`mvMapper`: interactive spatial mapping of genetic structures

Julian R. Dupuis[1,2], Forest T. Bremer[1,2], Thibaut Jombart[3], Sheina B. Sim[1], Scott M. Geib[1,*]

[1]U.S. Department of Agriculture-Agricultural Research Service, Daniel K. Inouye U.S. Pacific Basin Agricultural Research Center, Hilo, HI 96720, USA

[2]Department of Plant and Environmental Protection Services, University of Hawai'i at Mānoa, Honolulu, HI 96822, USA

[3]MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College, London W2 1PG, UK

Keywords: population genetics, multivariate analyses, ordinations in reduced space, data visualization, Python, software

*Corresponding author: scott.geib@ars.usda.gov, (808) 959-4335, fax: (808) 959-5470, address: USDA-ARS DKI-PBARC, 64 Nowelo Street, Hilo, Hawaii, USA 96720

Running title: interactive maps of genetic structures

ABSTRACT

Characterizing genetic structure across geographic space is a fundamental challenge in population genetics. Multivariate statistical analyses are powerful tools for summarizing genetic variability, but geographic information and accompanying metadata is not always easily integrated into these methods in a user-friendly fashion. Here, we present a deployable Python-based web-tool, `mvMapper`, for visualizing and exploring results of multivariate analyses in geographic space. This tool can be used to map results of virtually any multivariate analysis of georeferenced data and routines for exporting results from a number of standard methods have been integrated in the R package `adegenet`, including principal components analysis (PCA), spatial PCA (sPCA), discriminant analysis of principal components (DAPC), principal coordinates analysis (PCoA), non-metric dimensional scaling (NMDS), and correspondence analysis (CA). `mvMapper`'s greatest strength is facilitating dynamic and interactive exploration of the statistical and geographic frameworks side-by-side, a task that is difficult and time-consuming with currently available tools. Source code and deployment instructions, as well as a link to a hosted instance of `mvMapper`, can be found at https://popphylotools.github.io/mvMapper/.

INTRODUCTION

Assessing patterns of genetic structure is one of the foundational challenges of population genetics (Pritchard *et al.* 2000; Slatkin 1987; Verity & Nichols 2016; Wright 1949), and characterizing this structure across geographic space is one of the first steps in most population genetic studies. Such contextualization of genetic structure allows in-depth evolutionary investigations, such as characterizing dispersal and invasion pathways (Genton *et al.* 2005; Janes *et al.* 2014; Mori *et al.* 2016), assessing and prioritizing conservation efforts (Austin *et al.* 2011; Proshek *et al.* 2015; Zenboudji *et al.* 2016), quantifying hybridization (Chatfield *et al.* 2010; Dupuis & Sperling 2016), and even utilizing genomic information to predict human origins (Das *et al.* 2016; Elhaik *et al.* 2014; Flegontov *et al.* 2016). Some analyses explicitly incorporate spatial information in the assessment of population structure (e.g. TESS: Caye *et al.* (2016), BAPS: Cheng *et al.* (2013), GENELAND: Guillot *et al.* (2005), EEMS: Petkova *et al.* (2016), SCAT: Wasser *et al.* (2004), sPCA: Jombart *et al.* (2008)), and landscape genetics is a fast growing field of statistics combining population genetics and landscape ecology (Manel & Holderegger 2013; Manel *et al.* 2003; Storfer *et al.* 2007).

