Optimization of 454 sequencing library preparation from small amounts of DNA permits sequence determination of both DNA strands

Tomislav Maricic and Svante Pääbo

BioTechniques 46:51-57 (January 2009)
doi 10.2144/000113042

To increase the yield of DNA sequence generated by the 454 technology from small amounts of starting DNA, we investigated the efficiency of each step in the 454 library preparation process. We find that the last step, when the single-stranded library is released by NaOH, is inefficient and highly variable. When this step is replaced with heat treatment, library amounts dramatically increase. Furthermore, when sequencing templates are first isolated by NaOH treatment and subsequently by heat treatment, the sequences of both strands of individual template DNA molecules can be determined. Using this approach, we confirm that C/G base pairs observed as T/A base pairs in Neanderthal DNA sequences are due to a modification of the cytosine rather than guanine residues.

INTRODUCTION

Sequencing by synthesis on the 454 platform is one of several novel sequencing techniques that allows the determination of millions of base pairs per run using a single instrument (1). Although in most applications of this technology the amount of template DNA is not limiting, a number of applications start from small quantities of DNA or RNA. For example, when bacteria that cannot be cultured (2) or when cDNA libraries from a small number of cells (3) are sequenced, template DNA amounts limit the number of DNA sequences that can be determined. The same is true for many palaeontological samples, such as Neanderthal remains (4), where the amount and quality of available bone limits the amount of DNA sequence that can be generated. Furthermore, although some sequencing methods use both strands of template molecules as starting material, others use just one strand under the assumption that the two strands are complementary to each other. This may not always be the case. For example, when a molecule is hemimethylated (5,6), when bases are modified as is often the case with ancient DNA (7), or in other circumstances, DNA strands may not be complementary to each other.

To increase DNA sequence recovery from small amounts of biological or paleontological material, we studied the template recovery at each step in the 454 library production protocol. We find that although the efficiency of recovery in most steps varies between 40% and 96%, the final step—where the single-stranded library is retrieved—is highly variable and can entail losses of over 99%. Replacing this step with more efficient means of library recovery increases the library yield 5- to 200-fold. In addition, by combining two methods of retrieval of 454 library molecules, we show that one can retrieve both strands of the same ancient DNA molecules and thus study the occurrence of miscoding lesions by directly comparing the sequences of the two strands of ancient DNA molecules.

MATERIALS AND METHODS

Radioactive labeling of a PCR product

A PCR product 103 bp in length (Supplementary Figure 1, available at www.BioTechniques.com) was used as a template in two PCRs where in one only non-radioactive dATP (2.4 µM) was added, while in the other dATP (2.0 µM) and [α-^{32}P]dATP (0.4 µM, 3000 Ci/mmol; Hartmann Analytic, Braunschweig, Germany) were used essentially as described in Reference 8. A final reaction volume of 50 µL contained $MgCl_2$ (5 mM; Applied Biosystems, Foster City, CA, USA), Gold buffer (1×; Applied Biosystems), BSA (0.8 mg/mL; Sigma Aldrich, St. Louis, MO, USA), L164 primer (1 µM; Metabion, Martinsried, Germany), H221 primer (1 µM; Metabion), AmpliTaq Gold DNA Polymerase (2.5 U; Applied Biosystems), dTTP (200 µM; New England Biolabs, Ipswitch, MA, USA), dCTP (200 µM; New England Biolabs), and dGTP (200 µM; New

Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

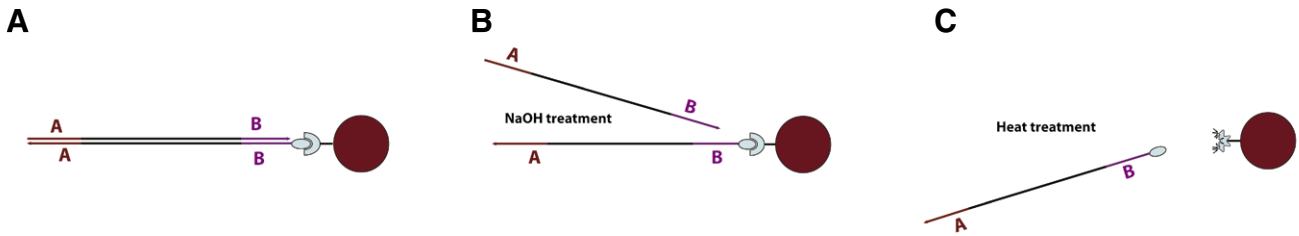

Figure 1. Interaction between 454 library molecules and streptavidin-coated beads. (A) Biotin-containing double-stranded library molecules are immobilized on streptavidin-coated beads. (B) The strand not containing biotin is released by NaOH treatment. (C) The streptavidin-biotin interaction can be broken by heat treatment.

