The Journal of Open Source Software

fastDiversity: an R package for population genetic diversity analysis of multi-species SNP datasets

Eilish S. McMaster [1¶]

1 School of Life and Environmental Sciences, University of Sydney, Australia ¶ Corresponding author

DOI: 10.21105/joss.11113

Software

- Review
- Repository
- Archive

Editor: Ethan White

Reviewers:

- @currocam
- @ethanwhite

Submitted: 18 May 2026
Published: 18 November 2026

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

Summary

Genetic diversity underpins evolutionary potential, adaptive capacity, and population persistence, making its quantification central to conservation genetics and population biology (Frankham et al., 2010). Reduced-representation sequencing approaches including DArTseq (Sansaloni et al., 2011) and RADseq (Baird et al., 2008) now routinely generate large biallelic single nucleotide polymorphism (SNP) datasets spanning hundreds of individuals and thousands of loci. Increasingly, these datasets include multiple closely related species or divergent lineages analysed within a comparative framework.

fastDiversity is an R package for rapid calculation of population genetic diversity statistics from biallelic SNP dosage matrices. The package operates directly on standard numeric genotype matrices encoded as 0, 1, 2, and NA, avoiding conversion to specialised genetic data classes. fastDiversity implements allele-sharing and private allele analyses; site- and population-level heterozygosity and inbreeding statistics (Hartl & Clark, 2007; Nei, 1978); population-specific F_{ST} (Weir & Goudet, 2017) and G_{ST} (Nei, 1973); rarefied allelic richness (Kalinowski, 2004); nucleotide diversity (Nei & Li, 1979), Watterson's theta (Watterson, 1975), and Tajima's D (Tajima, 1989), implemented following pixy (Korunes & Samuk, 2021); individual multilocus heterozygosity and inbreeding coefficients (Frankham et al., 2010); and Hardy–Weinberg equilibrium testing (Graffelman, 2015), with optional bootstrap confidence intervals.

A defining feature of the package is that locus-level filtering by minor allele frequency, missingness, and allele count is performed independently within user-defined genetic groups. This enables biologically appropriate analysis of multi-species datasets within a single workflow, avoiding artefacts introduced by pooled filtering across reproductively isolated groups.

fastDiversity has been under active development since 2024 and the software is archived on Zenodo (doi:10.5281/zenodo.12741582).

Statement of need

Several R packages provide population genetic analysis tools for biallelic SNP datasets, including hierfstat (Goudet, 2005), diveRsity (Keenan et al., 2013), adegenet (Jombart, 2008), dartR (Gruber et al., 2018), and poppr (Kamvar et al., 2014), collectively covering diversity statistics, ordination, and population structure inference. These tools are widely used but generally apply locus-level filtering globally across the pooled dataset, or require datasets to be manually split by species prior to analysis.

This presents a recurring problem for multi-species datasets. When MAF thresholds, missingness filters, or expected heterozygosity are calculated across pooled samples from divergent taxa, allele frequencies reflect a mixture of reproductively isolated gene pools rather than any biologically meaningful unit. As a result, loci that are polymorphic within a species may be incorrectly removed by pooled MAF filters, while loci that are fixed within species but

differentiated between them can inflate diversity or F_{IS} estimates when treated as within-population variation. Differences in locus presence and missingness across taxa further bias locus retention under standard workflows (Figure 1). Species-specific filtering avoids these artefacts by evaluating loci within genetic groups, preserving within-taxon signal while enabling comparable summaries across taxa.

Beyond filtering correctness, performing independent analyses per genetic group simultaneously, rather than through repeated manual subsetting, reduces friction and allows comparative diversity analyses to sit alongside other standard steps such as dimensionality reduction, ancestry estimation, and pairwise F_{ST} within a single analysis script.

fastDiversity was developed to address this gap. It applies filtering and diversity estimation independently within user-defined genetic groups while returning results in a unified tidy output, enabling scalable multi-species analyses without repetitive manual re-analysis. The package operates directly on standard SNP dosage matrices rather than requiring conversion to specialised object classes such as genind or genlight, lowering the barrier for users working within existing pipelines.

