Manuscript in press at *Developmental Psychobiology*

A primer on investigating the role of the microbiome in brain and cognitive development

AUTHORS: Caroline Kelsey[1*], Caitlin Dreisbach[2,3*], Jeanne Alhusen[3], Tobias Grossmann[1]

1. Department of Psychology, Gilmer Hall, University of Virginia, Charlottesville, Virginia 22903
2. Data Science Institute, 328 McCormick Road, University of Virginia, Charlottesville, Virginia 22903
3. School of Nursing, 225 Jeanette Lancaster Way, University of Virginia, Charlottesville, VA 22903

[*]Indicates shared first authorship

CORRESPONDING AUTHORS: Caroline Kelsey, cmk6jm@virginia.edu and Caitlin Dreisbach, cnd2y@virginia.edu

Abstract

Incorporating information regarding the gut microbiota into psychobiological research promises to shed new light on how individual differences in brain and cognitive development emerge. However, the investigation of the gut-brain axis in development is still in its infancy and poses several challenges, including data analysis. Considering that the gut microbiome is an eco-system containing millions of bacteria, one needs to utilize a breadth of methodologies and data analytic techniques. The present review serves two purposes. First, this review will inform developmental psychobiology researchers about the emerging study of the gut-brain axis in development and second, this review will propose methodologies and data analytic strategies for integrating microbiome data in developmental research.

Keywords: Brain development, Cognitive development, Gut-brain axis, Microbiome

A primer on investigating the role of the microbiome in brain and cognitive development

Within the human body, microorganisms, collectively called the microbiome, outnumber our own human body cells (Greenhalgh, Meyer, Aagaard, & Wilmes, 2016; Walker, 2013). In the gut microbiome alone, there are more than 1000 species that encode 200 times as many genes as the entire human genome (D'Argenio & Salvatore, 2015). This new knowledge about the human microbiome challenges existing views in multiple disciplines beyond biology, including concepts about the individual nature of the self (Rees, Bosch, & Douglas, 2018). Moreover, incorporating questions about the influence of the microbiome into research programs has the potential to significantly change the scientific landscape. The gut microbiome is thought to play a crucial role in everyday physiological functioning; and yet, relatively little is known about its specific contributions to health and disease (D'Argenio & Salvatore, 2015). With the emergence and improvement of next-generation genetic sequencing technology in recent years, the study of the microbiome has become feasible and more affordable. This increased access has kick-started large-scale scientific efforts to map the human microbiome, such as the *Human Microbiome Project,* funded by the National Institutes of Health (https://commonfund.nih.gov/hmp). These and future research efforts will help to uncover the role of the microbiome in human health and disease.

Given these advances in mapping the human microbiome, there is also growing interest in investigating how the microbiome affects developmental processes, specifically, brain and cognitive development (Borre et al., 2014; Carlson et al., 2018). Adding approaches that enable the study of gut microbiota to the developmental scientist's toolkit promises to shed new light on how individual differences psychological processes emerge. However, the investigation of how the gut-brain axis influences development is still in its infancy and poses several challenges,

including data analysis strategies. Considering that the gut microbiome is an eco-system containing millions of bacteria, applying traditional data analysis strategies is thus of limited use. The goals of the current review are: 1) to review the existing research investigating the role of the gut microbiome in brain and cognitive development, and 2) to provide an overview of research methodology and data analytic approaches utilized in the study of how the microbiome influences brain and cognitive development.

Investigating the gut-brain axis using animal models

A host of experimental evidence for the influence of the gut microbiome on brain development comes from animal studies comparing germ-free mice to conventional mice reared in a pathogenic-free environment (Heijtz et al., 2011). Germ-free mice do not have a microbiome, meaning that they are bacteria-, fungi-, and virus- free. These mice are created and maintained through specialized husbandry procedures, including birth through cesarean section and housing in sterile environments (Faith et al., 2010; Stilling, Dinan, & Cryan, 2014). Pathogen-free mice are reared through the most common husbandry practices, which includes screening to ensure that they are free from a specific list of disease-causing agents that would interfere with mouse health. In contrast to germ-free mice, pathogen-free mice are housed in bacteria-rich environments and maintain a diverse microbiome.

Germ-free mice exhibit marked physiological and behavioral differences from their conventional, pathogenic-free counterparts. More specifically, germ-free mice show a number of brain differences when compared to conventional mice, such as a significantly increased brain volume, decreased integrity of the blood brain barrier, increased serotonin synthesis, and increased myelination (Heijtz et al., 2011; Neufeld, Kang, Bienenstock, & Foster, 2011). Furthermore, a series of studies have shown that, compared to conventional mice, germ-free

mice display significantly reduced levels of anxiety-like behaviors resulting in increased risk-taking behaviors, such as increased open field exploration (De Palma et al., 2015; Hsiao et al., 2013).

The use of the germ-free animal model has illustrated the importance of the microbiome in psychobiological development. Specifically, the microbiome has been implicated in behavioral responses to early life stress. For example, germ-free mice but not pathogen-free control mice exhibit reduced species-typical anxiety-like behaviors in response to maternal separation (De Palma et al., 2015). This lack of anxiety seen in germ-free mice is thought to represent an aberrant response to the real-life threat of maternal separation. Moreover, the germ-free animal model has been used to study psychobiological effects of particular bacterial genera through administration during key phases during development. For example, Sudo and colleagues (2004) found that germ-free mice had increased release of corticosterone in response to an acute restraint stressor when compared to specific pathogen-free control mice. This study further showed that, in the germ-free mice, the HPA-axis response returned to normal levels after administration of *Bifidobacterium infantis*, whereas administration with *Escherichia coli* was associated with hyperactivity of the HPA axis. Critically, Sudo and colleagues also found that when during postnatal development administration occurred played an important role. In particular, this study showed that the administration of *Bifidobacterium infantis* was only able to return HPA axis activity to normal levels if administered to the mouse pups by six weeks (prior to sexual maturity) but not later during development (at eight weeks of age; the onset of sexual maturity). This research thus points to the possibility that there might be a sensitive period, prior to adolescence, when the microbiome may have the greatest impact on the development of the

stress system. Considering this kind of evidence it appears important to systematically investigate the role of the gut-brain axis in early psychobiological development.

