Methods in Ecology and Evolution

RZooRoH: an R package to characterize individual genomic autozygosity and identify homozygous-by-descent segments

Amandine R. Bertrand[1], Naveen K. Kadri[2], Laurence Flori[3], Mathieu Gautier[4], and Tom Druet[2]

February 12, 2019

[1]InBioS—PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, B-4000 Liège, Belgium

[2] Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, Liège, B-4000, Belgium

[3] SELMET, INRA, CIRAD, Montpellier Supagro, Univ. Montpellier, Montpellier, France

[4] INRA, UMR CBGP (INRA – IRD – Cirad – Montpellier SupAgro), Montferrier-sur-Lez, France

Corresponding author:

Tom Druet

Unit of Animal Genomics, GIGA (B34 +1), Quartier Hôpital, Avenue de l'Hôpital, 11, B-4000 Liège, Belgium

Tel: +3243669172; Fax:+3243664151

E-mail: tom.druet@uliege.be

Keywords: Autozygosity; Inbreeding; Runs of Homozygosity (ROH); homozygosity-by-descent (HBD); identity-by-descent (IBD)

Running headline: The RZooRoH package

1. Identifying homozygous-by-descent (HBD) regions in individual genomes is highly valuable to infer the recent history of populations and to provide insights into trait architecture.

2. Here we present the RZooRoH R-package that implements an efficient and accurate model-based approach to identify HBD segments. The underlying Hidden Markov Model partitions the genome-wide individual autozygosity into different age-related HBD classes while accounting for genotyping errors and genetic map information.

3. The RZooRoH package is user-friendly and versatile, accepting either genotyping or sequencing (including low-coverage) data in various formats. Through numerical maximization and parallelization, computational performances were improved compared to our initial Fortran implementation of the model. The package allows to evaluate and to compare various models defined by their number of HBD classes and it also provides several graphical functions that help interpretation of the results.

4. RZooRoH is an efficient tool that proves particularly suited for sub-optimal data sets (e.g., low marker density, individual low-coverage sequencing, uneven marker spacing) and for individuals from populations with complex demographic histories. RZooRoH is available from CRAN: https://CRAN.R-project.org/package=RZooRoH.

RZooRoH: un paquet R pour caractériser la consanguinité génomique et identifier des segments homozygotes-par-descendance

1. L'identification dans les génomes de segments homozygotes-par-descendance (HBD) est particulièrement utile pour comprendre l'histoire démographique récente des populations et l'architecture génétique des caractères.

2. RZooRoH est un paquet R qui repose sur un modèle précis et efficace pour identifier des segments HBD. La chaîne de Markov cachée sous-jacente répartit l'autozygotie individuelle dans des classes HBD d'âges différents tout en tenant compte des erreurs de génotypage et de la carte génétique.

3. RZooRoH est facile d'utilisation et polyvalent: le paquet accepte des données de génotypage et de séquençage (y compris à faible couverture) sous différents formats. Grâce à des techniques d'optimisation numérique, les performances ont été améliorées par rapport à la version initiale développée en Fortran. Le paquet permet d'évaluer et de comparer de nombreux modèles, définis par le nombre de classes HBD. Plusieurs fonctions graphiques sont incluses pour faciliter l'interprétation des résultats.

4. RZooRoH est donc un outil efficace qui est particulièrement bien adapté à des jeux de données sous-optimaux (p.ex., faible densité de marqueurs, séquençage à faible couverture, distances génétiques hétérogènes entre marqueurs) et aux populations ayant une histoire démographique complexe. RZooRoH est disponible sur le site du CRAN: https://CRAN.R-project.org/package=RZooRoH.

Introduction

Autozygous or homozygous-by-descent (HBD) segments in individual genomes result from the inheritance of two copies of a single chromosomal segment from an ancestor. Due to recombination, the length of an HBD segment is inversely related to the number of generations connecting these two chromosomal copies in the genealogy and thus to the time of living of the ancestor. In turn, the inbreeding level of an individual directly determines the proportion of its genome that is HBD. For instance, matings between parents and their offspring result on average in 25% inbreeding and HBD segments about 30 cM in length (see also Thomas *et al.* (1994) for expected IBD sharing between two related individuals). Partitioning individual genomes into HBD and non-HBD segments has become very popular in recent years due to its wide range of applications (Ceballos *et al.*, 2018). It allows indeed to estimate the inbreeding coefficient, a key parameter for the management of populations under conservation or included in breeding programs, and to study inbreeding depression or to map recessive deleterious variants using autozygosity mapping (Ferenčaković *et al.*, 2017; Lander & Botstein, 1987; Wang *et al.*, 2009). In addition, characterizing the distribution of HBD segments is informative about the recent demographic history of populations (MacLeod *et al.*, 2009; Palamara *et al.*, 2012) since the frequency of HBD segments depends on the effective population size (Ne) and their length reflects the time to the common ancestor.

