Supplementary online material - a practical example on how to use *bammds* and determine if a library is heavily contaminated

Anna-Sapfo Malaspinas*, Ole Tange*, J. Víctor Moreno-Mayar, Morten Rasmussen, Michael DeGiorgio, Yong Wang, Cristina E. Valdiosera, Gustavo Politis, Eske Willerslev, Rasmus Nielsen

1 Labwork experiments and mapping of raw data for "Gus"

The Arroyo de Frías human skeletons (Museo de La Plata MLP 5582) were recovered between 1870 and 1873 by Florentino Ameghino near the town of Mercedes, 100 km west from Buenos Aires, Argentina (see summary in [11]). The discoveries were made in the bed of the creek during a drought period [3]. Based on the stratigraphic profile reported by Florentino Ameghino, and on his original interpretations [1, 2], the bones may have been buried in the upper part of the Guerrero Member of the Luján Formation (named "Lujanense" in [1]) or in the upper part of the Pampean Formation, and below the paleosoil located between the Guerrero and Río Salado Members of the same formation (Puesto Callejón Viejo geosoil; [4, 5]). Florentino Ameghino found the skeletal remains of at least two individuals on the left bank of the creek, at a depth of 2.5 to 3 meters below ground level. The best preserved and complete human skeleton was identified as an adult female. It was almost entirely articulated, lying in a flexed position on its right side [3]. The position indicated an intentional, primary burial that had minor evidence of postdepositional disturbance. Few skeletal elements belonging to a taller and more robust individual (probably a male) were recovered for the second invidual. The female skull recovered during the 1870s has been missing since the early 1890s [7].

Two phalanges, probably from the same individual, recovered by Florentino Ameghino in 1873, were radiocarbon dated to $10,300 \pm 60^{14}$C yrs BP (CAMS-16598) and $9,520 \pm 75\ ^{14}C$ yrs BP (OxA-8545) [11]. To our knowledge, those are the oldest dated human remains in the Southern Cone of South America. We recovered DNA from one of the two phalanges ("Gus" in what follows).

1.1 Sample preparation and DNA extraction

A bone fragment was ground into powder using a multitool drill (Dremel), and two DNA extractions were done based on 200mg of bone powder each. The powder was digested overnight at 37°C with agitation in a 5ml of buffer containing 5M EDTA pH 8 and 10% proteinase K solution. The digested sample was further extracted using a silica binding method [18]. Final elution volumes were in $50\mu l$ EB buffer.

1.2 Library building, amplification and sequencing

The DNA extract was used to build an Illumina (San Diego, CA) index library (T-A overhang ligation method), using the Rapid Library kit from Roche-454 (Branford, CO), with the following modifications to the manufacturer's protocol. For each library, the fragmentation step was excluded, $16\mu l$ of DNA extract was used and mixed with $2.5\mu l$ RL 10x PNK buffer, $2.5\mu l$ of RL ATP, $1\mu l$ RL dNTP, $1\mu l$ RL T4 polymerase, $1\mu l$ RL PNK and $1\mu l$ of RL Taq polymerase. The mix was incubated at 25°C for 20 min, 72°C for 20 min and then placed at 4°C. One $1\mu l$ of Illumina indexing adaptor mix and $1\mu l$ RL ligase were added and the sample was incubated for 10 min at 25°C. Finally, the library was purified on a MinElute spin column according to protocol and eluted in $30\mu l$ of EB.

