typical omnivorous primates. Perhaps the most refined human study to date is that of Turnbaugh and colleagues (2009) on the gut microbiota of obese and lean twins. The strongest correlation was familial: Individuals from the same family had more similar microbiota, regardless of whether they cohabited—but the researchers detected greater similarity between samples taken from the same individual at a two-month interval. Obesity resulted in a shift in taxonomic structure (fewer Bacteroidetes and more Actinobacteria). Surprisingly, these authors' functional metagenomic analysis, based on pyrosequencing, showed that what is stable across time and between individuals is metabolic potential, not phylogenetic composition. They wrote:

> The hypothesis that there is a core human gut microbiome, definable by a set of abundant microbial organismal lineages that we all share, may be incorrect: by adulthood, no single bacterial phylotype was detectable at an abundant frequency in the guts of all 154 sampled humans. Instead, it appears that a core gut microbiome exists at the level of shared genes, including an important component involved in various metabolic functions. This conservation suggests a high degree of

> redundancy in the gut microbiome and supports an ecological view of each individual as an 'island' inhabited by unique collections of microbial phylotypes: as in actual islands, different species assemblages converge on shared core functions provided by distinctive components. (p. 483)

Dispersal. Microbial biogeography, an offspring of traditional biogeography of macroorganisms, has inherited a taxonomically centered framework to address the questions of spatial distribution of microorganisms and its causes. That is, the question is considered to be about species, and whether they are cosmopolitan (their global distribution being unrestrained, with local prevalence environmentally determined), or endemic (dispersal rates slow compared with divergence rates). A common belief is that bacterial species are cosmopolitan—"everything is everywhere but the environment selects" (O'Malley 2007)—because microbes are so much more easily transported, and have enormously larger populations, than kangaroos and elephants, for instance.

Unsurprisingly, numerous studies have confirmed that microbes are nonrandomly distributed across ecologically different environment types (see table 1 in Martiny et al. 2006)—the environment does select. But, moreover, in two seminal studies of microbial biogeography, it has been shown that populations in the same type of environments (hot springs) on different continents were significantly different (Papke et al. 2003, Whitaker et al. 2003). Martiny and colleagues (2008) were among the few to actually calculate the expected rates of sequence divergence versus cell dispersal produced as a biogeographic signal, and concluded that "*Prochlorococcus* cells may evolve faster than ocean currents can mix them, which results in locally distinct microdiversity" (p. 5).

Observed patterns are dependent on the units involved in examination of biogeographic patterns; however, the broader the examined units, the more cosmopolitan they become. Green and Bohannan (2006) proposed that the discussion concerning the spatial scaling of microbial biodiversity be recast. "Rather than ask the question, 'Do microbes have fundamentally different scaling relationships from those of plants and animals?', we suggest that the debate focus instead on the question, 'Is there a spatial scale, a degree of sampling effort and a level of taxonomic resolution at which microbial biodiversity scaling relationships approach those of macroorganisms?'" (p. 506). Such an approach is illustrated in figure 3, and metagenomic data will be uniquely useful in discovering if there is a "special" taxonomic cutoff at which correlations with geographic separation and environmental factors are found within and between gene data sets.

A new ontology: Ecosystems as units of biological organization

The island model for the human gut functional metagenome that Turnbaugh and colleagues (2009) sketched (see above) suggests a radical model for understanding the ontological status of communities, in which they are seen as "real"—as discrete spatiotemporal entities, targets of selection and categories within the hierarchy of biological organization—perhaps more real than the phylogenetic lineages or "species" that are their members. In other words, it may be community composition in terms of functional genes that is preserved by selection, not the phylogenetic identity of the bearers of the genes; the song, not the singer, matters. Supporting this view of the gut microbiome is a recent report, also from the Gordon group, that at least for two types of "carbohydrate-active" genes, gut bacterial and gut archaeal genomes have converged in gene content by LGT (Lozupone et al. 2008). It is (almost) as if this community were a superorganism, recruiting genes to maintain itself from a compositionally fluid collection of organismal lineages whose own evolutionary trajectories can be taken as largely irrelevant (to us and to the superorganism).

A similar but longer-term vision of the oceanic microbiota was articulated by Falkowski and colleagues (2008) in a review of the global biogeochemical cycles that keep our planet habitable. They wrote: "In essence, microbes can be viewed as vessels that ferry metabolic machines through strong environmental perturbations into vast stretches of relatively mundane geological landscapes. The individual taxonomic units evolve and go extinct, yet the core machines survive surprisingly unperturbed" (p. 1038).

