Integrating Evolutionary, Ecological and Statistical Approaches to Metagenomics

A proposal to the Gordon and Betty Moore Foundation

Jonathan A. Eisen
University of California, Davis
U. C. Davis Genome Center
Section on Evolution and Ecology
Department of Medical Microbiology and Immunology

Katherine S. Pollard
University of California, Davis
U. C. Davis Genome Center
Department of Statistics

Jessica L. Green
University of Oregon
Center for Ecology and Evolutionary Biology
Department of Biology

INTRODUCTION

Metagenomics[1] – the study of the genomes of many microbes in an environment simultaneously - has the potential to revolutionize our understanding of the hidden yet incredibly important world of microorganisms. This potential has been highlighted by a series of recent metagenomic-based studies [1-8] as well as multiple government reports [9] including in particular the recent National Academy of Sciences report "*The New Science of Metagenomics – Revealing the Secrets of our Microbial Planet*."

The great potential of metagenomics comes with enormous challenges in the analysis of the data[2]. These challenges include the fragmentary nature of sequence data, the sparse sampling of genomes, populations and communities and the unknown phylogenetic diversity and ecological structure of the communities being sampled [7]. Methods designed for analysis of single organism genomes simply do not work well on data sets sampled from complex ecological communities. To develop new methods, the NAS report suggested (and we agree) that integrated approaches involving interdisciplinary teams of researchers are needed in which the researchers both ask scientific questions and develop new data analysis tools.

Here we propose building exactly such an integrated, interdisciplinary effort, bringing together three labs with different relevant areas of expertise (Table 1) including statistics (to deal with the sparse sampling), comparative genomics (because the data is genomic in nature), evolutionary biology (to assess phylogenetic and genomic diversity), and ecological theory (to examine community structure).

Table 1. Principal investigators

PI	Areas of expertise
Jonathan Eisen	Evolutionary and comparative genomics, metagenomics, phylogenetics
Katherine Pollard	Statistical and computational genomics
Jessica Green	Applied and theoretical ecology, microbial community structure

The research we propose covers three major topics considered of fundamental importance in metagenomic studies[3]: biodiversity, evolutionary dynamics, and statistical measures (see Table 2). Our proposed work should lead to novel insights into microbial ecology and evolution. In addition, at the core of all of our work is the development of novel mathematical, statistical and computational methods for analyzing metagenomic data. Since these methods will be of use to the research community at large, we propose to work closely with the CAMERA team to the methods broadly available through the CAMERA database.

Table 2. Proposed research areas

1. Metagenomic based characterization of microbial biodiversity
   a. Guidelines and weightings for using different gene families in metagenomic based diversity assays
   b. Searching for novel phylogenetic types in metagneomic data
   c. Metagenomic analysis of community phylogenetic structure
   d. Estimating biodiversity from metagenomic samples
2. Metagenomic studies of microbial evolutionary dynamics
   a. Molecular evolutionary dynamics of gene families
   b. Population genomics
3. Statistical metagenomics: Correlation analysis of sequence data and metadata

[1] We use the term metagenomics to refer to shotgun sequencing DNA from environmental samples

[2] The NAS report identifies five challenges in metagenomics: need for interdisciplinary teams, role of government, methods development, complexities of data analysis and need for databases

[3] In the NAS report they identify four key questions: how can we find new functions, how diverse is life, how do microbes evolve and what role do microbes play in the health of their hosts

Though each of these projects can be considered separate activities, they are highly interdependent and the interdisciplinary nature of the labs involved is critical for the success of the project. For example, we propose to use phylogenetic analysis to search for new types of organisms and genes (project 1b led by Eisen). The results of this phylogenetic analysis will also be used to assess the phylogenetic structure of communities (project 1c led by Green) and to study evolution of gene families across environments (project 2b led by Pollard). Similarly, the statistical methods developed for comparative metagenomics (project 3 led by Pollard) will be used in the population genomic studies (project 2a led by Eisen) as well as in the development of biodiversity estimators (project 1d led by Green). To achieve this integration we plan to have active feedback between each of the PIs and each of the projects. To do this we propose a management plan that will help guide these interactions.

We believe that by taking this integrated approach – both in terms of the research topics and by combining separate fields of study, we will not only make important scientific discoveries about microbial communities but we will also build and develop novel methods and approaches of great utility to the metagenomics community.

1. METAGENOMIC BASED CHARACTERIZATION OF MICROBIAL BIODIVERSITY

Introduction and Background to the Problem

Two key questions come up over and over again in discussions of the role of microbes in the world. "*Who is out there?*" and "*What are they doing?*" Metagenomics has been justifiably praised for providing a revolution in the ability of researchers to answer the second question. This revolution is possible because one can take all of the data obtained from metagenomic sequencing and use similar functional prediction methods to those that have been developed for studying the genomes of single organisms to predict the functions present in a community. Though there are challenges with such functional predictions for communities, the insight provided has been and will continue to be enormous.

rRNA PCR revolution

The focus on "*what are they doing*" types of questions is in part because many consider the "*who is out there*?" issue to be largely solved through the use of PCR amplification of rRNA genes (rRNA-PCR) [10]. Using rRNA-PCR (and various downstream analyses) researchers have gotten a handle on many aspects of the "who is out there" for microbes. The most critical part of rRNA-PCR is that in a single reaction one can amplify rRNA genes from a wide diversity of organisms present in a mixed sample. And then one can use various types of analyses of the resulting PCR products to study the organisms that were present in the sample. In the following paragraphs we discuss how these analyses have helped address different aspects of the "who is out there question."

One key component of determining "who is out there" is the determination of what types of organisms are present. This task was nearly impossible for microbes until the advent of rRNA-PCR. With rRNA PCR, each PCR product can be assigned to a microbial phylogenetic group by sequencing it and then building a phylogenetic tree of this and other rRNA sequences and looking at where the PCR product sits in the tree [10, 11]. It is this type of phylogenetic classification analysis that led to the discovery of dozens of novel major subdivisions of cellular organisms, as well as hundreds of novel branches within particular groups [12-16].

Another use of analysis of rRNA-PCR products is in the estimation of the relative abundance of different organisms in a sample. One way of doing this is to sequence many PCR products from one sample, to then divide them into clusters of closely related sequences, and then to count the numbers of clones in each cluster. The clusters, also known as *phylotypes* or *operational taxonomic units (OTUs)* are commonly used as surrogates of species for studies of microbial ecology (although we note that in section 2a we discuss research on whether this is a valid assumption). When coupled to the phylogenetic classification analysis described above, one can estimate the relative abundance of different taxonomic groups (e.g., compare the abundance of cyanobacteria to spirochetes).

Another important component of "who is out there" is information on the total number of species present in a sample, something known as *richness*. Unfortunately, it is untenable to examine all the microbes present in any particular sample, even with rRNA-PCR. However, it is possible to *estimate* richness from analysis of rRNA-PCR products. For example, the Chao method estimates diversity by analyzing information on the fraction of OTUs that are represented by a single clone versus those with two clones [17].

Metagenomics can circumvent many limitations of rRNA-PCR

The use of rRNA-PCR to help answer questions about who is out there in microbial communities using methods such as those outlined in the previous section revolutionized environmental microbiology. And rRNA-PCR still plays a fundamentally important role in such studies. However, there are some limitations to rRNA-PCR that have led to gaps and even mistakes in our understanding of microbes in the environment. Most importantly for our purposes here, metagenomics circumvents most of these limitations. To understand the benefits of metagenomics and how it can complement rRNA PCR we discuss three examples below.

One severe limitation of rRNA-PCR is that it does not allow the sampling of viruses. Though viruses are not cellular organisms and thus not considered to be truly living, they play diverse and important roles in many ecosystems. In theory, one can study viral diversity using PCR of protein-coding genes. However this is not very robust for two reasons: protein-coding PCR is challenging to get to work well across large evolutionary distances and there are no genes shared by all viruses. Metagenomics surveys of viral genomes have provided many novel insights into viral biodiversity (see Rowher papers) and as more viral populations are surveyed it is likely that more insights await.

Another limitation of rRNA-PCR is that not all rRNA genes amplify equally well and some do not amplify at all. Thus analysis of the products of rRNA PCR can provide inaccurate information about what rRNA containing organisms were present in the sample. Since metagenomics does not require an amplification step, it can avoid the biases of PCR (although we note that this does not mean that metagenomic surveys have no biases). The potential power of metagenomics is best illustrated by comparing the rRNA genes seen in metagenomic data with those seen in rRNA-PCR from the same samples. For example, not only does one see quantitative differences in estimates of relative abundance of organisms, analysis of metagenomic data has identified whole lineages missed by PCR [18] .

