Loter: A Software Package to Infer Local Ancestry for a Wide Range of Species

Thomas Dias-Alves,1 Julien Mairal,2 and Michael G.B. Blum*1
1CNRS, TIMC-IMAG UMR 5525, Univ. Grenoble Alpes, Grenoble, France
2CNRS, Institute of Engineering Univ. Grenoble Alpes, LJK, Univ. Grenoble Alpes, Inria, Grenoble, France
*Corresponding author: E-mail: michael.blum@univ-grenoble-alpes.fr.
Associate editor: Rasmus Nielsen

Abstract
Admixture between populations provides opportunity to study biological adaptation and phenotypic variation. Admixture studies rely on local ancestry inference for admixed individuals, which consists of computing at each locus the number of copies that originate from ancestral source populations. Existing software packages for local ancestry inference are tuned to provide accurate results on human data and recent admixture events. Here, we introduce Loter, an open-source software package that does not require any biological parameter besides haplotype data in order to make local ancestry inference available for a wide range of species. Using simulations, we compare the performance of Loter to HAPMIX, LAMP-LD, and RFMix. HAPMIX is the only software severely impacted by imperfect haplotype reconstruction. Loter is the less impacted software by increasing admixture time when considering simulated and admixed human genotypes. For simulations of admixed Populus genotypes, Loter and LAMP-LD are robust to increasing admixture times by contrast to RFMix. When comparing length of reconstructed and true ancestry tracts, Loter and LAMP-LD provide results whose accuracy is again more robust than RFMix to increasing admixture times. We apply Loter to individuals resulting from admixture between Populus trichocarpa and Populus balsamifera and lengths of ancestry tracts indicate that admixture took place ~100 generations ago. We expect that providing a rapid and parameter-free software for local ancestry inference will make more accessible genomic studies about admixture processes.

Key words: admixture, optimization, local ancestry, population genetics.

Introduction
Admixture or hybridization between populations is a natural phenomenon that provides opportunity to map genomic regions involved in phenotypic variation and biological adaptation (Buerkle and Lexer 2008; Paysseur and Rieseberg 2016). Mapping can rely on Local Ancestry Inference (LAI) of admixed individuals, which consists of computing at a given locus the number of copies that originates from the ancestral source populations. LAI uses haplotypes from putative source populations and processes haplotypes or genotypes from admixed population to infer local ancestry of admixed individuals. Figure 1 shows local ancestry of four simulated Populus individuals resulting from admixture between two Populus species (Suarez-Gonzalez et al. 2016). Sequence and dense genotype data are now generated for a wide range of species besides humans for which LAI is relevant. LAI can be used to study patterns of introgression (Hufford et al. 2013; Suarez-Gonzalez et al. 2016; Medugorac et al. 2017), to map genes involved in reproductive isolation (Corbett-Detig and Nielsen 2017) and phenotypic variation (Lindtke et al. 2013; vonHoldt et al. 2016), and to decipher past admixture processes (Brandvain et al. 2014; Liu et al. 2014).

Although various LAI software are now available, they have been mainly tuned to human data in order to map disease-associated variants (Patterson et al. 2004; Seldin et al. 2011).

In this paper, we introduce the software package Loter for Local Ancestry Inference, which does not require specifications of statistical or biological parameters in order to make LAI available for a wide range of species. Several software packages for LAI have been developed including HAPMIX, LAMP-LD, and RFMix (Price et al. 2009; Baran et al. 2012; Maples et al. 2013). However, they require various parameters to be specified, which can hamper their practical use. HAPMIX requires specifications of several biological parameters that might be difficult to obtain such as genetic maps, recombination and mutation rate, average ancestry coefficients, and the average number of generations since admixture (Price et al. 2009). LAMP-LD requires a physical map and statistical parameters such as the number of hidden states in the hidden Markov Model and a window size where local ancestry is assumed to be constant (Baran et al. 2012). Default parameter values can be provided when using LAMP-LD. RFMix requires statistical and biological parameters, which are a genetic map, a window size (in centimorgan) where local ancestry is assumed to be constant, and the average number of generations since admixture.
Loter: A Software Package to Infer Local Ancestry

(Price et al. 2009). For an admixed individual, local ancestry at a given locus corresponds to the number of copies that has been inherited from the species Populus trichocarpa (Suarez-Gonzalez et al. 2016). For an admixed individual, local ancestry at a given locus corresponds to the number of copies that has been inherited from the species Populus trichocarpa. LAI software require haplotypes from putative source populations and process haplotypes or genotypes from admixed population to return local ancestry of admixed individuals. Details of the simulations are described in the Materials and Methods section.

