From anonymous, PhD

To the attention of: Thierry Jamin, Alain Bonnet

Object: Presentation and discussion of the detailed genomic analyses carried out on two unidentified species individuals found in Nazca desert, Peru, in 2015

1. Introduction : phenotypic characteristics of the two individuals.

1.1 Maria:

The individual named Maria for simplicity was found in Nazca, Peru, in 2015.

Maria is a 165-170 cm tall, unidentified gender humanoid creature in foetal like position. Despite obvious similarity with homo sapiens, the individual displays atypical phenotypic traits, among which: skull volume about 25% bigger than homo sapiens (at parietal level), three fingers on each hand and foot, very long phalanges. Two labs independently estimated the radiocarbon age of the individual to the same value of 1750 ± 30 BP.

1.2 Big Hand:

On the same site were found several hands with similar characteristics, three fingers, six phalanges and of considerable size. The age of one of them -- that will be named Big Hand in the current presentation for simplicity, has been estimated as 6420 ± 30 BP

1.3 Questions raised - authenticity and nature of the creatures.

Are these beings genuine biological creatures or only a refined arrangement of other already known species, animal and/or Human? has definitely been the immediate question raised by people having closely approached the material, and certainly by the public.

Before discussing this question, two points shall be kept in mind:
(i) The first point is that in science, when putting forward an hypothesis, the first step consists in reviewing its direct assumptions and implications and check if they are consistent with available evidence. If this prior stage is successful, then a second step consists in testing the predictions of such an hypothesis.
(ii) These bodies are 1700 and 6500 years old, respectively. The tissues are dried, hard and tend to crumble. Consequently, eventual surgical interventions cannot have been done recently, but rather at the death of the subjects, namely more than 1700 years ago for Maria, and more than 6000 years ago for the Big Hand.

So, how likely is the hypothesis of a *"refined arrangement of other already known species, animal or Human"?* This hypothesis implies the existence of traces (lesions, scars) that should be visible, since there is no scarring-over process after death. It also assumes certain technology and knowledge levels that are required to produce such individuals.
The careful anatomical observations, including CT scans, of the individual revealed particularly realistic and refined details (fingerprints, adult teeth, outer and inner surface of the skull including sutures, skin, vertebrae, ribs, joints and articulations, apparently internal organs). Additionally, no lesion on bones or skin tissues suggesting a surgical intervention could be detected. As such, the aforementioned hypothesis sounds unlikely because of
(i) the absence of evidence suggesting surgery or similar manipulation
(ii) the anatomical details that would require, for being emulated, the deployment of biotechnological means a priori not available at that time and even nowadays
(iii) the presence of another individual, an infant, that was found on the same site and displaying the same atypical characteristics as Maria. Details such as milk teeth and body/head proportions confirm this is a genuine baby, and not a small size adult.

A recurrent objection has been that the lesions or scars might be so subtle that we could have missed them out. Effectively, the resolution of the scanners used allowed to see many refined details, but was not the highest available on the market. However, this objection still presupposes a level of technology and science incompatible with the timeframe evoked by the radiocarbon analyses, and I would like to draw your attention to where this hypothesis is leading to. Which ancient surgery devices or tools could have be able to operate in such a subtle way that our modern scanners would miss out the traces left on the body? Can we reasonably **assume the existence of cutting-edge biotechnology lab facilities in Nazca desert between 1700 and 6500 years ago**. This is either just nonsensical, or at least not compatible with available evidence.

Rather, we will consider the predictions of this hypothesis.
Especially, if these creatures have been built with Human and/or animal remains, then DNA analyses should show, *after contaminant DNA removal* (i) either DNA 100% modern homo sapiens or (ii) partially modern homo sapiens and partially animal, probably those locally present in Peru. These are the analyses we carried out and are going to present today.

1.3.2 Plan of the presentation
After this long but necessary introduction, we will proceed as follows.

1) Material and Methods: a brief presentation of the sources of the data we have worked on (ancient DNA, DNA sequencing) and the methods (alignment analysis)

2) The outcome of the first analysis round: identification of contaminant, alignment analysis with modern Homo sapiens. Isolation of unmapped (undetermined) sequences

3) The outcome of the second analysis round, carried out on unmapped sequences: Comparison with other species

4) Conclusions and interpretations

5) Appendix

1.Data sources and Methods

The extraction and sequencing was done by another lab based in Mexico, BioTechMol http://biotecmol.mx/. The genomic analyses that were done did not mention the methods used and were not exhaustive. Hence we carried out the analyses of the raw data files in order to check the quality of the data first, and then to characterize the totality of the DNA sequenced -- contaminants included, on the sequences of the highest quality.

