Received: 30 August 2020 | Accepted: 12 January 2021

DOI: 10.1111/1755-0998.13339

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

SambaR: An R package for fast, easy and reproducible population-genetic analyses of biallelic SNP data sets

Menno J. de Jong[1,2] | Joost F. de Jong[3] | A. Rus Hoelzel[1] | Axel Janke[2,4,5]

[1]Department of Biosciences, Durham University, Durham, UK

[2]Biodiversity and Climate Research Centre, Senckenberg Institute, Frankfurt am Main, Germany

[3]Wildlife Ecology and Conservation Group, Wageningen University, Wageningen, The Netherlands

[4]Institute for Ecology, Evolution and Diversity, Goethe University, Frankfurt am Main, Germany

[5]LOEWE Centre for Translational Biodiversity Genomics, Frankfurt am Main, Germany

Correspondence
Menno J. de Jong and A. Rus Hoelzel, Department of Biosciences, Durham University, South Road, Durham, DH1 3LE, UK.
Email: Menno.de-Jong@senckenberg.de and a.r.hoelzel@dur.ac.uk

Funding information
LOEWE centre for translational biodiversity genomics (TBG); Kenneth Whitehead Trust; Leibniz-Gemeinschaft; British Deer Society

Abstract

SNP data sets can be used to infer a wealth of information about natural populations, including information about their structure, genetic diversity, and the presence of loci under selection. However, SNP data analysis can be a time-consuming and challenging process, not in the least because at present many different software packages are needed to execute and depict the wide variety of mainstream population-genetic analyses. Here, we present SambaR, an integrative and user-friendly R package which automates and simplifies quality control and population-genetic analyses of biallelic SNP data sets. SambaR allows users to perform mainstream population-genetic analyses and to generate a wide variety of ready to publish graphs with a minimum number of commands (less than 10). These wrapper commands call functions of existing packages (including adegenet, ape, LEA, poppr, pcadapt and StAMPP) as well as new tools uniquely implemented in SambaR. We tested SambaR on online available SNP data sets and found that SambaR can process data sets of over 100,000 SNPs and hundreds of individuals within hours, given sufficient computing power. Newly developed tools implemented in SambaR facilitate optimization of filter settings, objective interpretation of ordination analyses, enhance comparability of diversity estimates from reduced representation library SNP data sets, and generate reduced SNP panels and structure-like plots with Bayesian population assignment probabilities. SambaR facilitates rapid population genetic analyses on biallelic SNP data sets by removing three major time sinks: file handling, software learning, and data plotting. In addition, SambaR provides a convenient platform for SNP data storage and management, as well as several new utilities, including guidance in setting appropriate data filters. The SambaR source script, manual and example data set are distributed through GitHub: https://github.com/mennodejong1986/SambaR.

KEYWORDS
gene flow, genetic diversity, population assignment test, population genetics, R package, selection analyses, SNP data

1 | INTRODUCTION

Modern-day population geneticists risk spending as much time studying computer software as studying their actual scientific questions. They also risk spending as much time generating plots as generating new data. These time sinks can negatively affect the quality of research outcomes, as they eat away time needed for (a) understanding the theoretical underpinnings of analysis methods and (b) interpretation of analysis outcomes.

Integration of computer programs into one single software pipeline removes the necessity of getting acquainted with the technicalities of each program and therefore promotes increased efficiency and, by avoiding incorrect usage, increased accuracy. Efficiency will be improved further if this integrative software package automatically translates the results into ready-to-publish graphs. For two reasons a good candidate for such a wrapper and plotting software is an R package: many tools for population-genetic analyses are written in R (R Core Team, 2019), and R contains powerful graphing tools.

Here, we introduce the R package SambaR, which stands for: "Snp datA Management and Basic Analyses in R." SambaR is a collections of functions which increase the power of existing R tools for population-genetic analyses. SambaR aims to free users from the disproportionate time investment which currently is needed for tasks related to (a) managing input and output files, (b) learning the trivialities of computer software and (c) generating and polishing plots. SambaR automates the integrated usage of proven and widely used R packages for population genetic analyses and generates over 100 ready-to-publish graphs to depict data quality control and analyses outcomes. The pipeline consists of less than 10 commands, which suffice to perform a wide variety of population genetic analyses on SNP data sets, including quality control, population structure analyses, population differentiation analyses, genetic diversity analyses, and selection analyses. Users are guided through the workflow by an accompanying manual, as well as by built-in explicit error messages.

A major asset of SambaR is that the pipeline is designed with the aim to circumvent the trade-off between automation and customization. By default, SambaR runs most analyses using different methods and/or varying filter and parameter settings, allowing users to explore the data and parameter space and to choose appropriate filter settings. This way SambaR enables rapid data processing without taking relevant choices and decisions away from the users.

Apart from streamlining population-genetic analyses, SambaR is also meant to provide a convenient and user-friendly platform for SNP data management. SambaR stores the input data in three data objects. Analysis outcomes are added to these existing data objects, rather than stored in additional data objects. Output tables and plots are automatically exported to subdirectories, categorized by analysis type. In addition, SambaR contains tools which allow to subset (based on sample/locus names), subsample and intersect data sets (i.e., finding overlap between SNP data sets), and to detect small subsets of SNPs which are most informative with respect to population structure.

Here, we describe SambaR and test the software on previously published SNP data sets. We also discuss new tools implemented in SambaR, including: (a) output plots which can help users to optimize their filtering settings, (b) a Bayesian population assignment (BPA) test, (c) the "distinctive clustering-score," a metric for objective measurement of the distinctiveness of population clusters based on sample loadings on ordination axes, and (d) functions which extract and export reduced SNP panels of various sizes.

2 | MATERIALS AND METHODS

2.1 | Technical details

SambaR is implemented as an R package and can run on any operating system. The software has been tested on Windows, Linux and Mac computers. SambaR will install up to 2 GB of dependencies (i.e., other R packages needed by SambaR for plotting and data analysis). Due to this dependency on other packages, for full use SambaR requires recent R versions (currently 4.0.0 or higher).

