ABRAXAS
Biosystems

CONSULTING SERVICES REPORT:

Genomic and Bioinformatic analysis of high throughput DNA sequencing of samples extracted from desiccated bodies found at Nazca

Client:
GAIA INTERNATIONAL, INC.

JOSÉ JAIME MAUSSAN FLOTA

November/9th/2018

ABRAXAS
Bringing knowledge to life

Table of contents

OVERVIEW

The present document provides the details and relevant briefings of all the jobs, tasks and procedures involved in the service provided by **ABRAXAS BIOSYSTEMS S.A.P.I. DE C.V.** for **GAIA INTERNATIONAL, INC.** and **JOSE´ JAIME MAUSSAN FLOTA** for the "Genomic and Bioinformatic analysis of high throughput DNA sequencing of samples extracted from desiccated bodies found at Nazca". We present an ordered description of the main tasks and analysis developed for this project.

The tissue samples extracted from the dessicated bodies found at Nazca used for the analysis presented in this service were provided, managed and handled by JOSE´ JAIME MAUSSAN FLOTA and his scientific colleagues during all the stages previous to the DNA extraction described in this report while **CEN4GEN labs (6756 – 75 Street NW Edmonton, AB Canada T6E 6T9)** were in charge of performing all the tasks on the samples from the DNA extraction to the High throughput DNA sequencing, also known as Next Generation Sequencing, and clean sequencing data generation stages.

ABRAXAS BIOSYSTEMS S.A.P.I. DE C.V. performed all the Computational Genomics and Bioinformatic analysis.

For this project JOSE´ JAIME MAUSSAN FLOTA and his scientific colleagues provided for the delivery to CEN4GEN's labs 7 samples, 3 tissue samples and 4 DNA samples from the bodies found at Nazca, Peru. After the DNA extraction, quality control and MDA amplification procedures at CEN4GEN labs only 3 samples, from the original 7, passed the controls for NGS, the names of these samples, as coming from the original tubes sent for delivery, were as follows:

Sample name	Sample original label	Identity
Ancient-0002	Neck Bone Med Seated 00-12 Victoria 4	Victoria
Ancient-0003	1 Hand 001	Hand
Ancient-0004	Momia 5 -DNA	Victoria

Table 1: Sample name indicates the name that CEN4GEN assigned to the sample, Sample original label is the name in the tube were the sample was originally contained when it was delivered to CEN4GEN, Identity is the name of the body from which the sample is coming.

Hence all the analysis tasks referred in this report after the DNA extraction, quality control and MDA amplification tasks were only performed for these 3 samples.

SERVICES INDEX

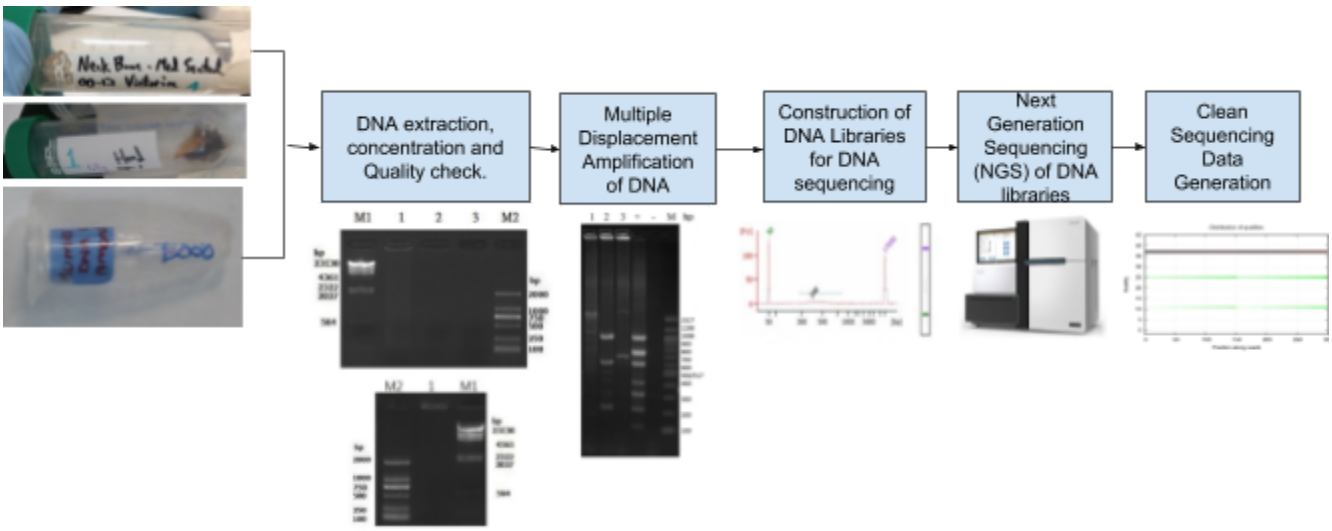
The comprehensive solution proposed by Abraxas Biosystems includes a wide range of services from ancient DNA extraction to sequencing and data analysis (bioinformatics), allowing the generation of accurate results from ancient samples analysis. The analysis tasks performed in this project were the following:

1. DNA extraction.
2. DNA Quality Check.
3. DNA Amplification by Multiple displacement Amplification.
4. DNA libraries construction.
5. Next Generation DNA sequencing (NGS).
6. Clean sequencing data generation.
7. QC of sequencing results.
8. Preliminar analysis by mapping of DNA reads to human genome reference.
9. Reads overlap analysis to detect short fragments common to ancient DNA.
10. Mapping of overlapped DNA reads from Ancient0003 to the most recent version of the human genome.
11. Mitochondrial analysis for detection of variants in D-loops and other informative regions to determine mitochondrial haplotypes.
12. Sex determination of the Ancient0003 sample.
13. Detection of possible organisms present in the sample by genomic dna sketching method (exact matches of groups of short fragments, k-mers, to public databases) and iterative filtering of reads by exact k-mer matches.
14. De novo assembly with strategies for mixed DNA of reads without matches to detected organisms in the sketch method.
15. Mapping of reads without any exact matches in the iterative filtering process to the resulting sequences in the de novo assembly.

