Gap Analysis and the Geographical Distribution of Parasites

Gap analysis and the geographical distribution of parasites

Mariah E. Hopkins[1] and Charles L. Nunn[2]

[1] Dept. of Anthropology, University of Texas at Austin, 1 University Station C3200, Austin, TX, 78712, contact: hopkins@mail.utexas.edu.

[2] Department of Anthropology, Harvard University, Peabody Museum, 11 Divinity Ave., Cambridge MA 02138, contact: cnunn@oeb.harvard.edu

WORD COUNT: 5192

SUMMARY

Sampling biases can have enormous impacts on studies of parasite biogeography. While complete sampling is sometimes possible for local or regional patterns of parasitism, continental and global analyses often rely on data collected in a heterogeneous manner. At these larger scales, spatially-explicit methods to quantify and correct for geographic sampling biases are necessary. Approaches based on "gap analysis" can contribute to the development of corrective measures by identifying geographical variation in our knowledge of parasites and quantifying how sampling varies in relation to host characteristics and habitat features. In this chapter, we review these methods and describe how they have been applied to study gaps in our knowledge of primate parasites.

INTRODUCTION

Studies of host-parasite biogeography, especially those looking beyond local analyses to regional or global scales, are influenced greatly by geographically inconsistent sampling patterns. A variety of factors result in heterogeneous sampling across space. Global parasite sampling is often limited by logistical factors that include a lack of suitable roads or airports, risks arising from unstable political climates, and difficulty in acquiring and preserving samples in remote locations. In addition, it is often easier to obtain funding to study parasites that have large economic impacts, including the potential for transmission to humans (zoonoses). Last, any of these factors can change through time, producing temporal variation that can further complicate studies of parasite biogeography.

In this chapter, we focus on quantifying these biases, regardless of their underlying causes. Specifically, we demonstrate how global gap analysis—a method used in conservation biology to identify conservation targets—can be applied to identify and quantify bias in geographic sampling for parasites. We review and illustrate the principles of gap analysis by describing its recent application to identify geographic gaps in our knowledge of primate parasites (Hopkins & Nunn, 2007) . We also provide new analyses to demonstrate how optimality techniques can be applied within a gap analysis framework to target sampling sites. The results from these analyses are important for those seeking to develop spatially-explicit corrections for sampling bias in regional and global analyses of host-parasite biogeography. Such analyses also serve as a guide to future sampling efforts, by targeting areas where additional sampling would reduce geographic bias in large-scale datasets.

Goals of spatial analyses of parasite sampling effort

Understanding the evolutionary diversification of parasites and their hosts requires a better understanding of the geographic distribution of parasites in relation to host characteristics or ecological factors that vary at continental or global scales (Gregory, 1990; Poulin, 1997; Poulin & Morand, 2000). At the most basic level, parasite distributions are inextricably linked to the distributions of their hosts (Guegan & Kennedy, 1996; Poulin, 1997; Tripet *et al.*, 2002; Lindenfors *et al.*, 2007). Thus, any analysis of geographic sampling patterns for parasites must first take into account host ranges. Detailed analyses of parasite sampling may also choose to take into account whether parasites have been sampled across a diversity of host characteristics such as body mass, population density, home range size, and diet, as these factors have been suggested to impact parasite species richness (Poulin & Morand, 2000; Nunn *et al.*, 2003; Araujo & Guisan, 2006). Furthermore, ensuring that parasite sampling incorporates environmental gradients such as distance from the equator, temperature, precipitation, and habitat variability may also be critical to understanding host-parasite biogeography (Nunn *et al.*, 2005; Nunn & Altizer, 2006; Lindenfors *et al.*, 2007; Krasnov *et al.*, 2008).

When evaluating host-parasite biogeography in relation to host or environmental factors that vary at continental scales, it is essential to ensure that sampling effort has been distributed evenly or in proportion to the expected abundance of parasites, and to take remedial action if sampling biases are present. In other words, when interpreting observed trends in parasite-host biogeography, one has to be careful that the patterns generated do not simply reflect taxonomic or geographic biases in research effort (Poulin & Morand, 2000; Burke, 2007). The likelihood of such biases is liable to increase with

the scale of the analysis. At a global scale, biases are especially likely because it is usually infeasible to achieve complete sampling; some regions, taxonomic groups, or subsets of parasites with different transmission modes are likely to be sampled better than others.

A number of methods to correct for uneven sampling effort have been developed, particularly in the context of studying parasite species richness (Walther *et al.*, 1995; Poulin, 1998; Walther & Morand, 1998b; Walther & Martin, 2001; Cam *et al.*, 2002; Robertson & Barker, 2006; Lobo, 2008). These methods range in complexity from simply including the number of sampled sites or individuals as independent factors in regression analyses, to more complex adjustments such as developing accumulation curve models and non-parametric estimators of species richness. Research has shown that certain performance estimates do better than others at low sampling effort (Walther & Morand, 1998a). However, few studies have incorporated *spatially-explicit* examinations of geographic variation in sampling effort when correcting for sampling bias.

In order to be most effective, measures of geographic bias must move beyond simply counting the number of sites or ecosystems in which a parasite species has been sampled to spatially-explicit analyses that take into account both the characteristics of each sampling site as well as its actual location. These spatially-explicit explorations of geographic bias can provide critical information that is currently unaccounted for in global studies of parasites, and is the essential first step in the development of effective corrections for geographic sampling effort in studies at these larger regional or global scales. For example, using a measure such as the number of publications on a particular

host or parasite species, or the number of sampling localities at which a species has been sampled, leaves out critical information such as the following. Were the sites at which sampling occurred extremely close together or far apart? Which hosts were present at that locality, and how many of these hosts were actually sampled? Were the sites at which sampling occurred representative of the range of environmental conditions in which the host/parasite occurs? What percentage of potential microhabitats in which the host occurs have been sampled?

Developing an approach to correct for geographic bias that can incorporate answers to these questions requires that geographic bias first be quantified within a spatially-explicit framework that can compare the distribution of sampling localities to the distribution of factors thought to influence host-parasite biogeography (e.g. host ranges, environmental characteristics). Gap analysis provides such a framework (Jennings, 2000; Funk & Richardson, 2002; Funk *et al.*, 2005).

Gap Analysis

Gap analysis provides a conceptual, technical, and organizational basis for identifying and quantifying gaps between two or more spatial distributions (Jennings, 2000). In evaluating gaps between distributions, gap analysis may incorporate traditional measures of spatial data analysis such as display mapping and spatial statistics. However, these traditional measures form only a part of the gap analysis framework. Gap analysis encompasses an entire process ranging from the establishment of an 'optimal' distribution to the comparison of the observed and 'optimal' distributions and the subsequent targeting of actions to remedy gaps between these distributions.

Early gap analyses were developed to solve location-allocation (i.e. distance-based) problems in a variety of fields including operations research and transportation engineering (Tansel *et al.*, 1983; Brandeau & Chui, 1989). For example, urban planners use gap analysis to determine where cities should build fire stations, with the goal to minimize the distance between all homes and the nearest fire station. Recent extensions of gap analysis have moved beyond purely distance-based location-allocation problems to incorporate attribute data as well. For example, conservation biologists use both locational information as well as geographic distributions of attribute values such as habitat type, species richness, and projected levels of environmental change in order to distribute protected areas in such a way that the highest amount of biodiversity is conserved (Scott *et al.*, 1987; Ferrier, 2002; Rodrigues *et al.*, 2004; Sarkar *et al.*, 2006).

Gap analysis has more recently been applied to guide spatial patterns in biological sampling efforts, including sampling for disease (Funk & Richardson, 2002; Funk *et al.*, 2005; Hortal & Lobo, 2005; Hopkins & Nunn, 2007). These studies typically use both location and attribute data to determine whether sampling has been distributed along critical host or environmental gradients. They also use optimization techniques to identify future sampling sites that have the highest probability of yielding additional biodiversity (Funk *et al.*, 2005; Hortal & Lobo, 2005).

