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,†]

[1]Department of Genetics, Harvard Medical School, New Research Building, 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

*co-first authors

[†] Corresponding authors: Nathan Nakatsuka (nathan_nakatsuka@hms.harvard.edu), Éadaoin Harney (harney@g.harvard.edu), and David Reich (reich@genetics.med.harvard.edu)

Abstract

We report a method, *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, leveraging 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 (e.g. less than 1.5% standard error in cases with <10% contamination and data from at least 500,000 sequences covering SNPs). This method is optimized for application to aDNA, leveraging characteristic aDNA damage patterns to provide calibrated contamination estimates. Availability: 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 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 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 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 typically carry characteristic damage. In addition, 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]. Another reliable set of methods for estimating rates of contamination in ancient DNA leverage polymorphism on the X chromosome in males, including the popular *ANGSD* method [6-9] and an improved methodology that enhances power for low-coverage samples [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

87 methods generally rely on access to uncontaminated genotype data from the individual of
88 interest or access to all possible contaminating individuals, neither of which is typically available
89 for aDNA. Another method estimated modern human autosomal contamination in aDNA from
90 archaic Denisovans [13] and Neanderthals [14] by producing maximum likelihood co-estimation
91 of sequence error, contamination, and parameters correlated with divergence and
92 heterozygosity. However, this method heavily relies on the significant divergence between
93 archaic and modern humans. A similar method, *DICE*, expanded on this method and jointly
94 estimates contamination rate and error rate along with demographic history based on allele
95 frequency correlation patterns [15]. However, this method requires both explicit demographic
96 modeling and high genome coverage. While this may be effective for estimation of
97 contamination in archaic genomes like Neanderthals and Denisovans that are highly genetically
98 diverged from likely contaminant individuals, it is not optimized for study of contamination
99 among closely related present-day human groups with complex demographic relationships
100 relative to each other, or contamination from individuals of the same population. In Racimo *et al.*
101 2016 [15], *DICE* required over 3x genome sequence coverage and solved the distinctive
102 problem of measuring contamination of present-day humans in a Neanderthal genome.
103
104 We report a method for estimating autosomal aDNA contamination using patterns of linkage
105 disequilibrium (LD) within a sample. This approach, implemented in our software *ContamLD*, is
106 based on the idea that when sequences from one or more contaminating individuals are present
107 in a sample, LD among sequences derived from that sample is expected to be diminished,
108 because the contaminant DNA derives from different haplotypes and therefore should have no
109 LD with the authentic DNA of the ancient individual of interest. Thus, the goal of the algorithm is
110 to determine the LD pattern the ancient individual would have had without contamination and
111 compare it to the LD pattern found in the sample. The LD patterns of ancient individuals are
112 determined using reference panels from 1000 Genomes Project populations to compute

113 approximate background haplotype frequencies, where haplotypes are defined as pairs of SNPs
114 with high correlation to each other. Contamination is then estimated by fitting a maximum
115 likelihood model of a mixture of haplotypes from an uncontaminated individual and a proportion
116 of contamination (to be estimated from the data) from an unrelated individual. *ContamLD*
117 corrects for mismatch of the ancestry of the ancient individual with the reference panels using
118 two different user-specified options. In the first option, mismatch is corrected using estimates
119 from damaged sequences (which, ideally, lack present-day contaminants). In the second option,
120 *ContamLD* performs an "external" correction by subtracting the sample's contamination estimate
121 from estimates for individuals of the same population believed to have negligible contamination
122 (the user could obtain this value from a *ContamLD* calculation on a male individual with a very
123 low estimate of contamination based on *ANGSD*). The second option has more power than the
124 first option and allows detection of cross-contamination by other ancient samples, but it could be
125 biased if a reliable estimate from an un-contaminated individual from the same population is not
126 available for the external correction.
127
128 We show that *ContamLD* accurately infers contamination in both ancient and present-day
129 individuals of widely divergent ancestries with simulated contamination coming from individuals
130 of different ancestries. The contamination estimates are highly correlated with estimates based
131 on X chromosome analysis in ancient samples that are male, as assessed using *ANGSD* [16].
132 *ContamLD* run with the first option has standard errors less than 1.5% in samples with at least
133 500,000 sequences covering SNPs (~0.5x coverage for data produced by in-solution
134 enrichment for ~1.2 million SNPs [2, 17], or ~0.1x coverage for data produced using whole-
135 genome shotgun sequences). With the second option, *ContamLD* has standard errors less than
136 0.5% in these situations, allowing users to detect samples with 5% or more contamination with
137 high confidence so they can be removed from subsequent analyses.
138

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 1.02x 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 Methods). 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 reagent [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 results from a larger panel (~5.6 million SNPs) provided with the software that we recommend for shotgun sequenced samples, and which provides more power to measure contamination.

165 We first analyzed data generated using a reference individual from the 1000 Genomes CEU
166 population (Utah Residents (CEPH) with Northern and Western European Ancestry) and the
167 SNP coverage profile of a 1.02x coverage ancient individual of West Eurasian ancestry (Iberian
168 Bronze individual I3756 who lived 2014-1781 calBCE; see Methods). Supplementary Figure 1
169 illustrates the distribution of LOD (logarithm of the odds) scores generated when *ContamLD* is
170 run on samples with 0%, 7% and 15% simulated contamination. Supplementary Figure 2 shows
171 the contamination rate estimates generated for data with simulated contamination rates
172 between 0 to 40%. At very high contamination (above 15%) *ContamLD* often overestimates
173 contamination, but in practice samples with above 10% contamination are generally removed
174 from population genetic analyses, so inaccuracies in the estimates at these levels are not a
175 concern in our view (the importance of a contamination estimate in many cases is to flag
176 problematic samples, not to accurately estimate the contamination proportion). *ContamLD*
177 assumes that the individual making up the majority of the sequences is the base individual, so
178 we do not explore contamination rates greater than 50% in these simulation studies.
179
180 We observe a linear shift in the contamination estimates such that most estimates are biased to
181 be slightly higher than the actual value, with even greater overestimates occurring at higher
182 contamination rates (Supplementary Figure 2). This is likely due to the difference between the
183 haplotype distribution of the test individual and that of the haplotype panel, since the magnitude
184 of this shift increases as the test individual increases in genetic distance from the haplotype
185 panel. Even in cases where the test individual is of the same ancestry as the haplotype panel
186 (as in Supplementary Figure 2) there is expected to be a shift, because the test individual's
187 haplotypes are a particular sampling of the population's haplotypes, and the difference between
188 having only frequencies of the haplotype panel and a particular instantiation of those
189 frequencies in the test individual will lead to the artificial need for an external source
190 ("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 inbred individuals, as expected because *ContamLD* assumes the paternal and maternal copy 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. If a reliable methodology for quantifying the proportion of the genome that is affected by inbreeding in ancient individuals becomes available, *ContamLD* could be further improved by using this information 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 under-estimated, 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.

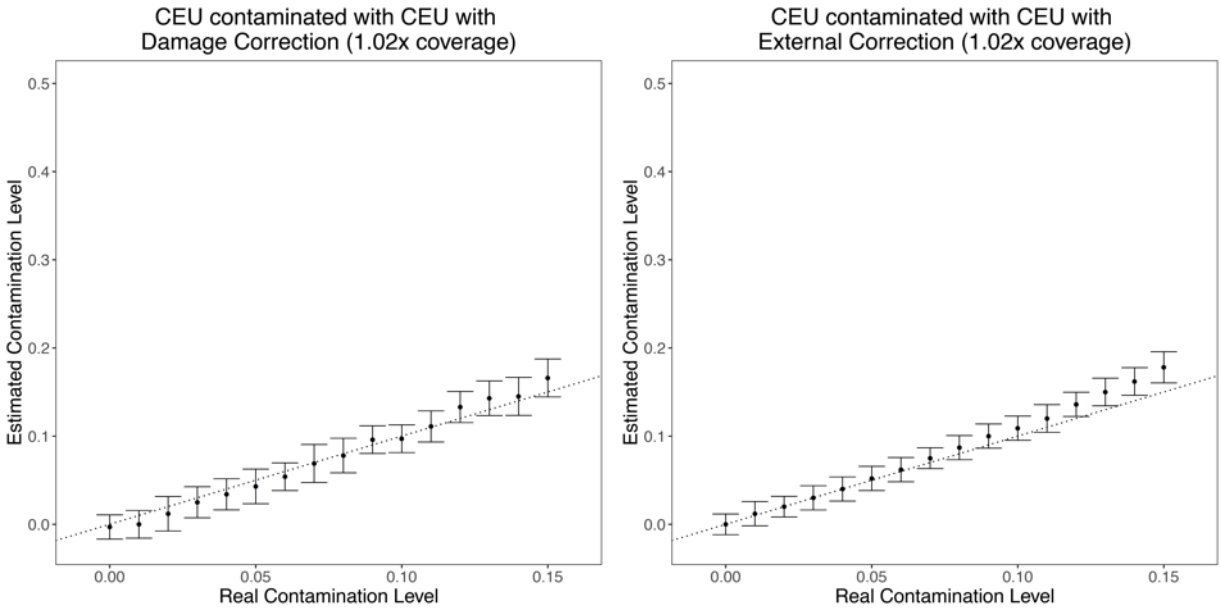
In our implementation, we correct for these systematic biases in two ways, implemented as different options in *ContamLD*.

