UNIVERSITY OF CALIFORNIA
SANTA CRUZ

DEVELOPING AND APPLYING MOLECULAR METHODS FOR DEGRADED DNA

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ECOLOGY AND EVOLUTIONARY BIOLOGY

by

Joshua D. Kapp

September 2022

The Dissertation of Joshua D. Kapp is approved:

Professor Beth Shapiro, Chair

Professor Richard E. Green

Professor Giacomo Bernardi

Assistant Professor Christopher Vollmers

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Developing and Applying Molecular Methods for Degraded DNA

by

Joshua D. Kapp

Researchers working with poor-quality samples have developed a suite of methods to maximize data generation from minute quantities of short and damaged DNA. Library preparation, the process of preparing extracted DNA for sequencing, is a particularly important step when working with degraded samples. Library preparation approaches optimized for degraded DNA produce more informative libraries but tend to be more laborious, costly, and have lower throughput compared to conventional approaches. In this dissertation, I present a rapid and cost-effective single-stranded DNA library preparation protocol to prepare degraded DNA for Illumina sequencing and apply the approach to several degraded sample types. In the first chapter, I present the first iteration of the library preparation method and demonstrate the effectiveness on cell-free DNA and synthetic oligonucleotides. In the second chapter, I optimize the library preparation method for highly degraded ancient samples and compare the efficacy to two commonly used degraded DNA optimized approaches. In the last chapter, I develop a workflow for whole genome sequencing of single hair shafts and characterize the variation of DNA recovered from the hairs of 50 anonymous volunteers.

Acknowledgements

I would like to sincerely thank Beth Shapiro and Richard E. Green for their invaluable mentorship and the endless opportunities they have provided me. They gave me the freedom to choose the topics I wanted to focus on in graduate school and taught me how to transform ideas into feasible projects. They were always willing to help solve a problem and kept me from getting too distracted, so I would finish projects. Knowing I had their support significantly increased my academic confidence over the years, which has helped me believe in my own work. I have been immensely lucky to such incredible PI's and I hope to continue working with them.

I was fortunate to be mentored by two passionate and encouraging postdocs early in my academic career, Pete Heintzman and Kelly Harkins Kincaid. Pete was my first mentor and provided me with the foundational wet-lab skills I rely upon every day. Additionally, Pete always made sure I understood the concepts behind the methods. Kelly is the reason I pursued method development. I worked on my first methods project with Kelly, where I began to learn how to translate ideas on a white board to tubes. Pete and Kelly were a consistent source of knowledge and support, which allowed me to develop as a researcher.

I learned from many PI's, graduate students, and postdocs in the lab, who were always willing to provide kind advice. These include Rachel Meyer, Lars Fehren-Schmitz, Nathan Schaefer, Nedda Saremi, James Cahill, Andre Soares, Brendan O'Connell, Jonas Oppenheimer, Sarah Crump, Merly Escalona, Jannine Forst, and

Megan Supple. A special thank you to Alisa Vershinina, who I went to for academic advice and every graduate student admin related question.

I have had the opportunity to work closely with many talented technicians and undergraduate researchers. I appreciate the time and energy the technicians have spent assisting my projects by keeping the lab together, running the machines, and providing protocol feedback. Additionally, I found mentoring undergraduates one of the most rewarding aspects of graduate school. I thank them for their trust, time, and effort. Thanks to you all: Lourdes Gomez, Nicholas Maurer, Molly Cassatt-Johnstone, Shelby Dunn, Sam Sacco, Talia Tzadikario, Ciara Wanket, Will Seligmann, Eric Beraut, Sam Cutler, Cameron Milne, Hayley Neadeau, and Manuel Varela.

I am grateful to all members of the UCSC Paleogenomics lab, past and present. I have been privileged to work with so many talented and kind researchers working on a wide array of topics. I have worked with some closer than others but have learned from everyone.

Additionally, I would like to thank everyone at Claret Bioscience. I have greatly enjoyed working with such skilled, methods focused researchers.

Lastly, I would like to thank my family who supported me throughout the extended academic journey. My parents always encouraged me to pursue a career I enjoyed and were supportive from community college through graduate school. My partner, Annie Roth, provided me with immense emotional support and happiness. There were many difficult days, but it was always a joy to go home.

The text of this dissertation includes reprints of the following previously published material: (Chapter 1) Troll CJ, Kapp J, Rao V, Harkins KM, Cole C, Naughton C, Morgan JM, Shapiro B, Green RE. 2019. A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. BMC Genomics. 20:1023; (Chapter 2) Kapp JD, Green RE, Shapiro B. A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. Journal of Heredity. 2021;112:241–9. Beth Shapiro and Richard E. Green directed and supervised the research which forms the basis for the dissertation.

Introduction

DNA has been found to persist for up to a million years under ideal conditions [1] but DNA molecules will accumulate damage over time, and potentially degrade rapidly, via several mechanisms. DNA fragmentation is the most common form of damage shared among samples thousands of years old and those still associated with a living organism. Researchers working with ancient DNA have observed degraded molecules are often flanked by purines on the reference strand [2–4], which suggests samples have accumulated single-strand breaks by hydrolytic depurination [5]. Endonucleases, such as DNase1, are responsible for the initial degradation of DNA still associated with an organism, such as cell-free DNA (cfDNA) from blood plasma [6] and DNA recovered from hair shafts [7]. Additionally, DNA bases can become chemically altered via cytosine deamination, which leads to the presence of uracils among DNA strands [5, 8]. Along with aged and natively degraded samples, the DNA contained in any sample can quickly degrade due to storage conditions and chemical treatments [9]. Working with degraded samples requires the efficient processing of small and damaged DNA molecules.

Molecular methods optimized for degraded samples have been developed for efficient recovery and sequence preparation of degraded DNA. DNA is first extracted from a sample and then sequence platform-specific adapters are added to the ends of the recovered DNA, which is known as library preparation. Preparing DNA for Illumina sequencing requires the ligation of sequence-specific adapters to both ends of the template DNA, where the 5’ end receives a ‘P5’ adapter and the 3’ end receives a

'P7' adapter. Generally, conventional methods assume samples contain large quantities of intact DNA and may purposefully exclude or unintentionally lose short and damaged molecules [10]. However, over the last several decades, methods have been developed to efficiently recover and process short and damaged DNA. Extraction and isolation methods have been developed to recover small quantities of fragmented DNA from several sample types [11, 12]. Degraded DNA optimized library preparation methods efficiently convert and retain small and damaged molecules [13, 14, 10, 15]. Finally, several approaches and optimizations have been developed to reduce contamination [16, 17] and enrich targets of interest [18, 19].

Library preparation is a particularly critical step when processing degraded samples. Typically, library preparation workflows convert just double-stranded DNA (dsDNA) to library molecules [13, 15] but the field of ancient DNA pioneered the development and use of library preparation methods that convert single-stranded DNA (ssDNA) [10, 14]. Compared to dsDNA library preparation methods, ssDNA methods have been found to more efficiently convert degraded DNA to library molecules while incorporating smaller molecules and a higher proportion of target molecules [9, 10, 20, 21]. Unlike dsDNA methods, some ssDNA library preparation methods do not manipulate the DNA termini, which preserves the native end information of the molecules. This feature is useful for the analysis of fragment end profiles to explore DNA damage patterns [4] and potential biomarkers for disease [22, 23]. Despite the performance benefits of ssDNA library preparation methods, their use is typically

reserved for the worst preserved samples due to their higher cost and laborious workflows.

In my dissertation, I present a rapid and affordable single-stranded DNA library preparation method (SRM here, alternative names indicated elsewhere) to prepare libraries for Illumina sequencing from cell-free DNA (Chapter 1), ancient DNA (Chapter 2), and DNA recovered from hair shafts (Chapter 3). The SRM simultaneously ligates Illumina's P5 and P7 adapter using splint ligation in the presence of single-stranded DNA binding proteins. The SRM uses widely available and affordable enzymes to modify the template DNA, T4 PNK drives the phosphorylation state of DNA ends, and T4 DNA ligase drives ligation. The SRM begins with a heat denaturation step, which allows for library conversion of double-stranded DNA, single-stranded DNA, and a combination of the two. Finally, the SRM does not manipulate the termini of the DNA, which preserves the native ends of the molecules in the DNA extraction. The SRM is an efficient, affordable, and fast library preparation method compared to commercial and field standard methods.

In chapter 1, I presented the first iteration of the single-stranded library preparation approach (SRSLY) described above, which was optimized for cell-free DNA (cfDNA) and oligonucleotides. Libraries were prepared using cfDNA extractions from healthy anonymous donors using SRSLY and several commercially available double-stranded and single-stranded DNA library preparation kits. SRSLY prepares similar or better-quality libraries compared to the commercially available kits. Additionally, SRSLY does not manipulate the native ends of DNA, which allows for better characterization

of fragmentation profiles from cfDNA and oligonucleotides. This project was a collaboration with Christopher J. Troll, Varsha Rao, Kelly M. Harkins, Charles Cole, Colin Naughton, Jessica M. Morgan, Beth Shapiro, and Richard E. Green. The text was originally published in BMC Genomics with the title 'A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos', where Christopher J. Troll, Varsha Rao, and I were co-first authors. I conceived, designed, and developed the method into a single-reaction library preparation, where simultaneous adapter ligation occurs in the presence of single-stranded DNA binding proteins, T4 PNK, and T4 DNA ligase.

In chapter 2, I presented an ancient DNA (aDNA) optimized version of the single-stranded library preparation approach (SCR), for a range of aDNA inputs. I compared the library preparation performance of the SCR to two field standard methods, one double-stranded DNA and one single-stranded DNA library preparation approach. The SCR prepares similar quality libraries compared to the field standard single-stranded approach and consistently higher quality libraries compared to the double-stranded approach. The SCR is a faster, less costly, and higher throughput library preparation approach compared to existing methods without sacrificing performance. This project was a collaboration with Richard E. Green and Beth Shapiro. The text was originally published in the Journal of Heredity with the title 'A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA'. I designed and performed experiments, prepared the DNA extractions, prepared the libraries, performed the analysis, and wrote the first draft of the manuscript. Beth Shapiro and

Richard E. Green provided input throughout the study, assisted in data analysis, and revised the manuscript.

In chapter 3, I presented a workflow to process hair shafts for whole genome sequencing and characterized the DNA recovered from single hair shafts of 50 anonymous volunteers. Hair shafts are a common evidence type at crime scenes but yield small quantities of degraded DNA. Due to poor DNA preservation, single hair shafts rarely generate full profiles using field standard methods. To assess hair shafts as a capable sample type to generate whole genomes, I optimized the DNA lysis of single-hair shafts and library preparation of picogram scale quantities of degraded DNA. Using this workflow, I collected, processed, and generated deep sequencing data from head and pubic hairs donated by 50 anonymous volunteers. I found hairs to be an appropriate sample type to generate multi-fold genome coverage, where 96% of hairs are likely to generate at least 2X genome coverage. This project was a collaboration with Hayley Neadeau, Ciara Wanket, and Richard E. Green. I designed and performed workflow optimization experiments, performed the DNA extractions, performed the library preparations, performed data analysis, and wrote the manuscript. Hayley Neadeau and Ciara Wanket assisted in the wet lab. Richard E. Green provided input throughout the study and performed data analysis. The data presented in this chapter will be included in a larger manuscript, which will also investigate the genotyping accuracy of the low coverage hair data compared to high coverage data generated from saliva.

References

1. van der Valk T, Pečnerová P, Díez-del-Molino D, Bergström A, Oppenheimer J, Hartmann S, et al. Million-year-old DNA sheds light on the genomic history of mammoths. Nature. 2021;591:265–9.

2. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, et al. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci USA. 2007;104:14616–21.

3. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. PLoS ONE. 2012;7:e34131.

4. Bokelmann L, Glocke I, Meyer M. Reconstructing double-stranded DNA fragments on a single-molecule level reveals patterns of degradation in ancient samples. Genome Res. 2020;30:1449–57.

5. Dabney J, Meyer M, Paabo S. Ancient DNA Damage. Cold Spring Harbor Perspectives in Biology. 2013;5:a012567–a012567.

6. Watanabe T, Takada S, Mizuta R. Cell-free DNA in blood circulation is generated by DNase1L3 and caspase-activated DNase. Biochemical and Biophysical Research Communications. 2019;516:790–5.

7. Fischer H, Szabo S, Scherz J, Jaeger K, Rossiter H, Buchberger M, et al. Essential Role of the Keratinocyte-Specific Endonuclease DNase1L2 in the Removal of Nuclear DNA from Hair and Nails. Journal of Investigative Dermatology. 2011;131:1208–15.

8. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29:1682–4.

9. Stiller M, Sucker A, Griewank K, Aust D, Baretton GB, Schadendorf D, et al. Single-strand DNA library preparation improves sequencing of formalin-fixed and paraffin-embedded (FFPE) cancer DNA. Oncotarget. 2016;7:59115–28.

10. Gansauge M-T, Gerber T, Glocke I, Korlević P, Lippik L, Nagel S, et al. Single-stranded DNA library preparation from highly degraded DNA using *T4* DNA ligase. Nucleic Acids Res. 2017;:gkx033.

11. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci USA. 2013;110:15758–63.

12. Rohland N, Glocke I, Aximu-Petri A, Meyer M. Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. Nat Protoc. 2018;13:2447–61.

13. Meyer M, Kircher M. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb Protoc. 2010;2010:pdb.prot5448.

14. Gansauge M-T, Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc. 2013;8:737–48.

15. Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding MHS, Samaniego JA, et al. Single-tube library preparation for degraded DNA. Methods Ecol Evol. 2018;9:410–9.

16. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010;328:710–22.

17. Boessenkool S, Hanghøj K, Nistelberger HM, Der Sarkissian C, Gondek AT, Orlando L, et al. Combining bleach and mild predigestion improves ancient DNA recovery from bones. Mol Ecol Resour. 2017;17:742–51.

18. Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, et al. Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. The American Journal of Human Genetics. 2013;93:852–64.

19. Zavala EI, Aximu-Petri A, Richter J, Nickel B, Vernot B, Meyer M. Quantifying and reducing cross-contamination in single- and multiplex hybridization capture of ancient DNA. Molecular Ecology Resources. 2022;22:2196–207.

20. Bennett EA, Massilani D, Lizzo G, Daligault J, Geigl E-M, Grange T. Library construction for ancient genomics: Single strand or double strand? BioTechniques. 2014;56:289–300.

21. Wales N, Carøe C, Sandoval-Velasco M, Gamba C, Barnett R, Samaniego JA, et al. New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA. BioTechniques. 2015;59:368–71.

22. Ding SC, Lo YMD. Cell-Free DNA Fragmentomics in Liquid Biopsy. Diagnostics. 2022;12:978.

23. Liu Y. At the dawn: cell-free DNA fragmentomics and gene regulation. Br J Cancer. 2022;126:379–90.

Chapter 1: A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos

Note: This project was a collaboration with Christopher J. Troll, Varsha Rao, Kelly M. Harkins, Charles Cole, Colin Naughton, Jessica M. Morgan, Beth Shapiro, and Richard E. Green. The text was originally published in BMC Genomics with the title 'A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos', where Christopher J. Troll, Varsha Rao, and I were co-first authors. I conceived, designed, and developed the method into a single-reaction library preparation, where simultaneous adapter ligation occurs in the presence of single-stranded DNA binding proteins, T4 PNK, and T4 DNA ligase. I have re-formatted the original manuscript.

1.1 Background

For high-throughput sequencing, DNA molecules must be converted into sequencing libraries, which requires ligation of sequencer-specific adapters [1]. Conventional methods for Next-Generation Sequencing (NGS) library preparation convert only double-stranded DNA (dsDNA) into library-ready molecules. Prior to adapter ligation, conventional dsDNA protocols perform end-polishing, which blunts the termini of each template molecule by using DNA polymerases to fill in 5-prime overhangs and digest 3-prime overhangs. In most cases, an additional polymerase will A-tail the 3-prime ends of template DNA to promote efficient ligation of the sequencer-specific adapters [2, 3]. While end-polishing is a prerequisite for efficient dsDNA NGS adapter ligation, it renders all molecules uniformly blunt, obscuring the native termini of molecules and changing their true lengths. Furthermore, conventional dsDNA

methods are unable to convert single-stranded DNA (ssDNA) or dsDNA nicked on both strands into sequencer compatible molecules. A variation of conventional dsDNA NGS library preparation uses Tn5 transposase to both cleave the DNA template and deliver adapters [4]. While not dependent on end-polishing or adapter ligation per se, transposase-based methods also fail to capture the native termini of molecules or convert ssDNA and nicked dsDNA into library molecules.

Single-stranded DNA library preparation methods offer several advantages over traditional dsDNA methods [5,6,7]. By denaturing the duplexed template DNA prior to adapter ligation and maintaining the DNA as single strands through at least an initial adapter ligation, single-stranded preparation methods are theoretically able to convert all of the molecules captured by traditional dsDNA library preparation methods as well as nicked dsDNA and ssDNA molecules. Originally developed for the genomic analysis of highly degraded ancient DNA [7, 8], ssDNA library preparation methods have been adopted for other fragmented sample types such as such as cell-free DNA (cfDNA) and DNA purified from Formalin Fixed Paraffin Embedded (FFPE) sections, due to their efficiency in converting a high fraction of input DNA fragments into sequencing library molecules and their ability to capture small DNA fragments. Further, the sequencing reads from some ssDNA library methods represent the natural 5-prime and 3-prime ends of the input DNA fragments. Thus, when mapped to a reference genome, these data reveal the exact genomic location of the input fragments; an important feature for cfDNA researchers studying biological fragmentation patterns.

Cell-free DNA found circulating in blood plasma and other bodily fluids contains a wealth of biomedical information that can be assayed by NGS with a minimally-invasive blood draw. A number of studies and commercial offerings use NGS data obtained from blood plasma-derived cfDNA to monitor prenatal health, organ transplant reception, cancer detection and progression, and other diseases [9,10,11,12,13,14]. In healthy individuals the vast majority of cfDNA recovered from blood is thought to originate from apoptotic lymphoid and myeloid cells, with a limited number of fragments deriving from other tissues [12, 15, 16]. However, during pregnancy or disease progression, studies have shown that blood plasma may also contain DNA fragments derived from e.g. fetal or tumor cells undergoing apoptosis, necrosis, or other forms of cell death [12, 17,18,19,20,21].

The length distribution of DNA extracted from blood plasma is centered around 167 base-pairs (bp). Thus, cfDNA fragments are thought to be mono-nucleosomal, the result of chromatosome (histone octamer core, also known as the nucleosome core particle, and an associated linker histone) imparted protection from nuclease degradation [12, 15, 16, 22,23,24]. In addition to DNA fragments centered around 167 bp, cfDNA also contains shorter DNA fragments (< 100 bp) that may not derive from nucleosome-bound DNA. Recent studies examining cfDNA within this smaller, sub-nucleosome size range show that these fragments may be the result of nuclease protection by other DNA binding proteins, such as transcription factors. Other components of cfDNA can include mitochondrial DNA and microbial DNA [16, 22, 25].

Several single-stranded library preparation methods have been described since 2013 [8, 22, 26,27,28,29,30]. However, widespread adoption by the NGS community has been hindered by the fact that they are more time consuming and require more enzymatic steps than traditional dsDNA methods. In addition, some ssDNA methods require exotic or expensive reagents [22, 26] and many necessitate the use of primer extension to create a second strand to facilitate sequence adapter ligation [8, 26, 28,29,30]. Also, in some cases special bioinformatic processing of the data is required to deal with artifacts introduced as a consequence of library prep [22, 28].

Here we describe a fast, simple, and efficient ligation-based single-stranded DNA library preparation method engineered to produce complex NGS libraries from as little as one nanogram (ng) of input DNA without altering the native ends of template molecules. Our method, called Single-Reaction Single-stranded LibrarY or SRSLY, requires no exotic reagents and can be completed in 2.5 h. SRSLY works by ligating uniquely designed NGS adapters in a single combined phosphorylation/ligation reaction without requiring end-polishing. Both SRSLY adapters are modified from the splint-adapter design introduced by Gansauge et al [26]. The approach of splint-ligation of both adapters was introduced by the SPLAT method, developed for bisulphite sequencing [26, 27]. SRSLY builds on these features with a streamlined workflow, a robust adapter design, and an optimized single-step ligation scheme that efficiently delivers both adapters.

We present standard sequencing metrics produced by SRSLY libraries made with cfDNA from healthy human donors and compare our results to those of

commercially available library preparation methods. We then highlight the benefits of ssDNA libraries generated using SRSLY compared to dsDNA preps using synthetic duplexed oligonucleotides. Next, we demonstrate the ability of SRSLY to capture short length ssDNA fragments, and the ability to assay oligonucleotide purity using single-stranded synthesized oligos of varying length and known sequence. Finally, we demonstrate how SRSLY libraries empower fragmentomic analyses of cfDNA data by capturing a wide range of DNA fragment lengths without altering their native 5-prime and 3-prime termini. Given its efficiency and ease of use, SRSLY is a drop-in replacement for both ssDNA and dsDNA library preparation methods for many applications.

1.2 Results

1.2.1 Library construction

The SRSLY method creates Illumina sequencing libraries from fragmented or degraded template (input) DNA (Figure. 1.1). Template DNA, which can be a complex mixture of dsDNA, ssDNA, and nicked dsDNA, is first heat denatured and then immediately cold shocked in order to render all template DNA molecules uniformly single-stranded. The DNA is maintained as single-stranded throughout the ligation reaction by the inclusion of a thermostable single-stranded binding protein (SSB). Next, the template DNA, which is now uniformly single-stranded and coated with SSB, is placed in a phosphorylation/ligation dual reaction with directional dsDNA NGS adapters that contain single-stranded overhangs.

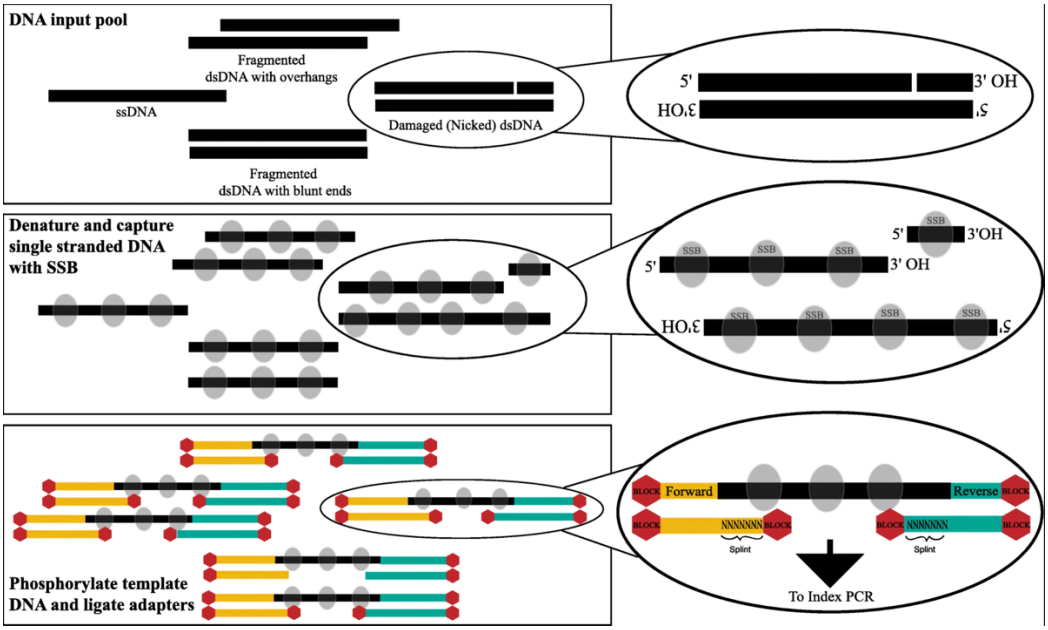

Figure 1.1: Schematic overview of SRSLY. A DNA input pool of diverse template molecules is denatured with heat and maintained as single-stranded molecules through a cold-snap and use of a thermostable single-stranded DNA binding protein (SSB). Template DNA is phosphorylated and SRSLY splint adapters are ligated in a combined phosphorylation/ligation reaction. Adapters contain a random single-stranded splint overhang and ligation blocking modifications on all termini except for the ones that facilitate correctly oriented library molecules. After clean up, molecules are ready for index PCR.

Both the forward and reverse sequencing adapters share similar structures but differ in which termini is unblocked in order to facilitate proper ligations. Both sequencing adapters are dsDNA, except for a single-stranded splint overhang of random nucleotides that occurs on the 3-prime termini of the bottom strand of forward adapter and the 5-prime termini of the bottom strand of the reverse adapter. In this way, the forward (P5) Illumina adapter is always delivered to the 5-prime end of template molecules and the reverse (P7) Illumina adapter is always delivered to the 3-prime end of template molecules. Thus, the native polarity of all input DNA molecules is retained.

During the dual phosphorylation/ligation reaction, T4 Polynucleotide Kinase (PNK) prepares template DNA termini for ligation by phosphorylating 5-prime termini and dephosphorylating 3-prime termini. T4 PNK works on both ssDNA and dsDNA molecules and has no activity on the phosphorylation state of proteins [31,32,33]. Simultaneously, the random nucleotides of the splint adapter anneal to the single-stranded template molecule. This creates a short, localized dsDNA molecule, enabling ligation of template to adapter with T4 DNA ligase, which has high ligation efficiency on double-stranded DNA templates but low efficiency on ssDNA [34]. After the single phosphorylation/ligation reaction is complete, the library DNA is purified and placed directly into standard NGS indexing PCR, compatible with both traditional single or dual index primers

1.2.2 Performance of the SRSLY protocol

To evaluate the quality and quantity of data produced by SRSLY we generated several sequencing libraries from two plasma cfDNA extracts obtained from two healthy human individuals (H-69 and H-81, respectively) using SRSLY, two standard commercially available end-polishing dsDNA library kits (New England Biolabs® NEBNext® Ultra™ II and TaKaRa SMARTer® ThruPLEX® Plasma-Seq) and a popular commercially available ssDNA library kit (Swift Bioscience Accel-NGS® 1S). After library preparation and quantification, libraries were paired-end sequenced on Illumina HiSeq X (2 × 150 bp) to roughly 400 million read pairs per cfDNA extract for SRSLY and NEBNext Ultra II and to roughly 100 million reads pairs per cfDNA extract for TaKara SMARTer and Swift 1S. Sequencing data from libraries generated

from the same cfDNA extract and library preparation method were combined for analysis. We merged the forward and reverse sequence reads when these reads overlap to generate single reads representing the original DNA fragment. Since the majority of sequence reads from cfDNA are about 167 bp long, only merged reads (where read 1 and read 2 overlap by at least 30 bp of complementarity) were used for downstream analyses. Table 1 and Table 2 contain the sequencing metrics for all sequenced libraries. The data generated resulted in about 15-fold coverage of the human genome for both SRSLY and NEBNext Ultra II samples per cfDNA extract and about 5-fold coverage for both TaKaRa SMARTer and Swift 1S per cfDNA extract.

As expected, libraries generated by SRSLY and NEBNext Ultra II cfDNA have length distribution features typical of cfDNA fragments. They both show fragment length distributions centered around the chromatosome length at 167 bp. They both show the sawtooth pattern in shorter fragments that are the result of DNase I cleaving the exposed minor grove of nucleosome bound DNA at a periodicity of 10.4 bp (Figure 1.2A, Figure 1.3). However, as shown in Figure 1.2A and its inset, the two preparation methods differ in the proportion of reads captured at different fragment lengths, as well as the length distribution of the sub-peaks present in the sawtooth pattern. SRSLY libraries have a higher abundance of shorter, i.e. sub-nucleosome length, reads with shorter sub-peaks in the sawtooth pattern versus NEBNext Ultra II. These observations are hallmark features of ssDNA preparation methods [16, 22]. The increased proportion of sub-nucleosome-sized reads reflect the increased ability of ssDNA methods to convert short and/or nicked DNA fragments into sequence library molecules. The

difference in sub-nucleosome peak sizes is likely due to the ability of SRSLY to retain native termini compared to dsDNA methods. In dsDNA library methods, 5-prime overhangs are filled in and 3-prime overhangs are removed. Thus, the observed length of a given DNA molecule will be dependent on what type of overhangs are present. This information is lost during the end-polishing step required in dsDNA library preps.

We compared the read coverage, estimated complexity (number of unique molecules in the library), and GC content of SRSLY versus NEBNext Ultra II libraries for both cfDNA extracts. Figure 1.2B shows that SRSLY produces fold-coverage similar to that of the NEBNext Ultra II kit and that both methods produce relatively uniform genomic coverage. Figure 1.2C shows that at a sequencing depth of 300 million reads, or roughly one HiSeq sequencing lane, SRSLY libraries are estimated to have higher molecular complexity than NEBNext Ultra II libraries. This difference might be a reflection of SRSLY's ability to recover nicked and ssDNA strands lost to traditional dsDNA library preparation. Figure 1.2D shows that the GC content of SRSLY libraries is similar to that of the NEBNext Ultra II kit. The GC content plots for both SRSLY and NEBNext Ultra II are shifted towards GC rich regions compared to the human genome reference (histogram, plotted in green) because cfDNA is biologically enriched for GC-rich regions [35]. The differences shown in regions of low GC content between SRSLY and NEB Ultra II could be either the result of reaction conditions or differences in polymerases used during index PCR.

Most dsDNA library preps, including NEBNext Ultra II, perform end-polishing on the input DNA molecules. Because the SRSLY prep delivers the sequencing

adapters to the native termini of DNA fragments, we can examine the base composition at and around the exact 5-prime and 3-prime end of each DNA fragment with single nucleotide resolution. Note that the end-polishing procedure retains the native 5-prime end of molecules. However, the 5-prime overhang "fill-in" and the 3-prime overhang exonuclease activity of T4 DNA polymerase generates a 3-prime end that is not representative of the original molecule when overhangs of either type are present. In this way, the end-polishing procedure is expected to make all 3-prime ends mirror what is present at the 5-prime end of the complementary strand.

