Differences in rDNA libraries of faecal bacteria derived from 10- and 25-cycle PCRs

Régis Bonnet,1 Antonia Suau,1 Joël Doré,1 Glenn R. Gibson2 and Matthew D. Collins2

Author for correspondence: Régis Bonnet. Tel: +33 1 34 65 23 04. Fax: +33 1 34 65 24 92.
e-mail: r.bonnet@caramail.com

In spite of its shortcoming, analysis of PCR-derived rDNA libraries is being employed increasingly to investigate microbial diversity within many ecosystems. In the present investigation, the effects of the number of PCR cycles (10 vs 25 cycles) on the inferred structure of a 16S rDNA library have been examined. Seventy-five 25-cycle sequences were retrieved and analysed in comparison with 284 10-cycle sequences already described in a previous study. The 359 clones obtained were classified into 94 molecular species (at least 98% sequence similarity). At the level of large phylogenetic groups, the two cloned rDNA libraries were not different. A mathematical model was developed in order to estimate the number of molecular species expected if further sequencing was performed. Coverage-based computing, projections and statistical analysis demonstrated that the structures of the two PCR-derived rDNA libraries were different and that the 25-cycle rDNA library displayed reduced diversity. It is suggested that the number of PCR cycles used for amplification of 16S rDNA genes for phylogenetic diversity studies must therefore be kept as small as possible.

Keywords: biodiversity, human gut flora, PCR, 16S rRNA

INTRODUCTION

The human large intestine harbours a complex community of micro-organisms that plays an important role in the health and physiological function of the host (Ducluzeau, 1988; Gibson & Roberfroid, 1995). There is increasing recognition that culture-dependent approaches provide an incomplete picture of the diversity of the gut microflora. The tools of molecular biology are now providing a means of bypassing culture techniques by accessing directly the diversity of bacterial rDNA in natural communities. Such rDNA PCR-based direct community analyses are being employed increasingly to elucidate microbial diversity within many ecosystems: e.g. rumen (Whitford et al., 1998), anaerobic digestors (Godon et al., 1997), hot springs (Barns et al., 1994), soil (Liesack & Stackebrandt, 1992) and marine sediments (Devereux & Mundfrom, 1994). Indeed, the recent application of such methodologies has indicated that 50–90% of the total human gut microflora has not been cultivated (Langendijk et al., 1995; Wilson & Blitchington, 1996; Wang et al., 1996; Suau et al., 1999) and that the human colon houses many unknown bacterial lineages.

Although rDNA PCR-based direct community analysis offers a powerful alternative to culture-based methodologies, it is now apparent that such a direct approach also has some shortcomings. Several factors may bias relative frequencies of rDNA genes from PCR products in mixed-template reactions (von Wintzingerode et al., 1997). For example, Chandler et al. (1997) have shown that the PCR template concentration may affect the composition of clone libraries. Similarly, the number of cycles carried out with PCR may also play a significant role. Defined mixtures of two templates have been used to study mechanisms that can lead to non-proportional amplification when a large number of cycles is used (Suzuki & Giovannoni, 1996; Mathieu-Daudé et al., 1996; Polz & Cavanaugh, 1998). Some investigations (Wilson & Blitchington,
R. Bonnet and others

1996; Whitford et al., 1998) have also reported a loss or distortion of rDNA sequence diversity when amplification was conducted using a large number of PCR cycles.

Recently, Suau et al. (1999) retrieved and performed a comprehensive phylogenetic analysis on a large number of 16S rDNA bacterial sequences generated with a 10-cycle PCR from a human faecal sample. In the present study, we have used the 10-cycle microflora inventory of Suau et al. (1999) as a comparative baseline for examining the effect of PCR cycle number on the inferred library structure (qualitative and quantitative). We herein report on mathematical tools that allow the statistical comparison of cloned rDNA libraries and demonstrate a significant difference in amplified rDNA libraries when using different numbers of cycles in PCR. We use the terms ‘PCR-derived rDNA libraries’ to describe the complete population of PCR products and ‘cloned rDNA libraries’ to describe the sequenced fraction of the former.

**METHODS**

**Extraction and purification of total DNA.** A faecal sample was taken from a healthy 40-year-old male. Immediately after collection, total DNA was extracted and purified from a 125 mg aliquot of the faecal sample as described by Godon et al. (1997). Chromosomal DNA integrity was confirmed using agarose gel electrophoresis [1.5% (w/v) agarose, 1 × TBE and 1 ng ethidium bromide ml⁻¹]. The DNA concentration was estimated to be 225 ng µl⁻¹.

