New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA

Nathan Wales¹, Christian Carøe¹.², Marcela Sandoval-Velasco¹, Cristina Gamba¹, Ross Barnett¹, José Alfredo Samaniego¹, Jazmin Ramos Madrigal¹, Ludovic Orlando¹, and M. Thomas P. Gilbert¹.³

¹Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark, ²Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark, and ³Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Australia

BioTechniques 59:368-371 (December 2015) doi 10.2144/000114364
Keywords: ancient DNA; paleogenomics; DNA library preparation; next-generation sequencing

An innovative single-stranded DNA (ssDNA) library preparation method has sparked great interest among ancient DNA (aDNA) researchers, especially after reports of endogenous DNA content increases >20-fold in some samples. To investigate the behavior of this method, we generated ssDNA and conventional double-stranded DNA (dsDNA) libraries from 23 ancient and historic plant and animal specimens. We found ssDNA library preparation substantially increased endogenous content when dsDNA libraries contained <3% endogenous DNA, but this enrichment is less pronounced when dsDNA preparations successfully recover short endogenous DNA fragments (mean size < 70 bp). Our findings can help researchers determine when to utilize the time- and resource-intensive ssDNA library preparation method.

One of the most promising methodological developments in ancient DNA (aDNA) analysis is a single-stranded DNA (ssDNA) library protocol that begins by denaturing double-stranded DNA (dsDNA) and subsequently converts all ssDNA templates into a sequencing library (1). This technique incorporates damaged molecules, including those with abasic sites, as well as short molecules frequently lost in dsDNA library preparation. While ssDNA preparation is more laborious and expensive than dsDNA protocols (Table 1), it has been integral in a number of aDNA and historic DNA studies, including the Neanderthal and Denisovan genome projects (2–6).

In two studies (2,5), researchers constructed ssDNA and dsDNA libraries from aliquots of the same DNA extract and determined that ssDNA libraries were more complex, enabling deeper sequencing from the same volume of starting material. In one of these studies, Prüfer et al. (5) report slightly higher endogenous content in Altai Neanderthal ssDNA libraries than in a corresponding dsDNA library. Even though both ssDNA and dsDNA libraries of the Altai Neanderthal had unusually high endogenous DNA content (~70%), for archaeological samples with more commonly observed endogenous levels (often < 1%) (7), a subtle change in absolute endogenous content—for example from <1% to ~3%—would correspond to a sizable proportional gain and could thereby make ssDNA preparation advantageous for many aDNA projects. To explore this and other issues, Bennett et al. (8) constructed libraries from 16 ancient and modern samples, finding substantial fold-increases (up to 25.9) in endogenous content with the ssDNA protocol. However, these increases did not follow a consistent pattern, and in some instances ssDNA libraries yielded considerably lower endogenous content.

In order to clarify when the ssDNA protocol might be worth undertaking, we prepared 23 corresponding dsDNA and ssDNA libraries from a range of samples: Late Pleistocene horses (Equus spp.), woolly rhinoceros (Coelodonta antiquitatis), saber-toothed cat (Smilodon populator), archeological human remains, archaeobotanical maize (Zea mays) and grapes (Vitis vinifera), and modern lion (Panthera leo) museum collections (Supplementary Tables S1 and S2). Libraries were sequenced on an Illumina HiSeq 2500 (Illumina, San Diego, CA), and
sequencing efforts were normalized across samples before mapping to reference genomes.

Endogenous DNA content of the corresponding libraries had a linear relationship ($R^2 = 0.996$) with slightly more ssDNA reads mapping to the respective genome (slope of line = 1.009, median fold-change = 1.10) (Figure 1, Supplementary Table S3). This increase was found to be statistically significant although the sample sizes were relatively small (mean increase in endogenous reads = 1.17%, paired t-test = -2.517, df = 22, $P = 0.020$). Factoring in the shorter length of endogenous DNA in ssDNA libraries (Supplementary Figure S1), the amount of bioinformatically informative data dropped appreciably: ssDNA preparation did not yield more endogenous nucleotides in 14 of the 23 libraries. Fold-increases in endogenous nucleotides were only observed when dsDNA libraries contained <3% endogenous DNA, suggesting that a threshold exists where ssDNA library preparation tends to enrich endogenous content. The precise value of this threshold is unclear, however, because only 1 sample yielded endogenous content in the 3%–25% range. In comparison to the large enrichments in endogenous content observed by Bennett et al. (8), fold-increases in endogenous nucleotides mostly ranged between 1.2–2.2,
with a single outlier with a fold-change of 5.0. It is also noteworthy that 2 dsDNA libraries with <2% endogenous content did not benefit from ssDNA library preparation (fold-changes of 0.87 and 0.88). We note that sample-specific degradation patterns likely play a significant role in the relative success of the library preparation strategies. For example, archeological skeletal remains studied by Bennett et al. (8) originated from warmer locations than most of our samples, where bones may have been exposed to more microbial action soon after the death of the animal, increasing DNA fragmentation and the relative amount of environmental contamination compared with frozen or desiccated tissues. Furthermore, in samples where endogenous DNA is more fragmented than exogenous DNA, the ssDNA protocol is expected to recover the shorter endogenous DNA molecules and be biased against longer contaminant molecules that may not denature at the beginning of the protocol, preventing their conversion to library molecules.

