Seeing the forest for the genes: using metagenomics to infer the aggregated traits of microbial communities

Noah Fierer1,2 *, Albert Barberán1 and Daniel C. Laughlin3

1 Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA
2 Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA
3 Environmental Research Institute, School of Science, University of Waikato, Hamilton, New Zealand

Most environments harbor large numbers of microbial taxa with ecologies that remain poorly described and characterizing the functional capabilities of whole communities remains a key challenge in microbial ecology. Shotgun metagenomic analyses are increasingly recognized as a powerful tool to understand community-level attributes. However, much of this data is under-utilized due, in part, to a lack of conceptual strategies for linking the metagenomic data to the most relevant community-level characteristics. Microbial ecologists could benefit by borrowing the concept of community-aggregated traits (CATs) from plant ecologists to glean more insight from the ever-increasing amount of metagenomic data being generated. CATs can be used to quantify the mean and variance of functional traits found in a given community. A CAT-based strategy will often yield far more useful information for predicting the functional attributes of diverse microbial communities and changes in those attributes than the more commonly used analytical strategies. A more careful consideration of what CATs to measure and how they can be quantified from metagenomic data, will help build a more integrated understanding of complex microbial communities.

Keywords: metagenomics, traits, community-aggregated traits, microbial diversity, microbial ecology

WHY MICROBIAL ECOLOGISTS ARE STILL LOST IN ‘TERRA INCOGNITA’

Ever since van Leeuwenhoek first peered through a microscope, it has been recognized that most environments harbor diverse and complex microbial communities. Whether we are studying soil, the human gut, marine sediments, or lake waters, we find many taxa that have unknown, or at least poorly described, ecological characteristics. Our descriptions of microbial communities are littered with question marks just as maps of continents were once littered with labels of ‘terra incognita’ by European explorers.

Recent methodological advances, most prominently advances in DNA sequencing, have provided unique insight into the structure and function of complex microbial communities, thereby improving our ability to chart ‘terra incognita.’ This is particularly true for those environments, like soils and sediments, which harbor many taxa that are resistant to laboratory isolation (Whitman et al., 1998; Fierer and Lennon, 2011). It is now routine for researchers to use shotgun metagenomics, randomly sequencing from a pool of whole-community DNA extracted from environmental samples (Handelsman, 2004), to characterize complex microbial communities and their functional capabilities. Shotgun metagenomics is by no means the only approach available – but, fueled by rapid declines in sequencing costs, shotgun metagenomes are an increasingly available source of data that can be mined to characterize microbial communities.

One of the explicit goals (and promises) of many shotgun metagenomic projects is to understand the broader functional and ecological characteristics of microbial communities, insight that cannot necessarily be gleaned from the sequencing of taxonomic or phylogenetic marker genes. These arguments are convincing because there is already evidence that such approaches can be used to better understand the function of complex microbial communities, whether those functional capabilities are related to phosphate removal from wastewater (Martin et al., 2006), carbon cycling in permafrost soils (Mackelprang et al., 2011), or arsenic geochemistry in marine sediments (Plewniak et al., 2013). While these and many other studies highlight the utility of shotgun metagenomic analyses, it is often difficult to distinguish tangible advances from the hype.

We can use metagenomic data to develop testable predictions regarding the ecological attributes of microbial communities, but the approach is no panacea and inferring the functional capabilities of communities from metagenomic data remains difficult. One reason for this is that community-level attributes are the emergent properties of a diverse array of organisms interacting directly and indirectly in a myriad of ways under environmental conditions that are rarely static. Even if we could predict the ecological attributes of all individual taxa living in a given community (a Sisyphean task in most microbial habitats), the overall functional capabilities of that community and how it responds to changes in biotic or abiotic conditions, will remain difficult to predict. If we want to know how rapidly soil communities will decompose soil organic matter or how efficiently a gut microbial community will ferment ingested polysaccharides, it is insufficient to document the genes associated with the metabolism of various organic carbon pools and their relative abundances. Even in simple communities composed of well-described microbial
taxa, predicting community-level metabolic properties from genomic or transcriptomic data can be challenging (Sieuwerts et al., 2010).

If the goal is to link microbial communities to processes, a critical step is to understand the ecological attributes of whole communities, not just the attributes of individual community members as the whole is rarely the sum of its parts. Here we argue that we can often do a better job of predicting the functional capabilities of whole communities by using the concept of community-aggregated traits (CATs) to glean more useful information from the terabases of shotgun metagenomic data being generated.

**WHAT ARE COMMUNITY-AGGREGATED TRAITS AND WHY ARE THEY USEFUL?**

Functional traits can be measured at various levels of organization – from the level of individual cells, to species, to whole communities. For example, one could determine the size of individual bacterial cells in a water sample, the mean cell sizes for different bacterial species found in that sample, or the mean size of cells found in the whole sample. Community-level traits can be quantified either as a “community-weighted mean trait,” where the mean trait values for all taxa in a community are weighted by their relative abundances, or as a “CAT,” where the traits are measured from a random sample of individuals irrespective of their taxonomic identities.