Evidence for gut-brain axis in human models

Preliminary evidence for the existence of the gut-brain axis in humans comes from correlational studies, showing that delivery method and breastfeeding influences the colonization of the gut with bacteria. These differential patterns in colonization may in turn, have downstream effects on psychological development. For example, vaginal birth has been shown to expose infants to a larger diversity of bacteria in the birth canal than seen in infants delivered by caesarean section, who predominately receive bacteria from their mothers' skin (Dominguez-Bello et al., 2010). Feeding method also appears to be a contributor to the type and diversity of bacteria that inhabit the infant gut. For instance, breastfeeding provides infants with both bacteria and prebiotics, or nutrients that support bacterial growth, leading to a larger number of keystone (health-promoting) bacteria in breastfed when compared to formula-fed infants (Heikkilä & Saris, 2003).

It is important to emphasize that direct causal links between gut bacteria, brain, and cognitive development have not yet been established in humans. Nonetheless, there is correlational evidence from epidemiological studies suggesting that delivery method and breastfeeding, which as outlined above affect changes in the microbiome, also impact brain and cognitive development in infants. A recent meta-analysis found that infants delivered by cesarean section, when compared to those delivered vaginally, show a modest increase in the risk of developing Autism Spectrum Disorder (ASD) and Attention Deficit/Hyperactivity Disorder (ADHD) (Curran et al., 2015). Similarly, in animal models, mice that are born through cesarean section when compared to mice delivered vaginally, display increased repetitive behaviors and

atypical social behaviors, characteristic of these neurodevelopmental disorders (Borre et al., 2014). Moreover, this is in line with the human studies showing differences in bacteria composition between children with ASD when compared to neurotypical children. Specifically, children with ASD show distinct patterns of broad classes of bacteria composition with an increase in some toxin-producing bacteria, such as *Clostridia* (Finegold et al., 2002; Parracho, Bingham, Gibson, & McCartney, 2005). However, research concerning ASD is inconsistent because other bacteria genera, such as *Bacteriodetes,* are reported as increased prevalence in ASD children one study and not increased in another study (Son et al., 2015; Tomova et al., 2015).

Breastfeeding, in addition to affecting the colonization of the infant gut with microbes, has also been shown to impact brain and cognitive development in infants (see Krol & Grossmann, 2018, for a review). Specifically, there is evidence suggesting that the absence or short duration of exclusive breastfeeding might be associated with the development of ASD. For example, a recent meta-analysis reports that those children diagnosed with ASD were significantly less likely to have been breastfed when compared to typically developing children (Tseng et al., 2017).

These associations seen between delivery and feeding experiences among infants and developmental outcomes obviously do not provide direct evidence that the microbiome is influencing brain and cognitive development. Moreover, there could be a host of alternative explanations for these results, one being that both breastfeeding and delivery method may impact the development of the oxytocin system; and consequently, the neurohormone oxytocin has been linked to various outcomes in social behavior (Carter, 2014). Nonetheless, given that changes in the microbiome are associated with breastfeeding and vaginal birth, it is likely that the associated

microbiome changes are relevant to psychological development in infancy and beyond. Developmental work, which directly assesses the role of the microbiome in early brain and cognitive development, is needed to arrive at a more mechanistic understanding of how the gut-brain axis functions in early development.

Direct assessment of how the human gut microbiome impacts cognitive development in infancy

To date, there is very little work in humans that has directly assessed the relation between the gut microbiome and early brain and cognitive development. Only very recently, Carlson and colleagues (2018) took a first step by assessing gut microbiome composition at 1 year of age, and testing the association with cognitive and motor development (measured by the Mullen Scales of Early Learning), and with brain volume (measured using structural Magnetic Resonance Imaging [MRI]), at both 1 and 2 years of age. This study characterized the gut microbiome composition in two ways: 1) using the mean bacteria species diversity per individual (alpha diversity) and 2) using cluster analysis, which identified three major groupings across infant microbial composition based on differences in the abundance of three key bacteria genera *Faecalibacterium, Bacterioides,* and *Ruminococcacea* (grouped with unclassified genera). Carlson et al.'s (2018) study shows that infants' overall score on cognitive and motor development tasks, the Early Learning Composite Score, differed significantly between the three groups. Specifically, infants with a relatively high abundance of *Bacterioides* achieved the highest score, whereas infants with a relatively high abundance of *Faecalibacterium* showed the lowest score with respect to their cognitive and motor development. Moreover, findings revealed that when the analysis was focused on specific subscales of the Mullen Scales of Early Learning, the difference across groups was most prominent with respect to their receptive language scores.

