Received: 23 July 2023 | Revised: 23 April 2024 | Accepted: 8 May 2024

DOI: 10.1111/1755-0998.13981

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Assessing the limits of local ancestry inference from small reference panels

Sandra Oliveira[1,2] | Nina Marchi[1,2,3] | Laurent Excoffier[1,2]

[1]CMPG, Institute for Ecology and Evolution, University of Bern, Berne, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]UMR7206 Eco-Anthropologie, CNRS-MNHN-Université Paris Cité, Paris, France

Correspondence
Sandra Oliveira, CMPG, Institute for Ecology and Evolution, University of Bern, 3012 Berne, Switzerland.
Email: sandra.dasilvaoliveira@unibe.ch

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 310030_188883

Handling Editor: Brent Emerson

Abstract

Admixture is a common biological phenomenon among populations of the same or different species. Identifying admixed tracts within individual genomes can provide valuable information to date admixture events, reconstruct ancestry-specific demographic histories, or detect adaptive introgression, genetic incompatibilities, as well as regions of the genomes affected by (associative-) overdominance. Although many local ancestry inference (LAI) methods have been developed in the last decade, their performance was accessed using large reference panels, which are rarely available for non-model organisms or ancient samples. Moreover, the demographic conditions for which LAI becomes unreliable have not been explicitly outlined. Here, we identify the demographic conditions for which local ancestries can be best estimated using very small reference panels. Furthermore, we compare the performance of two LAI methods (RFMix and MOSAIC) with the performance of a newly developed approach (simpLAI) that can be used even when reference populations consist of single individuals. Based on simulations of various demographic models, we also determine the limits of these LAI tools and propose post-painting filtering steps to reduce false-positive rates and improve the precision and accuracy of the inferred admixed tracts. Besides providing a guide for using LAI, our work shows that reasonable inferences can be obtained from a single diploid genome per reference under demographic conditions that are not uncommon among past human groups and non-model organisms.

KEYWORDS
admixture, local ancestry inference, MOSAIC, RFMix, simpLAI

1 | INTRODUCTION

Genetic admixture occurs when individuals from divergent populations interbreed. Due to recombination after the admixture event, the chromosomes inherited from each parent break down, forming genomes made up of a mosaic of DNA segments tracing back to the ancestral populations. Previous studies have shown that admixture took place recurrently in the evolutionary history of many populations and species and often played a prominent role in shaping their genomes (Edelman & Mallet, 2021; Martin & Jiggins, 2017; Moran et al., 2021). Admixture events can impact the fitness of individuals and contribute to an adaptation to local environments (often referred to as adaptive introgression) (Edelman & Mallet, 2021). In certain cases, admixed individuals benefit from enhanced fitness (hybrid vigour) due to overdominance or associative overdominance (Birchler et al., 2006), contributing to the retention of divergent haplotypes in the population and to elevated levels of heterozygosity in conserved regions of the genome. However, admixture might also lead to genetic incompatibilities and purifying selection, which can be predicted by the identification of introgression deserts (Martin & Jiggins, 2017). Decoding admixed genomes in terms of their ancestral origins is therefore of prime importance for investigating the consequences of introgression, reconstructing ancestry-specific demographic histories, and dating past admixture events (Browning et al., 2018; Ioannidis et al., 2020).

Several local ancestry inference (LAI) tools were developed for these purposes or for applications in medical genetics. Many of them use Hidden Markov Models (HMM) that account for linkage disequilibrium (LD) (e.g. SABRE, HAPAA, HAPMIX, ELAI, MOSAIC, ARCHes, FLARE) (Browning et al., 2023; Guan, 2014; Price et al., 2009; Salter-Townshend & Myers, 2019; Sundquist et al., 2008; Tang et al., 2006; Wang et al., 2021) or window-based approaches that do not explicitly model LD (e.g. LAMP, WINPOP) (Paşaniuc et al., 2009; Sankararaman et al., 2008); others use machine learning (ML) techniques such as random forests with a conditional random field (RFMix) (Maples et al., 2013), neural networks (AncestralPaths) (Pearson & Durbin, 2023), or a combination of different ML algorithms (GNOMIX) (Hilmarsson et al., 2021). Most LAI tools show high accuracy when applied to individuals whose admixing ancestries are well differentiated and when admixture is very recent (Geza et al., 2019). However, the amount of differentiation required to reach a certain accuracy level and the demographic conditions for which inferences become unreliable have not been explicitly outlined. In addition, masking regions of the genome for which LAI is uncertain based on marginal probabilities might be insufficient to remove LAI errors since these probabilities are not always well calibrated (Browning et al., 2023).

The most recent improvements in LAI methods aimed at increasing computational speed and accuracy while taking advantage of large-scale reference panels (Hilmarsson et al., 2021; Wang et al., 2021). Yet, large datasets consisting of high-coverage sequence or dense genotype data are often unavailable for non-model organisms or ancient samples. Using reference samples that are temporally closer to the admixture event could, however, improve LAI due to their higher genetic similarity to the actual sources, particularly when all extant populations from a region have experienced admixture and individuals that are representative of the ancestral genetic pool are no longer found.

In this work, we simulate various demographic models, including single admixture events and recurrent admixture, to identify the conditions for which local ancestries can be reasonably estimated when appropriate proxies for the ancestral populations are limited. Using a set of informative statistics, we compare the performance of three LAI methods: a widely used ML-based method (RFMix) (Maples et al., 2013), an HMM-based method that does not require direct surrogates for the admixing groups (MOSAIC) (Salter-Townshend & Myers, 2019), and a newly developed, simpler approach (that we named simpLAI), that is less affected by reference panel sizes. We show how the inferences vary with the choice of a prior admixture time (required in RFMix) or window size (required in simpLAI) and with the choice of reference populations; and propose post-painting filtering steps to reduce false-positive rates and improve the precision and accuracy of the estimated admixed tracts. Lastly, we applied simpLAI to admixed southern African modern humans, as well as admixed Neolithic farmers from Europe. LAI obtained using single genomes per reference was contrasted to inferences based on larger reference panels.

2 | METHODS

2.1 | Simulations

We simulated admixed populations resulting from one and two admixture events at varying times (Figure S1) using a modified version of *fastsimcoal2* (Excoffier et al., 2021) that records the local ancestry tracts of admixed individuals. The simulations include demographic models where two populations (S1 and S2) of equal size (2000, 5000, or 10,000 haploid individuals) diverge 500 or 1000 generations ago from an ancestral population of size 20,000. In the one pulse of admixture model, admixed populations of size 2000 are created 10, 100, or 300 generations ago, with 5%, 10%, 20%, or 30% of ancestry from population S1 and the remaining from population S2. In the two-admixture pulse model, an initial admixture event occurs 100 or 300 generations ago between populations S1 and S2, contributing 30% and 70%, respectively, to the ancestry of the admixed population (SA), which has the same size as the source populations. A second admixed population of 2000 haploid individuals is then formed 10 generations ago as a result of admixture between S1 (5%) and SA (95%). We sampled 10 diploid individuals from the admixed population (to be used as LAI targets) and one, two, four, or eight diploid individuals from each source to be used as references in the LAI for the one-pulse model and one or four diploid individuals for the two-pulse model. In our simulations, each haploid genome consists of a chromosome of length 100 Mb. We used uniform mutation and recombination rates, set to 1.25×10^{-8} mutations per base pair per generation (Scally & Durbin, 2012) and 1×10^{-8} per base pair per generation, respectively.

