Published in final edited form as:

J Microbiol Methods. 2014 October ; 105: 82–87. doi:10.1016/j.mimet.2014.07.001.

Improved resolution of bacteria by high throughput sequence analysis of the rRNA internal transcribed spacer

Paul M. Ruegger^a, Robin T. Clark^a, John R. Weger^b, Jonathan Braun^c, and James Borneman^{a,*}

^aDepartment of Plant Pathology and Microbiology, University of California, Riverside, California 92521, USA

^bDepartment of Botany & Plant Sciences, University of California, Riverside, California 92521, USA

^cDepartment of Pathology and Lab Medicine, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA

Abstract

Current high throughput sequencing (HTS) methods are limited in their ability to resolve bacteria at or below the genus level. While the impact of this limitation may be relatively minor in whole-community analyses, it constrains the use of HTS as a tool for identifying and examining individual bacteria of interest. The limited resolution is a consequence of both short read lengths and insufficient sequence variation within the commonly targeted variable regions of the small-subunit rRNA (SSU) gene. The goal of this work was to improve the resolving power of bacterial HTS. We developed an assay targeting the hypervariable rRNA internal transcribed spacer (ITS) region residing between the SSU and large-subunit (LSU) rRNA genes. Comparisons of the ITS region and two SSU regions using annotated bacterial genomes in GenBank showed much greater resolving power is possible with the ITS region. This report presents a new HTS method for analyzing bacterial composition with improved capabilities. The greater resolving power enabled by the ITS region arises from its high sequence variation across a wide range of bacterial taxa and an associated decrease in taxonomic heterogeneity within its OTUs. Although the method should be adaptable to any HTS platform, this report presents PCR primers, amplification parameters, and protocols for Illumina-based analyses.

Keywords

High throughput sequencing; rRNA; internal transcribed spacer; ITS; Illumina; bacteria

© 2014 Elsevier B.V. All rights reserved.

*Corresponding author: James Borneman, Department of Plant Pathology and Microbiology, University of California, Riverside, California 92521, USA, phone 1-951-827-3584, borneman@ucr.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

New and important experimental capabilities for examining bacterial communities have come from application of high throughput sequencing (HTS) technology to the small-subunit rRNA (SSU) gene. The combination of large numbers of sequencing reads and multiplexing technologies provides a cost-effective method to perform in-depth analysis of large numbers of samples. These capabilities facilitate investigations examining temporal and spatial variables (Caporaso et al., 2011a; Costello et al., 2009), and are fostering a broader understanding of the roles bacterial communities play in ecosystem functioning.

Although current HTS methods are limited in their ability to taxonomically classify bacterial sequences at or below genus levels (Caporaso et al., 2011b; Price et al., 2009; Wu et al., 2010), a related but less considered weakness is their limited ability to resolve such bacterial sequences. In this report, resolution, or resolving power, is defined as the ability to distinguish between different types of bacteria by their sequence differences, irrespective of whether taxonomy has been or can be imputed to them from the outset. Greater resolving power enables various analyses to be performed at finer levels, enhancing the ability to, for example, correlate specific bacteria with host or environmental traits, monitor population shifts, or assess microgeographical relationships through use of sequence-selective PCR primers and probes.

To achieve our goal of improving the resolving power of bacterial HTS, we developed an assay that targets the rRNA internal transcribed spacer (ITS) region. Because of its high sequence variation, the ITS region has been used for species and subspecies-level classification of bacteria (Barry et al., 1991; Frothingham et al., 1993; McLaughlin et al., 1993) and for ribosomal internal spacer analyses (RISA/ARISA) of community composition (Borneman et al., 1997; Fisher et al., 1999). Experimental protocols were created to facilitate the use of our new ITS HTS method, including PCR primers, amplification parameters and protocols for Illumina- based analyses. Using annotated bacterial genomes, the resolving power of our new method was assessed by a comparative analysis of intra-region sequence differences in the ITS region and two commonly used SSU regions. While our method's primary strengths do not rely on taxonomically identifying sequences it can be useful to do so. Therefore, we created an SSU-ITS sequence database and compare the accuracy of taxonomic assignments between the ITS and SSU regions in a mock community analysis. Moreover, we demonstrate a feasible way of performing a phylogenetic beta diversity analysis using ITS reads in a proof-of- concept experiment. After constructing a phylogenetic tree from the SSU genes of the SSU-ITS sequence database we used it to compare the experimental results obtained from ITS and SSU assays of the fecal bacteria of two mouse genotypes.

