GEOREFERENCED TREES AND THE PHYLOGENETIC SIMILARITY OF BIOLOGICAL COMMUNITIES

by

Donovan H. Parks

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
July 2012

© Copyright by Donovan H. Parks, 2012

DALHOUSIE UNIVERSITY

Faculty of Computer Science

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "GEOREFERENCED TREES AND THE PHYLOGENETIC SIMILARITY OF BIOLOGICAL COMMUNITIES" Donovan H. Parks in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Dated: July 31, 2012

External Examiner:

_____

Dr. Josh Neufeld

Research Supervisor:

_____

Dr. Robert Beiko

Examining Committee:

_____

Dr. Christian Blouin

_____

Dr. Stephen Brooks

Departmental Representative: _____

DALHOUSIE UNIVERSITY

DATE: July 31, 2012

AUTHOR: Donovan H. Parks

TITLE: GEOREFERENCED TREES AND THE PHYLOGENETIC SIMILARITY OF BIOLOGICAL COMMUNITIES

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: PhD CONVOCATION: October YEAR: 2012

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions. I understand that my thesis will be electronically available to the public.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

The author attests that permission has been obtained for the use of any copyrighted material appearing in the thesis (other than the brief excerpts requiring only proper acknowledgement in scholarly writing), and that all such use is clearly acknowledged.

_____

Signature of Author

Dedicated to

Mom, Dad, Sister,

&

Gwen

Table of Contents

List of Tables

List of Figures

Abstract

Culture-independent DNA sequencing is being used to recover genetic material directly from environmental samples. This has spurred large-scale community efforts to catalogue the diversity of life and its geographic distribution using molecular data. These initiatives stand to revolutionize our understanding of the processes that shape biodiversity and may ultimately provide critical information for setting public health, environmental, and economic policies. To achieve these aims new tools are required to effectively explore these large biogeographic datasets.

This thesis introduces a novel technique for visualizing hierarchically organized data in a geographic context that illustrates the influence of a geographic or environmental gradient on the phylogenetic relationships between organisms or the similarity of biological communities. This technique is incorporated into GenGIS, open-source software that supports the integration of digital map data with genetic sequences and environmental information from multiple sample sites. GenGIS addresses the need for an interactive geospatial analysis environment capable of handling large biogeographic datasets where a wealth of sequence data is available for each sample site. This is accomplished through a rich set of analysis options that produce georeferenced visualizations for data exploration and hypothesis generation. Studies conducted by myself and other research groups have used GenGIS to investigate the diversity of viruses, bacteria, plants, animals, and even language families.

I then explore measures of beta diversity that aim to assess the influence of geographic or environmental gradients on the similarity of biological communities. This thesis examines phylogenetic beta-diversity measures that determine community variation by considering the relationships between organisms in a phylogenetic tree. A large comparative study is performed in order to assess specific properties and performance characteristics of these measures. Many measures of phylogenetic beta diversity were found to be robust to sequence clustering, the addition of an outlying basal lineage, root placement, and the presence of rare organisms. Additionally, performance was found to differ substantially under different models of community variation. This thesis then describes how an important class of phylogenetic beta-diversity measures can be calculated over phylogenetic networks in order to account for uncertainty and conflict in inferred ancestral relationships.

List of Abbreviations and Symbols Used

BLAST	Basic local alignment search tool
BPR	Blue-absorbing proteorhodopsin
bp	Base pair
CL	Complete lineage
COI	Cytochrome *c* oxidase I
CT	Complete tree
DNA	Deoxyribonucleic acid
E-value	Expectation value
GAP	Geographic axis polyline
GDAL	Geospatial Data Abstraction Library
GIS	Geographic information system
GLL	Geographic layout line
GOS	Global Ocean Sampling Expedition
GPR	Green-absorbing proteorhodopsin
HIV	Human immunodeficiency virus
MLST	Multilocus sequence typing
MNND	Mean nearest neighbour distance
MPD	Mean phylogenetic distance
MRCA	Most recent common ancestor
mtDNA	Mitochondrial DNA
NAST	Nearest alignment space termination
NP	Nondeterministic polynomial
OLNO	Optimal leaf node ordering
OSCM	One-sided crossing minimization
OTU	Operational taxonomic unit
P	Deterministic polynomial
PC	Principal coordinate
PCoA	Principal coordinate analysis
RAxML	Randomized Accelerated Maximum Likelihood

RDP	Ribosomal Database Project
RNA	Ribonucleic acid
s.d.	Standard deviation
ST	Sequence type
rRNA	Ribosomal RNA
TLL	Tree layout line
UniFrac	Unique fraction
UPGMA	Unweighted pair group method with arithmetic mean
WAG	Whelan and Goldman

Acknowledgements

It is a pleasure to thank the many people who have supported me during my Ph.D. and made the journey worth traveling.

I would like to express my gratitude to my supervisor, Dr. Robert Beiko, whose insights and guidance have greatly strengthened this research and whose infectious optimism has kept me moving steadily forward. This research would not have come to fruition without Rob and his expertise in biology and computer science. He is truly a "man of two worlds".

This thesis has directly benefited from my committee members. Christian Blouin and Stephen Brooks have provided feedback throughout the development of my thesis, often raising questions that have directly influenced the direction of my research. Michael McAllister aided the development of GenGIS, and just as importantly has been a strong and spirited competitor on the Ultimate field.

I am indebted to my colleagues for their advice and support. Dennis Wong has been a fountain of knowledge, and aided me greatly in navigating the terminology and theories of biology. Norman MacDonald has provided valuable insights into many machine learning related issues. Michael Porter has been instrumental in the development of GenGIS and all things OS X related. Robert Eveleigh has acted as a sounding board, helping me to explore the possibilities of my research. Chris Whidden has helped me dip my toe into computational complexity theory. Timothy Mankowski has pushed forward the development of GenGIS. Conor Meehan, Morgan Langille, and Mahdi Shafiei have provided keen insights into the numerous papers we have discussed. All of my colleagues have given me an intellectually stimulating and enjoyable environment in which to work.

Funding for this research was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Killam Trusts. I am grateful for their support.

Without the love and support of my family, this thesis would not have been possible. My parents have always encouraged me to pursue with vigor what brings me joy. They have given me the courage and strength to do just that. I am thankful to my sister for all her love and her amazing capacity to find the best in people. Finally, I would like to express my admiration for Gwen Williams. Without her love, patience, and encouragement I would be lost.

Donovan Parks
Halifax, July 2012

Chapter 1

Introduction

1.1 Studying the Diversity of Life

Biodiversity is, in simple terms, the variety of life that exists in a region. It is often used in a broad sense to include not only variation in species, but also variation at the genomic level. Biodiversity studies may aim to enumerate all species and their relative abundance in a specified area or endeavour to account for the genetic variation of a particular gene across the globe. Major efforts have been put forth to catalogue both prokaryotic (i.e., bacterial and archaeal) and eukaryotic life by considering the diversity of specific genes. Studies of prokaryotic diversity focus almost exclusively on the small-subunit ribosomal RNA (16S rRNA) gene, a structural component of the ribosome, which is responsible for creating proteins based on an organism's DNA. As such, all known prokaryotes contain the 16S rRNA gene allowing it to be used as a ubiquitous "taxonomic marker" (Pace et al. 1984; Lane et al. 1985). The Ribosomal Database Project (RDP) currently contains over 2 million 16S rRNA gene sequences (Cole et al. 2009), and this likely represents only a small fraction of this gene's total diversity. The Consortium for the Barcode of Life parallels the earlier efforts to assess prokaryotic diversity by utilizing the cytochrome *c* oxidase I (COI) gene to study eukaryotic diversity (Hebert et al. 2003; Ratnasingham and Hebert 2007). Studies specific to the biodiversity of human populations have long relied on maternally inherited mitochondrial DNA (mtDNA) to infer patterns of human migration (Bandelt et al. 1995; Cooper et al. 2001), although the use of paternally inherited Y chromosome DNA and even complete genomes are becoming increasingly popular (Pakendorf and Stoneking 2005).

This thesis aims to advance 2 fields of study within the discipline of biodiversity, comparative biogeography and exploratory phylogeography, by using computer simulations to assess the properties of biodiversity measures, proposing extensions to

these measures of biodiversity, and introducing new algorithms and interactive visualizations for exploring the hierarchical relationships between biological units.

1.1.1 Comparative Biogeography

Biogeography is the study of geographic patterns of biodiversity. The field is based on the simple observation that life varies in a highly nonrandom fashion from place to place. Studies may focus on the biogeography of a particular gene, a specific species, or even entire biological communities, i.e., groups of interdependent organisms living and interacting with each other in a specific habitat or location. Several measures have been proposed for assessing the similarity of biological communities (Legendre and Legendre 1998; Magurran 2004). By comparing the measured variation among communities with geographic (e.g., elevation, ocean current patterns), geological (e.g., mountain-forming events, glaciation events), or environmental (e.g., temperature, salinity, or season) factors we can explore the influence of these factors on biodiversity (Martiny et al. 2006; Green et al. 2008; Lomolino et al. 2010). The advent of high-throughput culture-independent DNA sequencing has revolutionized our ability to assess the biogeography of microorganisms by permitting deep sequencing of marker genes directly from environmental samples (Fig. 1.1). For example, a landmark study by Lozupone and Knight (2007) assessed the variation of bacterial communities from a wide range of environments and found bacterial diversity to be largely driven by the salinity of habitats. Studies have also compared the biogeographic patterns of eukaryotes and prokaryotes, and have found instances where these patterns agree (Green and Bohannan 2006; Fuhrman et al. 2008) and instances where they disagree (Bryant et al. 2008; Wang et al. 2011). Ultimately, these studies aim to catalogue the contemporary distribution of life and to determine the processes that gave rise to this distribution. Such an understanding would inform conservation decisions, allow us to predict the consequences of global climate change on individual species, and enable us to more accurately assess the environmental impact of commercial and industrial activities.

Several notable initiatives are underway to characterize the global biogeography of prokaryotes and eukaryotes. The Earth Microbiome Project is a major international effort

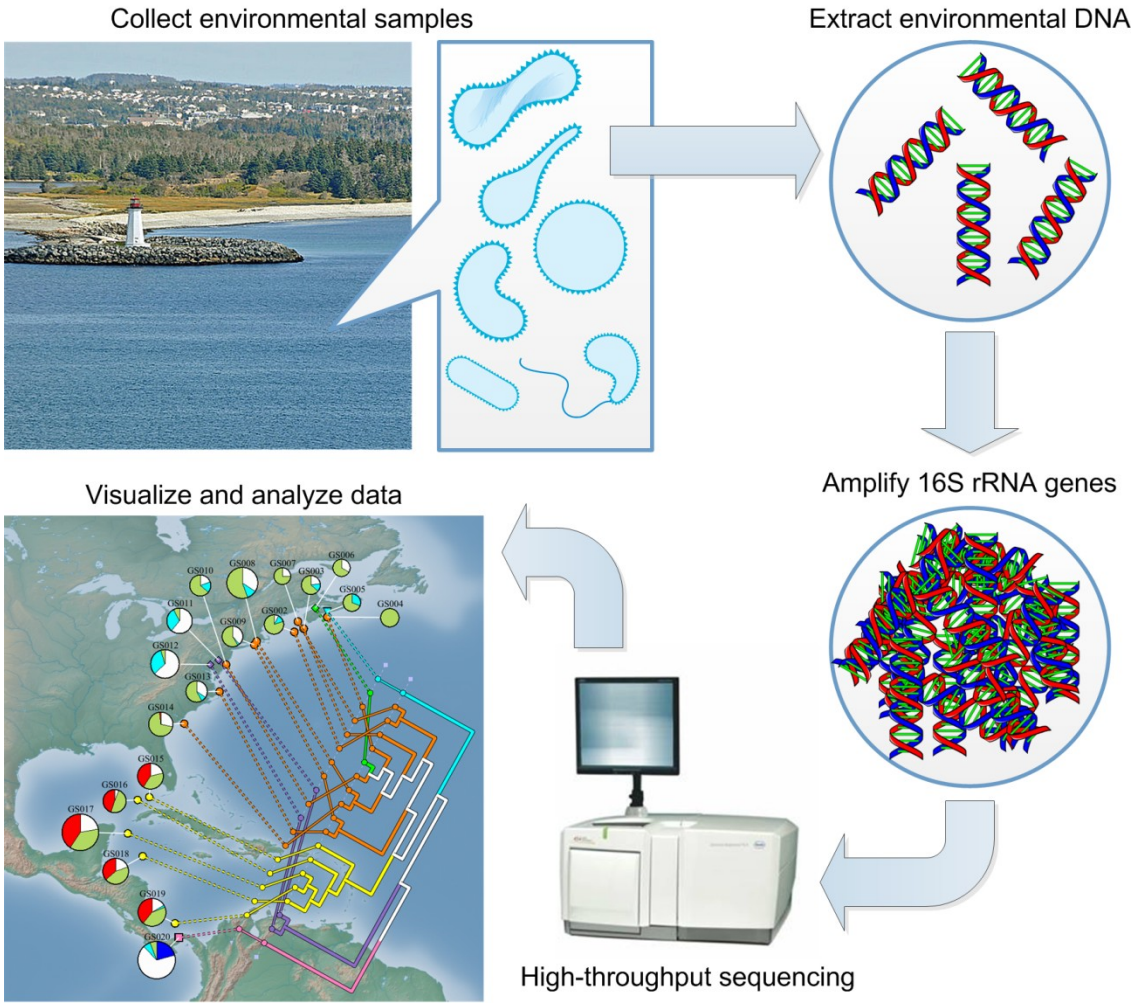


Figure 1.1. Standard workflow for high-throughput marker gene studies. Environmental samples are collected and brought back to the lab for bulk extraction of DNA. Specific marker genes are amplified (e.g., 16S rRNA gene for prokaryotes) from the extracted DNA using a well-established technique relying on barcoded, conserved primer pairs. High-throughput sequencing allows the cost-effective recovery of sequence data from dozens to hundreds of environmental samples. The resulting dataset is visualized and analyzed to gain insight into the biogeography of the sampled microbial communities.

to analyze microbial community across the globe from all of Earth's various biomes (Gilbert et al. 2010). This will include assessment of both taxonomic diversity based on 16S rRNA gene sequencing and functional diversity as determined using high-throughput "shotgun" metagenomic sequencing of bulk DNA from environmental samples. These efforts will produce a Gene Atlas of georeferenced genomic data along with metadata specifying environmental and other attributes of interest (Yilmaz et al. 2011). The

spatial distribution of microorganisms is also being studied in more abstract and restricted geographic contexts. Initiatives are underway to characterize the biogeography of prokaryotes on the human body (Turnbaugh et al. 2007; Costello et al. 2009; Caporaso et al. 2011) including detailed studies of specific organs (Nasidze et al. 2009; Stearns et al. 2011), and within indoor environments related to human health such as public washrooms and hospitals (Flores et al. 2011; Kembel et al. 2012).

The Map of Life initiative seeks to provide a single integrated service providing access to and visualization of the geographic distribution of all eukaryotic species (Jetz et al. 2012). By combining currently disparate sources of information, a more global view of biodiversity will be possible which will almost certainly reveal biogeographic patterns that are currently unknown and provide a tool for monitoring changes in global biodiversity. Within Canada, the Biomonitoring 2.0 project has been established for ecosystem monitoring which aims to take advantage of advances in high-throughput DNA sequencing in order to more thoroughly monitor changes in both microbial prokaryotes and eukaryotes (Baird and Hajibabaei 2012). Given Canada's economic reliance on natural resources, rapid biomonitoring will provide much needed improvements in assessing the impact of industrial processes on biodiversity and provide an early warning system indicating when environmental stresses are unduly affecting local ecosystems.

1.1.2 Exploratory Phylogeography

Phylogeography is the study of the evolutionary relationships between organisms in the context of their geographic distribution. It is the discipline within biogeography concerned with how evolutionary and ecological processes have given rise to the contemporary distribution of organisms (Avise 2000). Traditionally, phylogeography has been applied to understand intraspecific ("within species") patterns of spatial variation and this remains an active area of research. Recently, Kidd (2010) proposed a Map of Life which incorporates not only the distribution of species as discussed above, but also the evolutionary relationships between organisms. Kidd's Map of Life aims to contain all we know about biogeography "threaded through a dynamic earth history, capturing the spatiotemporal pathways that underlie current and past patterns of biodiversity." This

would allow direct investigation of geophylogenies, i.e., an evolutionary tree or phylogeny describing the ancestral relationships between geographically referenced organisms. Identifying similar geophylogenies spanning distinct sets of species will provide insights into processes driving speciation, the evolution of specific traits, and patterns of extinction.

Databases already exist which describe the evolutionary history of organisms (Piel et al. 2009), and extensive effort has been focused on the development of systems for visualizing and analyzing evolutionary data within a geospatial context (Kidd and Liu 2008; Hill and Guralnick 2010; Janies et al. 2010; Laffan et al. 2010; Bielejec et al. 2011). Initial efforts have resulted in the development of systems for exploring the global spread of emergent infectious diseases such as Influenza (Janies et al. 2007; Lemey et al. 2009; Parks, MacDonald, et al. 2009). These systems allow public health workers to explore how diseases are spreading, understand the distribution of different strains, and determine when and where pathogens have acquired mutations leading to drug resistance. This information is critical for proper distribution of medical resources to deal with existing outbreaks and for informing future policies on disease prevention.

Many of the exploratory techniques in phylogeography can also be applied to investigate the similarity of entire populations or communities within a spatiotemporal context. This is especially applicable within microbial ecology where communities have been proposed as the base unit of evolutionary studies (Doolittle and Zhaxybayeva 2010). However, the similarity of eukaryotes, including human populations, can also be usefully explored at the community or population level. The major difference in such analyses is that the hierarchical relationships between entities is no longer a strictly historical one reflecting the shared ancestry of organisms, but is instead a more abstract measure of the similarity between entire communities. Nevertheless, in both cases we are interested in determining processes that give rise to hierarchically organized units of biodiversity. In this thesis, I use the term *geotree* to refer to *any* tree where leaf nodes, and possibly internal nodes, are georeferenced and reserve the term geophylogeny for geographically referenced phylogenies.

1.1.3 Progress of Major Biodiversity Initiatives

Major initiatives such as the Barcode of Life, the Earth Microbiome Project, the Map of Life, and the Biomonitoring 2.0 project stand to revolutionize our ability to explore the current biodiversity of earth, identify and track the spread of emergent infectious diseases, and monitor changes in community structure. Not only will these projects advance our understanding of biodiversity, they stand to directly benefit public health, inform economic policy, and guide conservation efforts. These are lofty goals and extensive work remains to make these community efforts a reality.

One aim of this thesis is to help move these community initiatives forward by addressing 2 notable short-comings: 1) the lack of free and open-source software implementing efficient algorithms for visualizing and analyzing large molecular biogeographic datasets, and 2) an inadequate understanding of the properties and performance characteristics of different phylogenetic measures of community similarity. Current work has focused primarily on data collection (Ratnasingham and Hebert 2007; Turnbaugh et al. 2007; Gilbert et al. 2010), integration of disparate data sources (Kidd 2010; Jetz et al. 2012), and initial processing and analysis of genetic data (Schloss et al. 2009; Caporaso et al. 2010). Relatively little attention has been given to the development of interactive tools for exploring and analyzing large biogeographic datasets, and recent efforts are generally restricted in scope and reliant on proprietary software (Kidd and Liu 2008; Janies et al. 2010; Laffan et al. 2010; Bielejec et al. 2011). Measures of community similarity have a long history within ecology (Jaccard 1901; Bray and Curtis 1957) and are an integral part of many biogeographic analyses (Legendre and Legendre 1998; Magurran 2004; Anderson et al. 2011). Recent extensions have begun to incorporate phylogenetic information (Martin 2002; Lozupone and Knight 2005; Hardy and Senterre 2007; Webb et al. 2008) and properties of these extensions are less well understood, though a few notable studies have been performed (Nipperess et al. 2010; Root and Nelson 2011; Swenson 2011). Given the importance of these phylogenetic measures in biogeographic studies (Graham and Fine 2008; Faith et al. 2009), a thorough understanding of their properties and relative performances is essential as are efforts to improve their utility. In this thesis, I discuss how measures considering the phylogenetic relatedness of sequences can be extended to account for phylogenetic uncertainty.

1.2 Goals of Dissertation

Throughout my research, I have aimed to develop widely applicable methods for comparative biogeography and exploratory phylogeography. My research can be organized into 2 general goals: 1) to develop an interactive environment utilizing efficient algorithms for visualization and analyzing large biogeographic datasets and 2) to assess properties and performance characteristics of phylogenetic measures of community similarity which have not been adequately examined and to extend these measures to account for phylogenetic uncertainty.

This first goal has been realized through the development of GenGIS, a free and open-source software platform that combines digital map, environmental, and genetic datasets (Parks, Porter, et al. 2009; in preparation, Parks et al. 2012). GenGIS provides a wide range of visualization tools for exploring biogeographic data including a novel phylogeographic technique for exploring hierarchically organized units of biodiversity (Parks and Beiko 2009). Several common analytical techniques are provided in GenGIS including the calculation of biodiversity indices and standard statistical techniques such as linear regression and the Mantel test, a statistical test of the correlation between 2 matrices (Mantel 1967). The functionality of GenGIS can be extended using a plugin framework which allows for the easy development of custom visualizations and analyses. Further development of GenGIS aims to connect it with online data sources and major community initiatives. GenGIS would provide a powerful frontend for interacting with the Gene Atlas being created by the Earth Microbiome Project and is well-suited for navigating a Map of Life. The Biomonitoring 2.0 project recently received funding from Genome Canada to further develop GenGIS as a tool for ecosystem monitoring, and preliminary research has demonstrated how GenGIS can be used to study emergent infectious diseases such as influenza (Parks, MacDonald, et al. 2009).

To achieve the second goal, I performed a comparative analysis of 39 methods used to assess community similarity that explicitly account for the shared ancestry of organisms within a community (in press, Parks and Beiko 2012a). I then extended these measures to allow for the assessment of community similarity when sequences are related by a phylogenetic network, a generalization of phylogenetic trees, which can account for uncertainty in inferred evolutionary histories (in press, Parks and Beiko 2012b). A

phylogeographic assessment of community relationships can be performed within GenGIS along with direct visualization of the similarity between select pairs of communities. These similarity measures are essential for understanding spatiotemporal changes in biodiversity and as such will form the basis for many analyses within the Biomonitoring 2.0 project and other initiatives.

1.3 Background

This thesis is concerned with the visualization and analysis of molecular data resulting from the sequencing of environmental samples (Fig. 1.1). In this section, I introduce terminology and concepts from both the life and computational sciences that underlie the visualization and analysis workflow central to this thesis (Fig. 1.2). All datasets examined in this thesis were obtained from public databases (Figs. 1.2a and 1.2b). Basic terminology for discussing phylogenetic trees is given followed by an introduction to multiple sequence alignments (Fig. 1.2c), the first step in tree inference (Fig. 1.2d). Although this thesis is not concerned with the development of tree inference methods, it directly addresses the visualization of trees within a geographic context (Fig. 1.2e; Chapters 2 and 3) and aims to assess and generalize measures of biodiversity which explicitly account for phylogenetic relationships (Fig. 1.2f; Chapters 4 and 5). As such, measures used to assess biodiversity are discussed with particular attention given to phylogenetic beta-diversity measures. I then introduce multivariate statistical techniques used throughout this thesis for visualizing the results of beta-diversity analyses (Fig. 1.2g). Finally, I conclude this section with a brief discussion on computational complexity in order to aid the reader in understanding the algorithmic aspects of Chapter 2.

1.3.1 Describing Hierarchical Relationships

A central theme of this thesis is the development of methods for visualizing and defining the relationships between individual entities, which may represent specific organisms or entire communities of organisms. These relationships are often described in a hierarchical fashion in order to explore the association between groups or clusters of entities. Hierarchical relationships play a critical role in biology in the form of

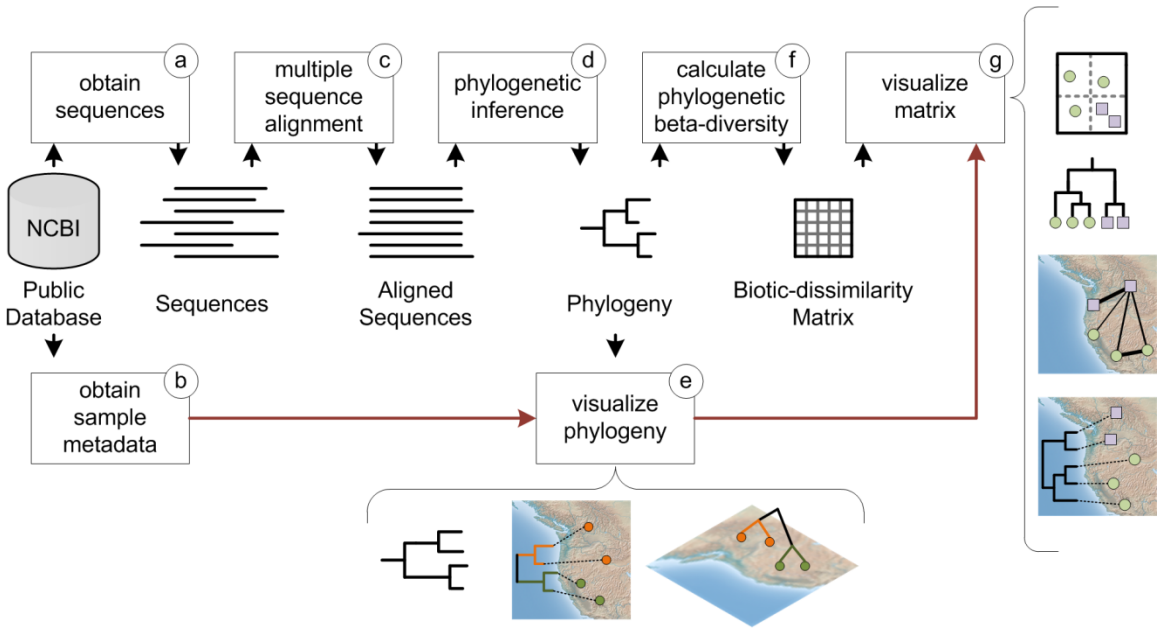


Figure 1.2. General workflow for exploratory data analysis. (**a**) Publically available genetic sequences are obtained from public databases. (**b**) Metadata describing the source of each sequence is obtained along with any other available information (e.g., geographic coordinates, environmental parameters of sample site, time information). (**c**) A multiple sequence alignment is inferred for all sequences. (**d**) A phylogenetic tree is inferred from the multiple sequence alignment. (**e**) A tree viewer is used to examine the phylogeny. Georeferenced phylogenies or "geophylogenies" can be visualized explicitly within a geographic context as either a two- or three-dimensional tree, with visual properties of the tree (e.g., colour of nodes or branches) set to reflect metadata attributes such as habitat type or sampling period. (**f**) The similarity of communities is assessed with a phylogenetic beta-diversity measure. (**g**) The relationship between communities is visualized using standard multivariate statistical techniques. Georeferenced communities can be visualized explicitly within a geographic context and metadata attributes used to set visual properties in order to help explore the data.

phylogenetic trees that describe the ancestral relationships between species or genetic sequences. This thesis describes new methods for visualizing phylogenetic trees in a geospatial context and develops statistics which provide a univariate summary of the phylogenetic relationships between taxa from different communities or populations. Here we introduce the basic terminology used to describe hierarchical relationships within the life and computational sciences.

Hierarchical relationships can be represented by either an *unrooted* or *rooted* tree

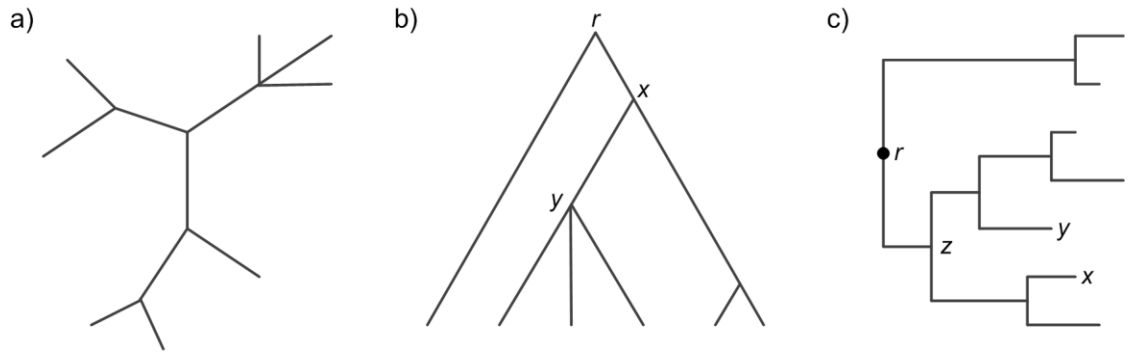


Figure 1.3. Examples of rooted and unrooted trees. (**a**) An unrooted tree with 8 leaf nodes, 5 internal nodes, and 12 branches. (**b**) A multifurcating rooted tree with 6 leaf nodes, 4 internal nodes, and 9 branches. Node *r* is the root of the tree. The tree is multifurcating as node *y* has 3 children. Node *x* has a height of 2. (**c**) A rooted bifurcating tree with 7 leaf nodes, 6 internal nodes, and 12 branches. Node *r* is the root of the tree. The tree is bifurcating as all internal nodes have exactly 2 children. The most recent common ancestor (MRCA) of nodes *x* and *y* is node *z*. These nodes are descendants of node *z*. Node *z* has a height of 3, nodes *x* and *y* have a height of zero.

where entities are depicted by *nodes* or *vertices* and the relationships between entities are depicted by *branches* or *edges* (Fig. 1.3). A tree is a *connected graph* with *no cycles* (i.e., closed paths). Rooted trees contain a unique node, called the *root*, which imputes a sense of direction to each branch and allows ancestral relationships to be defined. A rooted tree is a *directed acyclic graph* as each branch has a direction and a tree contains no cycles. The *children* of node *n* are all nodes connected to node *n* by branches directed away from the root, and node *n* is the *parent* of its children. *Leaf* or *terminal* nodes are nodes which have exactly zero children, and an *internal* node is any node which has children. Within a phylogenetic tree leaf nodes are often called *taxa* (singular: taxon). The *subtree* of node *n* is the rooted tree which is formed by removing the branch connecting node *n* to its parent node and making node *n* the root node. In phylogenetics, a rooted subtree is called a *clade*, and a group of taxa within a clade which contains no additional taxa are referred to as a *monophyletic group*. The *descendants* of node *n* are all nodes within the subtree rooted at node *n*. All nodes along the path from node *n* to the root node are *ancestors* of node *n*. The *height* of node *n* is the longest path from node *n* to a descendant leaf node of node *n*, and the height of a tree is the height of the root node. The *depth* of node *n* is the length of

the path from node *n* to the root node. The *most recent common ancestor* (MRCA) of 2 nodes *x* and *y* is the deepest node which is an ancestor of both nodes. The *degree* of a node is the number of branches connected to that node.

A rooted tree is called *bifurcating* if every internal node has exactly 2 children. Nodes with more than 2 children are called *multifurcating* or *k-ary* nodes. A rooted tree with one or more multifurcating nodes is a multifurcating tree. In a *complete k-ary tree* all internal nodes have exactly *k* children. A *layer* in a complete *k*-ary tree is the set of all nodes with the same height.

The above terminology can be applied to an unrooted tree by defining an imaginary root node along any branch. A rooted tree can be turned into an unrooted tree by removing the root node. For trees where ancestral relationships have explicit meaning, the placement of a root node within an unrooted tree must be done in a manner which correctly identifies this ancestry (Section 1.3.2).

Two representations are commonly used to depict trees: phylograms and cladograms. In a phylogram, the length of a branch is drawn to reflect the relative similarity of the nodes the branch connects. In contrast, cladograms simply depict the relationships between nodes and no meaning is assigned to the length of branches. In this thesis, phylograms and cladograms are almost exclusively represented as either two- or three-dimensional node-link diagrams with branches in a "slanted" (e.g., Fig. 1.3b) or "rectangular" (e.g., Fig. 1.3c) layout.

1.3.2 Inferring Ancestral Relationships

The evolutionary relationships between biological entities or taxa (e.g., species, individual organisms, or specific genes) are typically depicted as a rooted bifurcating tree. Leaf nodes represent present-day taxa, internal nodes correspond to hypothetical ancestors, and the lengths of branches indicate the amount of change or time between 2 taxa. Any set of phylogenetically informative data can be used to infer ancestral relationships such as changes in physical structure (e.g., a phenotypic character such as beak length) or gene content (e.g., a genotypic character such as a nucleotide substitution).

Aligning Molecular Sequences

All phylogenies in this thesis are inferred from DNA or protein sequences. These sequences are presumed to be *homologous*, i.e., evolved from a common ancestral sequence. Over evolutionary time, homologous DNA sequences will diverge as a result of:

- Substitutions: the change of a nucleotide from one character to another.
- Insertions: the addition of a nucleotide into a sequence.
- Deletions: the loss of a nucleotide from a sequence.

Two sequences are presumed to be homologous if they exhibit patterns of sequence similarity that are unlikely to be due to chance. Homology is most often inferred using the basic local alignment search tool (BLAST) algorithm which produces an expectation value (E-value) indicating the number of times an alignment as good as the one observed between 2 sequences will occur within a sequence database of a particular size (Altschul et al. 1990; Altschul et al. 1997). In practice, a pair of sequences producing a sufficiently small E-value are presumed to be homologous, though selecting an appropriate threshold can be challenging.

To assess the similarity of molecular sequences they must be aligned such that characters (i.e., nucleotides or amino acids) within a single column are homologous (Fig. 1.4). Given a substitution matrix indicating the "cost" of substitutions and insertions/deletions, the optimal alignment of 2 sequences can be found using a dynamic programming approach (Needleman and Wunsch 1970; Smith and Waterman 1981). BLAST also relies on a dynamic programming approach for aligning sequences, but makes use of a number of heuristics to substantially reduce the time required to identify good alignments within large sequence databases. Unfortunately, dynamic programming approaches scale poorly and are impractical for aligning multiple sequences (Wang and Jiang 1994). Many multiple sequence alignment heuristics have been proposed to address this limitation (Notredame et al. 2000; Edgar 2004; Bradley et al. 2009). The most widely used is the progressive alignment heuristic in which pairs of similar sequences are first aligned, followed by the pairwise alignment of groups of 2 or more aligned sequences (Thompson et al. 1994).

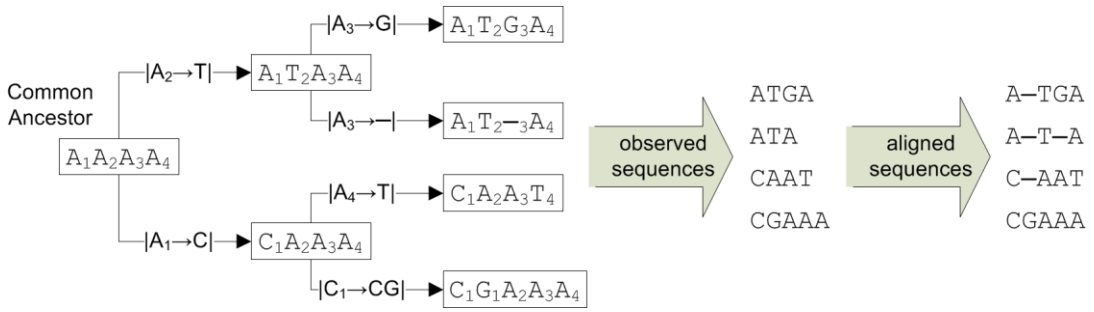


Figure 1.4. An example of 4 homologous sequences and their multiple sequence alignment. The evolution of each contemporary sequence is shown and includes several substitutions, a single deletion (−), and a single insertion. A multiple sequence alignment must be inferred for the observed sequences without explicit knowledge of the evolutionary history of these sequences. Ideally, each column in the aligned sequences will be homologous, i.e., derived from a single common ancestral nucleotide. The alignment shown here correctly aligns each homologous nucleotide.

Inferring Phylogenetic Trees

Phylogenetic inference typically begins with a multiple sequence alignment (Fig. 1.2c) with similarities and differences in each column of the alignment taken as a signal of the evolutionary relatedness of the aligned sequences. Both distance- and sequence-based methods have been proposed for constructing phylogenetic trees from a multiple sequence alignment. Distance-based methods begin by constructing a matrix indicating the evolutionary distance between each pair of sequences. Shared ancestry is determined from the distance matrix using methods such as neighbour-joining (Saitou and Nei 1987) or the hierarchical clustering algorithms discussed in Section 1.3.3 (notably, UPGMA). Distance-based methods tend to be computationally efficient, making it tractable to infer evolutionary relationships between tens of thousands of sequences. However, there is a loss of signal as inference of shared ancestry is based purely on the pairwise distance between sequences instead of the presumed homologous columns in the multiple sequence alignment.

Sequence-based methods infer a tree directly from the multiple sequence alignment using one of 3 main approaches: maximum parsimony, maximum likelihood, or Bayesian. Maximum parsimony and maximum likelihood attempt to search over the space of all possible trees and score the quality of each tree according to a specific model

of evolution. Since tree space is extremely large it is not practical to consider all possible trees and a search strategy must be used to identify promising trees to evaluate. For example, with 10 taxa there are approximately 2 million possible tree topologies. Maximum parsimony aims to identify the tree which can explain the sequence data with a minimum number of sequence changes (Camin and Sokal 1965), whereas maximum likelihood approaches aim to find a tree that maximizes the likelihood of the sequence data under a given model of evolution (Felsenstein 1981). Bayesian methods work in a different manner and attempt to compute the posterior probability for each tree given the aligned sequences, a model of evolution, and a prior distribution over tree space (Rannala and Yang 1996). Again, it is not practical to consider all trees so Bayesian approaches aim to produce a sample of trees that reflect the posterior probability distribution of trees. The set of sampled trees with high posterior probability can be combined to form a single phylogenetic consensus tree for the sequences under consideration. Several models of evolution can be used with both maximum likelihood and Bayesian approaches, and these methods are generally believed to outperform maximum parsimony if the selected model is a reasonable approximation of the processes that produced the sequences (Gaut and Lewis 1995; Swofford et al. 2001; Spencer et al. 2005).

In this thesis, several distance- and sequence-based methods have been used for phylogenetic inference either to explicitly compare results under different methods or due to advances in available inference programs. In general, I have preferred maximum likelihood approaches as they are currently more computationally efficient than Bayesian methods and still allow flexible sequence-based models of evolution to be incorporated. Inference has been performed using either RAxML (Stamatakis 2006) or FastTree (Price et al. 2009). The latter makes use of several heuristics to reduce computational requirements and produces trees which compare favourably with RAxML (Liu et al. 2011). Distance-based unweighted pair group method with arithmetic mean (UPGMA) and neighbour-joining trees have been used when exploring the influence of different inference methods.

Rooting Phylogenetic Trees

With the exception of UPGMA, all inference methods considered in this thesis, and most methods in general, produce unrooted trees as a time-reversible model of evolution

is used (i.e., the position of the root does not influence the model of evolution so other information must be used to infer the placement of the root). UPGMA differs in this regard by assuming all sequences evolve at a constant rate which implies that the similarity of sequences is directly related to the time since their last common ancestor. Unrooted trees are rooted by including sequences from an outgroup, a set of taxa closely related to the taxa under consideration that are known, or at least presumed, to be phylogenetically outside the set of taxa being studied (i.e., a sister to the group of interest). Given a creditable outgroup, a tree can be rooted at the midpoint of the branch separating the set of taxa being studied from the outgroup.

Visualizing Phylogenies

Numerous dedicated programs have been developed for visualizing phylogenies (Pavlopoulos et al. 2010), and best-of-class tree viewers can lay out trees in various ways (e.g., Dendroscope, Huson et al. 2007), provide specific functionality for handling trees with hundreds or thousands of taxa (e.g., TreeJuxtaposer, Munzner et al. 2003), and can colour nodes and branches to reflect additional attributions of interest (e.g., Radié, Whalley et al. 2009; iTOL, Letunic and Bork 2011). FigTree (http://tree.bio.ed.ac.uk) was used extensively during this thesis to inspect phylogenies for artifacts such as abnormally long branches, or outgroups that failed to form a clade separated from the group of interest. Although dozens of tree visualization tools have been developed, it remains an active field of research (Page 2012; see http://treevis.net for examples).

Recently, a number of programs have been developed explicitly for visualizing geophylogenies though these programs are suitable for visualizing any type of geotree (Kidd and Liu 2008; Hill and Guralnick 2010; Bielejec et al. 2011). Geotrees are a valuable exploratory tool in biogeographic studies and a novel quantitative method for visualizing geotrees is discussed in Chapter 2. Both two- and three-dimensional visualizations of geotrees are considered throughout this thesis for exploring biogeographic datasets (Fig. 1.2e and 1.2g).

Phylogenetic Networks

A tree inference algorithm will produce a tree even if there is conflict in the available phylogenetic signal which makes the evolutionary relationships between taxa unclear. At

the extreme, these methods will produce a tree even from a set of randomly generated molecular sequences. However, even for a set of taxa that has evolved in a strictly bifurcating manner, uncertainty in the ancestral relationships is likely to persist due to insufficient phylogenetic signal, violations in the selected model of evolution, or mechanisms such as incomplete lineage sorting which cause discordance in the phylogenies produced by different genes (Huson et al. 2010). Phylogenetic networks can be used to explicitly represent conflicting phylogenetic signal (Huson and Bryant 2006; Morrison 2011). Many methods for inferring phylogenetic networks have been proposed including the median network method which is commonly used to examine the relationships between human populations (Bandelt et al. 1995), and the computationally efficient neighbour-net (Bryant and Moulton 2004) algorithm which is often applied to provide further insights into poorly understood or complex phylogenies (Morrison 2005; Huson and Bryant 2006). Chapter 5 explores the use of phylogenetic networks for assessing the similarity of communities.

1.3.3 Assessing the Similarity of Biological Communities

Biogeographic studies rely on measures of diversity to assess the spatial distribution of species or communities. Here I discuss the different types of diversity that are typically considered and introduce essential terminology. This thesis is primarily concerned with phylogenetic beta diversity (Fig. 1.2f) and specific examples are provided to illustrate the use of these measures.

Types of Diversity: Alpha, Beta, and Gamma

Whittaker (1972) proposed 3 terms for describing biodiversity at varying spatial scales:

- *Alpha diversity*: the diversity within a specific community. The simplest measure of alpha diversity is species richness, the number of species within a community.
- *Beta diversity*: the variation in diversity between communities. For example, the change in species richness between 2 communities.
- *Gamma diversity*: the diversity of all communities within a region. For example, the richness of the regional species pool.

These terms are deliberately vague so that they can be adapted to different biological systems. Within microbial ecology alpha diversity is typically measured at the scale of individual environmental samples (e.g., lake water at a particular location and depth), beta diversity is measured between individual samples, and gamma diversity is measured over all collected samples. In contrast, within classical ecology alpha diversity may be measured over an entire lake with beta diversity being measured between lakes and gamma diversity measured over all lakes in a particular study.

Chapters 4 and 5 of this thesis are focused on the assessment and extension of beta-diversity measures. The term beta diversity has been used in the literature to describe a wide range of varying measures which has resulted in recent efforts to categorize these contrasting approaches (Tuomisto 2010a; Anderson et al. 2011). This thesis focuses solely on measures of beta diversity which assess similarity between pairs of communities. Such measures have been referred to as resemblance measures (Kuczynski et al. 2010), pairwise measures (Tuomisto 2010a; Tuomisto 2010b), and simply as beta-diversity measures (Koleff et al. 2003; Lozupone and Knight 2008). In this thesis, I use the broad term beta diversity to refer to this particular class of measures and concede that it is imprecise. The literature would benefit from adopting a classification scheme for "measures of variation".

A Diversity of Beta-diversity Measures

Traditional measures of beta diversity consider a vector indicating the species present within a community and assess the similarity of a pair of communities by comparing these vectors. For example, both the Manhattan and Euclidean metrics are valid beta-diversity measures though these are rarely used in practice. An impressive number of *taxon-based* measures of beta diversity have been proposed (Legendre and Legendre 1998; Magurran 2004), and there is no overall consensus regarding the most appropriate measures for addressing particular ecological questions (Koleff et al. 2003; Kuczynski et al. 2010). Within microbial ecology, taxon-based measures are applied by partitioning sequences into predefined clusters (e.g., named taxonomic groups or established gene families) or into *de novo* clusters (e.g., *operational taxonomic units* which are used as proxies for microbial species) based on the similarity of sequences (Schloss et al. 2009).

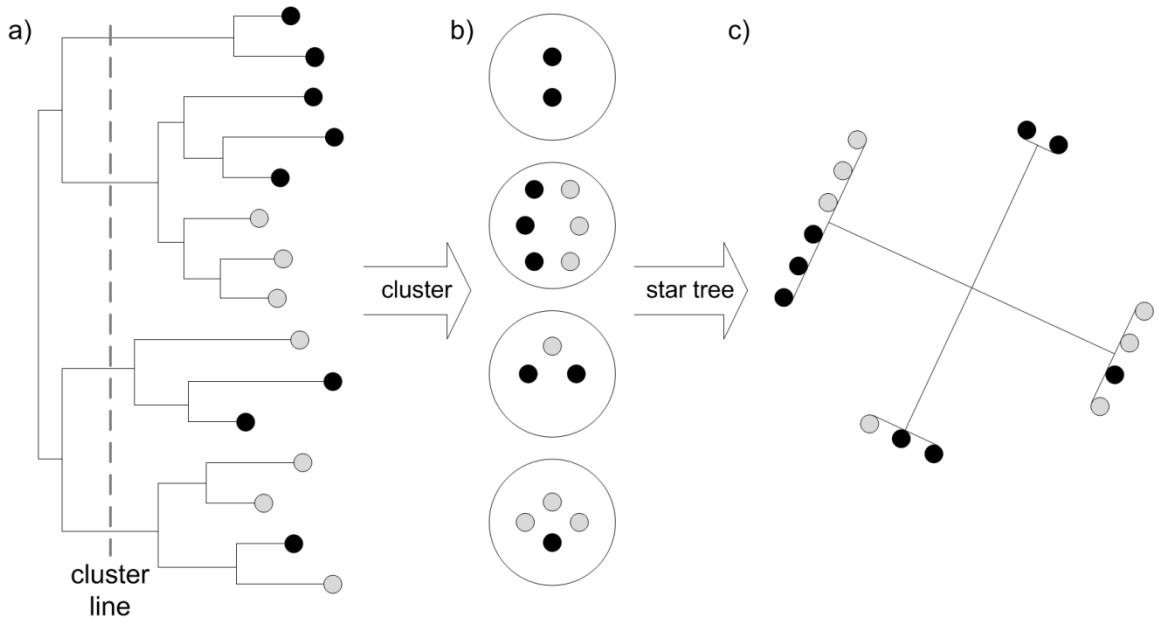


Figure 1.5. Sequences from 2 communities shown as black and grey circles. (**a**) The evolutionary relationship between the sequences is captured through a phylogenetic tree. In this example, clusters are formed by partitioning the tree at a fixed distance from the root as indicated by the dashed grey line. Clusters are more typically defined based on the similarity of sequences. (**b**) Clustering results in the evolutionary relationships between sequences being lost. (**c**) This is equivalent to relating sequences through a star tree where all branches are of equal length.

For example, operational taxonomic units (OTUs) are often obtained for prokaryotes by clustering 16S rRNA gene sequences at different sequence similarity thresholds (Stackebrandt and Goebel 1994; Schloss and Handelsman 2004). This taxon-based approach to measuring community variation has 2 notable limitations: (1) all species or clusters are considered equally distinct from each other, and (2) delineating "natural" clusters within microbial communities has proven to be challenging (Koeppel et al. 2008; Doolittle and Zhaxybayeva 2009). Phylogenetic beta-diversity measures address these limitations by using the similarity of sequences within a phylogeny to quantify the evolutionary divergence of communities (Graham and Fine 2008). Conceptually, taxon-based measures implicitly assume a star phylogeny and fail to account for the phylogenetic distance between taxa (Fig. 1.5). In practice, this can result in these 2 types of measures producing substantially different patterns of relationships between communities (Graham and Fine 2008; Hamady et al. 2010). This thesis is focused on

phylogenetic-based measures of beta diversity, but recent methods have also considered the incorporation offunctional information, i.e., traits that influence how an ecosystem operates, into the assessment of community similarity (Ricotta and Burrascano 2008; Swenson et al. 2011).

As a univariate summary statistic, no single beta-diversity measure can address all manners in which community similarity may be usefully defined. A key distinction is whether a measure is *qualitative* or *quantitative* (Legendre and Legendre 1998; Lozupone et al. 2007). Qualitative measures consider only distinct sequences and are indicative of whether ecological factors prohibit taxa or gene families from occupying certain habitats, whereas quantitative measures consider the relative abundance of each sequence and can be used to infer whether ecological differences between habitats cause the abundance of taxonomic groups or gene families to change. The most commonly used phylogenetic beta-diversity measure in microbial ecology is the unique fraction (UniFrac) measure. Both qualitative (unweighted) and quantitative (weighted) versions of UniFrac have been proposed, and these 2 measures can produce substantially different insights into the similarity of communities (Lozupone and Knight 2005; Lozupone et al. 2007). Unweighted UniFrac assesses the dissimilarity of a pair of communities based on the proportion of branch length that is unique to one community or the other (Fig. 1.6). More specifically, each branch in the phylogeny is classified as being shared by the communities, unique to one of the communities, or external to the 2 communities under consideration, and community dissimilarity is calculated as the total amount of unique branch length divided by the total amount of unique or shared branch length. Weighted UniFrac weights each branch according to the proportion of sequences which are descendent from that branch (Fig. 1.7). Specifically, a branch is weighted by the Manhattan distance between the sequence proportions of each community and these weighted branch lengths are summed over the entire tree to give a measure of dissimilarity.

1.3.4 Visualizing Biotic Dissimilarity Matrices

Beta-diversity measures produce a biotic dissimilarity matrix indicating the

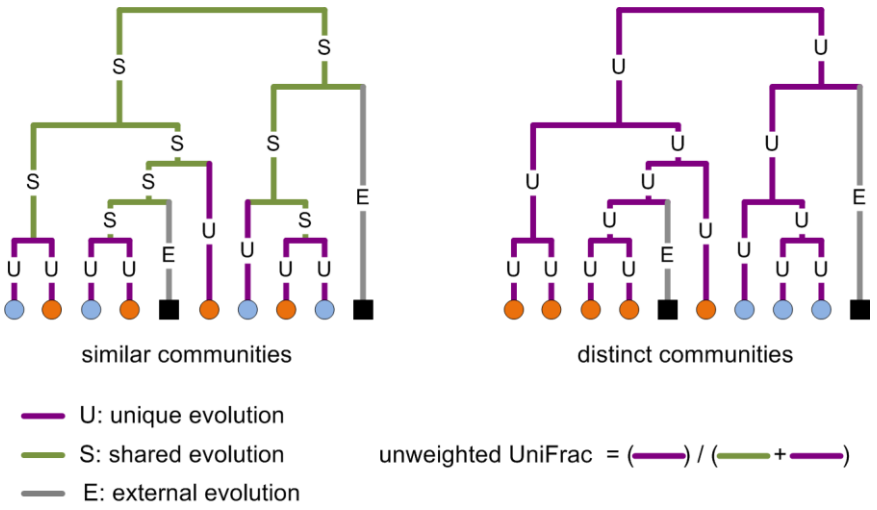


Figure 1.6. Examples of measuring beta diversity with unweighted UniFrac. In this example, sequences have been collected from 3 communities shown as blue circles, orange circles, and black squares. A phylogenetic tree is inferred from these sequences. Unweighted UniFrac is a qualitative measure of beta diversity. The dissimilarity between a pair of communities is defined as the proportion of branch length that is unique to one community or the other. Here we explicitly show which branches are unique, shared, or external to the 2 circular communities. Notice that external branches are effectively ignored in the calculation of similarity. The example on the left illustrates 2 communities that are relatively similar to one another and the example on the right shows 2 communities which are maximally distinct.

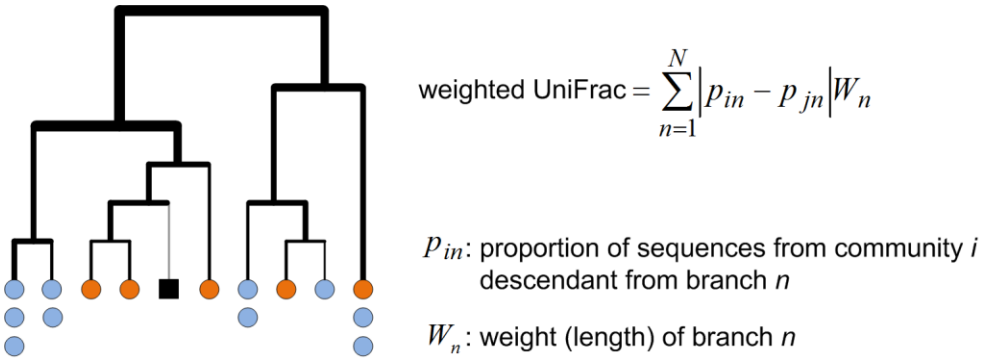


Figure 1.7. An example of measuring beta diversity with weighted UniFrac. In this example, sequences have been collected from 3 communities shown as blue circles, orange circles, and black squares. A phylogenetic tree is inferred from these sequences. Weighted UniFrac is a quantitative measure of beta diversity. The dissimilarity between the 2 circle communities depends on both the distribution of sequences within the phylogeny and the abundance of each sequence type. Formally, the relative proportion of sequences descendant from a branch is determined for each community. The Manhattan distance between these proportions is then calculated and weighted by the length of the branch. This is repeated for all branches in the phylogeny in order to assess the phylogenetic similarity of a pair of communities.

dissimilarity between any pair of communities. Since direct consideration of this matrix is generally ineffective for understanding the relationships between communities, this thesis makes extensive use of 2 multivariate statistical techniques for exploring the relative similarity of communities (Fig. 1.2g): hierarchical cluster trees and principal coordinate analysis. Each of these methods is introduced here and examples are provided in order to illustrate how to interpret the visualizations produced using these methods.

Hierarchical Cluster Trees

Hierarchical cluster trees are used to describe the relationships between a set of entities. Even though hierarchical clustering methods have been used to infer evolutionary trees, they are more commonly used in the life sciences to describe the relationships between entire communities or populations of organisms (Legendre and Legendre 1998). Methods used to build hierarchical clusters are based on 2 general approaches:

- *Agglomerative*: a "bottom up" approach which begins with each entity in its own cluster and progressively merge pairs of clusters until all entities are contained in a single cluster.
- *Divisive*: a "top down" approach which begin with all entities in a single cluster and progressively splits clusters into 2 parts until all clusters contain a single entity.

Both approaches are based on having a measure of dissimilarity between individual entities and a criterion for merging or splitting clusters. In this thesis, agglomerative clustering is used to visualize the relationships described by a biotic dissimilarity matrix. Regardless of the measure of beta diversity used to define the dissimilarity between communities, a linkage criterion is still required to decide when 2 clusters should be merged. There are 3 commonly used agglomerative linkage criteria for determining the dissimilarity between 2 clusters A and B (Legendre and Legendre 1998):

- *Complete linkage* or *furthest-neighbour*: the dissimilarity between A and B is the maximum dissimilarity between an entity in A and an entity in B, $D = \max\{d(a,b); a \in A, b \in B\}$.

- *Single linkage* or *nearest-neighbour*: the dissimilarity between A and B is the minimum dissimilarity between an entity in A and an entity in B, $D = \min\{d(a,b); a \in A, b \in B\}$.
- *UPGMA, average linkage,* or *average-neighbour*: the dissimilarity between A and B is the average dissimilarity between an entity in A and an entity in B, $D = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$.

where $d(a,b)$ is the dissimilarity between entities a and b as defined by the biotic dissimilarity matrix. Clustering proceeds by progressively combining the 2 most similar clusters until all entities are contained in a single cluster. The distance between clusters indicates the length of branches used to connect clusters in a rooted tree where leaf nodes correspond to communities and internal nodes indicate the order in which communities were clustered together (Fig. 1.8). Hierarchical cluster trees are typically depicted as phylograms in order to show the relative similarity of clusters, although cladograms are used on occasion to emphasize the hierarchical relationship between entities.

Each linkage criterion can produce a distinct clustering, and the selection of which criteria to use depends on the questions being addressed. This thesis makes extensive use of UPGMA clustering to explore the relationships between communities. This is the most commonly used linkage criterion in ecology as it is less sensitive to noise than the single or complete linkage criterion, and provides, arguably, a more intuitive depiction of the similarity of entities in 2 clusters (Legendre and Legendre 1998; Schloss and Westcott 2011). The UPGMA linkage criterion is attractive as many researchers are familiar with this clustering method, but recently developed approaches designed specifically for the hierarchical clustering of biological communities are an interesting alternative (Matsen et al. 2010).

Principal Coordinate Analysis

Principal coordinate analysis (PCoA) produces an ordination plot where each community in a biotic dissimilarity matrix is represented by a point and the Euclidean distance between 2 points is an approximation of the dissimilarity between the

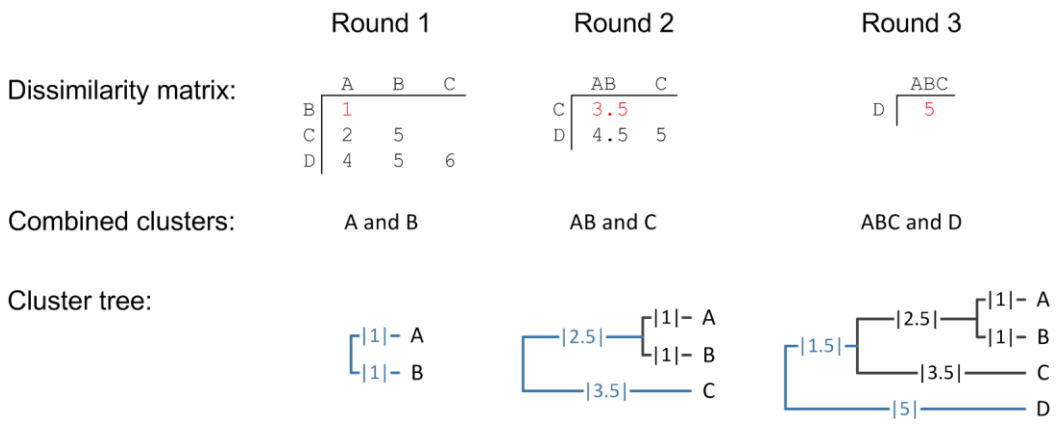


Figure 1.8. An example of constructing a UPGMA hierarchical cluster tree. In round 1, the entities A and B are the most similar and are clustered together. A partial tree is formed which connects these 2 entities by an internal node. For hierarchical cluster trees, the convention is to set the length of branches to an internal node to the distance between the merged clusters. This is in contrast to a UPGMA phylogeny where an internal node is placed halfway between its children so that the sum of branch lengths between 2 taxa will reflect their evolutionary divergence (i.e., the branch lengths in this example would be set to 0.5 instead of 1). In round 2, the distance between cluster AB and all other clusters is computed using the UPGMA linkage criteria. The closest pair under the UPGMA criterion is cluster AB and entity C with a dissimilarity of 3.5. These 2 clusters are merged by adding an additional internal node. In the final round, entity D is combined with the cluster ABC.

corresponding communities (Gower 1966). In general, PCoA plots are drawn in 2 or 3 dimensions to aid in visualizing the relationships between communities and, as a consequence, it is not possible to perfectly depict the dissimilarity between all pairs of communities. A useful conceptual model is to consider a set of points in a two-dimensional space with the goal of displaying these points along a one-dimensional line (Fig. 1.9). To achieve this dimensionality reduction, a line can be drawn in the two-dimensional space and each point projected onto this line. PCoA selects the line (referred to as the first principal coordinate) which captures as much variability in the data as possible. When points reside in higher dimensional spaces (i.e., > 3) direct visualization of the data is not possible, and PCoA is used to project or embed these points in a two- or three-dimensional space where the relationship between points can be visualized.

PCoA comes from a class of linear dimensionality reduction techniques which

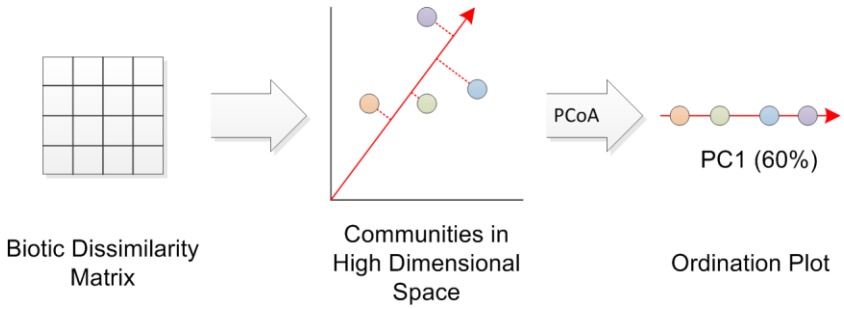


Figure 1.9. An example of constructing an ordination plot with principal coordinate analysis (PCoA). The biotic dissimilarity matrix indicates the pairwise distance between communities. These communities can be conceptualized as points in an *N*-dimensional space where the Euclidean distance between points corresponds to the dissimilarity between communities. In this example, $N = 2$ for visualization purposes. PCoA determines a set of orthogonal lines or principal coordinates (PCs) in a reduced dimensional space which will maximize the variation captured among the set of points. Here the points are being embedded in a single dimensional space. Within this reduced space, the Euclidean distance between points (communities) approximates the dissimilarity between communities as specified by the biotic dissimilarity matrix. The amount of variation captured by each PC can be used as a guide for assessing the quality of this approximation.

includes the more widely known principal component analysis method (Legendre and Legendre 1998). Unlike PCoA, principal component analysis is applied to vectors of data and implicitly calculates the distance between vectors using the Euclidean distance measure. Interestingly, PCoA and principal component analysis are identical when PCoA is applied to a biotic dissimilarity matrix formed under the Euclidean distance measure. PCoA is preferred in ecology precisely because it allows other measures of dissimilarity to be used when forming a biotic dissimilarity matrix. In these approaches, principal coordinates must be linear and orthogonal to one another. More recent methods permit nonlinear axes (Roweis and Saul 2000; Tenenbaum et al. 2000), although their use in ecology has received little attention with the notable exception of Mahecha et al. (2007).

1.3.5 Computational Complexity

In complexity theory, we are interested in the amount of resources (e.g., computation time or memory) needed to solve a problem. Such analyses are performed with regards to an abstract model of computation, most typically a Turing machine, and provide a comparative measure of the amount of resources required by different algorithms (Roos

and Rothe 2010). Using a model of computation is convenient as it abstracts away unnecessary details on how particular computations may be performed and makes results independent of a particular computing device. From a particular viewpoint, knowing the computational complexity of an algorithm indicates how it will scale as the input increases in size. It does not indicate the exact amount of time or memory that will be required as this naturally depends on the computing device.

Here we consider the worst-case complexity of an algorithm, the maximum amount of resource an algorithm will require for any possible input, although other complexity measures, such as average case, are possible (Roos and Rothe 2010). Worse-case complexity is stated using "big O" notation. For example, the Needleman-Wunsch algorithm used to align pairs of molecular sequences has a time complexity of $O(n^2)$, where n is the length of the 2 sequences. In the worst case, we should expect that aligning a pair of sequences of length $2n$ will not take twice as long, but in fact will take $(2n)^2/n^2 = 4$ times as long. In practice, this will only be true for large n as big O notation simplifies the actual complexity of an algorithm by removing constant factors and lower order terms. For example, an algorithm with a time complexity of $5n^2 + n + 4$ would be expressed as $O(n^2)$ since for sufficiently large n the running time is dominated by the term n^2. Additionally, the factor of 4 increase in running time is only for the worst-case so the performance may be better in practice and will certainly depend on the specifics of the computing device. Nonetheless, knowledge of the time complexity provides a good sense of how long an algorithm will take to run as the size of the input dataset increases.

Computational problems are organized into complexity classes, of which the 2 most prominent are deterministic polynomial (P) and nondeterministic polynomial (NP). A problem in the complexity class P can be solved in polynomial time, i.e., $O(n^k)$, where k is fixed for a given algorithm and n is the input size. In contrast, a problem in NP can only verify the solution to a given input in polynomial time. Problems are further classified as either NP-complete or NP-hard. A problem is NP-complete if (1) it is in NP and (2) every other problem in NP can be transformed into this problem in polynomial time. As such, finding a fast (i.e. polynomial time) solution to any NP-complete problem would indicate that all problems in NP can be solved quickly. Many NP-complete problems are currently known and none of them have an efficient solution, which

strongly suggests that fast solutions to these problems do not exist. Solutions to NP-complete problems typically require exponential time, i.e., $O(k^n)$, and may be intractable to compute even for relatively small datasets. A problem is NP-hard if it satisfies condition (2) above whether or not it is in NP. By definition, all NP-complete problems are also NP-hard. However, since NP-hard algorithms are not necessarily in NP, discovery of a fast solution to an NP-complete problem does not imply that all NP-hard problems can also be solved quickly. Informally, NP-hard problems are at least as challenging to solve as NP-complete problems. We will encounter NP-hard problems in Chapter 2 where heuristic and efficient search strategies are used to allow solutions to be found for the majority of datasets of interest.

