Decomposition of Individual SNP Patterns From Mixed DNA Samples

Gabriel Azhari
Bar Ilan University
Shamam Waldman
The Hebrew University of Jerusalem
Netanel Ofer
Bar Ilan University
Yosi Keller
Bar Ilan University
Shai Carmi
The Hebrew University of Jerusalem
Gur Yaari (gur.yaari@biu.ac.il)
Bar Ilan University

Research Article

Keywords: Single Nucleotide Polymorphism (SNPs), SNP, simplistic approaches

Posted Date: August 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-654059/v1

Decomposition of Individual SNP Patterns from Mixed DNA Samples

Gabriel Azhari[1], Shamam Waldman[2], Netanel Ofer[1], Yosi Keller[1], Shai Carmi[2], and Gur Yaari[1,*]

[1]Faculty of Engineering, Bar Ilan University, Ramat Gan, 5290002, Israel
[2]Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel
[*]Corresponding author: gur.yaari@biu.ac.il

ABSTRACT

Background: Single Nucleotide Polymorphism (SNPs) markers have great potential to identify individuals, family relations, biogeographical ancestry, and phenotypic traits. In many forensic situations, DNA mixtures of a victim and an unknown suspect exist. Extracting SNP profiles from suspect's samples can be used to assist investigation or gather intelligence. Computational tools to determine inclusion/exclusion of a known individual from a mixture exist, but no algorithm for extraction of an unknown SNP profile without a list of suspects is available.

Results: We present here AH-HA, a novel computational approach for extracting an unknown SNP profile from whole genome sequencing (WGS) of a two persons mixture. AH-HA utilizes techniques similar to the ones used in haplotype phasing. It constructs the inferred genotype as an imperfect mosaic of haplotypes from a reference panel of the target population. It outperforms more simplistic approaches, maintaining high performance through a wide range of sequencing depths (500x - 5x).

Conclusions: AH-HA can be applied in cases of victim-suspect mixtures and improve the capabilities of the investigating forces. This approach can be extended to more complex mixtures with more donors and less prior information, further motivating the development of SNP-based forensics technologies.

Background

Many studies in the field of forensic science have shown the benefits of using Single Nucleotide Polymorphism (SNP) data in DNA investigation and intelligence[1–4]. In the last 30 years, standard forensic DNA-based methods have been built on Short Tandem Repeats (STR). Such STR databases exist in most countries and direct comparison between DNA samples and these databases is admissible in court for human identification purposes. However, STRs do not store valuable information regarding forensically relevant traits such as ancestry and phenotype inference. The advances in DNA high throughput sequencing (HTS) have improved the study of SNPs and their potential uses for forensics purposes[5]. The ability to use millions of available SNPs, as opposed to a limited number of STR sites (13-26), opens the door for new forensics applications. Forensically relevant SNPs have been categorized into four groups of potential uses: (1) Individual identification, (2) Kinship search, (3) Biogeographical Ancestry, and (4) External Visible Characteristics[6]. Until recently, the main SNP genotyping technique for forensics has been a customized SNP array[7]. Such assays are available for a wide range of markers, most of them combining STRs and SNPs in the same kit[8]. These assays, despite their lower costs, are limited to the SNP positions they were designed for. Alternatively, as the cost of sequencing decreases, whole genome sequencing (WGS) of the sample can be conducted and then analyzed for all relevant markers.

The two primary challenges in forensics DNA are highly degraded samples and mixed samples (containing two or more individuals)[9]. When addressing degraded samples, SNPs have an advantage over STRs, as they require shorter amplicon lengths, and can overcome missing sections as there are many sites that are spread across the whole genome. Regarding mixture analysis, many studies have focused on the individualization problem, i.e., inferring the presence or absence of a known individual (POI - Person Of Interest) from a mixed DNA sample. STR based methods, such as STRmix[10], LRmix Studio[11] and EurForMix[12], are accurate only for 2-3 person mixtures. Gill et al.[13] and Bleka et al.[14] show how these STR packages, based on the widely used LR method, can be adapted to a SNP scenario. They are still limited to 2-3 person cases and lose accuracy when the contribution from the POI is low in uneven mix ratios. Other SNP based methods achieve good results even with complex mixtures of three or more contributors and various mixture ratios, outperforming STR based methods. These algorithms rely on deep sequencing, specific minor allele frequency (MAF)[15] and uneven mix ratios[16,17]. It should be noted that since most SNPs are typed as bi-allelic, it is harder to recognize the presence of a mix in a "SNP only" profile. STRs, on the other hand, are multi-allelic, enabling easier recognition of the presence of more than one donor. To confront the shortcomings of bi-allelic SNPs, Kidd et al.[18] have offered the use of a new marker type, Microhaplotype, a region with two or more SNPs that occur

within the length of an HTS read, effectively creating a pseudo "multi-allelic" marker. This "long-read" based approach was applied in Voskoboinik et al.[19].

An important problem that has been overlooked until now is a *de-novo* reconstruction of an unknown SNP profile from a DNA mixture. To address this limitation we introduce here AH-HA, a novel approach to infer an unknown SNP profile from a DNA mixture of two individuals. Our method receives as an input WGS data of a two person mixture, in which one genotype is known ("victim") and the other is unknown ("suspect"). AH-HA is compared with other computational approaches over varying sequencing depths. It is designed to cope with low coverage ($\sim$ 5x) and missing reads by incorporating an ancestral based coalescence model[20], similar to the one used in haplotype phasing[21] and imputation[22].

