Summer 8-12-2016

Bayesian Networks to Assess the Newborn Stool Microbiome

William E. Bennett Jr.
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science

Department of Computer Science and Engineering

Thesis Examination Committee:
Michael R. Brent, Chair
Phillip I. Tarr
Roman Garnett

Bayesian Networks to Assess the Human Newborn Stool Metatranscriptome

by

William E. Bennett, Jr.

A thesis presented to the School of Engineering
of Washington University in St. Louis in partial fulfillment of the
requirements for the degree of
Master of Science

August 2016

Saint Louis, Missouri

Contents

List of Figures

Acknowledgments

Thank you to the Tarr laboratory and the Genome Center at Washington University School of Medicine for assistance in sample preparation, sequencing, and analysis. Thank you to Nurmohammad Shaik, Ph.D. and Vincent Magrini, Ph.D. This work was funded in part by NIH grants F32 DK088516 (William Bennett), UH2 AI083266 (Phillip Tarr and Barbara Warner).

William E. Bennett, Jr.

Washington University in St. Louis
August 2016

Dedicated to Ellie, Liam, Oliver, Cecilia, and Silas.

ABSTRACT OF THE THESIS

Bayesian Networks to Assess the Human Newborn Stool Metatranscriptome

by

William E. Bennett, Jr.

Master of Science in Computer Science

Washington University in St. Louis, 2016

Thesis Committee: Professor Michael R. Brent,

Professor Phillip I. Tarr,

Professor Roman Garnett

In human stool, a large population of bacterial genes and transcripts from hundreds of genera coexist with host genes and transcripts. Assessments of the metagenome and transcriptome are particularly challenging, since there is a great deal of sequence overlap among related species and related genes. We sequenced the total RNA content from stool samples in a neonate using previously-described methods. We then performed stepwise alignment of different populations of RNA sequence reads to different indices, including ribosomal databases, the human genome, and all sequenced bacterial genomes. Each pool of RNA at each alignment step was subjected to compression to assess sequence complexity in bits per symbol. In order to account for the high degree of overlap among species, a Bayesian network tool (RNABayes) was constructed using a node based on 16S sequencing, and a large number of nodes based on alignment scores to bacterial genes. The following algorithm was then employed: (1) fit 16S census from a sample onto a Dirichlet distribution using maximum likelihood estimation to get the conjugate prior, (2) estimate probabilities of each bacterial genus for each bacterial mRNA alignment using BLAST alignment scores, (3) fit each of these probabilities to a Dirichlet distribution using maximum likelihood estimation, (4) perform inference iteratively to update the conjugate prior, with the result being the posterior probability distribution of metabolically active stool bacteria. This algorithm was then applied to three datasets: (1) a simulated data set with normally distributed mRNAs, (2) a simulated data set with skewed mRNAs for a single bacterial population, and (3) the RNASeq dataset from our newborn stool sample. Results indicate that a Bayesian network built in this fashion reliably adjusts the prior bacterial population distribution to more accurately reflect the transcriptionally active bacterial population. Application of this method to real world samples appears to show even more marked skew, indicating transcripts are not uniformly distributed by population.

Preface

The study of complex nucleic acid mixtures has made enormous strides in the past decade, but still suffers from the Big Data Curse: more information does not necessarily yield greater understanding of a complex process, and conversely, can often reveal further levels of complexity that only expand that which we don't know. I have found in most endeavors that believing something to be simple and easily solved is highly fallacious, and in almost every case, once we know more about a problem, we all too often discover that it's way more complicated that we thought.

In the 1971 the United States government passed the National Cancer Act, which was billed politically as a cure for cancer by the Nixon administration. The lay press, and many scientists and physicians, predicted that with the right amount of investment of resources, we would soon see an end to cancer in as little as a decade. Of course, this was strikingly optimistic, as we now see that cancer is an extremely varied clinical endpoint to hundreds of distinct physiologic processes. Our understanding of oncology remains incremental, but has benefited massively from genomic science.

Again in the late 1990's, the Human Genome Project was billed as a panacea for a whole host of human diseases. Once we knew the code for our genetic failures, we could cure a large number of ailments, or predict them. While the Human Genome Project was certainly a massive scientific achievement that has had ripples throughout science, the result has been, somewhat predictably, much more complexity than was anticipated, and a deep realization that there are many more details of our bodies that we do not understand. Knowing James Watson's genetic makeup is only a tiny piece of the puzzle. We must understand the diversity in our genomes and how they change over time, but even then, the puzzle contains many more layers of complexity.

I believe they way forward from this expanding complexity problem is the use of rigorous probabilistic methods. In this thesis I suggest a simple application of this idea by applying Bayesian networks to the study of complex, multi-species RNA mixtures in human stools. But, I think Bayesian statistics could be expanded to many more problems in medicine, and we would be all the better for it.

Chapter 1 - Background

The assessment of host and microbial transcripts in stool has received relatively little attention in the past several decades. Initial assessments have been relatively simplistic, but with the genomics revolution, massive datasets can be produced. Tools to illuminate useful biological information from these data are lacking. What follows is a survey of the published data regarding stool mRNA assessment.

Stool is a complex analyte containing a large number of bacterial and human cells, as well free protein and undigested food debris. Inhibition of downstream reactions, such as PCR, is common. A variety of methods have been used successfully to obtain pure RNA which is relatively free of inhibitors[1-3].

1.1 Viral Genomes

The first stool RNA species to be isolated in humans were those belonging to viral genomes. A large number of enteric viruses, such as hepatitis A[4] virus, poliovirus[5], and rotavirus[6] have single- or double-stranded RNA genomes, and the isolation of these viruses and their genomes have been accomplished for decades. A number of investigators have isolated total RNA from stool for the purposes of viral genotyping and discovery[7, 8], as a large percentage of enteric viruses have RNA genomes. Furthermore, it is now possible to take a broader approach and assess the metagenome of specific viruses or groups of viruses[9] both in infectious disease, but also in gastrointestinal disorders such as inflammatory bowel disease, which have been shown to be related to dysbiosis of both the microbiome and the virome[10].

While the techniques to isolate nucleic acid are similar, the information contained in genomic RNA is distinct from that in the multitude of rRNAs and mRNAs within stool.

1.2 Colon Cancer Biomarkers

A large portion of the investigations of mRNA in stool has been related to the detection of specific transcripts as putative biomarkers for colorectal adenocarcinoma in humans. Initial work was performed by Davidson, et al when they followed up their previous study on protein kinase C (PKC) expression in colon adenocarcinoma by assaying total stool mRNA from mouse stool for multiple isoforms of PKC using RT-PCR[11]. They concluded this method would be useful to screen for colon cancer, and their follow up studies in a rat tumor model showed stool PKC-β2 transcripts to be elevated in animals with tumors[12].

Subsequent work has found that a variety of other markers might be useful in colon cancer. Davidson, et al. applied their earlier methods to human samples, using RT-PCR to assay 11 different mRNAs in controls, patients with inflammation, and patients with colorectal cancer[13], although no clear pattern was seen among a small number of patients. Cyclo-oxygenase-2 (COX-2) is commonly overexpressed in aerodigestive tumors, and COX-2 mRNA has been found in the stools of a small series of patients with colorectal adenocarcinoma by Kanaoka, et al[14]. Subsequent work by the same group showed that sensitivity improved with the addition of matrix metalloproteinase-7 mRNA measurement[15], and that multiple other indicators (such as tumor size, number of exfoliated cells, and tumor expression of COX-2) correlated positively with detectable stool COX-2 mRNA[16]. Another group showed a somewhat lower sensitivity for detection of colon cancer using stool COX-2 mRNA[17]. Family history of colorectal cancer and polyps has also been associated with an increase in stool COX-2 mRNA in healthy adult volunteers[18].

Yamao, et al. showed that CD44 variant transcripts, which had been previously found to be elevated in tumor cells, were often present in adult humans with colorectal cancer, and post-operative samples showed a decrease in these stool mRNAs[19]. Fecal cytokeratin 19 (CK19) and ribosomal protein L19 (RPL19) mRNAs have been shown to be associated with an increased risk of metastatic colorectal cancer in a prospective cohort of adults[20]. In a recent study by Koga, et al., multiple stool mRNA markers (CEA, MMP7, MYBL2, PTGS2 and TP53) were assayed in a large sample of patients with and without colorectal cancer[21]. Their results were variable, with low sensitivity (58.3%), moderate specificity (88.1%), and mRNA detection was dependent on the number of sloughed colonocytes recovered in stool. Davidson's group extended their previous work recently by simultaneously measuring a large variety of stool mRNAs associated with insulin resistance and colorectal cancer, then applied linear discriminate analysis to determine the most useful combinations of genes[22]. While many researchers had expressed a need for analyses using multiple mRNA markers simultaneously, this represents the first effort to do so mathematically.

Additionally, stool micro RNA (miRNA) screening shows promise in illuminating mechanisms of physiology and disease when combined with other analyses, such as mRNA sequence[23]. miRNAs are probably more stable in complex analytes like stool, and so may have an important future role in noninvasive cancer screening[24].

