Applied & Translational Genomics 10 (2016) 10–15

Contents lists available at ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

Metabolic modeling with Big Data and the gut microbiome

Jaeyun Sung [a,1], Vanessa Hale [b,c,1], Annette C. Merkel [b], Pan-Jun Kim [a,d], Nicholas Chia [b,c,e,*]

[a] Asia Pacific Center for Theoretical Physics, Pohang, Gyeongbuk 37673, Republic of Korea
[b] Center for Individualized Medicine, Microbiome Program, Mayo Clinic, Rochester, MN 55905, USA
[c] Department of Surgery, Mayo Clinic, Rochester, MN 55905, USA
[d] Department of Physics, Pohang University of Science and Technology, Pohang, Gyeongbuk 37673, Republic of Korea
[e] Department of Biomedical Engineering, Mayo College, Rochester, MN 55905, USA

ARTICLE INFO

Article history:
Received 10 August 2015
Received in revised form 19 January 2016
Accepted 2 February 2016

Keywords:
Gut microbiome
Metabolic modeling
Microbial community
Data integration
Big Data

ABSTRACT

The recent advances in high-throughput omics technologies have enabled researchers to explore the intricacies of the human microbiome. On the clinical front, the gut microbial community has been the focus of many biomarker-discovery studies. While the recent deluge of high-throughput data in microbiome research has been vastly informative and groundbreaking, we have yet to capture the full potential of omics-based approaches. Realizing the promise of multi-omics data will require integration of disparate omics data, as well as a biologically relevant, mechanistic framework – or metabolic model – on which to overlay these data. Also, a new paradigm for metabolic model evaluation is necessary. Herein, we outline the need for multi-omics data integration, as well as the accompanying challenges. Furthermore, we present a framework for characterizing the ecology of the gut microbiome based on metabolic network modeling.

Contents

1. Introduction

The promise of the Big Data revolution has yielded an ever-increasing array of data and data types in many fields. In the medical field, the sequencing of the human genome in 2003 opened the door to truly individualized medicine, tailored to our genetic predispositions and risk factors (Collins et al., 2003). The first manifestations of the Big Data promise in medicine were necessarily surveys to identify biological markers of disease risk. While this resulted in databases upon databases of genetic events that explained risk behind hundreds of diseases, we quickly learned that genetics alone was not able to provide a full understanding of many health conditions (Lander, 2011). Researchers began to examine other factors, including the role of environmental influences such as the microbiome (Bultman, 2013; Zackular et al., 2013).

In 2008, the Human Microbiome Project (HMP) was established to characterize the role of human-associated microbial communities in human health and disease (Methé et al., 2012; The Human Microbiome Project, 2012). Efforts led by the HMP consortium thus far have yielded numerous insights regarding the microbial composition of the human body and the ecological structure and function of the

Abbreviations: GEMs, genome scale metabolic models; COMM, community-scale metabolic modeling; HMP, Human Microbiome Project.

* Corresponding author at: 200 First St. SW, Rochester, MN 55905, USA.
E-mail address: chia.nicholas@mayo.edu (N. Chia).

[1] These authors contributed equally.

http://dx.doi.org/10.1016/j.atg.2016.02.001

human microbiome. However, a shift from this "profiling" paradigm to one of mechanistic examination is now both warranted and feasible through the integration of multi-omics data onto a framework based on biomolecular pathways and networks.

The gut microbial community is increasingly well-characterized by various omics technologies – metagenomics, metatranscriptomics, metabolomics, metaproteomics – and offers much promise for data integration within a mechanistic framework (Erickson et al., 2012; Haiser et al., 2013a, 2013b; Weir et al., 2013). Gut microbes act as chemical transformers, converting host-acquired or host-produced nutrients into a milieu of metabolites (Lee and Hase, 2014). At the same time, the structure and function of the microbial community respond to changes in host diet or physiology (David et al., 2014; Kashyap et al., 2013; Liou et al., 2013), making microbes both modulators and reflections of the gut environment.

The gut microbiome contains over 3 million genes, or approximately 150-fold more than the human genome (Qin et al., 2010); thus, it becomes virtually impossible to obtain more independent samples than there are measured values within one individual's microbiome. The large data sets generated by the most recent omics technologies call for new methods of analysis. No longer can we afford to use a paradigm of statistical power where our insight dwindles with the amount of data we collect. Instead, we should rely on the fact that these variables are not independent of one another and therefore establish a more practical model for assessing the role of the microbiome.

