**APPLICATION OF ANCIENT DNA METHODOLOGIES TO FORENSIC SCIENCE**

**APPLICATION OF ANCIENT DNA METHODOLOGIES TO FORENSIC SCIENCE:**

**A TECHNOLOGICAL TRANSITION INTO**

**HIGH-THROUGHPUT SEQUENCING AND ENZYMATIC DNA REPAIR**

By

NATHALIE MAEVA MOUTTHAM, B.Sc.

A Thesis Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

McMaster University © Copyright by Nathalie Maeva Mouttham, December 2014

All rights reserved

**FULL TITLE**: Application of ancient DNA methodologies to forensic science: a technological transition to high-throughput sequencing and enzymatic DNA repair

**AUTHOR**: Nathalie Maeva Mouttham, B.Sc., University of Ottawa

**DEGREE**: Master of Science (2014)

**LOCATION**: McMaster University, Hamilton, Ontario, CANADA

**DEPARTMENT**: Biology

**SUPERVISOR**: Hendrik N. Poinar, PhD

**NUMBER OF PAGES**: xi, 64

# ABSTRACT

Forensic scientists and ancient DNA researchers face similar challenges with respect to genetic information acquisition and analysis. However, these communities differ in one critical aspect: while forensic science is regulated by the strict guidelines of the judicial community, ancient DNA is a research-based academic field free to explore emerging technologies as they arise. This thesis investigates the application of two methodologies, developed in ancient DNA research, to challenging extracts, in hopes of modernizing forensic models while maintaining compatibility with current standards. The first chapter focuses on blunt-end sequencing library preparation protocols previously optimized for ancient DNA specimens. Forensically-relevant extracts were converted into libraries and typed by short tandem repeats (STR) amplification. When compared to STR profiles from pre-library extracts, a significant decrease in the quality was observed, in the form of allelic drop-out, heterozygous peak imbalance and increased stutter ratios. The second chapter discusses the efficacy of two enzymatic DNA repair methods, “PreCR® Repair” and “Nelson”, on typical ancient DNA specimens. Based on endogenous sample content, fragment length variation and base misincorporation rates, some DNA repair was reported when using PreCR®. However, the use of the Nelson protocol is not recommended for use in its current state. Both sequencing library preparation and enzymatic DNA repair show potential application to forensic evidential material, but require further analyses to confirm hypotheses and observations outlined in this thesis.

# ACKNOWLEDGMENTS

To Dr. Hendrik Poinar: I owe this amazing opportunity to you. Your contagious energy, your endless capacity for forward-thinking and your strive for excellence has created a laboratory where students are encouraged to reach for the stars. You have pushed me out of my comfort zone and challenged me for the past two years, making me a stronger scientist and person. For that, I am ever grateful.

To Dr. Ron Fourney: if it weren't for your collaboration with Hendrik, I would have never been aware of this field or would have received this life-changing gift. Your patience and support during these two years was much appreciated.

To the current and formers members of the McMaster Ancient DNA Centre: we were all courageous (or crazy?) enough to embark on this wild journey that is ancient DNA research! Thank you for the pleasure of your company during my time here. P.S. You better put this new lab to good use and I look forward to all those Nature publications in the coming years!

To all my amazing friends around the world: you are the best support team a girl could ever wish for. You inspire me in so many ways and I am thankful to have you by my side in the arena, these past two years and long before that.

To my little sister Lara: nous faisons avancer le schmilblick de la science et c'est notre joie! And now, when people say "I have a Masters degree", we can both scoff and reply "Who doesn't?". Thank you for sciencing with me all these years.

To my dream team, Chunk, Ali, Mel and Jake: to say I could not have done this without you is a complete understatement. I have had the immense honour of having each of you in my life and I am a better scientist, friend and human being for it. This thesis is built on Fazari pizza, DCBs and enough laughter to shake the world.

A mes parents: je vous dédie cette thèse car sans vous, elle n'existerait pas. Il n'y a pas assez de pages pour énumérer l'immense liste de tout ce que vous avez fait pour moi mais elle se résume très bien: un amour inconditionnel pour votre fille. Il y a en moi plus que votre sang, il y a votre âme, qui brille dans tout ce que je fais et en tout ce que je suis. Je vous aime très fort.

N.

TABLE OF CONTENTS

[ABSTRACT iii](#_Toc405224655)

[ACKNOWLEDGMENTS iv](#_Toc405224656)

[FIGURES vii](#_Toc405224657)

[tables viii](#_Toc405224658)

[DECLARATION OF ACADEMIC ACHIEVEMENT ix](#_Toc405224659)

[INTRODUCTION - FORENSIC SCIENCE AND ANCIENT DNA 1](#_Toc405224660)

[CHAPTER I - HIGH THROUGHPUT SEQUENCING 8](#_Toc405224661)

[STR profiling from next-generation sequencing libraries: 8](#_Toc405224662)

[Abstract 9](#_Toc405224663)

[Introduction 10](#_Toc405224664)

[Materials and Methods 12](#_Toc405224665)

[Results 14](#_Toc405224666)

[Discussion 18](#_Toc405224667)

[Conclusion 19](#_Toc405224668)

[Acknowledgments 20](#_Toc405224669)

[References 21](#_Toc405224670)

[Supplementary Online Materials 25](#_Toc405224671)

[Supplementary Online Materials References 32](#_Toc405224672)

[CHAPTER II - ENZYMATIC DNA REPAIR 34](#_Toc405224673)

[Surveying the repair of damaged ancient DNA with high-throughput sequencing 34](#_Toc405224674)

[Abstract 35](#_Toc405224675)

[Methods Summary 35](#_Toc405224676)

[Introduction 36](#_Toc405224677)

[Materials and Methods 39](#_Toc405224678)

[Results and Discussion 41](#_Toc405224679)

[Author Contribution 46](#_Toc405224680)

[Acknowledgments 46](#_Toc405224681)

[References 47](#_Toc405224682)

[Supplementary Online Materials 50](#_Toc405224683)

[Supplementary Online Materials References 58](#_Toc405224684)

[conclusion 60](#_Toc405224685)

[REFERENCES 62](#_Toc405224686)

# FIGURES

**INTRODUCTION**

1 The structure of short tandem repeats (STR)................................................................... 2

2 An overview of types of DNA damage............................................................................... 3

3 Characterization of two metagenomic libraries................................................................ 4

4 A schematic representation of HTS of DNA from fossil remains....................................... 6

**CHAPTER I - hiGH THROUGHPUT SEQUENCING**

1 Number of alleles successfully designated from artificially-degraded DNA ................... 15

2 Peak height ratios of heterozygous alleles from artificially-degraded DNA ................... 16

3 Marker-specific stutter ratios ......................................................................................... 17

S1 Representation of blunt-end library preparation of a dsDNA fragment ........................ 29

S2 Fragment length distribution of human stock DNA extract ........................................... 30

S3 Number of alleles successfully designated from three-person DNA mixtures ............... 31

**CHAPTER ii - eNZYMATIC dna REPAIR**

1 Endogenous mitochondrial and nuclear content ............................................................ 42

2 Difference in median fragment length ............................................................................ 43

3 Occurrence of base misincorporation in mitochondrial reads ........................................ 44

S1 Number of adapted molecules quantified after library preparation............................... 55

S2 Comparison of fragment length distributions of total DNA content .............................. 56

S3 Comparison of fragment length distributions of the mapped nuclear DNA content ...... 57

# tables

**INTRODUCTION**

None.

**CHAPTER I - hiGH THROUGHPUT SEQUENCING**

None.

**CHAPTER ii - eNZYMATIC dna REPAIR**

1 Summary of all enzymes used in Nelson and PreCR® repair protocols .......................... 38

2 Summary of the differences observed between enzymatic repair protocols ................ 41

S1 Types of DNA damage .................................................................................................... 54

# DECLARATION OF ACADEMIC ACHIEVEMENT

Two multiple-author publications are included in this thesis and are presented herein with permission from all authors. Minor reformatting was necessary for consistency throughout the thesis. My contributions to each paper (noted NM) are detailed below.

**CHAPTER I - HIGH THROUGHPUT SEQUENCING**

N. Mouttham, J. Enk, R. Fourney and H. Poinar, *STR profiling from next-generation sequencing libraries: accuracy, reproducibility and the technological transition of forensic science.* Submitted for publication, 2014.

This project was conceived by HP and RF. NM, JE and HP designed the experiments and reviewed analyses. NM carried out all experiments, performed data analyses and prepared the manuscript (including the SOM). All authors edited the manuscript.

**CHAPTER II - ENZYMATIC DNA REPAIR**

N. Mouttham , J. Klunk, M. Kuch, R. Fourney and H. Poinar, *Surveying the repair of damaged ancient DNA with high-throughput sequencing.* Submitted for publication, 2014.

This project was conceived by HP and RF. NM, MK and HP designed the experiments and reviewed data analyses. JK performed the sequencing data curation. NM carried out all experiments, performed data analyses and prepared the manuscript (including the SOM). All authors edited the manuscript.

**Note**: References are listed after each manuscript and after each SOM. A fifth list at the end of the thesis includes additional references cited in the introduction and the conclusion. Each reference list is numbered independently in the style required by the journal and may include overlapping references.

Every human being carries with him from his cradle to his grave certain physical marks which do not change their character, and by which he can always be identified - and that without shade of doubt or question. These marks are his signature, his physiological autograph, so to speak, and this autograph cannot be counterfeited, nor can he disguise it or hide it away, nor can it become illegible by the wear and mutations of time. This signature is not his face - age can change that beyond recognition; it is not his hair, for that can fall out; it is not his height, for duplicates of that exist; it is not his form, for duplicates of that exist also, whereas this signature is this man's very own - there is no duplicate of it among the swarming populations of the globe.

*- Mark Twain, “Pudd’nhead Wilson”, 1894* [[1](#_ENREF_1)]

If we had called this “idiosyncratic Southern blot profiling”, nobody would have taken a blind bit of notice. Call it “DNA fingerprinting” and the penny dropped.

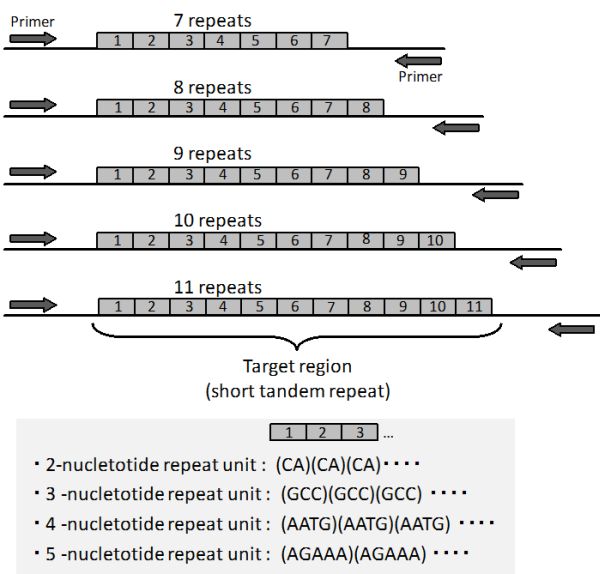
*- Professor Sir Alec Jeffreys, 1996* [[2](#_ENREF_2)]

# INTRODUCTION - FORENSIC SCIENCE AND ANCIENT DNA

Forensic science is defined as the application of scientific knowledge and methodologies to legal problems and criminal investigations. Its lies at the intersection of two vastly different but equally rigorous fields: the science and the justice communities. It has the responsibility of being conservative and always erring on the side of caution, with strict, objective and reproducible methodologies. Two landmark rulings, Frye v. United States in 1923, supplemented by Daubert v. Merrell Dow Pharmaceuticals in 1993, regulate the admission of new scientific evidence by imposing that new methodologies intended for the court must be deemed generally accepted by a meaningful segment of the associated scientific community. As such, forensic science is considered to be a very reliable, but innovation-adverse field. In the mid-1980s, Kary Mullis, along with others, developed PCR [[3](#_ENREF_3)] and Alec Jeffreys discovered short tandem repeat [[4](#_ENREF_4)], revolutionising the field of genetics. Yet, it would take another decade for this technology to replace restriction fragment length polymorphism (RFLP) as the profiling technique of choice in forensic laboratories.

Short tandem repeats (STR) consist of regions of highly repetitive short polynucleotides in the DNA sequence [[5](#_ENREF_5)]. These are the basis of STR typing, or DNA fingerprinting, currently the most widely used forensic DNA methodology [[6](#_ENREF_6)]. By annealing primers on sequences flanking the STR region and performing traditional PCR amplification, it produces amplicons whose sizes are positively correlated to the number of repeats (Fig. 1). While the units can range from two to 13 bases, forensic STR loci are typically tetra- or pentanucleotide repeats.

The polymorphic nature of STRs makes them ideal for genetic identification as they offer powerful discrimination between individuals. The likelihood of two profiles being identical is calculated by multiplying frequencies (i.e., the proportion of individuals in a population that carry X repeats at a given locus) for each STR locus in a profile. Current forensic DNA databases have 13 core loci that are required for a profile to be uploaded, but many laboratories use up to 21 loci. When combining the relative rarity of each allele and the large number of DNA regions observed, the probability of two random individuals having the same profile ranges from one-in-thousands to one-in-trillions.



**Figure 1. The structure of short tandem repeats (STR).** Figure from [[7](#_ENREF_7)].

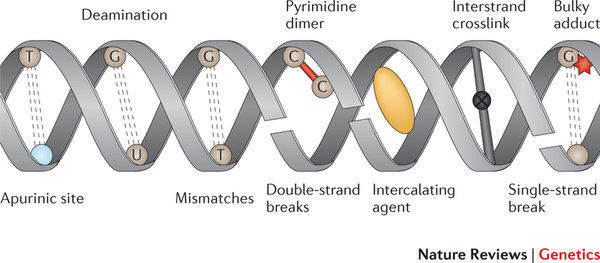
However, there are important limitations inherent to STR analysis. Firstly, the amplicon size ranges from approximately 100 to 450 base pairs (bp). Any sample with high degradation is unlikely to have many DNA fragments above 200 bp, which means that about half of the loci will not amplify, yielding partial profiles and low statistical values. Secondly, even if there are molecules long enough to be amplified at all sites, DNA damage present in the sequence can prevent proper amplification, either by stalling the DNA polymerase or by preventing the primer from annealing to its binding site [[10](#_ENREF_10)]. Finally and most importantly, forensic DNA analyses focus exclusively on the human component of casework extracts: the quantification targets human DNA and the STR amplifications are specific to the human genome [[5](#_ENREF_5)]. By doing so, it excludes all genetic information present in the rest of the metagenome (detailed below).

