



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

Che-Wei Chang <sup>1</sup> and Karl Schmid <sup>1</sup>

<sup>1</sup> University of Hohenheim, Stuttgart, Germany  Corresponding author

DOI: [10.21105/joss.05687](https://doi.org/10.21105/joss.05687)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Martin Fleischmann](#)  

## Reviewers:

- [@tkchafin](#)
- [@btmartin721](#)

Submitted: 01 May 2023

Published: 30 October 2023

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## Summary

Landscape genomics is an emerging field of research that integrates genomic and environmental information to explore the drivers of evolution. Reliable data on the geographical origin of biological samples is a prerequisite for accurate landscape genomics studies. Traditionally, researchers discover potentially questionable samples using visualization-based tools. However, such approaches cannot handle large sample sizes due to overlapping data points on a graph and can hinder reproducible research. To address this shortcoming, we developed **Geo-Genetic outlier** (GGoutlier), an R package of a heuristic framework for detecting and visualizing samples with unusual geo-genetic patterns. Outliers can be identified using either geography-based K-nearest neighbors (KNNs) or genetics-based KNNs. The framework calculates empirical p-values for each sample, allowing users to easily identify outliers in data sets with thousands of samples. The package also provides a plotting function to display the geo-genetic patterns of outliers on a geographical map. GGoutlier has the potential to significantly minimize the data cleaning required by researchers prior to conducting landscape genomics analyses.

## Statement of need

Landscape genomics is a thriving field in ecological conservation and evolutionary genetics ([Aguirre-Liguori et al., 2021](#); [Lasky et al., 2023](#)), providing insights into the links between genetic variation and environmental factors. This methodology requires reliable geographical and genomic information on biological samples. To determine whether data are reliable, researchers can examine associations between genetic similarities and the geographic origin of biological samples before proceeding with further studies. Under the assumption of isolation-by-distance, pairwise genetic similarities of samples are expected to decrease with increasing geographical distance between the sample origins. This assumption may be violated by long-distance migration or artificial factors such as human activity or data/sample management errors.

Visualization-based tools such as SPA ([Yang et al., 2012](#)), SpaceMix ([Bradburd et al., 2016](#)), unPC ([House & Hahn, 2018](#)) allow to identify samples with geo-genetic patterns that violate the isolation-by-distance assumption, but these tools do not provide statistics to robustly label outliers. Advances in genome sequencing technologies lead to much larger sample sizes, such as in geo-genetic analyses of genebank collections of rice ([Gutaker et al., 2020](#); [Wang et al., 2018](#)), barley ([Milner et al., 2019](#)), wheat ([Schulthess et al., 2022](#)), soybean ([Liu et al., 2020](#)) and maize ([Li et al., 2019](#)). Visualization-based approaches may not be suitable to display unusual geo-genetic patterns in big datasets due to the large number of overlapping data points on a graph. To overcome this problem, we developed a heuristic statistical framework for detecting **Geo-Genetic outliers**, named GGoutlier. Our GGoutlier package computes empirical p-values for violation of the isolation-by-distance assumption for individual samples according to prior information on their geographic origin and genotyping data. This feature allows researchers to

easily select outliers from thousands of samples for further investigation. In addition, GGoutlier visualizes the geo-genetic patterns of outliers as a network on a geographical map, providing insights into the relationships between geography and genetic clusters.

## Algorithm of GGoutlier

“Assuming isolation by distance, the geographical origins of samples can be predicted from their patterns of genetic variation, and vice versa (Battey et al., 2020; Guillot et al., 2016). In this context, prediction models should result in large prediction errors for samples that violate the isolation-by-distance assumption. Based on this concept, we developed the GGoutlier framework to model anomalous geo-genetic patterns.

Briefly, GGoutlier uses  $K$ -nearest neighbour (KNN) regression to predict genetic components with the  $K$  nearest geographical neighbours, and also predicts in the opposite direction. Next, the prediction errors are transformed into distance-based ( $D$ ) statistics and the optimal  $K$  is identified by minimising the sum of the  $D$  statistics. The  $D$  statistic is assumed to follow a gamma distribution with unknown parameters. An empirical gamma distribution is obtained as the null distribution by finding optimal parameters using maximum likelihood estimation. With the null gamma distribution, GGoutlier tests the null hypothesis that the geogenetic pattern of a given sample is consistent with the isolation-by-distance assumption. Finally, p-values are calculated for each sample using the empirical null distribution and prediction error statistics. The details of the GGoutlier framework are described step by step in the supplementary material (<https://github.com/kjschmidlab/GGoutlier/blob/master/paper/supinfo.pdf>).