A third limitation of rRNA-PCR is that rRNA is not the ideal gene family for all studies of biodiversity. For example, rRNA copy number is highly variable between species which makes estimates of relative abundance potentially biased [19]. In addition, trees based on rRNA do not always recapitulate the phylogeny of the rest of the genome [20]. For example, in analysis of the genome of *Hyphomonas neptunium* we found not only that the phylogenetic trees of this organism's 16s rRNA was anomalous compared to other genes in the genome, but that this had led to a major misclassification of the organism (it was placed in the wrong Order) [21]. The reasons for this and many other rRNA-based misclassifications are sometimes known and include lateral gene transfer, convergent evolution, and unequal rates of evolution. For cultured organisms, especially those for which genomes are available, it has been routine for many years to use other genes, especially protein coding genes, as a component of phylogenetic studies [22]. Unfortunately, extending this approach to uncultured organisms has been impossible because protein-coding genes are even more prone than rRNA is to the biases and limitations of PCR discussed in the previous section. Fortunately, metagenomics allows one to use genes other than rRNA for biodiversity studies of cellular organisms providing an important comparison for rRNA-based results. Interestingly, in the first analysis of this kind, we (Eisen) found that rRNA versus protein based estimates of various aspects of microbial diversity were somewhat similar but had both qualitative and quantitative differences [23].

Using metagenomic data has many challenges

Though the potential of metagenomics to avoid limitations of rRNA-PCR is clear, realizing this potential in full is challenging. One of the challenges is that different gene families are not equally useful

for characterizing the multiple dimensions of microbial diversity. For example, some gene families may provide very accurate measures of the types of organisms present (i.e., phylogenetic classification*)* but may not provide accurate estimates of the relative abundance of those types. We believe this is the case for rRNA due to copy number variation between taxa.

A second challenge is that the utility of a gene family for a task may be taxon dependent. For example, we have shown that the *recA* gene is a useful marker for phylogenetic classification of microbes [22, 24] and have used it in various analyses of metagenomic data [2, 23]. However, *recA* is missing from some cellular organisms, especially certain clades of endosymbionts. Thus it cannot be used to characterize these groups in any way much like rRNA cannot be used for viruses. The reverse is also true – there are many protein families which are only found in one or a few clades of microbes. These protein families may be quite useful in diversity studies of these clades but of no use in studies of any other group.

A third challenge in making use of metagenomic data for diversity assays lies in how to compare and combine results for different gene families. For example, if one wanted to compare and contrast species richness estimates using the 16s rRNA, RpoB and RecA gene families it would be helpful to normalize the definition of OTUs in these families. Since rRNA is more highly conserved at the DNA level than either RpoB or RecA, a 99% identity cutoff for defining rRNA OTUs (a commonly used setting) would be equivalent to a lower DNA identity cutoff for RpoB and RecA. But how much lower?

A final challenge is dealing with the fragmentary nature of metagenomic data. For example, imagine if one recovered 10 rRNA sequences from a sample: two covering the whole molecule, five covering the left two thirds and three covering the right two thirds. To determine if they are from the same or different OTU it might be necessary to analyze only the regions they have in common (the middle third). But if the middle third has a different average level of conservation than the whole molecule, using a 99% identity cutoff to define OTUs would mean that OTUs in this study would be different than those in a separate study of whole length molecules. Is it possible to somehow normalize the percent identity cutoff values to better compare and contrast the results from different genes?

Challenges in metagenomic analysis are not insurmountable

The challenges we raise above, as well as other challenges, is in large part what led the NAS group to recommend the development of interdisciplinary teams to work on metagenomic studies. For the purposes of addressing "who is out there?" types of question we believe that the challenges of metagenomics are solvable through such interdisciplinary work. In the following sections we propose four related projects to develop, make use of, and make available to others, novel methods for using metagenomic data to study biodiversity of microbes.

a. Guidelines and weightings for using different gene families in metagenomic based diversity assays (Eisen)

Many of the challenges in using metagenomic data for diversity studies outlined above relate to how to make use of data from different gene families or from different parts of the same gene family. We believe that these challenges can be largely ameliorated by using comparisons of complete or nearly complete genomes to develop guidelines and weightings for how different gene families can be used in metagenomic studies. For example, suppose for a particular gene family, each and every bacterial genome had identical duplicated genes in this family, while each and every archaeal genome had only a single gene in the family. If one wanted to use this gene family to calculate relative abundance of different OTUs in a metagenomic sample, one would have to divide the estimate in half for all bacterial OTUs. Similar types of weighting guidelines could be developed for other metagenomic based diversity assays (discussed in more detail below).

The reason to use complete or nearly complete genomes for this analysis is that they provide a wealth of information about genes, gene families and how genomes evolve. We recognize that complete genomes do not necessarily provide a perfect picture of evolution in all organisms. However there is good evidence that for many of the patterns in which we are interested (e.g., gene copy number,

phylogenetic history of genes), currently available genomes are good approximations for what might be seen in other species related to those from which the genomes come. There is one significant limitation of the current genomes in that they do not cover the phylogenetic diversity of species very well. However, Dr. Eisen is finishing one project (an NSF Funded Tree of Life Project) and coordinating another that is just beginning (a Genomic Encyclopedia Pilot Project at JGI) which will greatly increase the phylogenetic coverage of genomes.

It is important to point out that in addition to developing these methods, we will use the results of metagenomic analyses to provide some ground truth to the weightings developed. For example, if the weighting schemes are robust, then one should find that using the weighting schemes will lead to different genes providing similar results from within a metagenomic community. A schematic summarizing this is shown in Figure 1 and more detail is given in the following text. Our plan for developing the guidelines and weighting schemes is outlined below.

