environmental microbiology

Environmental Microbiology (2011) **13**(11), 3000–3009 doi:10.1111/j.1462-2920.2011.02577.x

On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys

Patricio Jeraldo,[1,2] Nicholas Chia[1,2] and Nigel Goldenfeld[1,2*]

[1]*Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 1206 West Gregory Drive, Urbana, IL 61801, USA.*

[2]*Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, IL 61801, USA.*

Summary

Pyrosequencing platforms have been widely used in 16S rRNA deep sequencing of organisms sampled from environmental surveys. Despite the massive number of reads generated by these platforms, the reads only cover short regions of the gene, and the use of these short reads has recently been called into question for phylogeny-based and diversity analyses. We explore the limits of the use of short reads by quantifying the loss of information, and its effect on phylogeny. Using available nearly-full-length reads from published clone libraries and databases, and simulated short reads created from these reads, we show that for selected regions of the gene, short reads contain a surprisingly high amount of biological information, making them suitable to resolve an approximate phylogeny. In particular, we find that the V6 region is significantly poorer than the V1–V3 region in its representation of phylogenetic relationships. We conclude that the use of short reads, combined with a careful choice of the gene region used, and a thorough alignment procedure, can yield phylogenetic information comparable with that obtained from nearly-full-length 16S rRNA reads.

Introduction

Advances in sequencing technology allow researchers to generate massive libraries of biological information. In particular, high-throughput sequencing methods (Margulies *et al*., 2005) are becoming widely used to analyse microbial communities (Sogin *et al*., 2006; Turnbaugh *et al*., 2006; 2009; Frias-Lopez *et al*., 2008; McKenna *et al*., 2008; Brazelton *et al*., 2010). One appealing aspect of recent advances in technology has been the deep environmental surveys of the microbial composition from a wide variety of environments, ranging from ocean (DeLong *et al*., 2006; Frias-Lopez *et al*., 2008), to soil (Elshahed *et al*., 2008), to mammal guts (Turnbaugh *et al*., 2009). In addition, 16S hypervariable tag sequencing has exposed the existence of a so-called 'rare biosphere' (Sogin *et al*., 2006), whose contribution to, and impact in, the microbial environment are only now beginning to be observable, quantified and appreciated (Brazelton *et al*., 2010). The potential significance of a previously unsuspected biosphere, rich in diversity but low in abundance, is that it may offer a major clue as to the response of ecosystems to change, and may well control their ability to adapt, by providing a large reservoir of genetic novelty to be tapped.

Despite this promise, the technology is still in its infancy. In particular, the reads generated by hypervariable tag pyrosequencing are short, spanning only hundreds of nucleotides. In the case of 16S rRNA, this has forced researchers to focus their studies on partial regions, usually including one or more of its hypervariable regions in an effort to capture the maximum possible amount of useful biological information. Naturally, there have been studies that compare the information obtained from these short reads with that obtained from nearly-full-length reads of SSU rRNA, quantifying the loss of information, dependence on the region of 16S rRNA being studied, and other possible biases (Liu *et al*., 2007; 2008; Huse *et al*., 2008; Youssef *et al*., 2009). In particular, the effects of the use of short reads in taxonomic assignments and ecological diversity indices have been documented. These studies make recommendations on which regions of the SSU rRNA are better suited to minimize the artefacts based on the observations of their studies, yet their recommendations are not fully consistent with each other, underlining some of the many complexities of the problem.

As pyrosequencing technology is maturing, the systematic artefacts that were present in earlier data sets have become less of an issue (Sogin *et al*., 2006). These artefacts were by no means minor in terms of their biological impact (Gomez-Alvarez *et al*., 2009; Quince *et al*., 2009; Kunin *et al*., 2010). For example, point errors present in reads and artificial replicates lead to spurious operational

Received 14 June, 2010; accepted 28 July, 2011. *For correspondence. E-mail nigel@uiuc.edu; Tel. (+1) 217 333 8027; Fax (+1) 217 333 9819.

taxonomic unit counts and overestimation of abundances in the operational taxonomic units (Gomez-Alvarez *et al*., 2009; Quince *et al*., 2009; Kunin *et al*., 2010). Fortunately, both of these artefacts can be easily removed with careful preprocessing (Gomez-Alvarez *et al*., 2009; Quince *et al*., 2009) and accurate alignment (DeSantis *et al*., 2006a; Cole *et al*., 2009) of the libraries. As these artefacts are being removed, we can grow more confident in pyrosequencing data and focus on the challenges imposed by the intrinsic information loss present in these data sets.

The purpose of this paper is to quantify the amount of phylogenetic information contained in short reads. In order to do this, we estimate the correlation between the phylogenetic information obtained using synthetic short reads, to that from nearly-full-length reads. To this end, we constructed an artificial clone library using 2000 nearly-full-length bacterial SSU rRNA sequences, randomly selected from the Greengenes (DeSantis *et al*., 2006b) 16S database, retrieved August 2009. The sequences in the library were trimmed in length to simulate data obtained using pyrosequencing, creating additional libraries. The libraries were then used to construct maximum likelihood (ML) phylogenetic trees. To quantify how much information is preserved in the trees made with short reads, a branch-length-based pairwise distance metric (Farris, 1967; Farahi *et al*., 2004), supplemented with Robinson–Foulds (RF) (Robinson and Foulds, 1981) and weighted RF (Robinson and Foulds, 1979) metrics, was used to correlate (Phipps, 1971; Farahi *et al*., 2004) and compare the structures between the different trees. We show that two different inferences of a phylogenetic tree using the same short read library can show marked discrepancies due to the stochastic nature of ML-based tree reconstruction methods the nature of the region being studied. We show that two different tree searches using the same short read library can show marked discrepancies due to the nature of the region being studied, given the randomized starting trees used by RAxML to perform such searches.

