Universidad Nacional Autónoma de México
Facultad de Ciencias
Genómica computacional
Sergio Hernández López, Maira Nayeli Luis Vargas y Rafael López Martínez

Proyecto: Identificación y clasificación de hongos en una muestra de eDNA mediante herramientas genómicas

Ruiz Almanza Nancy Selene, Gonzalez Dominguez Saul Fernando & Carrasco Reyes Bayron

---

Resumen:
El término ácido desoxirribonucleico ambiental (eDNA) se acuñó para definir al ácido desoxirribonucleico (DNA) que se puede recuperar o detectar del ambiente sin necesidad de que al menos un individuo de una especie concreta esté físicamente presente. El objetivo del presente trabajo es utilizar el eDNA para la evaluación de especies, lo que lo convierte en una herramienta versátil e importante para el futuro en investigación, permitiendo estudios de conservación, taxonómicos o de reconstrucción filogenética (incluyendo: la reconstrucción histórica de sus comunidades, la restauración del ecosistema y la salud humana). Para lograr esto, se usa el procedimiento de *metabarcoding*, el cual se basa en obtener DNA de cualquier origen (en este caso eDNA), en ausencia física o no del organismo, con apoyo de la reacción en cadena de la polimerasa (PCR), para finalmente, secuenciarse, obtener regiones del espaciador transcrito interno (ITS, del inglés *Internal transcribed spacer*), compararlas y analizarlas con otras ITS ya conocidas, para finalmente asignarles una categoría taxonómica identificable.

Introducción:
En el campo de la biología molecular desde hace más de dos décadas ha cobrado relevancia el estudio del genoma de los organismos, en los últimos años ha tenido un auge el uso de herramientas moleculares y bioinformáticas para realizar estudios genómicos sobre la biodiversidad contenida en diferentes ecosistemas. Se denomina eDNA al ácido desoxirribonucleico que se puede recuperar de una muestra ambiental, que puede provenir de suelo, aire o agua, sin necesidad de extraer el DNA directamente del organismo, el eDNA suele contener parte de la diversidad biológica que se encuentra en el ecosistema a estudiar. Estudiar el eDNA puede tener diferentes funciones, entre ellas, evaluar las especies reconstruyendo su filogenia, su taxonomía o estudiar el ecosistema con fines de restauración o conservación; el potencial de estos estudios puede ser útil no sólo para darle seguimiento a la biodiversidad que contiene un ecosistema, sino que puede tener aplicaciones en el análisis de la salud humana. Para realizar este tipo de estudios se utiliza una herramienta llamada *metabarcoding*, que se basa en obtener el eDNA, haciendo uso de otras herramientas moleculares como la reacción en cadena de la polimerasa (PCR); en el caso de organismos microscópicos como bacterias y hongos también suelen utilizarse técnicas de cultivo; posteriormente se realiza la secuenciación, que comúnmente se realiza por secuenciación de Sanger de nueva generación (NGS), ya que permite analizar, de manera masiva en un solo ensayo, un organismo a la vez para obtener su respectiva secuencia, esta técnica se basa en la terminación de la amplificación o extensión de cadenas de ácidos nucleicos, produciendo cadenas de diferentes tamaños que se pueden agrupar por una secuencia de nucleótidos en común que indica dónde terminó la reacción enzimática. Es importante destacar que en estudios de metabarcoding se utilizan secuencias evolutivamente conservadas, universales, secuencias específicas del genoma o genoma mitocondrial del organismo, a las que los investigadores suelen referirse como “código de barras”, pues para detectar los organismos presentes en el eDNA se requiere de estos “códigos de barras” porque sirven como una etiqueta molecular para identificar especies (Padilla-García, *et. al.*, 2021).

Es importante que entendamos que la biodiversidad no ha sido completamente estudiada ni descrita (debido principalmente a las tecnologías que ocupamos para cognizar), por lo que también los estudios de metabarcoding sirven para detectar mediante el eDNA a los organismos que con técnicas tradicionales no pueden ser estudiados. Los organismos fungales presentan una gran diversidad y plasticidad, lo cual sin duda les permite tener una ventaja competitiva sobre muchos de los otros grupos (De ahí que sea el segundo grupo con mayor éxito evolutivo de los eucariontes, tal vez el primero). Sin embargo, esto implica una gran dificultad en su estudio. Desde comprender taxonómicamente con qué organismos estamos trabajando, por lo que el uso del metabarcoding para su identificación y estudio es de gran utilidad

