Metagenomics and Its Applications in Agriculture, Biomedicine and Environmental Studies

Robert W. Li
Editor

Nova Science Publishers, Inc.
New York
Chapter 11

AN EVOLUTIONARY ECOLOGY PERSPECTIVE ON COMPARATIVE METAGENOMICS

Dieter Ebert and Weihong Qi

ABSTRACT

The field of metagenomics explores genetic material recovered directly from the environment. So far, the evolution and ecology of these microbial communities are poorly understood. Here, we outline guiding principles to understand host-associated metagenomes in the context of their ecology and evolution. Our framework originates from studies of parasites and pathogens, which are expected to show similar transmission dynamics as members of the microbial community of the host (=microbiota). First, we consider where the host acquires the members of its microbiota. We explore the differences between vertical and horizontal transmission among hosts and discuss the role of the hosts as isolated habitat patches in a metapopulation context. We outline the relation of metapopulation models, the theory of island-biogeography and colonization models for our understanding of microbiota diversity. Second, we discuss a coevolutionary approach to the study of metagenomics, based on a phylogenetic perspective. Tools developed for phylogenetics and comparative metagenomics can identify patterns of coevolution between hosts and symbionts. This approach allows one to understand specificity and coevolution of microorganisms in relation to closely related host species. Finally, we discuss advantages and challenges of the core metagencode concept and suggest ways that this concept can be more accurately understood in an ecological and evolutionary context.

GLOSSARY

Horizontal transmission: Transmission among hosts other than parent to offspring, i.e., among unrelated hosts or from the environment to a host.
Microbiota: The community of microorganisms associated with a host. The microbiota is sometimes called meta-community.

Mutualisms: Form of interactive relationship between members of two species where both members have a net benefit.

Symbionts: In this chapter symbionts are small organisms and viruses associated with a host. The term does not differentiate between possible functions, such as parasites (including pathogens), commensals (without a positive or negative effect on the host) or mutualists (with a positive effect on the host). It also does not differentiate between taxa, such as bacteria, protozoa or viruses.

Symbiosis: Form of an intimate and lasting relationship between members of two species. Mutualism and parasitism are forms of symbiosis.

Transmission: Passing of a symbiont from one host individual to another. It may also be used for passing of symbionts from the environment to a host.

Vertical transmission: Transmission of a symbiont from the mother (or more rarely the father) to the offspring. Vertical transmission may be tightly connected to reproduction (e.g., trans-ovarial transmission) or may be a consequence of the spatial structure of the host population, where offspring have a higher likelihood of receiving a symbiont from their parents than from other unrelated hosts.

INTRODUCTION

For centuries, organismal diversity could be distinguished only on the basis of visibly distinct characters. This limitation has been particularly problematic for the study of microorganisms. Although technological advances in microscopy have provided increasingly fine resolution, techniques based solely on visualisation are practically and inherently limited in scope. The advent of molecular genetics has radically changed the way we study biodiversity. Genetic polymorphisms can resolve taxa with no visible distinguishing features. Indeed, we can obtain and analyze DNA sequences without ever seeing the organism to which these sequences belong. Coupled with powerful new high-throughput sequencing technologies, this approach allows the qualitative and quantitative genetic description of entire microbial ecosystems (Beardsley, 2006). However, the conceptual background of understanding the biodiversity of microbiotas is not well developed.

Studies of metagenomes concern the analysis of samples collected from delimited habitat patches. Such habitat patches are usually characterized by selected criteria, such as geological features (e.g., soil types), chemical features (e.g., mine drainage, pH), ecological features (e.g., sea water, deep sea vents) or taxonomic classification of the host organism (e.g., human, mouse, Drosophila). Metagenomic studies of multicellular host organisms have been of particular interest, as humans host a rich diversity of microbes on various tissues including our skin, mouth and intestine. The nature of these microbes and their importance for human health and development is an area of active research (Eckburg et al., 2005; Ordovas and Mooser, 2006; Turnbaugh et al., 2007). Here we limit our discussion to the microbiotas of host organisms, excluding free-living communities such as those found in samples from water or soil, sometimes collectively called environmental samples. The community structure of within host microbiotas differs from strictly defined environmental samples in several
important aspects (Box 1). First, the host microbiota is temporally highly variable and can undergo dramatic changes throughout development and differentiation of the host. Second, hosts as habitat are heterogeneous. This is relevant for sexual host species, where every individual is genetically unique, but also for asexual hosts with population differentiation. If the microbiota interacts in a specific way with particular host genes or genotypes, it will differ from one individual host to the next. In this case, the degree to which host individuals resemble each other (relatedness) becomes important (Turnbaugh et al., 2009). This idea extends from genetic relatedness within species to phylogenetic relatedness across species and entire clades. For example, two mammals are likely to have more similar microbial metagenomes than a mammal and a fish. Heterogeneity among hosts is also a consequence of variation in host ecology, such as host habitat and diet (Ley et al., 2008). Finally, most hosts are spatially well-delimited habitat islands. This insularity of hosts has important consequences for the structure of the microbiota.

**Box 1. Aspects of host biology, which may shape the community structure of their microbiotas**

<table>
<thead>
<tr>
<th>Host development and differentiation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontogenesis: Multicellular hosts usually undergo development starting from a single, undifferentiated cell and ending with a complex organism. The single cell stage has no or very few symbionts, and the microbiota increases during the development of the host.</td>
</tr>
<tr>
<td>Aging and death: Hosts have a limited lifespan - they age and die. Microbiotas may be influenced by aging, and any feature that evolved specifically to a given individual may be lost upon the death of the host.</td>
</tr>
<tr>
<td>Tissue specificity: A key feature of development is the differentiation of specialized tissues and spatial structure, both of which are important for the microbiota, which is highly tissue specific.</td>
</tr>
<tr>
<td>Host gender: Many host species have separate sexes, which may differ in many aspects, including behavior, physiology (e.g., hormones, nutrition), ecological niche, body size, and tissue types.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Host genetics:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host individuals are often unique in their genetic make-up: Genetic recombination provides almost unlimited scope for building specific host genotypes, which may play a role in shaping the microbiota of an individual.</td>
</tr>
<tr>
<td>Host individuals are related to one another: Individuals may form kinships, belong to the same population, or belong to different sub-species. Species can be grouped in increasingly broader taxonomic units, ranging from genera to kingdoms. Parts of microbiotas may reflect the ancestry and taxonomic position of their hosts.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Host ecology:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habitat and transmission: Host habitats differ in the degree to which symbionts can be transmitted locally and globally. For example, transmission among aquatic hosts may be easier and more reliable than among terrestrial hosts.</td>
</tr>
<tr>
<td>Niche: Different host taxa, even distantly related species, may share similar niches and...</td>
</tr>
</tbody>
</table>
exist as part of the same transmission network. The microbiotas of unrelated hosts may resemble each other due to shared habitats or resources.

