Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences: Supplementary Materials

David A. W. Soergel, Neelendu Dey, Rob Knight, and Steven E. Brenner

November 18, 2011

Contents

1 Database coverage of environments

The validation procedure we developed holds out an entire study at a time, simulating the situation that each study was not yet incorporated in the reference database and needed to be classified. Figure S1 shows the proportion of sequences in each tested sample that are within a given percent identity of a sequence in the remainder of GreenGenes. Of course, now that the studies we evaluated are in GreenGenes, future samples of similar composition will be easier to classify than we report. This is especially relevant for studies that are the only representative in GreenGenes of a given environment type, such as the hypersaline mat sample. We would expect a second hypersaline mat sample to classify a good deal better than reported here, because matches can now be made to sequences from the first sample (assuming, of course, that those sequences now carry manually curated taxonomic annotations, including taxa that were not previously present in the database, and assuming that the second sample bears any resemblance to the first). But, many other environments remain as poorly represented now as hypersaline mat was previously. Thus our results are particularly sobering with respect to whatever environment types are underrepresented in the database at the moment that it is used as a basis for classification.

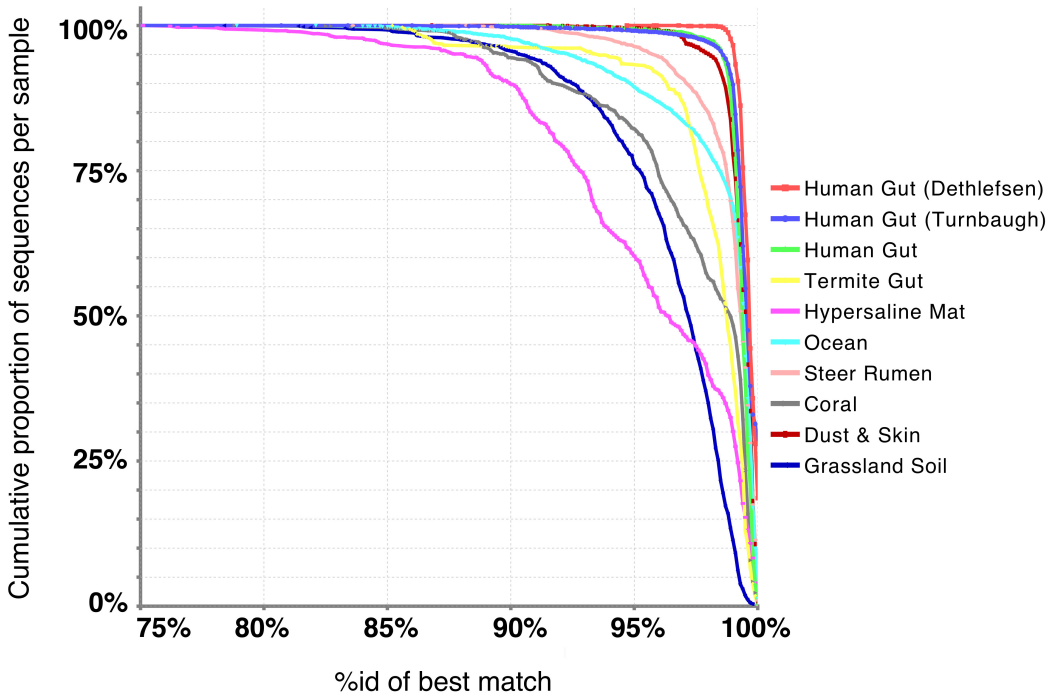

Figure S1: Coverage of different environments by GreenGenes. The plot shows the distribution of percent identity scores between environmental sequences and their closest matches in the remainder of GreenGenes (with each respective sample held out). The distributions are presented cumulatively from right to left, so that the Y value indicates the proportion of each sample that is within a given distance of any reference sequence. We see that GreenGenes provides excellent coverage of human gut and skin samples, but relatively poor coverage of the grassland soil and hypersaline mat samples. This plot is based on the full GreenGenes database (version of August 25, 2010), which had not been subjected to the 2-study filter or other filters for potential chimeras, as the version of GreenGenes used for the primary analysis was. It can be seen here, for instance, that more than half of the sequences from the hypersaline mat sample have no match within 97% identity in the rest of GreenGenes. Sequences from such 97% OTUs that are unique to one sample are not included in the main analysis.

2 Choice of query datasets

We wished to test all primer and read length combinations using consistent sets of underlying sequences. We therefore sought data sets containing many near-full-length sequences from the same environment. We defined "near-full-length" as including hypervariable regions V1 through V9 (specifically, extending from positions 69 to 1465 in E. coli coordinates). We wished the datasets to be as large as possible in order to limit stochastic variation in the proportions of sampled taxa, and so that rare species would be represented.

We downloaded the GreenGenes database (version of March 11, 2011) and identified eight appropriate studies contained within it, representing a variety of environments, shown in table S1.

We considered only those sequences passing the chimera filters of (McDonald et al., 2011). These filters included the requirement that each sequence be a member of an OTU, defined at 97% identity, which was present in at least two different samples. While some of the sequences excluded on this basis were likely indeed chimeras, the filter presumably also excluded legitimate sequences from OTUs that were unique to one sample. This circumstance is particularly likely when only one sample is available from a given environment type, as in the case of the hypersaline mat. Naturally, the excluded sequences would have been more difficult to classify because they lack a close match in the reference database. Hence, excluding them has the effect of increasing the apparent classification accuracy (measured with respect to the remaining sequences) compared to what would have been obtained from the whole sample. On the other hand, chimera filters are regularly applied to new environmental samples before classification.