Today, forensic science faces the same dilemma as it did in the mid-1980s, when new advances in DNA amplification challenged existing standards and protocols. A new technology – high-throughput sequencing (HTS) or next-generation sequencing (NGS) – is emerging from academic research and is quickly establishing itself as a powerful analytical tool, yielding orders of magnitude more genetic information than PCR–based methods, at an ever-decreasing cost. However alluring, a transition to this technology would mean a paradigm shift for forensic science, as STR and HTS are widely different methodologies. One cannot simply be replaced by the other, if for no other reason than current DNA profile databases contain hundreds of thousands of entries world-wide that would require conversion into an HTS-compatible format. Therefore, forensic science has to look towards the work of those in equivalent fields that have already successfully transitioned to HTS and use it to evaluate the feasibility of this shift. One such resource is the ancient DNA community.

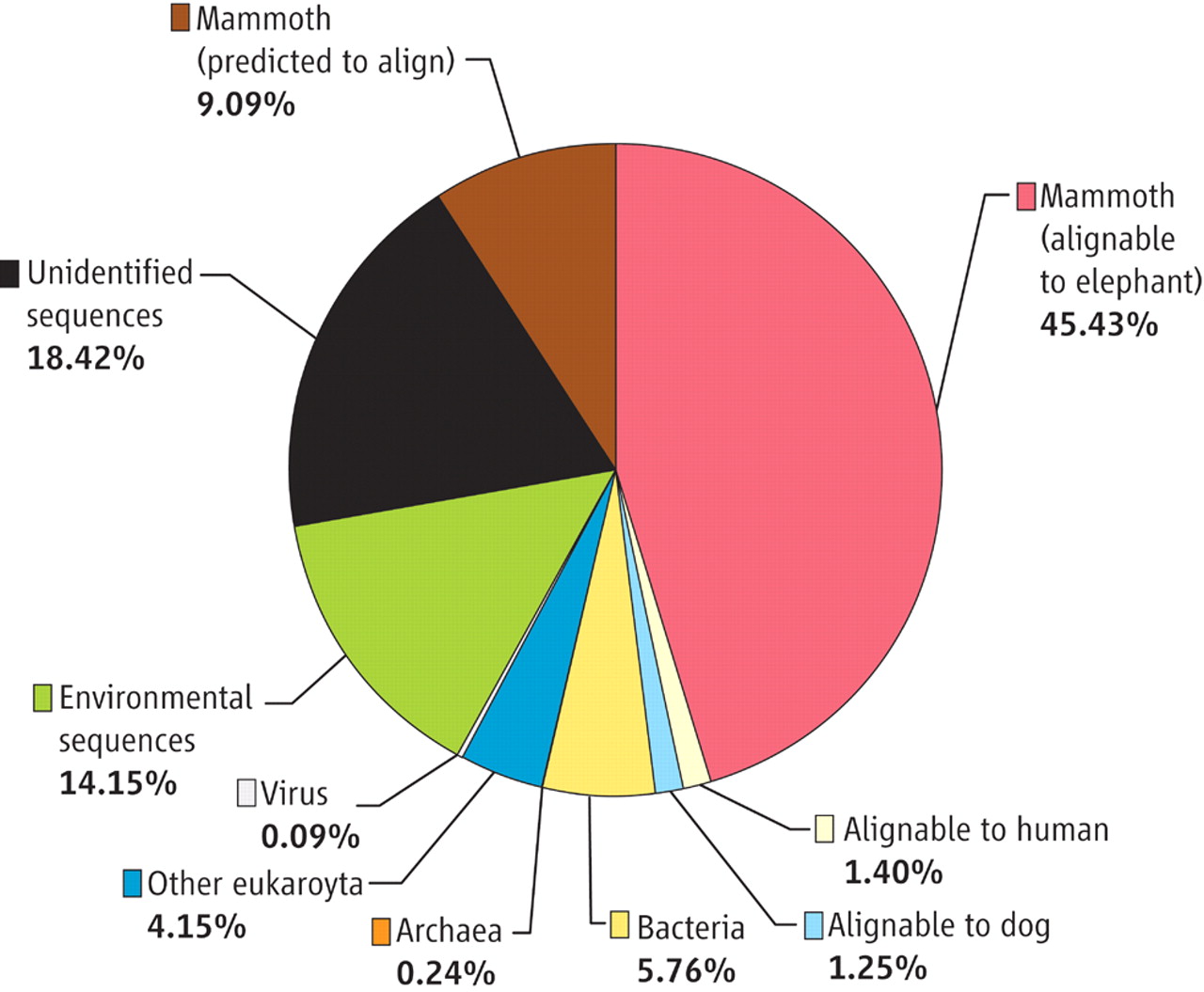
Ancient DNA has many commonalities with forensic DNA: a combination of challenging analyses performed on unique or irreplaceable specimens [[11](#_ENREF_11)], defined by:

* heavily fragmented DNA, where most strands are under 200 bp in length, with a median around 50-70 bp [[12](#_ENREF_12)];
* high amounts of DNA damage (Fig. 2), including 10-15 bp termini that tend to be single-stranded and that contain more damage than double-stranded DNA;
* low levels of endogenous DNA, in comparison to the exogenous background that is present (red boxes in Fig. 3).

Endogenous and exogenous content can be defined as the target and non-target molecules that together make up a metagenome. The exogenous background provides a wealth of information concerning the specimen's environment at the time of death: bacteria and viruses from the microflora are indicators of diet and lifestyle, while plants and fungi provide details of living conditions such as climate or habitat. This information is present in every DNA extract and while it is currently difficult and onerous to analyze, it should be carefully preserved until new analytical tools are developed or refined. One way to achieve this is to place extracts into sequencing libraries; the universal adapters ligated to either end of the fragment allow for repeated downstream amplification and effectively immortalizes the entire metagenome for future analyses.

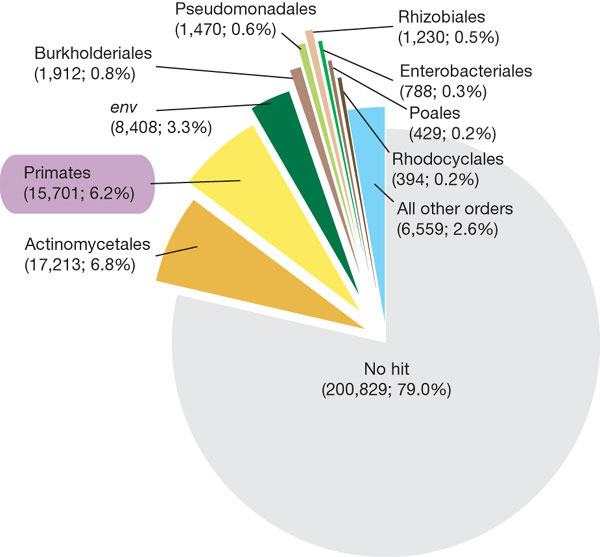


**Figure 2. An overview of types of DNA damage**. Figure from [[13](#_ENREF_13)].



**A**

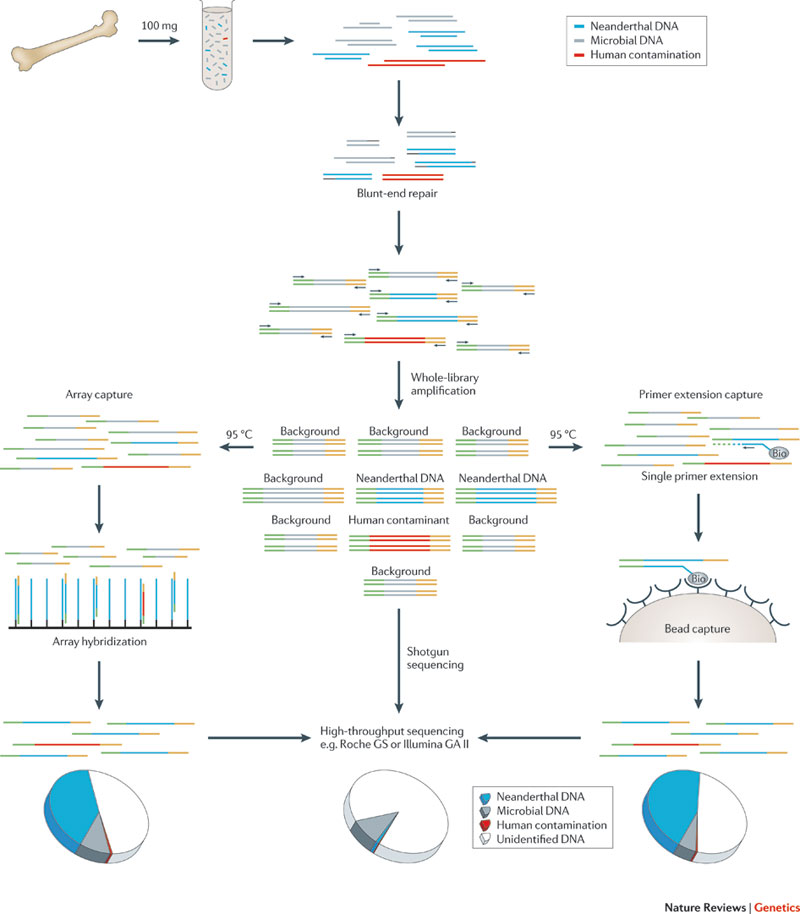
**B**



**Figure 3. Characterization of two metagenomic libraries, including percentage of read distributions to various taxa**. The most populous taxonomic orders are shown and the endogenous fraction is circled in red. Panel A represents a mammoth specimen and is modified from [[14](#_ENREF_14)]. Panel B represents a Neanderthal specimen and is modified from [[15](#_ENREF_15)].

Most ancient samples no longer contain sufficient DNA to be quantified by traditional methods, but the field of ancient DNA has circumvented this issue by developing innovative ways of exploiting extremely short fragments, such as those characteristic of very old or degraded samples. High-throughput sequencing can be combined with targeted genomic enrichment, where fragments from the desired endogenous genome are captured and separated from the exogenous background. This technique increases the proportion of reads available for sequencing, as highlighted in Fig. 4 by the relative increase of the blue Neanderthal portion in the left and right pie charts (showing enriched DNA extracts), compared to the center chart (depicting direct sequencing). Due to these technological developments, multiple ancient genomes have been reconstructed and analyzed, to great success, such as both mitochondrial and nuclear Neanderthal genomes [[16-18](#_ENREF_16)]; a high-coverage sequence of the Denisovan genome [[19](#_ENREF_19)]; *Yersinia pestis* genomes of strains responsible for the Plague of Justinian [[20](#_ENREF_20)] and the Black Death [[21](#_ENREF_21)]; *Vibrio cholerae* strains from the 1800s [[22](#_ENREF_22)], and complete mitochondrial genomes from Columbian mammoth specimens [[23](#_ENREF_23)]. These papers demonstrate that DNA can be obtained from hominins tens of thousands of years-old, centuries-old bacteria and species which have been extinct for more than 3,500 years. Thus, ancient DNA is a great example of the variety and depth of data to which forensic science can expect.

The objective of this thesis was to evaluate the inevitable transition of forensic science to HTS sequencing via the STR amplification of libraries and to assess whether novel repair strategies, like those used on ancient DNA specimens, may help render forensic samples more amenable to downstream analyses. Before samples can be sequenced, each must be converted into libraries, which are required for high throughput sequencing. The focus of this work, detailed in Chapter I, was to determine whether fragments converted into sequencing libraries could yield the same STR profile as pre-library extracts. Additionally, DNA damage is the most significant limitation in the retrieval of genetic information from many forensic and ancient specimens. To circumvent these limitations, it would be very beneficial to reverse or limit the effects of DNA decay before specimens are analysed. To that effect, two enzymatic DNA repair kits were evaluated on several mammoth samples (which act as a proxy to forensic extracts). Chapter II presents the results of these analyses, where the effect of active versus heat-inactivated enzymatic DNA repair are compared.



**Figure 4. A schematic representation of high-throughput sequencing of DNA from fossil remains, depicted as a Neanderthal bone**. The ancient DNA is first blunt-end repaired and then DNA adapters are added to each termini. This final sequencing library product serves as the input for various high-throughput sequencing strategies and technologies. All ancient DNA molecules in the library will first be PCR amplified using the adaptors as priming sites. Aliquots that contain copies of all original ancient DNA molecules can be directly sequenced on a high-throughput sequencer (centre panel) or used in targeted enrichment via array (left panel) or primer extension capture (right panel) methods. The pie charts illustrate the percentage of Neanderthal DNA obtained by each of these approaches. Figure and legend from [[24](#_ENREF_24)].

# CHAPTER I - HIGH THROUGHPUT SEQUENCING

## STR profiling from next-generation sequencing libraries:

*accuracy, reproducibility and the technological transition of forensic science*

Nathalie Mouttham a,b,Jacob Enk a,b,c,Ron Fourney d,Hendrik Poinar a,b,e

a McMaster Ancient DNA Centre, McMaster University, 1280 Main Street West , Hamilton, ON, L8S 4L8, Canada, b Department of Biology, McMaster University1280 Main Street West , Hamilton, ON, L8S 4L8, Canada, c MYcroarray, 5692 Plymouth Road, Ann Arbor, MI 48105, United States, d Royal Canadian Mounted Police, Science and Strategic Partnerships, Forensic Science and Identification Services, 1200 Vanier Parkway, K1A 0R2, Ottawa, ON, Canada, e Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, 1280 Main Street West , Hamilton, ON, L8S 4L8, Canada

This version of the manuscript was submitted for publication, December 2014.

Keywords: degraded DNA / library preparation / next-generation sequencing / STR amplification

Copyright ©2014 by the authors. All rights reserved.

Content used with permission.

### Abstract

### 

Historically, forensic DNA analysis methods are upgraded as new technologies emerge and are deemed reliable by comparable fields of study. Consequently, it is beneficial to preserve samples that fail to produce DNA profiles with current technologies until novel methods become available. One potential way to preserve DNA extracts for future characterization is to convert them into next-generation sequencing libraries. Library preparation effectively "immortalizes" all the constituents of a forensic extract by attaching known adapter sequences at either end of each fragment, which can serve as primer binding sites for repeated global PCR amplification or sequencing. However, for a method of preservation to be most effective, it should be compatible with existing methods. Here we assessed whether immortalized DNA libraries can still be profiled accurately using a traditional forensic STR assay. Blunt-end library preparation was performed on artificially-degraded, single-source extracts and on three-person mixtures, before being amplified using PowerPlex® 16HS. We observed that STR amplification profiles of library-prepared extracts are of markedly reduced quality, exhibiting increased allelic drop-out, heterozygous imbalance and extensive stuttering ratios. While existing library preparation protocols could be optimized to limit these artefacts, an alternative would be to encourage the forensic science community towards a new era of sequencing-based profiling, already successfully implemented in clinical diagnostics and genomics research.

