INTRODUCTION
Culture-independent studies based on obtaining 16S rRNA genes directly from the environment by broad-specificity primer PCR and cloning have greatly improved our understanding of microbial diversity (Hugenholtz et al., 1998; Pace, 1997). However, such PCR-based surveys have a number of recognized limitations (Hugenholtz & Goebel, 2001; von Wintzingerode et al., 1997), perhaps the most insidious of which is the formation of recombinant or chimeric sequences during PCR amplification. Chimera formation is thought to occur when a prematurely terminated amplicon reanneals to a foreign DNA strand and is copied to completion in the following PCR cycles. This results in a sequence composed of two or more phylogenetically distinct parent sequences and, when comparatively analysed with other 16S rRNA sequences, suggests the presence of a non-existent organism. This problem was recognized early on in the application of PCR-clone library studies (Kopczynski et al., 1994; Liesack et al., 1991) and significant efforts have been made both to quantify (and hopefully reduce) chimera formation (Qiu et al., 2001; Speksnijder et al., 2001; Wang & Wang, 1996, 1997) and to improve their detection (Komatsuolis & Waterman, 1997; Liesack et al., 1991; Maidak et al., 2001; Robinson-Cox et al., 1995). Despite these precautions, a surprising number of chimeric 16S rDNA sequences from molecular phylogenetic surveys were detected in the public databases during a recent collation (Hugenholtz, 2002).

METHODS

Phylogenetic analysis. 16S rDNA sequences from several PCR-clone library studies were obtained from the public databases and imported into an ARB (http://www.arb-home.de/) database, where they were automatically aligned against existing sequences using FAST ALIGNER (version 1.03) followed by a manual refinement of the alignment. Studies were selected on the basis that at least one sequence in the study had been putatively identified as chimeric during routine database updating. Only almost-complete 16S rDNA sequences (>1300 nt) were included in the analysis, because the phylogenetic placement of shorter sequences can be unreliable, particularly if they lack close relatives in the database (Hugenholtz et al., 1998).Datasets comprising all sequences (>1300 nt) from a single study and 341 or 200 reference sequences representing the bacterial or archaeal domains, respectively, were selected for phylogenetic inference (Hugenholtz, 2002). These datasets are available through the Ribosomal Database Project (RDP; Maidak et al., 2001; http://rdp.cme.msu.edu/html/alignments.html).

Evolutionary distance trees were inferred independently from 5' and 3' halves of each dataset (partial treeing) applying the Lane mask (Lane, 1991) from absolute positions 0 to 4000 (635 nt for comparative analysis equivalent to Escherichia coli positions 28–762; 5' half) and 4000 to 0 (653 nt equivalent to E. coli positions 762–1512; 3' half) using the 'column selection' option in the filter selection menu. The environmental clone sequences in the dataset were then marked and tree topologies were compared for branching incongruencies indicative of chimeric sequences (Wang & Wang, 1997). The alignments of putatively identified chimeras were examined against their closest 5' and 3' matches (at least two of each) and inspected for nucleotide signature shifts characteristic of chimeric sequences (Wang & Wang, 1997). Breakpoints (also known as chimeric junctions or recombination sites) were estimated as being halfway between the change of nucleotide signatures characteristic of each parent group. Exact breakpoints are difficult to determine because the parent sequences are usually identical around the recombination site (Hugenholtz & Goebel, 2001). Positively identified chimeras were flagged in the database by appending ‘#’ to the clone name and annotating the ‘warning’ field with the affiliations of the parent sequences and approximate breakpoint. This information is summarized in Table 1.

