

Short title: GenGIS 2

GenGIS 2: Geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework

Software Article

Donovan H. Parks^{1,2}, Timothy Mankowski¹, Somayyeh Zangoeei¹, Michael S. Porter¹, David G. Armanini^{3&4}, Donald J. Baird⁴, Morgan G. I. Langille¹ and Robert G. Beiko^{1,*}

¹Faculty of Computer Science, Dalhousie University

²Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland

³Prothea srl, Milan, Italy

⁴Environment Canada @ Canadian Rivers Institute, University of New Brunswick, Fredericton, Canada

*Corresponding author:

Robert G. Beiko

Faculty of Computer Science, Dalhousie University

6050 University Avenue

Halifax, Nova Scotia, Canada B3H 1W5

Tel: 1 902 494 8043

Fax: 1 902 492 1517

Email: beiko@cs.dal.ca

Abstract. GenGIS is free and open source software designed to integrate biodiversity data with a digital map and information about geography and habitat. While originally developed with microbial community analyses and phylogeography in mind, GenGIS has been applied to a wide range of datasets. A key feature of GenGIS is the ability to test geographic axes that can correspond to routes of migration or gradients that influence community similarity. Here we introduce GenGIS version 2, which extends the linear gradient tests introduced in the first version to allow comprehensive testing of all possible linear geographic axes. GenGIS v2 also includes a new plugin framework that supports the development and use of graphically driven analysis packages: initial plugins include implementations of linear regression and the Mantel test, calculations of alpha-diversity (e.g., Shannon Index) for all samples, and geographic visualizations of dissimilarity matrices. We have also implemented a recently published method for biomonitoring reference condition analysis (RCA), which compares observed species richness and diversity to predicted values to determine whether a given site has been impacted. The newest version of GenGIS supports vector data in addition to raster files. We demonstrate the new features of GenGIS by performing a full gradient analysis of an Australian kangaroo apple data set, by using plugins and embedded statistical commands to analyze human microbiome sample data, and by applying RCA to a set of samples from Atlantic Canada. GenGIS release versions, tutorials and documentation are freely available at <http://kiwi.cs.dal.ca/GenGIS>, and source code is available at <https://github.com/beiko-lab/gengis>.

Keywords: phylogeography, biogeography, geophylogeny, Kangaroo apples, *Solanum*, human microbiome, environmental assessment, biomonitoring, reference condition analysis

Introduction

Phylogeography aims to relate evolutionary processes to spatial, temporal and environmental factors in order to understand past and present biodiversity (Avice et al. 1987; Kidd 2010; Hickerson et al. 2010). Purpose-built software tools allow the geographic analysis of biodiversity by incorporating phylogenetic information in different ways. Several programs provide visualizations of 2D or 3D geophylogenies, along with functionality to explore the temporal structure of this data (Kidd and Liu 2008; Janies et al. 2011) or the geographic uncertainty associated with Bayesian techniques for phylogeographic inference (Bielejec et al. 2011). Other programs have focused on displaying the spatial distribution and diversity of taxa using pie charts or geographic heat maps, and incorporate phylogenetic information by focusing on specific lineages or assigning molecular sequences to haplotypes (Hijmans et al. 2001; Laffan et al. 2010; Tsai 2011). We developed GenGIS (Beiko et al. 2008; Parks et al. 2009a) as a tool to merge geographic, ecological and phylogenetic biodiversity data in a single interactive visualization and analysis environment which supports a wider range of visualization and analysis options than existing software packages. Since its release, GenGIS has been used to investigate the phylogeography of viruses (Parks et al. 2009b; Tucker et al. 2011), bacteria (Farikou et al. 2011), plants (Allal et al. 2011; Poczai et al. 2011), animals (Ruzzante et al. 2011; Shafer et al. 2011), insects (Schoville and Roderick 2010), humans (Loo et al. 2011), and language families (Walker et al. 2012).

Within GenGIS, potential migration routes or the influence of geography on community similarity can be explored by proposing linear or non-linear geographic axes, and visualizing the goodness of fit between a tree topology and the ordering of sample sites along the specified axis (Parks and Beiko 2009). The first version of GenGIS required the user to draw geographic axes by hand, allowing the testing of specific hypotheses but making it difficult to explicitly test all possible axes. Poczai et al. (2011) noted that "GenGIS does not allow broad testing of encoded hypotheses with automatic polyline enumeration" and manually tested a subset of all possible axes for a set of georeferenced kangaroo apple samples. We have developed version 2 of GenGIS, a major new release with improved stability and functionality on both the Windows and Macintosh platforms. GenGIS v2 now includes an automated test of *all* possible linear geographic axes and a new plugin architecture that allows the scripting and execution of custom data analyses. Included with GenGIS are plugins implementing widely used statistical

approaches such as the Mantel test, several alpha diversity measures, and geographic visualizations of dissimilarity matrices. We have also developed a plugin that implements the Atlantic Canada river reference model of Armanini et al. (2012), which demonstrates the use of GenGIS in ecosystem biomonitoring, an area currently being transformed by genomics approaches (Hajibabaei et al, 2011; Baird and Hajibabaei, 2012).

Here we illustrate the new features of GenGIS using three datasets. We first demonstrate the application of new gradient methods to the kangaroo apple dataset of Poczai et al. (2011) to find the optimal geographical axis and test its significance. The functionality of several plugins is demonstrated using microbial community data from 28 different body sites sampled by Costello et al. (2009), revealing alpha-diversity patterns by body site and clustering fecal samples from different individuals. Finally, we demonstrate the application of the Atlantic Canada reference model on benthic macroinvertebrate samples from a series of 16 potentially impaired sites in Nova Scotia.

Methods

Data Acquisition and Formats

GenGIS makes use of digital map data, sample site information, sequence data, and one or more phylogenetic trees. Digital map data can be freely obtained from several online sources as described on the GenGIS website (<http://kiwi.cs.dal.ca/GenGIS>), or using our freely available program, MapMaker. A wide range of raster file formats are supported, including both digital elevation maps for visualizing 3D terrain and georeferenced image files for displaying standard map or satellite imagery. GenGIS also supports vector files and allows these to be overlaid on a raster map. The locations file indicates the geographic coordinates of each sample site, along with other relevant site attributes (e.g., habitat parameters, sampling details). Information about sequences, such as taxonomy or molecular function, collected at each sample site can be specified in an optional sequence file. Tree files are automatically georeferenced by associating leaf node names with the unique identifier used to specify either the sample sites or sequences. Complete descriptions of file formats and example data files are available on the GenGIS website. GenGIS v2 also allows the state of a previous GenGIS session to be restored from a file.

Phylogeographic Techniques

A linear axis defines an ordering of sample sites based on their projection onto the specified geographic gradient (Fig. 1a). The fit between a tree and a geographic gradient is determined by finding the ordering of leaf nodes, subject to the constraints of the tree topology, which most closely matches the ordering of sample sites. When these two orderings are shown in parallel, any mismatches between them will cause crossings between the lines connecting leaf nodes to their associated sample sites. The optimal tree layout is the one resulting in the fewest crossings. We previously proposed a branch-and-bound algorithm which allows the globally optimal layout to be found for large, multifurcating trees (~1000 leaf nodes with an average node degree < 8) in interactive time, i.e. <100 ms (Parks and Beiko, 2009).