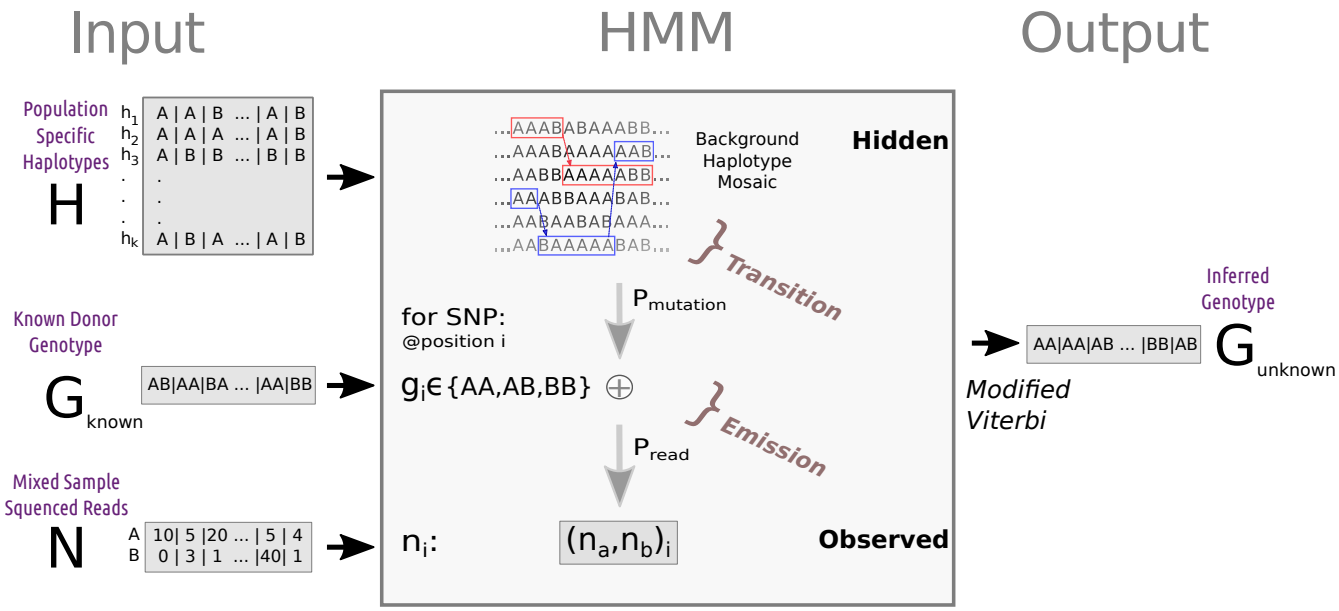

Figure 1. An outline of AH-HA. A mixed sequenced sample (N) is processed together with a reference panel (H) and a known donor genotype (G_{known}) in a Hidden Markov Model (HMM). The output is an inferred genotype of the unknown donor.

AH-Ha is designed to infer an unknown genotype from whole genome sequencing (WGS) data of a two person mixed sample. The problem is defined under the following assumptions:

1. The number of individuals in the mixed sample is known to be two.
2. The ethnic origin of the unknown person could be determined by a preliminary step.
3. The inference is designed for biallelic SNPs. Having the value of reference allele (tagged as 'A') or alternate allele (tagged as 'B') in respect to the GRCh37 build, also called "REF" or "ALT" respectively in this paper.

Model extensions that relax these assumptions are discussed below. The general flow of the presented approach is illustrated in Figure 1. Three types of input are required: 1) A REF-ALT allele count table for each SNP position (Figure 1N). This table is produced from WGS data after an alignment step. 2) A reference haploid dataset generated from a sample of the ethnic group of the unknown individual (Figure 1H). Phased cohort data for various populations can be obtained either by direct download from resources such as the 1000 Genomes Project[23], or by applying a computational tool (e.g., shapIT[24]) to a collection of WGS data samples from the target population. 3) A credible genotype of the known individual in the mixed sample (Figure 1G). These inputs enter AH-HA: a coalescence based HMM. The output of the model is an inferred genotype for the unknown individual in each SNP. Source code is available at `https://bitbucket.org/yaarilab/ah-ha/src/master/`.

Results

Performance Evaluation

Algorithm performance

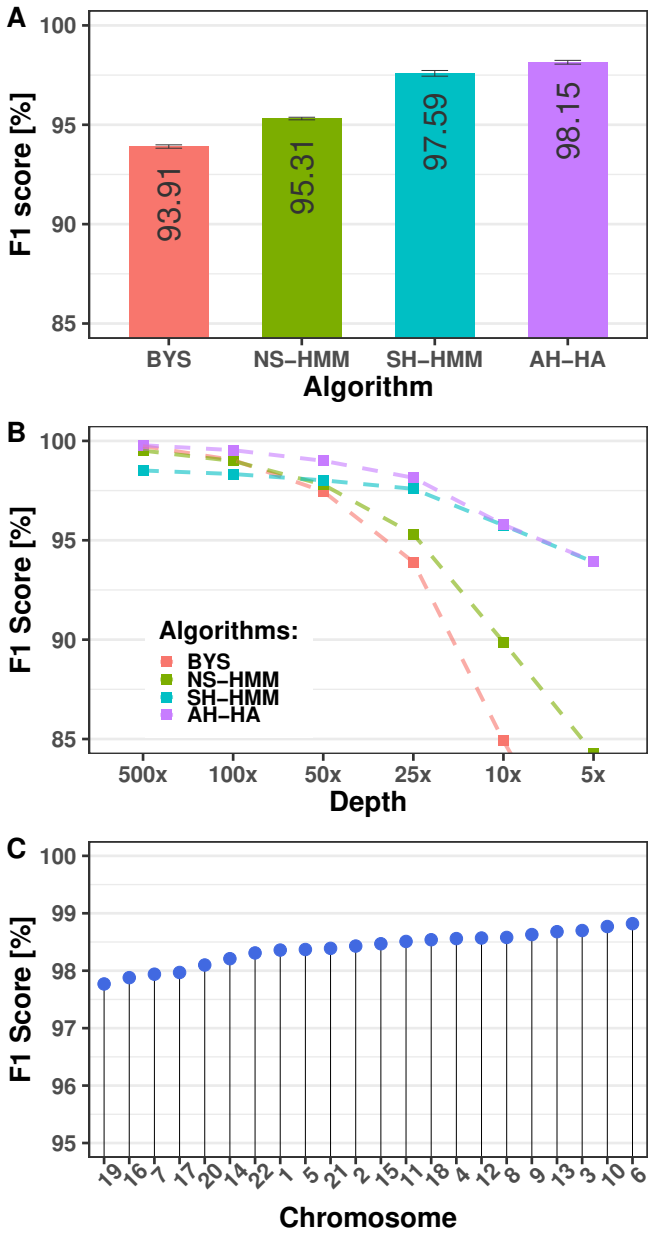

Figure 2. Performance comparison. F1 scores (Y axis) are shown (A) for the benchmark case for each algorithm as indicated by the X axis. Error bars represent the total variation between 10 random mixtures. (B) for varying sequencing depths (X axis) for the four algorithms (different colors as indicated in the legends), and (C) for AH-HA for all autosomal chromosomes (X axis). Chromosomes are ordered by their F1 score value.

