McDonald *et al. Microbiome* (2015) 3:52
DOI 10.1186/s40168-015-0117-2

Microbiome

REVIEW Open Access

Context and the human microbiome

Daniel McDonald[1,2], Amanda Birmingham[3] and Rob Knight[4,5*]

Abstract

Human microbiome reference datasets provide epidemiological context for researchers, enabling them to uncover new insights into their own data through meta-analyses. In addition, large and comprehensive reference sets offer a means to develop or test hypotheses and can pave the way for addressing practical study design considerations such as sample size decisions. We discuss the importance of reference sets in human microbiome research, limitations of existing resources, technical challenges to employing reference sets, examples of their usage, and contributions of the American Gut Project to the development of a comprehensive reference set. Through engaging the general public, the American Gut Project aims to address many of the issues present in existing reference resources, characterizing health and disease, lifestyle, and dietary choices of the participants while extending its efforts globally through international collaborations.

Keywords: Microbiome, American Gut Project, Reference database, Meta-analysis

Review

Background

In the last few years, the study of the bacteria, archaea, microbial eukaryotes, and viruses that inhabit the human body (particularly the large intestine) has revealed a remarkable biological and functional diversity [1–6]. These organisms, collectively known as the microbiome, potentially outnumber human cells in 10:1 [7] and vastly expand on the functional capabilities provided by our genomes. Disruption in these microbial communities, also known as dysbiosis, has been causatively associated by transferring microbiomes and phenotypes to mice associated with human Kwashiorkor [8] (a wasting disease endemic to Africa) and obesity [9]. Numerous correlative associations in humans and mouse models have also been observed in a broad spectrum of complex diseases including autism spectrum disorder [10], inflammatory bowel disease [11], type 2 diabetes [12], colorectal cancer [13], depression [14] (see [15] for a detailed review on the brain-gut-microbe axis), and more.

The implication of the microbiome in human health is immense, with prospects for novel medical products including therapeutics and clinical assays. This has led to large investments in both academia [3] and industry [16]. Although such research could have a profound impact on human society both in first- and third-world countries, we are just scratching the surface of understanding the complexity of this vital organ. As such, identifying means that improve the pace of research is arguably a matter of human health on a global scale.

A crucial and missing component of microbiome research is a robust and comprehensive reference set of microbiome samples and metadata about those samples that are available for public, unrestricted use. Such a dataset would characterize what we know about diversity of the human microbiome and its relationship to the health and lifestyle choices of individuals, providing much-needed context against which to compare findings of focused studies such as those on particular disease populations. This reference would allow researchers to place their study in the framework of what is already known in order to better interpret observed patterns (compelling examples of this can be found in [17, 18]). It would also enable stringent hypothesis testing and evaluation of effect sizes. A robust reference dataset must be built on top of a cross-sectional study design in order to understand the variation in the population, while also including rich longitudinal components to enable an understanding of how species structure changes over time.

In this review, we highlight the importance of reference sets in human microbiome research, limitations of

* Correspondence: robknight@ucsd.edu
[4]Department of Pediatrics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[5]Department of Computer Science and Engineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
Full list of author information is available at the end of the article

existing resources, technical challenges to employing reference sets, examples of prototypical reference usages, and contributions of the American Gut Project to addressing some of these issues. Discussion will focus on the 16S ribosomal RNA (16S rRNA) gene, which is a popular locus for use in microbiome studies over a wide range of environment types [19–23] and is the core locus assayed in the American Gut Project. Construction of references based on other loci is important for studying microbial eukaryotes, viruses, and interactions between these organisms, but high-throughput study of these other components of the community is not yet cost-effective.

Importance of reference sets in human microbiome research

The community structure of the human microbiome is the result of a multifactorial process that involves succession over time [24], is influenced by host genetics [25], and is affected by lifestyle choices [26, 27]. Communities are made up of thousands of microbial species, with the predominant microbial biomass residing in the human large intestine. Fascinatingly, within the human gastrointestinal tract, it appears that multiple organisms are capable of fulfilling common ecological niches, leading to remarkably different microbial communities that possess similar functional potential [3]. Furthermore, while variations in the human genome are minute across the population, variations in the human microbiome on geographical and temporal scales are immense [28, 29]. Despite investments of hundreds of millions of dollars, we still do not understand the distribution of community structures in healthy individuals [30], but we do know that when studies of the microbiome are performed without a concern for integration with existing studies, effects of significant biological importance can be easily missed [31].

