Supplementary data

Estimating the horse average genomic GC content for different read length

As different read size distributions were generated on the Helicos Sequencer and the Illumina GAIIx platforms, we estimated the expected GC content of the horse genome based on a mapping procedure. Random sequences were selected on each chromosome of eqCab2 (1 to 31, and X), with sequence length set to the respective median values of both read size distributions (31 nucleotides for Helicos, and 67 nucleotides for Illumina sequencing; see suppl Fig. 1). This resulted in 361,379 and 299,256 sequences that were further mapped against eqCab2 using BWA and following the procedure described in Material and Methods; of these, 322,960 and 290,558 presented mapping qualities superior to 25 and were found at unique positions in the reference genome, which represents 89.4% and 97.1% of the total number of sequences processed. The overall GC contents of both sequence sets were calculated to 41.41% and 41.38%, which provided respective estimates of the expected genomic GC content for Helicos and Illumina reads (Fig. 2 and 3).

Nucleotide composition of Helicos sequencing reads

We characterized possible bias resulting from denaturation, poly-A tailing, blocking, and oligo-T capture steps performed in the standard Helicos template preparation protocol using available Helicos sequence data generated from modern human genomic sequences (Pushkarev et al. 2009). We identified three typical features in the nucleotide composition of the genomic regions sequenced.

First, no thymine was found in the first position of sequence reads; this resulted from post-sequencing processing of the reads that were trimmed when starting with one or more thymine residues, as the latter could arise from sequencing the poly-A region after incomplete fill-in and lock reactions (suppl Fig. 5). This deficit in thymine caused the composition in other nucleotides to be increased, with preference for adenine, cytosine and, to a lesser extent, guanine residues.

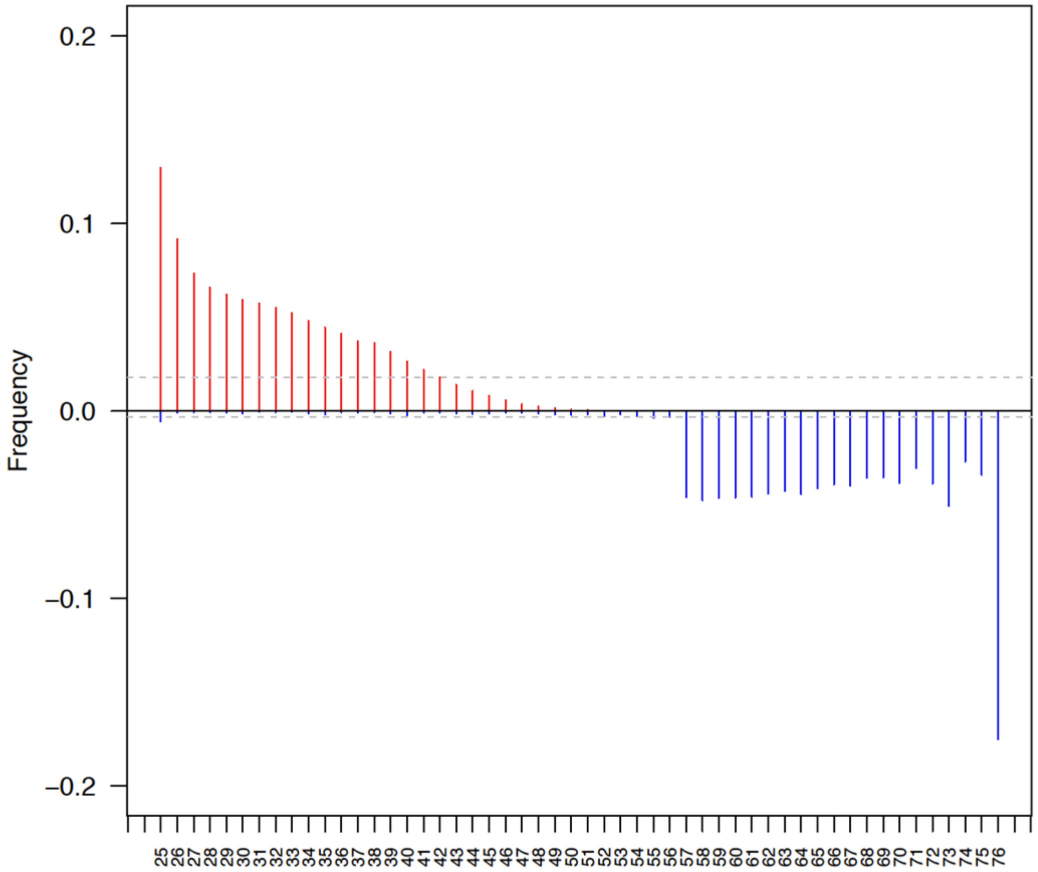
Second, the genomic coordinate that is located just before the sequencing start was preferentially enriched in guanine residues while impoverished in adenine and cytosine residues; the composition in thymine was not affected (suppl Fig. 5). This resulted from the fill-in/locking reactions that incorporated dTTP in the region complementary to the poly-A tail but dCTP, dGTP and dATP Virtual Terminator nucleotides at the first base beyond the tail (Fig. 1). The excess in guanine residues at the blocking site is caused by that base being most efficient in the fill-in/lock step. One consequence is that the first nucleotide sequenced is not necessarily the next to last nucleotide of the aDNA strand (the last corresponding to the site locked; suppl Fig. 6) but might be located a few nucleotides further into the preserved molecule.

The third deviation to the average nucleotide composition was observed in the genomic region preceding the blocking site, that showed a progressive excess in Thymine residues paralleled by a deficit in Adenine and to a lesser extent cytosine and guanine residues (suppl Fig. 5). This resulted from the poly-thymidine oligonucleotide capture step that is designed to probe poly-A tailed strands before fill-in, locking and sequencing; Adenine-rich regions outcompete other genomic regions in this capture step, hence, an enrichment for the complementary nucleotidic base, Thymine, was detected upstream of sequence reads (suppl Fig. 5). Interestingly, all three biases, except the excess in guanine residues at the locking-sites, are observed on the human reads generated from the ancient horse extracts (suppl Fig. 7**)**. Of note, the first nucleotide sequenced showed extremely low, but not null frequencies in thymine residues, as sequences starting with a minimum of two Thymine residues were trimmed post-sequencing.

