Comparing local ancestry inference models in populations of two- and three-way admixture

Ryan Schubert
Loyola University Chicago
Angela Andaleon
Loyola University Chicago
Heather E. Wheeler (hwheeler1@luc.edu)
Loyola University Chicago https://orcid.org/0000-0003-1365-9667

Research article

Keywords: Comparing local ancestry inference models, in populations of two- and three-way admixture

Posted Date: January 24th, 2020

DOI: https://doi.org/10.21203/rs.2.21798/v1

Version of Record: A version of this preprint was published at PeerJ on October 2nd, 2020. See the published version at https://doi.org/10.7717/peerj.10090.

RESEARCH

Comparing local ancestry inference models in populations of two- and three-way admixture

Ryan Schubert[1,2,3], Angela Andaleon[2,3] and Heather E. Wheeler[2,3,4,5*]

*Correspondence:
hwheeler1@luc.edu
[2]Program in Bioinformatics, Loyola University Chicago, 60660 United States
[3]Department of Biology, Loyola University Chicago, 60660 Chicago, IL, United States
[4]Department of Computer Science, Loyola University Chicago, 60660 Chicago, IL, United States
[5]Department of Public Health Sciences, Loyola University Chicago, 60153 Maywood, IL, United States
Full list of author information is available at the end of the article

Abstract

Background: Local ancestry estimation infers the regional ancestral origin of chromosomal segments in admixed populations using reference populations and a variety of statistical models. Integrating local ancestry into complex trait genetics has the potential to increase detection of genetic associations and improve genetic prediction models in understudied admixed populations, including African Americans and Hispanics. Five methods for local ancestry estimation are LAMP-LD (2012), RFMix (2013), ELAI (2014), Loter (2018), and MOSAIC (2019), but direct comparisons of accuracy, runtime, and memory usage of all these software tools have not previously been reported across common patterns of human admixture.

Results: We found that in cases of two-way admixture, RFMix and ELAI had the highest median accuracy depending on population structure, while in cases of three-way admixture, we found RFMix, MOSAIC, and LAMP-LD had the highest median accuracy. Additionally, we estimate the O(n) of both memory and runtime for each software and find that for both time and memory most software expand linearly with respect to sample size. The only exception is RFMix, which expands quadratically with respect to runtime and linearly with respect to memory.

Conclusions: Effective local ancestry estimation tools are necessary to combat population disparities in human genetics studies. RFMix performs the best across methods, however, depending on application, other methods perform similarly well with the benefit of shorter runtimes. Scripts used to format data, run software, and estimate accuracy can be found at `https://github.com/WheelerLab/LAI_benchmarking`.

Keywords: local ancestry; human genetics; admixture; benchmarking

Background

Humans are a chromosomal mosaic of their ancestors. Through sexual reproduction and recombination, chromosomes resemble a subset of their ancestors' chromosomes in varying sizes and locations across the genome [1]. Large scale studies of the genetics underlying human disease have been limited to predominantly European populations and thus lack global diversity, which exacerbates health disparities [2, 3]. It is well documented that prediction accuracy with polygenic risk scores decreases with increasing genetic distance [4, 5]. In addition, many underrepresented populations in human genetics include recently admixed individuals, meaning their ancestors were previously isolated from each other on different continents until the last few centuries. This leads to chromosomal tracts originating from different continental populations in modern populations like African Americans and Hispanics.

Population structure is a potential confounding factor in all genetic association studies. Global ancestry is the proportion of different ancestral populations represented across the entire genome. Genotypic principal components are used to adjust for these average genomic background effects in genetic association studies [6]. Correcting only for global ancestry does not precisely account for ancestry at any specific locus. Local ancestry is the number of alleles derived from distinct ancestral populations at a given locus and may improve power to detect genetic associations in admixed populations [7–10]. For example, a recent expression quantitative trait (eQTL) mapping study in African Americans found a greater replication rate of eQTLs discovered via models that adjust for local ancestry, rather than models that adjust for global ancestry [11].

Several models have been developed to estimate local ancestry in admixed populations [1, 12–17]. By leveraging population or continental-specific SNPs, chromosomal tracts can be differentiated into their ancestral segments. Chromosomal regions are compared to reference populations of non-admixed ancestry to find which sections of the chromosomes descend from which continental region [1, 18]. These estimates depend largely on the reference populations used, the genetic distance between the reference samples, the quality of the input genotypes, and, most importantly, the statistical models. LAMP-LD demonstrates strong ancestry estimation in recently admixed cohorts of African and Hispanic descent [12]. ELAI and Loter both report stable performance in instances of ancient admixture ($ngenerations \geq 100$), out competing methods that prioritize recent admixture[14, 16]. Additionally, Loter reports high performance in nonhuman species [16]. Similar to LAMP-LD, RFMix and MOSAIC each specialize in multi-way admixture. Unlike LAMP-LD, neither are constrained in the number of ancestral populations. Both RFMix and MOSAIC are reported to have robust performance even when reference panels are not closely related to the study population, though MOSAIC reported the added benefit of elucidating the relationship between all provided references and the study population and selecting the optimal references, thus circumventing the need to clarify the relationship between study and available reference populations [13, 17].

To satisfy the growing call for increased diversity in genome-wide association studies [2, 3], local ancestry estimation methods will become increasingly important in human genetics. While a recent review compared the underlying models of several local ancestry estimation software tools [19], accuracy and run time were not directly compared. A study from 2017 compared run time and memory usage of four older tools [20], but did not include the widely used RFMix [13] and newer tools MOSAIC [17] and Loter [16]. Here, we independently compare five local ancestry estimation methods for accuracy and feasibility by simulating admixed chromosomes from both two and three ancestral continental populations.