Screening strategies using cancer-associated DNA sequences have now entered into the clinic and received FDA-approval[25]. Recent work on targeted mRNA transcripts, such as integrins, seems to show more promise as a useful clinical test than previous work.[26]

1.3 Bacterial Gene Expression

The assessment of bacterial virulence factors in vivo is of significant clinical interest, and has been the target of several investigations. While measurement of transcripts in vitro from cultured bacteria from human feces is a mainstay of microbiology, the direct assessment of mRNA in stool has not received as much attention. In 2003, Fitzsimons et al. showed using quantitative RT-PCR that housekeeping genes in *Lactobacillus* can be detected in stool when as little as $2x10^7$ cells per gram of stool (about 0.1% of the total bacterial population)[27]. While this work did not attempt to assess the native activity of stool bacteria, it did show that bacterial mRNAs can be measured, and that they relate quantitatively to the bacteria present. In vivo assessment of toxin gene expression by assaying stool mRNAs in *Vibrio cholerae* infection could discern early from late phases of the disease[28], and in *Escherichia coli* O157:H7 showed differences between bovine and human infection[29]. More recently, Dunaev, et al. showed that stool mRNA assessments to determine the quantity of viable non-culturable bacteria (VNCB) may be a useful measure for quantifying potentially dangerous coliform bacteria, but that RNA isolation from these biosolids is a challenging task[30].

1.4 Other Conditions

A variety of other human conditions have shown changes in stool mRNA content. A subset of adults with HIV infection have been shown to have detectable HIV RNA and CD4 mRNA in their stools[31]. Recently, Kaeffer et al. studied stool samples and gastric aspirates obtained from premature infants, and found that several housekeeping mRNAs (β-actin, GAPDH, and PER1) are present in measurable quantities in a majority of these samples[32]. Bennett et al. showed that there is considerable inter-patient and inter-time variability in housekeeping (GAPDH) and inflammation-associated transcripts (IL-8, calprotectin) in stool, but that intra-specimen reproducibility is high[3]. The same group showed in a subsequent study that inflammatory transcripts were elevated in the stools of children with bacterial colitis when compared with age-matched controls[33]. These same approaches have been extended to other infections, such as C. difficile colitis, with similar results[34].

A similar disease process, environmental enteric dysfunction (EED), which is a complex interaction between the gut microbiome, the host immune system, and the host nutritional status, has been traditionally measured using more invasive and challenging techniques, such as lactulose permeability. Newer approaches using transcriptomic sequencing have identified specific mRNAs associated with EED, such as those associated with T-cell proliferation or mediators that dampen hormone response. These have been shown to be comparable to more traditional techniques at detecting EED.

1.5 Transcriptomes of Environmental Samples

Chen et al. showed that individual bacterial transcripts can be measured in waste water (presumably isolated from within biologically active cells), and that these correlate with nutrient availability and bacterial contamination[35]. Since that time, several groups have taken an agnostic, sequence-based approach to interrogation of total RNA in stool. Urich, et al. published results of the analysis of a single 454 sequencing run on total RNA isolated from a soil sample[36]. Their results show that a predominance of RNA reads are rRNA (74.8%), and that the remaining mRNAs are largely unaligned or unassigned to a specific gene ontology (8.2% map to a specific gene, 17.0 % are unassigned). Bacterial ribo-tags and mRNAs predominated in their samples. Bailly, et al. used a metatranscriptomic approach to describe eukaryotic diversity, and compared 18S rRNA sequence with metagenomic DNA sequencing [37]. Poor correlation was found between these two measures. Shrestha, et al. applied 3730 sequencing to total RNA enriched for mRNA using specific Shine-Dalgarno sequence primers and found significant functional diversity in a paddy soil community [38].

Marine bacterial communities have also been a focus of metatranscriptomic analysis. In 2005, Poretsky, et al. used selective hybridization to enrich freshwater lake samples for bacterial mRNAs, and found a great deal of functional diversity along with many previously unknown sequences[39]. In 2008, Frias-Lopez, et al. performed 454 sequencing on a sample of marine surface water[40]. They found ~ 85% of cDNA sequence (compared with ~ 50% of DNA sequence) did not align well to the NCBI non-redundant protein database. Those sequences that did align were predominantly related to photosynthesis and carbon fixation, consistent with the organisms' ecological niche. Gilbert et al. reported one of the first large-scale metatranscriptomic analyses of complex microbial communities by applying 454 sequencing to a marine community and then analyzing both cDNA and DNA sequence[41]. A surface water sample was subjected to an artificial "bloom," and then samples were obtained before, mid-bloom, and afterwards, so they could compare both DNA and cDNA among different time points. They found a marked difference between the changes in DNA sequence when compared to changes in cDNA sequence, and concluded this represents an alteration in gene transcription within taxa, rather than a change in community structure. More recently, Hollibaugh, et al. have demonstrated the strong potential of metatranscriptomic analysis to answer specific biological questions by querying a series of 454 sequences for evidence of ammonia-oxidation performed by a low-abundance organism (Crenarcheota)[42].

Fermentation samples have also been subjected to metatranscriptomic analysis, and have yielded interestng insights. Nam, et al. were able to show that a number of lactic acid bacilli are present in fermenting kimchi, and that, by metatranscriptomic analysis, these same bacteria are producing transcripts consistent with participation in fermentation, and in ratios similar to their average relative composition as determined by DNA sequencing[43]. Earlier this year, Weckx, et al. demonstrated a similar scenario using microarray data in fermenting sourdough[44].

1.6 Stool Transcriptomes

Stool is a convenient and readily available analyte for human studies, but until recently, the application of metatranscriptomics to fecal samples has been sparse. Microarrays have been used to compare the expression profiles of bifidobacteria in a small sample (4 patients) of breast-fed and formula-fed infants aged 1 week to 10 months, and may be effective at differentiating the two groups biologically[45]. This work was expanded upon by Chapkin, et al., using microarrays to measure human expression profiles in a larger number of breast-fed and formula-fed infants at 3 months of age[46]. Booijink, et al. applied RNA-fingerprinting (using restriction fragment length polymorphisms) to the stools of two healthy subjects, and found they could be accurately discriminated with this analysis[47]. Poroyko, et al. used RNA-Seq to characterize the functional diversity of the stool microbiome in formula-fed vs. mother-fed newborn piglets.

RNA sequencing has only more recently been applied to human stool samples. In the last year, Turnbaugh, et al. reported the first detailed, deep, sequencing-based assessment of human stool transcripts by comparing the expression profiles of two identical twins using RNA-Seq[48]. They found considerable inter-individual differences, as well as a great number of transcripts encoding hypothetical proteins. RNA-Seq continues to be used to assess specific disease states where invasive biopsy or other assessment is difficult, such as to determine the intestinal maturation of preterm infants[49].

Taken as a whole, the body of work studying stool RNA shows that a great deal of information is likely to be present, but that extraction of this information is hindered by a combination of relative instability of the analyte, as well as an unknown degree of uncertainty in alignment of reads and interpretation of biological data.

1.7 The Problem - Why Use a Probabilistic Model?

A significant problem with mRNA alignments to a large number of bacterial genomes is the high degree of homology among intestinal bacteria. *Shigella* and *E. coli* are virtually identical, for example.

Klebsiella and *E. coli* are also close evolutionary relatives with a high degree of gene homology. What's more, many unrelated bacteria may share plasmids or other functional genes acquired through gene transfer after a long period of co-evolution in a shared environment. This often results in many high scoring hits for each mRNA when cDNA sequence reads attempt to align to a large number of related bacterial genomes.

Our proposed solution to this problem is the use of a Bayesian network to improve alignment fidelity using two effects that are likely to change which mRNA-species pair is selected by an alignment: (1) the alignment score, (2) the prevalence of a bacteria by 16S sequencing (which as we will see below is probably more accurate than the alignments to rRNA in RNA-Seq). Such a model would be a large web of interconnected probabilities ideally suited for a Bayesian network.

Chapter 2 – Materials and Methods

2.1 Samples

Stool samples were taken from a single male preterm infant born at 24 weeks estimated gestational age. A single sample at 23 days of age, and another at 25 days of age were frozen at -80°C within 4 hours of collection. Informed consent was obtained from the subject's mother. Samples were collected as part of an NIH-funded study to investigate the microbial contribution to necrotizing enterocolitis (NIH UH2AI083266). The subject did not go on to develop necrotizing enterocolitis, so was considered a control patient in the study with normal acquisition of gut microbiota.

2.2 Nucleic Acid Extraction

DNA was isolated using the QIAamp DNA Stool Mini Kit (QIAGEN, Benlo, Netherlands – Product # 51504) in conjunction with the QIAcube automated nucleic acid isolation system (QIAGEN, Benlo, Netherlands – Product # 9001292). Aqueous DNA was then subjected to 16S sequencing as described below.

Total RNA was isolated from stools as described previously[3]. Briefly, frozen stool was subjected to a sequence of bead-beating, phenol-chloroform extraction, and then silica column extraction. The final silica column step was performed with the QIAcube, using the protocol provided by the manufacturer for RNA isolation. RNA integrity was assessed using the Agilent 6000 Nano Kit (Agilent Technologies, Santa Clara, CA, USA – Product # 5067-1511).