England Biolabs). After an initial incubation at 94°C for 9 min, 35 cycles at 94°C for 20 s, 53°C for 30 s, and 72°C for 30 s were performed. Both PCR products were purified using the MinElute PCR purification kit (Qiagen, Hilden, Germany); the non-radioactive PCR product was visualized in an agarose gel (Supplementary Figure 2, available at www.BioTechniques.com), and its concentration determined to be 14.8 ng/µL by absorbance (NanoDrop ND-1000 spectrometer, Thermo Scientific, Wilmington, DE, USA). This concentration was used as a proxy measure for the concentration of the radioactive PCR product.

454 library optimization

51.8 nanograms of the 103-bp radioactively labeled PCR product were used as the template for the parallel production of six 454 libraries using the 454 Life Sciences GS20 protocol (Branford, CT, USA) (see Supplementary Table 1, available at www.BioTechniques.com) but the initial nebulization step was omitted. In the final step—isolation of the single-stranded DNA library—three libraries were treated twice with 0.125 M NaOH according to the manufacturer's protocol and three were eluted in 50 µL of water for 2 min at 90°C, and then transferred to ice. Aliquots were collected from each fraction in each step, and then transferred to 20-mL scintillation vials (Carl Roth, Karlsruhe, Germany) containing 10 mL water. Their radiation content was measured (for more information on Cerenekov counting method, see las.perkinelmer.com/Content/ApplicationNotes/APP_TriCarbCerenkovCntingP32.pdf) with a scintillation counter (Packard Tri-Carb, Packard BioScience, Meriden, CT, USA) for 1 min. From the Cerenkov radiation of each aliquot, the total radiation for each fraction was calculated (for details see Supplementary Tables 2–4, available at www.BioTechniques.com).

DNA extracts from six different Neanderthal bones from Vindija Cave, Croatia, which may or may not have come from different individuals, were prepared (9) and used to generate 454 libraries using the manufacturer's protocol (without the nebulization step). After the two NaOH treatments, the streptavidin-coated beads were washed with 150 µL of 1× binding and washing buffer (B&W; 5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl) buffer and incubated for 2 min at 90°C in 45 µL of water. Tubes were transferred to a +4°C block, beads were pelleted using a magnetic particle concentrator, and the eluate was removed and mixed with 5 µL of 10× TE buffer. Concentrations of library molecules were measured using quantitative PCR as described in Reference 10.

Table 1. Numbers of Molecules in NaOH- and Heat-treated Neanderthal Libraries

Library	NaOH treatment	Heat treatment	Heat/NaOH
Neanderthal 1	5.9×10^5	1.6×10^8	264
Neanderthal 2	1.2×10^7	1.4×10^9	109
Neanderthal 3	3.6×10^7	1.5×10^9	40
Neanderthal 4	2.3×10^8	1.4×10^9	6
Neanderthal 5	5.7×10^7	4.2×10^8	7
Neanderthal 6	1.9×10^8	5.7×10^9	30

Retrieving both strands of Neanderthal molecules

DNA was extracted from five samples of a single Neanderthal bone discovered in Vindija Cave, Croatia, as described in Reference 9, and 454 libraries were made using the manufacturer's protocol with NaOH treatment. Two and three of these libraries were mixed to form pools 1 and 2, respectively. After NaOH treatment, the capture beads used for the preparation of the libraries in pool 1 were heat-treated, and the eluates were used to create pool 3. All pools were sequenced on the 454 GS FLX (454 Life Sciences) instrument with quality filters adjusted to allow for shorter sequences (11). Raw sequences were clustered according to sequence similarity (more than 90% identity over the complete read), and all clusters were reduced to a single sequence (11). The remaining singleton sequences were aligned to the reference genomes for human, chimpanzee, rhesus, and mouse, as well as GenBank "nt" and GenBank "env," using NCBI Mega BLAST with the word size (-W) reduced to 16 (11). For sequences where the best local alignment was to human and where the alignment was placed uniquely on the human genome, a semi-global alignment to that region of the reference was performed (11). About 10,000 such Neanderthal sequences were obtained from each of the two NaOH-treated pools 1 and 2 (standard protocol), and about 900,000 from heat-treated pool 3.

