Nakatsuka et al. Genome Biology (2020) 21:199
https://doi.org/10.1186/s13059-020-02111-2

Genome Biology

SOFTWARE Open Access

# ContamLD: estimation of ancient nuclear DNA contamination using breakdown of linkage disequilibrium

Nathan Nakatsuka[1,2,3*†], Éadaoin Harney[1,3,4*†], Swapan Mallick[1,3], Matthew Mah[1,3], Nick Patterson[3] and David Reich[1,3,5,6*]

* Correspondence: nathan_nakatsuka@hms.harvard.edu; harney@g.harvard.edu; reich@genetics.med.harvard.edu
†Nathan Nakatsuka and Éadaoin Harney are the co-first authors.
[1]Department of Genetics, New Research Building, Harvard Medical School, 77 Ave. Louis Pasteur, Boston, MA 02115, USA
Full list of author information is available at the end of the article

## Abstract

We report a method called ContamLD for estimating autosomal ancient DNA (aDNA) contamination by measuring the breakdown of linkage disequilibrium in a sequenced individual due to the introduction of contaminant DNA. ContamLD leverages the idea that contaminants should have haplotypes uncorrelated to those of the studied individual. Using simulated data, we confirm that ContamLD accurately infers contamination rates with low standard errors: for example, less than 1.5% standard error in cases with less than 10% contamination and 500,000 sequences covering SNPs. This method is optimized for application to aDNA, taking advantage of characteristic aDNA damage patterns to provide calibrated contamination estimates, and is available at https://github.com/nathan-nakatsuka/ContamLD.

**Keywords:** Ancient DNA, Linkage disequilibrium, Contamination, Autosomal DNA, Nuclear DNA

## Background

Ancient DNA (aDNA) data has emerged as a powerful tool for learning about ancient population history, allowing the direct study of the genomes of individuals who lived thousands of years in the past [1–3]. Unfortunately, these inferences can be distorted by contamination during the excavation and storage of skeletal material, as well as the intensive processing required to extract the DNA and convert it into a form that can be sequenced.

Accurate measurement of the proportion of contamination in ancient DNA data is important, because it can provide guidance about whether the analysis should be restricted to sequences that show the characteristic pattern of C-to-T mismatch to the reference genome of authentic aDNA (if contamination is high) [4], or carried out at all. When the analysis is restricted to focus only on sequences showing evidence of characteristic ancient DNA damage, the substantial majority of authentic sequences

are usually removed from the analysis dataset, as only a fraction of genuinely ancient sequences carry characteristic damage. Another limitation of restricting to damaged sequences is that if a sample is contaminated by another individual with damaged DNA—which can arise for example as a result of cross-contamination from other specimens handled in the same ancient DNA laboratory—it is impossible to distinguish authentic sequences from contaminating ones based on the presence or absence of characteristic ancient DNA damage.

Current methods for estimating contamination have significant limitations. Methods based on testing for heterogeneity in mitochondrial DNA sequences (which are almost always homogeneous in an uncontaminated individual) can be biased, because there are several orders of magnitude of variation in the ratio of the mitochondrial to nuclear DNA copy number across samples. Thus, samples that have evidence of mitochondrial contamination can be nearly uncontaminated in their nuclear DNA, while samples that have no evidence of mitochondrial contamination can have high nuclear contamination [5]. A more consistently reliable set of methods for estimating rates of contamination in ancient DNA measures the rate of polymorphism on the X chromosome in males assuming there should be none; the most commonly used implementation of this idea is the *ANGSD* software, although there are also other software packages [6–10]. However, these methods do not work in females.

Several methods for estimating contamination rates in nuclear DNA from modern genomes have been published, including *ContEst* [11] and *ContaminationDetection* [12]. However, these methods generally rely on access to uncontaminated genotype data from the individual of interest or access to all possible contaminating individuals, neither of which is typically available for aDNA. Another method estimated modern human autosomal contamination in aDNA from archaic Denisovans [13] and Neanderthals [14] by producing maximum likelihood co-estimation of sequence error, contamination, and parameters correlated with divergence and heterozygosity. However, this method heavily relies on the high level of genetic divergence between archaic and modern humans. A similar method, *DICE*, expanded on this approach and jointly estimates contamination rate and error rate along with demographic history based on allele frequency correlation patterns [15]. However, this method requires both explicit demographic modeling and high genome coverage. While this may be effective for estimation of contamination in archaic genomes like Neanderthals and Denisovans that are highly genetically diverged from likely contaminant individuals, it is not optimized for the study of contamination among closely related present-day human groups or contamination from individuals of the same population. In Racimo et al. [15], *DICE* required over 3× genome sequence coverage and solved the distinctive problem of measuring contamination of present-day humans in a Neanderthal genome.

We report a method for estimating autosomal aDNA contamination using patterns of linkage disequilibrium (LD) within a sample. This approach, implemented in our software *ContamLD*, is based on the idea that when sequences from one or more contaminating individuals are present in a sample, LD among sequences derived from that sample is expected to be diminished, because the contaminant DNA derives from different haplotypes and therefore should have no LD with the authentic DNA of the ancient individual of interest. Thus, the goal of the algorithm is to determine the LD pattern the ancient individual would have had without contamination and compare it

to the LD pattern found in the sample. The LD patterns of ancient individuals are determined using reference panels from 1000 Genomes Project populations to compute approximate background haplotype frequencies, where haplotypes are defined as pairs of SNPs with high correlation to each other. Contamination is then estimated by fitting a maximum likelihood model of a mixture of haplotypes from an uncontaminated individual and a proportion of contamination (to be estimated from the data) from an unrelated individual. *ContamLD* corrects for mismatch of the ancestry of the ancient individual with the reference panels using two different user-specified options. In the first option, the mismatch is corrected using estimates from damaged sequences (which we assume lack present-day contaminants). In the second option, *ContamLD* performs an "external" correction by subtracting the sample's contamination estimate from estimates for individuals of the same population believed to have negligible contamination (the user could obtain this value from a *ContamLD* calculation on a male individual with a very low estimate of contamination based on *ANGSD*). The second option has more power than the first option and allows detection of cross-contamination by other ancient samples, but it could be biased if a reliable estimate from an uncontaminated individual from the same population is not available for the external correction.

We show that *ContamLD* accurately infers contamination in both ancient and present-day individuals of widely divergent ancestries with simulated contamination coming from individuals of different ancestries. The contamination estimates are highly correlated with estimates based on X chromosome analysis in ancient samples that are male, as assessed using *ANGSD* [16]. *ContamLD* run with the first option has standard errors less than 1.5% in samples with at least 500,000 sequences covering SNPs (~ 0.5× coverage for data produced by in-solution enrichment for ~ 1.2 million SNPs [2, 17] or ~ 0.1× coverage for data produced using whole-genome shotgun sequences). With the second option, *ContamLD* has standard errors less than 0.5% in these situations, allowing users to detect samples with 5% or more contamination with high confidence so they can be removed from subsequent analyses.

## Results

### Simulations of contamination in present-day individuals

To test the performance of *ContamLD*, we simulated sequence level genetic data. For our first simulations, each uncontaminated individual was simulated based on genotype calls from a present-day individual from the 1000 Genomes Project dataset. To determine the sequence coverage at each site, we used data from an ancient individual for which we had data at 1.02× coverage and in each case generated the same number of simulated sequences at each site, with the allele drawn from the present-day individual (e.g., if the present-day individual is homozygous for the reference allele at a site, all simulated alleles are of the reference type, while if the present-day individual is heterozygous, simulated alleles are either of the reference or alternative variant, with 50% probability of each). The damage status (i.e., whether it carries the characteristic C-to-T damage often observed in ancient DNA sequences) of each sequence was also determined based on the status of the ancient reference individual. Contaminating sequences were then "spiked-in" at varying proportions (0 to 40%), using an additional present-

day individual from the 1000 Genomes Project to determine the contaminating allele type (see the "Materials and methods" section). All contaminating sequences were defined to be undamaged, as would be expected if the contamination came from a non-ancient source.

