

1 **New screening software shows most recent large 16S rRNA gene clone libraries**
2 **contain chimeras.**

3

4 **Running title**

5 Detecting chimeras within 16S rRNA gene libraries.

6

7 **Authors**

8 Kevin E. Ashelford*, Nadia A. Chuzhanova[†], John C. Fry, Antonia J. Jones[‡], and Andrew J.

9 Weightman

10

11 **Contact details**

12 Cardiff School of Biosciences, Cardiff University, Main Building, Park Place, PO Box 915, Cardiff,

13 CF10 3TL, UK

14 [†]Biostatistics & Bioinformatics Unit and Institute of Medical Genetics, Cardiff School of Medicine,

15 Cardiff University, Heath Park, Cardiff, CF14 4XN, UK

16 [‡]Cardiff School of Computer Science, Cardiff University, Queen's Buildings, 5 The Parade, Roath,

17 Cardiff, CF24 3AA, UK

18

19 **Correspondent footnote**

20 *Corresponding author: email, ashelford@cardiff.ac.uk; telephone, +44 (0)29 20 876002; Fax, +44

21 (0)29 20 874305.

ABSTRACT

1
2 A new computer program, called Mallard, is presented for screening entire 16S rRNA gene
3 libraries, of up to 1,000 sequences, for chimeras and other artifacts. Written in the Java computer
4 language, and capable of running on all major operating systems, the program provides a novel
5 graphical approach for visualizing phylogenetic relationships among 16S rRNA gene sequences. To
6 illustrate its use, we analyzed most of the large libraries of cloned bacterial 16S rRNA gene sequences
7 submitted during 2005. Defining a large library as one containing 100 or more sequences of 1,200
8 bases or greater, we screened 25 of the available 28 libraries and found all but three contained
9 substantial anomalies. Overall, 543 anomalous sequences were found, 90.8% of which had
10 characteristic chimeric patterns. Average anomaly content per clone library was 9.0%. One library
11 alone was found to contain 54 chimeras, representing 45.8% of its content. These figures far exceed
12 previous estimations of artifacts within public repositories and further highlight the urgent need for all
13 researchers to adequately screen their libraries prior to submission. To this end we offer Mallard to the
14 wider research community, which is freely available from our website at
15 <http://www.cardiff.ac.uk/biosi/research/biosoft/>.

INTRODUCTION

18 Recent papers (2, 6) have reported numerous corrupt 16S rRNA gene sequences within the
19 public repositories (3, 7, 9), and it has been estimated that, overall, 5% of records are likely to have
20 substantial anomalies (2). Whilst poor sequencing and errors during assembly have led to some of
21 these reported errors, most anomalies have been chimeras - artificial sequences generated from two or
22 more phylogenetically different DNA templates during PCR amplification (8, 11, 12, 13, 15, 16).

23 Our previous study has shown that chimeras, and other anomalies, are continuing to be
24 generated, and submitted without comment to the public repositories (2). The presence of such high

1 numbers of substantial anomalies in the public domain has serious implications for future efforts to
2 accurately estimate bacterial diversity, elucidate likely phylogenetic relationships, and form correct
3 taxonomic identifications. Anomalies must be excluded from 16S rRNA gene clone libraries prior to
4 submission, or at least be clearly annotated as such. Consequently, there is a requirement for effective
5 computer programs to simplify the screening process.

6 A number of useful, complementary approaches already exist, with Bellerophon (5) and
7 CHIMERA_CHECK (9) being two noteworthy examples. And in our previous paper we described a
8 new computer program, called Pintail, for screening individual sequences for errors (2). Now we
9 describe a further program, Mallard, which develops the Pintail algorithm further in order that whole
10 libraries of 16S rRNA gene sequences can be screened simultaneously and quickly.

11 We demonstrate the new program's ability to screen libraries of a range of sizes from different
12 sources. Through a detailed analysis of submissions made to public repositories during 2005, we show
13 that the problem of unrecognized anomalies within the public domain is getting substantially worse,
14 highlighting the need for immediate steps to be taken, by the research community at large, to minimize
15 further database contamination.

