Received: 15 August 2017 | Revised: 5 December 2017 | Accepted: 11 December 2017

DOI: 10.1111/1755-0998.12745

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

DARTR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing

Bernd Gruber[1] | Peter J. Unmack[1] | Oliver F. Berry[2] | Arthur Georges[1]

[1]Institute for Applied Ecology, University of Canberra, Canberra, ACT, Australia

[2]CSIRO Environomics Future Science Platform, Indian Ocean Marine Research Centre, The University of Western Australia (M097), Crawley, WA, Australia

Correspondence
Bernd Gruber, Institute for Applied Ecology, University of Canberra, Canberra, ACT, Australia.
Email: bernd.gruber@canberra.edu.au

Abstract

Although vast technological advances have been made and genetic software packages are growing in number, it is not a trivial task to analyse SNP data. We announce a new R package, DARTR, enabling the analysis of single nucleotide polymorphism data for population genomic and phylogenomic applications. DARTR provides user-friendly functions for data quality control and marker selection, and permits rigorous evaluations of conformation to Hardy–Weinberg equilibrium, gametic-phase disequilibrium and neutrality. The package reports standard descriptive statistics, permits exploration of patterns in the data through principal components analysis and conducts standard F-statistics, as well as basic phylogenetic analyses, population assignment, isolation by distance and exports data to a variety of commonly used downstream applications (e.g., NEWHYBRIDS, FASTSTRUCTURE and phylogeny applications) outside of the R environment. The package serves two main purposes: first, a user-friendly approach to lower the hurdle to analyse such data—therefore, the package comes with a detailed tutorial targeted to the R beginner to allow data analysis without requiring deep knowledge of R. Second, we use a single, well-established format—genlight from the ADEGENET package—as input for all our functions to avoid data reformatting. By strictly using the genlight format, we hope to facilitate this format as the de facto standard of future software developments and hence reduce the format jungle of genetic data sets. The DARTR package is available via the R CRAN network and GitHub.

KEYWORDS

next-generation sequencing, phylogenomics, population genomics, R package, RADSeq, SNPs

1 | INTRODUCTION

The genomic revolution in ecology and evolution has seen a shift in the components of workflow that govern productivity. The limiting factor to progress in research has moved from access to the technologies to generate DNA sequence data, to access to the computing facilities and availability of software to analyse the voluminous data emerging from next-generation sequencing technologies (Stephens et al., 2015). This shift has been accelerated by commercial providers of sequencing capacity and plummeting costs. Whereas in the recent past, genomic analysis would have required a sophisticated laboratory and trained staff, increasingly researchers require minimal laboratory preparation prior to submitting samples for analysis. Indeed, the major challenges have shifted from the laboratory to the bioinformatic treatment and analysis of sequence data.

One of the most prominent genomic data sought by ecologists are single nucleotide polymorphisms (SNPs) obtained through restriction site-associated DNA sequencing (RADSeq), double digest RADSeq (ddRADSeq) and DarTSeq (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Baird et al., 2008; Jaccoud, Peng, Feinstein, & Kilian, 2001; Kilian et al., 2012; Peterson, Weber, Kay, Fisher, & Hoekstra, 2012; Sansaloni et al., 2011; van Tassell et al., 2008).

These techniques enable genomewide studies of so-called non-model organisms, those for which there is limited genomic information. Common questions that are addressed with these kind of genomic data include those arising from studies on population structure (Beheregaray et al., 2017; Morin, Martien, & Taylor, 2009), isolation by distance, landscape genomic analyses (Brauer, Hammer, & Beheregaray, 2016; Swaegers et al., 2015), phylogenomics (Unmack et al., 2017; Zegura, Karafet, Zhivotovsky, & Hammer, 1999), loci under selection and genomewide association studies (Johnston et al., 2011; Santure et al., 2013). Typical SNP data sets consist of thousands to tens or even hundreds of thousands of loci, often more than two to three orders of magnitude more markers than were typically employed for equivalent microsatellite DNA analyses (Glover et al., 2010; Ytournel et al., 2011). This presents challenges for data management and analysis as popular stand-alone executable programs designed for smaller data sets (e.g., FSTAT, Goudet, 1995 and GENPOP, Raymond & Rousset, 1995) are often poorly suited. Increasingly, researchers are accessing the R programming language and the associated CRAN repository of R packages (for a current review refer to Paradis, Gosselin, Goudet, Jombart, and Schliep (2017)).

An increasing number of R packages are available for the analysis of SNP data sets (e.g., APE, PEGAS, ADEGENET, STAMPP, SNPASSOC, GAP and SNPRELATE), which are often developed for a specific type of analysis. A major difficulty is the array of data formats employed by the different packages, which complicates the construction of workflows combining multiple packages. This is especially true for researchers new to R, as it often requires a deep understanding of R structures and commands to be able to convert large genetic data sets from one format in the other. Another difficulty is a lack of packages that conduct the complete set of data evaluation analyses including quality control, Hardy–Weinberg equilibrium, gametic-phase disequilibrium and tests for markers under selection, as well as fundamental analyses such as isolation-by-distance analysis or assignment tests. In their recent review, Paradis et al. (2017) summarized the main packages and data formats that are available to analyse SNP data. Recently, many more packages often targeted to a specific analysis have been developed (e.g., PARALLELNYWHYBRID, SNPASSOC, RSNPSET, SNPRELATE, MIXMAP and SURFING), most of them using a proprietary format, which often challenges the use by researchers not fluent in R.