2.3 cDNA Production and Sequencing

DNA was subjected to 16S sequencing using protocols developed as part of the Human Microbiome Project Demonstration Project on Necrotizing Enterocolitis. The first major publication from that project[50] details the exact method employed:

> The V3-5 region of the 16S rRNA gene was sequenced on the Roche-454 platform to define the composition of the bacterial community. Sample preparation, DNA isolation, sequencing, data processing followed standardized protocols developed by the Human Microbiome Project consortium. The minimal sequence length was 200 bp, and chimeric sequences were removed by Chimera-Slayer. Average quality of scores of <35 were used as the minimum to remove low-quality reads. Sequences meeting the above criteria were further classified by the Ribosomal Database Project (RDP) Naive Bayesian Classifier version 2.5 using training set 9 from phylum to genus level[50].

The 16S profile of each sample was then recorded as a number of total reads aligning to the primer for a specific V3-5 region of a specific bacterial genus. We then knew the proportion of each sample constituting each bacterial genus.

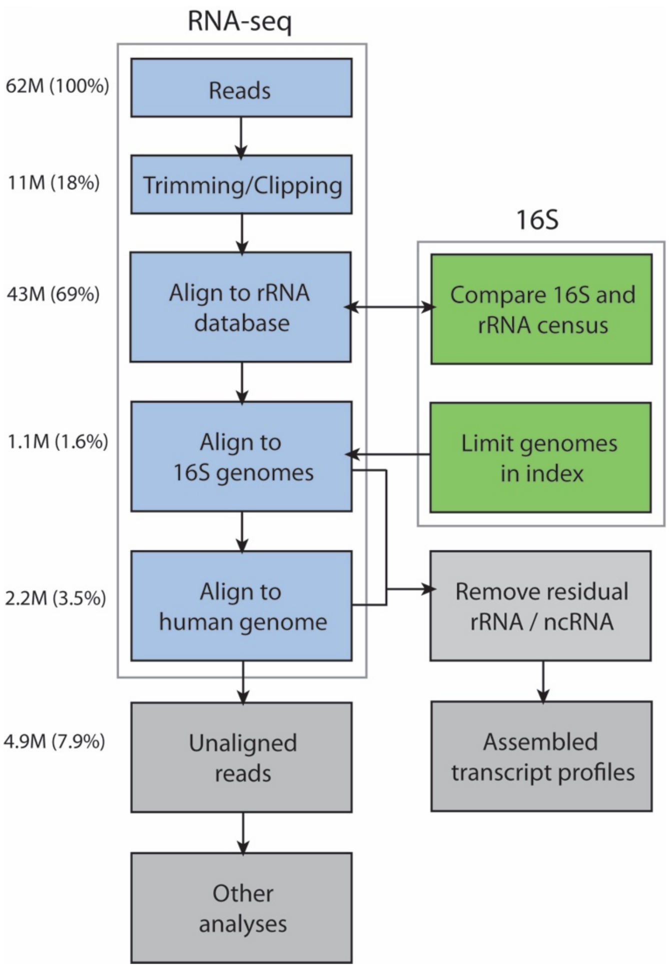
cDNA for RNA-Seq was created using both the Nu-Gen Ovation and Nu-Gen Ovation Prokaryotic enrichment systems. Sequencing was performed using the Illumina platform as described above. Sequences obtained from the Ovation Prokaryotic / Illumina system were then used to study the effect of Bayesian networks on the fidelity of assigning mRNA-organism pairs in complex mRNA mixtures.

2.4 Analysis

2.4.1 Overall Pipeline

The overall analysis pipeline is portrayed in Figure 1. Prior to alignment, reads were assessed for quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). All reads were first scrubbed of adapter and barcode sequences. Adapter sequences, bases with a Phred quality score < 28, and reads with a length < 40 (for 100 bp reads) and < 20 (for 40 bp reads) after trimming were removed using the FastX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Alignments were performed using Bowtie[51], with the addition of TopHat[52] for human alignments, to improve the likelihood of correctly mapping reads that span splice sites. Transcripts were assembled using CuffLinks. Ribosomal indices were built using the latest version of the SILVA database containing all known ribosomal RNA sequence[53]. Human indices were built using the February 2009 (hg27) NCBI assembly. Microbial indices were created using either all microbial sequence deposited in NCBI, or by creating custom indices based on the 16S census of individual samples. All genera (or families) with greater than a 0.1% contribution to the microbial census by 16S sequence analysis were included for each sample using all sequence deposited in the NCBI for that genus or family (both fully sequenced organisms and draft sequence). rRNA, mRNA, and unaligned bins were subjected to further analysis outlined below.

Figure 2.1 – Overall sequence analysis pipeline.



Reads which aligned to rRNA sequence were compared, by genus (or family, if necessary), to those genera and families which accounted for > 0.1% of the microbial census. In order to more accurately make comparisons with 16S sequence data, rRNA sequences that aligned to hypervariable regions were removed and compared separately. The percentage contribution to the microbial census was then compared between 16S and rRNA-Seq.

Reads which aligned to ribosomal RNA or other non-coding RNAs (and were missed by the previous ribosomal index alignment step) were removed prior to analysis of transcripts. For both microbial and human alignments likely to be mRNAs transcripts were assembled and binned according to KEGG category.

Additionally, a subset of specific human transcripts known to be involved in the maturation of the human intestine, and in the control of gut inflammation, were assessed separately, and then compared by relative abundance between samples and sequencing methods. These include: IL-1β, IL-6, IL-8, IL-10, S100A8, TGF-β, TNF-α, TLR-2, TLR-4, HMGBP-1, EGF, and GAPDH.

2.4.2 Complexity Assessment

We assessed the average complexity of our sequence at each step of the alignment process to help determine the likely contribution of each pool to unaligned reads. Complexity was determined by compressing the total sequence of each pool of reads using the expert model (XM) compression algorithm[54], and then computing the bits per symbol (BPS). This algorithm uses a series of experts: a simple Markov expert, a context Markov expert that uses the previous 512 symbols, and a hash table full of copy experts that encompass many combinations of potential sequence repeats. The relative weight of each expert is computed by:

$$\begin{aligned} w_{\theta_k,n} &= P(\theta_k|x_{1..n}) \\ &= \frac{\prod_{i=1}^{n} P(x_i|\theta_k, x_{1..i-1})P(\theta_k)}{\prod_{i=1}^{n} P(x_i|x_{1..i-1})} \end{aligned} \tag{2.1}$$

These weights are then combined using Bayesian averaging and their (relative) predictions used to compress the given sequence:

$$\begin{aligned} P(x_{n+1}|x_{1..n}) &= \sum_{k\in E} P(x_{n+1}|\theta_k, x_{1..n})w_{\theta_k,n} \\ &= \sum_{k\in E} P(x_{n+1}|\theta_k, x_{1..n})P(\theta_k|x_{1..n}) \end{aligned} \tag{2.2}$$

Sequences with a higher average BPS thus has less sequence complexity since more information can be stored per unit. For comparison, we used the same algorithm to compute the complexity of a simulated set of repeated adapters (where we would expect low complexity), as well as simulated random reads from human cDNA (where we would expect high complexity).

2.4.3 Creation of Simulated Reads

In order to test our Bayesian network, we needed to create simulated transcripts that varied in known ways, so that we could then determine if the model accounted for these changes in a predictable way. We simulated two sets of 1000 transcripts using built-in Python string functions:

1. A set of transcripts that were random at any coding portion of the genome and were normally distributed around the frequency of each genera. The model should have minimal effect on the posterior probability distribution of this data set.

2. A set of transcripts obtained identically as #1, but with a skewed frequency of genera. *Escherichia* and *Staphylococcus* were the dominant genera, rather than *Escherichia* and *Klebsiella*.

These simulated sets of transcripts with known relationships with the 16S proportions could then help determine the effectiveness of the model.

2.4.4 Fitting to and Sampling from the Dirichlet Distributions

Since our data consist of a distribution of the probability of several bacterial genera, we chose to use the Dirichlet distribution. Additionally, the Dirichlet distribution is ideally suited for use as a conjugate prior in a Bayesian network, which is described below.

All distributions were constructed and manipulated in our final software tool by modifying existing modules written in Python. The NumPy (www.numpy.org) and SciPy (www.scipy.org) packages both contain multiple classes helpful in the manipulation of Dirichlet distributions. Visualization of distributions was accomplished using the matplotlib Python package (www.matplotlib.org).

The Dirichlet distribution is useful when we have a set of conjugate probabilities, as we do with 16S data and data derived from mRNA alignment scores which are the conjugate probability of a set of bacterial genera:

$$\sum_{i} \Theta_1 + \Theta_2 + \Theta_3 + \dots = 1 \qquad (2.3)$$

Where θ represents the probability of each bacterial genus being present in the 16S population or the metabolically active mRNA population. These experimentally determined θ values can be used to construct a Dirichlet distribution using maximum likelihood estimation (MLE), since the parameters for the distribution (α), are present in the following conditional probability:

$$\mathrm{Dir}\left(\boldsymbol{\alpha}\right) \rightarrow \mathrm{p}\left(\boldsymbol{\theta} \mid \boldsymbol{\alpha}\right) = \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\alpha}_i\right)} \prod_{i=1}^{k} \boldsymbol{\theta}_i^{\boldsymbol{\alpha}_i - 1} \qquad (2.4)$$

MLE was performed on experimental θ values using an existing Python module written by Eric Suh (https://github.com/ericsuh/dirichlet). The code was modified slightly to allow for longer computation cycles, since many of the experimental data we attempted to fit had wide variability and high Dirichlet parameters, which would often result in failure of the MLE algorithm to converge.

