Systems Biology of the Human Microbiome

Beatriz Peñalver Bernabé[1], Lauren Cralle[1,2], and Jack A. Gilbert[1,2,3]

[1]The Microbiome Center, Department of Surgery, University of Chicago, Chicago, USA.
[2]Biosciences Division, Argonne National Laboratory, Lemont, IL, USA.
[3]Marine Biology Laboratory, Woods Hole, MA, USA.

Corresponding author:
Jack A. Gilbert
Surgery Brain Research Building
Room J557
5841 South Maryland Avenue, MC 5032
Chicago, IL 60637
773-834-8044
gilbertjack@uchicago.edu

Keywords: Microbiome, Systems Biology, Ecology, Network theory

Abstract

Recent research has shown that the microbiome—a collection of microorganisms, including bacteria, fungi, and viruses, living on and in a host—are of extraordinary importance in human health, even from conception and development in the uterus. Therefore, to further our ability to diagnose disease, to predict treatment outcomes, and to identify novel therapeutics, it is essential to include microbiome and microbial metabolic biomarkers in Systems Biology investigations. In clinical studies or, more precisely, Systems Medicine approaches, we can use the diversity and individual characteristics of the personal microbiome to enhance our resolution for patient stratification. In this review, we explore several Systems Medicine approaches, including Microbiome Wide Association Studies to understand the role of the human microbiome in health and disease, with a focus on 'preventive medicine' or P4 (i.e., personalized, predictive, preventive, participatory) medicine.

CHARACTERIZING THE HUMAN MICROBIOME

In recent years, Systems Biology has revolutionized our discovery of biomarkers to prevent, diagnose, and treat diseases. For example, the personalized diagnosis of HER2 breast cancer is one of the first examples implemented at the clinical level[1]. Systems Biology approaches allow us to make sense of the vast amount of data generated by "-omics" technologies, such as genomics, transcriptomics, metabolomics, and proteomics, through statistical, computational, and mathematical approaches that enable us to reveal the emergent properties of studied systems.

The Human Microbiome is heterogeneous between body sites (e.g. skin, gut, vagina), is distinctly personal[2], evolves over our life span[3], and has been implicated in, among other conditions, obesity[4] and depression[5]. Clinical studies to characterize the microbiome must consider numerous elements[6], including cohort selection, participant attrition, sample size, experimental design, sample collection, transportation and preservation, and more. Sample size is crucial to achieve statistical power, though few methods are currently available to establish *a priori* sample size for microbiome studies[7]. Many microbiome studies suffer from small sample sizes that may not capture the variability of the system, and we possess limited understanding of how to calculate sample size for longitudinal investigations. These limitations likely result from a lack of information about variability, which has led to a number of large scale efforts aimed at characterizing data from groups of participants in an attempt to quantify the variance in different traits[8,9]. For the microbiome, crowdsourcing efforts, such as American Gut (www.americangut.org), provide a unique opportunity to create data resources that can be used to predict statistical power for clinical studies.

To perform a Microbiome Wide Association Study (MWAS)[6], it is necessary to profile the microbiome to identify biomarkers that can be associated with host traits. The microbiome can be characterized using 16S/18S/ITS rRNA amplicon sequencing to identify the relative abundances of the different species, shotgun metagenomic sequencing to identify the organisms functional potential, metatranscriptomics (RNA-seq) to determine their functional response to change, metabolomics to identify microbial products, meta-proteomics[10] (UPLC-MS) to identify the enzymes being produced, and imaging (e.g. 3D cartography[11]) to visualize the spatial structure of the microbiome. The most common method is amplicon sequencing, usually using 16S rRNA[2,3] amplicons to describe bacterial and archaeal diversity, community structure, and composition of the microbiota. The benefit of amplicon sequencing is that it is inexpensive (<$20 a sample), is fast, and provides easy-to-interpret biomarker units. Traditionally, these biomarkers have been known as operational taxonomical units (OTUs) and were clusters of similar taxa (e.g. QIIME[12], Mothur[13]); however, new computational techniques have enabled this data to be probed at a greater taxonomic resolution[14-16], enabling the identification of biomarkers potentially at strain-level resolution (**Table.1**). Once that amplicon sequencing has been processed and annotated to known bacterial taxa, amplicon sequence data needs to be treated or normalized to avoid experimental and technical artifacts[12,17,18]. Subsequently, normalized amplicon data can be processed through computational pipelines (**Table.1**) to study the community structure (e.g., alpha and beta diversity) and to perform the statistical analysis that will link these biomarkers to host traits (e.g. phyloseq[19], QIIME[12], Mothur[13]).

Amplicon sequencing is limited, however, by the taxonomic resolution (i.e. you cannot usually identify microbes to the species or strain level), and it provides no information on the functional capacity of the microbes, although techniques exist to computationally predict microbial function for members of the ecological community that have a known sequence

(e.g. PICRUSt[20]). Therefore, to characterize microbial biomarkers such as genomic strain or functional gene, shotgun metagenomics is used, whereby the total genomic DNA of a sample is randomly sequenced[21,22]. While this provides less coverage of the total community composition, it does provide greater taxonomic resolution and potential functional information, which improves the ability to identify associations with host traits and patient stratification. However, shotgun metagenomic sequencing is expensive ($300-500 a sample), and analysis is more labor intensive than human genomics, mostly because there are no reference genomes for a majority of the organisms in a sample, which makes it harder to interpret the sequencing data.[23] However, there are a number computational pipelines, such as MetAMOS[24], Xander[25], and Anvi'o[26], that reduce the workload (**Fig.1**).

