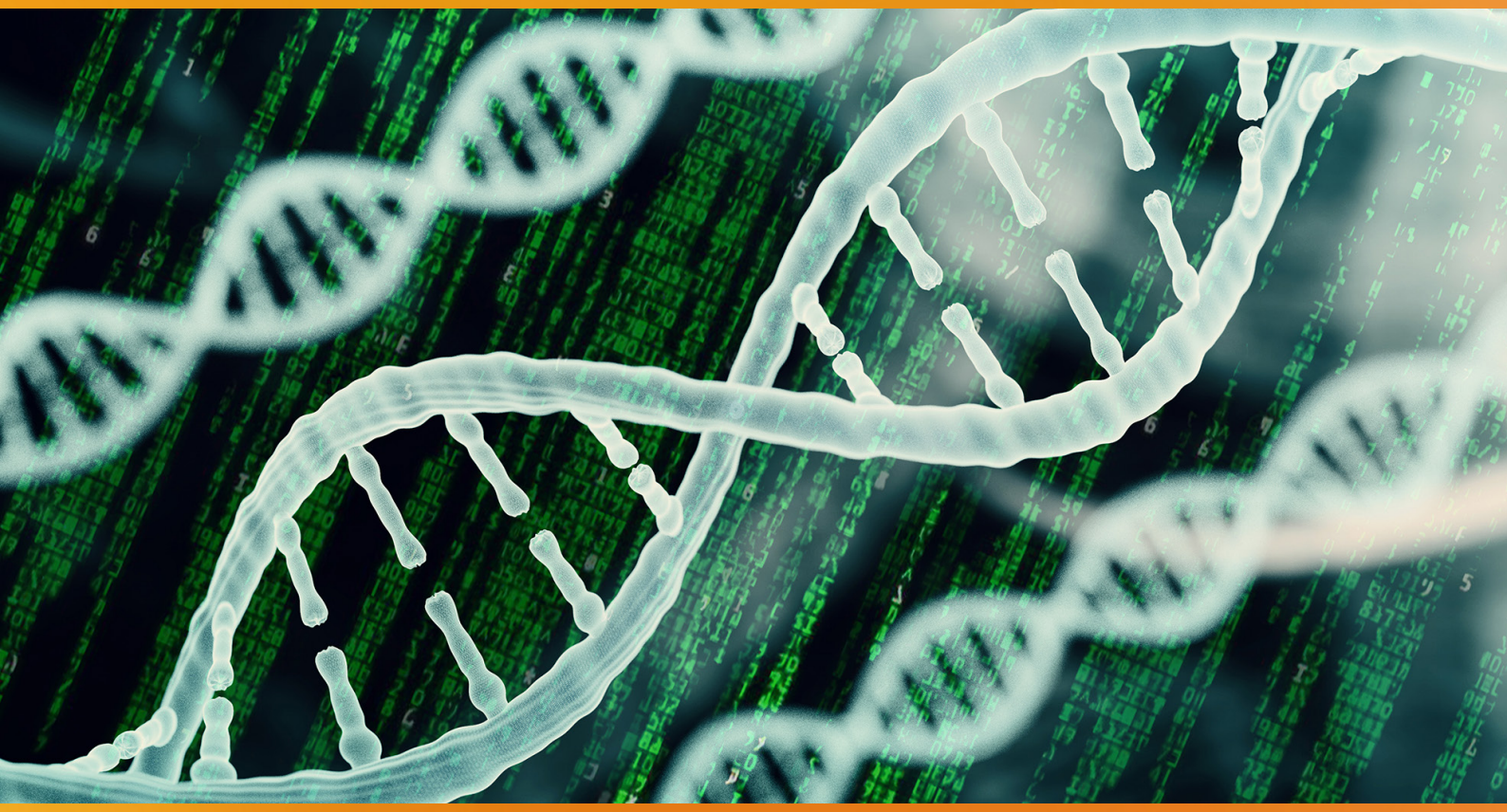
illumina®



How to get accurate insights from your metagenomics data – advanced microbiome bioinformatics

While metagenomic sequencing enables high-resolution access to the microbiome, the impact of bioinformatic profiling tools on interpreting metagenomic data to make it usable is often underestimated. Dr David Wood, Head of Bioinformatics Operations at Microba Life Sciences, discusses some of the common pitfalls of bioinformatics for microbiome profiling, and how these can be overcome to drive microbiome sciences forward.