3 Choice of primers

50 forward and 44 reverse primer sequences were obtained from from a survey of literature on primer choice. These were aligned to the 1541nt *E. coli* 16S sequence to confirm appropriate naming. The primers were also mapped to the 7682-column NAST coordinates (Desantis et al., 2006) by alignment to all GreenGenes sequences. In many cases, the primers began and ended in slightly different NAST columns (+-1-5nt) in different sequences, suggesting that there are errors in the GreenGenes NAST alignment; we therefore report the column with the largest number of hits. Our initial survey included 94 primers, but many of these were specific to Archaea or for some other reason hit a small fraction of sequences in the environments we tested. Here, we selected all 22 forward and 22 reverse primers that hit at least 40% of the sequences in at least one of the query datasets (Tables S2 and S3).

For PCR or paired-end sequencing, the selected primers could be combined into 374 viable pairs for very short reads; as the read length increases, pairings spaced more closely than the read length become unviable.

Environment	GreenGenes StudyID	All sequences		Chimera-free, non-unique sequences				Original citation
		Total	Near-full-length (NFL)	Total	Near-full-length (NFL)	Chimera-free percentage of all NFL	NFL with genus annotation	
Human Gut	42600	7013	6997	6776	6760	97%	5844	(Li et al., 2008)
Mattress Dust and Human Skin	42612	3192	3184	3042	3035	95%	2851	(Täubel et al., 2009)
Steer Rumen	42623	3251	1973	2892	1831	93%	558	(Brulc et al., 2009)
Termite Gut	33327	1208	1150	537	505	44%	379	(Warnecke et al., 2007)
Ocean	35248	6062	5226	4338	3985	76%	2388	(Shaw et al., 2008)
Coral	35251	1600	1531	1044	995	65%	192	(Sunagawa et al., 2009)
Hypersaline Mat	31588	1278	1242	525	510	41%	102	(Isenbarger et al., 2008)
Grassland Soil	42615	1021	918	687	654	72%	193	(Cruz-Martínez et al., 2009)

- Li, M., Wang, B., Zhang, M., et al. (2008). Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci U S A*, 105(6), 2117–22.
- Täubel, M., Rintala, H., Pitkäranta, et al. (2009). The occupant as a source of house dust bacteria. *J Allergy Clin Immunol*, 124(4), 834–40.e47.
- Brulc, J. M., Antonopoulos, D. A., Berg Miller, M. E., et al. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *PNAS*, 106(6), 1948.
- Warnecke, F., Luginbühl, P., Ivanova, N., et al. (2007). Metagenomic 19 and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450(7169), 560–5.
- Shaw, A. K., Halpern, A. L., Beeson, K., et al. (2008). It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol*, 10(9), 2200–2210.
- Sunagawa, S., Desantis, T. Z., Piceno, Y. M., et al. (2009). Bacterial diversity and white plague disease-associated community changes in the caribbean coral montastraea faveolata. *ISME J*.
- Isenbarger, T. A., Finney, M., Rios-Velazquez, C., et al. (2008). Miniprimer pcr, a new lens for viewing the microbial world. *Applied and environmental microbiology*, 74(3), 840.
- Cruz-Martínez, K., Suttle, K. B., Brodie, E. L., et al. (2009). Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. *ISME J*.

Table S1: Test datasets. The chimera-free columns represent sequences that have been filtered by various chimera-detection mechanisms, including requiring a $\geq$97% identical match in another study (McDonald et al., 2011). Sequences failing these filters were not included in our query datasets, leading to apparently higher classification performance than would have been seen using unfiltered sequences.