Amplification of the purified library was done using Platinum® Taq DNA Polymerase High Fidelity polymerase (Invitrogen) with a final mixture of 10X High Fidelity PCR Buffer, $50mM$ Magnesium Sulfate, $0.2mM$ dNTP, $0.5\mu M$ Multiplexing PCR primer 1.0 and $0.01\mu M$ Multiplexing PCR primer 2.0 and $0.5\mu M$ PCR primer Index 7, 3% DMSO, $0.02U/\mu l$ Platinum HiFi polymerase, 10-$20\mu l$ of template and water to $50\mu l$ final volume. Primers are part of Illumina's Multiplexing Sample Prep Oligo Kit. Cycling conditions were as follows: a 3 min activation step at 94 ^{o}C, followed by 15 cycles of 30 s at 94 ^{o}C, 20 s at 60 ^{o}C, 20 s at 68 ^{o}C, with a final extension of 7 min at 72 ^{o}C. The amplified library was run on a 2% agarose gel, and gel purified using Qiagen gel extraction kit, following manufactures guidelines. A second amplification of 12 cycles was performed using as template the amplicons of the first PCR using the same conditions. PCR products were finally gel purified before sequencing. Concentration and size profile was determined on a Bioanalyzer 2100 using a High Sensitivity DNA chip.

The libraries were sequenced on an Illumina HiSeq2000 (95 cycles) machine at the National High-throughput DNA Sequencing Center in Denmark (NSC, http://seqcenter.ku.dk/).

1.3 Mapping for Gus

Around 13 million single end reads were sequenced (*i.e.*, 6.5% of a lane by early 2014's standards, which corresponds to around 120 USD at the NSC). The raw reads were trimmed using *AdapterRemoval-1.1* [10] for adapter sequence and leading/trailing Ns

to a minimum length of 25 nucleotides (–minlength 25 –trimns). Statistics for each sample are given below. Hereafter the reads were mapped to the human reference genome build37.1 using *bwa-0.6.2* [8] with seed disabled to allow for better sensitivity [16]. Alignments were filtered for reads with a mapping quality of at least 30, sorted and merged to libraries using *Picard* (http://picard.sourceforge.net). Duplicates were removed using *Picard MarkDuplicates* at the library level. The mapped data was stored in a file "Gus.Ancient.bam" that we distribute with the tool for testing.

The level of endogenous DNA was determined as the percentage of mapped reads after filtering for a mapping quality of 30 (q30) compared to the raw amount of reads produced. Coverage and average read depth were estimated using *BEDtools* [13] and *pysam* (http://code.google.com/p/pysam/). Statistics related to coverage (*i.e.*, the number of positions in the genome covered by at least one read) and depth of coverage (DoC, *i.e.*, the average number of reads covering each position) for Gus are:

	#raw numbers	#mapped (q30)	dupli	Gus.Ancient.bam	endogenous	covered>1read	average DoC
Gus	13'137'520	67'260	5.34%	63'665	0.48%	0.13%	0.001

Only a small fraction of the DNA sequenced maps to the reference human genome (<1%) as would be expected for ancient human remains.

2 Is it worth sequencing more of this library?

2.1 *bamdamage*, damage plots and read length distribution

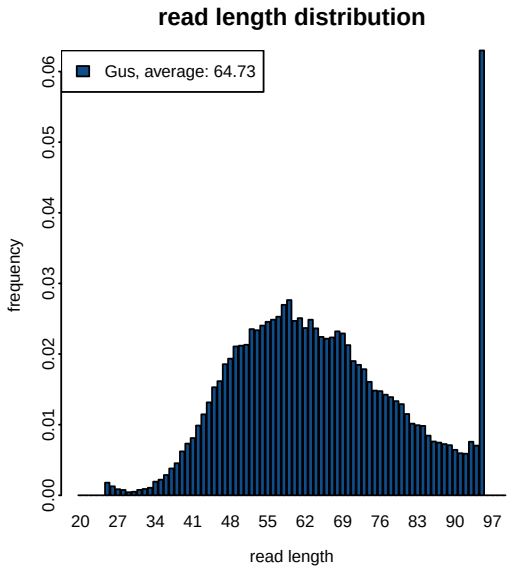
Ancient DNA is characterized by a number of features which include being fragmented and damaged (typically an increase of C to T and G to A changes at the read termini). These features can be used to diagnose if the DNA is endogenous or if it is the result of contamination (*e.g.*, [17]).

