Insights into study design and statistical analyses in translational microbiome studies

Jyoti Shankar

J. Craig Venter Institute, Rockville, Maryland, USA

Correspondence to: Jyoti Shankar, MBBS, PhD. J. Craig Venter Institute, Rockville, Maryland, USA. Email: jyoti.shankar@gmail.com.

Abstract: Research questions in translational microbiome studies are substantially more complex than their counterparts in basic science. Robust study designs with appropriate statistical analysis frameworks are pivotal to the success of these translational studies. This review considers how study designs can account for heterogeneous phenotypes by adopting representative sampling schemes for recruiting the study population and making careful choices about the control population. Advantages and limitations of 16S profiling and whole-genome sequencing, the two primary techniques for measuring the microbiome, are discussed followed by an overview of bioinformatic processing of high-throughput sequencing data from these measurements. Practical insights into the downstream statistical analyses including data processing and integration, variable transformations, and data exploration are provided. The merits of regularization and ensemble modeling for analyzing microbiome data are discussed along with a recommendation for selecting modeling approaches based on data-driven simulations and objective evaluation. The review builds on several recent discussions of study design issues in microbiome research but with a stronger emphasis on the downstream and often-ignored aspects of statistical analyses that are crucial for bridging the gap between basic science and translation.

Keywords: Microbiome; study design; statistical analysis; statistical models; translational research

Submitted Oct 26, 2016. Accepted for publication Nov 28, 2016.

doi: 10.21037/atm.2017.01.13

View this article at: http://dx.doi.org/10.21037/atm.2017.01.13

Introduction: the microbiome in translational medicine

The readily modifiable and dynamic nature of the microbiome and its close interactions with every major systemic component of the human body makes it a promising target for translation into medical practice. In the years following the completion of the National Institutes of Health's human microbiome project (1), there has been a large volume of research identifying and cataloging the microbiome in physiological and pathological states including metabolic disorders (2) [obesity (3,4), diabetes (5), metabolic syndrome (6)], autoimmune disorders (7,8) [inflammatory bowel diseases (9,10), multiple sclerosis (11), rheumatic disease (12)], infectious diseases (13) [upper and lower respiratory tract infections (14), candidiasis (15,16), urinary tract infections (17), secondary infections following cystic fibrosis (18) and HIV infections (19)], cancer (20,21) and cancer-related infections (22), lung and stem-cell (23,24) transplant-related infections and long-term effects of antibiotic-associated dysbiosis (25) [*Clostridium difficile* infections (26)].

Understanding the microbiome is becoming increasingly relevant across the translational spectrum because of its potential roles in multiple medical applications and as a therapeutic intervention by itself. The microbiome can serve as a diagnostic adjunct to traditional clinical and laboratory measures (23,27-29), a determinant of treatment-response (30), a window into the side-effects of exposure to antibiotics (25), a baseline measurement prior to the initiation of therapy (31), and a signature of immune processes such as inflammation (23,32). Furthermore, it can be used to guide nutrition and preventive interventions (33), monitor the severity, progression or recovery from disease (34), track

© Annals of Translational Medicine. All rights reserved.

treatment-outcomes (35), and identify small-molecule drugs (36) and mechanisms of pathogenesis which can then be targeted for therapeutic interventions (37).

The diverse ways in which the microbiome can be integrated into translational studies leads to the equally diverse technical challenges in designing these studies and the associated statistical analyses. A large body of work in the field has addressed day-to-day operations (38) involved in performing microbiome studies in research and epidemiological settings (39,40) including sample collection and storage (41-43), laboratory protocols for sample processing (40,44), choice of microbiome sequencing protocols (45) and the computational infrastructure required for data processing and storage (46,47). This review builds on these discussions while specifically focusing on the design of translational microbiome studies and development of downstream statistical analysis plans which answer complex questions in a translational setting and inform early and late-phase clinical trials.

Making informed decisions: study design

Because studies cannot include entire populations, it is crucial to define the target population of interest and then draw a representative study sample to ensure that the findings from the study are generalizable (48). Given the complex disease targets that translational studies attempt to understand, it is inevitable that diseased populations of interest are heterogeneous in their clinical phenotypes. While heterogeneous phenotypes could enable investigators to understand several facets of a disease spectrum, the presence of heterogeneity dilutes the statistical estimates of effect sizes of the microbiome. This problem is compounded by the fact that typical effect sizes of individual members of the microbiome are weak. The dilution of effect size estimates is even more acute for diseases with several complex phenotypes. If there is prior evidence for substantial clinical heterogeneity or if there are theoretically defined subpopulations with different disease characteristics within the population of interest, it is prudent to prioritize specific aspects of the disease to study and recruit a relatively homogeneous study population. For example, in a study examining the role of the microbiome in the development of pneumonia, it could be beneficial to focus on the most common type of pneumonia in the patient population rather than trying to profile all of the different etiological types. This can be achieved by specifying well-defined inclusion and exclusion criteria based on respiratory cultures such that only patients meeting the culture criteria for a specific etiology are included in the study (23). Such a targeted recruitment of patients increases the power of the study and is a judicious use of limited resources.