Nutrition: Host diet has a large influence on the gut microbiota. Hosts with similar diets may share parts of their microbiota.

**Spatial structure of hosts:**

Host population structure: The genetic and spatial structure of a host will determine to what degree symbionts are transmitted from host to host (e.g. from mother to offspring, among unrelated hosts, within or between local communities). Thus, microbiotas reflect the local transmission network.

Meta population structure: Hosts live in populations of various densities, where high densities may favor transmission. Transmission is more likely within than between host populations. Populations are arranged in clusters (metapopulations) or may be separated via long distances. Thus, hosts are nested in population, populations are nested in metapopulations, metapopulations may be nested in higher levels structure. With increasing spatial distance, transmission events become less likely.

In ecology, parasitology, and evolution, the species composition and richness of communities is of central relevance. Many studies take a comparative approach by quantifying differences or similarities of communities across units of interest. In metagenomic studies, these units are typically habitats or host individuals. Here, we integrate fundamental concepts from ecology, epidemiology, parasitology and evolutionary biology to outline a comparative approach to the study of microbial diversity associated with multicellular host organisms. We begin by considering the question "Where do they come from?", while in the second part we address the taxonomic composition of the microbiota from a coevolutionary perspective. Much of the reasoning of the approach we take in this study comes from ecological and epidemiological studies concerned with the spread and distribution of parasites (including pathogens). Parasites can be seen as part of the microbiota of hosts and their dynamics are similar to those of other members of the microbiota. Thus, we propose that much of the dynamics of microbiotas can be understood from an epidemiological perspective.

Functional aspects in an evolutionary framework are concerned with the effects of symbionts on host fitness and vice versa. A common approach to study fitness effects is by eliminating all symbionts associated with a host. This has been successful done in many taxa (Dufty, 1976; Xu and Gordon, 2003; Marques et al., 2006) and revealed that symbiont-free host organisms are often weak, but can recover after the addition of microbes from the normal environment (Rahat and Dimentman, 1982; Douillet, 2000; Marques et al., 2006). Among the typical symptoms seen in axenically raised organisms are undernutrition and a poorly developed immune system, leading to low fertility and survival (Marques et al., 2006). Such findings illustrate that symbionts can be essential for the normal health and development, and therefore fitness, of hosts, stressing their role as an important evolutionary agent. It is however, not clear to what degree these essential microbes are specific to the host species.
WHERE DO THEY COME FROM?

One of the most exciting insights from metagenomic studies is the overwhelming diversity of symbionts found in many host species (Woyke et al., 2006; Booijink et al., 2007; Cox-Foster et al., 2007; Schmitt et al., 2007). For example, the human intestine is colonized by more than one thousand microbial species. Where do these symbionts come from? There are three possible sources. First, symbionts can be transmitted vertically (Glossary) from mother to offspring. This is known to be true for some microbes, however, most hosts are not born with a diverse microbiota, indicating that parts of the symbiont diversity is acquired later in life. Second, symbionts can be transmitted horizontally (Glossary) from other hosts, such as family members or other unrelated individuals in the vicinity. Finally, members of the microbiota may also occur as free-living microbes, which can be taken up (also by horizontal transmission) by breathing, eating, or other contact with the environment.

Vertical Transmission

The question "Where do they come from?" is most easily answered for vertically transmitted symbionts. Vertical transmission refers to the passage of a symbiont from mother to offspring, or less commonly, from father to offspring. Vertical transmission in the strict sense, is limited to intracellular symbionts (i.e. those living in the cytoplasm of their host's cells), and transmission is therefore mostly uniparental. This is often called transovarial transmission. Chloroplasts and mitochondria are the hallmark examples of strict vertical transmission, but many other examples exist. For instance, aphids host several species of bacteria, which are passed from mother to offspring, with different mothers transmitting different communities (Moran et al., 2008; Vorburger et al., 2009). Many insects and nematodes vertically transmit the bacteria of the genus Wolbachia. While this symbiont is parasitic in many insects (Werren et al., 2008), they are mutualists in nematodes (Bandi et al., 1999). Modern techniques of visualization and genetic analysis continue to identify new examples of strictly vertically transmitted symbionts, particularly in invertebrate taxa (Vautrin and Vavre, 2009).

A clear distinction between horizontal and vertical transmission is not always possible. Defined more broadly, vertical transmission includes cases where symbionts are transmitted during birth, or during parental care, i.e. any form of maternal (rare paternal) transmission. A clear distinction between horizontal and vertical transmission is therefore not always possible, and it may be best to consider mixed modes of transmission. On a quantitative scale, it would be desirable to quantify the proportion of maternal transmission as part of the total transmission.