Additionally, since we usually had a single 16S or mRNA sample, we wanted to account for repeat sample / intrasample variation, since our analyte (stool) is not homogenous. Some previous work has shown intrasample variation of genus assignment to be low, about 5%[55]. So, when performing parameter estimation, 100 samples were generated with a variation of 3-7% in the most common genera, and subsequent genera adjusted randomly for that change, so that they all still summed to 1. This also resulted in greater likelihood of the MLE algorithm converging, as performance was noticeably poorer with smaller numbers of experimental samples for parameter estimation.

For the 16S census of the sample, the probabilities were straightforward, since the top 4 genera accounted for the vast majority of the likelihood. **As such, we built all subsequent distributions using the probabilities of only these top 4 genera: *Escherichia*, *Klebsiella*, *Staphylococcus*, and *Enterobacter*.** We thus created the prior θ distribution of:

$$\Theta_1+\Theta_2+\Theta_3+\Theta_4 = P(Escherichia)+ P(Klebsiella)+P(Staphylococcus)+ P(Enterobacter) \tag{2.5}$$

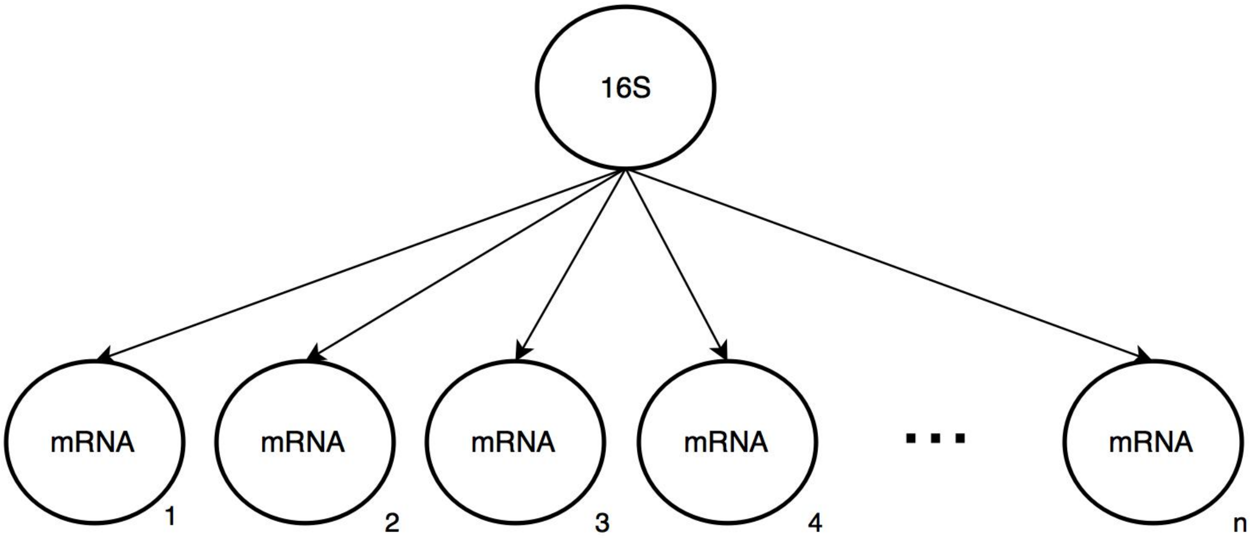
Assigning probabilities to mRNA alignments was more challenging. First, when aligning using BLAST, only the top 4 genera were used as the index. Second, once an alignment score was generated for each, probabilities were generated by determining the relative alignment score of each genus. If no alignment occurred, the θ was set to 0 for that genus. If there were multiple genes that aligned to a specific genus, the maximum alignment score was used.

We then had two data sets: (1) a 16S distribution that represented the prior probability of each genus, (2) a set of several thousand alignments that represented many examples of the likelihood of each mRNA being from one of those 4 genera. These θ's could then be converted to Dirichlet distributions and analyzed using a Bayesian network.

2.4.5 Bayesian Network Model for Prokaryotic mRNA

We constructed the Bayesian network according to the following schematic:

Figure 2.2 – Bayesian network model schematic.



The single 16S node layer represents the prior probability (Dirichlet prior) of which genera are metabolically active, which is simply those bacteria that we know to be present. As detailed in the overall algorithm below, this prior probability is then updated by the many distributions in the second layer via inference, resulting in a change in a successive change in the Dirichlet parameters, eventually resulting in the posterior probability, which is the model's estimate of the true metabolically active population of bacteria.

We can therefore find a new distribution that is a re-parameterization of the initial Dirichlet prior:

$$P(\Theta|X) = Dir(N + \alpha) \quad (2.6)$$

2.4.6 The Algorithm – RNABayes

The RNABayes software tool was written using Python 2.7 with the following features:

Input: Set of 16S probabilities (1 or more), Set of mRNA probabilities (1 or more)
Output: Set of estimated proportions of metabolically active bacteria.

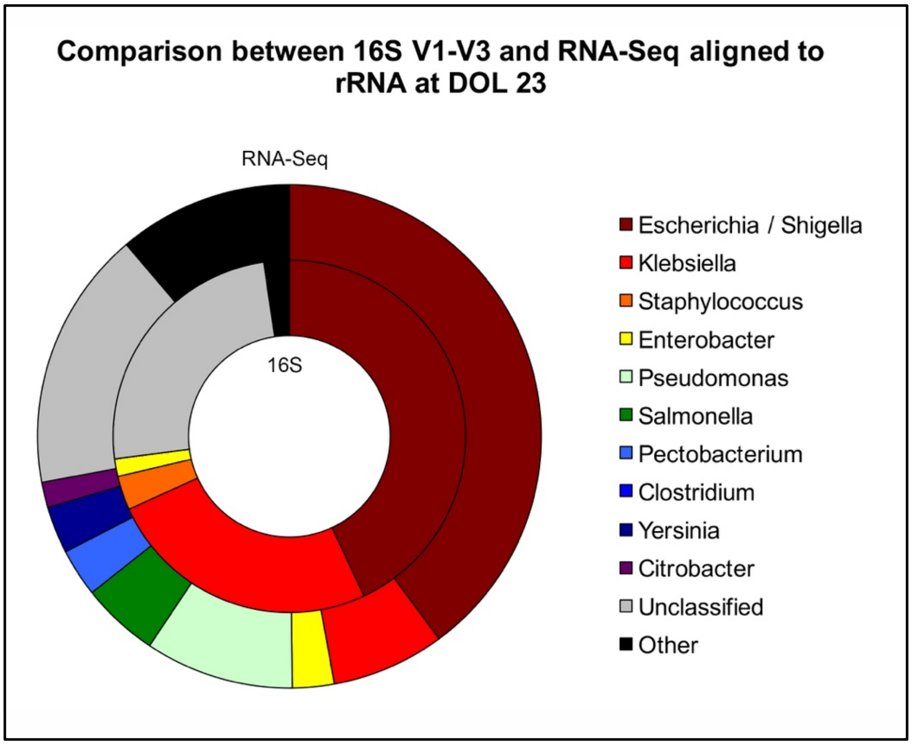
1. Read in 16S probabilities in the format ([a, b, c, d]).
2. Read in mRNA probabilities in the format ([a, b, c, d]).
3. Use maximum likelihood estimation to generate parameters for a Dirichlet distribution for each set of probabilities.
4. Assign each of these Dirichlet distributions to nodes in a Bayesian network as described above.
5. Perform inference on each node, updating the prior probability in layer 1 with the observations in layer 2.
6. Return the final updated distribution from layer 1 as the estimate of the metabolically active population of bacteria.

Chapter 3 – Results and Conclusions

3.1 rRNA and 16S Census Comparison

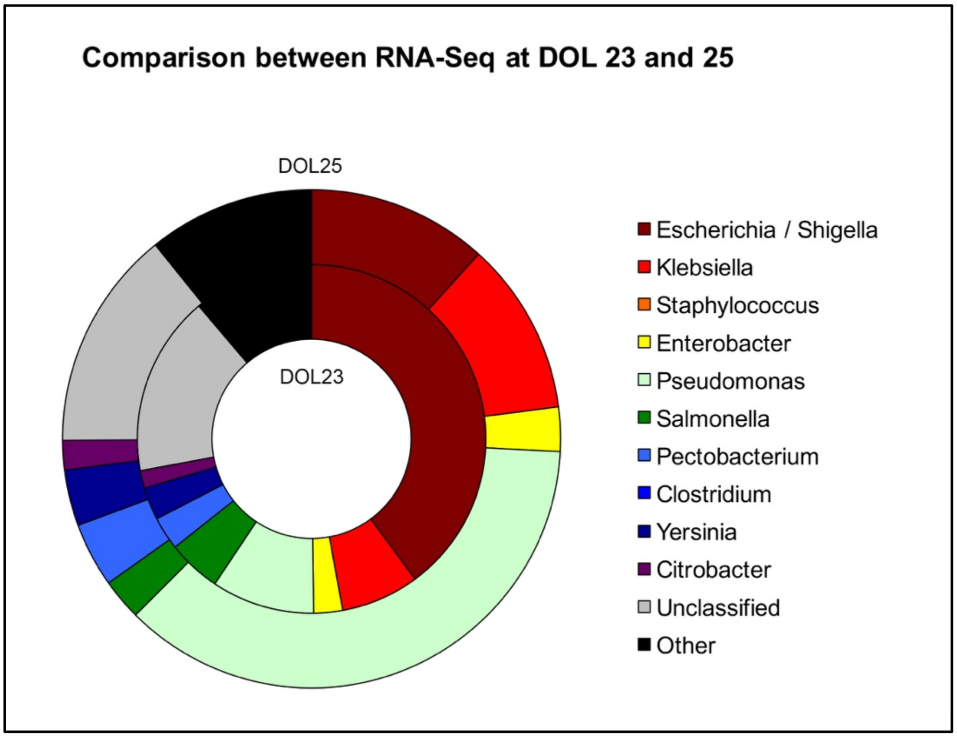
We made multiple comparisons between 16S genus profiles and rRNA profiles obtained via RNA-Seq. We took all genera with >1% of the total and compared them across multiple variables: rRNA vs. 16S, DOL 23 vs. DOL 25, between two runs of the same sample, and 16S V1-V3 vs. V3-V5. These results are shown in Figures 3.1, 3.2, 3.3, and 3.4, respectively.

