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

TABLE OF CONTENTS

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 methods of 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). To develop this integrated approach we have brought together three labs to work on this project (Table 1) with expertise in the following areas:

Table 1. Principal investigators involved.

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

We aim to address six distinct but interrelated computational metagenomics projects (see Table 2). Three of these projects focus on fundamental questions about microbes that we and others have identified as having the potential to be revolutionized by metagenomic studies.[3] The other three projects are more methodological in focus and are designed to provide a general framework for the analysis of metagenomic data.

Table 2. Proposed projects.

Fundamental questions about microbes
1. Microbial biogeography and biodiversity
2. Microbial population structure and evolutionary dynamics
3. Functions of microbial communities
Methodologically focused
4. A statistical framework for metagenomics
5. Assessment of sequencing methods used in metagenomics
6. Expanding the use of simulations in metagenomic studies

[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

For each project, we propose novel mathematical, statistical and computational studies utilizing metagenomic data that will provide new insight into microbial ecology and evolution. In addition, we propose to develop novel computational and statistical methods for analyzing metagenomic data that will be of use to the community at large. We note the all methods development will have a component that will focus on making the methods readily available to the research community (e.g., through CAMERA). In addition, though we are not proposing here to explicitly test any predictions by carrying out experimental studies, we would like to work with other groups to do this.

We propose to carry out this work over three years. We outline specific deliverables for this time frame for each project in terms of the Years in which deliverables will be produced. For each project the deliverables have been determined in part by the number of personnel we have allocated to each (see Table 6). Timelines could be changed and priorities shifted as request by the Moore Foundation. In addition, we outline the general resources that would be needed to carry out the whole project at the end of the proposal including in particular computational resources.

Though each of these projects can be considered separate activities, they are highly interdependent. For example, the simulations will be used to both design and test new methods and to create artificial communities as controls to compare to real data. The methods being developed will influence both the development of simulation approaches and the scientific studies. And the results of the scientific studies will guide the simulations (e.g., by allowing simulations to mimic real community structure) and the development of methods.

In addition, though each project has a lead PI who will be coordinating the work, all PIs will participate in each. For example, in the studies of spatial patterns in biodiversity, Dr. Green will coordinate the work and focus on the ecological components, Dr. Eisen will aid in the genome-informatics and phylogenetic components, and Dr. Pollard will aid in the statistical components.

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.

I. MICROBIAL BIODIVERSITY AND BIOGEOGRAPHY

Metagenomics offers an unprecedented opportunity to explore the biodiversity of microbes. These "who is out there" questions can be considered to have three main components: how many types of organisms are present at a particular study site, what types of organisms are those, and what are the relative abundances of the different types? With this information one can then answer questions regarding the ecology and evolution of the communities (e.g., how does community composition shift across the environmental landscape?). Previously, these questions have been answered primarily through the use of PCR amplification of rRNA genes and then analysis of the rRNA data [10]. Using this framework, the types of organisms present are studied by examining the relative position of clones in a phylogenetic tree of rRNA sequences including those from other organisms and environments [10]. The relative abundance of organisms is estimated by first dividing up the rRNA sequences into phylotypes (also known as operational taxonomic units or OTUs) and then counting the number of clones for each phylotype. The total number of phylotypes in a sample (richness) is estimated from the data using approaches such as the Chao diversity index. These types of analyses can be used both to study single samples as well as to compare and contrast different samples.

The potential of metagenomics lies in the fact that it circumvents two of the key limitations of rRNA PCR: bias in PCR amplification and inaccuracies of estimates from analysis of rRNA [11]. PCR bias is avoided because the sequencing is effectively random and the rRNA limitations can be avoided by also using other genes for the analysis. Here we propose to use metagenomic data to address fundamental questions relating to the phylogenetic diversity of microbes and the spatial structure of microbial communities. To do this properly requires the development of novel mathematical and computational methods for estimating diversity. Below we describe the scientific questions we will be asking and the methods we propose to develop. Importantly, these methods will be of use to anyone wishing to analyze the diversity of microbes from metagenomic data.

Understanding how and why microbial diversity varies over the surface of the Earth (Green)

A key goal of ecology is to understand the spatial scaling of biodiversity. Patterns in the spatial distribution of organisms provide important clues about the mechanisms that regulate diversity and are central to setting conservation priorities [12-14]. Although microorganisms comprise much of Earth's diversity, little is know about their biodiversity scaling relationships relative to that for plants and animals. It has been argued that the small size, large abundance, high dispersal capability and short generation times of microbes result in fundamentally different diversity patterns compared to larger organisms [15-17]. Advances in molecular methods for quantifying microbial biodiversity, mostly based on community fingerprinting and rRNA-PCR, have resulted in empirical evidence suggesting that microbes exhibit spatial scaling patterns akin to larger organisms [18]. Despite these recent advances, consensus regarding the generality and consequence of microbial biodiversity scaling patterns is lacking [19].

Approach

To quantify the spatial structure of marine microbial communities we will use five patterns central to ecology: the number of taxa within individual sites (*richness*), the number of taxa unique to individual sites (*endemism*), the distribution of rare and common taxa within sites (*relative abundance*), the number of co-occurring taxa within sites (*co-occurrence*), and the change of taxa composition between sites (*beta-diversity*). We will report these patterns focusing on phylogenetic diversity as measured by the observed numbers of 16S genes and protein-coding genes within individual metagenomic samples. Statistical developments discussed in Methods 2 below will yield more rigorous estimates of diversity. Analysis of beta-diversity requires samples gathered in a spatially-explicit manner with sites separated by a range of geographic distances. Preliminary analysis of the GOS GPS coordinates shows that the sample design is particularly well suited for beta-diversity spatial-scaling analyses (Figure 1).

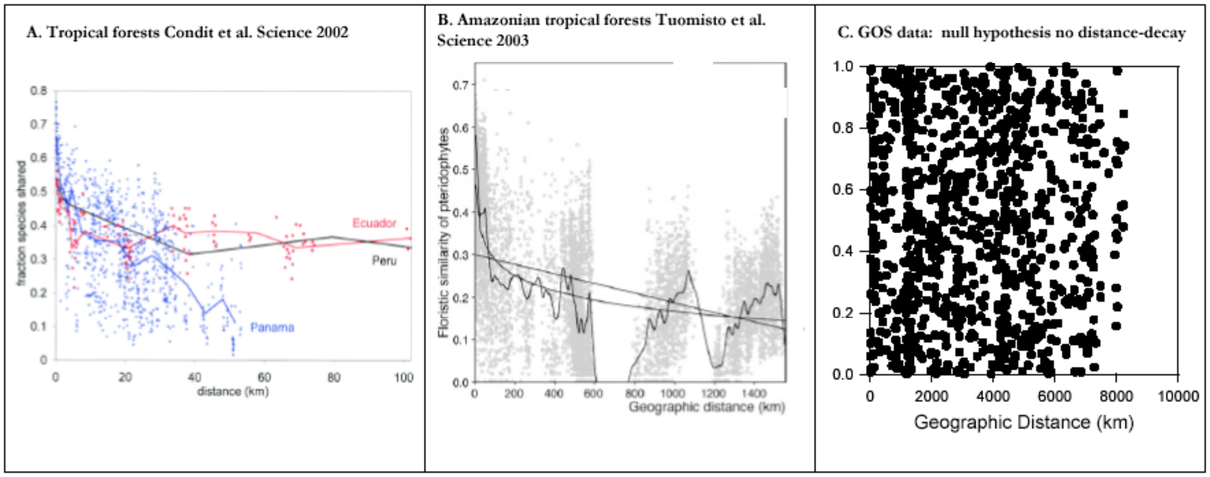

Figure 1. Spatial spread of GOS pair-wise comparisons (C) is well suited for classic distance-decay analyses (A & B). GOS data includes coastal and open ocean samples within 5 meters depth and assuming random values of compositional similarity. Coastal and open ocean data when plotted independently have comparable spatial spread.