To evaluate the performance of the different algorithms, the resulting inferred genotypes of these algorithms were summarized into an evaluation table. Table 2 shows the summarized results of AH-HA for the benchmark scenario: a 1:1 AJ-AJ mixed sample, Chromosome 22, with 25x coverage and a 240 haplotype reference panel. The two genotypes (known and unknown) were divided into 9 scenarios, from which an F1 score was calculated (see section F1 Score Calculation). A number of key notes from this table:

a. The AA-AA case accounts for 66.4% of all SNPs. This number is likely to remain high for all mixed samples. This case (together with the BB-BB case) is the most simple case for all algorithms, and hence the absolute performance results of all algorithms are generally high.

b. The unknown genotype is AA in 74.4% of all SNPs. A trivial algorithm that always outputs AA, gives a good concordance score (74.4%), but an F1 score of 0 %. That is the main motivation behind using an F1 score, instead of simple concordance.

c. Two other dummy predictors: 1. Inferred genotype equals the majority allele of the reference panel for each SNP. This predictor yields an F1 score of 52.87% for the benchmark case. 2. Inferred genotype equals the known genotype in each SNP. This predictor yields an F1 score of 72.31% for the benchmark case.

AH-HA was compared to three other algorithms: 1) A per-SNP bayesian algorithm (BYS), 2) A population based Next SNP HMM (NS-HMM), and 3) A simple haplotype based HMM (SH-HMM). See section Algorithms for more details. Ten different realizations of the benchmark mixture were generated. These realizations differ both in the reads that enter each mixture and in the order in which the reference panel is organized (this order affects SH-HMM and AH-HA). All four algorithms were applied to the ten realizations. Figure 2A shows the average results of the four algorithms applied to these benchmark cases. There is a clear order that ranks the four algorithms, where SH-HMM and AH-HA outperform the two other algorithms.

Figure 2B shows the performance of the four algorithms for different mixture coverage values. AH-HA outperforms all other algorithms, for all coverage values (500x-5x). BYS shows good performance in high coverage scenarios (>100x), but as coverage decreases the per SNP prediction becomes less reliable. Compared to BYS, NS-HMM manages to slightly improve the performance in lower coverage cases. This is due to the incorporation of allele statistics along the chromosome. However, also for NS-HMM, as coverage decreases, the performance significantly drops.

SH-HMM is outperformed by the above algorithms for high coverage scenarios (>100x), but for low coverage it utilizes the reference panel to attain better performance than BYS and NS-HMM. AH-HA combines two advantages: for low coverage it utilizes the coalescence model through the reference panel, and for high coverage it relies more on the observed allele count by its back-tracking phase. These features of AH-HA make it superior to all other algorithms considered here. It should be noted that forensic samples suffer from degradation. Therefore, they will have low coverage and high error rates, emphasizing the need for a reliable performance in these conditions specifically. To test the robustness of AH-HA, we calculated its performance on the rest of the autosomal chromosomes (Figure 2C). The algorithm shows consistent performance across all chromosomes (10,355,283 total SNPs). The score for chromosome 22 is close to the average, validating the benchmark mixture as a good indicator for a WGS case.

Computational runtimes

SH-HMM and AH-HA have relatively high computation cost; in memory and in running time. In these algorithms, the Viterbi solver runs over all possible combinations for every SNP in the "Forward" stage. It utilizes a "two-dimensional" hidden state matrix which scales quadratically with panel size. I.e, assuming J haplotypes are used in the panel (240 for the standard case), with L total SNPs, the running time scales like:

$$T_{runtime} \propto J \times J \times L \tag{1}$$

Running AH-HA on all of chromosome 22 ($\sim 150K$ SNPs) on one core of our server, Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz (256GB memory), takes about 100 hours. Even though emphasis is put on algorithm accuracy rather than efficiency, this runtime could be improved. In particular when considering BYS and NS-HMM that do the same task in about a minute on a home laptop (i5 processor, 8GB RAM). The first action that was taken is reducing the number of computations for each SNP. In position l for example, δ_l values from the Viterbi solver in the 2D state (h_i, h_j) are symmetrical. That is, $\delta_l(h_i, h_j) = \delta_l(h_j, h_i)$. Hence, it is sufficient to calculate only one of them (see also the Extended Mathematical Description section and tables S1 and S2 in the supplementary information).

The second action that was taken is to use multi-threading. Whole chromosomes were split into non-overlapping segments of 5Mbp, same as the approach taken by IMPUTE2[22]. 5Mbp is considered large enough for the SNPs in different chunks to be statistically independent from a linkage disequilibrium stand point. In IMPUTE2, they use a 250bp buffer between chunks, to prevent edge conditions. In our case, these buffers included zero to three SNPs, and including them did not influence the

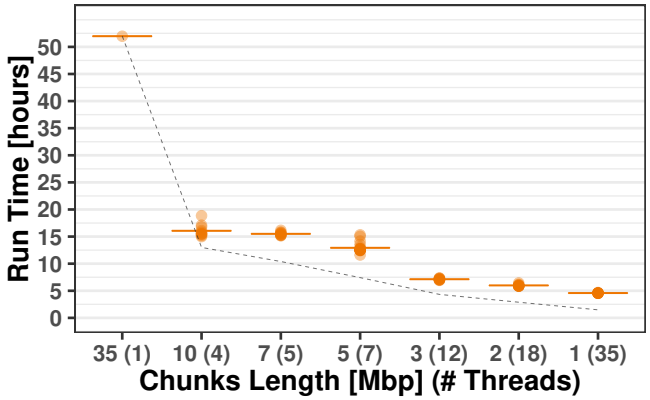

Figure 3. Algorithm runtimes. AH-HA runtimes are shown for different chunk sizes and thread count for the benchmark scenario. Average values (in hours) are indicated by a horizontal line. Theoretical, optimized runtime is shown by the dashed line. All runs were conducted on an Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz server with 48 cores and 256GB memory.

