Received: 24 February 2021 | Accepted: 17 July 2021

DOI: 10.1111/2041-210X.13684

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

SMARTSNP, an R package for fast multivariate analyses of big genomic data

Salvador Herrando-Pérez[1,2] | Raymond Tobler[1,3] | Christian D. Huber[1,4]

[1]Australian Centre for Ancient DNA, School of Biological Sciences, The University of Adelaide, Adelaide, SA, Australia

[2]Department of Biogeography and Global Change, Museo Nacional de Ciencias Naturales, Spanish National Research Council (CSIC), Madrid, Spain

[3]Evolution of Cultural Diversity Initiative, Australian National University, Canberra, ACT, Australia

[4]Department of Biology, The Pennsylvania State University, University Park, PA, USA

Correspondence
Salvador Herrando-Pérez
Email: salherra@gmail.com

Christian D. Huber
Email: cdh5313@psu.edu

Funding information
C.D.H. and S.H.P. funded by Australian Research Council's DECRA Fellowship DE180100883 to C.D.H.; S.H.P. funded by Australian Research Council Centre of Excellence for Australian Biodiversity and Heritage CE170100015, and European Union's LIFE18 NAT/ES/000121 LIFE DIVAQUA; R.T. funded by Australian Research Council's DECRA Fellowship DE190101069.

Handling Editor: Thomas White

Abstract

1. Principal component analysis (PCA) is a powerful tool for the analysis of population structure, a genetic property that is essential to understand the evolutionary processes driving biological diversification and (pre)historical colonizations, migrations and extinctions. In the current era of high-throughput sequencing technologies, population structure can be quantified from scores of genetic markers across hundreds to thousands of genomes. However, these big genomic datasets pose substantial computing and analytical challenges.
2. We present the R package SMARTSNP for fast and user-friendly computation of PCA on single-nucleotide polymorphism (SNP) data. Inspired by the current field-standard software EIGENSOFT, *smartsnp* includes appropriate SNP scaling for genetic drift and allows projection of ancient samples onto a modern genetic space while also providing permutation-based multivariate tests for population differences in genetic diversity (both location and dispersion).
3. Our extensive benchmarks show that *smartsnp*'s PCA is 2–4 times faster than EIGENSOFT's SMARTPCA algorithm across a wide range of sample and SNP sizes. All four *smartsnp* functions (*smart_pca*, *smart_permanova*, *smart_permdisp* and *smart_mva*) process datasets with up to 100 samples and 1 million simulated SNPs in less than 30 s and accurately recreate previously published SMARTPCA of ancient-human and wolf genotypes.
4. The package SMARTSNP provides fast and robust multivariate ordination and hypothesis testing for big genomic data that is also suitable for ancient and low-coverage modern DNA. The simple implementation should appeal to biological conservation, evolutionary, ecological and (palaeo)genomic researchers, and be useful for phenotype, ancestry and lineage studies.

KEYWORDS

ancient DNA, genetic drift, population structure, single nucleotide polymorphism, SMARTPCA

1 | INTRODUCTION

Determining the genetic make-up of populations ('population structure') is a major area of research in multiple disciplines of science (Habel et al., 2015; Helyar et al., 2011). Principal component analysis (PCA: Hotelling, 1933; Pearson, 1901) is a foundational analytical tool in evolutionary genetic research (Cavalli-Sforza & Piazza, 1975) and remains one of the most popular statistical methods for summarizing population structure in the genomic era—essentially, because the underlying mathematical theory is conceptually simple (Fenderson et al., 2020) and PCA outputs have a clear genetic interpretation (François & Gain, 2021; McVean, 2009; Peter, 2021). However, the magnitude of genetic data generated by modern high-throughput sequencing technologies poses substantial computing and analytical challenges for PCA and other genomics applications (Schork, 2018; Tripathi et al., 2016) that mandate the development of fast and robust programming pipelines in open-source platforms (e.g. Abraham & Inouye, 2014; Luu et al., 2017).

PCA is a fundamental step in the EIGENSOFT software suite—the current field standard for genetic research—which comprises two modules: (a) EIGENSTRAT (Price et al., 2006) accounts for ancestral relatedness in genome-wide disease studies contrasting affected individuals and controls and (b) POPGEN (Patterson et al., 2006) runs a PCA algorithm (SMARTPCA) that accounts for the expected allele-frequency dispersion caused by genetic drift in biallelic single nucleotide polymorphisms (SNP). The wide utility of this software is illustrated by >8,000 combined citations (*Scopus*; accessed November 2020) for the two seminal papers describing the functionality of the software (Patterson et al., 2006; Price et al., 2006), and citing publications include major areas of modern research like animal domestication (e.g. Orlando et al., 2013; Qiu et al., 2015) and extinction risk (e.g. Frandsen et al., 2020; Liu et al., 2018), and human population (pre)history (e.g. Lazaridis et al., 2014; Tishkoff et al., 2009) and disease (Khera et al., 2016; Zhang et al., 2009). However, SMARTPCA is currently only available for use in Unix command-line environments and therefore limited to scientists who are familiar with this bioinformatic language.

Here we present the R package SMARTSNP for fast and user-friendly computation of PCA on large SNP datasets (Herrando-Pérez et al., 2021; Huber & Herrando-Pérez, 2021). Crucially, *smartsnp* incorporates two of the most commonly used functionalities of SMARTPCA: (a) appropriate scaling of SNP genotypes to control for allele-frequency dispersion caused by genetic drift and (b) projection of ancient samples onto a genetic space generated from modern samples. Additionally, *smartsnp* includes functionality that allows users to contrast a PCA ordination against permutational multivariate ANOVA tests of population structure, which is currently unavailable in EIGENSOFT. The universality of the *R* language for scientific research (Tippmann, 2014), and the speed, simplicity and functionality of *smartsnp* should be attractive properties for the growing community of scientists investigating modern and ancient population structure in humans and other taxa.

2 | PACKAGE OVERVIEW

The SMARTSNP package is compatible with *R* versions from 3.6.3 (29/02/2020) upwards on *Linux*, *Mac* and *Windows* systems, and comprises four functions: *smart_pca*, *smart_permanova*, *smart_permdisp* and *smart_mva* (see summary of arguments in Table 1). In the following subsections, we explain and benchmark those functions, and provide descriptions of currently implemented input-data formats and SNP-scaling options.