To test the hypothesis that marine microbes have fundamentally different diversity patterns compared to larger organisms, we will also analyze data gathered by the H.M.S. Challenger (1872-1876) which circumnavigated the oceans in a manner parallel to GOS exploring the diversity of macroscopic marine life [20]. In the same spirit as CAMERA, the H.M.S. Challenger data is publicly available online. This analysis will entail georeferencing and taxonomically updating this historic data set.

To understand the mechanisms driving microbial diversity spatial scaling we will first leverage the phylogenetic beta-diversity patterns. We will use Mantel tests to quantify the role of geographic isolation (i.e. dispersal limitation) versus environmental heterogeneity (in instances where abiotic data is available) in generating observed patterns [21]. We will also leverage phylogenetic co-occurrence patterns to test mechanisms commonly invoked to understand the structure of diversity within ecological communities such as neutrality, competitive exclusion and abiotic filtering [22, 23].

Deliverables

1. Analysis of beta-diversity spatial-scaling for rRNA and protein-coding genes will be presented at a National Academy of Sciences Colloquium December 2007.
2. Comprehensive analysis of marine microbial metagenomic biodiversity (richness, endemism, relative abundance, co-occurrence, beta-diversity) using a spatially explicit framework. (Year 2)
3. Comparative analysis of microorganism metagenomic and microorganism morphospecies biogeography. (Year 2)
4. Insight into the mechanisms driving marine microbial metagenomic diversity. (Year 3)

What are the phylogenetic types of organisms present on the planet? (Eisen)

rRNA-PCR revolutionized studies of the tree of life by allowing uncultured organisms to be placed on the tree [10, 24]. This led to the discovery of dozens of novel major subdivisions of cellular organisms, as well as hundreds of novel branches within particular groups [25-29]. However, rRNA PCR poses limitations for three main reasons: trees based on rRNA are not always accurate (e.g., [30]), PCR does not amplify all cellular organism's rRNA genes [31] and some organisms do not have rRNA genes (i.e., viruses). 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 (e.g., [31]) and allowing the use of other genes to search for novel branches on the tree of life.

Approach

We have developed automated methods to build phylogenetic trees for various gene families and to search for novel metagenomics-only branches in these trees (Eisen et al., in preparation). We propose to further develop these methods and to use them for a select set of gene families to search for novel branches on the tree of life. Initially, the gene families will be a select set of 50 genes including all rRNAs and widely distributed protein-coding genes. These genes will be selected to cover bacterial, archaeal, eukaryotic and viral diversity. 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 identify in which environmental samples the unusual RpoB genes were found and then to look if in those samples other novel branches are found in analyses of other gene families. A truly novel group of organisms should show up in analyses of many gene families whereas artifacts may occur only for a small number of genes. Once novel lineages are found we will then characterize these lineages in more detail – such as by analyzing other genes that appear to be from these same organisms.

Deliverables

1. Automated system for searching for novel branches using rRNA sequences in metagenomic data (Year 1)
2. Automated systems for other marker genes (e.g., RpoB, HSP70, etc. as identified in Methods 3 below). (Year 2)
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. Targeted binning and the identification of other genes present in the genomes of these novel organisms (Years 2-3)

Methods 1: Identifying which genes to use and how to use them for metagenomic based diversity assays (Eisen)

It is clear that metagenomic data will be fundamental to studies of the biodiversity and biogeography of microbes. Above we outlined three key questions: how many types of organisms are present at a particular study site, what types of organisms are those, and what are the relative abundances of the different types? Using metagenomic data to answer these questions poses great challenges.

The first challenge is that all genes will not be equally useful for characterizing the multiple dimensions of microbial diversity. For example, a gene may be very useful in classifying organisms but it may not be useful in determining relative abundance due to copy number variation between species (we believe this is the case for rRNA). Other genes may be useful for both classification and relative abundance but may only work for certain taxa. The second challenge is how to normalize classification schemes between different gene families. For example, if one wanted to compare species richness estimates using 16s rRNA, RpoB and RecA it would be necessary to normalize ones definition of phylotypes for the different genes. For rRNA, frequently researchers use a 99% or 97% identity cutoff to define phylotypes. The percent identity cutoff for defining phylotypes using RecA or RpoB, two protein coding genes common to metagenomic diversity assays, may be very different. A final challenge is dealing with the fragmentary nature of metagenomic data. For example, if a fragment only covers a highly conserved portion of a gene a simple percent identity cutoff will lead to "overclustering" of this fragment with others in terms of defining phylotypes.

We believe there is a remarkably simple (though computationally costly) method for dealing with all of these challenges. The solution is a comparative analysis of all gene families found in complete genomes and the development of a database of weighting parameters for each gene family. These weighting parameters will be designed to quantify the relative contribution of each gene in various assays of microbial diversity. For example, if a gene is present in two copies per genome in all bacteria and one copy per genome in all archaea, then to calculate relative abundance of organisms from the number of copies of genes, one needs to divide the archaea abundance estimate in half compared to the bacterial abundance estimate.

Approach

We propose a two-tiered approach to identify a framework for utilizing a variety of genes in metagenomic based studies of diversity. First, we will analyze a set of 50-100 candidate genes selected to broadly cover bacteria, archaea, eukaryotes and viruses. This will include multiple protein coding genes (we have 20 in development) and 5S and 23S rRNA genes (which have mysteriously been neglected in this type of work). Second, rather than a priori selecting candidate genes we will examine all gene families found in complete genomes to assess and rank their utility for various purposes (using the methods described below). This approach will allow the identification of new metagenomic marker genes.

To derive a rank and weighting scheme of different genes for diversity assays we propose to build a likelihood matrix for all gene families that contains information about how useful the family is for phylotyping and estimating relative abundance and richness within samples. We will rank and weight genes in the following manner. We will start with complete genome sequences of both cellular organisms and a selected set of DNA based viruses. All proteins and non coding RNAs in these genomes will be placed into families and then subfamilies. Then alignments and phylogenetic trees will be generated for each subfamily. From these alignments and trees, multiple scores will be calculated to reveal how useful gene families are for various measures of diversity and to quantify weighting parameters for use in the assays of diversity that normalize the relative contribution of each gene. These weighting parameters will cover three diversity related tasks: phylogenetic classification, phylotyping, and estimates of relative abundance.

We note that using complete genomes as the basis for making these calculations will allow us to compare different gene families to each other without sampling biases due to having sequences from different genomic sets. A schematic summarizing this is shown in Figure 2 and more detail is given in the following text.

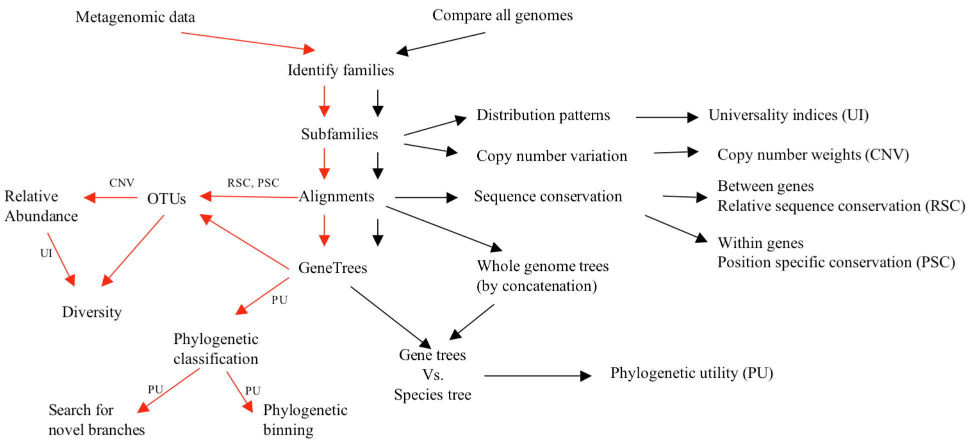

Figure 2. Development and use of weighting scores for metagenomic based measures of phylogenetic diversity. On the right of the figure, represented by black arrows, are the proposed analyses of complete genome sequences. On the left are the proposed anal yses of metagenomic data (arrows in red). First genes in these genomes will be divided into gene families and subfamilies. Then alignments and gene trees will be made for each subfamily. Analyses of these subfamilies, alignments and trees will produce we ighting scores for carrying out metagenomic analyses of diversity. From the distribution patterns of subfamilies we will calculate universality indices for families that will then be used to aid in converting relative abundance measures into diversity ind ices. 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 is of the alignments we will calculate sequence conservation indices. RSC will measure differences in conservation between gene families, which will be used to normalize the identification of OTUs between genes. PSC will measure variation in conservation within genes which 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 confidenc e levels on phylogenetic assignments.