The choice of controls is a challenging question and is determined by the purpose of the study. In a diagnostic study, the typical goal is to find a discriminating microbiome signature that would aid in the accurate differential diagnosis of two very close conditions. A good control population in this case includes patients with a clinical phenotype that is a clear contrast to either of the two conditions of interest. For example, when studying pneumonia and tracheobronchitis which are two closely related and almost indistinguishable culture-positive conditions that occur in patients admitted to the intensive care unit or in lung transplant recipients, the control population could comprise patients with asymptomatic colonization confirmed by culture (23). When the microbiome is utilized for monitoring the severity or progression of a disease or the efficacy of a treatment process, patient subsets with either less severe disease or lower treatment-response could be used as controls. In studies that investigate microbiome-host interactions to identify pathways and small-molecules of interest to pursue for drug discovery, a valuable control group could be one which is completely free of disease. Multiple control groups recruited on the basis of a variety of criteria and methods could offer more insight into the heterogeneous effects of the microbiome compared to studies with only a single control group (49).

An important aspect of study design is determining the timing and frequency of sample collection from the study population. If the goal is to discover and validate diagnostic microbiome signatures, it is most meaningful to collect cross-sectional samples from patients with clinically confirmed, early stages of the disease. On the other hand, if the goal is to monitor disease severity or treatment-response, an appropriate design would incorporate temporally separated samples and repeated measurements from the same study subject. The frequency of sample collection in temporal study designs is often determined by factors such as the monetary resources budgeted for sample collection and storage, invasiveness of the sampling procedure, subject compliance to study protocol and in the case of retrospective studies, availability of samples from a pre-existing biorepository (40). Samples retrieved from a biorepository might not be uniformly separated in time, an aspect that needs to be accounted for while interpreting study findings.

© Annals of Translational Medicine. All rights reserved.

While understanding the microbiome might be the central focus of the study, concurrent clinical, laboratory and "omics" measurements from the study play a crucial role in expanding the scope of microbiome-related findings by placing them in the context of disease pathways and pathological mechanisms. Non-microbiome measurements that encode disease phenotypes as continuous rather than categorical variables allow for more robust analyses and inference. For example, in a study investigating the role of the microbiome in obstructive or restrictive lung disease, periodic assessments of FEV1, lung density, airway dimensions and pulse oximetry, which are continuous surrogates of disease (50), are substantially more informative than a categorical assessment of disease severity based on clinical records. Studies can similarly benefit from the inclusion of high-resolution measurements of disease phenotypes obtained using smartphone-connected digital monitoring devices (51) that are increasingly becoming popular in healthcare.

Measuring the microbiome: taxonomy and function

A vast majority of studies have focused on the bacterial and, to a lesser extent, on the fungal microbiome. Research on the viral microbiome has been limited (52) and is, in part, due to the technical challenges associated with measuring viruses. We first discuss the measurement issues surrounding the bacterial and fungal microbiome and then briefly address the measurement of the virome.

The most common method of profiling the bacterial microbiome thus far has been to sequence one or more of the variable regions of the highly conserved gene that codes for the small-subunit (16S) of the ribosomal RNA (rRNA) in the bacterial kingdom (53). The variation in the base pairs within the less-conserved regions of the 16S gene enables the identification of bacteria. Similarly, the fungal microbiome has been profiled by sequencing the internal transcribed spacer (ITS) (54) DNA located between the small and the large ribosomal subunit genes. Profiling the 16S (and ITS) variable regions enables researchers to catalog the taxonomic composition of the bacteria (or fungi) in the samples, up to a genus-level resolution. However, the information contained in these short sequences do not directly reveal if these bacteria (or fungi) are alive, or provide information about their metabolic states or their functions within living systems (55,56). Additionally, sequence read-counts do not directly correlate with the absolute bacterial load in the samples, in part, due to the variable copy numbers of the 16S gene in any given organism, dissimilarities of the universal 16S primers to the target genes of some of the microbes, and amplification-related artefacts introduced by the polymerase chain reaction (PCR), which is used to extract and amplify the 16S DNA fragments prior to sequencing (38). Even though 16S profiling continues to become progressively inexpensive for large-scale sequencing of samples taken from sizeable clinical populations, it does not entirely meet the requirements of translational studies in which it is necessary to characterize the microbiome in terms of its strain-level composition (57), metabolic state, and transcriptional profiles. These facets help investigators understand host-pathogen interactions and inform the design of microbiome-based interventions in which specifically chosen microbial strains are administered to patients to repair dysbiotic disease-related microbial communities.

