

## Resumen

Los métodos de secuenciación utilizados por Charles River incluyen el ensamblaje automatizado de datos, el análisis y la revisión de la calidad de los datos, la medición de la distancia genética y la interpretación de los resultados. Estos elementos, junto con el uso de una biblioteca de referencia validada y relevante, garantizan la máxima precisión de los resultados de identificación microbiana.



## SOLUCIONES MICROBIANAS

# Características del método de identificación microbiana por secuenciación AccuGENX-ID®

## Importancia de la identificación microbiana en los programas de monitoreo ambiental

Cuando se logra aislar cepas bacterianas o fúngicas procedentes de entornos industriales, es sumamente importante identificar el organismo a nivel de especie de manera oportuna y con precisión. Para obtener resultados coherentes y precisos, se necesitan sistemas de identificación microbiana fiables que permitan rastrear el organismo, completar la investigación y evitar demoras en la distribución del producto. Los métodos que arrojan identificaciones inexactas, incongruentes o nulas no son útiles para rastrear las cepas aisladas hasta su origen, ni para generar informes de tendencias. En última instancia, esto puede dar lugar a una falsa sensación de control, la aplicación de medidas de descontaminación erróneas y costos operativos adicionales para la empresa. Los métodos de identificación precisos generarán una base de datos histórica fiable que permita comparaciones e interpretaciones correctas de los datos, así como la aplicación de medidas de mitigación de riesgos adecuadas.

Hay cuatro áreas principales que afectan específicamente a la precisión de las identificaciones microbianas por secuenciación: 1) la relevancia y la amplitud de la cobertura de la biblioteca de referencia, 2) la calidad de los datos generados por el sistema, 3) el ensamblaje y análisis de los datos y el cálculo de las mediciones

de distancia, y 4) la interpretación de los resultados (Figura 1). Los métodos utilizados por Charles River se remontan a los métodos utilizados por los taxónomos, que analizan datos de secuencias microbianas mediante un proceso de ensamblaje convencional. Este proceso incluye el autoensamblaje, la revisión de los datos y un análisis filogenético para generar una identificación.

Figura 1.



Hay cuatro áreas principales que afectan específicamente a la precisión de las identificaciones microbianas por secuenciación: 1) la calidad de los datos generados por el sistema, 2) la relevancia y la amplitud de la cobertura de la biblioteca de referencia, 3) el ensamblaje y análisis de los datos y 4) la interpretación de los resultados.

### Aumento de la precisión y la fiabilidad

La secuenciación comparativa de los genes de ARN ribosómico (ARNr) es la tecnología de genotipado de referencia para la clasificación taxonómica de bacterias y hongos. Asimismo, es el método más preciso y reproducible para identificar microorganismos desconocidos. Aunque los conocimientos científicos que respaldan la tecnología están ampliamente aceptados, todavía hay una serie de variables que pueden afectar a su implementación y al logro de una identificación precisa. El análisis de los datos de la secuencia de ADN es un proceso complicado y el rendimiento general de los diferentes sistemas o servicios de identificación de ADN ribosómico (ADNr) que se utilizan actualmente no es uniforme. Existen numerosos pasos que requieren que un científico experimentado tome decisiones para analizar e interpretar correctamente los datos. Estos pasos críticos son la evaluación de la calidad de la secuencia de ADN, el ensamblaje de la secuencia de ADN y la interpretación filogenética. Cuando estos pasos clave los lleva a cabo un filogenético microbiano experimentado cumpliendo los requisitos del proceso taxonómico, los resultados obtenidos son de mayor calidad. La precisión de la identificación depende no solo de los métodos utilizados para generar, analizar e interpretar los datos, sino también, y de forma significativa, de la base de datos de la biblioteca utilizada como referencia. Durante casi 20 años, el grupo Charles River Accugenix® ha identificado microorganismos mediante secuenciación de ADNr comparativa. Nuestra experiencia, junto con las mejoras que hemos introducido en el proceso de análisis y nuestras bibliotecas validadas y relevantes, nos permite proporcionar resultados de identificación de la más alta calidad para cepas aisladas desconocidas.