Species	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5
A	2	0	0	2	NA
A	1	0	0	1	NA
A	0	0	0	2	NA
A	2	0	0	0	NA
B	0	0	0	NA	1
B	1	0	0	NA	2
B	0	0	0	NA	0
B	2	0	0	NA	0
B	1	0	0	NA	1
C	2	1	2	2	NA
C	0	2	2	2	NA
C	1	0	2	2	NA

Figure 1: Effects of pooled vs species-specific filtering in multi-species SNP datasets. Example genotype matrix showing how locus informativeness differs across three species (A–C) under pooled filtering. Loci may be informative across all species (L1), polymorphic only within one species and potentially removed by pooled MAF filters (L2), alternatively fixed within species inflating pooled summary statistics such as FIS (L3), informative in one species but fixed in others and potentially removed by pooled MAF filters (L4), or present only in a subset of species and excluded under pooled missingness filtering (L5). Species-specific filtering retains different informative loci per taxon (A: L1, L4; B: L1, L5; C: L1, L2), illustrating how pooled filtering can distort comparative SNP datasets.

Software design and functionality

fastDiversity is designed around two principles: biologically appropriate filtering and minimal data-handling overhead.

The package accepts standard biallelic SNP dosage matrices produced by common variant-calling pipelines, together with plain metadata vectors describing sample groupings. This design allows integration into existing R analysis pipelines without object conversion or restructuring.

All locus-level quality filters, including missingness and minor allele frequency thresholds, are applied independently within each user-defined genetic group. This filtering framework propagates throughout the package and ensures that downstream diversity statistics are calculated only from loci informative within the relevant biological unit.

Because most SNP-calling pipelines report only polymorphic sites, the documentation

distinguishes statistics by their dependence on monomorphic sites. Ratio statistics (F_{IS}, G_{ST}, F_{ST}) and Tajima's D are unaffected by them. Heterozygosity and allelic richness are averages over the loci analysed, so they are comparable among sites within a genetic group but are not genome-wide estimates. Nucleotide diversity and Watterson's θ are calculated from segregating sites only and reported as sums rather than per-site values, because the number of invariant sites sequenced is unavailable for many reduced-representation datasets.

Implemented analyses span multiple biological scales, including: 1) allele-level summaries such as allele sharing and private alleles; 2) population-level diversity statistics including observed and expected heterozygosity, inbreeding coefficients, rarefied allelic richness (Kalinowski, 2004), and differentiation as population-specific F_{ST} (Weir & Goudet, 2017), the per-population counterpart of Weir and Cockerham's θ (Weir & Cockerham, 1984), and G_{ST} (Nei, 1973); 3) sequence diversity statistics including nucleotide diversity, Watterson's theta, and Tajima's D using missing-data-aware approaches (Korunes & Samuk, 2021); and 4) individual-level summaries including multilocus heterozygosity and Wright's inbreeding coefficient. Where appropriate, bootstrap and resampling procedures adapted from Goudet (2005) and Keenan et al. (2013) are implemented to estimate confidence intervals and standardise comparisons among uneven sample sizes.

Summary functions return tidy data frames (Wickham, 2014), with one row per genetic group and site and one column per statistic, which can be passed directly to standard R plotting and data-manipulation tools.

State of the field

hierfstat (Goudet, 2005), diveRsity (Keenan et al., 2013), adegenet (Jombart, 2008), poppr (Kamvar et al., 2014), dartR (Gruber et al., 2018), and tidypopgen (Carter et al., 2025) provide established implementations of diversity and differentiation statistics, but each retains a single set of loci for all individuals in a dataset. dartR can take populations into account when filtering by minor allele frequency, but the retained loci are still shared across all individuals, and locus selection in tidypopgen likewise applies to the whole dataset. Applying the same filters independently within each species therefore requires splitting the data and repeating the analysis for each group. fastDiversity applies the filters within every genetic group in a single call and returns the results in one table, complementing these packages for comparative multi-species SNP datasets.

Research impact statement

fastDiversity has been used in more than 30 confidential threatened-species assessments for the New South Wales Government Saving our Species program (Cascini et al., 2025; Doyle et al., 2025), where comparable diversity estimates across related taxa inform conservation management. It has also supported peer-reviewed conservation genomic studies of threatened Australian plants, including multi-species comparisons of diversity and population structure in *Homoranthus* (Myrtaceae) (McMaster, Pemberton, et al., 2025), subspecies delimitation in the critically endangered *Zieria obcordata* (Rutaceae) (McMaster, Duretto, et al., 2025), and evaluation of genetic rescue risks for Australia's rarest *Eucalyptus* (McMaster et al., 2026), as well as a comparison of kinship estimation methods for reduced-representation SNP data (McMaster, Lu-Irving, et al., 2025).

Acknowledgements

Development of fastDiversity benefited from discussions with collaborators in conservation genomics and population genetics at the University of Sydney, the Botanic Gardens of Sydney, and Queensland Herbarium.

AI usage disclosure

Claude Sonnet 4.6 (Anthropic) and ChatGPT (OpenAI GPT-5.5) were used to assist with editing and restructuring documentation and manuscript text based on author-provided drafts and outlines. All scientific content, software design decisions, implementations, analyses, and interpretations were developed, verified, and approved by the author.