For most of the analyses reported in this study, we simulate data for SNP sites targeted in the 1.24 million SNP capture reagents [2, 17] that intersect with 1000 Genomes sites, after removing sites on the X and Y chromosomes (this leaves ~ 1.1 million SNPs). The *ContamLD* software also allows users to make panels based on their own SNP sets, and in a later section, we report the results from a larger panel (~ 5.6 million SNPs) provided with the software that we recommend for shotgun-sequenced samples, which provides more power to measure contamination.

We first analyzed data generated using a reference individual from the 1000 Genomes CEU population (Utah Residents (CEPH) with Northern and Western European Ancestry) and the SNP coverage profile of a 1.02× coverage ancient individual of West Eurasian ancestry (Iberian Bronze individual I3756 who lived 2014–1781 calBCE; see the "Materials and methods" section). Additional file 1: Fig. S1 illustrates the distribution of logarithm of the odds (LOD) scores generated when *ContamLD* is run on samples with 0%, 7%, and 15% simulated contamination. Additional file 1: Fig. S2 shows the contamination rate estimates generated for data with simulated contamination rates between 0 and 40%. At very high contamination (above 15%), *ContamLD* often overestimates contamination, but in practice, samples with above 10% contamination are generally removed from population genetic analyses, so inaccuracies in the estimates at these levels are not a concern in our view (the importance of a contamination estimate in many cases is to flag problematic samples, not to accurately estimate the contamination proportion). *ContamLD* assumes that the individual making up the majority of the sequences is the base individual, so we do not explore contamination rates greater than 50% in these simulation studies.

We observe a linear shift in the contamination estimates such that most estimates are biased to be slightly higher than the actual value, with even greater overestimates occurring at higher contamination rates (Additional file 1: Fig. S2). This is likely due to the difference between the haplotype distribution of the test individual and that of the haplotype panel, since the magnitude of this shift increases as the test individual increases in genetic distance from the haplotype panel. Even in cases where the test individual is of the same ancestry as the haplotype panel (as in Additional file 1: Fig. S2), there is expected to be a shift, because the test individual's haplotypes are a particular sampling of the population's haplotypes, and the difference between having only frequencies of the haplotype panel and a particular instantiation of those frequencies in the test individual will lead to the artificial need for an external source ("contaminant") to fit the model properly.

In contrast to the upward bias in contamination estimates due to mismatch of the individual's haplotypes with the reference panel haplotype frequencies, we observe negative shifts for children of two closely related parents, as expected because *ContamLD* assumes the paternal and maternal copies of a chromosome are unrelated. In contrast, if the two chromosomes are related, extra LD will be induced and more contamination will be necessary to produce the expected LD pattern. In principle, this inbreeding effect could be corrected explicitly by estimating the total amount of ROH in each individual and applying this as a correction, although we

do not provide such functionality as part of our software. A reliable methodology for quantifying the proportion of the genome that is affected by inbreeding in ancient individuals has now become available (hapRHO), [18] and *ContamLD* could be further improved by using an estimate of ROH from software like this as an input parameter.

A final type of bias could be expected to arise if the contamination comes from an individual related to the target individual. In this case, the true contamination rate is expected to be underestimated, because *ContamLD* only detects contamination where the contaminant sequence differs from the target individual's sequence. If the contaminant carries the same haplotypes as the target individual, in the most extreme case as expected for an identical twin, then the existence of contamination will be missed altogether. In general, contamination from closely related individuals is unlikely to be a concern for many population genetic analyses, as close relatives usually (but with important exceptions) have very similar ancestry.

In our implementation, we correct for these systematic biases in two ways, implemented as different options in *ContamLD*.

The first option leverages information from sequences that contain evidence of the C-to-T damage that is characteristic of ancient sequences. This option assumes these sequences are authentically ancient and not derived from a contaminating source (assumed to be from present-day individuals), so the *ContamLD* estimate based on undamaged sequences is corrected by estimates based on the damaged sequences (see the "Materials and methods" section for more details). In the second option, we allow the user to subtract the contamination estimate from the estimate of an individual of the same ancestry assumed to be uncontaminated. An advantage of the second option compared to the first is that it has smaller standard errors (Fig. 1), reflecting the fact that it does not rely on estimates from damaged sequences (reliance on damaged sequences reduces power since it often reflects a very small subset of the data). Another

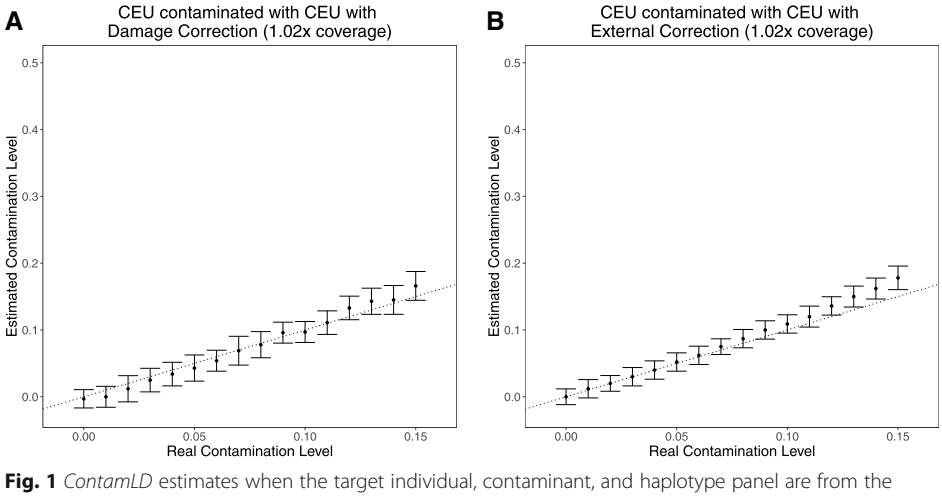


**Fig. 1** *ContamLD* estimates when the target individual, contaminant, and haplotype panel are from the CEU population. Contamination estimates when the simulated contamination rate is between 0.00 and 0.15. **a** Estimates with damage-restricted correction (option 1). **b** Estimates with an external correction from an uncontaminated sample (option 2). The black dotted line is $y = x$, which would correspond to a perfect estimate of contamination. Error bars are 1.96*standard error (95% confidence interval determined via jackknife resampling across chromosomes)

advantage of the second option is that it allows estimation of contamination in cases where the source of contamination is also ancient in origin, as would be expected if the contamination occurred thousands of years ago or due to cross-contamination with other ancient samples (the first option would be expected to produce an underestimate of contamination in such cases, since it assumes that sequences that contain C-to-T damage are not contaminated). On the other hand, a drawback of the second option is that it requires users to identify a relatively high coverage, uncontaminated, ancestry-matched samples for benchmarking purposes; the method is also only expected to work if there is minimal inbreeding in either the sample of interest or the matched sample. Identifying such benchmarking samples may be impossible when analyzing samples from previously unsampled contexts (e.g., early modern humans), and indeed verifying that a benchmarking sample is uncontaminated is very difficult if it is female (if it is male, a method like *ANGSD* can be used).

In what follows, we report the results of analyses based on the first option, but *ContamLD* includes both methods as options. The uncorrected score also forms the basis for warning output by the software, namely high contamination or possible contamination with another ancient sample leading to an inaccurate damage correction estimate.

### Simulated contamination of ancient samples with present-day samples

*ContamLD* is designed to work on ancient individuals, so we simulated contamination of real ancient individuals (Additional file 2: Table S1) with present-day individuals from the 1000 Genomes Project, a scenario that would occur when skeletal material from ancient individuals is contaminated by present-day individuals during excavation or at some point during the processing of the material. We used data from male individuals selected due to very low X chromosome contamination estimates (less than 1%) based on *ANGSD* [16] (developed first in Rasmussen et al. [9]; we used method 1 of that software). (We subtracted the *ANGSD* estimates from the *ContamLD* estimates to correct for any residual contamination.) Figure 2a shows the results from the Iberian Bronze Age sample [19] (I3756) with 1.02× coverage at the targeted ~ 1.24 million SNP positions, demonstrating that *ContamLD* produces highly accurate contamination estimates for this simulation.