We compare Loter to HAPMIX, LAMP-LD, and RFMIX using diploid accuracy, which is an error measure analogous to imputation error for LAI (Sankararaman, Sridhar, et al. 2008). We consider an admixture experiment using human genotypes from HAPMAP 3 where we simulate genotypes resulting from admixture between Europeans (CEU) and Africans (YRI) (International HapMap 3 Consortium 2010). We evaluate to what extent diploid accuracy is affected by the number of generations since admixture for the different approaches. We additionally consider a three-way admixture scenario between Chinese (CHB), Europeans (CEU), and Africans (YRI) from HAPMAP 3. We repeat admixture simulations using genotypes from two Populus species in North America (fig. 1) (Suarez-Gonzalez et al. 2016). We additionally evaluate to what extent length of ancestry tracts are accurately inferred by the different LAI software packages. Length of ancestry tracts is a biological information that is used to date and reconstruct admixture events and that should consequently be accurately inferred for reliable demographic

![Figure 1](https://academic.oup.com/mbe/article-abstract/35/9/2318/5040668)

**Figure 1.** Example of local ancestry inference for four simulated Populus individuals resulting from admixture between two Populus species, which are *Populus trichocarpa* and *Populus balsamifera* (Suarez-Gonzalez et al. 2016). For an admixed individual, local ancestry at a given locus corresponds to the number of copies that has been inherited from the species *P. trichocarpa*. LAI software require haplotypes from putative source populations and process haplotypes or genotypes from admixed population to return local ancestry of admixed individuals. Details of the simulations are described in the Materials and Methods section.

![Figure 2](https://academic.oup.com/mbe/article-abstract/35/9/2318/5040668)

**Figure 2.** Graphical description of Local Ancestry Inference as implemented in the software Loter. Given a collection of parental haplotypes from the source populations depicted in blue and red, Loter assumes that an haplotype of an admixed individuals is modeled as a mosaic of existing parental haplotypes. In this example, the first term of equation (1) (loss function) is equal to 1 because of a single mismatch between parental and admixed haplotype located at the next-to-last position, and the second term of equation (1) (regularization term) is equal to 2 because there are two switches between parental haplotypes. The displayed solution corresponds to the mathematical solution $(s_1, \ldots, s_{12}) = (5, 5, 5, 1, 1, 1, 2, 2, 2)$ where haplotypes are numbered from top to bottom, and $s_j = k$ if the admixed haplotype results from a copy of the $k$th parental haplotype at the $j$th SNP.

### Table 1. Differences between Several LAI Software.

<table>
<thead>
<tr>
<th></th>
<th>HAPMIX</th>
<th>LAMP-LD</th>
<th>RFMix</th>
<th>Loter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phasing of admixed individuals</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Number of ancestral pop.</td>
<td>2</td>
<td>2, 3, 5</td>
<td>≥ 2</td>
<td>≥ 2</td>
</tr>
<tr>
<td>Genetic map required (in cM)</td>
<td>Yes</td>
<td>No (physical position)</td>
<td>Yes</td>
<td>No (ordered SNPs)</td>
</tr>
<tr>
<td>Limitation of the number of SNPS</td>
<td>No</td>
<td>50,000</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Phasing error correction</td>
<td>No</td>
<td>Not required</td>
<td>Yes</td>
<td>Yes for two ancestral pop.</td>
</tr>
<tr>
<td>Parallel implementation</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Admixture time required</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Other biological param. required</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

*Note:*—The abbreviation param. stands for parameter and pop. for population. Other biological parameters required by HAPMIX are recombination and mutation rates.
reconstruction (Gravel 2012; Ni et al. 2016; Corbett-Detig and Nielsen 2017; Xue et al. 2017). Finally, we apply Loter to admixed Populus genotypes, and we estimate admixture time using length of reconstructed ancestry tracts.