### Impacto de la biblioteca de referencia

Charles River reconoce que la biblioteca de referencia es un componente clave para la identificación microbiana, por lo que otorga la máxima prioridad al mantenimiento y la recalificación anual de sus bibliotecas de ADN microbiano validadas, con el objetivo de proporcionar las identificaciones más precisas para todas las muestras bacterianas y fúngicas. El rendimiento, la fiabilidad y la relevancia superiores de los sistemas de identificación microbiana requieren bibliotecas que exhiban una cobertura amplia y detallada de las cepas aisladas que son importantes para las industrias manufactureras estériles y no estériles. Las diferencias entre las bibliotecas de referencia con las que compara sus datos pueden

afectar a la frecuencia de las identificaciones microbianas correctas. Para identificar correctamente un gran porcentaje de las cepas aisladas desconocidas en entornos industriales, la biblioteca debe contener secuencias de ADN de los organismos con mayor probabilidad de aparición. Charles River ha documentado los organismos hallados en estas instalaciones en todo el mundo, lo que aumenta la relevancia de nuestras bibliotecas.

Cada día se descubren, nombran y publican nuevas especies bacterianas y fúngicas. Estas nuevas especies pueden ser organismos hallados previamente en fábricas y podrían representar un riesgo para el producto. Además, la clasificación taxonómica de los organismos cambia constantemente. Con la aparición de la secuenciación del ADN como método de referencia para la clasificación e identificación de microorganismos, los científicos corrigen ahora errores taxonómicos cometidos en el pasado, cuando se empleaba únicamente la caracterización fenotípica y no la relación filogenética. Las bibliotecas Accugenix® se seleccionan y actualizan continuamente para reflejar los cambios taxonómicos y la inclusión de nuevos organismos hallados en entornos industriales. El hecho de actuar como laboratorio contratado para entornos industriales sometidos a normativas estrictas, obliga a Charles River a seguir un programa que cumple rigurosamente las BPF en vigor y que incluye nuestros procedimientos originales de validación de bibliotecas y continúa impulsando el mantenimiento de estas<sup>1</sup>.

### Ensamblaje y análisis de datos

Además de la biblioteca de referencia, otros factores que afectan a la precisión de las identificaciones son el método de análisis de los datos y las pautas de interpretación de los datos. La precisión se ve afectada por la capacidad del *software* del secuenciador para interpretar los diferentes tipos de anomalías de datos que afectan a la identificación de nucleótidos durante el proceso de secuenciación normal, así como por el método de conciliación de posibles errores en estas identificaciones de nucleótidos. Estos errores son causados por desplazamientos en la movilidad de los picos, inserciones o eliminaciones de bases, y posiciones de bases mixtas en genes de múltiples copias. La precisión también se ve afectada por la longitud de la secuencia de consenso resultante utilizada para la comparación con la biblioteca, donde la secuencia de consenso comprende los nucleótidos encontrados más comúnmente en cada ubicación en la alineación de la secuencia de ADN objetivo.

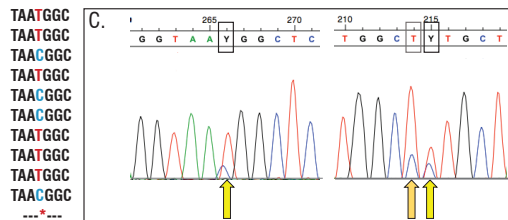
Durante el análisis de los datos, primero confirmamos que la calidad de estos es lo suficientemente alta como para generar una secuencia de consenso precisa. El control de calidad de datos inicial se realiza para clasificar los datos sin procesar como aceptables, no aceptables o que requieren verificación. La calidad de los datos se ve afectada por el propio operón de ARNr. El operón de ARNr comprende los genes para las tres moléculas de ARNr de diferentes tamaños (por ejemplo, los genes de ARNr 16S, 23S y 5S en bacterias o los genes de ARNr 18S, 5.8S y 28S en hongos). La mayoría de las especies tienen múltiples copias del operón ribosómico, hasta 15 copias en algunos casos. Las copias están dispuestas en tándem en el cromosoma y, a menudo, no todas las copias de cada especie tienen la misma secuencia o la misma longitud. Sin embargo, se amplifican todas las copias del gen, ya que se utilizan cebadores universales de la RCP para amplificar las regiones de ARNr objetivo con el fin de identificarlo. Cuando se secuencian estos productos de la RCP, la secuencia es una representación de la suma de todas las copias. Las diferencias en las secuencias entre las copias de ARNr 16S, por ejemplo, pueden dar como resultado polimorfismos o bases mixtas en diferentes posiciones de nucleótidos. En la Figura 2, se observa un polimorfismo en el electroferograma de datos sin procesar como nucleótidos distintos en la misma posición de la secuencia. Debe prestarse especial atención durante la revisión de los datos para identificar las posiciones mixtas de nucleótidos, ya que son importantes desde un punto de vista evolutivo.

