Computational Methods for Evaluating Microbial Diversity

by

David Alexander Wolfgang Soergel

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biophysics

and the Designated Emphasis

in

Computational and Genomic Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Steven Brenner, Chair
Professor Jillian Banfield
Professor Michael Eisen
Professor Michael Jordan

Fall 2010

Computational Methods for Evaluating Microbial Diversity

Abstract

Computational Methods for Evaluating Microbial Diversity

by

David Alexander Wolfgang Soergel

Doctor of Philosophy in Biophysics

Designated Emphasis in Genomic and Computational Biology

University of California, Berkeley

Professor Steven Brenner, Chair

The design and evaluation of methods for describing the diversity of microbial life in environmental samples is a critical step towards understanding life on earth and towards making prudent interventions in a wide variety of microbe-driven systems.

Microbes in the environment, including bacteria, archaea, viruses, and single-celled eukaryotes, are primary drivers of numerous geological and atmospheric processes, such as carbon fixation and sequestration, nutrient cycling, soil formation, and even cloud formation. Cyanobacteria in the surface of the ocean are estimated to be responsible for half of the primary production on earth. Microbes living in and on the human body are intimately involved in health and disease, even when they are not explicitly pathogenic; for instance, the gut is teeming with bacteria that are essential for digestion, but anomalies in this microbial community may contribute to disorders such as Crohn's disease. Environmental bacteria are critically important to climate change, agriculture, and public health, so understanding them has immediate practical importance, in addition to satisfying our scientific curiosity.

Environmental microbiology has long been limited by the fact that over 99% of bacteria found in the environment cannot yet be cultured, because the conditions required for growth have not yet been determined. In many cases, bacteria live in interdependent communities of species, making the growth conditions extremely complex and difficult to recreate, even if they could be determined. Thus, it is not possible to perform experiments on these organisms in the lab, or to acquire sufficient DNA to sequence their genomes in isolation. These limitations can be sidestepped through the use of culture-independent surveying techniques. With the availability of ever-cheaper DNA sequencing, methods that involve direct sequencing of DNA from environmental samples have now gained prominence, and are producing a deluge of data. However, the computational methods needed to make sense of these data are still in their infancy.

I evaluated methodological choices required for two kinds of culture-independent environmental sequencing techniques: taxonomic surveys using the 16S rRNA, and surveys of both taxonomy and function through shotgun sequencing. In both cases my goals were to increase the effectiveness of future studies in extracting biologically relevant information from environmental sequence datasets, and especially to head off misinterpretations of such datasets due to errors in methodology that have been overlooked to date.

Microbial community composition using the 16S ribosomal RNA sequence

PCR amplification and sequencing of the gene for the 16S ribosomal RNA subunit directly from environmental samples is a long-standing method of measuring species richness and relative abundance. I demonstrated that the use of sequencing reads that are much shorter than the gene itself (as has recently become economical and thus popular) has the potential to introduce substantial error in such studies. However, I also established, through exhaustive computational experiments, that a judicous choice of PCR and sequencing primers can avoid these errors. In particular, I found that the region following primer E517F provides the maximum available taxonomic information in diverse environments, and that sequencing more than 100nt provides little added value—a fact that justifies the use of next-generation sequencing technologies that are limited to a short read length. Notably, I obtained the same result both regarding supervised classification of sequences into known taxa and regarding unsupervised clustering of similar sequences into potentially unknown taxa. These are very different problems, so the congruence of results confirms that the region following E517F is indeed more informative than other regions.

Microbial species identification from environmental shotgun sequencing

The second culture-independent sequencing approach I addressed, known as "metagenomics" or "environmental genomics", does not target any specific gene but rather samples DNA sequences from the entire pool in an environment through shotgun sequencing. These data allow assessment of the range of metabolic functions present in a mixture of potentially many thousands of microbial species. A foundational problem in metagenomics is the assignment of sequences to known taxa, and the clustering of sequences into potentially unknown taxa. The surprising finding that sequence composition (i.e., statistical descriptions of the distribution of short words) can be discriminative of species identity has led to a wide range of proposed methods for both the supervised and the unsupervised variants of this "binning" problem, but the validation procedures applied to them have been both inconsistent and unrealistic. It has thus not been clear which method is best, or what performance can be expected in classifying real data. I reimplemented nearly all of the methods in the literature as special cases of a more general framework, allowing me to compare them on a common footing designed to mirror real circumstances.

Infrastructure for large-scale reproducible computational research

Each of the above projects relies on large-scale simulations, which require careful coordination of thousands of compute jobs and management of their inputs and outputs. This can be particularly daunting in the face of frequent updates to both input datasets and analysis programs, requiring recomputation of dependent results. To manage these computations, I developed Verdant (the "Versioned Data Analysis Tool"), a system for describing, sharing, and executing computational workflows on a cluster that guarantees reproducibility of results. It provides

a means of ensuring that a set of computational results are up-to-date with respect to the inputs and thus that they are internally consistent. It also provides a means of sharing inputs, intermediate results, and final outputs in a manner that facilitates collaboration while avoiding redundant computation.

For Ronan

Contents

1 Introduction **1**

1.1 Characterizing microbial community structure using the 16S ribosomal RNA sequence . . . 2

1.2 Shotgun sequencing of microbial communities . . . 4

Part I Methods for environmental diversity surveys using the 16S ribosomal sequence **11**

2 Selection of primers for optimal taxonomic classification of environmental 16S sequences **12**

2.1 Abstract . . . 12

2.2 Introduction and Background . . . 13

2.3 Results . . . 19

2.3.1 The confidence filter avoids spurious predictions . . . 19

2.3.2 Classification rate and precision vary widely among environments and primer/read length choices . . . 20

2.3.3 75nt reads from selected primers achieve near optimal classification performance . . . 20

2.3.4 Impact of the amplification primer . . . 24

2.3.5 Paired-end sequencing offers nearly no benefit over single-ended sequencing . . . 27

2.3.6 What read length is sufficient? . . . 27

2.4 Discussion . . . 31

2.5 Materials & Methods . . . 34

2.5.1 Choice of query datasets . . . 34

2.5.2 Preparation of reference databases 34
2.5.3 Note on primer nomenclature 35
2.5.4 Choice of primers 36
2.5.5 Simulation of sequencing reads 36
2.5.6 Classification procedure 39
2.5.7 Reconcilation of paired-end classifications 40
2.5.8 Precision vs Accuracy; “confident” predictions 40
2.5.9 Choice of representative optimal primers 41

3 Optimizing primer choice for OTU clustering of short-read environmental 16S sequences 42

3.1 Abstract 42
3.2 Introduction 42
3.3 Results 44
3.3.1 The response of clustering procedures to noise in the distance matrix . . 44
3.3.2 Choice of maximally informative primer and read length 51
3.4 Discussion 58
3.5 Materials & Methods 60
3.5.1 Construction of test datasets 60
3.5.2 Sampling of full-length sequence pairs 60
3.5.3 Extraction of sequence reads 61
3.5.4 SVM training 61
3.5.5 Choice of representative optimal primers 61

Part II Binning Methods 63

4 Supervised compositional binning methods are doomed on natural metagenomic samples 64

4.1 Abstract 64
4.2 Introduction 65
4.3 Results 66
4.3.1 Correct binning occurs only at very short phylogenetic distances 66

4.3.2 Fully-sequenced genomes dramatically undersample natural microbial communities . . . 77
4.4 Discussion . . . 82
4.5 Materials and Methods . . . 85
4.5.1 Placement of isolate genomes on a large 16S phylogenetic tree . . . 85
4.5.2 Computation of phylogenetic distances between sequences . . . 87
4.5.3 Compositional bias distances among genomes and genome fragments . 87
4.5.4 Environmental datasets and taxonomic classification . . . 88
4.5.5 Proportion of environmental sequences in the same genus as a fully-sequenced genome, by year . . . 88

5 A realistic and consistent methodology for comparison of metagenomic binning methods. 89
5.1 Abstract . . . 89
5.2 Results . . . 90
5.2.1 Choice of classification labels at different levels of the taxonomic hierarchy . . . 90
5.2.2 Separation of isolate genomes into test and training sets so as to produce realistic performance estimates . . . 91
5.2.3 Aggregating and sampling the training data . . . 96
5.2.4 Sampling the test data . . . 97
5.2.5 Summary of labelling choices . . . 98
5.2.6 A phylogenetic evaluation metric . . . 98
5.3 Discussion . . . 100

Part III Software 104

6 Verdant: a platform for computational research that guarantees reproducibility, internal consistency, and currency of results 105
6.1 Abstract . . . 105
6.2 Introduction . . . 105
6.3 Version control of input files . . . 106
6.4 Specification and computation of the workflow . . . 107

6.5 Standard programs and libraries 108
6.6 Collaboration and distributed workflows 108
6.7 Cluster computing 109
6.8 Continuous Integration 109
6.9 How this system guarantees reproducibility 109
6.10 TupleStreams and PlotBot 110
6.11 Conclusion 111
6.12 Related projects 112

Acknowledgments

I am grateful for the support and assistance of many people without whom this dissertation would not have been possible. I would like to thank my advisor, Dr. Steven Brenner, for providing an environment in which I could freely pursue a great variety of projects (however idiosyncratic some turned out to be), and for many years of incisive commentary on my work. My committee members, Dr. Jill Banfield, Dr. Michael Eisen, and Dr. Michael Jordan, also all provided helpful insights in committee meetings. In addition, Dr. Rob Knight was extremely helpful in the process of sorting out which of many ongoing projects to include in the dissertation, and provided extensive and insightful comments on the manuscript.

My wonderful labmates over the years all helped create a relaxed and collegial yet intellectually rigorous environment. My longtime officemate, Dr. Liana Lareau, was an invaluable source of entertainment and encouragement throughout. I am particularly grateful to Angela Brooks, Dr. Neelendu Dey, and Dr. Susanna Repo for their help with the submission process in the final weeks.

As an undergraduate researcher in the lab, David Tulga wrote a good deal of code in the early stages of the binning work, and was a valuable sounding board during that period. I am grateful to Robin Peters for all manner of administrative and editorial support, and particularly for navigating the university's byzantine funding apparatus on my behalf.

I appreciate the financial support of my research from the Howard Hughes Medical Institute, and from Leslie Chung and family via the Chang-Lin Tien Scholars program. My work also required a good deal of computer cluster time, for which I gratefully acknowledge Dr. Iddo Friedberg with the CAMERA project at UCSD and the Shared Research Computing Services pilot program managed by the University of California Office of the President.

Thousands of open-source programmers built software infrastructure and wrote libraries that are essential for my work and that of innumerable other researchers; anyone involved in computing owes them collectively a debt of gratitude.

Friends and family helped keep me sane through the travails of graduate school (though, of course, the responsibility for any remaining insanity rests solely with me). Special mention is due to Brent Eubanks and Dawn Pillsbury in this regard; their love and support has been a rock in our lives. My dear parents have always been a source of unwavering support and encouragement, for which I am deeply grateful. I also very much appreciate the advice and encouragement of Dr. Volker Soergel, Dr. Andrew Condey, and Drs. Rich and Judy Beery;

and I would like to acknowledge Dr. Irving Zucker for showing by example that the academic lifestyle may yet be tolerable.

None of this would have been remotely possible withot the love and support of my beautiful, funny, talented, industrious, and resilient wife, Dr. Annaliese Beery. I am forever grateful for her help, her patience, and her faith in me. Finally, I am grateful for and to our dear son Ronan, who (unbeknownst to him) tolerated a rather difficult year in excellent humor, and frequently provided much-needed reminders of life's great joys.

Chapter 1

Introduction

Microbes in the environment, including bacteria, archaea, viruses, and single-celled eukaryotes, are primary drivers of numerous geological and atmospheric processes, such as carbon fixation and sequestration (Bellamy et al. 2005; Monson et al. 2006), nutrient cycling (Arrigo 2005), soil formation (Oades 1993), and even cloud formation (Christner et al. 2008). Cyanobacteria in the surface of the ocean are estimated to be responsible for half of the primary production on earth (Bibby et al. 2003; Giovannoni and Stingl 2005). Environmental bacteria are critically important to climate change, agriculture, and public health; thus, understanding them has immediate practical importance, in addition to satisfying our scientific curiosity.

Environmental microbiology has long been limited by the fact that over 99% of bacteria found in the environment cannot yet be cultured, because the conditions required for growth have not yet been determined (Staley and Konopka 1985; Amann et al. 1995; Rappé and Giovannoni 2003; Riesenfeld et al. 2004). In many cases, bacteria live in interdependent communities of species (Bell et al. 2005; Tyson and Banfield 2005), making the growth conditions extremely complex and difficult to recreate, even if they could be determined. Thus, it is not possible to perform experiments on these organisms in the lab, or to acquire sufficient DNA to sequence their genomes in isolation. Indeed, it was difficult until recently even to estimate the species diversity and abundance distribution of these organisms.

1101 bacterial and 89 archaeal genomes have been fully sequenced to date by conventional large-insert or shotgun sequencing, but all of these genomes are of culturable strains. Incredibly, there are as many bacterial species in a single gram of soil (~3000-6000) as are listed in the entire NCBI species taxonomy (Daniel 2005; Gans et al. 2005; Tringe et al. 2005). Thus, the genetic diversity present in the environment is vastly undersampled, and the sample is biased as well—both due to intentional selection of "interesting" bacteria (such as human pathogens) for culturing and sequencing, and due to inherent biases in what can be cultured (Hugenholtz 2002). Furthermore, it is difficult to define "species" in the context of bacteria; a given clade consists of numerous strains that may be rapidly evolving and mixing genetic information through lateral gene transfer (Acinas et al. 2004a; Gevers et al. 2005; Ge et al. 2005; Konstantinidis and Tiedje 2005, 2007; Wilmes et al. 2009; Denef et al. 2010a).

For purposes of learning about microbes in the environment, the limitations of culturing can be sidestepped through the use of culture-independent surveying techniques. Various classes of such techniques have been proposed, including analyzing the diversity of phospholipids (Dowhan 1997), restriction fragment length polymorphisms (RFLPs) (Moyer et al. 1994), denaturing gradient gel electrophoresis (DGGE) (Muyzer and Smalla 1998), and automated ribosomal intergenic spacer analysis (ARISA) (Fisher and Triplett 1999; Popa et al. 2009; Kovacs et al. 2010). These methods study the diversity of biomarkers that are used as a proxy for species identity.

With the availability of ever-cheaper DNA sequencing, methods that involve direct sequencing of DNA from environmental samples have now gained prominence. The first such method is PCR amplification and sequencing of the gene for the 16S ribosomal RNA subunit; the resulting distribution of sequences can can be used to estimate the number and identity of microbial species present in an environmental sample with much greater precision and depth than the aforementioned methods (Olsen et al. 1986; Nocker et al. 2007). The second approach, commonly termed "metagenomics", is the study of DNA sequences uniformly sampled from an environment by shotgun sequencing (Handelsman 2004; Riesenfeld et al. 2004; Allen and Banfield 2005; DeLong 2005; Tringe and Rubin 2005).

In this dissertation I evaluate methodological choices required for both kinds of culture-independent environmental sequencing techniques, with the goals of increasing the effectiveness of future studies in extracting biologically relevant information from environmental sequence datasets, and especially of heading off misinterpretations of such datasets due to errors in methodology that have been overlooked to date.

1.1 Characterizing microbial community structure using the 16S ribosomal RNA sequence

The number and identity of microbial species present in an environmental sample can be estimated by PCR amplification and sequencing of the 16S ribosomal RNA subunit gene (Olsen et al. 1986; Britschgi and Giovannoni 1991; Curtis et al. 2002). This sequence is suitable for the task of identifying taxa because it must be present in all microbial cells; it is thought to be mostly vertically inherited (though this assumption has been called into question (Acinas et al. 2004b)); its function has not changed through evolution; and its overall mutation rate is fast enough (particularly in the hypervariable regions) to distinguish species and even strains from one another, but slow enough (particularly in the conserved regions) that sequence homology at much greater evolutionary distances is not obscured (Vandamme et al. 1996).

Also, the method is sensitive to contamination (Tanner et al. 1998), as well as to biases due to differing copy numbers of ribosomal RNA operons (Crosby and Criddle 2003; Acinas et al. 2004b). More profoundly, of course, the diversity of 16S ribosomal sequences present in a sample tells us little about the nature and distribution of the other genes. Nonetheless, interest in sequencing surveys of environmental microbes has exploded in recent years with the availability

of sequencing technologies that produce ever-larger data sets at ever-decreasing cost. As a result there has been an increasing need for evaluations of the methodological choices involved in performing these studies and for computational methods for interpreting the resulting data.

The microbial species definition. The definition of microbial "species" remains controversial. Methods of delineating species include DNA-DNA hybridization experiments; average nucleotide identity, especially among conserved housekeeping genes; laboratory characterization of metabolic functions; and division into "ecotypes" based on ecological niches. The correspondence between 16S sequences and all of these methods (and hence, the correspondence with traditional taxonomic names) is imperfect. Nonetheless, many studies adopt a functional definition of an "operational taxonomic unit" (OTU) as a group of organisms sharing 97% sequence identity among their 16S genes, corresponding roughly to the species rank in traditional taxonomies.

Measuring diversity of environmental microbes. Environmental diversity surveys may have different goals, each of which may be best served by different collection, sequencing, and analysis methods. Accordingly, there are three major distinctions that can be made among approaches to interpreting 16S sequence data sets that are common in the literature. The first is whether an environmental sample is analyzed with reference to a database of known sequences (a "supervised" method) or not ("unsupervised"). The second is whether the analysis is concerned only with distinguishing different types of microbes, usually at the species or genus level ("OTU-based" methods), or also with the phylogenetic relationships among these types ("tree-based" methods). The third is whether the unit of interest is the individual taxon or the community as a whole.

Common combinations include descriptions of microbial types present in a sample (supervised, OTU-based, taxon-centric) (Sogin et al. 2006; Sundquist et al. 2007; Wang et al. 2007; Huse et al. 2008; Liu et al. 2008; Wu et al. 2008; Hamp et al. 2009); species richness and evenness estimates (unsupervised, OTU-based, community-centric) (Schloss and Handelsman 2005); and the UniFrac beta diversity measure (unsupervised, tree-based, community-centric) (Lozupone and Knight 2005), though various other combinations arise regularly as well.

Sequencing technologies and the importance of primer choice. Sequencing of environmental 16S sequences requires that primers be chosen both for the PCR amplification step and for the sequencing reaction. Because the purpose of such studies is to measure sequence diversity, it is not obvious a priori that primers can be found that will amplify sequences from all microbial species. Fortunately the 16S sequence contains several highly conserved regions to which primers can be targeted. Nonetheless, even the highest-coverage primers are biased against some clades in which the targeted sequence contains slight variations (Baker and Cowan 2004).

In Chapters 2 and 3, I establish, based on exhaustive computational experiments, that environmental microbial diversity surveys based on short reads within the 16S rRNA sequence ought

to be done using primer E517F. Such surveys to date have been done with a wide variety of primers, chosen for reasons that are often not reported (e.g., because certain primers empirically amplify more DNA form certain samples than others). Systematic evaluations of the impacts of this choice (Baker and Cowan 2004; Liu et al. 2007, 2008; Frank et al. 2008; Hamp et al. 2009; Hong et al. 2009; Wang and Qian 2009) have not yet produced a community consensus about which primers to use under which circumstances. In particular, different primers provide sequences from different regions that evolve at different rates, and whose variation may be more or less correlated with the variation in the sequence as a whole. Thus, sequences from some primers are more informative about phylogenetic position than others. I found that reads from E517F provide the maximum available taxonomic information in diverse environments, and that sequencing more than 100nt provides little added value–a fact which justifies the use of Illumina sequencing for this problem compared with technologies that provide longer reads. Notably, I obtained the same result both regarding classification of sequences into known taxa and regarding clustering similar sequences into potentially unknown taxa. These are very different problems, so the congruence of results suggests that the region following E517F really is more informative than other regions.

1.2 Shotgun sequencing of microbial communities

The application of genome sequencing to microbial communities in recent years has produced an ever-increasing flood of "metagenomic" data, consisting of millions of reads sequenced directly from numerous environments. The collection and analysis of such data (Figure 1) is known as "metagenomics," because it concerns the study of genetic information pooled from multiple species. The goal of such projects is generally to answer questions about microbial ecology, including regarding the diversity of bacterial and archaeal (and less commonly, viral and eukaryotic) "species" in a community, the functional complement of proteins encoded, the evolutionary history of the represented populations (including both the "primary" phylogeny, to the extent that is meaningful, together with the distribution of lateral gene transfer events), and population genomics (i.e. regarding strain heterogeneity and recombination). However, for most environments, the generation of new sequence data continues to rapidly outpace our ability to interpret it. I therefore tackled the problem of developing and validating computational methods for analyzing metagenomic data.

In metagenomic sequencing projects, the species origin of each sequence read is unknown in advance; indeed, even the number of species and their abundance distribution is usually unknown. Thus, sequences from abundant species will appear more often than sequences from rare species; sequences that appear redundantly within the genome of some species will appear more often than those that do not; sequences conserved across taxa will appear more often than those that do not; and these three effects are confounded with one another. Sequencing depth is generally much less in metagenomic data sets than in isolate genome sequencing projects, where 8x average coverage is considered sufficient for good assembly. Indeed, after generating 100 Mbp of sequence from Minnesota farm soil, Tringe et al. (2005) found almost no overlaps

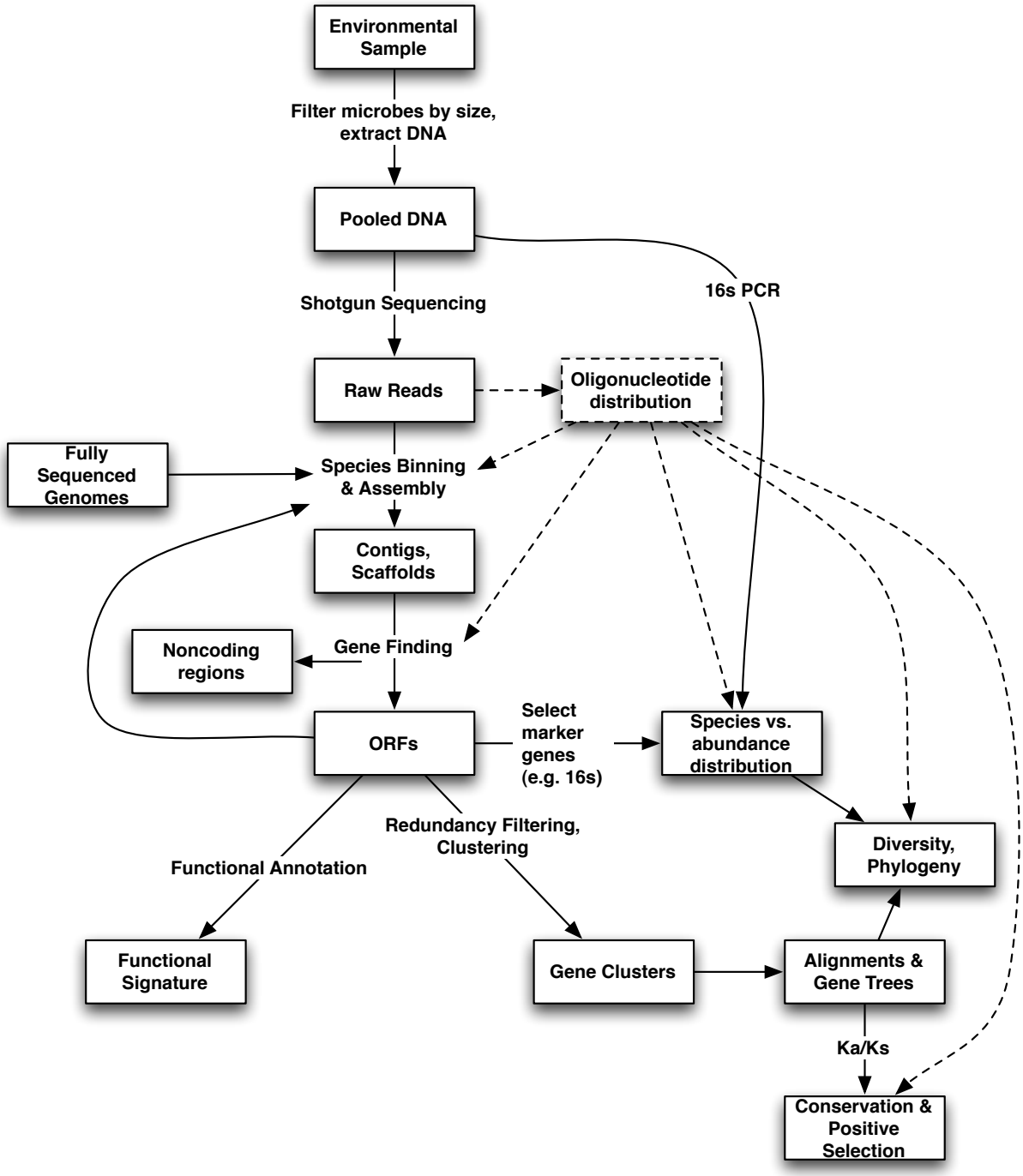

Figure 1.1: Common steps in the analysis of metagenomic data.

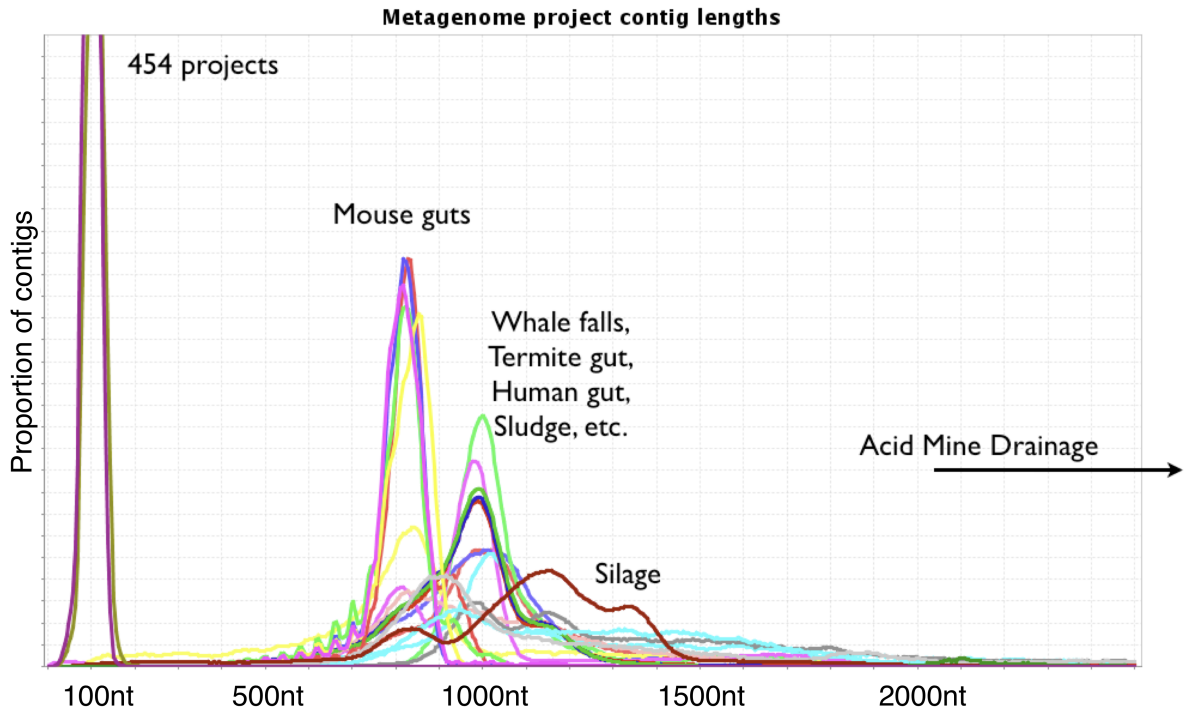

Figure 1.2: Contig length distributions from various metagenomic sequencing projects.

between reads, and estimated that 20-50 times more sequence would be required to assemble the genome of even the single most abundant species, to say nothing of the several thousand others. By contrast, Tyson et al. (2004) were able to assemble nearly complete genomes for two species and large genomic fragments for three others from a sample of biofilm found in acid mine drainage, a much less diverse environment, using about 75 Mbp of sequence. These early sequencing projects were performed using Sanger sequencing, producing reads of approximately 800 nucleotides each, sometimes from both ends of an insert. The recent and ongoing explosion in next-generation sequencing technologies from Roche/454, Illumina, and others is providing us with vastly greater quantities of sequence at a fraction of the cost of Sanger sequencing, albeit at shorter read lengths, currently in the range of 75 to 400 nt.

It may be that sequencing will soon become so cheap that we can expect to assemble full genomes from much more complex environments such as soil and seawater. However, assembly of metagenomic data is further complicated by the presence of substantial strain heterogeneity, recombination within a quasispecies, and lateral transfer of genetic material between species (Acinas et al. 2004b; Ge et al. 2005; Gevers et al. 2005; Vignuzzi et al. 2005). Sequence data from rare species will be sparse in any case.

The above concerns suggest that it will be useful to develop aggregate descriptions of metagenomic samples—that is, to learn how to characterize entire samples (and sets of samples) with respect to their diversity, evolutionary stability, complement of biological functions, and so forth, without first assembling single-species genomes from the shotgun sequence fragments.

A general approach to these problems is to consider the frequency distributions of sequence elements in the data. One kind of sequence element is the protein domain, the fundamental unit of protein structure and evolution. Protein domains can be identified in new sequences using probabilistic models, such as those in the Pfam collection (Bateman et al. 2004). Such models are typically tens to hundreds of nucleotides long, but accommodate substantial variability and thus match a range of related sequences. We took this approach in examining the distributions of Pfam hits in data from the Sorcerer II Global Ocean Survey (GOS) in collaboration with the J. Craig Venter Institute (Yooseph et al. 2007). Another kind of sequence element is the oligonucleotide (also known as a k-mer), which is simply an exact sequence of k nucleotides.

Classification using sequence compositional biases. It is well known that GC content varies across bacterial clades. Further, it has been shown that the frequency distribution of dinucleotides in the genome of a single species constitutes a "genome signature" that is unique to that species (at least, when making comparisons in an appropriate range of phylogenetic distances) (Karlin and Burge 1995; Karlin et al. 1998a). The dinucleotide signature is relatively invariant across a given genome (Campbell et al. 1999); thus, the distribution of dinucleotides in a 10-kb region is in many cases sufficient to identify the species of origin from a limited set of options (Nakashima et al. 1998; Abe et al. 2003). This discrimination becomes more precise and can be accomplished with sequences as short as 400 bases when tetranucleotides and perhaps larger oligonucleotides are used (Sandberg et al. 2001; Pride et al. 2003; Teeling et al. 2004b,a) (Figure 1.3). It may be that this startling combination of consistency within genomes and divergence between genomes can be explained in terms of codon bias, amino acid usage bias, biases in mutation and repair, and other known evolutionary phenomena, but this has not yet been established. The frequency distribution of k-mers in a genome, or more broadly any statistical description of its sequence composition, is known as the “genome signature”.

A basic question in any metagenomics project is "which species are present", or more generally, how the reads can be classified into phylogenetic groups at various ranks. Genome signatures can be exploited for this classification task despite our incomplete understanding of the underlying biological causes. A wide variety of "binning" methods have been proposed on this basis, but these have not been comprehensively compared to date. Indeed, since in general each method has been evaluated under different conditions and according to different criteria, it has not been meaningful to compare reported performance metrics such as sensitivity and specificity between papers in the literature. Thus it is a foundational problem in the field to establish a consistent evaluation methodology and to apply it to the whole range of binning methods, so as to make an informed decision about the "best" method to apply in a given context (perhaps depending both on features of the community and on the biological questions being asked).

Comparative evaluation of binning methods. The design of an informative evaluation methodology turns out to be surprisingly difficult for two reasons. First, there are very few real data sets that are well enough understood to form the basis of a benchmark. Evaluations are typically performed on simulated data for this reason, but it is not at all clear in turn how

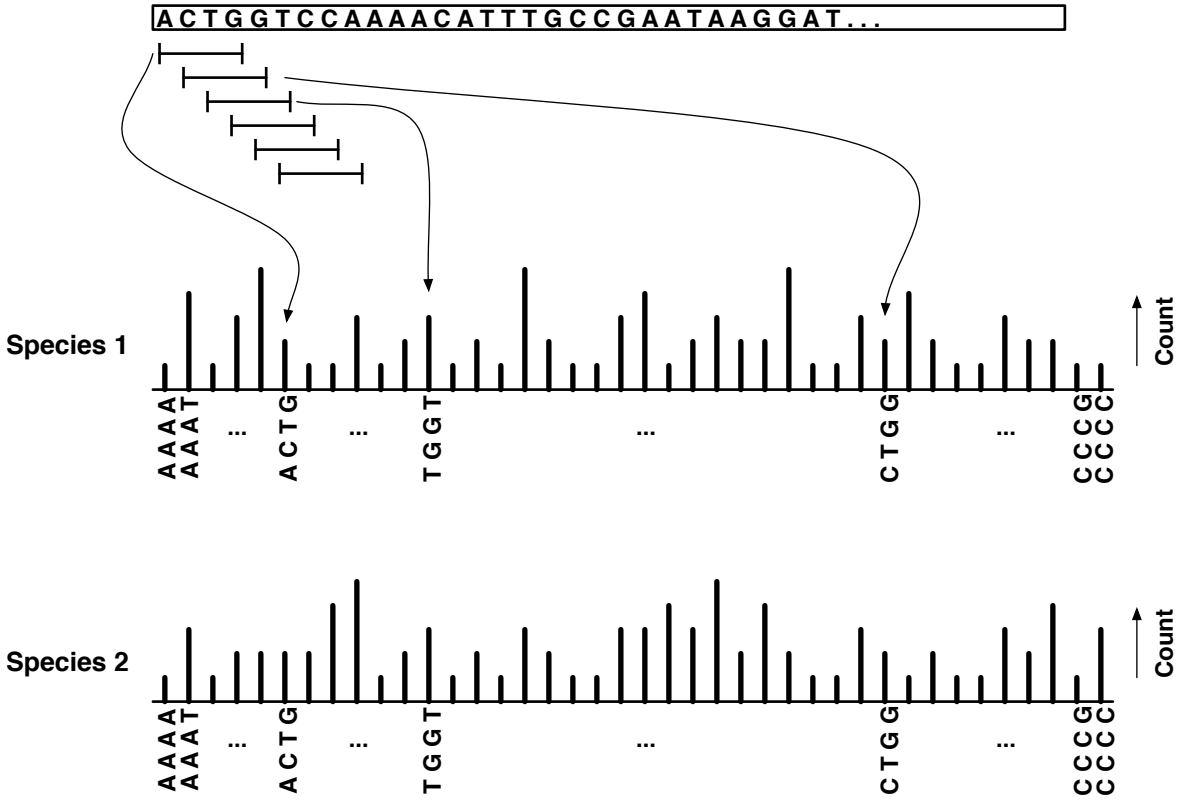

Figure 1.3: The genomes of different species contain distinctive distributions of 4-mers.

to simulate data with realistic properties. Indeed, I describe in Chapter 4 that validation procedures to date have made unrealistic assumptions about the composition of the communities that we wish to analyze, and as a result have produced substantially inflated estimates of binning accuracy. Correcting this overoptimism by designing a more realistic validation procedure is the topic of Chapter 5. Second, it is not obvious which metric should be optimized. For instance, measures commonly used to describe multiclass classifiers, such as class-normalized sensitivity and specificity, assume that the class labels are meaningfully chosen, mutually exclusive, and equally important. In the case of phylogenetic classification, the potential labels are hierarchically organized, and may have dramatically different weights associated with them (whether based on abundance, diversity, or importance by some other measure). Thus, Chapter 5 also proposes a measure for the quality of a phylogenetic classification.

There are very many parameters that can affect the accuracy of a binning process; these can be divided into three classes. The first set of parameters concerns the characteristics of the sampled community, over which we have no control, such as species richness, evenness, and phylogenetic distribution. The second set of parameters concerns the "wet" part of the sampling process, including the choice of sequencing technology (and the associated characteristic distribution of sequencing errors), any cloning or PCR biases, the read length, and the number of reads. These parameters clearly must be chosen early on in a project, and cannot be changed once the experiment has been done (or, can be "changed" only at great expense by starting over). The third set of parameters are the purely computational ones, including the choice of statistical model for compositional bias, the choice of a clustering method, the choice of a training set (in the case of supervised clustering), and so forth. These are the easiest parameters to change, in that the computations are typically fast enough that they can be repeated with different parameter sets at minimal expense.

I wished to estimate the classification accuracy that can be expected when different binning methods are applied to real environmental datasets. By varying all three classes of parameters—those describing the community, the sequencing process, and the classification procedure—I hope to choose the optimal binning method for a given community, i.e. to choose the third class of parameters as a function of the first two classes (which we can likely measure but not alter).

In order to evaluate the great diversity of possible binning methods on common footing, I first described them in terms of a general framework which includes nearly all of the published methods as special cases. I then implemented this framework with a mix-and-match plugin architecture, allowing me to test all of the published combinations of methodological choices, as well as a wide variety of combinations that have not yet been tested, with the goal of determining not only which set of parameters is best for given situation but also which of the parameters have the greatest impact on the results. This work is ongoing.

In addition to helping to answer basic questions about evolutionary, ecological, and biogeochemical processes, methods for analyzing environmental genomic data will be of increasing practical importance in public health (Breitbart et al. 2003; Duncan 2003; Eckburg et al. 2005; Vignuzzi et al. 2005), industry (Schloss and Handelsman 2003; Voget et al. 2003; Daniel 2005), agriculture (Yang et al. 2000a; Gur and Zamir 2004; Foley et al. 2005), bioremediation (Lovley 2003; Peacock et al. 2004; Sánchez et al. 2004; Brakstad and Lødeng 2005), and conservation

(Rauch and Bar-Yam 2004; Bell et al. 2005). The improvements I propose in the acquisition and analysis of data in both 16S surveys and environmental shotgun sequencing projects will increase the fidelity with which we can understand the composition and behavior of microbial communities in many environments.

Part I

Methods for environmental diversity surveys using the 16S ribosomal sequence

Chapter 2

Selection of primers for optimal taxonomic classification of environmental 16S sequences

2.1 Abstract

The composition of microbial communities has been studied for many years through sequencing of the 16S rRNA gene amplified from environmental samples. Recently there has been an explosion of interest in doing this using high-throughput short-read sequencing technologies. A common goal of such studies is to classify each observed sequence to a standard taxonomy, so as to enumerate the taxa present in the sample. There are dozens of "universal" primers that can be used for amplification and sequencing; these primer sequences are highly conserved and so are thought to amplify rRNA sequences from nearly all of the Bacterial species present, and sometimes the Archaeal species as well. However, different primers target different regions of the sequence, which may differ in the degree of taxonomic information they provide–for instance, because different regions evolve at different rates.

I developed an evaluation procedure that provides a realistic measure of the taxonomic precision that can be expected when classifying environmental sequence reads from a given primer. I then systematically tested thousands of combinations of amplification and sequencing primers and read lengths, simulating both single-ended and paired-end sequencing experiments. I thereby determined which regions of the 16S gene are most informative with respect to taxonomic classification. I found substantial variation in the information obtained from different primer and length choices, and observed that the most informative choice may differ depending on the environment being sequenced. Paired-end sequencing provides nearly no benefit in any environment or for any read length. For single-ended sequencing, an optimal choice of primer allows extraction of nearly all available taxonomic information using reads of only 75-90nt.

2.2 Introduction and Background

Measuring diversity of environmental microbes. Variation in the sequence of the 16S ribosomal subunit has been used since the mid-1980s (Olsen et al. 1986) to investigate the diversity of Bacteria and Archaea in many environments. Interest in sequencing surveys of environmental microbes has exploded in recent years with the availability of sequencing technologies that produce ever-larger data sets at ever-decreasing cost. As a result there has been an increasing need for evaluations of the methodological choices involved in performing these studies and for computational methods for interpreting the resulting data.

Variations in methodology. For many years, surveys were performed using Sanger sequencing, producing reads covering about half of the 16S sequence (~700-800nt) in many studies, and sometimes approaching the full length of the 16S sequence (~1500nt) if paired ends are used. Such studies provide a small number of reads–usually hundreds of sequences per sample, and at most few thousand. Next-generation sequencing technologies from 454, Illumina, and others produce shorter reads, currently in the 75-400nt range, but in vastly larger numbers (e.g., millions of sequences per sample) (Dethlefsen et al. 2008; Caporaso et al. 2010)– a compelling argument for diversity surveys (Tringe and Hugenholtz 2008). When performing an environmental survey, a choice must therefore be made about which region within the 16S sequence to target, and hence which amplification and sequencing primers to use, and whether to use paired-end sequencing. There is not yet a consensus on this topic; hence there are several dozen different primers that are commonly used in the literature (Baker et al. 2003; Wang and Qian 2009).

It is well known that the mutation rate varies widely within the 16S sequence (Van de Peer et al. 1994, 1996a; Cilia et al. 1996; Baker et al. 2003), largely driven by the structure of the RNA molecule, so that some regions are very highly conserved (allowing "universal" primer sequences to reside there) while others are hypervariable, such that different strains within a species have different sequences. The question arises, then, whether targeting different regions of the sequence may lead to different biological conclusions (Mills et al. 2006; Liu et al. 2007, 2008; Hamp et al. 2009; Youssef et al. 2009), and indeed whether the currently available short reads are sufficiently informative compared to near-full-length sequences.

One rather blatant example of this issue arises in many studies when sequences are clustered into OTUs of 97% sequence identity, intended to indicate the species level. This rule of thumb was established with respect to full-length 16S sequences, where it was found that the 97% identity threshold corresponds roughly with 70% DNA-DNA hybridization, a long-standing convention used to delineate species (Wayne et al. 1987; Vandamme et al. 1996; Hugenholtz et al. 1998; Gevers et al. 2005; Goris et al. 2007). Clearly, if a hypervariable region is sequenced, there will be far more 97% identical clusters than would have been found from full-length sequencing; and if a conserved region is sequenced, there will be fewer. Other methodological issues can have a similar impact, such as whether or not the Lane mask is applied to filter out hypervariable regions (Lane 1991; Desantis et al. 2006b); which alignment method is used (hypervariable regions naturally align poorly, so alignment variations essentially add noise to

the resulting identity scores)(Sun et al. 2009); which definition of percent identity is used (a question which is not at all as straightforward as it may seem); and whether the primer sequences themselves (which are of course guaranteed to be 100% identical for non-degenerate primers) are considered to be part of the read. There is nothing inherently wrong with using any arbitrary clustering threshold for any particular region, but it must be recognized that different studies may use wildly inconsistent standards to delineate OTUs, so the resulting species richness and other diversity measures are generally not comparable (Schloss 2010).

Robustness of results to such variations. In any case, whether short reads are "sufficiently informative" of course depends on the goal of each study. Liu et al. (2007) showed that 100-nt reads can be nearly as informative about beta diversity (as measured by UniFrac) as full-length sequences, provided that the primer is judiciously chosen; in particular, the authors recommend the use of primer E357R.

However, I (Chapter 3) and others (Engelbrektson et al. 2010; Schloss 2010) have shown that alpha diversity measures such as species richness and evenness can be quite sensitive to these issues.

Here, I address the extent to which supervised classification of environmental sequences (i.e. assignment of novel sequences to known taxa) can be reliable, given different choices of primers and read length.

