ABRAXAS
Biosystems

INFORME DE LOS SERVICIOS DE EXPERTOS

Análisis genómico y Bioinformático

de la secuenciación de ADN de

alto rendimiento de muestras extraídas de los cuerpos desecados encontrados en Nazca.

Cliente :

GAIA INTERNATIONAL, INC.

JOSÉ JAIME MAUSSAN FLOTA

9 de noviembre de 2018 1

ABRAXAS
Bringing knowledge to life

Tabla de contenidos

VISIÓN GENERAL

Este documento proporciona detalles de todos los trabajos, tareas y procedimientos involucrados en el servicio proporcionado por ABRAXAS biosistemas S.A.P.I. de C.V. para GAIA INTERNATIONAL, INC. y José JAIME MAUSSAN flotó para el proyecto "genómica y análisis Bioinformático de la secuenciación de ADN de alto rendimiento de muestras extraídas de cuerpos desecados encontrados en Nazca". Presentamos una descripción ordenada de las principales tareas y el análisis desarrollado para este proyecto.

Las muestras de tejido extraídas de los cuerpos desecados encontrados en Nazca y utilizadas para los análisis presentados en este servicio fueron proporcionadas, dirigidas y manejadas por JOSE JAIME MAUSSAN FLOTA y sus colegas científicos en todas las etapas anteriores a la extracción del ADN descrita en este informe mientras que los Laboratorios CEN4GEN (6756 - 75 Street NW Edmonton), AB) Canada T6E 6T9) fueron responsables de realizar todas las tareas en las muestras, desde la extracción del ADN hasta la secuenciación del ADN de alto rendimiento, también conocida como Secuenciación de Próxima Generación, y los pasos para generar datos de secuenciación limpia.

ABRAXAS BIOSYSTEMS S.A.P.I. DE C.V. ha llevado a cabo toda la genómica computacional y el Análisis de Bioinformática.

Para este proyecto, José JAIME MAUSSAN FLOTA y sus colegas científicos aseguraron la entrega a los laboratorios CEN4GEN, de 7 muestras, 3 muestras de tejido y 4 muestras de ADN de los cuerpos encontrados en Nazca, Perú. Después de la extracción del ADN, control de calidad y procedimientos de amplificación MDA en el laboratorio de CEN4GEN, sólo 3 muestras, de los 7 originales, pasaron los controles NGS, los nombres de estas muestras, de los tubos originales enviados para la entrega, fueron los siguientes:

Nombre de la muestra	Etiqueta original de la muestra	identidad
Antiguo -0002	Huesos de cuello da entidad sentada	Victoria
Antiguo -0003	1 Mano 001	Mano grande
Antiguo -0004	Momia 5 - ADN	Victoria

Tabla 1 : El nombre de la muestra indica el nombre asignado a la muestra por CEN4GEN. La etiqueta original de la muestra es el nombre en el tubo donde estaba originalmente contenida la muestra cuando se entregó a CEN4GEN, Identidad es el nombre del cuerpo del que proviene la muestra.

Esta es la razón por la que todas las tareas analíticas mencionadas en este informe después de la extracción del ADN, el control de calidad y las tareas de amplificación de MDA se realizaron solo para estas 3 muestras.

ÍNDICE DE SERVICIOS

La solución completa ofrecida por Abraxas Biosystems incluye una amplia gama de servicios que van desde la extracción de ADN antiguo hasta la secuenciación y el análisis de datos (bioinformática), permitiendo la generación de resultados precisos a partir de análisis de muestras antiguas. Las tareas de análisis realizadas en el marco de este proyecto son las siguientes:

1. Extracción de ADN.
2. Control de calidad del ADN.
3. Amplificación de ADN mediante amplificación por desplazamiento múltiple.
4. Construcción de bancos de ADN.
5. Secuenciación de ADN de última generación (NGS).
6. Generación de datos de secuenciación limpia.
7. QC de los resultados de la secuenciación.
8. Análisis cartográfico preliminar de las lecturas de ADN en referencia al genoma humano.
9. Análisis de referencias cruzadas para detectar fragmentos cortos comunes al ADN antiguo.
10. Mapeo de las lecturas de ADN superpuestas desde Ancient0003 a la versión más reciente del genoma humano.
11. Análisis mitocondrial para la detección de variantes en bucles D y otras regiones informativas para determinar haplotipos mitocondriales.
12. Determinación del sexo de la muestra Ancient0003.
13. Detección de cualquier organismo presente en la muestra mediante el método de esquema del ADN genómico (coincidencias exactas de grupos de fragmentos cortos, *k-mers*, con bases de datos públicas) y filtrado iterativo de lecturas de correspondencia exactas *k mer*.