The limitations of 16S profiling are, in part, addressed by metagenomic sequencing (56) in which whole genomes of microbes in biological samples are fractionated and sequenced in a shotgun manner. Since bacterial (and fungal) genomes can have a wide range of genomic sizes, a large number of sequencing reads per sample (i.e., sequencing depth) is required for attaining sufficient coverage of all genes in a community of microbes. Sequencing depth also determines the number of organisms in the samples for which whole-genome profiles can be completely indexed (58). The most distinct advantage of metagenomic sequencing is the ability to use the sequencing reads to construct a catalog of genes in a sample and use this information to arrive at species and strain-level identification (56). Additionally, the relative gene-content enables a direct profiling of the functional attributes of the organisms in the sample. The drawbacks of metagenomic sequencing stem from its substantial cost which increases with the depth of sequencing and the number of samples. Lower sequencing depths result in the complete profiling of only the most abundant species in a sample and thus limit the scope of this approach in translational studies where detection of sparsely distributed and yet substantially influential pathogenic strains is of interest. Since strains can be genetically and thus functionally quite different from each other, strain-level identification can pinpoint gene functions or single-nucleotide polymorphisms (SNPs) that are specific to the strain (57). This information can help detect genetic and metabolic changes in strains under varying interventions or, prospectively, over time. Strain analysis can also detect microevolutionary developments

© Annals of Translational Medicine. All rights reserved.

such as mutation hotspots or horizontal gene transfers in the microbial genetic framework (59).

The choice between 16S and metagenomic sequencing is dictated by both the core objectives of the study and the resources available to the investigator (56). If the goal is an in-depth characterization of the top 20 to 50 most abundant bacterial species or strains along with their functional and metabolic profiles, metagenome sequencing offers distinct advantages. However, if the objective is to develop diagnostic signatures or monitor changes in the entire microbiome community, 16S sequencing is a better alternative. Translational studies also have the choice of including concurrent measurements such as metabolomics and proteomics to supplement findings from 16S profiling. While these measurements increase the informativeness of a study, a carefully developed statistical framework is required to integrate these data sources to construct a more complete view of the microbiome along with its functional profile.

Unlike the bacterial and fungal microbiome, the virome can be measured only after substantial enrichment and purification of circulating viral particles and elimination of non-viral nucleic acids typically using nucleases (52,60). However, these methods do not recover intracellular latent viruses and double-stranded DNA viruses (61). A majority of virome studies have thus focused on RNA viruses and, as a result, the diversity and function of the human virome is not completely understood. After sequencing has been performed for profiling bacterial, fungal or viral populations, assigning these sequences to known taxonomic categories requires the availability of good public databases containing reference sequences of a large number of organisms with known positions in the phylogenetic tree. While these types of databases continue to improve at an impressive rate for bacterial and fungal microbiomes, databases for viromes have lagged behind. Once the composition and diversity of the human virome is more comprehensively cataloged, investigators can expect rapid advances in understanding the dynamics of the virome in diseased and healthy states.

Measuring the microbiome: quality control and data processing

Calibration and continual monitoring for sample contamination and sequencing artefacts are cornerstones to internal validity and accurate quantification of the microbiome. An effective strategy to track any contamination is to include negative controls such as blank water or reagent samples in the processing workflow (62). Negative controls are run through the same PCR-based template extraction and amplification steps as those used for processing study samples. Blank reagent controls test reagents for any inadvertent background contamination. After sequencing, comparisons of reads from the negative controls with the reads from the study samples provide information on whether reagent and/or processing-related contamination exists in the samples. Even if reads from the negative controls are negligible, the researcher may opt for a conservative route and remove any background signals by subtracting the negative control reads from the reads associated with study samples prior to downstream analysis.

Unlike negative controls, positive controls help calibrate the sequencing method. Two types of positive controls are usually included in the sequencing workflow. The simplest positive control is usually comprised of pure strains of *Escherichia coli* (*E. coli*) which produce strong PCR bands of a known size. Inspecting the PCR bands from the *E. coli* samples and the subsequent sequencing reads verifies whether the PCR amplification and the sequencing steps target both the expected organisms and the expected size of fragments of the 16S gene and, additionally, yield the desired number of reads of these organisms. Using the *E. coli* controls enables a qualitative comparison of PCR efficiency of the positive controls relative to the negative controls and the study samples.

The second type of positive control consists of synthetic mock microbial community samples from the Biodefense and Emerging Infectious Research (BEI) Resources of the American Type Culture Collection (ATCC) (Manassas, VA, USA) (63). Two variants of the mock community are usually included in the sequencing workflow, the first containing equimolar concentrations of rDNA operons within the genomic DNA from 21 bacterial strains and the second with the molar rDNA operon concentrations staggered by up to 1,000-fold across the 21 strains. By sequencing these controls in each sequencing plate and run, and examining the resulting reads, investigators can ensure that their amplification, sequencing and taxonomic classification protocols have not introduced substantial bias or distortions in the expected microbiome profiles.

