Proyecto integrador en Bioinformática

Entrega 2

Laura del Sol González
Sara Flórez Hernández
Maria Camila Gómez Villegas
Sergio Andrés Morales Toro

Universidad EIA
Envigado – Antioquia
Octubre 2025

Punto 1: Descripción de clasificación taxonómica

Descripción del flujo de trabajo:

Para este punto se hizo la clasificación taxonómica de las lecturas que no se mapearon al genoma de referencia. El análisis se realizó con Kraken2, Bracken y Krona, siguiendo el procedimiento descrito en el archivo Scripts.pdf del punto 1. Primero se convirtieron los archivos .sam a .bam y se extrajeron las lecturas no mapeadas usando Samtools. Después, con Kraken2 y la base de datos minikraken2_v2_8GB_201904_UPDATE, se asignaron las lecturas a sus posibles grupos taxonómicos. Luego, Bracken permitió refinar los conteos para obtener resultados más precisos a nivel de especie. Por último se generaron las visualizaciones con Krona, obteniendo los archivos evol1_krona.html y evol2_krona.html, donde se puede observar la composición taxonómica de las muestras.

Resultados obtenidos:

Evol1:

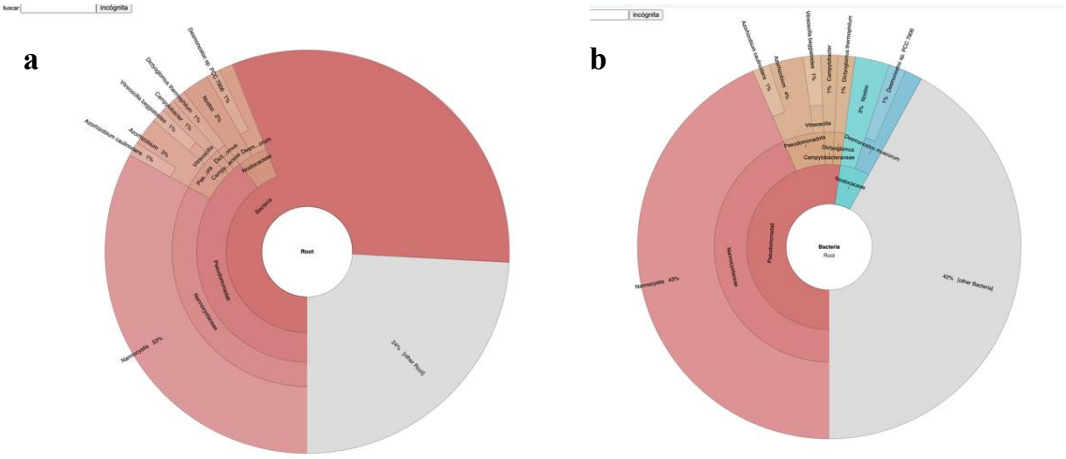

Figura1. Clasificación taxonómica de las lecturas no mapeadas de la muestra evol1 visualizada con Krona. (a) Vista general de la composición bacteriana. (b) Detalle ampliado de los géneros predominantes.

En la muestra evol1 se observó que la mayoría de las lecturas no mapeadas pertenecen al dominio Bacteria, lo que muestra que gran parte de las secuencias provienen de microorganismos bacterianos. El filo más abundante fue Pseudomonadota, con una alta presencia del género Nannocystis, que representó aproximadamente entre el 40 y el 45 % del total. Este resultado sugiere que las lecturas corresponden a bacterias típicas de ambientes naturales y suelos húmedos, ya que Nannocystis pertenece al grupo Myxococcota, conocido por formar estructuras cooperativas y participar en la degradación de materia orgánica.Estas bacterias suelen ser importantes en ecosistemas terrestres porque ayudan a reciclar nutrientes y descomponer compuestos complejos, lo que podría explicar su alta abundancia en la muestra. Además de Nannocystis, se encontraron otros géneros en menor proporción como Azorhizobium, Vitreoscilla, Dictyoglomus thermophilum y Campylobacter. Nos parecio interesante que aparezcan bacterias de ambientes tan diferentes, ya que eso podría indicar que la muestra contiene ADN ambiental mezclado o trazas de microorganismos de distintos entornos.