results. A comparison of run-times of AH-HA for the *benchmark* case for different chunk lengths and respectively the number of threads running in parallel is shown in Figure 3. The theoretical decrease in runtime is shown by the dashed line. The main reason for run-times to differ from the expected values is that in practice, the total runtime is determined by the segment with the largest number of SNPs. Since the chromosome is split based on genetic distance in bp, chunks have varying numbers of SNPs. The main thread waits for all threads to finish the run for every chunk before it can combine the results for the full chromosome, making the longest chunk length the bottle-neck of our process. The default chunk length used, 5Mbp, splits the ∼150K SNPs spread over 35Mbp in chromosome 22 into 7 chunks. It manages to cut algorithm runtime from about 50 hours to around 13 hours. Scoring was slightly affected by splitting. About a 0.5% decrease in F1 score was seen between single thread performance and multi-thread performance. Within the multi-thread runs, performance has minor differences, varying within 0.25%, where the 5Mbp chunk-split has the highest average score (Supplementary figure S1).

Model configuration

HMM parameters

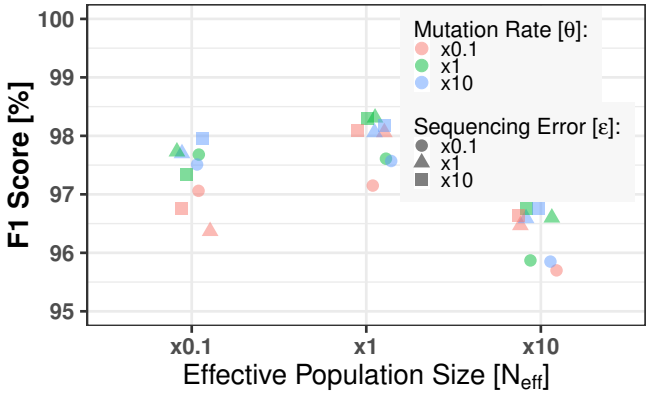

Figure 4. Parameter selection effect. F1 scores (Y axis) are shown for different AH-HA parameters for the benchmark scenario. The indicated fold change values refer to the fitted values of the parameters as described in the methods. N_{eff} values follow the X axis labels, θ is indicated by color, and ε is indicated by symbol shapes.

Figure 4 shows AH-HA's sensitivity to different model parameters. Since N_{eff}, θ and ε are estimated, the goal was to test: a) if these are optimal values, and b) the effect each parameter has on the performance. F1 scores of AH-HA running the benchmark case were calculated with parameters in different orders of magnitude relative to the optimal estimated values (x10, x1, and x0.1). All combinations of (N_{eff}, θ, ε) values were tested. The initial estimation (x1) indeed has the best F1 score. Larger θ and ε values in the model do not seem to negatively affect the F1 score so much.This can be attributed to the fact that

higher values in these parameters correspond to higher probabilities of changes from the observed data. Another observation from Figure 4 is that even when the parameters are estimated wrongly, AH-HA remains robust in terms of F1 score.

Reference panel

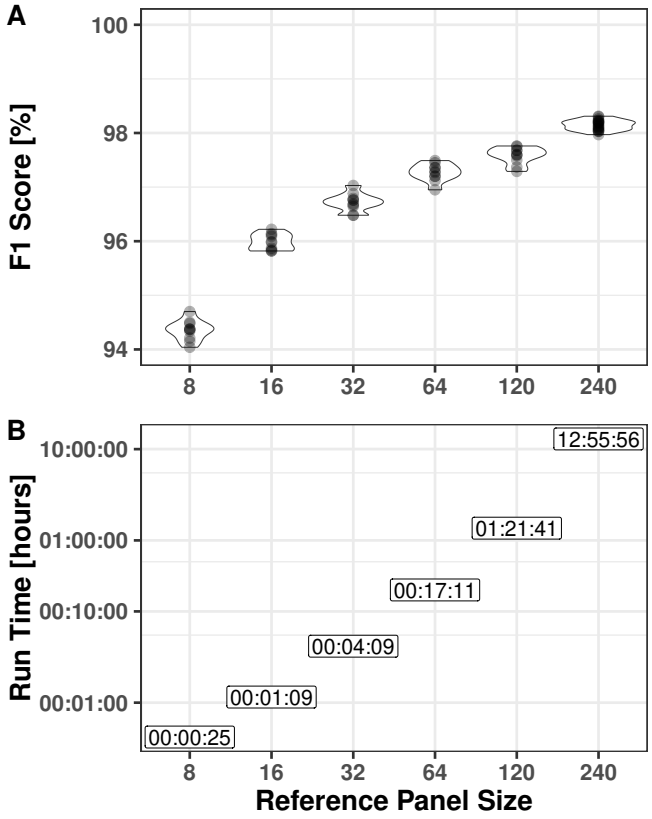

Figure 5. Panel size effect. F1 scores (Y axis in A) and mean runtimes (Y axis in B) are shown for varying reference panel sizes (X axis) for the AJ-AJ mixture. The different panels were created by sampling haplotypes from the AJ-panel**??**.

Another important factor affecting AH-HA's performance is the reference panel size. The *benchmark* scenario was run with varying panel sizes: 8, 16, 32, 64, 120, and 240 haplotypes. For each panel size, 10 realizations of J randomly chosen haplotypes from the 240 haplotypes in the AJ-panel were used. For the full panel (240), haplotype order was randomized between the 10 realizations. The results in Figure 5, indicate how the choice of panel size affects the accuracy (5A) and runtime (5B). There is a clear trade-off between the two. Running time seems to scale quadratically with panel size, while the F1 score seems to moderately increase with panel size. For example, decreasing the panel size from 240 to 120 (50%), cuts runtime by 89.5%, while F1 score decreases only by 0.6%. The decrease from 240 to 8 (97%), lowers the runtime by 99.9% but hurts the F1 score in just 3.8%. This could be a good trade-off for when run-times and memory are a consideration.

