4 bases de datos para realizar análisis bioinformático de comunidades microbianas

Laura Brenes-Guillén

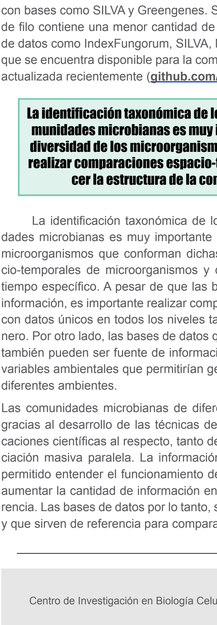
Las comunidades microbianas son posiblemente los ambientes más diversos y abundantes del planeta, todos los ecosistemas y organismos dependemos de una u otra forma de las actividades metabólicas microbianas. El estudio de estos ambientes utilizando herramientas de genética clásica no permiten identificar cepas de microorganismos que no sean cultivables o que dependen de condiciones de cultivo muy específicas. Es por lo tanto importante empezar a utilizar herramientas de secuenciación de nueva generación (NGS) que permitan identificar y caracterizar a los microorganismos cultivables y no cultivables, así como los que se encuentran en bajas abundancias.

La metagenómica es una herramienta reciente que utiliza la secuenciación de nueva generación (NGS) para estudiar y caracterizar las comunidades de microorganismos. Específicamente los estudios de librerías 16S tratan de identificar taxonómicamente los diferentes microorganismos y con ello comparar la diversidad y riqueza entre muestras y ambientes específicos. Uno de los pasos más importantes en este análisis es la asignación taxonómica de cada uno de los OTUS (Unidades Taxonómicas Operacionales), para lo cual se necesita de un algoritmo que pueda clasificar y alinear las secuencias y una base de datos de referencia. La selección de la base de datos debe considerar que esta contenga secuencias similares a la región de ADN que se quiera analizar en la investigación. Entre las bases de datos más comúnmente utilizadas se encuentran SILVA, Greengenes y NCBI. Sin embargo también existen otras bases como OTT, RDP, y UNITE.

SILVA

La base de datos SILVA contiene información taxonómica de bacterias, arqueas y eucariotas, basándose principalmente en filogenias de ARN ribosomal de subunidades pequeñas (16S / 18S, SSU) y de subunidades grandes (23S / 28S, LSU). SILVA utiliza como referencia el manual de sistemática de bacteriología de Bergey y las opiniones taxonómicas de la Sociedad Internacional de Protistas. La última versión (v132) fue publicada en diciembre de 2017 y cuenta con aproximadamente 7 millones de secuencias. Esta base se encuentra disponible de forma gratuita para todos los usuarios (**www.arb-silva.de**).

Greengenes



Otra de las bases ampliamente utilizada es Greengenes (**greengenes.secondgenome.com**), la cual contiene únicamente secuencias de bacterias y arqueas. Esta base se alimenta con datos de bases públicas como NCBI y CyanoDB.cz. A pesar de estar incluida dentro de paquetes muy utilizados en los análisis metagenómicos (e.g. Qiime), la última actualización de la Greengenes fue liberada en el 2013.

NCBI

La base de datos NCBI es una de las más utilizadas como referencia para análisis filogenéticos clásicos, no obstante también contiene información de bioproyectos y un repositorio de secuencias y alineamientos generados a partir de NGS. Esta plataforma incluye específicamente datos de secuenciación de librerías 16S, secuenciación masiva paralela, datos de genomas y transcriptomas. Está asociada a bases de datos como GNN (Genome News Network, Instituto J. Craig Venter, **www.genomenewsnetwork.org/**) y GOLD (Genomes OnLine Database, DOE-Joint Genome Institute, **gold.jgi.doe.gov/**). Este último pertenece al Programa de Genómica del Departamento de Energía de Estados Unidos.

OTT

Por otro lado, OTT (OpenTreeofLIFE) es una base que ha sido poco utilizada para los análisis de secuenciación de la región 16S, sin embargo cuenta con una mayor cantidad de secuencias clasificadas taxonómicamente hasta nivel de género en comparación con bases como SILVA y Greengenes. Sin embargo, en cuanto a la clasificación a nivel de filo contiene una menor cantidad de información. OTT utiliza información de bases de datos como IndexFungorum, SILVA, NCBI y Global Diversity Information. A pesar de que se encuentra disponible para la comunidad científica desde el año 2015, no ha sido actualizada recientemente (**github.com/opentreeoflife**).

La identificación taxonómica de los datos de secuenciación masiva de comunidades microbianas es muy importante para identificar y estudiar la diversidad de los microorganismos que conforman dichas comunidades, realizar comparaciones espacio-temporales de microorganismos y conocer la estructura de la comunidad en un tiempo específico

La identificación taxonómica de los datos de secuenciación masiva de comunidades microbianas es muy importante para identificar y estudiar la diversidad de los microorganismos que conforman dichas comunidades, realizar comparaciones espacio-temporales de microorganismos y conocer la estructura de la comunidad en un tiempo específico. A pesar de que las bases de datos comparten cierto porcentaje de información, es importante realizar comparaciones entre éstas, ya que cada una cuenta con datos únicos en todos los niveles taxonómicos, especialmente a nivel de filo y género. Por otro lado, las bases de datos que se utilizan como referencia para los análisis también pueden ser fuente de información, ya que algunas tienen datos asociados de variables ambientales que permitirían generar nuevas hipótesis y comparaciones entre diferentes ambientes.

Las comunidades microbianas de diferentes ambientes son ampliamente estudiadas gracias al desarrollo de las técnicas de secuenciación masiva, existen muchas publicaciones científicas al respecto, tanto de secuenciación del gen 16S, como de secuenciación masiva paralela. La información que se obtiene de este tipo de estudios ha permitido entender el funcionamiento de los ecosistemas microbianos y por otro lado, aumentar la cantidad de información en las bases de datos que se utilizan como referencia. Las bases de datos por lo tanto, son una herramienta en constante actualización y que sirven de referencia para comparar los diferentes estudios alrededor del mundo.

Laura Brenes-Guillén
Centro de Investigación en Biología Celular y Molecular (CIBCM), Universidad de Costa Rica
San José, Costa Rica

Figura encabezado: Courtesy of Pacific Northwest National Laboratory (CC BY-NC-SA 2.0)

Figura en texto: Comparación de taxonomías basadas en nombres de taxones encontrados a distintos niveles de clasificación, de las bases de datos SILVA, Greengenes, NCBI, OTT y RDP (no descrita en este blog); comúnmente usadas para análisis metagenómicos. Basado en (Balvočiūtė & Huson, 2017).

Referencias

Agarwala, R. et al. (2016). Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research, 44(D1), D7-D19.**

Balvočiūtė, M., & Huson, D. (2017). SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? **BMC Genomics, 18(2), 114.**

Glöckner, F. et al. (2017). 25 years of serving the community with ribosomal RNA gene reference databases and tools. **Journal of Biotechnology, 261, 169-176.**

Hinchliff, C. et al. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. **Proceedings of the National Academy of Sciences, 112(41), 12764-12769.**