### Effect of different haplotype panels

There are many potential cases in which ancient individuals can come from populations with very different genetic profiles compared to present-day 1000 Genomes populations, leading to an ancestry mismatch to the haplotype reference panels. *ContamLD* provides panels from all 1000 Genomes populations as well as tools to identify the panel most closely matching to the ancestry of their ancient individual (based on outgroup-$f_3$ statistics [20] to determine the most shared genetic drift), which they can then select for the analysis. However, due to the potential for ancestry mismatch to still occur, we tested the effect of choosing haplotype panels that are genetically diverged from the individual of interest (Fig. 2a). For the ancient Iberian sample, the CEU and Toscani in Italia (TSI) panels—representing northern and southern European ancestry, respectively—yielded contamination estimates that are close to the true contamination rate, especially for rates below 5%. However, *ContamLD* underestimates contamination

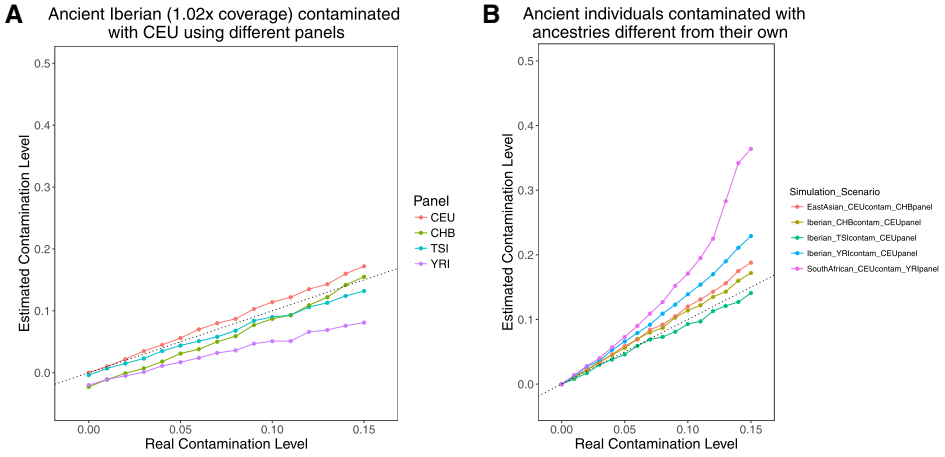


**Fig. 2** Genetic divergence between uncontaminated individual and contamination sources or haplotype panels impacts *ContamLD* estimates. **a** Ancient Iberian (I3756, 1.02× coverage) contaminated with CEU with haplotype panels generated from CEU, TSI, CHB, and YRI populations. **b** Contamination estimates from the same ancient Iberian contaminated with TSI, CHB, or YRI and analyzed with a CEU panel; from an ancient East Asian (DA362.SG, 1.10× coverage) contaminated with CEU and analyzed with a CHB panel; and from an ancient South African (I9028.SG, 1.21× coverage) contaminated with CEU and analyzed with a YRI panel. The black dotted line is $y = x$, corresponding to a perfect estimate of contamination. All estimates use the damage-restricted correction (option 1)

by ~ 2% when the CHB (Han Chinese in Beijing, China) and YRI (Yoruba in Ibadan, Nigeria) panels were used instead (though we view these as very pessimistic cases, because the user should usually be able to choose a panel more closely related to their ancient individual than these scenarios). We thus recommend that users take care to choose an appropriate panel that is within the same continental ancestry as their ancient individual. Nevertheless, we note that we were able to obtain reasonably accurate estimates for Upper Paleolithic European hunter-gatherers, such as the Kostenki14 individual [21], who is ~ 37,470 years old, even when using present-day European panels that have significantly different ancestry from the hunter-gatherers (Additional file 1: Fig. S3).

## Effect of mismatch between the ancestry of the true sample and contaminating individual

Contamination can come from a wide variety of sources, including, but not limited to, members of the archeological excavation team, the aDNA laboratory, or residual human DNA on the plastic and glassware or in laboratory reagents. Thus, we sought to understand the effect of mismatch in the ancestry of the true sample and the contaminating individual in our contamination estimates. We found that as the ancestry of the two diverged, *ContamLD* overestimated contamination (Fig. 2b and Additional file 1: Fig. S4). This occurred when we tested an ancient European with different contaminant ancestries and when we tested ancient East Asian [22] and ancient South African [23] samples contaminated with European DNA. Nevertheless, the overestimation was not severe at contamination levels below 5%, and samples above this proportion would likely be correctly flagged as problematic. We also explored scenarios where the ancestry of the panel matches the contaminant rather than the true sample (Additional file 1:

Fig. S4) and found a ~ 2% underestimate at low levels of contamination and an overestimate at high levels of contamination; these are modest effects and are unlikely to change our qualitative assessment. When we tested the effect of having multiple contaminant individuals (Additional file 1: Fig. S5), we found only a slight overestimate at higher levels of contamination, as expected given *ContamLD* normally assumes contamination from a single individual where the haplotypes are re-formed if they are created from two contaminant reads (which will happen at lower rates with more contaminant individuals).

### Estimating contamination in admixed individuals

*ContamLD* relies on measuring the difference between the LD pattern of the sample and that expected from an uncontaminated individual. However, individuals from groups recently admixed between two highly divergent ancestral groups have LD patterns that are similar in some ways to that of an unadmixed individual with contamination from a group with ancestry diverged from that of the individual of interest. To understand how this would impact *ContamLD*, we ran the software on an ASW (Americans of African Ancestry in Southwest USA) individual with different levels of added CEU contamination. When we ran *ContamLD* with a YRI panel and no correction on an individual with no contamination, the individual was inferred to have a contamination of ~ 20% (likely because the individual had ~ 15% European ancestry, and this was interpreted by the software as contamination). Using an ASW panel did not perform any better. However, the concerns were mostly addressed by the damage-restricted correction (option 1) at low contamination levels (Additional file 1: Fig. S6). The simulation with African-Americans represents an extreme of difficulty, because the individual is from a group with very recent admixture (~ 6 generations [24]) of ancestries highly divergent from each other with one of the ancestries very genetically similar to the reference panel. It highlights how the damage-restricted correction is still able to produce accurate estimates in these difficult cases.

### Effect of coverage

We tested the power of our procedure at different coverages with simulations of ancient West Eurasian ancestry individuals contaminated with CEU on the 1240K SNP set (Fig. 3). We found that while our estimates were not biased to produce estimates consistently above or below the true value, the standard errors increased significantly at lower coverages, as expected for the decreased power for accurate estimation in these scenarios. We provide a much larger panel with ~ 5.6 million SNPs (vs. ~ 1.1 million for the 1240K panel) that usually decreases standard errors for samples that are shotgun sequenced (Additional file 1: Fig. S7). This panel increases *ContamLD*'s compute time and memory requirements, so we recommend that it only be used for individuals with lower than 0.5× coverage. As an additional feature, we provide users tools to create their own panels to meet their specific needs.

### Effect of damage rate

We tested the power of ContamLD at different damage rates with simulations of an ancient West Eurasian ancestry individual (DA57.SG) down-sampled to 0.5x coverage

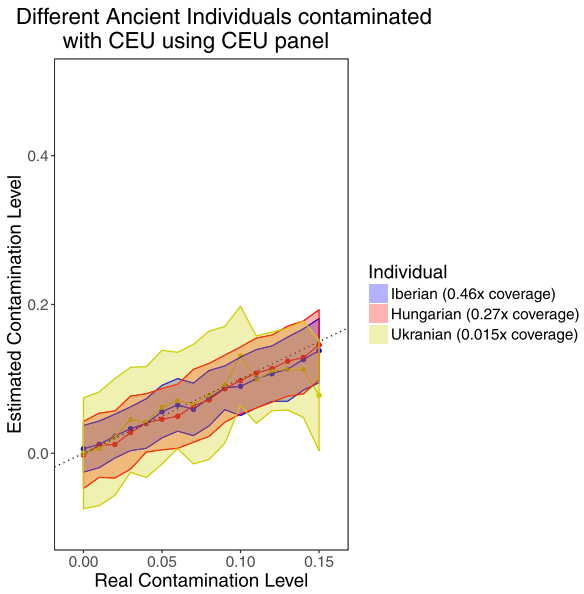


**Fig. 3** *ContamLD* estimates for ancient European samples of different coverages after damage-restricted correction (option 1). An ancient Iberian of 0.46× coverage, an ancient Hungarian of 0.27× coverage, and an ancient Ukrainian of 0.015× coverage (~ 16,000 SNPs) were contaminated with CEU and analyzed using a CEU panel with *ContamLD* option 1 (damage-restricted correction). The black dotted line is $y = x$. Error shading is 1.96*standard error (95% confidence interval)

contaminated with another ancient West Eurasian ancestry individual (I10895) as above with damage rates simulated to be between 0.005 to 0.075. We found that standard errors decreased as damage rate increased (Additional file 1: Fig. S8). The standard errors were below 3% at all damage rates above 0.01, which is lower than the damage rate of most ancient DNA samples even after partial UDG treatment.