2.1 | Functions

Functions *smart_pca*, *smart_permanova* and *smart_permdisp* implement PCA, and permutational multivariate analysis of variance (PERMANOVA, Anderson, 2001) and dispersion (PERMDISP, Anderson, 2006), respectively. The *smart_mva* function is a wrapper that runs any combination of the three standalone functions. The mathematical rationale of these methods is expanded in Supporting Information S2. Briefly, PCA recalculates the geometric position of multivariate data (*variables* × *samples*) by rigidly rotating a system of *j* orthogonal axes (variables) such that the dispersion of *i* points (samples) is maximized along the rotated axes. For genotype data, the SNPs are the variables and the genotyped individuals are the samples. PERMANOVA and PERMDISP are statistical tests for multivariate differences in the relative position (*location*) and spread (*dispersion*) of sample groups (populations) using permutations of a triangular matrix (*sample* × *sample*) containing pair-wise inter-sample proximities. Measuring inter-sample proximities as Euclidean distances allows global and pair-wise testing of the location and dispersion of sample groups within a PCA ordination via PERMANOVA and PERMDISP in the full *j*-multidimensional PCA space, or alternatively in a lower-dimensional PCA space (e.g. the first two or three principal axes that are typically subjected to visual inspection and inference). Importantly, appropriate application of PERMANOVA and PERMDISP tests requires that sample groups are defined a priori (before undertaking any analyses) using associated metadata and genetic theory, because a posteriori testing of sample groupings derived from visual inspection of a PCA ordination is philosophically and statistically flawed.

2.2 | Analytical sequence

Function *smart_pca* runs in seven steps (Figure 1): (1) loading data, (2) indexing samples (group assignment, modern versus ancient) and SNPs that will be used or removed for downstream analysis, (3) removing invariant SNPs, (4) imputing missing values (coded either *NA* or *9*), (5) scaling SNPs (unscaled, centred, scaled by *z*-scores or drift), (6) single value decomposition (SVD: canonical or truncated) and (7) optionally projecting ancient samples onto modern PCA space. In addition to steps (1)–(5), *smart_permanova* and *smart_permdisp* (8) partition the genetic variance in an ANOVA framework and (9) estimate the probability (α) of group location or dispersion using permutations given the null hypothesis of no genetic differences between groups. Function *smart_mva* can compute any combination of PCA, PERMANOVA and/or PERMDISP in a single run. All functions conclude their computations by extracting pertinent statistical results and storing them as named elements of a standard *R* list (Figure 1). This list can be assigned to an object within the

TABLE 1 Brief description of the arguments from the four SMARTSNP package functions (1) *smart_pca*, (2) *smart_permanova*, (3) *smart_permdisp* and (4) *smart_mva*. Input data consist of SNPS (rows) by samples (columns) as a text or EIGENSOFT file, without row or column headings. Computational flow is shown in Figure 1, and command-line examples are presented in Table S1

Function	Argument	Description
(1–4)	*snp_data*	Name of input genotype data
(1–4)	*packed_data*	EIGENSOFT data type (compressed, uncompressed)
(1–4)	*sample_group*	Sample assignment to groups
(1–4)	*sample_remove*	Sample exclusion from analysis
(1–4)	*snp_remove*	SNP exclusion from analysis
(1–4)	*missing_value*	Value for missing genotype (*9*, *NA*)
(1–4)	*missing_impute*	Handling missing SNP (removal, mean imputation)
(1–4)	*scaling*	SNP scaling (none, covariance, correlation, genetic drift)
(1–4)	*program_svd*	SVD computation (truncated/*RSpectra*, canonical/*bootSVD*)
(1–4)	*pc_axes*	Number of computed PCA axes
(1,4)	*sample_project*	Samples assigned to ancient or modern
(1,4)	*pc_project*	PCA space for ancient projection
(2–4)	*program_distance*	Inter-sample proximity calculation (*vegan*, *Rfast*)
(2–4)	*sample_distance*	Inter-sample proximity metric (e.g., Euclidean)
(2–4)	*target_space*	Variance-partition space (multidimensional, PCA)
(2–4)	*pairwise*	Computation of pair-wise tests
(2–4)	*pairwise_method*	Correction for multiple pair-wise testing (e.g. Holm)
(2–4)	*permutation_n*	Number of permutations for α value computation
(2–4)	*permutation_seed*	Random generator of permutations
(3–4)	*dispersion_type*	Group dispersion estimate (centroid, median)
(3–4)	*samplesize_bias*	Dispersion correction for unequal group size
(4)	*pca*	PCA computation
(4)	*permanova*	PERMANOVA computation
(4)	*permdisp*	PERMDISP computation

Abbreviations: SNP, single nucleotide polymorphism; SVD, single value decomposition.

R environment, and each element of the list can be accessed by its name.

Examples of how to run the four functions are explained in the documentation of our package, with simulated genotypes (README file) and real data (vignette) examining the flyways of a cosmopolitan bird (Kraus et al., 2013).

2.3 | Input data formats

The standard genotype (g) data taken by SMARTSNP are biallelic SNPs with genotypes [0|1|2] from diploid organisms based on the number of copies of non-reference alleles. For instance, a SNP with reference allele G and variant allele T will have genotypes $g(GG) = 0$ (homozygous reference), $g(GT) = 1$ (heterozygous) and $g(TT) = 2$ (homozygous non-reference). Genotypes from haploid or polyploid organisms can be similarly defined and used with our package.

The SMARTSNP package accepts three data formats: (1) a generic text file without row (SNP) or column (sample) names, or an EIGENSOFT **.geno* file in (2) uncompressed (EIGENSTRAT) *or* (3) compressed/binary (PACKEDANCESTRYMAP) format (https://reich.hms.harvard.edu/software). For users who have their genotype data stored in VCF or PLINK formats (Chang et al., 2015; Zhang, 2016), step-by-step instructions for converting these formats into a flat file that can be handled by our package are provided in a vignette in the *GitHub* repository of the *smartsnp package* (https://christianhuber.github.io/smartsnp/articles). Handling of missing values is achieved either by removal of SNPs with ≥1 missing value, or imputation with SNP means (Marchini & Howie, 2010). Users are required to provide a vector assigning samples to groups: for EIGENSOFT files, this vector can often be obtained from the 3rd column of the **.ind* file, which

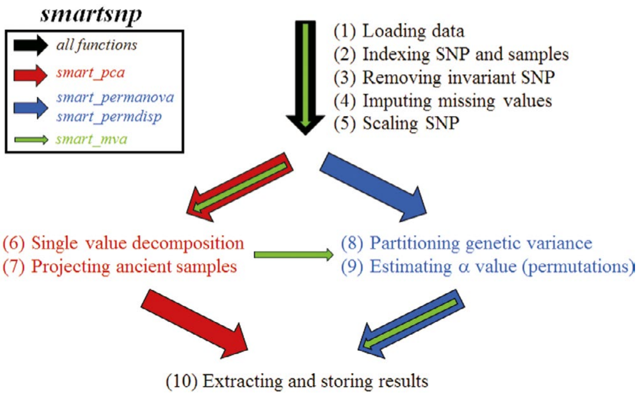
smartsnp
all functions
smart_pca
smart_permanova
smart_permdisp
smart_mva
(1) Loading data
(2) Indexing SNP and samples
(3) Removing invariant SNP
(4) Imputing missing values
(5) Scaling SNP
(6) Single value decomposition
(7) Projecting ancient samples
(8) Partitioning genetic variance
(9) Estimating α value (permutations)
(10) Extracting and storing results

FIGURE 1 Computational flow of package SMARTSNP when running principal component analysis (*smart_pca*), and permutational multivariate analysis of variance (*smart_permanova*) or dispersion (*smart_permdisp*). The wrapper function (*smart_mva*) can run any combination of these three primary analyses together. Function arguments described in Table 1

also includes alpha-numeric sample identifiers (1st column) and user-predefined descriptors like sexes (2nd column).