2.2 | SambaR pipeline

The SambaR pipeline consists of seven main functions (Figure 1, Table S1):

- The "getpackages" function installs and downloads dependencies. Optionally users can edit an automatically generated control file ("mypackageslist.txt") to prevent SambaR from attempting to install certain packages. The control file classifies packages into three categories: "essential," "recommended," and "optional." Essential packages are required for SambaR to run without errors. Recommended packages are needed for key analyses.
- The "importdata" function uses the read.PLINK function of the adegenet package (Jombart, 2008; Jombart & Ahmed, 2011) to import a SNP data set (from binary PED/MAP format) into R and to store this data as a genlight object (Jombart, 2008) named "mygenlight." Sample-specific and locus-specific information are stored in two auxiliary dataframes called "inds" and "snps," respectively (Figure 1). The function will incorporate in the "inds" dataframe sample information provided in an optional sample file. The function will also incorporate in the "snps" dataframe read depth and positional information found in optionally provided vcftools and STACKS output files. Monomorphic sites present in the input data file will be excluded from subsequent analyses.
- The "filterdata" function executes quality control. This function adds to the "inds" and "snps" dataframe boolean vectors (i.e., inds$filter, snps$filter and snps$filter2) which determine which samples and loci are included in subsequent analyses (Figure 1). Current filter options include: missing data per locus, missing data per sample, minor allele count per locus, locus specific deviation from HWE, read depth per locus, read depth per sample,

Sambar workflow

Load SambaR into R
source('C:/path/to/dir/SAMBAR.txt')
SambaR can be downloaded from Github:
https://github.com/mennodejong1986/SambaR

Install and load dependencies
getpackages(mylib=NULL)

Import data (binary PED/MAP) into R
setwd('C:/path/to/workdir/')

importdata(inputprefix='prefix',samplefile=NULL, sumstatsfile=NULL,depthfile=FALSE)

Or instead convert from an existing genlight object:
genlight2sambar(genlight_object=NULL)

Filter data
filterdata(indmiss=0.25,snpmiss=0.1,min_mac=2 ,dohefilter=TRUE,snpdepthfilter=TRUE, min_spacing=500,nchroms=NULL,silent=TRUE)
Included: quality control, F, HWE, relatedness

Analyze data

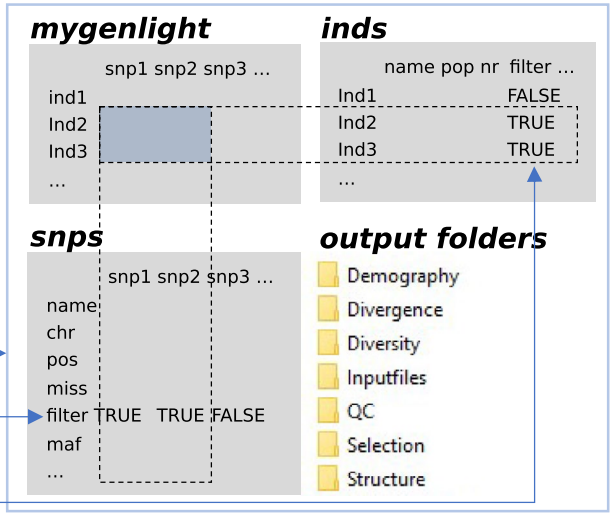
population structure
findstructure(Kmax=6,add_legend=TRUE, legend_pos='bottomright',legend_cex=3, symbol_size=3)
Included: PCoA, PCA, DAPC, CA, MDS, admixture analyses (LEA), Tess3r

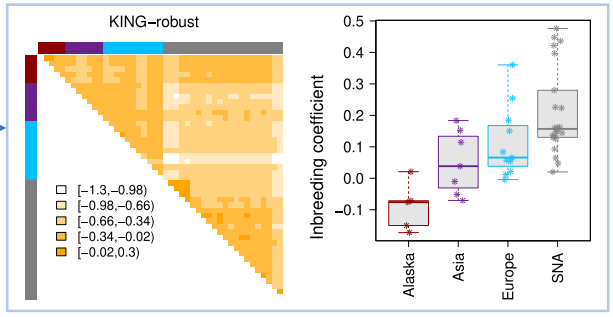
population differentiation
calcdistance(nchroms=NULL)
Included: Nei's D, (sliding window) Wright's Fst, Nei's Fst, Weir & Cockerham Fst, ABBA-BABA

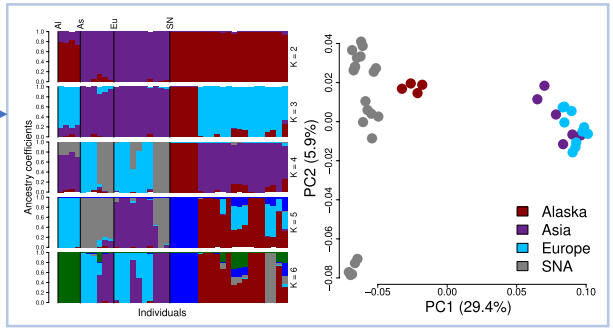
genetic diversity
calcdiversity(nrsites=NULL,legend_cex=2.5)
Included: SFS, heterozygosity, nucleotide diversity, Watterson's theta, Tajima's D, private alleles, proportion segregating sites

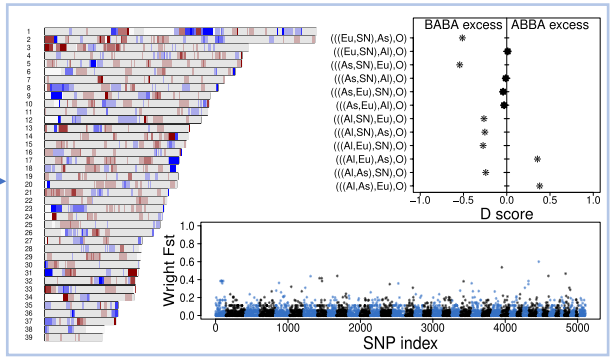
selection analyses
selectionanalyses(do_pcadapt=TRUE,do_outflank =TRUE,do_fsthet=TRUE,export='pdf')
Included: GWDS, FstHet, OutFLANK, PCadapt

Sambar output









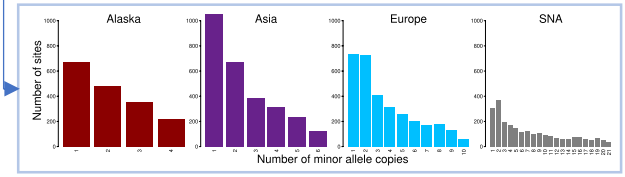


FIGURE 1 Schematic overview of the SambaR pipeline. [Colour figure can be viewed at wileyonlinelibrary.com]

transitions versus transversions, and, if genome locations are provided by the user, spacing between SNPs. Relatedness between samples is estimated by the kinship coefficient (Waples et al., 2018) and by the KING-robust measure (Waples et al., 2018).