### Simulations to compare *ContamLD* to *ANGSD* X chromosome estimates

We performed simulations where we randomly added contaminating sequences at increasing levels from 0 to 15% from an ancient West Eurasian individual (I10895) into the BAM files of 65 ancient male individuals of variable ancestries and ages (we set the damaged sequences to be only from the non-contaminant individual; see the "Materials and methods" section). We chose ancient male individuals that had average coverage over 0.5× and X chromosome contamination estimates under 2% (using method 1 of *ANGSD*) when no artificial contamination was added (and also corrected even for this baseline contamination by setting damaged reads to be a 5% down-sampling of the files that had no artificial contamination; see the "Materials and methods" section). We then analyzed the individuals with *ContamLD* and *ANGSD* and found that compared to *ANGSD*, *ContamLD* consistently had similar errors relative to the true contamination level (Fig. 4, Additional file 3: Table S2).

### Comparing *ContamLD*, *ANGSD*, and mitochondrial estimates (*ContamMix*) in ancient individuals without added contamination

We tested 439 ancient males with *ContamLD*, *ANGSD* (X chromosome contamination estimates), and *ContamMix* (mitochondrial contamination estimates) without adding

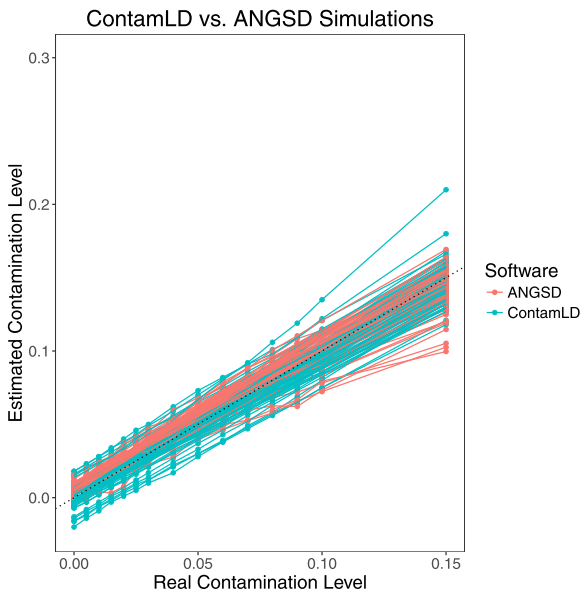


**Fig. 4** Contamination estimates with *ContamLD* and *ANGSD* for ancient individuals with different levels of contamination added. Sixty-five ancient individuals with average coverage over 0.5× had increasing levels of artificial contamination added in (from I10895, an ~ 1200BP ancient West Eurasian individual) and were then analyzed with *ContamLD* (with panels most genetically similar to the ancient individual and using damage-restricted correction, option 1) and *ANGSD*. Details of all estimates (including standard errors) are provided in Additional file 3: Table S2. The black dotted line is $y = x$, which would correspond to a perfect estimate of the contamination

additional contamination. For this analysis, we included published data generated with the ~ 1.24 million SNP enrichment reagent, as well as data from libraries that failed quality control due to evidence of contamination (Additional file 4: Table S3). Similar to prior studies [5], the mitochondrial estimates often differed from the nuclear (*ANGSD* and *ContamLD*) estimates, showing high contamination in some libraries with low

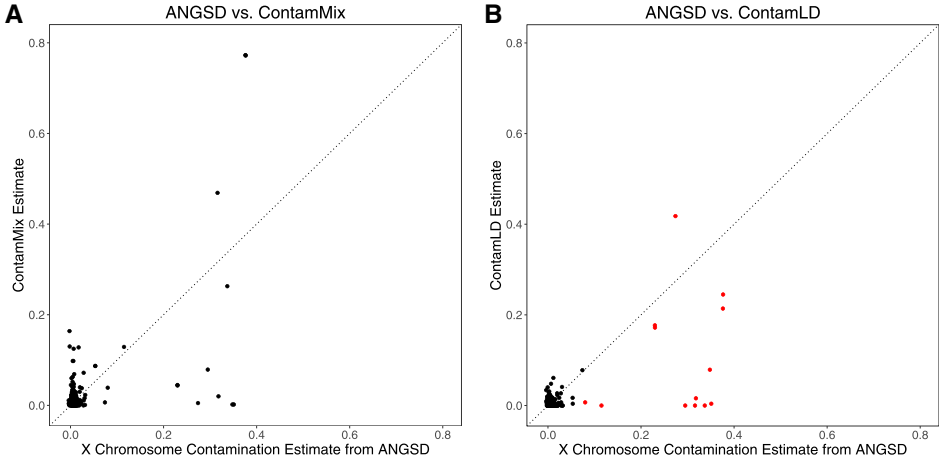


**Fig. 5** Contamination estimates from *ContamLD*, *ANGSD*, and *ContamMix* in 439 ancient individuals of variable ancestry. *ANGSD* estimates (method 1) are plotted on the *x*-axis, and on the *y*-axis are either **a** *ContamMix* or **b** *ContamLD* estimates. In red are samples that were flagged in *ContamLD* as "Very_High_Contamination" based on having uncorrected estimates over 15%. All *ContamLD* estimates below 0 were set to 0

nuclear contamination, and low mitochondrial contamination in some libraries with high nuclear contamination, likely reflecting the known biological phenomenon of orders of magnitude of variation in the mitochondrial to nuclear DNA ratio (Fig. 5a). In contrast, *ANGSD* and *ContamLD* had better concordance. However, we observed that some of the samples with high contamination estimates based on *ANGSD* had much lower *ContamLD* estimates, reflecting over-correction from analyzing the damaged sequences, perhaps because the contamination was actually cross-contamination from other ancient individuals, violating the assumptions of our damage correction (Fig. 5b). This problem was mitigated in part, however, because *ContamLD* produces a warning of "Very_High_Contamination" if the uncorrected estimate is above 15% (even in cases where the corrected estimate is very low), and all samples with X chromosome estimates over 5% were flagged with this warning and/or had estimates of over 5% contamination with *ContamLD* (all samples with less than 5% contamination in *ANGSD* had lower than 5% contamination with *ContamLD*). It is unfortunately not possible to know the true contamination of the samples we tested in Fig. 5, but the fact that our software produced results with good correlation to X chromosome estimates shows that it works well in real ancient data.

It is possible for there to be samples with moderately high contamination from another ancient individual but both a low damage-restricted correction estimate and no warning generated, because these would have high uncorrected estimates, yet not high enough to reach the threshold required for the warning. These samples would have to be identified with an external correction. Lowering the threshold for the "Very_High_Contamination" warning would produce too many false positives, because there are many cases with high uncorrected estimates that have low corrected estimates that are likely not contaminated (e.g., due to ancestry mismatches of the panel and the test individual). To understand these issues better, we performed a simulation in which an ancient Iberian (I3756) was contaminated with another ancient West Eurasian individual (I10895), and the damaged sequences were set to be a 5% down-sampling of the set of contaminated sequences (thus simulating a case in which all of the contamination is from another ancient individual who has the same damage proportion as the ancient individual of interest). We found that, as expected, the contamination from the ancient individual was not detected (the contamination estimates were always near 0%) by the damage-restricted correction version of *ContamLD* until the contamination reached 15% at which point the "Very_High_Contamination" flag came up (Additional file 1: Fig. S9). The contamination would have been detected with the external correction version of *ContamLD* (since the damage-restricted correction continued to go up with increasing contamination; see Additional file 5: Table S4), but without an uncontaminated ancient individual of the same group as the target individual, this would be difficult to do without bias in the contamination estimate.

