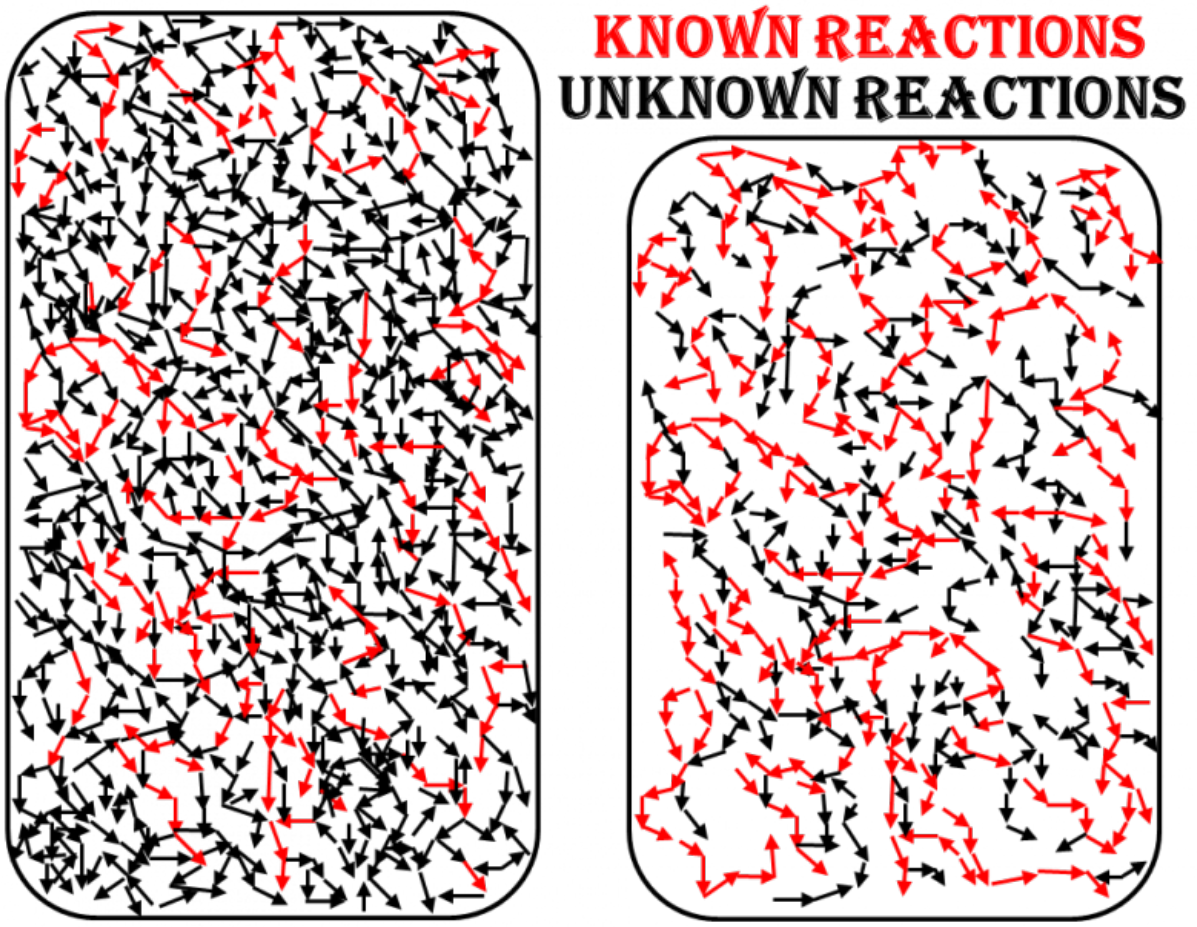
Scientists find way to remove 'noise' from big data in metabolomics study

September 19 2017



Metabolism is complicated. The good news is that it might not be as complicated as previously thought. New research from scientists at Washington University supports a picture more like the one on the right. Credit: Gary Patti lab

Not long ago, scientists placed wagers on the number of genes in the human genome. Some bets ranged upward of 100,000 genes being present. Once the human genome sequence was completed, a project led in part by the McDonnell Genome Institute at Washington University School of Medicine in St. Louis, even the lowest guess of 25,947 proved to be above the true number.