2.2 | Local ancestry inference

We inferred the local ancestries with RFMix v2 (Maples et al., 2013), MOSAIC (Salter-Townshend & Myers, 2019), as well as with a simple and newly developed LAI approach (simpLAI; available on https://github.com/CMPG/simpLAI). RFMix was run with the option *–reanalyze-reference* and three EM iterations, to account for cases where the reference population is already admixed. We performed three RFMix estimations per simulation assuming a prior of 10, 100, and 300 generations since admixture. For each haploid target, regions of the genome with a resulting marginal probability smaller than 1 were masked to minimize LAI errors. We ran MOSAIC for each simulation assuming two ancestral populations (−*a* 2) and without phase correction (−*nophase*). Since the phase is known and reference panels are small, allowing rephasing could have introduced phasing errors. By default, MOSAIC infers the relationship between reference and ancestral populations, recombination rates, mutation rates, and the timing and ancestry proportions of the admixture event. Therefore, we did not set any prior values for these parameters. LAI performance was estimated after converting the MOSAIC inferences at evenly spaced recombination distances to inferences at the single-nucleotide positions (SNP) with the MOSAIC function “grid_to_pos”. We used a uniform recombination map when performing LAI inferences on simulated data with RFMix and MOSAIC.

simpLAI is a window-based approach that leverages information on the matching of haplotypes from admixed and reference individuals to infer the ancestry across the genome. The programme performs two types of inferences. In the *min* mode, the ancestry assigned to a target haplotype of predefined size (−*s*) is that of the reference population that includes the chromosome displaying the lowest number of mismatches with the target. The *rec* mode accounts for intra-population ancestry switches that may have occurred within a haplotype through recombination. Each haplotype segment consists of *n* polymorphic sites, and the number of mismatches between a target and each chromosome of the reference population is computed on subsets of *t* linked polymorphic sites (assumed to be a non-recombining segment). The path consisting of the combination of segments (potentially identified on different chromosomes) among the *n* polymorphic sites showing the smallest number of mismatches within each reference population is selected to compare reference populations. The reference with the shortest path is selected for the ancestry of the target haplotype. The increment for the sliding window is defined in base pairs (−*i*) for the *min* mode and in number of polymorphic sites (−*m*) for the *rec* mode.

We tested simpLAI using windows ranging from 0.5 to 4 Mb and including 500–8000 polymorphic sites. The parameter *t* was set to five linked polymorphic sites since this value led to a better LAI performance compared to higher values in our exploratory analyses.

2.3 | Statistics

The LAI performance was evaluated for the three methods for each simulated condition. We calculate the *accuracy* to determine how close the inferred local ancestry is to the true local ancestry across the whole genome. This estimator has been commonly used to assess the performance of LAI methods but is highly affected by admixture proportions (e.g. if the minority ancestry accounts for 5% of the genome, the accuracy can be relatively high even if none of the segments assigned to the minority ancestry were correctly inferred). Therefore, we computed three additional statistics: the *precision* of the inference to quantify how much of the predicted minority ancestry is really true, the *true-positive rate* (TPR) to quantify how much of the minority ancestry is recovered (i.e. from the old vs. the recent pulse in the two-pulse case), and the *false-positive rate* (FPR) to quantify how much of the majority ancestry is incorrectly assigned to the minority ancestry (as defined in Figure S1).

We compared LAI results across runs performed with different parameters (time since admixture in RFMix) and different reference populations (RFMix, MOSAIC, simpLAI) for each simulated target individual. Regions of the genome for which an individual's LAI was inconsistent across runs were masked, the percentage of the genome remaining painted (i.e. the concordant subset) was recorded, and the statistics described above were additionally computed on the concordant subset of the genome.

We computed F_{ST} (Hudson et al., 1992) between the simulated reference populations at present using a custom R script (available on https://github.com/CMPG/simpLAI). To measure the level of differentiation between populations at different times in the past, we calculated the expected F_{ST} as a function of coalescence times (Slatkin & Voelm, 1990).

2.4 | Application to southern African individuals

To further evaluate the applicability of LAI in extreme cases where only a single diploid genome is available to represent a particular source population, we applied simpLAI to admixed modern humans from southern Africa. RFMix was previously used to infer LAI of the same present-day individuals from this region using references for three possible ancestries consisting of present-day individuals from East, West, and southern Africa (Oliveira et al., 2023). While the latter were chosen for their high amount of “original” southern African ancestry (i.e. the ancestry of southern Africans before contact with food-producing groups), all present-day southern African individuals have been shown to be admixed to some degree with western and eastern African-related groups. This condition is not ideal for LAI but is in principle accounted for in RFMix by additional EM steps that treat the reference haplotypes as query haplotypes, updating their own ancestry assignment. Here, we used simpLAI with the same reference populations and also performed an alternative LAI analysis using a single ancient genome (Ballito Bay A) as a proxy for the original southern African ancestry (Schlebusch et al., 2017). In contrast to present-day southern Africans, this individual (dated 1986–1831 BP) does not show any genetic contributions from East and West Africa. We used the same reference individuals as in Oliveira et al. (2023) to represent the East and West African ancestries (13 Somali and 13 Yoruba individuals, respectively). Yet, to avoid biased LAI results emerging from sample size asymmetries in the references (one individual for the ancient southern African reference and 13 individuals for each of the two other reference samples), we performed multiple LAI runs, each of them using as references the ancient Ballito Bay, one Yoruba, and one Somali individual. The LAI results from all possible combinations of reference individuals were then converted into a majority LAI. Note that by taking advantage of a larger number of individuals for two of the sources, we expect to obtain better inferences compared to using a single individual per reference.

To be able to compare the LAI results obtained with different approaches while avoiding differences due to phasing, we used previously phased data (Oliveira et al., 2023), consisting of 63 African populations genotyped on the Affymetrix Axiom Genome-Wide Human Origins Array (Lazaridis et al., 2014; Oliveira et al., 2023; Patterson et al., 2012; Pickrell et al., 2012). The diploid genotype calls for the ancient individual (Schlebusch et al., 2017) were first phased together with the same African dataset as in Oliveira et al. (2023) using Beagle 4.1 (Browning & Browning, 2007), and the phased ancient genome was then merged with the previously phased present-day individuals. A total of 492,413 polymorphic SNPs and 305 southern African target individuals were used in simpLAI. The software was run using the options −*n* 1000 −*m* 500 −*t* 5. The final chromosome paintings were plotted with *tagore* (Rishishwar et al., 2015), together with the paintings obtained in Oliveira et al. (2023).

We tested the relationship between global ancestry estimates obtained with different approaches by applying an Analysis of Covariance (ANCOVA) where the West and southern African ancestries are used as the categorical independent variable. Note that the East African ancestry was not included in the analysis since this ancestry is a linear combination of the other two.

2.5 | Application to Neolithic farmers from Europe

We then performed LAI in early Neolithic genomes from Europe. A previous paper (Marchi et al., 2022) showed that these European early farmers were a mixture of Western Hunter-Gatherers (WHG) and a population related to the ancestors of an Iranian early farmer (WC1), here referred to as Eastern Early Farmers (EEFs). European early farmers descend from Neolithic people from the Aegean region (Hofmanová et al., 2016) who already had ~30% of WHG ancestry, and further admixed (~5 to 10%) with WHGs when settling in Europe (Marchi et al., 2022). This scenario of multiple admixture pulses, in which an already admixed population receives further ancestry from the minority source, is analogous to the complex admixture scenario we simulated in Figure S1b. Using the ancient genomes sequenced at depth larger than 10X (Marchi et al., 2022), we identified WHG ancestry tracts in early European farmers based on (*i*) one unadmixed and one admixed source (corresponding to S1 and SA sources in Figure S1b, respectively) and (*ii*) based on two unadmixed sources (corresponding to S1 and S2 sources in Figure S1b, respectively). In the first approach (*i*), we used four European WHGs (Bichon, Loschbour, VLASA7, and VLASA32) and four early farmers from the Aegean region (Bar25, AKT16, Nea2, and Nea3) as a proxy for the WHG and farmer ancestry, respectively. In the second approach (*ii*), we used WC1 as a proxy for the unadmixed ancestral EEF component, and each of the four European HGs was used separately in four LAI runs as a proxy for the WHG ancestry to preserve balanced sample size of reference panels. As in the application to southern Africans, we report in this latter case the majority of LAI based on the four replicate runs. In case of potential ties in local inferred ancestries, we considered that the local ancestry was the same as that inferred in flanking SNPs if the majority of LAI at those SNPs was identical (i.e. we assumed no ancestry switch in this case). When the majority of LAI was different in flanking SNPs, we excluded these regions with LAI ties from further analyses.

simpLAI was run for each source set using the options −*n* 2000 −*m* 1000 −*t* 5. RFMix was run only for the first approach due to its sample size requirements, assuming 30 generations since admixture. The local and global ancestries were compared as described for the southern African application.

