SUPPLEMENTAL INFORMATION

Recharacterization of ancient DNA miscoding lesions: Insights in the era of the GS20

M Thomas P Gilbert, Jonas Binladen, Webb Miller, Carsten Wiuf, Eske Willerslev, Hendrik Poinar, John E. Carlson, James H. Leebens-Mack, Stephan C Schuster.

Theory underlying the identification of original strands of origin of damage and subsequent characterization of miscoding lesions.

Previous studies on miscoding lesions in ancient DNA have highlighted the difficulties that exist in identifying the underlying cause of the mutations, due to the complementary nature of PCR. For example, consider the example of a cytosine to thymine (C→T) transition observed on a Light (L) strand mitochondrial DNA sequence. Although initially the obvious cause would seem to be an original C→T transition on the ancestral DNA L strand molecule that was originally PCR amplified and subsequently sequenced, an alternative equally likely explanation could be a guanine to adenine (G→A) transition at the corresponding complementary nucleotide position on an original Heavy (H) strand molecule in the same PCR (S1). This observation has therefore lead to the common usage of grouping complementary miscoding lesions into pairs of data (c.f. S1-S3). As such, studies on DNA damage have attempted to argue the biochemical causes based on arguments that either draw on experimental evidence such as the treatment of aDNA extracts with enzymes chosen to cleave DNA at specific hypothetical damage-derived nucleotides (for example uracil-N-glycosylase cleavage of uracil, the hypothesized byproduct of cytosine deamination and as such the argued cause of C→T/G→A miscoding lesions; (S2-S4)), or attempt to draw possible solutions from information known about in vivo DNA damage biochemistry (e.g. the hypothesis that adenine deamination to hypoxanthine is the primary cause behind the complementary A→G/T→C miscoding lesions (S3))

These methods suffer obvious weaknesses, for example enzymatic assays will only be able to identify the presence of DNA damage derivatives that are specifically targeted, while attempts to draw explanations for post mortem damage from in vivo systems must rely on the unqualified assumption that similarities exist between the two systems. As such, the ability to directly identify the strand of origin of damage-derived miscoding lesions offers a much more powerful and accurate tool that can be used to investigate the underlying causes of observed miscoding lesions. Unlike conventional PCR, where even if a start from a single molecule can be guaranteed, it is not easily possible to identify which strand this is, DNA sequences produced using the GS20 offer such a possibility.

During initial stages of sample preparation, original double-stranded DNA molecules are first fragmented (Fig. S1a) and then bound to special double-stranded ligator sequences (known as ‘A’ and ‘B’). These new hybrids are in turn denatured to generate single-stranded molecules (Fig. S1b). Full details of this process can be obtained in the supplementary information to the original Margulies et al. (2005) paper that describes the GS20/454 platform (S5). The key point with regards to this study, is that in the subsequent stage of the DNA preparation process, individual single-stranded DNA sequences are isolated on to emPCR beads through binding with a probe that is permanently coupled to the bead (Fig. S1c). At this stage, emPCR commences on the individual emPCR-template hybrids. During the first cycle of emPCR, the complement to the captured molecule is generated (Fig. S1d). Subsequent cycles generate descendent double-stranded DNA molecules as with conventional PCR (Fig. S1e), however of these descendents, only molecules that are in the complementary orientation to the original single-stranded molecule bind to the emPCR bead. Post emPCR, additional purification and preparation steps are performed prior to pyrosequencing. These involve, among other things, the dissociation of the PCR products into single-stranded DNA, of which only the molecules that are bound to the emPCR bead are retained (Fig. 1f). Therefore, for each individual emPCR bead, the template molecules subjected to pyrosequencing occur exclusively in one orientation (for example with mtDNA, either H or L strands, but not both). The orientation does however differ between different emPCR beads within each GS20 run, thus when the overall data produced from the potentially hundreds of thousands of individual emPCR reactions is analysed, sequences are found representing both the H and L strand.

The pyrosequencing process itself is initiated with sequencing primers that are complementary to part of the 'B' ligator sequence (Fig. S2a). As with conventional Sanger sequencing, the reaction proceeds from 3'-5' along the captured template molecules, generating new sequence that is complementary to the bound molecules (Fig. S2b). Although in theory multiple template molecules could bind to the emPCR bead during initial sample preparation, various steps within the sample preparation and subsequent DNA analysis process ensure that this final sequence data only derives from single, single-stranded DNA molecules. Full details can be found within the supplemental information to reference S5.

Implications of sequencing-by-synthesis.

Using the above knowledge it is possible to highlight several key points upon which our analyses rest, and discuss their implications.

1) Each individual sequence derives from a single, single-stranded DNA template molecule. Therefore observed sequence variation can only be derived from damage, enzyme error or true sequence variation. Other explanations, such as recombination between template molecules during PCR are not possible.
2) Each individual sequence is the complement of template molecules captured on the emPCR bead. These captured molecules in turn are the descendents (through emPCR) of a single captured molecule, which was the complement of an original single-stranded template molecule. Therefore, as the 'complement of a complement', the sequence directly describes an original template molecule. As such, it is possible to explicitly identify the orientation of each original template molecule – it is simply the same as the sequence.

This in turn has several implications.

3) Firstly in our analyses we can divide the sequences up into two datasets, those derived originally from the H strand, and those from the L strand (*Analysis 3*). This is achieved by assessing whether each sequence directly aligns with an L

strand reference sequences, or whether it requires reverse complementation (Fig. S2c-f)

4) Secondly, it is possible to identify directly the cause of individual damage events (Fig. S3), as any miscoding lesion observed in generated sequences directly represents an original miscoding lesion event on the original template molecule. Thus, a G→A transition observed in a sequence derives from a G→A miscoding lesion on an original template molecule. Similarly, a C→T transition observed in a sequence derives from a C→T miscoding lesion on an original template molecule. This observation forms the basis of *Analysis 4.* Naturally this argument relies on the assumption that damage can be discriminated apart from other causes of sequence variation, such as PCR enzyme errors or heterogeny in the template molecules. We achieve this in two ways. Firstly, we investigate mtDNA, thus the effect of heteroplasmy can be assumed to be negligible (as far as this analysis is concerned). Secondly, we compare the miscoding lesions spectra of the ancient mammoth DNA with a modern chloroplast dataset, in which postmortem damage levels are also likely to be negligible, and statistically compare the data to demonstrate which miscoding lesions are over represented in the damage dataset, thus represent post mortem damage as opposed to enzyme error (*Analyses 1 and 2*).

SUPPLEMENTAL REFERENCES

S1. Hansen, A., E. Willerslev, C. Wiuf, T. Mourier, and P. Arctander. (2001) Statistical evidence for miscoding lesions in ancient DNA templates. Mol. Biol. Evol.,18, 262-265.

S2. Hofreiter, M., V. Jaenicke, D. Serre, A. von Haeseler, and S. Pääbo S. (2001) DNA sequences from multiple amplifications reveal artefacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res., 29, 4693-4799.

S3. Gilbert, M.T.P., Hansen, A.J., Willerslev, E., Rudbeck, L., Barnes, I., Lynnerup, N. and Cooper, A. (2003a) Characterisation of genetic miscoding lesions caused by *post mortem* damage. Am. J. Hum. Genet., 72, 48-61.

S4. Pääbo, S. (1989) Ancient DNA: Extraction, characterization, molecular cloning and enzymatic amplification. Proc. Natl. Acad. Sci. USA, 86, 1939-1943.

S5. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.T. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437, 376-380.