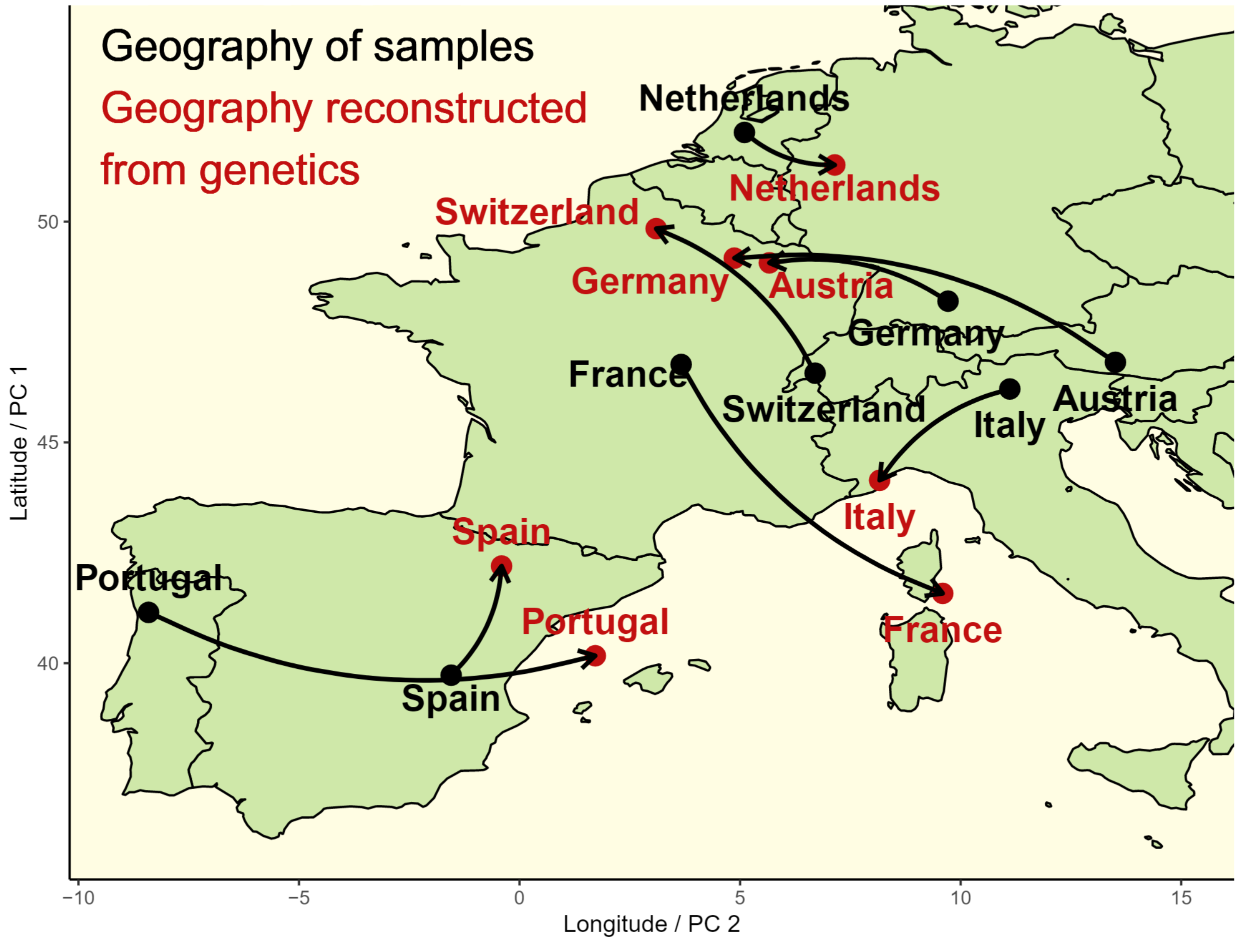
David Bass and Medha Prakash
Professor Noah Gade
STAT 3280: Data Visualization and Management
29 November 2022

Reconstructing fruit fly geography from genetics

Data visualization

Reconstructing *D. melanogaster* geography from genetics via PCA and Procrustes analysis