A well-characterized reference dataset can be used to test hypotheses and, conversely, to derive testable hypotheses from the reference itself. For instance, inflammatory bowel disease has been observed to be associated with a microbial dysbiosis index (MD-index) that is the ratio of the relative abundances of a set of pro-inflammatory taxa to a set of anti-inflammatory taxa [32]; a robust reference set would allow assessment of the hypothesis that diet or lifestyle factors are strongly correlated to this index within the general public as well. In an opposite example, a significant correlation between diversity and time of the year was observed in the American Gut reference set [33]. Because it appears that individuals have a higher diversity during the holiday season in the US, one might hypothesize that it is the holidays and not the time of the year that drives the correlation—possibly due to changes in exercise and diet patterns. This putative effect can then be tested once the project acquires sufficient samples from western countries in the southern hemisphere.

A comprehensive reference dataset will also help researchers make rational decisions about sample size by enabling power calculations, which can greatly impact the utility of a study [34]. Such a dataset is also crucially necessary to support characterization of the effect sizes of variables (e.g., antibiotic use). Within the microbiome field, effect size for many variables of interest is not yet well understood, and many that are important in diseases with complex etiologies such as autism [10] are likely to be small. Well-characterized references offer the possibility for a researcher to expand their dataset by pulling reference samples to augment their own [29], particularly when meta-analysis (i.e., combination of summarized data from multiple studies) is taken into consideration during the design phase for a study.

Limitations of existing reference sets

The $173 million NIH-initiated Human Microbiome Project (HMP) set out to characterize the human microbiome at a population scale and to define standard reference datasets to be used for human microbiome research [35]. The resulting 16S rRNA datasets are composed of samples from 242 individuals, all of whom were medical students in the USA and were certified healthy by medical professionals. Thousands of samples were collected from these individuals at one to three time points, covering 15 to 18 sampling sites depending on the sex of the individual. These samples were evaluated using two different regions of the 16S gene (leading to two distinct datasets—V1-3 and V3-5) [31] and were processed at four different sequencing centers. Phenotypic information about the individuals was collected, but while the sequence data associated with the samples are publically available, access to any de-identified information about the individuals requires rigorous approval mechanisms.

Although the HMP generated an incredible volume of data, numerous design, technical, and access decisions affecting the HMP dataset have made reuse challenging. For instance, the decision to sample a few people extensively rather than a large number of people minimally (i.e., a cross-sectional study design) led to observation of only a small fraction of the diversity present with the population [28] and resulted in small sample sizes for different stratifications in the dataset [36], effectively removing the potential to observe demographic or regional differences. The choice to sequence multiple loci within the 16S rRNA gene resulted in data that are impractical to combine due to technical bias as amplification performance differs between primers [31, 37]. Furthermore, because the study design was not sufficient to elucidate the effect of employing multiple sequencing centers (which has been observed in other contexts; see the

Microbiome Quality Control Project (MBQC) [38]), this issue must still be actively evaluated to assess the potential for technical biases. Host information, such as age and sex, are nearly prohibitive to access, requiring dean-level signatures for each individual piece of metadata, which makes explaining any systematic patterns in the data impossible without knowing in advance what pattern one expects to see. The end result is that use of the HMP 16S rRNA as a robust reference set has proven difficult.

In contrast to the HMP, the Global Gut project [28] set out to characterize microbial diversity at spatial and temporal scales. To do this, the researchers collected samples from three distinct populations (US citizens, Malawians, and Venezuelan Amerindians), the latter two of which are culturally distinct from western populations. Within each population, samples were collected cross-sectionally over an age gradient. Notably, the two non-western populations appear to be completely distinct from the western individuals, suggesting the limited population size and emphasis of the HMP grossly underestimate the variation in community structure across the human race. However, the populations do intersect on samples collected from infants, suggesting that it is potentially lifestyle, diet, or environmental choices that shape our microbiomes as we age (including interaction with our genetic predisposition [25]). Although the sequence data are readily available for reuse, the distribution of many of the study variables is not approved, limiting the long-term usefulness of the samples. (It should be noted that the Global Gut did not intend to be a reference for microbiome research, but the populations represented in the dataset are extremely difficult to collect samples from and have shown to be useful in adding perspective for independent projects [29, 39]).