Methods of classifying sequences to known taxa. Approaches to assigning environmental 16S sequences to known taxa naturally consist of two phases: first, finding matching sequences in a reference database, and second, using annotations on these reference sequences to infer the taxonomic identity of the query sequence.

For the search step, the most obvious solution is to use BLAST or FASTA, but these are computationally expensive due to the need to make alignments. An alternate method has been to find reference sequences containing a similar distribution of k-mers, usually words of 7 or 8 nt, which can be done much more rapidly (Chu et al. 2006; Sun et al. 2009). A recent hybrid solution is USEARCH, which makes alignments only of those sequences that are close according to a k-mer measure (Edgar 2010). In any case, an important parameter is the similarity threshold that is required: if our goal is to make genus-level annotations, then we ought to look for reference sequences within a percent identity to the query sequence (or equivalently, a proportion of matching k-mers) that corresponds to the genus level. However, it is not at all clear what that threshold should be, especially for fragmentary sequences. In practice, the choice of such thresholds has been largely arbitrary. Often the set of hits is further limited by choosing some number of the best ones (the "k-nearest-neighbor" or k-NN approach). Indeed, sometimes only the single best hit is used. This approach is especially dangerous if the similarity threshold is too permissive, since even the best hit may frequently be in a different taxon from the query sequence at the level of interest.

If all of the hits agree on the taxonomic annotation at some level, then it is usually simply transferred to the query sequence. When they disagree, as frequently happens even at higher

taxonomic levels, a voting procedure may be used (e.g., choosing an annotation if some proportion of the hits agree) (Sogin et al. 2006; Huse et al. 2007; Sundquist et al. 2007; Liu et al. 2008; Hamp et al. 2009). A more refined method of resolving such ambiguities is to consider the placement of the query sequence on a phylogenetic tree. The tree may be a previously computed one that relates all of the database sequences, as in the case of the ARB parsimony insertion tool (Ludwig et al. 2004) (a method that has been widely used, but that is applicable only to small numbers of sequences due to the substantal manual effort required). Some automated methods construct a local tree from each set of database hits (Wu et al. 2008).

Inadequate validation of classification methods: confidence measures. Confidence in the resulting taxonomy assignments should depend on confidence measures from each of the two phases. That is: if all of the hits are very similar to the query sequence and they all agree, then we will tend to have high confidence in the resulting classification; but lower confidence may result either from greater distances between the query and the hits, or from disagreement among the hits, or both. Published methods to date generally have not considered both sources of error, if they mention confidence at all.

For instance, choosing the top BLAST hit may provide confidence in the search phase, but only if the e-value threshold is stringent enough to strongly suggest that the hit is in fact in the same genus as the query. This approach does not consider the potential for disagreement among nearly-equivalent hits, which is all the more important when the search threshold is such that the wrong taxa may be matched.

Conversely, the RDP bootstrap confidence (Wang et al. 2007) is only a measure of agreement among annotations. It is applied to hits determined with a lax similarity threshold (because only subsets of the query k-mers are used, so many mismatches may be tolerated). The logic is essentially that, if our search threshold is permissive and the annotations agree anyway, then we should have high confidence in the outcome. This approach is particularly susceptible to database bias: the more permissive the search threshold, the more the set of hits will reflect such database bias, and the higher the apparent agreement will be among hits in overrepresented taxa. This confidence measure does not consider the likelihood that a hit with a certain k-mer score from a bootstrap subset is in fact in the same genus as the query in the first place.

Inadequate validation of classification methods: environments and database coverage. Because these methods are of the supervised variety, they are all subject to bias in the reference database. That is, all other things being equal, they are more likely to assign sequences to taxa that are highly represented in the database than to taxa that have rarely been seen before. Usually the reference database consists of a curated set drawn from all previously available 16S sequences, as provided by Silva (Pruesse et al. 2007), RDP (Cole et al. 2005), or GreenGenes (Desantis et al. 2006b). Thus the database favors taxa that have been most likely to be sequenced in the past. As of this writing, about one quarter of the ~500,000 GreenGenes sequences come from human skin; an additional quarter come from guts of humans and other mammals; about a quarter are not annotated as to origin– so less than one quarter are known to come from most of

the world's environments, such as oceans and soils. The databases are also likely to be biased in favor of taxa that can be cultivated. The finding in 2002 that only four divisions were well represented among isolate genomes Hugenholtz (2002) is still largely true today; indeed, only 24 of the ~100 known bacterial divisions have even a single isolate genome representative.[1] Efforts are being made to overcome this historical bias (GEBA, HMP), but the fact remains that we do not know how to culture most species of microbes. The cultivation bias has surely had an impact on the 16S databases as well, partly because the well-understood taxa are ultimately used to choose primers from which environmental samples are sequenced. At the very least, it is clear that model organisms are highly represented. In sum, while 500,000 sequences may seem like a large number, it seems likely that there are vastly more taxa that have not yet been observed. Indeed, novel bacterial and archaeal *divisions* are still being discovered at an alarming rate, often several in one sample (Chouari et al. 2005; Elshahed et al. 2008).

Especially in light of this potential for bias and the relatively sparse sampling of environments to date, it is important to thoroughly validate the predictions made by taxonomic classifiers. In particular, I am concerned that it is easy to "overreach", that is, to make a prediction that is more precise than is warranted. In the simplest case, some studies simply transfer the species annotation from the nearest BLAST hit for each query sequence, even if that hit differs by more than 3% from the query. Given the conventional correspondence of 97% sequence identity with the species level, the resulting species prediction is clearly wrong in this case (though the broader genus-level prediction may still be correct). Similarly, the GreenGenes and RDP web classifiers will commonly report a genus assignment for a sequence that is more than 5% different from any sequence in the database, even though it is obviously not a member of any known genus.

Inadequate validation of classification methods: consistency is not accuracy. A common validation procedure that can be quite misleading is to use unannotated sequences as the test queries, and to measure consistency between taxonomic assignments made from simulated short reads and taxonomic assignments made from full-length sequences from which the fragments were extracted(Huse et al. 2008; Liu et al. 2008). Agreement between these two predictions says nothing about whether the assignment is correct at all, especially not when a low sequence identity threshold is used in the database search (e.g. 75% in the case of (Liu et al. 2008)). Consider the case of a full-length sequence that is more than 5% different from any reference sequence. Even if an extracted fragment matches exactly the same set of database hits that the full-length sequence does, and even if those database hits agree on genus, the resulting genus prediction is clearly an artifact of database bias, not a legitimate assignment.

Inadequate validation of classification methods: the leave-one-out mistake. Authors of taxonomic classifiers have reported impressive estimates of precision and accuracy from their validations (Wang et al. 2007; Sundquist et al. 2007; Wu et al. 2008). However, these validations uniformly make a fatal mistake which inflates these results, which is that they are based

[1] http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html, http://greengenes.lbl.gov/cgi-bin/nph-browse.cgi

on "leave-one-out" experiments done at the sequence level. That is, they test prediction accuracy by extracting one sequence at a time from the reference database; classifying it using the remainder of the database; and comparing the predicted taxon with the "true" annotation on the query sequence. The mistake is that query sequences chosen randomly from the database reflect exactly the same biases that were present in the database to begin with, and so they are on average easier to classify than environmental sequences.

The reference databases are created by aggregating sequences from published studies, which in turn each typically report sequences from one sample or a few samples. Thus the database contains a few thousand sequences each from some samples, but more frequently a few hundred per sample, and very few sequences that were obtained independently from the others. Sequences within a given sample are more likely to have a close match within the same sample than to sequences from other samples (Acinas et al. 2004a); and obviously the same is true of environment types such as human gut. Thus, it is on average much easier to classify a sequence chosen from the database–which, despite the leave-one-out procedure, retains its sisters from the same sample and similar samples from the same environment–than it is to classify a sequence from a new sample (and especially a new environment) that is not yet in the database.

This issue comes into stark relief when one considers that natural environments contain many genuses, and even higher groups up to divisions, that are not yet represented in reference databases. The leave-one-out validations provide at best a measure of accuracy conditional on the knowledge that the query sequence comes from a known taxon. But environmental sequences do not meet that condition, and so the reported accuracy measure does not apply. In the case of the validation of the RDP classifier, the authors even imposed this condition intentionally:when a query sequence was the only representative of a genus in the database, such that leaving it out left no representatives, that data point was not counted in the accuracy measure (Wang et al. 2007).

Validations may be performed somewhat more fairly by leaving out an entire study at a time. In some cases, such as samples from human gut, there are enough sequences from the same environment left in the database that most of the sequences from the the held-out sample can be classified accurately. But in other cases, the held-out study is the only one of its kind, and consequently its sequences cannot be classified nearly as well.

I evaluated this issue using near-full-length sequences from eight large studies from diverse environments that are incorporated into the current GreenGenes. Figure 2.1shows the cumulative distribution of percent identity scores between query sequences from various studies and their closest match in the remainder of GreenGenes (holding out each study in turn). Nearly all sequences from the human gut sample are from species that have at least one representative in the rest of GreenGenes (i.e., they have a hit that is at least 97% identical over the full length), so we can expect human gut samples to be easy to classify. In the hypersaline mat sample, by contrast, fully 40% of the sequences have no match even at 95% identity, indicating that they represent novel genuses. Obviously, then, the RDP classifier's reported 88% accuracy at the genus level is not realistic in all cases. Even the ocean sample contains 10% novel genuses with respect to the remainder. Thus we should expect our ability to classify sequences to vary dramatically depending on the environment from which the sample is taken, and the degree to which that

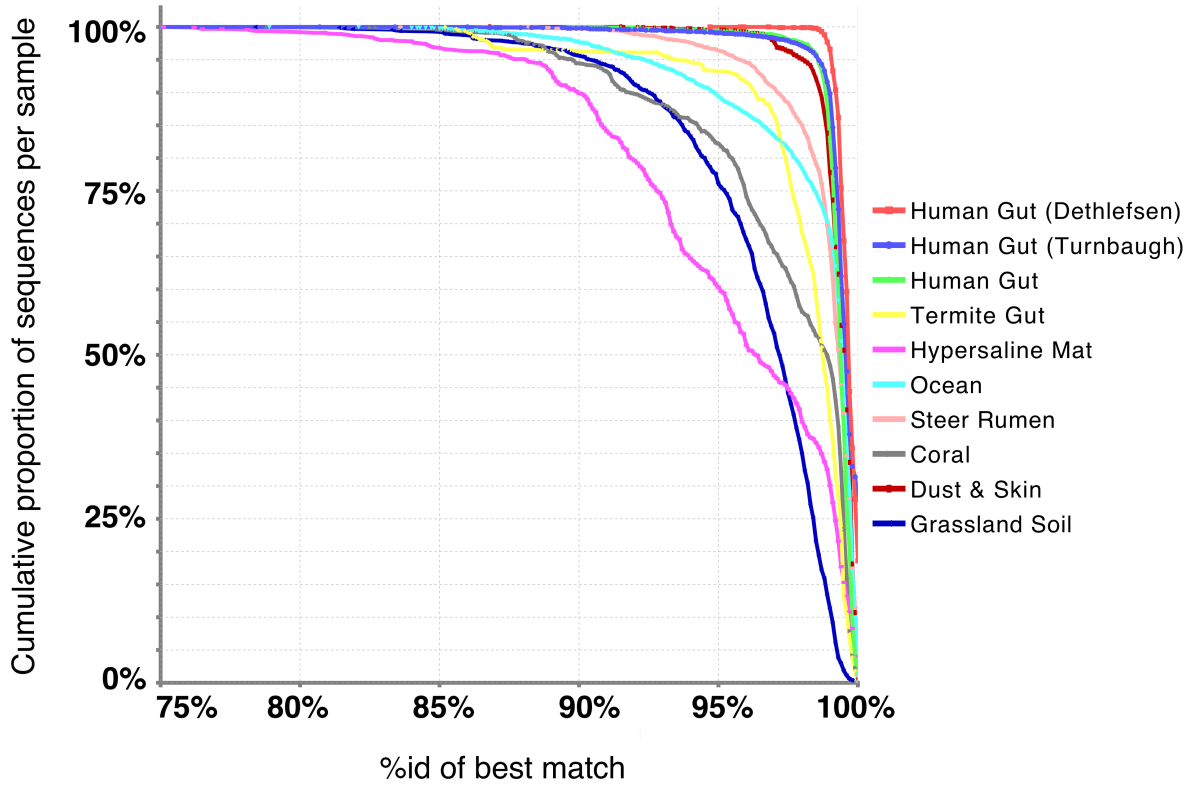

Figure 2.1: Coverage of different environments by GreenGenes. The plot shows the distribution of percent identity scores between environmental sequences and their closest matches in the remainder of GreenGenes (with each respective sample held out). The distributions are presented cumulatively from right to left, so that the Y value indicates the proportion of each sample that is within a given distance of any reference sequence. We see that GreenGenes provides excellent coverage of human gut and skin samples, but relatively poor coverage of the grassland soil and hypersaline mat samples.

environment has been previously characterized.

Open questions. In light of the above issues, I wished to reassess the precision and accuracy of taxonomic classification that can be expected, with respect to the environment being sequenced, the choice of primers, and the read length. In particular, I sought to determine which region within the 16S rRNA should be sequenced, given the read-length limitations of current technologies, so as to maximize the taxonomic information obtained. I also asked whether different environments call for different primers. Finally I asked whether paired-end sequencing can significantly improve taxonomic assignments compared to single reads.

2.3 Results

I performed a simulation study starting with environmental data sets of full length or near full length 16S sequences. I extracted reads of varying lengths at 44 universal primer locations commonly found in the literature, as well as pairs of these (simulating paired-end sequencing experiments). I searched for the nearest matches to each read in the GreenGenes database, in which many sequences have curated RDP taxonomic annotations. By looking for agreement among these annotations at each taxonomic rank, I then found a consensus taxonomic position for the query read.

The rank to which a prediction could be made varied from one read to the next, depending on the proximity of the query sequence to database sequences, on the ranks to which the database annotations described the hits at all (which is variable in GreenGenes), and on the depth to which those annotations agreed with one another. The classification procedure is similar to the GAST process (Sogin et al. 2006; Huse et al. 2008) and others (Sundquist et al. 2007; Liu et al. 2008; Hamp et al. 2009) (see Materials & Methods 2.5.6). Prior authors have reported the extent to which a classification can be made at all (i.e., precision), without regard for whether that classification is actually correct (accuracy). I applied a confidence filter (Section 2.5.8) so that I make classifications only to the level of the tree at which we can have a given level of confidence in the prediction (here, 80% or 95%) .

After classifying an entire environmental sample using a particular primer and length, I obtained the proportion of that sample that could be confidently classified to each taxonomic rank from domain through strain, as well as the proportions that could not be classified, either because the primer did not hit the query sequence at all, because the extracted read produced no close matches in the database, or because the database hits were insufficiently annotated.

The goal, then, was to choose a primer and read length for each environment that allowed the largest proportion of the sequences to be confidently classified to each level of the tree. To accomplish this, I simply computed these proportions for thousands of combinations of environments, primers, and read lengths by brute force.

I examined both single-ended and paired-end sequencing. PCR amplification is typically required prior to sequencing; thus, even in the single-ended case, a second primer is needed in the amplification step. This primer may limit the proportion of the original sample that can be sequenced and later classified, because sequences that do not contain it will not be amplified. I therefore tested each of the 44 sequencing primers in the context of all viable amplification partners, for a total of 794 combinations. In the paired-end case, I tested all 374 viable pairings of the 22 forward and 22 reverse primers.

2.3.1 The confidence filter avoids spurious predictions

I found that the classification procedure initially made many predictions which were later rejected by the confidence filter. For example, Figure 2.2 shows the proportions of classifications of ocean sequences made to each rank, using 44 primers and read lengths of 50, 75, 100, 125,

and 400nt. For single-ended sequencing, there are ~3500 viable combinations of PCR primer, sequencing primer, and length; for paired-end sequencing, there are ~1600 viable combinations of primer pair and length. All ~5100 single-ended and paired-end combinations are included in each plot, sorted along the X axis according to a rough measure of overall classification performance, so that the best combinations of primers and read length are to the left. The panels compare results using no confidence filter, an 80% confidence filter, and a 95% confidence filter.

The unfiltered classification provided strain- and species-level predictions for a few percent of the sample, and genus-level predictions for ~10-15% of the sample for many primer choices. At the 80% confidence level, all of the original strain- and species-level predictions, and many of the genus-level predictions, were deemed unreliable, though with a judicious choice of primers it is still possible to nearly match the unfiltered genus classification rates. When requiring 95% confidence, sequences could be confidently classified to the class level at best, regardless of primer and length choice. The overall proportion classified to the domain, phylum, and class levels does not change appreciably with these confidence thresholds (though those proportions do of course decrease at even higher confidence thresholds such as 99%; data not shown).

These results suggest that similar classification procedures that do not consider confidence will frequently overreach, making predictions at a lower taxonomic rank than is warranted.

2.3.2 Classification rate and precision vary widely among environments and primer/read length choices

The top two panels of figure 2.3 show the proportions of a human gut sample that could be classified to each taxonomic rank with 80% confidence and 95% confidence, respectively. The remaining panels (continuing in figure 2.4) follow similarly for samples from different environments.

Two conclusions can be drawn from these plots. First, within each environment, different primers and read lengths produce dramatically different classification performance; thus it is indeed important to choose a reasonably good one. Second, different environments can be classified to dramatically different levels. This is largely an effect of bias in the reference database, which remains dominated by only four bacterial divisions (Hugenholtz 2002). As suggested previously, samples that contain primarily sequences that are similar to known sequences are easier to classify than samples containing mostly novel sequences (Figure 2.1), a prediction that is borne out in these results (for instance, better classifications are obtained for sequences from human gut than from other environments).

2.3.3 75nt reads from selected primers achieve near optimal classification performance

The proportions of each sample that can be classified to the genus level with 80% confidence using each of the ~3500 single-ended primer/length combinations are available in the supplementary material (as are tables for other ranks and confidence levels). Table 2.1 presents selected

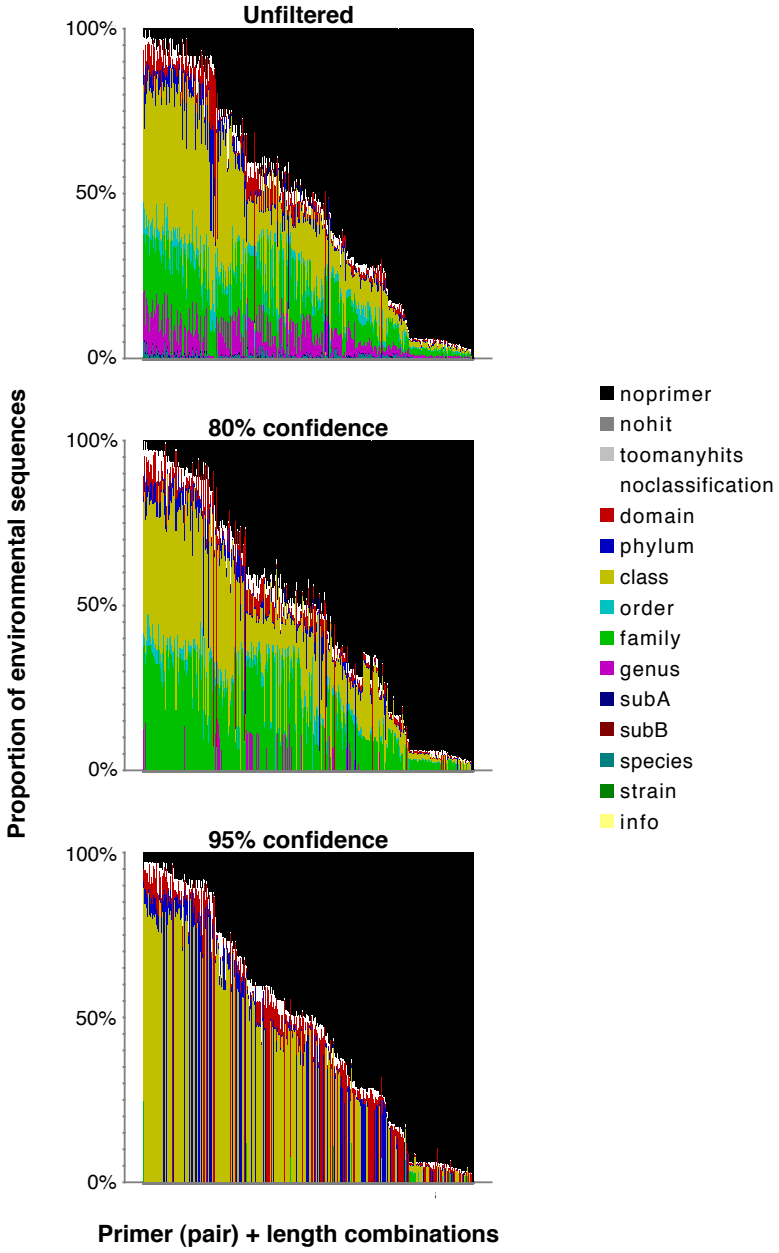

Figure 2.2: Impact of the confidence filter on classifications of an ocean sample. ~5100 possible choices of primers and read lengths are sorted on the X axis according to a rough measure of overall classification performance. The bar above each choice shows the proportion of the sample that can be classified to each taxonomic level. These proportions are stacked, so that the top of each colored section indicates how much of the sample can be classified to the given level or better. For instance, the red bars show that, for the best primers, ~5-10% of the sample can be classified *only* to the domain level, but that > 95% of the sample can be classified *at least* to the domain level.

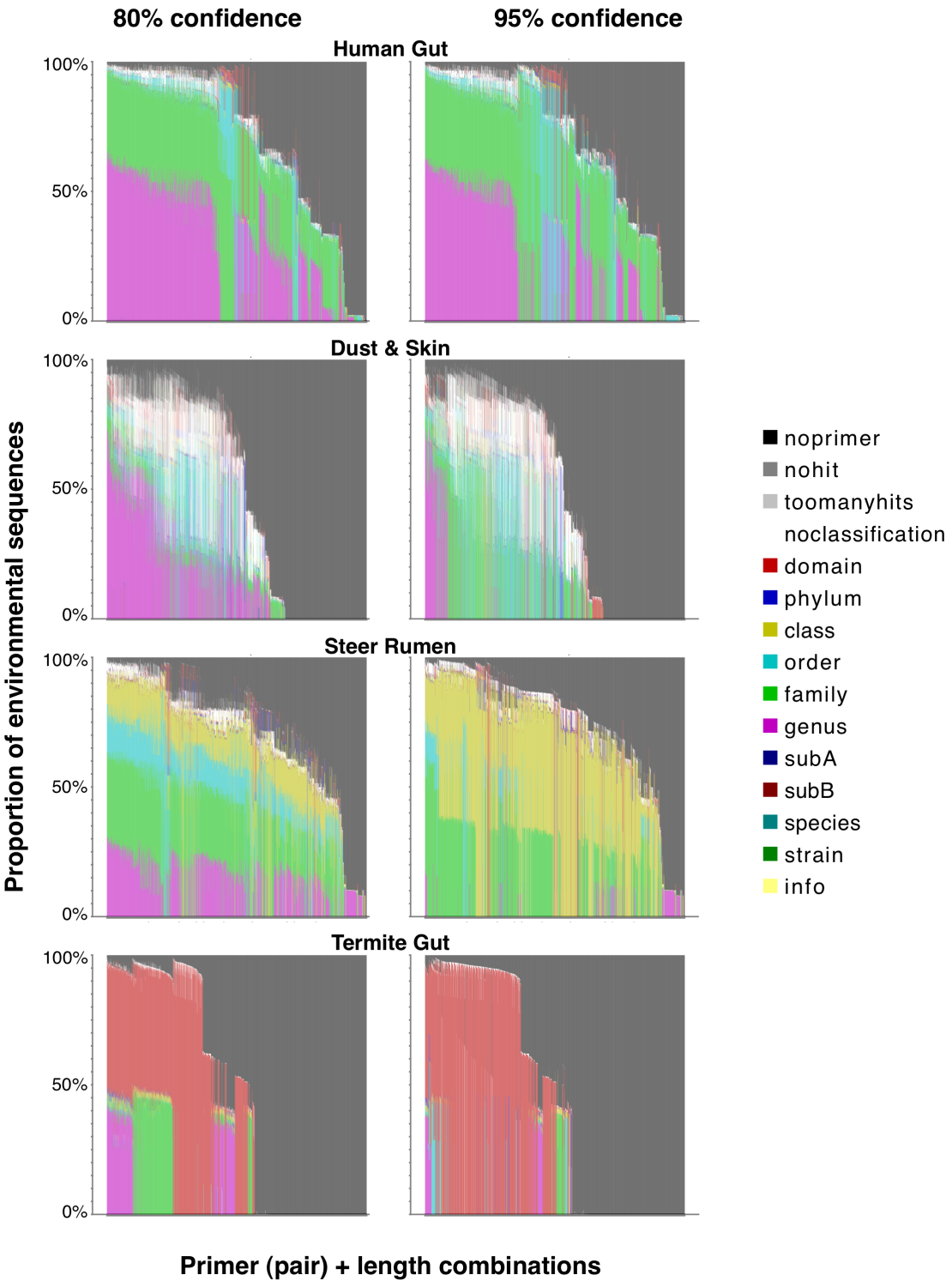

Figure 2.3: Classification performance of ~5100 possible choices of primers and read lengths for different environments, represented as in Figure 2.2.

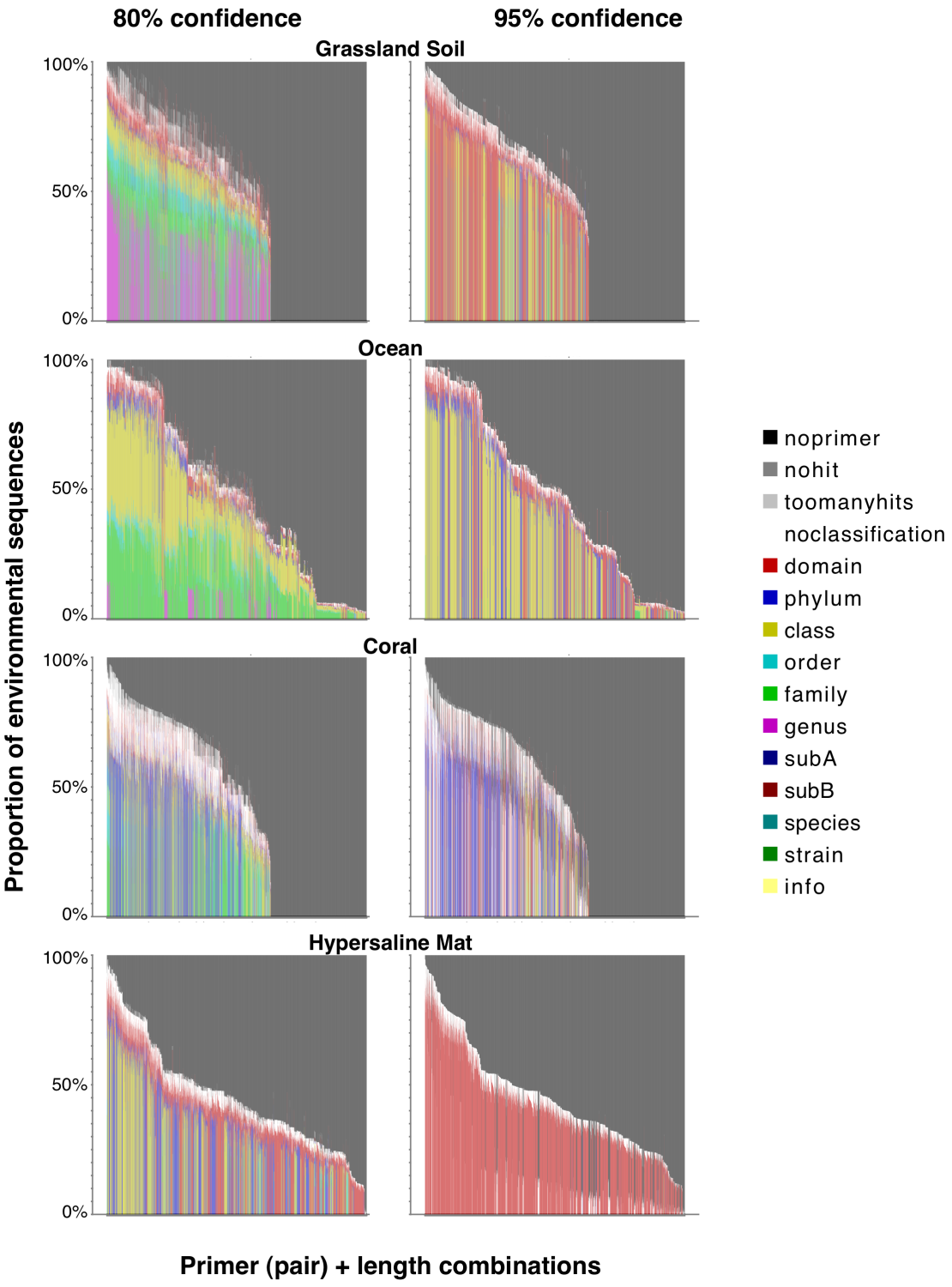

Figure 2.4: Classification performance of ~5100 possible choices of primers and read lengths for different environments, represented as in Figure 2.2.

combinations that are optimal for at least one environment per read length (see Materials & Methods 2.5.9). Table 2.2 presents optimal primers, similarly filtered, for making phylum-level predictions with 95% confidence.

From these tables I conclude that no one combination of primer and read length is best for all environments; for optimal performance, a different choice should be made depending on the environment being sequenced. I presume that this conclusion extends to other environments that I did not test. However, some primers perform well across a range of environments; for instance, E517F provides near-optimal performance in four or five out of the eight environments (depending on read length), both for genus-level and phylum-level classifications.

In most environments, sequencing 75nt from the best primer provides confident genus-level classifications within 5% of the maximum achievable from longer read lengths. For phylum-level predictions, the best classification rate in each environment is achieved with 50nt reads. I investigate the question of read length in more detail for two specific primers in section 2.3.6.

2.3.4 Impact of the amplification primer

The near full-length data sets from which I simulated short reads were originally produced using Sanger sequencing, typically using the highly conserved end primers E8F and E1406R, E1506R, or U1541R. Thus I was unable to take into account any organisms in the original samples whose 16S sequences did not contain these primers. I therefore initially assumed that using one of these primers as the PCR partner for each sequencing primer would produce the best available performance; using any other PCR primer can only limit the sequence pool further and thus reduce the classification rate. In most cases, this expectation was confirmed; that is why nearly every entry in the "amplification primer" column in tables 2.1 and 2.2 contains the word "end", indicating that an end primer is an optimal pairing for the respective sequencing primers.

However, in a few situations, using a more limiting PCR primer perversely improved classification performance. This can happen when the PCR primer preferentially excludes sequences that would have been misclassified anyway. This effect can cause the primer pair to pass the confidence filter, when the same sequencing primer paired with a more permissive PCR primer would have failed. Consider, for instance, classification of the grassland soil sample using

Table 2.1 (next page): Genus classification rates at 80% confidence for optimal choices of primers for amplification and single-ended sequencing. Thousands of combinations that produce suboptimal results are not shown (see Materials & Methods 2.5.9). Primers appearing together perform equivalently; i.e., for a given row, any choice among the given sequencing and amplification primers will produce the same result. "End" indicates an end primer such as E8F, E1406R, U1406R, E1407R, or E1506R. The highlighted cells indicate classification rates within 10% of the best achievable for each environment and read length.

			percentage of sample classified to genus level with >= 80% confidence							
read length	**sequencing primer**	**amplification primer**	**Human Gut**	**Dust & Skin**	**Steer Rumen**	**Termite Gut**	**Grassland Soil**	**Ocean**	**Coral**	**Hypersaline Mat**
	E1391F	end	32	70	0	36	42	0	0	0
	U529R E533R	end	46	56	16	0	51	0	0	0
	E969Fi	end	51	26	21	0	44	12	0	0
	E786F Eb787F	end E926Ra	52	15	17	40	30	0	0	0
50	E1492R	end	47	75	20	0	0	0	0	0
	E1492R	E969Fi	47	72	20	0	0	3	0	0
	E1064Ri	end E341F E517F U515F	45	25	20	0	49	0	0	0
	U515F E517F	end	37	0	0	42	34	0	0	0
	E1065R	end E517F U515F	2	0	10	35	43	11	0	0
	E1238R	end U515F E517F	0	8	0	35	11	0	0	16
	E517F	end E926Ra	58	60	19	41	51	0	0	0
	E1391F	end	60	58	23	42	0	0	0	0
	E786F Eb787F E805F	end E926Ra	54	20	21	41	35	11	0	0
	U519F	end E926Ra	40	60	25	0	48	0	0	0
75	E826R	end E341F E517F U515F	54	16	20	43	34	0	0	0
	E534R	end	39	53	25	0	50	0	0	0
	U515F	E1114R	24	56	17	0	43	3	0	0
	E926Ra	end E341F E517F U515F	28	0	19	42	43	8	0	0
	U529R E533R	end	58	55	25	0	0	0	0	0
	E1238R	E341F U341F E338F E1099F	1	13	0	37	26	0	0	19
	U515F / E517F	end / end E926Ra	61	61	26	42	0	0	0	0
	U519F	end	42	58	26	0	49	0	0	0
	E517F	E939R	55	17	18	41	37	7	0	0
	E926Ra	end E341F	56	23	23	0	53	13	0	0
	E1391F	end	60	65	0	40	0	0	0	0
100	E517F	E1114R	29	58	24	41	0	13	0	0
	E1391F	E1492R	60	64	23	0	0	0	0	0
	E917F	end E1407R	46	28	24	0	42	0	0	0
	E517F	E1065R	0	0	10	38	45	11	0	0
	E1238R	E786F E805F	4	15	16	31	20	0	17	0
	E338F	E1238R	2	39	6	0	32	0	0	21
	U515F	end	59	59	27	0	54	14	0	0
125	E1407R U1406R E1406R	end	58	72	24	0	49	0	0	0
	U515F E517F	end	60	61	27	0	0	14	0	0
400	E1492R	end	65	58	20	0	0	0	0	0
	U515F E517F / E926Ra	end E926Ra / end E341F	60	23	30	0	0	0	0	0

Table 2.1: Genus classification rates at 80% confidence for optimal choices of primers for amplification and single-ended sequencing. (Caption previous page)

			percentage of sample classified to phylum level with >= 95% confidence							
read length	**sequencing primer**	**amplification primer**	**Human Gut**	**Dust & Skin**	**Steer Rumen**	**Termite Gut**	**Grassland Soil**	**Ocean**	**Coral**	**Hypersaline Mat**
	E533Ra	end	89	70	95	0	87	87	71	0
	E805F	end	94	69	92	47	64	51	69	0
	U515F E517F	E1064Ri	96	66	96	45	0	88	78	0
	Eb787F	end	95	70	94	0	64	50	67	0
	U515F E517F	end	98	70	97	0	0	90	84	0
	U515F E517F	E826R	93	66	91	0	61	50	71	0
50	E1391F	end	99	86	94	0	0	89	63	0
	E926Ra	end	95	95	0	46	0	83	70	0
	E341F E343F	E926Rb U926R	92	73	81	45	0	0	50	0
	E338F E341F U341F E343F	E926Ra	92	74	91	46	0	0	0	0
	E1492R	end	97	94	76	0	0	0	0	0
	E338F E341F U341F	E1238R	29	56	57	39	0	43	0	30
	E926Ra	E517F U515F	86	68	94	47	84	85	71	0
	E517F	E1064Ri	96	65	95	45	78	87	56	0
	U515F	end	97	69	97	0	86	89	65	0
	E926Ra	end	87	71	95	0	86	88	73	0
75	E341F	end	92	64	94	0	85	86	75	0
	E355R	end	98	70	96	0	72	72	70	0
	E917F	end	98	71	94	0	51	84	65	0
	E1406R	end	97	89	94	0	0	91	69	0
	E1238R	E343F E341F U341F	34	57	59	40	46	59	51	30
	E338F	E1238R	34	50	60	0	46	58	57	30
	U515F E517F	end	98	67	93	46	0	82	0	0
100	E1406R	Eb787F E786F	93	82	91	0	60	0	56	0
	E1406R U1406R E1407R	E969Fi	98	83	94	40	0	0	0	0
125	no improvement									
400	no improvement									

Table 2.2: Phylum classification rates at 95% confidence for optimal choices of primers for amplification and single-ended sequencing. Results were filtered and represented as in table 2.1.

100nt reads from E517F. When amplified using an end primer or E926Ra, genus level predictions from these reads fail the 80% confidence test, so no predictions are made; but when paired with E939R, the predictions are deemed confident and 37% of the sample is classified. Another way to put this is that some primers are less universal, and conversely are more specific for taxa that we are able to identify. In such cases, a more universal primer only includes more novel taxa, thereby eroding our confidence that we can classify anything. This is probably also why the hypersaline mat sample yields no confident genus-level or even phylum-level predictions unless primer E1238R is involved.

2.3.5 Paired-end sequencing offers nearly no benefit over single-ended sequencing

For paired-end sequencing, predictions were made by a consensus of assignments to the two reads; however, when one read provided a more detailed classification than the other, it was accepted (Materials & Methods 2.5.7). I hoped thereby to obtain assignments that were more precise and accurate than were possible from one read alone. Tables 2.3 and 2.4 compare the maximum classification rates that can be achieved when optimal primers are chosen for single-ended or paired-end sequencing, when the goals are 80%-confident genus predictions and 95% confident phylum predictions, respectively. These demonstrate that the paired-end approach does not appreciably increase our ability to classify sequences. In each environment, the observed improvement in classification performance is a few percent at best, and likely does not justify the additional complexity and expense in most cases.

2.3.6 What read length is sufficient?

Longer reads are generally more difficult to obtain, more expensive, and less accurate (Schuster 2008); thus we would like to sequence the shortest reads that provide near-optimal classification performance. Because I concluded above that single-ended reads are sufficient, I explored the dependence of classification performance on read length in more detail for two of the best primers, E533R and E517F.

Figure 2.5 shows the dependence of 80%-confident classification performance on read length in the various environments. I found that genus classification performance using E517F on the human gut sample reaches a plateau at 80nt, and does not improve further even with 400nt reads. Using E533R, a plateau is reached at about 90nt, and scarcely improves again until 135nt. A similar plateau was reached for both primers in all environments, at read lengths varying from 55nt to 90nt. Only the Steer Rumen sample shows gains in genus classification past 90nt, and these are so modest and gradual that for most applications the cost of longer reads would not be justified. Classification rates at other taxonomic levels similarly approached their maximum values at read lengths of ~90nt, for both primers and in every environment. Finally, similar plateaus were reached in the vicinity of 90nt, though of course at a lower classification rate, for 95% confident predictions (data not shown).

		maximum percentage of sample classifiable to genus level with >= 80% confidence							
type	**read length**	**Human Gut**	**Dust & Skin**	**Steer Rumen**	**Termite Gut**	**Grassland Soil**	**Ocean**	**Coral**	**Hypersaline Mat**
single-ended	50	52	75	20	42	51	12	0	16
	75	60		25	43				19
	100	61		26			13	17	21
	125			27		54	14		
	400	65		30					
paired-end	50	56	78	22	41	49	13	0	20
	75	63		28	45	58		11	
	100			29		60	14		
	125						15	14	
	400	66		31					

Table 2.3: Paired-end sequencing offers little improvement in classification rate over single-ended sequencing for 80% confident genus classification. The classification rate shown in each cell is the maximum value observed for any choice of primers at each respective read length. Empty cells indicate no improvement over shorter reads.

		maximum percentage of sample classifiable to phylum level with >= 95% confidence							
type	**read length**	**Human Gut**	**Dust & Skin**	**Steer Rumen**	**Termite Gut**	**Grassland Soil**	**Ocean**	**Coral**	**Hypersaline Mat**
single-ended	50	99	95	97	47	87	90	84	30
	75	no improvement							
	100	no improvement							
	125	no improvement							
	400	no improvement							
paired-end	50	99	94	96	46	87	90	84	31
	75					89	91		33
	100	no improvement							
	125	no improvement							
	400	no improvement							

Table 2.4: Paired-end sequencing offers little improvement in classification rate over single-ended sequencing for 95% confident phylum classification. The classification rate shown in each cell is the maximum value observed for any choice of primers at each respective read length. Empty cells indicate no improvement over shorter reads.

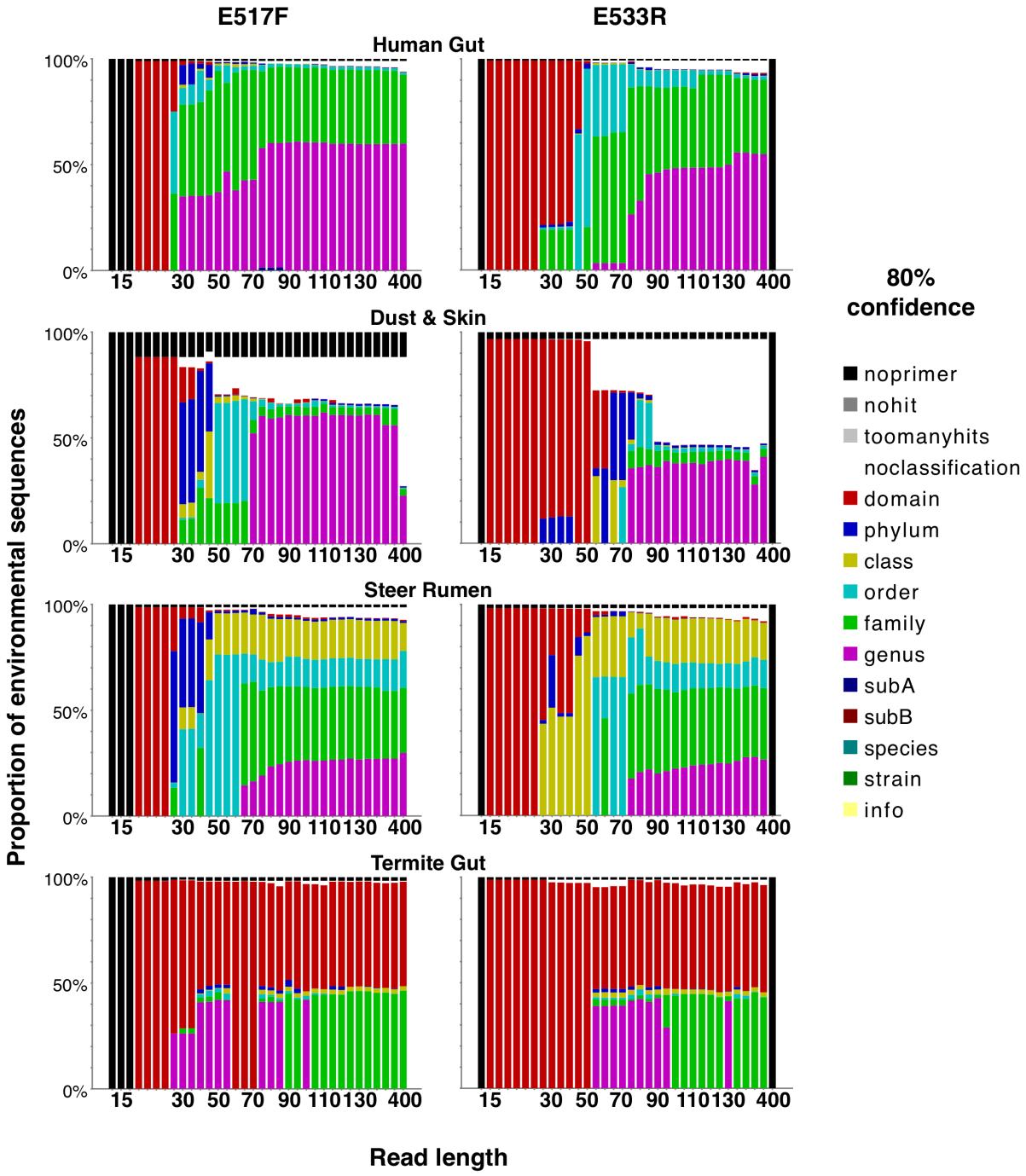

Figure 2.5: Classification performance resulting from different read lengths, starting from primers E517F and E533R, for various environments. Data are represented as in Figure 2.2.