Extraction:

DNA is fragile and ancient DNA samples are generally highly damaged and contaminated by viruses, bacteria, microorganisms and people/animals in direct contact with the sample. For old samples specific procedures have to be followed. A bone sample of 0.54g collected on Maria's body and a bone/unknown tissue of 2.38g collected on the big hand allowed to extract sufficient DNA (procedure by Shapiro B, Heslington M. 2012 + repair kit was used as well, Kit PreCR® Repair Mix de New England Biolabs M0309S).

Sequencing and analysis:

A few words maybe for non technical audience. DNA is made of mainly four nucleotides, or bases. Combined together, these nucleotides form (4*4*4) 64 triplets named "codons" that constitute an alphabet. We are far from mastering the syntax of DNA, but some rules are known, for instance some codons indicate the start of a coding sequence, while others indicate the end.

Sequencing the DNA means *determining the order of the nucleotides that compose DNA*. This is a molecular technique now carried out by machines and computers, and only monitored by a Human. A challenging aspect of sequencing is that DNA molecule is very long (3 billions of nucleotides for Human DNA), so sequencing cannot be done once in a row. Rather, DNA sequencing is carried out on "small pieces" of DNA which gives rise to *reads* (between 50 and 150 nucleotides, or bases). Afterwards, the genome is reconstructed using a procedure called *genome assembly*.

To facilitate the comprehension, one can imagine a text that would be cut off in a random way every 5-10 words. Then the chunks would be copied and finally the text would have to be reconstructed in a coherent and meaningful manner, while spotting eventual errors of copy.

To make reconstruction possible, the "small pieces" are of different lengths, sequenced and overlapped several times The overlaps, and the syntactic rules, allow to reconstruct the genome once sequenced. The overlap are also used to identify possible sequencing errors (for instance, to make it simple, if a region is sequenced 10 times and only 5 read are identical, one can consider there is a high error rate in these region, and maybe eliminate it from the next analysis).
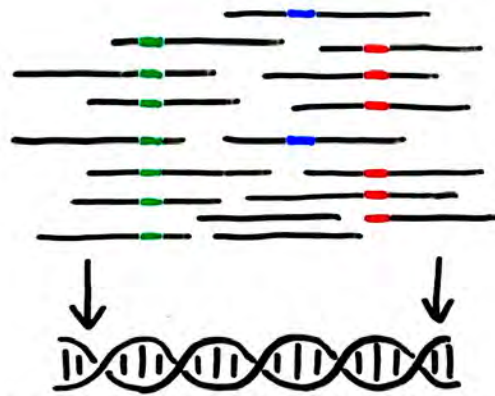
Further sequencing was carried out on Myseq Illumina platform (Illumina platforms are among the most used platforms).

The aim of these analyses was to characterize, quantitatively and qualitatively, the content of the DNA material, including contaminants.

Note that this procedure requires several steps after sequencing, especially genome assembly as mentioned earlier and alignment. Alignment is used for genome comparison when one needs identify a species for example. It consists in mapping the sequenced DNA onto the DNA of a all genomes available for a given species.

A C T C G C A A T A T G C T A G G C C A G C
A C T _ _ _ _ T T A T G C T A T G C _ _ G C

The resulting (Illumina fastq file) data were thus Quality-checked and parsed (see Appendix for further details) in order to determine the genomic content qualitatively and quantitatively in reference to the *RefSeq Complete Genomes database* https://www.ncbi.nlm.nih.gov/refseq/.

2. Identification of contaminants/virus, comparison with modern Homo sapiens, Isolation of unmapped (undetermined) sequences

2.1 Maria

The results are displayed in terms of percentage of <u>absolute mapping</u> to the regions in the genome of Human and Bacterial/Contaminant, the rest being considered as Unmapped/Unclassified.