While the inclusion of controls in the sample processing workflows can help monitor for contamination and correct for sequencing-related skews, HTS platforms come with their own share of challenges embedded in data generation (64,65). Errors can be introduced during the preparation of the template or sample libraries, sequencing, imaging or data analysis. Several of the errors associated

© Annals of Translational Medicine. All rights reserved.

with HTS platforms are stabilized and corrected by post-sequencing read demultiplexing (i.e., assigning sequence reads to samples) and quality-checking software designed by the platform manufacturers. However, several other errors and biases that persist after the initial quality checks need to be carefully identified and filtered out prior to taxonomic classification of the reads. Reads are classified into taxonomic categories using bioinformatic algorithms that operate in a sequential manner to (I) trim bases from the edges of the sequence reads that have been flagged as low-quality or less informative by the sequencing platform; (II) stitch paired reads together into one contiguous read (contigs), if the target DNA of interest was sequenced in two segments (also called paired-end sequencing), or construct several contigs if shotgun sequencing was performed; (III) remove sequencing artefacts such as chimeras, which consist of merged sequences from two distinct organisms inadvertently spliced together at the PCR stage; (IV) in the case of 16S profiling, filter out non-16S DNA by comparison with reference databases; and (V) assign taxonomic identities to the sequences in a probabilistic manner by comparing the sequences (or gene-content in case of shotgun sequencing) to those present in reference databases. Several such sequential bioinformatics workflows or pipelines for sequence processing and taxonomic assignment are available for 16S profiling (66-70) and metagenomic sequence analysis (71-76). While there are pros and cons to using each of these existing workflows, there are also large-scale evaluations and comparative studies that discuss the strengths and limitations of each of these pipelines under various conditions (77-80). An in-depth review of these evaluations, in consultation with bioinformatics experts, is recommended so that the investigators can adopt the most appropriate workflow for a given set of measurement choices made within a study.

Statistical considerations in microbiome analysis: a roadmap

A variety of downstream statistical analyses are necessary to analyze taxonomic and functional characteristics of microbiome communities that were measured by either 16S or metagenomic profiling. The design and choice of these statistical analyses is closely connected with the research objectives of the study. When examining the microbiome profiles of study samples, translational studies do not solely aim to catalog the microbial diversity or to arrive at a simple answer to the question of "What is different?" between any two groups of samples. Rather, translational study questions are substantially more complex such as, (I) how much does the microbiome add to the classification accuracy of existing clinical measures? (II) Which quantifiable aspect of the microbiome could be used to monitor treatment-response or the severity, progression or recovery from a disease? (III) Which strains, when introduced into an existing microbiome community, cause the least amount of disruption to the healthy microbiome while displacing the dysbiotic components? (IV) What are the microbiome signatures of a complex immune process such as inflammation or of particular aspects of disease pathogenesis? (V) Which components of the microbiome undergo dynamic changes, and do any of these changes reflect side-effects of a given treatment?

A multi-layered and iterative approach to analysis that appropriately integrates the non-microbiome sources of data with the microbiome data is required to effectively answer these questions. While the design of these analyses in translational studies is dictated by study questions, the design process involves some common elements that includes exploring the data by getting to know the attributes of variables collected, applying the necessary transformations, normalizing their values, inspecting and summarizing the missing values, examining the relationships among these variables by descriptive summaries and visualizations, selecting the dependent and independent variables and the form of statistical models and finally, developing and implementing the statistical models that explore the study hypotheses and questions. Each of these elements is discussed below.

A careful evaluation of the nature and type of variables informs the course of data processing. Notwithstanding careful sample processing, the number of sequencing reads (or read-counts) is inherently non-uniform across samples (81). Thus, analyses performed on the raw read-counts could identify significant differences in taxonomic composition between groups of samples, that are, in part, artefacts associated with the variability in sequencing depth across samples. Although there are analytical approaches, originally developed for RNA-sequencing and gene-expression data, that attempt to adjust for differences in sequencing depth across samples (82), a more robust option, specifically for microbiome data, is to convert the sequence reads to proportions and then log-transform these proportions to bring the values corresponding to abundant and rare taxa to a similar dynamic range (83). These transformations substantially smooth out the non-uniform

© Annals of Translational Medicine. All rights reserved.

variability of sequence reads across samples.

While the number of sequencing reads per sample depends on the body site being studied as well as the number of samples pooled in a given sequencing run, studies typically aim for a sequencing depth that saturates the rarefaction curve for the sequencing platform. However, rarefaction curves vary across platforms and extremely rare members of the microbiome which correspond to taxa with substantially fewer reads (sometimes as few as 20 to 50 reads) are more likely to be artefacts of sequencing errors or clustering algorithms rather than genuine members of the microbial communities being profiled (84). Dropping these taxa from analysis is an effective way to diminish the impact of these errors. Additionally, dropping these rare taxa after the read proportions have been calculated, partially eliminates the correlation of read proportions across microbes and elicits more robust estimates from downstream statistical analysis. It is beneficial to avoid categorization of variables that are either naturally continuous or have at least ≥5 unique values because categorizing a continuous variable results in loss of information by mapping the underlying complex process that created the variable to an artificial simplistic decision process that motivated the categorization (85). Once the continuous and categorical variables in the analysis are coded as numeric constructs, it is advantageous to scale the continuous variables to mean 0 and variance 1, which makes it easier for downstream statistical algorithms to detect associations in the data.