Multivariate analyses stand out as powerful tools for summarizing genetic variability (Jombart *et al.* 2009). A wide diversity of such methods exist, each with their own particular applications (reviewed in Jombart *et al.* 2009). As a whole, these statistics provide many analytical advantages for population genetics, including, but not limited to: few overarching assumptions regarding the data (e.g. Hardy-Weinberg expectations and linkage equilibria, which can mask subtle clinal population structure (Jombart *et al.* 2008)), low computational requirements for the analysis of large datasets (e.g. thousands of markers and individuals

(Jombart & Ahmed 2011; Patterson *et al.* 2006)), and the statistical flexibility to address complex population genetic questions (Jombart *et al.* 2009 and references therein). While some methods explicitly incorporate geographic information (e.g. spatial principal components analysis (sPCA) (Jombart *et al.* 2008) and spatial correspondence analysis (Dray *et al.* 2008)) and provide valuable geographic context to population genetic data, non-spatial analyses also benefit from visualization in geographic space (Cavalli-Sforza *et al.* 1994; Wang *et al.* 2012). However, incorporating geographic context into multivariate analyses often requires the laborious comparison of ordination plots to maps of sampling localities, or technical expertise in map-making or geographic information systems (GIS) that may be beyond the comfort zone of the average researcher. While some streamlined tools exist for specific geographic visualizations (e.g. the Geography of Genetic Variants browser (Marcus & Novembre 2017)), generalized tools for straightforward visualization are lacking.

Here, we present a tool for the visualization and exploration of multivariate analyses in geographic space. `mvMapper` is a Python-based, deployable web-based tool that can process outputs of virtually any multivariate analysis as well as sample locality information and allows users to interactively explore the statistical framework of the multivariate analysis in both ordination and geographical space (Figure 1). The input format is a simple comma-delimited tabular file that can either be assembled manually, or generated using `mvMapper`'s input generation function in the `adegenet` library (Jombart 2008) in R (R Core Team 2016), giving access to a wide range of commonly used methods.

METHODS

Implementation

`mvMapper` is implemented in Python v3.6 (Python Software Foundation 2017), and makes extensive use of the following libraries: Bokeh v0.12.4 for data visualization (Bokeh Development Team 2014), Pandas v0.19.2 for data structure and analysis (McKinney 2010), colorcet v0.9.1 for color utilities (Kovesi 2015), and pyproj v1.9.5.1 (Whitaker 2016), a python interface for cartographic transformations using PROJ.4 (Warmerdam 2001). Map tiles and map data are by Stamen Design under CC BY 3.0 (Stamen Design 2017) and OpenStreetMap under CC BY SA (OpenStreetMap contributors 2017), respectively, and use the WGS84 (ESPG 4326) spatial reference system. The automated data preparation script is implemented in the `adegenet` library (Jombart 2008) in R (R Core Team 2016). Links to `mvMapper`'s source code, documentation, a ready to deploy Docker container (Merkel 2014, see https://www.docker.com/), and a hosted instance of the web application can be found on our project page at https://popphylotools.github.io/mvMapper/. Although deploying a stand-alone instance of `mvMapper` provides a great deal of flexibility through the customization of the configuration file (default displayed statistical parameters, dataset, etc.), here, we generally refer to the default configuration available on our hosted instance. All modern desktop web browsers support `mvMapper`.

Data input

The primary input for `mvMapper` is a comma-delimited tabular file that contains individuals in rows and information about those individuals in columns. A typical file contains columns such as: specimen identification code (we refer to this unique identifier as `key`),

collection locality information (latitude and longitude, or `lat` and `lon`), a population identifier, results of the multivariate analysis (specimen coordinates across multiple dimensions of an analysis, e.g. principal components), and any other metadata related to the specimens (sex, host, morphological characteristics, etc.). Given that many of these analyses are conducted in R (R Core Team 2016), we have incorporated a data preparation function to the widely used R library `adegenet` (Jombart 2008). This function, `export_to_mvmapper`, combines an active R object from a multivariate analysis with locality information for each specimen. Currently, multivariate analyses conducted in `adegenet` and those based on the duality diagram (`dudi.*` functions) in `ade4` (Dray & Dufour 2007) are supported, including: sPCA and discriminant analysis of principal components (DAPC: Jombart *et al.* (2010)) in `adegenet`, and principal components analysis (PCA), principal coordinates analysis (PCoA), non-metric dimensional scaling (NMDS), correspondence analysis (CA), and others in `ade4`. Locality information is then incorporated into the multivariate analysis through another R object. This is most easily done by preparing an additional file with at least three columns, `key`, `lat`, and `lon`, where `key` matches the unique individual identifiers used in the multivariate analysis. After reading this locality file into R, `export_to_mvmapper` will combine the two R objects (the multivariate analysis and the locality information) into `mvMapper` input format, which can be manually written to a comma-delimited file (e.g. using R's `write.csv` function). Locality information can be incorporated via other means (e.g. when latitude and longitude are already part of a `genind` object), however the advantage of creating an additional file, as described here, is that any additional specimen-based information can be included in that file (named `localities.csv` in the following example), such as: specimen sex, host information, and

