

1 **At least one in twenty 16S rRNA sequence records currently held in public**
2 **repositories estimated to contain substantial anomalies.**

3

4 **Running title**

5 Substantial anomalies in public 16S rRNA repositories

6

7 **Authors**

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

9 Weightman

10

11 **Contact details**

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

13 Cardiff, CF10 3TL, UK

14 [†]Cardiff School of Computer Science, Cardiff University, Queen's Buildings, 5 The Parade,

15 Roath, Cardiff, CF24 3AA, UK

16 [‡]Biostatistics & Bioinformatics Unit and Institute of Medical Genetics, Cardiff School of

17 Medicine, Cardiff University, Heath Park, Cardiff, CF14 4XN, UK

18

19 **Correspondent footnote**

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

21 +44 (0)29 20 874305.

1 **Abstract**

2 A new method for detecting chimeras and other anomalies within 16S rRNA sequence
3 records is presented. Using this method we screened 1,399 sequences from 19 phyla, as defined
4 by the Ribosomal Database Project (RDP-II), release 9, update 22, and found 5.0% to harbour
5 substantial errors. Of these, 64.3% were obvious chimeras, 14.3% were unidentified sequencing
6 errors, and 21.4% were highly degenerate. In all, 11 phyla contained obvious chimeras,
7 accounting for 0.8 to 11% of these phyla's records. Many chimeras (43.1%) were formed from
8 parental sequences belonging to different phyla. Whilst most comprised of two fragments, 13.7%
9 were composed from at least three fragments, often from three different sources. Overall we
10 conclude that, as a conservative estimate, one in every twenty public database records is likely to
11 be corrupt. Our results support concerns recently expressed over the quality of the public
12 repositories. With 16S rRNA sequence data increasingly playing a dominant role in bacterial
13 systematics and environmental biodiversity studies, it is vital that steps are taken to improve
14 screening of sequences prior to submission. To this end, we have implemented our method as a
15 program with a simple-to-use graphic user interface that is capable of running on a range of
16 computer platforms. The program is called Pintail, is released under the terms of the GPL open
17 source license, and is freely available from our website at
18 <http://www.cardiff.ac.uk/biosi/research/biosoft/>.

19

20 **Introduction**

21 Analysis of the 16S rRNA gene is currently fundamental to an understanding bacterial
22 taxonomy, phylogeny and diversity (3, 5). Sequence anomalies, if undetected, can generate
23 misleading impressions of environmental diversity and complicate attempts to re-construct
24 bacterial evolutionary trees. It is vital, therefore, that public repositories such as those managed
25 by EMBL (9), GenBank (2), and the Ribosome Database Project, RDP-II (3) contain reliable

1 sequences if correct conclusions are to be made within studies that rely on 16S rRNA sequence
2 analysis.

3 Unfortunately, corrupt sequences such as chimeras formed during PCR amplification (12,
4 14, 15, 20, 21), or anomalies produced by other steps in the sequencing process, have long been
5 present in the public databases. Poor sequencing methodology often produces highly degenerate
6 sequences; these are easy to spot. More insidious are other sequencing errors that can not be
7 detected by a visual inspection of the sequence alone. Chimeras, sometimes referred to as
8 'jumping PCR products', 'shuffle-genes' or 'in vitro recombination products' have been a
9 recognised PCR amplification problem for some time (e.g. 17), with damage or degradation to
10 the DNA template, and contamination with other templates being likely causes of their formation
11 (e.g. 14). Chimeras have been shown to occur in PCR amplified gene libraries with frequencies
12 of 30% or more (12, 20, 21) and therefore pose a potentially significant problem.

13 Chimeric anomalies have long been recognised and several computational methods have
14 been developed over the years to detect and analyse suspect sequences (6, 7, 10 11, 13, 16).
15 Historically, the RDP's Chimera_Check program (13) has been used most widely, although the
16 more recent Bellerophon program (7) appears to be gaining in popularity. However, existing
17 tools for chimera detection, though often effective, have limitations (8, 11, 16 and 21). Also,
18 most of these tools have not been developed into sufficiently accessible computer programs that
19 can be used easily by researchers regardless of computing background. One reason for the
20 widespread use of RDP's Chimera_Check program is that it has a user-friendly interface and is
21 available to anyone with a web browser.

22 Most importantly, the problem of chimeras and other sequence anomalies is still
23 underestimated by the research community. Despite recent papers highlighting the problem,
24 some very obvious anomalies continue to be submitted to sequence repositories. Until the extent
25 of this problem is known, the impetus to improve screening procedures prior to submission and to
26 better curate those that have been submitted, is unlikely to come.

1 The aim of the current study was two fold. (i) To develop a 16S rRNA sequence anomaly
 2 detecting method currently used in our laboratory into a new software tool that is sufficiently
 3 user-friendly and reliable to be used easily by as many researchers as possible. (ii) To use this
 4 tool to estimate the true level of sequence corruption within public repositories. To this end we
 5 present our software to the wider community and detail the results from a survey of selected
 6 bacterial taxa as defined by the RDP database.

7

8 **Materials and Methods**

9 **Developing detection method**

10 All software was written in the Java computer language, using Sun's Java software
 11 development kit, J2SE SDK 1.4.2 (Java Technology [<http://java.sun.com/>]). The final program,
 12 called Pintail, was tested on RedHat 9.0 Linux, Microsoft Windows XP, and Apple Mac OS X
 13 v10.2. Pintail, along with its source code and help files, is freely available from
 14 <http://www.cardiff.ac.uk/biosi/research/biosoft/>, and is released under the terms of the GNU
 15 General Public License (GPL [<http://www.gnu.org/copyleft/gpl.html>]). The program uses
 16 ClustalW (19) to generate sequence alignments.

17 Our method works by aligning a 'query' sequence (S_q) with a trusted 'subject' sequence
 18 (S_s), then analysing differences between query and subject over the entire length of the 16S rRNA
 19 gene, by employing a sliding window of specified size w progressing a fixed number of bases l at
 20 a time along the resulting alignment S_{qs} of length n . The total number of windows will be

21 $m = \left\lceil \frac{n - w + 1}{l} \right\rceil$, where $\lceil \]$ signifies the ceiling of the enclosed expression, i.e. the smallest

22 whole number greater than or equal to the value of the expression. At the i th window w_i ($1 = i =$
 23 m), the percentage of mismatched bases is calculated, giving rise to an observed percentage
 24 difference o_i that can be thought of as an uncorrected measure of evolutionary distance between
 25 query and subject within w_i . The resulting set of observed percentage differences $O_{qs} = \{o_i; o_1,$

1 o_2, \dots, o_m when plotted provide a visual representation of the variation in evolutionary distance
 2 between S_q and S_s over the length of the 16S rRNA gene. The core algorithm for generating O_{qs}
 3 can be summarised as follows.

4 ALGORITHM 1

- 5 (i) Input query sequence S_q , the sequence to be checked for anomalies.
- 6 (ii) Input subject sequence S_s , a reliable sequence closely related to the query.
- 7 (iii) Globally align S_q with S_s using ClustalW to generate alignment S_{qs} of length n .
- 8 (iv) By sliding a window of size w with step l along S_{qs} , determine the percentage of
 9 mismatched bases o_i within window w_i as described above and compute the resulting
 10 dataset $O_{qs} = \{o_i: o_1, o_2, \dots, o_m\}$ of the observed percentage differences detected between S_q
 11 and S_s .
- 12 (v) Plot O_{qs} against base position i to display graphically the changes in evolutionary distance
 13 between S_q and S_s over their mutual length n .

14
 15 Note that the mean of the observed percentage differences $(\sum_i o_i)/m$ is essentially a
 16 measure of the overall uncorrected evolutionary distance between the two sequences. Although
 17 this value will not be exactly the same as that derived by a simple global alignment, for simplicity
 18 we will use the term 'overall evolutionary distance' to refer to this mean, as the distinction
 19 between the two concepts is irrelevant as far as the rest of the paper is concerned.

20 **Expected percentage differences**

21 To assess whether the observed percentage difference plot indicates an anomalous query,
 22 a method was developed for predicting 'expected' percentage differences that one might expect if
 23 both query and subject were reliable. To generate expected percentage differences $E_{qs} = \{e_i: e_1,$
 24 $e_2, \dots, e_m\}$ for any pair of sequences S_q and S_s , it was necessary to map accurately the hyper-
 25 variable regions within the 16S rRNA gene sequence. This was done as follows.

1 All type-strain sequences =1200 nucleotides were downloaded from the RDP web-site (3)
 2 as a single aligned file, with *Escherichia coli* U00096 included as a reference sequence. At the
 3 time of this study RDP release 9, update 22 (September 2004) was current, with 4383 'full-length'
 4 type-strain sequences available for downloading.

5 We totalled the number of each nucleotide residue r $\{r: A, C, G, T/U\}$ at each base
 6 position j ($1 = j = 1542$) within the RDP aligned type-strain sequences, using *E. coli* U00096 as
 7 reference (hence 1542 base positions). From these raw counts we identified the frequency f_j^r of
 8 the most common residue r at each base position j within the alignment (ignoring gap characters).
 9 Note that when position j is most variable, each of the four possible residues is equally likely to
 10 occur. By a simple correction, $p_j = \frac{f_j^r - 0.25}{0.75}$ relative frequencies were converted into
 11 probabilities and so the entire type-strain dataset was described by the probability profile $P = \{p_j:$
 12 $p_1, p_2, \dots, p_{1542}\}$ which reflects the probability of a 16S rRNA sequence being conserved at any
 13 particular residue position.

14 If p_j describes residue conservation at position j , then $q_j = 1 - p_j$ describes residue
 15 variability at that position. In other words, $Q = \{q_j: q_1, q_2, \dots, q_{1564}\}$ is a probability profile that
 16 reflects the variability of a 16S rRNA sequence at any particular residue position. Thus profile Q
 17 can be used to map accurately the hyper-variable regions within the 16S rRNA gene. The
 18 expected percentage differences E_{qs} can be generated from Q by applying the following
 19 algorithm.

