GenGIS: A geospatial information system for genomic data

Donovan H. Parks,[1] Michael Porter,[1] Sylvia Churcher,[2] Suwen Wang,[1] Christian Blouin,[1] Jacqueline Whalley,[3] Stephen Brooks,[1] and Robert G. Beiko[1,4]

[1]*Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia B3H 1W5, Canada;* [2]*Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia B3H 1W5, Canada;* [3]*Department of Computing and Mathematical Sciences, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand*

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 data sets requires effective tools for mathematical and visual analysis that can integrate digital maps, ecological data, and large genetic, genomic, or metagenomic data sets. GenGIS is a free and open-source software package that supports the integration of digital map data with genetic sequences and environmental information from multiple sample sites. Essential bioinformatic and statistical tools are integrated into the software, allowing the user a wide range of analysis options for their sequence data. Data visualizations are combined with the cartographic display to yield a clear view of the relationship between geography and genomic diversity, with a particular focus on the hierarchical clustering of sites based on their similarity or phylogenetic proximity. Here we outline the features of GenGIS and demonstrate its application to georeferenced microbial metagenomic, HIV-1, and human mitochondrial DNA data sets.

[Supplemental material is available online at http://www.genome.org. GenGIS, sample data files, and manual are available at http://kiwi.cs.dal.ca/GenGIS.]

Geography and habitat place constraints on the distributions of organisms. While 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% ribosomal DNA (rDNA) 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, rapid whole-genome sequencing, 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; Hughes 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). While 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 nine 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), human phylogenies based on mitochondrial DNA (mtDNA) (Ingman et al. 2000), and trees that track the spread of viruses such as human immunodeficiency

[4]Corresponding author.
E-mail beiko@cs.dal.ca; fax (902) 492-1517.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.095612.109.

19:1896–1904 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

virus-1 (HIV-1) through a host population (Hué et al. 2005). Such analyses, when coupled with geographic analysis tools such as Geophylobuilder (Kidd and Ritchie 2006) and Mesquite Cartographer (Maddison and Maddison 2008), 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.

GenGIS (Beiko et al. 2008b) is a new open source geospatial information system that is dedicated to the display and analysis of georeferenced genetic data. Existing tools such as megx.net (Lombardot et al. 2006) and Micro-Mar (Pushker et al. 2005) integrate marine microbial data with environmental variables and a world map, and the aforementioned geophylogeny tools allow a user to simultaneously visualize a three-dimensional tree and a map. With GenGIS we introduce 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 visual elements and data options implemented directly in GenGIS, users can extend its functionality by developing their own custom scripts or by installing add-on libraries for R or Python that implement population genetic or phylogenetic analyses. We illustrate the flexibility of GenGIS using three case studies: a series of marine metagenomic samples from the Global Ocean Sampling (GOS) expedition (Rusch et al. 2007), *pol* genes from nonrecombinant subtypes of HIV in Africa, and a human geophylogeny based on the mtDNA data sets of Hill et al. (2006, 2007).

Methods

In this section we describe the key features of GenGIS, including required input data types and functionality; complete details of the data sets and methods used in the examples below can be found in the online Supplemental Methods.

Functionality and implementation

GenGIS allows graphical summaries of data on a sample site-by-sample site basis. Location identifiers can be uniform or can be assigned distinct colors, shapes, or sizes based on any of their defined attributes, including latitude, longitude, or habitat parameters such as temperature or salinity. Information about each site can also be displayed on the screen as text, either associated with the location identifier or placed 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 charts, which can be assigned a size that is either constant or proportional to the corresponding sample size. The color scheme and positioning of pie charts can be modified by the user, with a range of predefined color 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 can draw georeferenced trees in two and three dimensions that indicate the ecological or phylogenetic similarity among samples collected from different sites. A key principle in the construction of these 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 (Supplemental Fig. 1A). Mismatches between these two gradients will lead to crossings between the lines that link the two. Fewer crossings imply a better fit between geography and phylogeny, so the best fit of a given tree to a geographic axis must be found, which requires a crossing minimization algorithm. To determine the optimal tree layout, GenGIS uses a branch-and-bound algorithm (Land and Doig 1960) to determine the ordering of leaves and internal nodes of a tree that minimizes the number of crossings (Parks and Beiko 2009). The idea of a linear geographic axis can be generalized to a multisegment line of arbitrary complexity, allowing the specification of piecewise, nonlinear geographic hypotheses (Supplemental Fig. 1B). Coupled with the axis layout functions is a statistical test, based on randomization of leaf nodes, 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 colored in accordance with the coloring of different environmental types: A given branch will be assigned a consistent color if all children of a given branch are associated with the same environment type, or a default color if its children cover multiple environments.