33.7% of the reads were aligned with **modern homo sapiens**
18.4% of the reads were **contaminants**
47.9% were **unmapped.**

Maria- General summary

Total number of reads (329037x2)	658074
Number of bases - READ1	46674295 4.6Mbp
Number of bases - READ2	46928318 4.6Mbp
Total number of bases	93602613 9.36Mbp

Human Genome	
Reads aligned to Human genome GrCh38	221623
Number of bases aligned (unique alignment)	5782559
Percent of reads mapped to human genome	33.6775

Bacterial and other genomes	
Reads aligned to Bacterial genomes	121186
Number of bases aligned (unique alignment)	17183634
Percent of reads mapped to bacterial and other genom·	18.415

Unknown	
Number of Unaligned/unmapped/unclassfied reads	315265
Number of total unclassified bases	44756142
Percent of total reads unmapped/unclassified	47.907

Maria - Mapping with Homo sapiens

CATEGORY	FIRST_OF_PAIR	SECOND_OF_PAIR	PAIR
TOTAL_READS	329037	329037	658074
FilterPassed_READS	329037	329037	658074
Percent_FilterPassed_READS	1	1	1
FilterPassed_NOISE_READS	0	0	0
FilterPassed_READS_ALIGNED	110328	111295	221623
Percent_FilterPassed_READS_ALIGNED	0.335306	0.338245	0.336775
FilterPassed_ALIGNED_BASES	2873043	2909516	5782559
FilterPassed_HQ_ALIGNED_READS	64305	64227	128532
FilterPassed_HQ_ALIGNED_BASES	1894551	1874692	3769243
FilterPassed_HQ_ALIGNED_Q20_BASES	1835735	1781625	3617360
FilterPassed_HQ_MEDIAN_MISMATCHES	0	0	0
FilterPassed_MISMATCH_RATE	0.002944	0.003927	0.003439
FilterPassed_HQ_ERROR_RATE	0.002706	0.003187	0.002945
FilterPassed_INDEL_RATE	0.00017	0.000299	0.000235
MEAN_READ_LENGTH	140.851205	141.623225	141.237215
READS_ALIGNED_IN_PAIRS	109707	109707	219414
Percent_READS_ALIGNED_IN_PAIRS	0.994371	0.985732	0.990033
BAD_CYCLES	0	0	0
STRAND_BALANCE	0.502737	0.498495	0.500607
Percent_CHIMERAS	0.003847	0.003948	0.003897
Percent_ADAPTER	0.061972	0.000243	0.031107

Maria- Mapping with Bacteria and other contaminant genomes

Total number of reads mapped	121186
Percent of reads aligned	18.415

Bacteria	No. of reads
Alteromonas_macleodii_str._'Ionian_Sea_U8'	21355
Caulobacter_sp._K31	8024
Phenylobacterium_zucineum_HLK1	5099
Delftia_acidovorans_SPH-1	4913
Delftia_sp._Cs1-4	4869
Caulobacter_segnis_ATCC_21756	4504
Caulobacter_crescentus_CB15	3769
Delftia	3762
Bradyrhizobium_sp._BTAi1	2997
Propionibacterium_acnes	2095
Caulobacter	1871
Brevundimonas_subvibrioides_ATCC_15264	1690
Ralstonia_pickettii	1503
Rhodopseudomonas_palustris_CGA009	1279
Ralstonia_pickettii_12J	1188
Alphaproteobacteria	1092
Proteobacteria	1026
Ralstonia_pickettii_12D	1012
Escherichia_coli	997
Enterobacteriaceae	922
Ralstonia_solanacearum	903
Propionibacterium_acnes_ATCC_11828	825
Bacteria_2	790
Caulobacteraceae	643
Others	34674
Virus and others	9384

2.2 Big Hand

As before, the results are displayed in terms of percentage of absolute mapping to the regions in the genome of Human and Bacterial/Contaminant, the rest being considered as Unmapped/Unclassified.

0.37% of the reads were aligned with **modern homo sapiens**
26.7% of the reads were **contaminants**
72.9% were **unmapped.**

Big_Hand- General summary

Total number of reads (341311x2)	682,622
Number of bases – READ1	51,160,199 5.1Mbp
Number of bases – READ2	51,205,429 5.1Mbp
Total number of bases	102,365,628 10.2Mbp

Human Genome	
Reads aligned to Human genome GrCh38	2,518
Number of bases aligned (unique alignment)	366,819
Percent of reads mapped to human genome	0.3689

Bacterial and other genomes	
Reads aligned to Bacterial, viral and other genomes	182,243
Number of bases aligned (unique alignment)	27,331,623
Percent of reads mapped to bacterial and other genome	26.7

Unknown	
Number of Unaligned/unmapped/unclassfied reads	497,836
Number of total unclassified bases	74,655,252
Percent of total reads unmapped/unclassified	72.93

