Report of the extended bioinformatic DNA analysis of the Nasca tridactyl bodies

---

Note: The updated version of this report is available at :
https://docs.google.com/document/d/1SSrKX-TibnrvCa55woImiPvmO3OUps3HYFVnWTaT9Bw/edit?usp=sharing

Report of the extended bio-informatic DNA analysis of the Nasca tridactyl bodies

Introduction

This report details the results obtained from multiple bioinformatic analysis extending previous independent analysis and reports by other groups in Mexico, the United States, and Russia to massive sequencing data obtained from tridactyl bodies of Nasca DNA. This report develops in particular the analysis performed on the CEN4GEN laboratory sequencing data. (Https://cen4gen.org/)

All analysis were done with open source software developed by large bioinformatics and genomics groups and upon request and with the consent to use data from Jaime Maussan's team. The samples analyzed for this report were those obtained from the body called Victoria labeled as Ancient002 and Ancient004 in CEN4GEN data that derive from MDA amplification of CEN4GEN DNA from both neck tissue (Ancient002) as well as DNA previously extracted from the body by Biotecmol (Ancient004), respectively.

The main analysis performed for this report are as follows:

1. Mapping with relaxed parameters of the complete set of original readings of the two Victoria samples to the human genome (hg19 version).

2. The most complete vicuña genome mapping available in the NCBI of Victoria's total readings.

3. Iterative analysis of overlap readings without significant correspondence with a selection of genomes from the NCBI DNA sequence refseq database

4. Analysis of "taxmaps" of the single readings proportions without significant correspondence with the reference and reference databases of NCBI DNA sequences.

5. Investigation of the presence of universal primers for identification of vertebrates in the sequencing data.

6. Finding coding capability in overlapping readings.

7. Quantitative summary of the proportions of readings observed in the main analysis database.

In the following sections, each of the analysis performed is detailed.

Human Genome Mapping (GRCH37 version) with relaxed parameters of the complete set of readings of the two Victoria samples.

A new mapping of the human genome was carried out using the bwa software (Heng Li, 2013) of all the complete readings sequenced by CEN4GEN with more flexible parameters, that is to say without rejecting low quality maps, in order to evaluate a wider margin on Victoria samples that could match with human DNA. The results are shown in the following table for the total readings obtained for each sample :

Measures	Ancient 002 with mapping	Ancient 002 percentages	Ancient 004 with mapping	Ancient 004 percentages
Initial complete readings	1,123,330,640	100	1,003,400,490	100
Readings with permissive maps to human	654,603,960	58,27	481,988,213	48,04

Table 1: Total mapped readings of the human genome using bwa, with no restrictions on mapping quality and default settings

This shows that even with a larger and more flexible mapping margin, it is not possible to map the total readings on the human genome and that in Ancient004 this represents less than 50% of the mapping. However, it should be noted that the increase in the mapped reading margin is largely dominated by poor quality mappings, as shown in the following graph of the quality distribution in the mapping of data corresponding to the mapping under this scheme of the example Ancient002 :

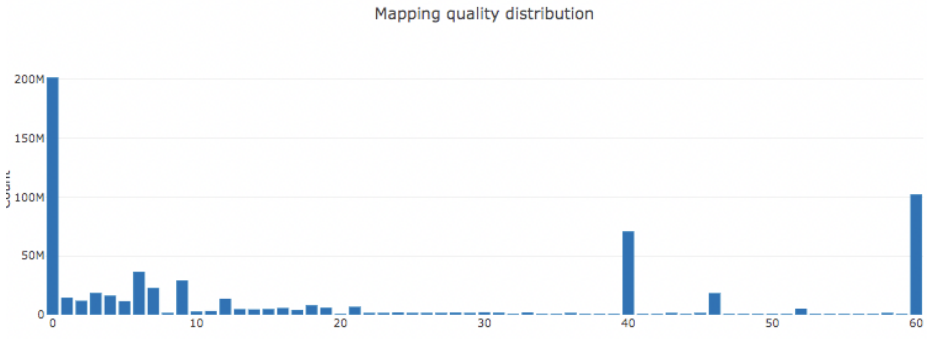

Figure 1: Distribution of ancient002 Ancient readings mapping qualities in the bwa result without quality restriction using Alfred with predetermined parameters.

In addition, in order to confirm the levels of human genome mapping between Victoria samples and the GRCH37 human genome, a re-mapping of the total readings of each sample was performed using Kart v2.4.5 software (Hsin -Nan Lin and Wen-Lian Hsu, 2017).) to get a faster result that is very similar to that provided by bwa but with greater specificity, including a greater restriction in the quality of the mapping.