The core GenGIS software is implemented using C++ and OpenGL, which supports the rendering of cartographic data in three dimensions. As a free and open-source application, GenGIS makes extensive use of other open-source software libraries, including Python, which provides GenGIS with a powerful scripting language that can be used to manipulate and analyze sequence and environmental data; RPy, which contains extensive support for performing statistical tests and analyses; and GDAL, which allows map data in nearly any format to be imported and manipulated (see below). Other included libraries are described in the GenGIS documentation.

Data acquisition and formats

There are several large public repositories of digital map data, including the Shuttle Radar Topography Mission (SRTM) (Farr et al. 2007) and GTOPO30 data sets hosted by the U.S. Geological Survey. Digital maps can be provided in any of a large number of file formats, and GenGIS uses the freely available GDAL libraries (http://www.gdal.org) to support a wide range of these formats. A GeoTIFF-formatted world map derived from the GTOPO30 set is included with the GenGIS release package, but GDAL can also be used as a preprocessing utility to directly manipulate maps from sources such as the SRTM, allowing a user to construct a more-detailed map that covers a smaller geographic area (see Supplemental Methods). Maps in GenGIS can be displayed using a number of different projections and source datums. While most file formats are interpreted as topographic data, three-channel input files are interpreted as colors with no assumption of elevation information, allowing the direct display of environmental features such as vegetation cover or salinity from an appropriate input file.

GenGIS also requires as input a comma-separated file containing data about geographic locations for a given data set. Each of these locations 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 (UTM) northing and easting values. Location coordinates need not

be unique, since a given site may have multiple samples associated with it (e.g., 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 and 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 identifier used to specify either the sample sites or sequences. Alternatively, a geographic location block can be specified in order to explicitly map leaf nodes to sample sites.

Results

Taxonomic diversity from the Global Ocean Sampling expedition

The Global Ocean Sampling expedition is using environmental shotgun sequencing to collect metagenome data from marine sample sites spread around the world. The initial publication (Rusch et al. 2007) analyzed 44 metagenome samples (0.1–0.8 μm fraction) collected from 41 sites, including Sargasso Sea sites examined previously by Venter et al. (2004). Data from these locations have been analyzed to reveal 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 rDNA-containing reads. Here we considered a set of 19 locations (sites GS002–GS020 from the original article) covering the Atlantic seaboard of North America, comprising all sites between Nova Scotia and the Panama Canal, including three 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 ~9°N and 45°N, allows the hypothesis proposed by Fuhrman et al. (2008) to be examined. The investigators of this study proposed that latitude is the primary determinant of species richness, which suggests 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 rDNA sequences from each sample using BLAST comparisons against the GreenGenes database (DeSantis Jr et al. 2006) and using the best full-length matches as a proxy for the fragmentary 16S sequences found in the GOS reads. To examine the possible relationship between normalized taxon richness (as indicated by the number of unique sequences or distinct operational taxonomic units (OTUs); see Supplemental Methods) and latitude, we visualized richness values in GenGIS and performed linear regression analysis. Supplemental Figure 2 shows a set of georeferenced bars indicating the normalized unique sequence count at all 19 locations, as well as a restricted subset covering only the 14 oceanic sites. When all 19 locations were included in the regression model, the relationship between taxon richness and latitude was significant ($0.003 \leq P \leq 0.05$) at all four levels of clustering (unique sequences, and OTU clustering at 97%, 95%, and 90% identity thresholds). The freshwater and estuarine sites did not produce the largest residuals, and deletion of five sites of unusual composition (5, 6, 11, 12, and 20, as identified above) from the analyses yielded models with worse fit and only marginal significance ($0.03 \leq P \leq 0.21$).

We used the unweighted and weighted UniFrac phylogenetic diversity measures (Lozupone et al. 2007), which compute phylogenetically weighted measures of species richness and evenness, to estimate the similarity between pairs of sites in this data set. A maximum-likelihood phylogenetic tree covering the proxy 16S sequences found at all sites was constructed and used as input to UniFrac. Figure 1 shows the clustering of these sites based on their phylogenetic similarity as determined using weighted UniFrac. The geographic axis in this figure, depicted as a pair of parallel lines, corresponds to the main axis along which sequence data sets 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 labels of the tree yielded four out of 1000 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 larger *P*-value (35 crossings, $P = 0.031$). However, these patterns may conflate geographic and habitat effects, and 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 data set, a reduced data set of 14 sites as above, and a further partitioning of the 14 sites into Atlantic seaboard (nine sites) and Caribbean Sea (five 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 (weighted UniFrac: 29 crossings, $P = 0.019$; unweighted UniFrac: 36 crossings, $P = 0.028$). While deletion of the "unusual" habitat types from the set (Supplemental Fig. 3) 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 (weighted UniFrac: six crossings, $P = 0.001$; unweighted UniFrac: eight crossings, $P = 0.003$). A further partitioning of sites into sets of nine and five as indicated above yielded results that were not statistically significant ($0.109 \leq P \leq 0.466$ for all combinations of weighted and unweighted UniFrac, and Atlantic and Caribbean 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 two regions.

