NIH Public Access
Author Manuscript
Metagenomics (Cairo). Author manuscript; available in PMC 2013 September 06.

Published in final edited form as:
*Metagenomics (Cairo)*. 2013 ; 2: 235646–.

# A microbial profiling method for the human microbiota using high-throughput sequencing

**Huei-Hun Elizabeth Tseng**[1,*], **Meredith A.J. Hullar**[2], **Fei Li**[2], **Johanna W. Lampe**[2], **Richard Sandstrom**[3], **Audra K. Johnson**[3], **Lisa L. Strate**[4], **Walter L. Ruzzo**[1,2,3], and **John Stamatoyannopoulos**[3]

[1]Department of Computer Science & Engineering, University of Washington, Seattle, WA 98195

[2]Fred Hutchinson Cancer Research Center, Seattle, WA 98109

[3]Department of Genome Sciences, University of Washington, Seattle, WA 98105

[4]Department of Medicine, Division of Gastroenterology, University of Washington, Seattle, WA 98104

## Abstract

Study of the human microbiota in relation to human health and disease is a rapidly expanding field. To fully understand the complex relationship between the human gut microbiota and disease risks, study designs that capture the variation within and between human subjects at the population level are required, but this has been hampered by the lack of cost-effective methods to characterize this variation. Illumina sequencing is inexpensive and produces millions of reads per run, but it is unclear whether short reads can adequately represent the microbial community of a human host. In this study, we examined the utility of a profiling method, microbial nucleotide signatures (MNS), focused on low-depth sampling of the human microbiota using Ilumina short reads. This method is intended to aid in human population-based studies where large sample sizes are required to adequately capture variation in disease or phenotype differences. We found that, by calculating the nucleotide diversities along the sequenced 16S rRNA gene region, which did not require assembly or phylogenetic identification, we were able to differentiate the gut microbial nucleotide signatures of 9 healthy individuals. When we further subsampled the reads down to 40,000 reads (51 bp long) per sample, the diversity profiles were relatively unchanged. Applying MNS to a public datasets showed that it could differentiate body site differences. The scalability of our approach offers rapid classification of study participants for studies with the sample sizes required for epidemiological studies. Using MNS to classify the microbiome associated with a disease state followed by targeted in-depth sequencing will give a comprehensive understanding of the role of the microbiome in human health.

## INTRODUCTION

Billions of bacteria live in the gastrointestinal tract and their symbiotic relationship with the host greatly affects human health [1, 2]. Over the past few years, high-throughput sequencing of the 16S rRNA gene has shown associations between the gut microbiota and diseases such as obesity [3], diabetes [4], inflammatory bowel disease [2], and cancer [5]. The16S rRNA gene is a widely used molecule for bacterial species identification for several reasons [6]: (1) it is present in all bacteria and archaea; (2) it has both conserved regions that

[*] To whom correspondence should be addressed. lachesis@cs.washington.edu.

**SOFTWARE AVAILABILITY**: The MNS source code is written in Python and available at: http://www.cs.washington.edu/homes/lachesis/.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

can be used as universal primer targets and variable regions that are species-specific; and (3) it has been extensively studied and comprehensive databases exist (e.g., SILVA [7], RDP [8], Greengenes [9]). To be able to assess the role of the microbiota in disease, studies need to be conducted prospectively and across different populations with a large enough sample size to encompass the natural variation in the gut microbiota [10]. However, clinical and epidemiological studies are often constrained by the expense of recruitment efforts. In addition, insufficient sequence depth will fail to adequately capture the variation of the microbiota within and between samples. A cost-effective approach is needed to profile the gut microbial community; this can then be followed by deep sequencing and phylogenetic analysis of selected individuals to relate differences in the microbiota with disease state or phenotype. For this screening process, we developed microbial nucleotide signatures (MNS) as a rapid, inexpensive approach to differentiate the gut microbiota of difference human disease states or phenotypes. MNS does not use phylogeny to identify the bacterial species (in contrast to the approach often taken using UniFrac [11] or through established pipelines provided by Mothur [12] and Qiime [13]). On the contrary, it is designed for short, fragmentary reads sequenced at a low depth that precludes successful assembly or phylogenetic analysis. MNS is not meant to replace deep sequencing and phylogenetic analysis but to aid in assessment of the role of the microbiota in health in clinical and human population-based studies. In this study, we focus on finding how few reads we would need to characterize the gut microbiota using MNS. While the eventual goal of gut microbiota studies is to find out what species are present, our approach provides an inexpensive strategy for characterizing samples—an approach that could be applied as a first-pass screening of large human populations. Once differences between samples have been identified and grouped using our profiling method, select samples could be picked from each group for further deep-sequencing and phylogenetic analysis. This provides a cost-effective, in-depth analysis of the role of the gut microbiota in human population studies.

## MATERIAL AND METHODS

Figure 1 outlines the MNS process: reads are filtered by quality scores and length, then aligned to a reference 16S rRNA gene database. A diversity index is calculated to represent the nucleotide diversity and evenness independently at each position in the sequenced 16S rRNA gene region (Figure 2). This method focuses on one variable region of the 16S rRNA gene but may be expanded to include the entire gene or other genes of interest. The total set of diversity indexes, which is a 1-d vector, is the MNS of the sample.

In the following sections we describe: 1) reference database creation, 2) read alignment, 3) calculation and clustering based on MNS, and 4) validation of our approach in a cross-sectional study. The computational bottleneck is aligning reads to the reference database. This step can be easily parallelized and be completed for an entire Illumina run (~80 million reads) in several hours. After alignment is done, computing and clustering the MNSs takes only several minutes.