GAP ANALYSIS METHODS: A CASE-STUDY EXAMING GEOGRAPHIC PATTERNS OF SAMPLING FOR PRIMATE PARASITES

In the following sections, we illustrate the most common steps of a gap analysis, providing illustrative examples based on our recent study of global primate parasite sampling (Hopkins & Nunn, 2007).

Isolation of an 'Optimal Distribution'

Gap analyses of parasite sampling effort rest on the assumption that while uniform data sampling may be suitable for studies of parasites at local or regional scales, uniform sampling is neither feasible nor the most representative sampling technique at continental or global scales. Instead, proportional sampling techniques may better inform studies of global biogeography (Schoereder *et al.*, 2004; Hortal & Lobo, 2005; Kery *et al.*, 2008). For example, since parasite distributions inevitably follow host distributions, one could argue that if the goal is to evenly sample parasite diversity, parasite sampling intensity should be allocated geographically according to host diversity. Alternatively, since ecological characteristics (such as the diversity of habitat types) have been hypothesized to increase parasite richness in a host species, it could be argued that comparatively more parasite sampling effort should be devoted to hosts in areas with higher host abundance and/or greater ecological diversity. We conducted a gap analysis based on the assumption that global patterns of parasite sampling should be allocated proportionally to host diversity (i.e. species richness) (Hopkins & Nunn, 2007). In the following sections, we provide examples from this research to illustrate typical gap analysis methods.

Data Acquisition

Once the theoretical optimal distribution has been selected, the next step in any global gap analysis becomes data acquisition. While this step may seem self-evident, we include discussion of it here because in studies of the global biogeography of host-parasite interactions, the acquisition of spatially-explicit sampling data may perhaps be the greatest challenge. Global GIS clearinghouses increasingly distribute information on environmental characteristics (e.g. Earth Resources Observation Systems Data Center, The Geography Network, National Geospatial Intelligence Agency Products and Services, Tropical Rain Forest Information Center). However, data on parasite sampling is often derived from literature searches where authors frequently fail to geo-reference their sampling sites. For example, we obtained data for our analyses on primate parasite sampling from the Global Mammal Parasite Database (Nunn & Altizer, 2005). Although this represents the most comprehensive database of parasites in wild primates, sufficient spatial information (coordinates or a unique locality name) to geo-reference parasite sampling sites was lacking in approximately one half of all primate parasite studies. This finding is consistent with previous studies of biodiversity sampling, which have urged scientist to adopt systematic georeferencing methods when collecting samples (Araujo & Guisan, 2006; Guralnick *et al.*, 2007). This lack of georeferenced sampling localities considerably limits the amount of data that can be used in spatially-explicit analyses of host-parasite biogeography. However, while spatially-explicit analyses of sampling gaps cannot incorporate all previous studies, they do illustrate geographic sampling trends and guide future data collection efforts.

Display Mapping

Increasingly widespread use of Geographic Information Systems (GIS) allows for the illumination of patterns that would be more difficult to discern if the data were analyzed in tabular form. For example, mapping the distribution of primate parasite sampling using data from Hopkins & Nunn (2007) indicates that primates have been sampled heavily in East Africa, and have been comparatively under-sampled in places such as Southeast Asia (Figure 1a). Gradient or proportional maps are commonly used to further illustrate numeric discrepancies between localities (Rodrigues *et al.*, 2004; Rinaldi *et al.*, 2006). In our example, applying a proportional circle mapping technique to the number of primate parasite sampling records at each locality further emphasizes the higher abundance of studies on primate parasites in Africa, as compared to Asia and South America (Figure 1b).

Quantifying Spatial Distributions

While display maps yield general trends, further statistical analysis is needed to quantitatively compare two spatial distributions. Spatial statistics can be applied to continuous data, point data, or tessellated ("polygon") data, in order to measure both first and second order spatial effects. First order effects refer to spatial variation in the mean value of a process (i.e. a global trend). Second order effects refer to the spatial correlation structure or spatial dependence within the dataset, which may cause deviations from the global trend in specific smaller regions. A number of spatial statistics software modules have recently become available, some of which are present within GIS frameworks (e.g.

ESRI's geostatistical analyst/spatial analyst, GEODA, Mapping and Spatial Statistics Toolboxes for Matlab and S-Plus, SpaceStat Pack). For a discussion of available spatial statistics GIS modules see: (Anselin, 2005). We discuss several of the most common spatial statistical measures here that can be used to quantify geographic patterns of parasite sampling by measuring the degree of spatial clumping or correlation present in the dataset(s):

a) *Summary Statistics:* When analyzing geographic datasets, it may be useful to identify anomalous regions or regions with high variability by calculating regional summary statistics. In this case, the entire dataset is usually divided into local regions (also called 'windows' or 'neighborhoods') of a size specified by the researcher. Rectangular windows are used for ease of calculation and the size of the window depends on the overall dimensions of the area being studied, as well as the average distances between data locations (Isaaks & Srivastava, 1989). For small datasets, or datasets in which the data are irregularly spaced, windows can be overlapped, offering a smoothing effect which can easily quantify regional trends and isolate outliers. For example, Figure 2a quantifies the first order intensity of primate parasite sampling points in dense areas such as East Africa relative to sparsely populated areas such as East Asia, by applying a moving mean statistic with a 5x5 decimal degree window.

b) *Measures of Clumping and Dispersion:* When summary statistics indicate clumping of data points as in the primate parasite example, spatial statistics

can be applied to quantitatively determine the degree of clumping present, or whether the data observed are significantly more clumped than an expected distribution (i.e. random or uniform distributions). Measures can be applied both globally and to local neighborhoods or regions. Examples include: Kernel estimation, nearest neighbor distances, K-function approaches, the Clark-Evans test, the Cuzick and Edwards method, the GAM/K method, and spatial scan statistics (Bailey & Gatrell, 1995; Kulldorff & Nagarwalla, 1995; Cuzick & Edwards, 1996; Openshaw *et al.*, 1999; Ward & Carpenter, 2000; Anselin, 2005). These estimates have been critical in identifying clusters of disease in human and animal populations (Kulldorff & Nagarwalla, 1995; Cuzick & Edwards, 1996; Rinaldi *et al.*, 2006; Wheeler, 2007).

In field research, however, the researcher decides on the distribution of sample points. Hence, the question is often not whether sample points are more clumped than random or uniform distributions. Rather, the question of interest becomes whether attribute values at some points are more similar to the values at closer points than the values at farther points. When attribute values are spatially clustered, this phenomena is termed spatial *autocorrelation.*

c) *Spatial Autocorrelation:* Measures of spatial autocorrelation can not only inform the researcher as to whether clustering in attribute values exists, but also to the scale and direction of that effect. In positive associations attribute

values increase in similarity the closer the sampling points. In negative associations, dissimilar values are found in close spatial association. Two of the most common measures of global spatial autocorrelation are Moran's I and Geary's C (Moran, 1950; Geary, 1954). The most common measures of local spatial autocorrelation are collectively known as the LISA statistics (Local Indicators of Spatial Association), and include both local Moran and Geary statistics as well as the Getis-Ord statistics (Getis & Ord, 1992; Anselin, 1995). These measures of spatial autocorrelation have been central to many epidemiological efforts seeking to identify spatial clumping in disease prevalence or intensity (Guernier *et al.*, 2004; Zhang & Lin, 2007; Crighton *et al.*, 2008; Jacob *et al.*, 2008). In studies of parasite sampling, measures of spatial autocorrelation may be most useful as a means to identify violation of standard statistical assumptions when applying non-spatial statistical models to spatial data (see subsequent section on *Measuring Spatial Correlation*).

Data Modeling: Comparing two or more distributions

a) *Measuring Spatial Correlation:* Studies of host-parasite biogeography often seek to correlate geographic attributes (e.g. host diversity or environmental characteristics) with parasite species richness, prevalence, intensity or abundance. Spatial correlation is also important for those seeking to understand factors driving spatial patterns of parasite sampling or seeking to understand how well observed sampling distributions correlate with optimal sampling distributions. Researchers often attempt to address correlation in

spatial variables with standard methods that do not incorporate a spatial component. However, spatial data often violate the central assumption of many standard statistical procedures, namely the independence of data points. Non-independent data can yield faulty statistical tests, and they can result in lower statistical power than models that incorporate spatial information.