### Introduction

Short tandem repeat (STR) amplification is the most commonly used technique for DNA profiling in forensic science [1]. While many samples will yield full STR profiles, a significant proportion fails to do so, due to low-template copy numbers, polymerase inhibition and nucleic acid damage [2] [3]. Forensic DNA tends to be heavily fragmented (i.e. 30-150bp), with median lengths often much lower than that required in a typical PCR amplification. Furthermore, the low endogenous content is mixed with substantial levels of non-target DNA from the environment or from the specimen’s natural microbiome. In addition, DNA degradation typical to forensic DNA includes abasic sites, single-stranded nicks and damaged nucleotides that can lead to base misincorporation and DNA polymerase stalling [4]. To mitigate these issues, a few avenues have been explored, including STR amplification kits optimized to overcome PCR inhibition [5] and the use of shorter loci, termed mini-STRs, with amplicons less than 200bp [6-8]. Recent work has focused on SNP-calling [9-11], as well as sequencing of STR amplicons [12, 13].

Though robust and widely implemented, there is little room for improvement on the current STR methodology beyond finding more discriminatory loci to use or improving on PCR enzymes and buffers. Additionally, since traditional STR amplification focuses on the human constituent of a forensic extract, it excludes the information available in the rest of the metagenomic profile. However, emerging technologies are bound to improve access to forensic DNA specimens in the future. Therefore it is in the interest of forensic science to preserve precious evidentiary material until it can be profiled with new technologies once validated.

Fortunately, comparable fields of study, in particular ancient DNA science, have succeeded in overcoming many of the difficult properties of forensic DNA by employing new technologies. Ancient DNA (aDNA) is typically defined as DNA extracted from archival, forensic, archaeological and paleontological remains [14-18]. Researchers have pursued various avenues to overcome the same difficulties as present in forensic DNA, by improving current techniques or by moving to novel genomic methodologies. As such, they have acquired genetic information from remains where the conventional forensic technical repertoire would have been unsuccessful. The most significant improvement the aDNA field has employed, to great success, is the use of next-generation sequencing (NGS). This emerging technology has quickly become a reliable staple in academic genetic research [19]. By increasing access to highly degraded and contaminated specimens, NGS has enabled complete genomic reconstruction of extremely damaged DNA extracts, such as that of a 40,000 years-old archaic Denisovan individual [20-22]. By coupling sequencing with hybridization-based targeted enrichment, aDNA research has overcome many of the issues related to the high exogenous background typical of degraded remains or environmentally-challenged DNA templates [23-25].

It stands to reason that investigative DNA science could benefit from the same technological developments already embraced by the ancient DNA field. Already the latest research in forensics forecasts a shift towards NGS-based profiling [11, 26]. However, any change in methodology must be tempered by the reality that, to stand up in court, techniques are required to be thoroughly validated, reproducible, and to meet standards agreed upon by both the scientific and the justice communities. Any transitional state should be compatible with current protocols, thereby maintaining continuous access to the considerable existing forensic DNA databases.

In order to convert a DNA extract to a sample that can be sequenced with NGS, universal “adapters” must be ligated to both ends of all individual DNA fragments. These short sequences serve as known primer-binding sites that enable subsequent whole-genome amplification and eventually sequencing. Additionally, library preparation has a peripheral benefit to forensic science, as it effectively "immortalizes" DNA extracts by allowing all the genetic material present in the original specimen (both target and non-target) to be amplified for repeated use. It also allows for sample-specific “barcoding”, which has important logistical benefits: it enables samples to be pooled for downstream reactions (i.e., NGS or re-amplification), to reveal sample-to-sample contamination and to detect the occurrence of PCR jumping in poor quality templates where base damage is prevalent [27].

There are multiple library preparation methods available, including sticky-end or forked-adapter [28] and single-stranded preparation [29]. However, the most widely-used protocol in ancient DNA is blunt-end library preparation, which ligates distinct double-stranded adapters to the termini of end-polished (blunt) double-stranded DNA fragments (Fig. S1) [30, 31]. The process is well-characterized and numerous protocols are available for various sequencing platforms.

This paper explores whether the gold standard of forensic profiling – STR analysis – could still be performed successfully to the same degree of precision on forensic DNA samples converted to NGS libraries. In order to survey an array of forensic samples seen in casework, the library preparation evaluation was performed on damaged, low-copy single source extracts as well as multiple-contributor mixtures, with the purpose of demonstrating whether a technological ‘transitional state’ could be achieved that allows forensic DNA extracts to be immediately compatible with both current and emerging technologies.

### Materials and Methods

Detailed descriptions of Materials & Methods can be found in the Supplementary Online Material.

*Sample Preparation - Degraded Samples*

Human DNA obtained from multiple buccal swabs was extracted using QIAamp Mini extraction columns (Qiagen, Venlo, Netherlands), according to manufacturer’s instructions. The degradation protocol used was not intended to reproduce the full array of damage common to ancient DNA and expected in forensic extracts, but to mirror the major types of degradation [18]. A simulation of natural decay was performed according to a modified version of Swango *et al*. [32]. Pooled DNA extracts were mixed with 10X Mg2+ DNase I buffer and H2O. Sample was sonicated, creating a fragment size ranging from 35 to 2000 bp (see Fig. S2). An aliquot was digested with DNase I (Fermentas, Burlington, ON) and then inactivated by EDTA addition and by heating. In order to simulate exogenous background DNA, the degraded human DNA was mixed with DNA extracted from a >30,000 years old woolly mammoth (*Mammuthus primigenius*) specimen, using bone demineralization, digestion and organic extraction described in Poinar *et al*. [33] and Schwarz *et al*. [34]. Artificially-degraded human DNA stock was diluted with mammoth DNA extract and H2O, to a final concentration of 23, 43, 90 and 172 pg/µl.

*Sample Preparation - Mixture Samples*

Mixture samples contained DNA from three modern human donors mixed in varying ratios. Human DNA obtained from buccal swabs was extracted using QIAamp Mini extraction columns (Qiagen), according to manufacturer’s instructions. The total amount of human DNA in each sample was standardized to 80-100 pg/µl. The final Female:Male1:Male2 donor ratios were: 1:1:1; 3:2:1; 6:3:1 and 9:4:1.

*Single-locus qPCR Quantification*

In order to track human DNA input into library preparation and STR amplification, samples were quantified using an 81 bp qPCR assay targeting the *c-Myc* nuclear sequence [[35](#_ENREF_35)]. Quantifications were carried out using a CFX-96 instrument, with real-time detection (Bio-Rad, Hercules, CA). The reaction was prepared using AmpliTaq Gold® DNA polymerase (Life Technologies, Carlsbad, CA), with 1 µl of DNA template.

*Library Preparation and Indexing Amplification*

The blunt-end library preparation protocol was taken from Meyer & Kircher [[30](#_ENREF_30)], with the modifications described in Kircher *et al*. [[31](#_ENREF_31)] for indexing amplification. All samples were indexed in a 10-cycle amplification reaction using a CFX-96 thermocycler with real-time fluorescence detection, with forward and reverse primers and Herculase II Fusion DNA polymerase (Agilent, Santa Clara, CA).

*STR Amplification and Data Analysis*

Up to 1.2 ng of DNA template was amplified using PowerPlex® 16HS (Promega, Madison, WI). Samples were amplified on a CFX-96, using a protocol from Laurin *et al* [[36](#_ENREF_36)]: 1min/96°C, 10 cycles of 30s/94°C, 30s/60.5°C (0.5°C/s), 45s/70°C (0.2°C/s), 19 cycles of 30s/90°C, 30s/60.5°C (0.5°C/s), 45s/70°C (0.2°C/s), 30min/60°C, hold at 4°C. Samples were prepared for capillary electrophoresis by mixing 1 µl of STR amplicon with 8.67 µl of Hi-Di™ Formamide (Life Technologies) and 0.33 µl of Internal Lane Standard 600 (Promega). Amplicon separation was performed on an ABI 3730 Genetic Analyzer (Applied Biosystems, Carlsbad, CA). STR profiles obtained were analyzed with GeneMapper® *ID* software v3.2 (Applied Biosystems), with the peak detection threshold set at 100 RFU (approximately twice the intensity of the background on ABI 3730) and marker-specific stutter ratios defined by Promega.

### Results

The success of performing STR amplification on blunt-end sequencing libraries was evaluated against three criteria: the presence or absence of alleles, heterozygous peak height balance, and loci-specific stutter ratios.

Allelic drop-out is defined as the absence of one or more expected alleles from a profile [2]. Based on known profiles, the number of alleles successfully designated before and after library preparation was established (Figure 1 and Figure S3). Given the median fragment lengths of the damaged DNA and the low input quantities, partial profiles were expected even prior to library preparation (Fig. 1). Notably however, there was a significant increase in allelic drop-out following library preparation, with at most only 17 out of 28 (60%) alleles successfully amplifying. That number decreases to 4 alleles out of 28 (14%) for the lowest input sample.

The same trend was observed in all mixture samples (Fig. S3), which were assessed without taking into account the respective contributions to the mixtures (i.e. there was no correction applied to take into consideration low copy numbers of the minor contributors). As with the degraded DNA extracts, the mixture samples showed increased allelic drop-out following library preparation, though not statistically significantly.

**Figure 1: Number of alleles successfully designated in STR profiles obtained with PowerPlex® 16HS from artificially-degraded DNA, before (in light grey) and after (in dark grey) library preparation.** Values are obtained from the median of four technical replicates and error bars represent one standard deviation. The dotted line represents the number of alleles in a complete profile (n = 28). A paired two-tailed student’s t-test indicates statistically significant increase in allelic dropout following library preparation (p = 0.01).

Next, heterozygous peak ratios were used to determine whether alleles in the STR profile are evenly represented (Fig. 2). Theoretically, in a balanced STR profile, two heterozygous alleles will contribute equal amounts, and the ratio of their peak heights should be close to one. As the peaks become increasingly imbalanced, this ratio will decrease. Heterozygous loci prior to library preparation have satisfactory peak ratios (as per SWGDAM guidelines) [[2](#_ENREF_2)]. However, post library preparation, there is a systematic decrease in heterozygous ratios, significantly in some loci (Amelogenin, TH01, D13S317 and FGA). Interestingly, peak imbalance does not appear to be dependent on amplicon size. In two loci (D16S539 and Penta D), both alleles dropped out.

**Figure 2: Peak height ratios of heterozygous alleles successfully designated in STR profiles obtained with PowerPlex® 16HS from artificially-degraded DNA, before (in light grey) and after (in dark grey) library preparation.** Loci (on X-axis) are listed from shortest to longest. Values are obtained from the mean of four technical replicates and error bars represent one standard deviation. An unpaired two-tailed student’s t-test indicates statistically significant decrease in heterozygous peak balance following library preparation in loci marked with an \* (p = 0.05).

Finally, marker-specific stutter ratios at given loci were evaluated (Fig. 3). Stuttering is an amplification artefact that creates an allele that is one repeat shorter (i.e., n- 4) than the expected peak. This phenomenon is well documented and each STR kit manufacturer publishes acceptable loci-specific stutter ratio values. They are calculated by dividing the peak height from the artefact over the peak height from the true allele (expressed in %). In nine of 16 loci, stutter ratios observed were above acceptable values as recommended when following optimal amplification conditions (in light grey). In some cases, the increase in stutter ratios was exceptional (i.e., D7S820 increased from the accepted 11% to over 18%). This phenomenon was not observed when amplifying directly from DNA extracts, positive amplification controls or non-indexed libraries (data not shown).

**Figure 3: Marker-specific stutter ratios observed in STR profiles obtained with PowerPlex® 16HS after library preparation.** Loci are listed from shortest to longest. Light grey bars indicate acceptable stutter ratios, as determined by the manufacturer. Dark grey bars show the values observed in STR profiles. The error bars indicate the standard deviation obtained from n samples (each n value indicated below locus name). The increase in stuttering is statistically significant with a one-tailed, one sample t-test (p = 0.05).

### Discussion

For both single source and multiple contributor mixture samples, library preparation resulted in an overall decrease in STR quality, with higher allelic drop-out, heterozygous peak imbalance and larger stutter ratios. This observation was noted even with a relatively high STR amplification input of one nanogram of human DNA, and therefore is unlikely due to the stochastic nature of low copy number samples. This is corroborated by the mixtures evaluation, which showed that allelic drop-out is not linked to the relative contribution of either constituent in the mixture. For example, there are instances where homozygous alleles from the major contributor dropped out while the heterozygous allele from the minor contributor was preferentially amplified.

These issues were not encountered when amplifying directly from pre-library (non-degraded) extracts, positive amplification controls or non-indexed libraries. This suggests that library amplification is a potential cause, where 8-10 cycles of indexing amplification compound with the 27-29 cycles of STR amplification that follow introduce bias in allele representation. Alternatively, this bias could have occurred further upstream, during adaptor ligation for instance. Further work is needed to test these hypotheses.

While the majority of the profile quality decrease seems to be due to library preparation, some of it can be attributed to the artificial degradation of the single-source samples. Sonication reduced fragment sizes to a median of ~280 bp, with a range of 35-2000 bp (verified by BioAnalyzer traces, Fig. S2) and DNase I with Mg2+ buffer produce single-stranded nicks [37]. This combination reduces the efficacy of amplification, due to DNA polymerase stalling on unrepaired gaps or to damage in the primer binding site. It likely accounts for the allelic drop-out and the slight heterozygous peak imbalance seen prior to library preparation of the extract (Fig. 1 and 2), both of these effects have been documented previously [2].

There are also limitations associated with blunt-end library preparation. First, the theoretical adaptation rate is 50% at best: any fragment with the same adapter at either end cannot be used for sequencing and is difficult to amplify [28]. Second, it requires two or three purification steps, each of which results in DNA loss [21].

These artefacts pose serious limitations for forensic profile interpretation. Allelic drop-out occurs when the number of template copies available is too low or when primers cannot anneal to their binding sequences. Both result in amplification failure and the absence of an expected allele at one or more loci (i.e., partial profiles). It can result in the misinterpretation of a profile by assuming the presence of a homozygous donor, when in fact a second allele is present. Heterozygous peak imbalance and increased stutter ratios will imply the presence of another contributor to the mixture or could skew the interpretation of a given donor's contribution (i.e., a minor component might seem like the major contributor). Such events reduce the discriminatory power of DNA typing and could present integrity and confidence issues for profile inclusion into forensic DNA databases.

In order to reproduce, to the greatest extent possible, the metagenomic content of a forensic extract, the degraded human DNA stock was diluted using both water and a mammoth DNA extract, thus creating an exogenous DNA background. The mammoth content increases the complexity of the forensic extract and the number of fragments available for library preparation, thereby reducing the stochastic effect due to low input copy number, without overpowering the human constituent of the sample, as the mammoth should have negligible effects on the human qPCR assay. This approach is a modification of a previously used technique termed "carrier blanks", commonly used as a control in aDNA research, and which may be considered for inclusion in forensic casework [38].