By comparing the genomic location of the semi-global alignments, pairs

of sequences from different libraries were selected that have the exact same start and end positions in the reference sequence. From the two related library pools (pools 1 and 3), 308 such pairs were obtained, while four were found when comparing the two unrelated library pools (pools 2 and 3). For the 308 pairs, the two semi-global alignments were merged into a single alignment and manually verified, whereby two were excluded since the correct alignment was not obvious. The remaining 306 pairs were aligned to the reference sequence and positions in which at least one sequence differed from the others were counted.

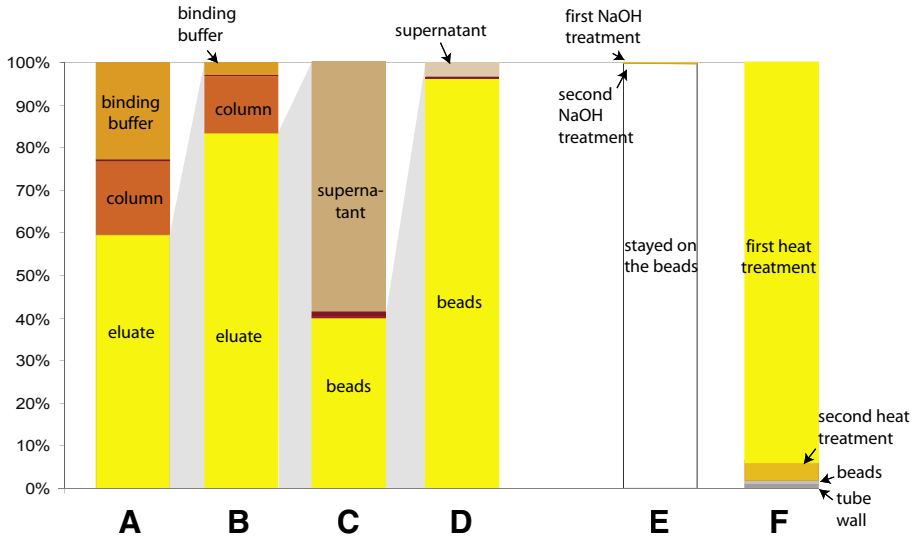

Figure 2. Retrieval of template DNA through the 454 library preparation process. (A–D) The first four steps in the 454 library preparation process: (A) end polishing, (B) adapter ligation, (C) library immobilization, and (D) fill-in reaction. Bars indicate the average amount of template present during each step as a percentage of the total amount of radioactivity entering a particular step. The fractions containing the relevant library molecules are indicated in yellow. Less than 1% of radioactivity ended in each washing buffer [thin red bars in (A–D)]. (E) shows isolation of the 454 library from the beads by two consecutive NaOH melting steps while (F) shows two consecutive heating steps (for details, see Supplementary Tables 2–4, available at www.BioTechniques.com).

RESULTS AND DISCUSSION

Library recovery

The 454 library preparation process consists of five steps (1): (Step 1) end polishing by *T4* DNA polymerase and *T4* polynucleotide kinase; (step 2) ligation using one non-biotinylated adaptor (A adaptor) and one biotinylated adaptor (B adaptor); (step 3) library immobilization; (step 4) fill-in reaction by *Bst* DNA polymerase; and (step 5) isolation of the single-stranded DNA library. After steps 1 and 2, the reaction products are isolated by adsorption to a solid matrix and subsequent elution (MinElute). In step 3, molecules that carry at least one biotinylated B adaptor are immobilized on streptavidin-coated beads (Figure 1A) on which the fill-in reaction (step 4) is performed. In step 5, non-biotinylated strands are separated from the beads by two consecutive incubations in 0.125 M NaOH (Figure 1B). Such molecules will carry an A adaptor on their 5′ ends and B adaptor sequences on their 3′ ends ("A-B strands"), whereas the complementary strands, which carry B adaptors on their 5′ ends and A adaptors on their 3′ ends ("B-A strands") will remain on the beads and be discarded.