Missing data points are a cause for concern in any experiment as they introduce bias in measurements. It is thus informative to examine the pattern of missingness in the data. If the missingness is not correlated with any particular variable and exists in only a few samples, it is worth adopting multiple imputation schemes (85) to impute the missing values in the interest of retaining as many samples as possible for data analysis. However, if a large number of values are missing for a specific variable, imputation is likely to bias the results. In these cases, it is a reasonable choice to exclude the variable from the analysis.

Following the initial data processing, it is necessary to identify the key independent and dependent variables related to the primary research questions in the study. Descriptive summaries of these variables stratified by clinical characteristics help examine the distribution of these variables. Similarly, summaries stratified by experimental or technical variables such as the sequencing run, batch, time of processing etc., enable the identification of any measurement biases. Visualization of the dependent variables in relation to the independent variables using a variety of graphical representations (86) such as boxplots, histograms etc. allows the investigator to assess the central and extreme values and the distribution of these variables, check assumptions about the data and verify that the assumptions are consistent with pre-existing knowledge in the field. Defining a color palette (87) for the variables of interest at the exploratory stage and consistently associating specific variables with this color palette throughout the analysis enhances clarity of the visualization. For microbiome datasets which are high-dimensional (i.e., number of measured microbes is far greater than the number of study subjects), it is advantageous at the exploratory stage to plot summary measures that represent the state of the microbiome such as the Shannon (88) or the inverse Simpson (89) diversity index, rather than visualize each microbe, one at a time.

The complexity of questions in a translational study makes it unlikely that a single statistical model would adequately answer all facets of the study question. In other words, no single hypothesis or model is guaranteed to capture the true relationships among the microbiome and other measurements that have been collected as part of the study (90). As a result, it is useful to design multiple statistical models using several combinations of independent-dependent variables that best address the study question based on domain knowledge and investigator-generated hypotheses. For example, in a study on pneumonia that has measured both the microbiome and immune-function in the form of multiplexed cytokine assays, the immune-function measurements can be seen as representing the intermediate functional pathways through which the disease processes of pneumonia are influenced by the microbiome or, alternatively, the mechanism through which the microbiome is shaped by the disease process. These hypotheses can be tested using two types of models. In the first, each of the disease-related phenotypes is modeled as the dependent variable with the microbiome, cytokines and other clinical measurements as the independent variables. In the second, each of the cytokines is modeled as the dependent variable. Thinking in advance about how the model findings would be interpreted and how they might answer study questions, simplifies the design of models.

Design choices made in the encoding of variables can help get the most out of statistical modeling. If a variable is continuous, using it directly in the model is substantially

© Annals of Translational Medicine. All rights reserved.

more informative than using either a categorical or binary encoding of the variable (85). This is true for both dependent (also called response) and independent variables. Using a binary response such as disease/no-disease constrains the analysis to use a classification model. While such a model is able to identify microbes that are differentially expressed across disease and no-disease categories, more robust estimates of the effects of microbes on the disease process can be obtained by using a model that employs a continuous phenotype of the disease as the response, and is thus closer to the underlying biology. If the model findings, obtained using a continuous response, reveal a monotonic relationship between the response and a given microbe, it strongly suggests that the microbe is dynamically related to the response. Such insights into the dynamic relationships between the microbiome and the disease process can, in part, address limitations of 16S profiling which is not able to unambiguously reveal if the microbes being studied are alive or dead. If gene-expression and/or functional profiles from metagenomic sequencing are also available, then a continuous response model could determine whether a microbe is upregulating specific genes and/or functional pathways. Carefully selected continuous responses that accurately represent the disease phenotype or treatment-response in translational studies could be valuable for bridging the gap between complex research findings and clinical trial designs with simple patient-response endpoints.

Just as there are several plausible models that can explain the underlying biology, there are several strategies to estimate these models. However, before proceeding to estimation, it is important to consider the unique characteristics of data comprised of microbiome variables and any accompanying "omics" variables measured using HTS platforms. These data are high-dimensional in nature, with the total number of variable measurements far exceeding the number of samples, often by two to three orders of magnitude. Given p variables, there is an exponential number (2^p) of variable combinations that could explain the response (91). Furthermore, these variables are multicollinear with complex covariance structures. Such high-dimensionality and multicollinearity leads to several challenges in model estimation. Univariate approaches that perform hypothesis testing, one variable at a time, underutilize data by not evaluating effects of variable combinations, assuming the orthogonality of variables and ignoring the covariance structure. While a multivariable regression model can address multicollinearity, it can result in a large number of inflated coefficients that overfit the data and do not generalize to new settings. Additionally, building a single multivariable model ignores the many alternative models that are also plausible. Given these challenges, model estimation for microbiome and HTS data in translational studies needs to employ towards a strategy that incorporates multicollinearity, avoids overfitting by incorporating coefficient penalties, combines evidence from a large number of models and estimates coefficients with low variability even with low sample sizes. Ensemble models with penalization (also called regularization) meet all of these criteria (83). By imposing an explicit cost on inflated coefficients, penalized regression improves the generalizability of model findings. Instead of using point estimates to summarize the data, ensemble models can aggregate coefficient estimates over a large collection of models to generate a stable list of influential variables ordered by their importance to the response while also generating confidence estimates for these coefficients. Examples of such ensemble models include frequentist approaches such as random forests (92) and elastic net regression (93) with stability selection (94), which aggregates information over multiple models estimated on bootstrap resamples of the data, and Bayesian approaches that employ either Markov chain Monte Carlo (MCMC) sampling (91) or variational approximations (95) to sample and combine estimates over a high-dimensional space of models.