The results were as follows :

Measures	Ancient 002 with mapping	Ancient 002 percentages	Ancient 004 with mapping	Ancient 004 percentages
Initial complete readings	1,123,330,640	100	1,003,400,490	100
Readings with permissive maps to human	115,329,515	10,27	88,285,395	8,80

Table 2: Total number of mapped GRCH37 genome readings of Victoria samples using Kart software with default settings.

This confirms previous reports indicating a weak coincidence with the human genome of the sequencing results of the Victoria samples when considering mappings with more demanding quality filters, as shown in the following graph of the distribution of qualities in the mapping of the data samples corresponding to this mapping scheme with the highest quality restriction level for the Ancient002 sample :

Figure 2 : Map distribution of the Ancient002 total readings in the Kart result with no restrictions on qualities and default settings.

The quality distribution analysis shown for the cartography were carried out with the software Alfred v0.1.17 (Tobias Rausch, Markus Hsi-Yang Fritz, Jan O Korbel, Vladimir Benes, 2018) on the classified and indexed bam files of each cartography.

The most complete vicuña genome mapping available on the NCBI of Victoria's total readings.

To evaluate the possibility that Victoria samples came from parts of a vicuña body, the mapping of the total sequenced readings of the two Victoria samples, was evaluated in comparison with the most complete genome assembly close to the available vicuña in the refseq database of NCBI, which was the *Vicugna pacos* or *alpaca* (Ensemble GCF_000164845.2). For this analysis, the Kart v2.4.5 software (Hsin-Nan Lin and Wen-Lian Hsu, 2017) was used to obtain a faster result very similar to that provided by bwa.

The results are presented in the following table :

Measures	Ancient 002 with mapping	Ancient 002 percentages	Ancient 004 with mapping	Ancient 004 percentages
Initial complete readings	1,123,330,640	100	1,003,400,490	100
Readings with vicuña mapping	8734233	0,78	22715081	2,26

Table 3: Total readings mapped on the Vicuña Assembly GCF_000164845.2 of Victoria samples using Kart software with default settings.

This shows that coincidence percentages with the Vicuña genome would be minimal for the analyzed samples from Victoria and do not appear to be significant.

In addition, to confirm these levels of coincidence between the readings and the vicuna genome, an analysis was performed by coincidence of *kmeros* using the "Seal" software of BBtools in its version 38-25 (Bushnell B. sourceforge .net / projects / bbmap /), developed by the US Department of Energy's Joint Genome Institute, to compare the number of size 31 (*kmeros*) subsequences between the vicuña genome and full readings of the Victoria sequencing data.

The results obtained are as follows :

Measures	Ancient 002 with mapping	Ancient 002 percentages	Ancient 004 with mapping	Ancient 004 percentages
Initial complete readings	1,123,330,640	100	1,003,400,490	100
Readings with vicuña *kmeros* correspondance	36,470,068	3,25	69,633,978	6,94

Table 4: Total readings with coincidences in at least one kmero with Vigogne assembly GCF_000164845,2 of Victoria samples using Seal software with parameters forbidn = t, ambigu = first, Rskip = 1 and other parameters by default.

Iterative analysis mapping of the ratios of overlapped unified reads without a significant match to a selection of genomes from the NCBI DNA sequence refseq database.

To explore the types of known genomes that sequenced reads might belong to, without mapping them or doing a BLAST of all these sequenced reads known genomes, as this would take too much time and would not be viable in time and computational capacity available, we started by extracting readings that could be related by overlaps, with PEAR v0.9.6 software (Zhang, Kobert, Flouri and Stamatakis, 2014) then with these readings of smaller volume but longer longitude, we proceeded to a technique of comparison by "sketching" and by looking for the *kmeros*, with the software "sendketch" and "seal" of the sets BBtools in version 38-25.

The "sendketch" software makes it possible to look for the similarity between large amounts of readings and the database is made faster and easier than with alignment techniques such as "mappings" or BLAST, while that "*seal*" makes it possible to look for any subsequence of 31 bases (*kmeros* of 31) contained in a genome. Using "sendketch", the genomes and types of organisms that may look like the overlapped sequential reads were identified and once the genomes with greater coincidences were identified, a genome group of the type having given the largest number of matches in "sendketch" has been downloaded. Then the readings, with respect to these genomes, were compared to look for those having a similarity in at least one *kmero* of 31 and those that did not have *kmero* of 31 were selected for a new series of searches with "sendketch" and a comparison with "seal", etc.