Results

We prioritize benchmarking each software in the context of recently admixed populations to assess accuracy and estimate previously unreported time and memory complexity. We selected five software for a combination of their novelty and relative popularity. LAMP-LD [12], ELAI [14], and RFMix [13] are each established local ancestry software that have been cited numerous times in the field of population genetics. Conversely, MOSAIC [17] and Loter [16] are fairly new, having been

published in the last two years at the time of writing. A brief summary of their differences can be found in Table 1.

Table 1 Software descriptions. Features and requirements of each software as described in the original publication. HMM: hidden Markov model

software	algorithm	pre-phasing	gen map	n ancestral pops
LAMP-LD	single layer HMM	not required	no	2,3,5
RFMix	random forest	required	yes	$n \geq 2$
ELAI	two layer HMM	not required	no	$n \geq 2$
Loter	single layer HMM	required	no	$n \geq 2$
MOSAIC	two layer HMM	required	yes	$n \geq 2$

Simulating admixed individuals

We simulated admixed populations with ancestry proportions similar to those observed in previous studies [18]. These include two-way admixture between YRI and CEU representing a common pattern of descent for African American individuals (AFA); two-way admixture between PEL and CEU representing one common pattern of descent for some Hispanic individuals (HIS); and three-way admixture between PEL, YRI, and CEU, representing another common pattern of descent among some Hispanic individuals (3WAY) [18]. For each admixture group, we simulated 1000 individuals and selected 100 that had European ancestry within 10% of the admixture proportions listed in Table 2. We summarize our workflow in Fig. 1, see Methods for details.

Table 2 Proportions of ancestry among simulated cohorts. We simulated 1000 admixed individuals. From these we selected 100 individuals with a true proportion of European ancestry within 10% of the proportions listed here for use in our accuracy testing.

Ancestral Panel	%YRI	%CEU	%PEL
AFA	80%	20%	0%
HIS	0%	70%	30%
3WAY	30%	60%	10%

Runtime and memory usage

Runtime increases with number of individuals

We simulated an additional 2000 individuals based on the AFA admixture proportions at 7 generations since admixture. We randomly subset this set of people to 2000, 1500, 1000, 500, 100, 50, and 20 individuals to test how each software scales with an increasing sample size (Fig. 2). We find that the runtimes of four of the five software tools scale linearly with the number of samples, with the exception of RFMix, which scales quadratically (Table 3). We also note that MOSAIC runtime decreases when n=2000. MOSAIC will exit early the iteration of its expectation-likelihood algorithm when the log-likelihood decreases resulting in in cases where it finishes faster than would be expected by a standard linear model [17].

Memory increases linearly with number of individuals

We simultaneously measured the memory burden expected for each level of sample size (Fig. 3). We found that all software expand linearly or near linearly (Table 4). Loter had the steepest memory requirement and ELAI had the smallest slope. ELAI has the most stable memory requirement across sample sizes. At high sample sizes ELAI had the lowest memory overhead, but at low sample sizes ($n \leq 100$) the memory requirement was third highest.

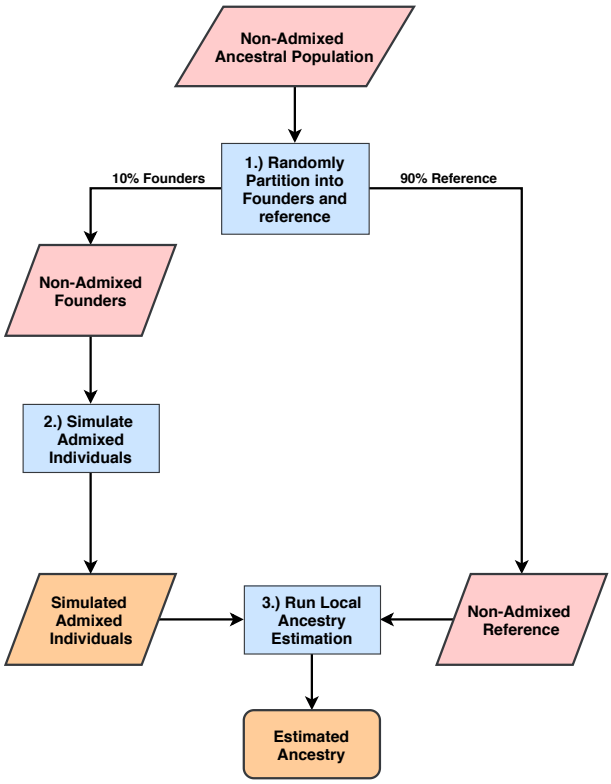


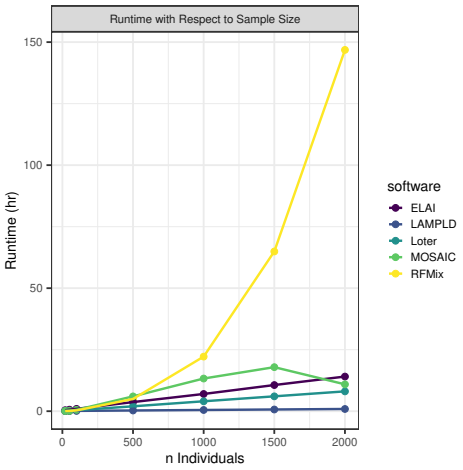
Figure 1 Process for simulating admixed individuals and estimating ancestry. 1) From non-admixed populations from 1000G we randomly select 10% of all individuals to use as founders for admixture simulation. The rest are used as reference panels for ancestry estimation. 2) We generate admixed individuals that are chromosomal mosaics of the founder group using the admixture simulation tool created by the authors of RFMix [13]. 3) Using the remaining our 1000G populations as reference panels, we estimate ancestry on our simulated population for all five softwares. Afterwards we compare estimation accuracy across software.

Table 3 Linear runtime estimated $O(n)$. We fit linear and quadratic models between the runtime and sample size for each software. We report the model R^2 and ANOVA p-value for each combination of software and model.