Here, we announce DARTR, a user-friendly R package designed as a workhorse for the preparation of SNP data sets for population genomics and phylogenomics, for the exploration of the data, and the production of framework analyses common in these types of analyses. A key part of the development of DARTR is to provide simple functions and a detailed manual to allow data analysis without requiring detailed knowledge of R to conduct analyses, thus broadening access to researchers. DARTR employs the compact genlight data format, making it readily combined in workflows with other popular R packages. A detailed description of the data format is provided in the tutorial for the package (see supporting material). Initially, DARTR was developed to provide custom functions to access and explore SNP data obtained from a leading commercial provider, Diversity Arrays Technologies Pty Ltd (DArT), but during its development, the focus became much broader and now it is suited to the analysis of large SNP data sets obtained from any provider or method. The package comes with a detailed vignette (tutorial) explaining the use of each function accompanied by sample analyses and guidelines for those wishing to contribute their scripts to the package. We outline typical steps of an analysis below. Finally, we invite researchers to follow the proposed data format and implement functions that allow nonexperts to run their analysis, without the need to take care to reformat their data.

2 | DATA AND FORMAT

2.1 | Genlight format

The genlight format of the ADEGENET package uses a bit-level coding scheme for SNP data, that is highly compact and brings access to exceptionally large SNP data sets to the desktop computer (Jombart & Ahmed, 2011). The format is accessible to the user, because ADEGENET comes with a range of methods that allow access to the SNP data analogous to those used to access a standard data matrix in R. DARTR extends the genlight format by adding two additional tables (data frames) to the genlight object; one that we refer to as loc.metrics contains metadata associated with each locus (e.g., clone id, call rate, trimmed sequence tag and SNP location), the other that we refer to as ind.metrics which contains metadata associated with each individual or sample (e.g., sample id, sex, population, latitude and longitude). For further information on accessing and manipulating genlight objects in R, including the associated metadata, refer to the user guides and manuals that accompany the ADEGENET package (Jombart & Ahmed, 2011).

2.2 | Importing data

Data can be imported to a genlight object in several ways. For those drawing on the services of Diversity Arrays Technologies Pty Ltd, a function in DARTR (gl.read.dart) intelligently combines the SNP csv file provided by DArT, which includes a range of locus metadata, and a user-generated metadata file that contains the individual or sample metadata. The metadata are attached to the genlight object via the @other slot and can be accessed using gl@other$loc.metrics and gl@other$ind.metrics. Here, possible extensions are data that store the coordinates of the sample (latlong), additional metadata for loci (loc.metrics) and individuals (ind.metrics) in data.frames.

If users have a data set that is not provided by DArT, import is possible via one of the many approaches outlined in Table 1. In many cases, this involves import to a genind object (Jombart, 2008), then conversion to a genlight object. Dataframes containing metadata can be added to the genlight object using conventional R code. Examples how to import such data are provided in the DARTR vignette.

TABLE 1 Possible import pathways to convert SNP data to genlight format

Import path	Package	Pathway[a]	Description
gl.read.dart	DARTR	–	Based on DaRT data [with optional meta data for individuals]
read.loci	PEGAS	loci2genind, gi2gl	Data set are provided as a csv text file (?read.loci)
read.vcfR	PEGAS	vcfR2genlight	vcf text file (vcfR package)
read.fstat	ADEGENET	gi2gl	Fstat format (version 2.9.3) by Jerome Goudet
read.genetix	ADEGENET	gi2gl	Format Belkhir K., Borsa P., Chikhi L., Raufaste N. & Bonhomme F. (1996–2004) GENETIX
read.structure	ADEGENET	gi2gl	Structure format of Pritchard, J.; Stephens, M. & Donnelly, P. (2000)
read.PLINK	ADEGENET	–	Data provided in PLINK format
fasta2genlight	ADEGENET	–	Extracts SNPs data from fasta format (?ADEGENET)
read.genetable	POPGENREPORT	gi2gl	csv text file based on df2genind Adamack and Gruber (2014) (?read.genetable)

[a] Pathway provides the order of functions needed to convert data to genlight, — indicates that the function directly converts to a genlight object.

