GGoutlieR: an R package to identify and visualize unusual geo-genetic patterns of biological samples

07 Feb 2023

Summary

Landscape genomics is a rising research field integrating genomic and environmental information to explore driving forces of evolution. Reliable geographical origin data of biological samples are a prerequisite for landscape genomics studies. Conventionally, researchers discover potentially questionable samples with visualization-based tools. However, such approaches are infeasible to handle large sample sizes due to overlapping data points on a graph and may encumber reproducible research. To address this shortage, we developed **G**eo-**G**enetic **outlier** (`GGoutlieR`), an R package of a heuristic framework to reveal and visualize samples with unusual geo-genetic patterns. `GGoutlieR` can calculate empirical p values for every sample, allowing researchers to easily spot outliers from thousands of samples. Furthermore, `GGoutlieR` provides a plotting function to display the geo-genetic patterns of outliers on a geographical map. `GGoutlieR` could greatly reduce the researcher's effort for data cleaning before conducting landscape genomics analyses.

Statement of need

Landscape genomics is a thriving field in ecological conservation and evolutionary genetics (Aguirre-Liguori, Ramírez-Barahona, and Gaut 2021; Lasky, Josephs, and Morris 2023), which provides insights into associations between genetic variation and environmental factors. This methodology requires reliable geographical and genomic information of biological samples. To recognize whether data are reliable, researchers may scrutinize associations between genetic similarities and geographical origins of biological samples before carrying out further studies. The pairwise genetic similarities of samples are expected to decline as geographical distances between origin habitats increase, so-called isolation-by-distance assumption. This assumption could be violated due to long-distance migration or artificial factors, such as human activities or mistakes in data and sample management.

Visualization-based tools, such as SPA (Yang et al. 2012), SpaceMix (Bradburd, Ralph, and Coop 2016), unPC (House and Hahn 2018), help researchers to unveil samples with geo-genetic patterns opposing the isolation-by-distance assumption, but those tools do not provide statistics to simply pinpoint outliers. This shortage could be detrimental to the reproducibility of research. Moreover, with the advances in genome sequencing technologies, researchers nowadays work on increasing sample sizes, for example, genebank collection studies in rice (Gutaker et al. 2020; W. Wang et al. 2018), barley (Milner et al. 2019), wheat (Schulthess et al. 2022), soybean (Y. Liu et al. 2020) and maize (J. Li et al. 2019). Visualization-based approaches may have difficulty in presenting unusual geo-genetic patterns in a large data set because of numerous overlapping data points on a graph. Therefore, a new approach is needed to facilitate the detection of unusual geo-genetic associations in biological samples. We developed a heuristic statistic framework to detect **G**eo-**G**enetic **outlier**s, named GGoutlieR. Our package GGoutlieR computes empirical p-values of violating the isolation-by-distance assumption for individual samples according to geographical origins and genotypic data. This feature enables researchers to easily select outliers from thousands of samples for further investigation. Furthermore, GGoutlieR visualizes the geo-genetic patterns of outliers in a network manner on a geographical map, providing insights into the relationships of geography and genetic clusters.

Concept of GGoutlieR

Under the isolation-by-distance assumption, the geographical origins are predictable from genetic variations (Battey, Ralph, and Kern 2020; Guillot et al. 2016), and vice versa. With this respect, prediction models should result in large prediction errors for samples that oppose the isolation-by-distance assumption. We developed the GGoutlieR framework based on this concept to model anomalous geo-genetic patterns.

In brief, GGoutlierR uses KNN regression to predict genetic components with the K nearest geographical neighbors and also does prediction contrariwise. Next, prediction errors are transformed into distance-based (D) statistics and the optimal K is identified by minimizing the sum of D statistics. D statistics is assumed following a Gamma distribution with unknown parameters. An empirical Gamma distribution is obtained as the null distribution by searching optimal parameters with maximum likelihood estimation. With the null Gamma distribution, GGoutlieR tests the null hypothesis that the geo-genetic pattern of a given sample agrees with the isolation-by-distance assumption. Finally, p values for every sample are computed with the empirical null distribution and statistics computed from prediction errors. The details of the GGoutlieR framework are described step-by-step in the supplementary material (GITHUB_LINK).

