Nucleic Acids Research

Unlocking hidden genomic sequence

Jonathan M. Keith, Duncan A. E. Cochran, Gita H. Lala, Peter Adams, Darryn Bryant and Keith R. Mitchelson

Nucleic Acids Res. 32:35-, 2004.
doi:10.1093/nar/gnh022

Published online February 18, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 3 ***e35***
DOI: 10.1093/nar/gnh022

Unlocking hidden genomic sequence

Jonathan M. Keith*, Duncan A. E. Cochran[1], Gita H. Lala[1], Peter Adams, Darryn Bryant and Keith R. Mitchelson[1,2]

Department of Mathematics, [1]Australian Genome Research Facility and [2]Institute of Molecular Bioscience, University of Queensland, St Lucia, Queensland 4072, Australia

Received October 21, 2003; Revised December 15, 2003; Accepted January 6, 2004

ABSTRACT

Despite the success of conventional Sanger sequencing, significant regions of many genomes still present major obstacles to sequencing. Here we propose a novel approach with the potential to alleviate a wide range of sequencing difficulties. The technique involves extracting target DNA sequence from variants generated by introduction of random mutations. The introduction of mutations does not destroy original sequence information, but distributes it amongst multiple variants. Some of these variants lack problematic features of the target and are more amenable to conventional sequencing. The technique has been successfully demonstrated with mutation levels up to an average 18% base substitution and has been used to read previously intractable poly(A), AT-rich and GC-rich motifs.

INTRODUCTION

Genome sequencing projects regularly encounter regions that yield no data with current sequencing strategies (1,2). These 'gaps' are present for several reasons. Some regions that cause gaps are unclonable or unstable in bacterial cells, and hence are under-represented in libraries. Sequencing using DNA polymerase-based extension products can be hindered by motifs that form secondary structures or other structural forms (3–6). These motifs are often GC-rich sequences with high thermal and structural stability (7–11), presumably because the high duplex melting temperature permits stable secondary structures to form, thus preventing completion of a sequencing reaction or causing band compressions in completed reactions. Similarly, AT-rich sequences (2,12) and other repetitive sequence motifs potentially allow extension from either aligned triplex strands or misaligned partially replicated duplex primed ends (13,14), thus preventing uniform sequencing. Other forms of simple repeat, homopolymer and stem–loop-forming or kinked DNA-forming regions (15,16) are known to limit or prevent entirely the procession of DNA polymerases. Indeed, Schlötterer and Tautz (17) reported that it is possible to synthesize all variant types of repetitious di- and trinucleotide simple sequence DNA motifs starting from short primers, a simple sequence template and a DNA polymerase *in vitro*.

Numerous methods that reduce the stability of duplex DNA have been used to overcome these problems. Some examples are: the inclusion of denaturing chemicals (18), sulfones (19) or dimethylsulfoxide (DMSO) (20); shearing of DNA into smaller pieces to disrupt the motif; and the introduction of non-mutagenic, strand-destabilizing nucleotide analogues into the sequencing reactions. Strand-destabilizing nucleotide analogues such as dITP (21,22), dUTP, 7-deaza-dGTP (10,11,23) and *N*4-methyl-2′-deoxycytidine 5′-triphosphate (24) have been widely used and some are now included in formulations of commercial sequencing kits. Improvement in sequencing can also be achieved in some refractory regions by using alternative sequencing enzymes (25) with modifications to cycling parameters, and by use of dye terminators instead of dye primers (26).

This paper describes a novel method with the potential to obtain sequence data from intractable regions. The method is called sequence analysis via mutagenesis (SAM) (27) because it involves generating and sequencing a number of mutated copies of the target DNA, then inferring the original sequence from the mutant sequences. The mutants must be altered sufficiently to no longer possess the characteristics that caused sequencing difficulties in the target. However, the mutants must also be sufficiently similar to the target to ensure that the original sequence can be inferred. Because the approach involves changing the target sequence, it can in theory address difficulties arising from any problematic sequence characteristic. Here we illustrate the technique by inferring the correct sequence of a problematic AT-rich motif from *Dictyostelium discoideum* and by sequencing a problematic GC-rich motif from the human genome. We also demonstrate generation of variants with random substitution of up to 18% of bases, and accurate recovery of original sequence from a surprisingly small number of such variants. Such high levels of mutation may be required to destroy problematic motifs in some applications. Finally, we use SAM to recover the sequence of an unclonable human mitochondrial gene.

MATERIALS AND METHODS

Mutagenic nucleotides

We used mutagenic dNTP analogues in PCRs (28) to generate the mutated libraries required for SAM. Nucleotide analogues

*To whom correspondence should be addressed. Tel: +61 7 3365 2309; Fax: +61 7 3365 1477; Email: j.keith1@mailbox.uq.edu.au

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

are compounds that mimic natural nucleotides in structural associations with natural nucleotides. The analogues are incorporated progressively during PCR at a relatively low frequency per cycle, and accumulate with increasing numbers of cycles. Although intensities of 5–10% mutation are readily achieved by several mutagenic analogues (29–32), few reach the 20–30% mutation intensities achieved by dPTP (28). Most mutagenic analogues tautomerize to establish equilibrium between amino and imino forms (30–32). The different tautomers can replace several different natural nucleotides rather than a single cognate nucleotide. This induces transition and transversion mutations in subsequent rounds of DNA replication when a novel native nucleotide, initially introduced opposite the nucleotide analogue, itself base-pairs with its natural cognate. The choice of mutagen for application to SAM depends upon base content of the problem sequence and the intensity of mutation required to remove the barrier to sequencing or cloning. Some analogues preferentially induce transition (or transversion) mutations of G:C to A:T (30,31); others principally induce mutations of A:T to G:C (28,29,32), whilst others are more indiscriminate, mutating all four bases to some extent (33–35). The tautomer ratio of some nucleotides can be influenced by pH, as can the efficiency of incorporation of both nucleotides and analogues by DNA polymerases, enabling additional control over the types and frequencies of the induced mutations (30,31). Manipulation of the DNA amplification conditions and the nucleotide analogue concentration also enables control over the intensity of mutation achieved when using high-intensity mutators (28–30).

