New Screening Software Shows that Most Recent Large 16S rRNA Gene Clone Libraries Contain Chimeras†

Kevin E. Ashelford,^{1*} Nadia A. Chuzhanova,² John C. Fry,¹ Antonia J. Jones,³ and Andrew J. Weightman¹

Cardiff School of Biosciences, Cardiff University, Main Building, Park Place, PO Box 915, Cardiff CF10 3TL, United Kingdom¹; Biostatistics and Bioinformatics Unit and Institute of Medical Genetics, Cardiff School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, United Kingdom²; and Cardiff School of Computer Science, Cardiff University, Queen's Buildings, 5 The Parade, Roath, Cardiff CF24 3AA, United Kingdom³

Received 8 March 2006/Accepted 9 June 2006

A new computer program, called Mallard, is presented for screening entire 16S rRNA gene libraries of up to 1,000 sequences for chimeras and other artifacts. Written in the Java computer language and capable of running on all major operating systems, the program provides a novel graphical approach for visualizing phylogenetic relationships among 16S rRNA gene sequences. To illustrate its use, we analyzed most of the large libraries of cloned bacterial 16S rRNA gene sequences submitted to the public repository during 2005. Defining a large library as one containing 100 or more sequences of 1,200 bases or greater, we screened 25 of the 28 libraries and found that all but three contained substantial anomalies. Overall, 543 anomalous sequences were found. The average anomaly content per clone library was 9.0%, 4% higher than that previously estimated for the public repository overall. In addition, 90.8% of anomalies had characteristic chimeric patterns, a rise of 25.4% over that found previously. One library alone was found to contain 54 chimeras, representing 45.8% of its content. These figures far exceed previous estimates of artifacts within public repositories and further highlight the urgent need for all researchers to adequately screen their libraries prior to submission. Mallard is freely available from our website at <http://www.cardiff.ac.uk/biosi/research/biosoft/>.

Recent papers (2, 14) have reported numerous corrupt 16S rRNA gene sequences within the public repositories (3, 16, 19), and it has been estimated that overall, 5% of records are likely to have substantial anomalies (2). While poor sequencing and errors during assembly have led to some of these reported errors, most anomalies have been chimeras—artificial sequences generated from two or more phylogenetically different DNA templates during PCR amplification (17, 22–24, 30, 31).

Our previous study showed that chimeras and other anomalies are continuing to be generated and submitted without comment to the public repositories (2). The presence of such high numbers of substantial anomalies in the public domain has serious implications for future efforts to accurately estimate bacterial diversity, elucidate likely phylogenetic relationships, and form correct taxonomic identifications. Consequently, there is a requirement for effective computer programs to simplify the screening process.

A number of useful, complementary approaches already exist, with Bellerophon (13) and CHIMERA_CHECK (19) being two noteworthy examples, and in our previous paper we described a new computer program, called Pintail, for screening individual sequences for errors (2). Now, we describe another program, Mallard, which develops the Pintail algorithm further so that whole libraries of 16S rRNA gene sequences can be screened simultaneously and quickly.

We demonstrate the new program's ability to screen libraries

of a range of sizes from different sources. Through a detailed analysis of submissions made to public repositories during 2005, we show that the problem of unrecognized anomalies within the public domain appears to be getting worse, highlighting the need for immediate steps to be taken, by the research community at large, to minimize further database contamination.

MATERIALS AND METHODS

Program development. Our new program, named Mallard, expands on the Pintail algorithm described previously (2). In brief, the Pintail algorithm works by undertaking a pairwise comparison between a query sequence, S_q , and a subject sequence, S_s , by aligning the sequence pair and then assessing changes in uncorrected evolutionary distance, o_n , between the two sequences within a sliding window of size w , moving l bases at a time along the alignment, resulting in m measurements. The resulting data set of observed percentage differences, $O_{qs} = \{o_i: o_1, o_2, \dots, o_m\}$, is compared with what might be expected for two reliable sequences of equivalent evolutionary distance, $E_{qs} = \{e_i: e_1, e_2, \dots, e_m\}$, and the resulting summarizing statistic, the deviation from expectation (DE) value,

$$DE = \sqrt{\frac{\sum_1^m (o_i - e_i)^2}{m - 1}}$$

quantifies the likelihood that an anomaly is present. A more thorough description of the Pintail algorithm, including an explanation of how E_{qs} is calculated, can be found in the help documentation accompanying the software and in our previous article (2).