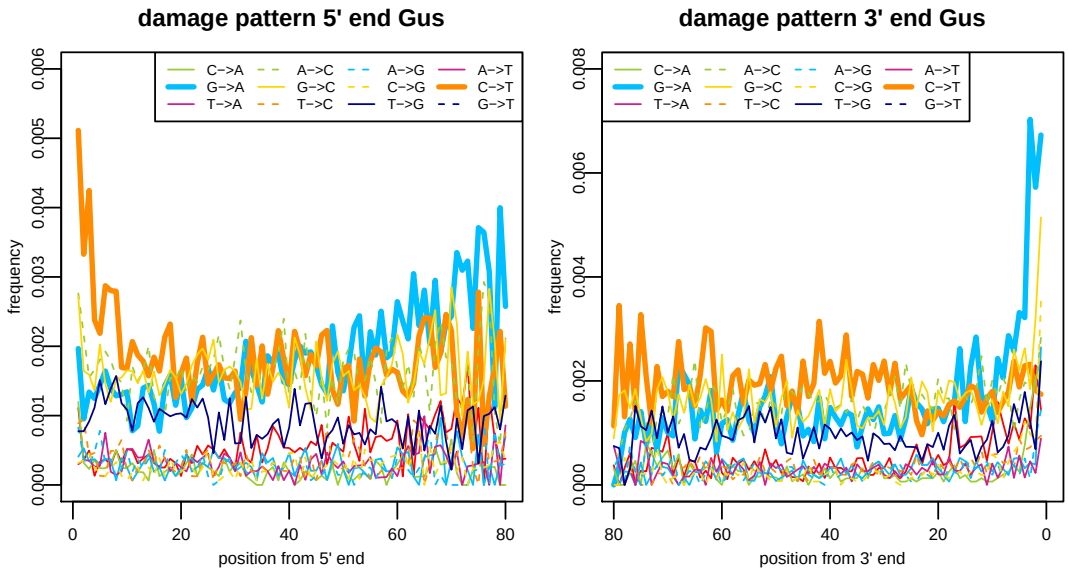
Once the data is mapped, we suggest as a first step to look into the read length distribution and the damage patterns across the reads. We added a simple script to plot those statistics in the package we distribute but such statistics can also be obtained with existing tools (*e.g.*, [6]).

By running:

```
bammds -s Gus Gus.Ancient.bam
```

We obtain the two following plots (`Gus.Ancient.dam.pdf`):





In terms of read length, these results suggest the reads are short (shorter than the number of cycles used for the sequencing) and that the read length distribution is unimodal, which is compatible with ancient DNA that has not been contaminated by present-day DNA.

Moreover, although C to T and G to A are increased compared to other substitutions at the read termini, the damage levels are much lower than expected, specifically around 0.5%. For example, Sawyer et al. [15] suggest that for a sample around 9'000 years old, the C to T misincorporation should on the order of 10%. This casts some doubt as to the origin of the DNA that maps to the human genome.