To test these expected differences in DNA termini information, we compared the base composition per position across the start coordinates for both the forward (read 1) and reverse (read 2) reads, inferred from the merged read dataset, for both the SRLSY and the NEBNext Ultra II cfDNA libraries (Figure 1.2E). There are four notable findings. First, for both SRSLY and NEBNext there is significant deviation from the average base composition at the start of each read, as well as upstream of the biological fragmentation point. This is a well-documented feature of the cfDNA nucleosome protection model [16, 36, 37], further discussed in the cfDNA results section below. Second, unlike the dsDNA library data, the average base composition for the start of the forward reads and the start of reverse reads differ in SRSLY libraries. This suggests that cfDNA fragments often contain overhangs that are altered during the end-polishing steps of dsDNA library prep. Third, the average base composition for the start of the forward read in NEBNext Ultra II libraries are exactly the reverse-complement of the average base composition for the start of the reverse read. This is

expected for molecules that are uniformly blunt ended, the byproduct of end-polishing during dsDNA library preparation. Finally, the average base composition for the start of the forward read in SRSLY libraries is nearly identical to that of NEBNext Ultra II libraries. This is the expected result as end-polishing retains the native 5-prime ends, as does the SRSLY direct ligation procedure.

We compared the length distribution, read coverage, complexity, GC content, and DNA termini results of the SRSLY prep to those of the TaKaRa SMARTer and Swift 1S methods as well. In order to do so we randomly down-sampled the SRSLY prep data to 5-fold coverage to adhere to the sequencing depth gathered from both the TaKaRa and Swift preps. The results are detailed in Figure 1.4.

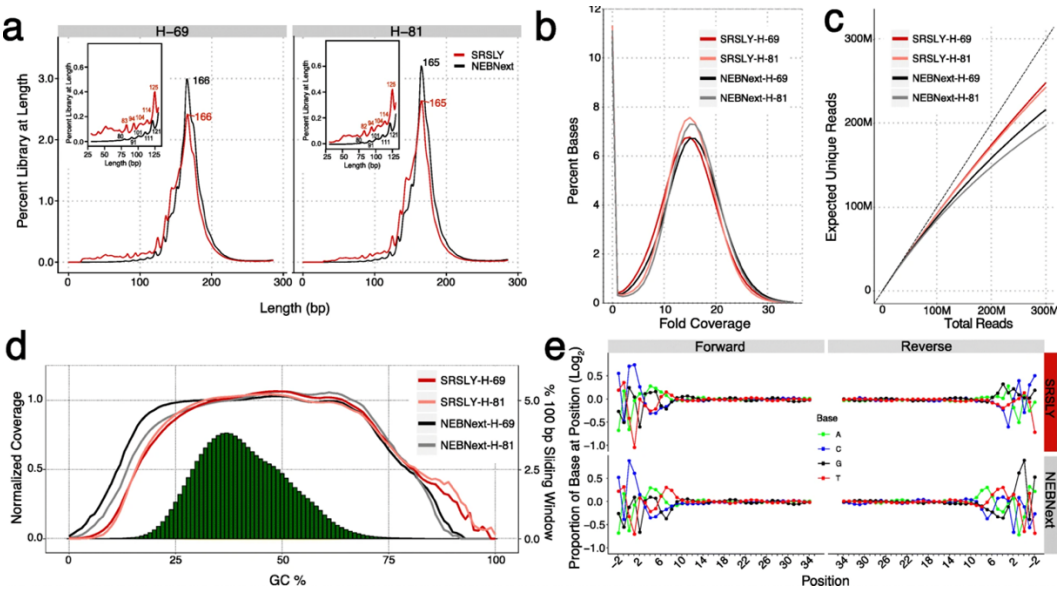

Figure 1.2: Standard NGS metrics for merged reads from SRSLY and NEBNext Ultra II libraries from healthy human cfDNA extracts H-69 and H-81. Unless otherwise stated, all libraries for each method were combined by cfDNA extract prior to analysis and filtered for PCR duplicates and a quality score equal to or greater than q20. **(A)** Insert distribution plots for cfDNA extracts H-69 and H-81, respectively. **(B)** Fold coverage by base percent across the human genome (*hg19*) for SRSLY and NEBNext by cfDNA extract. Combined libraries were subsampled to similar read depth prior to fold coverage calculations. Subsampled depth was set at 295 M reads, the limit of sequenced reads for SRSLY-H-81. **(C)** Preseq complexity estimate for SRSLY and NEBNext by cfDNA extract. Three libraries of equivalent sequencing depth per method were combined to estimate complexity, since more libraries were made via SRSLY than NEBNext. Files containing the PCR duplicate reads were used to facilitate complexity estimates **(D)** Normalized coverage as a function of GC content over 100 bp sliding scale across the human genome for SRSLY and NEBNext by cfDNA extract. Green histogram represents the human genome GC across the 100 bp sliding window. **(E)** Normalized, log-transformed base composition at each position of read termini starting 2 bp upstream and extending to 34 bp downstream of read start site for combined cfDNA extracts for SRSLY and NEBNext. All reads regardless of insert length considered.

1.2.3 5-prime and 3-prime overhangs

Given the base composition differences in cfDNA at the 5-prime and 3-prime ends, we designed an experiment to test whether SRSLY and dsDNA library preparation methods, like NEBNext Ultra II, are altering (or not altering) input DNA fragments as we expect. We constructed pools of 12 synthetic duplexed oligos, at equimolar concentrations, each having a specific length and type (5-prime or 3-prime) overhang. Each duplex contains a 50 nucleotide (nt) core sequence, unique to each overhang type and has a common structure: blunt terminus on one side, and a 5-prime or 3-prime overhang of a specific length of random sequence (one to six nt) on the other side (Fig. 1.5A; Table 1.3).

We generated both SRSLY and NEBNext Ultra II libraries by spiking this pool of oligos into cfDNA extracts. From the sequencing data, we identified reads that originate from the oligo pool by mapping the libraries to a reference file containing the known unique 50 nt core sequences of each oligo. We then calculate the depth of coverage at every position for each oligo in the pool, including the overhangs. Since the duplexed oligos are comprised of two single-stranded molecules with one strand that is one to six nt longer than its complement, we expected the SRSLY method to yield sequence data with 100% coverage across the complementary region and 50% coverage across the overhangs. The results (Fig. 1.5B) confirm that SRSLY produces reduced coverage across the overhanging regions compared to the double-stranded regions of the synthetic oligos illustrating the method's ability to yield stranded data that accurately characterizes the input DNA. By contrast, the libraries produced by NEB Ultra II demonstrate the expected result of end-polishing. Five-prime overhang

are filled-in, resulting in almost full coverage on the complementary strand of molecules with known 5-prime overhangs. Three-prime exonuclease activity, on the other hand, causes nearly complete loss of the 3-prime overhang sequence when it is present.

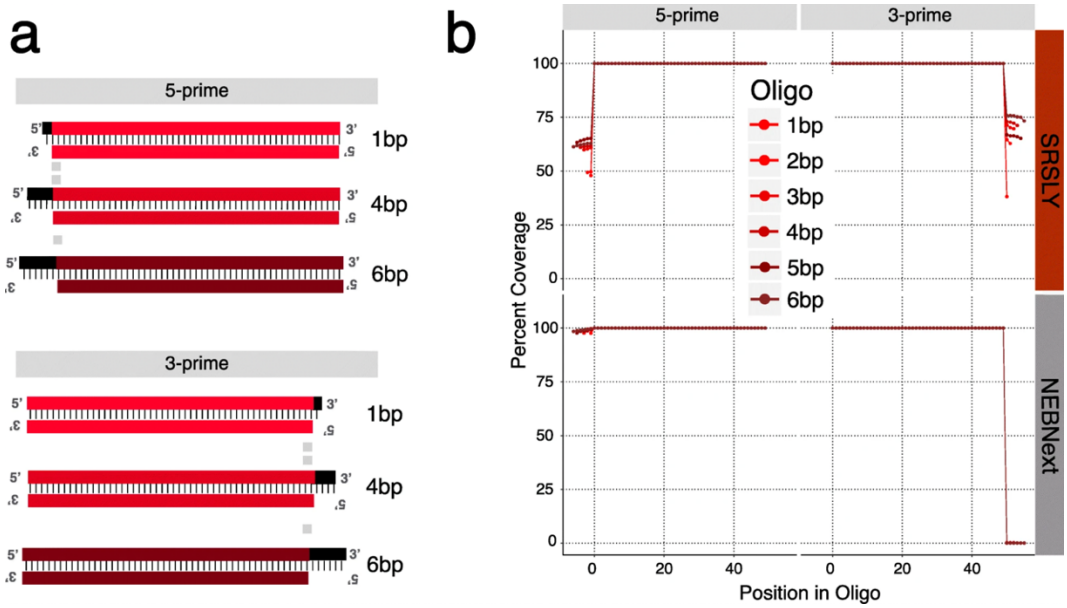

Figure 1.5. Coverage of duplexed oligos containing single-stranded overhangs for SRSLY and NEBNext. (**A**) Cartoon schematic of duplexed synthetic oligos – one blunt end, an identifiable 50 nt complementary region, and an overhang of specific length and type. (**B**) Average coverage per base across the length of all duplexed oligos for three technical replicates in 0 base coordinates for both SRSLY and NEBNext methods. Technical replicates were not statistically different from each other (Students t-test: SRSLY $p = 0.714$, NEBNext $p = 0.985$), error bars not shown for aesthetics. Each oligo sequenced > 5000 reads.

1.2.4 Single-stranded oligo libraries

To test the efficiency of SRSLY on a defined range of input DNA template lengths, we designed and ordered a set of 11 single-stranded oligos (standard desalt purification) of lengths ranging from 20 to 120 nucleotides at 10 nt length intervals

(Table 1.4). We made a pool using equimolar concentrations of each and generated SRSLY libraries from this pool. Analysis of the proportion of template lengths from sequencing these libraries shows that the SRSLY protocol generates ssDNA libraries across this length range (Figure. 1.6A). As a control, we attempted to generate a NEBNext Ultra II libraries from this pool of single-stranded oligos. As expected, this protocol fails to generate any library at all using a template of exclusively single-stranded input DNA (libraries contained adapter dimers but no detectable yield at expected size distributions).

There were several noteworthy observations from the SRSLY data analysis. First, the shortest test oligos (20 and 30 nt length) were under-presented in the libraries. This is likely due to the bead clean-up step after the ligation, which has a known length bias against DNA oligos in this size range. We note that DNA fragments less than 30 nt are often difficult to map uniquely within genomes and are thus of less value, even when present in actual cfDNA samples. Second, there is some variation in library conversion efficiency amongst the longer (>= 40 nt) test oligos. We suspect that this variation is likely due to subtle biases in our test oligos, which are a single, fixed sequence for each length. Finally, we observe a continuous background fraction of oligo lengths that do not correspond to the input oligo lengths. In fact, we observe at least some reads of every length between 20 and 120.

To test whether these reads of unexpected length are due to truncated and incomplete oligo synthesis or due to labile breakage of our longer single-stranded oligos we mapped all the reads in the SRSLY libraries to their respective oligo

reference (Figure. 1.6B, Table 1.5). Truncation products were present for each oligo. These truncated DNA fragments have lengths that are nearly uniformly distributed across the length of the oligo. The fraction of correct, full-length read mapping to each oligo decreases as a function of oligo length. We hypothesize that these two observations demonstrate the limits of the phosphoroamadite method of oligo synthesis. These observations are consistent with a model wherein nucleotide incorporation is less than 100% efficient in each chemical cycle of base addition.

To test whether SRSLY can assess the purity of oligos subjected to various purification methods, we ordered a 60 nt oligo purified using three common schemes: standard desalt, HPLC, and PAGE purification. We constructed SRSLY libraries, in duplicate, using the 60 nt oligo from all three purification methods. Mapping the sequence data to the 60 nt reference sequence (Figure. 1.6C) showed that the proportion of reads attributed to the expected full length sequence increases in both the HPLC and PAGE purified oligo libraries while truncation products, defined as reads at lengths shorter than 60 nt, decrease compared to the libraries generated from standard desalt oligos. These results are consistent with the expected quality of each purification method based on phosphoramidite synthesis (Integrated DNA Technologies Product Literature) and indicate that SRSLY can be used as a simple and sensitive assay to determine the purity of chemically synthesized DNA oligos.

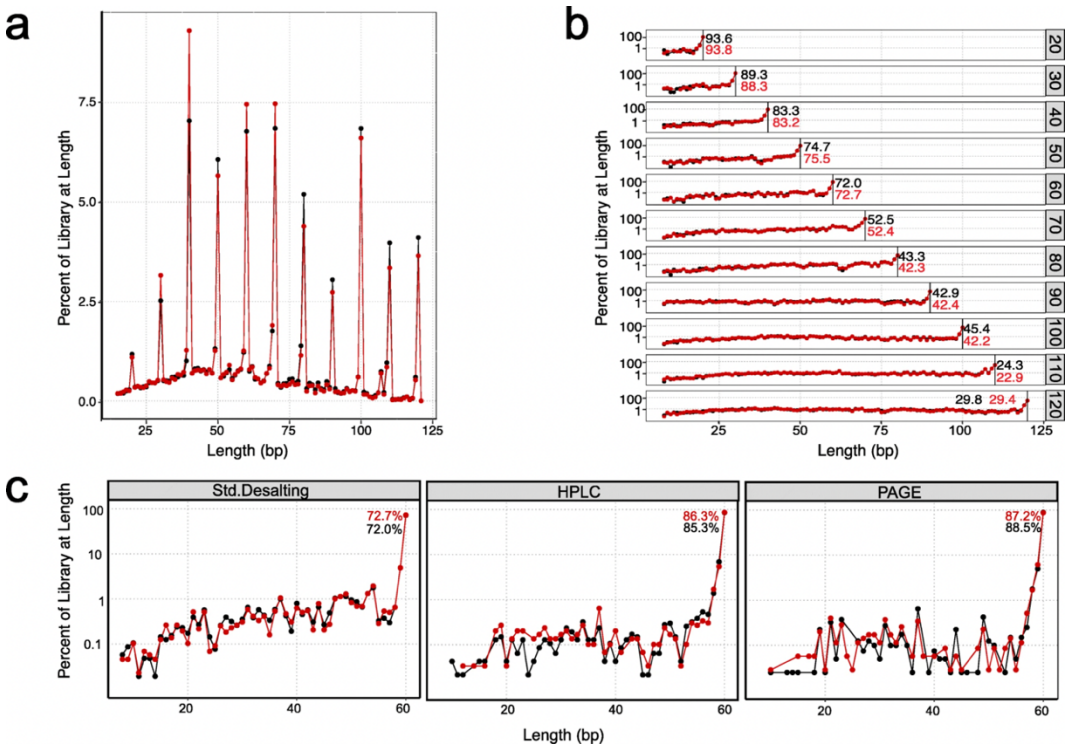

Figure 1.6. Single-stranded oligo analyses by the SRSLY method. Red and black lines and dots represent technical replicates **(A)** Insert distribution of equimolar pooled single-stranded oligo libraries. Oligos from 20 to 120 nt synthesized at 10 nt intervals were purified by standard desalting. Raw unfiltered sequencing data. **(B)** Mapped sequencing data for technical replicates separated by oligo. Represented as a function of oligo length. Black vertical bar and associated black and red numbers indicate percent of full-length product per oligo length present in the library pool. Each library was sequenced to a depth of ~ 100,000 read pairs (10,000 read pairs per oligo, excluding 20 and 30 nt lengths) **(C)** Effects for various purification methods on oligo purity as a function of oligo length for a 60 nt synthesized oligo. Associated black and red numbers indicate percent of full-length product per oligo. Data for the standard desalted 60 nt synthetic oligo pulled from **(B)**.

1.2.5 Analysis of SRSLY cfDNA libraries

The majority of cfDNA fragments derive from DNA wrapped around a nucleosome, a configuration that protects the DNA from nuclease degradation during cell death. Thus, the genomic map positions of cfDNA fragments can be used to infer

the positions of histones and other DNA binding proteins in the tissues that have given rise to a population of cfDNA molecules [16]. Single-stranded DNA library methods, like SRSLY, retain the native ends of cfDNA fragments and are thus maximally useful for inferring the positions of histones and other DNA-binding proteins insofar as these proteins protect the DNA from endonuclease activity. We combined SRSLY data from our two healthy individuals (H-69 and H-81), to obtain 30-fold average genome coverage. From these data, we explored the ability of SRSLY libraries to reveal aspects of the positioning of nucleosomes and other DNA-binding proteins.

It is well established that nucleosome positioning is at least partially encoded by the genome [36, 38]. For DNA bound to histones, A/T dinucleotides are favored when the minor grove faces towards the histone and G/C dinucleotides are favored when the minor grove faces outwards. Therefore, when analyzed in aggregate, DNA fragments originating from nucleosome protected DNA should contain an oscillating pattern of an A/T rich and G/C depleted region directly followed by a G/C rich and A/T depleted region within captured fragments, compared to the surrounding genomic regions. To test whether we observed this oscillation pattern in our SRSLY data we examined the A/T and G/C genomic dinucleotide in molecules of three fragment lengths, 167, 144, and 83 bp, including bases 100 nts upstream and downstream of each of the three read lengths (Figure. 1.7A). We centered each on the midpoint of the sequence. As noted, 167 bp corresponds to the length of DNA wrapped around the nucleosome core particle plus the associated linker region, 144 bp represents the length

of DNA wrapped around the nucleosome core particle only, and 83 bp may represents a degradation product originating from nucleosome-associated DNA.

Consistent with previous results from other ssDNA methods, we observe an oscillation enrichment for A/T and G/C dinucleotides within the sequenced molecule length compared to the surrounding genomic regions [16, 22]. We also observe a strong oscillation signal for ~ 55 bp upstream of the 83 bp fragment length indicating that these molecules are likely derived from degraded nucleosomal associated DNA. We also observe this dinucleotide oscillation within the defined fragment lengths for the NEBNext dsDNA method as well (Figure 1.7B). However, we do not observe the upstream oscillation signal in the 83 bp fragment length for the NEBNext data. This may be due to low recovery of short fragments in the dsDNA preparation methods or other differences in the ability of dsDNA preps to convert fragmented or nicked DNA into sequencing libraries.

An additional feature of dinucleotide-mediated histone wrapping is that DNase I mediated nicking occurs when the minor grove is accessible [36, 39,40,41]. This phenomenon leads to a specific enrichment for G/C dinculeotides at the terminal ends of nucleosome-associated fragments (Figure 1.7A-D). Due to the dsDNA end-polishing step, the terminal profile of the 5-prime and 3-prime ends in NEBNext data are mirror images of each other (Figure 1.7D). The fact that the dinucleotide frequency at 3-prime termini differs considerably between SRSLY and NEBNext suggests that a substantial population of diverse overhangs occurs in a population of nucleosome-associated cfDNA fragments (Figure 1.7C, D).

Next, we looked at nucleosome positioning using the window protection score (WPS) [16]. The WPS is a measure of whether a position in the genome tends to be protected from endonuclease activity or enriched for endonuclease activity. It is a function of how many reads span the given position (and thus were not cut) versus how many reads begin or end at that position (and thus were cut). We calculated the normalized WPS using SRLSY data at a region comprised of well-positioned nucleosomes on chromosome 12. Comparing our WPS results with previously published results using an alternative ssDNA library protocol, we observe good concordance with respect to the location of the peaks and troughs (Figure 1.7C; Overall Pearons Correlation: $r = 0.80$, $p < 0.0001$) [16, 36, 42].

We performed a second WPS validation of our SRSLY data by calculating normalized WPSs for fragments whose lengths fall into a long-sized bin (120–180 bp, the range of fragments lengths presumed to derived from histone protection) and a short-sized bin (35–80 bp, presumed to be enriched for fragments protected by other DNA-binding proteins) within 1 kb upstream or 1 kb downstream of experimentally determined binding sites for the transcription factor CTCF (Figure 1.7D). CTCF is a DNA-binding protein the occludes histones where it is bound and organizes histone positioning upstream and downstream [16, 43, 44]. Consistent with the previously described pattern, we find that the long fragment WPS shows a depression centered at the putative CTCF binding site (position 0) and oscillation patterns extending outward in both directions at a periodicity of ~ 180 bp indicating well-positioned nucleosomes. The short fragment results show a strong peak centered at the putative CTCF binding

site, presumably due to CTCF-protection from endonuclease activity. Upstream and downstream, the smaller amplitude oscillations are consistent with the absence of DNA-binding proteins other than nucleosomes.

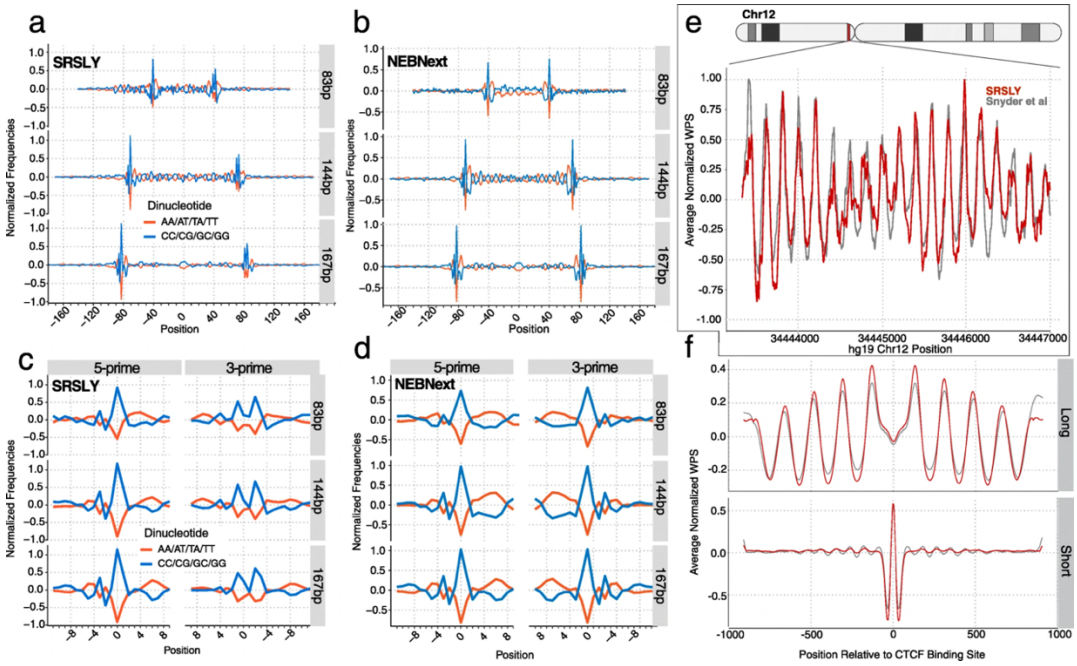

Figure 1.7. **(A)** Normalized genomic dinucleotide frequencies as a function of read length for SRSLY data for three discrete fragment lengths including 100 bp ± the read mapped coordinates. Read midpoint is centered at 0. Negative numbers denote genomic regions upstream (5-prime) of the midpoint and positive numbers denote genomic regions downstream (3-prime) of the midpoint. Input data is from the combined H-69 and H-81 SRSLY datasets. **(B)** Same as **(A)** except for NEBNext data. **(C)** Normalized genomic dinucleotide frequency as a function of read length for SRSLY data for the termini of three discrete fragment lengths including a 9 bp region into the read (positive numbers) and 10 bp outside the read (negative numbers). Read start and end coordinates are centered on 0. Input data is from the combined H-69 and H-81 SRSLY datasets. **(D)** Same as **(C)** except for NEBNext data. **(E)** Normalized WPS values (120 bp window; 120–180 bp fragments) for SRSLY data compared to sample CH01 [16] at the same pericentromeric locus on chromosome 12 used to initially showcase WPS. **(F)** Average normalized WPS score within ±1 kb of annotated CTCF binding sites for long fragment length binned data (120 bp window, 120–180 bp fragments) and short fragment length binned data (16 bp window, 35–80 bp fragments) for SRSLY data compared to sample CH01.

1.3 Discussion

Although the merits of single-stranded NGS library approaches have been well described elsewhere [16, 25], there are no simple, efficient, and widely accessible protocols for making ssDNA libraries. While SRSLY generates sequencing library

molecules from single-stranded DNA fragments, it can be used for DNA that is either single-stranded, double-stranded or a combination of the two. Thus, it is a drop-in replacement for a wide variety of NGS applications. It offers a fast, simple, ligation-based DNA library preparation that relies only on ubiquitously available reagents and an improved splint adapter design to create complex sequencing ready libraries in less than 3 h. The enhanced dual splint adapters allow SRSLY to benefit from the ligation efficiency of T4 DNA ligase. Because the adapter and splint oligos contain ligation-blocking modifications on every end except the ones where ligation should occur, the ligation reaction has been optimized for complete ligation. Unlike previous methods that use T4 DNA ligase to bind splint adapters to single-stranded template, our improved design eliminates the creation of a second strand via extension required for the final ligation, further reducing the possibility of introducing sequencing artifacts or errors into the preparation method.

We present validation of the SRSLY method via comparison to traditional dsDNA library preparation methods and a commercially available ssDNA preparation method showing that SRSLY produces sequencing libraries with uniform coverage, higher complexity, and base composition similar to those of the widely used NEBNext Ultra II kit. In contrast to dsDNA library methods, SRSLY converts a larger proportion of short DNA fragments into sequencing library molecules and retains the native termini of all input DNA fragments. On average, SRSLY cfDNA libraries are comprised of ~ 8% of DNA fragments in the 30–100 bp size bin compared to < 1% for the NEBNext kit. Like others, we have also observed increases in subnucleosomal

DNA content in plasma from cancer patients (data not shown) [45,46,47]. Notably, the proportion of short fragments recovered by SRSLY can be modulated by altering the clean-up step following index PCR (Figure 1.8).

We also demonstrated the utility of SRSLY's native termini retention using two groups of synthetic control oligos. By calculating the depth of coverage at each position for synthetic duplex oligos containing single-stranded DNA overhangs we showed that SRSLY is able to retain strand information from dsDNA and a more accurate characterization of the template molecules. By generating SRSLY libraries from synthetic single-stranded oligos we showed that SRLSY can assay synthetic oligos for artifacts of incomplete synthesis. While this approach is straight-forward and powerful, we note that our assay can only report on DNA fragments with 5-prime and 3-prime ends with the capacity to be ligated. Further exploration may be warranted for more complete analysis of synthetic or biologically-derived DNA fragments that lack ligatable ends.

cfDNA fragments are in many ways an ideal substrate for demonstrating the benefits of ssDNA library preps [16]. cfDNA is often present in low quantities and is comprised of short and often-nicked DNA fragments. Further, the precise mapping positions of cfDNA reads, powered by ssDNA library prep, can reveal an added dimension to sequence-based DNA analysis like the positions of nucleosomes or DNA-binding proteins [48]. We find that the base composition surrounding fragmentation points in cfDNA differs between the 5-prime and 3-prime ends. This observation is consistent with the hypothesis that many or most cfDNA fragments are not blunt-ended

since in that case every 5-prime end would have a corresponding 3-prime end. Further analysis of data generated from SRSLY may reveal further details of the nature of the overhangs present in cfDNA molecules and perhaps the identities of the active nucleases that generate them.

SRSLY is a simple and versatile tool for the preparation of sequencing libraries from fragmented single-stranded DNA. With only slight modifications, SRSLY could be adapted for use for other DNA sources besides cfDNA. The DNA present in FFPE samples is notoriously difficult for high-quality sequencing library preparation because it is fragmented and nicked. In preliminary tests, we have generated high-quality libraries from DNA recovered from FFPE samples and plan to adapt the protocol to the special challenges of this important input source. Another example is using SRSLY in a modified protocol for strand-accurate RNAseq libraries. Most methods for converting RNA into sequencing libraries either do not retain information about which DNA strand was transcribed or required additional steps to mark and destroy or mark and recover one strand of a double-stranded cDNA product. We have performed proof-of-concept experiments wherein first-strand reverse transcriptase products are used directly as input for SRSLY. These preliminary experiments, using a protocol much simpler than those currently available, generate RNAseq libraries retaining strand information as expected and are high complexity.

1.4 Methods

1.4.1 Human cell-free DNA preparation

Whole blood from healthy donors was commercially purchased from Stanford Blood Center, Palo Alto, CA. Donors were deidentified, no biographic or clinical information was provided to Claret Biosciences LLC. Blood plasma was extracted from whole blood by spinning the blood collection tubes at 1800 g for 10 min at 4 °C. Without disturbing the cell layer, the supernatant was transferred to microfuge tubes under sterile conditions in 2 ml aliquots and spun again at 16000 g for 10 min at 4 °C to remove cell debris. cfDNA was prepared from 4 ml plasma using the Circulating Cell-free DNA kit (Qiagen Technologies) following manufacturer's protocol. Concentration of the purified cell-free DNA (cfDNA) was measured using the Quant-iT high sensitivity dsDNA Assay Kit and a Qubit Fluorometer (ThermoFisher Scientific). cfDNA size distribution was analyzed using TapeStation and associated D5000 or D1000 high sensitivity products (Agilent).

1.4.2 Synthetic oligo preparation

Double-stranded synthetic oligos (Additional file 5) were designed using a random sequence generator at 50% GC content; sequences matching any known organism in public databases were removed. Each dsDNA oligo (n = 12) is a unique 50 nt sequence of double-stranded DNA with one blunt-end, and one 3-prime or 5-prime single-stranded overhang of random sequence, 1 to 6 nucleotides in length. Oligos were synthesized using standard desalting purification and duplexed by Integrated DNA Technologies (IDT); all random nucleotides were 'hand-mixed' to reduce synthesis bias. Control oligos were pooled together in an equimolar ratio for SRSLY library preparation.

Single-stranded synthetic oligos (Table 1.4) were generated in the same way as the double-stranded control oligos. Unless otherwise noted, oligos were synthesized using standard desalting purification for ssDNA oligos 20–80 nt in length and Ultramer purification for ssDNA oligos 90–120 nt in length.

1.4.3 SRSLY adapter preparation

The forward (P5) SRSLY adapter as well as the reverse (P7) SRSLY adapter are both double-stranded splint adapters. The forward SRSLY adapter contains a 5-prime overhang in the splint portion of the adapter and a free 3-prime OH end on the ligating end; all other ends contain ligation and/or extension blocking modifications. The reverse SRSLY adapter contains a 3-prime overhang in the splint portion of the adapter and is 5-prime phosphorylated for ligation; all other ends contain ligation and/or extension blocking modifications (Table 1.6). The SRSLY adapters are synthesized using standard desalting purification and duplexed by Integrated DNA Technologies (IDT). Working stocks of the adapters are made by diluting the adapters in TE + 50 mM NaCl.