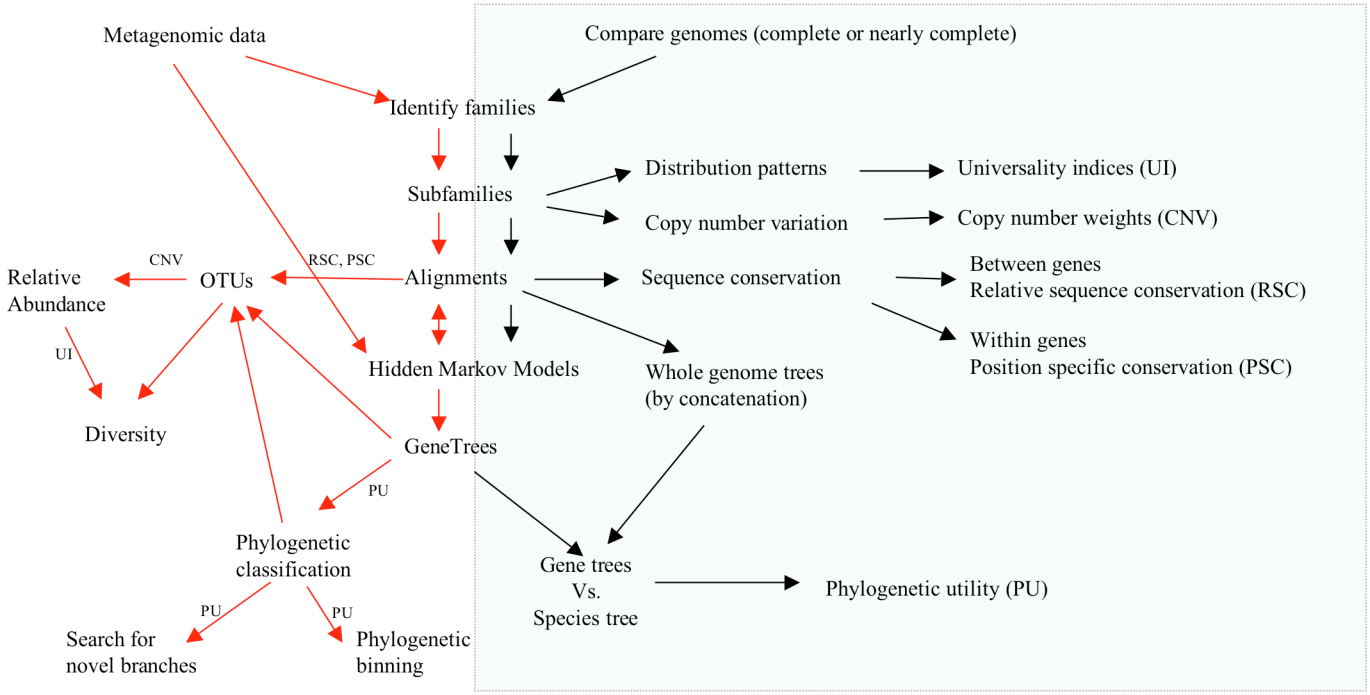

Figure 1. Development and use of weighting scores for metagenomic based measures of phylogenetic diversity. On the right of the figure, represented by black arrows (and in a shaded box), are the proposed analyses of complete genome sequences. On the left are the proposed analyses of metagenomic data (arrows in red). The goal of genome analysis is to develop weighting schemes, scores, and bioinformatic resources to use to aid in the anlaysis of metagenomic data. First, genes in these genomes will be divided into gene families and subfamilies (initially this will be done only for a subset of gene families). Then alignments, hidden markov models (HMMs) of the alignments and gene trees based on the alignments will be made for each subfamily. Analyses of these subfamilies, alignments,HMMs and trees will produce weighting scores for carrying out metagenomic analyses of diversity. From the distribution patterns of subfamilies we will calculate universality indices (UI) for families that wil then be used to aid in converting relative abundance measures into diversity indices. From measures of the copy number of gene families per genome we will calculate copy number variation (CNV) weights which will be used to convert counts of different operational taxonomic units (OTUs) into relative abundance estimates. From analysis of the alignments we will calculate sequence conservation indices. RSC will measure differences in conservation between gene families, which will be used t normalize the identification of OTUs between genes. PSC will measure variation in conservation within geneswhich will be used to handle fragmented sequence data in identifying OTUs. From comparison of trees of each gene family with species trees we will calculate phylogenetic utility (PU) measures for each gene family which will then be used to place confidence levels on phylogenetic assignments. We note that in addition, HMMs generated from gene family alignments will be used to aid in alignments of metagenomic data and can serve as a starting point for analyses.

General Approach

Our approach to developing the weighting schemes for different gene families is as follows. First, we will build a database of complete genomes including both cellular organisms and viruses (this will be a variation on a database already created by the Eisen lab called ComboDB). Then for each gene

family of interest we will (a) identify and extract the members of the gene family from all the genomes in the database, (b) generate a multiple sequence alignment for all members of each family, and (c) build phylogenetic trees for each family. Then these results will be analyzed and compared to generate the weightings and guidelines for using these gene families for three diversity related tasks: phylogenetic classification, phylotyping, and estimates of relative abundance. The specific methods for generating these weightings are discussed below.

Phylogenetic classification utility

Not all gene families are equally useful for phylogenetic classification. Some gene families are prone to lateral transfer. Others are difficult to align or too short for robust analysis. We can measure the phylogenetic utility of a gene family by comparing trees of the gene family to what is known as a "genome tree." Genome trees are used as a surrogate for a species tree. If a gene family is not useful (whatever the cause) its gene tree will be substantially different than the species tree. We propose two measures of gene family phylogenetic utility that will be calculated by comparing the gene family tree to the species trees. For the species tree we will create a tree of all organisms in our comparison using concatenated alignments of housekeeping genes (for cellular organisms) and the phage proteomic tree method for the viral genomes. Each gene family will be assigned a *global phylogenetic utility* (GPU) score that reflects how similar the gene tree is to the species tree. In addition, *taxon specific phylogenetic utility* (TPU) scores will be assigned for the gene family based on the similarity of the gene tree and species tree just for that subset of taxa.

These scores will be tested using metagenomic simulations to see whether they can be used to improve the accuracy of phylogenetic typing. In addition, we will test the PU scores by analyzing real metagenomic data using the following logic. For a single metagenomic sample, measurements of the types of organisms that are present should be the same for each gene family that have been given a high GPU score. That is, good phylogenetic markers should all give the same answer regarding who is present. If the markers considered to be high quality always give different answers for all metagenomic data sets, that would indicate that our GPU scores are not robust. Similarly, the TPU scores can be tested by examining how marker genes perform for particular taxonomic groups.

Phylotype/OTU weighting: sequence conservation and sequence fragments

As discussed above, if one wants to use different genes to simultaneously identify phylotypes in a metagenomic data set, one needs a way of normalizing the definition of phylotype between the different families. We propose to create two such normalization scores for each gene family by comparing the sequence alignment of the family to that of 16s rRNA genes from the same species. First, we will calculate a gene family specific *relative sequence conservation* (RSC) score by comparing the overall sequence conservation of the gene family to sequence conservation of rRNA genes. This will allow the use of equivalent percent similarity cutoffs for each gene when identifying phylotpyes.

Another challenge in phylotyping is the fragmentary nature of the sequence data which makes it difficult to even detect phylotypes within a single gene family. We propose two methods to better handle fragmentary data for phylotyping. First, for similarity-based determination of phylotypes we will create *position specific conservation* (PSC) scores for each gene family. Thus for the fragment that only covers a highly conserved region of a gene (discussed above) a higher percent identity will be required to place it into a phylotype with other fragments. Second, we will develop phylogenetic approaches, such as the use of supertrees, which will allow non overlapping fragments or fragments from different regions of a gene to be placed on the same gene tree. Then phylotypes can be determined from the tree itself and not by using a percent cutoff. As with the phylogenetic typing scores, these scores will be tested empirically using metagenomic simulations and real metagneomic data.

Relative abundance weighting

Even if one assumes that one can perfectly identify phylotypes and normalize estimates between different gene families, it is still not straightforward to use the number of occurrences of sequences from a particular phylotype to estimate the relative abundance of that phylotypes in the sample. In particular,

three critical factors influence the relationship of the occurrence frequency to relative abundance . First and perhaps most importantly, the copy number of particular genes varies between taxa. To account for this we will calculate both *global* and *taxon specific copy number variation scores* (GCNV and TCNV) for each gene family in the genomes analyzed. These scores will be tested empirically using metagenomic simulations.

A second factor in calculating relative abundance is that the size of a gene affects the ability to detect it robustly in metagenomic data (e.g., very short genes are hard to detect with certainty even when present). We will assess this effect using metagenomic simulations and develop weighting schemes to account for it.

A third factor is that not all genes show up with equal probability in metagenomic sequence data even when present in equal amounts. This variation is mostly due to differences in clonability when using clone-based sequencing. We will assess this using simulations that utilize real sequencing data as raw material for the simulations.

Finally, to estimate richness (the total number of phylotypes in a sample) one needs some information on the universality of a particular gene. Genes found only in particular phylogenetic groups can obviously only be used to estimate richness for that group. To aid in such richness estimates we will calculate for each gene family both a *global universality index* (GUI) and in a *taxon specific universality index* (TUI). One use of these indices would be for combining richness estimates from multiple genes. For example, if one found a set of universal genes for bacteria, and a separate set for eukaryotes and archaea, together they could be used to estimate total richness of cellular organisms.

A two tiered approach

We propose a two-tiered approach to build gene family rules for metagenomic analyses. The first analyses will be done with a set of 50-100 candidate gene families selected to broadly cover bacteria, archaea, eukaryotes and viruses. This will include gene families considered to be good phylogenetic markers (including each ribosomal RNA gene and multiple protein coding genes), a set of gene families of functional interest to microbial ecology researchers (e.g., proteorhodopsin) and a set of randomly selected genes (as a control). In the second round we will expand to at least 1000 gene families selected from analyses of complete genomes.

Deliverables

1. Generation of alignments and trees for targeted set of 50-100 gene families. Delivery of trees to projects 1b, 1c, 1d, 2a and 2b (Year 1).
2. Generation of scores for targeted set of 50-100 gene families (outlined above) (Year 1).
3. Testing the utility of these scores using simulated and real metagenomic data sets (Year 1).
4. Generation of alignments, trees, and scores for expanded set of ~500 families (Year 2).
5. Testing the utility of these scores using simulated and real metagenomic data sets (Year 2).
6. Integration of scores with development of diversity assays (see below) (Years 2-3).
7. Creation and updating of a database of the scores for different genes. Integration of database with CAMERA. (Years 1-3).

b. Searching for novel phylogenetic types in metagenomic data (Eisen)

Metagenomic sequencing has the potential to open up a new window into microbial phylogenetic diversity by revealing rRNA sequences that did not amplify with PCR and allowing the use of other genes to search for novel branches on the tree of life. We propose here to build automated systems to carry out such searches and to apply them to various gene families and multiple metagenomic data sets. The gene families we will focus on will be those analyzed in the previous section and will include both traditional phylogenetic marker genes (e.g., rRNA, RecA) and functionally important genes (e.g., proteorhodopsin). This will allow both for the discovery of novel types of organisms as well as novel subfamilies of functionally important gene families. The analysis of functionally important genes will also serve as a good point of comparison for the phylogenetic marker genes.

Approach

Previously we have used phylogenetic analysis to search for novel metagenomics-only branches for a variety of gene families. For example, we showed in analysis of the Sargasso Sea metagenomic data that many novel subfamilies of proteorhodopsin were present [23]. In addition, in the GOS data we have found many novel, very deep branches in trees of phylogenetically informative gene families (Eisen et al., in preparation). We propose to further develop these methods, make them more automated, and to use them to carry out a more systematic search for novel phylogenetic types in metagenomic data.