3 | VISUALIZATION

Visual exploration of data is an essential prerequisite to more detailed formal analysis. A very helpful and quick technique to check the underlying population structure is to use a dimension reduction method such as PCoA (principal coordinate analysis). The idea of such an approach is to reduce the population structure to two or three dimensions while retaining the maximal information of the data set. The amount of explained variance is the eigenvalue of the visualized axis expressed as a percentage of the sum of the eigenvalues. There are a large number of R scripts available for plotting data following dimension reduction, including glPCA in package ADEGENET, and DARTR does not attempt to reproduce these. The DARTR function gl.pcoa acts as a wrapper for glPca function of package ADEGENET with default settings, converting the eigenvalues to percentages, and adding some additional diagnostics. The command

```
pc <- gl.pcoa(gl, nfactors=5)
```

where gl is the genlight file and yields an object that contains the eigenvalues, factor scores and factor loadings that can be accessed for subsequent analyses. A scree plot can be used to decide on the number of dimensions. The function gl.pcoa.scree(pc) optionally plots the relative contribution of each dimension to total variance, for those dimensions that show an improvement over the original variables. The ordination can be plotted in two dimensions (gl.pcoa.plot) or three dimensions (gl.pcoa.plot.3d) and interacts intelligently with package PLOTLY to allow points to be identified by mouse-over (Sievert et al., 2017). Figure 1 provides an example output using the test data supplied with the package.

4 | ANALYSIS

More than 40 functions are included in DARTR to perform various analyses of SNP data (refer to Table S1). Depending on the aim of the study, these functions can serve as an additional filtering tool or for a final analysis of the data set. Functions that represent either simple filtering tools or are simply user-friendly implementations of already available functions are presented here only briefly (extended examples are provided in the tutorial). We focus below mainly on functions that are either new or allow for a very efficient analysis that is not possible using existing R packages.

4.1 | Filtering

Filtering the data on a range of criteria is often one of the first steps in an analysis, with the stringency of filtering depending on the requirements of subsequent analyses. The package DARTR provides a number of options and two examples follow.

SNP data sets often contain data where the SNP locus has not been called, referred to loosely as missing data. Failure to call a SNP state at a locus for a particular individual can arise because read depth is insufficient or variable such that the SNP state is ambiguous by consensus, or the sequence tag is missed in the sequencing phase. This is usually overcome by ensuring sufficient average read depth, say >20, made possible by the reduced representation which is determined by the choice of restriction enzymes in the digest, and the sequencing depth (e.g., samples per lane on an Illumina platform). The second reason for missing data, and the more common reason in robust data sets, is mutation at one or the other restriction enzyme recognition sites. In a sense, these missing data contain information, akin to the information derived from AFLPs.

Filtering on call rate is carried out using the function gl.filter.callrate, which can be applied on a locus by locus basis or an individual by individual basis. For example,

```
gl <- gl.filter.callrate(gl, method="loc", t=0.95)
```

retains only loci that have less than 5% missing data and their associated metadata. Using method=″ind″ will filter individuals with call rates lower than the specified threshold.

Reproducibility: DArT Pty Ltd. provides locus metadata that includes a measure of reproducibility. Thirty per cent of samples are run a second time, and average the proportion of technical replicate assay pairs for which the marker score is consistent over the two