Tabla 1. Principales géneros bacterianos encontrados en la muestra evol1 y su posible función ecológica.

Género / Especie	Porcentaje estimado	Descripción biológica y relevancia
Nannocystis	40–45 %	Bacteria del suelo perteneciente al filo *Myxococcota*. Participa en la degradación de materia orgánica y procesos cooperativos microbianos.
Azorhizobium caulinodans	3–5 %	Fijadora de nitrógeno, asociada a raíces de plantas leguminosas; su presencia puede reflejar restos de ADN ambiental.
Vitreoscilla beggiatoides	2–3 %	Bacteria aerobia productora de hemoglobina bacteriana (*Vitreoscilla hemoglobin*), importante para la respiración en ambientes pobres en oxígeno.
Dictyoglomus thermophilum	2–3 %	Bacteria termófila, habitante de ambientes calientes, posiblemente presente por contaminación cruzada o secuencias ambientales.
Campylobacter	1–2 %	Bacteria microaerófila presente en ambientes intestinales o acuáticos; su baja abundancia indica posible contaminación ambiental.
No asignados	20–25 %	Lecturas sin clasificación en la base de datos, posiblemente por falta de genomas de referencia o secuencias incompletas.

Con estos resultados se puede ver que la muestra evol1 contiene principalmente bacterias ambientales y saprófitas que son microorganismos que viven en el suelo o en medios naturales y participan en la descomposición de materia orgánica. El hecho de que una parte de las secuencias (20–25 %) no haya podido clasificarse es algo normal, ya que la base de datos MiniKraken2 no tiene todos los genomas disponibles.

Evol 2:

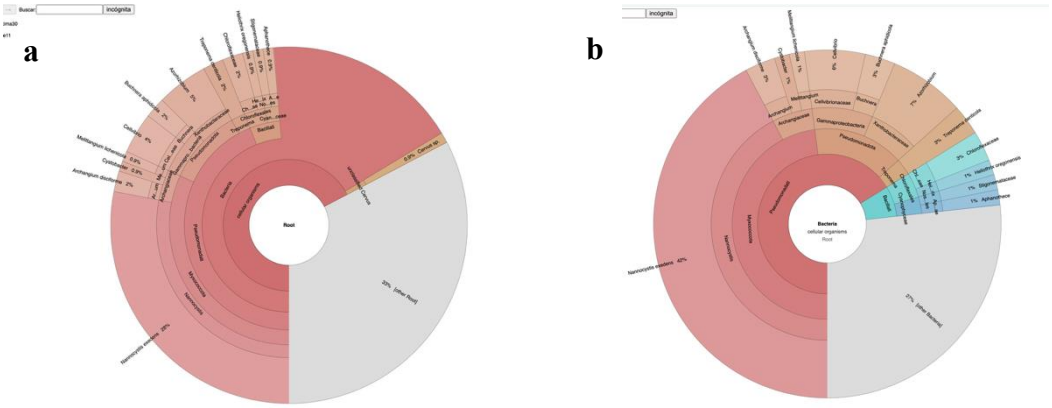

Figura 2.

Clasificación taxonómica de las lecturas no mapeadas de la muestra evol2 visualizada con Krona.(a) Vista general de la composición bacteriana. (b) Detalle ampliado de los géneros predominantes.