Thus, statistical analyses of spatially-distributed data should allow for the possibility that two data points may have similar values not due to one or more of the explanatory variables, but because the distance between these two data points is very small (i.e. spatial autocorrelation in the response variable is present). When using spatial data in statistical tests that are not spatially-explicit (e.g. OLS regression), tests for autocorrelation of standard model residuals should be completed. These tests can include the Moran's I or Geary's C methods above, as well as relatively recent calculations for this purpose using Lagrange Multipliers (Anselin, 1988; Anselin *et al.*, 1996).

If spatial autocorrelation is present, spatial models often result in better model fits (e.g. simultaneous autoregressive models (SAR) or conditional autoregressive models (CAR)). For example, when conducting a regression analysis of the relationship between species richness ('the optimal distribution') and the intensity of primate parasite sampling ('the observed distribution'), we found that the density of primate parasite sampling localities was correlated with host species richness, but that residuals from this regression model were spatially auto-correlated (Hopkins and Nunn 2007). When we applied an appropriate spatial regression model, the explanatory

power of the model increased from r^2=0.008 to R^2=0.05, but still was relatively poor, indicating that 95% of the variation in sampling effort could not be explained by host distributions. If the researchers are confident of model specifications, the residuals from spatial models can themselves serve as a quantification of the relationship between an optimal distribution and observed sampling distribution. However, traditional gap analyses also commonly use spatial layer manipulations within a GIS to quantify and illustrate differences between distributions.

b. Spatial Layer Manipulations: Spatial data within a GIS can be represented in two forms: feature data and raster data. Feature data are represented as the intersection of points, lines, and/or polygons. Raster data are represented in a grid format. When data are in raster format, two or more rasters can be combined by applying mathematical operations to each grid cell. In what follows, we provide two examples of spatial raster manipulations in which the distribution of primate parasite sampling points is compared to host distributions (from Hopkins and Nunn 2007).

Example 1: Quantile Subtraction

Quantile subtraction provides a straightforward means of quantifying dissimilarities between distributions, based on the concept of proportional sampling. Data from each layer are distributed evenly into bins ('quantiles') and the differences between layer quantile values are displayed according to

standard deviations from the mean. Resulting maps provide a geographic quantification of over- and under-sampling. Figure 2 (a-c) demonstrates this process for a comparison of primate parasite sampling distributions and primate host species richness. The resulting values clearly point to Central Africa, portions of the Amazon, and Borneo as the regions most in need of sampling for primate parasites.

Example 2: Sampling Factor Estimation

While quantile subtraction provides a straightforward means of quantifying dissimilarities in distributions, it fails to take into account pertinent factors other than geography, such as historical patterns of taxonomic sampling. The sampling factor approach is based upon conventional biodiversity gap analyses, which attempt to maximize species representation or complementarity in reserve networks. It combines geographic distributions of host species with historical sampling patterns to identify the sites that are the most under-sampled. Specifically, the sampling factor approach uses the percentage of hosts within each cell that have not been sampled at *any* geo-referenced location (Figure 3a) to determine the number of host species within a particular cell that need to be sampled in order to reach mean sampling levels (Figure 3b). Thus, it prioritizes areas both with high host species richness and high numbers of previously unsampled species, and pinpoints the geographic areas that would be most complementary to the current suite of sampled species and localities.

Sampling factor analyses applied to the distribution of primate parasite sampling revealed that while the overall mean percentage of unsampled primates at any given site is low (14%), this pattern varies extensively across regions. For example, up to 90% of the primates in large portions of Southeast Asia have not been sampled at a georeferenced location in the GMPD, and in order to reach mean sampling levels up to eight primate species would need to be sampled at some sites. Thus, by allocating sampling points according to species complementarity instead of species richness, a different optimal distribution was created, and results differed from the quantile subtraction method. Where quantile subtraction highlighted large portions of Africa as the most in need of sampling, this analysis indicated that Africa is comparatively over-sampled and instead allocated most research effort towards Asia.

Remedying Sampling Gaps

The final step in most gap analyses is to provide enough information on discrepancies between the observed and optimal distributions such that future sampling efforts can be targeted to remedy these discrepancies. Both measures of spatial correlation and spatial layer manipulations, such as the quantile subtraction and sampling factor methods listed above, provide quantitative measures of how geographic patterns of parasite sampling differ from geographic patterns of host species distributions. These methods are useful for quantifying geographic trends. However, researchers seeking to target just a few of the *most* under-sampled sites for future sampling efforts might benefit

from incorporating optimization techniques with traditional gap analysis techniques. For example, a number of sites in Southeast Asia and the Central Amazon have up to eight primate species that need to be sampled in order to reach global mean levels of parasite sampling. Optimization techniques can prioritize the sites to visit and even the order in which to visit them.

GUIDING FUTURE RESEARCH EFFORTS: TARGETING THE MOST UNDERSAMPLED SITES

Prioritizing sites for future sampling efforts

Optimization approaches derived from operations research and transportation engineering are increasingly being used in conservation biology to predict patterns of biodiversity (e.g. 'covering' problems, 'p-dispersion' problems, 'p-center problems', 'p-median problems', cluster analysis, compositional dissimilarity (Tansel *et al.*, 1983; Brandeau & Chui, 1989; Faith & Walker, 1996; Snelder *et al.*, 2006; Arponen *et al.*, 2008). With only minor modifications, these methods can be used to prioritize future parasite sampling sites.

Optimization techniques can be distance-based and/or attribute based. Distance methods use Euclidean distances between sites in order to place a site in an under-sampled area. Distance-based methods are most frequently applied to select sampling sites in epidemiological analyses conducted at smaller regional scales, where regular sampling is often a pre-requisite for statistical methods that create continuous disease-risk surfaces (e.g. Kriging or Bayesian Surface Estimation, (Best *et al.*, 2005; Rinaldi *et al.*,

2006)). In larger scale analyses, attribute values may have equal or greater weight than distance values. In these cases, non-Euclidean distances can be incorporated into analyses. Imagine, for example, that a researcher wishes to prioritize one of two sites that each contain one unsampled host species and an equal number of sampled species (the sites may or may not have host species in common). In such a case, it might be valuable to increase the phylogenetic breadth of sampling. In this example, phylogenetic distance between the unsampled species and its closest sampled relative at the site could serve as a 'non-Euclidean' distance.

In the next section, we use both Euclidean distances and non-Euclidean distances (phylogenetic relationships between unsampled and sampled host species) to illustrate two of the most common optimality approaches used currently for site selection for biological sampling. The first method follows a traditional gap analysis approach in which the site that is most different from the current suite of sites is selected. This approach has been used in a variety of contexts, and given a number of names (Faith & Norris, 1989; Belbin, 1993; Faith & Walker, 1996). Here, we refer to this approach as the F-N criteria after Faith and Norris (1989). The second optimality approach—'the p-median problem'—differs from the F-N criteria in that it seeks to identify the site (or a suite of *p* sites) that, if sampled, would reduce the overall mean distance between unsampled sites and the most similar sampled site (Tansel *et al.*, 1983; Faith & Walker, 1996). Thus, the p-median approach identifies the suite of sites that is most representative of all remaining target sites, whereas the F-N criteria selects the suite of sites that are most dissimilar from currently sampled sites. Both approaches allow for the incorporation of Euclidean and non-Euclidean (attribute based) distances. In this section, we apply

both approaches to the dataset from Hopkins & Nunn (2007) on primate parasite sampling to illustrate differences between these approaches.