Although theoretical results to suggest appropriate thresholds are lacking, the weighted and unweighted UniFrac trees display a wide range of jackknife support values (Supplemental Fig. 4). We complemented the analysis of jackknifed trees with pie chart

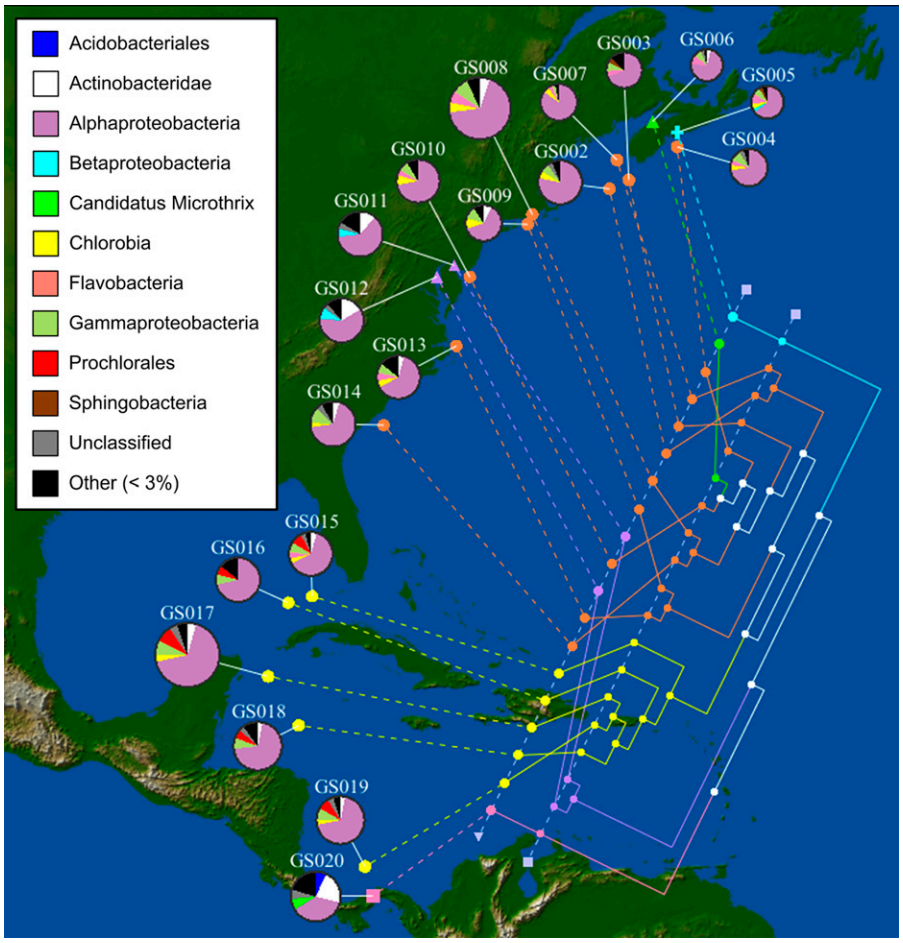

Figure 1. Clustering of Global Ocean Sampling (GOS) sites based on their shared phylogenetic diversity. Pie charts associated with each GOS site show the breakdown of 16S sequences by best-matching bacterial class, with rare groups collected together in the "other" category. Pie chart sizes are proportional to the total number of 16S sequences considered at each site. The clustering of sites obtained by weighted UniFrac is shown in the tree, with habitat type indicated by shape and color (cyan cross indicates embayment; circles, oceanic, [with orange, Atlantic Ocean; with yellow, Caribbean Sea]; purple triangle, estuary with low salinity; green triangle, estuary with typical oceanic salinity; pink square, freshwater lake). White branches in the tree indicate internal edges whose children cover multiple habitat types.

visualizations of the most highly variable taxonomic classes across sites (Supplemental Fig. 5). The grouping of three low-salinity sites, and Lake Gatun versus Delaware and Chesapeake Bays were supported with jackknife values ≥90, suggesting strong differentiation in both richness and relative abundance. 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 are overrepresented in Actinobacteridae (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, suggesting that differences in richness and relative abundance, while apparently 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 likely distinct from all other sites, which is potentially due to a relatively high proportion of betaproteobacterial sequences, a complete lack of Actinobacteridae, and possibly different relative proportions of certain ubiquitous taxonomic groups. Conversely, the Bay of Fundy estuary, with salinity levels that are similar to open ocean sites, was indistinguishable from other Atlantic Ocean sites in both analyses, although its closest neighbor was different in the weighted (GS010) and unweighted (GS007, with jackknife support of 76%) UniFrac analyses.