In Mallard, the Pintail algorithm is applied to all pairwise comparisons within a multiple alignment of size n , resulting in $(n^2 - n)/2$ separate DE values, each DE value presenting a unique pairwise comparison. DE values are plotted against their corresponding mean observed percentage differences, $(\sum o_i)/m$, which can be viewed as a simple measure of the evolutionary distance between sequences S_q and S_s . The larger the DE value, the greater the likelihood that either S_q or S_s (or perhaps even both) is in some way corrupt. Thus, by plotting DE values, one can immediately see which pairwise comparisons are likely to involve an anomalous sequence, since DE values generated from reliable se-

* Corresponding author. Mailing address: Cardiff School of Biosciences, Cardiff University, Main Building, Park Place, P.O. Box 915, Cardiff CF10 3TL, United Kingdom. Phone: 44 (0)29 20 876002. Fax: 44 (0)29 20 874305. E-mail: ashelford@cardiff.ac.uk.

† Supplemental material for this article may be found at <http://aem.asm.org/>.

quences will tend to cluster close to a DE value of zero, while DE values involving anomalous sequences will tend to appear as outliers.

In Mallard, outliers are identified as those DE values that appear above one of several possible cutoff lines, specified by the user and based on DE values calculated from comparisons of error-free sequences from type strains (2). Specifically, in our earlier study, we calculated DE values from a collection of 2,007 reliable type strain sequences, and the 75, 95, 99, 99.9, and 100% quantiles of the resulting plot were determined at each 1% interval along the x axis of the DE plot (2). These quantile data give roughly straight lines when plotted on a logarithmic scale, so for this study, the quantile data were simplified to the following equations: 75% quantiles, $y = 2.28 \log_{10} x + 1.00$; 95% quantiles, $y = 2.64 \log_{10} x + 1.46$; 99% quantiles, $y = 3.12 \log_{10} x + 1.66$; 99.9% quantiles, $y = 3.27 \log_{10} x + 2.07$; 100% quantiles, $y = 4.37 \log_{10} x + 1.81$. Cutoff lines, generated from these equations, are offered by the program.

DE outliers are caused by one, or even both, of the sequences involved in the corresponding pairwise comparison being anomalous. To identify which are the corrupt sequences, the following procedure is applied by the program. First, each sequence in the library is scored according to the number of DE outliers it is coreponsible for. The DE outliers are then ranked, in descending order, according to distance from the cutoff line. For each DE outlier, the two sequences responsible for that outlier are identified, and if neither sequence has previously been marked as anomalous, the sequence with the highest score is marked as such (or both are marked if they have the same score). In this way, a list of anomalous sequences is generated, with those that were identified first being the most likely anomalies.

Mallard was written in Java 1.4 (Java Technology) and tested on Redhat 9.0 Linux, Microsoft Windows XP, and Apple Mac OS X, version 10.2. The program, along with full instructions for use, help documentation, example files, and source code, is freely available from <http://www.cardiff.ac.uk/biosi/research/biosoft/>. Mallard is an open-source project and is released under the terms of the GNU General Public License (<http://www.gnu.org/copyleft/gpl.html>).

Analysis of 16S rRNA gene libraries. To demonstrate Mallard's utility, a selection of publicly available 16S rRNA gene libraries was analyzed. The procedure was the same for each library. A multiple-sequence alignment was prepared for each that included the sequence *Escherichia coli* U00096 (as the reference sequence). An explanation for the reference sequence is included in the accompanying help documentation. Each multiple sequence alignment was passed to the Mallard program and screened for putative anomalies. Each putative anomaly identified by the program was checked with BlastN (1), in conjunction with the Pintail program (2). First, a library of *Verrucomicrobia*-derived sequences, to exemplify a *Bacteria* phylum, was considered. A total of 222 near-complete ($\geq 1,200$ -base) representatives of the *Verrucomicrobia*, as identified by the Ribosome Database Project (RDP) (4) release 9 update 36, were downloaded, along with *E. coli* U00096 as a reference, as an aligned file from the website <http://rdp.cme.msu.edu/>.