2. Materials and methods

2.1. Sample acquisition and DNA extraction

Fecal samples were collected from mice of two genotypes, obtained from Jackson Laboratories (Bar Harbor, Maine, USA): C3H-IL10-KO = C3Bir.129P2 (B6)-Il10tm1Cgn/Lt and C3H = C3H/HeJ. Animal use protocols were approved by the

Institutional Animal Care and Use Committee at the University of California, Riverside. DNA was extracted from the samples using the PowerSoil DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA, USA), and a 30-second beat-beating step using a Mini-Beadbeater-16 (BioSpec Products, Bartlesville, OK, USA). High throughput sequencing analysis of bacterial rRNA genes was performed on the resulting DNA using two Illumina-based assays: one targeting the small-subunit (SSU) rRNA gene and the other targeting the rRNA internal transcribed spacer (ITS) region.

2.2. SSU assay

One hundred microliter amplification reactions were performed in an MJ Research PTC-200 thermal cycler (Bio-Rad Inc, Hercules, CA, USA) containing: 50 mM Tris (pH 8.3), 500 µg/ml bovine serum albumin (BSA), 2.5 mM MgCl₂, 250 µM of each deoxynucleotide triphosphate (dNTP), 400 nM of each PCR primer, 4 µl of DNA template, and 2.5 units JumpStart Taq DNA polymerase (Sigma-Aldrich, St. Louis, MO, USA). The PCR primers targeted a portion of the SSU rRNA gene containing the hypervariable V4 region, with the reverse primers including a 12-bp barcode (Table S1) (Caporaso et al., 2011b). We note that the 3' end of the forward PCR primer used in this assay (515F, GTGCCAGCMGCCGCGGTAA) is identical to the one used in the bioinformatics analyses described in this report (SSU-517F, GCCAGCAGCCGCGGTAA); thus, amplification for both primers begins at the same place. PCR primers were only frozen and thawed once. Thermal cycling parameters were 94°C for 5 min; 35 cycles of 94°C for 20 s, 50°C for 20 s, and 72°C for 30 s, and followed by 72°C for 5 min. PCR products were purified using the MinElute 96 UF PCR Purification Kit (Qiagen, Valencia, CA, USA). DNA sequencing was performed using an Illumina HiSeq 2000 (Illumina, Inc, San Diego, CA). Clusters were created using template concentrations of 1.9 pM and PhiX at 1.37 pM; PhiX is recommended by the manufacturer for samples with uneven distributions of A, C, G and T. One hundred base sequencing reads of the 5' end of the amplicons and seven base barcode reads were obtained using the sequencing primers listed in Table S1.

2.3. ITS assay

One hundred microliter amplification reactions were performed in an MJ Research PTC-200 thermal cycler (Bio-Rad Inc, Hercules, CA, USA) containing: 50 mM Tris (pH 8.3), 500 µg/ml bovine serum albumin (BSA), 2.5 mM MgCl₂, 250 µM of each deoxynucleotide triphosphate (dNTP), 400 nM of the forward PCR primer, 200 nM of each reverse PCR primer, 4 µl of DNA template, and 2.5 units JumpStart Taq DNA polymerase (Sigma-Aldrich, St. Louis, MO, USA). PCR primers targeted a portion of the SSU and LSU rRNA genes and the hypervariable internal spacer region, with the reverse primers including a 12-bp barcode (Table S2); primer binding sites are the reverse and complement of the commonly used SSU primer 1492R (Frank et al., 2008) and LSU primer 129F (Hunt et al., 2006). PCR primers were only frozen and thawed once. Thermal cycling parameters were 94°C for 5 min; 35 cycles of 94°C for 20 sec, 56°C for 20 s, and 72°C for 40 s, and followed by 72°C for 5 min. PCR products were purified using the MinElute 96 UF PCR Purification Kit (Qiagen). DNA sequencing was performed using an Illumina HiSeq 2000 (Illumina, Inc). Clusters were created using template concentrations of 2.5 pM. One hundred base

sequencing reads of the 5' end of the amplicons and seven base barcode reads were obtained using the sequencing primers listed in Table S2.

