EMBO reports

scientific report

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*10*, a component of the transposon Tn*10*, 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*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*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, Tn*10*, 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). Tn*10* contains inverted repeats, designated IS*10*, of 1329 base pairs at each end (for a review of Tn*10*, see Kleckner *et al.*, 1996). The repeats differ from each other at 16 positions (Chalmers *et al.*, 2000). Each IS*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*10* elements may be left behind (Ross *et al.*, 1979; Shen *et al.*, 1987). These can themselves transpose at a frequency of ~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 ('nr' database at http://www.ncbi.nlm.nih.gov/blast/) using the IS*10*R prototype as a query sequence (Halling *et al.*, 1982) uncovered 28 genome project clones that were contaminated by IS*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 Tn*10* that truncates the tetracycline resistance gene).

These 29 IS*10* copies were aligned (summarized in Table I); 18 perfect copies of IS*10*R were found, but only one copy of IS*10*L. The right copy, IS*10*R, is known to be more than 10 times as mobile as the left, IS*10*L (Foster *et al.*, 1981). The remaining 10 IS elements were distinct from both IS*10*L and IS*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 IS*10*R or IS*10*L at the 16 variable positions as correct.

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

Table I. Variation in IS*10* sequences

Position	24	25	27	47	83	107	203	410	425	903	1249	1268	1271	1272	1281	1300
IS*10*R	G	G	A	G	G	G	A	T	A	G	T	A	A	G	T	C
18 sequences	G	G	A	G	G	G	A	T	A	G	T	A	A	G	T	C
AC005317	G	G	A	G	G	G	A	*C*	*G*	G	T	A	A	G	T	C
9 sequences	*T*	*A*	*C*	*A*	*C*	*A*	*C*	*C*	*G*	*A*	T	A	A	G	T	C
AC006473	*T*	*A*	*C*	*A*	*C*	*A*	*C*	*C*	*G*	*A*	*C*	*G*	*T*	*A*	*C*	*T*
IS*10*L	*T*	*A*	*C*	*A*	*C*	*A*	*C*	*C*	*G*	*A*	*C*	*G*	*T*	*A*	*C*	*T*

Variable nucleotide sites in IS*10* sequences are numbered according to Halling *et al.* (1982) above each column. These variants are found either naturally in IS*10*R and IS*10*L (both from AF162223 described in Chalmers *et al.*, 2000), or contaminating genomic DNA sequences in public databases. The DDBJ/EMBL/GenBank accession Nos of the 18 sequences identical to IS*10*R are: AC007226; AC005410; AC008583; AC008102; AC007396; AC007396; AC009247; AC004617; AC007887; AC007630; AC007566; AC006210; AC005552; AC005215; AL136059; AL137818; AP000025; AP000245; AP000128. The accession numbers of the nine hybrid IS*10*L–IS*10*R sequences (fourth row) are: AC006650 (contains two IS*10* elements); AC012463; AC006305; AL133246; AL133224; AP000056; AP000330; AP000124.

It is well recognized that some DNA templates are easier to sequence accurately than others. To serve as a reference standard, IS*10* should be neither unusually difficult nor easy to sequence. To assess this, the error frequency in 40 single-pass sequences of ESTs containing IS*10* was determined, and found to be 3.1%. In addition, all other eukaryotic sequences containing full-length IS*10* 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 sequencing of IS*10* is not trivial.

Further contamination of the databases by IS*10* is unavoidable, 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 IS*10* 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 IS*10* elements that we found were annotated in any way, usually as Tn*10*. Worse are the occasional misleading annotations pointing out the similarity of the insertion sequence to cDNAs or ESTs that also contained IS*10*.

The presence of IS*10* 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.

IS*10* elements are not endogenous to K-12 *E. coli* strains, but the widespread use of Tn*10* 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 IS*10* and endogenous K-12 insertion sequences (such as IS*1*, IS*2*, etc.), as they can be for vector sequences (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html). We found, for example, complete copies of IS*1* 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 IS*10* 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

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1992) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Beck, S. (1993) Accuracy of DNA sequencing: should the sequence quality be monitored? *DNA Seq.*, **4**, 215–217.

Bentley, D.R. (1996) Genomic sequence information should be released immediately and freely in the public domain. *Science*, **274**, 533–534.

Bochner, B.R., Huang, H.C., Schieven, G.L. and Ames, G.N. (1980) Positive selection for loss of tetracycline resistance. *J. Bacteriol.*, **143**, 926–933.

Bogosian, G., Bilyeu, K. and O'Neil, J.P. (1993) Genome rearrangements by residual IS10 elements in strains of *Escherichia coli* K-12 which had undergone Tn10 mutagenesis and fusaric acid selection. *Gene*, **133**, 17–22.

Chalmers, R., Sewitz, S., Lipkow, K. and Crellin, P. (2000) Complete nucleotide sequence of Tn*10*. *J. Bacteriol.*, **182**, 2970–2972.

Davis, M.A. (1986) Determinants of the activity of transposon Tn*10*. PhD Thesis, Harvard University, Cambridge, MA, USA.

Foster, T.J., Davis, M.A., Roberts, D.E., Takeshita, K. and Kleckner, N. (1981) Genetic organization of transposon Tn10. *Cell*, **23**, 201–213.

Grant, S.G., Jessee, J., Bloom, F.R. and Hanahan, D. (1990) Differential plasmid rescue from transgenic mouse DNAs into *Escherichia coli* methylation-restriction mutants. *Proc. Natl Acad. Sci. USA*, **87**, 4645–4649.

Halling, S.M., Simons, R.W., Way, J.C., Walsh, R.B. and Kleckner, N. (1982) DNA sequence organization of IS10-right of Tn10 and comparison with IS10-left. *Proc. Natl Acad. Sci. USA*, **79**, 2608–2612.

Kleckner, N., Roth, J. and Botstein, D. (1977) Genetic engineering *in vivo* using translocatable drug-resistance elements. New methods in bacterial genetics. *J. Mol. Biol.*, **116**, 125–159.

Kleckner, N., Chalmers, R., Kwon, D., Sakai, J. and Bolland, S. (1996) Tn10 and IS10 transposition and chromosome rearrangements: mechanism and regulation *in vivo* and *in vitro*. *Curr. Top. Microbiol. Immunol.*, **204**, 49–82.

Maloy, S.R. and Nunn, W.D. (1981) Selection for loss of tetracycline resistance by *Escherichia coli*. *J. Bacteriol.*, **145**, 1110–1111.

Matsutani, S. (1991) Multiple copies of IS*10* in the *Enterobacter cloacae* MD36 chromosome. *J. Bacteriol.*, **173**, 7802–7809.

Ross, D.G., Swan, J. and Kleckner, N. (1979) Physical structures of Tn10-promoted deletions and inversions: role of 1400 bp inverted repetitions. *Cell*, **16**, 721–731.

Shen, M.M., Raleigh, E.A. and Kleckner, N. (1987) Physical analysis of Tn10- and IS10-promoted transpositions and rearrangements. *Genetics*, **116**, 359–369.

DOI: 10.1093/embo-reports/kvd015