Software	Linear R^2	Linear p-value	Quadratic R^2	Quadratic p-value
RFMix	0.853	0.00294	0.977	1.57×10^{-05}
MOSAIC	0.673	0.0146	0.387	0.0802
ELAI	0.999	8.97×10^{-11}	0.921	0.000383
Loter	0.999	2.62×10^{-13}	0.91	0.000438
LAMP-LD	0.999	7.90×10^{-14}	0.917	0.000427

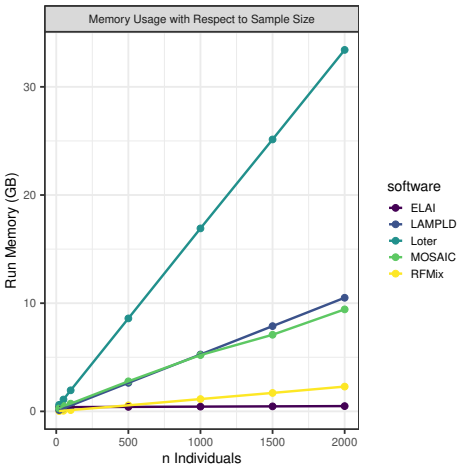
Table 4 Linear maximum memory usage estimated $O(n)$. We fit a linear model between the maximum memory usage and sample size for each software. We report the estimated β_1, model R^2, and ANOVA p-value for each combination of software and model.

Software	Linear R^2	Linear p-value	Quadratic R^2	Quadtratic p-value
RFMix	0.999	8.61×10^{-12}	0.9205	0.000394
MOSAIC	0.997	6.049×10^{-08}	0.894	0.000813
ELAI	0.999	1.03×10^{-11}	0.9154	0.000459
Loter	1	1.25×10^{-14}	0.9154	0.000459
LAMP-LD	1	$< 2.2 \times 10^{-16}$	0.9165	0.000445



v n Figure 2.pdf

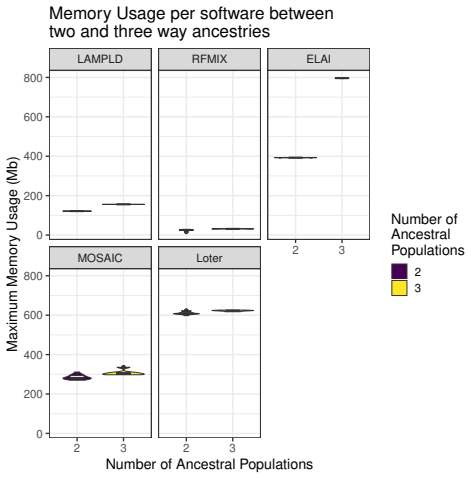
Figure 2 Software runtime versus sample size. We tested the runtime of each software on one core at a sample sizes of 20, 50, 100, 500, 1000, 1500, and 2000. Points represented sample sizes tested versus runtime in hours, which are connected by line segments colored by software. We simulated n African American individuals from CEU and YRI "founder" populations with average admixtures proportions of 20% and 80%, respectively. We find that the runtime of ELAI, LAMP-LD, MOSAIC, and Loter all increase linearly with the number of samples. The runtime of RFMix increases quadratically.



v n Figure 3.pdf

Figure 3 Software memory usage versus sample size. We tested the maximum memory usage of each software on one core at a sample size of 20, 50, 100, 500, 1000, 1500, and 2000. Points represented sample sizes tested versus memory, which are connected by line segments colored by software. We simulated n African American individuals from CEU and YRI "founder" populations with average admixtures proportions of 20% and 80%, respectively. We find that maximum memory usage for all software increases linearly with the number of samples.

Increasing number of ancestries can increase runtime and memory burden
We tested if increasing the number of ancestral populations increases the computational burden of each software. We found that in all software, increasing the number of ancestral populations resulted in a significant increase in memory usage (Fig. 4). However, increasing number of ancestries did not impact the runtime for two software: Loter and MOSAIC. For the three other software, increasing the number of ancestries did significantly increase the runtime (Fig. 5).

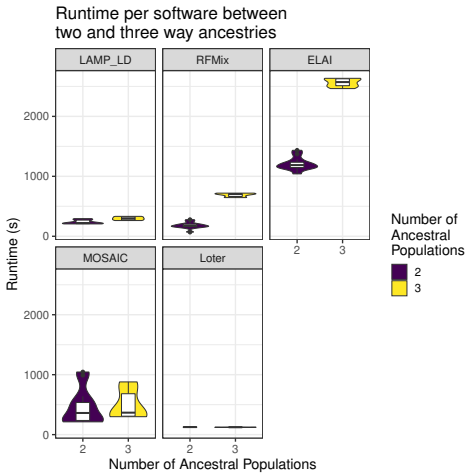


v n ancestries Figure 4.pdf

Figure 4 We test to see if there is a significant difference between the memory usage of each software when increasing the number of ancestries estimated. We estimated the ancestry of 20 unique individuals 3 times for both two and three ancestries and ran a one-way ANOVA between number of ancestries and the maximum memory usage. Significance by ANOVA: RFMix $p = 4 \times 10^{-4}$; LAMP-LD $p < 2 \times 10^{-16}$; ELAI $p =< 2 \times 10^{-16}$; MOSAIC $p = 2.5 \times 10^{-3}$; Loter $p = 1.17 \times 10^{-3}$)

