Nota: La versión actualizada de este reporte puede consultarse en:

https://docs.google.com/document/d/1SSrKX-TibnrvCa55woImiPvmO3OUps3HYFVnWTaT9Bw/edit?usp=sharing

Índice

Índice **1**

Introducción **2**

Mapeo al genoma humano (versión GRCH37) con parámetros relajados del set completo de lecturas de las dos muestras de Victoria . **3**

Mapeo al genoma más completo de Vicuña disponible en NCBI de las lecturas totales de Victoria . **5**

Análisis iterativo de las proporciones de lecturas unidas por sobrelape sin coincidencias significativas a una selección de genomas de la base de datos refseq de secuencias de ADN de NCBI. **6**

Análisis por “taxmaps” de las proporciones de lecturas totales únicas sin coincidencias significativas a las taxonomías y secuencias de DNA en las bases de datos refseq y nt de secuencias de ADN de NCBI. **9**

Búsqueda de presencia de cebadores (primers) universales de identificación de vertebrados en los datos de secuenciación. **11**

Busqueda de capacidad codificante en las lecturas sobrelapadas. **11**

Resumen cuantitativo de las proporciones de lecturas observadas en los principales análisis. **12**

Introducción

El presente reporte detalla los resultados obtenidos de múltiples análisis bioinformáticos que expanden los análisis y reportes previos hechos independientemente por parte de otros grupos en México, USA y Rusia a los datos de secuenciación masiva obtenidos del ADN de los cuerpos tridáctilos de Nasca. En particular este reporte expande sobre los análisis hechos a los datos de secuenciación del laboratorio CEN4GEN (https://cen4gen.org/) .

Todos los análisis fueron hechos con software de código abierto desarrollado por grupos líderes en bioinformática y genómica y bajo solicitud y consentimiento para usar los datos del equipo de Jaime Maussan. Las muestras analizadas para este reporte fueron las obtenidas del cuerpo llamado Victoria etiquetadas como Ancient002 y Ancient004 en los datos de CEN4GEN que provienen de la amplificación por MDA del DNA obtenido por CEN4GEN tanto del tejido del cuello (Ancient002) como de DNA extraído del cuerpo previamente por Biotecmol (Ancient004), respectivamente.

Los análisis primarios hechos para este reporte fueron los siguientes:

1. ***Mapeo con parámetros relajados del set completo de lecturas originales de las dos muestras de Victoria al genoma humano (versión hg19).***
2. ***Mapeo al genoma más completo de Vicuña disponible en NCBI de las lecturas totales de Victoria.***
3. ***Análisis iterativo de las proporciones de lecturas unidas por sobrelape sin coincidencias significativas a una selección de genomas de la base de datos refseq de secuencias de ADN de NCBI.***
4. ***Análisis por "taxmaps" de las proporciones de lecturas únicas sin coincidencias significativas a las bases de datos refseq y nt de secuencias de ADN de NCBI.***
5. ***Búsqueda de presencia de cebadores (primers) universales de identificación de vertebrados en los datos de secuenciación.***
6. ***Busqueda de capacidad codificante en las lecturas sobrelapadas.***
7. ***Resumen cuantitativo de las proporciones de lecturas observadas en los principales análisis.***

En las posteriores secciones se detallan cada uno de los análisis realizados.

Mapeo al genoma humano (versión GRCH37) con parámetros relajados del set completo de lecturas de las dos muestras de Victoria .

Se hizo un nuevo mapeo al genoma humano usando el software bwa (Heng Li, 2013) de todas las lecturas completas secuenciadas por CEN4GEN con parámetros más laxos, es decir sin descartar mapeos por bajas calidades de mapeo, para evaluar un margen más amplio en el que las muestras de Victoria podrían dar coincidencia a DNA humano. Los resultados se muestran en la siguiente tabla para las lecturas obtenidas totales para cada muestra:

Métrica	Ancient 002 con mapeo	Ancient 002 Porcentajes	Ancient 004 con mapeo	Ancient 004 Porcentajes
Lecturas completas iniciales	**1,123,330,640**	**100**	**1,003,400,490**	**100**
Lecturas con mapeo permisivo a humano	**654,603,960**	**58.27**	**481,988,213**	**48.04**

Tabla1: Lecturas totales mapeadas al genoma humano usando bwa sin restricciones de calidad de mapeo y parámetros predeterminados.