## Discussion and conclusion

We have presented a tool, *ContamLD*, for estimating rates of autosomal DNA contamination in aDNA samples. *ContamLD* is able to measure contamination accurately in both male and female individuals, with standard errors less than 1.5% for individuals with coverage above 0.5× on the 1240K SNP set (for contamination levels less than 10%) for the damage-restricted correction method (option 1). On the shotgun panel we provide,

standard errors are less than 1.5% for coverages above 0.1×. *ContamLD* is best suited to scenarios in which the contaminant and the ancient individual of interest are similar ancestries, which is useful, because *DICE* [15] and many population genetic tools (e.g., PCA or ADMIXTURE [25]) are better suited for detecting cases where the contaminant is of very different ancestry from the ancient individual of interest. *ContamLD* works even for recently admixed individuals. Lastly, *ContamLD* can detect cases of contamination from other ancient individuals, though this works best if it is large amounts of contamination that can reach the threshold required for the "Very_High_Contamination" flag.

We tested *ContamLD* in multiple simulation scenarios to determine when bias or less reliable results could be expected. When applied to the situation with a test individual (ancient or present-day), contaminant, and haplotype reference panel all from the same continental ancestry, *ContamLD* provides an accurate, unbiased estimate of contamination. When the contaminant comes from a population that is of a different continental ancestry from the population used for the base and haplotype panel, the contamination appears to be slightly overestimated, particularly for higher contamination. This should not be a large problem in analyses of real (i.e., non-simulated) data, because the effect is small at the contamination levels of interest (< 5%). When we varied haplotype panels, we found that the estimator is robust when applied to simulated datasets using haplotype panels that are moderately divergent from the base sample (within-continent levels of variation). We provide users tools for automatically determining the panel that shared the most genetic drift with the sample so that the user can select the panel most closely related to the sample. In other simulations, we found that the performance of the algorithm declines as the coverage of the sample decreases. The estimates are not biased, but the standard errors substantially increase when fewer than 300,000 sequences are available. In these cases, if the individual was shotgun sequenced, we recommend that users choose the shotgun panel, which will substantially increase the power for the analyses.

We applied the algorithm to estimate contamination levels in dozens of ancient samples and compared them to X chromosome-based contamination estimates. There was generally good correlation with the X chromosome estimates, except that when the true contamination was very high, the LD based estimates were sometimes estimated incorrectly, likely because the contamination was due to cross-contamination from another ancient individual and there was over-correction from the damage estimates. This problem is mitigated, however, because the software indicates if the uncorrected estimate is very high so users can identify highly contaminated samples and remove them from further analyses. A difficult case for the software is if there is contamination in part from another ancient sample. This can cause an over-correction and lead to an underestimate of the contamination. The "Very_High_Contamination" warning catches very high contamination from other ancient samples, but it will miss cases of moderate levels of contamination from other ancient samples, because it will not reach the threshold required for the warning. In theory, the user can determine the true contamination in these cases using the external correction, but the external correction can be difficult if the user does not have an adequate sample to correct the estimate of the sample of interest. The damage correction of the software also does not work if the samples have undergone full UDG treatment (which results in very few damaged sequences), and for this case, the external correction is the only option.

The software run time is dependent on SNP coverage. If ~ 1,000,000 SNPs are covered (the depth of the coverage on each SNP does not affect run time), the analysis takes approximately 2 h if 3 cores are available on CentOS 7.2.15 Linux machines (~ 25 GB of memory). The software is designed for samples to be run in parallel, so the total time for analysis even for large numbers of samples is often not much greater than the time for a single sample.

In summary, *ContamLD* is able to estimate accurately autosomal nuclear contamination in ancient DNA with standard errors that depend on the coverage of the sample. This will be particularly useful for female samples where X chromosome estimates are not possible. As a general recommendation for users, we believe in most cases all samples with a contamination estimate that is greater than 0.05 (5%) should be removed from further analyses, or the contamination should be explicitly modeled in population genetic analyses.

## Materials and methods

### Datasets

#### *Present-day samples*

Genome-wide datasets from individuals that were part of the 1000 Genomes Project [26] were used as present-day reference data. We restricted to autosomal sites included in the ~ 1.24 million SNP capture reagent [2, 17] and to SNPs at greater than 10% minor allele frequency in the pooled 1000 Genomes Project dataset [26]. However, the software allows users to make custom SNP panels. In the analyses presented here, we filtered for SNPs that were present in the 1000 Genomes dataset and also removed all sex chromosome SNPs leading to 1,085,678 SNPs in the final 1240K dataset and 5,633,773 SNPs in the final shotgun dataset.

#### *Ancient data set*

We analyzed mitochondrial and X chromosome contamination estimates [16, 27] from ancient individuals from previous studies generated by shotgun sequencing or targeted enrichment for 1.24 million SNPs, including many samples that failed quality control due to contamination but were from the same archeological sites [2, 23, 28–34]. Information about the ancient individual data is detailed in Additional file 2: Table S1 and below.

#### *Obtaining sequence information*

For each ancient individual, we generated the sequence depth data from the sample bam file, counting the number of reference and alternative alleles at each SNP site in the analysis dataset. Damage-restricted data were generated by restricting to sequences with PMD scores greater than or equal to 3 [4]. Our software can accommodate both genotype call data as well as sequence data (the sequence data adds additional power to the analyses), but all analyses were performed using the sequence-based method. We provide users with tools to pull down read count data from BAM files in the format required for *ContamLD*.

### Haplotype calculation

To create haplotype panels, we obtained all SNP pairs in high LD for each 1000 Genomes population using PLINK version 1.9 [35] with $r^2$ cutoff of 0.2. (Users can increase power slightly at the expense of increased computational time by creating their own haplotype panel with a lower $r^2$ cutoff.) We then calculated the frequencies of each SNP in all of these pairs as well as the haplotype frequencies at each of these pairs while holding out the present-day individuals used for contamination simulation.

### Algorithm to estimate contamination

#### *Overview*

Our goal is to estimate $\alpha$, the level of contamination, by examining the frequencies of allele pairs that should be in LD (we term this two-allele pair a haplotype) and determining how much their frequencies differ from what would be expected under no contamination. To estimate this, we need both the distribution underlying the haplotypes ($q$) that an uncontaminated test sample should have as well as the distribution of "unrelated haplotypes" ($\tilde{h}$) that would form by chance from background allele frequencies. Here and below, "distribution" refers to the set of frequencies of the different possible haplotypes (all possible combinations of ancestral and derived alleles at the SNP pair) across all haplotypes in the genome. Additional file 1: Fig. S10 is a schematic of the algorithm.

#### *Determining haplotype distributions based on reference panels*

To determine $q$, we must account for the fact that the test individual's genotypes do not have diploid calls and are not phased. Due to the low sequence depths at each SNP in many ancient DNA datasets, it is difficult to make confident heterozygous calls, so instead, we create pseudo-haploid calls by randomly choosing a sequence to represent the genotype at that position (this holds when we are using genotype calls or the sequence information directly, and when multiple sequences cover the same SNP, we use all of them and treat them as independent). Thus, for this analysis, when examining a pair of SNPs, it is equally likely for the SNP pair to have been formed from the true haplotype (if the same parental chromosome is sampled from in both SNPs of the haplotype) or the background distribution (if the opposite parental chromosome is sampled). We therefore can estimate $q$ as:

$$q = h/2 + \tilde{h}/2$$

where $h$ is the distribution of true haplotypes and $\tilde{h}$ is the distribution of unrelated haplotypes that would form by chance from background allele frequencies. For inbred samples, the weight on $h$ is more than 1/2, because the two parental chromosomes are more related, but this can generally be corrected (see below).