Figure 3.1



16S and RNA-Seq showed similar profiles, but RNA-Seq had far more variety, and probably reflects the fact that RNA-Seq sequences a large number of less variable RNA regions which might align well to multiple bacterial genomes, whereas V1-V3 provides greater resolution between genera. Furthermore, some portions of rRNA may be less stable in environmental samples such as stool.

Figure 3.2



The shift in bacterial species across a 48-hour period is consistent with comprehensive work describing the transition of bacterial genera across the first few weeks of life[50]. Two genera (*Pseudomonas* and *Escherichia*) show large reciprocal shifts, whereas the remaining portions are unchanged.

Figure 3.3

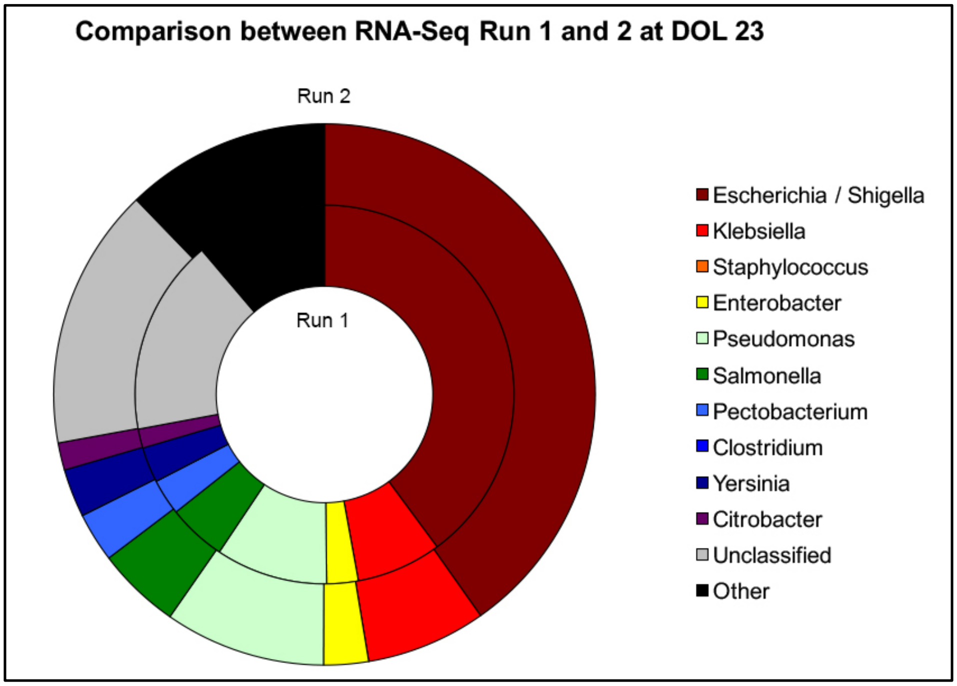
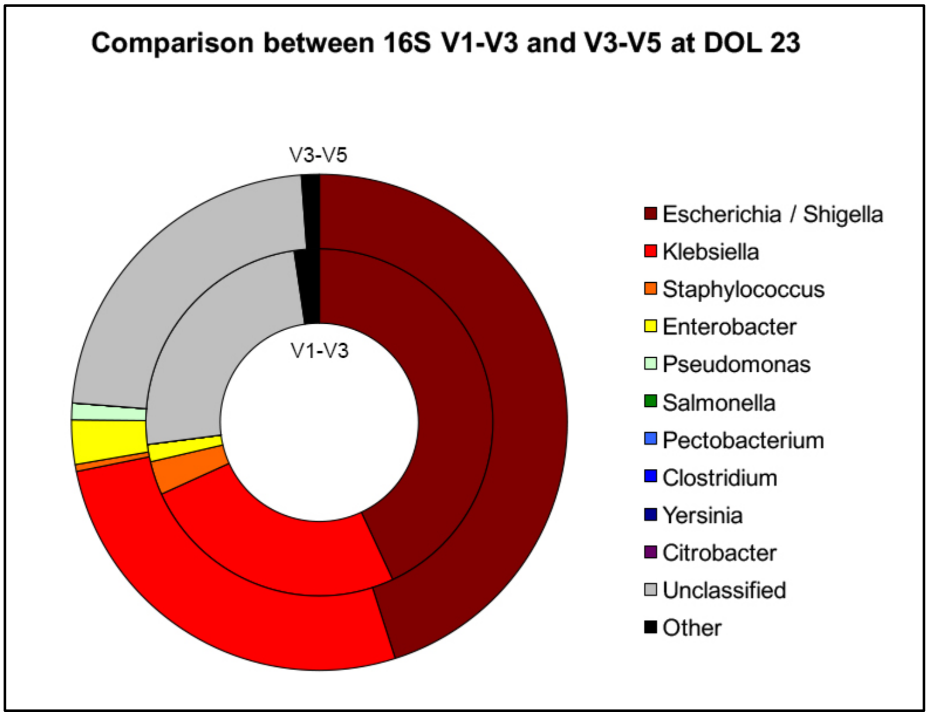


Figure 3.4



There are only minor variations between multiple runs of RNA-Seq and between 16S rRNA gene sequencing at different hypervariable regions, indicating that these are not important sources of variation.

3.2 Human mRNA

Since human mRNA was only present from a single individual from a single species, we did not subject it to further processing as we did with prokaryotic mRNA (see Bayesian Network Analysis for Prokaryotic mRNA, above). Using MEGAN[56], we determined the KEGG category for all aligned mRNA reads and compared DOL23 and DOL25. Results are shown in Figures 2.6 and 2.7 below. The relative distribution was similar for each sample. We further determined the KEGG categories for transcripts within the Metabolism category, since this was the dominant transcript type. The majority of these alignments were for mRNAs coding for carbohydrate (CHO) metabolism. These results are consistent with previous transcript profiles of individual gut microbiota, and with the overall metabolic activity of gram negative and anaerobic gut bacteria[57, 58].

Figure 3.5 – KEGG categories for DOL 23

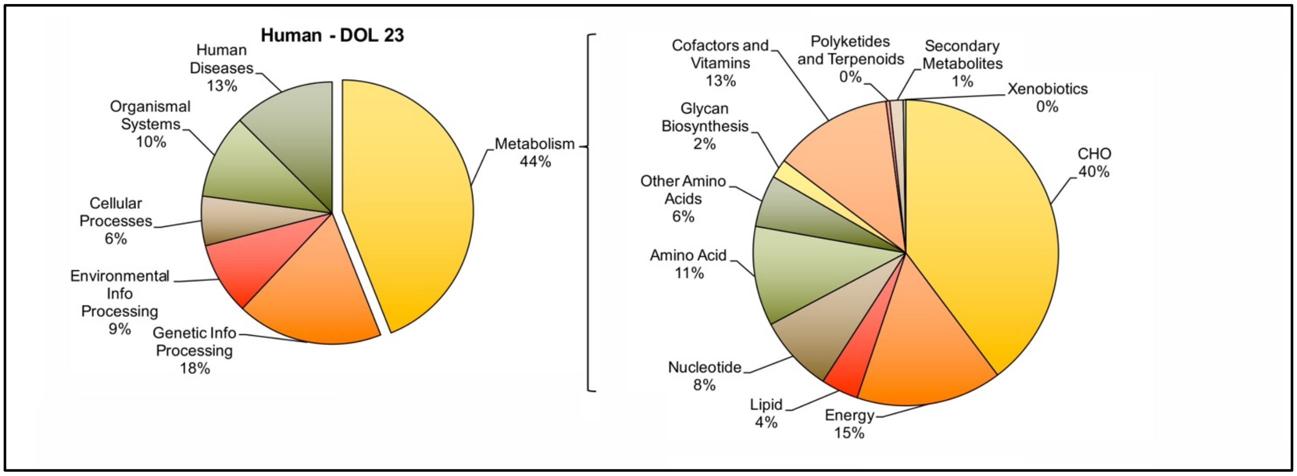
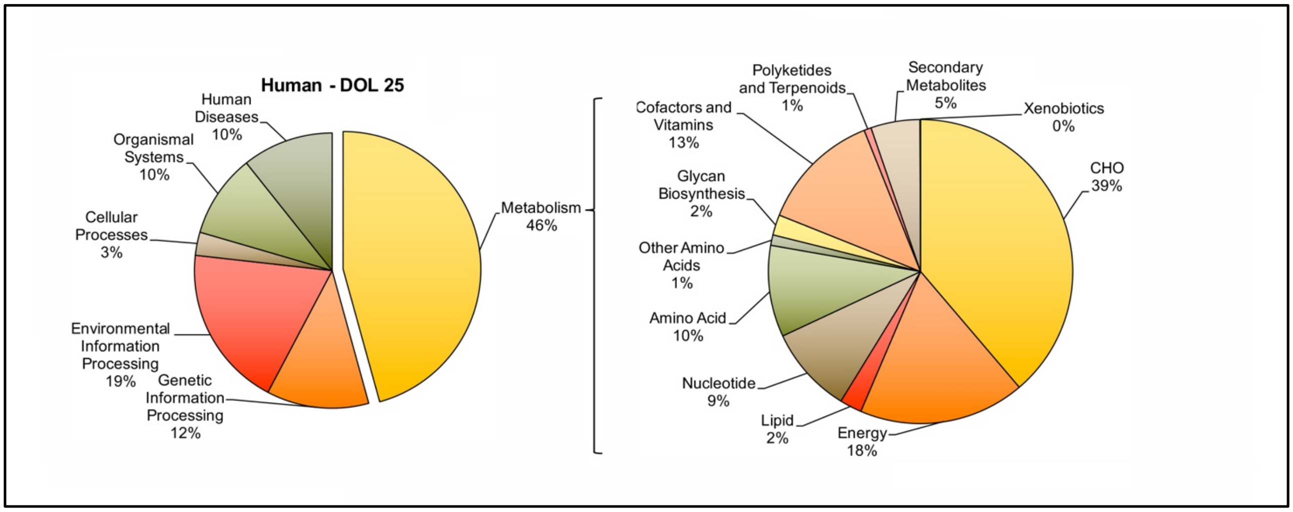