In genotyping or sequencing data sets, HBD segments are expected to consist of a stretch of homozygous genotypes called runs of homozygosity (ROH). Hence, several methods relying on ROH detection have been proposed to identify HBD segments (working most often with SNPs). First rule-based methods (e.g., McQuillan *et al.*, 2008) consider only stretches of genotypes fulfilling some pre-defined criteria (e.g., number of markers, number of heterozygous and missing genotypes, marker density, window length, marker spacing) as HBD segments. These criteria need to be optimized for every data set according to the characteristics of the studied population or the genotyping technology used. Second, model-based approaches relying on likelihood-ratio tests to classify windows of consecutive markers as HBD or non-HBD have been proposed (Pemberton *et al.*, 2012; Wang *et al.*, 2009). The likelihood accounts for marker allele frequencies and genotyping errors making decision less sensitive to marker ascertainment bias and an ad hoc procedure is used

to define the optimal window sizes in order to obtain a clear bimodal distribution of the LOD scores. Hidden Markov models (HMM) represent a valuable alternative by modeling individual genomes as mosaic of HBD and non-HBD segments (Leutenegger *et al.*, 2003). They indeed provide a framework to compute the HBD probabilities (as opposed to the binary classification obtained with ROH detection) at each marker position by integrating over all possible segment lengths while accounting for population allele frequencies, genotyping error rates and genetic distances between markers. They can also efficiently handle exome or whole-genome (including low-coverage) sequence data (Magi *et al.*, 2014; Narasimhan *et al.*, 2016; Vieira *et al.*, 2016).

Based on simulation studies, Narasimhan *et al.* (2016) found that HMM had lower false positive and false negative rates compared to ROH estimated with PLINK (Purcell *et al.*, 2007). Druet & Gautier (2017) concluded that the differences between the approaches was small when the number of markers per HBD segment was high whereas HMM performed better than window-based approaches to assess global or local (e.g., at each locus) autozygosity at lower marker density or for shorter HBD segments. The use of HMM is particularly valuable when information is sparser or less accurate as with low-coverage sequencing experiments, lower marker density, biased genotyping arrays, etc. (Druet & Gautier, 2017). The HMM are also particularly useful when the genotyping error or the recombination rates are variable, or the marker spacing is not uniform. For instance, Magi *et al.* (2014) showed that an HMM based approach outperforms PLINK when applied to whole-exome sequences data by considering the distances between consecutive markers. Vieira *et al.* (2016) demonstrated the importance of using genotype likelihoods (as integrated in some HMM approaches) instead of genotypes (as used in window-based approaches) when dealing with low-coverage sequencing data. HMM also present important conceptual differences with window-based methods since, by providing probabilities, they directly provide information about the uncertainty associated with the inference, which is particularly helpful when the information is degraded (e.g., low marker density, low MAF or low coverage). They also do not require the prior definition of various arbitrary thresholds (e.g., Yengo *et al.*, 2017) for marker spacing, window size, minimum number of markers, etc. that need to be re-defined for each data set with rule-based methods.

We herein present a new user-friendly R package implementing the HMM-based approach we recently developed (Druet & Gautier, 2017) to scan individual genomes for HBD segments. This method has been intensively tested on simulated data set with a wide range of characteristics (marker density, error rates, sequencing coverage, etc) and also compared to other methods (see Druet & Gautier (2017) for more details). The package works with different data types (genotypes, genotype likelihoods, sequence read counts) obtained with various technologies (genotyping arrays, whole genome-sequencing, genotyping-by-sequencing) and allows to explore and compare various model specifications. The package relies on a new optimization procedure implemented in

combination with a re-parametrization of the model that both improves parameter estimation and speed-ups computations. In addition, parallelization per individual allows further gain in speed. The package finally provides several graphical utilities that help interpretation of the results.