Analysis

We used both the F-N criterion and the p-median methods to calculate the top 5 sites most in need of future sampling, according to two variables: Euclidean distance and phylogenetic distance(Bininda-Emonds *et al.*, 2007, 2008). These distances were chosen as a negative relationship has been observed between parasite community similarity and both distance between sampled habitats and phylogentic distance between hosts (Poulin, 2003; Martiny *et al.*, 2006; Davies & Pedersen, 2008). While a number of other factors could be incorporated in these analyses (e.g. ecosystem type, temperature, levels of precipitation), we feel that using just these two values illustrates the differences between the F-N and p-median approaches well, while allowing for comparison to spatial layer manipulations conducted in the previous section.

The two distance metrics were calculated as follows:

1. ***Euclidean Distance****: Distance (km) between a potential future sampling site and the nearest already sampled site. Locations of primate parasite sampling were obtained from the Global Mammal Parasite Database, and include sampling through 2008.*
2. ***Phylogenetic Distance:*** *Phylogenetic distance (millions of years) between an unsampled species and its closest sampled relative, summed for all species present at a site. Note: if a species has been sampled, its phylogenetic distance is*

0. Phylogenetic distances between pairs of primate species were calculated as time to last common ancestor using the mammalian supertree from Bininda-Emonds et al. (2007, 2008).

Analyses were conducted in a similar grid format to Hopkins & Nunn (2007), with 5017 one degree^2 grid cells, to allow for appropriate comparisons. All unsampled grid cells were considered as target sites in a discrete analysis. Each cell's attribute value was normalized by the maximum value prior to calculations, and both calculations were executed using an iterative greedy algorithm (i.e. only one site was selected at a time). An iterative process was chosen, as the total number of future sites is unknown, and an iterative approach reduces the necessary computational power required for the analyses. Prioritization of sampled sites during each iteration proceeded as follows:

F-N Criterion:

$$k_{t+1} = max \prod_{i=1}^{n} \left| x_{ij} - x_{ik} \right|_t w_i \qquad \text{(Eq. 1)}$$

In the next time step (t+1), sample the cell that has the maximum overall distance to its nearest cell (j) for all distance metrics (i), in the current time step (t). Relative impacts of distance metrics can be specified by giving each metric (i) different weights (w_i).

P-Median:

$$k_{t+1} = min \sum_{k=1}^{n} \prod_{i=1}^{n} \left| x_{ij} - x_{ik} \right|_t w_i \qquad \text{(Eq. 2)}$$

In the next time step (t+1), sample the cell that would minimize the sum of the distances between all 5017 grid cells (k) and their nearest sampled neighbor (j) for all distance metrics (i). Relative impacts of distance metrics can be specified by giving each metric (i) different weights (w_i).

F-N and p-median values were generated for both distance metrics separately and together. Calculations using both criteria gave both Euclidean and phylogenetic distances equal weights.

Site placement

The placement and sequence of selected sites differed significantly depending upon how the optimal cell was calculated (F-N criterion or P-median), and which distance metrics were used (Euclidean Distance, Phylogenetic Distance, or Both). The initial target site fell in the same general area (East and Southeast Asia), regardless of the method or distance metric used (Figures 4-5). However, substantial differences occurred between the remaining four sites. For example, the F-N criteria using only Euclidean distance prioritized the sites that are farthest from existing sampling sites, resulting in the selection of sites at the most northern and southern tips of Africa. Sites in these regions were not prioritized by any of the other methods. The least amount of variation between the F-N criterion and P-median methods is evident when considering phylogenetic distances alone (Figure 4 b-c). Using this criterion, three out of five sites remained the same, regardless of selection method. Most methods placed only one site in the Americas, although they differed somewhat in regional placement of this site. When both criteria

(Euclidean and Phylogenetic Distances) were given equal weights, the p-median and F-N approaches only converged at two localities. The p-median method using both criteria placed the most undersampled site in Myanmar, whereas the F-N method placed it in Laos.

By selecting a series of five complementary sites, these optimization approaches provided different information than was evident in the previous spatial layer manipulations. Through the incorporation of phylogenetic distances, these approaches can prioritize sites with similar levels of sampling effort, as reflected by sampling factor calculations. For example, although the sampling factor approach identified a large area in South America that requires between 3-8 primate species to be sampled in order to reach mean sampling levels (Figure 3b), the Americas were rarely prioritized using either the P-median or F-N methods. This could be the case if more species in the Americas have closely sampled relatives, whereas unsampled species in East and Southeast Asia are comparatively unique from an evolutionary perspective. In addition, by using an iterative approach to select a complementary set of sites, we obtained an indication of how the static maps in the previous sections are likely to change with further sampling effort results.

CONCLUSIONS

In the previous sections, we used data on the geographic sampling patterns of primate parasites to illustrate the potential of gap analysis techniques to quantify geographic patterns of sampling effort. These methods are widely accessible due to their extensive use in fields ranging from conservation biology to transportation engineering,

and have user-friendly implementations within GIS frameworks. As a result, they have great potential to inform studies of host-parasite biogeography. At the most basic level, these studies can aid in illuminating geographic sampling biases. In addition, by quantifying geographic processes in a spatially-explicit way, these studies can provide the first necessary step towards developing quantitative measures to account for spatial biases in sampling effort.

Nevertheless, some qualification in the interpretation of results is necessary. Gap analyses ultimately rely on the optimal distribution selected. This selection invariably results from a subjective process that depends largely on the goals of the researcher. Thus, two gap analyses on the same dataset may differ in conclusions depending upon what host or environmental criteria are prioritized. In addition, due to a widespread lack of geo-referencing of sample sites, any spatially explicit analysis of sampling patterns must also be taken as a sample. Thus, results can only yield relative trends, and no absolute conclusions regarding historical sampling patterns. Yet, even with these limitations, the quantitative measures of sampling yielded have enormous potential to aid those seeking to better understand how patterns of sampling effort impact our knowledge of host-parasite biogeography.

FIGURE LEGENDS

Figure 1: Geographic distribution of sampling for primate parasites. a) Distribution of primate parasite sampling points redrawn from Hopkins & Nunn (2007), using primate parasite records added to the Global Mammal Parasite Database (Nunn and Altizer, 2005) prior to 2009. b) Distribution of primate parasite sampling points, weighed by the number of records in the GMPD at each locality.

Figure 2: a) Smoothed intensity of primate parasite sampling points (number of points per 1 degree^2 cell). Redrawn from Hopkins & Nunn (2007), using an updated version of the Global Mammal Parasite Database (Nunn and Altizer, 2005). A moving mean was calculated using overlapping windows of 5 x 5 decimal degrees. Resulting values are displayed in 10 quantiles; b) Distribution of primate species richness, displayed in 10 quantiles; c) Quantile subtraction of parasite sampling intensity from primate species richness, displayed as standard deviations from the mean. Positive values indicate under-sampling.

Figure 3. a) Percentage of unsampled primate species per 1x1 decimal degree cell. Redrawn from Hopkins and Nunn (2007), using an updated version of the Global Mammal Parasite Database (Nunn and Altizer, 2005). b) The number of species that need to be sampled in each cell in order to reach mean sampling levels. Values are distributed in 1 SD bins (mean value = 0.36) to allow for visual

comparison to the quantile subtraction approach (Figure 2c). Positive values indicate under-sampling.

Figure 4. Prioritization of five geographic areas in need of future sampling using two optimization methods: the F-N method (prioritization of the most dissimilar site) and the p-median method (prioritization of the site that reduces the mean dissimilarity of unsampled and sampled sites the most). Values are selected according to a) Euclidean distance (F-N and p-median methods); b) Phylogenetic distance (F-N method); and c) Phylogenetic distance (p-median method).

Figure 5. Prioritization of 5 geographic areas in need of future sampling by giving equal weights to Euclidean and phylogenetic distances. Calculations made using a) the F-N method and b) the p-median method.

REFERENCES

Anselin, L. (1988) Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, **20**, 1-17

Anselin, L. (1995) Local indicators of spatial association-LISA. *Geographical Analysis*, **27**, 93-115

Anselin, L. (2005) Spatial statistical modeling in a gis environment. *GIS, spatial analysis, and modeling* (ed. by D.J. Maguire, M. Batty and M.F. Goodchild), pp. 93-107. ESRI Press, Redlands, CA.