An objective approach to determine the most accurate model for the data of interest is to conduct an evaluation of all available methods on a simulated dataset that very closely resembles the characteristics of the microbiome and the HTS data under study (83). The simulation pre-specifies the true relationships between the response and the independent variables so that model performance can be calibrated against this pre-determined truth. Since the data used for simulation mimics actual data, it provides a meaningful assessment of model performance on real datasets. Embedding evaluation-based selection of estimation techniques into translational study designs reduces guesswork and enables investigators to choose a single best method for any given study dataset with an emphasis on consistent and accurate performance.

Conclusions

Microbiome research has been making steady inroads into translational studies over the past decade, at a rate that has far outstripped genomics and other high-throughput

© Annals of Translational Medicine. All rights reserved.

technologies. Appropriate study design and statistical analysis within translational studies have the potential to inform a number of aspects in clinical trial-design such as inclusion and exclusion criteria for patient populations, choice of agent or regimen to be used as control, and definitions of primary endpoints. The current broad review of the technical considerations in study design and integrative statistical analysis of microbiome datasets in translational studies attempts to familiarize investigators in the field with the practical limitations of the current techniques for microbiome measurement and analysis along with potential approaches to successfully address these challenges.

Acknowledgements

Funding: This research was supported by the Bill and Melinda Gates foundation grant (OPP1017579).

Footnote

Conflicts of Interest: The author has no conflicts of interest to declare.

References

1. NIH HMP Working Group, Peterson J, Garges S, et al. The NIH Human Microbiome Project. Genome Res 2009;19:2317-23.
2. Sonnenburg JL, Bäckhed F. Diet-microbiota interactions as moderators of human metabolism. Nature 2016;535:56-64.
3. Okeke F, Roland BC, Mullin GE. The role of the gut microbiome in the pathogenesis and treatment of obesity. Glob Adv Health Med 2014;3:44-57.
4. Boulangé CL, Neves AL, Chilloux J, et al. Impact of the gut microbiota on inflammation, obesity, and metabolic disease. Genome Med 2016;8:42.
5. Ussar S, Fujisaka S, Kahn CR. Interactions between host genetics and gut microbiome in diabetes and metabolic syndrome. Mol Metab 2016;5:795-803.
6. Arora T, Bäckhed F. The gut microbiota and metabolic disease: current understanding and future perspectives. J Intern Med 2016;280:339-49.
7. Forbes JD, Van Domselaar G, Bernstein CN. The Gut Microbiota in Immune-Mediated Inflammatory Diseases. Front Microbiol 2016;7:1081.
8. van der Meulen TA, Harmsen H, Bootsma H, et al. The microbiome--systemic diseases connection. Oral Dis 2016;22:719-34.
9. Dalal SR, Chang EB. The microbial basis of inflammatory bowel diseases. J Clin Invest 2014;124:4190-6.
10. Chang C, Lin H. Dysbiosis in gastrointestinal disorders. Best Pract Res Clin Gastroenterol 2016;30:3-15.
11. Rothhammer V, Quintana FJ. Environmental control of autoimmune inflammation in the central nervous system. Curr Opin Immunol 2016;43:46-53.
12. Coit P, Sawalha AH. The human microbiome in rheumatic autoimmune diseases: A comprehensive review. Clin Immunol 2016;170:70-9.
13. Tay WH, Chong KK, Kline KA. Polymicrobial-Host Interactions during Infection. J Mol Biol 2016;428:3355-71.
14. Taylor SL, Wesselingh S, Rogers GB. Host-microbiome interactions in acute and chronic respiratory infections. Cell Microbiol 2016;18:652-62.
15. Shankar J, Solis NV, Mounaud S, et al. Using Bayesian modelling to investigate factors governing antibiotic-induced Candida albicans colonization of the GI tract. Sci Rep 2015;5:8131.
16. Smeekens SP, van de Veerdonk FL, Netea MG. An Omics Perspective on Candida Infections: Toward Next-Generation Diagnosis and Therapy. Front Microbiol 2016;7:154.
17. Kline KA, Lewis AL. Gram-Positive Uropathogens, Polymicrobial Urinary Tract Infection, and the Emerging Microbiota of the Urinary Tract. Microbiol Spectr 2016;4(2).
18. Cribbs SK, Beck JM. Microbiome in the pathogenesis of cystic fibrosis and lung transplant-related disease. Transl Res 2017;179:84-96.
19. Williams B, Landay A, Presti RM. Microbiome alterations in HIV infection a review. Cell Microbiol 2016;18:645-51.
20. Wang X, Yang Y, Huycke MM. Microbiome-driven carcinogenesis in colorectal cancer: models and mechanisms. Free Radic Biol Med 2017;105:3-15.
21. Vogtmann E, Goedert JJ. Epidemiologic studies of the human microbiome and cancer. Br J Cancer 2016;114:237-42.
22. Taur Y, Pamer EG. Microbiome mediation of infections in the cancer setting. Genome Med 2016;8:40.
23. Shankar J, Nguyen MH, Crespo MM, et al. Looking Beyond Respiratory Cultures: Microbiome-Cytokine Signatures of Bacterial Pneumonia and Tracheobronchitis in Lung Transplant Recipients. Am J Transplant 2016;16:1766-78.
24. Sporrer D, Gessner A, Hehlgans T, et al. The Microbiome