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

217 authentically ancient and not derived from a contaminating source (assumed to be from present-
218 day individuals), so the *ContamLD* estimate based on undamaged sequences is corrected by
219 estimates based on the damaged sequences (see Methods for more details). In the second
220 option, we allow the user to subtract the contamination estimate from the estimate of an
221 individual of the same ancestry assumed to be uncontaminated. An advantage of the second
222 option compared to the first is that it has smaller standard errors (Figure 1), reflecting the fact
223 that it does not rely on estimates from damaged sequences (reliance on damaged sequences
224 reduces power since it often reflects a very small subset of the data). A second advantage of
225 the second option is that it allows estimation of contamination in cases where the source of
226 contamination is also ancient in origin, as would be expected if the contamination occurred
227 anciently or due to cross contamination with other ancient samples (the first option would be
228 expected to produce an underestimate of contamination in such cases, since it assumes that
229 sequences that contain C-to-T damage are not contaminated). On the other hand, a drawback
230 of the second option is that it requires users to identify a relatively high coverage,
231 uncontaminated, ancestry-matched sampled for benchmarking purposes; the method is also
232 only expected to work if there is minimal inbreeding in either the sample of interest or the
233 matched sample. Identifying such benchmarking samples may be impossible when analyzing
234 samples from previously unsampled contexts (e.g. early modern humans), and indeed verifying
235 that a benchmarking sample is uncontaminated is very difficult if it is female (if it is male a
236 method like *ANGSD* can be used).
237
238 In what follows, we report results of analyses based on the first option, but *ContamLD* includes
239 both methods as options. The uncorrected score also forms the basis for warning output by the
240 software, namely high contamination or possible contamination with another ancient sample
241 leading to an inaccurate damage correction estimate.
242

243

244 **A)** **B)**



245

246 **Figure 1. *ContamLD* estimates when the target individual, contaminant, and haplotype panel are**
247 **from the CEU population.** Contamination estimates when the simulated contamination rate is between
248 0.00-0.15. **A)** Estimates with damage restricted correction (option 1). **B)** Estimates with external
249 correction from an uncontaminated sample (option 2). The black dotted line is y=x, which would
250 correspond to a perfect estimate of contamination. Error bars are 1.96*standard error (95% confidence
251 interval determined via jackknife resampling across chromosomes).

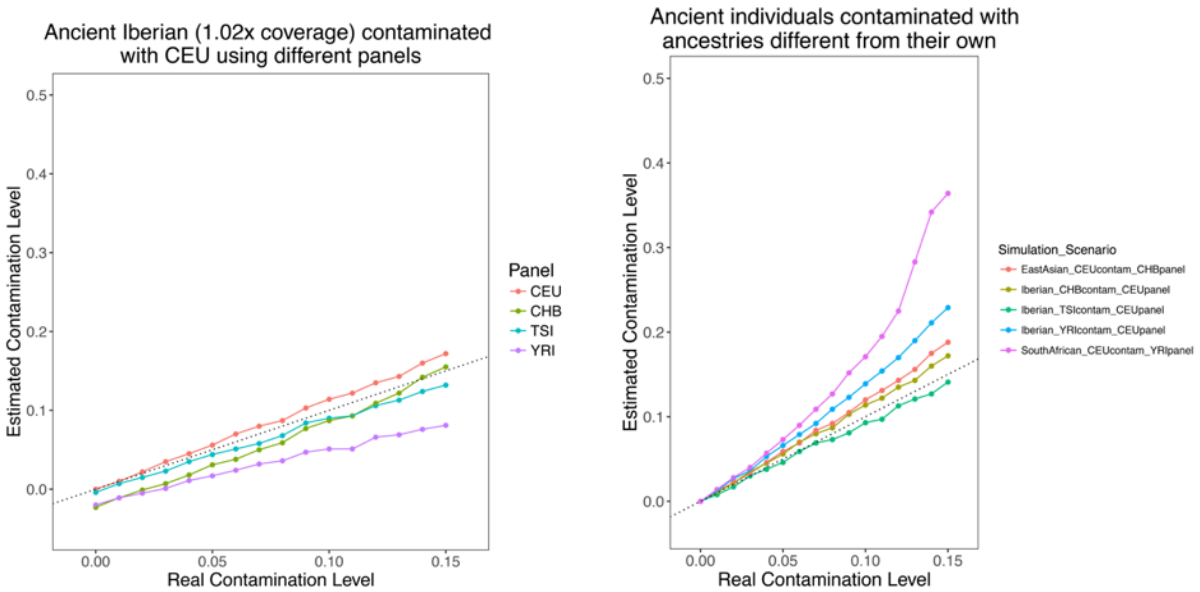
252

253 *Simulated Contamination of Ancient Samples with Present-Day Samples:*
254 *ContamLD* is designed to work on ancient individuals, so we simulated contamination of real
255 ancient individuals with present-day individuals from the 1000 Genomes Project, a scenario that
256 would occur when skeletal material from ancient individuals is contaminated by present-day
257 individuals during excavation or at some point during the processing of the material. We used
258 data from male individuals selected due to very low X chromosome contamination estimates
259 (less than 1%) based on *ANGSD* [16] (developed first in Rasmussen *et al*. [9]; we used method
260 1 of that software). (We subtracted the *ANGSD* estimates from the *ContamLD* estimates to

261 correct for any residual contamination.) Figure 2A shows results from the Iberian Bronze Age
262 sample [18] (I3756) with 1.02x coverage at the targeted ~1.24 million SNP positions,
263 demonstrating that *ContamLD* produces highly accurate contamination estimates for this
264 simulation.
265
266 *Effect of Different Haplotype Panels*
267 There are many potential cases in which ancient individuals can come from populations with
268 very different genetic profiles compared to present-day 1000 Genomes populations, leading to
269 an ancestry mismatch to the haplotype reference panels. *ContamLD* provides panels from all
270 1000 Genomes populations as well as tools to identify the panel most closely matching to the
271 ancestry of their ancient individual (based on outgroup-f_3 statistics [19] to determine the most
272 shared genetic drift), which they can then select for the analysis. However, due to the potential
273 for ancestry mismatch to still occur, we tested the effect of choosing haplotype panels that are
274 genetically diverged from the individual of interest (Figure 2A). For the ancient Iberian sample,
275 the CEU and TSI (Toscani in Italia) panels—representing northern and southern European
276 ancestry, respectively—yielded contamination estimates that are close to the true contamination
277 rate, especially for rates below 5%. However, *ContamLD* underestimates contamination by ~2%
278 when the CHB (Han Chinese in Beijing, China) and YRI (Yoruba in Ibadan, Nigeria) panels were
279 used instead (though we view these as very pessimistic cases, because the user should usually
280 be able to choose a panel more closely related to their ancient individual than these scenarios).
281 We thus recommend that users take care to choose an appropriate panel that is within the same
282 continental ancestry as their ancient individual. Nevertheless, we note that we were able to
283 obtain reasonably accurate estimates for Upper Paleolithic European hunter-gatherers, such as
284 the Kostenki14 individual [20], who is ~37,470 years old, even when using present-day
285 European panels that have significantly different ancestry from the hunter-gatherers
286 (Supplementary Figure 3).

287

288 **A)** **B)**



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

298

299 *Effect of Mismatch Between the Ancestry of the True Sample and Contaminating Individual*
300 Contamination can come from a wide variety of sources, including, but not limited to, members
301 of the archaeological excavation team, the aDNA laboratory, or residual human DNA on the
302 plastic and glassware or in laboratory reagents. Thus, we sought to understand the effect of
303 mismatch in the ancestry of the true sample and the contaminating individual in our
304 contamination estimates. We found that as the ancestry of the two diverged, *ContamLD* over-

305 estimated contamination (Figure 2B and Supplementary Figure 4). This occurred when we
306 tested an ancient European with different contaminant ancestries and when we tested ancient
307 East Asian [21] and ancient South African [22] samples contaminated with European DNA.
308 Nevertheless, the over-estimation was not severe at contamination levels below 5 percent, and
309 samples above this proportion would likely be flagged as problematic. We also explored
310 scenarios where the ancestry of the panel matches the contaminant rather than the true sample
311 (Supplementary Figure 4) and found a ~2% under-estimate at low levels of contamination and
312 an over-estimate at high levels of contamination; these are modest effects and are unlikely to
313 change our qualitative assessment. When we tested the effect of having multiple contaminant
314 individuals (Supplementary Figure 5), we found only a slight over-estimate at higher levels of
315 contamination, as expected given *ContamLD* normally assumes contamination from a single
316 individual where the haplotypes are re-formed if they are created from two contaminant reads
317 (which will happen at lower rates with more contaminant individuals).
318
319 *Estimating Contamination in Admixed Individuals*
320 *ContamLD* relies on measuring the difference between the LD pattern of the sample and that
321 expected from an uncontaminated individual. However, individuals from groups recently
322 admixed between two highly divergent ancestral groups have LD patterns that are similar in
323 some ways to that of an unadmixed individual with contamination from a group with ancestry
324 diverged from that of the individual of interest. To understand how this would impact *ContamLD*,
325 we ran the software on an ASW (Americans of African Ancestry in Southwest USA) individual
326 with different levels of added CEU contamination. When we ran *ContamLD* with a YRI panel
327 and no correction on an individual with no contamination, the individual was inferred to have a
328 contamination of ~20% (likely because the individual had ~15% European ancestry, and this
329 was interpreted by the software as contamination). Using an ASW panel did not perform any
330 better. However, the concerns were mostly addressed by the damage-restricted correction

331 (option 1) at low contamination levels (Supplementary Figure 6). The simulation with African-
332 Americans represents an extreme of difficulty, because the individual is from a group with very
333 recent admixture (~6 generations [23]) of ancestries highly divergent from each other with one
334 of the ancestries very genetically similar to the reference panel. It highlights how the damage-
335 restricted correction is still able to produce accurate estimates in these difficult cases.
336
337 *Effect of Coverage:*
338 We tested the power of our procedure at different coverages with simulations of ancient West
339 Eurasian ancestry individuals contaminated with CEU on the 1240K SNP set (Figure 3). We
340 found that while our estimates were not biased to produce estimates consistently above or
341 below the true value, the standard errors increased significantly at lower coverages, as
342 expected for the decreased power for accurate estimation in these scenarios. We provide a
343 much larger panel with ~5.6 million SNPs (vs. ~1.1 million for the 1240K panel) that usually
344 decreases standard errors for samples that are shotgun sequenced (Supplementary Figure 7).
345 This panel increases *ContamLD*'s compute time and memory requirements, so we recommend
346 that it only be used for individuals with lower than 0.5x coverage. As an additional feature, we
347 provide users tools to create their own panels to meet their specific needs.