Technique description

We worked geospatial data describing the genetic variation in biological samples of *Drosophila melanogaster* fruit flies from the Drosophila Evolution over Space and Time (DEST) dataset (Kapun et al. 2021, https://dest.bio/). Our visualization technique shows the geographic origins of *D. melanogaster* samples as black points and the results of an attempt to reconstruct geography solely from genetic variation information as red points, with arrows showing the mapping from physical space to genetic space for each pair of points corresponding to each sample. Inspired by the common method in population genetics of examining how the dominant eigenvectors from principal component analysis (PCA) on genetic variation data (e.g. Novembre et al. 2008, Kapun et al. 2021) describe geography, we used R (R Core Team 2022) and the tidyverse (Wickham et al. 2019) to apply PCA (Lê et al. 2008) to genetic variation information to *D. melanogaster* samples from Western Europe. We found that the two most dominant eigenvectors, PC 1 and PC 2, were strongly correlated with latitude and longitude, respectively, and then applied Procrustes analysis (Oksanen et al. 2022) to align geographic space with the space spanned by PC 1 and PC 2 under reasonable restrictions. Considering the widespread usage of PCA on genetic variation data in the literature, we expect our technique, which adds an additional step of statistical analysis and a novel visual representation of the difference between geographic and PC-spaces, to be generalizable to any dataset describing genetic variation in a geographically dispersed population (e.g. Sudlow et al. 2015).

Figure 7 of Kapun et al. (2020), contains a map overlaid with points showing the geographic origins of biological samples that are colored based on the points' locations in PC-space, resulting in a crowded color scheme in which the exact values of points along PCs are obscured. We resolve this issue by showing locations in PC-space as independent points in a single distinct color that are connected via arrows to the points representing samples' geographic origins. We recognize that this aspect of our technique only contributes to clarity when working with few samples; drawing arrows between many pairs of points would quickly crowd the figure as sample size grows. Figure 1 of Novembre et al. (2008) shows points in a transformed PC-space that is rotated to optimally resemble a map of Europe, with geography represented by country-specific point color and type. Circles represent median values for each country. With the actual map of Europe only shown in miniature beside the PC-space plot, it is difficult to conceptualize the variation between geography and PC-space. We resolve this issue both by showing a map in the background of our visualization and by representing the distances between pairs of mapped points directly with arrows. We avoid the excess of points present in either cited plot by only showing the mean location of all samples for each country.

In summary, our technique addresses existing shortcomings in the representation of PCA results on genetic variation data by **(1)** applying Procrustes analysis to align PC-space to geographic space, **(2)** plotting points for both geographic space and PC-space, **(3)** elucidating the transformation from geographic space to PC-space with arrows connecting pairs of corresponding points, **(4)** grouping the data so as to only display one point per country in either space. A shortcoming with our technique is the need to manually adjust each text label in order to avoid overlapping elements. Our code and a PDF version of our visualization can be found at our GitHub repository (https://github.com/DavidB256/Reconstructing-fruit-fly-geography-from-genetics).

Data description

We applied our visualization technique to a subset of the Drosophila Evolution over Space and Time (DEST) dataset, which contains information about genetic variation in more than 13,000 *D. melanogaster* fruit flies obtained from 272 samples taken from four continents over several years. While the DEST dataset contains myriad information about the environment, season, date, sex distribution, contamination, sequencing technique, chromosomal inversion status, locality, and collector of each sample, our technique only considers the latitude, longitude, country of origin, and genetic variation of samples. We plot the first three variables directly in black points and apply PCA and Procrustes analysis to the genetic variation information in order to calculate the coordinates of the red points. While values along PCs have no real-world meaning, our application of Procrustes analysis transforms PCs to best conform with latitude and longitude, endowing the coordinates of our points PC-space with standard geographic significance.

For the purpose of presenting a simple proof-of-concept with our visual, we chose to use only samples in DEST Western European (namely Portugal, Spain, France, Italy, Germany, Austria, and the Netherlands). We did not consider the samples' localities, instead choosing to collapse all data for each country into one observation. The curvature of the arrows in our visualization has no significance.

Works cited

Kapun, M., Nunez, J. C. B., Bogaerts-Márquez, M., Murga-Moreno, J., Paris, M., Outten, J., Coronado-Zamora, M., Tern, C., Rota-Stabelli, O., Guerreiro, M. P. G., Casillas, S., Orengo, D. J., Puerma, E., Kankare, M., Ometto, L., Loeschcke, V., Onder, B. S., Abbott, J. K., Schaeffer, S. W., … Bergland, A. O. (2021). Drosophila Evolution over Space and Time (DEST): A New Population Genomics Resource. *Molecular Biology and Evolution*, *38*(12), 5782–5805. https://doi.org/10.1093/molbev/msab259

Lê S, Josse J, Husson F (2008). "FactoMineR: A Package for Multivariate Analysis." *Journal of Statistical Software*, **25**(1), 1–18. doi:10.18637/jss.v025.i01

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), Article 7218. https://doi.org/10.1038/nature07331

Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P, Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette M, Ribeiro Cunha E, Smith T, Stier A, Ter Braak C, Weedon J (2022). *vegan: Community Ecology Package*. R package version 2.6-4, https://CRAN.R-project.org/package=vegan

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686