- The "findstructure" function uses various R packages to perform principal components analysis (PCA) and principal coordinates analysis (PCoA), multidimensional scaling (MDS), discriminant analysis of principal components (DAPC, Jombart et al., 2010), correspondence analyses (CA), admixture analyses (using the R package LEA), and in addition generates structure-like plots with Bayesian population assignment (BPA) probabilities (see below). PCoA analyses are performed on three different types of genetic distance estimates: Nei's genetic distance, Hamming's genetic distance, and pairwise sequence dissimilarity. If the user provides sample locations (i.e., geographical coordinates), SambaR also generates geographical maps and in addition runs Tess3r (Caye et al., 2018) to perform tessellation analyses.
- The "calcdistance" function generates population differentiation measures for all pairwise population comparisons. These estimates include D_{xy}, $F_{ST\pi}$ (Hudson et al., 1992), Nei's genetic distance and Weir and Cockerham (1984) F_{ST}, the latter two generated using functions of the StAMPP package (Pembleton et al., 2013). In addition to genome wide estimates, the function also generates locus specific F_{ST} estimates, using three different metrics (Nei et al., 1977; Cockerham & Weir, 1987; Wright, 1943, Supporting Information Methods). D statistics are generated using ABBA-BABA calculations described in Durand et al. (2011).
- The "calcdiversity" function performs 1D and 2D site frequency spectrum (SFS) analyses, calculates nucleotide diversity and pairwise sequence dissimilarity estimates, and screens the genome for runs of homozygosity (using the R package detectRUNS, Biscarini et al., 2018). If users provide the number of chromosomes to the nchroms flag (i.e., number of biggest scaffolds to include), SambaR will in addition generate karyotype plots (Gel & Serra, 2017) showing genome wide variation.
- The "selectionanalyses" function uses the R packages Fsthet (Flanagan & Jones, 2018), OutFLANK (Whitlock & Lotterhos, 2015), PCadapt 4.1.0 (Luu et al., 2017, 2019) and GWDS (De Jong et al., 2021) to search for SNPs under balancing or diversifying selection. The function also executes Fisher's exact tests for associations between allele frequencies and population assignment.

2.3 | Plotting

During execution of SambaR's main functions, results are automatically exported into ready-to-publish plots in four different file formats: eps, pdf, png, and, depending on the operating system, wmf. Layout settings, including font type, font size and colour coding matching population assignment, are coherent. Plots are generated with various settings allowing users to select plots according to personal preferences. Function arguments allow users to customize colour coding and font type, as well as the size and location of the legend. Default font and symbol sizes ensure readability even if plots are scaled down. Output files are stored in subdirectories named QC, Structure, Divergence, Diversity, Demography, Selection and Inputfiles. These subdirectories are located within a main directory called SambaR_output.

Several subsets of plots are automatically combined by SambaR into multitile figures and exported in the pdf format. For more advanced R users, SambaR provides a function to create custom multitile figures with user defined combinations of SambaR plots.

2.4 | List of R packages used by SambaR

Currently SambaR uses the following R packages to perform population-genetic analyses: adegenet (Jombart, 2008; Jombart & Ahmed, 2011), ape (Paradis & Schliep, 2018), detectRUNS (Biscarini et al., 2018), FactoMineR (Lê et al., 2008), Factoextra (Kassambara & Mundt, 2019), HybridCheck (Ward & Van Oosterhout, 2016), LEA (Frichot & François, 2015), OutFLANK (Whitlock & Lotterhos, 2015), pcadapt (Luu et al., 2017, 2019), poppr (Kamvar et al., 2014), StAMPP (Pembleton et al., 2013), qvalue (Storey et al., 2019), tess3r (Caye et al., 2018), SNPRelate (Zheng et al., 2012), and zoo (Zeileis & Grothendieck, 2005).

For plotting, Sambar makes use of the R packages: circlize (Gu et al., 2014), colorspace (Zeileis et al., 2019), gplots (Warnes et al., 2019), grid (Murrell, 2005), gridGraphics (Murrell & Wen, 2019), gridExtra (Auguie, 2017), karyoploteR (Gel & Serra, 2017), mapplots (Gerritsen, 2018), migest (Abel, 2019), plot3D (Soetaert, 2017), plyr (Wickham, 2011), RColorBrewer (Neuwirth, 2014), raster (Hijmans, 2019), rworldmap (South, 2011), scales (Wickham & Seidel, 2019), scatterplot3D (Ligges & Mächler, 2003), VennDiagram (Chen, 2018) and vioplot (Adler & Kelly, 2019).

2.5 | Analyses outside of the R environment which are supported by SambaR

SambaR also facilitates the usage of software outside of R. These include Admixture (Alexander et al., 2009), Bayesass (Mussmann et al., 2019), Bayescan (Foll & Gaggiotti, 2008), GCTA (Yang et al., 2011); PLINK (Chang et al., 2015; Gaunt et al., 2007; Purcell et al., 2007; for linkage disequilibrium, inbreeding and relatedness calculations) and Stairwayplot (Liu & Fu, 2015). SambaR does so by creating input files for these programs, such as site frequency spectrum vectors needed for Stairwayplot, and by generating plots from their output files.

2.6 | Highlighted features of SambaR

In the sections above we described the SambaR pipeline. In the following we will highlight particular features of SambaR, which include population-genetic tools uniquely implemented in SambaR.

2.6.1 | Data filtering recommendations

As outlined above, the main purpose of SambaR's "importdata" function is to import SNP data into R. In addition, the "importdata" function also generates plots which users can consult for choosing filter settings (with regard to levels of missing data) appropriate for their research questions.

Users are recommended to execute population structure analyses with a strict "snpmiss" filter that sets the maximum proportion of missing data of retained SNPs close to zero. This prevents distortion of ordination plots due to variation in levels of missing data between samples (Figure 2, Figures S1 and S2). The strictness of the SNP filter is however limited by the quality of the data, because a sufficient number of retained SNPs are needed to discern population structure. The "Data_quality"-plot (Figure 2a) shows the number of retained SNPs as a function of missing data thresholds, and there by allows users to choose the minimum threshold that is needed to retain the desired number of SNPs.