Name	Alternate names	Sequence	Source	Length	E. coli 5'	E. coli 3'	NAST 5'	NAST 3'
E8Fa	E8F, E27F	AGAGTTTGATCCTGGCTCAG	(Wuyts et al., 2002; Baker et al., 2003)	20	8	27	108	136
E8Fb	8F	AGAGTTTGATCMTGGCTCAG	(Youssef et al., 2009; Dethlefsen et al., 2008)	20	8	27	108	136
E9F		GAGTTTGATCCTGGCTCAG	(Baker et al., 2003)	19	9	27	109	136
E334F		CCAGACTCCTACGGGAGGCAGC	(Baker et al., 2003)	22	334	355	1864	1897
E338F		ACTCCTACGGGAGGCAGC	(Youssef et al., 2009)	18	338	355	1868	1897
E341F		CCTACGGGNGGCNGCA	(Baker et al., 2003)	16	341	356	1872	1899
U341F		CCTACGGGRSGCAGCAG	(Baker et al., 2003)	17	341	357	1872	1901
E343F		TACGGRAGGCAGCAG	(Wuyts et al., 2002)	15	343	357	1875	1901
E349F		AGGCAGCAGTGGGGAAT	(Wuyts et al., 2002)	17	349	365	1886	1916
U515F		GTGCCAGCMGCCGCGGTAA	(Baker et al., 2003)	19	515	533	2227	2263
E517F		GCCAGCAGCCGCGGTAA	(Wuyts et al., 2002)	17	517	533	2231	2263
U519F		CAGCMGCCGCGGTAATWC	(Baker et al., 2003)	18	519	536	2233	2268
E786F		GATTAGATACCCTGGTAG	(Baker et al., 2003)	18	786	803	4050	4081
Eb787F		ATTAGATACCCTGGTA	(Baker et al., 2003)	16	787	802	4052	4079
E805F		GGATTAGATACCCTGGTAGTC	(Youssef et al., 2009)	17	805	821	4049	4088
E917F		GAATTGACGGGGRCCC	(Wuyts et al., 2002)	16	917	932	4542	4563
E967F		CAACGCGAAGAACCTTACC	(Youssef et al., 2009)	19	967	985	4624	4653
E969F		ACGCGARRAACCTTACC	(Alan Walker, pers. comm.)	17	969	985	4626	4653
E1046F		AGGTGCTGCATGGCTGT	(Youssef et al., 2009)	16	1046	1061	4929	4955
U1053F		GCATGGCYGYCGTCAG	(Baker et al., 2003)	16	1053	1068	4940	4964
E1099F		GYAACGAGCGCAACCC	(Wuyts et al., 2002)	16	1099	1114	5012	5042
E1391F		TGTACACACCGCCCGTC	(Wuyts et al., 2002)	17	1391	1407	6427	6450

Table S2: Forward primers. *E. coli* and NAST columns indicate the start and end positions in forward-strand *E. coli* coordinates (out of 1541 columns) and NAST coordinates (out of 7682 columns), respectively, based on our alignment of the primer sequences to those references. We have named primers consistently according to their actual position in the *E. coli* sequence; we indicate a few cases where other names are in common usage.

Name	Alternate names	Sequence	Source	Length	E. coli 5'	E. coli 3'	NAST 5'	NAST 3'
E65R		TCGACTTGCATGTRTTA	(Wuyts et al., 2002)	17	49	65	176	200
E355R		GCTGCCTCCCGTAGGAGT	(Youssef et al., 2009)	15	341	355	1868	1897
E357R		CTGCTGCCTYCCGTA	(Wuyts et al., 2002)	15	343	357	1875	1901
U529R		ACCGCGGCKGCTGGC	(Baker et al., 2003)	15	515	529	2231	2260
E533Ra		TNACCGNNNCTNCTGGCAC	(Baker et al., 2003)	19	515	533	2227	2263
E533Rb	515R	TTACCGCGGCTGCTGGCAC	(Cho et al., 1996; Dethlefsen et al., 2008)	19	515	533	2227	2263
E534R		ATTACCGCGGCTGCTGGC	(Wuyts et al., 2002)	18	517	534	2231	2264
U534R		GWATTACCGCGGCKGCTG	(Baker et al., 2003)	18	517	534	2233	2268
E826R		GACTACCAGGGTATCTAATCC	(Youssef et al., 2009)	15	812	826	4049	4088
E926Ra	E926R	CCGNCNATTNNTTTNAGTTT	(Baker et al., 2003)	20	907	926	4521	4554
U926R		CCGTCAATTCCTTTRAGTTT	(Baker et al., 2003)	20	907	926	4521	4554
E926Rb		CCGTCAATTYYTTTRAGTTT	(Wuyts et al., 2002)	20	907	926	4521	4554
E939R		CTTGTGCGGGCCCCCGTCAATTC	(Baker et al., 2003)	23	917	939	4542	4580
E1064R	1046R	CGACARCCATGCASCACCT	(Dethlefsen et al., 2008)	19	1046	1064	4929	4958
E1065R		ACAGCCATGCAGCACCT	(Youssef et al., 2009)	19	1047	1065	4929	4955
E1114R		GGGTTGCGCTCGTTRC	(Wuyts et al., 2002)	16	1099	1114	5012	5042
E1115R		AGGGTTGCGCTCGTTG	(Baker et al., 2003)	16	1100	1115	5013	5044
E1238R		GTAGCRCGTGTGTMGCCC	(Youssef et al., 2009)	18	1221	1238	5883	5910
U1406R	1391R	GACGGGCGGTGTGTRCA	(Baker et al., 2003; Dethlefsen et al., 2008)	17	1390	1406	6427	6450
E1406R		GACGGGCGGTGWGTRCA	(Youssef et al., 2009)	17	1390	1406	6427	6450
E1407R		GACGGGCGGTGTGTRC	(Wuyts et al., 2002)	16	1392	1407	6428	6450
E1492R		ACCTTGTTACGACTT	(Youssef et al., 2009)	15	1478	1492	6792	6809

Table S3: Reverse primers. *E. coli* and NAST columns indicate the start and end positions in forward-strand *E. coli* coordinates (out of 1541 columns) and NAST coordinates (out of 7682 columns), respectively, based on our alignment of the primer sequences to those references. We have named primers consistently according to their actual position in the *E. coli* sequence; we indicate a few cases where other names are in common usage.

4 Preparation of reference databases

For each query data set, we built a reference database based on chimera-free GreenGenes, excluding all sequences from the same study as the query sequences (whether near-full-length or not).

We used the GreenGenes taxonomy (McDonald et al., 2011) to provide the taxonomic identity of the reference sequences.