Esto muestra que aún dando un margen de mapeo más amplio y laxo no se logra mapear el total de las lecturas al genoma humano y en Ancient004 es menor al 50% de mapeo. No obstante se debe hacer notar que el aumento en el margen de lecturas mapeadas está dominado significativamente por mapeos de mala calidad como muestra la siguiente gráfica de distribución de calidades en los mapeos de los datos correspondiente a los mapeos bajo este esquema de la muestra Ancient002:

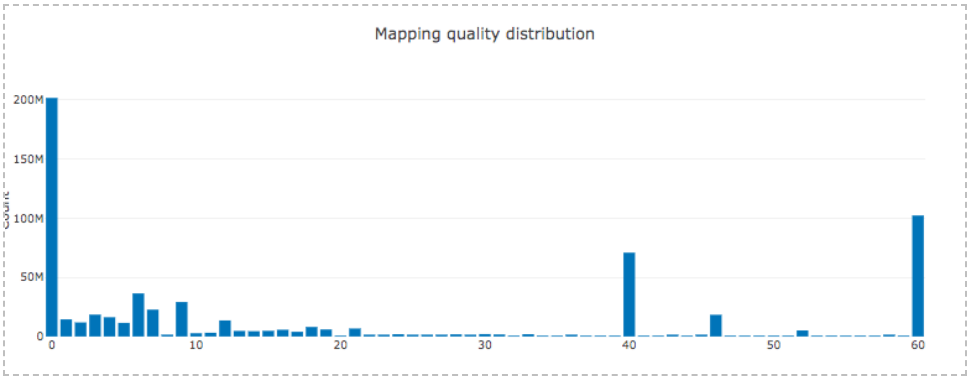

Figura 1: Distribución de calidades de mapeo de lecturas totales de Ancient002 en el resultado de bwa sin restricción de calidades usando Alfred con parámetros predeterminados.

Adicionalmente para confirmar los niveles de mapeo a genoma humano entre las muestras de Victoria y el genoma humano GRCH37 se hizo un remapeo de las lecturas totales de cada muestra usando el software Kart v2.4.5 (Hsin-Nan Lin & Wen-Lian Hsu,2017) para poder tener un resultado más rápido y muy cercano al que brinda bwa pero con una mayor especificidad , incluyendo una mayor restricción en la calidad de mapeo.

Los resultados fueron los siguientes:

Métrica	Ancient 002 con mapeo	Ancient 002 Porcentajes	Ancient 004 con mapeo	Ancient 004 Porcentajes
Lecturas completas iniciales	**1,123,330,640**	**100**	**1,003,400,490**	**100**
Lecturas con mapeo permisivo a humano	**115,329,515**	**10.27**	**88,285,395**	**8.80**

Tabla 2: Lecturas totales mapeadas al genoma GRCH37 de las muestras de Victoria usando el software Kart con los parámetros predeterminados.

Esto confirma los reportes previos de una baja coincidencia con el genoma humano de los resultados de secuenciación de las muestras de Victoria al considerar mapeos con filtros más exigentes de calidad, como muestra la siguiente gráfica de distribución de calidades en los mapeos de los datos correspondiente a este esquema de mapeo con mayor nivel de restricciones de calidad para la muestra Ancient002:

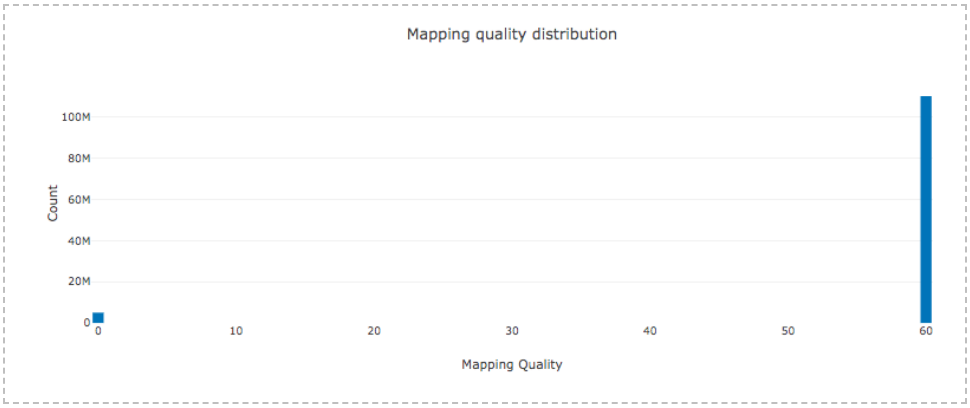

Figura 2: Distribución de calidades de mapeo de lecturas totales de Ancient002 en el resultado de Kart sin restricción de calidades y parámetros predeterminados.

