Advertencia, por favor lea cuidadosamente

El presente documento y su contenido ***son estrictamente para los fines previstos y no pueden ser utilizados, publicados o redistribuidos sin el consentimiento previo por escrito del autor.***

Las opiniones expresadas son de buena fe y, si bien se han tomado todas las precauciones para preparar el presente documento, el autor no puede ser considerado responsable del mal uso y mal interpretación de los datos y opiniones que contiene.

El autor declara que no hay intereses financieros en competencia y se autofinancia su propia parte en esta investigación.

El autor quiere mantener el anonimato. Para más consultas o información, como los archivos de secuencias no identificadas, Alain Bonnet recopilará los detalles de contacto, CV, calificaciones y una breve carta de presentación (idealmente en inglés) de las personas interesadas. https://www.the-alien-project.com/es/contacto/

De anónimo, Phd *(Doctorado)*
A la atención de : Thierry Jamin, Alain Bonnet.

Objeto: Presentación y discusión de los análisis genómicos detallados llevados a cabo en dos especies de individuos no identificados que se encontraron en el desierto de Nazca, Perú, en 2015.

1. Introducción: características fenotípicas de los dos individuos.

1.1 María:

El individuo llamado Maria en resumen fue encontrada en Nazca, Perú, en 2015.

María es una criatura humanoide de género no identificada de 165-170 cm de altura en posición fetal. A pesar de la obvia similitud con el homo sapiens, el individuo muestra rasgos fenotípicos atípicos, entre los cuales: volumen del cráneo aproximadamente un 25% mayor que el *Homo sapiens* (a nivel parietal), tres dedos en cada mano y pie, falanges muy largas. Dos laboratorios estimaron de forma independiente la edad de radiocarbono del individuo en el mismo valor de 1750 ± 30 PA.

1.2 Mano Grande:

En el mismo sitio se encontraron varias manos con características similares, tres dedos, seis falanges y de tamaño considerable. La edad de uno de ellos, que será llamado Mano Grande en la presentación actual por simplicidad, ha sido estimada en 6420 ± 30 BP.

1.3 Preguntas planteadas - autenticidad y naturaleza de las criaturas.

¿ Son estos seres criaturas biológicas genuinas o solo un arreglo refinado de otras especies ya conocidas, animales o humanos ? Ha sido definitivamente la pregunta inmediata planteada por la gente que se han acercado al material y, ciertamente, por el público.

Antes de discutir esta cuestión, se deben tener en cuenta dos puntos:
(i) El primer punto es que en la ciencia, cuando se presenta una hipótesis, el primer paso consiste en revisar sus suposiciones e implicaciones directas y verificar si son coherentes con la evidencia disponible. Si esta etapa previa es exitosa, entonces un segundo paso consiste en probar las predicciones de tal hipótesis.
(ii)Estos cuerpos tienen 1700 y 6500 años de antigüedad, respectivamente. Los tejidos están secos, duros y tienden a desmoronarse. En consecuencia, las intervenciones quirúrgicas eventuales no pueden haberse realizado recientemente, sino más bien a la muerte de los sujetos, es decir, hace más de 1700 años para María y hace más de 6000 años para la Mano Grande.

Entonces, ¿qué tan probable es la hipótesis de un *"arreglo refinado de otras especies ya conocidas, animales o humanas"* ? Esta hipótesis implica la existencia de rastros (lesiones, cicatrices) que deben ser visibles, ya que no hay proceso de cicatrización después de la muerte. También asume ciertos niveles de tecnología y conocimiento que se requieren para producir tales individuos.
Las cuidadosas observaciones anatómicas, incluidas las tomografías computarizadas, del individuo revelaron detalles particularmente realistas y refinados (huellas dactilares, dientes adultos, superficie externa e interna del cráneo incluyendo suturas, piel, vértebras, costillas, articulaciones y articulaciones, órganos aparentemente internos). Además, no se pudo detectar ninguna lesión en los huesos o tejidos de la piel que sugiriera una intervención quirúrgica. Como tal, la hipótesis antes mencionada suena improbable debido a que
i) la ausencia de pruebas que sugieran una cirugía o manipulación similar
ii)los detalles anatómicos que requerirían, para ser emulados, el despliegue de medios biotecnológicos a priori no disponibles en ese momento e incluso en la actualidad
(iii) la presencia de otro individuo, un bebé, que fue encontrado en el mismo lugar y que presenta las mismas características atípicas que María. Detalles como los dientes de leche y las proporciones cuerpo/cabeza confirman que se trata de un bebé genuino y no de un adulto de tamaño pequeño.