RESULTS AND DISCUSSION
A sampling of nine published and three unpublished studies with publicly available sequence data revealed 21 inter-phylum and 18 intra-phylum chimeras by partial treeing (Table 1). Numerous smaller local topological rearrangements of sequences in the partial trees were observed,
<table>
<thead>
<tr>
<th>Putative chimeric sequence (accession no.)</th>
<th>5' Parent sequence</th>
<th>3' Parent sequence</th>
<th>Approx. breakpoint (E. coli numbering)</th>
<th>Reference study</th>
<th>Chimera detection†</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inter-phylum</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arctic96AD-3 (AF354607)</td>
<td>Proteobacteria (Gamma)</td>
<td>Arctic96AD-9 (AF354608) (98%)</td>
<td>Gemmatimonas BD7-2 (AB015578) (92%)</td>
<td>935</td>
<td>Bano &amp; Hollibaugh (2002) C, P</td>
</tr>
<tr>
<td>Arctic95A-3 (AF355043)</td>
<td>Proteobacteria (Delta)</td>
<td>Arctic97A-12 (AF355044) (97%)</td>
<td>Marine group A Arctic96B-7 (AF355047) (91%)</td>
<td>675</td>
<td>Bano &amp; Hollibaugh (2002) C, P</td>
</tr>
<tr>
<td>AT425_EubG1 (AY053495)</td>
<td>OP1</td>
<td>KTK 32 (AJ133617) (87%)</td>
<td>Proteobacteria (Gamma) Pseudomonas sp. G2 (AF326356) (98%)</td>
<td>1215</td>
<td>Lanoil et al. (2001) N</td>
</tr>
<tr>
<td>BD7-9 (AB015584)</td>
<td>Proteobacteria (Epsilon)</td>
<td>BD7-6 (AB015582) (98%)</td>
<td>Actinobacteria BD2-10 (AB015539) (94%)</td>
<td>1085</td>
<td>Li et al. (1999) S</td>
</tr>
<tr>
<td>BPC043 (AF154080)</td>
<td>Firmicutes</td>
<td>BPC060 (AF154081) (98%)</td>
<td>OP9 Acidobacteria GCA018 (AF154105) (98%)</td>
<td>1070</td>
<td>Unpublished U</td>
</tr>
<tr>
<td>BPC061 (AF154092)</td>
<td>Planctomycetes</td>
<td>Pirellula sp. Schlesner 516 (X81940) (90%)</td>
<td>Acidobacteria BPC015 (AF154085) (96%)</td>
<td>930</td>
<td>Unpublished U</td>
</tr>
<tr>
<td>BPC005 (AF154090)</td>
<td>Proteobacteria (Delta)</td>
<td>Desulfocapsa sulfuressis (Y13672) (94%)</td>
<td>Acidobacteria BPC110 (AF154084) (98%)</td>
<td>875</td>
<td>Unpublished U</td>
</tr>
<tr>
<td>BPC066 (AF154095)</td>
<td>Acidobacteria</td>
<td>wb1_A08 (AF317741) (93%)</td>
<td>Chloroflexi BPC110 (AF154084) (91%)</td>
<td>380</td>
<td>Unpublished U</td>
</tr>
<tr>
<td>d011 (AF422631)</td>
<td>TM6</td>
<td>Ebpr8 (AF255643) (92%)</td>
<td>Firmicutes d154 (AF422677) (99%)</td>
<td>350</td>
<td>Stein et al. (2001) C</td>
</tr>
<tr>
<td>MNB2 (AF293011)</td>
<td>Proteobacteria (Alpha)</td>
<td>Methylasinos sp. LW4 (AY07293) (90%)</td>
<td>Nitrospira MNC2 (AF293010) (98%)</td>
<td>1055</td>
<td>Stein et al. (2001) C</td>
</tr>
<tr>
<td>MNF8 (AF293012)</td>
<td>Nitrospira</td>
<td>MNC2 (AF293010) (97%)</td>
<td>Proteobacteria (Gamma) SM2E06 (AF445725) (89%)</td>
<td>990</td>
<td>Unpublished U</td>
</tr>
<tr>
<td>MNA3 (AF293013)</td>
<td>Nitrospira</td>
<td>MNC2 (AF293010) (98%)</td>
<td>Proteobacteria (Gamma) Iron-oxidizer m-1 (AF387301) (93%)</td>
<td>930</td>
<td>Stein et al. (2001) C</td>
</tr>
<tr>
<td>UASB_TL84 (AF254390)</td>
<td>Chloroflexi</td>
<td>BPC110 (AF154084) (90%)</td>
<td>Proteobacteria (Delta) UASB_TL15 (AF254406) (99%)</td>
<td>900</td>
<td>Wu et al. (2001a) C, S, P</td>
</tr>
<tr>
<td>UASB_TL94 (AF254401)</td>
<td>OP9</td>
<td>BA021 (AF323760) (94%)</td>
<td>Spirochaetes 'Leptospira illini' (M88719) (99%)</td>
<td>1080</td>
<td>Wu et al. (2001a) C, S, P</td>
</tr>
<tr>
<td>UASB_TL56 (AF254405)</td>
<td>Proteobacteria (Delta)</td>