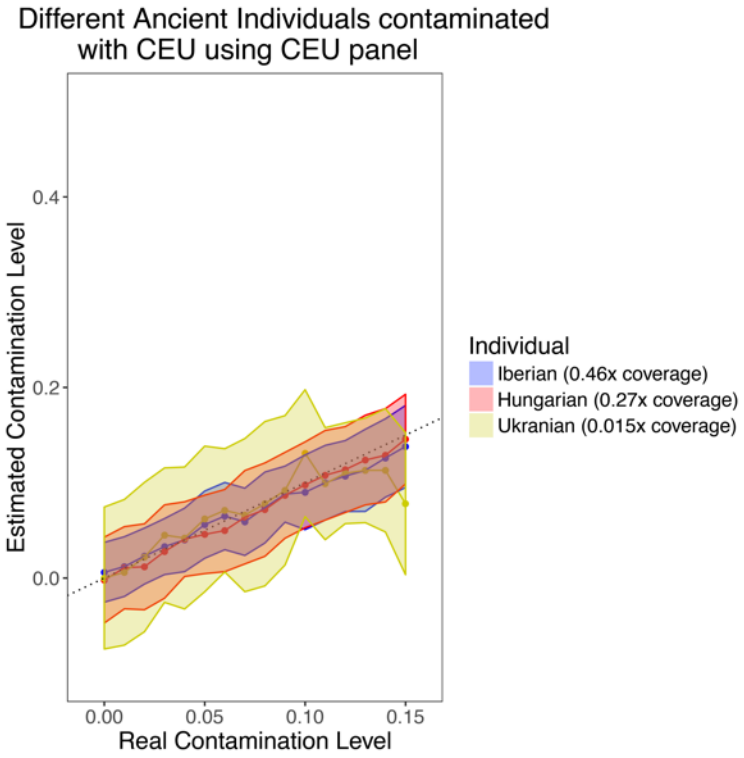


Figure 3. *ContamLD* estimates for ancient European samples of different coverages after damage restricted correction (option 1). An ancient Iberian of 0.46x coverage, an ancient Hungarian of 0.27x coverage, and an ancient Ukranian of 0.015x coverage (~16,000 snps) were contaminated with CEU and analyzed 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).

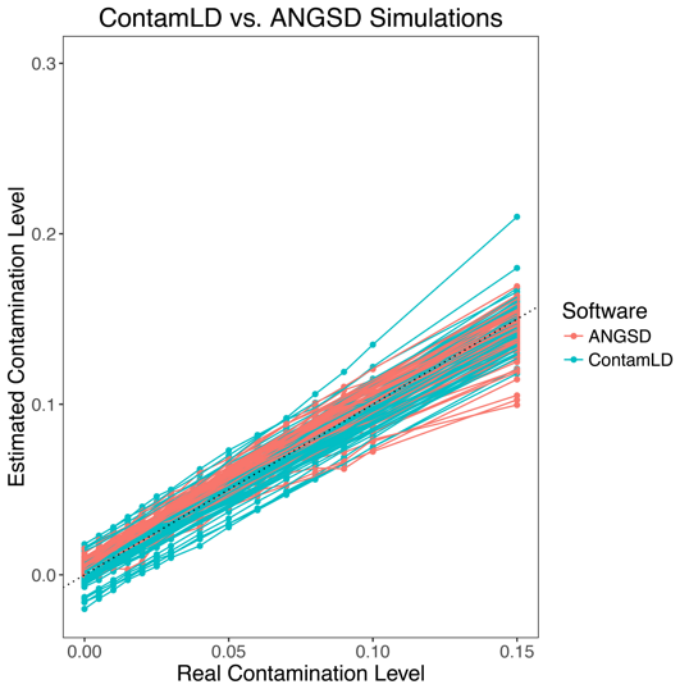
Simulations to Compare ContamLD to ANGSD X Chromosome Estimates

We performed simulations where we randomly added 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 Methods). We chose ancient male individuals that had average coverage over 0.5X 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 Methods). We then analyzed the individuals with *ContamLD* and

363 *ANGSD* and found that compared to *ANGSD*, *ContamLD* consistently had similar errors relative

364 to the true contamination level (Figure 4, Supplementary Online Table 2).

365



366

367 **Figure 4. Contamination estimates with *ContamLD* and *ANGSD* for ancient individuals with**

368 **different levels of contamination added.** 65 ancient individuals with average coverage over 0.5X had

369 increasing levels of artificial contamination added in (from I10895, an ~1200BP ancient West Eurasian

370 individual) and were then analyzed with *ContamLD* (with panels most genetically similar to the ancient

371 individual and using damage restricted correction, option 1) and *ANGSD*. Details of all estimates

372 (including standard errors) are provided in Supplementary Online Table 2. The black dotted line is y=x,

373 which would correspond to a perfect estimate of the contamination.

374

375 *Comparing ContamLD, ANGSD, and Mitochondrial Estimates (ContamMix) in Ancient*

376 *Individuals without Added Contamination*

377 We tested 439 ancient males with *ContamLD*, *ANGSD* (X chromosome contamination

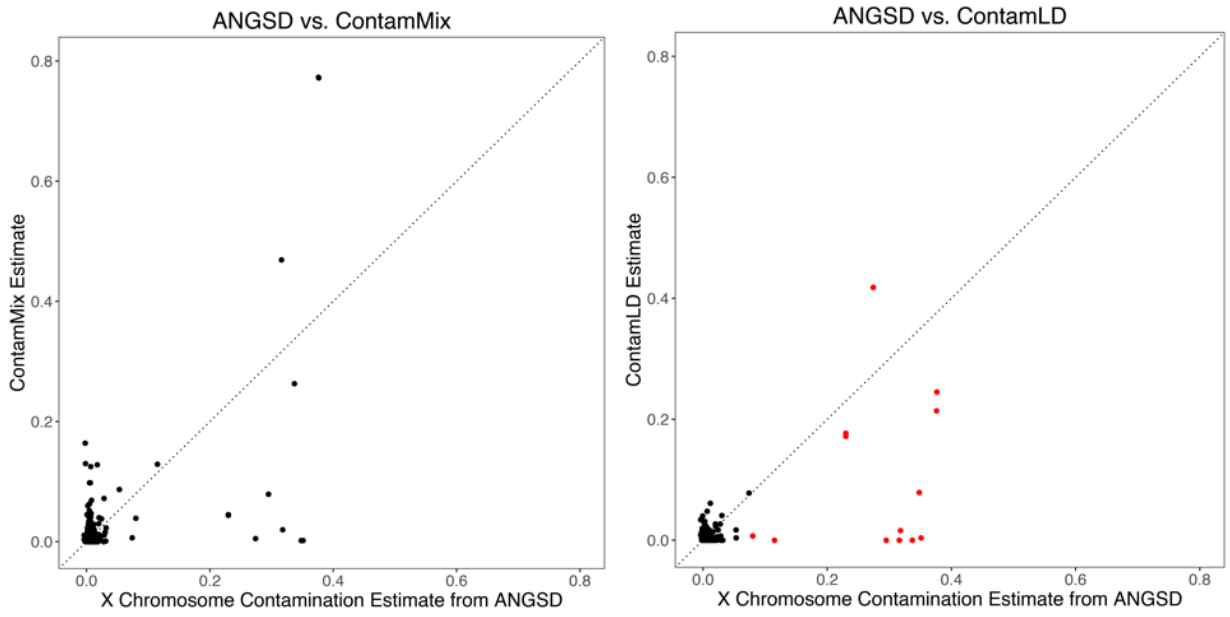
378 estimates), and *ContamMix* (mitochondrial contamination estimates) without adding additional

379 contamination. For this analysis, we included published data generated with the ~1.24 million

380 SNP enrichment reagent, as well as data from libraries that failed quality control due to
381 evidence of contamination (Supplementary Online Table 3). Similar to prior studies [5], the
382 mitochondrial estimates often differed from the nuclear (*ANGSD* and *ContamLD*) estimates,
383 showing high contamination in some libraries with low nuclear contamination, and low
384 mitochondrial contamination in some libraries with high nuclear contamination (Figure 5A). In
385 contrast, *ANGSD* and *ContamLD* had better concordance. However, we observed that some of
386 the samples with high contamination estimates based on *ANGSD* had much lower *ContamLD*
387 estimates, reflecting over-correction from analyzing the damaged sequences, perhaps because
388 the contamination was actually cross-contamination from other ancient individuals, violating the
389 assumptions of our damage-correction (Figure 5B). This problem was mitigated in part,
390 however, because *ContamLD* produces a warning of "Very_High_Contamination" if the
391 uncorrected estimate is above 15% (even in cases where the corrected estimate is very low),
392 and all samples with X chromosome estimates over 5% were flagged with this warning and/or
393 had estimates of over 5% contamination with *ContamLD* (all samples with less than 5%
394 contamination in *ANGSD* had lower than 5% contamination with *ContamLD*). It is unfortunately
395 not possible to know the true contamination of the samples we tested in Figure 5, but the fact
396 that our software produced results with good correlation to X chromosome estimates shows that
397 it works well in real ancient data.

398 It is possible for there to be samples with moderately high contamination from another
399 ancient individual but both a low damage restricted correction estimate and no warning
400 generated, because these would have high uncorrected estimates, yet not high enough to reach
401 the threshold required for the warning. These samples would have to be identified with an
402 external correction. Lowering the threshold for the "Very_High_Contamination" warning would
403 produce too many false positives, because there are many cases with high uncorrected
404 estimates that have low corrected estimates that are likely not contaminated (e.g. due to
405 ancestry mismatches of the panel and the test individual). To understand these issues better,

406 we performed a simulation in which an ancient Iberian (I3756) was contaminated with another
407 ancient West Eurasian individual (I10895) and the damaged sequences were set to be a 5%
408 down-sampling of the set of contaminated sequences (thus simulating a case in which all of the
409 contamination is from another ancient individual who has the same damage proportion as the
410 ancient individual of interest). We found that, as expected, the contamination from the ancient
411 individual was not detected (the contamination estimates were always near 0%) by the damage
412 restricted correction version of *ContamLD* until the contamination reached 15% at which point
413 the "Very_High_Contamination" flag came up (Supplementary Figure 9). The contamination
414 would have been detected with the external correction version of *ContamLD* (since the damage
415 restricted correction continued to go up with increasing contamination; see Supplementary
416 Online Table 4), but without an uncontaminated ancient individual of the same group as the
417 target individual, this would be difficult to do without bias in the contamination estimate.
418
419 **(A)** **(B)**



420
421 **Figure 5. Contamination estimates from *ContamLD*, *ANGSD*, and *ContamMix* in 439 ancient**
422 **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.