Mapping sensitivity to cytosine deamination

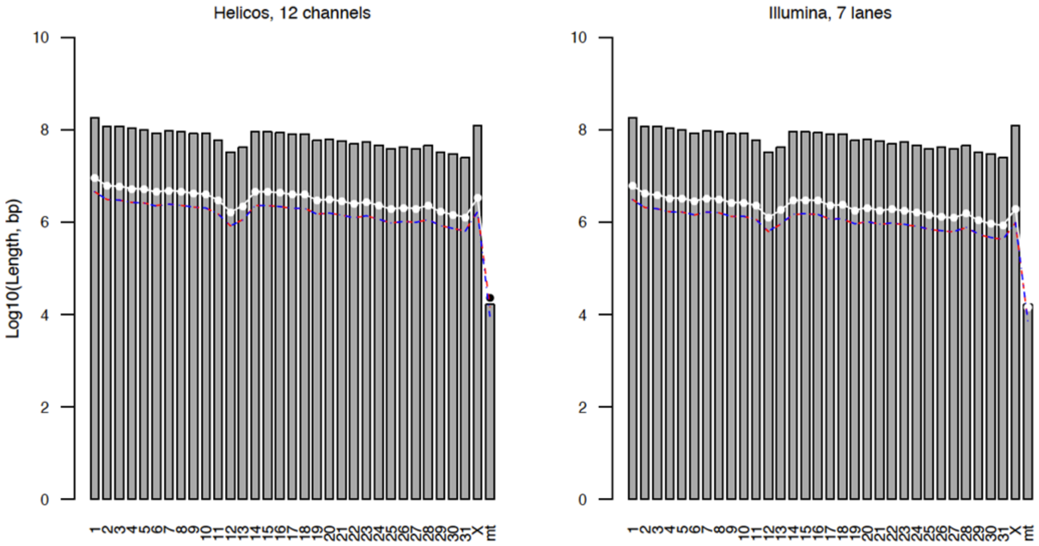
In order to evaluate how much cytosine deamination could affect the quality of sequence mapping against the horse reference genome eqCab2, we selected random sequences on each chromosome (1 to 31, and X). Sequence length was set to the median of Helicos reads (31 nucleotides); in addition, the first nucleotide was constrained to a guanine residue. This first dataset, consisting of 340,833 sequences, was mapped against eqCab2 using BWA, default parameters and mapping quality scores superior to 25. A total of 302,253 sequences (88.7%) were successfully mapped against eqCab2. The first nucleotide of all reads was then reverted to adenine in order to mimic guanine to adenine misincorporation, resulting from cytosine deamination. Using previous parameters, BWA resulted in a total of 294,257 sequences (86.33%) mapping successfully against eqCab2, suggesting that the mapping of 7,996 sequences was sensitive to the nature of the first nucleotide position (ie. 2.6% of the reads that previously mapped successfully were lost). The same procedure was reiterated for a new sequence dataset consisting of similar sequence length (31 nucleotides) but where the second nucleotide (not the first) was constrained to a guanine residue (388,007 sequences) and then reverted to an adenine residue. From the 344,016 (88.7%) sequences that successfully mapped, guanine to adenine reversion resulted in a loss of 8,384 sequences (2.4%). Overall, this suggests that the levels of cytosine deamination observed in the first two nucleotidic positions could affect mapping success, with 5.0% of the sequences concerned (ie. guanine to adenine deamination in the first or second position of Helicos reads) remaining unmapped. Albeit marginal, this frequency will certainly result in non-negligible numbers of unidentified genuine ancient sequences considering the tremendous amount of reads generated in paleogenome sequencing projects. When performed on 377,781 and 375,175 random sequences extracted from eqCab2 to mimick Illumina reads (median length = 67 nucleotides; Cytosine to Thymine reversion at the first and second nucleotide position, respectively), a significant decrease in the impact of cytosine deamination on mapping success was observed, with 0.4% in both cases (1,639 sequences out of 366,000 and 1,304 sequences out of 363,249 left with no significant mapping, respectively).