16. DNA databases search of de novo assembled DNA segments to detect similarity to known organisms.
17. Taxonomic Classification of unmatched sequencing in previous steps reads by match searches against comprehensive DNA databases.

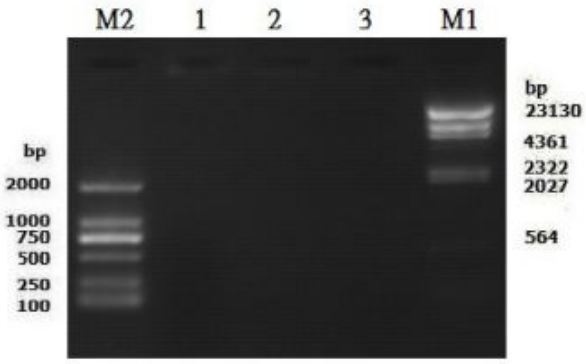
LABORATORY SERVICES WORKFLOW

For the laboratory analysis processes , tasks 1-6, the general workflow was as follows.

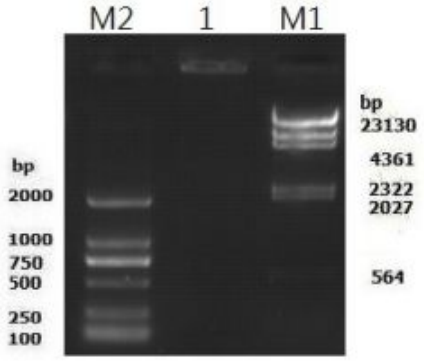


DNA EXTRACTION

The 3 tissue samples were run through a proprietary DNA extraction protocol customized for ancient samples and developed at CEN4GEN labs that was based on the protocols described at the following article (Gamba et al., 2016). After the DNA extraction process the extracted DNA was run through agarose gels to check the presence of bands indicating the presence of adequate amounts of DNA (indicated by visible illuminated horizontal bands on each lane corresponding to each sample). Also the already extracted DNA from Ancient-004 was checked by this method since it contained the DNA from a tissue sample that was not available anymore from Victoria's body. The results are shown in the next figure:



Lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Sample Integrity
M1	λ-Hind III digest(Takara)	1	3	
1	CEN4GEN-Ancient0001	1	3	Degraded completely
2	CEN4GEN-Ancient0002	1	3	Degraded completely
3	CEN4GEN-Ancient0003	1	3	Degraded completely
M2	D2000 (Tiangen)	1	6	



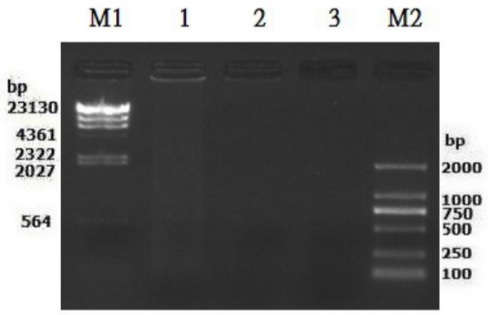
Lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Sample Integrity
M1	λ-Hind III digest(Takara)	1	3	
1	CEN4GEN-Ancient0004	1	3	
M2	D2000 (Tiangen)	1	6	

Figure 1: DNA extraction results (top) and DNA quality check of already extracted DNA(bottom) . M2 and M1 lanes on each gel are the molecular markers used to measure size of DNA fragments.

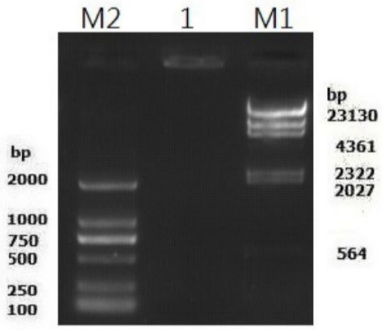
The Figure 1 shows that the bands at the extracted DNA samples were too low in visibility and hence had not enough DNA for proper NGS results. This took the labs to evaluate the 4 DNA samples to see if those had enough DNA.

DNA QUALITY ANALYSIS

After DNA extraction from the 3 tissue samples the 4 DNA samples were run through a quality check process to evaluate the presence of good amounts and sizes of DNA to find out if they could rescue the DNA needed for NGS. The quality check was also done in agarose gels as in the previous step. The results are shown below:



Lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Sample Integrity
M1	λ-Hind III digest(Takara)	1	3	
1	CEN4GEN-Momia1	1	3	Degraded completely
2	CEN4GEN-Momia3	1	3	N/A
3	CEN4GEN-Momia4	1	3	N/A
M2	D2000 (Tiangen)	1	6	



Lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Sample Integrity
M1	λ-Hind III digest(Takara)	1	3	
1	CEN4GEN-Ancient0004	1	3	Degraded completely
M2	D2000 (Tiangen)	1	6	

Figure 2: DNA quality check of already extract DNA samples not analyzed in the previous quality check (top) and DNA quality check of already extracted DNA previously analyzed along the tissue samples DNA extraction (bottom).

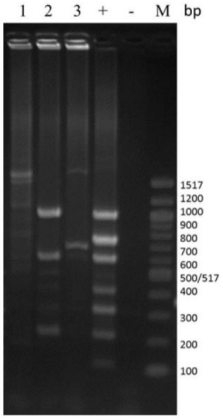
The results were the same as for the DNA extracted from the tissues by CEN4GEN labs showing very low presence of DNAe. This absence of high DNA amounts and the advantage of NGS to rescue data from low inputs of DNA with some amplification efforts led the CEN4GEN labs team to run a process called **Multiple Displacement Amplification** that had shown good results at their facilities with ancient samples to amplify the levels of available DNA needed for NGS sequencing.