Importantly, metagenomic analysis only describes the genetics and functional potential of the microbiome, as it does not characterize the genes that are actively transcribed and translated into proteins. Metatranscriptomics[27] and metaproteomics[28] can be used to explore these phenomena, but they are more expensive than amplicon or metagenomic sequencing—metatranscriptomics can cost more than $500 per sample, while metaproteomics can cost more than $1000 per sample. Metatranscriptomics is easier to implement experimentally and computationally[29] than meta-proteomics; in the latter, the cells have to be isolated and the extracted proteins must be analyzed using LC-MS methods[30]. Metaproteomics provides useful biomarkers, as these are the active proteins and enzymes that are influencing host traits, but cost and difficulty of sample preparation limit the application of this approach.

The culmination of genetics, transcriptomics, and proteomics is of course the metabolome, which represents the small molecules generated by the individual cell or community of microbes. The influence of microbial metabolites of human health is well recognized[31]. In fact, metabolite biomarkers can often show the strongest association with host traits, likely because they have direct influence on host function[31,32]. Microbial metabolites, such as short chain fatty acids, have been shown to have a significant influence on local inflammation[32], hormonal balance[9], and even on mitochondrial activity[33]. The presence of microbe-related metabolites is commonly determined by gas or liquid chromatography followed by mass spectrophotometry[29], and the cost can vary from a few dollars for single metabolites to more than $100 per sample for an untargeted analysis of the metabolome. Due to the large correlation and interdependence between metabolites, clustering methods are employed to reduce the data dimensionality for downstream analysis[34].

SYSTEMS MEDICINE APPROACHES

When analyzing individual variables or traits in isolation, it is likely that the emergent properties of their interactions will not be observed (e.g. quorum sensing). An emergent property of a system is defined as characteristic of a complex system that cannot be predicted from its individual components directly without knowing the relationships or interactions between them. Therefore, identifying the emerging properties of the ecological community is fundamental to deciphering the key molecules in the system for diagnosis and treatments. Cancer biology has greatly benefited from Systems Biology approaches, specifically in identifying markers for particular cancers and using these to predict treatment strategies for individual patients[1,35]. Including microbiome characteristics in these predictions could greatly enhance the potential of emergent property discovery through systems thinking.

Amplicon studies, such as 16S rRNA sequencing, generally provide relative abundance data, which generates a 'compositional' effect. Compositional data is in a non-Euclidian space, and therefore common Systems Biology approaches are not applicable, as they require the

properties that a Euclidian space possess—e.g. distances between 2 points. However, a number of methods have been developed that accommodate relative abundance, such as SparCc[36], which utilizes ratios of normalized abundance between OTUs to determine co-abundance correlations[37]. Correlation networks can lead to numerous indirect edges between the microorganisms, and graphical models have been proposed to remove these spurious connections—e.g. SPIEC-EASI[38], sparse neighborhood and inverse covariance selection. Silverman et al.[39] suggested a different approach to transforming the compositional data into Euclidian space, so that the standard Systems Medicine methods can be directly applied (**Table 1**). This method is comparable to other normalization approaches and it does allow the identification of OTUs that differentiate samples, providing a forum for the characterization of microbial biomarkers. Similarly, determining the microbial biomarkers that correlate with a given host trait is often performed by generalized linear models[17,18], linear discrimination analysis[40], and linear log-contrast models with *l*-1 penalization to accommodate compositional data[41]. For example, with LEfSE[40], biomarkers such as metabolic pathway and taxonomic signature that correlate a host trait such as diet, can be calculated from metagenomic data[27] (**Table 1**). To improve the prediction of potential metabolic function from 16S rRNA data, several publically available[27] and commerical[8] databases that link genetic function to taxonomy are being developed, so that Systems Medicine can benefit from more accurate predictions at a lower data generation cost.

Correlation networks (e.g., Spearman correlations) are also frequently used to determine the interactions between microbial components[36] of a system, or between microbes and metabolites, clinical variables, and host traits, such as inflammatory markers (**Table 1**). For example, Schirmer et al.[42] developed an elegant approach to establish whether microbes and genes in stool samples could describe population variance in individual immune cytokine expression following pathogenic stimulation. Using Spearman correlations, the authors were able to identify which cytokine changes were associated with which microbes and genes, and they demonstrated high interpersonal variability, suggesting that patient populations should be stratified by immune response and microbiome. Price et al.[8] demonstrated a 'personalized medicine'[6] approach in a cohort of 108 participants followed longitudinally over nine months, characterizing their genomes and clinical variables associated with disease in order to identify metabolic and microbial biomarkers of patient variables from blood, urine, and stool. Using Spearman correlations, they were able to identify clusters of patient traits that correlated with known clinical phenotypes and that were associated with specific microbial and metabolic biomarkers. Price et al.'s methodology can be extended to larger, more diverse samples for longer period of times and to include other more complex traits, such as those related with neurological disorders[5].

