An estimate of large-scale sequencing accuracy

Fergal Hill†, Christine Gemünd, Vladimir Benes, Wilhelm Ansorge and Toby J. Gibson

European Molecular Biology Laboratory, Meyerhofstraße 1, Heidelberg D-69117, Germany

Received April 27, 2000; revised May 23, 2000; accepted May 25, 2000

The accuracy of large-scale DNA sequencing is difficult to estimate without redundant effort. We have found that the mobile genetic element IS\textsuperscript{10}, a component of the transposon Tn10, has contaminated a significant number of clones in the public databases, as a result of the use of the transposon in bacterial cloning strain construction. These contaminations need to be annotated as such. More positively, by defining the range of sequence variation in IS\textsuperscript{10}, we have been able to determine that the rate of sequencing errors is very low, most likely surpassing the stated aim of one error or less in ten thousand bases.

INTRODUCTION

How accurate is the DNA sequence produced by the Genome projects? The aim, set out at the First International Strategy Meeting on Human Genome Sequencing, was to have one error or less in every 10 000 bases of finished sequence (Bentley, 1996). Checking such accuracy would require substantial redundant sequencing (Beck, 1993). Paradoxically, the contamination of genomic sequences by the bacterial mobile element IS\textsuperscript{10}, a cloning artifact described here, shows that the accuracy of large-scale sequencing surpasses the stated aim.

How does this cloning artifact arise? The tetracycline resistance transposon, Tn10, has been widely used in the construction of bacterial strains (Kleckner et al., 1977), including the common bacterial artificial chromosome (BAC) hosts DH10B, and its bacteriophage-resistant derivative, HS996 (Grant et al., 1990). Tn10 contains inverted repeats, designated IS\textsuperscript{10}, of 1329 base pairs at each end (for a review of Tn10, see Kleckner et al., 1996). The repeats differ from each other at 16 positions (Chalmers et al., 2000). Each IS\textsuperscript{10} copy can transpose independently of the other. The tetracycline resistance marker can conveniently be removed after strain construction (Bochner et al., 1980; Maloy and Nunn, 1981), as it was in the case of DH10B (Grant et al., 1990). Nevertheless, one or more IS\textsuperscript{10} elements may be left behind (Ross et al., 1979; Shen et al., 1987). These can themselves transpose at a frequency of \(\sim 10^{-4}\) per cell per bacterial generation (Shen et al., 1987).

RESULTS AND DISCUSSION

A BLAST search (Altschul et al., 1992) of the non-redundant NCBI database (\textit{nr} database at \url{http://www.ncbi.nlm.nih.gov/blast/}) using the IS\textit{10}R prototype as a query sequence (Halling et al., 1982) uncovered 28 genome project clones that were contaminated by IS\textit{10} during their propagation in Escherichia coli. Twenty-four were in human sequences, three in Arabidopsis thaliana clones and two were in the same Caenorhabditis elegans database entry (DDBJ/EMBL/GenBank accession No. AC006650; this entry contains an internal deletion within Tn10 that truncates the tetracycline resistance gene).

These 29 IS\textit{10} copies were aligned (summarized in Table I); 18 perfect copies of IS\textit{10}R were found, but only one copy of IS\textit{10}L. The right copy, IS\textit{10}R, is known to be more than 10 times as mobile as the left, IS\textit{10}L (Foster et al., 1981). The remaining 10 IS elements were distinct from both IS\textit{10}L and IS\textit{10}R, but appear to be hybrids of both. Such hybrid elements have been described before (Davis, 1986; Bogosian et al., 1993) and are most likely the result of recombination. Davis (1986) favoured gene conversion as the most likely cause, but the manner in which they are formed is unknown. Some hybrids have higher transposition rates (Davis, 1986), which might contribute to their rather striking frequency. We scored either of the bases found in 38 541 base pairs of sequence.

Aside from these variations, no alterations attributable to sequencing errors were found in any of the insertions, comprising in total 38 541 base pairs (29 × 1329). Using the Poisson distribution, we can estimate, for a given error rate, the probability of finding no errors in 38 541 base pairs of sequence. If the error rate is one mistake in 10 000 bases, there is only a 1 in 50 chance (2.1%) of finding no errors in 38 541 bases.