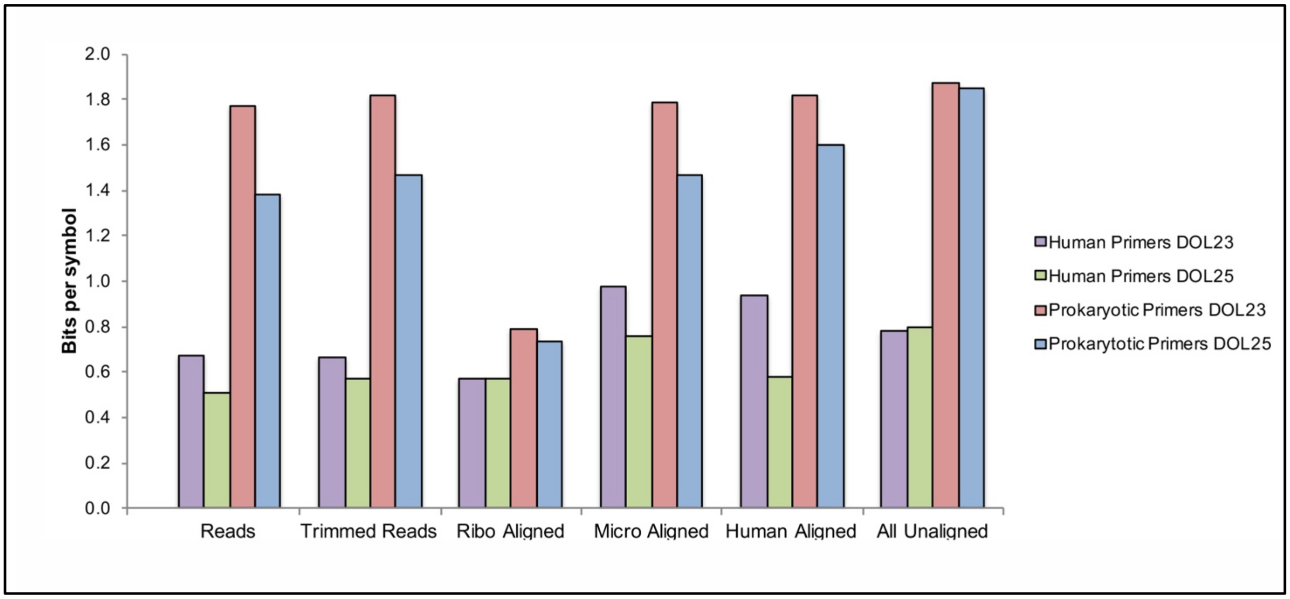
Figure 3.6 – KEGG categories for DOL 25



3.3 Complexity Analysis

While we might assume that more complex sequence would consistently be present in mRNAs compared to ribosomal sequence, this may not be true of all genes, and highly repetitive k-mers might make a larger contribution to the mRNA sequence pool. However, in our sample, there was substantially more complexity in non-rRNA (mean complexity = 1.24, median = 1.22) reads than in those aligning to rRNA (mean complexity = 0.667, median = 0.635), as we might expect. Furthermore, the striking difference between the average complexity at each alignment step when comparing DOL 23 and DOL 25 serves to underscore the heterogeneity one expects to find in the large, heterogenous total RNA pool within human stool.

Figure 3.7 – Read complexity with human and prokaryotic primers



3.4 Bayesian Network Model

3.4.1 Fitting Probabilities to the Dirichlet Distribution

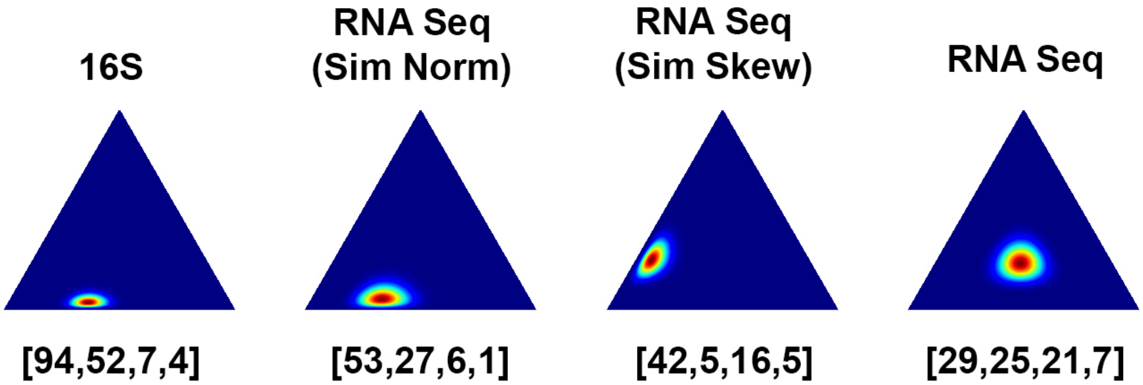
With a 3-7% variation in the most common bacteria, the Dirichlet distribution had a high scalar for parameters as determined by MLE. This resulted in a low variation in the probability density, as seen in Figure 3.8 "16S". When mRNA transcripts were simulated based on a normal distribution of the most common bacteria, the distribution was slightly more dispersed, as seen in Figure 3.8 "Sim Norm". When a skew was added decreasing the likelihood of a Klebsiella transcript, the probability distribution shifted to a new axis between Escherichia and Staphylococcus, but had a similar dispersion, as seen in Figure 3.8 "Sim Skew."

However, when actual RNA-Seq alignments were used, there was far more dispersion in probability density, which has several potential causes:

1. High degree of homology between Klebsiella and Escherichia, resulting in frequently similar alignment scores.
2. The frequent occurrence of a very low alignment score for 2 or more of the genera.

3. The occurrence of a high degree of homology between all genera, likely in the case of heavily conserved bacterial genes.

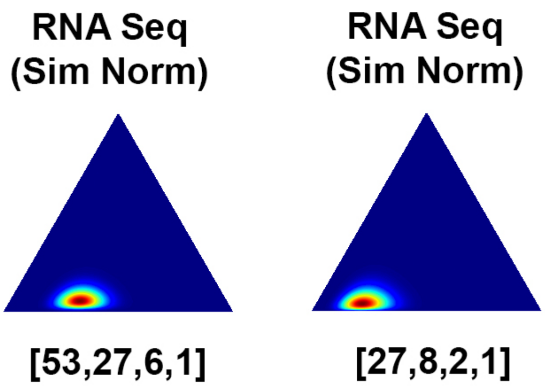
Figure 3.8 – Prior probability distributions for all datasets



3.4.2 Results from Inference on Simulated Data Set, Normally Distributed

The simulated transcript set with a normally distributed change in transcripts behaved as expected. There was a slightly more dispersion after the model adjustment than in the initial transcript set. The shift towards *Escherichia* can likely be explained by a disproportionate change in the other parameters (a 5% change in a higher probability was a 50% change in the lower probabilities).

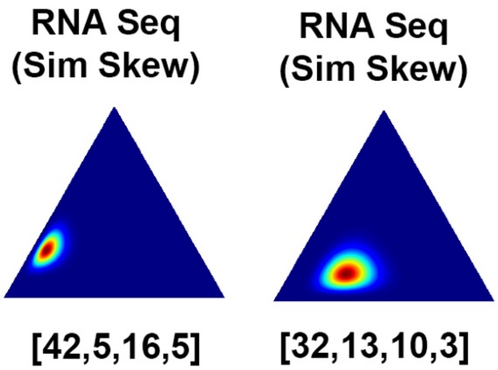
Figure 3.9 – Posterior probabilities for simulated data set, normal



3.4.3 Results from Inference on Simulated Data Set, Skewed

The skewed simulated transcript set also performed as expected. When the transcripts were skewed to be predominantly *Escherichia* and *Staphylococcus* rather than *Klebsiella*, the probability density shifted to the edge between the two dominant genera in the 16S data set (*Escherichia* and *Klebsiella*). The relative dispersion remained similar, since transcripts were not otherwise selectively drawn from one genome or another.