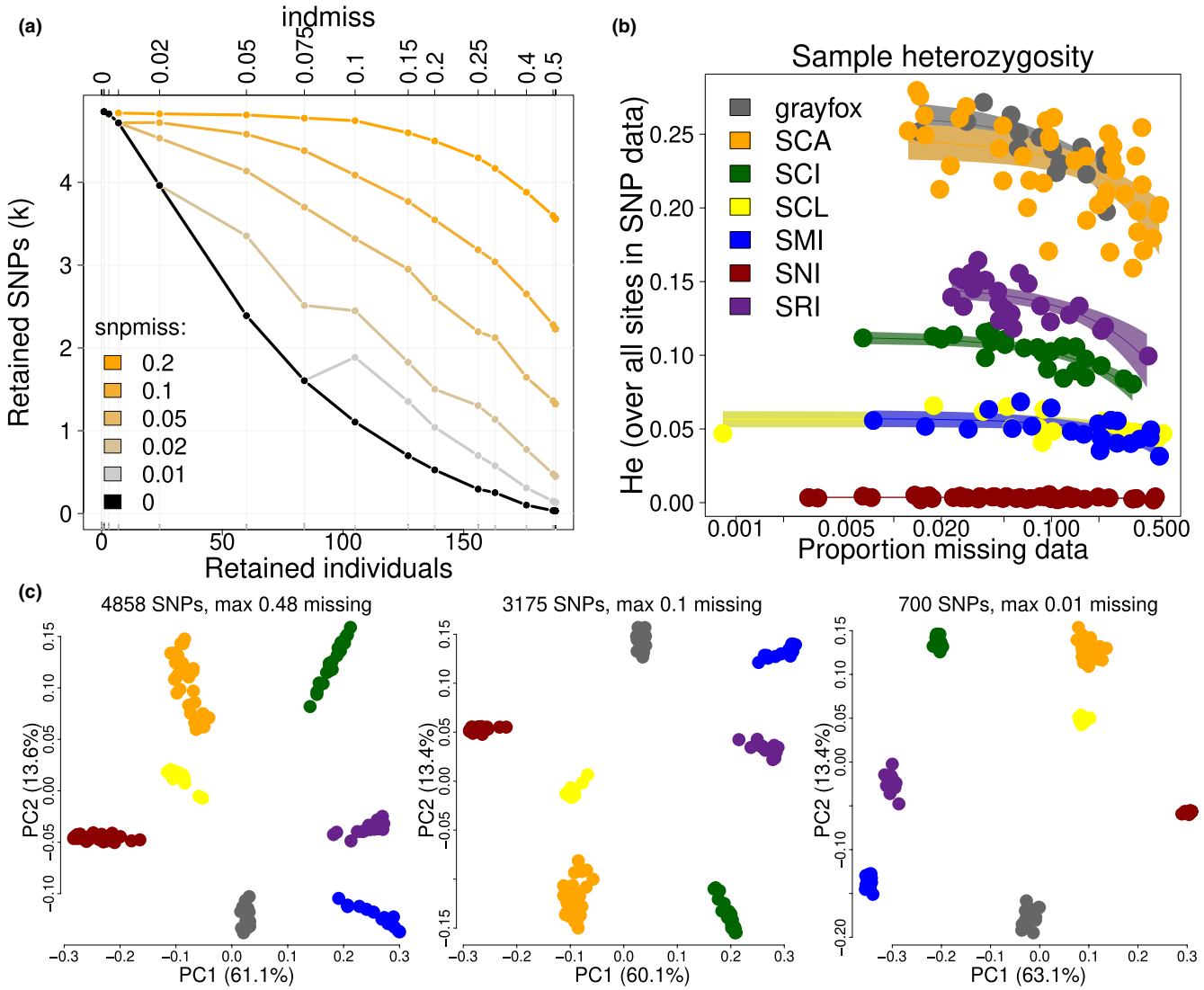


FIGURE 2 SambaR guides users choosing appropriate filter settings. SambaR output plots visualise the effect of filter settings on analyses outcomes, and thereby guide users towards selecting filter settings which return unbiased results, illustrated here using a RADseq data set of the island channel fox and the closely related mainland grey fox (Funk et al., 2016). (a) SambaR output plot depicting the number of retained SNPs and retained samples as a function of filter settings. Snpmiss stands for the maximum allowed proportion of missing data points per SNP. Indmiss stands for the maximum allowed proportion of missing data points per individual/sample). (b) SambaR output plot depicting sample heterozygosity against sample levels of missing data. Inclusion of samples with high proportions of missing data leads to underestimates of genetic diversity. (c) SambaR output plot depicting the outcome of principal coordinate analyses based on Hamming's genetic distance and using different SNP filter settings. Inclusion of SNPs with high levels of missing data leads to distorted PCoA plots, in which samples with high levels of missing data cluster towards the centre of the plot and samples with low levels of missing data cluster towards the plot edges. Colour coding according to Funk et al. (2016). The channel fox island populations are: Santa Catalina Island (sca, orange), Santa Cruz Island (sci, green), San Clemente Island (sli, yellow), San Miguel Island (smi, blue), San Nicolas Island (sni, red), and Santa Rosa Island (sri, purple). [Colour figure can be viewed at wileyonlinelibrary.com]

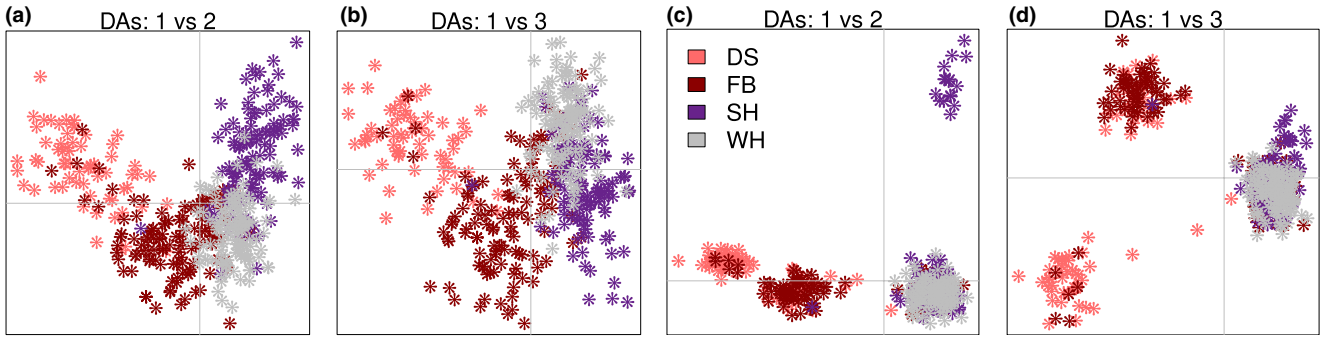


FIGURE 3 SambaR generates output plots for a range of settings, but leaves the final decision to the user. SambaR circumvents the trade-off between automation and customization by running analyses for a range of parameter settings. Afterwards, users can compare the output plots and select the output plots corresponding to the most appropriate setting. This exploration of parameter space is most exhaustive for DAPC analyses. By default, SambaR generates >50 output plot which depict the outcome of DAPC analyses given different combinations of parameter settings (including number of clusters, number of retained principal components, and inclusion of a priori population assignment information). The exploration of DAPC parameter space is here illustrated by a small subset of SambaR output plots for DAPC analyses on a SNP data set of 414 North American polar bears (Viengkone et al., 2017). (a) DAPC-plots depicting ordination axes 1–2 for K = 4, with a priori defined population assignment, and for 33 retained principal components (20% explained variance). B. Idem, but for ordination axes 1–3. (c) DAPC-plots depicting ordination axes 1–2 for K = 4, *without* a priori defined population assignment, and for 238 retained principal components (80% explained variance). (d) Idem, but for ordination axes 1–3. Comparison between plots makes the user aware that DAPC inferred clusters only agree with expected population structure if this information is provided a priori. Colour coding according to Viengkone et al. (2017). DS, Davis Strait (pink); FB, Foxe Basin (red); SH, Southern Hudson Bay (purple); WH, Western Hudson Bay (grey). [Colour figure can be viewed at wileyonlinelibrary.com]

For genetic diversity analyses, SambaR users are recommended to exclude samples for which the heterozygosity estimates are probably biased by relatively high proportions of missing data. These samples can be identified using the "Heterozygosity_vs_missingness" plot (Figure 2b).