14. Ensamblaje de novo con estrategias mixtas de ADN de lecturas sin correspondencia com los organismos detectados en el método de esquema.

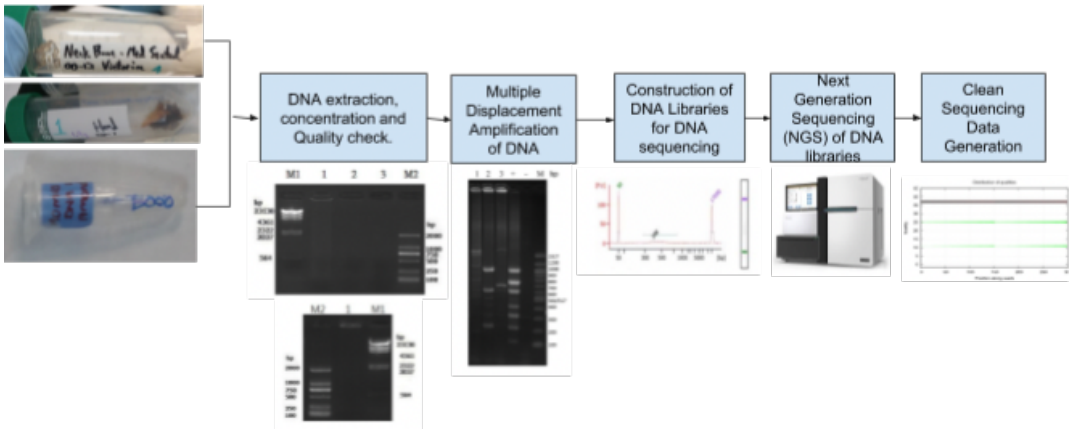
15. Mapeo de lecturas sin correspondencia exacta en el proceso de filtrado iterativo con las secuencias resultantes en el montaje de *novo (la síntesis de* novo *es la síntesis de uma molécula)*.

16. Búsqueda de bases de datos de ADN para segmentos de ADN ensamblados de *novo* para detectar similitudes con organismos conocidos.

17. Clasificación taxonómica de secuencias no coincidentes en pasos anteriores mediante búsquedas coincidentes em bases de datos genéticas completas.

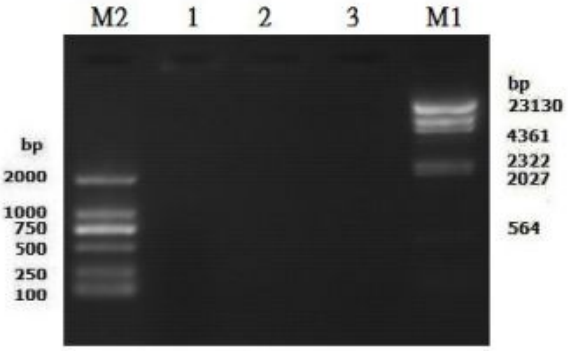
PROGRESO DE LAS OPERACIONES DE SERVICIOS DE LABORATORIO

Para los procesos de análisis de laboratorio, tareas 1 a 6, las operaciones generales fueron las siguientes.

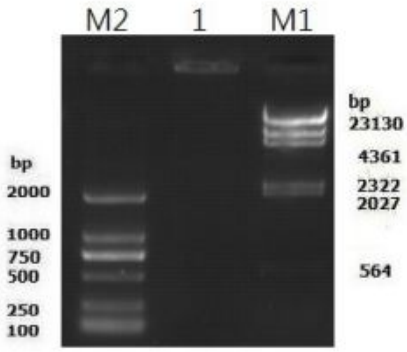


EXTRACCIÓN DEL ADN

Las 3 muestras de tejido fueron analizadas según un protocolo de extracción de ADN específico para muestras antiguas y desarrollado en los laboratorios CEN4GEN, basado en los protocolos descritos en el siguiente artículo (Gamba et al., 2016). Después del proceso de extracción de ADN, se pasó el ADN sobre geles de agarosa para comprobar la presencia de bandas que indicaban la presencia de cantidades adecuadas de ADN (indicadas por bandas luminosas horizontales visibles en cada pista correspondiente a cada muestra). Además, el ADN ya extraído de Ancient-004 fue verificado por este método porque contenía ADN de una muestra de tejido que ya no estaba disponible en el cuerpo de Victoria. Los resultados se presentan en la figura siguiente :