2.4. Sequence processing

QIIME (Caporaso et al., 2010b) was used to process reads from both ITS and SSU Illumina assays, which contained a total of 85 and 80 samples, respectively, from various microbial habitats in mice. For this study, we chose the largest subsets of comparable samples between the two assays, which were fecal samples originating from the same 14 mice (seven each of C3H and C3H-IL10-KO). Demultiplexing and removal of low-quality sequences was performed with default parameters of: minimum Q-score = 3, maximum number of N characters allowed = 0, maximum number of consecutive low-quality base calls allowed before truncating a read = 3. The minimum number of consecutive high-quality base calls to include a read as a fraction of input length = 0.99, which is higher than the default of 0.75. All filtered reads had a length of 101 bp. For the ITS assay, the total initial number of sequencing reads was 71,281,079. Numbers of sequences removed using the aforementioned quality control parameters were: barcode errors (3,923,456), reads too short after quality truncation (6,728,482), and too many Ns (21,520), leaving a total of 60,607,621 filtered reads. For the 14 fecal samples used in this study, filtered read counts ranged from 312,405 to 1,403,065. For the SSU assay, the total initial number of sequencing reads was 66,148,010. Numbers of sequences removed using the aforementioned quality control parameters were: barcode errors (3,559,863), reads too short after quality truncation (7,118,807), and too many Ns (16,866,051), leaving 38,603,289 filtered reads. For the 14 fecal samples used in this study, filtered read counts ranged from 384,301 to 603,227.

2.5. Simulated PCR

Simulated PCR was performed on annotated bacterial genome sequences with Geneious 6.1 (Biomatters Ltd) using the primers and number of allowed mismatches listed in Table 1.

2.6. Resolving power

For each region and simulated PCR amplicon length (100 bp and 400 bp), we calculated the average pairwise sequence difference of amplicons arising from all species within each genus, and all subspecies within each species. Only bacteria having at least two species were selected. Differences were counted at equal weighting after pairwise alignment, and included any character difference between aligned sequences. Resolving power for OTUs was determined using the original taxonomy of the representative sequence from each OTU.

2.7. Taxonomic database and phylogenetic tree

A sequence database was created from annotated bacterial genomes to contain both SSU and ITS regions (herein referred to as the SSU-ITSdb). Briefly, using custom Perl scripts, annotated genomic sequences spanning both regions were downloaded from GenBank (NCBI) with their corresponding taxonomic information and prepared for use by QIIME scripts to assign taxonomy. To impute phylogeny to ITS and SSU reads for performing comparative UniFrac analyses on both regions, a phylogenetic tree (herein referred to as the SSU-ITSpt) was constructed from the SSU portions of the sequences in the SSU-ITSdb.

Briefly, custom Perl scripts and Geneious 6.1 (Biomatters Ltd, Auckland, New Zealand) were used to extract SSU regions between the 27F (AGAGTTTGATCMTGGCTCAG) and 1492R (GGTTACCTTGTTACGACTT) universal bacterial primers. Following the recommended procedure for creating a phylogenetic tree with QIIME, the extracted SSU sequences were aligned with PyNAST (Caporaso et al., 2010a) against the greengenes core set and filtered using the greengenes lane mask (DeSantis et al., 2006). Finally, the SSU-ITSpt was constructed from the filtered alignment using FastTree (Price et al., 2009).

2.8. Mock bacteria communities and taxonomic assignments

The FASTA sequences of the SSU-ITSdb were used to create multiple mock reference databases (MRD) and mock bacterial communities (MBC). The SSU-ITSdb is comprised of 7,813 SSU-ITS operons from 536 genera, 1074 species and 1,795 subspecies, and range in size from 1,077 to 10,432 bp. From the SSU-ITSdb, ten 20-fold cross-validation sets were created whereby approximately 95% and 5% portions were used as an MRD and paired MBC, respectively. Custom Perl scripts and Geneious 6.1 (Biomatters Ltd) were used to simulate PCR and extract sequence from the MBCs for each region, extending 100 bp or 400 bp from the forward primers of primer pairs SSU-341F and SSU-1406R, SSU-517F and SSU-1406R, and ITS-1507F and 23SR (see Table 1). By iteratively using each MRD as a reference database, taxonomic assignments were made on its paired MBC for each region with QIIME using the RDP Classifier (Wang et al., 2007) at a confidence threshold of 80%. Taxonomic accuracy was assessed by comparing original and assigned taxonomies for each region.