Una objeción recurrente ha sido que las lesiones o cicatrices podrían ser tan sutiles que podríamos haberlas omitido. Efectivamente, la resolución de los escáneres utilizados permitía ver muchos detalles refinados, pero no era la más alta disponible en el mercado. Sin embargo, esta objeción sigue presuponiendo un nivel de tecnología y ciencia incompatible con el marco temporal evocado por los análisis de radiocarbono, y me gustaría llamar su atención sobre a dónde conduce esta hipótesis. ¿Qué dispositivos o herramientas quirúrgicas antiguas podrían haber funcionado de una manera tan sutil que nuestros escáneres modernos habrían perdido las huellas que quedaban en el cuerpo? ¿Podemos **asumir razonablemente la existencia de instalaciones de laboratorio de biotecnología de vanguardia en el desierto de Nazca entre 1700 y 6500 años atrás?** Esto es simplemente absurdo, o al menos no es compatible con la evidencia disponible.

Más bien, consideraremos las predicciones de esta hipótesis.
Especialmente, si estas criaturas han sido construidas con restos humanos y/o animales, entonces los análisis de ADN deben mostrar, ***después de la remoción del ADN contaminante*** (i) ya sea ADN 100% moderno de homo sapiens o (ii) parcialmente moderno de homo sapiens y parcialmente animal, probablemente aquellos presentes localmente en el Perú. Estos son los análisis que hemos realizado y que vamos a presentar hoy.

1.3.2 Plan de presentación.

Después de esta larga pero necesaria introducción, procederemos de la siguiente manera.

1) Material y métodos: una breve presentación de las fuentes de los datos en los que hemos trabajado (ADN antiguo, secuenciación de ADN) y los métodos (análisis de alineación)

2) El resultado de la primera ronda de análisis: identificación de contaminantes, análisis de alineación con el Homo sapiens moderno. Aislamiento de secuencias no mapeadas (indeterminadas)

3) El resultado de la segunda ronda de análisis, llevada a cabo en secuencias no mapeadas: Comparación con otras especies 4) Conclusiones e interpretaciones

5) Apéndice

1. Fuentes de datos y métodos.

La extracción y secuenciación fue realizada por otro laboratorio con sede en México, BioTechMol http://biotecmol.mx/. Los análisis genómicos que se realizaron no mencionaron los métodos utilizados y no fueron exhaustivos. Por lo tanto, realizamos los análisis de los archivos de datos sin procesar para verificar primero la calidad de los datos y luego caracterizar la totalidad del ADN secuenciado, incluidos los contaminantes, en las secuencias de la más alta calidad.

Extracción:

El ADN es frágil y las muestras antiguas de ADN generalmente están muy dañadas y contaminadas por virus, bacterias, microorganismos y personas / animales en contacto directo con la muestra. Para muestras antiguas se deben seguir procedimientos específicos. Se usó también una muestra de hueso de 0,54 g recolectada en el cuerpo de Maria y un tejido óseo / desconocido de 2,38 g recolectado en la mano grande que permitió extraer suficiente ADN (procedimiento realizado por Shapiro B, Heslington M. 2012 + kit de reparación, kit PreCR® Reparación Mix de New England Biolabs M0309S).

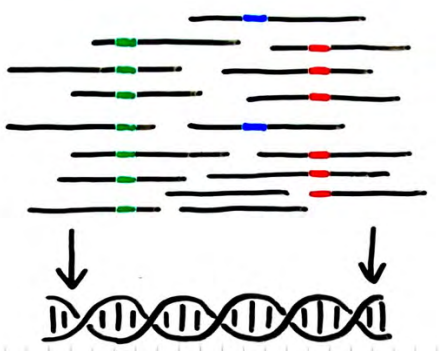
Secuenciación y análisis:

Unas pocas palabras quizás para un público no técnico. El ADN se compone principalmente de cuatro nucleótidos, o bases. Combinados juntos, estos nucleótidos forman (4*4*4) 64 trillizos llamados "codones" que constituyen un alfabeto. Estamos lejos de dominar la sintaxis del ADN, pero se conocen algunas reglas, por ejemplo, algunos codones indican el inicio de una secuencia de codificación, mientras que otros indican el final.

