Proyecto Integrador Bioinformática – Entrega 2

Clasificación taxonómica

Al realizar el mapeo, algunos *reads* no se alinearon con la referencia. Estos son de interés debido a su posible relación con cambios genéticos evolutivos o la presencia de ADN exógeno. Por consiguiente, a los *reads* no mapeados se les realizó una clasificación taxonómica con Kraken2 y una cuantificación con Bracken, con el fin de identificar qué organismos o secuencias se encontraban representadas y detectar posibles virus.

De acuerdo con las estadísticas de alineamiento, los porcentajes de *reads* no mapeados fueron los siguientes:

- Evolución 1: 99,92 % mapeados, 0,08 % no mapeados.
- Evolución 2: 97,78 % mapeados, 2,22 % no mapeados.

Kraken2 es un software rápido y confiable que clasifica lecturas por coincidencia exacta de *k-mers*, asignándolas al ancestro común más bajo entre los genomas de la base de datos, lo que permite una clasificación precisa. Bracken, a partir de los reportes de Kraken2, ajusta las abundancias corrigiendo sesgos por longitud del genoma y sobreasignaciones. Este ajuste resulta útil cuando especies muy similares coexisten en una misma muestra o fragmento. No considera los *reads unclassified*, ya que solo analiza las lecturas con clasificación válida. (Lu J, 2017).

En primera instancia al correr el código y comparar los datos de kraken2 y krona se observó que los datos no coincidían pues los porcentajes y el orden taxonómico que mantenía kraken2 sufría modificaciones al visualizarse en krona. Asimismo, al intentar leer el archivo de texto plano generado por krona aparecía el siguiente mensaje: "[WARNING] The following taxonomy IDs were not found in the local database and were set to root". Indicando que el software no reconocía algunos IDs taxonomicos que Kraken2 sí tenía. En pocas palabras las bases de datos con las que estaban trabajando los softwares no eran las mismas. Por lo tanto, fue necesario implementar unos comandos que resolvieran el problema.

```
http://ktUpdateTaxonomy.sh
cut -f2,3 kraken_output_evol1.3.txt > kraken_krona_input.txt
cut -f2,3 kraken_output_evol2.3.txt > kraken_krona_input2.3.txt
ktImportTaxonomy kraken_krona_input1.3.txt -o krona_evol1.3.html
ktImportTaxonomy kraken_krona_input2.3.txt -o krona_evol2.3.html
```

Con estos comandos se actualizo la base de datos interna de krona, también se extrajeron solamente dos columnas del archivo de texto de kraken en este caso la 2 y la 3 que son la columna del taxID del organismo y el nombre del organismo. Esto debido a que krona solo necesita esas columnas para realizar el archivo interactivo. Por último, se crean manualmente los reportes de krona.

Análisis

Evolución 1:

El análisis de Kraken2 determinó principalmente la presencia de organismos celulares, que abarcaron un 79.11% de la información analizada. En su mayoría se encontró al reino de las bacterias con el mismo valor de porcentaje. A su vez, se encontró un porcentaje pequeño de virus. Por otro lado, se encontró un 20.55% que no se clasificó.

Dentro del reino de las bacterias, se encontró un alto porcentaje de proteobacterias con un 77.05% y donde la familia más representativa fue la de Enterobacteriaceae, de aquí se desprendieron 3 géneros Escherichia, Salmonella y Citrobacter. En cada una de estas familias fue posible clasificar la especie y la subespecie indicando que los reads no mapeados conservan suficiente información genómica para permitir una asignación precisa. Al observar la presencia de Escherichia Coli se genera un poco de confusión dado que la clasificación proviene de los reads no mapeados de una muestra de la misma especie y debió haberse mapeado. Sin embargo, estudios han demostrado que sólo aproximadamente el 20 % de los genes de una cepa de E. coli corresponde a los genes que están presentes en todas las cepas de la especie, mientras que el resto constituye un conjunto de genes accesorios que varían ampliamente entre linajes (Nykrynova et al., 2022).