### 2.1 Creating a reference sequence database

The reference sequence database is a multiple sequence alignment of representative gut bacterial species. Accuracy of read placement is dependent on how well the composition of the database represents the gut microbiota. The more species we include in the database, the more likely a read will be correctly aligned. However, as the database size grows, both runtime and memory requirements for the aligners increase. We created a gut bacterial reference database by selecting non-redundant sequences from SILVA 104 that were annotated to be from fecal samples (Supp. section 2-3; Supp. Figure 1; Supp. Figure 2), which included sequences from both cultured and uncultured bacteria. We obtained 87,495 16S rRNA gene sequences that were mostly Firmicutes and Bacteroidetes (the two most

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

dominant phyla in the gut). Although creating this database from existing knowledge of the gut species may cause reads from novel species to be misaligned, we used this approach to maximize alignment efficiency. In our own dataset, we were able to align 60-90% of the reads, indicating that our reference database is representative of the gut microbiota.

## 2.2 Aligning reads using BowTie and BLAST

We used a multifaceted approach to optimize accuracy and efficiency of alignment to the reference database. BowTie [14] is a fast, memory-efficient aligner designed to align reads with very few mismatches to reference sequences. We ran BowTie with the maximum number of mismatches allowed (3 mismatches) against the gut bacterial reference database. Because BowTie uses a greedy, randomized search algorithm to find non-exact matches, it may not align all reads with at most 3 mismatches to the database or find the optimal match unless an exhaustive search is done. Because an exhaustive search is time-consuming, we set the parameter to -*k* 1 in BowTie which allows it to report the first match to the reference database for a given read. This less stringent criterion does not significantly affect the MNS of a sample (Supplement Section 3, Supp. Figure 4). To recover the small percentage of reads with at most 3 mismatches but missed by BowTie, we align the remaining unaligned reads through the same database, using the slower but more sensitive BLAST [15]. BLAST is run with parameters set so that gaps are not allowed and the best match to the database is reported. Both BowTie and BLAST can easily be parallelized for large numbers of reads to reduce runtime. For a full Illumina dataset (~80 million reads), alignment using 10 parallel processes takes at most several hours to complete. From the BowTie/BLAST output, we obtain an ungapped, base-to-base mapping of the read to a reference sequence—this mapping can then be extended to a universal, gapped, alignment for MNS calculation.

## 2.3 Calculating microbial nucleotide signatures as a vector of Simpson or Entropy Indexes

Simpson Diversity Index [16] is commonly used in ecology to measure both species richness and evenness. Species richness refers to the number of different species and evenness refers to the distribution of the species in a community. The same idea is applied here to account for the richness and evenness of nucleotides at each sequenced position in the 16S rRNA gene. For each position $i$, Simpson Index $D_i$ is calculated by:

$$D_i=1-\frac{\sum_{k=\{A,T,C,G\}} n_{k,i}\,(n_{k,i}-1)}{N_i\,(N_i-1)},\ N_i=\sum_{k=\{A,T,C,G\}} n_{k,i} \quad (1)$$

where $n_{k,i}$ is the number of nucleotides $k$ ($k = \{A, T, C, G\}$) observed at position $i$ and $N_i$ is the total number of nucleotides observed at position $i$. Equivalently, $D_i$ is the fraction of non-identical pairs of nucleotides in column i. When $N_i$ is large, we can approximate $D_i$ by:

$$D_i \approx 1-\sum_{k=\{A,T,C,G\}}\left(\frac{n_{k,i}}{N_i}\right)^2 \quad (2)$$

This creates a vector whose length is the number of nucleotide positions.

The second term in $D_i$ can be thought of as the sum of squared nucleotide proportions and is closer to 1 when the position is dominated by a certain nucleotide. Thus, $D_i$ is 0 when there is no diversity at position $i$ and increases with diversity (with a maximum of 3/4 for the 4-letter nucleotide alphabet using the simplified formula).

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Simpson Index encapsulates both the richness and evenness of nucleotides at each position. Another way to account for nucleotide diversity is to use information entropy. Entropy is often used in information theory to measure the expected amount of information contained in a message [17]. For each position *i*, the Entropy Index $E_i$ is calculated by:

$$\begin{gathered} E_i=\sum_{k=\{A,T,C,G\}}\left(p_{k,i}\times\log\frac{p_{k,i}}{q_k}\right)^2 \\ p_{k,i}=\frac{n_{k,i}}{N_i} \\ q_k=\frac{\sum_i n_{k,i}}{\sum_i\sum_k n_{k,i}} \end{gathered} \tag{3}$$

where $q_k$ is the nucleotide frequency across the entire sequenced region.

Once MNSs are calculated, pairwise distances between samples are calculated using their Euclidean distances. Clustering is done using average-linkage hierarchical clustering (UPGMA) [18].

## 2.4 Cross-sectional human population study

**2.4.1 Study participants—**Women and men were recruited from the Seattle, Washington area. Exclusion criteria included use of antibiotics during the 3-month period prior to sample collection and age less than 18 years. Fecal samples were collected from 9 participants (individual A-I) over two time points (3 months apart). The Institutional Review Board of the Fred Hutchinson Cancer Research Center, Seattle, WA approved all study procedures, and all participants provided informed written consent.

