Enrichment of Damaged DNA by Target Replacement of Deaminated Cytosines

Lourdes Gomez, Joshua D. Kapp, Dr. Ed Green, Dr. Beth Shapiro
Ecology and Evolutionary Biology
University of California, Santa Cruz

Explosive development of next generation sequencing technology has allowed for more extensive use of degraded samples in genomic analysis. Ancient DNA (aDNA) is extracted from samples several hundred to tens of thousands of years old and is often heavily fragmented, contaminated with modern DNA, and contains damaged bases. The most common form of detectable DNA damage is the deamination of cytosines to uracils and is used as an authentication method for aDNA. We aim to exploit this common form of aDNA damage by enzymatically replacing uracil bases with biotinylated nucleotides which allows for the target enrichment of authentic ancient DNA from modern contamination. After creating a single-stranded library and synthesizing a second strand, we enzymatically excise uracils and fill in the gap with a biotinylated nucleotides using a strand-displacing polymerase. With this biotin incorporation, we can pull down molecules with DNA damage onto a streptavidin bead, then PCR-amplify the molecules. We anticipate that the use of our protocol will increase the proportion of informative content in a DNA sequencing library. Libraries with a higher proportion of informative molecules increases the sequencing power of a single library. We hope this technique will provide a tool for researchers to better study the most poorly preserved specimens. With improved access to rare genomic data, we obtain a snapshot into the natural history of population structures and seek to understand current population dynamics in a quickly changing climate.

Background

Woolly mammoths, dodo birds, and passenger pigeons: there's something riveting about studying extinct species. But what about extant species such as polar bears and sea turtles? These may be the "woolly mammoths" of the future – the clock is ticking for these and numerous other struggling species in a time of uncertainty in the natural world. By studying populations that went extinct during historic climate change events, we gain a view of our world's ecological past. This history enables us to understand and predict the evolutionary trajectories of current population dynamics in a changing climate.

The field of ancient DNA research unravels the fascinating stories of past populations. In a study in 2016, Belmcheri et al. used ancient DNA (aDNA) and plant remains from sediment cores on St. Paul Island, Alaska to confirm the presence of woolly mammoths on the island 5,600 years ago (thousands of years more recently than previously thought). Further analysis of the aDNA paired with a mapping of stable isotopes and diatom levels suggests that mammoths disappeared due to island shrinking and limited freshwater access caused by climate change. In another recent study, a German research team used ancient DNA extracted from remains in the St. Jørgen Cemetery to link a human leukocyte antigen allele with susceptibility to leprosy, an epidemic-inducing disease, in Medieval Europeans (Krause-Kyora et al., 2018). The ability to gain information through the study of ancient DNA has rapidly accelerated with the recent development of next generation sequencing technology. Researchers are now able obtain more extensive and informative sequencing and genomic analysis of ancient DNA samples than previously thought possible. However, in order to maximize efficiency and results of genomic sequencing, scientists must employ specific methods for studying ancient DNA from extraction to sequencing. In the UCSC Paleogenomics Lab, we hope to develop a method that will increase

the amount of target DNA in a library and, thus, contribute to more efficient and informative sequencing data in the future.

Ancient DNA: Damage and Challenges of Study

Ancient DNA may come from sediment, bones, hair, or teeth that are several hundred to tens of thousands of years old. Over time, these samples are heavily damaged due to abiotic factors such as change in temperature or pH, as well as biotic factors, such as microbe degradation. Thus, the DNA recovered from archaeological remains is often heavily fragmented and contains damaged bases (Graham, 2016, Krause-Kyora, 2018, Mühlemann, 2018).

At the molecular level, the most common form of detectable DNA damage is the deamination of cytosines to uracils (Figure 1. Dabney, 2013). This may occur due to a variety of factors, such as temperature or pH at varying rates (Lindahl, 1974).

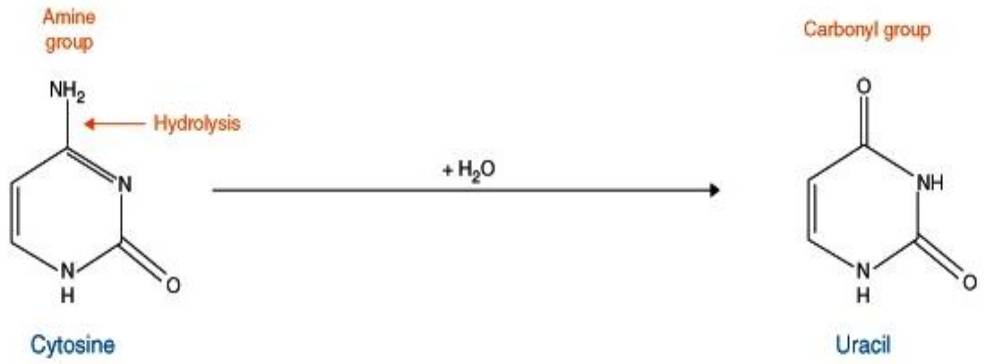


Figure 1. Cytosine Deamination. Cytosine (a DNA base) loses its amino group through hydrolysis, thus becoming uracil, an RNA base. (Dabney, 2013).

