Received: 11 February 2020 | Accepted: 4 August 2020

DOI: 10.1111/2041-210X.13467

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

APPLICATION

Unsupervised detection of ancestry tracks with the GHap R package

Yuri Tani Utsunomiya[1,2] | Marco Milanesi[1,2] | Mario Barbato[3] | Adam Taiti Harth Utsunomiya[1,2] | Johann Sölkner[4] | Paolo Ajmone-Marsan[3] | José Fernando Garcia[1,2,5]

[1]Department of Support, Production and Animal Health, School of Veterinary Medicine of Araçatuba, São Paulo State University (Unesp), Araçatuba/SP, Brazil

[2]International Atomic Energy Agency (IAEA) Collaborating Centre on Animal Genomics and Bioinformatics, Araçatuba/SP, Brazil

[3]Department of Animal Science Food and Nutrition—DIANA and Nutrigenomics and Proteomics Research Center, Università Cattolica del Sacro Cuore, Piacenza, Italy

[4]Division of Livestock Sciences, Department of Sustainable Agriculture System, BOKU—University of Natural Resources and Life Sciences, Vienna, Austria

[5]Department of Preventive Veterinary Medicine and Animal Reproduction, School of Agricultural and Veterinarian Sciences, São Paulo State University (Unesp), Jaboticabal/SP, Brazil

Correspondence
Yuri Tani Utsunomiya
Email: ytutsunomiya@gmail.com

Handling Editor: M. Gilbert

Abstract

1. The identification of ancestry tracks is a powerful tool to assist the inference of evolutionary events in the genomes of animals and plants. However, algorithms for ancestry track detection typically require labelled reference population data. This dependency prevents the analysis of genomic data lacking prior information on genetic structure, and may produce classification bias when samples in the reference data are inadvertently admixed.
2. We combined heuristics with *K*-means clustering to deploy a method that can detect ancestry tracks without the provision of lineage labels for reference population data. The resulting algorithm uses phased genotypes to infer individual ancestry proportions and local ancestry. By piling up ancestry tracks across individuals, our method also allows for mapping loci with excess or deficit ancestry from specific lineages.
3. Using both simulated and real genomic data, we found that the proposed method was accurate in inferring genetic structure, assigning chromosomal segments to lineages and estimating individual ancestry, especially in cases where ancestry tracks resulted from recent admixture of highly divergent lineages.
4. The method is implemented as part of the v2 release of the GHap R package (available at https://cran.r-project.org/package=GHap and https://bitbucket.org/marcomilanesi/ghap/src/master/).

KEYWORDS

admixture, chromosome painting, population structure, single-nucleotide polymorphism

1 | INTRODUCTION

Inference of genetic structure and estimation of individual ancestry are among the most prevalent analyses in studies involving human, animal and plant genomes. Both are usually obtained from genome-wide single-nucleotide polymorphism (SNP) marker data, covering either the whole genome (global analysis) or specific chromosomal segments (local analysis). Whereas global analysis can be conducted without the provision of reference data (Alexander, Novembre, & Lange, 2009; Falush, Stephens, & Pritchard, 2003; Pritchard, Stephens, & Donnelly, 2000; Tang, Peng, Wang, & Risch, 2005), local analyses have been largely limited to scenarios where purebred individuals of known genetic lineages are available to train prediction models (Baran et al., 2012; Churchhouse & Marchini, 2013; Durand, Do, Mountain, & Macpherson, 2014; Guan, 2014; Haals, McCarty, & Payseur, 2013; Hellenthal et al., 2014; Price et al., 2009). The growing interest in the development of unsupervised methods for the prediction of local ancestry is motivated by the

possibility of inferring bottlenecks, divergence, migrations and signatures of selection in populations with unknown genetic structure. Here, we combined heuristics with *K*-means clustering (Hartigan & Wong, 1979) to deploy an algorithm that can simultaneously perform predictions of haplotype structure and ancestry without the supervision of labelled reference population data. The new method is designed to detect the distinct genetic lineages underlying observed haplotypes and classify chromosomal segments according to those lineages.

2 | METHOD DESCRIPTION

2.1 | Implementation

The algorithm was implemented as part of the GHap package (Utsunomiya, Milanesi, Utsunomiya, Ajmone-Marsan, & Garcia, 2016) in R (R Core Team, 2020). The input data consist of phased genotypes, and the package provides conversion functions for popular file formats such as fastPHASE (Scheet & Stephens, 2006), VCF (Browning, Zhou, & Browning, 2018; Danecek et al., 2011) and Oxford HAPS/SAMPLE (Loh et al., 2016; O'Connell et al., 2014).