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 trees of the species from which they came. 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.

Phylotype/OTU weighting

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.

Relative abundance weighting

Even if one assumes that one can identify phylotypes perfectly and normalize estimates between different genes, it is still not straightforward to use the number of "hits" to a particular phylotype to estimate the relative abundance of that phylotypes in the sample. In particular, three critical factors influence this "hits" to relative abundance calculation.. 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 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 by using our in vitro simulations as well as in silico simulations that utilize real sequencing data as raw material for the simulations.

Finally, to estimate the total number of taxa in a sample requires some information on the universality of a particular gene. Genes found only in particular phylogenetic groups can obviously only be used to estimate the total numbers for that group. To aid in such estimates we will calculate universality indices for each gene family both at a global level (GUI) and in a taxon specific manner (TUI). Though some genes may not be found in all taxa, one could use iterative approaches to estimate the total number of organisms present. For example, if one a set of universal genes for bacteria, and a separate set for eukaryotes and archaea, together they could be used to estimate total numbers of species of cellular organisms.

Deliverables

1. Generation of scores for select marker genes. Testing the utility of these scores using simulated metagenomic communities (Year 1)
2. Generation of scores for all genes. Testing the utility of these scores using simulated metagenomic communities (Year 2)
3. Integration of score with development of diversity assays (see above). (Years 2-3)
4. Creation and updating of a database (possibly through CAMERA) of the scores for different genes (Years 1-3)

Methods 2: Estimating microbial 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 and biogeography studies rely on statistical estimators of diversity. The most commonly used estimators of diversity were first derived by Anne Chao and colleagues in the ecology literature [32, 33]. These estimators, which focus on richness within samples and beta-diversity between samples, originate from the mark-recapture models of mobile animals. Despite their tractability, their applicability to metagenomic data is unknown. In addition, the Chao estimators lack a framework for predicting patterns of relative abundance within samples.

We propose to evaluate currently available diversity estimators and to develop new estimators for the study of metagenomic biogeography patterns including richness, endemism, relative abundance, co-occurrence and beta-diversity. Theoretically, metagenomic data will be more reliable than PCR data for making such statistical estimates since there is no amplification step in the generation of the data, although this has not been tested. Challenges will include assessing the accuracy of estimators when applied to different genes and in dealing with variability between organisms in copy number of genes (or in the number of copies of the genome per cell).

Approach

We will first evaluate the Chao richness and beta-diversity estimators by sampling from in vitro and in silico microbial communities where diversity is known. We will quantify how sample effort (the proportion of the community sampled) and community structure (patterns of relative abundance, endemism and co-occurrence) influence the accuracy of each estimator. 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. Novel estimates of relative abundance will be guided by the mathematical framework outlined in [34]. All statistics will incorporate the effect of copy number variation by using the relative abundance weighting schemes described above.

Deliverables

1. Assessment of the fidelity of currently used richness and similarity estimators. (Year 1)
2. Development of novel biodiversity estimators geared towards metagenomics. Chao, Green and Jiang have planned their first meeting in California September 2007. (Year 2)
3. Publicly available software parallel to Estimate S for the metagenomics community (http://viceroy.eeb.uconn.edu/EstimateS). (Year 3)

II. Microbial Population Structure and Evolutionary Dynamics

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 the spatial patterns of genomic variation and studies of the connection between patterns of genome evolution and environmental/ecological properties of communities (both extending the studies of biodiversity spatial patterns proposed above).

Spatial patterns in genomic variation within microbial species (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 [35]. Some of this genomic variation is due to the occurrence of lateral gene transfer which can create large genomic differences between otherwise closely related types. Whatever the cause of the variation, this means that organisms that are placed into the same or related phylotypes can have significant biological differences. This fact has been the driving force behind the development of metagenomics – since the position of an organism in the rRNA tree of life is not always useful in predicting its biology. The genomic variation among close relatives means that even more than for plants and animals, to gain a full understanding of the evolution and ecology of microbes one needs to look at variation within species.

Though genome sequencing of cultured species has led to many new insights, it is limited in that it usually focuses on single populations from certain environments and it has a bias introduced by culturing. Metagenomics allows for the first time genomic variation patterns to be studied in multiple species simultaneously. This will not only be useful for understanding the structure within populations of microbes but will also be of value in designing binning methods (see below).

Approach

To carry out this work we will do two types of analyses. First, we will compare closely related populations/species of microbes within single samples. Second we will compare those same populations/species across multiple samples. 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 [36] implies that gene flow occurs at nearly an infinite rate in microbial communities. We will be able to measure gene flow using a metagenomics F_{ST}-like statistics (which measures within vs. between population genetic variation) to test predictions of the pangenome hypothesis.

Deliverables

1. The first genomic-based studies of multispecies geographic clines in microbes. In Year 1 we will focus on three phylotpyes from available transect data sets (e.g., HOTs, GOS). In Years 2-3 we will expand to more organisms.
2. Development of F_{ST} like measures of genomic variation within communities versus between communities (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 3)
4. 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 [37, 38] to do this for multiple microbes at once. (Years 2-3)

How do measures of genome evolution relate to the ecosystem? (Pollard)

Genome sequence analysis has revealed that cultured microbes differ significantly in their evolutionary properties including rates of mutation and recombination. These differences significantly affect the evolvability of organisms, such as their ability to genetically respond to environmental change. It has also been found that ecological niche can influence this evolvability – with the best example being intracellular microbes having high mutation rates and low levels of recombination. Metagenomic data allows for the first time 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). In addition metagenomics 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.

Approach

Genomic regions and genes with evolutionarily unique sequence distributions (rates and patterns of substitution, polymorphism, recombination, etc.) will be compared across metagenomic samples from different ecosystems. Correlation analysis will be used to identify associations between the sequence data and ecosystem variables, including demographic variables (e.g. population size), characteristics of the microbial community (e.g. species diversity, competitors, symbiotic relationships), and environmental variables (e.g. temperature, salinity, pH). Correlation analysis will also be done on general patterns of "evolvability" such as mutation rates, population size, recombination patterns with community characteristics. We will address questions such as: do certain gene families or functions evolve more quickly in particular ecosystems? Which biological processes are most stable? Which are most environmentally sensitive? Do some environmental variables foster rapid evolution more than others? 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 which 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 (or function of nearby genes for non-coding regions), using protein families [39] and publicly available ontologies [40]. This project will employ the evolutionary and population genetic methods described below, as well as the statistical approaches to correlation analysis developed in Project IV.