Description

The RZooRoH model: a multiple HBD classes HMM

The model is a HMM describing an individual genome as a succession of segments classified in $K-1$ HBD classes (defined by the expected length of the segments) and one non-HBD class, labeled K. To compute the probability of one such sequence, the model requires the probability to stay in the current segment or to start a new segment between two consecutive markers and the probability to observe particular genotype or sequence data conditionally on the class specificities. The probability to continue a segment is e^{-R_k}, where R_k is the rate specific to class k. As a result, the length of HBD segments from class k is exponentially distributed with a rate R_k. The expected length is then equal to $1/R_k$ Morgans. HBD classes with low rates (long HBD segments) correspond to HBD segments inherited from recent common ancestors whereas those with high rates correspond to autozygosity associated with ancient common ancestors. The rate of a class is approximately equal to twice the number of generations to the common ancestors associated with that class. Note that if the genetic distances are incorrect, HBD segments will still be identified but the relationship between their length and their age will be more complex and the rates of the HBD classes will no longer have a simple interpretation (see Druet & Gautier (2017) for more details).The probabilities to observe the genotype or sequence data depends on whether the class is HBD or not, on the allele frequencies and the genotyping error rate. In HBD segments, heterozygous genotypes are unlikely and result from mutations, gene conversions or genotyping errors, whereas in non-HBD segments, observed genotypes are expected in Hardy-Weinberg proportions. When a segment ends, the probability that the next segment starts in class k is a function of the mixing coefficients M_k. These parameters define the frequency of the segments from each class and don't have a straightforward biological interpretation.

On the benefits of using multiple HBD classes

The multiple HBD classes HMM relaxes the assumption made in previous works (Leutenegger *et al.*, 2003; Narasimhan *et al.*, 2016; Vieira *et al.*, 2016) that all HBD tracts belong to a single class (i.e., they have the same expected length) which might be interpreted biologically as considering that all the autozygosity traces back to one or several ancestors living in the same generation. Our multiple HBD classes model allows to fit more realistic situations where ancestors contributing

to autozygosity trace back to different generations in the past (e.g., Druet & Gautier, 2017; Solé *et al.*, 2017). The HBD classes are indeed related to the age of common ancestors contributing to them (Druet & Gautier, 2017): classes with longer (shorter) segments, i.e. lower (higher) rates, correspond to more recent (distant) common ancestors. We showed in previous studies that the use of multiple HBD classes results in a better fit of individual genetic data and more accurate estimations of autozygosity levels both locally (i.e. at each locus) and globally, particularly in complex populations (e.g., Druet & Gautier, 2017; Solé *et al.*, 2017). More precisely, single HBD class models may underestimate the autozygosity when multiple generations contribute to it (Druet & Gautier, 2017). We also illustrated with cattle data that the use of single HBD class models results in lower estimates of autozygosity and that the distribution of identified HBD segments is more concentrated at intermediary sizes (Solé *et al.*, 2017) because the smallest HBD segments are not captured while the long ones are fragmented in multiple smaller segments. Obtaining the correct length of HBD segments is essential to interpret the results, in particular to estimate the age of the ancestors. Interestingly, models with multiple HBD classes offer the possibility to reveal the recent demographic history by partitioning HBD segments in different age-related classes and to estimate the contribution of the different past generations to the current autozygosity (Druet & Gautier, 2017). At individual scale, high contributions of very recent HBD classes suggest that the parents were highly related which helps to understand mating behaviors in populations when pedigrees are unavailable. At the population level, large contributions of a class to autozygosity indicates a reduced effective population size (Ne) at the corresponding time period, possibly associated with a bottleneck or a founder effect. Conversely, a low contribution to autozygosity suggests a large Ne.

Model re-parametrization and implementation

In the HMM framework, we can use the forward-backward algorithm (Rabiner, 1989) to compute the probability to belong to each of the different classes at each marker position by integrating over all possible sequences of segments. These locus-specific HBD probabilities can be used for autozygosity mapping or can be averaged over all the positions to obtain genome-wide estimates (e.g., proportion of the genome within a given HBD class). They can be reported per HBD class or cumulated over several HBD classes with a rate smaller than a chosen threshold T. In that case, we estimate the proportion of autozygosity associated with common ancestors more recent than a selected time point (corresponding then to an inbreeding coefficient estimated with respect to a base population approximately $0.5*T$ generations ago). Alternatively, we can use the Viterbi algorithm (Rabiner, 1989) to identify the most likely sequence of classes and identify HBD segments as uninterrupted stretches of markers assigned to the same HBD class.