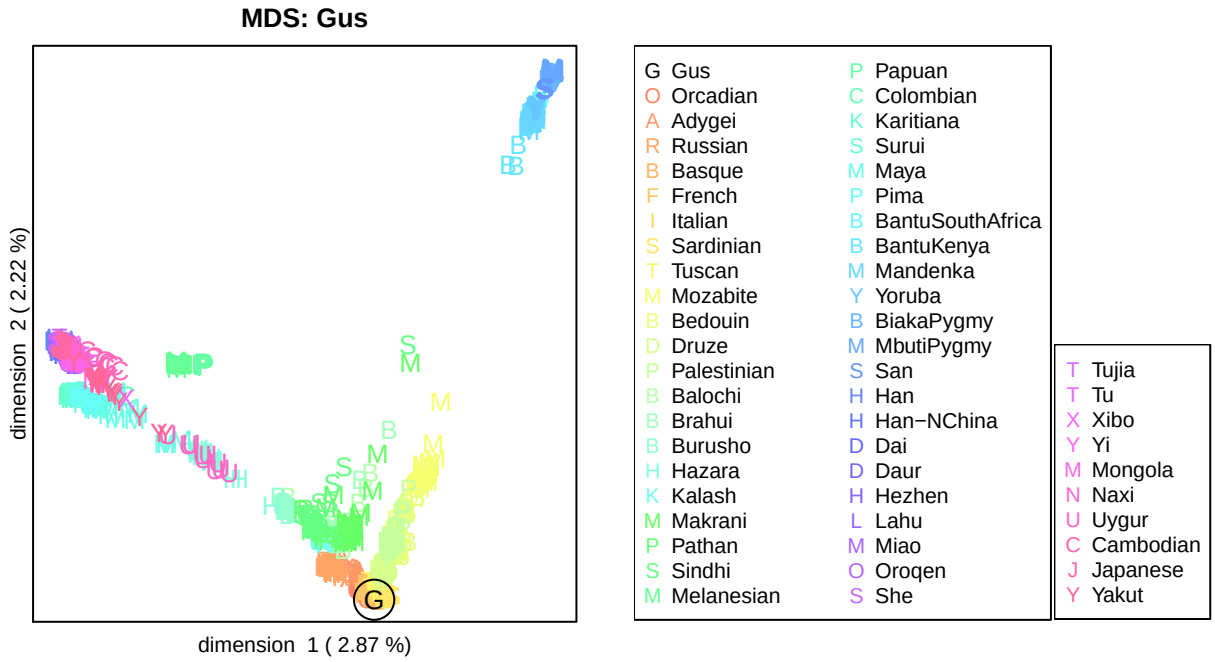
2.2 *bammds*, ancestry determination

Alongside the tool, we also distribute a reference panel in the native format we used to develop the tool (`HGDP_hg19_one_allele.txt`). This data is publicly available [9]. (Note that the tool can be run with a reference panel in several other formats, including *plink* [12]). Our simulations results (based on this same panel) suggest that a depth around 0.001X represents sufficient data to determine the broad geographic region of origin of the sample of interest. In some cases, it also allows to recover the closest population - but we will not attempt to do so here. Note that the required depth will also depend on the amount of damage and sequencing error - so it is best to consider 0.001X a lower bound in practice.

By running (on an 8 cores, 2.2 GHz machine: running time (wall clock time) 17:01.05, memory usage 18.1 Gb):

```
bammds Gus.Ancient.bam HGDP_hg19_one_allele.txt
```

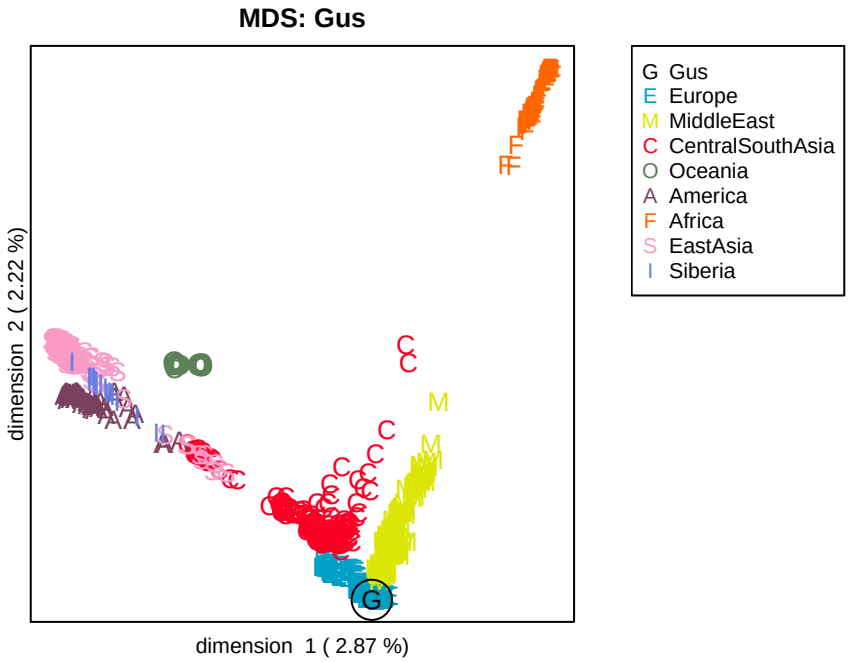
We obtain the following plot (`Gus.Ancient,HGDP_hg19_one_allele.pdf`, first two pages)