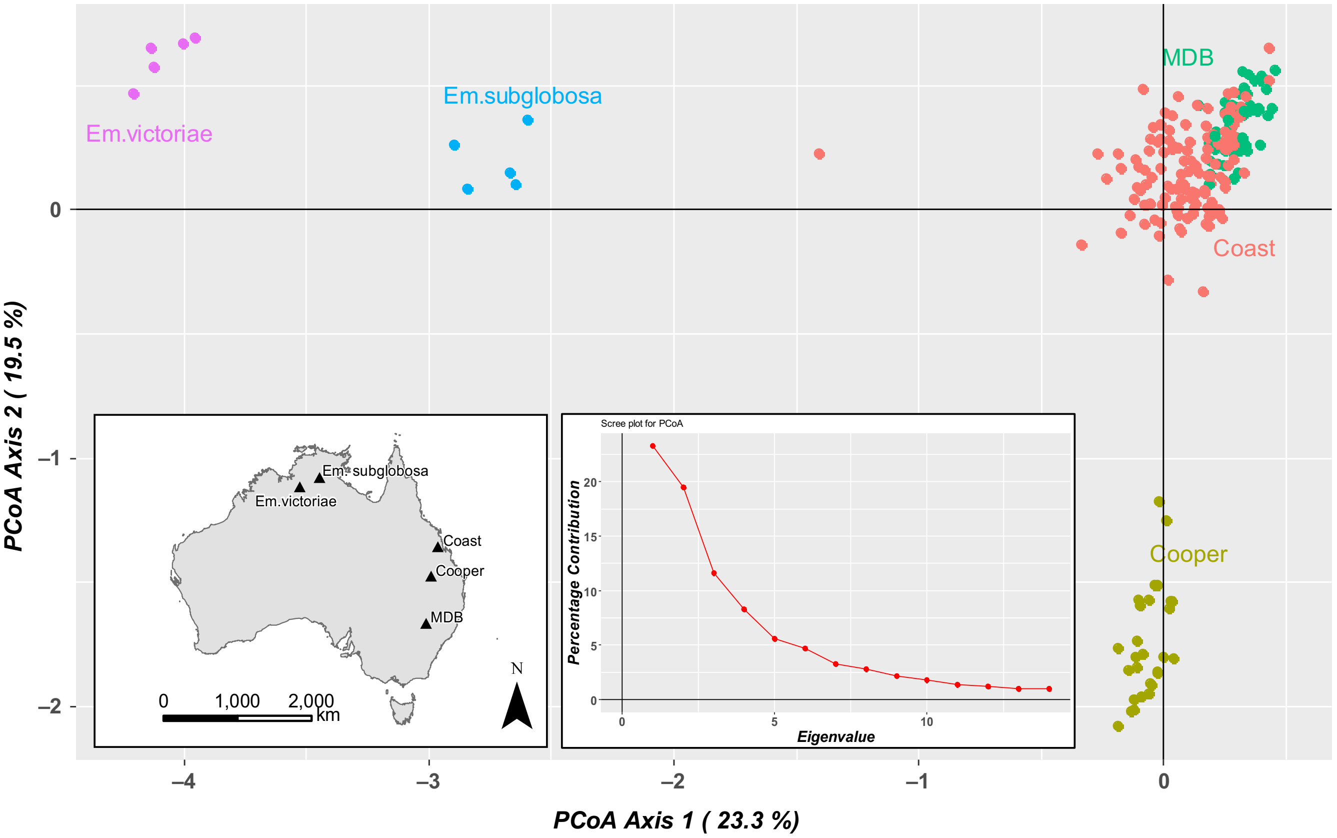

FIGURE 1 Sample PCoA plot and scree plot on the importance of eigenvalues of the test data set on fresh water turtles that is provided with the package. Please note the map insert of the physical locations of samples in Australia is added manually

alleles at each locus (loc.metric: RepAvg). To filter stringently on this quality statistic, insisting on perfect reproducibility, you would use the command

```
gl <- gl.filter.repavg(gl, t=1.00)
```

With non-DArT data, the user simply can provide the quality metric, add it to the genlight object in the right slot and use the function as usual.

```
gl@other$loc.metrics$repAvg <-
quality.metric.for.each.locus
gl2 <- filter.repavg(gl, t=0.95)
```

An overview on additional functions for filtering is given in Table S1. Note that for each filter function, there is a companion report function (e.g., gl.report.repavg) which summarizes the quality statistic, but makes no change to the target genlight object. This may be useful for informing judgements on the threshold to use in filtering.

4.2 | Subsetting and regrouping

The input data and associated individual metadata file provide for initial assignment of individual labels and populations, and connection back to the original extractions and tissue samples. Subsequently, there is often a need to reassign labels to individuals, to populations or to aggregate populations. Subsetting the data set may be necessary, by deleting individuals or populations. These actions can be achieved using DARTR recoding functions which draw upon csv recode tables (old label, new label). If the keyword “Delete” is used as the new label, the population or individual is removed from the genlight object.

For the more experienced user, the conventional R-syntax to subset data sets using the indexing function “[]” can be used. Functions in DARTR for recoding populations and individuals are shown in Table S1, and we provide examples on subsetting and recoding in our accompanying vignette.

4.3 | Isolation by distance

A common initial way to analyse genetic data is to check for isolation by distance. The approach aims to measure and test the relationship between genetic differentiation and geographic distance at neutral loci (Slatkin, 1993; Wright, 1943). Measurements of genetic differentiation and distances can be based on individuals or subpopulations. We implemented an isolation-by-distance analysis on the genlight object following the approach of Rousset (1997). If coordinates are provided as part of the individual metadata, the function gl.ibd() with default parameters set reprojects those coordinates from geographic coordinates (lat/long) to distances in metres (using the Mercator projection) and plots F_{st}/

$1 - F_{st}$ against log(geographic distance). Parameters can be set allowing any genetic or environmental distance (e.g., least-cost distances calculated via the package POPGENREPORT; Gruber & Adamack, 2015) to be used. A Mantel test based on bootstrapping is performed to test for significance of the association of both matrices. The function

```
gl.ibd(gl)
```

returns the pairwise genetic and Euclidean distance matrices and statistics of a Mantel test.

4.4 | Fixed difference analysis

If the interest of the study is in spatial population differentiation, there is some advantage in considering only fixed differences between populations, that is, allelic differences where the alleles have come to fixation to alternative states in populations taken pairwise (Davis & Nixon, 1992; Georges & Adams, 1996). A fixed difference between two populations at a specific locus occurs when the populations share no alleles at that locus. Gene frequencies may ebb and flow, but once a locus becomes fixed for an allele or suite of alleles, there is no returning. The accumulation of fixed differences between two populations is considered a robust indication of lack of gene flow. In a nutshell, fixed differences are summed over populations taken pairwise, and when two populations have no fixed differences (or insubstantial fixed differences), the populations are amalgamated and the process repeated until there is no further reduction (Georges & Adams, 1996). The final set of taxa are diagnosable by the presence or absence of a set of alleles at multiple loci. The script `gl.collapse.recursive(gl,t=0)` will ultimately yield a grouping of aggregated populations that are diagnosable from each other by one or more fixed allelic differences.