.

Los análisis de la distribución de calidades mostrados de mapeos se realizaron con el software Alfred v0.1.17 (Tobias Rausch, Markus Hsi-Yang Fritz, Jan O Korbel, Vladimir Benes,2018) sobre los archivos bam ordenados e indexados de cada mapeo.

Mapeo al genoma más completo de Vicuña disponible en NCBI de las lecturas totales de Victoria .

Para evaluar la posibilidad de que las muestras de Victoria provinieran de partes de cuerpo de Vicuña se evaluó el mapeo de las lecturas totales secuenciadas para las dos muestras de Victoria contra el ensamble de genoma más completo y más cercano a la Vicuña que se tuvo disponible en la base de datos de refseq de NCBI, que fue el de *Vicugna pacos* o alpaca (Ensamble GCF_000164845.2). Para este análisis se usó el software Kart v2.4.5 (Hsin-Nan Lin & Wen-Lian Hsu,2017) para poder tener un resultado más rápido y muy cercano al que brinda bwa.

Los resultados se muestran en la siguiente tabla:

Métrica	Ancient 002 con mapeo	Ancient 002 Porcentajes	Ancient 004 con mapeo	Ancient 004 Porcentajes
Lecturas completas iniciales	**1,123,330,640**	**100**	**1,003,400,490**	**100**
Lecturas con mapeo a vicuña	**8734233**	**0.78**	**22715081**	**2.26**

Tabla 3: Lecturas totales mapeadas al ensamble de Vicuña GCF_000164845.2 de las muestras de Victoria usando el software Kart con los parámetros predeterminados.

Esto muestra que los porcentajes de coincidencia con el genoma de la Vicuña serían mínimos para las muestras analizadas de Victoria y no parecen ser significativos.

Adicionalmente para confirmar estos niveles de coincidencia entre las lecturas y el genoma de la vicuña se hizo un análisis por coincidencias de kmeros usando el software “Seal” de la paquetería de BBtools en su versión 38-25 (Bushnell B. sourceforge.net/projects/bbmap/), desarrollada por el Joint Genome Institute del departamento de energía de USA, para comparar cuántas subsecuencias (kmeros) de tamaño 31 hay entre el genoma de la Vicuña y las lecturas completas de los datos de secuenciación de Victoria. Los resultados obtenidos fueron los siguientes:

Métrica	Ancient 002 con mapeo	Ancient 002 Porcentajes	Ancient 004 con mapeo	Ancient 004 Porcentajes
Lecturas completas iniciales	**1,123,330,640**	**100**	**1,003,400,490**	**100**
Lecturas con coincidencias a kmeros de vicuña	**36,470,068**	**3.25%**	**69,633,978**	**6.94%**

Tabla 4: Lecturas totales con coincidencias en al menos un kmero con el ensamble de Vicuña GCF_000164845.2 de las muestras de Victoria usando el software seal con los parámetros forbidn=t, ambiguous=first ,rskip=1 y los demás parámetros predeterminados.

Análisis iterativo de las proporciones de lecturas unidas por sobrelape sin coincidencias significativas a una selección de genomas de la base de datos refseq de secuencias de ADN de NCBI.