In GenGIS v2, we have developed a method, *linear axes analysis*, for efficiently determining the number of crossings which occur for any linear axis. The proposed method is a plane sweeping algorithm where a line (i.e., the linear axis) is rotated by 180° (de Berg et al. 1997). A 180° sweep is sufficient as the ordering of sample sites along a gradient at angle θ will be identical to the ordering at $\theta+180^\circ$ (i.e., gradients are treated as being undirected). The key insight of the proposed algorithm is observing that the ordering of sample sites along the sweep line changes only when the line between a pair of sites becomes perpendicular to the sweep line. This suggests an $O(n^2 \log n^2)$ algorithm for determining the number of crossings which occur for any linear axis. For each of the $n(n-1)/2$ pairs of sample sites, the slope of the line connecting the sites and the site information is stored in an array. This array is then sorted in ascending order of slope values. Starting from a horizontal sweep line where sample sites are ordered by their longitudinal (x-axis) position, the sorted slope array indicates the order in which sample sites must be swapped as the sweep line is rotated. For each permutation of the sample sites, the optimal tree layout is determined and the number of crossings for the current orientation of the sweep line stored (see Fig. 1c for a plot of such data). Pseudocode along with a discussion of degenerate cases (e.g., collinear points) is given in the Supplementary Material.

Visualization and Data Analysis

The Python console in GenGIS allows users to interact directly with data through the GenGIS application programming interface, and allows analyses to be performed using the scipy (<http://www.scipy.org/>) and numpy (<http://numpy.scipy.org/>) libraries. Users can also execute commands in the R statistical programming language (<http://www.r-project.org>) via the RPy2

library. New in GenGIS version 2 is a framework that allows users to develop plugins that provide a graphical interface for frequently used operations. Plugins can import external files and produce visualizations in both the plugin window itself and the map environment. Default plugins in GenGIS include regression analysis, the Mantel test, alpha-diversity calculation, and visualizations of external matrices indicating pairwise site dissimilarities.

Reference Condition Analysis

We used the Atlantic Canada reference model (Armanini et al., 2012) to assess whether a range of potentially impacted sites had reduced biodiversity below what would be expected based on a set of reference sites. The reference model was trained using a RIVPACS-style approach, with the incorporation of four geospatial variables, namely long-term annual temperature range (°C), % intrusive rocks, % sedimentary rocks, and % sedimentary and volcanic rocks. By means of comparison with an independent validation dataset, the precision and accuracy level, measured in terms of RMSE and standard deviation of the O/E ratio, were confirmed to align well with other comparable models. Test data sites were analyzed to compare observed biodiversity with the predictions of the reference model to generate the observed/expected (O/E) ratio. The following O/E measures were computed based on the model: (i) Taxon richness O/E; (ii) Shannon diversity O/E; (iii) Pielou's Evenness O/E; and (iv) Berger-Parker dominance index O/E. To compute the DF-based O/E measure for each index we have extended the work of Van Sickle et al. (2006, 2007) and Armanini et al. (2012). Since the focus of the analysis is on impacted sites with $O/E < 1.0$, all O/E ratios greater than 1.0 were constrained to the value 1.0 for the purposes of data visualization.

Test Datasets

The kangaroo apple analysis was conducted using the phylogeny and sample site information provided by Pocazi et al. (2011). Map data of Australia was obtained using MapMaker, a companion program to GenGIS that allows custom georeferenced maps to be derived from the digital map data provided by Natural Earth (<http://www.naturalearthdata.com/>). These data files are provided in the Supplemental Information.

Body site data from the Costello et al. (2009) dataset were obtained from the DNA Data Bank of Japan (ERA000159). The source data consisted of FASTQ files containing amplicons of

variable region 2 (V2) of the 16S ribosomal RNA gene. We used version 7 of the RDP classifier (Wang et al. 2007) as implemented in mothur 1.16.1 (Schloss et al., 2009) to assign taxonomy to all 16S sequences in this dataset. The resulting taxon counts generated were used to generate visual summaries for each body site. Bray-Curtis distances were calculated for each pair of samples, and the resulting distance matrix subjected to Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering in mothur. The background image was modified based on an original image obtained from Wikimedia Commons (http://commons.wikimedia.org/wiki/File:Human_body_silhouette.svg).

RCA was based on benthic macroinvertebrate samples collected between 2002 – 2011 in the Atlantic Maritime ecozone. Data employed for the example here are described in detail in Armanini et al. (2012). Samples were obtained from reference sites included in the calibration dataset used for model construction (n=128). Reference sites were distributed throughout New Brunswick, Nova Scotia, and Newfoundland. Test sites (n = 16) used for model testing in the present paper were collected in the Upper Mersey area of Nova Scotia. Most macroinvertebrate samples were collected using a standardized traveling kick method, in which the operator disturbs the river substrate to dislodge attached and unattached organisms, which are washed into a triangular net of 400- μ m mesh size while zig-zagging upstream. Samples were subsequently sorted in the lab and identified to the taxonomic level of family, to allow the identification of sites deviating from expected assemblage composition. Topographical data were obtained from the Shuttle Radar Topography Mission (SRTM) dataset, via the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC: <http://daac.ornl.gov/>). Overlaid on the topography map was vector data describing rivers in Atlantic Canada, obtained from Geobase (<http://www.geobase.ca>).

Results and Discussion

Biogeography of kangaroo apples

The biogeography of kangaroo apples (*Archaeosolanum*) in Australia and Papua New Guinea was recently examined by Poczai et al. (2011). In their analysis, GenGIS was used to investigate geographic structuring by manually testing a subset of all possible linear axes. Here we demonstrate how the linear axes analysis function in GenGIS v2 allows all possible linear

axes to be easily evaluated. Defining a strict west-east axis results in 23 crossings (Fig. 1a), whereas a strict north-south axis produces 57 crossings (Fig. 1b). While this reveals stronger longitudinal than latitudinal structuring, by evaluating all linear axes we can determine the globally optimal axis, and the range of angles over which statistically significant results are obtained. For the kangaroo apple phylogeny, the minimum crossing number of 23 occurs multiple times between 90° and 103° , while the maximum number of crossings (77) occurs around 172° (Fig. 1c). Even with a conservative critical value of $\alpha = 0.001$, a wide range of axis orientations (90° to $\sim 150^\circ$, and $\sim 220^\circ$ to 270°) result in significantly fewer crossings than expected under a random model (Fig. 1c), strongly supporting spatial structuring centered on a longitudinal gradient. A linear axes analysis can also be applied to specific lineages. On the *Similia* lineage (Fig. 1a), no linear axis resulted in fewer crossings than expected under the null model (Fig. 1d). For the *Avicularia/Laciniata* lineage marginally significant (p -value < 0.1) results were obtained for linear axes slightly south of due east and between $\sim 230^\circ$ to 270° (Fig. 1e). The absence of notable longitudinal structuring within either subclade suggests that the strong longitudinal structuring found for the full phylogeny is primarily due to species within the *Similia* subclade being to the west of those within the *Avicularia/Laciniata* subclade.