**2.4.2 Sample and DNA extraction—**Aliquots of fresh fecal samples (approx. 3 g) were collected into fecal collection containers (Fisher Scientific, Fair Lawn, NJ) containing 3 ml RNAlater (Ambion, Austin, TX) and shaken. Samples were shipped to the lab and stored at -80 C until DNA extraction. Before DNA extraction, samples were homogenized by OMNI tissue homogenizer 115 (OMNI Inc., Marietta, GA) and divided into 300 μl aliquots, centrifuged at 16,000 ×*g* with 300 μl phosphate buffered saline for 10 min and the supernatant containing RNAlater was discarded. Genomic DNA was extracted from fecal samples (QIAamp DNA Stool Mini Kit, QIAgen, Valencia, CA) with 1 min bead beating on setting 5.5 using a FastPrep system (MP Biomedicals, Solon, OH) [19].

**2.4.3 PCR primers and conditions—**The DNA of bacterial 16S rRNA genes was amplified with primer 330F (5'-ACT CCT ACG GGA GGC AGC AGT-3') and 530R (5'-GTATTACCGCGGCTGCTGGCAC-3') using PCR conditions as described in [20].

**2.4.4 Library construction—**Amplified sequences were finely fragmented using the published Covaris S2 protocol "DNA Shearing with microTubes (<1.5kb fragments)" with a target base pair peak of 200 for 6 minutes (Covaris, Inc., Woburn, MA). Following fragmentation, libraries were prepared following the protocol supplied with Illumina's Multiplexing Sample Preparation Oligonucleotide Kit scaled down for smaller DNA concentrations. The resulting "end-adaptered" DNA fragment population was subjected to massively parallel sequencing on an Illumina GA2 genome analyzer (Illumina, Inc., San Diego, CA).

## 2.5 Read alignment and quality filtering

**2.5.1 Pre-alignment quality filtering—**Reads were end-clipped using a Phred score cutoff of 2, which indicates ~63% chance of sequencing error. Clipped reads with length <30 bp were discarded and accounted for <1% of the reads. Any remaining read with at least

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

one base with a Phred score <10 was further removed. About 70-80% of reads remained after the above steps (see Supp. Table 1).

**2.5.2 Read alignment—**Reads were first aligned using BowTie with parameters −k 1 −n 3 which reports one best match with up to 3 mismatches to the reference sequence. We used the ungapped version of our refDB (section 2.1), but kept a mapping of the ungapped to gapped positions. Reads not aligned by BowTie were further aligned with the more sensitive BLAST, again allowing up to 3 mismatches. Reads whose first base aligned outside the V3 hypervariable region were discarded as they were likely to be contaminants. Reads whose first base aligned to the first or second position of V3 or whose last base aligned to the last or second to last position were also discarded because we observed an amplification bias that resulted in an abnormal excess of primer-containing reads (see Supp. Table 2). The mapped reads were then extended to the gapped alignment in preparation for MNS calculation.

**2.5.3 Calculating microbial nucleotide signatures—**We calculated diversity indices (either the Simpson or Entropy Index) using an "*E. coli* mask" on the sequenced region, i.e., we only included nucleotides that mapped to an *E. coli* reference nucleotide position in the SILVA database. We did this for practical reasons: the full alignment in the current SILVA database is very long (50,000 bp) due to long insertions (gaps) with respect to the *E. coli* reference. Since most gut species have very few insertions with respect to *E. coli* 16S rRNA gene, by excluding those insertions, we lost little information. Of the 87,495 reference sequences in SILVA, 86,447 had fewer than 5% non-*E. coli* position bases. Though we might have lost some information by only looking at the *E. coli* positions, we argue that for our method it is more important to obtain high nucleotide coverage at each position to calculate an accurate diversity index. By using *E. coli* positions, we maximized the amount of information and minimized potential noise.

# RESULTS

## Applying microbial nucleotide signatures to 9 healthy individuals

We sampled 9 healthy individuals (A-I) over two time points (3 months apart) and obtained a total of 20 samples (individual C had technical replicates). We PCR-amplified the V3 hypervariable region of each and sequenced the amplicons using Illumina sequencing. We decided to sequence the V3 hypervariable region of the 16S rRNA gene because our analysis showed that it was the most suitable for short read sequencing given its moderate length and high taxonomic identifiability (Supp. section 2; Supp. Table 3). We obtained 1-7 million 51 bp reads per sample. After quality filtering and alignment, 20-50% of the reads remained (Supp. Table 1 and 2). Of the total usable reads, less than 0.8% of the nucleotides mapped to non-*E. coli* positions.

Clustering based on MNSs showed that intra-individual differences were smaller than inter-individual differences. Samples from the same individuals always clustered closely (Figure 3) and the distances within the same individual were smaller than distances between individuals (Figure 4, *t*-test; p-value: $8 \times 10^{-10}$).

