**Title: Accurate sex identification of ancient elephant and other animal remains using low coverage DNA shotgun sequencing data**

Alida de Flamingh\*, Ashley Coutu†, ‡,§, Alfred L. Roca\*,\*\*,†† & Ripan S. Malhi \*,††,§§

***Author affiliation:***

\*Program in Ecology, Evolution and Conservation Biology, University of Illinois at Urbana-Champaign (UIUC), Urbana IL, 61801, USA

†Department of Archaeology, University of Cape Town, Cape Town, 7700, RSA

‡BioArCh, University of York, York YO10 5NG, UK

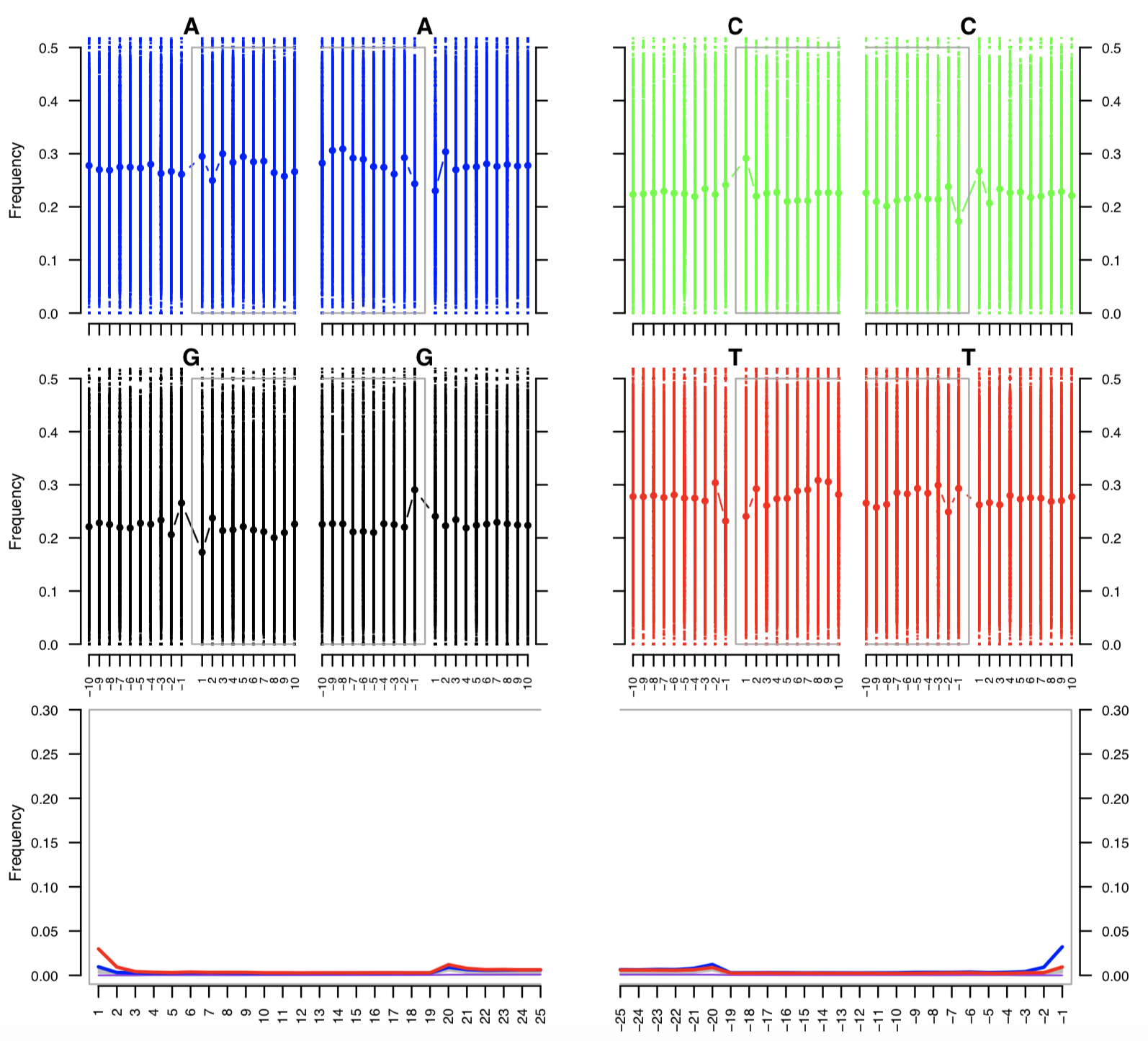
§Pitt Rivers Museum, University of Oxford, Oxford OX1 3PP, UK

\*\*Department of Animal Sciences, UIUC, Urbana IL, 61801, USA

††Carl R. Woese Institute for Genomic Biology, UIUC, Urbana IL, 61801, USA

§§Department of Anthropology, UIUC, Urbana IL, 61801, USA

**Supplementary Information – Tables and Figures**



Supplementary figure 1. DNA damage patterns showing fragmentation and misincorporation patterns. The top four plots show base frequency outside and in the read (grey box corresponds to the read), with base positions from the terminal position indicated below the graphs. The bottom plot shows base pair position-specific misincorporations from the 5’ (left) and the 3’ end (right), red shows C to T substitutions, and blue shows G to A substitutions.