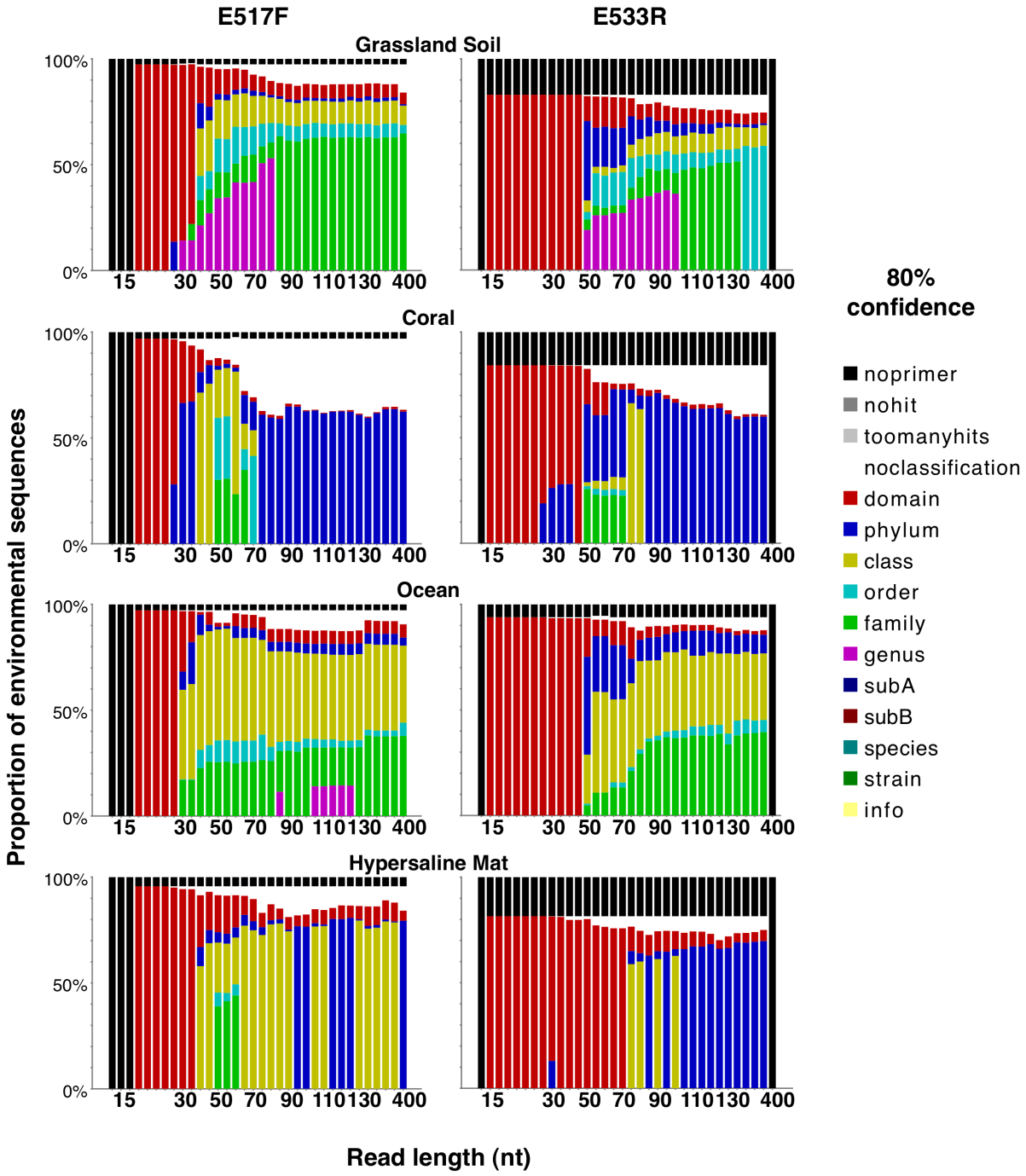

Figure 2.6: Classification performance resulting from different read lengths, starting from primers E517F and E533R, for various environments. Data are represented as in Figure 2.2.

In sum, I conclude that sequencing more than 90nt (including the primer) provides no benefit for taxonomic classification when using either E517F or E533R. Of course, I cannot conclude from this section alone that the same conclusion holds for other primers, because they target different regions of the sequence. However, we saw in section 2.3.3 (albeit in less detail) that reads longer than 75nt rarely offer large gains in classification for most of the optimal primers.

2.4 Discussion

Taxonomic classifications are less reliable than previously believed. I have shown that taxonomic classifications of short reads should be treated with substantial skepticism, especially when these classifications are made to the genus level. Claims of precise and accurate taxonomic assignments have been overstated to date, largely due to unrealistic assumptions made in the validation of classification procedures. Using a validation procedure designed to estimate real-world classification performance, I found that, in many environments, and for many choices of primer and read length, no genus-level predictions can be made with 80% confidence. Indeed, for some environments, the proportion that can be confidently classified to the genus level remains low even for the optimal primer choice.

Classifiers that do not adequately consider various sources of uncertainty in their predictions will commonly make genus-level assignments that are not supportable. This can occur due to poor database coverage of the environment in question, in which case the true genus of a query sequence is frequently not present in the reference database at all. It can also occur when the sequence region in question does not provide adequate taxonomic resolution to unambiguously identify the correct genus, even if it is present. In this circumstance, predictions are especially susceptible to database bias, resulting in classifications made preferentially to taxa that are abundant in the database.

Unfortunately, the widely used RDP and GreenGenes classifiers suffer from these difficulties– not only when applied to short reads, but potentially for near-full-length sequences as well. Consequently, I am concerned that the results of many hundreds of studies that employed these classifiers may be unreliable (depending, of course, on how the classification results were interpreted in each case).

The validation procedure I developed holds out an entire study at a time, simulating the situation that each study was not yet incorporated in the reference database and needed to be classified. Of course, now that the studies I evaluated are in GreenGenes, future samples of similar composition will be easier to classify than I report. This is especially relevant for studies that are the only representative in GreenGenes of a given environment type, such as the hypersaline mat sample. I would expect a second hypersaline mat sample to classify a good deal better than reported here, because matches can now be made to sequences from the first sample (assuming, of course, that those sequences now carry manually curated taxonomic annotations, including taxa that were not previously present in the database, and assuming that the second sample bears any resemblance to the first). But, many other environments remain as poorly represented now as hypersaline mat was previously. Thus my results are particularly sobering with respect to

whatever environment types are underrepresented in the database at the moment that it is used as a basis for classification.

The question of how samples become annotated upon incorporation into the reference database bears on an observation I made previously: that comparing fragment predictions to full-length predictions provides only a measure of consistency, not of accuracy. While some taxonomic annotations in reference databases are subject to careful manual curation, others may not have been so thoroughly treated. In cases where the annotations were largely the result of an automated classification procedure, they may not provide the robust ground truth that I assumed in performing my evaluations. In that case, my approach in fact also produces a measure of consistency between predictions from fragments and predictions from full-length sequence (with which database sequences were annotated), rather than a true accuracy. This potential for compounding errors would mean that my results are overly optimistic, so this concern should increase our skepticism about the accuracy of taxonomic classifiers still further.

Generalizability of results: environments. The environments I tested are diverse but by no means comprehensive. Thus I can hope, but cannot guarantee, that those primers that appear particularly informative in multiple environments here, such as E517F, E533R, E1391F, and others, would be effective in other samples as well. A related caveat is that the confidence filter I applied to the classifications from each sample was based on curated annotations on that sample itself; thus it is not possible to evaluate the confidence levels anew for a fresh sample. This is, of course, unavoidably how all validations of this type must work: we can predict the performance expected from a novel dataset only on the basis of that seen in prior datasets where the desired result is known. To the extent that the new dataset resembles the ones on which validations were performed (in my case, with respect to both the degree of database coverage as shown in figure 2.1 and the actual taxonomic composition), I assume that the conclusions from the validation process hold true for the actual experiment as well (i.e., regarding the confidence levels provided by different primers).

Generalizability of results: experimental design. The goal of this study was to identify the experimental parameters that allow the largest proportion of sequences from an environment to be confidently classified. Consequently, I have likely chosen primers that are present in taxa that are abundant in the tested environments. If the goal had been to find primers that identify the largest number of distinct genuses, including rare ones, the result may have been different.

Also, in the selection of optimal results to present in tables 2.1 and 2.2, I assumed that any pairing of primers and read lengths was equally viable. However, some sequencing technologies may impose constraints that I did not take into account, for instance preferring primer pairs separated by a given distance, or preferring longer primers to shorter ones. The complete tables available in the supplement can be searched for primer pair and read length choices that meet such criteria but that did not appear in my summary tables.

More generally, these results do not reveal which sequence region is most informative with respect to unsupervised clustering of a sample– a common goal of microbial diversity studies

that is quite different from the goal of taxonomic classification. I will address that question in Chapter 3.

Hypervariable regions need not be specifically targeted. No obvious pattern emerged from my analysis regarding whether hypervariable regions are particularly informative. This issue is hard to evaluate, however, because reads of 75nt or longer from any primer nearly inevitably include some hypervariable sequence. The reads found to be optimal in tables 2.1 and 2.2 frequently include small portions of V3, V4, or V9, but it is not clear how to determine whether this is significant.

Sequencing from E517F produces a read mostly between V3 and V4, but that does enter V4 at position 589, i.e., at position 72 of the read. In most environments there is no significant increase in classification as more of V4 is included (i.e., at read lengths of 75 and higher), though the human gut sample doesn't plateau until 85nt (including 13nt of V4). Overall this suggests that sequencing a hypervariable region is not essential for making near-optimal genus-level distinctions, as with 75nt reads from E517F.

For communities that are poorly represented in GreenGenes, increased read length actually decreased classification performance, because as the reads extend into more rapidly evolving regions, they accumulate ever more differences from their best matching reference sequence. For instance, in the case of grassland soil, genus classifications that can be made for reads up to 80 nt from E517F can no longer be made confidently for reads of 85 nt and higher. This may seem counterintuitive, but it does make sense due to my classification procedure and confidence filter. The increase in sequence variability beyond 72 nt means that, as the sequence length increases, the best database hits have an ever lower percent identity with the query sequence. Thus, database hits that differed from the query in the first 72 nt can be taken into account when they were not before. But these are more likely to have divergent genus annotations, resulting in the conclusion that we cannot make confident genus predictions. In the case of grassland soil, at least, I conclude that the region between V3 and V4 is more informative regarding genus identity than V4 itself.

Short single-ended reads provide near optimal classification. I found that, if appropriate primers are chosen, there is rarely a reason to sequence a single read longer than 90nt, and there is no benefit to paired-end sequencing. Indeed, assuming that the cost per base is roughly linear, one is often better off sequencing a single longer read than two shorter ones; for instance, table 2.3 shows that a single read of 100nt matches or slightly outperforms a pair of 50nt reads in seven of the eight environments tested.

I hope that these observations will allow future microbial diversity studies to be performed in the most informative and cost-effective way.

2.5 Materials & Methods

2.5.1 Choice of query datasets

I wished to test all primer and read length combinations using consistent sets of underlying sequences. I therefore sought data sets containing many near-full-length sequences from the same environment. I defined "near-full-length" as including hypervariable regions V1 through V9 (specifically, extending from positions 69 to 1465 in E. coli coordinates). I wished the datasets to be as large as possible in order to limit stochastic variation in the proportions of sampled taxa, and so that rare species would be represented.

I downloaded the GreenGenes database (version of August 25, 2010) and identified eight appropriate studies contained within it, representing a variety of environments, shown in table 2.5.

Environment	GreenGenes StudyID	Total sequences	Near-full-length sequences	Original citation
Human Gut	30418	7255	7234	(Li et al. 2008a)
Mattress Dust and Human Skin	36985	3294	3285	(Täubel et al. 2009)
Steer Rumen	35250	3369	2097	(Brulc et al. 2009)
Termite Gut	30924	1251	1167	(Warnecke et al. 2007)
Ocean	35248	6062	5222	(Shaw et al. 2008)
Coral	35251	1600	1520	(Sunagawa et al. 2009)
Hypersaline Mat	31588	1278	1174	(Isenbarger et al. 2008)
Grassland Soil	30925	1103	963	(Cruz-Martínez et al. 2009)

Table 2.5: Test datasets.

2.5.2 Preparation of reference databases

For each query data set, I built a reference database based on GreenGenes, excluding all sequences from the same study as the query sequences (whether near-full-length or not).

For the taxonomic identity of the reference sequences, I used the RDP taxonomy strings provided in GreenGenes, together with the species name from the "organism" field. I applied some

simple corrections to make these taxonomy strings more consistent in cases where different name variants clearly represented the the same taxon.

Each reference database was dereplicated at 99% using UCLUST 2.0.591 (Edgar 2010) such that for any cluster of sequences with 99% identity only the longest sequence was used. This reduced each database from approximately 500,000 sequences to approximately 140,000 representatives, thereby correcting for database bias at the strain level, and substantally improving performance of the downstream analyses.

The taxonomic identity of each reference cluster was usually unambiguous. For the occasional cluster containing sequences differing in taxonomic classification, I assigned taxonomic position at the deepest rank at which over half of the clustered sequences were in agreement.

2.5.3 Note on primer nomenclature

16S rRNA primers are conventionally named as follows. A one-letter prefix indicates the domain specificity of the primer (E = Eubacteria, A = Archaea, U = Universal). This is followed by the position of the primer, given in *E. coli* reference coordinates– that is, the position in the *E. coli* sequence homologous to the position in the sequence of interest where the primer is found. These positions may differ between *E. coli* and other species due to insertions and deletions, particularly at the beginning of the sequence. Finally, the suffix "F' or "R" indicates that the primer sequence is on the forward or the reverse strand, respectively.

Confusions commonly arise from differing interpretations of this naming scheme.

In my view, a primer should be named based on the starting position on the strand on which it occurs. That position is still given in forward coordinates even if the primer is on the reverse strand. For example, the commonly used primer name "U1046R" is a mistake, because the primer actually begins at position 1064 (and proceeds to the left). The confusion may arise because the primer ends at position 1046, or perhaps because a typographical error in an early paper has been propagated in the literature. Also, the reverse complement of U1064R is a forward primer, U1046F, so these may have become confounded. Even stranger, the primer beginning at position 8 is frequently called E27F instead of E8F. This primer ends at position 27, but since it is on the forward strand it is especially hard to see why it would be named based on the 3' end.

It is also worth noting that different primer sequences may start at the same position: they may have various levels of degeneracy (usually indicated by "N" characters in the sequence), or minor variations in the fixed characters (producing different clade specificity), or differing lengths. Thus a primer name does not unambiguously identify the primer; only the sequence itself can do that. I append suffixes (e.g. "a", "b") to distinguish such primer variants, but these are not conventional. Of course, it also commonly occurs that minor variants begin at different positions; for instance, E9F is essentially the same primer as E8F, shorter by one nucleotide at the 5' end.

The primer sequence should be given on the strand where it occurs. Occasionally I have seen a

primer described by the reverse complement of the actual sequence, which is clearly a mistake; so it is worth checking that provided sequences are on the expected strand.

Further bugs may arise because the conventional nomenclature makes the unfortunate choice of indicating positions in a 1-based system, and the interval is inclusive on both ends. For instance, in the sequence ABCDEF, the sequence BCD is at positions 2-4. That seems intuitive, but ends up being prone to off-by-one errors. For instance, the sequence has length 3, but 2 + 3 != 4. Also, if coordinates on the opposite strand are needed, one cannot simply subtract from the length of the sequence: D should be at position 3 on the reverse strand, but 6 - 4 = 2.

The long-standing solution to such problems in computer science is 0-based counting, inclusive on the left but exclusive on the right. This is easily visualized by the realization that what is counted is not the symbols themselves but rather the boundaries between the symbols. When beginning to read the sequence ABCDEF (i.e., when one is about to read "A"), one has consumed 0 symbols so far. Thus the sequence BCD is made up of positions 1, 2 and 3, so the end points should be specified as "1" and "4" (after one has read D, one has read 4 symbols total). Now 1 + 3 = 4, and the reverse coordinates are correct: 6 - 4 = 2 (the reverse position of D) and 6 - 1 = 5 (the reverse position after reading B). Sadly the 1-based system is now so entrenched in bioinformatics that it likely cannot be corrected.

2.5.4 Choice of primers

50 forward and 44 reverse primer sequences were obtained from from a survey of literature on primer choice. These were aligned to the 1541nt *E. coli* 16S sequence to confirm appropriate naming. The primers were also mapped to the 7682-column NAST coordinates by alignment to all GreenGenes sequences. In many cases, the primers began and ended in slightly different NAST columns (+- 1-5nt) in different sequences, suggesting that there are errors in the GreenGenes NAST alignment; I therefore report the column with the largest number of hits. Our initial survey included 94 primers, but many of these were specific to Archaea or for some other reason hit a small fraction of sequences in the environments we tested. Here, I selected the 22 forward and 22 reverse primers which hit at least 40% of the sequences in at least one of the query datasets (Tables 2.6 and 2.7).

For PCR or paired-end sequencing, the selected primers could be combined into 374 viable pairs for very short reads; as the read length increases, pairings spaced more closely than the read length become unviable.

2.5.5 Simulation of sequencing reads

Reads were extracted from the full length query sequences using the 374 viable primer pairs and read lengths of 50, 75, 100, 125, and 400, inclusive of primer length. Reads of 50-125 BP are common read lengths available now and in the near future using “next-generation” sequencing technologies such as Illumina; 400 BP sequences are available from Roche 454. In combination

Name	Sequence	Source	Length	E. coli 5'	E. coli 3'	NAST 5'	NAST 3'
E8Fa	AGAGTTTGATCCTGGCTCAG	(Wuyts et al. 2002) (Baker et al. 2003)	20	8	27	108	136
E8Fb	AGAGTTTGATCMTGGCTCAG	(Youssef et al. 2009)	20	8	27	108	136
E9F	GAGTTTGATCCTGGCTCAG	(Baker et al. 2003)	19	9	27	109	136
E334F	CCAGACTCCTACGGGAGGCAGC	(Baker et al. 2003)	22	334	355	1864	1897
E338F	ACTCCTACGGGAGGCAGC	(Youssef et al. 2009)	18	338	355	1868	1897
E341F	CCTACGGGNGGCNGCA	(Baker et al. 2003)	16	341	356	1872	1899
U341F	CCTACGGGRSGCAGCAG	(Baker et al. 2003)	17	341	357	1872	1901
E343F	TACGGRAGGCAGCAG	(Wuyts et al. 2002)	15	343	357	1875	1901
E349F	AGGCAGCAGTGGGGAAT	(Wuyts et al. 2002)	17	349	365	1886	1916
U515F	GTGCCAGCMGCCGCGGTAA	(Baker et al. 2003)	19	515	533	2227	2263
E517F	GCCAGCAGCCGCGGTAA	(Wuyts et al. 2002)	17	517	533	2231	2263
U519F	CAGCMGCCGCGGTAATWC	(Baker et al. 2003)	18	519	536	2233	2268
E786F	GATTAGATACCCTGGTAG	(Baker et al. 2003)	18	786	803	4050	4081
Eb787F	ATTAGATACCCTGGTA	(Baker et al. 2003)	16	787	802	4052	4079
E805F	GGATTAGATACCCTGGTAGTC	(Youssef et al. 2009)	17	805	821	4049	4088
E917F	GAATTGACGGGGRCCC	(Wuyts et al. 2002)	16	917	932	4542	4563
E967F	CAACGCGAAGAACCTTACC	(Youssef et al. 2009)	19	967	985	4624	4653
E969FIllumina	ACGCGARRAACCTTACC	Illumina	17	969	985	4626	4653
E1046F	AGGTGCTGCATGGCTGT	(Youssef et al. 2009)	16	1046	1061	4929	4955
U1053F	GCATGGCYGYCGTCAG	(Baker et al. 2003)	16	1053	1068	4940	4964
E1099F	GYAACGAGCGCAACCC	(Wuyts et al. 2002)	16	1099	1114	5012	5042
E1391F	TGTACACACCGCCCGTC	(Wuyts et al. 2002)	17	1391	1407	6427	6450

Table 2.6: Forward primers.

Name	Sequence	Source	Length	E. coli 5'	E. coli 3'	NAST 5'	NAST 3'
E65R	TCGACTTGCATGTRTTA	(Wuyts et al. 2002)	17	49	65	176	200
E355R	GCTGCCTCCCGTAGGAGT	(Youssef et al. 2009)	15	341	355	1868	1897
E357R	CTGCTGCCTYCCGTA	(Wuyts et al. 2002)	15	343	357	1875	1901
U529R	ACCGCGGCKGCTGGC	(Baker et al. 2003)	15	515	529	2231	2260
E533Ra	TNACCGNNNCTNCTGGCAC	(Baker et al. 2003)	19	515	533	2227	2263
E533Rb	TTACCGCGGCTGCTGGCAC	(Cho et al. 1996)	19	515	533	2227	2263
E534R	ATTACCGCGGCTGCTGGC	(Wuyts et al. 2002)	18	517	534	2231	2264
U534R	GWATTACCGCGGCKGCTG	(Baker et al. 2003)	18	517	534	2233	2268
E826R	GACTACCAGGGTATCTAATCC	(Youssef et al. 2009)	15	812	826	4049	4088
E926Ra	CCGNCNATTNNTTTNAGTTT	(Baker et al. 2003)	20	907	926	4521	4554
U926R	CCGTCAATTCCTTTRAGTTT	(Baker et al. 2003)	20	907	926	4521	4554
E926Rb	CCGTCAATTYYTTTRAGTTT	(Wuyts et al. 2002)	20	907	926	4521	4554
E939R	CTTGTGCGGGCCCCCGTCAATTC	(Baker et al. 2003)	23	917	939	4542	4580
E1064RIllumina	CGACARCCATGCASCACCT	Illumina	19	1046	1064	4929	4958
E1065R	ACAGCCATGCAGCACCT	(Youssef et al. 2009)	19	1047	1065	4929	4955
E1114R	GGGTTGCGCTCGTTRC	(Wuyts et al. 2002)	16	1099	1114	5012	5042
E1115R	AGGGTTGCGCTCGTTG	(Baker et al. 2003)	16	1100	1115	5013	5044
E1238R	GTAGCRCGTGTGTMGCCC	(Youssef et al. 2009)	18	1221	1238	5883	5910
U1406R	GACGGGCGGTGTGTRCA	(Baker et al. 2003)	17	1390	1406	6427	6450
E1406R	GACGGGCGGTGWGTRCA	(Youssef et al. 2009)	17	1390	1406	6427	6450
E1407R	GACGGGCGGTGTGTRC	(Wuyts et al. 2002)	16	1392	1407	6428	6450
E1492R	ACCTTGTTACGACTT	(Youssef et al. 2009)	15	1478	1492	6792	6809

Table 2.7: Reverse primers.

this produced 27,680 simulated datasets for the PCR-amplified single-ended case, and 12,920 datasets for the paired-end case.

Sequences which did not contain one of the requested primers at all were counted and included later in the fraction of sequences that could not be classified using that primer.

2.5.6 Classification procedure

Sequences from each of the 40,600 simulated environmental data sets were classified according to the following procedure.

For each read, the dereplicated reference database was searched using USEARCH 2.0.591 (Edgar 2010) with conservative parameters ("–allhits –maxaccepts 0 –maxrejects 128 –nowordcountreject") to locate all hits of at least 80% identity over the length of the read, using the USEARCH definition of identity.

This definition does not penalize insertions; that is, a query sequence that is identical to a target except for an insertion would receive an identity score of 100%. Insertions and deletions can be highly significant markers of taxonomic divergence, however, and furthermore they ought to be scored symmetrically in this context . Identity scores were therefore corrected, using the alignment information provided by USEARCH, to a definition in which insertions and deletions are simply counted as mismatches.

From the set of database hits thus obtained, a set of the best hits was selected by choosing those within 0.5% of the maximum %id observed. For read lengths < 200, this meant effectively that those reads were chosen that were tied for the maximum %id score. For 200nt reads, one extra mismatch was allowed (in addition to the number of mismatches between the query and the best hit), and for 400nt reads, two additional mismatches were accepted. The purpose here was to skim off a selection of the best available hits, while excluding more distant matches.

A consensus taxonomic position was then obtained from the taxonomic annotations of the selected top hits using a simple hierarchical voting procedure: at each taxonomic rank, a name was accepted if it was shared by at least half of the hits. In addition, the winning taxon was required to have at least twice as many hits as the runner up; otherwise a tie was declared, and no classification was made at that level. The resulting classification thus extended as far down the tree as there was at least 50% agreement among the database clusters and not too many dissenters. Database clusters that had no annotation were counted in the denominator; thus a query sequence whose best hits were more than 50% unannotated could not be classified even to the domain level.

I wished to avoid "overreaching", i.e., predicting taxonomic position past the point of correctness. This was controlled by voting threshold parameters: if I insisted on near-perfect agreement among all the hits, then the procedure would generally classify very shallowly, because a few hits anomalously disagree or simply have shallow annotations. Overly permissive thresholds, conversely, would produce predictions that are very precise, but wrong. I found, from a very coarse sampling of different thresholds, that only extremely stringent or extremely permissive

thresholds had a noticeable impact on my final results; in a large middle range, my results were not sensitive to the voting threshold. This is why I chose the simple majority-vote rule above.

Query sequences that hit more than 15,000 reference clusters were marked as overly generic and were not classified. Thus, the procedure does not work for reads that are very short (and thus nonspecific with respect to taxonomy), or perhaps for longer reads that are highly conserved in a phylum that is highly represented in the database.

In sum my approach is something like a *k*-nearest-neighbor classifier, except that different numbers of neighbors are used for each query sequence depending on how many database hits are (nearly) tied for the best %id score.

2.5.7 Reconcilation of paired-end classifications

In the case of paired ends, I found that requiring database hits to match both reads from a query sequence produced too few hits.

For this reason, the above procedure was first applied to each of the two reads independently. I then reconciled the annotations obtained from the two reads as follows. Starting from the root of the tree and walking down rank by rank, a classification was accepted if the two reads agreed. A classification was also accepted if it was present on one read, but the annotations on the other read did not extend to that depth (e.g., if one read had annotations only to the order level, then the annotations from the other read would be accepted to the species level, if present, provided that they agreed up to the order point). If either of the primers did not hit the sequence, of course, it was considered unclassifiable; similarly, if either of the sequenced reads had no database hits at all, then the pair as a whole was considered to have no hits.

2.5.8 Precision vs Accuracy; "confident" predictions

Prior authors have reported the extent to which a classification can be made at all (i.e., precision), without regard for whether that classification is actually correct (accuracy). An obviously problematic case is one in which all of the database hits to a sequence agree on genus, but these hits are more than 5% divergent over their full length from the query sequence, indicating that the query sequence is in fact not a member of that genus. It is not straightforward to limit the taxonomic level of the predictions on the basis of the observed %id of a sequence fragment, however, because the identity threshold associated with each level is variable throughout the sequence, and different fragments would give inconsistent results.

I therefore used annotations on the query sequences, where available, to evaluate the accuracy of the taxonomic predictions at each level. Within each query dataset, for each primer and read length, and for each taxonomic level, I computed the proportion of predictions that proved correct, out of those sequences that were annotated at that level at all. I then applied two confidence thresholds, 80% and 95%, to determine which primer/read length combinations produce trustworthy classifications under the given circumstances. I then removed all predictions that were

deemed unreliable; for instance, a genus-level prediction for some sequence might be truncated to the order level, because more detailed predictions were found to be wrong more than 20% of the time for the given primer, read length, and environment.

2.5.9 Choice of representative optimal primers

I exhaustively computed the classification rate for thousands of combinations of primer, read length, experiment type, environment, taxonomic level, and confidence level. I found that some choices of primer and read length provided more classifications (at a given confidence level) than certain other choices across all environments tested. My results suggest, for example, that one should not use primer E357R with 100nt reads for taxonomic classification, because primer E517F with 75nt reads is always at least as informative at the genus level (and usually much more so). In fact, for phylum level classifications, reads of only 50nt from E517F are substantially more informative. I filtered the results tables to exclude choices of primer and read length that were uniformly less informative than others of the same or shorter read length. I further filtered them, for the sake of tractable presentation, to include only primers that achieve at least 90% of the optimum classification rate (per read length) in at least one environment.

In a few cases, several choices provided nearly equivalent classification performance, particularly involving closely related primers such as E517F and U515F. I considered two choices to be equivalent (for a given taxonomic level and confidence level) if they provided classification rates within one percentage point in all environments. In these cases I list each alternative but report the classification performance of one arbitrarily chosen representative.

The entries that remained after this filter was applied highlight the trade-offs inherent in the choice of primer and read length. Each remaining entry is optimal according to some criterion. For instance, for genus predictions from 75 nt reads, E533R classifies more of the steer rumen sample than does E517F, but E517F is able to classify sequences from the termite gut sample, where E533R makes no confident predictions; and neither of them can classify any of the hypersaline mat sample, where only E1238R produces confident predictions. Note that other primers may also allow classification of some of the hypersaline mat sample, but they are not mentioned in the table because E1238R always outperforms them (in all environments).

Chapter 3

Optimizing primer choice for OTU clustering of short-read environmental 16S sequences

3.1 Abstract

Variation in the sequence of the 16S ribosomal RNA provides a means of assessing the phylogenetic diversity and structure of microbial communities through amplification and sequencing from environmental samples. The basic approach has been known since the mid-1980s, but the relatively recent development of low-cost, high throughput sequencing has produced a dramatic explosion of interest in the field. Environmental sequences are typically clustered into "operational taxonomic units"(OTUs), because highly similar sequences are likely to have originated from the same species. However, the fidelity of this clustering is compromised when the available sequences are short, as is the case with current sequencing technologies. Here, I assess how the choices of amplification and sequencing primers and of read length impact the clustering results, and conclude that, for the currently typical read lengths, the most accurate clusterings are achieved using reads sequenced from primer E517F.

3.2 Introduction

The application of next-generation sequencing technologies to survey the diversity of 16S ribosomal RNA sequences in environmental samples is revolutionizing the field of microbial ecology (Tringe and Hugenholtz 2008). Where the Sanger method was previously used to obtain hundreds of half- to full-length 16S sequences per sample, modern methods available from Roche/454, Illumina, and others can provide millions of sequences per sample, albeit at short read lengths of 75 to 400 nucleotides. This approach has already yielded fascinating insights

into the ecology of many environments (Caporaso et al. 2010), including the human gut (Andersson et al. 2008; Dethlefsen et al. 2008; Turnbaugh et al. 2010) and other body habitats (Sundquist et al. 2007; Fierer et al. 2008; Grice et al. 2009; Lazarevic et al. 2009; Nasidze et al. 2009), soils (Chu et al. 2010), and oceans (Huse et al. 2008; Galand et al. 2009).

One of the first steps in analyzing an environmental sequence data set is to cluster similar sequences together, typically using a program such as FastGroupII (Yu et al. 2006), CD-HIT (Li and Godzik 2006; Li et al. 2008b), or UCLUST (Edgar 2010). Depending on the purpose, sequences may initially be dereplicated, clustering them only if they are nearly identical (allowing for sequencing error and minor intragenomic variation). More frequently they are immediately grouped into "operational taxonomic units" (OTUs) at a lower level of identity such as 97%, associated roughly with the species level. This clustering provides the first glimpse of community structure, in the form of the number of species observed (richness) and their relative abundance distribution (evenness). It also makes downstream analyses more tractable (such as assignment of sequences to known taxa (Chapter 2), and computation of alpha and beta diversity measures (Magurran 2004) including UniFrac (Lozupone and Knight 2005)), by collapsing large numbers of redundant sequences, the analysis of which would be computationally expensive but uninformative, into a single representative.

The conventional rough correspondence of rRNA percent identity scores with taxonomic ranks (i.e., 97% = species, 95% = genus) was established with respect to full-length sequences (Wayne et al. 1987; Vandamme et al. 1996; Hugenholtz et al. 1998; Gevers et al. 2005; Goris et al. 2007). When only fragmentary sequences are available, the pairwise percent identity scores between these fragments do not perfectly mirror the percent identities that would have been found from full-length sequences, because of the well-known variations in mutation rate within the 16S sequence (Van de Peer et al. 1996a). Fragment percent identity scores should therefore be considered a noisy proxy for full length scores. Indeed, the use of sequence fragments is only one of many kinds of variation in how percent identity scores are obtained; others include variation in the alignment method, the use of the Lane mask (Lane 1991), and even the definition of "percent identity" itself (Schloss 2010). All of these can be thought of as sources of noise in the pairwise distances.

The question arises, then, to what extent the use of noisy distances impacts the results of the clustering procedure, and hence of downstream analyses. Here, I addressed this issue first with respect to noise in general, to get a sense of the overall magnitude of the problem. Next, I considered the choice of sequencing primer and read length, with the goal of recommending experimental choices that minimize noise and thereby produce a clustering that is as similar as possible to the full-length case.

This question also bears on taxonomic classification, which I treat explicitly in Chapter 2. The reason is that hypervariable regions mutate too rapidly to expect to find exact matches to environmental sequences in reference databases, at least given the level of coverage that these databases currently provide. Thus, the first step in annotating an environmental sequence is to identify database sequences that are likely to be in the same OTU.

3.3 Results

3.3.1 The response of clustering procedures to noise in the distance matrix

OTU counts depend primarily on the proportion of pairwise distances within the clustering threshold, not on the distances themselves

There are many important aspects of a clustering result, of course, but as a first step I simply considered the impact of noise on the one of the most basic outputs of a clustering procedure: the number of clusters (or OTUs) produced.

OTU clusterings are largely based on separating the set of all read pairs into two classes: those that are more than 97% identical, and those that are not. The single-linkage algorithm (aka "nearest neighbor" from DOTUR (Schloss and Handelsman 2005)) can be formulated as operating purely on this partition of read pairs, without reference to the actual distances. The complete-linkage ("furthest neighbor") clustering algorithm operates on only those sequence pairs in the close-distance set, and ignores the far pairs; the contribution of the distances is to choose the order in which nodes are agglomerated, thereby selecting one of possibly many clusterings that meet the complete-linkage criterion. Greedy clustering methods such as those provided by CD-HIT and UCLUST similarly consider only the close distances–in fact, only those between each input sequence and a limited number of "seed" sequences. The average-neighbor method (UPGMA) is the only one to which far distances may contribute, because they contribute to the distances computed between agglomerated clusters.

I first asked whether the exact values of the pairwise distances have an impact on the OTU count, assuming that the partitioning into near and far pairs is correct. Starting from a distance matrix relating sequences in a real dataset from a soil sample, I completely scrambled the distances less than 0.03 (97% identity) and those greater than 0.03, i.e. applying the transformation shown in Fig. 3.1. Any distance less than 0.03 was replaced with a distance uniformly sampled from the interval 0-0.03, and any distance greater than 0.03 was replaced with a distance uniformly sampled between 0.03 and 0.6. I then performed OTU clustering on this scrambled distance matrix; 100 repetitions produce a histogram of OTU counts.

Because the single-linkage method ignores distances once the partitioning is known, it naturally produced the same result in each case. The variation in OTU count using the complete-linkage method is due exclusively to the altered order of agglomeration of nodes into clusters. In the average-linkage case, agglomeration order plays a role, but so do the (now randomized) distance values; nonetheless the distribution is no wider than that produced by complete linkage.

The two types of error have different effects, depending on the clustering method (The clustering obtained responds differently to false positives and false negatives)

The results of the previous section show that the first step of clustering is simply a binary classification problem: the pairwise distances must be separated into those that are within a given

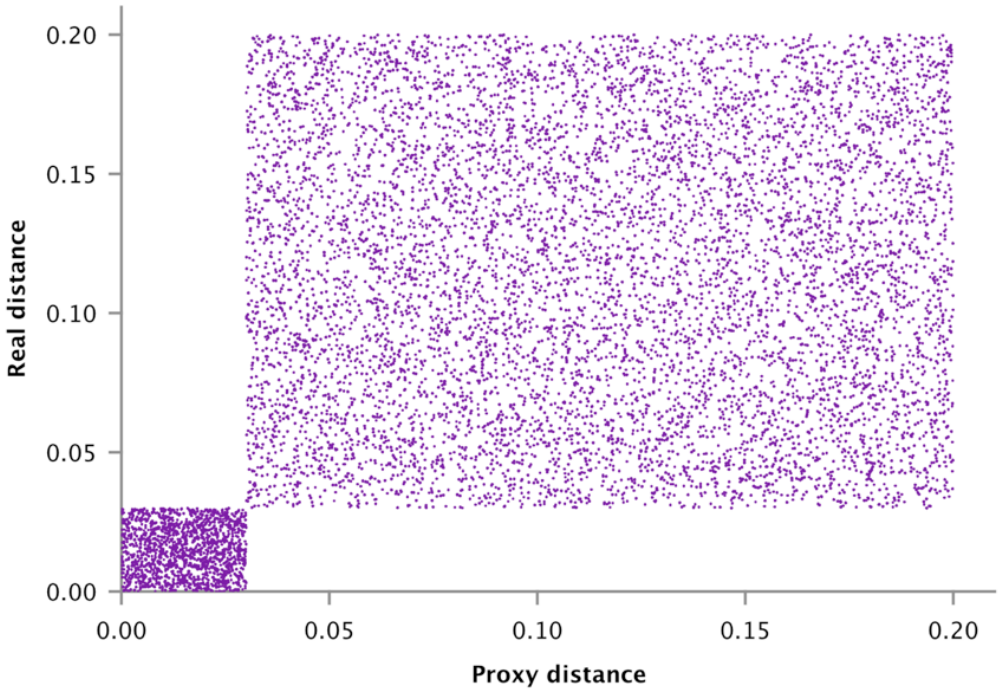

Figure 3.1: The distance-scrambling transformation

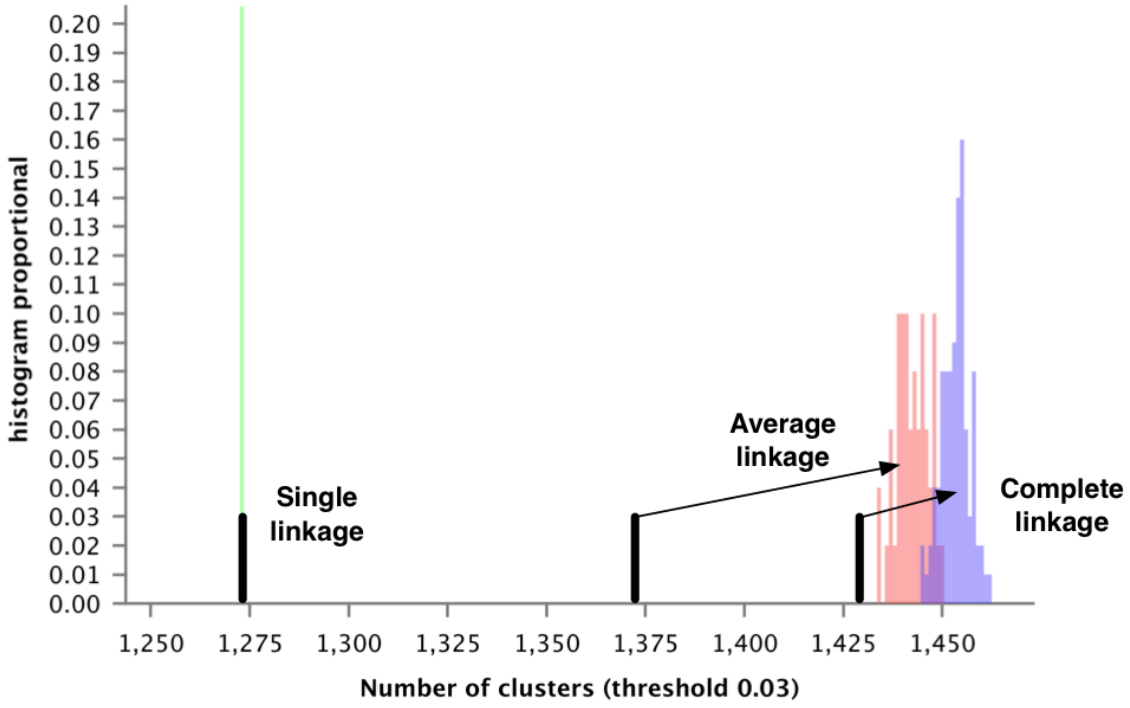

Figure 3.2: Distribution of OTU counts from a soil dataset, obtained from scrambling the distances, assuming correct partitioning. The solid bars show the number of clusters produced by three clustering procedures prior to distance scrambling. The histograms show the distributions of the number of clusters produced by each procedure after scrambling, for 100 replicates.

distance threshold, and those that are not. A noisy distance measure will make two kinds of mistakes: it will call some distant pairs close (false positives; reduced specificity) and will call some close pairs far (false negatives; reduced sensitivity). Because the effect of this mispartitioning on the resulting OTU count dominates the effect of the specific distance values, we can simply describe any noise distribution by the sensitivity and specificity with respect to this binary classification problem.

I empirically determined the influence of partitioning noise on OTU counts for an example soil dataset. For purposes of this experiment, I considered DNADIST distances between full-length sequences aligned by NAST (Desantis et al. 2006a) to be the reference distances (though this choice should not impact the outcome at all). Then I scrambled the distances as above, this time allowing false positives and false negatives in various proportions (fig 3.4), and then clustered the data based on the noisy distances. I repeated this process 100 times for each clustering method. The dependence of mean OTU count on the proportion of false positives and false negatives is shown in Figure 3.5.

The single-linkage clustering process reacts to false negatives by increasing the number of clusters, because often the close link that is missed is the only one that connects two clusters. Conversely, false positives rapidly reduce the number of clusters, because each false positive is likely to join two clusters that would otherwise have been separate.

The complete-linkage clustering process also reacts to false negatives by increasing the number of clusters, because removing any link from a clique makes it no longer a clique, so it must be broken in two. Complete linkage is less responsive to false positives, however. The reason for this is that joining two cliques together requires adding *n*m* links (the sizes of the two cliques, respectively), which is very unlikely even with a large number of false positives.