2.9. Beta diversity comparisons of ITS and SSU Illumina assays

Using QIIME, principal coordinates analyses (PCoA) of unweighted UniFrac beta diversity distances between the fecal bacterial communities of seven C3H and seven C3H-IL10-KO mice were performed for both the SSU and ITS regions. The results of PCoA analyses were then compared in a Procrustes analysis. Briefly, to equalize sample sizes of the fecal communities, 300,000 reads from each were extracted randomly and placed into two separate files (by region), resulting in a total of 2.1 M reads for each region. Closed-reference OTUs were picked from these using USEARCH (Edgar, 2010) to map all reads to the SSU-ITSdb. UniFrac beta diversity distances were calculated using the SSU-ITSpt as a common phylogenetic reference tree for the representative sequences of both ITS and SSU OTUs.

3. Results and discussion

3.1. Resolving power

To show the potential for improved resolving power of our ITS HTS method, we compared the average sequence distances in 100 bp and 400 bp simulated PCR amplicons from the ITS region and two commonly used SSU regions. From 1,795 subspecies within 536 genera with fully sequenced and annotated genomes available in GenBank (Table S3), we calculated the average pairwise sequence difference of simulated PCR amplicons arising from all species within each genus and all subspecies within each species, where $n \geq 2$ species or subspecies, respectively. Comparisons of species within each genus (Figure 1A and 1C) and

subspecies within each species (Figure 1B and 1D) showed significantly greater within-group average sequence distances between the amplicons of the ITS region compared to the SSU regions.

Clustering reads into OTUs is a commonly used step when analyzing the sequencing data of bacterial communities. To show how the sequence differences of Figure 1 are distributed among different bacteria and can affect resolving power, we created OTUs at a range of similarity thresholds with UCLUST (Edgar, 2010) using the same simulated PCR amplicons, and then determined how many of the original 1,795 subspecies could be detected in them. The results are shown in Figure 2 for both 100 bp and 400 bp read lengths. The two most obvious features in these results are that resolving power in the ITS region is superior to both SSU regions at all thresholds, and that 400 bp reads are better than 100 bp reads. Table 2 shows a more focused view of these results for the 100%, and more commonly used 97% thresholds. When clustering 100 bp reads at 100%, the fraction of detected-to-actual subspecies (resolving power) for the SSU-341F, SSU-517F and ITS-1507F regions, respectively, was 48%, 53% and 68%, and for the 400 bp reads it was 68%, 65% and 81%. These results highlight the difficulty of resolving bacteria with small amplicons from any region, as they are identical in many subspecies and are therefore indistinguishable. However, the ITS region fared better, and by a margin that outweighed the slightly smaller number of subspecies represented in its post-PCR amplicons – even slightly outperforming both SSU region's 400 bp reads with its 100 bp reads (Table 2). Unsurprisingly, resolving power decreased for all three regions when clustering at 97%, but the ITS region retained a clear advantage, surpassing the SSU regions by even wider margins than at the 100% threshold (Table 2).

While clustering below 97% similarity has little or no practical application, the greater resilience of the ITS region's resolving power can be more clearly seen at the lower thresholds, and provides insight into the magnitude and extent of its sequence variation across bacterial taxa compared to the SSU regions. For example, when examining resolution of the clustered reads at the genus level in 100 bp OTUs at 90% similarity (Figure 3), 515 genera can still be resolved in the ITS region versus only 178 and 276 for the SSU-341F and SSU-517F regions, respectively. Such high resolving power at similarities this low provides an indication of how robust the sequence differences are between taxa. And, although impossible to quantify precisely, it also implies that OTUs made at the more commonly used 97% threshold will be much less likely to contain amplicons from multiple genera (or any other level) than their SSU counterparts.

Some genera had no sequences amplified in the simulated PCRs, making them undetectable. This can be seen in the 400 bp reads at 100% similarity in Figure 3. Here, the number of resolvable genera in the SSU-341F, SSU-517F and ITS-1507F regions was 527, 535 and 525, respectively, out of a total of 536 (Table S3). The undetected genera in the SSU-341F and ITS-1507F regions had no simulated PCR amplicons from any representative species. However, a majority of these genera contained only one species in the data, though one not amplified in the ITS region had nine (Table S4). The single undetected genus in the SSU-517F region did have simulated PCR amplicons but they were indistinguishable from those of another genus.