Nonrecombinant 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 nonrecombinant subtypes A–D, F–H, J, and K, although we note the controversy surrounding the nonrecombinant nature of some of these subtypes (Abecasis et al. 2007). Only countries with at least 10 samples in this data set were retained, yielding a total of 40 countries with sequence counts between 12 (Guinea-Conakry) and 6576 (South Africa). Pie chart summaries of

subtypes by country are shown in Supplemental Figure 6; subtype counts by country, in Supplemental Table 3.

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 weighted UniFrac analysis: Since many subtypes are represented in many countries, an unweighted UniFrac analysis that ignores the relative abundance of different sequences would not discriminate well between locations. The initial set of sequences extracted from the HIV Sequence Database was reduced by identifying seed sequences and eliminating any other sequence in the set whose sequence identity to that seed was greater than 92%. This produced a set of 18 *pol* sequences, to which one sequence of group N and three of group O were added, to allow rooting of the group M subtree. The subtype reference tree was inferred using MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003). 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.

Figure 2 shows the clustering derived from a weighted UniFrac analysis based on the resulting tree of *pol* sequences. Three-dimensional trees such as this can be difficult to interpret in a static two-dimensional image, but we have colored four major groupings of countries that show a certain degree of geographic separation and appear to be largely driven by common subtypes seen in Supplemental Figure 6. Eastern and southern Africa are dominated by subtype C and constitute a cluster (colored purple in Fig. 2), 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 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, Congo-Kinshasa, Angola, Senegal, and Burkina Faso. A handful of west African countries are dominated by subtype G; two of these, Niger and Nigeria, constitute an early branch in the large cluster that also covers the rest of west Africa. The other early branch in this large cluster maps to the Gambia, which contains a rich 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 high proportions of A and G tend to be dominated by the circulating recombinant form AG(02). Supplemental Video 1 is an animation of the cluster tree, which was generated using five lines of Python code in GenGIS.

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

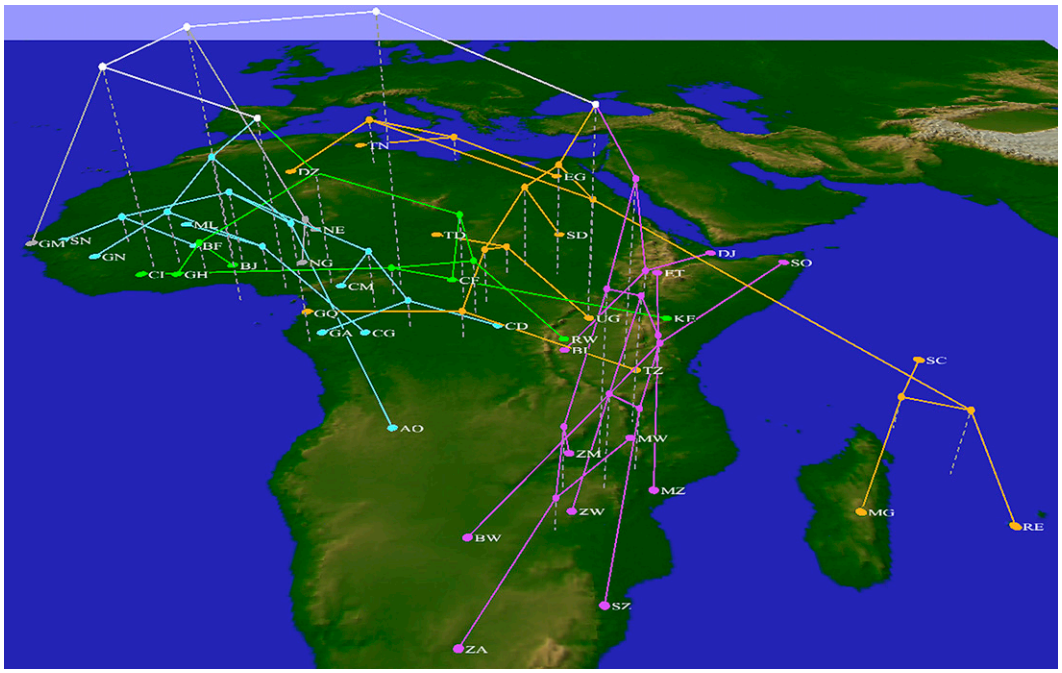

Figure 2. Clustering of African nations based on phylogenetic diversity of HIV subtypes. The UPGMA clustering of countries based on their UniFrac scores is shown using a three-dimensional tree; the four subclusters discussed in the text and the countries they cover are indicated by coloring different 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.

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 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.