A systems approach that utilizes metabolic networks may offer a potential solution. Network reconstruction is one such means of creating a scaffold for synthesizing multiple data types (Feist and Palsson, 2008; Lee et al., 2012; Reed et al., 2006; Töpfer et al., 2015). Metabolic models are composed of a collection of individual chemical reactions that are governed by the fundamental laws of mass conservation and thermodynamics. These models represent large-scale complex cellular dynamics and imply a network whose mechanistic chain of events can be computed to produce an outcome. Models are capable of converting large amounts of data – genetic, metabolic, biochemical – into phenotypes and interactions. The value of metabolic modeling for understanding the complex environment of the gut microbiome is in resolving biochemical relationships within and between microbial species and potentially predicting the effect of ecosystem-wide perturbations, such as antibiotics or pathogen invasion. There have been many recent efforts to model metabolic processes within microbial communities (Heinken and Thiele, 2015; Henry et al., 2009). However, the wealth of data available through multiple omics technologies remains underutilized in these models.

In this review, we discuss the promises and limitations offered by current mathematical paradigms for integrating disparate, yet complementary omics data, while pointing out the challenges that remain to be resolved. Finally, we offer our viewpoint on the need for an updated network-aware mathematical framework for statistical power — one that synthesizes multiple channels of information into a biological picture.

2. The Big Data paradox

The mathematical formalization of our knowledge is one of the most important aspects of any scientific study or clinical trial. As a practical tool, math is a means for taking pattern recognition and systematizing it. It is also a way for us to provide some form of communication and standard for comparing the results of different studies, and in the case of statistical significance, is meant to provide a measure of certainty against a null hypothesis.

Historically, clinical trials were developed around randomized treatment arms that were designed to answer the question, "Which treatment (A, B, or C) is better?" By selecting a straight forward metric, such as survival outcomes, statisticians could compare the efficacy of different treatments (Marubini and Valsecchi, 2004); however, this precluded our ability to ask what would happen if we combined treatment A and B, or B and C, or all 3 treatments, except by running yet another clinical trial. At the center of these often long and laborious trials was the notion of statistical power (Lachin, 1981). Just how many cases and controls does one need to ensure we can achieve significance? It is a simple question, but an important one that has been the subject of many sophisticated refinements. Here, there is a fundamental clash between Big Data science and classic clinical trial statistics. Paradoxically, the more data we collect on each subject, the more we decrease our likelihood of identifying statistically significant parameters as a result of multiple hypothesis correction. This is a fundamental flaw in the way that current statistical power calculations deal with large datasets.

Approaches to obtaining information from Big Data are different. Big Data is characterized by high volume, variety, and velocity of data generation (Costa, 2013). The strength of multi-omics is not merely the observation of many data points, but the discovery of biological mechanism through observation. Multi-omic Big Data grants us the power to examine disease in a human biological context, rather than extensively relying on murine models, which are limited in relevance to the human gut microbiome (Nguyen et al., 2015). In order to succeed, the Big Data movement in individualized medicine will require a holistic merger between large-scale data and biological mechanism.

3. Metabolic models for Big Data synthesis

To identify specific biological markers of disease, many studies utilize statistical correlations, which fall short of identifying underlying mechanisms (de Vos and de Vos, 2012). In the past, using Big Data to elucidate a biological mechanism involved generating a limited set of hypotheses that were then tested in the lab. While this approach has great value, it becomes less tenable as the number of measurements grows. The massive data sets generated from high-throughput omics technologies guarantee us more correlations arising purely from random chance. In the gut microbiome, this is especially problematic. The number of potential correlations increases with the hundreds of species and thousands of genes. Furthermore, the number of identified factors contributing to microbial composition including diet (David et al., 2014; De Filippo et al., 2010; Turnbaugh et al., 2009), sex (Chen et al., 2016), and even preservation method of the sample (Sinha et al., 2015), continue to grow and make it more difficult to differentiate the confounding from the causal.

A metabolic network provides a global picture of how metabolites and biochemical reactions are interconnected within a particular organism (Thiele and Palsson, 2010). Flux balance analysis on genome-scale metabolic models (GEMs) can be used to simulate microbial growth or to predict the production rate of a particular metabolite (Palsson, 2015). The power of this approach is not only that it recapitulates the mechanistic chemical flow through an entire organism, but also that it has the potential to integrate multiple data types. As the example shown in Fig. 1 indicates, reactions can be linked to genes, which are informed by DNA or RNA sequencing. RNA expression informs the amount of flux a reaction can carry, and metabolomics is a direct measurement of the metabolites. This makes metabolic models an ideal platform for organismal and community-scale data synthesis.

Increasing evidence suggests that integrating disparate, but complementary, data types can increase the power of one's analysis. Examples of this include the use of whole genome sequencing as a scaffold for RNA data (J. Wang et al., 2013; K. Wang et al., 2013) and the use of phosphorylation data to understand changes in metabolite concentration (Yugi et al., 2014). Within the microbiome field,16S rRNA data is combined with metagenomics to identify representative genomes and genome characteristics (PICRUSt: Langille et al., 2013; HUMAnN: Abubucker et al., 2012). Recent microbiome studies have also combined metagenome and metatranscriptome data to enable comparison between functional potential (metagenomic abundance/gene copy number) and functional activity (transcriptome level) (Franzosa et al.,