Mixed population

We further tested AH-HA for different populations, to evaluate the effect on the algorithm's performance when a mixture of donors of different origins is used, i.e. higher genetic diversity between individuals. Using data from the 1000Genomes project, a mixture of a YRI individual (NA18489) and AJ_{father} was generated. Using a YRI-panel (constructed from these data as well, see section Data Processing), relevant values of N_{eff} and θ were calculated with CHROMOPAINTER. The AJ-YRI mixture was processed by AH-HA, once with AJ_{father} as the unknown individual and once with NA18489 as the unknown. Each scenario was run over 10 realizations of the mixture with randomly ordered haplotype panels. Similarly, an AJ-AJ mixture of the same coverage with AJ_{father} as unknown was run by AH-HA. Figure 6 shows the scores for these three scenarios. AH-HA performs well for the AJ-YRI mixture in both cases of known-unknown selection and it is comparable to the AJ-AJ case, with a minor decrease in the average score.

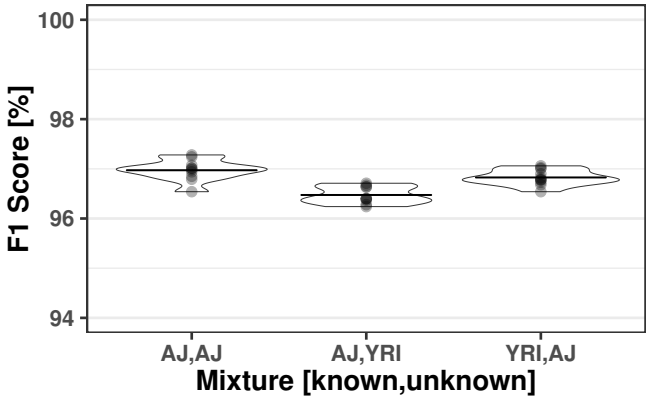

Figure 6. YRI-AJ Mixtures. F1 scores (Y axis) are shown for different population mixtures (X axis). Performance was evaluated on chromosome 20 with coverage 18x. Mean value for each mixture is indicated by a horizontal line.

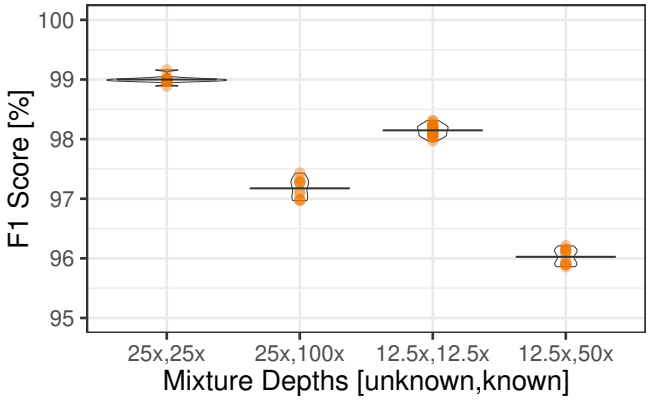

Figure 7. Uneven Mixture. F1 scores (Y axis) are shown for 1:1 and 1:4 mixture ratios (X axis). Mean value of the F1 score in each mixture is indicated by a horizontal line.

Uneven Mixtures

By modifying the calculation of the emission stage, AH-HA was adapted to analyze an uneven mixture of 20%-80% (1:4 ratio) for the unknown and known individuals respectively. The new proportion was weighed into P_{read}. Figure 7 shows a comparison of the results between a 1:1 mixture and a 1:4 mixture, for two coverage depths of the unknown contributor (25x and 12.5x). The results were generated from 10 randomly subsampled mixtures and 10 randomly ordered reference haplotypes. The uneven mixtures achieve lower F1 scores, about 1.8% below the 1:1 mixture (*benchmark* example), regardless of total coverage depths.

Discussion

In this paper we have introduced AH-HA, an approach to infer a SNP profile of an unknown individual from an HTS of a mixed sample. It outperforms other methods over varying coverage rates. In particular, the performance for low coverage is superior compared to more naïve algorithms. The robustness of AH-HA was shown over all chromosomes, mixed population cases and different hyper-parameters. AH-HA's runtime and memory was improved by utilizing multi-thread parallelization, splitting the chromosome into chunks and processing them simultaneously. The size of the reference panel strongly affects the runtime but has a minor impact on model's performance. Consequently AH-HA performs better than naïve methods while maintaining comparable run-times.

Other algorithms dealing with SNPs in forensic DNA mixtures, focus only on individual identification in a complex mixtures[9], mainly for inclusion or exclusion of a specific known individual from a mixture. While these other algorithms focus on a customized SNP panel of several thousand markers and require high coverage, AH-HA works on hundred of thousands of SNPs coming from "off-the-shelf" sequencers with moderate to low coverage (5x). Also, these algorithms require certain MAFs in their data and uneven mixture ratios to perform, while in our scenario all of the called SNPs were used.

AH-HA requires prior knowledge about the ethnic origin of the suspect and an accurate genotype of the victim. Available forensic and investigative methods can attain these preconditions[25,26]. Currently AH-HA handles only bi-allelic SNPs as this is the majority of SNPs It can be extended to multi-allelic SNPs the HMM's hidden states should change to cover all four allele values. For uneven mixture ratios, assuming the ratio is deduced beforehand with another technique, the emission probabilities can be adjusted for the new ratios. As shown in section Uneven Mixtures for a 1:4 case. AH-HA can confront noisy reads by changing ε. In case of an ad-mixed unknown individual, there would be a use of two reference panels for inferring each haplotype separately. These panels could be combined into a joint panel, and AH-HA can be modified to weight differently haplotypes that come from the different populations. Also, instead of the Viterbi solver that gives only the "best" path of the HMM, a softer solving method can be used. As discussed in Rabiner et al.[27], combining probabilities from the forward and backward stages the probability for each state value can be calculated, enabling a soft decision over all state values (per SNP) and even a "confidence" measure.

