BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2153–2155
doi:10.1093/bioinformatics/btr347

Sequence analysis

Advance Access publication June 9, 2011

mapDamage: testing for damage patterns in ancient DNA sequences

Aurelien Ginolhac*, Morten Rasmussen, M. Thomas P. Gilbert, Eske Willerslev and Ludovic Orlando

Centre for Geogenetics, Natural History Museum of Denmark, Copenhagen University, 1350 København K, Denmark

Associate Editor: Martin Bishop

ABSTRACT

Summary: Ancient DNA extracts consist of a mixture of contaminant DNA molecules, most often originating from environmental microbes, and endogenous fragments exhibiting substantial levels of DNA damage. The latter introduce specific nucleotide misincorporations and DNA fragmentation signatures in sequencing reads that could be advantageously used to argue for sequence validity. mapDamage is a Perl script that computes nucleotide misincorporation and fragmentation patterns using next-generation sequencing reads mapped against a reference genome. The Perl script outputs are further automatically processed in embedded R script in order to detect typical patterns of genuine ancient DNA sequences.

Availability and implementation: The Perl script mapDamage is freely available with documentation and example files at http://geogenetics.ku.dk/all_literature/mapdamage/. The script requires prior installation of the SAMtools suite and R environment and has been validated on both GNU/Linux and MacOSX operating systems.

Contact: aginolhac@snm.ku.dk

Supplementary information: Supplementary data available at *Bioinformatics* online

Received on March 11, 2011; revised on May 17, 2011; accepted on June 6, 2011

1 INTRODUCTION

The field of ancient DNA (aDNA) has a colorful history, with early claims of DNA survival from the Cretaceous (Woodward *et al.,* 1994) periods later recognized as the by-product of contamination (Zischler *et al.*, 1995). Extensive fragmentation and substantial chemical modifications of nucleotide bases are introduced *post-mortem* as a result of DNA damage (Pääbo, 1989). Some of these modifications, such as abasic sites, single-strand breaks and interstrand cross-links, are not replicated by standard *Taq* DNA polymerases (Hansen *et al.,* 2006; Mitchell *et al.,* 2005), hence the relatively better preserved modern contaminant DNA molecules will often outcompete ancient templates during polymerase chain reaction (PCR) amplification. For non-human species, contamination can be controlled by using specific PCR primers and/or human oligonucleotide PCR blockers (Gigli *et al.,* 2009) and by filtering out human-related sequences. But as long as ancient human fossils are concerned, including close relatives such as neandertals, contamination is still a risk.

Several recent methodological improvements have helped improve reliability when sequencing ancient hominids. Firstly, DNA libraries are tagged in clean labs with project-specific indexes in order to track back downstream contamination sources that could result from library amplification and/or sequencing (Briggs *et al.,* 2007). In addition, typical sequence patterns resulting from aDNA damage have been identified that help distinguish damaged ancient sequences from modern contaminants; these include: short sequence length; an excess of cytosine to thymine (C-to-T) misincorporations at 5′ ends of sequences, and complementary guanine to adenine (G-to-A) misincorporations at 3′-termini, due to enhanced cytosine deamination in single stranded 5′-overhanging ends (Green *et al.,* 2009; Krause *et al.,* 2010); and an excess of purines at the genomic coordinate located just before the sequencing start, indicative of *post-mortem* depurination, followed by strand fragmentation (Briggs *et al.,* 2007).

Here, we present a Perl script that analyzes the size distribution of next-generation sequencing reads, and dynamically recovers nucleotide misincorporation patterns in a position-dependent manner. In addition, the base composition of the genomic regions located up- and downstream of each read is recorded within windows of user-defined sizes. Using SAM files as input and corresponding reads mapped against a reference genome, the script mapDamage provides summary tables required to gauge aDNA authenticity based on DNA damage patterns.

2 RESULTS

Script outline: the following programs must be installed prior to running mapDamage: SAMtools (Li *et al.,* 2009a) and the R environment (R Development Core Team, 2010) that generates user-friendly tables and graphics using downstream scripts. Three main commands can be used to run mapDamage: *map*, *merge* and *plot*. The main command, *map*, takes as input a SAM file (if present, headers with @SQ lines will be skipped) and the reference genome used to map next-generation sequencing reads as a fasta file. Corresponding genomic regions from the reference are retrieved using SAMtools and aligned to sequencing reads using the CIGAR information (Li *et al.,* 2009a). The *merge* command compiles, in single tables, all individual tables generated with the *map* command for individual chromosomes. These tables can be further processed with the *plot* command that produces graphics without any requirements of R knowledge. The global pipeline is described in the Supplementary Figure S1.

*To whom correspondence should be addressed.

mapDamage options:

(1) *map* command, a second SAM file may be supplied with -j. Reads present in the first SAM file will be considered as long as missing from the second SAM file. This option is particularly useful to remove potential human contaminants when shotgun sequences from a non-human organisms are generated.

(2) *map* command, by default, the maximum length of reads is set to 70 nt. This could be adjusted using -l.

(3) *map* command, the size of the genomic window located upstream and downstream of the reads is set by default to 10 nt but can be changed with the -a option.

(4) *map* command, read alignments against the reference genome considered are output in a fasta format when the -f option is used.

(5) *map* command, multithreading is controlled by the -t option.

(6) *map* command, the option -u will remove the non-unique hits based on the optional field 'XA' for alternative hits from BWA.

(7) *map* command, via -d, users can provide the name of the subfolder where all output files will be stored.

(8) *merge* command, -d option indicates the main folder where individual files generated with the *map* command are stored.