1.4.4 SRSLY library preparation

1 ng of purified cfDNA or 5 ng of synthesized oligos, as measured by the Quant-iT, was combined with 10 mM Tris pH 8.0 and 8 ng of ET SSB (New England Biolabs) in a 22 µl denaturation reaction, on ice. The reaction was placed in a thermocycler preheated to 95 °C, incubated for 3 min, and then cold shocked on ice for at least 2 min. On ice, 1 pmol of the forward and 1 pmol of the reverse SRSLY adapters were added to the denaturation reaction, as well as PEG-8000, T4 DNA ligase Buffer, T4 PNK,

and T4 DNA ligase (all New England Biolabs) to a final volume of 50 µl. PEG-8000 was added to a final concentration of 18.5% v/v. T4 DNA ligase buffer was added to a final concentration of 1X. T4 PNK and T4 DNA ligase were added to a final concentration of 10 units and 800 units, respectively. This ligation reaction was incubated at 37 °C for one hour and purified using the MinElute PCR Purification Kit (Qiagen) and manufacturer's instructions with the following changes: The initial binding spin was performed at 6000 rpm on a desktop centrifuge. The wash spin was repeated for a total of two wash spins and both washes were performed at 6000 rpm. The DNA was eluted in 15 µl 10 mM Tris pH 8.0.

SRSLY libraries were indexed for sequencing by combining the purified ligated DNA with 1x Kapa HiFi HotStart ReadyMix (Roche) and 2 mM final concentration of universal primer and 2 mM final concentration of an index primer in a 50 µl reaction and amplified using the following thermal cycling conditions: 3 min at 98 °C for initial denaturation followed by 10 cycles at 98 °C for 20 s, 68 °C for 30 s, 72 °C for 30 s, and finally an elongation step of 1 min at 72 °C. After index PCR, SRSLY libraries were purified with a 1.2x AMPure clean (Beckman Coulter) and eluted in 20 µl og 10 mM Tris pH 8.0. Final molarity estimates were calculated using fragment length distribution and dsDNA concentration (Agilent Tapestation 4200 and Qubit Fluorometric Quantitation unit).

1.4.5 NEBNext ultra II library preparation

1 ng of purified cfDNA or 5 ng of synthesized oligos, as measured by the Quant-iT, was taken through library preparation (end-polishing, adapter ligation, index PCR)

as outlined in the NEBNext Ultra II manual using the supplied reagents, recommended AMPure cleanup ratios, and recommended index PCR cycles.

1.4.6 Sequencing

All cfDNA libraries were sequenced on an Illumina® HiSeqX at a 2 × 151 read length by Fulgent Genetics. All synthetic oligo libraries were sequenced on an in-house Illumina MiSeq benchtop sequencer at a read length of 2 × 151 bp following manufacturer's instructions.

1.4.7 Read processing

Sequencing data was first aligned to the PhiX genome using bwa mem [49] with default parameters. Reads that mapped to PhiX were discarded. Next we simultaneously removed adapter sequences and merged the reads as is standard practice in studies where short template molecules are expected [50]. This process consisted of collapsing forward and reverse reads into single sequences, based on sequence similarity, while trimming ends of reads that match known Illumina adapter sequences using SeqPrep (https://github.com/jstjohn/SeqPrep). Merged reads that remained after filtering were aligned to either the hg19 human reference genome (Table 1.1 and Table 1.2) downloaded from the UCSC genome browser [51], or to a custom fasta file corresponding to the synthesized oligo sequence (Table 1.5). We used bwa aln and bwa sampe [52] with default parameters for alignment and mapping. Mapping rates, for human libraries, were determined from samtools flagstat. Duplicate reads were then removed using samtools rmdup.

2.4.8 QC metrics

For most analyses bam files from individual libraries of same preparation method and same cfDNA extract were merged into sample- and library-specific bam files using samtools merge prior to analysis. For insert length distribution of merged reads, for the same preparation method and cfDNA extract insert length information was parsed from the bam files of individual libraries that were generated using samtools view -q20 -f66 and combined using a concatenate command. Frequency of reads per length was calculated and plotted as the percent reads of total library. Normalized genome coverage was extracted from down-sampled merged duplicate removed bam files using samtools view -s such that all libraries had the same number of mapped reads. Data was obtained by pipping downsampled bam files from samtools view -q20 -b into bedtools genomecov. Preseq complexity estimates were obtained by combining only 3 libraries for each cfDNA input sample per library preparation method prior to downsampling in order to not artificially inflate the complexity of SRSLY, which had more libraries per cfDNA extract than NEBNext Ultra II. Libraries combined for SRSLY H-69 were: SR-01, SR-02, SR-03. Libraries combined for SRSLY H-81 were: SR-06, SR-07, SR-08. Libraries combined for NEBNext Ultra II for H-69 and H-81 were NEB-01-03A and NEB-04-06A, respectively. After combining and downsampling the combined bam files to 100 M merged read-pairs, complexity estimates and extrapolation were performed using preseq lcextract [53]. GC coverage was obtained from down-sampled merged duplicate removed bam files utilizing Picard Tools (Broad Institute) CollectGCBiasMetrics. For each library type, fragment terminal nucleotide analysis was done by calculating the proportion of each base i.e.

the base composition, at every position for a region spanning from − 2 to + 34 bases on both reads of a fragment. The base composition per position was normalized with the mode for that base along the length of the region and log-2 transformed. The normalized, log-transformed proportions were calculated for both library types, for both reads and plotted. All plots were generated in R utilizing ggplot2.

1.4.9 Synthetic oligo analysis

Double-stranded synthetic oligo sequencing coverage at each position in the oligo was determined utilizing a custom script akin to samtools depth and plotted in R utilizing ggplot2 as a function of percent across the length of the oligos in 0 base coordinates. Fragment length analysis of single-stranded synthetic oligos was conducted analogous to that for cfDNA.

1.4.10 Biological analysis of cfDNA

For dinucleotide frequency calculations merged bam files from combined H-69 and H-81 libraries for each library preparation method were parsed using samtools view -bh -F 0X10 -m -M -q 20 to extract forward reads of specific insert lengths: 167 bp (chromatosome-wrapped DNA length), 144 bp (core particle-wrapped DNA length, and 83 bp (a shorter DNA length that occurs as a peak in Figure 2A). For each insert length, the dinucleotide counts around both fragmentation points were estimated using a custom python script for all 16 2-mer combination for either a 100 bp or 11 bp window, where 100 bp or 11 bp of genomic context at both 5-prime and 3-prime fragmentation points were added respectively. For the data generated with a 100 bp flanking window on both ends, the overlapping regions (which justifiably had the same

counts) were removed. The data was normalized using a median filter and dinucleotide frequency was plotted for weak (AA/AT/TA/TT) vs strong (CC/CG/GC/GG) dinucleotide interaction such that the center of the insert was at 0 and the regions upstream of the fragmentation point had negative values and downstream had positive values. For the data generated with a 11 bp flanking window, the data was normalized with a median filter and dinucleotide frequencies of weak vs strong dinucleotide were plotted for 5-prime and 3-prime ends using R.

WPS scores were calculated in the manner previously described [16]: The WPS score for each position in the genome was determined by collecting the reads which align in a window around that position, 120 bp in the case of large fragment analysis and 35 bp in the case of short fragment analysis. The score was calculated as follows: Every time an insert starts or end in that window, one is subtracted from the score. If an insert does not start or end in that window, but aligns to it nevertheless, one is added to the score. The normalized WPS score was calculated by taking the WPS scores over non-overlapping 1000 bp segments and adjusting to a median score of zero by subtracting the median WPS score. The scores were then smoothed by the Savitzky–Golay filter: second-order polynomials were fitted to median-adjusted scores over a 21 bp window at each position. The smoothed score is the value of that polynomial at that position. The Average WPS score is calculated over a set of regions of equal length by calculating the mean of the WPS scores over each position in each of the regions in our set, where position 1 is the first nucleotide of each region in our set, position 2 is the second nucleotide in each region, etc. CTCF sites were chosen in a method similar

to what was described previously [16]. A bed file containing a list of putative TF binding sites was downloaded from the JASPAR2018 table(hub_186875_JasparTFBS) from the UCSC Genome Browser Table Browser into a bed file and filtered to include only CTCF sites. These sites were compared with CTCF ChIP-Seq data from 19 cell lines [54]. Putative binding sites with overlapping ChIP-Seq peaks in all 19 cell lines were used for further analysis.

1.5 References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.

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

3. Bentley DR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9.

4. Adey A, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 2010;11:R119.

5. Bennett, E. A. et al. Library construction for ancient genomics: single strand or double strand? Biotechniques 56, 289–290, 292–286, 298, passim.

6. Dabney J, et al. Complete mitochondrial genome sequence of a middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci U S A. 2013;110:15758–63.

7. Meyer M, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338:222–6.

8. Gansauge MT, Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc. 2013;8:737–48.

9. Aravanis AM, Lee M, Klausner RD. Next-generation sequencing of circulating tumor DNA for early Cancer detection. Cell. 2017;168:571–4.

10. Agardh E, et al. Genome-wide analysis of DNA methylation in subjects with type 1 diabetes identifies epigenetic modifications associated with proliferative diabetic retinopathy. BMC Med. 2015;13:182.

11. De Vlaminck I, et al. Temporal response of the human virome to immunosuppression and antiviral therapy. Cell. 2013;155:1178–87.

12. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A. 2008;105:16266–71.

13. Jiang P, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci U S A. 2015;112:E1317–25.

14. Sun K, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A. 2015;112:E5503–12.

15. Jahr S, et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Res. 2001;61:1659–65.

16. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell. 2016;164:57–68.

17. Lo YM, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med. 2010;**2**:61ra91.

18. Murtaza M, et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature. 2013;497:108–12.

19. Newman AM, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. 2014;20:548–54.

20. Newman AM, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol. 2016;34:547–55.

21. Tie J, et al. Circulating tumor DNA as an early marker of therapeutic response in patients with metastatic colorectal cancer. Ann Oncol. 2015;26:1715–22.

22. Wu DC, Lambowitz AM. Facile single-stranded DNA sequencing of human plasma DNA via thermostable group II intron reverse transcriptase template switching. Sci Rep.

23. Mouliere F, et al. High fragmentation characterizes tumour-derived circulating DNA. PLoS One. 2011;6:e23418.

24. Quake S. Sizing up cell-free DNA. Clin Chem. 2012;58:489–90.

25. Burnham P, et al. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. Sci Rep. 2016;6:27859.

26. Gansauge MT, et al. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. Nucleic Acids Res. 2017;45:e79.

27. Raine A, Manlig E, Wahlberg P, Syvanen AC, Nordlund J. SPlinted ligation adapter tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. Nucleic Acids Res. 2017;45:e36.

28. Turchinovich A, et al. Capture and amplification by tailing and switching (CATS). An ultrasensitive ligation-independent method for generation of DNA libraries for deep sequencing from picogram amounts of DNA and RNA. RNA Biol. 2014;11:817–28.

29. Wu J, Dai W, Wu L, Wang J. SALP, a new single-stranded DNA library preparation method especially useful for the high-throughput characterization of chromatin openness states. BMC Genomics. 2018;19:143.

30. Wu J, et al. Decoding genetic and epigenetic information embedded in cell free DNA with adapted SALP-seq. Int J Cancer. 2019;145:2395–2406.

31. Soltis DA, Uhlenbeck OC. Isolation and characterization of two mutant forms of T4 polynucleotide kinase. J Biol Chem. 1982;257:11332–9.

32. Wang LK, Lima CD, Shuman S. Structure and mechanism of T4 polynucleotide kinase: an RNA repair enzyme. EMBO J. 2002;21:3873–80.

33. Wang LK, Shuman S. Domain structure and mutational analysis of T4 polynucleotide kinase. J Biol Chem. 2001;276:26868–74.

34. Kuhn H, Frank-Kamenetskii MD. Template-independent ligation of single-stranded DNA by T4 DNA ligase. FEBS J. 2005;272:5991–6000.

35. Kostyuk S, et al. GC-rich extracellular DNA induces oxidative stress, double-Strand DNA breaks, and DNA damage response in human adipose-derived Mesenchymal stem cells. Oxidative Med Cell Longev. 2015;2015:782123.

36. Gaffney DJ, et al. Controls of nucleosome positioning in the human genome. PLoS Genet. 2012;8:e1003036.

37. Harshman SW, Young NL, Parthun MR, Freitas MA. H1 histones: current perspectives and challenges. Nucleic Acids Res. 2013;41:9593–609.

38. Satchwell SC, Drew HR, Travers AA. Sequence periodicities in chicken nucleosome core DNA. J Mol Biol. 1986;191:659–75.

39. Boyle AP, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132:311–22.

40. Cousins DJ, et al. Redefinition of the cleavage sites of DNase I on the nucleosome core particle. J Mol Biol. 2004;335:1199–211.

41. Segal E, et al. A genomic code for nucleosome positioning. Nature. 2006;442:772–8.

42. Valouev A, et al. Determinants of nucleosome organization in primary human cells. Nature. 2011;474:516–20.

43. Fu Y, Sinha M, Peterson CL, Weng Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. 2008;4:e1000138.

44. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet. 2014;15:234–46.

45. Jiang P, Lo YMD. The long and short of circulating cell-free DNA and the ins and outs of molecular diagnostics. Trends Genet. 2016;32:360–71.

46. Lapin M, et al. Fragment size and level of cell-free DNA provide prognostic information in patients with advanced pancreatic cancer. J Transl Med. 2018;16:300.

47. Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nat Rev Cancer. 2011;11:426–37.

48. Sanchez C, Snyder MW, Tanos R, Shendure J, Thierry AR. New insights into structural features and optimal detection of circulating tumor DNA determined by single-strand DNA analysis. NPJ Genom Med. 2018;3:31.

49. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv*:1303.3997* (2013).

50. Kircher M. Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol. 2012;840:197–228.

51. Kent WJ, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.

52. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60.

53. Daley T, Smith AD. Predicting the molecular complexity of sequencing libraries. Nat Methods. 2013;10:325–7.

54. Wang H, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. 2012;22:1680–8.

Library ID	cfDNA extract	Raw read pairs	Pass filter read pairs	Merged read pairs	Mapped read pairs	Duplicate read pairs
SR-01	H-69	94,786,943	86,884,321 (91.7%)	74,059,050 (85.2%)	69,735,053 (77.7%)	6,646,784 (9.5%)
SR-02	H-69	94,297,123	89,775,122 (95.2%)	75,364,887 (83.9%)	74,408,496 (85.6%)	7,342,661 (9.9%)
SR-03	H-69	81,474,103	77,874,288 (95.6%)	65,201,784 (83.7%)	64,727,039 (83.1%)	5,958,851 (9.2%)
SR-04	H-69	98,450,841	90,642,659 (92.1%)	76,197,502 (84.1%)	74,101,090 (81.8%)	6,981,901 (9.4%)
SR-05	H-69	115,200,247	105,758,818 (91.8%)	86,567,929 (81.9%)	85,410,946 (80.8%)	9,571,162 (11.2%)
All	H-69	484,209,257	450,935,208 (93.1%)	377,391,152 (83.7%)	368,382,624 (81.7%)	36,501,359 (9.9%)
SR-06	H-81	84,140,424	80,948,813 (96.2%)	71,429,415 (88.2%)	68,958,103 (85.2%)	9,122,834 (13.2%)
SR-07	H-81	74,670,157	71,559,425 (95.8%)	63,111,643 (88.2%)	61,087,692 (85.4%)	7,380,490 (12.1%)
SR-08	H-81	77,438,201	74,583,049 (96.3%)	65,654,147 (88.0)	63,686,313 (85.4%)	8,372,356 (13.1%)
SR-09	H-81	84,600,059	81,361,847 (96.2%)	72,265,939 (88.8%)	70,187,322 (86.3%)	8,495,259 (12.1%)
SR-10	H-81	77,177,608	74,493,904 (96.5%)	66,365,109 (89.1%)	64,450,156 (86.5%)	8,256,944 (12.8%)
All	H-81	398,026,449	382,947,038 (96.2%)	338,826,253 (88.5%)	328,369,586 (85.7%)	41,627,883 (12.7%)

Table 1.1. SRSLY human cfDNA extract NGS statistics.

Library ID	Preparation kit	cfDNA extract	Raw read pairs	Pass filter read pairs	Merged read pairs	Mapped read pairs	Duplicate read pairs
NEB-01A	NEB Ultra II	S0069	56,804,742	56,581,216 (99.6%)	47,014,204 (83.1%)	49,075,274 (86.7%)	4,286,898 (8.7%)
NEB-02A	NEB Ultra II	S0069	62,159,092	61,939,780 (99.6%)	51,079,841 (82.5%)	53,969,240 (87.1%)	5,281,009 (9.8%)
NEB-03A	NEB Ultra II	S0069	54,665,148	54,415,552 (99.5%)	44,943,923 (82.6%)	47,100,881 (86.6%)	3,917,226 (8.3%)
NEB-04A	NEB Ultra II	S0081	45,040,369	44,945,092 (99.8%)	39,385,036 (87.6%)	40,229,499 (89.5%)	4,519,419 (11.2%)
NEB-05A	NEB Ultra II	S0081	40,276,528	40,208,800 (99.8%)	34,083,417 (84.8%)	35,894,023 (89.3%)	3,951,972 (11.0%)
NEB-06A	NEB Ultra II	S0081	38,184,951	38,107,032 (99.8%)	33,641,066 (88.3%)	34,066,689 (89.4%)	3,653,615 (10.7%)
NEB-01B	NEB Ultra II	S0069	145,586,438	145,391,68(99.9%)	124,430,880 (85.6%)	130,678,620 (89.9%)	21,702,629 (16.6%)
NEB-02B	NEB Ultra II	S0069	167,080,644	166,862,822 (99.9%)	141,337,108 (84.7%)	150,007,151 (89.9%)	27,522,314 (18.3%)
NEB-03B	NEB Ultra II	S0069	149,861,198	149,652,929 (99.9%)	128,254,214 (85.7%)	134,494,174 (89.9%)	21,894,091 (16.3%)
NEB-04B	NEB Ultra II	S0081	135,165,144	134,915,960 (99.8%)	117,427,639 (87.0%)	121,157,380 (89.8%)	20,317,676 (16.8%)
NEB-05B	NEB Ultra II	S0081	134,972,563	134,765,678 (99.8%)	112,962,099 (83.8%)	120,756,517 (89.6%)	20,380,321 (16.9%)
NEB-06B	NEB Ultra II	S0081	118,888,514	118686258 (99.8%)	104,471,069 (88.0%)	106,579,341 (89.8%)	17,229,569 (16.2%)
SWT-01A	Swift NGS 1S Accel	S0069	57,267,715	50,061,414 (87.4%)	42,649,149 (85.2%)	41,942,763 (83.8%)	4,410,212 (8.8%)
SWT-02A	Swift NGS 1S Accel	S0069	56,758,713	53,102,126 (93.6%)	45,451,975 (85.6%)	44,812,574 (84.4%)	4,538,421 (8.6%)
SWT-03A	Swift NGS 1S Accel	S0069	65,931,917	61,112,695 (92.7%)	51,810,429 (84.8%)	51,802,844 (84.8%)	5,815,032 (9.5%)
SWT-01B	Swift NGS 1S Accel	S0081	46,707,495	45.667,114 (97.8%)	41,471,348 (90.8%)	39,862,279 (87.3%)	5,161,043 (11.3%)
SWT-02B	Swift NGS 1S Accel	S0081	46,016,658	44,426,052 (96.5%)	40,374,114 (90.9%)	38,466,813 (86.6%)	4,755,327 (10.7%)
SWT-03B	Swift NGS 1S Accel	S0081	40,339,362	38,931,088 (96.5%)	35,423,108 (91.0%)	33,807,212 (86.8%)	3,917,002 (10.1%)

TKA-01A	TaKaRa ThruPLEX Plasma-Seq	S0069	70,159,339	69,775,464 (99.5%)	59,077,765 (84.7%)	59,758,638 (85.6%)	6,339,914 (9.1%)
TKA-02A	TaKaRa ThruPLEX Plasma-Seq	S0069	72,536,063	72,129,821 (99.4%)	60,625,330 (84.1%)	61,509,949 (85.3%)	6,299,368 (8.7%)
TKA-03A	TaKaRa ThruPLEX Plasma-Seq	S0069	63,650,274	63,315,928 (99.5%)	53,369,955 (84.3%)	54,307,565 (85.8%)	5,227,134 (9.5%)
TKA-01B	TaKaRa ThruPLEX Plasma-Seq	S0081	46,630,408	46,455,843 (99.6%)	41,764,957 (89.9%)	41,107,671 (88.5%)	5,220,268 (11.2%)
TKA-02B	TaKaRa ThruPLEX Plasma-Seq	S0081	46,385,711	46,171,497 (99.5%)	41,478,478 (89.8%)	40,770,738 (88.3%)	4,723,961 (10.2%)
TKA-03B	TaKaRa ThruPLEX Plasma-Seq	S0081	44,421,901	44,254,484 (99.6%)	40,012,865 (90.4%)	39,169,322 (88.5%)	4,961,394 (11.2%)

Table 1.2. Comparison library preps human cfDNA extract NGS statistics.

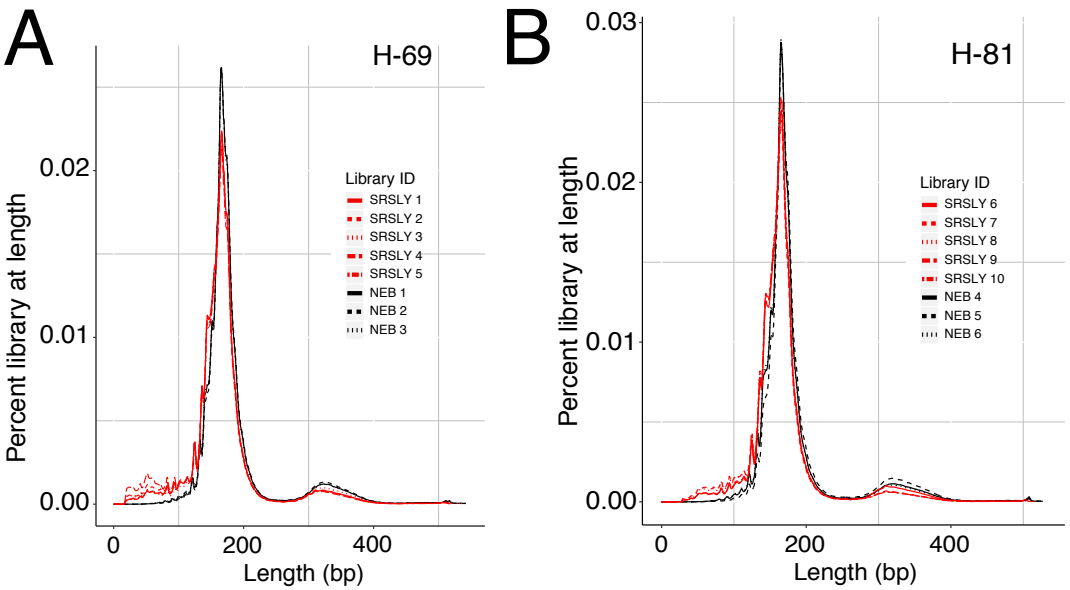

Figure 1.3. Insert distributions for replicate libraries for H-69 and H-81. (**A**) Insert distribution for all libraries made for cfDNA extract H-69. (**B**) Insert distribution for all libraries made for cfDNA extract H-81.

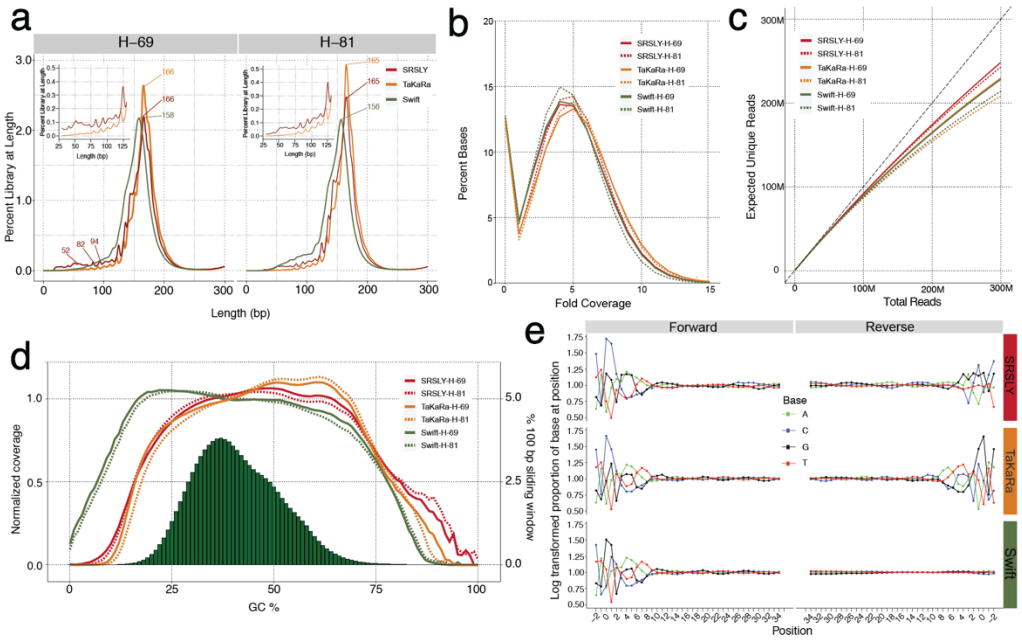

Figure 1.4. (A) Insert distribution plots for cfDNA extracts H-69 and H-81, respectively. TaKaRa data produces a similar distribution as the NEBNext Ultra II data (Figure 1.2). Due to data preprocessing, the main peak of the Swift data is artificially at a shorter length and the biologically relevant sawtooth pattern in fragments shorter than 130 bp is mostly lost. **(B)** Fold coverage by base percent across the human genome (hg19) for SRSLY, TaKaRa, and Swift based on 100 million merged read-pairs per cfDNA extract. All libraries produce similar fold-coverage and relatively uniform genomic coverage. **(C)** Preseq complexity estimate for SRSLY, TaKaRa, and Swift by cfDNA extract. All three methods produce high complexity libraries with SRSLY estimated complexity higher than TaKaRa or Swift for both cfDNA extracts. **(D)** Normalized coverage as a function of GC content over 100 bp sliding scale across the human genome for SRSLY, TaKaRa, and Swift by cfDNA extract. GC coverage for SRSLY and TaKaRa follow similar trends with TaKaRa having slightly higher coverage in genomic regions with 50% – 75% GC percent. Swift distribution is biased towards AT rich regions when compared to both SRSLY and TaKaRa data. **(E)** Normalized, log-transformed base composition at each position of read termini starting 2 bp upstream and extending to 34 bp downstream of read start site for combined cfDNA extracts for SRSLY, TaKaRa, and Swift. TaKaRa data reproduces the results seen by the NEBNext Ultra II data (Figure 1.2). Due to data preprocessing, the 3-prime signal (start of the reverse read) of the Swift data is lost. Also, the fragmentation biases around the forward read position 0 for the Swift data deviates for G and C bases from those observed in the SRSLY, TaKaRa, and NEBNext data.

Overhang Type	Sequence 1	Sequence 2
3' 1bp	CCATACTGTGGTCGTCACCTATTACCCCGCGTAAAGGTAGGCTATGTCATN$_1$	ATGACATAGCCTACCTTTACGCGGGGTAATAGGTGACGACCACAGTATGG
3' 2bp	GTGAATTGTTGATGTCCTGGGTGCCTCGTCCCAAAAGCTGTCCTCACGACN$_2$	GTCGTGAGGACAGCTTTTGGGACGAGGCACCCAGGACATCAACAATTCAC
3' 3bp	GCTTCTCGAACCCGCGATCCGGCCGATCCGGCATAATGGGTTGATTTAGAN$_3$	TCTAAATCAACCCATTATGCCGGATCGGCCGGATCGCGGGTTCGAGAAGC
3' 4bp	CGACACGGATATTCCATCAAGAGACGGGCCTATGGTCCCTGTGATGATGTN$_4$	ACATCATCACAGGGACCATAGGCCCGTCTCTTGATGGAATATCCGTGTCG
3' 5bp	ACCTTGTGTGTTGCTGAAGCAAAGCCGCGTGACCGTTTTAACCAGCGAACN$_5$	GTTCGCTGGTTAAAACGGTCACGCGGCTTTGCTTCAGCAACACACAAGGT
3' 6bp	ATTTTACCACGAGTTCCTTACGACGGCTGTGATGCCACGGTAGGCAGGTAN$_6$	TACCTGCCTACCGTGGCATCACAGCCGTCGTAAGGAACTCGTGGTAAAAT
5' 1bp	N$_1$CGCTTTACGGGTCCTGGGCCGGGGTGCGATACCTTGCAGAAATCGAGGCC	GGCCTCGATTTCTGCAAGGTATCGCACCCCGGCCCAGGACCCGTAAAGCG
5' 2bp	N$_2$AGGACTCTGCCGTCGACGAGTTCGTTAATTCACGGCATCACGTGCGTAGT	ACTACGCACGTGATGCCGTGAATTAACGAACTCGTCGACGGCAGAGTCCT
5' 3bp	N$_3$ACCTCCGTCGCGCTATGTTCTGTTGCATTCGACCTTCTCCGTTCTGTGGG	CCCACAGAACGGAGAAGGTCGAATGCAACAGAACATAGCGCGACGGAGGT
5' 4bp	N$_4$ACAAGAGGAGCATCCGTATTACCGCCTATATCGCCTACGTTTAGAGCATT	AATGCTCTAAACGTAGGCGATATAGGCGGTAATACGGATGCTCCTCTTGT
5' 5bp	N$_5$GTAAATCCCACACAGCTGTCGGCTTATATGGTCATTGGACGGCGTAATAG	CTATTACGCCGTCCAATGACCATATAAGCCGACAGCTGTGTGGGATTTAC
5' 6bp	N$_6$CCAGACAGCCATAGAGGTTACAAGCATAGCAATTTGCATCAGTTCGCAGA	TCTGCGAACTGATGCAAATTGCTATGCTTGTAACCTCTATGGCTGTCTGG

Table 1.3. Synthetic duplexed oligos sequences.