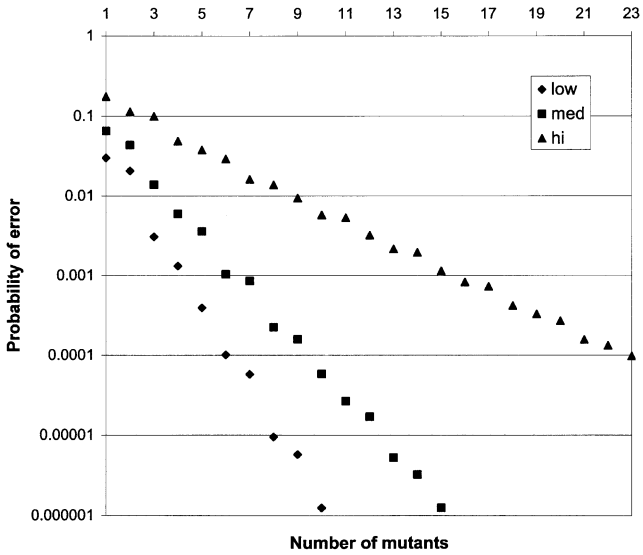

Figure 1. The error probability versus the number of mutants for three different intensities of mutation: low (3%), medium (7%) and high (18%). These graphs are used prior to sequencing to estimate the number of mutants required. The assumed substitution probabilities are shown in Supplementary tables 1–3, respectively.

Desalted primers

Enzymes and nucleotides were from Sigma Genosys, sequencing kits and PCR kits were from Applied Biosystems, and mutagenic nucleotide analogues were from Trilink (San Diego, CA). Cloning vectors pGEM-T EASY and pDrive were from Promega and Qiagen, respectively.

PCR amplifications involved 2 ng of DNA template, 1× AmpliTaq Gold buffer, 400 µM dNTPs, 2 mM $MgCl_2$, 0.4 µM each primer and 1 U of AmpliTaq Gold polymerase in 25 µl, supplemented with 50–400 µM of analogue nucleoside triphosphates and equimolar extra $MgCl_2$, essentially as described by Hill *et al.* (29). Amplifications were typically 30× cycles with 60 and 30 s periods of denaturation and annealing and extension for 10 min. Low and high mutation intensities were established by varying the cycling between 20× and 38×, respectively. In these experiments, A/T→G/C mutations were achieved in AT-rich motifs using the nucleotide analogues dPTP or 8-oxo-dGTP (28), whilst G/C→A/T mutations were achieved in GC-rich regions using the nucleotide 5-Br-dUTP (30) and a modified PCR buffer, which was supplemented with 20 mM glycine-KOH to elevate the pH to 8.6–8.8. Other mutagenic nucleotide analogues may also be used to achieve substitution mutations in recalcitrant DNA, but the above reagents were chosen for their particular mutagenic properties as well as the reasonable efficiency with which DNA polymerases may incorporate them into DNA.

DNA sequencing

Mutated targets were isolated as single plasmid clones using DNA Pure (Qiagen), then cycle sequenced with Applied BioSystems BigDye v 2.0, 3.0 and 3.1 sequencing kits using recommended conditions. Oligonucleotide primers for-20, rev-24, for-40 and rev-48 (catalogue nos S121s, S1201s, S1212s and S1233s; NE Biolabs) were used for PCR and cycle sequencing. Individual clones were analysed for mutation frequency. For sequence reconstruction, data from clone sets were analysed using SAM algorithms. Data sets are pre-screened to avoid sequence duplications.

SAM algorithms

The data analysis required by SAM consists of three components. The first is used prior to sequencing to estimate the number of mutants that will be required to reconstruct the original sequence with a specified level of accuracy. Such estimates may be obtained by plotting the expected proportion of errors against the number of mutants, under a simple model in which mutations are assumed to be independent, single-base substitutions and the four nucleotides are assumed to be present in approximately equal proportions. (These assumptions are relaxed for the second and third components.) The calculations depend on the probabilities of each type of substitution, which are specific to the mutagenesis protocol used and must be determined in separate experiments. We estimated substitution probabilities for the dPTP protocols used in this paper by aligning mutant sequences to known original sequences (see Supplementary tables 1–3, available at NAR Online). Figure 1 shows the dependence of the expected proportion of errors on the number of mutants for different concentrations of the nucleotide analogue dPTP. The number of mutants predicted by this method is best regarded as a lower bound on the number required to achieve the desired accuracy.

The second component is to infer the original sequence, once mutants have been generated and sequenced. A plausible