Supplementary Figures



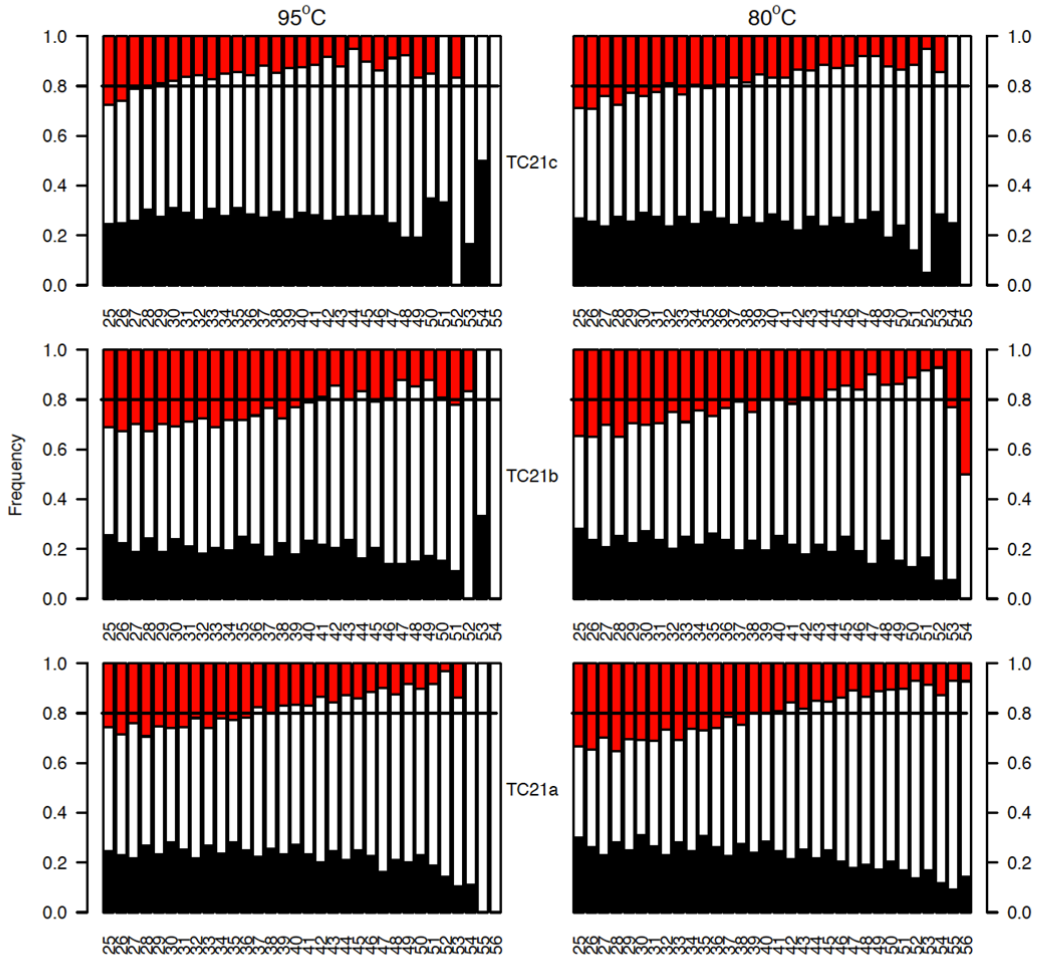
Supplementary Figure 1. Helicos *versus* Illumina: read length distributions.

The relative frequency of each read size is provided for the two sequencing technologies used in this study (Top, red: Helicos; Bottom, blue: Illumina). Dashed grey lines refer to the 5% and 95% quantiles in Illumina and Helicos distributions, respectively. Overall, 95% of Helicos reads are 42-nucleotides long or shorter while 95% of Illumina reads show read length superior or equal to 57 nucleotides.



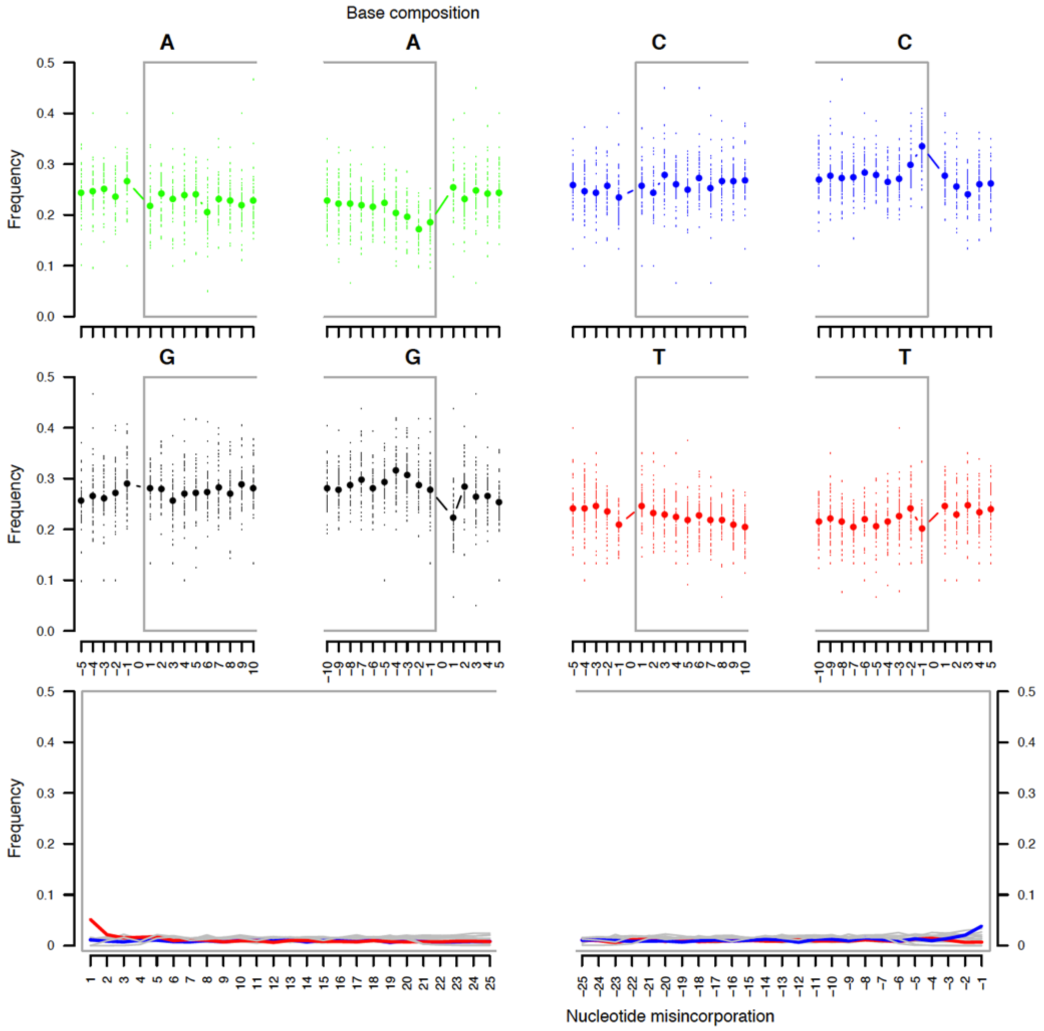
Supplementary Figure 2. Genomic coverage relative to chromosomal sizes.