When these damaged DNA fragments are sequenced, “U” uracil, an RNA base, is read as a thymine, or “T”. What once was a cytosine is now a thymine in amplified sequences, and scientists refer to this as “C to T substitution” (Figure 2).

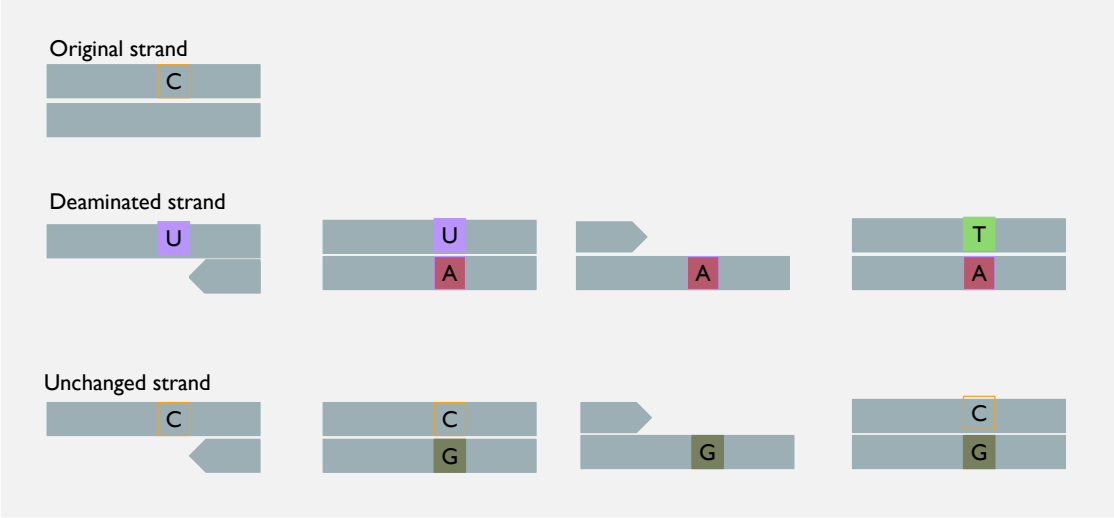


Figure 2. A side-by-side comparison of the polymerization and sequencing of a deaminated vs original DNA strand.

After sequencing, researchers are able to computationally quantify the disproportionate presence of thymine bases in a sequence by comparing the sequence to the reference genome of the species, and thus create a damage profile (Figure 3). This increase in thymine bases due to cytosine deamination is unique to damaged DNA and is thus used as an authentication method for ancient DNA (Dabney, 2013).

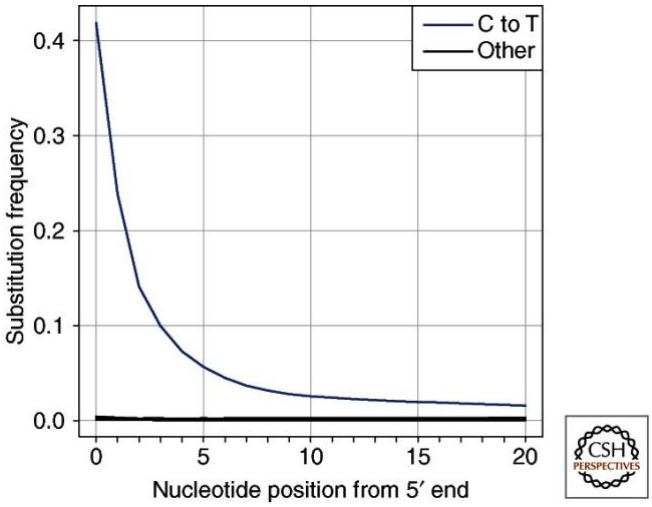


Figure 3. This figure of substitution frequency in relation to the 5’ end of a DNA fragment represents the drastically increased presence of thymine in a fragment sequence where the reference genome expresses a cytosine. This type of damage profile is used by aDNA researchers to authenticate ancient samples. (Dabney, 2013)

In addition to degradation, contamination poses a significant challenge in the field of ancient DNA. Contamination is the presence of DNA that does not map to the reference genome or does not represent the DNA of the living sample. This may come from a myriad of sources such as microbial degradation, plant pollen, or animal droppings that take place around or any time after an organism's death up until the recovery of a sample at an archaeological site and processing in a sequencing laboratory. A sample refers to an extract full of DNA, including the DNA of the organism where the extract came from (the target DNA) as well as contaminant DNA. In its natural environment over extended periods of time, a sample's contaminants also undergo damage, exhibiting similar damage profiles to the target DNA. Furthermore, in the time between the discovery of archaeological remains to the time of a sample's DNA extraction in a laboratory, the sample is passed between many human hands, further subjecting it to contamination. This contamination, however, can be considered "modern" because, when sequenced, it will most likely not have a damage profile. Commonly, researchers computationally isolate target DNA post-sequencing by comparing sequence data to a reference genome. A major drawback of this method, however, is that the proportion of desirable sequence data to extraneous data is extremely small; in many samples it is rare to recover higher than 1% endogenous DNA (DNA that is mapped to the reference genome of the target species). Most DNA in a sample is a collection of DNAs from non-target organisms. While technologies have been recently developed in an effort to increase accessibility to genomic sequencing, it is still a complex and expensive process. Due to its relatively high cost, sequencing an entire sample only to extract less than 1% of the genomic data is extremely inefficient for sequencing laboratories.