Con los estudios moleculares del DNA se puede tener una mayor precisión al identificar especies diferentes de hongos, lo cual es de gran ayuda porque no hay suficientes estudios que caractericen y describan la amplia diversidad de este reino, además gran parte de esta diversidad microscópica es sólo detectable por su DNA, pues en ocasiones no es posible cultivarlos . Los organismos fungales presentan claramente una inmensa diversidad, la cual sin duda le permitió tener una ventaja competitiva al explorar los recursos a su alcance (permitiendo establecerse como el segundo grupo de eucariotes más exitoso, inclusive ser el primero). Sin embargo, esto presenta grandes retos desde clasificar sus especies, comprender su filogenia, visualizar sus nichos ecológicos, entender sus historias de vida. En el caso del organismo modelo *Neurospora crassa*, que es un caso bien estudiado hasta hace poco sólo se conocía su fase miceliar, pero debido al cambio climático puede cambiar su forma a una levadura y causar coccidiomicosis. Debido a que siendo un organismo patógeno, una de sus adaptaciones es perder su pared de quitina y secuestrar la membrana del hospedero (una membrana de colesterol de los vertebrados) para así eludir el sistema inmunológico. Al perder una de las características diagnósticas de los hongos, esto tiene consecuencias a la hora de diagnosticar e implementar un tratamiento efectivo, causando 1.5 millones de fallecimientos. *(Nieto, A. & De la Rua, 2022)*

La región del espaciador transcrito interno (ITS) son las secuencias evolutivamente conservadas, universales y específicas de identificación más exitosa para la gama más amplia de especies fungales, con la brecha de código de barras más claramente definida entre la variación interespecífica e intraespecífica. La subunidad grande del ribosoma nuclear, un marcador filogenético popular en ciertos grupos, tenía una resolución de especie superior en algunos grupos taxonómicos, como los primeros linajes divergentes y las levaduras de ascomicetos, pero por lo demás era ligeramente inferior a la ITS. La subunidad pequeña ribosómica nuclear tiene una resolución pobre a nivel de especie en los hongos. ITS se propondrá formalmente para su adopción como el principal marcador de código de barras de hongos para el Consorcio para el código de barras de la vida, con la posibilidad de que se desarrollen códigos de barras complementarios para grupos taxonómicos particulares estrechamente circunscritos. como los primeros linajes divergentes y las levaduras de ascomicetos, pero por lo demás era ligeramente inferior a la ITS. La subunidad pequeña ribosómica nuclear tiene una resolución pobre a nivel de especie en los hongos. ITS se propondrá formalmente para su adopción como el principal marcador de código de barras de hongos para el Consorcio para el código de barras de la vida, con la posibilidad de que se desarrollen códigos de barras complementarios para grupos taxonómicos particulares estrechamente circunscritos (Schoch et al., 2012).

En este trabajo realizaremos un ejercicio de identificación taxonómica de hongos en una muestra de eDNA utilizando como marcador filogenético los ITS de ciertos phylum, para lo cual seguiremos un tutorial previamente descrito que nos permitirá entender cómo funciona el algoritmo mediante el cual se realizan los estudios de metabarcoding para abordar el problema de la identificación de la diversidad fungal en un ecosistema.

Pregunta de investigación
Al realizar un estudio de metabarcoding, ¿cómo funciona el algoritmo para clasificar la diversidad de hongos en un ecosistema partiendo de las secuencias obtenidas de una muestra de DNA ambiental?

Hipótesis
Los marcadores moleculares son una herramienta útil para crear filogenias, por lo que proponemos que usando marcadores filogenéticos específicos de hongos, como los ITS, podremos realizar la clasificación taxonómica de estos organismos, incluso sin saber a qué especie pertenecen, ya que la diversidad es tan amplia que podríamos encontrar organismos no descritos previamente en una muestra de eDNA

Objetivo
Emplear herramientas bioinformáticas para crear un programa que nos permita clasificar filogenéticamente las secuencias de hongos desconocidos obtenidos a partir de una muestra de DNA ambiental.

Metodología:
Para llevar a cabo este ejercicio de clasificación taxonómica seguimos algunos pasos del tutorial DADA2 ITS Pipeline Workflow (1.8), en el cual explican el flujo de trabajo que se realiza para la clasificación taxonómica de organismos utilizando ITS, a partir de archivos fastq de dos extremos secuenciados por Illumina que han sido divididos por muestras y se les han eliminado los "códigos de barras", utilizando para esto la paquetería de software de código abierto DADA2.

La paquetería DADA2 para R sirve para modelar y corregir errores de amplicón secuenciados por Illumina, puede inferir las secuencias de cada muestra y resuelve diferencias de hasta 1 nucleótidos, lo que a su vez permite que identifique variantes reales y que produzca menos secuencias espurias. Esta paquetería además amplía y mejora el algoritmo de DADA, que es un algoritmo de eliminación de ruido de amplicones divisivos basado en modelos para corregir errores de amplicones sin construir Unidades Taxonómicas Operativas (OTU). DADA2 implementa un modelo consciente de la calidad de los errores de amplicón de Illumina ya que cuantifica la tasa en la que se produce una lectura de amplicón en función de la composición y calidad de la secuencia, no tiene referencias y es aplicable a cualquier locus genético, además implementa el flujo de trabajo de amplicón completo (filtrado, desaplicación, inferencia de muestras, identificación de quimeras y combinación de lecturas de extremos emparejados). En comparación con otros métodos DADA2 ha demostrado ser más preciso que otros porque resuelve la variación a escala fina mejor que el método OTU

