Hindawi
International Journal of Genomics
Volume 2019, Article ID 1718741, 9 pages
https://doi.org/10.1155/2019/1718741

Research Article

A Robust Metatranscriptomic Technology for Population-Scale Studies of Diet, Gut Microbiome, and Human Health

Andrew Hatch, James Horne, Ryan Toma, Brittany L. Twibell, Kalie M. Somerville, Benjamin Pelle, Kinga P. Canfield, Matvey Genkin, Guruduth Banavar, Ally Perlina, Helen Messier, Niels Klitgord, and Momchilo Vuyisich

Viome Inc., Los Alamos, NM 87544, USA

Correspondence should be addressed to Momchilo Vuyisich; vuyisich@hotmail.com

Received 23 October 2018; Revised 17 May 2019; Accepted 6 June 2019; Published 1 October 2019

Academic Editor: Ferenc Olasz

A functional readout of the gut microbiome is necessary to enable precise control of the gut microbiome's functions, which support human health and prevent or minimize a wide range of chronic diseases. Stool metatranscriptomic analysis offers a comprehensive functional view of the gut microbiome, but despite its usefulness, it has rarely been used in clinical studies due to its complexity, cost, and bioinformatic challenges. This method has also received criticism due to potential intrasample variability, rapid changes, and RNA degradation. Here, we describe a robust and automated stool metatranscriptomic method, called Viomega, which was specifically developed for population-scale studies. Viomega includes sample collection, ambient temperature sample preservation, total RNA extraction, physical removal of ribosomal RNAs (rRNAs), preparation of directional Illumina libraries, Illumina sequencing, taxonomic classification based on a database of >110,000 microbial genomes, and quantitative microbial gene expression analysis using a database of ~100 million microbial genes. We applied this method to 10,000 human stool samples and performed several small-scale studies to demonstrate sample stability and consistency. In summary, Viomega is an inexpensive, high-throughput, automated, and accurate sample-to-result stool metatranscriptomic technology platform for large-scale studies and a wide range of applications.

1. Introduction

The human gut contains a vast number of commensal microorganisms performing a wide variety of metabolic functions. Metabolites produced by these microorganisms can have profound effects on human physiology, with direct links to health and disease status [1–4]. Gut dysbiosis likely contributes to the development and progression of many diseases and disorders, such as cardiovascular disease, hypertension, obesity, diabetes, and autoimmune diseases [5–9]. There is also strong evidence that the gut microorganisms directly interact with the nervous system, establishing the gut-brain axis [10]. The gut-brain axis has been shown to modulate the development of neurodegenerative diseases such as Alzheimer's disease, Autism Spectrum Disorder (ASD) and Parkinson's disease [11–14].

The gut microbiome plays a critical role in physiological homeostasis, resulting in increasing scientific investigation into the extent of the gut microbiome's role in human health and disease. Humans have coevolved with the microbiome and have become dependent on its biochemical output, such as certain vitamins and short-chain fatty acids [15, 16]. The gut microbiome can also produce harmful biochemicals that have been implicated in various disease states [15]. To fully understand the relationships between the gut microbiome and human health status, biochemical functions of the microorganisms must be identified and quantified. Several next generation sequencing-based methods have been used for analyzing the gut microbiome, each with clear advantages and disadvantages. The simplest, least expensive, and most common method is 16S rRNA gene sequencing [17], which sequences a small portion of the

highly conserved prokaryotic 16S ribosomal RNA gene [18]. This method can provide taxonomic resolution to the genus level [19, 20], but it does not measure the biochemical functions of the microorganisms [18] or distinguish living from dead organisms. In addition, traditional 16S rRNA sequencing excludes some bacteria, most archaea, and all eukaryotic organisms and viruses [21], resulting in a limited view of the gut microbiome ecosystem.

Metagenomic (shotgun DNA) sequencing provides strain-level resolution of all DNA-based microorganisms [18] (it does not detect RNA viruses or RNA bacteriophages). However, it can only identify the *potential* biochemical functions of the microbiome and can neither identify nor quantify the active biochemical pathways. This is a disadvantage for studying dysbiosis-related disease states, such as inflammatory bowel disease (IBD), which has been shown to have a disparity between metagenomic potential pathways and actual biochemical pathways expressed in disease and control populations [22].

Metatranscriptomic analysis (metatranscriptomics, RNA sequencing, and RNAseq) offers insights into the biochemical activities of the gut microbiome by quantifying expression levels of active microbial genes, allowing for the assessment of pathway activities, while also providing strain-level taxonomic resolution for all metabolically active organisms and viruses [23, 24]. To date, metatranscriptomic analyses of stool samples have been limited due to the cost and complexity of both laboratory and bioinformatic methods [25]. By removing less informative rRNA, more valuable transcriptome data can be generated with less sequencing depth [24, 26], resulting in reduced per-sample sequencing costs.

An automated technology has been developed for metatranscriptomic analysis of human clinical samples, called Viomega. In this study, Viomega was applied to 10,000 human stool samples to gain a better understanding of the strain-level taxonomies and microbial functions. Several small-scale studies were performed to quantify the metatranscriptomic stability in the lower colon and measure the intrasample variability of metatranscriptomic analyses.

2. Materials and Methods

2.1. Study Participants, Ethics, and Sample Collection and Transportation. For this study, Viome used data from 10,000 participants. All study participants gave consent to being in the study, and all study procedures were approved by a federally accredited Institutional Review Board (IRB). Participants were recruited from any age, gender, and ethnic group.

Stool samples were collected using Viome's Gut Intelligence kit by each study participant at their own residences. The kit included a sample collection tube with an integrated scoop, a proprietary RNA preservative, and sterile glass beads. A pea-sized stool sample was collected and placed inside the tube and vigorously shaken to homogenize the sample, exposing it to the RNA preservative. The sample was then shipped at room temperature using a common courier to Viome labs for analysis. Shipping times ranged from one to twelve days. Each participant completed a questionnaire with general lifestyle and health information.