Statistical Analysis of Human-Associated Microbes

Costello et al. (2009) profiled the microbial communities from nine individuals at 28 body sites and several time points using 16S rDNA genes as taxonomic markers, and concluded that community structure for a given individual at any site was relatively stable, although variation was seen among individuals. As reported by the original authors, taxonomic composition varies considerably by body site, with Actinobacteria and Bacilli predominant in many sites, other groups such as Proteobacteria found in some locations, but with gut (fecal) samples dominated by Bacteroidia and Clostridia (Fig. 2a). The dataset we used focused on the gut samples of six individuals, collected in two different ways (from toilet paper or directly from feces) and at two different time points: each individual was therefore associated with a total of four samples. UPGMA clustering of these communities based on differences in class abundance yielded perfectly clustered samples for two individuals (M1 and F1) with relatively high proportions of sequences assigned the label "Unclassified Bacteria" by the RDP classifier, but intermingled samples from the other four individuals based on the relative abundance of

Clostridia versus Bacteroidia (Fig. 2b). We then used the linear regression plugin (Fig. 3a) to examine the relationship between different classes: a regression of "Unclassified Bacteria" versus Bacteroidia produced a statistically significant ($p = 0.00001$) negative association between the two groups (Fig. 3b). This suggests that sequences labeled as "Unclassified Bacteria" may originate from phylum Bacteroidetes, but lack sufficiently similar representatives in the RDP database to allow for precise classification. The regression plugin also generates a geographic bar graph visualization in the main GenGIS window, which shows that individual F1 has the four lowest sample frequencies of Bacteroidia (Fig. 3c).

Heat maps offer an alternative way to visualize the relative abundance of taxa in different samples, and the "heatmap" function of the R statistical package generates two-dimensional heatmaps with clustering of rows and columns based on similarity. We developed a Python module, "Heatmap.py" (see Supplementary Material) which uses the Python environment within GenGIS to generate a heatmap by using the RPy2 libraries (<http://rpy.sourceforge.net>) to communicate with R. The resulting heatmap, generated for classes Clostridia, Bacteroidia, and "Unclassified Bacteria" (Fig. 4), shows the rough complementarity between the relative abundance of Bacteroidia and "Unclassified Bacteria" and a clustering similar but not identical to that displayed in Figure 2b.

Atlantic Reference Condition Model plugin

Reference condition approach (RCA) models have been developed to support bioassessment for river sites globally. Recently, Armanini et al. (2012) developed an approach for Atlantic Canada based on the River Invertebrate Prediction and Classification System (RIVPACS) methodology (Furse et al., 1981; Wright et al., 1984). The regional RIVPACS-based RCA model developed for Atlantic Canada uses biological monitoring data collected from wadeable streams paired with freely available, nationally consistent geospatial data layers with minimal anthropogenic influence. Biological classification of sites was carried out using a discriminant function (DF) approach and prediction of habitat group membership as described in Armanini et al. (2012). The model allows the derivation of ecological quality-ratio (O/E) data derived from the model. We implemented the Atlantic Canada RCA approach as a plugin for GenGIS that includes the trained reference model stored as a cross-platform R file object, which allows use of the plugin with RCA models from other regions. The RCA plugin uses the

abundance and taxon names from the sequence data and the location metadata that correspond to the parameters in the RCA model to generate various diversity measures. The O/E ratio for each diversity measure can be visualized as a bar graph for each location on the map, exported as location metadata for use with other GenGIS plugins (e.g. Linear Regression plugin), or output to a text file.

We applied the RCA plugin to 16 test sample sites from the Upper Mersey region in Nova Scotia, Canada (Fig 5A). The O/E Shannon diversity values calculated by the RCA plugin ranged from 0.563 to 1.136 (Fig. 5B) and were found to correlate with the measured Total Dissolved Oxygen Concentration (mg/L) from these same locations using the Linear Regression plugin ($r^2 = 0.431$, $p = 0.006$). As expected, community diversity was higher at those sites with well-oxygenated conditions, a property of healthy river habitats. By incorporating additional diagnostic tools based on relative abundance, it was possible to observe a stronger relationship between the environmental stressor and the biological endpoint. In fact, when compared to the results obtained by considering only O/E taxa richness, the amount of variance explained nearly doubled ($r^2 = 0.237$, $p = 0.047$; see Armanini et al., 2012 for details).

Conclusions

The above examples illustrate many of the new features of GenGIS. New phylogeographic tests allow for a more rigorous approach to the definition of axes. The implementation of an API for plugins allows users to develop their own tests either as Python / RPy2 scripts that can be executed in the console, or as a new graphical plugin that can easily be used by others. The RCA plugin demonstrates how sophisticated ecological modelling can now be carried out using the R functionality, and model results visualized within a plugin or directly on the geographical display. With the migration of the GenGIS source code to Github, we have made it easier for GenGIS users to make modifications to the software, particularly the development of new plugins, and contribute these changes to the GenGIS codebase.

Availability

GenGIS v2 is freely available under the GNU General Public License version 3.0. Executable binaries for Windows and Mac OS X can be obtained at <http://kiwi.cs.dal.ca/GenGIS>. Source code is hosted at Github (<https://github.com/beiko-lab/gengis>). The GenGIS website

contains an online manual, several written and video tutorials, and links to useful source for digital map data. MapMaker, software which provides maps compatible with GenGIS, can also be obtained on the website.

Acknowledgments

We thank Kathryn Dunphy, Conor Meehan, Brett O'Donnell, Kaila Lawton and Daniel Ruzzante for assistance with the development of GenGIS, and all the GenGIS users who have provided feedback. D.H.P. is supported by the Killam Trusts and Natural Sciences and Engineering Research Council of Canada; T.M., S. Z., and M.S.P. are supported by Genome Canada and the Ontario Genomics Institute; D. G. A. was supported by an Environment Canada contract; D. J. B. is supported by Environment Canada. M. G. I. L. is supported by the Canadian Institutes of Health Research. R.G.B. is supported by Genome Atlantic, the Canada Foundation for Innovation, and the Canada Research Chairs program. This project was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute through the Biomonitoring 2.0 project (OGI-050: see <http://biomonitoring2.org>) and grant number 2009-OGI-ABC-1405. The authors would like to thank all the users of the Canadian Aquatic Biomonitoring Network (CABIN) who kindly provided access to the data used for the RCA example described in this paper.

References

- Allal F, Sanou H, Millet L, Vaillant A, Camus-Kulandaivelu L, Logossa ZA, Lefèvre F, Bouvet J-M (2011) Past climate changes explain the phylogeography of *Vitellaria paradoxa* over Africa. *Heredity* 107: 174–186.
- Armanini DG, Monk WA, Carter L, Cote D, Baird DJ (2012) Towards generalised reference condition models for environmental assessment: a case study on rivers in Atlantic Canada. *Environ Monit Assess*, 1-13.
- Avisé JC, Arnold J, Ball RM Jr, Bermingham E, Lamb R, Neigel JE, Reed CA, Saunders NC (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annu Rev Ecol Syst* 18: 489–522.
- Baird DJ, Hajibabaei M (2012) Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol Ecol* 21: 2039-2044.

- Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R (2009) Bacterial community variation in human body habitats across space and time. *Science* 326: 1694-1697.
- de Berg M, van Kreveld M, Overmars M, Schwarzkopf O. (1997) *Computational geometry: algorithms and applications* (2nd edition). Springer-Verlag: Berlin, Germany.
- Farikou O, Thevenon S, Njiokou F, Allal F, Cuny G, Geiger A (2011) Genetic diversity and population structure of the secondary symbiont of tsetse flies, *Sodalis glossinidius*, in sleeping sickness foci in Cameroon. *PLoS Negl Trop Dis* 5: e1281.
- Furse MT, Wright JF, Armitage PD, Moss D (1981) An appraisal of pond-net samples for biological monitoring of lotic macroinvertebrates. *Water Res* 15: 679–689.
- Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One* 6: e17497.
- Hickerson MJ, Carstens BC, Cavender-Bares J, Crandall KA, Graham CH, Johnson JB, Rissler L, Victoriano PF, Yoder AD (2010) Phylogeography's past, present, and future: 10 years after Avise, 2000. *Mol Phylogenet Evol* 54: 291-301.
- Hujmans RJ, Guarino L, Cruz M, Rojas E (2011) Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. *Plant Genetic Resources Newsletter* 127: 15-19.
- Janies DA, Treseder T, Alexandrov B, Habib F, Chen JJ, Ferreira R, Çatalyürek Ü, Varón A, Wheeler WC (2011) The Supramap project: linking pathogen genomes with geography to fight emergent infectious diseases. *Cladistics* 27: 61-66.
- Kidd DM, Lio X (2008) Geophylobuilder 1.0: an ARCGIS extension for creating 'geophylogenies'. *Mol Ecol Res* 8: 88-91.
- Kidd DM (2010) Geophylogenies and the map of life. *Syst Biol* 59: 741-752.
- Laffan SW, Lubarsky E, Rosauer DF (2010) Biodiverse, a tool for the spatial analysis of biological and related diversity. *Ecography* 33: 643-647.
- Loo JH, Trejaut JA, Yen JC, Chen ZS, Lee CL, Lin M (2011) Genetic affinities between the Yami tribe people of Orchid Island and the Philippine Islanders of the Batanes archipelago. *BMC Genet* 12: 21.
- Maddison DR, Maddison WP (2008) Cartographer, a Mesquite package for plotting geographic data. <http://mesquiteproject.org/packages/cartographer>.

- Parks DH, Beiko RG (2009) Quantitative visualizations of hierarchically organized data in a geographic context. *17th International Conference on Geoinformatics*, Fairfax, VA.
- Parks DH, Porter M, Churcher S, Wang S, Blouin C, Whalley J, Brooks S, Beiko RG (2009a) GenGIS: A geospatial information system for genomic data. *Genome Res* 19: 1896-1904.
- Parks DH, MacDonald NJ, Beiko RG (2009b) Tracking the evolution and geographic spread of Influenza A. *PLoS Curr* 1: RRN1014.
- Poczai P, Hyvönen J, Symon DE (2011) Phylogeny of kangaroo apples (*Solanum* subg. *Archaeosolanum*, Solanaceae). *Mol Biol Rep* 38: 5243-5259.
- Tucker KP, Parsons R, Symonds EM, Breitbart M (2011) Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J* 5: 822-830.
- Ruzzante DE, Walde SJ, Macchi PJ, Alonso M, Barriga JP (2011) Phylogeography and phenotypic diversification in the Patagonian fish *Percichthys trucha*: the roles of Quaternary glacial cycles and natural selection. *Biol J Linn Soc* 103: 514-529.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75: 7537-41.
- Schoville SD, Roderick GK (2010) Evolutionary diversification of cryophilic *Grylloblatta* species (Grylloblattodea: Grylloblattidae) in alpine habitats of California. *BMC Evol Biol* 10: 163.
- Shafer A, White K, Côté S, Coltman D (2011) Deciphering translocations from relicts in Baranof Island mountain goats: is an endemic genetic lineage at risk? *Conserv Genet* 12: 1261-1268.
- Tsai YE (2011) PhyloGeoViz: a web-based program that visualizes genetic data on maps. *Mol Ecol Resour* 11: 557-561.
- Walker RS, Wichmann S, Mailund T, Atkisson CJ (2012) Cultural Phylogenetics of the Tupi Language Family in Lowland South America. *PLoS ONE* 7: e35025.
- Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73: 5261-5267.

Wright JF, Moss D, Armitage PD, Furse MT (1984) A preliminary classification of running-water sites in Great Britain based on macroinvertebrate species and the prediction of community type using environmental data. *Freshwater Biol* 14: 221–256.

Figure Legends

Figure 1. Phylogeography of kangaroo apples. a) A longitudinal gradient resulting in 23 crossings. Each of the eight species within the kangaroo apple phylogeny is assigned a unique color, and the two most substantial subclades are labelled. b) A latitudinal gradient results in 57 crossings. c) Results of a linear axes analysis on the kangaroo apple dataset. The number of crossings is only shown for axes between 90° and 270° as the graph has a period of 180° . Under the null model, only 10 of 10,000 permutations resulted in fewer than 34 crossings which is depicted by the red line (i.e. $\alpha = 0.001$). d) A linear axes analysis of the *Similia* subclade with the red line set to reflect a conservative critical value of $\alpha = 0.1$. e) A linear axes analysis of the *Avicularia/Laciniata* subclades ($\alpha = 0.1$).

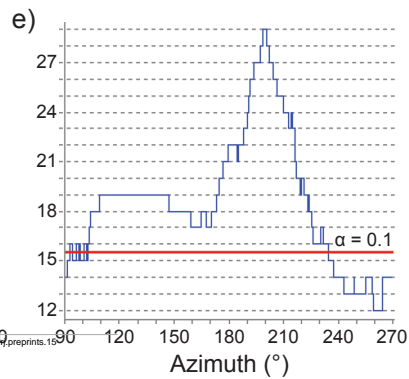
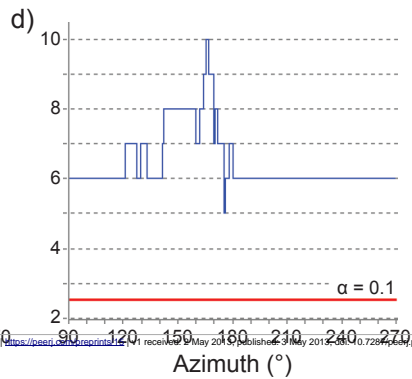
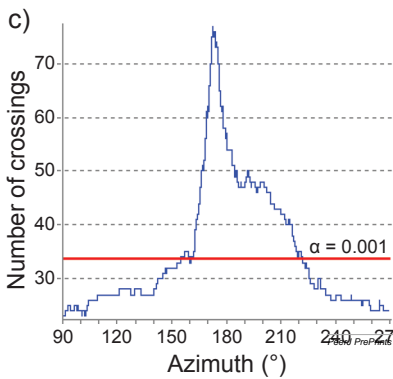
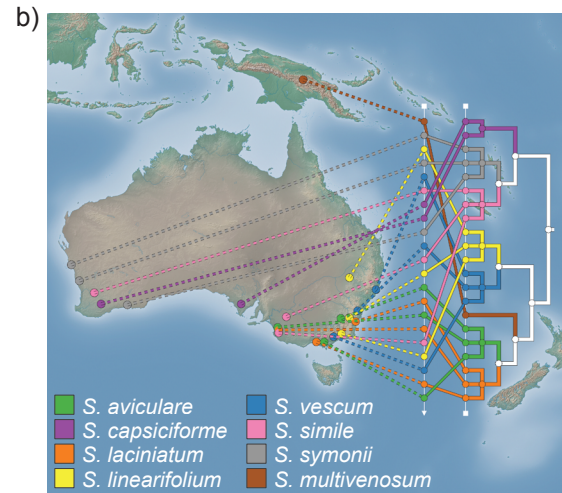
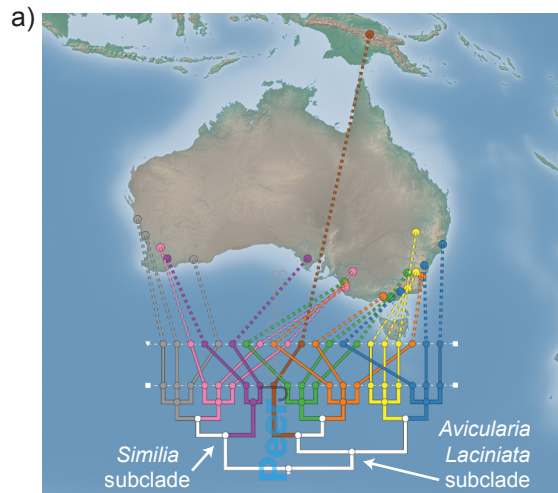
Figure 2. Biogeography of human-associated bacterial communities. a) Summary of taxonomic distribution across 28 different body sites. b) Detailed view of gut microbiota for six individuals (M1, M2, M3 = male; F1, F2, F3 = female). The third digit of each sample indicates time point 1 or 2 for a given individual, while the four-letter identifier indicates the sampling type (Fcsp = swab directly from stool sample, Fcsw = swab of toilet paper). The four samples for each individual are colored identically on the human body map. The UPGMA tree at right indicates the relative similarity of different samples; edges whose children all originate from a single individual are assigned the appropriate color, while black branches cover samples from more than one individual.