También se encontraron asignaciones al género Staphylococcus y Streptococcus además de la presencia de la especie de bacteria Cutibacterium acnes la cual está presente principalmente en la piel humana lo que podria indicar una contaminación, sin embargo, su porcentaje es tan bajo que no representa una contaminación considerable.

Por otro lado, el porcentaje de virus encontrado hace referencia a la especie Escherichia virus T1, un fago que infecta a bacterias del género Escherichia coli, lo cual tiene sentido y no se considera contaminación dado que se trataba de un cultivo de esa bacteria y el porcentaje es muy bajo dentro de la clasificación taxonómica y teniendo en cuenta que esa clasificación es a partir de los reads no mapeados su porcentaje decrece a un más del total de la muestra.

Evolución 2:

Se clasificó taxonómicamente el 98,79 % de las secuencias mediante Kraken2, lo que indica una alta proporción de lecturas con asignación confiable. De estas, el 95,53 % corresponde a virus, siendo el género T1virus el más representativo.

Dentro de los virus identificados, el más abundante fue Escherichia virus T1 con 14198 fragmentos asignados, un bacteriófago lítico perteneciente al género T1virus de la familia Drexlerviridae, conocido por infectar cepas de E. coli mediante la unión a receptores de la membrana externa. Su presencia sugiere la existencia de secuencias fágicas o restos genómicos derivados de una infección pasada. Por otro lado, Escherichia virus ADB-2 también pertenece al grupo de los fagos específicos de E. coli, aunque se encuentra menos representado en la muestra (99 fragmentos). Ambos virus son de tipo ADN bicatenario y participan en la dinámica evolutiva bacteriana, promoviendo procesos como transferencia horizontal de genes o resistencia frente a fagos competidores.

Este resultado es coherente con el origen de la muestra, ya que la bacteria de partida corresponde a una cepa de Escherichia coli. La presencia de fagos específicos de E. coli sugiere una posible infección viral histórica o la integración de secuencias fágicas en el genoma bacteriano, lo cual podría conferir ventajas adaptativas a la cepa, como resistencia frente a otros virus o mecanismos de defensa antivirales.

Por otro lado, un 3,26 % de las lecturas fue clasificado como cellular organisms, de las cuales el 3,19 % pertenece al filo Proteobacteria. Los géneros y especies restantes presentan porcentajes muy bajos (~0,01–0,04 %), lo que refleja una baja diversidad microbiana en la muestra y refuerza la predominancia de E. coli y sus fagos asociados.

El pequeño porcentaje no clasificado (1,21 %) podría corresponder a secuencias de baja calidad, regiones altamente conservadas o fragmentos de organismos no representados en la base de datos empleada. Los resultados visualizados con Krona confirmaron la dominancia del género T1virus, aunque se observaron ligeras variaciones respecto al reporte directo de Kraken2, posiblemente asociadas a los criterios de agrupamiento jerárquico del visualizador.