For each gene we will perform phylogenetic analysis of all available homologs (including those in Genbank, completed genomes and metagenomic data sets). We will then identify deep evolutionary lineages in these trees that are unique to metagenomic samples. When done with small subunit rRNA genes, this will allow the identification of organisms that have been missed by rRNA PCR surveys. When novel lineages are found in analyses of genes other than rRNA the interpretation will be a bit trickier. Suppose one built trees for all available RpoB genes including those from all metagenomic samples. And suppose that one found a novel deep branch in the RpoB tree which contained only sequences from metagenomic projects. While such a novel branch might indicate the existence of a previously unseen group of organism, it might also be caused by other factors including (i) the presence of a group that is known but that has not had its RpoB gene sequenced or (ii) the presence of paralogous RpoB genes or (iii) the occurrence of some type of phylogenetic or sequencing artifact. One way to help determine which of these possibilities is correct would be to see if the environmental samples that contained the novel *rpoB* genes also had novel forms of other gene families. A truly novel group of organisms should yield novel branches for many gene families whereas artifacts may occur only for a small number of genes.

We will also develop an alternative approach for characterizing the novelty of genes from metagenomic data by adapting the methods used in the Unifrac program [25]. This method analyzes phylogenetic trees of gene families by quantifying the amount of the total branch length in the tree that is unique to a particular subset of the data. This is commonly used to compare samples from different environments. However it could also be used to compare metagenomic versus genomic data sets or two metagenomic sets to each other. This could be used to give scores for which gene families have the most "uniqueness" in the metagenomic data compared to genomic data and could thus serve as a rapid screening tool for analyzing large numbers of gene families.

Once novel genes or subfamilies are found with whatever approach, we will then characterize the organisms from which these come in as much detail as possible, such as by characterizing other genes from the same sequence reads and/or metagenomic bins. In addition, we will also look for whether there are any types of environments (e.g., high salt, low temperature) are enriched for phylogenetic novelty.

Deliverables

1. Development of an automated system for searching for novel branches using rRNA sequences in metagenomic data (Year 1). Integration of methods into CAMERA (Year 2). We note that Eisen is already working with CAMERA to incorporate rRNA phylogenetic typing tools.
2. Development of an automated system for searching for novel branches for protein coding genes (Years 1-2). Integration of automated methods into CAMERA (Years 2-3).
3. Identification of novel branches for each gene family. New analyses will be carried out each year with new metagenomic data sets (Years 1-3).
4. Identification of other genes linked (either directly or through bins) to genes identified as novel (Years 2-3).
5. Development of a Unifrac-like system for identifying the amount of novelty in metagenomic data for a particular gene family (Years 2-3).
6. Comparison of the novelty of phylogenetic marker gene families versus functionally important families (Years 2-3).
7. Correlation analysis of gene family novelty with metagenomic sample metadata to determine if any types of environments are enriched for novelty (Year 3).

c. Metagenomic analysis of community phylogenetic structure (Green)

The increased availability of phylogenetic information has prompted development of 'community phylogenetics', a rapidly emerging field in ecology [26]. The aim of community phylogenetics is to analyze the degree of relatedness and diversity of species within local assemblages (local communities of potentially interacting individuals sampled from the larger region-scale metacommunity), and to understand the forces structuring those assemblages. For example, if a local assemblage is characterized by *phylogenetic overdispersion* of co-occurring species (i.e., species co-occur with other species that are more distantly related than expected by chance) this is consistent with the hypothesis that competition among ecologically similar species helped structure the local community. If a local assemblage is characterized by *phylogenetic clustering* of co-occurring species (i.e., species co-occur with other species that are more closely related than expected by chance), this is consistent with the hypothesis that environmental filters – filters that restrict community membership to species with certain conserved ecological traits - helped structure the local community. Empirical studies of phylogenetic overdispersion and clustering of ecological communities across a variety of habitats and taxonomic groups are on the rise, and emergent patterns from these studies continue to inform the evolving community phylogenetics framework [27, 28].

A limiting factor in the field of community phylogenetics is that many ecologists are deterred by the molecular techniques and phylogenetic methods required to study phylogenetic relationships. Similarly, many molecular biologists are unaware of the ecological questions that can be addressed using the phylogenies they produce [26]. Here we propose to bridge this gap by analyzing metagenomic data in a community phylogenetic context. Metagenomics offers a unique opportunity to examine community patterns in biodiversity data using a large suite of marker genes, including potentially "ecologically relevant" protein-coding genes [29]. As discussed below, this polyphasic approach has the potential to: (a) advance current thinking of community assembly in ecology, and (b) inform metagenomic-based studies on the utility of different gene families for building ecologically relevant phylogenetic trees (see section 1a above). Challenges will include addressing the ubiquitous issue of undersampling in biodiversity assays (see section 1d below).

Approach

The methodology for exploring community phylogenetic structure is well developed in the ecology literature [30] and is being continually refined by the Green lab [31] and others. The basic idea is to compare the phylogenetic dispersion of local communities to random species assemblages drawn from a more broadly defined species pool. Using established randomization methods, and the phylogenies derived from the research proposed above, we will quantify the phylogenetic structure of environmental metagenomic samples. We will begin our analyses with phylogenetic trees constructed using rRNA sequences. Fundamental questions to address include: Are the phylogenies of local assemblages over-dispersed, random or clustered? Does the phylogenetic structure of these assemblages differ among habitats? Does it differ with phylogenetic scale (e.g. considering all identified bacteria per sample versus betaproteobacteria only)? Answers to these basic questions have yet to be explored with metagenomic data, and emergent patterns from the resulting analyses will lay the foundation for future research aimed at understanding the mechanisms that drive ecological community assembly (e.g. see [32, 33]).

After answering the questions above using rRNA sequence-based trees, we will repeat all analyses using trees constructed with protein-coding genes that are commonly used as phylogenetic markers (e.g. ribosomal proteins, RecA, etc.), but not known to be major drivers of ecosystem function. We will then repeat all analyses using ecologically relevant functional genes (e.g. *pmoA* and *amoA*, which code for enzymes central to methane and ammonia oxidation, respectively). It has been hypothesized that phylogenetic overdispersion will be more prevalent in such data when the phylotypes sampled are performing the same function and require the same environmental resources. The degree to which this hypothesis holds may yield insight into the ecological similarity of taxa within protein families. For example, one might expect only highly unrelated members of the proteorhodopsin gene family to co-

occur in the photic zone to minimize niche overlap. A lack of overdispersion in assemblages of particular gene families could also suggest that those genes are too conserved to reveal ecological differences or that sequence variation at that phylogenetic scale does not reflect ecological differences [29].

This latter point is relevant to our broader goal, which is to incorporate information about phylogenetic community structure when assessing the utility of gene families for characterizing microbial biodiversity. Using an approach similar to that described for quantifying the usefulness of gene families for phylotyping and quantifying relative abundance and richness within samples (see 1a above), we will examine gene family utility in the context of understanding phylogenetic community structure. We will begin with a comparative analysis of community structure in metagenomic samples using phylogenies built with rRNA genes and protein coding genes.

Deliverables

1. Metagenomic community phylogenetic analysis of rRNA genes (Year 1).
2. Metagenomic community phylogenetic analysis of all 50 gene families discussed in Section 1a above (Year 2).
3. Assessment of utility of gene families from community phylogenetic perspective (Years 1-3).
4. Software development for metagenomics community in collaboration with CAMERA (Year 3).

d. Estimating biodiversity from metagenomic samples (Green)

A formidable challenge in the study of microbial diversity is that of undersampling. The extraordinary abundance of microorganisms makes the task of exhaustively sampling a full community even within a single environmental sample impossible. For this reason, microbial biodiversity studies rely on statistical estimators of diversity [34, 35]. The most commonly used estimators of diversity were first derived by Anne Chao and colleagues in the ecology literature [17, 36, 37]. These nonparametric estimators, which focus on species richness within local assemblages, originate from the mark-recapture models of mobile animals. Despite their tractability, their applicability to metagenomic data is unknown (although we note they are used extensively on such data). In addition, the Chao estimators lack a framework for predicting patterns of relative abundance within samples.

