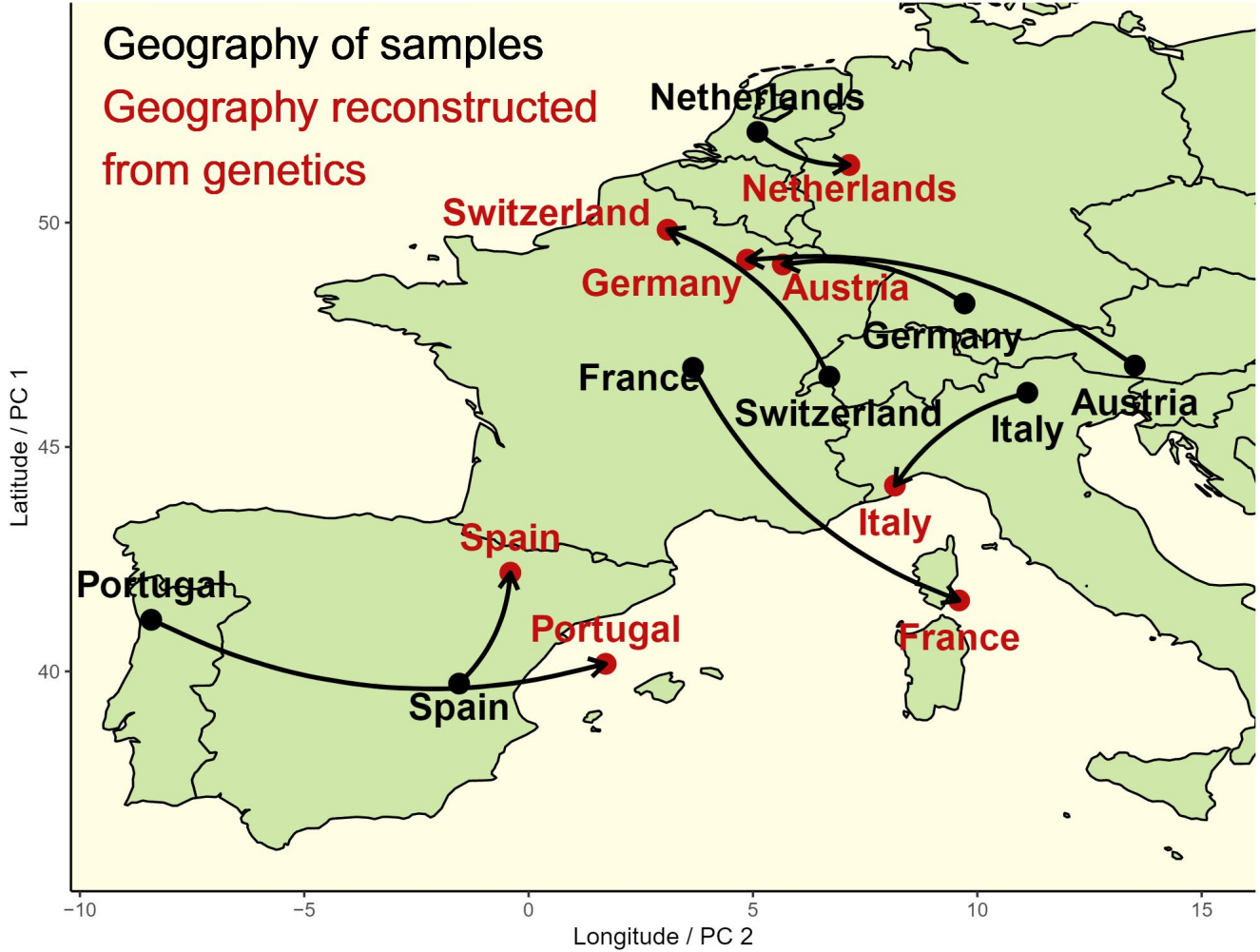
Reconstructing fruit fly geography from genetics

David Bass and Medha Prakash

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



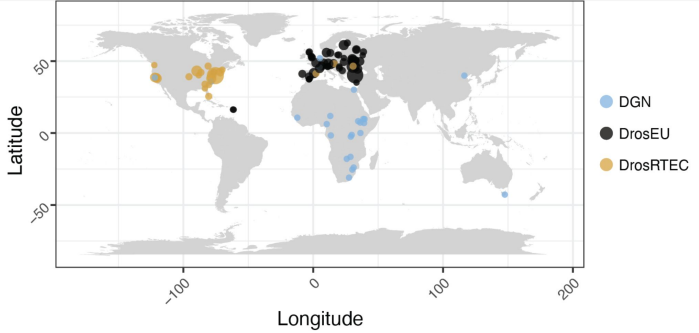
Project Objective

Develop a method for visualizing a type of geospatial data common in population genetics