As in Vieira *et al.* (2016), we use the L-BFGS-B method implemented in the optim function from the R stats package (R Core Team, 2013) to estimate the model parameters. In order to work with unconstrained parameters and to obtain ordered HBD classes (with increasing rates of exponential distributions), we defined new parameters (e.g., Zucchini & MacDonald, 2009):

$$\eta_k = \begin{cases} \log(R_k - R_{k-1}) & \text{if } 1 < \mathrm{k} < \mathrm{K} \\ \log(R_k) & \text{if k} = 1 \text{ or k} = \mathrm{K} \end{cases} \quad (1)$$

$$\tau_k = \log\left(\frac{M_k}{M_K}\right) \qquad \text{if k} < \mathrm{K} \quad (2)$$

Note that this optimization procedure was not implemented in our initial Fortran implementation that relied on a EM-algorithm described in Druet & Gautier (2017). This EM algorithm is also implemented in the R-package.

The package

The package consists of several functions. First, zoodata() reads genotype data and marker positions. It is compatible with the oxford gen format and accepts also data from whole-genome sequencing as genotypes (zformat = "gt"), genotype probabilities (zformat = "gp"), likelihoods as phred score (zformat = "gl") or read counts (zformat = "ad").

The zoomodel() function allows the user to define the model: i) the number of HBD classes; ii) whether the rate of each HBD class should be pre-defined or estimated; iii) the error rate associated with the genotyping or sequencing data.

Then the zoorun() function is used to estimate the global and local HBD probabilities or to obtain the list of HBD segments (see vignette for more details on the output). The function also allows to estimate the model parameters i.e., the K mixture coefficients and the K rates of each HBD class (when they are not pre-defined) with either the new optimization procedure (method = "opti") or with the EM-algorithm (method = "estem").

Finally, four additional plot functions can be used for visualization purposes: zooplot_partitioning(), zooplot_prophbd(), zooplot_hbdseg(), zooplot_individuals(). These functions plot the partitioning of the genome in different HBD classes per individual (as a barplot), report the proportion of the genome associated with different HBD classes at the population or individual levels (either with cumulative curves or not) and represent HBD segments identified in a specific genomic region.

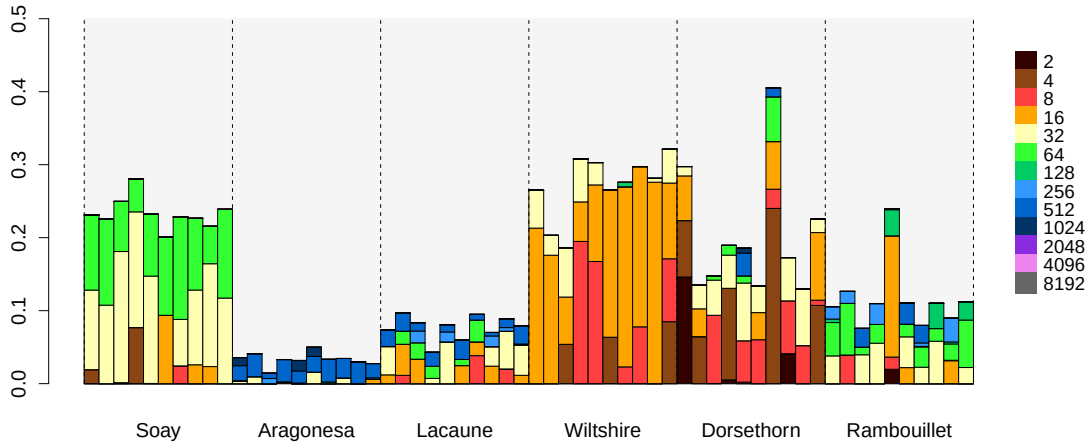

Figure 1: **Partitioning of the genome in different HBD classes in six sheep populations.** Results are plotted for ten randomly selected individuals per population. The height of each bar represents the proportion of the genome associated with the HBD class of the corresponding color.

Efficiency of the new implementation

To test the efficiency of the newly implemented optimization procedure, we compared it to the results obtained with the EM algorithm in terms of likelihood, number of iterations and computing time on two data sets we previously used (Druet & Gautier, 2017) and for four models (Tables 1 and 2). The first data set consists of a simulation with 500 individuals and the second data set consists of publicly available genotypes for 110 Soay sheep (Kijas *et al.*, 2012).

The new optimization procedure achieved higher or similar likelihoods than the EM algorithm for models with pre-defined classes or with a single HBD class, and performed clearly better for models with multiple HBD classes when rates are estimated (Table 1). Note that global autozygosity was always highly correlated across procedures (> 0.9995). Differences in HBD class partitioning between the two estimation procedures were subtle for models with pre-defined classes or a single HBD class but substantial for other models. This was expected since the estimation of HBD class rates makes the HBD classes no longer identical. Note however, that in these cases, the new optimization procedure leads to estimated rates that provides a better fit to the data (i.e., higher likelihoods). Similar trends were observed for analyses of simulated data sets and other real data sets (e.g., Bertrand, 2017). From a computational point of view, the optimization procedure always required less iterations (up to 4 times less) to converge than the EM algorithm for the simulated data set (Table 2) and for the Soay sheep data set when analyzed under the more complex model. However, one iteration of the optimization procedure requires a single call of the forward algorithm and thus remains two to three times faster than one iteration of the EM algorithm that requires in addition a call of the backward algorithm and additional computations.