Iteratively, this was done for all types of genomes that would mostly match with "sendketch" for readings without the presence of *kmeros* 31, from the genome type detected during the previous iteration until reaching the genomes of protozoa type. For genomes of micro-organisms, all genomes available in refseq were downloaded, while for animal genomes, only the genomes indicated by "sendketch" and others, suggested by various participants in the analysis, have been downloaded.

Below is a chart with a diagram of the procedure :

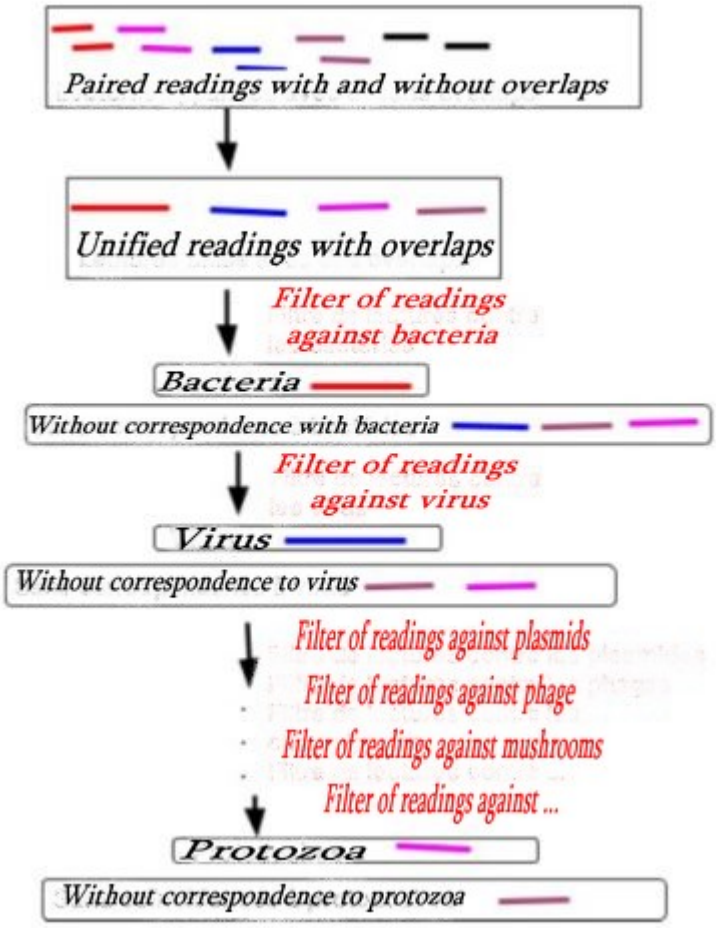

Figure 3: Iterative filtering strategy developed to find similarities between single union readings by *overlaps* and 37,877 of several types of refSeq genomes.

Following this strategy, the overlaps readings of each sample were compared to 37,877 genome sequences in total, and the results are presented below in a table with the numbers and percentages of reads with significant similarity, according to the software "seal", resulting from this comparison using the iterative search technique with "sendsketch" followed by comparison with "seal".

The table shows the types of genomes in the order in which they were compared to the superimposed results.

Report of the extended bioinformatic DNA analysis of the Nasca tridactyl bodies

Reading each step of the comparison	Ancient 002	%	Ancient 004	%
Complete readings	1,123,330,640	100	1,003,400,490	100
Overlapped readings (total to be taken into account for the following percentages)	103,600,552	18.45	338,014,892	67.37
United readings matching with bacterial	6,075,940	5.86	26,840,573	7.94
United readings matching with virus	241,181	0.23	2,734,450	0.81
United readings matching with plasmide	60,695	0.06	309,502	0.09
United readings matching with phage	954,886	0.92	652	0.00
United readings matching with mushrooms	707,072	0.68	6,857,348	2.03
United readings matching with platids	6,298,041	6.08	11,405	0.00
United readings matching with diatom	25,086	0.02	295,137	0.09
United readings matching with human	6,612,714	6.38	35,325,290	10.45
United readings matching with Bos Taurus	105,784	0.10	44,620,181	13.20
United readings matching with H. penzbergensis	139	0.00	31,208,623	9.23
United readings matching with P.vulgaris	53,527,894	51.67	25,633	0.01
United readings matching with various genomes	882,181	0,85	NA	NA
United readings matching with other vertebrates genomes	130,249	0.13	2245614	0.66
United readings matching with Protozoan	4,169	0.00	7768	0.00
United readings with total matching	75,626,031	73.00	150,482,176	44.52
United readings with total matching	27,974,521	27.00	187,532,716	55.48