In the Paleogenomics Lab at UC Santa Cruz, our proposed solution is to utilize the presence of damage in ancient DNA to filter out modern DNA contaminants *prior* to sequencing.

In an effort to remove the bias in sequencing data caused by cytosine deamination, scientists commonly follow a protocol of uracil-DNA-glycosylase treatment (Rohland, 2015). Uracil-DNA-Glycosylase (UDG) is an enzyme that removes uracils from DNA sequences, leaving an "abasic" site. In the next step of a UDG treatment, DNA is exposed to an endonuclease, which cleaves the phosphate backbone of the double helix, leaving a gap in double stranded DNA. Using a polymerase, those gaps are then filled in with cytosines, thus restoring the DNA to its original sequence. New England BioLabs produces and sells Uracil-Specific Excising Reagent (USER), which includes both UDG and an endonuclease. This enzyme mix was utilized for all uracil-excision in this project.

Project Aim

We aim to exploit the most common form of aDNA damage, cytosine deamination, by enzymatically replacing uracil bases with biotinylated nucleotides. This process allows for the target enrichment of authentic ancient DNA from modern contamination. Put simply, we aim to use the "damage signature" of uracils to filter out unwanted modern DNA contamination. If accomplished, this method will improve sequencing power by increasing the number of informative molecules in a library. By extracting maximal sequence data from ancient DNA samples, researchers are able to characterize how major events in natural history led to the demise of extinct populations. With improved access to rare genomic data, we can piece together the evolutionary stories of extinct populations, such as the mid-Holocene mammoth extinction in St. Paul, Alaska (Belmcheri et al., 2016). This empowers researchers to make inferences regarding current populations and ecosystems as we face a major climate shift, and thus, a major

worldwide evolutionary event that has the potential to adversely impact the survival of modern species.

Initial Protocol and Evolution of Method Workflow

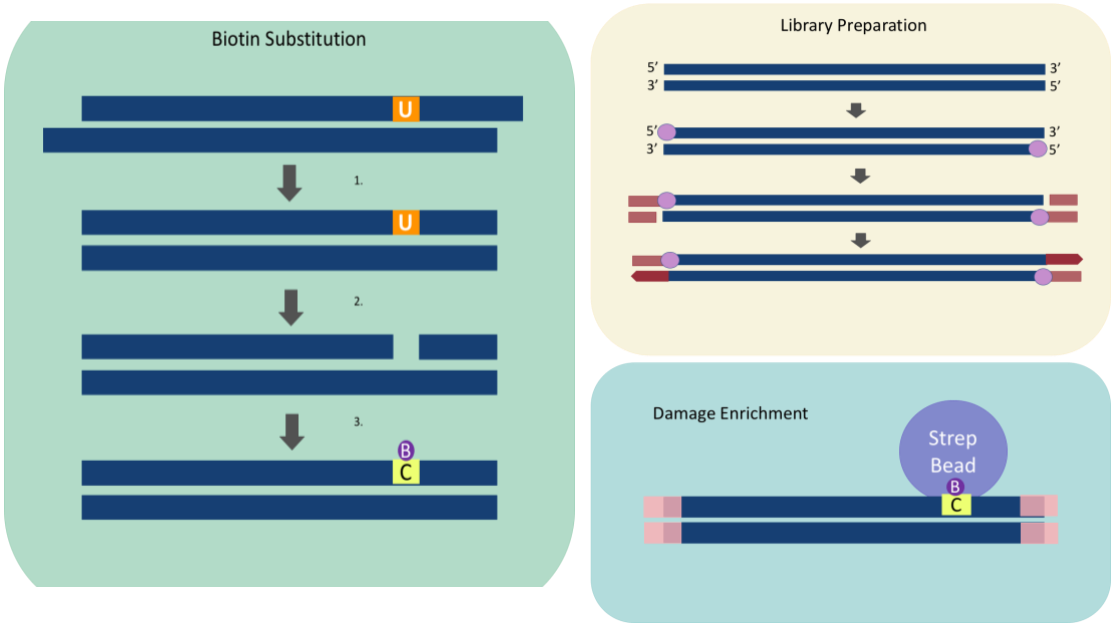


Figure 4. Method Workflow showing end repair of fragments creating blunt ends, uracil excision, incorporation of biotinylated cytosine, library preparation, and enrichment on a streptavidin bead.

Our initial protocol began with an end repair reaction, creating blunt ends on all molecules. Then, we excised uracils with NEB's USER enzyme. We filled the gap with a biotinylated nucleotide using T4 DNA Polymerase (NEB), then moved into a library preparation protocol before enriching for damaged fragments on streptavidin beads.