## Subsampling of reads show microbial nucleotide signatures informative at 40,000 reads

Microbial nucleotide signatures differentiated samples when the *effective* sequencing depth (i.e., number of high-quality, aligned reads; see last column in Supp. Table 2) was on the order of millions. To see if MNSs are equally informative when sequencing depth is lower, we randomly subsampled the pool of high-quality, aligned sequences for each individual, calculated the MNSs, and compared them to the original MNSs. The subsampling was

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

repeated 100 times for each subsample size. As we drew more reads from the original pool, the subsampled MNSs approached the original MNS calculated with ~1 million reads (Figure 5). Furthermore, the subsampled MNSs converged rather quickly for all samples (Figure 5, Supp. section 6). For example, sample A +0 originally had 1.8 million raw reads. After quality filtering, 39% or ~0.7 million reads were used to create the MNSs and were drawn from to create our subsamples. At a subsample size of 5,120 reads, the Euclidean distance between the subsampled and original MNS was <0.2, smaller than the distance between any pairs of samples shown in Figure 4. As proof, we generated a clustering for each of the subsampled MNSs to see how often we recovered the subtrees found in the full dataset. Figure 6 shows the frequency of the subtrees in the subsampled MNS overlain on the original clustering. A frequency of 1.0 means that this subtree appeared in the cluster in 100% of the subsampling tests. At 40,960 reads, except for one subtree, both the Simpson Index MNS and the Entropy Index MNS resulted in the same clustering as their original MNSs (Supp. section 7). The Simpson and Entropy Indexes are both good choices for calculating nucleotide diversity, although from our subsampling analysis it appears that the Entropy Index may be slightly preferable at lower effective sequencing depths. This might be because the Entropy Index accounts for the background nucleotide diversity (the denominator in the index calculation) and the Simpson Index does not.

### Applying MNS clustering to other datasets

We applied our MNS clustering approach to two other datasets: the MetaHit data from Qin et al. [21] and the atherosclerosis patient samples from Omry et al. [22]. The MetaHit data is publicly available and consists of metagenomics reads (75bp, Illumina) from 60 Spanish and 40 Danish human samples. To extract the 16S rRNA gene reads, we ran BowTie with our reference database and calculated MNSs using the same procedures. We obtained an average of 33,000 reads per sample that were distributed across the entire 16S rRNA gene. Using the whole reference *E. coli* region to calculate MNS, we showed that technical replicates were very well clustered for both datasets (Supp Figure 5 and 6). For the atherosclerosis dataset, we obtained the pyrosequencing data from the authors, which consisted of samples from feces, mouth, and plaque. The average number of samples is 5,000. The sequenced region was the 16S rRNA V2 region (*E. coli* position 220-320), which we used to align and calculate MNS. Supp Figure 6 shows that except for two samples, all samples cluster together by body site and not individuals.

## DISCUSSION

We found that with microbial nucleotide signatures, we can characterize gut microbiota differences in human hosts with as few as 40,000 reads, instead of the ~1 million reads we sequenced. While the MNS of a sample does not give the phylogenetic information, our study shows that enough information can be obtained even when assembly or taxonomic identification are not feasible. For long reads like those produced by pyrosequencing and the more recent Illumina reads (≥100bp), standard processing pipelines from RDP, Mothur, and Qiime may be the right tools. For short, low-coverage, fragmentary reads like the ones produced in our study and from MetaHit, MNS can be used to characterize the samples; all that is required is a reference database and a read aligner.

Extending from our results, we think MNS can be of great use in screening and identifying whether the gut microbial community in individuals with a particular disease differs from that of non-diseased controls. If differences between health and disease (or two different phenotypes) are noted, additional studies including deeper sequencing and phylogenetic analysis could be performed to further elucidate these differences and their functional significance. This may be particularly useful in diseases with diagnostic uncertainty or overlap. For example, inflammatory bowel disease (IBD) has an established correlation with

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the gut microbiota and in which diagnostic subtyping (Crohn's versus ulcerative colitis) can be difficult based on standard clinical testing [2, 23]. Furthermore, ulcerative colitis and Crohn's present a range of clinical manifestations. One could classify individuals with IBD into distinct subgroups based on MNSs which could then be targeted for further analysis. The same could be applied for forensic purposes as a recent study showed that pyrosequencing data from biological stains could be used in forensic identification [24].

We addressed the sensitivity and effectiveness of MNS as a diagnostic screening tool. We show that it can detect rare species at low abundances (see Supp. section 5; Supp. Figure 4 for additional analysis). For effectiveness, we showed using repeated subsampling that ~40,000 51 bp reads are just as effective at differentiating among healthy individuals' gut microbiota as using millions of reads (Figure 6, Supplement Section 6-7). This means that for single high-throughput sequencing runs, e.g., Illumina Hi-Seq which produces > 100 million reads, with multiplexing one could sequence hundreds of individuals, reducing screening costs while assessing variation at the human population level.

The number of sequences required to represent a microbial community differs depending on the sampled environment, the sequenced gene region, and the read length. In Caporaso et al. [25], the authors showed that 2,000 reads (of 100 bp) were necessary to represent environmental soil samples. For our 9-individual human gut microbiota dataset, we needed 40,000 51 bp reads. We speculate that the difference in the minimum number of reads is a result of read length and differences in phylogenetic diversity of the environments we are distinguishing. We applied our method to distinguish individual differences within the same environment (human gut; variation at species level) as opposed to comparing across different environments (gut, tongue, ocean, soil, etc; variation at phylum level). Since individual differences are more subtle than environmental biome differences, it was expected that we would need more reads. Currently MNS requires that all samples be sequenced from the same genomic region. For future work, we aim to develop methods that do not require sequencing from the same genomic region, though that would likely introduce new biases that need to be addressed, as it has been shown that there are region-specific sequencing biases [26]. We would also like to extend the approach beyond the 16S rRNA gene. Though the 16S rRNA gene is ideal for identifying bacteria, it is often more important to look at the profile of a microbial community from the perspective of functionality, as different species can have the same set of active genes. However, functional genes are not as comprehensively studied as the 16S rRNA gene and lack comprehensive databases. Our method will have to deal with identifying novel genes, comparing homologs and develop ways to account for the presence or absence of genes.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This work was supported by US NIH grants to the Fred Hutchinson Cancer Research Center (R03 CA115209, R01 DK084157).