A second approach that has gained popularity in microbial ecology uses parametric distributions to estimate the number of species within local assemblages (e.g. [23, 38]). Sample data are fit to models of relative abundance [39, 40], and the sample frequency distribution is projected to estimate the number of unobserved species in the community. This approach is grounded in the following assumptions: 1) the sample relative abundance distribution is a truncated version of the community-level distribution (a common assumption in ecology and microbial ecology is that the community from which a sample is drawn has a lognormal relative abundance distribution), and 2) individuals are randomly sampled from the community. In many ecological contexts, however, these assumptions may be violated. There is little empirical foundation for assuming the functional form of microbial relative abundance distributions, and in metagenomic data variability between organisms in copy number of genes (or in the number of copies of the genome per cell) may result in a heterogeneous (or non-random) sample of individuals. In other words, even if gene fragments are randomly sampled from environment, this may not correspond to a random sample of organisms from the environment.

We propose to evaluate these currently available diversity estimation techniques and to develop new estimators for the study of metagenomic biodiversity patterns. Congruent with the patterns explored in other aspects of this proposal, our efforts will focus on estimating phylotype richness, relative abundance, and phylogenetic structure within environmental samples.

Approach

We will first evaluate the Chao richness estimators by sampling from *in silico* communities where diversity is known. We will quantify how sampling effort (the proportion of the community sampled) and community structure (relative abundance and endemism) influence the accuracy and precision of each estimator. Next, in collaboration with Anne Chao and Yi-Huei Jiang at the University of Taiwan, we will

derive novel diversity statistics aimed at estimating microbial diversity from metagenomic data sets. These new indices will vary from the standard Chao estimators to better suit metagenomic data. For example, we will revisit the assumptions underpinning the mark-recapture model framework and consider the implications of: (a) hyperdiverse communities and large population sizes (orders of magnitude larger than those of macroorganisms), (b) short generation times (which may violate the assumption that organisms are not reproducing during the sampling period), (c) lateral gene transfer (which may violate the assumption that samples are drawn from a 'closed' community), and (d) variability between organisms in copy number of genes (or in the number of copies of the genome per cell). In instances where assumptions are violated, we will analytically and computationally examine the potential implications for estimating diversity (i.e. does this result in under- or over-estimating diversity?). Finally, using information garnered from the analyses above, we will attempt to modify the existing mark-recapture modeling framework (or develop new models) for estimating diversity with metagenomic data. As mentioned below, models will be tested and validated by sampling from in silico communities.

Estimates of relative abundance will draw upon the mathematical framework outlined in [41]. Green and Plotkin (2007) used statistical analysis of finite mixture models to derive the relationship between the abundance distribution in a local community and the distribution observed in a small sample from the community. The next step is to invert the problem and extrapolate sample relative abundance data (i.e. metagenomic data) to estimate larger-scale relative abundance patterns (i.e. at the scale of an environmental sample). We will expand our analyses of finite mixture models to include maximum likelihood estimates of abundance distribution parameters from sample data, by using the well established expectation-maximum (EM) algorithm [42]. Estimates of relative abundance can then be paired with traditional parametric approaches to estimate phylotype richness in metagenomic samples.

The most challenging aspect of the proposed research will involve estimating the structure of community phylogenies from sample data, which entail estimating not just phylotype richness and abundance, but also phylogenetic topology. Understanding how sample phylogenies relate to the community phylogeny from which they were drawn is essential for making inferences about the formation and maintenance of biodiversity using the community phylogentics framework described above. It is likely that metagenomic samples will reflect the phylogenetic structure of predominantly abundant organisms within communities; factoring in the presence of rare or moderately abundant phylotypes may significantly impact conclusions drawn from such analyses. As a preliminary approach we will consider the sampling properties of phylogenies generated under null hypotheses [43] such as a constant rate of cladogenesis [44], stochastic birth and death processes [45], or neutral assembly processes [46] -assuming poisson (or random) sampling of individuals from the tips of the phylogenetic tree. Next we will consider more complex/heterogeneous evolutionary models (e.g. [47]) and non-random sampling of individuals from the phylogeny (which may more accurately reflect the metagenomics framework). The sampling properties of phylogenies will be explored through traditional modes such as "species to genus ratios" [48] and also using topology metrics drawn from network theory.

Results from all aspects of the work proposed above will be tested and validated by sampling from in silico communities where biodiversity is known and also empirical data from completely censused ecology field sites such as the Smithsonian Institute's Center for Tropical Forest Study plots (http://www.ctfs.si.edu/doc/plots/).

Deliverables

1. Assessment of the fidelity of currently used species (or phylotype) richness estimators. (Year 1)
2. Development of novel biodiversity estimators geared towards metagenomics. (Years 2-3)
3. Publicly available software parallel to Estimate S for the metagenomics community developed in collaboration with CAMERA (http://viceroy.eeb.uconn.edu/EstimateS). (Year 3)

2. MICROBIAL POPULATION GENOMICS AND GENOME EVOLUTION

By providing random samples of the organisms present in a particular community, metagenomics has potential to reveal novel insights into the evolutionary dynamics of microbes in nature. Such

evolutionary dynamics must be understood if we are to make predictions about the response of organisms to environmental change and if we are to better understand and model microbial communities. We propose two research projects in this area: studies of genomic variation within populations and studies of the connection between patterns of genome evolution and environmental/ecological properties of communities. Both projects extend the studies of biodiversity proposed above.

a. Molecular evolution of gene families (Pollard)

Because of the vast and growing number of new genes, species, and environments represented by metagenomic data (e.g., [1, 2]), these sequences are a fertile ground for studies of genome evolution. Genome sequence analysis has revealed that cultured microbes differ significantly in their evolutionary properties including rates of mutation and efficiency of directional selection. These differences significantly affect the evolvability of organisms, such as their ability to genetically respond to environmental change. Ecological niche can influence this evolvability – with the best example being intracellular microbes having high mutation rates and low levels of recombination. Metagenomics data also allows one to look for parallel or convergent evolutionary events in multiple taxa at the same time. When events are seen in multiple separate lineages this is strong evidence for some environmental effect rather than a historical artifact. We are particularly interested in identifying rapidly evolving genes and gene families, extending approaches we have used to study mammalian evolution [49, 50].

In order to investigate these questions, we need a reasonable approach to computing molecular evolutionary parameters from metagenomic data. These parameters include mutation rates and patterns, substitution rates and patterns, and selection. The challenge lies in developing and/or adapting metrics of genetic divergence that are appropriate for metagenomic samples.

Approach

Most molecular evolutionary parameters are not straightforward to estimate from a single metagenomic sequence read or short assembled contig that lacks the concept of a genome. The main barrier for studies of metagenomic genome evolution is the lack of clear definitions of organism and species, and hence the absence of species phylogenies. Calculation of substitution rates (synonymous and non-synonymous), for example, typically requires estimation of an ancestral state based on a phylogenetic tree and the sequences of the extant species.

To address this problem, we propose an approach that does not rely on the concept of a species. We will quantify substitution rates and patterns using the phylogenies of related proteins from a collection of well-sampled protein families [51]. These will be the same protein families developed above (*Project 1a*) and utilized throughout this proposal. Specifically, we will make use of the multiple sequence alignments and phylogenies generated for each gene family in that project. While this approach will not easily distinguish paralogs from orthologs, it will enable estimation of models for sequence evolution. These models will allow us to estimate substitution rates on different branches of the protein tree. This approach will allow us to study genome evolution, for example, by comparing synonymous to non-synonymous rates within each protein family. To assess the performance of this approach, we will use metagenomic simulations where ancestral proteins are allowed to evolve under various models and a sample of extant proteins is generated in a way that reflects potential biases in metagenomic samples. Studying the reliability and repeatability of our estimation on these simulated data sets will allow us to benchmark the methods and calibrate our expectations for their performance on real data.

Next, we will apply these methods to sequence data from various metagenomic studies. These empirical studies will enable us to evaluate global patterns of genome evolution. We will investigate variation in these patterns both within and across protein families. The results of these studies will provide much needed information about similarities and differences in patterns of evolution between cultured and uncultured organisms.

Finally, specific cases of very rapid evolution will be investigated in an effort to identify examples of directional selection. A goal of this analysis will be to identify the genetic basis for various "keystone" traits often associated with species radiations - traits that allowed the ancestors of organisms

alive today to colonize new environments and utilize new resources. Interpretation of these associations will be aided by characterizing sequence data based on gene function using publicly available ontologies [52]. This project will benefit from the statistical approaches to correlation analysis developed in *Correlation analysis of sequence data and metadata.*

The molecular evolutionary methods we develop will be implemented in publicly available software similar to PHAST (http://compgen.bscb.cornell.edu/~acs/software.html). These methods will be integrated into CAMERA, facilitating their application to additional protein families and samples.