Deliverables

1. Assessment of global evolutionary patterns and trends. (Year 1)
2. Correlation of evolutionary patterns with ecosystem variables. (Year 2)
3. Detailed investigation of particular cases of very rapid genome. (Year 3)

Methods 1. Binning (Eisen)

A critical step in metagenomic analyses is binning in which one attempts to assign reads to bins that correspond to organisms from the sample. One reason this is critical is that one can treat bins much like one would treat a genome of a cultured organism and then one can use various computational tools developed for cultured species (e.g., the prediction of metabolic pathways) on the bin. Another reason binning is critical is for doing population genetic analysis of metagenomic data. For example, if one wants to look at allele diversity in relation to environmental parameters using standard population genetic methods requires calculating allele diversity within species.

Approach.

Although there is a diversity of approaches to binning we believe that ecological and evolutionary approaches can lead to great improvements. We propose to develop three novel approaches to binning. First, we propose to make use of the phylogenetic analysis of all gene families described above to develop and test the utility of phylogenetic binning, a method we pioneered in analysis of symbionts of the glassy winged sharpshooter [41] but which has not been systematically developed or tested. Second, we propose to develop methods that will explicitly link diversity assays of a sample to binning. For example, what type of binning method is ideal will vary depending on the number of species in a sample. In addition, binning should be greatly improved by comparing similar samples with differences in the relative abundance of particular taxa (the autocorrelations in abundance within vs. between samples can be used to identify bins). Third, we propose to develop methods that will use phylogenetic and population genetic analyses to assess whether proposed bins appear to contain more than one species. For example if a bin contains four closely related but non-interbreeding species, phylogenetic analysis of different genes should always show four subtypes. In contrast, if the bin contains four interbreeding populations, then different genes will likely show different patterns due to recombination. For each of these approaches, binning methods will be tested both with realistic simulations (see below) and through the comparison of the ecological and functional studies described throughout this proposal (e.g., unusually high functional diversity levels could indicate that a proposed bin contains many organisms).

Deliverables

1. Development and testing of methods for phylogenetic binning. (Year 1)
2. Development of methods for linking binning and diversity measures. (Year 2)
3. Development of phylogenetic and population genetic methods to deconvoluting and testing proposed bins. (Year 3)

Methods 2. Measuring genome evolution in metagenomic data (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. Since metagenomic data sets provide for the first time random samples of microbial communities, they also have enormous potential to be used to infer details of processes that shape the populations. Our goal is to identify a reasonable approach to computing molecular evolutionary and population genetic parameters from metagenomic data. These include mutation rates and patterns, recombination rates and patterns, and selection (both positive and negative). We are particularly interested in identifying rapidly evolving genes and genome sequences, which has been a fruitful approach in studies of mammalian evolution [42, 43]. 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. Similarly, population genetic parameters are difficult to estimate without knowledge about which sequences belong to which genomes. The main barrier for studies of metagenomic genome evolution is the lack of clear definitions of organism and species. For example, in this context population diversity is difficult to distinguish from species divergence. And patterns of linkage between alleles cannot be estimated.

It has been proposed that recombination rates and patterns may be estimated from inter- and intra-sample variation [3]. Binning will provide an alternative approach; by treating the bin as a “genome”, many standard parameter estimators for genome evolution can be applied. Both of these approaches await validation and testing, which our simulation approach (Project VI) will facilitate. There is considerable risk, however, that an inaccurate species concept will lead to flawed analysis. This is particularly true for less abundant species whose sequence reads are unlikely to be binned accurately. An alternative approach is to use simulations to investigate what sequence data would look like under a range of different assignments of sequence reads to populations. While these simulations are not likely to identify the exact population structure and allele distributions in a sample, they may help us rule out certain interpretations of the data that are highly unlikely. Finally, it may also be possible to estimate substitution rates and test for directional selection in the absence of a species concept (*e.g.* comparing synonymous to non-synonymous rates within a protein family).

Another roadblock in the analysis of metagenomics genome evolution is the lack of a species phylogeny. Calculation of substitution rates (synonymous and non-synonymous), for example, typically requires estimation of an ancestral state. One approach to this problem, for protein sequences, is to use the phylogeny of related proteins in a sample in place of the species phylogeny. While this method will not easily distinguish paralogs from orthologs, it will enable estimation of substitution rates in well-sampled protein families.

In addition to empirical studies, we will devote effort towards theoretical developments, including new population genetic models for metagenomics. In all of these analyses, care will be taken to account for sampling issues. These investigations will lead to recommendations for software development for CAMERA and other projects, including our own project to develop open source statistical software for molecular evolutionary analysis as part of the Bioconductor project (http://bioconductor.org [44]). Finally, we will apply the most promising methods to sequence data from GOS and other metagenomic studies and evaluate global patterns of genome evolution. These patterns will then be related to characteristics of the ecosystem, as described above.

Deliverables

1. Development of molecular evolutionary methodology for metagenomics. (Year 1)
2. Development of population genetics methodology for metagenomics. (Years 1-2)
3. Empirical evaluations of methods for measuring genome evolution. (Years 2-3)
4. Software development and applications. (Year 3)

III. Function(s) of microbial communities

The driving force behind the development of metagenomics has been the ability to use analysis of genome sequences of uncultured organisms to predict their biological properties. A hallmark discovery that launched the field of metagenomics was the finding of proteorhodopsin in uncultured microbes in surface ocean waters. Subsequently, analysis of metagenomic data has led to many fundamental insights into the biological properties (e.g., metabolism, pathogenicity, light sensitivity) of uncultured microbes, from endosymbionts to microbes in the human gut to those of the deep sea and open ocean. We believe that much greater insights about the functions present in microbial communities can be provided through the better integration of ecological, evolutionary and statistical methods in the analysis of the metagenomic data sets. For example, most current approaches to making functional inferences from metagenomic data involve treating the community like a bag of genes. However, it is clear that compartments (i.e., cells, populations, species) matter in making such predictions (e.g., [41]). We propose here to focus on two research areas that can particularly benefit from an integrated approach – the link between ecosystem function and biodiversity and the identification of novel, previously unknown functions from metagenomic data.

Quantifying the link between microbial diversity and ecosystem function across scales

Disentangling the link between biological diversity and ecosystem function is a fundamental goal of scientists and policy makers worldwide [45-47]. It has long been assumed that unlike plants and animals, microorganisms are functionally redundant, meaning that distinct microbial consortia are capable of the same ecosystem functions and services [48]. Under this paradigm microbial community composition is not important to ecosystem function, conservation of microbial populations is moot [19], and microbial extinction poses no threat to the natural flow ecosystem services. Metagenomics makes it possible to rigorously test these assumptions. We propose to quantify the link between microbial diversity and ecosystem function using an integrated approach that leverages classic distance-decay analyses from ecology and recently available metagenomic data sampled from ocean environments. Beyond addressing the question of functional redundancy in microbial communities, our analyses will yield insight into the role of environmental heterogeneity and spatial scale in driving phylogenetic and functional diversity patterns.

Approach

We will first examine the GOS data set, as it provides information for a large number of protein-encoding genes sampled across 41 marine microbial communities along an 8000 km transect; more metagenomic data sets will be studied as they become available through CAMERA. To begin our analysis, we will answer the following basic questions which have yet be explored: 1) which communities are the most and least diverse, both phylogenetically and functionally?, 2) is there a relationship between functional and phyogenetic diversity?, and 3) what environmental parameters and ecosystem properties are correlated with the diversity and signature of functional and phylogenetic types? Phylogenetic diversity in each sample will be estimated using the methods outline above in "Estimating microbial biodiversity from metagenomic samples", and functional diversity will be estimated using the methods described below

Next we will address the longstanding scientific assumption that microbial communities are functionally redundant. To do this will require estimating the similarity of different microbial communities both functionally and phylogenetically. As explained above for diversity estimates, phylogenetic and functional similarity (or beta-diversity) between samples will be estimated using newly developed metagenomics statistical techniques. We note that the GOS data may be organized into protein families and taxonomic groups to quantify functional similarity and phylogenetic distance between a large number of pairwise comparisons. We have derived a novel way to assess the degree of functional redundancy between microbial communities by examining the degree to which phylogenetically distinct communities harbor the same metabolic or biochemical potential. Our approach is briefly outlined in Figure 3.