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	



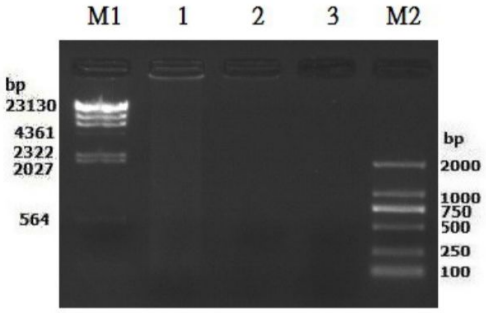
Lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Sample Integrity
M1	λ-Hind III digest(Takara)	1	3	
1	CEN4GEN-Ancient0004	1	3	
M2	D2000 (Tiangen)	1	6	

Figura 1: Resultados de la extracción del ADN (arriba) y control de calidad del ADN ya extraído (abajo). Las líneas M2 y M1 de cada gel son los marcadores moleculares utilizados para medir el tamaño de los fragmentos de ADN.

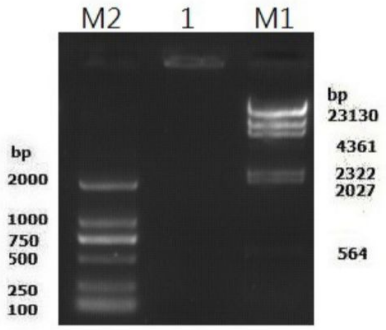
La Figura 1 muestra que las bandas a nivel de las muestras de ADN extraídas no tenían suficiente visibilidad, por lo tanto, no había suficiente ADN para obtener resultados correctos de la NGS. Esto obligó a los laboratorios a evaluar las 4 muestras de ADN para ver si tenían suficiente ADN.

ANÁLISIS DE LA CALIDAD DEL ADN

Después de la extracción de ADN de las 3 muestras de tejido, las 4 muestras de ADN fueron analizadas mediante un proceso de control de calidad para evaluar la presencia de buenas cantidades y tamaños de ADN para ver si podían guardar el ADN necesario para el NGS. El control de calidad también se llevó a cabo en los geles de agarosa como en el paso anterior. Los resultados se muestran a continuación:



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	

Figura 2: Control de calidad del ADN de muestras de ADN ya extraídas pero no analizadas durante el control de calidad anterior (arriba) y control de calidad del ADN ya extraído y previamente analizado con muestras de tejido de extracción de ADN (abajo).

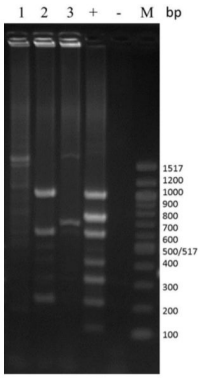
Los resultados son los mismos que para el ADN extraído de los tejidos por los laboratorios CEN4GEN, que muestran una presencia muy baja de ADNe. Esta falta de altas cantidades de ADN y la ventaja de NGS para recuperar datos de entradas bajas de ADN con algunos esfuerzos de amplificación llevó al equipo de laboratorio del CEN4GEN a realizar un proceso llamado Amplificación por Desplazamiento Múltiple que había mostrado un buen resultado en sus instalaciones con muestras antiguas para amplificar los niveles de ADN disponibles necesarios para la secuenciación de NGS.

AMPLIFICACIÓN DEL DNA

Por Amplificación de Desplazamiento Múltiple

Después de encontrar cantidades muy pequeñas resultantes de la condición degradada de las muestras de tejido, los laboratorios recurrieron al proceso MDA para amplificar las cantidades de fragmentos de ADN. Este proceso ha sido adaptado a las características del ADN extraído utilizando métodos propios de los Laboratorios CEN4GEN. Los resultados del MDA fueron aceptables para la prealimentación con NGS para 2 de las muestras de ADN extraídas por CEN4GEN (Ancient0002 y Ancient0003) y para 1 de las muestras ya entregadas extraídas (Ancient0004) como se muestra a continuación :

Electrophoretogram:



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	/

Figura 3: Resultados de la amplificación MDA, los 3 primeros canales corresponden a los valores amplificados con éxito de las muestras y los últimos 4 y 5 canales están previstos respectivamente al control negativo y a los marcadores moleculares.

