*¿En qué medida cambia el nivel de expresión génica de los tejidos preservados de un cánido fósil (lobo del pleistoceno) respecto a un cánido actual (perro)?¿cómo estas diferencias de expresión reflejan posibles cambios adaptativos en respuesta al ambiente del pasado?*

Esta propuesta de investigación considera como control al cánido moderno y como “tratamiento” al fósil del pleistoceno. A continuación se presenta el flujo de trabajo:

1. **Obtención de datos de RNA-seq.**
   1. Se emplearán los datos pertenecientes a las muestras del tejido hepático, esto debido a que los autores indican que fue este tejido el que mostró la mayor abundancia de transcritos y el mayor grado de similitud al compararlo con los datos modernos. Adicionalmentevarios de los genes altamente expresados ​​en el tejido hepático antiguo están asociados con la función hepática.
   2. Para tener claridad sobre el detalle de cada muestra, se realiza una búsqueda en NCBI Run Selector.
   3. Importación de datos a Galaxy: Se utilizará la herramienta EBI SRA ([European Nucleotide Archive](https://www.ebi.ac.uk/ena/data/search?GALAXY_URL=https%3A//usegalaxy.org/tool_runner%3Ftool_id%3Debi_sra_main)) para encontrar los códigos anteriormente descritos bajo la siguiente codificación:
      1. Control: PRJNA396033 {SRX3055179 (cartílago), SRX3055151 (hígado), SRX3055143 (piel) y SRX3055142 (músculos)}
      2. Tratamiento: PRJNA497993
2. **Verificación de calidad de los datos de RNA-seq.** 
   1. Herramienta **FastQC (Galaxy)**: Este paso es esencial pues permitirá evaluar la calidad de las bibliotecas de secuenciación y diagnosticar posibles problemas con la extracción de RNA, la construcción de la biblioteca o la secuenciación.
   2. Limpieza de muestras: En dependencia de los resultados, eliminar muestras de baja calidad y/o no relevantes al estudio. De ser necesario, se utilizará la herramienta **Cutadapt(Galaxy)** para la remoción de los adaptadores de secuenciación.
3. **Alineamiento y mapeo de secuencias de cada lectura con respecto al transcriptoma de referencia (CamFam3.1– perro moderno).**
   1. Herramienta **Bowtie2/Hisat2(Galaxy):** Bajo parámetros predeterminados para datos de un solo extremo.
   2. Herramienta **SAM-to-BAM** **(Galaxy)**: Con los BAM obtenidos, se realiza un control de calidad a través de un filtro por puntuación de calidad (q≥ 20).
4. **Identificación de genes de origen de los transcritos.**
   1. Herramienta **BLAST+ (NCBI)**: Búsqueda en la base de datos con parámetros predeterminados.
5. **Cuantificación de las lecturas asignadas a cada transcrito individuales** 
   1. Herramienta **featureCounts(Galaxy):** Consolida todas las lecturas en una tabla de conteo
6. **Identificación de genes específicos cuyos transcritos se expresan diferencialmente entre muestras.**
   1. Comparación directa de los niveles de expresión entre el control y las muestras antiguas, se mantendrán aquellos que se encuentren por encima del percentil 95 de los niveles de expresión.
7. **Interpretación biológica de genes expresados diferencialmente.**
   1. Análisis de vías: Enriquecimiento de categorías funcionales de genes o superposición con conjuntos de genes. Herramienta **GeneOntolgy (**[**http://geneontology.org**](http://geneontology.org)**).**

**PREGUNTAS:**

* **Influencia de edad en los cánidos.**

Una vez obtenidos los resultados de expresión diferencial de genes, se tendrán en cuenta sólo aquellos que corresponden a genes que no cambian su expresión a lo largo de su ontogenia. Es por ello que en el caso de observar una diferencia en la expresión de genes que se sabe que solo cambian su expresión en estados juveniles, estos no serán tenidos en cuenta.