**Figura 2.**

A. Organización genómica de los operones de ARNr

TAATGGC~~~TAATGGC~~~TAACGGC~~~TAACGGC~~~TAATGGC  
COPIA 1 COPIA 2 COPIA 3 COPIA 4 COPIA 5

B. Productos de la RCP que representan todas las copias diferentes



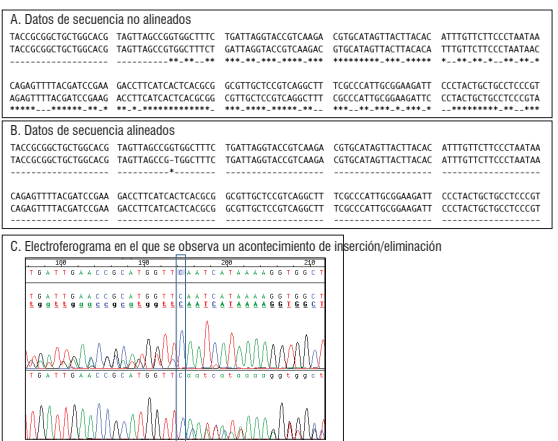
D. Secuencias de consenso GGTAA/GGC TC TGGCYTGCT

Visualización y determinación de identificaciones de nucleótidos polimórficos. A) Las copias de los operones de ARNr están dispuestas en tándem en el genoma y pueden tener diferentes secuencias. B) El objetivo del gen se amplifica con RCP y se visualiza mediante secuenciación Sanger. C) Ambas bases se pueden ver

en la misma posición en los electroferogramas, lo que indica polimorfismo de bases mixtas. El algoritmo de análisis identifica correctamente las posiciones indicadas con las flechas amarillas en los dos electroferogramas como «Y», una mezcla de T (pico rojo) y C (pico azul), mientras que la flecha púrpura indica una posición donde el *software* lector de nucleótidos no pudo reconocer el polimorfismo. C) En la revisión de los datos se detectan y concilian estos errores de lectura de nucleótidos, registrando la posición polimórfica de «Y» en las secuencias de consenso.

Además, para algunas de las cepas analizadas, las longitudes de los genes de ARNr no son las mismas porque a lo largo de la evolución del microorganismo se ha producido una inserción o eliminación de bases. Por ello, debe procederse con el mismo cuidado a la hora de revisar los datos, para capturar estos acontecimientos (Figura 3). Por ejemplo, los datos que se obtienen de un organismo que contiene múltiples copias del gen ARNr 16S, o la región ARNr ITS, que no tienen la misma longitud debido a un acontecimiento de inserción o eliminación, parecen muy similares a las secuencias de ADN derivadas de un cultivo mixto. Sin embargo, nuestros científicos experimentados en la revisión de datos pueden detectar fácilmente la diferencia entre secuencias de cultivos mixtos y datos de secuencia que contienen inserciones o eliminaciones, y utilizan un *software* patentado para conciliar estos acontecimientos. Esta capacidad permite el análisis y ensamblaje satisfactorios de un mayor número de muestras y un mayor rendimiento.

**Figura 3.**



Ejemplo de errores de lectura de nucleótidos resultado de un acontecimiento de inserción o eliminación en una copia de la región 16S o ITS. A) Los datos no alineados dan lugar a muchos desajustes (indicados con asteriscos). B) Los datos alineados compensan la inserción/eliminación e indican solo una falta de coincidencia. C) Imagen de la secuencia de datos sin procesar que muestra el efecto de una inserción/eliminación (en el cuadro

azul) en los patrones de pico, lo que hace que los datos a ambos lados del acontecimiento parezcan «mixtos».

Evaluar la calidad de los datos de la secuencia y conciliar las posiciones polimórficas garantiza que nuestro análisis utilice la secuencia completa generada a partir del producto de la RCP. En nuestro análisis no se omiten nucleótidos entre los sitios del cebador de la RCP y se genera una secuencia consenso completa. Nuestra experiencia ha demostrado que las mediciones de distancia más precisas, y por lo tanto las identificaciones más precisas, se generan utilizando la secuencia de ADN completa.