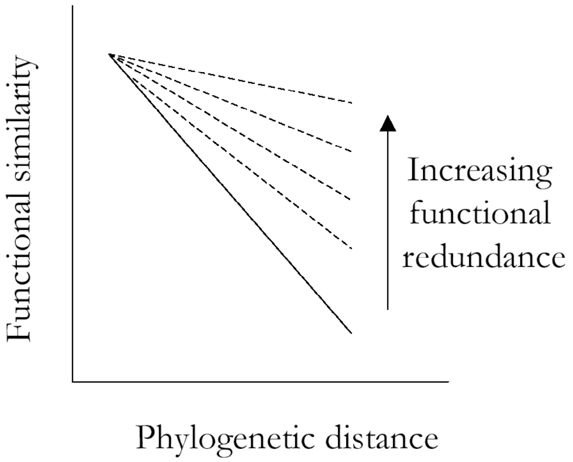

Figure 3. Example given the parsimonious assumption that phylogenetically similar communities harbor similar functional genes (upper left-hand corner of curves). If phylogenetically distinct communities harbor similar functional genes, this suggests functionally redundancy.

Deliverables

1. Statistical methods for assessing the functional diversity within samples and the similarity in function between samples. (Year 1)
2. A quantitative analysis of phylogenetic and functional diversity, and the relationship between the two, across environmental gradients and spatial scales. (Year 2)
3. A map of the correlation between functional or metabolic similarity and phylogenetic similarity, which will address the longstanding assumption that microbial communities are functionally redundant. (Year 3)

Using metagenomics to find novel functions (Eisen)

Genes and genomes of uncultured organisms were originally studied by those searching for new functions and processes that could be used for industrial and biotechnological purposes (e.g., Diversa, Microbia, etc). This is still one of the most important potential uses of metagenomic data and analyses. Metagenomics has potential for providing clues to novel functions and processes for a variety of reasons but especially due to the random nature of sampling genomes from uncultured organisms. For example, we previously showed in analysis of the Sargasso Sea data that a large number of novel proteorhodopsin families were present that had been missed by PCR based methods. In principal, embedded within metagenomic sequence data are sequences that encode completely novel pathways, activities, or specificities. We believe that searching for such novel activities can be aided by the integration of ecological, evolutionary and statistical approaches. Information on novel functions can be of great value in understanding the evolution of communities and the mechanisms underlying the origin of novelty.

Approach.

We propose to take a two-tiered approach to studying functional novelty. First, we will design new methods for identifying potential novel functions and for quantifying novelty within metagenomic samples (here by novelty we mean functions and processes that are rare or not seen in cultured organisms). This will include methods to identify novel protein families and subfamilies, novel proportions of different types of genes, novel gene combinations, and novel molecular processes (e.g., new genetic codes). In addition, we will develop and use methods to use non-homology functional prediction methods for metagenomic data (e.g.. co-occurrence of genes). Second, we will examine whether functional novelty is related to any community properties including phylogenetic novelty, population structure, and degree of isolation (e.g. unmixed vs. mixed). The key question here will be – are there communities that possess more novelty than others? And if so, why? Is it because new processes evolve more readily in those communities? Of it is simply due to the presence of novel phylogenetic types of organisms (as identified in Project 1 above?).

Deliverables

1. Development of functional novelty indices for metagenomic data and testing against simulations. (Year 1)
2. Comparison of protein family novelty (as identified by Yooseph et al. with phylogenetic novelty) for different GOS samples. (Year 2)
3. Identification of environmental parameters that are related to metagenomic novelty. (Year 3)

Methods. Functional diversity, novelty, and community metabolic potential (Eisen)

Perhaps the greatest informatics challenge in analyzing metagenomic data lies in making predictions about the functions present in individual organisms and their communities by analyzing the fragmented, unlinked, undersampled genomes. The real challenge is that functional prediction methods for genome sequence data are designed around the analysis of the genomes of single organisms not communities. Even for as simple a system as the two symbionts living inside the glassy winged sharpshooter, functional prediction for the community did not work well. In the symbiont case the only thing that worked was binning the data into pools corresponding to each symbiont, making separate predictions for each, and then integrating the information back together to make predictions for the community [41]. And these difficulties came up in a study of a single sample, with very very few organisms present. It is orders of magnitude more difficult to design methods for analyzing more complex ecosystems and for comparing samples to each other.

We believe it is currently virtually impossible to make complete, robust predictions of all the functions present in organisms or communities from metagenomic data. However, we believe there are some useful measures and indices that can be developed to aid in comparative studies of different metagenomic data sets, even with the caveat that specific functional inferences are almost certainly suspect.

Approach

Our general approach is to use simulations and statistical resampling of real metagenomic data sets to ask questions about how robust functional predictions can be and to aid in the design of functional diversity indices. For example, a simple question one might ask is, how robust are binning independent estimates of functional diversity. That is, if you do not assign metagenomic data to bins, how well can you make inferences about functional diversity? Another simple question is how well do different functional diversity estimates perform in various simulations? For example, one way of estimating functional diversity would be to map all predicted proteins in a system to gene ontology role categories and measure how much of the GO space is covered. Another method would be to map genes to gene families and see what proportion of gene families are represented. A third method would be to map functions to metabolic charts and measure the total network coverage. One could then create simulations with different functional mixes (e.g., one could have all photosynthetic species, another could have diverse metabolisms represented) and see which methods perform well. One could use a similar approach to comparing samples (using some of the methods described in the statistics section below).

Deliverables

1. Development and testing of functional diversity measures for metagenomic data including mapping of genes to GO categories, to PFAM families, and to metabolic charts (Year 1)
2. Design of methods to distinguish convergent similarities from homologous ones (Year 2)
3. Design of methods to weight functional diversity measures by how likely it is to find certain genes (e.g., a single nitrogen fixation gene might count more than a single carbohydrate transporter). (Year 2)

4. Testing of methods by comparing spatial patterning of phylogenetic diversity vs. functional diversity (Year 3)

IV. Statistical approaches to metagenomics

A Statistical framework for metagenomics (Pollard)

In a newly emerging field such as metagenomics it is essential to evaluate the reliability and repeatability of data analyses, since there is not typically a "gold standard" to which results can be compared. The simulations proposed in Project 6 are a critical component of methods assessment. Another is the development of measures of variability for each estimator (distance, correlation, and diversity), so that statistical significance can be assessed. Both simulation studies and significance testing rely on a sound statistical framework.

Approach

The first step to defining a statistical framework is to develop a concept of the sampling unit. What is the thing on which measurements are made? Is it the pool of DNA sequenced? Or the species? Or the gene family? The answer will depend to some extent on the application. Regardless, we must ask if the data we are analyzing represents a random sample from a population of interest. What is that population? Could we easily collect more data from the population? Were enough samples collected to assess variability? Were they collected and measured in an unbiased way or at least in a way in which we can quantify bias? With the concept of a statistical (not necessarily genetic) population in hand, we can define our data as realizations from a data generating distribution for this population. Then, quantities of interest, such as correlations and measures of diversity, can be viewed as parameters of this population distribution, which are estimated from observed data. This framework allows us to develop simulation-based and theoretical measures of variability for parameter estimates, thereby facilitating estimation of confidence intervals and assessment of statistical significance. This work will feed into the studies of similarity, correlation, and diversity described in Projects I and II as well as the simulation studies in Project VI. We will publish a paper about the statistical framework itself, as we did for microarray data [49]. Such a publication would include recommendations, along the lines of Yang & Speed [50], for optimal study designs for use in future data collection efforts.