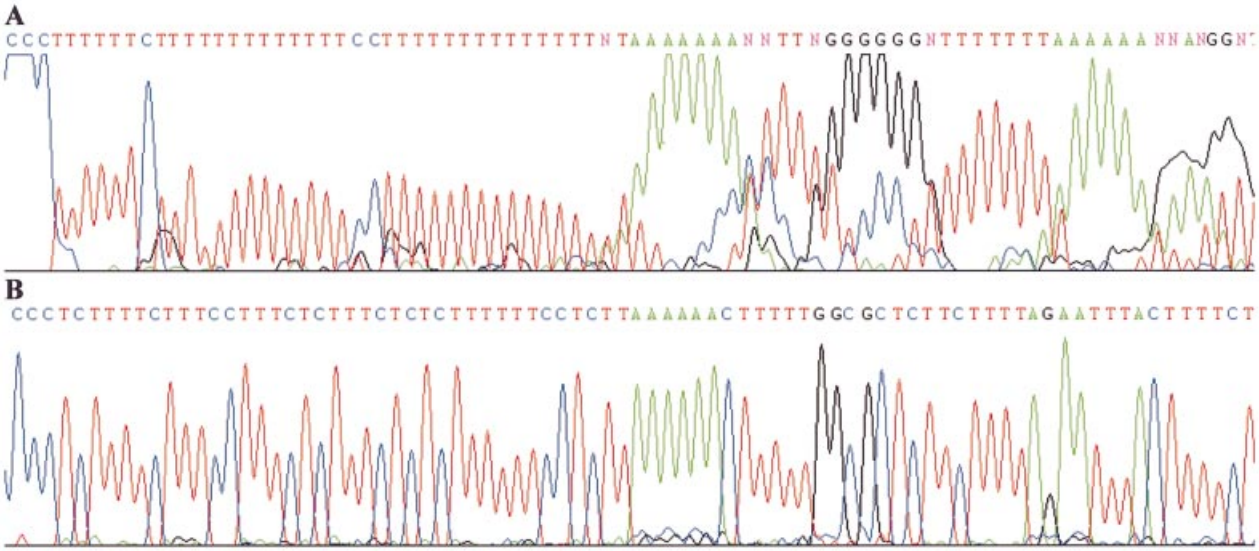

Figure 2. Sequence chromatograms of a *D.discoideum* shotgun clone (JC1a86h11) containing a homopolymer tract sequenced with BigDye v2.0 and M13-21 universal primer. (**A**) The sequence of the wild-type plasmid DNA showing the consequences of polymerase slippage within the homopolymer and resulting harmonic stutter peaks in the trace. (**B**) Introducing 12% random substitutions using dPTP reduced the uniformity of the problem motifs. The mutated variant of JC1a86h11 can then be readily sequenced.

approach is to form a multiple sequence alignment of the mutant sequences and determine a consensus character for each column of the alignment. However, this approach does not make best use of knowledge regarding the processes by which mutants were generated. We therefore developed a new algorithm (36–38) that incorporates a probabilistic model of these processes, and does not involve multiple sequence alignment. The algorithm uses the stochastic optimization technique simulated annealing (39) to search for the most probable original sequence with respect to a Bayesian probabilistic model (38). The model is more sophisticated than that used to predict the required number of mutants, and allows for insertions, deletions and non-uniform base composition. It also allows for different probabilities of mutation for each nucleotide, and different probabilities of the three possible substitutions for each of the four original nucleotides. These probabilities are estimated as discussed in the preceding paragraph. We found that sequences inferred using the new algorithm were significantly more accurate than those obtained using a popular multiple sequence alignment package (40). The algorithm produces highly accurate sequences using a small number of mutants, even with very high levels of mutation.

The third component is to assign each base in the inferred sequence a quality value, consistent with quality values generated by base-calling programs such as Phred (41). Computation of quality values involves generating sequences that differ from the inferred sequence at a given position and evaluating their respective posterior probabilities according to the model. We verified in computer simulations that these quality values are a good indication of the actual probability of error (see Supplementary fig. 1). If the quality values indicate a higher proportion of errors than was originally predicted, it may be necessary to sequence additional mutants.

RESULTS

Sequencing of a problematic AT-rich DNA

Dictyostelium discoideum has an AT-rich genome (~78% AT), and contains regions that are recalcitrant to conventional sequencing techniques (2). We mutated recalcitrant elements at several intensities to reduce AT content, and the mutants were then sequenced using conventional dye terminator chemistry. Figure 2A shows the sequence trace from a wild-type *Dictyostelium* 'unsequenceable' fragment: harmonic stutter from polymerase slippage obliterates the normal chromatograph pattern. Figure 2B shows the trace of a mutated fragment with ~12% base substitutions. Elimination of stutter peaks creates a well-defined chromatogram with uniform peak shape and good separations.

Sequencing of a problematic GC-rich DNA

We mutated recalcitrant GC-rich human genomic elements using 5-Br-dUTP to reduce the GC content, and the mutants were then sequenced using conventional dye terminator chemistry. Figure 3A shows the sequence trace from a wild-type human 'unsequenceable' fragment: here presumed polymerase blockage obliterates the normal chromatograph pattern. Figure 3B shows the trace of a mutated fragment with ~11% base substitutions, now displaying uniform peaks and good peak separations.

Inferring an original sequence

Figure 4 shows several reconstructions of the unsequenceable *Dictyostelium* fragment shown in Figure 2. Two different reconstruction algorithms were used: simulated annealing consensus (36) (SAC) and the Bayesian approach (38) (Bay). Sequences were inferred from four low-intensity (3%) mutants (4×Low), an independent set of six medium-intensity (7%) mutants (6×Med) and the collection of all 10 mutants