### Mediciones de distancia

La taxonomía bacteriana y fúngica refleja un esquema de clasificación filogenética basado en la secuencia del gen ARNr 16S o la región ITS, representado en dendrogramas o árboles, cuya estructura se basa en diferentes modelos de evolución. Nosotros utilizamos el mismo proceso para nuestras identificaciones. Una vez generada la secuencia de consenso, se compara con nuestra biblioteca de referencia utilizando la herramienta básica de búsqueda de alineación local (BLAST —*basic local alignment search tool*).

La búsqueda determina las coincidencias con los elementos de la biblioteca de referencia más similares a la secuencia desconocida. Estas entradas de la biblioteca de referencia inician el proceso de análisis filogenético con el cálculo de las diferencias porcentuales y las distancias genéticas.

En primer lugar, se alinean las secuencias para minimizar el número absoluto de diferencias entre ambas. A continuación, se comparan todas las posiciones de nucleótidos en ambas secuencias (comparación por pares) y se calcula la diferencia porcentual. Las diferencias porcentuales reflejan el número absoluto de bases que son diferentes entre la cepa desconocida y la de referencia en toda la secuencia. Un ejemplo sencillo: una diferencia de seis bases en una región de un gen de 500 pb es una diferencia del 1,2 % ( $6/500 \times 100 = 1,2\%$ ).

A continuación, se analizan las secuencias de referencia más estrechamente relacionadas y la muestra desconocida de acuerdo con un modelo de matriz de distancias.

Las mediciones de distancia son comparaciones entre dos secuencias y también se expresan en forma de porcentaje. Los modelos de matriz de distancias se basan en lo que se espera que genere diferencias entre las secuencias actuales comparándolas, contando las diferencias y aplicando correcciones para las mutaciones superpuestas. El número real de cambios puede subestimarse simplemente observando la diferencia observada en las secuencias actuales. Por lo tanto, las diferencias se transforman

en distancias utilizando modelos como el de Jukes-Cantor. Se calculan las mediciones de distancia evolutiva y, a partir de estos datos, se construye el árbol de Neighbor-joining, que es el modelo utilizado por Charles River y otros filogenéticos. El árbol es el núcleo del informe de identificación y tiene que interpretarse.

### Interpretación de los datos

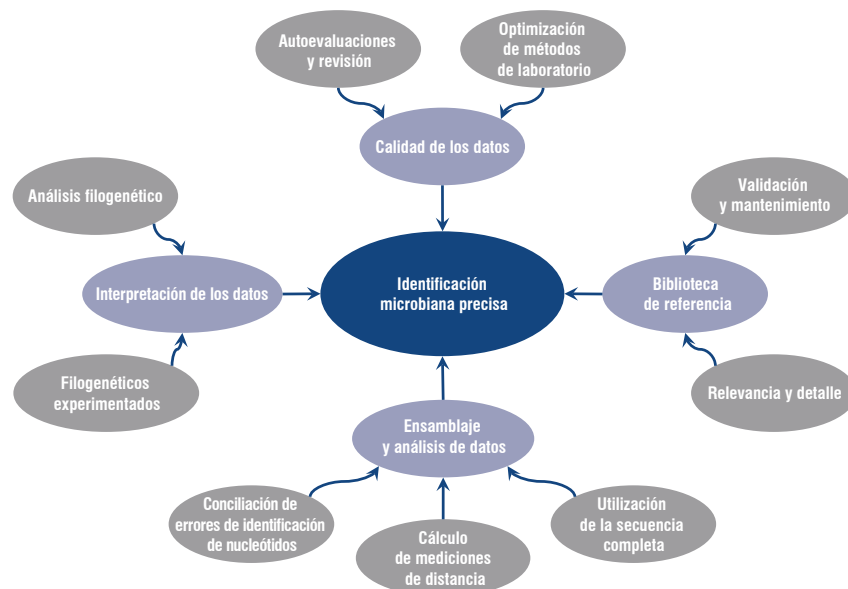
La interpretación del árbol filogenético es una parte crucial del proceso de análisis, ya que no se puede aplicar ninguna regla de interpretación universal. No existe ningún estándar filogenético para la delimitación de familias, géneros o especies. Tampoco existe ningún acuerdo universal sobre un valor umbral que constituya o defina de forma definitiva una especie. Debido a que no todos los géneros bacterianos evolucionan a la misma velocidad y a que muchos organismos se clasificaron erróneamente como pertenecientes a la misma especie o a especies diferentes antes de la evaluación de la secuencia de genes de ARNr, es necesario usar diferentes valores umbral en función del género de que se trate. Por lo tanto, el valor umbral varía según el grupo taxonómico, ya que algunas especies están demasiado estrechamente relacionadas o hay poca información sobre la variación normal de cepa a cepa dentro de una misma especie, o entre operones dentro de una misma cepa. La consideración más importante al hacer una identificación es el árbol filogenético. El árbol filogenético es una representación visual de la variabilidad genética entre los organismos de la biblioteca más estrechamente relacionados con la cepa desconocida. Tanto la distribución como el orden de ramificación indican cómo se relacionan entre sí los organismos y son importantes a la hora de hacer la interpretación final.