Human mitochondrial (mtDNA) geophylogeny

mtDNA evolves rapidly and has been used to reconstruct matrilineal patterns of human migration (e.g., Ingman et al. 2000). Many of the characteristic polymorphisms that have been used to define mtDNA haplogroups are found in the highly variable region that spans the origin of replication, and this region has served as the focus for many large-scale mtDNA studies (e.g., Helgason et al. 2001; Malhi et al. 2002). Although mtDNA mutations can be highly informative in phylogenetic and population genetic studies, most inhabited areas contain a diverse mixture of distinct haplogroups and subgroups. Understanding the anthropological history of a region requires the characterization of the different groups present, and the inference of population-level parameters to identify locations with high diversity and likely sources of migration at different times. For example, Island Southeast Asia (ISEA), which consists of the Indonesian and Philippine archipelagos, appears to have served as a source and a destination of migratory populations many times since its initial settlement ~50,000–60,000 yr ago (Macaulay et al. 2005; Barker et al. 2007). Recent studies of the region have identified at least 14 major haplogroups (Hill et al. 2007), including groups (E) and subgroups (e.g., M7c1c) that are restricted to ISEA and neighboring populations such as indigenous Taiwanese.

We used GenGIS to examine the phylogenetic distribution of mtDNA haplogroup E, using hypervariable segment I (HVS-I) sequences from ISEA. These sequences were part of a larger study that considered 929 sequences from ISEA and aboriginal Taiwanese individuals (Hill et al. 2007). A phylogenetic analysis of the entire set would not be informative, because many of the observed haplogroups have migrated in and out of the region in the last 50,000 yr and because lineages that appear in one location will subsequently be spread to others. Haplogroup E is largely restricted to ISEA and appears to have arisen ~25,000–35,000 yr ago (Hill et al. 2007; Soares et al. 2008). The subsequent divergence of this haplogroup into subclades E1a, E1b, E2a, and E2b covers a period from ~17,000 to ~4500 yr ago and may track important migrations that followed the last glacial maximum of 19,000 yr ago (Soares et al. 2008). Of particular importance is the potential influence of Wallace's Line (Mayr 1944), which splits Indonesia into a region (Sundaland) that was contiguous with the Asian mainland during the last glaciations, and a set of islands (collectively "Wallacea," which includes Sulawesi and Lombok) that were not. The native fauna differ significantly on either side of this line in spite of its narrowness (i.e., only 15 km between Bali and Lombok).

In this analysis, we focused on the geographic history of haplogroup E1. The tree of E1 sequences was rooted by including five sequences from the sister haplogroup E2 in the parsimony analysis. Given the rooting induced by the E2 sequences, the earliest split in the E1 tree separates a single individual in Palu, Sulawesi, from the rest of the E2 sequences; this is consistent (albeit weakly) with the hypothesis of Soares et al. (2008), that northeastern Sunda (i.e., eastern Borneo) or western Sulawesi served as the originating point for subclades of haplogroup E. All of the sequences from northern Borneo, the Philippines, and Taiwan constitute a clade in the tree of E1 sequences, which is also consistent with the spread of ancestral E1 sequences north from Sunda (roughly, western Indonesia), but there is considerable intermingling within the northern and southern groups (Supplemental Fig. 7). A linear geographic axis that is roughly parallel to Wallace's Line yields 90 crossings, which is nonetheless statistically significant ($P = 0.004$). This leaf ordering necessarily ignores any migration patterns that are not parallel to the defined linear axis, and comingles locations on the southeast and southwest migration routes. Using a strict north–south axis causes many locations in the southwest and east routes to overlap instead, and increases the number of crossings to 116. Choosing an east–west axis to optimize leaf ordering separates eastern from western migrations but fails to capture the strongly supported northern clade and yields 176 crossings. These examples illustrate the limitations of imposing a linear ordering on sequences that are unlikely to follow a linear migration pattern.

To approximate the hypothesis of Soares et al. (2008), we created a set of four geographic polylines that radiate out from Borneo in different directions: one toward Sulawesi, one heading north to Taiwan, and two spreading southeast or southwest away from Wallace's Line. The dashed lines in Figure 3 indicate the manually drawn polylines that correspond to these four migration paths: in each case, we used Borneo as a starting point for the polyline, and extended the line to locations in the appropriate direction that were progressively more distant from the starting location. For example, along the northern route, the polyline originates in central Borneo and passes through Kota Kinabalu in northern Borneo, the Philippines, and the three Taiwanese locations in a south-to-north order. The optimal layout produced by aligning the leaves of the E1 tree to this set of axes induces only 78 crossings ($P = 0.001$), suggesting that the proposed radial spread from eastern Sunda may indeed be a better explanation of observed E1 haplogroup sequence distribution in the region. Since the four-polyline model is more complex (i.e., has greater degrees of freedom), it is not necessarily surprising that this model induces fewer crossings than the linear axes described above, but the permutation test still suggests that the model fit is statistically significant. Therefore, the fact that the *P*-value under the more-complex model is less than that obtained from the linear axis (0.001 vs. 0.004) supports the use of the geographic polylines in this case. Supplemental Video 2 shows a complete session demonstrating the loading of source data into GenGIS, interactive configuration of the visualization options, and geographic analysis of the phylogenetic relationships among haplogroup E1 sequences.