Discounting for the transovarially transmitted microbes in some host taxa, hosts are expected to be born without microbes and their microbiota would need to be acquired from free-living bacteria in the external environment or from other hosts later in life. However, there is now growing evidence for maternal transmission as parts of the microbiota. For humans, it was speculated that most intestinal microbes are acquired from other humans, rather than the environment (Ley et al., 2005) and that mothers are the likely source of most of these microbes (Mändar and Mikelsaar, 1996; Grönlund et al., 1999). Part of the human
microbiota is transmitted with the mother’s milk (Martin et al., 2007). Significant maternal transmission was also suggested for animals, e.g. sponges (Lee et al., 2009) and corrals (Zilber-Rosenberg and Rosenberg, 2008). Mother-offspring transmission of mouse microbiota was suggested to be so stable over generations that kinship relationships are reflected in community composition (Ley et al., 2005; Ley et al., 2006). Thus, because of mother-offspring transmission, microbes may be in the position to evolve lineage specificity, i.e. evolve specific adaptation to a given host line. In case of mutualistic relationships (benefits for host and microbe) this in turn favours hosts-microbe system with more reliable mother-offspring transmission, because this is advantageous for both host and microbe. Thus, mutualistic interactions and maternal transmission are expected to coevolve. In those cases where strict associations persisted of long time periods, congruent evolution of host and microbe lineages is expected. This has been documented for *Wolbachia* in nematodes, *Buchnera* in Aphids, and *Wigglesworthia* in flies (Bandi et al., 1999; Chen et al., 1999; Moran and Baumann, 2000). The longer host and symbiont remain tightly linked by vertical transmission, the more intimate the interactions between the two species will become. This can sometimes lead to total interdependency, as is the case for *Buchnera* in Aphids (Moran, 1996), where none of the two partners can live without the other. Increased interdependencies have also been observed in evolutionary experiments with phages and bacteria where vertically transmission was artificially maintained for many generations (Bull et al., 1991; Bull and Molineux, 1992). In summary, the role of vertically transmission for host-symbiont ecology and evolution has certainly been underrated and deserves more attention with regard to the highly diverse microbiota.

**Horizontally Transmitted Symbionts: A Parasitological Epidemiological Approach**

Horizontal transmission is the transfer of symbionts from the environment or from one host to another, excluding the transfer among vertically related individuals, such as mother to offspring. Most pathogens and parasites are horizontally transmitted and it seems likely that this also holds for the majority of the members of microbioms. Parasitologists have long been concerned about questions related to the differences and similarities of communities of horizontally transmitted parasites when comparing host species and host populations. While there is hardly any multicellular organism that is not parasitized during some stage of its life, not every host or host population harbours every potential parasite species. Rather, most hosts coexist with only a small set of their potential parasites. A number of factors have been suggested to explain why some parasites are present in a given individual or population, while others are not (Price, 1980; Anderson and May, 1991; Hanski and Simberloff, 1997; Ebert et al., 2001; Poulin, 2007). These ideas can be extended to any symbiont community and therefore provide a good starting point for understanding symbiont richness in the context of metagenome studies.

**Epidemiology and the Metapopulation Perspective**

The first category of factors contributing to variation among symbiont communities relates to population size of the host. Larger host populations (often described by a larger
geographic range) have been shown to harbor more parasite species than smaller populations (Price and Clancy, 1983; Gregory, 1990; Dobson and Pacala, 1992). This may be because larger populations have higher absolute encounter rates with parasites, are less likely to pass through population bottlenecks (which may cause parasite and symbiont loss), are more likely to be above the parasite specific population threshold size (above which a parasite can persist (Anderson and May, 1991)) and may also be more dense on average. A parasite, or more generally, a symbiont can only persist in a host population if the number of secondary infections caused by transmission from a primary infection (usually denoted as $R_0$) is larger than one (Anderson and May, 1991). Thus, a member of the host microbiota has to be transmitted on average to more than one other host, or it will decline and eventually disappear from the population. Therefore, a key factor for transmission is the contact rate among hosts (direct contact, or indirect in case symbionts can survive in the environment for some period of time). Thus, higher host density is usually associated with higher transmission rates and therefore is expected to result in more diverse microbiotas. Related to the population size argument is the range of host species that a symbiont can colonize. With more hosts occurring in sympatry, the total host population size (and possibly total host density) becomes larger. Thus, more symbionts may be found on any given focal species because of a higher likelihood that a symbiont will persist in multi-host systems. The host population size-symbiont richness argument is analogous to the concept of species-area relationships from community ecology (Simberloff and Moore, 1996).

Individual hosts can be seen as habitat islands for symbiont populations (Kuris et al., 1980). Populations that are isolated by distance or some other barrier and are connected by dispersal form a metapopulation (Hanski and Gaggiotti, 2004). A population of hosts therefore comprises a metapopulation of symbionts. Different symbiont species may or may not be present on any given host island. Various ecological models concerning the presence or absence of species on islands have been described, based on factors influencing local extinction and colonization (MacArthur and Wilson, 1967; Kuris et al., 1980; Simberloff and Moore, 1996; Hanski and Simberloff, 1997; Hanski, 1999). Among the most discussed are the theory of island biogeography (MacArthur and Wilson, 1967) and the Levin-type metapopulation theory (Hanski and Simberloff, 1997; Hanski, 1999). The theory of island biogeography is based on islands of different size, located at different distances from a mainland, which contains all possible species. The number of species successfully colonizing the island from the mainland depends on both the size of the island and the distance from the mainland, with larger islands close to the mainland harboring more species. In metagenomics, the mainland translates to the pool of symbionts available in the environment. It is important to note, however, that this island biogeography model only applies to symbionts that can live or survive in a free-living state. In contrast, the Levin-type metapopulation model assumes that there is no continent, but only islands. Thus, a host can only receive colonizers from another host. Transmission can be either direct (e.g. skin contact) or indirect (e.g. via contaminated objects or surfaces or vectors). Because symbionts are spread from host to host, the density and spatial structure of the hosts (the islands) will have an important influence on the colonization dynamics. With increasing isolation, colonization becomes less likely. Therefore, it has been predicted that spatial isolation reduces symbiont richness (Kennedy and Bush, 1994; Simberloff and Moore, 1996). An interesting consequence of the metapopulation concept is that subdivided habitats connected by dispersal may contain species that would be unable to persist in a single patch (Hanski and Gilpin, 1997). In the context of metagenomics,
this may occur in cases where symbionts have a limited period during which they are associated with a certain host individual. This is typical for pathogens, which may be cleared by the host's immune system. While persistence within a host individual is limited, survival of the pathogen in the host population is possible, for example measles and rubella.