The length of each horse chromosome as represented in eqCab2 (black bars) and the total number of nucleotide sequenced (white line and points) is provided using Helicos (left) and Illumina reads (right). Red, dotted line: strand + (sequencing reads mapped in the same orientation as the reference genome). Blue, dotted line: strand – (sequencing reads mapped in the reverse complement orientation of the reference genome). The distributions of the number of sequencing reads mapping on strand + and strand – show no significant difference.



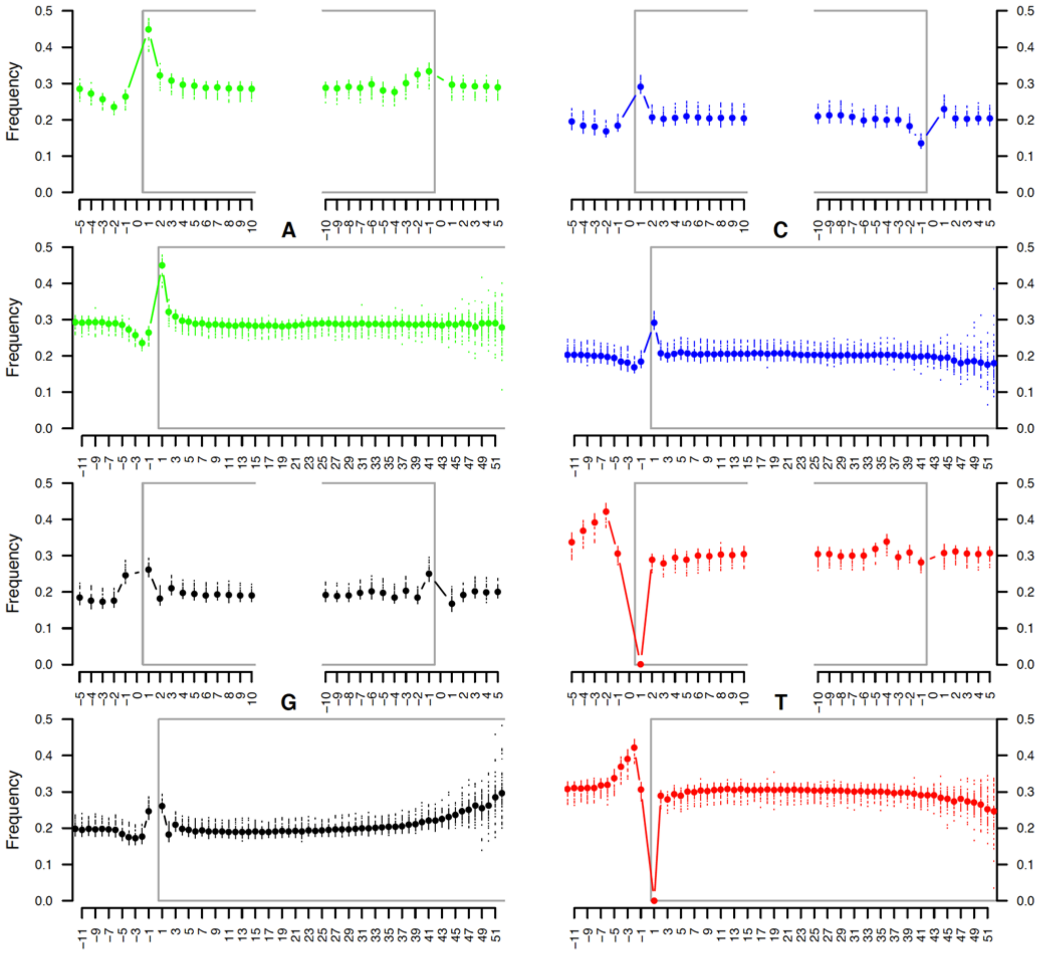
Supplementary Figure 3. GA contents of Helicos tSMS reads.

Three different extracts (top: TC21c; middle: TC21b; bottom: TC21a) have been sequenced on the same Helicos run (6 channels) following identical procedures, except that either high (95°C, left) or mild (80°C, right) temperatures were used for denaturation. Reads are classified according to their GA composition: black, GA% < 40.0; white, 40.0 < GA% < 60.0; red, GA% >60.0%; and the relative proportions of these three classes are reported as a function of read length. Lines are traced at 80% in order to facilitate comparisons among the different distributions.



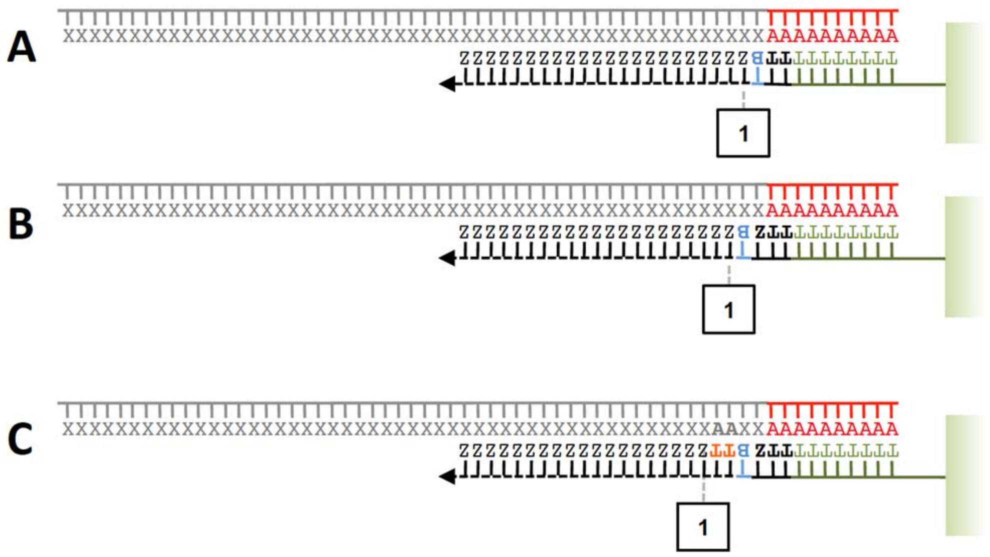
Supplementary Figure 4. Illumina sequencing: base composition and fragmentation pattern of human contaminant reads.