Comenzamos extrayendo 12 secuencias fastq que fueron descritas en el material suplementario del artículo "Maize stalk rot caused by *Fusarium graminearum* alters soil microbial composition and is directly inhibited by Bacillus siamensis isolated from rhizosphere soil" de Zhang y colaboradores (2022); en el cual realizaron un estudio de la diversidad microbiana de la rizosfera (parte del suelo próxima a las raíces de una planta) de *Zea mays*, comúnmente conocida como maíz, contenida en suelos con enfermedad severa de pudrición del tallo y el suelo de rizosfera libre de enfermedades; además en este trabajo se estudió la composición taxonómica de la misma mediante secuenciaciones por amplicones dirigidas para organismos de bacterias y hongos; para el caso de los organismos fungales amplificaron la región ITS del gen de RNA ribosomal utilizando pares de primers específicos de hongos como el ITS5F (forward) y el ITS2R (reverse). Posteriormente realizaron la amplificación de las secuencias mediante PCR para después secuenciarlas con el secuenciador Ilumina Miseq. En este estudio se obtuvieron un total de 1,205,999 secuencias del gen ITS de hongos clasificables, los

cuales estaban contenidos en 9 phylum más abundantes (*Ascomycota, Basidiomycota, Zygomycota, Chytridiomycota, Cercozoa, Ciliophora, Glomeromycota, Neocallimastigomy,y Rozellomycota*)

Para comenzar nuestro ensayo en R studio primero fue necesario identificar y eliminar los primers o cebadores de las secuencias de organismos identificados y verificar la orientación de esos mismos primers, ya que en cada secuencia se asignan dos tipos de primers: 1) forward, que polimeriza la hebra + o hebra guía y tiene una orientación de 5' a 3'; y 2) reverse, que polimeriza la hebra - o hebra molde y tiene una orientación de 3' a 5'. Para eliminar los primers contenidos en las secuencias tuvimos ayuda de la profesora Nayeli Vargas

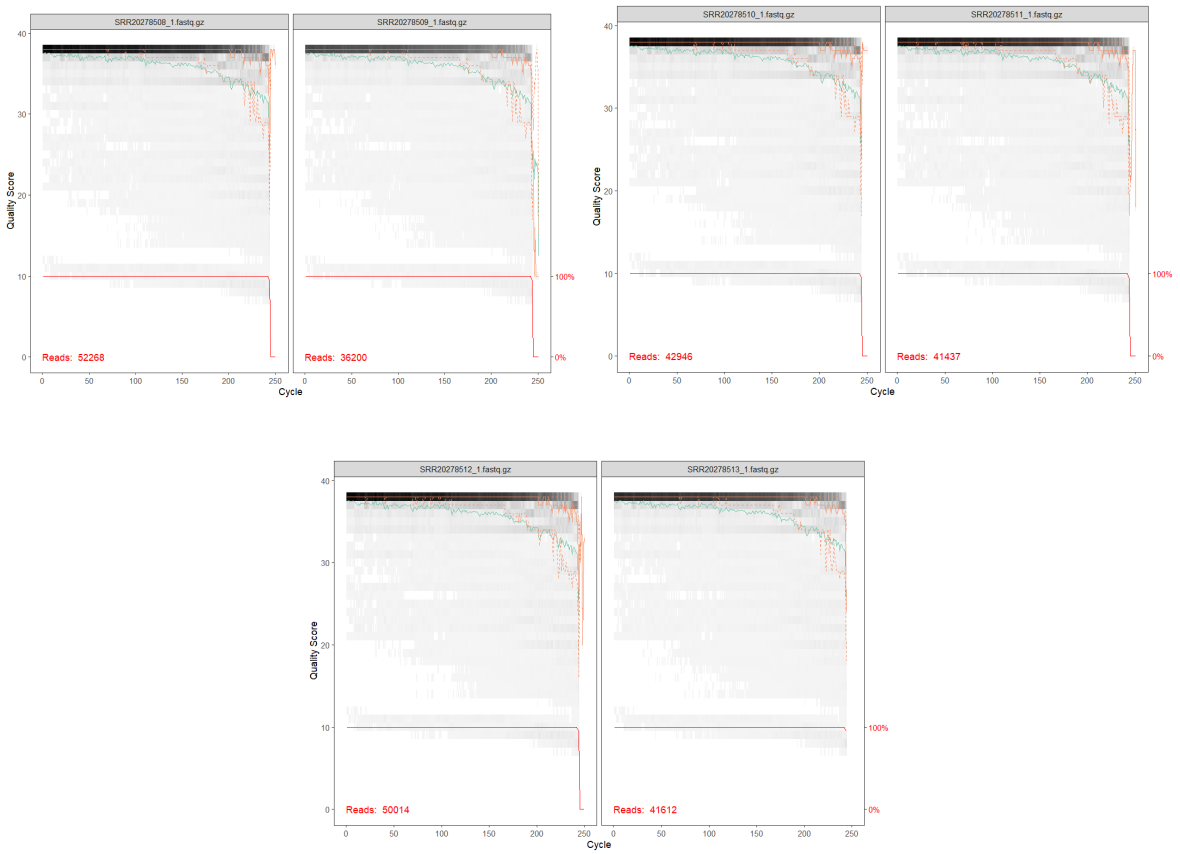
Una vez que nos proporcionaron los archivos de secuencias fastq sin cebadores procedimos a realizar una inspección de los perfiles de calidad de lectura. En ellos podemos ver un mapa en escala de grises de la frecuencia de cada puntuación de calidad en cada posición base. La puntuación de calidad mediana en cada posición se muestra en la línea verde y los cuartiles de la distribución de la puntuación de calidad en las líneas naranjas. La línea de lectura muestra la proporción escalada de lecturas que se extienden al menos hasta esa posición.