### Conclusion

Blunt-end library preparation for NGS sequencing was performed on artificially-degraded single-source and on three-person mixtures to mimic forensic DNA extracts. When libraries were amplified with a forensically-validated STR multiplex, there was a significant decrease in the quality of the profile, due to an increase in allelic drop-out, heterozygous peak imbalance, as well as an increase in stutter ratios. As such, until library preparation is significantly improved to reduce stutter and more faithfully reproduce the relative frequency of starting molecules, it is not recommended to perform STR profiling on blunt-end NGS libraries for forensic-based DNA applications. In order to step into a new era of profiling, forensic science community should consider seizing the full potential of NGS, in particular SNP-based DNA typing, which is less susceptible to the biases seen in multiplex amplification and tandem repeat counting.

### Acknowledgments

The authors thank current and former members of the McMaster Ancient DNA Centre for their invaluable guidance and input throughout this entire project. We also thank Dr. J. Robertson and S. Howlett at the FBI laboratories for graciously providing the initial degraded DNA. We thank M. Nesbitt and J. Slobodan at Coastal Genomics for their help in size selection. NM personally thanks donors who volunteered their DNA for the advancement of science. Comments from JB and SM greatly improved this manuscript. This work was funded by the Canadian Police Research Center 02-3027. Ethics approval for the study was granted by Hamilton Health Sciences and McMaster University.

### References

[1] J.M. Butler. Advanced Topics in Forensic DNA Typing: Interpretation: Academic Press; 2014.

[2] SWGDAM. SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories (2010)

[3] C. Capelli, F. Tschentscher, V.L. Pascali, "Ancient" protocols for the crime scene? Similarities and differences between forensic genetics and ancient DNA analysis, Forensic Sci Int 131 (2003) 59-64

[4] E.C. Friedberg, G.C. Walker, W. Siede. DNA repair and mutagenesis. Washington, D.C.: ASM Press; 1995.

[5] M.G. Ensenberger, J. Thompson, B. Hill, K. Homick, V. Kearney, K.A. Mayntz-Press, *et al*., Developmental validation of the PowerPlex 16 HS System: an improved 16-locus fluorescent STR multiplex, Forensic Sci Int Genet 4 (2010) 257-264

[6] V. Seidenberg, F. Schilz, D. Pfister, L. Georges, L. Fehren-Schmitz, S. Hummel, A new miniSTR heptaplex system for genetic fingerprinting of ancient DNA from archaeological human bone, Journal of Archaeological Science 39 (2012) 3224-3229

[7] P. Grubwieser, R. Muhlmann, B. Berger, H. Niederstatter, M. Pavlic, W. Parson, A new "miniSTR-multiplex" displaying reduced amplicon lengths for the analysis of degraded DNA, Int J Legal Med 120 (2006) 115-120

[8] J.M. Butler, Y. Shen, B.R. McCord, The development of reduced size STR amplicons as tools for analysis of degraded DNA, J Forensic Sci 48 (2003) 1054-1064

[9] C. Borsting, H.S. Mogensen, N. Morling, Forensic genetic SNP typing of low-template DNA and highly degraded DNA from crime case samples, Forensic Sci Int Genet 7 (2013) 345-352

[10] A. Freire-Aradas, M. Fondevila, A.K. Kriegel, C. Phillips, P. Gill, L. Prieto, *et al*., A new SNP assay for identification of highly degraded human DNA, Forensic Sci Int Genet 6 (2012) 341-349

[11] K.K. Kidd, A.J. Pakstis, W.C. Speed, R. Lagace, J. Chang, S. Wootton, *et al*., Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics, Forensic Sci Int Genet 12 (2014) 215-224

[12] S.L. Fordyce, M.C. Avila-Arcos, E. Rockenbauer, C. Borsting, R. Frank-Hansen, F.T. Petersen, *et al*., High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, Biotechniques 51 (2011) 127-133

[13] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, D. Deforce, Forensic STR analysis using massive parallel sequencing, Forensic Science International-Genetics 6 (2012) 810-818

[14] E. Willerslev, A. Cooper, Ancient DNA, Proc Biol Sci 272 (2005) 3-16

[15] E. Rizzi, M. Lari, E. Gigli, G. De Bellis, D. Caramelli, Ancient DNA studies: new perspectives on old samples, Genet Sel Evol 44 (2012) 21

[16] S. Paabo, Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification, Proc Natl Acad Sci U S A 86 (1989) 1939-1943

[17] S. Paabo, H. Poinar, D. Serre, V. Jaenicke-Despres, J. Hebler, N. Rohland, *et al*., Genetic analyses from ancient DNA, Annu Rev Genet 38 (2004) 645-679

[18] M. Hofreiter, D. Serre, H.N. Poinar, M. Kuch, S. Paabo, Ancient DNA, Nature Reviews Genetics 2 (2001) 353-359

[19] T.C. Glenn, Field guide to next-generation DNA sequencers, Mol Ecol Resour 11 (2011) 759-769

[20] M. Meyer, M. Kircher, M.T. Gansauge, H. Li, F. Racimo, S. Mallick, *et al*., A high-coverage genome sequence from an archaic Denisovan individual, Science 338 (2012) 222-226

[21] J. Dabney, M. Knapp, I. Glocke, M.T. Gansauge, A. Weihmann, B. Nickel, *et al*., Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments, Proc Natl Acad Sci U S A 110 (2013) 15758-15763

[22] L. Orlando, A. Ginolhac, G. Zhang, D. Froese, A. Albrechtsen, M. Stiller, *et al*., Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse, Nature 499 (2013) 74-78

[23] K.I. Bos, V.J. Schuenemann, G.B. Golding, H.A. Burbano, N. Waglechner, B.K. Coombes, *et al*., A draft genome of Yersinia pestis from victims of the Black Death, Nature 478 (2011) 506-510

[24] A.M. Devault, G.B. Golding, N. Waglechner, J.M. Enk, M. Kuch, J.H. Tien, *et al*., Second-pandemic strain of Vibrio cholerae from the Philadelphia cholera outbreak of 1849, N Engl J Med 370 (2014) 334-340

[25] J. Enk, J.M. Rouillard, H. Poinar, Quantitative PCR as a predictor of aligned ancient DNA read counts following targeted enrichment, Biotechniques 55 (2013) 300-309

[26] J. Weber-Lehmann, E. Schilling, G. Gradl, D.C. Richter, J. Wiehler, B. Rolf, Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing, Forensic Sci Int Genet 9 (2014) 42-46

[27] O. Handt, M. Hoss, M. Krings, S. Paabo, Ancient DNA: methodological challenges, Experientia 50 (1994) 524-529

[28] Z. Zheng, A. Advani, O. Melefors, S. Glavas, H. Nordstrom, W. Ye, *et al*., Titration-free 454 sequencing using Y adapters, Nat Protoc 6 (2011) 1367-1376

[29] M.T. Gansauge, M. Meyer, Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA, Nat Protoc 8 (2013) 737-748

[30] M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing, Cold Spring Harb Protoc 2010 (2010) pdb prot5448

[31] M. Kircher, S. Sawyer, M. Meyer, Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform, Nucleic Acids Res 40 (2012) e3

[32] K.L. Swango, M.D. Timken, M.D. Chong, M.R. Buoncristiani, A quantitative PCR assay for the assessment of DNA degradation in forensic samples, Forensic Sci Int 158 (2006) 14-26

[33] H.N. Poinar, C. Schwarz, J. Qi, B. Shapiro, R.D. Macphee, B. Buigues, *et al*., Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA, Science 311 (2006) 392-394

[34] C. Schwarz, R. Debruyne, M. Kuch, E. McNally, H. Schwarcz, A.D. Aubrey, *et al*., New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains, Nucleic Acids Res 37 (2009) 3215-3229

[35] P.A. Morin, K.E. Chambers, C. Boesch, L. Vigilant, Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus), Mol Ecol 10 (2001) 1835-1844

[36] N. Laurin, A. DeMoors, C. Fregeau, Performance of Identifiler Direct and PowerPlex 16 HS on the Applied Biosystems 3730 DNA Analyzer for processing biological samples archived on FTA cards, Forensic Sci Int Genet 6 (2012) 621-629

[37] J. Sambrook, Russell, D.W. . Molecular Cloning: A Laboratory Manual. 3rd ed. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2001.

[38] Z. Xu, F. Zhang, B. Xu, J. Tan, S. Li, L. Jin, Improving the sensitivity of negative controls in ancient DNA extractions, Electrophoresis 30 (2009) 1282-1285

### Supplementary Online Materials

**Materials & Methods**

*1. Sample Preparation*

*1.1 Mammoth DNA Extraction*

DNA was extracted from a mammoth (*Mammuthus primigenius*) specimen from the collection of the McMaster Ancient DNA Centre. 100-200 mg of bone material was broken down into ~ 1 mm2 pieces before undergoing bone demineralization and proteinase K digestion, followed by organic DNA extraction, based on protocols in Poinar *et al* [[1](#_ENREF_1)] and Schwarz *et al* [[2](#_ENREF_2)]. Bone fragments were demineralized twice by an overnight incubation (20-24h) on a shaker at 900 RPM and at room temperature with 0.5 M EDTA pH 8.0, after which the supernatant was removed. The bone pellets were digested with a buffer consisting of 10 mM Tris-Cl pH 8.0, 0.5% Sarcosyl, 250 µg/ml Proteinase K, 5 mM CaCl2, 50 mM DTT, 1% PVP and 2.5 mM PTB and incubated at 55°C for 5h at 700 RPM, to ensure homogenization. DNA from both demineralization and digestion supernatants was extracted using classic phenol-chloroform-isoamyl alcohol extraction methods. All aqueous phases were concentrated over 30 kDa Amicon®Ultra 0.5 ml columns (Millipore, Billerica, MA). Concentrates were purified following extraction using a modified MinElute PCR Purification protocol (Qiagen, Venlo, Netherlands), where samples are mixed to PB buffer in a 1:6 ratio, centrifuged at 3,300 x *g* for 1 min, followed by two 700 µl PE buffer washes at 3,300 x *g*. Normal protocol resumes when PE buffer is discarded and samples are centrifuged for 1 min at maximum speed. Samples were eluted in 10 µl of EB buffer by spinning at maximum speed for 1 min after standing for 5 min and pooled.

*1.2 Degraded Samples*

Four samples were created by mixing varying quantities of degraded human DNA into a mammoth DNA background. Human DNA obtained from multiple buccal swabs was extracted using QIAamp Mini extraction columns (Qiagen, Venlo, Netherlands), according to manufacturer’s instructions. Eluates were pooled and concentrated using a 30 kDa Amicon® Ultra 0.5 ml column, to a final concentration of 468 ng/µl. Degradation was performed according to a modified version of Swango *et al.* [[3](#_ENREF_3)]. 16.4 μg of total DNA was added to 10X DNase I buffer and ddH2O, for a total volume of 50 µl. Mg2+ DNase buffer was used to create random single-stranded nicks in the dsDNA fragments (digestion in a Mn2+ buffer creates double-stranded nicks) [[4](#_ENREF_4)]. Sample was sonicated on a Covaris S220, creating a fragment range from 35 to 2000 bp. An aliquot quantified at 9.9 μg of total DNA was digested with DNase I (Fermentas, Burlington, ON), by adding 0.003 U/μl to the sample, for a total volume of 25 μl, for 1 min at room temperature. DNase I was deactivated by adding 2.5µl of 50 mM EDTA to reaction and by heating sample to 65°C for 15 min, followed by 0.1°C/s decrease to 20°C, in order to mitigate denaturation. Final concentration of degraded DNA stock was 195 ng/µl of human DNA. Degradation was monitored via fragment analysis on a 2100 Bioanalyzer (Agilent, Santa Clara, CA), according to manufacturer’s instructions, to ensure that the fragment length distribution remained the same before and after enzymatic digestion. Degraded human DNA was added to a constant volume of mammoth DNA, for final concentrations of 23, 43, 90 and 172 pg/µl.

*1.3 Mixture Samples*

Samples were made by mixing DNA obtained from three modern human donors at varying ratios. Human DNA obtained from donors was extracted using QIAamp Mini extraction columns (Qiagen), according to manufacturer’s instructions. The total amount of human DNA in each sample was standardized to 1.2-1.5 ng. The final donor ratios for Female/Male1/Male2 were: 1:1:1, 3:2:1, 6:3:1, 9:4:1.

*2. Quantification*

Samples were evaluated using an 81 bp nuclear assay targeting *c-Myc* sequence, where primer sequences can be found [[5](#_ENREF_5)]. This assay was used for endogenous quantification of samples and did not provide information about the total amount of DNA found in the extracts. All quantification reactions were carried out using a CFX-96 instrument, with real-time detection, according to manufacturer's direction (Bio-Rad, Hercules, CA). The reaction was prepared as follows: 1X of 10X PCR Buffer II (Life Technologies), 2.5 mM of MgCl2, 250µM of dNTP mix, 1 mg/ml of BSA, 250nM of forward primer, 250 nM of reverse primer, 0.5X of EvaGreen® (Biotium, Hayward, CA), 0.05 U/µl of AmpliTaq Gold® (Life Technologies), in a final volume of 10 µl (including DNA). 1 µl of DNA extract was added to each well and samples were processed in triplicate, along with a standard curve ranging from 1.0E+00 up to 1.0E+05 theoretical copies, in duplicate. Samples were incubated with the following settings: 5 min/95°C, 45-50 cycles of 30 s/95°C, 30 s/58°C and 30s/72°C, 1 min/72°C and a melt curve from 55°-95°C with 0.5°C increments (to ensure the correct amplicon was obtained). Results were analyzed using Bio-Rad CFX Manager software, v3.1.

*3. Library Preparation*

The blunt-end library preparation protocol followed is a modified version of Meyer & Kircher [[6](#_ENREF_6)], with the modifications described in Kircher *et al*. [[7](#_ENREF_7)]. Whenever samples were purified, a modified MinElute PCR Purification protocol (Qiagen) was used, where samples are mixed to PB buffer in a 1:6 ratio, centrifuged at 3,300 x *g* for 1 min, followed by two 700 µl PE buffer washes at 3,300 x *g*. Normal protocol resumes when PE buffer is discarded and samples are centrifuged for 1 min at maximum speed. Samples were eluted in 30 µl of EB buffer for 5 min.

*4. Indexing Amplification (idAmp)*

All samples were indexed, including the human samples that were analyzed through STR profiling. The reactions were prepared as follows: 1 X Herculase buffer (Life Technologies), 250 nM of dNTP, 500 nM of each index primer, 0.75 X of EvaGreen®, 0.025 U/µl of Herculase polymerase, in a total reaction volume of 30µl. 10µl of DNA template was added to each well and samples were incubated as follows: 2min/95°C, 10 cycles of 15s/95°C, 30s/60°C and 45s/68°C, 90s/68°C. idAmp was performed on a CFX-96, with real-time fluorescence detection, according to manufacturer's direction. Results were analyzed using Bio-Rad CFX Manager software, v3.1. Each sample was purified using a modified MinElute PCR Purification protocol (Qiagen), described above, in a total volume of 25 µl.