Deliverables

1. Development and evaluation of molecular evolutionary methods for metagenomics (Years 1-2)
2. Assessment of global evolutionary patterns and trends. (Year 2)
3. Detailed investigation of particular cases of very rapid genome evolution. (Year 3)
4. Publicly available software similar to PHAST for the metagenomics community (in collaboration with CAMERA). (Year 3)

b. Population genomics (Eisen)

Genome sequencing of cultured isolates has revealed that the genomic variation among closely related microbial types is enormous. This it true even among what are considered to be different strains of the same species [53]. Factors that account for intraspecific genomic variation include lateral gene transfer, gene duplication and deletion, and high rates of mutation and evolution. Whatever the cause, the extensive genomic variation among closely related microbes has been one of the key driving forces behind the development of metagenomics. This is because the variation means that assigning an organism to a phylotype (such as by using rRNA-PCR) is not always useful in predicting the biology of that organism.

Although genomic variation among close relatives can complicate studies based on phylotypes, it can also be used to learn a great deal about the ecology, evolution, and general biology of particular organisms. The Eisen lab has helped pioneer such "phylogenomic" analyses for cultured isolates and for uncultured symbionts. Here we propose to carry out similar types of studies on uncultured organisms using metagenomic data. Since metagenomic data sets provide for the first time random samples of microbial communities, they have enormous potential to be used to infer details of processes that shape populations. In particular we propose to use metagenomic data to ask questions relating to the population genetics of uncultured microbes. We will focus on issues relating to gene flow, recombination, and mutation and also to how to define phylotypes and/or species with metagenomic data.

Approach

There is a significant challenge in this work in that most current population genetic models and methods analyze data where the source organism is known. This is important since frequently it is necessary to analyze multiple genes from different parts of one genome to make measures of important population genetic parameters such as recombination rates and linkage disequilibrium. Thus alternative approaches may be needed in metagenomic based studies. One alternative approach would be to generate sequence assemblies from metagenomic data and to then use those assemblies much as one would use contigs or chromosomes from genomes of cultured organisms. However this is fraught with difficulties since the assemblies may actually represent chimeras where different parts of the assmelby from different populations or even species. Therefore, we prefer to carry out our analyses on the sequence reads directly as has been done recently by the Banfield lab [3] and by other groups. In particular, we propose to use the following general approach: sequence reads from metagenomic data sets will be aligned to reference genomes (a process referred to as tiling) using relatively relaxed alignment stringencies. We will identify reference genomes for which there is deep coverage of these tiled metagenomic sequence reads in data sets from individual samples. These deep coverage tiled read data sets will then be used to estimate population genetic parameters such as mutation rate, recombination rate, and gene flow. For example, we propose to create a metagenomics F_{ST}-like statistic to measure within vs. between population genetic

variation which in turn can be used to estimate gene flow. This will allow the detection and identification of patterns of gene flow and the detection of genetic boundaries between populations. For example, the recently proposed pangenome concept [54] implies that gene flow occurs at nearly an infinite rate in microbial communities. We will be able to measure gene flow using the metagenomics F_{ST}-like statistic to test predictions of the pangenome hypothesis.

In addition we propose to develop methods to use phylogenetic trees to determine whether the sequence reads that have been tiled to the reference genome come from a single population or more. For example, if the tiled reads all came from a single recombining population, one would expect to see significant differences in the phylogenetic tree structure for different genes much as one sees for different genes within the human population. In contrast, if the tiled reads came from multiple non-recombining populations (or different species) then one would expect that there would be consistent structures in the trees of different genes (much like is seen when one compares genes from humans, chimp and gorilla). Of course with ancestral polymorphisms or lateral gene transfer the pattern will not be this simple. But we believe that with some simulations and modeling we can come up with probabilistic models for how to infer some aspects of population structure from trees of metagenomic data that is tiled to reference genomes.

It is from a similar approach to the phylogenetic analysis outlined above that the Banfield lab has begun to measure recombination rates and patterns in uncultured organisms [3]. We propose to develop new models that will build upon their work and that will enable us to make estimates of mutation rates, insertions or deletions, and rearrangements. Once such parameters are estimated then one can ask questions such as: "Do uncultured organisms follow the same rules as cultured organisms?" and "How much does the environment influence population genetic parameters (e.g., do deep sea microbes have different recombination rates than surface water organisms)?"

Deliverables

1. Development of F_{ST} like measures of genomic variation within communities versus between communities (Year 1).
2. Development of methods to quantify insertion-deletions, recombination, and rearrangement by comparison to reference genomes (Year 1).
3. Development of the genomic x spatial species concept for microbes. Geographic differential is critical in species concepts in animals and plants – and we will see whether it can be applied to microbes. Analysis of the pangenome concept (Year 2).
4. Development of phylogenetic sliding window approach for determining whether multiple populations are present in a single bin (Year 2).
5. Estimates of effective population (N_e) sizes for different microbes and design of methods to detect community-level bottlenecks that may make communities vulnerable as seen in endangered species. N_e is a critical parameter for population genetics and yet has been very difficult to estimate for microbes. We will use the approach of Lynch and Conery [55, 56] to do this for multiple microbes at once (Years 2-3)
6. Correlation analysis will be done on general patterns of "evolvability" such as mutation rates, population size, recombination patterns with community characteristics (Year 3).

3. STATISTICAL METAGENOMICS: CORRELATION ANALYSIS OF SEQUENCE DATA AND METADATA (POLLARD)

The full scientific benefit of metagenomic projects can only be realized if sequence data is carefully linked to other types of data. For example, metagenomic data allows a thorough assessment of the effect of ecosystem characteristics (e.g. physical parameters and levels of diversity) on evolutionary properties (e.g. mutation rates and effectiveness of natural selection). Identifying correlations between genomic variables and various metadata is at the core of many of the scientific questions in the metagenomics research community, including those posed in this proposal. Our goal is to develop models that allow us to quantify the magnitude and statistical significance of these correlations.

Approach

Methodology for association modeling and statistical testing is well developed in the statistical literature. The challenges for metagenomics will be to appropriately encode data and to account for variable sampling scenarios. Correlation analysis relies heavily on choices about data handling. Metagenomic data can be analyzed at the level of phylotype/bin, functional category, or sequence/allele. In addition, each type of data can be encoded as binary (present/absent), categorical, or quantitative (score or frequency, *e.g.* from a BLAST search). The optimal choice will depend on the application and quality of available data. Reducing data to presence/absence calls is useful when there is not good information about quantitative levels for at least some variables, but it can be less powerful than quantitative approaches when such data is available.

The appropriate statistical model for measuring correlation depends on how each variable (metagenomic or metadata) is encoded. For example, presence/absence of a protein family can be modeled as a function of a continuous environmental variable (*e.g.* temperature) through generalized linear models, such as logistic regression. When both variables are categorical, loglinear models may be appropriate. Most of these models allow for weights, which we will use to adjust for the uneven sampling that we see in metagenomic data, including uneven taxa sampling and possibly non-random missing data (*e.g.* metadata only measured on a subset of samples). These methods also allow us to include multiple variables into a single model, producing conditional estimates of association that are adjusted for the effects of other variables. A key feature of this statistical modeling approach is that estimates of correlation are accompanied by estimates of variability (due to sampling depth, sequence quality, sequencing method, etc.) that lead naturally to tests of statistical association (p-values). The performance of the proposed methods will be assessed through in silico simulations.

This correlation methodology will allow us to build upon several of the above projects including *1c* and *2a*. We are interested in studying the interrelationships between genome evolution, community structure, and the ecosystem. Gene families with evolutionarily unique rates and patterns of substitution will be compared across metagenomic samples from different ecosystems. Correlation analysis will be used to identify associations between sequence data and meta data, including demographic variables (e.g. population size), characteristics of the microbial community (e.g. species diversity, competitors, symbiotic relationships), and physico-chemical environmental variables (e.g. temperature, salinity, pH). We will address questions such as: Do certain gene families or functions evolve more quickly in particular ecosystems or with particular community structures? Which biological processes are most stable? Which are most environmentally sensitive? Do some environmental variables foster rapid evolution more than others?

It will be possible to address some of these questions with currently available data sets, such as GOS and vertical column data from the Delong lab. Others (e.g. genomic variation along environmental clines) await the collection of appropriate data. The methods we develop will be available to the research community when such data becomes available. Based on the results of these applications, we will adapt and extend the correlation methods as needed.

Deliverables

1. Survey of metagenomic data to determine the scope of data types (Year 1).
2. Methods for model selection in metagenomics (Year 1).
3. Development and evaluation of methods for weighted correlation analysis of metagenomic data (Year 2).
4. Applications of correlation analysis to publicly available data (Year 3).
5. Incorporation of methods into CAMERA (Years 2 and 3).

Note: Pollard will be the lead on this project. The postdoc will be co-supervised with Eisen..