This approach posed a problem, however, for instances in which uracils only reside on the single stranded ends of fragments. This is extremely common because cytosine deamination happens much more quickly in single stranded overhangs than in double stranded DNA (Lindahl, 1974). Blunt-ending our molecules in this protocol was a necessary step because the T4 DNA Polymerase that we used to incorporate biotinylated nucleotides also has an exonuclease activity

which chews back 3’ overhangs; if any 3’ ends were available, especially after uracil excision, we risked losing the entire fragment to the exonuclease activity of T4 Polymerase. Multiple experiments proved that this exonuclease activity is extremely aggressive. However, by creating blunt ends on all molecules, we lost the damage signal and ability to enrich for all fragments that only contain a uracil on the 3’ single stranded overhang. Further experimentation with the blunt end protocol and resulting sequencing suggests that uracils only exist on single stranded overhangs (See figure 8 in results). So, we adjusted the protocol in order to maintain all fragment overhangs.

Current Method Workflow

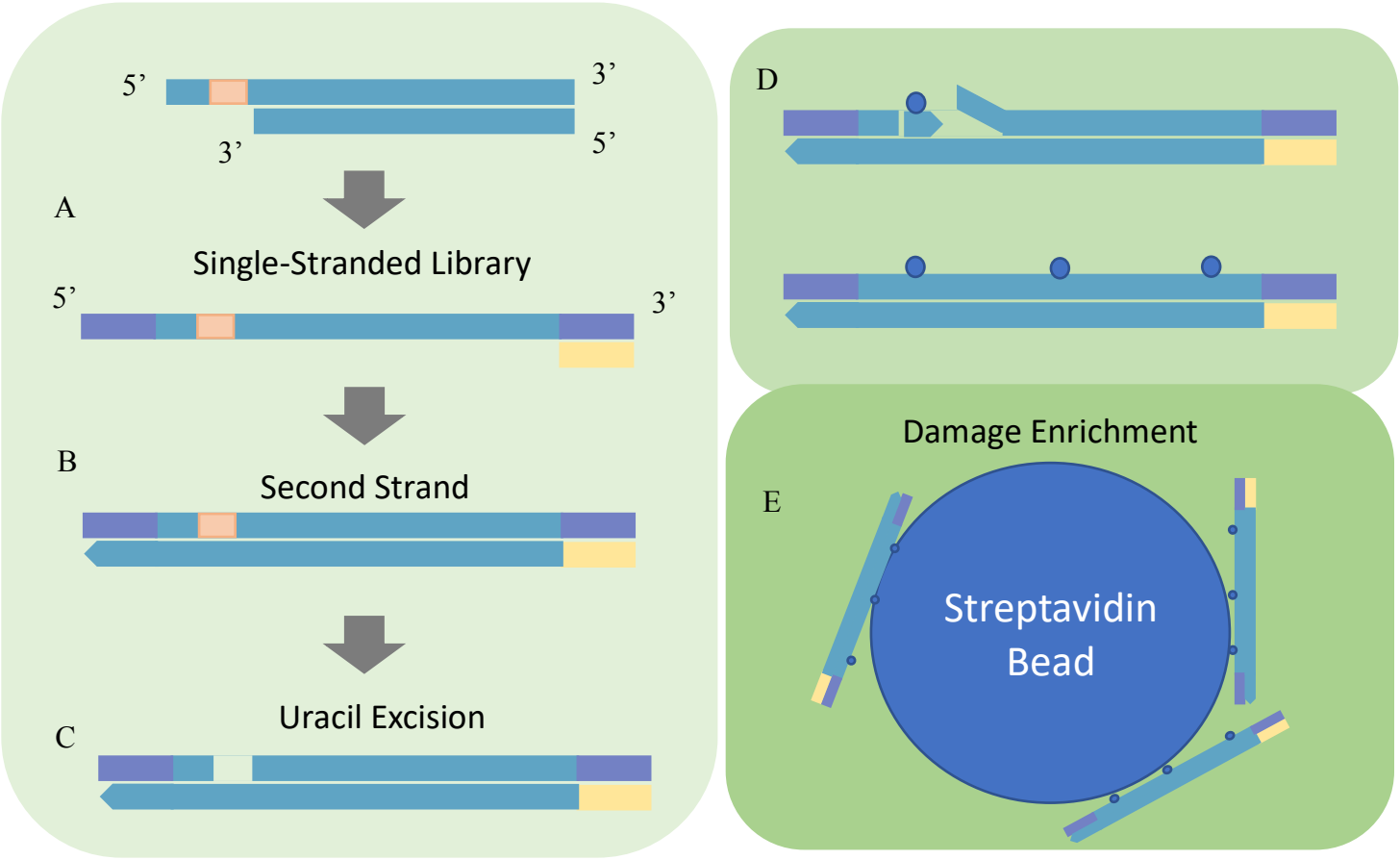


Figure 5. Method Workflow showing a. the ligation of adapters onto a template molecule containing uracil in a single-stranded library, b. the synthesis of a second strand, c. excision of uracils using USER enzyme, d. strand-displacing fill-in with biotinylated nucleotides using klenow exominus, e. enrichment on a streptavidin bead.