As a result, the approach with the numerical optimization was 1.22 to 10 times faster than with the EM algorithm (Table 2). Note also, that the zoorun() function can be run in parallel for several individuals (nT option) which allows approximately dividing the computation time by the number of threads.

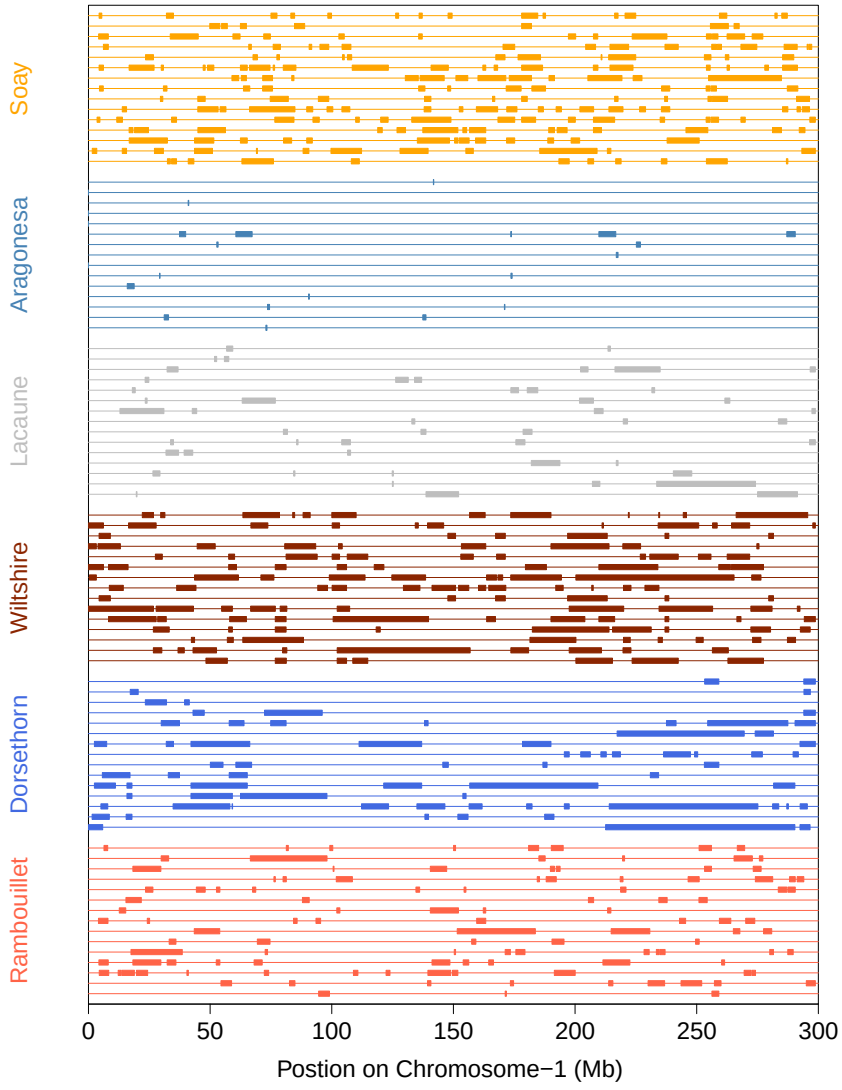

Figure 2: **HBD segments identified on chromosome 1 in six sheep populations.** Results are plotted for ten randomly selected individuals per population. One line represents one individual and HBD segments are indicated with wider boxes.

Illustration on real data

For illustration purposes, we ran the package on genotype data from six sheep populations with different demographic histories already studied in Druet & Gautier (2017). We defined a model with 13 HBD classes with pre-defined rates equal to {2, 4, 8, ... 8192}. The populations display variation in terms of numbers and age/length of HBD segments as shown by our results. Fig.