MANAGEMENT PLAN

Our proposed management plan is modeled under the philosophy that the most exciting science will result from cross disciplinary efforts of the PI's and their respective groups, whose expertise span metagenomics, evolutionary biology, statistics, bioinformatics, computer science, and ecology. For this reason, though each project has a lead PI who will be coordinating the research, all PIs will participate in each. As outlined below, our efforts and contributions will remain highly integrated throughout the duration of the project - from the hiring of personnel to disseminating results and deliverables.

Hiring personnel

Positions for the project will advertised jointly to convey the interdisciplinary nature of the proposed research. All applications will be evaluated by all PIs, although the location of the hires will be affiliated with each project lead's laboratory (i.e. Eisen, Pollard or Green). All personnel will benefit from the mentorship all three PIs through regular correspondence (see below). Postdoctoral researchers seeking a cross-institutional experience (UCD-UO or UO-UCD) will be afforded that opportunity, meaning that each PI agrees to host post-docs who wish to visit from their home institution. An example scenario is a post-doc may wish to spend 2 years at UO and 1 year at UCD (although their home institution will remain UO). Cross-institutional experiences will be an option, not a requirement.

Weekly video conferencing

To ensure regular correspondence regarding progress on deliverables, the PIs will meet by video conference once a week (for ~ 1 hour) throughout the tenure of the project. On a bi-weekly basis these meetings will include participation of each PI's respective lab members who are affiliated with the proposal. The primary objectives of the PI-only meetings will be to discuss: 1) scientific issues requiring immediate attention, 2) timelines and schedules, including the status of project deliverables, 3) budget related issues, and 4) personnel management. The larger, more inclusive group meetings will be coordinated to foster collaboration, optimize information exchange and create feedback opportunities. Each meeting will highlight the efforts of an individual member of the group, rotating from person to person across meetings. During these meetings personnel will first present the major questions they are tackling and then have time to receive feedback from others. PI-only meetings will be managed via iChat, and the larger bi-weekly meetings will be held through video conferencing.

UC Davis-University of Oregon quarterly meetings

UC Davis and UO are both on the quarter system. We will coordinate quarterly joint lab group meetings, with two meetings a year at UC Davis, one meeting a year at the University of Oregon, and one at UCSD as part of one of the CAMERA related meetings (see below). These joint lab meetings will last a day. In the spirit of NCEAS (National Center for Ecological Analysis and Synthesis) and SFI (Santa Fe Institute) working groups, the mornings will include short presentations (15-20 minutes each) from every person affiliated with the project. The afternoons will be dedicated to round-table discussions and small break-out groups to brainstorm on particular topics. PI Green is likely to be in or passing through the Davis area for additional independent site visits due to: 1) co-PI status on an MRI NSF grant at her former institution UC Merced to access specialized flow cytometry equipment, 2) current collaborations with faculty other than Eisen and Pollard at UC Davis, 3) current collaborations with faculty at UC Berkeley, and 4) family close to Davis.

CAMERA meetings and teleconferences

We propose to hold three meetings a year at UCSD with CAMERA personnel to discuss our work and implementation of our methods into the CAMERA database. We will attempt to coordinate these trips with other CAMERA associated meetings such as the Metagenomics Conferences. We are working with Paul Gilna, the Director of CAMERA, to come up with a schedule for these meetings. It is expected that some but not necessary all personnel from each lab will go to each meeting.

In addition we propose to hold monthly video teleconferences with CAMERA personnel to facilitate the integration of tools developed in this project with the CAMERA database.

Project Wiki for private and public communications

We plan to use a Wiki site for both private and public communication about this project. The Eisen lab already uses such a wiki site http://128.120.136.15/mediawiki/index.php/Main_Page that is based on the OpenWetWare site (http://openwetware.org/wiki/Main_Page) used by dozens of labs to collaborate and communicate. For the Eisen lab this site allows all members of the lab, as well as many collaborators, wherever they may be, to keep up with the work being done by others. We are working on making use of electronic notebooks so that all work is tracked on the wiki. Pages can be switched from private to public relatively easily to allow broad dissemination of information and to invite communication from others.

Project synergy

Our revised proposal has made the scientific links between the different subprojects much tighter. For example, many of the projects will make use of the same gene family alignments and phylogenetic trees. Though these will be used for different purposes, by using similar co-dependent data sets this will encourage and foster communication among the labs regarding results. We aim to leverage this and other synergies among the different projects to help foster exchange of personnel and collaborations between the labs.

Disseminating results

We are committed to as open and public a release of our results generated from this project as is feasible. First, all publications will be in Open Access journals or using Open Access options in other journal. Our goal is to publish in journals, such as those from PLoS and BMC, that use the least restrictive Creative Commons licenses available to ensure as widespread dissemination and use of the publications. Any software generated from this project will be released under Open Source principles.

Scientific Advisory Group

We propose to create a Scientific Advisory Group made up of researchers with experience in the field of metagenomics. We would hold semi-annual meetings with this group by teleconferencing or conference call. At these meetings we would present results of our work and our research plans and solicit the group's input on our progress. This group would also be very helpful in selecting metagenomics data sets to work with for this project. We would be pleased if the makeup of this group was determined by the Moore Foundation, although we would be more than willing to select the group ourselves.

OTHER SUPPORT FOR PIS

Pollard

Pending projects

What Made Us Human?
NIH-NIGMS
9/01/2007 – 8/31/2010
Principle Investigator – Katherine Pollard, UC Davis
Percent effort = 30%

CAREER: Statistical Inference for Proteomics and Metabolomics data
NSF
7/01/2008 – 6/30/2013
Principle Investigator – Katherine Pollard, UC Davis
Percent effort = 20%

Green

Active and ongoing projects

Unifying Current Theories of Ecology
NSF –DEB 0628281
07/01/06 – 06/30/08
Principle Investigator: Jessica Green
Percent effort = 10%

Spatial Scaling of Bacterial Biodiversity
NSF - DEB-452454
04/01/06 – 03/31/09
Principal Investigator: Jessica Green
Percent effort = 20%

Microbial diversity of Ny-Ålesund soils: a preliminary study
The European Centre for Arctic Environmental Research (ARCFAC-026129-74)
Principal Investigator: Lise Ovreas, University of Bergen
04/01/07 – 03/31/08
Percent effort = 10%

MRI: Acquisition of a flow cytometer for multiparametric analysis of environmental, microbial and aquatic samples at UC Merced
NSF MRI-0723268
Principal Invstigator: Marcos E. García-Ojeda
08/01/07 – 07/30/09
Percent effort = 0%

Eisen
Active and ongoing projects

Microstates to macrodynamics: a new mathematics of biology program
DARPA - FA9550-06-1-0478.
1/1/07-8/31/08
Principal Investigator – Simon Levin, Princeton University
Percent effort: 10%

Microbial Genome Sequencing " EST Survey of Charophycean green Algae
National Science Foundation EF0523719
12/01/2005-11/30/2008
Principal Investigator - Charles Delwiche, U. MD
Percent Effort: 0%

Comparative Genomics of Chemosynthetic Symbionts
National Science Foundation EF133125
09/15/2004-08/31/2008
Principal Investigator – Colleen Cavanaugh, Harvard University
Percent Effort: 0%

The PhyloFacts phylogenomic encyclopedia of microbial protein families
National Science Foundation
12/1/2007-11/30/2010
Principal Investigator – Kimmen Sjolander, U. C. Berkeley
Percent Effort: 2.5%

Projects in no-cost extensions and/or expiring soon:

Tetrahymena Genome Sequencing Project
NIH-NIGMS Grant# R01 GM067012
Principal Investigator – Jonathan A. Eisen
04/01/2003-03/31/2008
Percent effort 10%

Microbial Genome Sequencing: Shotgun Sequencing of Tetrahymena thermophila,
National Science Foundation Grant# NSF-EF-024036
03/15/2003-06/30/2008
Principal Investigator - Jonathan A. Eisen
Percent Effort 10%

Phylogenomics: A Genome Level Approach to Assembling the Bacterial Branches of the Tree of Life
National Science Foundation - Grant# DEB-022865
10/01/2002-09/30/2007
Principal Investigator - Jonathan A. Eisen
Percent Effort: 20%

Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis
Moore Foundation via subcontract from UCSD
9/15/06-12/14/07
Principal Investigator – Larry Smarr – University of California, Dan Diego