Discussion

By coupling digital map data with georeferenced sequence information, GenGIS has allowed us to visualize patterns of microbial species and viral subtype distribution and to explore migratory patterns via marker gene surrogates. 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

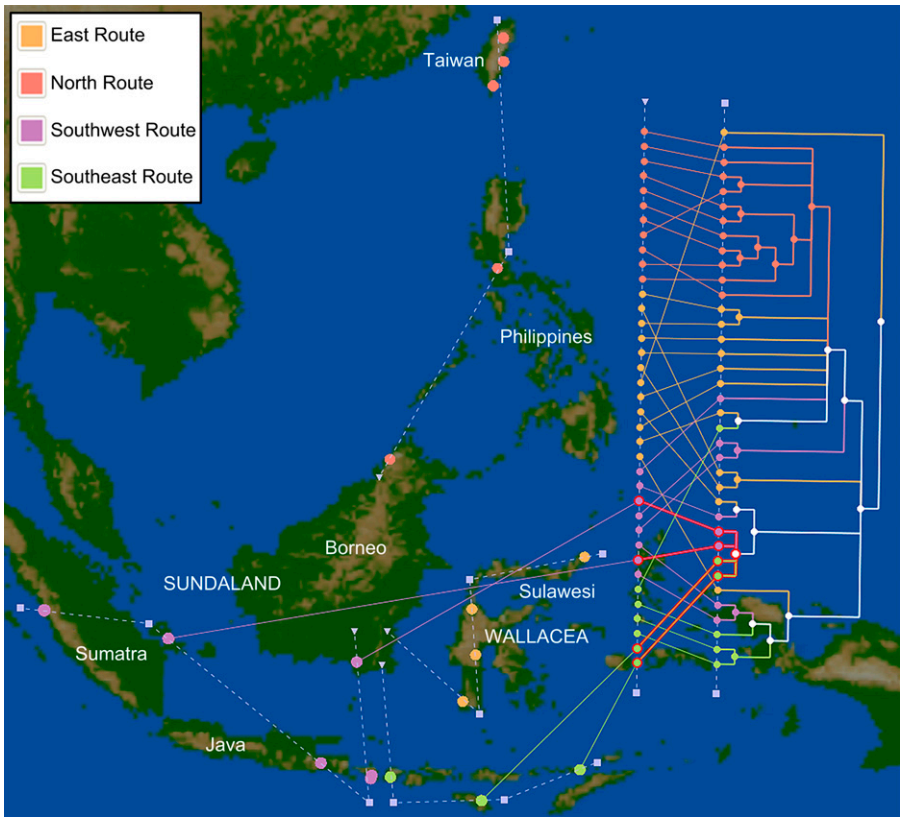

Figure 3. Map of Island Southeast Asia and Taiwan with hypothesized migration patterns of mtDNA haplogroup E1 encoded as geographic polylines. A triangle indicates the point of origin for each polyline on the island of Borneo. Most droplines connecting the geographic axis line to sample sites are omitted for clarity, but sites are colored according to their location relative to the proposed origin of haplogroup E1 and consequent migration route (Soares et al. 2008): Sulawesi (east) in orange, south Sundaland (south/southwest) in purple, south Wallacea (south/southeast) in green, and the northern points in pink. The optimal tree layout with 78 crossings is shown, with colored branches indicating leaves or groups of leaves that were exclusive to a single region once dereplication had been carried out (see Supplemental Methods). One southern subclade of E1 that contains both southeastern and southwestern sequences is highlighted, showing droplines that connect to its corresponding geographic locations.

implemented and carried out. The above analyses demonstrate the different interpretations that can be attached to hierarchical clusters of data. In the GOS example, latitudinal gradients of taxonomic richness were only supported when all habitat types were included, with *P*-values marginal or >0.05 when only oceanic sites were considered. Hierarchical analysis of shared 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 picocyanobacteria as the principal factor influencing this separation. Similarity in relative abundance (as assessed using weighted UniFrac) yielded a stronger clustering signal than similarity in richness (as assessed using unweighted UniFrac). 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. While 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. Finally, our phylogenetic analysis of mitochondrial haplogroup E1 produced visual and statistical support for the proposed migration pattern of this genotype. Our novel, sensitive approach highlighted regions where geographic structure was strong.