Accuracy varies by cohort composition

For each admixture group (Table 2), we simulated 100 individuals and ran local ancestry estimation and accuracy benchmarking. Each software performs with high fidelity in regards to two-way admixture, but we note a considerable difference in our simulated two-way AFA and HIS cohorts. We attribute this to the introduction of the PEL population as both founders and reference, as they contain a significant amount of admixture in and of themselves. As their admixture overlaps with the other two reference populations, it is expected that they will introduce noise into our local ancestry estimation. For two way admixture, RFMix and ELAI had the highest median accuracy for AFA and HIS, respectively, though all software performed competitively well. For three way admixture, RFMix had the highest median accuracy (Fig. 6). After assessing accuracy of each software we performed a Tukey's test to determine which pairs of software performed significantly differently. In the case of our simulated AFA cohort, it was found that both RFMix and ELAI performed significantly better than both LAMP-LD and Loter. All other pairs were not significantly different (Additional File 1). In the case of our HIS cohort, we found that RFMix performed significantly better than LAMP-LD, with all other pairs found to be not significantly different (Additional File 2). In the



v n ancestries Figure 5.pdf

Figure 5 We test to see if there is a significant difference between the runtime of each software when increasing the number of ancestries estimated. We estimated the ancestry of 20 unique individuals 3 times for both two and three ancestries and ran a two-way ANOVA between number of ancestries and the maximum memory usage. Significance by ANOVA: RFMix $p = 2.95 \times 10^{-13}$; LAMP-LD $p = 2.46 \times 10^{-3}$; ELAI $p = 7.64 \times 10^{-15}$; MOSAIC $p = 0.797$; Loter $p = 0.308$

case of our 3WAY cohort, we found that each of RFMix, MOSAIC, and LAMP-LD performed significantly better than both Loter and ELAI with all other pairs found not significantly different (Additional File 3).

Software is highly correlated on real data

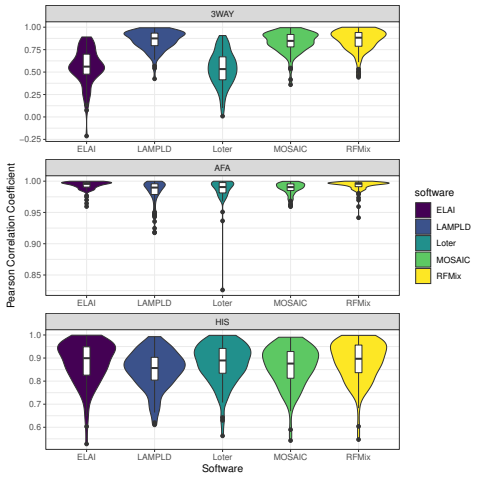
We ran each software as described on real admixed individuals from the ASW population of the 1000 genomes project with the YRI and CEU populations as reference panels. Local ancestry estimates were highly correlated between each software (Table 5). Additionally, to show the robustness of these estimates, we plot the mean local African ancestry estimated by each software against the first principal component of the genotypes, which is known to be an estimate of global African ancestry (Fig. 7) [11]. The local ancestry estimate was highly correlated with PC1 for all software tools ($R^2 > 0.960$, Table 5), with no significant difference between tools ($p > 0.621$).

Table 5 Between software Pearson correlation using real data. We ran all five software on 61 real admixed individuals from the 1000 Genomes ASW population. Here we report the squared pairwise Pearson correlations of local ancestry estimates. Additionally, in the last column, we report the squared correlation of each software's estimated mean African ancestry with genotypic principal component 1.

	ELAI	LAMP-LD	Loter	RFMix	MOSAIC	PC1
ELAI	1	0.977	0.976	0.974	0.959	0.965
LAMP-LD	0.977	1	0.980	0.977	0.959	0.968
Loter	0.976	0.981	1	0.977	0.962	0.967
RFMix	0.974	0.977	0.977	1	0.961	0.967
MOSAIC	0.959	0.959	0.962	0.961	1	0.960

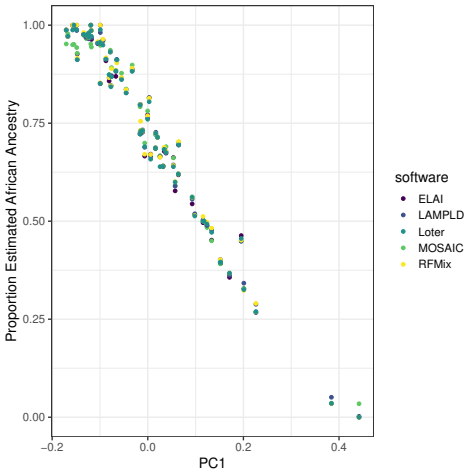
Discussion

Local ancestry estimation is key step in adjusting for potential population stratification in admixed populations and in elucidating the effect of ancestry specific



accuracy Figure 6.pdf

Figure 6 Distribution of accuracy for ancestry estimation. For each category of admixture we estimate ancestry on 100 simulated individuals. Accuracy is then calculated as the pearson correlation between true and estimated ancestry per each software. By Tukey's test we find significant differences between the following pairs in 3WAY: LAMP-LD $>$ ELAI ($p = 1.1 \times 10^{-10}$), MOSAIC $>$ ELAI ($p = 1.1 \times 10^{-10}$), RFMix $>$ ELAI ($p = 1.1 \times 10^{-10}$), LAMP-LD $>$ Loter ($p = 1.1 \times 10^{-10}$), MOSAIC $>$ Loter ($p = 1.1 \times 10^{-10}$), and RFMix $>$ Loter ($p = 1.1 \times 10^{-10}$). In AFA: ELAI $>$ LAMP-LD ($p = 1.7 \times 10^{-6}$), ELAI $>$ Loter ($p = 0.010$), RFMix $>$ LAMP-LD ($p = 3.1 \times 10^{-6}$), RFMix $>$ Loter ($p = 0.015$); In HIS: RFMix $>$ LAMP-LD ($p = 0.027$). All other pairs did not significantly differ. Note the y-axis differs in each plot to emphasize within simulated population differences.



v PC1 Figure 7.pdf

Figure 7 We plot the the mean local ancestry proportion of African ancestry estimated by each software against the first principal component of genotypes, a known estimate of global African ancestry, to validate the robustness of our local ancestry estimates. The local ancestry estimate was highly correlated with PC1 for all software tools ($R^2 > 0.96$), with no significant difference between tools ($p > 0.62$)

loci on complex traits. Given the wide variety of tools available to perform local ancestry estimation, it is necessary to explore how each performs in a particular context. Here, we focused on recent human admixture within African American and Hispanic populations, and performed complexity and accuracy analyses of five different software tools using simulated and real data.