## References

1. Eckburg PB. Diversity of the Human Intestinal Microbial Flora. Science. Jun; 2005 308(5728): 1635–1638. [PubMed: 15831718]

2. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A. Aug; 2007 104(34):13780–13785. [PubMed: 17699621]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

3. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. A core gut microbiome in obese and lean twins. Nature. Jan; 2009 457(7228):480–484. [PubMed: 19043404]

4. Larsen N, Vogensen FK, van den Berg FWJ, Nielsen DS, Andreasen AS, Pedersen BK, Al-Soud WA, Sørensen SJ, Hansen LH, Jakobsen M. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PLoS ONE. 2010; 5(2):e9085. [PubMed: 20140211]

5. Greer JB, O'Keefe SJ. Microbial induction of immunity, inflammation, and cancer. Front Physiol. 2011; 1:168. [PubMed: 21423403]

6. Clarridge JE. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. Clinical Microbiology Reviews. Oct; 2004 17(4): 840–862. [PubMed: 15489351]

7. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucl Acids Res. Dec; 2007 35(21):7188–7196. [PubMed: 17947321]

8. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research. Jan; 2009 37(Database):D141–D145. [PubMed: 19004872]

9. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. Mar; 2012 6(3):610–618. [PubMed: 22134646]

10. Lampe JW. The Human Microbiome Project: getting to the guts of the matter in cancer epidemiology. Cancer Epidemiol Biomarkers Prev. Oct; 2008 17(10):2523–2524. [PubMed: 18842991]

11. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J. Feb; 2011 5(2):169–172. [PubMed: 20827291]

12. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. Dec; 2009 75(23): 7537–7541. [PubMed: 19801464]

13. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. Nature Methods. Apr; 2010 7(5):335–336. [PubMed: 20383131]

14. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10(3):R25. [PubMed: 19261174]

15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. Oct; 1990 215(3):403–410. [PubMed: 2231712]

16. Simpson, Edward H. Measurement of diversity. Nature. 1949; 163:688.

17. Jaynes E. Information Theory and Statistical Mechanics. Phys Rev. May; 1957 106(4):620–630.

18. Sokal R, Michener CD. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin. 1958; 38:1409–1438.

19. Li F, Hullar MAJ, Lampe JW. Optimization of terminal restriction fragment polymorphism (TRFLP) analysis of human gut microbiota. J Microbiol Methods. Feb; 2007 68(2):303–311. [PubMed: 17069911]

20. Ahmed S, Macfarlane GT, Fite A, McBain AJ, Gilbert P, Macfarlane S. Mucosa-Associated Bacterial Diversity in Relation to Human Terminal Ileum and Colonic Biopsy Samples. Applied and Environmental Microbiology. Sep.2007 73:7435–7442. [PubMed: 17890331]

21. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, Wang J. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. Mar; 2010 464(7285):59–65. [PubMed: 20203603]

22. Koren O, Spor A, Felin J, Fåk F, Stombaugh J, Tremaroli V, Behre CJ, Knight R, Fagerberg B, Ley RE, Bäckhed F. Microbes and Health Sackler Colloquium: Human oral, gut, and plaque microbiota in patients with atherosclerosis. Proc Natl Acad Sci U S A. Oct.2010

23. Nikolaus S, Schreiber S. Diagnostics of inflammatory bowel disease. Gastroenterology. Nov; 2007 133(5):1670–1689. [PubMed: 17983810]

24. Brenig B, Beck J, Schütz E. Shotgun metagenomics of biological stains using ultra-deep DNA sequencing. Forensic Sci Int Genet. Jul; 2010 4(4):228–231. [PubMed: 20457050]

25. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A. Mar; 2011 108(Suppl 1):4516–4522. [PubMed: 20534432]

26. Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Welch DM, Relman DA, Sogin ML. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. Nov.2008 4(11):e1000255. [PubMed: 19023400]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