Readers interested in a more in-depth treatment of complexity classes, and computational complexity in general, are referred to the introductions by Tovey (2002) and Roos and Rothe (2010).

1.4 Structure of Thesis

This thesis is structured around 4 manuscripts that are accepted or published in peer-reviewed international conferences or journals, along with one additional manuscript that is currently in preparation. These manuscripts have been divided into 4 chapters:

- Chapter 2: Introduces a novel visualization technique for assessing the relationship between hierarchically organized data (i.e., a tree structure) and geography. Three case studies are used to illustrate how the proposed technique can be applied to studies of phylogeography. This chapter is an extended version of my *Geoinformatics 2009* (Parks and Beiko 2009) manuscript which includes additional content from the manuscript currently in preparation (Parks et al. 2012).
- Chapter 3: Discusses a geospatial information system, GenGIS, which allows digital map data, environmental data, and georeferenced genetic datasets to be combined into a single visualization and analysis platform. GenGIS provides an implementation of the technique introduced in Chapter 2. Two case studies are used to demonstrate how the visualizations and analyses provided in GenGIS can be used to explore biogeography. This chapter is an extended

version of my *Genome Research* (Parks, Porter, et al. 2009) manuscript which includes additional content from the manuscript currently in preparation (Parks et al. 2012).

- Chapter 4: Provides an extensive analysis of phylogenetic beta-diversity measures which are commonly used to assess variation in communities of organisms along geographic, environmental, and temporal gradients. This chapter derives new phylogenetic beta-diversity measures from existing taxon-based measures, provides an assessment of which measures produce correlated results, and demonstrates the robustness of measures under a variety of conditions. This chapter is based on a manuscript currently under revision at the *International Society of Microbial Ecology Journal* (in press, Parks and Beiko 2012a).
- Chapter 5: Proposes a framework for extending phylogenetic beta-diversity measures to phylogenetic networks known as split systems. This class of networks includes the commonly used median network and neighbour-net inference methods. Calculating community similarity over a split system provides a measure that accounts for uncertainty or conflict in the available phylogenetic signal. Three case studies are used to illustrate the benefits of the proposed framework. This chapter is based on a manuscript currently under revision at *Molecular Biology and Evolution* (in press, Parks and Beiko 2012b).

Each chapter begins with a description of my contributions to the manuscripts used as source material. Chapters are based heavily on the specified manuscripts, though modified where appropriate. This includes updating references, synchronizing vocabulary, expanding analyses, rewriting or removing paragraphs that are redundant in the context of previous chapters, and the removal of 2 case studies in order to help focus the thesis. Supplementary material for each manuscript has been placed in appendices at the end of the thesis or incorporated directly into the main text. Since these changes are extensive for some chapters, particularly Chapter 3, no effort has been made to indicate these modifications. All references are provided in a single comprehensive list at the end

of the thesis. The final chapter of this thesis, Chapter 6, provides a comprehensive discussion of the research presented in this thesis.

Chapter 2

Visualizing Hierarchically Organized Data in a Geographic Context

Parks DH, Beiko RG. 2009. Quantitative visualizations of hierarchically organized data in a geographic context. 17th International Conference on Geoinformatics (Fairfax, VA): 1-6.

Publication status: Published (August 14, 2009).

Contribution to research: **DHP** conceived of and carried out the research. RGB provided guidance throughout the project.

Contribution to writing: Written by **DHP** with suggestions and editorial advice provided by RGB.

Parks DH, Mankowski T, Porter MS, Beiko RG. 2012. GenGIS 2: Geospatial analysis of genetic and genomic datasets, with new gradient algorithms and an extensible framework.

Publication status: In preparation.

Contribution to research: **DHP** conceived of and implemented the linear axes analysis technique. **DHP** with the assistance of TM and MSP developed the plugin framework and other improvements available in GenGIS v2. The kangaroo apple analysis was performed by **DHP**. RGB performed the human microbiome analysis and oversaw the development of GenGIS.

Contribution to writing: Written jointly by **DHP** and RGB.

This chapter is a modified and expanded version of the Parks and Beiko (2009) paper, which incorporates new developments discussed in Parks et al. (in preparation, 2012).

2.1 Abstract

Here we introduce a novel quantitative technique for visualizing hierarchically organized data in a geographic context. In contrast to existing techniques, our visualization emphasizes the hierarchical relationships in the data by depicting them in standard tree formats. Our technique allows users to define a geographic axis and visualize how well a tree correlates with the ordering of geographic locations along this axis. This is accomplished by finding the ordering of leaf nodes that most closely matches the defined ordering of geographic locations. When these 2 orderings are shown in parallel, any mismatches will cause crossings between the lines connecting leaf nodes to their associated geographic locations. These crossings are a visual and quantitative indication of discordance between the topology of the tree and the user defined geographic axis. We have developed a branch-and-bound algorithm that allows the leaf ordering resulting in the fewest crossings to be determined quickly enough to support interactive exploration of different geographic axes even for large multifurcating hierarchies. The quantitative nature of our visualization allows a permutation test to be defined for determining if the relationship between a tree and a geographic axis is statistically significant. In this chapter, the utility of our visualization is demonstrated on biological datasets, but the method is applicable to any hierarchical data where structuring due to geography is of interest.

2.2 Introduction

Visualizing the hierarchical relationships within a georeferenced set of entities allows a user to explore the influence of geography on the patterns of similarity between these entities. Phylogenetic trees depict evolutionary relationships, with leaves typically corresponding to organisms or genetic sequences and internal edges showing common ancestry. There are many published tools to visualize these trees and several recent software packages allow the binding of leaf nodes to the geographic locations from which the corresponding entities were sampled (Kidd and Liu 2008; Maddison and Maddison 2008; Hill and Guralnick 2010; Janies et al. 2010; Bielejec et al. 2011). By testing the relationship between evolution and geography, these geophylogenies can yield valuable insights into speciation processes (Avise 2000; Kidd 2010), the origin and transmission

of viruses such as HIV (Gifford et al. 2007) and Influenza A (Janies et al. 2007; Lemey et al. 2009), and the long-term migration patterns of animals, including humans (Soares et al. 2008). Additionally, by displaying underlying environmental features such as habitat type, soil acidity, or population density, spatial and non-spatial hypotheses can be contrasted or combined.

Existing geophylogenies bind leaf nodes directly to sample sites and assign meaningful locations to internal nodes by inferring their position from evidence such as dated fossils, historical samples, or biogeographic reconstruction algorithms (Kidd and Ritchie 2006). Three-dimensional geophylogenies allow the depth of a node to be visualized as an offset from the geographic plane (Fig. 2.1a). This style of geophylogeny was first proposed by Kidd and Ritchie (2006) and later made available as Geophylobuilder (Kidd and Liu 2008), an extension to ArcGIS (http://www.ersi.com). Recently, Google Earth (http://earth.google.com) has been used to visualize three-dimensional geophylogenies. Although Google Earth lacks the spatial analysis and environmental data integration that is possible within a more complete geographic information system (GIS) framework, its free availability has encouraged its use and prompted the development of software for creating Google Earth-compatible geophylogenies (Hill and Guralnick 2010; Janies et al. 2010; Bielejec et al. 2011). Our software package supports three-dimensional representations of hierarchical data as we have found it to be a powerful visualization technique, especially when internal nodes can be assigned meaningful geographic positions.

In the absence of historical data or a plausible migration model, meaningful positions cannot be assigned to internal nodes and instead are typically placed at the spatial centroid of their children. In this case, the visualization can be misleading since it is difficult not to infer meaning from the position of internal nodes. This problem persists and perhaps is even emphasized when the phylogenetic tree is viewed along the vertical axis to obtain a two-dimensional visualization (Fig. 2.1b). A further weakness of three-dimensional geophylogenies is that the relationships between entities can be obscured since the tree structure relating the entities is distorted to fit the underlying geography. In contrast, our visualization emphasizes the hierarchical relationships in the data by

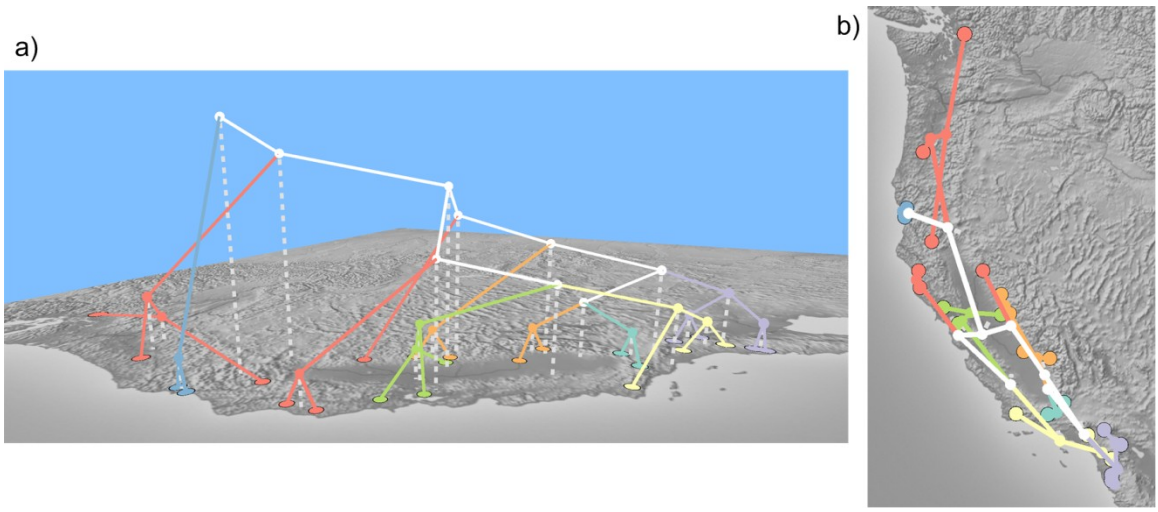


Figure 2.1. Examples of geophylogenies. (**a**) Visualization of a three-dimensional geophylogeny in GenGIS. These visualizations are typical of Geophylobuilder and programs using Google Earth as a backend visualization platform. (**b**) The same dataset viewed along the vertical axis order to produce a two-dimensional geophylogeny. Visualizations such as these are possible within Geophylobuilder and Mesquite Cartographer.

depicting them in standard two-dimensional tree formats where leaf nodes are visually related to their geographic locations through a series of lines which minimize visual clutter.

Our technique can be used as an interactive exploratory tool that allows users to define a geographic axis and visualize how well the topology of a tree correlates with the ordering of geographic locations along this axis. This is accomplished by finding the ordering of leaf nodes, subject to the constraints of the tree topology, which minimizes the number of crossings that occur between lines that connect leaf nodes to their associated geographic locations along the proposed axis. In this optimal layout, the number of crossings that occur between these lines is a visual and quantitative measure of the amount of discordance which exists between the topology of the tree and the user defined geographic axis. To allow interactive exploration of different geographic axes, we have developed a branch-and-bound algorithm that allows optimal leaf orderings to be determined in real time even for large multifurcating hierarchies. Our quantitative visualization is supported by a statistical test which determines whether the fit of tree leaves to a geographic axis is significantly better than random. Additionally, we have developed an algorithm for evaluating the fit between a tree and *all* possible linear

geographic axes which allows the best linear axes to be identified and the robustness of results to be assessed.

The proposed visualization technique is similar in principle to a tanglegram where 2 trees are placed parallel to each other and matching leaf nodes in the 2 trees are connected by lines (Holten et al. 2008; Venkatachalam et al. 2010). Any crossing between these lines indicates discordance between the 2 trees. However, tanglegrams are used to compare the similarity of 2 trees (e.g., a species and gene tree) whereas our visualization relates a georeferenced tree to a specific geographic axis. Minimizing the number of crossings in a tanglegram is a well-known NP-hard problem (Buchin et al., 2008).

Here we demonstrate the utility of our visualization using a series of biological datasets, but our method can be applied to any hierarchical dataset whose geographic structure may be of interest. Our technique has been implemented in GenGIS, a free and open-source GIS package that provides tools for visualizing and analyzing biological datasets (Beiko, Whalley, et al. 2008; Parks, Porter, et al. 2009).

2.3 Visual Design

2.3.1 Visualization Overview

Our visualization consists of a number of elements which allow for the rapid assessment of the goodness of fit between a tree topology and a user-defined geographic axis (Fig. 2.2). Exploration of a geographic axis begins by drawing a tree layout line (TLL) to indicate the desired position and orientation of the tree. Drawing a TLL causes a geographic layout line (GLL) to be generated with the geographic locations associated with the leaf nodes typically placed evenly along this line. The order of locations along the GLL corresponds to their ordering when they are projected onto the GLL. This facilitates the rapid investigation of linear geographic axes. To explore nonlinear geographic axes, an axis can be specified using a set of polylines. The order of geographic locations along the GLL will now reflect their ordering along this geographic axis polyline (GAP). For clarity, the start of the GLL and GAP are identified by a triangle. Location lines are drawn to visually associate geographic locations with their

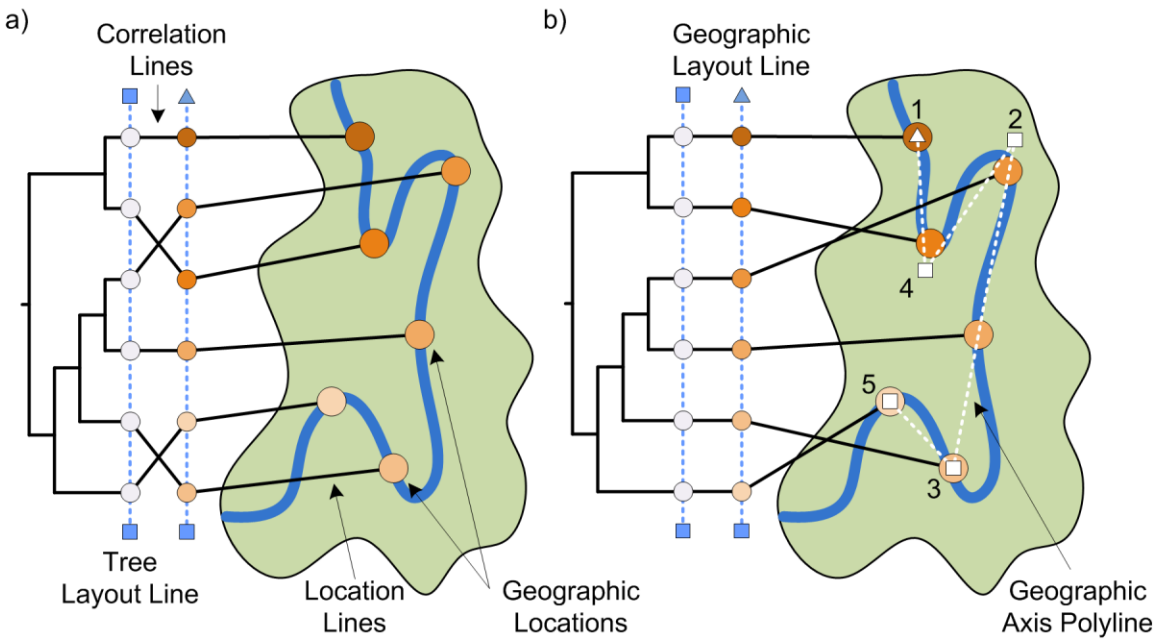


Figure 2.2. Optimal leaf layout along linear and nonlinear geographic axes. (**a**) A linear axis which optimizes the layout of leaf nodes in a tree with respect to a strict vertical geographic ordering of geographic locations (orange circles). Two crossings are induced because the tree cannot be perfectly reconciled with the imposed axis due to the presence of subtrees that are intermingled with respect to this geographic orientation. (**b**) A nonlinear axis in which geographic locations are ordered according to their positions along the river as specified by the geographic axis polyline: in this case, the absence of crossings between the geographic layout line and the tree layout line indicates a perfect reconciliation of the leaves of the tree with the ordering of geographic locations induced by the polyline. Numbers indicate the ordering of user-defined points used to construct the geographic axis polyline. Multiple geographic locations can be covered by a single polyline segment.

corresponding point on the GLL. Similarly, a geographic point is visually associated with its corresponding leaf node by drawing a correlation line. A key aspect of our visualization is that the layout of the tree is optimized to minimize the number of crossings that occur between these correlation lines. An algorithm for determining the optimal tree layout is described in Section 2.4. We also allow the visual properties (e.g., colour, thickness, visibility) of all elements to be customized in order to emphasize different aspects of the data.

2.3.2 Interactive Exploration of Geographic Axes

A number of features of our visualization support the rapid, interactive investigation of different geographic axes. Most importantly, the TLL, GLL, and GAP can be modified by dragging control points. All other elements of the visualization are automatically updated to reflect such a change. Emphasis has been placed on ensuring the visualization is updated at interactive rates (i.e., < 100 ms, Jacko and Sears 2002) in order to allow users to fluidly explore different geographic axes of interest.

Geographic locations can either be spread evenly along the GLL or positioned in proportion to their distance from the start of a geographic axis (Fig. 2.3). We support switching between these 2 visualization modes as they emphasize different aspects of the data. Evenly spreading out points along the GLL makes following and identifying crossings between correlation lines easier, whereas proportional positioning emphasizes the distance between geographic locations along the geographic axis.

To immediately determine the association between a leaf node, a point along the GLL, and a geographic location, users can select any of these elements in order to highlight the others (Fig. 2.4b). Selecting an internal node of the tree highlights all elements associated with the node's subtree. This allows users to determine if a subtree correlates strongly with a given geographic axis. For large trees, visual clutter can be

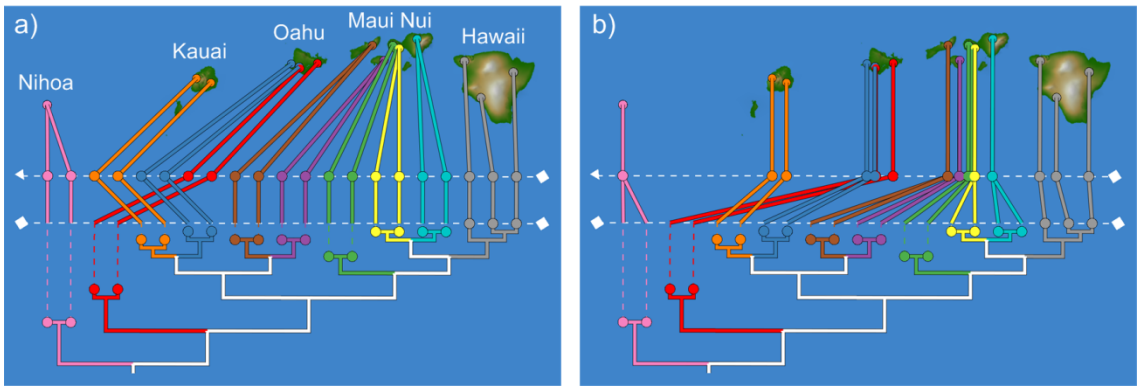


Figure 2.3. Visualizations of Shapiro et al.'s (2006) phylogeny of *Banza* katydids (acoustic insects) from the Hawaiian Islands with major geographic locations assigned unique colour. The low numbers of observed crossings in these images indicate that there is a strong linear geographic structure underlying katydid evolution. (**a**) Geographic locations are placed evenly along the GLL to emphasize crossings between correlation lines. (**b**) Geographic locations are directly projected onto the GLL to emphasize the distance between locations along the linear gradient.

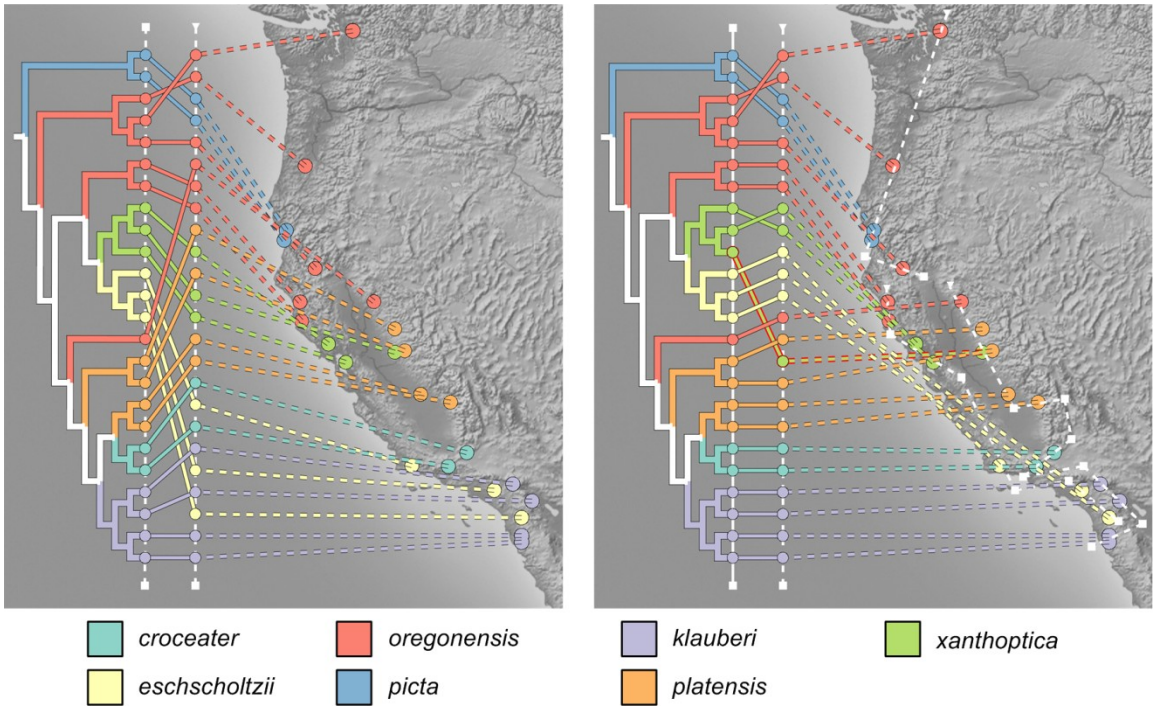


Figure 2.4. Visualizations of Moritz et al.'s (1992) phylogeny of *Ensatina eschscholtzii* salamanders from the western United States with each sub-species assigned a unique colour. (**a**) Visualization testing whether the phylogenetic tree correlates with a linear geographic axis along the coastline. This hypothesis results in 37 crossings. (**b**) Visualization evaluating a nonlinear geographic axis resulting from 2 diverging migration paths shown by the white dashed lines. This hypothesis results in only 11 crossings. The geographic location highlighted in red and marked by an arrow gives rise to 5 of these 11 crossings.

reduced by drawing only those location lines that connect selected elements. This is especially useful for complicated nonlinear geographic axes where location lines will necessarily cross.

Different colours can be assigned to geographic locations to emphasize important aspects of either the phylogenetic tree or the environment. Colours can be propagated from leaf nodes to internal nodes in the tree in either a discrete or continuous fashion. In the discrete case, an internal node is assigned the same colour as its children if all children have the same colour. Otherwise, an internal node is assigned a default colour. For continuous variables, internal nodes are assigned the average colour of their children. This allows the dispersal of discrete and continuous environmental variables (e.g., habitat type, temperature) along a tree to be visualized.

2.4 Optimal Leaf Ordering

Our visualization allows users to visually assess if a tree topology is strongly correlated with an underlying geographic axis by finding the ordering of leaf nodes that minimizes the number of crossings which occur between correlation lines. With the leaf nodes optimally ordered, the number of crossings that remain is a quantitative measure of how well the hierarchical data fits the geographic axis. Here we consider heuristic and approximation approaches to the optimal leaf node ordering (OLNO) problem. These approaches are used in a branch-and-bound algorithm which allows exact solutions to be determined in interactive time for large multifurcating trees. We then discuss extensions to our visualization for considering all possible linear axes or nonlinear axes consisting of multiple segments.

2.4.1 Important Theoretical Results

Dwyer and Schreiber (2004) have shown that determining the OLNO for a tree can be divided into a set of independent sub-problems. Specifically, they demonstrated that the optimal ordering of the children of any internal node can be found independently (Fig. 2.5). For binary trees, this has lead to $O(n \cdot \log n)$ algorithms which exploit this independence property (Dwyer and Schreiber 2004; Venkatachalam et al. 2010). However, minimizing the number of crossings for a general k-ary node is equivalent to finding a solution to a well-studied graph layout problem known as the one-sided crossing minimization (OSCM) problem.

The OSCM problem can be stated as follows: place vertices from one bipartition, V_{Fixed}, of a bipartite graph at prescribed positions along a straight line and find the position of the vertices from the other bipartition, V_{Free}, on a parallel line such that the number of straight line edge crossings will be minimized. The fixed bipartition, V_{Fixed}, is equivalent to the geographic locations along the GLL and the free bipartition, V_{Free}, is equivalent to the leaf nodes on the TLL (Fig. 2.5). This problem is NP-hard for general graphs (Eades and Wormald 1994), and known to be NP-hard even when all vertices of V_{Fixed} have degree 1 and all vertices of V_{Free} have degree 4 (Muñoz et al. 2002).

Finding a solution to the OSCM problem requires an algorithm for determining the

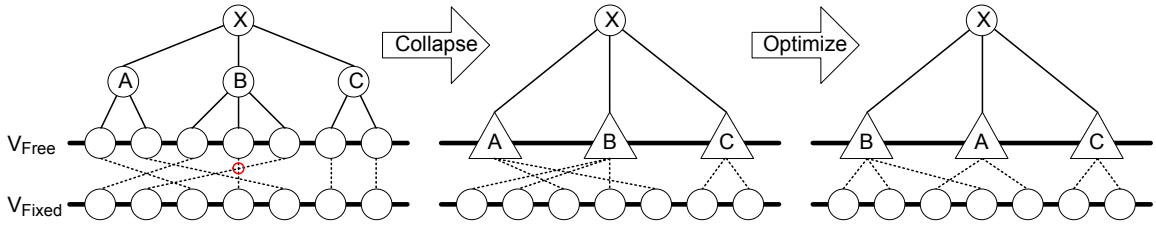


Figure 2.5. The optimal ordering of the children of node *X* can be determined in 2 steps. Initially, the subtree specified by each child is collapsed into a single node. This removes any crossings that occur within a subtree such as the one indentified by the red circle. It is clear that reordering the nodes within a subtree will not reduce the number of crossings between the subtrees. This indicates that each internal node can be optimized independently. Once the children of a node have been collapsed, finding the optimal ordering of the children is equivalent to solving the OSCM problem.

number of crossings that occur for a given ordering of vertices in V_{Free}. The most efficient algorithm known for counting edge crossings was proposed by Barth et al. (2002). This algorithm runs in $O(n \log k)$, where k is the number of children of an internal node N and n is the number of leaf nodes in the subtree rooted at N.

2.4.2 Heuristic and Approximation Algorithms

A number of heuristic and approximation algorithms have been proposed for the OSCM problem. These can be used to determine an upper bound on the minimum number of edge crossings. With a tight upper bound an exact solution to the OSCM problem can be efficiently solved using a branch-and-bound algorithm.

The barycentre and median heuristics have been widely used to allow computationally efficient layouts of large graphs to be obtained (Tollis et al. 1998). These heuristics determine the relative order of nodes in V_{Free} based on the mean (barycentre) or median positions of their neighbours in V_{Fixed}. A classic study by Jünger and Mutzel (1996) demonstrated that the barycentre heuristic often gives results that are extremely close to the actual minimum number of crossings and generally outperforms the median heuristic (along with all other heuristics considered in the study). However, given the low computational cost of these 2 heuristics, it is often reasonable to employ both and use the one resulting in the fewest crossings.

Approximation algorithms determine solutions which are guaranteed to be within a constant factor of the true answer. By convention, approximation algorithms for the OSCM problem are given as a constant factor above a canonical lower bound given by:

$$L = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \min\left(c_{ij}, c_{ji}\right)$$

where c_{ij} is the number of crossings which occur between nodes i and j when node i precedes node j on V_{Free}, and $m=|V_{Free}|$ (Eades and Wormald 1994; Nagamochi 2005). The median heuristic gives a 3-approximation, whereas the barycentre heuristic is less theoretically satisfying and has a $O\left(\sqrt{m}\right)$-approximation (Eades and Wormald 1994) or a (d-1)-approximation, where d is the maximum degree of nodes in V_{Free} (Li and Stallmann 2001). The best known approximation algorithm gives a 1.47-approximation (Nagamochi 2005), though it is rarely used in practice as it lacks the simplicity of the median and barycentre heuristics.

2.4.3 Branch-and-bound Algorithm

The above results suggest an *exhaustive search* algorithm can solve the OLNO problem for a complete k-ary tree in $O(h\ k!\ k^h \log k)$ time, where h is the height of the tree. This runtime can be derived by noting that an internal node at height i has k^i leaf nodes. To find the OSCM for an internal node, all $k!$ possible permutations of the k children must be considered. Using the Barth et al. (2002) algorithm to calculate edge crossings for all permutations requires $O(k!\ k^i \log k)$ work. Since there are k^{h-i} nodes at height i, the amount of work required for *each* layer of the tree is $O(k!\ k^h \log k)$.

Performing an exhaustive search is impractical in an interactive environment even for small trees when $k > 5$ and prohibitive for large trees when $k > 4$ (Fig. 2.6). In order to allow our visualization to be applied to larger trees of higher degree, we have developed a branch-and-bound algorithm (Land and Doig 1960) for solving the OSCM problem, which allows interactive exploration of different geographic axes to be performed on small trees when $k \leq 8$ or large trees when $k \leq 7$. Our branch-and-bound algorithm substantially reduces the amount of computation required to solve the OSCM problem by only considering nodes of a permutation tree that can produce a solution with fewer

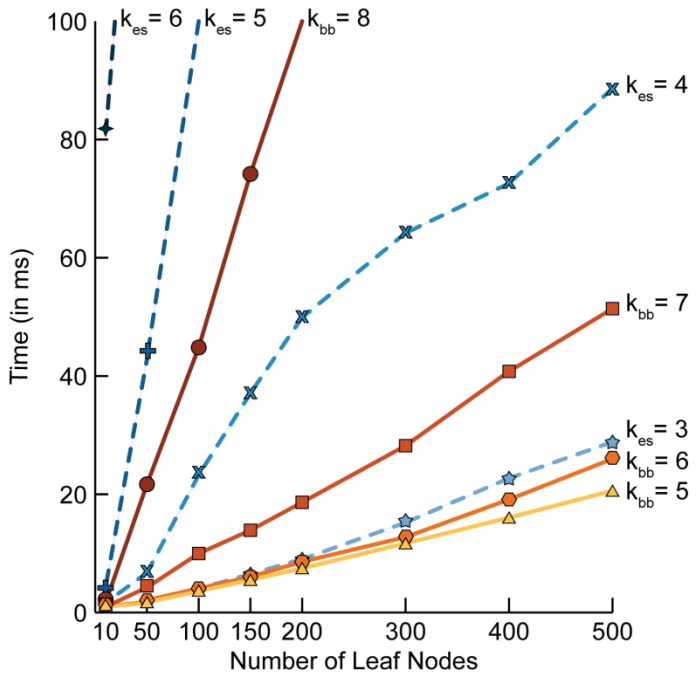


Figure 2.6. Running time to determine the optimal ordering of leaf nodes using a branch-and-bound (solid lines) or exhaustive search (dashed lines) algorithm for various complete *k*-ary trees. Times are averages over 250 independent permutations of the geographic locations. Experiments were performed on a single core of a 2.66 GHz Intel Core 2 Quad Q9450.

crossings than the currently specified upper bound (Fig 2.7, Algorithm 1). As such, it is important to seed the algorithm with a tight upper bound in order to minimize the portion of the permutation tree which must be considered. This can be efficiently done by making use of the heuristic and approximation algorithms discussed above. To solve the OLNO problem, Algorithm 1 must be applied to each internal node in the tree which can be done in parallel if multiple processors are available.

2.4.4 Linear Axes Analysis

The initial implementation of our visualization technique required the user to draw geographic axes by hand, allowing the testing of specific hypotheses but making it difficult to explicitly test all possible axes. Poczai et al. (2011) noted that our technique "does not allow broad testing of encoded hypotheses with automatic polyline enumeration" and manually tested a subset of all possible axes for a set of georeferenced

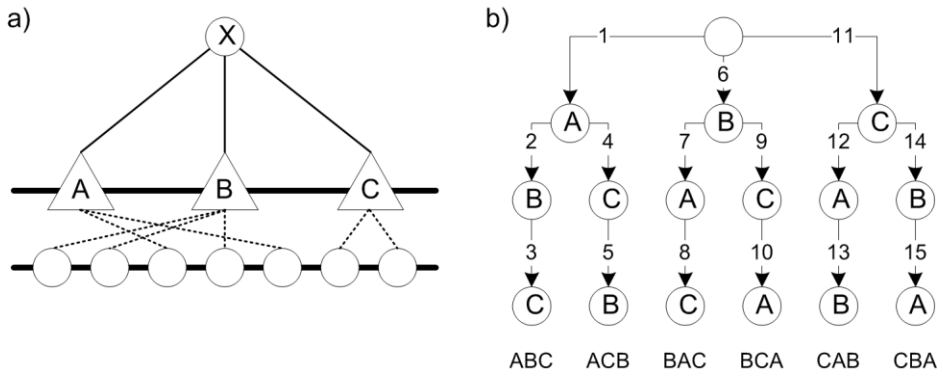


Figure 2.7. An example permutation tree considered by the branch-and-bound algorithm. (**a**) OSCM problem consisting of 3 free nodes. An exhaustive search would require evaluating the number of crossings under all 6 permutations of these free nodes. (**b**) Permutation tree for the example in (a). The permutation tree is explored in a depth first manner as indicated by the numbers on each branch. With a tight upper bound on the minimum number of crossings, large portions of the permutation tree do not need to be explored. For example, the partial permutation AB produces 5 crossings. If the upper bound is 5 or less then it is unnecessary to explore any of the permutations below branch 3 (which, in this example, is the single full permutation ABC).

ALGORITHM 1. OSCM BRANCH-AND-BOUND ALGORITHM

Input: *node* to optimize order of children, *upperBound* on number of crossings, *crossingMatrix* where *crossingMatrix*[*i*,*j*] = c_{ij}
Require: *node.children* is the vector of node's children, *sort*(*vector*,*i*) which sorts elements of *vector* that are ≥ *i* in descending order, *next_permutation*(*vector*) will give the next permutation of *vector* in lexicographically ascending order
Return: ordering of *node.children* which minimizes the number of crossings

```
procedure OptimalOrdering(node, upperBound, crossingMatrix)

{ a vector that indicates a permutation of the children nodes }
permutation = [1, 2, ..., node.children|]

do
  crossings = 0

  { count number of crossings for current permutation }
  for j = 2 to |node.children|
    for i = 1 to j
      crossings = crossings + countMatrix[permutation[i],permutation[j]]
    end for

    { check if the rest of the permutation can be skipped }
    if crossings ≥ upperBound then
      sort(permutation, j+1)
      break from for loop
    end if
  end for

  if crossings < upperBound then
    upperBound = crossings
    optimalOrder = permutation
  end if
while next_permutation(permutation) == TRUE

return optimalOrder
```

kangaroo apple samples. We have addressed this limitation by developing a method, *linear axes analysis*, for efficiently determining the number of crossings that occur for any linear axis. The proposed method is a plane sweeping algorithm where a line (i.e., the linear axis) is rotated by 180˚ (de Berg et al. 1997). A 180˚ sweep is sufficient as the ordering of geographic locations along a gradient at angle θ will be identical to the ordering at θ+180˚ (i.e., gradients are treated as being undirected). The key insight of the proposed algorithm is observing that the ordering of geographic locations along the sweep line changes only when a line between a pair of locations becomes perpendicular to the sweep line. This suggests an $O(n^2 \log n^2)$ algorithm for determining the number of crossings which occur for any linear axis under a constant cost model for solving the OLNO problem (Algorithm 2). For each of the $n(n-1)/2$ pairs of geographic locations, the slope of the line connecting the locations and the sample site information is stored in an array. This array is then sorted in ascending order of slope values. Starting from a horizontal sweep line where geographic locations are ordered by their longitudinal (x-axis) position, the sorted slope array indicates the order in which geographic locations must be swapped as the sweep line is rotated. For each permutation of the geographic locations, the optimal tree layout is determined and the number of crossings for the current orientation of the sweep line is stored. A linear axes analysis can be applied to the entire tree or any subtree.

There are 4 degenerate cases to consider when implementing the linear axes analysis algorithm that are not handled in Algorithm 2 (Fig. 2.8):

- *Multiple samples at the same geographic location*. Independent samples may be taken at the same geographic location. Sites with the same geographic location will project to a single point along a linear gradient. The *minNumberCrossings* method must be able to handle this case. In our implementation, we detect all such sites at the start of the algorithm and remove all except one from the *sampleSites* vector. Duplicate sample sites are then added back into the *sampleSites* vector just prior to calling *minNumberCrossings*.
- *Identical longitudinal coordinates (x-coordinates)*. If multiple sample sites have the same x-coordinate, extra work must be done when setting the initial ordering

ALGORITHM 2. LINEAR AXES ANALYSIS ALGORITHM

Input: *geoLocations*, a vector indicating the *x* and *y* position of each geographic location; *tree*, a tree where each leaf node is associated with a geographic location
Require: *calculateSlope*(*rise*, *run*) which calculates a slope between [180°, 360°), *sort* (*vector*, *field*) which sorts elements of *vector* in ascending order of the specified *field*, *minNumberCrossings*(*geoLocations, tree*) which returns the minimum number of crossings for *tree* and a set of geographic locations ordered according to the vector *geoLocations*, *swap*(*x*, *y*, *vector*) which swaps elements *x* and *y* in *vector*
Return: array indicating the number of crossings for each permutation of the geographic locations
Note on Notation: angles are measured using an azimuth where 90° is due east (i.e., standard compass directions)

```
procedure LinearAxesAnalysis(geoLocations, tree)

{ results of linear axes analysis }
results = [ ]

{ set initial ordering of geographic locations based on their x-coordinate }
sort(geoLocations, x)

{ calculate slope for each pair of geographic locations }
slopeInfoVector = [ ]
for i = 1 to |geoLocations|
  for j = i+1 to |geoLocations|
    slopeInfo.slope = calculateSlope(geoLocations[i].y – geoLocations[j].y, geoLocations[i].x – geoLocations[j].x)
    slopeInfo.geoLocI = geoLocations[i]
    slopeInfo.geoLocJ = geoLocations[j]
    slopeInfoVector[i] = slopeInfo
  end for
end for

{ sort vector in ascending order of slope }
sort(slopeInfoVector, slope)

{ save results for initial ordering }
numCrossings = minNumberCrossings(geoLocations, tree)
results[1].crossings = numCrossings
results[1].slope = 180

{ calculate number of crossings for each permutation of geographic locations }
for i = 1 to |slopeInfoVector|
  swap(slopeInfoVector[i].geoLocI, slopeInfoVector[i].geoLocJ, geoLocations)
  numCrossings = minNumberCrossings(geoLocations, tree)
  results[i+1].crossings = numCrossings
  results[i+1].slope = slopeInfoVector[i].slope
end for

return results
```

of sample sites. Sample sites should be placed in the ordering which occurs when the linear gradient is rotated a small ε amount in the clockwise direction (i.e., positioned based on their y-coordinate value).

- *Multiple pairs of sample sites with identical slopes.* Multiple pairs of sample sites may result in projection lines with the same slope. Handling sets of sample sites which are collinear is described below, but care must be taken even for noncollinear sample sites resulting in the same slope. The *swap* function must be called for all *slopeInfoVector* elements with the same *slope* before calling *minNumberCrossings* and storing the results.

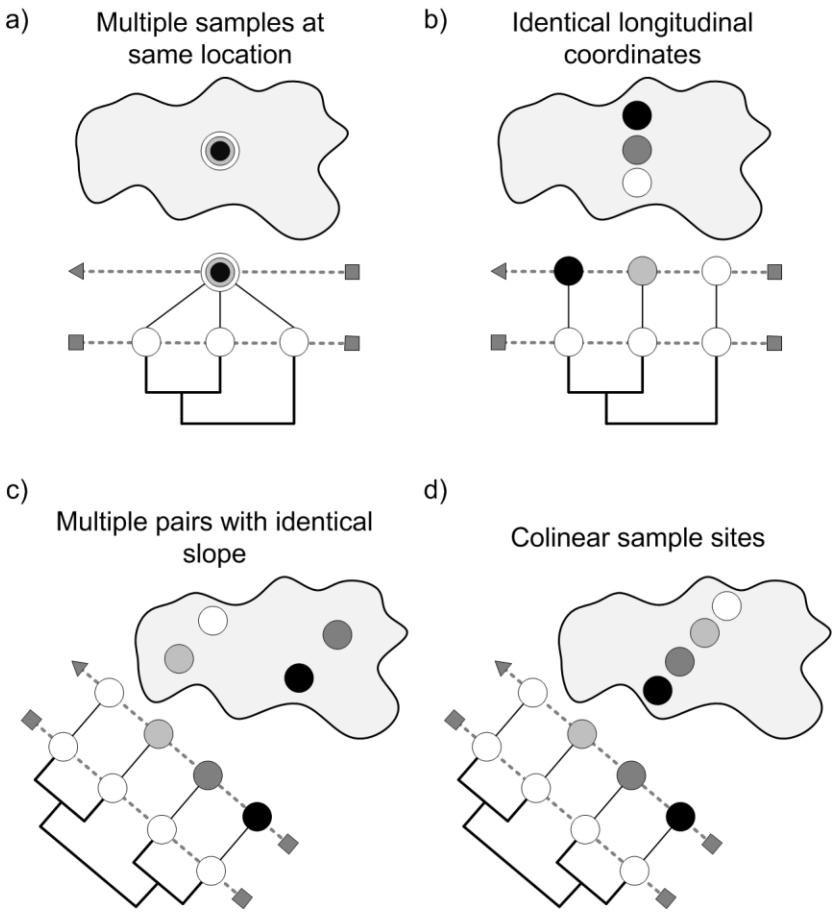


Figure 2.8. Degenerate cases for the Linear Axes Analysis algorithm. (**a**) Multiple samples may be taken from the same geographic locations. (**b**) Sample sites may have the same longitudinal coordinates. (**c**) Multiple pairs of sample sites may have a projection line with the same slope. (**d**) Sample sites may be collinear. (**b-d**) In cases b-d, sample sites are laid out along the GLL in the order they would appear after a small clockwise rotation passed the degenerate angle.

- *Collinear sample sites.* The degenerate case of multiple sample sites projecting to the same position along a gradient is only explicitly handled when sample sites have the same geographic location (see above). For collinear sample sites (or any pair of sample sites), we are only interested in the number of crossings that occur from an ε rotation in either direction. Let θ be the angle of a gradient resulting in 3 or more sample sites being collinear (i.e., along a line with an angle of θ+90°). For an angle of θ – ε, all sample sites will be in the correct order. At θ + ε, the

ordering of any set of collinear points needs to be reversed. This is the general case of having multiple points with *identical longitudinal coordinates*.

2.4.5 Nonlinear Axes with Multiple Polylines

Nonlinear axes can be composed of multiple polylines, where each polyline indicates the ordering for a subset of geographic locations (Fig. 2.9). Polylines are treated as directed in order to allow specific hypotheses to be evaluated. Each polyline is treated as being independent and we consider all $n!$ possible orderings of geographic locations which can be formed from the n polylines. Multiple orderings of the polylines may produce the minimum possible number of crossings. To aid visual clarity, the optimal ordering which minimizes the total length of all location lines is displayed and the results for all permutations are shown in tabular format.

Treating each polyline independently allows complex hypotheses to be evaluated, although in some circumstances it will be desirable to place more stringent constraints on the allowed ordering of polylines. For instance, if the example in Figure 2.9 depicted a migration route known to start in the north, it would be reasonable to constrain the evaluated orderings of geographic locations to those labeled 1 and 2. Although such constraints are often reasonable, recovering an expected migration path without imposing additional constraints provides stronger evidence in support of the hierarchical relationships having been influenced by a particular geographic axis.

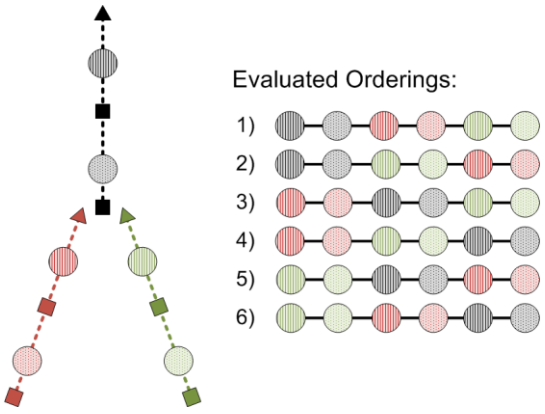


Figure 2.9. Nonlinear axes defined by 3 polylines. Polylines are directed and begin with a triangle. Each polyline induces an ordering for a subset of geographic locations. The 3 polylines specify 6 possible orderings of the geographic locations.

2.5 Monte Carlo Permutation Test

To support our visualization, we have developed a test of whether or not the fit of optimally ordered leaf nodes to geographic locations along a GLL is significantly better than expected by chance alone. A Monte Carlo permutation test can be used to test this null hypothesis by holding the tree topology, geographic axis, and the association between leaf nodes and geographic locations constant while permuting the ordering of locations along the GLL. After each random permutation, the number of crossings for the optimal ordering of leaf nodes is determined. By generating many random permutations, we obtain an estimate of the probability mass function of the null model. The reported *p*-value is the fraction of permutations that have a number of crossings fewer than or equal to the number of crossings in the original model. This test can be applied to the entire tree or any subtree.

2.6 Results

Allopatry is a widely accepted mechanism by which new species arise from populations that have become isolated due to physical barriers. Here we demonstrate how the proposed visualization technique can be used to illustrate allopatric speciation and test different hypotheses about the geographic axis under which a population may have evolved.

2.6.1 Banza katydids: Linear Geographic Axis

The phylogenetic tree of *Banza* katydids from the Hawaiian Islands has recently been recovered by Shapiro et al. (2006) using modern molecular techniques. To emphasize the geographic structure of this phylogeny, samples from each major geographic area within the Hawaiian Islands (e.g., Hawaii, East Maui, Lanai) were assigned a unique colour (Fig. 2.3). Testing a linear geographic axis along the island chain shows that the evolution of *Banza* katydids has been strongly influenced by geography. In fact, only a single subtree (highlighted in red) is not in perfect correlation with this geographic axis, which results in 8 crossings occurring between the correlation lines. Applying the proposed statistical test with 10,000 permutations indicates that the relationship between the leaf nodes and geographic locations along this geographic axis is significant ($p \leq 0.0001$).

The lack of correlation between a single subtree and the geographic structure of the Hawaiian Islands provides valuable evidence as to the biogeographic history of the *Banza* katydids. Shapiro et al. (2006) used a manually constructed geophylogeny in conjunction with information on the geographic history of the Hawaiian Islands and inferred dates of speciation events to suggest plausible scenarios in which katydids dispersed amongst the different islands. They suggest that the common ancestor of all *Banza* probably lived on Oahu which would account for the observed discordance between the *Banza* phylogeny and island geography.

2.6.2 Kangaroo Apples: Linear Axes Analysis

The biogeography of kangaroo apples (*Archaesolanum*) in Australia and Papua New Guinea was examined recently by Poczai et al. (2011). In their analysis, geographic structuring was evaluated by manually testing a subset of all possible linear axes. Here we demonstrate how the linear axes analysis algorithm allows all possible linear axes to be easily evaluated.

Defining a strict west-east axis results in 23 crossings (Fig. 2.10a), whereas a strict north-south axis produces 57 crossings (Fig. 2.10b). Although this suggests stronger longitudinal than latitudinal structuring, by evaluating all linear axes we can determine the globally optimal axis and the range of angles over which statistically significant results are obtained. Applying the Linear Axes Analysis to the kangaroo apple dataset only takes a few seconds on a modern computer despite needing to evaluate all 210 different orientations of the GLL which result in different orderings of the sample sites. This analysis shows that the minimum number of crossings is 23 and that this occurs multiple times between 90° and 103°; the maximum number of crossings (77) occurs around 172° (Fig. 2.10c). Even with a conservative critical value of $\alpha = 0.001$, a wide range of axis orientations (90° to ~150°, and ~220° to 270°) result in significantly fewer crossings than expected under a random model (Fig. 2.10c), strongly supporting spatial structuring centered on a longitudinal gradient. A linear axes analysis can also be applied to specific lineages. On the *Similia* lineage (Fig. 2.10a), no linear axis resulted in fewer crossings than expected under the null model (Fig. 2.10d). For the

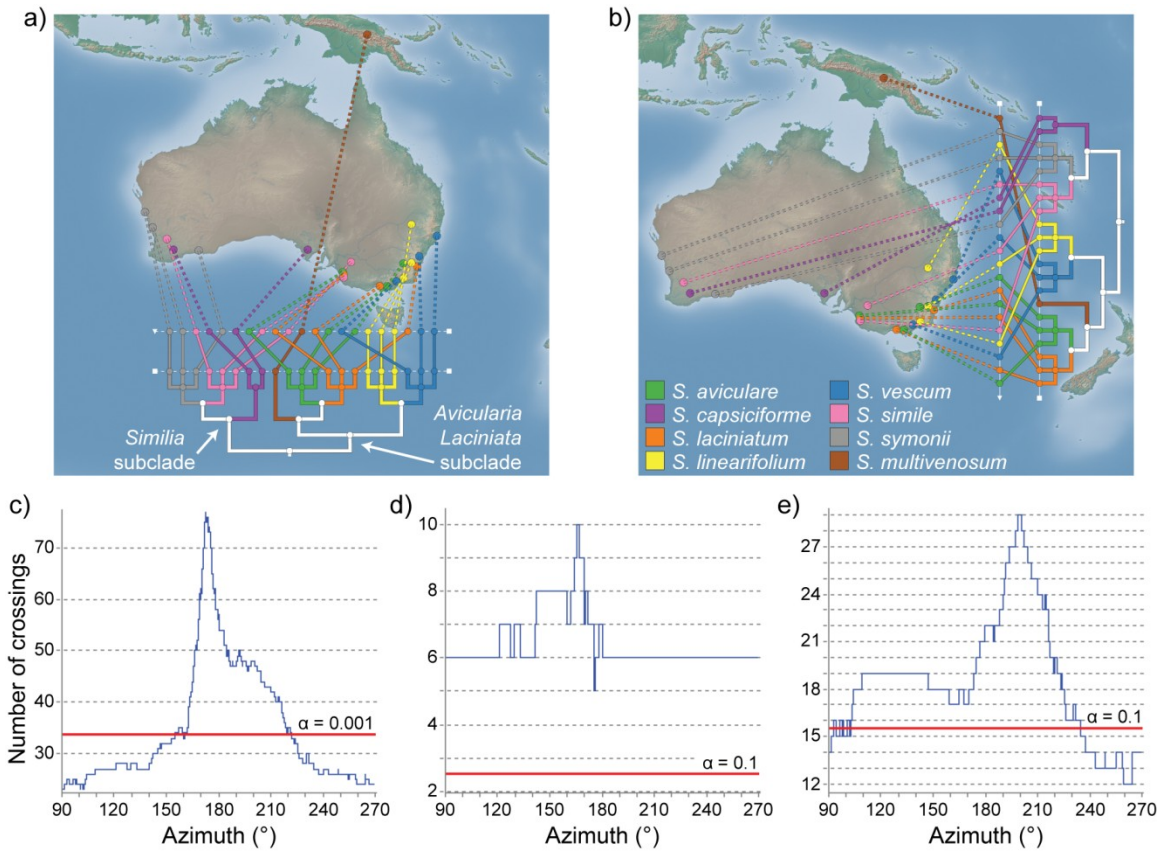


Figure 2.10. Phylogeography of kangaroo apples. (**a**) A longitudinal gradient resulting in 23 crossings. Each of the 8 species within the kangaroo apple phylogeny is assigned a unique colour, and the 2 most substantial subclades are labelled. (**b**) A latitudinal gradient results in 57 crossings. (**c**) Results of a linear axes analysis on the kangaroo apple dataset. The number of crossings is only shown for axes between 90° and 270° as the graph has a period of 180°. Under the null model, only 10 of 10,000 permutations resulted in fewer than 34 crossings as depicted by the red line (i.e. α = 0.001). (**d**) A linear axes analysis of the *Similia* subclade with the red line set to reflect a conservative critical value of α = 0.1. (**e**) A linear axes analysis of the *Avicularia*/*Laciniata* subclades (α = 0.1).

Avicularia/Laciniata lineage, marginally significant ($p < 0.1$) results were obtained for linear axes slightly south of due east and between ~230° to 270° (Fig. 2.10e). The absence of notable longitudinal structuring within either subclade suggests that the strong longitudinal structuring found for the full phylogeny is primarily due to species within the *Similia* subclade being to the west of those within the *Avicularia/Laciniata* subclade.

2.6.3 Ensatina eschscholtzii: Nonlinear Geographic Axis

The salamander *Ensatina eschscholtzii* of the western United States is a classic example of allopatric speciation (Stebbins 1949). Here we demonstrate how our visualization technique can be used to investigate 2 alternative hypotheses about the biogeographic history of these salamanders using the phylogeny inferred by Moritz et al. (1992). We first consider the hypothesis that these salamanders originated in northern Washington and migrated down the western United States during periods of greater humidity. This migration pattern was evaluated using a strictly north-south linear axis (Fig. 2.4a) which resulted in 37 crossings between the correlation lines. The Monte Carlo permutation test provides support for this hypothesis as only 4 of the 10,000 permutations resulted in 37 or fewer crossings ($p \leq 0.0004$). However, by assigning unique colours to each sub-species we can see that a strictly north-south axis results in many of the sub-species being highly intermixed as indicated by the heterogeneous distribution of colours along the GLL.

An alternative hypothesis is that the salamanders moved down the western United States and dispersed down separate coastal and inland ranges in California (Stebbins 1949). We evaluated this hypothesis using a nonlinear axis consisting of 3 segments: 1) migration from northern Washington to northern California, 2) migration along the California coast, and 3) migration along the east side of the California valley. This hypothesis results in only 11 crossings and none of the 10,000 random permutations from the Monte Carlo permutation test resulted in fewer crossings (Fig. 2.4b). The colour of points along the GLL is now far more homogeneous, providing further support for this hypothesis although it should be noted that the proposed method does not control for the additional model complexity associated with a nonlinear axis. Crossings that occur with this geographic axis have established biological and geographic interpretations. For example, the highlighted geographic point in Figure 2.4b is the cause of several crossings. This point is from the Sierran population of the *xanthoptica* sub-species which is found primarily along the coast. It is hypothesized that the Sierran population evolved from the coastal population during a mesic (moderately moist) period of the Pleistocene epoch (Moritz et al. 1992).

2.7 Discussion

Our initial visualization did not include a GLL and instead correlation lines directly connected leaf nodes to geographic locations. This obscured the order of points along a geographic axis and resulted in crossings occurring over a large visual area which made quickly judging the number of crossings difficult. Use of the GLL resolves both of these issues by explicitly showing the order of locations along a geographic axis and restricting the correlation lines to a small visual area. In addition, by spacing points evenly along the GLL, the angle between correlation lines and the GLL is a direct indication of how well geographic locations follow the optimal ordering of leaf nodes.

Unlike existing geotrees, our visualization clearly depicts the hierarchical relationships in the data by presenting these relationships in standard tree formats (i.e., as a phylogram or cladogram), which researchers have experience interpreting. These standard formats for depicting hierarchical relationships take advantage of many fundamental perceptual properties for visually grouping items. Specifically, they take advantage of the Gestalt properties of proximity and connectedness by place related items within close spatial proximity and by directly connect related items by lines (Palmer, 1999; Ware, 2004). However, for datasets where meaningful positions can be assigned to internal nodes three-dimensional geotrees may be more appropriate despite distorting the relationship between items in order to fit the underlying geography. For this reason, we recommend both types of geotree visualizations be supported by software when possible.

Finding the optimal ordering of leaf nodes allows our technique to be quantitatively interpreted. This quantitative property allows our visualization to be used as an exploratory tool for investigating how well alternative geographic axes correlate with a given tree topology. The interactive nature of our visualization encourages users to examine multiple hypotheses which can be evaluated based on the number of crossings they induce and the proposed statistical test. For linear gradients, we have proposed a method which allows investigation of all possible linear axes. A useful addition for linear gradients would be determining the optimal tree layout when crossings are assigned weights based on some external property of interest (e.g., geographic distance, temperature, or alpha diversity). Exploration of nonlinear geographic axes is supported by allowing axes to be specified by a set of polylines. This is flexible, but can be time

consuming when an axis is relatively complex. We plan to extend GenGIS to allow users to specify geographic axes by selecting one or more polylines in a shapefile (e.g., a river or coastline), which will also allow the exact same geographic hypothesis to be applied to different datasets, or by different users.

2.8 Acknowledgements

DHP is supported by the Killam Trusts and the Natural Sciences and Engineering Research Council of Canada; TM and MSP are supported by Genome Canada and the Ontario Genomics Institute; RGB is supported by Genome Atlantic, the Canada Foundation for Innovation, and the Canada Research Chairs program. This project was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute through the Biomonitoring 2.0 project (OGI-050: see http://biomonitoring2.org) and grant number 2009-OGI-ABC-1405.

Chapter 3

GenGIS: A Geospatial Information System for Genomic Data

Parks DH, Porter M, Churcher S, Wang S, Blouin C, Whalley J, Brooks S, Beiko RG. 2009. GenGIS: a geospatial information system for genomic data. Genome Res. 19:1896-1904.

Publication status: Published (July 27, 2009).

Contribution to research: GenGIS is the result of contributions by many individuals. RGB conceived of and initiated the GenGIS project. An initial Python/C++ implementation was developed by SW and MP. MP also integrated Python and RPy into the GenGIS environment. SC and RGB developed initial case studies that guided the development of GenGIS. SW designed and implemented the three-dimensional terrain engine under the guidance of SB. CB and JW provided helpful comments during the initial development of GenGIS. **DHP** ported initial development in Python to C++, brought GenGIS to a deliverable state, and expanded the functionality and usability of the software (including the design and implementation of all geotree visualizations). **DHP** analyzed the Global Ocean Sampling Expedition dataset and RGB analyzed the HIV-1 datasets.

Contribution to writing: **DHP** created the figures and wrote the Global Ocean Sampling Expedition section. RGB wrote the Discussion and HIV-1 sections. The remaining sections of the manuscripts were co-written by **DHP** and RGB. Suggestions and editorial advice was provided by the co-authors.

This chapter is a modified and expanded version of the Parks et al. (2009) paper, which incorporates new developments discussed in Parks et al. (in preparation, 2012) (see previous chapter).

3.1 Abstract

The increasing availability of genetic sequence data associated with explicit geographic and ecological information is offering new opportunities to study the processes that shape biodiversity. The generation and testing of hypotheses using these datasets requires effective tools for mathematical and visual analysis that can integrate digital maps, ecological data, and large genetic or genomic datasets. Existing software for visualizing biogeographic data has focused on one of 3 areas: displaying site-specific community data (e.g., pie charts), generating three-dimensional geophylogenies, *or* visualizing the spatial distribution of densely sampled species data (e.g. geographic heat maps). With GenGIS we introduce a free and open-source software package which provides a rich set of visualizations and analyses for exploring site-specific community data in addition to highly interactive and flexible displays of two- or three-dimensional geotrees. GenGIS also includes a plugin framework that supports the development of graphically driven custom visualizations and analyses which can make use of our custom programming interface as well as established bioinformatics and statistical libraries such as R. Initial plugins include implementations of linear regression and the Mantel test, calculation of alpha- and beta-diversity, and geographic visualizations of biotic dissimilarity matrices. Here we outline the features of GenGIS and demonstrate its application to georeferenced microbial metagenomic and HIV-1 datasets. The most recent version of GenGIS, including sample data files, an online manual, and video tutorials is available at http://kiwi.cs.dal.ca/GenGIS.

3.2 Introduction

Geography and habitat place constraints on the distributions of organisms. Although some of these barriers can be overcome by migration, the discipline of biogeography aims to quantify the long-term impacts of spatial separation on organismal adaptation and evolution. Different habitats offer a wide diversity of energy and nutrient sources but also present a range of biotic and abiotic challenges that must be overcome if an organism is to survive. Microbes pose significant challenges to ecological analysis due to their small size, immense population numbers, and relative lack of distinguishing physical characteristics. Microbial genomes are also highly diverse: a set of lineages that satisfy

the 97% 16S rRNA gene identity criterion for a bacterial species may in fact contain subsets of organisms with very different genetic complements and ecological roles (Gevers et al. 2005; Bapteste and Boucher 2008). Multicellular organisms present some of these challenges as well, particularly cryptic species that are morphologically similar but genetically distinct and reproductively isolated (Rissler and Apodaca 2007). Molecular techniques such as marker gene analysis, multilocus sequence typing, and environmental shotgun sequencing are now being used to explore competing hypotheses about the geographic distribution of organisms (Dick et al. 2004; Martiny et al. 2006; Margos et al. 2008).

Although the type of hypothesis under consideration differs between experiments and among data types, certain goals are common to many studies. One such goal is to assess the taxonomic diversity at one or more sites. The classical ecological measures of Shannon diversity and evenness have been applied to metagenomic data (Fierer and Jackson 2006; Dinsdale et al. 2008), but other measures have been developed to consider the similarity relationships between pairs of communities (e.g., Bray and Curtis 1957) and to account for the common phylogenetic structure between samples (Martin 2002; Lozupone and Knight 2005; Schloss and Handelsman 2006). Although it is clear that these measures capture different aspects of community diversity, recent comparative analyses demonstrate that a great deal remains to be learned about the nature, stability, and robustness of different measures (Schloss 2008; Shaw et al. 2008). Once computed, community diversity can be examined in light of variations in biotic and abiotic factors in the environment; such analyses have been used to demonstrate the effects of factors such as soil pH (Fierer and Jackson 2006), latitude (Fuhrman et al. 2008), elevation (Bryant et al. 2008), and season (Böer et al. 2009) on community composition. Genetic variation within a single named species or ecotype can also be examined using metagenomics (Simmons et al. 2008) or multilocus sequence typing (Konstantinidis et al. 2006).

The range of encoded biological functions can also depend on habitat location and type. DeLong et al. (2006) demonstrated a gradient of taxonomic composition and metabolic capabilities in a 3000 m range of ocean depths, while Green Tringe et al. (2005) used environmental genome tags to show the difference in functions encoded by communities of microorganisms in soil, marine, acid mine drainage, and whale fall

habitats. These approaches were recently extended to show significant functional distinctions in the microbial and viral communities sampled from 9 different habitat types (Dinsdale et al. 2008).

Given a set of homologous characters (e.g., molecular sequences) collected from distinct sites, one may also wish to relate the evolutionary history of these sequences to the relative proximity of sample sites (Avise et al. 1987). Examples of such geophylogenies include the salamander "ring species" *Ensatina eschscholtzii* (Moritz et al. 1992; Chapter 2), human phylogenies based on mitochondrial DNA (Ingman et al. 2000), and trees that track the spread of viruses such as human immunodeficiency virus-1 (HIV-1) through a host population (Hué et al. 2005). Such analyses, when coupled with phylogeographic tools such as Geophylobuilder (Kidd and Ritchie 2006), Mesquite Cartographer (Maddison and Maddison 2008), Supramap (Janies et al. 2010), and SPREAD (Bielejec et al. 2011), can demonstrate the relative rates of migration in different locations or at different times, suggest the locations of ancestral populations or refugia, and highlight evolutionary transitions that affect transmission dynamics. Other biogeographic tools have focused on displaying the spatial distribution and diversity of taxa using geographic heat maps, and incorporate phylogenetic information by focusing on specific lineages, named taxonomic groups, or by assigning molecular sequences to haplotypes (Hijmans et al. 2001; Laffan et al. 2010; Jetz et al. 2012). Georeferenced databases for marine microbial data (Pushker et al. 2005; Lombardot et al. 2006) and viral sequences (MacDonald et al. 2009) have also been compiled and allow rapid investigation of the geographic distribution of specific strains, sequence types, or lineages often in the context of relevant environmental data.

GenGIS (Beiko, Whalley, et al. 2008) is an open-source geospatial information system that is dedicated to the display and analysis of georeferenced genetic data. With GenGIS we provide a series of two-dimensional tree visualizations and analysis tools to complement existing three-dimensional approaches, provide a range of options for source data, and include a powerful analytical interface with the R (http://www.r-project.org) and Python (www.python.org) programming languages at its core. Thus, in addition to the range of visualization and data options implemented directly in GenGIS, users can extend the functionality of GenGIS by developing custom plugins and by installing add-

on libraries for R or Python that implement population genetic or phylogenetic analyses. Included with GenGIS are plugins implementing widely used statistical approaches such as the Mantel test, several measures of alpha and beta diversity, and geographic visualizations of dissimilarity matrices. Existing GIS software for biological data are designed for the relatively dense observational data typical of plant and animal datasets or focus exclusively on the display of geophylogenies using proprietary software. In contrast, GenGIS is an interactive visualization and analysis environment for examining datasets where spatial sampling is relatively sparse, but a wealth of sequence data is available for each sample site. GenGIS provides a rich set of visualization and analysis options for examining site-specific data in addition to highly interactive and flexible displays of geotrees indicating either the similarity of sample sites or phylogenetic relationships. Since its release, GenGIS has been used to investigate the phylogeography of viruses (Parks, MacDonald, et al. 2009; Tucker et al. 2011), bacteria (Farikou et al. 2011), plants (Allal et al. 2011; Poczai et al. 2011), animals (Ruzzante et al. 2011; Shafer et al. 2011), insects (Schoville and Roderick 2010), humans (Loo et al. 2011), and language families (Walker et al. 2012).