**Temporal Dynamics**

Recently founded host populations have been observed to have fewer parasite species than older populations, because in young host populations the parasite community is not yet saturated (Dobson and Pacala, 1992; Ebert et al., 2001). This is consistent with the observation that newly introduced host species have fewer parasite species than long-term residents (Guégan and Kennedy, 1993). The underlying mechanism for this relationship is that, during patch colonization, the host population goes through a bottleneck of a few individuals. Thus, only parasites and symbionts present in this small subset will be present in the new population. Over time, more symbionts will arrive by migration, either independently, or in association with immigrating hosts. An important consequence of the colonization time hypothesis is that host species that experience repeated bottlenecks over the course of range expansion will become progressively different from the original population. This will influence not only their own gene pool, but also their symbiont community (Moodley et al., 2009). This effect is known as "isolation by distance" and is an important tool to trace colonization and migration patterns of plants and animals, including humans (Linz et al., 2007). This method can be used both by tracing marker genes in selected symbionts, as well as by tracing the microbiome composition of a host species across wider geographic areas.

A different form of a temporal effect is related to host age. Most hosts are born with few or no symbionts (Breitbart et al., 2008). The diversity of the symbiont population increases over time, with increasing contact to the outside environment. This pattern is reflected in observations that older host individuals often have more parasites than younger individuals (Doigel, 1964). The expectation that symbiont richness increases with host age is also supported by metagenomics data (Kurokawa et al., 2007; Breitbart et al., 2008; Rajilic-Stojanovic et al., 2009), but only a few studies have analysed different host age classes. Based on predictions of island biogeographic theory, one would predict that a newborn host accumulates symbionts asymptotically - i.e. the initial increase in symbiont diversity will be rapid, but will slow to near-zero past a certain age. The rapid changes in human gut microbiota richness early in life and the comparatively stable richness later in life support this idea (Zoetendal et al., 1998; Kurokawa et al., 2007; Breitbart et al., 2008; Rajilic-Stojanovic et al., 2009). Some data suggest that the metagenome composition is dynamic over time (Rajilic-Stojanovic et al., 2009), but it is unclear if this dynamic is of qualitative or quantitative nature. Drawing again on phenomena observed in larger communities, many of the early arriving species (the colonizers) are later replaced by other species (Begon et al., 1990). Early species are often adapted for rapid growth and reproduction, but are poor competitors. In contrast late successional species are often competitively superior to the early colonizers, but grow slowly. Alternatively, species may simply accumulate over time without loss, resulting in a strictly monotonic increase in the richness curve. An interesting aspect of age related symbiont-host relationships is the role symbionts may play in host development (McFall-Ngai, 2002). From experiments with axenically raised hosts, it is known that development can be altered under the exclusion of symbionts (Marques et al., 2006). For
example, it is believed that gut symbionts are required for normal development of the human intestine (Neish, 2009). Currently, we have insufficient knowledge to predict how developmental processes will influence the richness of microbiota, and visa versa.

**HOST - SYMBIONT COEVOLUTION**

While some of the concepts outlined above treat present/absence patterns of symbionts as stochastic events (e.g. the theory of island biogeography), others emphasize finely coevolved associations between host and symbiont (e.g. vertical transmission). A recent report concluded that not only diet, but also phylogeny profoundly influences bacterial diversity in mammalian guts and that the intestinal communities codiversify with their hosts (Ley et al., 2008). In the following we will present a deeper look into the coevolutionary aspects of microbiotas and their hosts.

The coevolved relationships are of specific interest here because they may reveal some of the most intricate and important biological interactions. Coevolution among species has been invoked as an explanation for many biological phenomena. It has been suggested to occur in many situations, ranging from coevolving guilds (e.g. pollinators - plants), mutualism (e.g. aphids and their bacterial symbionts; ants - acacia), warning signals (e.g. mimicry), predators - prey associations (e.g. ants - ant lions) and host - parasites associations (e.g. human - malaria) (Darwin, 1859; Ehrlich and Raven, 1964; Futuyma and Slatkin, 1983; Thompson, 1994; Holland and Rice, 1999). In the last two decades, specific coevolution between hosts and symbionts has been a leading theme in evolutionary biology, parasitology, ecology, epidemiology, and lately also in applied fields, such as human and veterinary medicine, agriculture and biocontrol (Hunt et al., 2001; Visscher et al., 2002; Woolhouse et al., 2002; Levin and Bull, 2004; Mew et al., 2004; Thrall and Burdon, 2004). Coevolution between hosts and symbionts has been suggested to explain a diverse range of phenomena, such as genetic recombination (and sexual reproduction), sexual selection, autumn colors of trees, hypervariability loci, the extraordinary genetic diversity at genes related to immune function or resistance (e.g. MHC, R-genesis), spatial divergence and local adaptation, high rates of amino-acid replacements in resistance and virulence genes, rapid evolution at contingency loci, restriction enzymes in bacteria, multiple mating in social insects, polyploidy, RNA interference (RNAi), horizontal gene transfer, and host and symbiont specialization and speciation (Hamilton, 1980; Hill et al., 1991; Moxon et al., 1994; Ebert, 1998; Poulin, 1998; Baer and Schmid-Hempel, 1999; Bergelson et al., 2001; Thompson and Cunningham, 2002; Koella and Boete, 2003; Little et al., 2004; Fischer and Schmid-Hempel, 2005; Milinski et al., 2005; Ebert et al., 2007). Most of these examples are based on the interactions of hosts with one or, at most, a few symbionts, often pathogens or parasites. There is however no reason to assume that entire guilds of microbes associated with hosts would not produce equally interesting phenomena. Our growing awareness of symbiont diversity highlights the need to invest greater effort to understand coevolutionary dynamics between hosts and their rich symbiont communities.