2.2. Metatranscriptomic Analysis of Stool Samples. For the metatranscriptomic analysis of 10,000 stool samples, a proprietary sample-to-result automated platform called Viomega was used. Stool samples were lysed using bead beating in a strong chemical denaturant and then placed on an automated liquid handler, which performed all downstream laboratory methods. Samples were processed in batches in a 96-well microplate; each batch consisted of ninety-four human stool samples, a negative process control (NPC, water), and a positive process control (PPC, custom synthetic RNA). RNA was extracted using a proprietary method. Briefly, silica-coated beads and a series of washes were used to purify RNA after lysis and RNA was eluted in water. DNA was degraded using RNase-free DNase.

The majority of prokaryotic ribosomal RNAs (rRNAs: 16S and 23S) were removed using a custom subtractive hybridization method. Biotinylated DNA probes with sequences complementary to rRNAs were added to total RNA, the mixture was heated and cooled, and the probe-rRNA complexes were removed using magnetic streptavidin beads. The remaining RNAs were converted to directional sequencing libraries with unique dual-barcoded adapters and ultrapure reagents. Libraries were pooled and quality controlled with dsDNA Qubit (Thermo Fisher Scientific) and Fragment Analyzer (Advanced Analytical). Library pools were sequenced on Illumina NextSeq or NovaSeq instruments using 300 cycle kits.

Viomega's bioinformatics module operates on Amazon Web Services and includes quality control, taxonomic profiling, and functional analysis. Quality control tools trim and filter the raw reads and quantify the amounts of the sample-to-sample cross-talk and background contamination by microbial taxa. Viomega generates read-based taxonomy assignments using a multistep process. The sequencing reads are aligned to a proprietary database of precomputed genomic signatures at three taxonomic levels: strain, species, and genus. The unique signatures are computed from full-length genomes by removing short subsequences of a defined length, k (k-mers), shared among more than one genome and keeping unique k-mers that make up the signature [27]. The Viomega taxonomy database was generated from a large RefSeq database containing more than 110,000 microbial genomes. After the initial taxonomic assignments were generated, potential false positives were removed using the Auto-Blast algorithm that uses an even larger database of organisms.

The identity and relative activity of microbial genes and enzymatic functions in the stool samples were assessed using a proprietary algorithm. At a high level, this involves a multitiered approach to align the sample reads to the integrated gene catalog (IGC) [28] library of genes to first identify and then quantify the genes in the sample. Informative genes (i.e., non-rRNA) were quantified in units of transcripts per million (TPM) to allow for cross-sample comparisons. Using the Kyoto Encyclopedia of Genes and Genomes (KEGG) [29] annotation mapping of IGC genes to KEGG orthologies

(KOs), the enzymic functions and activity were quantified in these samples as the aggregate TPM. The KEGG mapping also allows for functional modules and pathway analysis.

2.3. Small-Scale Studies. For the validation of the sample lysis in the Viomega pipeline, the following organisms were grown in nutrient broth at 37°C and 450 rpm in a VWR incubating minishaker: *Bacillus subtilis* Marburg strain (ATCC 6051-U), *Corynebacterium stationis* strain NCTC 2399 (ATCC 6872), *Citrobacter freundii* strain ATCC 13316, NCTC 9750 (ATCC 8090), and *Serratia liquefaciens* strain CDC 1284-57 ATCC 12926 (ATCC 27592). In addition, the following organisms were grown in yeast mold broth at 37°C and 450 rpm in a VWR incubating minishaker: *Saccharomyces cerevisiae* strain S288C (ATCC 204508) and *Candida dubliniensis* strain CBS 7987 (ATCC MYA-646).

To illustrate the accuracy of taxonomic classification at the species level of the Viomega technology, the 10 Strain Even Mix Whole Cell Material (ATCC® MSA-2003™) product was utilized. As stated by the manufacturer, this product is comprised of an even mixture of the following organisms: *Bacillus cereus* (ATCC 10987), *Bifidobacterium adolescentis* (ATCC 15703), *Clostridium beijerinckii* (ATCC 35702), *Deinococcus radiodurans* (ATCC BAA-816), *Enterococcus faecalis* (ATCC 47077), *Escherichia coli* (ATCC 700926), *Lactobacillus gasseri* (ATCC 33323), *Rhodobacter sphaeroides* (ATCC 17029), *Staphylococcus epidermidis* (ATCC 12228), and *Streptococcus mutans* (ATCC 700610).

3. Results and Discussion

3.1. Validation of Viomega Technology

3.1.1. Sample Lysis. Uneven sample lysis can introduce major errors in any method since sample composition can vary widely in terms of easy-to-lyse microorganisms (viruses and Gram(-) bacteria) and difficult to very-difficult-to-lyse Gram(+) bacteria and yeast. Viomega utilizes a combination of chemical (denaturant) and physical (bead beating) sample lyses, which has been shown to have the best efficiency. To test this method, two strains of Gram(-) bacteria, two strains of Gram(+) bacteria, and two strains of yeast were grown to an optical density range of 0.4-0.8 AU. Equal amounts of each organism in triplicate then underwent chemical and physical sample lyses, and total RNA was extracted from each sample. RNA yields obtained were consistent and show no bias against Gram(+) bacteria or yeast (average yield: Gram$(-)$ = 93.3 ng/μL, 83.6 ng/μL; Gram$(+)$ = 131.9 ng/μL, 152.3 ng/μL; yeast = 113.1 ng/μL, 100.8 ng/μL) (Figure 1). The process was also reproducible, with a very small variability across technical replicates (standard deviation range = 3.4-11.0 ng/μL).