Here we illustrate the visualizations and analyses provided in GenGIS using 2 case studies: marine microbial communities sampled during the Global Ocean Sampling Expedition (Rusch et al. 2007) and *pol* genes from non-recombinant subtypes of HIV in Africa.

3.3 Methods

In this section we describe the key features of GenGIS, including required input data types and functionality, along with details of the datasets used in the case studies.

3.3.1 Functionality and Implementation

GenGIS provides graphical summaries of data on a site-by-site basis. Location identifiers can be uniform, or can be assigned distinct colours, shapes, or sizes based on any of their defined attributes including latitude, longitude, or habitat parameters such as temperature and salinity. Information about each site can also be displayed on the screen as text, either associated with the location identifier or in a metadata window. Summaries of the sequence properties (e.g., taxonomic distributions) at each site can be displayed

using two- or three-dimensional pie or bar charts, which can be assigned a size that is either constant or proportional to the corresponding sample size. The colour scheme and positioning of pie charts can be modified by the user, with a range of predefined colour palettes and linear or elliptical layout patterns available. Custom graphical visualizations of sample site data can be generated by exploiting the Python/RPy interface described below.

In addition to site-by-site summaries, GenGIS supports visualizing georeferenced trees in 2 and 3 dimensions that indicate the ecological or phylogenetic similarity among samples collected from different sites. A key principle in the construction of two-dimensional trees is the use of a geographic axis to define hypotheses that follow geographic gradients: for instance, mapping the leaf nodes of a tree to a linear geographic axis leads to a visualization of a one-dimensional gradient of similarity. The extent to which the data fit a given geographic axis can be expressed by the goodness of fit between the ordering of leaf nodes in the tree and the ordering of sample sites along the specified axis. Mismatches between these 2 gradients will lead to crossings between the lines that link these gradients. Fewer crossings imply a better fit between geography and phylogeny, so the best fit of a given tree to a geographic axis is found using a crossing minimization algorithm (Chapter 2; Parks and Beiko 2009). The idea of a linear geographic axis can be generalized to a multi-segment line of arbitrary complexity, allowing the specification of piecewise, nonlinear geographic hypotheses. Coupled with the axis layout functions is a statistical test that determines whether the fit of tree leaves to geography is significantly better than random (Parks and Beiko 2009). Branches of a tree can also be coloured to reflect discrete or continuous environmental variables of the sample sites.

The core GenGIS software is implemented using C++ and OpenGL, which supports the rendering of cartographic data in 3 dimensions. As a free and open-source application, GenGIS makes extensive use of other open-source software libraries, including GDAL (http://www.gdal.org) and Python (http://www.python.org). The Python console in GenGIS allows users to interact directly with data through the GenGIS application programming interface, and allows analyses to be performed using the SciPy (http://www.scipy.org/) and NumPy (http://numpy.scipy.org/) libraries. Users can also

execute commands in the R statistical programming language (http://www.r-project.org) via the RPy2 libraries (http://rpy.sourceforge.net). The functionality of GenGIS can be extended using a plugin framework which allows users to perform custom analyses, and produce novel visualizations in both the plugin window itself and the map environment.

3.3.2 Data Acquisition and Formats

GenGIS uses the freely available geospatial data abstraction library (GDAL) to support a wide range of digital map formats, including both digital elevation maps for visualizing three-dimensional terrain and georeferenced image files for displaying standard map or satellite imagery. There are several large public repositories of digital map data, including the Shuttle Radar Topography Mission (Farr et al. 2007) and GTOPO30 datasets hosted by the U.S. Geological Survey. GDAL can be used as a pre-processing utility to directly manipulate maps from these sources, allowing a user to construct detailed maps of specific geographic area. We have also developed a software tool, MapMaker, which allows custom maps to be generated based on the public domain map sets provided by Natural Earth (http://www.naturalearthdata.com). Maps in GenGIS can be displayed using a number of different projections and source datums. Arbitrary image files, such as a silhouette of the human body, can also be displayed in GenGIS in order to explore non-geographic spatial relationships (e.g., Loo et al. 2011; in preparation, Parks et al. 2012).

The geographic location of sample sites is specified using a comma-separated file. Each sample site must have a unique identifier and an associated set of geographic coordinates, represented using either decimal degrees of latitude and longitude, or Universal Transverse Mercator northing and easting values. Location coordinates need not be unique, since a given site may have multiple samples associated with it (for instance, a series of samples collected at different times, or samples collected by different individuals). Beyond these requirements, any set of attributes such as additional location identifiers, habitat parameters, or time information, may be specified.

Additional input files can supply information about the sequence data collected from each site or the trees that describe the relationships between sites. The format of the comma-separated sequence file is similar to that of the location file: each entity must

have a unique identifier and be associated with one of the entities from the location file, and can then have any number of defined fields potentially including the primary sequence data or inferred attributes such as taxonomy or functional properties of the sequences. Tree files are input to GenGIS in the widely used Newick format and automatically georeferenced if leaf node names correspond to the unique identifiers used to specify either the sample sites or sequences.

3.3.3 Availability

GenGIS is freely available under the GNU General Public License v3.0. Source code and executable binaries for Windows and OS X can be obtained at http://kiwi.cs.dal.ca/GenGIS. The website contains an online manual, several written and video tutorials, and links to useful sources for digital map data. MapMaker is also freely available from the GenGIS website.

3.3.4 Global Ocean Sampling Expedition Metagenome Analysis

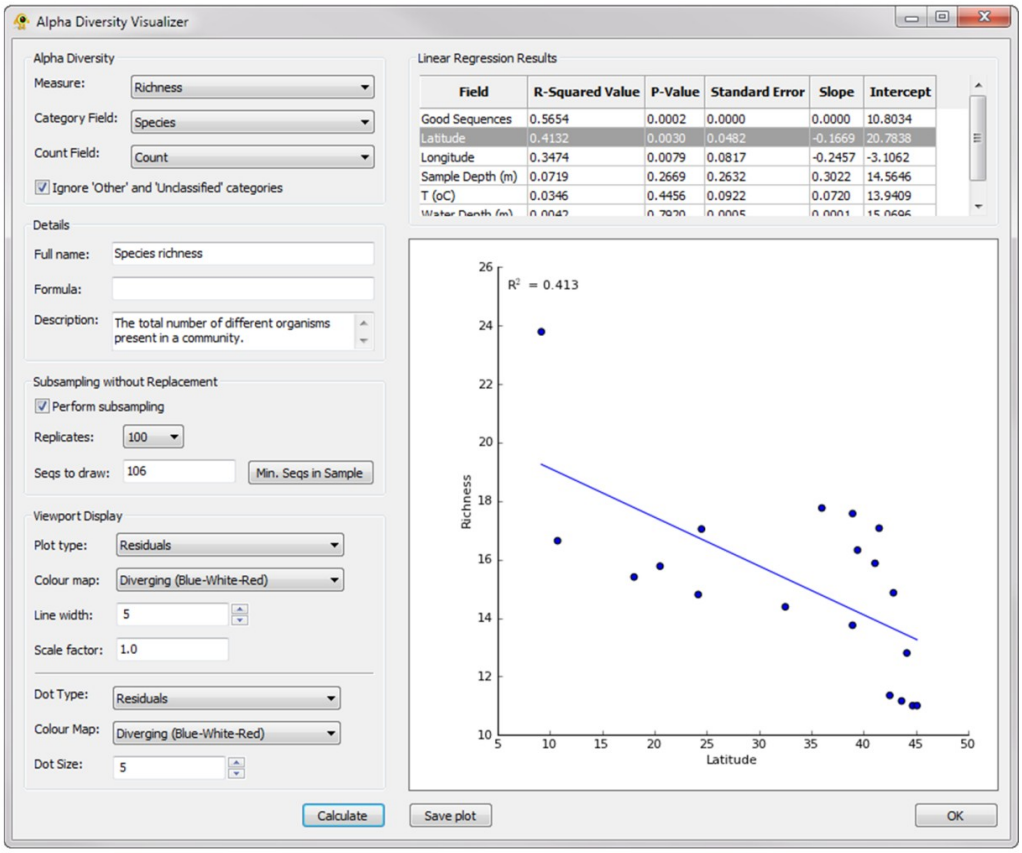
Sample site metadata including temperature and habitat type were obtained from the CAMERA website (Seshadri et al. 2007) and included in the location file that was loaded into GenGIS. An all-versus-all BLAST search between non-coding Global Ocean Sampling Expedition (GOS) sequences obtained from the CAMERA website and all 16S rRNA gene sequences with a length $\geq$ 1250 nt within the GreenGenes (DeSantis, Hugenholtz, Larsen, et al. 2006) database compiled on January 28, 2009 was performed using *blastall* with default parameters in order to identify 16S rRNA gene sequences in the GOS dataset (Camacho et al. 2009). The *blastall* results were filtered to remove any hits with E-values greater than 1.0×10^{-5}, alignments less than 50 nt in length, or a percent identity less than 70%. For these significant matches, the 16S rRNA gene sequences from the GreenGenes database were used as proxies for the corresponding short reads within the GOS dataset. Identical 16S rRNA gene sequences were removed before performing a multiple sequence alignment, but total count information was obtained in order to carry out downstream analyses (Table A.1 in Appendix A). Nearest alignment space termination (NAST: DeSantis, Hugenholtz, Keller, et al. 2006) was used to align the 16S rRNA gene sequences. Hyper-variable regions from the alignment were masked out using the mask columns tool at the GreenGenes portal. A maximum-

likelihood phylogenetic tree covering these aligned sequences was inferred using RAxML v7.0.3 (Stamatakis 2006). RAxML was configured to perform 1000 independent runs using rapid bootstrap analysis with a general time reversible model. All other parameters were left at their default values.

Latitudinal gradients of richness were tested by regressing taxon counts versus latitude. Taxon counts were established using both the taxonomic attributions assigned to sequences during the *blastall* procedure discussed above, and by performing a *de novo* clustering of sequences into OTUs. Taxonomic richness was established at different taxonomic ranks (species to phylum) using the *Alpha-Diversity Visualizer* plugin within GenGIS (Fig. 3.1). OTUs were built at percent identity thresholds of 97%, 95% and 90% using the furthest-neighbor clustering approach implemented in mothur v1.4.1 (Schloss et al. 2009). To correct for different numbers of sequences at different sites, we performed jackknifing subsampling (i.e., sampling without replacement) in order to reduce the number of sequences from each site to 106, the minimum number observed at any site. A total of 1000 jackknife replicates were performed for each site, and the average taxon count across these replicates was used as the response variable in regression analysis. Linear regression analysis was performed with the *Linear Regression* plugin in GenGIS which makes use of the SciPy libraries for regressing data. Independent analyses were performed on the full set of 19 sites and a reduced set of 14 sites which focused on the coastal and open ocean sites by excluding GS005, GS006, GS011, GS012, and GS020.

The relative similarity of GOS communities was examined using the unweighted UniFrac, normalized weighted UniFrac, and Bray-Curtis measures of beta diversity. Unweighted UniFrac measures the proportion of phylogenetic diversity that is unique to a pair of samples (Fig. 1.6; Lozupone and Knight 2005). Normalized weighted UniFrac computes a similar statistic except each branch is weighted by the proportion of taxa below a branch (Fig. 1.7; Lozupone et al. 2007). The Bray-Curtis index measures the compositional dissimilarity between sites using taxonomic profiles specifying the count of different taxa (Bray and Curtis 1957). Taxonomic profiles at the rank of species and genus were considered and the Bray-Curtis index calculated using the *Beta-Diversity* plugin within GenGIS (Fig. A.1 in Appendix A). Sequences which were unclassified

a)



b)

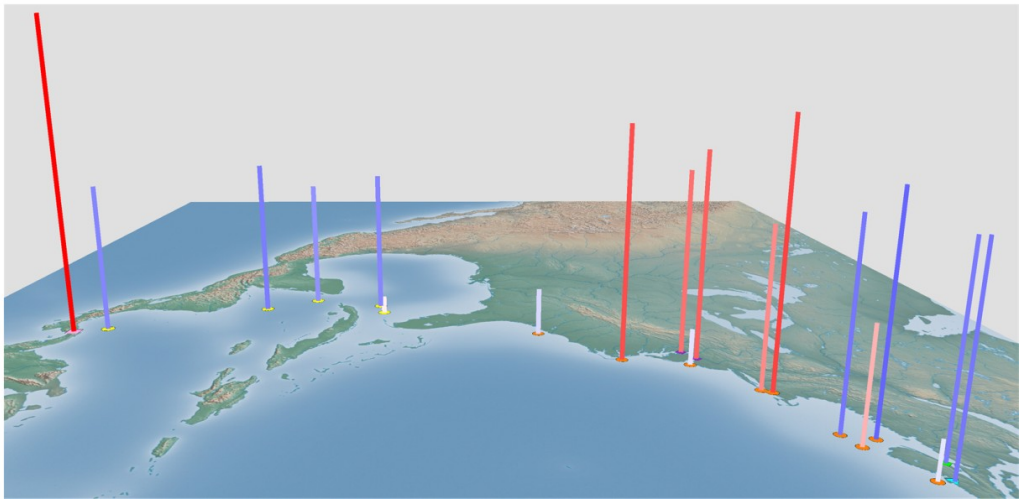


Figure 3.1. Investigating a latitudinal gradient of species richness. (**a**) The *Alpha-Diversity Visualizer* plugin in GenGIS used to calculate alpha-diversity indices, subsample communities, regress indices against environmental and geographic variables, and visualize results in a variety of manners. (**b**) Geographic visualization produced by the *Alpha-Diversity Visualizer* plugin showing geographically situated residuals from a linear regression of normalized (jackknifed) species richness versus latitudinal position (negative residuals = blue, near zero residuals = white, positive residuals = red). The map is oriented with north on the right-hand side.

at a given taxonomic rank were ignored when creating a taxonomic profile at that rank. The percentages of unclassified sequences were 51.6% and 29.6% at the rank of species and genus, respectively. To normalize for sampling effort, a jackknife analysis was used to construct taxonomic profiles consisting of 106 sequences per sample and the average Bray-Curtis index taken over 100 replicates.

Abiotic dissimilarity matrices were visualized by using the UPGMA algorithm to generate a hierarchical clustering of sample sites (Legendre and Legendre 1998). The jackknife analysis, as discussed by Lozupone and Knight (2005), with 1000 random permutations was used to assess how sample size and evenness affected the UPGMA clustering.

3.3.5 Non-recombinant HIV subtypes in Africa

Sequences in this analysis were collected from the HIV Sequence Database (http://www.hiv.lanl.gov/). The total number of instances of each strain collected from each country was acquired from the database; countries with fewer than 10 instances were excluded from the analysis. This yielded a total of 30,002 instances in 40 countries (Table B.1 in Appendix B). Each instance was included in the comma-separated GenGIS file as a separate line, with country association and subtype indicated in separate columns.

To build a reference tree for UniFrac, all sequences containing a full-length *pol* gene from a non-recombinant strain were first recovered from the database. This query yielded a reference alignment of 464 *pol* sequences upon which a dereplication procedure was carried out to reduce the number of sequences in the analysis, while retaining representatives of each subtype. In a procedure similar to CD-HIT (Li and Godzik 2006), we iterated through the set of sequences, choosing exemplars that were no greater than *x*% identical to the current set of exemplars. We chose a threshold value of 92%, which reduced our initial set of 464 sequences to 17 exemplars covering all subtypes of group M except subtype K. We reintroduced a subtype K sequence and added one sequence from group N and 3 from group O to serve as outgroups to the 18 M sequences. We used MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003) for phylogenetic inference, with a mixed prior on amino acid substitution models, an eight-category approximation of the

gamma distribution of rates across sites, 10 million sampling iterations, 3 heated chains and the first 1000 trees sampled in the analysis discarded as burn-in. The extended majority consensus tree was used as the reference for subsequent analysis. In cases where multiple representatives of a given subtype were present, the set of leaves was replaced with a single leaf whose length was a weighted average of the distance to all leaves in the subtree.

The resulting tree covering 10 subtypes of group M was used as the basis for a normalized weighted UniFrac analysis, with each of the 30,002 instances identified above mapped to the appropriate subtype in the reference tree. Distance matrix construction and UPGMA clustering were carried out as in the metagenome data analysis. Direct examination of abiotic dissimilarity matrices was also performed using the GenGIS *Dissimilarity-Matrix Viewer* plugin (Fig. B.1 in Appendix B).

3.4 Results

3.4.1 Taxonomic Diversity from the Global Ocean Sampling Expedition

The GOS used environmental shotgun sequencing to collect metagenomic data from marine sample sites spread around the world. The initial publication (Rusch et al. 2007) analyzed 44 metagenomic samples (0.1 – 0.8 μm fraction) collected from 41 sites, including the Sargasso Sea sites examined previously in Venter et al. (2004). Data analyzed from these locations has revealed an immense set of novel proteins and breadth of taxonomic and functional diversity in different habitats (Yooseph et al. 2007; Yutin et al. 2007; Zhang and Gladyshev 2008; Sharma et al. 2009).

Recently, Biers et al. (2009) found differences in taxonomic diversity between coastal, oceanic, and other habitat types based on unassembled 16S rRNA gene reads. Here we considered a set of 19 locations (sites GS002-GS020 from the original paper) covering the Atlantic seaboard of North America, comprising all sites between Nova Scotia and the Panama Canal, including 3 estuarine sites (GS006, GS011 and GS012), one embayment with substantial human impact (GS005), and one freshwater lake (GS020). The latitudinal gradient of these samples, between approximately 9°N and 45°N, allows the hypothesis proposed by Fuhrman et al. (2008) to be examined. The authors of this study suggest that latitude is a primary determinant of species richness,

indicating that the northernmost samples should be less diverse than those from southern locations, although the confounding effect of different habitat types must be carefully considered. In addition to the enumeration of species richness, clustering approaches such as UniFrac (Lozupone and Knight 2005) can be used to assess between-community similarity, also known as beta diversity. Since these sites have associated geographic points and habitat parameters, we can also consider the influence of site proximity on microbial community structure.

We estimated the diversity at each site by retrieving all 16S rRNA gene sequences from each sample and assigning taxonomic attributions to these sequences based on comparisons against the GreenGenes database (DeSantis, Hugenholtz, Larsen, et al. 2006). Normalized indices of taxon richness were established using OTU counts established at various percent identity thresholds and by considering the number of distinct taxa at different taxonomic ranks (see Section 3.3.4). To examine the possible relationship between taxon richness and latitude, we visualized richness indices in GenGIS and performed linear regression analyses (Figs. 3.1 and 3.2). When all 19 locations were included in the regression model, the relationship between taxon richness and latitude was significant ($0.003 \leq p \leq 0.029$; Table A.2 in Appendix A) at 4 distinct levels of OTU clustering (unique sequences, 97%, 95% and 90%). Significant results were also obtained for normalized taxon counts at all ranks from species to phylum ($0.0001 \leq p \leq 0.008$; Table A.3 in Appendix A). Removal of the 5 samples not taken from the coast or open ocean (i.e., 5, 6, 11, 12, and 20, as identified above) from the analyses yielded models with worse fit and of marginal significance for some indices (Tables A.2 and A.3 in Appendix A). When assessing richness with normalized OTU counts, results were only significant when considering unique sequences ($p = 0.034$), marginally significant at 97% clustering ($p = 0.052$), and insignificant at lower clustering thresholds ($p = 0.14$ at 95%, $p = 0.21$ at 90%). For normalized taxon counts all results were significant ($0.005 \leq p \leq 0.019$) except at the rank of class ($p = 0.061$).

We used the unweighted and normalized weighted UniFrac phylogenetic beta-diversity measures, which compute phylogenetically weighted measures of species richness and evenness, to estimate the similarity between pairs of sites in this dataset. A

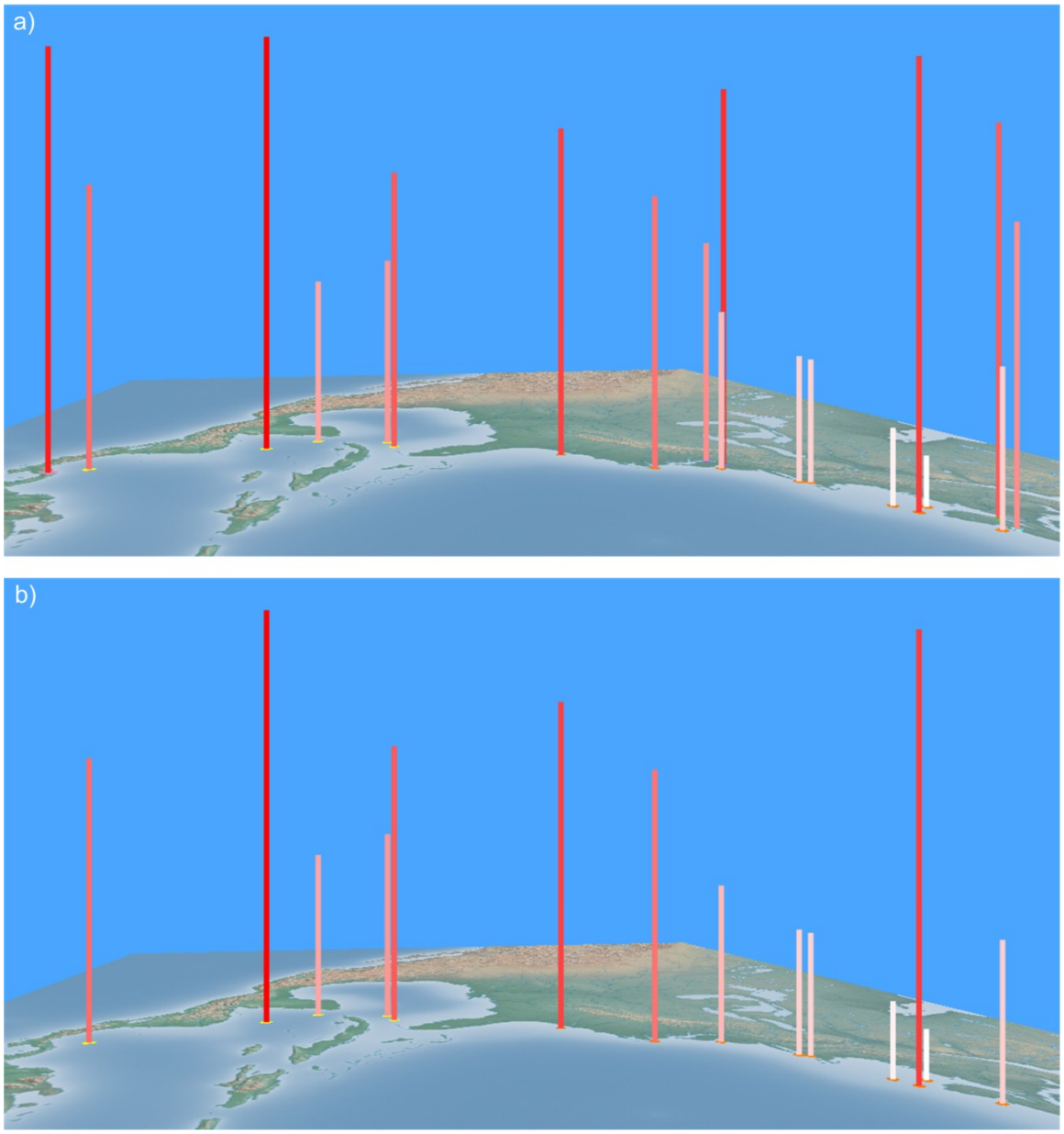


Figure 3.2. Georeferenced bar charts indicating normalized counts of unique 16S rRNA gene sequences for all 19 sample sites (a) or restricted to the 14 oceanic sample sites (b). The height and colour intensity of each bar is proportional to the normalized sequence count at that site. Maps are oriented with north on the right-hand side.

maximum-likelihood phylogenetic tree covering the proxy 16S rRNA gene sequences found at all sites was constructed and used as input to UniFrac. Figure 3.3 shows the clustering of these sites based on their phylogenetic similarity as determined using normalized weighted UniFrac. The geographic axis in this figure, depicted as a pair of

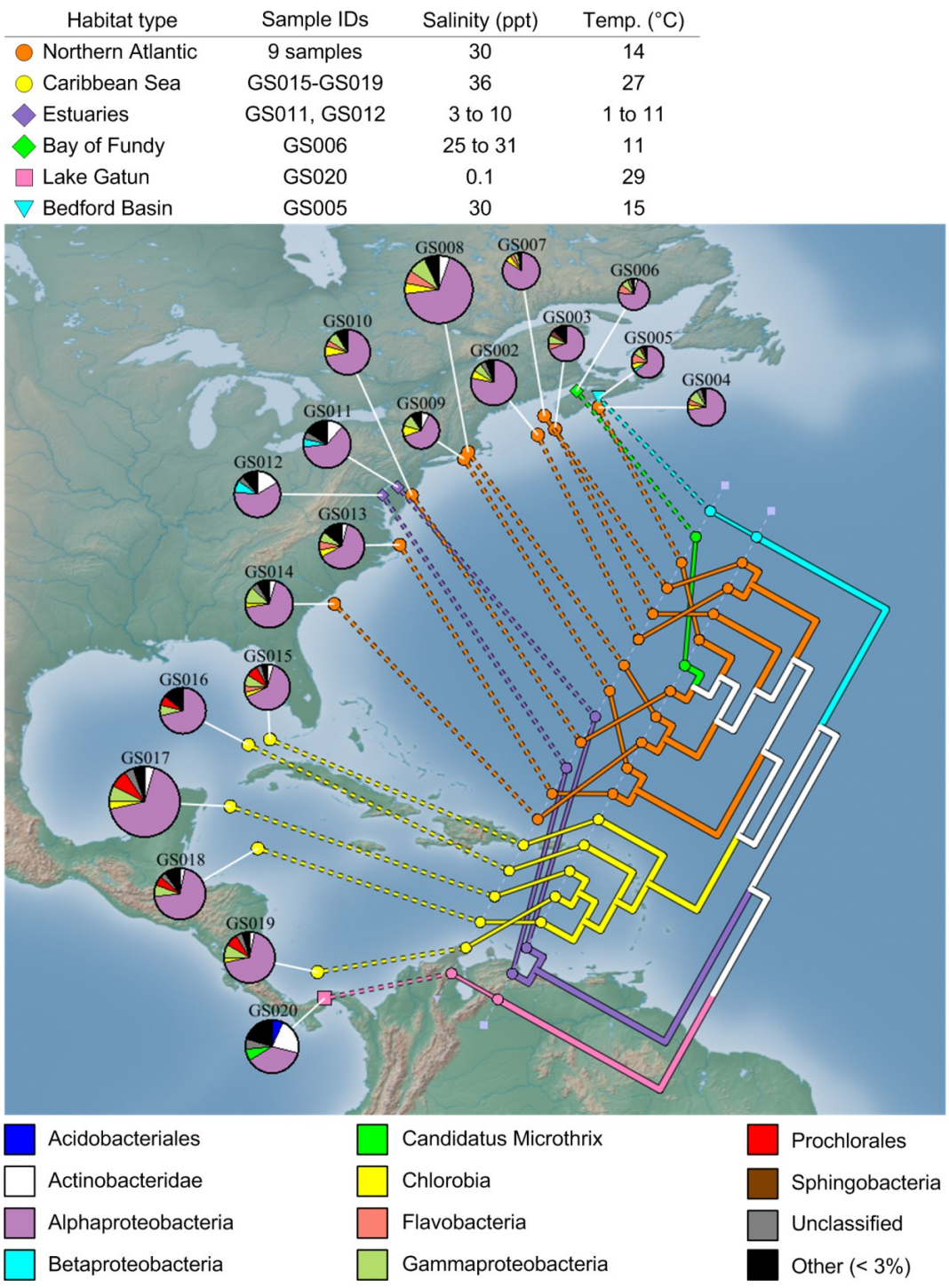


Figure 3.3. Clustering of GOS sites based on their shared phylogenetic diversity as determined by normalized weighted UniFrac. Pie charts associated with each GOS site show the breakdown of 16S rRNA gene sequences by best-matching bacterial group (common phylum, class, or genus) with rare groups collected together in the "other" category. Pie chart sizes are proportional to the total number of 16S rRNA gene sequences considered at each site. White branches in the tree indicate internal edges whose children cover multiple habitat types.

parallel lines, corresponds to the main axis along which sequence datasets were sampled. When geographic locations are mapped to the leaves of the optimized tree, a globally optimal minimum of 28 crossings is observed. A permutation test on the order of sample sites along this axis yielded 4/1,000 randomly generated permutations with 28 or fewer crossings, corresponding to a *p*-value of 0.004. Comparing this result against the typical $\alpha = 0.05$ threshold of significance leads to a rejection of the null hypothesis, suggesting that nearby sites may indeed have a stronger tendency toward mutual similarity. A corresponding unweighted UniFrac analysis yielded similar results, albeit with more crossings and a higher *p*-value (35 crossings, $p = 0.031$).

The above analysis conflates geographic and habitat effects, and a closer inspection is needed to understand the relative contribution of these factors to community similarity. To separate the effects of habitat type from those of geographic proximity, we performed the analysis on the full dataset, a reduced dataset of 14 sites as above, and a further partitioning of the 14 sites into those collected from either the Atlantic seaboard (9 sites) or the Caribbean Sea (5 sites). To facilitate comparisons we used a strict north-south axis for mapping of geographic points. The geographic fit of the full set to this axis was slightly worse than that shown above (normalized weighted UniFrac: 29 crossings, $p = 0.019$; unweighted UniFrac: 36 crossings, $p = 0.028$). Although deletion of the "unusual" habitat types from the set diminished the significance of the richness model reported above, the opposite effect was seen in the similarity-based UniFrac results on the 14-site set (Fig. 3.4; normalized weighted UniFrac: 6 crossings, $p = 0.001$; unweighted UniFrac: 8 crossings, $p = 0.003$). A further partitioning of sites into sets from the Atlantic seaboard and Caribbean Sea yielded results that were not statistically significant ($0.109 \leq p \leq 0.466$ for all combinations of the 2 UniFrac measures and the 2 sets of sites). Consequently, while there is a geographic signal in the similarity relationships between sites, most of this appears to be due to the partitioning of Atlantic seaboard versus Caribbean Sea sites, with no significant trend within either of these 2 regions.

The normalized weighted UniFrac and unweighted UniFrac trees display a wide range of jackknife support values (Fig. 3.5). We complemented the analysis of jackknifed

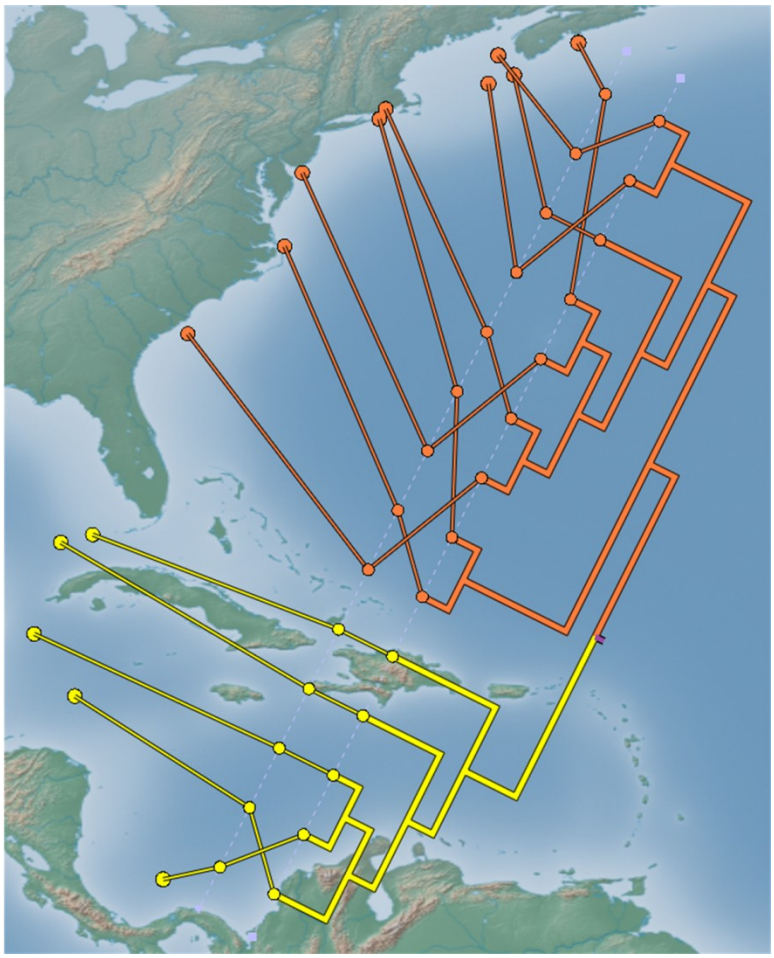


Figure 3.4. Clustering of a subset of 14 oceanic GOS sites based on their shared phylogenetic diversity as determined with normalized weighted UniFrac. Sample site colouring is consistent with Figure 3.3.

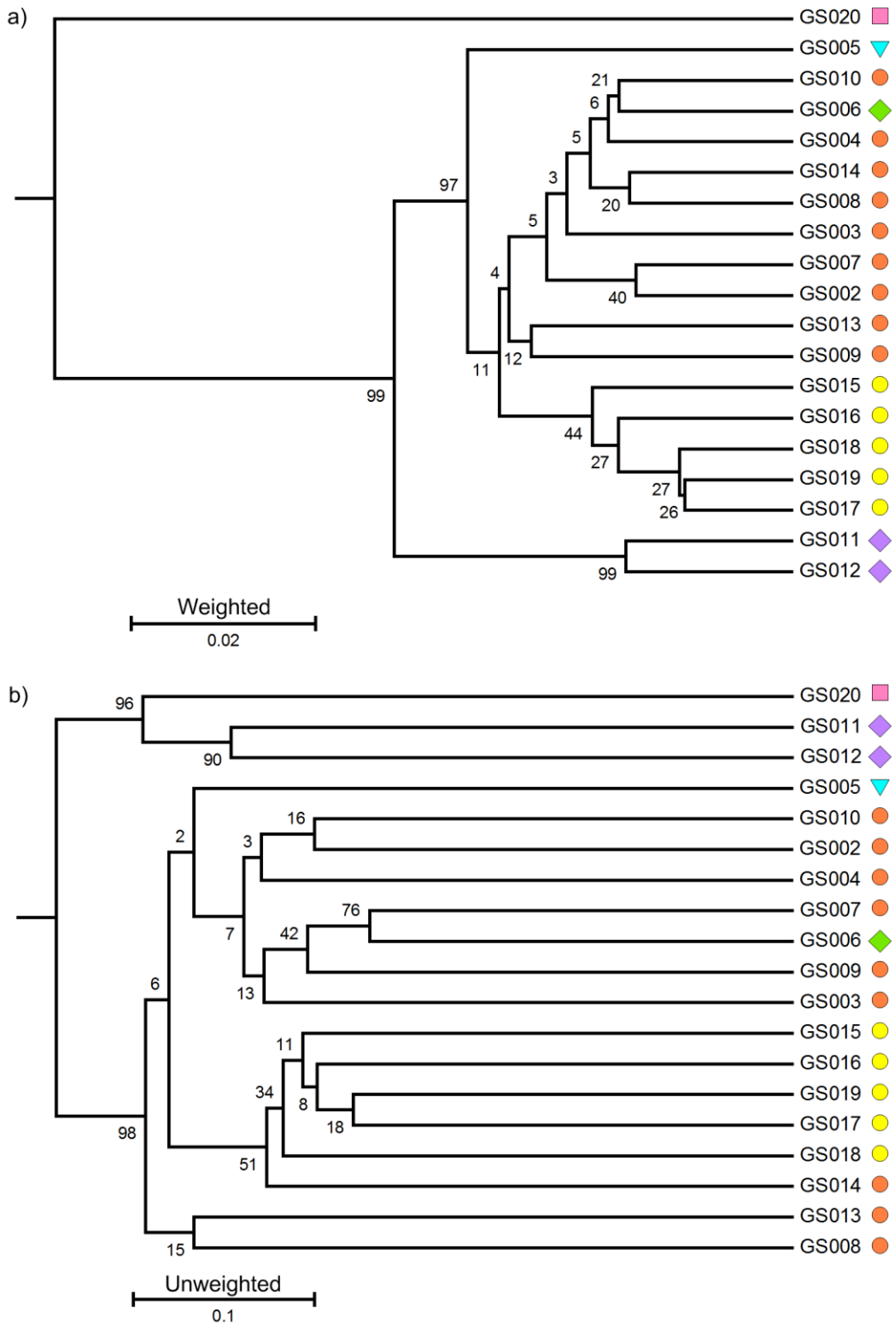


Figure 3.5. UPGMA clustering of GOS sites with associated jackknife support values as determined with normalized weighted UniFrac (a) and unweighted UniFrac (b). Sample site colouring and shape are consistent with Figure 3.3.

trees with pie chart visualizations of the most highly variable taxonomic groups across sites (Figs. 3.3 and 3.6). The grouping of 3 low-salinity sites were supported with jackknife values ≥ 90 as was the internal grouping of Delaware Bay and Chesapeake Bay, suggesting strong differentiation in both richness and relative abundance compared to the other samples. Lake Gatun is perhaps the most unusual site, uniquely having < 50% Alphaproteobacteria, and relatively high amounts of Acidobacteriales, Actinobacteridae and other groups that are rare or absent from other sites. Delaware and Chesapeake Bays were overrepresented by Actinobacteriadae (as with Lake Gatun) and Betaproteobacteria (unlike Lake Gatun). The higher proportion of Actinobacteridae at the low-salinity sites was previously reported by Biers et al. (2009). The similarity among the Caribbean sites can largely be attributed to the relatively high abundance of Prochlorales, specifically *Prochlorococcus*, which is consistent with an expected increased abundance of picocyanobacteria in warmer waters (Johnson et al. 2006). However, the separation of Caribbean sites is only supported by 44% and 34% of jackknife replicates in the weighted and unweighted UniFrac analyses, respectively, suggesting that differences in richness and relative abundance, although significant, are not as pronounced as those associated with the low-salinity sites. Our results also indicate that the Bedford Basin, Nova Scotia site is distinct from all other sites with a relatively high proportion of betaproteobacterial sequences and a complete lack of Actinobacteridae. This may reflect the inflow of wastewater into the Bedford Basin over the last 250 years (Metro Engineering Inc. 1993). Conversely, the Bay of Fundy, with salinity levels that are similar to open ocean sites, was indistinguishable from other Atlantic Ocean sites in both analyses, although its closest neighbour was different in the normalized weighted UniFrac (GS010) and unweighted UniFrac (GS007) analyses.

We also contrasted the phylogenetically weighted UniFrac results with those obtained with the Bray-Curtis index, a traditional taxon-based measure of beta diversity. Bray-Curtis results were obtained for normalized species and genus counts. At the species level, the Caribbean Sea samples still form a separate cluster in contrast to the Atlantic seaboard samples which show far less structuring than observed with either of the UniFrac measures (Fig. 3.7a). Interestingly, the Bedford Basin sample clusters with many

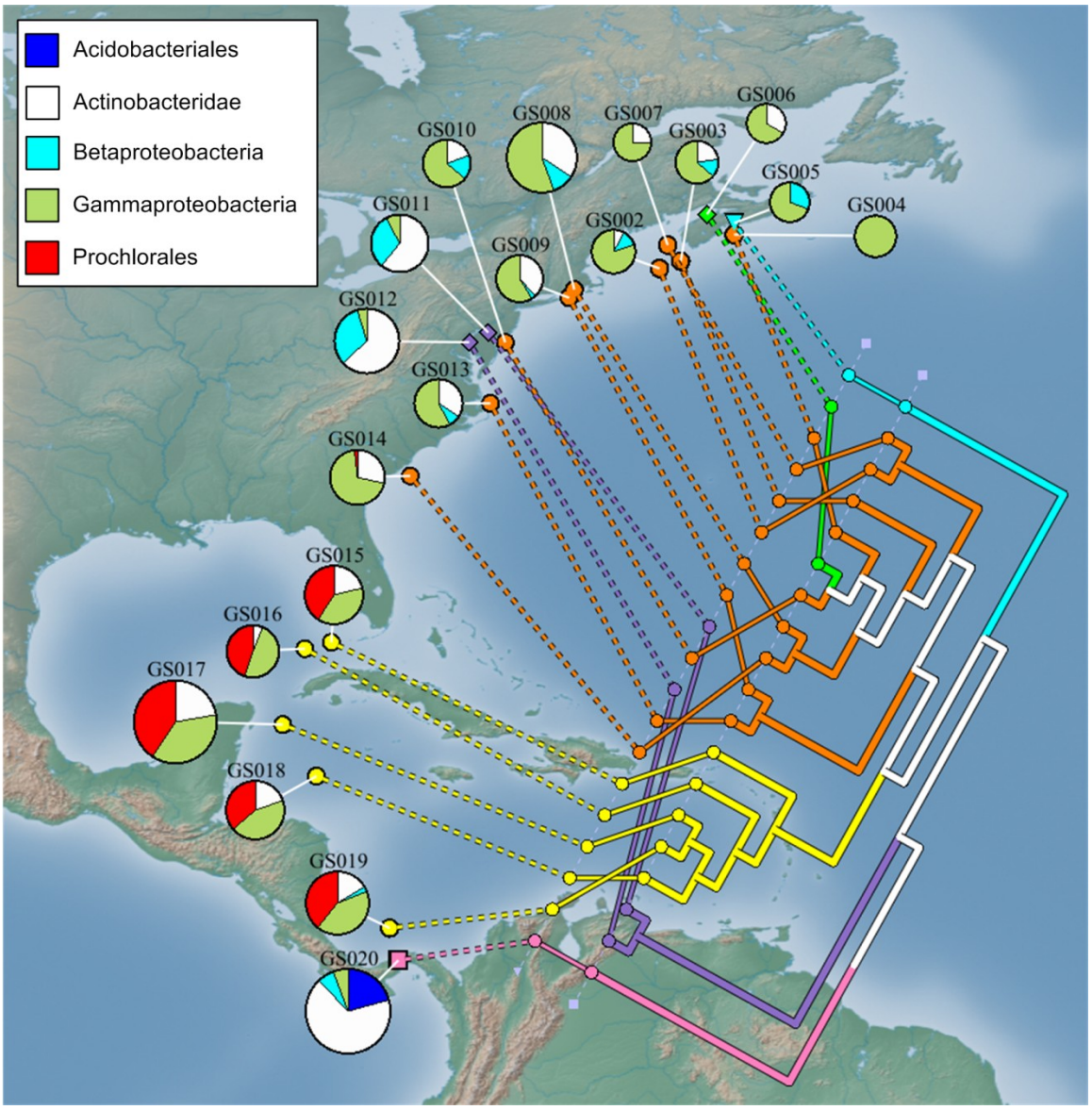


Figure 3.6. Relative abundance of 5 taxonomic groups whose distributions are highly variable across the 19 GOS sites considered. Taxon and sample site colouring are consistent with Figure 3.3.

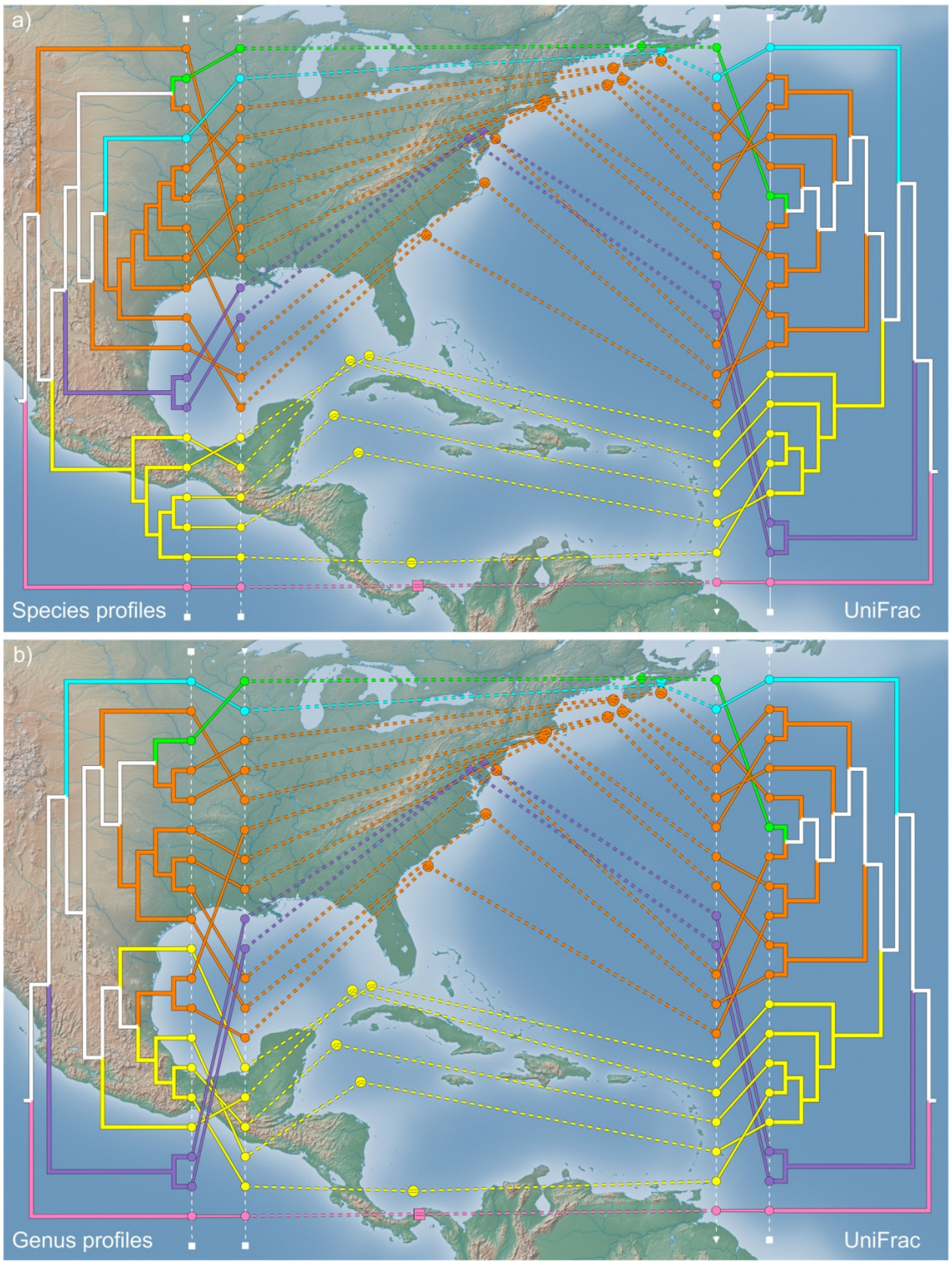


Figure 3.7. Taxonomic and phylogenetic similarity of GOS communities. (**a**) Comparison of community similarity determined by applying the Bray-Curtis index to species profiles or normalized weighted UniFrac to a 16S rRNA gene phylogeny. (**b**) Bray-Curtis index applied to genus profiles contrasted with normalized weighted UniFrac. Community relationships are shown as UPGMA cluster trees. Sample site colouring is consistent with Figure 3.3.

of the Atlantic seaboard samples when applying the Bray-Curtis index to species profiles which is in direct contrast to the results obtained on genus profiles or with normalized weighted UniFrac. This may suggest that the Bedford Basin contains taxa from relatively distinct lineages, or may simply be the result of the large number of sequences which were not assigned a species label. For genus-level profiles, the Caribbean Sea and Atlantic seaboard samples become intermixed demonstrating that the incorporation of phylogenetic information (either explicitly or by creating profiles at different taxonomic ranks) can substantially influence measured beta diversity (Fig. 3.7b). Nonetheless, some patterns are recovered across all considered cases: the clustering of the 2 estuary sites, the highly distinct composition of Lake Gatun, and the Bay of Fundy appearing more similar to Atlantic seaboard samples than to the 2 estuary samples.

3.4.2 Non-recombinant HIV-1 Subtypes in Africa

The reverse transcriptase-directed replication of HIV-1 is extremely error-prone, leading to very rapid rates of genomic change through mutation and recombination (Drake 1993; An and Telesnitsky 2002). The "major" or M group of HIV-1 is subdivided into several subtypes based on sequence similarity and likely shared ancestry within the M group; each of these subtypes is nonetheless genetically diverse and amino acid variation in the viral envelope protein within a subtype can approach 20% (Korber et al. 2001). Together with their derived recombinant forms such as CRF01(AE) and CRF02(AG), these subtypes are responsible for the vast majority of HIV infections worldwide. Subtype distributions vary dramatically by continent, country, and region (Kuiken et al. 2000; Peeters et al. 2003; Hemelaar et al. 2006), and there is considerable evidence and speculation that subtype differences influence the likelihood of detection, disease progression, and potential responses to antiviral treatment (Vasan et al. 2006; Taylor et al. 2008). The geographic origins of certain subtypes have been probed in depth: for instance, it is thought that the widely dispersed subtype B may have originated in Haiti during the 1960s (Gilbert et al. 2007).

To assess the extent to which HIV subtypes collected from different countries in Africa constitute distinct geographic clusters, we extracted full-length sequences of the HIV *pol* gene from the HIV sequence database (http://www.hiv.lanl.gov/). Given the

difficulties in computing phylogenetic diversity from sequences with ambiguous or conflicting phylogenetic signals, we restricted our analysis to the non-recombinant subtypes A-D, F-H, J, and K, although we note the controversy surrounding the non-recombinant nature of some of these subtypes (Abecasis et al. 2007). Only countries with at least 10 samples in this dataset were retained, yielding a total of 40 countries with sequence counts between 12 (Guinea-Conakry) and 6576 (South Africa).

Since the sampling depth varied dramatically among subtypes, we elected to use a rooted tree with one leaf representing each subtype as the basis for a normalized weighted UniFrac analysis (see Section 3.3.5). The results of this analysis were visualized in GenGIS as a three-dimensional UPGMA clustering in order to explore the relative similarity of HIV subtypes in each country (Fig. 3.8). Three-dimensional trees such as this can be difficult to interpret in a static two-dimensional image, but we have coloured 4 major groupings of countries that show a certain degree of geographic separation and appear to be largely driven by common subtypes (Fig. 3.9). Eastern and Southern Africa are dominated by subtype C and constitute a cluster (coloured purple in Fig. 3.8), with the notable exception of Tanzania, whose profile across 3010 sequences is nearly 50% subtype A and 25% each of subtypes C and D. Tanzania's closest affinities within the UPGMA clustering are with other countries that contain a substantial fraction of subtype D, including Equatorial Guinea, Uganda, Sudan, and Chad. The larger cluster that includes these countries also includes the B-dominated North African countries as well as the Indian Ocean islands, which contain a mixture of subtypes A, B, and C. The close proximity of North and Central African clusters appears to be an artifact arising from the partial affinities of each for the island countries. Other countries with a substantial representation of subtype A fall into either the green cluster which includes Kenya, Rwanda, the Central African Republic, Cote d'Ivoire, Ghana, and Benin, or the cyan cluster which includes the most diverse countries in the set such as Cameroon, the Democratic Republic of the Congo, Angola, Senegal and Burkina Faso. A handful of West African countries are dominated by subtype G; two of these, Niger and Nigeria, constitute a basal branch in the large cluster that also covers the rest of Western Africa. The other basal branch in this large cluster maps to the Gambia, which contains a rich

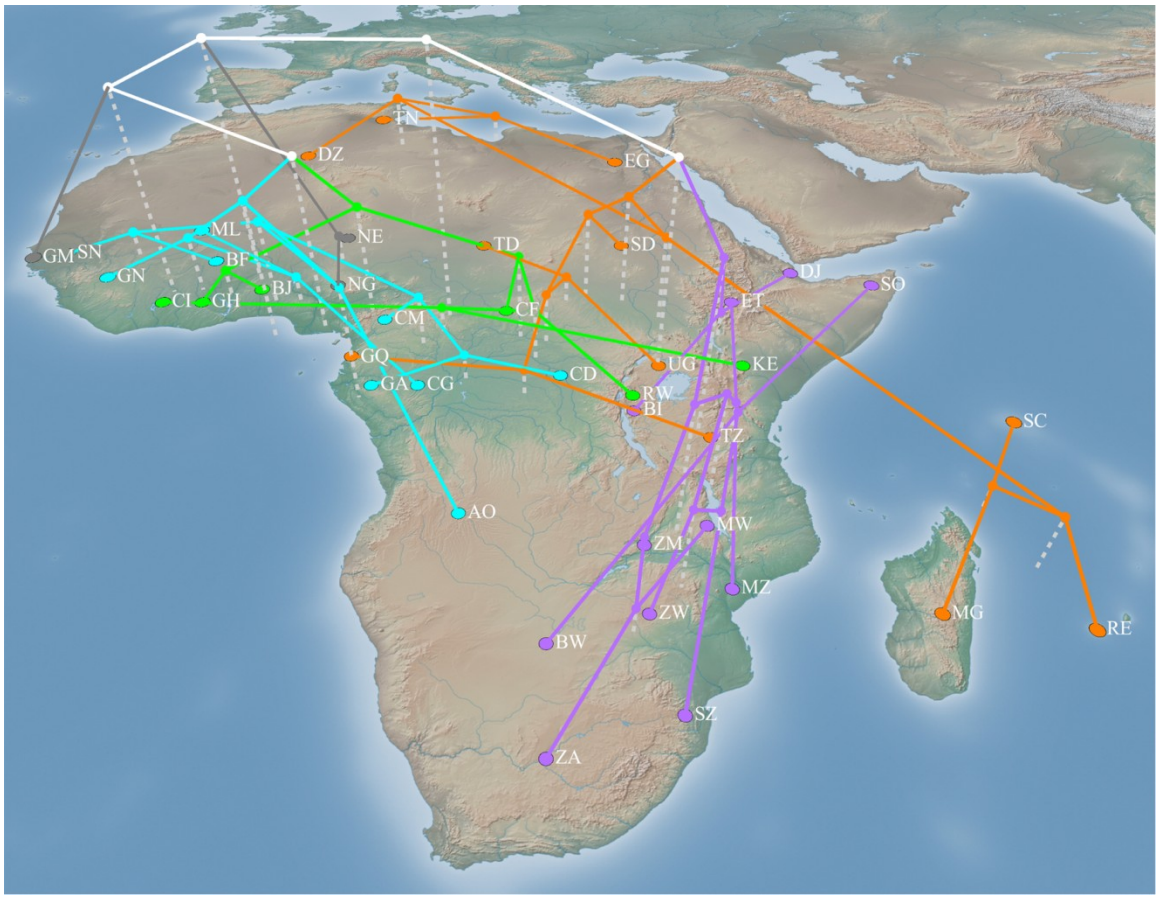


Figure 3.8. Clustering of African nations based on phylogenetic diversity of HIV subtypes. The UPGMA clustering of countries based on their UniFrac scores is shown using a three-dimensional tree; the 4 subclusters discussed in the text and the countries they cover are indicated by colouring different subtrees orange, cyan, purple, and green. Location identifiers are mapped to the geographic center of each country, which is also identified with the standard two-letter country code.

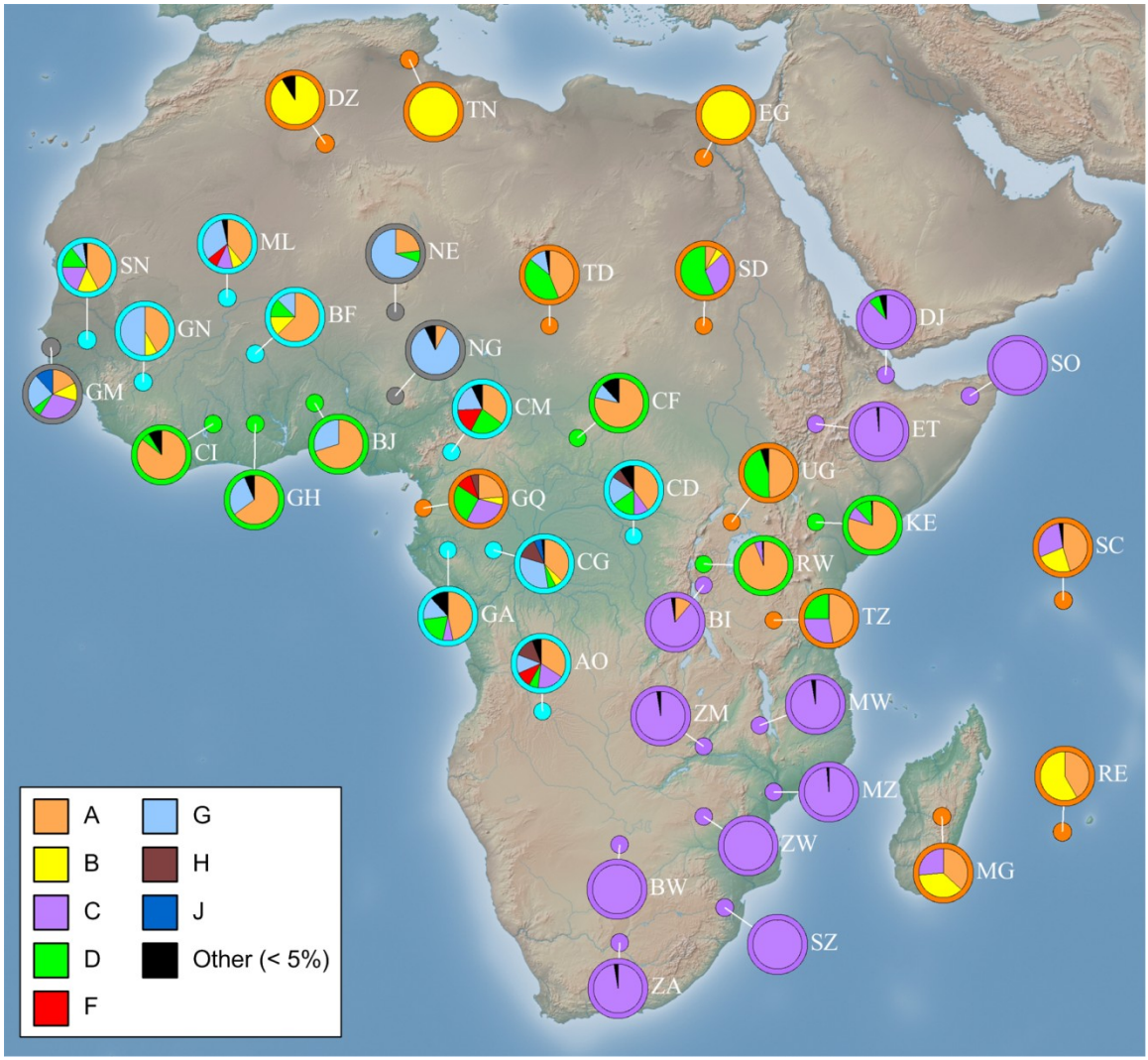


Figure 3.9. Distribution of non-recombinant HIV subtypes in 40 African countries. Pie charts indicate the breakdown of HIV *pol* gene subtypes by country, with subtype-to-colour mapping indicated in the legend. Two-letter country codes are attached to each pie chart to indicate the corresponding country, and droplines point to the geographic center of each country. Chart sizes are constant rather than proportional to the number of sequences. The colour assigned to each country and the outer ring of the pie charts reflects the 4 subclusters identified in Figure 3.8.

and evenly distributed set of subtypes (even though only 17 sequences are available from this country) and shows no strong affinity for any other country. Unsurprisingly, the West African nations with higher proportions of A and G tend to be dominated by the circulating recombinant form AG(02).

Although hierarchical clustering algorithms are a common and useful technique for visualizing biotic dissimilarity matrices, they necessarily depict only the most salient aspects of the data and can produce varying results as they emphasize different aspects of the data (Legendre and Legendre 1998). Direct visualization of dissimilarity values can help alleviate these limitations. The *Dissimilarity-Matrix Viewer* in GenGIS can be used to visualize portions of a biotic dissimilarity matrix specified either as a range of values, or selected based on the affinity of each location to a particular location of interest (Fig. B.1 in Appendix B). Using this plugin to visualize the 13 countries most similar to Tanzania shows similarities with the UPGMA clustering, but also shows high similarity between Tanzania and the West African countries of Senegal and Burkina Faso (Fig. 3.10). Like Tanzania, these countries are dominated by sequences of subtype A and have a relatively high proportion of subtype D sequences compared to other West African countries.

It is important to recognize that the HIV sequences considered in this analysis do not constitute random samples and the effects of differences in sampling effort in different regions has been noted before (Soares 2007). Also, some circulating recombinant subtypes (particularly AG) constitute a significant proportion of reported infections in many African countries, so their exclusion can potentially exert a large influence on the observed subtype diversity. Nonetheless, if the impact of unequal sampling efforts can be quantified and potentially mitigated through reweighting or georeferencing at resolutions higher than countries, then diversity patterns can be used to define and test epidemiological hypotheses concerning HIV and other pathogens such as Influenza A (Janies et al. 2007). For instance, the exceptional subtype distributions seen in Tanzania that lead it to cluster with countries in Central Africa is consistent with the hypothesis that events such as the Tanzania-Uganda War, which ended in 1979, were responsible for founder events that introduced non-C subtypes into Tanzania, while C arrived later from

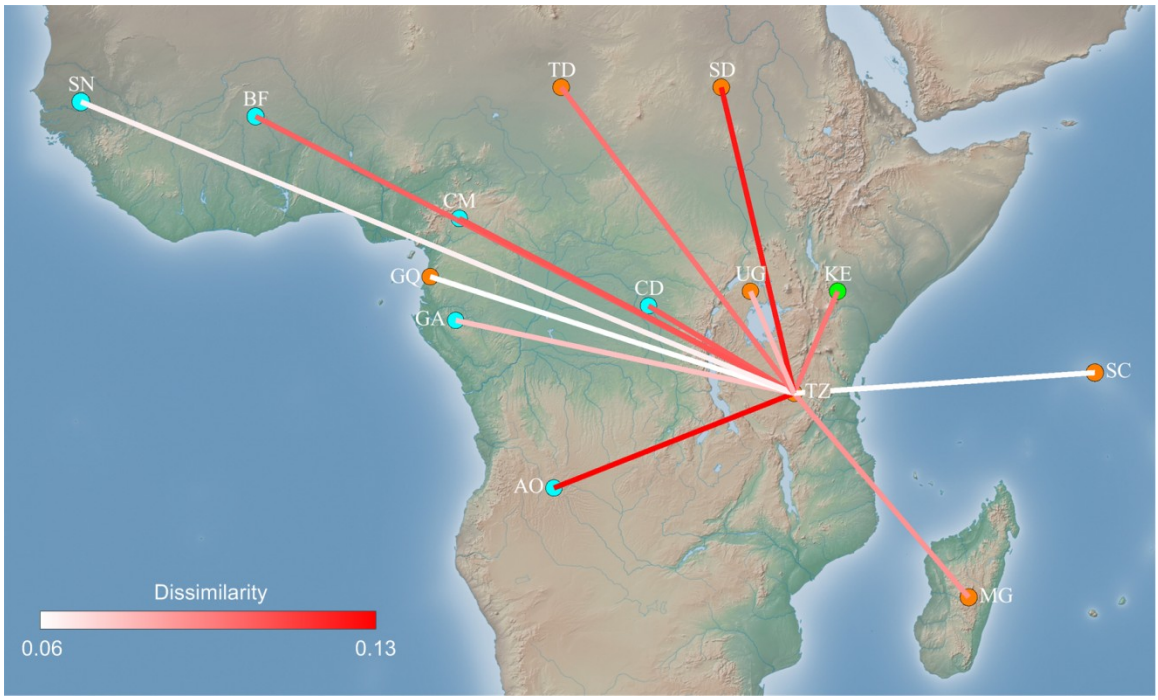


Figure 3.10. Thirteen countries most similar to Tanzania as determined by normalized weighted UniFrac. Two-letter country codes are shown next to each location marker. The colour assigned to each country reflects the 4 subclusters identified in Figure 3.8.

elsewhere in East Africa (Serwadda et al. 1985; Vasan et al. 2006). Additionally, Kenya and Tanzania may show distinct patterns due to the convergence of major north-south and east-west travel corridors (Bwayo et al. 1994; Robbins et al. 1999). In such cases, a clustered network may be a more appropriate representation of similarities than a tree.