To analyze the recovery of DNA in each step of the 454 library preparation process when small amounts of starting DNA is used, we used ~50 ng of a [α-^{32}P]dATP-labeled 103-bp-long DNA fragment to prepare six 454 libraries. For each step in library production, we measured the DNA recovery. In steps 1 and 2 of MinElute purification, an average of 60% and 84%, respectively, of the radioactively labeled DNA was retrieved, while in step 3, where 75% of molecules are expected to carry biotinylated adaptors, 40% (i.e., 53% of what is expected) were bound to the beads (yellow bars in Figure 2, A–C). In step 4, after the fill-in reaction and additional washings, 96% of the molecules stayed on the beads (yellow bar in Figure 2D). In the step 5, when three of the six libraries were eluted by two incubations in NaOH, less than 0.1% of the labeled DNA present on the streptavidin-coated beads was retrieved (first and second NaOH-treatment bars on Figure 2E). Thus, while losses are within what is reasonable to expect in steps 1–4, in step 5—where the single-stranded, adaptor-ligated library is released from the streptavidin beads—losses are immense. When we replaced the NaOH treatment with incubations at 90°C, recovery increased to 98% (Figure 2F).

To investigate if the heating step improves the yield in library production from DNA extracts from Neanderthal remains, DNA extracts from six different Neanderthal bones were used for library preparation. Each library was first treated twice with

NaOH according to the standard 454 protocol. Subsequently, the remaining streptavidin-coated magnetic beads were incubated in water at 90°C. The number of DNA molecules in the NaOH- and heat-treated libraries were determined by quantitative PCR using primers to the A and B adaptors (10). Six to 264 times more DNA molecules were recovered by heat treatment than by the previous NaOH treatment of the same beads (Table 1). Together with the high efficiency (98%) of heat treatment seen in the radioactive experiment (Figure 2F) this suggests that NaOH melting is not only inefficient but also highly variable.

Heat treatment differs from NaOH treatment in that it is expected not only to melt the DNA but also to disrupt the biotin-streptavidin interaction (12) and denature the streptavidin molecules (13), releasing not only DNA strands that do not contain biotinylated adaptors, but also the biotinylated strands. Thus, heat-treated libraries will contain biotinylated B-A strands as well as molecules carrying two B adaptors in addition to the non-biotinylated A-B strands. The B-A molecules will improve the possibility of exhaustively sequencing DNA molecules in a small sample while the B-B molecules will not impede sequencing since they will not be amplified in the emulsion PCR that constitutes the first step in the 454 sequencing process (data not shown).

Sequencing of complementary DNA strands

A-B strands, as well as B-A strands, can serve as templates in emulsion PCR and subsequent 454 sequencing. Therefore, it should be possible to determine the sequences of both strands of individual template molecules if one first isolates single-stranded A-B molecules from a library by NaOH treatment (Figure 1B) and subsequently isolates the complementary B-A strands by heat treatment of the remaining beads (as well as A-B strands remaining after the NaOH treatment) (Figure 1C). When differences to a reference sequence are observed as complementary base pairs in both strands of the library molecules, this serves to confirm such positions in the primary sequence. By contrast, when a difference to the reference sequence is observed on only one strand and, consequently, a non-standard base pair is observed in the library molecule, this indicates that some modification may affect just one of the two nucleotides that make up the base pairs in the library molecule, or, alternatively, that a sequencing error affects one of the sequences.

To test if this works, we searched for pairs of sequences from Neanderthal libraries where one sequence was recovered by NaOH treatment while the other was recovered by heat treatment and where both sequences started and ended at the same place when aligned to the human genome. Among ~10,000 Neanderthal sequences recovered by NaOH treatment, 306 corresponding sequences recovered by heat treatment were found. When aligned to the human reference sequence, these 306 Neanderthal sequences yielded 17,161 positions without gaps in any of the three sequences compared. At these positions, two types of differences to the human genome sequence were observed. Firstly, there were the 213 differences where NaOH-treated and heat-treated strands contained complementary base pairs that differed from the reference sequence. Secondly, there were the 73 differences where the two strands were not complementary to each other and thus only one of them differed from the reference sequence (Supplementary Figure 3, available at www.BioTechniques.com). In no cases were all three sequences different from each other.