morphological or ecological characters. Alternatively, rather than using `export_to_mvmapper`, the input data file can be generated manually from results of multivariate analyses in different programs or R libraries, as the tabular format is general and user-friendly.

Below we provide an example of data preparation from a DAPC, which in addition to standard multivariate analyses results (distribution of individuals along principal components) provides additional components recognized by `mvMapper`, such as membership to *a priori*-assigned and DAPC-assigned groups, and the posterior probabilities of the DAPC-assigned groups. See https://github.com/popphylotools/mvMapper/tree/master/dataPrepExampleFiles for an example of this file generated from a dataset of 783 autosomal microsatellite loci genotyped for 1,048 human individuals from 53 populations (Rosenberg *et al.* 2005).

```
> # An example using the microsatellite dataset of Rosenberg et al. 2005
> # Using adegenet devel version
> # Reading input file
> Rosenberg <- read.structure("Rosenberg_783msats.str", n.ind=1048,
n.loc=783,  onerowperind=F, col.lab=1, col.pop=2, row.marknames=NULL,
NA.char="-9", ask=F, quiet=F)

> # DAPC (n.pca determined using xvalDapc, see ??xvalDapc)
> dapc1 <- dapc(Rosenberg, n.pca=20, n.da=200)

> # read in localities.csv, which contains "key", "lat", and "lon" columns
with column headers (this example contains a fourth column "population" which
is a text-based population name based on geography)
> localities <- read.csv(file="localities.csv", header=T)

> # generate mvmapper input file and write to "rosenbergData.csv"
> out <- export_to_webapp(dapc1,localities)
> write.csv(out, "rosenbergData.csv", row.names=F)
```

By default, `mvMapper` is configured to display the microsatellite dataset of Rosenberg *et al.* (2005) from the example above. Users can upload their own datasets through the upload tab linked in the navigation bar at the top of the page (Figure 1, top). Files uploaded in this manner

158 are named using an alphanumeric random string that is integrated into the web address used to
159 select that dataset; users can return to a previously uploaded dataset using its unique web address
160 until it expires after 14 days.
161
162 *Interface and functionality*
163 The main interface of `mvMapper` consists of three components: a statistical panel, a
164 mapping panel, and a metadata panel (Figure 1). Aspects of these panels are linked, so that, for
165 example, selecting individuals in the ordination of the statistical panel will highlight those
166 individuals on the map and their metadata will appear in the metadata panel. Pull-down menus to
167 the left of the statistical panel allows users to select which data is displayed in the ordination
168 plot. In a general multivariate analysis, the most informative principal axes (or principal
169 components) would be plotted against each other (e.g. PC1 vs. PC2) (Figure 1); in `mvMapper`,
170 any of the multivariate analysis results (all principal axes) or specimen-based metadata can be
171 plotted in the statistical panel. For example, the distribution of individuals along a particular
172 principal component can be plotted against populations of origin (Figure 2A), assigned group
173 membership from DAPC, or latitude or longitude (Figure 2B). Individual specimen points in
174 both the statistical and mapping panels can be colored (with several palette choices) or sized
175 according to any column in the input data file, except when discrete data values outnumber
176 available colors/sizes, in which case those attributes are excluded from the dropdowns.
177 Automatic binning supports coloring and sizing of numeric attributes. Specific attributes can be
178 configured to be treated as discrete values, even if numeric, and by default these include `key`,
179 `grp`, and `assigned_grp`. These coloring and sizing abilities facilitate rapid exploration of

metadata with regard to population structure; for example, individuals can be colored by collection locality, group membership, host, sex, or other genetic attribute (Figure 2C), or be sized by the posterior probability of group membership in a DAPC, all with a few mouse clicks.