DNA AMPLIFICATION

By Multiple Displacement Amplification

After finding very low amounts, resulting from the degraded state of the tissue samples, the labs turned to the MDA process to amplify the amounts of the DNA fragments. This process was customized for the characteristics of the extracted DNA using the proprietary methods of CEN4GEN labs. The results of the MDA were acceptable to preed with NGS for 2 of the DNA samples extracted by CEN4GEN (Ancient0002 and Ancient0003) and for 1 of the samples already delivered extracted (Ancient0004) as shown below:

Electropnoretogram:



lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Number of housekeeping genes detected
1	CEN4GEN-Ancient0002	1	10	0
2	CEN4GEN-Ancient0003	1	10	3
3	CEN4GEN-Ancient0004	1	10	0
+	Positive control	1	10	7
-	Negative control	1	10	0
M	100bp DNA ladder (NEB)		6	/

Figure 3: MDA amplification results, the first 3 lanes correspond to the successfully amplified samples and the last fourth and fifth lanes are for the negative control and the molecular markers respectively.

The rest of the samples did not show amplification results so the rest of the analysis tasks were performed only for these 3 samples.

DNA LIBRARIES CONSTRUCTION

To proceed with the NGS run a library of the fragments to sequence has to be prepared first. This library preparation method was performed by CEN4GEN using a specialized protocol proprietary of CEN4GEN labs and reagents kits based on a commercial kit called Kapa Hyper Prep that were optimal to recover fragmented DNA for ancient samples. The results of the library construction methods were successfully performed for the 3 MDA amplified samples as shown below.

3a. QC Results Summary

Sample No.	KAPA NGS library construction	KAPA NGS library concentration (ng/µl)	Fragment size (bp)	Test Result
CEN4GEN-Ancient0002	Passed	113.0	398	Qualified
CEN4GEN-Ancient0003	Passed	112.3	515	Qualified
CEN4GEN-Ancient0004	Passed	208.1	423	Qualified

After this step everything was ready for the NGS run to sequence the amplified DNA of the 3 samples.

HIGH-THROUGHPUT DNA SEQUENCING

The constructed libraries were subjected to Paired end high throughput DNA sequencing (NGS) of read lengths of size 150 using the Hiseq X10 sequencing equipment from the Illumina company which is one the most powerful machines currently available for DNA sequencing. The sequencing run was performed with a 50X coverage, this means that if the samples came from a genome of size of around 3 billion bases then every base in the DNA of that genome would have been read by the machine on average 50 times. This is almost the double of

coverage that current human genomes are read for clinical applications and 5-50 times the coverage used for many scientific applications in human research.

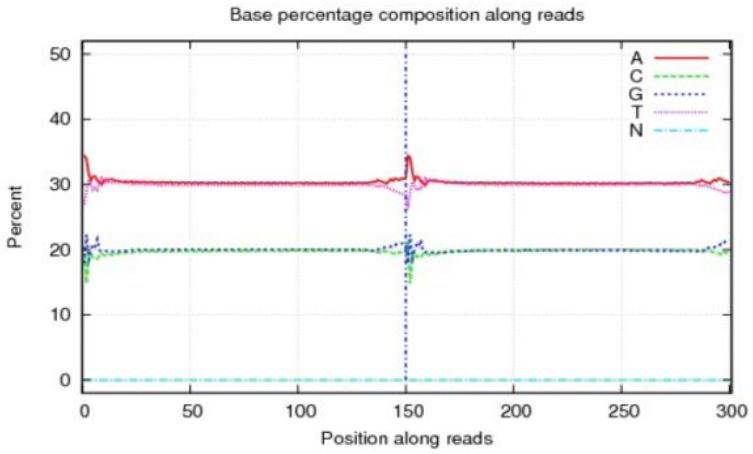
After completing the DNA sequencing run the results were cleaned for bad quality or artifact reads for all three samples after which the generated amount of DNA reads for each sample resulted in the following number of sequencing reads and bases:

Sample No.	Clean Reads	Clean Bases	Read length (bp)	Q20(%)	GC(%)
CEN4GEN-Ancient0002	1123330640	168499596000	150	98.12%	39.73%
CEN4GEN-Ancient0003	1295578732	194336809800	150	97.63%	41.35%
CEN4GEN-Ancient0004	1003400490	150510073500	150	98.12%	46.39%

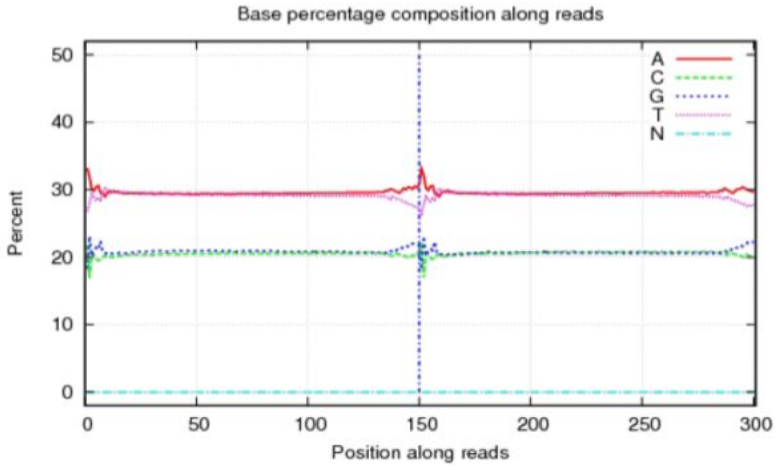
Samples Ancient0003 and Ancient0004 obtained the above reads in a single run result but the sample Ancient0002 needed of two sequencing runs so the numbers shown in this report correspond to the sum of both sequencing runs for sample Ancient0002.