Although missing amplicons adversely affect resolving power, the largest negative impact by far stemmed from the presence of identical or highly similar amplicons among divergent bacterial species. For instance, compared to the results shown above for the 400 bp reads at 100% similarity, resolvable genera at 97% similarity dropped from 527 to 504, and from 537 to 487 in the SSU-341F and SSU-517F regions, respectively, but only dropped from 525 to 523 in the ITS-1507F region.

Overall, these analyses show that a considerable increase in resolving power is possible by employing the more variable ITS region.

3.2. Taxonomic assignments of mock communities

Although the motivation of this research was to develop an HTS method with increased resolving power, we also examined the ability of ITS sequences to be accurately assigned at various taxonomic levels.

Comparison of taxonomic assignment accuracy was performed using mock reference databases to assign taxonomy to 100 bp and 400 bp simulated-PCR amplicons from ITS and SSU regions of mock bacterial communities. Figure S1 shows the results of the three regions at each taxonomic level. At the species level, assignments to the more variable ITS region were correct much more often than the two SSU regions in 100 bp reads, but only slightly more often in 400 bp reads. At the genus level in 100 bp reads, correct assignments for SSU-517F match those of the ITS while the SSU-314F still lags. In 400 bp reads, however, and at higher taxonomic levels, the ITS begins to perform progressively worse compared to the SSU regions, essentially levelling off from the family level onward while the SSU regions continue to improve. All reads produced from the simulated PCR are eventually assigned correctly at the Kingdom level as the sequences were drawn from and assigned with the SSU-ITSdb, which contains only bacterial sequences. Assignments to the subspecies level were not made because only one representative of each subspecies was present in the data, and each was either a part of the reference database or the mock community but not both.

These assignment results, especially at the higher taxonomic levels, are not surprising considering that the ITS region is not conserved enough to provide the deep phylogenetic information that is available in the SSU. This fact, coupled with the tendency of the ITS region to have higher sequence variation between species than the SSU, is also likely to make taxonomic assignment of ITS reads more sensitive to an incomplete reference database; for more distantly related bacteria, any correlation that may exist between sequence similarity and taxonomic similarity will tend to be lower for the ITS than for the SSU, thus increasing the odds of no assignment or a very wrong assignment being made whenever no closely related bacterial sequences exist in the reference database. Yet, for the same reasons, if the sequence of the same or a closely related bacterium is present in the reference database, taxonomic assignment accuracy for the ITS tends to be superior to the SSU, and seen most clearly at the species level.

It is important to emphasize that accuracy in taxonomic assignments is not the same thing as distinguishing power. Taxonomic assignment accuracy is, in large part, dependent upon the

completeness of the reference database. The properties of the sequences being assigned can have an impact as well, such as the amount of variation between them, and the previously mentioned phylogenetic information that they may or may not contain. In contrast, distinguishing power is dependent solely upon sequence differences between the reads produced in an assay.

3.3. Beta diversity comparisons of ITS and SSU assays

ITS reads can also be analyzed at the community-level by employing the SSU-ITSdb and its mated phylogenetic tree, the SSU-ITSpt. We demonstrated this in a Procrustes analysis comparing the PCoA results of unweighted UniFrac beta diversity distances, calculated separately on ITS and SSU reads derived from the fecal bacteria communities of two mouse genotypes ($M2 = 0.147$, $P < 0.0005$) (Figure S2). These animals are used as a model to study inflammatory bowel disease (IBD): C3H-IL10-KO mice develop intestinal colitis over time while the C3H mice do not. Both ITS and SSU assays reveal a similar ability to distinguish between the bacterial communities of C3H (red) and C3H-IL10-KO mice (blue), though the percent explained by the first of three principal components in the PCoAs differed (PC1:65.1, PC2:8.4 and PC3:5.4 for the ITS, and PC1:49.4, PC2:8.2 and PC3:6.9 for the SSU). We present this as a proof-of-concept experiment rather than an alternative to current methods, however, as the availability of SSU-ITS-LSU sequences is currently very small compared to existing SSU databases.