Traditionally, inference of coevolution is based on functional analyses of the two interacting species in question. In cases of mutualism, one could demonstrate the reciprocal benefits of interaction, while in cases of antagonism (e.g. pathogens and parasites) one could
Dieter Ebert and Weihong Qi
demonstrate asymmetry in costs and benefits between partners. However, such tests usually require an experimental approach, which often practically difficult or impossible. Species rich microbiotas pose a particular challenge for such analysis. The number of pairwise interactions can be very high, including not only the interactions between the hosts and the symbionts, but also among the symbionts themselves. The detection of coevolution may become more complex when entire groups of symbionts participate, making a functional analysis even more difficult. Furthermore, coevolution may occur in cases where the function of the symbionts is not known. Thus, alternative methods are needed, which do not rely on the quantification of costs and benefits. Metagenomics is a tool which would allow such an alternative approach.

The term coevolution is used to explain many different processes of how two (or more) species influence each other's evolution. In the context of metagenomics only those types of coevolutionary processes can be addressed which show a clearly visible genetic signature in the species considered. This excludes rapid forms of coevolution, such as antagonistic coevolution by negative frequency dependent selection (so called Red Queen dynamics), where only allele frequency changes at few loci are involved (Ebert, 2008). Such Red Queen dynamics are unlikely to be detectable with a metagenomic tool box.

**Diffuse Coevolution**

The term coevolution is often used in a colloquial sense to emphasize that many of the phenotypes we observe today are somehow shaped by coadaptation. For example, the human immune system is often claimed to have coevolved with parasites. While it is clear that parasites played an important role in the evolution of the vertebrate immune system, it is also certain that no single parasite accounts for the complexity of the immune system. Rather, many different parasites acting at the same or different time over millions of years contributed to its evolution. However, it is quite possible that none of the parasite clades have a phylogeny congruent with the hosts. Likewise, the human gut microbiota is said to be the result of millenia of coevolution with symbiotic bacteria (Neish, 2009). Again, no single symbiont species may show a signal of co-speciation. What we may see in these cases are associations among clades of hosts and symbionts, reflecting a certain degree of specificity, which is however lower than in case of specific coevolution. This is called diffuse coevolution because guilds of coevolving species interact (Fox, 1988). Symbionts may relatively frequently (on geological time scales) switch hosts or do not specialize on particular host species. This results in associations of host and parasites clades, but without patterns of co-speciation. For example, the malaria agent *Plasmodium* is only found in reptiles, birds and mammals and in dipteran insects as vectors. Members of the genus *Plasmodium* do not infect amphibians or blood-sucking insects other than mosquitoes. The closely related genus *Haemoproteus*, however, infects reptiles and birds, but not mammals. Such clade specific patterns of associations can be studied by mapping symbiont clades on host trees. A closer look into examples of clade specific associations may reveal however, patterns of co-speciation.
Box 2. How to find out who they are?

Metagenomics studies not only identify many new genes and gene functions, they can also uncover large numbers of previously undiscovered taxa. Phylogenetic analyses of metagenomes may also allow us to reconstruct events of past host/microbiota interactions. To undertake any such analysis requires the proper taxonomic classification of the members of the microbiota.

Predicting the source organisms or taxonomic origins of sequence fragments in a metagenomic dataset is technically challenging. Various approaches have been developed, which can roughly be group into two categories. The first category of methods is based on sequence similarity, in which sequence fragments are compared to known sequences present in reference databases and taxonomic assignments are made based on homologs of known sequences. The second category of approaches is based on the compositional characteristics of sequences, in which sequence fragments are clustered based on compositional signatures (McHardy and Rigoutsos, 2007; Raes et al., 2007; Kunin et al., 2008).

The most traditional similarity-based approach is construction of phylogenies using 16S rDNA sequences, which is the gold standard marker for the phylogenetic taxonomy of microbes (Woese and Fox, 1977). The first datasets of enviromental samples relied almost exclusively on the analysis of 16S rDNA sequences (Hugenholtz and Pace, 1996; Pace, 1997). 16S rDNA has been sequenced extensively from type strains, cultivable organisms and metagenomic samples, based on which large databases of reference sequences have been developed, such as the Ribosomal Database Project (Maidak et al., 1994), the Greengenes (DeSantis et al., 2006b), and SILVA (Pruesse et al., 2007), each of which provides a variety of software tools for sequence classification, alignments, and taxon identification (Ludwig et al., 2004; DeSantis et al., 2006a; Wang et al., 2007; Cole et al., 2009). Alternatively, taxonomic classification of sequence fragments can be based on mapping to a selective set of conserved, phylogenetically informative marker genes (Venter et al., 2004), ubiquitous single-copy genes (von Mering et al., 2007) to all Pfam (Finn et al., 2006) domain and protein families (Krause et al., 2008). With this approach, the set of marker genes are first identified in a reference data set, such as all available complete genomes. A reference sequence alignment of each marker gene is prepared. Then the shotgun metagenomic data set of interest is mined for the marker gene homologs. When found, the sequences are also included in the corresponding reference alignments, based on which phylogenetic trees are constructed to identify the relative positions of maker genes from the metagenome. Taxonomic composition estimates based on the phylogenetic inference of marker genes have a major limitation. Metagenome shotgun datasets represent sequences independently sampled from random regions of genomes randomly selected from a given community. Thus marker genes derived from metagenomic data sets are often fragmented and produce incomplete sequence alignments, which can compromise the resulting phylogenetic trees. Huson et al (2007) developed a metagenome analysis tool, MEGAN, which makes use of a greater fraction of a metagenomic dataset for taxon identification and allows the exploration of the taxonomic contents of a community sample based on the NCBI taxonomy. MEGAN was first used to describe the metabiome of a mammoth sample (Poinar et al., 2006). With this approach, a given metagenomic dataset is
first compared to a broad collection of reference sequences, such as all known nucleotide sequences or all known protein sequences using sequence similarity tool BLAST. Each identified sequence homolog is then associated to the lowest common ancestor (LCA) of the set of taxa it hits. Thus, species specific sequences are assigned to low order taxa such as species or strains, while widely conserved sequences are assigned to high-order taxa. A similar tool, SOrt-ITEMS (Monzoorul Haque et al., 2009), was developed to improve accuracy and specificity of taxon assignment. It uses multiple alignment parameters, in addition to the bit-score value, to ensure the accuracy of the hit assignment. Then it identifies the subset of hits that share an orthologous relationship with the metagenomic sequence. The LCA of this subset of hits is assigned to the metagenomic sequence, which is more specific than the LCA of all the obtained hits.