Second, a library of *Crenarchaeota*-derived sequences, to represent the *Archaea*, was analyzed. Near-complete sequences were identified from the National Center for Biotechnology Information (NCBI) online database (<http://www.ncbi.nlm.nih.gov/>) using the search phrase "16S[TITL] AND Archaea[ORGN] AND Crenarchaeota[ORGN] AND 1200[LEN]:1600[LEN]." The resulting data set of 270 sequences was checked and then aligned, along with *E. coli* U00096, using ClustalW (27).

Third, a coastal-marine 156-sequence clone library (AY354711 to AY354866), previously generated by our laboratory (20), was examined. Because this library consisted of partial 16S rRNA gene sequences, it was necessary to subdivide it according to the region of the 16S rRNA gene covered so that sensible alignments were obtained. All groups were aligned, along with *E. coli* U00096, using ClustalW.

Finally, a selection of clone libraries representing submissions to the public repositories over the last year was analyzed. Using the "View by Publication" facility on the RDP's online hierarchy browser, all libraries submitted during 2005 were identified. Of these, libraries containing ≥ 100 near-complete ($\geq 1,200$ -base) sequences were identified. Three libraries (with 2,062, 3,635, and 11,831 near-complete sequences) exceeded our 1,000-sequence limit and were discarded. In this way, 25 libraries were selected for analysis, the near-complete sequences of which were downloaded as RDP aligned datasets (each including *E. coli* U00096).

Comparison with Bellerophon. The most widely used program for checking whole gene libraries for chimeras is currently the server-based program Bellerophon (13). Bellerophon was used to analyze the *Verrucomicrobia*-, *Crenarchaeota*-, and coastal-water-derived (20) libraries described above, using the same input files prepared for Mallard.

In addition, we considered the performances of both programs in relation to

two further gene libraries. First, we considered the 18-sequence gene library of Stein et al. (26), which the Bellerophon website (<http://foo.maths.uq.edu.au/~huber/bellerophon.pl>) uses as an example file. Secondly, we considered the recently published gene library of Walker et al. (29), selected as an example of a library containing a mixture of near-complete and partial sequences. In the latter library, records AY911480, AY911482, AY911483, AY911485, AY911493, and AY911495, although labeled as *Alphaproteobacteria* in origin, were in fact found to closely resemble *Acanthamoeba* mitochondrial 16S ribosomal DNA and so were excluded from analysis. In addition, AY911496, an example of chloroplast 16S rRNA, was excluded.

For all comparisons, the default settings for both programs were used. The same aligned input files were used for both programs. Since Bellerophon is designed specifically to detect chimeras, we restricted our analyses to the detection of chimeric records. In all cases, chimeras were confirmed and false positives were identified by using the Pintail program (2).

RESULTS

Operation of the program with analysis of the *Verrucomicrobia* library as an example. Figure 1 shows a sample screenshot of the Mallard program, displaying an analysis of the *Verrucomicrobia* 16S rRNA gene library. The right-hand panel of the screenshot shows a plot (reproduced in Fig. 2A) of the 24,531 DE values calculated from the 222 *Verrucomicrobia* sequences. DE values above the cutoff line (in this case, 100%), superimposed on the plot, were judged by the program to be suspiciously high and were marked as outliers. DE outliers typically result from sequence comparisons in which at least one of the pair contains errors, so the DE values above the cutoff line in Fig. 1 (and Fig. 2A) are likely to be the result of anomalous sequences within the *Verrucomicrobia* library.

Each plotted DE value summarizes a separate Pintail plot (e.g., Fig. 2B and D) that is the result of applying the Pintail algorithm to a sequence pair. Within the program, Pintail plots for any DE value can be viewed. For example, in Fig. 2A, a suspiciously high DE value of 10.04 has been selected (by mouse clicking the data point). This particular DE value was generated by a comparison between sequences AY752110 and AF050561, and the accompanying Pintail plot is shown (Fig. 2B). Note how the observed percentage difference line (Fig. 2B), which reflects differences in evolutionary distance between the two sequences along their lengths, changes dramatically halfway along the x axis. This pattern is characteristically chimeric, where one of the sequences (in this case, AY752110) is closely related to the other (AF050561) for approximately half its length yet is distinctly different thereafter. Further analysis of AY752110 confirmed this to be the case, with the 5' end, up to the approximate breakpoint at position 920, of *Verrucomicrobia* origin but with the 3' end deriving from a *Betaproteobacteria* source, as represented by AY345578.