SambaR performs population structure, diversity and differentiation analyses on a thinned data set containing maximum one SNP per 500 bp (default settings). In contrast, selection analyses are performed on a nonthinned data set, because the detection of linked outlier SNPs strengthens inference about selection events. Although it is common practice to filter SNP data sets based on linkage disequilibrium considerations, SambaR users interested in genetic diversity and selection analyses are recommended to not thin their data set prior to importing the data into R, unless because of size limitations. Full, nonthinned, data sets facilitate the generation of dense Manhattan plots.

2.6.2 | Pairwise sequence dissimilarity, nucleotide diversity (π), Watterson's theta and D_{xy}

The SambaR-function "calcpi," which is invoked by several main functions, calculates for each pair of individuals pairwise sequence dissimilarity estimates (Supporting Information Methods). These estimates are subsequently used to calculate several dependent population-genetic measures, including nucleotide diversity (π), Watterson's theta, Tajima's D, $F_{ST\pi}$ and D_{xy} (Supporting Information Methods).

SambaR generates estimates of genome wide diversity scores to facilitate comparisons of genetic diversity between SNP data sets. Users can enable this estimation by providing input value to the nr-sites flag of the "calcdiversity" function. This input value should be an estimate of the total number of sequenced sites in the filtered sequencing read data set from which the SNP data set is derived. The calculation assumes that users did not select at maximum one SNP per read (pair), filtered their genotype file prior to extracting biallelic SNPs (not vice versa), and did not thin their data based on linkage disequilibrium calculations.

2.7 | Analyses

Accurate execution of discriminant analyses of principle components (DAPC, Jombart et al., 2010), as implemented in the "adegenet" package, depends on several parameters. These parameters include the number of principal components to retain, the number of clusters (K), and the inclusion or exclusion of a priori population structure information. SambaR explores DAPC parameter space by generating DAPC plots for various combinations of number of retained principal components, number of retained clusters (by default two to six), and inclusion and exclusion of a prior population structure information (Figure 3). SambaR runs DAPC for five different values of retained principal components, one based on the a-score, and the other four corresponding to various percentages (i.e., 20, 50, 80 and 95%) of explained variance.

To guide users in selecting the most appropriate DAPC plot, SambaR generates the following summary statistics plots:

- a-score as a function of number of retained PCs, generated by the function optim.a.score() (Figure S3)

- The estimated number of successful predictions as a function of number of retained PCs (i.e., x-value for cross-validation, generated by the function xvalDapc(), Figure S3)
- Explained variance as a function of number of retained PCs (Figure S3)
- BIC-value as a function of the number of clusters (Figure S3)
- Heatmaps depicting the overlap between predefined populations and DAPC inferred clusters (Figure S4)

SambaR also performs a chi-squared test for the goodness of fit between a priori defined populations and DAPC inferred clusters for K equalling the number of a priori defined populations.

2.7.1 | Selection of most informative SNPs

Depending on the study system, a relatively low number of highly informative SNPs can be sufficient to assign individuals to populations (Von Thaden et al., 2020). These subsets of highly informative SNPs allow for low-cost determination of sample ancestry. Thus, there is a need for software to detect the most informative SNPs within a SNP data set. SambaR contains a function, invoked by the "findstructure" function, which detects SNPs with the highest standard deviation in population minor allele frequencies (given a predefined set of populations). Not included are SNPs for which data is lacking for one or more populations, nor SNPs of which the minor allele is missing in or more populations.

SambaR exports PED and MAP files of subsets of various sizes (50, 100, 150, and 250 SNPs) of these most informative SNPs, alongside estimates of population allele frequencies. SambaR also generates PCoA plots (Figure 4) and Bayesian population assignment (BPA) test plots (Figure S5) showing population structuring inferred from these reduced SNP panels.

2.7.2 | Bayesian population assignment test

SambaR contains a function, invoked by the "findstructure" function, which calculates for each individual the posterior probability that this individual belongs to a set of predefined populations, using as input the population minor allele frequencies and individual genotypes. The question addressed by this Bayesian population assignment (BPA) test is: among a set of predefined populations, which population is most likely to be the origin of