Lack of access to the full set of metadata variables associated with these earlier studies is crippling, as interpretation of the observational data can only happen within the context of the collected variables. From a practical standpoint, if a systematic pattern is observed in the data, but there are not any variables that explain the pattern, then the researcher cannot support a hypothesis about the pattern without collecting new information (which may be impractical and impossible or introduce recall bias). Similarly, confidence in the face of confounding variables is reduced if only a limited number of variables are tracked. As a concrete example, if researchers broadly characterize subjects by diet type (e.g., vegan) and observe an effect, the researchers will be unable to assess whether the effect is due to the diet type itself, differential fiber consumption, protein source, etc., unless these other variables are recorded. Given that researchers typically do not know the answer in advance of a study, it is imperative that study designs strive to collect as much information as feasible.

Technical challenges to employing reference sets

Even a well-designed and carefully collected reference must be employed with caution in order to minimize spurious variation and contain necessary computational effort. The first of these needs arises since reference-based analyses assume that any systematic compositional differences inherent in the data outweigh any technical variation, which is particularly problematic when combining data generated from different protocols or platforms [31]. In fact, biological conclusions can be driven by technical variation even if the researchers are careful (as in [40], where samples were found to cluster by the extraction kit used), which underscores the need for accepted community standards for sample handling, sequencing, and data analysis in order to minimize the potential for introducing such variation. Bioinformatic strategies to mitigate any remaining variation, such as trimming sequences to a common length between studies, have shown to help normalize platform bias [29]. Sometimes, stronger measures are necessary: for example, the American Gut Project received samples from self-reported healthy individuals that contained levels of gammaproteobacteria beyond anything previously observed in healthy populations (although similar to those observed in samples from ICU patients [*manuscript in prep*]). It was determined that these blooms likely stemmed from the shipping conditions for some samples. The blooms can be bioinformatically subtracted from the dataset [*manuscript in prep*] by removing organisms observed to bloom (as has been observed to happen in storage [*unpublished observations*]). As a result, any meta-analysis that leverages the American Gut data must perform this same subtraction in order to equalize bias that the filter introduces. Ongoing studies of stability [41, 42] are explicitly exploring the effect of different types of storage effects so that they can be controlled for as necessary in the future.

Once technical variation has been minimized, the comparative analysis can begin. Many researchers, particularly those at remote sites, do not have access to large-scale compute instruments and must rely on commodity hardware for data analysis; this creates the temptation to employ analysis techniques that require as little computation as possible. However, some such techniques are particularly vulnerable to artifacts caused by combining dissimilar datasets.

An exemplar of this issue is the assignment of operational taxonomic units (OTUs). The primary data type used in analysis of a microbiome study is the OTU table [43, 44], a matrix in which the rows represent observations (OTUs), the columns represent samples, and the elements correspond to the number of counts of a given observation within a sample. In order to be comparable, a reference and a study must have their sequence data assigned to a shared set of OTUs (i.e., partitioned into a common set of bins). OTUs themselves are clusters of similar sequences, with the similarity threshold generally set at 97 % by sequence identity, and are typically determined in one of three ways as summarized in Table 1 (for a comprehensive review of OTU picking, please see [45]; each of these methods is named in terms of its OTU reference, but nota bene that this represents a distinct concept from that of the reference datasets discussed throughout). The first is a closed-reference approach in which all the sequence data for the input study and the microbiome reference set are compared against a curated 16S rRNA database such as Greengenes [46] to identify which known OTUs are represented. This is computationally tractable even for very large studies since the evaluation of every sequence is independent of every other and since the reference dataset's OTU assignments can be computed just once (and in advance). The second strategy, known as de novo picking, defines novel OTUs based on the sequences in a study. This is computationally expensive, as all the data must be maintained in memory in order to determine the clusters, and the process is very complex to parallelize. The third approach, open-reference picking, is a hybrid method in which sequences are first compared to a database of known OTUs as described above, after which those that fail to match to a known OTU are then put through a de novo step.