2.2 | Detection of ancestry tracks

The method starts by grouping all observed haplotypes into *K* pseudo-lineages using the *K*-means algorithm (Hartigan & Wong, 1979). By relying on a general-purpose clustering technique rather than on a specific coalescence or recombination model, our method is essentially a heuristic in approximating ancestral lineages. Prototype alleles are then built for every pseudo-lineage using the arithmetic mean of observed haplotypes initially assigned to that lineage, resulting in per-lineage centroids comprising SNP allele frequencies. Next, the genome is partitioned into a series of overlapping haplotype blocks (i.e. sliding windows), whose sizes and extent of overlap can be controlled by the user. For each block, a given tested haplotype is locally assigned to the pseudo-lineage containing the most similar prototype allele based on the smallest Euclidean distance. In the final step, ancestry tracks are formed by smoothing over the classifications of overlapping blocks using majority voting. Ties are treated as inconclusive classifications (i.e. unknown ancestry). Our algorithm provides default values for parameters such as the size of the haplotype block, the extent of overlap between successive blocks and the number of seeding markers used for *K*-means clustering based on prediction performance in the simulation study presented below. In this manner, the user can opt to run the analysis without tuning the parameters beforehand (albeit recommended for improved results). Importantly, due to the local convergence and stochastic nature of *K*-means, every analysis uses at least 10 random starts of the algorithm. We also encourage users to perform three to five independent runs to evaluate the consistency of the results. For a graphical summary of the classification algorithm, see Figure 1.

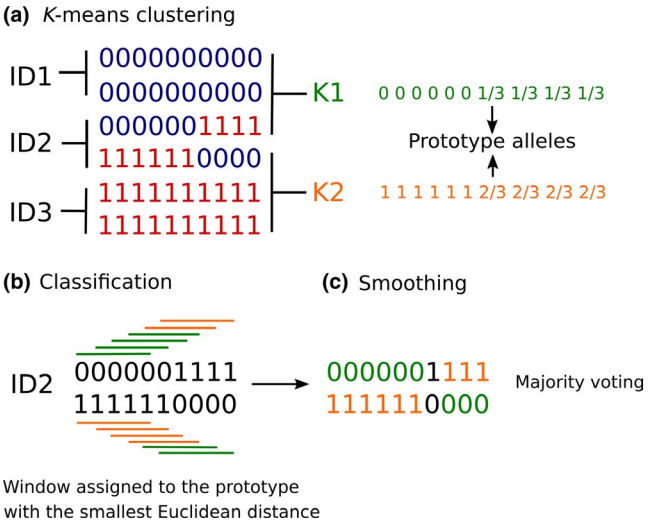

FIGURE 1 Graphical summary of the GHap method for unsupervised detection of ancestry tracks. A toy example is used here, considering two lineages and three diploid individuals (i.e. six haplotypes) with phased genotypes at 10 bi-allelic SNP markers (represented by alleles 0 and 1). Colours red and blue represent true ancestries, whereas green and orange represent predictions (with black meaning unknown). (a) Haplotypes are initially grouped using the *K*-means clustering algorithm. Prototype alleles are obtained as the centroids of the resulting clusters (values were presented as fractions for better visualization). (b) Sliding windows are then used to locally classify each haplotype. (c) In the final step, ancestry tracks are formed by using majority voting of overlapping windows

2.3 | Choice of *K*

The number of pseudo-lineages can be selected using different approaches. For instance, our algorithm can be applied to genomic data with increasing values of *K* in order to reveal ancestral population splits and thus reconstruct the divergence trajectories that gave rise to the observed data, similar to model-based clustering methods such as ADMIXTURE (Alexander et al., 2009). Alternatively, an optimal partitioning of the data can be inferred through the variance ratio criterion (Calinski-Harabasz index) or by applying the classic elbow method to the total within-cluster sum of squares of successive values of *K*.

2.4 | Estimation of ancestry proportions

For each pseudo-lineage, the global ancestry proportion is obtained by simply summing up the sizes of the ancestry tracks assigned to that lineage and dividing that sum by the genome size.

3 | SIMULATION STUDY

3.1 | Overview of the simulated data

The GHap method was applied to 1,215 simulated datasets of a single 100 cM autosome, which were generated with varying marker

densities (1 marker every 1, 2 or 20 kbp), number of divergent lineages (2, 4 or 8), degree of lineage divergence (F_{ST} of 0.10, 0.20 or 0.30) and age of admixture (5, 10 or 50 generations ago). The details of the simulation procedure and results on accuracy performance are described below.