Both the statistical and mapping panels are interactive with tools for panning, zooming in and out, and saving the image. Individuals can be selected singly with a mouse click, or multiply by shift clicking or using the dragged box tool. In the mapping panel, overlapping points can be separated with a jitter function, and the zoom tool is dynamic: zooming in or out will access finer-scale or coarser-scale map tiles with more or less detail, respectively (e.g. labeling countries, cities, roads, or other scale-appropriate geographical features). This allows `mvMapper` to function at both global and local geographic scales (Figure 2C). Selecting individuals in either the statistical or mapping panel displays their metadata in the lower panel, which can be sorted by clicking on column headers. Selected data can also be downloaded (as a comma-delimited file) to facilitate downstream analysis, for example re-analysis of individual population groups or hierarchical analysis (Vähä *et al.* 2007).

DISCUSSION

Visualizing population structure across geographic space is fundamental to most population genetic studies. However, combining multiple "data wrangling" tools (Kandel *et al.* 2011), including population genetic data processing, multivariate analysis, and particularly map-making or GIS, is a time-consuming, error-prone, and generally daunting task (e.g. Fletcher-Lartey & Caprarelli 2016; Rickles & Ellul 2014; Sipe & Dale 2003). `mvMapper` greatly facilitates this process by providing an accessible, open access, user-friendly interface for

exploring and visualizing results of multivariate analysis in geographic space, and perhaps most importantly facilitates dynamic and interactive exploration of these spaces. Interactivity, in particular, is key to enable users to quickly assess the geographic patterns of any combinations of principal components, population groupings, additional statistical parameters (assignments to groups based on discriminant functions in DAPC or lag-vectors of principal components in sPCA), and any other specimen-based metadata with a few mouse clicks in the drop-down menus to the left of the statistical panel. Given these characteristics, we envision `mvMapper` to be of wide interest to a broad range of researchers as well as for teaching and training purposes. Additionally, `mvMapper`'s highly generalized and modular approach allows it to be modified for more specific uses; for example, including metadata corresponding to whether specimens of an invasive species were collected in its native versus introduced range allows `mvMapper` to become a tool for source determination of intercepted material (Roderick 2004).

ACKNOWLEDGEMENTS

Funding for this project was provided by United States Department of Agriculture (USDA) Agricultural Research Service (ARS) and Animal and Plant Health Inspection Service (APHIS) Farm Bill Section 10007 projects "Diagnostic Resources to Support Fruit Fly Exclusion and Eradication, 2012-2014" and "Genomic approaches to fruit fly exclusion and pathway analysis, 2015-2016" to USDA-ARS, USDA-APHIS, and University of Hawai'i at Mānoa (projects 3.0251.02 and 3.01251.03 (FY 2014), 3.0256.01 and 3.0256.02 (FY 2015), and 3.0392.02 and 3.0392.03 (FY 2016)). TJ is funded by the Medical Research Council Centre for Outbreak Analysis and Modelling and the National Institute for Health Research - Health Protection

Research Unit for Modelling Methodology. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal opportunity employer.

LITERATURE CITED

Austin J. D., Jelks H. L., Tate B., Johnson A. R., Jordan F. (2011). Population genetic structure and conservation genetics of threatened Okaloosa darters (*Etheostoma okaloosae*). *Conservation Genetics*, 12, 981-989.