Imagen 2. Reporte de krona de los reads no mapeados de Evol1

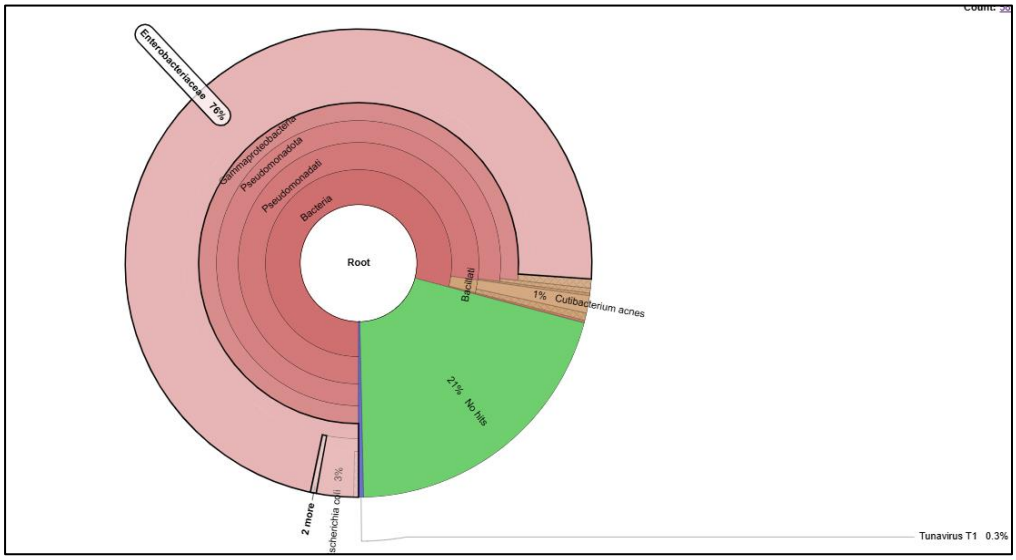
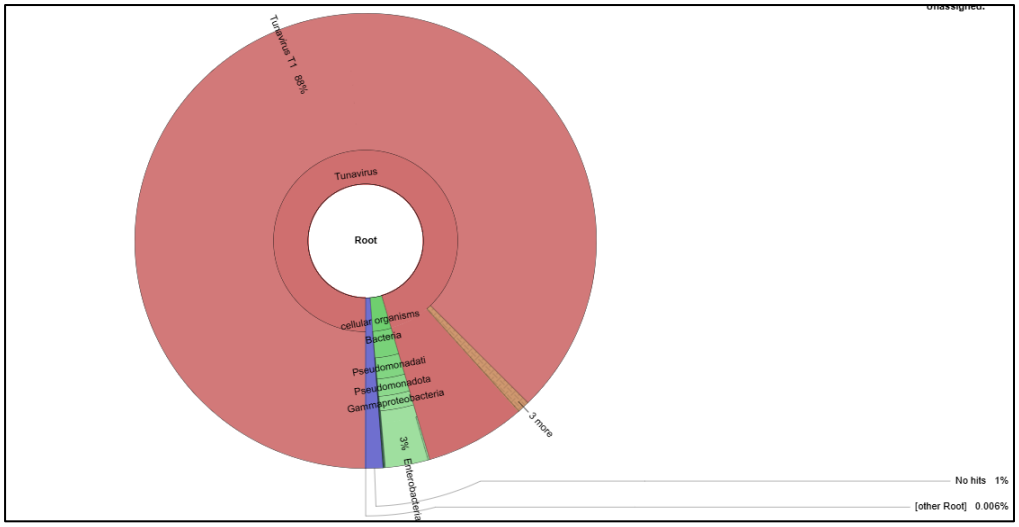


Imagen 3. Reporte de krona de la clasificación taxonómica de los reads no mapeados de Evol 2



Teniendo en cuenta los reportes de kraken2 y de krona es posible identificar diferencias entre las muestras de Evol 1 y Evol 2 puesto que en Evol 1 la mayoría de la taxonomía encontrada correspondía al reino de las bacterias sin embargo para la Evol 2 lo más representativo fue la presencia de Virus.

Tabla con variantes seleccionadas

Evolución 1:

Número	GENE	EFFECT	Impact	AA_CHANGE
1	JBCMMFKD_00636	frameshift_variant	HIGH	p.Ala292fs
2	JBCMMFKD_00659	frameshift_variant	HIGH	p.???159fs
3	JBCMMFKD_00839	frameshift_variant	HIGH	p.????73fs
4	JBCMMFKD_01798	frameshift_variant	HIGH	p.Lys513fs
5	JBCMMFKD_02585	frameshift_variant	HIGH	p.Glu58fs
6	JBCMMFKD_02815	stop_lost&splice_region_variant	HIGH	p.Ter279Lysext*?