EIGENSOFT was conceived for human genetics so the SMARTPCA suite accepts 22 (autosomal) chromosomes by default. If >22 chromosomes are provided and the parameter *numchrom* (number of chromosomes) is unmodified, SMARTPCA subsets chromosomes 1–22. Our package accepts any number of autosomes with/without the sex chromosomes, and can single out discrete sets of SNPs (by row number) or samples (by column number) to be excluded from PCA, PERMANOVA and/or PERMDISP. When projecting ancient samples onto modern PCA space, a vector specifying the column number of the *ancient* samples is required (Table 1).

2.4 | SNP scaling

Prior to SVD or ANOVA, the SMARTSNP package can scale SNPs in four different ways (command-line examples shown in Table S1): (1) unscaled, (2) centred by their mean (covariance-based PCA), (3) standardized by *z*-scores (correlation-based PCA: SNPs have zero mean and unitary variance; Jolliffe & Cadima, 2016) and (4) scaled to control for genetic drift as in SMARTPCA (Patterson et al., 2006) following the formula:

$$M(i, j) = C(i, j) - \mu(j)/\sqrt{p(j)\,(1 - p(j))},$$

where $C(i, j)$ is the raw genotype value for SNP j in sample i, $\mu(j)$ is the mean value for SNP j across samples, $p(j) = \mu(j)/2$ estimates the underlying allele frequency and $M(i, j)$ is the scaled genotype value per data cell. This scaling accounts for the expected dispersion of allele frequencies due to genetic drift being proportional to $\sqrt{p(j)\,(1 - p(j))}$ (Patterson et al., 2006)—effectively reweighting each SNP according to its heterozygosity (or, equivalently, penalising those alleles most prone to drift; that is, intermediate-frequency alleles).

3 | OPTIMIZATION AND BENCHMARKING

We expedited the runtime of *smartsnp* at the two key computational bottlenecks: data loading and SVD computation. For loading EIGENSOFT files, we use *vroom::vroom_fwg* (Hester & Wickham, 2020) for fast-conversion of fixed-width uncompressed files (EIGENSTRAT), and an internal C++ function customized to emulate *admixtools::read_packedancestrymap* (Maier & Patterson, 2020) for compressed/binary files (PACKEDANCESTRYMAP). For loading text files, we use *data.table::fread* (Dowle & Srinivasan, 2019), which automatically detects file extension and column separators. To reduce data load in memory, SNPs with zero variance (same genotype across samples) are removed by default, as invariant SNPs make no contribution to SVD or variance partitioning. Users can further reduce runtime by applying truncated SVD (calculation of a predefined number of principal axes) using *RSpectra::svds* (Qiu & Mei, 2019), rather than canonical SVD (calculation of all principal axes) using *bootSVD::fastSVD* (Fisher, 2015). Computation of the truncated SVD is much faster than canonical SVD for big data (see benchmarking below), and the use of either option will depend on the number of dimensions subjected to investigation.

We benchmarked our package using *R* function *microbenchmark::microbenchmark* (Mersmann, 2019) on 34 simulated datasets (described in Supporting Information S3) in two ways. We compared computing times taken by (1) the four functions run by *smartsnp* and (2) function *smart_pca* from *smartsnp* versus SMARTPCA from EIGENSOFT across different data sizes.

3.1 | Smartsnp functions

In Tables S2 and S3, we report mean and standard errors (10 runs) of computing times of SMARTSNP's four functions. Runtime increased through *smart_permdisp*, *smart_pca* and *smart_permanova*, indicating that the two most resource-consuming calculations were α-value estimation in PERMANOVA and SVD in PCA. Notable speed gains occurred with the wrapper function; so for any given dataset, *smart_mva* was 1–3 orders of magnitude faster than running the three standalone functions separately, because the former only needs to load the data once.

On average, for a dataset with 100 samples (Table S2), the full computation of any function took ≤30 s for ≤1 million SNPs, and <1 and <6 min for 5 and 10 million SNPs, respectively. For a dataset with 100,000 SNPs (Table S3), all functions took <2 min for ≤500 samples, 50 s to 7 min for 1,000 samples and 2 min to <5 hr for 5,000 samples. Truncated SVD was up to 3 and 19 times faster than canonical SVD across functions using an increasing number of SNPs (Table S2) and samples (Table S3), respectively.

3.2 | Smartsnp versus EIGENSOFT

In Figure S1 and Table S4, we report mean and standard errors of computing times of *smartsnp::smart_pca* against EIGENSOFT'S

SMARTPCA. When all PCA axes were computed (1 core), *smart_pca* was >4× faster than SMARTPCA; and when two PCA axes were computed, *smart_pca* was ~2× faster than SMARTPCA for the largest data sizes. When using multithreading (4 cores), *smart_pca* was 2× faster than SMARTPCA for 100 samples and varying amounts of SNPs, and both had similar speeds for 100,000 SNPs and varying sample sizes (Table S4). Runtime improvements come at a cost to memory efficiency, with *smartsnp* functions using more random-access memory (RAM) than EIGENSOFT for equivalently sized datasets (though memory usage levels are not onerous given modern RAM specifications; see below).