Notamos que las lecturas son de buena calidad, puesto que la linea verde, en general, la podemos encontrar arriba de los 30 puntos de Phred. Recordando que el nivel de calidad Phred (Phred quality) es una medida de calidad en la identificación de las nucleobases generadas por la secuenciación automatizada de ADN.

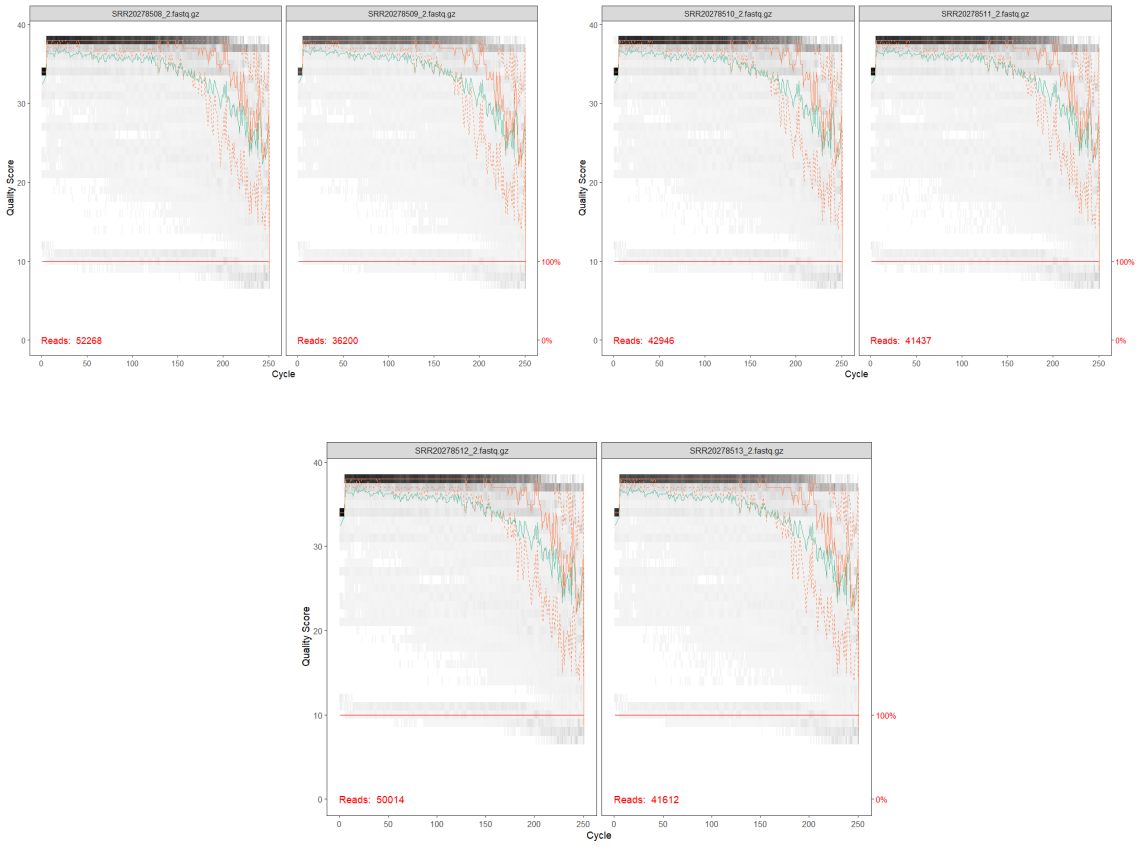
El nivel de calidad de Phred se utilizan para medir la precisión de las llamadas de base, esta es una de las métricas más comunes y evaluá la calidad de los datos de secuenciación de distintas tecnologías. Podemos resumir que los puntajes Phred bajos nos indicarían un aumento de las llamadas de variantes de falsos positivos, lo que resulta en conclusiones inexactas y costos más altos para los experimentos de validación. El propósito principal de estas puntuaciones es proporcionar evidencia adicional de que la secuencia, la alineación, el ensamblaje y el SNP son, de hecho, reales y no se deben a un problema al generar las secuencias.

La línea roja muestra que una parte significativa de las lecturas se adaptaron al corte a aproximadamente 250 de longitud, lo que probablemente refleja la longitud de la región ITS amplificada en uno de los taxones presentes en estas muestras. De igual forma, podemos notar que vemos el mismo pico de longitud en alrededor de 250 nts en las mediciones *reverse*, lo que representa una buena señal de coincidencia.