Table 5: Numbers and percentages of readings showing significant similarity in at least one cell of 31 bases returned by the "seal" software with respect to a selection of refseq genomes

Note: The complete list of sequences used for each type of organism can be found in the following link :

https://drive.google.com/file/d/15r-tKv94UgHGtQd9owHGWO3bqeZEIdpV/view?usp=sharing

This shows that by passing through a filter including a large line of known genomes and organism types, 100% of the genomic origin of the overlaps obtained from the Victoria samples is not found. In particular, about 27.00% and 55.48% of the overlaps of the Ancient002 and Ancient004 samples, respectively, can not be associated with any type of organism of all types of organisms and species accumulated in this analysis.

Analysis by "taxmaps" of the proportions of unique total readings without significant correspondences with the taxonomies and DNA sequences of the refseq databases and NCBI DNA sequences.

In order to evaluate with the widest possible margin the coincidence by similarity between the sequenced reads for Victoria and all DNA sequences recorded in NCBI, as well as their respective possible taxonomic levels, the duplicate readings of each Victoria sample were taken, that is to say, from the total readings, those that were redundant and that were very similar to others, in orfer to have a set of readings without redundancy between them. This was done separately for the "*forward*" and "*reverse*" readings of each set of paired reads.

From this set of readings, without duplicates, we used the corresponding reading to find the proportion of the sequenced DNA with coincidence by similarity with the most complete database available today, which is the one formed by the nt databases and refseq added to NCBI. This research was performed by generating a search from all sequences recorded in the "nt" and "Refseq" NCBI databases as implemented in the *taxmaps* v 0.2.1 software index (Corvelo, Clarke, Robine & Zody, 2018). This software was used for searching against the above index with "*forward*" reads of all data without duplicates of each sample and with the parameters -e 0.2 and – m s and the rest of the parameters by defaultt to eliminate low complexity readings or with the presence of indeterminate bases.

The results are as follows :

Report of the extended bioinformatic DNA analysis of the Nasca tridactyl bodies

Figure 4: Distribution diagrams of the proportions of readings showing a significant similarity to the nt and refseq database of the NCBI and without it (dark red color) for each sample.

In the bright red portion (unclassified section) of each chart, the percentage of single reads with no significant similarity to the Refseq and "nt" databases that would appear in the Victoria sequencing readings data is shown. This shows that there are over 50% sequential non-redundant reads with no indeterminate bases or low complexity levels that are unlike "nt" and "refseq" NCBI databases. This result also confirms the high levels of unclassified sequence percentages obtained from previous "Abraxas" analysis on overlapping reads.

Investigation of universal primers presence for identification of vertebrates in sequencing data.

To evaluate the possibility of identifying specific vertebrate species present in the sequenced samples, which could be specifically determined using the sequencing data, the presence of universal primers for the detection of vertebrate species has been identified in the complete readings of each sample. These primers are extracted from the publication "Two universal primer sets for species identification among vertebrates" (Kitano T, Umetsu K, Tian W, Osawa). M., 2007 https://www.ncbi.nlm.nih.gov/pubmed/16845543). The investigation for the presence of these universal sequences was carried out with the "seal" software of the set bbtools (Bushnell B. sourceforge.net/projects/bbmap/) with the parameters ambiguous = first, rskip = 1, K = 21 and minkmerhits. = 2 using as input the paired readings of each Victoria sample.

The results obtained returned 72 readings for Ancient 004 and 80 readings for Ancient002 with coincidences for the presence of universal sequences for vertebrates. However, during a search with BLAST in the nt database filtered only for vertebrates (taxid: 7742) of these readings, they all match only to humans. This analysis does not detect the presence of other vertebrate species.

Investigation for coding capability in overlapping reads.

Due to the lack of sequences with significant similarity in the DNA databases known for much of the sequenced reads, it was evaluated whether these reads could contain coding sequences and thus be part of new unknown genes in current databases. To evaluate this independently of the sequence similarity comparison, we performed the set of previously obtained overlaps that did not match with any of the 37.877.

references in the iterative filter of each sample and from this set the redundant sequences were removed using the "seal" software as it was done for the *taxmaps* software with the total readings. This set of attached overlaps readings without duplicates was used as input for the FragGeneScan software (Mina Rho, Haixu Tang, and Yuzhen Ye, 2010), which can be estimate using hidden Markov models, taking into account the possibility of reading errors and the use of codons to estimate the presence of gene coding fragments directly from mass-sequence reads even with reads having a considerable error rate.