In mixtures with more than two individuals, when only one of them is unknown, AH-HA will require all known genotypes, and the emission step should be adapted accordingly. Extending AH-HA to infer more than one unknown individual is a greater challenge for the currently used HMM. First of all, computation wise, the algorithm will need to process, for every SNP, all combinations of haplotype pairs for every unknown individual and the combinations between them. This will increase computation cost by two polynomial orders for each additional unknown individual. E.g., if we have two unknown individuals to infer from the mixture instead of one, AH-HA will effectively be calculating a 4D hidden state matrix (2x2) instead of a 2D matrix. In the case of all haplotypes coming from the same reference panel with J haplotypes, this would mean calculating and saving J^4 hidden state probabilities for each SNP. The second challenge is the possible switch errors between individual haplotype pairs. In this case the algorithm infers four haplotypes, without additional information on the target individuals it could be hard to assign these haplotypes into two genotypes correctly.

Further extensions of AH-HA can be done by exploring new implementation methods. First, incorporating "read-based" inference to the algorithm. This approach has the ability to accurately "stitch" SNPs from the same haplotype by overlapping read sequences (containing two or more SNPs)[28]. This will result in a better haplotype estimation for closely positioned SNPs, improving genotype inference. Second, keep the HMM model, but solve it using the Markov chain Monte Carlo (MCMC) algorithm, similar to the method used by SHAPEIT[24]. This has the potential to improve run-time and memory, but maintain accuracy. Another approach for improving run-times, could be to scale up from looking at per SNP states into using "representative" haplotype chunks as state values, similar to the method used in BEAGLE[29]. Another possible path forward is to use a machine learning approach to optimize the haplotype-chunk selection step, by learning from population specific data how to better label and categorize and even connect together inferred haplotype chunks.

AH-HA is built on the principles of Li and Stephans model[20]. Their model has revolutionized the methods used for phasing, imputation and ancestry studies, but not in the context of DNA mixture analysis. AH-HA opens the doors for future studies in DNA mixture analysis, which will develop as more and more HTS elements are being used for forensic work.

Methods

Data Sets

AJ reference panel (AJ-Panel). Data set from Carmi et al.[30] was used as the AJ reference panel, containing 128 sequenced individuals of Ashkenazi-Jewish ancestry (TAGC128). This panel was constructed from high coverage sequencing data (>50x) that were filtered, processed and phased as described therein (supplementary note 2). The first 120 members of TAGC128 (in a ".hap" IMPUTE2[22] format created by SHAPEIT[24]) were selected to form a 240 haplotype panel used throughout the paper.

Two deeply sequenced individuals (NA24143, NA24149). Two deeply sequenced individuals of Ashkenazi-Jewish ancestry, were taken from a mother-father-offspring trio created by the genome in a bottle project for genetic research[31]. For this paper, we took the mother (AJ_{Mother}) and the father (AJ_{Father}) samples, that are deeply covered (∼275x) and accurately genotyped. The AJ_{Mother} and AJ_{Father} were used as the "known" and "unknown" contributors, respectively, to the mixtures.

YRI samples. Samples of individuals from the Yoruba in Ibadan, Nigeria (tagged as YRI) were taken from the 1000Genomes database[23]. The NA18489 sample, that is relatively deep sequenced (∼8.7x), were used here for mixtures. From the remaining 106 YRI samples (NA18504 was also singled out, but not used), phased by SHAPEIT, a 212 haplotype panel was created. SNPs with more than two polymorphisms (non-biallelic SNPs) were filtered out.

Data Processing

AJ-AJ mixture. Synthetic mixtures of AJ_{Mother} and AJ_{Father} were generated, once for chromosome 22 and once for all autosomal chromosomes. Samtools' mpileup function and standard linux command line tools were applied to the AJ_{Mother} and AJ_{Father} .sam files to generate a table of nucleotide read composition for every position along the chromosome. SNPs that were not observed both in the AJ-trio and the TAGC128 panel were filtered out. Mixtures of 500x, 100x, 50x, 25x, 10x, and 5x coverage were created by randomly subsampling the high coverage reads (∼275x) to the desired coverage for each

G^{known}	G^{mix}			$\hat{p}_{\mathrm{A}}$		
AA	AAAA	AAAB	AABB	1	0.75	0.5
AB	AAAB	AABB	ABBB	0.75	0.5	0.25
BB	AABB	ABBB	BBBB	0.5	0.75	0

Table 1. Genotyping probability table. Binomial parameter for each genotype scenario, given the knwon genotype.

sample independently, and then combined to form a 1:1 mixed data set. For the shallow coverage mixtures null counts in certain positions may occur.

This method ensured a realistic sequencing error profile, that is known to vary between technologies and between REF and ALT allele reads[32]. The *"benchmark"* mixture used throughout most of the study is a 1:1 AJ-AJ mixture of chromosome 22, with a 25x coverage.

AJ-YRI mixture. Using the same concept described above, an AJ-YRI mixture of the AJ_{Father} and the NA18489 (YRI) sample was generated for chromosome 20. SNPs that were not observed in the AJ-trio, the TAGC128 panel, and the YRI panel were filtered out. Mixtures were created by sub-sampling AJ_{Father} reads by a 0.034 rate, giving a similar coverage to NA18489, and combining the two samples together to achieve a 1:1 mixture with an average coverage of $\sim$17.4x.

Legend files. The HMM algorithms rely on the distance between SNPs (in cM) for their recombination probability calculations. For every chromosome, a common ".legend" file was created containing distances between SNPs in cM as calculated from the human reference genome (Genome Reference Consortium human genome build 37, GRCh37.map) using a perl script.

Algorithms

Per SNP Bayesian Model (BYS). A "per SNP" bayesian approach for the unknown-donor genotype estimation was used. Briefly, a conjugate pair of a **Beta** prior with a binomial likelihood was used, resulting in a posterior probability for each possible genotype in each position. The genotype that maximizes this probability is the inferred genotype.

In more detail, for each SNP the unknown diploid genotype ($G^{unknown}$) is inferred by subtracting the known diploid genotype (G^{known}) from the inferred tetraploid genotype of the mixture (G^{mix}). The fraction of REF alleles in G^{mix} (second column in Table 1) corresponds to $\hat{p}_{\mathrm{A}}$ (third column there), the probability of success in the binomial distribution generating the REF allele counts of the mixed sample.