Then we show that, while a significant amount of information is preserved in the short read-based trees, the actual amount of information preserved seems to be not only a function of the length of the read, but also a function of the region sequenced. Our results indicate that the V1–V3 hypervariable region is a good estimator of phylogenetic information, and would be the preferred target for pyrosequencing assays of large communities, such as in large-scale environmental metagenomic surveys.

Results

As we aimed to compare phylogenetic information present in different regions of the 16S rRNA, we use a metric that can appropriately compare phylogenetic trees created from the different regions. Our metric measures the distance between a pair of sequences in the tree, and compares it with the corresponding distance in the other tree. The actual distance is computed as the length of the shortest path in the branches of the tree that goes from one member of the pair to the other one (Farris, 1967). This distance accounts for the different branch lengths calculated during the tree construction process. For a whole tree, this distance is computed for all possible pairs of sequences in the trees and it is stored as a distance matrix.

Now, to compare two trees, we create a tuple of the distances of corresponding pairs in both trees, and then we calculate the Pearson correlation (PC) R^2 for the set of distances (Phipps, 1971). We chose this method because it provides a balance between the computational expense of calculating more detailed comparison metrics, and the oversimplification of comparing single numbers coming from each tree, numbers that do not necessarily provide meaningful information, unless supplemented by extra details, such as the distribution of possible distances between randomly structured trees, which itself is something still very time-consuming to calculate.

We expect, for very similar trees, that the correlation R^2 will be very close to 1.0. This is a consequence of how close or far we expect different pairs of reads to be in different trees. In an ideal case, reads that are found to be close together in one tree are also expected to be found close in the other tree. Likewise, if two reads are found to be far apart in one tree, they are expected to be distant in the other tree. Thus, if we plot the distances found in one tree as a function of the distances found in a second tree, we would obtain a straight line in the case of identical trees, with $R^2 = 1.0$. For trees that are slightly different, the points will be scattered around this straight line. When the correlation is very poor, this straight line behaviour would be just a weak trend buried in a jungle of wildly scattered points. That said, we do not expect a situation where we observe very good, non-linear correlation, which can give very low values of R^2. We will call these graphs Sequence Correlation Plots.

As a first calculation, we applied this metric to different trees created from the same region. For the nearly-full-length case, this measurement would yield the minimum possible uncertainty in the structure of the resulting trees, resulting from the ML procedure we used. Thus, we construct an upper limit on the quality of our comparison methodology. From the resulting value of R^2 and the graph related to the comparison, we can base subsequent explorations.

The short reads are between 120 and 400 base pairs long, or approximately 15% and 30% of the full length of the 16S gene, and it is reasonable to expect a loss of phylogenetic information when using these reads. Now, when this distance metric is applied to trees created from the simulated short reads, we expect more deviations from the straight line behaviour of almost identical trees. This is because the ML trees created from shorter reads are harder to resolve – less sequence information leaves greater ambiguities to be resolved. The ML problem in these cases requires a more intense search of the solution space in order to converge to a solution, which itself is one of many equally good approximations to the 'real' solution. This suggests an interesting possibility: if it is harder to find a solution to the ML problem, which means the sequences would contain less phylogenetic information. This implies that the quality of a ML tree for fixed computational effort could be used as a proxy for measuring phylogenetic information content in the sequences used to create the trees.

As a way to supplement the patristic correlation metric, and also to gain further insight on the difference between the trees, we also compared the trees using a RF metric (Robinson and Foulds, 1981), and a weighted Robinson-Foulds (WRF) metric (Robinson and Foulds, 1979). The RF metric is the count of the splits present in one tree that are not present in the other. In other words, it is the symmetric difference of the sets of splits of the trees being compared. The WRF metric, on the other hand, multiplies each split count by a certain number. In our case, each split is weighted by the support value of said split, where the support value goes from 0 for no support, to 1.0 for full support for the split. Besides providing another measure of similarity between trees, using both RF and WRF metrics provide us some insight on the nature of the differences between the trees. If the RF distance is much larger than the WRF distance, we can infer that the differences between trees occur mostly on low-support subtrees, whereas if the WRF distance is closer to the RF distance, then the differences are mostly due to rearrangements of high-support subtrees (Pattengale *et al.*, 2010).

From this kind of comparison, we can gain some insight on two specific questions we have about the short reads: (i) how much information is contained in the hypervariable tag regions; and (ii) how the length of the read correlates with the phylogenetic information contained in the nearly complete gene. There has been previous work regarding the latter question, exploring simulated data sets to address the general question of length requirements for tree reconstruction (Bininda-Emonds *et al.*, 2001; Moret *et al.*, 2002), and although their concerns extend well beyond the question of short reads and into large-scale phylogeny in general, their findings are certainly relevant to our case and add to our observation of the complexity of the short reads situation.

In Fig. 1, we show Sequence Correlation Plots for the nearly-full-length (FL) tree, as well as for all the considered regions. Distances in the plot are normalized to a maximum value of 1.0 and, for clarity, they are also binned. The bins have a width of 0.05 distance units before normalization. The error bars correspond to the standard deviation calculated during the binning process.