20

21 ALGORITHM 2

22 (i) By sliding a window of size w with step l along the probability profile Q , determine the
 23 average probability a_i for each window w_i such that the resulting dataset $Q_{av} = \{a_i: a_1, a_2, \dots,$
 24 $a_m\}$ is a set of average probabilities that can be related directly to the observed percentage
 25 differences dataset O_{qs} generated by ALGORITHM 1.

1 (ii) Define a fitting coefficient a as the overall evolutionary distance between query and subject

2 (as defined by $(\sum_i o_i)/m$) divided by the mean of dataset Q_{av} . Thus, $\mathbf{a} = \frac{(\sum_i o_i)/m}{(\sum_i a_i)/m}$.

3 (iii) Multiply each element of Q_{av} by a to generate the expected percentage differences E_{qs} (i.e., e_i
4 $= a_i \cdot a$).

5 (iv) Plot E_{qs} alongside O_{qs} .

6

7 ALGORITHM 2 generates expected percentage differences for any query and subject pair.

8 By plotting the expected values E_{qs} against their observed values O_{qs} generated by ALGORITHM

9 1, a visual assessment of the quality of sequence S_q with respect to sequence S_s can be made. In

10 addition, subtracting e_i from o_i for each position i generates a series of deviations, the standard

11 deviation of which quantifies the overall deviation of O_{qs} from E_{qs} . This standard deviation we

12 refer to as the Deviation from Expectation (DE) statistic. Thus, $DE = \sqrt{\frac{\sum_i^m (o_i - e_i)^2}{m-1}}$.

13 **Calibrating the method**

14 Of the 4383 type-strain sequences from the RDP, 2361 contained at least one degenerate
15 base. As a means of discarding potentially unreliable records, these degenerate sequences were
16 removed leaving an RDP aligned dataset of 2022 sequences, plus the *E. coli* reference. The type-
17 strains were then analysed by applying the following two procedures.

18 PROCEDURE 1

19 (i) Applying ALGORITHMS 1 and 2, each sequence in the dataset was compared with each other

20 resulting in a DE value for each comparison. (ii) All DE values were plotted against their

21 corresponding overall evolutionary distances. (iii) Obvious outlier DE values were identified

22 from the plot. (iv) Sequences responsible for the outlier DE values were then identified. Since

23 each DE value was generated by a pair of sequences the sequence responsible for the high DE

1 value was identified using a ranking system that scored sequences according to the number of
2 times they were involved the generation of a DE outlier.

3 Identified sequences were then investigated by applying PROCEDURE 2.

4 PROCEDURE 2

5 (i) An NCBI BlastN search (1) was undertaken with each query sequence to identify its nearest
6 neighbours within the public database. (ii) A suitable nearest neighbour was chosen for
7 comparison (labelled First Subject). Sequences originating from different research groups, and
8 hence a different 16S rRNA gene library to that which had generated the query, were preferred.
9 (iii) The First Subject was compared with the query using the Pintail program and the output
10 assessed for evidence of any sequence anomaly. (iv) To confirm the reliability of the First
11 Subject, and hence the conclusion drawn, a second nearest neighbour was selected again from a
12 separate study. This Second Subject was compared with the First Subject using Pintail, and
13 output checked. (v) Finally, as a final check, the query was compared with the Second Subject.

14 It can be seen that, ideally, only three comparisons are necessary per query sequence to
15 unambiguously identify an anomaly. In practice this was not always possible, either because a
16 lack of suitable database entries meant that the only nearest neighbours available were those
17 generated by the same author(s) and thus probably from the same gene library, or because the
18 best available 'nearest neighbour' was only distantly related to the query. Under such
19 circumstances up to nine nearest neighbours were compared with the query sequence and each
20 other, and the final conclusion was made after assessing the overall trend in the resulting matrix
21 of pairwise comparisons. Where necessary, the NCBI's BLAST 2 SEQUENCES program
22 (bl2seq, 18) was used to resolve uncertainties.

23 PROCEDURES 1 and 2 were applied to the type-strain data and outlier DE values found
24 to be generated by anomalous sequences were excluded from subsequent analysis. The median,
25 upper quartile, 95, 99, 99.9 and 100% quantiles of the corrected DE plot were then determined
26 for each 1% interval along the x-axis of the plot. In this way, the corrected DE plot could be

1 described in terms of a series of quantile plots and included within the final Pintail program.
2 Thus, a DE value subsequently generated by Pintail could be compared with DE values
3 previously generated from the type-strain comparisons, and conclusions drawn as to the
4 likelihood of the new DE value being generated by a pair of non-anomalous sequences.

5 **Testing Pintail with known chimeras**

6 The Pintail program was tested with fifty known bacterial chimeric sequences originally
7 identified by Hugenholtz and Huber (8) and listed in the RDP database release 9, update 22. A
8 further five archaeal sequences listed by Hugenholtz and Huber (8) but not included on the RDP
9 website were also tested. Each chimera was analysed by following PROCEDURE 2.

10 **Screening selected bacterial phyla**

11 Using the RDP's online hierarchy browser, all bacterial phyla containing up to 200
12 sequence records were downloaded as separate aligned files. For each aligned dataset
13 PROCEDURE 1 was applied to identify putatively anomalous sequences. In this screening
14 outlier DE values were defined as those falling above the 99.9% quantile line calculated from the
15 type-strain data. Anomalous sequences identified in this way were checked by PROCEDURE 2.

16

17 **Results**

18 **Implementation of methodology**

19 The development of the methodology described in this paper culminated in the computer
20 program Pintail, the operation of which is now described. Fig. 1 shows a screenshot of Pintail,
21 showing the outcome of a typical analysis. The query sequence S_q (in this instance a chimera)
22 was entered into the top left text-box and the subject sequence S_s (a reliable sequence, identified
23 by BlastN as closely related to the query) was entered into the bottom left text-box. The results
24 of the analysis are displayed in the panel on the right and show graphically that the query is
25 indeed a chimera with its 5' end phylogenetically more distant from the subject sequence, than its

1 3' end. Fig. 2 illustrates in more detail typical graphs generated by the program, with panels A to
2 C showing the output from a reliable query sequence being compared with equally reliable
3 subject sequences of varying evolutionary distances. Conversely, panels D to F show typical
4 plots obtained when the query sequence is chimeric.

5 Each graph generated by the program consists of four plots. The plot of observed
6 percentage differences (O_{qs} ; black line in the Fig. 2 panels) shows the change in percentage
7 difference between query and subject as the sampling window moves along the alignment. In all
8 examples shown in Fig. 2 a window size w of 300 nucleotides was used, moving along the
9 alignment $l = 25$ bases at a time. This combination was found to be most suitable for displaying
10 overall trends. Reducing window size to $w = 100$ bases supplies more detail and is useful for
11 estimating chimeric breakpoints.

12 The mean of the observed percentage differences displayed by the program is roughly
13 equivalent to the uncorrected evolutionary distance between query and subject. From this mean
14 the expected percentage differences (E_{qs}) which might be expected for sequences of this
15 evolutionary distance are calculated. These expected percentage differences are displayed as a
16 second plot line within the program's output graph (Fig. 1) and as grey lines in Fig. 2. Similarly,
17 two further expected lines are plotted based on the mean observed percentage differences $\pm 5\%$,
18 and represents graphically this level of variation around the expected line as an area shaded light
19 grey (Fig. 2).

20 The expected line (E_{qs} plot) helps to indicate if and where the observed line deviates from
21 what might be expected from reliable sequences with the same overall evolutionary distance as
22 the query and subject. The Deviation from Expectation (DE) statistic calculated by the program,
23 quantifies this deviation. The higher the DE value, the greater will be the departure of the
24 observed data from that expected of trusted sequences. To aid interpretation, the DE statistic is
25 best viewed in the context of reliable query-versus-subject comparisons sharing similar
26 evolutionary distances. So the program summarises the DE values obtained between type-strains

1 of the same evolutionary distance as exhibited between query and subject, and from this
2 information the probability that the observed DE value is likely to have been generated by two
3 reliable sequences is inferred (Fig. 1).

4 **Development of methodology and testing the underlying assumption**

5 The assumption underlying the method implemented in Pintail is that two reliable (i.e.,
6 non-anomalous) 16S rRNA sequences of known overall evolutionary distance will vary by
7 roughly the same amount over the length of the gene, allowing for the effects of the hyper-
8 variable regions, when homologous bases are compared. Given the empirical nature of the
9 methodology it was necessary to test this assumption.

10 One test was to select pairs of reliable sequences at random, apply the method, and assess
11 the output for any contradiction of our assumption. Fig. 2A-C illustrates typical results obtained
12 this way. However, this approach was inevitably limited in scope. To test the assumption more
13 thoroughly and at the same time calibrate our method we needed to consider a much larger
14 dataset of reliable sequences. To do this necessitated finding a way of quantifying our
15 observations so that a more automated checking procedure could be employed. This led to the
16 concept of 'expected percentage differences', and the 'Deviation from Expectation' statistic,
17 described in Materials and Methods, and now considered in more detail below.

18 ***Expected Percentage Differences.*** To generate expected percentage differences for any
19 two sequences, it is necessary to take account of (i) the regions of conservation and variability
20 inherent in the 16S rRNA gene, and (ii) the evolutionary distance represented by sequence
21 dissimilarity between the two sequences. As Fig. 2 A-C illustrate, the character of the observed
22 percentage difference plot is informed by both of these concepts. Therefore we needed to model
23 16S rRNA intra-gene variability and then use this model to predict expected percentage
24 differences from overall evolutionary distance (as represented by the mean of the observed
25 percentage differences).