En la muestra evol2 se observó que, al igual que en evol1, la mayoría de las lecturas no mapeadas pertenecen al dominio Bacteria. Sin embargo, esta muestra presentó una mayor diversidad microbiana, ya que se identificaron más géneros bacterianos y varios filos distintos. El filo predominante fue Pseudomonadota, con una fuerte presencia de la clase Myxococcia y del género Nannocystis exedens, que representó aproximadamente entre un 40 % y 45 % del total de lecturas clasificadas. La alta abundancia de Nannocystis exedens muestra que probablemente se trata de una comunidad bacteriana típica de suelos y ambientes naturales, donde esta especie participa en la

degradación de materia orgánica y en procesos cooperativos entre microorganismos. Además, se detectaron géneros como Archangium disciforme, Melittangium lichenicola, Cystobacter, Cellvibrio y Buchnera aphidicola, entre otros. Estos resultados reflejan que en evol2 hay una mayor variedad de bacterias con funciones ecológicas diferentes, como la descomposición de compuestos vegetales, la fijación de nitrógeno o relaciones simbióticas con otros organismos. Nos pareció curioso que también se detectaran secuencias clasificadas como Cervus sp. ya que probablemente corresponden a contaminación cruzada o a errores de asignación por similitud entre fragmentos de ADN.

Tabla 2. Principales géneros bacterianos encontrados en la muestra evol2 y su posible función ecológica

Género / Especie	Porcentaje estimado	Descripción biológica y relevancia
Nannocystis exedens	40–45 %	Bacteria del suelo del filo *Myxococcota*, asociada a la degradación de materia orgánica.
Archangium disciforme	2–3 %	Bacteria depredadora del suelo que participa en la descomposición de materia orgánica.
Melittangium lichenicola	1–2 %	Bacteria con capacidad de formar esporas, común en ambientes naturales y húmedos.
Cystobacter	1–2 %	Bacteria del suelo con comportamiento cooperativo; forma cuerpos fructíferos.
Cellvibrio	6 %	Degrada polisacáridos vegetales, típica de suelos con abundante materia orgánica.
Buchnera aphidicola	3–4 %	Endosimbionte de insectos; su detección puede deberse a contaminación ambiental o ADN residual.
Treponema denticola	3 %	Bacteria anaerobia, presente en microbiotas animales; posible contaminación ambiental.
Heliothrix oregonensis / Aphanothece	1 %	Bacterias fotosintéticas, probablemente detectadas por ADN ambiental.
No asignados	27–33 %	Lecturas sin clasificación específica en la base de datos, posiblemente por falta de genomas de referencia.

Nos pareció interesante ver que, aunque evol1 y evol2 compartieron géneros similares, evol2 presentó más diversidad bacteriana y funciones ecológicas diferentes, lo que refleja que cada muestra tiene su propio perfil microbiano.

Punto 2: Tabla y discusión de variantes

En los archivos evol1_variantes.pdf, evol1_variantes_filtradas.pdf, evol2_variantes.pdf y evol2_variantes_filtradas.pdf se encuentran los resultados del análisis de variantes detectadas en cada muestra. Los archivos sin filtrar (evol1_variantes.pdf y evol2_variantes.pdf) contienen todas las variantes identificadas por bcftools call, mientras que los archivos filtrados (_variantes_filtradas.pdf) solo incluyen aquellas que cumplieron con los criterios de calidad establecidos (QUAL > 30 y DP > 10). Estas son las variantes consideradas más confiables y precisas.

Posteriormente, se utilizó Prokka, una herramienta ampliamente empleada para la anotación de genomas bacterianos, permitiendo identificar genes codificantes y otros elementos genómicos en el ensamblaje ancestral de Escherichia coli, generando archivos de salida en formato .gff y .faa con la información estructurada necesaria para la base de datos de SnpEff, herramienta utilizada para la predicción funcional de variantes genéticas. Aquí se integraron los archivos .vcf provenientes del análisis de variantes de las cepas evolucionadas (evol_1 y evol_2). Y, se construyó una base de datos a partir de la anotación generada por Prokka y las secuencias del genoma ancestral. Donde las variantes detectadas fueron clasificadas según su impacto funcional (HIGH, MODERATE, LOW) y se relacionaron con genes específicos. Con bcftools se extrajo la información más relévante de estos archivos (posición, referencia, alternativa, calidad y profundidad) y se convirtió en .csv para poder hacer una visualización en Jupyter Notebook y para hacer un filtrado manual.