4 | PROJECTION OF ANCIENT SAMPLES

Palaeogenomics is a rapidly growing field (Brunson & Reich, 2019) that uses *ancient* DNA (*a*DNA) recovered from specimens over at least the last 500,000 years (Pääbo et al., 2004; Slatkin & Racimo, 2016) to investigate (pre)historical population structure and other evolutionary questions. However, genetic degradation of *a*DNA results in abundant missing bases, challenging subsequent statistical analyses. Among the available approaches in PCA to handle missing data (reviewed by Ausmees, 2019; Günther & Jakobsson, 2019), function *smart_pca* implements the 'Projection to Model Plane' after Nelson et al. (1996)—the current standard method in the *a*DNA field, which is performed by SMARTPCA (Patterson et al., 2006). Briefly, PCA is computed using modern samples only, and ancient samples are projected onto the PCA space through linear regression. The projected coordinates of each ancient sample onto a particular subset of PCA axes equal the coefficient (slope) of a linear fit through the origin (Nelson et al., 1996), where the response is the vector of non-missing genotypes for that ancient sample, and the predictor is the vector(s) of the principal coefficients (loadings) assigned to each of the non-missing genotypes across modern samples. For example, projection onto PCA axes 1 and 2 equates with a linear model with two predictors (or vectors) of principal coefficients defined by modern data, and projection onto PCA axes 1–3 equates with a linear model with three such predictors. Our package provides users with the choice of any number and combination of PCA axes, for example, PCA 1 × PCA 2, PCA 1 × PCA 2 × PCA 3 × PCA 4, PCA 1 × PCA 3, PCA 2 × PCA 6, etc.

4.1 | Demonstration with empirical data

We analysed two previously published SNP datasets examined through SMARTPCA by Lazaridis et al. (2016) for anatomically modern humans *Homo sapiens* and Pilot et al. (2019) for grey wolves *Canis lupus*. For each dataset, we scaled SNPs to control for genetic drift and quantified the match between the ordination of samples using SMARTPCA (in EIGENSOFT) versus *smart_pca* (in *smartsnp*) using three statistical metrics (see Legendre & Legendre, 2012 for details of those tests): the Spearman correlation (Spearman, 1904) between the ranked sample positions along (1) PCA axis 1 and (2) PCA axis 2 from both analyses and (3) a Mantel test (Mantel, 1967) between pair-wise inter-sample Euclidean distances (*sample* × *sample* triangular matrix) in PCA 1 × PCA 2 space from both analyses based on Spearman correlations and α values obtained from 999 permutations. For the human dataset, we replicated Lazaridis et al.'s (2016) SMARTPCA by projecting ancient samples onto the modern genetic space (more details are provided in a GitHub vignette: https://christianhuber.github.io/smartsnp/articles). Additionally, for the wolf dataset, we formally tested whether the population structure reported by Pilot et al. (2019) was supported by PERMANOVA and PERMDISP analyses. The α values computed during the Mantel, PERMANOVA and PERMDISP analyses quantify the number of permuted datasets resulting in a test statistic equal to or larger than the observed statistic from the empirical data, with lower α values indicating lower probabilities that the observed statistic is due to chance.

Lazaridis et al. (2016) investigated the origin of farming using >500 thousand SNPs from 1,152 individuals sampled across West Eurasia—of which 278 were ancient hunter-gatherers (spanning 12,000–1,400 years BC). Our *smart_pca* ordination (Figure 2) mirrored the SMARTPCA ordination (figure 1b in Lazaridis et al., 2016), capturing the genetic gradient from European (left) to Near East (right) ancient groups in PCA 1, and the genetic gradients within these two groups in PCA 2. For modern samples, the Spearman correlation of sample positions along PCA 1 and 2 between both analyses were 0.999; while the Mantel correlation for inter-sample distances between *smart_pca* and SMARTPCA in PCA 1 × PCA 2 space was 0.969 (α = 0.001). We found the same magnitude of agreement between the ordination of ancient samples using *smart_pca* and SMARTPCA, with Spearman correlations of 0.999 (for both PCA 1 and PCA 2), and a Mantel correlation of 0.974 with α = 0.001 (PCA 1 × PCA 2) between both analyses. Such high correlations support a near-perfect match between the ordination of samples obtained by the two software packages. Function *smart_pca* used seven times more RAM memory and was three times faster than SMARTPCA (RAM allocation = 4,079 vs. 580 MB, and computation times = 2.0 min vs. 6.2 min, respectively).

Pilot et al. (2019) investigated phylogeographical patterns of grey wolves using 42,320 SNPs from 306 individuals (8 populations) sampled from Eurasia and North America. Among other predictions, they hypothesized that linkage disequilibrium (non-random association of alleles across loci) in East Eurasian populations increased proportionately with distance from West Eurasian and North American populations. Our *smart_pca* ordination (Figure 3) again mirrored the SMARTPCA ordination (see figure 3d in Pilot et al., 2019). PCA 1 recapitulated a genetic gradient from European (left) to North American wolves (right), with East Asian individuals and a single Pleistocene (~35,000 years BP) Taimyr wolf lying between these two population clusters, and PCA 2 separating Mexican wolves (bottom) from all other individuals (Figure 3). The Spearman ranked correlations of sample positions along PCA 1 and 2 were 0.999, respectively; while Mantel correlations between inter-sample distances in