The challenge of accurate profiling using metagenomic data

“Microbiome profiling aims to provide an accurate picture of the composition and functional potential of a microbial community. This can be done by comparing metagenomic reads from a sample to previously characterised sequence data in a reference database,” explains Dr Wood. “However, many species share genomic features with other species in their genus, and even across genera. This can cause incorrect sequence alignments, resulting in incorrect (false positive and false negative) species being reported.”

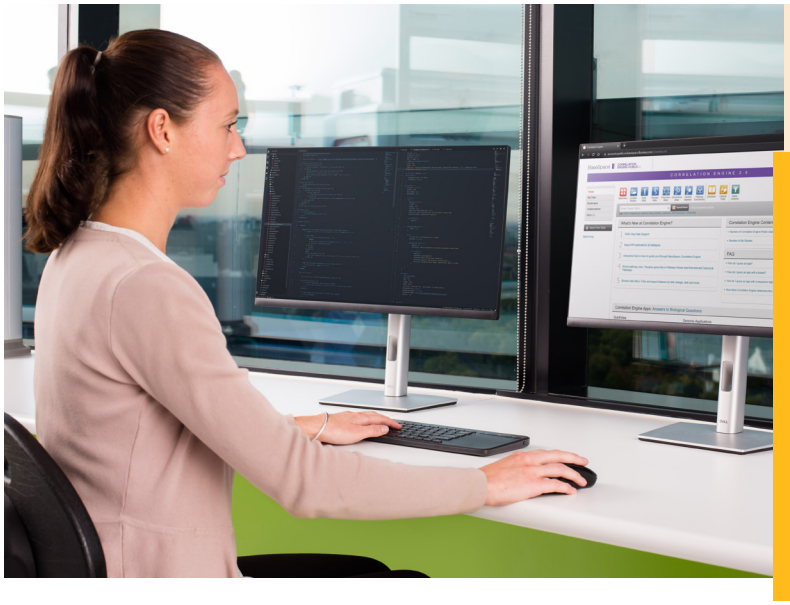
A major advantage of metagenomic sequencing over earlier sequencing approaches (e.g., 16S rRNA) is the ability to gain functional insights from microbial communities by measuring the genes in a sample that are known to perform functions. Of course, this ability also adds a layer of complexity to bioinformatic processing.

“Functional profiles are challenged by the accuracy and completeness of gene annotations,” says Dr Wood. “Many genes remain poorly annotated and therefore have no known function. There is also the difficulty of linking reported functions to their encoding species. These challenges need to be met by bioinformatic tools that are optimised for microbiome profiling.”

Why a comprehensive reference genome database is critical for accurate profiling

Dr Wood explains that having a comprehensive genome database (both at the species and strain level) is critical for addressing the challenges associated with microbiome profiling.

“A more complete database will enable better quality read alignments. This will increase both your recall, meaning the fraction of correctly identified species within your sample, and precision, or the fraction of species identified that are correct. If genomes are missing in the reference database, then the sequencing reads from your sample are more likely to misalign to an incorrect species.”



“Microbiome profiling aims to provide an accurate picture of the composition and functional potential of a microbial community.”

What about species that have yet to be added to a database?

Despite large-scale cultivation efforts, many microorganisms (including from the largely anaerobic human gut) remain uncultured and lack good quality reference genomes. This can be addressed by reconstructing high-quality metagenome-assembled genomes (MAGs) from metagenomic data, also known as de novo assembly.

"This is often the only viable way to obtain these reference genomes," says Dr Wood. He notes that MAGs can be added to your database when you taxonomically profile, increasing your coverage and decreasing both your false negative and false positive rates. If MAGs aren"t included in a reference database, a profiler may be more likely to assign sequences to another closely related species present in their database.

"The recovery of MAGs has drastically expanded the availability of draft genomes from uncultured organisms," he adds. "Our evidence indicates that with suitable quality control procedures, these MAGs can be very reliable. We control for genome completeness and contamination using tools such as CheckM, and ensure that technical artifacts that can arise in MAG construction are managed throughout our pipelines. Until high throughput culturing methodologies are broadly available, we strongly advocate mining for MAGs."

What to look for when choosing a profiler

Using a profiler and database without demonstrated performance can increase your risk of missing signals that are important to your experiment (false negatives), meaning potentially missed biomarkers for diagnostic or therapeutic applications. It can also lead to high rates of erroneous signals or false positives, causing you to waste effort and resources chasing leads where there was no true signal to begin with.