Top, Middle: The base composition of the reads is reported for the first 10 nucleotides sequenced (left: 1 to 10) as well as for the 5 nucleotides located upstream of the genomic region aligned to the reads (left: -5 to -1). In addition, the base composition of the last 10 nucleotides sequenced (right: -10 to -1) and of the 5 nucleotides located downstream of the reads in hg19 (right: 1 to 5) is provided. Nucleotide positions located within reads are reported with a grey frame. Each dot indicates base composition as reported from chromosome 1 to 22, X and Y. Full circles refer to average values. Bottom: The frequencies of all possible mismatches and indels observed between the horse genome and the reads are reported in grey as a function of distance for 5'- to 3'-ends (left, first 25 nucleotides sequenced) and 3'- to 5' (right, last 25 nucleotides), except for C→T, G→A that are reported in blue and red respectively. These frequencies are calculated by dividing the total number of occurrences of the modified base at a given position in a read by the total number of the unmodified base at the same position in the human reference genome hg19. For sequences shorter than 27 nucleotides, we detected a high proportion of misalignments against the human reference genome (hg19) at the 3'-ends of sequencing reads resulting from excessive gap extensions. Nucleotide misincorporation patterns were consequently plotted on mapping sam files generated after increasing the gap extension penalty from -4 to to -20 (BWA aln option –E) which improved the quality of such problematic alignments.



Supplementary Figure 5. Helicos tSMS of modern DNA: base composition.

The base composition of the reads is reported for the first 10 (or 51) nucleotides sequenced as well as for the 5 nucleotides located upstream of the genomic region aligned to the reads. In addition, the base composition of the last 10 nucleotides sequenced and of the 5 nucleotides located downstream of the reads in hg19 is provided, using the same numbering system as in suppl Fig 4. The base composition of the reads is calculated as a function of the position of the nucleotide along the read as well as for the nucleotides located upstream and downstream of the genomic region aligned to the reads. Each dot indicates base composition as reported from chromosome 1 to 22, X and Y. Full circles refer to average values. Nucleotide positions located within reads are reported with a grey frame.

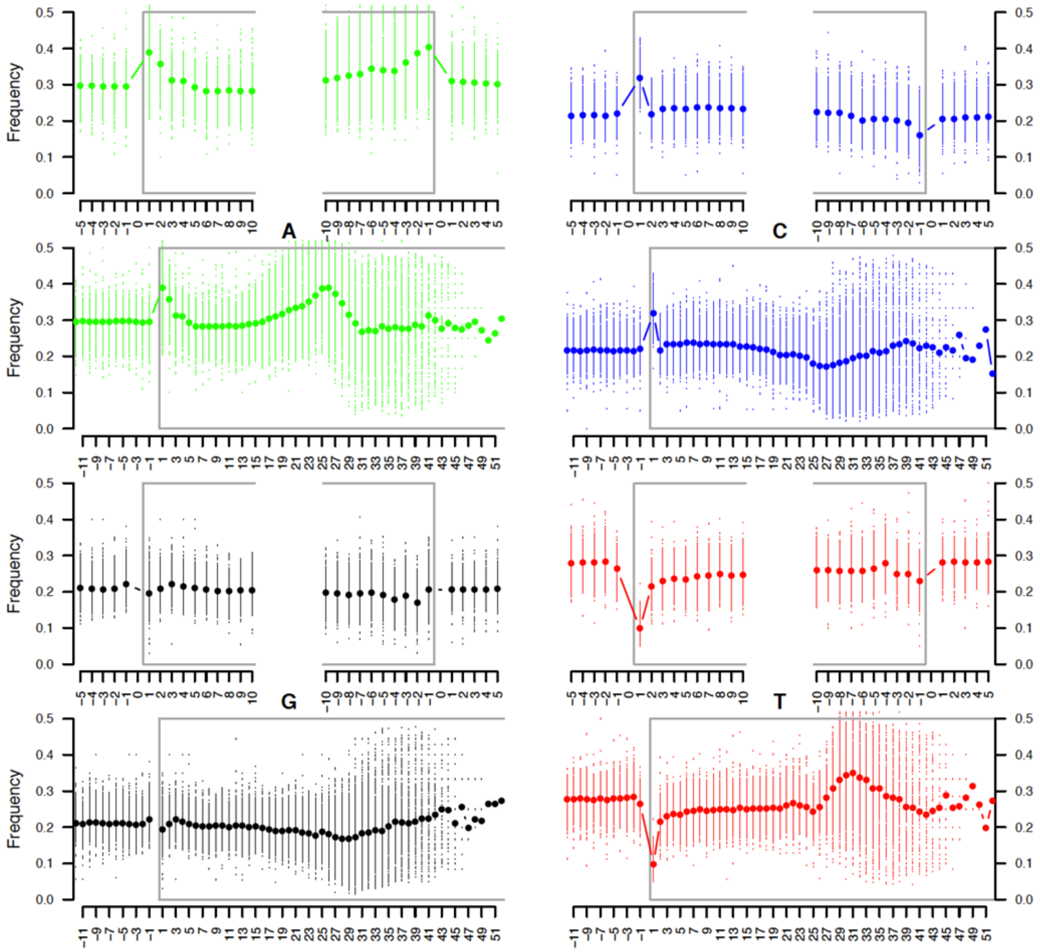


Supplementary Figure 6. Nucleotide position along a Helicos read: caveats.

Panel A: The fill-in and lock steps are designed to fill any remaining nucleotide complementary to the poly-A tail (as the first nucleotide in poly-A tails does not necessarily hybridize with the last oligo-dT) and hamper further extension prior to terminator cleavage and sequencing-by-synthesis.