Composition-based methods can be further divided into unsupervised and supervised approaches. Unsupervised composition-based methods do not depend on reference sequences. The taxonomic classifiers are trained using the same dataset being analyzed, thus novel metagenomic sequences lacking related known reference sequences can potentially be clustered by shared sequence composition features via unsupervised approaches. However the taxon assignment of the clustered sequences still relies on marker sequences of known origins. Supervised composition-based methods train taxonomic classifiers using reference data with known taxonomic origins to improve the taxon assignment accuracy. Table 1 lists the major composition-based methods. Unlike similarity-based methods, which can be applied to sequence reads as short as 100 bp, most composition-based methods are effective only for DNA fragments of 1 Kbp and longer, due to the limited number of words that are contained in short sequences. This means pre-processing of sequencing reads, such as assembly, are needed for most composition-based methods (Kunin et al., 2008). An exception is the most recently developed composition-based tool Phymm, which can accurately classify reads as short as 100 bp (Brady and Salzberg, 2009). The authors also presented a hybrid method PhymmBL, which incorporates the composition-based Phymm with the sequence similarity-based BLAST. It has been shown to outperform either of the individual methods (Brady and Salzberg, 2009).

Finally it should be noted, that besides the taxonomic classifier applied, many other factors can also affect the accuracy of taxon assignments, such as the complexity of the metagenomic sample, sequence length, pre-processing of the data, the presence of chimeric sequences, and the presence of horizontal gene transfer events (Mavromatis et al., 2007; Diaz et al., 2009). A few misclassified species in a large microbiota may not make a big difference, but in individual cases it is important to scrutinize the assignment of a species to a certain taxon to avoid mistakes.

Co-speciation

A powerful method to identify evidence of co-speciation is a co-phylogenetic approach (Page, 2003). This approach is based on the premise that coevolved symbionts of closely related hosts are expected to be closely related as well. In other words, the phylogenetic tree of a symbiont will match the tree of its host, showing a pattern of co-diversification (Hafner et al., 1994; Chen et al., 1999; Moran and Baumann, 2000). In contrast to diffuse coevolution, the pattern analysed to reveal co-speciation are based on very specific (often species specific)
host-symbiont interactions. For co-speciation to occur, it is necessary that the host-symbiont association has been stable since the last common ancestor of the hosts included in the dataset, as this stability is precisely what generates congruent host/symbiont phylogenies. However, it is not necessary that the last common ancestor be distantly related. The same approach can be taken with different populations of the same host species. In such cases, the last common ancestor may be within the same species. Figure 1 shows a schematic representation of co-speciation. The topology of the phylogenetic tree of the host (left side) resembles the topology of the symbiont tree, indicating that the symbiont clade underwent the same speciation events as the host clade. Associations between symbiont clades and host clades may be produced by ecological factors, such as common diet (Ley et al., 2008). It is unlikely that ecological factors alone would produce a pattern of coevolution at the species level, but care should be taken to consider such potentially confounding (but informative) factors. A refined co-phylogenetic approach would also estimate the timing of speciation events, with the expectation that symbionts should speciate at the same time or a bit later than their hosts. Co-phylogenetic comparisons can also identify important events in the coevolution of the species pair. For example, host species 6 in Figure 1 lacks a matching symbiont. Further investigation may elucidate the events that lead to this change. For instance, host 6 may have undergone a shift in habitat or diet. The comparison of real phylogenies rarely leads to scenarios as congruent as the hypothetical example in Fig. 1. However, even weak co-phylogenetic signals can be identified with statistical methods (Page, 2003).

**Box 3. Comparative Metagenomics**

Comparative metagenomics is a rapidly growing field. Most metagenomic data analysis tools are developed for the analysis of individual metagenomes. Fast and user-friendly tools with the capacity for rapid and highly interactive comparison of multiple datasets are increasingly in demand. Microbial communities were initially compared using environmental 16S rRNA clone sequences. Sequences are usually grouped into Operational Taxonomic Units (OTUs) if their 16S rRNA genes are above a certain sequence identity threshold (Hughes et al., 2001; Singleton et al., 2001; Martin, 2002; Schloss et al., 2004). The biodiversity between communities can then be compared using various ecological diversity indices and abundance models. Example software packages developed for this type of analysis include estimates (Colwell and Coddington, 1994), Mothur (Schloss et al., 2009) and the R package vegan (Oksanen et al., 2010). UniFrac is one of the first tools that compares microbial community diversity in a phylogenetic context. It computes differences between microbial communities by measuring the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both (Lozupone and Knight, 2005; Lozupone et al., 2006). The tool can be used to determine whether communities are significantly different, to compare many communities simultaneously using clustering and ordination techniques, and to measure the relative contributions of different factors, such as chemistry and geography, to similarities between samples.

MEGAN is a standalone tool that allows the comparative analysis of multiple datasets from metagenome shotgun sequencing (Huson et al., 2007; Huson et al., 2009) (Mitra et al.,
2009), for example (Qi et al., 2009). After taxonomic analysis of individual dataset with the LCA algorithm based on BLAST results, the software can display an arbitrary number of different datasets together on a subtree of the NCBI taxonomy. It also implements the Directed Homogeneity test for pairwise comparisons of two datasets, with Bonferroni and the Holm–Bonferroni multiple test corrections (Mitra et al., 2009). The metagenomics RAST server provides a web service for comparative metagenomics. It reconstructs the phylogeny of individual sample by using both the phylogenetic information contained in the SEED nr database and the similarities to the ribosomal RNA database. A taxonomic heat map is then produced to highlight the different taxonomic profiles in each sample (Meyer et al., 2008).