* + **El transcriptoma es dinámico >> Estrategia de normalización.**

Desde el punto de vista metodológico, las herramientas de normalización previa al ensamblaje y mapeo (como RPKM, FPKM y TPM) podrían ser útiles al momento de solucionar este punto; sin embargo, por la naturaleza de los datos que se van a usar para el cánido fósil, no es posible aplicar un análisis que considere tanto profundidad de la lectura como longitud del gen. Para este último no habría limitaciones, pero respecto a la profundidad de lectura hay que considerar que por ser material fósil el número de lecturas no es muy alto y la profundidad es difícil de determinar. A pesar de ello, el tejido hepático reporta una de las coberturas más altas del estudio (50%) lo que da la confianza suficiente para realizar este estudio aún sin normalizar.

* **Limitaciones/ventajas de usar una referencia actual.**

|  |  |
| --- | --- |
| **Limitaciones** | **Ventajas** |
| Es posible que se presente gaps en las lecturas, pues a pesar de que estamos empleando dos organismos del mismo género, presentan una distancia temporal de ~15.000 y cabe la posibilidad de que se hayan dado pequeñas variaciones en el genoma de cánido del pasado con respecto al del actual. | Corresponde al genoma de un organismo modelo, con un alto nivel de anotación y presentan una alta cercanía filogenética. |
| La degradación del material fósil lleva a la formación de secuencias cortas que, pueden generar variaciones de tipo no biológico con respecto al genoma de referencia. | Las muestras de los tejidos del pasado muestran una alta especificidad con respecto a las del genoma de referencia (actual) |
| La presencia de material genético contaminante (microbiano) que lleve a variaciones en las secuencias del cánido del pasado con respecto el actual | Las muestras del tejido del pasado, muestran un alto porcentaje de cobertura con respecto al genoma de referencia contribuyendo a mejor el poder estadístico del análisis. |

* **¿Por qué BLAST antes de cuantificar? ¿Qué información adicional esperamos obtener?**

Llevar a cabo un blast antes de proceder a la cuantificación, permitirá obtener información acerca de qué genes están expresándose, obtener id’s y con ello se podrá seleccionar o descartar aquellos que cumplan o no con el objetivo del análisis. Además, teniendo en cuenta algunas de las limitaciones presentadas anteriormente, para analizar fuentes potenciales de lecturas no asignadas se puede inspeccionar manualmente algunas secuencias de lectura seleccionadas al azar y ejecutar un BLAST con las lecturas no asignadas como consultas contra una base de datos de nucleótidos no redundante (NR) de NCBI. Esto nos indicará si las muestras han sido contaminadas por DNA/RNA de otras especies (por ejemplo, bacterias) o por artefactos de la construcción de la biblioteca.

Adicionalmente, esta herramienta nos podría dar una base de significancia estadística a partir del valor E proporcionado. En este sentido, si se encuentra que algún gen tiene un valor E bajo puede descartarse antes de tenerlo en cuenta en el conteo porque su efecto está dado por el azar y no por un fenómeno biológico. También podemos aplicar un filtro adicional de calidad asociado al score y centrarnos en valores altos que nos indiquen mayor similitud durante el alineamiento.

* **Efecto de Batch (lote) > Variación técnica genera cambios de expresión.**

El efecto Batch se da en datos procedentes de distinto laboratorio, estudio o tratados con distinto protocolo; los efectos sistemáticos introducidos sobre los datos hacen que difieran y la normalización no seaa capaz de compensarlos completamente. Frente a esto, la plataforma de NCBI Run Selector es clave para verificar el detalle de cada uno de los experimentos; las muestras que serán tenidas en cuenta para este estudio corresponden al mismo género, método (Illumina HiSeq 2500) y protocolo (Dr. Andrew M. Hoffman, Cummings School of Veterinary Medicine, Tufts University). Así, el único factor que no se puede controlar es la edad del organismo y ya se discutió anteriormente.