In Fig. 1A, we show the comparison between two tree searches performed on the alignment of nearly-full-length reads. The correlation of these two trees is very high, and this comparison can be thought of as an approximate minimum of the possible uncertainty in determining the phylogeny based on 16S rRNA, using RAxML as the treeing tool, with the NAST-based alignments. There are differences even in the nearly-full-length trees because of the approximate, probabilistic nature of the solutions found for the ML problem as solved by RAxML. Even in this 'benchmark' case, there are a couple of notable features in this plot. First, the error bars are very small in the lower 20% of the pairwise distances in the plot, meaning that the reads that were found to be close together in one tree search stayed in similar positions in the other search. As the pairwise distance is increased, there is an increase in the size of the error bars, yet the error bars stay relatively constant up to the maximum distance in the trees. This indicates that the relationship between distantly related pairs does not significantly deviate from the average, indicating that the relationships between pairs are preserved on average.

In Fig. 1B, we see a very similar situation for the reads encompassing hypervariable regions V1 to V3 (V1–V3), a library that can be experimentally obtained using 454 GS FLX Titanium platforms. The rather surprising observation is that the lower 15% of the pairwise distances are very well preserved between tree searches on the same alignment, the error bars being comparable with those present in Fig. 1A. Beyond this 15%, the error bars grow substantially, but their magnitude remains more or less constant throughout the set of pairwise distances, signalling some loss of information compared with the FL trees, yet, at the same time, still being able to resolve the long distance relationships between reads for a majority of the pairs. Interestingly, the PC in this case ($R^2 = 0.89$) is very similar to that obtained when comparing the V1–V3 tree with the nearly-full-length reads tree (Fig. 2). Overall, when compared with the nearly-full-length situation, the V1–V3 has produced very consistent trees, as indicated not only by visually comparing the graphs, but also by the high value of R^2.

The situation for the other two regions appears very different. There is poor consistency in the structures of the different tree searches, and the error bars tend to grow as the distance increases. In the V6 region graph, there is a major change in behaviour at very large distances, signalling substantial differences in the tree structures.

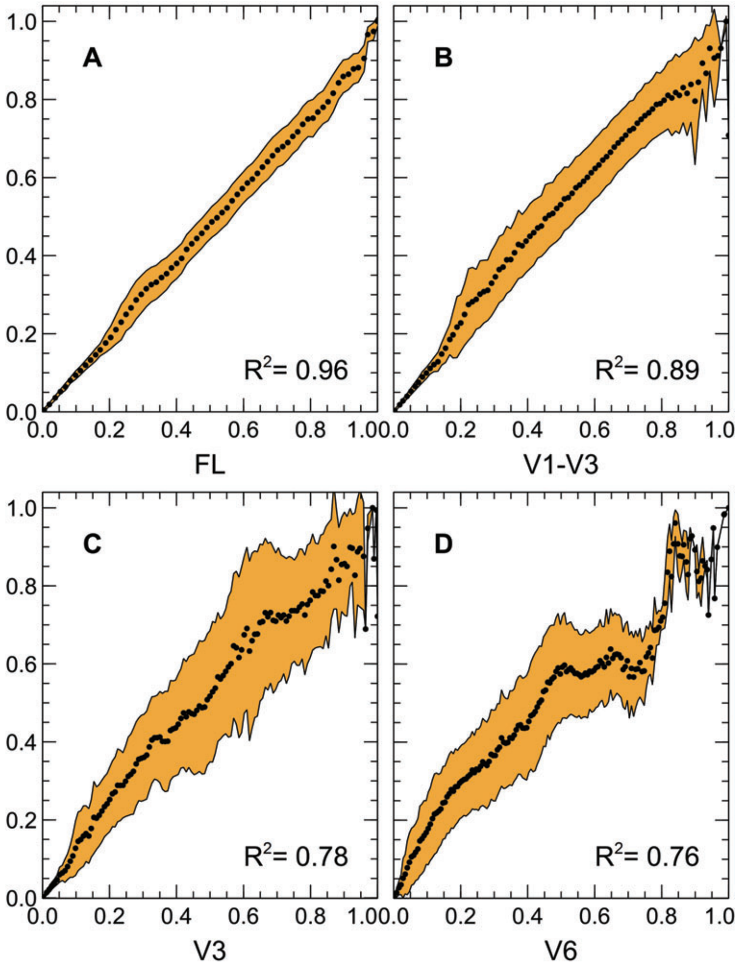

Fig. 1. Correlations between tree searches performed on full-length and partial reads. The sequence correlation plots show how similar are two ML tree searches for the different regions considered. The higher the correlations, the more similar the solutions to the ML problem. This measure can be used as a proxy for phylogenetic information content in the region used to construct the tree. The dots correspond to the average normalized distance in one tree as a function of the corresponding distance in the other tree, averaged over all distance pairs in a bin. The shaded area corresponds to the standard deviation for the points in each bin.

Figure 1C and D show the resulting plots for the shortest reads analysed. Beyond the very closely related reads, the error bars keep growing as the pairwise distances grow, in contrast to the FL and V1–V3 reads. This likely indicates an inability to consistently resolve the relationship between distant reads across tree searches. In particular, when focusing on the two trees we constructed using the V6 region (Fig. 1D), there are major differences between the distantly related taxa between both trees. This, despite the fact that the trees constructed using the short reads were calculated using more BS replicates during the construction process, and it is likely that adding more BS replicates to this process will not significantly change the outcome.