Machine learning approaches based on tree methods, e.g. Random Forest[3], are able to discriminate important biomarkers associated with a given host trait, even when non-linear relationships are present. For instance, to predict the glycemic response in a population of 800 participants, researchers[43] modelled postprandial blood glucose (PPGR) values as a function of different phenotypic and personal traits (e.g. nutritional intake, BMI, immune levels, glucose, and microbial composition and function) using a gradient bootstrapping tree (**Table 1**). PPGR predictions significantly aligned the experimental PPGR results from a new cohort of participants with a different dietary profile. The application of this method for high-dimensionality data has been used to form a company, DayTwo (www.daytwo.com), which predicts personalized diets for customers based on their blood chemistry and microbiome. Notably, these approaches can be used to predict biomarkers and create models for many disorders, such as cardiovascular diseases, hypertension, preeclampsia, anemia, and stress levels, just to name a few.

While we can create models of association between microbial taxa and host traits, microbial metabolic products generally appear to have greater predictive potential. While metabolomics is becoming more common in MWAS studies[6], it is possible to predict microbial metabolism from metagenomic data[44] and as a function of manually curated, genome-enabled metabolic models (GEMs). A database of semi-manually GEMs for 773 gut-associated microbial taxa just became available[45]. Using these approaches, it is possible to predict emergent metabolic properties from the interaction of the microbiota and to study metabolic connections with the host[9,46-48] through flux balance analysis—a linear optimization method[48] (**Table 1**). Although not yet applied to Systems Medicine, due to the limited amount of available data, it is also possible to apply GEMs to describe nodes in an artificial neural network[49], where the topology, such as edge connectivity of each node, can capture emergent properties and enable more accurate predictions of microbial biomarker activity[8].

PREDICTING CAUSATION WITH DYNAMIC DATA

Longitudinal studies, which provide time-resolved 16S rRNA amplicon sequencing data, are becoming common; yet, many studies do not use analyses that take advantage of the potential of these datasets to predict causality[8,22,48,50,51], most likely due to the lack of adequate methods to analyze longitudinal microbial data. Typically, correlation networks (e.g., SparCc[36], CCLasso[37]) or graphical models (e.g., SPEIC-EASI[38]) are instead employed, though they do not provide edge directionality. Several longitudinal studies have been investigated to provide predictions of the associations between microbial dynamics and host trait-dynamics (**Table 1**). For example, David et al.[27] followed participants on different diets over several days. Pearson correlation coefficients were grouped using dynamic hierarchical clustering, and those clusters were associated with the differing diets, corresponding to well-reported functions of microbiota in animal diets.

Machine learning methods can be extended to study dynamic microbial abundance data (**Table 1**). Sparse Variance Autocorrelation models (sVARs)—commonly used in econometrics—and Dynamic Bayesian networks (DBNs) have been employed to model longitudinal proteomics data[35]. Recently, sVARs have been used to model longitudinal stool microbial abundances, effectively characterizing the underlying network interactions that contribute to observed ecosystem dynamics[52]. However, sVAR models do not guarantee causation, since they predict Granger causality. DBNs allow us to include non-linear relationships, which is a distinct advantage over linear approaches, to account for noisy data, and to incorporate *prior knowledge* of the system. For example, McGeachie et al.[53] employed DBN to model the colonization of the gut microbiome of 58 low birth-weight infants in a neonatal intensive care facility. Employing lagged time correlations, whereby events at time point 1 are used to predict events at time point 2, microbial biomarkers were identified that were significantly associated with gestational age at delivery and the use of antibiotics. Therefore, DBNs and Bayesian statistics can generally be used to identify emergent properties that enable biomarker identification, but it is also essential to validate the proposed interactions, if possible.

Several other approaches exist to identify molecular mechanisms that underpin observed trends and can therefore be used to predict biomarkers. For example, it is possible to use Lotka–Volterra models (**Table 1**) to predict the individual growth rates of microorganisms in a community as a function of specific perturbations—e.g. antibiotics. Generalized Lotka–Volterra (gLV) parameters have been estimated through Tikhonov regularization[54], Bayesian statistics[54], or sparse linear regression with bootstrap aggregation[55]. While gLV analysis can be used to capture dynamics in multi-omic datasets, they generally require substantial longitudinal data to

achieve accuracy. Finally, agent-based models (**Table 1**) are useful to model systems and have been employed to predict interactions between host and microbial metabolism[56]. However, agent-based models require *a priori* knowledge of a relationship and quantification of that relationship, which we lack for many interactions. As systems become more parameterized, the use of agent-based modeling approaches might become more prevalent, enabling the prediction of complex behaviors that emerge from multicellular systems, like the human microbiome.