Instead of viewing communities as more or less stable assemblages of lineages, detectable through SSU rRNA phylotypes, we should consider them collections of biochemical activities and their respective gene families. Comparative studies are still in their infancy, but promising in this regard. Dinsdale and colleagues (2008), in comparing 15 million random reads obtained by pyrosequencing 45 metagenomes from 9 environment types, reported: "Most of the variance between the different environments (79.9% of the combined microbiome and 69.9% of the virome) was explained in this analysis, showing that metagenomes are highly predictive of metabolic potential within an ecosystem. In contrast, a recent analysis of 16S rRNA genes from multiple environments only explained about 10% of the variance, suggesting that different ecosystems cannot be distinguished by their taxa" (p. 630).

Appropriate analyses for studying communities include trait-based biogeography, as described by Green and colleagues (2008). By establishing correlations with gene family representation across a range of parameters, trait-based biogeography can identify interactions within communities in the same way that two-hybrid screening or synthetic lethal genetic interaction mapping has established an "interactome" of yeast. And the regularity and predictive values of such correlations should help us decide how cohesive and stable an assemblage of microorganisms needs to be before we call it a "community."

As with systems biology, as described by O'Malley and Dupré (2005), one can imagine two sorts of attitude to such practices, the pragmatic and the (eco)systems theoretic. The former sees a community as an organizing epistemology that

values multilevel explanation, while the latter would both enshrine communities as ontologically real, in the sense that organisms are considered to be, and urge that we contemplate making general statements about communities that go beyond the parameters by which we chose to define them. Here we must agree with Dupré and O'Malley (2007), who wrote:

> All biologists and philosophers of biology know the difficulties of uniquely dividing different groups of organisms into species or genomes into genes, and both communities of investigators are divided about the [inevitability] of pluralism or the possibility of defining natural kinds. Our view is that these problems reflect a more fundamental difficulty, that life is in fact a hierarchy of processes (e.g., metabolic, developmental, ecological, evolutionary) and that any abstraction of an ontology of fixed entities must do some violence to this dynamic reality. Moreover, while the mechanistic models that are constructed on the basis of these abstracted entities have been extraordinarily valuable in enhancing our understanding of life processes, we must remain aware of the idealized nature of such entities, and the limitations of analogies between biological process and mechanism. (p. 835)

In our own view, the ultimate acceptance of communities as real, ontologically, can come only with a demonstration that they serve as "units of selection." That is, we must be able to model metagenomes as "genomes of communities" rather than "communities of genomes." It may be that metagenomes are indeed both "replicators" and (through expression in the community metaphenome) "interactors" in the language of Dawkins as elaborated by Hull (1980). They may also comprise "lineages," if occupying stable niches and migrating collectively to newly established ones. But even where similar communities must be recruited from scratch (perhaps by amplification of appropriate members of the rare biosphere and assembling new genomes through LGT) each time a new niche of the same type appears, we might regard a community as a sort of virtual unit of selection. Theoreticians and biologists concerned with modeling symbioses as evolutionary units (McFall-Ngai 2001, Bouchard 2008), with treating humans as holobionts (comprising our animal bodies and microbial commensals; Zilber-Rosenberg and Rosenberg 2008), and with hierarchical selection theory more generally (Dupré and O'Malley 2009) can show us the way.

Acknowledgments

This work was supported through Canadian Institutes of Health Research (CIHR) grant MOP-4467 to W. F. D. We thank Maureen O'Malley for comments on the manuscript. The literature review and the writing of this manuscript were completed in January 2009.

References cited

Achtman M, Wagner M. 2008. Microbial diversity and the genetic nature of microbial species. Nature Reviews Microbiology 6: 431–440.

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF. 2004. Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430: 551–554.

Atwood KC, Schneider LK, Ryan FJ. 1951. Periodic selection in *Escherichia coli.* Proceedings of the National Academy of Sciences 37: 146–155.

Bohannan BJ, Hughes J. 2003. New approaches to analyzing microbial biodiversity data. Current Opinion in Microbiology 6: 282–287.

Bollmann A, Lewis K, Epstein SS. 2007. Incubation of environmental samples in a diffusion chamber increases the diversity of recovered isolates. Applied and Environmental Microbiology 73: 6386–6390.