Also a nucleotide frequency plot was run to make sure that there was not biases in certain bases or positions in the reads. These frequency plots are shown below as presented in the CEN4GEN's sequencing results report.

a) CEN4GEN-Ancient0002
- Base percentage distribution along reads after filtering sample



b) CEN4GEN-Ancient0003
- Base percentage distribution along reads after filtering sample



c) CEN4GEN-Ancient0004
- Base percentage distribution along reads after filtering sample

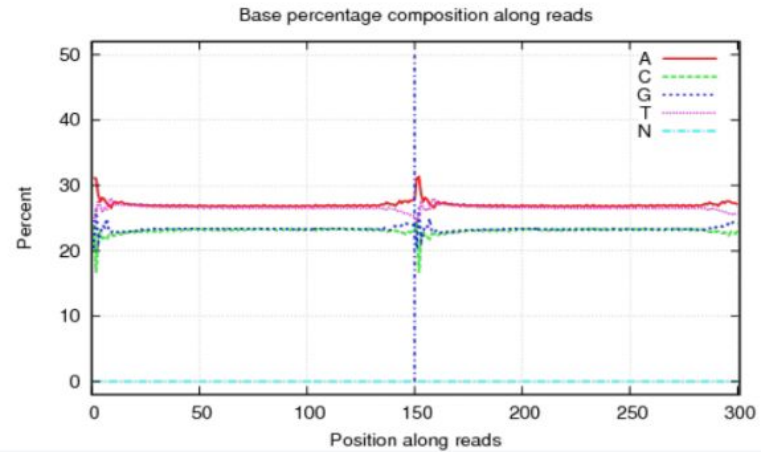


Figure 4: Nucleotide frequency plots for each sequenced sample.

The reads generated from this clean data were the raw data input used for the downstream genomic and bioinformatic analyses.

SEQUENCING QUALITY ANALYSIS

The sequenced data was checked for common quality metrics in sequencing assays like mean read lengths, GC content, overrepresentation of sequences and k-mer abundance using the **Fastqc** software :

(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

All the evaluated metrics were very good for all samples as shown in the complete reports stored at (Access request required):

https://drive.google.com/open?id=1-MiOTW4CC34wp6vV9FDjqBPOt9cNS_23

PRELIMINARY MAPPING ANALYSIS

To get a quick approximation of the relatedness of the samples to human DNA the QC checked reads were subsampled to a 25% portion (a fraction of 25 % of the reads were extracted randomly from the total of each of the sequencing runs corresponding to each sample, except for sample 2 for which only one of the two runs were sampled) mapped against the unmasked human genome reference in the most possibly updated version to the dates of the analysis that we could get, this was version GRCH38 release 93 downloaded from:

ftp.ensembl.org/pub/release-93/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz

This genome sequence allowed us to have a DNA sequence of reference to compare the reads for possible presence of human DNA in the samples by comparing each read against this reference. The comparison was done using the **software bwa mem mapper on its Version 0.7.17-r1188**. The mapping process showed the following results:

Sample	Sampled reads	Mapped reads	% of total	%of reads in pairs	MismatchRate
Ancient0002	217,932,960	31,147,853	14.2924%	87.2839%	0.006643
Ancient0003	156,666,974	153,046,994	97.6894%	99.6032%	0.006312
Ancient0004	250,850,122	38,276,901	15.2589%	88.292%	0.012726

The previous table shows that Ancient0002 and Ancient0004 have very few amounts of DNA that could be of human origin as compared to Ancient0003 sample that shows a high signal of relatedness to human DNA.

As a control for this analysis a set of DNA sequencing reads known to come purely from human origin, generated at Abraxas by a simulation of sequencing reads to 50 X coverage over the human genome version GRCH38 created with the **software ART version MountRainier** (Huang, Li, Myers, & Marth, 2012)**,** was applied to this same mapping method showing metrics very close to Ancient0003 thus confirming that the method functions correctly for detecting close relatedness to human genome in sequencing data.

READS OVERLAP ANALYSIS

Detecting short fragments common to ancient DNA

A common feature of ancient DNA samples is that they show very short fragment sizes in the sequencing data, this feature causes that the pairs of sequencing reads overlap to each other. This can be detected by searching for high similitude fragments in the opposing borders of each read pair in the paired-end sequencing data. We generated this overlap detection using for samples Ancient0002 and Ancient0004 the software **PEAR v0.9.6** (Zhang, Kobert, Flouri, & Stamatakis, 2014) which finds statistically significant overlaps in sequencing read pairs and joins them together in a single read while for Ancient0004 we used the software AdapterRemoval as implemented in the **Paleomix Pipeline** (Schubert et al., 2014)**.** The amounts of reads joined in this way, and thus showing presence of ancient DNA, for each sample were:

Sample	ANCIENT0002	ANCIENT0003	ANCIENT0004
Raw reads	1,123,330,640	1,295,578,732	1,003,400,490
Overlapped pairs	207,201,104	1,290,194,828	676,029,784
Overlapped reads	103,600,552	645,097,414	338,014,892

This analysis showed that we had a dataset with signals of coming from a degraded source which is coincident with ancient DNA samples. The overlapped reads generated in this task were used in the following analysis

READS MAPPING ANALYSIS

From the analysis of preliminary mapping we knew that the Ancient0003 sample was very probably coming from humans so we took all the 645,097,414 overlapped reads and mapped them to the human genome. This analysis showed a mapping of 613,314,913 / 95.07% of the total reads which is in agreement with the 97.69% obtained in the preliminary mapping analysis. This confirmed the relatedness to the human species of the sample Ancient0003. In addition we evaluated the quality of the mapping across all the chromosomes used for the mapping analysis to make sure that we had high quality matches in the mapping results that showed the 95.07% of match in the total amount of overlapped reads. The quality plot of the mapping is shown in the below:

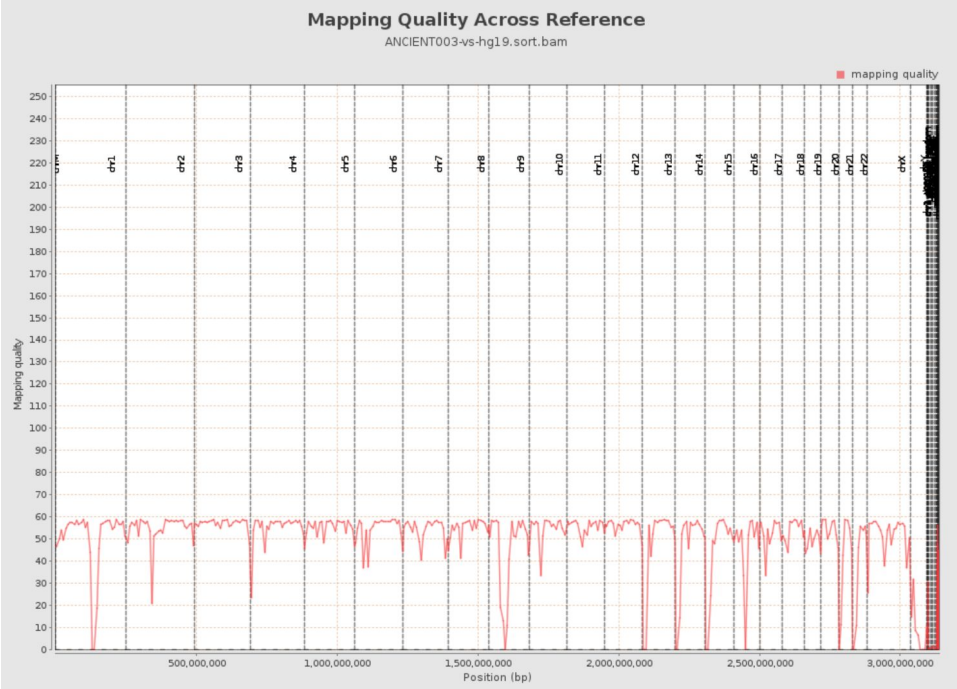


Figure 5: Mapping quality plot on the Y axis we have the quality level (on a 0-60 scale, 60 being the best) and in the X axis we have the positions along each chromosome.

For this analysis we used the Human Genome version hg19 for comparative purposes , which is one version earlier that the GRCH38 used for the preliminary analysis and also to be able to use this result in a more compatible way for further analysis.

MITOCHONDRIAL HAPLOTYPES ANALYSIS

The mapped results to the human genome hg19 obtained in the previous process were used to gather data about the mitochondrial regions recovered and with this to test the match to known regions associated to certain human populations known as haplogroup. To get this we applied the **web tool mtDNA-server** (Weissensteiner et al., 2016) using the portion of the mitochondrial mapping results of the bwa mem mappings extracted with the "samtools view" utility. The results showed the following information according to the mtDNA-server.

SampleID	Haplogroup	Quality	mean.COV.FOR	mean.COV.REV	min.COV	mtDNA.COVERED	HET.Level	HET.Count
ANCIENT003-vs-hg19.subChrM.sort	M20	91.9	143.6725	151.9117	1	16494	0.2260968	40

Also the regions with annotated loci were detected and plotted for their heteroplasmy in each detected and annotated loci which showed high heteroplasmy in the D-loop regions of the mtDNA:

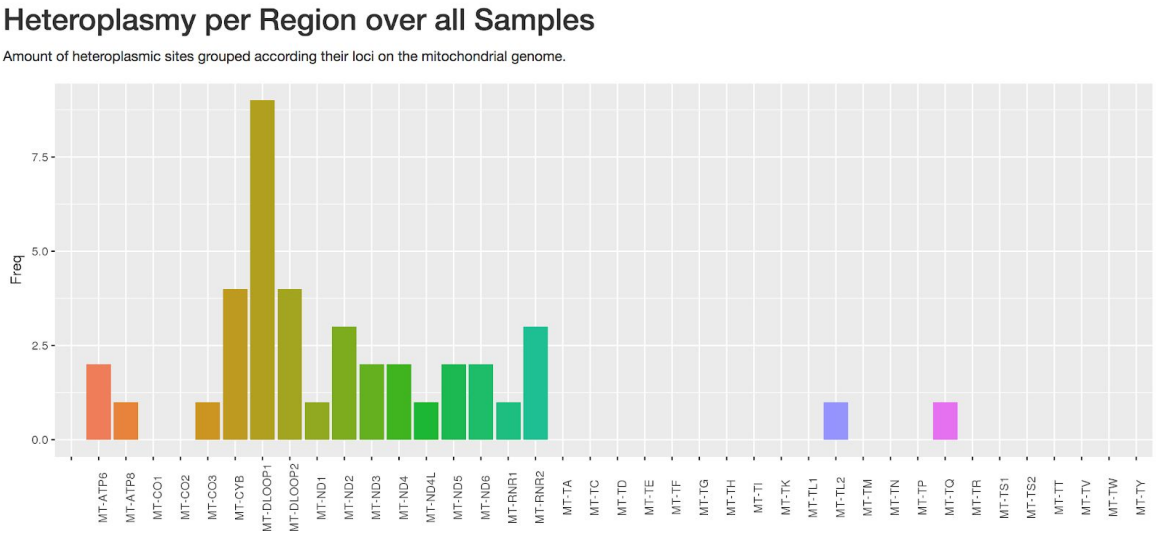


Figure 6: Plot of the heteroplasmic levels detected in each annotated region of the mt DNA as provided by the mt-DNA server web tool.