*5. STR Amplification*

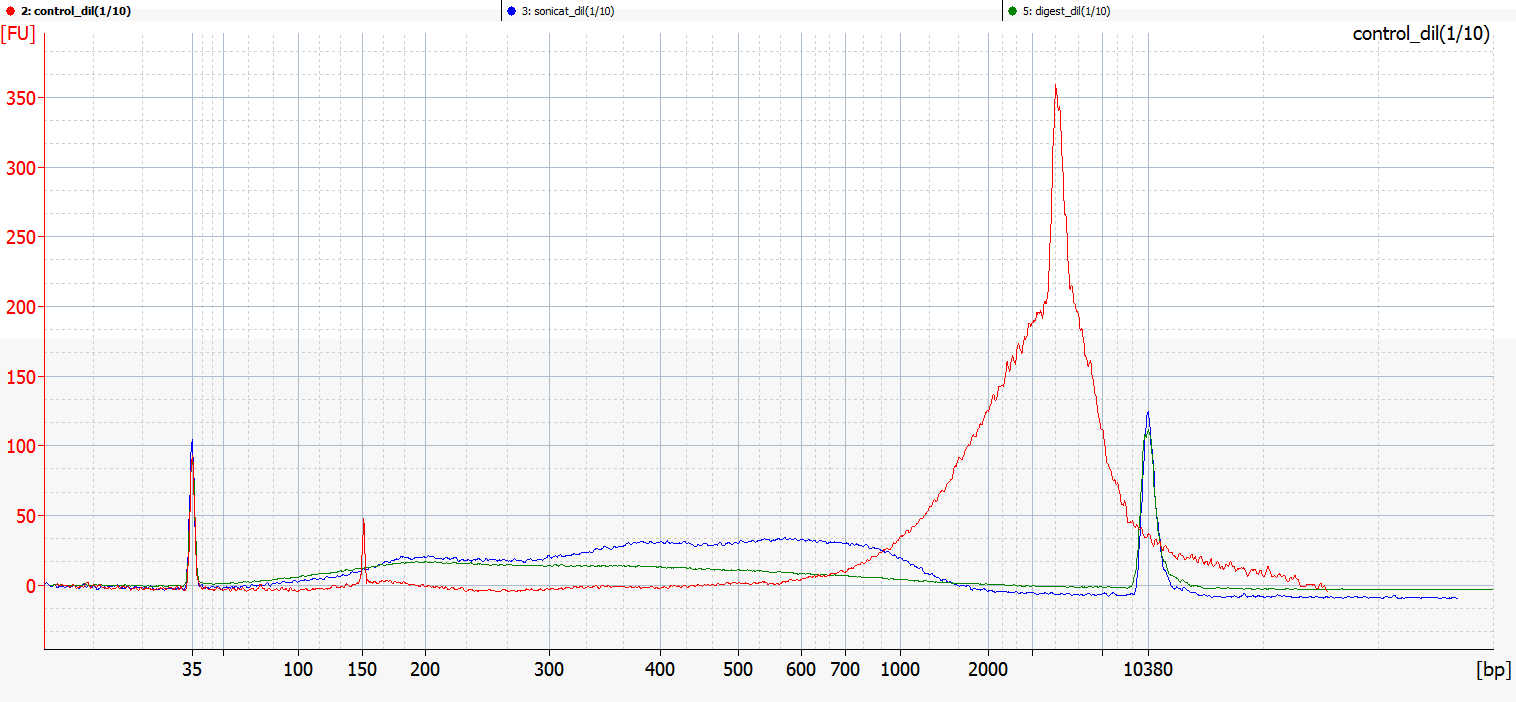
All human samples were analyzed with STR profiling, using Promega's PowerPlex® 16HS amplification system (Madison, WI). Reactions were prepared as follows: 5µl of 5X Gold Star Buffer, 2.5 µl of 10X PowerPlex®16 Primer Mix and ddH2O, in a total reaction volume of 25 µl. Up to 1.2 ng of DNA template was added to each well, depending on sample concentration. Samples were amplified using a protocol from Laurin *et al* [[8](#_ENREF_8)]: 1min/96°C, 10 cycles of 30s/94°C, 30s/60.5°C (0.5°C/s), 45s/70°C (0.2°C/s), 19 cycles of 30s/90°C, 30s/60.5°C (0.5°C/s), 45s/70°C (0.2°C/s), 30min/60°C, hold at 4°C.

*6. Capillary Electrophoresis and STR Data Analysis*

Amplicon separation by capillary electrophoresis was performed on ABI 3730 instruments (Applied Biosystems), according to manufacturer's recommendations. 1 µl of STR amplicon was mixed with 8.67 µl of Hi-Di™ Formamide (Life Technologies) and 0.33 µl of Internal Lane Standard 600 (Promega). Samples were incubated at 95°C for 3 min, followed by 3 min on ice, to denature DNA prior to electrophoresis. Injection parameters were 5s/3V. Run time was increased to 2500 s, as 50 cm capillaries, filled with POP-7 were used. STR profiles obtained were analyzed with GeneMapper® *ID* software v3.2 (Applied Biosystems), with the peak detection threshold set at 100 RFU (approximately twice the intensity of the background on ABI 3730) and marker-specific stutter ratios defined by Promega.



**Figure S1: Representation of blunt-end library preparation of a dsDNA fragment.** Blunt-end adapters have a common sequence (in purple) and a unique sequence (in red and in blue). They are ligated to the enzymatically-blunted ends of dsDNA fragments (in black). Fill-in step extends each adapter to its full length. Indexing primers have: one unique index sequence (in orange), the cluster sequence of the NGS platform (in this case, Illumina, in green) and a sequence that will target the blue or red sequence of the adapters.



**Figure S2: Fragment length distribution of a DNA extract untreated, after sonication and after enzymatic digestion.** Fragments lengths are shown on the X-axis, in bp. Number of fragments at each length are given on the Y axis, as fluorescence units (FU). Untreated DNA extract (in red) was sonicated (in blue) and then digested with DNAse I (in green). Peaks at 35 and 10,380 bp indicate sizing markers.

**Figure S3: Number of alleles successfully designated in STR profiles obtained with PowerPlex® 16HS from three-person DNA mixtures, before (in light grey) and after (in dark grey) library preparation.** The mixtures were based on one major female contributor and two minor male contributors, with the following ratios: 1:1:1 (A), 3:2:1 (B), 6:3:1 (C) and 9:4:1 (D). The black bars are the maximum number of alleles unique to the contributor (not accounting for amount of DNA input). Values are obtained from the mean of two technical replicates and the error bars represent one standard deviation. None of the results are statistically significant with a two-tailed, one sample t-test (p = 0.05)**.**

### Supplementary Online Materials References

[1] H.N. Poinar, C. Schwarz, J. Qi, B. Shapiro, R.D. Macphee, B. Buigues, *et al*., Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA, Science 311 (2006) 392-394

[2] C. Schwarz, R. Debruyne, M. Kuch, E. McNally, H. Schwarcz, A.D. Aubrey, *et al*., New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains, Nucleic Acids Res 37 (2009) 3215-3229

[3] K.L. Swango, M.D. Timken, M.D. Chong, M.R. Buoncristiani, A quantitative PCR assay for the assessment of DNA degradation in forensic samples, Forensic Sci Int 158 (2006) 14-26

[4] J. Sambrook, Russell, D.W. . Molecular Cloning: A Laboratory Manual. 3rd ed. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2001.

[5] P.A. Morin, K.E. Chambers, C. Boesch, L. Vigilant, Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus), Mol Ecol 10 (2001) 1835-1844

[6] M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing, Cold Spring Harb Protoc 2010 (2010) pdb prot5448

[7] M. Kircher, S. Sawyer, M. Meyer, Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform, Nucleic Acids Res 40 (2012) e3

[8] N. Laurin, A. DeMoors, C. Fregeau, Performance of Identifiler Direct and PowerPlex 16 HS on the Applied Biosystems 3730 DNA Analyzer for processing biological samples archived on FTA cards, Forensic Sci Int Genet 6 (2012) 621-629

# CHAPTER II - ENZYMATIC DNA REPAIR

## Surveying the repair of damaged ancient DNA with high-throughput sequencing

Nathalie Mouttham 1,2,Jennifer Klunk 1,2, Melanie Kuch 1, Ron Fourney 3, Hendrik Poinar 1,2,4

*1 McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, Hamilton, ON, Canada, 2 Department of Biology, McMaster University, Hamilton, ON, Canada, 3 Science and Strategic Partnerships, Forensic Science and Identification Services, Royal Canadian Mounted Police, Ottawa, ON, Canada and 4 Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada*

This version of the manuscript was submitted for publication, December 2014.

Keywords: ancient DNA; DNA damage; DNA repair; high-throughput sequencing

Copyright ©2014 by the authors. All rights reserved.

Content used with permission.

### Abstract

Upon cellular death, DNA continues to endure various attacks such as hydrolysis and oxidation, but repair pathways found *in vivo* no longer operate. Because of this, DNA damage in the form of abasic sites, chemically altered nucleotides, and strand fragmentation, is the foremost limitation in obtaining genetic information from many ancient samples. By incubating degraded DNA with specific enzyme combinations adopted from these pathways, it is possible to reverse some of the post-mortem nucleic damage prior to downstream analyses, such as library preparation, targeted enrichment and high-throughput sequencing. Here, we evaluate the performance of two available repair protocols, PreCR® Repair Mix and Nelson repair, on previously characterized DNA extracts from four mammoths. Both methods use endonucleases and glycosylases along with a DNA polymerase-ligase combination. PreCR® Repair Mix increases the number of molecules converted to sequencing libraries, leading to an increase in endogenous content and a decrease in cytosine to thymine transitions due to cytosine deamination. However, Nelson repair demonstrates potential formation of chimerical DNA fragments and remains inconclusive with respect to repair of nucleic damage.

### Methods Summary

Active and heat-inactivated enzymatic DNA repair with PreCR® and Nelson protocols was performed on four mammoth extracts (each in technical duplicates) and one mylodon carrier blank. Repaired extracts were placed into indexed libraries and high-throughput sequenced. Repair efficiency of each method was evaluated based on mitochondrial and nuclear endogenous contents, fragment length distributions and base misincorporation counts, established from three million read subsets of each library.

### Introduction

Ancient DNA (aDNA) is defined as DNA extracted from archival, archaeological, forensic or paleontological specimens ([1](#_ENREF_1)). It is typically highly degraded, as a result of having withstood chemical and physical assaults from post-mortem enzymatic digestion and environmental degradation ([2](#_ENREF_2)). Typical characteristics of an aDNA extract include: short fragment lengths distributed mainly between 30 and 150 bp, due to extensive fragmentation; single-stranded termini DNA where the majority of base damage is located ([3](#_ENREF_3)); low endogenous (or target) DNA content (i.e. mammoth DNA, here), though this can be variable ([4](#_ENREF_4),[5](#_ENREF_5)); high exogenous (or non-target) ancient DNA content, most often comprised of bacteria from the specimen’s microbiome and environment at the time of death and better preserved DNA from modern contamination. It is expected that the damaged DNA in both the endogenous and exogenous fractions will have similar degradation patterns due to similar preservation conditions. Although more contemporary DNA, which could still be centuries old, will likely have longer fragments and less intrinsic damaged sites, the age of a sample has no correlation with the extent of DNA degradation. Many factors, such as the specimen’s preservation conditions or the time until burial, are more reliable factors in predicting long-term DNA survival ([6](#_ENREF_6)). Most ancient samples no longer contain enough quantifiable DNA by traditional methods but the field of aDNA has circumvented this issue by finding innovative ways of exploiting genomic methodologies, such as the use of extremely short fragments via high-throughput sequencing (HTS) ([7](#_ENREF_7)).

Types of typical aDNA damage are summarized in Table S1 but the following three should be highlighted. Apurinic and apyrimidinic sites (AP sites) result from the cleavage of the β-N-glycosidic bond between the base and the deoxyribose ([8](#_ENREF_8)). They occur more frequently in guanine or adenine, due to a weak carbon-purine bond ([9](#_ENREF_9)) and they prevent extension during replication by stalling DNA polymerases. As such, while longer, damaged fragments may be preserved, there will be an amplification bias towards shorter, non-damaged DNA templates. Many nucleotide repair enzymes require an AP intermediate. If they are not removed by endonuclease IV, new abasic sites can be created during enzymatic repair, artificially increasing the level of pre-existing DNA damage. Additionally, single-stranded nicks (ss nicks) in double-stranded DNA (dsDNA) due to breaks in the phosphodiester bond of the deoxyribose backbone are prevalent ([8](#_ENREF_8)) and require the presence of the complementary strand to be repaired. Lastly, deaminated cytosines, caused by the loss of an amine group, become uracil bases, not normally found in DNA. This phenomenon occurs naturally over time and continues after the death of an organism. As ubiquitous as they are, cytosine deaminations have a minor impact in terms of DNA damage, as they will only induce a miscoding lesion in the DNA sequence, without preventing amplification or hindering library preparation. Because they can be observed in the sequencing data, deamination patterns are used as a sample authentication tool in aDNA studies ([3](#_ENREF_3)).

DNA damage occurs in every living organism and is used as a measure of stress or aging ([10](#_ENREF_10)). As such, there are repair mechanisms (nucleotide excision repair and base excision repair pathways) to counteract this perpetual deterioration ([11](#_ENREF_11)). *In-vivo* repair pathways are efficient and usually accurate, though some damaged bases can occasionally be omitted, leading to SNP formation. Upon death, however, DNA continues to face hydrolytic, oxidative and UV attacks that are not repaired post-mortem. The subsequent damage, commonly observed in ancient and forensic samples, is the single most important limitation to retrieving genetic information from deceased organisms. Fortunately, by using enzymatic mixtures very close to those found in the excision repair pathways, some DNA damage may theoretically be reversed. The efficacy of two *in vitro* repair protocols was evaluated on mammoth DNA extracts previously characterized by qPCR (unpublished data). Table 1 summarizes the enzymatic composition of both kits. The first one is the commercially-available PreCR® Repair Mix from NEB ([12](#_ENREF_12)). This kit has been previously tested for forensic application ([13-15](#_ENREF_13)). However, these evaluations focused on STR profiling, where this project, like the majority of current aDNA projects, is HTS-based. The second enzymatic mixture was obtained from a U.S. Department of Justice report, written by Dr. J. Nelson in 2009 ([16](#_ENREF_16)). It was chosen because of the similarities to the PreCR® protocol: they have four common endonuclease/glycosylases enzymes but have different ligase-DNA polymerase combinations and PreCR® contains a uracil-DNA-glycosylase.