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 0.5X 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 0.1x. *ContamLD* is best suited to scenarios in which the contaminant and the ancient individual of interest are similar ancestry, which is useful, because *DICE* [15] and many population genetic tools (e.g. PCA or ADMIXTURE [24]) 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

448 overestimated, particularly for higher contamination. This should not be a large problem in
449 analyses of real (i.e. non-simulated) data, because the effect is small at the contamination levels
450 of interest (<5%). When we varied haplotype panels, we found that the estimator is robust when
451 applied to simulated datasets using haplotype panels that are moderately divergent from the
452 base sample (within-continent levels of variation). We provide users tools for automatically
453 determining the panel that shared the most genetic drift with the sample so that the user can
454 select the panel most closely related to the sample. In other simulations, we found that the
455 performance of the algorithm declines as the coverage of the sample decreases. The estimates
456 are not biased, the standard errors substantially increase when fewer than 300,000 sequences
457 are available. In these cases, if the individual was shotgun sequenced, we recommend that
458 users choose the shotgun panel, which will substantially increase power for the analyses.
459
460 We applied the algorithm to estimate contamination levels in dozens of ancient samples and
461 compared them to X chromosome-based contamination estimates. There was generally good
462 correlation with the X chromosome estimates, except that when the true contamination was very
463 high, the LD based estimates were sometimes estimated incorrectly, likely because the
464 contamination was due to cross-contamination from another ancient individual and there was
465 over-correction from the damage estimates. This problem is mitigated, however, because the
466 software indicates if the uncorrected estimate is very high so users can identify highly
467 contaminated samples and remove them from further analyses. A difficult case for the software
468 is if there is contamination in part from another ancient sample. This can cause an over-
469 correction and lead to an under-estimate of the contamination. The “Very_High_Contamination”
470 warning catches very high contamination from other ancient samples, but it will miss cases of
471 moderate levels of contamination from other ancient samples, because it will not reach the
472 threshold required for the warning. In theory, the user can determine the true contamination in
473 these cases using the external correction, but the external correction can be difficult if the user

474 does not have an adequate sample to correct the estimate of the sample of interest. The
475 damage correction of the software also does not work if the samples have undergone full UDG
476 treatment (no damaged sequences), and for this case, the external correction is the only option.
477
478 The software run-time is dependent on SNP coverage. If ~1,000,000 SNPs are covered (the
479 depth of the coverage on each SNP does not affect run-time), the analysis takes approximately
480 2 hours if 3 cores are available on CentOS 7.2.15 Linux machines (~25 GB of memory). The
481 software is designed for samples to be run in parallel, so the total time for analysis even for
482 large numbers of samples is often not much greater than the time for a single sample.
483
484 In summary, *ContamLD* is able to estimate autosomal nuclear contamination in ancient DNA
485 accurately with standard errors that depend on the coverage of the sample. This will be
486 particularly useful for female samples where X chromosome estimates are not possible. As a
487 general recommendation for users, we believe in most cases all samples with a contamination
488 estimate that is greater than 0.05 (5%) should be removed from further analyses, or the
489 contamination should be explicitly modeled in population genetic analyses.
490

Supplementary Data:

The Supplementary Data include 4 Excel spreadsheets detailing all ancient samples used and the contamination estimates for this algorithm. Also included are 10 supplementary figures.

Materials and Methods

Datasets:

Present-day samples:

Genome wide datasets from individuals that were part of the 1000 Genomes Project [25] were used as present-day reference samples. We restricted to autosomal sites included in the aDNA ~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 [25]. However, the software allows users to make panels based on their own SNP set. 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, 26] from ancient individuals from previous studies generated by shotgun sequencing or targeted enrichment with 1.24 million SNP enrichment, including many samples that failed quality control due to contamination but were from the same archaeological sites [2, 22, 27-33]. Information about the ancient individual data are detailed in Supplementary Online Table 1 and below.

516

517 *Obtaining sequence information:*

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

525

526 **Haplotype Calculation**

527

528 To create haplotype panels, we obtained all SNP pairs in high LD for each 1000 Genomes
529 population using PLINK version 1.9 [34] with r^2 cut-off of 0.2. (Users can increase power slightly
530 at the expense of increased computational time by creating their own haplotype panel with a
531 lower r^2 cutoff.) We then calculated the frequencies of each SNP in all of these pairs as well as
532 the haplotype frequencies at each of these pairs while holding out the present-day individuals
533 used for contamination simulation.

534

535 **Algorithm to Estimate Contamination**

536

537 *Overview:*

538 Our goal is to estimate α, the level of contamination, by examining the frequencies of allele pairs
539 that should be in LD (we term this two-allele pair a haplotype) and determining how much their
540 frequencies differ from what would be expected under no contamination. To estimate this, we
541 need both the distribution underlying the haplotypes (q) that an uncontaminated test sample
542 should have as well as the distribution of "unrelated haplotypes" ($\tilde{h}$) that would form by chance

543 from background allele frequencies. Here and below, “distribution” refers to the set of
544 frequencies of the different possible haplotypes (all possible combinations of ancestral and
545 derived alleles at the SNP pair) across all haplotypes in the genome. Supplementary Figure 10
546 is a schematic of the algorithm.
547
548 *Determining Haplotype Distributions Based on Reference Panels:*
549 To determine *q* we must account for the fact that the test individual’s genotypes do not have
550 diploid calls and are not phased. Due to the low sequence depths at each SNP in many ancient
551 DNA datasets, it is difficult to make confident heterozygous calls, so instead we create pseudo-
552 haploid calls by randomly choosing a sequence to represent the genotype at that position (this
553 holds when we are using genotype calls or the sequence information directly, and when multiple
554 sequences cover the same SNP, we use all of them and treat them as independent). Thus, for
555 this analysis, when examining a pair of SNPs, it is equally likely for the SNP pair to have been
556 formed from the true haplotype (if the same parental chromosome is sampled from in both
557 SNPs of the haplotype) or the background distribution (if the opposite parental chromosome is
558 sampled from). We therefore can estimate *q* as:

559 $$q = h/2 + \widetilde{h}/2$$

560 where *h* is the distribution of true haplotypes and $\tilde{h}$ is the distribution of unrelated haplotypes
561 that would form by chance from background allele frequencies. For inbred samples, the weight
562 on *h* is more than 1/2, because the two parental chromosomes are more related, but this can
563 generally be corrected (see below).
564
565 $\tilde{h}$ can be determined by multiplying the SNP frequencies to obtain the haplotype frequencies
566 that would form after randomly pairing SNPs of unrelated individuals. *h* can be estimated from
567 an external reference panel using a maximum likelihood estimator (MLE) to obtain haplotypes

frequencies in the population from the counts (necessary because the panels are not phased). The MLE set-up is:

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

with:

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

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.

593 The MLE would be computationally intractable to solve due to our lack of knowledge of which
594 parent contributed to each count, so we instead used an EM algorithm to obtain *h*, where
595 knowledge of the parents' contribution is the unobserved latent variable. The algorithm involves
596 an expectation step of:
597

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

598
599
600 where n_1 is the expected number of times that the (*a, b*) configuration of the father's
601 chromosome contributed to a particular diploid count (this is the same value for the mother, n_2,
602 because they are assumed to be from the same haplotype distribution). In other words, given
603 the observed haplotype counts in the reference panel, how many times would it be expected
604 that a particular haplotype configuration (e.g. ancestral at SNP1, derived at SNP2) in one of the
605 parents contributed to those counts?
606
607 Once the counts (n_1 and n_2) of the haploid parents are obtained, they are added together to
608 produce the diploid individual (i.e. the expected number of all possible haplotype
609 configurations). Then the expected value of the haplotype distribution can be maximized by
610 averaging over the possible haplotype distributions. Thus, the maximization step is:

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

611

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

612
613
614 where *D*(*a,b*) is the sum of the probabilities of a particular haplotype configuration over all
615 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 α, we used the equation:

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

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 (α', where $'$ is used to indicate that this is an estimate of the real α). 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 α'^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

640 breaking up LD, so a fraction $2\alpha'(1 - \alpha')$ of the distribution should look like this (the probability

641 that the SNP pair is formed from a contaminated sequence and an uncontaminated sequence).

642

643 This expression can be used to solve for α' by maximizing the LOD (log of the odds) scores

644 under the null hypothesis that $\alpha' = 0$ and the alternative hypotheses of different α'. A LOD score

645 is assigned to each estimate of the contamination rate (α) between -0.1 to 0.5 (negative scores

646 are included to allow correction for inbreeding). The grid of α' is scaled by intervals of 0.0001.

647 The α' with the highest LOD score is the best estimate of α, and is returned. When we have

648 multiple sequences on the same SNP we assume independence of the sequences, which

649 provides additional power. The assumption of independence does not bias the error estimation

650 for the same reason as explained above for independence of SNP pairs.

651

652 *Correcting for Bias in Contamination Estimates:*

653 In practice, the α' we obtain is not equal to the true α, because the reference panel does not

654 perfectly capture the SNP and haplotype frequencies of the test sample. We found that this

655 difference causes a linear shift in contamination estimate where the mismatch between the

656 sample individual and the reference panel leads to a positive shift while inbreeding leads to a

657 negative shift. These biases can be addressed in either of two ways.

658

659 First, for the "damage correction" approach, we performed an α' estimate only on alleles from

660 sequences with evidence of damage characteristic of ancient samples. Under the assumption

661 that these sequences are not affected by present-day contamination, the inferred α' would be an

662 estimate of the bias, which can be subtracted out from the estimate based on all sites. We

663 separately analyzed the following pairs of SNPs: UU (both SNPs at undamaged sequences),

664 DU (one site damaged and the other undamaged), and DD (both SNPs at damaged

665 sequences). For the UU pairs, the value we calculate would be α + k, where k is the linear shift.
666 For DU pairs the value calculated would be $\alpha/2$ + k, and for DD pairs the value calculated would
667 be k. We added the likelihoods for these pairs and maximized the likelihood to solve for α and k.
668 After solving for α, we multiply by (1-damage rate) to obtain the contamination level across all
669 sequences, because α is the contamination rate at undamaged sequences.
670
671 Second, for the "external correction" approach, we took individuals from the test individual's
672 population that were high coverage and samples we believed had very low contamination
673 (based on X chromosome estimates with *ANGSD* using method 1 as developed first by
674 Rasmussen *et al*. [9]) and measured α'. We assumed a true contamination of 0 for these
675 samples and thus subtracted this α' from all other contamination estimates. We caution that this
676 method does not correct for uncertainty in the contamination estimate in the external sample
677 used for benchmarking.
678
679 *Comparison to a Similar Method:*
680 The approach of *ContamLD* is similar to that of Vohr *et al*. [35] except the two have the opposite
681 goals. Vohr *et al*. searches for LD in reads from two different samples in an attempt to
682 determine whether the two samples are from the same individual (or closely related individuals),
683 using a reference panel to determine LD patterns. In contrast, *ContamLD* searches for breaks in
684 LD in the sequences of a single sample to determine if sequences from other individuals are
685 present in the sample.
686
687
688 **Data simulation:**
689