3.5 Discussion

By coupling digital map data with georeferenced sequence information, GenGIS has allowed us to visualize patterns of microbial species and viral subtype distribution. In other work, GenGIS has been used to explore the emergence and global dispersal of the H1N1 (2009) "swine flu" pandemic (Parks, MacDonald, et al. 2009), the biogeography of plants and animals (Allal et al. 2011; Shafer et al. 2011), and the geographic distribution of language families (Walker et al. 2012). GenGIS is thus sufficiently flexible to be applied to many different types of genetic and genomic data, while at the same time allowing targeted analyses to be implemented and carried out. Biogeographic software typically focuses on one of 3 areas: 1) displaying site-specific community data, 2)

generating geophylogenies, *or* 3) visualizing the spatial distribution of densely sampled species data. GenGIS is unique in providing visualizations and analyses for both site-specific community data (area 1) and geotrees (area 2). Furthermore, it provides a flexible scripting and plugin framework for developing custom visualizations and analyses. Even though the existing GenGIS framework is robust enough to generate custom visualizations of the spatial distribution of densely sampled species data (area 3; see http://kiwi.cs.dal.ca/GenGIS/H1N1 for an example showing the temporal spread of the 2009 H1N1 "Swine Flu" outbreak), we plan to incorporate specific visualizations and analyses into GenGIS that address these data sources. Given the rich set of georeferenced molecular data becoming available, we believe an interactive visualization platform capable of utilizing all sources of data is essential for assessing biogeographic patterns.

The above analyses demonstrate the different interpretations that can be attached to hierarchical clusters of data. In the GOS example, hierarchical analysis of shared phylogenetic diversity using different subsets of sites indicated that habitat types were the primary separating feature, with a strongly supported split observed between low-salinity and high-salinity sites, consistent with the observations of Lozupone and Knight (2007). There were too few low-salinity sites to support a refined analysis within this group, but among oceanic sites the key driver of geographic structure was the separation of Atlantic from Caribbean sites, with the presence of picocyanobacteria as the principal factor influencing this separation. Similarity in relative abundance (as assessed using normalized weighted UniFrac) yielded a stronger clustering signal than similarity in richness (as assessed using unweighted UniFrac). Consideration of hierarchical clusters reflecting shared species or genera between samples was not always congruent with the above results indicating that taxon-based and phylogenetic-based measures of beta diversity provide complementary information (Graham and Fine 2008). Robust visualization environments, such as GenGIS, facilitate contrasting different measures and provide multiple views of the same data in order to provide deeper insights into the relationships between communities. Our clustering of countries based on their HIV-1 subtype profiles highlighted regions with similar patterns of diversity, which in some cases corresponded to previously observed trends that arose due to historical events. Although the clustering of some countries is likely unstable due to small sample sizes and

the imposition of a strict tree structure, the hybrid patterns in east Africa were clear and supported by several thousand sequences in each affected country.

Our chosen examples also illustrate some of the challenges that are well-known in population genetics and phylogenetics, including the use of trees to represent network-like data. The effects of forcing a tree structure on data that are not inherently treelike has been characterized for sequence alignments (Posada and Crandall 2002) and aggregate trees (Wiens 1998; Beiko, Doolittle, et al. 2008), and in many cases the recovered tree may contain features that are not present in the source data. Given the considerable evidence for network-like relationships in phylogenomic analyses (Beiko et al. 2005; Zhaxybayeva et al. 2006; Dagan et al. 2008) as well as population-level datasets such as our HIV example cited above, network visualizations will be a valuable future addition to GenGIS. Other potential problems such as uncertainty in tree inference, and the confounding effects of population migration and admixture, will need to be addressed through careful and thorough sampling and application of inferential techniques.

The number and size of genetic datasets that are available from public repositories is growing and all of the data used in this study were acquired from such resources. However, our vision for GenGIS includes not only the analysis of static datasets prepared in advance by a user, but also direct integration with emerging online repositories including the Barcode of Life database (Ratnasingham and Hebert 2007), the HIV sequence database (http://www.hiv.lanl.gov), the RDP (Cole et al. 2009), the Map of Life initiative (Jetz et al. 2012), and the Biomonitoring 2.0 web portal (Baird and Hajibabaei 2012). Querying online datasets will require extensions to the selection techniques currently available in GenGIS, but will then allow the monitoring of changes in community structure, and the emergence of novel pathogen genotypes, recombinants or environmental organisms. Beyond the automated acquisition of sequence data, another emerging opportunity lies in the increased availability of online ecological data with global scope (Kozak et al. 2008). Habitats present a complex combination of environmental features, and the acquisition of such data would offer the opportunity to test more candidate environmental factors such as nutrient concentrations and historical patterns of temperature, salinity, or rainfall, that may individually or collectively have a significant impact on community diversity and function.

3.6 Acknowledgements

We thank Harman Clair, Norman MacDonald, Gregory Smolyn, Kathryn Dunphy, Conor Meehan, Morgan Langille and Daniel Ruzzante for assistance with the development of GenGIS, and all the GenGIS users who have provided feedback. DHP is supported by the Killam Trusts and the Natural Sciences and Engineering Research Council of Canada; TM and MSP are supported by Genome Canada and the Ontario Genomics Institute; RGB is supported by Genome Atlantic, the Canada Foundation for Innovation, and the Canada Research Chairs program. This project was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute through the Biomonitoring 2.0 project (OGI-050: see http://biomonitoring2.org) and grant number 2009-OGI-ABC-1405.

Chapter 4

Measures of Phylogenetic Differentiation Provide Robust and Complementary Insights into Microbial Communities

Parks DH, Beiko RG. 2012a. Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities.

Publication status: In press at ISME J (July, 2012).

Contribution to research: **DHP** conceived of and carried out the research. RGB provided guidance throughout the project.

Contribution to writing: Written by **DHP** with suggestions and editorial advice provided by RGB.

4.1 Abstract

High-throughput sequencing techniques have made large-scale spatial and temporal surveys of microbial communities routine. Gaining insight into microbial diversity requires methods for effectively analyzing and visualizing these extensive datasets. Phylogenetic beta-diversity measures address this challenge by allowing the relationship between large numbers of environmental samples to be explored using standard multivariate analysis techniques. Despite the success and widespread use of phylogenetic beta-diversity measures, an extensive comparative analysis of these measures has not been performed. Here we compare 39 measures of phylogenetic beta diversity on 4 recently published microbial community datasets in order to establish the relative similarity of these measures along with key properties and performance characteristics. Although many measures are highly correlated, those commonly used within microbial ecology were found to be distinct from those popular within classical ecology, and from the recently recommended Gower and Canberra measures. Many of the measures are surprisingly robust to different rootings of the gene tree, the choice of similarity threshold used to define operational taxonomic units, and the presence of outlying basal lineages. We also established that measures differ considerably in their sensitivity to rare

organisms, and that the effectiveness of measures can vary substantially under alternative models of differentiation. Consequently, the depth of sequencing required to reveal underlying patterns of relationships between environmental samples depends on the selected measure. Our results demonstrate that using complementary measures of phylogenetic beta diversity can further our understanding of how communities are phylogenetically differentiated.

4.2 Introduction

Advances in DNA sequencing technology allow high-throughput recovery of genetic material directly from environmental samples. By using the 16S rRNA gene to establish the members of naturally occurring microbial communities, large-scale surveys have shed light on spatial and temporal patterns of microbial diversity (Martiny et al. 2006; Caporaso et al. 2011). Recent studies have revealed the relative influences of environmental factors on global patterns of diversity (Lozupone and Knight 2007; Lauber et al. 2009; Rousk et al. 2010), the impact of antibiotics on the gut microbiota of mice and humans (Dethlefsen et al. 2008; Ubeda et al. 2010), and established that human-associated communities differ between individuals and body habitats (Costello et al. 2009; Turnbaugh et al. 2009; Fierer et al. 2010). With surveys now encompassing hundreds of environmental samples, a primary challenge is to identify the biotic and abiotic factors that engender differences in microbial community structure. Beta-diversity measures address this challenge by providing a univariate statistic establishing the relative similarity of any pair of samples. Exploratory multivariate statistical techniques, such as hierarchical clustering and ordination, can then be used to identify trends across large numbers of samples.

Although beta-diversity measures have traditionally been determined on the basis of overlap between discretely defined sets of entities (species or OTUs), recent methods have incorporated phylogenetic information in order to establish the relative similarity of OTUs (Clarke and Warwick 1998; Martin 2002; Lozupone and Knight 2005; Graham and Fine 2008). By exploiting the hierarchical relatedness of organisms, phylogenetic beta-diversity measures are often more effective at revealing underlying ecological patterns (Hamady et al. 2010; Nipperess et al. 2010). However, as a univariate statistic a single beta-diversity measure cannot address all manners in which the similarity between

samples may be usefully defined. Consequently, many different measures of beta diversity have been proposed which vary in their treatment of community properties, such as the presence of rare OTUs or the relative abundance of OTUs (Legendre and Legendre 1998). This latter factor is commonly used to classify beta-diversity measures as either *quantitative*, where the relative abundance of each OTU influences the measured similarity, or *qualitative*, where only the presence or absence of an OTU is considered. These 2 classes of measures provide complementary information as quantitative measures indicate whether or not ecological differences between habitats cause the abundance of taxonomic groups to change, whereas qualitative measures suggest whether or not ecological factors prohibit a taxonomic group from occupying certain habitats.

Due to the complexity of naturally occurring communities and the wide range of mechanisms that can cause communities to differentiate, it can be beneficial to apply several phylogenetic beta-diversity measures. Our aim is to establish a set of properties and a methodology for determining a practical subset of measures which will provide complementary information on the similarity of microbial samples. We build upon initial surveys that have considered a limited number of phylogenetic beta-diversity measures under a restrictive set of conditions (Schloss 2008; Nipperess et al. 2010; Root and Nelson 2011; Swenson et al. 2011) by establishing key properties and performance measures for 24 quantitative and 15 qualitative measures (Table 4.1). We consider phylogenetic beta-diversity measures popular within microbial and classical ecology along with newly established phylogenetic extensions of commonly used taxon-based (non-phylogenetic) measures. Both the Gower and Canberra measures recently recommended by Kuczynski et al. (2010) are considered and are of particular interest as they have not been widely used or evaluated. We also consider all 3 variants of the UniFrac measures which are ubiquitous in the microbial ecology literature (Lozupone and Knight 2005; Lozupone et al. 2007): unweighted UniFrac, weighted UniFrac, and normalized weighted UniFrac. The F_{ST} measure popular for studying human migration patterns (Pakendorf and Stoneking 2005) is considered along with the mean phylogenetic distance (MPD) and mean nearest neighbour distance (MNND) measures which are widely used in ecological studies of multicellular eukaryotes such as plant and animal

Table 4.1. Phylogenetic beta-diversity measures.

Quantitative and qualitative measures are given below. Commonly used names for each measure are provided. For simplicity, measures are referred to by the first quantitative name listed in the table. Qualitative measures are specified by prefixing the associated quantitative name with a "u". All formulas specify a measure of dissimilarity. Names referring to a similarity measure are indicated as being the complement of the provided formula. Taxon-based measures extended in this manuscript to include phylogenetic information are shown in bold. References for each measure are given in Table C.1 in Appendix C. Formulas specify the dissimilarity between communities *i* and *j* using the following notation:

- p_{in} is the proportion of sequences from community *i* descendant from branch *n*.
- W_n is the weight or length of branch *n*.
- *N* the number of branches in the phylogeny.
- *a* is the amount of shared branch length, *b* is the amount of branch length unique to community *i,* c is the amount of branch length unique to community *j*, and *d* is the amount of branch length external to communities *i* and *j* (see Nipperess et al. 2010 for details).
- $p_{+n} = \sum_k p_{kn}$ is the proportion of sequences assigned to branch *n* across all communities.
- $p_{i+} = \sum_n p_{in}$ is the total proportion of sequences from community *i* across all branches.
- $W_{i+} = \sum_n p_{in} W_n$ is the weighted proportion of sequences from community *i* across all branches.
- $W_+ = \sum_n W_n$ is the total branch length.
- $\mathrm{cov}(p_i, p_j; W) = \frac{\sum_n W_n \left(p_{in} - W_{i+}/W_+\right)\left(p_{jn} - W_{j+}/W_+\right)}{W_+}$ is the weighted covariance between vectors p_i and p_j weighted by W_n.
- *X* is the set of leaf nodes containing sequences from community *i*.
- *Y* is the set of leaf nodes containing sequences from community *j*.
- $L_i = \sum_{x \in X} p_{ix}$ is the sum of sequence proportions across all leaf nodes. This is always equal to 1. It is explicitly shown in the formulas below to indicate measures that will differ when applied to raw count data.
- R_l is the phylogenetic distance from leaf node *l* to the root node.
- $\max_k \left(p_{kn}\right)$ is the maximum proportion of sequences descendant from branch *n* across all communities.
- $D_T = \sum_x^{X \cup Y} \sum_y^{X \cup Y} \frac{p_{ix} + p_{jx}}{2} \frac{p_{iy} + p_{jy}}{2} \delta_{xy}$
- $D_S = \frac{1}{2}\left(\sum_x^X \sum_z^X p_{ix} p_{iz} \delta_{xz} + \sum_y^Y \sum_z^Y p_{jy} p_{jz} \delta_{yz} \right)$
- δ_{xy} is the phylogenetic distance from sequence *x* to sequence *y*.

Table 4.1. Phylogenetic beta-diversity measures (continued).

Quantitative		Qualitative	
•Bray-Curtis •Normalized weighted UniFrac •Percentage difference	$\frac{\sum_n \lvert p_{in} - p_{jn} \rvert W_n}{\sum_n (p_{in} + p_{jn}) W_n}$	•Sørensen (complement) •PhyloSor •Dice's index (complement)	$\frac{b+c}{2a+b+c}$
•Bray-Curtis (MRCA restricted)	Bray-Curtis calculated over the most recent common ancestor subtree for a pair of samples.	-	(see quantitative)
•**Canberra**	$\sum_n \frac{\lvert p_{in} - p_{jn} \rvert}{p_{in} + p_{jn}} W_n$	•Canberra	$b+c$
•**Chi-squared**	$\sqrt{\sum_n \frac{W_n}{p_{+n}} \left(\frac{p_{in}}{L_i} - \frac{p_{jn}}{L_j} \right)^2}$	-	-
•**Coefficient of similarity** (complement)	$\sum_n \frac{\lvert p_{in} - p_{jn} \rvert}{\max(p_{in}, p_{jn})} W_n$	•Coefficient of similarity (complement)	$b+c$
•**Complete tree** (proposed here)	$\frac{\sum_n \lvert p_{in} - p_{jn} \rvert W_n}{\sum_n (\max_k(p_{kn}) - \min_k(p_{kn})) W_n}$	-	-
•**Euclidean** •Weighted Euclidean	$\sqrt{\sum_n W_n (p_{in} - p_{jn})^2}$	•Euclidean	$\sqrt{b+c}$
•F_{ST} •P_{ST}	$\frac{D_T - D_S}{D_T}$	-	-
•**Gower** (complement)	$\sum_n \frac{\lvert p_{in} - p_{jn} \rvert}{\max_k(p_{kn}) - \min_k(p_{kn})} W_n$	•Gower (complement)	$b+c$
•**Hellinger**	$\sqrt{\sum_n W_n \left(\sqrt{\frac{p_{in}}{L_i}} - \sqrt{\frac{p_{jn}}{L_j}} \right)^2}$	-	-
•**Kulczynski** (complement)	$1 - \frac{1}{2}\left(\frac{M}{W_{i+}} + \frac{M}{W_{j+}} \right)$ where, $M = \sum_n \min(p_{in}, p_{jn}) W_n$	•Kulczynski-Cody •Sokal-Sneath (complement)	$1 - \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c} \right)$
•Lennon compositional difference •Derived using Nipperess et al. (2010)	$\frac{\min(B,C)}{\sum_n \min(p_{in}, p_{jn}) W_n + \min(B,C)}$ where, $B = \sum_n (\max(p_{in}, p_{jn}) - p_{jn}) W_n$ $C = \sum_n (\max(p_{in}, p_{jn}) - p_{in}) W_n$	•Lennon compositional difference	$\frac{\min(b,c)}{a + \min(b,c)}$
•Manhattan •Weighted UniFrac	$\sum_n \lvert p_{in} - p_{jn} \rvert W_n$	•Hamming distance	$b+c$

Table 4.1. Phylogenetic beta-diversity measures (continued).

Quantitative		Qualitative	
•Mean nearest neighbour distance (MNND)	$\frac{\sum_{x\in X} p_{ix} \min_{y\in Y}(\delta_{xy}) + \sum_{y\in Y} p_{jy} \min_{x\in X}(\delta_{yx})}{2}$	•MNND	$\frac{\sum_{x\in X} \frac{\min_{y\in Y}(\delta_{xy})}{\lvert X\rvert} + \sum_{y\in Y} \frac{\min_{x\in X}(\delta_{yx})}{\lvert Y\rvert}}{2}$
•Mean phylogenetic distance (MPD) •Rao's D_p	$\frac{\sum_{x\in X}\sum_{y\in Y} p_{ix} p_{jy} \delta_{xy}}{\sum_{x\in X}\sum_{y\in Y} p_{ix} p_{jy}}$	•MPD	$\frac{\sum_{x\in X}\sum_{y\in Y} \delta_{xy}}{\lvert X\rvert \lvert Y\rvert}$
•**Morisita-Horn**	$\frac{2\sum_n p_{in} p_{jn} W_n}{\left(\frac{\sum_n p_{in}^2 W_n}{W_{i+}^2} + \frac{\sum_n p_{jn}^2 W_n}{W_{j+}^2}\right) W_{i+} W_{j+}}$	-	-
•Normalized weighted UniFrac	$\frac{\sum_n \lvert p_{in} - p_{jn}\rvert W_n}{\sum_{l\in L}(p_{il} + p_{jl}) R_l}$	-	-
•**Pearson dissimilarity**	$1 - \frac{\sum_n \left(p_{in} W_n - \frac{W_{i+}}{N}\right)\left(p_{jn} W_n - \frac{W_{j+}}{N}\right)}{\sqrt{\sum_n \left(p_{in} W_n - \frac{W_{i+}}{N}\right) \sum_n \left(p_{jn} W_n - \frac{W_{j+}}{N}\right)^2}}$	•Pearson dissimilarity	(see quantitative)
•Rao's H_p	$D_T - D_S$	-	-
•Soergel •Ružička (complement) •Percentage remoteness	$\frac{\sum_n \lvert p_{in} - p_{jn}\rvert W_n}{\sum_n \max(p_{in}, p_{jn}) W_n}$	•Jaccard (complement) •Unweighted UniFrac	$\frac{b+c}{a+b+c}$
•Tamàs coefficient	$\frac{\sum_n \lvert p_{in} - p_{jn}\rvert W_n}{\sum_n \max_k(p_{kn}) W_n}$	•Simple matching coefficient (complement)	$\frac{b+c}{a+b+c+d}$
•Unweighted UniFrac	-	•Unweighted UniFrac	See Jaccard
•**Weighted correlation** (complement)	$1 - \frac{\mathrm{cov}(p_i, p_j; W)}{\sqrt{\mathrm{cov}(p_i, p_i; W)\,\mathrm{cov}(p_j, p_j; W)}}$	•Weighted correlation (complement)	(see quantitative)
•Weighted UniFrac	See Manhattan	-	-
•**Whittaker index of association** (complement)	$\frac{1}{2}\sum_n \left\lvert \frac{p_{in}}{L_i} - \frac{p_{jn}}{L_j} \right\rvert W_n$	-	-
•**Yue-Clayton** (complement)	$1 - \frac{\sum_n p_{in} p_{jn} W_n}{\sum_n (p_{in} - p_{jn})^2 W_n + \sum_n p_{in} p_{jn} W_n}$	-	-

communities (Webb et al. 2008). We also consider the Pearson and weighted correlation dissimilarity measures as they are unique in assessing dissimilarity through correlation (Legendre and Legendre 1998). Although we focus primarily on these measures, many of the other measures are widely used (e.g., Morisita-Horn, Hellinger) and their inclusion helps provide context for the other measures. Complete results for all measures are provided in Appendix C which may be consulted by readers interested in a particular measure.

Performance measures are established over 2 distinct models of community differentiation which highlights critical aspects of the considered measures, the need to apply multiple measures, and shortcomings of the taxon-based measures previously recommended for assessing differences between microbial communities (Kuczynski et al. 2010). We contrast the performance of measures on complete and random subsets of 4 pyrosequencing datasets (Table 4.2): 1) fingertip and keyboard samples from 3 individual used for forensic identification (Fierer et al. 2010), 2) small and large intestinal samples taken from 4 groups of mice before and after antibiotic treatment (Ubeda et al. 2010), 3) samples collected from the navel, mouth, hair, and stool of 7 to 9 individuals (Costello et al. 2009), and 4) soil samples taken across a substantial pH gradient (Rousk et al. 2010). By considering random subsets of samples from these 4 distinct datasets we are able to evaluate measures over a range of tree topologies spanning samples with varying levels of inter- and intra-sample diversity.

Table 4.2. Details of empirical datasets.

Dataset	*Samples*	*Seqs/Sample (mean* ± s.d.)	*Study Design*	*Main Results*	*Reference*
Keyboard	89	1183±250	Samples taken from the fingertips and keyboards of 3 individuals	Samples from fingertips and keyboards clustered by individuals	Fierer et al., 2010
Mouse	40	901±394	Samples taken from the ileum and cecum of 4 groups of mice before and after treatment with antibiotics	Antibiotic treatment nearly completely displaced the normal microbiota of the small and large intestine	Ubeda et al., 2010
Human	76	1534±635	Samples taken from 27 body sites in 7 to 9 individuals on 4 occasions	Community composition was determined primarily by body habitat	Costello et al., 2009
Soil	22	1662±459	Samples collected from soil across a pH gradient ranging from 4.0 to 8.3	Relative abundance and diversity of bacteria were positively related to pH	Rousk et al., 2010

4.3 Methods

4.3.1 Empirical Datasets

The properties and effectiveness of phylogenetic beta-diversity measures were assessed using 4 empirical datasets (Table 4.2). Datasets were processed using a common pipeline, but with the dataset-specific filtering criteria specified in the original publications. Sequences were removed from the analysis if they were less than 200 bp or greater than a specific length (keyboard, soil: 300 bp; human, mouse gut: 400 bp), had a quality score less than 25, contained ambiguous characters, contained an unrecognized barcode, or did not contain the primer sequence. Sequences were aligned using the mothur v1.22.1 (Schloss et al. 2009) implementation of the NAST algorithm with the Greengenes reference alignment (DeSantis, Hugenholtz, Keller, et al. 2006). We removed sequences with an alignment length less than 150 or an identity with the reference alignment of less than 75% along with any samples containing an insufficient number of sequences (keyboard, human: 800; soil: 600; mouse gut: no filtering). Hypervariable columns of the alignment were removed using the PH Lane mask. Phylogenetic trees were inferred using FastTree v2.1.4 (Price et al. 2009) with a generalized time-reversible model. Trees were rooted with an outgroup of 3 archaeal sequences from distinct phyla. These 3 sequences formed a monophyletic group within each of the inferred trees, justifying their use as an outgroup to the sequences of interest.

4.3.2 Evaluating Properties of Phylogenetic Beta-diversity Measures

We evaluated 39 phylogenetic beta-diversity measures used within microbial and classical ecology along with newly established phylogenetic extensions of commonly used taxon-based measures (Table 4.1 and Methods C.1 in Appendix C). The properties of these measures were evaluated using 100 randomly selected subsets of 10 samples from each of the 4 empirical datasets. All sequences within a selected sample were used in all cases. Subsets of samples were considered in order to gauge the robustness of results in light of varying tree topologies, amounts of diversity spanned by a dataset, and patterns of phylogenetic similarity between samples. The hierarchical similarity of measures was determined by applying the UPGMA clustering algorithm to a matrix indicating the mean Pearson dissimilarity, $d=1-r$, between each pair of measures.

Correlation, *r*, was determined using Pearson's correlation coefficient as implemented in SciPy v0.9.0 (http://www.scipy.org). To evaluate the influence of sequence clustering, we clustered sequences using the furthest neighbour algorithm in mothur v1.22.1. Trees at different OTU thresholds were obtained by randomly selecting a representative sequence from each cluster and pruning the tree to the set of representative sequences. To evaluate the robustness of measures to root placement, phylogenies were randomly rerooted 100 times for each of the 100 subsets. Trees were rooted by randomly selecting a new node to be the root and using BioPython v1.58 (Cock et al. 2009) to reroot the tree. The addition of an outlying basal lineage was evaluated by appending a new lineage to the root of each dataset's phylogeny. This lineage consisted of a single branch whose length was set to the average distance from a leaf node to the root. Additional sequences were added to a sample and placed at the leaf node of the outlying lineage. Ordination plots indicating the similarity of samples were obtained using PCoA. UPGMA and PCoA results were obtained using software currently under development for the visualization and analysis of phylogenetic beta diversity.

4.3.3 Simulated Cluster Data

We simulated samples belonging to distinct clusters under 2 different models of differentiation, which we term the equal-perturbation and dominant-pair models. The equal-perturbation model extends the methodology proposed by Kuczynski et al. (2010) to the evaluation of phylogenetic beta-diversity measures. This model simulates microbial communities where a process stochastically influences the abundance of each OTU by an amount dependent on the initial abundance of that OTU. For each of the empirical datasets, we randomly selected a seed sample. We then perturbed this seed sample by multiplying the relative abundance of each OTU by a random number drawn from a normal distribution with unit mean and standard deviation $\sigma_1 = 1.0$. This was repeated 3 times in order to create starting distributions for 3 distinct clusters. These starting distributions were then renormalized to sum to 1.0. We generated 30 samples within each cluster by perturbing these 3 starting distributions, using a random number drawn from a normal distribution with unit mean and standard deviation $\sigma_2 = 0.5$. The simulated samples were then renormalized and sequence counts obtained by drawing,

with replacement, a specified number of sequences from each of these sample distributions. The values of σ_1 and σ_2 were set to 1.0 and 0.5, respectively, in order to approximate the clustering pattern of the keyboard dataset (Kuczynski et al. 2010).

The dominant-pair model simulates microbial communities where a process primarily influences the abundance of the 2 most abundant OTUs and only has a small stochastic effect on the remaining OTUs. This model is a simplified version of the shift seen in many communities where one predominant OTU is replaced with another. For example, in enhanced biological phosphorus removal (EBPR) communities, the primary strains of phosphate-accumulating *Candidatus* "Accumulibacter phosphatis" can decrease dramatically, with a concomitant increase in the frequency of other organisms such as *Candidatus* "Competibacter", possibly due to viral predation (Barr et al., 2010; Slater et al., 2010). To simulate this scenario, we initially perturb only the 2 most abundant OTUs in the seed sample in order to create the starting distributions for each cluster. The 2 most abundant OTUs were perturbed by an amount $\delta = d \cdot x - x$, where x is the abundance of the most abundant OTU in the seed sample and d is the central absolute moment of a normal distribution with $\sigma = 1.0$ (i.e., 1.797). The 3 starting distributions were created by modifying the 2 most abundant OTUs in the seed sample by $(+\delta,-\delta)$, $(-\delta,+\delta)$, and $(0,0)$. Samples within each cluster were then obtained as before. Since δ is the average expected change of an OTU under the equal-perturbation model, differences in the performance of a measure are expected to be the result of the models themselves and not an artifact of the relative distinctiveness of clusters. Randomizations under both models were repeated using 100 different seed samples from each empirical dataset.

4.3.4 Evaluation of Measures on Simulated Cluster Data

The ability of measures to recover simulated patterns of clustering was evaluated using 2 statistics. For the first statistic, we clustered samples with the *k*-medoids algorithm implemented in BioPython v1.58 and calculated the fraction of samples whose *k*-medoids clustering matched the known clustering of the samples. For the second statistic we calculated the UPGMA clustering of the simulated samples and determined the consistency index (Kluge and Farris 1969) of this hierarchical cluster tree. The consistency index is calculated by assigning each node in the UPGMA tree a state and

determining the number of state changes required to recover a particular distribution of states assigned to the leaf nodes. In this case, each leaf node is associated with a sample and the state is an identifier indicating the known cluster of the sample. If the samples come from *N* clusters, then *at least N*-1 state changes will be required to explain the distribution of samples. However, more than *N*-1 state changes will be required if the clustering in the UPGMA tree does not perfectly reflect the known clustering of the samples. The consistency index is the minimum number of required state changes (i.e. *N*-1) divided by the number of state changes required to explain the observed distribution of cluster identifiers in the UPGMA tree. For both statistics, perfect clustering gives a score of 1.

When sorting and summarizing the performance of measures we focused on results obtained with moderate sequence depth (1,000 sequences/sample) and the *k*-medoids statistic as it operates more directly on the dissimilarity matrix compared to the consistency index, which is calculated on the inferred UPGMA tree. Despite their differences, the 2 statistics were found to be highly correlated. Note that these measures differ from those previously proposed by Kuczynski et al. (2010) as we apply them directly to a measure's dissimilarity matrix as opposed to the distance between samples within an ordination plot which represents only one of many possible visualizations of a dissimilarity matrix.

4.3.5 Classifying measures by the branches they consider

Phylogenetic beta-diversity measures can be classified according to the set of branches that influence the calculation of community dissimilarity (Fig. 4.1). A measure only influenced by branches within the most recent common ancestor (MRCA) subtree spanned by a pair of samples is classified as an MRCA measure. In contrast, a measure also influenced by the "deep branches" which extend from the root of the MRCA subtree to the root of the tree spanning all samples is termed a complete lineage (CL) measure, and a measure influenced by all branches in the tree inferred from all data under consideration is termed a complete tree (CT) measure.

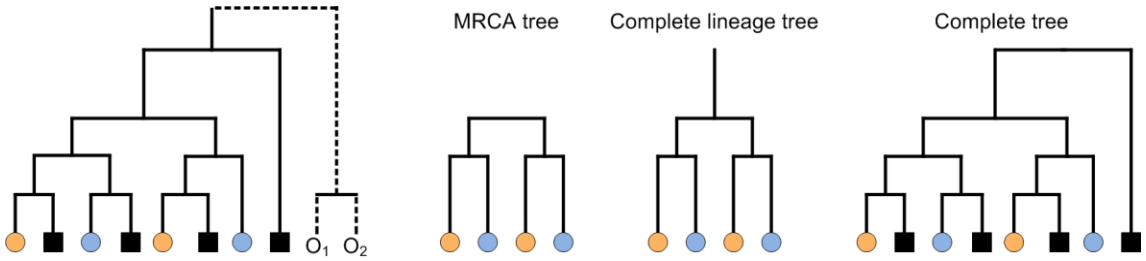


Figure 4.1 Phylogenetic measures can be classified as a most recent common ancestor (MRCA), complete lineage (CL), or complete tree (CT) measure based on the set of branches that influence the calculation of community dissimilarity. In this example, sequences have been collected from 3 communities shown as blue circles, orange circles, and black squares. A phylogenetic tree is inferred from all sequences. The MRCA subtree, CL subtree, and CT are shown for the two communities depicted by circles.

4.3.6 Software Availability and Verification

Express Beta Diversity is free and open-source software which implements the evaluated phylogenetic beta-diversity measures. Source code and executable binaries are available at http://kiwi.cs.dal.ca/Software/ExpressBetaDiversity. The software is designed to handle large datasets and provides functionality for clustering measures based on a user specified correlation threshold. Results of the MPD, MNND, and Rao's H_p measures along with their qualitative counterparts were verified against Phylocom v4.2 (Webb et al. 2008). The normalized weighted UniFrac (Bray-Curtis), weighted UniFrac (Manhattan), and unweighted UniFrac (qualitative Soergel) results were compared against the Fast UniFrac Web application (Hamady et al. 2010). Other measures were verified with a set of examples where ground-truth answers could be determined by hand calculation.

4.4 Results

4.4.1 Identifying Complementary Measures

Weakly correlated measures can provide complementary insights into the phylogenetic similarity of microbial communities. We assessed the degree of correlation between measures using 100 randomly selected subsets of 10 samples from each of the 4 empirical datasets. To explore the similarity of measures we then calculated statistics

over these trials and visualized the mean correlation between measures using hierarchical cluster trees (Fig. 4.2; see Fig. C.1 in Appendix C for dataset-specific results). This revealed a number of highly and perfectly correlated measures (Table C.2 in Appendix C). Notably, unweighted UniFrac (uSoergel) and PhyloSor (uBray-Curtis) are highly correlated (Pearson's r = 1.00±1.1·10^{-3} s.d.) whereas normalized weighted UniFrac is identical to the Bray-Curtis measure (Appendix D). Corresponding quantitative and qualitative measures were found to be only moderately correlated (Pearson's r = 0.66±0.16 s.d.) with the notable exception of the MNND, Canberra, Gower, and coefficient of similarity measures. Measures commonly used within microbial ecology (e.g., UniFrac variants) were found to be distinct from those popular for studying macroorganisms (i.e., MNND, MPD, Rao's H_p, F_{ST}), and from the Gower and Canberra measures recently recommended for assessing microbial community relationships from species profiles (Kuczynski et al. 2010).

4.4.2 Robustness to Sequence Clustering

In order to reduce computational requirements, similar sequences can be clustered and a single representative sequence from each cluster used during sequence alignment and phylogenetic inference. A common sequence similarity threshold for clustering full length 16S rRNA gene sequences is 97%, which roughly corresponds to the working definition of a microbial species (Stackebrandt and Goebel 1994), but OTUs may be usefully defined over a wide range of similarity thresholds, e.g., 80-99% (Schloss and Handelsman 2004; Bryant et al. 2008). To evaluate the influence of sequence clustering on phylogenetic beta-diversity measures, we assessed the degree of correlation between measures before and after clustering at various levels of sequence similarity. As above, we considered results for 100 randomly selected subsets of 10 samples from each of the 4 datasets.

Clustering sequences results in a substantial reduction in branch length and number of leaf nodes within a phylogeny (Fig. 4.3). At 97% sequence similarity, branch length was reduced by between 8% (soil study) and 44% (mouse gut study) suggesting that the inferred phylogenies for these datasets differ substantially. In particular, the mouse gut

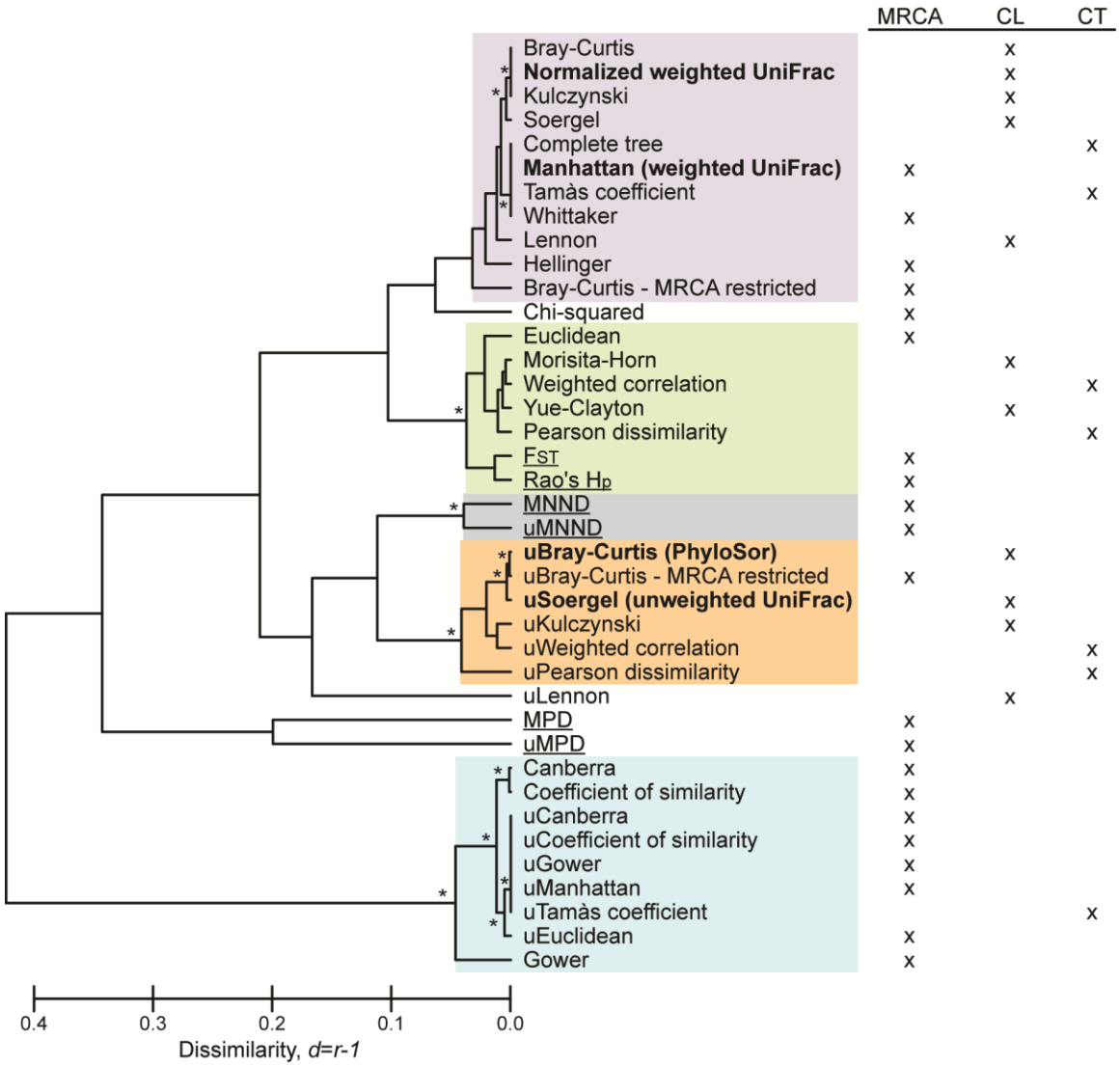


Figure 4.2. Similarity of phylogenetic beta-diversity measures. Branch lengths are transformed Pearson's *r* values, *d=r-1*, averaged over 100 random subsets of samples drawn from 4 empirical datasets. The hierarchical relationship between measures was obtained using the UPGMA clustering algorithm. Branches supported by at least 70% of the trials are indicated with asterisks. The five most highly correlated and consistently clustered groups of measures are highlighted in different colours. These clusterings are nearly perfectly recovered on all four datasets (Fig. C.1 in Appendix C). Phylogenetic measures commonly used within microbial ecology are shown in bold and measures popular in classical ecology are underlined. Measures are specified by their common quantitative name and qualitative counterparts indicated by prefixing a "u" for unweighted. Each measure is classified as a most recent common ancestor (MRCA), complete lineage (CL), or complete tree (CT) measure.

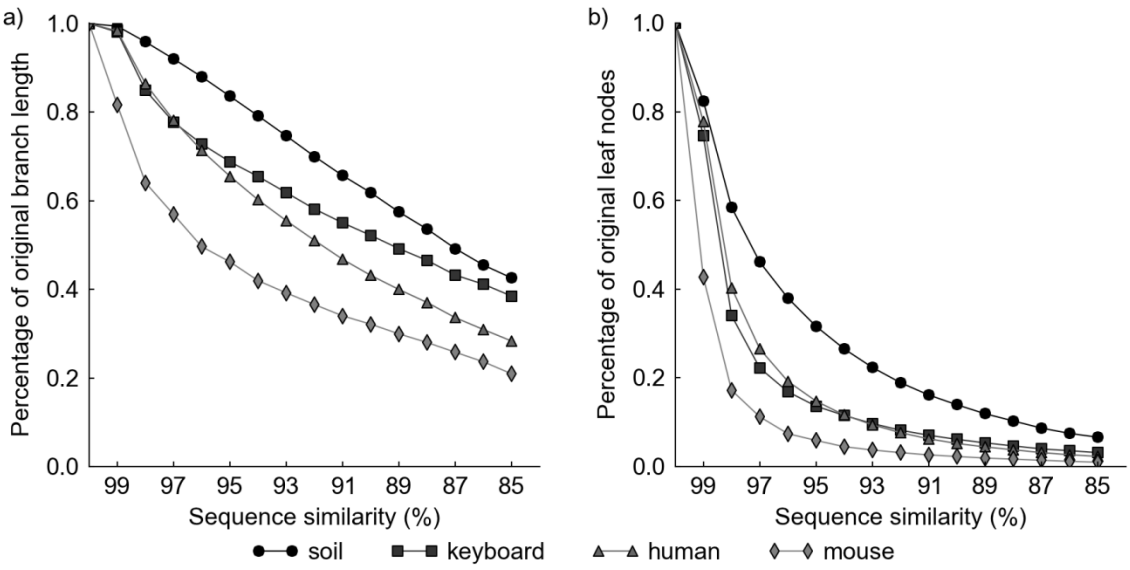


Figure 4.3. Influence of sequence clustering on 4 empirical phylogenies. (**a**, **b**) Percentage of retained branch length (a) and leaf nodes (b) as the sequence similarity threshold used to define clusters is relaxed.

phylogeny has more branch length associated with highly related sequences than the other three phylogenies considered. Nonetheless, the majority of quantitative phylogenetic beta-diversity measures exhibited only a slight decrease in correlation with the dissimilarity results obtained before clustering (Fig. 4.4a). Even at 85% sequence similarity, all measures remained highly correlated (Pearson's $r > 0.92$ for all trials) except for the Canberra, coefficient of similarity, Gower, and Pearson dissimilarity measures (Fig. 4.4b and Table C.3 in Appendix C). Qualitative measures were more sensitive to sequence clustering and varied more substantially between datasets (Fig. 4.4c). Although all qualitative measures were more sensitive to sequence clustering than their quantitative counterparts, the uMPD measure exhibited extreme sensitivity to the extent of being negatively correlated with the unclustered results for certain trials (Fig. 4.4d and Table C.4 in Appendix C). Although the majority of measures revealed the same biological patterns between microbial samples even at 85% sequence similarity (Figs. 4.4e and 4.4f), measures sensitive to clustering can fail to recover the same patterns (Figs. 4.4g and 4.4h). The observed robustness to the choice of OTU clustering threshold is a

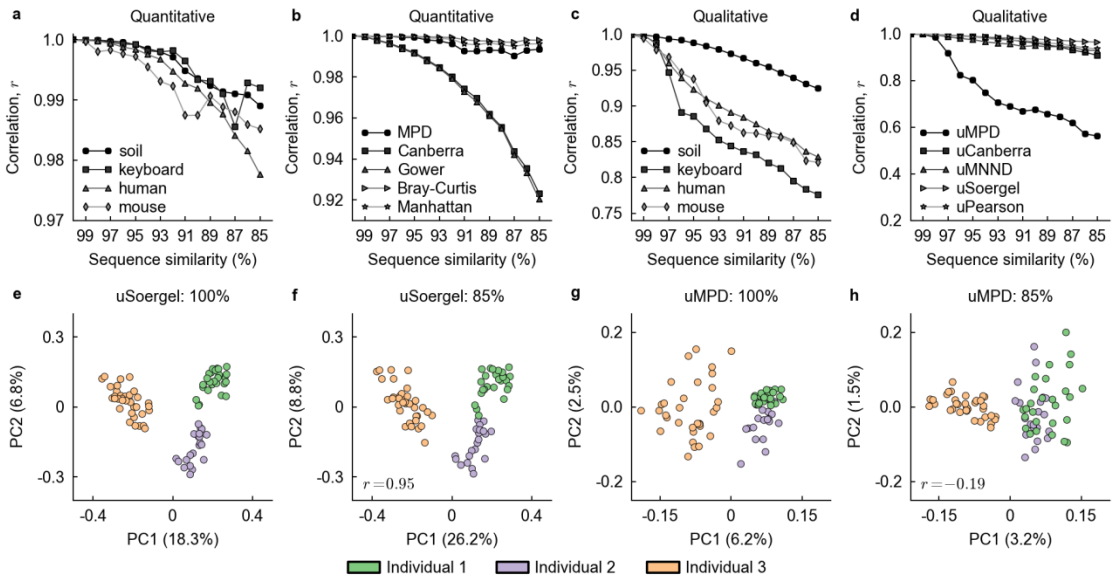


Figure 4.4. Influence of sequence clustering on phylogenetic beta-diversity measures. (**a**, **c**) Mean correlation across all quantitative (a) and qualitative (c) measures on subsets of samples from each empirical dataset. (**b**, **d**) Correlation of select quantitative (b) and qualitative (d) measures averaged over all 4 empirical datasets. (**e**, **f**) Ordination plots obtained by applying the qualitative Soergel measure to the keyboard dataset with sequences clustered at 100% (e) and 85% (f) sequence similarity. (**g**, **h**) Ordination plots for the qualitative MPD measure with sequences clustered at 100% (g) and 85% (h) sequence similarity. PCoA was used to generate the ordination plots. The percentage of total variance explained by each axis is shown in parentheses. Each data point represents a sample taken from one of the 3 individuals. Pearson's correlation coefficient, *r*, between dissimilarity values measured before and after clustering is given in the bottom-left corner of each plot.

positive attribute of these measures as it indicates computational requirements can be reduced by clustering sequences and that this will not substantially affect the measured dissimilarity between communities. However, these results also demonstrate that these measures lack sensitivity to fine-scale differences in community structure. In the remainder of this chapter we consider OTUs formed at 97% sequence similarity as this is the most commonly used clustering criterion in microbial ecology and all measures (with the exception of uMPD) were found to be highly correlated (Pearson's $r > 0.93$ for all trials) with their original dissimilarity values at this clustering threshold.

4.4.3 Robustness to Outlying Lineages

Outlying lineages may occur due to errors in sequence alignment or phylogenetic inference, or simply due to the stochastic detection of rare outlying taxonomic groups. To evaluate the robustness of measures to the addition of an outlying basal lineage, we added a single branch to the root node of each dataset's phylogeny and set the length of this branch to the average distance from each leaf node to the root. This simulates an outlying lineage such as a deeply branching phylum or superkingdom not always observed in the community. We then generated random subsets of samples from each of the empirical datasets and randomly selected half of these samples to contain sequences within the outlying lineage. Even with 5% of sequences assigned to the outlying lineage, nearly all quantitative measures were well correlated (Pearson's $r > 0.81$ for every trial except MNND, $r > 0.57$, and Pearson dissimilarity, $r > 0.67$) with their original dissimilarity values (Fig. 4.5 and Table C.5 in Appendix C). Despite the all-or-nothing nature of qualitative measures, they were robust (Pearson's $r > 0.88$ for every trial) to the addition of an outlying basal lineage (Fig. 4.5 and Table C.6 in Appendix C). The sole exception is the uPearson dissimilarity measure, which was highly sensitive (mean Pearson's $r = 0.42$) to the inclusion of an outlying lineage and failed to recover the expected relationship between samples on both the human (Fig. 4.5) and soil (Fig. C.2 in Appendix C) datasets.

In general, measures are robust to a moderate percentage of sequences being assigned to an outlying basal lineage. Although the length of the outlying basal branch is long compared to other branches in the phylogeny, it represents only a small portion of the total branch length and as a result does not substantially influence the calculated dissimilarity between samples under most measures. Most measures will be robust to any perturbation of the phylogeny influencing only a small portion of the total branch length.

Exceptions are the MNND, Pearson and uPearson measures which showed severe sensitivity to sequences being assigned to an outlying lineage. The degree of sensitivity is highly dependent on the length of the outlying basal branch (data not shown). The weighted correlation measure can be used instead of the Pearson measure when a correlation-based measure of dissimilarity is desired for either quantitative or qualitative data.

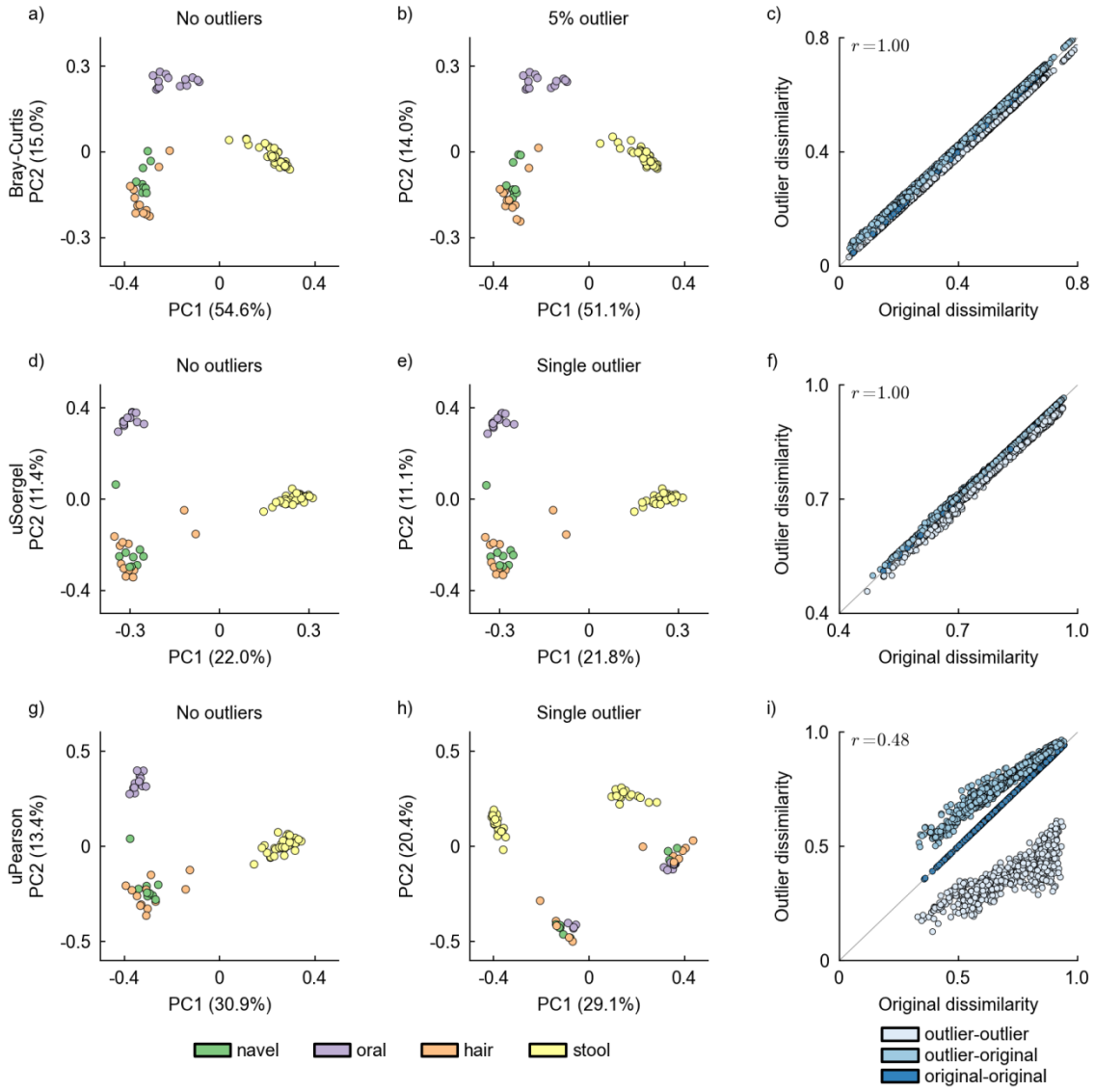


Figure 4.5. Recovery of clusters is influenced by a measure's robustness to outlying basal lineages. (**a**-**i**) The quantitative Bray-Curtis (a-c), qualitative Soergel (d-f), and qualitative Pearson dissimilarity (g-i) measures were applied to the human dataset. (**a**, **d**, **g**) All 3 methods revealed 3 clusters: a stool cluster, an oral cluster, and a mixed navel and hair cluster. (**b**, **e, h**) The addition of an outlying basal lineage to half the samples did not substantially affect the Bray-Curtis (b: 5% of sequences assigned to the outlying lineage) or uSoergel (e) measures, but obscured the underlying biological clusters for the uPearson dissimilarity (h) measure. (**c**, **f**, **i**) Each data point in the scatter plots indicates the dissimilarity measured between a pair of samples before (x-axis) and after (y-axis) adding sequences to the outlying lineage. Pearson's correlation coefficient, *r*, between dissimilarity values measured before and after addition of the outlying lineage is given in the upper-left corner of each scatter plot.

4.4.4 Robustness to Root Placement

In the absence of a credible outgroup for rooting a phylogeny, it is beneficial to have measures which can be applied to unrooted trees. Several of the evaluated measures are invariant to root placement. Since the distribution of sequences across leaf nodes and the phylogenetic distance between leaf nodes are invariant to where a tree is rooted, it follows that the quantitative and qualitative F_{ST}, MNND, MPD, and Rao's H_p measures are root invariant. Other root invariant quantitative measures can be proven as follows. Every branch within a tree induces a bipartition on the set of taxa within a tree. If the proportion of taxa in one set induced by a branch is p_{in} and p_{jn}, then the proportion of taxa in the other set is $1 - p_{in}$ and $1 - p_{jn}$. As such, the following terms produce the same measure of dissimilarity regardless of where a tree is rooted:

- $\left|(1-p_{in})-(1-p_{jn})\right| = \left|p_{jn}-p_{in}\right| = \left|p_{in}-p_{jn}\right|$
- $\left((1-p_{in})-(1-p_{jn})\right)^2 = \left(p_{jn}-p_{in}\right)^2 = \left(p_{in}-p_{jn}\right)^2$

Furthermore, the term $\max_k(p_{kn}) - \min_k(p_{kn})$ is also invariant to root placement. If $\max_k(p_{kn}) = x$ and $\min_k(p_{kn}) = y$ for one of the sets induced by branch *n*, then $\max_k(p_{kn}) = 1 - y$ and $\min_k(p_{kn}) = 1 - x$ in the other set. Since $x - y = (1-y) - (1-x)$, this term is also root invariant. The complete tree, Euclidean, Gower, Manhattan, and Whittaker measures are composed of only these terms or simple normalizations of these terms and are therefore root invariant (Fig. 4.6).

We evaluated the robustness of the remaining measures to root placement by considering dissimilarity values obtained on random subsets of samples before and after randomly rerooting their corresponding phylogeny (Tables C.7 and C.8 in Appendix C). The quantitative Canberra and coefficient of similarity measures were found to be highly robust to root placement (Pearson's $r > 0.99$ for all trials). The remaining quantitative measures showed sensitivity to at least some random root placements, including the Bray-Curtis measure, i.e., normalized weighted UniFrac (minimum Pearson's $r = 0.30$) measure. All qualitative measures were found to be robust to root placement (Pearson's *r*

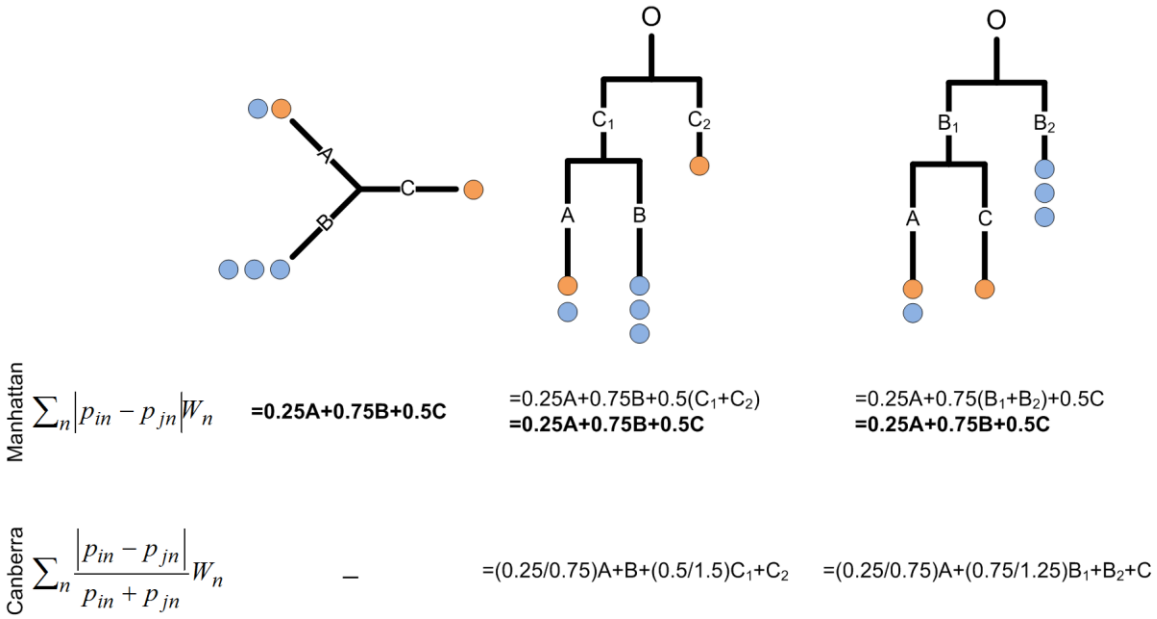


Figure 4.6. An example of root invariant and root dependent measures. In this example, sequences have been collected from 2 communities shown as blue and orange circles. The phylogeny for these sequences is shown as both an unrooted tree and with a root at the mid-point of 2 of the branches. Since root invariant measures, such as the Manhattan measure, produce the same measure of dissimilarity between samples regardless of where a tree is rooted they can be applied to unrooted trees. Measures sensitive to where a tree is rooted, such as the Canberra measure, require a tree to be rooted before they can be calculated.

> 0.92 for all trials) with the exception of the uPearson dissimilarity measure (minimum Pearson's $r = 0.73$). Most measures are robust to small changes in root placement as only branches along the path from the original root to the new root will differ in their contribution to the dissimilarity measured between a pair of samples (Fig. 4.7).

4.4.5 Robustness to Rare OTUs

Measures vary in their treatment of rare OTUs. We assessed a measure's robustness to rare OTUs using randomly selected subsets of samples from each dataset. For each trial, we determined the correlation between dissimilarity values obtained before and after filtering OTUs containing only a single sequence, less than 0.1% of sequences, or less than 1% of sequences. All quantitative measures were relatively insensitive to the removal of lineages containing only rare OTUs (Pearson's $r > 0.85$ for all trials at 0.1%

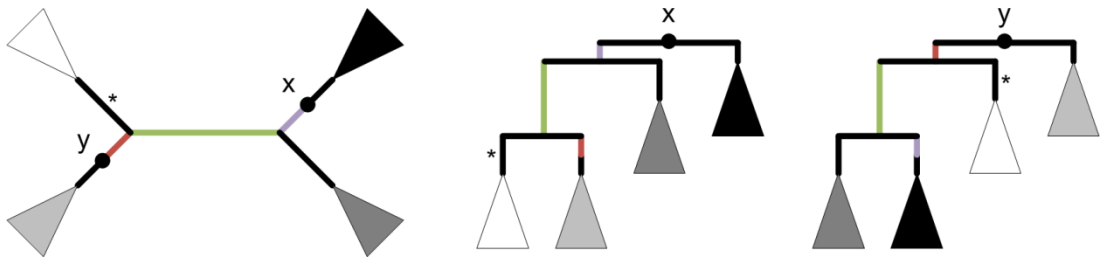


Figure 4.7. An example illustrating the influence of root placement on dissimilarity measures. The unrooted tree is rooted at either position *x* or position *y*. Only branches along the path from position *x* to position y (i.e., the coloured branches) will have their descendants changed when moving the root between these 2 positions. All other branches have the same descendants for both root placements (one of these branches has been marked with an asterisk for easy identification in all trees). For large phylogenies, the descendants of most branches will not be affected by rerooting the tree and as a consequence many dissimilarity measures are generally robust to changes in root placement.

filtering) with the exception of the Canberra (minimum Pearson's $r = 0.17$), coefficient of similarity (minimum Pearson's $r = 0.08$), and Gower (minimum Pearson's $r = 0.07$) measures which were highly sensitive (Table C.9 in Appendix C). In contrast, all qualitative measures can be substantially affected by the removal of rare OTUs as they are sensitive to the removal of lineages irrespective of the number of sequences assigned to a lineage (Table C.10 in Appendix C).

4.4.6 Revealing Clusters of Samples

We assessed the ability of measures to identify discrete clusters of samples under 2 models of phylogenetic differentiation. Under the equal-perturbation model, an initial seed sample was selected from one of the 4 datasets and the relative abundance of *each* OTU perturbed by a random percentage in order to create 3 starting distributions. We then applied a relatively small perturbation to these starting distributions in order to generate 3 clusters consisting of 30 distinct samples. Model parameters were selected to mimic the clustering pattern of the keyboard dataset. Under the dominant-pair model, the initial perturbation of the seed sample was restricted to the 2 most abundant OTUs. This was followed by a more subtle stochastic process applied to all OTUs in order to again generate 3 clusters of 30 distinct samples. Clusters were simulated under both models for 100 randomly selected seed samples from each dataset and at varying sequencing depths.

The relative effectiveness of measures was dependent on the simulated model (Table C.11 in Appendix C). Measures sensitive to lineages containing rare OTUs such as Canberra and Gower performed strongly under the equal-perturbation model, but failed to identify clustering patterns under the dominant-pair model (Fig. 4.8 and Tables C.12-C.19 in Appendix C). The most effective measures under the dominant-pair model were those highly sensitive to the most abundant OTUs such as Morisita-Horn and Euclidean (Magurran 2004). Measures that are relatively insensitive to rare OTUs while not being overly sensitive to the most abundant OTUs, such as the Bray-Curtis, Soergel and Manhattan measures, performed moderately well under both models (Table C.20 in Appendix C). Although the performance of each measure depended on the empirical dataset from which the initial seed sample was drawn, the relative performance of the measures was remarkably stable under both models (Tables C.21 and C.22 in Appendix C). Consequently, we are confident these results are not an artifact of the phylogenetic structure or diversity of a particular dataset. We also found the hierarchical clustering of measures under these two models of differentiation to largely resemble those obtained on the empirical datasets (Fig. C.3 in Appendix C).

4.5 Discussion

Phylogenetic measures of beta diversity can be classified as MRCA, CL, or CT based on the set of branches which influence the dissimilarity calculated between a pair of communities. Measures operating over different sets of branches can be highly correlated (Fig. 4.2) whereas measures operating over the same set of branches can differ substantially in their properties and effectiveness at revealing patterns of clustering (Table 4.3). Although using the shared absence of species has been criticized in classical ecology as being uninformative to the ecological similarity of sites (Legendre and Legendre 1998), we found CT measures to perform relatively well under both models of differentiation considered (Table C.20 in Appendix C) and we contend that with sufficiently deep sampling shared lineage absence is informative. Recently, the use of deep branches for conservation assessment has been debated (Crozier et al. 2005; Faith and Baker 2006). Our results indicate that MRCA, CL, and CT measures can all efficiently recover biologically informative patterns. Interestingly, implementations of the

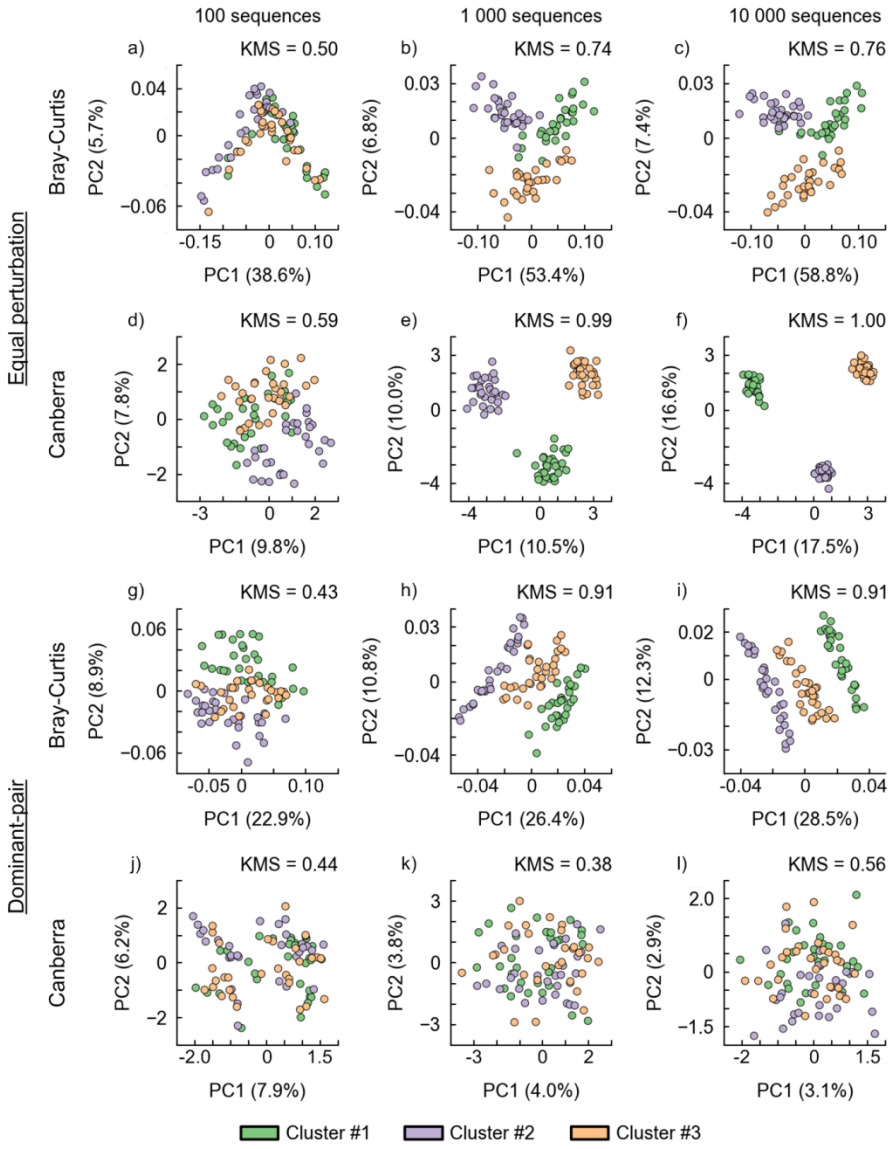


Figure 4.8. Effectiveness of measures depends on the mechanism of phylogenetic differentiation and sequencing depth. (**a**-**f**) The Bray-Curtis (a-c) and Canberra (d-f) measures were applied to clusters obtained under the equal-perturbation model at sequencing depths of 100, 1,000, or 10,000 sequences per sample. (**g**-**l**) These measures were also applied to clusters generated under the dominant-pair model. The *k*-medoids score (KMS) is given in the upper-right corner of each ordination plot.