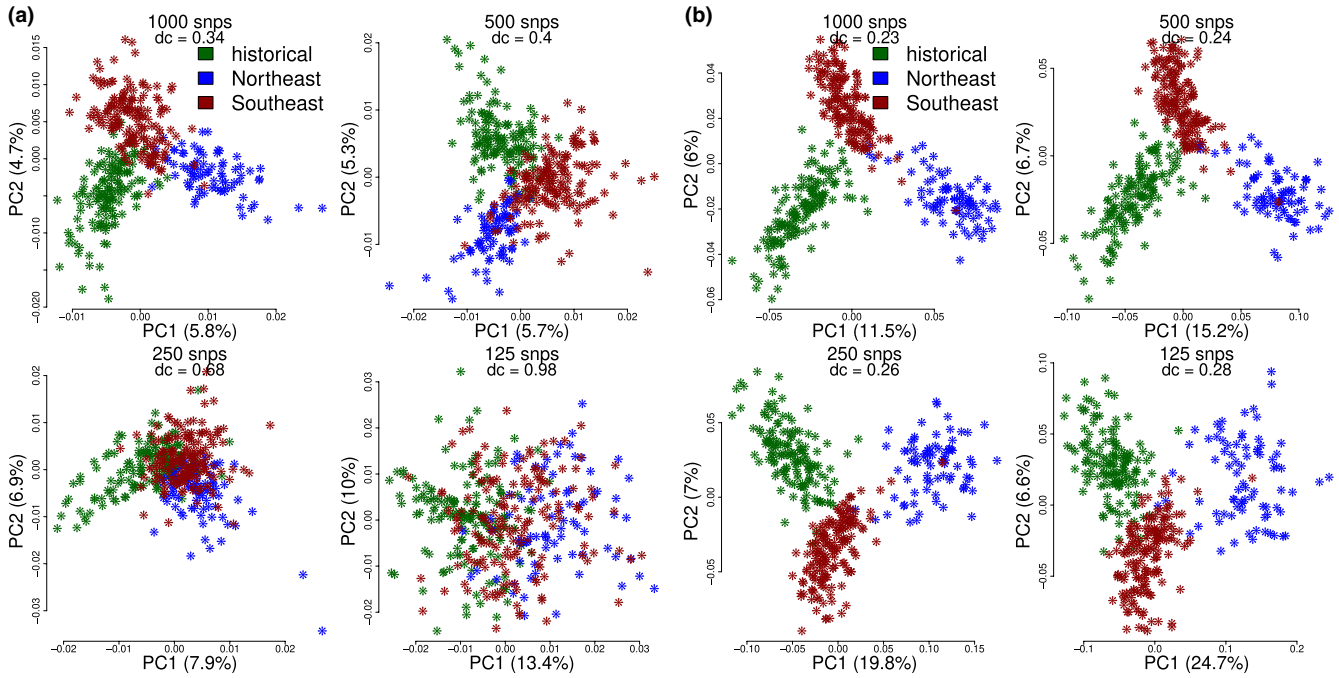


FIGURE 4 SambaR generates small subsets of highly informative SNPs, which could be used for low-cost population assignment. SambaR generates subsets of SNPs which are highly informative with regard to an a priori defined population structure, and which could be used for the design of SNP panels aimed at low-cost population assignment of samples of unknown origin. Furthermore, SambaR calculates a so called "dc-score," designed to objectively quantify the distinctiveness with which populations cluster away from each other in ordination plots. Low dc-scores are indicative of distinct clustering. The usefulness of these low information SNP panels and the dc-score is here illustrated for a ~ 22 K SNP data set of 394 coyotes (*Canis latrans*) sampled throughout the United States (Heppenheimer et al. 2018). (a) PCoA plots depicting population structure according to small subsets of randomly selected SNPs. Associated dc-scores are depicted above each plot. (b) PcoA plots depicting population structure according to small subsets of highly informative SNPs generated by SambaR. Associated dc-scores are depicted above each plot. Unlike subsets of randomly selected SNPs, small subsets of selected SNPs separate out populations distinctly. A subset of 125 informative SNPs generates a dc-score of 0.28, compared to 0.98 for a subset of 125 random SNPs. Colour coding according to Heppenheimer et al. (2018). [Colour figure can be viewed at wileyonlinelibrary.com]

a particular individual? The test is similar to previously published methods (Baudouin et al., 2004; Peatkau et al., 1995), with minor modifications (see Supporting Information Methods for more details).

Because the BPA test assumes independency of loci, SambaR performs the calculations on a thinned data set (if genomic locations of SNPs are provided). By default, the thinned data set includes maximum 1 SNP per 500 bp. This threshold is arbitrary and can be changed by the user when running the "filterdata" function.

Also excluded from the calculation are all SNPs for which one of either allele is missing in one or multiple populations, because these loci make the assignment probability converge to 0 or 1 immediately. A limitation of the BPA test is therefore that the test can only be applied to SNP data sets in which a sufficient number of SNPs have both alleles present in all populations. Depending on the study system, a few hundred biallelic SNPs might suffice (Figure S5).

The reliability of the BPA test depends on the precision of the allele frequency estimates, which in turn depends on sample sizes. SambaR users are therefore advised to exercise caution when interpreting the BPA test results for data sets with a small or uneven number of individuals per population. Reliable estimation of population allele frequencies generally requires 30 or more individuals per population (Fung & Keenan, 2014).

Another potential shortcoming of the BPA test is circular reasoning. This occurs if the population specific allele frequency estimates are calculated based on a data set which includes the individual for which the population assignment is being investigated. SambaR therefore recalculates population minor allele frequencies by excluding data for the investigated individual before running the BPA test (see Supporting Information Methods).

2.7.3 | dc-score

SambaR aims to facilitate an objective interpretation of ordination analyses by calculating the "dc-score," which we introduce here. The dc-score, or "distinct clustering"-score, measures the overlap between population clusters in a two-dimensional space defined by two ordination axes (by default the first and second). The score is calculated by dividing the mean Euclidian distance of samples from their population centre by the mean Euclidian distances between population centres (Figure S6, Supporting Information Methods). A dc-score close to zero indicates the absence of overlap between population clusters, whereas a dc-score greater than one indicates that the mean distance between population centres is smaller than the mean distance of samples to their population centre. Percentage of explained variance per ordination axis is not considered in the calculation. The dc-score is not a substitute to population differentiation measures such as F_{ST} and D_{xy}. The single purpose of the dc-score is to objectively quantify the clustering of samples within ordination plots, to avoid subjective statements such as: "populations clearly clustered separately."

3 | RESULTS AND DISCUSSION

3.1 | Data size limits and run time

To explore the data size limitations and run time of SambaR, we used SambaR to analyse online available whole genome sequencing (WGS) and reduced representation library (RRL) SNP data sets on three computers with different capacities. The findings indicate that the run time of SambaR, and whether it completes without encountering memory allocation errors, mainly depends on the capacities of the computer (Table S2). The run time estimates (Table S2) can provide guidance for users to match computational capacity to the data set in question, or alternatively to filter down data sets to computational capacity.

On High Performance Clusters, SambaR can process data sets of more than 100.000 SNPs and more than hundred individuals within hours. On ordinary desktop computers, data sets containing both over 200 K SNPs and over 100 individuals are likely to result in memory allocation errors, depending on the memory dimensions of the computer. Data sets of less than 100 K SNP and less than 100 individuals are typically processed in less than an hour on an average desktop computer (Table S2). Due to the data size limitations, SambaR cannot be applied to whole genome resequencing data sets (which typically contain millions of SNPs, even after stringent filtering), unless the data is thinned.

3.2 | Data filtering recommendations

We explored the effect of levels of missing data on the outcome of ordination analyses using a published RADseq SNP data set of the island channel fox (*Urocyon littoralis*, Funk et al., 2016). PCoA analyses plots based on Hamming's genetic distance using SNP data sets with a relaxed SNP filter threshold resulted in distorted ordination plots (Figure 2c), with sample loadings on ordination axes being a function of their proportion of missing data points (Figure S1). This distortion was not observed for a relatively small data set of 700 SNPs, which was obtained after excluding all SNPs with more than one percent missing data points. Similar findings were observed when running principal component analyses (PCA, Figure S2). These findings support SambaR's recommendation to perform structure analyses with a relatively low number of high-quality SNPs rather than with a high number of low-quality SNPs.

For diversity analyses, in contrast, SambaR users are recommended to use a data set which exhibits no relationship between the proportion of missing data points and heterozygosity per sample. For the island channel fox data set, it can be argued that individuals with more than ten procent missing data should be excluded from the analyses (Figure 2b).

3.3 | DAPC analyses

We calculated the goodness of fit between a priori defined populations and DAPC inferred populations for a RADseq data set of 414 polar

bears (*Ursus maritimus*, Viengkone et al., 2017). DAPC analyses with 33 retained principal components, K = 4, and with prior population information, resulted in a graph similar to Figure 1 in Viengkone et al. (2017). DAPC analyses with 238 retained principal components, K = 4, and without prior population information, resulted in a graph similar to Figure 2 in Viengkone et al. (2017). For both settings overlap between predefined populations and DAPC clusters was poor (Figure S3), and this was reflected in the highly significant goodness of fit test *p*-values (X^2 = 298, df = 9, p = 0 and X^2 = 278, df = 9, p = 0).