Studies employing a reference set typically rely on the closed-reference approach to minimize compute since only the input study need be evaluated and can be done so in an embarrassingly parallel fashion. Another benefit is that the closed-reference strategy is unlikely to result in OTUs composed of non-16S sequence, as the reference is expected to only contain 16S exemplars; furthermore, comprehensive references like Greengenes typically contain only near-full-length reads, thus allowing researchers to combine data represented by multiple variable regions. Of course, any annotation information about the reference, such as the phylogenetic relationship between the data contained or annotations such as taxonomy, can be attached to the input study data "for free." Unfortunately, this strategy can only classify sequences that are reasonably similar to those in the reference database. Combining studies with differential representation in the reference (e.g., samples from different environments) can lead to statistically significant patterns in the data that are not driven by the underlying biology. As an example, imagine three samples A, B, and C where A is composed of *Escherichia coli* and both B and C are composed of *Escherichia coli* coli and *Bacillus subtilis*. If the reference is only composed of *Escherichia coli*, then all three samples will appear to be quite similar. However, if the reference includes *Bacillus subtilis*, then the conclusion drawn is quite different as A would be less similar to B and C.

In contrast to closed-reference, a de novo approach consumes more computational resources but requires no pre-existing reference and allows a researcher to assign OTUs to as much of the data as possible, including OTUs never before observed. It is capable of producing phylogenies

Table 1 A comparison of OTU-picking strategies

Strategy	Pros	Cons	Data combination bias
Closed-reference	• Is extremely parallelizable	• Is limited to finding diversity present in OTU reference	• May show large bias if combining studies with differential representation in the reference
	• Computes reference assignments only once		
	• Is highly unlikely to retain non-16S sequences		
	• Supports and reads fragments from multiple loci		
	• Gets the phylogeny and taxonomy for free		
De novo	• Utilizes all of the sequences	• Must hold all sequence data in memory	• May generate spurious OTUs if combining studies with differential error profiles
	• Requires no OTU database	• Is very complex to parallelize	
	• Can group organisms distinct from anything seen before	• Produces spurious OTUs without pre-filtering	
	• May produce phylogenies sensitive to subtle differences in OTUs	• Is infeasible if data are from multiple loci	
		• Must redo OTU picking with all data being combined	
Open-reference	• Leverages an OTU database but also utilizes sequences that do not match to that database	• Produces spurious OTUs without pre-filtering	• Shows less bias due to differential diversity representation than closed-reference
		• Is infeasible if data are from multiple loci	
	• Is modestly parallelizable	• Must redo OTU picking with all data being combined	• Shows less bias due to differential error profiles than de novo

sensitive to subtle differences in OTUs (as the representative member for an OTU is an actual study sequence), but this very sensitivity means that contamination in the data (e.g., non-16S sequence such as phiX) will also be clustered into OTUs unless the contamination is explicitly filtered out prior to OTU picking and that the method is not suitable for data drawn from multiple variable regions (as highlighted in [31]). Additionally, the distinct error profiles of the studies being combined (which can stem from the 16S protocol, variation in the master PCR mix, error profiles of the sequencing instrument used, etc.) may lead to spurious, study-specific OTUs (for example, the GC bias of the Illumina platform can lead to sequences that contain more GC than another platform which can result in OTUs specific to the platform even if the biological origin of the amplicon is the same). As a result, a meta-analysis that uses a de novo strategy must redo OTU picking after combining the sequence data from the studies.

The hybrid open-reference method steers a middle course: Since data that are not represented in the reference are recovered, bias driven by differential representation in the reference is reduced. In addition, since the amount of data being fed into the de novo step is minimized, the impact of study-specific error profiles is diminished. Open-reference OTU picking is modestly parallelizable and can be augmented with techniques such as use of a random subsample when constructing the intermediate de novo reference in order to accelerate its performance (details on the procedure can be found in [45] but are generally handled in by software without end-user intervention). However, open-reference picking shares a number of drawbacks with the de novo strategy, including necessity for prefiltering, unsuitability for data from multiple variable regions, and necessity for re-picking when combining studies together. Of course, given the continued expansion of computational resources, sequencing throughput, and completed genomes available, optimal strategies for overcoming technical hurdles and enabling meta-analysis will require ongoing re-assessment.