A partir de los datos obtenidos del filtrado manual, se eligieron las variantes con los valores de calidad más altos para construir las tablas de resultados que se muestran a continuación. En ellas se resumen las variantes más representativas de cada muestra y su posible interpretación biológica.

En las tablas se pueden ver los campos principales obtenidos del archivo VCF, como la secuencia (contig), la posición genómica, el cambio de bases, el nombre del gen afectado y el tipo de variante detectada, junto con una breve descripción del posible efecto biológico asociado.

Tabla 3. Variantes seleccionadas en la muestra evol1 y su posible interpretación biológica

Secuencia	Posición (bp)	Cambio de bases	Tipo de Variante	Gen Afectado	Posible efecto biológico
NODE_1_length_58499_cov_4.208781,19176	19176	A - T	Stop lost and splice region variant	KAGFDIAF_00021	Perdida del codón de parada, que resulta en pérdida o ganancia de función
NODE_2_length_51549_cov_4.432322,36552	36552	GGTAAGTAA - GGTAA	Frameshift variant	KAGFDIAF_00083	Deleción de 4 nucleótidos, cambio en el marco de lectura, y por ende en secuencia de aminoácidos, resultara en proteína no funcional.
NODE_16_length_28105_cov_4.645312,23540	23540	T - A	Stop gained	KAGFDIAF_00530	Codon de parada prematuro, la cadena de aminoácidos se acorta resultado en perdida de función de la proteína resultante.
NODE_1_length_58499_cov_4.208781,7996	7996	C - T	Missense variant	KAGFDIAF_00010	Altera sutilmente la función de la proteína por cambio en el aminoácido, esto puede ser una mejora adaptativa.
NODE_37_length_21976_cov_4.474750,3594	3594	T - A	Stop lost and splice region variant	KAGFDIAF_01027	Pérdida del codón de parada, cambio disruptivo en la función de la proteína

Tabla 4. Variantes seleccionadas en la muestra evol2 y su posible interpretación biológica

Secuencia	Posición (bp)	Cambio de bases (REF→ALT)	Tipo de variante	Gen afectado	Posible efecto biológico
NODE_1_length_58499_cov_4.208781	7996	C → T	Missense variant	KAGFDIAF_00010	Cambia un aminoácido, puede alterar la función
NODE_12_length_31822_cov_4.522649	3530	T → TG	Frameshift variant	KAGFDIAF_00408	Cambio del marco de lectura, puede ser disruptivo y tener un efecto severo
NODE_16_length_28105_cov_4.645312	23540	T → A	Stop gained	KAGFDIAF_00530	Genera un codón de parada, lo que es muy dañino
NODE_106_length_12735_cov_4.212618	1060	T → A	Stop lost and splice region variant	KAGFDIAF_02079	Elimina el codón de parada y altera el splicing. Puede generar una proteína más larga y con estructura alterada
NODE_334_length_3908_cov_4.445886	2924	A → G	Stop lost and splice region variant	KAGFDIAF_03748	Elimina el codón de parada y altera el splicing. Puede generar una proteína más larga y con estructura alterada