3 | RESULTS

3.1 | LAI performance based on simulations

We first investigated the performance of RFMix when the time since admixture—a required parameter – is incorrectly specified (Figure 1 and Figure S2). While assuming an admixture time that is equal to the true value maximizes the recovery of ancestry from the minority source (measured by the TPR), we found that in general assuming old times since admixture (i.e. 100 or 300 generations instead of 10 generations) leads to lower precision and higher FPR, even if the true admixture time is old (Figure 1). This pattern is consistent for different levels of differentiation between source populations (Figure S2), except in the limiting case where F_{ST} between sources at the time of admixture is very low (~0.01) since inferences become extremely poor regardless of the assumed admixture time (Figure S2c). Our results thus suggest that even if admixture is known to be old (e.g. from historical records), using the true time of admixture is not always ideal. Instead, for some downstream applications of the LAI, it might be better to assume a more recent admixture time to reduce false positives. Since admixture times are often unknown and estimating them can be challenging, an alternative solution is to perform inferences assuming different admixture times and then consider only regions of the genome for which the LAI is concordant. For instance, we obtained an improved performance when combining inferences from 10 and 100 generations (10 ∩ 100 gen), as well as 10 and 300 generations (10 ∩ 300 gen), for simulations in which admixture occurred 10, 100, or 300 generations ago (Figure 1 and Figure S2). After masking the LAI that was inconsistent between runs with an assumed time since admixture of 10 and 100 generations, 92%–99% of the genome remained painted (Figure S4). A slightly lower percentage of the genome (81%–94%) remained painted for assumed times since admixture of 10 and 300 generations (Figure S4).

We also investigated the influence of key parameters in the performance of simpLAI (Figure S5). Under the tested models, windows of 2 Mb or containing 4000 polymorphic sites provide the best results for very recent admixture times (10 generations) across all

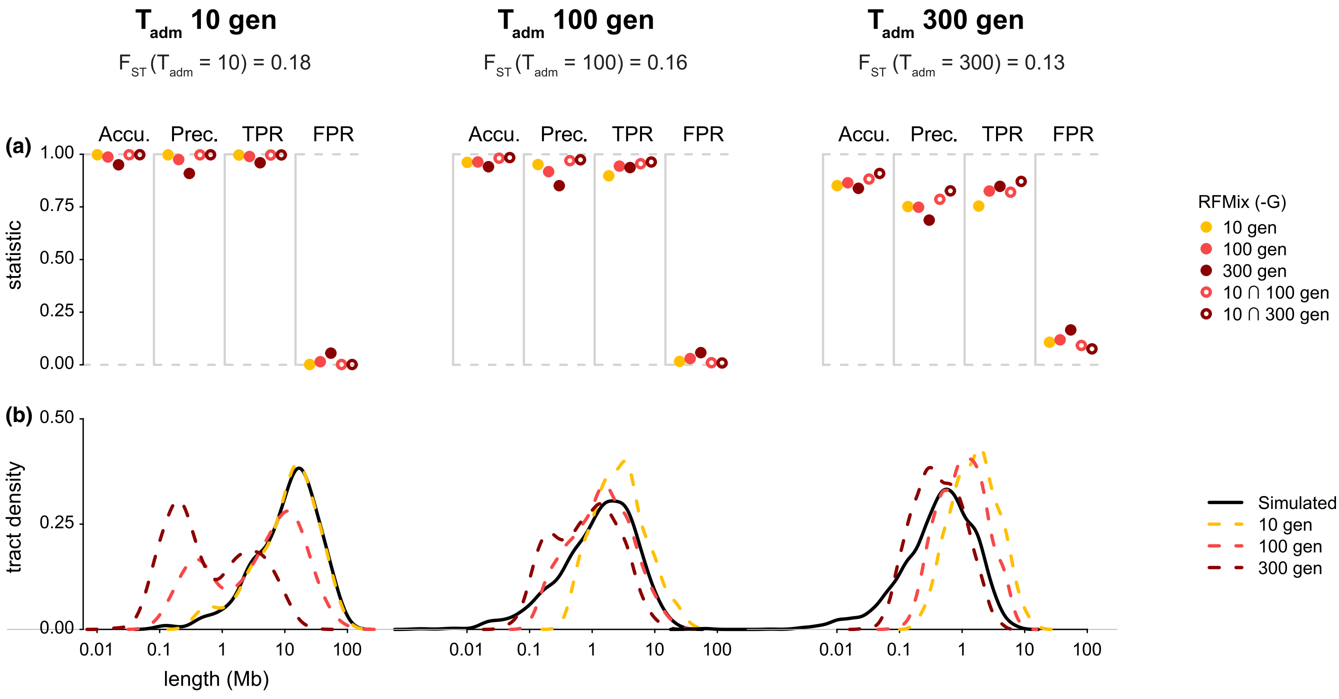

FIGURE 1 Impact of mis-specifying the time since admixture on the performance of RFMix. Results for simulations of one pulse of admixture (30%), occurring 10, 100, and 300 generations ago (T_{adm}) between two populations with effective size 5000 that diverged 1000 generations ago (F_{ST} at present = 0.18). (a) Measures of performance using eight diploid individuals per reference and assuming various times since admixture (−G 10, 100, 300) and combining the LAI from runs obtained under two different times since admixture (10 ∩ 100; 10 ∩ 300). (b) Densities of log-transformed tract lengths for the (true) simulated data and for the inferred local ancestry under different parameters. The x-axis is in log scale (see the untransformed exponential densities in Figure S3). Accu., accuracy; Prec., precision; TPR, true-positive rate; FPR, false-positive rate; gen, generations.

statistics, while windows of 1 Mb or 1000–2000 polymorphic sites provide a good trade-off between TPR and FPR for older admixture times (100–300 generations). Our results indicate that similar performance can be obtained using the *rec* and the *min* inference modes. However, for some of the demographic conditions, the *rec* mode slightly outperforms the *min* mode.

The length distribution of genomic segments from different ancestral populations is informative about the time since admixture and has been used for dating admixture events (Chimusa et al., 2018). Yet, the identification of segments in most LAI approaches (except MOSAIC (Salter-Townshend & Myers, 2019) and Ancestry HMM (Corbett-Detig & Nielsen, 2017)) requires specifying a prior admixture time or some window length-related parameters. This dependency is often overlooked, and its impact on the resulting admixture estimates remains unclear. We show that assuming a time since admixture that is far from the true admixture time (Figure 1b and Figure S2b,d) or using an inadequate setup for the LAI windows (Figure S5b) will often result in biased distributions of tract lengths, suggesting that these parameters should be carefully chosen when LAI is used to estimate admixture times. Although similar tract length distributions can be obtained under different admixture scenarios, our results suggest that a strongly bimodal distribution (with one peak around very short tracts and one around large tracts) is often observed when the assumed time since admixture in RFMix is much older than the true admixture or when one is using too few polymorphic sites per window in simpLAI. Besides, when admixture occurred recently (~10 generations ago), the track length mode should be larger than 10 Mb (see true, simulated data in Figure 1b, Figure S2b,d, and S5b). Observing smaller modes when the assumed time since admixture is ~10 generations or when using ~4000 polymorphic sites per window indicates that admixture is older. It can thus be beneficial to inspect the tract length distributions obtained for different parameters to rule out a poor choice of parameters for the two methods. Figure S2 shows that the distribution of true (simulated) tract lengths and inferred tract lengths under different prior admixture times only start to roughly converge for quite highly differentiated sources, corresponding to an F_{ST} ~ 0.3 (compare Figure S2b,d).