Para explorar a qué tipos de genoma conocidos podrían pertenecer las lecturas secuenciadas sin hacer un mapeo o un BLAST a todos los genomas conocidos de todas las lecturas secuenciadas, ya que esto tardaría mucho tiempo y no sería viable en tiempo y capacidad de cómputo disponible, se procedió primero a extraer aquellas lecturas que se pudieran unir por sobrelape, con el software PEAR v0.9.6 (Zhang, Kobert, Flouri, & Stamatakis, 2014) y con estas lecturas de menor volumen pero mayor longitud se procedió a usar una técnica de comparación por “sketching” y búsqueda de kmeros, con el software “sendsketch” y “seal” de la paquetería BBtools en su versión 38-25. El software “sendsketch” permite buscar similitud entre grandes cantidades de lecturas y la base de datos de refseq de manera mucho más rápida y sencilla que con técnicas de alineamiento como los mapeos o BLAST mientras que “seal” permite buscar si alguna subsecuencia de 31 bases (kmero de 31), cualquiera, está contenida en algún genoma. Usando “sendsketch” se identificaron los genomas y tipos de organismos que pudieran parecerse a las lecturas secuenciadas unidas por sobrelape y una vez identificados los genomas con mayores coincidencias se procedió a descargar un grupo de genomas del tipo que dió mayor número de coincidencias en “sendsketch”. Posteriormente se compararon las lecturas contra dichos genomas para buscar aquellas que tuvieran similitud en al menos un kmero de 31 y aquellas que no tuvieran ningún kmero de 31 se seleccionaron para una nueva ronda de búsqueda con “sendsketch” y comparación con “seal” y así de forma iterativa se hizo para todos los tipos de genomas que dieran mayor coincidencia con “sendsketch” para las reads sin presencia de kmeros de 31 del tipo de genoma detectado en la iteración anterior hasta llegar a los genomas de tipo protozoarios. Para los genomas de microorganismos se descargaron todos los genomas disponibles en refseq mientras que para los genomas de animales se descargaron sólo los genomas indicados por “sendsketch” y otros cuantos más sugeridos por diversos participantes en el análisis.

A continuación se muestra una figura con un diagrama del procedimiento:

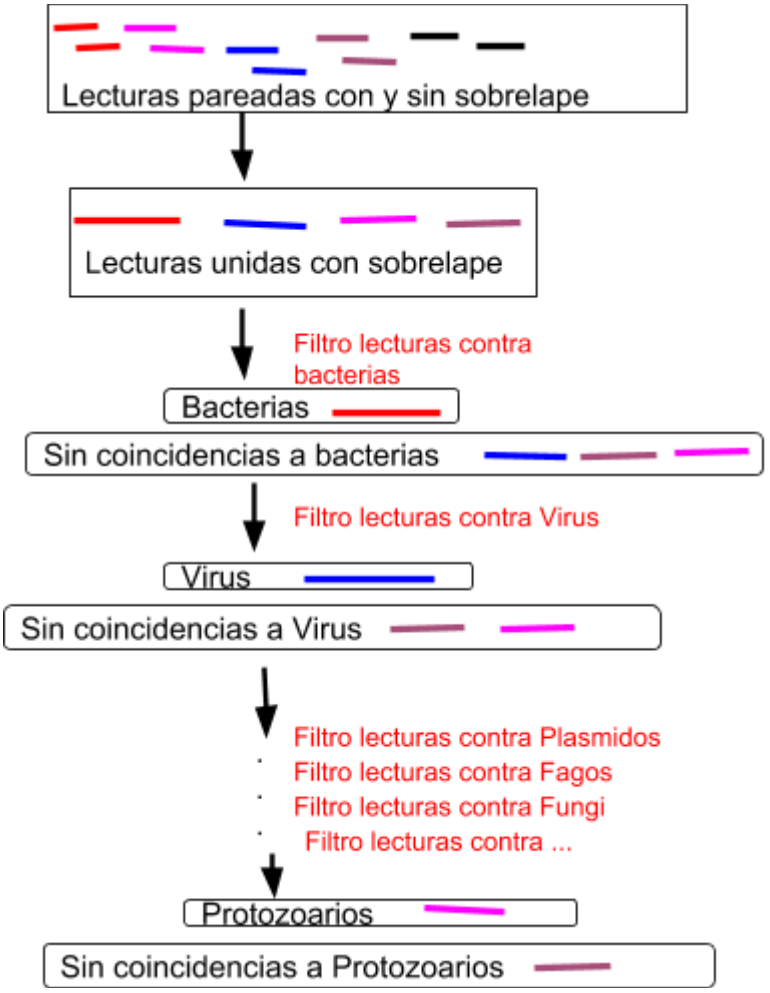

Figura 3: Estrategia de filtro Iterativo desarrollada para encontrar similitudes entre las lecturas únicas de unión por sobrelape y 37,877 de múltiples tipos de genomas de refseq.

Siguiendo esta estrategia se compararon las lecturas sobrelapadas de cada muestra contra 37,877 secuencias de genomas en total y los resultados están mostrados a continuación con una tabla con los números y porcentajes de lecturas con similitud significativa, de acuerdo al software “seal”, resultantes de esta comparación usando la técnica iterativa de búsqueda con “sendsketch” seguida por la comparación con “seal”.