690 To test the accuracy of the algorithm, we applied it to a variety of scenarios with both present-
691 day DNA as well as real aDNA samples that had simulated present-day DNA contamination. In
692 all our simulations with 1000 Genomes individuals, we removed the individual being used from
693 our haplotype panel before performing the analyses.
694
695 *Simulating Contamination of Present-day Individuals:*
696 We first simulated contamination of present-day individuals with other present-day individuals as
697 contaminants (this allowed us to be sure that there was no baseline contamination). In order to
698 best approximate the distribution of both the damaged and undamaged sequences that is
699 characteristic of aDNA data, we used sequence-depth information from an ancient individual as
700 a reference. At each SNP, the total number of simulated “damaged” and “undamaged”
701 sequences was determined based on the number of damaged and undamaged sequences at
702 the SNP in the reference ancient individual. The identity of each allele for the present-day
703 “base” sample was randomly chosen based on the genotype of the “base” present-day 1000
704 Genomes individual at each SNP, as described above for the contamination. The addition of
705 contaminant sequences to the dataset was performed using the method described above. In
706 order to reduce bias caused by the damage correction procedure, the damage restricted dataset
707 was generated only once for each simulation type (which included multiple simulations across
708 varying contamination rates) and combined with the undamaged dataset to produce the overall
709 dataset. This method was used to generate a simulated individual using present-day CEU
710 (NA06985) or ASW (NA19625) from the 1000 Genomes dataset as the “base” sample from the
711 sequence distributions of a 1.02x coverage ancient Iberian individual (I3756) (the “reference”)
712 [18]. The CEU (NA06984) individual was used as “contaminant” in each case.
713
714 We generated simulated data with contamination from multiple sources by adjusting the
715 present-day contamination simulation method to randomly sample from two or more present-

716 day source contaminant genomes with equal probability. In each case, a 1000 Genomes Project
717 CEU individual (NA06985) was used as a “base” genome with the sequence distribution of
718 I3756 (the “reference”). In the case of 2 sources of contamination (Supplementary Figure 5), two
719 CEU individuals from the 1000 Genomes Project dataset (NA06984 and NA06986) were used
720 as contamination sources, and in the case of three contamination sources, an additional CEU
721 individual was used (NA06989). Data was generated for all combinations of undamaged
722 contamination rates, α, from 0-15%.
723
724 *Simulated contamination of ancient individuals:*
725 We performed two sets of simulations contaminating different ancient individuals. In both cases
726 we selected ancient male individuals with minimal contamination (as assessed by X
727 chromosome contamination levels from *ANGSD* [16]) to act as the “base” uncontaminated
728 genome. In the first simulation set, we tested *ContamLD*’s performance with different ancient
729 individuals and different present-day contaminant individuals from the 1000 Genomes dataset
730 [25] to assess the impact of contaminant ancestry and coverage of the ancient individual. In this
731 case we were only using *ContamLD* and thus we performed the simulated contamination on the
732 genotype level. In the second simulation set, we compared *ContamLD* to *ANGSD* and used a
733 ~1200BP ancient West Eurasian individual (I10895) to contaminate the BAM files directly.
734
735 In the first simulation set, we assumed that sequences with C-to-T damage are highly unlikely to
736 be the product of contamination (this assumption would be falsified in the context of cross-
737 contamination by another ancient DNA sample). Thus, we exclusively added contamination to
738 the “undamaged” fraction of sequences. At each SNP site, we classified sequences present in
739 the damage-restricted dataset as “damaged” and added to the simulated data. We classified all
740 other sequences as “undamaged” and also added them to the simulated data, but for each
741 “undamaged sequence” we added a contaminant sequence to the simulated SNP data with

742 probability $\alpha/(1-\alpha)$, where α is equal to the contamination rate (since the added sequences
743 contribute to the total number of sequences, we needed to add a higher proportion than the
744 contamination rate to obtain our desired contamination rate). The identity of the added
745 contaminant allele was randomly chosen based on the genotype of the chosen "contaminant"
746 present-day genome at the site (i.e. if the contaminant individual was homozygous at the site,
747 the allele it possesses would be added to the simulated individual, while if it were heterozygous
748 at the site, either the reference or alternative allele would be selected randomly and added to
749 the simulated individual). This method maintains the underlying distribution of "uncontaminated"
750 reference and alternative alleles at each SNP site, while adding additional "contaminant" alleles
751 to each site, producing an overall contamination rate of α in the undamaged sequences.
752 For each simulation, we generated two output files: (1) a file reporting the total number of
753 sequences carrying reference and alternative alleles at each SNP and (2) a damage restricted
754 file reporting the total number of damaged sequences carrying reference and alternative alleles
755 at each SNP. We used a 1.02x coverage ancient Iberian individual (I3756) (Supplementary
756 Online Table 1) with contamination from either the 1000 Genomes CEU individual NA06984, the
757 TSI individual NA20502, the CHB individual NA18525, or the YRI individual NA18486. We also
758 used 5 other ancient individuals: I1845 (an ancient Iberian sample of 0.46x coverage) [18],
759 I2743 (an ancient Hungarian of 0.27x coverage) [30], I5891 (a Neolithic Ukrainian individual of
760 0.016x coverage) [36], DA362.SG (a Russian early Neolithic Shamanka East Asian individual of
761 1.10x coverage) [21], and I9028.SG (a South African individual of 1.21x coverage) [22]. In each
762 case, we simulated individuals with 0-15% contamination.
763
764 For the second simulation set, we analyzed 65 ancient individuals of average coverage over
765 0.5X and baseline *ANGSD* estimates under 2% (Supplementary Online Table 2). In these
766 cases, we added artificial contamination with sequences from a ~1200BP ancient West
767 Eurasian individual (I10895) into the BAM files at the amounts: (0.000, 0.005, 0.010, 0.020,

768 0.025, 0.030, 0.040, 0.050, 0.060, 0.070, 0.080, 0.090, 0.100, 0.150). We removed two base
769 pairs from the end of each sequence of partial UDG treated samples and ten nucleotides for
770 non-UDG treated samples and pulled down the genotypes by randomly selecting a single
771 sequence at each site covered by at least one sequence in each individual to represent the
772 individual's genotype at that position ("pseudo-haploid" genotyping). To ensure that the damage
773 sequences were only from the non-contaminant individual (so that we could use the damage
774 restricted correction mode, option 1, of *ContamLD* without bias), we created the "damaged"
775 sequence set as a randomly chosen 5% of the sequences from the non-contaminant individual.
776 We then analyzed the data with *ContamLD* (damage restricted correction version, option 1) and
777 *ANGSD* using default settings (Method 1). We also performed simulations with a 1.0x coverage
778 ancient West Eurasian ancestry individual (DA57.SG, an ancient Krgyzstanian individual) [37]
779 down-sampled to 0.5x coverage and contaminated with I10895. To simulate different damage
780 rates, we varied the damage rate to the proportions (0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.075)
781 by setting the amount of "damaged" sequences to be those proportions.
782
783 As a last simulation, we examined the case of an ancient individual contaminating another
784 ancient individual where some of the damaged sequences would also come from the
785 contaminating individual. In this simulation, we analyzed a 1.02x coverage ancient Iberian
786 individual (I3756) and contaminated the BAM with sequences from a ~1200BP ancient West
787 Eurasian individual (I10895) in the proportions (0.000, 0.005, 0.010, 0.020, 0.025, 0.030, 0.040,
788 0.050, 0.060, 0.070, 0.080, 0.090, 0.100, 0.150, 0.200, 0.300). We then down-sampled the
789 BAM, taking a random 5% of the sequences of these contaminated BAM files to act as the
790 "damaged" sequences, because this would correct for any baseline contamination in the I3756
791 individual yet would simulate additional contamination of I3756 by an ancient individual with the
792 same damage rate as I3756 (i.e. if there is 5% contamination, then also 5% of the damaged
793 sequences would be from the contaminant individual in this simulation). We then performed the

794 standard processing of both the full contaminated BAMs and the 5% down-sampled BAMs
795 (simulated to be “damaged” sequences), removing two base pairs from the end of each
796 sequence and carrying out a “pseudo-haploid” genotype pulldown. We ran *ContamLD* on the
797 resulting data with damage restricted correction, option 1.
798
799 *Direct Analyses of Contamination Levels in Ancient Individuals:*
800 As our last set of analyses, we directly measured contamination levels in ancient individuals
801 without simulated contamination. We used *ContamLD* to examine shotgun sequenced
802 individuals analyzed at the1240K SNP set and the large 5.6 million SNP shotgun panel. The
803 ancient shotgun sequenced individuals were of 0.1-0.5x coverage from Allentoft *et al*., 2015
804 [31], Damgaard *et al.*, Nature 2018 [37], and Damgaard *et al*., Science 2018 [21]. In addition, we
805 analyzed 439 individuals from a variety of ancestries with *ContamLD* (damage corrected
806 version), *ANGSD* [16, 38] using default settings (we report the results from Method 1), and
807 *contamMix* [39] with the settings: down-sampling to 50X for samples above that coverage, --
808 trimBases X (2 bases for UDG-half samples and 10 bases for UDG-minus samples), 8 threads,
809 4 chains, and 2 copies, taking the first one that finishes. Supplementary Online Table 1 includes
810 all information from these individuals.
811

812 Declarations

813
814 **Ethics Approval and Consent to Participate**
815 Not applicable (all samples were from previously published studies).
816
817 **Consent for publication**

Not applicable.

Availability of Data and Materials:

All data analyzed in this article are available in [2, 21, 22, 27-33, 37]. The software is available at: https://github.com/nathan-nakatsuka/ContamLD. 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 (DOI: 10.5281/zenodo.3736774) Scripts for data simulations are available in the Github folder “Simulation_Scripts”.

Competing interests:

The authors declare that they have no competing interests.

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.

Authors’ Contributions:

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

Acknowledgements:

We thank Iosif Lazaridis and Mark Lipson for helpful discussions.