With the large number of loci typically generated in SNP data sets, there is a risk of generating false positives, that is, fixed differences arising by chance in the samples when they do not occur in the populations from which the samples are drawn. It is difficult to calculate the expected frequency of false positives for two samples of a given size without detailed knowledge of the allele frequency distribution of the populations from which they were drawn, and there is the issue of compounding error. However, the probability of a false positive becomes vanishingly small regardless of these two influences provided samples sizes for each population are ten or more ($2n = 20$). This should be a target for sampling intensity, and manual amalgamation of populations before fixed difference analysis should be considered where sample sizes are below 5. Two parameters in `gl.collapse.recursive` provide control over sample size. Parameter `tpop` sets the number of fixed differences that are tolerated when amalgamating two populations. The default is `tpop=0`, but `tpop=1` is recommended for corroborated fixed differences. The second parameter is `nlimit` (default=2), which is the combined sample size of the two populations being compared that is required for an assessment of fixed differences. One might choose for example, to set `nlimit=10` to ensure an adequate sample size, taking into account missing values, in each of the paired comparisons.

Definition of an absolute fixed difference can be relaxed to allow fixed differences to be defined at some specified level of allele frequency, say 0.05 to score two populations with SNP allele frequencies of 95:5% vs. 5:95% to be regarded as fixed. This enables examination of structure among populations using allele frequencies that have come nearly, but not yet, to fixation.

4.5 | Population assignment

Assigning individuals of unknown provenance to populations of known provenance is a challenging exercise, and several approaches have been suggested (Blanchong, Scribner, & Winterstein, 2002; Götz & Thaller, 1998; Manel, Gaggiotti, & Waples, 2005; Paetkau, Slade, Burden, & Estoup, 2004). A first approach is to eliminate from consideration those target populations where a SNP allele is present in the unknown individual but not in the target. When the unknown individual possesses such a private allele, the target population is unlikely to be the source population. This analysis can be performed with function `gl.report.pa`. In many cases, examining private alleles will narrow down the possible source populations considerably, and depending on the spatial resolution required for the assignment, may provide a satisfactory answer.

A second approach is to examine the position of the unknown individual relative to the target populations in a reduced ordinated locus space using PCoA. This graphic representation is provided by `gl.assign`. Addition of confidence ellipses then allows a decision to eliminate some populations from consideration as the source of the unknown individual. The converse is not true. This approach does not allow assignment of the unknown to populations that contain the unknown within their confidence ellipse. The overall confidence envelope is multidimensional, and separation of the unknown from a target population may occur in deeper dimensions. Hence, as with the private alleles approach, this graphical approach serves to narrow down the candidates for the source of the unknown and may, in that sense, provide a satisfactory answer.

A third approach is to eliminate from consideration those populations for which the unknown has private alleles, and then calculate the probability or likelihood of yielding the genotype of the unknown individual given the observed allele frequencies in each remaining target population. Using this approach, the individual is assigned notionally to those populations for which this probability is highest; populations for which the probability is lower than some level of significance are eliminated from further consideration. This approach was first applied in a study of microsatellite markers in bear populations (Paetkau et al., 2004) and subsequently applied using classical and Bayesian approaches to estimating probabilities (Blanchong et al., 2002; Götz & Thaller, 1998).

Unfortunately, the sheer number of SNPs generated by next-generation sequencing technologies, often in the 10s or 100s of thousands, makes the assumption of independence of the loci untenable. The nonindependence (linkage) of loci is problematic, and without additional genomic information, it is not possible to overcome this directly. In addition, the sample sizes used in studies of population assignment are typically small, and inappropriate for tests of sufficient power to identify loci that can be regarded as independent. This lack of independence leads also to a statistical problem, because independence is a prerequisite for combining the probabilities (or likelihoods) of the observed genotype at each locus as a simple product to yield an overall probability of assignment for the unknown genotype. Nevertheless, this approach can be used as an index by which to make a judgement on assignment, so long as the index is not regarded as an accurate estimate of probability of assignment.

A final option is a fourth approach that addresses the issue of nonindependence (linkage) among the SNP loci by ordinating the space defined by those loci. The resultant axes, linear combinations of the information contained in each locus, are orthogonal and so can be regarded as independent. Subsequent standardization can achieve independent and identically distributed variates, which simplifies analysis of probabilities and likelihoods. This new approach, outlined below, is now implemented in R—it is likely to have wide applicability.

The script `gl.assign` first eliminates populations on the basis of private alleles. It then ordinates the space defined in locus space for the remaining populations. A limited number of dimensions are retained in the final solution, and the decision is based on consideration of (i) the number of substantive eigenvalues (greater in explanatory power than the original variables before ordination), (ii) the number of populations including the unknown, (iii) an operational maximum number of dimensions as specified in the script (dim=7) or (iv) a user specified value. The script selects the minimum of these values to set the dimension of the reduced ordination space used subsequently.

A 95% confidence envelope (or any other level of confidence as specified by the user) is defined in the reduced ordinated space, and the likelihood of the unknown genotype occurring is estimated for each dimension under normal distribution assumptions. These likelihoods are logged for computational reasons, weighted by the eigenvalue for their respective dimension and summed to yield an Assignment Index for the unknown against each population. Summing the weighted logged likelihoods is supported by the independence of each of the ordinated axes, but the result should be nevertheless regarded as an assignment index rather than an accurate likelihood.