We next compared the performance of RFMix, MOSAIC, and simpLAI for different reference sizes. The results for RFMix and simpLAI were obtained using the true time since admixture as a prior and an optimal number of polymorphic sites per window, respectively (Figure 2). Within the tested range and conditions, the performance of RFMix is the most affected by reference size. While RFMix outperforms the other methods when four or more diploid individuals are used per reference, MOSAIC and simpLAI are better when only two diploid individuals are available. Contrarily to RFMix, simpLAI and MOSAIC can additionally be used with a single diploid individual per reference, albeit with a lower performance. Our results indicate that under such limited reference size, simpLAI outperforms MOSAIC. Yet,

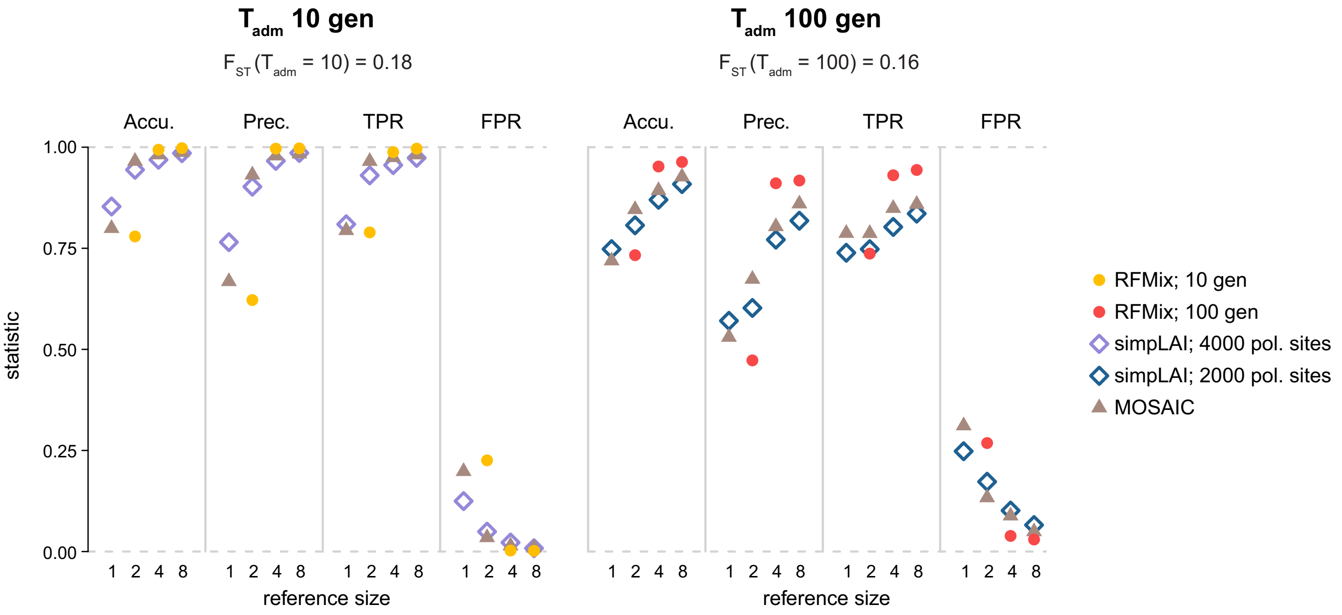

FIGURE 2 Influence of reference size in the performance of RFMix, MOSAIC, and simpLAI. Measures of performance for simulations of one pulse of admixture (30%), occurring 10 and 100 generations ago (T_{adm}) between two populations with effective size 5000 that diverged 1000 generations ago. The reference size corresponds to the number of diploid individuals. Accu., accuracy; Prec., precision; TPR, true-positive rate; FPR, false-positive rate; gen, generations; pol. sites, number of polymorphic sites.

we caution that LAI should only be applied using a single diploid per source if there is high differentiation between sources and admixture is very recent (e.g. F_{ST} ≥0.18 and T_{adm} = 10 gen) since the FPR becomes very high when admixture occurred 100 generations ago. Moreover, in demographic scenarios that are favourable to LAI, the FPR of RFMix becomes close to zero when as few as four diploid individuals are used per reference. This favourable outcome might be partially due to the joint inference of local ancestry for 10 (diploid) admixed individuals, since RFMix incorporates inferred ancestry assignments from the admixed panel to augment the training set information.

We also evaluated LAI performance under a more complex admixture scenario, where an already admixed population receives additional gene flow from the minority source (Figure 3, Figures S6 and S7). This scenario is commonly observed when a population undergoes spatial expansion, entering territories that are already occupied by another population. If admixture occurs as the population expands, it leads to a cumulative assimilation of the local ancestry. In this case, LAI for the recurrently admixed population could be done based on non-admixed ancestral reference populations (if available) or the more recent parental groups, even if already admixed. By comparing the performance of RFMix and simpLAI using different reference populations, each with four diploid individuals, we find that even though the proportion of tracts recovered from the older event is much higher when using unadmixed sources (S1–S2), the precision and the FPR are better when using the most recent ancestry contributors (S1–SA) (Figure 3). As expected, both LAI methods can identify most of the minority ancestry assimilated in the most recent event, but they recover much less from older events. Even though RFMix was designed to discover latent admixture in the reference panels and should therefore be able to infer the local ancestry of population SA, the very low proportion of “blue” ancestry (Figure 3) recovered from the older event (TPR1 for sources S1–SA) shows that this is not always the case, in particular when the initial admixture is old. MOSAIC performs similarly to simpLAI when using unadmixed sources (Figure S6a). Contrastingly, MOSAIC performs poorly when using an admixed population as a source: it recovered less than 10% of the true total “blue” ancestry from each admixture event (TPR1 and TPR2 in Figure S6a), and only 25% of the inferred “blue” ancestry is actually “blue” (low Precision in Figure S6a).

Reducing the sample size of the reference populations to a single diploid individual in MOSAIC and simpLAI leads to an overall lower performance (Figure S7a). Note that even though MOSAIC recovers most of the “blue” ancestry (high TPR1 and TPR2) under these conditions, this unexpected result is actually due to an erroneous assignment of most of the genome to the “blue” ancestry, as reflected by a very high FPR (~75%).

The distribution of tract lengths inferred by MOSAIC is largely shifted towards small values compared to the simulated data when a reference population is admixed (Figure S6b), further highlighting the inadequacy of already admixed sources for this method (although this is not always apparent; see Figure S7b). The true tract length distribution for the simulated data under this two-pulse model also shows that even when admixture events are very distant in time, we do not observe an obvious bimodality in the admixed tract length distribution (Figure 3b–c). If such a pattern emerges in the inferences (e.g. Figure 1b and Figure 3b–c), it is most likely the result of LAI errors and thus provides a way to recognize if LAI parameters are not well chosen. The discrepancy between the true (simulated) distributions and the distributions inferred by simpLAI and RFMix is more obvious for smaller than for larger tract lengths, with many short tracts not being detected. This pattern is expected since the minimum inferred tract length depends on the window-related parameters used in these methods. Even though this bias might affect the dating of admixture, the ancestry of short tracts is