Table 1. Composition-based classifiers for metagenomic data

<table>
<thead>
<tr>
<th>Package name</th>
<th>Learning techniques</th>
<th>Features</th>
<th>Minimum sequence length</th>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>TETRA</td>
<td>Unsupervised, z-score correlations</td>
<td>Tetranucleotide frequencies</td>
<td>40 Kbp</td>
<td>Both web service and standalone tool; pre-computed tetranucleotide usage patterns for 166 prokaryote genomes</td>
</tr>
<tr>
<td>SOM</td>
<td>Unsupervised, self-organizing map</td>
<td>Tetranucleotide frequencies</td>
<td>5 Kbp</td>
<td>SOMs constructed using sequence fragments obtained from 1,502 prokaryotes; visualization of phylogenetic diversity on a single map.</td>
</tr>
<tr>
<td>S-GSOM</td>
<td>Semi-supervised, seeded growing self-organizing map</td>
<td>Tetranucleotide frequencies</td>
<td>8 Kbp</td>
<td>Flanking sequences of 16S rRNA as seeds; sorting according to phylogenetic relationships inferred from seeds; visual identification of relatively abundant of unknown species.</td>
</tr>
<tr>
<td>Phylopythia</td>
<td>Supervised, a hierarchical set of support vector machines</td>
<td>Over-represented oligonucleotide patterns</td>
<td>1 Kbp</td>
<td>Classifier trained on 340 completed genomes; can learn the relevant class-specific characteristics in the presence of considerable noise (such as global G+C content and thermophily at the domain level).</td>
</tr>
<tr>
<td>TACOA</td>
<td>Supervised, k-nearest neighbor approach with a smoother kernel function</td>
<td>Oligonucleotide frequencies</td>
<td>800 bp</td>
<td>Classifier trained on 373 complete genomes; the reference set can be easily updated without the need of retraining; standalone tool.</td>
</tr>
<tr>
<td>Phymm</td>
<td>Supervised, interpolated Markov</td>
<td>Variable-length oligonucleotide</td>
<td>100 bp</td>
<td>Classifier trained on 539 complete, curated genomes; use information from multiple</td>
</tr>
</tbody>
</table>
It is not possible to construct the sort of phylogenetic trees shown in Fig. 1 using metagenomic data from a single host species. To test for coevolutionary patterns, several datasets from hosts of known relatedness are needed. The first step is to classify all symbionts into clades, based on phylogenetic relatedness. The methods for doing this are rapidly improving and several analysis programs are available. Box 2 gives a summary of some of the methods used currently. Modern methods for taxon assignment do not require that phylogenetic marker genes are sequenced (although this helps), but random sequences from the symbiont genomes can be used to assign taxa with some degree of confidence (see Box 2 for a summary of methods). Using multiple metagenomic datasets and a phylogeny of hosts, one can select clades of symbionts and produce symbiont trees for comparison with the host tree. This approach can be cumbersome. It requires that many independent trees be built and statistical power can be low due to multiple comparisons. To our knowledge, there are currently no software packages available designed specifically for this type of analysis. There are however, various methods available to compare metagenome data sets and to find clades represented in more than one sample. Box 3 gives an overview of some methods useful for this approach.

Comparative genomics software can identify clades of symbionts that are shared among hosts. Figure 2 shows an example from three species of water fleas of the genus *Daphnia* (Qi et al., 2009) produced with the software MEGAN (Huson et al., 2007). The pie charts show the proportions of symbionts belonging to the different host species. The diameter of the chart scales with the total number of reads assigned to this clade. As can be seen in Fig. 2, bacteria in the clade Rhodobacteraceae are found in all three *Daphnia* species, while most of the other taxa show up only in one or two *Daphnia* species. Thus, symbionts within the...
Rhodobacteraceae are potential candidates for coevolution with *Daphnia*. It should be noted, that the other symbiont/*Daphnia* associations shown in Fig. 2 cannot be ruled out as examples for coevolution, but the signal picked up by the tree is not strong enough to suggest this. Once a symbiont clade of interest is identified, a more detailed analysis is necessary. The next step requires a higher resolution of the symbiont taxa, down to species or strain level. Unfortunately, most metagenomic datasets contain few reads which map to any particular clade and many of the microbes are new to science. Furthermore, most metagenome studies include too few host species to allow strong conclusions. These limitations can be circumvented by designing primers specific for the symbiont clades of interest, which can be used to amplify target genes from a larger number of different hosts (Fraune and Bosch, 2007). The co-phylogenetic approach outlined here can help elucidate coevolutionary patterns in the microbioms of closely related host species. For example, in the comparative metagenome study of three waterflea species (Qi et al., 2009) more than 50 % of all clades would classify as candidates for further studies of coevolution.

As stressed before, co-speciation is most likely in cases where hosts and symbionts form long lasting stable associations. The hallmark of such relationships are exclusively vertically transmitted symbionts. Vertically transmitted symbionts have the most intimate relationships with their hosts and each other and are therefore expected to show strong genetic signatures of co-speciation. Their genomes are co-transmitted with the host genome, providing an opportunity to develop long-term, stable associations. In the absence of horizontal transmission, vertically transmitted symbionts may become entirely dependent on their hosts. Long-term stability of such associations can result from both parasitic and mutualistic relationships. Harmful vertically transmitted symbionts (e.g. germ line parasites and cytoplasmic parasites) need to increase their representation in the next generation above the average representation of a nuclear host allele, as otherwise these symbionts would decline together with the host. The most common means of achieving this are sex ratio distortion and cytoplasmic incompatibility (Hurst, 1993). If a harmful vertically transmitted symbiont cannot increase its representation relative to the host, it must evolve avirulence (Fine, 1975; Ebert and Herre, 1996). Some symbionts may evolve to become mutualists, thereby entering into potentially extremely long-lasting, stable relationships. For example, the association between Aphids and their *Buchnera* symbionts is thought to be between 150 and 200 million years old (Moran and Baumann, 2000). Such long-term host-symbiont associations are likely to lead to a genetic signature of coevolution. This effect is so strong, that it is reasonable to assume that patterns of co-speciation revealed by genetic analysis indicate the presence of vertically transmitted symbionts (Chen et al., 1999; Moran and Baumann, 2000). This is however, not always be the case as indicated by the Sepiolid squid-*Vibrio* symbioses, where the squid lives in obligate symbiosis with a horizontally acquired light producing bacterium (Nishiguchi et al., 1998).