FUTURE PERSPECTIVES

The application of Systems Biology modeling to medical microbiome studies is still in its infancy, in large part due to the availability of adequately powered studies. However, much of the analytical infrastructure and the systems thinking does exist, albeit with provisos. While it is possible to calculate and leverage correlations between microbial and host variables to demonstrate significant associations, we still have substantial knowledge gaps regarding what these associations actually represent. These knowledge gaps will be filled through a combination of Systems Biology on different scales, from *in vitro* cellular studies to *in vivo* community dynamic studies. An obvious gap that should be easy to fill is the absence of extensive datasets characterizing the fungal[27] and viral[57] components of the human microbiome. This is essential, as we attempt to predict bacterial-host trait associations, since viral or fungal variables may explain the host trait more effectively. In addition, where appropriate, there should be further investment into longitudinal datasets, especially prospective longitudinal investigations. These temporal association studies will enable the prediction of time-lagged relationships and feedback loops, which are particular useful to uncovering emergent properties that could be used to forecast clinical outcomes for specific diseases. Gut microbiome-produced metabolites can influence distal organs in the body, and have long-term effects that may not be manifest for many years. For example, bacteria can produce neurotransmitters, such as serotonin pre-cursors and gamma aminobutyric acid (GABA), which can influence neurophysiological development in infants[5] and result in cognitive disruption in childhood. Also, such studies need more effective strategies to integrate data types, such as immunological, endocrinological, and neurological variables, with multi-omic microbiome variables to feed the systems modeling approaches outlined above.

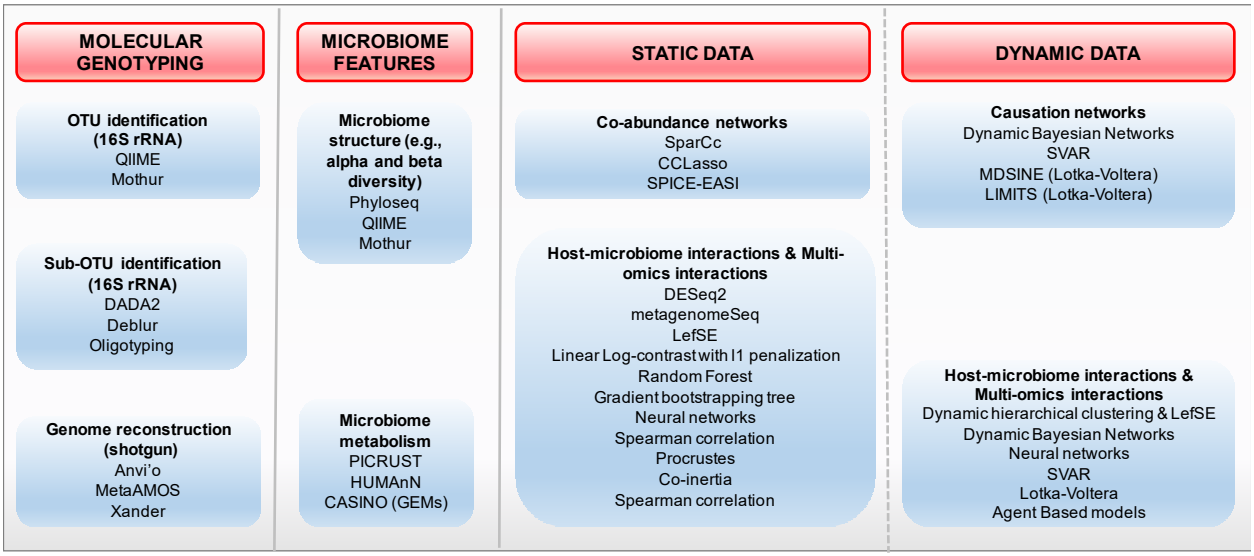
The goal of building large-scale networks that accommodate all known data types and create predictions of putative functional associations is not enough. There needs to be a strategy, not just intention, to aid experimental validation of proposed associations. This is often difficult due to the number of associations identified, resulting in a necessity to sub-select observed associations for validation[42]. It is also necessary to validate predictive models on relevant data, whereas current validation methods rely on simulated data to determine the accuracy of their methods. Simulated data, while being a great initial approach for method optimization, rarely resembles reality. Examples of validation methods are functional genomics. Functional genomics can identify host-microbiome interactions through transfection of cloned microbial DNA fragments into *E. coli*, for instance, to identify microbially-derived molecules—peptides, metabolites, etc.—that influence the host[58] or the development of *in vitro* analogues for simulating the gut environment[59]. Again, there is a substantial need to integrate data across scales, and as such, single cell analyses are becoming invaluable. For example, mathematical combinatorial approaches based on ordinary linear regression to generate random communities from single cell isolation experiments have proven successful in predicting likely strains responsible for observed phenotypes[60]. Progress in single cell isolation and subsequent analysis[61] will open doors to new biological insights, as well as novel methods to validate experimental results.