3.1.2. Ambient Temperature Sample Transportation. A noted shortcoming of metatranscriptomics is that it analyzes labile RNA molecules. This is most apparent in the case of dead organisms, as the existing RNA is rapidly degraded, while no new transcripts are made. In living organisms, however, RNA is continuously made and degraded. By exposing living organisms to appropriate reagents, this dynamic equilibrium

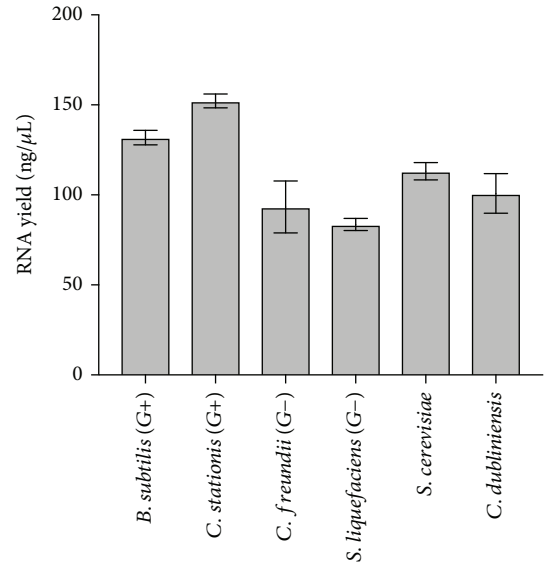

Figure 1: Average RNA yields (ng/μL) of model organisms after sample lysis and RNA extraction (*B. subtilis*: 131.9 ± 4.0 ng/μL; *C. stationis*: 152.3 ± 3.9 ng/μL; *C. fruendii*: 93.3 ± 14.4 ng/μL; *S. liquefaciens*: 83.6 ± 3.4 ng/μL; *S. cerevisiae*: 113.1 ± 4.8 ng/μL; *C. dubliniensis*: 100.8 ± 11.0 ng/μL; $n = 3$).

of gene expression can be "frozen in time" at the time of sample collection and quantitatively analyzed later. To achieve this, Viomega uses a chemical denaturant/RNA stabilizing solution that ensures the preservation of RNA integrity during sample transport at ambient temperatures. Fourteen aliquots were made from a single donor sample; four aliquots were processed using Viomega immediately, while three samples were stored at room temperature (RT) for four weeks prior to processing. Seven aliquots were shipped through a standard courier and held on-site at the laboratory for a total time of four weeks prior to processing. All comparisons show very strong correlation with a Spearman correlation value of 0.8 or greater (Figure 2) [30]. No difference was found in taxonomic profiling or functional composition between time to processing or shipping conditions prior to processing (Figure 2).

3.1.3. Sample-to-Sample Cross-Talk (STSC). STSC (also known as barcode switching, barcode hopping, or read misassignment) can cause significant errors when sequencing many samples on a single sequencing run. Standard library preparation methods for sample barcoding have high error rates, from 0.2 to 5%, especially on the newest generation of Illumina platforms that use ExAmp technology [31, 32]. This phenomenon can cause errors in the reported taxonomies, *e.g.*, abundant taxa in a sequencing run being assigned to samples in which those taxa did not exist. Viomega minimizes STSC by a combination of specially produced barcode oligos, dual unique barcode sequences of 11 bps each, and only reporting the taxa that did not exceed the rate of measured STSC on each batch of 96 samples. STSC was quantified by introducing a synthetic, nonnatural RNA sample (PPC) in each microplate and measuring its quantity in each

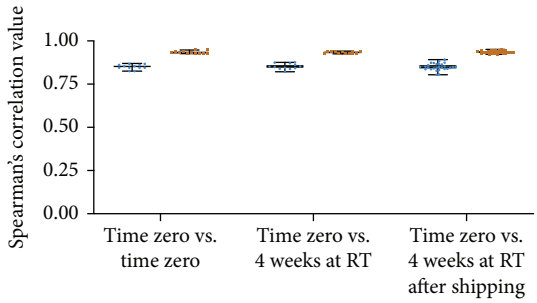

Figure 2: Spearman's correlation of taxonomic data (blue) and functional data (orange) between and within storage conditions. Median correlation of taxonomic data: T_0 vs. $T_0 = 0.8493$; T_0 vs. $T_{4\text{weeksRT}} = 0.8503$; T_0 vs. $T_{4\text{weeksSHIP}} = 0.8493$. Median correlation values of functional data: T_0 vs. $T_0 = 0.932$; T_0 vs. $T_{4\text{weeksRT}} = 0.9341$; T_0 vs. $T_{4\text{weeksSHIP}} = 0.9349$.

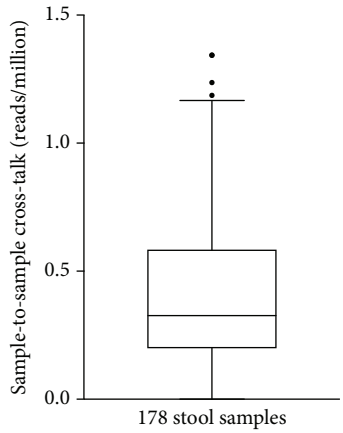

Figure 3: Sample-to-sample cross-talk (reads per million) for 178 stool samples sequenced on the Illumina NovaSeq platform using Viomega dual unique barcode sequences. Sample-to-sample cross-talk determined by measuring the occurrence of nonnatural PPC reads in each stool sample. Range = 0-1.34 reads per million; median = 0.33 reads per million.

of the other samples. The PPC sample was randomly positioned in each plate. STSC in Viomega mostly falls under 1 read per million reads (0.0001%) (Figure 3) on the NovaSeq platform (S1 flow cell, 300 cycle kits), which is more than 1,000-fold lower than in data obtained using commercial library preparation kits [31–33].

3.1.4. Background Contamination of Samples. Since any metagenomic or metatranscriptomic analysis identifies all taxa in a sample, nucleic acid contamination of the reagents (especially purification kits and enzymes), instruments, and poor laboratory practices can lead to the inclusion of contaminating taxa into scientific results [34–36]. To minimize this, Viomega uses ultrapure reagents, good laboratory practices, and fully automated liquid handling systems. Every plate of ninety-six samples contained a positive process control (PPC) sample, which is a synthetic RNA that was subjected to the same process as the rest of the samples (from kit manufacturing to bioinformatics). This sample was sequenced and analyzed like all other samples on the plate, allowing any microbial contamination to be detected. Over the course of twenty consecutive batches (1,880 stool samples, 20 PPC samples), the level of background contamination observed in PPC samples was extraordinarily low, with an average of 1.4 contaminating reads (std. dev. 2.6; $n = 20$) out of 5-15 million sequencing reads and 0.3 contaminating taxa (std. dev. 0.5; $n = 20$). Across twenty batches, the number of contaminating taxa was either zero or one, with a maximum of ten sequencing reads (out of an average of ~10 million) assigned to the taxon. These values were below the threshold for reporting any microorganism from the Viomega analysis and therefore do not cause any false positives to the results.