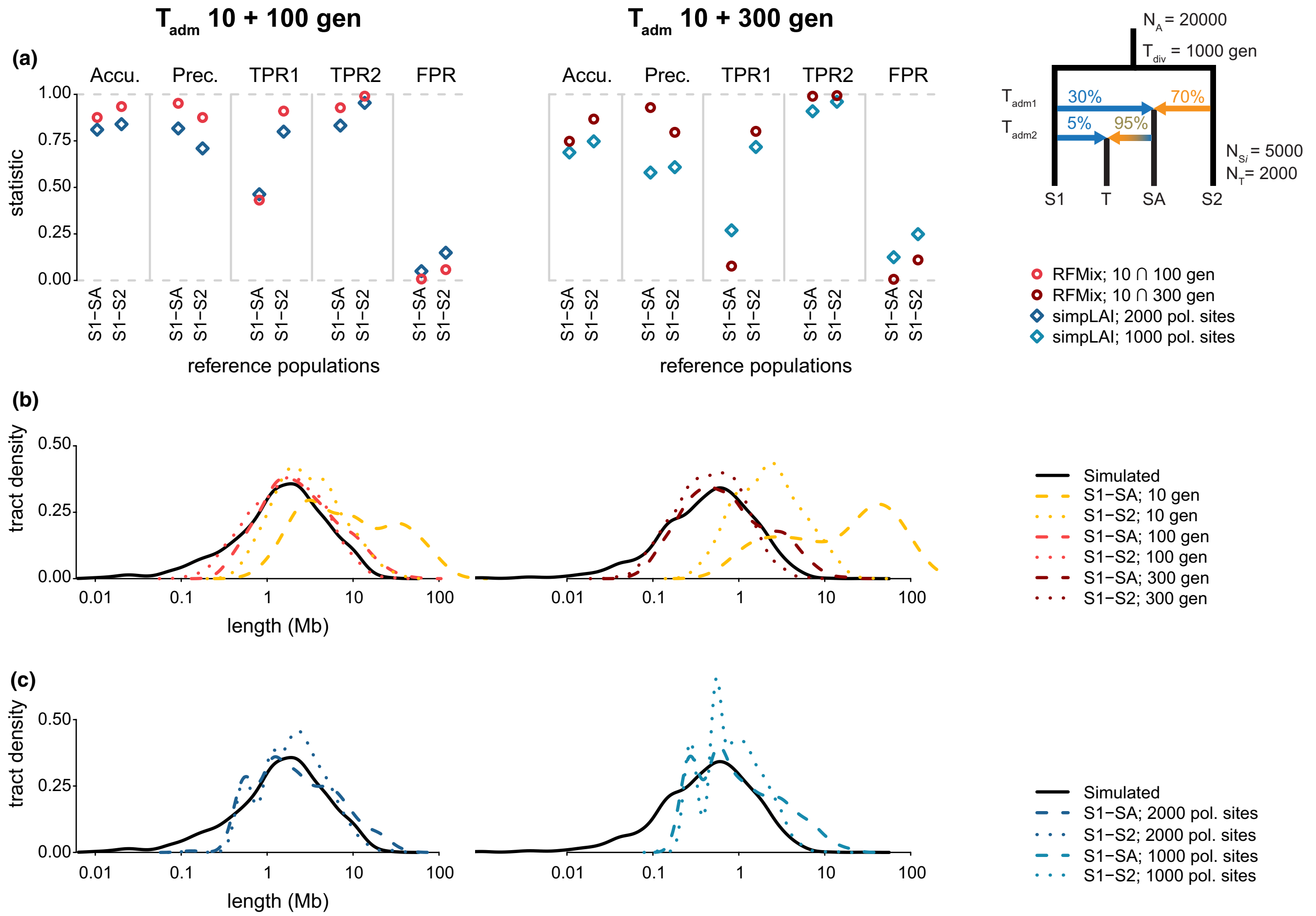

FIGURE 3 LAI performance using an admixed versus non-admixed reference for a model with two admixture pulses. The demographic model is displayed on the top right. (a) Measures of performance for RFMix and simpLAI, using four diploid individuals per reference population (S1 and SA vs. S1 and S2; Figure S1). (b, c) Densities of log-transformed tract lengths for the (true) simulated data, and for the local ancestry inferred by RFMix (b) and simpLAI (c) under different parameters. Accu., accuracy; Prec., precision; TPR1, true-positive rate for ancestry introduced in the oldest admixture pulse; TPR2, true-positive rate for ancestry introduced in the recent admixture pulse; FPR, false-positive rate; gen, generations; pol. sites, number of polymorphic sites.

often incorrectly inferred (Figure S8), and some tract pruning based on length could be beneficial for other LAI applications.

We observe that below a certain tract length, which depends on specific demographic conditions, the tract accuracy quickly drops (Figure S8). Tracts whose inferred ancestry is completely wrong (accuracy = 0) were detected in all three methods tested here. Among the tested scenarios, the lengths of wrongly assigned tracts tend to be smaller than 2 Mb when admixture occurred 10 generations ago and smaller than 4 Mb when at least part of the admixture occurred 100 generations ago. Therefore, we also calculated the overall LAI accuracy after masking tracts that are smaller than these thresholds (Figures S9, S10, and S11). When the total sum of these mis-identified short tracts constitutes a relatively small portion of a genome, as is the case for RFMix inferences using four individuals per reference, the overall accuracy does not change much (Figure S9). However, by excluding simpLAI-inferred tracts smaller than 2 Mb, we not only obtain higher accuracy (Figure S10) but more importantly higher precision, while FPR drops to values closer to zero (Figure 4). For the demographic conditions presented in Figure 4, the proportion of the genome that remains painted after excluding segments smaller than 2 Mb ranges from 87% (S1–SA) to 83% (S1–S2). The same exclusion criteria also result in a better performance when inferences are based on a single individual per reference in simpLAI (S1–SA or S1–S2) and MOSAIC (S1–S2), although the resulting FPR is still relatively high (~12%; Figure S12f).

We find that an additional way to improve LAI performance in simpLAI and RFMix (but not MOSAIC) consists of overlapping inferences obtained using admixed (S1–SA) versus unadmixed references (S1–S2), and masking regions of the genome for which the LAI is inconsistent (Figure 4 and Figure S12). While the FPR reduces to 6% when simpLAI is used with a single individual per reference, we caution that combining multiple masking options (including the generation overlap in RFMix) might result in a large amount of LAI data loss (Figure 4a, Figures S12a and S13a).

3.2 | Application to southern African admixed individuals

The genetic diversity of southern African populations has been impacted by two major migratory movements in the past 2000 years:

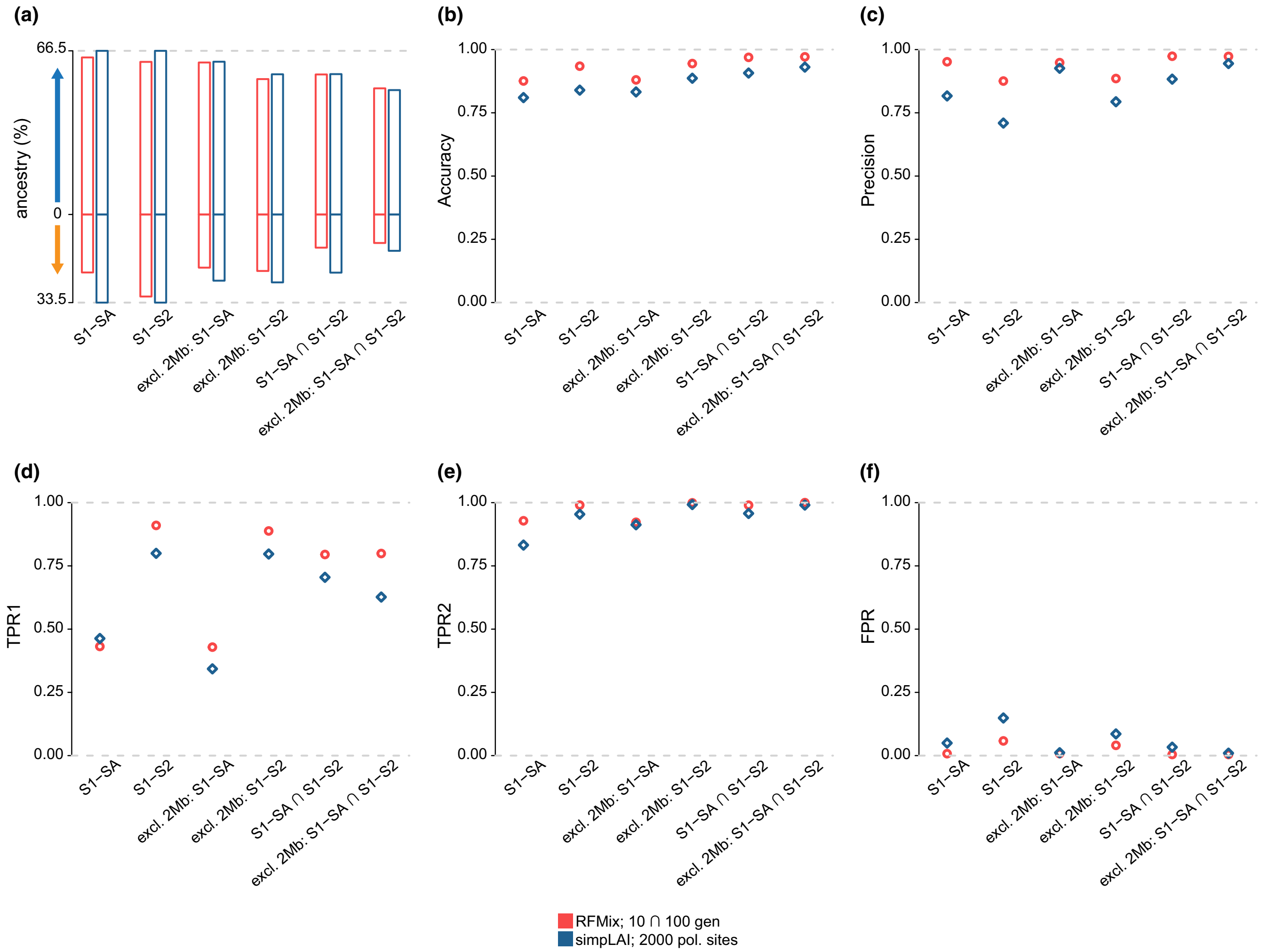

FIGURE 4 LAI performance after filtering ancestry tracts smaller than 2 Mb and after intersecting LAI results obtained with different reference populations of size 4. Results are shown for the two-pulse demographic model displayed in Figure 3, for $T_{\text{adm1}} = 100$ and $T_{\text{adm2}} = 10$, when an admixed population is included as reference (S1–SA) versus when only non-admixed populations are used (S1–S2). (a) Percentage of the painted genome from the minority (orange) and majority (blue) ancestry remaining after each processing step. (b–f) Measures of performance for RFMix and simpLAI. TPR1, true-positive rate for ancestry introduced in the oldest admixture pulse; TPR2, true-positive rate for ancestry introduced in the recent admixture pulse; FPR, false-positive rate; gen, generations; pol. sites, number of polymorphic sites.

one associated with the spread of pastoralism from Eastern Africa, and another associated with the expansion of Bantu-speaking farmers out of West Africa (Pickrell et al., 2012; Schlebusch et al., 2012). With the analysis of an ancient genome from southern Africa (Ballito Bay A, 1986–1831 BP), which lacks ancestry related to these recent movements, Schlebusch et al. (2017) showed that all of the so-called Khoisan-speaking groups carry some level of East African-related ancestry (9%–30%), and that most of them additionally carry West African-related ancestry. Given the lack of present-day unadmixed representatives of the original southern African ancestry, this ancestry component was previously estimated by using the least admixed southern African individuals as references (Oliveira et al., 2023). Yet, based on our simulations (Figure 3), we expect that some of the East and West African-related ancestries will not be detected with this strategy. By using the single unadmixed ancient genome from southern Africa as an alternative reference for the original southern African ancestry, local ancestry estimates of simpLAI broadly match the previous RFMix estimates based on more than 10 modern admixed individuals (average of 79% overlap per target individual, or 82% overlap based on the 93% of the genome that remains painted after excluding tracts <2 Mb; Figure 5 and Figure S14). A comparison between simpLAI results based on the two alternative reference sets shows a similar level of LAI overlap (average of 80% per target individual, or 81% overlap based on the 98% of the genome that remains painted after excluding tracts <2 Mb; Figure S14). The global ancestry estimates for each southern African population obtained using the ancient reference in simpLAI are also extremely correlated with those obtained using modern references in RFMix ($R^2 = .998$; p-value < .001; Figure S15 and Table S1) and simpLAI ($R^2 = .998$; p-value < .001; Figure S16 and Table S2), and using the frequency-based method *qpAdm* ($R^2 = .992$; p-value < .001) (Oliveira et al., 2023). For the selected LAI parameters (assuming 25 generations since admixture in RFMix and 1000 polymorphic sites per window in simpLAI), we detect higher amounts of West African ancestry

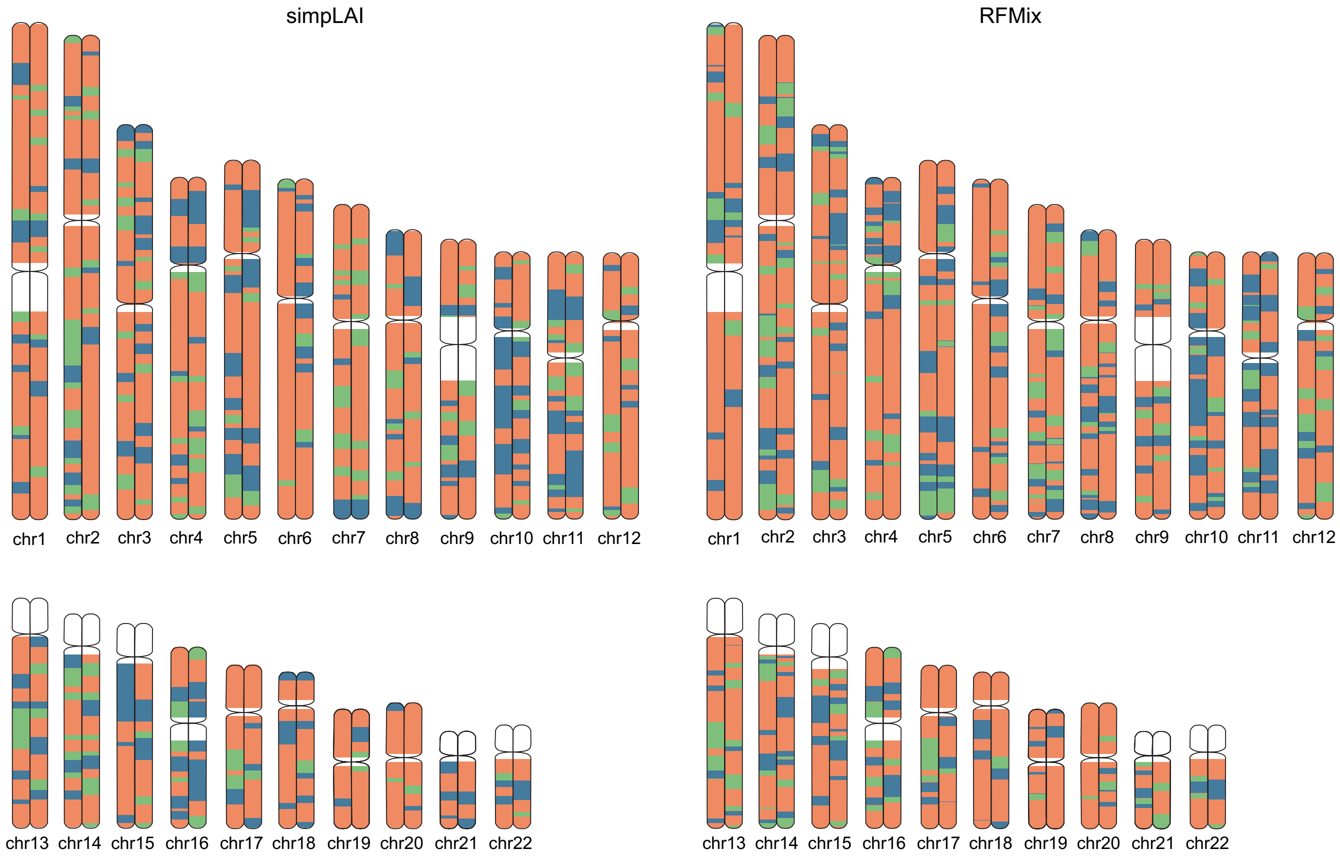

FIGURE 5 LAI results from simpLAI and RFMix for an individual (Kwepe) from southwestern Angola. The blue, green, and orange colours represent the southern, East, and West African-related ancestries, respectively. The white colour corresponds to regions without SNPs in our dataset. Discrepancies between inferences are highlighted in Figure S14.

recovered by using a non-admixed southern African source in simpLAI compared to an admixed source in RFMix (7.6% difference in intercept, *p*-value < .001 and 2.8% difference in slope, *p*-value = .011; Table S1) or simpLAI (5.2% difference in intercept, *p*-value < .001 and a non-significant difference in slope, *p*-value = .101; Table S2). Even though these differences are in line with our predictions that the use of individuals slightly admixed with West Africans as source for Southern African ancestry would tend to underestimate West African ancestry in present populations, we caution that a different outcome might be observed when using other LAI parameters. Regardless of this, our results confirm that simpLAI can be used in cases of limiting reference sizes when sources are relatively well differentiated.