We now apply the same comparison metric to trees made from different regions to explore cross-correlation between different regions of the 16S gene. Unlike the previous case in which we used the correlation metric as a proxy to estimate the intrinsic phylogenetic information content of a given region, this cross-correlation between different regions can be used to estimate the phylogenetic information content from a region, relative to the information present in another region.

Figure 2 shows the normalized pairwise distances between the reads in one tree, as a function of the corresponding distance of a second tree. The top row contains all the comparisons against the nearly-full-length reads, where the V1–V3 tree shows good agreement with the FL tree. Other comparisons for this same region show similarly high correlations (Table 1). A different picture can be observed for V3 (Fig. 2A) and V6 (Fig. 2C), where correlations are not as good. The values of the correlations also vary with tree searches, such that the better match between a short read and the nearly-full-length tree is dependent on the found trees, signalling a rather important loss of phylogenetic information.

The bottom row shows the comparison between the different simulated short read libraries. As expected, comparisons with the V1–V3 region reads give a better correlation, comparable even with the correlations with nearly-full-length reads. But the comparison between the V3 and V6 is not as good, highlighting substantial

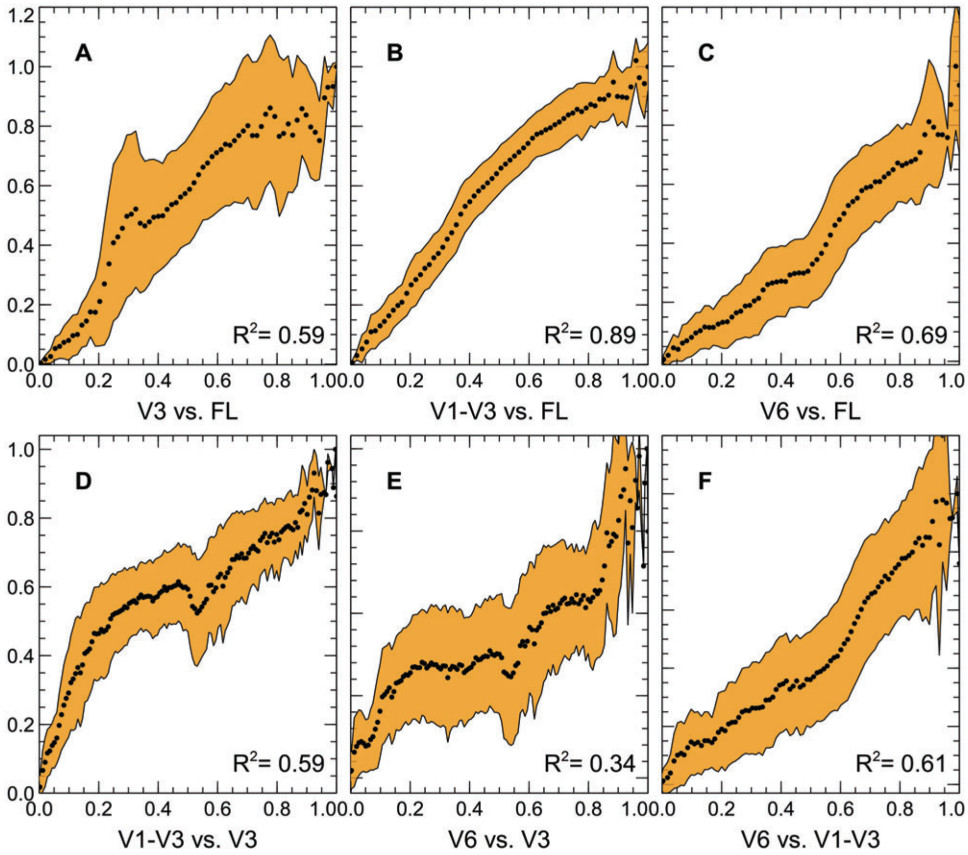

Fig. 2. Correlations between trees constructed using different regions. The top row shows pairwise distance correlations between trees from the different regions and a full-length tree. The bottom row shows the correlations between all the trees made from the different regions. These graphs illustrate the possible variations in tree correlation, ranging from very good (B) to poor (E). The dots correspond to the average normalized distance in one tree as a function of the corresponding distance in the other tree, averaged over all distance pairs in a bin.
The shaded area corresponds to the standard deviation for the points in each bin.

Table 1. Correlations and distances between all trees studied.

Tree pair	PC	RF	WRF	Tree pair	PC	RF	WRF
FL(1) vs. FL(2)	0.96	1012	150.98	V1–V3(1) vs. V3(1)	0.59	3302	746.68
FL(1) vs. V1–V3(1)	0.89	2700	861.52	V1–V3(1) vs. V3(2)	0.64	3336	750.52
FL(1) vs. V1–V3(2)	0.85	2724	871.21	V1–V3(1) vs. V6(1)	0.61	3634	942.84
FL(1) vs. V3(1)	0.59	3310	980.40	V1–V3(1) vs. V6(2)	0.53	3600	946.27
FL(1) vs. V3(2)	0.68	3308	969.56	V1–V3(2) vs. V3(1)	0.63	3306	746.42
FL(1) vs. V6(1)	0.69	3472	1083.62	V1–V3(2) vs. V3(2)	0.63	3322	744.12
FL(1) vs. V6(2)	0.60	3444	1083.58	V1–V3(2) vs. V6(1)	0.55	3638	943.23
FL(2) vs. V1–V3(1)	0.86	2702	871.29	V1–V3(2) vs. V6(2)	0.46	3596	946.51
FL(2) vs. V1–V3(2)	0.88	2742	886.41	V3(1) vs. V3(2)	0.78	2560	159.82
FL(2) vs. V3(1)	0.65	3300	979.69	V3(1) vs. V6(1)	0.34	3716	675.30
FL(2) vs. V3(2)	0.70	3292	969.87	V3(1) vs. V6(2)	0.33	3710	687.56
FL(2) vs. V6(1)	0.65	3476	1093.31	V3(2) vs. V6(1)	0.47	3704	666.35
FL(2) vs. V6(2)	0.58	3450	1091.76	V3(2) vs. V6(2)	0.42	3700	680.00
V1–V3(1) vs. V1–V3(2)	0.89	1578	171.42	V6(1) vs. V6(2)	0.76	2772	195.53