Las muestras restantes no mostraron los resultados de la amplificación, por lo que las tareas analíticas restantes se realizaron sólo para estas tres muestras.

PRELIMINAR ANÁLISIS DE MAPEO

Para obtener una aproximación rápida de la relación de las muestras con el ADN humano, las lecturas verificadas con QC se submuestrearon a una porción del 25% (una fracción del 25% de las lecturas se extrajeron al azar del total de cada una de las ejecuciones de secuenciación correspondientes a cada muestra , excepto en la muestra 2 para la cual solo se realizó una muestra de una de las dos ejecuciones) mapeada contra la referencia del genoma humano desenmascarado en la versión más posiblemente actualizada a las fechas del análisis que pudimos obtener, esta fue la versión GRCH38 versión 93 descargada de: ftp.ensembl.org/pub/release-93/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz

Esta secuencia del genoma nos permitió tener una secuencia de referencia de ADN para comparar las lecturas de la posible presencia de ADN humano en las muestras comparando cada lectura con esta referencia. La comparación se realizó utilizando el **software bwa mem mapper en su versión 0.7.17-r1188**. El proceso de mapeo mostró los siguientes resultados:

Muestra	Numeros de lecturas	Nros de lecturas coindidentes	%	% de lecturas en parejas	Tasa de brecha
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

La tabla anterior muestra que Ancient0002 y Ancient0004 tienen muy pocas cantidades de ADN que podrían ser de origen humano en comparación con la muestra de Ancient0003 que muestra una alta señal de relación con el ADN humano.

ANÁLISIS DE SIMILITUD CON ORGANISMOS CONOCIDOS

Los *cóntigos* resultantes se analizaron contra la base de datos NT mediante BLASTn v 2.7.1 (utilizando un valor E de 10, un tamaño de palabra de 20, y un porcentaje de identidad de 30) para buscar posibles coincidencias con organizaciones conocido en la base de datos NT y el número de resultados fue contado para determinar si los fragmentos ensamblados tenían mejores resultados consiguiendo un fósforo con los organismos sabidos. En total, para la muestra de Ancient0002, solamente 1 256 (de 60 852) *cóntigos* no tenía ninguna correspondencia y para Ancient0004 solamente 1 768 (de 54 273) *cóntigos* no recibió ninguna correspondencia.

También otra ensamblaje de *novo* que utilizó como entrada los dos sistemas de las lecturas únicas incomparables de Ancient0002 y de Ancient0004 junto, pero los resultados de montaje fueron mucho menos ensamblados y tuvieron resultados mucho más fragmentados, por lo que este montaje extra fue descartado.