3.3 | Application to Neolithic farmers from Europe

The Neolithic transition in Europe was marked by a demographic expansion of early farmers (EFs), whose origins can be traced back to Neolithic populations from the Aegean region (Hofmanová et al., 2016). Although this expansion was accompanied by relatively low levels of admixture with local Western Hunter-Gatherer (WHG) populations in central Europe, EFs from the Aegean region already have ~30% of their ancestry related to WHGs (Marchi et al., 2022). When using four of these already admixed Aegean EFs as a proxy for the EF source, together with a reference consisting of four European WHG, we found an average of 74% matching between the raw LAI obtained with simpLAI and RFMix for European EFs. After excluding tracts <2 Mb from both LAI results, 58% of the genome is masked but the matching increases to 95%, indicating that for a substantial part of the genome it is possible to obtain a good agreement between methods with relatively small reference panels. Moreover, simpLAI results obtained with the four admixed EF references match 59% of the simpLAI results obtained by using a single unadmixed ancient EF from Iran. After excluding tracts <2 Mb, the matching increases to 72%, but only 14% of the genome remains painted. The 25% increase (on average) in global WHG ancestry inferred when using the unadmixed EF source rather than the admixed source by simpLAI (Figure S17) probably reflects a higher recovery of HG tracts from early admixture pulses occurring in the Aegean ancestors.

4 | DISCUSSION

The demographic models and parametric conditions tested here are informative on the limits of LAI and can aid researchers to understand if LAI can be applied to their data and to which extent results can be trusted. While previous studies suggest that most LAI tools perform relatively well when admixture between highly differentiated populations occurred recently (Geza et al., 2019), the exact amount of divergence required between sources (Split time/Ne) to reach a certain level of accuracy, precision or other performance indicator was not clearly outlined. This lack of information was even more critical for small reference panels. Moreover, the evaluation of LAI performance was often based on the construction of admixed genomes using data from present-day human groups (e.g. Hilmarsson et al., 2021; Maples et al., 2013; Paşaniuc et al., 2009; Schubert et al., 2020; Uren et al., 2020). However, since the demographic history of many of them is not fully understood and is often complex, the performance results could hardly be extrapolated to other populations or species.

By simulating a relatively wide range of demographic scenarios, we show that if the source populations are highly differentiated (F_{ST} at time of admixture ≥0.3) and each reference sample is beyond a critical size (>4 diploids), RFMix inferences have a high precision and low FPR, even when admixture is old (~300 generations; Figure S2a). On the other hand, if differentiation between sources is low (F_{ST} at time of admixture = 0.05), a good performance can still be expected when admixture is very recent (~10 generations)—a condition that can be diagnosed through inspection of tract length distributions. Indeed, recent admixture is characterized by a unimodal distribution of inferred tract lengths with a large mode (>10 Mb) when assuming recent times since admixture and this distribution is contrastingly strongly bimodal when assuming older times (100–300 generations). A reduction in source differentiation from F_{ST} ~ 0.3 to F_{ST} ~ 0.15 when admixture is old leads to a considerable decrease in the performance of RFMix (Figure 1a and Figure S2a). For such intermediate levels of differentiation (i.e. F_{ST} ~ 0.15), RFMix outperforms MOSAIC and simpLAI (Figure 2) if four diploid individuals or more are used per reference population. However, simpLAI performs better than RFMix and MOSAIC when only one diploid individual is available as reference for each source population (Figure 2), which might occur in non-model organisms or ancient DNA samples.

In case of multiple sequential admixture events, our ability to recover tracts from the oldest events is severely reduced (Figure 3), but the inclusion of a sample (or multiple ones when available) with an age that is closer to the admixture event should lead to better inferences. Under such scenarios of complex admixture, if some level of admixture in reference populations cannot be excluded, MOSAIC performs poorly (Figures S6a and S7a). This suggests that the relationship between panels and ancestral populations is not adequately inferred when sources are admixed, probably because the copying matrix relating reference panels to ancestries is difficult to estimate, and therefore the use of MOSAIC in this case is not recommended.

Inferences from simpLAI can be greatly improved by the use of post-processing measures, such as masking small LAI tracts or tracts whose ancestry assignment differs depending on the choice of LAI parameters or reference populations. The improvement achieved by repeating LAI for different but closely related populations that could represent one specific source (Figure 4) is ideal for applications for which having a small FPR is more important than having LAI for the whole genome (e.g. studying introgression in highly conserved vs. non-conserved regions). Likewise, the positive relationship observed between tract length and tract accuracy (Figure S8) suggests that studying the properties of long LAI tracts could still be valuable when demographic conditions are not ideal for LAI. Yet, it is important to consider that by restricting any subsequent analysis of the consequences of introgression to large LAI tracts, we will be limiting ourselves to studying the most recent admixture events or other evolutionary processes, like selection, that could maintain long tracts for many generations. We also caution that the same filtering steps might not lead to similar improvements under very different demographic scenarios or different methods. For example, the post-processing measures tested in this work do not seem to always improve RFMix results as much (see Figure 4 and Figure S9). However, in this case we strongly recommend users to compute the intersection of RFMix results under different assumed admixture times (Figure S13), given the high reduction in FPR, unless it is evident from historical sources that admixture is very recent.

Many studies used LAI without considering the effect of the assumed time since admixture in their applications. Corbett-Detig and Nielsen (2017) reported a bias in LAI due to uncertainty in admixture times with RFMix, while Uren et al. (2020) reported no significant differences in results within the narrow range of 10–20 generations since admixture using the same method. Our work underlines the importance of carefully selecting this time and other window-related parameters in LAI tools, as these parameters drastically impact inferences. The sharp differences between tract length distributions for different assumed times and window lengths put into question the use of some state-of-the-art LAI tools for dating admixture. For these purposes, LAI methods that directly estimate relevant parameters from the data, such as MOSAIC or Ancestry HMM, should be preferred.

Our simulations suggest that reasonable LAI can be obtained by simpLAI based on one or two diploid genomes under demographic conditions that are not uncommon among past human groups and other organisms (Figure 2). This result is confirmed empirically by finding a high level of matching between the LAI of southern African individuals based on a single southern African ancient genome using simpLAI and based on a larger reference panel using RFMix. The lower amount of matching among European Neolithic farmers when using a single unadmixed individual versus several admixed individuals as a proxy for the EF ancestry could be due to the more complex admixture history of these populations and to a higher recovery of HG ancestry from early admixture events. However, further studies using additional ancient individuals as sources would be helpful to confirm the validity of the inferred admixed tracts. Our study nevertheless suggests that simpLAI should be useful for studying the consequences of introgression assessed from a few ancient DNA samples or from genomes of non-model organisms, for which high-coverage data, accurate genotype calling, and chromosome phasing are still challenging to obtain.

AUTHOR CONTRIBUTIONS

SO, NM, and LE designed the study. SO performed the research and analysed the data. LE contributed new analytical tools. SO, NM, and LE wrote the manuscript.

ACKNOWLEDGEMENTS

This work was supported by a Swiss NSF grant (No 310030_188883) to LE. We thank Carina Schlebusch for sharing the genotype calls for the ancient Ballito Bay A, Stephan Peischl for statistical advice, and the CMPG members for their comments during lab meetings. Open access funding provided by Universitat Bern.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The software developed in this study and instructions to generate the simulated data are available from: https://github.com/CMPG/simpLAI. The genotype data for present-day southern Africans and ancient western Eurasians have been previously deposited in the European Genome-Phenome Archive and European Variant Archive, under the accession numbers EGAS00001007011 and PRJEB51919, respectively. The genotype data for the ancient Ballito Bay A were provided by Carina Schlebusch. Raw data for this individual are available from the European Nucleotide Archive under accession number PRJEB22660.

