Translating microbial community profiles into meaningful scientific insights

Tertiary analysis for microbiome research

With the appropriate laboratory and bioinformatic tools, metagenomic sequencing enables high resolution access to the microbiome. However, the resulting microbiome profiles have characteristics that complicate the statistical analysis required to translate these data into scientific discoveries. This means that nuanced analysis approaches are required to support robust discoveries. Dr Martha Cooper, Senior Data Scientist at Microba, discusses some of the common hurdles in the tertiary analysis of metagenomics data and how to overcome them.

Common challenges of tertiary analysis for microbiome research

"Metagenomics is incredibly powerful because of the large amount of data it can provide," says Dr Cooper. "However, microbiome data present intrinsic characteristics that pose challenges for statistical analysis. First, metagenomic datasets are enormous, and the number of features we're able to measure, species or microbial functions, are typically much larger than the number of samples. It can be difficult to extract and interpret meaningful patterns from such datasets, especially when many features are highly correlated, as we often see in microbiome data."

Dr Cooper further explains that microbiome data is compositional, as the abundance of each species or function is usually reported relative to the total abundance within a sample.

This is challenging because the abundance of each of the individual markers is not independent from the abundance of others.

"Finally, the distribution of microbiome community profiles is also not what is expected by many standard statistical tests," she adds. "Metagenomics data is sparse, with a large number of zero values, and skewed, with a high number of data points with low relative abundance values and fewer observations with extremely high values."

Metagenomic data is not only enormous, but also inherently noisy, and analysis of metagenomics data is further complicated by technical variation (e.g. different laboratory processes or sequencing platforms), inter-individual variation and the large number of potential confounding factors that can impact gut microbiome composition. We need tools to see through the noise and instead uncover meaningful biological variation associated with the effect of an intervention or the phenotype of interest.

"Finally, the distribution of microbiome community profiles is also not what is expected by many standard statistical tests"

For Research Use Only. Not for use in diagnostic procedures.

M-APJ-00216 | 2

What happens if we fail to address these challenges?

"The goal of tertiary analysis of microbiome data is often to identify members or functions of a microbial community that are either associated with a disease or experimental condition, or that respond to an intervention," says Dr Cooper. The results of microbiome studies frequently lead to investment in larger research and development programs, including biomarker discovery, drug development, and clinical trials.
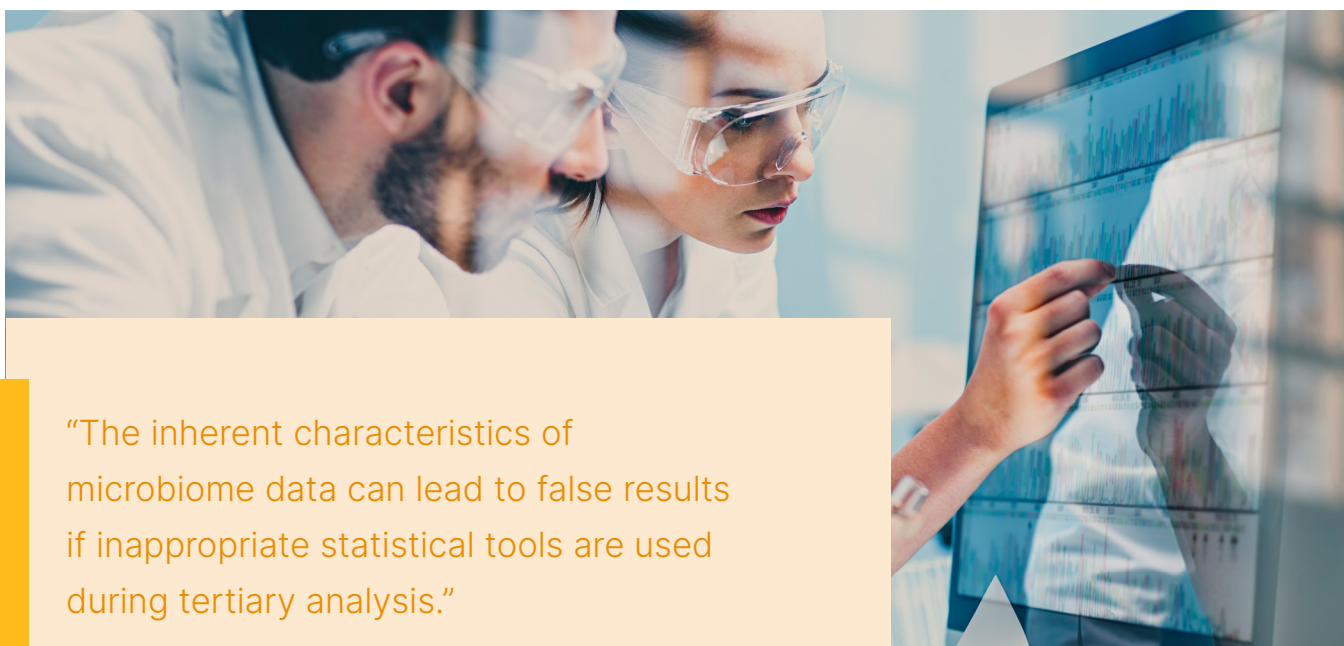
"However, the inherent characteristics of microbiome data can lead to false results if inappropriate statistical tools are used during tertiary analysis. This includes false positives, which can cause researchers to pursue erroneous leads, and false negatives, which could miss valuable biomarkers or therapeutic opportunities. Failing to effectively address the statistical challenges of microbiome data analysis could result in the inability to distinguish the genuine biological signals in your data from mere noise, risking the success of large research programs."