26 Type-strain sequences, *a priori*, can be considered reliable, in that they will normally

1 have been generated from pure cultures and therefore will have been less prone to the errors
2 common to environmental samples, due to quality and purity of the template. RDP release 9,
3 update 22, contains 4383 type strain sequences with a length =1200 nucleotides. We downloaded
4 all 4383 records from the RDP website retaining the RDP's alignment, along with a reliable
5 *Escherichia coli* record (U00096) as reference sequence. From this we were able to allocate to
6 each base position in the *E. coli* reference sequence a frequency for the most common nucleotide
7 residue (A, C, G or T/U; Fig. 3A). For example, a position that is occupied by an adenine in all
8 type-strain sequences would have a frequency of 1. Conversely, a position where all four bases
9 are equiprobable would have a frequency of 0.25.

10 Smoothing these data revealed peaks and troughs which corresponded to the known
11 hyper-variable and conserved regions for the 16S rRNA gene (Fig. 3B), matching peaks and
12 troughs in observed percentage difference plots. Converting these frequencies to a probability
13 profile – allocating a probability to each 16S rRNA base position – created a profile of the 16S
14 rRNA intra-gene variability, for use in the final program. Expected percentage differences for
15 any two sequences were generated from this profile by multiplying each probability by the fitting
16 coefficient a so as to ensure the resulting dataset had the same mean as the observed data.

17 ***Deviation from Expectation (DE) statistic.*** Subtracting a set of expected values from
18 corresponding observed data points generated a set of 'error' values, the standard deviation of
19 which summarised the extent to which observation deviated from expectation. This is how the
20 DE statistic was derived and used in this study as a way of summarising any analysis of sequence
21 pairs as a single value.

22 We were now in a position to automate our method and consider a much larger dataset of
23 reliable sequences. The 4383 type-strain sequences initially served as this dataset; however,
24 since our method detects any sequence anomaly, it quickly became apparent that high levels of
25 type-strain degeneracy were hampering our survey and needed to be discounted. Only 2022 out
26 of 4383 type strain sequences were completely without degenerate base characters. Of the

1 remaining 2361, levels of degeneracy as high as 483 bases were detected, although 2173 had = 50
2 degenerate characters. Further analysis concentrated on the 2022 degeneracy-free sequences,
3 since these were considered to be least likely to have anomalies.

4 **Calibration.** Pairwise comparisons of the 2022 sequences without degeneracies generated
5 2,043,231 DE values. Plotting all these against the mean of the observed percentage differences
6 for each comparison (Fig. 4) revealed that most DE values, and hence most comparisons,
7 clustered together. However, a number of outlier clusters quite distinct from the main cluster
8 were also observed (Fig. 4A) and investigation showed the same 15 sequences responsible for
9 these outliers (Table 1).

10 Application of PROCEDURE 2 (Fig. 5) showed two of these 15 sequences to be chimeric.
11 Record AJ272391 (classified as *Lactobacillus psittaci*) is a two-fragment chimera with 5'-end
12 practically identical to *Lactobacillus jensenii* (AF243159) and 3'-end similarly close to
13 *Lactobacillus vaginalis* (AF243154). Record U10877 (classified *Riemerella anatipestifer* ATCC
14 11845) is a three-fragment chimera with fragments 1 and 3 deriving from a member of the
15 *Bacteroidetes* and fragment 2 of *Gammaproteobacteria* origin (Fig. 2E). The remaining thirteen
16 sequences contained anomalies most likely to be sequencing errors. Eight originated from the
17 same research group and all contained some sort of sequencing error in the first 220 to 240 bases
18 at the 5'-end. Intriguingly, two of these anomalies were observed when the original 2022 type-
19 strain RDP-alignment was used but not when checked with ClustalW. Further investigation by
20 eye confirmed these anomalies to be real confirming the RDP alignment to be the more accurate
21 than the ClustalW alignment.

22 When the 15 anomalous sequences were removed from the dataset, the plotted DE values
23 clustered together as one group (Fig. 4B). Fig. 4C shows the same data reduced to a series of
24 quantile plots, which were used to estimate the probability of the query sequence being
25 anomalous, as indicated in Fig. 1.

26 **Testing program with known chimeras**

1 We tested our approach with 39 chimeric 16S rRNA sequences identified by Hugenholtz
2 and Huber (8) and applied PROCEDURE 2 as summarised in Fig. 5. All were confirmed as
3 chimeric by our method. In addition, we found that Hugenholtz and Huber had incorrectly
4 characterised record AF254401 as a two-fragment chimera, whereas our method reveals it to be a
5 three-fragment (Fig. 6). AF254401 sequence up to *E. coli* position 340 is of *Firmicutes* origin
6 (closely matching AF323775). Bases from 341 to 1080 come from an unknown source, the
7 closest match being AF323760, previously identified as from the OP9 phylum (8) but remaining
8 unclassified by the RDP. The remainder of AF254401 derives from the *Spirochaetes* phylum and
9 closely matches M88719.

10 We also tested an additional 15 chimeras identified by Hugenholtz and Huber and listed
11 within the RDP hierarchy browser but not included in their paper (8). We confirmed twelve to be
12 chimeric. However, we could not find evidence that X84498, AF333535, or AY082475 were
13 chimeric (although with AY082475 there is evidence of a possible sequencing anomaly at the
14 extreme 5'-end), and a series of comparisons using bl2seq (18) under a range of parameter
15 settings failed to contradict this analysis.

16 **Database analysis**

17 The RDP website hierarchy browser (3) classifies 16S rRNA sequence records according
18 to the current Bergey's 16S rRNA-based classification system (5). We used this facility to obtain
19 aligned sequence files for 19 phyla amounting to 1399 records in all. Phyla were selected purely
20 by size, with any phylum containing ≥ 200 sequences chosen. Thus, all were selected without
21 prior knowledge of any sequence anomalies.

22 Initial screening by DE value, as detailed in PROCEDURE 1 identified 73 putatively
23 anomalous sequences. Application of PROCEDURE 2 showed 70 out of these 73 to be
24 unambiguously anomalous and distributed within 16 of the 19 phyla (Fig. 7, Table 2). The three
25 false positives all occurred within the *Aquificae* and were caused by the absence of sufficiently
26 closely related subject sequences for comparison with the query sequences concerned.

1 Of the 70 confirmed anomalies, 45 were clearly chimeric. A further 15 were highly
2 degenerate. The remaining 10 contained other sequence anomalies, such as that found within the
3 *Aquificae* record AY268103, the 5'-end of which, up to *E. coli* position 560, was the reverse-
4 complement of 16S rRNA.

5 Pintail identified 22 of the 45 chimeras as derived from parents belonging to different
6 phyla. For example, sequence AF523990 is part *Acidobacteria*, part *Actinobacteria*. A further
7 16 chimeras contained one parent of either unknown (no close record in current database) or
8 unclassified (RDP unable to classify according to Bergey's classification) origin. Thirteen out of
9 45 were formed from parents belonging to the same phylum.

10 Whilst most chimeras were composed of two fragments from unrelated source sequences,
11 9 three-fragment chimeras were also detected. A striking example of this is the '*Fusobacteria*'
12 sequence AJ289180 with its 5'-end originating from a *Fusobacterium*, the middle region being of
13 *Spirochaete* origin, and the 3'-end belonging to a member of the *Bacteroidetes*.

14 Table 2 lists a further 10 anomalous sequences discovered during our investigations but
15 not included in our original 19-phylum dataset. All but two are obvious chimeras. One is
16 another example of the 5'-end being a reverse-complement of the correct sequence. Three of
17 these records were submitted to the public repositories during our study.

18 **Chimera breakpoints**

19 Approximate breakpoints for chimeras in this study were determined by analysing the
20 plots produced by Pintail. Reducing window size to 50-100 was most effective in providing
21 sufficient visual detail in order to make this assessment. Breakpoints were most easily assessed
22 when both 'parent' sequences were identified (e.g., Fig. 5) since their corresponding observed
23 percentage differences plots could easily be superimposed on one another and breakpoints
24 identified where the lines crossed.

25 Identified breakpoint positions were combined with values identified by Hugenholtz and
26 Huber (8) and plotted alongside the known hyper-variable regions within the 16S rRNA gene

1 (Fig. 3C). Most were found to fall between hyper-variable regions. Given that variability of each
2 16S rRNA base position can be described in terms of the frequency of the most common residue
3 at that position (Fig. 3A), the overall median and 95% confidence interval 'notches' of these
4 frequencies is 0.931 ± 0.013 . In contrast, the median of those frequencies corresponding to
5 breakpoint positions was significantly higher at 0.975 ± 0.015 .

6

7 **Discussion**

8 It has long been recognised that corrupt sequences are present within the public
9 repositories. What has not been known is how many there may be. In the current study 5% of
10 records were found to be corrupt and most of these (78.6%) were chimeras or similarly insidious
11 sequencing errors. Eleven of the 19 phyla investigated contained obvious chimeras with chimeric
12 content ranging from 0.8 to 11.8% of the total. Six phyla contained sequence anomalies
13 presumably generated during sequencing. Five phyla contained records with highly degenerate
14 sequences. In total, sixteen out of the nineteen phyla considered contained some sort of
15 substantial sequence anomaly. Extrapolating our results to the public database as a whole this
16 would suggest, at a conservative estimate, one in twenty sequences have substantial errors. We
17 believe these figures underestimate the true number of anomalous records given that we
18 concentrated our efforts on uncovering the more obvious sequence anomalies.

19 This study confirms that anomalous sequences continue to be added to the public
20 databases; of the chimeras identified in this study, 27.7% were submitted to the NCBI during
21 2004 alone (Fig. 8) and 91.5 % of these were submitted in the last five years. These figures
22 reflect recent interest in many of the phyla considered in this study and the steady year-on-year
23 increase in sequence submissions generally. They also highlight the ongoing nature of the
24 problem. Indeed, we noted five chimeric additions to the RDP database whilst our study
25 progressed (two added to *Nitrospira* , one to *Verrucomicrobia*, two to the *Betaproteobacteria*, a

1 taxon not otherwise investigated in this study).