**Table 1**. Summary of all enzymes used in Nelson and PreCR® repair protocols

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Enzyme | Nelson | PreCR® | Substrate | Function |
| Endonuclease IV | ✓ | ✓ | AP sites in dsDNA | Cleaves phosphodiester bond at 5' of AP site; removes unsaturated aldehyde and phosphate from 3' termini; leaves 5' deoxyribose-5-phosphate and 3' hydroxyl groups |
| T4 Pyrimidine- dimer- glycosylase (T4 PDG) | ✓ | ✓ | cis-syn-cyclobutane pyrimidine dimers | Cleaves N-glycosydic bond at 5' end of dimer to create AP site; removes one thymine from dimer (second thymine remains as first base at 3' end of AP site) |
| Endonuclease VIII | ✓ | ✓ | Damaged pyrimidines  in dsDNA | Cleaves N-glycosydic bond at 5' end of damaged base to create AP site; cleaves at 5' and 3' end of AP site to create 3' phosphate (via 3' α,β-unsaturated aldehyde) and 5' phosphate termini, respectively |
| Formaminidopyrimidine- DNA- glycosylase (FPG) | ✓ | ✓ | Damaged purines in dsDNA | Cleaves N-glycosydic bond at 5' end of damaged base to create AP site; cleaves at 5' and 3' end of AP site to create 3' phosphate (via 3' α,β-unsaturated aldehyde) and 5' phosphate termini, respectively |
| *Taq* ligase |  | ✓ | dsDNA | Catalyzes formation of phosphodiester bond between juxtaposed 5´ phosphate and 3´ hydroxyl termini in dsDNA; cannot repair ss nicks; activity less optimal at 37°C |
| T4 ligase | ✓ |  | dsDNA | Catalyzes formation of phosphodiester bond between juxtaposed 5' phosphate and 3' hydroxyl termini in dsDNA; can repair ss nicks in dsDNA |
| *Bst* DNA polymerase |  | ✓ | dsDNA | DNA polymerase with 5'-3' exonuclease nick translation; no 3'-5' exonuclease proof-reading; no strand displacement |
| *E.coli* DNA polymerase I | ✓ |  | dsDNA | DNA polymerase with 3'-5' exonuclease proof-reading activity and 5'-3' exonuclease nick translation; requires 3' hydroxyl termini and dsDNA template; cannot translate nicks with 5' deoxyribose-5-phosphate termini |
| Uracil-DNA- glycosylase (UDG) |  | ✓ | Deaminated cytosines  in ss or ds DNA | Cleaves N-glycosydic bond of uracil to create AP site |
| AP site: apurinic or apyrimidinic site; ssDNA : single-stranded DNA; dsDNA: double-stranded DNA | | | | |

### Materials and Methods

Detailed descriptions of the Materials and Methods can be found in the Supplementary Online Material.

*Experimental design*

When deciding on controls for repair efficacy, two options were proposed: to compare active repair reactions to either un-repaired extracts or extracts incubated with a heat-inactivated reaction. The heat-inactivated controls were critical in minimizing which changes to aDNA extracts were caused by non-enzymatic reactions, due to pH, ionic concentration or buffer components. As such, we opted to compare heat-inactivated and active enzymatic repair mixtures.

Two types of negative controls were used in this study: extractions and carrier blanks. Extraction or reagent blanks undergo the same process as the normal samples but do not contain any DNA, allowing for the detection of contamination introduced by reagents. Carrier blanks have the same role as extraction blanks, except they contain DNA from another species (*Mylodon darwinii,* an extinct ground sloth, here). This control, commonly found in aDNA studies, favors the detection of minute amounts of cross-contaminating DNA by creating an artificial exogenous background ([17](#_ENREF_17)).

*Sample choice and DNA extraction*

DNA was extracted from technical duplicates of four mammoth (*Mammuthus* genus) specimens, chosen for their varying endogenous contents, and one mylodon specimen, used as a carrier blank ([17](#_ENREF_17)). Multiple 100-200 mg subsamples of bone material were demineralized and digested and DNA was extracted using a modified proteinase K digestion and subsequent phenol-chloroform-isoamyl alcohol method ([18](#_ENREF_18),[19](#_ENREF_19)).

*DNA repair*

PreCR® repair mix was prepared in 30 µl, according to the manufacturer's instructions (New England Biolabs, Ipswitch, MA) ([12](#_ENREF_12)), while Nelson repair mix was prepared in 30 µl, according to the author's instruction ([16](#_ENREF_16)). Heat-inactivated reactions mixes were incubated at 95°C (PreCR®) or 85°C (Nelson) for 20 min while the active mixtures were kept on ice. Aliquots of 40-50 ng of total DNA were added and reactions were incubated at 37°C for 20 min. To inactivate the reaction, repaired extracts were purified over a MinElute column, following the manufacturer's instructions (Qiagen, Venlo, The Netherlands).

*Library preparation and quantification*

The protocol followed is a modified version of Meyer & Kircher ([20](#_ENREF_20)), with the modifications described in Kircher *et al* ([21](#_ENREF_21)) for double-indexing. A quantification assay targeting the cluster generation sequence was performed, which measured the number of adapted molecules. Based on these values, each sample was normalized to the number of copies in 1µl of the lowest, non-extraction blank sample and was pooled to equal copy number.

*Sequencing and data curation*

All sequencing was performed using a rapid flow cell on a HiSeq® 1500 (Illumina, San Diego, CA), using a 2 x 100 bp HiSeq® Reagent Kit, according to manufacturer's recommendations. Raw data was processed with HiSeq® Control Software (HCS, Illumina) v1.5.15.1 and Real-Time Analysis (RTA, Illumina) v1.13.48.0. Filer conversion and demultiplexing using each 7 bp reverse index (requiring 100% match) were performed using CASAVA (Illumina) v1.8.2.

Sequencing data was curated and aligned, based on the parameters in Enk *et al* ([22](#_ENREF_22)), using cutadapt v.1.2 ([23](#_ENREF_23)), FLASH v.1.0.3 ([24](#_ENREF_24)), Burrows-Wheeler Aligner v.0.6.1-r104 (BWA) ([25](#_ENREF_25)), using the mitochondrial mammoth (*Mammuthus primigenius*) genome (GenBank Accession No JF912200.1) ([26](#_ENREF_26)) and the nuclear African elephant (*Loxodonta africana*) consensus genome (Loxafr3.0 from the Broad Institute, RefSeq Assembly ID: GCF\_000001905.1) as references. Random three million read subsets were taken from each library to analyze endogenous content, fragment length distribution and base mis-incorporations using mapDamage v.2.0 ([27](#_ENREF_27)). A paired, two-sample t-test (p = 0.05) was used to determine statistical significance for all analyses.

### Results and Discussion

A summary and interpretation of the results is presented in Table 2. It is rapidly clear that both active Nelson and PreCR® reaction mixtures have distinct effects on ancient extracts, based on endogenous contents, fragment length distributions and base misincorporation counts.

**Table 2.** Summary of the differences observed between inactive to active Nelson and PreCR® enzymatic repair protocols.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Figure | Nelson | PreCR® |
| # of adapted molecules | S1 | **↘** | Inconclusive |
| Endogenous mitochondrial content | 1 | **↘** | **↗** |
| Endogenous nuclear content | 1 | **↘** | **↗** |
| # of total reads shorter than median | S2 | **↘** | **↗** |
| # of total reads longer than median | S2 | **↗** | **↘** |
| # of nuclear reads shorter than median | S3 | **↘** | **↗** |
| # of nuclear reads longer than median | S3 | **↘** | **↗** |
| Total reads median length | 2 | **↗** | **↘** |
| Mitochondrial reads median length | 2 | **↗** | **↘** |
| Nuclear reads median length | 2 | **↗** | **↘** |
| # of C to T misincorporations | 3 | Inconclusive | **↘ \*** |
| # of G to T misincorporations | 3 | Inconclusive | Inconclusive |
| ↗: increase, ↘: decrease, \* : statistically significant with paired, two-tailed t-test (p=0.05), S: Supplementary figure | | | |

Quantitative PCR measurements targeted to the 5' end of each index show that, given the same library preparation DNA input, there was no significant variation in adaptation or indexing, as the number of adapted molecules is within the same order of magnitude for all extracts (Figure S1).

Endogenous mitochondrial and nuclear contents were calculated based on the number of reads that mapped to the mammoth mitochondrial genome and a masked repeat African elephant nuclear reference genome (Figure 1). Not surprisingly, only 0.008-0.05% (~190-1400 reads) of the total fragments map to the *M. primigenius* mitochondrial genome. Conversely, the nuclear endogenous content is higher, representing almost 25% of the reads (~620,000 reads) for sample A and as low as 2.5% (~75,000 reads) for sample B. This heterogeneity is characteristic of the diversity seen in most aDNA specimens. The presence of "mammoth" reads in the mylodon carrier blank (sample E) are likely due to the mapping of mylodon reads to conserved regions found in both genomes, as the extraction (reagent) blanks yielded a negligible amount of both total (~300-43,000 reads, data not shown) or nuclear mammoth reads (< 15 reads, data not shown).

**Figure 1. Endogenous mitochondrial and nuclear content of four mammoth samples and one mylodon blank, after undergoing active and heat-inactivated DNA repair with Nelson or PreCR® protocols.** Four mammoth extracts (A-D) and one mylodon carrier blank (E) were incubated with active or heat-inactivated DNA enzymatic repair mixtures, either Nelson (in dark and light orange, respectively) or PreCR® (in dark and light blue, respectively). Reads from repaired extracts were mapped to the mitochondrial (panel A) and nuclear reference genomes (panel B). Values are expressed as a percentage of the total number of reads available for each extract, are calculated as the mean of two technical replicates and have error bars showing one standard deviation (when available). Statistically significant variations are marked with a red bracket.

There is an increase in the proportion of shorter fragments from total reads (Figure S2) and from mapped nuclear reads (Figure S3) in active PreCR® when compared to heat-inactivated. Likewise, the average decrease in median fragment length seen in total DNA reads (-5.79 bp) is accentuated in mitochondrial (-6.60 bp) and nuclear (-7.52 bp) mammoth reads (Figure 2), implying that length reduction is occurring preferentially in degraded molecules and presumably in the fragile single-stranded termini. Given the number of sequential enzymatic activities that need to take place to repair heavily degraded DNA molecules, it is probable that additional AP sites and ss nicks are formed during the repair process.

**Figure 2. Difference in median fragment length of the total, mitochondrial and nuclear content of four mammoth samples and one mylodon blank, from heat-inactivated to active DNA repair with Nelson or PreCR® protocols.** Four mammoth extracts (samples A-D) and one mylodon blank (sample E) were incubated with Nelson (in orange) or PreCR® (in blue) enzymatic repair mixtures. Median fragment length for total (Tot.), mitochondrial (Mito.) and nuclear (Nuc.) reads were calculated. The values for each median length can be inferred from the dashed lines of Figures S2 and S3.

In a successful enzymatic repair, one would expect an increase in overall fragment length, as previously unavailable longer fragments are now accessible, and a stable or increasing number of adapted molecules and endogenous content, as more reads are mapped to the reference genome. However, the opposite is seen in Figure S3, corroborated by Figure 1: the nuclear content decreases after Nelson repair. This would imply that the reaction is increasing the fragment length of all reads while simultaneously decreasing the number of endogenous fragments. From these contradictory statements, no plausible hypothesis could be formulated to explain this data when assuming successful repair.

A qualitative assessment of the DNA post repair reactions was established by reads mapped against the *M. primigenius* mitochondrial genome using mapDamage2.0. Cytosine to thymine misincorporations, which are the result of cytosine deamination, are repaired by uracil-DNA-glycosylase (UDG) (Figure 3, panel A). The absence of UDG from the Nelson repair mix accounts for the expected similarities between active and inactive reactions. The decrease in numbers of C to T in active compared to inactive PreCR® is significant in most samples (statistics cannot be established for sample A due to a failed technical replicate). This data suggests at minimum that UDG is functioning properly in the repair reactions.

**Figure 3. Occurrence of base misincorporation in mitochondrial reads from four mammoth samples, after undergoing active and heat-inactivated DNA repair with Nelson or PreCR® protocols.** Four mammoth extracts (A-D) were incubated with active or heat-inactivated DNA enzymatic repair mixtures, either Nelson (in dark and light orange, respectively) or PreCR® (in dark and light blue, respectively). Two types of misincorporation were counted: transition of cytosine to thymine (panel A) and transition of guanine to thymine (panel B). Values are expressed as a percentage of the total number of bases that mapped to the mitochondrial genome, are calculated as the mean of two technical replicates and have error bars showing one standard deviation**.** Statistically significant variations are marked with a red bracket.

Guanine to thymine misincorporations are representative of guanine oxidation (leading to thymine glycol) and are repaired by formaminidopyrimidine-DNA-glycosylase (FPG) (Figure 3, panel B). For both active Nelson and PreCR®, the high variation, large error bars and lack of clear trend in G to T mutations indicate that the activity of FPG is either hindered, highly variable, or the damage is not frequent enough to notice clear trends in repair. FPG requires dsDNA as a substrate in order to repair oxidized purines and this may be a limiting factor in its efficacy in aDNA extracts, which tend to contain many single-stranded overhangs. This observation would apply to any enzyme with the same dsDNA substrate requirement, namely all enzymes in these repair reactions except UDG.

Given these results, it appears that the majority of PreCR® observations stems from UDG activity. It is the only enzyme that can repair the damage-prone 10-15 bp single-stranded fragment termini common in aDNA. UDG cleaves off uracil but leaves behind an unrepaired AP site, that neither endonuclease IV nor *Bst* DNA polymerase is able to repair, as they require a dsDNA substrate. This decreases the length of all fragments but affects the ancient endogenous content more readily than the modern exogenous content (Figure 2). As the phospodiester backbone of unrepaired AP sites eventually breaks, the remaining ssDNA stretch would be end-filled during the DNA library preparation protocol or break off during the reaction. Whether this accounts for the increase in adapted molecules (Figure S1) and endogenous DNA content (Figure 1) is unclear. The results presented here are consistent with the previous STR-based evaluations, where peak height decrease or allelic drop-out were observed in more complex DNA samples equivalent to aDNA specimens ([14](#_ENREF_14),[15](#_ENREF_15)).

The interpretation of the Nelson repair is less straightforward. Since it does not contain UDG, the enzymes can only repair damaged bases in dsDNA. Endonuclease IV creates 5' deoxyribose-5-phosphate termini when it repairs AP sites. Consequently, DNA polymerase I is unable to translate nicks due to AP site cleavage, unlike *Bst* polymerase in PreCR®. Unrepaired AP sites can lead to blunt end double-stranded breaks. Since T4 ligase in Nelson is more efficient at binding blunt-end fragments together than *Taq* ligase in PreCR®, it is hypothesized that active Nelson repair ligates random blunt-end fragments together, creating mammoth-exogenous chimerical molecules. This may explain the overall increase in fragment length (Figure 2) and the decrease in endogenous content (Figure 1 and Figure S3) as these hybrids may be problematic when mapping to the reference genome.