1 represents the partitioning of individuals genomes in these 13 different HBD classes (obtained with the zooplot_partitioning() function). The populations display clear differences in terms of partitioning. For instance, autozygosity in individuals from the Aragonesa breed is limited and associated with small (ancient) HBD segments (represented in blue). Individuals from the Soay population exhibit higher levels of autozygosity than Dorsethorn individuals but associated with more distant ancestors (mostly HBD classes with rates equal to 32 and 64). The Rambouillet population displays more variation in total autozygosity and length of HBD segments (more HBD classes contribute to autozygosity).

The HBD segments identified on chromosome 1 are plotted in Fig. 2 (obtained with the zooplot_hbdseg() function). They are in agreement with the partitioning described above: few and short segments in the Aragonesa breed, high autozygosity levels with long HBD segments in the Wiltshire individuals, many small segments in Soay sheep, etc. Finally, Fig. 3 shows variation in individual levels of autozygosity, cumulated over several HBD classes (summing all the autozygosity associated with HBD classes with a rate smaller than a selected threshold, associated with common ancestor more recent than a selected period). As explained in Solé *et al.* (2017) these values can be interpreted as inbreeding coefficients estimated with respect to different base populations. We observe clear differences in terms of total autozygosity, generations contributing to autozygosity and individual variation (some breeds being more homogeneous than others).

Conclusion

Identifying HBD segments is essential for many applications in quantitative and population genetics. Methods should handle both genotyping and sequencing data available in both model and non-model organisms. In addition, they should account for complex demographic histories (with multiple common ancestors in multiple generations), marker allele frequencies, genetic distances, genotyping errors probabilities and confidence in genotype calling (amount of information). To that end, we developed RzooRoH, an efficient user-friendly R-package, that implements a HMM model-based approach partitioning autozygosity in age-related HBD classes.

Acknowledgements

We thank the International Sheep Genomics consortium for data sharing. Tom Druet is Senior Research Associate from the F.R.S.-FNRS. This work has been supported by the Fonds de la Recherche Scientifique - FNRS (F.R.S.-FNRS) under Grant J.0134.16 and J.0154.18. We used the supercomputing facilities of the "Consortium d'Equipements en Calcul Intensif en Fédération Wallonie-Bruxelles" (CECI), funded by the F.R.S.-F.N.R.S.