Figure 3. Linear regression analysis of class frequencies in fecal samples from six individuals. a) Linear regression plugin in GenGIS, showing data selection panels in the upper left, quantitative results display (e.g., p-value) in the lower left, and scatterplot on the right. b) Scatterplot saved from regression plugin, showing the relationship of frequencies of sequences labeled "Unclassified Bacteria" (y axis) with sequences assigned to Bacteroidia (x axis). The best fit regression line is shown. c) Geographic visualization of Bacteroidia frequencies as bar charts

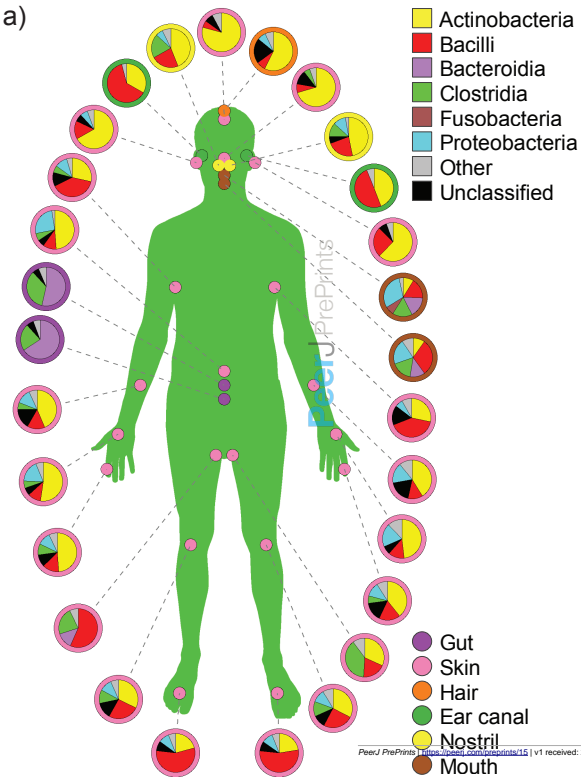
with a color gradient (blue = low frequency, white = intermediate frequency, red = high frequency). The arrangement of samples is identical to that in Figure 2b.

Figure 4. Heatmap of frequencies of three taxonomic groups (Bacteroidia, Clostridia, and "Unclassified Bacteria") from 24 fecal samples. Dark colors correspond to low frequencies, while yellow, tan and pink indicate high frequencies. Hierarchical clustering of samples and taxonomic groups are shown along both dimensions of the heatmap. Sample labels are explained in the legend of Figure 2b.

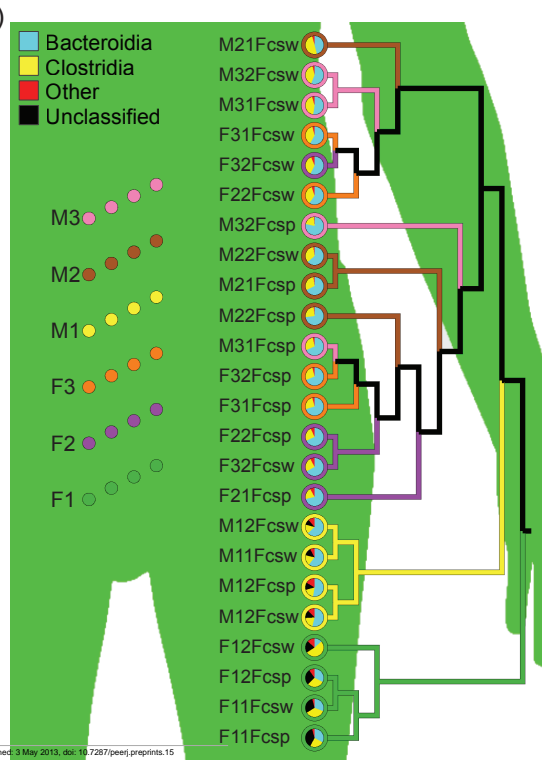
Figure 5. Reference condition analysis. A) The Atlantic RCA model was developed using 128 calibration sites from across the Atlantic Provinces, Canada (red circles) and was run on 16 test sites from the Upper Mersey region, Nova Scotia, Canada (green circles). B) Calculation of the O/E ratio using various diversity measures (table), and O/E for Shannon diversity visualized for each geographic location (bar plots). C) Linear regression of Shannon diversity O/E versus Dissolved Oxygen concentration for each location ($r^2 = 0.431$, $p = 0.006$).



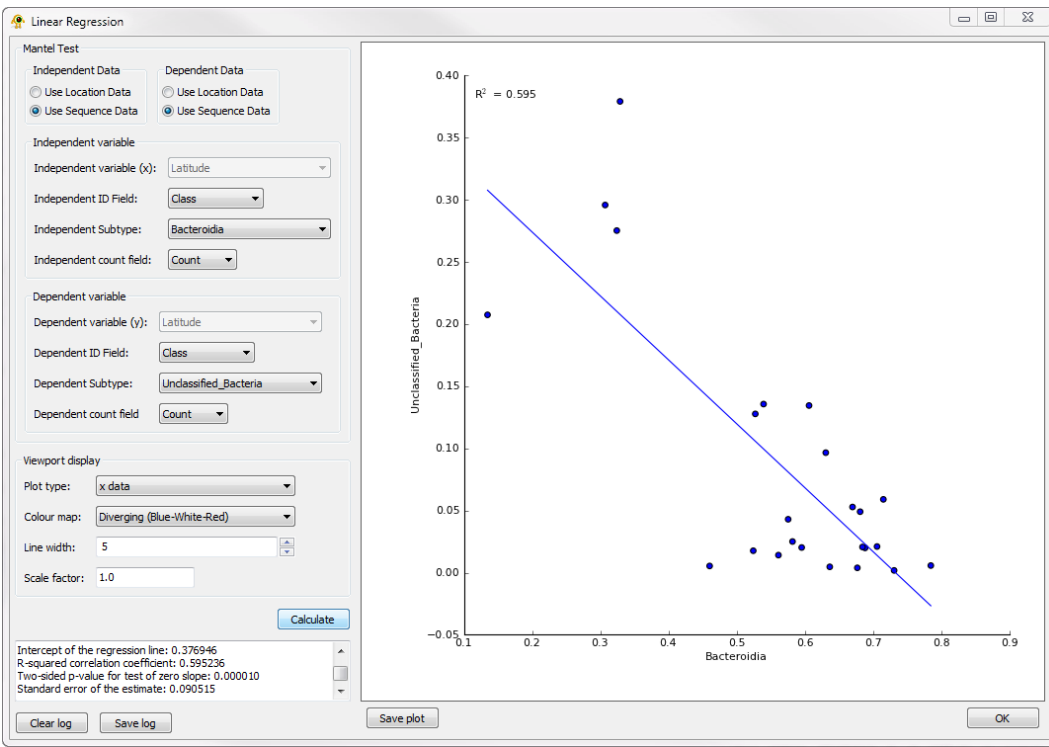
a)