Repairing damaged DNA could have useful applications in many other fields, such as forensic science analysis of highly degraded biological evidence. While the idea of having multiple repair enzymes in a one-step reaction remains attractive for large-scale applications that require robust reproducibility and consistency, it is advised that sequential repair protocols, where one enzymatic reaction is performed at a time, should be explored for optimal DNA repair of ancient extracts.

In conclusion, the repair of ancient DNA extracts with PreCR® Repair Mix yields shorter fragments but minimizes the number of base misincorporations by virtue of deaminated cytosine removal by UDG. The efficiency of other enzymatic components could not be clearly measured and thus it is difficult to assess their efficacy during PreCR® repair. Conversely, DNA repair with the Nelson protocol shows inconclusive results and is not recommended until further analyses have been performed, to evaluate the potential formation of chimerical fragments. Optimizing enzymatic repair will be critical for the genetic analysis of previously unattainable samples, including heavily damaged fossil DNA extracts, forensic biological evidence and medical bio-banks.

### Author Contribution

NM, MK and HP designed the experiments and performed analyses. JK performed the sequencing data curation. NM carried out the experiments and prepared the manuscript. All authors edited the manuscript.

### Acknowledgments

We thank current and former members of the McMaster Ancient DNA Centre for their invaluable guidance and input throughout this entire project. Comments from LM greatly improved this manuscript. This research was funded by the Canadian Police Research Center 02-3027.

### References

1. **Capelli, C., F. Tschentscher, and V.L. Pascali.** 2003. "Ancient" protocols for the crime scene? Similarities and differences between forensic genetics and ancient DNA analysis. Forensic science international *131*:59-64.

2. **Rizzi, E., M. Lari, E. Gigli, G. De Bellis, and D. Caramelli.** 2012. Ancient DNA studies: new perspectives on old samples. Genetics, selection, evolution : GSE *44*:21.

3. **Briggs, A.W., U. Stenzel, P.L. Johnson, R.E. Green, J. Kelso, K. Prufer, M. Meyer, J. Krause, *et al*.** 2007. Patterns of damage in genomic DNA sequences from a Neandertal. Proceedings of the National Academy of Sciences of the United States of America *104*:14616-14621.

4. **Green, R.E., J. Krause, S.E. Ptak, A.W. Briggs, M.T. Ronan, J.F. Simons, L. Du, M. Egholm, *et al*.** 2006. Analysis of one million base pairs of Neanderthal DNA. Nature *444*:330-336.

5. **Meyer, M., M. Kircher, M.T. Gansauge, H. Li, F. Racimo, S. Mallick, J.G. Schraiber, F. Jay, *et al*.** 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science *338*:222-226.

6. **Paabo, S., H. Poinar, D. Serre, V. Jaenicke-Despres, J. Hebler, N. Rohland, M. Kuch, J. Krause, *et al*.** 2004. Genetic analyses from ancient DNA. Annual review of genetics *38*:645-679.

7. **Stoneking, M. and J. Krause.** 2011. Learning about human population history from ancient and modern genomes. Nature reviews. Genetics *12*:603-614.

8. **Lindahl, T.** 1993. Instability and Decay of the Primary Structure of DNA. Nature *362*:709-715.

9. **Olafsson, P.G., A.M. Bryan, G.C. Davis, and S. Anderson.** 1974. A comparative study of the relative ease of thermolytic depurination vs. depyrimidination in 2'-deoxynucleosides. Canadian journal of biochemistry *52*:997-1002.

10. **Ames, B.N.** 1989. Endogenous DNA damage as related to cancer and aging. Mutation research *214*:41-46.

11. **Frosina, G., P. Fortini, O. Rossi, F. Carrozzino, G. Raspaglio, L.S. Cox, D.P. Lane, A. Abbondandolo, and E. Dogliotti.** 1996. Two pathways for base excision repair in mammalian cells. The Journal of biological chemistry *271*:9573-9578.

12. **NewEnglandBiolabs.** 2014. PreCR Repair Mix.

13. **Diegoli, T.M., M. Farr, C. Cromartie, M.D. Coble, and T.W. Bille.** 2012. An optimized protocol for forensic application of the PreCR Repair Mix to multiplex STR amplification of UV-damaged DNA. Forensic science international. Genetics *6*:498-503.

14. **Robertson, J.M., S.M. Dineen, K.A. Scott, J. Lucyshyn, M. Saeed, D.L. Murphy, A.J. Schweighardt, and K.A. Meiklejohn.** 2014. Assessing PreCR repair enzymes for restoration of STR profiles from artificially degraded DNA for human identification. Forensic science international. Genetics *12*:168-180.

15. **Ambers, A., M. Turnbough, R. Benjamin, J. King, and B. Budowle.** 2014. Assessment of the role of DNA repair in damaged forensic samples. International journal of legal medicine.

16. **Nelson, J.** 2009. Repair of Damaged DNA for Forensic Analysis*In* U.S.D.o. Justice (Ed.).

17. **Xu, Z., F. Zhang, B. Xu, J. Tan, S. Li, and L. Jin.** 2009. Improving the sensitivity of negative controls in ancient DNA extractions. Electrophoresis *30*:1282-1285.

18. **Schwarz, C., R. Debruyne, M. Kuch, E. McNally, H. Schwarcz, A.D. Aubrey, J. Bada, and H. Poinar.** 2009. New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. Nucleic acids research *37*:3215-3229.

19. **Poinar, H.N., C. Schwarz, J. Qi, B. Shapiro, R.D. Macphee, B. Buigues, A. Tikhonov, D.H. Huson, *et al*.** 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science *311*:392-394.

20. **Meyer, M. and M. Kircher.** 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor protocols *2010*:pdb prot5448.

21. **Kircher, M., S. Sawyer, and M. Meyer.** 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic acids research *40*:e3.

22. **Enk, J., J.M. Rouillard, and H. Poinar.** 2013. Quantitative PCR as a predictor of aligned ancient DNA read counts following targeted enrichment. BioTechniques *55*:300-309.

23. **Martin, M.** 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*:10-12.

24. **Magoc, T. and S.L. Salzberg.** 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*:2957-2963.

25. **Li, H. and R. Durbin.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*:1754-1760.

26. **Enk, J., A. Devault, R. Debruyne, C.E. King, T. Treangen, D. O'Rourke, S.L. Salzberg, D. Fisher, *et al*.** 2011. Complete Columbian mammoth mitogenome suggests interbreeding with woolly mammoths. Genome biology *12*:R51.

27. **Ginolhac, A., M. Rasmussen, M.T. Gilbert, E. Willerslev, and L. Orlando.** 2011. mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics *27*:2153-2155.

### Supplementary Online Materials

**Materials & Methods**

*Sample Preparation*

This study used ancient DNA from two extinct species as the substrate for the repair enzyme assays. Four previously well-characterized samples were obtained from mammoth (*Mammuthus primigenius*) bones and tusks preserved under varying conditions from the collection of the McMaster Ancient DNA Centre. One mylodon (*Mylodon darwinii*) sample was used as a carrier blank, along with one extraction blank. 100-200 mg of bone or tusk material was broken down into ~ 1 mm2 pieces before undergoing bone demineralization and proteinase K digestion, followed by organic DNA extraction, based on protocols in Poinar *et al* ([1](#_ENREF_1)) and Schwarz *et al* ([2](#_ENREF_2)). Bone and tusk fragments were demineralized twice with an overnight incubation (20-24 h) on a shaker at 900 RPM and at room temperature with 0.5M EDTA pH 8.0, after which the supernatant was removed. The bone pellets were digested with a buffer consisting of 10 mM Tris-Cl pH 8.0, 0.5% Sarcosyl, 250 µg/ml Proteinase K, 5 mM CaCl2, 50 mM DTT, 1% PVP and 2.5 mM PTB and were incubated at 55°C for 5 h at 700 RPM, to ensure homogenization. DNA from both demineralization and digestion supernatants were pooled and extracted using classic phenol-chloroform-isoamyl alcohol extraction methods. All aqueous phases were concentrated over 30 kDa Amicon®Ultra 0.5 ml columns (Millipore, Billerica, MA). Concentrates were purified following extraction using a modified MinElute PCR Purification protocol (Qiagen, Venlo, Netherlands), where samples are mixed to PB buffer in a 1:6 ratio, centrifuged at 3,300 x *g* for 1 min, followed by two 700 µl PE buffer washes at 3,300 x *g*. Normal protocol resumes when PE buffer is discarded and samples are centrifuged 1 min at maximum speed. Samples were eluted in 10 µl of EB buffer after standing for 5 min. All extracts from the same sample were combined and brought up to 175 µl in EB buffer.

*Quantification*

Total DNA was measured using QuantiFluor® dsDNA system (Promega, Madison, WI), in the following reaction: 25 µl of 1X of QuantiFluor® dye, 2.5 µl of DNA extract and 22.5 µl of EB buffer (Qiagen). Samples were incubated for 5 min at room temperature. A standard curve of 1.0, 0.5, 0.1 and 0.05 ng/µl was created from a stock concentration of 100 ng/µl of Lambda DNA (Promega). 50 µl of the sample-dye mixture were pipetted into appropriate cuvettes and fluorescence measurements were taken on a TBS-380 Mini-fluorometer, according to manufacturer's directions (Turner BioSystems, Sunnyvale, CA). Measurements were taken in duplicate and the standard was measured between each round, to account for variations in fluorescence over time.

*PreCR® Repair*

Based on the total DNA quantification values, each extract was diluted to a final concentration of 50 ng in 20 µl (or 2.5 ng/µl) with 1X ThermoPol buffer (New England Biolabs, Ipswitch, MA). Master mix for each repair reaction was prepared to a final concentration of: 1X of ThermoPol buffer (NEB), 1 mg/ml of BSA, 100 µM of dNTP mix , 1X of NAD+ , 1 µl of Repair Mix (kept on ice) and molecular grade H2O (double-distilled, de-ionized, autoclaved and free of nucleases), in a total volume of 30 µl ([3](#_ENREF_3)). Heat-inactivated reaction mixes were incubated at 95°C for 20 min while the active reaction mixes were kept on ice. Heated enzymes master mixes were slowly cooled to 20°C, to prevent denaturation of DNA in the subsequent step. 20 µl of DNA extract prepared earlier was added to each tube. Reactions were incubated at 37°C for 20 min. To inactivate the reaction and eliminate the need to keep the reaction mix at 4°C, repaired extracts were purified over a MinElute column, following the manufacturer's instructions.

*Nelson Repair*

Based on the total DNA quantification values, each extract was diluted to a final concentration of 50 ng in 20 µl (or 2.5 ng/µl) with 1X repair buffer. Master mix for each repair reaction was prepared to a final concentration of: 1X of repair buffer, 0.1 mg/ml of BSA, 200 µM of dNTP mix, 1 mM of ATP, 0.33 U/µl of *E. coli* DNA Polymerase I, 13.3 U/µl of T4 DNA Ligase, 3.3 mU/µl of T4 PDG, 3.3 mU/µl of Endonuclease IV, 3.3 mU/µl of Endonuclease VIII, 0.3 mU/µl of FPG (all enzymes from NEB) and molecular grade H2O, in a total volume of 30 µl ([4](#_ENREF_4)). Heat-inactivated reaction mixes were incubated at 85°C for 20 min while the active reaction mixes were kept on ice. Heated enzymes master mixes were slowly cooled to 20°C, to prevent denaturation of DNA in the subsequent step. 20 µl of DNA extract prepared earlier was added to each tube. Reactions were incubated at 37°C for 20min. To inactivate the reaction and eliminate the need to keep the reaction mix at 4°C, repaired extracts were purified over a MinElute column, following the manufacturer's instructions.

*Blunt-end Library Preparation and Indexing Amplification (idAmp)*

The protocol followed is a modified version of Meyer & Kircher ([5](#_ENREF_5)), with the modifications described in Kircher *et al* ([6](#_ENREF_6)) for indexing. All samples were double-indexed, to allow for pooling during sequencing. The reactions were prepared as follows to a final concentration of: 1X of KAPA SYBR® FAST Bio-Rad iCycler 2X qPCR master mix (KAPA Biosystems, Wilmington, MA), 150 nM of Forward Index, 150 nM of Reverse Index and molecular grade H2O, in a final volume of 20 µl. 14 µl of DNA template was added to each well and samples were incubated as follows: 5 min at 95°C, 10 cycles of 30 s at 95°C and 45 s at 60°C, followed by 3 min at 60°C. idAmp was performed on a CFX-96, with real-time fluorescence detection, according to manufacturer's direction. Each sample was purified using a MinElute column and was eluted in 20 µl of 10 mM Tris-Cl, pH 8.5, with a 5 min RT incubation.

*Pooling*

To facilitate sequencing, all indexed samples were pooled to equal copy number, using an assay targeting the cluster generation sequence. The reactions were prepared as follows: 1X of KAPA SYBR® FAST Bio-Rad iCycler 2X qPCR master mix (KAPA Biosystems), 200 nM of each primer, 4 µl of DNA template (diluted 1:250 with 10 mM Tris-Cl, pH 8.5 with 0.05% Tween 20) and molecular grade H2O, in a final volume of 10 µl. Samples were quantified in triplicate, along with PhiX standard ranging from 100 to 1 pM, with the following profile: 5 min at 95°C, 10 cycles of 30 s at 95°C and 45 s at 60°C, followed by a melt curve from 60°-95°C with 0.5°C increments. Quantification was carried out using qPCR on a CFX-96 instrument, with real-time detection according to manufacturer's directions. Values were size-corrected to adjust for the difference between the mean sample fragment length (obtained from BioAnalyzer traces) and the 500 bp PhiX standard. Each sample was normalized to the number of copies in 1 µl of the lowest, non-blank sample.

*High-Throughput Sequencing and Data Analysis*

All sequencing was performed on two separate lanes of a HiSeq® 1500 Rapid flowcell (Illumina, San Diego, CA), using a 2 x 100 bp HiSeq® Reagent Kit, according to manufacturer's recommendations. Raw data was processed with HiSeq® Control Software v1.5.15.1 and Real-Time Analysis v1.13.48.0 (both from Illumina). File conversion and demultiplexing using each 7 bp reverse index (requiring 100% match) were performed using CASAVA v1.8.2. Reads were curated and aligned using the parameters in Enk *et al (*[*7*](#_ENREF_7)*)*. Sequencing reads were trimmed of 3' universal adapters using cutadapt v1.2 ([8](#_ENREF_8)). Paired-end reads were merged with FLASH v1.0 ([9](#_ENREF_9)). Merged reads were combined with un-merged Read 1 sequences to create the final data set used for alignment. Reads 24 bp and longer were then aligned to both the mitochondrial mammoth (*Mammuthus primigenius*) genome (GenBank Accession No JF912200.1) ([10](#_ENREF_10)) and the hard masked (i.e. no repeats) nuclear African elephant (*Loxodonta africana*) consensus genome (Loxafr3.0 from the Broad Institute, RefSeq Assembly ID: GCF\_000001905.1) using BWA v.0.6.1-r104 ([11](#_ENREF_11)). Random 3 million read subsets were taken from each library to analyze endogenous content and fragment length distribution. Base misincorporations were calculated using mapDamage2.0 ([12](#_ENREF_12)).