Oligo	Sequence (5' -> 3')
20mer	GTA AAG GTA GGC TAT GTC AT

30mer	GTG CCT CGT CCC AAA AGC TGT CCT CAC GAC
40mer	GCT TCT CGA ACC CGC GAT CCG GCC GAT CCG GCA TAA TGG G
50mer	CGA CAC GGA TAT TCC ATC AAG AGA CGG GCC TAT GGT CCC TGT GAT GAT GT
60mer	ATT TTA CCA CAC CTT GTG TGT TGC TGA AGC AAA GCC GCG TGA CCG TTT TAA CCA GCG AAC
70mer	CCA TTC GGG CAT AAT ATG AAC TAT ACG CAG CTT ATC CCG GGC CCG TAA CAA ACA ATT TGC GTG AGG TAT G
80mer	GTC CCA CTC AGA GAA TTA GCA GCC CTG GTC TAG CGA GGG ATG CCG CTT AGC GTC GGT TGA ATT TCG CTG CAC TAC AGA CG
90mer	CGC TTT ACG GGT CCT GGG CCG GGG TGC GAT ACC TTG CAG AAT CTG CGC CTC TTG GTG GCG CCC CAT CAG TAG TGT CTA CAC GGG CGC TGT
100mer	GTA AAT CCC ACA CAG CTG TCG GCT TAT ATG GTC ATT GGA CGG CGT AAT AGA CAA GAG GAG CAT CCG TAT TAC CGC CTA TAT CGC CTA CGT TTA GAG CAT T
110mer	GGT TCC TAA CAG GTG ATT ACC AGT GCA GTT AGC CAT TTA TCC TCG TCA AAA AGC CAC GTT CCA GAC AGC CAT AGA GGT TAC AAG CAT AGC AAT TTG CAT CAG TTC GCA GA
120mer	GAC GGC CCT AGT CTG CTT CTC GAG ACA ATC TGC TAG AAC TCG GAC GCC TCG CAC TGT ACT GAT GCA TGG TCC GTA ATC GAG GTG AAA ACT ACA CGG TAT GAC ATC AGC GAT AAC TGG TTT

Table 1.4. Synthetic single-stranded oligos sequences.

Library ID	**Oligo length**	**Raw mapped reads**	**Library ID**	**Oligo Length**	**Raw mapped reads**
Replicate 1	20 bp	1241	Replicate 2	20 bp	876
Replicate 1	30 bp	2864	Replicate 2	30 bp	2918
Replicate 1	40 bp	9802	Replicate 2	40 bp	10144
Replicate 1	50 bp	9340	Replicate 2	50 bp	6481
Replicate 1	60 bp	10437	Replicate 2	60 bp	8761
Replicate 1	70 bp	15275	Replicate 2	70 bp	13055
Replicate 1	80 bp	14465	Replicate 2	80 bp	9531
Replicate 1	90 bp	8229	Replicate 2	90 bp	5678

Replicate 1	100 bp	18801	Replicate 2	100 bp	14240
Replicate 1	110 bp	20577	Replicate 2	110 bp	13938
Replicate 1	120 bp	17567	Replicate 2	120 bp	12004
Replicate 1	All	128598 (96.47%)	Replicate 2	All	97626 (96.24%)
Replicate 1	60 bp HPLC	4686	Replicate 2	60 bp HPLC	2966
Replicate 1	60 bp PAGE	4038	Replicate 2	60 bp PAGE	3520

Table 1.5. Synthetic single-stranded oligo raw read counts.

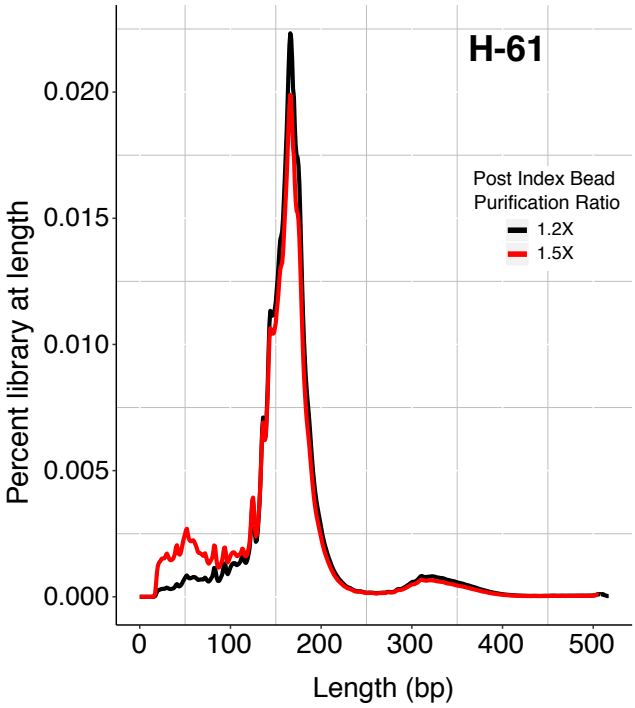

Figure 1.8. Effect of post index PCR DNA purification on SRSLY fragment length retention. SRSLY libraries for cfDNA H-69 were purified using either a 1.2x or 1.5x DNA purification bead volume:Index PCR reaction volume ratio. Recovery of <100 bp fragments changed from 9.3% to 14.7% for the higher ratio from the lower ratio.

Adapter	Sequence 1 (Adapter)	Sequence 2 (Splint)
Forward (P5)	/5AmMC12/ACACTCTTTCCCTACACGACGCTCTTCCGATCT	/5AmMC6/NNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT/3AmMO/
Reverse (P7)	/5Phos/AGATCGGAAGAGCACACGTCTGAACTCCAGTCA/3ddC/	/5AmMC12/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNN/3AmMO/

Table 1.6. SRSLY adapter design.

Chapter 2: A fast and efficient single-stranded genomic library preparation method optimized for ancient DNA

This project was a collaboration with Richard E. Green and Beth Shapiro. The text was originally published in the Journal of Heredity with the title 'A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA'. I designed and performed experiments, prepared the DNA extractions, prepared the libraries, performed the analysis, and wrote the first draft of the manuscript. Beth Shapiro and Richard E. Green provided input throughout the study, assisted in data analysis, and revised the manuscript. I have re-formatted the original manuscript.

2.1 Abstract

We present a protocol to prepare extracted DNA for sequencing on the Illumina sequencing platform that has been optimized for ancient and degraded DNA. Our approach, the Santa Cruz Reaction or SCR, uses directional splinted ligation of Illumina's P5 and P7 adapters to convert natively single-stranded DNA and heat denatured double-stranded DNA into sequencing libraries in a single enzymatic reaction. To demonstrate its efficacy in converting degraded DNA molecules, we prepare five ancient DNA extracts into sequencing libraries using the SCR and two of the most commonly used approaches for preparing degraded DNA for sequencing: BEST, which targets and converts double-stranded DNA, and ssDNA2.0, which targets and converts single-stranded DNA. We then compare the efficiency with which each approach recovers unique molecules, or library complexity, given a standard amount of DNA input. We find that the SCR consistently outperforms the BEST protocol in recovering unique molecules and, despite its relative simplicity to perform and low cost

per library, has similar performance to ssDNA2.0 across a wide range of DNA inputs. The SCR is a cost- and time-efficient approach that minimizes the loss of unique molecules and makes accessible a taxonomically, geographically, and temporally broader sample of preserved remains for genomic analysis.

2.2 Introduction

Ancient DNA, or DNA that persists after organismal death, can provide unique insights into evolutionary history. Over the last three decades, ancient DNA has been used to place extinct taxa in phylogenetic trees (Shapiro et al. 2002; Bunce et al. 2005; Orlando et al. 2008; Bunce et al. 2009) to reconstruct dynamics of extinct populations and communities (Shapiro et al. 2004; Stiller et al. 2010; Lorenzen et al. 2011), and to reveal past ecological changes such as extinction events or turnovers in community composition (Graham et al. 2016; Pedersen et al. 2016). With the advent of Next Generation Sequencing (NGS) technologies and consequent ability to sequence much shorter DNA fragments, the temporal and geographic scope of ancient DNA has expanded (Orlando et al. 2013; Brace et al. 2016; Meyer et al. 2017) and it has become feasible to sequence entire genomes from extinct species, which has facilitated the reconstruction of fine-scale evolutionary histories for many species, including our own (Green et al. 2010; Lazaridis et al. 2014). Despite these successes, the field remains limited by challenges in efficiently recovering short fragments of often damaged DNA from preserved biological remains, all of which are ultimately finite resources (Shapiro and Hofreiter 2014; Green and Speller 2017).

After organismal death, DNA damage accumulates in every cell via environmental, enzymatic, and chemical mechanisms (Dabney, Meyer, et al. 2013).

Most commonly, DNA strands become fragmented, likely via the accumulation of single-stranded breaks by hydrolytic depurination followed by β elimination (Lindahl 1993; Briggs et al. 2007). Additionally, the resulting fragments become chemically altered via deamination of cytosines to uracils at strand termini (Hofreiter et al. 2001; Briggs et al. 2007). The rate of depurination is influenced by temperature (Lindahl and Nyberg 1972), which means that preservation is environment-dependent, with the slowest degradation occurring in cold, temperature stable, and dry environments (Smith et al. 2003). Following collection from the environment, formalin-fixation (Van Beers et al. 2006; Stiller et al. 2016) or storage in warm or moist environments can also promote degradation. As a consequence of accumulating DNA damage, the number of recoverable molecules decays over time and, consequently, so does the sample's utility for ancient DNA analysis.

Over the last three decades, methods have been developed that optimize recovery and processing of the short and damaged DNA fragments preserved in organismal remains. Ancient DNA optimized extraction protocols, for example, use in-solution silica binding (Rohland and Hofreiter 2007), silica spin columns (Dabney, Knapp, et al. 2013), or magnetic beads (Rohland et al. 2018), to retain short DNA molecules. Approaches have also been developed that increase the quality and proportion of extracted authentically ancient molecules, for example, by repairing or excising DNA damage (Briggs et al. 2009; Mouttham et al. 2015; Rohland et al. 2015), by enzymatic depletion of microbial DNA (Green et al. 2010), or by enriching for damaged (and therefore authentically ancient) molecules (Gansauge and Meyer 2014).

After extraction, ancient DNA molecules must be converted into sequencing libraries via the addition of platform-specific DNA sequencing adapters at the ends of each molecule. Most commercially available library preparation approaches for Illumina sequencing perform poorly with damaged and degraded DNA (Stiller et al. 2016; Gansauge et al. 2017). For example, the commonly used New England Biolabs (NEB) Ultra II DNA library preparation kit discards short fragments during clean-up steps, cuts uracil bases (which form naturally in ancient extracts via depurination) with the USER enzyme cocktail, and uses a non-uracil tolerant polymerase, all of which will reduce recovery of ancient DNA molecules. These challenges have led to development of several ancient DNA specific approaches to library preparation.

The most commonly used library preparation methods optimized for degraded DNA process double-stranded DNA (dsDNA). The Meyer and Kircher (MK) approach (Meyer and Kircher 2010), for example, includes an end-repair step that fills-in or chews back bases present as single-stranded overhangs to create blunt-ended molecules onto which the sequencing adapters can be ligated. However, the blunt-end ligation of the two sequencing adapters is non-directional, which means that either of the two adapters can be added to each end of the molecule. Because only molecules that have one of each adapter in the correct orientation can be sequenced, half of the molecules are lost due to incompatible adapter combinations. Additionally, MK requires three purification steps prior to amplification, all of which lose unique molecules. A more recently developed double-stranded DNA library preparation protocol, BEST (Carøe et al. 2017) (Figure 2.1A), is performed in a single tube, using heat denaturation rather

than purification between reaction steps. BEST has been shown to yield higher complexity libraries compared to other double-stranded library protocols (Carøe et al. 2017), which may be partly explained by the reduction in unique molecule loss from limiting the number of purification steps. However, like MK, BEST also uses blunt-end repair and non-directional ligation scheme to add the sequencing adapters to double-stranded DNA molecules.

Some ancient-DNA specific library preparation approaches target and convert single-stranded DNA (ssDNA) rather than dsDNA (Gansauge and Meyer 2013; Gansauge et al. 2017). Single-stranded library preparation methods begin with a denaturation step in which all DNA molecules in the extract are converted to single-stranded form. This allows conversion of DNA that is preserved in a single-stranded state as well as separate conversion of both strands of DNA preserved in a double-stranded state. When working with degraded DNA, ssDNA library preparation methods are more efficient, converting more DNA fragments into adapter-ligated form, compared to double-stranded approaches (Bennett et al. 2014; Wales et al. 2015; Gansauge et al. 2017). Additionally, some ssDNA library preparation methods leave the ends of DNA molecules unaltered, which makes it possible to explore patterns of stranded DNA fragmentation in aDNA extracts (Bokelmann et al. 2020).

Although ssDNA library preparation approaches improve DNA library conversion compared to dsDNA library preparation approaches, ssDNA approaches have yet to be widely adopted in ancient DNA research, mainly because of their higher cost and longer protocol duration compared to double-stranded approaches. For

example, the first ssDNA library preparation approach introduced for ancient DNA (Gansauge and Meyer 2013) required CircLigase (Lucigen), a single-stranded DNA ligase that is both expensive and difficult to obtain, and the protocol required two days to complete. A revised approach, ssDNA2.0 (Gansauge et al. 2017) (Figure 2.1B), reduced the expense and protocol duration by replacing the single-stranded ligation step with splinted ligation in which a double-stranded ligation junction is created via hybridization of a double-stranded adapter with a single-stranded degenerate overhang (Kwok et al. 2013). This made it possible to use the widely available and inexpensive T4 DNA ligase rather than CircLigase. While ssDNA2.0 is simpler to implement than the original version, it still requires four enzymatic steps and three clean-up steps, the latter of which creates opportunities for loss of unique molecules.

We present the Santa Cruz Reaction, or SCR, a fast and inexpensive single-reaction single-stranded DNA library preparation approach that we optimized for ancient DNA (Figure 2.1C). The SCR is an ancient DNA-specific version of the approach presented by Troll et al. (Troll et al. 2019), in which different enzymatic concentrations, a distinct hybridization strategy, and the use of a dilution series facilitates high-throughput processing of degraded samples. The SCR uses splinted adapters to simultaneously ligate both of the Illumina sequencing adapters in the correct orientation. Because we combine all steps into a single enzymatic reaction, we avoid multiple clean-up steps associated with the loss of unique molecules. To demonstrate the efficacy of the SCR in converting damaged DNA molecules, we use DNA extracted from five ancient specimens and prepare libraries using the SCR, BEST, and

ssDNA2.0. The SCR converts more molecules than BEST and performs with similar efficiency compared to ssDNA2.0 despite its relative simplicity.

2.3 Materials and Methods

2.3.1 DNA extraction

To compare the efficacy of the SCR to other commonly used library preparation approaches in ancient DNA, we prepared DNA extracts from five previously characterized ancient bones (four bison and one horse) that varied in DNA concentration, average fragment length, and deamination frequency (Table 2.1). We powdered each bone using a MM 400 ball mill (Retsch) and performed four extractions, each with 100-120mg of bone powder, from each sample following the silica column based method described in Dabney et. al (Dabney, Knapp, et al. 2013). We eluted DNA from the column using 50μL of EBT buffer (10mM Tris-HCl, 0.05% Tween-20) and pooled the four extracts from each sample into a single tube. We then quantified the DNA extraction pools with a Qubit 1X dsDNA HS Assay Kit (Invitrogen) using 5μL of DNA extract and a Qubit 4 Fluorometer (Invitrogen). Using these data, we calculated pmols/μL of dsDNA in each pooled extract using an estimated average length of 90bp for all samples, and pmols/μL of ssDNA or dsDNA ends by multiplying the dsDNA pmol/μL value by 2.

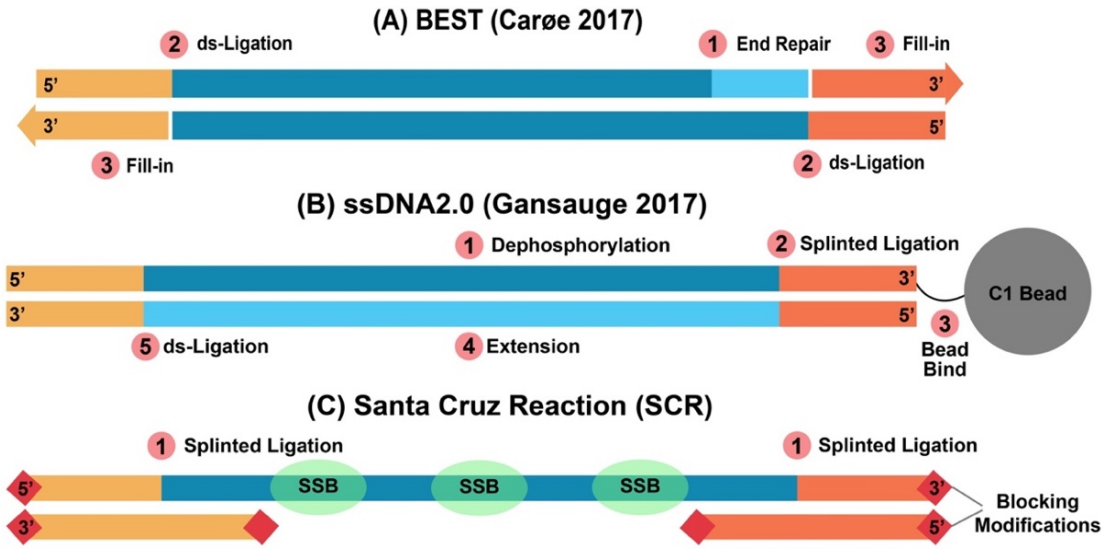

Figure 2.1: Key steps of the three library preparation methods compared here. **A.** The BEST protocol begins with an input DNA end-repair step where 3' overhangs are digested and 5' overhangs are filled in (1). Then**,** double-stranded adapters are ligated to the 5' ends of the input DNA (2), followed by adapter fill-in with a polymerase extension step, which initiates at the 3' nick present among adapter ligated input DNA molecules (3). **B.** ssDNA2.0 begins with the dephosphorylation of the input DNA (1), then a biotinylated adapter is ligated via splinted ligation to the 3' end of the input DNA (2), which is then bound to streptavidin beads (3). After annealing an extension primer, a second strand is synthesized (4), and then the 5' end of a double-stranded adapter is ligated to the 3' end of the synthesized strand (5). **C.** The Santa Cruz Reaction simultaneously ligates Illumina's P5 and P7 adapter using splinted ligation (1).

2.3.2 Library Preparation

We prepared libraries using the SCR, BEST, and ssDNA2.0 library preparation protocols (Figure 2.1). To assess whether library performance varied by DNA input amounts, we first prepared SCR, BEST, and ssDNA2.0 libraries from the extraction PH158 using six different inputs: 1.00 pmols (29.70ng), 0.50 pmols (14.85ng), 0.25 pmols (7.43ng), 0.125 pmols (3.71ng), 0.063 pmols (1.85ng), and 0.032 pmols (0.93ng) of ssDNA or dsDNA ends. Next, we assessed library performance consistency among samples by preparing SCR, BEST, and ssDNA2.0 libraries from each of the four

remaining DNA extracts using an input of 0.125 pmols (3.71ng) of DNA. All final pre-amplified libraries were eluted in 50uL of EBT buffer.

Below, we describe briefly the three library preparation protocols. A detailed description of the SCR is provided as a Supplemental Protocol (see section 2.6).

2.3.3 BEST

The BEST protocol (Figure 2.1A) is a single-tube double-stranded DNA (dsDNA) library preparation protocol optimized for ancient DNA. We prepared BEST libraries as outlined in Carøe et al. (Carøe et al. 2017) with the modifications described in Mak et al. (Mak et al. 2017), using a 25:1 adapter:template ratio. We used a MinElute column for the final clean-up prior to amplification.

Briefly, BEST libraries are prepared by first performing an end-repair reaction with T4 Polynucleotide Kinase (NEB) and T4 DNA Polymerase (NEB) which blunt-ends the input DNA. Following end-repair, a blunt-end ligation reaction is performed using T4 DNA Ligase (NEB) which facilitates the ligation of the 5′ ends of template molecules to the 3′ ends of blunt-end adapters. Then, an adapter fill-in reaction is performed with Bst 2.0 DNA Polymerase (NEB), which initiates at the ligation junction nick present at the 3′ ends of the template and 5′ ends of the non-ligated adapter strand. Heat inactivation of enzymes occurs between reactions, but a MinElute (Qiagen, Hilden, Germany) column clean-up step is performed following the fill-in reaction.

The BEST protocol flanks the native input DNA molecules with adapters, which will include uracil bases. A uracil-tolerant polymerase must therefore be used during library amplification.

2.3.4 ssDNA2.0

SsDNA2.0 (Figure 2.1B) is a single-stranded library preparation method optimized for damaged and degraded DNA and is the current state-of-the-art ssDNA method for highly degraded samples. We prepared ssDNA2.0 libraries as described in Gansauge et al. (Gansauge et al. 2017) using the TL136 splinter oligo (Table 2.2).

Briefly, ssDNA2.0 libraries are prepared by first dephosphorylating the 5' and 3' termini of the input DNA with FastAP (Thermo Scientific). The DNA is then heat denatured at 95˚C for 1 minute and then rapidly cooled in an ice bath. Once cooled, a biotinylated splinted adapter is ligated to the 3' end of input DNA using T4 DNA Ligase (Thermo Scientific). The biotinylated adapters, including the ligation products, are then immobilized on C1 beads (Invitrogen), pulled down, and washed. An extension primer is then annealed to the ligated adapter and a second strand is synthesized using the Klenow Fragment (Thermo Scientific), followed by a second C1 bead pull-down and wash step. T4 DNA Ligase (Thermo Scientific) is then used to ligate a double-stranded blunt-end adapter to the 3' end of the synthesized strand, followed by a third C1 bead pull-down and wash step. Finally, the reactions are heat denatured and the pre-amplified library is collected with the supernatant.

The final adapter-flanked product of an ssDNA2.0 library is the synthesized second strand. Because this will not contain uracil bases, a high fidelity and/or non-uracil tolerant polymerase can be used during library amplification. However, non-standard Illumina oligonucleotide design differences lead to a truncated P5 adapter, which requires use of a non-standard Illumina sequencing primer.

2.3.5 The Santa Cruz Reaction

The Santa Cruz Reaction (SCR; Figure 2.1C) uses directional splinted ligation of Illumina's P5 and P7 adapters to convert natively single-stranded DNA and heat denatured double-stranded DNA into Illumina libraries in one enzymatic reaction. Similar to other library preparation protocols, including BEST and NEB Ultra II, the SCR scales the concentration of reaction components to the amount of input DNA to reduce the proportion of adapter-dimers. In the case of the SCR, that includes Extreme Thermostable Single-Stranded Binding Proteins (ET SSB, NEB), which scales with the amount of single-stranded DNA in the reaction. We recommend preparing several splinted adapter and ET SSB dilutions to be used for specific ranges of input DNA (see Supplemental Protocol, 2.6).

The SCR begins by combining 20µL of a DNA extract with 2µL ET SSB (NEB) at a dilution optimized for the amount of input DNA (see Supplementary Protocol, 2.6) to create a sample mixture. The sample mixture is then denatured by heating to 95˚C for three minutes, followed by rapid cooling in an ice bath. Next, 1µL each of P5 and P7 splinted adapters (also at dilutions optimized for the amount of input DNA; see Supplemental Protocol, 2.6) are added to the sample mix. Finally, 26µL of SCR master mix containing 3.75µL SCR Buffer (666mM Tris-HCl, 132mM MgCl2), 0.5µL 100mM ATP (Thermo Scientific), 0.5µL 1M DTT (Thermo Scientific), 0.625µL 2,000,000 U/mL T4 DNA Ligase (NEB), 0.625µL 10,000 U/mL T4 Polynucleotide Kinase (NEB), and 20µL 50% PEG 8000 (NEB) is added to the sample mixture, creating a 50µL reaction. The reaction is pulse-vortexed for 30 seconds, incubated at

37˚C for 45 minutes, and then cleaned with a MinElute column following the manufacturer's instructions.

Because the SCR ligates adapters directly to the native input molecules, an uracil tolerant polymerase must be used during library amplification.

The SCR is an ancient DNA-specific version of SRSLY, which was described by Troll et al. (Troll et al. 2019). Several alterations make the SCR more appropriate than SRSLY for converting damaged DNA. For example, the SCR uses DTT and ATP in place of T4 DNA Ligase Buffer (NEB), which appears to better stimulate T4 PNK. Because adapter-dimers are problematic when working with degraded and low-input samples, the SCR also recommends a series of splinted adapter and ET SSB dilutions for lower DNA input volumes, and implements an asymmetric P5:P7 adapter molar ratio that reduces adapter-dimer formation. Finally, like ssDNA 2.0, the SCR adapter hybridization strategy uses a molar excess of splints to reduce the chance of splintless adapters in the reaction (see Supplementary Protocol, 2.6).

2.3.6 Quantitation, indexing, and sequencing

We quantified the amount of library molecules in each library by performing quantitative PCR (qPCR) on a 1:50 dilution of each library using the primers IS7 and IS8 (Gansauge and Meyer 2013), which amplify adapter-ligated templates. We then prepared a 25µL qPCR for each library using 1µL of diluted library, 12.5µL 2X Maxima SYBR Green Master Mix (Thermo Scientific), 10.5µL H2O, 0.5µL 10µM IS7 primer, and 0.5uL 10µM IS8 primer. Reactions were cycled with the following conditions: 95˚C for 10 minutes, followed by 40 cycles of 95˚C for 30 seconds, 60˚C

for 30 seconds, and 72˚C for 30 seconds. Fluorescence was measured at the end of each extension step.

We then performed library amplification and double indexing using the indexing primers described in Kircher et al (Kircher et al. 2012). For each library we prepared a 100µL PCR using 2µL undiluted library, 50µL Amplitaq Gold 360 Master Mix (Applied Biosystems), 1µL unique 100µM i7 indexing primer, 1µL unique 100µM i5 indexing primer, and 46µL H2O. We amplified each library with the following cycling conditions: 95˚C for 10 minutes, followed by a library specific number of cycles of 95˚C for 30 seconds, 60˚C for 30 seconds, and 72˚C for 60 seconds, and a final extension of 72˚C for 7 minutes. We inferred the optimal cycle number for each library from the qPCR results (Table 2.3). We used each library's CT value, rounded to the nearest cycle, to determine the optimal number of cycles for indexing PCR.

We purified the amplified libraries using 120µL (1.2x) of a SPRI bead mixture, which we prepared according to and performed as described in Rohland and Reich (Rohland and Reich 2012). We quantified the purified libraries with the Qubit 1X dsDNA HS Assay Kit and Qubit 4, and visualized the products using a D1000 ScreenTape (Agilent) and Tapestation 2200 (Agilent).

We sequenced each library at the University of California, Santa Cruz Ancient and Degraded DNA Processing Center using 150 cycle mid output kits on an Illumina NextSeq 550. Because they needed a non-standard primer, we sequenced libraries prepared with ssDNA2.0 on separate runs with a complete replacement of Illumina's read 1 sequencing primer with the oligo CL72, as described in Gansauge et al.

(Gansauge and Meyer 2013). We performed base calling using Illumina's bcl2fastq2 software.

2.3.7 Data Analysis

To compare the performance of the three library preparation protocols, we downsampled fastq files from each library to the number of reads generated from the least deeply sequenced library per library preparation approach. We merged reads that overlapped by at least 15 bases, trimmed adapters, and removed reads that were under 30 bp long using SeqPrep (https://github.com/jstjohn/SeqPrep). We then mapped merged and unmerged reads separately using Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) v0.7.12 aln algorithm with seed disabled to either the equCab2 (Wade et al. 2009) or bison_umd1.0 (GCA_000754665.1) reference genomes, depending on whether the sample was a horse or a bison. We collapsed PCR duplicates and generated mapping summary statistics using SAMtools (Li et al. 2009). We calculated read lengths directly from the merged reads and, for unmerged reads, inferred total lengths based on mapping coordinates. We computed cytosine deamination frequencies using mapDamage2.0 (Jónsson et al. 2013).

2.4 Results

2.4.1 Library Conversion Efficiency

We used qPCR and the proportion of unique mapped reads to estimate library conversion efficiency. As the number of amplifiable molecules in a library increases, fewer cycles are necessary for the library to reach a detection threshold during qPCR. The cycle threshold (CT value) is the cycle number at which a library reaches this detection threshold. A library that is detected one CT value sooner than another has

approximately twice the number of amplifiable starting molecules than the later-detected library.

When comparing CT values for libraries from PH158 prepared using six different DNA input volumes, both ssDNA library preparation methods converted more molecules compared to the double-stranded approach (Figure 2.2A). The ssDNA library approaches performed similarly, with SCR recovering more molecules among the higher input libraries and ssDNA2.0 recovering more molecules among the lower input libraries. At the highest DNA input, 29.70ng or 1.00pmol ssDNA, the SCR library reached the detection threshold 2.8 cycles earlier than ssDNA2.0, suggesting that 7.0X more DNA was converted. At increasingly lower DNA inputs, the CT value difference between SCR and ssDNA2.0 decreased. At the lowest DNA input, 0.93ng or 0.037pmol ssDNA, the ssDNA2.0 library reached the detection threshold 0.4 cycles earlier than SCR, suggesting that ssDNA2.0 converted 1.3X more DNA than the SCR.

Because qPCR cannot discriminate between adapter-dimers and adapter-flanked molecules, we next compared, as a measure of library complexity, the proportion of reads that mapped to the reference genome that are duplicates (1 – (# unique mapped reads) / (# total mapped reads)). This allows us to distinguish libraries that convert more unique molecules as those that have a lower proportion of mapped duplicated reads. After down sampling each library to equal numbers of reads, we observed a trend similar to that from qPCR in which the SCR and ssDNA2.0 libraries contain a lower proportion of mapped duplicates (and therefore a higher proportion of unique reads) compared to the BEST libraries (Figure 2.2B). While the two ssDNA

library preparation approaches performed similarly to each other, we observed some differences with DNA input. The SCR libraries contained a lower proportion of mapped duplicates compared to the ssDNA2.0 libraries at the higher DNA inputs, while ssDNA2.0 libraries contained a lower proportion of mapped duplicates compared to SCR at low (1.86ng and 0.93ng) inputs. Because the two ssDNA library preparation protocols have similar conversion efficiency at higher DNA inputs, we selected a higher DNA input for the remaining library comparisons.