Among the former 213 cases where the two Neanderthal DNA strands contained complementary bases that differed from the reference sequence, two substitutions stand out in frequency: those where C and G in the reference sequence are T and A, respectively, in the Neanderthal sequences (Figure 3). Both are of similar frequency (92 and 96 cases, respectively), ~10 times more frequent than the other two transitions, and ~20 times more frequent than the transversions. When C-to-T and G-to-A substitutions are plotted with respect to where they occur in the Neanderthal sequences, C-to-T substitutions appear predominantly in the first six bases from the 5′ ends of the molecules (Figure 4A), while G-to-A substitutions appear primarily in the six bases from the 3′ ends of the molecules (Figure 4B).

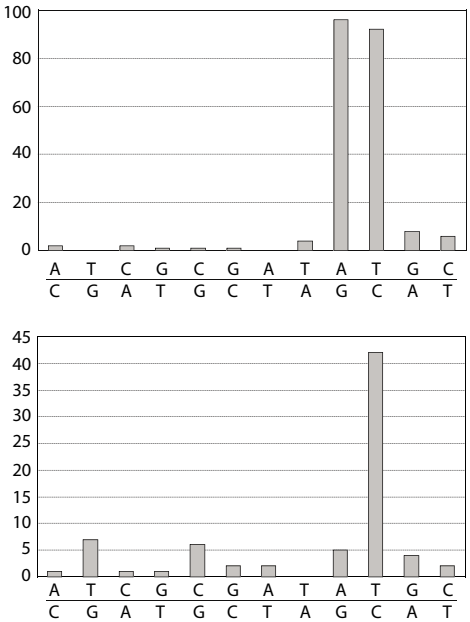

Figure 3. Differences to the human genome observed in both strands of Neanderthal sequences (above) and in only one strand (below). Human bases are in the lower row, while Neanderthal bases are in the upper row.

The latter 73 cases where the strands sequenced do not contain complementary bases and one is identical to the reference sequence, the discrepancy between the two sequenced strands could be due to either sequencing errors or to modifications that cause a nucleotide misincorporation on one strand of the molecule. Among the 12 possible nucleotide differences, 11 are observed between 0 and 7 times (Figure 3). By contrast, in 42 cases, Cs are seen in the reference sequence as Ts in the Neanderthal sequences. Notably, in only five cases do Gs in the references sequence appear as As in the Neanderthal sequences. When the C-to-T substitutions are plotted with respect to where they occur in the Neanderthal sequences, they appear to occur with approximately even frequencies across the molecules, except for the first and last six bases toward the ends of the molecules, where they are almost absent (Figure 4, C and D).