Bokeh Development Team (2014) *Bokeh: Python library for interactive visualization.* http://www.bokeh.pydata.org/

Cavalli-Sforza L. L., Menozzi P., Piazza A. (1994) *The History and Geography of Human Genes* Princeton University Press, Princeton, NJ.

Caye K., Deist T. M., Martins H., Michel O., Francois O. (2016). TESS3: fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources*, 16, 540-548.

Chatfield M. W., Kozak K. H., Fitzpatrick B. M., Tucker P. K. (2010). Patterns of differential introgression in a salamander hybrid zone: inferences from genetic data and ecological niche modelling. *Molecular Ecology*, 19, 4265-4282.

Cheng L., Connor T. R., Siren J., Aanensen D. M., Corander J. (2013). Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution*, 30, 1224-1228.

Das R., Wexler P., Pirooznia M., Elhaik E. (2016). Localizing Ashkenazic Jews to primeval villages in the ancient Iranian lands of Ashkenaz. *Genome Biology and Evolution*, 8, 1132-1149.

Dray S., Dufour A.-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22, 1-20.

Dray S., Saïd S., Débias F. (2008). Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science*, 19, 45-56.

Dupuis J. R., Sperling F. A. H. (2016). Hybrid dynamics in a species group of swallowtail butterflies. *Journal of Evolutionary Biology*, 29, 1932-1951.

Elhaik E., Tatarinova T., Chebotarev D.*, et al.* (2014). Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nature Communications*, 5, 3513.

Flegontov P., Kassian A., Thomas M. G.*, et al.* (2016). Pitfalls of the Geographic Population Structure (GPS) Approach Applied to Human Genetic History: A Case Study of Ashkenazi Jews. *Genome Biology and Evolution*, 8, 2259-2265.

Fletcher-Lartey S. M., Caprarelli G. (2016). Application of GIS technology in public health: successes and challenges. *Parasitology*, 143, 401-415.

Genton B. J., Shykoff J. A., Giraud T. (2005). High genetic diversity in French invasive populations of common ragweed, *Ambrosia artemisiifolia*, as a result of multiple sources of introduction. *Molecular Ecology*, 14, 4275-4285.

Guillot G., Mortier F., Estoup A. (2005). Geneland: a computer package for landscape genetics. *Molecular Ecology Notes*, 5, 712-715.

Janes J. K., Li Y., Keeling C. I.*, et al.* (2014). How the mountain pine beetle (*Dendroctonus ponderosae*) breached the Canadian Rocky Mountains. *Molecular Biology and Evolution*, 31, 1803-1815.

Jombart T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24, 1403-1405.

Jombart T., Ahmed I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27, 3070-3071.

Jombart T., Devillard S., Balloux F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11, 94.

Jombart T., Devillard S., Dufour A. B., Pontier D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity (Edinb)*, 101, 92-103.

Jombart T., Pontier D., Dufour A. B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity (Edinb)*, 102, 330-341.

Kandel S., Heer J., Plaisant C.*, et al.* (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10, 271-288.

Kovesi P. (2015). Good colour maps: how to design them. *CoRR*, abs/1509.03700.

Manel S., Holderegger R. (2013). Ten years of landscape genetics. *Trends in Ecology & Evolution*, 28, 614-621.

Manel S., Schwartz M. K., Luikart G., Taberlet P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, 18, 189-197.

Marcus J. H., Novembre J. (2017). Visualizing the geography of genetic variants. *Bioinformatics*, 33, 594-595.

McKinney W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51-56.

Merkel D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 239, 2.

Mori B. A., Davis C. S., Evenden M. L. (2016). Genetic diversity and population structure identify the potential source of the invasive red clover casebearer moth, *Coleophora deauratella*, in North America. *Biological Invasions*, 18, 3595-3609.

OpenStreetMap contributors (2017) *Planet dump retrieved from* https://planet.osm.org/. https://www.openstreetmap.org/

Patterson N., Price A. L., Reich D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2, e190.