SEX DETERMINATION

Ancient 0003 Sample

By means of coverage analysis as implemented in the tool indexcov contained in the **software goleft v0.1.18** using default parameters and with the mapping results of the

Ancient0003 sample against the hg19 human genome as inputs. The results give evidence of a male origin of the human DNA in Ancient0003 sample as shown in the next plot that the software generated:

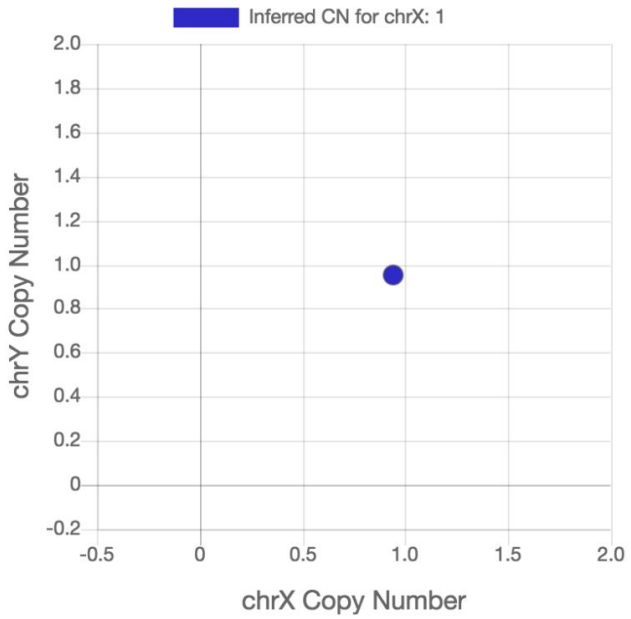


Figure 7: Plot of the copy number ratios of the X and Y chromosomes detected by the indexcov tool in Ancient0003 mapping results against the hg19 human genome.

DNA SKETCHING AND ITERATIVE FILTERING ANALYSIS

For other organisms detection in Ancient 0002 and Ancient0004 samples

To get an approximate idea of the possible types of organisms present in Ancient0004 and Ancient0002 samples we performed a genomic dna analysis called genomic **DNA sketching** (Ondov et al., 2016) which uses exact matches of groups of short fragments, k-mers, to public databases and codification of these matches in data structures called hashes that are fast to compare to public databases of published genomes converted to these structures and then we performed and iterative filtering of reads by exact k-mer matches to the genomes retrieved by the sketching method.

For this analysis we used the **bbtools software (https://jgi.doe.gov/data-and-tools/bbtools/)** on its version 38-25 to perform the sketching and the iterative filtering.

This filtering also allowed us to test the hypothesis that the sequenced DNA came from microbial origins (bacteria, archaea,virus, fungus or plasmids).

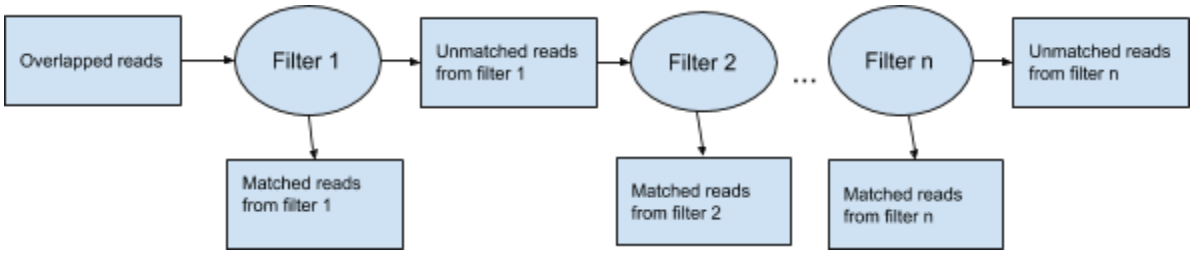
Briefly the strategy a database of all the bacterial, archeal, viral, fungal, plasmidic, model eukaryotic (including human) and some non-model eukaryotic genomes as well as the top genomes indicated by the sketching method available in public databases to the date of the analyses. These genomes were used to build a local database to compare the overlapped reads of each sample in an iterative manner that separated the reads having at least one exact sequence match of size 31 along their length to each type of genomes compared. The sets of genomes used for the iterative filtering were as follows:

1. Bacteria
2. Virus
3. Plasmids
4. Phages
5. Fungi
6. Plastid
7. Diatoms
8. Human
9. Bos Taurus
10. H penzbergensis
11. Phaseolus Vulgaris
12. *Mix2: Label for the following genomes:
    - Lotus japonicus chloroplast, complete genome
    - Canis lupus familiaris cOR9S3P olfactory receptor family 9 subfamily S pseudogene (cOR9S3P) on chromosome 25
    - Vigna radiata mitochondrion, complete genome
    - Millettia pinnata chloroplast, complete genome
    - Curvibacter lanceolatus ATCC 14669 F624DRAFT_scaffold00015.15, whole genome shotgun sequence
    - Asinibacterium sp. OR53 scaffold1, whole genome shotgun sequence

- Bacillus firmus strain LK28 32, whole genome shotgun sequence
- Bupleurum falcatum chloroplast, complete genome
- Alicycliphilus sp. B1, whole genome shotgun sequence
- Bacillus litoralis strain C44 Scaffold1, whole genome shotgun sequence
- Chryseobacterium takakiae strain DSM 26898, whole genome shotgun sequence
- Paenibacillus sp. FSL R5-0490
- Bacillus halosaccharovorans strain DSM 25387 Scaffold3, whole genome shotgun sequence
- Rhodospirillales bacterium URHD0017, whole genome shotgun sequence
- Bacillus onubensis strain 10J4 10J4_trimmed_contig_26, whole genome shotgun sequence
- Radyrhizobium sp. MOS004 mos004_12, whole genome shotgun sequence
- Bacillus sp. UMB0899 ERR1203650.17957_1_62.8, whole genome shotgun sequence

13. Vertebrates: Label for the following genomes:
    - Amblyraja-radiata_sAmbRad1_p1.fasta
    - bStrHab1_v1.p_Kakapo.fasta
    - bTaeGut1_v1.p_ZebraFinch.fasta
    - GCA_000978405.1_CapAeg_1.0_genomic_CapraAegagrus.fna
    - GCA_002863925.1_EquCab3.0_genomic_Horse.fna
    - GCF_000002275.2_Ornithorhynchus_anatinus_5.0.1_genomic.fna
    - GCF_000002285.3_CanFam3.1_genomic.fna
    - Macaco_GCF_000772875.2_Mmul_8.0.1_genomic.fna
    - rGopEvg1_p1_Gopherus_evgoodei_tortuga.fasta
14. Protozoa

The method can be depicted as follows:



The set of reads left after passing all the filters for each sample contained 27,974,521 reads for Ancient0002 and 304,785,398 for Ancient0004. This indicated that by this strategy around **27% of the DNA from overlapped read in Ancient0002 sample and 90% of the DNA from overlapped reads in Ancient0004 sample could not be identified to any of the organisms in the compiled database for this strategy even though the databases included a wide variety of genomes and a comprehensive amount of organisms.**

To make the following analysis faster and more efficient the previously mentioned last set of reads resulting from this filtering strategy was depleted of redundant sequencing reads by using the **software "dedupe" which is part of the "bbtools" software**.