In addition, Carlson et al. (2018) report structural brain differences, whereby infants in the group with a relatively high abundance of *Bacterioides* showed a larger right superior occipital gyrus at age one but smaller caudate nucleus when compared to infants in the other two groups. On the one hand this study suggests that there are some specific structural brain differences; however, it should be noted that the majority of structural brain measures such as intracranial volume, total white or gray matter, total cerebrospinal fluid, or lateral ventricle volume did not reveal any differences between infants in the different bacterial composition groups. Moreover, from these data it is unclear how these differences in brain structure are linked to brain and cognitive function.

Carlson et al.'s (2018) study also showed that gut microbial diversity at the age of one year predicted cognitive development at the age of two years. The longitudinal association found in this study was such that increased microbial diversity was associated with lower cognitive performance as measured in the Early Learning Composite Score and lower scores on the specific subscales of visual reception and expressive language. This finding is surprising considering that higher microbial diversity in adults has typically be shown to be predictive of positive health outcomes (Abrahamsson et al., 2014; Kostic et al., 2015). Carlson and colleagues (2018) suggest that microbial diversity may affect cognitive functions differently in infancy than later in development. This points to the importance of developmental research which maps associations between microbial characteristics and brain and cognitive development across the human lifespan. In the following, we would like to briefly outline how researchers may use new methodological and statistical approaches to explore the influence of the microbiome on brain and cognitive development.

Generating microbiome data

After reviewing existing research on the role that the microbiome may play in brain and cognitive development, this section of the review will discuss sampling, data collection, and genomic sequencing of microbiome data for use in psychobiological research. Here, we outline how the microbiome is collected, sequenced, and processed in a data stream to address questions about composition and function of microbes.

Collecting microbiome data

There are several methods for collecting microbiome samples from study participants. The two major collection methods for assessing the microbiome of the distal gastrointestinal tract, a proxy for understanding the community structure of the gut, are rectal swabbing and collection of a stool/fecal samples. A rectal sample includes the participant utilizing a small q-tip swab after a recent bowel movement to collect the microbiota that are more focused at the rectum. Participants should insert the swab approximately 1-2 centimeters beyond the rectum for optimum collection (Bassis et al, 2017). For fecal samples, sterilized containers with small scoops should be used for cleanliness. Tools, such as toilet inserts to catch samples, can help participants to feel comfortable with collection. Moreover, infant researchers may ask parents to bring in a diaper (note, researchers may choose to provide a sterile plastic insert to parents to put into the diapers to optimize collection) and transfer the sample from the diaper to a storage container in the lab. Both fecal samples and swabs should be labeled appropriately with date and time of collection and study identification number. In the interest of gathering more robust data for fecal samples, charts such as the Bristol Stool chart (BSC) can be used to allow participants to classify their sample into 7 distinct types illustrated by representative pictures of various textures and consistencies. Classification of the types using the BSC, including amount and consistency, is important because early research has identified stool consistency being associated

with gut microbiota richness and composition (Vandeputte et al, 2016). Complete sampling kits, with all the necessary materials, are available for purchase from a wide range of medical and research distributors. Both methods, including swabs or fecal samples, should be considered for research and the specific method should be selected on the basis of participant population (e.g, for infants, fecal samples may be easier to collect and may have higher compliance from families as compared to swabs), resources available to data collection team, and consultation with the data sequencers.

After collection, storage options can vary depending on study question and availability. Options include liquid buffers to help stabilize samples and long-term freezing (typically in temperatures ranging from -80 to -4 degrees celsius). Most importantly, consideration should be used when freezing and thawing samples as this could have an effect on bacterial growth and/or DNA damage (Hugerth & Andersson, 2017). With the expansion of large-scale cohort studies and biospecimen banking, long-term freezing is common to ensure sample availability for future research. A recent comparison study found that these storage methods, including freezing temperature and stabilizing agents, can be used interchangeably with similar diversity metrics (Bassis et al, 2017). However, for consistency, studies should utilize the same technique for all samples.

Sequencing microbiome and processing samples

A gut microbiome sample, either a stool sample or a rectal swab, can be sequenced in two major protocols that result in different types of output data. The first and most common method to studying taxonomy and phylogeny of the microbiome, due to cost and efficiency, is 16s RNA sequencing (Janda & Abbott, 2007). In this method, short strands of DNA called primers, which are designed to target specific variable regions of the 16s ribosomal RNA gene, are used to

classify taxonomic units of microbiota within a sample (Illumina, 2018). The 16s RNA gene is highly conserved, or passed through generations, and acts as a microbe-specific genetic signature. The protocol begins with purified DNA from the fecal samples (Illumina, 2018). Primers are tagged with indexing barcodes and samples are pooled into a single library, or a collection of the primer nucleic acid targets, for sequencing (Illumina, 2018). Taxonomic profiling on the Illumina MiSeq system, a type of popular sequencing equipment and the industry standard platform, is typically cycled to generate paired 250-base pair reads in each protocol (Illumina, 2018). Other platforms include the Roche 454 GS FLX and the Ion Torrent PGM which both include different library preparations, procedures for barcodes and adapters as well as amplification (Allali et al, 2017). A recent study found that microbiome community profiles were comparable across platforms but that the relative abundance of specific microbiota varied depending on the sequencing platform, library preparation procedures, and analytic approach (Allali et al, 2017). However, the longer read lengths provided by the Illumina platform offer a high-quality analysis of the rRNA gene to ensure the most accurate classification available. Additionally, because chimeric sequences, sequences originating from two transcripts, and mismatched primers are considered to be contaminant within the analysis, they are filtered out using the standard Human Microbiome Project search and clustering program, USEARCH (Shaw et al, 2017). Raw sequence data in the form of fastq files are the output product of this processing pipeline, which can then be entered for further analysis.