When comparing CT values given a static 3.71ng DNA input across the five DNA extracts, we observed similar trends between library preparation approaches to those reported above. The two single-stranded approaches reach the detection threshold before BEST (Figure 2.2C) across all five extracts. The SCR libraries reached the detection threshold between 0.1 and 1.5 cycles before the ssDNA2.0 libraries, suggesting that SCR converted 1.1X - 2.8X more molecules at this input volume.

After down sampling each library to an equal number of reads, we observed the two single-stranded approaches contained a lower proportion of mapped duplicate reads than the double-stranded approach for all five extracts (Figure 2.2D). The single-stranded approach that produced the lowest proportion of duplicates varied by extract, suggesting that the two ssDNA approaches are similarly efficient at this DNA input (3.71ng). While the SCR produced libraries with a lower proportion of mapped duplicates in three of five extracts, qPCR suggested that SCR converts more DNA to library compared to ssDNA2.0 across all five extracts. The discrepancy between the

qPCR and sequencing results is most likely due to the higher proportion of adapter-dimers in the SCR libraries compared to the ssDNA2.0 libraries.

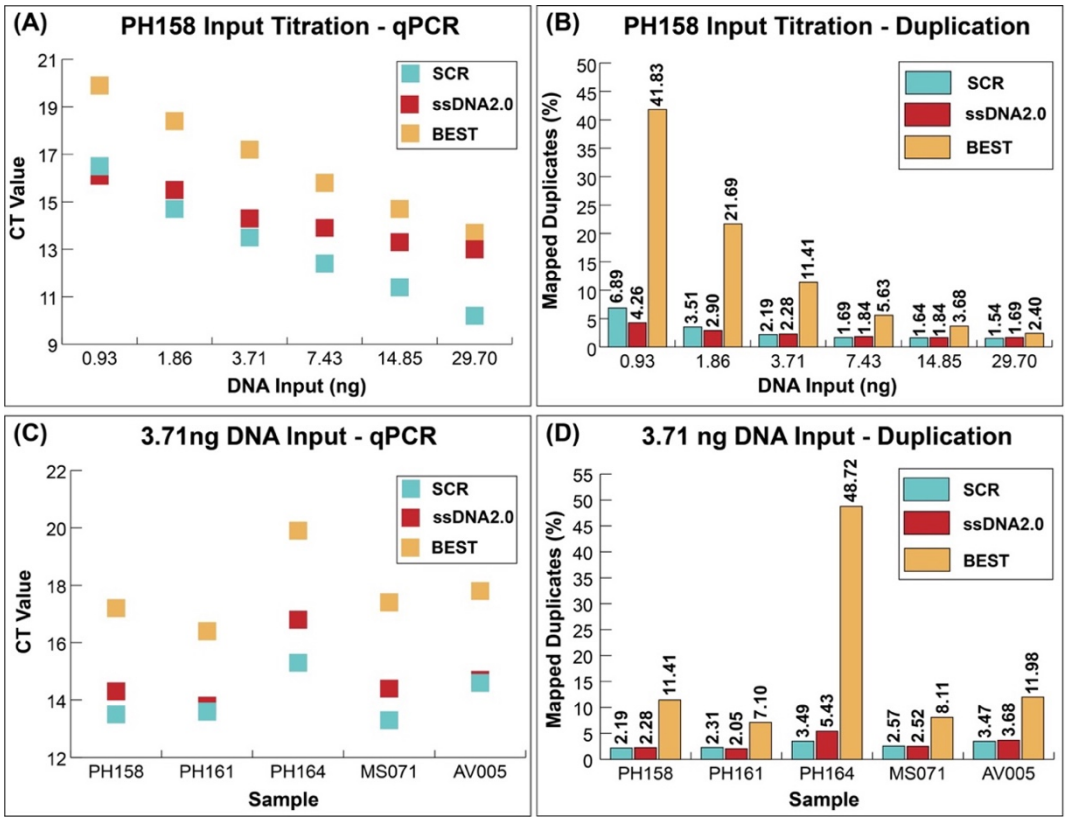

Figure 2.2: Library preparation complexity comparison. **(A)** Quantitative PCR CT values for libraries prepared from sample PH158 using a titration of six DNA inputs ranging from 0.93ng – 29.70ng. Lower CT values indicate more starting library molecules in the reaction **(B)** Proportion of mapped reads that are duplicates prepared from sample PH158 at the different titrations of DNA input. **(C)** Quantitative PCR CT values for libraries prepared from five ancient DNA extracts using a static DNA input of 3.71ng. **(D)** Proportion of duplicated reads from libraries prepared from these five ancient DNA extracts using the static 3.71ng DNA input.

2.4.2 Endogenous content, average fragment length, and terminal deamination frequency

Next, we compared the endogenous DNA content, the proportion of DNA that mapped to the relevant reference genome, in each library. The two ssDNA library

approaches recovered either more or a similar proportion of endogenous DNA compared to the dsDNA library approach for all five extracts (Figure 2.3A). SsDNA2.0 recovered the highest proportion of endogenous DNA for all extracts, and SCR recovered between 72.8% and 93.1% of that recovered by ssDNA2.0.

The two ssDNA methods produced libraries with similar average fragment lengths and, for most samples, shorter average fragment lengths compared to BEST. The average fragment length difference between the SCR and ssDNA2.0 libraries ranged from 0.18bp to 3.52bp (Figure 2.3B), and both approaches resulted in a similar fragment length distributions (Figure 2.4). The BEST libraries had a noticeably higher average fragment length compared to SCR when the extracted DNA was less fragmented. However, BEST libraries produced from the two most heavily fragmented samples, AV005 and MS071, had a similar average fragment length compared to the SCR.

Libraries prepared with ssDNA2.0 have a consistently higher frequency of terminal deamination on both the 5’ and 3’ ends compared to the SCR and BEST libraries (Figure 2.3C). We also observed a terminal deamination asymmetry in nearly all libraries in which the 5’ end contains a higher rate of deamination compared to the 3’ end. Libraries prepared with the SCR exhibit the highest deamination asymmetry in all but AV005, which also has the shortest average fragment length and is likely to be the most degraded of the five extracts.

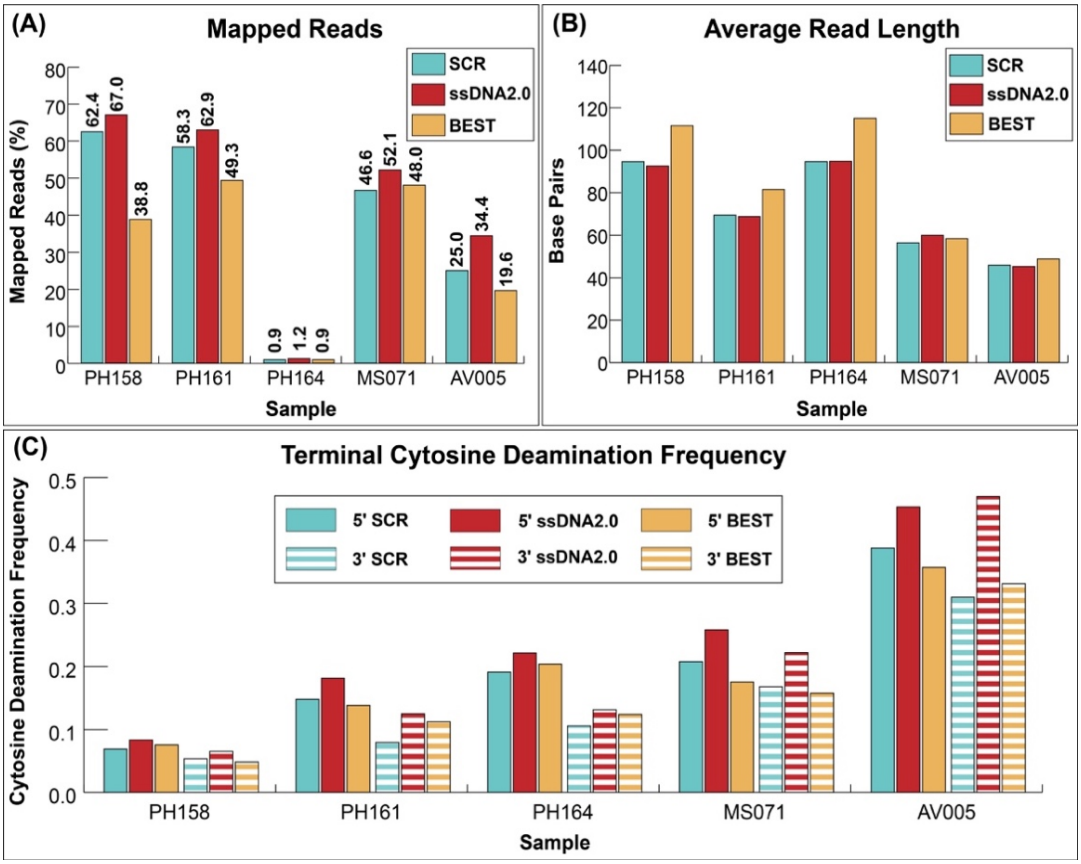

Figure 2.3. Sequencing statistics for libraries prepared from five samples using 3.71ng of input DNA. **(A)** The percentage of reads mapped to the reference genome. **(B)** The average length of all mapped reads. **(C)** The terminal 5' and 3' cytosine deamination frequencies of the mapped reads.

2.5 Discussion

Ancient DNA research often involves screening tens to hundreds of samples for preserved DNA at the outset of a research project. For most samples, it is impossible to know whether the sample will be sufficiently well-preserved to generate genome-scale data without extracting DNA, preparing that DNA extract into a sequenceable library, and sequencing that library. Although single-stranded library preparation approaches are understood to be more efficient in converting ancient DNA molecules

into sequenceable form than are double-stranded library preparation approaches (Bennett et al. 2014; Wales et al. 2015; Gansauge et al. 2017), ssDNA library preparation approaches have yet to be widely adopted in ancient DNA research labs because they are more expensive and take considerably longer to complete compared to dsDNA library preparation methods. The Santa Cruz Reaction solves this problem by converting ancient DNA molecules into sequencing libraries with efficiency that is comparable to the current state-of-the-art ssDNA approach, ssDNA2.0. Compared to ssDNA2.0, however, SCR reduces ssDNA library preparation to a single cost-effective enzymatic step that focuses on the primary goal of fragmented DNA library preparation, adapter ligation. Reducing the number of protocol steps reduces the duration of the pre-amplification protocol by 2.5 times compared to ssDNA2.0. We also note that time to completion can be further reduced by replacing the column-based cleanup used here with a magnetic bead clean-up.

We highlight several challenges associated with the SCR protocol. First, because we add four oligonucleotides to a single reaction, including a phosphorylated adapter, adapter-dimers form more readily compared to ssDNA2.0 and BEST (Table 2.3). To reduce the proportion of adapter-dimers in the final library, titration of adapters to input DNA is beneficial. We have developed an adapter dilution series, where five adapter concentrations are used across specific ranges of DNA inputs. Second, batches of synthesized splint oligonucleotides often include synthesis artifacts that render DNA fragments capable of ligation at ends that should be blocked for ligation (Figure 2.5). While this issue was noted previously along with an oligonucleotide purification

strategy (Gansauge et al. 2017; Gansauge et al. 2020), we were not able to successfully adopt an artifact removal scheme from our synthesized splint oligonucleotides. Instead, we implemented several oligonucleotide usage optimizations such as a quality control procedure (see Supplemental Protocol, 2.6), which allows users to identify poor quality splint batches prior to use. Furthermore, we have optimized the adapter:splint hybridization ratio and use asymmetrical P5:P7 adapter concentrations in the reaction, this reduces the most detrimentally volatile reaction component, the P7 splint, without hindering library preparation performance. Future oligo design improvements may allow for further streamlining of SCR reagent preparation and usage.

In agreement with previous studies (Bennett et al. 2014; Wales et al. 2015; Gansauge et al. 2017), we found the ssDNA library preparation methods convert more molecules to library compared to the dsDNA method across all DNA input amounts and the five ancient DNA extracts used here. The differences in conversion efficiencies between the two single-stranded methods are more nuanced. Both qPCR and sequencing results from the input titration experiment suggest the SCR and ssDNA2.0 outperform each other at opposite ends of the DNA input spectrum, with ssDNA2.0 outperforming SCR at the lowest input amounts. The SCR's lower library conversion efficiency at the lowest input amounts is likely due to too low P5 and P7 adapter concentrations in the reaction at lowest adapter dilution tier. The scaling of the adapters with the amount of input DNA is a challenge and, at present, a necessity. We note that the higher proportion of adapter-dimers in the SCR libraries may lead to an inflation of qPCR-based estimates of converted DNA. This could be explored further by

sequencing each library to exhaustion, however the high complexity of most ssDNA libraries made this impractical.

Interestingly, the two ssDNA library methods appear to convert slightly different populations of input molecules to the final library. In particular, ssDNA2.0 libraries consistently have higher terminal deamination frequencies compared to the SCR libraries (Figure 2.6). This may indicate that ssDNA2.0 is better able to convert and retain molecules containing a terminal uracil, which may partly explain the higher endogenous content of ssDNA2.0 libraries compared to SCR. The differences in terminal deamination frequency may be driven by ligation scheme, in which the splinted adapter targeting the 3' end during the SCR is in approximately 6X molar excess compared to the input DNA and in approximately 80X molar excess in ssDNA2.0. Splint species that are highly reactive to uracil containing termini, such as those with an adenine at the ligation junction, may become limiting when splint molar excess is low. This hypothesis could be tested by altering the base composition of the splint bases near the ligation junction to contain higher adenine content. We also observed that the mapped reads from the SCR libraries have an average GC content that more closely reflects the reference genome compared to libraries prepared with ssDNA2.0 (Figure 2.7). This may be caused by polymerase GC bias during the second strand synthesis of ssDNA2.0.

We present a protocol for fast and simple DNA library preparation that can recover degraded molecules preserved in both single-stranded and double-stranded form. Although ssDNA2.0 outperforms the SCR at the lowest input volumes and may

be more appropriate for the most degraded samples, the SCR performs as well as ssDNA2.0 across a wide range of input volumes and is an appropriate and more efficient replacement for commonly used dsDNA library preparation approaches.

2.6 Supplementary Protocol

Oligonucleotides (IDT Format, 5' → 3')

Note: All oligonucleotides are ordered with HPLC purification conducted by IDT.
Note: We recommended ordering two separate batches of each splint oligonucleotide. See oligonucleotide quality control recommendations in section 11 of this protocol.

scr_P5_adapter: /5AmMC12/ACACTCTTTCCCTACACGACGCTCTTCCGATCT
scr_P7_adapter: /5Phos/AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC/3AmMO/
scr_P5_splint:
/5AmMC6/NNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT/3AmMO/
scr_P7_splint:
/5AmMC12/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNN/3AmMO/

Reaction Reagents

Reagent	Provider	Cat #
T4 DNA Ligase (2,000,000 U/mL)	NEB	M0202M
T4 PNK (10,000 U/mL)	NEB	M0201L
ET SSB (500 ng/µL)	NEB	M2401S
ATP (100 mM)	Thermo Scientific	R0441
DTT (1 M)	Thermo Scientific	P2325
T4 RNA Ligase Buffer	NEB	B0216L
PEG 8000 (50%)	NEB	B0216L
Tris-HCl pH 8.0 (1M)	Invitrogen	15568025
EDTA pH 8.0 (0.5M)	Invitrogen	15575020
Tween-20 (10%)	Teknova	T0710
NaCl (5M)	Sigma-Aldrich	S5150-1L
Glycerol (50%)	Invitrogen	15514011
$MgCl_2$ (1M)	Invitrogen	AM9530G

Buffer Preparation

- **TE Buffer (Store at RT):** 10 mM Tris-HCl (pH 8.0), 1 mM EDTA (pH 8.0)
- **EBT Buffer (Store at RT):** 10 mM Tris-HCl (pH 8.0), 0.05% Tween-20
- **Adapter Dilution Buffer (Store at -20˚C):** 1X T4 RNA Ligase Buffer, 0.05% Tween-20
- **SSB Dilution Buffer (Store at -20˚C):** 20 mM Tris-HCl (pH 8.0), 20 mM NaCl, 0.5 mM DTT, 0.1 mM EDTA (pH 8.0), 50% Glycerol
- **SCR Buffer (Store at -20 °C):** 666 mM Tris-HCl, 132 mM $MgCl_2$

Adapter-Splint Hybridizations (Store hybridized stocks at -20 °C)

1. Resuspend all oligonucleotides to 100 µM using TE buffer.

2. Add the following components to a 0.2 mL tube labeled P5:
   - **H_2O:** 30.6 µL
   - **10X T4 RNA Ligase Buffer:** 5 µL
   - **100 µM scr_P5_adapter:** 6 µL
   - **100 µM scr_P5_splint:** 8.4 µL

3. Add the following components to a 0.2 mL tube labeled P7:
   - **H_2O:** 30.6 µL
   - **10X T4 RNA Ligase Buffer:** 5 µL
   - **100 µM scr_P7_adapter:** 6 µL
   - **100 µM scr_P7_splint:** 8.4 µL

4. In a thermocycler with heated lid (105 °C**)**, hybridize the adapters and splints by incubating at 95 °C for 1 minute before ramping down to 10 °C at 0.1 °C per second.

5. Store hybridized P5 (12 µM) and P7 (12 µM) adapter stock solutions at -20 °C.

Adapter-Splint Dilutions (Store dilutions at -20 °C)

Note: Freeze-thaw adapter dilutions no more than five times.

1. Prepare working **P5** adapter solutions by diluting hybridized 12 µM stock solutions to the following molarities. Dilute with Adapter Dilution Buffer.
   - **6.0 µM**
   - **3.0 µM**
   - **1.5 µM**
   - **0.75 µM**

2. Prepare working **P7** adapter solutions by diluting hybridized 12 µM stock solutions to the following molarities. Dilute with Adapter Dilution Buffer.
   - **6.0 µM**
   - **3.0 µM**
   - **1.5 µM**
   - **0.75 µM**
   - **0.375 µM**

S6. ET SSB Dilutions (Store dilutions at -20 °C)

1. Prepare the following **ET SSB** dilutions from **500 ng/µL** stock solution using SSB Dilution Buffer:
   - **328 ng/µL**
   - **164 ng/µL**
   - **82 ng/µL**
   - **41 ng/µL**
   - **20.5 ng/µL**

Reaction Mix Preparation (Store reaction mix at -20 °C)

Notes:

- Freeze-thaw the Reaction Mix no more than five times.
- Warm PEG 8000 to 50 °C before pipetting into the Reaction Mix tube as the first component.
- Equilibrate PEG 8000 to RT before adding remaining reagents.
- Thoroughly mixing the Reaction Mix is essential, vortexing is recommended.

Reagent	Stock Conc.	1 RXN (µL)	Final Conc.	20 RXNs (µL)
PEG 8000	50%	20	20%	400
SCR Buffer	13.3X	3.75	1X	75
DTT	1M	0.5	10mM	10
ATP	100mM	0.5	1mM	10
T4 PNK	10,000 U/mL	0.625	0.125 U/mL	12.5
T4 DNA Ligase	2,000,000 U/mL	0.625	25 U/mL	12.5

Reaction Adapter-Splint and SSB Input Tiers

picomoles ssDNA	dsDNA* (ng)	P5 (µM)	P7 (µM)	SSB (ng/µL)
1.0 - 2.5	29 - 75	12	6	328
0.5 - 0.99	15 - 29	6	3	164
0.25 - 0.49	7 - 15	3	1.5	82
0.12 - 0.24	3 – 7	1.5	0.75	41
< 0.12	< 3	0.75	0.375	20.5

***Approximate degraded dsDNA input. The average size used to calculate the number of picomoles is deliberately overestimated due to the difficulties of accurately visualizing highly degraded samples.**

The Santa Cruz Reaction (SCR) Workflow

Note: The SCR uses a tiered adapter and SSB system based on the DNA input into the reaction. Use the table in **Section 8** to select the correct dilution set for each DNA extract.

Critical Note: Thoroughly mixing the reaction in steps seven and eight is essential to consistently achieve high ligation efficiency. Inadequately mixing the reaction is the most common failure mode. Vortexing is recommended.

1. Gather the following reagents and equipment:
   - Thaw and equilibrate the Reaction Mix Preparation and appropriate SSB dilution to room temperature before pipetting.
   - Thaw the appropriate P5 and P7 adapter-splint dilutions and place them on ice.
   - Prepare an ice bath.

2. Prepare the Sample Mix by combining the following in a 0.2 mL PCR tube, 8-tube strip, or 96-well plate:

- **DNA Extract:** 20 µL (if less than 20 µL, fill to 20 µL with buffer EBT)
- **Appropriate SSB Dilution:** 2 µL

3. Pulse-vortex the Sample Mix five times at maximum speed prior to briefly spinning down in a mini centrifuge.

4. Incubate the Sample Mix at 95 °C for 3 minutes before immediately placing the Sample Mix into an ice bath.

5. Allow the Sample Mix to cool on ice for 2 minutes, spin down briefly in a mini centrifuge, and return the Sample Mix to the ice bath.

6. While on ice, add the following to the cooled Sample Mix:
   - **Appropriate P5 Adapter Dilution:** 1 µL
   - **Appropriate P7 Adapter Dilution:** 1 µL
   - **Reaction Mix Preparation:** 26 µL

7. Pulse-vortex the reaction five times at maximum speed and then spin down briefly in a mini centrifuge.

8. Repeat step 7 two more times.

9. Incubate the reaction at 37 °C for 45 minutes.

10. Proceed directly to reaction clean-up:
    - To retain the shortest molecules, use a MinElute PCR Purification Kit (Qiagen Catalog No. 28004) following the manufacturer's instructions. Elute in 50 µL buffer EBT.

11. The purified reaction products are ready for indexing and amplification by PCR.
    - **Note:** Use of a uracil-tolerant polymerase is required when working with degraded DNA as a starting material. Examples include: AmpliTaq Gold (ThermoFisher), Pfu Turbo cx (Agilent), KAPA Uracil+ (Roche), and Q5U (NEB).

12. A 1.2X SPRI purification is the recommended post-PCR purification strategy.

Oligonucleotide Quality Control - Workflow

Note: The oligonucleotides for the Santa Cruz Reaction are designed with blocking modifications to limit undesirable ligation products. However, the splint oligonucleotides may arrive with one or more subspecies containing unblocked termini. Currently, a consistent purification strategy to eliminate poor quality splint

species does not exist. However, this section presents a ligation based method to identify poor quality splint batches.

Briefly, splint oligonucleotides are spiked into Santa Cruz Reactions. Ligatable and amplifiable species within the splint spike will convert to library molecules, which will be identifiable during post amplification visualization. A splint batch with a high proportion of ligatable and amplifiable species should not be used for the library preparation of ancient samples (see section 11 for trace interpretation guidance).

Each new batch of P5 and P7 splint should undergo the quality control scheme below.

1. Follow the adapter hybridization, reagent preparation, and dilution guidelines in Sections 3-7.

2. Prepare a number of Santa Cruz Reactions equal to the number of freshly synthesized splint batches requiring quality control plus one blank
- **Example**: 1 P5 Splint + 1 P7 Splint + 1 Blank = 3 reactions.

3. In a 0.2 mL PCR 8-tube strip add the following to wells 1-3.
- **Well 1:** 1 pmol P5 splint and fill to 20 µL (5 µL of 0.2 µM P5 splint + 15 µL EBT)
- **Well 2:** 1 pmol P7 splint and fill to 20 µL (5 µL of 0.2 µM P7 splint + 15 µL EBT)
- **Well 3:** 20 µL EBT

4. Add 2 µL of the 20.5 ng/µL SSB dilution to each well.

5. Follow steps 3-5 of the SCR workflow in section eight to mix, heat denature, and chill the reactions.

6. Add the following components to each chilled reaction:
- **0.75 µM P5 Hybridized Adapter Dilution:** 1 µL
- **0.375 µM P7 Hybridized Adapter Dilution:** 1 µL
- **Reaction Mix:** 26 µL

7. Follow steps 7-9 of the SCR workflow in section eight to properly mix and incubate the reactions.

8. Following incubation at 37 °C, clean each reaction using the MinElute PCR Purification Kit according to the manufacturer's instructions. Elute in 50 µL buffer EBT.

9. Amplify and index the entire eluate from step 8 for 18 cycles using the PCR scheme of your choice.

10. Clean each amplified library with the MinElute PCR Purification Kit according to the manufacturer's instructions.

11. Visualize each library, including the negative control, on a Fragment Analyzer, TapeStation, or BioAnalyzer automated electrophoresis system.

2.7 References

Bennett EA, Massilani D, Lizzo G, Daligault J, Geigl EM, Grange T. 2014. Library construction for ancient genomics: single strand or double strand? Biotechniques. 56:289–90, 292.

Bokelmann L, Glocke I, Meyer M. 2020. Reconstructing double-stranded DNA fragments on a single-molecule level reveals patterns of degradation in ancient samples. Genome Res. 30:1449–1457.

Brace S, Thomas JA, Dalén L, Burger J, MacPhee RD, Barnes I, Turvey ST. 2016. Evolutionary history of the Nesophontidae, the last unplaced recent mammal family. Mol Biol Evol. 33:3095–3103.

Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, et al. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci U S A. 104:14616–14621.

Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. 2009. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 38:e87.

Bunce M, Szulkin M, Lerner HR, Barnes I, Shapiro B, Cooper A, Holdaway RN. 2005. Ancient DNA provides new insights into the evolutionary history of New Zealand's extinct giant eagle. PLoS Biol. 3:e9.

Bunce M, Worthy TH, Phillips MJ, Holdaway RN, Willerslev E, Haile J, Shapiro B, Scofield RP, Drummond A, Kamp PJ, et al. 2009. The evolutionary history of the extinct ratite moa and New Zealand Neogene paleogeography. Proc Natl Acad Sci U S A. 106:20646–20651.

Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding MHS, Samaniego JA, Wales N, Sicheritz-Pontén T, Gilbert MTP. 2017. Single-tube library preparation for degraded DNA. Methods Ecol Evol. 9:410–419.

Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, García N, Pääbo S, Arsuaga JL, et al. 2013a. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci U S A. 110:15758–15763.

Dabney J, Meyer M, Pääbo S. 2013b. Ancient DNA damage. Cold Spring Harb Perspect Biol. 5:1–7.

Gansauge MT, Aximu-Petri A, Nagel S, Meyer M. 2020. Manual and automated preparation of single-stranded DNA libraries for the sequencing of DNA from ancient biological remains and other sources of highly degraded DNA. Nat Protoc. 15:2279–2300.

Gansauge MT, Gerber T, Glocke I, Korlevic P, Lippik L, Nagel S, Riehl LM, Schmidt A, Meyer M. 2017. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. Nucleic Acids Res. 45:e79.

Gansauge MT, Meyer M. 2013. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc. 8:737–748.

Gansauge MT, Meyer M. 2014. Selective enrichment of damaged DNA molecules for ancient genome sequencing. Genome Res. 24:1543–1549.

Graham RW, Belmecheri S, Choy K, Culleton BJ, Davies LJ, Froese D, Heintzman PD, Hritz C, Kapp JD, Newsom LA, et al. 2016. Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. Proc Natl Acad Sci U S A. 113:9310–9314.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, et al. 2010. A draft sequence of the Neandertal genome. Science. 328:710–722.

Green EJ, Speller CF. 2017. Novel substrates as sources of ancient DNA: prospects and hurdles. Genes (Basel). 8:180.

Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Pääbo S. 2001. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res. 29:4793–4799.

Jónsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 29:1682–1684.

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

Kwok CK, Ding Y, Sherlock ME, Assmann SM, Bevilacqua PC. 2013. A hybridization-based approach for quantitative and low-bias single-stranded DNA ligation. Anal Biochem. 435:181–186.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 513:409–413.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25:2078–2079.

Lindahl T. 1993. Instability and decay of the primary structure of DNA. Nature. 362:709–715.

Lindahl T, Nyberg B. 1972. Rate of depurination of native deoxyribonucleic acid. Biochemistry. 11:3610–3618.

Lorenzen ED, Nogués-Bravo D, Orlando L, Weinstock J, Binladen J, Marske KA, Ugan A, Borregaard MK, Gilbert MT, Nielsen R, et al. 2011. Species-specific responses of Late Quaternary megafauna to climate and humans. Nature. 479:359–364.

Mak SST, Gopalakrishnan S, Carøe C, Geng C, Liu S, Sinding MS, Kuderna LFK, Zhang W, Fu S, Vieira FG, et al. 2017. Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. Gigascience. 6:1–13.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010:pdb.prot5448.

Meyer M, Palkopoulou E, Baleka S, Stiller M, Penkman KEH, Alt KW, Ishida Y, Mania D, Mallick S, Meijer T, et al. 2017. Palaeogenomes of Eurasian straight-tusked elephants challenge the current view of elephant evolution. Elife.

Mouttham N, Klunk J, Kuch M, Fourney R, Poinar H. 2015. Surveying the repair of ancient DNA from bones via high-throughput sequencing. Biotechniques. 59:19–25.

Orlando L, Calvignac S, Schnebelen C, Douady CJ, Godfrey LR, Hänni C. 2008. DNA from extinct giant lemurs links archaeolemurids to extant indriids. BMC Evol Biol. 8:121.

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, et al. 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 499:74–78.

Pedersen MW, Ruter A, Schweger C, Friebe H, Staff RA, Kjeldsen KK, Mendoza ML, Beaudoin AB, Zutter C, Larsen NK, et al. 2016. Postglacial viability and colonization in North America's ice-free corridor. Nature. 537:45–49.

Rohland N, Glocke I, Aximu-Petri A, Meyer M. 2018. Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. Nat Protoc. 13:2447–2461.

Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. 2015. Partial uracil – DNA – glycosylase treatment for screening of ancient DNA. Philos Trans R Soc B Biol Sci. 370:20130624 .