Deliverables

1. Formulation of a statistical framework for metagenomic studies. (Year 1)
2. Development of study design/sampling recommendations. (Year 2)
3. Extensions and applications to other project aims. (Year 3)

Similarity measures for metagenomics (Pollard)

As the means for collecting and storing increasingly large amounts of metagenomic data develop, it is essential to have sound methods for measuring similarity between samples of sequence data. Similarity measures are useful at many levels of analysis, from binning to phylogenetics. Furthermore, a concrete definition of distance between samples will enable analysis of associations between sequence data and metadata, such as ecological and geographic variables. The goal of this project is to understand which measures perform best under a range of realistic data scenarios.

Approach

Similarity measures quantify the distance between any pair of samples. The Pollard lab and others have developed and employed a variety of approaches for other fields of genomics [44, 51]. The first issue to resolve with regard to developing robust and useful similarity measures for metagenomics is what level of sequence data to use: phylotype/bin, protein family, protein, or sequence read. For example, consider the goal of clustering samples based on sequence data to see if these clusters correlate with environmental variables of interest. Do we want to call two samples similar if they contain the same organisms? Or the same kinds of genes (*i.e.* functional distance)? Or actually the same genes/sequences?

A second consideration is how to encode the sequence data for computations: presence/absence (binary variable) or score (quantitative variable, 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. For each choice of data encoding (binary or quantitative), there are many different formulas for quantifying similarity, including metric (*e.g.* Minkowski) and non-metric (*e.g.* correlation) distances. These equations can be extended to include weights that may be used to adjust for differences in quality, sampling depth, etc. between samples generated with different technologies and in a variety of labs. The results of this investigation will be published and will lead to recommendations for software implementation through the CAMERA and Bioconductor projects.

Deliverables

1. Survey of metagenomic data to determine the scope of data types and similarity concepts. (Year 1)
2. Development of weighted similarity measures for metagenomics. (Year 2)
3. Simulation-based comparison of different similarity measures. (Year 3)
4. Software development and applications. (Year 3)

Correlation of genome sequence data with metadata (Pollard)

In addition to comparing metagenomic samples to each other, we are also interested in measuring association between sequence data and a variety of different types of metadata, such as environmental clines and measures of ecosystem complexity. Identifying correlations between genomic variables and environmental variables is at the core of many of the scientific questions in the metagenomics research community, including those posed in Projects 1-3. Our goal is to develop models that allow us to quantify the magnitude and statistical significance of 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, like similarity analysis, relies on several choices about data handling. As discussed above, metagenomic data can be analyzed at the level of phylotype, functional category, or sequence/allele. In addition, each type of data can be encoded as binary (present/absent) or quantitative (score or frequency). Similarly, metadata may be categorical or continuous. 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). Results will be disseminated through publications and conference presentations. Promising methods will be included in software implementations through CAMERA and Bioconducor.

Deliverables

1. Methods for model selection in metagenomics. (Year 1)
2. Methods for weighted correlation analysis of metagenomic data. (Year 2)
3. Simulation studies comparing correlation methods. (Year 2)

4. Software development and applications. (Year 3)

V. Assessment of Sequencing and Sampling Methods Used in Metagenomics (Eisen)

Many important questions in metagenomic studies relate to the methods used for sequencing and gathering samples. We propose here to use a combination of simulations and analysis of real data sets to address some of the most pressing questions in the techniques used in metagenomics. Two areas we believe are most important to evaluate at this time are the design of the metagenomic sequencing and the selection and sequencing of reference genomes. We discuss both here but in addition propose to use similar approaches to address other metagenomic methods questions as they arise (e.g., these could come from the CAMERA Scientific Advisory Board).

First, we consider methods in metagenomic sequencing. Metagenomic sequencing includes a diversity of methods and options within each method. A critical question for the field is which methods and options should be used for particular purposes. For example, there are at least three major classes of methods that can be used to carry out the sequencing: Sanger clone based, 454 cloning independent, and massively parallel but short read methods (e.g., ABI-solid or Illumina/Solexa). These methods have different read lengths, error rates, and costs. Which is best? Even if one chooses a method (e.g., based upon availability of a machine) there are still many options to choose from. For example, for clone based capillary sequencing, one has to choose the library insert size.

One method that has been used more and more recently in metagenomic analysis is the comparison of metagenomic data sets to the “reference” genomes of cultured species that are closely related to phylotypes found in the metagenomic data. Such reference genomes have been in a variety of ways including as scaffolds for assembly (e.g., [52]), as tools in binning data (e.g., [2], [53]), to enable studies of recombination (e.g., [3, 54]) and to identify genomic islands that are absent from particular environmental samples. It is the potential uses of such reference genomes that has led to “Moore 175” project to sequence relevant marine microbe references, an NHGRI project on the human microbiome, and multiple other projects to inform particular environmental studies. However, there has been no significant effort to determine how to select reference genomes (e.g., how closely related do they have to be to phylotypes from the environment to be useful) nor to determine just what information is needed for these genomes (e.g., do the genomes have to be complete, do they have to be annotated?). In addition, the information one has about reference genomes will influence which sequencing methods are most useful for the metagenomic data itself and thus questions about reference genomes are directly linked to questions about sequencing raised above.

Approach

We propose to use a combination of simulations to assess the methods used for sequencing metagneomic samples and the methods for sequencing and selection of reference genomes. For metagenomc sequencing, we will compare different sequencing methods and different methods and options within each method. To compare methods we will generate and then compare data sets with either same total amount of sequence data generated or the same total cost of generating the data. For each simulation we will ask how well the data can be used to answer some of the questions raised in the three scientific research project areas. We will focus our analysis in particular on questions relating to read length, sequence quality, insert size, and availability of mate pairs.

For the analysis of reference genomes, using our simulations and analysis of real data, we will conduct comparisons of results varying which reference genomes are used and the state of the genome (e.g., closed vs. 8x coverage, Sanger vs. 454 draft sequencing). Among the questions we will test: what is the effect of phylogenetic distance of the reference genome to the phylotypes in the data set?, what is the effect of the distance of the closest phylotype to other phylotypes in the data set?, and how much does using finished reference genomes improve ones inferences compared to unfinished genomes (different levels of coverage will also be analyzed)?

Finally, we will combine the two questions to ask if the use of reference genomes affects which sequencing method is useful. Theoretically if one had enough reference genomes then short sequence read methods might become quite useful for metagenomic studies much in the way they are being used for genotyping in single organisms. This can be assessed through simulations.

Deliverables

1. Comparison of Sanger, 454 and Illumina/ABI methods for measuring phylogenetic diversity. (Year 1)
2. Comparison of library insert sizes for Sanger sequencing. (Year 1)
3. Comparison of sequencing methods for prediction of community functions. (Year 2)
4. Comparison of single-sided versus mate pair based methods. (Year 2)
5. Generation and release of in-vitro and in-silico data sets from comparisons of different sequencing methods. (Years 1-2)
6. Analysis of new sequencing methods as they become available. (Years 1-3)
7. Generation of “comparison” simulation data sets for publicly available metagenomics projects. For example, release of a simulation of GOS with Sanger vs. 454. (Years 1-3)
8. Analysis of the effects of sequencing errors on inferences (especially important for some of the new sequencing methods) (Year 2)
9. Analysis of the effect of phylogenetic distance of reference genomes. (Year 1)
10. Analysis of the effect of different levels of coverage of reference genomes. (Years 1-2)
11. Analysis of the use of multiple reference genomes for particular groups. (Year 3)

VI. Expanding the use of simulations in metagenomic studies (all PIs)

Simulations are a critical component of theoretical and computational biology. While simulations are leveraged in many scientific disciplines such as physics, ecology and genomics, this is not the case in metagenomics (although there are notable exceptions [55]). We propose to integrate the use of simulations with our methods development and scientific studies. In addition, we will release simulated data sets for use by others in the field.