Bouchard FDR. 2008. Causal processes, fitness, and the differential persistence of lineages. Philosophy of Science 75: 560–570.

Brent R. 2000. Genomic biology. Cell 100: 169–183.

Cohan FM, Perry EB. 2007. A systematics for discovering the fundamental units of bacterial diversity. Current Biology 17: R373–R386.

Dinsdale EA, et al. 2008. Functional metagenomic profiling of nine biomes. Nature 452: 629–632.

Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. Genome Research 19: 744–756.

Dupré J, O'Malley MA. 2007. Metagenomics and biological ontology. Studies in History and Philosophy of Science C: Biological and Biomedical Sciences 38: 834–846.

———. 2009. Varieties of living things: Life at the intersection of lineage and metabolism. Philosophy and Theory in Biology 1: e003.

Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320: 1034–1039.

Feil EJ, Spratt BG. 2001. Recombination and the population structures of bacterial pathogens. Annual Review of Microbiology 55: 561–590.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. Science 315: 476–480.

Friedmann HC. 2004. From "Butyribacterium" to "*E. coli*": An essay on unity in biochemistry. Perspectives in Biology and Medicine 47: 47–66.

Gevers D, et al. 2005. Opinion: Re-evaluating prokaryotic species. Nature Reviews Microbiology 3: 733–739.

Giovannoni S. 2004. Evolutionary biology: Oceans of bacteria. Nature 430: 515–516.

Giovannoni S, Stingl U. 2007. The importance of culturing bacterioplankton in the 'omics' age. Nature Reviews Microbiology 5: 820–826.

Gould SJ, Raup DM, Sepkoski JJ, Schopf TJM, Simberloff DS. 1977. The shape of evolution: A comparison of real and random clades. Paleobiology 3: 23–40.

Green JL, Bohannan BJ. 2006. Spatial scaling of microbial biodiversity. Trends in Ecology and Evolution 21: 501–507.

Green JL, Bohannan BJ, Whitaker RJ. 2008. Microbial biogeography: From taxonomy to traits. Science 320: 1039–1043.

Hanage WP, Fraser C, Spratt BG. 2006a. Sequences, sequence clusters and bacterial species. Philosophical Transactions of the Royal Society B 361: 1917–1927.

Hanage WP, Spratt BG, Turner KM, Fraser C. 2006b. Modelling bacterial speciation. Philosophical Transactions of the Royal Society B 361: 2039–2044.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. 1998. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. Chemistry and Biology 5: R245–R249.

Handelsman J, et al. 2007. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. National Academies Press.

Harvey PH, May RM, Nee S. 1994. Phylogenies without fossils. Evolution 48: 523–529.

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML. 2007. Microbial population structures in the deep marine biosphere. Science 318: 97–100.

Hull DL. 1980. Individuality and selection. Annual Review of Ecology, Evolution, and Systematics 11: 311–332.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320: 1081–1085.

Jacob F. 1977. Evolution and tinkering. Science 196: 1161–1166.

Kettler GC, et al. 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genetics 3: e231.

Koeppel A, et al. 2008. Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. Proceedings of the National Academy of Sciences 105: 2504–2509.

Konstantinidis KT, DeLong EF. 2008. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME Journal 2: 1052–1065.

Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences 102: 2567–2572.

Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the genomic era. Philosophical Transactions of the Royal Society B 361: 1929–1940.

Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. 2008. Worlds within worlds: Evolution of the vertebrate gut microbiota. Nature Reviews Microbiology 6: 776–788.

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proceedings of the National Academy of Sciences 101: 11013–11018.

Lozupone CA, Knight R. 2008. Species divergence and the measurement of microbial diversity. FEMS Microbiology Reviews 32: 557–578.

Lozupone CA, Hamady M, Cantarel BL, Coutinho PM, Henrissat B, Gordon JI, Knight R. 2008. The convergence of carbohydrate active gene repertoires in human gut microbes. Proceedings of the National Academy of Sciences 105: 15076–15081.

Martin AP. 2002. Phylogenetic approaches for describing and comparing the diversity of microbial communities. Applied and Environmental Microbiology 68: 3673–3682.

Martin AP, Costello EK, Meyer AF, Nemergut DR, Schmidt SK. 2004. The rate and pattern of cladogenesis in microbes. Evolution 58: 946–955.

Martiny AC, Tai AP, Veneziano D, Primeau F, Chisholm SW. 2008. Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. Environmental Microbiology 11: 823–832.