Evolución 2:

Número	GENE	EFFECT	Impact	AA_CHANGE
1	JBCMMFKD_00636	frameshift_variant	HIGH	p.Ala292fs
2	JBCMMFKD_00659	frameshift_variant	HIGH	p.???159fs

3	JBCMMFKD_00839	frameshift_variant	HIGH	p.????73fs
4	JBCMMFKD_01452	frameshift_variant	HIGH	p.Thr570fs
5	JBCMMFKD_01452	frameshift_variant&stop_gained	HIGH	p.Glu587fs
6	JBCMMFKD_01798	frameshift_variant	HIGH	p.Lys513fs
7	JBCMMFKD_02815	stop_lost&splice_region_variant	HIGH	p.Ter279Lysext*?

Evolución 1: En el análisis de anotaciones generado con **SnpEff** para el genoma evolutivo (Evol1), se identificaron varias variantes clasificadas con **impacto HIGH**, lo cual indica que estas mutaciones probablemente afectan de manera drástica la función de las proteínas codificadas. Según la documentación de SnpEff, las categorías con impacto **HIGH** incluyen principalmente:

- **Frameshift_variant**: inserciones o deleciones que cambian el marco de lectura del gen, alterando todos los codones posteriores y usualmente produciendo proteínas truncadas o no funcionales.
- **Stop_lost**: mutaciones que eliminan un codón de parada, generando una proteína extendida con posibles consecuencias negativas en su función o estabilidad.

En casi todas las variantes detectadas corresponden a **frameshift_variant**, lo cual sugiere que el tipo de mutación predominante en Evol1 afecta directamente la integridad estructural de las proteínas. Estos eventos suelen ser más perjudiciales que las sustituciones puntuales, ya que generan cambios masivos en la secuencia de aminoácidos o incluso pérdida completa de la función proteica.

Estos efectos se evaluaron a través del formato GFF de Prokka, que relaciona los nombres de los genes de la bacteria de referencia con un código de identificación, el cual se muestra en el reporte de variantes. De esta manera, a cada gen con variante se le identificó su función y se comenzó con una interpretación del cambio que sufrió por la variante.

Para el gen mutado número uno, que sufre de un frameshift, se reportó que este está encargado de la producción de histidina quinasa, una enzima de señalización celular sensible a cambios para funciones como la regulación de la presión osmótica. De esta misma manera, el gen mutado número cuatro, que generalmente codifica la enzima reguladora de descomponedores de la maltosa en *E.coli,* probablemente ha quedado inhibido o dañado, lo que puede generar cambios graves en el procesamiento de maltosa como fuente energética en los que no se

produzcan los componentes que la degraden o se sobreproduzcan y se genere un desbalance energético sin importar el gradiente en la célula. Por su lado, el gen mutado cinco, llamado nadR, que se encarga de la regulación de biosintesis del NAD^+ cambió en funcionalidad, lo que continúa con la mutación de la línea energética de la célula. El resto de los genes de "HIGH importance" mencionados en este reporte de esta evolución se clasificaron como "hypothethical proteins", porteínas no estudiadas a fondo de las que todavía no se conoce muy bien su función y como pudo haber cambiado.

Evolución 2: En el análisis de anotaciones generado con **SnpEff** para el genoma evolutivo **Evol2**, se identificaron varias variantes clasificadas con **impacto HIGH**, lo que indica la presencia de mutaciones con un alto potencial de alterar la función de las proteínas codificadas. Según la documentación de SnpEff, las variantes de impacto **HIGH** suelen corresponder a alteraciones que modifican de manera severa la secuencia o estructura proteica, comprometiendo su función biológica.

Dentro de las categorías encontradas destacan principalmente los tipos:

- **frameshift_variant: Como en la primera evolución.**
- **frameshift_variant & stop_gained:** combinación de un corrimiento del marco de lectura con la aparición de un codón de parada prematuro. Este evento produce proteínas incompletas y no funcionales, que en muchos casos son degradadas por los mecanismos celulares de control de calidad del ARNm.
- **stop_lost & splice_region_variant:** pérdida del codón de parada junto con alteraciones en regiones involucradas en el corte y empalme (splicing). Este tipo de mutación puede generar proteínas extendidas o mal procesadas, afectando su estabilidad o localización celular.