3.1.5. Depletion of Ribosomal RNAs. The vast majority of RNA molecules in any biological sample are ribosomal RNAs (rRNA). Approximately 96% of all reads from stool samples align to microbial rRNAs (Table S1), leaving only ~4% of sequencing data aligning to microbial messenger RNAs. Since rRNA sequences are not very informative (housekeeping functions, poor taxonomic resolution), and a key goal of Viomega is to deeply probe the functional landscape of the gut microbiome (*i.e.*, quantify the messenger RNAs), a subtractive hybridization method for rRNA depletion has been implemented in the Viomega process. This fully automated method reduces rRNA to $60.4 \pm 14.9\%$ ($n = 90$), thus providing an average enrichment of microbial messenger RNAs of ~10-fold by increasing sequencing data aligning to microbial messenger RNAs to ~40%.

3.2. Viomega: Accuracy of Taxonomic Classification. Given the large amounts of metatranscriptomic data obtained from each sample (over one giga-base pairs) and a very large database (more than 110,000 genomes), it is extremely challenging to have a high-throughput, fully automated, cost-effective, cloud-based, and highly accurate bioinformatic pipeline. Viomega is a fully automated cloud application whose efficiency comes from using a precomputed database of microbial signatures. This approach reduces the amount of searchable sequence space by roughly two orders of magnitude and largely eliminates false positive results. Viomega technology was used to analyze a commercially available mock community (10 Strain Even Mix Whole Cell Material, ATCC® MSA-2003™). For identification at the species level, Viomega shows 100% accuracy consistent with the mock community, with no false positive or false negative calls (Figure 4). The whole cell relative abundance of these microorganisms was reported by the supplier as identical. However, the relative RNA amounts (measured as relative activity by Viomega) may not be the same due to potential differences in how each monoculture was grown, processed, and stored prior to the preparation of the mock community. It is also possible that *Deinococcus radiodurans* contains more RNA per cell than other bacteria, due to its diploid genome. The somewhat higher relative abundance cannot

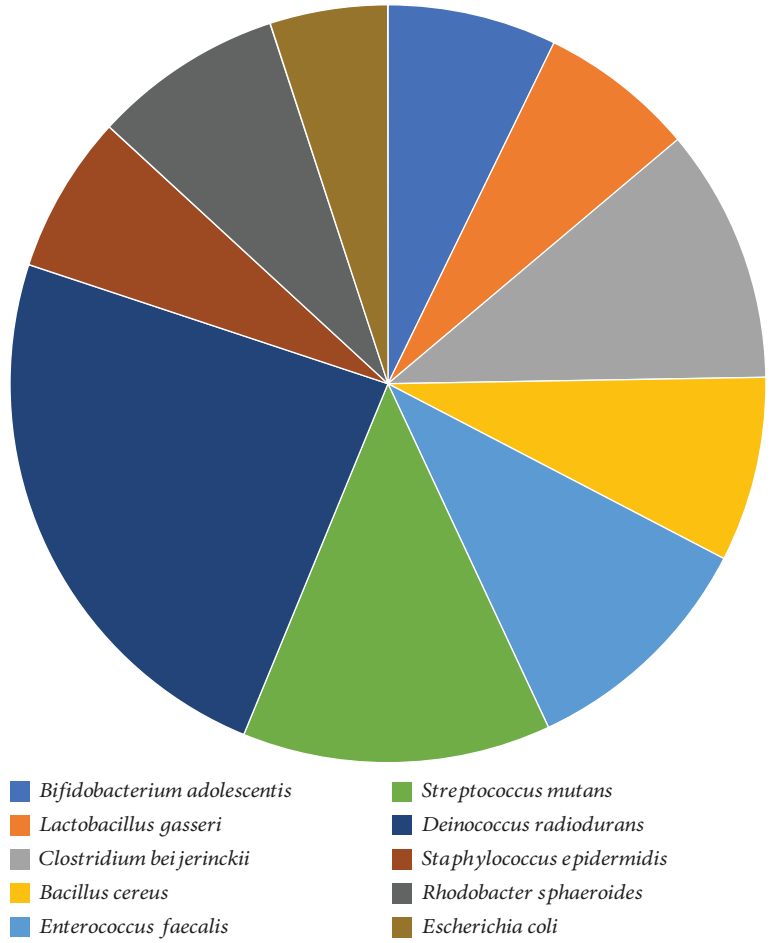

Figure 4: Accuracy and relative abundance of Viomega analysis on a commercial mock community. Species contents of the mock community as listed by the supplier are shown (left); supplier lists relative abundance of whole cells as equal among all species. Viomega achieves 100% accuracy at species level identification.

be explained with facile lysis (it is a Gram-positive organism) or low GC content (67%).

3.3. Findings from the Viomega Taxonomic Classification. Using the Viomega taxonomy classification pipeline, a total of 2,723 microbial strains, 1,946 microbial species, and 528 microbial genera have been identified in 10,000 human stool samples. The identified microorganisms include bacteria, archaea, viruses, bacteriophages, and eukaryotes (Table 1).

3.4. Viomega: Quantification of Microbial Biochemical Functions. Using Viomega's functional analysis tools, the expression of >100,000 microbial open reading frames (ORFs), which were grouped into 6,879 KEGG functions, was identified and quantified from 10,000 human stool samples. The top ten KEGG functions are shown in Table 2.