While the average length of mapped reads is nearly always shorter in our ssDNA libraries than in our dsDNA libraries, the difference is far less than in Bennett et al. (10 bp versus 29 bp) (Supplementary Table S3, Supplementary Figure S1) (8). The similarity between our dsDNA and ssDNA libraries in read length sheds light on why the fold-increases observed here are not as impressive as those in the aforementioned study. We suggest that when dsDNA library preparations yield endogenous DNA with a mean length <70 bp, it is unlikely that ssDNA library preparation will increase the endogenous content more than 2–3-fold in most cases (Figure 1). Conversely, if dsDNA libraries contain <0.5% endogenous DNA with mean read lengths of 70–100 bp, large fold-increases may be obtained with ssDNA preparations, as in Bennett et al. (8). Therefore, optimization of dsDNA protocols may be more cost-effective than ssDNA preparations, and if the primary objective is to enrich endogenous DNA, other methodological approaches may be preferable, such as pre-digestion or targeted capture (both reviewed in Reference 9).

As noted in other studies, ssDNA libraries are biased toward short DNA fragments, and a large proportion of the sequencing capacity is wasted if amplified libraries are not size-fractionated before sequencing (2,5,8). Here, an average of 15.1% of ssDNA sequencing reads were <28 bp (range: 4.4%–36.3%) compared with only 2.2% for dsDNA libraries (range: 0.2%–8.3%) (Supplementary Figures S2 and S3). High proportions of molecules that are too short to be bioinformatically useful may undermine efforts to enrich endogenous content (Supplementary Figure S4) unless size fractionation is performed.

By certain criteria, ssDNA libraries should outperform dsDNA libraries, especially in terms of library complexity. At a normalized sequencing level of 7 million reads, PCR clonality levels are low (mean = 2.4%), and neither preparation has a consistent advantage (Supplementary Figure S5). However, based on estimates from preseq (10), 81% of ssDNA libraries would yield more unique endogenous reads than the corresponding dsDNA libraries at a sequencing depth of 200 million reads (Supplementary Figure S6), which may provide substantial increases in depth of coverage for full genome sequencing projects or targeted capture experiments. Also, consistent with previous findings (2,5), we observe ssDNA libraries are less biased toward high GC-content than the paired dsDNA libraries (Supplementary Figure S7). This is an important consideration for heavily degraded samples (mean fragment length < 50 bp), where dsDNA library preparation will underrepresent AT-rich regions, yielding less homogenous genome coverage than ssDNA libraries.

Like Bennett et al. (8), we observe that ssDNA library preparation can yield fold-enrichments in endogenous content, particularly in samples with low initial endogenous content. In contrast to previous findings of endogenous content enrichments of >20-fold, we suggest that much lower enrichment levels, on the order of 2–5-fold are more common, especially if dsDNA library preparations yield endogenous DNA with a mean length <70 bp. Given our findings and those of other studies, we propose using the guidelines listed in Table 1 for determining the types of aDNA studies that are most likely to benefit from the available library preparation protocols and recommend that most ancient samples should undergo a preliminary screening with cost-effective conventional dsDNA library-building protocols before ssDNA preparations are considered.

Author contributions
N.W. and M.T.P.G. planned the study. N.W., C.C., M.S.V., C.G., and R.B. performed experiments. N.W., J.A.S., and J.R.M. analyzed the data. N.W. wrote the manuscript with assistance from all other authors.

Acknowledgments
This work was supported by the Lundbeck Foundation (R52-A5062), the Danish Council for Independent Research (10-081390), and the Danish National Research Foundation (DNRF94). C.G. was supported by FP7-IEF-328024, M.S.V. was funded by Marie Curie Actions FP7 People grant 290344 ‘EUROTAST’, and R.B. was supported by Marie Curie Actions Intra-European Fellowship FP7-PEOPLE-2011-IEF-298820. The authors thank collaborators who provided samples for methodological testing: Chinese mumified human remains from V. H. Mair, University of Pennsylvania; grape samples from L. Bouby, CNRS, Centre de Bio-Archéologie et d’Ecologie; equid samples from J. Weinstock, University of Southampton; museum lion samples from J. Cuisin, F. Renoud, and D. Robineau, Muséum National d’Histoire Naturelle, Paris; maize samples from B. D. Smith, National Museum of Natural History, Smithsonian Institution; Smilodon sample from R. van Zeist and C. Pepermans, Naturalis, Leiden; and St. Helena human remains from K. Robson Brown, Department of Archaeology and Anthropology, University of Bristol, and A. Pearson, Pearson Archaeology Ltd.. We acknowledge the Danish National High-throughput Sequencing Centre for assistance in generating the Illumina data. Many thanks to Maria C. Ávila-Arcos, Shyam Gopalakrishnan, Inge K. C. Lundstrom, Fernando Peñaloza Figueroa, Hannes Schroeder, and...
Mikkel Schubert for discussions about library preparation and programming advice. We thank E. Andrew Bennett, Diyendo Massilani, Eva-Maria Geigl, and Thierry Grange for fruitful discussions about their experiments on ssDNA preparation. Special thanks to Timothy Daley and Andrew Smith, University of Southern California, for providing for suggestions regarding the preseq software package. The authors also thank the three anonymous reviewers who provided insights and recommendations on this study.

Competing interests
The authors declare no competing interests.

References

Received 11 December 2014; accepted 09 September 2015.

Address correspondence to Nathan Wales, Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7 1350 Copenhagen K, Denmark. E-mail: nathan.wales@snm.ku.dk

To purchase reprints of this article, contact: biotechniques@fosterprinting.com