Calidad forward



Calidad reverse



Posteriormente, usando la paquetería DADA2, utilizando la función *filterAndTrim,* en donde como su nombre lo dice filtramos y recortamos los archivos fastq de entrada en función de varios criterios definibles por el usuario, y generamos archivos fastq (comprimidos de forma predeterminada) que contienen las lecturas recortadas que pasaron los filtros. Los archivos fastq directos e inversos correspondientes se pueden proporcionar como entrada, en cuyo caso el filtrado se realiza en las lecturas directas e inversas de forma independiente, y ambas lecturas deben pasar para que se emita el par de lectura. Los parametros que se usaron fueron:

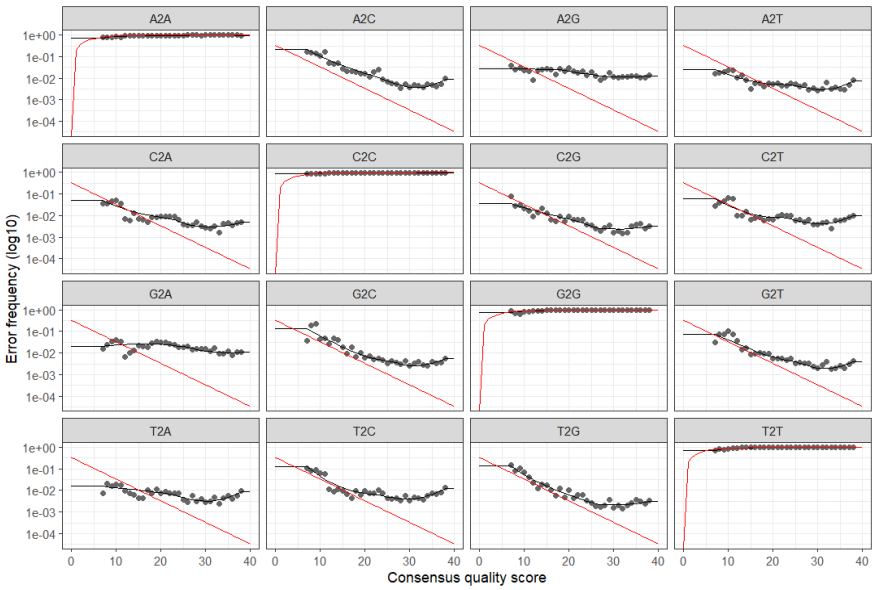
- trunQ: Trunca las lecturas en la primera instancia de un puntaje de calidad menor o igual a truncQ.
- maxN: Valor predeterminado 0. Después del truncamiento, las secuencias con más de maxN se descartarán.
- maxRR: Predeterminado Inf(sin filtrado EE). maxEEDespués del truncamiento, se descartarán las lecturas con más de "errores esperados".
- minLen:. Predeterminado 20. Eliminar lecturas con una longitud inferior a minLen.
- compress: Determina si la salida, seran valores comprimidos
- multithread: Habilita el uso de multiples hilos(Solo se puede usar en linux)

La paquetería DADA2, incorpora información de calidad en su modelo de error, lo que hace que el algoritmo sea robusto en secuencias de menor calidad, y al medir la diversidad a través de diferentes parámetros de filtrado y tasas de error mejora la sensibilidad del algoritmo a variantes de secuencias raras. Este algoritmo utiliza un modelo de error paramétrico, por lo que cada conjunto de datos de amplicón tiene un conjunto diferente de tasas de error. Para obtener este error empleamos el método *leansErrors* que aprende este modelo de error de los datos, alterando la estimación de las tasas de error.

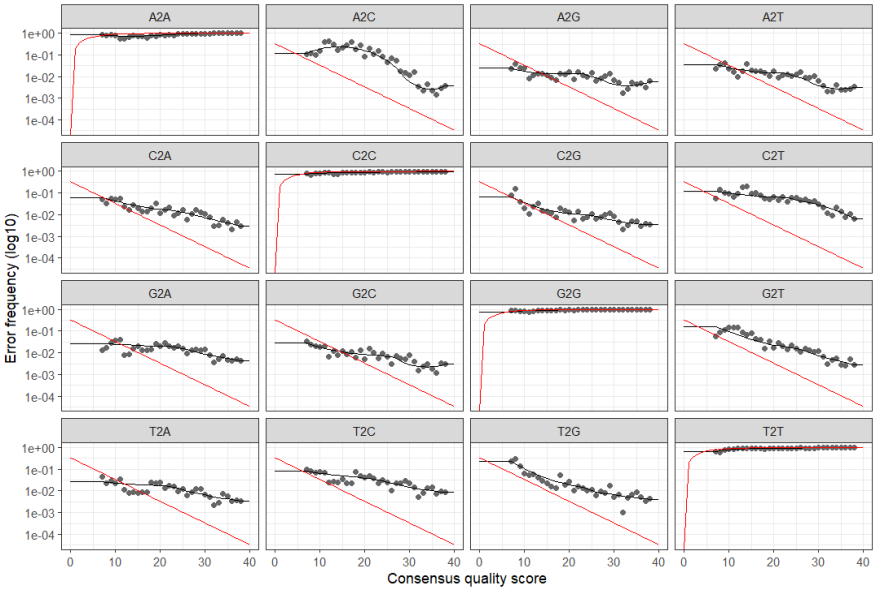
Posteriormente aplicamos el método de DADA2 que implementa el algoritmo Divisive Amplicon Denoising Algorithm, que es un algoritmo de eliminación de ruido que infiere tanto los genotipos de muestra como los parámetros de error que produjeron un conjunto de datos de metagenoma. Elimina la necesidad de datos de entrenamiento para establecer parámetros de error, utiliza completamente la información de abundancia de secuencias y permite la inclusión de tasas de error de PCR dependientes del contexto.