3.5. Intrasample Variability of Metatranscriptomic Analyses. Because large-scale studies would preferably analyze a single stool sample (instead of an average of multiples), it was important to understand the variability of microbial taxonomy and functions across individual stool samples. To understand this variability, three volunteers (P11, P12, and P13) collected samples from three parts of their stool samples: (1) from one end, (2) from the opposite end, and (3) from the middle. Each biological sample was split into three technical replicates (a, b, and c). All samples were analyzed using Viomega, followed by unsupervised clustering analysis (Kendall's correlation). All biological and technical replicates from the same stool sample (in-group) clustered by participant with very high similarity, and were different from the outgroup samples, especially at the strain-level taxonomy (Figure 5). This ministudy shows high uniformity of the metatranscriptomic data across stool samples. While there have been claims of large intrasample variability [37], these were likely based on biased methods, and not real differences in microbial taxonomy [38]. For large-scale studies, it is cost prohibitive to collect and analyze multiple samples per collection time; Viomega metatranscriptomic analysis provides reproducible results across stool samples.

3.6. Short-Term (Minutes) Stability of Stool Metatranscriptomes. To identify any changes in the measured microbial taxonomy and functions in the first few minutes after a stool sample was produced, three participants (P12, P13, and P14) were asked to collect samples from the

Table 1: The top ten strains, species, and genera identified in 10,000 human stool samples, based on their prevalence. See supplementary materials for all taxa identified in 10,000 human stool samples (Table S2 for strain, Table S3 for species, Table S4 for genera).

Top 10 genera		Top 10 species		Top 10 strains	
Genus	Prevalence (%)	Species	Prevalence (%)	Strains	Prevalence (%)
Clostridium	99.7	*Bacteroides vulgatus*	97.1	*Eggerthella lenta* 1_1_60AFAA	97.1
Bacteroides	99.6	*Acinetobacter baumannii*	96.8	(*Eubacterium*) *hallii* DSM 3353	93.9
Blautia	97.6	*Faecalibacterium prausnitzii*	96.4	*Veillonella dispar* ATCC 17748	92.3
Acinetobacter	97.5	*Bacteroides uniformis*	95.4	*Anaerotruncus colihominis* DSM 17241	91.4
Eubacterium	97.2	*Eggerthella lenta*	91.8	*Clostridium phoceensis* strain GD3	90.8
Parabacteroides	96.9	(*Eubacterium*) *hallii*	91.5	*Blautia obeum* ATCC 29174	89.7
Lactococcus	96.9	*Anaerotruncus colihominis*	91.4	(*Eubacterium*) *eligens* strain 2789STDY5834875	88.9
Faecalibacterium	96.4	*Clostridium phoceensis*	90.8	*Faecalibacterium* cf. *prausnitzii* KLE1255	88.3
Roseburia	95.7	*Veillonella dispar*	89.0	*Faecalibacterium prausnitzii* A2-165	86.0
Alistipes	95.1	*Fusicatenibacter saccharivorans*	87.3	*Roseburia hominis* A2-183	83.4

Table 2: The top ten KEGG functions identified in 10,000 human stool samples. See supplementary material for the top 100 KEGG functions (Table S5).

KO ID	Name	Prevalence in 10,000 samples (%)
K00936	pdtaS	100.00
K01190	lacZ	99.99
K03046	rpoC	99.99
K02355	fusA, GFM, EFG	99.99
K00540	fqr	99.99
K03695	clpB	99.99
K03296	TC.HAE1	99.99
K01362	OVCH	99.98
K02358	tuf, TUFM	99.98
K06950	K06950	99.98

same stool (a) immediately, (b) three minutes later, and (c) ten minutes later. Unsupervised clustering analysis (Kendall's correlation) was performed on the nine samples, and all samples clustered with high similarity based on the sample, and not the time of collection (Figure 6).

3.7. Long-Term (Weeks) Stability of Stool Metatranscriptomes. Because gene expression can change rapidly due to environmental changes, a ministudy was performed to look for metatranscriptome changes in stool microbiome over time. Seven volunteers were asked to maintain their normal diet and lifestyle for two weeks. During this period, three stool samples were collected from each participant: time zero, one week later, and two weeks later. The twenty-one samples were analyzed with Viomega, and unsupervised clustering analysis (Kendall's correlation) was performed based on taxonomy and KEGG functions (Figure 7). For both taxonomy (Figure 7(a)) and KEGG functions (Figure 7(b)), the samples clustered by the participant, confirming that both gut microbiome composition and biochemical functions were stable over the course of the study while maintaining a consistent diet. These data clearly demonstrate the utility of Viomega technology, as the microbial metatranscriptome was maintained with a consistent diet over a period of weeks.

4. Conclusions

In this report, Viomega, a sample-to-result, automated, and robust stool metatranscriptomic analysis technology is described. Viomega includes at-home sample collection, stability at ambient temperatures during transport (for up to twenty-eight days), complete sample lysis, RNA extraction, physical removal of noninformative (ribosomal) RNAs, sequencing library preparation, Illumina sequencing, and a quantitative bioinformatic analysis platform that includes taxonomic classification and functional analysis. Almost all laboratory steps are performed in a 96-well format using automated liquid handlers. All bioinformatic analyses are automatically performed on cloud servers. Viomega includes several critically important quality control steps, both per sample (number of base pairs generated for microbial messenger RNAs (% rRNA) and sample-to-sample cross-talk) and per batch of ninety-six samples (background contamination, process control samples, RNA yields, etc.).

Using a commercial mock community, Viomega shows 100% accuracy (no false positives or negatives) at the species level. Since the ground truth for the RNA content of each member of the mock community cannot be obtained from the manufacturer, it is unclear whether the small differences in the relative abundance of the ten microorganisms provided are an artifact of the sample or the method of producing the mock community.