Reconstruct the geography of biological samples of *D. melanogaster* fruit flies solely from genetic variation information

Visualize the transformation from geographic space to PC-space

Dataset used



Kapun et al. (2021), Fig. 1a

Drosophila Evolution over Space and Time (DEST)

Genetic variation of more than 13,000 *D. melanogaster* fruit flies

272 samples across four continents over several years

We use, latitude, longitude, country of origin, and genetic variation of samples

Unused variables include environmental information, genome sequencing technique, locality, and date

We focus on Western Europe

Our method

Apply principal component analysis (PCA) to genetic variation information of fruit fly samples

Find that PC1 and PC2 strongly correlated with latitude and longitude

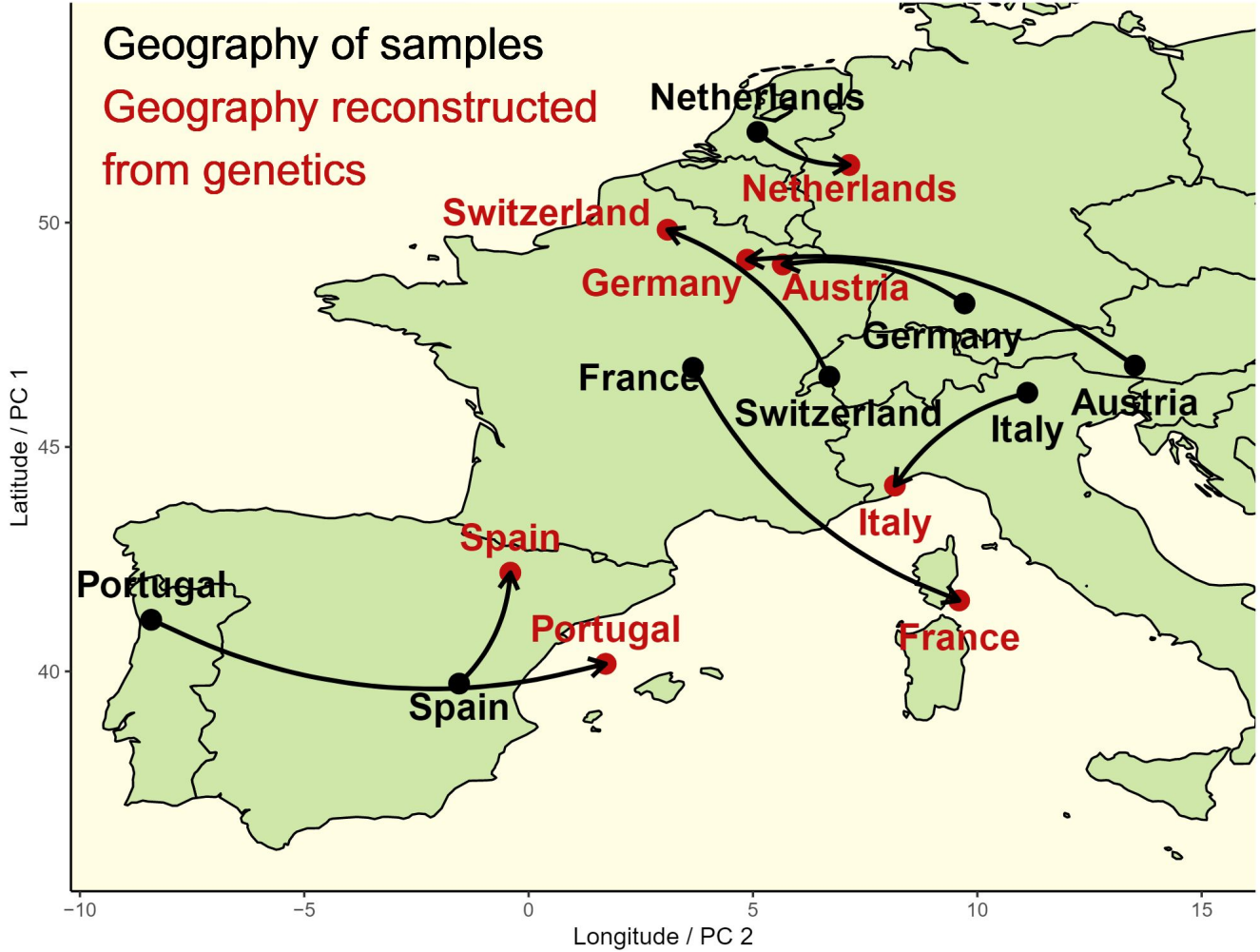
Average latitude and longitude of each country to display geographic information in one point per country **(black)**