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 et al. 2008a), 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 data sets 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 tree-based approaches to diversity currently implemented in GenGIS allow the identification of optimal matches between a clustering hierarchy and geography. This parallels earlier work in biogeography and cospeciation (Page 1994; Ronquist 1998; Charleston 1998), and GenGIS extends the notion of fixed geographic intervals by allowing user-defined polylines that correspond to proposed directions of migration, speciation, or other processes. While *P*-values reflect the probability of the observed number of crossings occurring by chance in a randomized tree, we have not yet explored the impact of increasing the degrees of freedom (i.e., by introducing multiple polylines as in Fig. 3) on the expected decrease in the number of crossings. While the calculated *P*-value expresses the significance of a given choice of geographic axis or polyline and thus should be comparable across different types of axis, the number of different linear axes or geographic polylines that can be specified is very large, and it may be difficult for a researcher to choose one from a set of subtly different candidate axes. Further refinements to our approach will include options to encode a hypothesis in broad terms (e.g., thick arrows pointing in different directions from a shared ancestral location) and have the program automatically enumerate all candidate orderings so implied, in the end showing summaries of the goodness-of-fit of these distinct orderings.

The number and size of genetic data sets that are available from public repositories is growing, and all of the data used in this

study were acquired from such resources (see Supplemental Methods). However, our vision for GenGIS includes not only the analysis of static data sets prepared in advance by a user but also direct integration with emerging online repositories including NCBI, the Barcode Online Database (Ratnasingham and Hebert 2007), RDP (Cole et al. 2009), and the HIV sequence database. Querying online data sets 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, or recombinants, or environmental organisms. Automated Web or FTP interfaces to some of these sites already exist: For instance, BioPerl modules exist for automated access to the HIV sequence database, and RDP offers a series of Web services based on the SOAP standard. In such cases, automated access and acquisition of data by GenGIS could be achieved by executing scripts in the Python console. 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.

Acknowledgments

We thank Harman Clair, Norman MacDonald, and Gregory Smolyn for assistance with the development of GenGIS. The development of GenGIS has been supported by Genome Atlantic, the Natural Sciences and Engineering Research Council of Canada, the Canada Foundation for Innovation, and the Dalhousie Faculty of Computer Science. R.G.B. is supported by the Canada Research Chairs program.

References

Abecasis AB, Lemey P, Vidal N, de Oliveira T, Peeters M, Camacho R, Shapiro B, Rambaut A, Vandamme AM. 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.

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

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.

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

Barker G, Barton H, Bird M, Daly P, Datan I, Dykes A, Farr L, Gilbertson D, Harrisson B, Hunt C. 2007. The "human revolution" in lowland tropical Southeast Asia: The antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). *J Hum Evol* **52:** 243–261.

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

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

Beiko RG, Whalley J, Wang S, Clair H, Smolyn G, Churcher S, Porter M, Blouin C, Brooks S. 2008b. Spatial analysis and visualization of genetic biodiversity. FOSS4G 2008, Cape Town, South Africa. http://conference.osgeo.org/index.php/foss4g/2008/paper/view/105.

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.

Böer SI, Hedtkamp SI, van Beusekom JE, 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.

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

Bryant JA, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ, Green JL. 2008. Colloquium paper: Microbes on mountainsides: Contrasting elevational patterns of bacterial and plant diversity. *Proc Natl Acad Sci* **105:** 11505–11511.

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.

Charleston MA. 1998. Jungles: A new solution to the host/parasite phylogeny reconciliation problem. *Math Biosci* **149:** 191–223.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, et al. 2009. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37:** D141–D145.

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

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, et al. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496–503.

DeSantis TZ Jr, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72:** 5069–5072.

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, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, et al. 2008. Functional metagenomic profiling of nine biomes. *Nature* **452:** 629–632.

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

Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, Kobrick M, Paller M, Rodriguez E, Roth L, et al. 2007. The shuttle radar topography mission. *Rev Geophys* **45:** RG2004. doi: 10.1029/2005RG000183.

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

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* **105:** 7774–7778.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, et al. 2005. Re-evaluating prokaryotic species. *Nat Rev Microbiol* **3:** 733–739.

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* **104:** 18566–18570.

Green Tringe S, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, et al. 2005. Comparative metagenomics of microbial communities. *Science* **308:** 554–557.

Helgason A, Hickey E, Goodacre S, Bosnes V, Stefánsson K, Ward R, Sykes B. 2001. mtDna and the islands of the North Atlantic: Estimating the proportions of Norse and Gaelic ancestry. *Am J Hum Genet* **68:** 723–737.

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–W23.

Hill C, Soares P, Mormina M, Macaulay V, Meehan W, Blackburn J, Clarke D, Raja JM, Ismail P, Bulbeck D, et al. 2006. Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol Biol Evol* **23:** 2480–2491.