Table S1. Types of DNA damage

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Damage | Example | Ancient DNA1 | Caused by | Effect | Characteristics |
| Apurinic or apyrimidinic (AP) sites |  | Expected | Hydrolysis,  glycosylases | DNA polymerase stalling; base misincorporation | Can lead to preferential misincorporation of adenine in complementary strand |
| Damaged purines | 8-oxoguanine | Expected | Oxidation | DNA polymerase stalling; base misincorporation | Causes guanine to thymine misincorporations in coding stand |
| Damaged pyrimidines | 5-hydroxy-5-methylhydantoin | Expected | Oxidation | DNA polymerase stalling |  |
| Base deamination | Cytosine deamination | Observed | Hydrolysis | Base misincorporation | Causes cytosine to thymine misincorporations in coding stand |
| Nicks / fragmentation |  | Observed | Hydrolysis,  nucleases | DNA polymerase stalling | Cannot be repaired once dsDNA is denatured (i.e. fragmentation) |
| ssDNA termini |  | Expected | Hydrolysis,  nucleases | DNA polymerase stalling | Short length (< 10 bp) of ssDNA at either fragment end |
| Dimerization | Cis-syn-cyclobutane pyrimidine dimers | Not observed or expected | UV radiation | DNA polymerase stalling | Rarely seen in ancient DNA |
| 1 shows whether damage type has been observed or is expected in ancient DNA; ssDNA : single-stranded DNA; dsDNA: double-stranded DNA | | | | | |

**Figure S1. Number of adapted molecules quantified after library preparation of extracts treated with inactive and active versions of Nelson or PreCR® repair mix.** Four mammoth extracts (A-D) and one mylodon carrier blank (E) were incubated with active or heat-inactivated DNA enzymatic repair mixtures, either Nelson (in dark- active and light- inactive orange, respectively) or PreCR® (in dark- active and light- inactive blue, respectively). Values are obtained from the mean of two size-corrected technical replicates and the error bars represent one standard deviation. Differences that were deemed statistically significant with a paired, two-tailed t-test (p = 0.05) are marked with a red bracke



**Figure S2. Comparison of fragment length distributions (FLD) of total DNA content of four mammoth samples and one mylodon blank, after undergoing active and heat-inactivated DNA repair with Nelson or PreCR® protocols.** Four mammoth extracts (panels A-D) and one mylodon carrier blank (panel E) were incubated with active or heat-inactivated DNA enzymatic repair mixtures, either PreCR® (in dark and light blue, respectively) or Nelson (in dark and light orange, respectively). For eac FLD, the median of all fragments was calculated (dashed line).

**Figure S2. Comparison of fragment length distributions (FLD) of total DNA content of four mammoth samples and one mylodon blank, after undergoing active and heat-inactivated DNA repair with Nelson or PreCR® protocols.** Four mammoth extracts (panels A-D) and one mylodon carrier blank (panel E) were incubated with active or heat-inactivated DNA enzymatic repair mixtures, either PreCR® (in dark and light blue, respectively) or Nelson (in dark and light orange, respectively). For eac FLD, the median of all fragments was calculated (dashed line).

# 

**Figure S3. Comparison of fragment length distributions (FLD) of the mapped nuclear DNA content of four mammoth samples and one mylodon blank, after undergoing active and heat-inactivated DNA repair with Nelson or PreCR® protocols.** Four mammoth extracts (panels A-D) and one mylodon carrier blank (panel E) were incubated with active or heat-inactivated DNA enzymatic repair mixtures, either PreCR® (in dark and light blue, respectively) or Nelson (in dark and light orange, respectively). Repaired extracts were mapped to the nuclear reference genome. For each FLD, the median of all fragments was calculated (dashed line).

### Supplementary Online Materials References

1. **Poinar, H.N., C. Schwarz, J. Qi, B. Shapiro, R.D. Macphee, B. Buigues, A. Tikhonov, D.H. Huson, *et al*.** 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science *311*:392-394.

2. **Schwarz, C., R. Debruyne, M. Kuch, E. McNally, H. Schwarcz, A.D. Aubrey, J. Bada, and H. Poinar.** 2009. New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. Nucleic acids research *37*:3215-3229.

3. **NewEnglandBiolabs.** 2014. PreCR Repair Mix.

4. **Nelson, J.** 2009. Repair of Damaged DNA for Forensic Analysis*In* U.S.D.o. Justice (Ed.).

5. **Meyer, M. and M. Kircher.** 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor protocols *2010*:pdb prot5448.

6. **Kircher, M., S. Sawyer, and M. Meyer.** 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic acids research *40*:e3.

7. **Enk, J., J.M. Rouillard, and H. Poinar.** 2013. Quantitative PCR as a predictor of aligned ancient DNA read counts following targeted enrichment. BioTechniques *55*:300-309.

8. **Martin, M.** 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*:10-12.

9. **Magoc, T. and S.L. Salzberg.** 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*:2957-2963.

10. **Enk, J., A. Devault, R. Debruyne, C.E. King, T. Treangen, D. O'Rourke, S.L. Salzberg, D. Fisher, *et al*.** 2011. Complete Columbian mammoth mitogenome suggests interbreeding with woolly mammoths. Genome biology *12*:R51.

11. **Li, H. and R. Durbin.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*:1754-1760.

12. **Ginolhac, A., M. Rasmussen, M.T. Gilbert, E. Willerslev, and L. Orlando.** 2011. mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics *27*:2153-2155.

# 

# conclusion

This thesis demonstrates some of the challenges that lie ahead of the forensic science community upon the transition to HTS-based DNA profiling. While both projects made important contributions to our knowledge of the application of ancient DNA methodologies to forensic science, further work is required to confirm hypotheses presented here and to investigate the root cause of these observations.

In Chapter I, this thesis demonstrated that STR amplification of artificially degraded single source extracts and on three-person mixtures converted to blunt-end sequencing libraries causes a significant decrease in profile quality. These observations are due to allelic drop-out, heterozygous peak imbalance and increased stutter ratios. The source of this problem has not yet been identified, but is presumed to occur during library preparation, with its effects observable only after multiple rounds of amplification (either indexing or STR).

In Chapter II, the DNA repair efficiency of two enzymatic kits was evaluated on four mammoths and one mylodon extract. Based on nuclear and mitochondrial mammoth content, variations in median fragment lengths and rate of base misincorporation, PreCR® demonstrates potential for future use; it produced previously unavailable endogenous content upon removal of cytosine deamination, which largely compensates for the overall decrease in fragment length. Conversely, Nelson exhibits no significant evidence of repair and may cause chimerical fragments. However attractive these multi-enzymatic kits might be for casework applications, it is advised that sequential repair should be explored for optimal DNA repair.

In both projects, established ancient DNA methods provided a framework to direct the experiments. The blunt-end library preparation protocol was developed for ancient DNA specimens and is adapted to its associated challenges. For DNA repair, mammoth samples acted as a proxy for forensic extracts: both have similar degradation patterns but there is no risk of modern contamination of mammoth extracts. The analyses for both projects were performed in dedicated clean rooms, in strict adherence to published ancient DNA guidelines [[25](#_ENREF_25)].

Available literature for library preparation shows successful sequencing of STR amplicons [[26](#_ENREF_26), [27](#_ENREF_27)], potential use for microhaplotypes [[28](#_ENREF_28), [29](#_ENREF_29)] and the development of SNP assays [[30](#_ENREF_30)] but the scope of the work still revolves around the traditional STR loci found in DNA databases. Similarly, publications evaluating enzymatic repair kits were performed with STR amplification on artificially damaged samples [[31-33](#_ENREF_31)]. This thesis work provides additional qualitative data on how these technologies would perform on actual casework samples (as demonstrated with the mammoth extracts) and complements the available literature.

While the results presented herein are preliminary work, their value resides in the fact that this research has never been performed and published before. Thus, the conclusions of this thesis lay the groundwork on which to build stronger and better-adapted protocols for forensic, ancient or challenging DNA extracts.

# REFERENCES

[1] M. Twain, Roy J. Friedman Mark Twain Collection (Library of Congress). Pudd'nhead Wilson : a tale. London: Chatto & Windus; 1894.

[2] S.A. Cole. Suspect identities : a history of fingerprinting and criminal identification. Cambridge, MA: Harvard University Press; 2001.

[3] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, H. Erlich, Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction, Cold Spring Harb Symp Quant Biol 51 Pt 1 (1986) 263-273

[4] A.J. Jeffreys, V. Wilson, S.L. Thein, Individual-specific 'fingerprints' of human DNA, Nature 316 (1985) 76-79

[5] SWGDAM. SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories (2010)

[6] J.M. Butler. Advanced Topics in Forensic DNA Typing: Interpretation: Academic Press; 2014.

[7] M. Hashiyada. DNA biometrics, Biometrics: InTech; 2011.

[8] M.G. Ensenberger, J. Thompson, B. Hill, K. Homick, V. Kearney, K.A. Mayntz-Press, et al., Developmental validation of the PowerPlex 16 HS System: an improved 16-locus fluorescent STR multiplex, Forensic Sci Int Genet 4 (2010) 257-264

[9] D.Y. Wang, C.W. Chang, R.E. Lagace, N.J. Oldroyd, L.K. Hennessy, Development and validation of the AmpFlSTR(R) Identifiler(R) Direct PCR Amplification Kit: a multiplex assay for the direct amplification of single-source samples, J Forensic Sci 56 (2011) 835-845

[10] E.C. Friedberg, G.C. Walker, W. Siede. DNA repair and mutagenesis. Washington, D.C.: ASM Press; 1995.

[11] C. Capelli, F. Tschentscher, V.L. Pascali, "Ancient" protocols for the crime scene? Similarities and differences between forensic genetics and ancient DNA analysis, Forensic Sci Int 131 (2003) 59-64

[12] S. Paabo, H. Poinar, D. Serre, V. Jaenicke-Despres, J. Hebler, N. Rohland, et al., Genetic analyses from ancient DNA, Annu Rev Genet 38 (2004) 645-679

[13] T. Helleday, S. Eshtad, S. Nik-Zainal, Mechanisms underlying mutational signatures in human cancers, Nat Rev Genet 15 (2014) 585-598

[14] H.N. Poinar, C. Schwarz, J. Qi, B. Shapiro, R.D. Macphee, B. Buigues, et al., Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA, Science 311 (2006) 392-394

[15] R.E. Green, J. Krause, S.E. Ptak, A.W. Briggs, M.T. Ronan, J.F. Simons, et al., Analysis of one million base pairs of Neanderthal DNA, Nature 444 (2006) 330-336

[16] A.W. Briggs, J.M. Good, R.E. Green, J. Krause, T. Maricic, U. Stenzel, et al., Targeted retrieval and analysis of five Neandertal mtDNA genomes, Science 325 (2009) 318-321

[17] H.A. Burbano, E. Hodges, R.E. Green, A.W. Briggs, J. Krause, M. Meyer, et al., Targeted investigation of the Neandertal genome by array-based sequence capture, Science 328 (2010) 723-725

[18] R.E. Green, J. Krause, A.W. Briggs, T. Maricic, U. Stenzel, M. Kircher, et al., A draft sequence of the Neandertal genome, Science 328 (2010) 710-722

[19] M. Meyer, M. Kircher, M.T. Gansauge, H. Li, F. Racimo, S. Mallick, et al., A high-coverage genome sequence from an archaic Denisovan individual, Science 338 (2012) 222-226

[20] D.M. Wagner, J. Klunk, M. Harbeck, A. Devault, N. Waglechner, J.W. Sahl, et al., Yersinia pestis and the plague of Justinian 541-543 AD: a genomic analysis, Lancet Infect Dis 14 (2014) 319-326

[21] K.I. Bos, V.J. Schuenemann, G.B. Golding, H.A. Burbano, N. Waglechner, B.K. Coombes, et al., A draft genome of Yersinia pestis from victims of the Black Death, Nature 478 (2011) 506-510

[22] A.M. Devault, G.B. Golding, N. Waglechner, J.M. Enk, M. Kuch, J.H. Tien, et al., Second-pandemic strain of Vibrio cholerae from the Philadelphia cholera outbreak of 1849, N Engl J Med 370 (2014) 334-340

[23] J. Enk, A. Devault, R. Debruyne, C.E. King, T. Treangen, D. O'Rourke, et al., Complete Columbian mammoth mitogenome suggests interbreeding with woolly mammoths, Genome Biol 12 (2011) R51

[24] M. Stoneking, J. Krause, Learning about human population history from ancient and modern genomes, Nat Rev Genet 12 (2011) 603-614

[25] A. Cooper, H.N. Poinar, Ancient DNA: do it right or not at all, Science 289 (2000) 1139

[26] S.L. Fordyce, M.C. Avila-Arcos, E. Rockenbauer, C. Borsting, R. Frank-Hansen, F.T. Petersen, et al., High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, Biotechniques 51 (2011) 127-133

[27] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, D. Deforce, Forensic STR analysis using massive parallel sequencing, Forensic Science International-Genetics 6 (2012) 810-818

[28] K.K. Kidd, J.R. Kidd, W.C. Speed, R. Fang, M.R. Furtado, F.C. Hyland, et al., Expanding data and resources for forensic use of SNPs in individual identification, Forensic Sci Int Genet 6 (2012) 646-652

[29] K.K. Kidd, A.J. Pakstis, W.C. Speed, R. Lagace, J. Chang, S. Wootton, et al., Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics, Forensic Sci Int Genet 12 (2014) 215-224

[30] A. Freire-Aradas, M. Fondevila, A.K. Kriegel, C. Phillips, P. Gill, L. Prieto, et al., A new SNP assay for identification of highly degraded human DNA, Forensic Sci Int Genet 6 (2012) 341-349

[31] T.M. Diegoli, M. Farr, C. Cromartie, M.D. Coble, T.W. Bille, An optimized protocol for forensic application of the PreCR Repair Mix to multiplex STR amplification of UV-damaged DNA, Forensic Sci Int Genet 6 (2012) 498-503

[32] J.M. Robertson, S.M. Dineen, K.A. Scott, J. Lucyshyn, M. Saeed, D.L. Murphy, et al., Assessing PreCR repair enzymes for restoration of STR profiles from artificially degraded DNA for human identification, Forensic Sci Int Genet 12 (2014) 168-180

[33] A. Ambers, M. Turnbough, R. Benjamin, J. King, B. Budowle, Assessment of the role of DNA repair in damaged forensic samples, Int J Legal Med (2014)