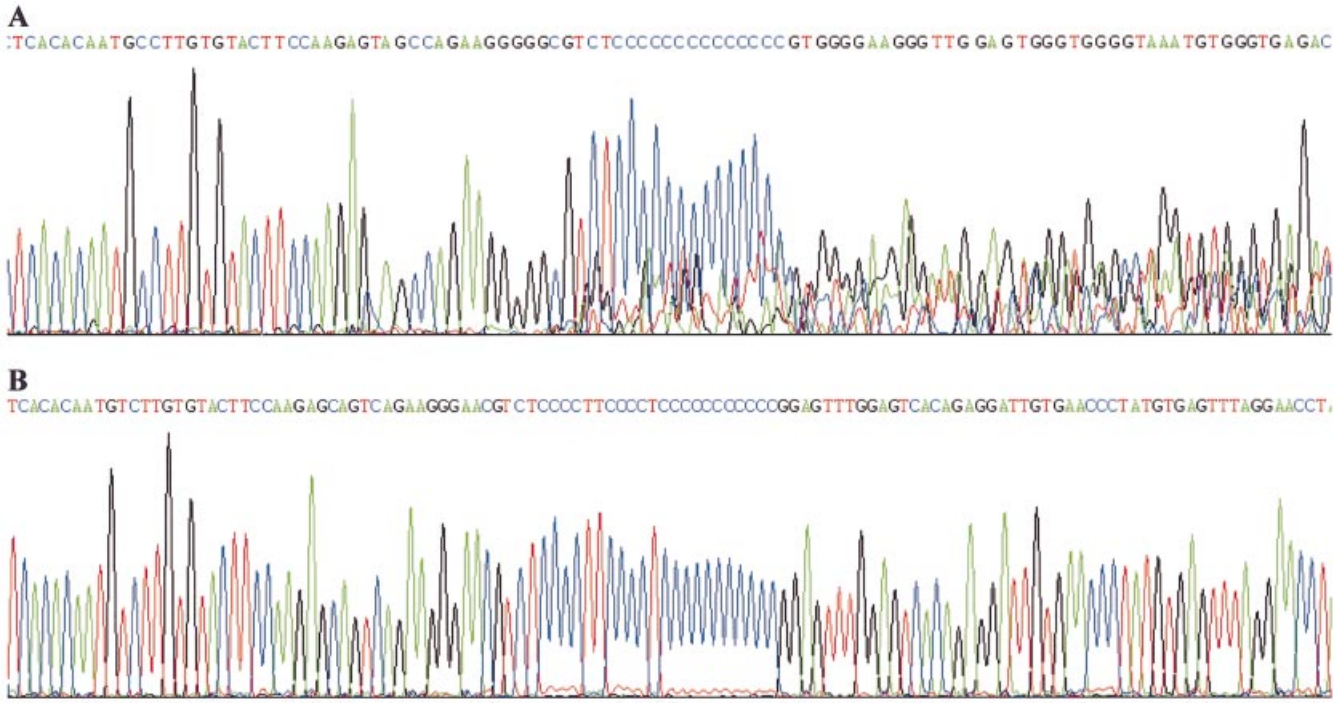

Figure 3. Sequence chromatograms of a human clone (A9A05J2) containing a GC-rich tract sequenced with BigDye v3.1 and M13-21 universal primer on a ABI 3730xl capillary sequencer. (**A**) The sequence of the wild-type plasmid DNA showing the consequences of polymerase slippage within the homopolymer and resulting harmonic stutter peaks in the trace. (**B**) Introducing 11% random substitutions using 5-Br-dUTP reduced the uniformity of the problem motifs. The mutated variant of A9A05J2 can then be readily sequenced.

(10×All). The expected proportion of errors (see Fig. 1) using the low- and medium-intensity data sets is ~1 per 1000 bases in both cases. The most accurate reconstruction should be that obtained using Bay and the full set of 10 mutant sequences. Note that this sequence is identical to the reconstruction obtained using the 6×Med mutant set, which suggests that the additional 4xLow sequences merely contribute to the confidence that this is the correct original sequence. This is borne out by the quality values for the three Bayesian reconstructions, which are given in Supplementary figure 2. Here, a quality value of x indicates a $10^{-x/10}$ probability that the base has been miscalled. Note that the quality of the 10xAll sequence is high everywhere except in one poly(T) region, where one or more bases might have been omitted. A high proportion of insertions and deletions is observed in this region of the mutants, due perhaps to replication slippage during PCR mutagenesis. Additional mutants, possibly generated using different mutagenesis protocols, may be needed to improve quality values in this region.

SAM reconstruction at high mutation intensities

To test whether SAM correctly reconstructs an original sequence, we applied the technique to non-problematic DNA with known sequence. Fourteen mutants were generated with an average of 18% substitution (the proportion of substitutions ranged from 5 to 28%). The original sequence was reconstructed by aligning the sequences using ClustalW (40) and determining the consensus (i.e. most frequent) character for each column (Fig. 5A). A second reconstruction was performed using the Bayesian approach. The sequence reconstructed using ClustalW differs in 10 places from the original sequence. In eight of these, there is no clear consensus and an 'n' is shown. The Bayesian approach correctly identifies these bases but contains one error where an 'A' has been called as a 'G'. Figure 5B shows the quality values assigned to each base of the Bayesian inferred sequence. The miscalled base has one of the lowest quality values at 13. Some of the other quality values are also low, but the corresponding bases are called correctly. The miscalled base does not necessarily reflect a mutation 'hotspot'; that a single error is observed using such high substitution probabilities is unsurprising given that the expected proportion of errors in reconstruction is ~0.002 (see triangular markers in Fig. 1). If the experiment were repeated many times, the correct reconstruction would be expected in only ~60% of the repeats. Better accuracy would be expected for a larger number of mutants.

Unclonable mitochondrial DNA

SAM can be used to overcome an important practical impediment to sequencing: missing or unclonable DNA fragments. To illustrate this, we used SAM to sequence human mitochondrial $tRNA^{Thr}$, which is unclonable in *Escherichia coli*. Mita *et al.* (42) found that mutations within two 5 and 6 bp 'hotspots' of the cloverleaf structure rendered the DNA clonable. Presumably, these regions are implicated in the unclonability of the original DNA, as the original sequence was altered in all cloned sequences. We sequenced 16 mutants, all of which were found to have one or more mutations in these hotspots (Fig. 6). Regions flanking the

```
SAC 4xLow:      .....................-.........................
SAC 6xMed:      .....................-.........................
SAC 10xAll:     ..................................................
Bay 4xLow:      ..................................................
Bay 6xMed:      ..................................................
Bay 10xAll:     CCCTTTTTTCTTTTTTTTTTTTTTCTTTTTTTTTTTTTTTTTAAAAAACT

SAC 4xLow:      ..................................................
SAC 6xMed:      ..................................................
SAC 10xAll:     ..................................................
Bay 4xLow:      ..................................................
Bay 6xMed:      ..................................................
Bay 10xAll:     TTTTGGCGTTTTTTTTTTAAAATTTATTTTTTTTTTTTATTTTTTTTTT

SAC 4xLow:      ....-.............................................
SAC 6xMed:      ...C..............................................
SAC 10xAll:     ....-.............................................
Bay 4xLow:      ....-.............................................
Bay 6xMed:      ..................................................
Bay 10xAll:     TTTTTCAAAAACTTTTTTTTCAACACACAGTATTGCAATTACTAGCAGGC

SAC 4xLow:      ..................................................
SAC 6xMed:      ..................................................
SAC 10xAll:     ..................................................
Bay 4xLow:      ..................................................
Bay 6xMed:      ..................................................
Bay 10xAll:     AAAAACGTTAACGAAAATCAAAAATAAAAGTACTTATAAGTACCGAGCTC
```

Figure 4. Inferred original sequence of the *Dictyostelium* fragment JC1a86h11 obtained using simulated annealing consensus (SAC) and probabilistic Bayesian (Bay) approaches. Sequences were inferred using four mutants with 3% mutation intensity (4×Low), six mutant sequences with 7% mutation intensity (6×Med), and the collection of all 10 mutant sequences (10×All). Bases marked with a period are identical to the base at the bottom of that column. Mutations were induced using dPTP at low (4×Low) or medium (6×Med) concentration.

hotspots had a mutation intensity of 0.7%, whereas the hotspots had ~12% mutations. The Bayesian approach (38) was used to reconstruct the original sequence (see Fig. 6A). The reconstruction agreed with the published gene sequence, except at one base. All quality values were 99 (the maximum) except at the miscalled base, which has a low quality value of 14. A large number of A to G mutations occurred at the miscalled base. This is an unusual substitution for the nucleotide analogue used to generate mutants 1-1 to 1-13 [8-oxo-dGTP (28)], which suggests that this mutation significantly improves clonability. The large number of mutants used in this analysis was unnecessary. In fact, high-quality sequence could be achieved with fewer mutants. As an example, the quality values achieved using the first six mutants are shown in Figure 5B.