**PCR conditions.** Amplification of bacterial 16S rRNA genes by PCR was carried out as described previously by Suau et al. (1999), using the same primers. The same DNA sample was used and the number of PCR cycles was increased to 25.

**Cloning and sequencing.** PCR products were purified and cloned as described by Suau et al. (1999). Sequencing reactions were performed with the dRhodamine terminator cycle sequencing ready reaction kit with AmpliTaq DNA polymerase FS (Perkin-Elmer) in a Gene Amp PCR system 9600 thermocycler (Perkin-Elmer Cetus). Plasmid inserts were sequenced with an automated ABI Prism 377 DNA sequencer (Applied Biosystems, Perkin-Elmer). Two primers were used: S*-Univ-0536-a-A-18 (5'GTATTACCGCGGCTGCT-3'); Giovannoni et al., 1988) and S-D-Bact-0947-a-19 (5'TTCGAATTAAACCACATGC-3'). The mean fragment length was 850 bases.

**Sequence analysis.** Phylogenetic analysis and tree construction were conducted as reported by Suau et al. (1999). The terms operational taxonomic unit (OTU) or ‘rDNA species’ were given to a group of sequences with at least 98% similarity. This threshold was used by Suau et al. (1999) and is consistent with 16S rRNA homology and DNA–DNA reassociation values (Stackebrandt & Goebel, 1994) and other observations (Godon et al., 1997). An OTU represented by a unique clone (e.g. sampled only once) was called a single-clone OTU. The stability of phylogenetic relationships was assessed using the bootstrap method with the programs SEQBOOT (250 replicates), DNADIST, NEIGHBOR and CONSENSE (Felsenstein, 1993). To detect chimeras, sequences were analysed with the RDP CHECKCHIMERA program (Maidak et al., 1999). Branching of trees based on the 5' end and on the 3' end of the sequences were also compared (Kopczynski et al., 1994). A sequence retrieved from both the 10- and 25-cycle studies was assumed not to be chimeric.

**Calculation of coverage and theoretical computing.** The coverage gives an estimate of the proportion of the amplified rDNA community represented in a clone library. It can also be described as the probability that the next sequence analysed belongs to an already identified OTU. Using the coverage formula of Good (1953), three mathematical functions were computed, \( N_s \), \( C_s \) and \( O_s \), which respectively expressed the number of single-clone OTUs, the coverage and the number of OTUs found after sequencing \( n \) clones. Functions are noted as \( (F_s) \), and the corresponding values as \( F_s \).

\( C_s \) has been estimated by Good (1953) as:

\[
C_s = 1 - \left( \frac{N_s}{n} \right)
\]

If we assume coverage to be estimated by the previous formula, \( N_s+1 \) can be inferred from \( N_s \); to \( N_s \), add 1 if the \( n+1 \)th clone sequenced has not been retrieved previously (the probability of retrieving a new OTU after sequencing of \( n \) clones is \( N_s/n \) or subtract 1 if the \( n+1 \)th clone has been retrieved previously [probability: \( 1-(N_s/n) \)] and if this clone falls into a single-clone OTU (probability: \( N_s/n \)). \( N_s+1 \) can thus be described as:

\[
N_s+1 = N_s+(N_s/n) - (1-(N_s/n)) \times (N_s/n)
\]

which is equal to:

\[
N_s+1 = N_s+(N_s/n)^\circ
\]

Likewise, the number of OTUs after \( n+1 \) sequences can be defined as the sum of the number of OTUs after \( n \) sequences and the expected number of OTUs not previously retrieved when sequencing the \( n+1 \)th clone, which is equal to \( (N_s/n) \).

Thus:

\[
O_s+1 = O_s + (N_s/n)
\]

The expected \( O_s \) values for the 25-cycle study are reported in Fig. 2. This comparison indicated that diversity was higher in the 10-cycle study. The coverage curves led to the same conclusion (data not shown): coverage for the 25-cycle study rose swiftly and reached a value that the 10-cycle clone library would not attain before sequencing 150 clones.