Mallard lists those sequences identified as likely causes for the observed DE outliers. For example, 13 sequences are listed in the screenshot (Fig. 1); these were judged by the program to be suspicious. Mallard identifies these records only as likely anomalies, so they need to be further checked to confirm that they have not been falsely identified. To do this, Pintail is used to check individual records, as described previously (2). In this example, 11 of the 13 sequences were confirmed to be chimeras (AY942760, AM040116, AJ617868, AJ401133, AF316731, AJ401123, AB179538, AF449257, AF351215, AJ401131, and the already considered AY752110). A further sequence (Z94005) was shown to be poorly assembled, with roughly 130 bases

FIG. 1. Mallard program screenshot, illustrating a typical analysis. In this example, the library containing 222 16S rRNA gene sequences representing the *Verrucomicrobia* phylum is being considered. Each sequence within the library was compared with every other sequence, generating 24,531 separate DE values that were plotted against the mean percentage differences (a simple measure of evolutionary distance). Unusually high DE values are those plotted above the superimposed dotted line, and they represent comparisons in which one (or both) of the sequences is likely to be anomalous. From these outlier DE values, a list of suspected anomalies is generated (upper left-hand panel of the screenshot). Clicking on a listed sequence record causes associated DE values to be highlighted in red in the right-hand panel. Clicking on individual plotted DE values displays the underlying Pintail plot in a separate panel (not shown), and from this information, the nature of any anomaly may be discerned.

missing from the middle of the gene. Pintail analysis of the remaining sequence (AJ401106) failed to confirm an anomaly, so this was deemed a false positive.

Rerunning the analysis with the 12 confirmed anomalies removed generated the plot illustrated in Fig. 2C. Note how only DE values below the cutoff line remain, representing comparisons between reliable sequences only. For example, by selecting the DE value indicated in Fig. 2C, the Pintail plot illustrated in Fig. 2D is obtained. Note how in this plot the observed percentage difference between the two sequences is essentially constant along the length of the 16S rRNA gene; this is typical of comparisons between reliable sequences.

The 100% cutoff line, as shown in Fig. 1 and 2, provides a conservative estimate of anomaly numbers: some true anomalies will be missed. Typically, more anomalies can be uncovered with lower cutoff lines, but at the cost of more false positives (Fig. 3). With the *Verrucomicrobia* example, dropping

the cutoff line to 99.9% (Fig. 3A) revealed two further anomalies (AJ244308 and AJ401118) that were previously undetected, but also one further false positive (Fig. 3B). Dropping to 99% (Fig. 3A) identified another chimera (DQ015833), but now seven false positives were identified (Fig. 3B). Reducing the cutoff line still further failed to identify any more anomalies, but the number of false positives increased greatly (Fig. 3B). Thus, choosing a cutoff line will often be a compromise between the numbers of false positives and false negatives.

In summary, the analysis of the phylum *Verrucomicrobia* resulted in 15 anomalies being identified (6.8% of the records), 14 of which were chimeras and 1 a poorly assembled sequence.

Analysis of remaining gene libraries. An equivalent analysis of 270 near-complete sequences from the archaeal taxon *Crenarchaeota* revealed 21 anomalies (7.8% of the records). Of these, nine were clearly chimeric (AY882843, AY861964, AY882689, AB113633, AB113628, AY882728, AB113635,

FIG. 2. Mallard-generated DE plot in detail. (A) Reproduced DE plot of the *Verrucomicrobia* phylum library shown in Fig. 1 with the dotted line (the 100% cutoff line) identifying unusually high DE values (outliers), which lie above the line. Each plotted DE value represents a separate sequence comparison using the Pintail algorithm, and clicking on a plotted point within the program reveals the underlying Pintail plot. (B) The plot generated from one such comparison (between the chimera AY752110 and the error-free AF050561). The solid black line represents changes in evolutionary distance between these two sequences, when aligned, as determined from a 300-base sampling window moving 25 bases at a time along the alignment (2). The solid dark-gray line represents those evolutionary distances that one might have expected had both sequences been error free (2). The disparity between these two lines reflects the chimeric nature of AY752110. Excluding this and other chimeras identified by the program from the analysis produces the plot in panel C. DE values below the dotted cutoff line result from comparisons between error-free sequences; panel D represents a typical example, with AY212657 being compared with AB154319.

AB113631, and AB113630), seven were assembly errors with missing sequence (AF425659, U71116, U71111, U71110, X99558, AY861962, and AY861949), and five were highly degenerate (AY247896, X99559, AF425658, AF169012, and AY264344).