(9) *plot* command, users may change the range of the *y*-axis in nucleotide misincorporation graphs with -m option. The number of nucleotides to plot for the read and reference regions are controlled through options -l and -a, respectively.

Runtime: the *map* command is the most-time consuming step, as it requires multiple disk access. Typically, ca. 650 000 reads are processed within 6 h on a laptop with 8 GB RAM and a quad core 1,6 GHz processors. For optimal performance, several threads should be used. Both *merge* and *plot* commands are performed within a few minutes.

Outputs: Typical file outputs with corresponding examples are described in details at http://geogenetics.ku.dk/all_literature/mapdamage/. Briefly, the *map* command outputs seven file categories for each chromosome plus the length distribution per strand. The *merge* command outputs seven tabular files compiling the previous information for the whole set of chromosomes. Finally, *plot* outputs one pdf file and the respective R script corresponding to fragmentation and misincorporation patterns.

Application on the paleo-Eskimo Saqqaq genome: the first ancient human genome was sequenced on Illumina GAIIx platforms to 20X average depth (Rasmussen *et al.,* 2010). DNA libraries were PCR amplified with Phusion polymerase in order to limit nucleotide misincorporations resulting from cytosine deamination (Fogg *et al.,* 2002). An additional lane of sequencing data was generated from a DNA library amplified with the Platinum *Taq* DNA polymerase High Fidelity (Invitrogen). From the latter, after tag trimming and duplicate template removal, a total number of 15 918 505 sequences were generated but were not included in the final genome assembly. One million sequences were randomly selected and mapped against the human genome (hg19, obtained from UCSC Genome Browser) using BWA (Li and Durbin, 2009b) with default parameters. Resulting SAM files were filtered for a minimum mapping quality of 25 with SAMtools (Li *et al.,* 2009a) resulting in 626 425 uniquely mapped reads. Regarding the age of this specimen, i.e. 4000 years, cytosine to thymine misincorporations are drastically increased at 5′-termini when using a Hifi *Taq* DNA polymerase (Supplementary Fig. S2, top right), suggesting a substantial level of cytosine deamination was present in the ancient human DNA fragments. In contrast, random sampling of 1 million sequences from 15 179 000 sequences generated on one GAIIx lane using a Phusion amplified DNA library resulted in 610 149 unique maps against hg19. The damage pattern showed a 3.8-fold reduction considering the 25 first nucleotides of cytosine to thymine misincorporation (Supplementary Fig. S2, top left), suggesting that the selection of a single enzyme, here Phusion polymerase, drastically improved the quality of the genome by neutralizing DNA damage-related nucleotide misincorporations. The DNA fragmentation pattern remains unaffected with an excess of purines at the genomic position preceding the read start (Supplementary Fig. S2, bottom).

3 DISCUSSION

mapDamage reports DNA damage patterns in sequences generated from aDNA using next-generation sequencing platforms. We have validated the script both on simulated sequence datasets with significant levels of DNA damage (Prüfer *et al.*, 2010; Supplementary Fig. S3) and on real aDNA datasets. With mapDamage, nucleotide misincorporations driven by cytosine deamination can be easily observed and quantified. Similarly, the nucleotide position at which aDNA molecules have been fragmented can be monitored. The complete set of analyses can be performed globally, or per chromosome or DNA strand in order to detect possible skews. In addition, coverage information is provided in hitPerChrom files that report the total number of nucleotides sequenced on each strand and per chromosome. In summary, mapDamage provides useful summary tables, which together with R plotting script (or other data analysis software), extracts information for authenticating aDNA based on DNA damage signatures.

ACKNOWLEDGEMENTS

We thank Tobias Mourier for testing mapDamage.pl on the MacOSX operating system, Uffe Gram Wilken for technical assistance and three reviewers for fruitful suggestions.

Funding: Danish Council for Independent Research; Natural Sciences (FNU); Danish National Research Foundation (Danmarks Grundforskningsfond).

Conflict of Interest: none declared.

REFERENCES

Briggs,A.W. *et al*. (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA*, **104**, 14616–14621.

Fogg,M.J. *et al.* (2002) Structural basis for uracil recognition by archaeal family B DNA polymerases. *Nat. Struct. Biol.*, **9**, 922–927.

Gigli,E. *et al*. (2009) An improved PCR method for endogenous DNA retrieval in contaminated Neandertal samples based on the use of blocking primers. *J. Arch. Sci*, **36**, 2676–2679.

Green,R.E. *et al*. (2009) The Neandertal genome and ancient DNA authenticity. *EMBO J.*, **28**, 2494–2502.

Hansen,A.J. *et al.* (2006) Crosslinks rather than strand breaks determine access to ancient DNA sequences from frozen sediments. *Genetics*, **173**, 1175–1179.

Krause,J. *et al*. (2010) A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr. Biol.*, **20**, 231–236.

Li,H. *et al.* (2009a) 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,H. and Durbin R. (2009b) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Mitchell,D.L. *et al*. (2005) Damage and repair of ancient DNA. *Mutat. Res.*, **571**, 265–276.

Pääbo,S. (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification *Proc. Natl Acad. Sci. USA*, **86**, 1939–1943.

Prüfer,K *et al.* (2010) Computational challenges in the analysis of ancient DNA. *Genome Biol.*, **11**, R47.

R Development Core Team (2010) *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen,M. *et al.* (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, **463**, 757–762.

Woodward,S.R. *et al*. (1994) DNA sequence from Cretaceous period bone fragments. *Science*, **266**, 1229–1232.

Zischler,H. *et al*. (1995) Detecting dinosaur DNA. *Science*, **268**, 1192–1193.