**Reference strains and cloned sequences used in phylogenetic analysis.** Strains used for phylogenetic analysis and not reported in Suau et al. (1999) are as follows (GenBank accession numbers in parentheses): unidentified rumen bacterium (URB) JW23 (AF018552), URB JW33 (AF018559), URB RC7 (AF001700), URB RC26 (AF001715), URB RC39 (AF001724), URB RFN18 (AB009177). The accession numbers of the 10-cycle clones are AF132232–AF132286.

**Nucleotide sequence accession numbers.** Partial sequences of one representative of each OTU (except some OTUs that had already been recovered among the 10-cycle clones) were deposited in GenBank (accession numbers AF153851–AF153873).

**RESULTS**

The mean length of the 75 sequences analysed derived from the 25-cycle PCR was 850 bases (Escherichia coli positions 50–900). However, comparison between the 25-cycle and 10-cycle PCR-generated sequences was based on approximately 400 aligned, homologous
nucleotides (E. coli positions 50–450) because fewer nucleotides were determined during the sequencing phase of the 10-cycle-based study (Suau et al., 1999).

**Chimeras**

Six chimeras were detected amongst the 75 rDNA clones, representing 8% of all sequences, which was not significantly higher than the chimera frequency observed during the 10-cycle study of Suau et al. (1999). Four of the detected chimeras were composed of two sequences, both parents of which were identified in this study (hybrid sequences composed of 44.25c and 71.25c, 12.25c and 61.25c, 48.25c and 79.25c, 52.25c and 68.25c). The two remaining chimeric sequences detected were more complex to analyse. They did not consist of hybrids of retrieved 25-cycle sequences. However, one parental sequence of each chimera was found in the 10-cycle clones (Suau et al., 1999) (i.e. 309.10c or 217.10c).

**Global phylogenetic analyses**

Excluding the six chimeric sequences, the remaining 69 25-cycle sequences were recovered into 35 OTUs, which were distributed between three monophyletic groups: the Bacteroides group, the Clostridium cocoides group and the Clostridium leptum subgroup (RDP registration 2.15.1.2, 2.30.4.1 and 2.30.9.1.3, respectively). In the study of Suau et al. (1999), 11 OTUs (out of a total of 83) were recovered outside the aforementioned three major phylogenetic groups. No equivalents of these ‘exceptional’ OTUs were detected in the 25-cycle PCR. Chi-squared analysis of the relative distribution of OTUs and clones among the three major groups did not reveal any significant differences.

Only a quarter of the delimited OTUs contained recognized species (exhibiting < 2% sequence divergence from rDNA of reference type strains), encompassing a third of all clones. Similar proportions were found for the 10-cycle clones (Suau et al., 1999). Unknown clones (not related to any recognized species) were spread among groups in proportion to group size. Half of the OTUs contained only one clone, which gave a coverage of 75% using Good’s formula (Good, 1953) compared with 85% recovered in the 10-cycle investigation of 284 sequences. Among the 94 different OTUs recovered from the two studies, 25 were found to be common to both clonal populations. These represented 56 sequences (81%) of the 25-cycle study and 171 sequences (62%) of the 10-cycle study (Suau et al., 1999). Distribution of these common OTUs among the three groups did not differ significantly from the distribution of all OTUs (by chisquared analysis). Unknown OTUs represented 77% of all 25-cycle OTUs and encompassed two-thirds of the clones. Comparable data were reported in the 10-cycle study of Suau et al. (1999). Ten OTUs were unique to the 25-cycle study. Only three of these were represented by more than one clone (OTUs 75.25c, 30.25c and 35.25c; for phylogenetic placement, see Fig. 1).

**DISCUSSION**

Few methods are currently available for comparing diversity from different clone libraries. When using 16S rRNA gene cloning and sequencing strategies, the biodiversity of a natural sample is assessed from OTU diversity (i.e. the clonal library). Diversity in this case is a function of the number of OTUs present (OTU richness) and the evenness with which clones are distributed among these OTUs (OTU evenness) (Hurlbert, 1971). In this paper, distortion is defined as a difference in OTU evenness, and lack of diversity as lower OTU richness, when comparing two rDNA libraries. The use of coverage as well as standard indices is straightforward, but indices are often difficult to interpret. The OTU rarefaction curve (Fig. 2) is of special interest in this respect, as it gives a dynamic picture of the data and provides information on how estimations of diversity relate to the number of sequences determined.