La tabla muestra los tipos de genomas en el orden en el que se fueron comparando contra las lecturas unidas por sobrelape.

Lecturas de cada etapa de la comparación	Ancient 002	%	Ancient 004	%
Lecturas completas	**1,123,330,640**	**100**	**1,003,400,490**	**100**
Lecturas unidas por sobrelape (Total a considerar para los siguientes porcentajes)	**103,600,552**	**18.45**	**338,014,892**	**67.37**
Lecturas unidas con coincidencia a bacteria	**6,075,940**	**5.86**	**26,840,573**	**7.94**
Lecturas unidas con coincidencia a virus	**241,181**	**0.23**	**2,734,450**	**0.81**
Lecturas unidas con coincidencia a plasmidos	**60,695**	**0.06**	**309,502**	**0.09**
Lecturas unidas con coincidencia a Fagos	**954,886**	**0.92**	**652**	**0.00**
Lecturas unidas con coincidencia a Fungi	**707,072**	**0.68**	**6,857,348**	**2.03**
Lecturas unidas con coincidencia a plastidos	**6,298,041**	**6.08**	**11,405**	**0.00**
Lecturas unidas con coincidencia a diatomeas	**25,086**	**0.02**	**295,137**	**0.09**
Lecturas unidas con coincidencia a Humano	**6,612,714**	**6.38**	**35,325,290**	**10.45**
Lecturas unidas con coincidencia a Bos Taurus	**105,784**	**0.10**	**44,620,181**	**13.20**
Lecturas unidas con coincidencia a H. penzbergensis	**139**	**0.00**	**31,208,623**	**9.23**
Lecturas unidas con coincidencia a P. vulgaris	**53,527,894**	**51.67**	**25,633**	**0.01**
Lecturas unidas con coincidencia a genomas diversos	**882,181**	**0.85**	**NA**	**NA**
Lecturas unidas con coincidencia a genomas de otros vertebrados	**130,249**	**0.13**	**2245614**	**0.66**
Lecturas unidas con coincidencia a Protozoarios	**4,169**	**0.00**	**7768**	**0.00**
Lecturas unidas con coincidencia totales	**75,626,031**	**73.00**	**150,482,176**	**44.52**
Lecturas unidas sin coincidencia totales	**27,974,521**	**27.00**	**187,532,716**	**55.48**

Tabla 5: Números y porcentajes de lecturas con similitud significativa en al menos un kmero de 31 bases devueltos por el software “seal” contra una selección de genomas de refseq.

Nota: La lista completa de secuencias que se usaron para cada tipo de organismo se puede encontrar en la liga siguiente:

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

Esto muestra que aún pasando por un filtro que incluye a un gran umbral de genomas y tipos de organismos conocidos no se logra encontrar el 100% del orígen genómico de las lecturas con sobrelape obtenidas de las muestras de Victoria. En particular el 27.00% y el 55.48 % aproximadamente de lecturas con sobrelape de las muestras

Ancient002 y Ancient004 , respectivamente, no se logra asociar a ningún tipo de organismos de todos los tipos de organismos y especies acumulados en este análisis.

Análisis por “taxmaps” de las proporciones de lecturas totales únicas sin coincidencias significativas a las taxonomías y secuencias de DNA en las bases de datos refseq y nt de secuencias de ADN de NCBI.

Para evaluar en un margen lo más amplio posible la coincidencia por similitud entre las lecturas secuenciadas para Victoria contra todas las secuencias de DNA dadas de alta en NCBI y sus respectivos niveles taxonómicos posibles se tomaron las lecturas sin duplicados de cada muestra de Victoria, es decir de las lecturas totales se quitaron aquellas lecturas redundantes que tenían alta similitud con alguna otra, de manera que se tuviera un set de lecturas sin redundancia entre ellas. Esto se hizo por separado para las lecturas “forward” y las lecturas “reverse” de cada set de lecturas pareadas. De este set de lecturas sin duplicados se usó el correspondiente a las para buscar la proporción del ADN secuenciado con coincidencia por similitud a la base de datos más completa disponible a la fecha que es la conformada por las bases de datos nt y refseq juntas de NCBI. Esta búsqueda se realizó al generar una búsqueda contra todas las secuencias registradas en las base de datos “nt” y “Refseq” de NCBI en su forma implementada en el índice del software taxmaps v 0.2.1 (Corvelo, Clarke, Robine & Zody. 2018). Este software se usó para buscar contra el índice anteriormente mencionado con las lecturas “Forward” del set de datos sin duplicados de cada muestra y con los parámetros de -e 0.2 y -m s y el resto de los parámetros default para descartar lecturas de baja complejidad o con presencia de bases no determinadas. Los resultados son los siguientes:

Figura 4: Diagramas de distribución de proporciones de las lecturas con similitud significativa a la base de datos nt y refseq de NCBI y sin ella (color rojo intenso) para cada muestra.