4. Conclusions

In sum, this report presents a new HTS method for analyzing bacterial composition with improved resolving power. The greater resolving power enabled by the ITS region arises from its high sequence variation across a wide range of bacterial taxa and an associated decrease in taxonomic heterogeneity within its OTUs (Figures 2 and 3). This feature enables various analyses to be performed at finer levels, such as efforts to correlate specific phylotypes with host or environmental traits and/or monitoring population shifts, or assessing microgeographical relationships through use of sequence-selective PCR primers and probes. We anticipate that these capabilities will contribute to a better understanding of systems where bacteria play important roles. Potential shortcomings include the possibility of compositional skew arising from differential PCR efficiencies of the variably-sized ITS amplicons (Table S5) and (Ihrmark et al., 2012). In addition, the method will not detect bacteria that do not have the typical rRNA gene order (SSU-ITS-LSU). However, our simulated PCR results indicate that at least 97.9% of the bacterial subspecies analyzed have the necessary (typical) order (Table S3).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research is supported in part by NIH grant AI078885 and Crohn's and Colitis Foundation of America grant 3153.

Abbreviations

ARISA	Automated ribosomal intergenic spacer analysis
HTS	high throughput sequencing
ITS	internal transcribed spacer
LSU	large subunit
OTU	operational taxonomic unit
PCoA	principle coordinates analysis
RISA	ribosomal intergenic spacer analysis
SSU	small subunit
SSU-ITSdb	a sequence database was created from annotated bacterial genomes to contain both SSU and ITS regions
SSU-ITSpt	a phylogenetic tree used to impute phylogeny to ITS and SSU reads for performing comparative UniFrac analyses on both regions

References

- Barry T, Collieran G, Glennon M, Dunican LK, Gannon F. The 16S/23S ribosomal spacer region as a target for DNA probes to identify eubacteria. *PCR Methods Appl.* 1991; 1:51–56. [PubMed: 1726852]
- Borneman J, Triplett EW. Molecular microbial diversity in soils from eastern Amazonia: evidence for unusual microorganisms and microbial population shifts associated with deforestation. *Appl Environ Microbiol.* 1997; 63:2647–2653. [PubMed: 9212415]
- Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics.* 2010a; 26:266–267. [PubMed: 19914921]
- 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. *Nat Meth.* 2010b; 7:335–336.
- Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon JI, Knight R. Moving pictures of the human microbiome. *Genome Biol.* 2011a; 12:R50. [PubMed: 21624126]
- 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 USA.* 2011b; 108:4516–4522. [PubMed: 20534432]
- Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. *Science.* 2009; 326:1694–1697. [PubMed: 19892944]
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ Microbiol.* 2006; 72:5069–5072. [PubMed: 16820507]
- Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* 2010; 26:2460–2461. [PubMed: 20709691]
- Fisher MM, Triplett EW. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol.* 1999; 65:4630–4636. [PubMed: 10508099]

- Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA, Olsen GJ. Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl Environ Microbiol.* 2008; 74:2461–2470. [PubMed: 18296538]
- Frothingham R, Wilson KH. Sequence-based differentiation of strains in the *Mycobacterium avium* complex. *J Bacteriol.* 1993; 175:2818–2825. [PubMed: 8491701]
- Hunt DE, Klepac-Ceraj V, Acinas SG, Gautier C, Bertilsson S, Polz MF. Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. *Appl Environ Microbiol.* 2006; 72:2221–5. [PubMed: 16517676]
- Ihrmark K, Bödeker IT, Cruz-Martinez K, Friberg H, Kubartova A, Schenck J, Strid Y, Stenlid J, Brandström-Durling M, Clemmensen KE, Lindahl BD. New primers to amplify the fungal ITS2 region--evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiol Ecol.* 2012 Dec; 82(3):666–77. [PubMed: 22738186]
- McLaughlin GL, Howe DK, Biggs DR, Smith AR, Ludwinski P, Fox BC, Tripathy DN, Frasci CE, Wenger JD, Carey RB, Hassan-King M, Vodkin MH. Amplification of rDNA loci to detect and type *Neisseria meningitidis* and other eubacteria. *Mol Cell Probes.* 1993; 7:7–17. [PubMed: 8455644]
- Price LB, Liu CM, Melendez JH, Frankel YM, Engelthaler D, Aziz M, Bowers J, Rattray R, Ravel J, Kingsley C, Keim PS, Lazarus GS, Zenilman JM. Community analysis of chronic wound bacteria using 16S rRNA gene-based pyrosequencing: impact of diabetes and antibiotics on chronic wound microbiota. *PLoS One.* 2009; 4:e6462. [PubMed: 19649281]
- Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 2009; 26:1641–1650. [PubMed: 19377059]
- Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol.* 2007; 73:5261–5267. [PubMed: 17586664]
- Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, Bittinger K, Hwang J, Chen J, Berkowsky R, Nessel L, Li H, Bushman FD. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol.* 2010; 10:206. [PubMed: 20673359]

Highlights

New method for analyzing bacterial composition with improved resolving power.