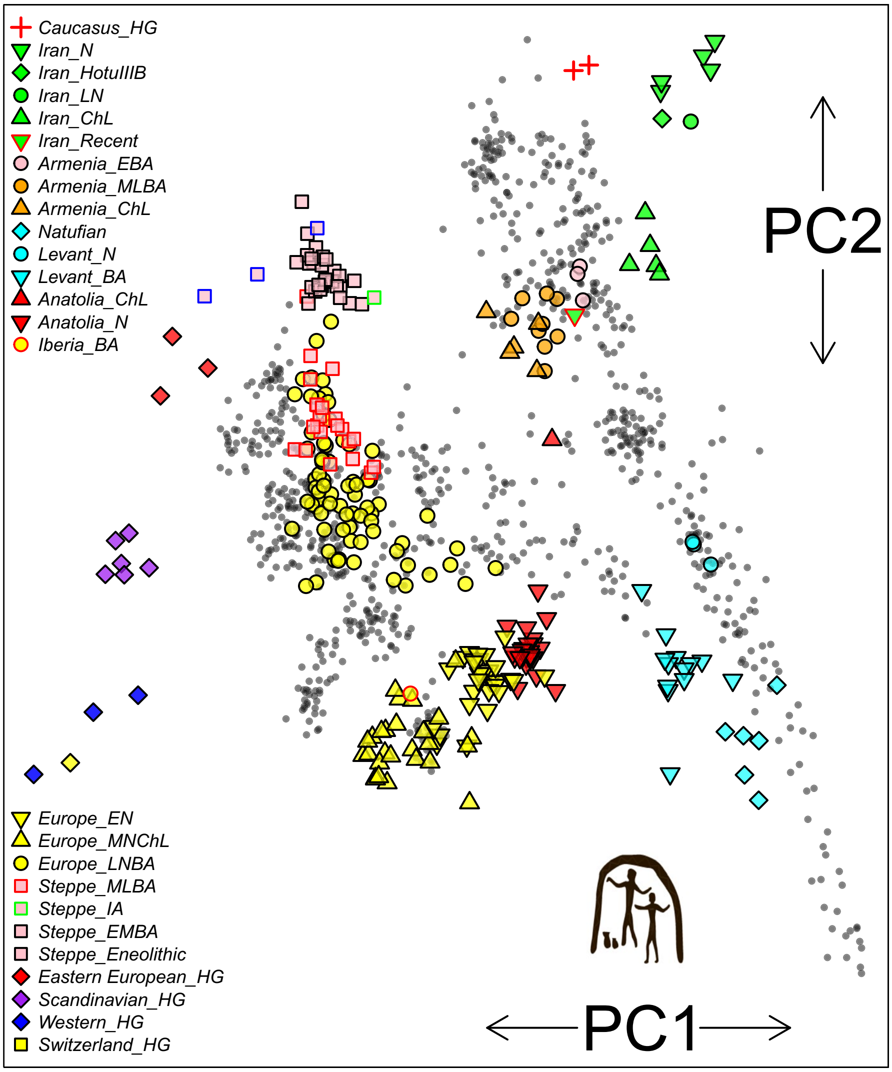

FIGURE 2 Population structure of ancient West Eurasian farmers and hunter-gatherers using principal component analysis in R package SMARTSNP. Data comprise 874 modern West Eurasians (grey circles), 278 projected ancient individuals (26 populations, coloured symbols) and 548,749 single nucleotide polymorphisms scaled to control for genetic drift (Lazaridis et al., 2016). The ordination explains 1.2% in genetic diversity across individuals (PCA 1 = 0.8%; PCA 2 = 0.4%). Population acronyms: ChL = Chalcolithic, BA = Bronze Age, E = Early, HG = Hunter-Gatherer, IA = Iron Age, L = Late, M = Middle, N = Neolithic

PCA 1 × PCA 2 space between both analyses was 0.999 ($\alpha = 0.001$). For this relatively small dataset, runtimes for *smarp_pca* (2 s) and SMARTPCA (4 s) were comparable, and *smart_pca* used 15 times more RAM memory than SMARTPCA (442 vs. 29 MB).

Based on the original SMARTPCA results and related analyses, Pilot et al. (2019) surmised that the Taimyr wolf is a sister lineage of modern Eurasian wolves but its relationship with North American wolves remains uncertain. After excluding the Taimyr wolf, PERMANOVA and PERMDISP global tests supported that SNP genetic diversity differed in both location and dispersion in PCA 1 × PCA 2 space among the other seven wolf populations ($\alpha = 0.0001$ for PERMANOVA and PERMDISP global tests; Table 2). The probability of the differences in group location given the null hypothesis of groups having the same location was α = 0.0021 (with correction for multiple testing) for all pair-wise PERMANOVA comparisons (Table 2), and a total of 16 (no multiple-testing correction) and 9 (multiple-testing correction) of 21 PERMDISP pair-wise comparisons had α < 0.1 (Table 2). The median dispersion of samples to group spatial medians was lowest for Minnesota and West Asian wolves and highest for Mexican and North American wolves, with European and East Asian populations exhibiting intermediate dispersions (Figure S2). Increased genetic heterogeneity in North American wolfs might indicate the wider geographical range of the selected individuals compared to the other study populations while the Mexican wolfs form a reintroduced population that have experienced genetic bottlenecks and strong genetic drift, which should magnify to the variability of SNP composition among populations (Małgorzata Pilot, pers. comm., May 2021). Runtimes for PERMANOVA and PERMDISP tests totalled 28 and 23 s, respectively.

Differences in genetic heterogeneity between populations at neutral markers reflect differences in demographic history and geographical structure (Charlesworth et al., 2003), and can be used to detect processes that decrease (e.g. bottlenecks) or increase (e.g. admixture) genetic variation. More generally, in ecology, multivariate dispersion in species composition is interpreted as a measure of beta diversity (Anderson et al., 2006) that quantifies species turnover across different assemblages, and also serves as a measure of stress (Warwick & Clarke, 1993) where increased variability in species composition signals the impact of environmental perturbations. Both interpretations have analogous applications in population genetics. For instance, increased genetic heterogeneity in the PCA space is interpreted as increased genetic diversity when

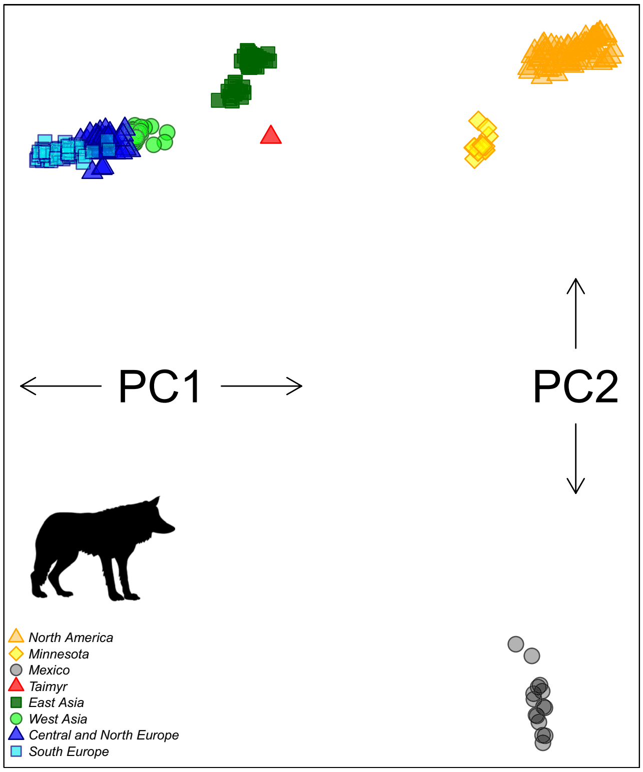

FIGURE 3 Population structure of Eurasian and North American grey wolves using principal component analysis in R package SMARTSNP. Data comprise 306 individuals (8 populations) and 42,320 single nucleotide polymorphisms scaled to control for genetic drift (Pilot et al., 2019). The ordination explains 9.8% in genetic diversity across individuals (PCA 1 = 6.7%; PCA 2 = 3.1%)

the most variable loci have the strongest effects on the phenotype of certain ethnic groups (Fadhlaoui-Zid et al., 2015; Solovieff et al., 2010; Yu et al., 2020), and this property has also been used as an indicator of disease profiles and their ancestral origin (Horne et al., 2016; Ioannidis et al., 2004; Manichaikul et al., 2012; Turajlic et al., 2019).