Secuenciar el ADN significa *determinar el orden de los nucleótidos que componen el ADN*. Se trata de una técnica molecular que ahora se lleva a cabo mediante máquinas y ordenadores, y que sólo es supervisada por un humano. Un aspecto desafiante de la secuenciación es que la molécula de ADN es muy larga (3 mil millones de nucleótidos para el ADN humano), por lo que la secuenciación no se puede hacer una sola vez. Más bien, la secuenciación del ADN se lleva a cabo en "pequeñas piezas" de ADN que dan lugar a lecturas (entre 50 y 150 nucleótidos, o bases). Posteriormente, el genoma se reconstruye mediante un procedimiento llamado *ensamblaje del genoma*.

Para facilitar la comprensión, se puede imaginar un texto que se cortaría de forma aleatoria cada 5-10 palabras. Luego se copiarían los trozos y, finalmente, el texto tendría que ser reconstruido de manera coherente y significativa, a la vez que se detectarían eventuales errores de copia.

Para hacer posible la reconstrucción, las "pequeñas piezas" son de diferentes longitudes, secuenciadas y superpuestas varias veces. Las superposiciones y las reglas sintácticas permiten reconstruir el genoma una vez secuenciado. La superposición también se utiliza para identificar posibles errores de secuenciación (por ejemplo, para hacerlo simple, si una región está secuenciada 10 veces y sólo 5 leídas son idénticas, se puede considerar que hay una alta tasa de error en esta región, y tal vez eliminarla del siguiente análisis).



Se llevó a cabo una secuenciación adicional en la plataforma Myseq Illumina (las plataformas Illumina se encuentran entre las más utilizadas).

Análisis:
El objetivo de estos análisis era caracterizar, cuantitativa y cualitativamente, el contenido del material de ADN, incluidos los contaminantes.
Tenga en cuenta que este procedimiento requiere varios pasos después de la secuenciación, especialmente el ensamblaje del genoma como se mencionó anteriormente y la alineación. La alineación se utiliza para la comparación del genoma cuando se necesita identificar una especie, por ejemplo. Consiste en mapear el ADN secuenciado en el ADN de todos los genomas disponibles para una especie determinada.

Los datos resultantes (archivo Illumina fastq) se comprobaron y analizaron (véase el Apéndice para más detalles) para determinar el contenido genómico cualitativa y cuantitativamente en referencia a la *base de datos RefSeq Complete Genomes* https://www.ncbi.nlm.nih.gov/refseq/.

2.Identificación de contaminantes/virus, comparación con el Homo sapiens moderno, Aislamiento de virus no mapeado. Secuencias (indeterminadas)

2.1 María

Los resultados se muestran en términos de porcentaje de mapeo absoluto a las regiones del genoma humano y bacteriano/contaminante, y el resto se considera como no mapeado/no clasificado.

33.7% de las lecturas estaban alineadas con el homo sapiens moderno
18.4% de las lecturas eran contaminantes
47,9% no fue mapeado.

Maria - Resumen general

Número total de Reads *(lecturas)* (329037x2)	658074	
Número de bases - READ1	46674295	4.6Mbp
Número de bases - READ2	46928318	4.6Mbp
Número total de bases	93602613	9.36Mbp

Genoma humano	
Reads alineados al genoma humano GrCh38	221623
Número de bases alineadas (alineación única)	5782559
Porcentaje de Reads mapeadas al genoma humano	33.6775

Bacterias y otros genoma	
Reads alineados a genomas bacterianos.	121186
Número de bases alineadas (alineación única)	17183634
Porcentaje de Reads asignadas a bacterias y otros genomas	18.415

Desconocido	
Número de Reads no alineadas/no mapeadas/no clasificadas	315265
Número total de bases no clasificadas	44756142
Porcentaje del total de Reads no mapeado/no clasificado	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

Mapeo de Maria con Bacterias y otros genomas contaminantes

Número total de lecturas mapeadas	121186
Porcentaje de lecturas alineadas	18.415

Bacteria	Número de 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
Otros	34674
Virus y otros	9384

2.2 Mano Grande

Como antes, los resultados se muestran en términos de porcentaje de mapeo absoluto a las regiones del genoma humano y bacteriano/contaminante, mientras que el resto se considera como no mapeado/no clasificado.

0.37% de las lecturas estaban alineadas con el homo sapiens moderno
26.7% de las lecturas eran contaminantes
72,9% no fue mapeado.