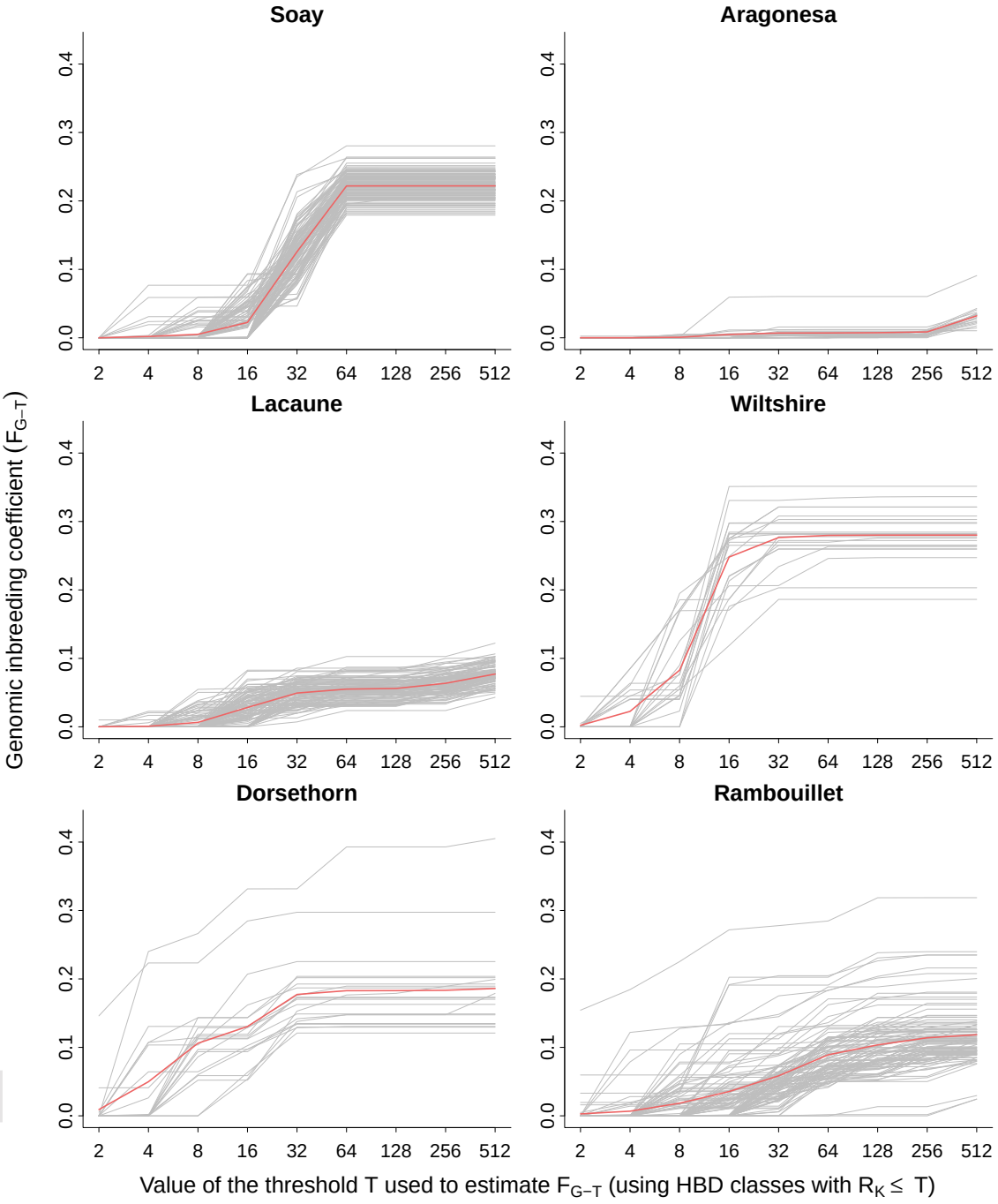

Figure 3: **Genomic inbreeding coefficients estimated with respect to different base populations (F_{G-T}).** These were estimated as the probability of belonging to any of the HBD classes with a rate $\leq T$, setting the base population approximately 0.5^*T generations ago.

Author's contributions

T.D., M.G., N.K.K and A.R.B. conceived and designed the package. T.D., A.R.B. and L.F. performed experiments and extensively tested the package on simulated and real data sets. T.D. and M.G. wrote the manuscript and documentation with the input from the other authors.

Data Accessibility

The data sets used in the present study come from previous studies and are publicly available. The Sheep Diversity panel (Kijas *et al.*, 2012) data was downloaded from the WIDDE data base at http://widde.toulouse.inra.fr/widde (Sempéré *et al.*, 2015). Both data sets can also be obtained at https://doi.org/10.5281/zenodo.2562716.

References

Bertrand, A. (2017) *Etude de la consanguinité des bisons d'Europe avec un modèle de Markov caché à multiples classes autozygotes*. Master's thesis, Université de Liège, Liège, Belgique.

Ceballos, F.C., Joshi, P.K., Clark, D.W., Ramsay, M. & Wilson, J.F. (2018) Runs of homozygosity: windows into population history and trait architecture. *Nature Reviews Genetics*.

Druet, T. & Gautier, M. (2017) A model-based approach to characterize individual inbreeding at both global and local genomic scales. *Molecular ecology*, **26**, 5820–5841.

Ferenčaković, M., Sölkner, J., Kapš, M. & Curik, I. (2017) Genome-wide mapping and estimation of inbreeding depression of semen quality traits in a cattle population. *Journal of dairy science*, **100**, 4721–4730.

Kijas, J.W., Lenstra, J.A., Hayes, B., Boitard, S., Porto Neto, L.R. *et al.* (2012) Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology*, **10**, e1001258.

Lander, E.S. & Botstein, D. (1987) Homozygosity mapping: a way to map human recessive traits with the dna of inbred children. *Science*, **236**, 1567–1570.

Leutenegger, A.L., Prum, B., Genin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F. & Thompson, E.A. (2003) Estimation of the inbreeding coefficient through use of genomic data. *American Journal of Human Genetics*, **73**, 516–23.

MacLeod, I., Hayes, B., Goddard, M. *et al.* (2009) A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. *Genetics research*, **91**, 413–426.

Magi, A., Tattini, L., Palombo, F., Benelli, M., Gialluisi, A., Giusti, B., Abbate, R., Seri, M., Gensini, G.F., Romeo, G. *et al.* (2014) H 3 m 2: detection of runs of homozygosity from whole-exome sequencing data. *Bioinformatics*, **30**, 2852–2859.

McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., MacLeod, A.K., Farrington,

s.R., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S.H., Dunlop, M.G., Wright, A.F., Campbell, H. & Wilson, J.F. (2008) Runs of homozygosity in european populations. *American Journal of Human Genetics*, **83**, 359–372.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C. & Durbin, R. (2016) Bcftools/roh: a hidden markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, **32**, 1749–1751.