Materials and Methods

Materials:

Reagent
0.5 μM single stranded oligonucleotide
0.5 μM 60N oligonucleotide
50 μM Primer complement to library adapter
10X Cutsmart Buffer
USER Enzyme (New England BioLabs)
Klenow exominus Polymerase (New England BioLabs)
5mM stock dNTP (total rx [200 μM])
Biotinylated dNTP (total rx [200 μM])
Streptavidin C1 Beads (ThermoFisher)
0.1X BWT+SDS, 0.1X BWT, Stringency Wash

Oligonucleotide Design: consistent DNA is imperative to our assay

We designed and ordered the following oligonucleotides from Integrated DNA Technologies (IDT), which acted as an imperative indicator of the success of our assay:

1. This "top" oligonucleotide strand was used for all single-stranded experiments and library preparations:

5' ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA GTA TGG CTT CAT TCA GCT CCG GTT CCC AAU GAT CAA GGC GTC GGA AGA GCA CAC GTC TGA ACT CCA GTC A3'

2. The "top" and "bottom" oligonucleotide strands were hybridized and used for all double-stranded experiments:

5' ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA G T A TGG CTT CAT TCA GCT
3' C* A* T*ACC GAA GTA AGT CGA

CCG GTT|CCC AAU GAT CAA GGC GTC GGA AGA GCA CAC GTC TGA ACT CCA GTC A 3'
GGC CAA GGG TTG CTA GTT CCG CAG CCT TCT CGT GTG CAG ACT TGA GGT CAG T 5'

The IDT-synthesized oligonucleotides contain uracils and thus, represent damage DNA in our experiments. We also utilized a "60N" oligonucleotide synthesized by IDT that is a collection of molecules of nonspecific sequences, all 60 bases long that contain adenine, thymine, cytosine, and guanine bases, but no uracils. These oligonucleotides represented modern, or undamaged DNA in our experiments.

Past Methods

The following methods were used until the researchers had reason to believe that cytosine deamination may exist only on single stranded DNA fragment ends. The methods below are provided as a template which represents many experiments in which one condition was tested or altered at a time. Altered conditions are provided in tables below the methods template.

End Repair Creating Blunt Ends
We created blunt ends using the end repair step outlined at the beginning of the Blunt-End Single Tube (BEST) protocol (Caroe, 2017).

USER Treatment of oligonucleotides
Each oligo of 1 pM (2 μL of 0.5 uM stock dilution) was combined with 5 μL of EBT, 1 μL 10X Cutsmart Buffer, and 2 μL of Uracil-Specific Excision Reagent (USER) Enzyme (New England Biolabs). The reaction was incubated at 37°C for 20 minutes. Immediately after incubation, 1 μL of Uracil Glycosylase Inhibitor (UGI) was added.

Incorporation of nucleotides
I added the following reagents to the previous USER-treated reaction: 2 μL 10X Cutsmart Buffer, 4 μL 2mM stock dNTP (total reaction concentration of 200 μM), 2 μL T4 DNA Polymerase, 2 μL Polynucleotide Kinase (PNK), 0.8 μL of 50 mM ATP, and 4 μL of 50 mM DTT. I filled the remaining reaction to 40 μL total volume with EBT. The reaction was then incubated at 12°C for 30 minutes, then inactivated at 65°C for 20 minutes.

Ligation of cleaved phosphate backbone
I added the following reagents to the previous incorporation reaction: 1 μL of T4 DNA Ligase, 0.2 μL of 50 mM ATP, μL of 10X Cutsmart Buffer. I then filled the remaining reaction to a total volume of 50 μL with EBT. I then incubated the reaction at 16°C for 30 minutes, then inactivated at 65°C for 20 minutes.

Library Preparation
We used a single-stranded library preparation protocol, but any single or double stranded library prep can be used.

Modifications:

USER:

Time titration:

Time (minutes):	15	30	60	120

USER titration:

USER (μL):	1	2	3	4	5

Incorporation:

Time titration:

Time (minutes):	15	30	60	120

Deoxynucleotide (dNTP) titration:

dNTP (μM)	25	50	100	200	400

Ligation:

Ligase titration:

Ligase (μL)	1	2	3	4	5

[ATP] titration:

[ATP] μM	0	50	100	200

[PNK] titration:

[PNK] μM	0	50	100	200

Current Methods

The methods below are provided as a template which represents many experiments in which one condition is or will be tested or altered at a time.

Create Library from Sample
Any library preparation method can be used, we used a single stranded library preparation. A single stranded library was created that consisted of a mixture of 60N and single-stranded "top" oligonucleotides.

Primer Spike
I denatured 20 μL of each library at 95°C for 3 minutes, then added 10 μL mix comprised of 2 μLof 50 μM Primer (complement to 3' library adapter), 3 μL of Cutsmart Buffer, and 5 μLof EBT. I then incubated the reaction at 95°C for 30 seconds, then ramped down to a 60°C for 2 minutes at 0.1°C/second.

Second Strand Synthesis
I added a 20 μL reaction mix containing 2 μL Klenow exominus polymerase (NEB), 2 μL Cutsmart Buffer, 2 μL 5mM dNTP (total [200 μM]), and 14 μL of EBT to the previous reaction. The reaction was incubated at 30°C for 30 minutes.

USER Spike
I added a 10 μL reaction mix to the previous reaction that included 2 μL USER II (NEB), 1 μL Cutsmart Buffer, 3.75 μL 400 μM biotinylated dCTP (total [25 μM]), and 3.25 EBT. The reaction was incubated at 37°C for 30 minutes.
Enrichment
The reaction was pulled down on Dynabeads MyOne Streptavidin C1 Beads according to the ThermoFisher Scientific protocol and specifications.