ORCID

Sandra Oliveira https://orcid.org/0000-0002-1133-7130
Nina Marchi https://orcid.org/0000-0001-6624-5922
Laurent Excoffier https://orcid.org/0000-0002-7507-6494

REFERENCES

Birchler, J. A., Yao, H., & Chudalayandi, S. (2006). Unraveling the genetic basis of hybrid vigor. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(35), 12957–12958. https://doi.org/10.1073/pnas.0605627103

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Browning, S. R., Browning, B. L., Daviglus, M. L., Durazo-Arvizu, R. A., Schneiderman, N., Kaplan, R. C., & Laurie, C. C. (2018). Ancestry-specific recent effective population size in the Americas. *PLoS Genetics*, *14*(5), e1007385. https://doi.org/10.1371/journal.pgen.1007385

Browning, S. R., Waples, R. K., & Browning, B. L. (2023). Fast, accurate local ancestry inference with FLARE. *American Journal of Human Genetics*, *110*(2), 326–335. https://doi.org/10.1016/j.ajhg.2022.12.010

Chimusa, E. R., Defo, J., Thami, P. K., Awany, D., Mulisa, D. D., Allali, I., Ghazal, H., Moussa, A., & Mazandu, G. K. (2018). Dating admixture events is unsolved problem in multi-way admixed populations. *Briefings in Bioinformatics*, *21*(1), 144–155. https://doi.org/10.1093/bib/bby112

Corbett-Detig, R., & Nielsen, R. (2017). A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. *PLoS Genetics*, *13*(1), e1006529. https://doi.org/10.1371/journal.pgen.1006529

Edelman, N. B., & Mallet, J. (2021). Prevalence and Adaptive Impact of Introgression. https://doi.org/10.1146/annurev-genet-021821

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics*, *37*(24), 4882–4885. https://doi.org/10.1093/bioinformatics/btab468

Geza, E., Mugo, J., Mulder, N. J., Wonkam, A., Chimusa, E. R., & Mazandu, G. K. (2019). A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. In *Briefings in bioinformatics* (Vol. *20*, Issue 5, pp. 1709–1724). Oxford University Press. https://doi.org/10.1093/bib/bby044

Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. *Genetics*, *196*(3), 625–642. https://doi.org/10.1534/genetics.113.160697

Hilmarsson, H., Kumar, A. S., Rastogi, R., Bustamante, C. D., Montserrat, M., & Ioannidis, A. G. (2021). High resolution ancestry deconvolution for next generation genomic data. BioRxiv. https://doi.org/10.1101/2021.09.19.460980

Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-Del-Molino, D., Van Dorp, L., López, S., Kousathanas, A., Link, V., Kirsanow, K., Cassidy, L. M., Martiniano, R., Strobel, M., Scheu, A., Kotsakis, K., Halstead, P., Triantaphyllou, S., Kyparissi-Apostolika, N., … Burger, J. (2016). Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(25), 6886–6891. https://doi.org/10.1073/PNAS.1523951113

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2), 583–589. https://doi.org/10.1093/genetics/132.2.583

Ioannidis, A. G., Blanco-Portillo, J., Sandoval, K., Hagelberg, E., Miquel-Poblete, J. F., Moreno-Mayar, J. V., Rodríguez-Rodríguez, J. E., Quinto-Cortés, C. D., Auckland, K., Parks, T., Robson, K., Hill, A. V. S., Avila-Arcos, M. C., Sockell, A., Homburger, J. R., Wojcik, G. L., Barnes, K. C., Herrera, L., Berríos, S., … Moreno-Estrada, A. (2020). Native American gene flow into Polynesia predating Easter Island settlement. *Nature*, *583*(7817), 572–577. https://doi.org/10.1038/s41586-020-2487-2

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*(7518), 409–413. https://doi.org/10.1038/nature13673

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*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Marchi, N., Winkelbach, L., Schulz, I., Wegmann, D., Burger, J., Correspondence, L. E., Brami, M., Hofmanová, Z., Blö, J., Reyna-Blanco, C. S., Diekmann, Y., Thié, A., Kapopoulou, A., Link, V., Rie Piuz, V., Kreutzer, S., Figarska, S. M., Ganiatsou, E., Pukaj, A., … Excoffier, L. (2022). The genomic origins of the world's first farmers. *Cell*, *185*, 1842–1859. https://doi.org/10.1016/j.cell.2022.04.008

Martin, S. H., & Jiggins, C. D. (2017). Interpreting the genomic landscape of introgression. In *Current opinion in genetics and development* (Vol. *47*, pp. 69–74). Elsevier Ltd. . https://doi.org/10.1016/j.gde.2017.08.007

Moran, B. M., Payne, C., Langdon, Q., Powell, D. L., Brandvain, Y., & Schumer, M. (2021). The genomic consequences of hybridization. In *eLife* (Vol. *10*). eLife Sciences Publications Ltd, 10:e69016. https://doi.org/10.7554/ELIFE.69016

Oliveira, S., Fehn, A.-M., Amorim, B., Stoneking, M., & Rocha, J. (2023). Genome-wide variation in the Angolan Namib Desert reveals unique pre-bantu ancestry. *Science Advances*, *9*(38), eadh3822. https://doi.org/10.1126/SCIADV.ADH3822

Paşaniuc, B., Sankararaman, S., Kimmel, G., & Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, *25*(12), i213–i221. https://doi.org/10.1093/bioinformatics/btp197

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Pearson, A., & Durbin, R. (2023). Local ancestry inference for complex population histories. BioRxiv. https://doi.org/10.1101/2023.03.06.529121

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S. W., Nakagawa, H., Naumann, C., Lipson, M., Loh, P. R., Lachance, J., Mountain, J., Bustamante, C. D., Berger, B., Tishkoff, S. A., Henn, B. M., Stoneking, M., … Pakendorf, B. (2012). The genetic prehistory of southern Africa. *Nature Communications*, *3:1143*. https://doi.org/10.1038/ncomms2140

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., & Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, *5*(6), e1000519. https://doi.org/10.1371/journal.pgen.1000519

Rishishwar, L., Conley, A. B., Wigington, C. H., Wang, L., Valderrama-Aguirre, A., & King Jordan, I. (2015). Ancestry, admixture and fitness in Colombian genomes. *Scientific Reports*, *5:12376*. https://doi.org/10.1038/srep12376

Salter-Townshend, M., & Myers, S. (2019). Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics*, *212*(3), 869–889. https://doi.org/10.1534/genetics.119.302139

Sankararaman, S., Sridhar, S., Kimmel, G., & Halperin, E. (2008). Estimating local ancestry in admixed populations. *American Journal of Human Genetics*, *82*(2), 290–303. https://doi.org/10.1016/j.ajhg.2007.09.022

Scally, A., & Durbin, R. (2012). Revising the human mutation rate: Implications for understand ing human evolution. *Nature Reviews Genetics*, *13*(10), 745–753. https://doi.org/10.1038/nrg3295

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H., Lombard, M., & Jakobsson, M. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, *358*(6363), 652–655. www.sahumanities.org/ojs/

Schlebusch, C. M., Skoglund, P., Sjodin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G. B., Soodyall, H., & Jakobsson, M. (2012). Genomic variation in seven Khoe-san groups reveals adaptation and complex African history. *Science*, *338*(6105), 374–379. https://doi.org/10.1126/science.1227721

Schubert, R., Andaleon, A., & Wheeler, H. E. (2020). Comparing local ancestry inference models in populations of two- and three-way admixture. *PeerJ*, *8*, e10090. https://doi.org/10.7717/peerj.10090

Slatkin, M., & Voelm, L. (1990). FST in a hierarchical Island model. *Genetics*, *127*(3), 627–629.

Sundquist, A., Fratkin, E., Do, C. B., & Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, *18*(4), 676–682. https://doi.org/10.1101/gr.072850.107

Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. In. *American Journal of Human Genetics*, *79*). www.ajhg.org, 1–12.

Uren, C., Hoal, E. G., & Möller, M. (2020). Putting RFMix and ADMIXTURE to the test in a complex admixed population. *BMC Genetics*, *21*(1), 40. https://doi.org/10.1186/s12863-020-00845-3

Wang, Y., Song, S., Schraiber, J. G., Sedghifar, A., Byrnes, J. K., Turissini, D. A., Hong, E. L., Ball, C. A., & Noto, K. (2021). Ancestry inference using reference labeled clusters of haplotypes. *BMC Bioinformatics*, *22*(1), 459. https://doi.org/10.1186/s12859-021-04350-x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Oliveira, S., Marchi, N., & Excoffier, L. (2024). Assessing the limits of local ancestry inference from small reference panels. *Molecular Ecology Resources*, *00*, e13981. https://doi.org/10.1111/1755-0998.13981