Once the sequencing is completed, 16S rRNA gene sequence data in the form of fastq will be input into the Quantitative Insights Into Microbial Ecology (QIIME) 1.8.0 software package (Caporaso et al., 2010). QIIME is a big data, open-source software built for microbiome analysis from raw fasq sequencing data on Illumina platforms. QIIME groups the genomic

sequence into operational taxonomic units (OTUs). OTUs are groups of similar 16s sequences that become proxies for a species of a microbe (Caporaso et al., 2010). OTUs group the genomic sequences to identify which taxonomic group it belongs to. Genome reference databases such as Green Genes should be used to provide standardization of OTU assignment with publically available and published taxonomies (DeSantis et al., 2006). In addition to QIIME, several other bioinformatics packages are available including mothur and MetaGenome Rapid Annotation using Subsystem Technology (MG-RAST). Both MG-RAST and mothur offer a comparable data processing pipeline for 16s microbial comparisons. Another recent bioinformatic comparison study found that researchers arrived at largely comparable results regardless of which of the three existing pipelines were used (Plummer et al, 2015). It is worth noting that the main difference revealed by the pipeline comparison carried out in this study was the significantly increased computational speed for QIIME compared to mothur and MG-RAST, taking approximately 1 hour, 10 hours, and 2 days respectively for processing a sample of 35 specimens (Plummer et al, 2015). Similar to the sequencing methods, for the purposes of this primer, the focus will be on processing with QIIME due to its widespread use.

From QIIME, data can be read into R to be manipulated using a package called ‘phyleoseq’. A full microbial analysis workflow is provided open access through Bioconductor (Callahan et al, 2016). Bioconductor in R is the most common package for bioinformatic analysis with inclusion of packages such as ‘dada2’, ‘phyloseq’, ‘DESeq2’, ‘ggplot2’ and ‘vegan’ to normalize, visualize, test, and compare microbial data samples. At this point, questions about community analysis, including which microbes are present and how do they compare in abundance to others, can be elucidated.

SixteenS rRNA sequencing is not the only method for extracting valuable insight from microbiome data. Metagenomic sequencing, also known as shotgun metagenomics, uses next-generation sequencing technology to understand functional gene composition rather than just viewing the 16s RNA conserved gene (Thomas et al., 2012). The sequence pathway begins with extracting DNA from the fecal samples similar to 16s. By sequencing the community DNA and comparing it to reference gene catalogs, metagenomics offers improved precision and allows for genetic observation of variants such as single nucleotide polymorphisms (Morgan & Huttenhower, 2012). Function can then be assessed and assigned using other publicly-available databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Kanehisa et al, 2008). The ability to sequence the entire genome comes at a significant expense, which may double the costs of 16S sequencing. With the additional cost and expertise comes the ability to generate more data to elucidate information on not only community structure and prevalence of microbiota but also on the function of the microbes present (Sharpton, 2014). Metagenomic profiling can offer answers to questions of not only 'what microbiota are present?' but also 'what can the community do?'. For the purposes of this primer, we have outlined methodologies to approach microbiome research through the cost-effective and widespread use of 16S sequencing.

Outlining the process of microbiome collection, sequencing, and the bioinformatics pathway for the raw data analysis as done here in brief is important in preparing microbiome data to be used in a research study. Once the taxonomic data, including which species of bacteria are present within the samples, has been identified, further data analysis can be pursued to answer questions of developmental and clinical relevance.

Analysis using data science methods to analyze the human microbiome

In the following section, we will discuss suitable research data science methods including machine learning, data mining, and deep learning that can be applied to explore heterogeneous microbiome data sets. The introduction of such analysis techniques to study the role of the microbiome in human development has the potential to capture the complexity, allow for relatively unbiased statistical inferences, generate testable predictions, and ultimately, may result in clinical applications.

Traditional approaches and correlational statistics

The primary output from microbiome sequencing processing is counts of genomic reads that are taxonomically assigned to specific microbiota through reference genome sets. This count data is often normalized or processed additionally to remove systematic variance in the data. The count data is then considered for each microbiota in terms of its abundance within a certain feature, in the case of 16S sequencing it is in terms of the taxonomic classification. The counts for each classification are typically reported as proportions, which reflect fractions of specific species rather than absolute abundances (Lovell et al, 2015). Unfortunately, these proportions are difficult to predict and interpret in relation to the absolute abundance and confounding factors within an environment (Gloor et al, 2017; Friedman et al, 2012). This is a particularly challenging problem for the investigation of maternal microbiome samples because of the known community and diversity changes in pregnancy related to hormonal fluctuations, which occur independently from microbial dysbiosis or pathology.

More than just understanding the role of read counts as data output and microbiota proportions, there are several other traditional methods that should be reconsidered as the science moves toward understanding more than just community composition. One particular issue is the reliance on proportions as the major data analytic method when analyzing microbiome data --

this results in the assumption that abundance, or relative low abundance, is the key driver for functional differences. In addition, many standard statistical approaches assume an independence between microbiota, which does not exist (see Xia & Sun, 2017).

Data mining

As an overarching method comprised of several approaches in data science, data mining has become a popular computing and statistical process to discover patterns in large, mixed source datasets (Hendler, 2014). Data mining helps to explore information in large datasets where patterns emerge, which cannot be adequately captured by traditional linear regression models because of the highly non-linear complexity present within the data (Zhang & Zaki, 2006). We will now briefly describe some of the existing data mining techniques and in turn discuss the use of simulated data, time analysis, and clustering.