2 It is fair to say that many researchers have been insufficiently cognisant of the problem of
3 sequence anomalies within the public databases. This situation is changing, however, as
4 evidenced by the renewed burst of activity in generating software tools for recognising chimeras.
5 Within the last year or so three new tools have been introduced (6, 7, 10), presumably driven by
6 these authors' desire, like us, to screen sequences generated through their own researches.
7 Certainly, our experiences with chimeric sequences within 16S rRNA clone libraries led us to
8 develop Pintail.

9 It is important that the extent of sequence anomalies within public repositories is fully
10 realised. The research community's phylogenetic view of the bacterial world is increasingly
11 informed by 16S rRNA information (3, 5, 15). At least half of the 53 phyla named in 2003 are
12 currently known only from 16S rRNA gene sequences amplified from the environment by PCR
13 (15) and this number is growing (4). It is notable that, of the six proposed new taxa analysed in
14 this study, four harboured chimeras, some of which were extreme. For example, a third of the
15 'OP11' sequence AY693838 derives from a *Betaproteobacterium*. Another 'OP11' sequence,
16 AY218572, is almost half an *Episilonproteobacterium*. The 5'-end of 'WS3' bacterium
17 AY592328 is *Actinobacteria* in origin.

18 In all, 48.9% of identified chimeras were derived from bacteria belonging to different
19 phyla (a particularly striking example being AJ289180 – a jumble of *Fusobacteria*, *Spirochaetes*
20 and *Bacteroidetes*). This figure is undoubtedly an underestimate as, for further 35.6 %, we either
21 could not identify the source (no suitable subject record in the database) or the source was as-yet
22 unclassified. Some of these chimeras were so extreme it is surprising that they have not been
23 detected before. We find this worrying, as our concern is that there are far more subtle chimeras
24 in the database, constructed from close phylogenetic neighbours, that have less chance of being
25 spotted and could give rise to all sorts of spurious intra-taxon clustering errors.

26 Our study also shows that a significant proportion of chimeras were generated from three

1 fragments, often from three separate sources (consider AJ289180, above). Chimeras with more
2 than three fragments may also be possible since the positions of chimeric breakpoints in
3 conserved regions suggested that there are several areas within the 16S rRNA gene where
4 splicing may occur (Fig. 3C).

5 The methodology presented here depends on the type-strain 16S rRNA database used.
6 Clearly, current type-strain sequences are not representative of all *Bacteria*; our RDP-derived
7 type-strain database reflects past cultivation successes and there is a definite slant towards
8 *Bacteria* of medical interest. Furthermore as this study shows, the quality of some type-strain
9 sequences is not good. Nevertheless our method was effective over a wide phylogenetic range,
10 and could even be applied to *Archaea* sequences, as analysis of those archaeal chimeras listed in
11 Hugenholtz and Huber's paper (8) proved. Since we used sequence alignments from the RDP
12 database that currently only lists members of the *Bacteria*, our model and calibration data were
13 constructed from members of this domain only. However, there is no theoretical reason why a
14 more comprehensive model incorporating *Archaea* sequences could not be created, or indeed
15 generate models for specific domains, phyla or other taxa to improve sensitivity.

16 DE values generated from type-strain data, once anomalous sequences were removed,
17 proved useful in calibrating our method; that is placing observed DE values in the context of
18 sequences identified as reliable. This raises the possibility of screening database records on a
19 much larger scale than that tackled in this study.

20 How should the research community tackle the problem of monitoring anomalous
21 sequences in databases? Curators have a role to play. For example we found three chimeras
22 within the NCBI, labelled as such, yet not similarly flagged within the RDP database (though an
23 understandable omission given the RDP's automated nature). But the practicalities of current
24 database management are such that the curators' contribution must be limited. Primary
25 responsibility must, indeed should, lie with researchers submitting sequences. To this end
26 software tools must be available and used by researchers to assist in screening PCR generated

1 sequences for anomalies before database deposition. Any tool produced for this purpose must be
2 easily accessible to encourage as widespread use as possible. For a tool to be accessible it should
3 be easy to use, easy to understand and interpret, transparent in how it comes to its conclusions,
4 freely available, and capable of running on whatever computer platform a user might have.
5 Unless chimeras and other anomalous sequences can be eliminated from public databases
6 microbial ecologists will have an erroneous picture of natural prokaryotic biodiversity.

7

8 **Acknowledgments**

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

11

12 **References**

- 13 1. **Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman.**
14 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search
15 programs. *Nucleic Acids Research* **25**:3389-3402.
- 16 2. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L.**
17 **Wheeler.** 2000. GenBank. *Nucleic Acids Research* **28**:15-18.
- 18 3. **Cole, J., B. Chai, T. Marsh, R. Farris, Q. Wang, S. Kulum, S. Chandra, D. McGarrell,**
19 **T. Schmidt, G. Garrity, and J. Tiedje.** 2003. The Ribosomal Database Project (RDP-II):
20 previewing a new autoaligner that allows regular updates and the new prokaryotic
21 taxonomy. *Nucleic Acids Research* **31**:442-443.
- 22 4. **Fox, J. L.** 2005. Ribosomal gene milestone met, already left in dust. *ASM News* **71**:6-7.
- 23 5. **Garrity, G. M., M. Winters, A. W. Kuo, and D. Searles.** 2002. Taxonomic outline of the
24 prokaryotes, *Bergey's Manual of Systematic Bacteriology*, Second Edition. Springer-Verlag,
25 New York.

- 1 6. **Gonzalez, J. M., J. Zimmerman, and C. Saiz-Jimenez.** 2005. Evaluating putative chimeric
2 sequences from PCR-amplified products. *Bioinformatics* **21**:333-337.
- 3 7. **Huber, T., G. Faulkner, and P. Hugenholtz.** 2004. Bellerophon: a program to detect
4 chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**:2317-2319.
- 5 8. **Hugenholtz, P., and T. Huber.** 2003. Chimeric 16S rDNA sequences of diverse origin are
6 accumulating in the public databases. *International Journal of Systematic and Evolutionary*
7 *Microbiology* **53**:289-293.
- 8 9. **Kanz, C., P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. v. d.**
9 **Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F.**
10 **G. Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M.**
11 **McHale, F. Nardone, V. Silventoinen, S. Sobhany, P. Stoehr, M. A. Tuli, K. Tzouvara,**
12 **R. Vaughan, D. Wu, W. Zhu, and R. Apweiler.** 2005. The EMBL Nucleotide Sequence
13 Database. *Nucleic Acids Research* **33**:D29-D33.
- 14 10. **Klepac-Ceraj, V., M. Bahr, B. C. Crump, A. P. Teske, J. E. Hobbie, and M. F. Polz.**
15 2004. High overall diversity and dominance of microdiverse relationships in salt marsh
16 sulphate-reducing bacteria. *Environmental Microbiology* **6**:686-698.
- 17 11. **Komatsoulis, G. A., and M. S. Waterman.** 1997. A new computational method for
18 detection of chimeric 16S rRNA artifacts generated by PCR amplification from mixed
19 bacterial populations. *Applied and Environmental Microbiology* **63**:2338-2346.
- 20 12. **Kopczynski, E. D., M. M. Bateson, and D. M. Ward.** 1994. Recognition of chimeric
21 small-subunit ribosomal DNAs composed of genes from uncultured microorganisms.
22 *Applied and Environmental Microbiology* **60**:746-748.
- 23 13. **Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker Jr, P. R. Saxman, R. J. Farris, G.**
24 **M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje.** 2001. The RDP-II (Ribosomal
25 Database Project). *Nucleic Acids Research* **29**:173-174.
- 26 14. **Paabo, S., D. M. Irwin, and A. C. Wilson.** 1990. DNA damage promotes jumping between

- 1 templates during enzymatic amplification. *Journal of Biological Chemistry* **265**:4718-4721.
- 2 15. **Rappe, M. S., and S. J. Giovannoni.** 2003. The uncultured microbial majority. *Annual*
3 *Review of Microbiology* **57**:369-394.
- 4 16. **Robison-Cox, J. F., M. M. Bateson, and D. M. Ward.** 1995. Evaluation of nearest-
5 neighbor methods for detection of chimeric small- subunit rRNA sequences. *Applied and*
6 *Environmental Microbiology* **61**:1240-1245.
- 7 17. **Shuldiner, A., A. Nirula, and J. Roth.** 1989. Hybrid DNA artifact from PCR of closely
8 related target sequences. *Nucleic Acids Research* **17**:4409.
- 9 18. **Tatusova, T. A., and T. L. Madden.** 1999. Blast 2 sequences - a new tool for comparing
10 protein and nucleotide sequences. *FEMS Microbiology Letters* **174**:247-250.
- 11 19. **Thompson, J., D. Higgins, and T. Gibson.** 1994. Clustal W: improving the sensitivity of
12 progressive multiple sequence alignment through sequence weighting, positions-specific gap
13 penalties and weight matrix choice. *Nucleic Acids Research* **22**:4673-4680.
- 14 20. **Wang, G. C.-Y., and Y. Wang.** 1996. The frequency of chimeric molecules as a
15 consequence of PCR co-amplification of 16S rRNA genes from different bacterial species.
16 *Microbiology* **142**:1107-1114.
- 17 21. **Wang, G. C.-Y., and Y. Wang.** 1997. Frequency of formation of chimeric molecules as a
18 consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes.
19 *Applied and Environmental Microbiology* **63**:4645-4650.
- 20

1 **Figure legends**

2 **Fig. 1.** Program screenshot illustrating a typical analysis. In this example, query AY693838 (top
3 left) is compared with subject AJ551147 (bottom left) generating a plot of evolutionary distances
4 that demonstrate high similarity between these two sequences at the 5' end only. AY693838,
5 introduced into the NCBI on 30th August 2004, is classified by the RDP as belonging to the
6 proposed new OP11 phylum. AJ551147, in contrast, belongs to the *Betaproteobacterium* genus,
7 *Janthinobacterium*.