To demonstrate the effectiveness of Mallard in handling partial sequences, a library of 156 sequences, generated from our laboratory (20), was investigated. This library contained partial sequences ranging from 655 to 1,115 bases and four near-complete ($\geq 1,200$ -base) sequences. The partial sequences fell into two groups: those located at the 5' end of the 16S rRNA gene (82 sequences) and those derived from the 3' end (70 sequences). In total, 11 anomalies (all chimeras) were found (AY354817, AY354789, AY354824, AY354794, AY354776, AY354718, AY354851, AY354749, AY354852, AY354811, and

AY354804). A detailed breakdown of this analysis is included as a worked example with the Mallard program help documentation.

Finally, a selection of libraries generated by other authors over the preceding year (2005) were screened. Here, analysis was restricted to putative anomalies identified by a cutoff line of 100% only; thus, our results (Fig. 4; see Table S2 in the supplemental material) have underestimated the true anomaly numbers. All but three of the 25 libraries identified were found to contain anomalies. Mallard identified 714 putative anomalies; of these, 543 were subsequently confirmed to be anomalous, 493 of which showed clear chimeric patterns (see the supplemental material for a complete list of confirmed anomalies). The average (confirmed) anomaly content per library was 9.0%, with the highest

FIG. 3. Impact of cutoff line choice on correct identification of anomalies. (A) DE values from the phylum *Verrucomicrobia* analysis are plotted, with the five possible cutoff lines superimposed. (B) The numbers of true anomalies and false positives recorded for each cutoff line show that reducing the cutoff line allows more actual anomalies to be correctly identified as such but also leads to an increased number of falsely identified anomalies. The default cutoff line for the Mallard program is 99.9%, which provides a reasonable compromise between detecting as many anomalies as possible and producing the smallest number of false positives.

content recorded as 45.8% (Fig. 4; see Table S2 in the supplemental material).

Figure 4 also shows the distribution of false positives among the libraries. False positives generally occurred (i) when the library in question contained particularly high numbers of anomalies, (ii) when the DE values responsible were found to be very close (<1 DE unit) to the cutoff line (reflecting the empirical nature of the line), (iii) when conclusions could be drawn only from comparisons between distantly related ($>20\%$) sequences, or (iv) when the alignment used was inaccurate. Of the 17 libraries with associated papers, 9 stated that a chimera analysis was undertaken (see Table S2 in the supplemental material), with CHIMERA_CHECK (19) and Bellerophon (13) being used in all but two of these instances, either together or separately. No obvious trend could be discerned. Where available, we also looked at the number of PCR cycles used to generate the libraries (see Table S2 in the supplemental material), since this was considered an important chimera-generating factor (30); however, we could find no correlation between the number of PCR cycles and the number of chimeras ($r = 0.419$; $P = 0.154$; $n = 12$).

Comparison with Bellerophon results. Mallard was consistently better at correctly detecting chimeras than Bellerophon, with an average of 73.1% of known chimeras being detected per library using default settings only, in contrast to Bellerophon, where only 59.8% of chimeras were correctly identified (Table 1). Mallard was also consistently better at avoiding false positives than Bellerophon, with an average of 1.9% of library records being falsely identified as chimeric in contrast to Bellerophon's mean figure of 7.2% (Table 1). Although Mallard was consistently better at detecting chimeras, Bellerophon would sometimes detect anomalies missed by Mallard. For example, Bellerophon correctly identified the *Crenarchaeota* records AY882694 and AY882830 as chimeric, whereas they were missed by Mallard.

In using Bellerophon, we noted that different results were obtained depending on the presence of line break characters

within the Fasta-formatted input file. For example, in considering the library of Stein et al. (26), different results were obtained depending on whether sequences within the Fasta file (i) contained line breaks (resulting in two true chimeras and three false positives being identified), (ii) contained no line breaks (three chimeras; one false positive), or (iii) contained extra line breaks (two chimeras; four false positives). Clearly, there is a problem in the parsing of input files, which results in line break characters being treated as additional base characters in the analysis.