"To avoid this pitfall," Dr Wood explains, "researchers should always evaluate profilers independently. You want to be confident that a profiler both accurately reports species that are present and has demonstrated precision to minimise false reporting of species."

He adds that mock communities are a good tool for assessing profilers, but not all mock communities are equal. "Make sure you use mocks that are properly representative of the community to be profiled. This includes species and strains that are not represented in your database so that you can determine how your profiler performs with previously uncharacterised species. We have published a detailed set of freely available mock communities that researchers can use to evaluate profilers on their own."

"The recovery of MAGs has drastically expanded the availability of draft genomes from uncultured organisms."

When it comes to evaluating a profiler"s performance, Dr Wood says that precision and recall are two of the most important metrics.

"Look for a profiler that has robustly demonstrated a high F1 score, which combines precision and recall (also known as sensitivity) to provide a measure of accuracy. There is typically a trade-off between precision and recall, so using a profiler with an optimised balance is important for overall performance¬. You also want to consider the detection limit, or the point at which a target false discovery rate (FDR) can be achieved. Your target FDR will depend on your research goals; in most cases, an FDR of <0.5% is necessary. However, for clinical purposes such as infectious disease detection, this may need to be even lower."

Using a profiler that performs well for each of these metrics using mock communities will give you more confidence when analysing samples of your own.

"Recognising the major limitations in existing bioinformatic profilers, we developed the Microba Community Profiler, or MCP"

What are Microba's tools and how do they perform?

"Recognising the major limitations in existing bioinformatic profilers, we developed the Microba Community Profiler, or MCP, which uses a genome alignment approach to comprehensively profile microbiome samples," Dr Wood explains. "One of the major strengths of MCP is the use of the Microba Genome Database, or MGDB, which consists of over 73,600 dereplicated genomes. Genomes in MGDB are taxonomically classified based on the Genome Taxonomy Database, which provides improved taxonomic resolution relative to the NCBI taxonomy."

"When we benchmarked MCP against other commonly used metagenomic profilers using 140 in silico mock microbial communities, we found that MCP had an F1 score of 0.97 (out of 1), compared to scores of 0.76–0.91 among other profilers," he adds.

MCP favours a slight increase in the percent of false negatives in order to substantially reduce the percent of false positives, resulting in 4–16 times less false positives than other profilers. It also assigns at least 25% more DNA reads per sample than other evaluated profilers when profiling human gut microbiome samples. This is partly because MGDB contains many uncultured gut microbiome species that are absent from other reference databases.

Dr Wood notes that MCP is continually being improved, particularly as more genomes are added to MGDB.

"In addition to MCP, we also have a range of other tools, including subspecies and functional profilers, isolate characterisation, and an infectious disease diagnostic profiler, which can be used in the clinic to detect pathogens, virulence factors and antimicrobial resistance genes," he adds.

So, is it important to consider bioinformatics in your study design?

“Bioinformatics should always be a key consideration when scoping study design,” Dr Wood emphasises. “First and foremost, identify your research questions and consider if you have enough samples and if you are going to generate enough data to answer those questions. Then think about what sample types you are analysing: is the microbiome composition of these sample types sufficiently represented in reference databases? What level of resolution of markers do you need to attain? Ensure you choose a profiler that has demonstrated performance for your type of sample and research question. Finally, ensure that you have budgeted for both computational and bioinformatics analysis costs.”

As a final piece of advice, Dr Wood adds, “We recommend you leverage a domain expert to contribute to your exploratory and statistical analysis plan, as there are additional features unique to microbiome data that need to be considered during tertiary analysis. This will help you get the most accurate signals out of your data.”

Getting the profiles right will enable robust discoveries from your microbiome study. While there are many complex factors to consider, you don”t need to do it alone. Learn how Microba can help.

Further reading

1. Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36(10):996-1004.
2. Parks DH, Rigato F, Vera-Wolf P, Krause L, Hugenholtz P, Tyson GW, Wood DLA. Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets from the Human Gut Microbiome. Front Microbiol. 2021;12:643682.

Illumina & Microba: Empowering microbiome research

Microba Life Sciences and Illumina work together to accelerate microbiome research. Combining Microba’s high-quality proprietary gut microbiome Analysis Platform with Illumina’s revolutionary Next Generation Sequencing tools, researchers have access to world-leading, accurate metagenomic data to drive new discovery from the microbiome.

illumina®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel | techsupport@illumina.com | www.illumina.com