<td>UASB_TL44 (AF254395) (99%)</td>
<td>OP8 OPB95 (AF027060) (98%)</td>
<td>620</td>
<td>Wu et al. (2001a) C, S, P</td>
</tr>
<tr>
<td>UASB_TL15 (AF254406)</td>
<td>OP11</td>
<td>TA18 (AF229791) (99%)</td>
<td>Proteobacteria (Delta) UASB_TL11 (AF254397) (99%)</td>
<td>930</td>
<td>Wu et al. (2001a) C, S, P</td>
</tr>
<tr>
<td>VC.2.1 Bac7 (AF068788)</td>
<td>Aquificae</td>
<td>VC2.1 Bac13 (AF068793) (99%)</td>
<td>Proteobacteria (Epsilon) VC2.1 Bac8 (AF068789) (99%)</td>
<td>420</td>
<td>Reysenbach et al. (2000) C, H</td>
</tr>
<tr>
<td>VC.2.1 Bac9 (AF068790)</td>
<td>Aquificae</td>
<td>VC2.1 Bac13 (AF068793) (100%)</td>
<td>Proteobacteria (Epsilon) VC2.1 Bac8 (AF068789) (99%)</td>
<td>420</td>
<td>Reysenbach et al. (2000) C, H</td>
</tr>
<tr>
<td>VC.2.1 Bac32 (AF068806)</td>
<td>Proteobacteria (Epsilon)</td>
<td>VC1.2-cd26 (AF367490) (98%)</td>
<td>Aquificae VC2.1 Bac10 (AF068791) (99%)</td>
<td>1090</td>
<td>Reysenbach et al. (2000) C, H</td>
</tr>
<tr>
<td>Putative chimeric sequence (accession no.)</td>
<td>Parent sequence</td>
<td>Parent sequence</td>
<td>Approx. breakpoint (E. coli numbering)</td>
<td>Reference study</td>
<td>Chimera detection†</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>----------------------------------------</td>
<td>-----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>VC2.1 Bac43 (AF068810)</td>
<td>Ferribacter</td>
<td>5′ Parent sequence</td>
<td>Proteobacteria (Epsilon)</td>
<td>675</td>
<td>C, H</td>
</tr>
<tr>
<td>YNPFPP89 (AF391983)</td>
<td>Acidobacteria</td>
<td>3′ Parent sequence</td>
<td>Actinobacteria</td>
<td>550</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Intra-phyllum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arctic96B-6 (AF353224)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>530</td>
<td>C, P</td>
</tr>
<tr>
<td>Arctic96BD-1 (AF354605)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>615</td>
<td>C, P</td>
</tr>
<tr>
<td>AT425_EubD5 (AY053489)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>555</td>
<td>N</td>
</tr>
<tr>
<td>BD3-1 (AB015547)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>790</td>
<td>S</td>
</tr>
<tr>
<td>BD6-4 (AB015574)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>520</td>
<td>S</td>
</tr>
<tr>
<td>BPC023 (AF154087)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>1070</td>
<td>Unpublished</td>
</tr>
<tr>
<td>d035 (AF422644)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>283</td>
<td>U</td>
</tr>
<tr>
<td>MNG7 (AF292997)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>1030</td>
<td>C</td>
</tr>
<tr>
<td>MNH4 (AF293002)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>550</td>
<td>C</td>
</tr>
<tr>
<td>MNA5 (AF293005)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>1165</td>
<td>C</td>
</tr>
<tr>
<td>pIVWA101 (AB019722)</td>
<td>Crenarchaeota</td>
<td></td>
<td></td>
<td>975</td>
<td>C</td>
</tr>
<tr>
<td>SAGMA-C (AB050207)</td>
<td>Crenarchaeota</td>
<td></td>
<td></td>
<td>520, 790</td>
<td>C</td>
</tr>
<tr>
<td>SAGMA-Z (AB050231)</td>
<td>Crenarchaeota</td>
<td></td>
<td></td>
<td>720</td>
<td>C</td>
</tr>
<tr>
<td>SAGMA-3 (AB050234)</td>
<td>Crenarchaeota</td>
<td></td>
<td></td>
<td>520</td>
<td>C</td>
</tr>
<tr>
<td>TA07 (AF229780)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>790</td>
<td>N</td>
</tr>
<tr>
<td>TA09 (AF229782)</td>
<td>Proteobacteria</td>
<td></td>
<td></td>
<td>970</td>
<td>N</td>
</tr>
</tbody>
</table>