DISCUSSION

The nucleotide analogues mentioned in the Introduction have previously been used to reduce the thermal stability of duplex DNA within a refractory sequence, and to achieve a more uniform distribution of local thermal stability across the entire DNA region. However, these methods are either non-mutagenic or do not introduce mutations at a sufficient intensity to substantially alter the sequence-related structural characteristics of these regions and are, consequently, often inefficient. For example, the analogue 7-deaza-dGTP is not reported to be mutagenic to DNA (9,13,43). Similarly, although dITP is used in error-prone PCR for randomly mutating genes by altering the concentrations of the respective dNTPs (44,45), it induces only low-level (~4×10^{-3}) mutations after strong incorporation with elevated cycles of PCR amplification (45).

Direct sequencing of PCR products is commonly used to resolve examples of unclonable DNA regions, and one important application of SAM occurs when classical PCR fails to amplify a bridging PCR product, or the product is also unsequenceable. Other examples where SAM might be expected to display advantages are in unclonable regions flanked by or containing repeated motifs, which may preclude identification of unique priming sites, or in unclonable regions that are larger than the range of readily PCR-amplifiable fragments.

The SAM process selects for mutants that are clonable and sequenceable, and this can produce mutation 'hotspots' in the regions responsible for the problems. If the hotspot is restricted to only one or two bases, as it apparently is in our final example, our algorithms for inferring the original

A

```
Clustal W   ............................n............................n
Bayesian    ..............................................................
Original    GGCGTAATAATACTATTTGTTGTGTCAATTTTCTTGGTTCCTGACTAAAACATTAAGGTT

Clustal W   ....................................n.....................
Bayesian    ..............................................................
Original    TCTCAGTTAAGCTATATACGATAAATATTGGCATCTTTCTATTGCAGGATGATTTCTAGT

Clustal W   ............................n..........n..G................
Bayesian    ..............................................................
Original    GCTAAGCATTATAGCCAGGAGTAAAGGAAATAACGCTTTAACGATACCACCATTAATTTA

Clustal W   ..n.n.....................nG.............................
Bayesian    ............................G.............................
Original    AAAAATGGAGTCTGAAATGGAAAAAGAAGAAAAAAGCAATCTCATCTACGATAAAGATCC
```

B

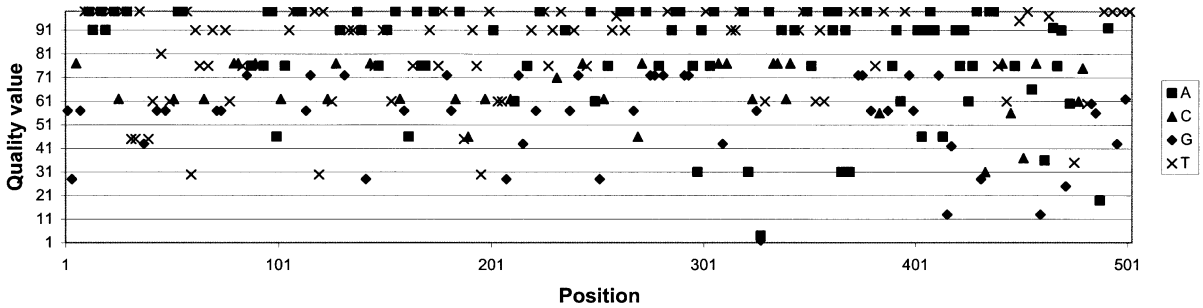

Figure 5. SAM reconstruction of a known sequence (pTEST), using 14 mutant copies of the sequence. The dPTP-induced mutants were found to differ from the original sequence on average in ~18% of bases. (**A**) Inferred original sequence based on alignment using ClustalW (ClustalW) and the Bayesian approach (Bayesian). The known original sequence (Original) is also shown. Bases marked with a period are identical to the base at the bottom of that column. (**B**) Quality values (vertical axis) for the Bayesian reconstruction. Quality values are assigned to each base and to the hypotheses that there are no additional bases between each pair of adjacent characters, or at the ends of the inferred sequence. On the horizontal axis, odd numbers represent positions between characters and at the ends of the sequence, whereas even numbers represent base positions. The same convention is used in Figure 6B. Quality values for odd-numbered positions are not shown; all are 99.

sequence may fail at those bases (although correctly reconstructing flanking sequence). In that example, the possibly miscalled base is easily detected by noting the low quality score and the implied frequency of an unusual substitution (G to A). In general, the quality scores should be compared with the expected error; if they are substantially lower, a putative mutation hotspot is indicated. In this case, the expected proportion of errors is minute, so a quality value of 14 is a sure indication of a hotspot. When only the first six mutants were used (see Fig. 6B), hotspots were less obvious. Nevertheless, there are two positions with comparatively low quality scores and unusual substitutions for the analogue used (A to G and G to A). This indicates two possible hotspots, one of which corresponds to the miscalled base observed with the larger data set. In future applications, automated detection of hotspots should be possible using statistical tests based on the expected distribution of quality values, discrepancies between sequences reconstructed using different mutagenesis protocols, and the occurrence of unusual substitutions. The ability to detect mutation hotspots is important because many genomic sequences obtained from conventional cloned libraries potentially contain mutant hotspots that are currently undetected (46). Sequencing of additional mutants, or alternative methods such as SNP analysis, might be used to determine residual 1–2 bp inaccuracies. There may also be advantages in analysing data from two or more independent sets of mutants, each generated using a different mutagen. In regions where one set of mutants is extensively modified, other sets may reproduce the original sequence more accurately, and thus the various sets provide complementary information.