REFERENCES

1. Yooseph, S., et al., *The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families.* PLoS Biol, 2007. **5**(3): p. e16.
2. Rusch, D.B., et al., *The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.* PLoS Biol, 2007. **5**(3): p. e77.
3. Denef, V.J., et al., *Implications of Strain- and Species-Level Sequence Divergence for Community and Isolate Shotgun Proteomic Analysis.* J Proteome Res, 2007.
4. Tyson, G.W., et al., *Community structure and metabolism through reconstruction of microbial genomes from the environment.* Nature, 2004. **428**(6978): p. 37-43.
5. DeLong, E.F., et al., *Community genomics among stratified microbial assemblages in the ocean's interior.* Science, 2006. **311**(5760): p. 496-503.
6. DeLong, E.F., *Microbial community genomics in the ocean.* Nat Rev Microbiol, 2005. **3**(6): p. 459-69.
7. Casas, V. and F. Rohwer, *Phage metagenomics.* Methods Enzymol, 2007. **421**: p. 259-68.
8. Delwart, E.L., *Viral metagenomics*. Rev Med Virol, 2007. **17**(2): p. 115-31.
9. Pennisi, E., *Metagenomics. Massive microbial sequence project proposed.* Science, 2007. **315**(5820): p. 1781.
10. Pace, N.R., *A molecular view of microbial diversity and the biosphere.* Science, 1997. **276**(5313): p. 734-40.
11. Olsen, G.J., et al., *Microbial ecology and evolution: a ribosomal RNA approach.* Annu Rev Microbiol, 1986. **40**: p. 337-65.
12. Hugenholtz, P., B.M. Goebel, and N.R. Pace, *Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity.* J Bacteriol, 1998. **180**(18): p. 4765-74.
13. Hugenholtz, P., et al., *Novel division level bacterial diversity in a Yellowstone hot spring.* J Bacteriol, 1998. **180**(2): p. 366-76.
14. Hugenholtz, P. and N.R. Pace, *Identifying microbial diversity in the natural environment: a molecular phylogenetic approach.* Trends Biotechnol, 1996. **14**(6): p. 190-7.
15. Giovannoni, S.J., et al., *16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the Green Non-Sulfur bacteria.* Proc Natl Acad Sci U S A, 1996. **93**(15): p. 7979-84.
16. Giovannoni, S.J., et al., *Genetic diversity in Sargasso Sea bacterioplankton.* Nature, 1990. **345**(6270): p. 60-3.
17. Chao, A., *Non-parametric estimation of the number of classes in a population.* Scandanavian Journal of Statistics, 1984. **11**: p. 265-270.
18. Baker, B.J., et al., *Lineages of acidophilic archaea revealed by community genomic analysis.* Science, 2006. **314**(5807): p. 1933-5.
19. Eisen, J.A., *Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes.* PLoS Biol, 2007. **5**(3): p. e82.
20. Weisburg, W.G., S.G. Giovannoni, and C.R. Woese, *The Deinococcus-Thermus phylum and the effect of rRNA composition on phylogenetic tree construction.* Syst. Appl. Microbiol., 1989: p. 128-134.
21. Badger, J.H., J.A. Eisen, and N.L. Ward, *Genomic analysis of Hyphomonas neptunium contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders 'Rhodobacterales' and Caulobacterales.* Int J Syst Evol Microbiol, 2005. **55**(Pt 3): p. 1021-6.
22. Eisen, J.A., *The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16s rRNAs from the same species.* J Mol Evol, 1995. **41**(6): p. 1105-1123.

23. Venter, J.C., et al., *Environmental Genome Shotgun Sequencing of the Sargasso Sea.* Science, 2004: p. 1093857.
24. Gruber, T.M., et al., *The phylogenetic relationships of Chlorobium tepidum and Chloroflexus aurantiacus based upon their RecA sequences.* FEMS Microbiol Lett, 1998. **162**: p. 53-60.
25. Lozupone, C. and R. Knight, *UniFrac: a new phylogenetic method for comparing microbial communities.* Applied and Environmental Microbiology, 2005. **71**: p. 8228-8235.
26. Webb, C.O., et al., *Phylogenies and community ecology.* Annual Review of Ecology and Systematics, 2002. **33**: p. 475-505.
27. Webb, C.O., J.B. Losos, and A.A. Agrawal, *Species issue: integrating phylogenies into community ecology.* Ecology, 2006. **87**: p. S1-S165.
28. Kraft, J.B.N., et al., *Trait evolution, community assembly, and the phylogenetic structure of ecological communities.* American Naturalist, 2007. **170**: p. 271-283.
29. Horner-Devine, M.C. and B.J.M. Bohannan, *Phylogenetic clustering and overdispersion in bacterial communities.* Ecology, 2006. **87**: p. S100-S108.
30. Webb, C.O., D.D. Ackerly, and S.W. Kembel, *Phylocom: software for the analysis of community phylogenetic structure and trait evolution. Version 3.41.* http://www.phylodiversity.net/phylocom/. 2007.
31. Bryant, J., et al. *Phylogenetic diversity across a landscape*. in *Ecological Society of America*. 2007. San Jose, California.
32. Kembel, S.W. and S.P.H. Hubbell, *The phylogenetic structure of a Neotropical forest tree community*. Ecololgy, 2006. **87**: p. S86-S99.
33. Swenson, N.G., et al., *The problem and promise of scale dependency in community phylogenetics.* Ecology, 2006. **87**: p. 2418-2424.
34. Green, J.L. and B.J.M. Bohannan, *Spatial scaling of microbial biodiversity.* Trends in Ecology and Evolution, 2006. **21**: p. 501 - 507.
35. Green, J.L., et al., *Spatial scaling of microbial eukaryote diversity.* Nature, 2004. **432**: p. 747-750.
36. Chao, A., et al., *A statistical approach to estimate soil ciliate diversity and distribution based on data from five continents.* Oikos, 2006. **114**: p. 479-493.
37. Chao, A., *Species estimation and applications*, in *Encyclopedia of Statistical Sciences*, N. Balakrishnan, C.B. Read, and B. Vidakovic, Editors. 2005, Wiley: New York. p. 7907-7916.
38. Bohannan, B.J.M. and J. Hughes, *New approaches to analyzing microbial biodiversity data.* Current Opinion in Microbiology, 2003. **6**(3): p. 282-287.
39. Curtis, T.P., W.T. Sloan, and J.W. Scannell, *Estimating prokaryotic diversity and its limits.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(16): p. 10494 - 10499.
40. Hong, S.-H., et al., *Predicting microbial species richness.* Proceedings of the National Academy of Sciences, 2006. **103**: p. 117-122.
41. Green, J.L. and J.B. Plotkin, *A statistical theory for sampling species abundances.* Ecology Letters, 2007. **10**: p. xxx-xxx.
42. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society B, 1977. **39**: p. 1-38.
43. Maddison, W.P. and M. Slatkin, *Null models for the number of evolutionary steps in a character on a phylogenetic tree.* Evolution, 1991. **45**: p. 1184-1197.
44. Martin, A.P., et al., *The rate and pattern of cladogenesis in microbes.* Evolution, 2004. **58**(5): p. 946-955.
45. Hahn, M.W., et al., *Estimating the tempo and mode of gene family evolution from comparative genomic data.* Genome Research, 2005. **15**: p. 1153-1160.
46. Hubbell, S.P., *The unified neutral theory of biodiversity and biogeography*. Monographs in Population Biology. Vol. 32. 2001, Princeton: Princeton University Press. 1-375.
47. Iwasaki, W. and T. Takagi, *Reconstruction of highly heterogeneous gene-content evolution across the three domains of life.* Bioinformatics, 2007. **23**: p. i230-i239.

48. Gotelli, N.J. and R.K. Colwell, *Quantifying biodiversity: proceedures and pitfalls in the measurement and comparison of species richness.* Ecology Letters, 2001. **4**: p. 379-391.
49. Pollard, K.S., et al., *Forces shaping the fastest evolving regions in the human genome.* PLoS Genet, 2006. **2**(10): p. e168.
50. Pollard, K.S., et al., *An RNA gene expressed during cortical development evolved rapidly in humans.* Nature, 2006. **443**(7108): p. 167-72.
51. Brown, D. and K. Sjolander, *Functional classification using phylogenomic inference.* PLoS Comput Biol, 2006. **2**(6): p. e77.
52. Consortium, T.G.O., *Gene Ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**: p. 25-29.
53. Perna, N.T., et al., *Genome sequence of enterohaemorrhagic Escherichia coli O157:H7.* Nature, 2001. **409**(6819): p. 529-33.
54. Tetz, V.V., *The pangenome concept: a unifying view of genetic information.* Med Sci Monit, 2005. **11**(7): p. HY24-9.
55. Lynch, M., *Streamlining and simplification of microbial genome architecture.* Annu Rev Microbiol, 2006. **60**: p. 327-49.
56. Lynch, M. and J.S. Conery, *The origins of genome complexity.* Science, 2003. **302**(5649): p. 1401-4.