The greedy clustering algorithms are already heuristic in nature, because they do not consider even all of the "close" pairwise distances. Rather, they compute distances only to limited set of representative sequences, the selection of which is highly sensitive to the order in which the inputs are presented. I did not evaluate the relative contribution of that source of noise compared to noise in the distances. Nonetheless it is clear that a false positive "close" distance may prevent the creation of a new cluster, by causing a sequence to be added to an existing cluster when otherwise it would have become a new seed. On the other hand, when the next input arrives that would have been added to the missed cluster, it will likely seed an an analogous cluster of its own, thereby muting the effect of the original false positive. A false negative may have the reverse effect, resulting in the creation of a new seed, but only if the match that is missed is the only one available; otherwise the sequence will simply be associated with the next-nearest seed. The relative frequencies of these various kinds of events depends on the number of seeds and the (eventual) distribution of cluster sizes. Overall, it seems likely that noise in the distances may produce many minor errors (i.e., assignment of a sequence to the wrong cluster), but far fewer major ones (creation of too few or too many clusters), because the vast majority of errors have consequences only for a single sequence before they are forgotten. This is in contrast to the non-heuristic methods, where individual errors (especially false negatives) compound, and thus are more likely to impact the cluster count. In any case, the entangled and opposing contributions of various sources of noise call for a more thorough empirical investigation of the

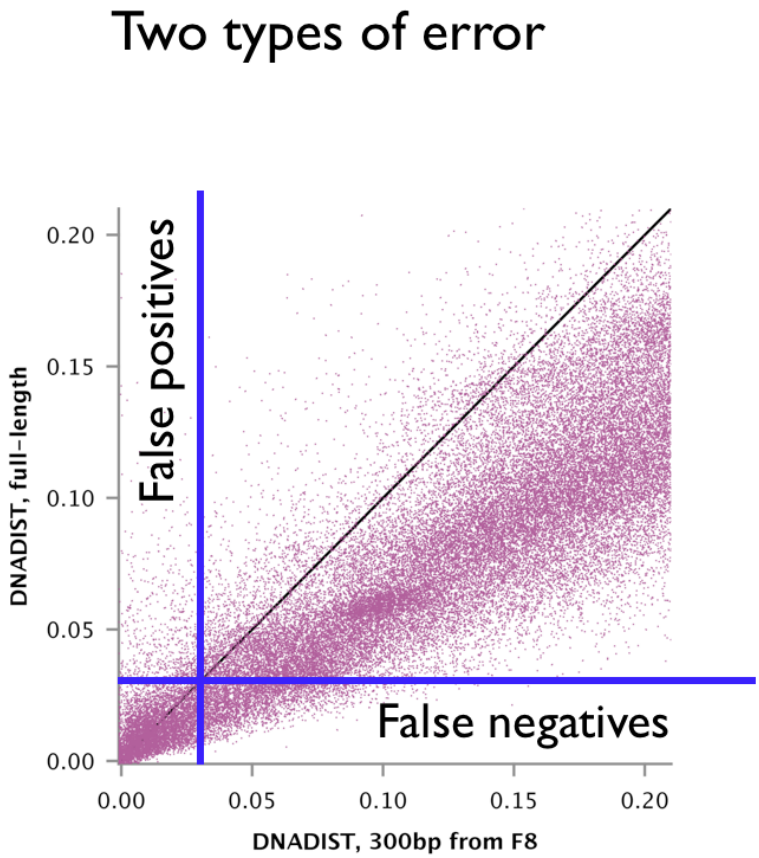

Figure 3.3: Example of noise when predicting full-length %id from a proxy distance. Each point represents a pair of sequences. Full-length percent difference appears on the Y axis, and percent difference between 300bp reads from E8F appears on the X. This region evolves more quickly than average, so that the distribution falls below the y=x line. If one were to predict pairs whose full-length %id is 97% or better simply by choosing those whose fragment %id is 97% or better, one would thus incur substantially more false negatives than false positives. The error types could be balanced, and perhaps the total number of errors decreased, by using a lower threshold on the fragment axis (e.g., 95%) to predict 97% identity on the full-length axis.

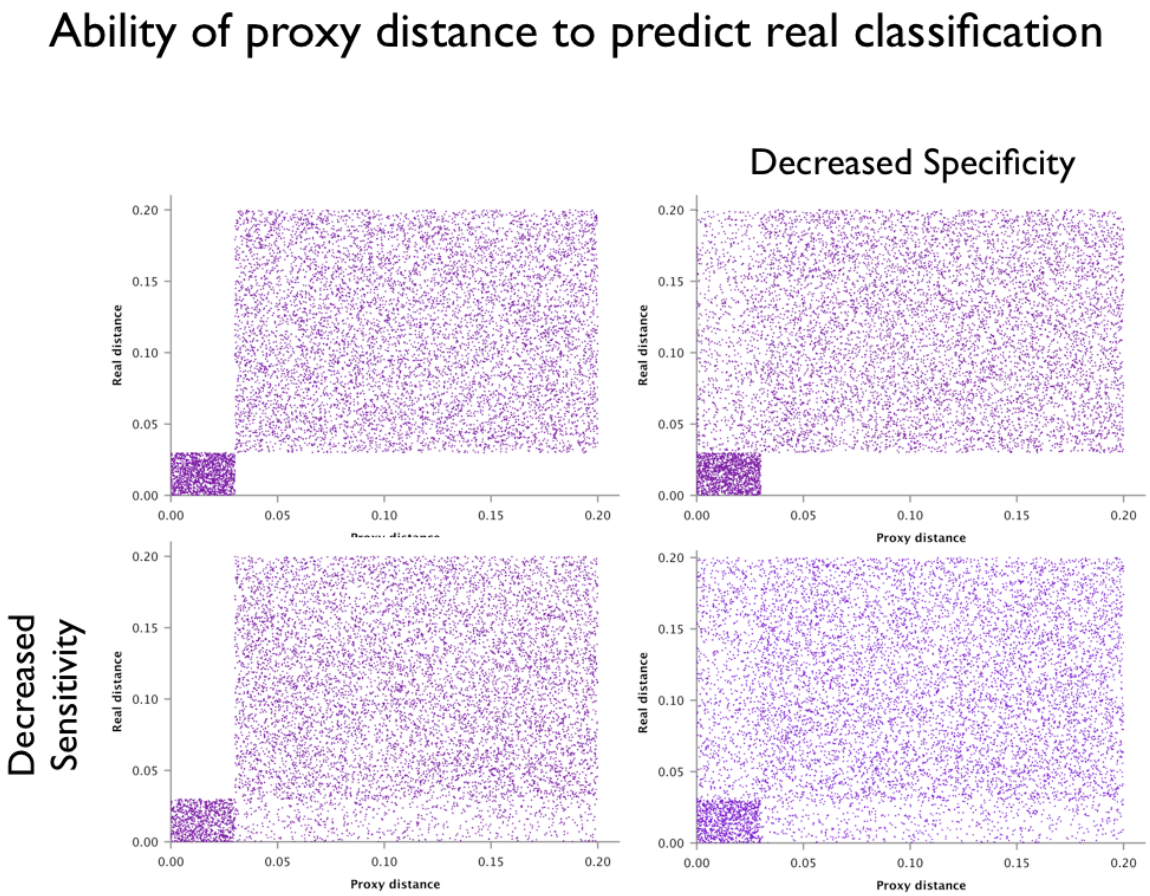

Figure 3.4: The distance-scrambling transformation, with noise.

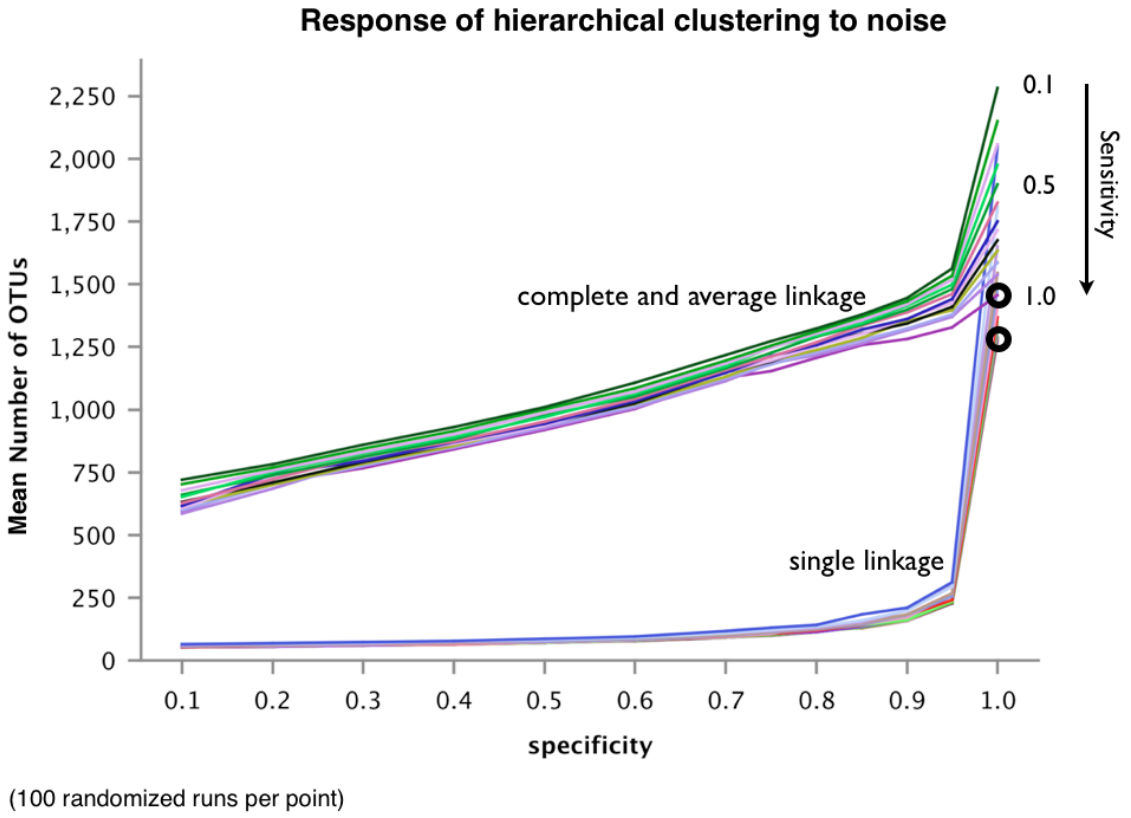

Figure 3.5: Mean OTU count vs. different amounts of noise for all three linkage methods.

greedy heuristic approach.

Complete linkage clustering is the only one that makes sense with respect to phylogeny

We cluster similar sequences together to indicate that they belong to the same taxon, i.e. that they are members of a subtree on a phylogeny descending from a common ancestor. In this view, the most appropriate clustering procedure would be simply to build phylogenies and choose appropriately sized subtrees. However, the computational cost of doing this is prohibitive given the very large numbers of sequences that can now be produced.

Of the simple clustering methods discussed here, only the complete linkage approach is compatible with the phylogenetic view of the problem. This is a simple consequence of the conventional definitions that two sequences can be in the same OTU only if they are within a given distance of one another, and that taxa are monophyletic. Under these definitions, an OTU is a subtree with the property that the greatest distance between any two leaves is no greater than a given threshold (e.g., 3%). That is almost exactly the complete-linkage criterion, assuming that the tree distances closely reflect distances based on sequence identity scores.

The average linkage, single linkage, and star methods all commonly create clusters containing pairs of sequences that are more distant than the allowable threshold. Placed on a phylogeny, the common ancestor of sequences in such a cluster would therefore be higher in the tree than in the complete-linkage case. Worse, these methods provide no assurance that the clusters are nonoverlapping on the tree. Even complete linkage does not guarantee monophyletic clusters, because one taxon may be contained within another (i.e., the common ancestor of one taxon may descend from the common ancestor of another taxon); but at least it cannot happen that the clusters overlap, so that descendants of both common ancestors include members of both taxa.

3.3.2 Choice of maximally informative primer and read length

Some sequence regions are much better at predicting "near" vs. "far" than others

I wished to evaluate the ability of each sequence region to distinguish same-taxon from different-taxon pairs–that is, the ability of each region to predict, from a pair of fragments, whether the percent identity score for the corresponding pair of full-length sequences is within a given threshold or not. I sought to predict thresholds of 95%, 97%, and 98.5%, roughly corresponding to the genus, species, and strain levels, respectively.

I started from from near-full-length sequences in GreenGenes, and extracted those associated with three environment types: human gut, soil, and ocean. I then simulated short read sequencing experiments, extracting reads of various lengths starting from 44 primers. For each choice of primer and read length, I determined the threshold percent identity between the fragments that provided the best classification performance with respect to taxon clusters determined from the full-length sequences. For fragments that evolve more rapidly than average, the fragment %id threshold should be higher than the target full-length %id, and conversely for more conserved

regions. For consistency with the next section, I performed the optimization using an SVM (even though that is overkill for this problem). Figure 3.6 shows the classification performance obtained for each combination of primer and read length.

These figures demonstrate that the choice of primer and read length does indeed have a large impact on the accuracy with which pairs of sequences can be determined to be in the same taxon or not. I had expected genus level distinctions to be easier to make than strain level distinctions, and was surprised to find that the best primers were able to make classifications of comparable accuracy across all three taxonomic levels.

I used the class-normalized sensitivity as a simple measure of classification performance: that is, the average of the sensitivity with respect to identifying same-taxon pairs and the sensitivity with respect to identifying different-taxon pairs. This number can also be thought of as the accuracy of predictions in the case that the classes are balanced, i.e. when there are equally many positive and negative examples.

This measure is not meant to reflect the composition of real environments, which can differ greatly in the proportion between the classes. For instance, in a sample that is both very diverse and very even, the proportion of sequence pairs originating from the same species will be very small (indeed, for a small sample size, it may happen that no two sequences come from the same species). In a sample that is dominated by a few species, conversely, the proportion of within-species pairs may be quite large. The test datasets employed here aggregate sequences from multiple sources, and are sampled nonuniformly (see section 3.5.2); thus, the resulting accuracy measure does not predict classification performance for any particular real environment. Rather, it is a relative measure that allows me to rank the primer/read length combinations to determine which one is the most informative. I assume that this ranking of primers will be consistent between environments of different composition.

Note too that this measure takes into account the fact that some primers hit fewer sequences than others. Because the class sensitivities are measured with respect to the total sample, reduced primer coverage simply reduces both of them.

Table 3.1 lists the fragment class-normalized sensitivity values obtained for the primers that produce the best results for each read length in each of the three environment types, and Table 3.2 lists the corresponding %id thresholds. For read lengths of 75, 100, and 125nt, primer E517F achieved optimal performance for nearly every combination of environment and taxonomic level. Notably, primer E969Fi proved substantially better at distinguishing strains in human gut, but dramatically worse in the other two environments; indeed 75nt reads from E517F performed better under all other circumstances.

Paired-end sequencing offers no benefit

We hypothesized that paired-end reads might be more effective than single reads at determining whether two sequences come from the same species or not, both because there is simply twice as much sequence to work with for a given read length, and because we thought that the reads might provide different kinds of information. For instance, one fairly conserved region might

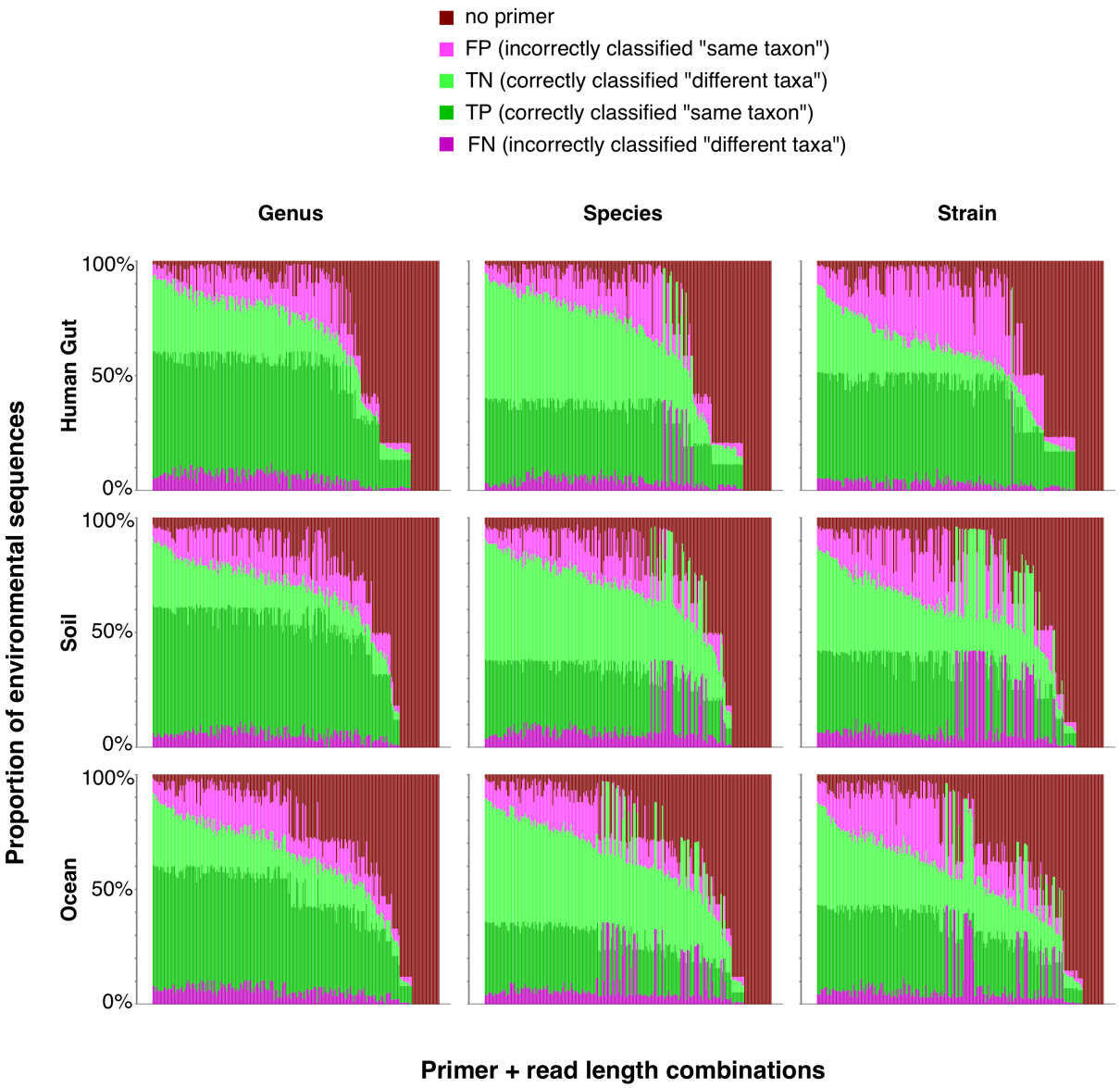

Figure 3.6: Impact of primer and read length on classification of sequence pairs into within-taxon and between-taxon pairs, for different environment types and taxon sizes. 231 possible choices of primers and read lengths are sorted on the X axis by overall accuracy. The bars above each choice show the proportions of pairs of sequences, sampled as described in section 3.5.2, that are correctly or incorrectly classified.

		Class-normalized sensitivity for making taxon distinctions								
		Genus			Species			Strain		
read length	primer	Human Gut	Ocean	Soil	Human Gut	Ocean	Soil	Human Gut	Ocean	Soil
50	E1064Ri	0.73	0.63	0.63	0.76	0.71	0.71	0.71	0.67	0.67
	E1391F	0.74	0.65	0.67	0.76	0.72	0.68	0.58	0.65	0.47
	U529R, E533Ra	0.69	0.69	0.67	0.68	0.74	0.71	0.56	0.61	0.59
	E969Fi	0.76	0.44	0.51	0.79	0.49	0.55	0.73	0.44	0.53
75	E517F	0.79	0.71	0.66	0.85	0.76	0.75	0.66	0.70	0.63
	U515F	0.70	0.65	0.68	0.81	0.72	0.72	0.59	0.59	0.55
	U529R	0.66	0.67	0.66	0.77	0.74	0.74	0.58	0.67	0.66
	E1064Ri	0.74	0.63	0.66	0.76	0.71	0.72	0.70	0.67	0.70
	E926Ra	0.70	0.64	0.69	0.72	0.64	0.71	0.60	0.62	0.59
	E969Fi	0.78	0.44	0.54	0.80	0.51	0.57	0.75	0.47	0.54
100	E517F	0.82	0.76	0.73	0.87	0.79	0.77	0.71	0.70	0.69
	E969Fi	0.78	0.44	0.54	0.80	0.51	0.57	0.75	0.47	0.54
125	E517F, U515F	0.78	0.77	0.73	0.87	0.79	0.78	0.74	0.72	0.73
400	U529R, E533Ra	0.83	0.84	0.84	0.91	0.86	0.87	0.83	0.84	0.81

Table 3.1: Primers producing optimal clustering fidelity at each read length. Hundreds of combinations that produce suboptimal results are not shown (see Materials & Methods 3.5.5). The highlighted cells indicate the best achievable classification rates for each environment and read length.

		Optimal difference threshold for making taxon distinctions								
		Genus (5%)			Species (3%)			Strain (1.5%)		
read length	primer	Human Gut	Ocean	Soil	Human Gut	Ocean	Soil	Human Gut	Ocean	Soil
50	E1064Ri	10.1%	8.0%	11.7%	4.5%	3.1%	4.6%	3.1%	2.6%	3.1%
	E1391F	2.0%	3.1%	5.3%	1.0%	1.7%	1.0%	1.0%	1.0%	0.0%
	U529R, E533Ra	2.0%	2.9%	3.1%	1.0%	1.0%	1.6%	1.0%	1.0%	1.0%
	E969Fi	9.8%	8.9%	9.9%	6.0%	4.2%	4.2%	3.1%	2.0%	2.0%
75	E517F	2.0%	2.0%	2.7%	0.6%	0.6%	1.3%	0.6%	0.6%	0.6%
	U515F	1.3%	1.3%	2.0%	0.6%	0.6%	1.3%	0.6%	0.6%	0.6%
	U529R	4.2%	8.0%	7.2%	1.6%	4.2%	3.6%	1.3%	2.0%	2.4%
	E1064Ri	12.6%	9.5%	12.9%	5.7%	4.2%	5.7%	4.2%	2.0%	3.4%
	E926Ra	2.0%	2.7%	2.0%	0.9%	1.3%	0.6%	0.6%	0.6%	0.6%
	E969Fi	13.6%	10.3%	12.6%	5.9%	4.2%	5.7%	4.2%	2.0%	2.7%
100	E517F	3.6%	3.6%	4.1%	1.5%	1.5%	2.0%	0.5%	1.5%	1.0%
	E969Fi	9.5%	8.1%	9.2%	4.7%	3.1%	4.2%	3.1%	1.5%	2.0%
125	E517F, U515F	4.1%	4.5%	5.4%	1.6%	1.6%	2.0%	0.8%	1.2%	1.2%
400	U529R, E533Ra	4.7%	6.2%	5.9%	2.8%	3.7%	3.5%	1.2%	1.7%	1.6%

Table 3.2: Percent difference threshold between sequence fragments corresponding to full-length percent difference thresholds. These thresholds were found to produce the best available discrimination of within-taxon and between-taxon pairs. Values that are higher than the corresponding full-length value indicate a sequence region with a higher average mutation rate than the sequence as a whole. For instance, primer E1064Ri targets a hypervariable region, so that 8%-13% sequence difference is required to conclude that two sequences originate from different genera, where 5% would be sufficient given full-length sequence. Conversely, primer E517F targets a conserved region, reflected here in percent difference thresholds that are lower than the full-length thresholds that they predict. Only optimal primer choices from table 3.1 are listed; the complete table for all choices of primer and read length is available in the supplementary material. Values seem to have more significant figures than is possible given the read length (e.g., for 50nt reads, thresholds ought to be quantized in 2% increments); this reflects the fact that the SVM learns a decision boundary that lies between the input data points. Also, the occasional insertion event causes the inputs to be imperfectly quantized.

		maximum achievable class-normalized sensitivity								
		Genus			Species			Strain		
type	read length	Human Gut	Soil	Ocean	Human Gut	Soil	Ocean	Human Gut	Soil	Ocean
single-ended	50	0.76	0.69	0.67	0.79	0.74	0.71	0.73	0.67	0.67
	75	0.79	0.71	0.69	0.85	0.76	0.75	0.75	0.70	0.70
	100	0.82	0.76	0.73	0.87	0.79	0.77			
	125		0.77				0.78		0.72	0.73
	400	0.83	0.84	0.84	0.91	0.86	0.87	0.83	0.84	0.81
paired-end	50	0.76	0.69	0.67	0.79	0.74	0.71	0.73	0.67	0.67
	75	0.79	0.71	0.72	0.85	0.76	0.75	0.75	0.71	0.70
	100	0.83	0.77	0.75	0.87	0.79	0.77		0.73	0.74
	125		0.79	0.77	0.89	0.8	0.78		0.76	0.76
	400	0.90	0.84	0.84	0.91	0.86	0.87	0.86	0.84	0.81

Table 3.3: Paired-end sequencing offers little improvement over single-ended sequencing in class-normalized sensitivity for predicting co-clustering. The class-normalized sensitivity shown in each cell is the maximum value observed for any choice of primers at each respective read length. Empty cells indicate no improvement over shorter reads. Cells are highlighted when paired-end sequencing provides an improvement over single-ended sequencing using the same read length. Despite having twice as much sequence to work with, only about one third of the paired-end cells are thus highlighted, and even these provide minimal gains compared to the corresponding single-end experiments.

be good at making high-level phylum distinctions, while another, more variable region might be good at making fine-grained species distinctions within a phylum, but less good at the phylum level because too many mutations accumulate over long evolutionary distances. In such a case, combining information from both reads could provide improved discrimination.

Under that hypothesis, one way to use paired-end reads to predict whether two sequences have a common species origin would be to be compute the %id scores for each read individually, and then to use a weighted mixture of these to predict whether the full-length %id falls within 97% or not. That is, we would want to count mutations in a conserved region as more significant than mutations in a variable region.

I evaluated this question using a simple SVM classifier with a linear kernel. This procedure learned the weights to be assigned to each read in order to maximize discrimination performance. The cartoon in Figure 3.7 helps to visualize this approach: for a large set of pairs of full-length sequences, and a given pair of primers producing short reads, the task is to choose the line that best separates the green points (within-taxon) from the red ones (between taxa). Because the green and red points are not linearly separable, some number of misclassifications must be accepted, but this number is to be minimized.

Table 3.3 shows the best class-normalized sensitivity achievable using any pair of primers for

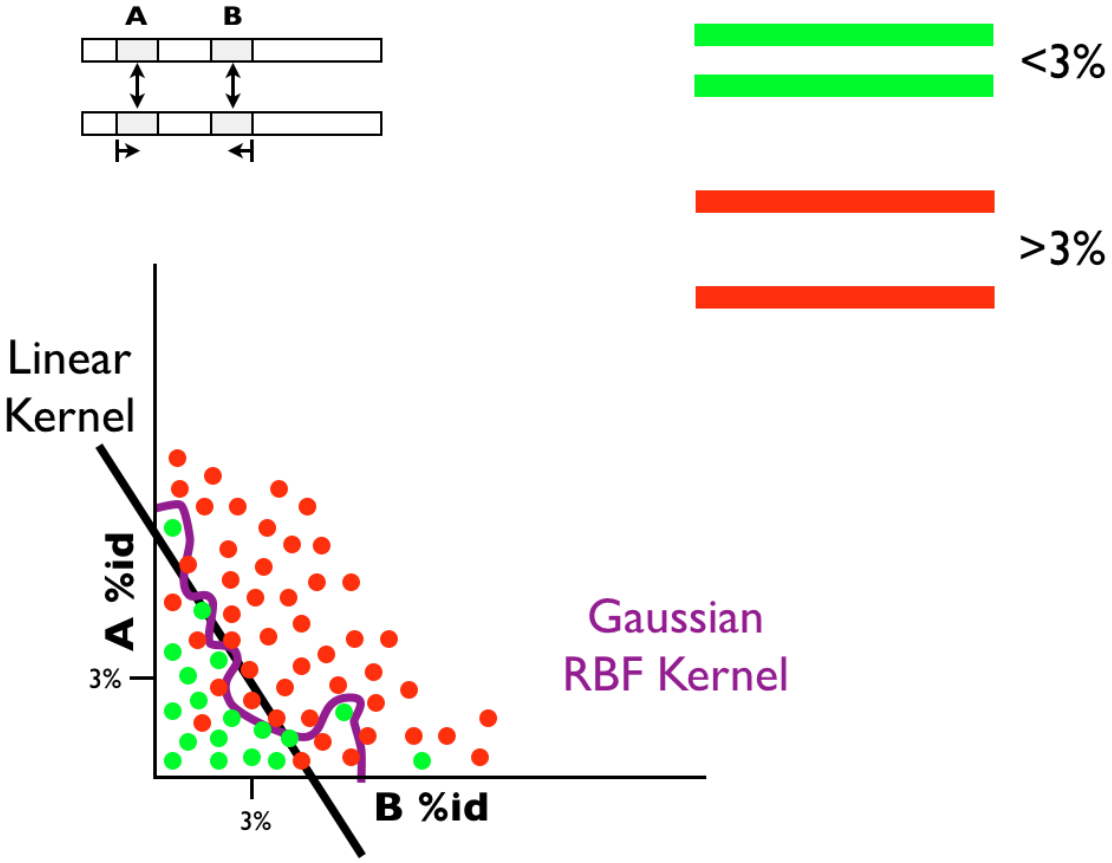

Figure 3.7: Cartoon of the SVM task for predicting common taxon origin from paired-end sequence fragments. Each point represents a pair of sequences, colored according to whether the full-length percent identity is at least 97% (green) or not (red). The X and Y axes represent the percent identities of short reads extracted from those sequence pairs, using a forward and a reverse primer, respectively. If a line could be found that separates the red points from the green, that would indicate that paired-end sequencing using the chosen combination of primers and read length could be used to accurately predict whether the sequences originate from the same taxon or not. In practice, perfect separation cannot be achieved; thus the task is to choose the line to minimize prediction errors.

each read length, and compares these with the best sensitivities achievable from single-ended reads. I was surprised to find that paired-end sequencing does not appreciably improve classifications. Compared with the best single-ended primer choices, the best paired-end experiments (using twice as much sequence in total) show an improvement of at most a few percent, and none at all at all in the majority of cases. The result is more dramatic when the total amount of sequence is held fixed (on the argument that cost scales roughly per nucleotide): a single read of 100nt always outperforms a pair of 50nt reads, and a single 125nt read always matches or outperforms a pair of 75nt reads.

I repeated the paired-end experiments using the Gaussian RBF kernel, which can learn nonlinear boundaries between the two classes. I performed a grid search to optimize the parameters C and gamma. I found that this did not offer improved performance over the linear kernel (data not shown). This suggests that the fragment percent identity values for the two ends are independent given the full-length percent identity. In other words, I detected no nonlinear dependence among mutation rates between any two regions that might be exploitable for the classification task.

In combination, these findings falsify our hypothesis that information from two regions of the 16S sequence might be combined to provide substantially greater taxon clustering accuracy than a single region.

3.4 Discussion

I have shown that environmental 16S surveys should be performed using primer E517F, in order to distinguish within-OTU from between-OTU sequence pairs as accurately as possible given the read length limitations of current sequencing technologies. Because OTU clustering depends largely on this distinction, making this optimal choice will provide the clustering result that most closely resembles that which would be obtained from full-length sequencing.

This is a somewhat indirect argument; what is important, ultimately, is whether the choice of primer and read length impacts downstream biological conclusions.

Ecological diversity measures such as richness, evenness, and many others form one frequently reported class of results from environmental surveys. I showed that the simplest of these, the OTU richness, is quite sensitive to noise in the pairwise distances between environmental sequences, and that a suboptimal choice of primer will introduce such noise in abundance. This suggests that richness and other diversity measures derived from short read sequencing using suboptimal primers should be treated with substantial skepticism.

On the other hand, it was previously shown that the unweighted UniFrac beta diversity measure is not very sensitive to the choice of primer (using a more limited set) (Liu et al. 2007). This finding may be explained by the fact that UniFrac is a phylogenetic method, where the contribution of long branch lengths nearer the root of the tree dominates the result. The errors in OTU clustering which I have attempted to minimize here most likely to occur near the leaves of the tree: for instance, noise in the pairwise distances may substantially alter the number of

species detected, but it is much less likely to alter the number and abundance distribution at the phylum or class level. Thus, because UniFrac measures primarily the concordance between two communities at higher levels of the tree, it makes sense that it is not particularly sensitive to errors in distinguishing species or genera from one another. On the basis of this argument, I speculate (though I have not confirmed) that UniFrac values should be fairly robust to the OTU clustering threshold; that is, the clustering of similar environments based on UniFrac distances may be largely the same when only class-level OTUs are considered (for example) instead of the more commonly used species-level OTUs.

The authors of that study recommended primer E357R on the basis that UniFrac-based clusters of environmental samples sequenced from that primer recapitulate the clusters obtained from full-length sequencing (for example, grouping gut samples from mouse littermates together with their mother) better than other primers. I cannot explain this finding using the present results. Primer E357R provides middling performance in our evaluation– substantially worse than several of the other primers they tested, especially of course E517F. I would expect that more accurate clustering of sequences within a sample would lead to more accurate clustering of samples based on UniFrac distances.

Another type of biological result of widespread interest is taxonomic classification of environmental sequences. I addressed the question of primer choice for that problem explicitly in Chapter 2. Nearly all of the primers reported here as optimal for unsupervised clustering were found in that study to be near optimal for supervised classification as well. I also found in both studies that paired-end sequencing provides no benefit. In particular, single-ended reads from primer E517F provided the most accurate results in both studies across a range of read lengths and environments. This concordance is particularly notable because the prior study was based on completely different methods and different (though partly overlapping) input datasets, and had a different goal.

One possible explanation for the previous classification results is as follows. I showed here that the region following E517F is particularly good at identifying close relatives. Thus, compared to other regions, this region tended to match sequences in the reference database that were both more likely to be in the same taxon as the query sequence and consequently more likely to agree with one another with respect to taxonomic annotation. Since both supervised and unsupervised clustering procedures depend on a distance measure, then, more accurate distance estimates produce more accurate results in both cases.

As in the prior study, I found no clear indication that hypervariable regions should or should not be targeted for sequencing. For 50nt reads, the best strain-level distinctions are made using either E969Fi (only for the human gut dataset) or E1064Ri, both of which produce reads including about 20nt of the V6 hypervariable region. The high variability of these sequences is reflected in the percent difference thresholds shown in table 3.2, which are much higher than the full-length thresholds that they predict. The best strain-level distinctions in human gut using 100nt sequences were also obtained with primer E969Fi, including all of V6; but for every other combination of environment and taxonomic level, even 75nt reads from E517F performed better, despite consisting almost entirely of the relatively conserved region between V3 and V4. Because of this sequence conservation, lower percent difference thresholds are required

for reads from E517F; for instance, a threshold of 2.0%-2.7% sequence difference produces the most accurate genus-level distinctions (corresponding to 5% difference for the full-length sequence). Similarly, reads of 75nt from U529R include most of V3 (~40nt), yet 50nt reads from E1391F perform largely comparably while including only 10nt of V9. The argument has been made that hypervariable regions (especially V3 and V6) are likely to make cleaner distinctions among strains within a sample than is possible using more conserved regions (Wang et al. 2007; Huse et al. 2008). Our results do not support this assertion, except perhaps for very short reads.

I conclude that 16S primers and read lengths should be chosen on the basis of their empirical performance for the task at hand, not on the basis of emphasizing hypervariable regions or even necessarily of maximizing read length. Based on simulated experiments using thousands of combinations of primers, read lengths, and environments, I found that primer E517F provides optimal or near-optimal accuracy (with respect to analogous results from full-length sequencing) for both supervised and unsupervised clustering tasks in nearly every circumstance.

3.5 Materials & Methods

3.5.1 Construction of test datasets

I downloaded the GreenGenes database (version of August 25, 2010) and selected near-full-length sequences by requiring that each sequence span hypervariable regions V1 through V9 (specifically, extending from positions 69 to 1465 in *E. coli* coordinates). This condition is met by 188,580 of the 508,194 total sequences in GreenGenes. I then determined the environment type from which each sequence originated, by simple text matching on the "isolation_source" field. For instance, I assigned the label "Ocean" to any sequences whose isolation source contained one of the words "marine", "harbor", "plankton", and so on. I thereby extracted sets of full-length sequences from human gut, soil, and ocean.

I partially dereplicated each of these datasets by clustering them at 97% identity using UCLUST. For clusters with fewer than ten members, all were retained; for larger clusters, I randomly selected ten representatives. This step removed some degree of database bias at the species level within each set, while retaining many within-species sequence pairs.

3.5.2 Sampling of full-length sequence pairs

I sampled pairs of sequences from each environment in a manner that allowed us to test how well within-taxon pairs could be distinguished from between-taxon pairs. First, I reasoned that very divergent pairs of sequences, those with less than 90% identity, can easily be identified as coming from different genuses, and thus need not be included in the experiment. I therefore sampled 5000 sequence pairs per environment such that their percent identity scores were uniformly distributed between 90% and 100%. Thus, for the genus level (95% identity) experiments, there were roughly equal numbers of positive examples (within-taxon pairs) and negative

examples (between-taxon pairs). For the species level (97% identity) experiments, roughly 30% of the examples were in the positive class. For the strain level (98.5% identity) experiments, I sampled 5000 sequence pairs uniformly distributed between 95% and 100% identity, so that the positive examples represented roughly 30% of the total here also.

3.5.3 Extraction of sequence reads

I simulated single-ended next-generation sequencing experiments by extracting reads of lengths 50, 75, 100, 125, and 400 nt starting from 44 universal bacterial primers, for a total of 220 viable combinations. I also simulated paired-end experiments by generating the 1619 viable pairings of those reads. The primers were chosen as previously described (Section 2.5.4).

3.5.4 SVM training

I trained a support vector machine for each environment, primer or pair of primers, and read length, and measured classification performance using 5-fold cross-validation. For each of the 5000 sequence pairs in each dataset, the true within- or between-taxon relationship was known from the full-length percent identity. For paired-end reads, the inputs were the percent identity scores computed independently for each end. For single reads, there was only one input: the percent identity of the read. Thus the SVM reduced to finding the percent identity threshold for the fragment that best corresponds to the percent identity for the full-length sequence (effectively, a measure of the local mutation rate).

I used my jLibSvm software (http://dev.davidsoergel.com/jlibsvm, a refactored Java port of LIBSVM (Chang and Lin 2001)) with a linear kernel. For each scenario, I performed a grid search to obtain the optimal value of the cost parameter C. I adjusted C for the positive and negative examples to correct for the size imbalance between the two classes, so as to make misclassifications on either side equally important.

For the paired-end scenarios, I repeated the experiment using a Gaussian (RBF) kernel, performing a grid search over the parameters C and gamma.

For each scenario, then, I obtained the proportion of sequence pairs that could not be compared because one of the primers did not hit one of the sequences, and the proportions of true positives, true negatives, false positives, and false negatives produced by the optimized classifier.

3.5.5 Choice of representative optimal primers

I exhaustively computed the best achievable classification performance for thousands of combinations of primer, read length, environment, and taxonomic level. I found that some choices of primer and read length provided better classifications than certain other choices across all levels and environments tested. My results suggest, for example, that there is no reason to use primer E357R with 100nt reads, because primer E517F with 75nt reads is always more informative. I

filtered the results table to exclude choices of primer and read length that were uniformly less informative than others of the same or shorter read length.

In a few cases, several choices provided nearly equivalent classification performance, particularly involving closely related primers such as E517F and U515F. I considered two choices to be equivalent (for a given taxonomic level and confidence level) if they provided class-normalized sensitivities within one percentage point for all levels and in all environments. In these cases I list each alternative but report the classification performance of one arbitrarily chosen representative.

Part II

Binning Methods

Chapter 4

Supervised compositional binning methods are doomed on natural metagenomic samples

4.1 Abstract

The application of genome sequencing to microbial communities in recent years has produced an ever-increasing flood of "metagenomic" data, consisting of millions of shotgun reads sequenced directly from numerous environments. An essential step in analyzing these data is to classify the reads into taxonomic groups, both in order to estimate the species composition of the community (for comparison with 16S surveys) and more importantly to link metabolic functions with the organisms that contain them. The ability to solve this "binning" problem accurately would have great consequences for detection and treatment of human disease, biosensing, bioprospecting, bioremediation, and understanding biogeochemical cycles. Conversely, the ability to observe the species distribution of specific genes of unknown function may aid in determining their role.

Numerous procedures have been proposed for supervised taxonomic classification, in which environmental sequence reads are assigned to known taxa based on similarity of compositional biases. Here I report that these methods cannot be expected to work nearly as well in practice as prior studies suggest, at least not when fully-sequenced genomes are used for training. This is a consequence of two observations: supervised binning is accurate only when the query sequence is very closely related to one of the training bins, but a large proportion of microbes in the environment are phylogenetically distant from any fully-sequenced genome.

I conclude that accurate binning will require at least one of three approaches: 1) a vastly more comprehensive set of training genomes (hard to come by due to culturing difficulties); 2) unsupervised binning methods; or 3) self-supervised binning methods, in which the training sequences are taken from the very dataset that is to be binned.

4.2 Introduction

Since the advent in 1995 of full-genome sequencing of microbes, approximately 1190 bacterial and archaeal strains have been sequenced. The selection of strains for sequencing has been largely driven by ease of cultivation and medical relevance, leading to biases in the sampling of the microbial universe (Hugenholtz 2002). At present, only 24 of the ~100 known bacterial divisions have even a single isolate representative.[1] It is thus widely accepted that isolate sequencing efforts have barely scratched the surface of microbial diversity. The plummeting price of sequencing is now spurring efforts to sample both more widely, as in the GEBA project (Wu et al. 2009), and more deeply in a targeted manner with respect to a specific environment, as in the Human Microbiome Project (Turnbaugh et al. 2007; Consortium et al. 2010).

In parallel, recent years have seen the development of a genomic approach to microbial ecology, based on shotgun sequencing of environmental samples. The set of fully-sequenced isolate genomes provides an important reference for the interpretation of these metagenomic data. Methods for solving various problems in metagenomic data analysis, such as taxonomic classification of sequence reads, analysis of community species composition based on marker genes, and comparative assembly, are sensitive to the phylogenetic proximity between each environmental sequence read and a fully-sequenced isolate (Kunin et al. 2008).

Here I concentrate on the "binning" problem, where the task is to classify short nucleotide sequences from the environment into known taxonomic groups. The most obvious procedure to do this is simply to perform sequence similarity searches (e.g., using BLAST), and to classify the query sequence to a taxon in which similar sequences are found. Many classification procedures have been designed around this idea (Mavromatis et al. 2007; Hanekamp et al. 2007; Krause et al. 2008; Monzoorul et al. 2009; Essinger and Rosen 2010; Horton et al. 2010). A general concern with this approach is that it can be difficult to know at what taxonomic level it is appropriate to make an assignment. The best BLAST hit may be annotated to the strain level, but may in fact be in a different family or even division from the query sequence due to limited database coverage and database bias. The degree of conservation of different sequences is of course highly variable, so the similarity score between a query sequence and a database hit cannot in general be used to infer the phylogenetic distance. These issues can be addressed to some extent by aggregating the annotations from a number of database hits and by considering their e-value scores in concert; but the tradeoffs inherent in these decisions have not yet been thoroughly explored.