16

17

MATERIALS AND METHODS

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

1 resulting summarizing statistic – the Deviation from Expectation (DE) value, $DE = \sqrt{\frac{\sum_1^m (o_i - e_i)^2}{m-1}}$ –
 2 quantifies the likelihood that an anomaly is present. A more thorough description of the Pintail
 3 algorithm, including an explanation on how E_{qs} is calculated can be found in the help documentation
 4 accompanying this software and in (2).

5 In our new program, the Pintail algorithm is now applied to all pairwise comparisons within a
 6 multiple alignment, of size n , resulting in $(n^2-n)/2$ separate DE values, each DE value presenting a
 7 unique pairwise comparison. DE values are plotted against their corresponding mean observed
 8 percentage differences $(\sum_i o_i)/m$, which can be viewed as a simple measure of evolutionary distance
 9 between sequences S_q and S_s . The larger the DE value, the greater the likelihood that either S_q or S_s (or
 10 perhaps even both) is in some way corrupt. Thus, by plotting DE values one can immediately see
 11 which pairwise comparisons are likely to involve an anomalous sequence, since DE values generated
 12 from reliable sequences will tend to cluster close to the x -axis, whilst DE values involving anomalous
 13 sequences will be plotted relatively distant from the x -axis and thus appear as outliers.

14 In Mallard, outliers are identified as those DE values which appear above one of several
 15 possible cut-off lines, specified by the user, and based on DE values calculated from comparisons of
 16 error-free sequences from type strains (2). Specifically, in our earlier study we calculated DE values
 17 from a collection of 2,007 reliable type-strain sequences; the 75, 95, 99, 99.9, and 100% quantiles of
 18 the resulting plot were determined at each 1% interval along the x -axis (2). These quantile data give
 19 roughly straight lines when plotted in a logarithmic scale, so for this study, the quantile data was
 20 simplified to the following equations: 75% quantiles, $y = 2.28\text{Log}_{10}x + 1.00$; 95% quantiles, $y =$
 21 $2.64\text{Log}_{10}x + 1.46$; 99% quantiles, $y = 3.12\text{Log}_{10}x + 1.66$; 99.9% quantiles, $y = 3.27\text{Log}_{10}x + 2.07$;
 22 100% quantiles, $y = 4.37\text{Log}_{10}x + 1.81$. Cut-off lines, generated from these equations, are offered by
 23 the program.

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

10 Mallard was written in Java 1.4 (Java Technology; <http://java.sun.com/>) and tested on Redhat
11 9.0 Linux, Microsoft Windows XP, and Apple Mac OS X, version 10.2. The program, along with full
12 instructions for use, help documentation, example files, and source code, is freely available from
13 <http://www.cardiff.ac.uk/biosi/research/biosoft/>. Mallard is an open source project and is released
14 under the terms of the GNU General Public Licence (<http://www.gnu.org/copyleft/gpl.html>).

15 **Analysis of 16S rRNA gene libraries.** To demonstrate Mallard's utility, a selection of publicly
16 available 16S rRNA gene libraries was analyzed. The procedure was the same for each library; a
17 multiple sequence alignment was prepared for each that included the sequence *Escherichia coli*
18 U00096 (as reference sequence). Each multiple sequence alignment was passed to the Mallard
19 program and screened for anomalies. A full description of how to use the Mallard program, along with
20 an explanation for the reference sequence, is included in the accompanying help documentation. Each
21 putative anomaly, identified by the program, was checked with BlastN (1) and the Pintail program (2).

22 First a library of *Verrucomicrobia*-derived sequences, to exemplify a *Bacteria* phylum, was
23 considered. A total of 222 near-complete ($\geq 1,200$ base) representatives of the *Verrucomicrobia*, as
24 identified by the Ribosome Database Project (RDP; 4) release 9 update 36, were downloaded, along

1 high and marked as outliers. DE outliers typically result from sequence comparisons where at least one
2 of the pair contains errors, so the DE values above the cut-off line in Figure 1 (and 2A) are likely to be
3 the result of anomalous sequences within the *Verrucomicrobia* library.