Table 4.3. Properties of phylogenetic beta-diversity measures.

Quantitative Measures

Contributing branches	*Root invariant*	*Sensitive to rare OTUs*	*Robust to outlying lineages*	*Highly effective on equal-perturbation model*	*Highly effective on dominant-pair model*	*Measure(s)*
MRCA	**Yes**	**Yes**	**Yes**	**Yes**	No	Gower
MRCA	**Yes**	No	**Yes**	No	No	Manhattan, MPD, Whittaker,
MRCA	**Yes**	No	**Yes**	No	**Yes**	Euclidean, Rao's H_p, F_{ST}
MRCA	**Yes**	No	**Yes**	**Yes**	No	MNND
CT	**Yes**	No	**Yes**	No	No	Complete tree, Tamàs coefficient
MRCA	No	**Yes**	**Yes**	**Yes**	No	Canberra, Coefficient of similarity
MRCA	No	No	**Yes**	**Yes**	No	Hellinger
MRCA	No	No	**Yes**	No	No	Bray-Curtis (MRCA restricted)
CT	No	No	**Yes**	No	No	Pearson dissimilarity
CT	No	No	**Yes**	No	**Yes**	Weighted correlation
CL	No	No	**Yes**	No	No	Bray-Curtis, Kulczynski, Lennon, Soergel
MRCA	No	No	**Yes**	**Yes**	No	Chi-squared
CL	No	No	**Yes**	No	**Yes**	Yue-Clayton, Morisita-Horn

Qualitative Measures

Contributing branches	*Root invariant*	*Sensitive to rare OTUs*	*Robust to outlying lineages*	*Highly effective on equal-perturbation model*	*Highly effective on dominant-pair model*	*Measure(s)*
MRCA	**Yes**	**Yes**	**Yes**	No	No	uMNND, uMPD
MRCA	No	**Yes**	**Yes**	No	No	uCanberra, uEuclidean, uGower, uManhattan, uCoefficient of similarity, uBray-Curtis (MRCA restricted)
CT	No	**Yes**	**Yes**	No	No	uTamàs coefficient, uWeighted correlation
CT	No	**Yes**	No	No	No	uPearson dissimilarity
CL	No	**Yes**	**Yes**	No	No	uBray-Curtis, uKulczynski, uLennon, uSoergel

CL = complete lineage; MRCA = most recent common ancestor; CT = complete tree. A measure was deemed sensitive to rare OTUs if the minimum Pearson's correlation coefficient was less than 0.8 on any subset at 0.1% filtering. A measure was considered robust to outlying lineages if the minimum Pearson's correlation coefficient on any subset of samples was greater than 0.8 when 5% of sequences were assigned to the outlying lineage. A measure was considered highly effective at identifying the underlying clustering pattern for a given model of differentiation only if it was within 10% of the top performing measure on all 4 empirical datasets.

normalized weighted UniFrac (Bray-Curtis) measure have differed in their inclusion of deep branches, e.g., the Fast UniFrac web services (Lozupone et al. 2006; Hamady et al. 2010) calculated diversity over the CL subtree whereas mothur (Schloss et al. 2009) considers the MRCA subtree by default. Restricting the Bray-Curtis measures to the MRCA subtree can have a notable influence on the dissimilarity measured between communities (Table C.2 in Appendix C), highlighting the importance of explicitly specifying the set of branches a measure is calculated over.

The evaluated phylogenetic beta-diversity measures differed in their properties and ability to reveal clustering patterns under alternative models of differentiation (Table 4.3). For example, the Canberra and Gower measures easily identified clusters under the equal-perturbation model with only 1,000 sequences per sample whereas the Morisita-Horn measure generally failed to reveal clusters even with 10,000 sequences per sample. In contrast, Morisita-Horn readily identifies clusters under the dominant-pair model where the Canberra and Gower measures proved ineffective. The performance of a measure on samples which have differentiated according to a particular model can often be inferred from its properties. Five of the 6 most effective measures on the equal-perturbation model are either sensitive to rare OTUs (Canberra, Coefficient of similarity, Gower) or downweight the contribution of abundant OTUs (Hellinger, Chi-squared). Since these are properties of the measures themselves, it is unsurprising that our results on phylogenetic-based measures are in general agreement with those obtained for taxon-based measures under this model (Kuczynski et al. 2010). These results illustrate the need to consider the performance of a measure under multiple models of differentiation, and indicate that the Canberra and Gower measures can perform poorly under some models of community variation and must be interpreted with regards to their high sensitivity to rare OTUs. These results likely apply to the taxon-based variants of the Canberra and Gower measures recently recommended by Kuczynski *et al*. (2010) as high sensitivity to rare OTUs is an inherent property of these measures.

Our results suggest that the depth of sequencing required to reveal clusters depends not only on the selected measure, but also the prominence of the underlying clusters. On subsets from the mouse gut dataset, dominant-pair clustering could be readily identified by effective measures such as the Morisita-Horn and Euclidean with only 100 sequences

per sample, but generally required 1,000 sequences per sample on the other datasets. This suggests that even measures suited to the underlying mechanism of differentiation may require deep sequencing to reveal subtle patterns. Although the performance of quantitative measures generally increased with sequencing depth, the performance of qualitative measures often decreases when increasing from 1,000 to 10,000 sequences per sample (Table C.20 in Appendix C). Deeper sampling results in increased detection of rare OTUs causing samples from distinct clusters to share additional lineages. Although, to an extent, this is a result of not explicitly modeling lineage loss between clusters, it highlights the sensitivity of qualitative measures to sampling depth and rare OTUs, and emphasizes the benefits of applying both qualitative and quantitative measures.

The variation in the performance of measures under alternative models of differentiation is the direct result of measures focusing on different aspects of phylogenetic relatedness. Measures may produce contrasting biological patterns indicating the relative importance of factors such as rare OTUs, root placement, or abundance information. As such, complementary information on the phylogenetic similarity of communities may be obtained by applying several measures. For example, when applied to cecal microbiota from lean and obese mice, the unweighted UniFrac (uSoergel) measure identified high similarity between the microbiota of mothers and their offspring, whereas the weighted UniFrac (Manhattan) measure indicated that community composition was associated with obesity genotype (Lozupone et al. 2007). Our observed mean correlation between these measures was 0.83, suggesting that even a relatively high correlation between two measures does not necessarily preclude the recovery of contrasting results in parallel ordination analyses.

A number of measures were found to be highly correlated under both random sampling of empirical datasets (Fig. 4.2 and Fig. C.1 in Appendix C) and the evaluated models of differentiation (Fig. C.3 in Appendix C). This suggests that these measures will be highly correlated for many datasets. Here we recommend specific measures based on this clustering. The blue, purple, and green clusters appear to be driven by the sensitivity of measures to rare or abundant OTUs. We recommend the Gower, Soergel, and Morisita-Horn measures as representative measures as their taxon-based variants are well studied and widely used (Legendre and Legendre 1998; Magurran 2004). The Gower

measure is sensitive to rare OTUs, the Soergel measure takes a more balanced approach, and the Morisita-Horn measure places additional emphasis on highly abundant OTUs. If root-invariant measures are required the Manhattan and Euclidean measures may be preferred to Soergel and Morisita-Horn, respectively. All measures within the orange cluster are qualitative and we recommend the uSoergel measure (i.e., Jaccard index) as its taxon-based variant is well studied and widely used (Koleff et al. 2003, Magurran 2004), as is its phylogenetic variant under the guise of unweighted UniFrac (Lozupone and Knight, 2005; Lozupone *et al.* 2007). Notably, the Soergel measure is equivalent to the Jaccard index when applied to qualitative data (Pielou 1984), motivating its use over other measures in the purple cluster. Both the root invariant MNND and uMNND (grey cluster) often produce only weakly correlated results compared to other measures and are of interest due to their wide use in classical ecology (Webb *et al.* 2008). The Chi-squared, MPD, uMPD, and uLennon measures tended to produce rather distinct results from all other measures so may warrant consideration by studies conducting a thorough analysis of beta diversity. Although the above clustering of measures was highly similar across all four empirical datasets, and under the two models of variation considered, it may differ for specific datasets, especially those considering alternative genes or specific lineages within a 16S rRNA gene phylogeny. As such, our Express Beta Diversity software provides functionality for identifying dataset-specific subsets of measures within a given correlation threshold and inferring *de novo* hierarchical cluster trees based on the dataset-specific correlation between measures.

We have explored a number of important properties of phylogenetic beta-diversity measures and their performance under 2 models of differentiation, with a focus on the correlation between measures. Additional work is required to assess how the magnitude of dissimilarity values change under different conditions. For example, although the dissimilarity values of most measures remain highly correlated with the addition of an outlying lineage, we observed that the magnitude of dissimilarity values changed more substantially for measures sensitive to rare OTUs. Further efforts to relate the performance of measures to different mechanisms of differentiation would also be of substantial benefit. We would especially welcome efforts to model communities along environmental gradients or models illustrating the effect of selective lineage loss.

4.6 Acknowledgements

DHP is supported by the Killam Trusts; RGB acknowledges the support of Genome Atlantic, the Canada Foundation for Innovation, and the Canada Research Chairs program.

Chapter 5

Measuring Community Similarity with Phylogenetic Networks

Parks DH, Beiko RG. 2012b. Measuring community similarity with phylogenetic networks.

Publication status: In press at Molecular Biology and Evolution (July, 2012).

Contribution to research: **DHP** conceived of and carried out the research. RGB provided guidance throughout the project.

Contribution to writing: Written by **DHP** with suggestions and editorial advice provided by RGB.

5.1 Abstract

Environmental drivers of biodiversity can be identified by relating patterns of community similarity to ecological factors. Community variation has traditionally been assessed by considering changes in species composition and more recently by incorporating phylogenetic information in order to account for the relative similarity of taxa. Here we describe how an important class of measures including Bray-Curtis, Canberra, and UniFrac can be extended to allow community variation to be computed on a phylogenetic network. We focus on *phylogenetic split systems*, networks that are produced by the widely used median network and neighbour-net methods, which can represent incongruence in the evolutionary history of a set of taxa. Calculating community similarity over a split system provides a measure which is averaged over uncertainty or conflict in the available phylogenetic signal. Our freely available software, Network Diversity, provides 11 qualitative (presence-absence, unweighted) and 14 quantitative (weighted) network-based measures of community similarity which model different aspects of community richness and evenness. We demonstrate the broad applicability of network-based diversity approaches by applying them to 3 distinct datasets: pneumococcal isolates from distinct geographic regions, human mitochondrial DNA data from the Indonesian island of Nias, and proteorhodopsin sequences from the

Sargasso and Mediterranean Seas. Our results show that network-based measures can recover patterns of community variation similar to those recovered using tree-based measures. However, tree- and network-based community similarity can differ substantially when discordant phylogenetic signals are present in the underlying data. Network-based measures provide a methodology for assessing the robustness of beta-diversity results in light of discordant signal, and suggest new measures of beta diversity which can be applied to widely used network structures such as median networks.

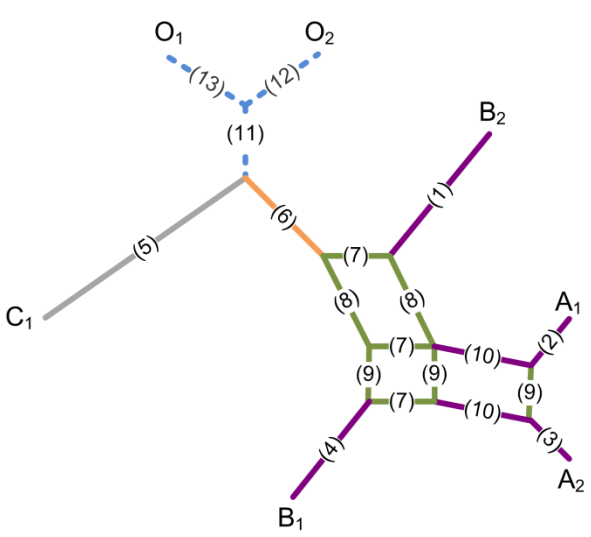
5.2 Introduction

Beta-diversity measures are used to assess variation in community composition between sample sites. Examining patterns of beta diversity across environmental gradients, between treatment conditions, or over time provides insights into the spatiotemporal dynamics of communities and the influence of ecological factors on biodiversity (Anderson et al. 2011). Measures of beta diversity have been used in studies which range from the investigation of human migration patterns (Kayser et al. 2008; HUGO Pan-Asian SNP Consortium 2009) to the exploration of the structure of human-associated bacterial communities (Costello et al. 2009; Caporaso et al. 2011). Although community variation has traditionally been assessed by considering changes in species composition with *taxon-based measures* such as the Bray-Curtis or Canberra indices, more recent methods account for the relative similarity of taxa with *sequence-based measures* that consider the genetic distance between aligned sequences (e.g., F_{ST}, Holsinger and Weir 2009) or with *phylogenetic-based measures* that compute distances between taxa using a phylogenetic tree (e.g., UniFrac, Lozupone and Knight 2005). Phylogenetic-based measures are becoming increasingly popular as they are complementary to taxon-based measures (Graham and Fine 2008; Hamady et al. 2010), and eliminate the need to separate units of diversity into predefined groups. Although there is evidence suggesting phylogenetic-based measures are robust to the methods used for phylogenetic inference (Lozupone et al. 2007), they are currently restricted to being calculated over a single estimate of the true phylogeny. Here we describe how an important class of beta-diversity measures, which includes the Bray-Curtis, Canberra, and UniFrac measures, can be extended in order to account for phylogenetic uncertainty and conflict.

Phylogenetic networks are a generalization of phylogenetic trees and can be divided into 2 distinct classes (Huson and Bryant 2006): (1) *implicit* networks which represent phylogenetic uncertainty and conflict in the available phylogenetic signal, and (2) *explicit* networks that represent evolutionary histories that can contain reticulate events such as hybridization or lateral gene transfer. Phylogenetic uncertainty and conflict can arise when there is insufficient signal to adequately resolve all branches of a bifurcating evolutionary process (Ho and Jermiin 2004; Kennedy et al. 2005), or when the underlying evolutionary history of a set of taxa contains reticulate events. Implicit networks aim to represent all, or at least many, plausible evolutionary scenarios for a set of taxa (Huson and Scornavacca 2011; Morrison 2011). In contrast, explicit networks aim to describe a single biologically plausible phylogeny which may include reticulate connections.

The network-based measures we have developed allow beta diversity to be calculated over split systems, a widely used class of implicit networks (Huson and Scornavacca 2011). A *split* $X|Y$ is a bipartition of the taxa into 2 non-empty, disjoint subsets X and Y. Splits systems that can be represented as a phylogenetic tree are termed *compatible*. More generally, a split system may contain *incompatible* splits in which case it can be represented graphically as a split network or as a table enumerating all splits (Fig. 5.1). Graphically, a split is represented by one or more parallel lines. Each split has a weight representing, among other possibilities, the amount of evolutionary change that has occurred between the taxa on either side of the split.

We have focused on split systems as they are widely used within the biological sciences and include phylogenetic trees as a special case. Character-based (Bandelt and Dress 1993; Bandelt et al. 1995), distance-based (Bandelt and Dress 1992; Bryant and Moulton 2004), and tree-based (Huson et al. 2004; Holland et al. 2005; Holland et al. 2007) methods for inferring incompatible split systems have all been proposed. Implementations of many split system algorithms are available in Spectronet (Huber et al. 2002) and SplitsTree4 (Huson and Bryant 2006). Two popular inference methods are the character-based median network method (Bandelt et al. 1995) commonly used to study the relationships between human populations (Bandelt et al. 1995; Herrnstadt et al.



Split #	Split	Classification
1	$B_2 \mid A_1A_2B_1C_1O_1O_2$	U_B
2	$A_1 \mid A_2B_1B_2C_1O_1O_2$	U_A
3	$A_2 \mid A_1B_1B_2C_1O_1O_2$	U_A
4	$B_1 \mid A_1A_2B_2C_1O_1O_2$	U_B
5	$C_1 \mid A_1A_2B_1B_2O_1O_2$	E
6	$A_1A_2B_1B_2 \mid C_1O_1O_2$	R
7	$A_1A_2B_2 \mid B_1C_1O_1O_2$	S
8	$A_1A_2B_1 \mid B_2C_1O_1O_2$	S
9	$A_2B_1 \mid A_1B_2C_1O_1O_2$	S
10	$A_1A_2 \mid B_1B_2C_1O_1O_2$	U_A
11	$O_1O_2 \mid A_1A_2B_1B_2C_1$	O
12	$O_2 \mid A_1A_2B_1B_2C_1O_1$	O
13	$O_1 \mid A_1A_2B_1B_2C_1O_2$	O

Figure 5.1. An example of a rooted split system depicted as a split network and a table of splits. Each split has been assigned a number and is represented by one or more parallel edges in the split network (e.g., the 2 lines labeled 10 represent the split $A_1A_2|B_1B_2C_1O_1O_2$). The split system contains taxa from 3 communities labeled *A*, *B*, and *C*, and has been rooted using the outgroup taxa labeled *O*. Splits are classified with respect to communities *A* and *B*: unique to community *A* (U_A: purple), unique to community *B* (U_B: purple), shared (S: green), root (R: orange), external (E: grey), or outgroup (O: blue).

2002), and the distance-based neighbour-net method (Bryant and Moulton 2004) that is often applied to provide further insights into poorly understood or complex phylogenies (Morrison 2005; Huson and Bryant 2006). Neighbour-nets have also been used within conservation planning to account for conflicting phylogenetic signal (Spillner et al. 2008; Minh et al. 2009), and as a method for visualizing the relationships between communities determined by a beta-diversity analysis (Mitra et al. 2010).

Our proposed extension allows phylogenetic beta diversity to be measured over unrooted and rooted (i.e., as defined by a set of outgroup taxa) split systems. For rooted systems, both qualitative (i.e., presence-absence, unweighted) and quantitative (i.e., weighted) measures can be computed in order to allow changes in both community richness and evenness to be investigated (Legendre and Legendre 1998; Lozupone et al. 2007). Many of the quantitative measures can also be applied to unrooted split systems. We demonstrate the validity and utility of calculating network-based diversity by analyzing 3 distinct datasets: pneumococcal isolates from distinct geographic regions,

mitochondrial DNA data from the Indonesian island of Nias, and proteorhodopsin sequences from the Sargasso and Mediterranean Seas. In all cases, we found that the previously determined patterns of variation can be recovered on an incompatible split system indicating the robustness of these results to phylogenetic uncertainty and conflict. Nonetheless, finer-scale analyses revealed that the relative dissimilarity of communities can differ substantially when the evolutionary history of taxa is allowed to include incompatible splits, or with the measure of beta diversity considered.

5.3 Methods

5.3.1 Measuring Qualitative Beta Diversity over a Rooted Split System

Here we propose a classification scheme for splits within a rooted split system which allows qualitative phylogenetic beta-diversity measures to be applied to such systems. For each pair of communities, i and j, all splits within a rooted split system can be classified as either *unique*, *shared*, *root*, *external*, or *outgroup* using the following definitions (Fig. 5.1 and Table E.1 in Appendix E):

Unique split: A split (i.e., $X|Y$) is unique to community i only when a subset induced by the split (i.e., either X or Y) contains a taxon or taxa from community i, and does not contain any taxa from community j or the outgroup.

Shared split: A split is shared by communities i and j only when 1) there is a subset induced by the split that contains taxa from communities i and j and no taxa from the outgroup, and 2) the other subset contains at least one taxon from community i or j.

Root split: A split is a root of communities i and j only when 1) there is a subset induced by the split that contains all the taxa from these communities and no taxa from the outgroup, and 2) the other subset contains at least one ingroup taxon.

External split: A split is external to communities i and j only when there is a subset induced by the split that contains only taxa from other communities within the study and no taxa from the outgroup.

Outgroup split: A split belongs to the outgroup when there is a subset induced by the split that contains only outgroup taxa. Additionally, any split where both induced subsets contain outgroup taxa is defined as an outgroup split even if both subsets also include ingroup taxa. In practice, such splits often occur even for a set of credible outgroup taxa.

The terms *a*, *b*, *c*, and *d* are used to define many qualitative taxon-based measures (Koleff et al. 2003; Kuczynski et al. 2010). Using the above classifications, these terms can be applied to a split system by defining *a* as the total split weight shared by communities *i* and *j*, *b* as the total split weight unique to community *i*, *c* as the total split weight unique to community *j*, and *d* as the total split weight external to communities *i* and *j*. Ferrier et al. (2007) and Nipperess et al. (2010) have proposed similar definitions for extending these taxon-based terms to phylogenetic trees. Ferrier and colleagues do not specify how root or external branches were to be treated whereas Nipperess and colleagues implicitly treat root branches as shared (i.e., contribute to *a*).

Our definitions are more general as they can be applied to split systems and explicitly differentiate between shared and root splits. This latter distinction is critical as it allows calculations to be restricted to the MRCA tree of a pair of communities, which can substantially influence the dissimilarity measured between a pair of communities (Chapter 4; in press, Parks and Beiko 2012a). For example, unweighted UniFrac is defined as $(b+c)/(a+b+c)$, but produces 2 distinct measures depending on whether root splits contribute to *a* or *d*. Ambiguity already exists with both the Fast UniFrac web interface (Hamady et al. 2010) and QIIME (Caporaso et al. 2010) implicitly treating root branches as contributing to *a*, whereas mothur (Schloss et al. 2009) treats these branches as contributing to *d* by default (though it does give users the choice to treat these branches as contributing to *a*).

5.3.2 Measuring Quantitative Beta Diversity over a Rooted Split System

Many quantitative phylogenetic beta-diversity measures consider the proportion of taxa from communities *i* and *j* descendant from branch *n* (p_{in} and p_{jn}, respectively) along with the length of the branch (W_n). These terms can be adapted to a rooted split system by defining p_{in} as the proportion of taxa from community *i* within the subset induced by split *n* which contains no outgroup taxa. Splits where both induced subsets contain outgroup taxa are ignored. Within a split system, W_n is simply the weight of split *n*. For example, on a split system the commonly used Manhattan or weighted UniFrac (Lozupone et al. 2007) measure is given by $\sum_n \left| p_{in} - p_{jn} \right| W_n$, where the summation is over all splits within

the split system. We have recently evaluated a large number of measures which are a strict function of p_{in}, p_{jn}, and W_n (Chapter 4; in press, Parks and Beiko 2012a).

5.3.3 Measuring Phylogenetic Beta Diversity over an Unrooted Split System

In practice, a credible rooting for a split system cannot always be established and phylogenetic beta diversity must be determined with a root invariant measure, i.e., a measure producing the same result regardless of root placement. Although root invariant qualitative measures such as the mean nearest neighbour distance and mean phylogenetic distance do exist (Chapter 4; Webb et al. 2008; in press, Parks and Beiko 2012a), any qualitative measures which must distinguish between shared and unique splits require a rooted split system. To our knowledge, all measures which are a function of the terms *a*, *b*, *c*, and *d* require distinguishing between these 2 types of splits. This is unfortunate as these measures are among the most commonly used and best understood qualitative measures. In contrast, many of the most commonly used quantitative measures such as Manhattan, Euclidean, and Gower are root invariant (Chapter 4; in press, Parks and Beiko 2012a).

5.3.4 Interpretation of Beta Diversity Measured over a Split System

Calculating beta diversity over a split system allows incompatible splits representing phylogenetic uncertainty or conflict to be considered (Fig. 5.2). When splits are assigned a weight proportional to the amount of phylogenetic signal supporting the split and accurately reflect the available signal, the dissimilarity measured between a pair of communities will be an average over phylogenetic uncertainty and conflict. In practice, it is often unclear how to best represent the available phylogenetic signal and many inference methods have been proposed. In this chapter, we consider split systems inferred with the UPGMA (Legendre and Legendre 1998), neighbour-joining (Saitou and Nei 1987), neighbour-net (Bryant and Moulton 2004), and median network (Bandelt et al. 1995) methods. UPGMA assumes a constant rate of evolution and produces an ultrametric tree where branch lengths must be adjusted to account for any discordant phylogenetic signal. Neighbour-joining relaxes the molecular clock assumption and aims to infer a tree of minimal length under the "balanced minimum evolution" criterion

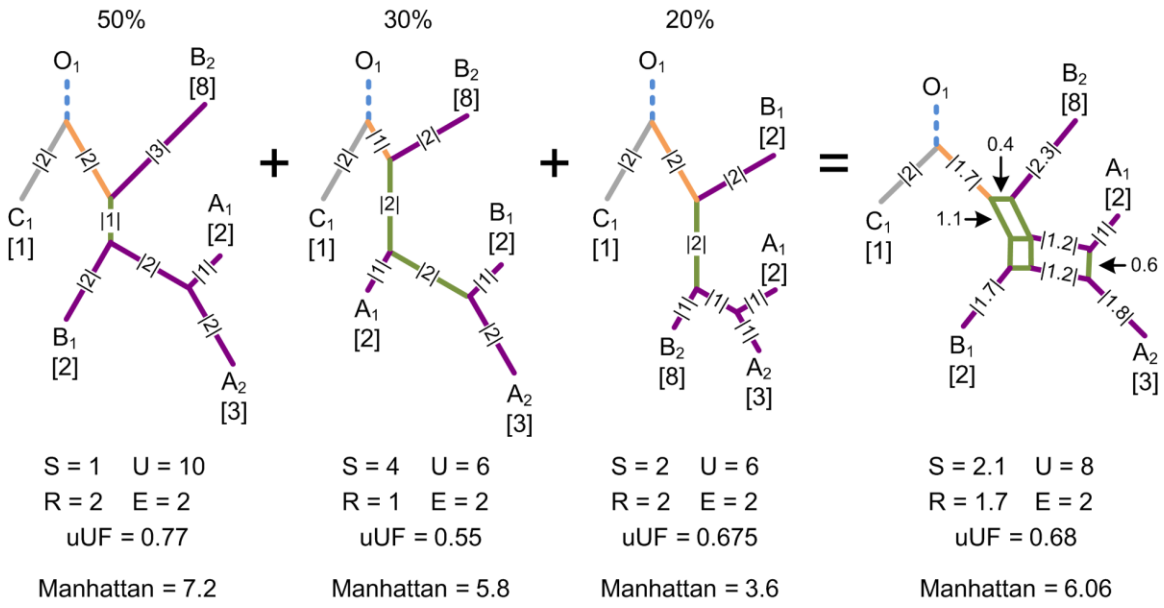


Figure 5.2**.** Measuring beta diversity over a split system provides an average over phylogenetic uncertainty and conflict. A set of trees representing alternative evolutionary histories for a set of taxa is given on the left along with the percentage of data supporting each tree. This could represent the percentage of trees within a collection of trees or the percentage of sites within a multiple sequence alignment supporting each tree. The splits contained within this set of trees can be represented by the split network on the right. Splits within the network have a weight equal to their mean length within the set of trees. The total weight of shared (S), unique (U), root (R), and external (E) splits is given below each phylogeny. Splits are coloured in the same manner as Figure 5.1. Dissimilarity values obtained with the unweighted UniFrac (uUF) and quantitative Manhattan measures for communities *A* and *B* are also given. The uUF distance measured on the split network is *not* the mean of the uUF distances measured over the set of tree. When calculating beta diversity over a split system, each split within the set of trees is considered and weighted by the amount of phylogenetic signal supporting the split.

(Gascuel and Steel 2006). Neighbour-net infers a more general split system consisting of a collection of *circular* splits which can contain mutually incompatible splits. A set of circular splits can always be represented as a planar split network (Dress and Huson 2004) and neighbour-net tends to produce well-resolved networks (Bryant and Moulton 2004). Notably, neighbour-net will infer a tree if the provided distance matrix is additive; i.e., can be perfectly represented by a tree (Bryant et al. 2007). A median network allows *all* splits within a set of *binary* sequences to be inferred. Although the resulting split systems are often too complicated to visualize (Huber et al. 2001), beta diversity can still

be calculated over the inferred splits. The split systems inferred with alternative methods can differ substantially, which will directly influence the phylogenetic dissimilarity measured between a pair of communities. These differences in community similarity must be interpreted with regard to how phylogenetic uncertainty and conflict are treated, and the varying assumptions made by each method.

5.3.5 Pneumococcus Dataset

A multilocus sequence typing (MLST) dataset of *Streptococcus pneumoniae* serotype 1 isolates from 29 geographic regions was compiled from previous studies (Table F.1 in Appendix F). These studies characterized isolates by sequencing fragments of 7 genes according to the protocol of Enright and Spratt (1998). Each allele was assigned a distinct number and each set of 7 unique alleles specifies a sequence type. The distance between sequence types is the number of alleles in which they differ. Using these distances a UPGMA tree, neighbour-joining tree, and neighbour-net were inferred with SplitsTree v4.12.3 (Huson and Bryant 2006). The relationship between samples was visualized using a UPGMA tree. Jackknife values were calculated over 100 independent trials in order to assess the robustness of results to sequence subsampling (Lozupone and Knight 2005). For each trial, sequences were randomly drawn in order to reduce each sample to the number of sequences contained in the smallest sample.

5.3.6 Mitochondrial DNA Dataset

Mitochondrial DNA (mtDNA) sequences covering the first hypervariable region (HVR1) were collected by van Oven et al. (2011) from 9 groups on Nias, Indonesia. This region of the mtDNA is widely used in human migration studies as it contains large numbers of mutations which can be used to infer common descent in subpopulations. mtDNA has been used extensively to define human haplogroups, sets of sequences sharing a few key substitutions that are presumed to have arisen in a common ancestor. We obtained these sequences from the NCBI PopSet database and analyzed all groups with a sample size greater than 10. Sequences were aligned with MUSCLE v3.8.31 (Edgar 2004) using default parameters. A median network was inferred from a binary character representation of the aligned sequences using SplitsTree v4.12.3. Of the 59 variable sites, 53 consisted of only 2 character states yielding a natural binary encoding.

The remaining 6 variable sites contained more than 2 character states. These sites were converted to binary characters using an R/Y encoding. This favours transversional changes by converting bases A and G to Y (pyrimidine), and bases C and T to R (purine). This encoding scheme is used by Spectronet (Huber et al. 2002) and is similar to the encoding proposed by Bandelt et al. (1995). Pairwise F_{ST} values were calculated using Arlequin 3.5.1.3 (Excoffier et al. 2005). PCoA plots were generated using custom scripts. Geographic visualizations showing the similarity of Hia to other groups on Nias were generated with GenGIS (Parks, Porter, et al. 2009).

5.3.7 Proteorhodopsin Dataset

Proteorhodopsin sequences from 12 samples collected by Sabehi et al. (2007) and 3 additional Mediterranean Sea samples submitted by Sabehi and Béjà (ABD84734-ABD85012) were retrieved from GenBank. Protein sequences were aligned with MUSCLE v3.8.31 using default parameters. Initial and trailing columns were trimmed if they contained missing data for >90% of the sequences. Maximum likelihood distances between sequences were calculated with the Whelan and Goldman (WAG) model of protein substitution (Whelan and Goldman 2001), and a neighbour-net was inferred with SplitsTree v4.12.3. The phylogeny was rooted with a eukaryotic rhodopsin from *Pyrocystis lunula* (AAO14677; de la Torre et al. 2003; Sabehi et al. 2005; Sabehi et al. 2007). The relationship between samples was visualized using a UPGMA tree and jackknife values calculated as described for the pneumococcus dataset. Genotype-specific samples were also considered by assigning each protein sequence as either preferentially blue-absorbing or green-absorbing based on whether the amino acid at position 105 was a glutamine or leucine, respectively (Table G.1 in Appendix G). Eight green-absorbing sequences containing a methionine at position 105 were ignored in order to replicate the dataset considered by Sabehi et al. (2007). When calculating unweighted UniFrac values we treat root splits as contributing to *a* (i.e., shared between the 2 communities).

5.3.8 Software Availability

Our software, Network Diversity, for calculating 11 qualitative and 14 quantitative network-based measures of phylogenetic beta diversity is freely available at http://kiwi.cs.dal.ca/Software/NetworkDiversity. The software is open source and

released under the GNU General Public License. It is compatible with split systems generated by the widely used SplitsTree4 software. The software is designed for large datasets and the limiting factor for most analyses will be the computational resources required to infer the underlying split system. On a split system containing 10,000 taxa from 500 environmental samples, calculating the beta diversity between all pairs of samples requires approximately 10 minutes on a standard desktop computer.

5.4 Results

5.4.1 Pneumococcal Biogeography: Alternative Phylogenies Influence Beta Diversity

Pneumococcal infections are a major cause of morbidity and mortality, causing diseases ranging in severity from otitis media to meningitis and pneumonia. Serotype 1 pneumococci are increasingly responsible for invasive pneumococcal diseases in several countries (Henriques Normark et al. 2001; McChlery et al. 2005; Obando et al. 2008) and a major cause of pneumonia and pulmonary empyema in children (Esteva et al. 2011). Here we consider the biogeography of serotype 1 isolates from 29 distinct geographic regions spanning 4 continents (Table F.1 in Appendix F). We investigated the geographic distribution of serotype 1 by applying the Manhattan phylogenetic beta-diversity measure to a neighbour-net, neighbour-joining tree, and UPGMA tree inferred from serotype 1 MLST sequence types (STs). Differences between these results reveal that the underlying phylogeny can substantially influence the similarity measured between communities.

Hierarchical clustering of the Manhattan distances obtained over a neighbour-net shows clear geographic structuring (Fig. 5.3a). A well-supported clustering separates North American and European serotype 1 populations from those in Africa, Asia, and South America. The North American and European cluster is itself separated into its respective continents with the exception of England which falls within the North American cluster. Isolates collected from South America appear highly distinct from those obtained in other regions whereas isolates collected from countries in Africa and Asia are intermixed indicating they contain similar serotype 1 clones. These results are in agreement with an earlier study of serotype 1 isolates collected from 14 countries

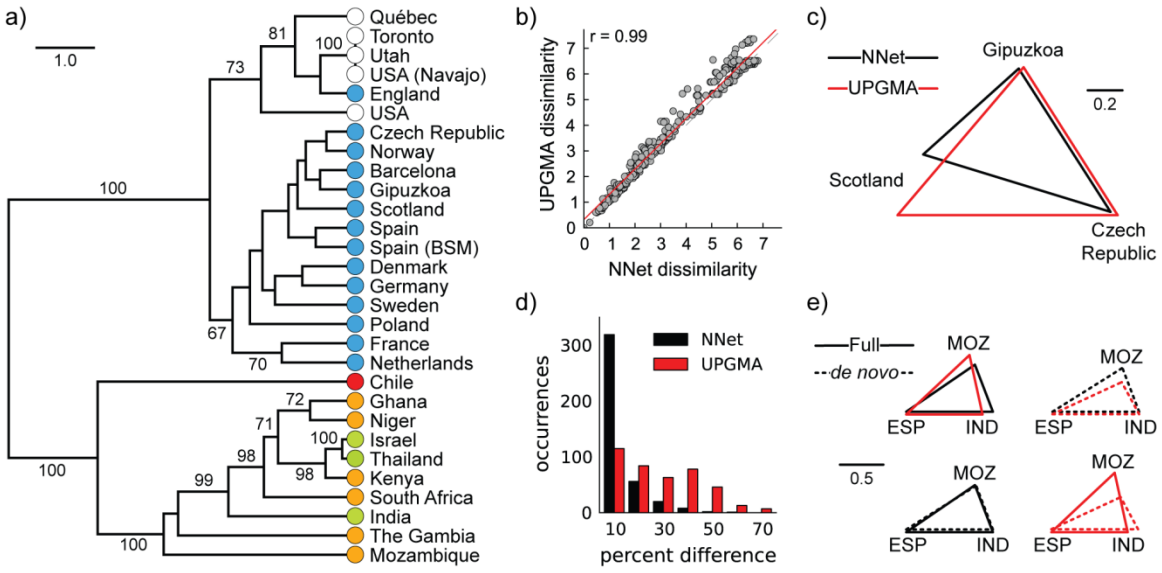


Figure 5.3. Similarity of pneumococcus isolates from 29 countries measured over a neighbour-net or UPGMA tree. (**a**) Hierarchical clustering of the neighbour-net distance matrix with jackknife values greater than 60 shown. Colours indicate the continent of each sample. (**b**) Scatter plot relating the dissimilarity values for every pair of countries over a neighbour-net (x-axis) or UPGMA tree (y-axis). The y=x line is shown as a dashed line, the linear regression line is shown in red, and Pearson's correlation coefficient, *r*, is given in the top-left corner. (**c**) Distances measured over a neighbour-net and UPGMA tree which differ substantially for select pairs of geographic regions. (**d**) Histogram of the percentage difference between distances measured for every pair of geographic regions when the full dataset is considered or when a *de novo* analysis is performed for each pair. (**e**) Comparison of distances measured between Mozambique (MOZ), Spain (ESP), and India (IND) when considering the full dataset or performing a *de novo* analysis using only STs from these 3 countries. To show that a linear scaling does not make the distances congruent, each set of measurements was normalized by the longest distance between any pair of countries.

(Brueggemann and Spratt 2003). This is highly encouraging as Bruggemann and Spratt drew their conclusions by considering a UPGMA phylogeny of serotype 1 STs along with a table indicating the number of times each ST is observed in a country. Phylogenetic beta-diversity measures require the same data as input, but provide a less subjective and more scalable methodology for exploring the relative similarity between populations.

Calculating beta diversity over a UPGMA tree gives similar results to those obtained on a neighbour-net (Fig. 5.3b) despite the neighbour-net containing an additional 67

(47% increase) splits (Fig. 5.4). Although the majority of pairwise distances between geographic regions are robust to the underlying phylogeny differences do exist. For example, the distance measured between isolates from Scotland and the Spanish city of Gipuzkoa are substantially larger when considering a UPGMA tree instead of a neighbour-net (Fig. 5.3c; 1.12 vs. 0.74). A simple scaling cannot be used to resolve such incongruencies as many pairwise distances are already in agreement (e.g., Gipuzkoa and Czech Republic). To further investigate how beta-diversity results differ under alternative phylogenies, we compared the distances measured between every pair of geographic regions when all STs are considered or when a *de novo* analysis is performed using only the STs from the pair of regions under consideration. Distances changed more drastically when they were calculated over a UPGMA tree (Fig. 5.3d). For example, even though the distances measured between Mozambique, Spain, and India are nearly identical when measured over a neighbour-net inferred on the full dataset or *de novo*, they change drastically when a UPGMA tree is used as the underlying phylogeny (Fig. 5.3e). This suggests that considering the distance between specific subsets of samples will generally be more robust when beta diversity is measured over a neighbour-net. Repeating the above analysis with a neighbour-joining tree also indicates that measuring beta diversity over a neighbour-net will readily recover strong patterns of variation, but that consideration of phylogenetic uncertainty and conflict can substantially change the dissimilarity measured between certain samples, e.g., Québec (Fig. 5.5).

5.4.2 mtDNA Diversity in Nias: Contrasting Sequence- and Phylogenetic-based Measures

Nias is an Indonesian island located approximately 120 km off the western coast of Sumatra. Its inhabitants are known for their distinctive architecture (Bontaz 2009) and unique Austronesian dialect (Gray et al. 2009). Within Nias, there is a cultural and linguistic division between the northern and southern portions of the island (Beatty 1993; Bontaz 2009). A recent genetic study of Niasans found paternally inherited Y chromosome haplogroups to be strongly differentiated between south and north populations, whereas maternally inherited mtDNA haplogroups were more evenly

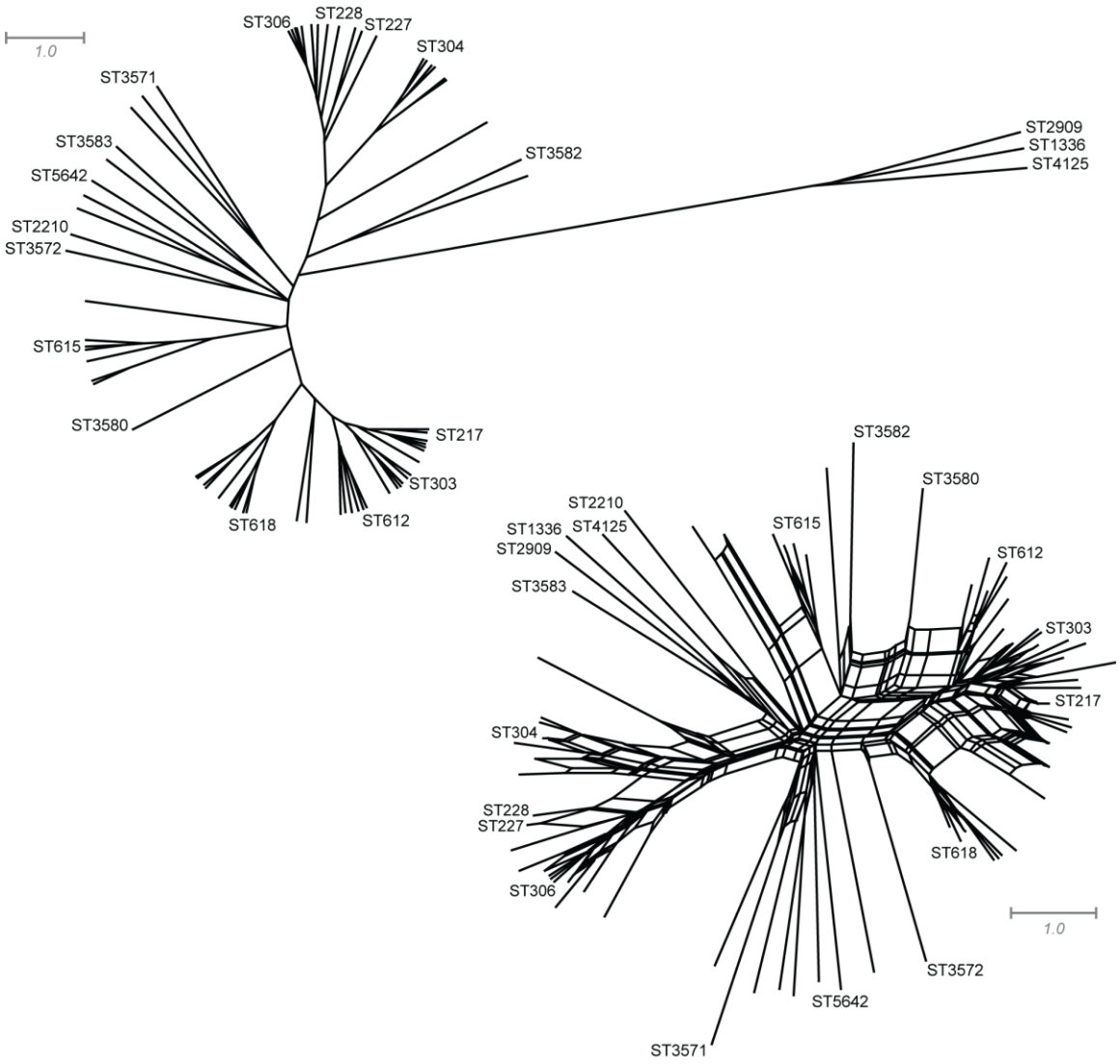


Figure 5.4**.** UPGMA tree and neighbour-net of pneumococcus serotype 1 isolates from 29 geographic regions. For clarity, only the most common sequence types are shown along with sequence types useful as landmarks. The UPGMA tree consists of 143 splits and has a fit of 0.81, whereas the neighbour-net consists of 210 splits and has a fit of 0.93. Fit is defined as the sum of all pairwise distances between taxa in the tree divided by the sum of all pairwise distances in the input distance matrix.

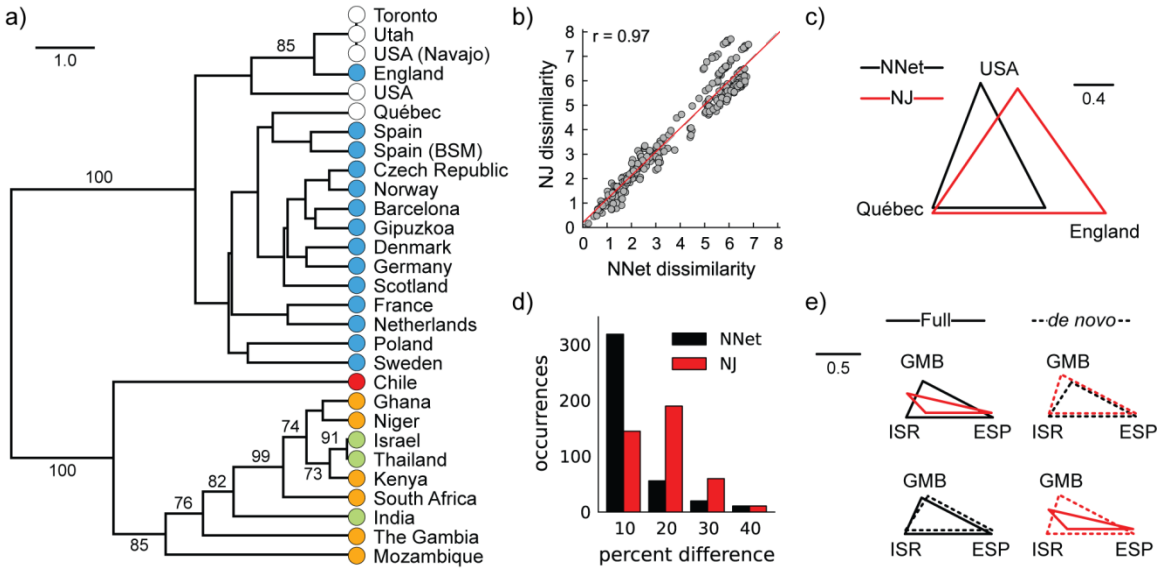


Figure 5.5. Similarity of pneumococcus isolates from 29 countries measured over a neighbour-net or neighbour-joining tree. (**a**) Hierarchical clustering of the neighbour-joining distance matrix with jackknife values greater than 60 shown. Colours indicate the continent of each sample. (**b**) Scatter plot relating the dissimilarity values for every pair of countries over a neighbour-net (x-axis) or neighbour-joining tree (y-axis). The y=x line is shown as a dashed line and is nearly identical to the linear regression line shown in red. Pearson's correlation coefficient, *r*, is given in the top-left corner. (**c**) Distances measured over a neighbour-net and neighbour-joining tree differ substantially for select pairs of geographic regions. (**d**) Histogram of the percentage difference between distances measured for every pair of geographic regions when the full dataset is considered or when a *de novo* analysis is performed for each pair. (**e**) Comparison of distances measured between Israel (ISR), Spain (ESP), and Gambia (GMB) when considering the full dataset or performing a *de novo* analysis using only STs from these 3 countries. To show that a linear scaling does not make the distances congruent, each set of measurements was normalized by the longest distance between any pair of countries.

distributed throughout the island (van Oven et al. 2011). The more uniform distribution of mtDNA haplogroups is likely the result of Niasan culture whereby a woman must marry a man from a different clan and relocate to her husband's village. Here we determine if consideration of the phylogenetic relationships between mtDNA HSV1 sequences changes the conclusions of van Oven et al. (2011).

We assessed the similarity of Niasan groups by applying both the phylogenetic-based Manhattan measure to a median network and the sequence-based F_{ST} measure. These 2 dissimilarity measures produce distinct patterns of relationships between Niasan groups

(Figs. 5.6a and 5.6b). Most notably, under F_{ST} the Si'ulu population from southern Nias appears relatively distinct from all other populations, whereas the Manhattan measure identifies this population as being closely related to the southern Sarumaha population. This striking difference occurs despite the 2 measures being reasonably correlated (Fig. 5.6c). Differences are not confined to the Si'ulu population, nor are they an artifact of the PCoA plots. For example, the relative similarity of the centrally located Hia group to the other Niasan groups depends strongly on the dissimilarity measure used (Figs. 5.6d and 5.6e), and subsets of groups exist whose dissimilarity are only poorly correlated between the 2 measures (Fig. 5.6f; Pearson's r = 0.37). Although consideration of the phylogenetic structure between mtDNA sequences substantially changes the relative similarity between Niasan groups, our results largely corroborate the conclusions of van Oven et al. (2011). Neither the F_{ST} nor Manhattan measures support a north-south division in mtDNA diversity.

5.4.3 Distribution of Proteorhodopsins: Qualitative and Quantitative Beta Diversity

Proteorhodopsin proteins provide a diverse range of aquatic bacteria with a light-driven proton pump suggesting that photosynthesis may play a significant role in the metabolism of aquatic ecosystems (Béjà et al. 2000; Sharma et al. 2008). To investigate the distribution of proteorhodopsin proteins, Sabehi et al. (2007) collected sequences from 12 environmental samples. Three samples were taken at the BATS station in the Sargasso Sea in March (1998 and 2003), when deep water mixing occurs, and another 3 samples were collected in July (1998) when the water is highly stratified. Samples were taken at depths of 0, 40, and 80 m. Analogous samples were taken from the H01 station in the Mediterranean Sea in January 2006 (mixed) and May 2003 (stratified) at depths of 0, 20, and 50-55 m. Here we analyze these samples along with 3 additional Mediterranean samples taken at the same depths and collected 90 km away at the TB04 station in February 2006 (Table G.1 in Appendix G). We consider a rooted neighbour-net in order to establish that the conclusions of Sabehi et al. (2007) regarding the geographic,

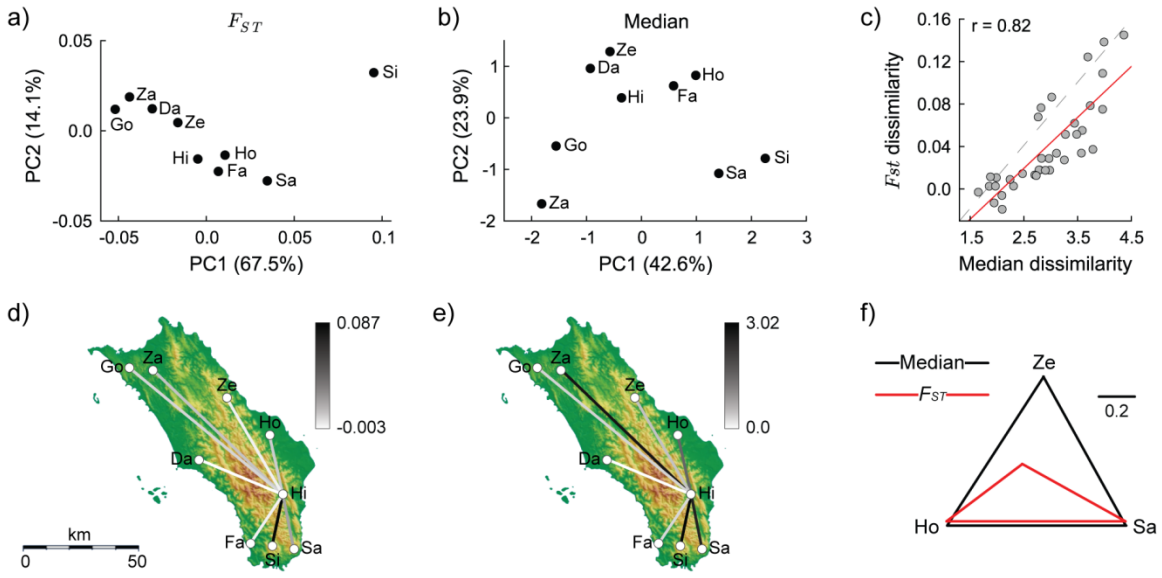


Figure 5.6**.** Similarity of 9 groups residing on Nias determined using F_{ST} or applying the Manhattan phylogenetic beta-diversity measure to a median network. Ethnic groups abbreviated as follows: Daeli (Da), Fau(Fa), Gözö (Go), Hia (Hi), Ho (Ho), Sarumaha (Sa), Si'ulu (Si), Zalukhu (Za), Zebua (Ze). (**a, b**) PCoA plot of F_{ST} values (a) and Manhattan distances (b). The percentage of total variance explained by each axis is shown in parentheses. (**c**) Scatter plot relating the dissimilarity values for every pair of groups measured over the median network (x-axis) or using F_{ST} (y-axis). The y=x line is shown as a dashed line, the linear regression line is shown in red, and Pearson's correlation coefficient, *r*, is given in the top-left corner. (**d, e**) Dissimilarity measured between Hi and each of the other 8 groups as determined using the F_{ST} (d) or Manhattan (e) measures. (**f**) Comparison of F_{ST} values and Manhattan distances between the Ho, Sa, and Ze ethnic groups. To aid in comparing the relative magnitude of values, each set of measurements was normalized by the longest distance between any pair of groups.

seasonal, and depth structuring of these samples can be recovered using both qualitative and quantitative phylogenetic beta-diversity measures. Furthermore, we investigate the structuring of specific proteorhodopsin genotypes.

Hierarchical clustering of the unweighted UniFrac dissimilarities among proteorhodopsin communities revealed clustering by geographic location and a strong seasonal dependence (Fig. 5.7a). Samples taken at different depths during periods of water mixing were more similar to each other than the stratified samples, but both sets of samples showed geographic structuring (Fig. 5.7b). Under the quantitative Manhattan measure, stratified communities were no longer separated perfectly by geography (Fig. 5.8a) and showed high variation in the dissimilarity measures between samples (Fig.

5.8b). In contrast, mixed communities were strongly geographically structured and even showed small-scale structuring between the H01 and TB04 samples taken only 90 km apart. The average Manhattan distance between samples is 0.14 for the mixed H01 samples and 0.12 for TB04 samples, in comparison to 0.23 across these 2 sets of samples. Together, the qualitative and quantitative results suggest that during periods of water mixing the proteorhodopsin proteins at these 2 stations are similar, but differ in their relative abundance.

Proteorhodopsins are preferentially blue-absorbing (BPR) or green-absorbing (GPR) based largely on the amino acid at position 105 (Béjà et al. 2001; Man et al. 2003). The structuring determined above may be reliant on both spectral genotypes, primarily due to a single genotype, or observed individually for each genotype. To test this, we define "genotype communities" by dividing each sample into 2 sets consisting exclusively of BPR or GPR proteins. This results in 24 "genotype communities" as only a single GPR sequence was found in the Sargasso Sea. Applying the unweighted UniFrac measure to these communities indicated that these 2 genotypes are phylogenetically distinct (Figs. 5.7c and 5.7d). Congruent with the results for the undivided samples, strong seasonal dependencies were observed for both genotypes. However, perfect geographic clustering was not observed between the BPR samples indicating that this structuring is largely due to the absence of GPR proteins in the Sargasso Sea. Similar results were obtained under the Manhattan measure. Samples clustered perfectly by genotype and showed strong seasonal structuring within each genotype (Figs. 5.8c and 5.8d). Even though there was evidence of small-scale geographic structuring between the mixed BPR samples from H01 and TB04, the mixed GPR samples showed little structuring. This may be a result of higher dispersal of GPR proteins relative to BPR proteins. However, the small number of GPR proteins found within these samples may be unduly influencing the results and deeper sequencing is required in order to further explore the structuring of specific proteorhodopsin genotypes.

5.5 Discussion

Tree inference algorithms produce a single tree model even if there is substantial incongruent phylogenetic signal, whereas split systems allow uncertainty and conflict to

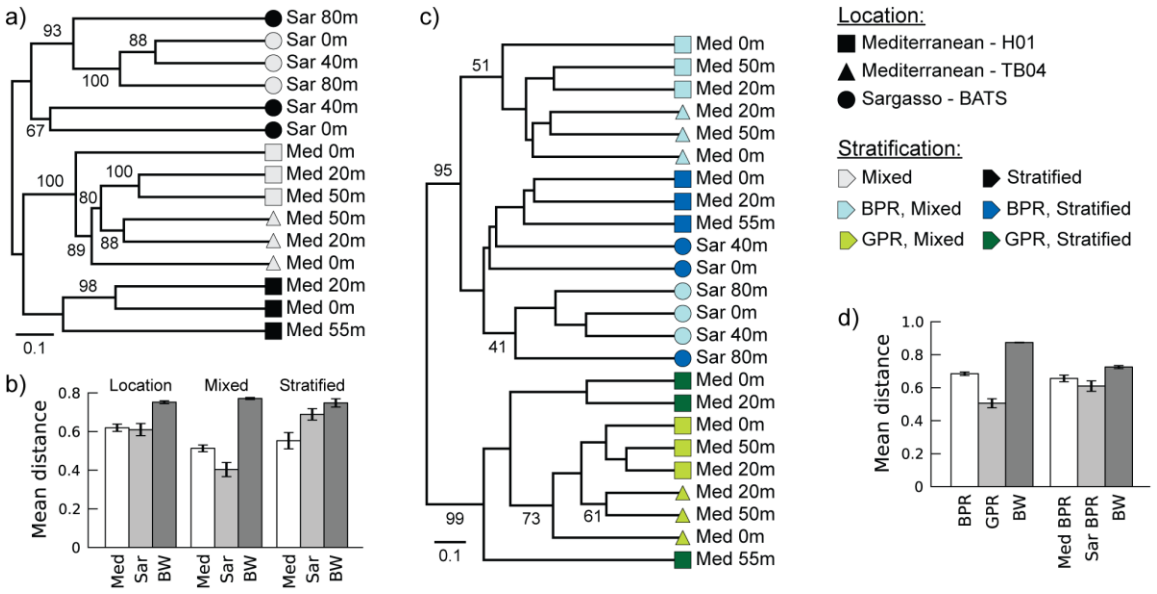


Figure 5.7. Relationship between proteorhodopsin communities and genotype-specific samples determined by applying unweighted UniFrac to a rooted neighbour-net. (**a**) Hierarchical clustering of 15 proteorhodopsin samples collected from the Sargasso (Sar) or Mediterranean (Med) Seas with jackknife values greater than 50 shown. (**b**) Mean (±SEM) unweighted UniFrac dissimilarity values for *all* samples either within the same or between (BW) the 2 sampling locations. The same analysis was also repeated using only samples taken during periods of deep water mixing or when the water is highly stratified. (**c**) Hierarchical clustering of blue (BPR) and green (GPR) proteorhodopsin genotype samples with jackknife values greater than 40 shown. (**d**) Mean (±SEM) unweighted UniFrac dissimilarity values for samples from the same genotype or between samples from different genotypes. Results are also shown for BPR samples collected from the Sargasso or Mediterranean Seas.

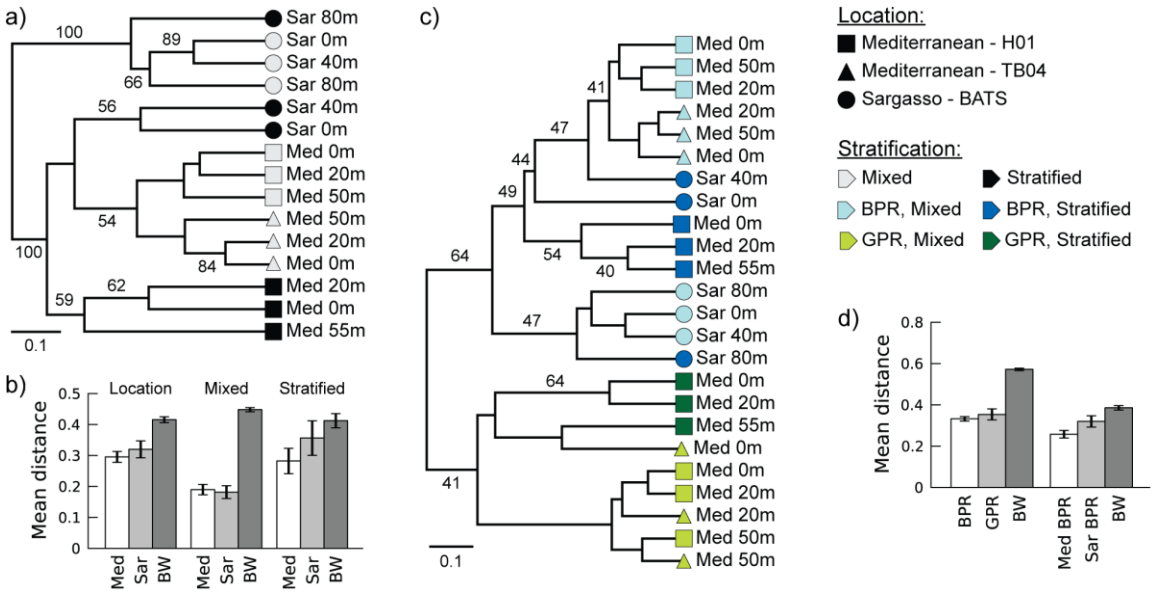


Figure 5.8. Relationship between proteorhodopsin communities and genotype specific samples determined by applying the quantitative Manhattan measure to an unrooted neighbour-net. (**a**) Hierarchical clustering of 15 proteorhodopsin samples collected from the Sargasso (Sar) or Mediterranean (Med) Seas with jackknife values greater than 50 shown. (**b**) Mean (±SEM) Manhattan distances for *all* samples either within the same or between (BW) the 2 sampling locations. The same analysis was also repeated using only samples taken during periods of deep water mixing or when the water is highly stratified. (**c**) Hierarchical clustering of blue (BPR) and green (GPR) proteorhodopsin genotype samples with jackknife values greater than 40 shown. (**d**) Mean (±SEM) Manhattan distances for samples from the same genotype or between samples from different genotypes. Results are also shown for BPR samples collected from the Sargasso or Mediterranean Seas.

be modeled. Phylogenetic beta-diversity measures have been shown to produce similar results under varying tree inference methods (Lozupone et al. 2007) and we have observed, at least when strong patterns of community variation exist, that dissimilarity values will remain highly correlated even when measuring beta diversity over an incompatible split system (data not shown). Identifying similar patterns of community variation on an inferred tree *and* on a split system provides strong evidence that results are not due to systematic errors (e.g., the forcing of data into a tree model). When community relationships are sensitive to the underlying method used for phylogenetic inference results must be interpreted carefully. For example, even though the primary conclusions regarding the biogeographic distribution of pneumococcus serotype 1 were

recovered using UPGMA, neighbour-joining, and neighbour-net, the placement of the Québec sample depended on whether phylogenetic inference was performed with neighbour-joining or neighbour-net (contrast Figs. 5.3a and 5.5a). This may be the result of the neighbour-joining algorithm forcing incongruent signal into a tree model, or the neighbour-net algorithm inferring incompatible splits that would best be viewed as phylogenetic noise.

Our analysis of mtDNA diversity in Nias demonstrates that dissimilarity values obtained using the sequence-based F_{ST} measure can vary considerably from those obtained by applying the Manhattan measure to a median network (Pearson's $r = 0.82$). This disagreement between beta-diversity measures, although not necessarily surprising, is of particular interest as it contrasts the population- and lineage-based approaches commonly used in studies of human biogeography (Pakendorf and Stoneking 2005). In a population-based analysis, F_{ST} or a related measure is used to explore the relationship between populations whereas in a lineage-based analysis a (often simplified) median network is used to investigate the evolutionary history of mtDNA or Y-chromosome haplogroups (Bandelt et al. 1995; Huber et al. 2001). By extending beta-diversity measures to median networks, population-based analyses can be performed directly over a median network instead of on the underlying sequence data. Interestingly, the dissimilarity values obtained with F_{ST} or by applying the Manhattan measure to a median network can be highly negatively correlated indicating that these measures are capturing fundamentally different notions of beta diversity (Fig. 5.9). Further investigation is warranted to determine which phylogenetic beta-diversity measures are best suited for studying human populations, the influence biologically motivated simplifications of median networks have on measured beta diversity (i.e., reduced and pruned networks, Bandelt et al. 1995; Huber et al. 2001), and how phylogenetic-based measures complement traditional sequence-based measures.