†Corresponding author. Tel: +49 6221 387474; Fax: +49 6221 387306; E-mail: Fergal.Hill@embl-heidelberg.de

Present address: Avidis S.A., Saint Beauzire, 63730 France
It is well recognized that some DNA templates are easier to sequence accurately than others. To serve as a reference standard, IS10 should be neither unusually difficult nor easy to sequence. To assess this, the error frequency in 40 single-pass sequencing of IS10 was determined, and found to be 3.1%. In addition, all other eukaryotic sequences containing full-length IS10 insertions were examined for errors. Five insertions were found. No errors were found in the three entries originating from large-scale cDNA sequencing projects (DDBJ/EMBL/GenBank accession Nos AF181652, AL117609 and AK001627), but four errors were found in the remaining pair (G for C at position 151 in AF199339; two deletions of a single C residue at positions 1 and 1012 and substitution of A for T at position 20 in AJ001004). We conclude that accurate positioning of characteristic 9 bp duplications flanking each insertion in the IS10 elements is not trivial.

Further contamination of the databases by IS10 is avoidable, since the vast majority of the clones to be sequenced have already been prepared. Indeed, a search of genomic DNA sequences whose release is pending (but unfinished) has revealed >50 IS10 sequences in the Drosophila melanogaster genome (available from http://edgp.ebi.ac.uk/www-blast.html using ‘All Drosophila nucleic’ as database), and more than three times this number, so far, in human sequences (http://www.sanger.ac.uk/HGP/blast_server.shtml using ‘unfinished human genomic sequence’ as database). In itself, this poses few problems if such insertions are automatically annotated, as genuine repeated elements are. However, only a quarter of the NCBI genomic IS10 elements that we found were annotated in any way, usually as Tn10. Worse are the occasional misleading annotations pointing out the similarity of the insertion sequence to cDNAs or ESTs that also contained IS10.

The presence of IS10 in genomic clones creates the risk of further rearrangements, such as inversions or deletions, induced by the element (Shen et al., 1987). We confirmed the presence of characteristic 9 bp duplications flanking each insertion in the 28 genomic sequences from NCBI; these direct repeats would be lost following secondary rearrangements. In practice, each insertion site should be checked by sequencing amplified genomic DNA fragments encompassing the site.

IS10 elements are not endogenous to K-12 E. coli strains, but the widespread use of Tn10 in strain construction has resulted in their presence in many laboratory cloning strains, including those that are no longer tetracycline resistant, for example JM109 (Matsutani, 1991). Database entries should be automatically screened for both IS10 and endogenous K-12 insertion sequences (such as IS1, IS2, etc.), as they can be for vector sequences (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html). We found, for example, complete copies of IS1 in both Drosophila (DDBJ/EMBL/GenBank accession Nos: AE002757 and AC007176) and human sequences (DDBJ/EMBL/GenBank accession Nos: AF020504, AC005386, AC007948 and AC005684).

In summary, insertions of the mobile element IS10 are not uncommon in large genomic clones and their presence needs to be annotated. More positively, we can conclude from an analysis of these insertion sequences, that the target accuracy of less than one sequencing error in 10,000 base-pairs is being met in large-scale sequencing centres.

ACKNOWLEDGEMENTS
We thank the referees for constructive comments.

REFERENCES


Bogosian, G., Bilyeu, K. and O’Neil, J.P. (1993) Genome rearrangements by the element (Shen et al., 1987). We confirmed the presence of characteristic 9 bp duplications flanking each insertion in the 28 genomic sequences from NCBI; these direct repeats would be lost following secondary rearrangements. In practice, each insertion site should be checked by sequencing amplified genomic DNA fragments encompassing the site.

IS10 elements are not endogenous to K-12 E. coli strains, but the widespread use of Tn10 in strain construction has resulted in their presence in many laboratory cloning strains, including those that are no longer tetracycline resistant, for example JM109 (Matsutani, 1991). Database entries should be automatically screened for both IS10 and endogenous K-12 insertion sequences (such as IS1, IS2, etc.), as they can be for vector sequences (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html). We found, for example, complete copies of IS1 in both Drosophila (DDBJ/EMBL/GenBank accession Nos: AE002757 and AC007176) and human sequences (DDBJ/EMBL/GenBank accession Nos: AF020504, AC005386, AC007948 and AC005684).

In summary, insertions of the mobile element IS10 are not uncommon in large genomic clones and their presence needs to be annotated. More positively, we can conclude from an analysis of these insertion sequences, that the target accuracy of less than one sequencing error in 10,000 base-pairs is being met in large-scale sequencing centres.

ACKNOWLEDGEMENTS

We thank the referees for constructive comments.

REFERENCES


Bogosian, G., Bilyeu, K. and O’Neil, J.P. (1993) Genome rearrangements by the element (Shen et al., 1987). We confirmed the presence of characteristic 9 bp duplications flanking each insertion in the 28 genomic sequences from NCBI; these direct repeats would be lost following secondary rearrangements. In practice, each insertion site should be checked by sequencing amplified genomic DNA fragments encompassing the site.


DOI: 10.1093/embo-reports/kvd015