Now, nearly 15 years later, scientists at Washington University are seeing a reminiscent trend in the newest type of big data known as metabolomics. They estimate that the number of metabolites present in a data set could be 90 percent smaller than previously estimated.

The study was published online Sept. 15 in *Analytical Chemistry*.

Like its genomic predecessor, metabolomics seeks to profile all of the metabolites present in a sample. Unlike genes, however, metabolites are not made from common building blocks and are much more chemically diverse. Familiar metabolites include molecules such as glucose and cholesterol, many of which are a product of diet. Thus, trying to pin down the exact number of metabolites in humans has been a tough challenge. Because of its strong nutritional dependence, some scientists have argued that it's not even the relevant question to be asking.

There has been interest in measuring metabolites for nearly as long as there has been interest in human health. Analysis of glucose in diabetes probably dates back centuries. Handfuls of other metabolites have been used to diagnose diseases broadly referred to as "inborn errors of metabolism" since the 1960s. Metabolomics tries to measure all of these metabolites, and more. The question is: How many more are there?

The scene for metabolomics was set with the advent of sophisticated devices called mass spectrometers. These instruments are like tiny scales that can measure the weights of molecules, such as sugars. By using

databases and computational algorithms, scientists can convert measured weights into compound names, like glucose.

A decade ago, when metabolomics started to become mainstream, scientists were surprised to discover that the number of signals in a typical metabolomics experiment greatly exceeds the number of known metabolites in biochemistry textbooks. Said Gary Patti, associate professor of chemistry in Arts & Sciences and senior author of the study: "Of course, the knee-jerk reaction is to assume that most of the signals that do not return matches in databases correspond to unknown metabolites that have never been reported before."

The implications of such an assumption are major: tens of thousands of metabolites remain to be discovered, an order of magnitude more than what is included on your common wall chart of comprehensive metabolism (see image below).

"It is routine to detect tens of thousands of signals in metabolomics, but only 1,000 to 2,000 have been identified in any experiment to date," said Nathaniel Mahieu, a postdoctoral fellow in Patti's lab, who led the study.

Said Patti: "The million dollar question is: How many metabolites do all of these metabolomic signals actually correspond to?"

Mahieu and Patti, who was announced last week as an awardee of an eight-year, $5.85 million inaugural grant in environmental health from the National Institutes of Health, developed new experimental and computational approaches to interrogate metabolomics data sets. They arrived at a striking conclusion. They found that the actual number of metabolites in a typical metabolomics analysis may be one-tenth as large as previously suggested, with much of the data coming from "noise." Thousands of signals arise from contamination, artifacts, and something called "degeneracy"—say, when one metabolite shows up as many

different signals. The research team found that some metabolites show up as more than 150 signals.

"It turns out that more than 90 percent of the signals we see in E. coli data are essentially noise," Mahieu said. "This greatly reduces the number of unknown metabolites that we thought we were detecting."

"I think this is sort of a wake-up call, a reality check if you will, on what metabolomics suggests about the size of the metabolome," Patti said. "I believe it is a good thing. It means we're a lot closer to understanding metabolism than we probably thought we were."

As for the next step, Patti's lab intends to extend their techniques to human samples.

"The ultimate goal is to do analogous experiments for humans," Patti said. "Our work here is an important step forward."

So what do all of these noise signals mean to other scientists performing metabolomics? The Patti lab has started curating what they term "reference data sets" in a database called creDBle (creDBle.wustl.edu). They hope that it will facilitate experiments for other scientists performing metabolomics.

"The way metabolomics is currently performed is terribly inefficient. We waste a lot of time trying to interpret signals that provide minimal biological insight," Mahieu said. "We hope that these reference data sets in creDBle will help prevent scientists from having to identify the same noise signals over and over again now that we have annotated them."

More information: Nathaniel G. Mahieu et al. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites, *Analytical Chemistry* (2017). DOI:

10.1021/acs.analchem.7b02380