Review

CellPress

Omics: Fulfilling the Promise

Ordering microbial diversity into ecologically and genetically cohesive units

B. Jesse Shapiro[1] and Martin F. Polz[2]

[1] Département de Sciences Biologiques, Université de Montréal, Montréal, QC H3C 3J7, Canada
[2] Parsons Laboratory for Environmental Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

We propose that microbial diversity must be viewed in light of gene flow and selection, which define units of genetic similarity, and of phenotype and ecological function, respectively. We discuss to what extent ecological and genetic units overlap to form cohesive populations in the wild, based on recent evolutionary modeling and on evidence from some of the first microbial populations studied with genomics. These show that if recombination is frequent and selection moderate, ecologically adaptive mutations or genes can spread within populations independently of their original genomic background (gene-specific sweeps). Alternatively, if the effect of recombination is smaller than selection, genome-wide selective sweeps should occur. In both cases, however, distinct units of overlapping ecological and genotypic similarity will form if microgeographic separation, likely involving ecological tradeoffs, induces barriers to gene flow. These predictions are supported by (meta)genomic data, which suggest that a 'reverse ecology' approach, in which genomic and gene flow information is used to make predictions about the nature of ecological units, is a powerful approach to ordering microbial diversity.

Introduction and motivation

It is often said that species are fundamental units of ecology because they comprise individuals that are phenotypically and hence ecologically more similar to each other than to other species [1,2]. This notion was extended in Mayr's biological species concept [3], which states that species are reproductively isolated units, implying that adaptive mutations can spread within a species leaving other coexisting species unaffected. Although recent evidence has shown that reproductive boundaries can be

Corresponding authors: Shapiro, B.J. (jesse.shapiro@umontreal.ca); Polz, M.F. (mpolz@mit.edu).

Keywords: population genomics; ecological differentiation; reverse ecology; gene flow; selective sweeps; mosaic sympatric speciation.

0966-842X/$ – see front matter

 http://dx.doi.org/10.1016/j.tim.2014.02.006

Glossary

Allopatric: a set of sampled isolates or genomes from different geographic areas, where barriers to migration and gene flow are significant.
Clonal frame: the portion of the genome transmitted by vertical (clonal) evolution, unimpacted by HGT. Mutations in the clonal frame should all fall parsimoniously on a single phylogenetic tree.
Core genome: the portion of the genome that is present (or in practice, that can be aligned) in all of a given set of sequenced isolates or metagenomes.
Flexible genome: the set of genes or DNA that is present in only a fraction of a given set of sequenced isolates or metagenomes.
Gene-specific selective sweep: the process in which an adaptive gene or allele (possibly a niche-specifying variant) spreads in a population by recombination faster than by clonal expansion. The result is that the adaptive variant is present in more than a single clonal background, and that diversity is not purged genome-wide.
Genome-wide association study (GWAS): a technique commonly used in eukaryotic genetics to identify genomic variants that are associated with a phenotype of interest. In highly structured populations (e.g., clonal microbes), it is essential to correct for false associations due to phylogenetic structure.
Genome-wide selective sweep: the process in which an adaptive gene or allele (possibly a niche-specifying variant) spreads in a population by clonal expansion of the genome that first acquired it. The result is that diversity is purged genome-wide, and that the adaptive variant is linked in the same clonal frame as the rest of the genome.
Horizontal gene transfer (HGT): the incorporation of foreign DNA into a genome. Incorporation can be mediated by either homologous recombination or nonhomologous recombination of DNA that enters a cell via transformation, transduction, or conjugation. In bacteria and archaea, all gene transfer is horizontal (i.e., always unidirectional).
Homologous recombination: a mechanism of DNA integration requiring at least short tracts of identity between the genome and the foreign DNA, mediated by RecA (protein necessary for DNA repair, recombination, and maintenance) and mismatch repair machinery. The integrated DNA can result in single nucleotide changes and, in some cases, addition or loss of relatively long stretches of DNA including entire genes.
Metagenome: the total set of all genomic DNA in a particular environment or sample.
Negative frequency-dependent selection: a type of natural selection that favors rare phenotypes in a population.
Niche: a specific set of ecological parameters (environments, resources, physical and chemical characteristics, biotic interactions, etc.) to which an organism is adapted. This does not necessarily imply (but does not exclude) physical separation between niches.
Niche-specifying variant: a mutation, gene, or allele that allows a cell to be part of a particular niche. These variants are under positive selection within the particular niche, but not outside it.
Nonhomologous recombination: integration of DNA with no homologous allele already present in the genome, often mediated by phage and integrative elements. This results in the acquisition of entirely new genes.
Population: a group of individuals sharing genetic and ecological similarity, and coexisting in a sympatric setting.
Positive selection: a type of natural selection that favors variants conferring a fitness advantage, causing them to increase in frequency in a population.
Sympatric: a set of sampled isolates or genomes from the same geographic area, where barriers to migration and gene flow are low or nonexistent.

leaky [4–6], species are still regarded as congruent genetic and ecological units for sexual eukaryotes, even if hybrids and intermediate forms are common [7]. For bacteria and archaea, however, the situation has been marred by several complicating factors that question whether such units can be defined.