3.2 | Simulation of divergent lineages

We used the QMSim v1.10 software (Sargolzaei & Schenkel, 2009) to simulate a genome consisting of a single autosome of 100 cM with *M* evenly spaced bi-allelic SNP markers. This genome was initialized in an historic population of 1,000 individuals (500 males and 500 females) that randomly mated for 1,000 generations in order to create linkage disequilibrium between markers. In each mating season, gametes were produced by crossing-over haplotypes at random sites across the 100 cM sequence. The number of recombination events per gamete was sampled from a Poisson distribution with mean parameter $\lambda = 1$. After 1,000 generations, the historic population suffered a bottleneck and divided into *K* divergent lineages following a star-shaped genealogy. Each lineage was established with 200 individuals (100 males and 100 females) and let to neutrally evolve in parallel under genetic drift for *G* generations. Variable quantities *M*, *K* and *G* are discussed below.

3.3 | Simulation of test samples

After the divergent lineages were originated, ancestral haplotypes obtained with QMSim were used to simulate the test data that were effectively used to assess the accuracy of the method. A total of 50 purebred individuals for each one of the *K* divergent lineages (i.e. $K \times 50$ purebred individuals) and 100 admixed individuals were created. Each simulated individual was generated following the pseudocode below (adapted from Guan, 2014; Lawson, van Dorp, & Falush, 2018; Leslie et al., 2015; Price et al., 2009):

```
Sample ancestry proportions q_1, q_2, ..., q_K from Q ~ Dir(α);
For each one of the two new haplotypes:
  Initialize j = 0;
  While j < M:
    Set i = j + 1;
    Sample lineage k ∈ {1,2 ... K} with probabilities q_1, q_2, ..., q_K;
    Sample ancestral haplotype h from lineage k;
    Sample window size x from X ~ Exp(β);
    Evaluate j = min(i + x, M);
    Copy the segment [i; j] from haplotype h;
```

Hyperparameters $\boldsymbol{\alpha}$ for the Dirichelet and β for the exponential distributions were chosen as follows. For admixed individuals, $\boldsymbol{\alpha}$ was set as a *K*-dimensional vector with all elements equal to 1, which yields a wide range of configurations of ancestry proportions and thus individuals with very different ancestry profiles within a single simulation replicate (including some individuals with $q_k = 0$). For the simulation of a purebred individual from lineage *k*, $\boldsymbol{\alpha}$ was set as a *K*-dimensional vector with the *k*th element equal to 1 and all remaining elements equal to 0. The mean parameter of the exponential distribution was defined as $\beta = L^{-1}$, where *L* is the average ancestry track length expressed in number of markers. Recalling that the expectation of the ancestry track length in Morgans after *g* generations of the admixture event is $1/(2g)$, we have $L = d[100/(2g)]$, where *d* is the marker density in the genome expressed in number of markers per cM.

3.4 | Simulation replicates

We used the above simulation framework to generate data under a variety of scenarios by changing the values of the following parameters:

Marker density

M = [5,000 50,000 100,000]: reflecting SNP panels of low (1 marker/20 kbp), moderate (1 marker/2 kbp) and high density (1 marker/1 kbp) respectively.

Number of divergent lineages

K = [2 4 8]: translating into small, moderate and large number of divergent lineages respectively.

Divergence time in generations

G = [100 250 500]: representing low, moderate and high divergence among lineages respectively. We verified that the selected values of *G* resulted in average divergences of F_{ST} = [0.10 0.20 0.30], with deviations no greater than 0.03 between replicates.

Age of admixture in generations

g = [5 10 50]: indicating very recent, recent and remote admixture respectively.

Since our method requires the selection of a haplotype block size L' for local classifications, we created an extra parameter g' = [5 10 50] such that in a given analysis we could calculate $L' = d[100/(2g')]$. This was done in order to see the influence of the choice of haplotype block sizes in the accuracy of predictions. Each combination of *M*, *K*, *G*, *g* and g' was tested in five independent replicates, resulting in 1,215 different simulations.