Examples of reference set usage

One of the first studies to combine multiple microbiome datasets (which these researchers are aware of) was the work by Lozupone and Knight [47], which aggregated sequence data from hundreds of studies in order to determine environmental factor(s) that explained the observed differences in microbial community structure. They discovered that data from samples collected in the natural environment across a multitude of gradients (e.g., pH, temperature, atmospheric pressure) separated primarily based on whether the samples originated from saline or non-saline environments—despite the substantial technical differences between studies. Fascinatingly, when these same data were combined with samples collected from vertebrate guts, the primary variation in the data was explained by whether the samples were environmental or host associated [1], implying that an extremely high degree of specialization has occurred in the microbial communities of vertebrate guts (which is particularly interesting given the difference in evolutionary time that environmental microbial communities have had to specialize relative to the time that vertebrates have existed). While this meta-analysis did not employ a reference set of the type discussed here, it has itself become a de facto reference set that has subsequently been employed for comparison with numerous other studies [48–50].

More recently, a re-evaluation of a longitudinal study aimed at exploring succession in microbial communities within an infant (for the original study, see [24]) was performed using the HMP as context [31]. While the original work showed a distinct increase in the diversity of the infant's fecal community through the first few years of life, putting its results in context immediately clarified the trajectory of succession by showing that the microbiome moved from resembling a vaginal community (which makes sense given the mode of birth, see [51] for a study on the effects of delivery mode on the infant microbiome) to resembling a fecal community. Visualizing longitudinal microbiome studies as animations (see [52] for a movie of the re-evaluation of the aforementioned infant longitudinal data), particularly in the context of a reference, has been so useful that the ability was recently added into EMPeror [53], a common visualization tool for ordination plots generated from microbiome data.

Meta-analyses are becoming more widespread as computational power increases, sometimes employing past studies that were not intended as reference sets in that new role. Moeller et al. [29] reused the Global Gut [28] data to paint a compelling picture of the coevolution of hominids and their gut communities, highlighting a departure that humans have appeared to take with respect to our closest ancestors. The data suggest that the rate of change in the human microbiome is significantly higher since divergence with chimpanzee, particularly in US adults, including a significant decrease in alpha diversity. The motivation to reuse the Global Gut data was access to samples collected from hunter-gatherer groups as well as western adults, enabling the researchers to test the hypothesis that hunter-gatherer groups are more similar from a microbial perspective to our closest ancestors potentially due to the dramatic dietary differences that exist between these groups and western populations. However, the sample size for any given age group and population combination (e.g., infant Malawians) within Global Gut was relatively small, so it would be interesting to revisit this and see what the pattern of coevolution is against a reference that contains a larger number of samples for different age groups.

Contributions of the American Gut Project

The American Gut Project set out to build a comprehensive open-source and open-access microbiome 16S rRNA reference dataset for the scientific community to use. It relies on a crowd-funding model that allows for broad reach across the US population and is set up so that virtually anyone can participate (with the exception of convicted felons and children younger than 6 weeks old). Individuals can elect to receive a collection kit in exchange for a contribution to the project. Though the sample population is not free from bias (being shifted toward older Caucasians interested in their own health), the variability encompassed by the project vastly exceeds that of the HMP [36]. In addition, the project has recently expanded internationally to the UK and Australia to reduce participant overhead for shipping samples (although, to minimize the introduction of technical variability, all samples are extracted at one site, UC San Diego). All participants in the project are consented under protocol #141853 approved by the University of California San Diego's Human Research Protection Program (HRPP); the protocol specifies that all non-identifying data collected will be deposited into the public domain. Each participant is presented with a HRPP-approved questionnaire that covers diet, lifestyle, and health history, including a NIH-validated food frequency questionnaire [54]. The infrastructure to support electronic consent, questionnaires, localization for international portals, and management of over 22,000 bar-coded samples has opened the doors for external researchers and the general public alike to perform their own experiments using the framework of the American Gut Project.