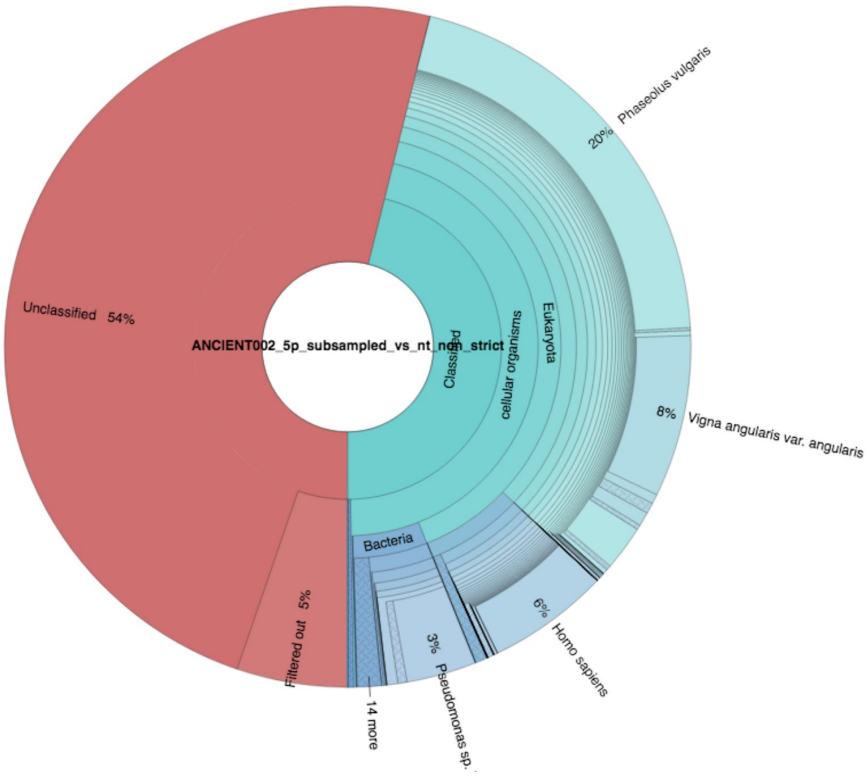
CLASIFICACIÓN TAXONÓMICA ULTRA COMPLETA

Para lecturas de submuestras únicos y en bruto.

Las tareas anteriores consistían en encontrar la cantidad de lecturas de ADN clasificables de las muestras de los modelos Ancient0002 y Ancient0004 para comprender hasta qué punto las muestras con una cartografía débil del genoma humano se parecen a organismos conocidos a nivel del ADN. Aunque nuestro esquema genómico y el enfoque de filtrado iterativo eran completos para muchos rangos taxonómicos y tipos de organismos, que necesitábamos para producir una clasificación aún más detallada de las lecturas de secuenciación de ADN en todos los niveles taxonómicos posibles, y con una base de datos aún mayor con un algoritmo de comparación más flexible para ampliar, en la medida de lo posible, los espectros y el poder de detección de nuestros métodos de comparación de ADN con organismos conocidos y secuenciados en la Tierra. Para lograr este mayor espectro de detección y clasificación detallada en todos los rangos taxonómicos posibles, estamos implementando una estrategia ultra-comprensiva y altamente sensible basada en la creación de una nueva base de datos con aún más entidades, con uno de los conjuntos más completos de datos de secuencias conocidos en bioinformática como la base de datos NCBI, construido utilizando algoritmos de compresión y redundancia y también basado en la implementación de una búsqueda óptima de coincidencias inexactas comparable en sensibilidad a la altamente sensible BLAST pero, en la práctica, esta búsqueda utilizando una búsqueda BLAST habría llevado meses, extendiendo así el poder de búsqueda del boceto y el filtrado iterativo aunque esta estrategia se limita a la precisión de las coincidencias. Esta estrategia se implementó utilizando el software taxmaps v 0.2.1 (Corvelo, Clarke, Robine, & Zody, 2018) y se aplicó a un subconjunto del 5% de todas las lecturas brutas no filtradas de las muestras de los modelos Ancient0002 y Ancient0003.

El mismo análisis se repitió para un subconjunto del 25% de la muestra de Ancient0004, demasiado justo para confirmar lo que nuestros métodos predijeron correctamente para las proporciones de lecturas clasificadas y no clasificadas a medida que las muestras se acercaban a las lecturas generales (lo que requería docenas de días adicionales).

Esta estrategia también nos permitió comparar si el comportamiento de los procesos de superposición y filtrado era diferente de las lecturas no filtradas originalmente secuenciadas en su correspondencia con organismos conocidos. ADN Los resultados se muestran a continuación :