Simulated data. Simulated data is a particularly powerful data mining technique when applied prior to large-scale, cost-intensive experimental studies as it can help formalize conceptual models that can then be tested empirically. Step-by-step workflows are available to assist researchers in creating simulated data for specific purposes (Hallgren, 2013). Available coding software, such as R and python, can be used to simulate data according to a formalized model before investing in large-scale experimental studies. Implementation of simulated data in the context of a relevant research questions can help with answering specific questions in model building, estimation of beta coefficients, and better tuning of parameters of machine learning algorithms such as gamma values or learning rates (Schloss, 2008; Chen, 2012). For example, simulations of what parameters of the microbiome in an animal model (mouse) impact a given outcome such as social behaviors seen in ASD, could be used to help design experimental studies with humans. The largest limitation of this method is that the use of simulated data critically

relies on prior information, which is needed to build a simulation model. Considering that, apart from animal models, very little prior information is available in human translational research limiting the utility of this technique until more literature is published in this area.

Time Analysis. Previous research in humans and model organisms have predominately collected and analyzed microbial data cross-sectionally (Caporaso et al, 2010; Parks et al., 2014; Fukuyama et al., 2017). However, in order to arrive at a mechanistic understanding of microbial influence on outcome variables, it is of critical importance to understand how microbial patterns change in development and in response to certain events or interventions (Faust, et al., 2015), making longitudinal research designs the method of choice for fostering such an understanding (Morgan & Huttenhower, 2012). For example, an important unanswered question is how does the human microbiome change due to feeding and mode of delivery, and whether and how do these changes in the microbiome predict brain and cognitive development in children. New computational tools (software packages) have emerged to help visualize microbial time-series data, which can also be applied to longitudinal data. One such application is Temporal Insights into Microbial Ecology (TIME), a web-based software for longitudinal microbiome data analysis, offering a wide range of input data types and capability to identify potential taxonomic markers through analysis and visualization (Baksi, Kuntal, & Mande, 2018). Another web-based software tool is called BURRITO (https://github.com/borenstein-lab/burrito), which also offers time-series based visualization and analysis, coupled with taxonomic and functional profiling to elucidate the contribution of the microbiota to a biological function such as neurotransmitter transport, GABA-A receptor agonists/antagonists or systemic inflammatory responses (McNally, Eng, Noecker, Gagne-Maynard & Borenstein, 2018; Kanehisa et al, 2008). Applying these

methods to longitudinal data promises to innovatively capture and visualize the link between microbial and developmental changes.

Clustering. Clustering is a common technique to describe the proximity between subjects or samples (Cameron, 2012). Interestingly, centroid-based clustering algorithms, such as k-means using euclidean distance metrics, which group samples based on distance to the computed centroid, have shown to perform well on clinical and simulated microbial datasets (Cameron, 2012). Beyond distanced-based clustering algorithms, other data science methods are also able to account for complex biological data. For example, hierarchical clustering, which is a set of descriptive techniques used for grouping by similarity, has been particularly useful when applied to metagenomic data (McMurdie, 2016). Clustering algorithms may help researchers to profile similarity across microbiome samples and identify boundaries based on function, and thus help uncover clusters of microbes that best characterize any given developmental outcome.

Machine learning

For developmental, psychological, and clinical researchers, machine learning algorithms have been proposed to be effective in addressing questions concerning classification and prediction of biological and behavioral variables (Yarkoni & Westfall, 2017). Large datasets can be used to train models to answer classification problems or provide probabilities of an outcome. This section outlines some techniques for machine learning and areas for exploration in this new domain that focuses on prediction rather than description.

Reduction of Features. Feature reduction methods are extremely important in high-dimensional datasets. One widely used technique in microbial analysis to achieve a reduction in relevant features is Principal Component Analysis (PCA), which uses orthogonal transformation to reduce features and create a smaller set of components (Meng, Zeleznik, Thallinger, Kuster,

Gholami et al, 2016). PCA relies on linear methodologies which may not best describe the underlying truth. However, feature reduction can also be harnessed through neural networks using autoencoders, which provide a neural network structure for unsupervised learning of encoded nodes (Tan, Hammon, Hogan & Greene, 2015). The encoded nodes represent a component of the original data. Tools such as the Analysis Using Denoising Autoencoders of Gene Expression (ADAGE), allow researchers to train an autoencoder on a dataset to derive nodes thereby reduce features to highlight highly-active genes (Tan et al., 2015). Autoencoding is particularly relevant in datasets with a large number of participants and a wide range of behavioral and brain measurements. From a data science perspective, high dimensional problems arising from such data sets are thought to be best addressed using autoenconding methods.

Classification. In terms of classification, random forest models tend to be popular because they have been proven to be high in their prediction accuracy (Touw et al, 2012). Identifying conditional relations, such as the presence or absence of a certain microbes accounts for the presence or absence of certain outcomes, are prime purpose for using random forest modeling. Random forest models are supervised learning algorithms that generate decision trees allowing for classification on the basis of deterministic rather than random relations between a certain microbe and an outcome variable (Touw et al, 2012). Random forest models can be thus be used to better characterize which microbial species or Operational Taxonomic Units (OTUs) are most important for a particular classification task. This could include a classification problems in cognitive and brain development of clinical relevance such as the diagnosis of ASD. This can be implemented in the data processing stream through packages available in R such as ‘randomForest’ to be used in conjunction with ‘phyloseq’, which allows for the general analysis and visualization of microbial communities.