$\tilde{h}$ can be determined by multiplying the SNP frequencies to obtain the haplotype frequencies that would form after randomly pairing SNPs of unrelated individuals. $h$ can be estimated from an external reference panel using a maximum likelihood estimator (MLE) to obtain haplotype frequencies in the population from the counts (necessary because the panels are not phased). The MLE setup is:

$$log(L(h|c)) = \sum_{j=1}^{n} \sum_{i=1}^{4} c_{ij} log(P(i, j|h))$$

with

$$P(i, j|h) = \sum_{a1,a,b1,b2=0,1;a,b\rightarrow(i,j)} h_{(a_1,b_1)} * h_{(a_2,b_2)}$$

where $P(i, j|h)$ is the (unknown) diploid count distribution of the haplotypes of the population the test individual is from (approximated by the external panel), $n$ is the number of SNP pairs, $c$ is the vector of observed haplotypes in the diploid count panel (from 1000 Genomes), $i$ sums over all 4 haplotype possibilities, $h_{(a,b)}$ are the (also unknown) haplotype distributions of the parents of the test individual (the haploid chromosomes they pass on to their child), and $a, b \rightarrow (i, j)$ implies that $a_1 + a_2 = i$ and $b_1 + b_2 = j$, meaning that one adds up all cases where the haplotype combination would lead to a particular diploid count (e.g., in the notation, for example, 01,11 means the first parent contributes a haplotype that has 0 alternative alleles at the first SNP and 1 alternative allele at the second SNP, and the second parent contributes a haplotype where both SNPs have the alternative allele. The test individual with these parents would then have a 12 diploid count, which means at the first SNP the individual has 1 alternative allele and at the second, SNP the individual has 2 alternative alleles. Since our observed data are not phased, both 01,11 and 11,01 would lead to a 12 diploid count. This assumes independence of SNP pairs, which is not true, but because our standard errors are based on jackknife resampling across chromosomes, correlation among SNP pairs is corrected for in our error estimates.

The MLE would be computationally intractable to solve due to our lack of knowledge of which parent contributed to each count, so we instead used an EM algorithm to obtain $h$, where knowledge of the parents' contribution is the unobserved latent variable. The algorithm involves an expectation step of:

$$n_1 = \frac{C(i, j) * \sum_{a,b\rightarrow(i,j)} h_{(a,b)} * h_{(a_2,b_2)}}{P(i, j|h)}$$

where $n_1$ is the expected number of times that the $(a, b)$ configuration of the father's chromosome contributed to a particular diploid count (this is the same value for the mother, $n_2$, because they are assumed to be from the same haplotype distribution). In other words, given the observed haplotype counts in the reference panel, how many times would it be expected that a particular haplotype configuration (e.g., ancestral at SNP1, derived at SNP2) in one of the parents contributed to those counts?

Once the counts ($n_1$ and $n_2$) of the haploid parents are obtained, they are added together to produce the diploid individual (i.e., the expected number of all possible haplotype configurations). Then, the expected value of the haplotype distribution can be maximized by averaging over the possible haplotype distributions. Thus, the maximization step is:

$$D_{(a,b)} = \sum_{(i,j)} C_{(i,j)} * [n_1 + n_2]$$

$$h_{(\hat{a},b)} = \frac{D_{(a,b)}}{\sum_{a,b} D_{(a,b)}}$$

where $D(a,b)$ is the sum of the probabilities of a particular haplotype configuration over all diploid count configurations.

We initially set all $h(a,b)$ to be 0.25 and then iterated through the algorithm until convergence (using a squared distance summed over all SNPs and a threshold of 0.001). We then used this estimate of $h$ to get an estimate of $q$ (based on the first equation above).

#### *Estimating contamination based on haplotype distributions and test individual's haplotypes*

To estimate $\alpha$, we used the equation:

$$T = \left(1 - 2\alpha' + 2\alpha'^2\right)q + 2\alpha'(1-\alpha')\tilde{h}$$

Here, $T$ is the expected distribution underlying the observed haplotypes of the sample, which is a mix of the test individual and contaminant. This means that assuming the test individual comes from a population with a haplotype distribution (frequency of the different haplotype possibilities at each SNP pair throughout the genome) that can be approximated by the chosen reference panel (and estimated as above), $T$ is the haplotype distribution expected for the sample given a particular amount of contamination ($\alpha'$, where $'$ is used to indicate that this is an estimate of the real $\alpha$). $q$ is the haplotype distribution for an uncontaminated sample. A fraction $(1-\alpha')^2 + \alpha'^2$ of the distribution should look like this, where $(1-\alpha')^2$ is the probability that two uncontaminated sequences form the SNP pair and $\alpha'^2$ is the probability that two contaminated sequences form the SNP pair, assuming the contaminating sequences are from a single individual, which would "re-form" a SNP pair with LD (note: this also makes the simplifying assumption that the contaminant and the test individual have the same background haplotype and SNP distribution). $\tilde{h}$ is the distribution of unrelated "haplotypes" that would form by chance from background allele frequencies in the population. Contamination would form these unrelated haplotypes by breaking up LD, so a fraction $2\alpha'(1-\alpha')$ of the distribution should look like this (the probability that the SNP pair is formed from a contaminated sequence and an uncontaminated sequence).

This expression can be used to solve for $\alpha'$ by maximizing the log of the odds (LOD) scores under the null hypothesis that $\alpha' = 0$ and the alternative hypotheses of different $\alpha'$. A LOD score is assigned to each estimate of the contamination rate ($\alpha$) between −0.1 and 0.5 (negative scores are included to allow correction for inbreeding). The grid of $\alpha'$ is scaled by intervals of 0.0001. The $\alpha'$ with the highest LOD score is the best estimate of $\alpha$ and is returned. When we have multiple sequences on the same SNP, we assume independence of the sequences, which provides additional power. The assumption of independence does not bias the error estimation for the same reason as explained above for the independence of SNP pairs.

#### *Correcting for bias in contamination estimates*

In practice, the $\alpha'$ we obtain is not equal to the true $\alpha$, because the reference panel does not perfectly capture the SNP and haplotype frequencies of the test sample. We found that this difference causes a linear shift in contamination estimate where the mismatch between the sample individual and the reference panel leads to a positive

shift while inbreeding leads to a negative shift. These biases can be addressed in either of two ways.

First, for the "damage correction" approach, we performed an $\alpha'$ estimate only on alleles from sequences with evidence of damage characteristic of ancient samples. Under the assumption that these sequences are not affected by present-day contamination, the inferred $\alpha'$ would be an estimate of the bias, which can be subtracted out from the estimate based on all sites. We separately analyzed the following pairs of SNPs: UU (both SNPs at undamaged sequences), DU (one site damaged and the other undamaged), and DD (both SNPs at damaged sequences). For the UU pairs, the value we calculate would be $\alpha + k$, where $k$ is the linear shift. For DU pairs, the value calculated would be $\alpha/2 + k$, and for DD pairs, the value calculated would be $k$. We added the likelihoods for these pairs and maximized the likelihood to solve for $\alpha$ and $k$. After solving for $\alpha$, we multiply by (1 damage rate) to obtain the contamination level across all sequences, because $\alpha$ is the contamination rate at undamaged sequences.

Second, for the "external correction" approach, we took individuals from the test individual's population that were high coverage and samples we believed had very low contamination (based on X chromosome estimates with *ANGSD* using method 1 as developed first by Rasmussen et al. [9]) and measured $\alpha'$. We assumed a true contamination of 0 for these samples and thus subtracted this $\alpha'$ from all other contamination estimates. We caution that this method does not correct for uncertainty in the contamination estimate in the external sample used for benchmarking.

#### *Comparison to a similar method*

The approach of *ContamLD* is similar to that of Vohr et al. [36] except the two have opposite goals. Vohr et al. searches for LD in reads from two different samples in an attempt to determine whether the two samples are from the same individual (or closely related individuals), using a reference panel to determine LD patterns. In contrast, *ContamLD* searches for breaks in LD in the sequences of a single sample to determine if sequences from other individuals are present in the sample.

### Data simulation

To test the accuracy of the algorithm, we applied it to a variety of scenarios with both present-day DNA as well as real aDNA samples that had simulated present-day DNA contamination. In all our simulations with 1000 Genomes individuals, we removed the individual being used from our haplotype panel before performing the analyses.