La forma de interpretar estos valores de error, es cuando se tienen las transiciones por ejemplo A2C, implica una transición de Adenina a Citocina, la columna indica el numero de apariciones de errores en cierto valor de calidad.

Error reverse



Error forward



Después, anulamos la replicación de secuencias de amplicones de archivos fastq o fastq comprimidos, al mismo tiempo que controlamos los requisitos de memoria máxima para admitir archivos de gran tamaño, con la función *derepFastq*.

Luego con ayuda de la función *mergePairs*, permitimos fusionar cada par sin ruido de lecturas directas e inversas, rechazando cualquier par que no se superponga lo suficiente o que contenga demasiadas discrepancias (>0 de forma predeterminada) en la región de superposición. Nota: esta función asume que los archivos fastq para las lecturas hacia adelante y hacia atrás estaban en el mismo orden.

Ahora podemos construir una tabla de variantes de secuencia de amplicón (ASV), una versión de mayor resolución de la tabla OTU producida por métodos tradicionales. La tabla de secuencias es una

matriz con filas correspondientes a las muestras y columnas correspondientes a las variantes de secuencia. Esta tabla contiene 293 ASV, y todas las longitudes de nuestras secuencias combinadas se encuentran dentro del rango esperado para este amplicón V4. Esto con ayuda del método *makeSequenceTable*.

El método central DADA2 corrige los errores de sustitución e indel, pero quedan quimeras. Afortunadamente, la precisión de las variantes de secuencia después de eliminar el ruido hace que identificar ASV quiméricos sea más simple que cuando se trata de OTU difusas. Las secuencias quiméricas se identifican si pueden reconstruirse exactamente combinando un segmento izquierdo y un segmento derecho de dos secuencias "primarias" más abundantes. Para esto se emplea el metodo *removeBimeraDenovo.*

Finalmente, para poder construir la taxonomía se emplea el método *assignTaxonomy*, el cual emplea el método de clasificación de Naive Bayes, para poder asignar la taxonomía. En la sección de discusión se explicará este método con mayor detalle. Algo que es importante mencionar, es que la ejecución de este método puede ser algo lenta, en el caso de nuestras secuencias tomó un par de horas, pero puede ser que en secuencias de mayor longitud o al clasificar un mayor número de secuencias, pueda tomar varios días.

Resultados

Como resultado final de toda la ejecución del algoritmo,se pudieron clasificar las diversas secuencias encontrando la clasificación taxonómica a la que pertenecen, dicha clasificación se presenta a continuación.

Secuencia	**Reino**	**Filo**	**Clase**	**Orden**
SRR20278508	**Fungi**	**Ascomycota**	**Sordariomycetes**	**Hypocreales**
SRR20278509	**Fungi**	**Basidiomycota**	**Tremellomycetes**	**Cystofilobasidiales**
SRR20278510	**Fungi**	**Ascomycota**	**Sordariomycetes**	**Hypocreales**
SRR20278511	**Fungi**	**Basidiomycota**	**Tremellomycetes**	**Cystofilobasidiales**
SRR20278512	**Fungi**	**Ascomycota**	**Sordariomycetes**	**Sordariales**
SRR20278513	**Fungi**	**Ascomycota**	**Sordariomycetes**	**Hypocreales**

Secuencia	**Familia**	**Genero**	**Especie**
SRR20278508	**Nectriaceae**	**Fusarium**	**NA**
SRR20278509	**Cystofilobasidiaceae**	**Guehomyces**	***pullulans***

SRR20278510	**Nectriaceae**	**Fusarium**	**NA**
SRR20278511	**Cystofilobasidiaceae**	**Guehomyces**	***pullulans***
SRR20278512	**Chaetomiaceae**	**Humicola**	***nigrescens***
SRR20278513	**Nectriaceae**	**Fusarium**	**NA**