Qualitative and quantitative measures of beta diversity provide complementary information on community variation (Legendre and Legendre 1998; Lozupone et al. 2007). On the proteorhodopsin dataset, communities sampled during periods of water mixing were found to be strongly geographically structured under both the unweighted

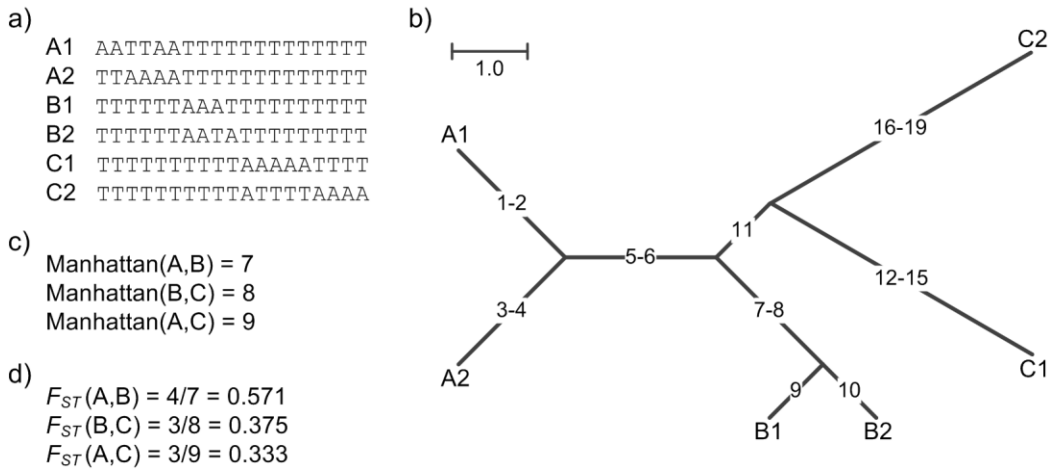


Figure 5.9. An example illustrating poor correlation between dissimilarity values obtained by applying either F_{ST} to sequence data or the quantitative Manhattan measure to a median network. (**a**, **b**) Aligned sequence data (a) and the inferred median network (b). Edges in the network are labeled by the mutated position in the sequence alignment. (**c**) The quantitative Manhattan distance between the 3 communities A, B, and C. Each community consists of 2 sequences, e.g., community A consists of the sequences A1 and A2. (**d**) F_{ST} dissimilarity values between the 3 communities. Manhattan and F_{ST} dissimilarity values are rank-order inverted and have a Pearson's correlation coefficient of r = -0.937. These results are not restricted to median networks, as both the neighbour-joining and neighbour-net methods produce the same phylogeny for these sequences when inferred from a distance matrix indicating the number of sites that differ between 2 sequences.

(qualitative) UniFrac and quantitative Manhattan measures. This suggests that unique proteorhodopsin lineages are present in the Sargasso and Mediterranean Seas (qualitative results), and that such lineages do not consist only of rare proteins (quantitative results). In contrast, the small-scale geographic structuring between the H01 and TB04 samples is only recovered under the quantitative Manhattan measure suggesting that these sites contain similar proteins, but in different abundances. Even though the additional insights provided by applying both a qualitative and quantitative measure are clear, in practice applying qualitative measures is complicated by requiring a rooted split system (Chapter 4; in press, Parks and Beiko 2012a). Rooting an incompatible split system can be problematic as there will often be several low-weight splits where both induced subsets contain ingroup and outgroup taxa. As such, our proposed definitions for calculating qualitative and quantitative beta diversity on a split system ignore any splits where both

induced subsets contain outgroup taxa. Nonetheless, care should be taken to ensure that outgroup taxa are tightly clustered within the split system. For closely related taxa, an interesting alternative is to root a split system based on neutral coalescent theory (Castelloe and Templeton 1994).

We have described how an important class of phylogenetic beta-diversity measures can be applied to split systems containing incompatible splits. This allows phylogenetic uncertainty and conflict to be considered in the assessment of community variation, and provides a methodology for testing if the assumption of a bifurcating evolutionary history is causing erroneous patterns of relationships between communities. Although split systems are the mostly widely used class of phylogenetic networks, they suffer from providing only an abstract representation of the relationship between taxa. Extending beta-diversity measures to networks which explicitly model recombination or hybridization would allow the influence of these evolutionary events on community variation to be explored. We believe the branch classification scheme proposed here is a strong starting ground for extending beta-diversity measures in this direction.

5.6 Acknowledgements

DHP is supported by the Killam Trusts and the Tula Foundation; RGB acknowledges the support of Genome Atlantic, the Canada Foundation for Innovation, and the Canada Research Chairs program.

Chapter 6

Discussion

Major community initiatives now underway stand to vastly increase our understanding of the diversity of life (e.g., Gilbert et al. 2010; Baird and Hajibabaei 2012; Jetz et al. 2012). Ultimately, these initiatives aim to advance our understanding of the contemporary distribution of life and the processes that give rise to this distribution. Tools for exploring the large biogeographic datasets being generated by these initiatives are essential for generating hypotheses and furthering our understanding of biodiversity. Recognition of these needs has resulted in the recent development of several biogeographic visualization packages (Hijmans et al. 2001; Kidd and Liu 2008; Hill and Guralnick 2010; Janies et al. 2010; Laffan et al. 2010; Bielejec et al. 2011). In Chapters 2 and 3 of this thesis, I introduced GenGIS and a novel method for visualizing geotrees. This research was motivated by specific limitations of currently available software packages and the need for interactive geospatial visualization and analysis software. Attention was then given to beta-diversity measures which were first introduced in ecology at the start of the 20th century (Jaccard 1901) and continue to play an important role in assessing the environmental factors that influence community variation (Anderson et al. 2011). Recent efforts have begun incorporating phylogenetic information into these measures in order to account for the relative similarity of taxa (Clarke and Warwick 1998; Martin 2002; Lozupone and Knight 2005). Chapters 4 and 5 of this thesis were motivated by the success of phylogenetic beta-diversity measures in revealing environmental factors which appear to be driving community variation, but are not readily identifiable using traditional taxon-based measures (Graham and Fine 2008; Hamady et al. 2010). In Chapter 4, I performed a large comparative study in order to assess specific properties and performance characteristics of phylogenetic measures of beta diversity which have not previously been adequately examined. An extension of these measures to phylogenetic networks was then discussed in Chapter 5.

The specific contributions of this thesis are given in the following section, and I conclude this thesis with a discussion of promising future avenues of research.

6.1 Contributions of Thesis

6.1.1 Chapter 2: Visualizing Hierarchically Organized Units of Biodiversity

In Chapter 2 I proposed a novel visualization for two-dimensional tree structures which emphasizes the hierarchical structure of data while still associating leaf nodes with specific geographic locations (Parks and Beiko 2009). This two-dimensional tree may represent a geophylogeny intended to convey how geography has influenced a set of taxa or indicate the similarity of communities as determined by applying a hierarchical clustering algorithm to a biotic dissimilarity matrix. This work was motivated by the success of three-dimensional geophylogenies which have been popularized by programs such as Geophylobuilder (Kidd and Liu 2008) and the ability to display such structures within Google Earth (Hill and Guralnick 2010; Janies et al. 2010; Bielejec et al. 2011). These visualizations have been successfully used to study migration patterns (Kidd and Ritchie 2006; Kidd 2010) and the spread of infectious diseases (Janies et al. 2007; Lemey et al. 2009; Parks, MacDonald, et al. 2009). Although these three-dimensional geophylogenies are appropriate when the geographic structure of the data is of primary importance or when ancestral nodes can be assigned meaningful geographic positions (e.g., using fossil evidence or inferred routes of migration), they necessarily distort the hierarchical relationships between entities in order to fit the tree to the underlying geography. In contrast, the visualization I have proposed is specifically designed to maintain the hierarchical relationships in the data by depicting them in standard two-dimensional tree format.

The key principle of the proposed visualization is the determination of an optimal tree layout that minimizes the number of crossings which occur when connecting leaf nodes to sample sites that are ordered according to a specific geographic axis. This is similar in principle to a tanglegram where 2 trees are placed parallel to each other and matching leaf nodes in the 2 trees connected by lines (Holten et al. 2008; Venkatachalam et al. 2010). Here crossings between lines indicate discordance between the 2 trees, whereas in my visualization they indicate discordance with the underlying geography. In order to permit interactive exploration of different geographic axes, I developed a branch-and-bound algorithm to efficiently determine the optimal layout of a tree. Extensions to this

visualization technique permit all possible linear geographic axes to easily be considered and allow nonlinear geographic axes to be explored. In addition, a Monte Carlo permutation test has been proposed which uses the number of crossing between correlation lines as a test statistic.

The specific contributions of Chapter 2 of this thesis are:

- Introduction of a quantitative visualization technique for examining hierarchically organized data within a geographic context.
- Development of the OSCM Branch-and-Bound algorithm for efficiently solving the OSCM problem and an evaluation of the efficiency of this algorithm relative to an exhaustive search of the solution space.
- Development of the Linear Axes Analysis algorithm for efficiently determining the number of crossings which occur for all possible linear geographic axes.
- An extension of the proposed visualization to nonlinear geographic axes.
- Proposal of a Monte Carlo permutation test for assessing whether the fit of a tree to a geographic axis is significantly better than random.
- Demonstration of the proposed visualization and its extensions on 3 case studies.

6.1.2 Chapter 3: Visualization and Analysis of Molecular Biogeography

In Chapter 3 I introduced GenGIS, a free and open-source geospatial environment for interactively visualizing and analyzing the geographic distribution of genetic data (Parks, Porter, et al. 2009; in preparation, Parks et al. 2012). Existing GIS software has been designed for relatively dense observational data typical of multicellular eukaryotes such as plants and animals (Hijmans et al. 2001; Laffan et al. 2010; Jetz et al. 2012) or has focused exclusively on the display of three-dimensional geophylogenies using proprietary software (Kidd and Liu 2008; Hill and Guralnick 2010; Janies et al. 2010; Bielejec et al. 2011). GenGIS addresses the need for a geospatial analysis environment capable of handling large biogeographic datasets where a wealth of sequence data is obtained at each sample site. This is accomplished through a rich set of interactive visualizations which includes visualizing sequence distributions as pie or bar charts, mapping of ecological factors to the visual properties of sample markers, and displaying geographic line graphs of biotic dissimilarity matrices. In addition, GenGIS provides a highly

customizable implementation of the visualization technique developed in Chapter 2. GenGIS is unique in providing visualizations of two- or three-dimensional geotrees along with site-specific visualizations. These visualizations are provided within an interactive environment which supports widely used analysis techniques (e.g., linear regression, the Mantel test, and calculation of biodiversity indices), each of which produces statistical plots and georeferenced visualizations to aid in data exploration. The functionality of GenGIS can also be extended using a plugin framework which allows for the easy development of graphically driven custom visualizations and analyses or by using the built-in Python interpreter which has direct access to all data within GenGIS.

The specific contributions of Chapter 3 of this thesis are:

- Design and development of GenGIS.
- An interactive and highly customizable implementation of the two-dimensional tree visualization technique developed in Chapter 2.
- Application of GenGIS to distinct datasets to demonstrate the flexibility of the software and to illustrate how novel insights can be gained through the synthesis of genetic, ecological, and digital map data.

6.1.3 Chapter 4: Assessing Phylogenetic-based Measures of Beta Diversity

In Chapter 4 I conducted an extensive analysis of 39 measures of phylogenetic beta diversity. I focused on phylogenetic-based measures of beta diversity because they have received far less attention in the literature and are quickly growing in popularity. This analysis complements recent comparative studies of phylogenetic beta diversity which have focused on a limited number of measures (1 to 8) to investigate specific properties: the ability of a measure to identify significantly different communities (Schloss 2008), the correlation between taxon- and phylogenetic-based measures (Nipperess et al. 2010), the success of phylogenetic-based measures in recovering known ecological gradients (Root and Nelson 2011), or the sensitivity of measures to terminal or basal branches (Swenson 2011).

The comparative analysis performed here is the first large-scale analysis of phylogenetic beta-diversity measures. I have evaluated measures popular within both microbial and traditional ecology along with newly established phylogenetic extensions

of commonly used taxon-based measures. Measures commonly used within microbial ecology (e.g., UniFrac variants) were found to be distinct from those popular in classical ecology (i.e., MNND, MPD, Rao's H_p, F_{st}) and from the Gower and Canberra measures whose taxon-based variants were recently recommended by Kuczynski et al. (2010). This result strongly suggests exploratory work should consider multiple measures. I also showed that the performance of measures varies under different models of differentiation. In particular, the Gower and Canberra measures recommended by Kuczynksi and colleagues fail under certain plausible models of differentiation. An investigation of specific properties showed that many measures are robust to sequence clustering, the addition of an outlying basal lineage, root placement, and the presence of rare OTUs. I also demonstrated that some measures are root invariant and can be applied to unrooted trees.

The specific contributions of Chapter 4 of this thesis are:

- Proposal of phylogenetic extensions of commonly used taxon-based measures.
- Demonstration that phylogenetic beta-diversity measures favoured by different disciplines show divergent tendencies that can impact the conclusions of a community analysis.
- Demonstration that the performance of a phylogenetic beta-diversity measure depends on the underlying processes causing communities to differentiate.
- Evaluation of 4 key properties of phylogenetic beta-diversity measures which provide practical insights into: 1) the practice of clustering sequencing into OTUs, 2) the sensitivity of measures to "outliers" mistakenly assigned to a basal lineage, 3) the use of unrooted trees and the relative sensitivity of measures to root placement, and 4) the influence of rare taxa on measured dissimilarity.
- Design and development of Express Beta Diversity, a free and open-source software package for calculating 39 measures of phylogenetic beta diversity and assessing which measures are likely to provide complementary insights into a particular dataset.

6.1.4 Chapter 5: Measuring Beta Diversity over Phylogenetic Networks

In Chapter 5 I described how a large and important class of phylogenetic beta-diversity measures can be calculated over rooted or unrooted split systems. This allows uncertainty and conflict in the available phylogenetic signal to be taken into account when determining the relative similarity of communities or populations. My work was motivated by the success and increased use of phylogenetic-based beta-diversity measures for assessing biogeographic relationships (Graham and Fine 2008; Lozupone and Knight 2008) and recent efforts to use split systems in conservation planning (Spillner et al. 2008; Minh et al. 2009). I have focused on split systems as they are the most prevalent class of implicit phylogenetic networks within the biological sciences (Huson et al. 2010; Morrison 2011) and can be readily inferred using existing software (Huber et al. 2002; Huson and Bryant 2006). To my knowledge, this is the first work which considers the calculation of beta diversity over a phylogenetic network.

I have demonstrated the use of network-based measures of beta diversity by analyzing 3 distinct datasets: pneumococcal isolates from distinct geographic regions, mitochondrial DNA data from Indonesia, and proteorhodopsin sequences from the Sargasso and Mediterranean Seas. The resulting patterns of variation between populations were contrasted with those previously determined using a more subjective methodology. In all cases, I showed that the primary patterns of variation can be recovered by calculating beta diversity over a split system. This indicates both the robustness of these results to phylogenetic uncertainty and conflict and the general applicability of network-based measures of beta diversity in recovering biologically informative patterns. Although phylogenetic beta-diversity measures can recover the primary patterns of variation for each dataset, I also demonstrated that the inferred phylogeny for a set of taxa can substantially influence a beta-diversity measure.

The specific contributions of Chapters 5 of this thesis are:

- Extension of a large class of beta-diversity measures to allow community variation to be calculated over split systems.
- Application of network-based measures of beta diversity to 3 distinct datasets in order to demonstrate that 1) these methods can recover expected patterns of

community variation and that 2) the underlying phylogeny can substantially influence measures of beta diversity.

- Design and development of Network Diversity, a free and open-source software package for calculating over 25 measures of phylogenetic beta diversity over a split system.

6.2 Future Work

This thesis makes several contributions, and the published work in Chapters 2 and 3 have already been applied to other ecology studies (e.g., Bloomquist et al. 2010; Kidd 2010; Page 2012). Nevertheless, there are several promising directions for future work which would increase the usefulness of the proposed methods or address specific limitations of this research. Here I outline future work for each chapter in turn.

The quantitative visualization proposed in Chapter 2 has been used by several research groups within microbial and classical ecology (e.g., Farikou et al. 2011; Poczai et al. 2011; Ruzzante et al. 2011; Tucker et al. 2011). The majority of work making use of this visualization technique has been biogeographic analyses, although Loo et al. (2011) illustrate how the technique can be applied to an abstract gradient of expected relationships and my recent manuscript demonstrates how the technique can be applied within arbitrary spatial contexts such as the human body (in preparation, Parks et al. 2012). A natural extension of the proposed method is the development of a program (or extension to GenGIS) for determining optimal tree layouts for a particular quantitative attribute of interest (e.g., temperature, salinity, season, alpha diversity). That is, instead of using a geographic axis to define an ordering of geographic locations, the ordering could be based on a particular attribute of interest. Currently, the optimal layout of a tree is the one resulting in the fewest crossings between correlation lines. A natural extension to this, particularly useful for axes of general quantitative attributes, would be to determine the layout which minimizes the *weighted* crossings of correlation lines, where the weighting reflects the severity of a given crossings (e.g., the absolute difference of an attribute such as temperature). Initial work has considered the problem of minimizing the cost of crossings where each correlation line has a particular weight (Çakıroğlu et al. 2009), but the problem of minimizing crossings for an arbitrary cost function remains an open problem (Fernau et al. 2010). In terms of running time, it would be interesting to

compare the proposed branch-and-bound technique with the linear programming formulation of Jünger and Mutzel (1996) or with recent fixed-parameter solutions (Nagamochi 2005; Dujmović et al. 2008). Although the branch-and-bound approach is sufficient for most trees, it cannot efficiently handle trees with nodes that have a degree of 9 or higher. The proposed Monte Carlo permutation test requires further investigation in order to assess the statistical power of the test and the required number of permutations for accurate estimation of *p*-values. It should also be noted that when multiple axes are considered, which occurs during the Linear Axes Analysis, no effort is made to account for multiple tests.

GenGIS, as discussed in Chapter 3, is still in active development with recent funding provided through a Genome Canada grant as part of the Biomonitoring 2.0 initiative. Several research groups have used GenGIS to explore both site-specific sequence distributions and the relationship between taxa using either two- or three-dimensional geophylogenies (e.g., Schoville and Roderick 2010; Allal et al. 2011; Shafer et al. 2011). Although the existing GenGIS framework has been used to generate custom visualizations of the spatial distribution of densely sampled species data (namely, the spread of the 2009 H1N1 "Swine Flu" outbreak) additional work is required to adequately handle this type of data. Specifically, GenGIS would benefit from the ability to produce georeferenced heat maps showing interpolated values from a relatively dense underlying geographic sampling of species abundance data. Work also remains to improve the general applicability of geotrees. Recent work has begun exploring the use of geotrees for studying species range data where a given leaf node may be associated with several geographic locations (Kidd 2010). Visualizing these one-to-many relationships cannot currently be done in an efficient manner using GenGIS. The related situation where a single geographic location is associated with multiple leaf nodes also warrants further attention although the two-dimensional geotrees in GenGIS can already accommodate this situation. Methods have also begun incorporating geographic uncertainty into the display of geophylogenies (Bielejec et al. 2011). It would be worth extending GenGIS to include such approaches along with investigating methods for displaying phylogenetic uncertainty. Finally, geophylogenies often haven an explicit notion of when each internal node occurred in time. Visualizations of geophylogenies in

Google Earth allow these temporally annotated geotrees to be "grown" from the root to the leaf nodes as time progresses. GenGIS currently lacks functionality for displaying temporally annotated geotrees.

Handling of nonlinear geographic axes could be improved. Ideally, GenGIS would be extended to support vector files and users could select specific geographic elements to define a nonlinear axis (e.g., the set of lines forming a meandering stream). Additionally, it would be interesting to allow nonlinear hypotheses to be encoded in broad terms (e.g., with thick arrows pointing in different directions from a shared ancestral location), and have GenGIS automatically enumerate all candidate orderings so implied, in the end showing summaries of the goodness of fit of these distinct orderings. The current method for considering nonlinear axes consisting of several polylines also requires additional consideration. In particular, in some circumstances it will be desirable to allow users to place more stringent constraints on the allowed ordering of polylines. Extension of the Linear Axis Analysis algorithm to nonlinear axes would clearly be beneficial although this would require substantial research effort.

The comparative analysis of phylogenetic beta-diversity measures conducted in Chapter 4 raises additional research questions. My work necessarily focused on a small number of models of community variation and I would welcome efforts to examine other biologically motivated models of differentiation in order to try and establish if there is a core set of measures which generally perform well. Of particular interest would be models which consider the turnover of species along an environmental gradient (see Root and Nelson 2011 for initial work), the contrasting of patterns of underdispersion versus overdispersion, or the effect of selective lineage loss. The correlation between measures determined in my work provides guidance as to the relative similarity of measures and hence which measures are likely to produce different patterns of community similarity. Further work is required to establish the generality of the observed correlations to particular datasets. A substantial contribution would be a mathematical treatment showing under what conditions different measures will necessarily be highly correlated. Perhaps easier, would be explicit efforts to explain the correlation between measures under a specific set of biologically motivated conditions. Other properties of these measures are also worth exploring including their robustness to long internal branches

arising from misaligned sequences, the effect of using a static reference tree as opposed to inferring a *de novo* phylogeny in order to reduce computational requirements (see Hamady et al. 2010 for such an assessment of UniFrac), or the influence of varying inference methods including the inference of incompatible split systems. My work has focused on the correlation between the biotic dissimilarity matrices produced by different methods or under varying conditions. Additional work is required to assess how the magnitude of dissimilarity values change under different conditions.

Chapter 5 discusses how phylogenetic beta-diversity measures can be extended to allow community variation to be computed on a split system. This work focuses exclusively on split systems as they are currently the most widely used type of phylogenetic network. My work may suggest how measures can be extended to other types of networks such as those which explicitly model recombination or hybridization (see Huson et al. 2010 for a discussion on these types of networks). Work in this direction is encouraged as such beta-diversity measures would allow the influence of these evolutionary events on community variation to be directly explored. Continued work is also needed to further establish the benefits and limitations of network-based approaches when applied to split systems. In particular, the prevalent use of median networks in biogeographic studies of human populations raises the question of how measures of beta diversity calculated directly over these networks may complement more traditional measures such as F_{ST}.

6.3 Advances to the Field

GenGIS advances the sophistication of biogeography software. It provides novel methods for exploring the hierarchical relationships between georeferenced samples and a rich set of standard GIS visualizations that would otherwise require the use of proprietary software. Evidence of the need for free and open-source biogeography software specifically targeted at biologists is evident from the rapid adoption of GenGIS by several research groups in a range of biological disciplines. The visualizations and analyses provided by GenGIS are already helping ecologists understand how geographic and environmental factors influence biodiversity. Equally as important, the publication-quality images produced by GenGIS help to convey these results not only to fellow researchers, but also to the greater public. Further development of biogeographic

software is required in order to further our understanding of the processes given rise to biodiversity and to handle the increasing availability of georeferenced molecular datasets. Looking forward, the extensibility of GenGIS makes it an ideal research platform for further developing biogeographic visualization and analyses techniques. This will help ensure GenGIS remains a powerful tool for practicing ecologists.

The increasing number of environmental samples considered in a typical biogeographic study also suggests the continued need for summary statistics of community variation. Given the long-standing use and success of beta-diversity measures, I believe these measures will continue to play a critical role in the field. The work here demonstrates the need to consider multiple measures of beta diversity. With a better understanding of the variation emphasized by alternative measures, we stand to greatly enhance our understanding of how communities are differentiated. The development of network-based measures and their application to studies of human diversity provides a compelling example of how alternative measures of beta diversity may further our understanding into the patterns and processes of biodiversity.

References

Abecasis AB, Lemey P, Vidal N, de Oliveira T, Peeters M, Camacho R, Shapiro B, Rambaut A, Vandamme A-M. 2007. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J. Virol. 81:8543–8551.

Allal F, Sanou H, Millet L, Vaillant A, Camus-Kulandaivelu L, Logossa ZA, Lefèvre F, Bouvet J-M. 2011. Past climate changes explain the phylogeography of Vitellaria paradoxa over Africa. Heredity 107:174–186.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

An W, Telesnitsky A. 2002. HIV-1 genetic recombination: experimental approaches and observations. AIDS Rev. 4:195–212.

Anderson MJ, Crist TO, Chase JM, et al. 2011. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecol. Lett. 14:19–28.

Antonio M, Hakeem I, Awine T, et al. 2008. Seasonality and outbreak of a predominant *Streptococcus pneumoniae* serotype 1 clone from The Gambia: expansion of ST217 hypervirulent clonal complex in West Africa. BMC Microbiol. 8:198.

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Syst. 18:489–522.

Avise JC. 2000. Phylogeography: The History and Formation of Species. Cambridge, Massachusetts: Harvard University Press.

Baird DJ, Hajibabaei M. 2012. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. Mol. Ecol. 21:2039–2044.

Bandelt HJ, Dress AW. 1992. A canonical decomposition theory for metrics on a finite set. Adv. Math. 92:47–105.

Bandelt HJ, Dress AW. 1993. A relational approach to split decomposition. Technical Report, Universität Bielefeld, Bielefeld, Germany.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. Genetics 141:743–753.

Bapteste E, Boucher Y. 2008. Lateral gene transfer challenges principles of microbial systematics. Trends Microbiol. 16:200–207.

Barr JJ, Slater FR, Fukushima T, Bond PL. 2010. Evidence for bacteriophage activity causing community and performance changes in a phosphorus-removal activated sludge. FEMS Microbiol. Ecol. 74:631-642.

Barth W, Jünger M, Mutzel P. 2002. Simple and efficient bilayer cross counting. In: 10th International Symposium on Graph Drawing. London, UK: Springer-Verlag.

Beatty A. 1993. Nias. In: Levinson D, Hockings P, editors. Encyclopedia of world cultures, volume V-East and Southeast Asia. New York: G.K. Hall & Co.

Beiko RG, Doolittle WF, Charlebois RL. 2008. The impact of reticulate evolution on genome phylogeny. Syst. Biol. 57:844–856.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. U.S.A. 102:14332–14337.

Beiko RG, Whalley J, Wang S, Clair H, Smolyn G, Churcher S, Porter M, Blouin C, Brooks S. 2008. Spatial analysis and visualization of genetic biodiversity. In: Free and Open Source Software for Geospatial (FOSS4G) Conference. Cape Town, South Africa.

Béjà O, Aravind L, Koonin EV, et al. 2000. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289:1902–1906.

Béjà O, Spudich EN, Spudich JL, Leclerc M, DeLong EF. 2001. Proteorhodopsin phototrophy in the ocean. Nature 411:786–789.

Bielejec F, Rambaut A, Suchard MA, Lemey P. 2011. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. Bioinformatics 27:2910–2912.

Biers EJ, Sun S, Howard EC. 2009. Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. Appl. Environ. Microbiol. 75:2221–2229.

Bloomquist EW, Lemey P, Suchard MA. 2010. Three roads diverged? Routes to phylogeographic inference. Trends Ecol. Evol. 25:626–632.

Böer SI, Hedtkamp SIC, van Beusekom JEE, Fuhrman JA, Boetius A, Ramette A. 2009. Time- and sediment depth-related variations in bacterial diversity and community structure in subtidal sands. ISME J. 3:780–791.

Bontaz D. 2009. The megaliths in Nias. In Gruber P, Herbig U, editors. Traditional architecture and art on Nias, Indonesia. Vienna, Austria.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast Statistical Alignment. PLoS Comput. Biol. 5:e1000392.

Bray R, Curtis T. 1957. An ordination of the upland forest communities of Southern Wisconsin. Ecol. Monogr. 27:325–349.

Brueggemann AB, Spratt BG. 2003. Geographic distribution and clonal diversity of *Streptococcus pneumoniae* serotype 1 isolates. J. Clin. Microbiol. 41:4966–4970.

Bryant D, Moulton V, Spillner A. 2007. Consistency of the neighbor-net algorithm. Algorithms Mol. Biol. 2:8.

Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Bryant JA, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ, Green JL. 2008. Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. Proc. Natl. Acad. Sci. U.S.A. 105 Suppl. 1:11505–11511.

Buchin K, Buchin J, Byrka J, Nöllenburg M, Okamoto Y, Silveira RI, Wolff A. 2008. Drawing (complete) binary tanglegrams: hardness, approximation, fixed-parameter tractability. In: Graph Drawing. Crete, Greece: Springer-Verlag.

Bwayo J, Plummer F, Omari M, Mutere A, Moses S, Ndinya-Achola J, Velentgas P, Kreiss J. 1994. Human immunodeficiency virus infection in long-distance truck drivers in east Africa. Arch. Intern. Med. 154:1391–1396.

Byington CL, Samore MH, Stoddard GJ, et al. 2005. Temporal trends of invasive disease due to *Streptococcus pneumoniae* among children in the intermountain west: emergence of nonvaccine serogroups. Clin. Infect. Dis. 41:21–29.

Çakıroğlu OA, Erten C, Karataş Ö, Sözdinler M. 2009. Crossing minimization in weighted bipartite graphs. J. Discrete Algorithms 7:439–452.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Camin J, Sokal R. 1965. A method for deducing branching sequences in phylogeny. Evolution 19:311–326.

Caporaso JG, Kuczynski J, Stombaugh J, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7:335–336.

Caporaso JG, Lauber CL, Costello EK, et al. 2011. Moving pictures of the human microbiome. Genome Biol. 12:R50.

Castelloe J, Templeton AR. 1994. Root probabilities for intraspecific gene trees under neutral coalescent theory. Mol. Phylogenet. Evol. 3:102–113.

Clarke KR, Warwick RM. 1998. A taxonomic distinctness index and its statistical properties. J. Appl. Ecol. 35:523–531.

Cock PJA, Antao T, Chang JT, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423.

Cole JR, Wang Q, Cardenas E, et al. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37:D141–145.

Cooper A, Rambaut A, Macaulay V, Willerslev E, Hansen AJ, Stringer C. 2001. Human origins and ancient human DNA. Science 292:1655–1656.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science 326:1694–1697.

Crozier RH, Dunnett LJ, Agapow P-M. 2005. Phylogenetic biodiversity assessment based on systematic nomenclature. Evol. Bioinform. Online 1:11–36.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc. Natl. Acad. Sci. U.S.A. 105:10039–10044.

DeLong EF, Preston CM, Mincer T, et al. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503.

DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, Phan R, Andersen GL. 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res. 34:W394–399.

DeSantis TZ, Hugenholtz P, Larsen N, et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72:5069–5072.

Dethlefsen L, Huse S, Sogin ML, Relman DA. 2008. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol. 6:e280.

Dick CW, Roubik DW, Gruber KF, Bermingham E. 2004. Long-distance gene flow and cross-Andean dispersal of lowland rainforest bees (Apidae: Euglossini) revealed by comparative mitochondrial DNA phylogeography. Mol. Ecol. 13:3775–3785.

Dinsdale EA, Edwards RA, Hall D, et al. 2008. Functional metagenomic profiling of nine biomes. Nature 452:629–632.

Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. Genome Res. 19:744–756.

Doolittle WF, Zhaxybayeva O. 2010. Metagenomics and the Units of Biological Organization. BioScience 60:102–112.

Drake JW. 1993. Rates of spontaneous mutation among RNA viruses. Proc. Natl. Acad. Sci. U.S.A. 90:4171–4175.

Dress AWM, Huson DH. 2004. Constructing splits graphs. IEEE/ACM Trans. Comput. Biol. Bioinform. 1:109–115.

Dujmović V, Fernau H, Kaufmann M. 2008. Fixed parameter algorithms for one-sided crossing minimization revisited. J. Discrete Algorithms 6:313–323.

Dwyer T, Schreiber F. 2004. Optimal leaf ordering for two and a half dimensional phylogenetic tree visualisation. In: Proceedings of the 2004 Australasian symposium on Information Visualisation. Darlinghurst, Australia.

Eades P, Wormald NC. 1994. Edge crossings in drawings of bipartite graphs. Algorithmica 11:379–403.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Enright MC, Spratt BG. 1998. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology 144:3049–3060.

Esteva C, Selva L, de Sevilla MF, Garcia-Garcia JJ, Pallares R, Muñoz-Almagro C. 2011. *Streptococcus pneumoniae* serotype 1 causing invasive disease among children in Barcelona over a 20-year period (1989-2008). Clin. Microbiol. Infect. 17:1441–1444.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform. Online 1:47–50.

Faith DP, Baker AM. 2006. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. Evol. Bioinform. Online 2:121–128.

Faith DP, Lozupone CA, Nipperess D, Knight R. 2009. The Cladistic Basis for the Phylogenetic Diversity (PD) Measure Links Evolutionary Features to Environmental Gradients and Supports Broad Applications of Microbial Ecology's "Phylogenetic Beta Diversity" Framework. Int. J. Mol. Sci. 10:4723–4741.

Farikou O, Thevenon S, Njiokou F, Allal F, Cuny G, Geiger A. 2011. Genetic diversity and population structure of the secondary symbiont of tsetse flies, Sodalis glossinidius, in sleeping sickness foci in Cameroon. PLoS Negl. Trop. Dis. 5:e1281.

Farr TG, Rosen PA, Caro E, et al. 2007. The shuttle radar topography mission. Rev. Geophys. 45:33 pages.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Fernau H, Kaufmann M, Poths M. 2010. Comparing trees via crossing minimization. J. Comput. Syst. Sci. 76:593–608.

Ferrier S, Manion G, Elith J, Richardson K. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Divers. Distrib. 13:252–264.

Fierer N, Jackson RB. 2006. The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. U.S.A. 103:626–631.

Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. Proc. Natl. Acad. Sci. U.S.A. 107:6477–6481.

Flores GE, Bates ST, Knights D, Lauber CL, Stombaugh J, Knight R, Fierer N. 2011. Microbial biogeography of public restroom surfaces. PLoS ONE 6:e28132.

Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH. 2008. A latitudinal diversity gradient in planktonic marine bacteria. Proc. Natl. Acad. Sci. U.S.A. 105:7774–7778.

Gascuel O, Steel M. 2006. Neighbor-joining revealed. Mol. Biol. Evol. 23:1997–2000.

Gaut BS, Lewis PO. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12:152–162.

Gevers D, Cohan FM, Lawrence JG, et al. 2005. Re-evaluating prokaryotic species. Nat. Rev. Microbiol. 3:733–739.

Gifford RJ, de Oliveira T, Rambaut A, Pybus OG, Dunn D, Vandamme A-M, Kellam P, Pillay D. 2007. Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of human immunodeficiency virus type 1. J. Virol. 81:13050–13056.

Gilbert J, Meyer F, Antonopoulos DA, et al. 2010. The terabase metagenomics workshop and the vision of an earth microbiome project. Stand. Genomic Sci. 3:243–248.

Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc. Natl. Acad. Sci. U.S.A. 104:18566–18570.

Gower JC. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338.

Graham CH, Fine PVA. 2008. Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. Ecol. Lett. 11:1265–1277.

Gray RD, Drummond AJ, Greenhill SJ. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science 323:479–483.

Green J, Bohannan BJM. 2006. Spatial scaling of microbial biodiversity. Trends Ecol. Evol. 21:501–507.

Green JL, Bohannan BJM, Whitaker RJ. 2008. Microbial biogeography: from taxonomy to traits. Science 320:1039–1043.

Green Tringe S, von Mering C, Kobayashi A, et al. 2005. Comparative metagenomics of microbial communities. Science 308:554–557.

Hamady M, Lozupone C, Knight R. 2010. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J. 4:17–27.

Hardy OJ, Senterre B. 2007. Characterizing the phylogenetic structure of communities by an additive partitioning of phylogenetic diversity. J. Ecol. 95:493–506.

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. 270:313–321.

Hemelaar J, Gouws E, Ghys PD, Osmanov S. 2006. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS 20:W13–23.

Henriques Normark B, Kalin M, Ortqvist A, et al. 2001. Dynamics of penicillin-susceptible clones in invasive pneumococcal disease. J. Infect. Dis. 184:861–869.

Herrnstadt C, Elson JL, Fahy E, et al. 2002. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am. J. Hum. Genet. 70:1152–1171.

Hijmans R, Guarino L, Cruz M, Rojas E. 2001. Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. Plant Genet. Res. Newsletter 127:15–19.

Hill AW, Guralnick RP. 2010. GeoPhylo: an online tool for developing visualizations of phylogenetic trees in geographic space. Ecography 33:633–636.

Ho SY, Jermiin L. 2004. Tracing the decay of the historical signal in biological sequence data. Syst. Biol. 53:623–637.

Holland B, Conner G, Huber K, Moulton V. 2007. Imputing supertrees and supernetworks from quartets. Syst. Biol. 56:57–67.

Holland B, Delsuc F, Moulton V. 2005. Visualizing conflicting evolutionary hypotheses in large collections of trees: using consensus networks to study the origins of placentals and hexapods. Syst. Biol. 54:66–76.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat. Rev. Genet. 10:639–650.

Holten D, Wijk V, J J. 2008. Visual Comparison of Hierarchically Organized Data. In: Eurographics/IEEE-VGTC Symposium on Visualization. Crete, Greece.

Huber KT, Langton M, Penny D, Moulton V, Hendy M. 2002. Spectronet: a package for computing spectra and median networks. Appl. Bioinformatics 1:159–161.

Huber KT, Moulton V, Lockhart P, Dress A. 2001. Pruned median networks: a technique for reducing the complexity of median networks. Mol. Phylogenet. Evol. 19:302–310.

Hué S, Pillay D, Clewley JP, Pybus OG. 2005. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc. Natl. Acad. Sci. U.S.A. 102:4425–4429.

HUGO Pan-Asian SNP Consortium. 2009. Mapping human genetic diversity in Asia. Science 326:1541–1545.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Huson DH, Dezulian T, Klöpper T, Steel MA. 2004. Phylogenetic super-networks from partial trees. IEEE/ACM Trans. Comput. Biol. Bioinform. 1:151–158.

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:460.

Huson DH, Rupp R, Scornavacca C. 2010. Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge, UK: Cambridge University Press.

Huson DH, Scornavacca C. 2011. A survey of combinatorial methods for phylogenetic networks. Genome Biol. Evol. 3:23–35.

Ingman M, Kaessmann H, Pääbo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713.

Jaccard P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles 37:547–579.

Jacko JA, Sears A eds. 2002. The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications, second edition. 1st ed. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Janies DA, Hill AW, Guralnick R, Habib F, Waltari E, Wheeler WC. 2007. Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). Syst. Biol. 56:321–329.

Janies DA, Treseder T, Alexandrov B, Habib F, Chen JJ, Ferreira R, Çatalyürek Ü, Varón A, Wheeler WC. 2010. The Supramap project: linking pathogen genomes with geography to fight emergent infectious diseases. Cladistics 27:61–66.

Jetz W, McPherson JM, Guralnick RP. 2012. Integrating biodiversity distribution knowledge: toward a global map of life. Trends Ecol. Evol. (Amst.) 27:151–159.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. 2006. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science 311:1737–1740.

Jünger M, Mutzel P. 1996. Exact and heuristic algorithms for 2-layer straightline crossing minimization. Brandenburg F, editor. Graph Drawing 1027:337–348.

Kayser M, Choi Y, van Oven M, Mona S, Brauer S, Trent RJ, Suarkia D, Schiefenhövel W, Stoneking M. 2008. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. Mol. Biol. Evol. 25:1362–1374.

Kembel SW, Jones E, Kline J, Northcutt D, Stenson J, Womack AM, Bohannan BJ, Brown GZ, Green JL. 2012. Architectural design influences the diversity and structure of the built environment microbiome. ISME J. (advanced access): doi:10.1038/ismej.2011.211

Kennedy M, Holland BR, Gray RD, Spencer HG. 2005. Untangling long branches: identifying conflicting phylogenetic signals using spectral analysis, neighbor-net, and consensus networks. Syst. Biol. 54:620–633.

Kidd DM, Liu X. 2008. Geophylobuilder 1.0: an ArcGIS extension for creating "geophylogenies." Mol. Ecol. Resour. 8:88–91.

Kidd DM, Ritchie MG. 2006. Phylogeographic information systems: putting the geography into phylogeography. J. Biogeogr. 33:1851–1865.

Kidd DM. 2010. Geophylogenies and the map of life. Syst. Biol. 59:741–752.

Kluge A, Farris J. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 40:446–457.

Koeppel A, Perry EB, Sikorski J, et al. 2008. Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. Proc. Natl. Acad. Sci. U.S.A. 105:2504–2509.

Koleff P, Gaston KJ, Lennon JJ. 2003. Measuring beta diversity for presence–absence data. J. Anim. Ecol. 72:367–382.

Konstantinidis KT, Ramette A, Tiedje JM. 2006. Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. Appl. Environ. Microbiol. 72:7286–7293.

Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. 2001. Evolutionary and immunological implications of contemporary HIV-1 variation. Br. Med. Bull. 58:19–42.

Kozak KH, Graham CH, Wiens JJ. 2008. Integrating GIS-based environmental data into evolutionary biology. Trends Ecol. Evol. 23:141–148.

Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N, Knight R. 2010. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat. Methods 7:813–819.

Kuiken C, Thakallapalli R, Esklid A, de Ronde A. 2000. Genetic analysis reveals epidemiologic patterns in the spread of human immunodeficiency virus. Am. J. Epidemiol. 152:814–822.

Laffan SW, Lubarsky E, Rosauer DF. 2010. Biodiverse, a tool for the spatial analysis of biological and related diversity. Ecography 33:643–647.

Land A, Doig A. 1960. An automatic method for solving discrete programming problems. Econometrica 28:497–520.

Lane D, Pace B, Olsen G, Stahl D, Sogin ML, Pace N. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc. Natl. Acad. Sci. U.S.A. 82:6955–6959.

Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl. Environ. Microbiol. 75:5111–5120.

Legendre P, Legendre L. 1998. Numerical ecology, second edition. Elsevier Science.

Leimkugel J, Adams Forgor A, Gagneux S, et al. 2005. An outbreak of serotype 1 *Streptococcus pneumoniae* meningitis in northern Ghana with features that are characteristic of Neisseria meningitidis meningitis epidemics. J. Infect. Dis. 192:192–199.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. PLoS Comput Biol 5:e1000520.

Letunic I, Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39:W475–W478.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Li XY, Stallmann M. 2001. New bounds on the barycenter heuristic for bipartite graph drawing. Inform. Process. Lett. 82:293–298.

Liu K, Linder CR, Warnow T. 2011. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. PLoS ONE 6:e27731.

Lombardot T, Kottmann R, Pfeffer H, Richter M, Teeling H, Quast C, Glöckner FO. 2006. Megx.net--database resources for marine ecological genomics. Nucleic Acids Res. 34:D390–393.

Lomolino MV, Riddle BR, Whittaker RJ, Brown JH. 2010. Biogeography, Fourth Edition. Fourth. Sunderland, Massachusetts: Sinauer Associates, Inc.

Loo J-H, Trejaut JA, Yen J-C, Chen Z-S, Lee C-L, Lin M. 2011. Genetic affinities between the Yami tribe people of Orchid Island and the Philippine Islanders of the Batanes archipelago. BMC Genet. 12:21.

Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl. Environ. Microbiol. 73:1576–1585.

Lozupone CA, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. 71:8228–8235.

Lozupone CA, Knight R. 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. U.S.A. 104:11436–11440.

Lozupone CA, Knight R. 2008. Species Divergence and the Measurement of Microbial Diversity. FEMS Microbiol. Rev. 32:557–578.

MacDonald NJ, Parks DH, Beiko R. 2009. SeqMonitor: influenza analysis pipeline and visualization. PLoS Curr. 1:RRN1040.

Maddison D, Maddison W. 2008. Cartographer, a Mesquite package for plotting geographic data. Available from: mesquiteproject.org/packages/cartographer.

Magurran AE. 2004. Measuring biological diversity. Oxford, UK: Blackwell Publishing.

Mahecha MD, Martínez A, Lischeid G, Beck E. 2007. Nonlinear dimensionality reduction: Alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. Ecol. Inform. 2:138–149.

Man D, Wang W, Sabehi G, Aravind L, Post AF, Massana R, Spudich EN, Spudich JL, Béjà O. 2003. Diversification and spectral tuning in marine proteorhodopsins. EMBO J. 22:1725–1731.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27:209–220.

Margos G, Gatewood AG, Aanensen DM, et al. 2008. MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. U.S.A. 105:8730–8735.

Marimon JM, Ercibengoa M, Alonso M, Zubizarreta M, Pérez-Trallero E. 2009. Clonal structure and 21-year evolution of *Streptococcus pneumoniae* serotype 1 isolates in northern Spain. Clin. Microbiol. Infect. 15:875–877.

Martin AP. 2002. Phylogenetic approaches for describing and comparing the diversity of microbial communities. Appl. Environ. Microbiol. 68:3673–3682.

Martiny JBH, Bohannan BJM, Brown JH, et al. 2006. Microbial biogeography: putting microorganisms on the map. Nat. Rev. Microbiol. 4:102–112.

Matsen F, Hoffman N, Evans S. 2010. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. arXiv 1107.5095v1.

McChlery SM, Scott KJ, Clarke SC. 2005. Clonal analysis of invasive pneumococcal isolates in Scotland and coverage of serotypes by the licensed conjugate polysaccharide pneumococcal vaccine: possible implications for UK vaccine policy. Eur. J. Clin. Microbiol. Infect. Dis. 24:262–267.

Metro Engineering Inc. 1993. Halifax Harbour Cleanup Project Pre-Design Engineering. Halifax, NS.

Minh BQ, Klaere S, von Haeseler A. 2009. Taxon selection under split diversity. Syst. Biol. 58:586–594.

Mitra S, Gilbert JA, Field D, Huson DH. 2010. Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. ISME J. 4:1236–1242.

Moritz C, Schneider CJ, Wake DB. 1992. Evolutionary relationships within the *Ensatina eschscholtzii* complex confirm the ring species interpretation. Syst. Biol. 41:273–291.

Morrison DA. 2005. Networks in phylogenetic analysis: new tools for population biology. Int. J. Parasitol. 35:567–582.

Morrison DA. 2011. Introduction to phylogenetic networks. Uppsala, Sweden: RJR Productions.

Muñoz X, Unger W, Vrto I. 2002. One sided crossing minimization is NP-hard for sparse graphs. In: 9th International Symposium on Graph Drawing. London, UK.

Munzner T, Guimbretière F, Tasiran S, Zhang L, Zhou Y. 2003. TreeJuxtaposer: scalable tree comparison using Focus+Context with guaranteed visibility. In: ACM SIGGRAPH 2003. New York, NY.

Nagamochi H. 2005. An improved bound on the one-sided minimum crossing number in two-layered drawings. Discrete Comput. Geom. 33:569–591.

Nasidze I, Li J, Quinque D, Tang K, Stoneking M. 2009. Global diversity in the human salivary microbiome. Genome Res. 19:636-643.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.

Nipperess DA, Faith DP, Barton K. 2010. Resemblance in phylogenetic diversity among ecological assemblages. J. Veg. Sci. 21:809–820.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–217.

Obando I, Muñoz-Almagro C, Arroyo LA, et al. 2008. Pediatric parapneumonic empyema, Spain. Emerging Infect. Dis. 14:1390–1397.

van Oven M, Hämmerle JM, van Schoor M, et al. 2011. Unexpected island effects at an extreme: reduced Y chromosome and mitochondrial DNA diversity in Nias. Mol. Biol. Evol. 28:1349–1361.

Pace N, Stahl D, Lane D, Olsen G. 1984. The analysis of microbial populations by ribosomal RNA sequences. Adv. Microb. Ecol. 9:1–55.

Page RDM. 2012. Space, time, form: viewing the Tree of Life. Trends Ecol. Evol. 27:113–120.

Pakendorf B, Stoneking M. 2005. Mitochondrial DNA and human evolution. Annu. Rev. Genom. Hum. G. 6:165–183.

Palmer, SE. 1999. Vision science: photons to phenomenology. Cambridge, Massachusetts: MIT Press.

Parks DH, Beiko RG. 2009. Quantitative visualizations of hierarchically organized data in a geographic context. In: 17th International Conference on Geoinformatics. Fairfax, VA.

Parks DH, Beiko RG. 2012a. Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities. ISME J., accepted July 2012.

Parks DH, Beiko RG. 2012b. Measuring community similarity with phylogenetic networks. Mol. Biol. Evol., accepted July 2012.

Parks DH, MacDonald N, Beiko RG. 2009. Tracking the evolution and geographic spread of Influenza A. PLoS Curr. 1:RRN1014.

Parks DH, Mankowski T, Porter MS, Beiko RG. 2012. GenGIS 2: Geospatial analysis of genetic and genomic datasets, with new gradient algorithms and an extensible framework. In preparation.

Parks DH, Porter M, Churcher S, Wang S, Blouin C, Whalley J, Brooks S, Beiko RG. 2009. GenGIS: a geospatial information system for genomic data. Genome Res. 19:1896–1904.

Pavlopoulos GA, Soldatos TG, Barbosa-Silva A, Schneider R. 2010. A reference guide for tree analysis and visualization. BioData Min. 3:1.

Peeters M, Toure-Kane C, Nkengasong JN. 2003. Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials. AIDS 17:2547–2560.

Piel W, Chan L, Dominus M, Ruan J, Vos R, Tannen V. 2009. TreeBASE v. 2: A Database of Phylogenetic Knowledge. In: e-BioSphere 2009. London, UK.

Pielou EC. 1984. The interpretation of ecological data: a primer of classification and ordination. John Wiley and Sons.

Poczai P, Hyvönen J, Symon DE. 2011. Phylogeny of kangaroo apples (Solanum subg. Archaesolanum, Solanaceae). Mol. Biol. Rep. 38:5243–5259.

Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. 54:396–402.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26:1641–1650.

Pushker R, D'Auria G, Alba-Casado JC, Rodríguez-Valera F. 2005. Micro-Mar: a database for dynamic representation of marine microbial biodiversity. BMC Bioinformatics 6:222.

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Ratnasingham S, Hebert PDN. 2007. BOLD: the barcode of life data system (http://www.barcodinglife.org). Mol. Ecol. Notes 7:355–364.

Ricotta C, Burrascano S. 2008. Beta diversity for functional ecology. Presilia 80:61–71.

Rissler LJ, Apodaca JJ. 2007. Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*). Syst. Biol. 56:924–942.

Robbins KE, Kostrikis LG, Brown TM, Anzala O, Shin S, Plummer FA, Kalish ML. 1999. Genetic analysis of human immunodeficiency virus type 1 strains in Kenya: a comparison using phylogenetic analysis and a combinatorial melting assay. AIDS Res. Hum. Retroviruses 15:329–335.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Roos M, Rothe J. 2010. Introduction to Computational Complexity. In: Mathematical Programming Glossary. INFORMS Computing Society. p. 1–23.

Root HT, Nelson PR. 2011. Does phylogenetic distance aid in detecting environmental gradients related to species composition? J. Veg. Sci. 22:1143–1148.

Rousk J, Bååth E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG, Knight R, Fierer N. 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J. 4:1340–1351.

Roweis ST, Saul LK. 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326.

Rusch DB, Halpern AL, Sutton G, et al. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5:e77.

Ruzzante DE, Walde SJ, Macchi PJ, Alonso M, Barriga JP. 2011. Phylogeography and phenotypic diversification in the Patagonian fish *Percichthys trucha*: the roles of Quaternary glacial cycles and natural selection. Biol. J. Linn. Soc. 103:514–529.

Sabehi G, Kirkup BC, Rozenberg M, Stambler N, Polz MF, Béjà O. 2007. Adaptation and spectral tuning in divergent marine proteorhodopsins from the eastern Mediterranean and the Sargasso Seas. ISME J. 1:48–55.

Sabehi G, Loy A, Jung K-H, Partha R, Spudich JL, Isaacson T, Hirschberg J, Wagner M, Béjà O. 2005. New insights into metabolic properties of marine bacteria encoding proteorhodopsins. PLoS Biol. 3:e273.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Schloss PD, Handelsman J. 2004. Status of the microbial census. Microbiol. Mol. Biol. Rev. 68:686–691.

Schloss PD, Handelsman J. 2006. Introducing TreeClimber, a test to compare microbial community structures. Appl. Environ. Microbiol. 72:2379–2384.

Schloss PD, Westcott SL, Ryabin T, et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75:7537–7541.

Schloss PD, Westcott SL. 2011. Assessing and improving methods used in OTU-based approaches for 16S rRNA gene sequence analysis. Appl. Environ. Microbiol. 77:3219–3226.

Schloss PD. 2008. Evaluating different approaches that test whether microbial communities have the same structure. ISME J. 2:265–275.

Schoville SD, Roderick GK. 2010. Evolutionary diversification of cryophilic *Grylloblatta* species (Grylloblattodea: Grylloblattidae) in alpine habitats of California. BMC Evol. Biol. 10:163.

Serwadda D, Mugerwa RD, Sewankambo NK, et al. 1985. Slim disease: a new disease in Uganda and its association with HTLV-III infection. Lancet 2:849–852.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. 2007. CAMERA: a community resource for metagenomics. PLoS Biol. 5:e75.

Shafer A, White K, Côté S, Coltman D. 2011. Deciphering translocations from relicts in Baranof Island mountain goats: is an endemic genetic lineage at risk? Conserv. Genet. 12:1261–1268.

Shapiro LH, Strazanac JS, Roderick GK. 2006. Molecular phylogeny of Banza (Orthoptera: Tettigoniidae), the endemic katydids of the Hawaiian Archipelago. Mol. Phylogenet. Evol. 41:53–63.

Sharma AK, Sommerfeld K, Bullerjahn GS, Matteson AR, Wilhelm SW, Jezbera J, Brandt U, Doolittle WF, Hahn MW. 2009. Actinorhodopsin genes discovered in diverse freshwater habitats and among cultivated freshwater Actinobacteria. ISME J. 3:726–737.

Sharma AK, Zhaxybayeva O, Papke RT, Doolittle WF. 2008. Actinorhodopsins: proteorhodopsin-like gene sequences found predominantly in non-marine environments. Environ. Microbiol. 10:1039–1056.

Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC, Martiny JBH. 2008. It's all relative: ranking the diversity of aquatic bacterial communities. Environ. Microbiol. 10:2200–2210.

Simmons SL, Dibartolo G, Denef VJ, Goltsman DSA, Thelen MP, Banfield JF. 2008. Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. PLoS Biol. 6:e177.

Slater FR, Johnson CR, Blackall LL, Beiko RG, Bond PL. 2010. Monitoring associations between clade-level variation, overall community structure and ecosystem function in enhanced biological phosphorus removal (EBPR) systems using terminal-restriction fragment length polymorphism (T-RFLP). Water Res. 44: 4908-4923.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J. Mol. Biol. 147:195–197.

Soares MA. 2007. Geographical biases and convenience sampling in HIV molecular epidemiology estimates. AIDS 21:2359–2360.

Soares P, Trejaut JA, Loo J-H, et al. 2008. Climate change and postglacial human dispersals in southeast Asia. Mol. Biol. Evol. 25:1209–1218.

Spencer M, Susko E, Roger AJ. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. 22:1161–1164.

Spillner A, Nguyen BT, Moulton V. 2008. Computing phylogenetic diversity for split systems. IEEE/ACM Trans. Comput. Biol. Bioinform. 5:235–244.

Stackebrandt E, Goebel BM. 1994. A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. 44:846–849.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stearns JC, Lynch MDJ, Senadheera DB, Tenenbaum HC, Goldberg MB, Cvitkovitch DG, Croitoru K, Moreno-Hagelsieb G, Neufeld JD. 2011. Bacterial biogeography of the human digestive tract. Sci. Rep. 1: doi:10.1038/srep00170.

Stebbins RC. 1949. Speciation in salamanders of the plethodontid genus Ensatina. Berkeley, CA: Univ. of California Press.

Swenson NG, Anglada-Cordero P, Barone JA. 2011. Deterministic tropical tree community turnover: evidence from patterns of functional beta diversity along an elevational gradient. Proc. R. Soc. B 278:877–884.

Swenson NG. 2011. Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. PLoS ONE 6:e21264.

Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM. 2008. The challenge of HIV-1 subtype diversity. N. Engl. J. Med. 358:1590–1602.

Tenenbaum JB, Silva VD, Langford JC. 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Tollis IG, Battista GD, Eades P, Tamassia R. 1998. Graph drawing: algorithms for the visualization of graphs. 1st ed. New Jersey, USA: Prentice Hall.

de la Torre JR, Christianson LM, Béjà O, Suzuki MT, Karl DM, Heidelberg J, DeLong EF. 2003. Proteorhodopsin genes are distributed among divergent marine bacterial taxa. Proc. Natl. Acad. Sci. U.S.A. 100:12830–12835.

Tovey CA. 2002. Tutorial on computational complexity. Interfaces 32:30–61.

Tucker KP, Parsons R, Symonds EM, Breitbart M. 2011. Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. ISME J. 5:822–830.

Tuomisto H. 2010a. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography 33:2–22.

Tuomisto H. 2010b. A diversity of beta diversities: straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. Ecography 33:23–45.

Turnbaugh PJ, Hamady M, Yatsunenko T, et al. 2009. A core gut microbiome in obese and lean twins. Nature 457:480–484.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The Human Microbiome Project. Nature 449:804–810.

Ubeda C, Taur Y, Jenq RR, et al. 2010. Vancomycin-resistant Enterococcus domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. J. Clin. Invest. 120:4332–4341.

Vasan A, Renjifo B, Hertzmark E, Chaplin B, Msamanga G, Essex M, Fawzi W, Hunter D. 2006. Different rates of disease progression of HIV type 1 infection in Tanzania based on infecting subtype. Clin. Infect. Dis. 42:843–852.

Venkatachalam B, Apple J, St. John K, Gusfield D. 2010. Untangling Tanglegrams: Comparing Trees by Their Drawings. IEEE/ACM Trans. Comput. Biol. Bioinform. 7:588–597.

Venter JC, Remington K, Heidelberg JF, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74.

Walker RS, Wichmann S, Mailund T, Atkisson CJ. 2012. Cultural phylogenetics of the Tupi language family in Lowland South America. PLoS ONE 7:e35025.

Wang J, Soininen J, Zhang Y, Wang B, Yang X, Shen J. 2011. Contrasting patterns in elevational diversity between microorganisms and macroorganisms. J. Biogeogr. 38:595–603.

Wang L, Jiang T. 1994. On the complexity of multiple sequence alignment. J. Comput. Biol. 1:337–348.

Ware C. 2004. Information visualization: perception for design, second edition. 2nd ed. San Francisco, CA: Morgan Kaufmann.

Webb CO, Ackerly DD, Kembel SW. 2008. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24:2098–2100.

Whalley J, Brooks S, Beiko RG. 2009. Radié: visualizing taxon properties and parsimonious mappings using a radial phylogenetic tree. Bioinformatics 25:672–673.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

Whittaker R. 1972. Evolution and measurement of species diversity. Taxon 21:213–251.

Yilmaz P, Kottmann R, Field D, et al. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 29:415–420.

Yooseph S, Sutton G, Rusch DB, et al. 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. 5:e16.

Yutin N, Suzuki MT, Teeling H, Weber M, Venter JC, Rusch DB, Béjà O. 2007. Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. Environ. Microbiol. 9:1464–1475.

Zemlickova H, Jakubu V, Urbaskova P, Motlova J, Musilek M, Adamkova V. 2010. Serotype-Specific Invasive Disease Potential of *Streptococcus pneumoniae* in Czech Children. J. Med. Microbiol. 59:1079–1083.

Zhang Y, Gladyshev VN. 2008. Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project. PLoS Genet. 4:e1000095.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 16:1099–1108.

Appendix A

Global Ocean Sampling Expedition

Table A.1. Number of 16S rRNA gene sequences from sample sites before and after dereplication.

Sample Site	No. of Sequences	No. of Sequences after Dereplication
GS002	315	88
GS003	175	68
GS004	200	60
GS005	118	40
GS006	106	39
GS007	188	51
GS008	639	190
GS009	181	54
GS010	307	96
GS011	344	136
GS012	322	109
GS013	310	110
GS014	346	133
GS015	314	116
GS016	321	107
GS017	678	220
GS018	396	168
GS019	393	141
GS020	425	173
Min	**106**	**39**
Total	**6028**	**2099**

Table A.2. Model fit (R^2) and statistical significance of regression models of OTU counts versus latitude for 4 different thresholds.

Sample Set	Clustering Level	R^2 value	p-value
Full	Identical	20.7%	0.0288
Full	97%	37.5%	0.0032
Full	95%	29.3%	0.0098
Full	90%	23.0%	0.0218
Reduced	Identical	26.8%	0.0335
Reduced	97%	22.0%	0.0515
Reduced	95%	10.0%	0.1434
Reduced	90%	5.4%	0.2108

Table A.3. Model fit (R^2) and statistical significance of regression models of taxonomic richness counts versus latitude at various taxonomic ranks.

Sample Set	Clustering Level	R^2 value	p-value
Full	Species	41.3%	0.0030
Full	Genus	49.0%	0.0009
Full	Family	56.5%	0.0002
Full	Order	61.7%	0.0001
Full	Class	34.5%	0.0082
Full	Phylum	51.2%	0.0006
Reduced	Species	40.7%	0.0141
Reduced	Genus	51.6%	0.0038
Reduced	Family	65.3%	0.0005
Reduced	Order	63.5%	0.0006
Reduced	Class	26.4%	0.0611
Reduced	Phylum	37.9%	0.0191

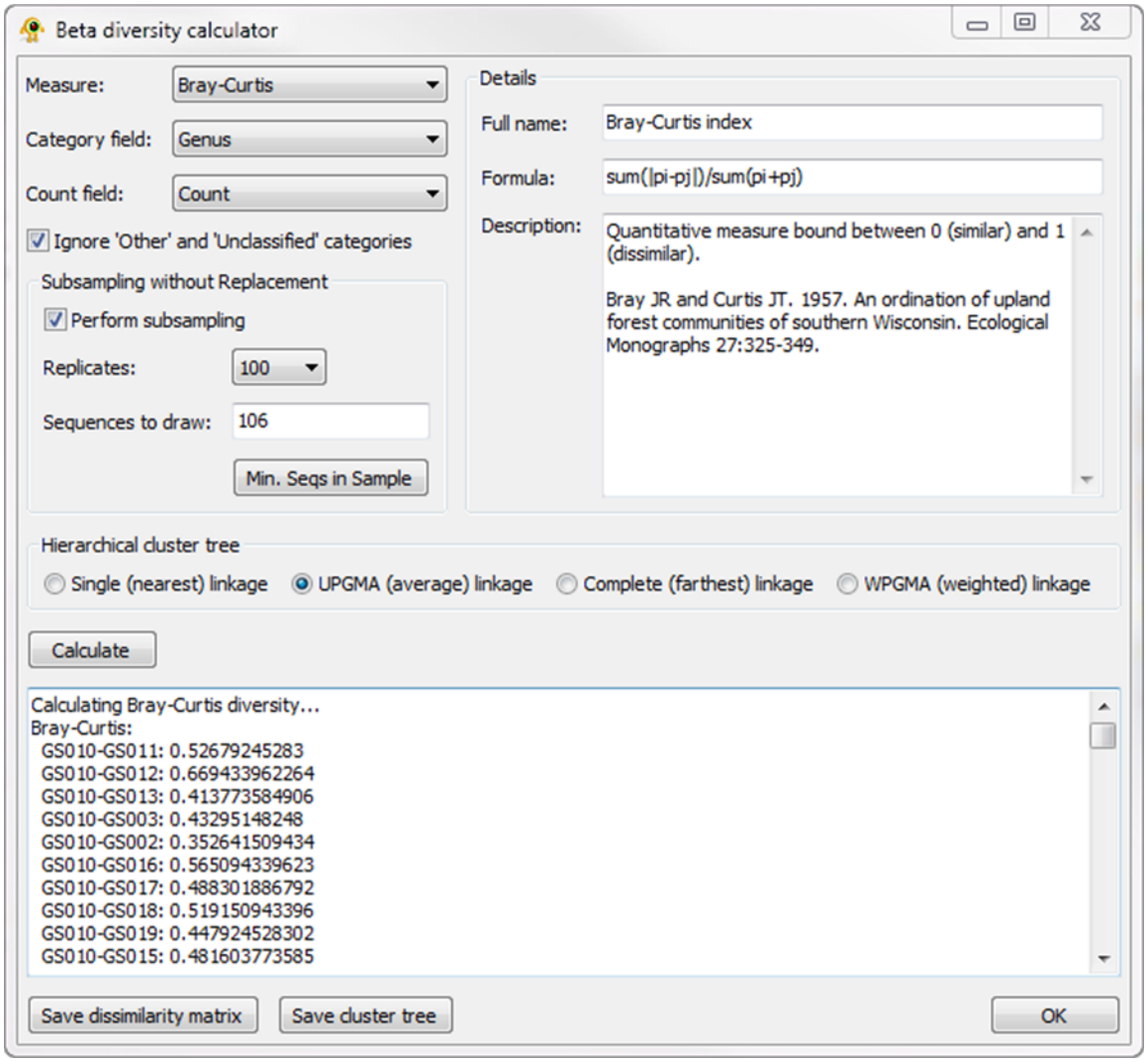


Figure A.1. Exploring the taxonomic similarity of GOS communities. The Beta-Diversity Calculator plugin in GenGIS is used to calculate beta-diversity indices, subsample communities, and generate hierarchal cluster trees.

Appendix B

Non-recombinant HIV Sequences

Table B.1. Country codes and non-recombinant HIV sequence counts for each country in the African dataset acquired from the HIV Sequence Database. C.-K. = Democratic Republic of the Congo, C.A.R. = Central African Republic, C.-B. = Republic of Congo.