© Annals of Translational Medicine. All rights reserved.

and Allogeneic Stem Cell Transplantation. Curr Stem Cell Rep 2015;1:53-9.
25. Langdon A, Crook N, Dantas G. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. Genome Med 2016;8:39.
26. Johanesen PA, Mackin KE, Hutton ML, et al. Disruption of the Gut Microbiome: Clostridium difficile Infection and the Threat of Antibiotic Resistance. Genes 2015;6:1347-60.
27. Gevers D, Kugathasan S, Denson LA, et al. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. Cell Host Microbe 2014;15:382-92.
28. Raes J. Microbiome-based companion diagnostics: no longer science fiction? Gut 2016;65:896-7.
29. Gilbert JA, Quinn RA, Debelius J, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature 2016;535:94-103.
30. Shaw KA, Bertha M, Hofmekler T, et al. Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. Genome Med 2016;8:75.
31. Bizzarro S, Laine ML, Buijs MJ, et al. Microbial profiles at baseline and not the use of antibiotics determine the clinical outcome of the treatment of chronic periodontitis. Sci Rep 2016;6:20205.
32. Gallo RL, Hultsch T, Farnaes L. Recognizing that the microbiome is part of the human immune system will advance treatment of both cancer and infections. J Am Acad Dermatol 2016;74:772-4.
33. Zhang LS, Davies SS. Microbial metabolism of dietary components to bioactive metabolites: opportunities for new therapeutic interventions. Genome Med 2016;8:46.
34. Singh P, Teal TK, Marsh TL, et al. Intestinal microbial communities associated with acute enteric infections and disease recovery. Microbiome 2015;3:45.
35. Enright EF, Gahan CG, Joyce SA, et al. The Impact of the Gut Microbiota on Drug Metabolism and Clinical Outcome. Yale J Biol Med 2016;89:375-82.
36. Donia MS, Fischbach MA. Human microbiota. Small molecules from the human microbiota. Science 2015;349:1254766.
37. Grady NG, Petrof EO, Claud EC. Microbial therapeutic interventions. Semin Fetal Neonatal Med 2016;21:418-23.
38. Bik EM. The Hoops, Hopes, and Hypes of Human Microbiome Research. Yale J Biol Med 2016;89:363-73.
39. Hanson BM, Weinstock GM. The importance of the microbiome in epidemiologic research. Ann Epidemiol 2016;26:301-5.
40. Robinson CK, Brotman RM, Ravel J. Intricacies of assessing the human microbiome in epidemiologic studies. Ann Epidemiol 2016;26:311-21.
41. Kia E, Wagner Mackenzie B, Middleton D, et al. Integrity of the Human Faecal Microbiota following Long-Term Sample Storage. PLoS One 2016;11:e0163666.
42. Song SJ, Amir A, Metcalf JL, et al. Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. mSystems 2016;1(3). pii: e00021-16.
43. Shaw AG, Sim K, Powell E, et al. Latitude in sample handling and storage for infant faecal microbiota studies: the elephant in the room? Microbiome 2016;4:40.
44. Hsieh YH, Peterson CM, Raggio A, et al. Impact of Different Fecal Processing Methods on Assessments of Bacterial Diversity in the Human Intestine. Front Microbiol 2016;7:1643.
45. Clooney AG, Fouhy F, Sleator RD, et al. Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. PLoS One 2016;11:e0148028.
46. Wagner J, Paulson JN, Wang X, et al. Privacy-preserving microbiome analysis using secure computation. Bioinformatics 2016;32:1873-9.
47. Muir P, Li S, Lou S, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol 2016;17:53.
48. Keiding N, Louis TA. Perils and potentials of self-selected entry to epidemiological studies and surveys. J R Stat Soc Ser A Stat Soc 2016;179:319-76.
49. Pomp ER, Van Stralen KJ, Le Cessie S, et al. Experience with multiple control groups in a large population-based case-control study on genetic and environmental risk factors. Eur J Epidemiol 2010;25:459-66.
50. Han MK, Dransfield MT, Martinez F. Chronic obstructive pulmonary disease: Definition, clinical manifestations, diagnosis, and staging UpToDate. Waltham, MA: UpToDate; 2016. Cited 2016 Nov 6. Available online: http://www.uptodate.com/contents/chronic-obstructive-pulmonary-disease-definition-clinical-manifestations-diagnosis-and-staging
51. Vashist SK, Schneider EM, Luong JH. Commercial Smartphone-Based Devices and Smart Applications for Personalized Healthcare Monitoring and Management. Diagnostics (Basel) 2014;4:104-28.
52. Zou S, Caler L, Colombini-Hatch S, et al. Research on the human virome: where are we and what is next. Microbiome 2016;4:32.
53. Woo PC, Lau SK, Teng JL, et al. Then and now: use of

© Annals of Translational Medicine. All rights reserved.