Each reference database was dereplicated at 99% using UCLUST 4.1.93 (Edgar, 2010) such that for any cluster of sequences with $\geq$99% identity only the most abundant sequence was used. This reduced each database from approximately 500,000 sequences to approximately 140,000 representatives, thereby correcting for database bias at the strain level, and substantally improving performance of the downstream analyses.

The taxonomic identity of each reference cluster was usually unambiguous. For the occasional cluster containing sequences differing in taxonomic classification, we assigned taxonomic position at the deepest rank at which over half of the clustered sequences were in agreement.

5 RTAX: Rapid and accurate taxonomic classification of short paired-end sequence reads from the 16S ribosomal RNA gene.

5.1 Introduction

The rapid advance in sequencing technology continues to motivate researchers to obtain larger numbers of shorter reads. As platforms such as the 454 GS 20, the Illumina GA IIx and HiSeq, and the Ion Torrent reach ~100 bases of reasonable-quality sequence, they begin to be applicable to taxonomic profiling of microbial communities (Sogin et al., 2006; Lazarevic et al., 2009; Claesson et al., 2010; Degnan & Ochman, 2011). A key challenge is using these short reads, especially paired-end reads that do not overlap, to obtain taxonomic assignments.

Accordingly, we report a new method of performing taxonomic classifications of non-overlapping paired-end reads from the 16S ribosomal RNA gene. This gene is frequently used to delineate bacterial and archaeal taxa, because it present in all bacteria and archaea, and has an overall mutation rate that is amenable to phylogenetic analysis (Pace, 1997; Tringe & Hugenholtz, 2008).

Existing classification methods require sequences that are contiguous (except for small indels): that is, single reads, or paired-end reads that overlap and so can be assembled prior to classification (Gloor et al., 2010; Caporaso et al., 2011; Zhou et al., 2011; Bartram et al., 2011; Ram et al., 2011). However, some current sequencing technologies (e.g., Illumina) can provide non-overlapping paired-end reads in high volume and at low cost. To date, the two ends of each sequence have been classified independently, without taking advantage of mate pair information (Caporaso et al., 2011). We therefore asked whether including this information might substantially improve classifier performance.

To address this question, we first attempted to reconcile taxonomic assignments made for each read independently; we hoped that cases in which one read provided an ambiguous assignment might be "rescued" by a precise assignment from the other read. However, we found that this procedure provided essentially no advantage in classification rate or accuracy over single-ended data. This counterintuitive result arises because existing reference databases are relatively sparse and biased. Thus, given the degree of variation present in different regions of the 16S sequence, a single short read often has seemingly optimal matches in the database that are in fact phylogenetically distant (as revealed by poor matching to the mate pair). The use of reference databases that are not filtered for chimeras can exacerbate this effect.

Classification of paired-end reads therefore requires finding reference sequences that match both reads simultaneously. Essentially, the goal is to search using paired-end reads concatenated with N's in between. However, existing sequence similarity search software such as BLAST (Altschul et al., 1990) and USEARCH (Edgar, 2010) do not perform well with this type of query.

5.2 Results and conclusions

We developed RTAX to provide accurate paired-end taxon assignment. RTAX is a script that drives and interprets results from multiple underlying USEARCH runs. Like pure USEARCH, the identification of best matches is alignment-based, but is nonetheless very fast because of an initial k-mer based filtering step.

The speed and accuracy of USEARCH depends strongly on search parameters, including the minimum percent identity (%id) threshold that is accepted to report a match. In the typical usage, a list of query sequences is provided to USEARCH; for each sequence it then returns all hits that match better than the given threshold. Because database coverage is biased towards some taxa, some query sequences have nearly perfect matches; in those cases a lower %id threshold would dilute the perfect hits and consequently decrease classification accuracy. Conversely, other query sequences have only very distant matches (although these distant matches may still be sufficiently informative to classify at higher taxonomic levels), which would be excluded by a stringent %id threshold. RTAX must thus be adaptive to different %id levels for each query sequence, taking a selection of the best available hits at any %id level.

RTAX runs two USEARCH jobs concurrently—one for each sequence region. That is: RTAX takes two FASTA files as input, containing, respectively, all reads from the forward primer and all reads from the reverse primer. Each underlying USEARCH job takes one of these files as input, and reports database matches to each read in turn. As hits to each read are found, the outputs of the two jobs are intersected to find reference sequences that matched both reads of a mate pair (on the basis of matching sequence identifiers in the forward and reverse FASTA files). For each such matching sequence, the average %id is computed. Finally, reference sequences are selected where the average %id to the query reads is within 0.5% of the maximum observed for that query sequence pair.

One technical challenge is that USEARCH requires a %id threshold per read, but assignments are improved by finding hits that have maximal average %id for the pair. Using a permissive %id threshold would allow us to comprehensively find all sequences that match both query reads, but is time-consuming because of the large number of hits returned. It is therefore much faster to take an iterative approach, starting with a stringent %id threshold. Any query sequence pairs that have good matches can be immediately classified. A list is made of those queries that do not initially produce a match. This list is used to construct input files for subsequent USEARCH runs with a less stringent threshold. The queries are thus passed through a "sieve", iteratively processing unmatched sequences with successively lower thresholds until either a match is found for each query or the lowest permissible threshold is exceeded. The advantage of this approach is that many query sequences can be processed in the early iterations (using fast, high-%id searches), so searches with a low %id threshold (which are slow and can produce voluminous output) are performed only for few query sequences—i.e., those for which no better match was found in earlier iterations.