Deep Learning. Due to the known complexity of microbiome and developmental data, other data science methods are needed to further pinpoint health or disease-relevant outcomes. Deep learning is a collection of machine learning methods that are designed to carry out non-linear algorithms in an artificial neural network. Similar to autoencoders, alluded to above, deep learning methods make use of multiple connected layers in which output from the previous layer is employed to denoise and reconstruct the original data, while unveiling nodes, or representations of the data. Importantly, deep learning is generally considered as one of the most rigorous data science methods also due to its unbiased (and non-linear) nature of capturing patterns in complex data sets. Deep learning can be applied to both OTU or metagenomic data and is typically implemented through python-based software packages such as Keras and Tensorflow, but it can also be realized in R.

Taken together, this brief summary of some of the available data science practices is intended to provide a general guide for what analysis strategies might be useful in studying microbiome effects on brain and cognitive development. The review of the data science practices presented here is by no means exhaustive. Moreover, to date, there is no standardized procedure or platform available that integrates across these data science practices, and it is important to emphasize that the exact data science-based approach to be employed has to be specifically tailored to the particular research questions being addressed.

Conclusion

The growing body of research reviewed here provides first insights into how the gut microbiome influences early brain and cognitive development. We have seen that incorporating information regarding the gut microbiome into psychobiological research promises to further our understanding of how individual differences in brain and cognitive development emerge. While

the investigation of the gut-brain axis in development is still in its infancy, we have argued that an approach using data science methods has the potential to help us make progress in describing and predicting how the gut microbiome, as an eco-system containing millions of bacteria, influences brain and cognitive development. Applying data science methods including machine learning, data mining, and deep learning to mixed methods microbiome data sets will advance the study of the gut-brain axis in early human development. By summarizing some basic principles in microbiome analysis, data analytics and its application to brain and cognitive development, this review is meant to offer a brief introduction into this new frontier in developmental psychobiology. This is done in the hope that this review will help inspire the bold research efforts needed in the coming years to realize advances in our understanding of the microbiome's role in development.

References

Abrahamsson, T., Jakobsson, H., Andersson, A. F., Björkstén, B., Engstrand, L., & Jenmalm, M. (2014). Low gut microbiota diversity in early infancy precedes asthma at school age. *Clinical & Experimental Allergy, 44*(6), 842-850. doi:10.1111/cea.12253

Allali, I., Arnold, J. W., Roach, J., Cadenas, M. B., Butz, N., Hassan, H. M., … Azcarate-Peril, M. A. (2017). A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiology*, *17*, 194. http://doi.org/10.1186/s12866-017-1101-8

Bercik, P., Denou, E., Collins, J., Jackson, W., Lu, J., Jury, J., . . . Collins, S. M. (2011). The intestinal microbiota affect central levels of brain-derived neurotropic factor and behavior in mice. *Gastroenterology, 141*(2), 599-609, 609.e591-593. doi:10.1053/j.gastro.2011.04.052

Borre, Y. E., O'Keeffe, G. W., Clarke, G., Stanton, C., Dinan, T. G., & Cryan, J. F. (2014). Microbiota and neurodevelopmental windows: implications for brain disorders. *Trends in molecular medicine, 20*(9), 509-518. doi:10.1016/j.molmed.2014.05.002

Baksi, K., Kuntal, B. & Mande, S. (2018). A Web Application for Obtaining Insights into Microbial Ecology Using Longitudinal Microbiome Data. *Frontiers in Microbiology, 9*, 1-36. doi: 10.3389/fmicb.2018.00036

Bassis, C., Moore, N., Lolans, K., Seekatz, A., Weinstein, R., Young, V., & Hayden, M. (2017). Comparison of stool versus rectal swab samples and storage conditions on bacterial community profiles. *BMC Microbiology, 17*(78). doi: 10.1186/s12866-017-0983-9.

Callahan, B., Sankaran, K., Fukuyama, J. *et al.* (2016). Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research*, 5, 1492. doi: 10.12688/f1000research.8986.1

Cameron, E., "Optimal clustering techniques for metagenomic sequencing data" (2012). Electronic Thesis and Dissertation Repository. 707. https://ir.lib.uwo.ca/etd/707

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Carlson, A. L., Xia, K., Azcarate-Peril, M. A., Goldman, B. D., Ahn, M., Styner, M. A., . . . Knickmeyer, R. C. (2018). Infant gut microbiome associated with cognitive development. *Biological psychiatry, 83*(2), 148-159. doi:10.1016/j.biopsych.2017.06.021

Chen, J. (2012). Statistic methods for human microbiome data analysis. *UPenn Dissertation Repository*, 1-1-2012.

Curran, E. A., O'neill, S. M., Cryan, J. F., Kenny, L. C., Dinan, T. G., Khashan, A. S., & Kearney, P. M. (2015). Research review: birth by caesarean section and development of autism spectrum disorder and attention-deficit/hyperactivity disorder: a systematic review and meta-analysis. *Journal of Child Psychology and Psychiatry, 56*(5), 500-508. doi:10.1111/jcpp.12351

D'Argenio, V., & Salvatore, F. (2015). The role of the gut microbiome in the healthy adult status. *Clinica Chimica Acta, 451*, 97-102.