An alternative approach is to classify sequences on the basis of statistical descriptions of their composition, typically involving the frequency distribution of short words ("*k*-mers") within the sequence. These compositional signals can be exploited for taxonomic classification on the basis of the surprising observation that microbial genomes each have a characteristic "genome signature" that may be detectable even in fairly short reads (Jeffrey 1990; Karlin et al. 1998a; Sandberg et al. 2001; Mavromatis et al. 2007) . One motivation for compositional binning is the idea that signatures may be detectable for higher-level taxa, even at the phylum level, allowing

[1] http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html, http://greengenes.lbl.gov/cgi-bin/nph-browse.cgi

at least a coarse classification of environmental sequences that are not closely related to known organisms in the reference database (McHardy et al. 2007).

The biological basis of the phenomenon of genome signatures is not well understood. For instance, it is not known how quickly signatures diverge after speciation; whether they diverge continuously or in a punctuated manner; whether cases of convergence are influenced by environmental factors (Foerstner et al. 2005; Perry and Beiko 2010) or are simply due to crowding of the signature space (Mrázek 2009); how much consistency can be expected among the fragments taken from a given genome with respect to fragment length; and so forth. It is thought that the signatures arise out of differences in mutation and repair biases (Chen et al. 2004; Lind and Andersson 2008), but they are also clearly entangled with GC content (which can change fairly rapidly and thus carries little phylogenetic signal), amino acid usage biases, and codon usage biases (Vetsigian and Goldenfeld 2009). The fact that horizontally transferred material adopts the signature of its host genome over time (Lawrence and Ochman 1997) supports the view that signatures are maintained by ongoing genome-wide processes.

Lacking a thorough understanding of the mechanisms that drive the evolution of oligonucleotide distributions, we cannot design a classification procedure using them on theoretical grounds. Rather, a wide variety of procedures have been proposed and empirically tested, employing different statistical models of sequence composition, different metrics of distance between sequences, different clustering procedures, and so forth. These methods are typically evaluated by cross-validation, using simulated metagenomic datasets derived from fully sequenced genomes; results from such studies have suggested that classification performance can be quite good (Sandberg et al. 2001; McHardy et al. 2007; McHardy and Rigoutsos 2007; Mavromatis et al. 2007; Zhou et al. 2008; Brady and Salzberg 2009). However, I am concerned that these validations have inadvertently incorporated unrealistic assumptions, and thus that the reported measures of accuracy are substantially higher than can be expected when classifying real environmental datasets.

4.3 Results

4.3.1 Correct binning occurs only at very short phylogenetic distances

In this section I explore the relationship between compositional bias distances and phylogenetic distances, with the goal of building intuition about why compositional binning works at all. I will use the Euclidean distance between tetramer frequency vectors as a prototypical distance metric throughout, because it is widely used and easy to understand (see Materials & Methods 4.5.3). As I discuss in section 4.3.1, the qualitative results of this section hold true for other distance metrics as well.

Compositional biases are consistent within genomes and distinct between genomes

In order for compositional biases to be useful for taxonomic classification, two conditions must hold. First, biases must be consistent between different regions of a given genome. Second, biases must be distinct between genomes.

Many simple random models predict the divergence of compositional bias between entire genomes over time. If sites mutate independently, then the bias of a genome as a whole simply follows brownian random walk through the course of evolution. Since the space of distinguishable compositional biases is fairly large, different populations following random paths will naturally diverge from one another. However, under such a model, we would expect two nonoverlapping regions of a genome to diverge from each other over time as well. Thus, there is no reason to believe a priori that the consistency condition holds.

Figure 4.1 illustrates that both conditions hold between *Escherichia coli* and each of *Halobacterium salinarum* and *Wigglesworthia glossinidia*. First consider the genome *of E. coli* alone, shown in green. The compositional bias distances between sliding windows of 5 kb and the genome as a whole are consistently near zero. The distances between sliding windows on the two other genomes and the *E. coli* genome as a whole are consistently greater than the corresponding distances within *E. coli*. That is, nearly all reads from *H. salinarum* are more distant from *E. coli* than is any read from *E. coli* to itself. This illustrates that the distinctness condition holds for this pair of genomes: given the task of classifying a read into one genome or the other simply by distinguishing those with a signature like *E. coli* from those with a different signature, only a small proportion of reads will have ambiguous assignment.

In Figure 4.2, the magenta histogram shows that the compositional bias distance between many pairs of reads of 50 kb (panel A) or 1kb (panel B) randomly sampled from *E. coli* are also consistently near zero. This demonstrates that nonoverlapping regions of the genome have not diverged from one another in terms of their compositional bias; rather, some mechanism must be acting to maintain a specific compositional bias over the entire length of the genome.

Of course, some pairs of genomes are more distinguishable than others. In addition to the within-*E. coli* histogram, the figure shows histograms of the compositional distances between reads randomly selected from various genomes and reads randomly selected from *E. coli*. Regions where the histograms overlap correspond to compositional bias distances at which classification is ambiguous. When the histograms do not overlap, that indicates that compositional bias distance is sufficient to correctly distinguish reads from the two genomes. For closely related strains, the histograms will overlap completely. In this example, 50-kb reads known to originate from one of the five example species can easily be classified as belonging to *E. coli* or not. As we would expect, shorter reads provide less discrimination between species. With 1-kb reads, substantial overlap is seen between the histograms, with the exception of *Wigglesworthia glossinidia*, which remains nearly perfectly distinguishable from *E. coli*. We might hope that the ease of distingishing genomes increases with phylogenetic distance, but we will see below that this is not the case.

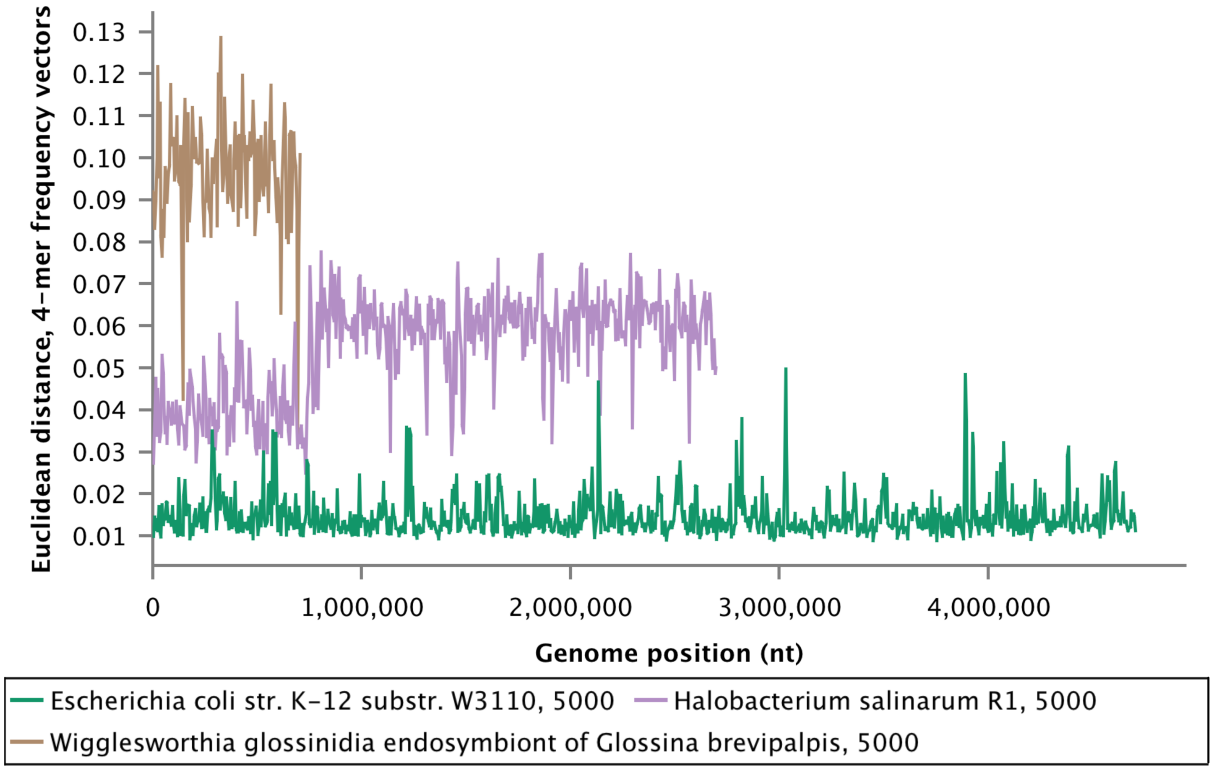

Figure 4.1: Using the genome signature of the entire *E. coli* genome as a target, compositional bias distance from tiled 5kb windows of that genome to the target (green) are consistently less than distances from tiled 5kb windows of other genomes to the target. The overall genome signature is of course the average of the 5kb windows, but (with a few short exceptions) it is not the case that different regions of the E. coli genome have different signatures, since in that case the green line would not be horizontal. The genome of *Halobacterium salinarum* (purple) contains several large plasmids(Pfeiffer et al. 2008) , which here I have given position indexes preceding the main chromosome. The plasmids have substantially different GC content from the main chromosome, resulting in the evident distinction between two regions of different signature. The *Wigglesworthia glossinidia* genome (brown) is very short, as is typical of endosymbionts, and very different from E. coli in tetramer signature.

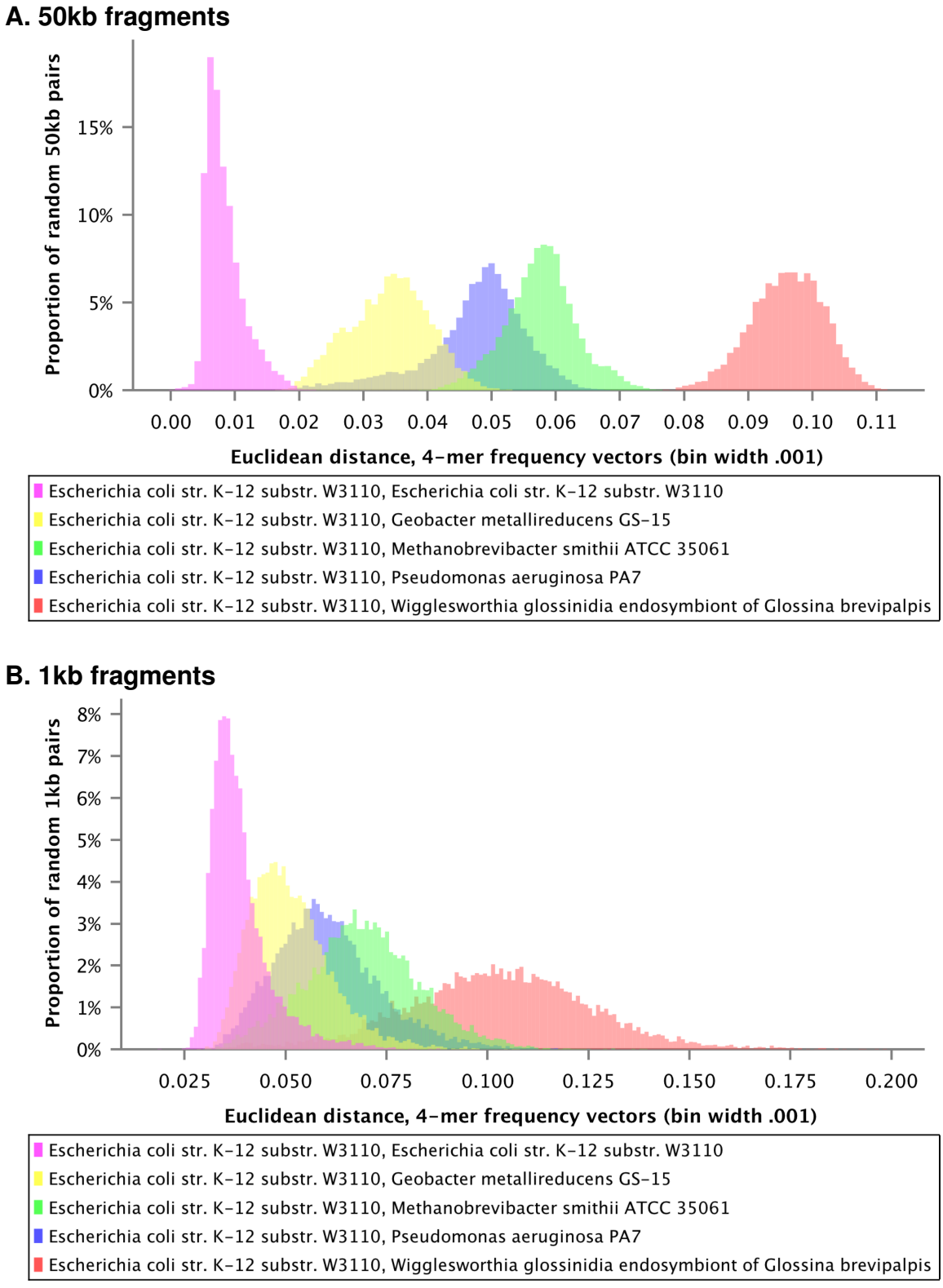

Figure 4.2: Some pairs of genomes are more distinguishable than others.

Absolute compositional bias distances do not correlate with phylogenetic distances

If species have distinctive compositional biases, and if that bias evolves relatively slowly with respect to speciation, then we would expect closely related species to have similar compositional biases. Certainly it must be the case that closely related strains of the same species have nearly identical compositional biases, since they have nearly identical sequences; on the other hand there are cases where different strains of the same "species" share fewer than half of their genes (Welch et al. 2002), so their compositional biases may or may not agree on a whole-genome scale. On the whole we would expect compositional bias distance to increase with phylogenetic distance, at least for short distances up to some correlation length, beyond which all bets are off because enough time has passed for the compositional bias to become unrecognizable.

Various authors have explored this idea, and have even gone so far as to build phylogenetic trees on the basis of compositional bias distance (Pride et al. 2003; Qi et al. 2004a; Chapus et al. 2005; Pride et al. 2006; Gao et al. 2007; Sims et al. 2009).

I investigated the relationship of compositional bias distance to phylogenetic distance by sampling pairs of subsequences from the fully sequenced genomes available from NCBI, computing compositional bias distances, and plotting these against the phylogenetic distance between the genomes. A diagonal trend on such a scatterplot would indicate that phylogenetic distance can be predicted from compositional bias distance.

A representative scatterplot is shown in Figure 4.3. In this example, “reads” of 50kb sampled from the 1190 isolate genomes are compared with full-genome signatures. Because the signatures of short reads are naturally more variable, the use of very long reads here represents a very conservative scenario, where we expect the signature of the read to closely match the signature of its source genome (see section 4.3.1). The distance measure is the Euclidean distance between tetramer frequency vectors, and the phylogenetic tree is the FastTree ribosomal tree (see Materials & Methods 4.5.1).

As can been seen from the plot, there is no correlation between the compositional bias distance and the phylogenetic distance, except at very short distances (i.e., a diagonal tail is visible in the lower left corner). This suggests that compositional biases do indeed diverge through speciation, but that the differences rapidly saturate, so that past a correlation length in the vicinity of 0.1 branch-length units (i.e., nucleotide substitutions per site), they carry no phylogenetic information. Thus, compositional bias distances cannot in general be used to predict phylogenetic distance or to build phylogenetic trees. The plot shows, however, that a very small compositional bias distance implies phylogenetic proximity (e.g., points within a compositional bias distance of 0.1 are nearly always within 0.1 branch-length units). The converse is however less clear; compositional bias distances can be large even when the phylogenetic distance is small.

Environmental sequences may be binned to very closely related reference genomes

Most binning methods to date have used a nearest-neighbor (1-NN) approach; thus binning depends only on the distance to the best bin being shorter than the distances to all the others.

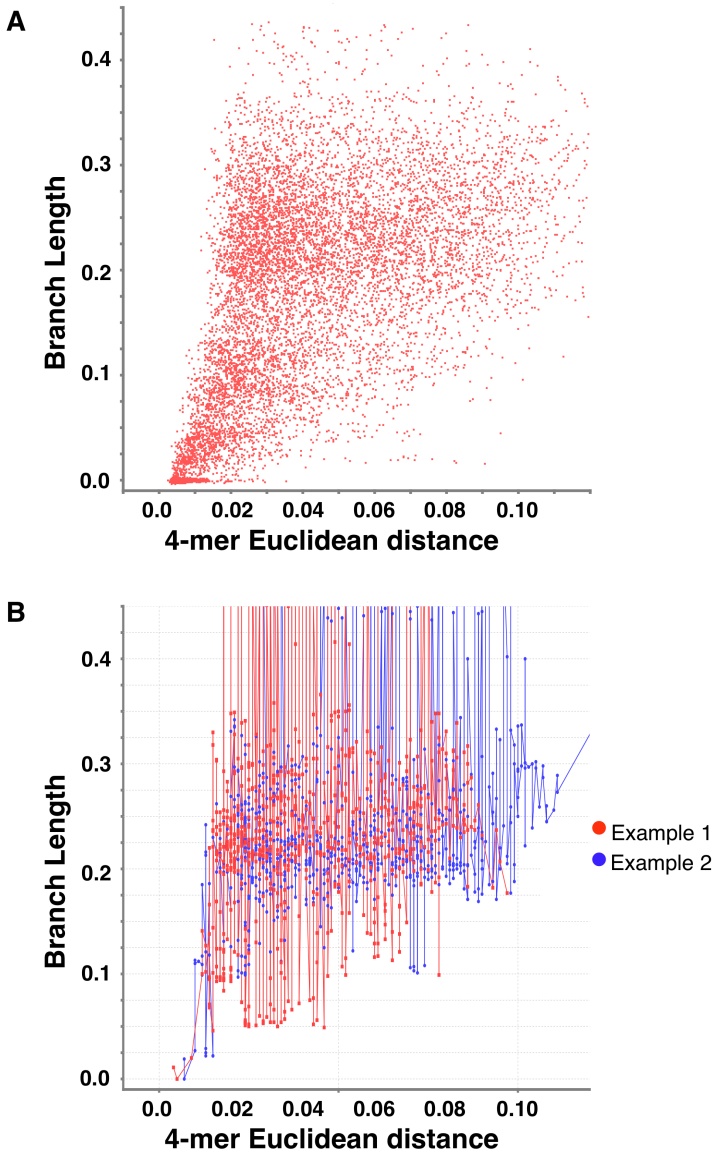

Figure 4.3: A) A representative scatterplot of real phylogenetic distance vs. a compositional distance measure, comparing randomly sampled 50kb reads with full genomes. The sequence pairs were sampled so as to produce a more or less uniform distribution of phylogenetic distances from 0.0 to 0.4. B) Two example scatterplots of real phylogenetic distance vs. compositional distance between a randomly chosen 50kb fragment and all genomes in the dataset. Panel A can in a sense be thought of as the overlaying of thousands of such examples. This suggests that the genome with the smallest compositional distance to a given read is in fact phylogenetically the nearest as well; the fact that these distances differ from one read to the next explains why region near the origin of panel A shows substantial decorrelation. That is: while the *absolute* compositional distance between a read and a genome may not allow predicting phylogenetic proximity, the *relative* distance (i.e., the choice of the nearest genome) is nonetheless informative.

Beyond that, the rank order of the bins is irrelevant. Fortunately that's exactly the situation we have above: the best few bins should be correctly ranked, and all the others have greater distances and are effectively randomly ranked.

Apparently different test reads are more or less similar to genome averages as a whole: that is, although reads from a given genome are generally consistent, some reads are anomalous, in that their signatures are quite divergent from the genome background. In these cases, the compositionally nearest genome may not be the source genome at all, but some other genome– perhaps one related to the source genome if the signature divergence is not too great, or perhaps an essentially random one chosen on the basis of noise.

Figure 4.4 A shows the cumulative distribution of phylogenetic distances between 50kb query sequences randomly sampled from all 1190 genomes and the genome with the most similar signature (i.e., the prediction of a 1-NN classifier). When I permitted matching the source genome itself (or a sister strain of the same species), producing a phylogenetic distance of zero, the correct genome was identified for 85% of the query reads. But, even for such long reads, the best match is more than 0.1 branch-length units distant from the source genome (i.e., in a different family) about 4% of the time.

This is however an unrealistic indicator of how well a binning procedure will work when applied to environmental sequences, because very few of the species in any environment are represented in the set of isolate genomes. The simplest way to account for this fact is a "leave-one-out" procedure (also known as *n*-fold cross-validation, where *n* is the number of data points). In a supervised classification context, this is normally achieved by holding out one genome prior to training, and testing using samples from this held-out genome. This procedure is then repeated for each genome.

For all of the clustering procedures considered here, the training of each bin is independent of the other bins, as is the computation of a distance from a test sample to a bin. Thus, we can train all of the bins once, and achieve the leave-one-out effect in the testing phase simply by refusing to classify a sample to the same bin from which it came—instead choosing the second-closest bin in this case. Figure 4.4 shows that phylogenetic distances between the query seuence and the best reference genome are substantially increased in this case, to the point that the correct genus (i.e., branch length <= 0.05) is identified only 45% of the time. This plot is of course completely dependent on the phylogenetic distribution of the available genomes; for instance, when a genus has only one representative genome, then a query from that genome can never be classified to the correct genus in a leave-one-out setting.

Signature variation within genomes impacts classification performance

I showed in section 4.3.1 that the compositional biases are generally consistent within a given genome. Nonetheless, classification performance may suffer if a genome does not have a consistent signature throughout, but instead contains multiple regions of distinct signature (which may occur for various reasons, most obviously plasmids, viruses, and horizontal transfers), as we saw in the case of *Halobacterium salinarum* (Figure 4.1). Computing the average signature

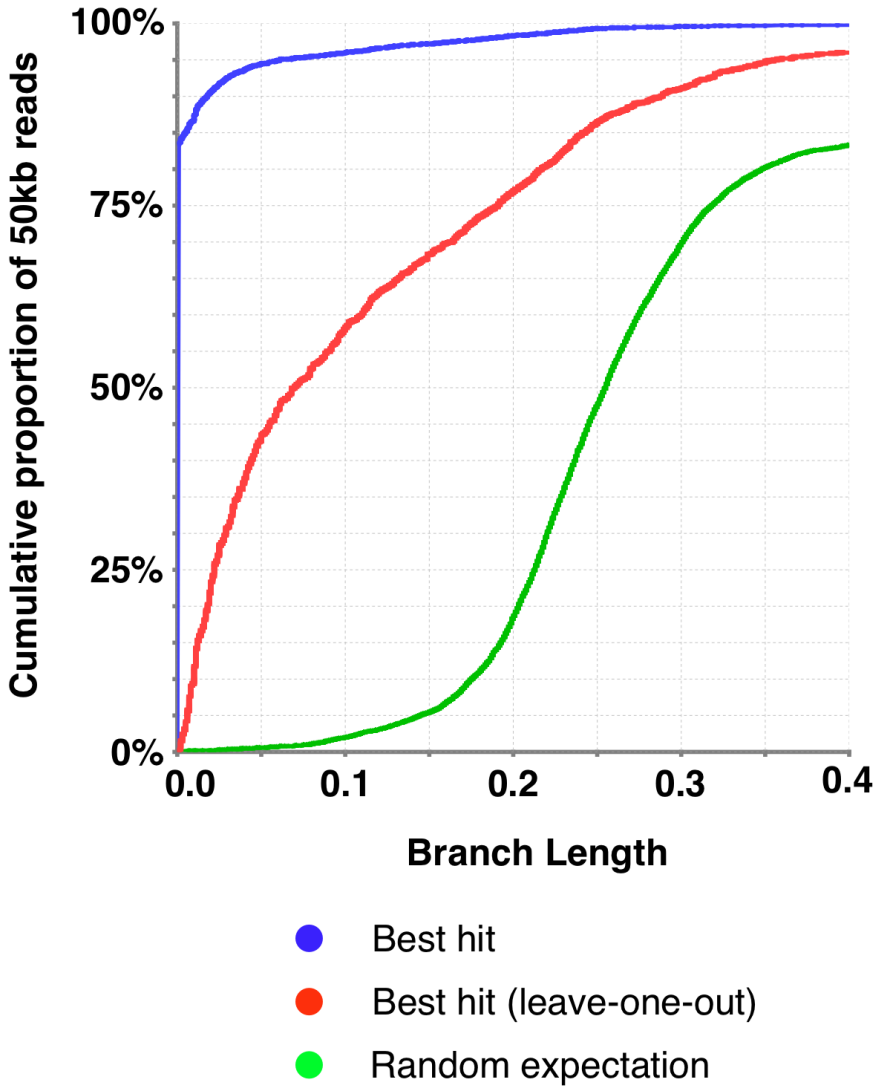

Figure 4.4: (Blue) Cumulative histogram of phylogenetic distances from randomly sampled 50kb fragments to the most similar genome in the dataset, measured as the Euclidean distance between tetramer frequency vectors. The source genome from which the 50kb read was sampled is correctly identified 85% of the time. (Red) Cumulative histogram of phylogenetic distances from randomly sampled 50kb fragments to the most similar genome in the dataset, excluding the source of the fragment. (Green) Random expectation, i.e., the distribution of branch-length distances between pairs of leaves selected randomly and uniformly.

of the entire genome in these cases loses important information. For example, if the signatures are in fact bimodal, then no sample will match the average very well. Conversely, when comparing short reads against each other, this situation would produce short distances when the reads come from the same signature region, and long distances when they come from different regions.

If a genome consists of multiple signature regions, then one might expect the most relevant compositional bias distance for binning a short read to be the *shortest* distance between the read and any one of the regions. To model this situation, I fragmented each genome into 50kb regions, and defined a distance from a read to that genome as the minimum distance to any subsequence (taking care to exclude any sequence overlapping the query, of course).

Indeed, we see in Figure 4.5 that about 4% of the sequences are classified nearer to their source genome when fragmented targets are used than in figure 4.4, where the targets were whole-genome averages. These are of course sequences from the very regions that were previously found to be anomalous with respect to their genome context.

When I use 5kb reads for both the queries and the targets, the effect is even greater: even in the leave-one-out scenario, using 5kb targets instead of whole-genome targets increases the genus-level accuracy from 36% to 43% (Figure 4.6).

(Note that the leave-one-out experiment producing these "accuracy" values is a contrived example that does not reflect performance on real datasets, as I discuss below and in Chapter 5. Nonetheless I believe that the relative increase in accuracy when using multiple targets per genome is a legitimate point that carries over to other settings).

From these results I conclude that, although genome signatures are generally consistent within genomes, enough variation remains that classification can be improved by considering each genome to contain multiple characteristic signatures rather than just one. In the simplest case, this involves computing signatures for tiled windows across each genome for use as targets of a 1-NN classifier. I presume that the principle holds for other types of classifiers (k-NN, SVM, etc.) as well, on the argument that taking the genome average evidently discards valuable information.

The fact that this observation holds for leave-one-out evaluations demonstrates that distinctive patterns in multiple genome regions survive speciation, because genus-level classification of a read from one species is frequently improved by considering multiple regions *in a sister species*. That is, it must be the case that, some of the time, homologous regions in the genomes of two congeners are more similar to each other than to their respective genomes as a whole. We would expect this result, for example, if a plasmid of distinctive composition were present in several related species. More generally, it may be that local signature heterogeneities within the ancestor genome have not yet been overcome in the descendants, because signature divergence after speciation did not occur quickly enough or consistently enough to obscure the preexisting signal. It is important to keep this possibility in mind when performing unsupervised classifications as well, since the resulting clusters may represent related regions in multiple taxa rather than separating reads primarily by taxon.

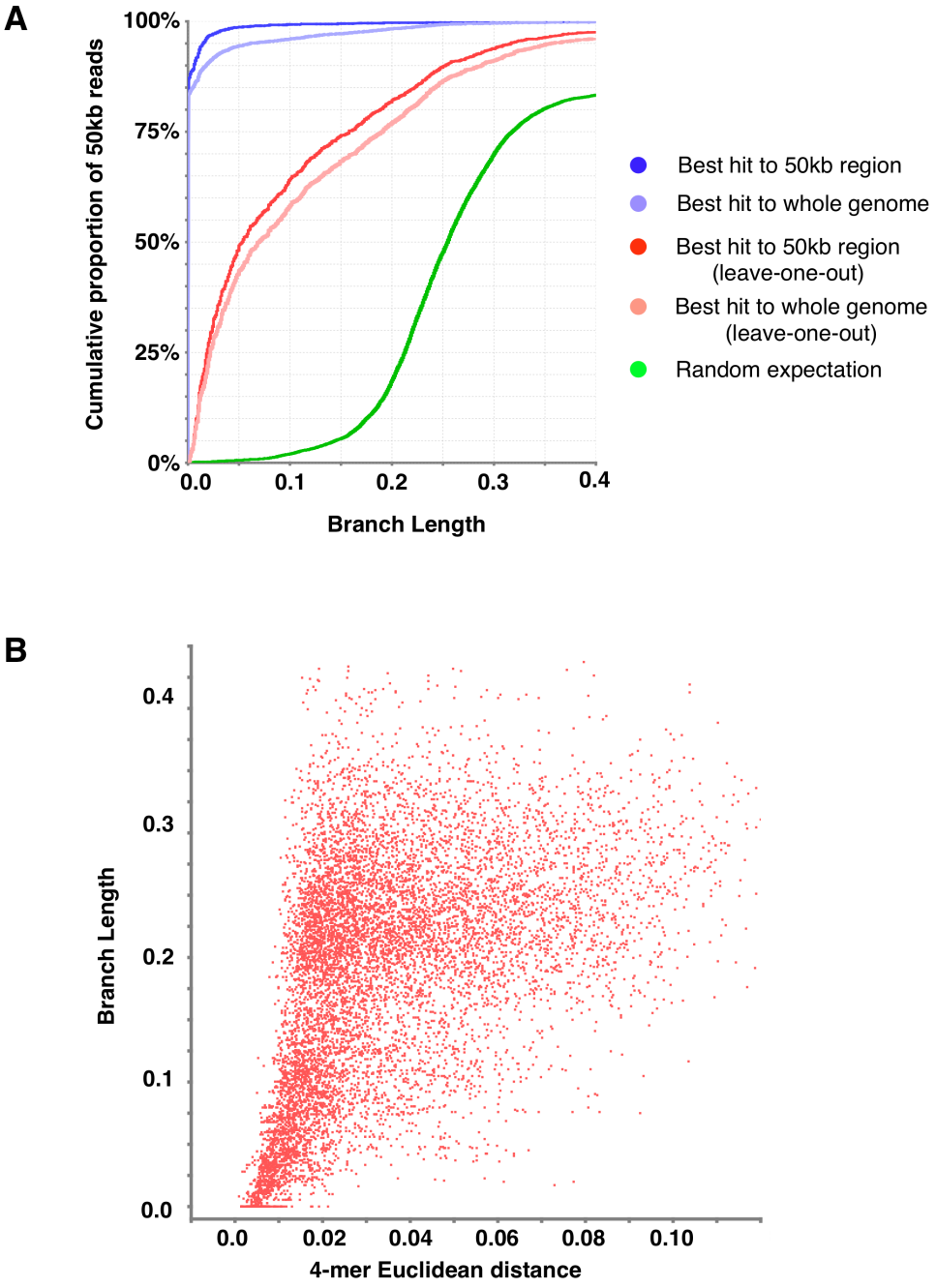

Figure 4.5: 50kb fragment targets provide improved classification of 50kb reads vs. whole-genome targets. A) Cumulative histogram of phylogenetic distances from randomly sampled 50kb fragments to the most similar non-overlapping 50kb fragment in the dataset, according to the Euclidean distance between tetramer frequency vectors. In the leave-one-out experiments, entire genomes were left out as previously. B) A representative scatterplot of real phylogenetic distance vs. a compositional distance measure, comparing pairs of 50kb reads. The sequence pairs were sampled so as to produce a more or less uniform distribution of phylogenetic distances fom 0.0 to 0.4.

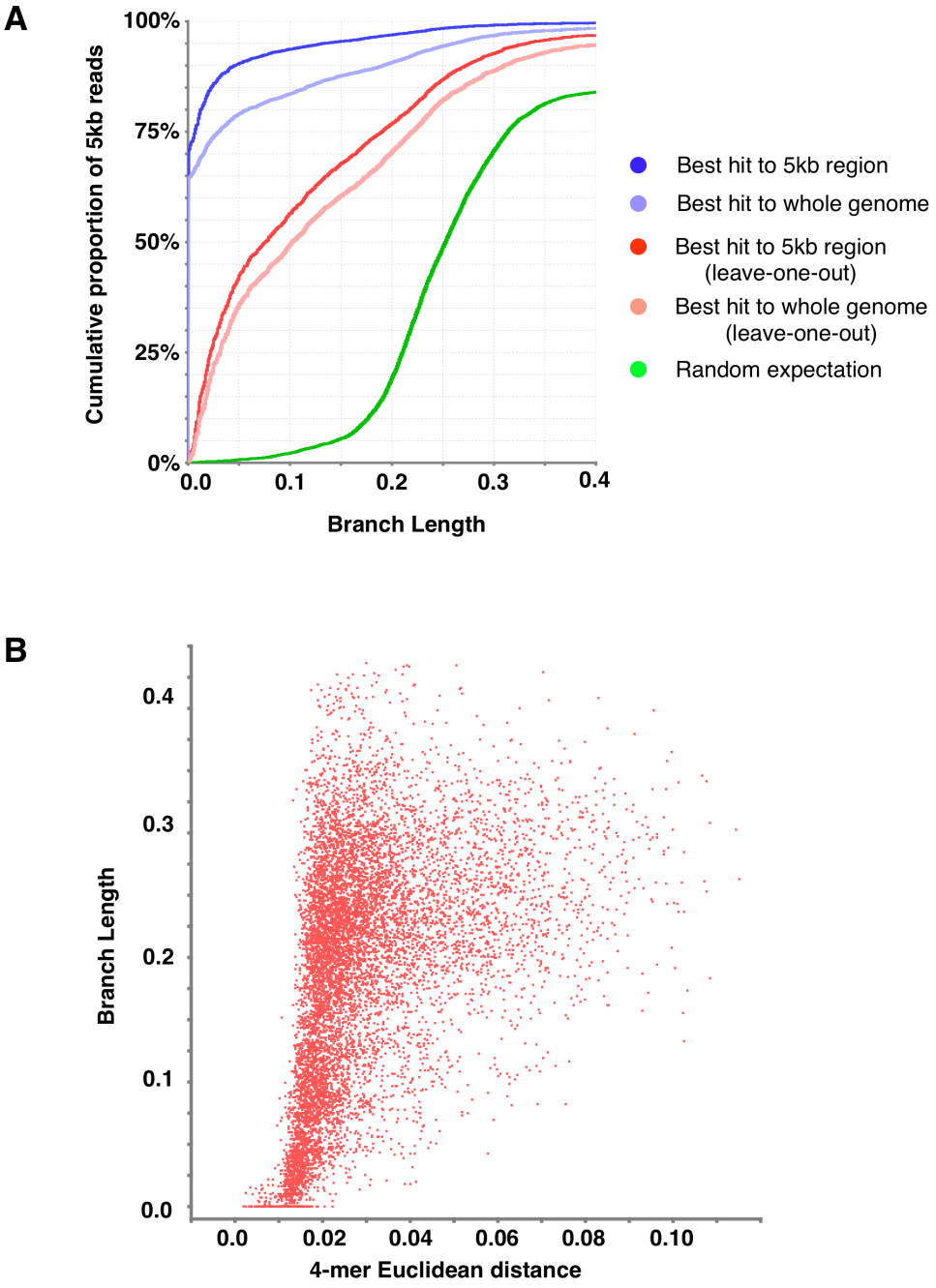

Figure 4.6: 5kb fragment targets provide improved classification of 5kb reads vs. whole-genome targets. A) Cumulative histogram of phylogenetic distances from randomly sampled 5kb fragments to the most similar non-overlapping 5kb fragment in the dataset, according to the Euclidean distance between tetramer frequency vectors. In the leave-one-out experiments, entire genomes were left out as previously. B) A representative scatterplot of real phylogenetic distance vs. a compositional distance measure, comparing pairs of 5kb reads. The sequence pairs were sampled so as to produce a more or less uniform distribution of phylogenetic distances fom 0.0 to 0.4.

Generalization to other distance metrics

I performed the experiments in this chapter hundreds of times, varying all manner of parameters including read length, word length, distance metric, and so forth. To guard against possible anomalies in the phylogenetic tree, I performed the tests using both the FastTree ribosomal tree (see Materials & Methods 4.5.1) and a tree based on conserved protein sequences (Ciccarelli et al. 2006); I also used 16S percent difference values as a direct estimate of phylogenetic distance without reference to a tree. In all cases the qualitative results are the same: there is no correlation between any computed distance based on compositional biases and phylogenetic distance, except at very short distances. A few representative examples using different distance metrics are shown in Figure 4.7. Because I am ultimately concerned with binning performance but not with the strength of these correlations *per se*, I leave the quantative comparison of distance metrics to a future paper (Soergel, 2011, in prep.)

I conclude that compositional biases diverge fairly rapidly, so that they become entirely uncorrelated after 0.05-0.1 branch-length units on the 16S tree, corresponding very roughly to the genus or family level. Thus, binning should not be performed to higher taxonomic levels when no genus (or at least family) representative is present in the training set (i.e., when an attempt to bin at the genus or family level fails).

On the other hand, the signatures rarely converge, so the leaves of the tree have distinct signatures. Because some degree of signature consistency within each genome is maintained, whole-genome signatures may be used to classify sequences at the leaves; but more accurate classifications are obtained when the assumption of consistency is dropped and each target genome is represented as a cloud of signatures.

4.3.2 Fully-sequenced genomes dramatically undersample natural microbial communities

I demonstrated in the previous section that compositional biases are informative about phylogenetic relationships only at short distances, suggesting that a supervised classification procedure will work only if the test samples are phylogenetically near the training bins. The question arises, then, how well different environments are represented by the set of isolate full-genome sequences.

Here I quantify how well isolate genome sequencing to date covers microbiota from eight diverse environments. I take an approach based purely on branch-length distances along a tree relating 16S ribosomal sequences, without considering traditional taxonomic ranks such as “family” and “genus”, to ask the question: given a randomly sampled microbial cell from some environment, what phylogenetic proximity to a fully-sequenced isolate genome can we expect? (Figure 4.8).

I found substantial variation among different environments in how well each has been described by isolate genome sequencing to date. I computed the branch-length distance on the 16S phylogeny between each environmental sequence and its nearest fully-sequenced isolate genome;

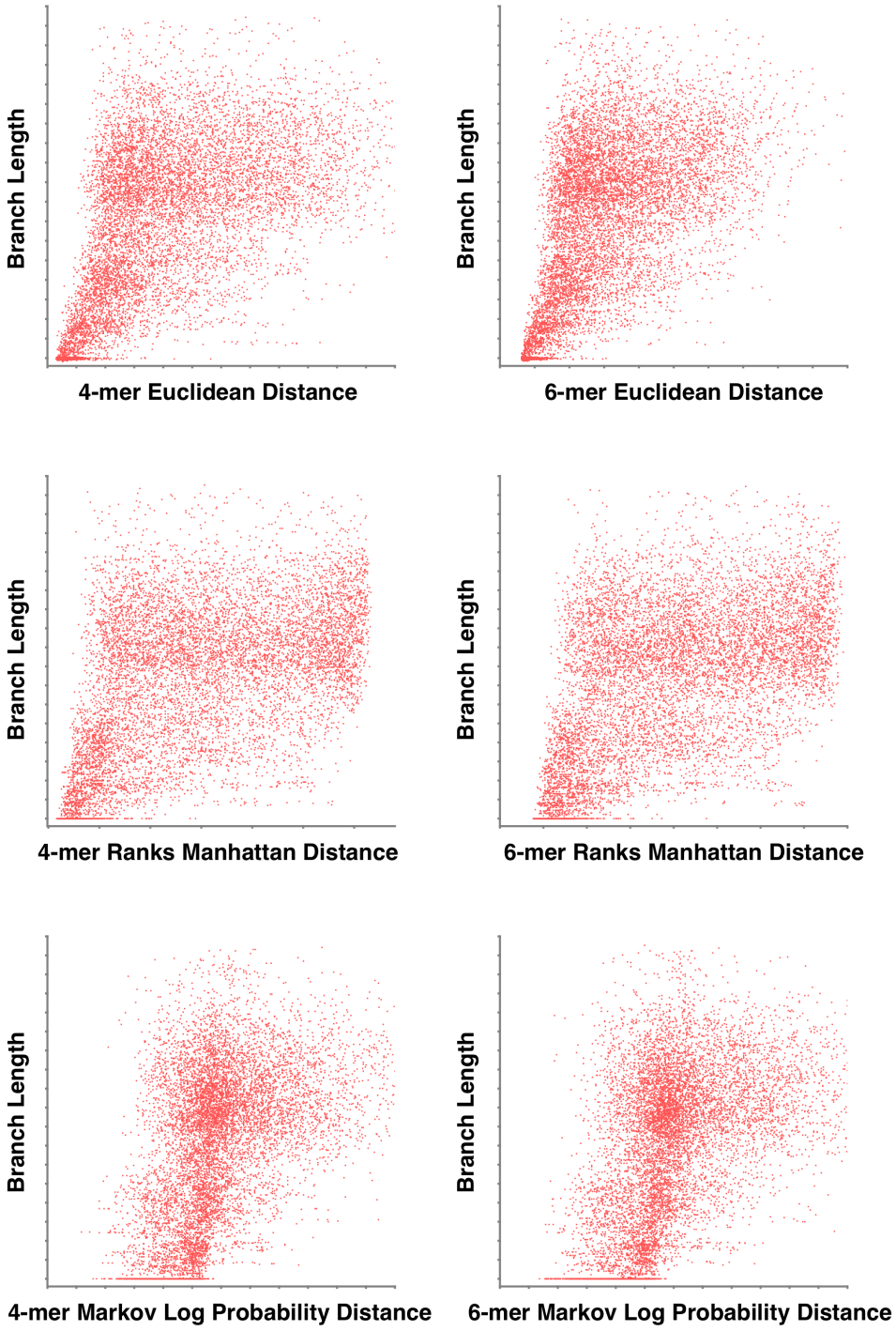

Figure 4.7: Representative scatterplots of real phylogenetic distance vs. several compositional distance measures, comparing 50kb reads with full genomes.

