Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Are microbiome studies ready for hypothesis-driven research?

Anupriya Tripathi[1], Clarisse Marotz[1], Antonio Gonzalez[1], Yoshiki Vázquez-Baeza[1], Se Jin Song[1], Amina Bouslimani[5], Daniel McDonald[1], Qiyun Zhu[1], Jon G Sanders[1], Larry Smarr[2,3,6], Pieter C Dorrestein[1,3,4,5] and Rob Knight[1,2,3]

Hypothesis-driven research has led to many scientific advances, but hypotheses cannot be tested in isolation: rather, they require a framework of aggregated scientific knowledge to allow questions to be posed meaningfully. This framework is largely still lacking in microbiome studies, and the only way to create it is by discovery-driven, tool-driven, and standards-driven research projects. Here we illustrate these issues using several such non-hypothesis-driven projects from our own laboratories, including spatial mapping, the American Gut Project, the Earth Microbiome Project (which is an umbrella project integrating many smaller hypothesis-driven projects), and the knowledgebase-driven tools GNPS and Qiita. We argue that an investment of community resources in infrastructure tasks, and in the controls and standards that underpin them, will greatly enhance the investment in hypothesis-driven research programs.

Addresses
[1] Department of Pediatrics, University of California San Diego, La Jolla, CA, USA
[2] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA
[3] Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA
[4] Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, CA, USA
[5] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
[6] California Institute for Telecommunications and Information Technology, University of California San Diego, La Jolla, CA, USA

Corresponding author: Knight, Rob (robknight@ucsd.edu)

Current Opinion in Microbiology 2018, **44**:61–69

This review comes from a themed issue on **Microbiotia**

Edited by **Jeroen Raes**

https://doi.org/10.1016/j.mib.2018.07.002

Introduction

Microbiome research is making dramatic progress, with thousands of papers now published each year linking specific microbes and/or host–microbe co-metabolites to specific diseases, physiological properties, or environmental parameters. Much of this research is performed in a traditional, hypothesis-driven way, or at least presented as a rational reconstruction that fits this model, much as Darwin re-wrote much of his discovery-driven work as hypothesis driven to increase its respectability under the influence of contemporary philosophers of science such as William Whewell [1•]. However, it should be noted that hypothesis-driven science was not always so respectable — Isaac Newton famously wrote '*Hypotheses non fingo*', or 'I feign no hypotheses', in an essay appended to the second edition of the *Principia* [2] — so the tradition of modifying how science is framed to meet respectability criteria dates back at least 300 years. What can be framed as a testable hypothesis suffers important limitations based on what we can measure and what we already know.

Ten years ago Chris Anderson, editor of Wired magazine, set off an international debate with his article 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete' [3]. The idea was that with enough data, hypotheses will emerge ('Let the data speak for itself') has become widely discussed in the rapidly growing data science profession. A thoughtful review of this topic was written in EMBO Reports in 2015 — 'Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science' [4•]. As the author points out:

> 'Francis Bacon, the 'father of the scientific method' himself, in his Novum Organum (1620), argued that scientific knowledge should not be based on preconceived notions but on experimental data. Deductive reasoning, he argued, is eventually limited because setting a premise in advance of an experiment would constrain the reasoning so as to match that premise. Instead, he advocated a bottom-up approach: In contrast to deductive reasoning, which has dominated science since Aristotle, inductive reasoning should be based on facts to

generalize their meaning, drawing inferences from observations and data.'

We recently reviewed experimental design considerations for traditional hypothesis-driven microbiome studies elsewhere [5,6•], and do not discuss these issues further in this review. Here we describe the danger of jumping too soon into hypothesis testing, and describe the need for four major categories of non-hypothesis-driven research: better spatial and abstract maps, better tools, and better standards. Given space constraints, we illustrate these primarily using the American Gut Project [7••], the Earth Microbiome Project [8••], and tools we developed in our laboratories.

The challenge of unknown unknowns

In microbiome research, a recurring challenge has been that factors intuitively suspected to drive differences in the microbiome are less important than other, more surprising factors. For example, sex has a small impact on microbiomes across the human body [9,10••] and has a much weaker effect than many other variables such as age (even within adults), or the time of year the sample was collected [11,12]. However, sex is far more frequently reported than time of year. Similarly, although long-term dietary habits are correlated with the overall composition of the human microbiome within and between populations [7••,13••,14•,15,16••], and dietary changes over months can lead to changes in overall microbiome composition larger than the differences between arbitrarily chosen individuals [17•,18], but short-term changes have transient effects smaller than typical differences between individuals [14•,19•]. However, many studies focus on short-term rather than long-term diet. Perhaps even more surprisingly, factors such as temperature and pH have much smaller impacts on environmental microbiomes than salinity [8••,20], and even the saline versus non-saline difference is much smaller than the host-associated versus free-living difference [8••,21•]. Samples from different parts of the same person's body differ more from one another in their overall microbial communities than radically different free-living microbial communities, such as soils versus oceans [8••]. Differences of this magnitude can also occur within the gut of a single person, with sufficiently large perturbation [7••].

Because factors of large effect are often unknown and unreported, studies testing hypotheses concerning intuitively obvious factors of small effect are often subject to important confounding variables, that, when uncovered, prompt complete reinterpretation of the study. For example, suppose an investigator is unaware that cage effects are important in the microbiome [22], and profiles microbiomes in two cages each of two different genotypes of mice. The results will likely be driven by which cages happens to resemble each other more closely. If the variable of cage is not measured, or not tested in an unsupervised model, this important confounding variable will likely remain undiscovered, and the interpretation of the experiment entirely incorrect.

Similarly, a frequent practice is to discard unannotated microbes or unannotated molecules, focusing on the subset of microbes or molecules that can be matched to an existing database. Because databases of both microbes and molecules are heavily biased (microbes, by studies of known pathogens that come from only a few taxonomic groups, and molecules, by commercially available compounds), the entities that best discriminate among classes of samples may be lost in the analysis: often, only 60% of sequences and 2% of molecular features from an untargeted metabolomics experiment can be annotated by existing references [23,24]. However, a rational reconstruction of why the annotatable microbes or molecules are plausibly connected to a phenotype of interest can frequently be developed, especially given the characteristics of these highly multivariate datasets that can lead to high false discovery rates when the true number of implicitly tested hypotheses is considered [25••].

The need for spatial maps

An important metaphor in science and information visualization is the idea of the map. As data volumes increase, it is frequent that the main research activity in a field moves from tests of hypotheses of differences in individual variables among sites, to tests of these hypotheses with replicates at each site, to spatially or temporally explicit sampling, to detailed spatial maps that reveal otherwise unsuspected patterns. This progression has occurred in 16S rRNA amplicon-based microbiome studies over the past decade [8••,26], and increasingly characterizes mass spectrometry-based metabolome studies over the past four years [27,28•,29,30•,31,32].

The value of spatial maps is so self-evident that the results are often cursed by obviousness. For example, the finding that metabolomes cluster by individual, as revealed by principal coordinates analysis (PCoA), is interesting (Figure 1a). However, the finding that a given molecule such as lauryl sulphate (m/z 355.219) is distributed across the body of one of the two individuals, but is absent from the other individual is obvious (Figure 1b), especially when subject A, who is male, reports using the skin care product Nivea for Men, the source of the molecule [28•]. Similarly, the finding that samples from four individuals differ significantly in levels of specific purines between and within subjects might well prompt further investigation. However, a spatial map with dense sampling of the same individuals (Figure 1c) makes it obvious that the molecule is something that is touched and consumed, and sometimes spilled, allowing one to guess that it is caffeine; similarly, the spatial map reveals