Example

Outlier identification

For demonstration, we used the genotypic data and passport data of the global barley landrace collection with 1,661 accessions from the IPK genebank (Milner et al. 2019; König et al. 2020). The full analysis of the barley data set with `GGoutlieR` is available in the vignette (GITLAB_LINK). The outlier identification was done with the function `ggoutlier`. The function `summary_ggoutlier` was then used to obtain a summary table of outliers by taking the output of `ggoutlier`.

```r
library(GGoutlieR)
data("ipk_anc_coef") # get ancestry coefficients
data("ipk_geo_coord") # get geographical coordinates

pthres = 0.025 # set a p-value threshold

## run GGoutlieR
ggoutlier_result <- ggoutlier(geo_coord = ipk_geo_coord,
                              gen_coord = ipk_anc_coef,
                              plot_dir = "./fig",
                              p_thres = pthres,
                              cpu = 4,
                              klim = c(3,50),
                              method = "composite",
                              verbose = F,
                              min_nn_dist = 1000)

## print out outliers
head(summary_ggoutlier(ggoutlier_result))

#>                 ID     method      p.value
#> 1  BRIDGE_HOR_2827     geoKNN 0.0002533251
#> 2 BRIDGE_HOR_12795     geoKNN 0.0002871882
#> 3    BRIDGE_BCC_37     geoKNN 0.0003011807
#> 4 BRIDGE_HOR_10557     geoKNN 0.0003500990
#> 5 BRIDGE_HOR_10555     geoKNN 0.0003697789
#> 6        BTR_FT519 geneticKNN 0.0003816026
```

Visualization of unusual geo-genetic patterns

The unusual geo-genetic patterns detected by `GGoutlieR` can be presented on a geographical map with the function `plot_ggoutlier`.

```
plot_ggoutlier(ggoutlier_res = ggoutlier_result,
               gen_coord = ipk_anc_coef,
               geo_coord = ipk_geo_coord,
               p_thres = pthres,
               map_type = "both",
               select_xlim = c(-20,140),
               select_ylim = c(10,62),
               plot_xlim = c(-20,140),
               plot_ylim = c(10,62),
               pie_r_scale = 1.2,
               map_resolution = "course",
               adjust_p_value_projection = F)
```

References

Aguirre-Liguori, Jonas A, Santiago Ramírez-Barahona, and Brandon S Gaut. 2021. "The Evolutionary Genomics of Species' Responses to Climate Change." *Nature Ecology & Evolution* 5 (10). Nature Publishing Group UK London: 1350–60.

Battey, Christopher J, Peter L Ralph, and Andrew D Kern. 2020. "Predicting Geographic Location from Genetic Variation with Deep Neural Networks." *eLife* 9. eLife Sciences Publications, Ltd: e54507.

Bradburd, Gideon S, Peter L Ralph, and Graham M Coop. 2016. "A Spatial Framework for Understanding Population Structure and Admixture." *PLoS Genetics* 12 (1). Public Library of Science San Francisco, CA USA: e1005703.

Guillot, Gilles, Hákon Jónsson, Antoine Hinge, Nabil Manchih, and Ludovic Orlando. 2016. "Accurate Continuous Geographic Assignment from Low-to High-Density Snp Data." *Bioinformatics* 32 (7). Oxford University Press: 1106–8.

Gutaker, Rafal M, Simon C Groen, Emily S Bellis, Jae Y Choi, Inês S Pires, R Kyle Bocinsky, Emma R Slayton, et al. 2020. "Genomic History and Ecology of the Geographic Spread of Rice." *Nature Plants* 6 (5). Nature Publishing Group UK London: 492–502.

House, Geoffrey L, and Matthew W Hahn. 2018. "Evaluating Methods to Visualize Patterns of Genetic Differentiation on a Landscape." *Molecular Ecology*