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

16 Mallard lists those sequences identified as likely causes for the observed DE outliers. For
17 example, 13 sequences are listed in the screenshot (Fig. 1); these were judged by the program to be
18 suspicious and needed further checking. In doing so, 11 were found to be chimeras (AY942760,
19 AM040116, AJ617868, AJ401133, AF316731, AJ401123, AB179538, AF449257, AF351215,
20 AJ401131, and the already considered AY752110). A further sequence (Z94005) was poorly
21 assembled, with roughly 130 bases missing from the middle of the gene. Analysis of the remaining
22 sequence (AJ401106) failed to confirm an anomaly and so this was deemed a false positive.

23 Re-running the analysis, with the 12 confirmed anomalies removed, generates the plot
24 illustrated (Fig. 2C). Note how only DE values below the cut-off line remain, representing as they do,

1 comparisons between reliable sequences only. For example, selecting the DE value indicated in Figure
2 2C, the Pintail plot illustrated in Figure 2D is obtained. Observe how in this plot the observed
3 percentage difference between the two sequences is essentially constant along the length of the 16S
4 rRNA gene; this is typical of comparisons between reliable sequences.

5 The 100% cut-off line, as shown in Figures 1 and 2, provides a conservative estimate of
6 anomaly numbers: some true anomalies will be missed. Typically, more anomalies can be uncovered
7 with lower cut-off lines, but this is at the cost of more false positives (Fig. 3). With the
8 *Verrucomicrobia* example, dropping the cut-off line to 99.9% (Fig. 3A) revealed two further anomalies
9 (AJ244308 and AJ401118) previously undetected, but also one further false positive (Fig. 3B).
10 Dropping to 99% (Fig. 3A) identified another chimera (DQ015833), but now seven false positives were
11 identified (Fig. 3B). Reducing the cut-off line further still failed to identify any more anomalies, but
12 the number of false positives increased greatly (Fig. 3B). Thus in choosing a cut-off line will often be
13 a compromise between number of false positives and false negatives.

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

17 **Analysis of remaining gene libraries.** An equivalent analysis of 270 near-complete sequences
18 from the archaeal taxon *Crenarchaeota* revealed 21 anomalies (7.8% of records). Of these, 9 were
19 clearly chimeric (AY882843, AY861964, AY882689, AB113633, AB113628, AY882728, AB113635,
20 AB113631, AB113630), 7 were assembly errors, with missing sequence (AF425659, U71116, U71111,
21 U71110, X99558, AY861962, AY861949), and 5 were highly degenerate (AY247896, X99559,
22 AF425658, AF169012, AY264344).

23 To demonstrate the effectiveness of our program in handling partial sequences, a library of 156
24 sequences, generated from our laboratory (10), was investigated. This library contained partial

1 sequences ranging from 655 to 1,115 bases and 4 near-complete ($\geq 1,200$ base) sequences. The partial
2 sequences fell into two groups; those located at the 5' end of the 16S rRNA gene (82 sequences), and
3 those derived from the 3' end (70 sequences). In total, 11 anomalies (all chimeras) were found
4 (AY354817, AY354789, AY354824, AY354794, AY354776, AY354718, AY354851, AY354749,
5 AY354852, AY354811 and AY354804). A detailed breakdown of this analysis is included, as a
6 worked example, with the Mallard program help documentation.

7 Finally, a selection of libraries generated by other authors over the preceding year (2005) were
8 screened. Here analysis was restricted to putative anomalies identified by a cut off line of 100% only;
9 thus our results (Fig. 4) will have underestimated true anomaly numbers. All but three of the 25
10 libraries identified were found to contain anomalies. Mallard identified 714 putative anomalies; of
11 these, 543 were subsequently confirmed anomalous of which, 493 showed clear chimeric patterns. See
12 supplementary data for a complete list of confirmed anomalies. The average (confirmed) anomaly
13 content per library was 9.0% with the highest content being recorded as 45.8% (Fig. 4).