Petkova D., Novembre J., Stephens M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, 48, 94-100.

Pritchard J. K., Stephens M., Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945-959.

Proshek B., Dupuis J. R., Engberg A., *et al.* (2015). Genetic evaluation of the evolutionary distinctness of a federally endangered butterfly, Lange's Metalmark. *BMC Evolutionary Biology*, 15, 73.
Python Software Foundation (2017) *Python Language Reference, version 3.6.* http://www.python.org/
R Core Team (2016). R: a language and environment for statistical computing. *R Foundation for Statistical Computing, VIenna, Austria. URL* https://www.r-project.org/.
Rickles P., Ellul C. (2014). A Preliminary Investigation into the Challenges of Learning GIS in Interdisciplinary Research. *Journal of Geography in Higher Education*, 39, 226-236.
Roderick G. K. (2004) Tracing the origin of pests and natural enemies: genetic and statistical approaches. In: *Genetics, Evolution and Biological Control* (eds. Ehler L. E., Sforza R., Mateille T.). CABI Publishing, New York, NY.
Rosenberg N. A., Mahajan S., Ramachandran S., *et al.* (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, 1, e70.
Sipe N. G., Dale P. (2003). Challenges in using geographic information systems (GIS) to understand and control malaria in Indonesia. *Malaria Journal*, 2, 36.
Slatkin M. (1987). Gene flow and the geographic structure of natural populations. *Science*, 236, 787-792.
Stamen Design (2017) *Map tiles by Stamen Design.* http://maps.stamen.com/
Storfer A., Murphy M. A., Evans J. S., *et al.* (2007). Putting the "landscape" in landscape genetics. *Heredity (Edinb)*, 98, 128-142.
Vähä J. P., Erkinaro J., Niemela E., Primmer C. R. (2007). Life-history and habitat features influence the within-river genetic structure of Atlantic salmon. *Molecular Ecology*, 16, 2638-2654.
Verity R., Nichols R. A. (2016). Estimating the Number of Subpopulations (K) in Structured Populations. *Genetics*, 203, 1827-1839.
Wang C., Zollner S., Rosenberg N. A. (2012). A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genetics*, 8, e1002886.
Warmerdam F. (2001) *PROJ.4.* http://proj4.org/index.html
Wasser S. K., Shedlock A. M., Comstock K., *et al.* (2004). Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proceedings of the National Academy of Sciences*, 101, 14847-14852.
Whitaker J. (2016) *pyproj, GitHub repository.* https://github.com/jswhit/pyproj
Wright S. I. (1949). The genetic sturcture of populations. *Annals of Eugenics*, 15, 323-354.
Zenboudji S., Cheylan M., Arnal V., *et al.* (2016). Conservation of the endangered Mediterranean tortoise *Testudo hermanni hermanni*: The contribution of population genetics and historical demography. *Biological Conservation*, 195, 279-291.

DATA ACCESIBILITY

Software, documentation, and example data are available at https://popphylotools.github.io/mvMapper/, and a stable release at the time of publishing is available at http://zenodo.org DOI: XXXXXXX.

AUTHOR CONTRIBUTIONS

JRD, FTB, SSB, and SMG conceptualized software; JRD, FTB, and TJ implemented software. JRD wrote the manuscript with input from all authors.

FIGURE LEGENDS

Figure 1. The user interface of `mvMapper` in a web browser, displaying the human microsatellite dataset of Rosenberg *et al.* (2005). Features include the statistical panel (left), mapping panel (right), metadata panel (lower), and navigation bar (top).

Figure 2. Various visualization options for the human and swallowtail butterfly microsatellite datasets of Rosenberg *et al.* (2005) (A and B) and Dupuis and Sperling (2016) (C), respectively. A) population grouping vs. principal component 3, B) latitude vs. principal component 3, and C) principal component 2 vs. 1, colored by COI clade and zoomed in to the Red Deer River valley in southeast Alberta, Canada.