Como parte de la evidencia del proceso, colocamos de igual forma una imagen de nuestra clasificación como se obtuvo directamente de la terminal, notando que en este caso, existe una notación en donde el primer caracter corresponde a si la cadena hace referencia a un reino(Kingdom), un filo (Phylum), una clase(Class), orden(Order), familia(Family), género(Genus) o especie(Species).

```
     Kingdom    Phylum             Class                Order
[1,] "k__Fungi" "p__Ascomycota"    "c__Sordariomycetes" "o__Hypocreales"
[2,] "k__Fungi" "p__Basidiomycota" "c__Tremellomycetes" "o__Cystofilobasidiales"
[3,] "k__Fungi" "p__Ascomycota"    "c__Sordariomycetes" "o__Hypocreales"
[4,] "k__Fungi" "p__Basidiomycota" "c__Tremellomycetes" "o__Cystofilobasidiales"
[5,] "k__Fungi" "p__Ascomycota"    "c__Sordariomycetes" "o__Sordariales"
[6,] "k__Fungi" "p__Ascomycota"    "c__Sordariomycetes" "o__Hypocreales"
     Family                   Genus           Species
[1,] "f__Nectriaceae"         "g__Fusarium"   NA
[2,] "f__Cystofilobasidiaceae" "g__Guehomyces" "s__pullulans"
[3,] "f__Nectriaceae"         "g__Fusarium"   NA
[4,] "f__Cystofilobasidiaceae" "g__Guehomyces" "s__pullulans"
[5,] "f__Chaetomiaceae"       "g__Humicola"   "s__nigrescens"
[6,] "f__Nectriaceae"         "g__Fusarium"   NA
```

Discusión:

Para poder realizar la asignación taxonómica y como parte de nuestra pregunta de investigación hay que entender cómo funciona el algoritmo del clasificador de Naive-Bayes, para ello, es necesario entender el teorema de Bayes, teorema por el cual este clasificador, recibe su nombre, y para ello primero tenemos que entender cómo funciona la probabilidad condicional.

La probabilidad condicional es la probabilidad de que ocurra un evento A, sabiendo que también sucede otro evento B. La probabilidad condicional se escribe P(A|B) y se lee «la probabilidad de A dado B». La idea intuitiva, es que tenemos el evento en donde ocurre la intersección de los dos eventos y vamos a "escalar" esta probabilidad. La fórmula para poder encontrarlo es la siguiente.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Ahora que tenemos la fórmula de probabilidad condicional, podemos entonces deducir el teorema de bayes, notemos que la fórmula de arriba se puede escribir de la siguiente forma, debido a que estamos buscando la probabilidad de que A ocurra cuando sucede B.

$$P(A \cap B \mid B) = \frac{P(A \cap B)}{P(B)}$$

Ademas nuevamente por la formula de probabilidad condicional también ocurre que:

$$P(A \cap B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Notamos entonces que podemos despejar $P(A \cap B)$ e igualar las ecuaciones.

$$P(A \cap B \mid A)P(A) = P(A \cap B \mid B)P(B)$$

Ademas eliminando la notación de la intersección en la probabilidad condicional, tenemos que

$$P(B \mid A)P(A) = P(A \mid B)P(B)$$

Finalmente, despejando la probabilidad condicional, obtenemos el teorema de Bayes.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Ahora que tenemos el teorema de Bayes, podemos pasar a entender el clasificador de Naive Bayes. Un clasificador(en el análisis de datos) es un algoritmo utilizado para asignar un elemento entrante no etiquetado en una categoría concreta conocida. Dicho algoritmo, permite asignar una clase a los objetos entrantes, tomando en cuenta ciertas características de este. De forma que modelan este problema con probabilidad condicional, podemos considerar a las características como una serie de variables aleatorias independientes entre sí, y queremos saber la probabilidad de que al ocurrir estas variables ocurra una variable dependiente de estas variables es decir pertenecer a una clase. Empleando el teorema de bayes, tendremos

$$P(C_i \mid F_1 \cap \ldots \cap F_n) = \frac{P(F_1 \cap \ldots \cap F_n \mid C_i)P(C_i)}{P(F_1 \cap \ldots \cap F_n)}$$

Recordando que lo que queremos hacer es asignar una clase a nuestro objeto, por lo que querríamos comparar las probabilidades obtenidas para cada clase. Notemos que al realizar esta comparación el denominador siempre será igual para todas las clases, por lo que entonces procederemos a despreciar.

Por otro lado, podemos desarrollar el numerador empleando por la fórmula de probabilidad condicional, es decir, que tenemos

$$P(F_2 \cap \ldots \cap F_n \mid C_i, F_1) = \frac{P(F_1 \cap \ldots \cap F_n \mid C_i)}{P(F_1 \mid C_i)}$$

Es decir que

$$P(F_1 \cap \ldots \cap F_n \mid C_i) = P(F_1 \mid C_i)P(F_2 \cap \ldots \cap F_n \mid C_i, F_1)$$

Y nuevamente volvemos a aplicar la formula de probabilidad condicional

$$P(F_3 \cap \ldots \cap F_n \mid C_i, F_1, F_2) = \frac{P(F_2 \cap \ldots \cap F_n \mid C_i, F_1)}{P(F_2 \mid C_i, F_1)}$$

Es decir que

$$P(F_3 \cap \ldots \cap F_n \mid C_i, F_1) = P(F_2 \mid C_i, F_1)P(F_3 \cap \ldots \cap F_n \mid C_i, F_1, F_2)$$