The SAM algorithms do not reconstruct the original sequence with certainty. Indeed, this is not possible; one can only estimate the probability that a candidate sequence was the original. However, the same could be said of all existing sequencing methods, since all methods are subject to error. The uncertainties caused by the introduction of mutations are no more problematic than uncertainties caused by other types of random error, and can be reduced in the same way: by additional sequencing. The number of mutants required to achieve any desired level of accuracy can be estimated in advance using graphs such as that shown in Figure 1. If, after reconstruction, the level of accuracy indicated by the quality values is lower than that desired, additional mutants can be generated and the analysis repeated.

The results presented here and elsewhere (36–38) indicate that our reconstruction algorithms are delivering highly accurate sequence (see also Supplementary fig. 3) and that the Bayesian approach is superior to simulated annealing consensus, which is superior to the alignment approach (37).

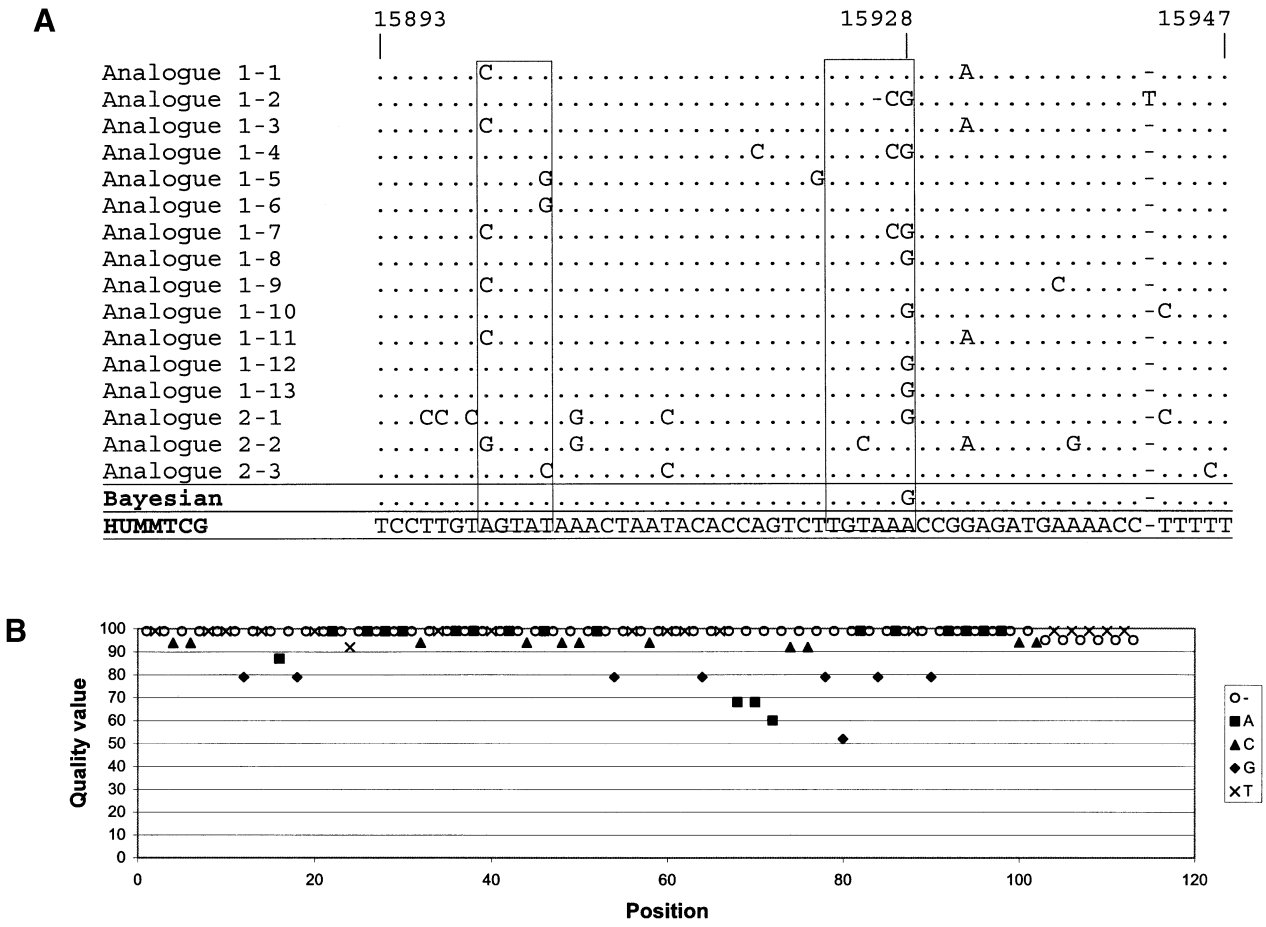

Figure 6. Alignment of DNA sequences of 16 individual clones of the 'unclonable' human mitochondrial $tRNA^{Thr}$ gene (42) to the inferred original sequence (Bayesian). (**A**) The putative 'mutation hotspots' necessary for clone stability in *E.coli* are outlined (large boxes). Thirteen mutants (1-1–1-13) were generated using 8-oxo-dGTP (24) and three mutants (2-1–2-3) were generated using dPTP at a high concentration. The inferred sequence agreed with known mitochondrial gene sequence (accession no. HUMMTCG) across both the bulk (0.7% mutated) and hotspot (12% mutated) regions except in one base. Bases marked with a period are identical to the base at the bottom of that column. (**B**) Quality values for the Bayesian reconstruction using the first six mutants only from (A). The inferred sequence is correct.

Nevertheless, several improvements may be possible. For example, the algorithms do not currently use quality values obtained for the mutant sequences using base-calling programs such as Phred (41). Such information could be used to reduce the undesirable effects of sequencing errors. At present, the algorithms ignore sequencing errors and effectively treat them as mutations. The mutation model could be improved. In laboratory work, we have observed mutations that do not conform to the model. Specifically, a small amount of replication slippage sometimes occurs during mutagenic PCR of homopolymer regions, resulting in context-dependent probabilities of insertions and deletions. Our model could in principle be generalized to account for these and other types of mutation. However, this may be unnecessary, as the current model seems to produce accurate inferences even where slippage has occurred.