8
9 **Fig. 2.** Typical 16S rRNA gene sequence comparison plots generated by Pintail (all graphs
10 generated with window size 300 and step size 25). Panels A-C show plots between pairs of
11 trusted sequences of increasing evolutionary distance, whilst D to F show examples where the
12 query sequence is a chimera. Observed percentage differences between sequences are plotted as
13 black lines. Gray lines show the expected percentage differences for the sequence pairs. Light
14 grey shading indicates expected percentage differences plus or minus 5%. *Escherichia coli*
15 ATCC 11775T (X80725) is compared with *Escherichia vulneris* ATCC 33821T (X80734) in
16 panel A, with *Pseudomonas aeruginosa* LMG 1242T (Z76651) in panel B, and with *Aquifex*
17 *pyrophilus* (T) Kol5a (M83548) in panel C. Panels D to F show three typical chimeric patterns.
18 In panel D the three-fragment *Nitrospira* chimeric sequence AY373422 (estimated breakpoints;
19 340, 740) is compared with its Blast identified 'nearest neighbour' X82559. In panel E, the three
20 fragment chimeric record U10877 generated from *Riemerella anatipestifer* (T) ATCC 11845 is
21 shown to diverge from the sequence of its nearest neighbour *R. anatipestifer* strain 115/02
22 (AY856450) around *E. coli* positions 790 to 1130. In panel F, the two-fragment *Fusobacteria*
23 chimeric sequence AY548989 (estimated breakpoint, 800), is compared to the sequence from its
24 nearest neighbour, AY548984.

25

1 **Fig. 3.** Illustrating variable regions within the 16S rRNA gene and location of chimeric
2 breakpoints. Panel A displays the frequency of occurrence of the most common nucleotide
3 residue at each base position within the 16S rRNA gene, as determined from RDP listed 4383
4 type strains, with *E. coli* U00096 as reference. These frequencies are measures of variability
5 within the gene. Smoothing the data, by taking the mean frequency within a window of 50 bases,
6 moving one base at a time along the gene, creates the plot in panel B. In B, the locations of the
7 hyper-variable regions are labelled, with grey bars on the x-axis defining these regions as V1-V9
8 (The Comparative RNA Web Site [<http://www.rna.icmb.utexas.edu/>]). Panel C is a histogram of
9 all chimera breakpoints identified in this study and that of Hugenholtz and Huber (8).

10

11 **Fig. 4.** Deviation from Expectation (DE) values generated from type-strain dataset containing
12 2022 16S rRNA gene sequences without any degenerate base positions (see text). DE value was
13 generated for each of the 2,043,231 pairwise sequence comparisons and plotted against
14 evolutionary distance between sequences. Panel A illustrates the dataset prior to the removal of
15 the 15 anomalous sequences (see text) and panel B shows the plot after removal. Panel C shows
16 the quantile values used to describe this data, and incorporated into the Pintail program as a
17 means of calibration.

18

19 **Fig. 5.** Illustrating PROCEDURE 2 for unambiguously confirming a chimeric sequence (all
20 graphs generated with window size 300 and step size 25). In this example the query, an
21 *Acidobacteria* (AF523990), is compared with its nearest neighbour (AF523976) identified by
22 BlastN search, and an anomaly at the 5'-end is identified (Panel A). AF523976 is next compared
23 with its nearest neighbour, AY234512, to confirm that it is reliable (Panel B). No anomaly is
24 detected. As a final check, AF523990 is compared with AY234512 and, as expected, the 5'-end
25 anomalous feature is seen (Panel C). To determine whether this anomaly is chimeric, the
26 identified 5' region is excised, a Blast search undertaken, and the identified nearest neighbour (in

1 this case *Actinobacteria*, X68459) is compared with AF523990 (Panel D). Again an anomaly is
2 detected, but this time the reverse of that seen in panel A, clearly indicating our query to be a
3 chimera. Comparing X68459 with its neighbour, AF498683, confirms its reliability (panel E),
4 and as expected, comparing the original query with AF498683 generates the same profile as seen
5 in panel D. Chimeric breakpoint can be estimated by superimposing A on D.

6

7 **Fig. 6.** Analysis of the three fragment chimera AF254401 (all graphs generated with window size
8 100 and step size 25). The query is shown compared with AF323775 (panel A), AF323760
9 (panel B) and M88719 (panel C).

10

11 **Fig. 7.** Distribution of sequence anomalies with the nineteen *Bacteria* phyla, as defined by the
12 Ribosome Database project (RDP-II; 3). Numbers in brackets after the phylum (or candidate
13 division) name are the total number of sequences within that phylum present in RDP release 9,
14 update 22, September 2004.

15

16 **Fig. 8.** First appearance in the NCBI database of the anomalous records identified by this study.

1 TABLE 1. Anomalous *Bacteria* 16S rRNA gene sequence records from type-strains.

Accession	Name	Location of anomaly relative to <i>E. coli</i>	Description
D17751	<i>Leucobacter komagatae</i> IFO15245T	60 to 220	Anomaly near 5' end - likely sequencing error.
D21342	<i>Microbacterium imperiale</i> IFO 12610T	230	Anomaly at 5' end - likely sequencing error.
D21344	<i>Microbacterium laevaniformans</i> IFO 14471T	90 to 220	Anomaly near 5' end - likely sequencing error.
AJ242532	<i>Arthrobacter flavus</i> CMS-19Y	1130 to 1420	Anomaly near 3' end - likely sequencing error.
AJ233946	<i>Nannocystis exedens</i> Na e1	730 to 840	Anomaly near middle - likely sequencing error.
D21245	<i>Luteococcus japonicus</i> IFO12422	240, 680 to 790	Anomaly at 5' end and in middle - likely sequencing errors.
AF195797	<i>Thermoanaerobacter subterraneus</i> SEBR 7858; LA61	800 to 960	Anomaly near middle - likely sequencing error.
D21343	<i>Microbacterium lacticum</i> IFO 14135T	70 to 240	Anomaly near 5' end - likely sequencing error.
Z49116	<i>Halanaerobium saccharolyticum</i> subsp. <i>senegalense</i> DSM 7379	1320 to 1450	Anomaly near 3' end - likely sequencing error.
D21339	<i>Microbacterium arborescens</i> IFO 3750T	230	Practically identical to D21342.
D21341	<i>Microbacterium dextranolyticum</i> IFO 14592T	60 to 240	Anomaly near 5' end - likely sequencing error (only visible with RDP alignment).
AB013297	<i>Vibrio rumoiensis</i> S-1	500?	Anomaly near 5' end - likely sequencing error (only visible with RDP alignment).
D17527	<i>Kineococcus aurantiacus</i> IFO 15268	70 to 240	Anomaly near 5' end - likely sequencing error.
AJ272391	<i>Lactobacillus psittaci</i>	790	Two fragment chimera with 5' end practically identical to <i>Lactobacillus jensenii</i> (AF243159) and 3' end practically identical to <i>Lactobacillus vaginalis</i> (AF243154).
U10877	<i>Riemerella anatipestifer</i> ATCC 11845	790, 1130	Three fragment chimera with middle fragment of <i>Gammaproteobacteria</i> origin. Fragments one and three derive from the same <i>Bacteroidetes</i> origin.

1 TABLE 2. Anomalous sequences identified by this study.

Accession	Phylum	Approx. break position relative to <i>E. coli</i>	Details
AY268103	<i>Aquificae</i>	560	PCR or sequencing error with first ~465 bases the reverse complement of what they should be.
AB183857	<i>Aquificae</i>	425	Anomaly at 5' end, though origin unknown - either chimera or sequencing error.
AF018191	<i>Aquificae</i>	1080	Two fragment chimera, with both fragments <i>Aquificae</i> in origin.
AJ237665	<i>Thermotogae</i>	930	Two fragment chimera, with 3' end <i>Firmicutes</i> in origin.
L10662	<i>Thermodesulfobacteria</i>	-	Degenerate sequence - several large blocks of N bases.
Z15060	<i>Deinococcus-Thermus</i>	-	Degenerate sequence - one large block of N bases.
X58340	<i>Deinococcus-Thermus</i>	-	Degenerate sequence - several large blocks of 'N' bases.
AF317775	<i>Nitrospira</i>	-	Degenerate sequence - one large block of N bases.
AF317779	<i>Nitrospira</i>	-	Degenerate sequence - one large block of N bases.
L14619	<i>Nitrospira</i>	-	Degenerate sequence - several large blocks of N bases.
AY661410	<i>Nitrospira</i>	320, 540	Two, possibly three, fragment chimera, with 3' end of unknown origin.
AF543500	<i>Nitrospira</i>	250	Sequencing anomaly only visible when RDP alignment used.
AY373422	<i>Nitrospira</i>	340, 740	Three fragment chimera, with 5' end <i>Gammaproteobacteria</i> , middle <i>Alphaproteobacteria</i> , and 3' end unknown in origin.
AY661421	<i>Nitrospira</i>	370	Two fragment chimera, with 3' end unclassified (candidate division OP5 according to NCBI).
AF485343	<i>Nitrospira</i>	1080	Two fragment chimera, with 3' end unclassified. Record now replaced in database.
AY297986	<i>Nitrospira</i>	700	Two fragment chimera, with 5' end <i>Firmicutes</i> in origin. Record already marked as chimeric in database.
AY796049	<i>Nitrospira</i> *	790	Two fragment chimera with 5' end <i>Betaproteobacteria</i> in origin.
AY762631	<i>Nitrospira</i> *	660, 940	Three fragment chimera derived from two parents, with middle fragment of unclassified origin.