Específicamente, las variantes de "HIGH impact" para esta evolución incluyen en el gen uno, igual al gen uno en Evol 1, la mutación del gen para la proteína de regulación de la histidina quinasa, que se encarga de señalización celular y reacciones a cambios de ambiente. Luego, también se encuentra la mutación del gen cuatro y cinco, encargado de la codificación de las proteínas que transporta ferricromo e iones de hierro en la membrana externa y genera los receptores de fagos, lo que así como puede mostrar un mal siendo mutada por los necesarios iones de hierro, esto puede mostrarse como una evolución que elimina algunos aceptores de fagos que pueden hacer daño a la bacteria. Como el gen cuatro en Evol1, el gen seis en Evol2 muestra un cambio que puede afectar la descomposición y utilización de maltosa como fuente de energía. El resto de las proteínas modificadas de "HIGH importance" en Evol2 son proteínas hipotéticas no estudiadas.

Ahora haciendo un breve análisis de otras variantes que son interesantes, sin que sean de alto impacto son las siguientes:

Número	Evolución	Gen	Cambio
1	Evol1	JBCMMFKD_00284	Lys634Asn|

2	Evol2	JBCMMFKD_00407	.

En el número uno, según los reportes de SnpEff, el cambio de una lisina a un aspártico en la posición 634, que codifica la subunidad C del complejo Rsx, corresponde a un SNP (Polimorfismo de un Solo Nucleótido) de tipo *missense*. Esto significa que no altera la estructura general de la proteína, pero sí produce un cambio en un aminoácido específico, lo que podría afectar su funcionamiento. Este tipo de mutaciones puede tener dos posibles efectos: uno adaptativo, en el que el cambio resulta beneficioso para el microorganismo, o uno negativo, que no aporta ninguna ventaja o incluso podría perjudicarlo. Las variantes *missense* son, de hecho, las más comunes en los organismos vivos.

En el número dos, se identificaron tres SNV (variantes de un solo nucleótido) que, a diferencia de los SNP, no suelen ser tan frecuentes. El efecto es similar al del primer caso: en determinadas posiciones se sustituye un aminoácido por otro. Además, estas mutaciones son heterocigotas.

Cabe aclarar que se sabe que son heterocigotas gracias a la columna “GT”, donde aparece el valor “0/1”, que indica precisamente ese tipo de variación.

Referencias

Bracken. (2022). Jhu.edu. https://ccb.jhu.edu/software/bracken/

jenniferlu717. (2025, February 26). *GitHub - jenniferlu717/Bracken: Bracken (Bayesian Reestimation of Abundance with KrakEN) is a highly accurate statistical method that computes the abundance of species in DNA sequences from a metagenomics sample.* GitHub. https://github.com/jenniferlu717/Bracken?tab=readme-ov-file

Kraken Manual -. (2017). Jhu.edu. https://ccb.jhu.edu/software/kraken/MANUAL.html#sample-reports

Lu, J. (2017, abril 27). *Why use Bracken instead of Kraken?* microBEnet: The Microbiology of the Built Environment Network. https://microbe.net/2017/04/27/why-use-bracken- instead-of-kraken/

Lu, J., Rincon, N., Wood, D. E., Breitwieser, F. P., Pockrandt, C., Langmead, B., Salzberg, S. L., & Steinegger, M. (2022). Metagenome analysis using the Kraken software suite. *Nature Protocols*. https://doi.org/10.1038/s41596-022-00738-y

Cingolani, P. (s. f.). *SnpEff: Input and output files*. Recuperado el 20 de octubre de 2025, de http://pcingola.github.io/SnpEff/snpeff/inputoutput/