The resulting posterior probability of p_{A} is a **Beta** probability density function with the hyper parameters $a_{post} = a_{prior} + \mathrm{n}_A$ and $b_{post} = b_{prior} + \mathrm{n}_B$, where a_{prior} and b_{prior} are the percentage of REF (A) and ALT (B) alleles in the reference panel at this SNP position respectively and $(\mathrm{n}_A, \mathrm{n}_B)$ are the allele count observed in the mix (equation (2)).

$$P(p_{\mathrm{A}}|\mathrm{n}_A,\mathrm{n}_B)_{\mathrm{Beta}} = \frac{P(\mathrm{n}_A,\mathrm{n}_B|p_{\mathrm{A}})_{\mathrm{Bin}} P(p_{\mathrm{A}})_{\mathrm{Beta}}}{P(\mathrm{n}_A,\mathrm{n}_B)} = \mathrm{Beta}_{a_{post},b_{post}}(p_{\mathrm{A}}) \tag{2}$$

By comparing the posterior probabilities for different models ($\hat{p}_{\mathrm{A}}$ values), we select the model that has the highest probability as the model of G^{mix} from which $G^{unknown}$ is inferred (equation (3)).

$$\hat{p}_{\mathrm{A}} = \underset{p^*}{\mathrm{argmax}}\{\mathrm{Beta}_{a_{post},b_{post}}(p^*)\}, \quad p^* \in \frac{\{1,0.75,0.5,0.25,0\}+\varepsilon}{1+2\cdot\varepsilon} \tag{3}$$

where ε represents alignment and amplification errors ("sequencing errors").

Next SNP based HMM (NS-HMM). In an attempt to confront low coverage samples, a simple HMM based on the genotypes in the AJ population and their "next SNP" transitions was considered. This model utilizes statistical connections between neighboring SNPs. For each SNP position there is a hidden state, with three possible values- AA, AB, BB. The transition probabilities are calculated by averaging transitions between consecutive SNPs in each individual genotype in the AJ panel. Lastly, the emission probabilities are calculated from the REF-ALT read count in each SNP. The inferred genotypes are calculated by applying the Viterbi algorithm[27] to the HMM. **Simple and Advanced Haplotype based HMM Algorithm (SH-HMM and AH-HA).** These two algorithms are based on the model introduced by Li and Stephens[20]. Their model is used in ancestry studies, haplotype phasing, and SNP imputation. It captures key features in genealogical processes, while remaining computationally tractable for large datasets. The model assumes that a chromosome is built as an imperfect mosaic of a set of fixed haplotypes (Figure 1). Originating from the approximate coalescent model and the linkage disequilibrium model, an HMM is built and then solved.

In the current study, a two-dimensional HMM is applied with the following components: *states* - haplotype pair from the reference panel, *transition probability* - the recombination rate, *observed data* - REF-ALT read count table, and *emission*

probability - experimental errors and mutation rate. These parameters can be fixed, or estimated and then optimized (see section Parameter Estimation).

Both algorithms presented here (SH-HMM and AH-HA) are based on the same HMM formulation, but differ in their solving approaches. SH-HMM utilizes a standard Viterbi algorithm. It infers the maximum likelihood pair of haplotypes constructed from chunks of the haplotype panel. This pair of haplotypes is the output of the algorithm and serves as the inferred genotype of the unknown individual in the mixed sample. This output ignores possible mutations in specific SNPs. I.e, the inferred genotype is constructed from the reference panel, without considering a possibility for mutations in certain SNPs.

In AH-HA, a post-processing step is added to the Viterbi back-track. In this step, for each SNP it chooses the most likely genotype from the *emission probabilities*, conditioned with the state-pair suggested by SH-HMM.

Parameter Estimation

Following the guidelines set by Li and Stephans at el.[20], the coalescence-based HMM was created with these key parameters:

- N_{eff}- represents the "Effective Population Size", impacting the probability for a recombination event that is analogous to a transition probability in HMM. We have used CHROMOPAINTER[33] and its built-in E-M functionality to optimize this parameter over the haploids in the relevant reference panel.

- θ- represents the mutation rate, impacting the probability of a mutation event that is analogous to an emission probability in HMM. Similar to N_{eff}, this parameter was also optimized by using CHROMOPAINTER over the relevant haplotype panel.

- ε- represents the base pair error rate, caused by amplification, alignment, and sequencing errors. In modern NGS technologies (ILLUMINA and CG) there is at least a 0.1% discordance rate[32]. Hence, this was the default value used in the calculations of genotype probability given reads value. This parameter adds to θ in determining the emission probability in HMM.

- *Reference panel*- derived as described above (section Data Processing). For assessing the effect on runtime and performance we have used different panel sizes and haploid orders, as shown in the results section.

Known	Unknown	AA [#]	AB [#]	BB [#]	Total [#]	Percentage [%]
AA	**AA**	**95057**	284	1	95342	66.42
AA	**AB**	73	**12068**	66	12207	8.5
AA	**BB**	8	210	**1744**	1962	1.37
AB	**AA**	**9120**	284	0	9404	6.55
AB	**AB**	190	**8803**	73	9066	6.32
AB	**BB**	0	249	**3524**	3773	2.63
BB	**AA**	**2034**	82	0	2116	1.47
BB	**AB**	47	**3774**	9	3830	2.67
BB	**BB**	3	39	**5798**	5840	4.07

Table 2. Evaluation table, for the 9 "tetra-ploid" cases. As inferred by AH-HA, running the "benchmark" scenario. In the right column is the percentage of each case from the total SNPs processed.

F1 Score Calculation

After attaining an inferred genotype, performance is assessed by dividing the results into 9 categories, covering all cases of $< known, unknown >$ combinations: $< \{AA, AB, BB\}, \{AA, AB, BB\} >$, as shown in Table 2. For the data analyzed here, in $\sim 70\%$ of the SNPs there is a trivial correct inference - AA. Thus, a simple concordance measure is not sufficient to assess the performance of the algorithm. A different approach is to view the problem as a detection problem where the goal is to detect ALT alleles correctly. A REF allele is labeled as "Negative" and a ALT allele as "Positive". Heterozygous cases are labeled as half negative-half positive. A confusion matrix, shown in Table 3, is used to calculate - True Negative (TN), False Negative (FN), True Positive (TP) and False Positive (FP) counts. From these measures Precision and Recall values are calculated. Where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$. Finally, from these values, an F1 score is calculated. The equation for F1 score is: $F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$

Truth \ Predicted	AA	AB	BB
AA	TN	½TN + ½FP	FP
AB	½TN + ½FN	½TN + ½TP	½FP + ½TP
BB	FN	½FN + ½TP	TP

Table 3. Genotyping confusion matrix. F1 scores are calculated by applying this confusion matrix on the results.