We can calculate community-weighted mean traits from shotgun metagenomic data by reconstructing genomes, or parts of genomes, and using this genomic information to predict the characteristics of individual community members. This ‘bottom up’ approach has already been demonstrated to be useful for describing the putative functions of undescribed microbial taxa (e.g., Walsh et al., 2009; Hug et al., 2013) or for documenting taxa likely responsible for specific microbial processes (e.g., Howard et al., 2008). Likewise, we can calculate community-weighted mean traits from *a priori* information on the traits of specific microbial taxa (or lineages). While this approach is certainly appealing (e.g., Langille et al., 2013), it has some clear disadvantages if the traits of many microbial taxa, even closely related taxa, are not already known, as would be the case in many habitat types. Likewise, for those traits that are capable of being transferred across distantly related taxa via horizontal gene transfer (e.g., antibiotic resistance; Forsberg et al., 2012), trying to determine community-level traits solely from taxonomic or phylogenetic information would be problematic.

An alternative approach is to calculate CATs directly from a community of interest without collecting any information on the identities of the taxa found in a given plot. For example, leaf traits in a plant community can be determined by remote sensing of canopy spectra (Homolová et al., 2013) or through random taxon-free sampling (Gaucherand and Lavorel, 2007) without having to measure trait averages for each plant species. In a similar manner, we can calculate CATs from shotgun metagenomic data as long as we assume that our metagenomes represent a random sampling of all microbial genomes present in that sample and that the traits of interest can actually be inferred from the genomic information. Such a ‘top–down’ approach has been widely used in microbial environments ranging from soil (Fierer et al., 2012) to marine waters (Ganesh et al., 2014) to the human gut (Greenblum et al., 2012). In all of these cases, much of the insight into community attributes came not from assigning genes to taxa, but rather by determining the relative abundances or presence/absence of genes and gene categories at the community-level of inquiry, not at the level of individual taxa.

CATs are often useful predictors of community-level properties because, according to the mass ratio hypothesis, species controls on community-level processes are in proportion to their relative abundances (Grime, 1998). The functional traits of abundant taxa will have more important influences on the functional properties of a community than the traits of subordinate taxa. There is strong empirical support for the mass ratio hypothesis from those studies that have explicitly tested the hypothesis in plant communities (reviewed in Lavorel and Grigulis, 2012). For example, Mokany et al. (2008) demonstrated that mean trait values were a good predictor of multiple ecosystem properties in grasslands, including litter decomposition rates and aboveground net primary productivity. Likewise, other studies have demonstrated the utility of using plant CATs to predict a wide range of processes and ecosystem properties from nitrification rates (Laughlin, 2011) to soil carbon dynamics (Garnier et al., 2004), and green biomass production (Lavorel et al., 2011).

In many cases, the mass ratio hypothesis should also apply to microbial communities. For example, we might expect that the efficiency by which microbes mineralize nitrogen in a composting bio-reactor would be a function of the community mean for those traits associated with nitrogen mineralization rates. The appeal of the mass ratio hypothesis to ecologists working with highly diverse microbial communities is that we do not necessarily need to know the traits of all taxa to predict a process of interest, rather we just need to know the traits of the more dominant taxa (Grime, 1998). Of course, not all community-level processes will be predictable from the mass ratio hypothesis. For example, relatively rare ‘keystone’ taxa can have a disproportionate influence on certain processes (Paine, 1995) and biotic interactions could invalidate the mass ratio hypothesis (Vile et al., 2006). Moreover, the diversity of traits (the range or variance in trait values found in a given community) could often be more relevant (Violette et al., 2012) than just the mean CAT value (Figure 1). For example, under fluctuating environmental conditions, like those found in wet soils exposed to wide swings in O2 levels, the distribution of traits associated with O2 tolerance could be more informative for predicting community responses over time than simply measuring a mean community-level O2 preference (Pett-Ridge and Firestone, 2005). In theory, we should be able to use shotgun metagenomic data to calculate both means and variances of traits (Figure 1), based on the distribution of the relevant genes or gene categories indicative of the trait of interest within the community.

CATs have already proven useful for predicting some microbial processes. For example, soil microbial ecologists have long considered bacterial:fungal ratios to be a useful metric for understanding soil carbon dynamics because bacteria and fungi are considered to have distinct carbon use efficiencies (Bailey et al., 2002), an assumption that has been called into question...
one approach to simplify this complexity. For example, microbial taxa have been divided into groups based on shared life history characteristics (Fierer et al., 2007; Ho et al., 2013), groups defined by the source of carbon or energy (lithotrophs, heterotrophs, autotrophs), or divided into groups based on specific functional capabilities (N\textsubscript{2}-fixers, photosynthesizers, methylotrophs). While pragmatic, these functional categories over-simplify trait-level variability and it is likely more appropriate to define microbial trait space as a set of continuous, quantifiable variables, where microbes, or microbial communities, sit within this multidimensional space.