Anselin, L., Bera, A.K., Florax, R. & Yoon, M.J. (1996) Simple diagnostic tests for spatial dependence. In, pp. 77-104. Elsevier Science Bv

Araujo, M.B. & Guisan, A. (2006) Five (or so) challenges for species distribution modeling. *Journal of Biogeography*, **33**, 1677-1688

Arponen, A., Moilanen, A. & Ferrier, S. (2008) A successful community-level strategy for conservation prioritization. *Journal of Applied Ecology*, **45**, 1436-1445

Bailey, T.C. & Gatrell, A.C. (1995) *Interactive spatial data analysis.* Pearson Education, Edinburgh, Scotland.

Belbin, L. (1993) Environmental representativeness: Regional partitioning and reserve selection. *Biological Conservation*, **66**, 223-30

Best, N., Richardson, S. & Thomson, A. (2005) A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, **14**, 35-59

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., Macphee, D.E., Beck, R.M.D., Grenyer, R. & Price, R.D. (2007) The delayed rise of present-day mammals. *Nature*, **446**, 507-512

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., Macphee, D.E., Beck, R.M.D., Grenyer, R. & Price, R.D. (2008) Corrigendum: The delayed rise of present day mammals. *Nature*, **456**, 274

Brandeau, M.L. & Chui, S.S. (1989) An overview of representative problems in location research. *Management Science*, **35**, 654-74

Burke, A. (2007) How sampling effort affects biodiversity measures in an arid succulent karoo biodiversity hotspot. *African Journal of Ecology*, **46**, 488-499

Cam, E., Nichols, J.D., Hines, J.E., Sauer, J.R., Alpizar-Jara, R. & Flather, C.H. (2002) Disentangling sampling and ecological explanations underlying species-area relationships. *Ecology*, **83**, 1118-1130.

Crighton, E.J., Elliott, S.J., Kanaroglou, P., Moineddin, R. & Upshur, R.E.G. (2008) Spatio-temporal analysis of pneumonia and influenza hospitalizations in ontario, canada. *Geospatial Health*, **2**, 191-202

Cuzick, J. & Edwards, R. (1996) Clustering methods based on k nearest neighbour distributions. *Methods for investigating localized clustering of disease.* (ed. by F.E. Alexander and P. Boyle). International agency for research on cancer., Lyon, France.

Davies, T.J. & Pedersen, A.B. (2008) Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proceedings of the Royal Society B-Biological Sciences*, **275**, 1695-1701

Faith, D.P. & Norris, R.H. (1989) Correlation of environmental variables with patterns of distribution and abundance of common and rare fresh-water macroinvertebrates. *Biological Conservation*, **50**, 77-98

Faith, D.P. & Walker, P.A. (1996) Environmental diversity: On the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. *Biodiversity and Conservation*, **5**, 399-415

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? *Systematic Biology*, **51**, 331-363

Funk, V.A., Richardson, K. & Ferrier, S. (2005) Survey-gap analysis in expeditionary research: Where do we go from here? *Biological Journal of the Linnean Society*, **2005**, 549-567

Funk, V.A. & Richardson, K.S. (2002) Systematic data in biodiversity studies: Use it or lose it. *Systematic Biology*, **51**, 303-316

Geary, R.C. (1954) The contiguity ratio and statistical mapping. *The incorporated statistician.*, 115-145

Getis, A. & Ord, J.K. (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis*, **24**, 189-206

Gregory, R.D. (1990) Parasites and host geographic range as illustrated by water fowl. *Functional Ecology*, **4**, 645-54

Guegan, J.-F. & Kennedy, C.R. (1996) Parasite richness/sampling effort/host range: The fancy three-piece jigsaw puzzle. *Parasitology Today*, **12**, 367-369

Guernier, V., Hochberg, M.E. & Guegan, J.F. (2004) Ecology drives the worldwide distribution of human diseases. *PLoS Biology*, **2**, 740-746

Guralnick, R.P., Hill, A.W. & Lane, M. (2007) Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters*, **10**, 663-672

Hopkins, M.E. & Nunn, C.L. (2007) A global gap analysis of infectious agents in wild primates. *Diversity and Distributions*, **13**, 561-572

Hortal, J. & Lobo, J.M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2913-2947

Isaaks, E.H. & Srivastava, R.M. (1989) *An introduction to applied geostatistics*. Oxford University Press, Oxford.

Jacob, B.G., Muturi, E.J., Caamano, E.X., Gunter, J.T., Mpanga, E., Ayine, R., Okelloonen, J., Nyeko, J.P.M., Shililu, J.I., Githure, J.I., Regens, J.L., Novak, R.J. & Kakoma, I. (2008) Hydrological modeling of geophysical parameters of arboviral and protozoan disease vectors in internally displaced people camps in gulu, uganda. *International Journal of Health Geographics*, **7**

Jennings, M.D. (2000) Gap analysis: Concepts, methods, and recent results. *Landscape Ecology*, **15**, 5-20

Kery, M., Royle, J.A. & Schmid, H. (2008) Importance of sampling design and analysis in animal population studies: A comment on Sergio et al. *Journal of Animal Ecology*, **45**, 981-986

Krasnov, B.R., Shenbrot, G.I., Khokhlova, I.S., Mouillot, D. & Poulin, R. (2008) Latitudinal gradients in niche breadth: Empirical evidence from haematophagous ectoparasites. *Journal of Biogeography*, **35**, 592-601

Kulldorff, M. & Nagarwalla, N. (1995) Spatial disease clusters:Detection and inference. *Statistics in Medicine.*, **14**, 799-810

Lindenfors, P., Nunn, C.L., Jones, K.E., Cunningham, A.A., Sechrest, W. & Gittleman, J.L. (2007) Parasite species richness in carnivores: Effects of host body mass,

latitude, geographical range and population density. *Global Ecology and Biogeography*, **16**, 496-509

Lobo, J.M. (2008) Database records as a surrogate for sampling effort provide higher species richness estimations. *Biodiversity and Conservation*, **17**, 873-881

Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., Green, J.L., Horner-Devine, M.C., Kane, M., Krumins, J.A., Kuske, C.R., Morin, P.J., Naeem, S., Ovreas, L., Reysenbach, A.-L., Smith, V.H. & Staley, J.T. (2006) Microbial biogeography: Putting microorganisms on the map. *Nature*, **4**, 102-112

Moran, P.A.P. (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17-23

Nunn, C.L., Altizer, S., Jones, K.E. & Sechrest, W. (2003) Comparative tests of parasite species richness in primates. *The American Naturalist*, **162**, 597-614

Nunn, C.L. & Altizer, S.A. (2006) *Infectious diseases in primates: Behavior, ecology, and evolution*. Oxford University Press, Oxford.

Nunn, C.L. & Altizer, S.M. (2005) The global mammal parasite database: An online resource for infectious disease records in wild primates. *Evolutionary Anthropology*, **14**, 1-2

Nunn, C.L., Altizer, S.M., Sechrest, W. & Cunningham, A.A. (2005) Latitudinal gradients of disease risk in primates. *Diversity and Distributions*, **11**, 249-256

Openshaw, S., Turton, E. & Macgill, J. (1999) Using the geographical analysis machine to analyze limiting long-term illness census data. *Geographical and Environmental Modeling.*, **3**, 83-99

Poulin, R. (1997) Species richness of parasite assemblages: Evolution and patterns. *Annual Review of Ecology and Systematics*, **28**, 341-358

Poulin, R. (1998) Comparison of three estimators of species richness in parasite component communities. *Journal of Parasitology*, **84**, 485-490

Poulin, R. (2003) The decay of similarity with geographical distance in parasite communities of vertebrate hosts. *Journal of Biogeography*, **30**, 1609-1615

Poulin, R. & Morand, S. (2000) The diversity of parasites. *The Quarterly Review of Biology*, **75**, 277-293

Rinaldi, L., Musella, V., Biggeri, A. & Cringoli, G. (2006) New insights into the application of geographical information systems and remote sensing in veterinary parasitology. *Geospatial Health*, **1**, 33-47