Cleanup
I used the Monarch Cleanup Kit with 5:1 binding buffer to reaction volume ratio. I performed two washes with the kit's ethanol-based wash buffer and eluted the DNA in 50 μL of EBT buffer.

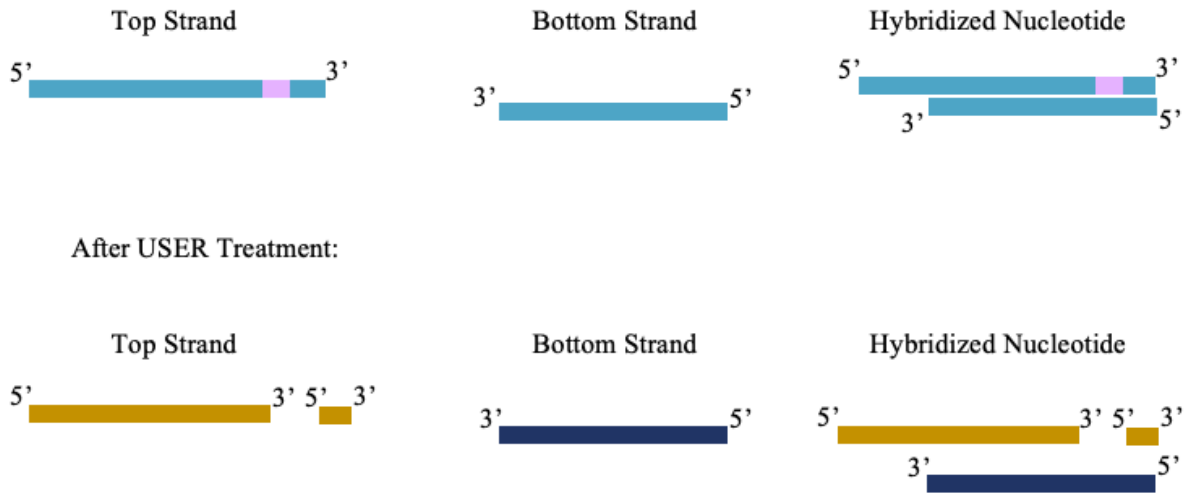
Analysis
I used a quantitative polymerase chain reaction (qPCR), Polyacrylamide Gel Electrophoresis (PAGE), TapeStation, and Fragment Analyzer to analyze reactions.

Results

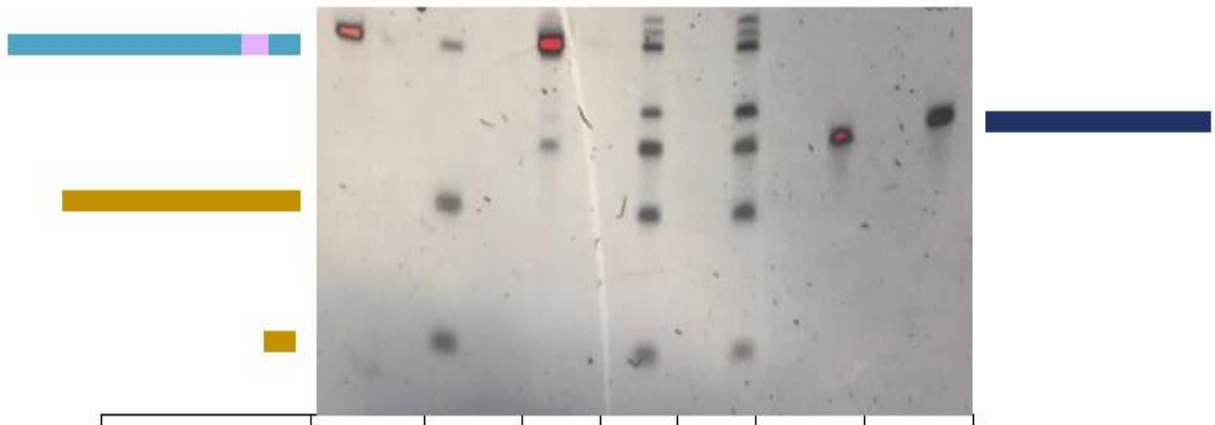
USER®: Uracil-Specific Cleaving Reagent

In a USER-efficiency test, we compared unhybridized bottom and top single strands as well as hybridized double-stranded samples of our oligonucleotide in two conditions: treated with USER (cut) and untreated (uncut). USER successfully excises most uracils, but not all. There seem to be remnants of uncut top strand in the top strand and hybridized lanes. Additionally, there is an unknown band in the hybridized cut lane.

Single strands and double-stranded hybridized oligonucleotide



Polyacrylamide Gel Electrophoresis of Treated & Untreated Denatured Oligonucleotides



Oligonucleotide	Top	Top	Hyb	Hyb	Hyb	Bottom	Bottom
USER Treatment	-	+	-	+	+	-	+

Figure 7. Denatured DNA run on Polyacrylamide Gel Electrophoresis (PAGE) with animated strands representing oligonucleotide fragments.