The American Gut Project is a subset of the Earth Microbiome Project (EMP) [19], which has been instrumental in advocating for adherence to the standards of the Genomics Standards Consortium, including minimum information about a marker gene sequence (MIMARKS) [55]—a suite of standards defining variables to be collected within a marker gene survey for virtually any environment imaginable. The EMP and American Gut also follow published sequencing protocols [56] that aim to normalize technical bias for microbiome studies and employ the Biological Observation Matrix (BIOM) [44] specification as a standard and computationally efficient means to represent the resulting large, sparse-omics datasets and their sample and observation metadata. All data are de-identified and deposited into the public domain as quickly as possible via the European Bioinformatics Institute (EBI), which is part of the International Nucleotide Sequence Database Consortium (INSDC). American Gut has taken a further step by providing executable IPython [57] Notebooks allow others to reproduce and modify the analyses being performed on the data. All code for the project is hosted on Github in the "biocore" organization and is available under the BSD license, and all code and binaries used by the project are open-source.

Although the American Gut is useful, by design it is not intended to provide an unbiased population but rather to harness crowd funding and public enthusiasm to uncover the range of extant microbiomes. Given this fact, many questions could be best addressed by instead adding microbiome components to existing carefully designed cohorts such as NHANES [58], the Nurses' Health Study [59], and TwinsUK [60]. Relevant areas of inquiry include relating the microbiome to heart disease, cancer, stroke, cognitive abilities, and host genetics, as well as leveraging new avenues to assess sources of technical variation. These studies offer the unique potential to build off of their already well-characterized populations.

Conclusions

Research is never performed in isolation. It is built upon the foundations laid by prior knowledge and evaluated in the context of present knowledge. However, if data are not collected with a view toward integration, or if rich reference points do not exist, research is effectively performed in a vacuum. These are some of the challenges that a common reference can help to address, and the American Gut is a widely collaborative, carefully structured project that aims to provide such a reference. The establishment of a comprehensive reference encourages widespread use of standard protocols, since normalization of technical variation is essential when comparing results to the reference and assessing the significance of a study against the background population. Application of context-aware study designs that adhere to community-accepted standards used by references like the American Gut should minimize the time until microbiome research findings become medically actionable.

Abbreviations

BIOM: Biological Observation Matrix; EBI: European Bioinformatics Institute; EMP: Earth Microbiome Project; HMP: Human Microbiome Project; HRPP: Human Research Protection Program; INSDC: International Nucleotide Sequence Database Consortium; MBQC: Microbiome Quality Control Project; MD-index: microbial dysbiosis index; MIMARKS: minimum information about a marker gene sequence; OTU: operational taxonomic unit.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DM conceived of the review and drafted the manuscript. RK participated in the structure of the review and provided critical feedback. AB helped in the design and draft of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by NSF IGERT award 1144807. The author would like to thank Aaron Clauset, Robin Dowell, and Merete Eggesbø for their constructive comments during the review of the manuscript.

Author details

[1]Department of Computer Science, University of Colorado at Boulder, Boulder, CO 80304, USA. [2]BioFrontiers Institute, University of Colorado, 3415 Colorado Avenue, Boulder, CO 80304, USA. [3]Center for Computational Biology & Bioinformatics, Department of Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA. [4]Department of Pediatrics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA. [5]Department of Computer Science and Engineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA.

Received: 15 June 2015 Accepted: 28 September 2015
Published online: 04 November 2015