The ultimate goal of systems microbiome medicine is to develop diagnostic tools based on the microbiome, and treatments—probiotics and prebiotics—to restore loss-of function or to elicit specific host responses. Therefore, it is important to develop mathematical theory, or adopt statistical approaches from other fields, that can facilitate the identification of emergent properties and specific associations to define biomarkers[62]. Finally, Systems Microbial Medicine is the ultimate multidisciplinary field, and the ability to translate basic research into clinical applications will require integration of expertise across microbiologists, geneticists, mathematicians, statistics, engineers, computational biologist, nutritionists, immunologist, neurologist, endocrinologist, etc. In essence, we need a system of scientists to study the systems of life.

Funding

BPB is funded by the Arnold and Mabel Beckman Foundation (Arnold O. Beckman Postdoctoral Fellow)

Table 1. Summary of the most common experimental and computational methods employed in Systems Microbiome Medicine



REFERENCES

1 van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *New Engl J Med* **347**, 1999-2009, doi:DOI 10.1056/NEJMoa021967 (2002).

2 Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449**, 804-810, doi:10.1038/nature06244 (2007).

3 Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222-227, doi:10.1038/nature11053 (2012).

4 Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology - Human gut microbes associated with obesity. *Nature* **444**, 1022-1023, doi:10.1038/nature4441022a (2006).

5 Sharon, G., Sampson, T. R., Geschwind, D. H. & Mazmanian, S. K. The Central Nervous System and the Gut Microbiome. *Cell* **167**, 915-932, doi:10.1016/j.cell.2016.10.027 (2016).
***This review paper presents a comprehensive summary of the human gut microbiome effects in appropriate development of the neurological system through what has been coined as the "brain-gut-microbiome" axis. The authors described compelling human observations and germ-free murine experiments that showed the relationships between the dynamic development of the gut microbiome and adequate development of the neurological system—blood-brain barrier, myelination, neurogenesis, and microglia maturation—starting with conception in the uterus. Further, the authors relate this initial gut microbial dysbiosis with other mental and neurological disorders, such as autism and depression.**

6 Gilbert, J. A. *et al.* Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**, 94-103, doi:10.1038/nature18850 (2016).
***The authors describe the concept of microbiome-wide association studies (MWA), which are similar to genome-wide association studies with the caveat that in MWA, the genome associated with phenotype is generally partially known and is changing over time as well as between individuals. Several MWA studies are described, including multi-omics integrations, and the possible clinic applications in precision medicine are delineated. Specifically, the authors proposed to use the individual changes in meta-genomics, what they call "microbial Global Positioning System" (GPS), to stratify individuals and to guide their treatment.**

7 La Rosa, P. S. *et al.* Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. *Plos One* **7**, doi:ARTN e52078 10.1371/journal.pone.0052078 (2012).

8 Price, N. D. *et al.* A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat Biotechnol* **35**, 747-756, doi:10.1038/nbt.3870 (2017).
****This is a pilot study to demonstrate the initial feasibility of the 100,000 (100K) person wellness project. Price et al. followed 108 subjects for 9-months and they measured microbial abundance, metabolomes and proteomes of the individuals, whole genome sequences, and the results of their clinical tests and daily activity tracking (personal data clouds). Using all of this data and Systems Biology approaches (i.e. correlation networks), the authors were able to establish relationships between their obtained data and the participant risk for a given disease—cardio metabolic disease, inflammatory bowel disease, etc.**

9 Nielsen, J. Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine. *Cell Metab* **25**, 572-579, doi:10.1016/j.cmet.2017.02.002 (2017).
****Nielsen describes how genome-wide metabolic models (GEM) and overlapping meta-transcriptomics and meta-metabolomics data in GEMs are potential tools to relate the human host and its microbiome, as metabolite concentration is commonly employed in clinical setups as a marker for diagnosis (e.g., insulin and glucose). Several examples are depicted in which studies have shown how to couple this important interaction and how personalized medicine could be implemented.**

10 Hettich, R. L., Pan, C. L., Chourey, K. & Giannone, R. J. Metaproteomics: Harnessing the Power of High Performance Mass Spectrometry to Identify the Suite of Proteins That Control Metabolic Activities in Microbial Communities. *Anal Chem* **85**, 4203-4214, doi:10.1021/ac303053e (2013).

11 Bouslimani, A. *et al.* Molecular cartography of the human skin surface in 3D. *Proc Natl Acad Sci U S A* **112**, E2120-2129, doi:10.1073/pnas.1424409112 (2015).
***Using a combination of mass spectrophotometry and 16S rRNA data, the authors mapped the distribution of microbes and metabolites in more than 400 locations of the human skin. The 3D reconstruction allowed them to visualize the co-abundance preference between microbial species, microbes, and metabolites and their preferred location.**

12 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335-336, doi:10.1038/nmeth.f.303 (2010).

13 Schloss, P. D. *et al.* Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl Environ Microb* **75**, 7537-7541, doi:10.1128/Aem.01541-09 (2009).