**WHAT IS A CORE METAGENOME?**

In asking the question "Who is there?" researchers observed large differences among the microbiotas of hosts of the same species. To distinguish accidental passengers from the more basic components of the microbe community, the concept of a core metagenome or core
microbiota is used. In its most simple definition it is the intersection of the communities from a group of comparable hosts (e.g. the gut microbiota of a sample from people from various location, who consume diverse diets, and have diverse genetics and lifestyles). By identifying a host specific core microbiota, the huge diversity of microbes is reduced to a smaller, more manageable set with a presumably more important role in the function of microbiota. A description of the core microbiota could also be used as a reference to help identify abnormalities in hosts. However, the definition of the core microbiota should be applied with caution, as it can be both conceptually and statistically misleading (Fierer et al., 2008; Turnbaugh and Gordon, 2009).

The first problem with the core microbiota is related to sampling intensity. The smaller the sample size, the lower the probability will be to detect rare components of the microbiota. Thus, at the intersection of two microbiotas, rare members will be less likely to be present. However, rare does not mean less important. The loss of rare symbionts from the core microbiota becomes stronger as more microbiotas are included in the analysis. Therefore, the size of a core microbiota will increase with sample size per individual host, but will decline with the number of hosts included. This problem leads to an unresolvable asymmetry. Inclusion of a species in the core microbiota because it is omnipresent cannot be a mistake, but the exclusion of a species because it occurs only in a subset of samples can be highly misleading.

A second problem with the concept of a core microbiota is that it may not exist. Currently we have no evidence for this. Ecological studies of plant and animal communities have shown that the same niche can be occupied by different species with similar functional characteristics. These species may even play key roles in the ecosystem function. If the same is true for microbiotas we should not expect a core microbiota. Indeed, with regard to functional genetic groups, we may find a core metabiome instead of a core microbiota (Turnbaugh and Gordon, 2009).

A third problem pertains to heterogeneity among host microbiotas. Biogeographic studies show that species are highly variable across the face of the planet, and there is little doubt that this is also true for the members of the microbiota. This geographic variation may be the result of adaptation to different local environments, genetic drift, or historic movement patterns of hosts and symbionts. Such variation would require that the core microbiota be independently defined for each different location or environments. While such differences are interesting in their own right, they work against the clear definition of the core microbiota.

A fourth challenge relates to the difficulty of defining the “normal” core microbiota. Among the earliest studies of the human metagenome were reports of differences among individuals with distinct physical characteristics, like obese versus lean (Turnbaugh et al., 2008). In cases where these conditions fall clearly into two or more categories, we can select the category we associate with the normal or healthy phenotype and define a core microbiota from there. However, host phenotypes are often continuously distributed, making it difficult to define the group of hosts and their associated symbionts, which will be used as a reference.

Finally, the core microbiota is only a statistical reality. It will always represent only a subset of what is found within a host. It is incorrect to assume that the core microbiota represents the microbial constituents for proper function of the host species. In fact, it may not even include the most important members of the microbial symbiont community.
Figure 2. Comparison of the symbiont community of three waterflea species of the genus *Daphnia*. Only the section of the Alphaproteobacteria are shown (Qi et al., 2009). The graph was produced with the software MEGAN (Huson et al., 2007). The pie charts show the representation of symbionts belonging to the three host species. The size of the chart scales with the total number of reads assigned to this clade. Bacteria in the taxon Rhodobacteraceae (arrow) are found in all three *Daphnia* species, while most of the other families show up only in one or two *Daphnia* species. The grey shades of each pie chart is as the following: dark grey for *D. pulex* sequences, pale grey for *D. pulicaria* sequences, light grey for *D. magna* sequences. Note that smaller pies include often less host species. This is because of stochastic effects of sampling.
Given these problems related with the core-microbiota concept it is not surprising that in one study it was reported that there was not a single microbe with more than 0.5% abundance shared by all gut microbiotas in samples from 154 humans (Turnbaugh et al., 2009). Does this invalidate the concept? We believe that despite all problems, the core microbiota is a useful idea, as it provides structure for complex data sets. This structure can help to generate hypotheses and to identify interesting patterns, but only if the aforementioned hazards are taken into consideration. Care has to be taken that the factors mentioned above can produce spurious patterns and misguide interpretations and further research. Some of the obvious problems can be easily avoided and may then lead to stronger interpretations and conclusion, yet, not all of them can be circumvented. We cannot avoid the problems, but we need to be aware of them.

CONCLUSION

The perspective offered in this chapter can only be realized in a comparative context. By comparing microbiotas of different hosts and of the same host at different points in space and time, we can answer questions apart from the mostly functionally oriented current approaches to metagenomics. A comparative approach will deepen our insight into the intricate dynamics of complex communities and illuminate our co-evolutionary past. It will shed light on the incredible biodiversity inside us and inside every other organism on this planet and provide a structural framework for these complex communities based on the ecological and evolutionary forces that shape them.

In this it will also allow us to ask what features limit symbiont richness. For instance, why does the human microbiota include more than 1,000 species, but that of a waterflea (Daphnia) only about 100? The enormous diversity of microbiotas makes it a challenging task to understand the factors that drive and/or limit the diversity of these communities. A comparative approach is able to shed light on these questions and can help us to structure the diversity seen in microbiotas. Evolutionary theory is the overarching concept in all of biology and a powerful tool in understanding biological complexity.

ACKNOWLEDGEMENTS

We express our thanks to Jason Andras who spent many hours editing the style and English of this chapter.

REFERENCES


An Evolutionary Ecology Perspective on Comparative Metagenomics