#### *Simulating contamination of present-day individuals*

We first simulated contamination of present-day individuals with other present-day individuals as contaminants (this allowed us to be sure that there was no baseline contamination). In order to best approximate the distribution of both the damaged and undamaged sequences that is characteristic of aDNA data, we used sequence depth information from an ancient individual as a reference. At each SNP, the total number of simulated "damaged" and "undamaged" sequences was determined based on the number of damaged and undamaged sequences at the SNP in the reference ancient

individual. The identity of each allele for the present-day "base" sample was randomly chosen based on the genotype of the "base" present-day 1000 Genomes individual at each SNP, as described above for the contamination. The addition of contaminant sequences to the dataset was performed using the method described above. In order to reduce bias caused by the damage correction procedure, the damage-restricted dataset was generated only once for each simulation type (which included multiple simulations across varying contamination rates) and combined with the undamaged dataset to produce the overall dataset. This method was used to generate a simulated individual using present-day CEU (NA06985) or ASW (NA19625) from the 1000 Genomes dataset as the "base" sample from the sequence distributions of a 1.02× coverage ancient Iberian individual (I3756) (the "reference") [19]. The CEU (NA06984) individual was used as a "contaminant" in each case.

We generated simulated data with contamination from multiple sources by adjusting the present-day contamination simulation method to randomly sample from two or more present-day source contaminant genomes with equal probability. In each case, a 1000 Genomes Project CEU individual (NA06985) was used as a "base" genome with the sequence distribution of I3756 (the "reference"). In the case of 2 sources of contamination (Additional file 1: Fig. S5), two CEU individuals from the 1000 Genomes Project dataset (NA06984 and NA06986) were used as contamination sources, and in the case of three contamination sources, an additional CEU individual was used (NA06989). Data was generated for all combinations of undamaged contamination rates, $\alpha$, from 0 to 15%.

#### *Simulated contamination of ancient individuals*

We performed two sets of simulations contaminating different ancient individuals. In both cases, we selected ancient male individuals with minimal contamination (as assessed by X chromosome contamination levels from *ANGSD* [16]) to act as the "base" uncontaminated genome. In the first simulation set, we tested *ContamLD*'s performance with different ancient individuals and different present-day contaminant individuals from the 1000 Genomes dataset [26] to assess the impact of contaminant ancestry and coverage of the ancient individual. In this case, we were only using *ContamLD*, and thus, we performed the simulated contamination on the genotype level. In the second simulation set, we compared *ContamLD* to *ANGSD* and used a ~ 1200BP ancient West Eurasian individual (I10895) to contaminate the BAM files directly.

In the first simulation set, we assumed that sequences with C-to-T damage are highly unlikely to be the product of contamination (this assumption would be falsified in the context of cross-contamination by another ancient DNA sample). Thus, we exclusively added contamination to the "undamaged" fraction of sequences. At each SNP site, we classified sequences present in the damage-restricted dataset as "damaged" and added to the simulated data. We classified all other sequences as "undamaged" and also added them to the simulated data, but for each "undamaged sequence," we added a contaminant sequence to the simulated SNP data with probability $\alpha/(1-\alpha)$, where $\alpha$ is equal to the contamination rate (since the added sequences contribute to the total number of sequences, we needed to add a higher proportion than the contamination rate to obtain our desired contamination rate). The identity of the added contaminant allele was

randomly chosen based on the genotype of the chosen "contaminant" present-day genome at the site (i.e., if the contaminant individual was homozygous at the site, the allele it possesses would be added to the simulated individual, while if it were heterozygous at the site, either the reference or alternative allele would be selected randomly and added to the simulated individual). This method maintains the underlying distribution of "uncontaminated" reference and alternative alleles at each SNP site, while adding additional "contaminant" alleles to each site, producing an overall contamination rate of $\alpha$ in the undamaged sequences.

For each simulation, we generated two output files: (1) a file reporting the total number of sequences carrying reference and alternative alleles at each SNP and (2) a damage-restricted file reporting the total number of damaged sequences carrying reference and alternative alleles at each SNP. We used a 1.02× coverage ancient Iberian individual (I3756) (Additional file 2: Table S1) with contamination from either the 1000 Genomes CEU individual NA06984, the TSI individual NA20502, the CHB individual NA18525, or the YRI individual NA18486. We also used 5 other ancient individuals: I1845 (an ancient Iberian sample of 0.46× coverage) [19], I2743 (an ancient Hungarian of 0.27× coverage) [31], I5891 (a Neolithic Ukrainian individual of 0.016x××coverage) [37], DA362.SG (a Russian early Neolithic Shamanka East Asian individual of 1.10× coverage) [22], and I9028.SG (a South African individual of 1.21× coverage) [23]. In each case, we simulated individuals with 0–15% contamination.

For the second simulation set, we analyzed 65 ancient individuals of average coverage over 0.5× and baseline *ANGSD* estimates under 2% (Additional file 3: Table S2). In these cases, we added artificial contamination with sequences from a ~ 1200BP ancient West Eurasian individual (I10895) into the BAM files at the following proportions: 0.000, 0.005, 0.010, 0.020, 0.025, 0.030, 0.040, 0.050, 0.060, 0.070, 0.080, 0.090, 0.100, and 0.150. We removed two base pairs from the end of each sequence of partial UDG-treated samples and ten nucleotides for non-UDG-treated samples and pulled down the genotypes by randomly selecting a single sequence at each site covered by at least one sequence in each individual to represent the individual's genotype at that position ("pseudo-haploid" genotyping). To ensure that the damage sequences were only from the non-contaminant individual (so that we could use the damage-restricted correction mode, option 1, of *ContamLD* without bias), we created the "damaged" sequence set as a randomly chosen 5% of the sequences from the non-contaminant individual. We then analyzed the data with *ContamLD* (damage-restricted correction version, option 1) and *ANGSD* using default settings (method 1). We also performed simulations with a 1.0× coverage ancient West Eurasian ancestry individual (DA57.SG, an ancient Krgyzstanian individual) [38] down-sampled to 0.5× coverage and contaminated with I10895. To simulate different damage rates, we varied the damage rate to the proportions 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, and 0.075 by setting the amount of "damaged" sequences to be those proportions.

As the last simulation, we examined the case of an ancient individual contaminating another ancient individual where some of the damaged sequences would also come from the contaminating individual. In this simulation, we analyzed a 1.02× coverage ancient Iberian individual (I3756) and contaminated the BAM with sequences from a ~ 1200BP ancient West Eurasian individual (I10895) in the proportions 0.000, 0.005,

0.010, 0.020, 0.025, 0.030, 0.040, 0.050, 0.060, 0.070, 0.080, 0.090, 0.100, 0.150, 0.200, and 0.300. We then down-sampled the BAM, taking a random 5% of the sequences of these contaminated BAM files to act as the "damaged" sequences, because this would correct for any baseline contamination in the I3756 individual yet would simulate additional contamination of I3756 by an ancient individual with the same damage rate as I3756 (i.e., if there is 5% contamination, then also 5% of the damaged sequences would be from the contaminant individual in this simulation). We then performed the standard processing of both the full contaminated BAMs and the 5% down-sampled BAMs (simulated to be "damaged" sequences), removing two base pairs from the end of each sequence and carrying out a "pseudo-haploid" genotype pulldown. We ran *ContamLD* on the resulting data with damage-restricted correction, option 1.

### *Direct analyses of contamination levels in ancient individuals*

As our last set of analyses, we directly measured the contamination levels in ancient individuals without simulated contamination. We used *ContamLD* to examine the shotgun-sequenced individuals analyzed at the1240K SNP set and the large 5.6 million SNP shotgun panel. The ancient shotgun sequenced individuals were of 0.1–0.5× coverage from Allentoft et al. [32]; Damgaard et al., Nature [38]; and de Barros Damgaard et al., Science [22]. In addition, we analyzed 439 individuals from a variety of ancestries with *ContamLD* (damage-corrected version), *ANGSD* [16, 39] using default settings (we report the results from Method 1), and *ContamMix* [40] with the settings: down-sampling to 50× for samples above that coverage, --trimBases X (2 bases for UDG-half samples and 10 bases for UDG-minus samples), 8 threads, 4 chains, and 2 copies, taking the first one that finishes. Additional file 2: Table S1 includes all information from these individuals.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s13059-020-02111-2.

**Additional file 1: Supplementary Figures.** 10 supplementary figures showing results of additional analyses.

**Additional file 2: Table S1.** Ancient Individual Meta-Data. Excel spreadsheet detailing information about the ancient individuals analyzed.