DISCUSSION

In our previous paper, we described Pintail, a novel program for screening individual sequence records for chimeras and other anomalies (2). While it is useful to have a tool that can consider a single sequence in detail, there is also a need for software that specializes in screening whole gene libraries quickly and accurately. Mallard was therefore written to meet this separate need. This paper demonstrates the ability of the Mallard program to identify putative anomalies, which can then be checked further with the Pintail tool (2), and shows that Mallard successfully detects anomalies within bacterial taxa, archaeal taxa, libraries of near-complete sequences, and libraries of partial sequences. Currently, the most widely used program for screening whole gene libraries is Bellerophon (13). In comparing Mallard with Bellerophon, the former consistently performed better than the latter in correctly identifying chimeras while minimizing false-positive results. We therefore conclude that Mallard is a significant improvement in chimera detection. However, it should also be noted that Bellerophon sometimes identified chimeras that Mallard missed. In light of this, and acknowledging the conclusions of previous studies investigating chimeras (14, 31), we believe that more than one method should be employed where feasible to detect as many chimeras as possible.

Like Bellerophon and most sequence comparison methods

FIG. 4. Analysis of near-complete ($\geq 1,200$ -base) sequences from 25 16S rRNA gene clone libraries submitted to the public repositories during 2005 (5–12, 15, 18, 21, 25, 28, 32, 33). Gene libraries are identified by the first author surname and the RDP REFID number, with the number of near-complete sequences (library size) in parentheses. The bars indicate the number of detected anomalies (identified with the 100% cutoff line) as a percentage of library size, with those anomalies confirmed as such by further investigation and false positives shown.

generally, Mallard uses aligned sequence data and is dependent on the quality of these alignments to arrive at the correct answer. In this study, we used a mixture of ClustalW alignments and alignments downloaded from the RDP website. Unlike ClustalW, the RDP's alignment procedure takes into account 16S rRNA secondary structure when constructing an alignment. Theoretically, this should make RDP alignments more accurate than ClustalW alignments; however, in practice we found that RDP alignments were sometimes inferior. An example is the RDP alignment for the gene library of Spear et al. (25), which successfully identified four chimeras but also generated 28 false-positive results; further investigation revealed that these false positives were caused by poor alignment. Realignment with ClustalW resolved the problem, and the four correctly identified chimeras were identified without extra false positives. We recommend that the user pay particular attention to the quality of the alignment when using Mallard, Bellerophon, or indeed any other alignment-based method.

In our previous study (2), we estimated that, overall, around 5% of *Bacteria* 16S rRNA gene sequence records within the public repositories have substantial errors. In our current study, we found anomaly levels of 6.8% among *Verrucomicrobia* records (*Bacteria*) and 7.8% among *Crenarchaeota* records (*Archaea*). More significantly, however, in our survey of 16S rRNA clone gene libraries submitted during 2005, we showed that the average number of anomalies per submitted library had risen to 9.0% over the course of that year. This is very likely an underestimate. Using a 100% cutoff line alone to identify putative anomalies resulted in a conservative estimate of true anomalies, and as a result, some more subtle (and not so subtle) chimeras that we know exist were excluded from our final counts.

The submitted 2005 clone libraries varied greatly in chimera content, ranging from 0 to 45.8% of the total sequence records considered. Of the 25 libraries, only 17 are currently associated with papers, and of these, the amounts of information on how libraries were constructed and checked vary greatly (for example, only nine papers actually stated that chimera detection methods were used, preventing any conclusion as to the efficacy of existing methods based on these libraries). Consequently, it is difficult to draw any conclusions as to why such a variation in chimera content has occurred. It has been speculated that increasing the number of cycles when PCR amplify-

TABLE 1. Comparison of the performance of Mallard with that of Bellerophon

Library	Size (no. of sequences)	Total no. of known chimeras ^a	Bellerophon		Mallard	
			% of known chimeras identified ^b	% of library falsely identified as chimeric ^c	% of known chimeras identified ^b	% of library falsely identified as chimeric ^c
<i>Verrucomicrobia</i>	222	15	66.7	9.5	73.3	0.5
<i>Crenarchaeota</i>	270	11	54.5	8.5	81.8	1.5
O'Sullivan et al. (20)	156	11	63.6	6.4	81.8	0.6
Stein et al. (26)	18	7	42.9	5.6	42.9	5.6
Walker et al. (29)	68	7	71.4	5.9	85.7	1.5

^a Libraries were checked manually for chimeras using the Pinatil program.

^b Means, 59.8% (Bellerophon) and 73.1% (Mallard). The number of these chimeras (expressed as a percentage of total known chimeras) detected by each program running with its respective default settings was subsequently determined.