Country Code	Country Name	Count of non-recombinant subtype									
		A	B	C	D	F	G	H	J	K	Sum
AO	Angola	116	4	60	21	35	42	45	16	0	339
BF	Burkina Faso	10	2	0	2	0	2	0	0	0	16
BI	Burundi	36	0	278	5	0	1	2	0	0	322
BJ	Benin	33	0	0	0	0	14	0	0	0	47
BW	Botswana	0	0	294	0	0	0	0	0	0	294
CD	C.-K.	543	1	132	209	53	252	91	46	28	1355
CF	C.A.R.	194	3	1	8	3	24	6	8	0	247
CG	C.-B.	22	3	0	3	1	19	8	3	0	59
CI	Cote d'Ivoire	138	7	0	9	0	7	0	0	0	161
CM	Cameroon	420	15	17	267	197	224	32	12	8	1192
DJ	Djibouti	3	0	57	5	0	0	0	0	0	65
DZ	Algeria	4	225	0	6	1	10	1	0	0	247
EG	Egypt	0	22	0	0	0	0	0	0	0	22
ET	Ethiopia	5	3	832	4	0	0	0	0	0	844
GA	Gabon	55	3	8	23	4	18	5	2	0	118
GH	Ghana	249	2	18	5	0	110	0	0	0	384
GM	Gambia	3	2	5	1	0	4	0	2	0	17
GN	Guinea	5	1	0	0	0	6	0	0	0	12
GQ	Equ. Guinea	9	2	11	10	4	0	2	0	0	38
KE	Kenya	4897	1	523	690	0	44	0	0	0	6155
MG	Madagascar	7	7	5	0	0	0	0	0	0	19
ML	Mali	11	2	3	1	2	9	0	0	0	28
MW	Malawi	10	0	828	13	0	0	0	0	0	851
MZ	Mozambique	0	0	487	3	1	3	0	0	0	494
NE	Niger	3	0	0	1	0	9	0	0	0	13
NG	Nigeria	16	2	7	3	0	169	0	2	0	199
RE	Reunion	5	7	0	0	0	0	0	0	0	12
RW	Rwanda	831	7	45	0	1	0	0	0	0	884
SC	Seychelles	52	27	33	3	0	0	0	0	0	115
SD	Sudan	3	2	12	22	0	0	0	0	0	39
SN	Senegal	88	29	39	30	1	17	2	2	0	208
SO	Somalia	0	0	21	0	0	0	0	0	0	21
SZ	Swaziland	0	0	47	0	0	0	0	0	0	47
TD	Chad	47	0	0	45	3	12	0	0	0	107
TN	Tunisia	0	23	0	0	0	0	0	0	0	23
TZ	Tanzania	1423	9	832	746	0	0	0	0	0	3010
UG	Uganda	1642	2	162	1494	0	11	0	0	0	3311
ZA	South Africa	37	102	6405	23	1	7	1	0	0	6576
ZM	Zambia	10	0	1657	3	0	28	0	2	0	1700
ZW	Zimbabwe	0	0	411	0	0	0	0	0	0	411
	Sum	**10927**	**515**	**13230**	**3655**	**307**	**1042**	**195**	**95**	**36**	**30002**

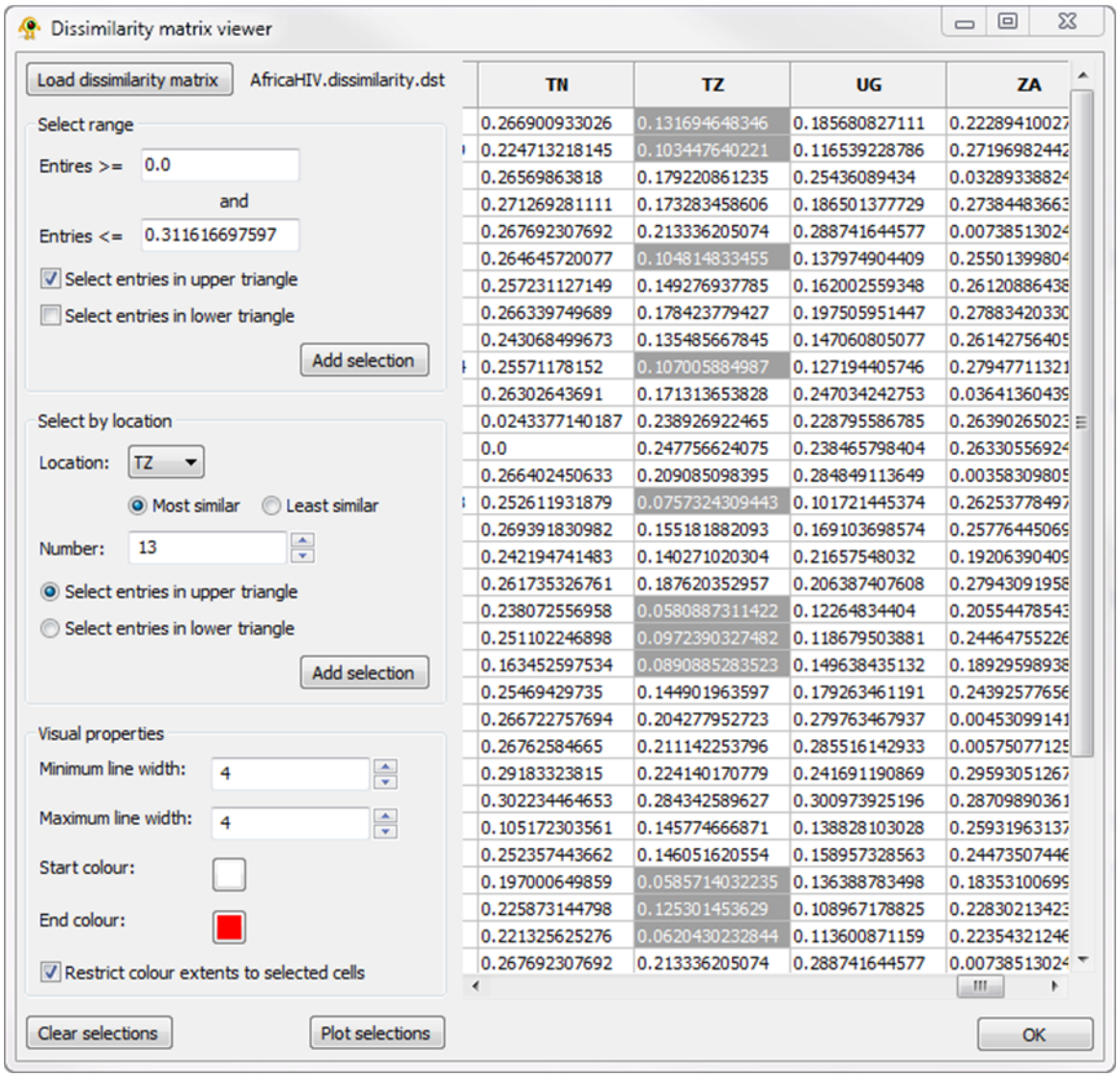


Figure B.1. Dissimilarity-Matrix Viewer plugin for GenGIS. Cells within the dissimilarity matrix can either be manually selected, selected by specifying a range of values, or selected based on the similarity of locations to a location of interest. A line will be drawn between each pair of locations specified by a selected cell. The colour and width of these lines can be configured by the user. In this example, the 13 locations most similar to Tanzania (TZ) were selected.

Appendix C

Comparison of Phylogenetic Beta-diversity Measures

Methods C.1 Deriving Phylogenetic Beta-diversity Measures

In this manuscript, we consider previously proposed measures of phylogenetic beta-diversity along with phylogenetic extensions of commonly used taxon-based measures. Phylogenetic extensions of all qualitative measures and a few of the quantitative measures can be derived using the framework proposed by Nipperess et al. (2010), whereas other quantitative measures were derived here (see Table 4.1). In general, a given taxon-based measure can be extended to incorporate phylogenetic information in multiple ways and many of these will result in reasonable measures of phylogenetic beta-diversity. We have opted to extend measures such that branch lengths provide an *unscaled* weighting of community proportions. For example, the Euclidean distance between two vectors p_i and p_j is:

$$\sqrt{\sum_n \left(p_{in} - p_{jn}\right)^2}$$

In a phylogenetic framework, each vector indicates the proportion of sequences descendant from a given branch and our goal is to weight the calculated distance between vectors by the length of each branch, W_n. This can be done in at least two ways:

1) $$\sqrt{\sum_n \left(W_n p_{in} - W_n p_{jn}\right)^2} = \sqrt{\sum_n W_n^2 \left(p_{in} - p_{jn}\right)^2}$$

2) $$\sqrt{\sum_n W_n \left(p_{in} - p_{jn}\right)^2}$$

We have selected to investigate the second definition as this does not scale branches. This is a sound strategy as there is no evidence or reason to believe that scaling branch lengths will provide superior measures of beta-diversity. Moreover, the second definition often results in well-known weighted measures as exemplified here where the second definition is known as the weighted Euclidean distance.

The exception to this rule is the two proposed correlation measures. It is unclear how best to add in branch lengths to Pearson's correlation measure so two distinct measures

were considered: weighted correlation and Pearson dissimilarity. The first has been proposed as a method for accounting for repeated observations (Bland and Altman, 1995) whereas the second is a straight-forward derivation. For each community create a "branch" vector where each branch in the phylogeny defines an element in the vector whose value is the proportion of sequences descendant from the branch multiplied by the length of the branch. The Pearson dissimilarity between a pair of communities is one minus the correlation between their "branch" vectors.

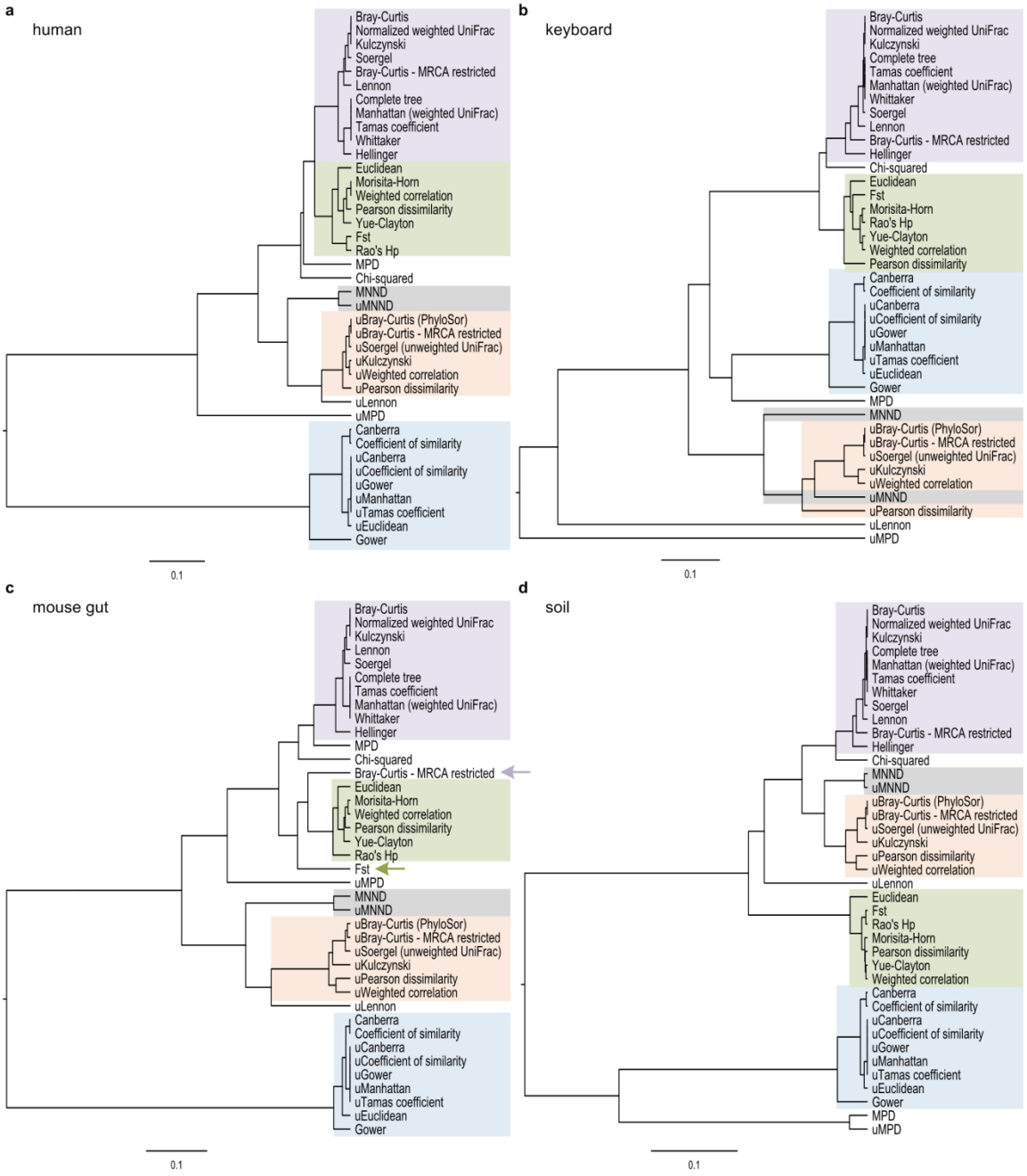


Figure C.1. Hierarchical cluster trees indicating the mean correlation of phylogenetic beta-diversity measures on the (a) human, (b) keyboard, (c) mouse gut, and (d) soil datasets. Trees were inferred using the UPGMA criterion. Branch lengths are transformed Pearson's *r* values, *d*=*r*–1, averaged over 100 random subsets of samples. Pearson's correlation coefficients between the mean correlation matrices are: human vs. keyboard, *r*=0.47; human vs. mouse gut, *r* = 0.89; human vs. soil, *r*=0.79; keyboard vs. mouse gut, *r*=0.45; keyboard vs. soil, *r* = 0.74; and mouse gut vs. soil, *r*=0.73. The five most highly correlated and consistently clustered groups of measures are highlighted in different colours. Colours correspond to those used in Figure 4.2.

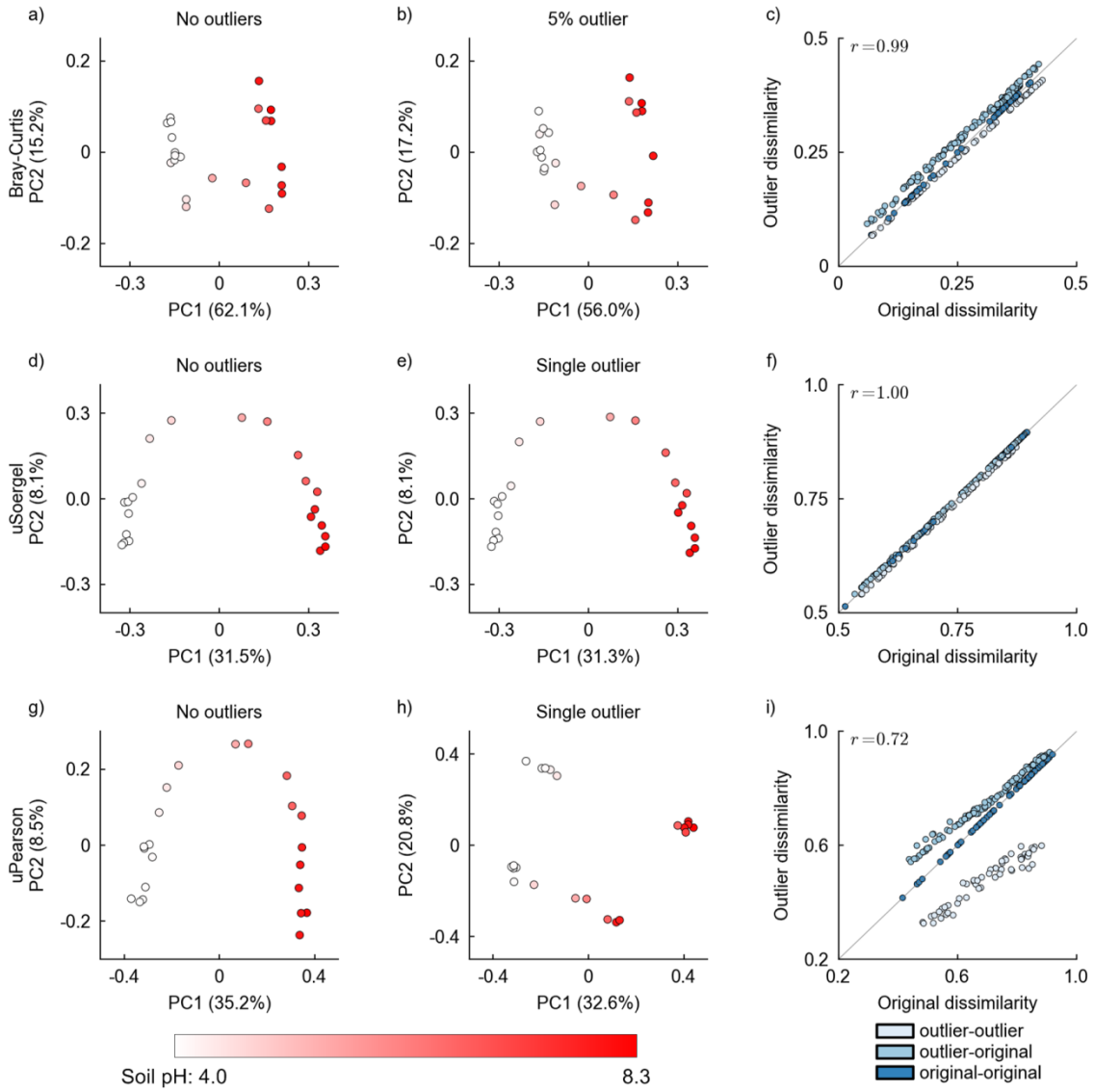


Figure C.2. Recovery of gradients is influenced by a measure's robustness to outlying basal lineages. (**a**-**i**) The quantitative Bray-Curtis (a-c), qualitative Soergel (d-f), and qualitative Pearson dissimilarity (g-i) measures were applied to the soil dataset. All 3 methods reveal a pH gradient (a, d, g). (**b**, **e**, **h**) The addition of an outlying basal lineage to half the samples did not significantly affect the Bray-Curtis (b: 5% of sequences assigned to the outlying lineage) or Soergel (e) measures, but obscured the underlying pH gradient for the Pearson dissimilarity (h) measure. (**c**, **f**, **i**) Each data point in the scatter plots indicates the dissimilarity measures between a pair of samples before (x-axis) and after (y-axis) adding sequences to the outlying lineage. Pearson's correlation coefficient, *r*, between dissimilarity values measured before and after addition of the outlying lineage is given in the upper-left corner of each scatter plot.

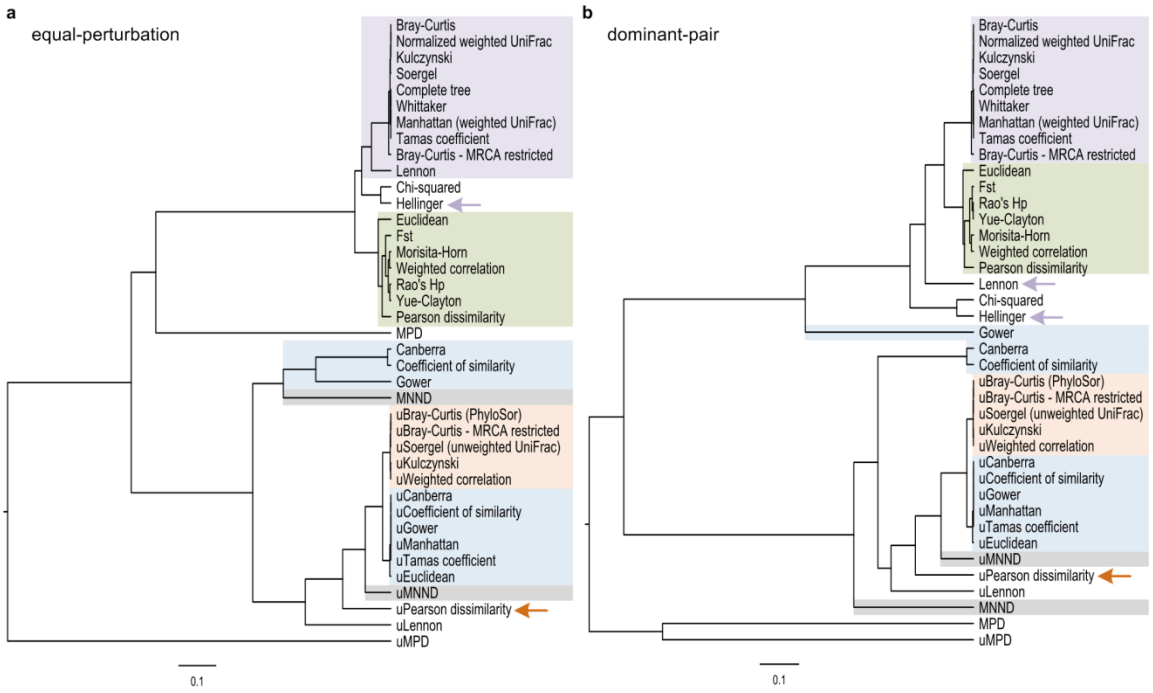


Figure C.3. Correlation between measures under the equal-perturbation (a) and dominant-pair (b) models of differentiation. Hierarchical cluster trees were inferred using the UPGMA criterion. Branch lengths are transformed Pearson's *r* values, $d = r - 1$, averaged over 100 independent simulations. The equal-perturbation simulations were performed with σ_1=1.0 and σ_2=0.5, and the dominant-pair simulations with d=1.797. For both models, simulations were performed with 1,000 sequences being drawn per sample. Colours correspond to those used in Figure 4.2 to highlight highly correlated and consistently clustered groups of measures.

Table C.1. References for phylogenetic beta-diversity measures.

Quantitative Measure	References	Qualitative Measure	References
•Bray-Curtis •Percentage difference	1-3	•Sørensen (complement) •PhyloSor •Dice's index (complement)	1,-6
•Bray-Curtis (MRCA restricted)	-	-	-
•Canberra	1,7,8	•Canberra	9
•Chi-squared	1	-	-
•Coefficient of similarity (complement)	10	•Coefficient of similarity (complement)	-
•Complete tree	(proposed here)	-	-
•Euclidean •Weighted Euclidean	1	•Euclidean	9
•F_{ST} •P_{ST}	11,12,13	-	-
•Gower (complement)	1,14	•Gower (complement)	-
•Hellinger	15,16	-	-
•Kulczynski (complement)	1	•Kulczynski-Cody •Sokal-Sneath (complement)	1,17
•Lennon compositional difference	18	•Lennon compositional difference	19,20
•Manhattan •Weighted UniFrac	1,21	•Hamming distance	-
•Mean nearest neighbour distance (MNND)	22	•MNND	-
•Mean phylogenetic distance (MPD) •Rao's D_p	22	•MPD	
•Morisita-Horn	23,24	-	-
•Normalized weighted UniFrac	21	-	-
•Pearson dissimilarity	25	•Pearson dissimilarity	-
•Rao's H_p	22	-	-
•Soergel •Ružička (complement) •Percentage remoteness	26,27	•Jaccard (complement) •Unweighted UniFrac	1,19,28
•Tamàs coefficient	18	•Simple matching coefficient (complement)	1
•Weighted correlation (complement)	29	•Weighted correlation (complement)	-
•Whittaker index of association (complement)	1,30	-	-
•Yue-Clayton (complement)	31	-	-

1. Legendre P, Legendre L. 1998. *Numerical Ecology*, 2nd English edn. Elsevier: Amsterdam, The Netherlands.

2. Bray JR, Curtis JT. 1957. An ordination of upland forest communities of southern Wisconsin. Ecol. Monogr. 27:325-349.

3. Odum EP. 1950. Bird populations of the Highlands (North Carolina) plateau in relation to plan succession and avian invasion. Ecology 41:395-399.

4. Sørensen T. 1948. A method of establishing groups of equal amplitude in plan sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biol. Skr. 5:1-34.

5. Bryant JA, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ, Green JL. 2008. Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. Proc. Natl. Acad. Sci. U.S.A. 105:11505-11511.

Table C.1. References for phylogenetic beta-diversity measures (continued).

6. Dice LR. 1945. Measures of the amount of ecologic association between species. Ecology 26:297-302.

7. Lance GN, Williams WT. 1967. Mixed-data classificatory programs. II. Divisive systems. Aust. Computer. J. 1:82-85.

8. Faith DP, Minchin PR, Belbin L 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69:57-68.

9. Greig-Smith, P. 1983. Quantitative plant ecology. Blackwell Scientific Publications.

10. Pinkham CFA, Pearson JO. 1976. Applications of a new coefficient of similarity to pollution surveys. J. Water. Poll. Contr. Fed. 48:717-723.

11. Martin AP. 2002. Phylogenetic approaches for describing and comparing the diversity of microbial communities. Appl. Environ. Microbiol. 68:3673-3682.

12. Lozupone CA, Knight R. 2008. Species divergence and the measurement of microbial diversity. FEMS Microbiol. Rev. 32:557-578.

13. Hardy OJ, Senterre B. 2007. Characterizing the phylogenetic structure of communities by an additive partitioning of phylogenetic diversity. J. Ecol .95:493-506.

14. Gower JC. 1971. A general coefficient of similarity and some of its properties. Biometrics 23:623-637.

15. Rao CR. 1995. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Qüestiió 19:2363.

16. Legendre P, Gallagher, E. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129:271-280.

17. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75: 7537-7541.

18. Nipperess DA, Faith DP, Barton K. 2010. Resemblance in phylogenetic diversity among ecological assemblages. J. Veg. Sci .21:809-820.

19. Koleff P, Gaston KJ, Lennon JJ. 2003. Measuring beta diversity for presence-absence data. J. Anim. Ecol. 72:367-382.

20. Lennon JJ, Koleff P, Greenwood JDD, Gaston KJ. 2001. The geographical structure of British bird distributions: diversity, spatial turnover and scale. J. Anim. Ecol. 70: 966-979.

21. Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Appl. Environ. Microbiol. 73:1576-1585.

22. Webb CO, Ackerly DD, Kembel SW. 2008. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24:2098-2100.

Table C.1. References for phylogenetic beta-diversity measures (continued).

23. Magurran AE. 2004. Measuring biological diversity. Blackwell Publishing.

24. Wolda, H. 1981. Similarity indices, sample size and diversity. Oecologia 50:296-302.

25. Rodgers JL, Nicewander WA. 1988. Thirteen ways to look at the correlation coefficient. The American Statistician 42:59-66.

26. Fechner U, Schneider G. 2004. Evaluation of distance metrics for ligand-based similarity searching. ChemBioChem 5:538.

27. Pielou EC. 1984. The interpretation of ecological data. John Wiley & Sons.

28. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. 71:8228-8235.

29. Bland MJ, Altman DG. 1995. Calculating correlation coefficients with repeated observations: Part 2 – correlation between subjects. BMJ 310:633.

30. Whittaker RH. 1952. A study of summer foliage insect communities in the Great Smoky Mountains. Ecol. Monogr. 22:1-44.

31. Yue JC, Clayton MK. 2005. A similarity measure based on species proportions. Commun. Stat A-Theor. 34:2123-2131.

Table C.2. Minimum correlation between measures over all trials.

Perfectly correlated measures are given in the first column and the first listed measure used as a representative in all other columns.

Measure	>0.99	>0.97	>0.95	>0.90
Bray-Curtis* Normalized weighted UniFrac*	Kulczynski	Soergel	Lennon	Manhattan Hellinger
Bray-Curtis (MRCA restricted)				
Canberra	Coefficient of similarity			uEuclidean, uManhattan
Chi-squared				
Coefficient of similarity	Canberra			Gower
Euclidean			Rao's H_p Yue-Clayton	Morisita-Horn Weighted correlation
F_{ST}				Rao's H_p
Gower				Coefficient of similarity
Hellinger				Bray-Curtis, Kulczynski Manhattan, Soergel
Kulczynski	Bray-Curtis	Lennon, Soergel		Hellinger, Manhattan
Lennon		Kulczynski	Bray-Curtis, Soergel	Manhattan
Manhattan Complete tree Tamàs coefficient Whittaker				Bray-Curtis Hellinger Kulczynski Lennon, Soergel
MNND				
MPD				
Morisita-Horn		Yue-Clayton	Weighted correlation	Euclidean, Rao's H_p
Pearson dissimilarity				Rao's H_p Weighted correlation Yue-Clayton
Rao's H_p			Euclidean	F_{ST}, Morisita-Horn Pearson dissimilarity Weighted correlation Yue-Clayton
Soergel		Bray-Curtis, Kulczynski	Lennon	Hellinger, Manhattan
Weighted correlation		Yue-Clayton	Morisita-Horn	Euclidean Pearson dissimilarity, Rao's H_p
Yue-Clayton		Morisita-Horn Weighted correlation	Euclidean	Pearson dissimilarity Rao's H_p
uBray-Curtis		uBray-Curtis (MRCA restricted) uSoergel		uWeighted correlation
uBray-Curtis (MRCA restricted)		uBray-Curtis uSoergel		
uEuclidean		uManhattan		Canberra
uKulczynski				uWeighted correlation
uLennon				
uManhattan uCanberra uCoefficient of similarity uGower uTamàs coefficient		uEuclidean		Canberra
uMNND				
uMPD				

* Measures are mathematically equivalent

Table C.2. Minimum correlation between measures over all trials (continued).

Measure	>0.99	>0.97	>0.95	>0.90
uPearson dissimilarity				
uSoergel		uBray-Curtis uBray-Curtis (MRCA restricted)		uWeighted correlation
uWeighted correlation				uBray-Curtis uKulczynski, uSoergel

Table C.3. Robustness of quantitative measures to sequence clustering.

Results where the mean or minimum Pearson's *r* equals 1.00 are marked with asterisks.

Mean Pearson's r	**human**			**keyboard**			**mouse gut**			**soil**		
Min. Pearson's r	0.97	0.95	0.85	0.97	0.95	0.85	0.97	0.95	0.85	0.97	0.95	0.85
Bray-Curtis	*	*	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	*	*	*	0.99	*	*	*
Bray-Curtis (MRCA restricted)	*	*	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	0.99	*	*	0.99	*	*	*
Canberra	*	0.99	0.86	*	*	0.95	0.99	0.99	0.94	*	*	0.94
	0.99	0.97	0.63	0.99	0.98	0.87	0.96	0.94	0.81	*	0.99	0.88
Chi-squared	*	*	0.99	*	*	*	*	*	0.99	*	*	0.99
	*	*	0.98	*	*	0.98	*	*	0.96	*	*	0.99
Coefficient of similarity	*	0.99	0.86	*	*	0.96	0.99	0.99	0.95	*	*	0.95
	0.99	0.97	0.62	0.99	0.98	0.88	0.96	0.94	0.82	*	0.99	0.89
Complete tree	*	*	*	*	*	*	*	*	0.99	*	*	*
	*	*	0.99	*	*	0.99	*	*	0.97	*	*	0.99
Euclidean	*	*	*	*	*	*	*	*	0.99	*	*	*
	*	*	0.99	*	*	0.99	*	*	0.98	*	*	0.99
F_{ST}	*	*	*	*	*	*	*	*	0.99	*	*	*
	*	*	0.99	*	*	0.99	*	*	0.98	*	*	0.99
Gower	*	0.99	0.83	*	*	0.97	0.99	0.99	0.94	*	*	0.94
	0.97	0.94	0.56	0.99	0.98	0.90	0.94	0.92	0.76	*	0.99	0.85
Hellinger	*	*	*	*	*	*	*	*	0.99	*	*	*
	*	*	0.99	*	*	*	*	*	0.99	*	*	*
Kulczynski	*	*	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	*	*	*	0.99	*	*	*
Lennon	*	*	*	*	*	*	*	*	0.99	*	*	0.99
	*	*	0.99	*	*	0.99	*	*	0.98	*	*	0.99
MNND	*	*	0.99	*	*	0.99	*	*	0.98	*	*	0.99
	*	*	0.98	0.99	0.98	0.92	0.99	0.99	0.95	*	*	0.99
MPD	*	*	0.99	*	*	*	*	*	0.99	*	*	0.99
	*	*	0.97	*	*	*	*	*	0.95	*	*	0.98
Manhattan	*	*	*	*	*	*	*	*	0.99	*	*	*
	*	*	0.99	*	*	0.99	*	*	0.97	*	*	0.99
Morisita-Horn	*	*	*	*	*	*	*	*	*	*	*	*
	*	*	0.99	*	*	0.99	*	*	0.99	*	*	0.99
Pearson dissimilarity	*	*	0.97	*	*	0.98	*	*	0.97	*	*	0.99
	*	*	0.83	*	0.99	0.89	*	0.99	0.91	*	*	0.98
Rao's H_p	*	*	*	*	*	*	*	*	0.99	*	*	0.99
	*	*	0.99	*	*	0.99	*	*	0.98	*	*	0.99
Soergel	*	*	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	*	*	*	0.99	*	*	*
Tamàs coefficient	*	*	*	*	*	*	*	*	0.99	*	*	*
	*	*	0.99	*	*	0.99	*	*	0.97	*	*	0.99
Weighted correlation	*	*	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	0.99	*	*	0.99	*	*	0.99
Whittaker	*	*	*	*	*	*	*	*	0.99	*	*	*
	*	*	0.99	*	*	0.99	*	*	0.97	*	*	0.99
Yue-Clayton	*	*	*	*	*	*	*	*	0.99	*	*	*
	*	*	0.99	*	*	0.99	*	*	0.99	*	*	0.99

Table C.4. Robustness of qualitative measures to sequence clustering.

Results where the mean or minimum Pearson's *r* equals 1.00 are marked with asterisks.

Mean Pearson's r	**human**			**keyboard**			**mouse gut**			**soil**		
Min. Pearson's r	0.97	0.95	0.85	0.97	0.95	0.85	0.97	0.95	0.85	0.97	0.95	0.85
uBray-Curtis	*	*	0.98	*	*	0.96	0.99	0.98	0.95	*	*	0.98
	*	*	0.95	0.99	0.99	0.91	0.96	0.94	0.86	*	*	0.97
uBray-Curtis	*	*	0.98	*	*	0.96	0.98	0.97	0.94	*	*	0.98
(MRCA restricted)	*	*	0.95	0.99	0.99	0.91	0.96	0.92	0.83	*	*	0.97
uCanberra	*	0.99	0.85	*	0.99	0.93	0.99	0.98	0.93	*	0.99	0.93
	0.99	0.97	0.56	0.99	0.97	0.85	0.94	0.92	0.82	*	0.99	0.79
uCoefficient of	*	0.99	0.85	*	0.99	0.93	0.99	0.98	0.93	*	0.99	0.93
similarity	0.99	0.97	0.56	0.99	0.97	0.85	0.94	0.92	0.82	*	0.99	0.79
uEuclidean	*	0.99	0.85	*	0.99	0.93	0.99	0.98	0.93	*	0.99	0.94
	0.99	0.97	0.60	0.99	0.97	0.85	0.93	0.91	0.82	*	0.99	0.81
uGower	*	0.99	0.85	*	0.99	0.93	0.99	0.98	0.93	*	0.99	0.93
	0.99	0.97	0.56	0.99	0.97	0.85	0.94	0.92	0.82	*	0.99	0.79
uKulczynski	*	*	0.98	*	0.99	0.95	0.98	0.97	0.92	*	*	0.98
	*	*	0.95	0.99	0.98	0.85	0.95	0.93	0.84	*	*	0.95
uLennon	*	*	0.97	*	0.99	0.93	0.97	0.94	0.80	*	*	0.97
	*	0.99	0.91	0.98	0.95	0.76	0.93	0.87	0.50	*	0.99	0.93
uMNND	0.99	0.98	0.95	0.98	0.95	0.89	0.98	0.97	0.88	*	*	0.99
	0.98	0.95	0.89	0.94	0.86	0.64	0.96	0.93	0.76	*	*	0.97
uMPD	0.94	0.85	0.69	0.77	0.43	-0.08	0.97	0.93	0.72	*	0.99	0.93
	0.80	0.43	-0.24	0.28	-0.26	-0.72	0.89	0.80	0.46	0.99	0.99	0.82
uManhattan	*	0.99	0.85	*	0.99	0.93	0.99	0.98	0.93	*	0.99	0.93
	0.99	0.97	0.56	0.99	0.97	0.85	0.94	0.92	0.82	*	0.99	0.79
uPearson	*	0.99	0.97	*	0.99	0.90	0.98	0.97	0.90	*	*	0.97
dissimilarity	*	0.99	0.91	0.99	0.97	0.74	0.96	0.94	0.81	*	0.99	0.95
uSoergel	*	*	0.98	*	*	0.96	0.98	0.97	0.94	*	*	0.98
	*	0.99	0.94	0.99	0.99	0.91	0.96	0.92	0.84	*	*	0.97
uTamàs coefficient	*	0.99	0.85	*	0.99	0.93	0.99	0.98	0.93	*	0.99	0.93
	0.99	0.97	0.56	0.99	0.97	0.85	0.94	0.92	0.82	*	0.99	0.79
uWeighted	*	*	0.98	*	0.99	0.96	0.99	0.97	0.93	*	*	0.99
correlation	*	0.99	0.96	0.99	0.99	0.89	0.97	0.93	0.83	*	*	0.96

Table C.5. Robustness of quantitative measures to outlying lineages.

The mean correlation and standard deviation is given in the first row. The second row gives the minimum correlation over all trials. Results where the mean or minimum Pearson's *r* equals 1.00 are marked with asterisks.

Mean *Minimum*	**human**		**keyboard**		**mouse gut**		**soil**	
	1 sq	*5%*	*1 sq*	*5%*	*1 sq*	*5%*	*1 sq*	*5%*
Bray-Curtis	* *	1.00 0.993	* *	0.98 0.956	* *	1.00 0.990	* *	0.99 0.975
Bray-Curtis (MRCA restricted)	0.99 0.977	0.99 0.970	0.94 0.879	0.94 0.871	0.94 0.830	0.93 0.819	0.97 0.957	0.97 0.949
Canberra	* 0.997	1.00 0.996	1.00 0.990	1.00 0.991	1.00 0.984	1.00 0.981	* *	* 0.999
Chi-squared	* *	1.00 0.996	* *	0.99 0.970	* *	1.00 0.997	* *	0.99 0.990
Coefficient of similarity	* 0.997	1.00 0.996	1.00 0.990	1.00 0.991	1.00 0.984	1.00 0.980	* *	* 0.999
Complete tree	* *	1.00 0.993	* *	0.98 0.959	* *	1.00 0.991	* *	0.99 0.975
Euclidean	* *	* 0.996	* *	0.99 0.983	* *	1.00 0.997	* *	0.99 0.985
F_{ST}	* *	0.99 0.990	* *	0.99 0.978	* *	0.99 0.980	* *	0.99 0.982
Gower	1.00 0.995	1.00 0.989	1.00 0.987	0.99 0.984	1.00 0.978	1.00 0.951	* *	* 0.998
Hellinger	* *	1.00 0.993	* *	0.97 0.886	* *	1.00 0.992	* *	0.99 0.978
Kulczynski	* *	1.00 0.993	* *	0.98 0.956	* *	1.00 0.990	* *	0.99 0.975
Lennon	* *	1.00 0.991	* *	0.98 0.952	* *	0.99 0.988	* *	0.98 0.974
Manhattan	* *	1.00 0.993	* *	0.98 0.959	* *	1.00 0.991	* *	0.99 0.975
MNND	* *	1.00 0.989	* *	0.89 0.575	* *	1.00 0.992	* *	0.99 0.976
MPD	* *	0.99 0.972	* *	0.98 0.955	* *	0.99 0.990	* *	0.94 0.854
Morista-Horn	* *	* *	* *	1.00 0.986	* *	1.00 0.999	* *	1.00 0.989
Pearson dissimilarity	* *	0.99 0.965	* *	0.90 0.678	* *	1.00 0.991	* *	0.62 0.415
Rao's H_p	* *	1.00 0.996	* *	0.99 0.988	* *	1.00 0.997	* *	1.00 0.987
Soergel	* *	1.00 0.993	* *	0.98 0.949	* *	1.00 0.991	* *	0.99 0.975
Tamàs coefficient	* *	1.00 0.993	* *	0.98 0.959	* *	1.00 0.991	* *	0.99 0.975
Weighted correlation	* *	* *	* *	* 0.999	* *	* *	* *	* 0.997
Whittaker	* *	1.00 0.993	* *	0.98 0.959	* *	1.00 0.991	* *	0.99 0.975
Yue-Clayton	* *	* 0.999	* *	1.00 0.991	* *	1.00 0.999	* *	1.00 0.992

Table C.6. Robustness of qualitative measures to outlying lineages.

The mean correlation and standard deviation is given in the first row. The second row gives the minimum correlation over all trials. Results where the mean or minimum Pearson's *r* equals 1.00 are marked with asterisks.

Mean ± s.d. *Minimum*	**human** *≥1 sq*	**keyboard** *≥1 sq*	**mouse gut** *≥1 sq*	**soil** *≥1 sq*
uBray-Curtis	1.00±0.001 0.992	1.00±0.002 0.990	0.98±0.006 0.966	0.98±0.006 0.955
uBray-Curtis (MRCA restricted)	1.00±0.002 0.991	1.00±0.002 0.989	0.98±0.006 0.964	0.98±0.006 0.956
uCanberra	* 0.996	* 0.996	1.00±0.002 0.988	1.00±0.002 0.990
uCoefficient of similarity	* 0.996	* 0.996	1.00±0.002 0.988	1.00±0.002 0.990
uEuclidean	* 0.996	* 0.996	1.00±0.002 0.988	1.00±0.002 0.990
uGower	* 0.996	* 0.996	1.00±0.002 0.988	1.00±0.002 0.990
uKulczynski	0.99±0.002 0.987	0.99±0.003 0.986	0.98±0.010 0.945	0.98±0.011 0.938
uLennon	0.99±0.005 0.976	0.99±0.005 0.979	0.97±0.019 0.887	0.97±0.018 0.911
uMNND	* *	* 1.000	1.00±0.002 0.990	1.00±0.002 0.990
uMPD	1.00±0.001 0.996	1.00±0.001 0.996	1.00±0.005 0.963	1.00±0.005 0.960
uManhattan	* 0.996	* 0.996	1.00±0.002 0.988	1.00±0.002 0.990
uPearson dissimilarity	0.50±0.100 0.203	0.52±0.101 0.319	0.35±0.139 -0.007	0.34±0.142 0.052
uSoergel	1.00±0.002 0.990	1.00±0.002 0.990	0.98±0.006 0.965	0.98±0.006 0.956
uTamàs coefficient	* 0.996	* 0.996	1.00±0.002 0.988	1.00±0.002 0.990
uWeighted correlation	1.00±0.002 0.991	0.98±0.007 0.960	0.97±0.012 0.925	* 0.999

Table C.7. Robustness of quantitative measures to root placement.

The mean and standard deviation is given in the first row. The second row gives the minimum correlation over all trials. Results where the mean or minimum Pearson's *r* equals 1.000 are marked with asterisks.

Measure	**human**	**keyboard**	**mouse gut**	**soil**
Bray-Curtis	0.964±0.0473 0.7254	0.984±0.0309 0.7517	0.909±0.1149 0.2991	0.995±0.0061 0.9645
Bray-Curtis (MRCA restricted)	0.942±0.0593 0.2275	0.948±0.0446 0.6202	0.854±0.1121 0.3911	0.972±0.0319 0.6034
Canberra	1.000±0.0002 0.9983	1.000±0.0002 0.9955	1.000±0.0003 0.9960	* 0.9999
Chi-squared	0.995±0.0118 0.7807	0.997±0.0043 0.9337	0.992±0.0161 0.7824	0.999±0.0020 0.9726
Coefficient of similarity	1.000±0.0001 0.9986	1.000±0.0002 0.9934	1.000±0.0003 0.9963	* 0.9999
Complete tree	Invariant			
Euclidean	Invariant			
F_{ST}	Invariant			
Gower	Invariant			
Hellinger	0.998±0.0031 0.9700	0.997±0.0035 0.9344	0.989±0.0131 0.9027	1.000±0.0006 0.9912
Kulczynski	0.964±0.0472 0.6992	0.984±0.0315 0.7041	0.907±0.1206 0.3046	0.995±0.0067 0.9554
Lennon	0.932±0.1002 0.3434	0.906±0.1615 -0.0739	0.820±0.2317 0.1275	0.977±0.0340 0.7611
Manhattan	Invariant			
MNND	Invariant			
MPD	Invariant			
Morista-Horn	0.961±0.0467 0.6045	0.979±0.0419 0.4975	0.880±0.1579 0.1027	0.988±0.0180 0.8331
Pearson dissimilarity	0.952±0.0609 0.4027	0.955±0.0866 0.1073	0.886±0.1791 0.0427	0.988±0.0203 0.8240
Rao's H_p	Invariant			
Soergel	0.965±0.0494 0.7058	0.985±0.0300 0.7437	0.921±0.1062 0.3302	0.995±0.0065 0.9580
Tamàs coefficient	Perfect positive correlation			
Weighted correlation	0.961±0.0456 0.5805	0.979±0.0412 0.4733	0.911±0.1348 0.0544	0.987±0.0190 0.8219
Whittaker	Invariant			
Yue-Clayton	0.969±0.0355 0.7716	0.989±0.0227 0.7444	0.933±0.0966 0.3572	0.995±0.0074 0.9332

Table C.8. Robustness of qualitative measures to root placement.

The mean and standard deviation is given in the first row. The second row gives the minimum correlation over all trials. Results where the mean or minimum Pearson's *r* equals 1.000 are marked with asterisks.

Measure	human	keyboard	mouse gut	soil
uBray-Curtis	1.000±0.0007 0.9912	0.999±0.0012 0.9798	0.993±0.0076 0.9504	* 0.9994
Bray-Curtis (MRCA restricted)	1.000±0.0006 0.9933	0.999±0.0018 0.9729	0.989±0.0116 0.9232	1.000±0.0001 0.9994
uCanberra	1.000±0.0002 0.9968	1.000±0.0003 0.9962	1.000±0.0007 0.9933	* 0.9999
uCoefficient of similarity	1.000±0.0001 0.9968	1.000±0.0003 0.9952	1.000±0.0007 0.9932	* 0.9999
uEuclidean	1.000±0.0001 0.9971	1.000±0.0003 0.9961	1.000±0.0006 0.9903	* 0.9999
uGower	1.000±0.0002 0.9970	1.000±0.0003 0.9971	1.000±0.0007 0.9932	* 0.9998
uKulczynski	1.000±0.0009 0.9893	0.999±0.0013 0.9714	0.992±0.0088 0.9368	* 0.9993
uLennon	0.999±0.0014 0.9845	0.999±0.0011 0.9839	0.991±0.0109 0.9266	1.000±0.0001 0.9991
uManhattan	1.000±0.0001 0.9970	1.000±0.0003 0.9962	1.000±0.0007 0.9931	* 0.9999
uMNND	Invariant			
uMPD	Invariant			
uPearson dissimilarity	0.999±0.0023 0.9768	0.998±0.0031 0.9097	0.990±0.0139 0.7339	1.000±0.0003 0.9918
uSoergel	1.000±0.0007 0.9907	0.999±0.0014 0.9774	0.995±0.0057 0.9588	* 0.9994
uTamàs coefficient	1.000±0.0002 0.9969	1.000±0.0003 0.9950	1.000±0.0007 0.9932	* 0.9999
uWeighted correlation	1.000±0.0007 0.9916	0.999±0.0011 0.9861	0.994±0.0068 0.9562	* 0.9996

Table C.9. Sensitivity of quantitative measures to rare OTUs.

Results where the mean or minimum Pearson's *r* equals 1.00 are marked with asterisks.

Mean Pearson's r *Min. Pearson's r*	**human**			**keyboard**			**mouse gut**			**soil**		
	1sq	*0.1%*	*1%*	*1sq*	*0.1%*	*1%*	*1sq*	*0.1%*	*1%*	*1sq*	*0.1%*	*1%*
Bray-Curtis	*	*	0.99	*	*	0.98	*	*	*	*	*	0.75
	*	0.99	0.94	0.99	0.99	0.91	*	*	0.99	0.99	0.99	0.38
Bray-Curtis (MRCA restricted)	0.99	0.99	0.97	0.98	0.99	0.96	0.98	*	0.95	0.99	0.99	0.71
	0.97	0.97	0.92	0.95	0.95	0.90	0.94	*	0.80	0.96	0.96	0.32
Canberra	0.94	0.61	0.62	0.89	0.71	0.73	0.98	0.96	0.93	0.97	0.70	0.41
	0.77	0.22	0.27	0.68	0.26	0.27	0.95	0.89	0.75	0.90	0.17	0.05
Chi-squared	*	*	0.98	*	*	0.97	*	*	*	*	*	0.75
	*	*	0.92	0.99	0.99	0.86	*	*	0.99	0.99	0.99	0.23
Coefficient of similarity	0.93	0.59	0.60	0.89	0.70	0.73	0.99	0.96	0.93	0.97	0.68	0.33
	0.76	0.18	0.23	0.66	0.22	0.30	0.96	0.89	0.74	0.90	0.08	-0.05
Complete tree	*	*	0.99	*	*	0.98	*	*	*	*	*	0.78
	*	0.99	0.93	0.99	0.99	0.90	*	*	0.99	0.99	0.99	0.42
Euclidean	*	*	0.98	*	*	0.98	*	*	*	0.99	0.99	0.62
	0.99	0.99	0.91	0.99	0.99	0.93	*	*	0.99	0.98	0.98	0.21
F_{ST}	*	*	0.97	*	*	0.97	*	*	0.99	0.99	0.99	0.42
	0.99	0.99	0.91	0.99	0.99	0.90	*	*	0.98	0.98	0.98	-0.04
Gower	0.93	0.61	0.59	0.92	0.80	0.80	0.99	0.96	0.93	0.97	0.70	0.35
	0.75	0.19	0.12	0.63	0.37	0.41	0.94	0.88	0.71	0.88	0.07	-0.11
Hellinger	*	*	0.99	*	*	0.95	*	*	0.99	*	*	0.78
	*	*	0.93	0.99	0.99	0.83	*	*	0.97	0.99	0.99	0.44
Kulczynski	*	*	0.99	*	*	0.98	*	*	*	*	*	0.75
	*	0.99	0.94	0.99	0.99	0.91	*	*	0.99	0.99	0.99	0.39
Lennon	*	*	0.99	*	*	0.98	*	*	*	*	*	0.77
	0.99	0.99	0.92	0.99	0.99	0.90	*	*	0.99	0.99	0.99	0.43
MNND	0.97	0.97	0.92	0.96	0.95	0.82	0.95	0.98	0.89	0.99	0.99	0.70
	0.92	0.90	0.81	0.86	0.85	0.41	0.85	0.91	0.68	0.98	0.98	0.27
MPD	*	*	0.97	*	*	0.98	*	*	*	*	*	0.74
	0.97	0.98	0.73	*	0.99	0.93	*	*	0.99	0.99	0.99	0.23
Manhattan	*	*	0.99	*	*	0.98	*	*	*	*	*	0.78
	*	0.99	0.93	0.99	0.99	0.90	*	*	0.99	0.99	0.99	0.42
Morisita-Horn	*	*	0.98	*	*	0.97	*	*	*	0.99	0.99	0.53
	*	0.99	0.92	0.99	0.99	0.91	*	*	0.99	0.98	0.98	0.02
Pearson dissimilarity	*	*	0.98	*	*	0.97	*	*	0.99	0.99	0.99	0.57
	0.99	0.99	0.92	0.99	0.99	0.90	*	*	0.98	0.98	0.98	0.08
Rao's H_p	*	*	0.97	*	*	0.97	*	*	*	0.99	0.99	0.49
	0.99	0.98	0.91	0.99	0.99	0.91	*	*	0.98	0.98	0.98	-0.00
Soergel	*	*	0.99	*	*	0.98	*	*	*	*	*	0.78
	*	0.99	0.94	*	0.99	0.91	*	*	0.99	0.99	0.99	0.41
Tamàs coefficient	*	*	0.99	*	*	0.98	*	*	*	*	*	0.78
	*	0.99	0.93	0.99	0.99	0.90	*	*	0.99	0.99	0.99	0.42
Weighted correlation	*	*	0.98	*	*	0.97	*	*	*	0.99	0.99	0.55
	0.98	0.98	0.88	0.99	0.99	0.90	*	*	0.99	0.98	0.98	0.08
Whittaker	*	*	0.99	*	*	0.98	*	*	*	*	*	0.78
	*	0.99	0.93	0.99	0.99	0.90	*	*	0.99	0.99	0.99	0.42
Yue-Clayton	*	*	0.98	*	*	0.97	*	*	*	0.99	0.99	0.59
	*	0.99	0.91	0.99	0.99	0.91	*	*	0.99	0.98	0.98	0.13

Table C.10. Sensitivity of qualitative measures to rare OTUs.

Results where the mean or minimum Pearson's *r* equals 1.00 are marked with asterisks.

Mean Pearson's r	**human**			**keyboard**			**mouse gut**			**soil**		
Min. Pearson's r	*1sq*	*0.1%*	*1%*	*1sq*	*0.1%*	*1%*	*1sq*	*0.1%*	*1%*	*1sq*	*0.1%*	*1%*
uBray-Curtis	0.98	0.96	0.91	0.88	0.85	0.62	0.98	0.99	0.92	0.99	0.98	0.80
	0.95	0.90	0.81	0.60	0.50	0.17	0.96	0.98	0.79	0.97	0.96	0.54
uBray-Curtis	0.98	0.96	0.91	0.88	0.85	0.62	0.98	0.99	0.90	0.99	0.98	0.77
(MRCA restricted)	0.95	0.90	0.80	0.59	0.49	0.16	0.96	0.98	0.73	0.97	0.96	0.45
uCanberra	0.94	0.60	0.63	0.86	0.70	0.64	0.99	0.96	0.92	0.96	0.71	0.52
	0.78	0.29	0.22	0.61	0.36	0.10	0.96	0.91	0.77	0.89	0.21	0.20
uCoefficient of	0.94	0.60	0.63	0.86	0.70	0.64	0.99	0.96	0.92	0.96	0.71	0.52
similarity	0.78	0.29	0.22	0.61	0.36	0.10	0.96	0.91	0.77	0.89	0.21	0.20
uEuclidean	0.94	0.64	0.65	0.87	0.71	0.65	0.99	0.97	0.92	0.97	0.76	0.53
	0.80	0.31	0.26	0.62	0.40	0.07	0.96	0.93	0.78	0.90	0.29	0.25
uGower	0.94	0.60	0.63	0.86	0.70	0.64	0.99	0.96	0.92	0.96	0.71	0.52
	0.78	0.29	0.22	0.61	0.36	0.10	0.96	0.91	0.77	0.89	0.21	0.20
uKulczynski	0.98	0.97	0.90	0.85	0.88	0.57	0.97	0.99	0.91	0.99	0.98	0.80
	0.94	0.93	0.78	0.61	0.61	0.15	0.94	0.97	0.75	0.97	0.96	0.53
uLennon	0.96	0.93	0.84	0.69	0.53	0.32	0.95	0.97	0.86	0.97	0.88	0.72
	0.87	0.83	0.63	0.13	-0.09	-0.16	0.90	0.91	0.62	0.93	0.77	0.42
uMNND	0.98	0.97	0.92	0.90	0.89	0.56	0.96	0.98	0.89	0.99	0.99	0.66
	0.95	0.93	0.80	0.68	0.68	0.10	0.86	0.89	0.76	0.98	0.97	0.24
uMPD	0.94	0.95	0.78	0.86	0.90	0.53	0.99	*	0.97	0.98	0.99	0.65
	0.78	0.80	0.31	0.57	0.60	0.06	0.96	0.98	0.93	0.96	0.97	0.22
uManhattan	0.94	0.60	0.63	0.86	0.70	0.64	0.99	0.96	0.92	0.96	0.71	0.52
	0.78	0.29	0.22	0.61	0.36	0.10	0.96	0.91	0.77	0.89	0.21	0.20
uPearson	0.95	0.93	0.83	0.74	0.74	0.46	0.97	0.99	0.91	0.95	0.94	0.81
dissimilarity	0.91	0.86	0.60	0.36	0.31	-0.01	0.92	0.98	0.75	0.88	0.86	0.59
uSoergel	0.97	0.96	0.90	0.88	0.85	0.62	0.98	0.99	0.92	0.99	0.98	0.81
	0.93	0.87	0.78	0.61	0.51	0.19	0.96	0.98	0.79	0.97	0.96	0.58
uTamàs coefficient	0.94	0.60	0.63	0.86	0.70	0.64	0.99	0.96	0.92	0.96	0.71	0.52
	0.78	0.29	0.22	0.61	0.36	0.10	0.96	0.91	0.77	0.89	0.21	0.20
uWeighted	0.97	0.97	0.91	0.88	0.88	0.62	0.98	0.99	0.92	0.98	0.97	0.81
correlation	0.95	0.93	0.80	0.56	0.67	0.31	0.96	0.97	0.80	0.96	0.94	0.57

Table C.11. Performance of measures depends on the model of diversification.

Comparison of *k*-medoids score results under the equal-perturbation and dominant-pair models of diversification for samples containing 1,000 sequences. See Supplementary Tables C.12-C.19 for results on individual measures and different sampling depths.

Dataset	*Paired t-test*	*Pearson's correlation*	*Spearman's correlation*
human	$1.30 \cdot 10^{-6}$	0.10	0.03
keyboard	$3.16 \cdot 10^{-5}$	0.30	0.39
mouse gut	$5.24 \cdot 10^{-3}$	0.41	0.52
soil	$6.10 \cdot 10^{-18}$	0.31	0.23

Table C.12. Results for equal-perturbation model on samples from the human dataset.

The mean and standard deviation are given for the *k*-medoid score (KMS) and consistency index (CI) statistics. Pearson's correlation coefficient between the 2 statistics is r = 0.76, 0.80, 0.86 for 100, 1,000, and 10,000 sequences, respectively.

Results are sorted by KMS for samples with 1,000 sequences.

Mean ± s.d.	**100 sequences**		**1,000 sequences**		**10,000 sequences**	
	KMS	*CI*	*KMS*	*CI*	*KMS*	*CI*
MNND	0.58±0.110	0.15±0.076	0.91±0.123	0.83±0.286	0.95±0.114	0.90±0.236
Coefficient of similarity	0.55±0.103	0.14±0.073	0.90±0.123	0.80±0.286	0.98±0.051	0.96±0.142
Hellinger	0.70±0.134	0.20±0.117	0.89±0.134	0.72±0.319	0.92±0.118	0.75±0.312
Canberra	0.56±0.101	0.13±0.059	0.89±0.131	0.80±0.289	0.99±0.056	0.97±0.107
Chi-squared	0.68±0.134	0.16±0.091	0.89±0.129	0.68±0.322	0.92±0.113	0.73±0.321
Gower	0.58±0.119	0.14±0.061	0.89±0.141	0.78±0.298	0.98±0.051	0.89±0.218
Lennon	0.65±0.128	0.14±0.071	0.80±0.148	0.39±0.276	0.85±0.140	0.49±0.301
uCoefficient of similarity	0.52±0.086	0.09±0.029	0.80±0.146	0.52±0.308	0.73±0.178	0.35±0.291
uEuclidean	0.51±0.082	0.09±0.031	0.78±0.152	0.52±0.299	0.74±0.181	0.35±0.285
uCanberra	0.51±0.089	0.09±0.029	0.78±0.155	0.52±0.308	0.76±0.170	0.35±0.291
Kulczynski	0.65±0.124	0.12±0.059	0.77±0.141	0.31±0.226	0.81±0.146	0.40±0.282
uTamàs coefficient	0.50±0.077	0.09±0.029	0.77±0.158	0.52±0.308	0.75±0.175	0.35±0.291
Soergel	0.64±0.124	0.12±0.057	0.77±0.144	0.31±0.228	0.81±0.143	0.41±0.289
uManhattan	0.51±0.080	0.09±0.029	0.77±0.160	0.52±0.308	0.75±0.174	0.35±0.291
Bray-Curtis	0.66±0.118	0.12±0.057	0.77±0.143	0.31±0.229	0.81±0.150	0.40±0.283
Bray-Curtis (MRCA restricted)	0.65±0.118	0.12±0.051	0.77±0.147	0.31±0.233	0.81±0.148	0.40±0.275
Complete tree	0.65±0.120	0.12±0.055	0.77±0.141	0.32±0.243	0.81±0.142	0.41±0.285
Whittaker	0.65±0.124	0.12±0.055	0.77±0.143	0.32±0.243	0.80±0.151	0.41±0.285
Tamàs coefficient	0.65±0.124	0.12±0.055	0.77±0.139	0.32±0.243	0.81±0.150	0.41±0.285
uGower	0.51±0.086	0.09±0.029	0.76±0.154	0.52±0.308	0.78±0.158	0.35±0.291
Manhattan	0.65±0.113	0.12±0.055	0.76±0.142	0.32±0.243	0.80±0.148	0.41±0.285
uBray-Curtis (MRCA restricted)	0.53±0.088	0.10±0.040	0.74±0.146	0.51±0.297	0.74±0.177	0.33±0.278
uBray-Curtis	0.53±0.083	0.10±0.030	0.73±0.157	0.51±0.298	0.75±0.173	0.33±0.279
uSoergel	0.53±0.076	0.10±0.030	0.73±0.154	0.51±0.298	0.74±0.179	0.33±0.279
uMNND	0.54±0.093	0.10±0.036	0.73±0.155	0.55±0.304	0.74±0.172	0.39±0.305
Euclidean	0.63±0.122	0.11±0.047	0.72±0.136	0.19±0.092	0.73±0.136	0.22±0.144
Weighted correlation	0.63±0.125	0.11±0.047	0.71±0.139	0.20±0.103	0.72±0.145	0.23±0.128
Morisita-Horn	0.63±0.119	0.11±0.047	0.71±0.138	0.20±0.106	0.72±0.143	0.24±0.145
uWeighted correlation	0.52±0.080	0.10±0.030	0.71±0.151	0.51±0.290	0.71±0.183	0.34±0.284
uKulczynski	0.52±0.076	0.10±0.028	0.71±0.149	0.51±0.297	0.74±0.175	0.34±0.284
Rao's H_p	0.62±0.121	0.11±0.047	0.71±0.140	0.19±0.090	0.72±0.141	0.22±0.143
F_{ST}	0.62±0.122	0.11±0.043	0.71±0.135	0.19±0.090	0.73±0.135	0.22±0.142
Yue-Clayton	0.62±0.122	0.11±0.047	0.70±0.139	0.19±0.092	0.72±0.140	0.22±0.125
Pearson dissimilarity	0.61±0.121	0.10±0.041	0.69±0.139	0.17±0.085	0.70±0.141	0.20±0.110
uPearson dissimilarity	0.49±0.070	0.07±0.015	0.63±0.116	0.17±0.120	0.70±0.160	0.15±0.081
uLennon	0.48±0.069	0.08±0.033	0.63±0.137	0.32±0.228	0.57±0.170	0.41±0.275
MPD	0.48±0.105	0.07±0.026	0.49±0.125	0.09±0.041	0.49±0.130	0.10±0.060
uMPD	0.40±0.055	0.05±0.005	0.38±0.052	0.05±0.006	0.36±0.042	0.06±0.008

Table C.13. Results for equal-perturbation model on samples from the keyboard dataset.

The mean and standard deviation are given for the *k*-medoids score (KMS) and consistency index (CI) statistics. Pearson's correlation coefficient between the 2 statistics is r = 0.79, 0.90, 0.90 for 100, 1,000, and 10,000 sequences, respectively.

Results are sorted by KMS for samples with 1,000 sequences.