En la porción de color rojo intenso (Sección Unclassified) de cada figura se muestra el porcentaje de lecturas únicas sin similitud significativa a las bases de datos Refseq y

“nt “que habría en los datos de las lecturas de secuenciación de Victoria. **Esto demuestra que hay más del 50% de lecturas de secuenciación, no redundantes y sin presencia de bases no determinadas o bajos niveles de complejidad, que no se parecen a nada en las bases de datos juntas de “nt” y “refseq” del NCBI**. También este resultado confirma los altos niveles de porcentajes de secuencias no clasificadas obtenidos en análisis previos hechos en “Abraxas” sobre las lecturas sobrelapadas.

Búsqueda de presencia de cebadores (primers) universales de identificación de vertebrados en los datos de secuenciación.

Para evaluar la posibilidad de identificar especies concretas de vertebrados presentes en las muestras secuenciadas que pudieran determinarse de forma específica mediante los datos de secuenciación se buscó en las lecturas completas de cada muestra la presencia de cebadores (primers) universales para la detección de especies de vertebrados. Estos primers se tomaron de la publicación “Two universal primer sets for species identification among vertebrates” (Kitano T, Umetsu K, Tian W, Osawa M. 2007. https://www.ncbi.nlm.nih.gov/pubmed/16845543). La búsqueda de la presencia de estas secuencias universales se hizo con el software seal de la paquetería de bbtools (Bushnell B. sourceforge.net/projects/bbmap/) con los parámetros ambiguous=first , rskip=1, K=21, y minkmerhits=2 y usando como input las lecturas pareadas de cada muestra de Victoria.

Los resultados obtenidos devolvieron 72 lecturas para Ancient 004 y 80 lecturas para Ancient002 con coincidencias para la presencia de las secuencias universales para vertebrados. No obstante al hacer una búsqueda por BLAST contra la base de datos nt filtrada sólo para Vertebrados (taxid:7742) de estas lecturas todas ellas corresponden a humano únicamente. No se detecta la presencia de otras especies de vertebrados por este análisis.

Busqueda de capacidad codificante en las lecturas sobrelapadas.

Debido a la ausencia de secuencias con coincidencias significativas por similitud en las bases de datos conocidas de DNA para una gran parte de las lecturas secuenciadas se evaluó si dichas lecturas pudieran contener secuencias codificantes y por ende formar parte de nuevos genes desconocidos en las actuales bases de datos. Para evaluar esto en una forma independiente a la comparación por similitud de secuencias se procedió a tomar el set obtenido previamente de lecturas unidas por sobrelape que no dió

coincidencia a ninguna de las 37,877 referencias en el filtro iterativo de cada muestra y de este set se eliminaron las secuencias redundantes usando el software “seal” tal como se hizo para el software taxmaps con las lecturas totales. Este set de lecturas unidas por sobrelape sin duplicados se usó como entrada para el software FragGeneScan (Mina Rho, Haixu Tang, and Yuzhen Ye. 2010), el cual permite estimar mediante modelos ocultos de Markov que toman en cuenta la posibilidad de errores en las lecturas y el uso de codones para estimar la presencia de fragmentos codificantes de genes directamente de lecturas de secuenciación masiva incluso con lecturas que tengan una tasa de error considerable.

Los resultados obtenidos de FragGeneScan se procesaron mediante cálculos desarrollados en código de R para obtener el potencial codificante del set de lecturas, que definimos como el cociente de lecturas con regiones codificantes sin interrupciones, señalados como “*” por FragGeneScan, entre el total de lecturas únicas.