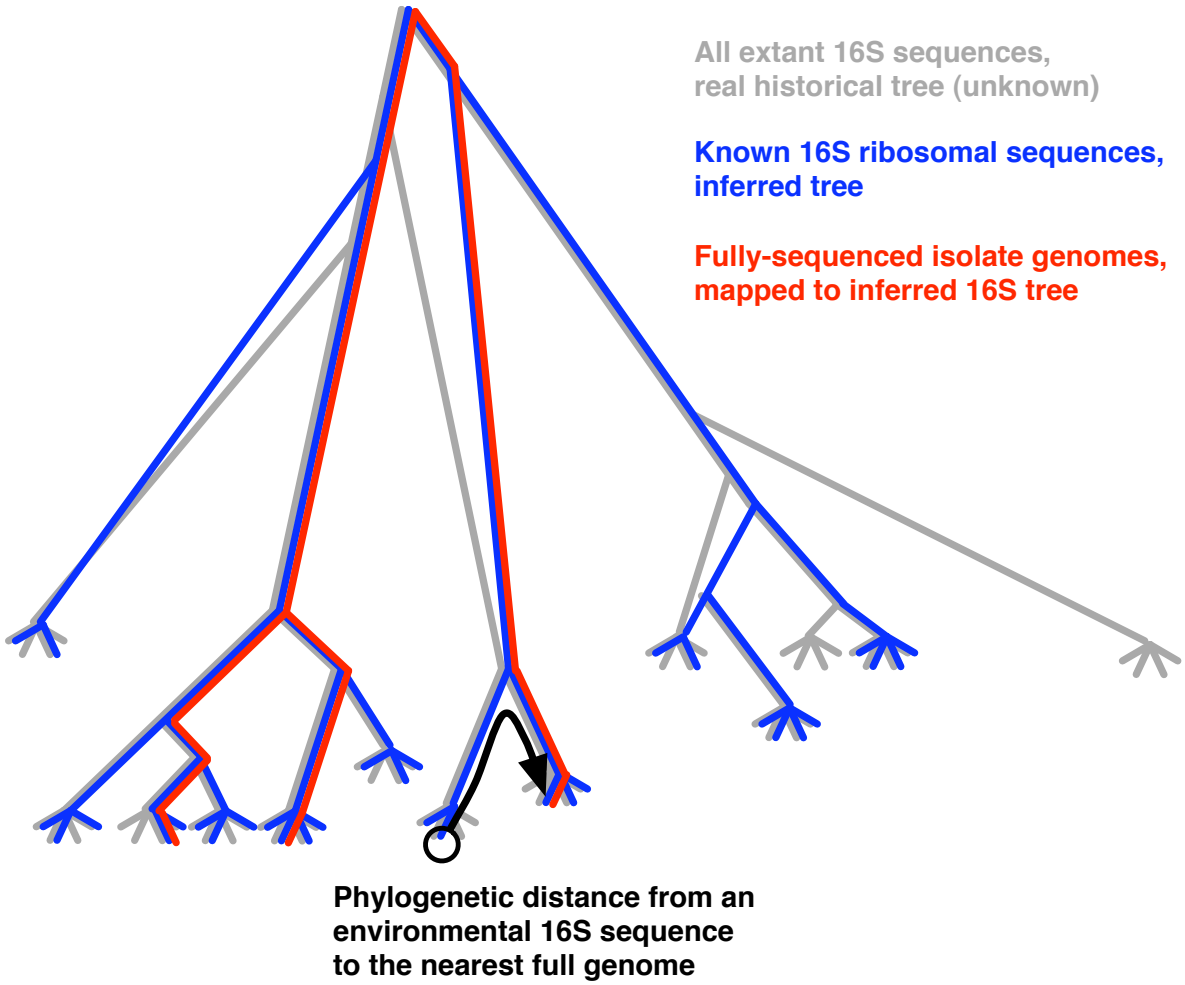

Figure 4.8: Undersampling of the tree of life by isolate full-genome sequences. In this cartoon, the 16S rRNA tree (blue) provides the best available sampling of the real underlying 16S tree (gray), but the tree relating isolate genomes (red) is extremely limited by comparison. The degree of undersampling can be measured by considering the distribution of phylogenetic distances between environmental sequences that can be placed on the 16S tree (e.g., at the leaf circled in black) and their nearest fully-sequenced isolates.

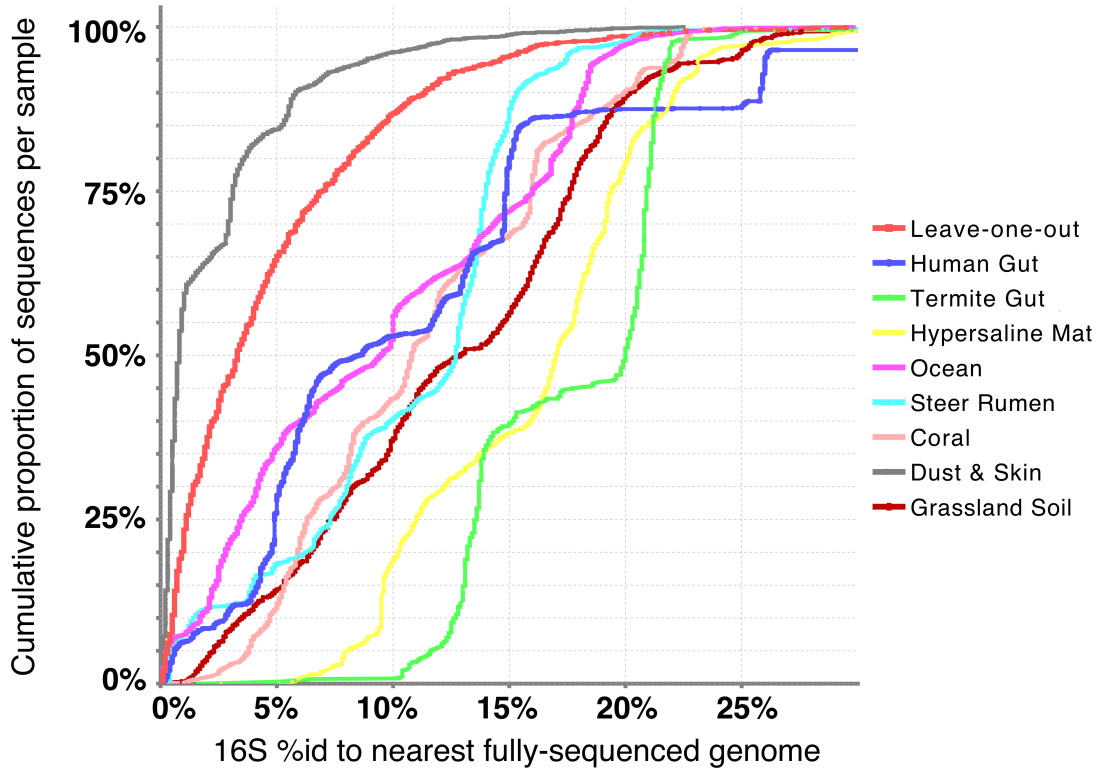

Figure 4.9: The proportion of environmental sequences that are within a given distance of a fully-sequenced isolate genome. This can be thought of as a description of beta (between-community) diversity, related to the lineage-vs-time plot describing alpha (within-community) diversity (Martin, 2002), in that it plots the proportion of lineages in one tree (the query community) that are represented in the other (the reference database) vs. tree depth. We would naturally expect that essentially all environmental strains differ from laboratory strains at the fine-grained level (Acinas et al., 2004) (i.e., that the lines begin at the origin); what this plot shows is that much larger clades and even entire divisions have no sequenced isolates, and that different environments contain these unsequenced taxa in different proportions.

Figure 4.9 shows the cumulative distributions of these distances for eight near-full-length 16S datasets from a variety of environments.

The sample of human skin and mattress dust is unusual in that over 60% of the sequences are from species for which there is an isolate representative. This may be because the isolates are highly enriched in species found on the skin compared to other environments. Even so, this sample seems so anomalous that I would not be surprised to learn of contamination or some unusual bias in the sample preparation; but I did not investigate this further.

In the other samples, between 0% and 22% of sequences are in the same species as a fully-sequenced genome (branch length < 0.03); between 0% and 36% are in a represented genus (branch length < 0.05); and between 1% and 55% are in a represented family (branch length < 0.10).

The results of section 4.3.1 show that taxonomic binning of shotgun reads based on compositional bias has any hope of success only when the 16S branch-length distance between the source organism and the nearest isolate genome is less than 0.05 or at best 0.1; but here I found that, in many environments, at most 55% and as few as 1% of the sequences qualify for binning on this basis; the remaining 45%-99% should therefore be classified "unknown". Binning methods to date have been evaluated using a leave-one-out procedure, which does not reflect the phylogenetic diversity found in real environments, even when higher-level clades are left out.

Neglecting the suspicious human skin sample, no more than about a third of the sequences in even the best-covered environments are from genera with a sequenced representative. Thus we cannot hope to make genus-level predictions for more than about one third of the sequences, even given a perfect classification procedure. In this light claims of 75% or greater accuracy in genus-level classifications (Sandberg et al. 2001; McHardy et al. 2007; McHardy and Rigoutsos 2007; Mavromatis et al. 2007; Zhou et al. 2008; Brady and Salzberg 2009) make no sense.

Slow improvement of coverage over time

As more microbial species have been sequenced over time (Figure 4.10A), the nearest-isolate distances have naturally decreased. Figure 4.10B shows the proportion of 16S sequences from various environments for which a genus representative had been sequenced (defined as a branch-length distance <= 0.05 between a query and a representative), over time (terminating in October 2010 at the same values that can be read for a branch length of 0.05 in Figure 4.9).

With the exception of the human gut sample, the coverage of environments by isolate genomes has remained strikingly stagnant since 2007, despite the doubling in the number of sequenced genera during that time.

The number of sequenced genomes is increasing exponentially, but the number of taxa with respect to tree depth is exponential as well. Thus we would expect, if species were selected for sequencing at random, that the nearest-isolate distances would improve more or less linearly. In fact, of course, the strong bias in organisms chosen for sequencing means that some taxa will be saturated while others remain neglected. The GEBA project aims to correct for this bias by

choosing organisms for sequencing based on their phylogenetic diversity(Wu et al. 2009), and some of the genomes thus chosen are included in the October 2010 dataset, but the impact of this new approach is not yet evident in my results.

In the case of the human gut, a major improvement in coverage resulted from genomes sequenced in 2009. I expect that the ongoing targeted sequencing of gut bacteria as part of the HMP (Turnbaugh et al. 2007; Consortium et al. 2010) will produce similar jumps in future years. The project remains at an early stage, however, and the contribution of HMP genomes during 2010 (to date) did not produce a substantial improvement in genus coverage.

4.4 Discussion

The pace of microbial genome sequencing continues to increase rapidly—so much so that the number of available genomes could easily double between the time of this writing and the date of publication. The present results are based on the full-genome set as of October 2010; more genomes, particularly from more diverse isolates, may improve the situation considerably in the future. At the same time, it is sobering that the improvement in nearest-genome distances has been fairly slow in recent years despite the large number of genomes that have been sequenced. This can be explained first by a simple diminishing-returns argument: given the rapid increase in the number of taxa with respect to tree depth as one approaches the leaves of the tree, ever more genomes are required to achieve the same benefit in terms of branch length. Second, these results may be an indication that the biases in the selection of strains described by Hugenholtz (2002) are still with us: most obviously, only those strains that can be cultured can be fully sequenced, and the pace of novel strain isolations is not increasing very rapidly to the best of my knowledge (especially not for large clades—including even divisions—without any cultured representative.)

It is frequently claimed that, in general, fewer than 1% of microbial cells found in the environment can be cultured (Staley and Konopka 1985; Amann et al. 1995). This belief is at least somewhat called into question by my finding that up to 22% of 16S sequences in the ocean have better than 97% sequence identity with an rRNA from strains that are not only isolated but even fully sequenced. For several other environments, including soil and human gut, the proportion of sequences in the same species as a fully-sequenced isolate was in the range of 4-12%. On the other hand, almost none of the sequences in the termite gut and hypersaline mat samples were even in the same genus as any fully-sequenced genome. The apparent inconsistency between these findings and the 1% claim may be explained by the fact that some strains within a species may be culturable while others are not. Also, the 16S sequence datasets may not sample the cells from an environment uniformly, both due to measurement issues such as PCR bias and due to various forms of filtering and dereplication which, if applied, would alter the apparent relative abundances of types. Nonetheless, it appears overall that coverage of different environments by fully-sequenced genomes (and, presumably, by cultured isolates in general) is highly variable, to the point that the oft-cited 1% figure is not meaningful.

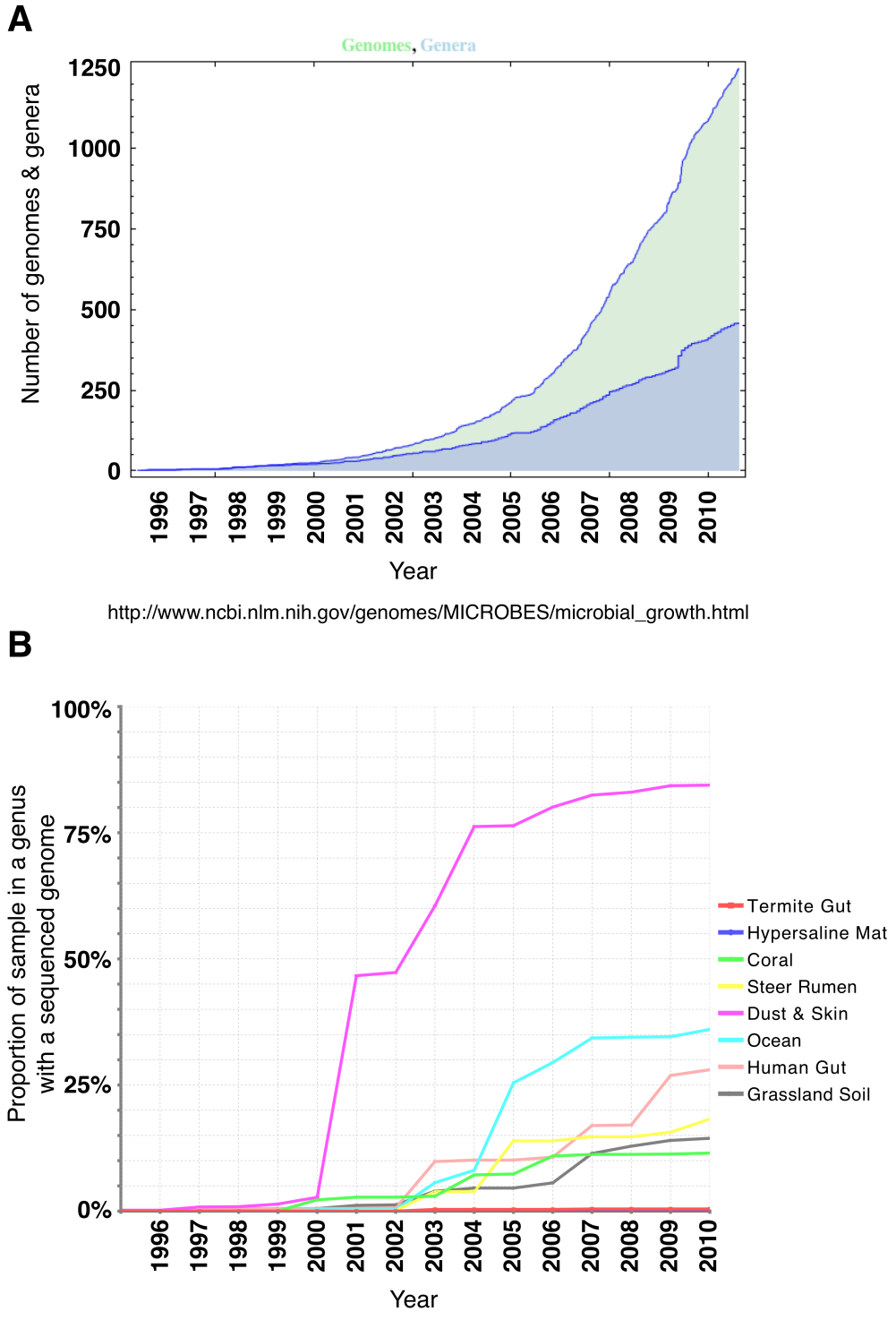

Figure 4.10: The proportion of sequenced genera in many environments is stagnant, despite the rapidly increasing number of total sequenced genera. A) number of genomes and genera sequenced, by year (1995-2010) B) Proportion of sequences from various environmental samples with a fully sequenced congener, by year (1995-2010).

Consequences for evaluating metagenomic analysis methods. Because the available isolate genomes are extremely unevenly distributed over the tree of life, the phylogenetic distance from any isolate genome to its nearest neighbor is likely to be very small. Thus, testing a binning method using the leave-one-out method produces accuracy estimates that are far higher than is realistic for supervised binning of environmental sequences trained on isolate genomes. Ironically, the use of leave-one-out evaluations may have been somewhat more realistic in the past, when the available genomes were fewer and more distant from one another on the tree.

Because binning works only at short evolutionary distances, I find that it is not sensible to try to bin sequences to taxonomic ranks above genus or perhaps family. In all likelihood, reports that such binning can be successful are largely driven by the leave-one-out evaluation approach in combination with the highly uneven distribution of fully-sequenced genomes on the tree of life. In those evaluations, it is artifically easy to bin a sequence at the order or phylum level, because the reference genomes in fact do not represent higher-level taxa as a whole, but rather only a biased subset of the species within the taxon. Because the query sequences are subject to the same biases, they are likely to be much more closely related to a reference genome than can be expected in a natural environment.

This is the reason why some authors recommend training bins using fragments from the environmental dataset under consideration that contain marker genes (McHardy and Rigoutsos 2007; McHardy et al. 2007; Chan et al. 2008a). This produces training bins such that (ideally) the phylogenetic distance from any fragment to a bin is very low, as in the same-genome and leave-one-out cases.

Binning environmental sequences against fully-sequenced genomes may nonetheless be useful to identify those sequences that are likely to be closely related to a known target. In this case, we should expect, depending on the environment, that a relatively small proportion of the sample can be binned at all; the remainder should be labelled "unknown". Binning procedures must therefore be calibrated with a threshold of compositional distance beyond which no classification is made. This threshold should be empirically chosen, based on more realistic simulations (Chapter 5), at a level that makes reasonable tradeoff between the number of sequences that are classified at all and the accuracy of those classifications.

Finally, if phylogenetic proximity is required for compositional binning, it is not clear whether or why compositional binning provides any benefit over alignment-based methods, which are also naturally most accurate at close distances. Past evaluations of alignment-based binning methods have also been done by a leave-one-out procedure and so are subject to the same difficulties discussed above; these experiments should also be repeated in a more realistic setting in order to obtain a valid comparison. In any case, it may be that compositional methods can make classifications that alignment-based methods miss, because of the substantial variation in gene content even among closely related genomes: a reference genome may simply contain no sequence homologous to a query sequence from a sister species or even strain.

4.5 Materials and Methods

4.5.1 Placement of isolate genomes on a large 16S phylogenetic tree

1190 complete isolate genomes were downloaded from NCBI on October 14, 2010. My analyses required a measure of phylogenetic distance among these genomes and between these genomes and environmental species represented by 16S sequences. While measures of distance between whole genomes have been developed based on highly conserved protein-coding genes (Ciccarelli et al. 2006), I instead computed distances purely on the basis of a 16S tree. To build the reference tree, I downloaded the GreenGenes database of ~500,000 aligned 16S rRNA sequences on August 25, 2010. I applied the Lane mask (Lane 1991; Desantis et al. 2006b) to these sequences and then built a tree from them using FastTree 2.1.3MP (Price et al. 2009). Branch length distances between pairs of sequences on the resulting tree are denominated in units of nucleotide substitutions per unmasked site along the full length of the rRNA. Near the leaves of the tree, branch length distances between pairs of leaves roughly equal the percent difference between the sequences (i.e., DNADIST values), as would be expected (Figure 4.11). Tree distances may be shorter than percent difference distances because the tree is based on masked sequence while the DNADIST values are not. Conversely, the branch length distances are often greater than the sequence percent difference, especially at greater distances, presumably because the tree-building process is able to resolve reversions. Evidently, substantially more reversions are thus identified than would be inferred using the Jukes-Cantor correction (which is incorporated in the DNADIST measure); this makes sense because most of the mutations are concentrated in the most variable regions of the sequence that remain after masking. As a result, the tree building procedure infers a greater number of mutations than are evident from the pairwise percent identity score alone.

It should be noted that the 16S tree is not strictly a species tree, partly due to the ongoing controversies regarding the definition of microbial species, and more obviously because a single genome may contain multiple copies of the 16S sequences that can differ by as much as a few percent. Conversely, the tree does not always resolve taxa corresponding to conventional species names. It is not clear to what extent these issues may be due to misannotations in GreenGenes, failure of the tree building approach, or a legitimate inability to cleanly distinguish the species on the basis of their 16S sequences. I observed the same phenomena on the Hugenholtz ARB tree (Hugenholtz 2002; Desantis et al. 2006b), so they are not purely artifacts of the tree building procedure.

With those caveats, I nonetheless wished to place the fully sequenced genomes on the tree, to obtain rough estimates of phylogenetic distances among them and from environmental sequences. To do this, I first extracted all 16S ribosomal sequences from the genomes using RNAMMER (Lagesen et al. 2007). I then used USEARCH (Edgar 2010) to identify nearly identical sequences in GreenGenes. For this task, I first clustered GreenGenes using UCLUST (Edgar 2010) with a 99% identity threshold; the resulting ~143,000 representative sequences served as the reference database. Of the 1190 genomes, most matched exactly one reference cluster with >99% identity (even when multiple rRNA copies were present). A few hit more

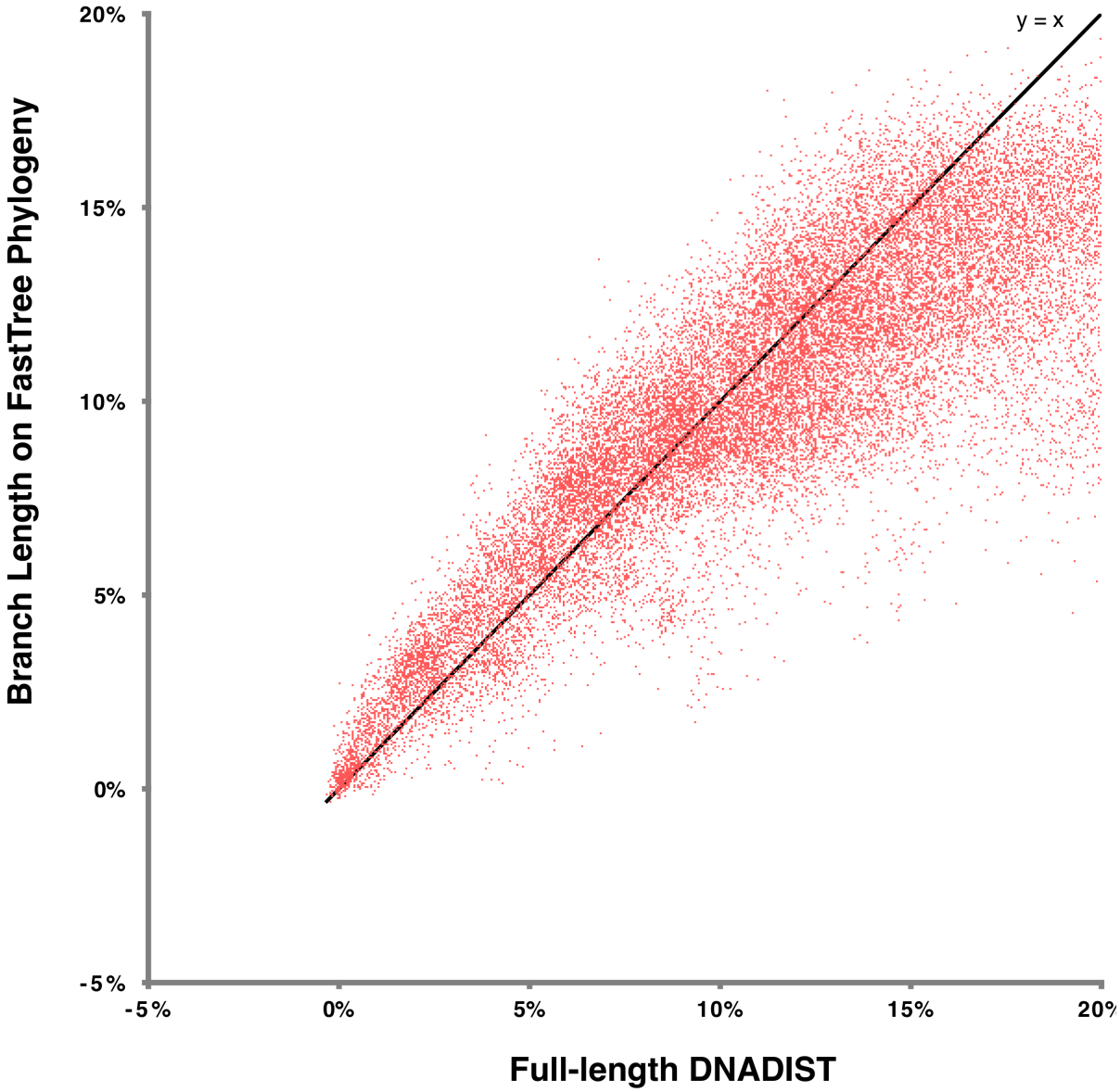

Figure 4.11: The FastTree phylogeny of full-length 16S sequences (based on Lane-masked pairwise DNADIST values) correlates well with the unmasked pairwise DNADIST values. This occurs due to two opposing effects: masking hypervariable regions tends to reduce the percent difference between two sequences, but building a tree in this case tends to produce branch lengths greater than the pairwise distances, because the tree resolves reversions.

than one cluster, indicating that at least two copies were present that differed by more than 1% from each other. Conversely, in many cases, multiple genomes have been sequenced of the same strain or very closely related strains, so these map to the same GreenGenes cluster. In total, 1139 of the genomes were mapped to 808 representative GreenGenes sequences. All of these 808 sequences were used downstream to represent the isolate genomes; thus, in the case of multiple representative matches for a single genome, the phylogenetic distance from a query sequence to the nearest one was chosen to represent the distance to the genome as a whole. The 51 genomes that did not match a GreenGenes cluster were not considered in the downstream analyses.

4.5.2 Computation of phylogenetic distances between sequences

I recognize that the microbial phylogeny is by no means settled, and that there are numerous curators with differing opinions (Desantis et al. 2006b). Here I assume only that the topology produced by FastTree provides some approximation to the real historical tree, so that the branch lengths on this tree are a meaningful (if heuristic) measure of phylogenetic divergence. I computed branch length distances between nodes on the tree using my Phyloutils package (Soergel, 2009). For each left-out genome or environmental sequence, the branch-length distance between its leaf and the leaves associated with the isolate genomes was computed, and the minimum selected. Cumulative distributions of the resulting "nearest-isolate distances" were prepared using my Verdant software (Chapter 6).

4.5.3 Compositional bias distances among genomes and genome fragments

I extracted sequence regions of various lengths from the fully-sequenced genomes, both by tiling the genomes and by sampling randomly within them, as indicated in the text. There are very many possible distance metrics between sequences based on compositional bias; here, for the sake of example, I used the Euclidean distance between tetramer frequency distributions. To compute this, I counted how many times each possible tetramer appeared in each sequence (counting overlapping tetramers as if they were independent), and divided by the total number of tetramers ($n-3$, where n is the sequence length). I made no correction for edge effects at the start and end of each sequence. The "distance" between two sequences was then computed as

$$\sqrt{\sum_{i=0}^{255}(a_i - b_i)^2}$$

where a and b are the frequency distributions found in the two sequences being compared and i indexes the 256 possible 4-nucleotide words.

I am not recommending this distance metric over many other alternatives; its use here is purely a pedagogical device, in that it allows us to demonstrate issues that are common to all compositional distance metrics.

Two additional distance metrics are shown for the sake of comparison in Figure 4.7. The "Ranks Manhattan" distance is computed as follows (Reva and Tümmler 2004). For each sequence, all possible words (i.e., 256 4-mers or 4096 6-mers) are ranked by frequency. The ranks thereby obtained are stored, indexed by word: e.g., if the most frequent 4-mer in sequence x is "accg", then $r(a)_{accg} = 1$. Words of the same frequency are all assigned the same "rank", computed as the average of their original ranks (which were naturally arbitrary with respect to one another). The absolute differences between the ranks for each word are then summed (i.e., the Manhattan distance is then taken between the two rank vectors):

$$\sum_{i=0}^{255} |r(a)_i - r(b)_i|$$

Finally the Markov Log Probability distance is obtained by building a Markov model from one sequence and computing the probability of the other sequence under that model. This measure is asymmetric; I consider the target sequence (i.e., the full genome to which we wish to classify) to provide the model, and compute the probability of the query sequence (i.e., the simulated read). The probability of the sequence is the product of the probabilities of each character within it; these in turn depend on the sequence context. For instance, when using a 4-mer model, the probability of a character is taken to be conditional on the prior three characters. Because the resulting probabilities are extremely small, the computation is done in log space. The log values are increasingly negative as the sequence becomes less probable; to interpret the result as a distance, I simply take the negative. Finally I normalize this value to account for sequence length: in probability space, this normalization would be acheived by taking the geometric mean of the per-character probabilities; in log space, then, I simply divide by the sequence length.

4.5.4 Environmental datasets and taxonomic classification

Sets of 16S sequences from eight diverse environments were extracted from GreenGenes as described in section 2.5.1.

4.5.5 Proportion of environmental sequences in the same genus as a fully-sequenced genome, by year

The year of sequencing of each genome was obtained from NCBI [2]; this data was used to assemble the complete set of genomes available at the end of each year from 1995 through 2009, and in October 2010. The distribution of nearest-isolate distances from the environmental datasets to the genomes available at each time point was computed as described in section 4.5.2, and the proportion obtained.

[2] http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi

Chapter 5

A realistic and consistent methodology for comparison of metagenomic binning methods.

5.1 Abstract

A wide variety of "binning" methods have been proposed to perform taxonomic classification of environmental shotgun sequence reads on the basis of sequence compositional biases, but these have not been comprehensively compared to date. Indeed, since in general each method has been evaluated under different conditions and according to different criteria, it has not been meaningful to compare reported performance metrics such as sensitivity and specificity between papers in the literature. Thus it is a foundational problem in the field to establish a consistent evaluation methodology and to apply it to the whole range of binning methods, so as to make an informed decision about the best method to apply in a given context (perhaps depending both on features of the community and on the biological questions being asked).

The design of an informative evaluation methodology turns out to be surprisingly difficult for two reasons. First, there are very few real data sets that are well enough understood to form the basis of a benchmark. Evaluations are typically performed on simulated data for this reason, but it is not at all clear in turn how to simulate data with realistic properties. In particular, evaluations to date have not adequately considered the phylogenetic distance between query sequences and the training bins, with the consequence that classification accuracy is nearly always overestimated, even when the “leave-one-out” approach is used. I correct this problem by showing how to simulate training and test data so as to produce results mirroring those that can be expected from natural environments.

Second, it is not obvious which metric should be optimized. For instance, measures commonly used to describe multiclass classifiers, such as class-normalized sensitivity and specificity, assume that the class labels are meaningfully chosen, mutually exclusive, and equally important. In the case of phylogenetic classification, the potential labels are hierarchically organized, and

may have dramatically different weights associated with them (whether based on abundance, diversity, or importance by some other measure).

Here I propose the "bep95" measure for the quality of a phylogenetic classification, which is simply the phylogenetic distance (measured on a reference 16S tree) within which 95% of the classifications are correct. The measure incorporates the tradeoff between accuracy and precision: division-level classifications will usually be more accurate than species-level classifications, but are clearly less informative.

In combination, I propose an evaluation procedure which may be applied uniformly to various binning methods to provide a fair comparison, with results that are meaningful to biologists analyzing real data sets. The method may be applied to supervised, self-supervised, and semi-supervised methods. While fully unsupervised methods cannot be evaluated directly, choices prior to the clustering step (i.e. the choices of statistical model, smoothing, and distance metric) may be optimized in a supervised setting, and then applied in the unsupervised setting.

5.2 Results

In order to test the accuracy of the wide variety of binning methods described in chapter **??**, I wish to use a typical strategy of training the classifiers on some data set with known labels and then testing them on another set with known (but hidden) labels, taking care to avoid unfair biases (as would arise, for instance, if the test and training sets were not disjoint). Such a procedure requires a number of choices, which here I will consider in turn. First, what are the classification labels that we wish to predict? Second, how are test and training data separated from one another? Third, how are training points sampled from the underlying training data set, and are the individual samples aggregated into bins that will be used as the classification targets? Fourth, how are test points sampled from the underlying test data set? Finally, how are predicted and actual labels compared with one another?

5.2.1 Choice of classification labels at different levels of the taxonomic hierarchy

To my knowledge, all evaluations of metagenomic classification procedures to date make a binary distinction between "correct" and "wrong" predictions, in order to compute standard measures such as accuracy (i.e., the proportion of correct predictions), sensitivity, and specificity. These do not take phylogenetic distance into account. If the set of target bins includes two closely related species or strains, a read from one that is assigned to the other is counted as "wrong", producing just as large a negative impact on the resulting accuracy measure as if it had been assigned to the wrong kingdom.

Consequently, the resulting accuracy measures are highly sensitive to the number and phylogenetic resolution of the classification labels. Consider a hypothetical procedure that performs perfect binning at the species level, but which cannot distinguish strains. If one target bin is

learned for each species (i.e., from a single isolate genome), then the performance is excellent. But if the genomes of ten isolate strains are available for each species (a circumstance that is already true of several species such as *Escherichia coli*, *Staphylococcus aureus*, and *Prochlorococcus marinus*, and which will rapidly become more common), and if each available genome carries a distinct label, then nine times out of ten the procedure will choose the wrong strain by chance—so the "accuracy" will plummet to 10%.

One way of addressing this situation is to perform the classification at a higher level of the tree. In the typical evaluations, bins are trained and tested separately at each taxonomic rank from genus through division. Usually the result is that classification to higher ranks is more accurate (i.e., previous studies report that it is easier to choose the right division than the right genus (Sandberg et al. 2001; McHardy et al. 2007; McHardy and Rigoutsos 2007; Mavromatis et al. 2007; Zhou et al. 2008; Brady and Salzberg 2009)). However, Chapter 4 shows that compositional binning works best when the test sequences are phylogenetically very near training bins, and it is well known (and I quantify below) that the the set of available isolate genomes is highly biased. In combination, these facts suggests that apparent good classification performance at high levels of the tree may be at least partly an artifact of the specific choices of test and training sets that were used. In particular, it will be easier to choose the correct division if both the test and the training samples from that division do not represent its full diversity but instead are largely drawn from a few families.

It is also worth remembering that the tree relating the isolate genomes is not well established. Usually taxonomic labels are chosen at specific ranks (e.g. genus, family, etc.) according to a curated taxonomy, such as those provided by the RDP or GreenGenes. An alternate approach would be to choose internal nodes on a phylogenetic tree for use as labels. This would allow choosing the tree level more continuously rather then from only a few discrete options, and would lend more confidence that each label has more or less the same phylogenetic scope, in contrast to the traditional taxonomic names (which, for historical reasons, may name taxa of widely differing internal diversity at the same rank (Cohan 2002; Gevers et al. 2005; Konstantinidis and Tiedje 2005, 2007)). On the other hand, the accuracy of reconstructed phylogenies with respect to the true historical tree remains quite uncertain.

In short, accuracy scores from binning evaluations are meaningless unless one also knows which set of classification labels was used, as well as the distribution of phylogenetic distances between test samples and training bins. Thus, such scores cannot be meaningfully compared between evaluations that employed different training and test sets.

5.2.2 Separation of isolate genomes into test and training sets so as to produce realistic performance estimates

Binning methods are typically trained on fully-sequenced isolate genomes. With the exception of the Acid Mine Drainage studies (Tyson et al. 2004; Denef et al. 2010b), there are no environmental shotgun datasets for which the taxonomic assignment of each sequence read is certain. Since real datasets can therefore not be used as test sets, simulated datasets are produced for

testing by sampling sequence reads from fully-sequenced isolates. The question arises, then, how best to perform this sampling so as to estimate the performance of the binning methods that can be expected on real data.

Leave-one-out cross-validation

A common strategy is the "leave-one-out" method (aka *n*-fold cross-validation), in which bins are trained using all of the isolate genomes but one, and testing is done using reads sampled from the held-out genome. The process is then repeated, holding out different genomes in turn.

The leave-one-out method is overly optimistic. I showed in Chapter 4 that compositional distance measures correlate with phylogenetic distance only when the two sequences in question come from very closely related organisms; if the computed distance is greater than some (fairly low) threshold, then all bets are off regarding the phylogenetic proximity. Thus, I expect that the accuracy of a binning procedure will be extremely sensitive to the distribution of phylogenetic distances between test samples and training bins. For instance, if all of the test samples come from the same species as were used for training, then we can expect to do very well; and if the test samples are phylogenetically distant from the training bins, then the classification results will be largely random.

For this reason I am concerned that the leave-one-out strategy may give unrealistic results, depending on the taxonomic rank at which test samples are left out. If one genome is left out, for instance, but another strain of the same species is available, then the binning performance for that purportedly held-out genome will be excellent. Previous studies have generally been careful to leave out all strains of the species being tested, but have not done the leaving-out at higher taxonomic ranks (e.g., at the genus level). Because of the dramatic bias in which taxa have isolate genomes available, the result is that, regardless of which genome is held out, there is usually a genome available for a closely related species—at least, more closely related than would be expected by chance. Consequently, evaluations based on leaving out individual species are likely to be overly optimistic about the accuracy that a binning procedure may achieve.

Varying the leave-one-out level. When a leave-one-out evaluation is performed, a taxonomic level must be selected at which the leaving-out occurs that is lower than the classification level (otherwise no correct classification could be achieved). Usually this is done at the species level regardless of the classification level, e.g. to see whether the correct division can be predicted when there is no species match. However, Chapter 4 suggests that we should also be concerned about whether the correct division can be predicted when the test sample comes from an order or class that is not represented in the training set. Varying the leave-one-out level would thus provide a means of testing the correlation length of compositional biases along the tree, for instance to determine whether samples from different families within a given order share

identifiable signature patterns. However, it seems clear from Chapter 4 that this is not the case, so I did not pursue this question further.

Choosing the leave-one-out level can be thought of as setting a threshold phylogenetic distance between a test sample and a training bin, below which I do not make a classification on the grounds that such proximity is likely an artifact of database bias. However, setting the level too high in the tree–in particular, higher than the correlation length of compositional biases–amounts to an assumption that no environmental sequences resemble isolate genomes at all, with the result that no classifications can be made.

Cross-validation with random subsamples

In fact, the real distribution of phylogenetic distances between environmental sequences and isolate genomes varies continuously, so no one discrete cutoff is realistic. I therefore account for this concern differently, by randomly selecting some number of species to hold out. In general, the greater the number of held-out species, the greater the phylogenetic distances between the test samples and the nearest training bin, because the remaining training genomes are ever more sparse on the tree. Thus, by choosing this number appropriately, it is possible to obtain a distance distribution for use in testing that resembles the distance distribution expected in a real environment, and thereby overcomes the overoptimism of the leave-one-out approach.

I do in addition recommend leaving-out at the level of individual strains (i.e., merging genomes at a phylogenetic distance of 0.0), in order to avoid testing a sample against its source genome or against a technical replicate of the same genome.

Starting with the set of fully-sequenced isolate genomes, I held out test sets of different sizes and measured the branch-length distance on the 16S tree between each held-out genome and the nearest remaining training genome, excluding strain-level matches. Figure 5.1 shows the distributions of such distances obtained by aggregating 100 replicates for each test-set size.

In Figure 5.2, selected curves from this simulation are overlaid on distance distributions expected from real environment types, as previously described (Section 4.3.2). The purpose of this plot was to establish an analogy "*808 : real environments :: x : (808 - x)*" regarding the relationship between the training and test sets. That is, given that real environments are to be classified using all 808 available training bins, how should we partition those same 808 species into training and test sets for evaluation? The hypersaline mat and termite gut samples have essentially no fully-sequenced congeners; on the basis of chapter 4, then, classification of these samples is hopeless. The dust and skin sample is anomalously even easier to classify than a leave-one-out evaluation would suggest. For the remaining samples, the overlaid curves show that simulations using training sets of 25-200 species produce nearest-isolate distributions that are very roughly comparable to the distributions from real environments. In order to simulate an environmental dataset from isolate genomes in such a way that the nearest-isolate distance is realistic, then, we should choose between 25 and 200 genomes from the isolate set to act as training bins, and then test using the remainder.

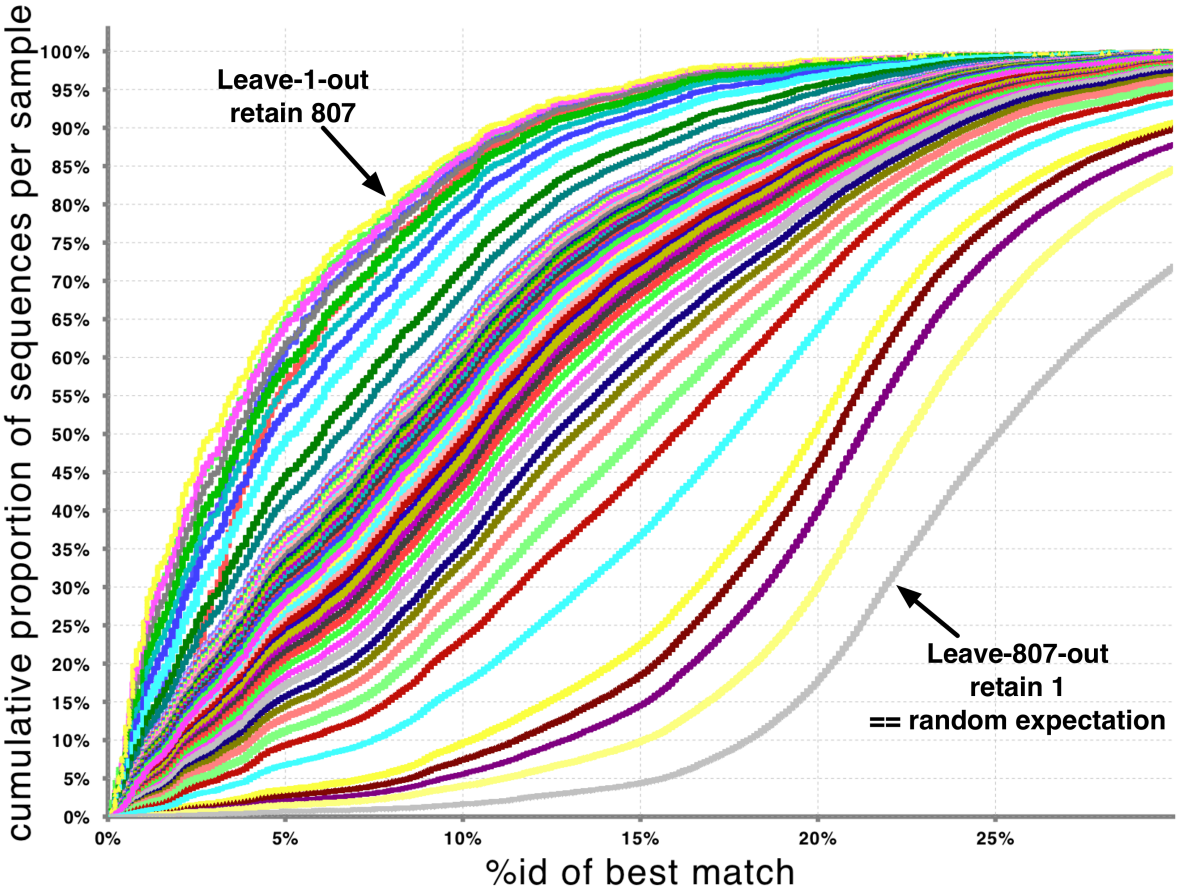

Figure 5.1: Distributions of phylogenetic distances between query and target species in leave-n-out simulations, for many values of n (averaged over 100 replicates each). Starting from 808 taxa, a leave-one-out experiment has 807 targets, and thus produces small distances; naturally, choosing fewer targets (e.g., 100 targets == leave-708-out) produces greater distances.