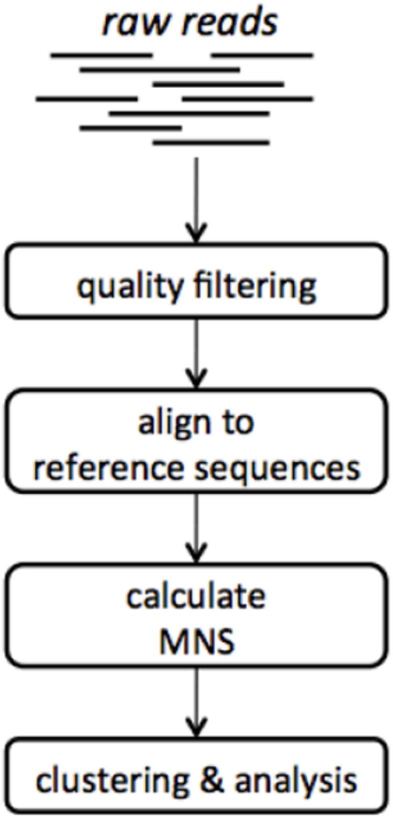
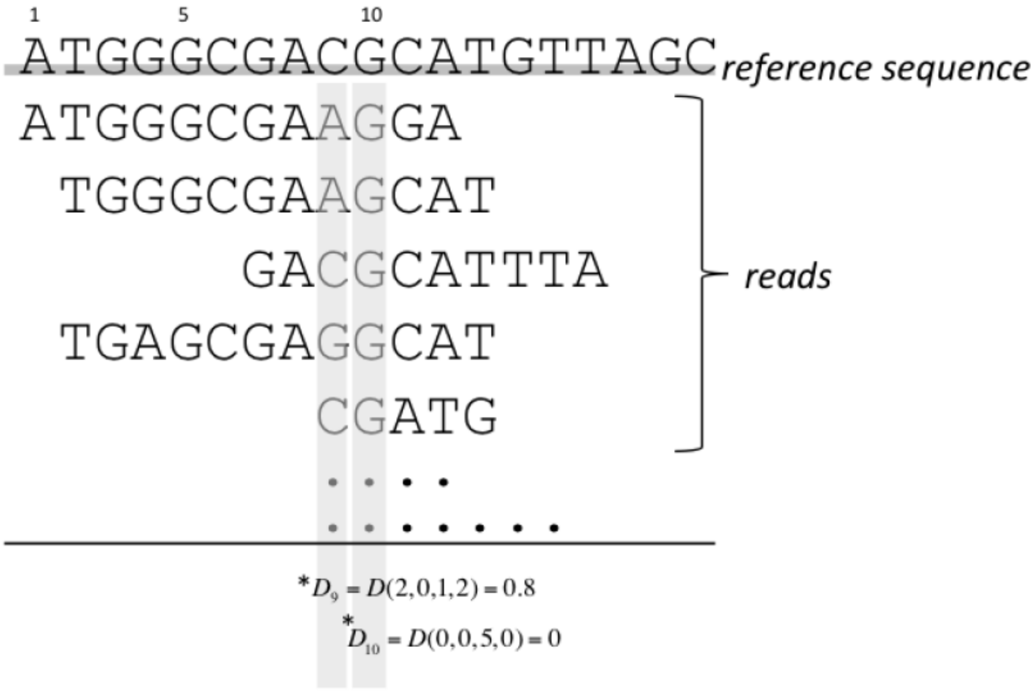


**Figure 1.**
Diagram of microbial nucleotide signatures (MNS) analysis.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript



$$*\; D(n_A, n_T, n_G, n_C) = 1 - \frac{n_A(n_A-1) + n_T(n_T-1) + n_G(n_G-1) + n_C(n_C-1)}{N(N-1)}$$

