# pooledpeaks: Open-Source Fragment Scoring and Analysis Workflow for Population Genetic Analyses of Microsatellite Markers in Pooled Samples

**Kathleen M. Kuesters** [1,¶], **Jessica M. Blanton** [1], **Jeffrey D. Kovach** [2,3], **Walter A. Blank**[4], **Lúcio M. Barbosa** [5,6], **Luciano K. Silva** [6], **Mitermayer G. Reis** [6,7,8], and **Ronald E. Blanton** [1]

**1** Department of Tropical Medicine and Infectious Disease, Tulane University Celia Scott Weatherhead School of Public Health and Tropical Medicine, New Orleans, LA 70112, United States of America ᴿᴼᴿ **2** Case Western Reserve University School of Medicine, Cleveland, OH 44106, United States of America ᴿᴼᴿ **3** Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, United States of America ᴿᴼᴿ **4** Independent Researcher, United States of America **5** Bahiana School of Medicine and Public Health, Av. Silveira Martins, n 3386, Salvador, Bahia, 41150-100, Brazil ᴿᴼᴿ **6** Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Ministério da Saúde, Rua Waldemar Falcão, 121, Candeal, CEP 40296-710, Salvador, Bahia, Brazil **7** Faculdade de Medicina, Universidade Federal da Bahia, Praça XV de novembro, s/n - Largo do Terreiro de Jesus, CEP 40026-010, Salvador, Bahia, Brazil ᴿᴼᴿ **8** Department of Epidemiology of Microbial Diseases, Yale School of Public Health, 60 College St, New Haven, Connecticut, 06510, United States of America ᴿᴼᴿ ¶ Corresponding author

## Summary

Microsatellite markers are short, highly variable, multi-repeat DNA sequences (aka short tandem repeats) that appear throughout the genome and can be used to estimate population genetic metrics (Silva et al., 2006; Vieira et al., 2016). These markers are frequently evaluated using fragment analysis which is based on Sanger sequencing. The pooledpeaks R package provides tools to analyze fragment analysis results (.fsa files). It provides functions that fall in three subcategories: 1) peak scoring, 2) data manipulation, and 3) genetic analysis. The package was designed for the use of microsatellite markers on pooled parasite samples, but the peak scoring functions are applicable to any fragment analysis. The peak scoring functions were partially adapted from Fragman, a package designed to score microsatellite markers in cranberries (Covarrubias-Pazaran et al., 2016). Although Fragman works for the older file version, newer versions cannot be read. In addition to revising this outdated function, we also added features including expanded scoring parameter options and exporting resulting scoring plots as a pdf file for review. The data manipulation functions were created to clean and format the data from the called peaks and transform them into allele frequencies. These frequencies can then be input into the genetic analysis functions for calculation of diversity and differentiation measures adapted from a range of papers (Chao et al., 2008; Foulley & Ollivier, 2006; Jost, 2008; Long et al., 2022; Nei, 1973). An in-depth walk-through of how to use the analysis pipeline can be found in the vignette.

## Statement of Need

While a plethora of methods exist for downstream statistical analysis of allele frequencies, processing raw fragment data is limited by available software. Of the limited software that can read the .fsa binary raw data file format, nearly all require purchase or registration, are primarily built for Windows, are inefficient for analyzing large batches of files, and are highly dependent on individual researcher experience. Additionally, a previous R package allowing for

the analysis of .fsa files is incompatible with the updated file version. When using fragment analysis for microsatellite markers on pooled samples, once the raw data is extracted and scored, it must be cleaned and transformed into allele frequencies using a second software, such as Excel, which is limited in its capacity for automation and version control. Another platform shift is often required to analyze the resulting allele frequencies. These factors highlight the need for a comprehensive scoring and analysis pipeline that is open-source, offline, reproducible, consistent between researchers, and that does not require platform switching between steps.

## Ongoing Research Projects

This package is currently being used to analyze genetic clustering of *Schistosoma mansoni* pooled egg samples from four Brazilian communities, as well as the relatedness of *Schistosoma haematobium* populations around Lac de Guiers in Senegal and from Gabon.

## Acknowledgements (Financial)

## References

Chao, A., Jost, L., Chiang, S. C., Jiang, Y. -H., & Chazdon, R. L. (2008). A Two-Stage Probabilistic Approach to Multiple-Community Similarity Indices. *Biometrics*, *64*(4), 1178–1186. https://doi.org/10.1111/j.1541-0420.2008.01010.x

Covarrubias-Pazaran, G., Diaz-Garcia, L., Schlautman, B., Salazar, W., & Zalapa, J. (2016). Fragman: An R package for fragment analysis. *BMC Genetics*, *17*(1), 62. https://doi.org/10.1186/s12863-016-0365-6

Foulley, J.-L., & Ollivier, L. (2006). Estimating allelic richness and its diversity. *Livestock Science*, *101*(1-3), 150–158. https://doi.org/10.1016/j.livprodsci.2005.10.021

Jost, L. (2008). $G_{ST}$ and its relatives do not measure differentiation. *Molecular Ecology*, *17*(18), 4015–4026. https://doi.org/10.1111/j.1365-294X.2008.03887.x

Long, J. C., Taylor, S. E., Barbosa, L. M., Silva, L. K., Reis, M. G., & Blanton, R. E. (2022). Cryptic population structure and transmission dynamics uncovered for Schistosoma mansoni populations by genetic analyses. *Scientific Reports*, *12*(1), 1059. https://doi.org/10.1038/s41598-022-04776-0

Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences*, *70*(12), 3321–3323. https://doi.org/10.1073/pnas.70.12.3321

Silva, L. K., Liu, S., & Blanton, R. E. (2006). Microsatellite analysis of pooled *Schistosoma mansoni* DNA: An approach for studies of parasite populations. *Parasitology*, *132*(3), 331–338. https://doi.org/10.1017/S0031182005009066

Vieira, M. L. C., Santini, L., Diniz, A. L., & Munhoz, C. D. F. (2016). Microsatellite markers: What they mean and why they are so useful. *Genetics and Molecular Biology*, *39*(3), 312–328. https://doi.org/10.1590/1678-4685-GMB-2016-0027