Even with this iterative technique, and especially in the later iterations, the lists of hits matching each read are much larger than their intersection; writing these to disk as large files would be very slow. Thus, we process piped outputs from the two USEARCH processes in a streaming fashion. This approach works because input queries are processed in order, so the two USEARCH output streams can be kept synchronized (i.e., so that they provide results for mate-paired reads simultaneously). We thus find the best matches to the two reads taken jointly, using an alignment score that does not penalize the large gap between reads.

Given these hits, we examine the taxonomic annotations provided in the reference database (in our case, chimera-filtered GreenGenes with consistent taxonomic annotations: (McDonald et al., 2011)). At each taxonomic level, we accept an annotation that is present on at least 50% of the hits. Others have required 67% agreement (Huse et al., 2008; Hamp et al., 2009); we found in preliminary experiments (not shown) that classification performance is fairly insensitive to this threshold, except at the extremes.

Fully utilizing information from mate pairs in this way markedly improves classification rate and accuracy (Figure S2). In an extreme case, we found in a human gut sample that a pair of 32nt reads provided confident genus-level classifications of 90% of the sequences—far more than the 41% of the sample that could be classified using a single 64nt read, and more than could be classified using single-ended reads of any length. The apparent paradox—that a single long read might encompass both short reads, and yet provide less accurate classifications—arises from the fact that the sequence in the middle is less phylogenetically informative than that at the ends, so including it dilutes the signal.

Classification accuracy depends strongly on which region within the 16S rRNA gene is sequenced, and on the environment from which the sample was taken. Accordingly, our paired-end approach provides a benefit in some but not all circumstances, as demonstrated in the main text.

Because of the iterative procedure, the speed of classification varies with database coverage. Overall, in testing datasets from a variety of environments, we observed rates in the vicinity of 10 sequence pairs per second on an Amazon EC2 “large” instance, corresponding roughly to a 2007-era 2.4GHz dual-core Xeon machine.

The availability of RTAX as a component in QIIME (Caporaso et al., 2010) makes it broadly available to microbial ecologists studying a range of environments, and greatly increases the utility of short reads produced on Illumina and other emerging platforms.

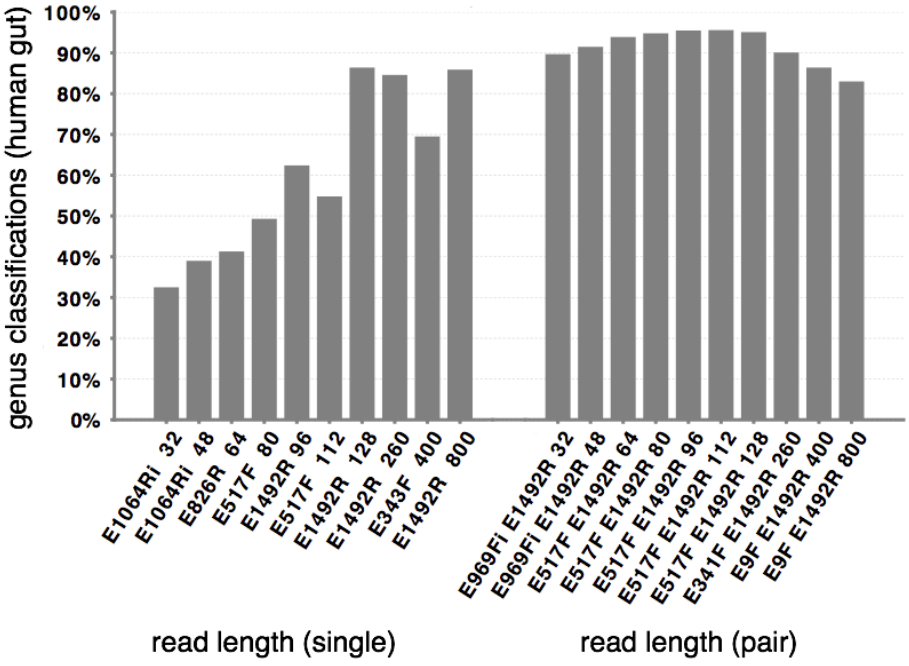

Figure S2: Maximum achievable genus-level classification rate for read lengths from 32nt through 800nt, comparing single-ended and paired-end experiments. Sequence reads from many primers were simulated from a human gut sample of near-full-length 16S rRNA gene sequences (Li et al., 2008) and classified using RTAX, taking care to exclude the query sequences from the reference database. For each read length, the primer or primer pair producing the greatest number of genus classifications is shown. Primers and primer pairs were excluded when they produced genus classifications that were less than 95% accurate with respect to GreenGenes annotations on the original full-length sequences.

6 Benefit of paired-end sequencing

Paired-end sequencing often, but not always, allows accurate classification of a greater proportion of a sample to the genus level. Tables S4 and S5 compare classification rates between single-and paired-end experiments for the same total read length, highlighting those cases where paired-end sequencing provides a benefit.