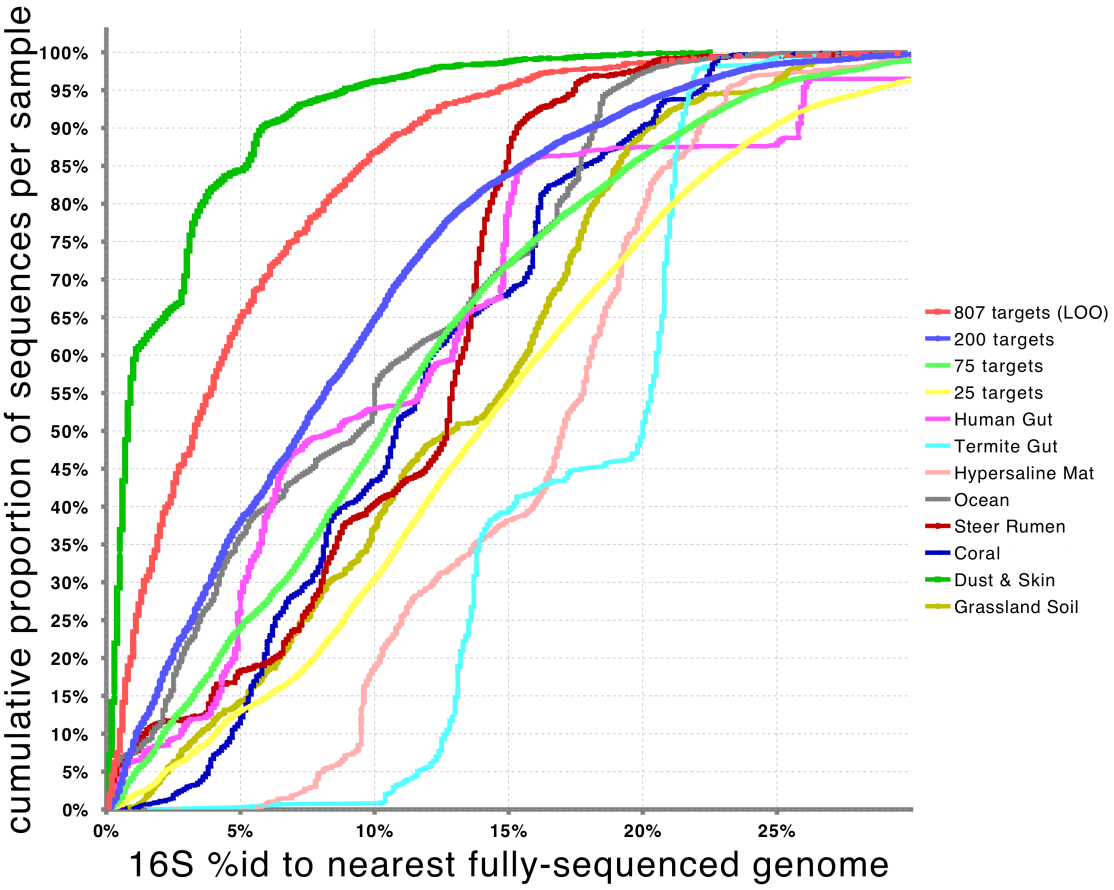

Figure 5.2: The distribution of phylogenetic distances between environmental query sequences and reference sequences can be simulated by choosing different numbers of target species. Cumulative distributions of phylogenetic distances between real environmental samples and isolate genomes are shown, overlaid with four simulated curves selected from figure 5.1. A leave-one-out experiment (red) effectively assumes that environmental sequences are more closely related to fully-sequenced isolates than is realistic in most environments. Instead, the distributions observed in many of the environmental samples can be roughly simulated by choosing only 25-200 target species to represent the reference database, and by then extracting simulated query sequences from the remaining 608-783 species.

One might raise the objection that we cannot possibly achieve accurate binning at the genus level with only 25 target bins; indeed the choice of 25 targets will surely neglect entire divisions. This is exactly the point: figure 5.2 shows that, when classifying real environments using all 808 training bins, nearly all genus-level taxa and even some divisions present in the environment will not be represented in the training set; it is this mismatch between the test and training sets that I wish to simulate.

5.2.3 Aggregating and sampling the training data

Once the test and training genomes have been partitioned, the next step is to use the training data to learn models of sequence composition to act as targets in the classification procedure; I refer to these as "training bins". Most evaluations to date have employed one training bin per classification label. For instance, when classifying samples into divisions, all training samples from each division are used to train a single model of sequence composition for that division. If the nature of this aggregation is effectively to average the genome signatures of the training samples (i.e., by representing the bin as the centroid of the samples), then this averaging is likely to eliminate any phylogenetic signal that was present, by increasing the compositional bias distances from individual samples to the centroid past the noise threshold (Chapter 4).

Varying the training and classification levels independently. Thus, an alternate approach would be to perform the classification initially using fine-grained bins (e.g., at the species or genus level), but to assign only the corresponding course-grained labels (e.g. at the division level) to the test samples. The results of Chapter 4suggest that using multiple target bins per classification label in this way can produce substantially more accurate predictions. The benefit is likely to be smaller when the process of aggregating training samples into a bin in some way remembers their internal distribution, as in the case of the SVM classifier.

The training bins may be even finer-grained than individual genomes, which would be useful if the genomes are subdivided into regions of potentially different signature. Carried to the extreme, the use of fine-grained training bins produces the 1-nearest-neighbor classifier, where effectively every read produces a distinct training bin. I found in Section 4.3.1 that doing this is likely to improve binning accuracy on the whole. However, we have also seen that randomly selected pairs of reads may have similar composition despite being phylogenetically distant. This noise may be ameliorated through the use of a k-nearest-neighbor classifier, which considers many compositionally nearby reads and can thereby overrule the few anomalous ones. The same effect can be achieved with an SVM classifier, where anomalous reads can be overruled during training through the use of a "soft margin" (Noble 2006).

Balancing training classes. One might be concerned that the phylogenetic bias in the set of isolate genomes will result in a biased set of training bins, thereby imposing an unrealistic prior on the classifier. One approach to correcting this bias is to factor it out as part of the classification procedure, as in the case of the naïve Bayesian classifier. A second approach,

which I take for the sake of consistent evaluation across classifier types, is simply to ensure up front that each bin is trained on the same number of samples.

The goal of both approaches is to assume a uniform prior on the training bins. Of course, a prior that is uniform with respect to genus labels will be highly nonuniform with respect to division labels, because some divisions contain many genera represented by isolate genomes, while others contain few or none.

How much training sequence is enough? In light of the above issues, my typical procedure is to choose the phylogenetic level of training bins (e.g., the species level), then to aggregate all training sequence that might contribute to each bin (e.g., the genomes of multiple strains in the same species), and finally to sample a consistent number of training reads from each sequence pool. Because of the consistency of compositional biases within each genome, it is not necessary to train a species-level bin using all of the available sequence. I found that training using 100 kb of sequence per species produced classification performance nearly equalling that seen when training using entire genomes (data not shown). For my evaluations I therefore adopted a conservative standard of training using 100 reads of 10 kb each per species (1 Mb total). This sampling procedure corrected for differences in genome size between species, and sped up the training phase of the computations.

5.2.4 Sampling the test data

The phylogenetic distribution of the test samples may affect the outcome of the evaluation, depending on which scoring scheme is used. In the simplest case, if test samples are drawn uniformly from all of the available genomes, then they will be strongly biased towards some clades and against others. Total classification accuracy is likely to be artificially high in this case, because the test samples will tend to come from the very classes that are easiest to distinguish.

"Class-normalized" measures of accuracy, sensitivity, and specificity attempt to correct for this effect by giving each class an equal contribution to the total score, regardless of the number of samples in it. When using such a measure we need to ensure that enough test samples are taken from each class that its class-specific sensitivity and specificity are accurately measured. This is not the case when genomes are sampled uniformly, because the number of genomes per division (for example) is extremely unbalanced.

For this reason I allow choosing a phylogenetic level at which uniform sampling of test sequences takes place. When using a class-normalized score, it makes the most sense for this level to match the prediction level: for instance, if we are predicting division labels, then we should test with equally many samples from each division. In this case, genomes are sampled uniformly within each balanced class.

The choice of training and test sets described above is designed to produce a realistic estimate of total classification accuracy. Note that while the real environments being simulated surely contained species in different abundances, the simulations of section 5.2.2 used uniform distributions of the test species. Thus, we must now provide test samples drawn uniformly from the

test genomes in order to simulate a realistic distribution of phylogenetic distances to the nearest training bin (i.e., in order to maintain the validity of the analogy esablished in that section). This approach is related to the argument that the diversity of a community with many unevenly distributed species can, for many purposes, be described as having a smaller "effective number of species", taken to be uniformly distributed (Jost 2006, 2007).

5.2.5 Summary of labelling choices

In summary, we can choose label sets at different taxonomic levels for four distinct purposes (Figure 5.3): balancing the test samples, choosing the training bins (and balancing the training samples), prohibiting classifying a test sample to a bin that is too closely related (the leave-one-out process), and finally making predictions and measuring their accuracy. In combination, these choices encompass most of the variation among evaluation methodologies employed in the literature to date.

5.2.6 A phylogenetic evaluation metric

Classification methods are typically evaluated using a test set consisting of data for which the correct class is known; samples from this set are classified and the predictions compared to the right answer. In the case where the classes are related by a phylogenetic tree, we need not call predictions absolutely "correct" or "wrong"; rather, we have a natural continuous measure of wrongness, namely the sum of the branch lengths between the predicted taxon and the true taxon. We call this quantity the "binning error with respect to phylogeny" or "bep". In computing these values, we always consider the true taxon to be a leaf on the tree (i.e., a strain), while the predicted taxon may be an internal node at any level of the tree (Fig 1a,b). After classifying a test set, we can plot a histogram of the bep distances, or better yet a cumulative histogram. When the error distribution is plotted for a single species, the result is a plot of class-conditional sensitivity vs. phylogenetic resolution. We call this the "bep profile" of the class. Plotted for an entire test set, this plot shows the proportion of test samples classified correctly within a given phylogenetic distance (Figure 5.5).

A common complaint about binning methods to date is that they are insufficiently accurate—a belief that arises from evaluations which fix a set of target labels and then report (for example) 85% correct classification. The bep plot allows us to ask the reverse question: given that we require a certain level of confidence in our predictions, what phylogenetic resolution can we achieve? Let us assume that for many purposes a confidence of 95% is desirable. Using the bep plot for a given binning procedure, we can simply read off the phylogenetic distance within which 95% of the predictions are correct. I call this number "bep95" and use it as a single score describing the phylogenetic precision of a procedure.

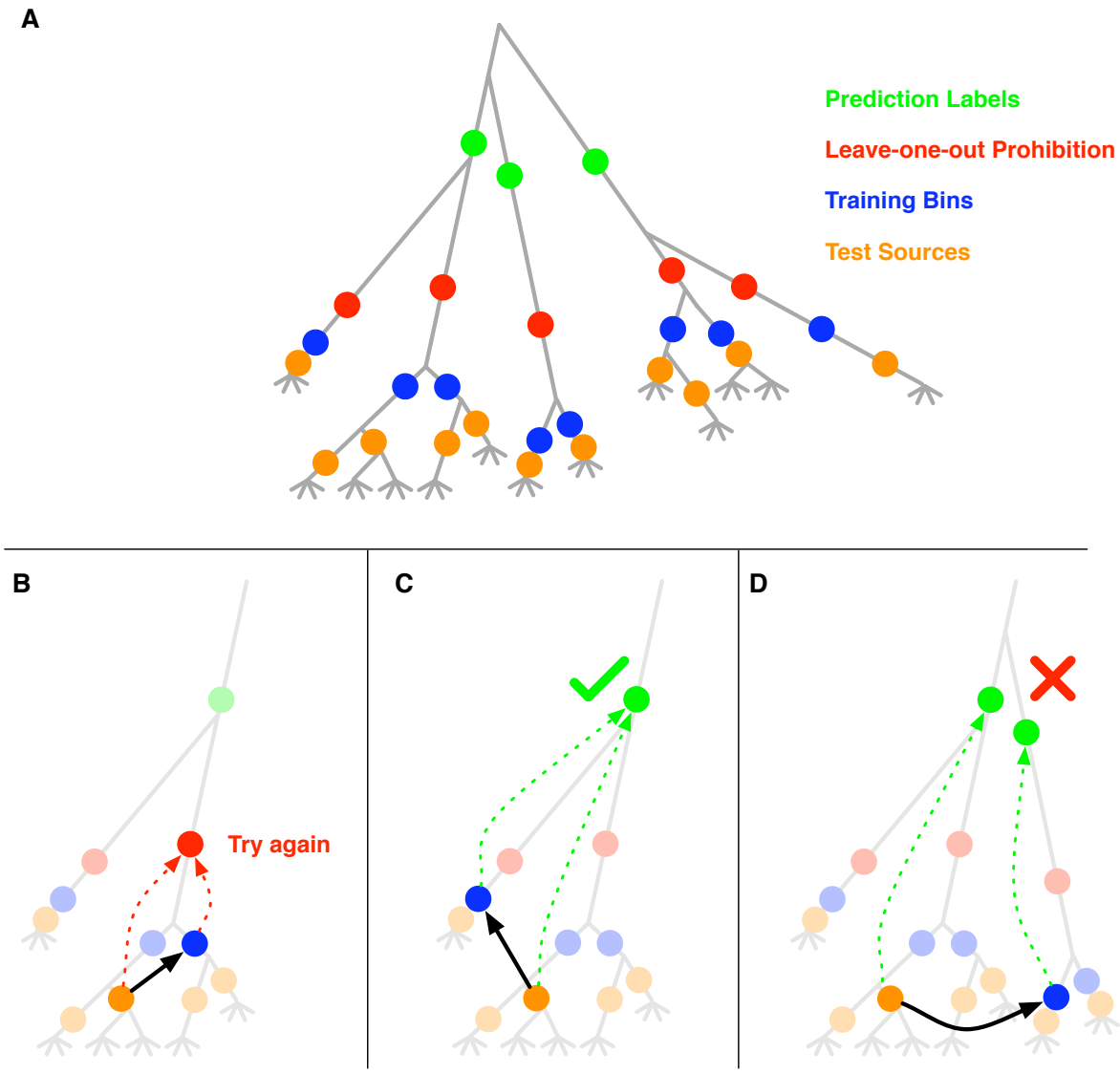

Figure 5.3: The label-based evaluation infrastructure. A) Sets of labels are defined for different purposes at different levels of the tree. B) A test sample is initially classified to a training bin that is prohibited by the leave-one-out label; this classification is rejected. C) The test sample is instead classified to the next-best training bin, producing a correct prediction. D) The test sample is classified to a training bin associated with the wrong prediction label.

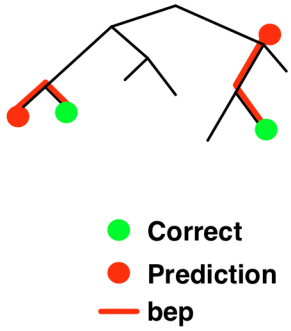

Figure 5.4: The "bep" score. When classifying a single sample, the binning error with respect to phylogeny ("bep") is simply the branch length between the predicted taxon (red) and the true taxon (green). A classification procedure may choose an internal node (i.e., a higher-level taxon); this incurs a bep cost even if the true taxon is a descendant of the predicted taxon, but this cost may be less than the cost of a misclassification to a leaf of the tree. The bep value thus incorporates a tradeoff between accuracy and precision.

5.3 Discussion

To allow meaningful comparisons among binning procedures, I introduced a measure of binning accuracy that has a consistent meaning regardless of the choice of training bins. In particular, it will be useful to compare binning performance at different phylogenetic levels in a manner that incorporates the tradeoff between precision and accuracy. Quantifying this tradeoff will help decide whether an increase in accuracy at the order level compared to the genus level is worth the cost of reduced precision.

It might be tempting to think of a binning method as a function which may be applied to a training set and a test set, i.e. method(training set, test set). However, it will in fact be more instructive to consider a binning procedure to consist of the combination of the method and the training set; this combination forms a function which may be applied to the test set, i.e. (method + training set)(test set). The reason is that we wish to score the performance of different procedures when applied to the same test set, in order to choose the best one. One question we hope to answer in this way is which training set to use (e.g., how many training genomes are necessary? Should they be fragmented or considered whole? At what phylogenetic level should we make predictions?). Considering the training set to be a part of the procedure that is being evaluated, and producing a score that has a consistent meaning and can be compared across training sets, makes it possible to answer these questions. This is in contrast to the conventional accuracy measures, which are comparable only within the context of a given training set.

The bep95 score has numerous advantages:

Consistent units; resolution vs accuracy. The bep95 score consistently uses branch length units, and so is comparable across all combinations of binning methods and training labels. In particular, it straightforwardly incorporates into the score the tradeoff between label resolution

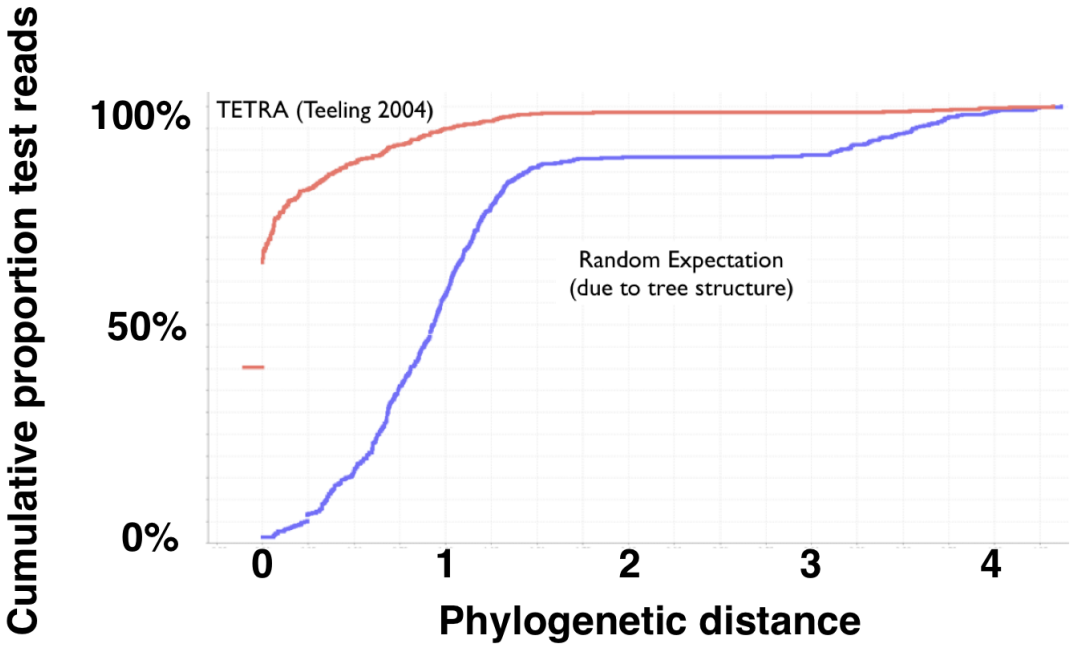

Figure 5.5: Cumulative histogram of bep scores for the TETRA method compared with the random expectation. For this example, reads of 1600nt selected randomly from the entire tree of life were used. The shape of the cumulative histogram curve for the "random" binning method is completely dictated by the distribution of branch lengths on the tree, which for this example was taken from Ciccarelli et al. (2006). In particular, the long plateau from bep ~1.5 to ~3.25 arises from the long branch separating Bacteria from Archaea. For better binning methods, the curve will approach the upper left corner. The bep95 score for a given method is the bep value at which the cumulative histogram reaches 95% (in this example, almost exactly 1.0 Ciccarelli units); the lower this score, the better the phylogenetic resolution of the method.

and accuracy. If the binning procedure attempts to classify only at the division level, then even if it gets every division right, there is still some error arising from the branch length from the division node down to the strain. However, if it attempts to classify at the strain level and gets it wrong, that's even worse, because the branch lengths from the common ancestor to both the prediction and the true species are counted. Thus the measure naturally incorporates the idea that binning correctly but with lower precision is better than aiming for higher precision but getting it wrong.

Disjoint test and training sets. The bep95 score can describe the precision of binning procedure even in the case where the test taxa are completely disjoint from the training taxa (a circumstance which would normally produce "0% accuracy").

Variable granularity. The bep95 score can meaningfully score a procedure which classifies at various levels of the tree: i.e., rather than choosing one of a set of disjoint classes, a procedure might be designed that sometimes chooses a species, but other times chooses only a division, depending on the procedure's internal confidence in its predictions.

Weighted by test set composition, vs. uniform class-normalized Sp/Sn. Sensitivity and specificity are defined only with respect to a single class in a binary classification setting. A common means of reporting sensitivity and specificity values for a multiclass classifier is to think of it as a set of binary classifiers, each distinguishing one class from all the others; this produces sensitivity and specificity values for each class, which can then be averaged to produce the "class-normalized sensitivity" and "class-normalized specificity".(Mavromatis et al. 2007; McHardy et al. 2007). This procedure weights the classes equally, when in fact we are interested in quality measures that reflect the composition of the test sample. I.e., the class-normalized specificity may appear to be excellent due to the contributions of rare classes, while the specificity with respect to a few dominant classes is less good. The bep95 score automatically takes the composition of the test sample into account.

This property may also be considered a detriment, since in some circumstances the bep95 score for a given binning procedure and training set may vary widely based on the choice of test set. In the case of a highly dominated population, for instance, the bep95 score will be driven by the bep profile of the dominant class (i.e., even for an ideal binning procedure, the bep95 score will be at least the distance from the dominant population to the nearest training bin). Thus two test sets with similar richness and evenness but with different dominant populations may produce different bep95 scores.

In summary, the computation of the bep95 score using test and training sets designed to mirror real circumstances will allow estimation of binning performance in terms that are directly relevant to biologists. Furthermore, the fact that the bep95 score has a consistent meaning in all circumstances allows the comparison and optimization of a wide variety of binning procedures, taking into account both the binning method itself and the choices of training set and classification level. It will thus finally become possible, after two decades of discussion of compositional

biases and proposed binning procedures, to choose the method with the best performance on the basis of a comprehensive and systematic evaluation.

Part III

Software

Chapter 6

Verdant: a platform for computational research that guarantees reproducibility, internal consistency, and currency of results

6.1 Abstract

Verdant is a system for describing, sharing, and executing computational workflows in a manner that guarantees reproducible results. It provides a means of ensuring that a set of computational results are up-to-date with respect to the inputs and thus that they are internally consistent. It also provides a means of sharing inputs, intermediate results, and final outputs in a manner that facilitates collaboration while avoiding redundant computation.

Widespread use of a system of this type would lead to many compelling benefits for the scientific community in all disciplines, including computer science, bioinformatics, neuroscience, physics, climate science, epidemiology, economics, sociology, and so forth.

My prototype implementation demonstrates that many of the ideas presented do in fact work in practice, but a substantially more robust and user-friendly implementation will be required to achieve widespread adoption.

6.2 Introduction

The need for tools to make computational research projects fully reproducible is becoming ever more evident (Mesirov 2010; Barnes 2010; Merali 2010), but the technical and social issues involved in accomplishing that goal are surprisingly complex (Stodden 2009a,b). A number of "scientific workflow" tools have been developed over the years (section 6.12), but their use is not yet standard practice.

Here I demonstrate several features that to my knowledge have been absent or underemphasized in workflow systems to date:

- The use of cryptographic hashing to uniquely identify derivation paths.
- "Deep" dependencies on every piece of software used in a derivation, including compilers and system libraries.
- Automatic distribution of those deep dependencies to whatever machines are performing the computations.
- Storage of inputs in a distributed version control system, allowing various approaches to collaboration, including branches and private branches.
- Automatic updating of results when inputs change ("truth maintenance"), even when these inputs come from collaborators or third parties.

This informal overview may eventually be followed by a document that goes into the conceptual underpinnings of collaborative workflow tools and solutions to various pitfalls in more depth. However, before discussing design principles for workflow managers in general, it will be helpful to have a basic idea of what my prototype software actually does.

Verdant, the "Versioned Data Analysis Tool", itself consists of a relatively small amount of glue code that serves to coordinate the activities of two other programs, Mercurial and Nix, which do the heavy lifting.

6.3 Version control of input files

All inputs to a computational workflow are stored in a set of Mercurial repositories. By "inputs" I mean any computer files that cannot be automatically derived from other files, such as files created by humans (source code and documentation) as well as raw data obtained from experiments.

Mercurial is a distributed version control system (DVCS) very similar to Git. Unlike their predecessors (such as CVS and Subversion), DVCSs do not require a central repository location; rather, the entire version history is distributed to all clients. It can however be convenient, though it is not necessary, to store repositories for a given project or lab together in a central location.

Mercurial assigns version numbers (really cryptographic hashes) to repository snapshots as a whole, not to individual files. Thus, each input that may vary independently (i.e., each "module") should be stored in a separate repository. This approach produces a large number of separate repositories, each usually containing a relatively small amount of information such as the code for a single program, or a single data file.

I employ a simple standardized naming scheme to keep these many repositories organized, based on reversed domain names (like Java package naming). For instance, I keep the Muscle multiple alignment program in a module called “com.drive5.muscle”.

6.4 Specification and computation of the workflow

I have repurposed the Nix package manager (http://www.nixos.org) to drive my computational workflows. Nix was designed to build Linux programs from their sources with rigorous dependency tracking and resolution; it is conceptually similar to Make, but operates on a larger scale and with more rigorous guarantees of deterministic results. For instance, when building the Apache web server from sources, Nix first builds prerequisites such as openssl, downloading sources as needed.

In the intended usage of Nix, the inputs to the Apache build are the source code files and the prerequisite libraries; the computation being performed is simply compilation (typically via Make); and the output is the binary “httpd” program. In my usage, by analogy, data files and other inputs to some computation are like source code; the computation is whatever we like; and the output plays the role of the compiled binary, in that it is derived from the inputs.

To perform a computation, then, I write an expression in the Nix language that specifies which program to run and on which inputs. I specify both the programs and the data inputs by their module names as described above. These module names map to Mercurial repositories, but not to specific versions; these will be specified later. I store the Nix expressions themselves as modules like any other input, so these too are versioned.

The input to one step of the workflow may be the output of another step, which is also described by a Nix expression. Thus, the Nix dependency resolution mechanism ensures that intermediate results are computed in the required order. (For readers with CS background: Nix is a pure functional language with lazy evaluation.)

Workflows need not be linear (as might be implied if we called them “pipelines”); each step may have multiple inputs and multiple outputs, and intermediate results may be used in multiple places downstream; the only constraint is that the specified dependency network may obviously have no cycles.

When I want to run the workflow, I provide a file that maps module names to version numbers (typically, the most recent version available in each of the module repositories), and request a specific output. Verdant then derives the output, computing prerequisite intermediate results as needed. Verdant automatically retrieves the specified version of each input module from its Mercurial repository, without referring to or interfering with any working copies. This ensures that the computation is performed on a committed version of each input module, and that the same version of a given module is used throughout the computation if it appears in multiple places.

Nix caches the results of each computation, employing cryptographic hashing to uniquely identify each derived artifact based on all of its dependencies. Thus, if one of several inputs to a

workflow is updated to a new version, only those nodes that depend on that input need to be recomputed. The Nix cache may contain multiple output files from the same workflow resulting from runs with different input versions. These will have different hash values, and the input versions that were used in each case can be traced at any time.

6.5 Standard programs and libraries

Inputs to a computation may include not only my own programs and data stored in Verdant modules as described above, but also Nix expressions from the Nixpkgs collection; these can build thousands of common Linux packages. Importantly, Nix enforces that every input must be specified explicitly; for instance, scripts cannot refer to programs expecting them to be on the path, or to libraries expected to be in /usr/lib. Thus, for instance, a Nix expression that runs a Perl script must include the Perl interpreter itself as an input. The version of Perl that is used will be the one specified in Nixpkgs; this will be downloaded and compiled from source automatically, and any version of Perl already available on the build machine will be ignored. This provides a hard guarantee that exactly the same versions of every upstream program and library will be used on every machine where a derivation is computed (e.g., on the desktop machines of multiple collaborators, and on a production cluster, etc.)– and this happens completely automatically, with no danger of library incompatibilities and no system administration effort.

A potentially troublesome consequence is that, if a Verdant user updates Nixpkgs after a new Perl version has become available, then any derivations that involve a Perl script anywhere upstream must be recomputed. This may seem like overkill, but is formally the correct behavior: there is no guarantee that existing scripts executed by a new Perl interpreter will give the same result that they previously did. I take the conservative approach that if any input changes, even including incremental updates to distant upstream libraries, then all bets are off as to my results, so I must recompute them. In practice, a solution for now is to update Nixpkgs infrequently if at all.

6.6 Collaboration and distributed workflows

Because all inputs are stored in Mercurial repositories, all of the collaboration features provided by distributed version control apply to the Verdant workflow as well, including branching and merging, private branches, straightforward backups, and so forth.

Also, because Nix derivations are uniquely identifiable by their cryptographic hash, the outputs of computations can be freely shared among machines or users without fear of unknown variability in the inputs. Nix provides a means of sharing outputs via a web server (intended as a mechanism for distributing compiled binaries instead of sources). Thus, for instance, if a user requests the (potentially expensive) evaluation of an expression that has already been computed, then Nix can simply download the output instead of recomputing it, with confidence that the result would have been identical anyway. This can work regardless of where the result was

previously computed: whether by the same user on a different machine, or by a collaborator, or by a researcher somewhere in the world who makes his Nix cache publicly accessible.

6.7 Cluster computing

Verdant is ideally suited to cluster and cloud computing, because a) it guarantees a consistent software environment and b) the pure, functional, and lazy nature of the evaluation of Nix expressions means that independent derivations may be computed in parallel. The Nix engine automatically makes use of multicore machines, or, with minor configuration, ad-hoc clusters of machines. I have successfully computed complex derivation networks in parallel on a large research cluster at UC Berkeley.

6.8 Continuous Integration

A common piece of infrastructure in a software engineering environment is a "continuous integration" server, which automatically recompiles code, often from multiple modules, and downloads new versions of libraries from remote sources. The resulting programs are then subjected to automated testing to ensure that the combination of all the latest versions of the software components behaves as expected. This is done either continuously (i.e., whenever a change is detected in a version control system) or on a recurring schedule, i.e. nightly.

Verdant inherently provides this functionality for free. It is necessary only to place test cases in a module that depends on the software to be tested; the output of the derivation is then a report of the test results. All that remains is to trigger the test derivation regularly (or upon a version control update); the test results will then naturally reflect the most recent versions of all inputs.

6.9 How this system guarantees reproducibility

In order to reproduce a given result, such as a plot made for a paper, we need a) the version numbers for all of the input modules upstream of the plot, and either b1) access to the Mercurial repositories from which we may extract those versions, or b2) archived tarballs of just those versions of the inputs.

Verdant makes it trivial to archive everything that is needed to reproduce the plot: it can simply compute the closure of the expression that generates the image. This includes the expression itself, any upstream Nix expressions used in generating intermediate results, and a list of all input modules and their versions. Given this list, one can copy all of the indicated artifacts (i.e., the tarballs of the specified versions extracted from the Mercurial repositories) to an archival location.

The deterministic nature of this whole procedure, verified by cryptographic hashing, guarantees that the entire computational network leading to the plot can be recomputed later from the archived inputs, and that all intermediate and final results will be identical. The exact versions of any programs used (or more likely the source code of these programs) are naturally part of the archive, so these can always be examined later to fully understand the provenance of the result.

Thus, this mechanism can be used to publish a complete set of artifacts together with each paper, allowing readers to regenerate the plots exactly as published, or to generate alternate versions by changing parameters or providing alternate input files.

6.10 TupleStreams and PlotBot

Many data manipulation tasks that might naturally arise in the course of a scientific workflow would be most easily performed in a database environment, because SQL is well suited to such tasks. However, database manipulations are not reproducible, and thus cannot be incorporated into a Verdant workflow (except perhaps in the form of manipulations to an in-memory database that is loaded anew at the beginning of a derivation).

Verdant therefore calls for a means to make simple data rearrangements that is functional in nature. I provide a very simple tool that fills this need, called TupleStreams. As the name suggests, it functions on streams of tuples, which can be thought of as rows of a table (in practice, often stored in tab-delimited text files). TupleStreams executes scripts, written in a simple custom language, that specify transformations to such streams, including selecting and rearranging columns; computing new columns as functions of existing columns; sorting; filtering; joining; zippering; and so forth.

Some of these functions (such as sorting) inherently require reading the entire input stream before operating, and thus are memory-limited. However, in many cases, the transformation of each row is independent of the others, or perhaps dependent on an aggregate computation based on the preceding rows. In these cases, a streaming approach is taken, allowing the processing of arbitrarily large data files.

TupleStreams does nothing that could not be accomplished (with more effort) in Perl or any other language; but it does dramatically simplify a number of common data-manipulation tasks.

Another task that is frequently performed in a stateful (i.e. non-functional) manner is the preparation of plots. In order to produce plots as outputs of Verdant workflows, we require a program that takes two inputs, a data file and a file describing the appearance of the desired plot (a "plot spec"), and produces a graphic file as output. I was surprised to find a dearth of scriptable plotting applications available, and so wrote PlotBot to fill this need.

PlotBot is simply a wrapper around the JFreeChart plotting library which allows it to participate in Verdant workflows as described. JFreeChart provides a wide variety of plot types, including line plots, scatterplots, bar charts, and so forth, as well as very many options for customizing

the plot appearance. As a result, the plot spec files are unfortunately quite complex; providing a simpler syntax for these (or better yet, a graphical editor) will make the program substantially more user-friendly in the future.

PlotBot need not incorporate any features for data manipulations that might normally be considered to be in the purview of a plotting program (e.g., histogramming, adding noise, sorting and aggregating data, computing error bars, etc.) because all such transformations can be made in TupleStreams. It is thus a common pattern in Verdant workflows to use TupleStreams to generate the exact values to be plotted, which are then passed to PlotBot for drawing.

6.11 Conclusion

Verdant is a system for describing, sharing, and executing computational workflows in a manner that guarantees reproducible results. My prototype implementation is difficult to use and suffers from various technical hiccups. Nonetheless, it is a proof of principle that computational results can be simultaneously reproducible, internally consistent, and up-to-date, even when multiple collaborators work in parallel. In fact, the system as described will work even when different inputs to a system are maintained by different parties, who are thus effectively collaborators even if they don't know each other.

A future implementation, not dependent on Nix, can be made far more user-friendly and efficient in various ways. Indeed, many similar projects already provide polished user interfaces (section 6.12)

The combination of version control with reproducible computation means that we can both examine the provenance of old results and compute up-to-date results at any time.

Workflow inputs can be shared in a relatively fine-grained way through Mercurial repositories. Many programs and data files are of interest to thousands of researchers, and thus could be stored in publicly accessible Mercurial repositories, providing those researchers with automatic access to the latest version (or whatever specific versions they choose) of each third-party input to their computations.

Computational results can be uniquely identified using cryptographic hashes (both of the derivation path and of the result itself) and shared (presently via "Nix channels"). Such outputs may also be of widespread interest when they are published in journal articles; for instance, researchers may wish to use published outputs as inputs to their own computations. Thus these artifacts could also be made publicly available. Furthermore, electronic availability of the workflow used to compute results for a given publication would facilitate complete transparency about exactly what was done, and would allow third parties to test whether the conclusions of the paper (or any downstream results) remain true when new versions of the inputs become available.

In sum, widespread use of a system of this type would produce a worldwide ecosystem of versioned interdependent digital artifacts. Performing computational research in this way would

guarantee reproducibility, would enable consistency within and between studies, would encourage openness, and would allow work in all disciplines where computation is used again to be performed according to the scientific method.

I used Verdant to perform nearly all of the computations reported in this dissertation, and to generate nearly all of the plots. Indeed, but for one minor technical hiccup, it was nearly possible for the dissertation PDF itself to be the output of a Verdant derivation, dependent on those plots and hence in turn on the computations.

6.12 Related projects

A number of projects that aim to manage workflows for reproducible research are far more polished than Verdant (Table 6.1), but I am not yet convinced that any of them provide the same combination of flexibility, rigor, and openness that Verdant does. In particular, Verdant demonstrates the utility of integrating version control systems, cryptographic hashing, and functional programming, as weorders of magnitudell as a deep dependency network including compilers and system libraries; and it supports a fully distributed collaboration model. To my knowledge no other project has this combination of features.

Name	URL	Citation
GenePattern	http://www.broadinstitute.org/cancer/software/genepattern/	(Reich et al. 2006; Mesirov 2010)
Galaxy	http://galaxy.psu.edu/	(Giardine et al. 2005; Blankenberg et al. 2007; Taylor et al. 2007; Kosakovsky Pond et al. 2009; Blankenberg et al. 2010; Bock et al. 2010)
Trident	http://research.microsoft.com/en-us/collaboration/tools/trident.aspx	(Simmhan et al. 2008)
Kepler	https://kepler-project.org/	(Ludäscher et al. 2009)
Taverna	http://www.taverna.org.uk	(Oinn et al. 2004; Hull et al. 2006; Oinn et al. 2006; Lanzén and Oinn 2008; Sroka et al. 2009)
myExperiment	http://www.myexperiment.org/	(Roure et al. 2009)
SciWalker	http://laser.cs.umass.edu/tools/sciwalker.shtml	(Osterweil et al. 2006)

Table 6.1: Existing workflow management tools with similar goals.

Bibliography

Takashi Abe, Shigehiko Kanaya, Makoto Kinouchi, Yuta Ichiba, Tokio Kozuki, and Toshimichi Ikemura. Informatics for unveiling hidden genome signatures. *Genome Res*, 13(4):693–702, 4 2003. ISSN 1088-9051. doi: 10.1101/gr.634603.

Silvia G. Acinas, Vanja Klepac-Ceraj, Dana E. Hunt, Chanathip Pharino, Ivica Ceraj, Daniel L. Distel, and Martin F. Polz. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, 430(6999):551–4, 7 2004a. ISSN 1476-4687. doi: 10.1038/nature02649.

Silvia G. Acinas, Luisa A. Marcelino, Vanja Klepac-Ceraj, and Martin F. Polz. Divergence and redundancy of 16s rrna sequences in genomes with multiple rrn operons. *J Bacteriol*, 186(9): 2629–35, 5 2004b. ISSN 0021-9193.

Eric E. Allen and Jillian F. Banfield. Community genomics in microbial ecology and evolution. *Nat Rev Microbiol*, 3(6):489–98, 6 2005. ISSN 1740-1526. doi: 10.1038/nrmicro1157.

R. I. Amann, W. Ludwig, and K. H. Schleifer. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev*, 59(1):143–69, 3 1995. ISSN 0146-0749.

Anders F. Andersson, Mathilda Lindberg, Hedvig Jakobsson, Fredrik Bäckhed, Pål Nyrén, and Lars Engstrand. Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE*, 3(7):e2836, 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0002836.

Kevin R. Arrigo. Marine microorganisms and global nutrient cycles. *Nature*, 437(7057):349–55, 9 2005. ISSN 1476-4687. doi: 10.1038/nature04159.

G. C. Baker and D. A. Cowan. 16 s rdna primers and the unbiased assessment of thermophile diversity. *Biochem Soc Trans*, 32(Pt 2):218–21, 4 2004. ISSN 0300-5127. doi: 10.1042/.

G. C. Baker, J. J. Smith, and D. A. Cowan. Review and re-analysis of domain-specific 16s primers. *J Microbiol Methods*, 55(3):541–55, 12 2003. ISSN 0167-7012.

Nick Barnes. Publish your computer code: it is good enough. *Nature*, 467(7317):753, 10 2010. ISSN 1476-4687. doi: 10.1038/467753a.

Alex Bateman, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L. L. Sonnhammer, David J. Studholme, Corin Yeats, and Sean R. Eddy. The pfam protein families database. *Nucleic Acids Res*, 32(Database issue):D138–41, 1 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh121.

Thomas Bell, Jonathan A. Newman, Bernard W. Silverman, Sarah L. Turner, and Andrew K. Lilley. The contribution of species richness and composition to bacterial services. *Nature*, 436(7054):1157–60, 8 2005. ISSN 1476-4687. doi: 10.1038/nature03891.

Pat H. Bellamy, Peter J. Loveland, R. Ian Bradley, R. Murray Lark, and Guy J. D. Kirk. Carbon losses from all soils across england and wales 1978-2003. *Nature*, 437(7056):245–8, 9 2005. ISSN 1476-4687. doi: 10.1038/nature04038.

T. S. Bibby, I. Mary, J. Nield, F. Partensky, and J. Barber. Low-light-adapted prochlorococcus species possess specific antennae for each photosystem. *Nature*, 424(6952):1051–4, 8 2003. ISSN 1476-4687. doi: 10.1038/nature01933.

Daniel Blankenberg, James Taylor, Ian Schenck, Jianbin He, Yi Zhang, Matthew Ghent, Narayanan Veeraraghavan, Istvan Albert, Webb Miller, Kateryna D. Makova, Ross C. Hardison, and Anton Nekrutenko. A framework for collaborative analysis of encode data: making large-scale analyses biologist-friendly. *Genome Res*, 17(6):960–4, 6 2007. ISSN 1088-9051. doi: 10.1101/gr.5578007.

Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*, Chapter 19:Unit 19.10.1–21, 1 2010. ISSN 1934-3647. doi: 10.1002/0471142727.mb1910s89.

Christoph Bock, Greg Von Kuster, Konstantin Halachev, James Taylor, Anton Nekrutenko, and Thomas Lengauer. Web-based analysis of (epi-) genome data using epigraph and galaxy. *Methods Mol Biol*, 628:275–96, 2010. ISSN 1940-6029.

Arthur Brady and Steven L. Salzberg. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nat Methods*, 8 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1358.

O. G. Brakstad and A. G. G. Lødeng. Microbial diversity during biodegradation of crude oil in seawater from the north sea. *Microb Ecol*, 49(1):94–103, 1 2005. ISSN 0095-3628. doi: 10.1007/s00248-003-0225-6.

Mya Breitbart, Ian Hewson, Ben Felts, Joseph M. Mahaffy, James Nulton, Peter Salamon, and Forest Rohwer. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol*, 185(20):6220–3, 10 2003. ISSN 0021-9193.

T. B. Britschgi and S. J. Giovannoni. Phylogenetic analysis of a natural marine bacterioplankton population by rrna gene cloning and sequencing. *Appl Environ Microbiol*, 57(6):1707–13, 6 1991. ISSN 0099-2240.

J. M. Brulc, D. A. Antonopoulos, M. E. Berg Miller, M. K. Wilson, A. C. Yannarell, E. A. Dinsdale, R. E. Edwards, E. D. Frank, J. B. Emerson, and P. Wacklin. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of the National Academy of Sciences*, 106(6):1948, 2009.

A. Campbell, J. Mrázek, and S. Karlin. Genome signature comparisons among prokaryote, plasmid, and mitochondrial dna. *Proc Natl Acad Sci U S A*, 96(16):9184–9, 8 1999. ISSN 0027-8424.

J. Gregory Caporaso, Christian L. Lauber, William A. Walters, Donna Berg-Lyons, Catherine A. Lozupone, Peter J. Turnbaugh, Noah Fierer, and Rob Knight. Microbes and health sackler colloquium: Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A*, 6 2010. ISSN 1091-6490. doi: 10.1073/pnas.1000080107.

Chon-Kit Kenneth Chan, Arthur L. Hsu, Saman K. Halgamuge, and Sen-Lin L. Tang. Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics*, 9(1):215, 4 2008a. ISSN 1471-2105. doi: 10.1186/1471-2105-9-215.

Chih-Chung . C. Chang and Chih-Jen . J. Lin. *LIBSVM: a library for support vector machines*, 2001.