**New Approaches**

We describe the optimization problem, which accounts that haplotypes from admixed individuals are described as a mosaic of haplotypes originating from the source populations (fig. 2). We assume that there are \( n \) individuals in the source populations resulting in \( 2n \) haplotypes denoted by \( (H_1, \ldots, H_{2n}) \). The value (0 or 1) of the \( i \)th haplotype at the \( j \)th SNP is denoted by \( H^j_i \). Haplotypes can be obtained from genotypes using computational phasing software such as fastPHASE or Beagle (Scheet and Stephens 2006; Browning and Browning 2007). A vector \( (s_1, \ldots, s_p) \) describes the sequence of haplotype labels from which the haplotype \( h \) of an admixed individual can be approximated (fig. 2). For the \( j \)th SNP in the data set, \( s_j = k \) if haplotype \( h \) results from a copy of haplotype \( H_k \). The optimization problem consists of minimizing the following cost function:

\[
C(s_1, \ldots, s_p) = \sum_{j=1}^{p} |H^j - H^j_k| + \lambda \sum_{j=1}^{p-1} 1_{s_j \neq s_{j+1}}, \tag{1}
\]

where \( (s_1, \ldots, s_p) \in \{1, \ldots, 2n\}^p \). The first term in equation (1) is a loss function that sums over all possible loci of a \{0, 1\}-valued function equals to 1 if haplotype \( h \) is different from the copied haplotype and to 0 otherwise. The second term is a regularization function that is equal to the regularization parameter \( \lambda \) times the number of switches between parental haplotypes. The optimization problem of equation (1) can be visualized using the graph of figure 3. A solution to minimize equation (1) can be found using dynamic programming and is provided in the Materials and Methods section.

Once a solution has been provided about the sequence \( (s_1, \ldots, s_p) \) of parental haplotypes, local ancestry values can be deduced automatically from this sequence because each parental haplotype belongs to one of the source populations (fig. 2). The formulation described in equation (1) is valid for \( K = 2 \) or more ancestral source populations.

The optimization problem described in equation (1) involves a regularization parameter \( \lambda \). Large values of \( \lambda \) strongly penalize switches between parental haplotypes such that solutions have long stretches of constant local ancestry. To avoid the difficult choice of \( \lambda \), solutions for local ancestry are averaged by running the optimization method several times on different values of \( \lambda \). Besides, to improve the stability of the solution, for each value of \( \lambda \), we consider a bagging technique where 20 different solutions are found based on 20 different data sets generated using bootstrap resampling (Breiman 1996). We consider \( \lambda = 1 + 0.5i, i = 1, \ldots, 8 \), resulting in \( 160 = 8 \times 20 \) different solutions and the final choice for local ancestry is obtained using a majority rule. When the most frequent vote has less than 75% of the votes, ancestry is imputed using local ancestry values of the closest SNPs with a preference for the SNP on the left in case of ambiguity. Finally, an additional smoothing procedure is considered in order to account for switch errors when phasing admixed individuals (see Materials and Methods).

**Results**

**Simulated Human Admixed Individuals**

We consider simulated admixed individuals resulting from admixture between Africans (YRI population in HAPMAP 3) and Europeans (CEU population in HAPMAP 3). Accuracy obtained with several LAI software packages varies depending on time since admixture occurs (fig. 4). For recent admixture where admixture occurred five generations ago, LAMP-LD and RFMix obtain the best result with median diploid accuracies of 99.6% and 99.8% whereas median diploid accuracy of Loter is equal to 99.3%. For ancient admixture where admixture occurred 500 generations ago, Loter obtains the largest median diploid accuracy of 86.7% followed by LAMP-LD with a diploid accuracy of 80.6% and RFMix with a diploid accuracy of 72.0%. For the smallest admixture times (\( \mu \leq 20 \) generations), diploid accuracy obtained with the three approaches is >95% and RFMix and LAMP-LD outperform LOTER. By contrast, for the largest admixture times (\( \mu \geq 150 \) generations), Loter is the most accurate software for LAI (fig. 4).

To evaluate these LAI software packages in a different context, we consider an additional setting where haplotypes of the reference panel are not drawn from the source populations but from populations that have diverged from the

---

**Fig. 3.** Graph that represents the optimization problem of equation (1). An optimal solution for \((s_1, \ldots, s_p)\) is found by finding the shortest path from node \(a\) to node \(b\). We assume that there are \( n \) individuals in the source populations resulting in \( 2n \) haplotypes denoted by \((H_1, \ldots, H_{2n})\). The value (0 or 1) of the \( i \)th haplotype at the \( j \)th SNP is denoted by \( H^j_i \). A vector \((s_1, \ldots, s_p)\) describes the sequence of haplotype labels from which the haplotype \( h \) of an admixed individual can be approximated. For the \( j \)th SNP in the data set, \( s_j = k \) if haplotype \( h \) results from a copy of haplotype \( H_k \).
The distribution of the 20 values for the mean diploid accuracy. For each value of the number of generations since admixture, 20 sets of 48 admixed individuals are generated. Boxplots show the diploid accuracy obtained with LAMP-LD, Loter, and RFMix for simulated admixed human individuals as a function of the number of generations since admixture. Admixed individuals are simulated by constructing haplotypes using Beagle software and then applying true haplotype phases. Local ancestry inference is performed using trio information. For local ancestry inference, haplotypes are obtained using Beagle individuals that are not used for simulating admixed individuals. For each value of the number of generations since admixture, 20 sets of 48 admixed individuals are generated. Boxplots show the distribution of the 20 values for the mean diploid accuracy.