The reconstruction method assumes that the mutants were generated independently. In practice, it is efficient to generate mutants via a single PCR in the presence of nucleotide analogues. This may result in some degree of dependence among the mutants, since two or more mutants may have a common ancestor generated at some cycle of the PCR. However, this is unlikely to be a major source of error, since the number of variants subjected to each cycle is large, except during early cycles. Thus mutants are only likely to have common ancestors generated in early PCR cycles, when all variants were similar to the original. Nevertheless, such dependencies can potentially be taken into account via phylogenetic methods. Some existing phylogenetic algorithms could be used for this purpose. However, one concern with most existing phylogenetic methods is that they depend on an initial multiple sequence alignment, and this may introduce bias if there is uncertainty about how to align the mutants. Often the problematic sequences for which SAM is intended are repetitive, and there are likely to be uncertainties about how to align mutants of such sequences. The authors are currently developing phylogenetic methods specialized for SAM reconstruction.

Other approaches for sequencing gaps have been reported recently. For example, PACE (47) is a method for extension of contig ends, although it is restricted by the limits of the PCR. Whilst such advances are useful for extracting maximum

information from available data, they do not address some of the fundamental causes of sequencing problems. These causes can potentially be eliminated using the SAM technique.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The following individuals and companies kindly provided reagents: Gernot Glöckner provided *Dictyostelium* clones; Bruce Roe provided human genomic clones; and David Loakes and Trilink Biotechnologies Inc. each provided analogue nucleotides. We thank the Centre for Advanced Mathematics and Computing, University of Queensland for access to its Sun computing cluster. D.B. is an ARC QEII fellow in combinatorial mathematics. K.M., D.B. and P.A. are supported by ARC discovery grant DP0208534, Biotechnology Innovation Fund grant no. BIF02620 and a development grant from Uniseed.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al*. The International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Glöckner,G., Eichinger,L., Szafranski,K., Pachebat,J.A., Bankier,A.T., Dear,P.H., Lehmann,R., Baumgart,C., Parra,G., Abril,J.F. *et al*. (2002) Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature*, **418**, 79–85.
3. Ji,J., Clegg,N.J., Peterson,K.R., Jackson,A.L., Laird,C.D. and Loeb,L.A. (1996) *In vitro* expansion of GGC:GCC repeats: identification of the preferred strand of expansion. *Nucleic Acids Res.*, **24**, 2835–2840.
4. Tabor,S. and Richardson,C.C. (1987) DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl Acad. Sci. USA*, **84**, 4767–4771.
5. Donlin,M.J. and Johnson,K.A. (1994) Mutants affecting nucleotide recognition by T7 DNA polymerase. *Biochemistry*, **33**, 14908–14917.
6. Weinshenker,B.G., Hebrink,D.D., Gacy,A.M. and McMurray,C.T. (1998) DNA compression caused by an upstream point mutation. *Biotechniques*, **25**, 68–72.
7. Mizusawa,S. Nishimura,S. and Seela,F. (1986) Improvement of the dideoxy chain termination method of DNA sequencing by use of deoxy-7-deazaguanosine triphosphate in place of dGTP. *Nucleic Acids Res.*, **14**, 1319–1324.
8. McConlogue,L., Brow,M.A. and Innis,M.A. (1988) Structure-independent DNA amplification by PCR using 7-deaza-2′-deoxyguanosine. *Nucleic Acids Res.*, **16**, 9869.
9. Haqqi,T.M., Sarkar,G. David,C.S. and Sommer,S.S. (1988) Specific amplification with PCR of a refractory segment of genomic DNA. *Nucleic Acids Res.*, **16**, 11844.
10. Fernandez-Rachubinski,F., Eng,B., Murray,W.W., Blajchman,M.A. and Rachubinski,R.A. (1990) Incorporation of 7-deaza dGTP during the amplification step in the polymerase chain reaction procedure improves subsequent DNA sequencing, *DNA Seq.*, **1**, 137–140.
11. Motz,M., Paabo,S. and Kilger,C. (2000) Improved cycle sequencing of GC-rich templates by a combination of nucleotide analogs. *Biotechniques*, **29**, 268–270.
12. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S., *et al*. (2002) Genome sequence of the human malaria parasite. *Plasmodium falciparum*. *Nature*, **419**, 498–511.
13. Baran,N., Lapidot,A. and Manor,H. (1991) Formation of DNA triplexes accounts for arrests of DNA synthesis at d(TC)n and d(GA)n tracts. *Proc. Natl Acad. Sci. USA*, **88**, 507–511.
14. Thoraval,D., Asakawa,J., Kodaira,M., Chang,C., Radany,E., Kuick,R., Lamb,B., Richardson,B., Neel,J.V. Glover,T. and Hanash,S. (1996) A methylated human 9-kb repetitive sequence on acrocentric chromosomes is homologous to a subtelomeric repeat in chimpanzees. *Proc. Natl Acad. Sci. USA*, **93**, 4442–4447.
15. Razin,S.V., Ioudinkova,E.S., Trifonov,E.N. and Scherrer,K. (2001) Non-clonability correlates with genomic instability: a case study of a unique DNA region. *J. Mol. Biol.*, **307**, 481–486.
16. Kurahashi,H., Shaikh,T.H., Hu,P., Roe,B.A., Emanuel,B.S. and Budarf,M.L. (2000) Regions of genomic instability on 22q11 and 11q23 as the etiology for the recurrent constitutional t(11;22). *Hum. Mol. Genet.*, **9**, 1665–1670.
17. Schlotterer,C. and Tautz,D. (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.*, **20**, 211–215.
18. Varadaraj,K. and Skinner,D.M. (1994) Denaturants or cosolvents improve the specificity of PCR amplification of a G + C-rich DNA using genetically engineered DNA polymerases. *Gene*, **140**, 1–5.
19. Chakrabarti,R. and Schutt,C.E. (2002) Novel sulfoxides facilitate GC-rich template amplification. *Biotechniques*, **32**, 866, 868, 870–872, 874.
20. Seto,D., Seto,J., Deshpande,P. and Hood,L. (1995) DMSO resolves certain compressions and signal dropouts in fluorescent dye labeled primer-based DNA sequencing reactions. *DNA Seq.*, **5**, 131–140.
21. Kawase,Y., Iwai,S., Inoue,H., Miura,K. and Ohtsuka,E. (1986) Studies on nucleic acid interactions. I. Stabilities of mini-duplexes (dG2A4XA4G2-dC2T4YT4C2) and self-complementary d(GGGAAXYTTCCC) containing deoxyinosine and other mismatched bases. *Nucleic Acids Res.*, **14**, 7727–7736.
22. Bergstrom,D.E., Zhang,P. and Johnson,W.T. (1997) Comparison of the base pairing properties of a series of nitroazole nucleobase analogs in the oligodeoxyribonucleotide sequence 5′-d(CGCXAATTYGCG)-3′. *Nucleic Acids Res.*, **25**, 1935–1942.
23. Dierick,H., Stul,M., De Kelver,W., Marynen,P. and Cassiman,J.J. (1993) Incorporation of dITP or 7-deaza dGTP during PCR improves sequencing of the product. *Nucleic Acids Res.*, **21**, 4427–4428.
24. Li S., Haces,A., Stupar,L., Gebeyehu,G. and Pless,R.C. (1993) Elimination of band compression in sequencing gels by the use of *N*4-methyl-2′-deoxycytidine 5′-triphosphate. *Nucleic Acids Res.*, **21**, 2709–2714.
25. Kukanskis,K.A., Siddiquee,Z., Shohet,R.V. and Garner,H.R. (2000) Mix of sequencing technologies for sequence closure: an example. *Biotechniques*, **28**, 630–632, 634.
26. Robbins,C.M., Hsu,E. and Gillevet,P.M. (1996) Sequencing homopolymer tracts and repetitive elements. *Biotechniques*, **20**, 862–864, 866–868.
27. Keith,J.M., Adams,P. and Bryant,D. (2001) Method for sequence analysis. US Patent Application PCT/AU02/00397. Priority Date 28th March 2001.
28. Zaccolo,M., Williams,D.M., Brown,D.M. and Gherardi,E.E. (1996) An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J. Mol. Biol.*, **255**, 589–603.
29. Hill,F., Williams,D.M., Loakes,D. and Brown,D.M. (1998) Comparative mutagenicities of N6-methoxy-2,6-diaminopurine (dK) and N6-methoxyaminopurine 2′-deoxyribonucleotides (dZ) and their 5′-triphosphates. *Nucleic Acids Res.*, **26**, 1144–1149.
30. Yu,H., Eritja,R., Bloom,L.B. and Goodman,M.F. (1993) Ionization of bromouracil and fluorouracil stimulates base mispairing frequencies with guanine. *J. Biol. Chem.*, **268**, 15935–15943.
31. Suen,W., Spiro,T.G., Sowers,L.C. and Fresco,J.R. (1999) Identification by UV resonance Raman spectroscopy of an imino tautomer of 5-hydroxy-2′-deoxycytidine, a powerful base analog transition mutagen with a much higher unfavored tautomer frequency than that of the natural residue 2′-deoxycytidine. *Proc. Natl Acad. Sci. USA*, **96**, 4500–4505.
32. Schuerman,G.S., Van Meervelt,L., Loakes,D., Brown,D.M., Kong,T., Lin,P., Moore,M.H. and Salisbury,S.A. (1998) A thymine-like base analogue forms wobble pairs with adenine in a Z-DNA duplex. *J. Mol. Biol.*, **282**, 1005–1011.
33. Loakes,D. and Brown,D.M. (1994) 5-Nitroindole as a universal base analogue. *Nucleic Acids Res.*, **22**, 4039–4043.
34. Loakes,D. (2001) The applications of universal DNA base analogues. *Nucleic Acids Res.*, **29**, 2437–2447.
35. Seela,F. and Debelak,H. (2000) The *N*(8)-(2′-deoxyribofuranoside) of 8-aza-7-deazaadenine: a universal nucleoside forming specific hydrogen bonds with the four canonical DNA constituents. *Nucleic Acids Res.*, **28**, 3224–3232.