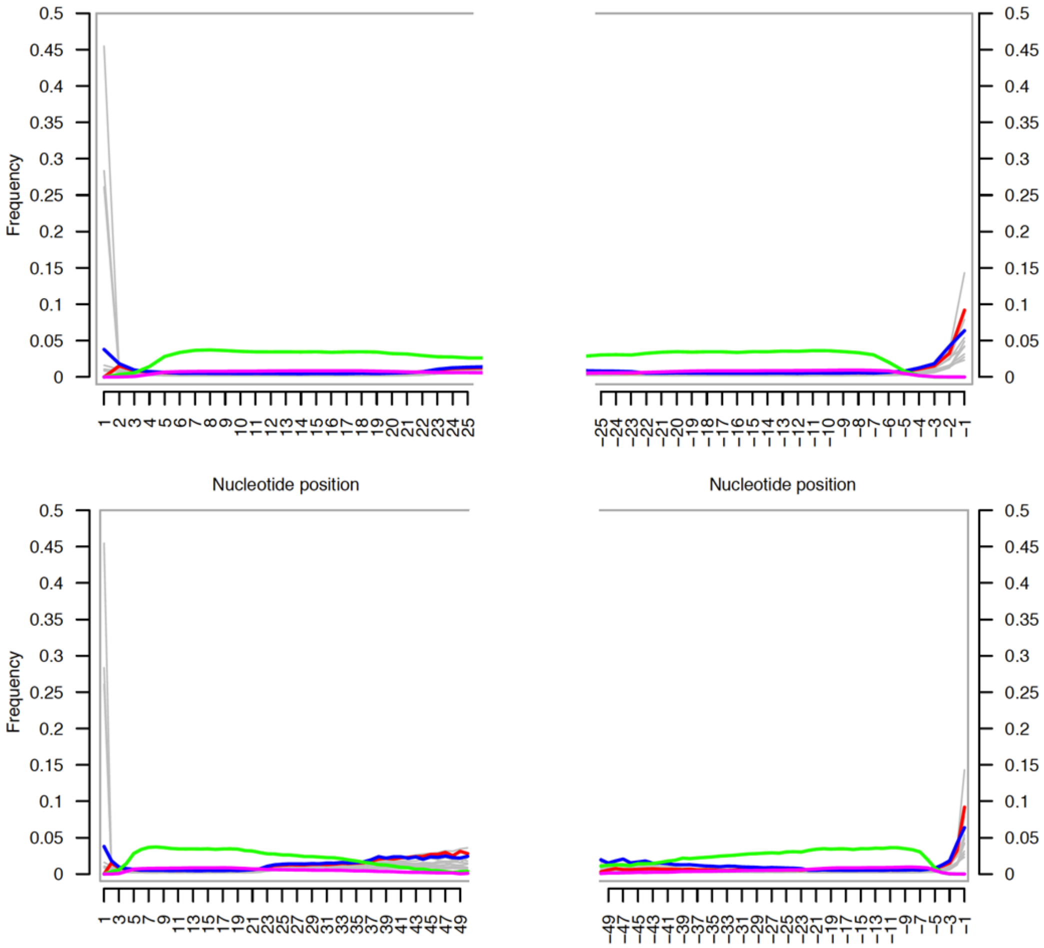
Panel B: The lock process is not 100%-efficient and locked nucleotides can be located further in the strand to be sequenced. In these cases, the first nucleotide sequenced does not correspond to the next to the last nucleotide on the original aDNA strand.

Panel C: In addition, raw Helicos sequences starting with a minimum number of two Ts are trimmed before being analyzed, as they could have arisen from incomplete fill and lock reactions. Yet, the 3’-end of a fraction of endogenous fragments might be A-rich. *In silico* trimming of the corresponding T-stretch in the sequence generated (here, in orange) will shift the nucleotide positions within the reads (here, the first nucleotide conserved in the sequence file would correspond to the one located fifth from the end of the aDNA strand).



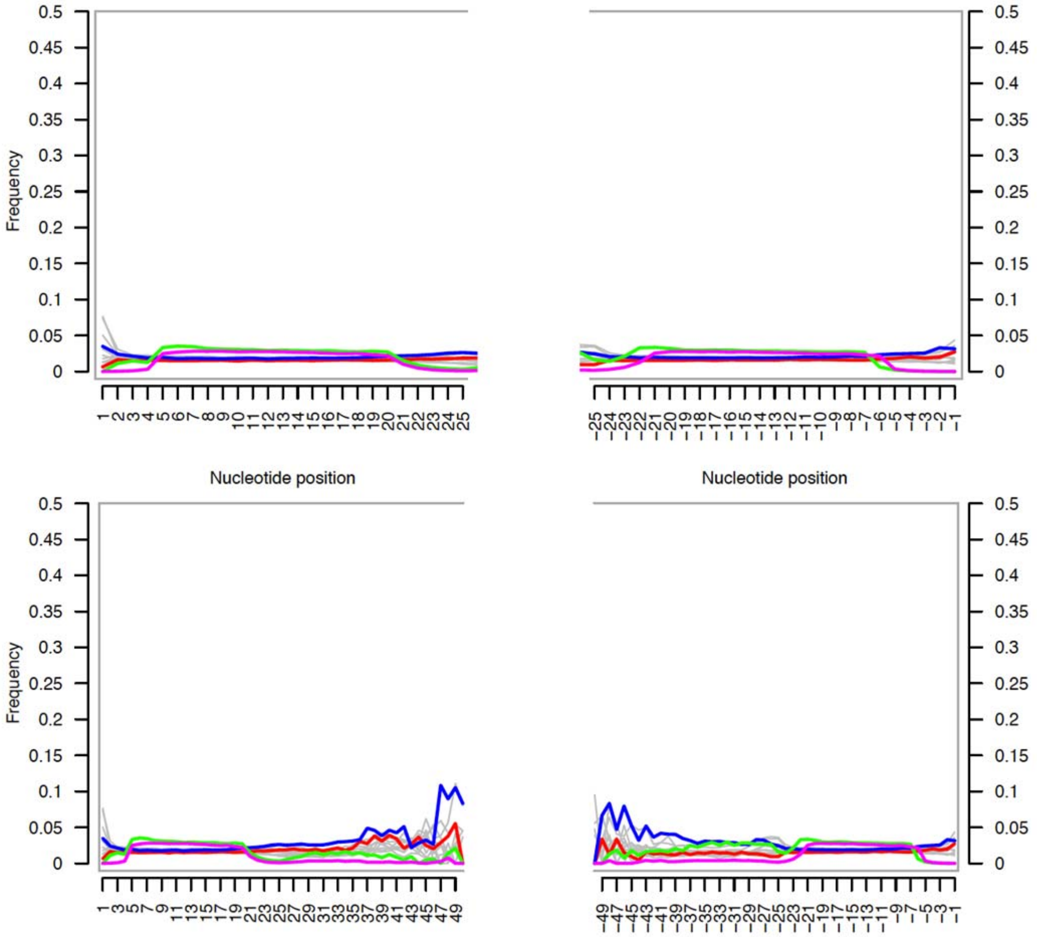
Supplementary Figure 7. Helicos tSMS of human contaminant reads in the horse ancient DNA extract: base composition.