14 Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581-+, doi:10.1038/Nmeth.3869 (2016).
****Using statistical models to account for sequencing errors, the authors developed an approach to merge sequences robustly, which avoids the mapping into databases (Green Genes) to cluster them in operational taxonomical units (OTUs).**

15 Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**, doi:10.1128/mSystems.00191-16 (2017).
***Deblur aims to pick sub-OTUs from 16s rRNA sequencing data, using a computational approach based on an error rate of sequencing determined by the number of indels (insertions and deletions in sequences) that were obtained from alignments of the reads. Through an iterative process, from more abundant to less abundant sequences, DeBlur determines the most likely sequences and the total accounts associated with them.**

16 Eren, A. M. *et al.* Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *Isme J* **9**, 968-979, doi:10.1038/ismej.2014.195 (2015).
***The authors used oligotyping to align sequences and determine the nucleotide variability at each position of the 16S rRNA sequences, similarly to single-nucleotide polymorphism (SNPs) in GWAS. The variation of entropy at each location can be used to differentiate between cohorts.**

17 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
18 Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods* **10**, 1200-+, doi:10.1038/Nmeth.2658 (2013).
19 McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *Plos One* **8**, doi:ARTN e6121710.1371/journal.pone.0061217 (2013).
20 Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**, 814-+, doi:10.1038/nbt.2676 (2013).
21 Kuczynski, J. *et al.* Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet* **13**, 47-58, doi:10.1038/nrg3129 (2012).
22 Vatanen, T. *et al.* Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **165**, 842-853, doi:10.1016/j.cell.2016.04.007 (2016).
****This outlines a study conducted to understand the difference in autoimmune disease prevalence in three closely related communities–Russia, Finland and Estonia—using 16S rRNA and shot-gun sequencing of infant gut microbiome as well as cytokine abundances. The authors demonstrate that the microbially produced LPS is specific, generating LPS subtypes that differently trigger the immune system to act.**

23 Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 8, doi:10.1186/s40168-016-0154-5 (2016).
***This paper is a detailed summary of the most commonly used methods for contig assembling of short reads from shot-gun sequencing, binning (group contigs into taxonomical bins), curation, and validation of reconstructed community genomes. The authors provide several pros and cons for each of the described methods.**

24 Treangen, T. J. *et al.* MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol* **14**, doi:ARTN R210.1186/gb-2013-14-1-r2 (2013).
25 Wang, Q. *et al.* Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome* **3**, doi:ARTN 3210.1186/s40168-015-0093-6 (2015).
***Xander is a computational pipeline that aims to determine the abundance of genes that encode for specific proteins out of shotgun sequences from microbial communities. Thus, Xander does not aim to provide reconstructed community genomes. Xander is available at https://github.com/rdpstaff/Xander_assembler.**

26 Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319, doi:10.7717/peerj.1319 (2015).
***Anvi'o is a platform for processing multiple data types, such as metagenomics and metatranscriptomics, and overlapping "-omics" sets. Anvi'o also incorporates multiple visualization tools for large "-omics" datasets. Anvi'o is available at https://merenlab.org**

27 David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559-+, doi:10.1038/nature12820 (2014).

28 Verberkmoes, N. C. *et al.* Shotgun metaproteomics of the human distal gut microbiota. *Isme J* **3**, 179-189, doi:10.1038/ismej.2008.108 (2009).

29 Aguiar-Pulido, V. *et al.* Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis. *Evol Bioinform Online* **12**, 5-16, doi:10.4137/EBO.S36436 (2016).
***This review underscores the current experimental methods to measure microbial abundance, transcriptional activity, and microbially-produced metabolites. The review presents common statistical methods to filter and pre-process the raw data generated by "-omics" techniques.**

30 Zhang, X. *et al.* MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4**, 31, doi:10.1186/s40168-016-0176-z (2016).
***This approach limits the peptide space that is possible in a sample by *a priori* information, establishing the possible proteins that could be present based on the genomes of the murine and human microbiome.**

31 Donia, M. S. & Fischbach, M. A. Small molecules from the human microbiota. *Science* **349**, doi:ARTN 125476610.1126/science.1254766 (2015).
***This review explores types of microbially-produced metabolites and includes a summary of experimental methodologies to determine microbiome-host interactions based on target and non-target metabolomics analysis in murine models.**

32 Smith, P. M. *et al.* The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. *Science* **341**, 569-573, doi:10.1126/science.1241165 (2013).

33 Franco-Obregon, A. & Gilbert, J. A. The Microbiome-Mitochondrion Connection: Common Ancestries, Common Mechanisms, Common Goals. *mSystems* **2**, doi:10.1128/mSystems.00018-17 (2017).
***This perspective article is focus on the relationship between gut microbiome and mitochondria. Based on multiple research, the authors argued that through several microbially produced metabolites (i.e., small fatty acids (SCFA), urolithins and lactate), the gut microbiome help mitochondria well-functioning. As the adequate performance of the mitochondria is key in human health, this new axis, microbiome-mitochondria, open a new avenue for diagnostics and treatment, according to the authors.**

34 McHardy, I. H. *et al.* Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* **1**, 17, doi:10.1186/2049-2618-1-17 (2013).

35 Hill, S. M. *et al.* Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods* **13**, 310-318, doi:10.1038/Nmeth.3773 (2016).
****This paper outlines the results from a Systems Biology competition—DREAM (http://dreamchallenges.org/) —to develop methods to predict breast cancer signaling networks using dynamic target proteomics. The competition organizers compared all the proposed methods among each other, which provided a unique setting for an adequate discrimination between the different computational, statistical and combination of thereof systems biology approaches.**

36 Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *Plos Comput Biol* **8**, doi:ARTN e100268710.1371/journal.pcbi.1002687 (2012).