The table shows PC between all the trees used in this study, as well as their corresponding RF and WRF distances. The number in parentheses represents the tree search used when multiple trees were made from the same library. We see that the PC values for the longest short read libraries (V1–V3) are consistently high, and the corresponding values for the V3 and V6 regions show a high degree of fluctuation. Also of interest is the increase of RF/WRF when moving from V1–V3 towards V6, noting that the differences between trees start involving high-support subtrees, specially in the V6 case.

differences between the structures of the trees created using these regions.

The RF and WRF distances calculated for all tree pairs, besides complementing the data obtained with the correlations and giving us more insight on the differences between the trees, should also be consistent with the trends shown in the figures. Table 1 contains the PC, RF distances and WRF distances for all tree pairs analysed in this study. The number in parentheses indicates which of the tree searches for that particular region is being compared. The purpose of the table is to show all the range of R^2 values present in the data set, and also compare them with the corresponding RF and WRF distances. The first point we want to show is the range of differences for trees created from the same alignment. The PC values drop from the FL trees down to the V3/V6 case, as the RF distances increase, and the WRF distances only show a slight increase. We can interpret these differences as an increasing discrepancy between topologies and branch lengths as the size of the regions being used to construct the trees become shorter, but given the WRF distances most of the differences seem to be concentrated on the low-support subtrees. The next trend we notice is that the R^2 values between the V1–V3 trees and the FL trees are consistently high, ranging from 0.85 to 0.89. Their corresponding RF distances are very similar, ranging from 2700 to 2742, and their respective WRF distances also follow this pattern. These numbers show that, while there is quantifiable phylogenetic information loss, the tree structure is rather well resolved and consistent across tree searches. The next point we want to highlight is the fluctuation of the R^2 values for the trees found for V3 and V6, when correlated to the FL trees. From these numbers alone we cannot reliably tell if one of these regions is more suitable than the other. However, by looking at the RF and WRF values, we can see that the topological differences between the trees are due to rearrangements of high-support subtrees. The fact that the RF and WRF distances don't fluctuate as much compared with the PC values shows that, while the differences in subtree structure might be relatively comparable, the branch length differences have a measurable effect, signalling potentially damaging loss of phylogenetic information for these regions. The final point is that the magnitude of the values of R^2 for the V3 and V6 trees, when correlated with the V1–V3 trees, are only slightly lower than for the FL case, the corresponding RF distances are similar, and their WRF distances are slightly lower, meaning that the topological differences are slightly more biased towards lower-support subtrees, giving another point of support for V1–V3 as being a good proxy for the FL reads.

Finally, we can say something about the trees themselves. In Table 2 we see the tree lengths, defined as the sum of all branch lengths in the trees, and the logarithm of the likelihood (LogLk) value for the trees, as calculated using the original full-length alignment. The values of the tree lengths don't seem to follow a particular trend with respect to the read length. On the other hand, the LogLk values seem to get closer to zero as the read length decreases, meaning that the trees obtained are a progressively worse description of the data in the full-length alignments, when decreasing the read lengths, thus adding another layer of support to the observations about the discrepancies when using shorter reads.

Table 2. Parameters characterizing the trees.

Tree	Length	LogLk
FL(1)	189.38	–473754.97
FL(2)	188.15	–473750.26
V1–V3(1)	252.90	–494070.55
V1–V3(2)	254.76	–493119.36
V3(1)	172.54	–519286.99
V3(2)	187.78	–520317.05
V6(1)	132.72	–552733.64
V6(2)	136.26	–550485.72

The table shows the tree length, defined as the sum of all branch lengths of the tree, and the logarithm of the likelihood for the tree (LogLk) using the full-length alignment as the input data, as calculated during the tree search using a maximum likelihood method. The tree lengths don't seem to follow a trend with decreasing read length. The likelihood values, on the other hand, indicate that the trees made with short reads get worse at describing the relationships inferred from the full-length alignment as the reads become shorter.

Discussion

We have examined trees created from simulated short read libraries that were constructed using a random sample from the Greengenes database, comparing them using a Pearson Correlation of patristic distances between leaves of the trees, supplemented with RF and WRF distances. This comparison give us insight on the phylogenetic information content of the short reads, and it is a complement to other studies that quantify the pros and cons (Quince *et al*., 2009; Kunin *et al*., 2010) of pyrosequencing technology. This is specially relevant now, in the light of the huge influx of environmental data coming from big projects such as the ocean environmental studies (Shi *et al*., 2009), Human Microbiome Project (Costello *et al*., 2009; Turnbaugh *et al*., 2010) or studies on other more particular environments (Jones *et al*., 2010; Koopman *et al*., 2010) that would be difficult to accomplish if not for pyrosequencing.