X86774	<i>Nitrospira</i>	790, 1220	Three fragment chimera derived from two parents, with middle fragment <i>Gammaproteobacteria</i> in origin. Already identified as chimera.
AF543509	<i>Nitrospira</i>	500, 790	Three fragment chimera, with middle fragment also <i>Nitrospira</i> in origin.
AB176700	<i>Nitrospira</i>	500	Two fragment chimera, with 5' end of unknown origin.
AF543503	<i>Nitrospira</i>	540	Two fragment chimera, with both fragments <i>Nitrospira</i> in origin.
AF543511	<i>Nitrospira</i>	760	Two fragment chimera, with both fragments <i>Nitrospira</i> in origin.
L22045	<i>Nitrospira</i>	-	Degenerate sequence - one large block of N bases.
M79383	<i>Nitrospira</i>	-	Degenerate sequence - one large block of N bases.
Y10652	<i>Chlorobi</i>	-	Degenerate with Ns clustered at 5' and 3' end giving superficial appearance of a chimera.
Y10643	<i>Chlorobi</i>	-	Degenerate with Ns clustered at 5' and 3' end giving superficial appearance of a chimera.
Y10651	<i>Chlorobi</i>	-	Degenerate with Ns clustered at 5' and 3' end giving superficial appearance of a chimera.
Y10647	<i>Chlorobi</i>	-	Degenerate sequence - one large block of N bases, and numerous other Ns.
Y10640	<i>Chlorobi</i>	-	Degenerate with Ns clustered at 5' and 3' end giving superficial appearance of a chimera.
AY661796	<i>Chlamydiae</i>	-	Sequencing anomaly only visible when RDP alignment used.
AY661795	<i>Chlamydiae</i>	-	Sequencing anomaly only visible when RDP alignment used.
AB179510	<i>Acidobacteria</i>	940-1100	Two fragment chimera, with 3' end of unclassified origin. Lack of clear break due to number of degenerate bases.
AY326570	<i>Acidobacteria</i>	600-1000	Two fragment chimera, with 3' end of unclassified origin. No obvious reason for lack of clear break.
AF523990	<i>Acidobacteria</i>	370	Two fragment chimera, with 5' end of <i>Actinobacteria</i> origin.
AJ536862	<i>Acidobacteria</i>	280	Two fragment chimera, with 5' end of unclassified origin.
Y07575	<i>Acidobacteria</i>	560	Two fragment chimera, with 3' end of unknown origin.
AY548989	<i>Fusobacteria</i>	800	Two fragment chimera, with 3' end <i>Deltaproteobacteria</i> in origin.
AJ289180	<i>Fusobacteria</i>	930, 1210	Three fragment chimera, with 5' end <i>Fusobacteria</i> , middle <i>Spirochaetes</i> , and 3' end <i>Bacteroidetes</i> in origin.
AJ441248	<i>Fusobacteria</i>	~580	Two fragment chimera, with 3' end of unclassified origin. Exact position of break unclear due to lack of full-length subjects.

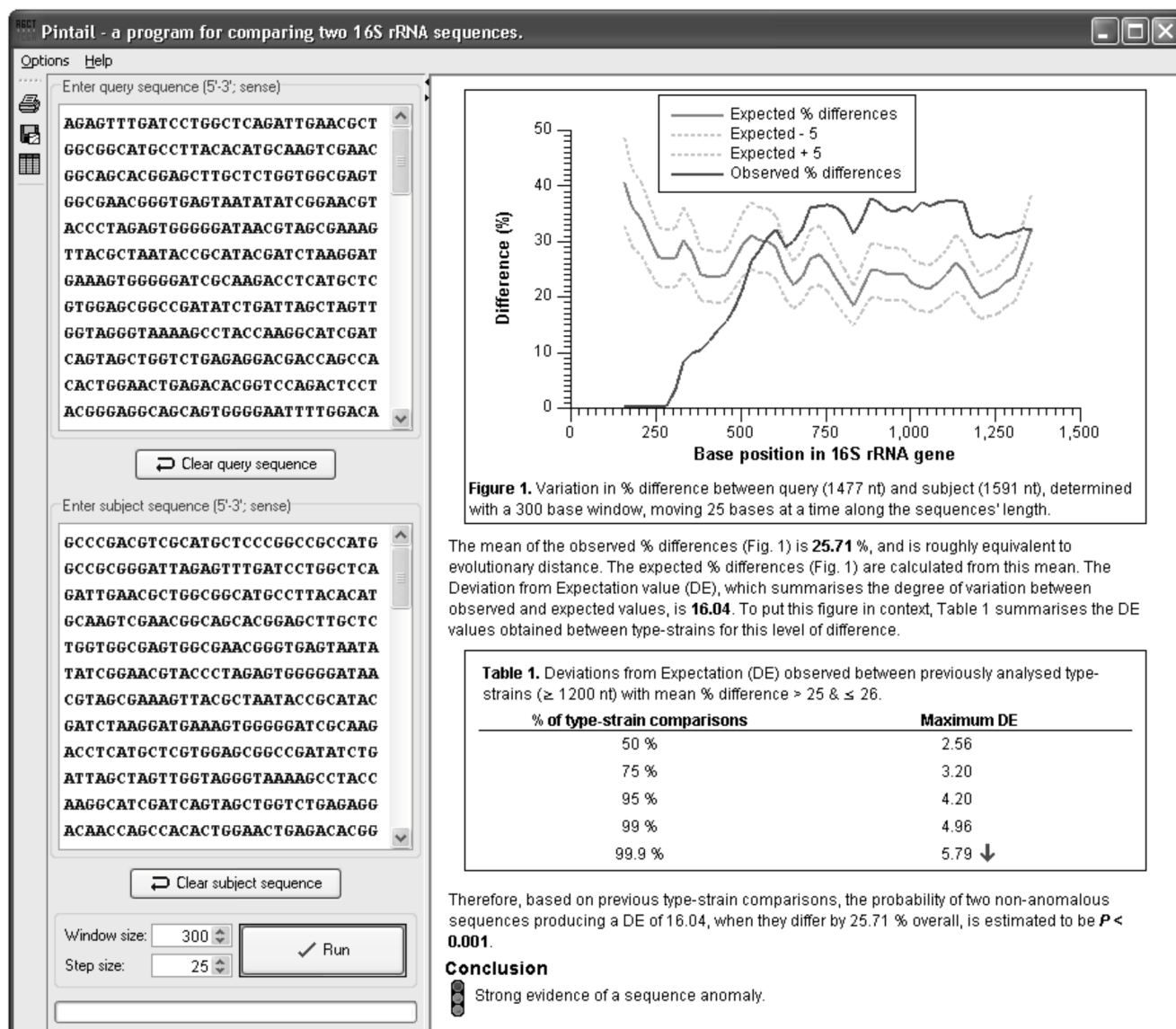
AY548992	<i>Fusobacteria</i>	280, 790	Three fragment chimera, with 5' end <i>Epsilonproteobacteria</i> , middle <i>Fusobacteria</i> , and 3' end <i>Epsilonproteobacteria</i> in origin.
AJ441228	<i>Fusobacteria</i>	150, 930	Two, possibly three, fragment chimera with 5' end unknown, middle unclassified, and 3' end <i>Epsilonproteobacteria</i> in origin.
AY548985	<i>Fusobacteria</i>	1140	Two fragment chimera, with 3' end of <i>Spirochaetes</i> origin.
AF287807	<i>Fusobacteria</i>	350	Two fragment chimera, with both fragments of <i>Fusobacteria</i> origin.
AF287808	<i>Fusobacteria</i>	920	Two fragment chimera, with both fragments of <i>Fusobacteria</i> origin.
AF366272	<i>Fusobacteria</i>	1160	Two fragment chimera, with both fragments of <i>Fusobacteria</i> origin.
AF385542	<i>Fusobacteria</i>	350	Very similar to AF287807.
Z94005	<i>Verrucomicrobia</i>	1025, 1150	Region 1025-1150 is alien to sequence but no close match found within database. Unusual nature of plot suggests sequencing error.
AJ401133	<i>Verrucomicrobia</i>	550	Two fragment chimera, with both fragments of <i>Verrucomicrobia</i> origin.
AJ401131	<i>Verrucomicrobia</i>	920	Two fragment chimera, with 5' end of unknown origin.
AF316731	<i>Verrucomicrobia</i>	300	Two fragment chimera, with 5' end of unclassified origin.
AJ401123	<i>Verrucomicrobia</i>	590	Two fragment chimera, with both fragments of <i>Verrucomicrobia</i> origin.
AB179538	<i>Verrucomicrobia</i>	570	Two fragment chimera, with 3' end of unknown origin.
AF351215	<i>Verrucomicrobia</i>	1080	Two fragment chimera, with 3' end of <i>Deltaproteobacteria</i> origin.
AJ617868	<i>Verrucomicrobia</i> *	1080	Two fragment chimera, with 3' end of <i>Deltaproteobacteria</i> origin.
AF234140	<i>Gemmatimonadetes</i>	-	Degenerate sequence - one large block of N bases.
AF009987	<i>Gemmatimonadetes</i>	-	Degenerate sequence - two large blocks of N bases.
AY218634	<i>Gemmatimonadetes</i>	700	Two fragment chimera, with both fragments of <i>Gemmatimonadetes</i> origin.
AY221051	<i>Gemmatimonadetes</i>	~900	Two fragment chimera, with both fragments of <i>Gemmatimonadetes</i> origin; break-point uncertain due to quality of available subject sequences.
AJ582052	<i>Gemmatimonadetes</i>	600, 950	Likely sequencing error.
AY218706	<i>Gemmatimonadetes</i>	275	Likely sequencing error at 5' end.
AF368188	<i>OP10</i>	1100	Two fragment chimera, with 3' end of probable <i>Bacteroidetes</i> origin.
AF368185	<i>OP10</i>	~260, 970	Likely three fragment chimera with 5' and 3' ends originating from some unknown source.
AF368184	<i>OP10</i>	~260, 970	Same as AF368185.
AY693838	<i>OP11</i>	520	Two fragment chimera, with 5' end of <i>Betaproteobacteria</i> origin.