14 Figure 4 also shows the distribution of false positives among the libraries. False positives
15 generally occurred when: (i) the library in question contained particularly high numbers of anomalies;
16 or (ii) when the DE values responsible were found to be very close (< 1 DE unit) to the cut-off line
17 (reflecting the empirical nature of the line); or (iii) when conclusions could only be drawn from
18 comparisons between distantly related ($> 20\%$) sequences; or (iv) when the alignment used was
19 inaccurate.

21 DISCUSSION

22 This paper demonstrates the ability of the Mallard program to detect anomalies within bacterial
23 taxa, archaeal taxa, libraries of near-complete sequences, and libraries of partial sequences. The
24 software was developed to be user-friendly and capable of running on as many computer platforms as

1 possible to encourage its use. We offer our software free to the wider research community in the hope
2 that it will complement existing methods for chimera detection.

3 In our previous study (2) we estimated that, overall, around 5% of *Bacteria* 16S rRNA gene
4 sequence records within the public repositories will have substantial errors. In our current study we
5 find anomaly levels of 6.8% among *Verrucomicrobia* records (*Bacteria*) and 7.8% among
6 *Crenarchaeota* records (*Archaea*). More significantly however, from our survey of 16S rRNA clone
7 gene libraries submitted during 2005, we show that the average number of anomalies per submitted
8 library has risen to 9.0% over the last year. This is surely an underestimate. By using a 100% cut-off
9 line alone to identify putative anomalies, we arrived at a conservative estimate of true anomalies and,
10 as a result, some more subtle (and not so subtle) chimeras, that we know to exist, were excluded from
11 our final counts.

12 Overall, we conclude that the problem of erroneous sequences in the public databases for PCR-
13 generated 16S rRNA gene sequences is becoming more acute. Moreover, our results show that the vast
14 majority of these errors will be chimeras – the most insidious and misleading of anomalies. At least
15 90.8% of the anomalies considered in this study had chimeric patterns, which contrasts dramatically to
16 the 64.3% of anomalies reported previously (2). This suggests to us that recent research trends have
17 resulted in the widespread adoption of methodologies which, whilst undeniably useful, have
18 nevertheless also led to an explosion in chimera generation.

19 Chimeras within 16S rRNA gene clone libraries are inevitable, at least with current PCR
20 methodologies (8, 15, 16). All of us who generate clone libraries, need to be aware of this fact.
21 Previously, it has been estimated that up to 30% of PCR generated clones will be chimeras (8, 15, 16).
22 This current study has shown that libraries, with up to 45.8% chimeras, are not only being generated
23 but also being submitted, without comment, to the public repositories. It is vital that this situation does
24 not persist. Serious anomalies are polluting the public repositories to an extent that their usefulness is

1 being surreptitiously and progressively compromised. The effects are already being felt. In this study
2 for example, some putative chimeras were especially difficult to check because so many anomalies had
3 already been submitted for the taxa they 'represent'.

4 Current screening procedures are clearly not working (and it is clear to us from this study that in
5 some cases screening is not occurring at all). This is not to criticize existing individual methods for
6 chimera detection. Due to the nature of the problem, no one method can be foolproof and so no one
7 method should be used in isolation. This caution applies to the software presented in this study as
8 much as any other. We do believe our new software is a useful new approach that is capable of
9 detecting chimeras that other approaches sometimes miss – but we also know it should not be viewed
10 as a panacea. Ultimately we all, as researchers, need to employ a suite of anomaly detection methods
11 to be as certain as possible we have only reliable sequences. It is the responsibility of all of us to
12 ensure that every possible effort is taken to ensure only error-free sequences are added to the public
13 repositories.