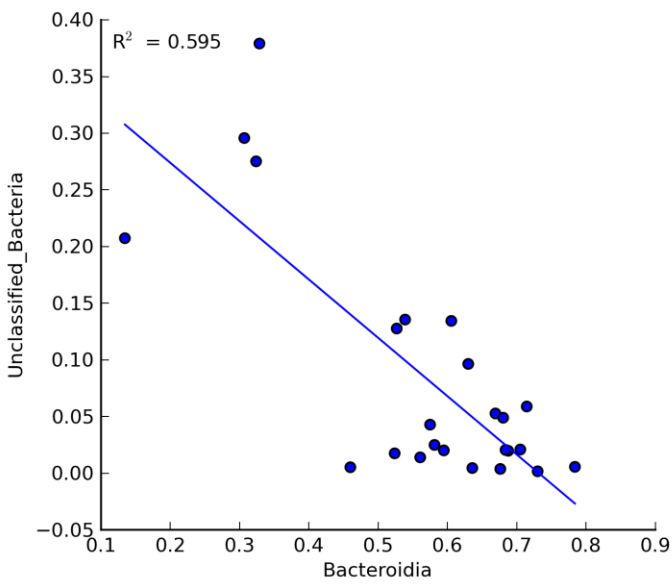
b)



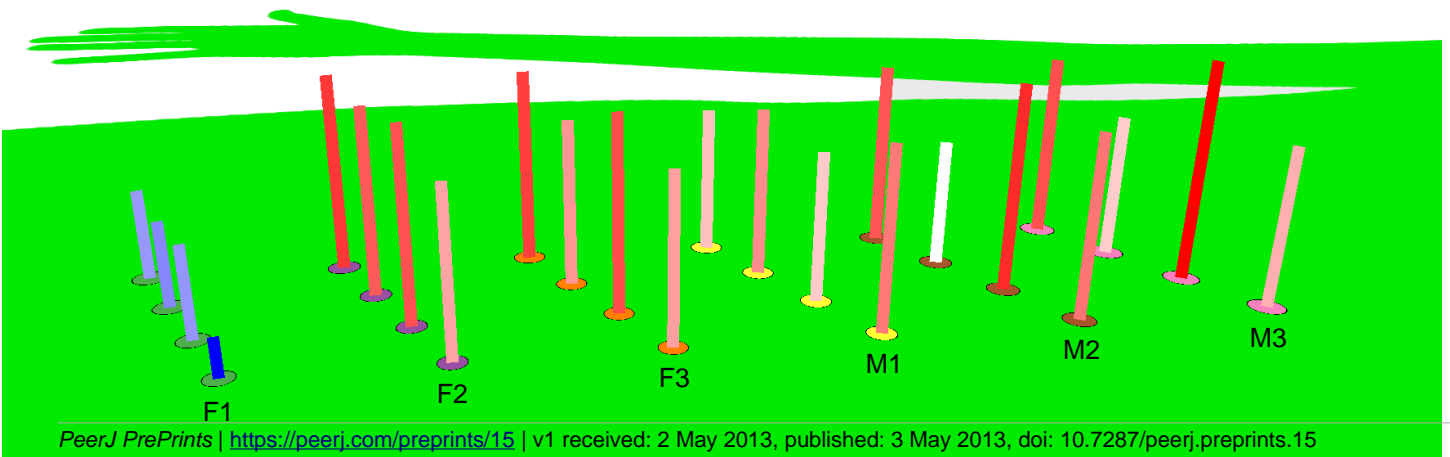
a)