In the time titration of USER treatment incubation, the quantitative polymerase chain reaction (qPCR) results are as follows:

uL of USER	qPCR CT value	replicate
1	13.67	13.94
1	14.16	14.01
2	14.34	14.36
2	14	13.98
3	14.59	14.48
3	14.33	14.39
4	14.63	14.57
4	14.4	14.4

Damage Profile of ancient DNA samples with and without end repair

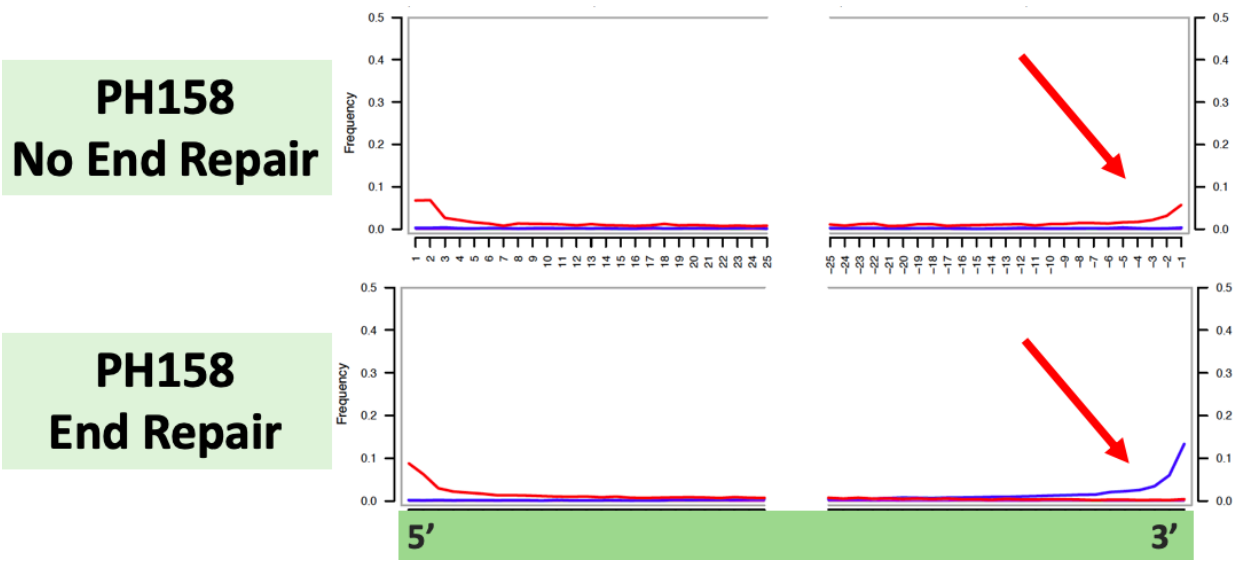


Figure 8. Map Damage Plot of an ancient horse extract that was sequenced under two conditions: with and without end repair prior to library preparation. The x axis represents nucleotide position with respect to the 5' and 3' ends of fragments and the y axis represents the frequency of substitutions. These are the results that prompted the shift from the previous workflow to the current methods.