Maximum achievable genus classification rate, 80% estimated accuracy
single vs. paired-end

normal = single-ended reads
bold = paired-end reads
gray italics = no improvement over shorter reads
light green = paired-end performs better than single for same total read length
saturated green = more than 3% better
light red = paired-end performs worse than single for same total read length
saturated red = more than 3% worse

Single-end read length	Paired-end read length	Ocean (35248)		Coral (35251)		Hypersaline Mat (31588)		Grassland Soil (42615)		Steer Rumen (42623)		Human Gut (42600)		Dust & Skin (42612)		Termite Gut (33327)	
32		41		32		0		39		57		35		60		83	
48		55		44		28		49		75		66		73		*83*	
64	**32**	59	64	57	52	42	50	*49*	53	78	77	*66*	90	77	79	*83*	76
80		63		64		59		58		*78*		78		78		*83*	
96	**48**	*63*	68	*64*	67	61	58	64	59	83	82	80	92	80	80	*83*	76
112		*63*		77		*61*		69		87		81		*80*		*83*	
128	**64**	*63*	75	*77*	76	62	58	72	67	88	87	86	94	85	86	*83*	76
	80		*75*		*76*		65		75		92		95		87		*76*
	96		*75*		*76*		*65*		78		93		*95*		88		*76*
	112		78		79		*65*		*78*		*93*		96		*88*		*76*
260	**128**	*63*	*78*	*77*	*79*	*62*	*65*	*72*	*78*	91	*93*	*86*	*88*	86	*88*	*83*	76
400		*63*		*77*		*62*		*72*		*91*		*86*		*86*		*83*	
	260		*78*		81		*65*		*78*		94		*96*		89		*76*
800	**400**	66	*78*	*77*	*81*	*62*	*65*	*72*	*78*	*91*	*94*	*86*	*88*	*86*	*89*	*83*	76
	800		*78*		*81*		*65*		*78*		*94*		*96*		*89*		*76*
	Maximum	78		81		65		78		94		96		89		83	
	Max(single 96, pair 48) / Max(all)	87%		83%		94%		82%		88%		96%		90%		100%	

Table S4: Maximum achievable genus classification rate, 80% estimated accuracy, comparing paired-ending with single-ended sequencing. The classification rate shown in each cell is the maximum value observed for any choice of primers at each respective read length.

Maximum achievable genus classification rate, 95% estimated accuracy
single vs. paired-end

normal	= single-ended reads
bold	= paired-end reads
gray italics	= no improvement over shorter reads
light green	= paired-end performs better than single for same total read length
saturated green	= more than 3% better
light red	= paired-end performs worse than single for same total read length
saturated red	= more than 3% worse

Single-end read length	Paired-end read length	Ocean (35248)		Coral (35251)		Hypersaline Mat (31588)		Grassland Soil (42615)		Steer Rumen (42623)		Human Gut (42600)		Dust & Skin (42612)		Termite Gut (33327)	
32		41		12		0		0		55		33		60		83	
48		50		*12*		0		21		*55*		39		73		*83*	
64	**32**	59	**64**	*12*	**35**	28	**45**	25	**34**	78	**77**	41	**90**	77	**79**	*83*	**76**
80		63		44		*28*		*25*		*78*		49		78		*83*	
96	**48**	*63*	**66**	*44*	**35**	*28*	**56**	*25*	**57**	83	**82**	62	**92**	80	**80**	*83*	**76**
112		*63*		49		*28*		27		87		*62*		*80*		*83*	
128	**64**	*63*	**75**	*49*	**35**	44	**57**	*27*	**65**	88	**87**	86	**94**	85	**86**	*83*	**76**
	80		*75*		**52**		**65**		*65*		**92**		**95**		**87**		*76*
	96		*75*		**55**		*65*		*65*		**93**		*95*		**88**		*76*
	112		**78**		*55*		*65*		**77**		*93*		**96**		*88*		*76*
260	**128**	*63*	*76*	76	**77**	*54*	*65*	55	*77*	91	*93*	*86*	*96*	86	*88*	*83*	**76**
400		*63*		*76*		*60*		*55*		*91*		*86*		*86*		*83*	
	260		*78*		**80**		*65*		*77*		**94**		*96*		**89**		*76*
800	**400**	66	*78*	*76*	*80*	*60*	*65*	*55*	*77*	*91*	*94*	*86*	*96*	*86*	*89*	*83*	**76**
	800		*78*		*80*		*65*		*77*		*94*		*96*		*89*		*76*
	Maximum	78		80		65		77		94		96		89		83	
	Max(single 96, pair 48) / Max(all)	85%		55%		86%		74%		88%		96%		90%		100%	

Table S5: Maximum achievable genus classification rate, 95% estimated accuracy, comparing paired-ending with single-ended sequencing. The classification rate shown in each cell is the maximum value observed for any choice of primers at each respective read length.

7 Precision vs Accuracy; "confident" predictions

Prior authors have often reported the extent to which a classification can be made at all (i.e., precision), without regard for whether that classification is actually correct (accuracy). An obviously problematic case is one in which all of the database hits to a sequence agree on a genus (corresponding roughly to a 95% OTU), but these hits are more than 5% divergent over their full length from the query sequence—indicating that the query sequence is in fact not a member of that OTU. It is not straightforward to limit the taxonomic level of the predictions on the basis of the observed %id of a sequence fragment, however, because the identity threshold associated with each level is variable throughout the sequence, and different fragments would give inconsistent results.