There are 100s of traits that could possibly be measured for any organism – but it is often not necessary to measure every one of these traits to place species into multidimensional trait space (Laughlin, 2014). The key is to identify and measure those traits that are most relevant to the system in question and select those traits that are independent and that most effectively discriminate between taxa or communities. Such an approach has proven useful in plant ecology where a handful of plant traits can be used to accurately place species into multidimensional trait space (Laughlin et al., 2010; de Bello et al., 2013) as many other plant traits are often correlated with this subset of traits. There are clearly similar trade-offs in microbial traits (Gudelj et al., 2010). For example, there are well-established trade-offs between growth rate and yield (Pfeiffer et al., 2001), between stress tolerance and the ability to compete for substrates (Ferenci and Spira, 2007), and between cell size and nutrient uptake rates (Young, 2006). Although the list of possible traits is enormous, there are inescapable morphological, physiological, and genetic constraints that narrow the list of possible trait combinations associated with different life history strategies.

Selecting CATs to study can often be done a priori. For example, we can assume that traits that confer tolerance to changes in water activity will likely be important if we are trying to predict soil C dynamics in arid or semi-arid systems (Lennon et al., 2012; Evans and Wallenstein, 2014). Likewise, we could assume that traits associated with nutrient uptake will have important controls on phytoplankton growth in many freshwater systems (Edwards et al., 2013). Sometimes the traits that might be relevant to predicting the community-level function are unknown or unexpected because the process itself or specific controls on the process are not well understood. For example, it would be difficult to predict a priori which CATs would be most relevant to understanding controls on nitrification by archaea given that physiologies remain poorly understood (You et al., 2009). In these cases, one could determine which CATs are most relevant to a process of interest by first measuring the community-level process across space and/or time and then empirically determining what CAT, or set of CATs, appears to be correlated with the measured changes in the process. In this manner, one could generate specific, testable predictions about how changes in CATs relate to changes in community-level processes and test them experimentally.

**INFERRING COMMUNITY-AGGREGATED TRAITS FROM METAGENOMIC DATA**

Many microbial traits could be inferred from genomic or metagenomic data (Table 1) and there are many other examples of traits...
Table 1 | Selected examples of microbial traits and the genes or genomic characteristics that could be used to calculate community-aggregated trait (CAT) values from shotgun metagenomic data.

<table>
<thead>
<tr>
<th>Microbial trait</th>
<th>Selected genes, gene categories, or genomic characteristics that could be used to infer the trait value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum growth rate</td>
<td>rRNA operon number, codon usage bias in highly expressed genes, rRNA/tRNA position</td>
<td>Lauro et al. (2009), Vieira-Silva and Rocha (2010)</td>
</tr>
<tr>
<td>Dormancy</td>
<td>Sporulation proteins, toxin-antitoxin systems, resuscitation-promoting factors</td>
<td>Lennon and Jones (2011)</td>
</tr>
<tr>
<td>Osmoregulation</td>
<td>Trehalose and peptidoglycan production</td>
<td>Culligan et al. (2012)</td>
</tr>
<tr>
<td>Ability to catabolize recalcitrant organic compounds</td>
<td>Genome size, secondary metabolite transport/metabolism</td>
<td>Konstantinidis and Tiedje (2004), Fierer et al. (2012)</td>
</tr>
<tr>
<td>Stress resistance (general)</td>
<td>Sigma factor subunits of RNA polymerases (e.g., σB)</td>
<td>Kazmierczak et al. (2003), Marles-Wright and Lewis (2007)</td>
</tr>
<tr>
<td>Cold tolerance</td>
<td>Cold shock proteins, trehalose synthesis proteins</td>
<td>Varin et al. (2012)</td>
</tr>
<tr>
<td>Motility</td>
<td>Chemoreceptor, flagellar genes</td>
<td>Girgis et al. (2007)</td>
</tr>
<tr>
<td>Oxidative stress tolerance</td>
<td>Catalase, peroxidase, and polyketide synthase genes</td>
<td>Qin et al. (2012), Zhang et al. (2013)</td>
</tr>
<tr>
<td>Nitrogen/phosphorus affinities</td>
<td>Genes for membrane-bound nutrient uptake/transporters</td>
<td>Hewson et al. (2009), Eloé et al. (2011)</td>
</tr>
<tr>
<td>Resistance to toxic metals</td>
<td>COGs associated with heavy metal efflux pumps</td>
<td>Hemme et al. (2010), Eloé et al. (2011)</td>
</tr>
<tr>
<td>Antibiotic resistance</td>
<td>Genes for efflux pumps, ribosomal protection, enzymatic inactivators</td>
<td>Hu et al. (2013), Forsberg et al. (2014)</td>
</tr>
</tbody>
</table>

References include examples of studies where that trait was inferred from either genomic or metagenomic data.
We thank the graduate students at the University of Colorado who participated in the “Genomes and Traits” seminar for useful discussions on some of the topics discussed here. Support was provided by a grant from the National Science Foundation (DEB0953331) to Noah Fierer, a James S. McDonnell (JSMF) Postdoctoral Fellowship to Albert Baran, and a grant (UOW1201) from the Royal Society of New Zealand Marsden Fund awarded to Daniel C. Laughlin.

REFERENCES


Fierer et al.