We have also only considered focusing on a *de novo* approach to constructing phylogenetic trees, instead on doing a survey on all popular methods for reconstructing phylogenies, including insertion of reads into a pre-existing tree. Although the question of the reliability and comparison of large-scale phylogenies is certainly interesting, as evidenced by published studies on the subject [for example Liu *et al*., (2008)] we wanted to focus on a specific problem of phylogenetic information loss, not doing a comparison of methods used in community analysis, which is beyond the scope of this paper.

Our study identifies in detail the limitations of the short reads, from a phylogenetic information point of view, complementing other short read studies (Huse *et al*., 2008; Liu *et al*., 2008; Youssef *et al*., 2009), which conclude that short reads less than 200 bp long show significant topological differences between tree searches, signalling phylogenetic information loss. From this observation, we can say that any conclusions derived using phylogeny-based tools [most notably Unifrac (Lozupone *et al*., 2006)] that used very short reads and *de novo* phylogenies as their input data, have to be interpreted with some significant degree of uncertainty, independent of the region of 16S sequenced. Similar concerns have also been expressed elsewhere in the published literature (Schloss, 2010).

On the other hand, based on the results shown here, the prospect looks much better when using appropriately chosen longer reads, which are already accessible using FLX Titanium pyrosequencing technology. These longer reads make it possible to extract phylogenetic information with high degree of reliability. The type of analysis we performed can be extended to other genes of interest, such as proteorhodopsins (Frigaard *et al*., 2006), which show a high degree of environmental correlation.

Conclusion

In this paper, we generated synthetic short reads from complete 16S rRNA databases, and compared the complete phylogenetic trees with those obtained from the synthetic short reads. Our results show unequivocally that the different hypervariable regions are not equally suitable for this purpose, and that the V1–V3 region is the one that represents the best proxy for the complete 16S rRNA gene.

Experimental procedures

Selection of sequence sample and alignment

The single data set used in this study consisted of 2017 bacterial, nearly-full-length 16S rRNA sequences randomly selected from the Greengenes database (DeSantis *et al*., 2006b), as of August 2009. The reason for choosing 2017 sequences as the sample size of the database comes from the desire to perform a realistic comparison in the light of the sizes of existing read libraries obtained for nearly-full-length 16S sequences (for example, using Sanger sequencing). Although pyrosequencing is now able to create libraries with sizes in the order of millions of reads, a comparison study is certainly not realistic due to the lack of full-length libraries of that size in studies, and also bumping the separate problem of creating a phylogeny for such a big number of reads. Thus, we chose to limit ourselves to a simple case and small size, which is relevant to already published studies.

Table 3. PC, RF and WRF metrics for a non-masked (NM) and Lane-masked versions of a test alignment.

Tree pair	PC	RF	WRF
NM(1) vs. NM(2)	0.97	1042	179.50
NM(1) vs. LM(1)	0.93	1934	512.35
NM(1) vs. LM(2)	0.94	1900	495.39
NM(2) vs. LM(1)	0.94	1910	503.73
NM(2) vs. LM(2)	0.95	1864	484.81
LM(1) vs. LM(1)	0.96	1304	183.44

It is customary to perform Lane-masking (Lane *et al*., 1985) of 16S rRNA alignments when constructing neighbour-joining trees. In our work, we use the more accurate ML algorithms. No masking of the alignment is needed, because a ML approach would place little weight on extremely variable regions. To demonstrate this point, a Lane-masked version of an almost-full-length 16S rRNA alignment was tested using the standard RF and WRF metrics. The RF metric measures the number of splits present in one tree that are not present in the other tree, that is, the symmetric difference of the two sets of splits. The WRF metric has the extra feature of weighting the splits by their support value. This alignment was used in two tree searches, and the corresponding results compared using these metrics. As shown in Table 3, the trees from the Lane-masked alignment show a RF distance of 1304, whereas the trees from the non-masked alignment show a RF distance of 1042. In the case of the WRF distance, the Lane-masked trees show a distance of 179.50, and the non-masked version shows a distance of 183.44. Although there is greater variation in the Lane-masked trees found by ML, these differences have low-support values, as evidenced by the very similar WRF distances for both the Lane-masked and non-masked trees. This shows that Lane masking creates a measurable deterioration of the quality of the tree from a topological point of view, and therefore would create an uncontrolled artefact in our topologically focused analysis. For these reasons we do not use Lane masking.

Influence of sequence alignment on tree metrics

The sequences extracted for this study were obtained from the Greengenes database, and as such those sequences are profile-aligned using NAST to Greengenes own 16S rRNA alignment template. If, for example, we profile-align the same sequences to a different template, such as the SILVA bacterial template, or an altogether different method, such as the Infernal aligner present in the Ribosomal Database Project's 16S pipeline, it is reasonable to expect differences between the obtained phylogenies. In this study we are interested in the relative differences between phylogenies processed using the same pipeline, so a valid question is, what differences can we expect in the tree comparison metrics if we use different alignment strategies?

Table 4. Greengenes (GG) and SILVA (S) alignment templates compared using the PC, RF and WRF metrics.