**Fig. 4.** Diploid accuracy obtained with LAMP-LD, Loter, and RFMix for simulated admixed human individuals as a function of the time since admixture occurred. Admixed individuals are simulated by constructing their genomes from a mosaic of true African (YRI) and European (CEU) haplotypes (International HapMap 3 Consortium 2010). For performing simulations, true haplotypes are obtained using trio information. For local ancestry inference, haplotypes are obtained with Beagle using individuals that are not used for simulating admixed individuals. For each value of the number of generations since admixture, 20 sets of 48 admixed individuals are generated. Boxplots show the distribution of the 20 values for the mean diploid accuracy.

For individuals resulting from admixture between Yoruba (YRI) and Northern-European ancestry (CEU), reference panels are the Luhya sample from Kenya (LWK) and the Toscani sample from Italy (TSI). Compared with Loter, we again find that RFMix and LAMP-LD provide more accurate reconstructions of local ancestry for the recent admixture times of $\mu = 5, 10, 20$ generations but less accurate reconstructions for more ancient admixture events (supplementary fig. S1, Supplementary Material online). Loter is more accurate than RFMix when $\mu \geq 50$ generations; for example, Loter accuracy is of $92\%$ for $\mu = 100$ generations whereas RFMix accuracy is equal to $87\%$. The loss of accuracy of LAMP-LD for ancient admixture times is less pronounced but Loter still provides more accurate reconstructions than LAMP-LD when $\mu \geq 100$ generations.

We evaluate the benefit of bagging and of averaging local ancestry values obtained with different values of the regularization parameter, which are implemented by default in Loter. First, diploid accuracy depends on the choice of the regularization parameter (supplementary fig. S2, Supplementary Material online). As expected, smallest values of $\lambda$ provide the best result for ancient admixture event; $\lambda = 2$ is optimal when $\mu = 400$ and $\mu = 500$ generations, and $\lambda = 5$ is optimal otherwise. Second, for all values of the admixture times, averaging local ancestry values obtained with different values of the regularization parameter $\lambda$ instead of considering a single value improves inference (supplementary fig. S2, Supplementary Material online). Last, averaging over bootstrap replicates (bagging) further improves local ancestry inference (supplementary fig. S2, Supplementary Material online). Additionally, we vary the number of haplotypes in the reference panels to investigate to what extent Loter is sensitive to the reference panels. When using the TSI and LWK reference samples, the diploid accuracy varies substantially from 41% with a single haplotype in each population to 98% when using 100 haplotypes in each population.

We additionally evaluate the diploid accuracy of HAPMIX, which is another LAI software package based on the copying model. Compared with the three other LAI software packages, its diploid accuracy is the smallest with values of diploid accuracy ranging from 42% to 57% (supplementary fig. S3, Supplementary Material online). The fact that haplotypes used for LAI are not exact but can be computationally phased using Beagle causes the severe reduction of diploid accuracy obtained with HAPMIX (supplementary fig. S4, Supplementary Material online). When considering haplotypes reconstructed with Beagle instead of true haplotypes, diploid accuracy is reduced by 32% on an average.

Additionally, we compare diploid accuracy of Loter, RFMix, and LAMP-LD on data simulated under a three-way admixture model (supplementary fig. S5, Supplementary Material online). As for two-way admixture model, we find that RFMix and LAMP-LD have larger diploid accuracies than Loter for recent admixture and smaller ones for ancient admixture. When admixture took place five generations ago, the diploid accuracy of RFMix is of 99.8%, the accuracy of LAMP-LD is of 99.8%, and the accuracy of Loter is of 99.5%. By contrast, when admixture took place 500 generations ago, the diploid accuracy of RFMix is of 84.6%, the accuracy of LAMP-LD is of 89.1%, and the accuracy of Loter is of 92.3%.