How can we ensure analysis success?

Dr Cooper explains that a study's success starts with a clearly defined hypothesis and a good experimental design. An experiment must include appropriate cohorts, sample sizes sufficient to achieve the study goals, study design that minimises bias, and appropriate data quality control. Ideally, results from early discovery studies are replicated in an independent cohort.

"Once high quality data from a well-designed experiment has been generated, it is important to first perform appropriate data QC and filtering pre-processing steps, then to choose statistical methods that are appropriate for the hypothesis and that address the unique challenges of microbiome data."

Dr Cooper explains that several statistical analysis approaches can be used to extract meaningful insights from microbiome data. "The first approach uses data transformations to help metagenomic data look more like the type of data expected by standard, univariate parametric statistical tests. These tests are simple to perform, but the validity of the results will depend on the success of the transformations, which can be tricky to interpret."

"The inherent characteristics of microbiome data can lead to false results if inappropriate statistical tools are used during tertiary analysis."

For Research Use Only. Not for use in diagnostic procedures.

M-APJ-00216 | 3

"Another approach uses statistical tests that make less assumptions about the data. For example, nonparametric tests are more robust against skewed data, outliers and small sample sizes, which are often seen in microbiome data. However, such tests are less likely to detect small differences, which are also common in microbiome studies."

"Finally, we can use statistical tests that are specifically designed for metagenomic data. This is an active area of research and new tools are coming out all the time to address not only the challenges discussed above, but many more."

While there are many options available, it is important to be aware of the pros and cons of each of these approaches, and to keep their specific limitations in mind when interpreting results.

How Microba addresses these challenges

"At Microba, we prioritise developing clear research hypotheses and well-designed experiments to maximise the chance of analysis success," says Dr Cooper. "We apply rigid quality control prior to statistical analysis, identify potentially problematic samples and account for biases using cohort matching or by including confounders during statistical testing."

"While there are many statistical tools to address the challenges of microbiome data, none are perfect solutions. Even more, there is no one size fits all approach, and the best solution may be different from one study to the next based on the researcher's specific hypothesis."

Microba takes a consensus approach to statistical analysis, Dr Cooper adds. "We apply multiple statistical tools and make sure that we get robust results that remain consistent across these methods. Meaningful signals should be strong and reproducible and where possible, we replicate results in independent cohorts. We also apply a combination of univariate and multivariate statistics, and machine-learning approaches, which can identify complex and non-linear signals in microbiome data."

Summary

Application of the right statistical tools for metagenomics analysis can enable extraction of genuine, tangible signals from your study. However, it is critical to be aware of the nuances of microbiome data and how to address them when designing your experiment and statistical analysis plan, and how to apply and interpret the various statistical approaches available.

Looking for help with your tertiary analysis plan? Learn how Microba can help.

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®

For Research Use Only. Not for use in diagnostic procedures.

M-APJ-00216 | 4