Supplementary figure 2. Rx is the ratio of sequence read alignments to the *X* chromosome compared to sequence read alignments to all autosomes. Rx values are shown for low coverage ancient elephant genomes (black), and for datafiles that were subsampled to include ~10 000 (gray) and ~1000 reads (white). An Rx ratio with an upper 95% CI of less than 0.6 indicates male sex, and an Rx ratio with a lower 95% CI that is greater than 0.8 indicates female sex. For samples that have 95% CI outside of the Rx cut-off values, the Rx method provides the sex identification output as “sample is consistent with XX” or “sample is consistent with XY” rather than definitively scoring the sample as either “male” or “female”. This occurred in 5 cases in the subsampled datafiles (indicated with asterisks).

Supplementary table 1. Examples of organisms for which the reference genome was generated from a female animal or where only the *X* and not the *Y* chromosome is reported in the GenBank reference genome assembly.

|  |  |  |
| --- | --- | --- |
| Common name | Scientific name | GenBank assembly number |
| Domestic cat | *Felis catus* | GCA\_000181335.4 |
| Domestic sheep | *Ovis aries* | GCA\_002742125.1 |
| Domestic horse | *Equus caballus* | GCA\_002863925.1 |
| Dromedary camel | *Camelus dromedarius* | GCA\_000803125.2 |
| European rabbit | *Oryctolagus cuniculus* | GCA\_000003625.1 |
| Domestic goat | *Capra hircus* | GCA\_000317765.2 |
| Sumatran orangutan | *Pongo abelii* | GCA\_002880775.3 |
| Western lowland gorilla | *Gorilla gorilla gorilla* | GCA\_000151905.3 |
| Gelada | *Theropithecus gelada* | GCA\_003255815.1 |
| Meerkat | *Suricata suricatta* | GCA\_006229205.1 |

Supplementary table 2. Genome alignment statistics for modern and ancient elephant genomes. Breadth of coverage was calculated as the proportion of the genome covered by at least 1 read. Average coverage was calculated as the depth of coverage across the genome for all positions in the genome, including those with zero coverage.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample ID | Sample type | Breadth of coverage (%) | Average  coverage (X-fold) |
| DS1531 | Modern | 45.4242 | 0.721645 |
| DS1548 | Modern | 56.2786 | 0.982199 |
| DS1514 | Modern | 39.1834 | 0.564377 |
| DS1543 | Modern | 50.7346 | 0.826475 |
| DS1506 | Modern | 55.9282 | 0.969139 |
| LO3503 | Modern | 52.5948 | 0.864645 |
| LO3509 | Modern | 53.8217 | 0.897441 |
| LO3511 | Modern | 49.8096 | 0.791459 |
| LO3521 | Modern | 47.3061 | 0.737514 |
| LO3514 | Modern | 49.4394 | 0.793958 |
| GR0041 | Modern | 51.025 | 0.821681 |
| GR0037 | Modern | 49.8363 | 0.806842 |
| WA4013 | Modern | 12.3095 | 0.141394 |
| WA4020 | Modern | 61.1214 | 1.09633 |
| BE4059 | Modern | 63.3544 | 1.16019 |
| B6079 | Ancient | 7.23345 | 0.153929 |
| B6025 | Ancient | 1.4834 | 0.0300563 |
| B8030 | Ancient | 7.96246 | 0.169617 |
| B6059 | Ancient | 5.23486 | 0.109229 |
| B6074 | Ancient | 1.27702 | 0.0258847 |
| B6022 | Ancient | 2.69254 | 0.0550849 |
| B6062 | Ancient | 1.35766 | 0.0275851 |
| B6051 | Ancient | 0.637552 | 0.0128607 |
| B6130 | Ancient | 20.6747 | 0.496995 |
| B6520 | Ancient | 0.193728 | 0.00388982 |

Supplementary table 3. Rx ratio values for the ancient ivory samples calculated using all sequence reads (Rx all), approximately 10 000 sequence reads (Rx ~ 10 000), and approximately 1000 sequence reads (Rx ~ 1000).

|  |  |  |  |
| --- | --- | --- | --- |
| Ancient sample ID | Rx all | Rx ~ 10 000 | Rx ~ 1000 |
| B6022 | 0.495573 | 0.4483153 | 0.3957278 |
| B6025 | 0.491787 | 0.5043557 | 0.1873106 |
| B6051 | 0.490834 | 0.4407088 | 0.4415854 |
| B6059 | 0.482732 | 0.493083 | 0.3499586 |
| B6062 | 0.968406 | 0.9666384 | 1.029646 |
| B6074 | 0.491222 | 0.4447996 | 0.2836775 |
| B6079 | 0.477581 | 0.5970461 | 0.6939766 |
| B6130 | 0.474589 | 0.5681135 | 0.8121833 |
| B6520 | 0.491588 | 0.4147076 | 0.5607132 |
| B8030 | 1.00599 | 0.9682274 | 0.8207848 |