4.2 | Conclusions

Our R package SMARTSNP can be used to conduct exploratory analyses and confirm hypotheses about population structure in large genomic SNP datasets. It can be applied to all living systems for which haploid, diploid or polyploid genotype datasets are available to visualize complex genetic relationships resulting from evolutionary and demographic processes, and be useful for phenotype, ancestry and lineage studies. The implemented projection method can be applied to any dataset with large amounts of missing data (*a*DNA or low-coverage modern data) as long as a high-quality reference dataset with little missing data is available. Importantly, our PCA functionality provides results that mirror SMARTPCA analyses but runs 2–4 times faster for big datasets in a user-friendly, platform-independent context. By providing multivariate tests for differences in the location and dispersion of genetic data across predefined groups, the SMARTSNP package also makes it possible for users to formally detect population structure and other differences potentially caused by evolutionary, ecological or sociocultural factors.

ACKNOWLEDGEMENTS

We thank Julia A. Pilowsky, Simon (J.) Tuke and Joshua M. Schmidt for guidance for package editing for *GitHub*/*CRAN*, Robert Maier for sharing the code of his function *admixtools::read_packedancestrymap*, Małgorzata Pilot and Wiesław Bogdanowicz for their kind provision of the wolf dataset and assistance to interpret genetic-dispersion results, Robert H. S. Kraus for instructions for the correct use of the bird-flyways dataset in one of the package VIGNETTES, and Fernando Valladares for proof-reading the Spanish Abstract (available in the online version of the article). The human dataset was obtained from https://reich.hms.harvard.edu/datasets.

CONFLICT OF INTEREST

None declared.

AUTHORS' CONTRIBUTIONS

C.D.H. and S.H.-P. conceived the idea and carried out the benchmarking; S.H.-P. wrote the first draft of manuscript, package functions and manuals; R.T. optimized function performance for big genomic data and revised the manuals; C.D.H. provided supervision throughout, expanded the *R* code to read PACKEDANCESTRYMAP genotype data and submitted the package to *CRAN*, *GitHub* and *Zenodo*. All authors contributed to manuscript revisions and approved submission.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13684.

DATA AND PACKAGE AVAILABILITY

The SMARTSNP package is freely available as an open-source R package and a user's manual under MIT licence at the Comprehensive *R* Archive Network (*CRAN*: https://cran.r-project.org/web/packages/smartsnp), *GitHub* (https://github.com/ChristianHuber/smartsnp) and *Zenodo* (Huber & Herrando-Pérez, 2021: https://doi.org/10.5281/zenodo.5124765). Vignettes providing several package-usage examples can be found at https://christianhuber.github.io/smartsnp/articles. The package includes a simulated dataset (name = 'dataSNP') with 100 samples (columns) and 100,000 randomly generated SNPs (rows) comprising 9,886 missing values coded as 9s. A further empirical dataset is available on the *GitHub* repository that comprises 364 SNPs (rows) from 696 individuals (columns) of mallards, *Anas platyrhynchos*, obtained from Kraus et al. (2013). Website hyperlinks listed in Supporting Information S1.

ORCID

Salvador Herrando-Pérez https://orcid.org/0000-0001-6052-6854
Raymond Tobler https://orcid.org/0000-0002-4603-1473
Christian D. Huber https://orcid.org/0000-0002-2267-2604

TABLE 2 Global and pair-wise testing for differences in location (PERMANOVA) and dispersion (PERMDISP) of seven wolf populations in PCA 1 × PCA 2 space (Figure 3) using the R package SMARTSNP. Data comprise 305 individuals (7 populations) and 42,320 single nucleotide polymorphisms scaled to control for genetic drift (Pilot et al., 2019). The probability (α) of the observed location or dispersion was obtained from 9,999 permutations given the null hypothesis that all groups had the same location or dispersion, respectively, without (α_1) and with (α_2) Holm's correction for multiple testing. Statistical metrics also include R^2 = percentage variance explained and F = Fisher's ANOVA statistic. The asterisk (*) indicates pair-wise comparisons with $\alpha_2 < 0.1$. Population acronyms are CN = Central and North, E = East, N = North, S = South and W = West

	PERMANOVA				PERMDISP		
	R^2	F	α_1	α_2	F	α_1	α_2
Global test →	16.8	10.1	0.0001	—	11.5	0.0001	—
Pair-wise tests ↓							
CN Europe–E Asia	6.2	7.3	0.0001	0.0021	5.2	0.0236	0.1935
CN Europe–Mexico*	13.2	11.2	0.0001	0.0021	18.3	0.0002	0.0038
CN Europe–Minnesota	8.1	6.5	0.0001	0.0021	1.0	0.3190	1.0000
CN Europe–N America*	10.0	15.9	0.0001	0.0021	45.0	0.0001	0.0021
CN Europe–S Europe	4.2	4.6	0.0001	0.0021	6.2	0.0133	0.1584
CN Europe–W Asia	5.2	5.0	0.0001	0.0021	1.1	0.2903	1.0000
E Asia–Mexico	15.9	12.1	0.0001	0.0021	3.9	0.0478	0.3346
E Asia–Minnesota	9.5	6.7	0.0001	0.0021	3.6	0.0598	0.3588
E Asia–N America*	9.1	13.4	0.0001	0.0021	10.0	0.0023	0.0345
E Asia–S Europe	9.1	9.6	0.0001	0.0021	0.0	0.9044	1.0000
E Asia–W Asia	7.9	6.8	0.0001	0.0021	6.0	0.0132	0.1584
Mexico–Minnesota*	25.5	9.6	0.0001	0.0021	8.9	0.0028	0.0392
Mexico–N America	11.6	12.8	0.0001	0.0021	0.2	0.6527	1.0000
Mexico–S Europe	18.3	13.5	0.0001	0.0021	5.6	0.0191	0.1910
Mexico–W Asia*	20.8	11.5	0.0001	0.0021	15.6	0.0003	0.0051
Minnesota–N America*	5.6	5.9	0.0001	0.0021	19.4	0.0002	0.0038
Minnesota–S Europe	12.1	8.2	0.0001	0.0021	5.2	0.0215	0.1935
Minnesota–W Asia	13.4	6.8	0.0001	0.0021	0.1	0.8233	1.0000
N America–S Europe*	12.0	17.8	0.0001	0.0021	12.4	0.0006	0.0096
N America–W Asia*	10.2	12.9	0.0001	0.0021	35.1	0.0001	0.0021
S Europe–W Asia*	7.9	6.5	0.0001	0.0021	8.2	0.0045	0.0585