Tree pair	PC	RF	WRF
GG(1) vs. GG(2)	0.98	936	157.11
GG(1) vs. S(1)	0.95	1766	648.35
GG(1) vs. S(2)	0.92	1782	646.06
GG(2) vs. S(1)	0.96	1820	663.00
GG(2) vs. S(2)	0.92	1780	644.58
S(1) vs. S(2)	0.96	890	193.67

Two tree searches were made from each alignment.

To this end, we set up a simple control test to measure the differences between two different alignment templates, the Greengenes template and the SILVA template, starting from the same original data. We expect trees constructed using different alignments of the same library to be somewhat different, of course, but it is essential that trees resulting from searches performed on the same alignment should be more similar to each other (for full-length 16S trees) than to trees from a different alignment.

To see if this is the case, we realigned the full-length library to the SILVA bacterial profile using Mothur's NAST and compared the resulting trees with each other and with the trees from the 'unaligned' library. The results are shown in Table 4. The trees were compared using Pearson Correlation, RF and WRF metrics. All three distance metrics between different alignments are greater than those for trees from the same alignment. This analysis shows that indeed there is consistency between trees made from the same alignment, and thus realignment is not necessary.

Although the RF scores of the SILVA and Greengenes alignments are comparable, the WRF scores are significantly different, presumably reflecting the presence of subtrees in the SILVA trees with higher support values than the Greengenes trees. This may also be reflected in the slight difference in the PC. In summary, realignment is not necessary for our analysis, and we proceed just using the original Greengenes alignment.

Trimming of sequences to create the libraries of simulated short reads

The trimming procedure necessary to create the artificial libraries was performed after the alignment of the source sequences was completed, instead of arbitrarily set the lengths before alignment, which would indiscriminately remove some information give the uneven starting and ending points for the reads. The reason for doing this is to maximize the information content of the reads, and also to use the standard starting and ending points for the studied regions, which are defined by the primers used at sequencing time. Also, for the almost-full-length library, the end points were trimmed in such a way to maximize the overlap between all reads.

The selected sequences were imported into the alignment manipulation program Jalview (Waterhouse *et al.*, 2009). The sequences were then trimmed to the lengths expected for pyrosequencing reads coming from the 454 Life Sciences Genome Sequencers GS 20, GS FLX Standard and GS FLX Titanium platforms, making sure they contain the hypervariable regions of interest.

Construction of phylogenetic trees

To construct the ML phylogenetic trees from the sequences in the libraries, we used RAxML (Stamatakis, 2006b) version 7.0.4, multithreaded using the Pthreads library (Ott *et al.*, 2007), using the rapid bootstrap (Stamatakis *et al.*, 2008) option and the CAT model of rate heterogeneity (Stamatakis, 2006a). The trees constructed using the nearly-full-length sequences were created from 300 BS replicates, and the ones created from the simulated short reads were created using 1000 BS replicates. We performed two tree searches for each library in the set using different seeds for RAxML's random number generator, which then we used to calculate pairwise distances.

The actual command line used for the tree searches reads as:

```
raxmlHPC-PTHREADS -T <threads> -fa -m GTRGA
MMA -N <replicates> -x <seed1> -p <seed2> -s
<alignment> -n <name>
```

where 'threads' is the number of threads per computer node, 'replicates' is the number of BS replicates in the tree search, 'seed1' and 'seed2' are integer numbers used to seed RAxML's random number generators, 'alignment' is the alignment file name, and 'name' is a name to identify the output file.

Distance and correlation calculations

To compare the different trees, we used a definition of pairwise distance, which depends on the structure of the tree. The pairwise distance between two sequences is defined as the sum of the branch lengths of the shortest path connecting the leaves representing the sequences in the tree (Farris, 1967). For each tree we calculated the pairwise distances between all possible pairs of sequences, with branch-lengths calculated under GAMMA, thus creating a patristic distance matrix for a particular tree.

Then we calculate the PC for all possible pairs of patristic distance matrices. For that, we create tuples from the corresponding elements in a pair of matrices, and then we calculated the PC R^2 for this set of tuples.

We also supplemented this metric with standard metrics for tree comparison, namely the RF metric (Robinson and Foulds, 1981) and a WRF metric (Robinson and Foulds, 1979). The RF distance between two trees is the number of splits present in one tree that are not present in the other tree, that is, the symmetric difference of the two sets of splits. The WRF distance differs from the unweighted RF distance in that it assigns a weight to each of the splits present in the symmetric difference, and the actual value is then the sum of these weights. To calculate these distances between all trees created we used RAxML version 7.2.6, which uses the support values of the splits as the weight for WRF distances.

Plotting of distance data

As a data reduction step, we took the tuples created from the two distance matrices being compared, and proceeded to do a binning step. This step consisted in performing an average over intervals of size 0.01 distance units, and also obtaining the standard deviation for each interval. After normalizing the maximum distance to 1.0, and rescaling the standard deviations accordingly, the data sets were now ready for plotting.

Acknowledgements

We thank Bryan White and Maksim Sipos for insightful discussions during the preparation of this manuscript. This work was supported by the U.S. Department of Energy under grant number DOE-2005–05818, and the National Science Foundation under grant number NSF-EF-0526747. P.J acknowledges support by the L.S. Edelheit Family Biological Physics Fellowship. N.C acknowledges support from the Institute of Genomic Biology Fellows Program.

References

Bininda-Emonds, O.R., Brady, S.G., Kim, J., and Sanderson, M.J. (2001) Scaling of accuracy in extremely large phylogenetic trees. *Pac Symp Biocomput* **6:** 547–558.