We did not consider instances of ancient admixture despite ELAI and Loter reporting robust performance in such instances [14, 16], which could be one reason they underperformed in our 3-way simulation (Fig 6). In addition, Loter was designed to be compatible with many different species and both Loter and ELAI may require more fine-tuning of software parameters beyond the default settings than the other methods, especially in cases of 3-way admixture.

Here we report on how memory and time usage scale with number of individuals and not SNPs, as it is simpler and more common to scale studies by population size than by genome size. However, it is expected that most if not all software will increase in both time and memory usage given an increased number of SNPs. We find that all software perform with high accuracy in cases of two-way admixture, with RFMix and ELAI performing the best. In cases of three-way admixture, RFMix had the highest median accuracy and RFMix, MOSAIC, and LAMP-LD all performed significantly better than ELAI and Loter. While RFMix has a relatively low memory overhead, its runtime scales quadratically, severely limiting its scalability at standard GWAS sample sizes.

An important consideration in all cases is the availability of high quality reference data. Currently, Native American genetic data is not widely available due to cultural and historical incidents that have raised barriers between the tribal communities and the genetic community [21, 22]. Here we use the PEL population as a proxy for non-admixed individuals of Native American descent as PEL has the highest portion of NAT ancestry among 1000G populations. However, PEL introduces noise as it contains significant admixture. This noise likely causes our HIS and 3WAY simulated populations to underperform. Still, our simulations show robust performance of several software.

Conclusion

We find that in cases of two-way simulated admixture, each software performs similarly well with RFMix and ELAI having the highest median performance depending on the population structure. In our three-way simulated admixed population, we see marked difference in performance, with RFMix performing best overall, followed by LAMP-LD and MOSAIC. While RFMix performs the best across methods, its scalability with regards to time may give weight to considering other software. Robust, scalable local ancestry estimation software are crucial for equitable implementation of genetics and genomics in medicine.

Methods

Simulating Genotypes

Our workflow is summarized in Fig 1. We chose three 1000 genomes (1000G) populations [23] to serve as non-admixed ancestral populations. From each of these

populations we randomly selected 10% of individuals to use as founders for simulation of admixed individuals and the remaining individuals made up the non-admixed reference populations. The three 1000G populations from which we drew samples are: Utah residents with Northern and Western European ancestry (CEU) for use as our European ancestral group; Yoruba in Ibadan, Nigeria (YRI) for use as our African ancestral group; and Peruvians from Lima, Peru (PEL) for use as our Native American ancestral group. We note that individuals in the PEL population have Native American, European and African admixture, however, the PEL have more Native American ancestry than all of the other American populations in 1000G ($\mu = 0.77$, 95% CI = [0.75–0.80] [4]. The PEL population thus serves as a reasonable proxy for our Native American ancestral population. Simulated admixed populations fall into one of three categories; two-way admixture between YRI and CEU representing a common pattern of descent for African American individuals (AFA); two-way admixture between PEL and CEU representing one common pattern of descent for some Hispanic individuals (HIS); and lastly three-way admixture between PEL, YRI, and CEU, representing another common pattern of descent among some Hispanic individuals (3WAY) as observed in [18]. For each admixture group, we simulated 1000 individuals and selected 100 that had European ancestry within 10% of the admixture proportions described in Table 2. Global ancestry percentages across individuals are shown in Additional Files 4-6.

We used the admixture simulation tool developed by the creators of RFMix to generate simulated admixed chromosomes [13]. We limited our simulation to SNPs on chromosome 22, for a total of 158,159 SNPs. LAMP-LD v 1.0 has a computational limit of 50,000 random SNPs. In keeping with this, after simulating the entirety of chromosome 22, we independently selected 50,000 SNPs from each cohort using the `--thin-count 50000` option in PLINK [24] and subset each cohort accordingly. The code used to run simulation can be found at `https://github.com/WheelerLab/LAI_benchmarking`.

Running Each Software

We used individuals remaining within the non-admixed ancestral group after founder selection as the required reference group for running each of the five software. We ran each software using default parameters or using the minimum number of settings necessary as this is representative of how most new users will interact with each software. We ran each software as follows:

LAMP-LD v1.0

```
unolanc 300 15 <snp position file> <ancestral haploypes 1>
<ancestral haploypes 2> <ancestral haploypes 3> <admixed
haploypes> <output name>

unolanc2way 300 15 300 15 <snp position file> <ancestral
haploypes 1> <ancestral haploypes 2> <admixed haploypes>
<output name>
```

MOSAIC v1.3

```
Rscript mosaic.R <admixed population name> <folder
containing required input> -c <chr range> -a <number of
ancestries to infer> -m <maximum number of cores> --gens
<number of generations>
```

Loter

```
loter_cli -r <reference panel genotype/haplotype> -a
<admixed genotype/haplotype> -f <genotype file format>
-o <output name> -n <number of cores> -v
```

ELAI v1.01

```
elai-lin -g <ancestral haploypes 1> -p 10 -g <ancestral
haploypes 2> -p 11 -g <ancestral haploypes 3> -p 12 -g
<admixed haploypes> -p 1 -pos <snp position file> -C 3
-o <output name>

elai-lin -g <ancestral haploypes 1> -p 10 -g <ancestral
haploypes 2> -p 11 -g <admixed haploypes> -p 1 -pos
<snp position file> -C 2 -o <output name>
```

RFMix v1.5.4

```
python RunRFMix.py -e 2 -w 0.2 --num-threads <maximum
number of cores> --forward-backward PopPhased
<population haploypes> <population classes file> <snp
position file> -o <output name>
```

In all cases we ran software on one core. In cases with three ancestries, 11 was used for number of generations. In cases with two ancestries, 8 was used for number of generations. In most cases each software requires a genetic map file or SNP position file, the number of generations since admixture, and reference/admixed genotypes in a software specific format. As our genotype data was already phased, we do not consider phasing in this paper, though it could be considered a necessary step 0 of this process. As each software carries different requirements for formatting, we have constructed a brief pipeline for formatting and running each software. All scripts used to run each software can be found at `https://github.com/WheelerLab/LAI_benchmarking`.