Nuestros biólogos filogenéticos especializados en microbiología tienen mucha experiencia y reconocen dónde pueden estar los posibles problemas y cómo estos pueden hacer que la interpretación sea confusa. Una de las áreas problemáticas es que la clasificación filogenética de muchos organismos es actualmente incorrecta. Los organismos están mal clasificados y mal nombrados, lo que crea una situación muy compleja. Otra de las áreas problemáticas es que, en ocasiones, el análisis filogenético de las regiones del gen ribosómico no es suficiente para la resolución de la especie. Esto es lo que sucede con el complejo *Burkholderia cepacia*. Hay muchos organismos que están estrechamente relacionados y presentan un alto grado de conservación en las regiones de ARN ribosómico, de modo que la

secuenciación de 16S o ITS2 no es suficiente para que la precisión de la clasificación vaya más allá del nivel de grupo o complejo (lo que corresponde a nuestro nivel de confianza «especie\*», que indica que «\* El organismo desconocido coincide con dos o más especies estrechamente relacionadas»). Para aumentar la discriminación, se puede utilizar la información de secuenciación de otros marcadores genéticos, como los objetivos de genes que codifican proteínas como gyrB, recA o TUB2. Sin embargo, en muchos casos ni siquiera la secuenciación de genes que codifican proteínas es suficiente para lograr la resolución a nivel de especie debido a que la clasificación taxonómica es compleja o está mal definida. En estos casos, tenemos que aceptar una identificación a nivel de grupo o complejo (nivel de confianza especie\*), ya que es la respuesta más precisa.

Al interpretar los informes y asignar el nivel de confianza taxonómica, nuestros filogenéticos microbianos utilizan una combinación de la variabilidad genética, el orden de ramificación en el árbol de Neighbor-joining y el conocimiento de la variación entre especies. Gracias a nuestra experiencia en identificación (más de 1 000 000 de microorganismos desconocidos), Charles River tiene la veteranía y el conocimiento para convertir esta información compleja en identificaciones rutinarias.

**Figura 4.**



Entre los factores que afectan a la precisión de una identificación microbiana, se incluyen la calidad de los datos, el ensamblaje y el análisis de los datos, la interpretación de los datos y la biblioteca de referencia.

## Resumen

Los métodos de identificación filogenética por secuenciación utilizados por Charles River para nuestro servicio AccuGENX-ID se remontan a los métodos utilizados por los taxónomos; con ellos se analizan los datos de las secuencias microbianas mediante un proceso convencional. Este proceso incluye el autoensamblaje, la revisión de los datos y un análisis filogenético para generar una identificación. Los métodos patentados empleados por Charles River comparan secuencias de muestra con bibliotecas patentadas de cobertura total, dan como resultado una interpretación de datos concluyente y una identificación con niveles de confianza asignados basados en análisis filogenéticos. La utilización de estos métodos científicamente probados y las bibliotecas patentadas validadas y continuamente seleccionadas proporcionan las identificaciones microbianas más precisas (Figura 4). A su vez, estos datos permitirán un seguimiento y análisis de tendencia precisos de las cepas aisladas, así como una gran confianza en la interpretación de los resultados.

## Bibliografía

1. Creating and Maintaining Validated Microbial Identification Libraries. 2018. Charles River Technical Note.