AY218572	<i>OP11</i>	660	Two fragment chimera, with 5' end of <i>Epsilonproteobacteria</i> origin.
AJ582211	<i>OP11</i>	~560	Likely two fragment chimera, with 3' end of unknown origin.
AF513093	<i>TM7</i>	900	Two fragment chimera, with 3' end of unclassified origin.
AJ318135	<i>TM7</i>	1090	Two fragment chimera, with 3' end of <i>Actinobacteria</i> origin.
AY592328	<i>WS3</i>	380	Two fragment chimera, with 5' end of <i>Actinobacteria</i> origin.
AY217439	<i>Dehalococcoides</i>	500, 675	Likely sequencing error.
AY133080	<i>Dehalococcoides</i>	1400	Likely sequencing error.
AY548991	<i>Epsilonproteobacteria</i> **	320	Two fragment chimera, with 5' of <i>Gammaproteobacteria</i> origin, 3' of <i>Epsilonproteobacteria</i> origin.
AJ441247	Unclassified Bacteria**	380	Two fragment chimera, with 5' end of <i>Deltaproteobacteria</i> origin, 3' end of <i>Chloroflexi</i> origin.
AY762628	<i>Betaproteobacteria</i> **	780	Two fragment chimera, with both fragments of <i>Betaproteobacteria</i> origin.
AY762632	<i>Betaproteobacteria</i> **	780	Practically identical to AY762628.
AJ582208	Unclassified Bacteria**	540	Two fragment chimera, with 5' of <i>Firmicutes</i> origin, and end 3' of <i>Gemmatimonadetes</i> origin.
AY218710	Unclassified Bacteria**	630	Sequencing error in which first ~152 bases are the reverse complement.
AY280419	Unclassified Bacteria**	520	Two fragment chimera, with 5' of <i>Bacteroidetes</i> origin, and 3' end of <i>WS3</i> origin.
AB007420	<i>Actinobacteria</i> **	40 to 250	Likely sequencing error.

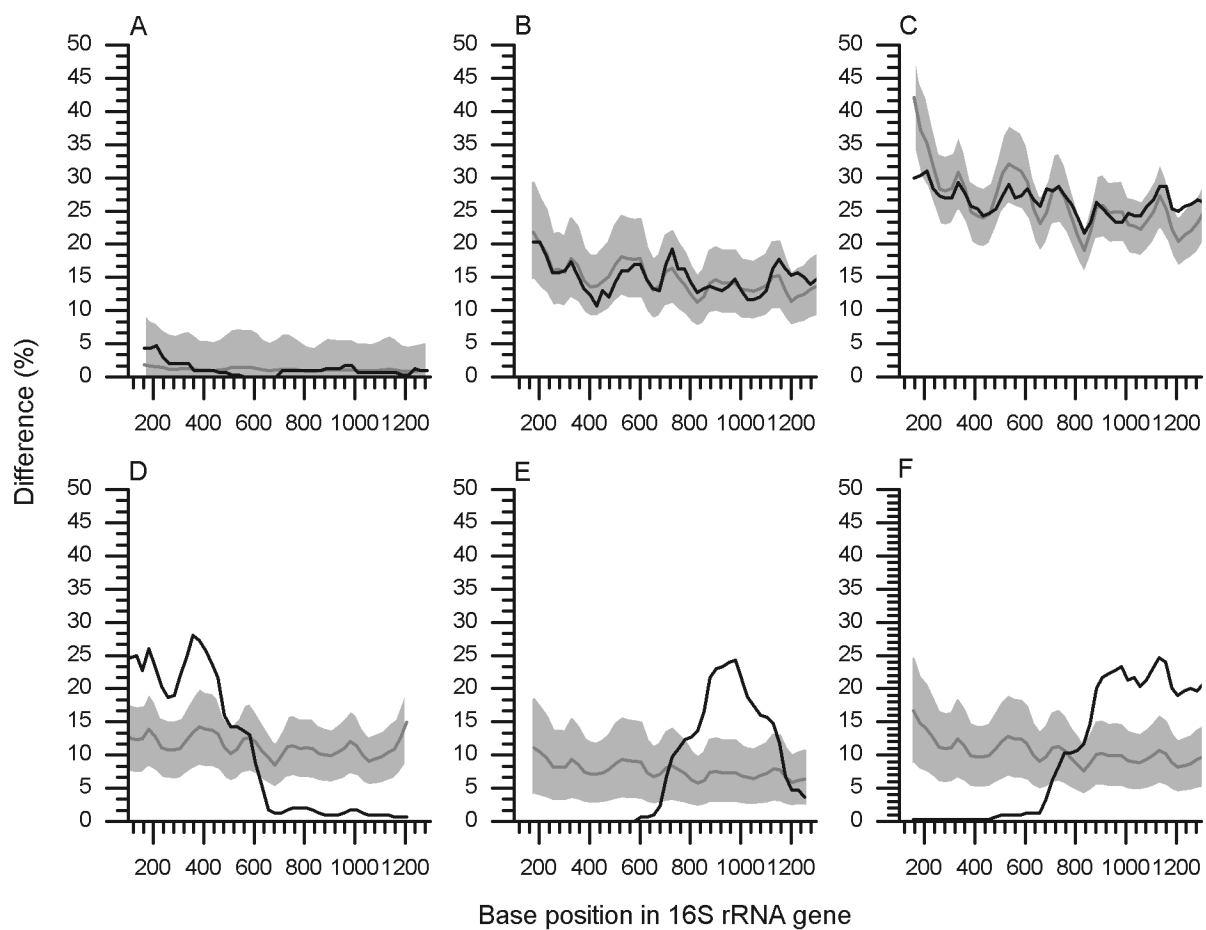
1 * Added after September release (not included in calculations).

2 ** Uncovered during analysis (not included in calculations).