In addressing whether we can identify genetically and ecologically congruent units, we need to take into account the peculiarities of bacterial and archaeal evolution, that is, the varying modes and rates of genetic exchange. In these organisms, incorporation of new genetic material is always unidirectional and leads either to gene conversion by homologous recombination or gene addition by nonhomologous recombination (see Glossary). (In fact, the distinction might not be so clear: there is mounting evidence that homologous recombination is often involved in gene addition and loss [8–11].) Importantly, the rates and bounds of this gene transfer can vary considerably. Although some lineages follow a highly clonal mode of evolution, in others, rates of recombination can differ by several orders of magnitude. Regardless of the overall rate of gene flow, genetic material can, in principle, be incorporated from distantly related organisms. This variation in genetic exchange and its effect on genotypic integrity and ecological adaptation is at the heart of the debate about what constitutes ecological and genetic units for bacteria and archaea.

In particular, horizontal gene transfer (HGT) among distantly related organisms can create genotypes that vary in properties of ecological relevance by acquiring functions, such as antibiotic resistance or nitrogen fixation, that distinguish them from otherwise closely related genotypes [4,12]. At the same time, the recipient genotype has also become ecologically similar, in at least one niche dimension, to the organism from which it acquired the novel pathway. In fact, such functional differentiation is observed among closely related environmental isolates [13] and, in combination with high gene turnover, has been taken as evidence that gene acquisition and loss is so high as to quickly erode any niche association of lineages [12]. By extension, the very notion of a lineage has been questioned on the same grounds – with the consequence that nearly each genotype might represent its own, independent ecological unit [14] that can only be recognized by the functional genes it carries [15].

In recent years, however, analysis of environmental isolates and metagenomes has shown that microbial communities consist of genotypic clusters of closely related organisms and that these can display cohesive environmental associations and dynamics that clearly distinguish them from other such clusters coexisting in the same samples. Despite also showing evidence for extensive gene flow, genetically distinguishable clusters have been observed among closely related environmental and pathogenic isolates by multilocus sequence analysis and genomics [1,16,17] and by metagenomics [18–20]. Moreover, cohesive ecological dynamics and associations have been demonstrated for a growing number of cases, including for vibrios, sulfate-reducing bacteria, and cyanobacteria, as well for organisms represented in several marine, freshwater, and acid-mine drainage community metagenomes. These observations suggest congruence of genotypic and ecological units and are, in principle, consistent with the notion of populations as locally coexisting members of a species. As we will discuss below, selection and recombination are paramount in shaping and maintaining such units, although the effects of biogeography, on both local [20,21] and global [22,23] scales may also come into play.

The idea that genotypic clusters should be rapidly eroded by HGT might in part be an artifact of early comparative studies of quite anciently diverged genomes. In these, only a fraction of genes in the core genome showed phylogenetic congruence, and the flexible genome seemed to be completely unrelated [12,24]. Moreover, we often call organisms closely related if their 16S rRNA genes, which are commonly used as taxonomic markers, show few percent nucleotide differences, yet such difference may indicate millions of years of separate evolution with associated large genome changes [25]. But even as closely related genomes (e.g., identical in 16S rRNA genes) began to be sequenced, these usually were not isolated from the same habitat and hence were not part of the same populations of interacting genotypes. This means that the effect of environmental selection might not be easily disentangled from genetic divergence due to geographic separation [26]. For example, in the marine cyanobacterium *Prochlorococcus*, populations in the Atlantic contain genes responsible for efficient phosphorus acquisition that are absent from populations in the Pacific [27]. Hence these genes are part of the core genome of Atlantic populations but would be judged flexible genes if closely related isolates were compared from both ocean regions. We therefore believe that an important step forward will be to emphasize population thinking in microbiology by assembling genomic datasets that represent clusters of close relatives co-occurring in the same environment – because only these will allow interpretation of how environmental selection acts on genomes from within the same population.

The challenge is then to develop an understanding of how genotypic clusters originate and are maintained, and whether they are selectively optimized to occupy sufficiently different niches to coexist with other clusters. Importantly, any such attempt needs to take into account the considerable genotypic diversity encountered in environmental populations, which often consist of genomes differing by a considerable fraction of their gene content and displaying large allelic diversity even if most of their genes suggest close relationships [26].

In this review, we begin by discussing the extent to which ecological and genetic units overlap, and under what circumstances genetic units can be used as a proxy for ecological units. We argue that although it is essential to sequence populations of microbial genomes and record ecological metadata, a powerful alternative is represented by a 'reverse ecology' approach in which genomic and gene flow information is used to make predictions about the nature of ecological units (Box 1). What distinguishes reverse ecology from the broader field of ecological genomics is its focus on simultaneously predicting ecological and genetic units, rather than mapping ecological data onto predefined genetic units. These predictions can then be tested using ecological metadata and experimental