The results obtained from FragGeneScan were processed by computations developed in R code to obtain the coding potential of the set of reads, defined as the quotient of reads with uninterrupted coding regions, marked "*" by FragGeneScan, among the total of the unique reads.

Depending on the size of the threshold of the coding region, different possible coding potentials have been obtained. By taking a threshold at least 150 bases in the coding region, or 50 amino acids, without it covering all the reading, we have a coding potential of 23.66% for Ancient002 and 56.99% for Ancient004. However, to be more strict in the possibility that reading is part of a gene, it is necessary to take a threshold that covers the total size of the reading and below this threshold and for readings at least 150 bases of overlapped readings together, a lower coding potential was obtained for both samples, but remains considerable.

This analysis shows considerable DNA coding potential for overlapping readings without duplication of the two sets of sequenced sample readings, particularly for Ancient002 overlaps readings with a percentage of about 6% and 25.7% of the readings with coding capacity and for Ancient004 of 18.15% and 58.03%.

Quantitative summary of the readings proportions observed in the main analysis

Finally, the proportions observed for the readings included in the previously described results for each sample were compiled in the following chart to demonstrate the high number of readings analyzed and the high proportion of readings falling into the relevant categories of each analysis, such as unmatched duplicated reads in the NCBI databases for taxmaps as well as for bbtools in the refseq database or the total number of reads resulting from the overlapping union.

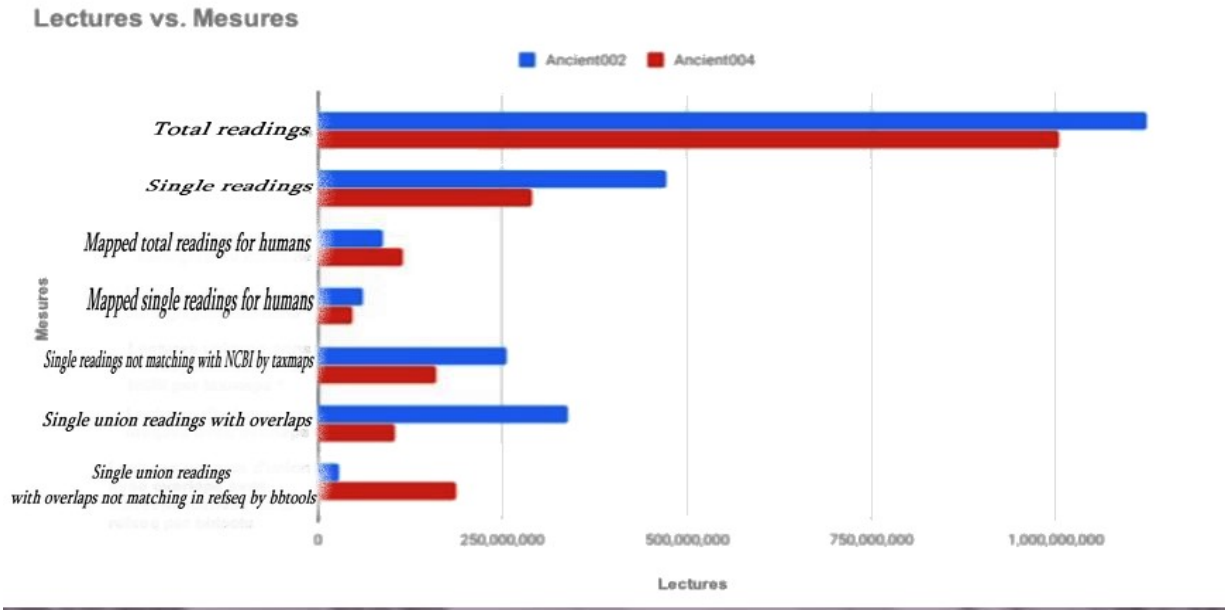

Figure 5: Quantities observed for the readings present in the results of the analysis described in this report for each sample. The mapped single readings for humans were obtained using Kart and its default single readings. Single readings not matching with NCBI by taxmaps were estimated by multiplying the proportion indicated by the taxmaps in the numbers indicated by the total of the unique readings of each sample.

Realized by :

Salvador Ángel Romero Martínez.

Graduate in Genomic Sciences.

April 9th 2019