Robertson, M.P. & Barker, N.P. (2006) A technique for evaluating species richness maps generated from collections data. *South African Journal of Science*, **102**, 77-85

Rodrigues, A.S.L., Akcakaya, H.R., Andelman, S.J., Bakarr, M.I., Boitani, L., Brooks, T.M., Chanson, J.S., Fishpool, L.D.C., Fonseca, G.A.B.D., Gaston, K.J., Hoffmann, M., Marquet, P.A., Pilgrim, J.D., Pressey, R.L., Schipper, J., Sechrest, W., Stuart, S.N., Underhill, L.G., Waller, R.W., Watts, M.E.J. & Yan, X. (2004) Global gap analysis: Priority regions for expanding the global protected-area network. *BioScience*, **54**, 1092-1100

Sarkar, S., Pressey, R., Faith, D., Margules, C., Fuller, T., Stoms, D., Moffett, A., Wilson, K., Williams, K., Williams, P. & Andelman, S. (2006) Biodiversity conservation planning tools: Present status and challenges for the future. *Annual Review of Environmental Resources*, **31**, 123-159

Schoereder, J.H., Glabiati, C., Ribas, C., Sobrinho, T., Sperber, C., Desouza, O. & Lopes-Andrade, C. (2004) Should we use proportional sampling for species-area studies? *Journal of Biogeography*, **31**, 1219-1226

Scott, J.M., Csuti, B., Jacobi, J.D. & Estes, J.E. (1987) Species richness. *BioScience*, **37**, 782-788

Snelder, T., Dey, K. & Leathwick, J. (2006) A procedure for making optimal selection of input variables for multivariate environmental classifications. *Conservation Biology*, **21**, 365-375

Tansel, B.C., Francis, R.L. & Lowe, T.J. (1983) Location on networks: A survey. Parti: The p-center and p-median problems. *Management Science*, **29**

Tripet, F., Christe, P. & Moller, A.P. (2002) The importance of host spatial distribution for parasite specialization and speciation: A comparative study of bird fleas. *Journal of Animal Ecology*, **71**, 735-748

Walther, B.A., Cotgreave, P., Price, R.D., Gregory, R.D. & Clayton, D.H. (1995) Sampling effort and parasite species richness. *Parasitology Today*, **11**, 306-310

Walther, B.A. & Martin, J.-L. (2001) Species richness estimation of bird communities: How to control for sampling effort. *Ibis*, **143**, 413-419

Walther, B.A. & Morand, S. (1998a) Comparative performance of species richness estimation methods. *Parasitology* **116**, 395-405

Walther, B.A. & Morand, S. (1998b) Comparative performance of species richness estimation methods. *Parasitology*, **116**, 395-405

Ward, M.P. & Carpenter, T.E. (2000) Techniques for analysis of disease clustering in space and in time in veterinary epidemiology. *Preventative Veterinary Medicine*, **45**, 257-284

Wheeler, D.C. (2007) A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in ohio, 1996-2003. *International Journal of Health Geographics*, **6**, 13

Zhang, T.L. & Lin, G. (2007) A decomposition of moran's i for clustering detection. *Computational Statistics & Data Analysis*, **51**, 6123-6137

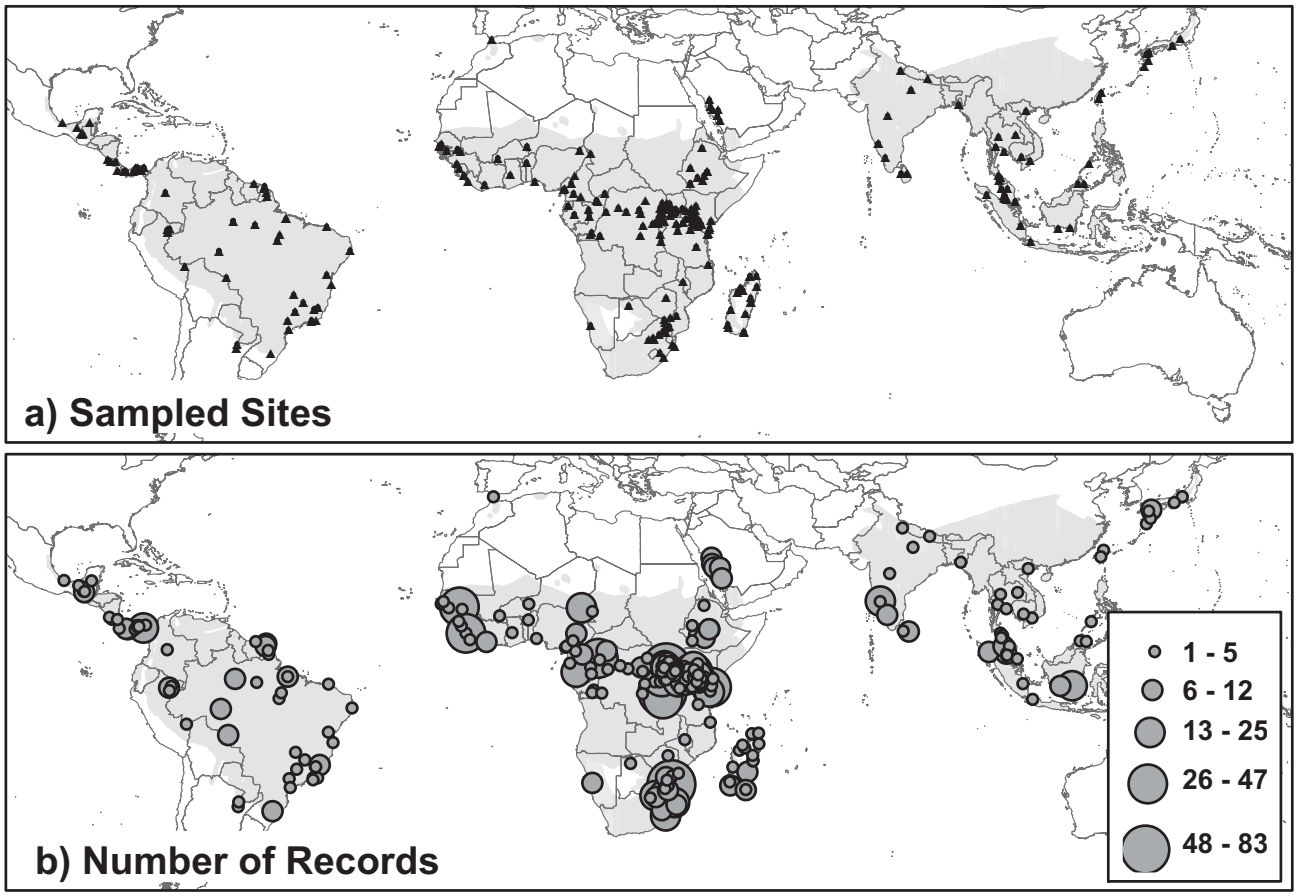

Figure 1.