		Chimera-free, near-full-length sequences annotated to:							
Environment	GreenGenes StudyID	Total	kingdom	phylum	class	order	family	genus	species
Human Gut	42600	6760	6760	6760	6760	6760	6612	5844	1620
Mattress Dust and Human Skin	42612	3035	3035	3035	3033	3030	2976	2851	1346
Steer Rumen	42623	1831	1831	1831	1830	1829	1164	558	81
Termite Gut	33327	505	505	505	505	451	408	379	0
Ocean	35248	3985	3985	3985	3985	3935	2055	2388	714
Coral	35251	995	995	995	900	890	619	192	13
Hypersaline Mat	31588	510	510	510	477	401	182	102	20
Grassland Soil	42615	654	654	654	610	482	317	193	7

Table S6: Number of near-full-length sequences annotated to each rank, per environment.

We therefore used annotations on the query sequences, where available, to evaluate the accuracy of the taxonomic predictions at each level. Within each query dataset, for each primer and read length, and for each taxonomic level, we computed the proportion of predictions that proved correct, out of those sequences that were annotated at that level at all. We term this "estimated accuracy" because it is computed using only a subset of the query set (Table S6). We then applied two thresholds, 80% and 95%, to determine which primer/read length combinations produce trustworthy classifications under the given circumstances. We then removed all predictions that were deemed unreliable; for instance, a genus-level prediction for some sequence might be truncated to the order level, because more detailed predictions were found to be wrong more than 20% of the time for the given primer, read length, and environment.

8 Ranking of primer pair and read length combinations (Figure 1)

Because each panel of Figure 1 contains 9678 vertical bars (which are not resolvable in the image), it was necessary to sort the bars on the X axis in such a way that similar colors were adjacent; otherwise the plots became visually uninterpretable. The classification rates do not correlate well enough between different ranks to solve this problem by a simple sort at one rank: for example, when we sorted by genus classification rate, the proportions associated with other ranks remained a jumble. We found by trial and error that sorting by a weighted mixture of the proportions at different ranks produced comprehensible plots, though some degree of scrambled appearance inevitably remains. The function was:

$$sortorder = 8*strain + 8*species + 8*genus + 4*family + 2*order + 1*class + 1*phylum + 1*domain + 0.1*noclassification$$

where the proportion at each rank excludes that at lower ranks (e.g., here *family* means "family but not genus").

This sort was performed for each panel independently, so the resulting order is not consistent between panels.

9 Choice of representative optimal primers (Table 1)

We exhaustively computed the classification rate for thousands of combinations of primer, read length, experiment type, environment, taxonomic level, and estimated accuracy level. We found that some choices of primer and read length provided more classifications (at a given estimated accuracy level) than certain other choices across all environments tested. Our results suggest, for example, that one should not use primer E1046F with 128nt reads for taxonomic classification, because primer E517F with 80nt reads is always at least as informative at the genus level (and usually much

more so). In fact, for phylum level classifications, reads of only 32nt from E533R are substantially more informative. We filtered the results tables to exclude choices of primer and read length that were uniformly less informative than others of the same or shorter read length. We further filtered them, for the sake of tractable presentation, to include only primers that achieve at least 90% of the optimum classification rate (per read length) in at least one environment.

In a few cases, several choices provided nearly equivalent classification performance, particularly involving closely related primers such as E517F and U515F. We considered two choices to be equivalent (for a given taxonomic level and estimated accuracy level) if they provided classification rates within two percentage points in all environments. In these cases we list each alternative but report the classification performance of the representative with the best average performance.

The entries that remained after this filter was applied highlight the trade-offs inherent in the choice of primer and read length. Each remaining entry is optimal according to some criterion. For instance, for genus predictions from 128 nt reads, U519F classifies more of the ocean sample than does E341F, but E341F is able to classify sequences from the termite gut sample, where U519F makes few confident predictions; and neither of them are near optimal on the dust and skin sample, where E1406R produces substantially more confident predictions.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, *215*(3), 403–10.

Baker, G. C., Smith, J. J., & Cowan, D. A. (2003). Review and re-analysis of domain-specific 16s primers. *J Microbiol Methods*, *55*(3), 541–55.

Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G., & Neufeld, J. D. (2011). Generation of multimillion-sequence 16s rrna gene libraries from complex microbial communities by assembling paired-end illumina reads. *Appl Environ Microbiol*, *77*(11), 3846–52.

Brulc, J. M., Antonopoulos, D. A., Berg Miller, M. E., Wilson, M. K., Yannarell, A. C., Dinsdale, E. A., Edwards, R. E., Frank, E. D., Emerson, J. B., & Wacklin, P. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of the National Academy of Sciences*, *106*(6), 1948.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., & Knight, R. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nat Methods*, *7*(5), 335–6.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., & Knight, R. (2011). Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A*, *108 Suppl 1*, 4516–22.

Cho, J. C., Lee, D. H., Cho, Y. C., Cho, J. C., & Kim, S. J. (1996). Direct extraction of dna from soil for amplification of 16s rrna gene sequences by polymerase chain reaction. *J. Microbiology*, *34*(3), 229–235.

Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., & O'Toole, P. W. (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16s rrna gene regions. *Nucleic Acids Res*, *38*(22), e200.

Cruz-Martínez, K., Suttle, K. B., Brodie, E. L., Power, M. E., Andersen, G. L., & Banfield, J. F. (2009). Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. *ISME J*.

Degnan, P. H., & Ochman, H. (2011). Illumina-based analysis of microbial community diversity. *ISME J*.

Desantis, T. Z., Hugenholtz, P., Keller, K., Brodie, E. L., Larsen, N., Piceno, Y. M., Phan, R., & Andersen, G. L. (2006). Nast: a multiple sequence alignment server for comparative analysis of 16s rrna genes. *Nucleic Acids Res*, *34*(Web Server issue), W394–9.

Dethlefsen, L., Huse, S., Sogin, M. L., & Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rrna sequencing. *PLoS Biol*, *6*(11), e280.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, *26*(19), 2460–1.

Gloor, G. B., Hummelen, R., Macklaim, J. M., Dickson, R. J., Fernandes, A. D., MacPhee, R., & Reid, G. (2010). Microbiome profiling by illumina sequencing of combinatorial sequence-tagged pcr products. *PLoS One*, *5*(10), e15406.

Hamp, T. J., Jones, W. J., & Fodor, A. A. (2009). Effects of experimental choices and analysis noise on surveys of the "rare biosphere". *Appl Environ Microbiol*, *75*(10), 3263–70.

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., & Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing. *PLoS Genet*, *4*(11), e1000255.

Isenbarger, T. A., Finney, M., Rios-Velazquez, C., Handelsman, J., & Ruvkun, G. (2008). Miniprimer pcr, a new lens for viewing the microbial world. *Applied and environmental microbiology*, *74*(3), 840.

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Osterås, M., Schrenzel, J., & François, P. (2009). Metagenomic study of the oral microbiota by illumina high-throughput sequencing. *J Microbiol Methods*, *79*(3), 266–71.

Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H., Zhang, Y., Shen, J., Pang, X., Zhang, M., Wei, H., Chen, Y., Lu, H., Zuo, J., Su, M., Qiu, Y., Jia, W., Xiao, C., Smith, L. M., Yang, S., Holmes, E., Tang, H., Zhao, G., Nicholson, J. K., Li, L., & Zhao, L. (2008). Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci U S A*, *105*(6), 2117–22.

McDonald, D., Price, M., Goodrich, J., Nawrocki, E., DeSantis, T., Probst, G., Andersen, G., Knight, R., & Hugenholtz, P. (2011). An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, *276*(5313), 734–40.

Ram, J. L., Karim, A. S., Sendler, E. D., & Kato, I. (2011). Strategy for microbiome analysis using 16s rrna gene sequence analysis on the illumina sequencing platform. *Syst Biol Reprod Med*, *57*(3), 162–70.

Shaw, A. K., Halpern, A. L., Beeson, K., Tran, B., Venter, J. C., & Martiny, J. B. H. (2008). It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol*, *10*(9), 2200–2210.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., & Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A*, *103*(32), 12115–20.

Sunagawa, S., Desantis, T. Z., Piceno, Y. M., Brodie, E. L., Desalvo, M. K., Voolstra, C. R., Weil, E., Andersen, G. L., & Medina, M. (2009). Bacterial diversity and white plague disease-associated community changes in the caribbean coral montastraea faveolata. *ISME J*.

Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16s rrna gene. *Curr Opin Microbiol*, *11*(5), 442–6.

Täubel, M., Rintala, H., Pitkäranta, M., Paulin, L., Laitinen, S., Pekkanen, J., Hyvärinen, A., & Nevalainen, A. (2009). The occupant as a source of house dust bacteria. *J Allergy Clin Immunol*, *124*(4), 834–40.e47.

Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., Cayouette, M., McHardy, A. C., Djordjevic, G., Aboushadi, N., Sorek, R., Tringe, S. G., Podar, M., Martin, H. G., Kunin, V., Dalevi, D., Madejska, J., Kirton, E., Platt, D., Szeto, E., Salamov, A., Barry, K., Mikhailova, N., Kyrpides, N. C., Matson, E. G., Ottesen, E. A., Zhang, X., Hernández, M., Murillo, C., Acosta, L. G., Rigoutsos, I., Tamayo, G., Green, B. D., Chang, C., Rubin, E. M., Mathur, E. J., Robertson, D. E., Hugenholtz, P., & Leadbetter, J. R. (2007). Metagenomic

and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, *450*(7169), 560–5.

Wuyts, J., Van de Peer, Y., Winkelmans, T., & De Wachter, R. (2002). The european database on small subunit ribosomal rna. *Nucleic Acids Res*, *30*(1), 183–5.

Youssef, N., Sheik, C. S., Krumholz, L. R., Najar, F. Z., Roe, B. A., & Elshahed, M. S. (2009). Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16s rrna gene-based environmental surveys. *Appl Environ Microbiol*, *75*(16), 5227–36.

Zhou, H.-W. W., Li, D.-F. F., Tam, N. F.-Y., Jiang, X.-T. T., Zhang, H., Sheng, H.-F. F., Qin, J., Liu, X., & Zou, F. (2011). Bipes, a cost-effective high-throughput method for assessing microbial diversity. *ISME J*, *5*(4), 741–9.