MANO GRANDE - Resumen General

Número total de Reads *(lecturas)* (341311x2)	682,622	
Número de bases - READ1	51,160,199	5.1Mbp
Número de bases - READ2	51,205,429	5.1Mbp
Número total de bases	102,365,628	10.2Mbp

Genoma humano	
Reads alineados al genoma humano GrCh38	2,518
Número de bases alineadas (alineación única)	366,819
Porcentaje de Reads mapeadas al genoma humano	0.3689

Bacterias y otros genoma	
Lecturas alineadas con bacterias, virus y otros genomas	182,243
Número de bases alineadas (alineación única)	27,331,623
Porcentaje de Reads asignadas a bacterias y otros genomas	26.7

Desconocido	
Número de Reads no alineadas/no mapeadas/no clasificadas	497,836
Número total de bases no clasificadas	74,655,252
Porcentaje del total de Reads no mapeado/no clasificado	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

Mapeo de la Mano Grande con Bacterias y otros genomas contaminantes

Bacteria	Número de 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
Otras bacterias	167154
Virus y plasmidosos	1361

3. Secuencias no mapeadas: Comparación con otras especies

Realizamos una segunda ronda para caracterizar las secuencias que no estaban mapeadas.
Varias especies fueron utilizadas para la comparación tanto para María como para la Mano Grande, incluyendo: Alpaca, Babuino, Perro, Gato, Caballo, Chimpancé, Rhesus Macaque.
Los resultados fueron negativos para ambos sujetos.

Organismo	Total de lecturas	Lecturas (Reads) alineadas	% Alineación
Maria sin mapeo			
Alpaca	315265	1	0.00032
Babuino	315265	0	0.00000
Perro	315265	366	0.11609
Gato	315265	31	0.00983
Caballo	315265	0	0.00000
Chimpancé	315265	1	0.00032
Macaco Rhesus	315265	1	0.00032
Mano Grande no mapeado			
Alpaca	497861	11	0.00221
Babuino	497861	0	0.00000
Perro	497861	240	0.04821
Gato	497861	38	0.00763
Caballo	497861	2	0.00040
Chimpancé	497861	25	0.00502
Macaco Rhesus	497861	3	0.00060

NB: Tenga en cuenta que el número de lecturas alineadas para Perro no es significativo para la alineación, pero es significativamente mayor que el de las otras especies de referencia. Supusimos que los huaqueros que encontraron los cuerpos tenían perros.

Todavía se están llevando a cabo análisis de alineación para identificar la naturaleza de estas secuencias no mapeadas/no clasificadas - casi la mitad de la muestra. En este momento, están retenidos como no identificados.

4.Conclusiones generales

Durante estas investigaciones se han planteado varias cuestiones. Los más frecuentes fueron los siguientes: (a) la autenticidad de los individuos, (b) sus eventuales vínculos o similitudes con el homo sapiens, (c) sus orígenes.

a) la autenticidad de los individuos:

Los puntos siguientes:

(i) Los detalles anatómicos sutiles (articulaciones y articulaciones, tejidos de la piel, superficie interna del cráneo, suturas craneales, gradiente de densidad ósea, órganos internos, huellas dactilares...).

ii) la ausencia de cicatrices o de lesiones mecánicas o quirúrgicas detectadas en los tejidos

iii)el hecho de que no se haya encontrado ADN animal (varias especies locales sometidas a pruebas)

iv)el hecho de que las secuencias modernas de ADN humano estaban presentes en un individuo en un porcentaje menor

v) la presencia de un bebé individual en el mismo lugar, con las mismas características atípicas

sugieren que aunque no podemos ser formales sobre la autenticidad, ninguna evidencia material acredita la "hipótesis falsa". Más bien, este creciente conjunto de pruebas sugiere más bien que **podríamos estar en presencia de especies biológicamente indefinidas, que merecen más investigaciones.**

b) Sus posibles vínculos o similitudes con el homo sapiens

Aquí hay que establecer las definiciones. Las especies no se definen sobre la base de su apariencia o morfología similares. Dos individuos pertenecen a la misma especie si, y sólo si pueden cruzarse. En el marco de la teoría de la evolución, se puede decir que hay un continuo entre especies. Dos especies con un antepasado común se consideran totalmente diferenciadas si no pueden cruzarse, lo que incluye aparearse y tener descendencia fértil. Por ejemplo, los leones y los tigres pueden aparearse, pero su descendencia es estéril. También lo son los caballos y los burros. Ambos tienen un antepasado común del cual evolucionaron (el proceso se llama especiación) de manera diferente, pero aún no están totalmente diferenciados, en la medida en que todavía pueden aparearse.