*Phylum (Class)*

†Chimera detection

http://ijs.sgmjournals.org
suggesting the presence of additional chimeras created from closely related parent sequences (data not shown). Interphylum chimeras are particularly problematic in phylogenetic inference, as they can result in novel lines of descent that may be misinterpreted as representing novel species, genera or even families of prokaryotes, although they should also be the easiest to detect (von Wintzingerode et al., 1997). The most popular method for detecting chimeric 16S rDNA sequences is the CHIMERA_CHECK program available through the RDP (Maidak et al., 2001). This is one of a number of nearest-neighbour methods that detect chimeric sequences by determining whether fragments of two independent database entries have a higher overall similarity to the query sequence than a single, full-length database entry (Komatsoulis & Waterman, 1997; Robinson-Cox et al., 1995). Unfortunately, once a chimeric sequence is added to the RDP database it becomes invisible to CHIMERA_CHECK because it is simply compared against itself in the analysis. Currently, there is no way to use CHIMERA_CHECK against subsets of the database, such as 16S rDNA sequences from cultivated organisms, to bypass this problem.

One instance of a chimeric sequence with two breakpoints was detected, SAGMA-C (Table 1). PCR-mediated recombinant sequences with multiple recombination sites have been documented previously (Bradley & Hillis, 1997) and are more likely to occur between closely related sequences, as seen in this instance. The sequences presented in Table 1 reduce the quality of the public databases and should be removed, or divided at the breakpoint and resubmitted as separate entries designated A and B to distinguish the chimeric fragments.

This study is by no means an exhaustive search of the public databases and simply serves to illustrate that chimeric 16S rDNA sequences are being overlooked in molecular phylogenetic surveys, despite a general appreciation of PCR-generated chimeras amongst researchers. Nearest-neighbour chimera detection methods should be routinely supplemented with partial treeing analysis, as this method is less sensitive to the absence of closely related parent sequences in the databases and is relatively simple to implement using ARB. In addition, we have recently written a program, called BELLEROPHON, that detects chimeric sequences in aligned datasets based on partial treeing analysis; this program is available online (http://cassandra.visac.uq.edu.au/perl/bellerophon.pl).

NOTE ADDED IN PROOF
One of the three unpublished studies addressed in this paper has now been published (Lowe et al., 2002; sequences d011 and d035 in Table 1).

REFERENCES
Chimeric 16S rDNA sequences in the public databases