De Palma, G., Blennerhassett, P., Lu, J., Deng, Y., Park, A., Green, W., . . . Sanz, Y. (2015). Microbiota and host determinants of behavioural phenotype in maternally separated mice. *Nature communications, 6*, 7735. doi:10.1038/ncomms8735

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., … Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. http://doi.org/10.1128/AEM.03006-05

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proceedings of the National Academy of Sciences, 107(26), 11971-11975. doi:10.1073/pnas.1002601107

Faith, J. J., Rey, F. E., O'donnell, D., Karlsson, M., McNulty, N. P., Kallstrom, G., ... & Gordon, J. I. (2010). Creating and characterizing communities of human gut microbes in gnotobiotic mice. *The ISME journal*, *4*(9), 1094-1098. doi: 10.1038/ismej.2010.110

Faust, K., Lahti, L., Gonze, D., de Vos, W. & Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current Opinion in Microbiology*, *25*, 56-66.

Finegold, S. M., Molitoris, D., Song, Y., Liu, C., Vaisanen, M. L., Bolte, E., . . . Kaul, A. (2002). Gastrointestinal microflora studies in late-onset autism. *Clin Infect Dis, 35*(Suppl 1), S6-s16. doi:10.1086/341914

Friedman, J., & Alm, E. J. (2012). Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology*, *8*(9), e1002687. http://doi.org/10.1371/journal.pcbi.1002687

Fukuyama, J., Rumker, L., Sankaran, K., Jeganathan, P., Dethlefsen, L., Relman, D. A., & Holmes, S. P. (2017). Multidomain analyses of a longitudinal human microbiome

intestinal cleanout perturbation experiment. *PLoS Computational Biology*, *13*(8), e1005706.

Gareau, M. G., Wine, E., Rodrigues, D. M., Cho, J. H., Whary, M. T., Philpott, D. J., . . . Sherman, P. M. (2011). Bacterial infection causes stress-induced memory dysfunction in mice. *Gut, 60*(3), 307-317. doi:10.1136/gut.2009.202515

Goehler, L. E., Park, S. M., Opitz, N., Lyte, M., & Gaykema, R. P. (2008). Campylobacter jejuni infection increases anxiety-like behavior in the holeboard: possible anatomical substrates for viscerosensory modulation of exploratory behavior. *Brain Behav Immun, 22*(3), 354-366. doi:10.1016/j.bbi.2007.08.009

Greenhalgh, K., Meyer, K. M., Aagaard, K. M., & Wilmes, P. (2016). The human gut microbiome in health: establishment and resilience of microbiota over a lifetime. *Environmental microbiology, 18*(7), 2103-2116. doi:10.1111/1462-2920.13318

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, *8*, 2224. http://doi.org/10.3389/fmicb.2017.02224

Hallgren, K. (2013). Conducting Simulation Studies in the R Programming Environment. *Tutorials in quantitative methods for psychology*, 9(2), 43-60.

Heijtz, R. D., Wang, S., Anuar, F., Qian, Y., Björkholm, B., Samuelsson, A., . . . Pettersson, S. (2011). Normal gut microbiota modulates brain development and behavior. *Proceedings of the National Academy of Sciences, 108*(7), 3047-3052. doi:10.1073/pnas.1010529108

Heikkilä, M., & Saris, P. (2003). Inhibition of Staphylococcus aureus by the commensal bacteria of human milk. *Journal of applied microbiology, 95*(3), 471-478. doi:10.1046/j.1365-2672.2003.02002.x

Hendler, J. (2014). Data Integration for Heterogenous Datasets. *Big Data*, 2(4), 205-215. doi:10.1089/big.2014.0068.

Hirschfeld, R. M. A. (2001). The Comorbidity of Major Depression and Anxiety Disorders: Recognition and Management in Primary Care. *Primary Care Companion to The Journal of Clinical Psychiatry, 3*(6), 244-254.

Hsiao, E. Y., McBride, S. W., Hsien, S., Sharon, G., Hyde, E. R., McCue, T., . . . Petrosino, J. F. (2013). Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell, 155*(7), 1451-1463. doi:10.1016/j.cell.2013.11.024

The Human Microbiome Project. (2012). Retrieved from https://hmpdacc.org/

Hugerth, L., & Andersson, A.(2017). Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. *Frontiers in Microbiology, 8.* 1561. doi:10.3389/fmicb.2017.01561

Illumina. (2018). Demonstrated workflow for 16S rRNA sequencing. Retrieved from https://www.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/16s-rrna-sequencing.html

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls . *Journal of Clinical Microbiology*, *45*(9), 2761–2764. http://doi.org/10.1128/JCM.01228-07

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., … Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, *36*(Database issue), D480–D484. http://doi.org/10.1093/nar/gkm882

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., . . . Mattila, I. (2015). The dynamics of the human infant gut microbiome in development and

in progression toward type 1 diabetes. *Cell host & microbe, 17*(2), 260-273. doi:10.1016/j.chom.2015.01.001

Krol, K.M., & Grossmann, T. (2018). Psychological effects of breastfeeding on children and mothers. *Bundesgesundheitsblatt, 61(8),* 977-985.

Krol, K. M., Monakhov, M., San Lai, P., Ebstein, R. P., & Grossmann, T. (2015). Genetic variation in CD38 and breastfeeding experience interact to impact infants' attention to social eye cues. *Proceedings of the National Academy of Sciences, 112*(39), E5434-E5442. doi:10.1073/pnas.1506352112

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., & Bähler, J. (2015). Proportionality: A Valid Alternative to Correlation for Relative Data. *PLoS Computational Biology*, *11*(3), e1004075. http://doi.org/10.1371/journal.pcbi.1004075

Lyte, M., Varcoe, J. J., & Bailey, M. T. (1998). Anxiogenic effect of subclinical bacterial infection in mice in the absence of overt immune activation. *Physiol Behav, 65*(1), 63-68.