37 Fang, H. Y., Huang, C. C., Zhao, H. Y. & Deng, M. H. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31**, 3172-3180, doi:10.1093/bioinformatics/btv349 (2015).
***The authors present a method, CCLasso, to determine correlations between 16S rRNA compositional data by using least squares with *l*1 penalty after log ratio transformation. The method has several theoretical advantages, such as the fact that their optimization equation is convex, their minimum is global, and the compositional data correlation matrix is positive definite. CCLasso is available at https://github.com/huayingfang/CCLasso.**

38 Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *Plos Comput Biol* **11**, doi:UNSP e100422610.1371/journal.pcbi.1004226 (2015).
***SPIEC-EASI uses a sparse graphical model to reconstruct the microbial network using sparse neighborhood and inverse covariance selection. Graphical models are based on conditional independence—two nodes are conditionally independent if there is no other alternate network that can better predict their abundance. Conditional independence aims to avoid reporting indirectly connected microorganisms. SPIEC-EASI is available at https://github.com/zdk123/SpiecEasi.**

39 Silverman, J. D., Washburne, A. D., Mukherjee, S. & David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* **6**, doi:ARTN e2188710.7554/eLife.21887 (2017).
****The authors present a new proposed transformation method for compositional data such that common Systems Biology approaches can be employed without the need to modify them to account for the non-Euclidian space. The method, PhILR, is based on Isometric Log-Ratio transformation using phylogenetic trees for partitioning to avoid the singular covariance matrix that centered log-ratio transform renders, making the latter transformation inadequate for common statistical models.**

40 Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol* **12**, R60, doi:10.1186/gb-2011-12-6-r60 (2011).

41 Lin, W., Shi, P. X., Feng, R. & Li, H. Z. Variable selection in regression with compositional covariates. *Biometrika* **101**, 785-797, doi:10.1093/biomet/asu031 (2014).

42 Schirmer, M. *et al.* Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell* **167**, 1125-1136 e1128, doi:10.1016/j.cell.2016.10.020 (2016).
****The authors implement an experimental methodology to understand the relationships between host microbiome and host systems. In this paper, the authors specifically focus on the effects of the gut microbiome and its metabolic functions on the response of the host immune system under well-defined microbial challenges. The predictions from Spearman correlations were experimentally validated.**

43 Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079-1094, doi:10.1016/j.cell.2015.11.001 (2015).
****Zeevi et al. complete a large, comprehensive study to explore the relationship between postprandial (post-meal) glycemic responses (PPGR) of more than 800 participants as a function of multiple measurements (e.g. diet, microbiome abundance, etc.). Using the data to generate a predictive tree model, the authors select a personalized diet for newly recruited participants. The results are similar**

to those obtained in diets made by certified dietitians, demonstrating the potential of statistical and computational methods, i.e. Systems Medicine, for personalized medicine.

44 Larsen, P. E. *et al.* Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. *Microb Inform Exp* **1**, 4, doi:10.1186/2042-5783-1-4 (2011).

45 Magnusdottir, S. *et al.* Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol* **35**, 81-89, doi:10.1038/nbt.3703 (2017).
****This is the largest effort to date to generate semi-automatic metabolic models for microorganism present in the human gut microbiome. The human gut metabolic models are available at https://vmh.uni.lu/#microbes/search**

46 Bauer, E., Laczny, C. C., Magnusdottir, S., Wilmes, P. & Thiele, I. Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires. *Microbiome* **3**, doi:ARTN 5510.1186/s40168-015-0121-6 (2015).
***Using metabolic reconstructions of 301 human gut microorganisms, the authors explores the properties of the human microbiome with several statistical and computational techniques—linear regression, correlations, and t-distributed stochastic neighbor embedding (t-SNE). Among other properties, Bauer et al. observe that phylogenic similarity does not guarantee metabolic similarity.**

47 Heinken, A. & Thiele, I. Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. *Gut Microbes* **6**, 120-130, doi:10.1080/19490976.2015.1023494 (2015).
***The authors generate a model that includes 11 published manually-curated gut microbe reconstructions and one of the latest human metabolic reconstruction, Recon2. Using this metabolic modeling framework, the authors explore the metabolic behavior of the system under pathogenic infection as well as secretion and absorption patterns between host and microbiome.**

48 Shoaie, S. *et al.* Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome. *Cell Metab* **22**, 320-331, doi:10.1016/j.cmet.2015.07.001 (2015).
****Shoaie et al. develop a toolbox called CASINO (Community And Systems-level Interactive Optimization) for modelling metabolic communities. Using metabolic reconstructions of 5 gut microorganisms, the authors predict the changes in fecal and serum metabolism upon diet changes in more than 40 participants, with low and high gut microbial abundance. The metabolic model was able to predict the gut microbial metabolism and some of the serum metabolites—acetate and some amino acids.**

49 Larsen, P. E., Field, D. & Gilbert, J. A. Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods* **9**, 621-+, doi:10.1038/Nmeth.1975 (2012).