Chi-squared tests for goodness of fit between DAPC inferred and three a priori defined European roe deer populations (De Jong et al., 2020) resulted in nonsignificant *p*-values (X^2 = 4.35, df = 4, p = 0.36 for both 20% and 80% explained variance), indicating DAPC inferred clusters generally corresponded to the predefined population structure (Figure S4).

3.4 | Selection of high informative SNPs and BPA test

We tested the power of reduced SNP panels to infer population structure in a data set of 394 coyotes (*Canis latrans*) sampled throughout the United States (Heppenheimer et al., 2018). SambaR generates these reduced SNP panels by selecting the SNPs with the highest standard deviation in minor allele frequency across populations. PCoA analyses indicated that random subsets of ≤500 SNPs generally gave poor power in resolving population structure (Figure 4). In contrast, PCoA analyses using small subsets of SNPs with the highest standard deviation in minor allele frequencies among populations, separated out predefined populations (Figure 4). Similar findings were observed when comparing the results of BPA tests on random and nonrandom SNP subsets (Figure S5). These findings illustrate that reduced SNP panels which are generated by selecting SNPs with high standard deviation in population minor allele frequencies, have the potential of low-cost population assignment.

3.5 | dc-score

To evaluate the usefulness of the dc-score, we compared the dc-score of the PCoA analyses on random and selected SNP subsets of the coyote data set (Heppenheimer et al., 2018). A strong correlation was observed between dc-score and the size of the SNP data set, ranging from 0.34 for 1,000 SNPs to 0.98 for 125 SNPs (Figure 4). For subsets comprised of most informative SNP's, the dc-score was less dependent on number of retained SNPs, and ranged between 0.24 for 500 SNPs and 0.28 for 125 SNPs (Figure 4).

4 | CONCLUSION

SambaR facilitates rapid population genetic analyses on biallelic SNP data sets by removing three major time sinks: file handling, software learning, and data plotting. In addition, SambaR provides a convenient platform for SNP data storage and management, guides users to adapt appropriate filter settings with regard to levels of missing data, and provides new tools. These newly developed utilities allow for generating reduced SNP panels, for generating structure-like plots with Bayesian populations assignment probabilities, and for objective interpretation of ordination analyses using the so called "distinct clustering"-score.

ACKNOWLEDGEMENTS

We thank Sofia Esteves da Silva, Erandi Bonillas Monge, Matt Newbould, Vania Fonseca da Silva, Daniel Moore, Sarah Mueller, Maria Nilsson-Janke, Thomas Parker, Dennis Schreiber, Yinhla Shihlomule, Biagio Violi, Magnus Wolf and Paige Yates for testing SambaR and providing feedback. We thank Tilman Schell and Christoph Sinai for installation instructions.

This work was supported by the British Deer Society, by the Kenneth Whitehead Trust, by Hesse's funding program LOEWE and by the Leibniz Association.

AUTHOR CONTRIBUTIONS

Menno de Jong and Joost de Jong developed the software. Menno de Jong wrote the paper and the manual, and developed the dc-score and the BPA-test. A. Rus Hoelzel and Axel Janke provided funding, feedback/advice, and input into the writing.

DATA AVAILABILITY STATEMENT

Data sets used in this study, as well as the SambaR script, can be found at: https://github.com/mennodejong1986/SambaR.

ORCID

Menno J. de Jong https://orcid.org/0000-0003-2131-9048
Joost F. de Jong https://orcid.org/0000-0002-8042-6022
A. Rus Hoelzel https://orcid.org/0000-0002-7265-4180

REFERENCES

Abel, G. J. (2019). *migest: Methods for the indirect estimation of bilateral migration*. Retrieved from https://CRAN.R-project.org/package=migest

Adler, D., & Kelly, S. T. (2019). *vioplot: Violin plot*. Retrieved from https://github.com/TomKellyGenetics/vioplot

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109.

Auguie, B. (2017). *gridExtra: Miscellaneous functions for "grid" graphics*. Retrieved from https://CRAN.R-project.org/package=gridExtra

Baudouin, L., Piry, S., & Cornuet, J. M. (2004). Analytical Bayesian approach for assigning individuals to populations. *Journal of Heredity*, *95*(3), 217–224.

Biscarini, F., Cozzi, P., Gaspa, G., & Marras, G. (2018). *detectRUNS: Detect runs of homozygosity and runs of heterozygosity in diploid genomes*. CRAN (The Comprehensive R Archive Network).

Caye, K., Jay, F., Michel, O., & François, O. (2018). Fast inference of individual admixture coefficients using geographic data. *The Annals of Applied Statistics*, *12*(1), 586–608. https://doi.org/10.1214/17-AOAS1106

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, H. (2018). *VennDiagram: Generate high-resolution venn and euler plots*. Retrieved from https://CRAN.R-project.org/package=VennDiagram

Cockerham, C. C., & Weir, B. S. (1987). Correlations, descent measures: Drift with migration and mutation. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(23), 8512–8514.

De Jong, M., Li, Z., Qin, Y., Quemere, E., Baker, K., Wang, W., & Hoelzel, A. R. (2020). Demography and adaptation promoting evolutionary transitions in a mammalian genus diversifying during the Pleistocene. *Molecular Ecology*, *29*(15):2777–2792. https://doi.org/10.1111/mec.15450

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252.

Flanagan, S. P., & Jones, A. G. (2018). *fsthet: Fst-heterozygosity smoothed quantiles*. Retrieved from https://CRAN.R-project.org/package=fsthet

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, *180*(2), 977–993. https://doi.org/10.1534/genetics.108.092221

Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*(8), 925–929. https://doi.org/10.1111/2041-210X.12382

Fung, T., & Keenan, K. (2014). Confidence intervals for population allele frequencies: the general case of sampling from a finite diploid population of any size. *PLoS One*, *9*(1), e85925. https://doi.org/10.1371/journal.pone.0085925

Funk, W. C., Lovich, R. E., Hohenlohe, P. A., Hofman, C. A., Morrison, S. A., Scott Sillett, T., Ghalambor, C. K., Maldonado, J. E., Rick, T. C., Day, M. D., Polato, N. R., Fitzpatrick, S. W., Coonan, T. J., Crooks, K. R., Dillon, A., Garcelon, D. K., King, J. L., Boser, C. L., Gould, N., & Andelt, W. F. (2016). Adaptive divergence despite strong genetic drift: genomic analysis of the evolutionary mechanisms causing genetic differentiation in the island fox (Urocyon littoralis). *Molecular Ecology*, *25*(10), 2176–2194.