References

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Carter, E., Tysall, E., Hodgson, J., & Manica, A. (2025). Tidypopgen: Tidy population genetics in R. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210x.70204

Cascini, M., Doyle, C. A. T., Mulcahy, A., McMaster, E. S., Dimon, R., Hogbin, P. M., Merwe, M. van der, Yap, J.-Y. S., & Rossetto, M. (2025). The impact of taxonomic confusion on conservation resources – why population genomics should inform threatened species determination. *Biological Conservation*, *306*, 111113. https://doi.org/10.1016/j.biocon.2025.111113

Doyle, C. A. T., Cascini, M., Yap, J.-Y. S., Matthews, H., Hogbin, P. M., Wilson, T. C., Mahon, E., Brown, D., Mulcahy, A., Brown, R., & Rossetto, M. (2025). Conservation genomics within government led conservation planning: An Australian case study exploring cost and benefit for threatened flora. *Annals of Botany*, *135*(6), 1229–1242. https://doi.org/10.1093/aob/mcae222

Frankham, R., Ballou, J. D., & Briscoe, D. A. (2010). *Introduction to conservation genetics* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511809002

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184–186. https://doi.org/10.1111/j.1471-8286.2004.00828.x

Graffelman, J. (2015). Exploring diallelic genetic markers: The HardyWeinberg package. *Journal of Statistical Software*, *64*(3), 1–23. https://doi.org/10.18637/jss.v064.i03

Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2018). dartR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*, *18*(3), 691–699. https://doi.org/10.1111/1755-0998.12745

Hartl, D. L., & Clark, A. G. (2007). *Principles of population genetics* (4th ed.). Sinauer Associates.

Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Kalinowski, S. T. (2004). Counting alleles with rarefaction: Private alleles and hierarchical sampling designs. *Conservation Genetics*, *5*(4), 539–543. https://doi.org/10.1023/B:COGE.0000041021.91777.1a

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2*, e281. https://doi.org/10.7717/peerj.281

Keenan, K. G., McGinnity, P., Cross, T. F., Crozier, W. W., & Prodöhl, P. A. (2013). DiveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution*, *4*(8), 782–788. https:

//doi.org/10.1111/2041-210x.12067

Korunes, K. L., & Samuk, K. (2021). Pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*(4), 1359–1368. https://doi.org/10.1111/1755-0998.13326

McMaster, E. S., Duretto, M., Yap, J.-Y. S., & Rossetto, M. (2025). Conservation genomics uncovers disjunct subspecies and critically low diversity in Zieria obcordata A.Cunn. (Rutaceae). *Australian Systematic Botany*, *38*(1). https://doi.org/10.1071/SB24034

McMaster, E. S., Lu-Irving, P., Merwe, M. M. van der, Ho, S. Y. W., & Rossetto, M. (2025). Evaluating kinship estimation methods for reduced-representation SNP data in non-model species. *Molecular Ecology Resources*, e70038. https://doi.org/10.1111/1755-0998.70038

McMaster, E. S., Pemberton, P. J., Bruhl, J. J., Fawcett, A., Hunter, J. T., Saunders, M. E., Wandrag, E. M., Yap, J.-Y. S., Telford, I. R. H., Rossetto, M., & Andrew, R. L. (2025). Geographically proximate rare species exhibit strong population divergence while maintaining intraspecific genetic diversity in Homoranthus (Myrtaceae). *Annals of Botany*, mcaf316. https://doi.org/10.1093/aob/mcaf316

McMaster, E. S., Wright, G., Martino, A., Yap, J.-Y. S., & Rossetto, M. (2026). Genomic evidence reframes the risks of genetic rescue for Australia's rarest Eucalyptus. *Biological Conservation*, *320*, 111929. https://doi.org/10.1016/j.biocon.2026.111929

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

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, *89*(3), 583–590. https://doi.org/10.1093/genetics/89.3.583

Nei, M., & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, *76*(10), 5269–5273. https://doi.org/10.1073/pnas.76.10.5269

Sansaloni, C., Petroli, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., & Kilian, A. (2011). Diversity arrays technology (DArT) and next-generation sequencing combined: Genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. *BMC Proceedings*, *5*, P54. https://doi.org/10.1186/1753-6561-5-S7-P54

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://doi.org/10.1093/genetics/123.3.585

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256–276. https://doi.org/10.1016/0040-5809(75)90020-9

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Weir, B. S., & Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics*, *206*(4), 2085–2103. https://doi.org/10.1534/genetics.116.198424

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10