Martiny JB, et al. 2006. Microbial biogeography: Putting microorganisms on the map. Nature Reviews Microbiology 4: 102–112.

Mayr E. 1942. Systematics and the Origin of Species. Columbia University Press.

———. 1996. What is a species, and what is not? Philosophy of Science 63: 262–277.

McFall-Ngai MJ. 2001. Identifying 'prime suspects': Symbioses and the evolution of multicellularity. Comparative Biochemistry and Physiology B: Biochemistry and Molecular Biology 129: 711–723.

O'Malley MA. 2007. The nineteenth century roots of "everything is everywhere." Nature Reviews Microbiology 5: 647–651.

O'Malley MA, Dupré J. 2005. Fundamental issues in systems biology. Bioessays 27: 1270–1276.

Pace NR. 1997. A molecular view of microbial diversity and the biosphere. Science 276: 734–740.

Pagel M, Venditti C, Meade A. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. Science 314: 119–121.

Papke RT, Ramsing NB, Bateson MM, Ward DM. 2003. Geographical isolation in hot spring cyanobacteria. Environmental Microbiology 5: 650–659.

Papke RT, Zhaxybayeva O, Feil EJ, Sommerfeld K, Muise D, Doolittle WF. 2007. Searching for species in haloarchaea. Proceedings of the National Academy of Sciences 104: 14092–14097.

Prosser JI, Nicol GW. 2008. Relative contributions of archaea and bacteria to aerobic ammonia oxidation in the environment. Environmental Microbiology 10: 2931–2941.

Quince C, Curtis TP, Sloan WT. 2008. The rational exploration of microbial diversity. ISME Journal 2: 997–1006.

Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson RV. 1998. Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. Journal of Bacteriology 180: 5003–5009.

Schmidt TM, DeLong EF, Pace NR. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. Journal of Bacteriology 173: 4371–4378.

Schubert S, Darlu P, Clermont O, Wieser A, Magistro G, Hoffmann C, Weinert K, Tenaillon O, Matic I, Denamur E. 2009. Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species. PLoS Pathogens 5: e1000257.

Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC, Martiny JB. 2008. It's all relative: Ranking the diversity of aquatic bacterial communities. Environmental Microbiology 10: 2200–2210.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proceedings of the National Academy of Sciences 103: 12115–12120.

Stackebrandt E, et al. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology 52: 1043–1047.

Staley JT, Konopka A. 1985. Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annual Review of Microbiology 39: 321–346.

Stanley SM, Signor PW, Lidgard S, Karr AF. 1981. Natural clades differ from "random" clades; simulations and analyses. Paleobiology 7: 115–127.

Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF. 1996. Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. Journal of Bacteriology 178: 591–599.

Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF. 2005. Genotypic diversity within a natural coastal bacterioplankton population. Science 307: 1311–1313.

Ting CS, Rocap G, King J, Chisholm SW. 2002. Cyanobacterial photosynthesis in the oceans: The origins and significance of divergent light-harvesting strategies. Trends in Microbiology 10: 134–142.

Touchon M, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genetics 5: e1000344.

Townsend JP, Nielsen KM, Fisher DS, Hartl DL. 2003. Horizontal acquisition of divergent chromosomal DNA in bacteria: Effects of mutator phenotypes. Genetics 164: 13–21.

Turnbaugh PJ, et al. 2009. A core gut microbiome in obese and lean twins. Nature 457: 480–484.

Welch RA, et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proceedings of the National Academy of Sciences 99: 17020–17024.

Whitaker RJ, Banfield JF. 2006. Population genomics in natural microbial communities. Trends in Ecology and Evolution 21: 508–516.

Whitaker RJ, Grogan DW, Taylor JW. 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science 301: 976–978.

Woese CR. 1987. Bacterial evolution. FEMS Microbiology Reviews 51: 221–271.

Zhaxybayeva O, Gogarten JP. 2004. Cladogenesis, coalescence and the evolution of the three domains of life. Trends in Genetics 20: 182–187.

Zilber-Rosenberg I, Rosenberg E. 2008. Role of microorganisms in the evolution of animals and plants: The hologenome theory of evolution. FEMS Microbiology Reviews 32: 723–735.

W. Ford Doolittle (ford@dal.ca) is with the Department of Biochemistry and Molecular Biology at Dalhousie University in Halifax, Nova Scotia, Canada. Olga Zhaxybayeva is with Environmental Proteomics NB, Inc., in Sackville, New Brunswick, Canada.