References

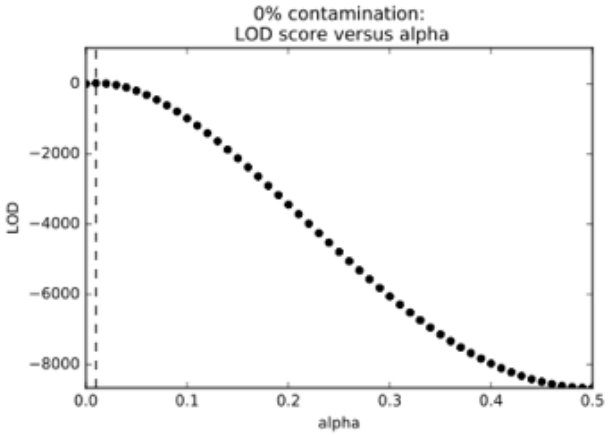
1. Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, Garcia N, Paabo S, Arsuaga JL, Meyer M: **Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments.** *Proc Natl Acad Sci U S A* 2013, **110:**15758-15763.
2. Haak 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:**207-211.
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 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:**2229-2234.
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:**15696-15700.
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:**2223-2227.
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:**416-426.
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:**710-722.
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:**94-98.
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:**828-841.
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:**2601-2602.
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:**839-848.
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:**222-226.
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:**43-49.
15. Racimo F, Renaud G, Slatkin M: **Joint estimation of contamination, error and demography for nuclear DNA from ancient humans.** *PLoS genetics* 2016, **12**.

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. 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:**1230-1234.
19. 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:**1065-1093.
20. 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:**1113-1118.
21. de Barros Damgaard P, 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.
22. 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.
23. 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.** *The American Journal of Human Genetics* 2004, **74:**979-1000.
24. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res* 2009, **19:**1655-1664.
25. 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.
26. 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.
27. 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:**409-413.
28. 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:**419-424.
29. 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:**214-218.
30. 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:**368-372.
31. 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:**167-172.
32. 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.

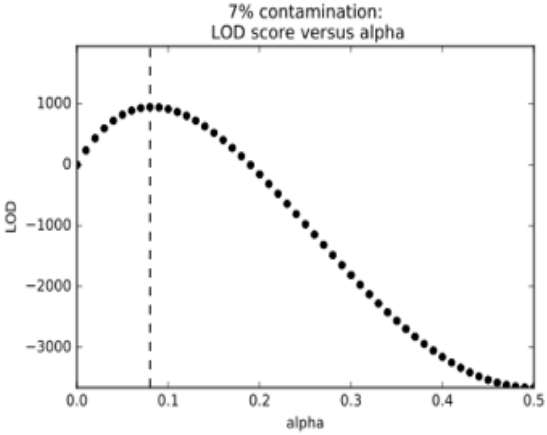
33. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szécsényi-Nagy A, Mittnik A: **The Beaker phenomenon and the genomic transformation of northwest Europe.** *Nature* 2018, **555:**190.
34. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller 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:**559-575.
35. 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.
36. 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.
37. de Barros Damgaard P, 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.
38. 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:**1449-1454.
39. 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:**445-449.

Supplementary Figures

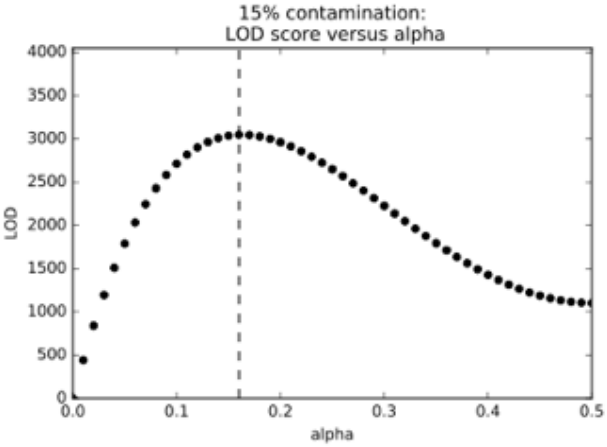
A)



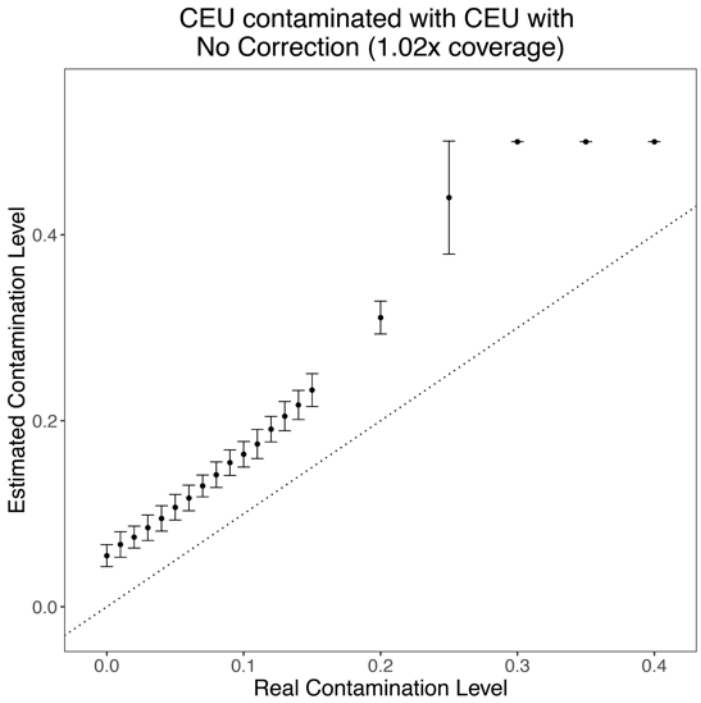
B)



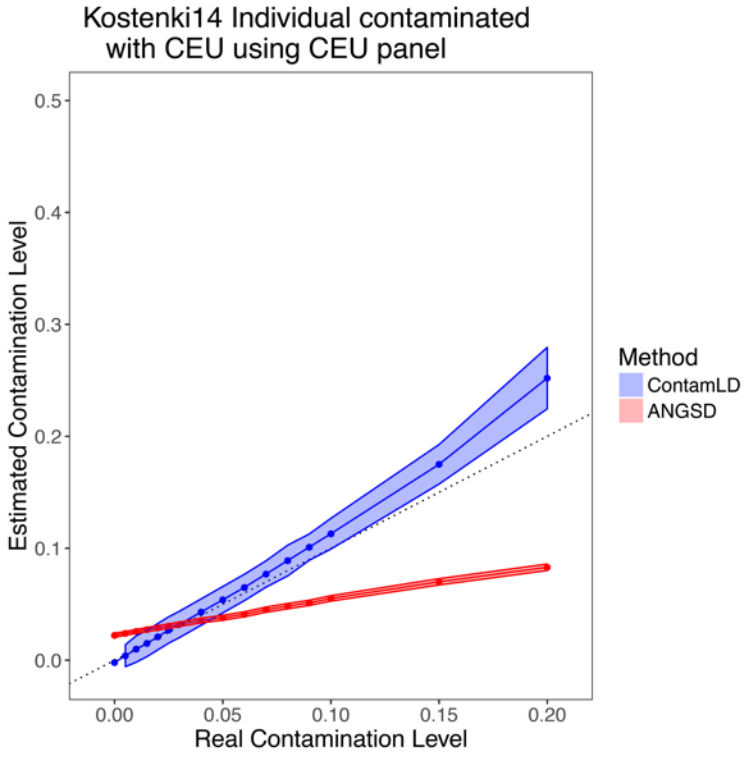
C)



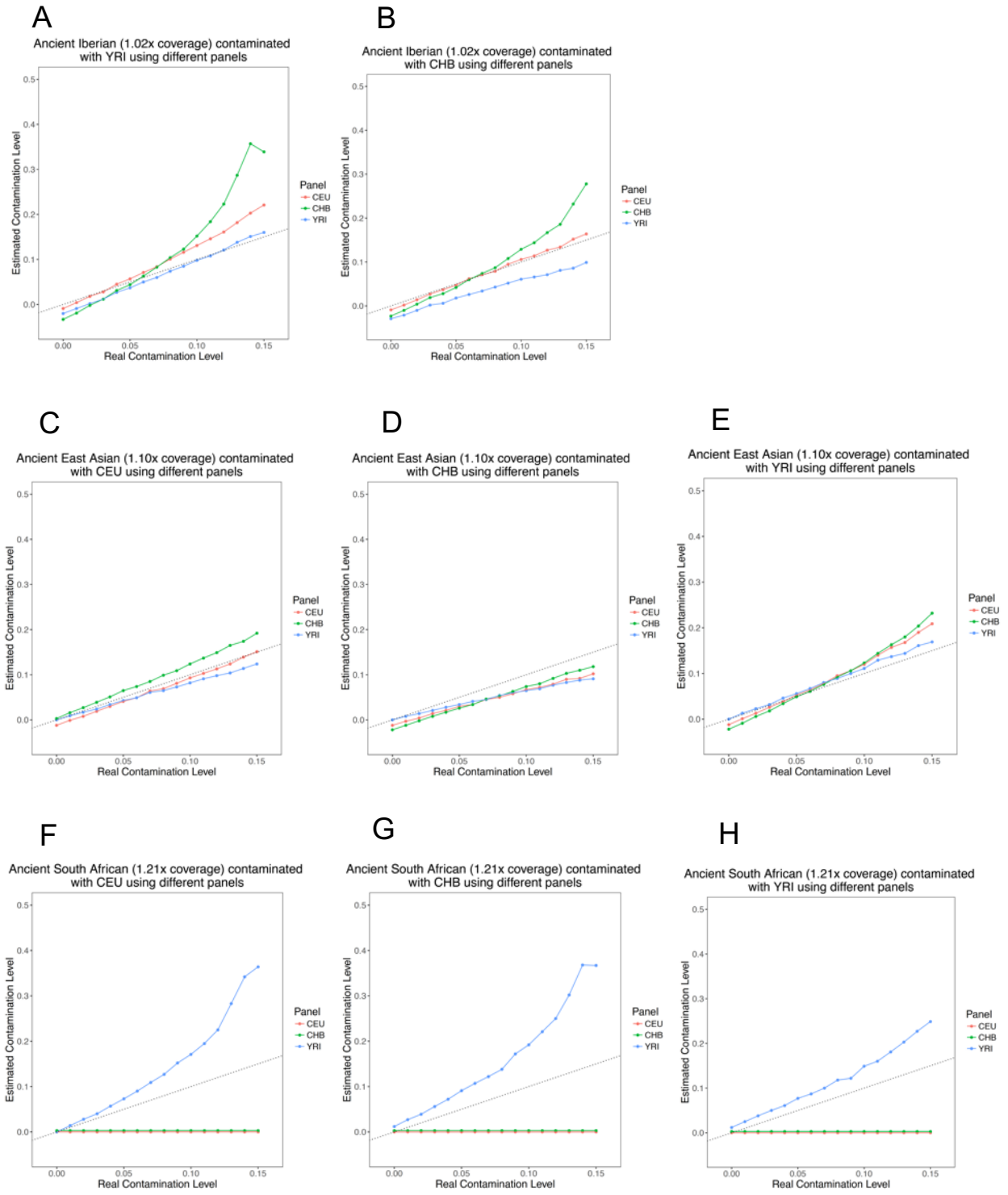
Supplementary Figure 1. Distribution of LOD scores in simulated data. The distribution of LOD scores is depicted for samples with **A)** 0%, **B)** 7%, and **C)** 15% simulated contamination. These data were generated as part of tests using 1000 Genomes CEU individuals as the sample and contaminant DNA and for the haplotype panel.