**Accounting for Phasing Errors**

Loter requires phased haplotypic data for both the target as well as the reference panels. Usually haplotype phases are found using computational phasing programs such as Beagle or shapeIT that may have different phasing accuracies (see Fig. S.24 of Malinsky et al. 2017). To investigate the effects of phasing on ancestry reconstruction, we have conducted an additional experiment using the CEU-YRI admixture simulation data (admixture time of five generations). We have modified haplotype phases of target and admixed samples until there is a 30% switch error rate with the true haplotypic phase. We found that diploid accuracy is the same (97%) whether or not true haplotype phases are used, which shows that Loter is able to account for phasing errors.

**Simulated Populus Admixed Individuals**

We simulate individuals that result from admixture between *Populus trichocarpa*, which is adapted to relatively humid, moist, and mild conditions west of the Rocky Mountains and *Populus balsamifera*, which is a boreal species (Suarez-Gonzalez et al. 2016). As for human data, we compare Loter, RFMix, and LAMP-LD using diploid accuracy as a criterion for comparison. Again, the diploid accuracy of RFMix decreases with increasing admixture time. It ranges from a diploid accuracy of 92.0% when admixture occurs five generations ago...
to a diploid accuracy of 65.3% when admixture occurs 500 generations (fig. 5). By contrast to the simulations of human data, we find that the diploid accuracy of Loter and of LAMP-LD does not change with admixture time. For LAMP-LD, diploid accuracies ranges from 91.3% to 90.1% and for Loter it ranges from 89.9% to 89.0% when admixture increases from 5 to 500 generations (fig. 5).

Length of Ancestry Tracts

For simulated admixed Populus individuals, we additionally compare the length of *P. balsamifera* reconstructed ancestry tracts to the length of true ancestry tracts (fig. 6). When admixture took place 10 generations ago, RFMIX provides a distribution of ancestry tracts that is closer to the true distribution. For Populus simulations, the true median length of ancestry tract is of 4.26 cM and RFMIX finds 5.20 cM whereas LAMP-LD and Loter find median lengths of 0.05 and 2.31 cM, respectively. Both LAMP-LD and Loter return spurious chunks of local ancestry of small lengths and that contribute to reduce the mean length of local ancestry (supplementary fig. S6, Supplementary Material online). Additionally, several long blocks of ancestry chunks are cut into smaller pieces when using Loter or LAMP-LD (supplementary fig. S6, Supplementary Material online). When admixture took place 10 generations ago in the human simulations, both Loter and RFMIX provide the most accurate results; the true median length of ancestry tract is of 9.03 cM and RFMIX, Loter, and LAMP-LD reconstruct ancestry tracts of median length 9.99, 12.20, and 9.02 cM, respectively.

When admixture took place 200 or of 500 generations, length of ancestry chunks are more accurately reconstructed with Loter and with LAMP-LD than with RFMIX (fig. 6). For both human and Populus simulated data, RFMIX, by contrast to Loter and LAMP-LD, reconstructs ancestry tracts that are too long compared with true ancestry tracts when admixture is ≥200 generations. When admixture took place 200 generations ago, true median ancestry tracts is of 1 cM or less whereas RFMIX reconstructs tracts of 2 cM or more. For Populus simulations, the true median length of ancestry tract is of 0.45 cM when admixture took place 200 generations ago, and RFMIX, Loter, and LAMP-LD find, respectively, 2.00, 0.56, and 0.30 cM (supplementary fig. S6, Supplementary Material online). When admixture took place 500 generations ago, the true median length of ancestry tract is of 0.17 cM, and RFMIX, Loter, and LAMP-LD find, respectively, 1.60, 0.33, and 0.17 cM.

Application of Loter to Admixed Populus Individuals

We applied Loter to 36 individuals that are admixed between *P. balsamifera* and *P. trichocarpa*. When averaging local ancestry coefficients, we find that admixed individuals have on an average 87% of *P. trichocarpa* ancestry and 13% of *P. balsamifera* ancestry. We find that the median length of *P. balsamifera* ancestry tracts is equal to 0.76 cM and the first and third quartiles are equal to 0.25 cM and 1.47 cM. We also perform simulations of admixed individuals based on true genotypes from *P. balsamifera* and *trichocarpa* individuals. When admixture time varies from 10 to 500 generations; median reconstructed *P. balsamifera* ancestry tracts vary from 2.3 to 0.32 cM, the first quartile varies from 0.28 to 0.19 cM, and the third quartile varies from 4.18 to 0.55 cM (fig. 7). Of the six admixture times, we considered (µ ∈ {10, 50, 100, 200, 300, 500}) in the simulations, we find that µ = 100 generations provide the most similar distribution of *P. balsamifera* ancestry tracts; when µ = 100 generations, the three quartiles of *P. balsamifera* ancestry tracts are equal to 0.21 cM, 0.78 cM, and 1.29 cM.

