Abstract

Summary: We present a new R package admixr, which provides a convenient interface for performing reproducible population genetic analyses ($f_2$, $D$, $f_4$, $f_4$-ratio, qpWave and qpAdm), as implemented by command-line programs in the ADMIXTOOLS software suite. In a traditional ADMIXTOOLS workflow, the user must first generate a set of text configuration files tailored to each individual analysis, often using a combination of shell scripting and manual text editing. The non-tabular output files then need to be parsed to extract values of interest prior to further analyses. Our package simplifies this process by automating all low-level configuration and parsing steps, making analyses as simple as running a single R command. Furthermore, we provide a set of R functions for processing, filtering and manipulating datasets in the EIGENSTRAT format. By unifying all steps of the workflow under a single R framework, this package enables the automation of analytic pipelines, significantly improving the reproducibility of population genetic studies.

Availability and implementation: The source code of the R package is available under the MIT license. Installation instructions, reference manual and a tutorial can be found on the package website at https://bioinf.eva.mpg.de/admixr.

Contact: mp@bodkan.net or kelso@eva.mpg.de

Supplementary information: Supplementary data are available at Bioinformatics online.
possible to implement and share fully automated, reproducible analytic pipelines.

2 Implementation

The admixr package is implemented using the R programming language. It consists of several wrapper functions (calling ADMIXTOOLS commands internally from R), and a set of complementary functions for filtering and processing datasets in the EIGENSTRAT file format required by ADMIXTOOLS (Patterson et al., 2012).

An EIGENSTRAT dataset is represented by an S3 object of the class EIGENSTRAT, which is created using the eigenstrat() constructor function, and encapsulates the paths to a trio of 'ind', 'snp' and 'geno' files:

```r
> snps <- eigenstrat("~/path/to/eigenstrat/data")
> snps
EIGENSTRAT object

components:
ind file: ~/path/to/eigenstrat/data.ind
snp file: ~/path/to/eigenstrat/data.snp
genone: ~/path/to/eigenstrat/data.geno
```

All other functions in the package accept this object as their first argument, and perform either a requested calculation on it (returning an R data frame for further analysis), or return a new, modified EIGENSTRAT S3 object (in case of filtering and processing functions) or return a new, modified argument, and perform either a requested calculation on it (returning an R data frame for further analysis and plotting):

```r
result <- f4ratio(
  X = c("French", "Han", "Papuan"),
  A = "Altai", B = "Vindija", C = "Mbuti", O = "Chimp",
  data = eigenstrat("<path to EIGENSTRAT data>")
)
```

Internally, the f4ratio() function performs all configuration and parsing work, and returns an R data frame which can be immediately used for further statistical analysis and plotting:

```r
> result

A   B X C O alpha   stderr Zscore
Altai Vindija French Mbuti Chimp 0.0196966 0.0031344 6.324
Altai Vindija Han Mbuti Chimp 0.0243790 0.0033640 7.248
Altai Vindija Papuan Mbuti Chimp 0.0321670 0.0034990 9.193
```

All other admixr wrapper functions have a similar interface and are described in the tutorial vignette on the package website in more detail.

3 Example usage

Performing even the most trivial analysis using ADMIXTOOLS presents a significant amount of overhead for the user. For example, to estimate the proportion of Neandertal ancestry in a set of individuals, X, the user would typically calculate an f4-ratio statistic such as:

$$f_4(\text{Altai, Chimp; X, Mbuti})$$

$$f_4(\text{Altai, Chimp; Vindija, Mbuti})$$

The fact that ADMIXTOOLS requires the data to be in EIGENSTRAT format presents additional challenges for quality control, processing and filtering, as this format is not supported by standard bioinformatics tools. Our R package therefore provides additional functionality to simplify the processing and filtering of EIGENSTRAT genotype data. This includes:

- Reading and writing of ind, snp and geno file components.
- Filtering of SNPs based on regions specified in a BED file.
- Restricting analyses to sites carrying transversion SNPs.
- Renaming samples or grouping them into larger population groups.
- Merging of EIGENSTRAT datasets.
- Counting the number of sites present or missing in each sample.

Acknowledgements

We would like to thank Benjamin Peter and Hassan Shafiey for helpful comments, suggestions and testing.

Funding

This work was supported by funding from the Max Planck Society.

Conflict of Interest: none declared.

References

Fu, Q. et al. (2016) The genetic history of Ice Age Europe. Nature, 534, 200–205.
Haak,W. et al. (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. Nature, 522, 207–211.
Hajdinjak,M. et al. (2018) Reconstructing the genetic history of late Neanderthals. Nature, 555, 652–656.
Lazaridis,I. et al. (2016) Genomic insights into the origin of farming in the ancient Near East. Nature, 536, 419–424.
Patterson,N. et al. (2012) Ancient admixture in human history. Genetics, 192, 1065–1093.