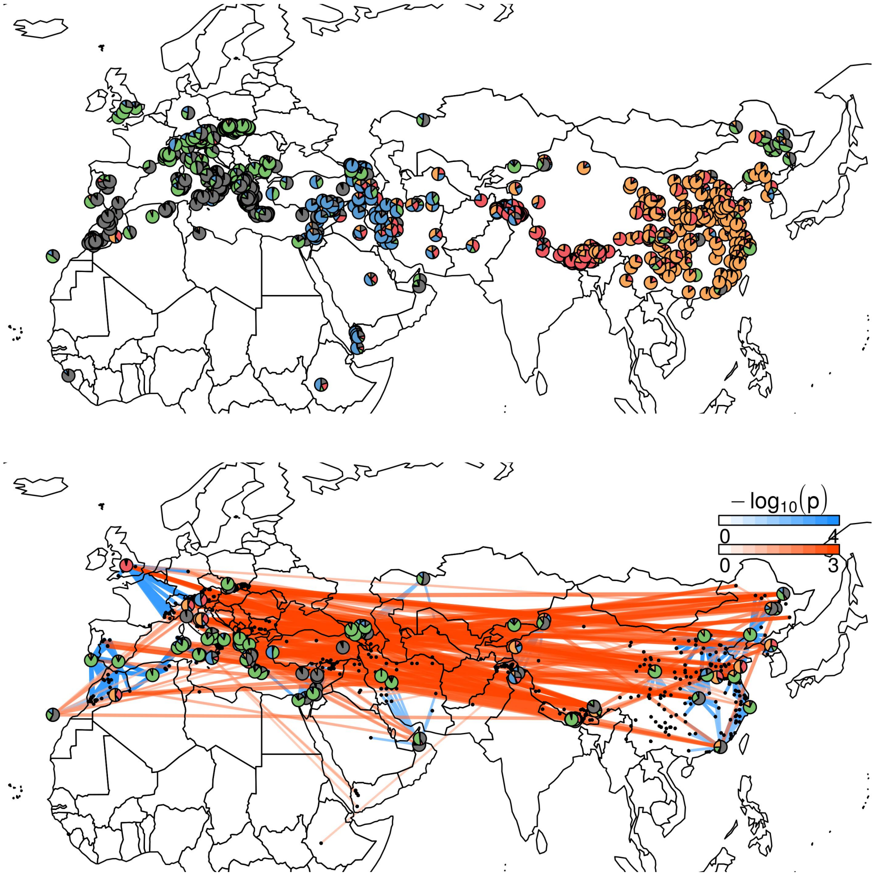

Figure 1: Visualization example of GGoutlieR with IPK barley landrace data. The red lines show the individual pairs with unusual genetic similarities across long geographical distances. The blue lines indicate the unusual genetic differences between geographical neighbors. Pie charts present the ancestry coefficients of outliers identified by GGoutlieR.

Resources 18 (3). Wiley Online Library: 448–60.

König, Patrick, Sebastian Beier, Martin Basterrechea, Danuta Schüler, Daniel Arend, Martin Mascher, Nils Stein, Uwe Scholz, and Matthias Lange. 2020. "BRIDGE–a Visual Analytics Web Tool for Barley Genebank Genomics." *Frontiers in Plant Science* 11. Frontiers Media SA: 701.

Lasky, Jesse R, Emily B Josephs, and Geoffrey P Morris. 2023. "Genotype–Environment Associations to Reveal the Molecular Basis of Environmental Adaptation." *The Plant Cell* 35 (1). Oxford University Press: 125–38.

Li, Jing, Guo-Bo Chen, Awais Rasheed, Delin Li, Kai Sonder, Cristian Zavala Espinosa, Jiankang Wang, et al. 2019. "Identifying Loci with Breeding Potential Across Temperate and Tropical Adaptation via Eigengwas and Envgwas." *Molecular Ecology* 28 (15). Wiley Online Library: 3544–60.

Liu, Yucheng, Huilong Du, Pengcheng Li, Yanting Shen, Hua Peng, Shulin Liu, Guo-An Zhou, et al. 2020. "Pan-Genome of Wild and Cultivated Soybeans." *Cell* 182 (1). Elsevier: 162–76.

Milner, Sara G, Matthias Jost, Shin Taketa, Elena Rey Mazón, Axel Himmelbach, Markus Oppermann, Stephan Weise, et al. 2019. "Genebank Genomics Highlights the Diversity of a Global Barley Collection." *Nature Genetics* 51 (2). Nature Publishing Group US New York: 319–26.

Schulthess, Albert W, Sandip M Kale, Fang Liu, Yusheng Zhao, Norman Philipp, Maximilian Rembe, Yong Jiang, et al. 2022. "Genomics-Informed Prebreeding Unlocks the Diversity in Genebanks for Wheat Improvement." *Nature Genetics* 54 (10). Nature Publishing Group US New York: 1544–52.

Wang, Wensheng, Ramil Mauleon, Zhiqiang Hu, Dmytro Chebotarov, Shuaishuai Tai, Zhichao Wu, Min Li, et al. 2018. "Genomic Variation in 3,010 Diverse Accessions of Asian Cultivated Rice." *Nature* 557 (7703). Nature Publishing Group UK London: 43–49.

Yang, Wen-Yun, John Novembre, Eleazar Eskin, and Eran Halperin. 2012. "A Model-Based Approach for Analysis of Spatial Structure in Genetic Data." *Nature Genetics* 44 (6). Nature Publishing Group: 725–31.