An Assignment Index is calculated in the same way for a notional individual residing on the boundary of the confidence envelope. Comparing the Assignment Index for each population with that of the notional boundary individual provides a basis for a decision on assignment. If the Assignment Index for the unknown is less than the critical value for the Assignment Index (that of the boundary individual), then the unknown is assigned to that population. Where more than one population is selected, the population with the greatest Assignment Index is the most likely.

4.6 | Hardy–Weinberg equilibrium and gametic-phase disequilibrium

Testing data for conformation to Hardy–Weinberg and gametic-phase disequilibrium expectations prior to analyses has been a cornerstone of population genetic analyses because many downstream analyses depend on those assumptions (Allendorf & Luikart, 2007; Hedrick, 2011). Yet, our observation is that these analyses are infrequently reported for SNP data sets. One explanation for this may be the high computational power required to undertake these analyses for typical SNP data sets, and the lack of user-friendly software packages geared to population genomic analyses where analyses must be conducted at the population level. The function gl.report.hwe and gl.filter.hwe can be used to test and then filter loci on the basis of meeting HWE expectations for each locus within each population or overall. Users can then evaluate the consistency of these departures across populations before deciding whether to exclude loci.

Tests for gametic-phase disequilibrium are a computational challenge because of the requirement to run an analysis on the pairwise linkage disequilibrium of all pairs of loci for every sampled population. We implemented a very fast and efficient version that takes advantage of the multicore architecture of modern processors. Although this kind of analysis was available before, it was almost impossible to implement for a normal user as it first required to subset the data set for all subpopulation and then run often several hundreds of thousands pairwise comparisons between loci. The function

```
gl.report.ld(gl)
```

returns a matrix that identifies loci under linkage disequilibrium for each subpopulation. Users can evaluate consistency of departures before deciding whether to exclude loci. Analyses of 2,000–5,000 SNP loci for populations of 20-30 individuals typically take 20-60 min per population on a 25-core cluster.

4.7 | Detecting loci under selection

High marker coverage of genomes also provides opportunities to detect signals of selection through the detection of outlier loci ["genome scans"; for a review see Excoffier, Hofer, and Foll (2009)]. A potential downstream analysis is to study the association of those loci with environmental and trait data (Hecht, Matala, Hess, & Narum, 2015). Alternatively, many analyses assume markers are neutral, and selected markers must be removed prior to data analysis. A variety of approaches to detecting signals of selection have been developed. Lotterhos and Whitlock (2015) showed that their statistical motivated approach fitting a curve to the distribution of a trimmed subset of neutral loci provides the most reliable identification of loci experiencing directional selection, and implemented the method in the R package Outflank [(Lotterhos & Whitlock, 2015), available on GitHub]. We re-implemented the existing function in our package (with permission of the authors) to link it to genlight objects. The code for such an analysis now simplifies to:

```
gl.outflank(gl)
```

This function returns the same output as the original function implemented by Lotterhos and Whitlock (2015).

5 | EXPORTING

Although there are a number of analysis available within R, it is often desirable to run specialized software that require often a quite particular, proprietary format. For those cases, we provide several export functions that create data files for use as input for those programs. The package provides an export function to convert genlight objects to be used for the following software programs: STRUCTURE (Pritchard, Stephens, & Donnelly, 2000), FASTSTRUCTURE (Raj, Stephens, & Pritchard, 2014), NEWHYBRIDS (Anderson & Thompson, 2002).

In addition, we provide a function that exports the data set in the commonly required FASTA format if sequence information is available. Here, we implemented four different versions that output the actual data set as the concatenated full sequences with ambiguity codes, majority codes in the case of missing data and only the SNP positions. This function needs to have information on the sequence and on the type and position of the SNP within the sequence. If DArT data are used, the genlight object already contains this information in the right slots. If data sets from other sources are used, the user needs to put the information into the genlight object. The vignette provided with the package gives a detailed example how to achieve this.

Finally, there are three important R packages commonly used to analyse population genomic data sets that use their own format: APE, DEMERELATE, and SNPRELATE. The DARTR package provides functions to convert a genlight object into data structures suitable for those packages. In addition, we provide an internal conversion from genlight to genind and vice versa to allow the interchange between those formats. For example, the package POPGENREPORT (Adamack & Gruber, 2014) calculates landscape resistance matrices and conducts isolation–by-distance analysis based on genind objects, and the package MMOD (Winter, 2012) calculates a variety of genetic distance metrics also based on the genind format. These conversion functions facilitate the construction of a single R workflow concatenating functions from several popular R packages.