Discussion

As dense genotype or sequencing data become more affordable, local ancestry inference provides an opportunity for admixture mapping and for deciphering admixture processes as well. We have introduced the software package Loter in order to make local ancestry available for a wide range of species for which biological parameters such as admixture times or recombination rates are not available. The regularization parameter λ, which controls smoothing, depends in a complicated manner on several biological and statistical parameters including mutation and recombination rates. To avoid the difficult choice of the regularization parameter, Loter implements an averaging procedure where we average solutions for different values of the regularization parameter. Thanks to an appropriate model averaging strategy, Loter does not require parameter tuning which makes it easy to use.

Loter is based on a mathematical formulation, which involves minimizing an objective function using dynamic programming. The Viterbi algorithm, which is a particular instance of dynamic programming, has already been used to reconstruct local ancestry in Hidden Markov models (HMM) (Sankararaman, Kimmel, et al. 2008; Baran et al. 2012).
Compared with other HMM formulations of local ancestries, Loter formulation is more parsimonious because it involves a single regularization parameter (Scheet and Stephens 2006; Baran et al. 2012). In addition to local ancestry reconstruction, dynamic programming has also been used to partition chromosomes into haplotype blocks, which correspond to discrete blocks of limited haplotype diversity (Zhang et al. 2002; Kimmel et al. 2003).

We compared Loter to the local ancestry software packages HAPMIX, RFMIX, and LAMP-LD. We found that the diploid accuracy obtained with HAPMIX is reduced by 32% in average when haplotypes are not known perfectly using trio-phasing but only reconstructed using phasing software such as Beagle. By contrast, RFMIX, LAMP-LD, and Loter are robust to imperfect haplotype reconstruction, which explains why only RFMIX, LAMP-LD, and Loter were further considered in software comparisons. When admixture took place five generations ago, RFMIX and LAMP-LD provide the largest diploid accuracies but all three software were found to provide an accurate reconstruction of local ancestry with diploid accuracies always >99% for the simulations of Afro-American admixed haplotypes and >89.9% for the simulations of Populus individuals. Compared with RFMIX and LAMP-LD, results obtained with Loter are more robust with respect to the time since admixture occurred. For simulated human data, the diploid accuracy of Loter decreases to 87% for the most ancient admixture times of 500 generations whereas it decreases to 72% and 81% when using RFMIX and LAMP-LD. For Populus data, diploid accuracy does not depend on admixture times when using Loter and LAMP-LD whereas it is severely impacted when using RFMIX. The fact that diploid accuracy is not impacted by the considered range of admixture times is encouraging and suggests that local ancestry inference is possible for admixture that occurred hundreds of generations ago when SNP density is large enough, as for Populus data; the mean distance between two SNPs being of 9.8 × 10⁻⁶ cM for Populus data and of 2.7 × 10⁻³ cM for the human data. A large enough genetic differentiation between parental populations is another explanation about why LAI can be accurate for Populus even for ancient admixture events; mean $F_{ST}$ between Populus balsamifera and Populus trichocarpa is equal to 0.22 (A. Geraldes, personal communication) whereas it is only of 0.14 when comparing CEU and YRI populations (Bhatia et al. 2013). Diploid accuracy may be impacted by admixture time and explanations may be related to the smoothing procedure of the different software. For instance, RFMIX has been tuned to genotypes resulting from recent admixture that occurred 10
generations ago or less such as admixture between African and Europeans (Gravel 2012; Bryc et al. 2015). When using default parameters of RFMix for data resulting from ancient admixture events, reconstructed ancestry tracts are inadequately long (fig. 6 and supplementary fig. S6, Supplementary Material online) and oversmoothing can affect diplodiplo
ccuracy. Even when providing true admixture time to RFMix, diplodiplo
ccuracy of RFMix remains more impacted by admixture time possibly because of the default of 0.2 cM for window size (supplementary fig. S7, Supplementary Material online). When using HAPMIX, a choice of window lengths or of the time since admixture should also be provided. However, choice of admixture time can be very difficult to estimate and impacts biological results. For instance, lengths of ancestry tracts found with HAPMIX depend on the input value for admixture time (Patin et al. 2014). The model averaging procedure implemented in Loter has the advantage to avoid putting a strong prior on a particular length of ancestry tract. In addition, model averaging improves parameter inference, which has already been observed when phasing genotypes using a statistical model of Linkage Disequilibrium (Scheet and Stephens 2006).