En este análisis se encontraron muchas variantes en las muestras evol1 y evol2, lo que muestra que las lecturas mapeadas tienen bastantes diferencias con el genoma de referencia. Esto puede deberse a la diversidad natural de las bacterias presentes o a que algunas especies no están completamente representadas en la base de datos usada. Después del filtrado, se conservaron solo las variantes más confiables. Todas estas variantes retenidas y seleccionadas (en los archivos data_evol1.csv y data_evol2.csv) codifican proteínas, tienen un nivel de impacto alto según SnpEff (ósea que la mutación probablemente tenga consecuencias severas sobre la función de un gen) como frameshift, stop_gained o stop_lost (y en algunos casos variantes en regiones de splicing) y todas tienen codón de inicio (lo que nos asegura que es de buena calidad y se leyó bien. Estas categorías indican una alta probabilidad de alterar la proteína: los frameshift cambian el marco de lectura y suelen producir proteínas truncadas o no funcionales; los stop_gained introducen terminadores prematuros; los stop_lost extienden la traducción más allá del final esperado. Realizar su detección es útil porque ayuda a entender la variabilidad genética que existe entre las muestras y confirma que el filtrado funcionó bien. La identificación de este tipo de variantes es fundamental para comprender los posibles cambios funcionales que ocurren en el genoma, detectar señales de adaptación o evolución en poblaciones bacterianas o reconocer genes con gran importancia biológica.

Bibliografía:

- Center for Computational Biology, Johns Hopkins University. (2019). *Kraken2: Metagenomic sequence classification system*. Recuperado de https://ccb.jhu.edu/software/kraken2/
- Center for Computational Biology, Johns Hopkins University. (2019). *Bracken: Bayesian reestimation of abundance with Kraken*. Recuperado de https://ccb.jhu.edu/software/bracken/
- National Center for Biotechnology Information (NCBI). (2018). *Krona: Interactive metagenomic visualization*. Recuperado de https://github.com/marbl/Krona/wiki
- Wellcome Sanger Institute. (2021). *SAMtools: Sequence Alignment/Map tools*. Recuperado de http://www.htslib.org/
- Victorian Bioinformatics Consortium. (2023). *Prokka: Rapid prokaryotic genome annotation*. Recuperado de https://github.com/tseemann/prokka
- Universidad Nacional de Córdoba. (2023). *SnpEff: Genetic variant annotation and functional effect prediction toolbox*. Recuperado de https://pcingola.github.io/SnpEff/
- MicrobeWiki. (2023). *Nannocystis exedens*. Recuperado de https://microbewiki.kenyon.edu/index.php/Nannocystis_exedens
- MicrobeWiki. (2023). *Azorhizobium caulinodans*. Recuperado de https://microbewiki.kenyon.edu/index.php/Azorhizobium_caulinodans
- National Center for Biotechnology Information (NCBI). (2023). *Vitreoscilla beggiatoides (Taxonomy ID: 41428)*. Recuperado de https://www.ncbi.nlm.nih.gov/
- Madigan, M. T., Bender, K. S., Buckley, D. H., Sattley, W. M., & Stahl, D. A. (2019). *Brock: Biología de los microorganismos* (15.a ed.). Pearson Educación.
- MicrobeWiki. (2023). *Archangium disciforme*. Recuperado de https://microbewiki.kenyon.edu/index.php/Archangium_disciforme
- MicrobeWiki. (2023). *Melittangium lichenicola*. Recuperado de https://microbewiki.kenyon.edu/index.php/Melittangium_lichenicola
- MicrobeWiki. (2023). *Cystobacter fuscus*. Recuperado de https://microbewiki.kenyon.edu/index.php/Cystobacter_fuscus
- MicrobeWiki. (2023). *Cellvibrio mixtus*. Recuperado de https://microbewiki.kenyon.edu/index.php/Cellvibrio_mixtus
- National Center for Biotechnology Information (NCBI). (2023). *Buchnera aphidicola (Taxonomy ID: 9)*. Recuperado de https://www.ncbi.nlm.nih.gov/
- National Center for Biotechnology Information (NCBI). (2023). *Treponema denticola (Taxonomy ID: 158)*. Recuperado de https://www.ncbi.nlm.nih.gov/
- National Center for Biotechnology Information (NCBI). (2023). *Heliothrix oregonensis (Taxonomy ID: 1097)*. Recuperado de https://www.ncbi.nlm.nih.gov/
- National Center for Biotechnology Information (NCBI). (2023). *Aphanothece sp. (Taxonomy ID: 1194)*. Recuperado de https://www.ncbi.nlm.nih.gov/