6 | DISCUSSION

In developing the DARTR package, our primary aim was to make it simple for geneticists unfamiliar with R to capitalize on the power and flexibility of this rapidly growing platform to analyse large SNP data sets for the study of population genomic and phylogenomic problems. Although several packages addressing these types of analyses exist, none provide a single framework where a comprehensive analysis can be completed starting from data quality assessment to fundamental analyses common to many research problems. DARTR achieves this by a combination of newly developed functions, and by packaging functions available in other packages but often based on unique data formats, which are more difficult to integrate into a single workflow and may lead to conversion errors. An often overlooked but very important step for the beginner is being able to filter their data set in terms of quality and on the basis of individual metrics such as population identifier, location and other grouping variables. DARTR provides more than 20 new filtering functions to conveniently subset SNP data. Moreover, DARTR also provides functions and types of analysis that have not been implemented before in R (fixed difference analysis, population assignments, fast linkage disequilibrium per population). An overview of all available functions is given in Table S1. In addition, DARTR permits an efficient data processing workflow by utilizing a single standard format—the genlight data format. The genlight format has several very useful attributes—namely, it is very compact allowing efficient data storage, and core functions are programmed using C, which is one of the fastest computer languages. Finally, the genlight structure is expandable via the @other slot which permits additional metadata to be linked to genotypes, and loci, and therefore a straight forward procedure to subset data sets and there are already a number of pathways to convert existing formats into a genlight object. This makes genlight the prime candidate to become the de facto standard to analyse SNP data, and we advocate its use for the development of new population genomic methods.

ACKNOWLEDGEMENTS

We would like to thank all beta-testers from various universities for their valuable feedback on previous versions of the manuscript, package and tutorial, and especially Jason Bragg, Paul Sunnucks, Alexandra Pavlova, Terry Bertozzi, Alan Couch, David Berman, Pierre Feutry, Rachel Binks and Jim Underwood. Aaron T. Adamack and Anthony Davidson assisted and provided code for additional functions.

DATA ACCESSIBILITY

The current version of the DARTR package can be downloaded and installed via CRAN R repository [`install.packages("dartR")`]. The latest development version is hosted on GitHub under: https://github.com/green-striped-gecko/dartR accompanied by a detailed description how to install the latest version and a change log. Errors, feature requests and contributions should be submitted via the GitHub repository.

AUTHOR CONTRIBUTIONS

All authors contributed to the general design of the package. The package was mainly coded by B.G. and A.G., and all authors contributed to the preparation of the manuscript.

ORCID

Bernd Gruber http://orcid.org/0000-0003-0078-8179
Arthur Georges http://orcid.org/0000-0003-2428-0361

REFERENCES

Adamack, A. T., & Gruber, B. (2014). POPGENREPORT: Simplifying basic population genetic analyses in R. *Methods in Ecology and Evolution*, *5*, 384–387. https://doi.org/10.1111/2041-210X.12158

Allendorf, F. W., & Luikart, G. (2007). *Conservation and the genetics of populations*. Malden, MA: Wiley-Blackwell.

Anderson, E. C., & Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, *160*, 1217–1229.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*, 81–92. https://doi.org/10.1038/nrg.2015.28

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

Beheregaray, L. B., Pfeiffer, L. V., Attard, C. R. M., Sandoval-Castillo, J., Domingos, F. M. C. B., Faulks, L. K., … Unmack, P. J. (2017). Genome-wide data delimits multiple climate-determined species ranges in a widespread Australian fish, the golden perch (*Macquaria ambigua*). *Molecular Phylogenetics and Evolution*, *111*, 65–75. https://doi.org/10.1016/j.ympev.2017.03.021

Blanchong, J. A., Scribner, K. T., & Winterstein, S. R. (2002). Assignment of individuals to populations: Bayesian methods and multi-locus genotypes. *The Journal of Wildlife Management*, *66*, 321. https://doi.org/10.2307/3803164

Brauer, C. J., Hammer, M. P., & Beheregaray, L. B. (2016). Riverscape genomics of a threatened fish across a hydroclimatically heterogeneous river basin. *Molecular Ecology*, *25*, 5093–5113. https://doi.org/10.1111/mec.13830

Davis, J. I., & Nixon, K. C. (1992). Populations, genetic variation, and the delimitation of phylogenetic species. *Systematic Biology*, *41*, 421–435. https://doi.org/10.1093/sysbio/41.4.421

Excoffier, L., Hofer, T., & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity*, *103*, 285–298. https://doi.org/10.1038/hdy.2009.74

Georges, A., & Adams, M. (1996). Electrophoretic delineation of species boundaries within the short-necked freshwater turtles of Australia (Testudines: Chelidae). *Zoological Journal of the Linnean Society*, *118*, 241–260. https://doi.org/10.1111/j.1096-3642.1996.tb01266.x

Glover, K., Hansen, M., Lien, S., Als, T., Hoyheim, B., & Skaala, O. (2010). A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics*, *11*, 2. https://doi.org/10.1186/1471-2156-11-2

Götz, K.-U., & Thaller, G. (1998). Assignment of individuals to populations using microsatellites. *Journal of Animal Breeding and Genetics*, *115*, 53–61. https://doi.org/10.1111/j.1439-0388.1998.tb00327.x

Goudet, J. (1995). FSTAT (version 1.2): A computer program to calculate F-statistics. *Journal of Heredity*, *86*, 485–486. https://doi.org/10.1093/oxfordjournals.jhered.a111627

Gruber, B., & Adamack, A. T. (2015). LANDGENREPORT: A new R function to simplify landscape genetic analysis using resistance surface layers. *Molecular Ecology Resources*, *15*, 1172–1178. https://doi.org/10.1111/1755-0998.12381

Hecht, B. C., Matala, A. P., Hess, J. E., & Narum, S. R. (2015). Environmental adaptation in Chinook salmon (*Oncorhynchus tshawytscha*) throughout their North American range. *Molecular Ecology*, *24*, 5573–5595. https://doi.org/10.1111/mec.13409

Hedrick, P. W. (2011). *Genetics of populations*. Burlington, MA: Jones & Bartlett Learning.