The simulation results show that accuracy obtained with Loter and LAMP-LD is less sensitive to admixture times compared with the accuracy obtained with RFMix. LAMP-LD is more accurate that Loter for recent admixture times when using human data and for all values of admixture times when considering the Populus data. However, LAMP-LD has limitations for large-scale NGS data that contains a large number of molecular markers. Its current implementation has a memory allocation limit of 50,000 SNPs and the software can be computer intensive. To perform local ancestry inference, it is recommended to use LAMP-LD and of 6 minutes using Loter. This small running time shows that Loter can handle massive genomic data set using a desktop or laptop computer. However, although there are differences between LAI software packages, using different LAI software packages can be a wise strategy to provide enhanced evidence for an association or a selection signal (Zhou et al. 2016). We expect that providing a parameter-free and rapid software for local ancestry inference will make more accessible genomic studies about admixture processes.

Materials and Methods

Dynamic Programming

The optimization problem of equation (1) is solved using dynamic programming. The solution of the problem with \(p\) SNPs can be derived from the solution with \((p-1)\) SNPs. Two configurations are possible. Either the admixed haplotype copies from the same haplotype at the \((p-1)\)th and \(p\)th SNP and

\[
C(s_1, \ldots, s_p) = C(s_1, \ldots, s_{p-1}) + |H^p - H^p_{p-1}|, \tag{2}
\]

or it uses different template haplotypes and

\[
C(s_1, \ldots, s_p) = C(s_1, \ldots, s_{p-1}) + |H^p - H^p_{p-1}| + \lambda. \tag{3}
\]

The optimal solution is then found by computing the shortest path on a graph displayed in supplementary figure S3, Supplementary Material online. To find the shortest path, dynamic programming computes at each node a quantity \(Q(i, j)\) that corresponds to the optimal solution for the first \(j\) SNPs and when the template haplotype at SNP \(j\) is the \(i\)th haplotype \(s_j\). The quantity \(Q(i, j)\) is updated as followed:

\[
Q(i, j) = \min \left( Q(i, j-1), \min_{\gamma \in \{1, \ldots, n\}} \{ Q(i', j-1) \} + \lambda \right) + |H^i - H^j|.
\]

Because we store the value of \(Q(i, j)\) at locus \(j-1\), the value of \(Q(i, j)\) can be computed as a minimum between two values. For each admixed haplotype, the complexity of this algorithm is therefore \(O(n \times p)\) where \(n\) is the number of individuals in the ancestral populations and \(p\) is the number of SNPs. The path \((s_1, \ldots, s_p)\) is then converted to an haploid ancestry sequence \(a = (a_1, \ldots, a_p) \in \{1, \ldots, K\}^p\) where \(a_j\) is the population of origin of the \(j\)th haplotype.

Accounting for Phase Errors of Admixed Genotypes

Reconstructed haplotypes from an admixed population may contain switch errors (Pasaniuc et al. 2009; Browning and Browning 2011). As considered in RFMix, it is possible to redistribute ancestry chunks among the two haplotypes from the same individual in order to correct for switch errors. For now, the software Loter accounts for phase error when there are two ancestral populations only. Once local ancestry values for each of the two haplotypes have been found after solving equation (1), we compute the sum of the two haplotypic local ancestries resulting in diploid local ancestry values \(d = (d_1, \ldots, d_p) \in \{0, 1, 2\}^p\) as returned by the software HAPMIX. Local ancestry values are then reconstructed using an internal ancestry phasing algorithm. The phasing procedure considers that the two haplotypic local ancestry sequences are a mosaic of two possible ancestry sequences \(A_1 = (0, \ldots, 0)\) and \(A_2 = (1, \ldots, 1)\) corresponding to the two possible ancestral populations. Two vectors \((s_1, \ldots, s_p) \in \{0, 1\}^p\) and \((s'_1, \ldots, s'_p) \in \{0, 1\}^p\) describe the sequence of labels, two vectors \(a \in \{0, 1\}^p\) and \(a' \in \{0, 1\}^p\) are the haploid local ancestry values, and \(\Theta\) is a compound parameter equal to \((s_1, \ldots, s_p, s'_1, \ldots, s'_p, a, a')\). The ancestry phasing algorithm consists of minimizing the following cost function:

\[
C'(\Theta) = \sum_{j=1}^{p} |d - A_j| + \sum_{j=1}^{p} |a| - A'_j + \lambda \sum_{j=1}^{p-1} \sum_{l=1}^{p-1} 1_{s_l \neq s_{l+1}}
\]

subject to the constraint that the sum of the haploid local ancestries \(a\) and \(a'\) is equal to the diploid local ancestry \(d\).