**Additional file 3: Table S2.** *ContamLD* and *ANGSD* Estimates. Excel spreadsheet providing *ContamLD* and *ANGSD* estimates from simulations of ancient individuals with meta-data of the ancient individuals also included.

**Additional file 4: Table S3.** *ContamLD*, *ContamMix*, and *ANGSD* Estimates. Excel spreadsheet providing *ContamLD*, *ContamMix*, and *ANGSD* estimates from ancient male individuals with meta-data of the ancient individuals also included.

**Additional file 5: Table S4.** Simulation with Damaged Contaminated Reads. Excel spreadsheet providing *ContamLD* estimates for a simulation where the damaged reads of the contaminant individual were also included in the contamination.

**Additional file 6.** Review history.

#### Acknowledgements

We thank Iosif Lazaridis and Mark Lipson for helpful discussions.

#### Review history

The review history is available as Additional file 6.

#### Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

#### Authors' contributions

N.N., E.H., N.P., and D.R. conceived the study. N.N., E.H., and S.M. performed the analysis. N.N., E.H., and D.R. wrote the manuscript with the help of all co-authors. The authors read and approved the final manuscript.

**Authors' information**
Twitter handles: @EadaoinSays (Éadaoin Harney).

**Funding**
Funding was provided by an NIGMS (GM007753) fellowship to NN and a MHAAM fellowship to EH. DR is an investigator of the Howard Hughes Medical Institute; this work was supported by grants HG006399 and GM100233 from the National Institutes of Health, by an Allen Discovery Center grant from the Paul Allen Foundation, and by grant 61220 from the John Templeton Foundation.

**Availability of data and materials**
All data analyzed in this article are available in [2, 22, 23, 28–34, 38]. The software is available at https://github.com/nathan-nakatsuka/ContamLD [41] with the following open source license: https://github.com/nathan-nakatsuka/ContamLD/blob/master/LICENSE. It requires Python 3 and R (any version should suffice). Archived version (1.0) used for analyses in this manuscript: https://zenodo.org/record/3736774#.XoTbj257mgQ (https://doi.org/10.5281/zenodo.3736774) [42].
Scripts for data simulations are available in the GitHub folder "Simulation_Scripts."

**Ethics approval and consent to participate**
Not applicable (all samples were from previously published studies).

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
[1]Department of Genetics, New Research Building, Harvard Medical School, 77 Ave. Louis Pasteur, Boston, MA 02115, USA. [2]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA. [3]Department of Human Evolutionary Biology, Harvard University, 16 Divinity Ave., Cambridge, MA 02138, USA. [4]Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Ave., Cambridge, MA 02138, USA. [5]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02141, USA. [6]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA.

Received: 7 February 2020 Accepted: 16 July 2020
Published online: 10 August 2020

**References**

1. Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, Garcia N, Paabo S, Arsuaga JL, Meyer M. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci U S A. 2013;110:15758–63.
2. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522:207–11.
3. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos Trans R Soc Lond Ser B Biol Sci. 2015;370:20130624.
4. Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Paabo S, Krause J, Jakobsson M. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc Natl Acad Sci U S A. 2014;111:2229–34.
5. Sawyer S, Renaud G, Viola B, Hublin JJ, Gansauge MT, Shunkov MV, Derevianko AP, Prufer K, Kelso J, Paabo S. Nuclear and mitochondrial DNA sequences from two Denisovan individuals. Proc Natl Acad Sci U S A. 2015;112:15696–700.
6. Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Paabo S. DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci U S A. 2013;110:2223–7.
7. Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell. 2008;134:416–26.
8. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, et al. A draft sequence of the Neandertal genome. Science. 2010;328:710–22.
9. Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science. 2011;334:94–8.
10. Moreno-Mayar JV, Korneliussen TS, Dalal J, Renaud G, Albrechtsen A, Nielsen R, Malaspinas A-S. A likelihood method for estimating present-day human contamination in ancient male samples using low-depth X-chromosome data. Bioinformatics. 2020;36:828–41.
11. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics. 2011;27:2601–2.
12. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet. 2012;91:839–48.
13. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, De Filippo C. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338:222–6.
14. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, De Filippo C. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505:43–9.
15. Racimo F, Renaud G, Slatkin M. Joint estimation of contamination, error and demography for nuclear DNA from ancient humans. PLoS Genet. 2016;12(4):e1005972. https://doi.org/10.1371/journal.pgen.1005972. eCollection 2016 Apr.

16. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014;15:356.
17. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528:499–503.
18. Ringbauer H, Novembre J, Steinrücken M. Detecting runs of homozygosity from low-coverage ancient DNA bioRxiv. https://doi.org/10.1101/2020.05.31.126912.
19. Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, Dulias K, Edwards CJ, Gandini F, Pala M. The genomic history of the Iberian Peninsula over the past 8000 years. Science. 2019;363:1230–4.
20. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. Ancient admixture in human history. Genetics. 2012;192:1065–93.
21. Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, Albrechtsen A, Ko A, Margaryan A, Moiseyev V. Genomic structure in Europeans dating back at least 36,200 years. Science. 2014;346:1113–8.
22. de Barros DP, Martiniano R, Kamm J, Moreno-Mayar JV, Kroonen G, Peyrot M, Barjamovic G, Rasmussen S, Zacho C, Baimukhanov N. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science. 2018;360:eaar7711.
23. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, Salie T, Rohland N, Mallick S, Peltzer A, et al. Reconstructing prehistoric African population structure. Cell. 2017;171:59–71 e21.
24. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D. Methods for high-density admixture mapping of disease genes. Am J Hum Genet. 2004;74:979–1000.
25. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
26. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68–74.
27. Renaud G, Slon V, Duggan AT, Kelso J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol. 2015;16:224.
28. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513:409–13.
29. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016;536:419–24.
30. Lazaridis I, Mittnik A, Patterson N, Mallick S, Rohland N, Pfrengle S, Furtwangler A, Peltzer A, Posth C, Vasilakis A, et al. Genetic origins of the Minoans and Mycenaeans. Nature. 2017;548:214–8.
31. Lipson M, Szecsenyi-Nagy A, Mallick S, Posa A, Stegmar B, Keerl V, Rohland N, Stewardson K, Ferry M, Michel M, et al. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. Nature. 2017;551:368–72.
32. Allentoft ME, Sikora M, Sjogren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlstrom T, Vinner L, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522:167–72.
33. Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun. 2012;3:698.
34. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szécsényi-Nagy A, Mittnik A. The Beaker phenomenon and the genomic transformation of Northwest Europe. Nature. 2018;555:190.
35. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
36. Vohr SH, Najar CFBA, Shapiro B, Green RE. A method for positive forensic identification of samples from extremely low-coverage sequence data. BMC Genomics. 2015;16:1034.
37. Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, Olalde I, Broomandkhoshbacht N, Candilio F, Cheronet O. The genomic history of southeastern Europe. Nature. 2018;555:197.
38. de Barros DP, Marchi N, Rasmussen S, Peyrot M, Renaud G, Korneliussen T, Moreno-Mayar JV, Pedersen MW, Goldberg A, Usmanova E. 137 ancient human genomes from across the Eurasian steppes. Nature. 2018;557:369.
39. Durvasula A, Hoffman PJ, Kent TV, Liu C, Kono TJ, Morrell PL, Ross-Ibarra J. ANGSD-wrapper: utilities for analysing next-generation sequencing data. Mol Ecol Resour. 2016;16:1449–54.
40. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PL, Aximu-Petri A, Prufer K, de Filippo C, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514:445–9.
41. Nakatsuka, Nathan; Harney, Eadaoin; Mallick, Swapan; Mah, Matthew; Patterson, Nick; Reich, David. Estimation of ancient nuclear DNA contamination using breakdown of linkage disequilibrium. Github. 2020. https://github.com/nathan-nakatsuka/ContamLD.
42. Nakatsuka, Nathan; Harney, Eadaoin; Mallick, Swapan; Mah, Matthew; Patterson, Nick; Reich, David. Estimation of ancient nuclear DNA contamination using breakdown of linkage disequilibrium. Zenodo. https://doi.org/10.5281/zenodo.3736774. (2020).

## Publisher's Note