Apply Procrustes analysis to PC-space to align it with geographic space

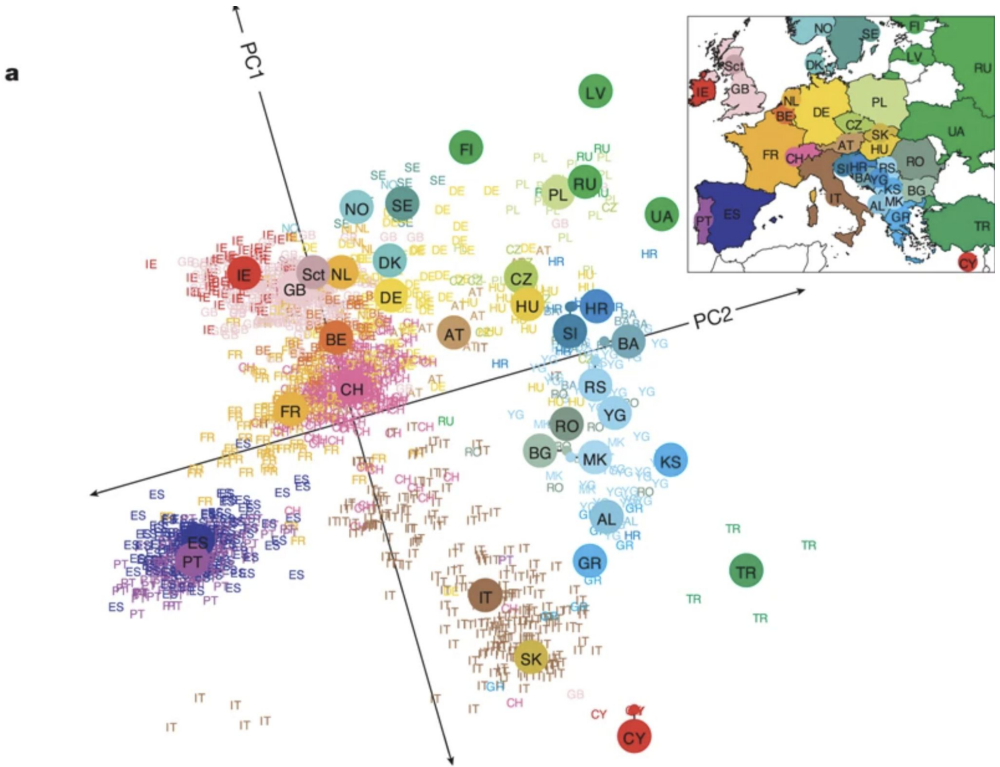
Display PC-space in single points divided by country of relevance **(red)**

Connect the geographic information and PC information with arrows

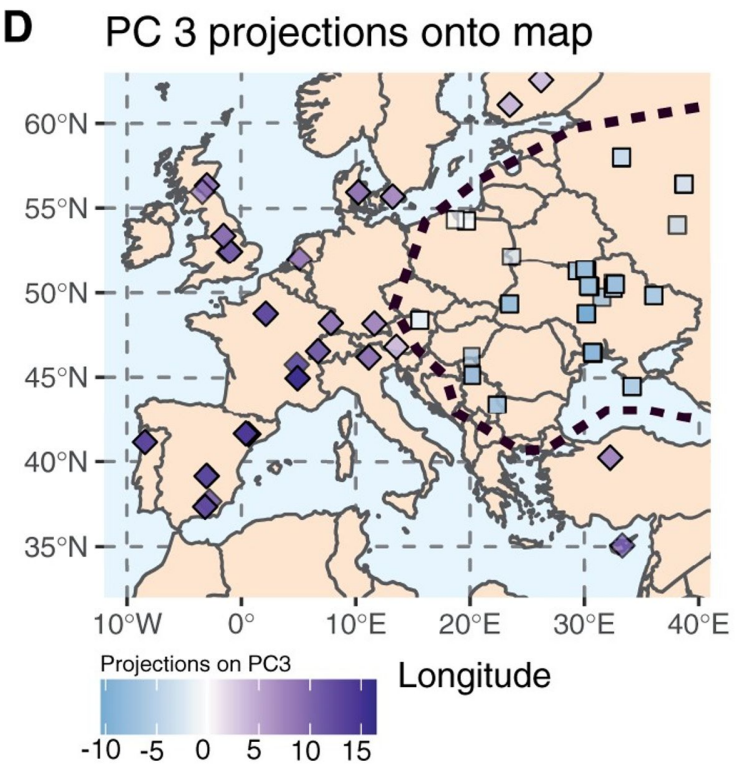
Reconstructing *D. melanogaster* geography from genetics via PCA and Procrustes analysis