The resulting number of reads after the duplicate removal process where 16,412,862 for Ancient0002 and 30,823,217 for Ancient 0004.

DE NOVO ASSEMBLY FOR MIXED DNA

Applied for readas without matches to detect organisms in the sketch-iterative filtering method.

To reduce even further the complexity of the unmatched DNA found in the Ancient0002 and Ancient0003 samples the unmatched reads in the previous task were conciliated to generate consensus sequences by de novo assembly using a metagenomic assembler which has capabilities of working with samples that contain a mix of different organisms which could be the case of the DNA samples in Ancient0002. The unmatched unique reads were de novo assembled for each sample using the software **megahit v1.1.3** (Li et al., 2016). The results for the assembly were as follows:

Ancient0002: 60852 contigs, total 50459431 bp, min 300 bp, max 24990 bp, avg 829 bp, N50 868 bp, 884,385 (5.39%) assembled reads

Ancient0003: 54273 contigs, total 52727201 bp, min 300 bp, max 35094 bp, avg 972 bp, N50 1200 bp, 20,247,568 (65.69%) assembled reads

SIMILARITY ANALYSIS TO KNOWN ORGANISMS

The resulting contigs were blasted against the nt database using the blastn v 2.7.1, (using an E-value of 10, word size of 20 and a percent identity of 30) to search for possible matches against known organisms in the nt database and the number of hits was counted to determine if the assembled fragments had better results in getting a match to known organisms. In total for sample Ancient0002 just 1,256 (out of 60,852) contigs didn't get any match and for Ancient0004 just 1,768 (out of 54,273) contigs didn't get any match.

Also a further de novo assembly that used as input the two sets of unmatched unique reads from both Ancient0002 and Ancient0004 together but the assembly results were much less assembled and had much more fragmented results so this additional assembly was discarded.

ULTRA COMPREHENSIVE TAXONOMIC CLASSIFICATION

For unmatched unique and raw subsample reads

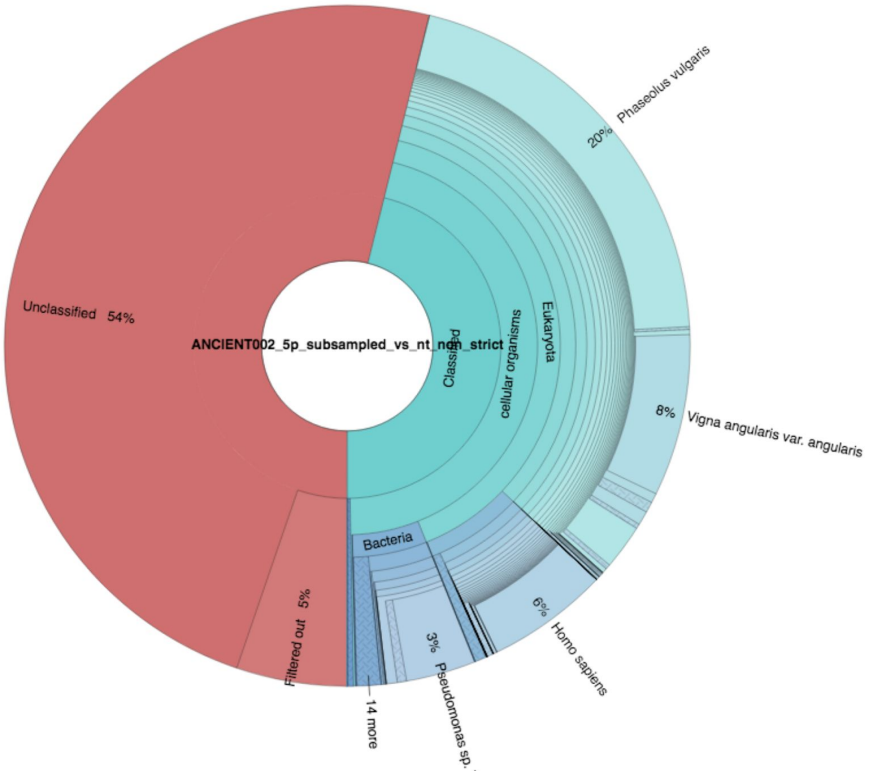
The previous tasks were performed to find the amount of classifiable DNA reads from the Ancient0002 and Ancient0004 samples to get an understanding of the extent to which the samples with low mapping mathes to the human genome resemble known organisms at the DNA level. Even though our genomic sketch and iterative filtering approach was comprehensive at many taxonomic ranges and types of organisms we needed to provide an even more detailed classification of the DNA sequencing reads at every possible taxonomic rank and with an even larger database along with a more flexible matching algorithm to broaden, as much as it could be reasonably possible, the spectra and detection power of our methods of DNA matching to known and sequenced organisms in earth. To accomplish this increased spectra of detection and detailed classification at every possible taxonomic rank we implement an ultra comprehensive and highly sensitive strategy based on the creation of a new database with even more entities , actually with one of the most comprehensive sequence datasets known in bioinformatics as the **NCBI nt database**, constructed using

compression and redundancy removal algorithms and based also on the implementation of an optimal non-exact match mapping search comparable in sensitivity to the **highly sensitive BLAST search** but in practical times, this search using a BLAST search would have taken months thus broadening the search power of the sketch and iterative filtering since that strategy was constrained to exact matches only.