3.5 | Accuracy of ancestry predictions

Our method was accurate in inferring genetic structure, assigning chromosomal segments to lineages and estimating individual ancestry, especially in cases where ancestry tracks resulted from recent admixture of highly divergent lineages. However, an analysis of variance (ANOVA) indicated that accuracy of local ancestry predictions significantly decreased ($p < 0.05$) as the number of divergent lineages increased, the degree of divergence among lineages decreased and the length of ancestry tracks shortened (Figure 2a). Reduced accuracy was also observed when large haplotype blocks were used to predict ancient admixture or small

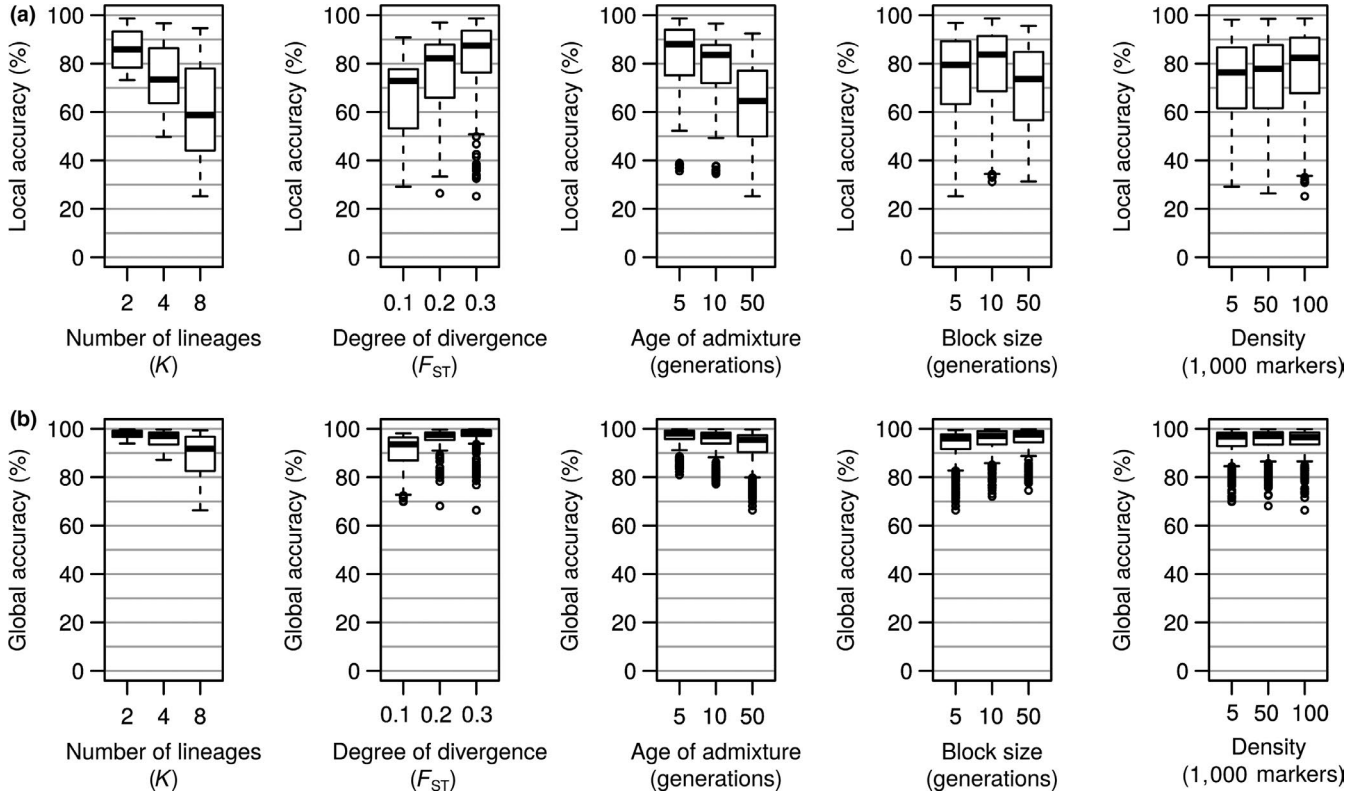

FIGURE 2 Accuracy of ancestry predictions using the GHap unsupervised method in 1,215 simulation replicates of a 100 cM chromosome. Local (a) and global (b) ancestry predictions were assessed according to the number of ancestral lineages (*K*), the degree of divergence between lineages (F_{ST}), the number of generations since the admixture event, the choice of haplotype block size, and the number of markers

haplotype blocks were used for the prediction of very recent admixture. Nevertheless, the use of haplotype blocks compatible with recent admixture (~10 generations) retained reasonable accuracy across most scenarios, regardless of the age of admixture. Conversely, marker density in the range of values tested here did not have a significant effect ($p > 0.05$) on accuracy. Global ancestry estimation was not significantly impacted by any of the mentioned factors and remained highly accurate in most scenarios (Figure 2b). Furthermore, the inference of *K* was correct in approximately 70% of all simulations. The prediction accuracy and the inferred value of *K* obtained in each analysis replicate can be seen in the Supplementary Results.