Benchmarking Each Software

We used the bash command `time -v` to benchmark time and memory of each software run. To benchmark time and memory usage with increasing sample size, we used the methods described above and simulated an additional 2000 two-way admixed AFA individuals to test time and memory burden at each level of 20, 50, 100, 500, 1000, 1500, and 2000 individuals. We performed regression analysis of time and memory complexity in base R for each software.

We defined accuracy as the Pearson correlation for each individual in a simulated population. For each individual, we calculated Pearson correlation of all SNPs tested between the known ancestry output by the ancestry simulation tool and the ancestry inferred by a given software.

Additional Files

Additional file 1. 95% family confidence intervals for Tukey's test when running each software on a two way AFA admixed population. We find significant differences between the following pairs: LAMPLD and ELAI (p=1.72e-06), Loter and ELAI (p=1.01e-02), RFMix and LAMPLD (p=3.08e-06), RFMix and Loter (p=1.47e-02). All other pairs were found to not have significantly different means.

Additional file 2. 95% family confidence intervals for Tukey's test when running each software on a two way HIS admixed population. We find significant differences between RFMix and LAMPLD (p=2.66e-2). All other pairs were not found significantly different.

Additional file 3. 95% family confidence intervals for Tukey's test when running each software on three way admixed population. We find significant differences between the following pairs: LAMPLD and ELAI (p=1.10e-10), MOSAIC and ELAI (p=1.10e-10), RFMix and ELAI (p=1.10e-10), Loter and LAMPLD (p=1.10e-10), MOSAIC and Loter (p=1.10e-10), and RFMix and Loter (p=1.10e-10). All other pairs were not found significantly different.

Additional file 4. Distribution of global ancestries in simulated admixed AFA population. Each bar is an individual with their proportion of global ancestries represented per haplotype.

Additional file 5. Distribution of global ancestries in simulated admixed HIS population. Each bar is an individual with their proportion of global ancestries represented per haplotype.

Additional file 6. Distribution of global ancestries in simulated admixed 3WAY population. Each bar is an individual with their proportion of global ancestries represented per haplotype.

Abbreviations

1000G - 1000 Genomes.
3WAY - Admixture group signifying three way admixture between CEU, PEL, and YRI.
eQTL - expression Quantitative Trait Loci.
AFA - Admixture group signifying two way admixture between CEU and YRI individuals.
CEU - 1000 Genomes Utah residents with Northern and Western European ancestry.
HIS - Admixture group signifying two way admixture between PEL and CEU.
PEL - 1000 Genomes Peruvians from Lima, Peru.
YRI - 1000 Genomes Yoruba in Ibadan, Nigeria.

Declarations

Ethics approval and consent to participate

This project was confirmed exempt from human subjects federal regulations under exemption number 4 by the Loyola University Chicago Institutional Review Board (project number 2014).

Consent for publication

Not Applicable

Availability of data and materials

All genomic data used is publically available 1000 Genomes Project data from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/`.
LAMP-LD is available at `lamp.icsi.berkeley.edu/lamp/lampld/`.
RFMix is available at `sites.google.com/site/rfmixlocalancestryinference/`.
ELAI is available at `https://haplotype.org/software.html`.
MOSAIC is available at `https://maths.ucd.ie/~mst/MOSAIC/`.
Loter is available at `https://github.com/bcm-uga/Loter`.
Admixture simulation software used is available at `github.com/slowkoni/admixture-simulation`.
All scripts used in analyses are available at `https://github.com/WheelerLab/LAI_benchmarking`

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by the NIH National Human Genome Research Institute Academic Research Enhancement Award R15 HG009569 (HEW), the Loyola Carbon Undergraduate Research Fellowship (AA), and the Loyola MS Bioinformatics Research Assistant Fellowship (AA).

Author's contributions

AA, RAS, and HEW designed the study. RAS and AA performed the analyses. RAS, AA, and HEW wrote the manuscript. All authors have read and approved the manuscript.

Acknowledgements

We thank Dr. Catherine Putonti for her feedback during development of this project.

Author details
[1]Department of Mathematics and Statistics, Loyola University Chicago, 60660 Chicago, IL, United States. [2]Program in Bioinformatics, Loyola University Chicago, 60660 United States. [3]Department of Biology, Loyola University Chicago, 60660 Chicago, IL, United States. [4]Department of Computer Science, Loyola University Chicago, 60660 Chicago, IL, United States. [5]Department of Public Health Sciences, Loyola University Chicago, 60153 Maywood, IL, United States.