36. Keith,J.M., Adams,P., Bryant,D., Kroese,D.P., Mitchelson,K.R., Cochran,D.A.E. and Lala,G.L. (2002) A simulated annealing algorithm for finding a consensus sequence. *Bioinformatics*, **18**, 1494–1499.
37. Keith,J.M., Adams,P., Bryant,D., Mitchelson,K.R., Cochran,D.A.E. and Lala,G.L. (2003) Inferring an original sequence from erroneous copies: two approaches. *Asia-Pacific BioTech News*, **7**, 107–114.
38. Keith,J.M., Adams,P., Bryant,D., Mitchelson,K.R., Cochran,D.A.E. and Lala,G.L. (2003) Inferring an original sequence from erroneous copies: a Bayesian approach. In Chen,Y.-P.P. (ed.), *Proceedings of the 1st Asia-Pacific Bioinformatics Conference* (APBC2003), vol. 19, pp. 23–28.
39. Kirkpatrick,S., Gelatt,C.D. and Vecchi,M.P. (1982) Optimization by simulated annealing. *Science*, **220**, 671–680.
40. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
41. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
42. Mita,S., Monnat,R.J.,Jr and Loeb,L.A. (1988) Direct selection of mutations in the human mitochondrial $tRNA^{Thr}$ gene: reversion of an 'uncloneable' phenotype. *Mutat. Res.*, **199**, 183–190.
43. McClary J., Ye,S.Y., Hong,G.F. and Witney,F. (1991) Sequencing with the large fragment of DNA polymerase I from *Bacillus stearothermophilus*. *DNA Seq.*, **1**, 173–180.
44. Leung,S. and Miyamoto,N.G. (1989) Point mutational analysis of the human c-fos serum response factor binding site. *Nucleic Acids Res.*, **17**, 1177–1195.
45. Spee,J.H., de Vos,W.M. and Kuipers,O.P. (1993) Efficient random mutagenesis method with adjustable mutation frequency by use of PCR and dITP. *Nucleic Acids Res.*, **21**, 777–778.
46. Harris,D.J. (2003) Can you bank on GenBank? *Trends Ecol. Evol.*, **18**, 317–319.
47. Carraro,D.M., Camargo,A.A., Salim,A.C., Grivet,M., Vasconcelos,A.T. and Simpson,A.J. (2003) PCR-assisted contig extension: stepwise strategy for bacterial genome closure. *Biotechniques*, **34**, 626–628, 630–632.