4 | APPLICATION TO REAL DATA

We applied the new method to the Human HapMap 3 dataset, which included 1,011 individuals from 11 different populations genotyped at 1,387,394 SNPs (Altshuler et al., 2010). Prior to the analysis, genotypes were phased with Eagle v2.4.1 (Loh et al., 2016). We paralleled our analysis with supervised predictions made by Random Forest (RFMix) (Maples, Gravel, Kenny, & Bustamante, 2013) and support vector machines (SVM; Durand et al., 2014; Haasl et al., 2013), methods that are currently being used in large scale commercial applications by personal genomics services such as 23andMe (https://www.23andme.com/) and Embark (https://embarkvet.com/). We obtained RFMix v2 from its GitHub repository (https://github.com/slowkoni/rfmix), whereas SVM was implemented in GHap by building upon the e1071 package (Meyer et al., 2019). In our implementation of SVM, instead of using the string kernel that requires pairwise comparisons of every possible *k*-mer in a segment, we followed the simpler approach of Haasl et al. (2013) and used the Gaussian radial basis function kernel, since it also indirectly captures complex interactions among feature vectors (i.e. markers within a segment). In addition, we compared our global ancestry predictions with those yielded by the ADMIXTURE software (Alexander et al., 2009). Our unsupervised method selected $K = 3$ as the optimal partitioning of the data, separating human populations into three lineages: Africans, Europeans and Asians (Figure 3). These results are consistent with previous inference of lineage structure in human populations (Alexander et al., 2009). ADMIXTURE, SVM, RFMix and GHap presented highly correlated predictions of global ancestry (Figure 4). Our method further produced local ancestry results that were similar to those obtained with SVM or RFMix, with average concordance over 90%, but with the advantage of not requiring prior information on the source of ancestry (Figures 5 and 6). Our method was also capable of replicating the African signature of selection in the MHC locus previously found in Mexicans (Figure 7) by Guan (2014).

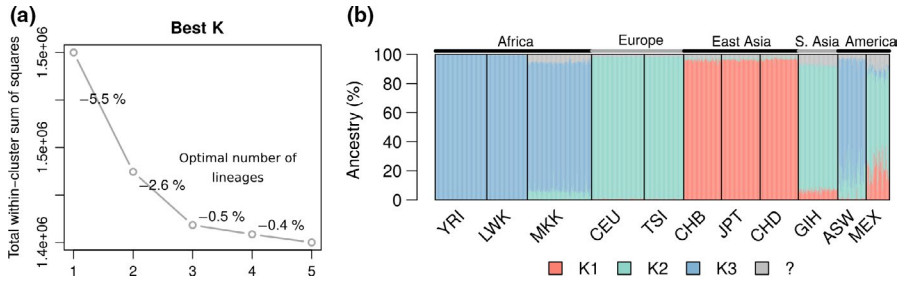

FIGURE 3 Application of the GHap unsupervised method to the Human HapMap 3 build 36 dataset. (a) Identification of the optimal number of lineages underlying the observed haplotypes. (b) Global ancestry results for 11 human populations (YRI = Yoruba in Ibadan, Nigeria; LWK = Luhya in Webuye, Kenya; MKK = Maasai in Kinyawa, Kenya; CEU = Utah residents with European ancestry; TSI = Toscani in Italy; CHB = Han Chinese in Beijing, China; JPT = Japanese in Tokyo, Japan; CHD = Chinese in Metropolitan Denver, CO; GIH = Gujarati Indians in Houston, TX; ASW = African ancestry in Southwest USA; MEX = Mexican ancestry in Los Angeles, CA)