14

15

ACKNOWLEDGMENT

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

18

19

REFERENCES

- 20 1. **Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman.** 1997.
21 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic*
22 *Acids Research* **25**:3389-3402.
- 23 2. **Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman.** 2005. At
24 least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to

- 1 contain substantial anomalies. *Applied and Environmental Microbiology* **71**:7724-7736.
- 2 3. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler.**
3 2000. GenBank. *Nucleic Acids Research* **28**:15-18.
- 4 4. **Cole, J., B. Chai, T. Marsh, R. Farris, Q. Wang, S. Kulum, S. Chandra, D. McGarrell, T.**
5 **Schmidt, G. Garrity, and J. Tiedje.** 2003. The Ribosomal Database Project (RDP-II): previewing
6 a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids*
7 *Research* **31**:442-443.
- 8 5. **Huber, T., G. Faulkner, and P. Hugenholtz.** 2004. Bellerophon: a program to detect chimeric
9 sequences in multiple sequence alignments. *Bioinformatics* **20**:2317-2319.
- 10 6. **Hugenholtz, P., and T. Huber.** 2003. Chimeric 16S rDNA sequences of diverse origin are
11 accumulating in the public databases. *International Journal of Systematic and Evolutionary*
12 *Microbiology* **53**:289-293.
- 13 7. **Kanz, C., P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. v. d.**
14 **Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F. G. Diez,**
15 **N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone,**
16 **V. Silventoinen, S. Sobhany, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu,**
17 **and R. Apweiler.** 2005. The EMBL Nucleotide Sequence Database. *Nucleic Acids Research*
18 **33**:D29-D33.
- 19 8. **Kopczynski, E. D., M. M. Bateson, and D. M. Ward.** 1994. Recognition of chimeric small-
20 subunit ribosomal DNAs composed of genes from uncultured microorganisms. *Applied and*
21 *Environmental Microbiology* **60**:746-748.
- 22 9. **Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker Jr, P. R. Saxman, R. J. Farris, G. M.**
23 **Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje.** 2001. The RDP-II (Ribosomal Database
24 Project). *Nucleic Acids Research* **29**:173-174.

- 1 10. **O’Sullivan, L. A., K. E. Fuller, E. M. Thomas, C. M. Turley, J. C. Fry and A. J. Weightman**
2 (2004) Distribution and culturability of the uncultivated ‘AGG58 cluster’ of the *Bacteroidetes*
3 phylum in aquatic environments. *FEMS Microbiology Ecology* **47**:359-370.
- 4 11. **Pääbo, S., D. M. Irwin, and A. C. Wilson.** 1990. DNA damage promotes jumping between
5 templates during enzymatic amplification. *Journal of Biological Chemistry* **265**:4718-4721.
- 6 12. **Rappe, M. S., and S. J. Giovannoni.** 2003. The uncultured microbial majority. *Annual Review of*
7 *Microbiology* **57**:369-394.
- 8 13. **Shuldiner, A., A. Nirula, and J. Roth.** 1989. Hybrid DNA artifact from PCR of closely related
9 target sequences. *Nucleic Acids Research* **17**:4409.
- 10 14. **Thompson, J., D. Higgins, and T. Gibson.** 1994. Clustal W: improving the sensitivity of
11 progressive multiple sequence alignment through sequence weighting, positions-specific gap
12 penalties and weight matrix choice. *Nucleic Acids Research* **22**:4673-4680.
- 13 15. **Wang, G. C.-Y., and Y. Wang.** 1996. The frequency of chimeric molecules as a consequence of
14 PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology* **142**:1107-
15 1114.
- 16 16. **Wang, G. C.-Y., and Y. Wang.** 1997. Frequency of formation of chimeric molecules as a
17 consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Applied*
18 *and Environmental Microbiology* **63**:4645-4650.