Viomega was applied to 10,000 human stool samples and identified several thousand taxa at the strain, species, and genus ranks. More than 100,000 open reading frames (ORFs) were identified, quantified, and grouped into thousands of KEGG functions. The large bioinformatic data outputs of Viomega are being used to learn how gut microbiome taxonomy and functions are affected by the diet, develop improved models of how to precisely control the gut microbiome using

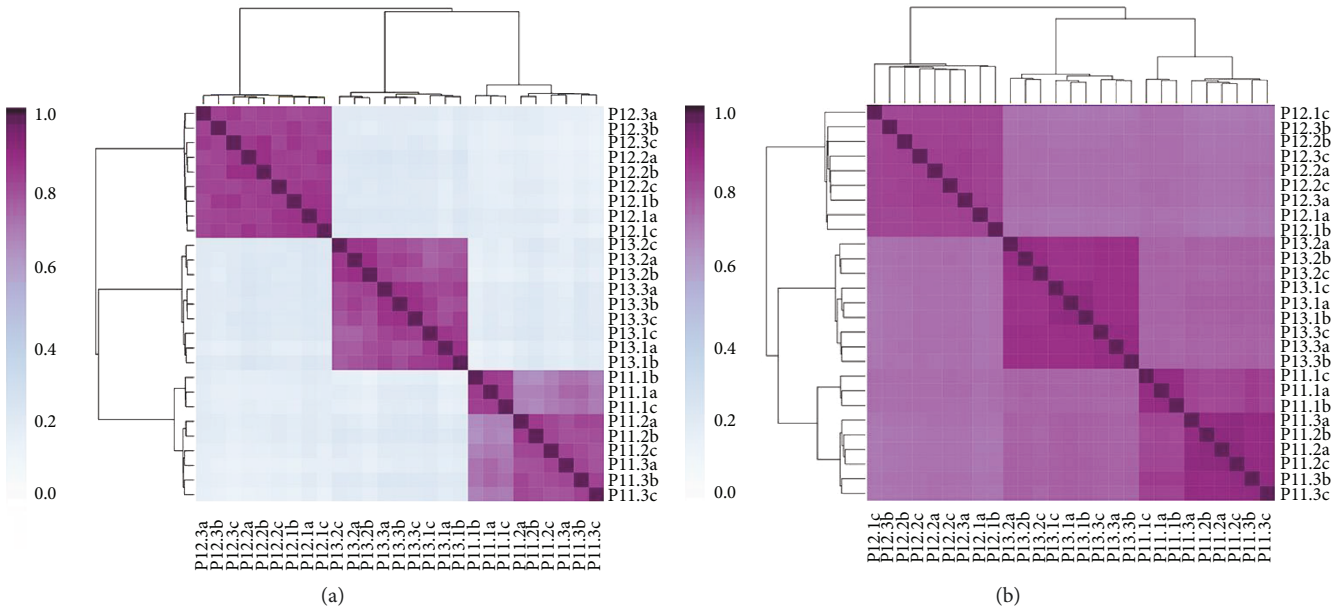

FIGURE 5: Intrasample variability of microbial taxonomy and functions using Viomega analysis. Three participants (P11, P12, and P13) provided three stool samples each, from three different parts of the stool (1, 2, and 3). Each biological sample was analyzed as three technical replicates (a, b, and c). Following Viomega analysis, unsupervised clustering analysis (Kendall's correlation) was performed on microbial taxonomy (at the strain level) (a) and biochemical functions at the KEGG level (b).

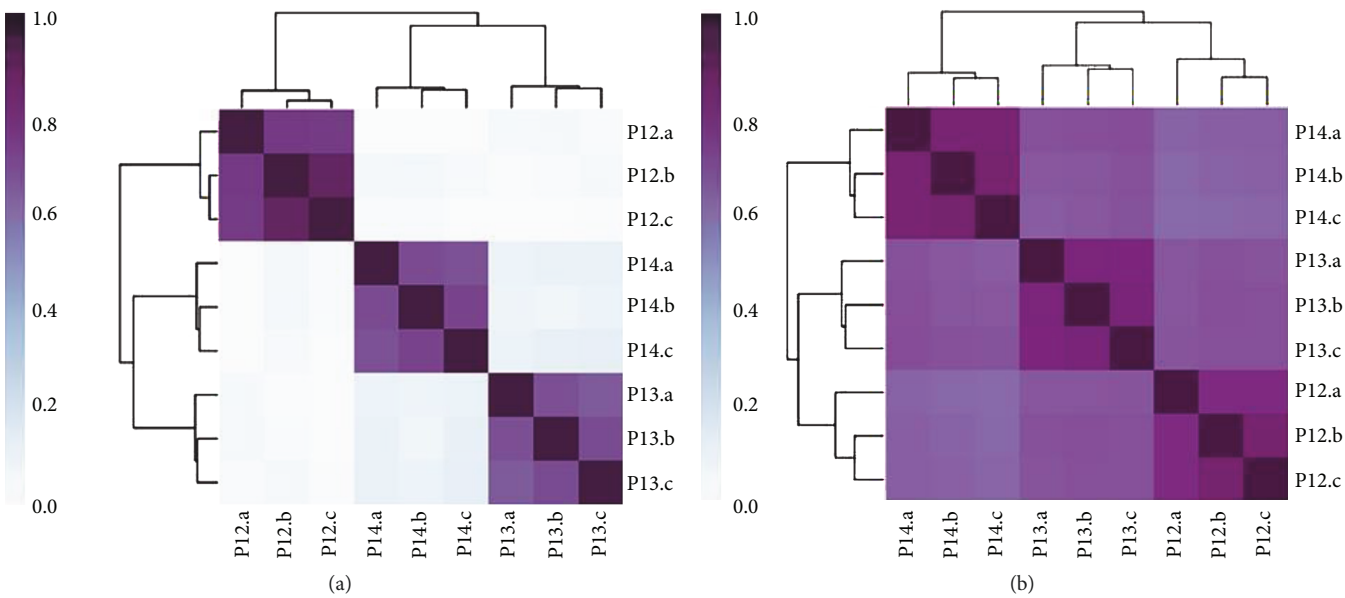

FIGURE 6: Stool samples were collected immediately and at three and ten minutes (a, b, and c, respectively) after the stool was produced by three participants (P12, P13, and P14). Unsupervised clustering analysis (Kendall's correlation) of samples shows a high similarity of strain-level taxonomy (a) and KEGG-based microbial functions (b) based on the donor, and not based on the time of collection.

diet, and learn how the gut microbiome correlates with human health and disease. These analyses will be described in upcoming publications. While Viomega was specifically designed for stool sample analysis, modifications may be made for alternative pipelines for other types of human clinical applications in the future.