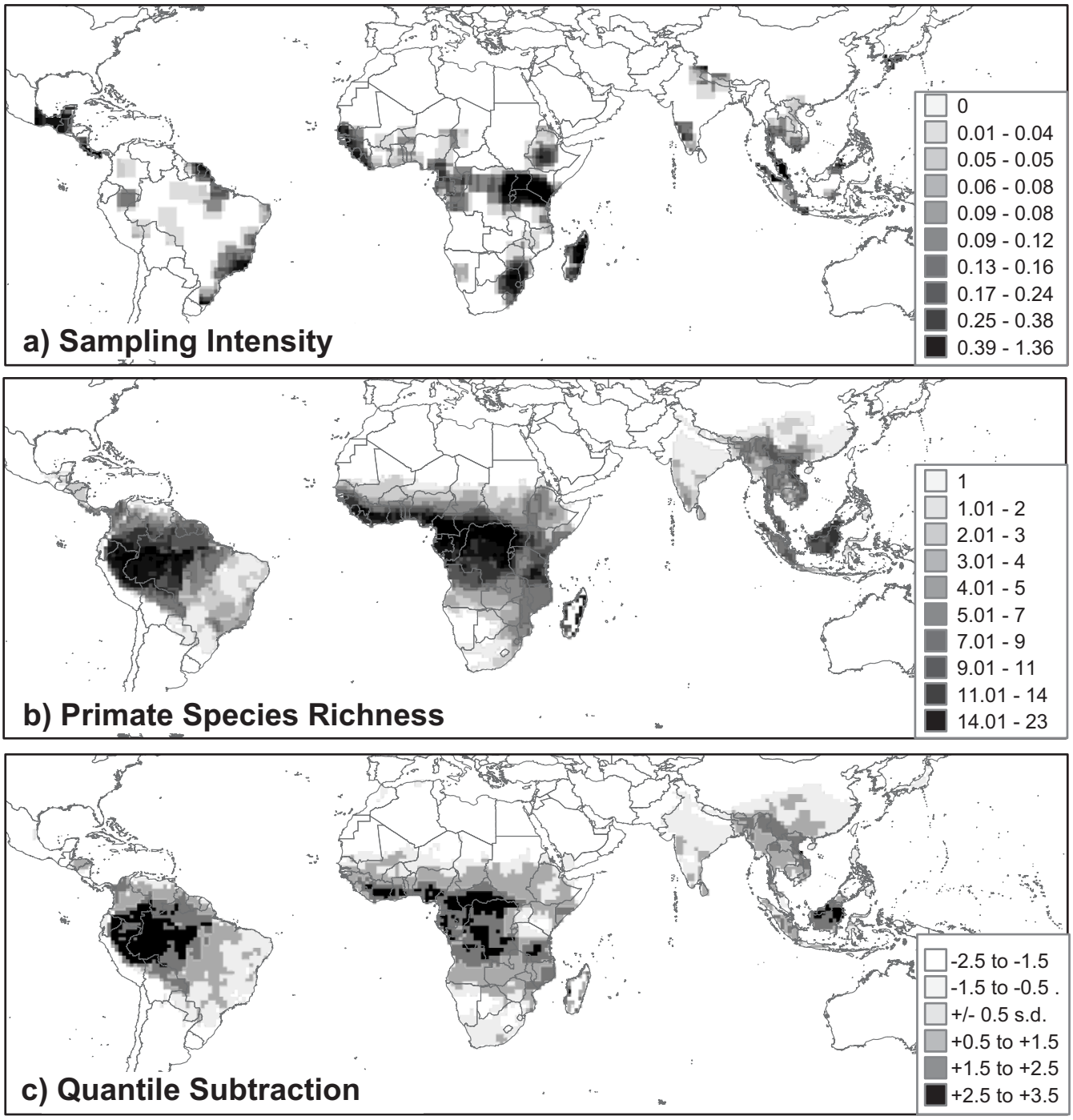
a) Sampling Intensity
0
0.01 - 0.04
0.05 - 0.05
0.06 - 0.08
0.09 - 0.08
0.09 - 0.12
0.13 - 0.16
0.17 - 0.24
0.25 - 0.38
0.39 - 1.36
b) Primate Species Richness
1
1.01 - 2
2.01 - 3
3.01 - 4
4.01 - 5
5.01 - 7
7.01 - 9
9.01 - 11
11.01 - 14
14.01 - 23
c) Quantile Subtraction
-2.5 to -1.5
-1.5 to -0.5 .
+/- 0.5 s.d.
+0.5 to +1.5
+1.5 to +2.5
+2.5 to +3.5

Figure 2.

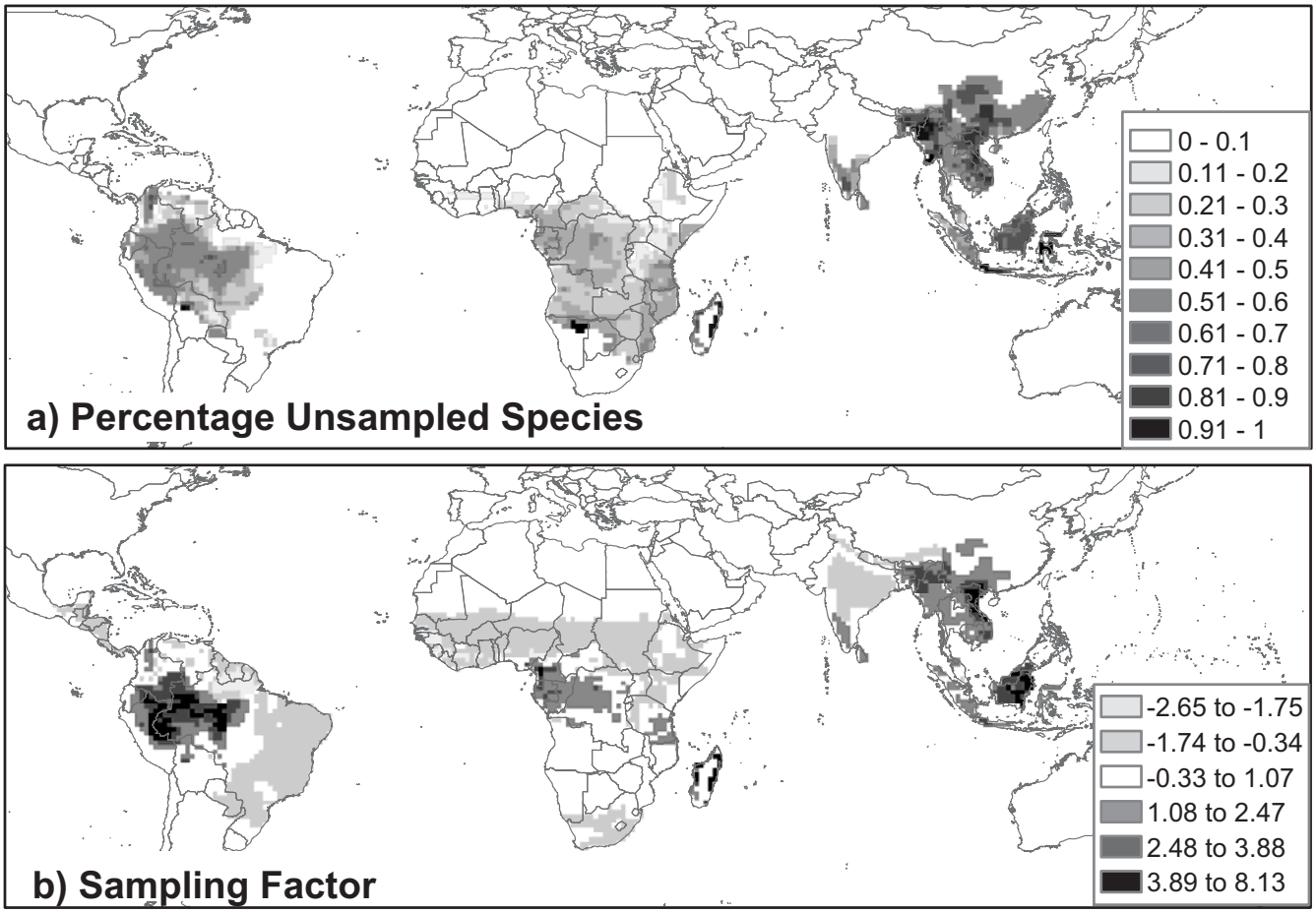
a) Percentage Unsampled Species
0 - 0.1
0.11 - 0.2
0.21 - 0.3
0.31 - 0.4
0.41 - 0.5
0.51 - 0.6
0.61 - 0.7
0.71 - 0.8
0.81 - 0.9
0.91 - 1
b) Sampling Factor
-2.65 to -1.75
-1.74 to -0.34
-0.33 to 1.07
1.08 to 2.47
2.48 to 3.88
3.89 to 8.13

Figure 3.