The base composition of the reads is reported for the first 10 (or 51) nucleotides sequenced as well as for the 5 nucleotides located upstream of the genomic region aligned to the reads. In addition, the base composition of the last 10 nucleotides sequenced and of the 5 nucleotides located downstream of the reads in hg19 is provided, using the same numbering system as in suppl Fig 4. The base composition of the reads is calculated as a function of the position of the nucleotide along the read as well as for the nucleotides located upstream and downstream of the genomic region aligned to the reads. Each dot indicates base composition as reported from chromosome 1 to 22, X and Y. Full circles refer to average values. Nucleotide positions located within reads are reported with a grey frame.



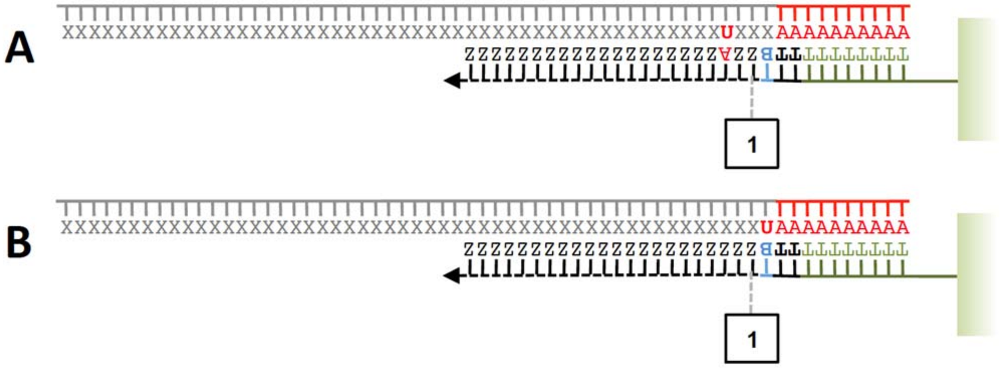
Supplementary Figure 8. Helicos tSMS of modern DNA: nucleotide misincorporation.

The frequencies of all possible mismatches and indels observed between the human genome and the reads are reported in grey as a function of distance for 5’- to 3’-ends (Left, first 25 and 50 nucleotides sequenced on top and bottom, respectively) and 3’- to 5’- (Right: last 25 and 50 nucleotides on top and bottom, respectively), except for C→T, G→A, insertions, and deletions that are reported in blue, red, pink, and green respectively. These frequencies are calculated by dividing the total number of occurrences of the modified base at a given position in a read by the total number of the unmodified base at the same position in the human reference genome (hg19).



Supplementary Figure 9. Helicos tSMS of human contaminant reads in the horse ancient DNA extract: nucleotide misincorporation.

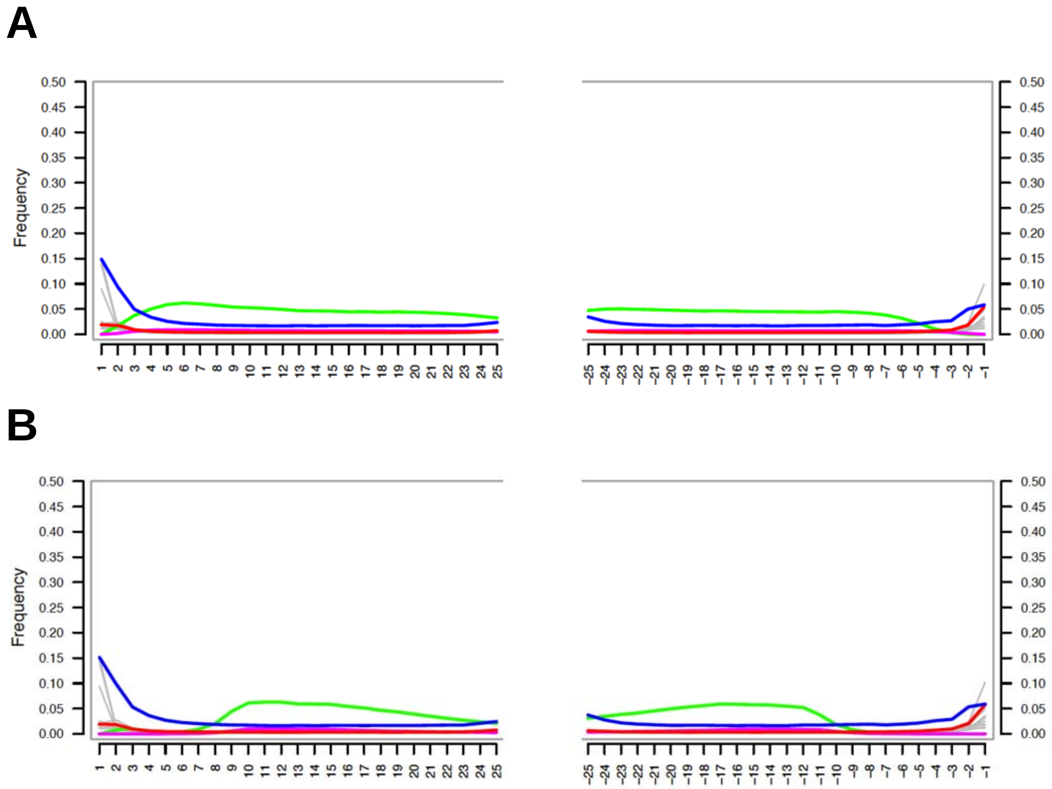
The frequencies of all possible mismatches and indels observed between the human genome and the reads are reported in grey as a function of distance for 5’- to 3’-ends (Left: first 25/50 nucleotides sequenced on top and bottom, respectively) and 3’- to 5’ (Right: last 25/50 nucleotides on top and bottom, respectively), except for C→T, G→A, insertions, and deletions that are reported in blue, red, pink, and green respectively. Nucleotide positions located within reads are reported with a grey box. These frequencies are calculated by dividing the total number of occurrence of the modified base at a given position in a read by the total number of the unmodified base at the same position in the human reference genome (hg19).