Viomega was used to perform several small-scale studies to demonstrate the robustness of stool metatranscriptomic analysis when the methods introduce minimal biases. These studies show that it is possible to collect a single stool sample as representative of the entire colonic microbiome. The studies also establish that the gut metatranscriptome exhibits

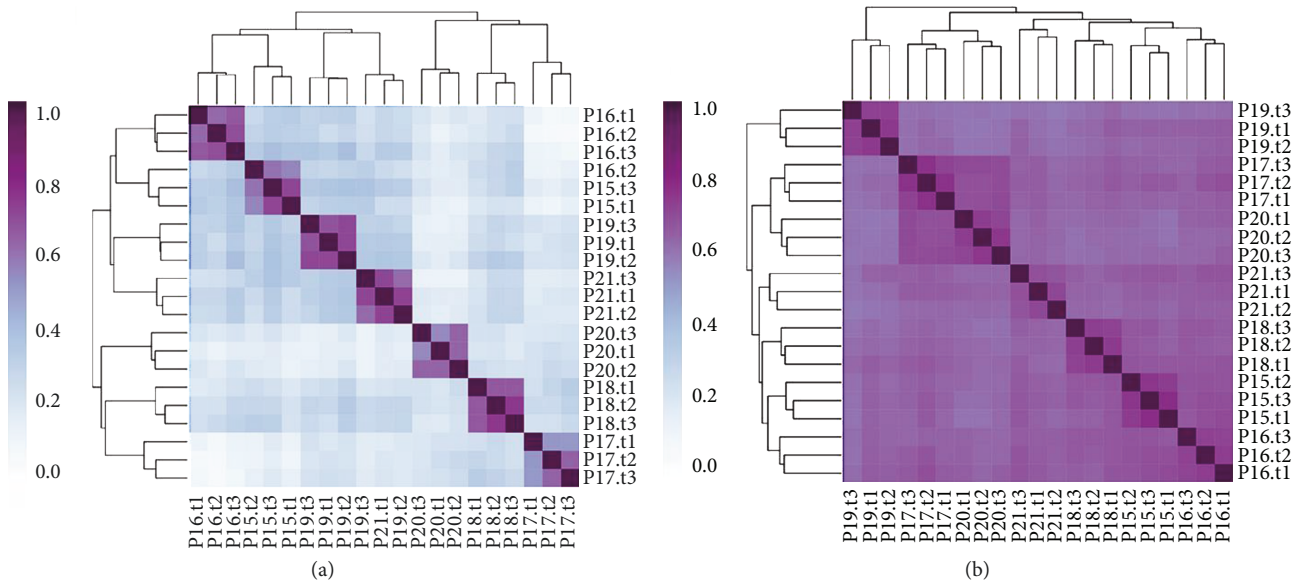

Figure 7: Unsupervised clustering analysis (Kendall's correlation) of gut microbiome samples collected from seven participants (P15-P21) at three time points two weeks apart (weeks 0, 1, and 2), while not changing the diet. (a) Image shows a high degree of clustering of microbial taxonomy (at the strain level) by person, longitudinally, with high in-group similarity. (b) Image shows a high degree of clustering of microbial functions (KEGG) by person, longitudinally, with high in-group similarity.

stability for several weeks without a diet change both compositionally and functionally. It should be noted that all participants involved in the above ministudies were self-reported healthy individuals. In each study, clustering by taxonomy showed much lower outgroup similarity than clustering by function. While taxonomy has been shown to vary from person to person among healthy individuals [39], the observed clustering patterns (Figures 5, 6, and 7) suggest similar functionality between healthy individuals; although different organisms are present, they are performing similar biochemical functions.

Viomega is a robust technology that offers a rapid and comprehensive taxonomic *and* functional readout of the gut microbiome. In addition, the cost to process a human stool sample through the Viomega pipeline ($199 for the Viome Gut Intelligence™ Test at the time of submission) is inexpensive compared to similar services ($15,000 for up to five samples through The Human Microbiome Project—"What are they actually doing" service) [40] largely due to batched processing, removal of rRNAs, and the unique Viomega taxonomy database. This technology will increase the overall understanding of the interplay among diet, gut microbiome, and human health, and is enabling gut microbiome-based personalized nutrition as an emerging field. These advances may fuel mitigation and treatment for a variety of human health conditions, such as cardiovascular disease, obesity, autoimmune disease, ASD, and Parkinson's disease.

Data Availability

The raw data used to support the findings of this study are not publicly available because they are proprietary to Viome Inc.

Conflicts of Interest

All authors are also employees of Viome Inc., which sponsored the study.

Acknowledgments

Viome Inc. funded these studies.

Supplementary Materials

Supplementary Table S1: data for the percent ribosomal RNA in stool samples that are processed through Viomega without the custom rRNA depletion method. Supplementary Tables S2–S4: all taxa identified by Viomega in 10,000 human stool samples (strains, species, and genera, respectively) are shown. Supplementary Table S5: the top 100 KEGG functions identified by Viomega in 10,000 human stool samples are shown. *(Supplementary Materials)*

References

[1] C. Nasca, B. Bigio, F. S. Lee et al., "Acetyl-l-carnitine deficiency in patients with major depressive disorder," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 34, pp. 8627–8632, 2018.

[2] H. J. Flint, "Gut microbial metabolites in health and disease," *Gut Microbes*, vol. 7, no. 3, pp. 187-188, 2016.

[3] T. S. Postler and S. Ghosh, "Understanding the Holobiont: how microbial metabolites affect human health and shape the immune system," *Cell Metabolism*, vol. 26, no. 1, pp. 110–130, 2017.

[4] J. C. Clemente, L. K. Ursell, L. W. Parfrey, and R. Knight, "The impact of the gut microbiota on human health: an integrative view," *Cell*, vol. 148, no. 6, pp. 1258–1270, 2012.

[5] J. Li, F. Zhao, Y. Wang et al., "Gut microbiota dysbiosis contributes to the development of hypertension," *Microbiome*, vol. 5, no. 1, 2017.

[6] P. J. Parekh, L. A. Balart, and D. A. Johnson, "The influence of the gut microbiome on obesity, metabolic syndrome and gastrointestinal disease," *Clinical and Translational Gastroenterology*, vol. 6, no. 6, p. e91, 2015.