\[
\sum_{j=1}^{p} a_j = d, \quad \sum_{j=1}^{p} a'_j = d \tag{4}
\]
The solution for Θ is found using dynamic programming. For each admixed individual, the complexity of this algorithm is \(O(p)\). Once a solution has been found, haplotypic local ancestry, which has been corrected for phasing errors, consists of the two sequences \((A^1_p, \ldots, A^p_p)\) and \((A^1_p, \ldots, A^p_p)\).

Parameter Settings for Local Ancestry Software

For RFMix, LAMP-LD, and HAPMIX, we use the default parameter settings.

Admixture between Populus Species

We simulate admixed individuals by constructing their genomes from a mosaic of real P. balsamifera and P. trichocarpa individuals (Suarez-Gonzalez et al. 2016). We consider a probabilistic model that has already been used to simulate admixed individuals and to evaluate the performances of HAPMIX (Price et al. 2009). Simulations are based on 20 haplotypes (first 50, 000 SNPs) from chromosome 6 from the species P. balsamifera (balsam poplar) and of 20 haplotypes from the species P. trichocarpa (black cottonwood) (Suarez-Gonzalez et al. 2016). Haplotypes were obtained from genotypes using Beagle (Browning and Browning 2007). The P. trichocarpa ancestry \(x_i\) of a simulated admixed individual is drawn randomly according to a Beta distribution of mean 0.8 and of variance 0.1. At the first marker, the haplotype of an admixed individual \(i\) is assumed to originate from P. trichocarpa with a probability \(x_i\) and from P. balsamifera otherwise. For each simulated haplotype, we associate one P. balsamifera haplotype and one P. trichocarpa haplotype. For a given admixed individual, haplotypes are exclusively copied from these two source haplotypes that are chosen at random. The length (measured in Morgans) of an ancestry chunk is drawn according to an exponential distribution of rate \(\mu\) generations. In the simulations, we consider values for \(\mu\) ranging from 5 to 500 generations. The species origin of the new ancestry tract is again determined using the \((x_i, 1-x_i)\) admixture coefficients and the copying process for haplotype is repeated as before. To reconstruct local ancestry of simulated admixed individuals, we consider 30 haplotypes from P. balsamifera and 30 haplotypes from P. trichocarpa that were not used when simulating admixed individuals. Haplotypes were again phased using the software Beagle. To evaluate diploid accuracy for a given value of \(\mu\), we consider 20 sets of simulations consisting of 20 admixed haplotypes each. When considering reference panels different from the source populations, we consider 100 haplotypes from the HAPMAP 3 Luhya sample and 176 samples from the HAPMAP 3 TSI samples. Local ancestry reconstruction was based on true haplotypes from the LWK and TSI sample.

We consider an additional set of simulations where three populations admixed. Admixture is assumed to occur between Chinese (CHB), Europeans (CEU), and Africans (YRI). The exponential distribution for European chunks is of rate \(\mu \in \{5, 100, 200, 500\}\) generations and the exponential distribution for African and Asian chunks is equal to \(\mu/2\) generations. By contrast to the two-way admixture models, we use trio-phased haplotypes for inference and not haplotypes reconstructed with Beagle.

Admixed Populus Individuals

We consider 36 individuals that are admixed between P. balsamifera and P. trichocarpa (Suarez-Gonzalez et al. 2016) and use Beagle to phase them. We use the first 500,000 SNPs of chromosome 6 to reconstruct ancestry tracts. We simulate 16 admixed individuals based on 20 genotypes from P. balsamifera and P. trichocarpa species. Instead of considering true ancestry ancestry tracts, we rather replicate the same pipeline as for real data such as the bias of ancestry tract reconstruction should be same for data and simulations. We phase simulated individuals using Beagle and reconstruct ancestry tracts using Loter.

Software

The Loter software package and its source code are available at https://github.com/bcm-uga/Loter.

Data


Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments

Authors acknowledge Grenoble Alpes Data Institute, supported by the French National Research Agency under the “Investissements d’avenir” program (ANR-15-IDEX-02), the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01), the MACARON project (ANR-14-CE23-0003), and the FROGH project (ANR-16-CE12-0033). JM is also supported by a grant from the European Research Council (SOLARIS project, number 714381).

References