**OTU richness**

The easiest way to compare the diversity of two clonal libraries is to study OTU richness. When libraries differ in size, which is the case in the present study, rarefaction is used (Hurlbert, 1971; Simberloff, 1978). It was estimated that the number of OTUs after the sequencing of 69 10-cycle clones was 38.8 ± 3 (Fig. 2, Table 1). We found 35 OTUs after the sequencing of 69 25-cycle clones, which falls outside the range from the 10-cycle study (35.8–41.8). The error associated with the estimate of 35 OTUs in the 25-cycle library could not be calculated. Therefore, no statistical test supported this difference. Another approach is to estimate the OTU richness of the whole PCR-derived rDNA library (and not only that of the clone library). This can be done, using the rarefaction curve. Recently, Sekiguchi et al. (1998) used a parabolic equation to approximate that curve and inferred $O_c$ from this approximation. We used the same range of $n$ (from 1 to 69) to compare the two studies. Strong correlation were demonstrated ($r^2 = 0.98$). It was estimated that $O_{25c} = 21$ was approximately 82 (lower estimation) for the 10-cycle rDNA amplified community and approximately 67 (upper estimation) for the 25-cycle rDNA amplified community. One can infer that the OTU richness of the 25-cycle community is lower. Predictions computed with our model (Fig. 2) confirmed that the OTU richness appears to be greatly affected when the number of cycles is increased during PCR amplification, leading to a smaller number of OTUs being amplified.

**OTU evenness**

Analysis of diversity using OTU evenness can be carried out by comparing, for each OTU, its frequency...
Fig. 1. For legend see facing page.
The number reported for the 10-cycle study is a lower estimate; none of the non-common 10-cycle OTUs was composed of more than eight clones. By relying upon visual inspection of the trees, it is possible to conclude inadvertently there is a large discrepancy between the two rDNA amplified communities (which expresses the fact that only 24 OTUs are common to both clone libraries). Statistical analysis, however, revealed that only three OTUs showed significant differences between the two studies. Likewise, comparison of the size of OTUs within one study could lead to a determination of main OTUs (i.e. OTUs the size of which is significantly greater than others). By focusing on the 25-cycle OTUs, those represented by fewer than five clones are not significantly more important in size (by Fisher’s exact test) than OTUs that were not retrieved (size equal to 0). As only one OTU included more than four clones, it was not possible to identify main OTUs. This emphasizes that statistical analyses are needed as adjuncts to treeing when comparing data that are acquired from clone libraries. The necessity to retrieve large numbers of sequences to be able to observe significant differences is also emphasized. Nevertheless, distortion was demonstrated and agrees with Whitford et al. (1998), who compared two 16S rDNA libraries obtained after 12- and 30-cycle PCR amplification of rumen contents. Diversity indices also showed that the evenness of the two communities was different, but they failed to demonstrate which of the two was the more even (Table 1).

---

Table 1. Diversity indices based on OTU distribution after 25- and 10-cycle PCR amplification

<table>
<thead>
<tr>
<th>Index</th>
<th>10-cycle study</th>
<th>25-cycle study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – SI*</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>Shannon H†</td>
<td>0.79</td>
<td>0.83</td>
</tr>
<tr>
<td>SD†</td>
<td>0.79</td>
<td>0.76</td>
</tr>
<tr>
<td>NM§</td>
<td>0.25</td>
<td>0.34</td>
</tr>
<tr>
<td>E (S)§</td>
<td>38.8 (3.0)</td>
<td>35</td>
</tr>
<tr>
<td>( O_{n=1,69} ) ‡</td>
<td>89</td>
<td>67</td>
</tr>
</tbody>
</table>

* Simpson’s index (Simpson, 1949).
† Shannon–Weiner function; for example, see Pielou (1975).
‡ Standard deviation; for example, see Fager (1972).
§ Index based on the number of ‘moves’ that would have to be made to convert an observed distribution of clones among OTUs into an even distribution (Fager, 1972).
|\| Calculated by rarefaction (Simberloff, 1978; Hurlbert, 1971) for a standardized sample size of 69 clones for each clone library. The value in parentheses is the variance arising from inherent statistical uncertainty during the estimation of OTU richness for sample sizes smaller than the actual sample size.
‡ Estimates the total number of OTUs that would be retrieved if further sequencing was carried out. A parabolic equation was used to approximate the curve representing the cumulative number of OTUs over the same range of \( n \) (from 1 to 69) for both studies, from which \( O_n \) is derived (Sekiguchi et al., 1998). The number reported for the 10-cycle study is a lower estimate; the number reported for the 25-cycle study is an upper estimate.