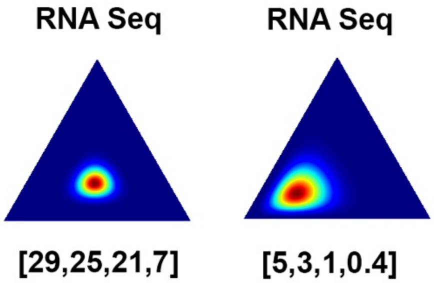
Figure 3.10 – Posterior probabilities for simulated data set, skewed



3.4.4 Results from Inference on RNA Seq Data Set

As the final analysis, we sought to determine how the model performed on a real data set. Both of the simulated examples performed as we would expect, so we would also expect for the model to adjust both the probabilities by genus, but also to likely increase the amount of uncertainty in the distribution, since the two pieces of information (16S and RNA-Seq) contained such different predictions. Our hypothesis was confirmed. The probability distribution expanded substantially, indicating a large amount of uncertainty, and the region shifted towards *Escherichia* as we would expect.

Figure 3.11 – Posterior probability distribution for RNA Seq



3.5 Discussion

We have described a novel approach to improving the fidelity of alignment of complex bacterial mRNA populations in human stool using Bayesian inference. The method relies on the construction of a small Bayesian network using Dirichlet distributions to model the conjugate probabilities of the most common bacterial constituents in a stool sample. Most work has focused on detection of the host transcriptome, but we hope that further work building on probabilistic improvement of these complex and difficult transcriptomes will help us gain further insight into the metabolic activity of these important bacterial communities.

The results of the model are not surprising, and confirm the utility of Bayesian networks for this application. In general, whichever direction the two sets of distributions (16S vs. RNA-Seq) tended to differ, the model found a distribution with both (a) more uncertainty and (b) a shift towards the other distribution. This may especially helpful in specific disease states where one or more populations of bacteria are having a disproportionately large effect on the host.

3.5.1 Weaknesses

There are several important weaknesses to this work. First, we had no gold standard to determine efficacy of our Bayesian model in improving fidelity of aligning mRNA-species pairs correctly. While the model performed modestly well on simulated reads, we are unable to determine if that performance held for real world sequence, especially given how different the real samples were from the simulated ones. There could be a cause of variation completely unrelated to the actual metabolic activity of the bacteria in the sample (such as increased degradation of some mRNAs, different locations within the stool of some bacteria vs. others). Secondly, only a small portion of total RNA was mRNA, which was the eventual analyte we chose. Of those mRNAs, a minority were prokaryotic. Given the massive community that is the human intestinal microbiome, it is surprising that we are able to see only a tiny fraction of the bacterial transcriptome. However, we have no reason to believe there is disproportionate degradation of one gene's mRNA over another.

3.5.2 Future Directions

Future work in the use of probabilistic methods to assist in assessment of complex transcriptomes is essential. Overall, the use of Bayesian methods to provide continually updated probabilities in complex systems tends to improve outcomes at many levels. There are several research directions that would continue to improve our understanding of complex transcriptomes:

1. Better systems biology at the level of bacterial communities, especially those occupying the human host.
2. Better isolation methods and depth of sequencing. As sequencing technology continues to decrease in cost, the coverage of single samples that can be achieved will only improve.
3. Expansion of Bayesian methods to analyze a large number of samples simultaneously.

Based on our preliminary results with this method, there are several improvements that could be made to the software in later versions:

1. Parsing of reads or alignments without the need for preprocessing of data.
2. Ability to test the performance of different graph structures.
3. Additional analysis of Bayesian network performance.
4. Graphical interface.

References

1. Alexander, R.J. and R.F. Raicht, *Purification of total RNA from human stool samples.* Dig Dis Sci, 1998. **43**(12): p. 2652-8.

2. Ahmed, F.E., et al., *Improved methods for extracting RNA from exfoliated human colonocytes in stool and RT-PCR analysis.* Dig Dis Sci, 2004. **49**(11-12): p. 1889-98.

3. Bennett, W.E., Jr., et al., *A method for isolating and analyzing human mRNA from newborn stool.* J Immunol Methods, 2009. **349**(1-2): p. 56-60.

4. Melnick, J.L., *Properties and classification of hepatitis A virus.* Vaccine, 1992. **10 Suppl 1**: p. S24-6.

5. Holland, J.J., *Enterovirus Entrance into Specific Host Cells, and Subsequent Alterations of Cell Protein and Nucleic Acid Synthesis.* Bacteriol Rev, 1964. **28**: p. 2-13.

6. McNulty, M.S., *Rotaviruses.* J Gen Virol, 1978. **40**(1): p. 1-18.

7. Holtz, L.R., et al., *Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea.* Virol J, 2008. **5**: p. 159.

8. Nakamura, S., et al., *Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach.* PLoS One, 2009. **4**(1): p. e4219.

9. Sun, G., et al., *Viral metagenomics analysis of picobirnavirus-positive feces from children with sporadic diarrhea in China.* Arch Virol, 2016. **161**(4): p. 971-5.

10. Perez-Brocal, V., et al., *Metagenomic Analysis of Crohn's Disease Patients Identifies Changes in the Virome and Microbiome Related to Disease Status and Therapy, and Detects Potential Interactions and Biomarkers.* Inflamm Bowel Dis, 2015. **21**(11): p. 2515-32.

11. Davidson, L.A., et al., *Noninvasive detection of putative biomarkers for colon cancer using fecal messenger RNA.* Cancer Epidemiol Biomarkers Prev, 1995. **4**(6): p. 643-7.

12. Davidson, L.A., et al., *Non-invasive detection of fecal protein kinase C betaII and zeta messenger RNA: putative biomarkers for colon cancer.* Carcinogenesis, 1998. **19**(2): p. 253-7.

13. Davidson, L.A., et al., *Quantification of human intestinal gene expression profiles using exfoliated colonocytes: a pilot study.* Biomarkers, 2003. **8**(1): p. 51-61.

14. Kanaoka, S., et al., *Potential usefulness of detecting cyclooxygenase 2 messenger RNA in feces for colorectal cancer screening.* Gastroenterology, 2004. **127**(2): p. 422-7.

15. Takai, T., et al., *Fecal cyclooxygenase 2 plus matrix metalloproteinase 7 mRNA assays as a marker for colorectal cancer screening.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(6): p. 1888-93.

16. Hamaya, Y., et al., *Factors that contribute to faecal cyclooxygenase-2 mRNA expression in subjects with colorectal cancer.*

Br J Cancer, 2010. **102**(5): p. 916-21.

17. Leung, W.K., et al., *Detection of hypermethylated DNA or cyclooxygenase-2 messenger RNA in fecal samples of patients with colorectal cancer or polyps.* Am J Gastroenterol, 2007. **102**(5): p. 1070-6.

18. Kato, I., et al., *DNA/RNA markers for colorectal cancer risk in preserved stool specimens: a pilot study.* Tumori, 2009. **95**(6): p. 753-61.

19. Yamao, T., et al., *Abnormal expression of CD44 variants in the exfoliated cells in the feces of patients with colorectal cancer.* Gastroenterology, 1998. **114**(6): p. 1196-205.

20. Yang, S.H., et al., *Fecal RNA detection of cytokeratin 19 and ribosomal protein L19 for colorectal cancer.* Hepatogastroenterology, 2010. **57**(101): p. 710-5.

21. Koga, Y., et al., *Detection of colorectal cancer cells from feces using quantitative real-time RT-PCR for colorectal cancer diagnosis.* Cancer Sci, 2008. **99**(10): p. 1977-83.

22. Zhao, C., et al., *Noninvasive detection of candidate molecular biomarkers in subjects with a history of insulin resistance and colorectal adenomas.* Cancer Prev Res (Phila), 2009. **2**(6): p. 590-7.

23. Ahmed, F.E., et al., *Diagnostic microRNA markers for screening sporadic human colon cancer and active ulcerative colitis in stool and tissue.* Cancer Genomics Proteomics, 2009. **6**(5): p. 281-95.

24. Ahmed, F.E., *miRNA as markers for the diagnostic screening of colon cancer.* Expert Rev Anticancer Ther, 2014. **14**(4): p. 463-85.

25. Dickinson, B.T., et al., *Molecular markers for colorectal cancer screening.* Gut, 2015. **64**(9): p. 1485-94.

26. Beaulieu, J.F., et al., *Use of integrin alpha 6 transcripts in a stool mRNA assay for the detection of colorectal cancers at curable stages.* Oncotarget, 2016.

27. Fitzsimons, N.A., et al., *Bacterial gene expression detected in human faeces by reverse transcription-PCR.* J Microbiol Methods, 2003. **55**(1): p. 133-40.

28. Larocque, R.C., et al., *Transcriptional profiling of Vibrio cholerae recovered directly from patient specimens during early and late stages of human infection.* Infect Immun, 2005. **73**(8): p. 4488-93.