Rohland N, Hofreiter M. 2007. Ancient DNA extraction from bones and teeth. Nat Protoc. 2:1756–1762.

Rohland N, Reich D. 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res. 22:939–946.

Shapiro B, Drummond AJ, Rambaut A, Wilson MC, Matheus PE, Sher AV, Pybus OG, Gilbert MT, Barnes I, Binladen J, et al. 2004. Rise and fall of the Beringian steppe bison. Science. 306:1561–1565.

Shapiro B, Hofreiter M. 2014. A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. Science. 343:1236573.

Shapiro B, Sibthorpe D, Rambaut A, Austin J, Wragg GM, Bininda-Emonds OR, Lee PL, Cooper A. 2002. Flight of the dodo. Science. 295:1683.

Smith CI, Chamberlain AT, Riley MS, Stringer C, Collins MJ. 2003. The thermal history of human fossils and the likelihood of successful DNA amplification. J Hum Evol. 45:203–217.

Stiller M, Baryshnikov G, Bocherens H, Grandal D'Anglade A, Hilpert B, Münzel SC, Pinhasi R, Rabeder G, Rosendahl W, Trinkaus E, et al. 2010. Withering away-

25,000 years of genetic decline preceded cave bear extinction. Mol Biol Evol. 27:975–978.

Stiller M, Sucker A, Griewank K, Aust D, Baretton GB, Schadendorf D, Horn S. 2016. Single-strand DNA library preparation improves sequencing of formalin-fixed and paraffin-embedded (FFPE) cancer DNA. Oncotarget. 7:59115–59128.

Troll CJ, Kapp J, Rao V, Harkins KM, Cole C, Naughton C, Morgan JM, Shapiro B, Green RE. 2019. A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. BMC Genomics. 20:1023.

Van Beers EH, Joosse SA, Ligtenberg MJ, Fles R, Hogervorst FB, Verhoef S, Nederlof PM. 2006. A multiplex PCR predictor for aCGH success of FFPE samples. Br J Cancer. 94:333–337.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, et al. ; Broad Institute Genome Sequencing Platform; Broad Institute Whole Genome Assembly Team. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 326:865–867.

Wales N, Carøe C, Sandoval-Velasco M, Gamba C, Barnett R, Samaniego JA, Madrigal JR, Orlando L, Gilbert MT. 2015. New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA. Biotechniques. 59:368–371.

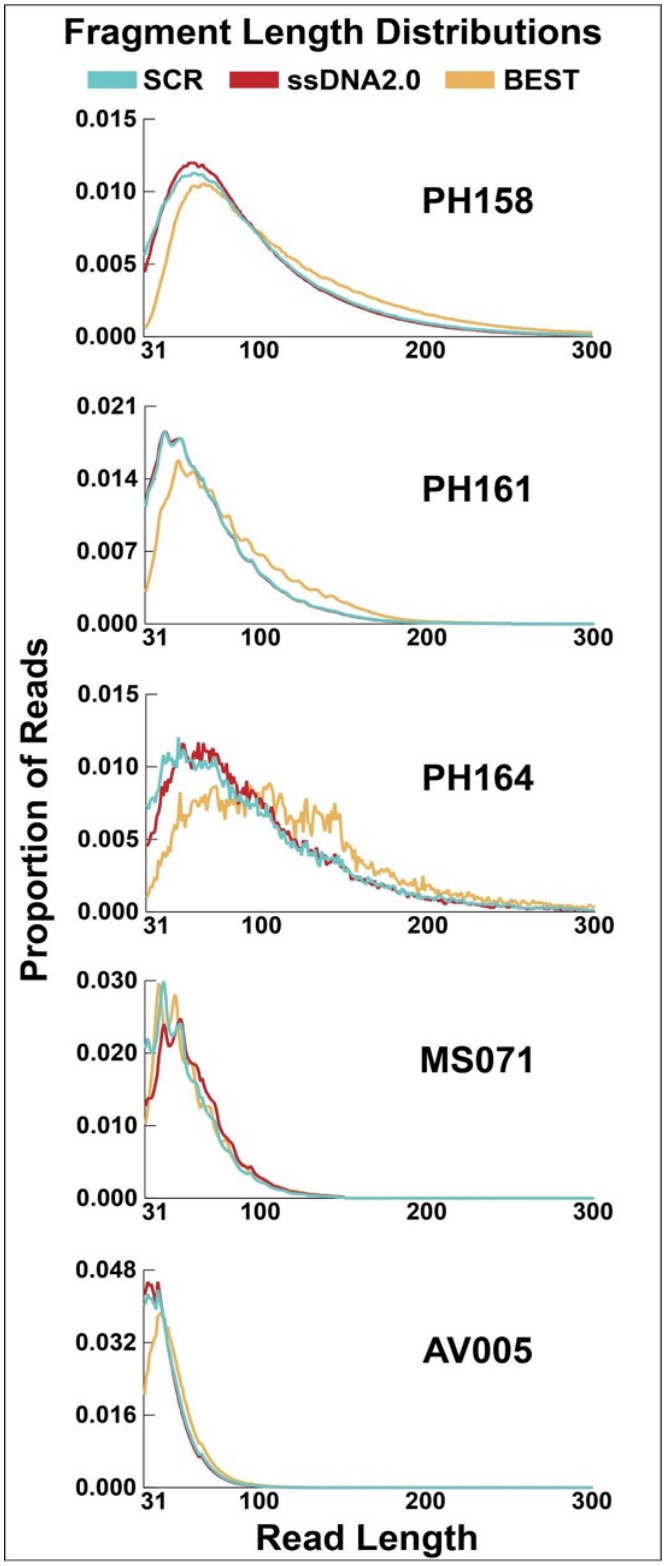

Figure 2.4. Length distribution of reads mapped to the reference genome.

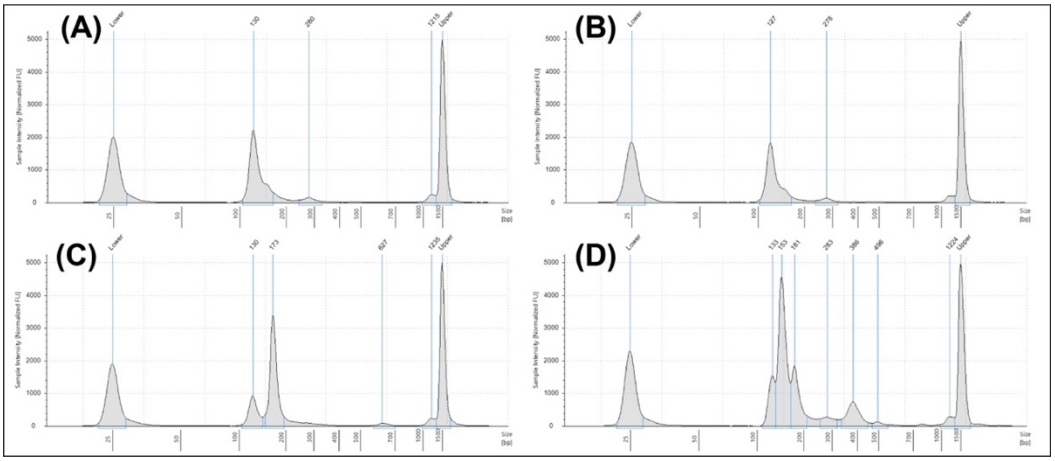

Figure 2.5. Four P5 splint oligonucleotide spike-in reaction TapeStation traces. **(A)** and **(B)** are P5 splint batches that contain acceptable levels of oligonucleotide secondary artifact species. **(C)** and **(D)** are P5 splint batches that have high levels of oligonucleotide secondary artifact species and should be quarantined from normal library use or discarded.

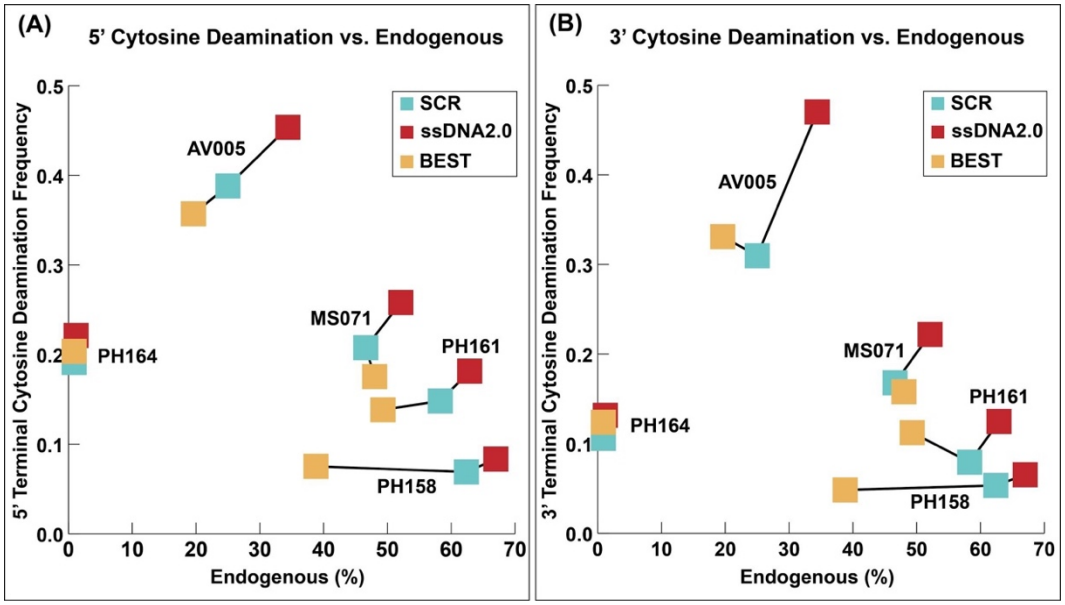

Figure 2.6. Scatter plots of **(A)** 5' terminal deamination frequency vs. the percentage of endogenous reads in the library and **(B)** 3' terminal deamination frequency vs. the percentage of endogenous reads library.

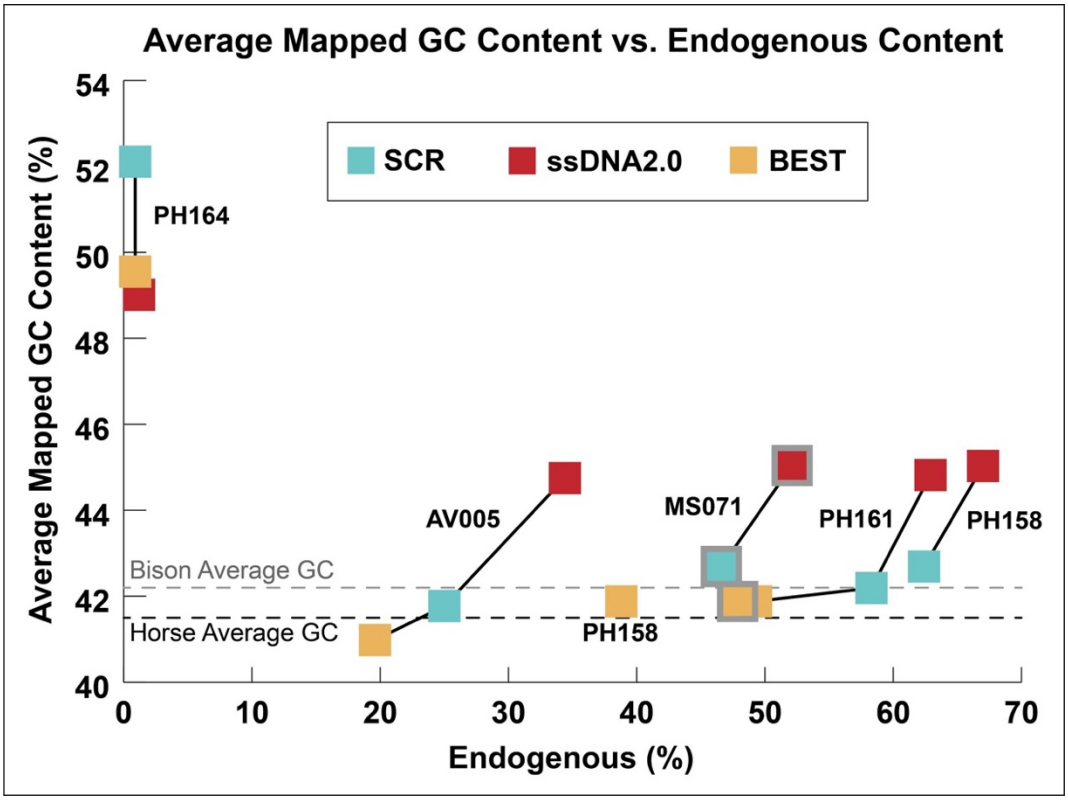

Figure 2.7. Scatter plot of the average GC content of mapped reads vs. the percentage of endogenous reads in the library.

UCSC ID	Museum ID	Organism	Age (C14 date)	Location	Bone Powder Per Extract (mg)	Extracts Performed	Pool Conc. (ng/µL)
PH158	F-2431	Horse	40,680	Yakutiya, Russia	100	4	3.77
PH161	F-2530	Horse	Not Dated	Yakutiya, Russia	120	4	1.26
PH164	F-2538	Horse	Not Dated	Yakutiya, Russia	120	4	4.07
MS071	YG 303.666	Bison	16,390	Yukon Territory, Canada	120	4	0.52
AV005	YG412.17	Horse	Not Dated	Yukon Territory, Canada	120	4	0.60

Table 2.1. Overview of the ancient samples used for DNA extraction and library preparation.

Method	Oligo Name	Provider	Purification	Sequence 5'->3'
SCR	scr_p5_adapter	IDT	HPLC	/5AmMC12/ACACTCTTTCCCTACACGACGCTCTTCCGATCT
	scr_p5_splint	IDT	HPLC	/5AmMC6/NNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT/3AmMO/
	scr_p7_adapter	IDT	HPLC	/5Phos/AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC/3AmMO/
	scr_p7_splint	IDT	HPLC	/5AmMC12/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNN/3AmMO/
ssDNA2.0	CL78	Sigma	Ion-exchange HPLC (Dual)	Pho-AGATCGGAAG[C3Spacer]10-TEG-biotin
	TL136	Sigma	RP-HPLC	SpacerC12-AA[SpacerC12]CTTCCGATCTNNNNNNN-AmC6
	CL53	Sigma	RP-HPLC	CGACGCTCTTC-ddC
	CL73	Sigma	RP-HPLC	Pho-GGAAGAGCGTCGTGTAGGGAAAGAG*T*G*T*A
	CL130	Sigma	RP-HPLC	GTGACTGGAGTTCAGACGTGTGCTCTTCC*GA*TC*T
BEST	lS1_BEDC3	IDT	HPLC	A*C*A*C*TCTTTCCCTACACGACGCTCTTCCG*A*T*C*T
	lS2_BEDC3	IDT	HPLC	G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T
	lS3_BEDC3	IDT	HPLC	A*G*A*T*CGGAA*G*A*G*C/3SpC3/
qPCR	lS7	IDT	Desalted	ACACTCTTTCCCTACACGAC
	lS8	IDT	Desalted	GTGACTGGAGTTCAGACGTGT

Table 2.2. Information for oligonucleotides used during library preparation.

Method	Lib ID	Sample ID	Organism	Lib Input (ng)	Lib Input (pmol)	qPCR (CT)	Pre-Amp Vol (µL)	iPCR Input (µL)	Cycles Amplified	Ref Genome	Raw Reads	Adapter-Dimers (%)	Sequences >30bp (%)	Mapped Sequences Over 30bp (%)	Mapped Unique Sequences (%)	Average Size of Mapped Sequences (bp)	C-to-T Sub Freq 5' end (%)	C-to-T Sub Freq 3' end (%)
SCR	JKFC1	PH158	Horse	29.70	1.000	10.2	50	2	10	equCab2	11,000,000	1.82	93.07	61.65	98.46	94.90	7.05	6.04
	JKFC2	PH158	Horse	14.85	0.500	11.4	50	2	11	equCab2	11,000,000	1.69	93.18	61.28	98.36	95.70	7.06	5.96
	JKFC3	PH158	Horse	7.43	0.250	12.4	50	2	12	equCab2	11,000,000	1.32	93.96	61.06	98.31	96.44	7.02	5.88
	JKFC4	PH158	Horse	3.71	0.125	13.5	50	2	14	equCab2	11,000,000	1.18	94.19	62.39	97.81	94.48	6.90	5.34
	JKFC5	PH158	Horse	1.86	0.063	14.7	50	2	15	equCab2	11,000,000	1.11	94.59	63.95	96.49	94.49	6.24	4.15
	JKFC6	PH158	Horse	0.93	0.031	16.5	50	2	16	equCab2	11,000,000	2.25	92.13	62.44	93.11	93.13	5.79	3.68
	JKFC8	PH161	Horse	3.71	0.125	13.6	50	2	14	equCab2	13,000,000	0.82	92.25	58.33	97.69	69.35	14.82	7.93
	JKFC10	PH164	Horse	3.71	0.125	15.3	50	2	15	equCab2	7,200,000	7.31	83.50	0.92	96.51	94.50	19.11	10.60
	JKFC12	MS071	Bison	3.71	0.125	13.3	50	2	13	bison_umd1.0	12,000,000	0.81	90.82	46.57	97.43	56.38	20.77	16.80
	JKFC14	AV005	Horse	3.71	0.125	14.6	50	2	15	equCab2	12,000,000	1.82	86.68	25.04	96.53	45.82	38.81	30.97
ssDNA2.0	JKFC57	PH158	Horse	29.70	1.000	13.0	50	2	13	equCab2	11,000,000	0.14	98.02	69.78	98.31	101.22	8.93	7.00
	JKFC58	PH158	Horse	14.85	0.500	13.3	50	2	13	equCab2	11,000,000	0.18	97.50	69.02	98.32	98.14	8.67	6.98
	JKFC59	PH158	Horse	7.43	0.250	13.9	50	2	14	equCab2	11,000,000	0.19	96.76	68.01	98.16	92.95	8.58	6.89
	JKFC60	PH158	Horse	3.71	0.125	14.3	50	2	14	equCab2	11,000,000	0.19	95.52	67.02	97.72	92.46	8.33	6.53
	JKFC61	PH158	Horse	1.86	0.063	15.5	50	2	15	equCab2	11,000,000	0.21	92.64	66.54	97.10	89.12	8.35	6.58
	JKFC62	PH158	Horse	0.93	0.031	16.1	50	2	16	equCab2	11,000,000	0.25	89.99	66.14	95.74	90.46	8.32	6.49
	JKFC64	PH161	Horse	3.71	0.125	13.8	50	2	14	equCab2	13,000,000	0.16	92.91	62.90	97.95	68.69	18.15	12.48
	JKFC66	PH164	Horse	3.71	0.125	16.8	50	2	17	equCab2	7,200,000	0.42	89.37	1.18	94.57	94.68	22.14	13.16
	JKFC68	MS071	Bison	3.71	0.125	14.4	50	2	14	bison_umd1.0	12,000,000	0.21	93.70	52.09	97.48	59.90	25.81	22.19
	JKFC70	AV005	Horse	3.71	0.125	14.7	50	2	15	equCab2	12,000,000	0.26	85.10	34.40	96.32	45.21	45.35	47.01
BEST	JKFC29	PH158	Horse	29.70	1.000	13.7	50	2	14	equCab2	11,000,000	0.50	98.70	40.33	97.60	109.56	8.41	5.31
	JKFC30	PH158	Horse	14.85	0.500	14.7	50	2	15	equCab2	11,000,000	0.37	98.94	40.08	96.32	114.54	8.09	4.97
	JKFC31	PH158	Horse	7.43	0.250	15.8	50	2	16	equCab2	11,000,000	0.34	98.86	38.80	94.37	111.43	7.90	5.10
	JKFC32	PH158	Horse	3.71	0.125	17.2	50	2	17	equCab2	11,000,000	0.18	99.06	38.83	88.59	111.43	7.58	4.88
	JKFC33	PH158	Horse	1.86	0.063	18.4	50	2	18	equCab2	11,000,000	0.07	99.22	38.82	78.31	112.65	7.05	4.72
	JKFC34	PH158	Horse	0.93	0.031	19.9	50	2	20	equCab2	11,000,000	0.03	99.26	39.51	58.17	109.95	6.49	4.37
	JKFC36	PH161	Horse	3.71	0.125	16.4	50	2	16	equCab2	13,000,000	0.03	97.72	49.34	92.90	81.43	13.84	11.24
	JKFC38	PH164	Horse	3.71	0.125	19.9	50	2	20	equCab2	7,200,000	2.02	96.61	0.91	51.28	115.01	20.36	12.39
	JKFC40	MS071	Bison	3.71	0.125	17.4	50	2	17	bison_umd1.0	12,000,000	0.05	96.25	47.95	91.89	58.33	17.54	15.79
	JKFC42	AV005	Horse	3.71	0.125	17.8	50	2	18	equCab2	12,000,000	0.05	95.85	19.63	88.02	48.80	35.71	33.10

Table 2.3. Summary of library sequencing statistics.

Chapter 3: Developing a workflow to process single hair shafts for next-generation sequencing and characterizing the recovered DNA

This project was a collaboration with Hayley Neadeau, Ciara Wanket, and Richard E. Green. I designed and performed workflow optimization experiments, performed the DNA extractions, performed the library preparations, performed data analysis, and wrote the manuscript. Hayley Neadeau and Ciara Wanket assisted in the wet lab. Richard E. Green provided input throughout the study and performed data analysis. The data presented in this chapter will be included in a larger manuscript, which will also investigate the genotyping accuracy of the low coverage hair data compared to high coverage data generated from saliva.

3.1 Introduction

Hair is a common evidence type collected at crime scenes [1] and the vast majority of collected hairs do not contain a root [2]. Despite the abundance of hair shafts as a potential forensic sample type, forensic labs rarely successfully generate data using short tandem repeat (STR) typing from hair and often do not pursue the sample type [3]. However, small quantities of degraded nuclear DNA can be recovered from hair shafts [3] and hair has been shown to be reliably resistant to contamination [4]. Data has been challenging to generate from hair with typical amplicon-based forensic techniques, but hair represents an underused sample type in the next generation sequencing (NGS) era.

The high-quality DNA contained in the hair root rapidly degrades during keratinization, which leaves the DNA present in the hair shaft highly degraded. Rapid

cell turnover occurs in hair, where the root is metabolically active, but cells become fully keratinized approximately 1 mm from the follicle base after 2.5 days [5]. The keratinization process causes cell cytolysis and subsequent nuclease attack, which fragments DNA [6]. For example, the enzyme DNase1L2 has been detected in hair follicles and is associated with nuclear DNA-specific degradation during keratinization, via cleavage of unprotected DNA between nucleosomes [7].

The short length of the surviving DNA fragments from hair shafts is a key challenge for forensic labs performing STR typing assay. PCR-based STR typing has been the primary method to generate molecular data for criminal identification for decades [8–10]. STR typing assays are simple to multiplex, cost-effective, and the data can be compared to over 19 million profiles in the National DNA Index (NDIS) [11]. However, typical STR assays have been shown to only produce full profiles from up to 20% of hair shafts while consistently yielding full profiles from hair roots [12, 13]. Typical STR assays require target sizes of several hundred base pairs to produce amplicons, but the average length of DNA recovered from fresh hair shafts has been reported to be less than 100 bp [3]. The development of reduced size STR amplicons, many amplicons less than 100 bp, increased the success of generating profiles from some degraded sample types [14] but hair shafts remained challenging [15]. Furthermore, reduced size STR amplicons assays for hair shafts have been reported to be cumbersome to perform [3] without returning consistent results. The low success rate of hair shafts has led forensics labs to rarely perform STR typing on hair shafts [3] despite the abundance at crime scenes [1].

Investigative genetic genealogy has become a growing field within forensics, which overcomes some limitations of STR typing when working with degraded samples. Genetic genealogy relies on the generation of single nucleotide polymorphism (SNP) profiles comprised of hundreds of thousands of observations [16]. SNP profiles can produce investigative leads through familial matches when compared to SNP profiles uploaded to GEDmatch [17], a public genetic genealogy site. The SNP profiles contained in GEDmatch are primarily generated by direct-to-consumer ancestry tests using high-density array technology and then uploaded by the consumer. Array technology is a rapid and affordable method to generate SNP data but requires hundreds of nanograms of high-quality DNA (Illumina Infinium Global Diversity Array-8 v1.0). However, SNP profiles can be generated from whole genome sequencing (WGS) data, which has been a routine data generation approach for fields working with degraded DNA for over a decade [18]. Generating high coverage WGS data from degraded samples can be challenging and costly but imputation approaches have been developed to improve SNP calling accuracy from low genome coverage, down to 1X [19–21].

The success of generating enough genomic data from degraded samples hinges on applying appropriate laboratory methods. Sample pre-treatment before DNA extraction, such as sodium hypochlorite washes [4, 22, 23], are commonly applied to degraded samples to remove contaminating DNA but may also lead to the loss of endogenous molecules [23]. Several DNA extraction methods for degraded samples are used for sediment and bone [24–26] but sample-to-sample performance variation [26] typically results in project-specific method comparisons. Furthermore, library

preparation choice influences the amount and type of molecules converted to library from poor quality samples, which demands the use of degraded DNA optimized methods [27–29]. Notably, single-stranded DNA library preparation methods have been shown to yield better quality libraries from degraded samples compared to double-stranded DNA methods [27, 29–31]. Processing degraded samples for NGS has become routine but method choice and occasionally sample-specific optimizations influence the success of data generation.

NGS methods have been applied to ancient hair samples for over a decade, but often multiple hairs are processed in the same DNA extraction, and the samples likely degraded over time. Clumps of permafrost preserved ancient hair shafts were processed using early sequencing by synthesis technology to generate whole mitochondrial genomes [32, 33]. Hair remains an infrequent sample type used in the field to generate mitochondrial genomes and small amounts of nuclear data [34–37]. More recently, efforts to characterize the content of single human hair shafts have confirmed the DNA recovered from even fresh hairs is heavily degraded [3]. However, the expansive characterization of DNA recovered from a variety of fresh hair shafts has not been fully explored.

Here, we describe an optimized workflow for processing single rootless hairs and use the workflow to generate whole genome sequencing data from the head and pubic hairs of 50 anonymous volunteers. We describe the characteristics of DNA recovered and the range of coverage that can be generated from a single head and pubic hairs no more than 5cm and 3cm respectively. Additionally, we explore the differences

between the DNA recovered from head and pubic hairs and how hair volume relates to DNA content. We find most single hair shafts yield enough informative content to generate multi-fold genome coverage and believe fresh hair shafts are a valuable forensic sample type when NGS approaches are applied.

3.2 Methods

3.2.1 Sample Collection and Pre-Processing

We collected head and pubic hair from 50 anonymous volunteers under an IRB-approved protocol (HS3382). Participants anonymously picked up and dropped off a collection kit containing an OGR-500 (DNA Genotek) saliva collection device, a plastic bag for head hair, a plastic bag for pubic hair, and a set of instructions. Each participant was requested to donate at least 10 head hairs, saliva following the OGR-500 instructions, and was given the option to provide at least 3 pubic hairs.

We removed identifiable roots, trimmed, and measured each hair to calculate volume. First, we examined each hair and removed identifiable roots. If a root was not identified, then we trimmed approximately 1cm from each end of the hair. Following root removal, we trimmed the head hairs to 5cm and pubic hairs to 3cm. The length was recorded if a hair was shorter than the desired length. Following calibration using the provided reference slide, we imaged each hair using a Zeiss Primo Star microscope and SwiftCam microscope camera. The diameter of each hair was found using the SwiftImaging 3.0 software. Finally, we calculated the volume of each hair using the measured length and diameter.

3.2.2 Extraction Method Descriptions

We tested two ancient DNA optimized isolation approaches described in Rohland et al. [26], a silica spin column and silica-coated magnetic bead-based solution.

The silica column isolation method is an optimization of the commonly used ancient DNA isolation approach outlined in Dabney et al. [24]. Briefly, 1 mL of lysate is combined with 10.4 mL of a high concentration guanidine and isopropanol binding buffer (binding buffer D). The binding buffer and lysate mixture is passed through a large volume silica spin column assembly by centrifugation. Next, the silica membrane is washed twice with 750 µL of an ethanol-based wash buffer, Qiagen buffer PE. Finally, the DNA is eluted in 50 µL of buffer EBT (10 mM Tris, 0.5% Tween-20).

The bead isolation method uses identical buffers as the column approach but different buffer volumes and relies on silica-coated magnetic beads. Briefly, 0.15 mL of lysate, 15% of a typical lysate volume, is combined with 1.57 mL of binding buffer D. The binding buffer and lysate mixture is incubated with silica-coated magnetic beads before magnetic separation. Next, the beads are washed three times with 250 µL of Qiagen buffer PE. Finally, the DNA is eluted in 30 µL of buffer EBT.

3.2.3 Library Preparation Method Description - Spotlight

Spotlight libraries are prepared as described in Kapp et al. [29] with the following modifications: 9 µL DNA extract, 1 µL 76 ng/µL ET SSB (NEB), 1 µL 2 µM P5 splinted adapter, 1 µL 0.4 µM P7 splinted adapter, and 12 µL reaction mix (per reaction: 9.20 µL 50% 8000, 1.25 µL 1M Tris-HCl, 0.25 µL 1M MgCl2, 0.25 µL 1M

DTT, 0.25 µL 100mM ATP, 0.55 µL 10,000 U/mL T4 PNK (NEB), and 0.25 µL 2,000,000 U/mL T4 DNA Ligase(NEB)). Following ligation, a SPRI clean is performed as described in Rohland and Reich [38] using 35 µL buffer EBT and 60 µL SPRI solution for the initial incubation. Cleaned libraries are eluted in 21 µL of buffer EBT before proceeding to qPCR and indexing PCR.

3.2.4 DNA Extraction Method Comparison

We extracted and isolated DNA from three hairs for eight individuals using the silica column and silica magnetic bead approaches previously described in 3.2.2, for a total of 24 hair extractions per method. Individuals were randomly selected from a pool of individuals that had donated more than 10 hairs at least 5cm in length. Six hairs per individual were trimmed of any identifiable root, approximately 1 cm, and then further trimmed to 5 cm. To clean the hair exterior, we submerged each hair in 0.5% sodium hypochlorite for 10 seconds and then submerged the hairs in three water baths for 10 seconds each.