List of abbreviations

Single Nucleotide Polymorphism (SNP)
Whole Genome Sequencing (WGS)
Short Tandem Repeats (STR)
High Throughput Sequencing (HTS)
Person Of Interest (POI)
Minor Allele Frequency (MAF)
Hidden Markov Model (HMM)
True Negative (TN)
False Negative (FN)
True Positive (TP)
False Positive (FP)
Markov chain Monte Carlo (MCMC)

Declarations

Ethics approval and consent to participate

All data used in this study are public and do not require special approval.

Consent for publication

Not applicable.

Availability of data and materials

All data analyzed here were downloaded from public domains as indicated in the Methods section. Source code is available at https://bitbucket.org/yaarilab/ah-ha/src/master/.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

GY conceived the project and developed the approach. GA, SC, SW, and GA formulated the mode. GA and SW implemented the algorithms and analyzed the data. GY supervised the project. GY and GA wrote the paper. NO and YK contributed to data analysis. All authors edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Pazit Polak for helpful discussions and for commenting on the manuscript.

References

1. Gill, P. An assessment of the utility of single nucleotide polymorphisms (snps) for forensic purposes. *Int. journal legal medicine* **114**, 204–210 (2001).
2. Sobrino, B., Bríon, M. & Carracedo, A. Snps in forensic genetics: a review on snp typing methodologies. *Forensic science international* **154**, 181–194 (2005).

3. Butler, J. M., Coble, M. D. & Vallone, P. M. Strs vs. snps: thoughts on the future of forensic dna testing. *Forensic science, medicine, pathology* **3**, 200–205 (2007).

4. Butler, J. M. *et al.* Report on isfg snp panel discussion. *Forensic Sci. Int. Genet. Suppl. Ser.* **1**, 471–472 (2008).

5. Liu, Y.-Y. & Harbison, S. A review of bioinformatic methods for forensic dna analyses. *Forensic Sci. Int. Genet.* **33**, 117–128 (2018).

6. Budowle, B. & Van Daal, A. Forensically relevant snp classes. *Biotechniques* **44**, 603–610 (2008).

7. Daniel, R. *et al.* A snapshot of next generation sequencing for forensic snp analysis. *Forensic Sci. Int. Genet.* **14**, 50–60 (2015).

8. Jäger, A. C. *et al.* Developmental validation of the miseq fgx forensic genomics system for targeted next generation sequencing in forensic dna casework and database laboratories. *Forensic Sci. Int. Genet.* **28**, 52–70 (2017).

9. Yang, J. *et al.* The advances in dna mixture interpretation. *Forensic science international* (2019).

10. Buckleton, J. S. *et al.* The probabilistic genotyping software str mix: Utility and evidence for its validity. *J. forensic sciences* **64**, 393–405 (2019).

11. Haned, H., Slooten, K. & Gill, P. Exploratory data analysis for the interpretation of low template dna mixtures. *Forensic Sci. Int. Genet.* **6**, 762–774 (2012).

12. Bleka, Ø., Storvik, G. & Gill, P. Euroformix: an open source software based on a continuous model to evaluate str dna profiles from a mixture of contributors with artefacts. *Forensic Sci. Int. Genet.* **21**, 35–44 (2016).

13. Gill, P. *et al.* The open-source software lrmix can be used to analyse snp mixtures. *Forensic Sci. Int. Genet. Suppl. Ser.* **5**, e50–e51 (2015).

14. Bleka, Ø. *et al.* Open source software euroformix can be used to analyse complex snp mixtures. *Forensic Sci. Int. Genet.* **31**, 105–110 (2017).

15. Voskoboinik, L., Ayers, S. B., LeFebvre, A. K. & Darvasi, A. Snp-microarrays can accurately identify the presence of an individual in complex forensic dna mixtures. *Forensic Sci. Int. Genet.* **16**, 208–215 (2015).

16. Isaacson, J. *et al.* Robust detection of individual forensic profiles in dna mixtures. *Forensic Sci. Int. Genet.* **14**, 31–37 (2015).

17. Ricke, D. O. *et al.* The plateau method for forensic dna snp mixture deconvolution. *bioRxiv* 225805 (2017).

18. Kidd, K. *et al.* Microhaplotype loci are a powerful new type of forensic marker. *Forensic Sci. Int. Genet. Suppl. Ser.* **4**, e123–e124 (2013).

19. Voskoboinik, L., Motro, U. & Darvasi, A. Facilitating complex dna mixture interpretation by sequencing highly polymorphic haplotypes. *Forensic Sci. Int. Genet.* **35**, 136–140 (2018).

20. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).

21. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).

22. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genet.* **1**, 457–470 (2011).

23. Consortium, . G. P. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (2012).

24. Delaneau, O., Coulonges, C. & Zagury, J.-F. Shape-it: new rapid and accurate algorithm for haplotype inference. *BMC bioinformatics* **9**, 540 (2008).

25. Kidd, K. K. *et al.* Progress toward an efficient panel of snps for ancestry inference. *Forensic Sci. Int. Genet.* **10**, 23–32 (2014).

26. Wei, Y.-L. *et al.* A single-tube 27-plex snp assay for estimating individual ancestry and admixture from three continents. *Int. journal legal medicine* **130**, 27–37 (2016).

27. Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).

28. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. *The Am. J. Hum. Genet.* **93**, 687–696 (2013).

372 **29.** Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome
373 association studies by use of localized haplotype clustering. *The Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

374 **30.** Carmi, S. *et al.* Sequencing an ashkenazi reference panel supports population-targeted personal genomics and illuminates
375 jewish and european origins. *Nat. communications* **5**, 4835 (2014).

376 **31.** Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. data*
377 **3** (2016).

378 **32.** Wall, J. D. *et al.* Estimating genotype error rates from high-coverage next-generation sequence data. *Genome research* **24**,
379 1734–1739 (2014).

380 **33.** Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS*
381 *genetics* **8**, e1002453 (2012).