$$N = n_A + n_T + n_G + n_C$$

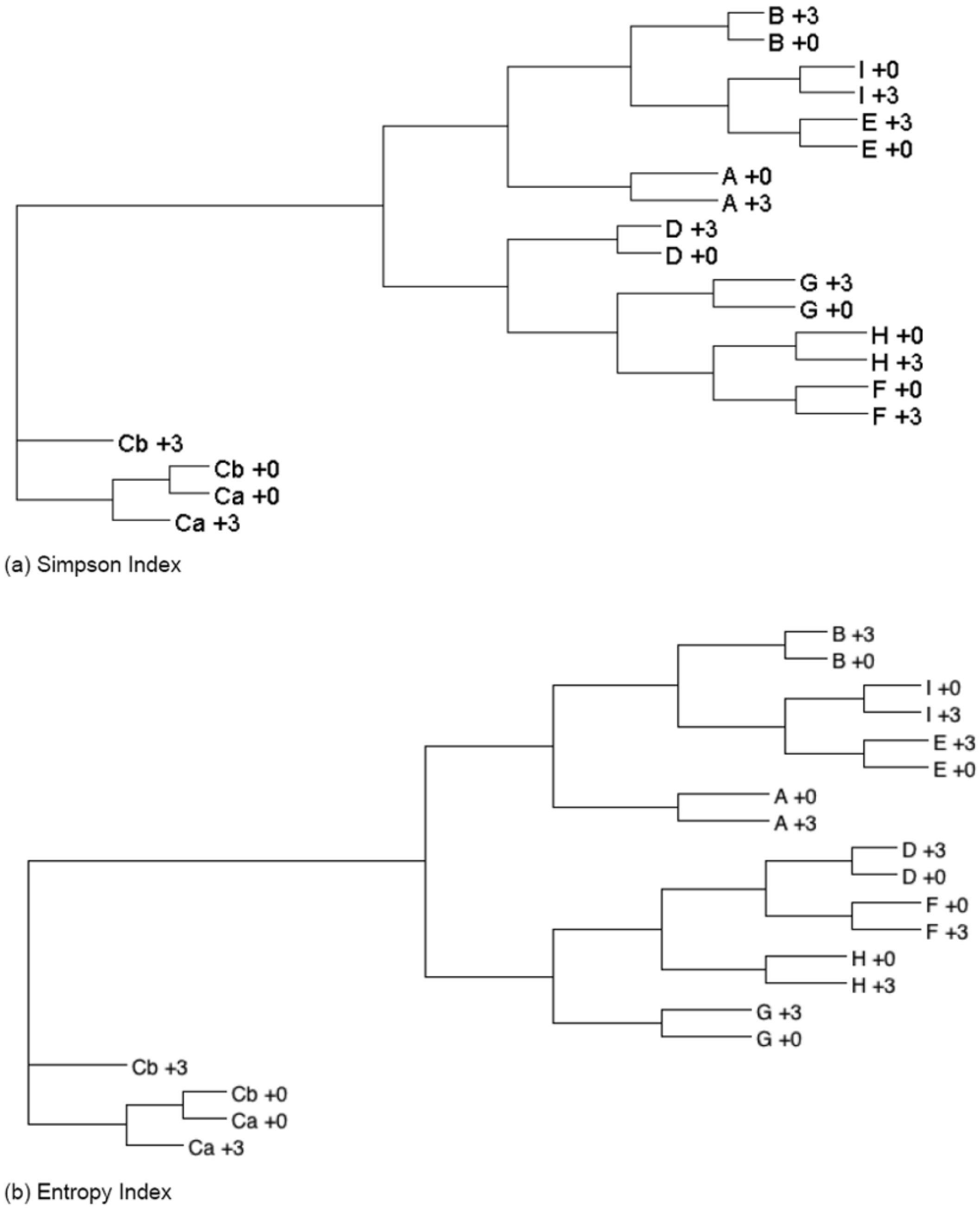
**Figure 2.**

Concept of microbial nucleotide signatures as a vector of nucleotide diversity indexes calculated independently for each position in the sequenced region. Here nucleotide diversity is calculated using the Simpson Index, although alternative indices such as the Entropy Index (see Methods) could also be used.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



(a) Simpson Index

(b) Entropy Index

**Figure 3.**

Clustering of 20 samples from 9 individuals (A-I) at two time points (+0, +3) using microbial nucleotide signatures calculated with (a) Simpson Index or (b) Entropy Index. Ca/ Cb are technical replicates. Samples from the same individuals all cluster closely together.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

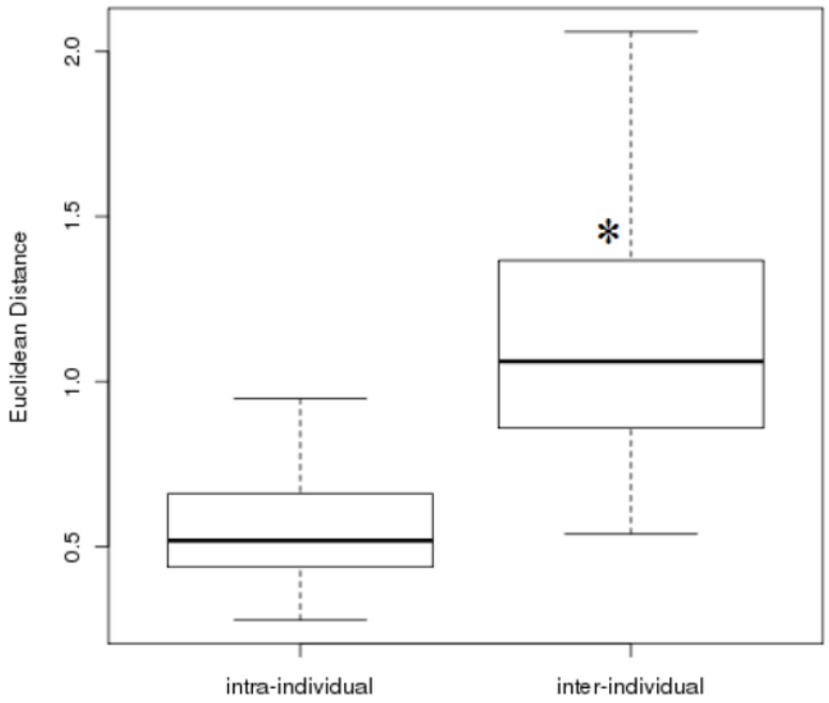


**Figure 4.**

Intra- and inter-individual distances between 9 healthy individuals. The Euclidean Distance is calculated using the microbial nucleotide signatures. The mean is 0.54±0.18 for intra-individual samples and 1.14±0.35 for inter-individual samples. (*t-test p-value: $8\times10^{-10}$).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