## Example

### Outlier identification

For demonstration, we used the genotypic and passport data of the global barley landrace collection of 1,661 accessions from the IPK genebank (König et al., 2020; Milner et al., 2019). The full analysis of the barley dataset with GGoutlier is available in the vignette ([https://github.com/kjschmidlab/GGoutlier/blob/master/vignettes/outlier\\_detection.pdf](https://github.com/kjschmidlab/GGoutlier/blob/master/vignettes/outlier_detection.pdf)). Outliers were identified using the `ggoutlier` function. The function `summary_ggoutlier` was then used to obtain a summary table of outliers by taking the output of `ggoutlier`.

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

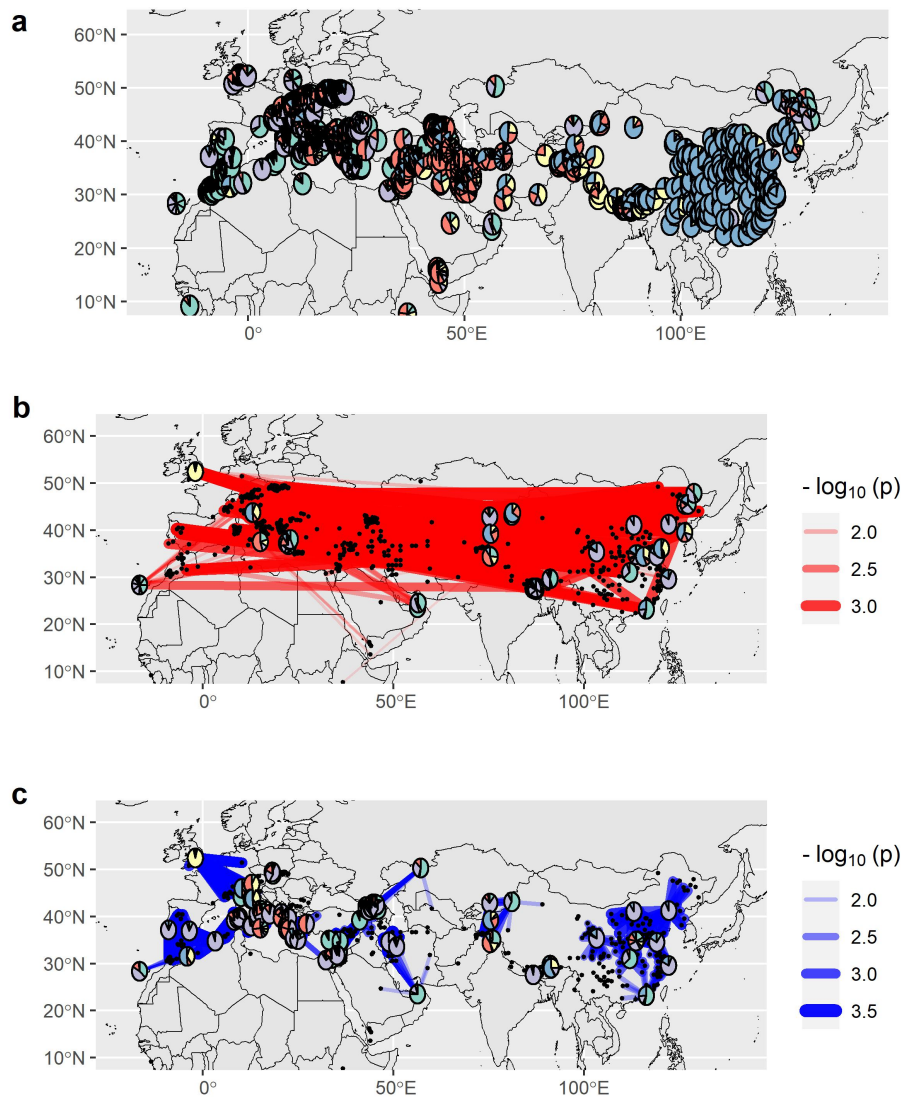
pthres = 0.025 # set a p-value threshold