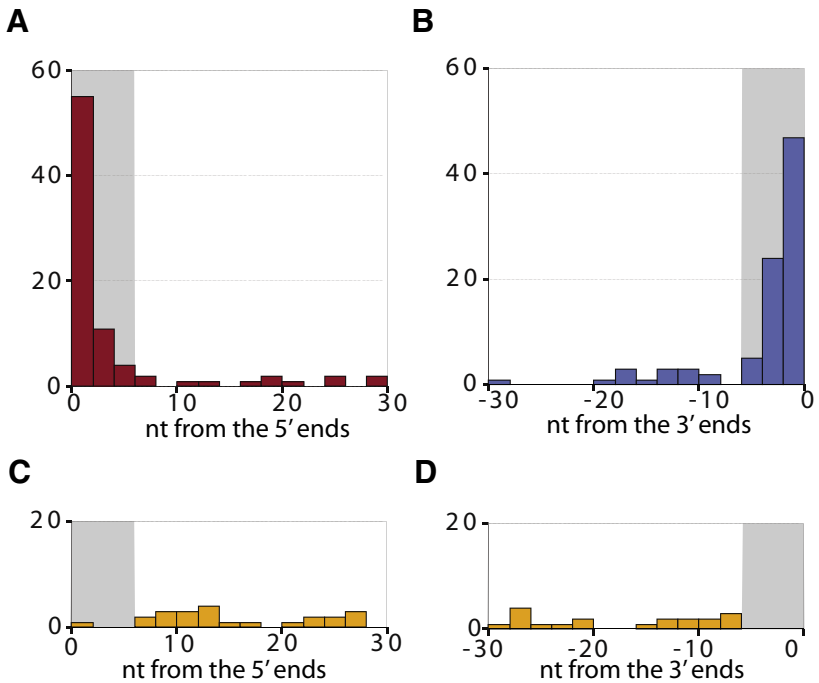

Figure 4. Positions of C-to-T and G-to-A substitutions in Neanderthal DNA molecules. Above, C-to-T substitutions (A) and G-to-A substitutions (B) seen on both NaOH- and heat-treated strands are plotted as a function of their distance from the 5′ and 3′ ends, respectively. Below, C-to-T substitutions on only one strand are plotted as a function of their distance from the 5′ (C) and 3′ ends (D). Gray shading indicates the first and last six bases of Neanderthal DNA strands.

In conclusion, four observations stand out from these experiments. Firstly, the substitutions seen on both strands are dominated by C/G to T/A substitutions. Secondly, whereas the C-to-T substitutions are concentrated toward the 5′ ends of DNA strands, the G-to-A substitutions are concentrated toward the 3′ ends. Thirdly, the substitutions seen on only one strand are dominated by C-to-T substitutions, whereas G-to-A substitutions are not overrepresented. Fourthly, the C-to-T substitutions are distributed randomly, it appears, along the molecule, except toward the ends of the molecules where they are rare.

It has previously been observed that C-to-T substitutions accumulate at the 5′ ends of ancient DNA sequences determined by the 454 technology whereas G-to-A substitutions accumulate at the 3′ ends (11,14). It has also been observed that cytosine deamination is a dominant miscoding lesion in ancient DNA (11,14,15). The accumulation of the two misincorporations at the ends of molecules has been suggested to be due to the fact that the ends of ancient DNA molecules may often be single-stranded (11,14). Since cytosine residues are much more susceptible to deamination in single-stranded than in double-stranded DNA (16), these should accumulate at the ends of molecules (11,14). While C-to-T substitutions are directly observed at the 5′ ends of molecules, they appear as G-to-A substitutions at the 3′ ends, where they derive from deaminated cytosine residues in the 5′-overhanging ends that serve as templates during the fill-in by T4 DNA polymerase during the end polishing used to create blunt ends before adaptor ligation. The current data directly confirms this model proposed by Briggs et al. (11) and Brotherton et al. (14) showing that inside the molecules, differences to the reference genome are largely C-to-T substitutions seen on only one strand (Figure 4, C and D), which demonstrates that cytosine rather than guanine residues are responsible for the misincorporations inside the molecules. Double-stranded differences that occur at the ends of molecules (Figure 4, A and B) support the view that deaminated cytosine residues in overhanging single-stranded ends cause the accumulation of C-to-T and G-to-A substitutions in the 5′ and 3′ ends of ancient molecules, respectively.

For double-stranded DNA, the current data allow the fraction of deaminated cytosine residues in this Neanderthal specimen to be estimated at ~0.97% (95% confidence interval, 0.71–1.32%), in agreement with a previous estimate (11). Since the proportion of 5′-overhanging ends relative to other types of ends is unknown, the rate of cytosine deamination in single-stranded Neanderthal DNA cannot be easily calculated. However, the frequency of C-to-T and G-to-A substitutions is increased in the first and last 5–6 nucleotides of Neanderthal molecules, suggesting that the lengths of 5′-overhanging ends of the molecules extracted from Neanderthal bones vary between 0 and 6 bases. The frequency with which cytosine residues are observed as thymine residues decreases from ~35% at the first position, ~22% at the second position, ~10% at the third and fourth positions, to ~5% in the sixth position (Supplementary Figure 4, available at www.BioTechniques.com), suggesting that the relative lengths of overhanging ends approximately decrease exponentially with length.