Immediately after washing, we submerged each hair in a lysis buffer composed of 2% SDS, 10 mM Tris-HCl, 2.5 mM EDTA, 10 mM NaCl, 5 mM CaCl2, 40 mM DTT, and 2 mg/mL Proteinase K. Three hairs from each individual were submerged in 1 mL (high lysis volume) of lysis buffer and another three in 0.145 mL (low lysis volume). The hairs were incubated at 55°C for approximately 18 hours. Following the overnight incubation, we added 3 µL of 1M DTT and 7 µL of 20 mg/mL Proteinase K to each low lysis volume sample and incubated them for an additional 1 hour at 55°C.

Next, we incubated the low lysis volume samples on ice for 5 minutes and then spun the tube for 2 minutes at 16,000 rcf in a microcentrifuge.

DNA was isolated from the entire volume of the high lysis volume samples using the column approach and from 0.15 mL of the low lysis volume samples using the bead approach. We eluted the column samples in 50 µL EBT and the bead samples in 29 µL EBT. Next, we transferred the bead extractions to a new tube and then added 21 µL of EBT, for a final volume of 50 µL. We quantified 20 µL of each extraction using a Qubit 1X dsDNA HS Assay Kit and a Qubit 4.

From each extraction, we prepared Spotlight libraries from 10 µL of extract as described in 3.2.3. Following ligation and cleanup, we performed qPCR and double-indexed each library as described in Kapp et al. [29]. We purified the amplified libraries using a SPRI ratio of 1.2X and eluted the cleaned libraries in 22 µL. Finally, we quantified the libraries using a Qubit 1X dsDNA HS Assay Kit and a Qubit Flex fluorometer before pooling equal-nanogram for sequencing on a NextSeq 550 using 150 cycle mid output reagent kits.

3.2.5 Panel – DNA Extraction of Hair Samples

For 50 individuals, we extracted and isolated DNA from two head hairs up to 5cm and, when available, one pubic hair up to 3cm. To clean the hair exterior, we submerged each hair in 0.5% sodium hypochlorite for 10 seconds and then submerged the hairs in three water baths for 10 seconds each. We placed each trimmed hair in 145 µL of lysis buffer (2% SDS, 10 mM Tris-HCl, 2.5 mM EDTA, 10 mM NaCl, 5 mM CaCl2, 40 mM DTT, and 2 mg/mL Proteinase K). Next, we incubated the hairs

overnight in a 750rpm, 55°C thermoshaker. Following the overnight incubation, we added 3 µL of 1M DTT and 7 µL of 20 mg/mL Proteinase K and incubated each sample for an additional 1h in a 750rpm, 55°C thermoshaker. We incubated each sample on ice for 5 m and spun the tubes for 2 m at 16,000 rcf in a microcentrifuge. Finally, we transferred 150 µL of the supernatant to a new tube while avoiding the pellet. We isolated DNA from 150 µL of the supernatant following the silica magnetic bead approach described in Rohland et al. [26] using Buffer D and eluting in 28 µL of buffer EBT.

3.2.6 Panel – Library Preparation of Hair Samples

For each individual, we prepared three Illumina libraries per extract for up to two head hair and one pubic hair DNA extraction, depending on availability. We prepared Illumina libraries as described in section 3.2.3 with the following modifications: 9 µL DNA extract, 2 µL 38 ng/µL ET SSB (NEB), 2 µL adapter mix containing 1 µM P5 splinted adapter and 0.2 µM P7 splinted adapter.

We performed quantitative PCR on each library using primers IS7 and IS8 described in Gansauge and Meyer [39] to inform the optimal number of cycles for indexing PCR. We prepared 50 µL qPCR reactions containing: 1 µL library, 25 µL 2X Maxima SYBR Green Master Mix (Thermo Scientific), 0.5 µL 100 µM primer IS7, 0.5 µL 100 µM primer. Next, we cycled the reactions using a BioRad CFX Opus 96 using the following conditions: 95°C for 10m, followed by 40 cycles of 95°C for 30s, 60°C for 30s, and 72°C for 30s. Fluorescence was measured at the end of each extension step.

Following qPCR, we amplified and double-indexed each library using the primers described in Kircher et al [40]. We prepared 50 µL indexing reactions containing: 20 µL library, 25 µL 2X Amplitaq Gold 360 Master Mix (Applied Biosystems), 2.5 µL unique 20 µM i7 indexing primer, and 2.5 µL unique 20 µM i5 indexing primer. Next, we amplified the libraries with the following conditions: 95°C for 10m, followed by a library-specific number of cycles of 95°C for 30s, 60°C for 30s, and 72°C for 60s, and a final extension of 72°C for 7m. We purified the amplified libraries using a SPRI ratio of 1.2X and eluted the cleaned libraries in 22 µL. Finally, we quantified the libraries using a Qubit 1X dsDNA HS Assay Kit and a Qubit Flex fluorometer before pooling equal-nanogram for sequencing.

We developed protocols and performed the pre-PCR liquid handling steps on an Agilent Bravo with an NGS option A configuration.

3.2.7 Panel – Sequencing

For quality control, we sequenced all libraries at the University of California, Santa Cruz Ancient and Degraded DNA Processing Center on an Illumina Next 550 using 150 cycle mid-output reagent kits. For head hair extractions from 50 individuals and pubic hair extractions from 30 individuals, we sent the 3 libraries associated with each extraction to the Duke Center for Genomic and Computational Biology for sequencing on a NovaSeq 6000 using S4 200 cycle reagent kits.

3.2.8 Panel – Analysis

The optimization libraries and the panel libraries were processed with the same pipeline. First, we merged reads that overlapped by at least 15 bases, trimmed the reads

of tailed adapter sequences, and discarded reads less than 30 bp using SeqPrep [41]. Following read pre-processing, we mapped the merged and unmerged reads separately using Burrow-Wheeler Aligner (BWA) [42] aln algorithm to the human reference genome, build hg38 (GCA_000001405.15). We used SAMtools [43] to collapse duplicate reads and generate mapping statistics. We either calculated read lengths directly from the merged reads or inferred the length based on mapping coordinates for the unmerged reads. We calculated the expected total library complexity using PreSeq [44].

3.3 Results

3.3.1 DNA Extraction Method Comparison

When comparing extraction DNA yields, we observed the bead-based method to consistently recover more DNA compared to the column method. The bead method recovered more DNA averaged over 3 hair extract replicates for all but 1 individual (Figure 3.1A), where the bead method recovered 0.95X the DNA of the column approach. For the remaining individuals, the bead-based method recovered between 1.28X and 3.57X more DNA compared to the column method. However, the column approach failed to produce quantifiable extracts for individual S047, which prevented a direct comparison of yield. The bead method yielded between 0.50ng and 2.86 ng of DNA, or 17 pg/μL – 99 pg/μL given our elution volume of 29 μL. The lowest extract concentration is near the limit of detection for a Qubit 1X HS assay with 20 μL of sample input, which makes consistent quantification difficult without consuming most

of the extract. We determined quantification did not provide a performance-based benefit and chose not to quantify extractions for the remainder of the study.

Following library preparation, we used qPCR to compare the amount of amplifiable library molecules among the libraries. Libraries with a higher amount of amplifiable library molecules will reach the threshold of detection earlier compared to the library with fewer molecules. When comparing qPCR cycle threshold values (CT value), a library detected one CT value (one cycle) earlier has approximately twice the starting molecules as the lagging library.

In agreement with the DNA yields, we found the extractions prepared with the bead method resulted in libraries with lower CT values averaged over the 3 library replicates (Figure 3.1B). The CT values indicate the bead method produced libraries with 0X to 4.3X more amplifiable molecules (0 to 2.1 cycles lower) compared to the column method.

Next, we compared the mapped and unmapped content of the libraries generated from the bead and column DNA extraction approaches. We found the column approach generated libraries with 1.01X to 1.14X more reads mapping to the human genome compared to all reads (Figure 3.2A). However, the percentage of mapped reads among just the merged reads is similar between the two approaches, which may indicate both methods add low amounts of short molecule background contamination (Figure 3.2B). The difference in total mapped content is primarily explained by the increased recovery of molecules <35bp by the bead method (Figure 3.3). The increased recovery of the shortest molecules results in a lower average mapped fragment length

(Figure 3.2C) and a greater percentage of reads 1bp-25bp (Figure 3.2D), which are discarded before mapping.

While the column method produces a more informative library at shallow sequencing depths, the bead approach results in a higher complexity library. We chose to adopt the bead method, due to library complexity concerns when deep sequencing picogram DNA input libraries.

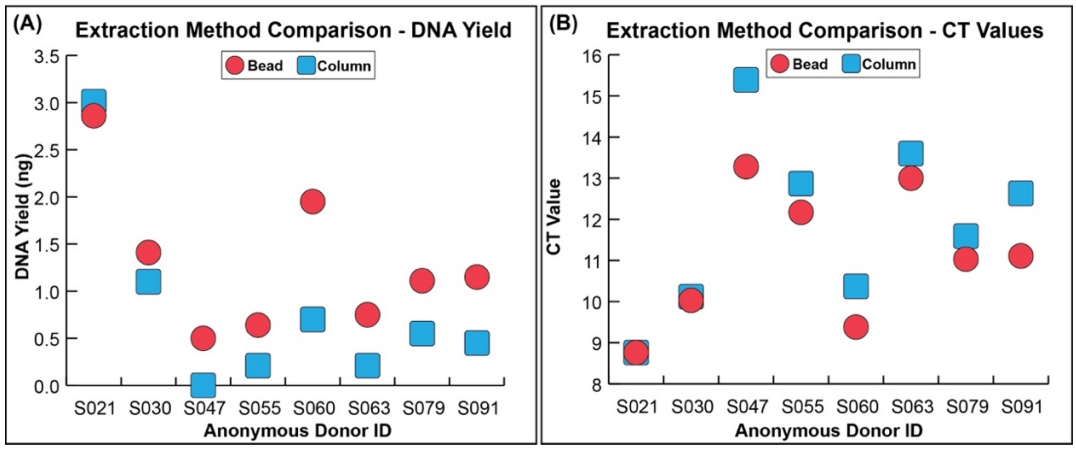

Figure 3.1: Comparison of DNA yield and library preparation efficiency between the silica coated magnetic bead and silica column-based DNA extraction methods. **(A)** Comparison of the average total DNA yield from 3 hairs for each donor per extraction method. Each extract was quantified using 20 µL of eluate and a Qubit 4. **(B)** Comparison of the average qPCR values from 3 libraries, one for each extraction, for each donor per extraction method. Libraries were prepared using 10 µL of each extraction. Lower CT values indicate more starting molecules in the qPCR reaction compared to higher CT values.

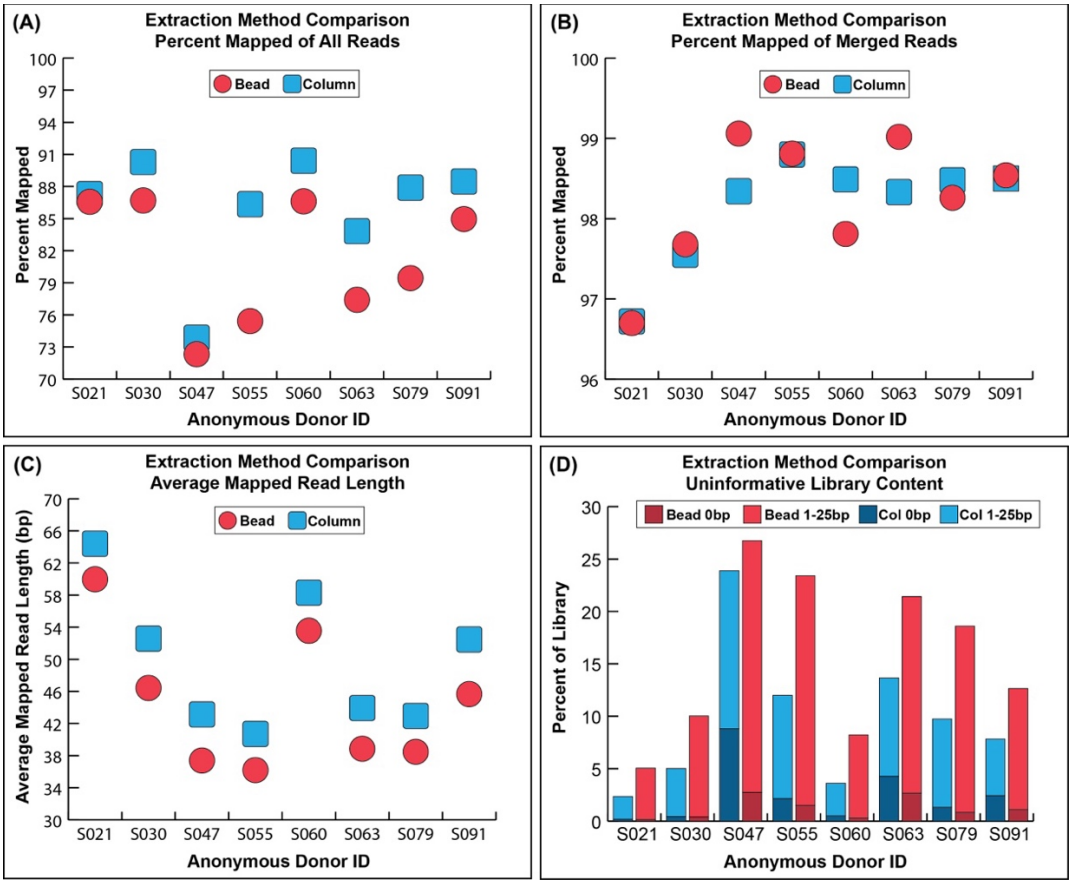

Figure 3.2: Comparison of sequencing metrics between the silica coated magnetic bead and silica column-based DNA extraction methods. **(A)** Comparison of the average percent of mapped reads to the human genome compared to all reads and **(B)** only merged reads across the 3 libraries generated from each donor. **(C)** Comparison of the average mapped read length and **(D)** the average percent of adapter-dimers and reads 1-25bp, which were discarded prior to mapping.

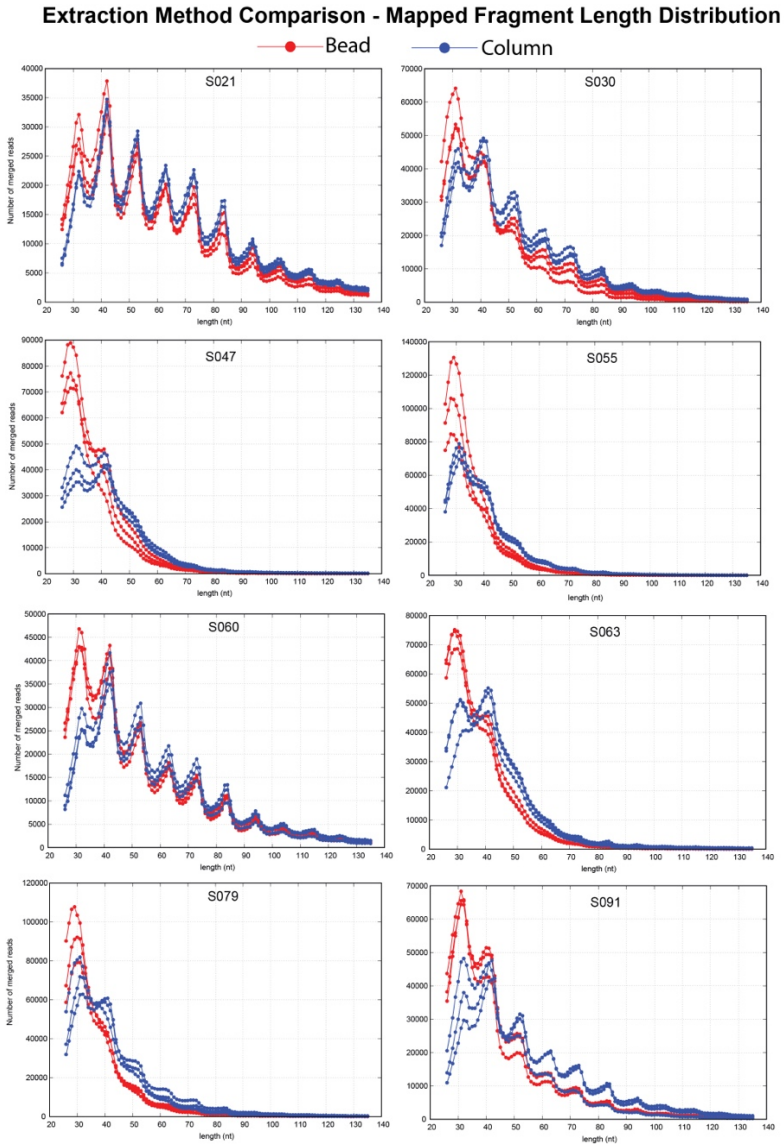

Figure 3.3: Comparison of mapped read length distributions between libraries generated from silica bead and silica column-based extraction methods. Library triplicates were prepared and sequenced for both methods for each of the 8 individuals.

3.3.2 Characterizing Hair Library Information Content

We generated between 227M to 507M (352M average) reads over 3 libraries from 50 head hairs and 30 pubic hairs to explore hair as a sample to generate whole genome

data. The average genome coverage for a single hair was calculated from the total length of the unique mapped reads across the three libraries divided by the length of the human genome. The libraries from the head hairs generated between 0.40X and 3.86X coverage, where 27 of 50 (54%) hairs generated over 2X coverage (Figure 3.4A). The libraries from the pubic hairs generated between 1.45X and 4.47X coverage where 21 of 30 (70%) hairs generated over 2X coverage.

Next, we evaluated the hair libraries potential performance if sequenced deeper. We calculated the theoretical genome coverage using PreSeqs [44] library complexity estimate after 300M reads per library, for a total of 900M reads per hair. With additional sequencing, 47 of 50 (94%) head hairs and 30 of 30 pubic hairs are estimated to generate at least 2X coverage (Figure 3.4B). Additionally, 61 of 80 (76%) hairs are estimated to generate over 4X coverage.

Estimated library coverage = ([PreSeq Unique Estimate]*[Proportion of Mapped Reads]*[Average Mapped Read Length] / [Length of Human Genome])

The hair libraries vary in the percentage of reads mapped to the human genome and the proportion of nuclear and mitochondrial mapped reads. Among the head hair libraries, between 15.6% and 77.9% of reads map to the human genome (Figure 3.5A), where nuclear reads are 9.3X to 227.7X more abundant than mitochondrial reads. Among the pubic hair libraries, between 42.9% and 82.3% of reads map to the human genome, where nuclear reads are 6.9X to 516X more abundant than mitochondrial

reads (Figure 3.6A). Neither hair type showed a significantly higher nuclear to mitochondrial read ratio compared to the other type (paired t-test, $p > 0.05$).

We observed short average mapped read lengths across all samples, where the nuclear reads were often shorter compared to the mitochondrial reads. The average mapped read length ranged between 38.05bp and 62.69bp for head hairs and 40.32bp and 103.77bp for pubic hairs (Figure 3.6B). Compared to the mitochondrial reads, the reads mapping to chromosome 1 were 4.97bp and 6.38bp shorter on average for head and pubic hairs respectively. The nuclear reads for both the pubic and head hairs were significantly shorter compared to the mitochondrial reads (paired t-test, $p < 0.05$).

Wide variation exists among the hairs of the individuals sampled but only minor differences between head and pubic hairs generally. Neither hair type produced libraries with significantly more molecules (paired t-test, $p > 0.05$), as shown in the library CT values (Figure 3.7A). While the differences are small, the pubic hairs produced libraries with a significantly higher percentage of human reads (Figure 3.7B), average mapped read length (Figure 3.7C), and observed coverage per 100M reads (figure 3.7D) (paired t-tests, $p < 0.05$).

Observed coverage per 100M = ([Coverage] / [Raw Reads] * 1000000)

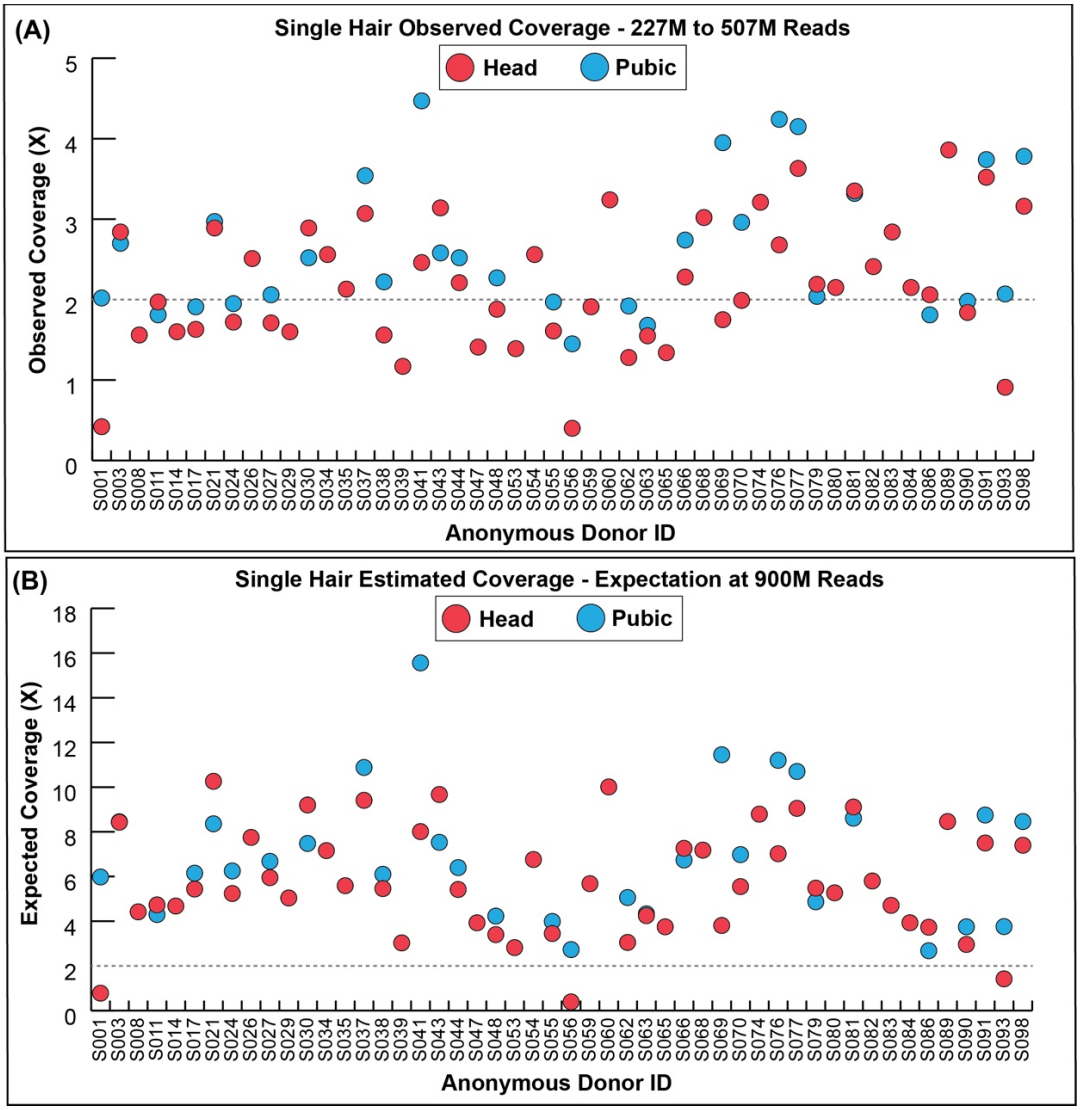

Figure 3.4: Observed and expected average coverage of the human genome for 50 head hairs and 30 pubic hairs. **(A)** The average observed coverage generated from a single hair after sequencing 227M to 507M reads across the 3 libraries. The total length of unique mapped reads across the 3 libraries generated from each hair DNA extract was divided by the size of the human genome to find the average observed coverage per hair. **(B)** The estimated average genome coverage if each library was sequenced to a depth of of 300M reads, 900M reads per hair. Estimated coverage was calculated using the number of unique reads at 300M read depth estimated by PreSeq, the percentage of mapped reads, and average read length for each library. The dotted lines for **(A)** and **(B)** mark 2X coverage.

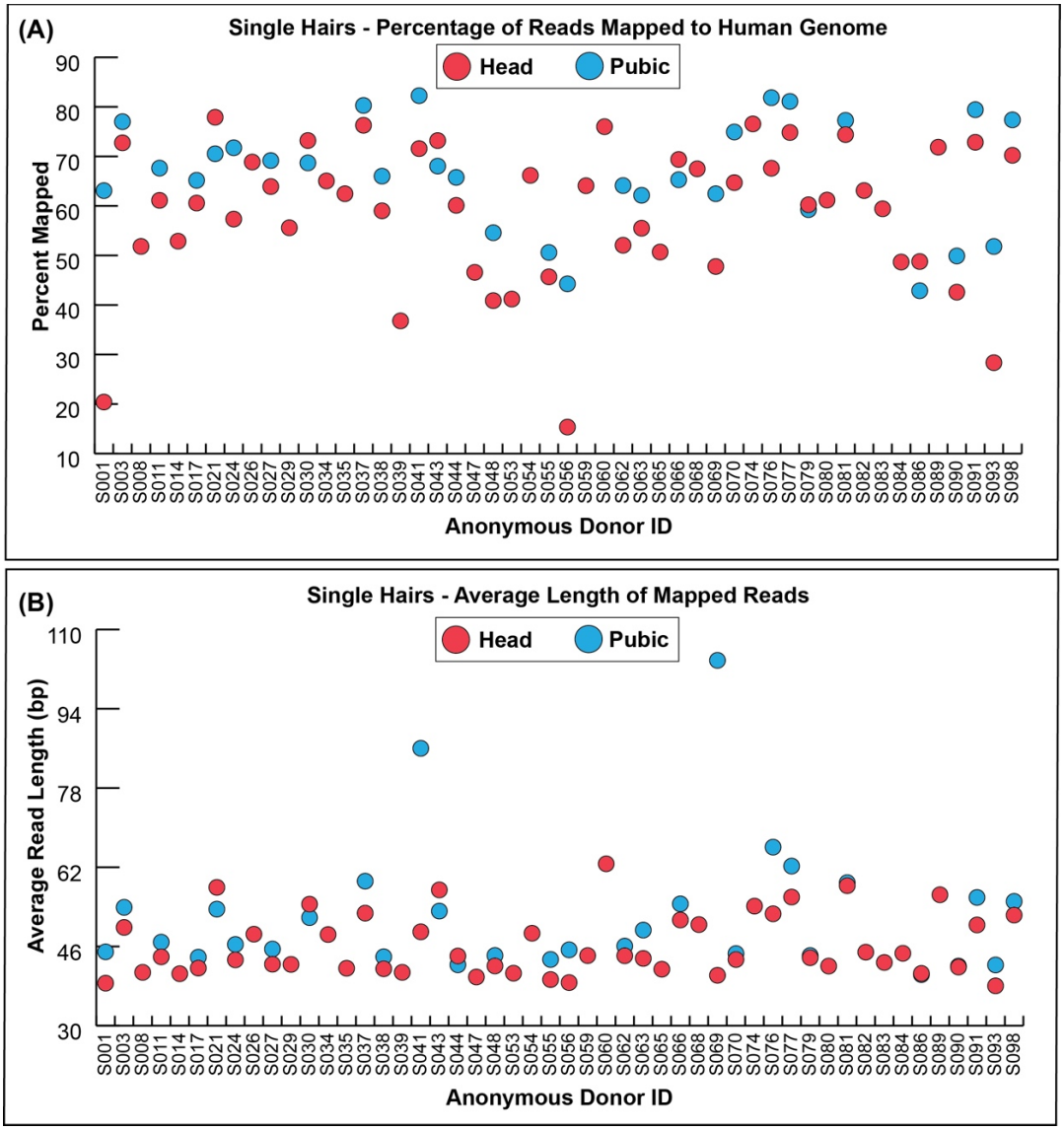

Figure 3.5: Characteristics of the informative library content for 50 head hairs and 30 pubic hairs. **(A)** The percentage of reads that mapped to the human genome and **(B)** the average length of the mapped reads, all values are averages across the 3 libraries generated from each hair.

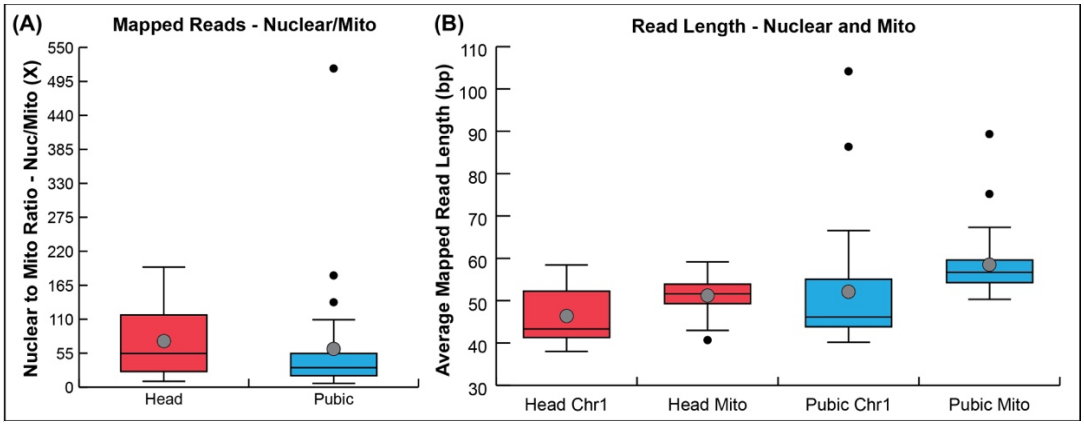

Figure 3.6: Comparison of reads mapped to the human nuclear and mitochondrial genomes. **(A)** The difference between the amount of mapped nuclear and mitochondrial reads (nuclear / mitochondrial) for head and pubic hairs. **(B)** The average length of reads mapped to chromosome 1 and the mitochondrial genome for head and pubic hairs. The comparisons for both **(A)** and **(B)** were limited to the 30 individuals with both head and pubic hairs sequenced.