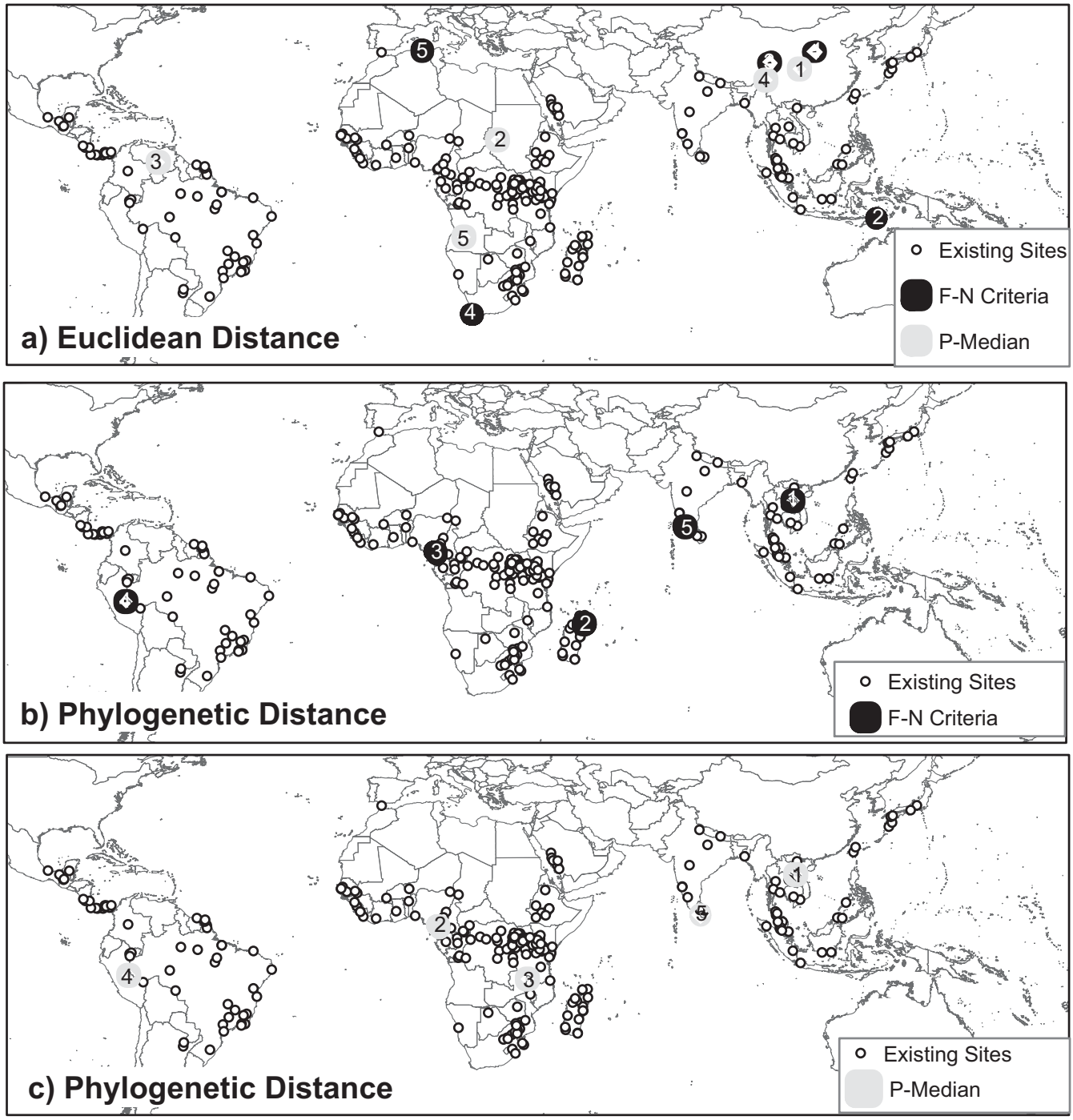
a) Euclidean Distance
Existing Sites
F-N Criteria
P-Median
b) Phylogenetic Distance
Existing Sites
F-N Criteria
c) Phylogenetic Distance
Existing Sites
P-Median

Figure 4.

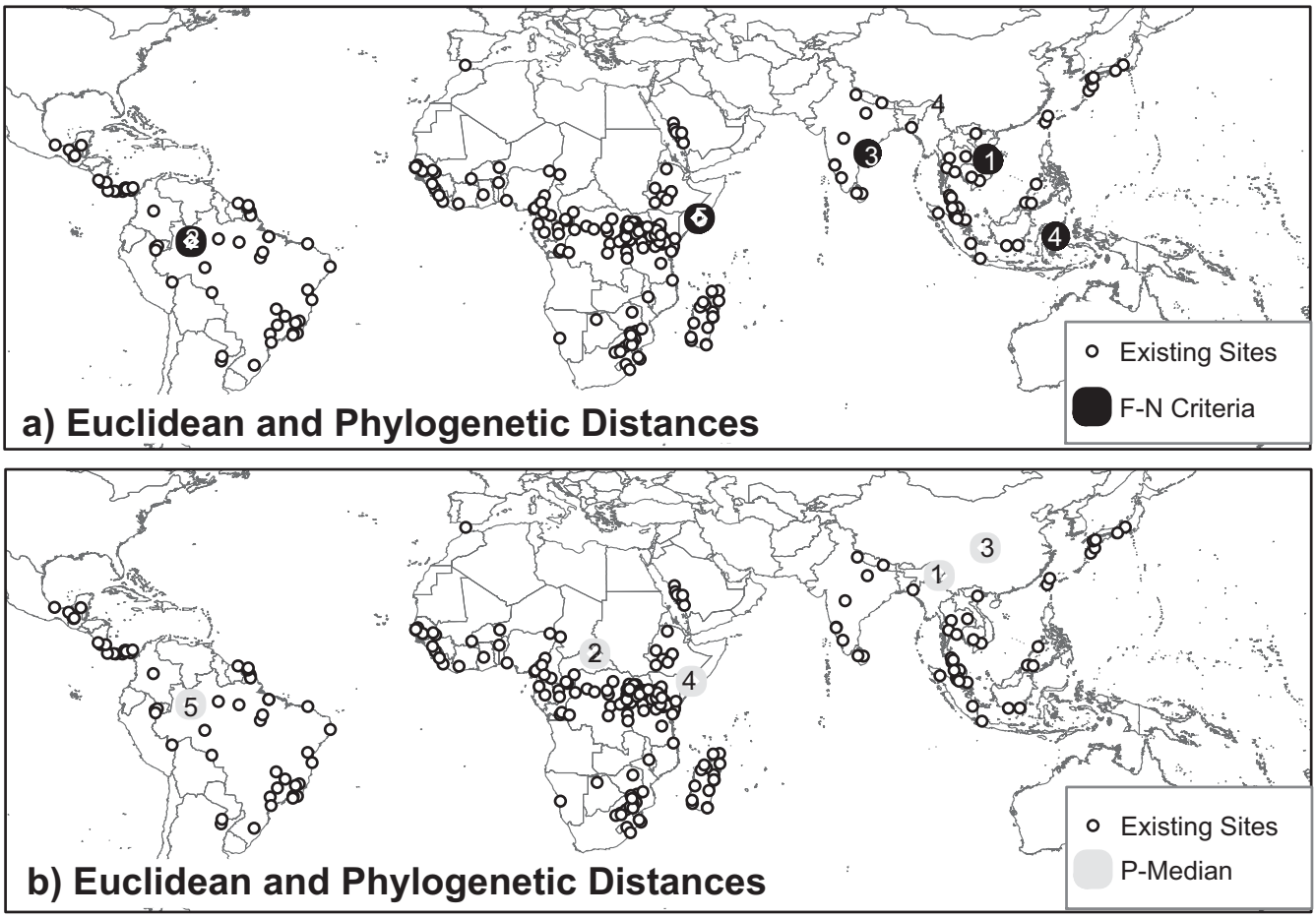
a) Euclidean and Phylogenetic Distances
Existing Sites
F-N Criteria
b) Euclidean and Phylogenetic Distances
Existing Sites
P-Median

Figure 5.