Por lo que se cumple la siguiente igualdad, en además podemos aplicar la definición de probabilidad condicional de manera iterativa.

$$P(C_i)P(F_1 \cap \ldots \cap F_n \mid C_i) = P(C_i)P(F_1 \cap \ldots \cap F_n \mid C_i)$$

$$= P(C_i)P(F_1|C_i)P(F_2 \cap \ \dots \ \cap \ F_n \mid C_i, F_1)$$
$$= P(C_i)P(F_1|C_i)P(F_2|C_i, F_1)P(F_3 \cap \ \dots \ \cap \ F_n \mid C_i, F_1 F_2)$$
$$\dots$$
$$= P(C_i)P(F_1|C_i)P(F_2|C_i, F_1)\dots P(F_n| \ C_i, F_1, F_2, \dots \ F_{n-1})$$

Pero como las variables son independientes entonces podemos reescribir la expresión como:

$$P(C_i)P(F_1|C_i)P(F_2|C_i, F_1)\dots P(F_n| \ C_i, F_1, F_2, \dots \ F_{n-1})$$
$$= P(C_i)P(F_1|C_i)P(F_2|C_i)\dots P(F_n| \ C_i)$$

Que sería el valor que tendríamos que calcular para determinar y comparar para determinar a qué clase pertenece cada uno de los objetos que poseen las características.

En el artículo de Wang [5], explica que el funcionamiento de la asignación taxonómica, funciona de manera muy parecida a la clasificación de textos, en una clasificación de textos, se toman palabras y se considera como características el número de ocurrencias de las palabras en dicho textos. En el caso del artículo, señala que para poder analizar el texto, dividió las enormes secuencias en "palabras" de longitud 8. Y como datos de referencia, tomó todas las secuencias anatómicas en NCBI que fueron obtenidas antes del 2004, estas secuencias están clasificadas en alrededor de 1,187 clases, y el promedio de la longitud de estas cadenas es de 1,454 bases. Sin embargo, el caso concreto de los organismo fungales presenta otras problemáticas, como es la carencia de definiciones de especies que nos permitan distinguir entre grupos con alta variabilidad clinial y variación interespecífica y esto sin duda tiene consecuencias para la identificación y descripción de nuevas especies. En el caso de los organismos cuya asignación de especie presenta NA, podría sin duda ser el primer registro de una especie no descrita o que no ha sido secuenciada anteriormente, y esto también implicaciones en los estudios de diversidad ya que no se conservan muchos de los holotipos de algunas especies por lo cual se puede discutir cuál es la secuencia a comparar dentro de un grupo con alta variabilidad clinial (como es el caso de los complejos de especies).

Conclusiones

La metodología de metabarcoding es un método muy verificable y consistente, para poder asignar categorías taxonómicas a secuencias eDNA. Además es útil para ampliar las bases de datos en las bibliotecas de secuencias genómicas ya que como mencionamos, no hemos descrito ni descubierto toda la biodiversidad existente en la biosfera, por lo que tampoco hemos secuenciado a todas las especies ya descubiertas. La ampliación de las bases de datos es importante para ampliar el conocimiento y la caracterización de los organismos que hasta ahora no han sido estudiados. Este tipo de estudios son transdisciplinarios porque involucran desde un trabajo de campo en el que se toma la muestra de eDNA, pasando por las metodologías de biología molecular que se han perfeccionado a lo largo de los años para permitir hacer estudios de secuenciación masiva implementando herramientas computacionales para su análisis como lo son los software y algoritmos que facilitan la interpretación de esos datos y nos permiten darles una interpretación evolutiva a partir de la elaboración de clasificaciones filogenéticas

Referencias:

[1] Padilla-García, C. Y., Camacho-Sánchez, F. Y., & Reyes-López, M. Á. (2021). Metabarcoding de DNA ambiental: un enfoque para el seguimiento de la biodiversidad. *CienciaUAT*, *16*(1), 136-149. Retrieved December 14, 2022, from https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-78582021000200136

[2] Callahan, B. (n.d.). *DADA2: Fast and accurate sample inference from amplicon data with single-nucleotide resolution*. Dada2: Fast and accurate sample inference from AMPLICON data with single-nucleotide resolution. Retrieved December 14, 2022, from https://benjjneb.github.io/dada2/index.html

[3] Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods*, *13*(7), 581-583. Retrieved December 14, 2022, from https://www.nature.com/articles/nmeth.3869#methods

[4] Nieto, A. & De la Rua, A.. (10 de Diciembre de 2022). *Vídeo | No estamos preparados para una pandemia de hongos: las 19 especies mortales que preocupan a la OMS*.. . El País Recuperado from https://elpais.com/ciencia/2022-10-26/video-no-estamos-preparados-para-una-pandemia-de-hongos-las-19-especies-mortales-que-preocupan-a-la-oms.html

[5] Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, *73*(16), 5261-5267. Retrieved December 14, 2022, from https://pubmed.ncbi.nlm.nih.gov/17586664/