Charles Chapus, Christine Dufraigne, Scott Edwards, Alain Giron, Bernard Fertil, and Patrick Deschavanne. Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol Biol*, 5:63, 11 2005. ISSN 1471-2148. doi: 10.1186/1471-2148-5-63.

Swaine L. Chen, William Lee, Alison K. Hottes, Lucy Shapiro, and Harley H. McAdams. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci U S A*, 101(10):3480–5, 3 2004. ISSN 0027-8424. doi: 10.1073/pnas.0307827100.

J. C. Cho, D. H. Lee, Y. C. Cho, J. C. Cho, and S. J. Kim. Direct extraction of dna from soil for amplification of 16s rrna gene sequences by polymerase chain reaction. *J. Microbiology*, 34 (3):229–235, 1996.

Rakia Chouari, Denis Le Paslier, Patrick Daegelen, Philippe Ginestet, Jean Weissenbach, and Abdelghani Sghir. Novel predominant archaeal and bacterial groups revealed by molecular analysis of an anaerobic sludge digester. *Environ Microbiol*, 7(8):1104–15, 8 2005. ISSN 1462-2912. doi: 10.1111/j.1462-2920.2005.00795.x.

Brent C. Christner, Rongman Cai, Cindy E. Morris, Kevin S. McCarter, Christine M. Foreman, Mark L. Skidmore, Scott N. Montross, and David C. Sands. Geographic, seasonal, and precipitation chemistry influence on the abundance and activity of biological ice nucleators in rain and snow. *Proc Natl Acad Sci U S A*, 105(48):18854–9, 12 2008. ISSN 1091-6490. doi: 10.1073/pnas.0809816105.

Haiyan Chu, Noah Fierer, Christian L. Lauber, J. G. Caporaso, Rob Knight, and Paul Grogan. Soil bacterial diversity in the arctic is not fundamentally different from that found in other biomes. *Environ Microbiol*, 6 2010. ISSN 1462-2920. doi: 10.1111/j.1462-2920.2010.02277.x.

K. H. Chu, C. P. Li, and J. Qi. Ribosomal rna as molecular barcodes: a simple correlation analysis without sequence alignment. *Bioinformatics*, 22(14):1690–701, 4 2006. ISSN 1460-2059. doi: 10.1093/bioinformatics/btl146.

Francesca D. Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–7, 3 2006. ISSN 1095-9203. doi: 10.1126/science.1123061.

V. Cilia, B. Lafay, and R. Christen. Sequence heterogeneities among 16s ribosomal rna sequences, and their effect on phylogenetic analyses at the species level. *Mol Biol Evol*, 13(3): 451–61, 3 1996. ISSN 0737-4038.

Frederick M. Cohan. What are bacterial species? *Annu Rev Microbiol*, 56:457–87, 2002. ISSN 0066-4227. doi: 10.1146/annurev.micro.56.012302.160634.

J. R. Cole, B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje. The ribosomal database project (rdp-ii): sequences and tools for high-throughput rrna analysis. *Nucleic Acids Res*, 33(Database issue):D294–6, 1 2005. ISSN 1362-4962. doi: 10.1093/nar/gki038.

The Human Microbiome Jumpstart Reference Strains Consortium, Karen E. Nelson, George M. Weinstock, Sarah K. Highlander, Kim C. Worley, Heather Huot Creasy, Jennifer Russo Wortman, Douglas B. Rusch, Makedonka Mitreva, Erica Sodergren, Asif T. Chinwalla, Michael Feldgarden, Dirk Gevers, Brian J. Haas, Ramana Madupu, Doyle V. Ward, Bruce W. Birren, Richard A. Gibbs, Barbara Methe, Joseph F. Petrosino, Robert L. Strausberg, Granger G. Sutton, Owen R. White, Richard K. Wilson, Scott Durkin, Michelle Gwinn Giglio, Sharvari Gujja, Clint Howarth, Chinnappa D. Kodira, Nikos Kyrpides, Teena Mehta, Donna M. Muzny, Matthew Pearson, Kymberlie Pepin, Amrita Pati, Xiang Qin, Chandri Yandava, Qiandong Zeng, Lan Zhang, Aaron M. Berlin, Lei Chen, Theresa A. Hepburn, Justin Johnson, Jamison McCorrison, Jason Miller, Pat Minx, Chad Nusbaum, Carsten Russ, Sean M. Sykes, Chad M. Tomlinson, Sarah Young, Wesley C. Warren, Jonathan Badger, Jonathan Crabtree, Victor M. Markowitz, Joshua Orvis, Andrew Cree, Steve Ferriera, Lucinda L. Fulton, Robert S. Fulton, Marcus Gillis, Lisa D. Hemphill, Vandita Joshi, Christie Kovar, Manolito Torralba, Kris A. Wetterstrand, Amr Abouellleil, Aye M. Wollam, Christian J. Buhay, Yan Ding, Shannon Dugan, Michael G. Fitzgerald, Mike Holder, Jessica Hostetler, Sandra W. Clifton, Emma Allen-Vercoe, Ashlee M. Earl, Candace N. Farmer, Konstantinos Liolios, Michael G. Surette, Qiang Xu, Craig Pohl, Katarzyna Wilczek-Boney, and Dianhui Zhu. A catalog of reference genomes from the human microbiome. *Science*, 328(5981):994–999, 5 2010. ISSN 1095-9203. doi: 10.1126/science.1183605.

Laurel D. Crosby and Craig S. Criddle. Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity. *Biotechniques*, 34(4):790–4, 796, 798 passim, 4 2003. ISSN 0736-6205.

Karelyn Cruz-Martínez, K. Blake Suttle, Eoin L. Brodie, Mary E. Power, Gary L. Andersen, and Jillian F. Banfield. Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. *ISME J*, 3 2009. ISSN 1751-7370. doi: 10.1038/ismej.2009.16.

Thomas P. Curtis, William T. Sloan, and Jack W. Scannell. Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci U S A*, 99(16):10494–9, 8 2002. ISSN 0027-8424. doi: 10.1073/pnas.142680199.

Rolf Daniel. The metagenomics of soil. *Nat Rev Microbiol*, 3(6):470–8, 6 2005. ISSN 1740-1526. doi: 10.1038/nrmicro1160.

Edward F. DeLong. Microbial community genomics in the ocean. *Nat Rev Microbiol*, 3(6): 459–69, 6 2005. ISSN 1740-1526. doi: 10.1038/nrmicro1158.

Vincent J. Denef, Linda H. Kalnejais, Ryan S. Mueller, Paul Wilmes, Brett J. Baker, Brian C. Thomas, Nathan C. Verberkmoes, Robert L. Hettich, and Jillian F. Banfield. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proc Natl Acad Sci U S A*, 2 2010a. ISSN 1091-6490. doi: 10.1073/pnas.0907041107.

Vincent J. Denef, Ryan S. Mueller, and Jillian F. Banfield. Amd biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J*, 2 2010b. ISSN 1751-7370. doi: 10.1038/ismej.2009.158.

T. Z. Desantis, P. Hugenholtz, K. Keller, E. L. Brodie, N. Larsen, Y. M. Piceno, R. Phan, and G. L. Andersen. Nast: a multiple sequence alignment server for comparative analysis of 16s rrna genes. *Nucleic Acids Res*, 34(Web Server issue):W394–9, 7 2006a. ISSN 1362-4962. doi: 10.1093/nar/gkl244.

T. Z. Desantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Appl Environ Microbiol*, 72(7):5069–72, 7 2006b. ISSN 0099-2240. doi: 10.1128/AEM.03006-05.

Les Dethlefsen, Sue Huse, Mitchell L. Sogin, and David A. Relman. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rrna sequencing. *PLoS Biol*, 6(11):e280, 11 2008. ISSN 1545-7885. doi: 10.1371/journal.pbio.0060280.

W. Dowhan. Molecular basis for membrane phospholipid diversity: why are there so many lipids? *Annu Rev Biochem*, 66:199–232, 1997. ISSN 0066-4154. doi: 10.1146/annurev.biochem.66.1.199.

Margaret J. Duncan. Genomics of oral bacteria. *Crit Rev Oral Biol Med*, 14(3):175–87, 2003. ISSN 1544-1113.

Paul B. Eckburg, Elisabeth M. Bik, Charles N. Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael Sargent, Steven R. Gill, Karen E. Nelson, and David A. Relman. Diversity of the human intestinal microbial flora. *Science*, 308(5728):1635–8, 6 2005. ISSN 1095-9203. doi: 10.1126/science.1110591.

Robert C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–1, 10 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq461.

Mostafa S. Elshahed, Noha H. Youssef, Anne M. Spain, Cody Sheik, Fares Z. Najar, Leonid O. Sukharnikov, Bruce A. Roe, James P. Davis, Patrick D. Schloss, Vanessa L. Bailey, and Lee R. Krumholz. Novelty and uniqueness patterns of rare members of the soil biosphere. *Appl Environ Microbiol*, 74(17):5422–8, 7 2008. ISSN 1098-5336. doi: 10.1128/AEM.00410-08.

Anna Engelbrektson, Victor Kunin, Kelly C. Wrighton, Natasha Zvenigorodsky, Feng Chen, Howard Ochman, and Philip Hugenholtz. Experimental factors affecting pcr-based estimates of microbial species richness and evenness. *ISME J*, 1 2010. ISSN 1751-7370. doi: 10.1038/ismej.2009.153.

Steven D. Essinger and Gail L. Rosen. Benchmarking blast accuracy of genus/phyla classification of metagenomic reads. *Pac Symp Biocomput*, pages 10–20, 2010. ISSN 1793-5091.

Noah Fierer, Micah Hamady, Christian L. Lauber, and Rob Knight. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci U S A*, 11 2008. ISSN 1091-6490. doi: 10.1073/pnas.0807920105.

M. M. Fisher and E. W. Triplett. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol*, 65(10):4630–6, 10 1999. ISSN 0099-2240.

Konrad U. Foerstner, Christian von Mering, Sean D. Hooper, and Peer Bork. Environments shape the nucleotide composition of genomes. *EMBO Rep*, 6(12):1208–13, 12 2005. ISSN 1469-221X. doi: 10.1038/sj.embor.7400538.

Jonathan A. Foley, Ruth Defries, Gregory P. Asner, Carol Barford, Gordon Bonan, Stephen R. Carpenter, F. Stuart Chapin, Michael T. Coe, Gretchen C. Daily, Holly K. Gibbs, Joseph H. Helkowski, Tracey Holloway, Erica A. Howard, Christopher J. Kucharik, Chad Monfreda, Jonathan A. Patz, I. Colin Prentice, Navin Ramankutty, and Peter K. Snyder. Global consequences of land use. *Science*, 309(5734):570–4, 7 2005. ISSN 1095-9203. doi: 10.1126/science.1111772.

Jeremy A. Frank, Claudia I. Reich, Shobha Sharma, Jon S. Weisbaum, Brenda A. Wilson, and Gary J. Olsen. Critical evaluation of two primers commonly used for amplification of bacterial 16s rrna genes. *Appl Environ Microbiol*, 74(8):2461–70, 4 2008. ISSN 1098-5336. doi: 10.1128/AEM.02272-07.

Pierre E. Galand, Emilio O. Casamayor, David L. Kirchman, and Connie Lovejoy. Ecology of the rare microbial biosphere of the arctic ocean. *Proc Natl Acad Sci U S A*, 106(52): 22427–32, 12 2009. ISSN 1091-6490. doi: 10.1073/pnas.0908284106.

Jason Gans, Murray Wolinsky, and John Dunbar. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*, 309(5739):1387–90, 8 2005. ISSN 1095-9203. doi: 10.1126/science.1112665.

Lei Gao, Ji Qi, JianDong Sun, and BaiLin Hao. Prokaryote phylogeny meets taxonomy: an exhaustive comparison of composition vector trees with systematic bacteriology. *Sci China C Life Sci*, 50(5):587–99, 10 2007. ISSN 1006-9305. doi: 10.1007/s11427-007-0084-3.

F. Ge, L. S. Wang, and J. Kim. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol*, 3(10):e316, 8 2005. ISSN 1545-7885. doi: 10.1371/journal.pbio.0030316.

Dirk Gevers, Frederick M. Cohan, Jeffrey G. Lawrence, Brian G. Spratt, Tom Coenye, Edward J. Feil, Erko Stackebrandt, Yves Van de Peer, Peter Vandamme, Fabiano L. Thompson, and Jean Swings. Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol*, 3(9): 733–9, 9 2005. ISSN 1740-1526. doi: 10.1038/nrmicro1236.

Belinda Giardine, Cathy Riemer, Ross C. Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, Webb Miller, W. James Kent, and Anton Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, 15(10):1451–5, 10 2005. ISSN 1088-9051. doi: 10.1101/gr.4086505.

Stephen J. Giovannoni and Ulrich Stingl. Molecular diversity and ecology of microbial plankton. *Nature*, 437(7057):343–8, 9 2005. ISSN 1476-4687. doi: 10.1038/nature04158.

Johan Goris, Konstantinos T. Konstantinidis, Joel A. Klappenbach, Tom Coenye, Peter Vandamme, and James M. Tiedje. Dna-dna hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol*, 57(Pt 1):81–91, 1 2007. ISSN 1466-5026. doi: 10.1099/ijs.0.64483-0.

Elizabeth A. Grice, Heidi H. Kong, Sean Conlan, Clayton B. Deming, Joie Davis, Alice C. Young, NISC Comparative Sequencing Program, Gerard G. Bouffard, Robert W. Blakesley, Patrick R. Murray, Eric D. Green, Maria L. Turner, and Julia A. Segre. Topographical and temporal diversity of the human skin microbiome. *Science*, 324(5931):1190–2, 5 2009. ISSN 1095-9203. doi: 10.1126/science.1171700.

Amit Gur and Dani Zamir. Unused natural variation can lift yield barriers in plant breeding. *PLoS Biol*, 2(10):e245, 10 2004. ISSN 1545-7885. doi: 10.1371/journal.pbio.0020245.

Timothy J. Hamp, W. Joe Jones, and Anthony A. Fodor. Effects of experimental choices and analysis noise on surveys of the "rare biosphere". *Appl Environ Microbiol*, 75(10):3263–70, 5 2009. ISSN 1098-5336. doi: 10.1128/AEM.01931-08.

Jo Handelsman. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*, 68(4):669–85, 12 2004. doi: 10.1128/MMBR.68.4.669-685.2004.

Kristian Hanekamp, Uta Bohnebeck, Bánk Beszteri, and Klaus Valentin. Phylogena–a user-friendly system for automated phylogenetic annotation of unknown sequences. *Bioinformatics*, 23(7):793–801, 4 2007. ISSN 1460-2059. doi: 10.1093/bioinformatics/btm016.

Sunhee Hong, John Bunge, Chesley Leslin, Sunok Jeon, and Slava S. Epstein. Polymerase chain reaction primers miss half of rrna microbial diversity. *ISME J*, 3(12):1365–73, 8 2009. ISSN 1751-7370. doi: 10.1038/ismej.2009.89.

Matthew Horton, Natacha Bodenhausen, and Joy Bergelson. Marta: a suite of java-based tools for assigning taxonomic status to dna sequences. *Bioinformatics*, 26(4):568–9, 2 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp682.

P. Hugenholtz, B. M. Goebel, and N. R. Pace. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol*, 180(18):4765–74, 9 1998. ISSN 0021-9193.

Philip Hugenholtz. Exploring prokaryotic diversity in the genomic era. *Genome Biol*, 3(2): REVIEWS0003, 2002. ISSN 1465-6914.

Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R. Pocock, Peter Li, and Tom Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, 34(Web Server issue):W729–32, 7 2006. ISSN 1362-4962. doi: 10.1093/nar/gkl320.

Susan M. Huse, Julie A. Huber, Hilary G. Morrison, Mitchell L. Sogin, and David Mark Welch. Accuracy and quality of massively parallel dna pyrosequencing. *Genome Biol*, 8(7):R143, 2007. ISSN 1465-6914. doi: 10.1186/gb-2007-8-7-r143.

Susan M. Huse, Les Dethlefsen, Julie A. Huber, David Mark Welch, David A. Relman, and Mitchell L. Sogin. Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing. *PLoS Genet*, 4(11):e1000255, 11 2008. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000255.

T. A. Isenbarger, M. Finney, C. Rios-Velazquez, J. Handelsman, and G. Ruvkun. Miniprimer pcr, a new lens for viewing the microbial world. *Applied and environmental microbiology*, 74(3):840, 2008. ISSN 0099-2240.

H. J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Res*, 18(8):2163–70, 4 1990. ISSN 0305-1048.

L. Jost. Entropy and diversity. *Oikos*, 113(2):363, 2006.

Lou Jost. Partitioning diversity into independent alpha and beta components. *Ecology*, 88(10): 2427–39, 10 2007. ISSN 0012-9658.

S. Karlin and C. Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, 11(7):283–90, 7 1995. ISSN 0168-9525.

S. Karlin, A. M. Campbell, and J. Mrázek. Comparative dna analysis across diverse genomes. *Annu Rev Genet*, 32:185–225, 1998a. ISSN 0066-4197. doi: 10.1146/annurev.genet.32.1.185.

Konstantinos T. Konstantinidis and James M. Tiedje. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A*, 102(7):2567–72, 2 2005. ISSN 0027-8424. doi: 10.1073/pnas.0409727102.

Konstantinos T. Konstantinidis and James M. Tiedje. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol*, 10(5):504–9, 10 2007. ISSN 1369-5274. doi: 10.1016/j.mib.2007.08.006.

Sergei Kosakovsky Pond, Samir Wadhawan, Francesca Chiaromonte, Guruprasad Ananda, Wen-Yu Y. Chung, James Taylor, Anton Nekrutenko, and The Galaxy Team. Windshield splatter analysis with the galaxy metagenomic pipeline. *Genome Res*, 19(11):2144–53, 10 2009. ISSN 1549-5469. doi: 10.1101/gr.094508.109.

Amir Kovacs, Keren Yacoby, and Uri Gophna. A systematic assessment of automated ribosomal intergenic spacer analysis (arisa) as a tool for estimating bacterial richness. *Res Microbiol*, 161(3):192–7, 2 2010. ISSN 1769-7123. doi: 10.1016/j.resmic.2010.01.006.

Lutz Krause, Naryttza N. Diaz, Alexander Goesmann, Scott Kelley, Tim W. Nattkemper, Forest Rohwer, Robert A. Edwards, and Jens Stoye. Phylogenetic classification of short environmental dna fragments. *Nucleic Acids Res*, 2 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn038.

Victor Kunin, Alex Copeland, Alla Lapidus, Konstantinos Mavromatis, and Philip Hugenholtz. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev*, 72(4):557–78, Table of Contents, 12 2008. ISSN 1098-5557. doi: 10.1128/MMBR.00009-08.

Karin Lagesen, Peter Hallin, Einar Andreas Rødland, Hans-Henrik H. Staerfeldt, Torbjørn Rognes, and David W. Ussery. Rnammer: consistent and rapid annotation of ribosomal rna genes. *Nucleic Acids Res*, 35(9):3100–8, 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm160.

David A. Lane. *16S/23S rRNA sequencing*, pages 115–175. John Wiley & Sons, Chichester, 1991.

Anders Lanzén and Tom Oinn. The taverna interaction service: enabling manual interaction in workflows. *Bioinformatics*, 24(8):1118–20, 4 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn082.

J. G. Lawrence and H. Ochman. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol*, 44(4):383–97, 4 1997. ISSN 0022-2844.

Vladimir Lazarevic, Katrine Whiteson, Susan Huse, David Hernandez, Laurent Farinelli, Magne Osterås, Jacques Schrenzel, and Patrice François. Metagenomic study of the oral microbiota by illumina high-throughput sequencing. *J Microbiol Methods*, 9 2009. ISSN 1872-8359. doi: 10.1016/j.mimet.2009.09.012.

Min Li, Baohong Wang, Menghui Zhang, Mattias Rantalainen, Shengyue Wang, Haokui Zhou, Yan Zhang, Jian Shen, Xiaoyan Pang, Meiling Zhang, Hua Wei, Yu Chen, Haifeng Lu, Jian Zuo, Mingming Su, Yunping Qiu, Wei Jia, Chaoni Xiao, Leon M. Smith, Shengli Yang, Elaine Holmes, Huiru Tang, Guoping Zhao, Jeremy K. Nicholson, Lanjuan Li, and Liping Zhao. Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci U S A*, 2 2008a. ISSN 1091-6490. doi: 10.1073/pnas.0712038105.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–9, 7 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158.

Weizhong Li, John C. Wooley, and Adam Godzik. Probing metagenomics by rapid cluster analysis of very large datasets. *PLoS ONE*, 3(10):e3375, 2008b. ISSN 1932-6203. doi: 10.1371/journal.pone.0003375.

Peter A. Lind and Dan I. Andersson. Whole-genome mutational biases in bacteria. *Proc Natl Acad Sci U S A*, 11 2008. ISSN 1091-6490. doi: 10.1073/pnas.0804445105.

Zongzhi Liu, Catherine Lozupone, Micah Hamady, Frederic D. Bushman, and Rob Knight. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res*, 35(18):e120, 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm541.

Zongzhi Liu, Todd Z. Desantis, Gary L. Andersen, and Rob Knight. Accurate taxonomy assignments from 16s rrna sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res*, 8 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn491.

Derek R. Lovley. Cleaning up with genomics: applying molecular biology to bioremediation. *Nat Rev Microbiol*, 1(1):35–44, 10 2003. ISSN 1740-1526.

Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*, 71(12):8228–35, 12 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.12.8228-8235.2005.

Wolfgang Ludwig, Oliver Strunk, Ralf Westram, Lothar Richter, Harald Meier, Yadhukumar, Arno Buchner, Tina Lai, Susanne Steppi, Gangolf Jobb, Wolfram Förster, Igor Brettske, Stefan Gerber, Anton W. Ginhart, Oliver Gross, Silke Grumann, Stefan Hermann, Ralf Jost, Andreas König, Thomas Liss, Ralph Lüssmann, Michael May, Björn Nonhoff, Boris Reichel, Robert Strehlow, Alexandros Stamatakis, Norbert Stuckmann, Alexander Vilbig, Michael Lenke, Thomas Ludwig, Arndt Bode, and Karl-Heinz H. Schleifer. Arb: a software environment for sequence data. *Nucleic Acids Res*, 32(4):1363–71, 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh293.

Bertram Ludäscher, Ilkay Altintas, Shawn Bowers, Julian Cummings, Terence Critchlow, Ewa Deelman, David De Roure, Juliana Freire, Carole Goble, Matthew Jones, Scott Klasky, Timothy McPhillips, Norbert Podhorszki, Claudio Silva, Ian Taylor, and Mladen Vouk. Scientific process automation and workflow management, 2009.

Anne E. Magurran. *Measuring Biological Diversity*. Blackwell, Oxford, 2004.

K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*, 4(6):495–500, 4 2007. ISSN 1548-7091. doi: 10.1038/nmeth1043.

Alice C. McHardy and Isidore Rigoutsos. What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr Opin Microbiol*, 10 2007. ISSN 1369-5274. doi: 10.1016/j.mib.2007.08.004.

Alice Carolyn McHardy, Héctor García Martín, Aristotelis Tsirigos, Philip Hugenholtz, and Isidore Rigoutsos. Accurate phylogenetic classification of variable-length dna fragments. *Nat Methods*, 4(1):63–72, 1 2007. ISSN 1548-7091. doi: 10.1038/nmeth976.

Zeeya Merali. Nature, 10 2010. ISSN 1476-4687.

Jill P. Mesirov. Computer science. accessible reproducible research. *Science*, 327(5964):415–6, 1 2010. ISSN 1095-9203. doi: 10.1126/science.1179653.

Deetta K. Mills, James A. Entry, Joshua D. Voss, Patrick M. Gillevet, and Kalai Mathee. An assessment of the hypervariable domains of the 16s rrna genes for their value in determining microbial community diversity: the paradox of traditional ecological indices. *FEMS Microbiol Ecol*, 57(3):496–503, 9 2006. ISSN 0168-6496. doi: 10.1111/j.1574-6941.2006.00135.x.

Russell K. Monson, David L. Lipson, Sean P. Burns, Andrew A. Turnipseed, Anthony C. Delany, Mark W. Williams, and Steven K. Schmidt. Winter forest soil respiration controlled by climate and microbial community composition. *Nature*, 439(7077):711–4, 2 2006. ISSN 1476-4687. doi: 10.1038/nature04555.

Haque M. Monzoorul, Shankarghosh Tarini, Komanduri Dinakar, and Mande Sharmila S. Sortitems : Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 5 2009. ISSN 1460-2059. doi: 10.1093/bioinformatics/btp317.

C. L. Moyer, F. C. Dobbs, and D. M. Karl. Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16s rrna genes from a microbial mat at an active, hydrothermal vent system, loihi seamount, hawaii. *Appl Environ Microbiol*, 60(3):871–9, 3 1994. ISSN 0099-2240.

Jan Mrázek. Phylogenetic signals in dna composition: Limitations and prospects. *Mol Biol Evol*, 2 2009. ISSN 1537-1719. doi: 10.1093/molbev/msp032.

G. Muyzer and K. Smalla. Application of denaturing gradient gel electrophoresis (dgge) and temperature gradient gel electrophoresis (tgge) in microbial ecology. *Antonie Van Leeuwenhoek*, 73(1):127–41, 1 1998. ISSN 0003-6072.

H. Nakashima, M. Ota, K. Nishikawa, and T. Ooi. Genes from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res*, 5(5):251–9, 10 1998. ISSN 1340-2838.

Ivan Nasidze, Jing Li, Dominique Quinque, Kun Tang, and Mark Stoneking. Global diversity in the human salivary microbiome. *Genome Res*, 2 2009. ISSN 1088-9051. doi: 10.1101/gr.084616.108.

W. S. Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, 2006. ISSN 1087-0156.

A. Nocker, M. Burr, and A. K. Camper. Genotypic microbial community profiling: A critical technical review. *Microb Ecol*, 54(2):276–89, 3 2007. ISSN 0095-3628. doi: 10.1007/s00248-006-9199-5.

J. M. Oades. The role of biology in the formation, stabilization and degradation of soil structure. *Geoderma*, 56(1-4):377–400, 1993. ISSN 0016-7061.

Thomas Oinn, Mark Greenwood, Matthew Addis, Nedim Alpdemir, Justin Ferris, Kevin Glover, Carole Goble, Antoon Goderis, Duncan Hull, Darren Marvin, Peter Li, Phillip Lord, Matthew Pocock, Martin Senger, Robert Stevens, Anil Wipat, and Christopher Wroe. Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100, 8 2006. ISSN 1532-0626. doi: 10.1002/cpe.v18:10.

Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat, and Peter Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–54, 11 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth361.

G. J. Olsen, D. J. Lane, S. J. Giovannoni, N. R. Pace, and D. A. Stahl. Microbial ecology and evolution: a ribosomal rna approach. *Annu Rev Microbiol*, 40:337–65, 1986. ISSN 0066-4227. doi: 10.1146/annurev.mi.40.100186.002005.

L. Osterweil, A. Wise, L. Clarke, A. Ellison, J. Hadley, E. Boose, and D. Foster. Process technology to facilitate the conduct of science. *Unifying the Software Process Spectrum*, pages 403–415, 2006.

A. D. Peacock, Y. J. Chang, J. D. Istok, L. Krumholz, R. Geyer, B. Kinsall, D. Watson, K. L. Sublette, and D. C. White. Utilization of microbial biofilms as monitors of bioremediation. *Microb Ecol*, 47(3):284–92, 4 2004. ISSN 0095-3628. doi: 10.1007/s00248-003-1024-9.

Scott C. Perry and Robert G. Beiko. Distinguishing microbial genome fragments based on their composition: evolutionary and comparative genomic perspectives. *Genome Biol Evol*, 2010: 117–31, 2010. ISSN 1759-6653. doi: 10.1093/gbe/evq004.

F. Pfeiffer, S. C. Schuster, A. Broicher, M. Falb, P. Palm, K. Rodewald, A. Ruepp, J. Soppa, J. Tittor, and D. Oesterhelt. Evolution in the laboratory: the genome of halobacterium salinarum strain r1 compared to that of strain nrc-1. *Genomics*, 91(4):335–46, 4 2008. ISSN 1089-8646. doi: 10.1016/j.ygeno.2008.01.001.

Radu Popa, Rodica Popa, Matthew J. Mashall, Hien Nguyen, Bradley M. Tebo, and Suzanna Brauer. Limitations and benefits of arisa intra-genomic diversity fingerprinting. *J Microbiol Methods*, 78(2):111–8, 8 2009. ISSN 1872-8359. doi: 10.1016/j.mimet.2009.06.005.

Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol*, 26(7):1641–50, 7 2009. ISSN 1537-1719. doi: 10.1093/molbev/msp077.

David T. Pride, Richard J. Meinersmann, Trudy M. Wassenaar, and Martin J. Blaser. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res*, 13(2): 145–58, 2 2003. ISSN 1088-9051. doi: 10.1101/gr.335003.

David T. Pride, Trudy M. Wassenaar, Chandrabali Ghose, and Martin J. Blaser. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics*, 7:8, 2006. ISSN 1471-2164. doi: 10.1186/1471-2164-7-8.

Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M. Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic Acids Res*, 35 (21):7188–96, 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm864.

Ji Qi, Hong Luo, and Bailin Hao. Cvtree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res*, 32(Web Server issue):W45–7, 7 2004a. ISSN 1362-4962. doi: 10.1093/nar/gkh362.

Michael S. Rappé and Stephen J. Giovannoni. The uncultured microbial majority. *Annu Rev Microbiol*, 57:369–94, 2003. ISSN 0066-4227. doi: 10.1146/annurev.micro.57.030502.090759.

Erik M. Rauch and Yaneer Bar-Yam. Theory predicts the uneven distribution of genetic diversity within species. *Nature*, 431(7007):449–52, 9 2004. ISSN 1476-4687. doi: 10.1038/nature02745.

Michael Reich, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P. Mesirov. Nat genet, 5 2006. ISSN 1061-4036.

Oleg N. Reva and Burkhard Tümmler. Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics*, 5:90, 7 2004. ISSN 1471-2105. doi: 10.1186/1471-2105-5-90.

Christian S. Riesenfeld, Patrick D. Schloss, and Jo Handelsman. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet*, 38:525–52, 2004. ISSN 0066-4197. doi: 10.1146/annurev.genet.38.072902.091216.

David De Roure, Carole Goble, and Robert Stevens. The design and realisation of the myexperiment virtual research environment for social sharing of workflows. *Future Generation Computer Systems*, 25:561–567, 2 2009.

R. Sandberg, G. Winberg, C. I. Bränden, A. Kaske, I. Ernberg, and J. Cöster. Capturing whole-genome characteristics in short sequences using a naïve bayesian classifier. *Genome Res*, 11 (8):1404–9, 8 2001. ISSN 1088-9051. doi: 10.1101/gr.186401.

Patrick D. Schloss. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16s rrna gene-based studies. *PLoS Comput Biol*, 6(7): e1000844, 7 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000844.

Patrick D. Schloss and Jo Handelsman. Biotechnological prospects from metagenomics. *Curr Opin Biotechnol*, 14(3):303–10, 6 2003. ISSN 0958-1669.

Patrick D. Schloss and Jo Handelsman. Introducing dotur, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol*, 71(3): 1501–6, 3 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.3.1501-1506.2005.

Stephan C. Schuster. Next-generation sequencing transforms today's biology. *Nat Methods*, 5 (1):16–8, 1 2008. ISSN 1548-7105. doi: 10.1038/nmeth1156.

Allison K. Shaw, Aaron L. Halpern, Karen Beeson, Bao Tran, J. Craig Venter, and Jennifer B. H. Martiny. It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol*, 10(9):2200–2210, 8 2008. ISSN 1462-2920. doi: 10.1111/j.1462-2920.2008.01626.x.

Y. L. Simmhan, B. Plale, and D. Gannon. Karma 2: Provenance management for data-driven workflows. *International Journal of Web Services Research*, 5(2):1–22, 2008.

Gregory E. Sims, Se-Ran R. Jun, Guohong A. Wu, and Sung-Hou H. Kim. Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proc Natl Acad Sci U S A*, 2 2009. ISSN 1091-6490. doi: 10.1073/pnas.0813249106.

M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A*, 103(32):12115–20, 7 2006. ISSN 0027-8424. doi: 10.1073/pnas.0605127103.

J. Sroka, J. Hidders, P. Missier, and C. Goble. A formal semantics for the taverna 2 workflow model. *Journal of Computer and System Sciences*, 2009. doi: 10.1016/j.jcss.2009.11.009.

J. T. Staley and A. Konopka. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology*, 39(1):321–346, 1985. ISSN 0066-4227.

V. Stodden. Enabling reproducible research: licensing for scientific innovation. *Int'l J. Comm. L. & Pol'y*, 13:1, 2009a.

V. Stodden. The legal framework for reproducible scientific research: Licensing and copyright. *Computing in science & engineering*, 11(1):35–40, 2009b.

Yijun Sun, Yunpeng Cai, Li Liu, Fahong Yu, Michael L. Farrell, William McKendree, and William Farmerie. Esprit: estimating species richness using large collections of 16s rrna pyrosequences. *Nucleic Acids Res*, 5 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp285.

Shinichi Sunagawa, Todd Z. Desantis, Yvette M. Piceno, Eoin L. Brodie, Michael K. Desalvo, Christian R. Voolstra, Ernesto Weil, Gary L. Andersen, and Mónica Medina. Bacterial diversity and white plague disease-associated community changes in the caribbean coral montastraea faveolata. *ISME J*, 1 2009. ISSN 1751-7370. doi: 10.1038/ismej.2008.131.

Andreas Sundquist, Saharnaz Bigdeli, Roxana Jalili, Maurice L. Druzin, Sarah Waller, Kristin M. Pullen, Yasser Y. El-Sayed, M. Mark Taslimi, Serafim Batzoglou, and Mostafa Ronaghi. Bacterial flora-typing with targeted, chip-based pyrosequencing. *BMC Microbiol*, 7:108, 2007. ISSN 1471-2180. doi: 10.1186/1471-2180-7-108.

M. A. Sánchez, M. Vásquez, and B. González. A previously unexposed forest soil microbial community degrades high levels of the pollutant 2,4,6-trichlorophenol. *Appl Environ Microbiol*, 70(12):7567–70, 12 2004. ISSN 0099-2240. doi: 10.1128/AEM.70.12.7567-7570.2004.

M. A. Tanner, B. M. Goebel, M. A. Dojka, and N. R. Pace. Specific ribosomal dna sequences from diverse environmental settings correlate with experimental contaminants. *Appl Environ Microbiol*, 64(8):3110–3, 8 1998. ISSN 0099-2240.

James Taylor, Ian Schenck, Dan Blankenberg, and Anton Nekrutenko. Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics*, Chapter 10:Unit 10.5, 9 2007. ISSN 1934-340X. doi: 10.1002/0471250953.bi1005s19.

Hanno Teeling, Anke Meyerdierks, Margarete Bauer, Rudolf Amann, and Frank Oliver Glöckner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol*, 6(9):938–47, 9 2004a. ISSN 1462-2912. doi: 10.1111/j.1462-2920.2004.00624.x.

Hanno Teeling, Jost Waldmann, Thierry Lombardot, Margarete Bauer, and Frank Oliver Glöckner. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, 5:163, 10 2004b. ISSN 1471-2105. doi: 10.1186/1471-2105-5-163.

Susannah G. Tringe and Philip Hugenholtz. A renaissance for the pioneering 16s rrna gene. *Curr Opin Microbiol*, 9 2008. ISSN 1369-5274. doi: 10.1016/j.mib.2008.09.011.

Susannah Green Tringe and Edward M. Rubin. Metagenomics: Dna sequencing of environmental samples. *Nat Rev Genet*, 6(11):805–14, 11 2005. ISSN 1471-0056. doi: 10.1038/nrg1709.

Susannah Green Tringe, Christian von Mering, Arthur Kobayashi, Asaf A. Salamov, Kevin Chen, Hwai W. Chang, Mircea Podar, Jay M. Short, Eric J. Mathur, John C. Detter, Peer Bork, Philip Hugenholtz, and Edward M. Rubin. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–7, 4 2005. ISSN 1095-9203. doi: 10.1126/science.1107851.

Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. The human microbiome project. *Nature*, 449(7164):804–10, 10 2007. ISSN 1476-4687. doi: 10.1038/nature06244.

Peter J. Turnbaugh, Christopher Quince, Jeremiah J. Faith, Alice C. McHardy, Tanya Yatsunenko, Faheem Niazi, Jason Affourtit, Michael Egholm, Bernard Henrissat, Rob Knight, and Jeffrey I. Gordon. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci U S A*, 107(16):7503–8, 4 2010. ISSN 1091-6490. doi: 10.1073/pnas.1002355107.

Gene W. Tyson and Jillian F. Banfield. Cultivating the uncultivated: a community genomics perspective. *Trends Microbiol*, 13(9):411–5, 9 2005. ISSN 0966-842X. doi: 10.1016/j.tim.2005.07.003.

Gene W. Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, and Jillian F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 3 2004. ISSN 1476-4687. doi: 10.1038/nature02340.

Martin Täubel, Helena Rintala, Miia Pitkäranta, Lars Paulin, Sirpa Laitinen, Juha Pekkanen, Anne Hyvärinen, and Aino Nevalainen. The occupant as a source of house dust bacteria. *J Allergy Clin Immunol*, 124(4):834–40.e47, 10 2009. ISSN 1097-6825. doi: 10.1016/j.jaci.2009.07.045.

Y. Van de Peer, I. Van den Broeck, P. De Rijk, and R. De Wachter. Database on the structure of small ribosomal subunit rna. *Nucleic Acids Res*, 22(17):3488–94, 9 1994. ISSN 0305-1048.

Y. Van de Peer, S. Chapelle, and R. De Wachter. A quantitative map of nucleotide substitution rates in bacterial rrna. *Nucleic Acids Res*, 24(17):3381–91, 9 1996a. ISSN 0305-1048.

P. Vandamme, B. Pot, M. Gillis, P. de Vos, K. Kersters, and J. Swings. Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev*, 60(2):407–38, 6 1996. ISSN 0146-0749.

Kalin Vetsigian and Nigel Goldenfeld. Genome rhetoric and the emergence of compositional bias. *Proc Natl Acad Sci U S A*, 106(1):215–20, 1 2009. ISSN 1091-6490. doi: 10.1073/pnas.0810122106.

M. Vignuzzi, J. K. Stone, J. J. Arnold, C. E. Cameron, and R. Andino. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, 439 (7074):344–8, 12 2005. ISSN 1476-4687. doi: 10.1038/nature04388.

S. Voget, C. Leggewie, A. Uesbeck, C. Raasch, K-E E. Jaeger, and W. R. Streit. Prospecting for novel biocatalysts in a soil metagenome. *Appl Environ Microbiol*, 69(10):6235–42, 10 2003. ISSN 0099-2240.

Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Appl Environ Microbiol*, 73(16):5261–7, 8 2007. ISSN 0099-2240. doi: 10.1128/AEM.00062-07.

Yong Wang and Pei-Yuan Y. Qian. Conservative fragments in bacterial 16s rrna genes and primer design for 16s ribosomal dna amplicons in metagenomic studies. *PLoS One*, 4(10): e7401, 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0007401.

Falk Warnecke, Peter Luginbühl, Natalia Ivanova, Majid Ghassemian, Toby H. Richardson, Justin T. Stege, Michelle Cayouette, Alice C. McHardy, Gordana Djordjevic, Nahla Aboushadi, Rotem Sorek, Susannah G. Tringe, Mircea Podar, Hector Garcia Martin, Victor Kunin, Daniel Dalevi, Julita Madejska, Edward Kirton, Darren Platt, Ernest Szeto, Asaf Salamov, Kerrie Barry, Natalia Mikhailova, Nikos C. Kyrpides, Eric G. Matson, Elizabeth A. Ottesen, Xinning Zhang, Myriam Hernández, Catalina Murillo, Luis G. Acosta, Isidore Rigoutsos, Giselle Tamayo, Brian D. Green, Cathy Chang, Edward M. Rubin, Eric J. Mathur, Dan E. Robertson, Philip Hugenholtz, and Jared R. Leadbetter. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450(7169):560–5, 11 2007. ISSN 1476-4687. doi: 10.1038/nature06269.

L. G. Wayne, D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E. Murray, and E. Stackebrandt. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic and Evolutionary Microbiology*, 37(4):463, 1987.

R. A. Welch, V. Burland, G. Plunkett, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S-R R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. T. Mobley, M. S. Donnenberg, and F. R. Blattner. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic escherichia coli. *Proc Natl Acad Sci U S A*, 99(26):17020–4, 12 2002. ISSN 0027-8424. doi: 10.1073/pnas.252529799.

Paul Wilmes, Sheri L. Simmons, Vincent J. Denef, and Jillian F. Banfield. The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev*, 33(1):109–32, 1 2009. ISSN 0168-6445. doi: 10.1111/j.1574-6976.2008.00144.x.

Dongying Wu, Amber Hartman, Naomi Ward, and Jonathan A. Eisen. An automated phylogenetic tree-based small subunit rrna taxonomy and alignment pipeline (stap). *PLoS ONE*, 3 (7):e2566, 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0002566.

Dongying Wu, Philip Hugenholtz, Konstantinos Mavromatis, Rüdiger Pukall, Eileen Dalin, Natalia N. Ivanova, Victor Kunin, Lynne Goodwin, Martin Wu, Brian J. Tindall, Sean D. Hooper, Amrita Pati, Athanasios Lykidis, Stefan Spring, Iain J. Anderson, Patrik D'haeseleer, Adam Zemla, Mitchell Singer, Alla Lapidus, Matt Nolan, Alex Copeland, Cliff Han, Feng Chen, Jan-Fang F. Cheng, Susan Lucas, Cheryl Kerfeld, Elke Lang, Sabine Gronow, Patrick Chain, David Bruce, Edward M. Rubin, Nikos C. Kyrpides, Hans-Peter P. Klenk, and Jonathan A. Eisen. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*, 462(7276):1056–60, 12 2009. ISSN 1476-4687. doi: 10.1038/nature08656.

Jan Wuyts, Yves Van de Peer, Tina Winkelmans, and Rupert De Wachter. The european database on small subunit ribosomal rna. *Nucleic Acids Res*, 30(1):183–5, 1 2002. ISSN 1362-4962.

Y. Yang, J. Yao, S. Hu, and Y. Qi. Effects of agricultural chemicals on dna sequence diversity of soil microbial community: A study with rapd marker. *Microb Ecol*, 39(1):72–79, 1 2000a. ISSN 0095-3628.

S. Yooseph, G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter. The sorcerer ii global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biol*, 5(3):e16, 3 2007. ISSN 1545-7885. doi: 10.1371/journal.pbio.0050016.

Noha Youssef, Cody S. Sheik, Lee R. Krumholz, Fares Z. Najar, Bruce A. Roe, and Mostafa S. Elshahed. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16s rrna gene-based environmental surveys. *Appl Environ Microbiol*, 75(16):5227–36, 8 2009. ISSN 1098-5336. doi: 10.1128/AEM.00592-09.

Yanan Yu, Mya Breitbart, Pat McNairnie, and Forest Rohwer. Fastgroupii: a web-based bioinformatics platform for analyses of large 16s rdna libraries. *BMC Bioinformatics*, 7:57, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-57.

Fengfeng Zhou, Victor Olman, and Ying Xu. Barcodes for genomes and applications. *BMC Bioinformatics*, 9(1):546, 12 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-546.