Big_Hand- Mapping with Homo sapiens

CATEGORY	FIRST_OF_PAIR	SECOND_OF_PAIR	PAIR
TOTAL_READS	341311	341311	682622
FilterPassed_READS	341311	341311	682622
Percent_FilterPassed_READS	1	1	1
FilterPassed_NOISE_READS	0	0	0
FilterPassed_READS_ALIGNED	1274	1244	2518
FilterPassed_ALIGNED_BASES	186810	180009	366819
Percent_FilterPassed_READS_ALIGNED	0.003733	0.003645	0.003689
FilterPassed_HQ_ALIGNED_READS	1184	1152	2336
FilterPassed_HQ_ALIGNED_BASES	174577	167615	342192
FilterPassed_HQ_ALIGNED_Q20_BASES	170494	159118	329612
FilterPassed_HQ_MEDIAN_MISMATCHES	0	0	0
FilterPassed_MISMATCH_RATE	0.003014	0.004873	0.003927
FilterPassed_HQ_ERROR_RATE	0.002326	0.004135	0.003212
FilterPassed_INDEL_RATE	0.000252	0.000361	0.000305
MEAN_READ_LENGTH	148.893203	149.025721	148.959462
READS_ALIGNED_IN_PAIRS	1214	1214	2428
Percent_READS_ALIGNED_IN_PAIRS	0.952904	0.975884	0.964257
BAD_CYCLES	0	0	0
STRAND_BALANCE	0.515699	0.47508	0.495631
Percent_CHIMERAS	0.007143	0.007347	0.007243
Percent_ADAPTER	0.102754	0.000319	0.051537

Big_Hand-Mapping with Bacteria and other contaminant genomes

Bacteria	No of reads
Acinetobacter baumannii	6993
Ralstonia sp. MD27	831
Franconibacter helveticus	784
Pseudomonas sp. UBA6753	438
Acinetobacter pittii	413
Acinetobacter sp. 1542444	389
Acinetobacter sp. UNC434CL69Tsu2525	382
Acinetobacter sp. 826659	370
Acinetobacter sp. 742879	331
Delftia sp. 67-8	272
Acinetobacter sp. LMB-5	267
Achromobacter denitrificans	213
Acinetobacter sp. UBA1297	201
Clostridium cochlearium	192
Clostridium novyi	184
Clostridium botulinum	175
Caulobacter mirabilis	173
Bradyrhizobium sp. BTAi1	171
Acinetobacter sp. UBA4567	170
Acinetobacter nosocomialis	162
Caulobacter henricii	160
Acinetobacter lactucae	127
Acinetobacter sp. UBA3098	119
Acinetobacter sp. WC-141	111
Pseudomonas sp. Irchel 3E13	100
Other bacteria	167154
Viruses and Plasmids	1361

3. Unmapped sequences: Comparison with other species

We carried out a second round in order to characterize the sequences that were unmapped. Several species were used for comparison for both Maria and the Big Hand, including: Alpaca, Baboon, Dog, Cat, Horse, Chimpanzee, Rhesus Macaque.

The outcomes were negative for both subjects.

Organism	Total Reads	Aligned Reads	% Alignment
Maria unmapped			
Alpaca	315265	1	0.00032
Baboon	315265	0	0.00000
Dog	315265	366	0.11609
Cat	315265	31	0.00983
Horse	315265	0	0.00000
Chimpanzee	315265	1	0.00032
Rhesus macaque	315265	1	0.00032
Big_Hand unmapped			
Alpaca	497861	11	0.00221
Baboon	497861	0	0.00000
Dog	497861	240	0.04821
Cat	497861	38	0.00763
Horse	497861	2	0.00040
Chimpanzee	497861	25	0.00502
Rhesus macaque	497861	3	0.00060

NB: Note that the number of reads aligned for Dog is not significant for alignment but significantly higher than the other species of reference. We supposed that the huaqueros who found the bodies had dogs.

Alignment analyses are still being run in order to identify the nature of these unmapped/unclassified sequences – almost half of the sample. At the moment, they are held as unidentified.

4. General conclusions

Several questions have been raised during these investigations. The most frequent were the following: (a) the authenticity of the individuals, (b) their eventual links or similarity with homo sapiens, (c) their origins.

(a) the authenticity of the individuals

The following points :

(i) The subtle anatomical details (joints and articulations, skin tissues, inner surface of the skull, cranial sutures, bone density gradient, internal organs, fingerprints..)