Fragment Analyzation Results of current method

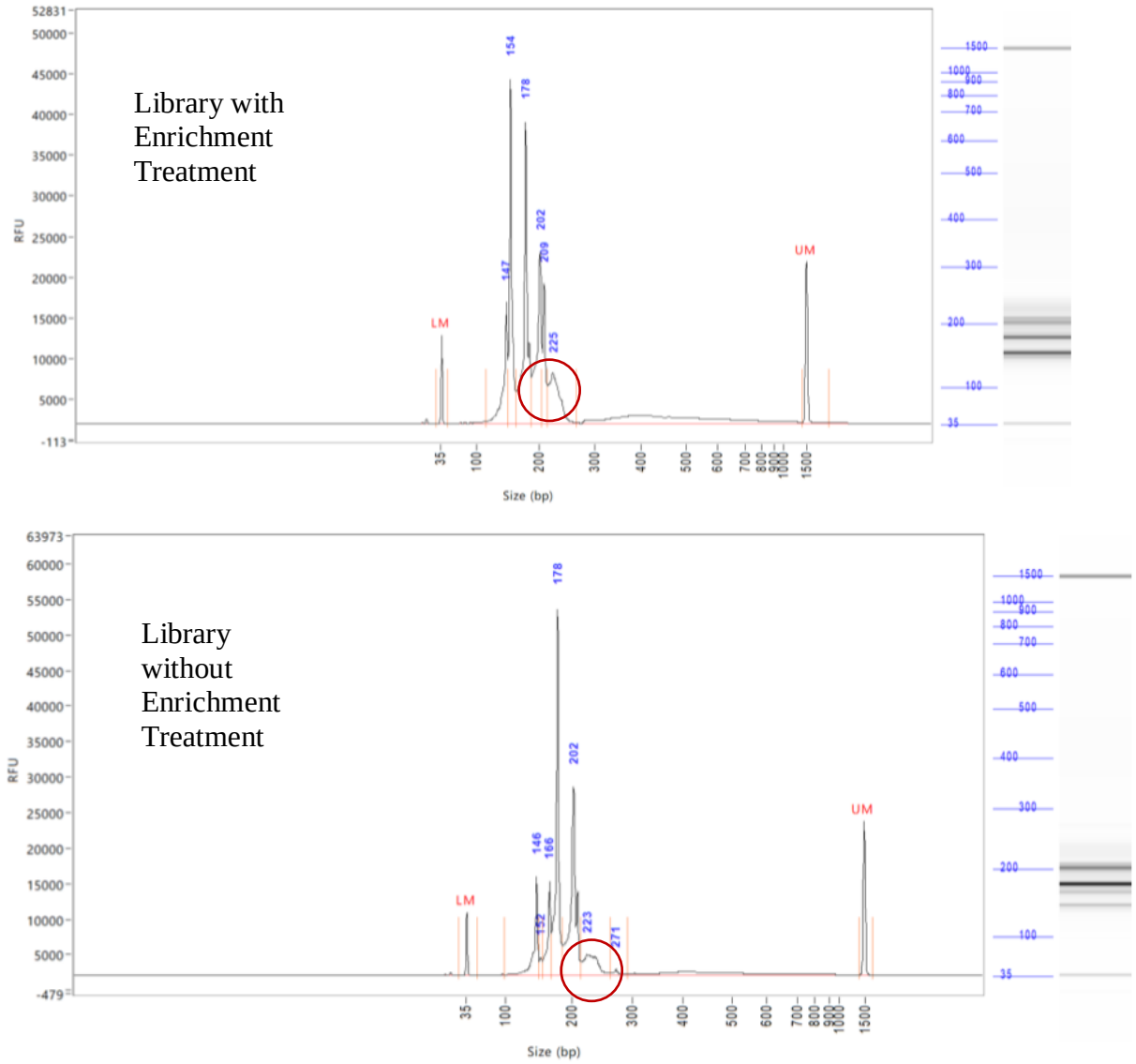


Figure 9. Fragment analyzation report with comparison of libraries with and without enrichment treatment.

In a comparison of libraries with and without enrichment treatment, the uracil-containing oligonucleotides representing damaged DNA exist in higher frequency in the treated sample than in the non-treated sample.

Discussion:

USER®: Uracil-Specific Cleaving Reagent

Based on the results of the PAGE, it is clear that USER is cleaving uracils with acceptable efficiency. Although, there are still visual remnants of un-cleaved oligos, our USER volume titration did not suggest that an increased amount of USER would significantly affect the efficiency of USER. Based on our USER volume titration experiment, we were able to set a baseline of qPCR CT values to compare future experiments to. Determining the USER efficiency was the first step in developing our assay and protocol.

Whether or not cytosine deamination damage exists at all in double stranded DNA or if it only exists in single stranded DNA is a question that we hope to pursue in future research. In the meantime, our preliminary results (figure 8) do suggest that cytosine deamination may only exist on single-stranded overhangs and thus, these results were a significant factor in our decision to change our methods.

Fragment analyzation results of libraries with and without treatment suggest that there has been some enrichment for the oligonucleotide that represents damaged DNA. However, this is preliminary data and we are working to generate substantial data that will confirm this enrichment. Furthermore, we will continue to develop and optimize this method.

We move forward with confidence that our method will improve our ability to capture maximal data from ancient, damaged DNA samples and thus, increase our knowledge of species of the past. By learning from our ecological history, we can move forward with confidence in a changing world.

Works Cited

Bernick, David. “Programming for Life Sciences.” BME 160. University of California, Santa Cruz, Santa Cruz.

Briggs, A. et al., 2010, Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA, Nucleic Acids Research;38(6)

Caroe, C. et al., 2017, Blunt-End Single-Tube Illumina library building for modern and ancient DNA, Methods in Ecology and Evolution 2: 410-419

Dabney, J et al., 2013, Ancient DNA Damage, Cold Spring Harbor Perspectives in Biology, 7

Gansauge, M. and Meyer, M., 2013, Single-Stranded DNA library preparation for the sequencing of ancient or damaged DNA, Nature Protocols 8: 737-748

Gansauge, M. and Gerber, T., 2017, Single-Stranded DNA library preparation from highly degraded DNA using T4 DNA ligase, Nucleic Acids Research 10: e79

Graham, R. Belmcheri et al., 2016, Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska, Proceedings of the National Academy of Sciences 33: 9310-9314

Kapp, Joshua. Paleogenomics Lab. University of California, Santa Cruz

Korlević, P. et al., 2018, A combined method for DNA analysis and radiocarbon dating from a single sample, Scientific Reports 4127

Krause-Kyora et al., 2018 Ancient DNA study reveals HLA susceptibility locus for leprosy in Medieval Europeans, Nature Communications 1569

Lindahl, Tomas & Nyberg, Barbro. 1974, Heat-induced deamination of cytosine residues in DNA. Biochemistry. 13. 3405-10. 10.1021/bi00713a035.

Meyer, M. and Kircher, M., 2010, Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing, Cold Spring Harbor Laboratory Press 6

Mühlemann, B. and Jones, T., 2018, Ancient Hepatitis B Viruses from the Bronze Age to the Medieval Period, Nature. 557: 418-423