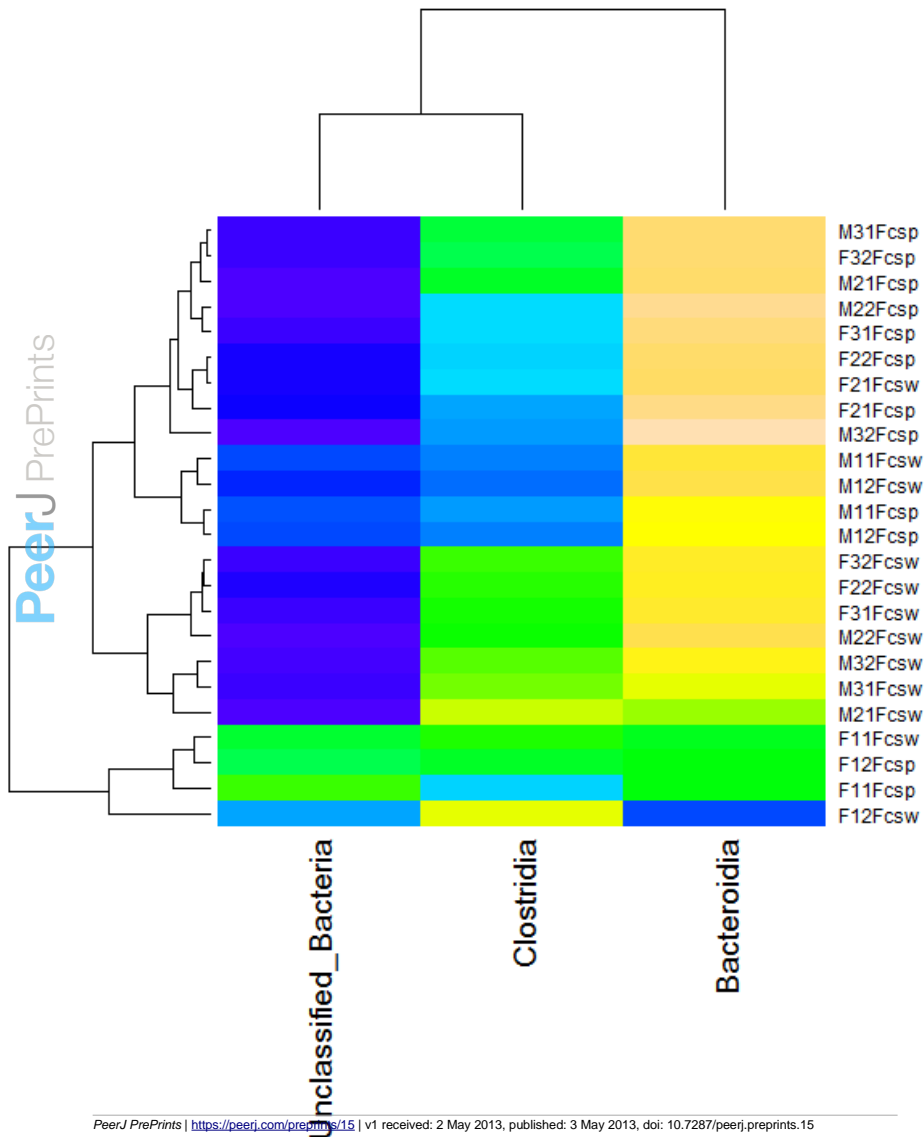


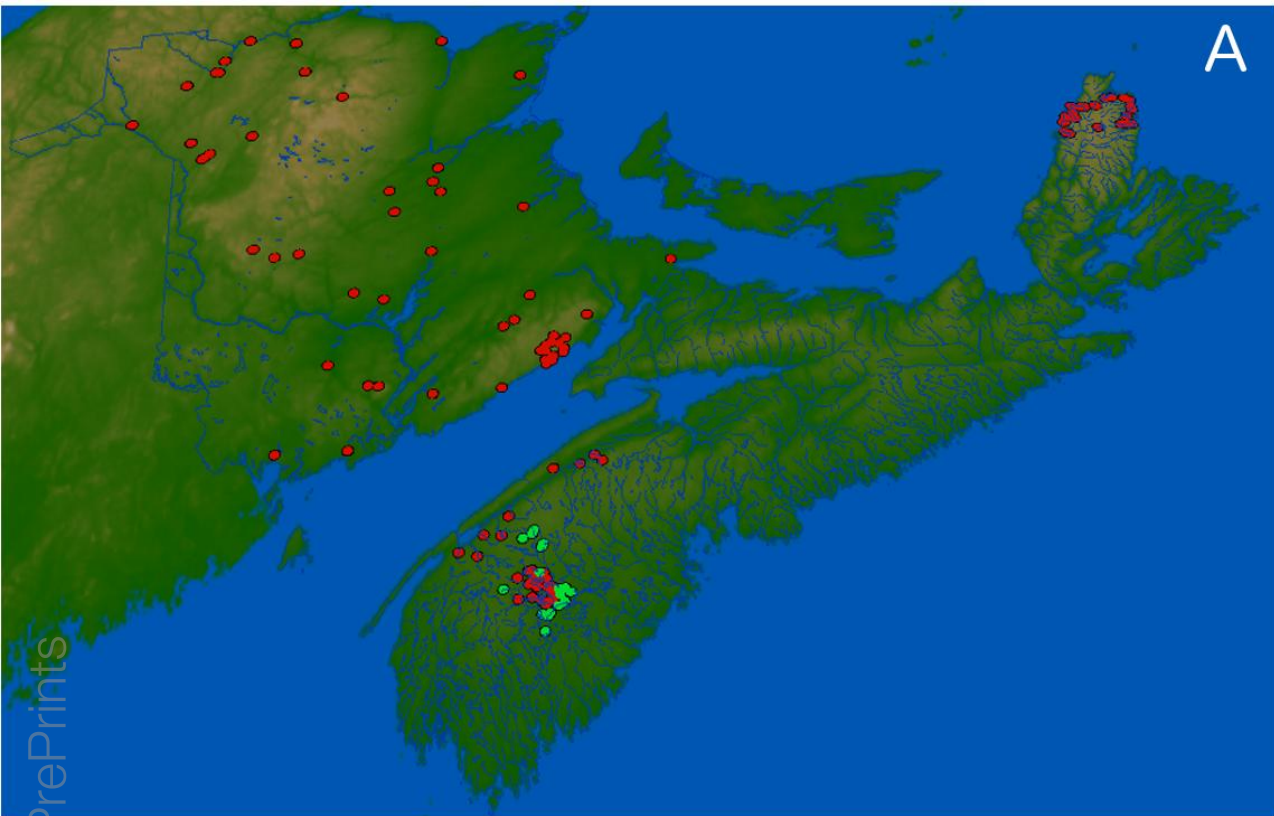
b)



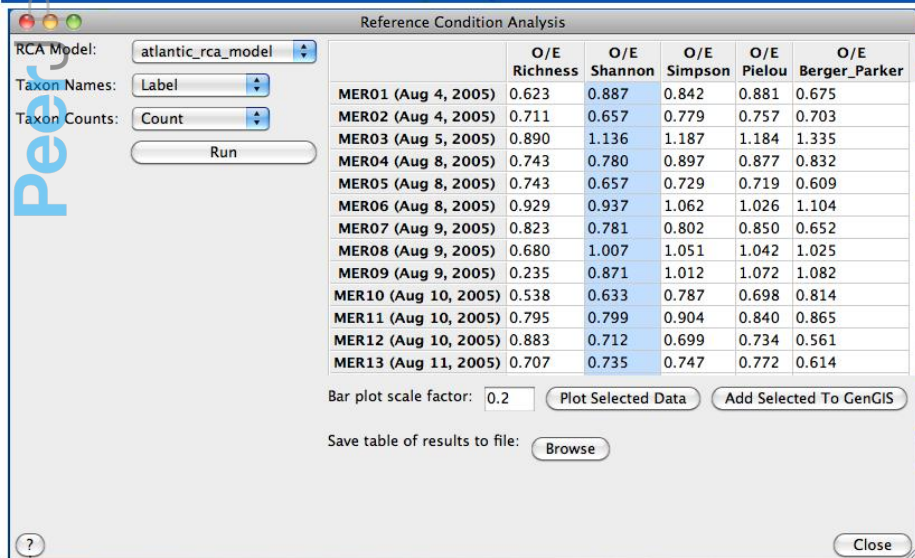
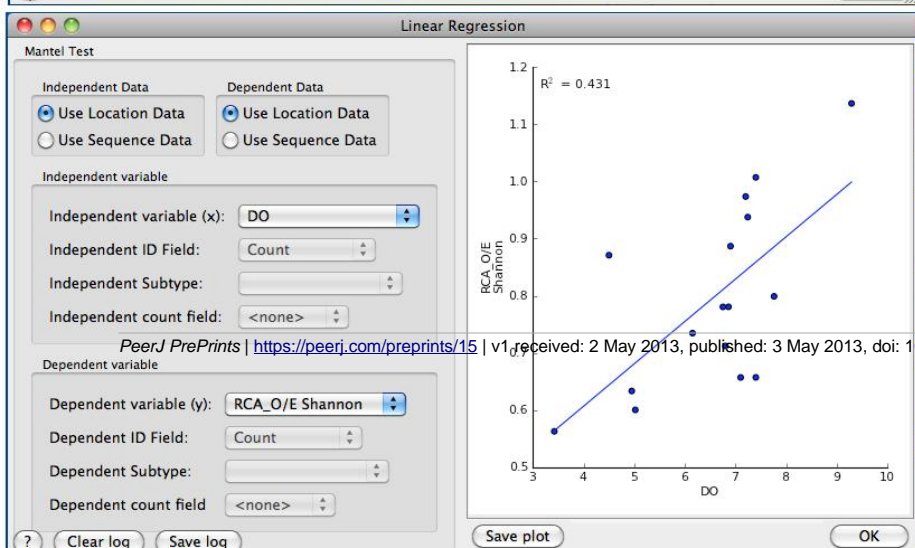
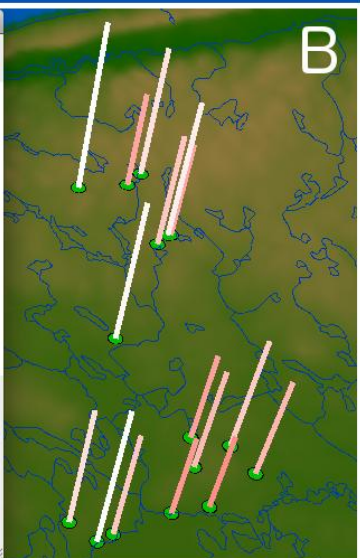
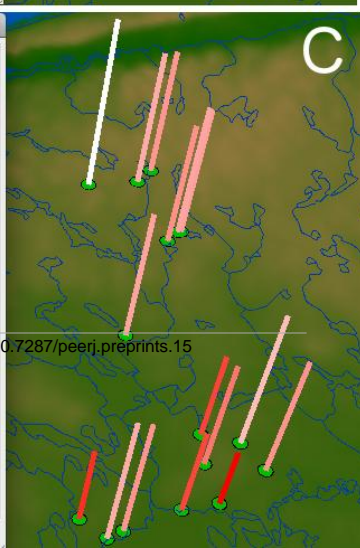
c)