The last page indicates the number of markers that overlap between the sample and the reference panel, where the panel includes >600'000 SNPs that overlap with Gus at ~1'000 positions. *bammds* assigns colors to each population using a color wheel by default. One can also run the code with user defined colors and by grouping the populations into broader regions. This is done by modifying a csv file generated by *bammds* under a tmp directory. We provide such a file as an example, named `HGDP_hg19_one_allele.legend.csv`. To run:

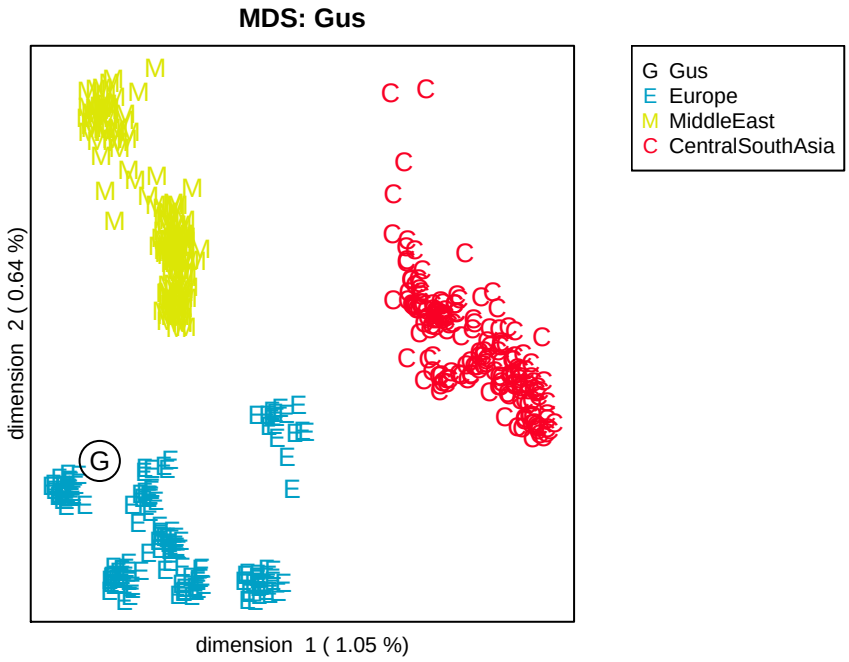
```
bammds -l HGDP_hg19_one_allele.legend.csv --nosum Gus.Ancient.bam HGDP_hg19_one_allele.txt
```

We obtain the following plot (running time (wall clock time) 0:58.54, memory usage 2.3 Gb: in this case the allele sharing matrix is reused and the `--nosum` option allows to skip the step where the overlapping SNPs are computed).



One can also remove populations or individuals by modifying the csv file `HGDP_hg19_one_allele.legend.csv` to "zoom in". As an example, we restrict the plot to samples from Europe, Middle East and Central South Asia:

```
bammds -l HGDP_hg19_one_allele.legend.EuropeAsia.csv --nosum Gus.Ancient.bam HGDP_hg19_one_allele.txt
```