Hill C, Soares P, Mormina M, Macaulay V, Clarke D, Blumbach PB, Vizuete-Forster M, Forster P, Bulbeck D, Oppenheimer S, et al. 2007. A mitochondrial stratigraphy for island southeast Asia. *Am J Hum Genet* **80:** 29–43.

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* **102:** 4425–4429.

Hughes Martiny JB, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, et al. 2006. Microbial biogeography: Putting microorganisms on the map. *Nat Rev Microbiol* **4:** 102–112.

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

Janies D, 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.

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.

Kidd DM, Ritchie MG. 2006. Phylogeographic information systems; putting the geography into phylogeography. *J. Biogeography* **33:** 1851–1865.
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.
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.
Land AH, Doig AG. 1960. An automatic method for solving discrete programming problems. *Econometrica* **28:** 497–520.
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–D393.
Lozupone C, Knight R. 2005. UniFrac: A new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71:** 8228–8235.
Lozupone C, Knight R. 2007. Global patterns in bacterial diversity. *Proc Natl Acad Sci* **104:** 11436–11440.
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.
Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308:** 1034–1036.
Maddison DR, Maddison WP. 2008. Cartographer, a Mesquite package for plotting geographic data, version 1.3. http://mesquiteproject.org/packages/cartographer.
Malhi RS, Eshleman JA, Greenberg JA, Weiss DA, Schultz Shook BA, Kaestle FA, Lorenz JG, Kemp BM, Johnson JR, Smith DG. 2002. The structure of diversity within New World mitochondrial DNA haplogroups: Implications for the prehistory of North America. *Am J Hum Genet* **70:** 905–919.
Margos G, Gatewood AG, Aanensen DM, Hanincová K, Terekhova D, Vollmer SA, Cornet M, Piesman J, Donaghy M, Bormane A, et al. 2008. MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. *Proc Natl Acad Sci* **105:** 8730–8735.
Martin AP. 2002. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* **68:** 3673–3682.
Mayr E. 1944. Wallace's line in the light of recent zoogeographic studies. *Q Rev Biol* **19:** 1–14.
Moritz C, Schneider CJ, Wake DB. 1992. Evolutionary relationships within the *Ensatina eschscholtzii* complex confirm the ring species interpretation. *Syst Biol* **41:** 273–291.
Page RDM. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst Biol* **43:** 58–77.
Parks DH, Beiko RG. 2009. *Quantitative visualizations of hierarchically organized data with a geographic context*. Geoinformatics 2009, Fairfax, VA.
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.
Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol* **54:** 396–402.
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. doi: 10.1186/1471-2105-6-222.
Ratnasingham S, Hebert PDN. 2007. BOLD: The Barcode of Life Data System. *Mol Ecol Notes* **7:** 355–364.
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. 1998. Phylogenetic approaches in coevolution and biogeography. *Zool Scr* **26:** 313–322.
Ronquist F, Huelsenbeck J. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572–1574.
Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K. 2007. The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5:** e77. doi: 10.1371/journal.pbio.0050077.
Soares S. 2007. Geographical biases and convenience sampling in HIV molecular epidemiology estimates. *AIDS* **21:** 2359–2360.
Soares P, Trejaut JA, Loo JH, Hill C, Mormina M, Lee CL, Chen YM, Hudjashov G, Forster P, Macaulay V, et al. 2008. Climate change and postglacial human dispersals in southeast Asia. *Mol Biol Evol* **25:** 1209–1218.
Schloss PD. 2008. Evaluating different approaches that test whether microbial communities have the same structure. *ISME J* **2:** 265–275.
Schloss PD, Handelsman J. 2006. Introducing TreeClimber, a test to compare microbial community structures. *Appl Environ Microbiol* **72:** 2379–2384.
Serwadda D, Mugerwa RD, Sewankambo NK, Lwegaba A, Carswell JW, Kirya GB, Bayley AC, Downing RG, Tedder RS, Clayden SA. 1985. Slim disease: A new disease in Uganda and its association with HTLV-III infection. *Lancet* **19:** 849–852.
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.
Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC, Hughes Martiny JB. 2008. It's all relative: Ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10:** 2200–2210.
Simmons SL, Dibartolo G, Denef VJ, Goltsman DS, 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. doi: 0.1371/journal.pbio.0060177.
Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM. 2008. The challenge of HIV-1 subtype diversity. *N Engl J Med* **358:** 1590–1602.
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.
Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74.
Wiens JJ. 1998. Combining data sets with different phylogenetic histories. *Syst Biol* **47:** 568–581.
Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, et al. 2007. The *Sorcerer II* Global Ocean Sampling expedition: Expanding the universe of protein families. *PLoS Biol* **5:** e16. doi: 10.1371/journal.pbio.0050016.
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.
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. doi: 10.1371/journal.pgen.1000095.
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.

Received May 3, 2009; accepted in revised form July 16, 2009.