Mean ± s.d.	**100 sequences**		**1,000 sequences**		**10,000 sequences**	
	KMS	*CI*	*KMS*	*CI*	*KMS*	*CI*
MNND	0.61±0.120	0.16±0.081	0.90±0.126	0.76±0.277	0.89±0.151	0.79±0.275
Coefficient of similarity	0.55±0.098	0.15±0.084	0.90±0.110	0.74±0.298	0.97±0.058	0.88±0.187
Canberra	0.57±0.103	0.14±0.074	0.90±0.116	0.73±0.302	0.97±0.061	0.89±0.178
Hellinger	0.70±0.146	0.20±0.138	0.89±0.120	0.63±0.296	0.90±0.103	0.67±0.277
Gower	0.57±0.117	0.14±0.071	0.87±0.136	0.61±0.335	0.95±0.072	0.73±0.265
Chi-squared	0.66±0.145	0.17±0.136	0.86±0.139	0.57±0.322	0.88±0.123	0.62±0.303
Whittaker	0.66±0.141	0.13±0.069	0.76±0.172	0.33±0.260	0.77±0.179	0.41±0.310
Tamàs coefficient	0.65±0.142	0.13±0.069	0.75±0.172	0.33±0.260	0.76±0.181	0.41±0.310
Kulczynski	0.65±0.138	0.13±0.068	0.75±0.171	0.32±0.254	0.76±0.179	0.41±0.309
Complete tree	0.65±0.139	0.13±0.069	0.75±0.171	0.33±0.260	0.76±0.186	0.41±0.310
Lennon	0.64±0.140	0.13±0.066	0.75±0.171	0.32±0.235	0.77±0.183	0.42±0.298
Bray-Curtis	0.65±0.140	0.13±0.064	0.75±0.170	0.31±0.252	0.77±0.183	0.41±0.307
Bray-Curtis (MRCA restricted)	0.65±0.142	0.13±0.069	0.75±0.173	0.31±0.250	0.76±0.182	0.41±0.308
Manhattan	0.65±0.141	0.13±0.069	0.75±0.178	0.33±0.260	0.76±0.182	0.41±0.310
Soergel	0.65±0.134	0.13±0.064	0.75±0.172	0.31±0.252	0.77±0.180	0.42±0.312
uTamàs coefficient	0.53±0.086	0.11±0.035	0.73±0.150	0.36±0.224	0.67±0.168	0.17±0.135
uCanberra	0.54±0.092	0.11±0.035	0.73±0.145	0.36±0.224	0.69±0.166	0.17±0.135
uCoefficient of similarity	0.52±0.080	0.11±0.035	0.72±0.152	0.36±0.224	0.68±0.169	0.17±0.135
uManhattan	0.54±0.087	0.11±0.035	0.71±0.153	0.36±0.224	0.68±0.170	0.17±0.135
uBray-Curtis	0.57±0.111	0.11±0.041	0.71±0.145	0.34±0.223	0.66±0.158	0.17±0.123
uEuclidean	0.54±0.080	0.11±0.037	0.71±0.155	0.36±0.222	0.67±0.177	0.17±0.134
uGower	0.54±0.094	0.11±0.035	0.70±0.152	0.36±0.224	0.67±0.177	0.17±0.135
uBray-Curtis (MRCA restricted)	0.57±0.103	0.11±0.044	0.69±0.153	0.34±0.224	0.66±0.170	0.17±0.123
uWeighted correlation	0.56±0.100	0.11±0.037	0.69±0.149	0.34±0.231	0.67±0.165	0.17±0.126
Euclidean	0.61±0.136	0.11±0.051	0.68±0.164	0.19±0.122	0.70±0.165	0.23±0.175
uKulczynski	0.54±0.090	0.11±0.036	0.68±0.141	0.34±0.228	0.67±0.175	0.17±0.124
uSoergel	0.56±0.104	0.11±0.041	0.68±0.154	0.34±0.223	0.68±0.163	0.17±0.123
uMNND	0.51±0.079	0.11±0.037	0.68±0.160	0.39±0.255	0.66±0.185	0.19±0.167
F_{ST}	0.62±0.137	0.11±0.052	0.68±0.162	0.19±0.118	0.69±0.172	0.22±0.152
Yue-Clayton	0.61±0.138	0.11±0.056	0.67±0.163	0.19±0.119	0.68±0.174	0.23±0.171
Rao's H_p	0.61±0.134	0.11±0.052	0.67±0.166	0.19±0.118	0.68±0.173	0.23±0.172
Weighted correlation	0.60±0.136	0.11±0.062	0.67±0.168	0.19±0.115	0.67±0.175	0.24±0.176
Morisita-Horn	0.61±0.137	0.11±0.050	0.66±0.167	0.18±0.115	0.67±0.175	0.22±0.164
Pearson dissimilarity	0.60±0.129	0.10±0.048	0.65±0.161	0.17±0.102	0.67±0.168	0.19±0.108
uPearson dissimilarity	0.50±0.087	0.07±0.016	0.63±0.121	0.12±0.040	0.65±0.163	0.12±0.051
uLennon	0.48±0.077	0.08±0.026	0.57±0.141	0.23±0.213	0.51±0.151	0.24±0.209
MPD	0.42±0.085	0.08±0.034	0.42±0.121	0.09±0.049	0.42±0.109	0.09±0.052
uMPD	0.39±0.043	0.05±0.005	0.38±0.053	0.05±0.008	0.36±0.024	0.05±0.010

Table C.14. Results for equal-perturbation model on samples from the mouse gut dataset.

The mean and standard deviation are given for the k-medoids score (KMS) and consistency index (CI) statistics. Pearson's correlation coefficient between the 2 statistics is r = 0.92, 0.87, 0.81 for 100, 1,000, and 10,000 sequences, respectively.

Results are sorted by KMS for samples with 1,000 sequences.

Mean ± s.d.	**100 sequences**		**1,000 sequences**		**10,000 sequences**	
	KMS	*CI*	*KMS*	*CI*	*KMS*	*CI*
Gower	0.62±0.133	0.13±0.075	0.84±0.139	0.47±0.315	0.92±0.111	0.66±0.287
Hellinger	0.75±0.137	0.21±0.133	0.84±0.128	0.46±0.260	0.86±0.120	0.49±0.249
Coefficient of similarity	0.60±0.133	0.14±0.074	0.84±0.124	0.55±0.316	0.96±0.058	0.83±0.246
Canberra	0.59±0.130	0.14±0.078	0.83±0.149	0.55±0.303	0.98±0.035	0.87±0.220
Chi-squared	0.71±0.129	0.18±0.095	0.82±0.128	0.39±0.247	0.84±0.120	0.44±0.241
MNND	0.60±0.129	0.15±0.087	0.80±0.151	0.43±0.263	0.82±0.167	0.44±0.227
Lennon	0.65±0.135	0.12±0.063	0.72±0.148	0.21±0.152	0.74±0.155	0.26±0.214
Whittaker	0.65±0.131	0.12±0.060	0.71±0.149	0.19±0.138	0.72±0.147	0.23±0.194
Tamàs coefficient	0.65±0.128	0.12±0.060	0.71±0.153	0.19±0.138	0.72±0.149	0.23±0.194
Soergel	0.66±0.134	0.13±0.065	0.71±0.151	0.19±0.141	0.72±0.146	0.23±0.195
Manhattan	0.65±0.129	0.12±0.060	0.71±0.149	0.19±0.138	0.72±0.150	0.23±0.194
Bray-Curtis	0.65±0.131	0.13±0.064	0.71±0.148	0.19±0.141	0.72±0.151	0.23±0.196
Complete tree	0.65±0.133	0.12±0.060	0.71±0.150	0.19±0.138	0.72±0.148	0.23±0.194
Kulczynski	0.65±0.132	0.12±0.063	0.71±0.150	0.19±0.141	0.72±0.147	0.23±0.196
Bray-Curtis (MRCA restricted)	0.65±0.134	0.12±0.061	0.71±0.148	0.19±0.139	0.72±0.147	0.23±0.193
Euclidean	0.65±0.130	0.13±0.076	0.70±0.143	0.17±0.131	0.70±0.146	0.18±0.116
F_{ST}	0.66±0.131	0.13±0.068	0.69±0.144	0.17±0.129	0.69±0.141	0.18±0.131
Morisita-Horn	0.65±0.129	0.13±0.070	0.68±0.144	0.17±0.131	0.68±0.142	0.18±0.111
Weighted correlation	0.66±0.130	0.13±0.072	0.68±0.149	0.17±0.113	0.69±0.143	0.19±0.139
Yue-Clayton	0.65±0.128	0.13±0.079	0.67±0.146	0.17±0.132	0.69±0.143	0.18±0.135
Rao's H_p	0.65±0.129	0.13±0.072	0.67±0.140	0.17±0.132	0.68±0.141	0.18±0.117
Pearson dissimilarity	0.63±0.121	0.12±0.057	0.66±0.137	0.16±0.098	0.67±0.132	0.17±0.114
uCoefficient of similarity	0.52±0.074	0.09±0.022	0.66±0.143	0.18±0.177	0.47±0.190	0.15±0.101
uGower	0.52±0.082	0.09±0.022	0.66±0.133	0.18±0.177	0.47±0.192	0.15±0.101
uCanberra	0.52±0.079	0.09±0.022	0.65±0.136	0.18±0.177	0.47±0.183	0.15±0.101
uEuclidean	0.52±0.082	0.09±0.021	0.64±0.140	0.18±0.176	0.46±0.177	0.15±0.094
uManhattan	0.52±0.078	0.09±0.022	0.64±0.136	0.18±0.177	0.46±0.179	0.15±0.101
uTamàs coefficient	0.52±0.088	0.09±0.022	0.64±0.137	0.18±0.177	0.46±0.182	0.15±0.101
uSoergel	0.54±0.084	0.09±0.023	0.63±0.126	0.17±0.142	0.45±0.182	0.15±0.088
uBray-Curtis	0.54±0.092	0.09±0.024	0.63±0.128	0.17±0.144	0.45±0.174	0.15±0.088
uBray-Curtis (MRCA restricted)	0.54±0.090	0.09±0.023	0.63±0.135	0.17±0.144	0.47±0.187	0.15±0.088
uKulczynski	0.53±0.083	0.09±0.021	0.63±0.139	0.17±0.152	0.45±0.176	0.15±0.093
uWeighted correlation	0.53±0.084	0.09±0.023	0.63±0.140	0.17±0.149	0.46±0.174	0.15±0.091
uMNND	0.53±0.085	0.09±0.023	0.60±0.126	0.19±0.152	0.46±0.187	0.16±0.130
uPearson dissimilarity	0.50±0.078	0.08±0.015	0.60±0.131	0.11±0.054	0.45±0.168	0.13±0.049
uLennon	0.47±0.061	0.07±0.020	0.52±0.129	0.14±0.121	0.39±0.118	0.18±0.150
MPD	0.47±0.114	0.08±0.032	0.47±0.125	0.09±0.038	0.47±0.118	0.09±0.036
uMPD	0.39±0.051	0.05±0.005	0.38±0.054	0.05±0.008	0.35±0.022	0.09±0.047

Table C.15. Results for equal-perturbation model on samples from the soil dataset.

The mean and standard deviation are given for the *k*-medoids score (KMS) and consistency index (CI) statistics. Pearson's correlation coefficient between the 2 statistics is r = 0.63, 0.83, 0.90 for 100, 1,000, and 10,000 sequences, respectively.

Results are sorted by KMS for samples with 1,000 sequences.

Mean ± s.d.	**100 sequences**		**1,000 sequences**		**10,000 sequences**	
	KMS	*CI*	*KMS*	*CI*	*KMS*	*CI*
Hellinger	0.68±0.133	0.26±0.150	0.99±0.034	0.99±0.047	1.00±0.007	0.99±0.047
Chi-squared	0.65±0.138	0.16±0.124	0.99±0.030	0.98±0.092	1.00±0.009	1.00±0.033
Canberra	0.56±0.098	0.17±0.065	0.97±0.085	0.99±0.057	1.00±0.032	1.00±0.000
Coefficient of similarity	0.56±0.100	0.17±0.070	0.96±0.091	0.99±0.057	1.00±0.000	1.00±0.000
MNND	0.65±0.108	0.24±0.117	0.96±0.103	0.98±0.079	1.00±0.011	0.98±0.100
Complete tree	0.63±0.137	0.14±0.093	0.96±0.063	0.83±0.240	0.98±0.046	0.93±0.167
Whittaker	0.64±0.134	0.14±0.093	0.95±0.058	0.83±0.240	0.99±0.025	0.93±0.167
Soergel	0.64±0.131	0.15±0.089	0.95±0.058	0.82±0.238	0.98±0.053	0.94±0.160
Bray-Curtis (MRCA restricted)	0.61±0.125	0.14±0.083	0.95±0.063	0.80±0.243	0.98±0.046	0.94±0.159
Tamàs coefficient	0.62±0.139	0.14±0.093	0.95±0.061	0.83±0.240	0.97±0.077	0.93±0.167
Manhattan	0.63±0.131	0.14±0.093	0.95±0.063	0.83±0.240	0.98±0.058	0.93±0.167
Kulczynski	0.65±0.134	0.15±0.089	0.95±0.068	0.81±0.241	0.98±0.031	0.94±0.160
Gower	0.59±0.112	0.18±0.077	0.95±0.118	1.00±0.000	1.00±0.000	1.00±0.000
Bray-Curtis	0.63±0.136	0.15±0.089	0.94±0.090	0.82±0.238	0.99±0.028	0.94±0.162
Lennon	0.62±0.129	0.13±0.057	0.92±0.093	0.83±0.251	0.98±0.054	0.93±0.168
uTamàs coefficient	0.55±0.089	0.12±0.042	0.91±0.110	0.86±0.197	0.88±0.146	0.77±0.313
uEuclidean	0.54±0.087	0.12±0.043	0.91±0.103	0.86±0.197	0.89±0.145	0.77±0.315
uCanberra	0.55±0.087	0.12±0.042	0.91±0.114	0.86±0.197	0.90±0.141	0.77±0.313
uGower	0.55±0.089	0.12±0.042	0.90±0.110	0.86±0.197	0.90±0.137	0.77±0.313
uManhattan	0.54±0.094	0.12±0.042	0.90±0.114	0.86±0.197	0.91±0.131	0.77±0.313
uCoefficient of similarity	0.55±0.095	0.12±0.042	0.89±0.130	0.86±0.197	0.88±0.149	0.77±0.313
uKulczynski	0.58±0.093	0.14±0.056	0.88±0.127	0.87±0.211	0.89±0.148	0.76±0.319
uWeighted correlation	0.58±0.088	0.15±0.061	0.88±0.131	0.86±0.211	0.89±0.153	0.76±0.319
uMNND	0.58±0.101	0.17±0.082	0.88±0.143	0.92±0.172	0.88±0.148	0.83±0.284
uSoergel	0.57±0.088	0.15±0.058	0.88±0.137	0.87±0.220	0.89±0.136	0.76±0.319
uBray-Curtis (MRCA restricted)	0.58±0.090	0.15±0.061	0.87±0.137	0.87±0.220	0.90±0.134	0.76±0.319
uBray-Curtis	0.58±0.086	0.15±0.060	0.85±0.146	0.87±0.217	0.90±0.140	0.76±0.319
Weighted correlation	0.59±0.124	0.11±0.057	0.83±0.115	0.38±0.216	0.90±0.101	0.59±0.289
Rao's H_p	0.59±0.119	0.10±0.050	0.82±0.119	0.35±0.187	0.89±0.107	0.57±0.284
Euclidean	0.59±0.128	0.10±0.048	0.82±0.117	0.35±0.193	0.89±0.110	0.57±0.282
Morisita-Horn	0.59±0.121	0.11±0.067	0.82±0.118	0.37±0.202	0.90±0.094	0.59±0.283
F_{ST}	0.58±0.116	0.10±0.050	0.81±0.119	0.36±0.194	0.89±0.106	0.57±0.277
Yue-Clayton	0.59±0.123	0.10±0.052	0.81±0.123	0.36±0.189	0.89±0.105	0.58±0.288
Pearson dissimilarity	0.58±0.119	0.10±0.040	0.81±0.122	0.32±0.182	0.87±0.120	0.53±0.263
uLennon	0.51±0.090	0.12±0.050	0.75±0.164	0.75±0.278	0.73±0.177	0.78±0.287
uPearson dissimilarity	0.49±0.064	0.08±0.016	0.73±0.144	0.36±0.235	0.82±0.152	0.32±0.206
MPD	0.43±0.090	0.06±0.013	0.44±0.125	0.07±0.023	0.46±0.133	0.08±0.027
uMPD	0.39±0.044	0.05±0.004	0.37±0.049	0.05±0.006	0.36±0.027	0.05±0.007

Table C.16. Results for dominant-pair model on samples from the human dataset.

The mean and standard deviation are given for the *k*-medoids score (KMS) and consistency index (CI) statistics. Pearson's correlation coefficient between the 2 statistics is r = 0.97, 0.99, 0.99 for 100, 1,000, and 10,000 sequences, respectively.

Results are sorted by KMS for samples with 1,000 sequences.

Mean ± s.d.	**100 sequences**		**1,000 sequences**		**10,000 sequences**	
	KMS	*CI*	*KMS*	*CI*	*KMS*	*CI*
Weighted correlation	0.64±0.193	0.19±0.250	0.80±0.210	0.50±0.375	0.83±0.225	0.64±0.388
Morisita-Horn	0.65±0.196	0.19±0.249	0.80±0.221	0.48±0.373	0.82±0.233	0.62±0.384
Euclidean	0.63±0.201	0.20±0.262	0.79±0.210	0.47±0.370	0.83±0.216	0.60±0.379
Yue-Clayton	0.64±0.200	0.19±0.260	0.79±0.214	0.47±0.369	0.83±0.220	0.60±0.380
F_{ST}	0.64±0.197	0.19±0.259	0.79±0.211	0.47±0.369	0.82±0.222	0.60±0.379
Rao's H_p	0.63±0.201	0.19±0.260	0.78±0.218	0.46±0.366	0.82±0.222	0.61±0.385
Soergel	0.62±0.205	0.19±0.258	0.78±0.212	0.43±0.375	0.82±0.222	0.57±0.405
Tamàs coefficient	0.62±0.207	0.19±0.259	0.77±0.213	0.43±0.373	0.81±0.221	0.57±0.405
Bray-Curtis	0.62±0.204	0.19±0.258	0.77±0.213	0.43±0.375	0.81±0.220	0.57±0.405
Complete tree	0.62±0.209	0.19±0.259	0.77±0.212	0.43±0.373	0.81±0.220	0.57±0.405
Manhattan	0.63±0.204	0.19±0.259	0.77±0.211	0.43±0.373	0.80±0.225	0.57±0.405
Whittaker	0.62±0.208	0.19±0.259	0.77±0.216	0.43±0.373	0.79±0.234	0.57±0.405
Kulczynski	0.62±0.206	0.19±0.258	0.77±0.222	0.43±0.375	0.81±0.227	0.57±0.405
Bray-Curtis (MRCA restricted)	0.63±0.203	0.19±0.261	0.76±0.220	0.42±0.370	0.80±0.229	0.57±0.405
Lennon	0.61±0.210	0.19±0.269	0.76±0.214	0.44±0.384	0.79±0.221	0.54±0.400
Pearson dissimilarity	0.62±0.203	0.19±0.258	0.76±0.220	0.39±0.372	0.79±0.234	0.53±0.400
Hellinger	0.56±0.201	0.17±0.239	0.71±0.209	0.39±0.377	0.76±0.211	0.54±0.376
Chi-squared	0.56±0.204	0.12±0.164	0.70±0.209	0.39±0.378	0.76±0.222	0.57±0.388
Gower	0.49±0.147	0.07±0.060	0.53±0.161	0.19±0.279	0.54±0.148	0.23±0.305
MPD	0.47±0.126	0.16±0.235	0.47±0.166	0.30±0.396	0.50±0.180	0.32±0.396
Coefficient of similarity	0.44±0.077	0.06±0.017	0.46±0.107	0.08±0.083	0.51±0.145	0.14±0.185
Canberra	0.42±0.063	0.05±0.010	0.44±0.081	0.06±0.027	0.49±0.128	0.10±0.115
uSoergel	0.40±0.028	0.05±0.003	0.41±0.028	0.05±0.003	0.37±0.018	0.05±0.006
uTamàs coefficient	0.40±0.028	0.05±0.003	0.41±0.031	0.05±0.003	0.37±0.020	0.05±0.006
uManhattan	0.40±0.032	0.05±0.003	0.41±0.028	0.05±0.003	0.37±0.023	0.05±0.006
uMNND	0.40±0.027	0.05±0.003	0.41±0.032	0.05±0.003	0.37±0.023	0.05±0.006
uGower	0.40±0.026	0.05±0.003	0.40±0.030	0.05±0.003	0.37±0.021	0.05±0.006
uEuclidean	0.40±0.026	0.05±0.003	0.40±0.029	0.05±0.003	0.37±0.021	0.05±0.006
uPearson dissimilarity	0.40±0.026	0.05±0.003	0.40±0.027	0.05±0.003	0.37±0.020	0.05±0.007
MNND	0.41±0.032	0.05±0.003	0.40±0.028	0.05±0.003	0.41±0.030	0.05±0.008
uCanberra	0.40±0.030	0.05±0.003	0.40±0.027	0.05±0.003	0.37±0.023	0.05±0.006
uWeighted correlation	0.40±0.030	0.05±0.003	0.40±0.026	0.05±0.003	0.37±0.022	0.05±0.006
uBray-Curtis (MRCA restricted)	0.40±0.026	0.05±0.003	0.40±0.030	0.05±0.003	0.37±0.021	0.05±0.006
uBray-Curtis	0.40±0.030	0.05±0.003	0.40±0.027	0.05±0.003	0.37±0.019	0.05±0.006
uCoefficient of similarity	0.40±0.031	0.05±0.003	0.40±0.029	0.05±0.003	0.37±0.019	0.05±0.006
uKulczynski	0.40±0.029	0.05±0.003	0.40±0.027	0.05±0.003	0.37±0.021	0.05±0.006
uLennon	0.40±0.029	0.05±0.003	0.40±0.030	0.05±0.003	0.36±0.012	0.05±0.008
uMPD	0.37±0.027	0.05±0.003	0.36±0.018	0.05±0.003	0.35±0.006	0.05±0.007

Table C.17. Results for dominant-pair model on samples from the keyboard dataset.

The mean and standard deviation are given for the *k*-medoids score (KMS) and consistency index (CI) statistics. Pearson's correlation coefficient between the 2 statistics is r = 0.98, 0.99, 0.99 for 100, 1,000, and 10,000 sequences, respectively.

Results are sorted by KMS for samples with 1,000 sequences.

Mean ± s.d.	**100 sequences**		**1,000 sequences**		**10,000 sequences**	
	KMS	*CI*	*KMS*	*CI*	*KMS*	*CI*
Euclidean	0.57±0.105	0.07±0.022	0.78±0.149	0.30±0.296	0.85±0.139	0.48±0.339
Yue-Clayton	0.56±0.109	0.07±0.022	0.78±0.145	0.29±0.289	0.84±0.153	0.48±0.340
Weighted correlation	0.57±0.106	0.07±0.023	0.78±0.155	0.30±0.305	0.85±0.148	0.49±0.344
F_{ST}	0.57±0.105	0.07±0.022	0.78±0.150	0.29±0.289	0.86±0.133	0.48±0.343
Rao's H_p	0.57±0.104	0.07±0.023	0.78±0.149	0.30±0.296	0.85±0.149	0.47±0.340
Manhattan	0.54±0.110	0.07±0.021	0.77±0.148	0.27±0.278	0.84±0.135	0.44±0.340
Morisita-Horn	0.57±0.105	0.07±0.023	0.77±0.156	0.30±0.293	0.86±0.130	0.47±0.334
Kulczynski	0.54±0.113	0.07±0.020	0.77±0.155	0.26±0.278	0.83±0.143	0.44±0.344
Tamàs coefficient	0.54±0.113	0.07±0.021	0.77±0.151	0.27±0.278	0.82±0.159	0.44±0.340
Bray-Curtis (MRCA restricted)	0.54±0.109	0.07±0.020	0.76±0.154	0.27±0.286	0.84±0.142	0.44±0.345
Whittaker	0.55±0.107	0.07±0.021	0.76±0.155	0.27±0.278	0.83±0.150	0.44±0.340
Complete tree	0.54±0.113	0.07±0.021	0.76±0.156	0.27±0.278	0.84±0.138	0.44±0.340
Soergel	0.54±0.113	0.07±0.020	0.76±0.153	0.26±0.277	0.83±0.153	0.44±0.345
Bray-Curtis	0.54±0.105	0.07±0.020	0.76±0.157	0.27±0.278	0.83±0.141	0.44±0.344
Pearson dissimilarity	0.55±0.105	0.07±0.020	0.74±0.154	0.23±0.236	0.82±0.148	0.40±0.326
Lennon	0.53±0.097	0.07±0.019	0.73±0.166	0.24±0.251	0.80±0.158	0.40±0.336
Chi-squared	0.47±0.075	0.06±0.011	0.65±0.168	0.22±0.232	0.72±0.194	0.43±0.339
Hellinger	0.46±0.074	0.06±0.015	0.63±0.161	0.21±0.225	0.73±0.198	0.41±0.336
Gower	0.44±0.057	0.05±0.007	0.50±0.090	0.08±0.041	0.54±0.103	0.11±0.070
Coefficient of similarity	0.42±0.039	0.05±0.005	0.43±0.049	0.05±0.009	0.49±0.074	0.07±0.027
Canberra	0.42±0.033	0.05±0.005	0.43±0.048	0.05±0.006	0.47±0.080	0.07±0.018
MPD	0.40±0.088	0.06±0.013	0.41±0.116	0.14±0.186	0.42±0.125	0.21±0.270
uSoergel	0.40±0.031	0.05±0.003	0.40±0.032	0.05±0.003	0.36±0.015	0.05±0.008
MNND	0.41±0.026	0.05±0.004	0.40±0.026	0.05±0.003	0.40±0.024	0.05±0.007
uBray-Curtis	0.40±0.029	0.05±0.003	0.40±0.029	0.05±0.003	0.36±0.016	0.05±0.008
uCoefficient of similarity	0.40±0.030	0.05±0.003	0.40±0.032	0.05±0.003	0.36±0.016	0.05±0.008
uBray-Curtis (MRCA restricted)	0.40±0.028	0.05±0.003	0.40±0.027	0.05±0.003	0.36±0.015	0.05±0.008
uKulczynski	0.40±0.029	0.05±0.003	0.40±0.026	0.05±0.003	0.36±0.016	0.05±0.008
uTamàs coefficient	0.40±0.029	0.05±0.003	0.40±0.031	0.05±0.003	0.36±0.018	0.05±0.008
uEuclidean	0.40±0.025	0.05±0.003	0.40±0.032	0.05±0.003	0.36±0.016	0.05±0.008
uGower	0.40±0.028	0.05±0.003	0.40±0.027	0.05±0.003	0.36±0.016	0.05±0.008
uMNND	0.40±0.031	0.05±0.003	0.40±0.026	0.05±0.003	0.36±0.017	0.05±0.007
uWeighted correlation	0.40±0.029	0.05±0.003	0.40±0.028	0.05±0.003	0.36±0.014	0.05±0.008
uManhattan	0.40±0.029	0.05±0.003	0.40±0.028	0.05±0.003	0.36±0.016	0.05±0.008
uPearson dissimilarity	0.40±0.026	0.05±0.003	0.40±0.024	0.05±0.003	0.36±0.015	0.05±0.007
uLennon	0.40±0.028	0.05±0.003	0.40±0.026	0.05±0.003	0.35±0.009	0.05±0.008
uCanberra	0.40±0.028	0.05±0.003	0.39±0.028	0.05±0.003	0.36±0.018	0.05±0.008
uMPD	0.37±0.025	0.05±0.003	0.36±0.016	0.05±0.003	0.35±0.009	0.05±0.007

Table C.18. Results for dominant-pair model on samples from the mouse gut dataset.

The mean and standard deviation are given for the *k*-medoids score (KMS) and consistency index (CI) statistics. Pearson's correlation coefficient between the 2 statistics is r = 0.98, 0.99, 0.98 for 100, 1,000, and 10,000 sequences, respectively.

Results are sorted by KMS for samples with 1,000 sequences.

Mean ± s.d.	**100 sequences**		**1,000 sequences**		**10,000 sequences**	
	KMS	*CI*	*KMS*	*CI*	*KMS*	*CI*
Weighted correlation	0.80±0.192	0.38±0.287	0.88±0.181	0.76±0.306	0.88±0.199	0.80±0.279
Morisita-Horn	0.80±0.195	0.38±0.287	0.88±0.179	0.74±0.309	0.87±0.210	0.78±0.294
Rao's H_p	0.79±0.191	0.37±0.285	0.87±0.180	0.73±0.314	0.86±0.217	0.76±0.297
Yue-Clayton	0.80±0.194	0.38±0.296	0.87±0.181	0.73±0.315	0.86±0.213	0.76±0.300
Euclidean	0.79±0.193	0.37±0.278	0.87±0.185	0.74±0.314	0.89±0.178	0.77±0.299
F_{ST}	0.80±0.188	0.36±0.271	0.87±0.187	0.73±0.312	0.87±0.194	0.76±0.302
Pearson dissimilarity	0.75±0.222	0.31±0.255	0.82±0.227	0.65±0.357	0.82±0.237	0.69±0.354
Hellinger	0.67±0.179	0.20±0.168	0.81±0.217	0.65±0.364	0.82±0.230	0.72±0.333
Manhattan	0.73±0.227	0.27±0.225	0.80±0.233	0.60±0.385	0.80±0.237	0.64±0.383
Chi-squared	0.69±0.184	0.19±0.158	0.80±0.231	0.68±0.353	0.83±0.221	0.73±0.325
Kulczynski	0.73±0.214	0.27±0.230	0.80±0.235	0.59±0.384	0.80±0.245	0.64±0.385
Whittaker	0.73±0.228	0.27±0.225	0.80±0.233	0.60±0.385	0.80±0.239	0.64±0.383
Complete tree	0.74±0.222	0.27±0.225	0.80±0.232	0.60±0.385	0.80±0.240	0.64±0.383
Bray-Curtis (MRCA restricted)	0.73±0.225	0.27±0.234	0.79±0.240	0.59±0.384	0.80±0.245	0.64±0.386
Tamàs coefficient	0.73±0.224	0.27±0.225	0.79±0.238	0.60±0.385	0.81±0.234	0.64±0.383
Soergel	0.73±0.222	0.27±0.228	0.79±0.242	0.59±0.383	0.80±0.245	0.64±0.385
Bray-Curtis	0.73±0.227	0.27±0.228	0.79±0.248	0.59±0.383	0.81±0.235	0.64±0.385
Lennon	0.62±0.185	0.16±0.111	0.73±0.238	0.52±0.378	0.75±0.243	0.56±0.383
Gower	0.50±0.112	0.07±0.022	0.53±0.133	0.13±0.122	0.57±0.173	0.16±0.181
MPD	0.49±0.150	0.11±0.081	0.52±0.178	0.32±0.341	0.47±0.151	0.37±0.380
Coefficient of similarity	0.46±0.069	0.06±0.014	0.48±0.092	0.09±0.099	0.54±0.140	0.14±0.131
Canberra	0.44±0.055	0.06±0.011	0.46±0.070	0.07±0.026	0.55±0.138	0.11±0.070
uBray-Curtis	0.41±0.027	0.05±0.003	0.41±0.035	0.05±0.003	0.35±0.007	0.13±0.060
uMNND	0.40±0.028	0.05±0.003	0.40±0.032	0.05±0.003	0.35±0.007	0.13±0.059
uEuclidean	0.40±0.031	0.05±0.003	0.40±0.032	0.05±0.003	0.35±0.004	0.13±0.060
uGower	0.40±0.030	0.05±0.003	0.40±0.033	0.05±0.003	0.35±0.007	0.13±0.060
uKulczynski	0.40±0.027	0.05±0.003	0.40±0.028	0.05±0.003	0.35±0.007	0.13±0.060
uCanberra	0.40±0.031	0.05±0.003	0.40±0.031	0.05±0.003	0.35±0.006	0.13±0.060
uManhattan	0.41±0.029	0.05±0.003	0.40±0.032	0.05±0.003	0.35±0.007	0.13±0.060
uBray-Curtis (MRCA restricted)	0.41±0.031	0.05±0.003	0.40±0.033	0.05±0.003	0.35±0.008	0.13±0.060
uWeighted correlation	0.41±0.030	0.05±0.003	0.40±0.028	0.05±0.003	0.35±0.008	0.13±0.060
uCoefficient of similarity	0.40±0.028	0.05±0.003	0.40±0.030	0.05±0.003	0.35±0.009	0.13±0.060
MNND	0.40±0.030	0.05±0.004	0.40±0.029	0.05±0.003	0.39±0.022	0.12±0.055
uPearson dissimilarity	0.40±0.027	0.05±0.003	0.40±0.032	0.05±0.003	0.35±0.005	0.13±0.060
uTamàs coefficient	0.40±0.029	0.05±0.003	0.40±0.032	0.05±0.003	0.35±0.007	0.13±0.060
uSoergel	0.41±0.027	0.05±0.003	0.40±0.033	0.05±0.003	0.35±0.007	0.13±0.060
uLennon	0.40±0.023	0.05±0.003	0.40±0.030	0.05±0.003	0.35±0.007	0.14±0.075
uMPD	0.37±0.027	0.05±0.003	0.36±0.012	0.05±0.003	0.35±0.005	0.13±0.056

Table C.19. Results for dominant-pair model on samples from the soil dataset.

The mean and standard deviation are given for the *k*-medoids score (KMS) and consistency index (CI) statistics. Pearson's correlation coefficient between the 2 statistics is r = 0.98, 0.99, 0.99 for 100, 1,000, and 10,000 sequences, respectively.

Results are sorted by KMS for samples with 1,000 sequences.

Mean ± s.d.	**100 sequences**		**1,000 sequences**		**10,000 sequences**	
	KMS	*CI*	*KMS*	*CI*	*KMS*	*CI*
Morisita-Horn	0.54±0.152	0.07±0.034	0.69±0.220	0.29±0.341	0.72±0.224	0.39±0.371
Euclidean	0.53±0.147	0.07±0.033	0.69±0.217	0.30±0.352	0.72±0.221	0.40±0.384
Weighted correlation	0.54±0.153	0.07±0.034	0.68±0.223	0.31±0.353	0.72±0.223	0.39±0.379
F_{ST}	0.53±0.155	0.07±0.033	0.68±0.219	0.30±0.347	0.71±0.232	0.39±0.384
Rao's H_p	0.53±0.146	0.07±0.032	0.68±0.217	0.29±0.342	0.71±0.228	0.39±0.382
Yue-Clayton	0.53±0.148	0.07±0.035	0.67±0.219	0.29±0.342	0.71±0.228	0.40±0.387
Soergel	0.52±0.140	0.08±0.041	0.67±0.221	0.29±0.354	0.69±0.226	0.38±0.393
Tamàs coefficient	0.51±0.136	0.07±0.039	0.66±0.225	0.30±0.365	0.70±0.230	0.38±0.397
Kulczynski	0.51±0.139	0.07±0.041	0.66±0.220	0.29±0.350	0.69±0.223	0.38±0.397
Bray-Curtis	0.52±0.137	0.08±0.041	0.66±0.220	0.29±0.354	0.70±0.219	0.38±0.393
Complete tree	0.50±0.126	0.07±0.039	0.66±0.221	0.30±0.365	0.70±0.220	0.38±0.397
Bray-Curtis (MRCA restricted)	0.51±0.133	0.07±0.040	0.66±0.215	0.30±0.360	0.71±0.227	0.38±0.393
Pearson dissimilarity	0.53±0.157	0.07±0.036	0.66±0.233	0.29±0.359	0.69±0.235	0.37±0.391
Whittaker	0.51±0.137	0.07±0.039	0.65±0.225	0.30±0.365	0.70±0.226	0.38±0.397
Manhattan	0.51±0.136	0.07±0.039	0.65±0.226	0.30±0.365	0.69±0.225	0.38±0.397
Lennon	0.49±0.122	0.07±0.034	0.62±0.209	0.24±0.306	0.67±0.224	0.34±0.379
Chi-squared	0.46±0.092	0.05±0.009	0.59±0.196	0.25±0.324	0.66±0.214	0.38±0.399
Hellinger	0.46±0.081	0.06±0.020	0.57±0.189	0.26±0.339	0.64±0.212	0.38±0.396
Gower	0.43±0.054	0.05±0.006	0.44±0.060	0.06±0.016	0.48±0.084	0.07±0.030
MPD	0.43±0.110	0.05±0.009	0.44±0.148	0.07±0.034	0.46±0.164	0.09±0.066
Coefficient of similarity	0.41±0.038	0.05±0.005	0.42±0.047	0.05±0.009	0.45±0.063	0.06±0.024
Canberra	0.41±0.034	0.05±0.005	0.41±0.035	0.05±0.006	0.44±0.064	0.06±0.016
uTamàs coefficient	0.40±0.025	0.05±0.003	0.41±0.029	0.05±0.003	0.38±0.025	0.05±0.004
uCoefficient of similarity	0.40±0.025	0.05±0.003	0.41±0.030	0.05±0.003	0.38±0.026	0.05±0.004
uGower	0.40±0.029	0.05±0.003	0.41±0.032	0.05±0.003	0.38±0.031	0.05±0.004
uEuclidean	0.40±0.027	0.05±0.003	0.41±0.033	0.05±0.003	0.38±0.026	0.05±0.003
uSoergel	0.40±0.029	0.05±0.003	0.41±0.029	0.05±0.003	0.38±0.027	0.05±0.004
uMNND	0.40±0.028	0.05±0.003	0.40±0.033	0.05±0.003	0.38±0.026	0.05±0.004
uWeighted correlation	0.40±0.029	0.05±0.004	0.40±0.031	0.05±0.003	0.38±0.029	0.05±0.004
uCanberra	0.40±0.026	0.05±0.003	0.40±0.031	0.05±0.003	0.38±0.026	0.05±0.004
uBray-Curtis	0.41±0.033	0.05±0.003	0.40±0.028	0.05±0.003	0.38±0.026	0.05±0.004
MNND	0.41±0.032	0.05±0.003	0.40±0.027	0.05±0.003	0.39±0.024	0.05±0.003
uBray-Curtis (MRCA restricted)	0.40±0.025	0.05±0.003	0.40±0.025	0.05±0.003	0.38±0.029	0.05±0.004
uPearson dissimilarity	0.40±0.028	0.05±0.003	0.40±0.030	0.05±0.003	0.38±0.026	0.05±0.003
uKulczynski	0.41±0.029	0.05±0.003	0.40±0.029	0.05±0.003	0.38±0.028	0.05±0.004
uManhattan	0.40±0.023	0.05±0.003	0.40±0.027	0.05±0.003	0.38±0.026	0.05±0.004
uLennon	0.40±0.032	0.05±0.003	0.39±0.028	0.05±0.003	0.37±0.025	0.05±0.004
uMPD	0.37±0.025	0.05±0.003	0.36±0.013	0.05±0.003	0.35±0.005	0.05±0.004

Table C.20. Mean performance over both models of diversification and all empirical datasets.

The mean is given for the *k*-medoids score (KMS) and consistency index (CI) statistics. Pearson's correlation coefficient between the 2 statistics is r = 0.98, 0.89, 0.96 for 100, 1,000, and 10,000 sequences, respectively.

Results are sorted by KMS for samples with 1,000 sequences.

Mean	**100 sequences**		**1,000 sequences**		**10,000 sequences**	
	KMS	*CI*	*KMS*	*CI*	*KMS*	*CI*
Hellinger	0.622	0.170	0.791	0.539	0.829	0.619
Chi-squared	0.610	0.136	0.788	0.520	0.826	0.613
Complete tree	0.623	0.139	0.772	0.409	0.802	0.501
Kulczynski	0.625	0.140	0.772	0.400	0.800	0.501
Soergel	0.625	0.142	0.772	0.400	0.802	0.504
Whittaker	0.626	0.139	0.771	0.409	0.800	0.501
Tamàs coefficient	0.621	0.139	0.771	0.409	0.800	0.501
Manhattan	0.624	0.139	0.770	0.409	0.799	0.501
Bray-Curtis (MRCA restricted)	0.621	0.139	0.769	0.399	0.802	0.501
Bray-Curtis	0.625	0.142	0.769	0.401	0.805	0.501
Euclidean	0.625	0.145	0.756	0.339	0.789	0.431
Weighted correlation	0.629	0.146	0.754	0.351	0.782	0.446
Lennon	0.601	0.126	0.754	0.399	0.794	0.492
F_{ST}	0.628	0.143	0.751	0.337	0.782	0.428
Morisita-Horn	0.630	0.146	0.751	0.341	0.780	0.436
Rao's H_p	0.624	0.144	0.748	0.335	0.776	0.429
Yue-Clayton	0.625	0.145	0.745	0.336	0.777	0.431
Pearson	0.609	0.133	0.724	0.297	0.754	0.385
Gower	0.527	0.104	0.694	0.415	0.747	0.481
Coefficient of similarity	0.499	0.103	0.674	0.419	0.737	0.510
Canberra	0.496	0.099	0.666	0.412	0.736	0.509
MNND	0.509	0.113	0.646	0.400	0.636	0.405
uCoefficient of similarity	0.464	0.076	0.585	0.265	0.527	0.215
uTamàs coefficient	0.463	0.076	0.584	0.265	0.527	0.215
uCanberra	0.465	0.076	0.583	0.265	0.535	0.215
uEuclidean	0.464	0.076	0.581	0.265	0.527	0.215
uGower	0.465	0.076	0.579	0.265	0.535	0.215
uManhattan	0.465	0.076	0.579	0.265	0.532	0.215
uSoergel	0.476	0.081	0.568	0.261	0.528	0.211
uBray-Curtis	0.480	0.081	0.566	0.261	0.528	0.211
uBray-Curtis (MRCA restricted)	0.479	0.081	0.566	0.261	0.529	0.211
uWeighted correlation	0.475	0.081	0.564	0.260	0.524	0.213
uKulczynski	0.473	0.080	0.563	0.261	0.526	0.213
uMNND	0.470	0.084	0.563	0.281	0.525	0.231
uPearson	0.448	0.063	0.524	0.120	0.510	0.125
uLennon	0.443	0.069	0.508	0.205	0.454	0.237
MPD	0.449	0.084	0.457	0.146	0.461	0.169
uMPD	0.381	0.050	0.369	0.050	0.354	0.066

Table C.21. Relative performance of measures on different datasets under the equal-perturbation model.

Comparison among KMS results with 1,000 sequences/sample. The statistically significant *p*-values of the paired t-test indicate a directional change in the absolute performance of measures between datasets. However, the relative performance of measures is fairly stable across datasets as indicated by the high Pearson's and Spearman's correlation values.

Dataset	*Paired t-test* (p-value)	*Pearson's correlation (r)*	*Spearman's correlation (r)*
human vs. keyboard	$4.61 \cdot 10^{-6}$	0.95	0.83
human vs. mouse gut	$3.49 \cdot 10^{-8}$	0.84	0.66
human vs. soil	$6.81 \cdot 10^{-16}$	0.93	0.84
keyboard vs. mouse gut	$2.39 \cdot 10^{-7}$	0.95	0.89
keyboard vs. soil	$2.25 \cdot 10^{-18}$	0.91	0.94
mouse gut vs. soil	$3.89 \cdot 10^{-17}$	0.83	0.82

Table C.22. Relative performance of measures on different datasets under the dominant-pair model.

Comparison of KMS results with 1,000 sequences/sample. The statistically significant *p*-values of the paired t-test indicate a directional change in the absolute performance of measures between datasets. However, the relative performance of measures is fairly stable across datasets as indicated by the high Pearson's and Spearman's correlation values.

Dataset	*Paired t-test* (p-value)	*Pearson's correlation (r)*	*Spearman's correlation (r)*
human vs. keyboard	$7.60 \cdot 10^{-2}$	0.99	0.93
human vs. mouse gut	$8.66 \cdot 10^{-6}$	0.99	0.91
human vs. soil	$1.53 \cdot 10^{-7}$	0.99	0.94
keyboard vs. mouse gut	$7.29 \cdot 10^{-4}$	0.98	0.91
keyboard vs. soil	$6.13 \cdot 10^{-7}$	1.00	0.92
mouse gut vs. soil	$2.80 \cdot 10^{-7}$	0.98	0.93

Appendix D

Normalized Weighted UniFrac is Equivalent to the Phylogenetic Bray-Curtis Semimetric

Weighted UniFrac is defined as follows (Lozupone et al. 2007):

$$\mu_{ij} = \sum_{n=1}^{N} \left| p_{in} - p_{jn} \right| W_n$$

where N is the number of branches in the tree, W_n is the length of branch n, and p_{in} is the proportion of sequences from community i which are descendant from branch n. The following scaling factor is used to obtain a normalized measure of dissimilarity:

$$\Lambda_{ij} = \sum_{s=1}^{S} d_s \left(q_{is} + q_{js} \right)$$

where S is the number of leaf nodes in the tree, d_s is the distance from leaf node s to the root, and q_{is} is the proportion of sequences from community i assigned to leaf node s. This normalization factor can be expressed in terms of weighted branch lengths as follows:

$$\sum_{s=1}^{S} d_s \left(q_{is} + q_{js} \right) = \sum_{s=1}^{S} \left(q_{is} + q_{js} \right) \sum_{n \in B_s} W_n$$

where B_s is the set of branch indices which lead from leaf node s to the root. Each branch n will be present in one or more of the sets B_1, B_2,…, B_{S-1}, B_S. We denote the indices of the sets B_1, B_2,…, B_{S-1}, B_S which contain branch n as T_n. Using these sets, the above expression can be written as:

$$= W_1 \sum_{s \in T_1} \left(q_{is} + q_{js} \right) + W_2 \sum_{s \in T_2} \left(q_{is} + q_{js} \right) + \cdots + W_{N-1} \sum_{s \in T_{N-1}} \left(q_{is} + q_{js} \right) + W_N \sum_{s \in T_N} \left(q_{is} + q_{js} \right)$$

$$= \sum_{n=1}^{N} W_n \left(\sum_{s \in T_n} q_{is} + \sum_{s \in T_n} q_{js} \right)$$

The term $\sum_{s \in T_n} q_{is}$ is the proportion of sequences from community *i* which are descendant from branch *n* which is the definition of p_{in}. As such, the above equation can be expressed as:

$$\sum_{n=1}^{N} \left(p_{in} + p_{jn} \right) W_n$$

Normalized weighted UniFrac can therefore be expressed as:

$$\frac{\sum_{n=1}^{N} \left| p_{in} - p_{jn} \right| W_n}{\sum_{s=1}^{S} d_s \left(q_{is} + q_{js} \right)} = \frac{\sum_{n=1}^{N} \left| p_{in} - p_{jn} \right| W_n}{\sum_{n=1}^{N} \left(p_{in} + p_{jn} \right) W_n}$$

which is the phylogenetic extension of the Bray-Curtis semimetric.

Appendix E

Classification of Splits

Table E.1. Enumeration of all possible splits.

Enumeration of all possible splits with respect to communities *i* and *j* within a split system containing sequences from one or more other ingroup communities, and sequences from an outgroup. The set of sequences from community *i* is denoted by *A*, sequences from community *j* are denoted by *B*, sequences from any other ingroup community are denoted by *C*, and sequences from the outgroup are denoted by *O*. Splits are classified as unique to community *i* (U_i), unique to community *j* (U_j), shared (S), root (R), external (E), or outgroup (O). Splits which cannot occur are marked as "not possible" (NP). A monophyletic outgroup is assumed. For example, the row ABC indicates a subset of taxa which contains all ingroup taxa and at least one taxon from another community. The column CO indicates a subset of taxa which contains all outgroup taxa and at least one taxon from another community. The cell at the intersection of this row and column indicates a split of type ABC|CO which is a root split.

	A	*B*	*C*	*O*	*AB*	*AC*	*AO*	*BC*	*BO*	*CO*	*ABC*	*ABO*	*ACO*	*BCO*	*ABCO*
A	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	Ui	Ui
B	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	Uj	NP	Uj
C	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	E	NP	NP	E
O	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	O	NP	NP	NP	O
AB	NP	NP	NP	NP	NP	NP	NP	NP	NP	R	NP	NP	S	S	S
AC	NP	NP	NP	NP	NP	NP	NP	NP	Ui	NP	NP	Ui	NP	Ui	Ui
AO	NP	NP	NP	NP	NP	NP	NP	Uj	NP	NP	S	NP	NP	NP	NP
BC	NP	NP	NP	NP	NP	NP	Uj	NP	NP	NP	NP	Uj	Uj	NP	Uj
BO	NP	NP	NP	NP	NP	Ui	NP	NP	NP	NP	S	NP	NP	NP	NP
CO	NP	NP	NP	NP	R	NP	NP	NP	NP	NP	R	NP	NP	NP	NP
ABC	NP	NP	NP	O	NP	NP	S	NP	S	R	NP	S	S	S	S
ABO	NP	NP	E	NP	NP	Ui	NP	Uj	NP	NP	S	NP	NP	NP	NP
ACO	NP	Uj	NP	NP	S	NP	NP	Uj	NP	NP	S	NP	NP	NP	NP
BCO	Ui	NP	NP	NP	S	Ui	NP	NP	NP	NP	S	NP	NP	NP	NP
ABCO	Ui	Uj	E	O	S	Ui	NP	Uj	NP	NP	S	NP	NP	NP	NP

Appendix F

Pneumococcus Samples

Table F.1. Sampling of *Streptococcus pneumoniae* serotype 1 isolates.

Sample Id	Location	Continent	# Seqs	Reference
Québec	Québec, Canada	North America	10	Brueggemann and Spratt 2003
Toronto	Toronto, Canada	North America	10	Brueggemann and Spratt 2003
Chile	Chile	South America	11	Brueggemann and Spratt 2003
Czech Republic	Czech Republic	Europe	8	Zemlickova et al. 2010
Denmark	Denmark	Europe	9	Brueggemann and Spratt 2003
England	England	Europe	10	Brueggemann and Spratt 2003
France	France	Europe	10	Brueggemann and Spratt 2003
Germany	Germany	Europe	10	Brueggemann and Spratt 2003
Ghana	Ghana	Africa	68	Leimkugel et al. 2005
India	India	Asia	13	MLST.net[2]
Israel	Israel	Asia	12	Brueggemann and Spratt 2003
Kenya	Kenya	Africa	12	Brueggemann and Spratt 2003
Mozambique	Mozambique	Africa	14	MLST.net[2]
Niger	Niger	Africa	31	MLST.net[2]
Norway	Norway	Europe	10	Brueggemann and Spratt 2003
Poland	Poland	Europe	7	Brueggemann and Spratt 2003
Scotland	Scotland	Europe	41	McChlery et al. 2005
South Africa	South Africa	Africa	10	Brueggemann and Spratt 2003
Spain	Spain	Europe	14	Brueggemann and Spratt 2003
Spain (BSM)	Spain[1]	Europe	42	Obando et al. 2008
Barcelona	Barcelona, Spain	Europe	55	Esteva et al. 2011
Gipuzkoa	Gipuzkoa, Spain	Europe	134	Marimon et al. 2009
Sweden	Sweden	Europe	20	Henriques Normark et al. 2001
Thailand	Thailand	Asia	14	MLST.net[2]
The Gambia	The Gambia	Africa	163	Antonio et al. 2008
The Netherlands	The Netherlands	Europe	13	Brueggemann and Spratt 2003
Utah	Utah, USA	North America	21	Byington et al. 2005
USA	USA	North America	10	Brueggemann and Spratt 2003
Navajo Indians	USA	North America	8	Brueggemann and Spratt 2003

[1] isolates collected from Barcelona, Seville, and Malaga

[2] isolates and associated metadata obtained from MLST.net

Appendix G

Proteorhodopsin Samples

Table G.1. Proteorhodopsin samples from the Mediterranean and Sargasso Seas.

Sample Id	# Seqs	Spectrum	Stratification	Location	Depth (m)	Date	Station
Med0-B-S	58	Blue	Stratified	Mediterranean	0	May-03	H01
Med 0-B-M	58	Blue	Mixed	Mediterranean	0	Jan-06	H01
Med20-B-S	95	Blue	Stratified	Mediterranean	20	May-03	H01
Med20-B-M	71	Blue	Mixed	Mediterranean	20	Jan-06	H01
Med55-B-S	69	Blue	Stratified	Mediterranean	55	May-03	H01
Med50-B-M	61	Blue	Mixed	Mediterranean	50	Jan-06	H01
Med0-B-M	77	Blue	Mixed	Mediterranean	0	Feb-06	TB04
Med20-B-M	83	Blue	Mixed	Mediterranean	20	Feb-06	TB04
Med50-B-M	74	Blue	Mixed	Mediterranean	50	Feb-06	TB04
Sar0-B-S	91	Blue	Stratified	Sargasso	0	Jul-98	BATS
Sar0-B-M	58	Blue	Mixed	Sargasso	0	Mar-98	BATS
Sar40-B-S	82	Blue	Stratified	Sargasso	40	Jul-98	BATS
Sar40-B-M	58	Blue	Mixed	Sargasso	40	Mar-98	BATS
Sar80-B-S	82	Blue	Stratified	Sargasso	80	Jul-98	BATS
Sar80-B-M	40	Blue	Mixed	Sargasso	80	Mar-03	BATS
Med0-G-S	91	Green	Stratified	Mediterranean	0	May-03	H01
Med0-G-M	32	Green	Mixed	Mediterranean	0	Jan-06	H01
Med20-G-S	74	Green	Stratified	Mediterranean	20	May-03	H01
Med20-G-M	32	Green	Mixed	Mediterranean	20	Jan-06	H01
Med55-G-S	7	Green	Stratified	Mediterranean	55	May-03	H01
Med50-G-M	42	Green	Mixed	Mediterranean	50	Jan-06	H01
Med0-G-M	15	Green	Mixed	Mediterranean	0	Feb-06	TB04
Med20-G-M	9	Green	Mixed	Mediterranean	20	Feb-06	TB04
Med50-G-M	19	Green	Mixed	Mediterranean	50	Feb-06	TB04

Appendix H

Publications

H.1 Published or Accepted Manuscripts (Discussed in Thesis)

1. **Parks DH**, Porter M, Churcher S, Wang S, Blouin C, Whalley J, Brooks S, Beiko RG. 2009. GenGIS: a geospatial information system for genomic data. Genome Res. 19:1896-1904.
   - Discussed in Chapter 3

2. **Parks DH**, Beiko RG. 2010. Quantitative visualizations of hierarchically organized data in a geographic context. 17th International Conference on Geoinformatics (Fairfax, VA): 1-6.
   - Discussed in Chapter 2

H.2 Submitted Manuscripts (Discussed in Thesis)

1. **Parks DH**, Mankowski T, Porter MS, Beiko RG. 2012. GenGIS 2: Geospatial analysis of genetic and genomic datasets, with new gradient algorithms and an extensible framework. In preparation.
   - Discussed in Chapters 2 and 3

2. **Parks DH**, Beiko RG. 2012a. Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities. In press at ISME J, July 2012.
   - Discussed in Chapter 4

3. **Parks DH**, Beiko RG. 2012b. Measuring community similarity with phylogenetic networks. In press at Mol. Biol. Evol., July 2012.
   - Discussed in Chapter 5

H.3 Published or Accepted Manuscripts (Not Discussed in Thesis)

1. **Parks DH**, Beiko RG. 2010. Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26:715-721.

Motivation: Metagenomics is the study of genetic material recovered directly from environmental samples. Taxonomic and functional differences between metagenomic samples can highlight the influence of ecological factors on patterns of microbial life in a wide range of habitats. Statistical hypothesis tests can help us distinguish ecological influences from sampling artifacts, but knowledge of only the *p*-value from a statistical hypothesis test is insufficient to make inferences about biological relevance. Current reporting practices for pairwise comparative metagenomics are inadequate, and better tools are needed for comparative metagenomic analysis.

Results: We have developed STAMP, a new software package for comparative metagenomics that supports best practices in analysis and reporting. Examination of a pair of iron mine metagenomes demonstrates that deeper biological insights can be gained using the statistical techniques available in our software. An analysis of the functional potential of "*Candidatus* Accumulibacter phosphatis" in two enhanced biological phosphorus removal metagenomes identified several subsystems that differ between the *A. phosphatis* stains in these related communities, including phosphate metabolism, secretion and metal transport.

2. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartman M, Hollister EB, Lesniewski RA, Oakley BB, **Parks DH**, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: Open source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75:7537-41.

"mothur" aims to be a comprehensive software package that allows users to use a single piece of software to analyze community sequence data. It builds upon previous tools to provide a flexible and powerful software package for analyzing sequencing data. As a case study, we used mothur to trim, screen, and align sequences, calculate distances, assign sequences to OTUs, and describe the alpha- and beta-diversity of eight marine samples previously characterized by pyrosequencing of 16S rRNA gene

fragments. This analysis of more than 222,000 sequences was completed in less than 2 hours using a laptop computer.

3. **Parks DH**[*], MacDonald NJ[*], Beiko RG. 2011. Classifying short genomic fragments from novel lineages using composition and homology. Bioinformatics 12:328.
[*] these authors contributed equally to this work

Background: The assignment of taxonomic attributions to DNA fragments recovered directly from the environment is a vital step in metagenomic data analysis. Assignments can be made using *rank-specific* classifiers, which assign reads to taxonomic labels from a predetermined level such as named species or strain, or *rank-flexible* classifiers, which choose an appropriate taxonomic rank for each sequence in a dataset. The choice of rank typically depends on the optimal model for a given sequence and on the breadth of taxonomic groups seen in a set of close-to-optimal models. Homology-based (e.g., LCA) and composition-based (e.g., PhyloPythia, TACOA) rank-flexible classifiers have been proposed, but there is at present no hybrid approach that utilizes both homology and composition.

Results: We first develop a hybrid, rank-specific classifier based on BLAST and Naïve Bayes (NB) that has comparable accuracy and a faster running time than the current best approach, PhymmBL. By substituting LCA for BLAST or allowing the inclusion of suboptimal NB models, we obtain a rank-flexible classifier. This hybrid classifier outperforms established rank-flexible approaches on simulated metagenomic fragments of length 200 bp to 1000 bp and is able to assign taxonomic attributions to a subset of sequences with few misclassifications. We then demonstrate the performance of different classifiers on an enhanced biological phosphorous removal metagenome, illustrating the advantages of rank-flexible classifiers when representative genomes are absent from the set of reference genomes. Application to a glacier ice metagenome demonstrates that similar taxonomic profiles are obtained across a set of classifiers which are increasingly conservative in their classification.

Conclusions: Our NB-based classification scheme is faster than the current best composition-based algorithm, Phymm, while providing equally accurate predictions. The rank-flexible variant of NB, which we term ε-NB, is complementary to LCA and can be

combined with it to yield conservative prediction sets of very high confidence. The simple parameterization of LCA and ε-NB allows for tuning of the balance between more predictions and increased precision, allowing the user to account for the sensitivity of downstream analyses to misclassified or unclassified sequences.

4. MacDonald NJ*, **Parks DH***, Beiko RG. 2011. Rapid identification of high-confidence taxonomic assignments for metagenomic data. Nucleic Acids Res., advanced access, April 24, 2012.
   * these authors contributed equally to this work

Determining the taxonomic lineage of DNA sequences is an important step in metagenomic analysis. Short DNA fragments from next-generation sequencing projects and microbes that lack close relatives in reference sequenced genome databases pose significant problems to taxonomic attribution methods. Our new classification algorithm, RITA (Rapid Identification of Taxonomic Assignments), uses the agreement between composition and homology to accurately classify sequences as short as 50 nt in length by assigning them to different classification groups with varying degrees of confidence. RITA is much faster than the hybrid PhymmBL approach when comparable homology search algorithms are used, and achieves slightly better accuracy than PhymmBL on an artificial metagenome. RITA can also incorporate prior knowledge about taxonomic distributions to increase the accuracy of assignments in datasets with varying degrees of taxonomic novelty, and classified sequences with higher precision than the current best rank-flexible classifier. The accuracy on short reads can be increased by exploiting paired-end information, if available, which we demonstrate on a recently published bovine rumen dataset. Finally, we develop a variant of RITA that incorporates accelerated homology search techniques, and generate predictions on a set of human gut metagenomes that were previously assigned to different "enterotypes". RITA is freely available in Web server and standalone versions.

H.4 Non-refereed Manuscripts (Not Discussed in Thesis)

1. **Parks DH**, MacDonald NJ, Beiko RG. 2009. Tracking the evolution and geographic spread of influenza A. PLoS Currents: Influenza, RRN1014.

The 2009 swine-origin strain of Influenza A H1N1 has spread to nearly all parts of the world, with 175 countries reporting confirmed cases thus far. Consistent with seasonal flu outbreaks, the current pandemic strain has shown rapid dispersal, with multiple examples of introduction into different geographic regions. Here we use an automated pipeline to collect data for analysis in the geospatial package GenGIS, which allows the geographic and temporal tracking of new sequence types and polymorphisms. Using this approach, we examine a pair of amino acid changes in the neuraminidase protein that are implicated in antibody recognition, and exhibit global dispersal with little or no geographic structure.

2. MacDonald NJ, **Parks DH**, Beiko RG. 2009. SeqMonitor: Influenza analysis pipeline and visualization. PLoS Currents: Influenza, RRN1040.

Unprecedented sequencing effort has led to daily submissions of influenza genomes to public repositories such as the NCBI GenBank. With the decreasing cost of genome sequencing, it is expected that rapidly evolving viruses such as influenza will be sampled in even greater depth in the future. Keeping analyses up to date and managing this data is a prime concern for researchers and public-health officials alike. We have developed an influenza sequence pipeline, polymorphism data warehouse, and an interactive web-based analysis program to assist in managing the flow of sequence data. The system provides a framework for studying polymorphic associations with various metadata, for downloading subsets based on metadata criteria, as well as for tracking polymorphisms geographically and temporally.

Appendix I

Copyright Permission Letters

This thesis contains 5 manuscripts that are in preparation, have been submitted to or have been published in peer-reviewed international conferences or journals. Permission to reuse these manuscripts has been granted by all co-authors, and permission of reuse forms with co-author signatures are provided for each manuscript. I, and my co-authors, retain the copyright of all manuscripts. Publishers of accepted manuscripts have been granted an exclusive license to publish, and potential publishers of submitted manuscripts will also be granted an exclusive license to publish should the manuscripts be accepted. These licenses do not restrict reuse of manuscripts by the original authors. Details of publisher policies are given in this appendix.

Publication:

Parks DH, Beiko RG. 2009. Quantitative visualizations of hierarchically organized data in a geographic context. 17th International Conference on Geoinformatics (Fairfax, VA): 1-6.

- Conference proceedings were published by the Institute of Electrical and Electronics Engineers (IEEE).

Copyright Policy:

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant. Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

- The following IEEE copyright/credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
- In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

For more information, please consult the IEEE website:
http://www.ieee.org/publications_standards/publications/rights/rightslink_usetypes.html

Publication:

Parks DH, Porter M, Churcher S, Wang S, Blouin C, Whalley J, Brooks S, Beiko RG. 2009. GenGIS: a geospatial information system for genomic data. Genome Res. 19:1896-1904.

- Genome Research is published by Cold Spring Harbor Laboratory Press.

Copyright Policy:

Copyright © 2012, published by Cold Spring Harbor Laboratory Press.

- Authors of articles published in *Genome Research* retain copyright in the articles but grant Cold Spring Harbor Laboratory Press exclusive right to publish the articles. This grant of rights lasts for six months following full-issue publication and includes the rights to publish, reproduce, distribute, display, and store the article in all formats; to translate the article into other languages; to create adaptations, summaries, extracts, or derivations of the article; and to license others to do any or all of the above.
- Authors of articles published in *Genome Research* can reuse their articles in their work as long as *Genome Research* is credited as the place of original publication. They can also archive the article with their institution.
- Beginning six months from the full-issue publication date, or immediately upon publication for articles that carry the journal's Open Access icon, articles published in *Genome Research* are distributed under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described athttp://creativecommons.org/licenses/by-nc/3.0/. This license permits non-commercial use, including reproduction, adaptation, and distribution of the article provided the original author and source are credited.

For further information, please consult the Genome Research website:

http://genome.cshlp.org/site/misc/terms.xhtml

Publication:

Parks DH, Beiko RG. 2012a. Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities.

- In press at ISME J. Accepted July 2012. ISME J is published by the Nature Publishing Group.

Copyright Policy:

If you are the author of this content (or his/her designated agent) please read the following. Since 2003, ownership of copyright in in original research articles remains with the Authors*, and provided that, when reproducing the Contribution or extracts from it, the Authors acknowledge first and reference publication in the Journal, the Authors retain the following non-exclusive rights:

- To reproduce the Contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).
- They and any academic institution where they work at the time may reproduce the Contribution for the purpose of course teaching.
- To reuse figures or tables created by them and contained in the Contribution in other works created by them.
- To post a copy of the Contribution as accepted for publication after peer review (in Word or Tex format) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the Journal article on NPG's web site (eg through the DOI).

* Commissioned material is still subject to copyright transfer conditions

For further information, please consult the Nature Publishing Group website:
http://www.nature.com/authors/policies/license.html

Publication:

Parks DH, Beiko RG. 2012b. Measuring community similarity with phylogenetic networks.

- In press at Molecular Biology and Evolution. Accepted July 2012. Molecular Biology and Evolution is published by the Oxford University Press.

Copyright Policy:

Rights retained by ALL Oxford Journal Authors:

- The right, after publication by Oxford Journals, to use all or part of the Article and abstract, for their own personal use, including their own classroom teaching purposes;
- The right, after publication by Oxford Journals, to use all or part of the Article and abstract, in the preparation of derivative works, extension of the article into book-length or in other works, provided that a full acknowledgement is made to the original publication in the journal;
- The right to include the article in full or in part in a thesis or dissertation, provided that this not published commercially;

For the uses specified here, please note that there is no need for you to apply for written permission from Oxford University Press in advance. Please go ahead with the use ensuring that a full acknowledgment is made to the original source of the material including the journal name, volume, issue, page numbers, year of publication, title of article and to Oxford University Press and/or the learned society.

For further information, please consult the Oxford University Press website:
http://www.oxfordjournals.org/access_purchase/publication_rights.html

Permission of Reuse

I grant Donovan Parks permission to reuse the article "Quantitative visualizations of hierarchically organized data in a geographic context" (17th International Conference of Geoinformatics, Fairfax, VA) in his thesis:

Robert G. Beiko _____

Permission of Reuse

I grant Donovan Parks permission to reuse the article "GenGIS: A geospatial information system for genomic data" (Genome Research, 19, 1896-1904) in his thesis:

Michael Porter _____

Sylvia Churcher _____

Suwen Wang _____

Christian Blouin _____

Jacqueline Whalley _____

Stephen Brooks _____

Robert G. Beiko _____

Permission of Reuse

I grant Donovan Parks permission to reuse the article "Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities" in his thesis:

Robert G. Beiko _____

Permission of Reuse

I grant Donovan Parks permission to reuse the article "Measuring community similarity with phylogenetic networks" in his thesis:

Robert G. Beiko _____

Permission of Reuse

I grant Donovan Parks permission to reuse the article "GenGIS 2: Geospatial analysis of genetic and genomic datasets, with new gradient algorithms and an extensible framework" in his thesis:

Michael Porter _____

Timothy Mankowski _____

Robert G. Beiko _____