## 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 0.0002534661
#> 2 BRIDGE_HOR_12795    geoKNN 0.0002875591
#> 3 BRIDGE_BCC_37      geoKNN 0.0003014085
#> 4 BRIDGE_HOR_10557    geoKNN 0.0003502037
#> 5 BRIDGE_HOR_10555    geoKNN 0.0003697646
#> 6      BTR_FT519  geneticKNN 0.0003828147
```

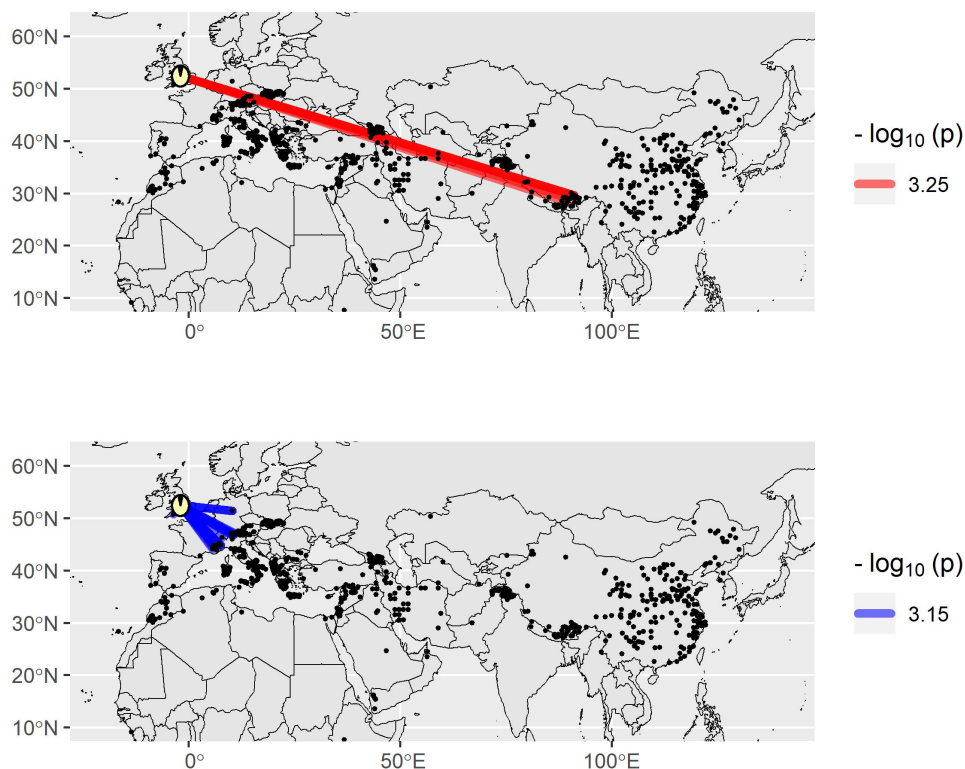
### Visualization of unusual geo-genetic patterns



**Figure 1:** Visualization example of GGoutlierR with IPK barley landrace data. (a) Geographical map with ancestry coefficients of landraces presented by pie charts. (b) and (c) Unusual geo-genetic associations identified by GGoutlierR. 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.

The unusual geo-genetic patterns detected by GGoutlierR can be presented on a geographical map with the function `plot_ggoutlier` (Figure 1).

Moreover, the function `plot_ggoutlier` allows users to gain insight into outliers from a selected geographical region (Figure 2).



**Figure 2:** Visualization example of IPK barley landrace data with a highlight of samples from UK. The red lines show that the outliers in UK are genetically similar to accessions from Southern Tibet.

```
## Visualize GGoutlierR results
## Figure 1: visualize all outliers
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 = 2,
               map_resolution = "medium")