References

1. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol. 2008;6(10):776–88. doi:10.1038/nrmicro1978.
2. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. Science. 2009;326(5960):1694–7. doi:10.1126/science.1177486.
3. Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14. doi:10.1038/nature11234.
4. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59–65. doi:10.1038/nature08821.
5. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature. 2010;466(7304):334–8. doi:10.1038/nature09199.
6. Andersen LO, Vedel Nielsen H, Stensvold CR. Waiting for the human intestinal Eukaryotome. ISME J. 2013;7(7):1253–5. doi:10.1038/ismej.2013.21.
7. Savage DC. Microbial ecology of the gastrointestinal tract. Annu Rev Microbiol. 1977;31:107–33. doi:10.1146/annurev.mi.31.100177.000543.
8. Smith MI, Yatsunenko T, Manary MJ, Trehan I, Mkakosya R, Cheng J, et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. Science. 2013;339(6119):548–54. doi:10.1126/science.1229000.
9. Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, Kau AL, et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science. 2013;341(6150):1241214. doi:10.1126/science.1241214.
10. Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell. 2013;155(7):1451–63. doi:10.1016/j.cell.2013.11.024.
11. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. Nature. 2011;474(7351):307–17. doi:10.1038/nature10209.
12. Everard A, Cani PD. Diabetes, obesity and gut microbiota. Best Pract Res Clin Gastroenterol. 2013;27(1):73–83. doi:10.1016/j.bpg.2013.03.007.
13. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe. 2013;14(2):207–15. doi:10.1016/j.chom.2013.07.007.
14. Maes M, Kubera M, Leunis JC, Berk M. Increased IgA and IgM responses against gut commensals in chronic depression: further evidence for increased bacterial translocation or leaky gut. J Affect Disord. 2012;141(1):55–62. doi:10.1016/j.jad.2012.02.023.
15. O'Mahony SM, Clarke G, Borre YE, Dinan TG, Cryan JF. Serotonin, tryptophan metabolism and the brain-gut-microbiome axis. Behav Brain Res. 2014;277:32–48. doi:10.1016/j.bbr.2014.07.027.
16. Reardon S. Microbiome therapy gains market traction. Nature. 2014;509(7500):269–70. doi:10.1038/509269a.
17. Antonio Gonzalez YVB, Rob Knight. The assembly of an infant gut microbiome framed against healthy human adults. 2012. https://www.youtube.com/watch?v=Pb272zsixSQ. Accessed Nov 2014.
18. Antonio Gonzalez YVB, Rob Knight. Gut ecosystem restoration via fecal transplantation. 2014. https://www.youtube.com/watch?v=-FFDqhM4pks. Accessed Nov 2014.
19. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC Biol. 2014;12:69. doi:10.1186/s12915-014-0069-1.
20. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505(7484):559–63. doi:10.1038/nature12820.
21. Lamendella R, Strutt S, Borglin S, Chakraborty R, Tas N, Mason OU, et al. Assessment of the Deepwater Horizon oil spill impact on Gulf coast microbial communities. Front Microbiol. 2014;5:130. doi:10.3389/fmicb.2014.00130.
22. Willing B, Halfvarson J, Dicksved J, Rosenquist M, Jarnerot G, Engstrand L, et al. Twin studies reveal specific imbalances in the mucosa-associated microbiota of patients with ileal Crohn's disease. Inflamm Bowel Dis. 2009;15(5):653–60. doi:10.1002/ibd.20783.
23. Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T, et al. The seasonal structure of microbial communities in the Western English Channel. Environ Microbiol. 2009;11(12):3132–9. doi:10.1111/j.1462-2920.2009.02017.x.
24. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, et al. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A. 2011;108 Suppl 1:4578–85. doi:10.1073/pnas.1000081107.
25. Goodrich Julia K, Waters Jillian L, Poole Angela C, Sutter Jessica L, Koren O, Blekhman R, et al. Human genetics shape the gut microbiome. Cell. 2014;159(4):789–99. doi:10.1016/j.cell.2014.09.053.
26. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011;334(6052):105–8. doi:10.1126/science.1208344.
27. Kang SS, Jeraldo PR, Kurti A, Miller ME, Cook MD, Whitlock K, et al. Diet and exercise orthogonally alter the gut microbiome and reveal independent associations with anxiety and cognition. Mol Neurodegener. 2014;9:36. doi:10.1186/1750-1326-9-36.
28. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012;486(7402):222–7. doi:10.1038/nature11053.
29. Moeller AH, Li Y, Mpoudi Ngole E, Ahuka-Mundeke S, Lonsdorf EV, Pusey AE, et al. Rapid changes in the gut microbiome during human evolution. Proc Natl Acad Sci U S A. 2014;111:16431–5. doi:10.1073/pnas.1419136111.
30. Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, et al. Rethinking "enterotypes". Cell Host Microbe. 2014;16(4):433–7. doi:10.1016/j.chom.2014.09.013.
31. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vazquez-Baeza Y, et al. Meta-analyses of studies of the human microbiota. Genome Res. 2013;23(10):1704–14. doi:10.1101/gr.151803.112.
32. Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15(3):382–92. doi:10.1016/j.chom.2014.02.005.
33. American-Gut-Project. Alpha Diversity Notebook. 2015. http://nbviewer.ipython.org/github/biocore/American-Gut/blob/master/ipynb/Alpha diversity notebook.ipynb. Accessed Feb 2015.
34. Kuczynski J, Costello EK, Nemergut DR, Zaneveld J, Lauber CL, Knights D, et al. Direct sequencing of the human microbiome readily reveals community differences. Genome Biol. 2010;11(5):210. doi:10.1186/gb-2010-11-5-210.
35. Human Microbiome Project C. A framework for human microbiome research. Nature. 2012;486(7402):215–21. doi:10.1038/nature11209.
36. American-Gut-Project. mod1. 2015. http://microbio.me/americangut/img/mod1_main.pdf. Accessed Nov 2014.
37. Liu Z, DeSantis TZ, Andersen GL, Knight R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. 2008;36(18):e120. doi:10.1093/nar/gkn491.
38. MBQC. Microbiome Quality Control Project. http://www.mbqc.org. Accessed Nov 2014.
39. American-Gut-Project. Website. 2015. http://americangut.org. Accessed June 10 2015.
40. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12(1):87. doi:10.1186/s12915-014-0087-z.
41. Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiol Lett. 2010;307(1):80–6. doi:10.1111/j.1574-6968.2010.01965.x.
42. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, Bittinger K, et al. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiol. 2010;10:206. doi:10.1186/1471-2180-10-206.

43. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6. doi:10.1038/nmeth.f.303.
44. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaSci. 2012;1(1):7. doi:10.1186/2047-217X-1-7.
45. Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM, et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ. 2014;2:e545. doi:10.7717/peerj.545.
46. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012;6(3):610–8. doi:10.1038/ismej.2011.139.
47. Lozupone CA, Knight R. Global patterns in bacterial diversity. Proc Natl Acad Sci U S A. 2007;104(27):11436–40. doi:10.1073/pnas.0611525104.
48. Sullam KE, Essinger SD, Lozupone CA, O'Connor MP, Rosen GL, Knight R, et al. Environmental and ecological factors that shape the gut bacterial communities of fish: a meta-analysis. Mol Ecol. 2012;21(13):3363–78. doi:10.1111/j.1365-294X.2012.05552.x.
49. Lozupone C, Faust K, Raes J, Faith JJ, Frank DN, Zaneveld J, et al. Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. Genome Res. 2012;22(10):1974–84. doi:10.1101/gr.138198.112.
50. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, et al. Evolution of mammals and their gut microbes. Science. 2008;320(5883):1647–51. doi:10.1126/science.1155725.
51. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A. 2010;107(26):11971–5. doi:10.1073/pnas.1002601107.
52. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome Biol. 2011;12(5):R50. doi:10.1186/gb-2011-12-5-r50.
53. Vazquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. EMPeror: a tool for visualizing high-throughput microbial community data. GigaSci. 2013;2(1):16. doi:10.1186/2047-217X-2-16.
54. Kristal AR, Kolar AS, Fisher JL, Plascak JJ, Stumbo PJ, Weiss R, et al. Evaluation of web-based, self-administered, graphical food frequency questionnaire. J Acad Nutr Diet. 2014;114(4):613–21. doi:10.1016/j.jand.2013.11.017.
55. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29(5):415–20. doi:10.1038/nbt.1823.
56. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012;6(8):1621–4. doi:10.1038/ismej.2012.8.
57. Pérez FG, Brian E. IPython: a system for interactive scientific computing. Comput Sci Eng. 2007;9:21–9. doi:10.1109/MCSE.2007.53.
58. (CDC) CfDCaP. NHANES. http://www.cdc.gov/nchs/nhanes.htm. Accessed September 4 2015.
59. Nelson NJ. Nurses' health study: nurses helping science and themselves. J Natl Cancer Inst. 2000;92(8):597–9.
60. Moayyeri A, Hammond CJ, Hart DJ, Spector TD. The UK Adult Twin Registry (TwinsUK Resource). Twin Res Hum Genet. 2013;16(1):144–9.