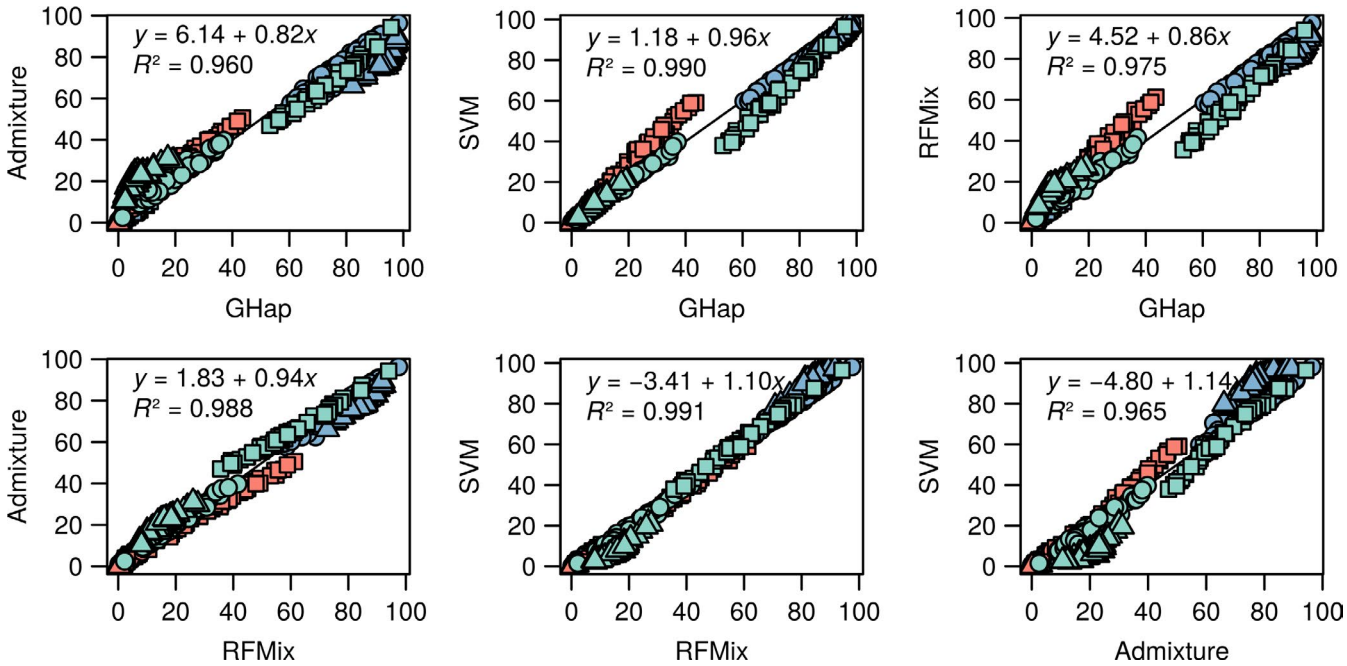

FIGURE 4 Correlations of global ancestry estimation between ADMIXTURE (unsupervised), SVM (supervised), RFMix (supervised) and GHap (unsupervised) in the Human HapMap 3 build 36 dataset. Correlations of African (blue), Asian (red) and European (green) ancestry were computed using data from ASW (dot), MEX (square) and MKK (triangle). Each individual is represented by three points in the plot, corresponding to the three sources of ancestry. For the supervised analyses in RFMix and SVM, the following reference data were used: African (YRI and LWK), Asian (CHB, JPT and CHD) and European (CEU and TSI). GHap and ADMIXTURE were ran with the entire HapMap data assuming $K = 3$

5 | BENCHMARKING

All analyses in this study were performed with R v3.4.4 under Ubuntu 16.04.5 LTS (xenial) installed in a Dell PowerEdge VRTX M630 blade equipped with 256 GB RAM and 2 × Intel Xeon E52640 2.4 GHz processors. The analysis of the Human HapMap 3 data took 78 and 18 min to complete using 4 and 40 cores respectively. In order to evaluate how the method scaled with data size and analysis complexity, we used our simulation procedure to generate additional datasets consisting of 100, 500 and 1,000 individuals. Since the degree of divergence and the length of ancestry tracks influenced only the time spent to generate the data but not the analysis per se, our benchmarking was performed on replicates with fixed values of $G = 100$ and $g = 10$. We let all remaining parameters, namely K, M and g', to vary according to the previously described simulations. We then analysed the resulting data using 4, 8 and 16 cores, mimicking resources often available in home computers. Each combination of parameters was replicated five times, generating 1,215 additional analyses. The results of these analyses are found in the Supplementary Results. Briefly, using only four cores, the average processing time of 100 individuals and 5,000 markers was <4 s, whereas the analysis of 1,000 individuals and 100,000 markers required 5 min to complete.