Supplementary Figure 10. Helicos nucleotide misincorporation pattern. Any deaminated-cytosine (uracil, U) located at the 3’-end of the aDNA strand will be read as a thymine, generating a G→A misincorporation during sequencing.

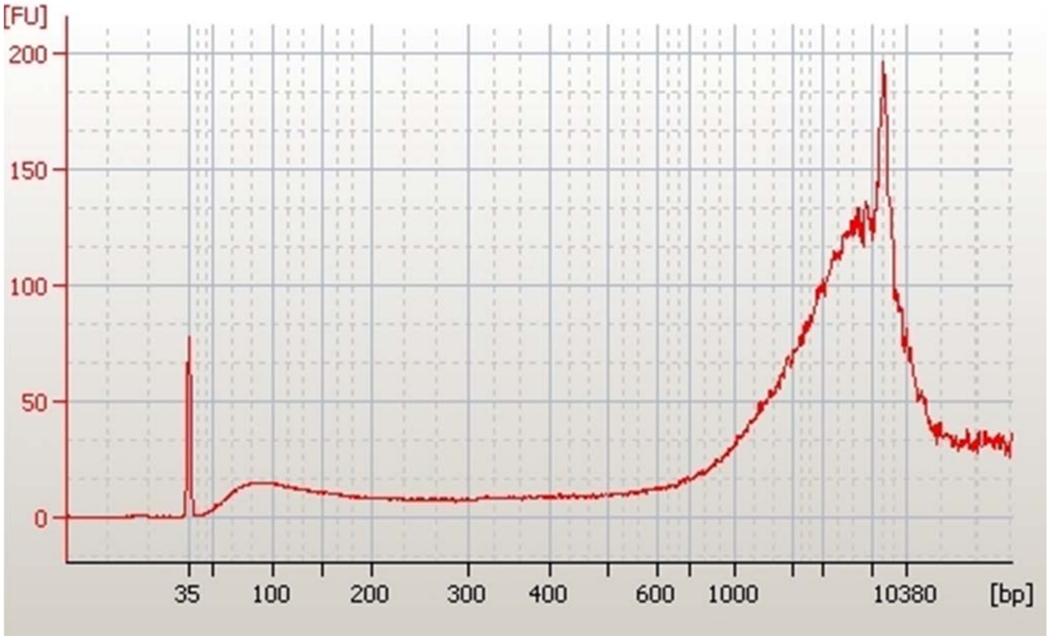
Panel A: Due to the fill-in and locking reactions, the position of the misincorporation within the read (here 3 nucleotides from the sequencing start) will not be phased with its original location in the aDNA strand (4 nucleotides from the end).

Panel B: Consequently, in cases when the last nucleotide located 3’- of the aDNA strand has been deaminated, no G→A misincorporation will be recorded in the Helicos read, suggesting that Helicos reads provide a minimal estimate for cytosine deamination in aDNA templates.



Supplementary figure 11. Helicos guanine to adenine misincorporation pattern: insensitivity to the presence and location of indels.

The frequencies of all possible mismatches and indels observed between the horse genome and the reads are reported in grey as a function of distance for 5’- to 3’-ends (first 25 nucleotides sequenced) and 3’- to 5’- (last 25 nucleotides), except for C→T and G→A that are reported in red and blue respectively. Different mapping procedures have been used in order to test if nucleotide misincorporation patterns were robust to the presence/location of indels. Helicos reads generated from extract TC21a denatured at 80°C and showing a minimal size of 25 nucleotides were mapped using different options available in BWA and mapped reads were filtered for a minimal mapping quality of 25. Panel A: Indels are allowed in the first/last 5 nucleotide positions of mapped sequencing reads (BWA aln option –i 0). Panel B: Indels are disallowed in the first/last 10 nucleotide positions of mapped sequencing reads (BWA aln option –i 10). Both procedures result in guanine to adenine misincorporation patterns virtually identical to the standard mapping procedure reported in Figure 4, bottom (BWA aln option –i 5), suggesting that the detected pattern is not a mis-alignment by-product.



Supplementary figure 12. Example of a Agilent 2100 Bioanalyzer profile for extract TC21c. One microliter of the raw DNA extract has been run on an Agilent High Sensitivity DNA chip. DNA concentration and fragment length are reported on the y-axis (FU: Fluorescent Units) and x-axis (base pairs, bp), respectively. Although the major fraction of the extract consists of high molecular weight fragments that correspond to modern environmental DNA, fragments of a much lower molecular weight are detected and could indicate the presence of endogenous ancient DNA material. This has been further confirmed through PCR amplification of a horse-specific short mitochondrial fragment (see Methods), and shotgun sequencing on both Illumina and Helicos platforms.