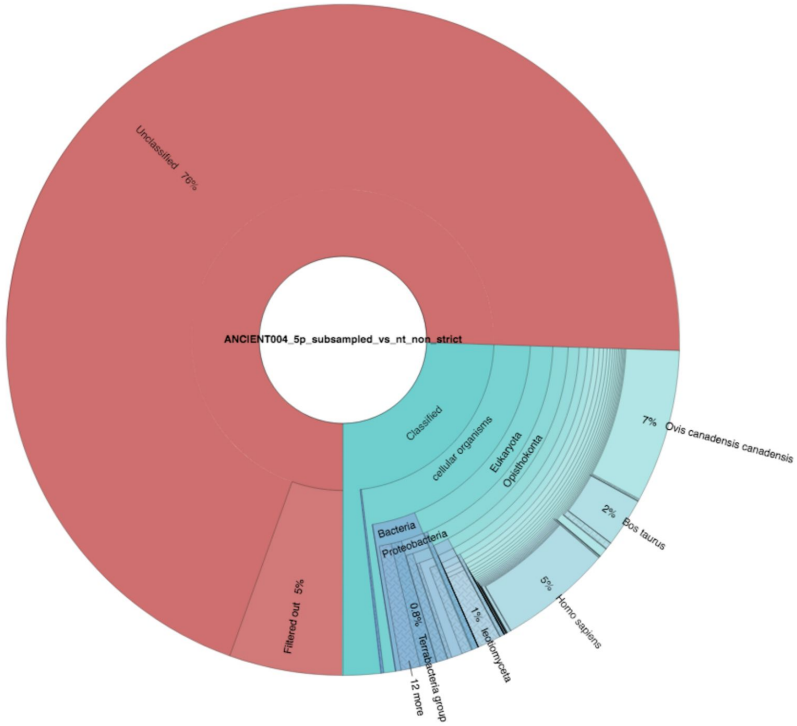


Figura 8: Proporción de lecturas clasificadas y no clasificadas de una submuestra del 5% (28.073.655 lecturas para Ancient0002 y 25.084.962 para Ancient0004) de las lecturas de secuenciación bruta completa para Ancient 0002 (arriba) y Ancient0004 (abajo) en comparación con la base de datos del NCBI, tal como se implementa en las tarjetas de impuestos 0.2.1, que incluye 34.904.805 secuencias de ADN que representan 1.109.518 taxones.

Este enfoque confirmó la presencia de tasas muy altas de ADN no correspondiente y no clasificado contenido en las muestras secuenciales en comparación con una de las bases de datos públicas más completas de información genómica en los parámetros considerados (una distancia máxima de modificación permitida de 0,2 entre los k mers buscados por *taxmaps* y la base de datos no redundante implementada para la base de datos nt).

CONCLUSIONES

Abraxas Biosystems ha realizado una amplia gama de análisis bioinformáticos y genómicos para identificar el posible origen biológico y la ascendencia de las muestras proporcionadas por Jaime Maussan y sus colegas científicos y extraídas/secuenciadas en los laboratorios CEN4GEN. Después de diseñar un protocolo meticulosamente personalizado para maximizar la tasa de éxito de la antigua extracción de ADN, la secuenciación (con los laboratorios CEN4GEN) y el análisis bioinformático de las muestras, los resultados muestran una correlación muy baja con los datos del genoma humano

para las muestras Ancient0002 y Ancient0004, a diferencia de la muestra Ancient0003 que mostró un mapeo muy alto correspondiente al genoma humano. También se debe tener en cuenta que los ejemplos de Ancient0002 y Ancient0004 muestran muy pocas coincidencias con una de las bases de datos más confiables y precisas (de NCBI). Sin embargo, las bases de datos del NCBI no contienen todos los organismos conocidos en el mundo, por lo que podría haber muchos organismos posibles que podrían corresponder a este ADN o algunas regiones que podrían ser excluidas o difíciles de secuenciar, comunes a muchos organismos en estas muestras y en los protocolos aplicados a los genomas informados por el NCBI.

Los protocolos de laboratorio y computacionales para el análisis de ADN antiguo, dada la naturaleza de las muestras, incluyen varios pasos que pueden interferir con los datos y tener un impacto directo en los resultados. Uno de los ejemplos más comunes es la manipulación de tejidos por varios individuos y en un ambiente abierto antes de su aislamiento, lo que complica la posibilidad de que todo el ADN secuenciado provenga del ADN endógeno de los cuerpos individuales muestreados. Una forma de evitar este tipo de ruido y obtener mejores resultados es secuenciar muestras de hueso interno y no tejidos expuestos.

Por último, las bases de datos actuales del NCBI están en constante desarrollo, por lo que es posible que pronto se pueda desarrollar una base de datos mejor y más completa que incluya más genomas microbianos y/o eucariotas disponibles que puedan arrojar luz sobre la naturaleza de las muestras de ADN no correspondientes. Además, se podría desarrollar un análisis específico de los segmentos de ADN no correspondientes para confirmar que no están secuenciando restos o protocolos de amplificación. Los antiguos protocolos de ADN se mejoran continuamente debido a su sensibilidad a las características de degradación de este tipo de muestras. Se recomiendan estudios adicionales para aceptar o rechazar cualquier otra conclusión.