Supplementary Figure 2. *ContamLD* estimates when the individual, contaminant, and haplotype panel are all from CEU. *ContamLD* was run with no correction. The black dotted line is y=x, which would correspond to a perfect estimation of the contamination. Error bars are 1.96*standard error (95% confidence interval).



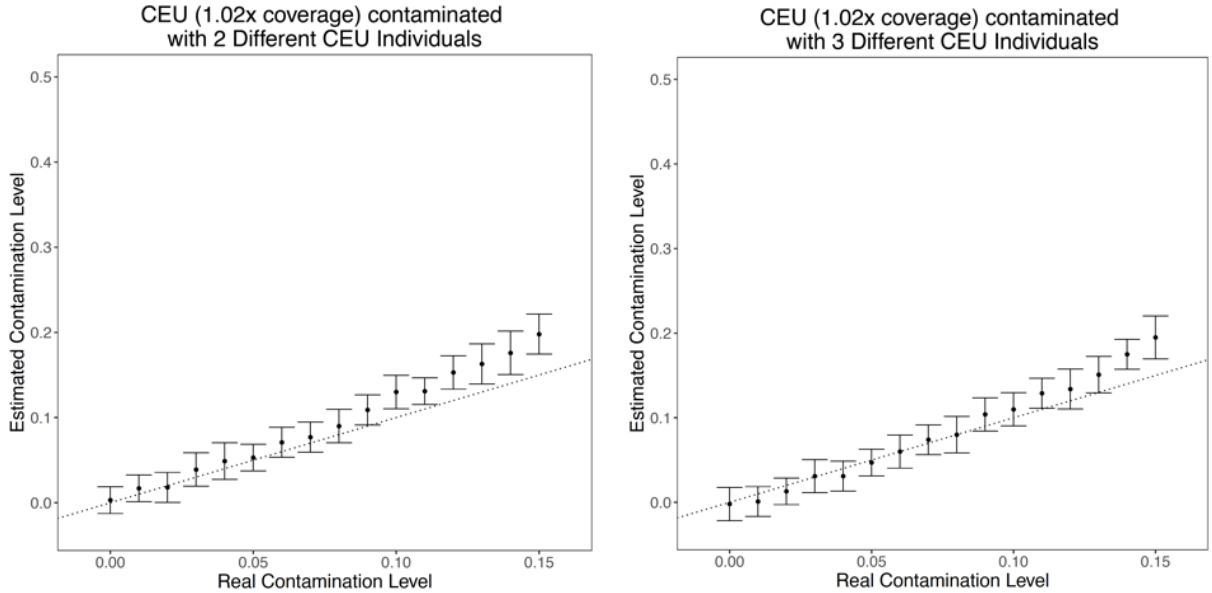
Supplementary Figure 3. *ContamLD* estimates for Upper Paleolithic European individual after damage restricted correction (option 1). Kostenki14 (2.81x coverage) was contaminated with CEU and analyzed using a CEU panel with *ContamLD* using damage correction and *ANGSD* [16] (Method 1). The black dotted line is y=x, which would correspond to a perfect estimation of the contamination. Error shading is 1.96*standard error (95% confidence interval).



Supplementary Figure 4. *ContamLD* estimates where the sample, contaminant, and haplotype panels have varying ancestries. *ContamLD* was run with damage restricted correction (option 1). An ancient Iberian of 1.02x coverage (I3756) was analyzed after contamination with **A)** CHB or **B)** YRI. An ancient East Asian of 1.10x coverage (DA362.SG) is analyzed after contamination with **C)** CEU, **D)** CHB,

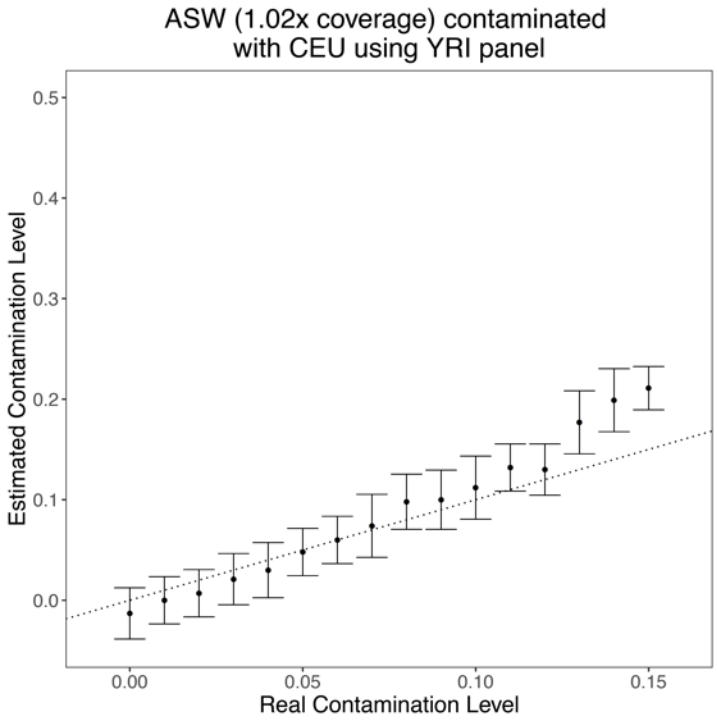
or **E)** YRI. An ancient South African of 1.21x coverage (I9028.SG) is analyzed after contamination with **F)** CEU, **G)** CHB, or **H)** YRI. The black dotted line is y=x, which would correspond to a perfect estimation of the contamination.

A) **B)**



Supplementary Figure 5. *ContamLD* estimates with CEU as the sample and multiple CEU individuals as contaminants. *ContamLD* was run with CEU haplotype panels and damage restricted correction (option 1). A CEU individual of 1.02x coverage (from the sequence distribution of the ancient Iberian above) is contaminated with **A)** two CEU individuals or **B)** three CEU individuals. The black dotted line is y=x, which would correspond to a perfect estimation of the contamination. Error bars are 1.96*standard error (95% confidence interval).

1014



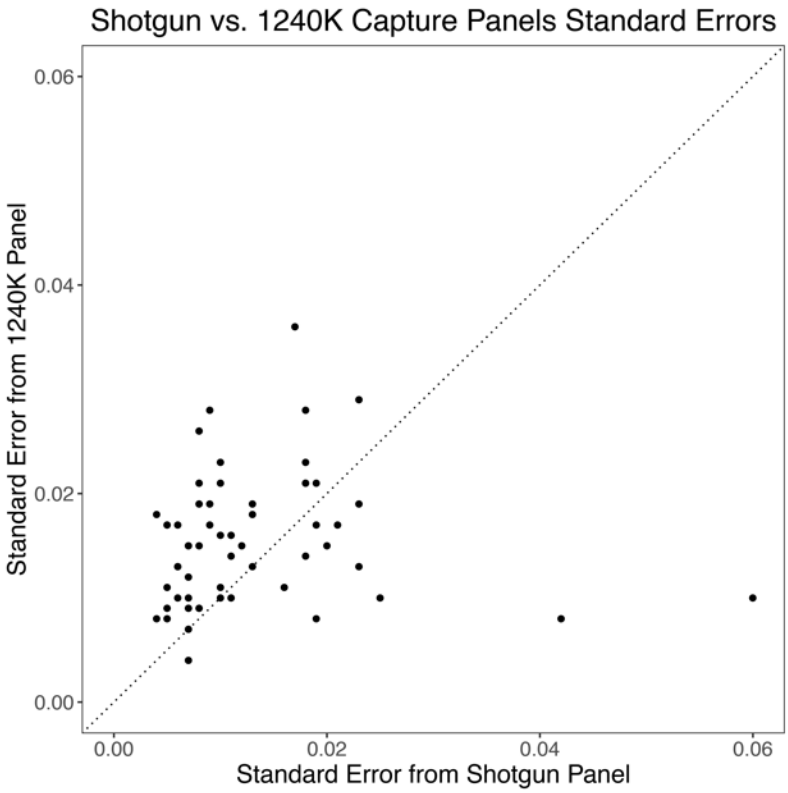
1015

1016 **Supplementary Figure 6. *ContamLD* estimates with an ASW (African-American) individual and YRI**

1017 **panel.** *ContamLD* was run with the 1240K SNP set and damage restricted correction (option 1). The

1018 black dotted line is y=x, which would correspond to a perfect estimate of the contamination. Error bars are

1019 1.96*standard error (95% confidence interval).