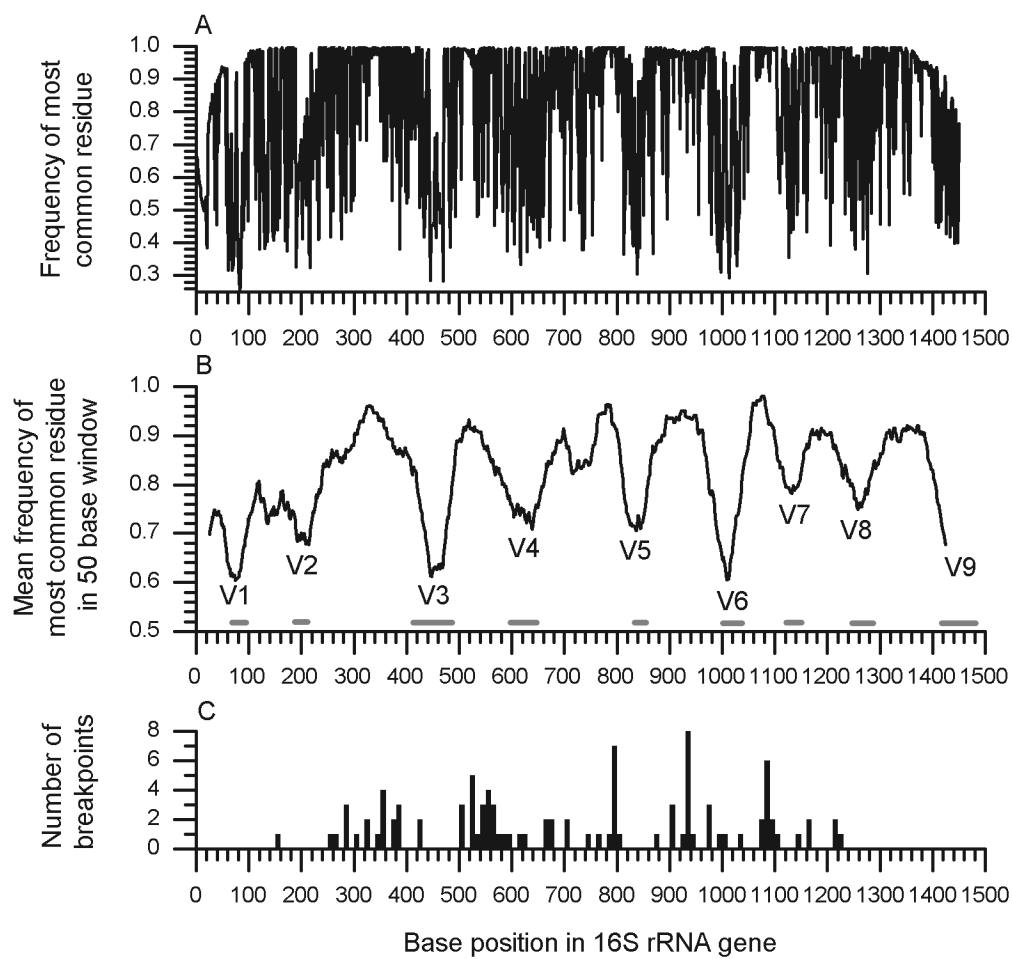
1 Figure 1



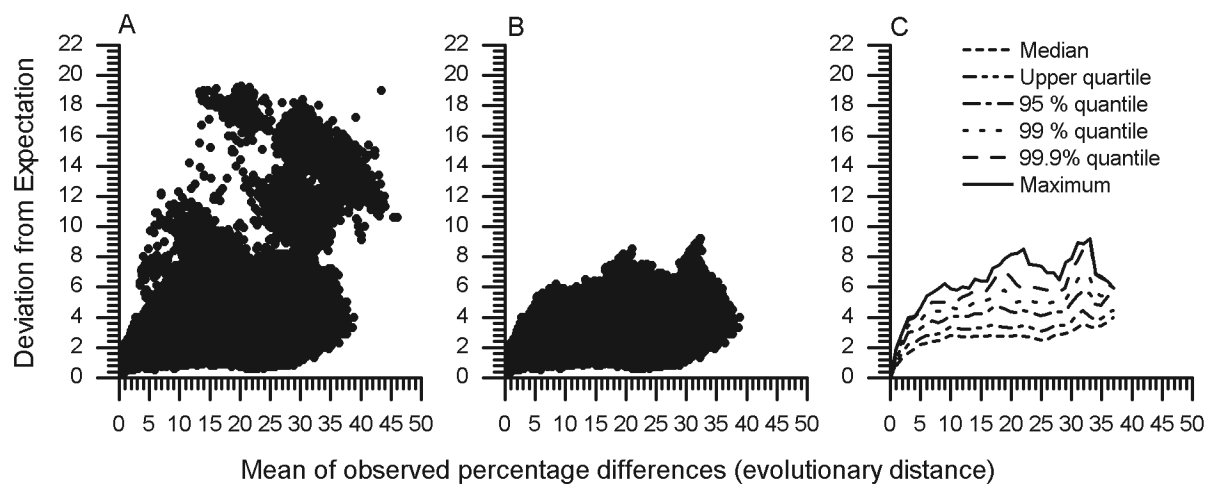
1 Figure 2



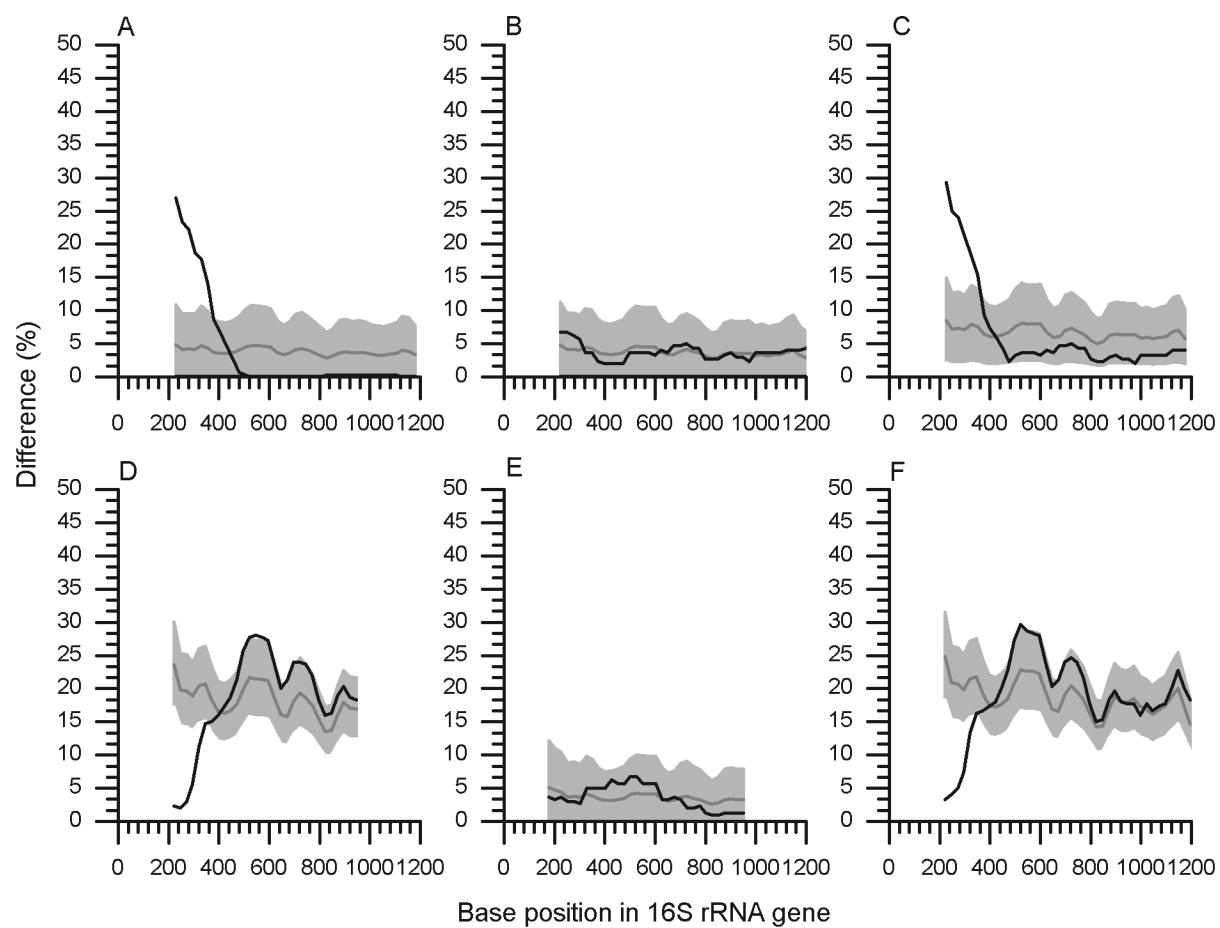
1 Figure 3



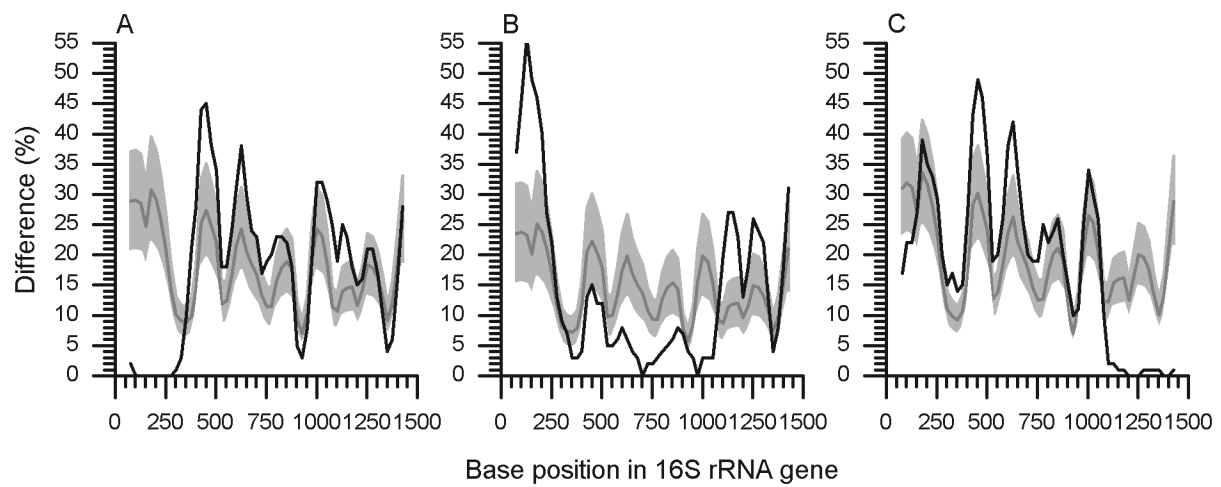
1 Figure 4



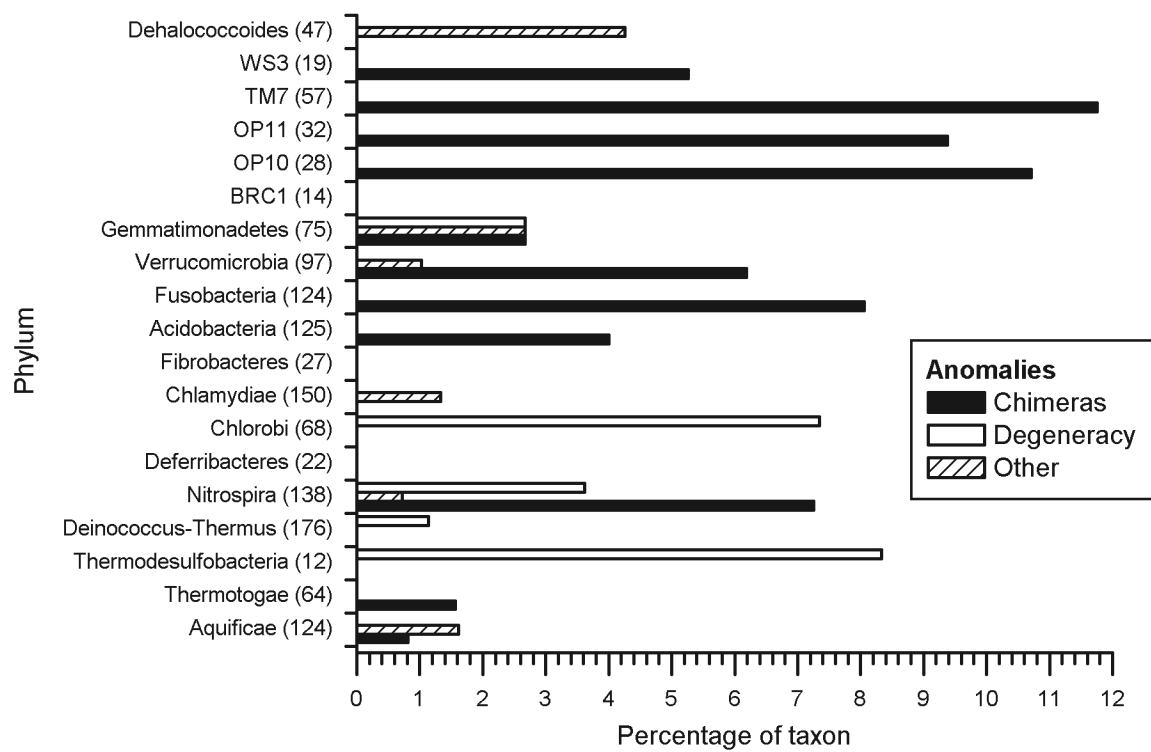
1 Figure 5



1 Figure 6



1 Figure 7



1 Figure 8