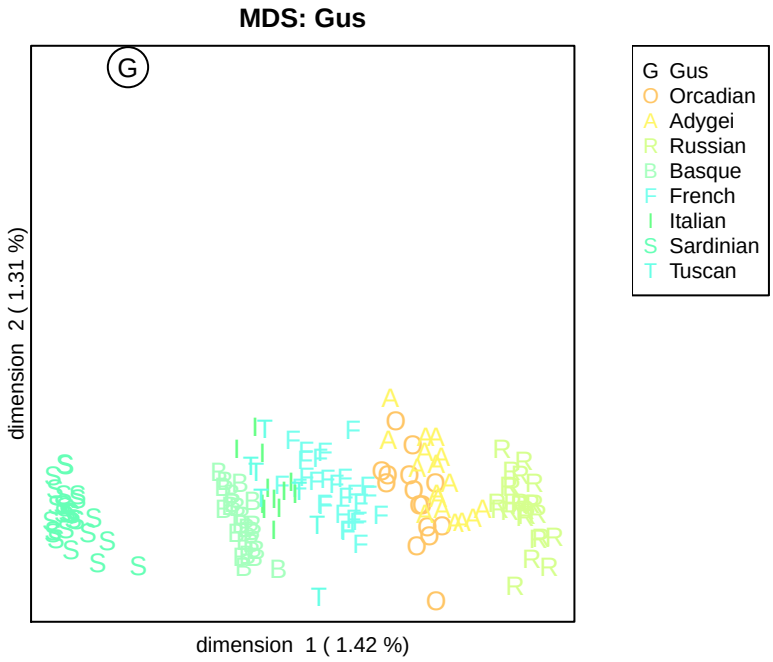


Those results suggest that the sample Gus is highly contaminated by a European descent individual(s) (since it dates to pre-Columbian times and therefore we do no expect any European ancestry, see *e.g.*, [14]), as suggested by the damage plots above. Depending on resources, one might consider different strategies from this point on. In our case, we discarded these libraries.

Our experience is that for heavily contaminated samples, such as this one, this approach allows to determine within an hour (once the sequence data is produced and mapped) if a library is worth pursuing or not - provided one has the appropriate reference panel. Ideally, this panel should include ancestral populations for the contaminant and the sample. When the contamination is not as high, more effort is needed to establish if a sample is contaminated or not. Indeed, a partially contaminated sample would tend to look like an individual with admixed ancestry between the contaminant ancestral population and the actual ancestral population, making the signal less clear.

For each reference panel used, it is important to determine if more data is needed to establish the ancestry of the sample - *i.e.*, guidelines on how much data is enough data. Cases where the goal is to distinguish two divergent populations are of course easier (here Native Americans versus Europeans). Simulations similar to the ones presented in the main text for another reference panel should allow to establish these guidelines. In other words, if possible, one would consider published higher depth genomes of known ancestry. One would first determine the level of resolution while considering all the data of these high depth genomes. The idea is to then (1) downsample the high depth genome until the assignment becomes incorrect, (2) consider the level of resolution for a depth similar to the newly sequenced data. The hope is that (1) the option `-z` to downsample the bam file, (2) the population assignment in as determined by the distance to the centroid of each populations and (3) the flexibility in the legend file would help determining appropriate levels of resolution (in the case of the HGDP panel, natural groups are geographical regions and countries). For those simulations, to minimize the bias introduced by different mapping strategies, one should also map the high depth genome in the same way as the newly generated data (with same reference genome).

In our case for example, we do not believe we have enough data to determine where in Europe the individual comes from (with this method). We compute a plot including only Europeans:



This plot shows a result typical of cases where the "missingness" of the data dominates the signal (cases where the allele sharing distance between the ancient sample and the individuals in the reference panel is too far from the true value): Gus looks like an outlier to all other European populations. In this case, we would conclude that more data is needed to determine the closest European population for Gus.

Finally, the tool allows to plot several bamfiles on the same plot. We suggest to always consider running the tool for each bamfile separately since several low depth bamfiles are likely to have very few overlapping positions. It may happen that the bamfiles cluster together on the mds plot because they are all outliers, *i.e.*, by chance only and not because they have shared ancestry.

References

[1] F. Ameghino. *Contribucion al Conocimiento de los Mamiferos Fosiles de la Republica Argentina*, volume 6. Actas de la Academia Nacional de Ciencias de Cordoba, Buenos Aires, 1889.

[2] F. Ameghino. *La Antiguedad del Hombre en el Plata. In Obras Completas y Correspondencia Cientifica de Florentino Ameghino*, volume III. La Plata, Argentina, taller de impresiones oficiales edition, 1915.

[3] F. Ameghino. *Los problemas geo, arqueo y paleoantropologicos de la Argentina. In Obras Completas y Correspondencia Cientifica de Florentino Ameghino*, volume XIX. La Plata, Argentina, taller de impresiones oficiales edition, 1935.