This strategy was implemented using the **taxmaps v 0.2.1 software** (Corvelo, Clarke, Robine, & Zody, 2018) and applied to a subset of 5 % of all the raw unfiltered reads for the Ancient0002 and Ancient0003 samples.

The same analysis was repeated for a 25% subset of the Ancient0004 sample too just to confirm that our methods were predicting correctly the proportions of classified and unclassified reads as the samples reads approached to the full set of reads (which would have required dozens of more days to complete).

This strategy also allowed us to compare if the overlapping and filtering processes behaved differently to the unfiltered reads originally sequenced in their matches to known organisms' DNA. The results are as shown below:



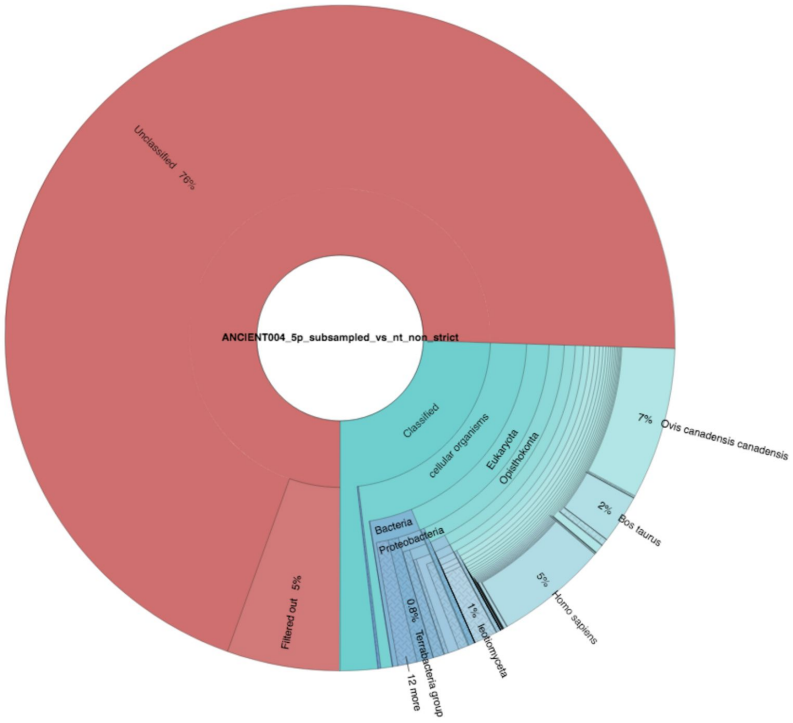


Figure 8: Proportions of classified and unclassified reads from a 5% subsample (28,073,655 reads for Ancient0002 and 25,084,962 for Ancient0004) of the complete raw sequencing reads for Ancient 0002 (Top) and Ancient0004 (Bottom) against the NCBI's nt Database as implemented in the taxmaps 0.2.1 which includes 34,904,805 DNA sequences that represent 1,109,518 of Taxa.

This approach confirmed that there are very high levels of unmatched and unclassified DNA content in the sequenced samples when compared against one of the most comprehensive datasets compiled publicly for genomic information under the parameters considered (an allowed edit distance of maximum 0.2 between the kmers searched by taxmaps against the non redundant database implemented for the nt dataset).

CONCLUSIONS

Abraxas Biosystems performed a wide range of bioinformatic and genomic analysis in order to identify the possible biological origin and the ancestry of the samples provided by Jaime Maussan and his scientific colleagues and extracted/Sequenced at CEN4GEN labs. After the design of a meticulously customized protocol for maximizing the success rate of ancient DNA extraction, sequencing (with CEN4GEN Labs) and bioinformatic analysis of the samples, the

results show a very low mapping match with human genome data for samples Ancient0002 and Ancient0004 contrary to the Ancient0003 sample that did show very high mapping matches to the human genome. Also it is notable that Ancient0002 and Ancient0004 samples show very low rates of matches to one of the most trusted and accurate databases (nt from NCBI). However, NCBI databases does not contain all the known organisms existing in the world so there could be a lot of possible organisms that account for the unmatched DNA or could be some regions excluded, or difficult to sequence, common to many of the organisms accounting for the samples in the applied protocols for the genomes reported at NCBI.

Laboratory and computational protocols for ancient DNA analysis, given the nature of the samples, include several steps that could bring noise to the data and directly impact in the results. One of the most common examples is tissue manipulation by multiple individuals and left to the open environment previous to its isolation, complicating the possibilities that all the sequenced DNA comes from the endogenous DNA of the individual bodies sampled. One way to avoid this kind of noise and obtain better results is to sequence internal bone samples and not exposed tissues.

Finally, current databases at NCBI are constantly growing so it could be that a better and even more comprehensive databases can soon be constructed that includes more available microbial and/or eukaryotic genomes that can shed light on the nature of the unmatched DNA samples. Even more a focused analysis on just the unmatched DNA segments could be developed to double confirm that these are not artifacts of the sequencing or amplification protocols. Ancient DNA protocols are in continuous improvement given its sensible and degradative characteristics of this kind of samples. We recommend additional studies to accept or discard any other conclusions.

REFERENCES

Corvelo, A., Clarke, W. E., Robine, N., & Zody, M. C. (2018). taxMaps: comprehensive and highly accurate taxonomic classification of short-read data in reasonable time. *Genome Research*, *28*(5), 751–758.

Gamba, C., Hanghøj, K., Gaunitz, C., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., … Orlando, L. (2016). Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Molecular Ecology Resources*, *16*(2), 459–469.

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics* , *28*(4), 593–594.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., … Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* , *102*, 3–11.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132.

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., … Orlando, L. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols*, *9*(5), 1056–1082.

Weissensteiner, H., Forer, L., Fuchsberger, C., Schöpf, B., Kloss-Brandstätter, A., Specht, G., … Schönherr, S. (2016). mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Research*, *44*(W1), W64–W69.

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* , *30*(5), 614–620.