References

1. Gravel, S.: Population genetics models of local ancestry. Genetics **191**(2), 607–619 (2012). doi:10.1534/genetics.112.139808
2. Popejoy, A.B., Fullerton, S.M.: Genomics is failing on diversity. Nature **538**, 161–164 (2016). doi:10.1038/538161a
3. Sirugo, G., Williams, S.M., Tishkoff, S.A.: The Missing Diversity in Human Genetic Studies. Cell **177**, 26–31 (2019). doi:10.1016/j.cell.2019.02.048
4. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., Kenny, E.E.: Human demographic history impacts genetic risk prediction across diverse populations. The American Journal of Human Genetics **100**(4), 635–649 (2017)
5. Mogil, L.S., Andaleon, A., Badalamenti, A., Dickinson, S.P., Guo, X., Rotter, J.I., Johnson, W.C., Im, H.K., Liu, Y., Wheeler, H.E.: Genetic architecture of gene expression traits across diverse populations. PLoS genetics **14**(8), 1007586 (2018)
6. Price, A.J., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D.: Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38**, 904–909 (2006). doi:10.1038/ng1847
7. Duan, Q., Xu, Z., Raffield, L.M., Chang, S., Wu, D., Lange, E.M., Reiner, A.P., Li, Y.: A robust and powerful two-step testing procedure for local ancestry adjusted allelic association analysis in admixed populations. Genetic Epidemiology **42**(3), 288–302 (2017). doi:10.1002/gepi.22104
8. Martin, E.R., Tunc, I., Liu, Z., Slifer, S.H., Beecham, A.H., Beecham, G.W.: Properties of Global and Local Ancestry Adjustments in Genetic Association Tests in Admixed Populations. Genetic Epidemiology **42**(2), 214–229 (2017). doi:10.1002/gepi.22103
9. Zhang, J., Stram, D.O.: The role of local ancestry adjustment in association studies using admixed populations. Genetic Epidemiology **96**(6), 502–515 (2014). doi:10.1002/gepi.21835
10. Liu, J., Lewinger, J.P., Gilliland, F.D., Gauderman, W.J., Conti, D.V.: Confounding and heterogeneity in genetic association studies with admixed populations. American Journal of Epidemiology **177**(4), 351–360 (2013). doi:10.1093/aje/kws234
11. Zhong, Y., Perera, M.A., Gamazon, E.R.: On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. American Journal of Human Genetics **104**, 1097–1115 (2019). doi:10.1016/j.ajhg.2019.04.009
12. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., Rodriguez-Santana, J., Burchard, E.G., Halperin, E.: Fast and accurate inference of local ancestry in Latino populations. Bioinformatics **28**(10), 1359–1367 (2012). doi:10.1093/bioinformatics/bts144
13. Maples, B.K., Gravel, S., Kenny, E.E., Bustamante, C.D.: RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. American Journal of Human Genetics **93**(2), 278–288 (2013). doi:10.1016/j.ajhg.2013.06.020
14. Guan, Y.: Detecting structure of haplotypes and local ancestry. Genetics **196**(3), 625–642 (2014). doi:10.1534/genetics.113.160697. 1304.1834
15. Thornton, T.A., Bermejo, J.L.: Local and global ancestry inference and applications to genetic association analysis for admixed populations. Genetic epidemiology **38**(S1), 5–12 (2014)
16. Dias-Alves, T., Mairal, J., Blum, M.B.B.: Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. Molecular Biology and Evolution **35**(9), 2318–2326 (2018). doi:10.1093/molbev/msy126
17. Salter-Townshend, M., Myers, S.: Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. Genetics **212**(3), 869–889 (2019). doi:10.1534/genetics.119.302139
18. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., Mountain, J.L.: The genetic ancestry of African Americans, Latinos, and European Americans across the United States. American Journal of Human Genetics **96**(1), 37–53 (2015). doi:10.1016/j.ajhg.2014.11.010
19. Geza, E., Mugo, J., Mulder, N.J., Wonkam, A., Chimusa, E.R., Mazandu, G.K.: A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Briefings in Bioinformatics **20**(5), 1709–1724 (2018). doi:10.1093/bib/bby044. http://oup.prod.sis.lan/bib/article-pdf/20/5/1709/31555028/bby044.pdf
20. Hui, D., Fang, Z., Lin, J., Duan, Q., Li, Y., Hu, M., Chen, W.: LAIT: A local ancestry inference toolkit. BMC Genetics **18**(1), 1–5 (2017). doi:10.1186/s12863-017-0546-y
21. Garrison, N.: Genomic justice for native americans: Impact of the havasupai case on genetic research. Science, Technology, & Human Values **38**(2), 201–223 (2013)
22. TallBear, K.: Native American DNA: Tribal Belonging and the False Promise of Genetic Science. University of Minnesota Press, ??? (2013)
23. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., Mardis, E.R., Marth, G.T., McVean, G.A., Nickerson, D.A., Schmidt, J.P., Sherry, S.T., Wang, J., Wilson, R.K., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y.Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Gupta, N., Gharani, N., Toji, L.H.,

