Emu software uses common gene to profile microbial communities

June 30 2022

Rice University computer scientists introduced Emu, an algorithm that uses long reads of genomes to identify the species of bacteria in a community. The program could simplify sorting harmful from helpful bacteria in microbiomes like those in the gut or in agriculture and the environment. Credit: Kristen Curry/Rice University

Part of a gene is better than none when identifying a species of microbe. But for Rice University computer scientists, part was not nearly enough in their pursuit of a program to identify all the species in a microbiome.

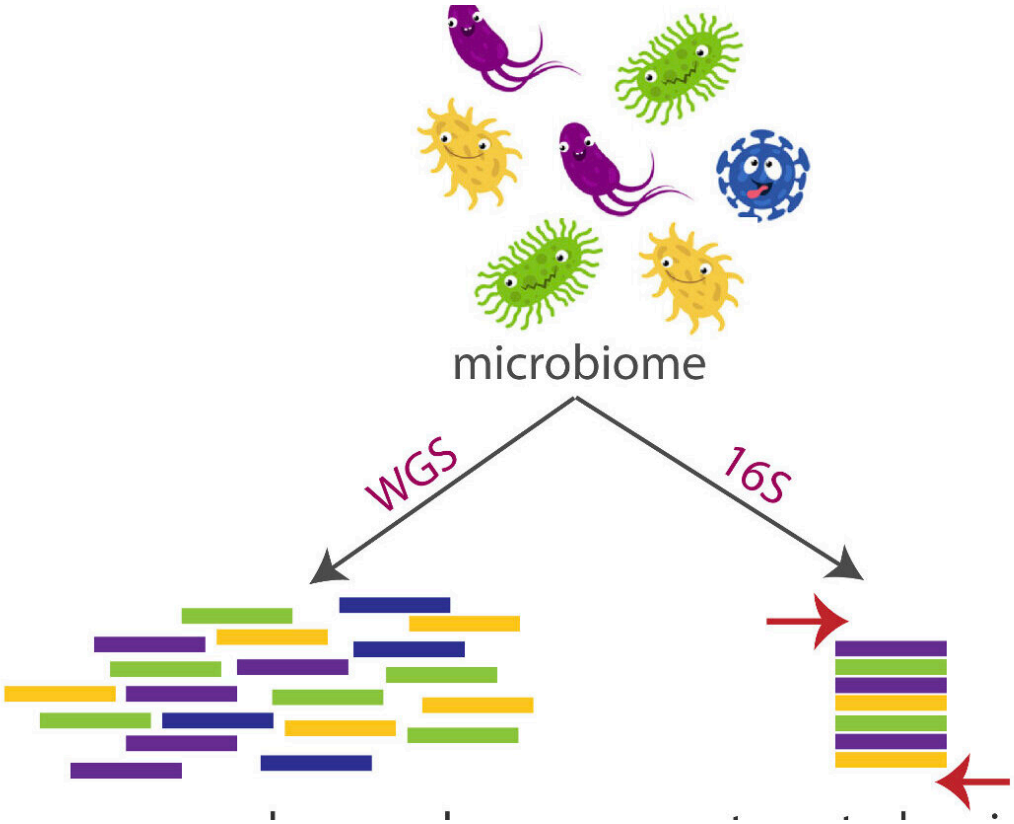
Emu, their microbial community profiling software, effectively identifies bacterial species by leveraging long DNA sequences that span the entire length of the gene under study.

The Emu project led by computer scientist Todd Treangen and graduate student Kristen Curry of Rice's George R. Brown School of Engineering facilitates the analysis of a key gene microbiome researchers use to sort out species of bacteria that could be harmful—or helpful—to humans and the environment.

Their target, 16S, is a subunit of the rRNA (ribosomal ribonucleic acid) gene, whose usage was pioneered by Carl Woese in 1977. This region is highly conserved in bacteria and archaea and also contains variable regions that are critical for separating distinct genera and species.

"It's commonly used for microbiome analysis because it's present in all bacteria and most archaea," said Curry, in her third year in the Treangen group. "Because of that, there are regions that have been conserved over the years that make it easy to target. In DNA sequencing, we need parts of it to be the same in all bacteria so we know what to look for, and then we need parts to be different so we can tell bacteria apart."

The Rice team's study, with collaborators in Germany and at the Houston Methodist Research Institute, Baylor College of Medicine and Texas Children's Hospital, appears in the journal *Nature Methods*.



A schematic illustrates the relative simplicity of more random shotgun sequencing (WGS) and Emu, a technique developed at Rice University to identify bacterial species by leveraging long DNA sequences of the common 16S gene, which is highly conserved in bacteria. The program could simplify sorting harmful from helpful bacteria in microbiomes. Credit: Kristen Curry/Rice University

"Years ago we tended to focus on bad bacteria—or what we thought was bad—and we didn't really care about the others," Curry said. "But there's been a shift in the last 20 years to where we think maybe some of those other bacteria hanging out mean something.

"That's what we refer to as the microbiome, all the microscopic organisms in an environment," she said. "Commonly studied

environments include water, soil and the intestinal tract, and microbes have shown to affect crops, carbon sequestration and human health."

Emu, the name drawn from its task of "expectation-maximization," analyzes full-length 16S sequences from bacteria processed by an Oxford Nanopore MinION handheld sequencer and uses sophisticated error correction to identify species based upon nine distinct "hypervariable regions."

"With previous technology we could only read part of the 16S gene," Curry explained. "It has roughly 1,500 base pairs, and with short-read sequencing you can only sequence up to 25%-30% of this gene. However, you really need the full-length gene to attain species-level precision."

But even the newest technology isn't perfect, allowing errors to slip into sequences.

"While error rates have dropped in recent years, they can still have up to 10% error inside an individual DNA sequence, while species can be separated by a handful of differences in their 16S gene" said Treangen, an assistant professor of computer science who specializes in tracking infectious disease. "Distinguishing sequencing error from true differences represented the main computational challenge of this research project.

"One issue is that a lot of the error is nonrandom, meaning it can occur repeatedly in specific positions, and then start to look like true differences instead of sequencing error," he said.

"Another issue is there can be thousands of bacterial species in a given sample, creating a complex mixture of microbes that can exist at abundances well below the sequencing error rate," Treangen said. "This

means we can't simply rely on ad hoc cutoffs to distinguish signal from error."

Instead, Emu learns to distinguish between signal and error by comparing a multitude of long sequences, first against a template and then against each other, refining its error-correction iteratively as it profiles microbial communities. In the performed experiments, false positives dropped significantly in Emu in comparison to other approaches when analyzing the same data sets.

"Long-reads represent a disruptive technology for microbiome research," Treangen said. "The goal of Emu was to leverage all of the information contained across the full-length 16S gene, without masking anything, to see if we could achieve more accurate genus- or species-level calls. And that's exactly what we accomplished with Emu, thanks to a fruitful, multidisciplinary collaborative effort."

Alexander Dilthey, a professor of genomic microbiology and immunity at Heinrich Heine University, Düsseldorf, Germany, is co-corresponding author of the paper.

More information: Kristen Curry, Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data, *Nature Methods* (2022). DOI: 10.1038/s41592-022-01520-4.
www.nature.com/articles/s41592-022-01520-4

Provided by Rice University

profile.html