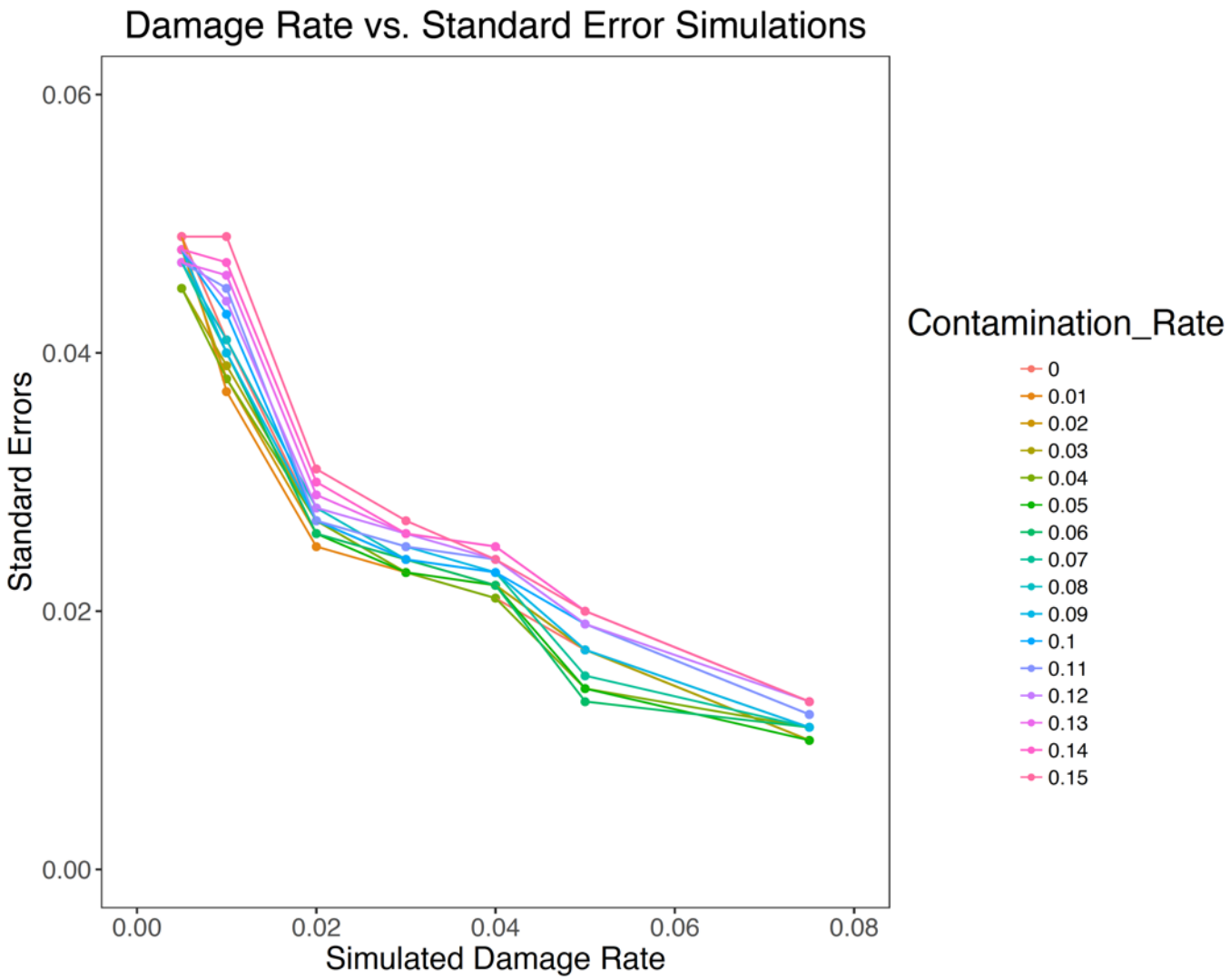
1020

1021 **Supplementary Figure 7. Contamination estimate standard errors of shotgun sequenced ancient**
1022 **individuals comparing the 1240K and shotgun panels.** Ancient shotgun sequenced individuals of 0.1-
1023 0.5x coverage from Allentoft *et al*., 2015 [31], Damgaard *et al.*, Nature 2018 [37], and Damgaard *et al*.,
1024 Science 2018 [21] were analyzed with *ContamLD* damage restricted correction (option 1) using the
1025 1240K SNP set and a shotgun panel created using all variants above 10% frequency in the 1000
1026 Genomes dataset. This test shows that analyses with the shotgun panel generally have smaller error bars
1027 relative to those done with the 1240K panel, though it is unclear why there are two outliers with high
1028 standard errors on the shotgun panel and low standard errors on the 1240K panel. All estimates are in
1029 Supplementary Online Table 1.

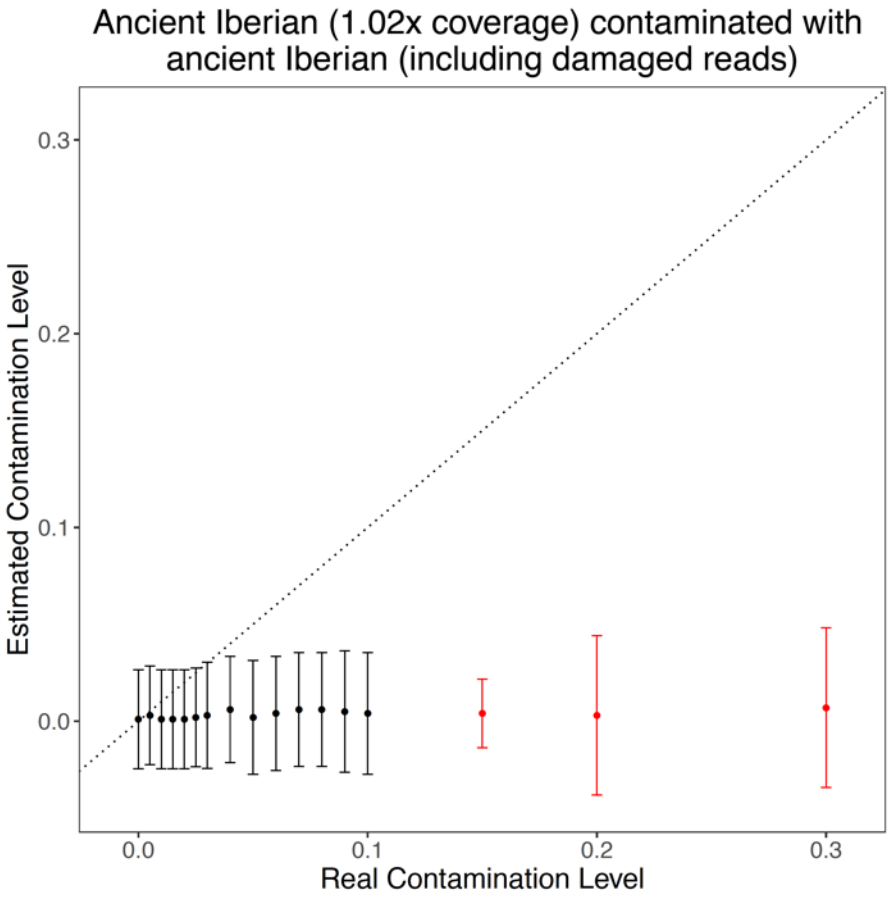
1030

1031

1032



Supplementary Figure 8. *ContamLD* standard errors from an ancient individual contaminated with another individual at different damage rates. The target individual was a 0.5x coverage ancient West Eurasian-related individual (DA57.SG), and the contaminating individual was an ancient Iberian (I10895). *ContamLD* was run with the 1240K SNP set with the CEU panel. The damaged sequences were simulated as 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, and 0.075.



1039

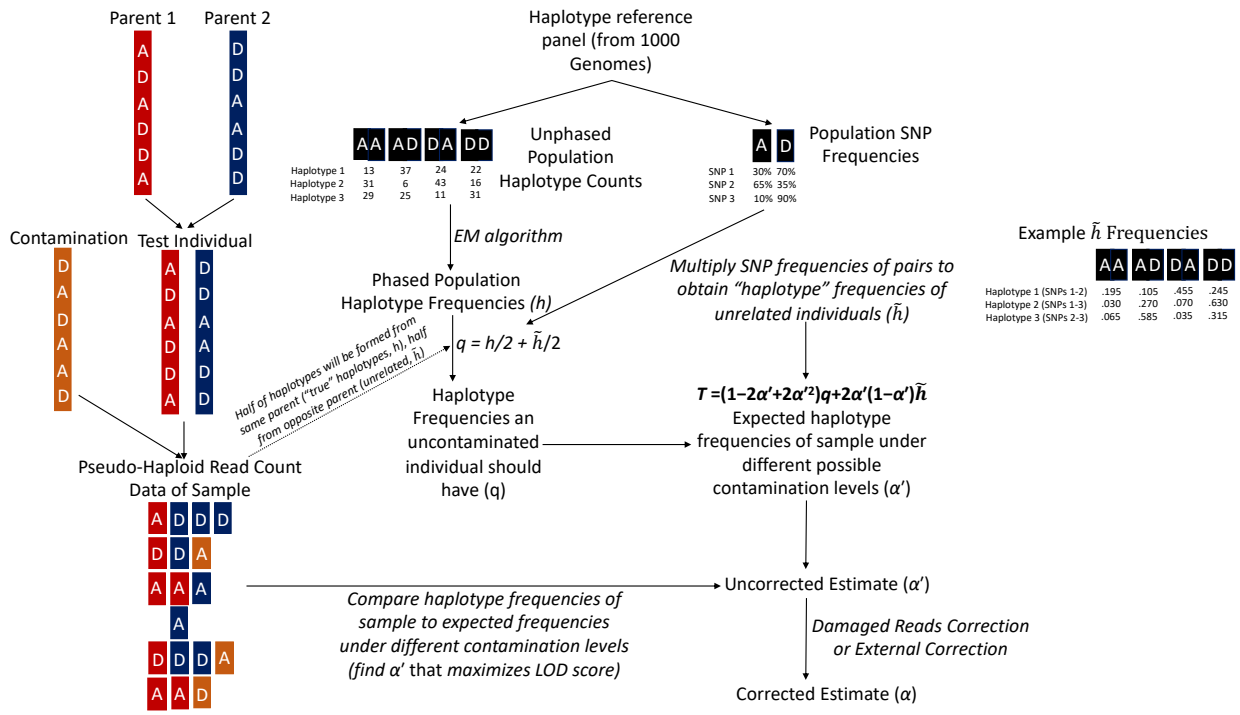
1040 **Supplementary Figure 9. *ContamLD* estimates with an ancient individual contaminated with**
1041 **another ancient individual including its damaged sequences.** The target individual was an ancient
1042 Iberian (I3756) and the contaminating individual was another ancient Iberian (I10895). *ContamLD* was run
1043 with the IBS panel and 1240K SNP set using damage restricted correction (option 1). The damaged
1044 sequences were simulated as a 5% down-sampling of each respective contaminated BAM file. IBS are
1045 1000 Genomes Project present-day Iberians from Spain. The black dotted line is y=x, which would
1046 correspond to a perfect estimation of the contamination. Error bars are 1.96*standard error (95%
1047 confidence interval). Points in red are those flagged with “Very_High_Contamination” by the software.
1048 See Supplementary Online Table 4 for all values.

1049

1050

1051

1052



Supplementary Figure 10. Schematic of *ContamLD* algorithm (see text for additional details).