Brazelton, W.J., Ludwig, K.A., Sogin, M.L., Andreishcheva, E.N., Kelley, D.S., Shen, C.C., *et al*. (2010) Archaea and bacteria with surprising microdiversity show shifts in dominance over 1,000-year time scales in hydrothermal chimneys. *Proc Natl Acad Sci USA* **107:** 1612–1617.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., *et al*. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37:** D141–D145.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009) Bacterial community variation in human body habitats across space and time. *Science* **326:** 1694–1697.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U., *et al*. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496–503.

DeSantis, T.Z., Hugenholtz, P., Keller, K., Brodie, E.L., Larsen, N., Piceno, Y.M., *et al*. (2006a) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* **34:** W394–W399.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., *et al*. (2006b) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ Microbiol* **72:** 5069–5072.

Elshahed, M.S., Youssef, N.H., Spain, A.M., Sheik, C., Najar, F.Z., Sukharnikov, L.O., *et al*. (2008) Novelty and uniqueness patterns of rare members of the soil biosphere. *Appl Environ Microbiol* **74:** 5422–5428.

Farahi, K., Pusch, G.D., Overbeek, R., and Whitman, W.B. (2004) Detection of lateral gene transfer events in the prokaryotic tRNA synthetases by the ratios of evolutionary distances method. *J Mol Evol* **58:** 615–631.

Farris, J.S. (1967) The Meaning of Relationship and Taxonomic Procedure. *Syst Zool* **16:** 44–51.

Frias-Lopez, J., Shi, Y., Tyson, G.W., Coleman, M.L., Schuster, S.C., Chisholm, S.W., and Delong, E.F. (2008) Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA* **105:** 3805–3810.

Frigaard, N.U., Martinez, A., Mincer, T.J., and DeLong, E.F. (2006) Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439:** 847–850.

Gomez-Alvarez, V., Teal, T.K., and Schmidt, T.M. (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **3:** 1314–1317.

Huse, S.M., Dethlefsen, L., Huber, J.A., Mark Welch, D., Relman, D.A., and Sogin, M.L. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* **4:** e1000255.

Jones, R.T., Knight, R., and Martin, A.P. (2010) Bacterial communities of disease vectors sampled across time, space, and species. *ISME J* **4:** 223–231.

Koopman, M.M., Fuselier, D.M., Hird, S., and Carstens, B.C. (2010) The carnivorous pale pitcher plant harbors diverse, distinct, and time-dependent bacterial communities. *Appl Environ Microbiol* **76:** 1851–1860.

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12:** 118–123.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* **82:** 6955–6959.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D., and Knight, R. (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* **35:** e120.

Liu, Z., DeSantis, T.Z., Andersen, G.L., and Knight, R. (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* **36:** e120.

Lozupone, C.A., Hamady, M., and Knight, R. (2006) UniFrac – an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7:** 371.

McKenna, P., Hoffmann, C., Minkah, N., Aye, P.P., Lackner, A., Liu, Z., *et al*. (2008) The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog* **4:** e20.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Moret, B., Roshan, U., and Warnow, T. (2002) Sequence-Length Requirements for Phylogenetic Methods. In *WABI 2002, Volume 2452 of Lecture Notes in Computer Science*. Guigó, R., and Gusfield, D. (eds). Berlin-Heidelberg, Germany: Springer-Verlag, pp. 343–356.

Ott, M., Zola, J., Stamatakis, A., and Aluru, S. (2007) Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. In *Proceedings of the 2007 ACM/IEEE Conference on Supercomputing – SC '07*. New York, USA: ACM Press.

Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R.P., Moret, B.M.E., and Stamatakis, A. (2010) How many bootstrap replicates are necessary? *J Comput Biol* **17:** 337–354.

Phipps, J.B. (1971) Dendrogram topology. *Syst Zool* **20:** 306–308.

Quince, C., Lanzen, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., *et al*. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6:** 639–641.

Robinson, D., and Foulds, L. (1979) Comparison of Weighted Labelled Trees. In *Combinatorial Mathematics IV, Volume 748 of Lecture Notes in Mathematics*. Horadam, A.F., and Wallis, W.D. (eds). Berlin-Heidelberg, Germany: Springer-Verlag, pp. 119–126.

Robinson, D., and Foulds, L. (1981) Comparison of phylogenetic trees. *Math Biosci* **53:** 131–147.

Schloss, P.D. (2010) The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. *PLoS Comp Biol* **6:** e1000844.

Shi, Y., Tyson, G.W., and DeLong, E.F. (2009) Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459:** 266–269.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., *et al*. (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103:** 12115–12120.

Stamatakis, A. (2006a) Phylogenetic models of rate heterogeneity: a high performance computing perspective. In Proceedings 20th IEEE International Parallel & Distributed Processing Symposium, pages 1–8. IEEE.

Stamatakis, A. (2006b) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688–2690.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008) A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol* **57:** 758–771.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444:** 1027–1031.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., *et al*. (2009) A core gut microbiome in obese and lean twins. *Nature* **457:** 480–484.

Turnbaugh, P.J., Quince, C., Faith, J.J., McHardy, A.C., Yatsunenko, T., Niazi, F., *et al*. (2010) Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci USA* **107:** 7503–7508.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009) Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25:** 1189–1191.

Youssef, N., Sheik, C.S., Krumholz, L.R., Najar, F.Z., Roe, B.A., and 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:** 5227–5236.