Palamara, P.F., Lencz, T., Darvasi, A. & Pe'er, I. (2012) Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, **91**, 809–822.

Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A. & Li, J.Z. (2012) Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics*, **91**, 275–292.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**, 559–575.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabiner, L.R. (1989) A tutorial on hidden markov models and selected applications in speech recognition. *PROCEEDINGS OF THE IEEE*, pp. 257–286.

Sempéré, G., Moazami-Goudarzi, K., Eggen, A., Laloë, D., Gautier, M. & Flori, L. (2015) Widde: a web-interfaced next generation database for genetic diversity exploration, with a first application in cattle. *BMC genomics*, **16**, 940.

Solé, M., Gori, A.S., Faux, P., Bertrand, A., Farnir, F., Gautier, M. & Druet, T. (2017) Age-based partitioning of individual genomic inbreeding levels in belgian blue cattle. *Genetics Selection Evolution*, **49**, 92.

Thomas, A., Skolnick, M.H. & Lewis, C.M. (1994) Genomic mismatch scanning in pedigrees. *Mathematical Medicine and Biology*, **11**, 1–16.

Vieira, F.G., Albrechtsen, A. & Nielsen, R. (2016) Estimating ibd tracts from low coverage ngs data. *Bioinformatics*, **32**, 2096–2102.

Wang, S., Haynes, C., Barany, F. & Ott, J. (2009) Genome-wide autozygosity mapping in human populations. *Genet Epidemiol*, **33**, 172–80.

Yengo, L., Zhu, Z., Wray, N.R., Weir, B.S., Yang, J., Robinson, M.R. & Visscher, P.M. (2017) Detection and quantification of inbreeding depression for complex traits from snp data. *Proceedings of the National Academy of Sciences*, **114**, 8602–8607.

Zucchini, W. & MacDonald, I. (2009) Hidden markov models for time series, volume 110 of monographs on statistics and applied probability.

Data	Model	Mean difference	Minimal Difference	Maximal difference
Soay	Pre-defined - 4 classes	0.001	0.000	0.038
Soay	Pre-defined - 14 classes	0.007	-0.006	0.140
Soay	Rate estimation - 1 HBD class	0.000	0.000	0.000
Soay	Rate estimation - 4 classes	-2.879	-7.406	0.001
Simulated	Pre-defined - 4 classes	0.000	-0.003	0.010
Simulated	Pre-defined - 14 classes	0.004	-0.013	0.186
Simulated	Rate estimation - 1 HBD class	0.000	0.000	0.000
Simulated	Rate estimation - 4 classes	-4.091	-17.305	0.369

Table 1: **Difference in likelihood maximization.** The differences are equal to the maximum log(likelihood) achieved with the EM algorithm minus the value achieved with the numerical maximization procedure and the model re-parametrization. Negative values indicate that the new procedure achieved higher likelihood than the EM algorithm.

Data	Model	Mean number of iterations			Total CPU time (hh::mm:ss)			Elapsed time (hh::mm:ss)
		f-EM	r-EM	r-optim	f-EM	r-EM	r-optim	r-optim with 8 threads
Soay	Pre-defined - 4 classes	78.7	78.7	182.6	00:09:11	00:10:49	00:08:54	00:01:15
Soay	Pre-defined - 14 classes	429.9	429.9	1037.5	05:41:30	06:18:42	04:19:44	00:35:07
Soay	Rate estimation - 1 HBD class	42.2	42.2	51.5	00:03:14	00:03:53	00:01:41	00:00:23
Soay	Rate estimation - 4 classes	952.4	953.3	495.3	02:33:02	02:54:25	00:30:57	00:03:27
Simulated	Pre-defined - 4 classes	534.3	534.4	151.8	02:27:39	02:44:00	00:16:33	00:02:16
Simulated	Pre-defined - 14 classes	717.7	718.0	652.7	12:15:26	13:38:39	03:33:34	00:27:24
Simulated	Rate estimation - 1 HBD class	101.5	101.5	49.2	00:18:20	00:20:17	00:03:29	00:00:35
Simulated	Rate estimation - 4 classes	988.6	989.6	524.7	06:17:50	07:13:17	00:55:37	00:07:10

Table 2: **Comparison of running times.** The compared software are ZooRoH.f90 (f-EM), RZooRoH using numerical maximization and re-parametrisation (r-optim) or using the EM-algorithm (r-EM). Computations have been performed without constraints. The last column indicates the elapsed running time when the new procedure is run in parallel with eight cores. The maximum number of iterations is constrained to 1,000 with both EM algorithms.