## Figure 2: highlight outliers in UK with `select_xlim` and `select_ylim`
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(-12,4),
select_ylim = c(47,61),
plot_xlim = c(-20,140),
plot_ylim = c(10,62),
pie_r_scale = 2,
map_resolution = "medium",
add_benchmark_graph = F,
plot_labels = NA)
```

## Availability

The GGoutlierR package and vignette are available in our GitHub repository (<https://github.com/kjschmidlab/GGoutlierR>) and CRAN (<https://cran.r-project.org/web/packages/GGoutlierR/index.html>).

## Acknowledgements

We are grateful to Dr. Martin Mascher and Max Haupt of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) for providing raw VCF data of barley landraces used in the example. This work was supported by the funds from the Federal Ministry of Food and Agriculture (BMEL) according to a decision of the parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the Federal Programme for Ecological Farming and Other Forms of Sustainable Agriculture (Project number 2818202615). C.W.C. was supported by the Study Abroad Fellowship from the Education Ministry of Taiwan (R.O.C.) (Project number 1100123625).

## References

- Aguirre-Liguori, J. A., Ramírez-Barahona, S., & Gaut, B. S. (2021). The evolutionary genomics of species' responses to climate change. *Nature Ecology & Evolution*, 5(10), 1350–1360. <https://doi.org/10.1038/s41559-021-01526-9>
- Batthey, C. J., Ralph, P. L., & Kern, A. D. (2020). Predicting geographic location from genetic variation with deep neural networks. *eLife*, 9, e54507. <https://doi.org/10.7554/eLife.54507>
- Bradburd, G. S., Ralph, P. L., & Coop, G. M. (2016). A spatial framework for understanding population structure and admixture. *PLoS Genetics*, 12(1), e1005703. <https://doi.org/10.1371/journal.pgen.1005703>
- Guillot, G., Jónsson, H., Hinge, A., Manich, N., & Orlando, L. (2016). Accurate continuous geographic assignment from low-to high-density SNP data. *Bioinformatics*, 32(7), 1106–1108. <https://doi.org/10.1093/bioinformatics/btv703>
- Gutaker, R. M., Groen, S. C., Bellis, E. S., Choi, J. Y., Pires, I. S., Bocinsky, R. K., Slayton, E. R., Wilkins, O., Castillo, C. C., Negrão, S., & others. (2020). Genomic history and ecology of the geographic spread of rice. *Nature Plants*, 6(5), 492–502. <https://doi.org/10.1038/s41477-020-0659-6>
- House, G. L., & Hahn, M. W. (2018). Evaluating methods to visualize patterns of genetic differentiation on a landscape. *Molecular Ecology Resources*, 18(3), 448–460. <https://doi.org/10.1111/1755-0998.12747>
- König, P., Beier, S., Basterrechea, M., Schüler, D., Arend, D., Mascher, M., Stein, N., Scholz, U., & Lange, M. (2020). BRIDGE—a visual analytics web tool for barley genebank genomics. *Frontiers in Plant Science*, 11, 701. <https://doi.org/10.3389/fpls.2020.00701>

- Lasky, J. R., Josephs, E. B., & Morris, G. P. (2023). Genotype–environment associations to reveal the molecular basis of environmental adaptation. *The Plant Cell*, 35(1), 125–138. <https://doi.org/10.1093/plcell/koac267>
- Li, J., Chen, G.-B., Rasheed, A., Li, D., Sonder, K., Zavala Espinosa, C., Wang, J., Costich, D. E., Schnable, P. S., Hearne, S. J., & others. (2019). Identifying loci with breeding potential across temperate and tropical adaptation via EigenGWAS and EnvGWAS. *Molecular Ecology*, 28(15), 3544–3560. <https://doi.org/10.1111/mec.15169>
- Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M., & others. (2020). Pan-genome of wild and cultivated soybeans. *Cell*, 182(1), 162–176. <https://doi.org/10.1016/j.cell.2020.05.023>
- Milner, S. G., Jost, M., Taketa, S., Mazón, E. R., Himmelbach, A., Oppermann, M., Weise, S., Knüpfner, H., Basterrechea, M., König, P., & others. (2019). Genebank genomics highlights the diversity of a global barley collection. *Nature Genetics*, 51(2), 319–326. <https://doi.org/10.1038/s41588-018-0266-x>
- Schulthess, A. W., Kale, S. M., Liu, F., Zhao, Y., Philipp, N., Rembe, M., Jiang, Y., Beukert, U., Serfling, A., Himmelbach, A., & others. (2022). Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. *Nature Genetics*, 54(10), 1544–1552. <https://doi.org/10.1038/s41588-022-01189-7>
- Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R. R., Zhang, F., & others. (2018). Genomic variation in 3,010 diverse accessions of asian cultivated rice. *Nature*, 557(7703), 43–49. <https://doi.org/10.1038/s41586-018-0063-9>
- Yang, W.-Y., Novembre, J., Eskin, E., & Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics*, 44(6), 725–731. <https://doi.org/10.1038/ng.2285>