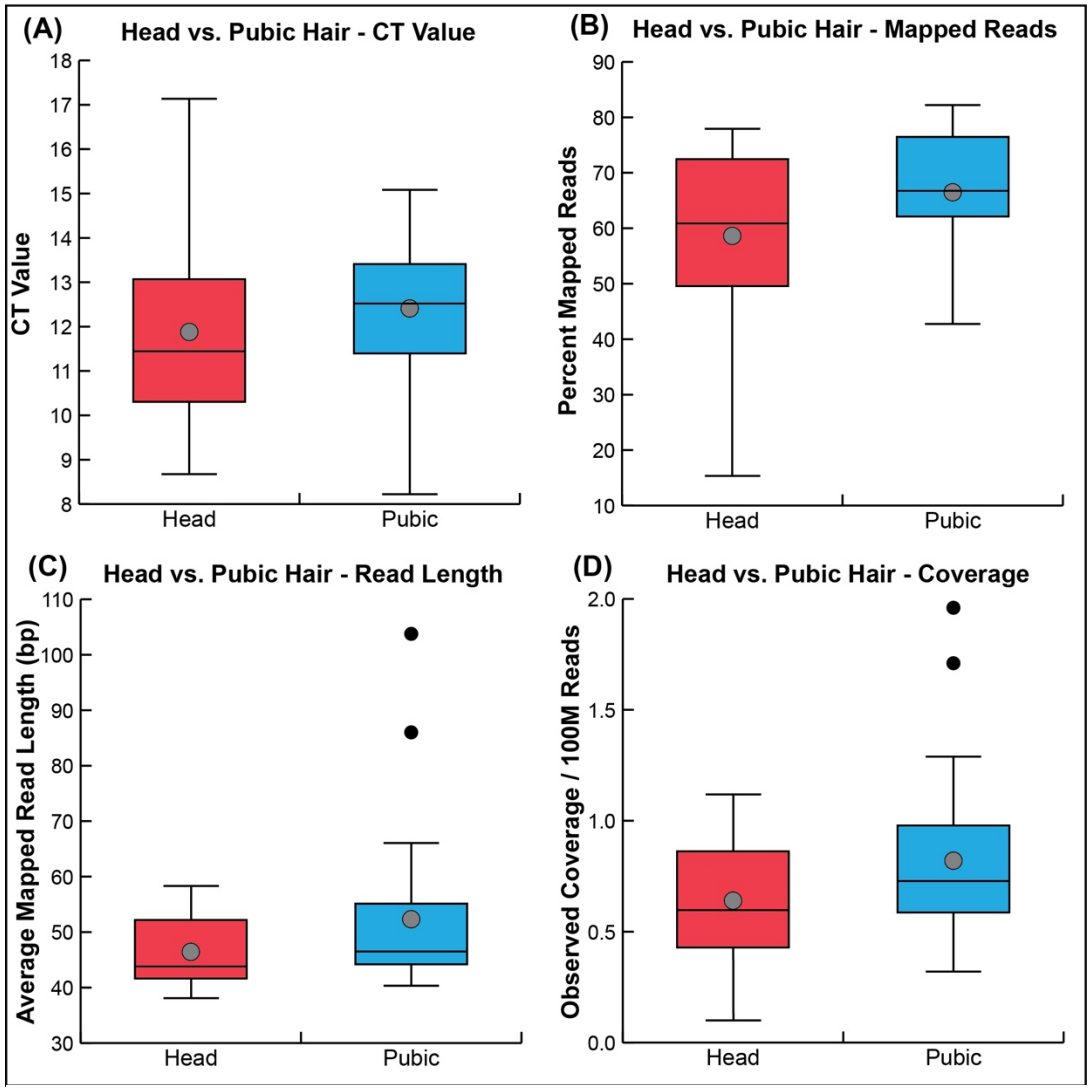

Figure 3.7: Comparison of **(A)** library CT value, **(B)** percentage of mapped reads, **(C)** average mapped read length, and **(D)** observed coverage per 100M reads generated between head and pubic hairs. The comparison was restricted to the 30 donors with both head and pubic hair sequenced.

3.3.3 Hair to Hair Variation

To evaluate hair-to-hair variation within an individual, we compared the predicted coverage between the 2 head hairs processed from each individual using approximately 1M reads generated during QC sequencing. Additionally, for the

individuals with both head and pubic hairs sequenced, we compared the observed coverage per 100M reads between the hair types.

Predicted Coverage = ([Deep Sequencing Read Target] / [Number of QC Sequencing Reads])*[Total Mapped Bases] / [Human Genome Size])

When comparing the two head hairs from an individual, we found the difference in predicted coverage to range from 1.00X to 2.5X (average 1.18X) (Figure 3.8A). The difference in observed coverage per 100M reads between an individual's head and pubic hair ranged from 1.01X to 6.34X (average 1.62X) with the pubic hair having higher coverage in 23 of 30 individuals (Figure 3.8B).

Next, we assessed if the hair with greater volume was expected or observed to generate greater coverage. We compared the difference in volume ([Vol of Higher Cov Hair] – [Vol of Lower Cov Hair]) to the difference in coverage and found hair volume is not a consistent predictor of coverage (Figure 3.8).

Volume does not appear to be a reliable indicator of a hair's potential to generate genomic data. We found a weak correlation between hair volume and the amount of library molecules generated from the DNA extract (Figure 3.9). In an extreme example, the head hairs for individual S001 and S065 had the same volume but produced libraries with approximately 53X (5.7 cycles) difference in amplifiable molecules. Additionally, we found a weak correlation between hair volume and coverage, both observed (Figure 3.10A) and estimated coverage after simulating deeper sequencing depth with PreSeq

(Figure 3.10B). Compared to hair volume, library CT value correlates better with observed (Figure 3.10C) and estimated coverage (Figure 3.10D). However, CT value measures amplifiable molecules without discriminating between on-target and off-target molecules, so should not be heavily relied upon as a sample screening metric.

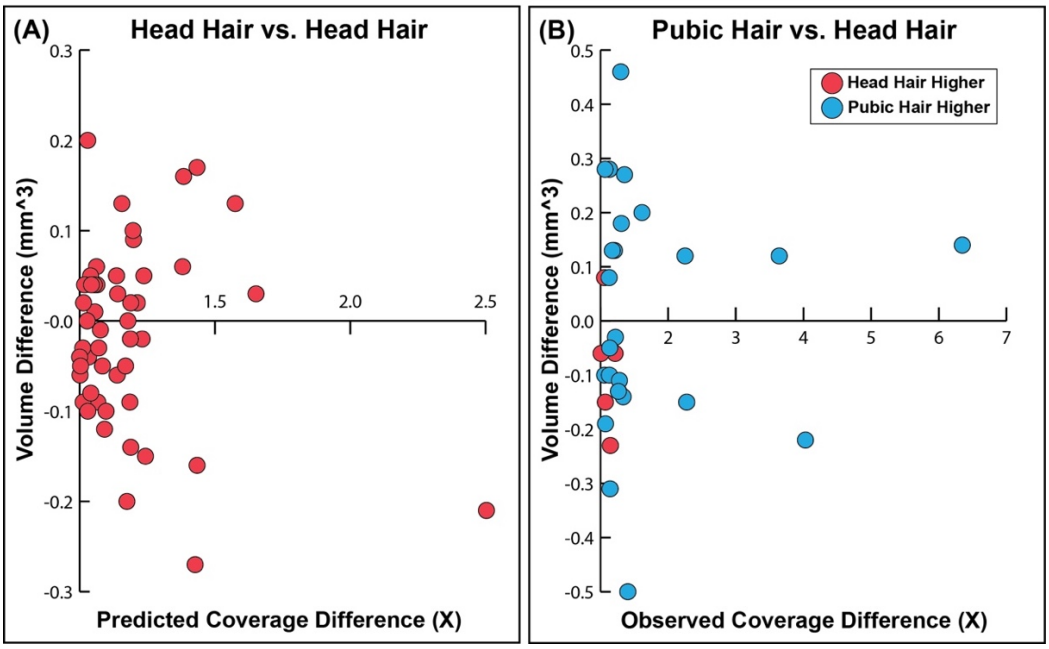

Figure 3.8: Comparison of hair variation within an individual. **(A)** The difference in predicated coverage after 300M compared to the volume difference between two head hairs from the same individual. Predicated coverage was calculated from approximately 1M reads generated during library QC. **(B)** The difference in observed coverage per 100M reads compared to the difference in volume between head and pubic hairs from the same individual. A negative volume difference for **(A)** and **(B)** indicates the higher coverage hair was lower volume compared to the lower predicated coverage hair.

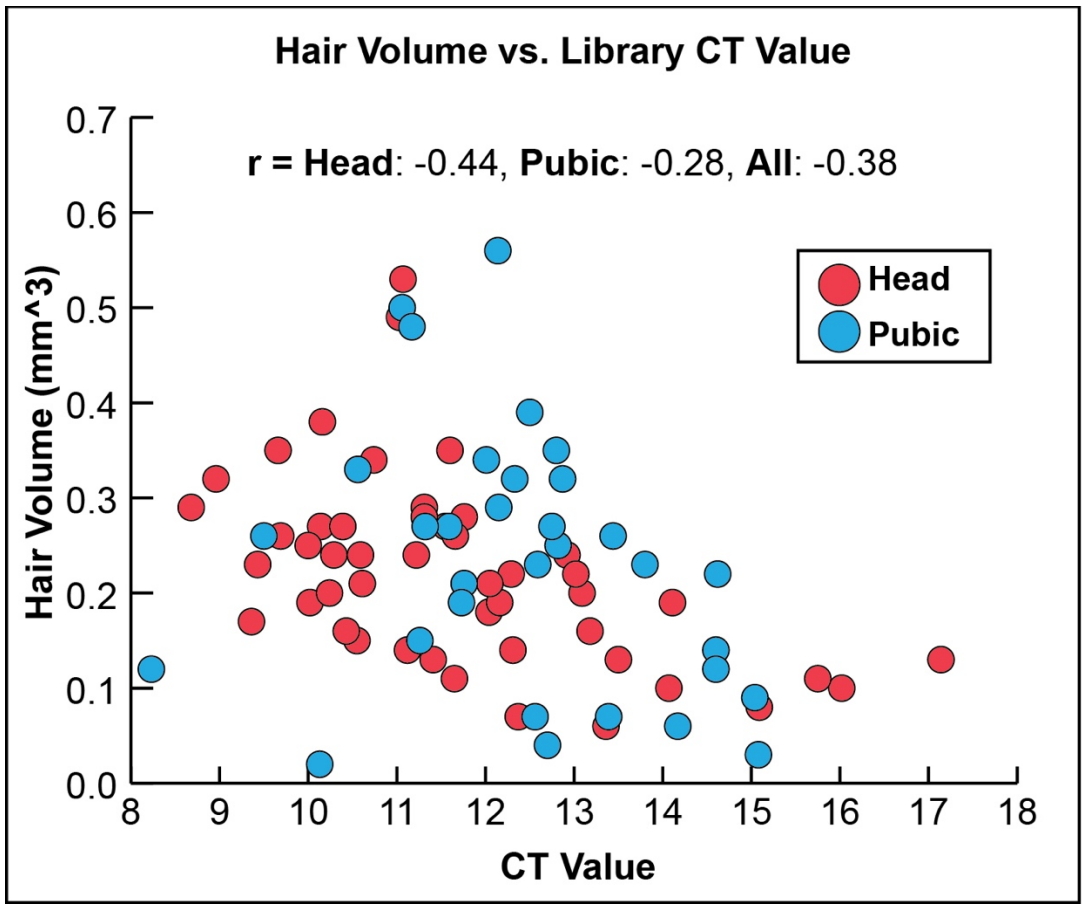

Figure 3.9: The volume of the single hair used for DNA extraction compared to the average CT value of the 3 libraries generated from the resulting extract. A lower CT value indicates more starting molecules in the reaction compared to a higher CT value. Pearson correlation coefficient is reported for head hairs, pubic hairs, and all hairs.

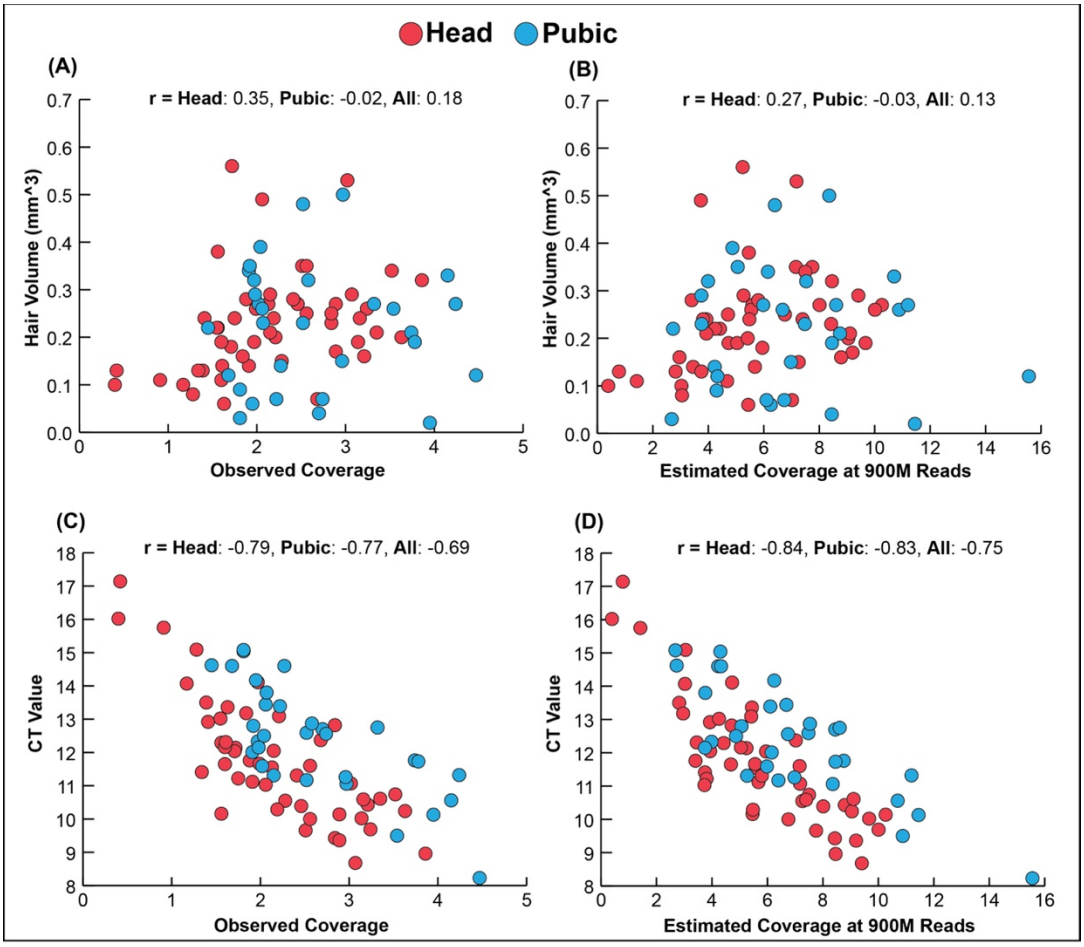

Figure 3.10: Investigating the power of volume or library CT value alone to predict the coverage potential of a single hair. Hair volume was compared to **(A)** observed coverage and **(B)** estimated coverage. Library CT value compared to **(C)** observed coverage and **(D)** estimated coverage. Estimated genome coverage was calculated from the PreSeq estimate, proportion of mapped reads, and average read length. Pearson correlation coefficient is reported for head hairs, pubic hairs, and all hairs.

3.4 Discussion

Hair shafts are an underused forensic sample type due to frequent STR assay failures but whole genome data can be confidently generated using degraded DNA approaches. As previously observed across several samples [3], we found the DNA

recovered from fresh hair shafts to be highly fragmented. On average, the mapped reads were 46bp and 52bp for head and pubic hairs respectively (Figure 3.7C), which leaves little chance for STR assay success when combined with the picogram DNA yields from single hairs. However, we generated 2X average genome coverage from 60% of the 80 hair shafts sampled (Figure 3.4A) by maximizing the recovery and retention of the shortest DNA fragments during processing. Additionally, up to 96% of single hairs appear capable of generating over 2X coverage with additional sequencing (Figure 3.4B).

We found small differences between head and pubic hairs and variation between the nuclear and mitochondrial content. Pubic hairs often had a lower volume than head hairs but had significantly more reads mapping to the human genome and longer reads, which contributed to greater genome coverage when adjusted for sequencing depth (Figure 3.7). While the differences between pubic and head hairs are small, we found the three individuals (S001, S056, S093) with the poorest performing head hairs had pubic hairs capable of generating multi-fold genome coverage (Figure 3.4A). The ratio of nuclear to mitochondrial reads varied widely among the hairs but the mitochondrial reads were significantly longer compared to the nuclear reads (Figure 3.6). Better preservation of mitochondrial reads has been previously assumed due to the greater success rate of assays targeting the mitochondrial genome [45] and more recently directly overserved among several fresh hairs [3].

The high variation among hairs and weak correlation between hair volume and quality requires the use of a conservative wet lab workflow. We found hair volume to

be a weak predictor of hair quality when comparing the performance of hairs from a single individual, where the higher volume hair outperformed the lower volume hair in only 44% of comparisons (Figure 3.8). Furthermore, across individuals, we found a weak correlation between volume and the amount of library molecules generated from the resulting DNA extraction (Figure 3.9). Without a simple physical characteristic to aid in sample evaluation, conservative wet lab approaches should be used to maximize data generation from the worst samples. The implementation of a DNA extraction method that retains the smallest fragments (Figure 3.2) in tandem with a single-stranded DNA library preparation method to efficiently convert the recovered DNA is essential when processing hair shafts. Conservative processing approaches generate libraries with more uninformative content, but post amplification removal of fragments <30bp may be an effective way to reduce sequencing costs.

The hairs processed in this study represent the best-case scenario, and casework samples will present additional challenges. DNA in hair rapidly degrades during keratinization and has been shown to continue to degrade while still associated with an individual, as shown in hair segment analysis [3]. Additionally, DNA will continue to fragment over time and may accumulate base damage and contaminants. Fresh hair shafts appear to be a sample type capable of generating multi-fold genome coverage when using appropriate methods, but aged hair may fail to deliver the same results or be too costly to sequence.

Whole genome shotgun sequencing of hair samples is prohibitively expensive, which will prevent widespread adoption of the technique within forensics.

Disregarding sample preparation cost and expertise, the cost of deep sequencing represents a key barrier to adoption. The cost of sequencing 300M reads can vary from less than $500 to greater than $1200 based on factors such as sequencing center, machine choice, and sample number. Furthermore, we have shown that more than 300M reads are often required to generate multi-fold coverage for fresh hair shafts (Figure 3.4). The development and testing of hybridization enrichment panels to selectively sequence the most informative genomic regions for identification will be critical for the cost-effective processing of casework hair samples using NGS approaches.

This work highlights the potential for single hairs to be used in forensics and other fields to generate whole genomes. We have shown most fresh hair shafts contain enough informative content to generate multi-fold genome coverage. In addition to human hair, hair should be broadly considered as a non-invasive sample type in fields working with mammals, such as conservation genomics. Hair can be opportunistically or deliberately collected in the field using hair snags and trimmings [46]. We believe this study highlights the relevance of hair shafts as a sample type to generate whole genomes.

3.5 References

1. Pfeiffer H, Hühne J, Ortmann C, Waterkamp K, Brinkmann B. Mitochondrial DNA typing from human axillary, pubic and head hair shafts - success rates and sequence comparisons. International Journal of Legal Medicine. 1999;112:287–90.

2. Bender K, Schneider PM. Validation and casework testing of the BioPlex-11 for STR typing of telogen hair roots. Forensic Science International. 2006;161:52–9.

3. Brandhagen MD, Loreille O, Irwin JA. Fragmented Nuclear DNA is the Predominant Genetic Material in Human Hair Shafts. Genes. 2018;9:640.

4. Gilbert MTP, Menez L, Janaway RC, Tobin DJ, Cooper A, Wilson AS. Resistance of degraded hair shafts to contaminant DNA. Forensic Science International. 2006;156:208–12.

5. Bengtsson CF, Olsen ME, Brandt LØ, Bertelsen MF, Willerslev E, Tobin DJ, et al. DNA from keratinous tissue. Part I: Hair and nail. Annals of Anatomy - Anatomischer Anzeiger. 2012;194:17–25.

6. McNevin D, Wilson-Wilde L, Robertson J, Kyd J, Lennard C. Short tandem repeat (STR) genotyping of keratinised hair. Forensic Science International. 2005;153:237–46.

7. Fischer H, Szabo S, Scherz J, Jaeger K, Rossiter H, Buchberger M, et al. Essential Role of the Keratinocyte-Specific Endonuclease DNase1L2 in the Removal of Nuclear DNA from Hair and Nails. Journal of Investigative Dermatology. 2011;131:1208–15.

8. Gill P, Jeffreys AJ, Werrett DJ. Forensic application of DNA 'fingerprints.' Nature. 1985;318:577–9.

9. Jobling MA, Gill P. Encoded evidence: DNA in forensic analysis. Nat Rev Genet. 2004;5:739–51.

10. Butler JM. The future of forensic DNA analysis. Phil Trans R Soc B. 2015;370:20140252.

11. CODIS - NDIS Statistics. Federal Bureau of Investigation. https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics. Accessed 23 Jul 2022.

12. Graham EAM. DNA reviews: hair. Forens Sci Med Pathol. 2007;3:133–7.

13. Ottens R, Taylor D, Abarno D, Linacre A. Successful direct amplification of nuclear markers from a single hair follicle. Forensic Sci Med Pathol. 2013;9:238–43.

14. Hill CR, Kline MC, Coble MD, Butler JM. Characterization of 26 MiniSTR Loci for Improved Analysis of Degraded DNA Samples: 26 MINISTR LOCI. Journal of Forensic Sciences. 2007;53:73–80.

15. Müller K, Klein R, Miltner E, Wiegand P. Improved STR typing of telogen hair root and hair shaft DNA. Electrophoresis. 2007;28:2835–42.

16. Kling D, Phillips C, Kennett D, Tillmar A. Investigative genetic genealogy: Current methods, knowledge and practice. Forensic Science International: Genetics. 2021;52:102474.

17. Kennett D. Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. Forensic Science International. 2019;301:107–17.

18. Orlando L, Allaby R, Skoglund P, Der Sarkissian C, Stockhammer PW, Ávila-Arcos MC, et al. Ancient DNA analysis. Nat Rev Methods Primers. 2021;1:14.

19. Hui R, D'Atanasio E, Cassidy LM, Scheib CL, Kivisild T. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. Sci Rep. 2020;10:18542.

20. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021;53:120–6.

21. Sousa da Mota B, Rubinacci S, Cruz Dávalos DI, Amorim CEG, Sikora M, Johannsen NN, et al. Imputation of ancient genomes. preprint. Genomics; 2022.

22. Basler N, Xenikoudakis G, Westbury MV, Song L, Sheng G, Barlow A. Reduction of the contaminant fraction of DNA obtained from an ancient giant panda bone. BMC Res Notes. 2017;10:754.

23. Boessenkool S, Hanghøj K, Nistelberger HM, Der Sarkissian C, Gondek AT, Orlando L, et al. Combining bleach and mild predigestion improves ancient DNA recovery from bones. Mol Ecol Resour. 2017;17:742–51.

24. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci USA. 2013;110:15758–63.

25. Glocke I, Meyer M. Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. Genome Res. 2017;27:1230–7.

26. Rohland N, Glocke I, Aximu-Petri A, Meyer M. Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. Nat Protoc. 2018;13:2447–61.

27. Gansauge M-T, Gerber T, Glocke I, Korlević P, Lippik L, Nagel S, et al. Single-stranded DNA library preparation from highly degraded DNA using *T4* DNA ligase. Nucleic Acids Res. 2017;:gkx033.

28. Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding MHS, Samaniego JA, et al. Single-tube library preparation for degraded DNA. Methods Ecol Evol. 2018;9:410–9.

29. Kapp JD, Green RE, Shapiro B. A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. Journal of Heredity. 2021;112:241–9.

30. Bennett EA, Massilani D, Lizzo G, Daligault J, Geigl E-M, Grange T. Library construction for ancient genomics: Single strand or double strand? BioTechniques. 2014;56:289–300.

31. Wales N, Carøe C, Sandoval-Velasco M, Gamba C, Barnett R, Samaniego JA, et al. New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA. BioTechniques. 2015;59:368–71.

32. Gilbert MTP, Tomsho LP, Rendulic S, Packard M, Drautz DI, Sher A, et al. Whole-Genome Shotgun Sequencing of Mitochondria from Ancient Hair Shafts. Science. 2007;317:1927–30.

33. Gilbert MTP, Kivisild T, Grønnow B, Andersen PK, Metspalu E, Reidla M, et al. Paleo-Eskimo mtDNA Genome Reveals Matrilineal Discontinuity in Greenland. Science. 2008;320:1787–9.

34. Tobler R, Rohrlach A, Soubrier J, Bover P, Llamas B, Tuke J, et al. Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. Nature. 2017;544:180–4.

35. Meachen J, Wooller MJ, Barst BD, Funck J, Crann C, Heath J, et al. A mummified Pleistocene gray wolf pup. Current Biology. 2020;30:R1467–8.

36. Andreeva T, Manakhov A, Kunizheva S, Rogaev E. Genetic Evidence of Authenticity of a Hair Shaft Relic from the Portrait of Tsesarevich Alexei, Son of the Last Russian Emperor. Biochemistry Moscow. 2021;86:1572–8.

37. Loreille O, Tillmar A, Brandhagen MD, Otterstatter L, Irwin JA. Improved DNA Extraction and Illumina Sequencing of DNA Recovered from Aged Rootless Hair Shafts Found in Relics Associated with the Romanov Family. Genes. 2022;13:202.

38. Rohland N, Reich D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res. 2012;22:939–46.

39. Gansauge M-T, Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc. 2013;8:737–48.

40. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Research. 2012;40:e3–e3.

41. John JS. jstjohn/SeqPrep. C. 2022.

42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

44. Daley T, Smith AD. Predicting the molecular complexity of sequencing libraries. Nat Methods. 2013;10:325–7.

45. Melton T, Dimick G, Higgins B, Lindstrom L, Nelson K. Forensic mitochondrial DNA analysis of 691 casework hairs. J Forensic Sci. 2005;50:73–80.

46. Long RA, MacKay P, Ray J, Zielinski W. Noninvasive Survey Methods for Carnivores. Island Press; 2012.

Synthesis

Fields working with samples yielding poor-quality DNA have benefited from advances in short-read sequencing technology but have relied on methodological advances within the field to effectively take advantage of the increased sequencing throughput. Most samples processed for sequencing contain high quantities of high-quality DNA, so methods are optimized for those characteristics. Conversely, many poor-quality samples yield small quantities of short and damaged fragments, while also being physically limited in sample availability. To maximize data generation, researchers working with degraded sample types have developed a suite of high-efficiency methods to recover and then process small and damaged molecules for sequencing. However, the highest-performing methods have typically been more costly and laborious, which results in researchers choosing between performance and usability.

In this dissertation, I have introduced a rapid and efficient single-stranded DNA library preparation method (hereafter, SCR) to prepare highly fragmented and damaged DNA for sequencing. I have presented applications to cell-free DNA (Chapter 1), ancient DNA isolated from bone samples (Chapter 2), and DNA isolated from hair shafts (Chapter 3), but the SCR can be broadly applied to a variety of degraded samples. DNA extracted from sediment cores (eDNA), formalin-fixed and paraffin-embedded (FFPE) tissue, museum specimens, and any improperly preserved sample type will yield degraded DNA appropriate for the SCR. Compared to existing commercial and field standard approaches, the SCR produces similar or higher quality libraries while

being less costly, less time-consuming, and higher throughput. The SCR allows researchers to adopt an affordable and fast single-stranded library preparation method without sacrificing performance, to process a higher number of samples while generating more data from each sample.

The SCR ligates adapters without prior manipulation of the termini to both ends of the native DNA, which produces a library that better reflects the DNA molecules isolated from degraded samples. By not manipulating the native end of DNA the SCR allows for accurate analysis of nucleosome positions and fragment end frequencies from cell-free DNA samples [1]. Furthermore, when library efficiency and sequencing depth are high the reconstruction of double-stranded fragments from single-stranded libraries can be performed to explore DNA damage patterns [2].

Further development of the SCR will simplify method adoption, improve library quality from the worst samples, and provide new molecular tools to researchers. Single-stranded DNA contamination among adapter and splint oligonucleotides is variable, which increases costs and ultimately results in method development compromises. Increasing oligonucleotide purity, either within the manufacturing process or after receiving the oligonucleotides, will allow for re-optimization of the adapters to reduce library artifacts and increase library efficiency across a wider range of DNA inputs. Finally, the SCR ligation approach can be applied to non-shotgun sequencing approaches for the affordable library preparation of DNA templates with defined ends, such as amplicons for viral genome sequencing.

In Chapter 3, I used an optimized version of the SCR to prepare libraries from picogram scale DNA extractions to characterize the DNA recovered from the hair shafts of 50 anonymous volunteers. Before this work, deep sequencing of single hair shafts from a large panel of individuals using a workflow optimized for degraded DNA had not been performed. I found 96% of hair shafts appear capable of generating multi-fold genome coverage when appropriate methods are applied. The data generated from the hair samples will be used to explore the accuracy of low coverage genotyping by imputation. However, hair shafts appear to be more confidently processed for next-generation sequencing compared to PCR-based short tandem repeat typing, the primary data generation scheme used in forensics. This work shows the potential of single-hair shafts to generate whole genomes in forensics and fields that benefit from non-invasive or opportunistic sample collection. Additional improvements to the efficiency and purity of DNA isolation and library preparation will maximize the success of generating data from the worst preserved samples, such as aged hair shafts. Finally, this wet lab workflow and project design can be used to re-assess other forensic sample types using NGS techniques, such as touch DNA.

References

1. Liu Y. At the dawn: cell-free DNA fragmentomics and gene regulation. Br J Cancer. 2022;126:379–90.

2. Bokelmann L, Glocke I, Meyer M. Reconstructing double-stranded DNA fragments on a single-molecule level reveals patterns of degradation in ancient samples. Genome Res. 2020;30:1449–57.