McMurdie, P. Lecture 7: Machine Learning of the microbiome [PDF document]. Retrieved from University of Washington, Department of Biostatistics website: *https://www.biostat.washington.edu/sites/default/files/modules//2016-SISMID-14-07.pdf*

McNally, C., Eng, A., Noecker, C., Gagne-Maynard, W. & Borenstein, E. (2018). BURRITO: An interactive multi-omic tool for visualizing taxa-function relationships in microbiome data. *Frontiers in Microbiology.* doi: 10.3389/fmicb.2018.00365.

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., & Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, *17*(4), 628–641. http://doi.org/10.1093/bib/bbv108

Morgan, X. & Huttenhower, C. (2012). Chapter 12: Human Microbiome Analysis. *PLOS Computational Biology. https://doi.org/10.1371/journal.pcbi.1002808*

Muthukrishnan, R. & R. Rohini. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Coimbatore, 18-20. doi: 10.1109/ICACA.2016.7887916

Neufeld, K., Kang, N., Bienenstock, J., & Foster, J. (2011). Reduced anxiety-like behavior and central neurochemical change in germ-free mice. *Neurogastroenterology & Motility, 23*(3), 255. doi:10.1111/j.1365-2982.2010.01620.x

Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30, 3123–3124. doi: 10.1093/bioinformatics/btu494

Parracho, H. M., Bingham, M. O., Gibson, G. R., & McCartney, A. L. (2005). Differences between the gut microflora of children with autistic spectrum disorders and that of healthy children. *J Med Microbiol, 54*(Pt 10), 987-991. doi:10.1099/jmm.0.46101-0

Plummer, E., Twin, J., Bulach, D., Garland, S., & Tabrizi, S. (2015). A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. *Journal of Proteomics & Bioinformatics*, 8. 10.4172/jpb.1000381.

Rees, T., Bosch, T., & Douglas, A. E. (2018). How the microbiome challenges our concept of self. *PLoS biology, 16*(2), e2005358.

Schloss, P. (2008). Evaluating different approaches that test whether microbial communities have the same structure. ISME J., 2(3), 265–275.

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, *5*, 209. http://doi.org/10.3389/fpls.2014.00209.

Shaw, A., Black, N., Rushd, A., Sim, K., Randell, P., Kroll, S., & Epstein, J. (2017). Assessing the Colonic Microbiota in Children: Effects of Sample Site and Bowel Preparation. *J Pediatr Gastroenterol Nutr*, *64*(2), 230-237. doi: 10.1097/MPG.0000000000001233.

Son, J. S., Zheng, L. J., Rowehl, L. M., Tian, X., Zhang, Y., Zhu, W., . . . Li, E. (2015). Comparison of Fecal Microbiota in Children with Autism Spectrum Disorders and Neurotypical Siblings in the Simons Simplex Collection. *PLoS One, 10*(10), e0137725. doi:10.1371/journal.pone.0137725

Stilling, R. M., Dinan, T. G., & Cryan, J. F. (2014). Microbial genes, brain & behaviour–epigenetic regulation of the gut–brain axis. *Genes, Brain and Behavior*, *13*(1), 69-86. doi: 10.1111/gbb.12109

Sudo, N., Chida, Y., Aiba, Y., Sonoda, J., Oyama, N., Yu, X. N., ... & Koga, Y. (2004). Postnatal microbial colonization programs the hypothalamic–pituitary–adrenal system for stress response in mice. *The Journal of physiology*, *558*(1), 263-275. doi: 10.1113/jphysiol.2004.063388

Tan, J., Hammond, J., Hogan, D. & Greene, C. (2016). ADAGE-Based Integration of Publicly Available *Pseudomonas aeruginosa*Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions. *mSystems*, *1*(1), e00025-15; DOI: 10.1128/mSystems.00025-15

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, *2*, 3. http://doi.org/10.1186/2042-5783-2-3

Tomova, A., Husarova, V., Lakatosova, S., Bakos, J., Vlkova, B., Babinska, K., & Ostatnikova, D. (2015). Gastrointestinal microbiota in children with autism in Slovakia. *Physiol Behav, 138*, 179-187. doi:10.1016/j.physbeh.2014.10.033

Touw, W., Bayjanov, J., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. (2013). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315–326, https://doi.org/10.1093/bib/bbs034

Tseng, P.-T., Chen, Y.-W., Stubbs, B., Carvalho, A. F., Whiteley, P., Tang, C.-H., . . . Chu, C.-S. (2017). Maternal breastfeeding and autism spectrum disorder in children: A systematic review and meta-analysis. *Nutritional neuroscience*, 1-9. doi:10.1080/1028415X.2017.1388598

Vandeputte, D., Falony, G., Vieira-Silva, S., et al. (2016). Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut,* 65, 57-62.

Walker, W. A. (2013). Initial intestinal colonization in the human infant and immune homeostasis. *Annals of Nutrition and Metabolism, 63*(Suppl. 2), 8-15. doi:10.1159/000354907

Yarkoni, T. & Wastfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect Psychol Sci, 12*(6), 1100-1122.

Xia, Y., & Sun, J. (2017). Hypothesis testing and statistical analysis of the microbiome. *Genes & Diseases*, *4*, 138-148.

Yazdani, M., Taylor, B., Debelius, J., Weizhong, L., Knight, R., & Smarr, L. (2016). Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease. 1272-1280. 10.1109/BigData.2016.7840731.

Zhang, S., & Zaki, M. (2006). Mining Multiple Data Sources: Local Pattern Analysis. *Data Mining and Knowledge Discovery,* 12(2-3), 121-125.