16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin Microbiol Infect 2008;14:908-34.
54. Schoch CL, Seifert KA, Huhndorf S, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A 2012;109:6241-6.
55. Willis AL, Calton JB, Carr TF, et al. Dead or alive: Deoxyribonuclease I sensitive bacteria and implications for the sinus microbiome. Am J Rhinol Allergy 2016;30:94-8.
56. Ranjan R, Rani A, Metwally A, et al. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun 2016;469:967-77.
57. Zhang C, Zhao L. Strain-level dissection of the contribution of the gut microbiome to human metabolic disease. Genome Med 2016;8:41.
58. Sims D, Sudbery I, Ilott NE, et al. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 2014;15:121-32.
59. Wang J, Jia H. Metagenome-wide association studies: fine-mining the microbiome. Nat Rev Microbiol 2016;14:508-22.
60. Conceição-Neto N, Zeller M, Lefrère H, et al. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. Sci Rep 2015;5:16532.
61. Rascovan N, Duraisamy R, Desnues C. Metagenomics and the Human Virome in Asymptomatic Individuals. Annu Rev Microbiol 2016;70:125-41.
62. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 2014;12:87.
63. Kozich JJ, Westcott SL, Baxter NT, et al. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 2013;79:5112-20.
64. Kircher M, Heyn P, Kelso J. Addressing challenges in the production and analysis of illumina sequencing data. BMC Genomics 2011;12:382.
65. Zhou X, Rokas A. Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. Mol Ecol 2014;23:1679-700.
66. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75:7537-41.
67. Kuczynski J, Stombaugh J, Walters WA, et al. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Bioinformatics 2011;Chapter 10:Unit 10.7.
68. White JR, Maddox C, White O, et al. CloVR-ITS: Automated internal transcribed spacer amplicon sequence analysis pipeline for the characterization of fungal microbiota. Microbiome 2013;1:6.
69. Hildebrand F, Tadeo R, Voigt AY, et al. LotuS: an efficient and user-friendly OTU processing pipeline. Microbiome 2014;2:30.
70. Albanese D, Fontana P, De Filippo C, et al. MICCA: a complete and accurate software for taxonomic profiling of metagenomic data. Sci Rep 2015;5:9743.
71. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;15:R46.
72. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 2015;12:902-3.
73. Ounit R, Wanamaker S, Close TJ, et al. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 2015;16:236.
74. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 2016;7:11257.
75. Silva GG, Green KT, Dutilh BE, et al. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics 2016;32:354-61.
76. Ulyantsev VI, Kazakov SV, Dubinkina VB, et al. MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data. Bioinformatics 2016;32:2760-7.
77. Peabody MA, Van Rossum T, Lo R, et al. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinformatics 2015;16:363.
78. Majaneva M, Hyytiäinen K, Varvio SL, et al. Bioinformatic Amplicon Read Processing Strategies Strongly Affect Eukaryotic Diversity and the Taxonomic Composition of Communities. PLoS One 2015;10:e0130035.
79. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep 2016;6:19233.
80. Richardson RT, Bengtsson-Palme J, Johnson RM. Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. Mol Ecol Resour

© Annals of Translational Medicine. All rights reserved.

2017;17:760-9.
81. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 2014;10:e1003531.
82. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106.
83. Shankar J, Szpakowski S, Solis NV, et al. A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses. BMC Bioinformatics 2015;16:31.
84. He Y, Caporaso JG, Jiang XT, et al. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. Microbiome 2015;3:20.
85. Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, 2015.
86. Wickham H. ggplot2: Elegant Graphics for Data Analysis (Use R!). 1st ed. New York: Springer-Verlag New York Incorporated, 2010.
87. Harrower M, Brewer CA. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. Cartogr J 2003;40:27-37.
88. Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal 1948;27:379-423.
89. Simpson EH. Measurement of Diversity. Nature 1949;163:688.
90. Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat Sci 2001;16:199-231.
91. George EI, McCulloch RE. Approaches for Bayesian variable selection. Statistica Sinica 1997;7:339-73.
92. Breiman L. Random Forests. Mach Learn 2001;45:5-32.
93. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005;67:301-20.
94. Meinshausen N, Bühlmann P. Stability selection: Stability Selection. J R Stat Soc Series B Stat Methodol 2010;72:417-73.
95. Blei DM, Kucukelbir A, McAuliffe JD. Variational Inference: A Review for Statisticians. 2016. Available online: http://arxiv.org/abs/1601.00670

Cite this article as: Shankar J. Insights into study design and statistical analyses in translational microbiome studies. Ann Transl Med 2017;5(12):249. doi: 10.21037/atm.2017.01.13

© Annals of Translational Medicine. All rights reserved.