---

Fig. 1. Phylogenetic tree derived from partial 16S rDNA sequence data for members of the C. coccoides group. Phylogenetic analysis was based on analysis of approximately 450 nucleotides (E. coli positions 33–498). Names of 25-cycle clones are in bold. Sequences from the 25-cycle and 10-cycle studies are designated by the extensions ‘.25c’ and ‘.10c’. Numbers given between dashes (e.g. -3) represent numbers of similar sequences. Bar, 2% sequence divergence. The tree was constructed using the programs SIMILARITY and NEIGHBOR. Bootstrap values (%) are given for 250 replications. Clones that were significantly over-represented in one study compared with the second study (\( P < 0.05 \) by Fisher’s exact test) are indicated by asterisks.
Coverage

Coverage for the 25-cycle clones was 75%. This means that 75% of the rDNA sequences of the 25-cycle amplified community belonged to one of the OTUs of the clonal library. If we assume that no bias was introduced by PCR, 75% of the 10-cycle clones should fall into an OTU present in the 25-cycle clones. The 212 10-cycle sequences expected to belong to an OTU found among the 25-cycle clones (since 284 sequences had been analysed for the 10-cycle experiment of Suau et al., 1999) was significantly more ($P < 0.05$) than the 176 new sequences observed. This indicated either that the coverage was poorly estimated or that the assumption that both clonal libraries represented the same rDNA community was incorrect. Similarly, coverage for the 10-cycle clones was 85.5%; we therefore expected to observe 59 sequences from the 25-cycle clones to belong to an OTU observed in the 10-cycle study. Fifty-five of these sequences were actually retrieved; this difference was not significant.

Conclusions

Numerous studies have investigated the effect of DNA rehybridization by PCR, and all have concluded that reannealing of major PCR products becomes more probable than primer binding as PCR proceeds. Suzuki & Giovannoni (1996) and Mathieu-Daudé et al. (1996) compared the template-to-product ratios in PCRs involving two 16S rDNAs. They found that a discrepancy between template proportion and product proportion became strongly biased against major templates as the number of cycles increased, allowing rarer rDNAs to be sampled more efficiently. The product diversity of multi-template PCR should then increase with the number of replication cycles; however, faecal samples contain so many bacterial taxa that even the most frequent may constitute only a small proportion of the entire population. As pointed out by Wang & Wang (1996), there is therefore little chance that, as the PCR proceeds, the less-abundant, prematurely terminated DNA molecules would anneal with rDNAs of the same species. This phenomenon may explain the lack of diversity when faecal DNA samples are amplified with a large number of PCR cycles. On the other hand, three 25-cycle OTUs were over-represented compared with the 10-cycle study, which is contrary to the above argument. This may be due to variation in amplification during the early cycles, when primers still target the genomic DNA. Polz & Cavanaugh (1998) showed that this phenomenon could result in little bias, but also observed that some occasional over-amplification could occur.

It is now recognized that culture-dependent techniques can strongly bias and underestimate the diversity of environmental microbial populations. Molecular tools based on 16S rDNA amplification circumvent the need to culture organisms in such investigations. However, the data presented here further demonstrate that the number of cycles used during PCR amplification may affect the diversity (mainly in terms of richness) of the amplified 16S rDNAs. The distortion observed when comparing two clone libraries may also be due in part to stochastic amplification during the early cycles. In conclusion, we have shown that comparison of clone libraries must be based on statistical analysis, since the clone libraries are only a partial estimate of PCR-derived rDNA libraries. Therefore, a significant number of clones must be sequenced and the coverage value used to estimate whether the information retrieved is sufficient.

ACKNOWLEDGEMENTS

We thank Roger Hutson for help in sequencing. This study was supported in part by the European Research Project FLAIR CT97-3035. The 25-cycle analysis was initiated during the Molecular Ecology Spring Session Laboratory, Chair of Biology, Conservatoire National des Arts et Metiers, under Dr P. Pochart’s supervision.

REFERENCES


16S rDNA libraries derived from 10- and 25-cycle PCRs