29. Rashid, R.A., et al., *Expression of putative virulence factors of Escherichia coli O157:H7 differs in bovine and human infections.* Infect Immun, 2006. **74**(7): p. 4142-8.

30. Dunaev, T., S. Alanya, and M. Duran, *Use of RNA-based genotypic approaches for quantification of viable but non-culturable Salmonella sp. in biosolids.* Water Sci Technol, 2008. **58**(9): p. 1823-8.

31. Chakrabarti, A.K., et al., *Detection of HIV-1 RNA/DNA and CD4 mRNA in feces and urine from chronic HIV-1 infected subjects with and without anti-retroviral therapy.* AIDS Res Ther, 2009. **6**: p. 20.

32. Kaeffer, B., et al., *Recovery of exfoliated cells from the gastrointestinal tract of premature infants: a new tool to perform "noninvasive biopsies?".* Pediatr Res, 2007. **62**(5): p. 564-9.

33. Bennett, W.E., Jr., et al., *Proinflammatory fecal mRNA and childhood bacterial infections.* Gut Microbes, 2010.

1(4): p. 209-212.

34. El Feghaly, R.E., et al., *Intestinal inflammatory biomarkers and outcome in pediatric Clostridium difficile infections.* J Pediatr, 2013. **163**(6): p. 1697-1704 e2.

35. Chen, H., et al., *Culture-independent analysis of fecal enterobacteria in environmental samples by single-cell mRNA profiling.* Appl Environ Microbiol, 2004. **70**(8): p. 4432-9.

36. Urich, T., et al., *Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome.* PLoS One, 2008. **3**(6): p. e2527.

37. Bailly, J., et al., *Soil eukaryotic functional diversity, a metatranscriptomic approach.* ISME J, 2007. **1**(7): p. 632-42.

38. Shrestha, P.M., et al., *Transcriptional activity of paddy soil bacterial communities.* Environ Microbiol, 2009. **11**(4): p. 960-70.

39. Poretsky, R.S., et al., *Analysis of microbial gene transcripts in environmental samples.* Appl Environ Microbiol, 2005. **71**(7): p. 4121-6.

40. Frias-Lopez, J., et al., *Microbial community gene expression in ocean surface waters.* Proc Natl Acad Sci U S A, 2008. **105**(10): p. 3805-10.

41. Gilbert, J.A., et al., *Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities.* PLoS One, 2008. **3**(8): p. e3042.

42. Hollibaugh, J.T., et al., *Metatranscriptomic analysis of ammonia-oxidizing organisms in an estuarine bacterioplankton assemblage.* ISME J, 2011. **5**(5): p. 866-78.

43. Nam, Y.D., et al., *Metatranscriptome analysis of lactic acid bacteria during kimchi fermentation with genome-probing microarrays.* Int J Food Microbiol, 2009. **130**(2): p. 140-6.

44. Weckx, S., et al., *Metatranscriptome analysis for insight into whole-ecosystem gene expression during spontaneous wheat and spelt sourdough fermentations.* Appl Environ Microbiol, 2011. **77**(2): p. 618-26.

45. Klaassens, E.S., et al., *Mixed-species genomic microarray analysis of fecal samples reveals differential transcriptional responses of bifidobacteria in breast- and formula-fed infants.* Appl Environ Microbiol, 2009. **75**(9): p. 2668-76.

46. Chapkin, R.S., et al., *Noninvasive stool-based detection of infant gastrointestinal development using gene expression profiles from exfoliated epithelial cells.* Am J Physiol Gastrointest Liver Physiol, 2010. **298**(5): p. G582-9.

47. Booijink, C.C., et al., *Metatranscriptome analysis of the human fecal microbiota reveals subject-specific expression profiles, with genes encoding proteins involved in carbohydrate metabolism being dominantly expressed.* Appl Environ Microbiol, 2010. **76**(16): p. 5533-40.

48. Turnbaugh, P.J., et al., *Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins.* Proc Natl Acad Sci U S A, 2010. **107**(16): p. 7503-8.

49. Knight, J.M., et al., *Non-invasive analysis of intestinal development in preterm and term infants using RNA-Sequencing.* Sci Rep, 2014. **4**: p. 5453.

50. La Rosa, P.S., et al., *Patterned progression of bacterial populations in the premature infant gut.* Proc Natl Acad Sci U S A, 2014. **111**(34): p. 12522-7.

51. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

52. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

53. Pruesse, E., et al., *SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.* Nucleic Acids Res, 2007. **35**(21): p. 7188-96.

54. Cao, M.D., et al. *A simple statistical algorithm for biological sequence compression*. in *Data Compression Conference, 2007. DCC'07*. 2007. IEEE.

55. Guo, F., et al., *Taxonomic precision of different hypervariable regions of 16S rRNA gene and annotation methods for functional bacterial groups in biological wastewater treatment.* PloS one, 2013. **8**(10): p. e76185.

56. Huson, D.H., et al., *Integrative analysis of environmental sequences using MEGAN4.* Genome Res, 2011. **21**(9): p. 1552-60.

57. Le Bihan, G., et al., *Transcriptome analysis of Escherichia coli O157:H7 grown in vitro in the sterile-filtrated cecal content of human gut microbiota associated rats reveals an adaptive expression of metabolic and virulence genes.* Microbes Infect, 2015. **17**(1): p. 23-33.

58. Bruchmann, S., et al., *Deep transcriptome profiling of clinical Klebsiella pneumoniae isolates reveals strain and sequence type-specific adaptation.* Environ Microbiol, 2015. **17**(11): p. 4690-710.

Vita

William E. Bennett, Jr.

Degrees

M.D., Univ. of Texas Southwestern Medical School, 2005
B.S., Biology and Biochemistry, Baylor University, 2000

Professional Societies

American Academy of Pediatrics
North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition
Society for Medical Decision Making
American Medical Informatics Association

Publications

Roth J, Keenan A, Carroll AE, Cain MP, Whittam BW, **Bennett WE Jr**. Readmission Characteristics of Elective Pediatric Circumcisions Using Large-Scale Administrative Data. J Pediatr Urology. 2015 Nov 14. pii: S1477-5131(15)00398-8.

Benneyworth BD, **Bennett WE Jr**, Carroll AE. Cross-sectional comparison of critically ill pediatric patients across hospitals with various levels of pediatric care. BMC Res Notes. 2015 Nov 19;8(1):693. doi: 10.1186/s13104-015-1550-9.

Bennett WE Jr. Quantitative Risk-Benefit Analysis of Probiotic Use for Irritable Bowel Syndrome and Inflammatory Bowel Disease. Drug Safety. 2015 Oct 14.

Jones PM, Rosenman MB, Pfefferkorn MD, Rescorla FJ, **Bennett WE Jr**. "Gallbladder Ejection Fraction is Unrelated to Gallbladder Pathology in Children and Adolescents." J Pediatr Gastroenterol Nutr. 2015 Dec 14.

LaRosa PS, Warner BB, Zhou Y, Weinstock GM, Sodergren E, Hall-Moore CM, Stevens HJ, **Bennett WE Jr**, Shaikh N, Linneman LA, Hoffmann JA, Hamvas A, Deych E, Shands

BA, Shannon WD, Tarr PI. "The Patterned Progression of Bacterial Populations in the Premature Infant Gut." Proc Natl Acad Sci U S A. 2014 Aug 26;111(34):12522-7

Bennett WE Jr, Hendrix KS, Thompson-Fleming RT, Carroll AE, Downs SM. "The Natural History of Weight Percentile Changes in the First Year of Life." JAMA Pediatr. 2014 Jul 1; 168(7).

Papic JC, Finnell SM, Leys CM, **Bennett WE Jr**, Downs SM. "Referring Physicians Decision-Making for Anti-Reflux Procedures". Surgery. 2014 May; 155(5):851-9.

Bennett WE Jr, Hendrix KS, Thompson-Fleming RT, Downs SM, Carroll AE. "Early Cow's Milk Introduction is Associated with Failed Personal-Social Milestones After One Year of Age." Eur J Pediatr. 2014 Jul; 173(7): 887-92.

Thompson RT, **Bennett WE Jr**, Finnell SM, Downs SM, Carroll AE. "Increased Length of Stay and Costs Associated with Weekend Admissions for Failure to Thrive." Pediatrics. 2013 Mar; 131(3):e805-10.

Bennett WE Jr, Heuckeroth RO. "Hypothyroidism is a rare cause of isolated constipation: 5-year review of all thyroid tests in a pediatric gastroenterology office." J Pediatr Gastroenterol Nutr. 2012 Feb;54(2):285-7.

Bennett WE Jr, Gonzalez-Rivera R, Puente, BN, Shaikh N, Stevens HJ, Mooney JC, Klein EJ, Draghi A II, Sylvester FA, Tarr PI. "Proinflammatory fecal mRNA and childhood bacterial enteric infections." Gut Microbes. 2010 July/Aug. 1(4). PMCID: PMC3023602

Bennett WE Jr, González-Rivera R, Shaikh N, Magrini V, Boykin M, Warner BB, Hamvas A, Tarr PI. "A method for isolating and analyzing human mRNA from newborn stool." J Immunol Methods. 2009 Sep 30;349(1-2):56-60. PMCID: PMC2850193