REFERENCES

Abraham, G., & Inouye, M. (2014). Fast principal component analysis of large-scale genome-wide data. *PLoS ONE*, *9*, e93766. https://doi.org/10.1371/journal.pone.0093766

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*, 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, *62*, 245–253. https://doi.org/10.1111/j.1541-0420.2005.00440.x

Anderson, M. J., Ellingsen, K. E., & McArdle, B. H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecology Letters*, *9*, 683–693. https://doi.org/10.1111/j.1461-0248.2006.00926.x

Ausmees, K. (2019). Evaluation of methods handling missing data in PCA on genotype data: Applications for ancient DNA. Technical Report, 2019-009, 1–10. Department of Information Technology, Uppsala University, Sweden. Retrieved from http://www.it.uu.se/research/publications/reports/2019-009/

Brunson, K., & Reich, D. (2019). The promise of paleogenomics beyond our own species. *Trends in Genetics*, *35*, 319–329. https://doi.org/10.1016/j.tig.2019.02.006

Cavalli-Sforza, L. L., & Piazza, A. (1975). Analysis of evolution: Evolutionary rates, independence and treeness. *Theoretical Population Biology*, *8*, 127–165. https://doi.org/10.1016/0040-5809(75)90029-5

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

Charlesworth, B., Charlesworth, D., & Barton, N. H. (2003). The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution, and Systematics*, *34*, 99–125. https://doi.org/10.1146/annurev.ecolsys.34.011802.132359

Dowle, M., & Srinivasan, A. (2019). *data.table*: extension of 'data.frame'. Retrieved from https://CRAN.R-project.org/package=data.table (version 1.12.8).

Fadhlaoui-Zid, K., Garcia-Bertrand, R., Alfonso-Sánchez, M. A., Zemni, R., Benammar-Elgaaied, A., & Herrera, R. J. (2015). Sousse: Extreme genetic heterogeneity in North Africa. *Journal of Human Genetics*, *60*, 41–49. https://doi.org/10.1038/jhg.2014.99

Fenderson, L. E., Kovach, A. I., & Llamas, B. (2020). Spatiotemporal landscape genetics: Investigating ecology and evolution through space and time. *Molecular Ecology*, *29*, 218–246. https://doi.org/10.1111/mec.15315

Fisher, A. (2015). *bootSVD: Fast, exact bootstrap principal component analysis for high dimensional data*. Retrieved from https://CRAN.R-project.org/package=bootSVD

François, O., & Gain, C. (2021). A spectral theory for Wright's inbreeding coefficients and related quantities. *PLoS Genetics*, *17*, e1009665. https://doi.org/10.1371/journal.pgen.1009665

Frandsen, P., Fontsere, C., Nielsen, S. V., Hanghøj, K., Castejon-Fernandez, N., Lizano, E., Hughes, D., Hernandez-Rodriguez, J., Korneliussen, T. S., Carlsen, F., Siegismund, H. R., Mailund, T., Marques-Bonet, T., & Hvilsom, C. (2020). Targeted conservation genetics of the endangered chimpanzee. *Heredity*, *125*, 15–27. https://doi.org/10.1038/s41437-020-0313-0

Günther, T., & Jakobsson, M. (2019). Population genomic analyses of DNA from ancient remains. In D. J. Balding, I. Moltke, & J. Marioni (Eds.), *Handbook of statistical genomics* (pp. 295–240). Wiley.

Habel, J. C., Zachos, F. E., Dapporto, L., Rödder, D., Radespiel, U., Tellier, A., & Schmitt, T. (2015). Population genetics revisited – Towards a multidisciplinary research field. *Biological Journal of the Linnean Society*, *115*, 1–12. https://doi.org/10.1111/bij.12481

Helyar, S. J., Hemmer-hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, A., Maes, G. E., Diopere, E., Carvalho, G. R., & Nielsen, E. E. (2011). Application of SNPs for population genetics of nonmodel organisms: New opportunities and challenges. *Molecular Ecology Resources*, *11*, 123–136. https://doi.org/10.1111/j.1755-0998.2010.02943.x

Herrando-Pérez, S., Tobler, R., & Huber, C. D. (2021). *smartsnp: Fast multivariate analyses of big genomic data* (v. 1.1.0). Retrieved from https://CRAN.R-project.org/package=smartsnp

Hester, J., & Wickham, H. (2020) *vroom: read and write rectangular text data quickly*. https://CRAN.R-project.org/package=vroom (Version 1.3.0).

Horne, H. N., Chung, C. C., Zhang, H., Yu, K., Prokunina-Olsson, L., Michailidou, K., Bolla, M. K., Wang, Q., Dennis, J., Hopper, J. L., Southey, M. C., Schmidt, M. K., Broeks, A., Muir, K., Lophatananon, A., Fasching, P. A., Beckmann, M. W., Fletcher, O., Johnson, N., … Figueroa, J. D. (2016). Fine-mapping of the 1p11.2 breast cancer susceptibility locus. *PLoS ONE*, *11*, e0160316. https://doi.org/10.1371/journal.pone.0160316

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417–441. https://doi.org/10.1037/h0071325

Huber, C. D., & Herrando-Pérez, S. (2021). ChristianHuber/*smartsnp*: first release (version v1.0.0). *Zenodo*, https://doi.org/10.5281/zenodo.5124765

Ioannidis, J. P. A., Ntzani, E. E., & Trikalinos, T. A. (2004). 'Racial' differences in genetic effects for complex diseases. *Nature Genetics*, *36*, 1312–1318. https://doi.org/10.1038/ng1474

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*, 20150202. https://doi.org/10.1098/rsta.2015.0202

Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P., Bick, A. G., Cook, N. R., Chasman, D. I., Baber, U., Mehran, R., Rader, D. J., Fuster, V., Boerwinkle, E., Melander, O., Orho-Melander, M., Ridker, P. M., & Kathiresan, S. (2016). Genetic risk, adherence to a healthy lifestyle, and coronary disease. *New England Journal of Medicine*, *375*, 2349–2358. https://doi.org/10.1056/NEJMoa1605086

Kraus, R. H. S., van Hooft, P., Megens, H.-J., Tsvey, A., Fokin, S. Y., Ydenberg, R. C., & Prins, H. H. T. (2013). Global lack of flyway structure in a cosmopolitan bird revealed by a genome wide survey of single nucleotide polymorphisms. *Molecular Ecology*, *22*, 41–55. https://doi.org/10.1111/mec.12098

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., Connell, S., Stewardson, K., Harney, E., Fu, Q., Gonzalez-Fortes, G., Jones, E. R., Roodenberg, S. A., Lengyel, G., Bocquentin, F., … Reich, D. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature*, *536*, 419–424. https://doi.org/10.1038/nature19310

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K. I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer, K., … Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*, 409–413. https://doi.org/10.1038/nature13673

Legendre, P., & Legendre, L. F. J. (2012). *Numerical ecology* (3rd ed.). Elsevier.