Por el contrario, los Homo sapiens y los simios (chimpancés, por ejemplo) que se considera que tienen un antepasado común, no pueden aparearse y tener descendencia. Se mantienen como dos especies diferentes. Sin embargo, las especies conocidas, especialmente los mamíferos, tienen un alto porcentaje de secuencias de ADN en común, más del 95% dependiendo del caso. Basándonos únicamente en el genoma, no podemos responder a esta pregunta de manera precisa.

(c) Sus orígenes ET

No identificado no significa extraterrestre. Tenga en cuenta que María parece estar completamente equipada para sobrevivir y moverse en la biosfera de la Tierra. Por lo tanto, desde un punto de vista biológico, a priori no hay nada que sugiera que ella vendría de otro planeta. Además, no existe una base de datos para genomas ET o exobiología en general. Por lo tanto, no podemos comparar su genoma con nada que se tenga como ET y, por lo tanto, no podemos hablar de sus posibles orígenes. Hoy en día, los orígenes de ET NO son una información genómica (todavía).

Las secuencias no mapeadas están disponibles a pedido: Los datos de contacto y la prueba de cualificaciones y recursos computacionales (*Cloud* está bien) se pueden dar a Thierry Jamin.

Apéndice

Materiales y métodos

Datos.

Los archivos Illumina de extremo emparejado estaban disponibles para la muestra en formato fastq. Los archivos de hebras hacia adelante y hacia atrás contenían 329037 lecturas para María, y 341311 las lee para la Mano Grande. Los archivos fastq se convirtieron a formato fasta usando scripts personalizados para contar el número de bases.

Datos sin procesar QC

Los archivos finales emparejados se sometieron a control de calidad por calidad de secuencia, calidad por base, contenido de k-mer, adaptador y otra contaminación. La herramienta de control de calidad NGS Fastqc se utilizó para este propósito. Los adaptadores Illumina y la lista de contaminantes estándar se utilizaron para filtrar cualquier ruido conocido. Las lecturas fueron de diferentes longitudes que van desde 35 hasta 151 bases.
Más del 90% de las lecturas superaron la puntuación de phred de Q30 y un número mínimo por debajo de Q20. En general, los datos sin procesar fueron de buena calidad de secuencia y no mostraron signos de agotamiento de reactivos o artefactos de secuenciación.

Alineación al genoma humano (GRCh38)

Después de pasar el control de calidad, las lecturas se alinearon con el ensamblaje más reciente y más completo del genoma humano, GRCh38. La alineación se realizó utilizando el algoritmo bwa-mem con opciones estrictas. Bwa-mem es el algoritmo de alineación de elección para lecturas de más de 100 pb de longitud debido a su velocidad y precisión, especialmente con los genomas de mamíferos. Aproximadamente el 33% y menos del 1% de las lecturas se alinearon con el genoma humano GRCh38 y la alineación se almacenó en formato de mapa de alineación binaria (bam).

Bamtools se usó para filtrar las lecturas asignadas y no asignadas. Las lecturas no asignadas extraídas de este archivo de alineación se convirtieron a formato fastq para su posterior análisis.

Clasificación de lecturas no mapeadas por alineaciones exactas de k-mers

Las lecturas no mapeadas del paso de alineación anterior se sometieron a un análisis de clasificación utilizando Kraken. Kraken asigna etiquetas taxonómicas a las lecturas de secuenciación basadas en la alineación exacta de k-mers contra grupos de genomas (bacterianos, plásmidos, virus, etc.). Una base de datos de referencia fue construida por primera vez a partir de genomas bacterianos, arqueológicos y virales completos en RefSeq. Esta base de datos era considerablemente grande y medía unos 8 gigabytes de tamaño. Para eliminar la fuente primaria de falsos positivos como las secuencias de baja complejidad en los genomas mismos; por ejemplo, una cadena de 31 o más A's consecutivas, ejecutamos el programa'dust' en todos los genomas y luego construimos la base de datos a partir de estos genomas'dust'. El análisis kraken se ejecutó durante 6 horas en una instancia t2.xlarge EC2 en una nube Amazon AWS con 4 CPUs de alta potencia y 16 GiB de RAM. Al final, alrededor del 27,7% de las lecturas no mapeadas fueron clasificadas. Esto contribuyó al 18,5% de las lecturas totales de la muestra. Se asignaron varias clasificaciones taxonómicas bacterianas y virales y las lecturas no clasificadas se obtuvieron como archivo fastq.

En total, alrededor del 33,7% de las lecturas estaban alineadas con el genoma humano, el 18,4% de las lecturas mapeadas con genomas bacterianos y el 47,9% restante no estaban clasificadas. El mismo procedimiento se siguió para la Mano Grande, con diferentes porcentajes.