50 DiGiulio, D. B. *et al.* Temporal and spatial variation of the human microbiota during pregnancy. *P Natl Acad Sci USA* **112**, 11060-11065, doi:10.1073/pnas.1502875112 (2015).
***DiGuilio et al. follow women from early pregnancy until delivery on a weekly schedule and measured the gut, saliva, and vaginal microbiome, aiming to predict preterm birth. The authors describe several microbial clusters in the vaginal microbiome that were associated with this negative outcome.**

51 Schirmer, M. *et al.* Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell* **167**, 1897, doi:10.1016/j.cell.2016.11.046 (2016).
****The authors implement an experimental methodology to understand the relationships between host microbiome and host systems. In this paper, the authors specifically focus on the effects of the gut microbiome and its metabolic functions on the response of the host immune system under well-defined microbial challenges. The predictions from Spearman correlations were experimentally validated.**

52 Gibbons, S. M., Kearney, S. M., Smillie, C. S. & Alm, E. J. Two dynamic regimes in the human gut microbiome. *Plos Comput Biol* **13**, e1005364, doi:10.1371/journal.pcbi.1005364 (2017).
****This is one of the first times that sparse vector autoregression (sVAR) is employed in Systems Medicine. Specifically, the authors model the dynamic of the gut microbiome and are able to identify that the diet accounted for non-autoregressive dynamics, while anaerobes allowed the microbiome recovery.**

53 McGeachie, M. J. *et al.* Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks. *Sci Rep* **6**, 20359, doi:10.1038/srep20359 (2016).
***This paper describes one of the first times in which dynamic Bayesian network approaches is used to understand the dynamic evolution of a human microbial community. The authors use one lag between data points.**

54 Bucci, V. *et al.* MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol* **17**, doi:ARTN 12110.1186/s13059-016-0980-6 (2016).
***This is a method based on generalized Lotka-Voltera equations. The main difference with previous approaches is in the manner that model parameters (e.g. interaction terms between species, growth rates) are determined (e.g. Bayesian estimation approach).**

55 Fisher, C. K. & Mehta, P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *Plos One* **9**, doi:ARTN e10245110.1371/journal.pone.0102451 (2014).

56 Shashkova, T. *et al.* Agent Based Modeling of Human Gut Microbiome Interactions and Perturbations. *Plos One* **11**, e0148386, doi:10.1371/journal.pone.0148386 (2016).
***This is an application of Agent Based models in determining the interactions between two bacteria and the gut upon antibiotic treatment.**

57 Lim, E. S. *et al.* Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat Med* **21**, 1228-+, doi:10.1038/nm.3950 (2015).
***Lim et al. is one of the first to study the dynamic change of the viral microbiome—bacteriophages and eukaryotic RNA and DNA viruses—and the bacterial microbiome in infant twins. The authors observed that the viral microbiome was more similar between twins and that it grew in diversity, opposite to the bacteriophage virome. Dynamic interactions between bacteriophage and bacteria were modeled using a generalized Lotka-Volterra model.**

58 Cohen, L. J. *et al.* Functional metagenomic discovery of bacterial effectors in the human microbiome and isolation of commendamide, a GPCR G2A/132 agonist. *P Natl Acad Sci USA* **112**, E4825-E4834, doi:10.1073/pnas.1508737112 (2015).
****This explores a synthetic biology application to determine the interaction of the human microbiome with their human host. The authors clone large fragments of commensal DNA and transfected then to *E. coli*. By contacting the produced media with human cells lines, they observe which cloned DNA segments produced the desired phenotype, e.g. activation of NFKB, and the microbially-produced molecules that elicit the change.**

59 Kim, H. J., Li, H., Collins, J. J. & Ingber, D. E. Contributions of microbiome and mechanical deformation to intestinal bacterial overgrowth and inflammation in a human gut-on-a-chip. *P Natl Acad Sci USA* **113**, E7-E15, doi:10.1073/pnas.1522193112 (2016).
****This paper is an example of how microfluidics systems can be constructed to simulate different organs in the body, i.e., the human gut. This system allows us to determine *in vitro* effects of microbes (i.e., bacteria, fungi, virus) in a well-defined and control environment.**

60 Faith, J. J., Ahern, P. P., Ridaura, V. K., Cheng, J. Y. & Gordon, J. I. Identifying Gut Microbe-Host Phenotype Relationships Using Combinatorial Communities in Gnotobiotic Mice. *Sci Transl Med* **6**, doi:ARTN 220ra1110.1126/scitranslmed.3008051 (2014).

61 Tolonen, A. C. & Xavier, R. J. Dissecting the human microbiome with single-cell genomics. *Genome Med* **9**, 56, doi:10.1186/s13073-017-0448-7 (2017).
***This perspective presents the new advances in single-cell genomics as well as the possibility of single-cell genomics, e.g. microfluidics, to understanding microbial communities.**

62 Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350**, 663-666, doi:10.1126/science.aad2602 (2015).
****Coyte et al. present a theory to understand community dynamics and robustness under external perturbations.**