FIGURE LEGENDS

21 **Figure 1**

22 Mallard program screenshot, illustrating a typical analysis. In this example, the 222 16S rRNA gene
23 sequence library, representing the *Verrucomicrobia* phylum, is being considered. Each sequence
24 within the library has been compared with each other, generating 24,531 separate DE values that have

1 been plotted against mean percentage differences (a simple measure of evolutionary distance).
2 Unusually high DE values are those plotted above the superimposed dotted line, and represent
3 comparisons where one (or both) of the sequences are likely to be anomalous. From these outlier DE
4 values, a list of suspected anomalies is generated (upper left-hand panel of the screenshot). Clicking on
5 a listed sequence record causes associated DE values to be highlighted red in the right-hand panel.
6 Clicking on individual plotted DE values displays the underlying Pintail plot in a separate panel (not
7 shown), and from this information the nature of any anomaly may be discerned.

8

9 **Figure 2**

10 Mallard generated DE plot in detail. Panel A reproduces the DE plot of the *Verrucomicrobia* phylum
11 library shown previously (Fig. 1), with the black dotted line (the 100% cut-off line) identifying
12 unusually high DE values (outliers) above this line. Each plotted DE value represents a separate
13 sequence comparison using the Pintail algorithm, and clicking on a plotted point within the program
14 reveals the underlying Pintail plot. Panel B shows the plot generated from one such comparison
15 (between the chimera AY752110 and the error-free AF050561). The solid black line represents
16 changes in evolutionary distance between these two sequences, when aligned, as determined from a
17 300 base sampling window moving 25 bases at a time along the alignment (2). The solid dark grey line
18 represents those evolutionary distances that one might have expected had both sequences been error-
19 free (2). The disparity between these two lines reflects the chimeric nature of AY752110. Excluding
20 this, and other chimeras identified by the program, from the analysis, produces the plot in panel C. DE
21 values below the dotted cut-off line result from comparisons between error-free sequences, panel D
22 representing a typical example with AY212657 being compared with AB154319.

23

1 **Figure 3**

2 Impact of cut-off line choice on correct identification of anomalies. DE values from the
3 *Verrucomicrobia* phylum analysis are plotted, with the 5 possible cut-off lines superimposed (panel A).
4 The number of true anomalies (black bars) and false positives (white bars) recorded for each cut-off
5 line show that reducing the cut-off line allows more actual anomalies to be correctly identified as such,
6 but also leads to an increasing number of falsely identified anomalies (panel B). The default cut-off
7 line for the Mallard program is 99.9% - providing, as it does, a reasonable compromise between
8 detecting as many anomalies as possible with the least number of false positives.

9

10 **Figure 4**

11 Analysis of near-complete ($\geq 1,200$ base) sequences from 25 16S rRNA gene clone libraries submitted
12 to the public repositories during 2005. Gene libraries are identified by first author surname and RDP
13 REFID number, with number of near-complete sequences (library size) in parentheses. Bars indicate
14 number of detected anomalies (identified with the 100% cut-off line), as a percentage of library size,
15 with black bars showing those anomalies confirmed as such by further investigation, and white bars
16 showing false positives.

17

18



List of anomalous sequences

Name	Highest DE difference	No. of outliers	Selected	Result
Z94005	11.58	221	✓	!
AY752110	5.47	19	✓	!
AY942760	4.77	3	✓	!
AM040116	2.75	11	✓	!
AJ617868	2.00	7	✓	!
AJ401133	1.87	5	✓	!
AF316731	1.87	7	✓	!
AJ401123	0.74	4	✓	?
AB179538	0.61	3	✓	?
AF449257	0.59	3	✓	?
AF351215	0.52	4	✓	?
AJ401131	0.38	2	✓	?
AJ401106	0.38	2	✓	?