[4] F. Fidalgo, F. O. De Francesco, and U. R. Colado. Geologia superficial de las hojas castelli, j. m. cobo y monasterio (provincia de buenos aires). In *Actas del 5 Congreso Geologico Argentino*, pages 27–39, Buenos Aires, Argentina, 1973.

[5] E. Fuchs and C. M. Deschamps. Depositos continentales cuaternarios en el noreste de la provincia de buenos aires. 63:326–343, 2008.

[6] Aurelien Ginolhac, Morten Rasmussen, M. Thomas P. Gilbert, Eske Willerslev, and Ludovic Orlando. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics*, June 2011. PMID: 21659319.

[7] A. Hrdlicka. *Early Man in South America*. Bureau of American Ethnology Bulletin, Washington, D.C., USA, 1912.

[8] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. PMID: 19451168 PMCID: PMC2705234.

[9] Jun Z. Li, Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, Gregory S. Barsh, Marcus Feldman, Luigi L. Cavalli-Sforza, and Richard M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104, February 2008. PMID: 18292342.

[10] Stinus Lindgreen. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res Notes*, 5:337, July 2012. PMID: 22748135 PMCID: PMC3532080.

[11] G. Politis, G. and T. Stafford. *Revisiting Ameghino: new C14 dates from ancient human skeletons from the Argentine Pampas. In Peuplements et Prehistoire en Amerique.* Vialou, D., Paris, France, editions du comite des travaux historiques et scientifiques edition, 2011.

[12] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, September 2007. PMID: 17701901.

[13] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010. PMID: 20110278 PMCID: PMC2832824.

[14] Morten Rasmussen, Sarah L. Anzick, Michael R. Waters, Pontus Skoglund, Michael DeGiorgio, Thomas W. Stafford Jr, Simon Rasmussen, Ida Moltke, Anders Albrechtsen, Shane M. Doyle, G. David Poznik, Valborg Gudmundsdottir, Rachita Yadav, Anna-Sapfo Malaspinas, Samuel Stockton White V, Morten E. Allentoft, Omar E. Cornejo, Kristiina Tambets, Anders Eriksson, Peter D. Heintzman, Monika Karmin, Thorfinn Sand Korneliussen, David J. Meltzer, Tracey L. Pierre, Jesper Stenderup, Lauri Saag, Vera M. Warmuth, Margarida C. Lopes, Ripan S. Malhi, SÃžren Brunak, Thomas Sicheritz-Ponten, Ian Barnes, Matthew Collins, Ludovic Orlando, Francois Balloux, Andrea Manica, Ramneek Gupta, Mait Metspalu, Carlos D. Bustamante, Mattias Jakobsson, Rasmus Nielsen, and Eske Willerslev. The genome of a late pleistocene human from a clovis burial site in western montana. *Nature*, 506(7487):225–229, February 2014.

[15] Susanna Sawyer, Johannes Krause, Katerina Guschanski, Vincent Savolainen, and Svante Pääbo. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*, 7(3):e34131, March 2012.

[16] Mikkel Schubert, Aurelien Ginolhac, Stinus Lindgreen, John F. Thompson, Khaled AS AL-Rasheid, Eske Willerslev, Anders Krogh, and Ludovic Orlando. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, 13(1):178, May 2012. PMID: 22574660.

[17] Pontus Skoglund, Bernd H. Northoff, Michael V. Shunkov, Anatoli P. Derevianko, Svante Pääbo, Johannes Krause, and Mattias Jakobsson. Separating endogenous ancient DNA from modern day contamination in a siberian neandertal. *PNAS*, page 201318934, January 2014. PMID: 24469802.

[18] Dongya Y. Yang, Barry Eng, John S. Waye, J. Christopher Dudar, and Shelley R. Saunders. Improved DNA extraction from ancient bones using silica-based spin columns. *American Journal of Physical Anthropology*, 105(4):539–543, 1998.