Dependiendo del umbral del tamaño de la región codificante se obtuvieron distintos potenciales codificantes posibles. Tomando un umbral de al menos 150 bases en la región codificante, o 50 aminoácidos, sin que esta abarque toda la lectura se tiene un potencial codificante de 23.66% para Ancient002 y 56.99% para Ancient004. No obstante para ser más estrictos en la posibilidad de que la lectura forme parte de un gen conviene tomar un umbral que abarque el tamaño total de la lectura y bajo este umbral y para lecturas de al menos 150 bases del set lecturas unidas por sobrelape se obtuvo un menor potencial codificante para ambas muestras pero aún considerable, de 5.71% para Ancient002 y 18.13% para Ancient004.

Este análisis muestra un considerable potencial codificante del ADN para las lecturas sobrelapadas no duplicadas de ambos sets de lecturas de las muestras secuenciadas, en particular para las lecturas sobrelapadas sin duplicados de Ancient002 se observó un porcentaje de alrededor de 6% y 25.7% de lecturas con capacidad codificante y para Ancient004 de 18.15% y 58.03%.

Resumen cuantitativo de las proporciones de lecturas observadas en los principales análisis.

Finalmente las proporciones observadas para los las lecturas incluidas en los resultados previamente descritos para cada muestra se compilaron en la siguiente figura para demostrar la alta cantidad de lecturas analizadas y la alta proporción de lecturas que caen en las categorías relevantes de cada análisis como las lecturas

duplicadas sin coincidencias en las bases de datos de NCBI por taxmaps así como por bbtools en la base de datos de Refseq o la cantidad de lecturas totales resultantes de la unión por sobrelapeña.

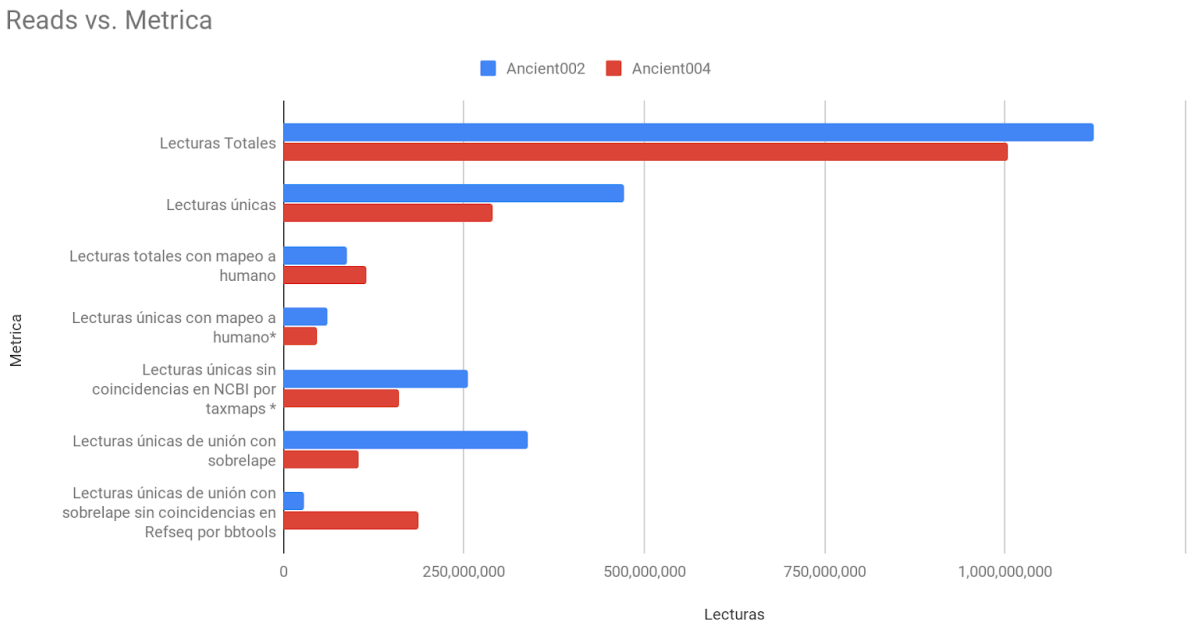

Figura 5: Cantidades observadas para las lecturas presentes en los resultados de los análisis descritos en este reporte para cada muestra. Las lecturas únicas mapeadas a humano se obtuvieron mapeando con Kart y sus parámetros predeterminados las lecturas únicas. Las lecturas únicas sin coincidencias en NCBI por taxmaps se estimaron multiplicando la proporción mostrada por taxmaps en las figuras mostradas por el total de las lecturas únicas de cada muestra.

Realizado por:
Lic en Ciencias Genómicas.
Salvador Ángel Romero Martínez.
09-Abril-2019.