Gaunt, T. R., Rodríguez, S., & Day, I. N. (2007). Cubic exact solutions for the estimation of pairwise haplotype frequencies: Implications for linkage disequilibrium analyses and a web tool "CubeX". *BMC Bioinformatics*, *8*, 428. https://doi.org/10.1186/1471-2105-8-428.

Gel, B., & Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*, *33*(9), 3088–3090.

Gerritsen, H. (2018). *mapplots: Data visualisation on maps*. Retrieved from https://CRAN.R-project.org/package=mapplots

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811–2812.

Heppenheimer, E., Brzeski, K. E., Hinton, J. W., Patterson, B. R., Rutledge, L. Y., DeCandia, A. L., Wheeldon, T., Fain, S. R., Hohenlohe, P. A., Kays, R., White, B. N., Chamberlain, M. J., & vonHoldt, B. M. (2018). High genomic diversity and candidate genes associated with a range expansion in eastern coyote (*Canis latrans*) populations. *Ecology and Evolution*, *8*(24):12641–12655. https://doi.org/10.1002/ece3.4688

Hijmans, R. J. (2019). *raster: Geographic data analysis and modeling*. Retrieved from https://CRAN.R-project.org/package=raster

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2), 583–589.

Jombart, T. (2008). adegenet: An 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*(21), 3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminat analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*, 94.

Kamvar, Z. N., Tabima, J. F., & Grunwald, 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

Kassambara, A., & Mundt, F. (2019). *factoextra: Extract and visualize the results of multivariate data analyses*. Retrieved from https://CRAN.R-project.org/package=factoextra

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Ligges, U., & Mächler, M. (2003). Scatterplot3d—An R package for visualizing multivariate data. *Journal of Statistical Software*, *8*(11), 1–20.

Liu, X., & Fu, Y.-X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics*, *47*(5), 555–559. https://doi.org/10.1038/ng.3254

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: An R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77. https://doi.org/10.1111/1755-0998.12592

Luu, K., Blum, M., & Prive, F. (2019). *pcadapt: Fast principal component analysis for outlier detection*. Retrieved from https://CRAN.R-project.org/package=pcadapt

Murrell, P. (2005). *R graphics*. Chapman & Hall/CRC Press.

Murrell, P., & Wen, Z. (2019). *gridGraphics: Redraw base graphics using "grid" graphics*. Retrieved from https://CRAN.R-project.org/package=gridGraphics

Mussmann, S. M., Douglas, M. R., Chafin, T. K., & Douglas, M. E. (2019). BA3-SNPs: Contemporary migration reconfigured in BayesAss for next-generation sequence data. *Methods in Ecology and Evolution*, *10*(10), 1808–1813. https://doi.org/10.1111/2041-210X.13252

Nei, M., Chakravarti, A., & Tateno, Y. (1977). Mean and variance of FST in a finite number of incompletely isolated populations. *Theoretical Population Biology*, *11*(3), 291–306.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes*. Retrieved from https://CRAN.R-project.org/package=RColorBrewer

Paradis, E., & Schliep, K. (2018). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526–528.

Peatkau, D., Calvert, W., Stirling, I., & Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, *4*(3), 347–354. https://doi.org/10.1111/1j.1365-294X.1995.tb00227.x

Pembleton, L. W., Cogan, N. O. I., & Forster, J. W. (2013). StAMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources*, *13*, 946–952. https://doi.org/10.1111/1755-0998.12129

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Soetaert, K. (2017). *Plot3D: Plotting multi-dimensional data*. Retrieved from https://CRAN.R-project.org/package=plot3D

South, A. (2011). rworldmap: A new R package for mapping global data. *The R Journal*, *3*, 35–43. https://doi.org/10.32614/RJ-2011-006

Storey, J. D., Bass, A. J., Dabney, A., & Robinson, D. (2019). *qvalue: Q-value estimation for false discovery rate control*. Retrieved from http://github.com/jdstorey/qvalue

Viengkone, M., Derocher, A. E., Richardon, E. S., Malenfant, R. M., Miller, J. M., Obbard, M. E., Dyck, M. G., Lunn, N. J., Sahanatien, V., & Davis, C. S. (2017). Assessing polar bear (*Ursus maritimus*) population structure in the Hudson Bay region using SNPs. *Ecology and Evolution*, *6*(23), 8474–8484.

Von Thaden, A., Nowak, C., Tiesmeyer, A., Reiners, T. E., Alves, P. C., Lyons, L. A., Mattucci, F., Randi, E., Cragnoloni, M., Galian, J., Hegyeli, Z., Kitchener, A. C., Lambinet, C., Lucas, J. M., Mölich, T., Ramos, L., Schockert, V., & Cocchiararo, B. (2020). Applying genomic data in wildlife monitoring: Development guidelines for genotyping degraded samples with reduced single nucleotide polymorphism panels. *Molecular Ecology Resources*, *20*(3), 662–680. https://doi.org/10.1111/1755-0998.13136

Waples, R. K., Albrechtsen, A., & Moltke, I. (2018). Allele frequency-free inference of close familial relationships from genotypes or low depth sequencing data. *Molecular Ecology*, *28*, 35–48.

Ward, B. J., & Van Oosterhout, C. (2016). HYBRIDCHECK: Software for the rapid detection, visualization and dating of recombination regions in genome sequence data. *Molecular Ecology Resources*, *16*(2), 534–539.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., & Venables, B. (2019). *gplots: Various R programming tools for plotting data*. Retrieved from https://CRAN.R-project.org/package=gplots

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370.

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of F(ST). *The American Naturalist*, *186*(Suppl 1), S24–36. https://doi.org/10.1086/682949

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H., & Seidel, D. (2019). *scales: Scale functions for visualization*. Retrieved from https://CRAN.R-project.org/package=scales

Wright, S. (1943). Isolation by distance. *Genetics*, *28*(2), 114–138.

Yang, J., Hong Lee, S., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1), 76–82.

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., Stauffer, R., & Wilke, C. O. (2019). colorspace: A toolbox for manipulating and assessing colors and palettes (ArXiv 1903.06490). arXiv.org E-Print Archive. http://arxiv.org/abs/1903.06490

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, *14*(6), 1–27. https://doi.org/10.18637/jss.v014.i06

Zheng X., Levine D., Shen J., Gogarten S. M., Laurie C., Weir B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*, (24), 3326–3328. http://dx.doi.org/10.1093/bioinformatics/bts606

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Jong MJ, Jong JF, Hoelzel AR, Janke A. SambaR: An R package for fast, easy and reproducible population-genetic analyses of biallelic SNP data sets. *Mol Ecol Resour*. 2021;21:1369–1379. https://doi.org/10.1111/1755-0998.13339