^c Means, 7.2% (Bellerophon) and 1.9% (Mallard). Both programs also falsely identified some sequences as chimeric (confirmed by further in silico analysis), and these are quoted as a percentage of the total library size.

ing DNA can increase the chances of chimera formation (30), although no correlation between chimera generation and cycle number could be detected in the current study. The harshness of the DNA extraction method used has also been implicated in chimera formation, but even recourse to "gentle" DNA extraction methods involving detergents or enzymes does not appear to reduce the problem (17), and certainly there is insufficient information available to draw any conclusions in this regard from the 2005 clone libraries considered.

It would appear, therefore, that chimeras within 16S rRNA gene clone libraries are inevitable, at least with current PCR methodologies. Previously, it had been estimated that up to 30% of individual PCR-generated clone libraries were likely to be chimeric (17, 30, 31). We cannot comment on how many chimeras were originally generated by the researchers considered in this study, but we note that libraries with up to 45.8% chimeras are being submitted without comment to the public repositories. Serious anomalies are polluting the public repositories to such an extent that their usefulness is being surreptitiously and progressively compromised. The effects are already being felt; for example, some putative chimeras were especially difficult to check during the current study because so many anomalies had been submitted for the taxa they supposedly represent.

This study indicates that most libraries submitted during 2005 contain misleading anomalies, and the average anomaly content per library is estimated to be 4% higher than the 5% estimated previously for the public repository overall. Moreover, our results show that the vast majority of these errors are now chimeras—the most insidious and misleading of anomalies. At least 90.8% of the anomalies considered in this study had chimeric patterns, which contrasts dramatically with the 64.3% of anomalies reported previously (2). Our previous study showed that between 1993 and 2004 a steadily increasing number of chimeras were submitted to the NCBI database (2), at least among the phyla investigated by that study. Overall, we conclude that the specific problem of chimeric 16S rRNA sequences in the public databases is at best not improving and at worst is becoming more acute. We offer our software free to the wider research community in the hope that it will complement existing methods to ensure that as few chimeras and other anomalies as possible are submitted in future.

ACKNOWLEDGMENT

This study was supported by grant BBS/B/11494 from the Biotechnology and Biological Sciences Research Council (BBSRC).