A

PeerJ PrePrints

**B****C**

Supplementary Material

GenGIS 2: Geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework

Linear Axes Analysis Algorithm

Input: *sampleSites*, a vector indicating the *x* and *y* position of each sample site; *tree*, a tree where each leaf node is associated with a sample site
Require: *calculateSlope*(*rise*, *run*) which calculates a slope between [90°, 270°), *sort* (*vector*, *field*) which sorts elements of *vector* in ascending order of the specified *field*, *minNumberCrossings*(*sampleSites*, *tree*) which returns the minimum number of crossings for *tree* and a set of sample sites ordered according to the vector *sampleSites*, *swap*(*i*, *j*, *vector*) which swaps the the elements at positions *i* and *j* in *vector*
Return: array vector indicating the number of crossings for each permutation of the sample sites
Notational note: angles are measured using an azimuth where 90° is due east (i.e., standard compass directions)

procedure *LinearAxesAnalysis*(*sampleSites*, *tree*)

```
{ calculate slope for each pair of sample sites }
slopeInfoVector = [ ]
slopeInfoIndex = 1
for i = 1 to |sampleSites|
  for j = i+1 to |sampleSites|
    slopeInfo.slope = calculateSlope(sampleSite[i].y - sampleSite[j].y, sampleSite[i].x - sampleSite[j].x)
    slopeInfo.indexI = i
    slopeInfo.indexJ = j
    slopeInfoVector[slopeInfoIndex] = slopeInfo
    slopeInfoIndex = slopeInfoIndex + 1
  end for
end for

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

{ set initial ordering of sample sites based on their x-coordinate }
sort(sampleSites, x)
numCrossings = minNumberCrossings(sampleSites, tree)
results = [ ]
results[1].crossings = numCrossings
results[1].slope = 90

{ calculate number of crossings for each permutation of sample sites }
for i = 1 to |slopeInfoVector|
  swap(slopeInfoVector[i].indexI, slopeInfoVector[i].indexJ, sampleSites)
  numCrossings = minNumberCrossings(sampleSites, tree)
  results[i+1].crossings = numCrossings
  results[i+1].slope = slopeInfoVector[i].slope
end for

return results
```

DEGENERATE CASES FOR LINEAR AXES ANALYSIS ALGORITHM

There are 4 degenerate cases which must be considered when implementing the linear axes analysis algorithm (Figure S1):

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

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

- *Multiple pairs of points with identical slopes.* Multiple pairs of sample sites may result in lines with the same slope. Handling sets of sample sites which are collinear is described below, but care must be taken even for noncollinear sample sites resulting in the same slope. The *swap* function must be called for all *slopeInfoVector* elements with the same *slope* before calling *minNumberCrossings* and storing the results.
- *Collinear sample sites.* The degenerate case of multiple sample sites projecting to the same position along a gradient is only explicitly handled when sample sites have the same geographic location (see above). For collinear sample sites (or any pair of sample sites), we are only interested in the number of crossings which occur from an ϵ rotation in either direction. Let θ be the angle of an axis resulting in 3 or more sample sites being collinear (i.e., along a line with an angle of $\theta+90^\circ$). For an angle of $\theta - \epsilon$, all sample sites will be in the correct order. At $\theta + \epsilon$, the ordering of any set of collinear points needs to be reversed.

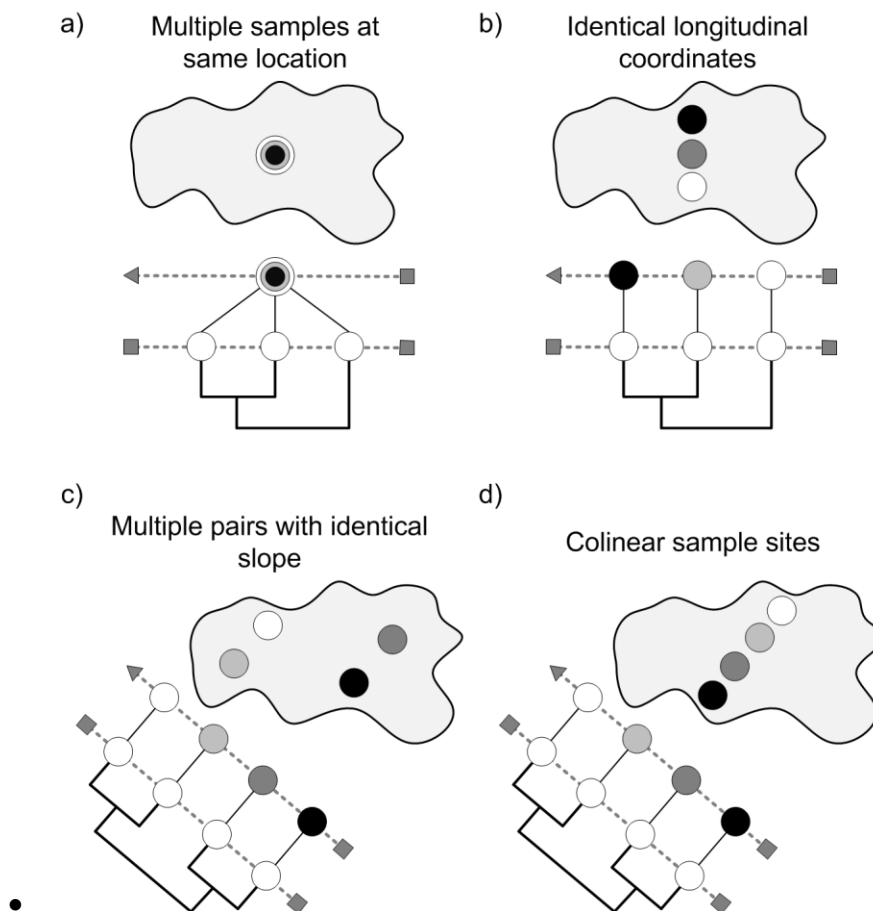


Figure S1. Degenerate cases for the Linear Axes Analysis algorithm. **(a)** Multiple samples may be taken from the same geographic locations. **(b)** Sample sites may have the same longitudinal coordinates. **(c)**

Multiple pairs of sample sites may have a projection line with the same slope. **(d)** Sample sites may be collinear. **(b-d)** In cases b-d, sample sites are laid out along the GLL in the order they would appear after a small clockwise rotation passed the degenerate angle.

SUPPLEMENTARY DATA FILES

FILE S1 – (Kangaroo-Apple.zip) – Data files for kangaroo apple analysis.

- *Australia.tif*: map of Australia and surrounding region.
- *Australia.tfw*: file describing geographic information for *Australia.tif*.
- *Kangaroo.samples.txt*: locations file for the 21 kangaroo apple samples.
- *Kangaroo.tre*: kangaroo apple phylogeny in Newick format.

FILE S2 (Costello-Body.zip) – Data files used in human body site analysis.

- *Body_locations_All.csv*: location file for all 28 body sites.
- *Body_locations_Fecal_only.csv*: location file for fecal samples, with locations separated by individual, time point, and sampling technique.
- *Costello_sequences_All.csv*: sequence file for all body sites, with samples aggregated by site.
- *Costello_sequences_Fecal.csv*: sequence file for fecal samples, separated by individual, time point, and sampling technique.
- *Costello_UPGMA_Fecal_only.tre*: Newick-formatted tree file showing the clustering relationships among all 24 fecal samples.
- *Heatmap.py*: Python script to compute R heatmap for selected taxonomic groups.
- *Human_body_silhouette_green.tif*: "map" file for body sites.