Limitations of current methods previous figures



Novembre et al. (2009), Fig. 1a



Kapun et al. (2021), Fig. 7d

Resolving limitations

We apply Procrustes analysis to align PC-space to geographic space

We plot points for both geographic space and PC-space

We emphasize the transformation from geographic space to PC-space with arrows connecting pairs of corresponding points

We group the data so as to only display one point per country in either space

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

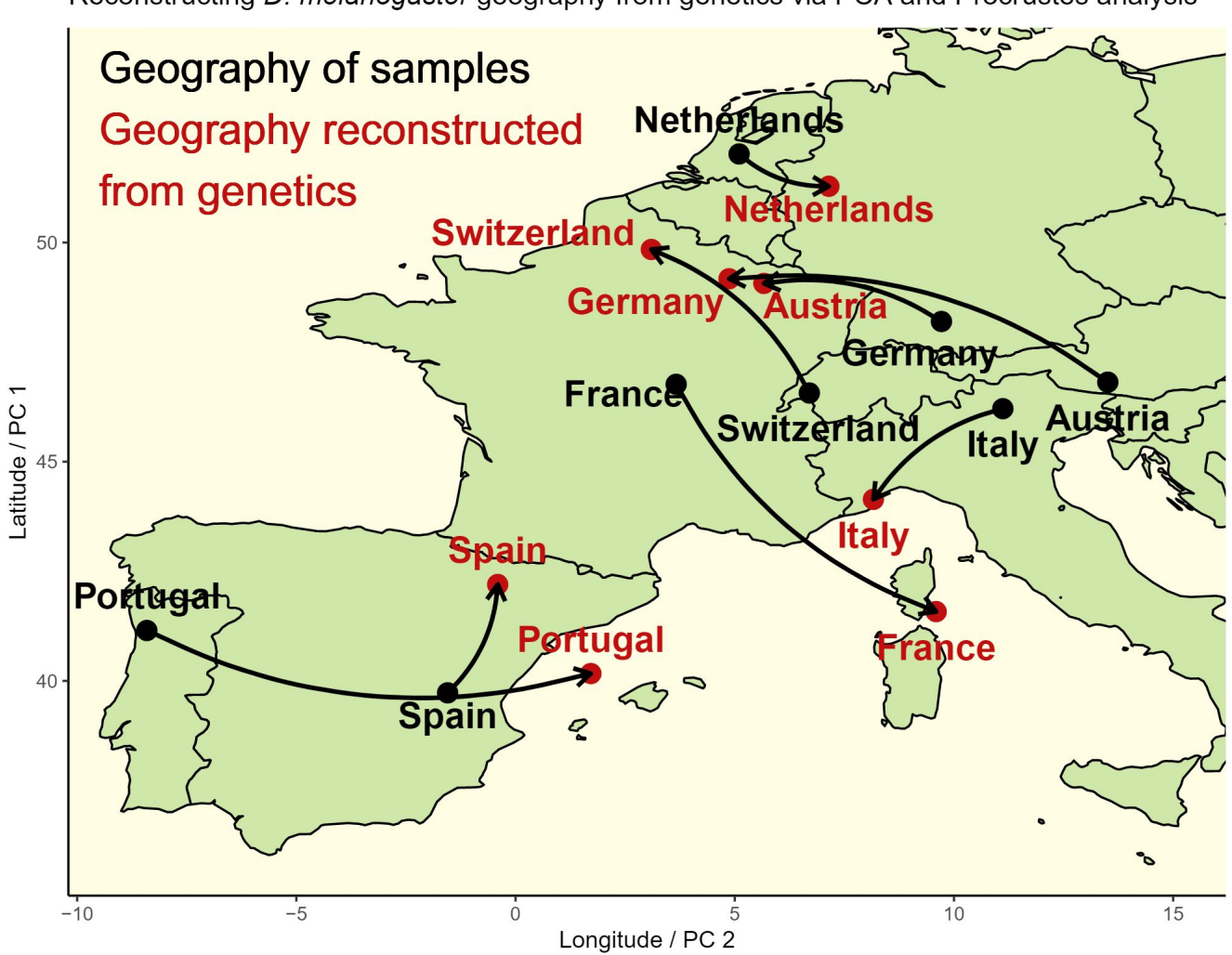
Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P, Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette 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

Questions?


Reconstructing *D. melanogaster* geography from genetics via PCA and Procrustes analysis
Geography of samples
Geography reconstructed from genetics
Netherlands
Netherlands
Switzerland
Germany
Austria
Germany
France
Switzerland
Italy
Austria
Italy
Spain
Portugal
Portugal
France
Spain
50
45
40
Latitude / PC 1
−10
−5
0
5
10
15
Longitude / PC 2