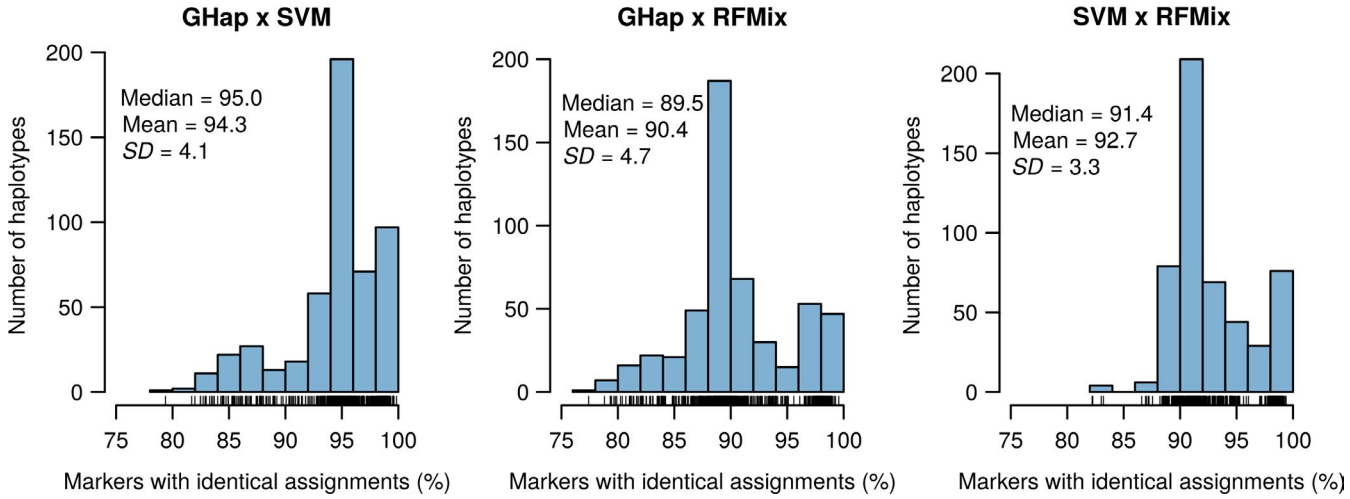

FIGURE 5 Concordance of local ancestry between GHap (unsupervised), SVM (supervised) and RFMix (supervised) in the Human HapMap 3 build 36 dataset. For each haplotype, concordance was measured as the percentage of markers with identical assignments between methods, and was computed using data from MKK, ASW and MEX. For the supervised analyses in RFMix and SVM, reference populations were defined as: African (YRI and LWK), Asian (CHB, JPT and CHD) and European (CEU and TSI)

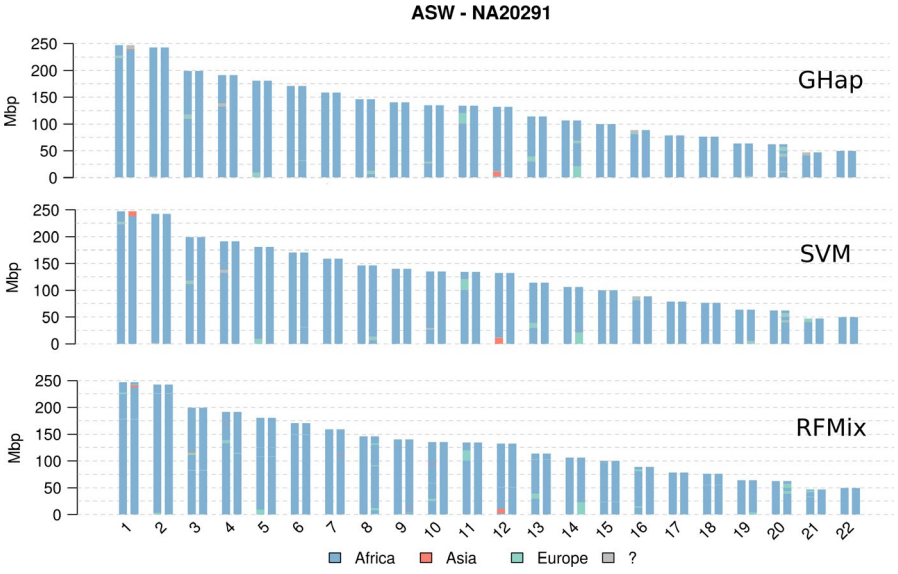

FIGURE 6 Karyoplot (chromosome painting) of one individual from the ASW (African ancestry in Southwest USA) population. For the supervised analyses in RFMix and SVM, reference populations were defined as: African (YRI and LWK), Asian (CHB, JPT and CHD) and European (CEU and TSI)

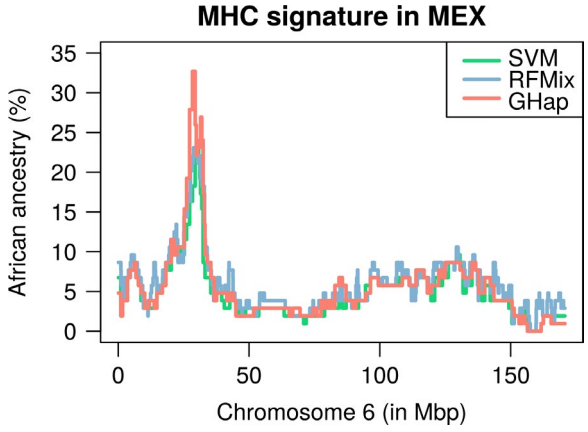

FIGURE 7 Ancestry mapping showing excess African ancestry towards the MHC locus in Mexicans