Selected sequence

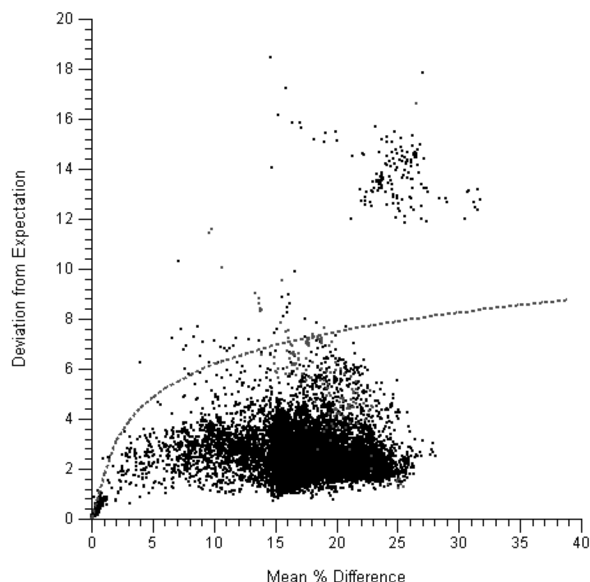
>AY752110

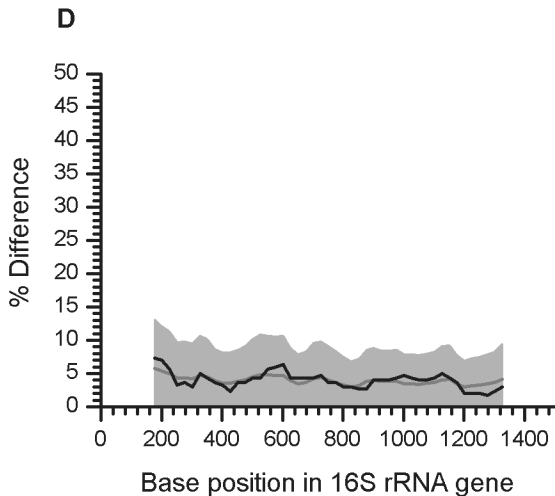
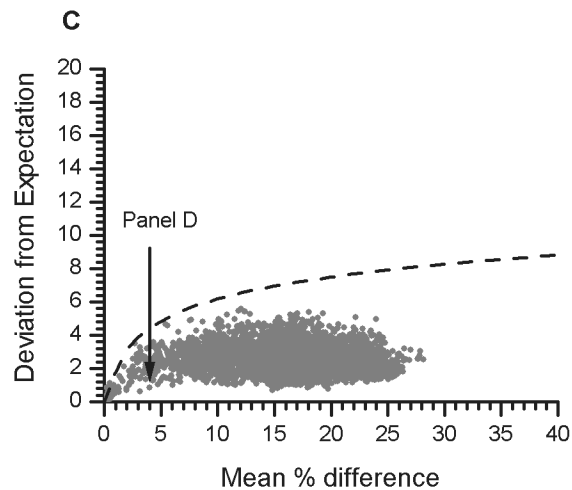
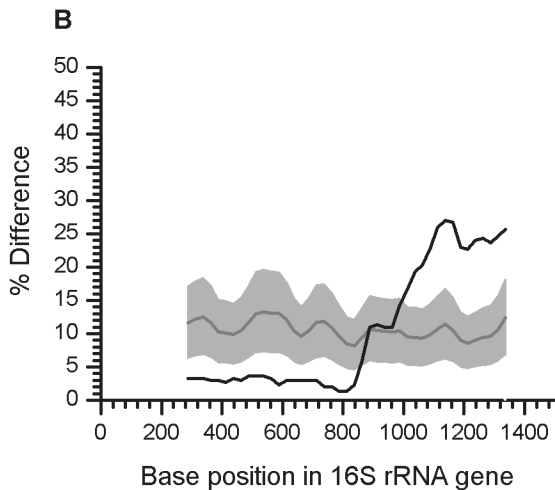
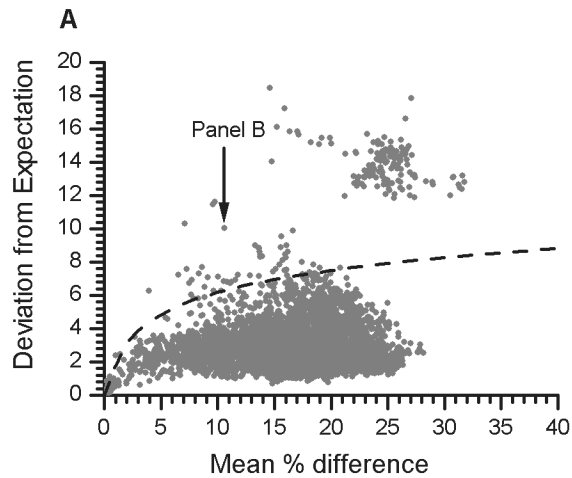
```

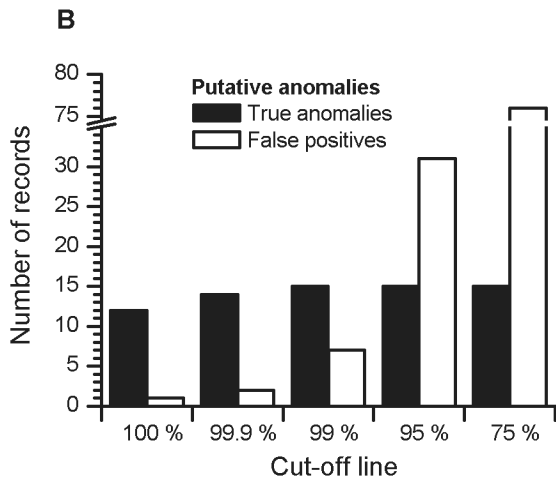
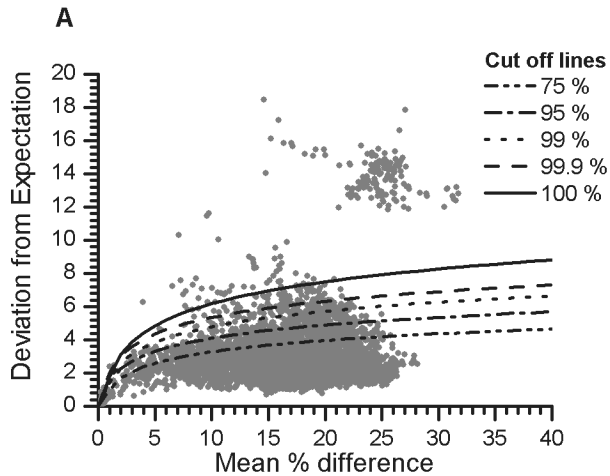
AACACGTGAGTAATCTGCCGGGAAGTGGGGGATAGCTCGCCGAAAGGCGAA
TTAATACCGCATATGCAGGGAAGACATCTTTCCGAAGTCAAAGTCGCAA
GACGCTTCTCAGTACGCTCGCGCCTATCAGCTAGTTGGTGAGGTAACGGC
TCACCAAGGCAATGACGGGTAGCTGGTCTGAGAGGACGACCGCACACTG
GAACTGAGACACGGTCCAGACACCTACGGGTGGCAGCAGTCGAGAAATTTT
CACAAATGGGGAAAACCTGTATGAGCGCAGCCGCCGTGGAGGATGAAGGTCT
TCGGATTGTAAACTCTCTGTATCGCGAGAACAAGAAAGTGATAGTATCGCAA
GAGGAAGAGACGGCTAACTCTGTGCCAGCAGCCGCGTAATACAGAGGTCT
CAAGCGTTGTTCCGGATTCATTGGGGTAAGGGTGCCTAGGTGGCGTGGAAA
GTTGAGTGTGAAATCTCAGGGCTTAACTTAGAAGTGCACCTCAATACTCCC
ATGCTAGAGGAATGTAGAGGAGAGTGGAAATTCACGGGTGAGCAGTAAATG
CGTAGATATCGTGAGGAAGACCAGTTGCCAAGGGCAGCTCTCTGGGCATTTT
CTGACACTGAGGCACGAAGGCCAGGGGAGCAAATGGGATTAGATACCCCGG
TAGTCTGGCAGTAAACGGTGCACGTTTGGTGTGGGGGGCTCAGACCCCG

```

Analysis







16S rRNA gene library