Approach

The main purpose behind the use of metagenomic simulations is to create data sets of known entities that can be used to test the accuracy, consistency and robustness of metagenomic methods. We propose to use three types of simulations (Table 3) of metagenomic data sets: (i) in silico creation of artificial communities using genome sequences of isolates, (ii) in-vitro simulations generated by mixing DNA or cells of organisms and then submitting these to standard metagenomic processes, and (iii) resampling simulations in which real metagenomic data sets are resampled in various ways.

Table 3. Types of Metagenomic Simulations.

Type	Details
In Silico	In silico communities of organisms whose genome sequence is known will be created and metagenomic sequencing of these communities will be simulated. For this we will use both simulated sequence reads, and real reads from the trace archive.
In Vitro	DNA, cells or libraries from organisms whose genome sequence is known will be mixed in vitro and these mixes will then be used for metagenomic sequencing.
Resampling	Metagenomic data from real communities will be resampled (e.g., paired ends will be broken, a subset of reads will be randomly selected). For example, simple symbiont communities where binning is well-resolved will be used to test binning methods.

The simulations will be used to control specific variables that will then allow the testing of how well methods perform in the face of such variation. Such variables are easiest to control in the in silico simulations and those we plan to use are summarized in Table 4.

Table 4. Variables for computational (in silico) simulations

Variable	Details
Community structure	
Number of taxa	The total number of organisms used will be varied. This could include both real organisms and simulated genomes (e.g., simulated recombinants).
Phylogenetic distance between types	Many metagenomic methods should work well when taxa are all distantly related but will start to fail with closely related sets of organisms.
Relative abundance of taxa	Keeping the total number of individual cells and total taxa richness constant but changing the relative abundance of taxa.
Particular isolate used for specific phylotypes	Sets of simulations will be created where all variables are held constant (e.g., number of types and relative abundance of those types) but where specific types are replaced by close relatives (e.g., *E. coli* K12 will replace *E. coli* O157:H7).
Type but not functions	We will create artificial communities with functionally similar dominant taxa but from different taxonomic groups. For example, photosynthetic communities dominated by cyanobacteria vs. green sulfur bacteria vs. chloroflexi vs. algae. This will be used to assess functional analysis methods.
Functions but not type	We will create artificial communities with phylogenetically similar dominant taxa but with different functional groups. For example, communities dominated by photosynthetic, chemosynthetic or heterotrophic proteobacteria.
Genome features	We will vary features such as GC content, number of genetic elements, presence of repetitive DNA, number of copies of particular genes (e.g., rRNA), and genome size.
Creating alternative genomes	
Creating recombinants from known genomes	Multiple sequenced genomes will be mixed to create recombinants.
Simulation of evolution	Mutation, deletion, recombination will be simulated to create more realistic populations of genomes for the simulations.
Lateral gene transfer	Foreign genes will be artificially inserted into some of the genomes.
Generation of sequence data	
Depth of sequencing	We will test how the depth of sequencing affects various inference methods.
Sequencing methods	We will simulate different methods of sequencing, holding either the total number of bases constant or the total costs constant. We will also simulate errors for each method, such as clone chimeras and sequencing errors.
Library properties	We will simulate different insert sizes for shotgun libraries. This variable is likely important for both methods assessment and design of metagenomics projects. For example we have shown that binning methods work better with paired end sequences from 20 kb libraries than from 2 kb libraries.
Other issues	
Resampling	Multiple simulations will be run on each data set to assess the effects of sampling on conclusions.
Environmental mimics	For specific metagenomic projects we will create simulations that try to mimic the natural community, both in terms of species diversity and also by using genomes from organisms related to those from the environment when possible.

We also propose to perform a series of in vitro simulations but since these are more expensive these will be narrower in scope. Most important in these simulations will be to generate data sets using multiple sequencing methods from the same samples.

A critical part of the use of these simulations is that they will be done in conjunction with the other Tiers and thus the simulation design will be geared towards assessing specific methods and as controls for particular scientific questions. For example, to test methods for estimating the number of species in a metagenomic sample, we will create simulations with different numbers and relative abundances of taxa as well as the depth of sequencing performed. To address questions concerning the ideal sequencing methods to use, we will create simulations of different methods and then ask how metagenomic analyses (e.g., estimates of diversity) are affected by the different sequencing methods.

Deliverables:

1. A large suite of publicly available simulated datasets with known parameters for use by the metagenomics community to test scientific hypotheses. (Years 1-3)
2. Software for generating in silico simulations. (Year 2)
3. Mimic simulations of select metagenomic data sets (e.g., GOS, AMD). (Years 1-3)

REQUIREMENTS AND BUDGET

We provide here a summary of what we believe it will take in terms of personnel and supplies to carry out these tasks. Each of the tasks could be either expanded or contracted depending on the goals of the Moore Foundation but we believe this outline provides a good estimate of what it would take to get interesting and useful results in each area. First we provide a Table showing the allocation of personnel to the different projects and groups. Included in here are personnel dedicated to general support for the whole project who will be supervised by Dr. Eisen but will provide support for all.

Table 5. Personnel involved.

Projects	Green	Eisen	Pollard
I. Biodiversity			
A. Spatial variation	100% Post doc #1		
B. Novel types		50% Post doc #2	
C. Methods – Genes		50% Post doc #2 25% Software engineer	
D. Methods – Diversity	100% Post doc #3		
II. Population structure			
A. Spatial patterns		50% PhD #1	50% PhD #1
B. Genome evolution			100% Post doc #4
C. Methods – Binning	50% Post doc #5	50% Post doc #5	
D. Methods – Measuring evolution			100% Post doc #6
III. Functions			
A. Function vs. phylogeny	100% PhD #2		
B. Novel functions		50% PhD #3	
C. Methods – Functional diversity		50% PhD #3	
IV. Statistical estimators	50% Post doc #7		50% Post doc #7
V. Metagenomic methods		50% Post doc #8 25% Lab technician	
VI. Simulations		50% Post doc #8 25% Software engineer 75% Lab technician	
General	Summer Salary	Summer Salary 50% Software engineer 50% Project scientist #1 50% Project scientist #2	Summer Salary

Total personnel

- Post doctoral researchers 800%
- Lab technician 100%
- Software engineer 100%
- PhD students 300%
- Project scientists 100% (50% Martin Wu and 50% Dongying Wu)
- Summer salary x 3 PIs

Other requirements

- Computational resources
  - Linux cluster nodes: ~$30,000
  - Large memory machine: ~ $15,000
  - Disk space: ~ $5,000
  - Desktop computers x 10: ~ $20,000
  - Web server: ~ $5,000

    - Misc. equipment and supplies: ~ $5,000
    - Support services (e.g., cluster maintenance): ~ $5,000
- Sequencing expenses
    - Sequencing for in vitro simulations
        - Sanger runs: ~ $40,000
        - 454 runs: ~$30,000
        - ABI Solid/Illumina runs: ~ $20,000
    - Lab supplies and equipment: ~$5,000
- Travel: ~ $15,000
- Publication costs: ~ $10,000

COORDINATION AMONG GROUPS AND DISSEMINATION OF RESULTS AND METHODS

We realize that we are proposing an ambitious plan in the above project descriptions. However, we believe that this type of large scale, interdisciplinary effort is critically needed in metagenomics at this time. Each of the PIs project is committed to making this a true collaborative project and each has significant experience with large-scale collaborations involving multiple disciplines.

In terms of specific details, the overall coordination of the project will be carried out by Dr. Eisen who has coordinated on dozens of large scale collaborative genome projects and has been working on metagenomic related projects for many years. Dr. Eisen will supervise personnel to be shared between the projects and will coordinate the communications among the groups. To facilitate communications among the group we will have weekly video-conferences among the PIs and monthly group meetings where methods, research and technical issues will be discussed. In addition we will utilize the Eisen – Lab wiki site to share information about projects (the Eisen lab uses this site for all research activities in the group and to communicate with dozens of outside collaborators on various projects). Finally, we plan to have personnel rotate between groups to generate cross – fertilization of ideas and to aid in the training of these individuals.