Greater resolving power is enabled by sequence analysis of the rRNA ITS region.

Greater resolving power enables population analyses to be performed at finer levels.

Figure 1.

Average sequence distances of SSU and ITS simulated PCR amplicons. Distances between 100 bp amplicons of, (A) species within a genus, and (B) subspecies within a species are shown. Likewise, (C) and (D) are for 400 bp amplicons. Data were from all bacterial genera ($n = 165\text{--}167$ per assay) having successfully simulated PCR reads from at least two species in the annotated genomes available in GenBank (NCBI). Bars are standard error. *Indicates column was statistically different than the other two using the Kruskal-Wallis test.

Figure 2.

Subspecies level resolving power in OTUs by region and similarity thresholds.

Simulated PCR amplicons from annotated bacterial genomes, of size (A) 100 bp or (B) 400 bp, were clustered into operational taxonomic units (OTUs) from 60% to 100% sequence similarity with UCLUST(18). The taxonomy of the representative sequence of each OTU was used to determine the number of detectable subspecies.

Figure 3.

Genus level resolving power in OTUs by region and similarity thresholds.

Simulated PCR amplicons from annotated bacterial genomes, of size (A) 100 bp or (B) 400 bp, were clustered into operational taxonomic units (OTUs) from 60% to 100% sequence similarity with UCLUST(18). The taxonomy of the representative sequence of each OTU was used to determine the number of detectable subspecies.

Table 1

Primers for simulated PCR

Primer Name	Sequence	Length	Allowed Mismatches
SSU-341F	CCTACGGGNGGCNGCA	16	2
SSU-517F	GCCAGCAGCCGCGGTAA	17	2
SSU-1406R	GACGGGCGGTGWGTRCA	17	2
ITS-1507F	GGTGAAGTCGTAACAAGGTA	20	3
ITS-23SR	GGGTTBCCCCATTCTRG	16	2

SSU-1406R was used as the reverse primer for both SSU forward primers. The criteria for a successful hybridization of a primer to genomic DNA was: mismatches $\leq$ allowed mismatches, length of match = primer length.

Table 2

The effect of clustering on resolving power

Amplicon Size (bp)	Simulated PCR					
	100			400		
Primer Pairs	SSU-3 41F 1406R	SSU-5 17F 1406R	ITS-1 507F 23SR	SSU-3 41F 1406R	SSU-5 17F 1406R	ITS-1 507F 23SR
Subspecies pre-PCR	1795	1795	1795	1795	1795	1795
Subspecies post-PCR	1758	1792	1757	1758	1792	1757
Amplification rate	97.9%	99.8%	97.9%	97.9%	99.8%	97.9%
Clustering post-PCR @100% similarity						
Number of OTUs	943	1051	2112	1716	1524	3504
Subspecies detected in OTUs	860	956	1229	1219	1157	1448
Resolving power	47.9%	53.3%	68.5%	67.9%	64.5%	80.7%
Clustering post-PCR @97% similarity						
Number of OTUs	556	680	1364	767	676	1995
Subspecies detected in OTUs	547	670	1017	756	671	1138
Resolving power	30.5%	37.3%	56.7%	42.1%	37.4%	63.4%

Simulated PCR amplicons from annotated bacterial genomes, of size 100 bp or 400 bp, were clustered into operational taxonomic units (OTUs) at both 100% and 97% sequence similarity. From the initial 1795 subspecies, SSU-517F primers amplified sequences from all but three, while SSU-341F and ITS primers missed 37 and 38 subspecies, respectively. The results when clustering at 100% similarity reveal the difficulty of resolving bacteria with small amplicons, as they are identical in many subspecies and are therefore indistinguishable. However, resolving power in the ITS region fairs better by a margin that outweighs the smaller number of subspecies amplified initially. Not surprisingly, all three regions show a decrease in resolving power when clustering at 97% versus 100% similarity, but the ITS region retains a clear advantage, besting the SSU regions by even wider margins.