Gerry, N.P., Resch, A.M., Barker, J., Clarke, L., Gil, L., Hunt, S.E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W.M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R.E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Sudbrak, R., Albrecht, M.W., Amstislavskiy, V.S., Borodina, T.A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.L., Fulton, L., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T.M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Davies, C.J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Campbell, C.L., Kong, Y., Marcketta, A., Yu, F., Antunes, L., Bainbridge, M., Sabo, A., Huang, Z., Coin, L.J.M.M., Fang, L., Li, Q., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Garrison, E.P., Kural, D., Lee, W.P., Leong, W.F., Stromberg, M., Ward, A.N., Wu, J., Zhang, M., Daly, M.J., DePristo, M.A., Handsaker, R.E., Banks, E., Bhatia, G., Del Angel, G., Genovese, G., Li, H., Kashin, S., McCarroll, S.A., Nemesh, J.C., Poplin, R.E., Yoon, S.C., Lihm, J., Makarov, V., Gottipati, S., Keinan, A., Rodriguez-Flores, J.L., Rausch, T., Fritz, M.H., Stütz, A.M., Beal, K., Datta, A., Herrero, J., Ritchie, G.R.S.S., Zerbino, D., Sabeti, P.C., Shlyakhter, I., Schaffner, S.F., Vitti, J., Cooper, D.N., Ball, E.V., Stenson, P.D., Barnes, B., Bauer, M., Cheetham, R.K., Cox, A., Eberle, M., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E.E., Batzer, M.A., Konkel, M.K., Walker, J.A., MacArthur, D.G., Lek, M., Herwig, R., Ding, L., Koboldt, D.C., Larson, D., Ye, K.K., Gravel, S., Swaroop, A., Chew, E., Lappalainen, T., Erlich, Y., Gymrek, M., Willems, T.F., Simpson, J.T., Shriver, M.D., Rosenfeld, J.A., Bustamante, C.D., Montgomery, S.B., De La Vega, F.M., Byrnes, J.K., Carroll, A.W., DeGorter, M.K., Lacroute, P., Maples, B.K., Martin, A.R., Moreno-Estrada, A., Shringarpure, S.S., Zakharia, F., Halperin, E., Baran, Y., Cerveira, E., Hwang, J., Malhotra, A., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Hyland, F.C.L.L., Craig, D.W., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A.A., Sinari, S.A., Squire, K., Xiao, C., Sebat, J., Antaki, D., Gujral, M., Noor, A., Ye, K.K., Burchard, E.G., Hernandez, R.D., Gignoux, C.R., Haussler, D., Katzman, S.J., Kent, W.J., Howie, B., Ruiz-Linares, A., Dermitzakis, E.T., Devine, S.E., Kang, H.M., Kidd, J.M., Blackwell, T., Caron, S., Chen, W., Emery, S., Fritsche, L., Fuchsberger, C., Jun, G., Li, B., Lyons, R., Scheller, C., Sidore, C., Song, S., Sliwerska, E., Taliun, D., Tan, A., Welch, R., Wing, M.K., Zhan, X., Awadalla, P., Hodgkinson, A., Li, Y.Y., Shi, X., Quitadamo, A., Lunter, G., Marchini, J.L., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Kretzschmar, W., Iqbal, Z., Mathieson, I., Menelaou, A., Rimmer, A., Xifara, D.K., Oleksyk, T.K., Fu, Y.Y., Liu, X., Xiong, M., Jorde, L., Witherspoon, D., Xing, J., Browning, B.L., Browning, S.R., Hormozdiari, F., Sudmant, P.H., Khurana, E., Tyler-Smith, C., Albers, C.A., Ayub, Q., Chen, Y., Colonna, V., Jostins, L., Walter, K., Xue, Y., Gerstein, M.B., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Y.Y., Harmanci, A.O., Jin, M., Lee, D., Liu, J., Mu, X.J., Zhang, J., Zhang, Y.Y., Hartl, C., Shakir, K., Degenhardt, J., Meiers, S., Raeder, B., Casale, F.P., Stegle, O., Lameijer, E.W., Hall, I., Bafna, V., Michaelson, J., Gardner, E.J., Mills, R.E., Dayama, G., Chen, K., Fan, X., Chong, Z., Chen, T., Chaisson, M.J., Huddleston, J., Malig, M., Nelson, B.J., Parrish, N.F., Blackburne, B., Lindsay, S.J., Ning, Z., Zhang, Y.Y., Lam, H., Sisu, C., Challis, D., Evani, U.S., Lu, J., Nagaswamy, U., Yu, J., Li, W., Habegger, L., Yu, H., Cunningham, F., Dunham, I., Lage, K., Jespersen, J.B., Horn, H., Kim, D., Desalle, R., Narechania, A., Sayres, M.A.W., Mendez, F.L., Poznik, G.D., Underhill, P.A., Mittelman, D., Banerjee, R., Cerezo, M., Fitzgerald, T.W., Louzada, S., Massaia, A., Yang, F., Kalra, D., Hale, W., Dan, X., Barnes, K.C., Beiswanger, C., Cai, H., Cao, H., Henn, B., Jones, D., Kaye, J.S., Kent, A., Kerasidou, A., Mathias, R.A., Ossorio, P.N., Parker, M., Rotimi, C.N., Royal, C.D., Sandoval, K., Su, Y., Tian, Z., Tishkoff, S., Via, M., Wang, Y., Yang, H., Yang, L., Zhu, J., Bodmer, W., Bedoya, G., Cai, Z., Gao, Y., Chu, J., Peltonen, L., Garcia-Montero, A., Orfao, A., Dutil, J., Martinez-Cruzado, J.C., Mathias, R.A., Hennis, A., Watson, H., McKenzie, C., Qadri, F., LaRocque, R., Deng, X., Asogun, D., Folarin, O., Happi, C., Omoniwa, O., Stremlau, M., Tariyal, R., Jallow, M., Joof, F.S., Corrah, T., Rockett, K., Kwiatkowski, D., Kooner, J., Hien, T.T., Dunstan, S.J., ThuyHang, N., Fonnie, R., Garry, R., Kanneh, L., Moses, L., Schieffelin, J., Grant, D.S., Gallo, C., Poletti, G., Saleheen, D., Rasheed, A., Brooks, L.D., Felsenfeld, A.L., McEwen, J.E., Vaydylevich, Y., Duncanson, A., Dunn, M., Schloss, J.A.: A global reference for human genetic variation. Nature **526**(7571), 68–74 (2015). doi:10.1038/nature15393. 15334406

24. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., J., L.J.: Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience **4** (2015). doi:10.1186/s13742-015-0047-8

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 3WAYTukeyFigureS3.pdf
- 3WAYglobalancestryFigureS6.pdf
- AFAglobalancestryFigureS4.pdf
- AFATukeyFigureS1.pdf
- SoftwareaccuracyFigure6.pdf
- Methodsflowchartfigure1.pdf
- MemoryvnFigure3.pdf
- MemoryvnancestriesFigure4.pdf
- HisTukeyFigureS2.pdf
- RuntimevnancestriesFigure5.pdf
- AncestryvPC1Figure7.pdf
- HlSglobalancestryFigureS5.pdf
- RuntimevnFigure2.pdf