Conclusions

The retrieval of 454 libraries from streptavidin-coated beads by heat rather than NaOH increases the sequence yield dramatically when small amounts of template DNA are analyzed. In combination with other improvements, such as quantitative PCR to determine suitable amounts of 454 libraries for emulsion PCR (10), this allows much more efficient use of 454 libraries and thus the application of high-throughput sequencing to samples where only limited amounts of template DNA is available. In addition, the sequential isolation of single-stranded library molecules by NaOH and by heat allows both strands of individual double-stranded molecules to be isolated and separately sequenced. This makes it possible to detect miscoding DNA modifications that affect only one strand in template molecules. By this approach, we confirm that

in Neanderthal DNA fragments, miscoding lesions that cause transitions affect cytosine residues much more often than guanine residues in C-G base pairs.

ACKNOWLEDGMENTS

We especially thank Udo Stenzel for help with the bioinformatics analyses; Adrian Briggs, Hernan Burbano, Ed Green, Michael Hofreiter, Michael Lachman, Johannes Krause, Matthias Meyer, Kay Prüfer, and Susan Ptak for helpful discussions; Victor Wiebe and Esther Lizano for radioactive labeling; Christine Green for comments on the manuscript; Barbara Höber, Barbara Höffner, and Antje Weihmann for sequencing; the Croatian Academy of Sciences and Arts and the Berlin-Brandenburg Academy of Sciences for collaboration on the Neanderthal Genome Project; and the Max Planck Society for financial support.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

1. **Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*:376-380.
2. **Rappe, M.S. and S.J. Giovannoni.** 2003. The uncultured microbial majority. Annu. Rev. Microbiol. *57*:369-394.
3. **Schutze, K. and G. Lahr.** 1998. Identification of expressed genes by laser-mediated manipulation of single cells. Nat. Biotechnol. *16*:737-742.
4. **Green, R.E., J. Krause, S.E. Ptak, A.W. Briggs, M.T. Ronan, J.F. Simons, L. Du, M. Egholm, et al.** 2006. Analysis of one million base pairs of Neanderthal DNA. Nature *444*:330-336.
5. **Clark, S.J., J. Harrison, C.L. Paul, and M. Frommer.** 1994. High sensitivity mapping of methylated cytosines. Nucleic Acids Res. *22*:2990-2997.
6. **Pradhan, S. and G.D. Kim.** 2002. The retinoblastoma gene product interacts with maintenance human DNA (cytosine-5) methyltransferase and modulates its activity. EMBO J. *21*:779-788.
7. **Hofreiter, M., D. Serre, H.N. Poinar, M. Kuch, and S. Paabo.** 2001. Ancient DNA. Nat. Rev. Genet. *2*:353-359.
8. **Sambrook, J. and D.W. Russel.** 2001. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
9. **Rohland, N. and M. Hofreiter.** 2007. Ancient DNA extraction from bones and teeth. Nat. Protocols *2*:1756-1762.
10. **Meyer, M., A.W. Briggs, T. Maricic, B. Hober, B. Hoffner, J. Krause, A. Weihmann, S. Paabo, and M. Hofreiter.** 2008. From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. Nucleic Acids Res. *36*:e5.
11. **Briggs, A.W., U. Stenzel, P.L. Johnson, R.E. Green, J. Kelso, K. Prufer, M. Meyer, J. Krause, et al.** 2007. Patterns of damage in genomic DNA sequences from a Neanderthal. Proc. Natl. Acad. Sci. USA *104*:14616-14621.
12. **Holmberg, A., A. Blomstergren, O. Nord, M. Lukacs, J. Lundeberg, and M. Uhlen.** 2005. The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. Electrophoresis *26*:501-510.
13. **Gonzalez, M., L.A. Bagatolli, I. Echabe, J.L. Arrondo, C.E. Argarana, C.R. Cantor, and G.D. Fidelio.** 1997. Interaction of biotin with streptavidin. Thermostability and conformational changes upon binding. J. Biol. Chem. *272*:11288-11294.
14. **Brotherton, P., P. Endicott, J.J. Sanchez, M. Beaumont, R. Barnett, J. Austin, and A. Cooper.** 2007. Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Res. *35*:5717-5728.
15. **Hofreiter, M., V. Jaenicke, D. Serre, A. Haeseler Av, and S. Paabo.** 2001. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res. *29*:4793-4799.
16. **Lindahl, T.** 1993. Instability and decay of the primary structure of DNA. Nature *362*:709-715.

Received 3 August 2008; accepted 24 October 2008.

Address correspondence to Tomislav Maricic, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany. e-mail: tomislav_maricic@eva.mpg.de