Liu, Y.-C., Sun, X., Driscoll, C., Miquelle, D. G., Xu, X., Martelli, P., Uphyrkina, O., Smith, J. L. D., O'Brien, S. J., & Luo, S.-J. (2018). Genome-wide evolutionary analysis of natural history and adaptation in the world's tigers. *Current Biology*, *28*, 3840–3849.e3846. https://doi.org/10.1016/j.cub.2018.09.019

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*, 67–77. https://doi.org/10.1111/1755-0998.12592

Maier, R., & Patterson, N. (2020). *admixtools: Inferring demographic history from genetic data*. Retrieved from https://rdrr.io/github/uqrmaie1/admixtools

Manichaikul, A., Palmas, W., Rodriguez, C. J., Peralta, C. A., Divers, J., Guo, X., Chen, W.-M., Wong, Q., Williams, K., Kerr, K. F., Taylor, K. D., Tsai, M. Y., Goodarzi, M. O., Sale, M. M., Diez-Roux, A. V., Rich, S. S., Rotter, J. I., & Mychaleckyj, J. C. (2012). Population structure of hispanics in the United States: The multi-ethnic study of atherosclerosis. *PLoS Genetics*, *8*, e1002640. https://doi.org/10.1371/journal.pgen.1002640

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*, 209–220.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*, 499–511. https://doi.org/10.1038/nrg2796

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, *5*, e1000686. https://doi.org/10.1371/journal.pgen.1000686

Mersmann, O. (2019). *microbenchmark: Accurate timing functions* (version 1.4-7). Retrieved from https://CRAN.R-project.org/package=microbenchmark

Nelson, P. R. C., Taylor, P. A., & MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, *35*, 45–65. https://doi.org/10.1016/S0169-7439(96)00007-X

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L. F., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliussen, T., Malaspinas, A.-S., Vogt, J., Szklarczyk, D., Kelstrup, C. D., … Willerslev, E. (2013). Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, *499*, 74–78. https://doi.org/10.1038/nature12323

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., & Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annual Review of Genetics*, *38*, 645–679. https://doi.org/10.1146/annurev.genet.37.110801.143214

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*, e190. https://doi.org/10.1371/journal.pgen.0020190

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*, 559–572. https://doi.org/10.1080/14786440109462720

Peter, B. M. (2021). Modelling complex population structure using *F*-statistics and principal component analysis. *bioRxiv*, 1–19. https://doi.org/10.1101/2021.07.13.452141

Pilot, M., Moura, A. E., Okhlopkov, I. M., Mamaev, N. V., Alagaili, A. N., Mohammed, O. B., Yavruyan, E. G., Manaseryan, N. H., Hayrapetyan, V., Kopaliani, N., Tsingarska, E., Krofel, M., Skoglund, P., & Bogdanowicz, W. (2019). Global phylogeographic and admixture patterns in grey wolves and genetic legacy of an ancient Siberian lineage. *Scientific Reports*, *9*, 17328. https://doi.org/10.1038/s41598-019-53492-9

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909. https://doi.org/10.1038/ng1847

Qiu, Q., Wang, L., Wang, K., Yang, Y., Ma, T., Wang, Z., Zhang, X., Ni, Z., Hou, F., Long, R., Abbott, R., Lenstra, J., & Liu, J. (2015). Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nature Communications*, *6*, 10283. https://doi.org/10.1038/ncomms10283

Qiu, Y., & Mei, J. (2019). *RSpectra: solvers for large-scale eigenvalue and SVD problems* (version 0.16-0). Retrieved from https://CRAN.R-project.org/package=RSpectra

Schork, N. J. (2018). The big data revolution and human genetics. *Human Molecular Genetics*, *27*, R1. https://doi.org/10.1093/hmg/ddy123

Slatkin, M., & Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 6380–6387. https://doi.org/10.1073/pnas.1524306113

Solovieff, N., Hartley, S. W., Baldwin, C. T., Perls, T. T., Steinberg, M. H., & Sebastiani, P. (2010). Clustering by genetic ancestry using genome-wide SNP data. *BMC Genetics*, *11*, 108. https://doi.org/10.1186/1471-2156-11-108

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101. https://doi.org/10.2307/1412159

Tippmann, S. (2014). Programming tools: Adventures with *R*. *Nature*, *517*, 109–110. https://doi.org/10.1038/517109a

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., … Williams, S. M. (2009). The genetic structure and history of Africans and African Americans. *Science*, *324*, 1035–1044. https://doi.org/10.1126/science.1172257

Tripathi, R., Sharma, P., Chakraborty, P., & Varadwaj, P. K. (2016). Next-generation sequencing revolution through big data analytics. *Frontiers in Life Science*, *9*, 119–149. https://doi.org/10.1080/21553769.2016.1178180

Turajlic, S., Sottoriva, A., Graham, T., & Swanton, C. (2019). Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics*, *20*, 404–416. https://doi.org/10.1038/s41576-019-0114-6

Warwick, R. M., & Clarke, K. R. (1993). Increased variability as a symptom of stress in marine communities. *Journal of Experimental Marine Biology and Ecology*, *172*, 215–226. https://doi.org/10.1016/0022-0981(93)90098-9

Yu, C., Ni, G., van der Werf, J., & Lee, S. H. (2020). Detecting genotype-population interaction effects by ancestry principal components. *Frontiers in Genetics*, *11*, 379. https://doi.org/10.3389/fgene.2020.00379

Zhang, F.-R., Huang, W., Chen, S.-M., Sun, L.-D., Liu, H., Li, Y. I., Cui, Y., Yan, X.-X., Yang, H.-T., Yang, R.-D., Chu, T.-S., Zhang, C., Zhang, L., Han, J.-W., Yu, G.-Q., Quan, C., Yu, Y.-X., Zhang, Z., Shi, B.-Q., … Liu, J.-J. (2009). Genomewide association study of leprosy. *New England Journal of Medicine*, *361*, 2609–2618. https://doi.org/10.1056/NEJMoa0903753

Zhang, H. (2016). Overview of sequence data formats. In E. Mathé & S. Davis (Eds.), *Statistical genomics: Methods and protocols* (pp. 3–17). Springer.

SUPPORTING INFORMATION

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

How to cite this article: Herrando-Pérez, S., Tobler, R., & Huber, C. D. (2021). SMARTSNP, an R package for fast multivariate analyses of big genomic data. *Methods in Ecology and Evolution*, 00, 1–10. https://doi.org/10.1111/2041-210X.13684