6 | CONCLUSION

We developed a method for the unsupervised inference of global and local lineage structure in genomes of unknown ancestry. Although the method can be used to analyse human populations, we conjecture that research groups working with SNP data of both domesticated and wildlife species of diploid animals and plants are expected to benefit the most from our implementation.

ACKNOWLEDGEMENTS

The authors declare that there are no conflicts of interest.

AUTHORS' CONTRIBUTIONS

Y.T.U. conceived the method and the simulation procedures; Y.T.U., M.M. and M.B. implemented the package; Y.T.U., M.M. and A.T.H.U.

tested the package in the simulated and real datasets; J.S., P.A.-M. and J.F.G. contributed with discussions to improve the method; Y.T.U. wrote the manuscript. All authors revised and approved the final version of the manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13467.

DATA AVAILABILITY STATEMENT

Human HapMap 3 dataset is publicly available and can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/hapmap/. The method reported here is implemented as part of the v2 release of the GHap R package, which is available at https://cran.r-project.org/package=GHap and https://bitbucket.org/marcomilanesi/ghap/src/master/.

ORCID

Yuri Tani Utsunomiya https://orcid.org/0000-0002-6526-8337

REFERENCES

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

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., … McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*, 52–58.

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., … Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, *28*, 1359–1367. https://doi.org/10.1093/bioinformatics/bts144

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A one-penny imputed genome from next generation reference panels. *American Journal of Human Genetics*, *103*, 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Churchhouse, C., & Marchini, J. (2013). Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic Epidemiology*, *37*, 1–12. https://doi.org/10.1002/gepi.21692

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Durand, E. Y., Do, C. B., Mountain, J. L., & Macpherson, J. M. (2014). Ancestry composition: A novel, efficient pipeline for ancestry deconvolution. *BioRxiv*, 010512.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*, 1567–1587.

Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. *Genetics*, *196*, 625–642. https://doi.org/10.1534/genetics.113.160697

Haasl, R. J., McCarty, C. A., & Payseur, B. A. (2013). Genetic ancestry inference using support vector machines, and the active emergence of a unique American population. *European Journal of Human Genetics*, *21*, 554–562. https://doi.org/10.1038/ejhg.2012.258

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A *K*-means clustering algorithm. *Applied Statistics*, *28*, 100–108. https://doi.org/10.2307/2346830

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*, 747–751. https://doi.org/10.1126/science.1243518

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*, 3258. https://doi.org/10.1038/s41467-018-05257-7

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., … Bodmer, W. (2015). The fine-scale genetic structure of the British population. *Nature*, *519*, 309–314. https://doi.org/10.1038/nature14230

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., … Price, A. L. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*, 1443–1448. https://doi.org/10.1038/ng.3679

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *American Journal of Human Genetics*, *93*, 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2019). *Misc functions of the department of statistics (e1071), probability theory group (Formerly: E1071), TU Wien*. R package version 1.7-3. Retrieved from https://cran.r-project.org/web/packages/e1071/index.html

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., … Marchini, J. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics*, *10*, e1004234. https://doi.org/10.1371/journal.pgen.1004234

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., … Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, *5*, e1000519. https://doi.org/10.1371/journal.pgen.1000519

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

R Core Team. (2020). *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: A large-scale genome simulator for livestock. *Bioinformatics*, *25*, 680–681. https://doi.org/10.1093/bioinformatics/btp045

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, *78*, 629–644. https://doi.org/10.1086/502802

Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, *28*, 289–301. https://doi.org/10.1002/gepi.20064

Utsunomiya, Y. T., Milanesi, M., Utsunomiya, A. T. H., Ajmone-Marsan, P., & Garcia, J. F. (2016). GHap: An R package for genome-wide haplotyping. *Bioinformatics*, *32*, 2861–2862. https://doi.org/10.1093/bioinformatics/btw356

SUPPORTING INFORMATION

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

How to cite this article: Utsunomiya YT, Milanesi M, Barbato M, et al. Unsupervised detection of ancestry tracks with the GHap R package. *Methods Ecol Evol*. 2020;00:1–7. https://doi.org/10.1111/2041-210X.13467