We also believe that communication with other researchers is critical for this project. In addition to the "normal" publications and presentations we are committed to "Open Science" in terms of software, data, and publications. In addition, we plan to work closely with the CAMERA database for dissemination of results and methods.

Finally, we believe that the landscape of metagenomics is changing very rapidly and realize that there may be new projects or areas of interest to the Moore Foundation in our area. We are committed to doing research related to critical needs of the community and with our interdisciplinary team would be able to move into new projects quickly if so requested.

SUMMARY

In this proposal, we have outlined a collaborative interdisciplinary approach to the analysis of metagenomic sequence data. In particular we have designed projects that integrate statistical, mathematical, evolutionary and ecological approaches because we believe that these are critical for handling the complexities of metagenomic data. In total, we propose six projects – three focusing on specific scientific questions about microbes and three focusing on general methods for metagenomic studies. In all of the projects, one of the driving forces is the development of methods and tools to aid the community of scientists working with metagenomic data. Thus a critical part of our project is the dissemination of results and methods to the broader community. We are committed to making this dissemination as wide and as user friendly as possible both through publications and presentations but also through development of software and integration of results and methods into the CAMERA database. We are excited about the possibility of working on this project and working with the Moore Foundation to make metagenomic data a more valuable resource for the scientific community.

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. Eisen, J.A., *Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes.* PLoS Biol, 2007. **5**(3): p. e82.
12. Ricketts, T.H., et al., *Pinpointing and preventing imminent extinctions.* PNAS, 2005. **102**(51): p. 18497-18501.
13. Brooks, T.M., et al., *Global Biodiversity Conservation Priorities.* Science, 2006. **313**(5783): p. 58-61.
14. Lamoreux, J.F., et al., *Global tests of biodiversity concordance and the importance of endemism.* Nature, 2006. **440**(7081): p. 212.
15. Fenchel, T. and B.J. Finlay, *The ubiquity of small species: patterns of local and global diversity.* Bioscience, 2004. **54**: p. 777 - 784.
16. Finlay, B.J., *Global dispersal of free-living microbial eukaryote species.* Science, 2002. **296**(5570): p. 1061-1063.
17. Hillebrand, H., *On the generality of the latitudinal diversity gradient.* American Naturalist, 2004. **163**: p. 192-211.
18. Green, J.L. and B.J.M. Bohannan, *Spatial scaling of microbial biodiversity.* Trends in Ecology and Evolution, 2006. **21**: p. 501 - 507.
19. Pearson, H., *Microbe meeting promotes habitat conservation.* Nature, 2007. **447**(7141): p. 127.
20. Gross, L., *Untapped Bounty: Sampling the Seas to Survey Microbial Biodiversity.* PLoS Biology, 2007. **5**(3): p. e85.
21. Green, J.L., et al., *Spatial scaling of microbial eukaryote diversity.* Nature, 2004. **432**: p. 747-750.
22. Horner-Devine M.C., et al., *A comparison of taxon oc-occurrence patterns for macro- and microorganisms.* Ecology, 2007. **88**: p. 1345-1353.
23. Swenson N.G., et al., *The influence of spatial and size scale on phylogenetic relatedness in tropical forest communities.* Ecology, 2007. **88**: p. 1770-1790.
24. Olsen, G.J., et al., *Microbial ecology and evolution: a ribosomal RNA approach.* Annu Rev Microbiol, 1986. **40**: p. 337-65.
25. 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.

26. Hugenholtz, P., et al., *Novel division level bacterial diversity in a Yellowstone hot spring.* J Bacteriol, 1998. **180**(2): p. 366-76.

27. 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.

28. 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.

29. Giovannoni, S.J., et al., *Genetic diversity in Sargasso Sea bacterioplankton.* Nature, 1990. **345**(6270): p. 60-3.

30. 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.

31. Baker, B.J., et al., *Lineages of acidophilic archaea revealed by community genomic analysis.* Science, 2006. **314**(5807): p. 1933-5.

32. Chao, A., et al., *A new statistical approach for assessing similarity of species composition with incidence and abundance data.* Ecology Letters, 2005. **8**: p. 148-159.

33. 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.

34. Green, J.L. and J.B. Plotkin, *A statistical theory for sampling species abundances.* Ecology Letters, In press.

35. Perna, N.T., et al., *Genome sequence of enterohaemorrhagic Escherichia coli O157:H7.* Nature, 2001. **409**(6819): p. 529-33.

36. Tetz, V.V., *The pangenome concept: a unifying view of genetic information.* Med Sci Monit, 2005. **11**(7): p. HY24-9.

37. Lynch, M., *Streamlining and simplification of microbial genome architecture.* Annu Rev Microbiol, 2006. **60**: p. 327-49.

38. Lynch, M. and J.S. Conery, *The origins of genome complexity.* Science, 2003. **302**(5649): p. 1401-4.

39. Brown, D. and K. Sjolander, *Functional classification using phylogenomic inference.* PLoS Comput Biol, 2006. **2**(6): p. e77.

40. Consortium, T.G.O., *Gene Ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**: p. 25-29.

41. Wu, D., et al., *Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters.* PLoS Biol, 2006. **4**(6): p. e188.

42. Pollard, K.S., et al., *Forces shaping the fastest evolving regions in the human genome.* PLoS Genet, 2006. **2**(10): p. e168.

43. Pollard, K.S., et al., *An RNA gene expressed during cortical development evolved rapidly in humans.* Nature, 2006. **443**(7108): p. 167-72.

44. Pollard, K.S. and M.J. van der Laan, *Cluster Analysis of Genomic Data*, in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, et al., Editors. 2005, Springer p. 209-229.

45. Hooper, D.U., et al., *Effects of biodiversity on ecosystem functioning: a concensus of current knowledge.* Ecological Monographs, 2005. **75**: p. 3-35.

46. Kinzig, A., Pacala, S., Tilman, D, ed. *Functional Consequences of Biodiversity: Empirical Progress and Theoretical Extensions*. 2002, Princeton University Press: Princeton.

47. Cardinale, B.J., et al., *Effects of biodiversity on the functioning of trophic groups and ecosystems.* Nature, 2006. **443**(7114): p. 989.

48. Finlay, B.J., S.C. Maberly, and J.I. Cooper, *Microbial diversity and ecosystem function.* Oikos, 1997. **80**: p. 209-213.

49. Pollard, K.S. and M.J. van der Laan, *Statistical inference for simultaneous clustering of gene expression data.* Math Biosci, 2002. **176**(1): p. 99-121.

50. Yang, Y.H. and T. Speed, *Design issues for cDNA microarray experiments.* Nat Rev Genet, 2002. **3**(8): p. 579-88.

51. van der Laan, M.J. and K.S. Pollard, *A new algorithm for hybrid clustering with visualization and the bootstrap.* Journal of Statistical Planning and Inference, 2003. **117**: p. 275-303.

52. Salzberg, S.L., et al., *Serendipitous discovery of Wolbachia genomes in multiple Drosophila species.* Genome Biol, 2005. **6**(3): p. R23.

53. Giovannoni, S.J., et al., *Genome streamlining in a cosmopolitan oceanic bacterium.* Science, 2005. **309**(5738): p. 1242-5.

54. Eppley, J.M., et al., *Genetic exchange across a species boundary in the archaeal genus Ferroplasma.* Genetics, 2007.

55. Mavromatis, K., et al., *Use of simulated data sets to evaluate the fidelity of metagenomic processing methods.* Nat Methods, 2007. **4**(6): p. 495-500.