[7] Y. Zhang and H. Zhang, "Microbiota associated with type 2 diabetes and its related complications," *Food Science and Human Wellness*, vol. 2, no. 3-4, pp. 167–172, 2013.

[8] F. H. Karlsson, F. Fåk, I. Nookaew et al., "Symptomatic atherosclerosis is associated with an altered gut metagenome," *Nature Communications*, vol. 3, no. 1, article 1245, 2012.

[9] A. V. Chervonsky, "Microbiota and autoimmunity," *Cold Spring Harbor Perspectives in Biology*, vol. 5, no. 3, article a007294, 2013.

[10] E. A. Mayer, K. Tillisch, and A. Gupta, "Gut/brain axis and the microbiota," *Journal of Clinical Investigation*, vol. 125, no. 3, pp. 926–938, 2015.

[11] H. Wong and C. Hoeffer, "Maternal IL-17A in autism," *Experimental Neurology*, vol. 299, pp. 228–240, 2018.

[12] J. M. Yano, K. Yu, G. P. Donaldson et al., "Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis," *Cell*, vol. 161, no. 2, pp. 264–276, 2015.

[13] T. Harach, N. Marungruang, N. Duthilleul et al., "Reduction of Abeta amyloid pathology in APPPS1 transgenic mice in the absence of gut microbiota," *Scientific Reports*, vol. 7, no. 1, 2017.

[14] F. Scheperjans, V. Aho, P. A. B. Pereira et al., "Gut microbiota are related to Parkinson's disease and clinical phenotype," *Movement Disorders*, vol. 30, no. 3, pp. 350–358, 2015.

[15] B. Wang, M. Yao, L. Lv, Z. Ling, and L. Li, "The human microbiota in health and disease," *Engineering*, vol. 3, no. 1, pp. 71–82, 2017.

[16] J. Lloyd-Price, G. Abu-Ali, and C. Huttenhower, "The healthy human microbiome," *Genome Medicine*, vol. 8, no. 1, 2016.

[17] P. C. Y. Woo, S. K. P. Lau, J. L. L. Teng, H. Tse, and K.-Y. Yuen, "Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories," *Clinical Microbiology and Infection*, vol. 14, no. 10, pp. 908–934, 2008.

[18] M. G. I. Langille, J. Zaneveld, J. G. Caporaso et al., "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences," *Nature Biotechnology*, vol. 31, no. 9, pp. 814–821, 2013.

[19] R. Knight, A. Vrbanac, B. C. Taylor et al., "Best practices for analysing microbiomes," *Nature Reviews Microbiology*, vol. 16, no. 7, pp. 410–422, 2018.

[20] R. Poretsky, L. M. Rodriguez-R, C. Luo, D. Tsementzi, and K. T. Konstantinidis, "Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics," *PLoS One*, vol. 9, no. 4, article e93827, 2014.

[21] K. Raymann, A. H. Moeller, A. L. Goodman, and H. Ochman, "Unexplored archaeal diversity in the great ape gut microbiome," *mSphere*, vol. 2, no. 1, article e00026-17, 2017.

[22] M. Schirmer, E. A. Franzosa, J. Lloyd-Price et al., "Dynamics of metatranscription in the inflammatory bowel disease gut microbiome," *Nature Microbiology*, vol. 3, no. 3, pp. 337–346, 2018.

[23] M. J. Gosalbes, A. Durbán, M. Pignatelli et al., "Metatranscriptomic approach to analyze the functional human gut microbiota," *PLoS One*, vol. 6, no. 3, article e17447, 2011.

[24] S. Bashiardes, G. Zilberman-Schapira, and E. Elinav, "Use of metatranscriptomics in microbiome research," *Bioinformatics and Biology Insights*, vol. 10, article BBI.S34610, 2016.

[25] R. Knight, J. Jansson, D. Field et al., "Unlocking the potential of metagenomics through replicated experimental design," *Nature Biotechnology*, vol. 30, no. 6, pp. 513–520, 2012.

[26] S. He, O. Wurtzel, K. Singh et al., "Validation of two ribosomal RNA removal methods for microbial metatranscriptomics," *Nature Methods*, vol. 7, no. 10, pp. 807–812, 2010.

[27] T. A. K. Freitas, P.-E. Li, M. B. Scholz, and P. S. G. Chain, "Accurate read-based metagenome characterization using a hierarchical suite of unique signatures," *Nucleic Acids Research*, vol. 43, no. 10, article e69, 2015.

[28] MetaHIT Consortium, J. Li, H. Jia et al., "An integrated catalog of reference genes in the human gut microbiome," *Nature Biotechnology*, vol. 32, no. 8, pp. 834–841, 2014.

[29] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, vol. 44, no. D1, pp. D457–D462, 2016.

[30] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, 2018.

[31] R. D'Amore, U. Z. Ijaz, M. Schirmer et al., "A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling," *BMC Genomics*, vol. 17, no. 1, 2016.

[32] R. Sinha, G. Stanley, G. S. Gulati et al., *Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing*, bioRxiv, 2017.

[33] Illumina, *Effects of index misassignment on multiplexing and downstream analysis*, Illumina, 2017.

[34] S. J. Salter, M. J. Cox, E. M. Turek et al., "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses," *BMC Biology*, vol. 12, no. 1, p. 87, 2014.

[35] A. Glassing, S. E. Dowd, S. Galandiuk, B. Davis, and R. J. Chiodini, "Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples," *Gut Pathogens*, vol. 8, no. 1, 2016.

[36] R. W. Lusk, "Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data," *PLoS One*, vol. 9, no. 10, article e110808, 2014.

[37] P. B. Eckburg, E. M. Bik, C. N. Bernstein et al., "Diversity of the human intestinal microbial flora," *Science*, vol. 308, no. 5728, pp. 1635–1638, 2005.

[38] G. D. Wu, J. D. Lewis, C. Hoffmann et al., "Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags," *BMC Microbiology*, vol. 10, no. 1, p. 206, 2010.

[39] The Human Microbiome Project Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[40] American GutMay 2019, http://humanfoodproject.com/americangut/.