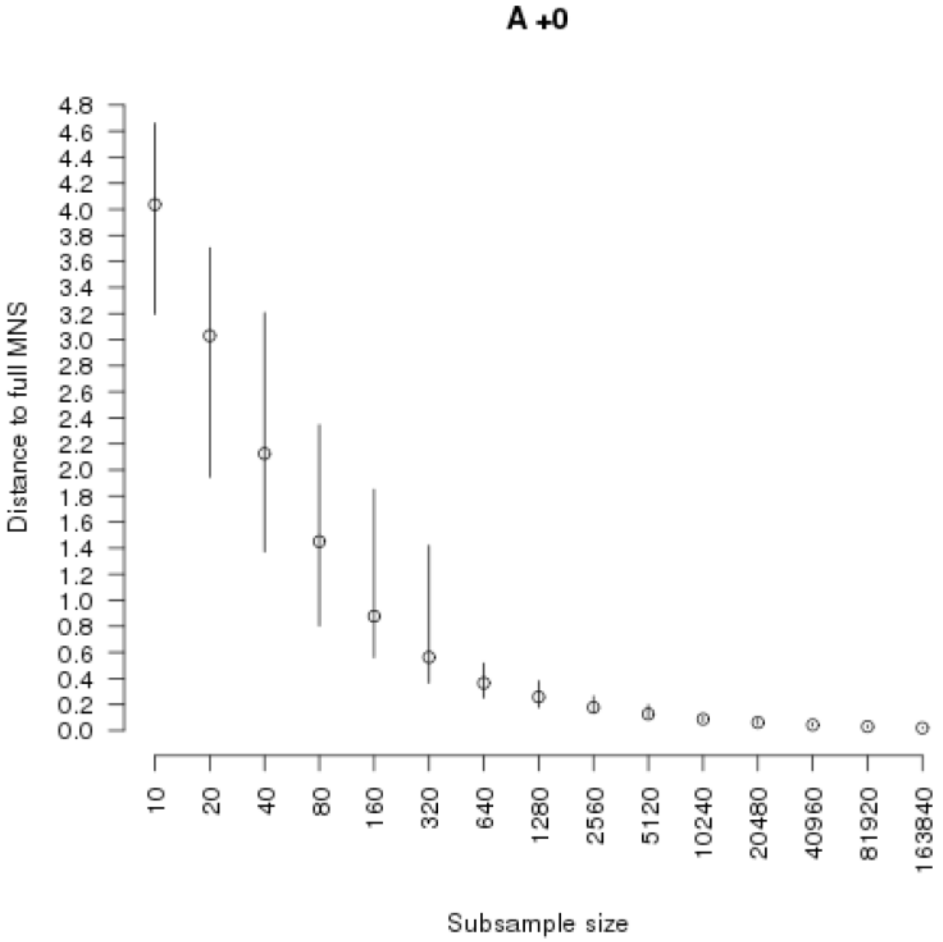


**Figure 5.**

Subsampled microbial nucleotide signature (MNS) rapidly approaches the original MNS as the subsample size grows. The X-axis is the subsample size and the Y-axis is the Euclidean distance between the subsampled and original MNS. The means are drawn as circles and error bars are plotted, based on 100 repetitions. Sample A +0 shown; others are similar. See supplementary materials for the analogous plot for the rest of the samples.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

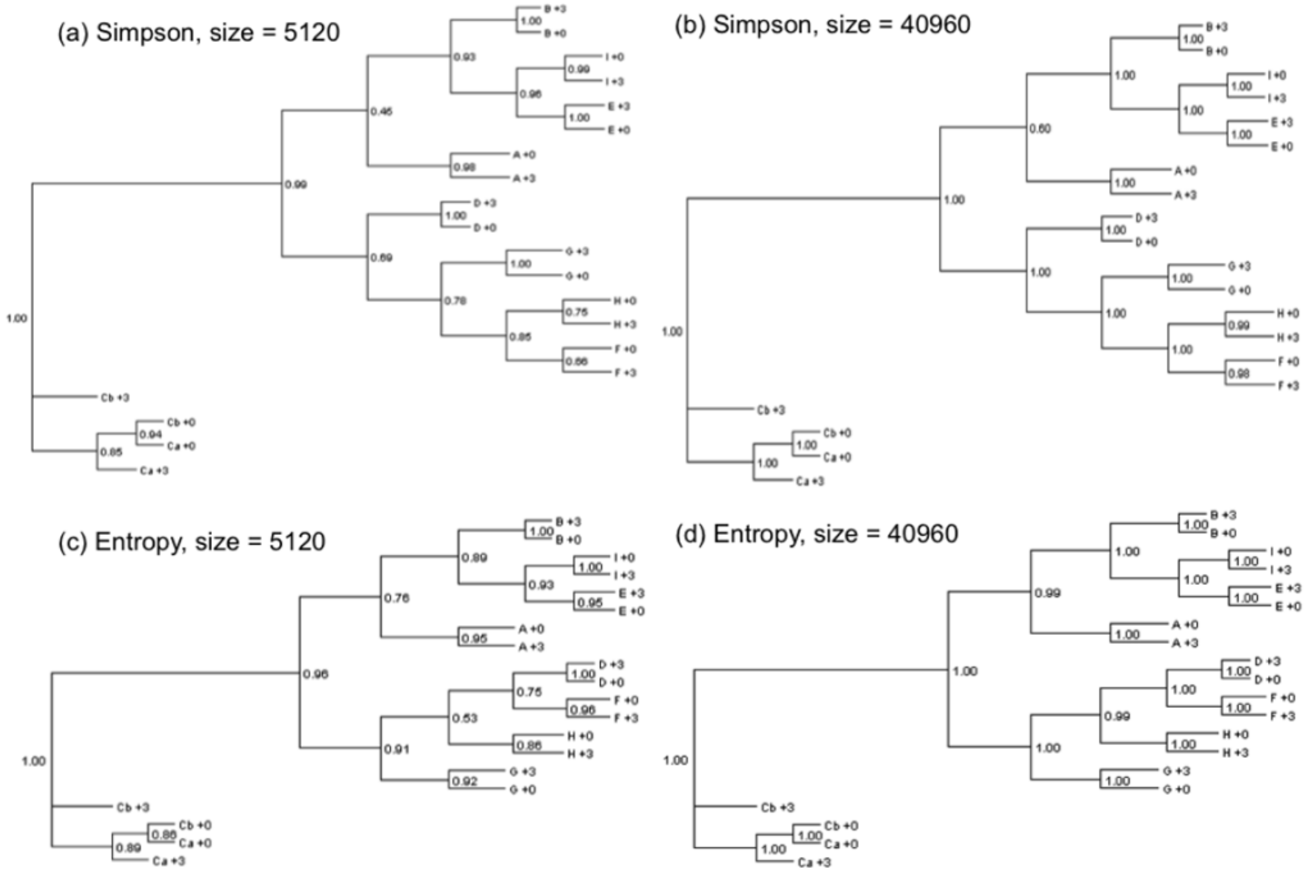


**Figure 6.**

Clustering of 9 healthy individual’s Illumina samples using either Simpson Index (a,b) or Entropy Index (c,d) at subsample size 5,120 (a,c) or 40,960 (b,d). Tree is the same as in Figure 3. Numbers are the fraction of 100 random subsamples in which the given partition of the individuals appeared in the clustering. See supplementary materials for the analogous plots for the rest of the subsample sizes.