REFERENCES

- Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389–3402.
- Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* **71**:7724–7736.
- Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. 2000. GenBank. *Nucleic Acids Res.* **28**:15–18.
- Cole, J., B. Chai, T. Marsh, R. Farris, Q. Wang, S. Kulm, S. Chandra, D. McFarrell, T. Schmidt, G. Garrity, and J. Tiedje. 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res.* **31**:442–443.
- Crump, B. C., and J. E. Hobbie. 2005. Synchrony and seasonality in bacterioplankton communities of two temperate rivers. *Limnol. Oceanogr.* **50**:1718–1729.
- Dunn, A. K., and E. V. Stabb. 2005. Culture-independent characterization of the microbiota of the ant lion *Myrmeleon mobilis* (Neuroptera: Myrmeleontidae). *Appl. Environ. Microbiol.* **71**:8784–8794.
- Eckburg, P. B., E. M. Bik, C. N. Berstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman. 2005. Diversity of the human intestinal microbial flora. *Science* **308**:1635–1638.
- Fuchs, B. M., D. Woeckel, M. V. Zubkov, P. Burkill, and R. Amann. 2005. Molecular identification of picoplankton populations in contrasting waters of the Arabian Sea. *Aquat. Microb. Ecol.* **39**:145–157.
- Gihring, T., D. P. Moser, L.-H. Lin, M. Davidson, T. C. Onstott, L. Morgan, M. Millese, T. L. Kieft, E. Trimarco, D. L. Balkwill, and M. E. Dollhopf. The distribution of microbial taxa in the subsurface water of the Kalahari Shield, South Africa. *Geomicrobiol. J.*, in press.
- Graff, A., and R. Conrad. 2005. Impact of flooding on soil bacterial communities associated with poplar (*Populus* sp.) trees. *FEMS Microbiol. Lett.* **234**:401–415.
- Hongoh, Y., P. Deevong, T. Inoue, S. Moriya, S. Trakulnaleamsai, M. Ohkuma, C. Vongkalluang, N. Noparatnaraporn, and T. Kudo. 2005. Intra- and interspecific comparisons of bacterial diversity and community structure support coevolution of gut microbiota and termite host. *Appl. Environ. Microbiol.* **71**:6590–6599.
- Hongoh, Y., L. Ekporpraisit, T. Inoue, S. Moriya, S. Trakulnaleamsai, M. Ohkuma, N. Noparatnaraporn, and T. Kudo. 2006. Intracolony variation of bacterial gut microbiota among castes and ages in the fungus-growing termite *Macrotermes gilvus*. *Mol. Ecol.* **15**:505–516.
- Huber, T., G. Faulkner, and P. Hugenoltz. 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**:2317–2319.
- Hugenoltz, P., and T. Huber. 2003. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int. J. Syst. Evol. Microbiol.* **53**:289–293.
- Hyman, R. W., M. Fukushima, L. Diamond, J. Kumm, L. C. Giudice, and R. W. Davis. 2005. Microbes on the human vaginal epithelium. *Proc. Natl. Acad. Sci. USA* **102**:7952–7957.
- Kanz, C., P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. V. D. Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F. G. Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, S. Sobhani, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler. 2005. The EMBL nucleotide sequence database. *Nucleic Acids Res.* **33**:D29–D33.
- Kopczynski, E. D., M. M. Bateson, and D. M. Ward. 1994. Recognition of chimeric small-subunit ribosomal DNAs composed of genes from uncultured microorganisms. *Appl. Environ. Microbiol.* **60**:746–748.
- Ley, R. E., B. Backhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight, and J. I. Gordon. 2005. Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. USA* **102**:11070–11075.
- Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker, Jr., P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje. 2001. The RDP-II (Ribosomal Database Project). *Nucleic Acids Res.* **29**:173–174.
- O'Sullivan, L. A., K. E. Fuller, E. M. Thomas, C. M. Turley, J. C. Fry, and A. J. Weightman. 2004. Distribution and culturability of the uncultivated 'AGG58 cluster' of the *Bacteroidetes* phylum in aquatic environments. *FEMS Microbiol. Ecol.* **47**:359–370.
- Ozutsu, Y., K. Tajima, A. Takenaka, and H. Itabashi. 2005. The effect of protozoa on the composition of rumen bacteria in cattle using 16S rRNA gene clone libraries. *Biosci. Biotechnol. Biochem.* **63**:499–506.
- Pääbo, S., D. M. Irwin, and A. C. Wilson. 1990. DNA damage promotes jumping between templates during enzymatic amplification. *J. Biol. Chem.* **265**:4718–4721.
- Rappe, M. S., and S. J. Giovannoni. 2003. The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**:369–394.
- Shuldiner, A., A. Nirula, and J. Roth. 1989. Hybrid DNA artifact from PCR of closely related target sequences. *Nucleic Acids Res.* **17**:4409.
- Spear, J. R., J. J. Walker, T. M. McCollom, and N. R. Pace. 2005. Hydrogen and bioenergetics in the Yellowstone geothermal ecosystem. *Proc. Natl. Acad. Sci. USA* **102**:2555–2560.
- Stein, L. Y., M. T. La Duc, T. J. Grundl, and K. H. Nealson. 2001. Bacterial and archaeal populations associated with freshwater ferromanganese micronodules and sediments. *Environ. Microbiol.* **3**:10–18.
- Thompson, J., D. Higgins, and T. Gibson. 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenoltz, and E. M. Rubin. 2005. Comparative metagenomics of microbial communities. *Science* **308**:554–557.
- Walker, J. J., J. R. Spear, and N. R. Pace. 2005. Geobiology of a microbial endolithic community in the Yellowstone geothermal environment. *Nature* **434**:1011–1014.
- Wang, G. C.-Y., and Y. Wang. 1996. The frequency of chimeric molecules as

- a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology* **142**:1107–1114.
31. **Wang, G. C.-Y., and Y. Wang.** 1997. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl. Environ. Microbiol.* **63**:4645–4650.
32. **Yoshida, N., N. Takahashi, and A. Hiraishi.** 2005. Phylogenetic characterization of a polychlorinated-dioxin-dechlorinating microbial community by use of microcosm studies. *Appl. Environ. Microbiol.* **71**:4325–4334.
33. **Zeng, R., J. Zhao, R. Zhang, and N. Lin.** 2005. Bacterial community in sediment from the western Pacific Warm Pool and its relationship to environment. *China Environ. Sci.* **48**:282–290.