(ii) the absence of scars, or any mechanical or surgical lesions detected on the tissues

(iii) the fact that no animal DNA was found (several local species tested)

(iv) the fact that modern Human DNA sequences were present in one individual in minor percentage

(v) the presence of an infant individual found on the same site, with the same atypical characteristics

suggest that although we cannot be formal about the authenticity, no material evidence accredits the "fake hypothesis". Rather, this growing set of evidence suggests more that **we might be in presence of biologically undefined species, that deserve further investigations**.

(b) Their eventual links or similarity with homo sapiens

Here definitions have to be set up. Species are not defined on the basis of their similar appearance or morphology. Two individuals belong to the same species if, and only if they can interbreed. In the evolution theory framework, one can say there is a continuum between species. Two species with a common ancestor are considered as fully differentiated if they cannot interbreed -- which includes mating and having fertile offspring. For example, lions and tigers can mate, but their offspring is sterile. So are horses and donkeys. They both have a common ancestor from which they evolved (the process is called speciation) differently, but they are not totally differentiated yet, as far as they still can mate.

By contrast, Homo sapiens and apes (chimpanzees for instance) who are held to have a common ancestor, cannot mate and have any offspring. They are held as two different species. Nevertheless, known species, especially mammals, have a high percentage of DNA sequences in common – more than 95% depending on the case.

Based on the genome alone, we therefore cannot answer this question in an accurate way.

(c) Their ET origins

Unidentified does not mean Extraterrestrial. Note that Maria seems to be fully equipped to survive and move in Earth biosphere. Therefore, from a biological point of view, there is a priori nothing suggesting that she would come from another planet.

Also, there is no database for ET genomes or exobiology in general. We thus cannot compare her genome with anything held as ET, and consequently are unable to tell about her possible ET origins. Today, ET origins is NOT a genomic information (yet).

Unmapped sequences are available upon request -- contact details and proof of qualifications and computational resources (cloud is okay) can be given to Alain Bonnet.

Materials and methods

Data

Paired-end Illumina files were available for the sample in fastq format. The forward and reverse strand files contained 329037 reads each for Maria, and 341311 reads each for the Big Hand. The fastq files were converted to fasta format using custom scripts to count the number of bases.

Raw Data QC

The paired end files were subjected to quality control for per sequence quality, per base quality, k-mer content, adapter and other contamination. NGS quality control tool Fastqc was used for this purpose. Illumina adapters and standard contaminants list was used to filter any known noise. The reads were of varying lengths ranging from 35 to 151 bases. More than 90% of the reads passed the phred score of Q30 and a minimal number below Q20. Overall the raw data was of good sequence quality and did not show any signs of reagent depletion or sequencing artifacts.

Alignment to human genome (GRCh38)

After passing quality control the reads were aligned with the latest and more comprehensive assembly of human genome, GRCh38. Alignment was performed using bwa-mem algorithm with strict options. Bwa-mem is the alignment algorithm of choice for reads more than 100bp in length due to its speed and accuracy especially with mammalian genomes. Approximately 33% and less than 1% of the reads aligned to human genome GRCh38 and the alignment was stored in binary alignment map (bam) format.

Bamtools was used to filter out mapped and unmapped reads. The unmapped reads extracted from this alignment file were converted to fastq format for further analysis.

Classification of unmapped reads by exact alignments of k-mers

The unmapped reads from the previous alignment step were subjected to classification analysis using Kraken. Kraken assigns taxonomic labels to sequencing reads based on the exact alignment of k-mers against groups of genomes (bacterial, plasmids, viruses, etc.). A reference database was first constructed from complete bacterial, archaeal, and viral genomes in RefSeq. This database was considerably a large one measuring about 8 gigabytes in size. To eliminate the primary source of false positive hits like low-complexity sequences in the genomes themselves; e.g., a string of 31 or more consecutive A's, we ran the 'dust' program on all genomes and then building the database from these 'dusted' genomes. The kraken analysis ran for 6 hours on a t2.xlarge EC2 instance on Amazon AWS cloud with 4 high power CPUs and 16 GiB of RAM. At the end, about 27.7% of unmapped reads were classified. This contributed to 18.5% of the total reads in the sample. Various bacterial and viral taxonomic classifications were assigned and the unclassified reads were obtained as fastq file.

Overall about 33.7% of the reads were aligned to the human genome, 18.4% of the reads mapped to bacterial genomes and the rest 47.9% were unclassified.
The same procedure was followed for the Big Hand, with different percentages;