Jaccoud, D., Peng, K., Feinstein, D., & Kilian, A. (2001). Diversity Arrays: A solid state technology for sequence information independent genotyping. *Nucleic Acids Research*, *29*, e25. https://doi.org/10.1093/nar/29.4.e25

Johnston, S. E., McEwav, J. C., Pickering, N. K., Kijas, J. W., Beraldi, D., Pilkington, J. G., … Slate, J. (2011). Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Molecular Ecology*, *20*, 2555–2566. https://doi.org/10.1111/j.1365-294X.2011.05076.x

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

Jombart, T., & Ahmed, I. (2011). ADEGENET 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*, 3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., … Uszynski, G. (2012). Diversity arrays technology: A generic genome profiling technology on open platforms. *Methods in Molecular Biology*, *888*, 67–89. https://doi.org/10.1007/978-1-61779-870-2

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, *24*, 1031–1046. https://doi.org/10.1111/mec.13100

Manel, S., Gaggiotti, O. E., & Waples, R. S. (2005). Assignment methods: Matching biological questions with appropriate techniques. *Trends in Ecology and Evolution*, *20*, 136–142. https://doi.org/10.1016/j.tree.2004.12.004

Morin, P. A., Martien, K. K., & Taylor, B. L. (2009). Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, *9*, 66–73. https://doi.org/10.1111/j.1755-0998.2008.02392.x

Paetkau, D., Slade, R., Burden, M., & Estoup, A. (2004). Genetic assignment methods for the direct, real-time estimation of migration rate: A simulation-based exploration of accuracy and power. *Molecular Ecology*, *13*, 55–65. https://doi.org/10.1046/j.1365-294X.2004.02008.x

Paradis, E., Gosselin, T., Goudet, J., Jombart, T., & Schliep, K. (2017). Linking genomics and population genetics with R. *Molecular Ecology Resources*, *17*, 54–66. https://doi.org/10.1111/1755-0998.12577

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for De Novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*, e37135. https://doi.org/10.1371/journal.pone.0037135

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). FASTSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, *197*, 573–589. https://doi.org/10.1534/genetics.114.164350

Raymond, M., & Rousset, F. (1995). GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. *Journal of Hereditry*, *86*, 248–249. https://doi.org/10.1093/oxfordjournals.jhered.a111573

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, *145*, 1219–1228.

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*, 54. https://doi.org/10.1186/1753-6561-5-S7-P54

Santure, A. W., De Cauwer, I., Robinson, M. R., Poissant, J., Sheldon, B. C., & Slate, J. (2013). Genomic dissection of variation in clutch size and egg mass in a wild great tit (*Parus major*) population. *Molecular Ecology*, *22*, 3949–3962. https://doi.org/10.1111/mec.12376

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2017). PLOTLY: *Create interactive web graphics via 'plotly.js'*. R package version 4.6. 0.

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, *47*, 264–279. https://doi.org/10.1111/j.1558-5646.1993.tb01215.x

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., ... Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLoS Biology*, *13*, 1–11.

Swaegers, J., Mergeay, J., Van Geystelen, A., Therry, L., Larmuseau, M. H. D., & Stoks, R. (2015). Neutral and adaptive genomic signatures of rapid poleward range expansion. *Molecular Ecology*, *24*, 6163–6176. https://doi.org/10.1111/mec.13462

van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., ... Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, *5*, 247–252. https://doi.org/10.1038/nmeth.1185

Unmack, P. J., Sandoval-Castillo, J., Hammer, M. P., Adams, M., Raadik, T. A., & Beheregaray, L. B. (2017). Genome-wide SNPs resolve a key conflict between sequence and allozyme data to confirm another threatened candidate species of river blackfishes (Teleostei: Percichthyidae: Gadopsis). *Molecular Phylogenetics and Evolution*, *109*, 415–420. https://doi.org/10.1016/j.ympev.2017.02.013

Winter, D. J. (2012). MMOD: An R library for the calculation of population differentiation statistics. *Molecular Ecology Resources*, *12*, 1158–1160. https://doi.org/10.1111/j.1755-0998.2012.03174.x

Wright, S. (1943). Isolation by distance. *Genetics*, *28*, 139–156.

Ytournel, F., Bedõhom, B., Gut, I., Lathrop, M., Weigend, S., & Simianer, H. (2011). Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations. *Animal Genetics*, *43*, 419–428.

Zegura, S. L., Karafet, T. M., Zhivotovsky, L. A., & Hammer, M. F. (1999). High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Molecular Biology and Evolution*, *21*, 164–175.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Gruber B, Unmack PJ, Berry OF, Georges A. DARTR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Mol Ecol Resour*. 2018;00:1–9. https://doi.org/10.1111/1755-0998.12745