DNA Sequence, October/December 2004 Vol. 15 (5/6), pp. 362–364

Taylor & Francis
healthsciences

Short Communication

DNA Sequence Error Rates in Genbank Records Estimated using the Mouse Genome as a Reference

PHILIPP L. WESCHE, DANIEL J. GAFFNEY and PETER D. KEIGHTLEY*

University of Edinburgh, School of Biological Sciences, Ashworth Laboratories, West Mains Road, Edinburgh EH9 3JT, UK

(Received 6 May 2004)

We estimate DNA sequence error rates in Genbank records containing protein-coding and non-coding DNA sequences by comparing sequences of the inbred mouse strain C57BL/6J, sequenced as part of the mouse genome project and independently by other laboratories. C57BL/6J was produced by more than 100 generations of brother–sister mating, and can be assumed to be virtually free of residual polymorphism and mutational variation, so differences between independent sequences can be attributed to error. The estimated single nucleotide error rate for coding DNA is 0.10% (SE 0.012%), which is substantially lower than previous estimates for error rates in Genbank accessions. The estimated single nucleotide error rate for intronic DNA sequences (0.22%; SE 0.051%) is significantly higher than the rate for coding DNA. Since error rates for the mouse genome sequence are very low, the vast majority of the errors we detected are likely to be in individual Genbank accessions. The frequency of insertion–deletion (indel) errors in non-coding DNA approaches that of single nucleotide errors in non-coding DNA, whereas indel errors are uncommon in coding sequences.

Keywords: DNA sequence error rates; Genbank; Mouse strain; Indels

Sequencing errors submitted to public databases can have serious consequences for future research for at least three reasons. First, sequencing errors can cause automated annotation algorithms to fail to detect sequence features such as initiation and termination signals and introns. Second, sequencing errors can lead to spurious annotation of single nucleotide polymorphisms. Third, sequencing errors can lead to inaccurate or erroneous evolutionary inferences, such as in estimating the rate of nucleotide substitution or recombination (Clark and Whittam, 1992). The consequences of sequencing error will be most serious in intraspecific comparisons involving species with low natural genetic variability such as humans (Li and Sadler, 1991), or in interspecific comparisons between closely related taxa.

There have been several previous analyses of sequencing error rates in public databases. Karlin *et al.* (2001) compared *Drosophila* SwissProt protein sequences with the *Drosophila* genome published by Celera Genomics (Adams *et al.*, 2000), and found that 45% of the 1,059 sequences sampled "had differences of more than 1%, including mismatches, insertions and deletions". Hill *et al.* (2000) studied contaminating mobile genetic element sequences derived from cloning artifacts in Genbank, and estimated the error rate in large-scale genome projects to be less than 0.01%. This confirms that multiple sequence passes result in very high quality sequence. However, single pass sequencing of these elements revealed an error rate as high as 3.1% (Hill *et al.*, 2000). This is similar to the error rates of 3.6 and 3.2% reported by Kristensen *et al.* (1992) and Lamperti *et al.* (1992), respectively, in vector sequences that had contaminated the sequence databases. Krawetz (1989), however, reports an error rate of only 0.29% for GenBank on the basis of the frequency of annotated conflicts and revisions, and Beck (1993) reports error rates averaging 0.46% for resequencing of cosmids containing human genomic DNA by independent laboratories.

Here, we use independent sequences from the inbred *Mus musculus* strain C57BL/6J to estimate

*Corresponding author. Tel.: +44-0-131-650-5443. E-mail: peter.keightley@ed.ac.uk

ISSN 1042-5179 print/ISSN 1029-2365 online © 2004 Taylor & Francis Ltd
DOI: 10.1080/10425170400008972

the sequencing error rate in stretches of coding and non-coding DNA. The C57BL/6J inbred strain was developed in the 1920s by C. C. Little at the Jackson Laboratory (Morse, 1978) initially by 125 generations of brother–sister matings (Foster *et al.*, 1981). DNA sequences of many loci from this strain or its inbred derivatives have been deposited in GenBank. The Mouse Genome Sequencing Consortium's (MGSC) sequence of the mouse genome (MGSC, 2002) was from the C57BL/6J strain (the Jackson Laboratory's reference strain), and was produced by an initial phase of whole genome shotgun sequencing. The whole genome shotgun sequence (WGSS) was produced by 7.7-fold sequence coverage of the euchromatic regions (MGSC, 2002); with such a high degree of sequence coverage, the mouse genome is an accurate reference against which the accuracy of GenBank records may be assessed.

We compiled data sets of C57BL/6J sequences, excluding expressed sequence tags, sequence tagged sites, genome survey sequence, and working draft, comprising sequences from 375 protein-coding genes plus 63 intron sequences from 46 loci. The average intronic DNA sequence length was 460 bases per locus. From the gene sequences, we extracted coding sequences and their 5′ and 3′ flanking sequences (average lengths 1,201, 288 and 745 bases, respectively). We then extracted sequences for the identical loci from the mouse WGSS using BLAST searches, with repetitive elements masked using RepeatMasker. We aligned these pairs of sequences using ClustalW v1.82 (Thompson *et al.*, 1994). The accession numbers of the sequences analysed are available on request. The alignments were inspected visually before proceeding. We calculated k_n and k_g, the frequency of mismatches and gaps, respectively, per aligned base, and θ, the ratio of gaps to mismatches.

Estimates of k_n, k_g, and θ are shown in Table I. Our estimate for the error rate for individual Genbank coding DNA accessions is close to 0.1%. However, the estimate for k_n in intronic sequences of 0.22% is significantly greater than that for coding sequences ($P < 0.0001$). Error rates for 5′ and 3′ UTRs are also substantially higher than for coding DNA. Two possible reasons for this are: (1) less attention being paid to the sequencing of non-coding DNA when the primary goal is to obtain the coding sequence, and, (2) a higher intrinsic difficulty in accurately sequencing non-coding DNA.

Estimates of error rate in the mouse WGSS are better than 0.01% (MGSC), which implies that the vast majority of these nucleotide sequencing errors are in the individual Genbank records rather than the WGSS. We independently checked the error rate in the mouse WGSS by estimating the number of intronic splice donor and acceptor nucleotides that have been incorrectly sequenced. These sites are believed to be practically invariant, and, as such, deviations from known splice site types can indicate potential sequencing errors. Donor and acceptor sites were considered only if they belonged to the known canonical types. We aligned a total of 6,956 mouse mRNA sequences to the WGSS, then counted the number of sites adjacent to correctly aligned exons that differed by one base from the canonical GT/GC 5′ donor and AG 3′ acceptor dinucleotides. These putative errors were then confirmed by comparing with the rat ortholog, extracted from the rat genome sequence (RGSPC, 2004). We observed six such differences in 24,569 introns, giving an adjusted error rate of 4.1×10^{-5}. This is similar to the sequencing error rate that has been estimated for the mouse WGSS (MGSC, 2002), and confirms that the majority of the errors between C57BL/6J accessions can be attributed to errors in individual Genbank records.

Coding sequence alignments contained gaps at a frequency of less than 0.01%. In both coding and non-coding DNA, gaps are mostly due to single bp indels (Fig. 1). The relative frequency of gaps (θ) in intronic sequences is about 40 times higher than in coding sequences. This is a likely consequence of indels being more easily spotted and corrected in coding DNA, due to the tendency for indels that are not multiples of 3 bp to cause downstream premature stop codons.

Residual polymorphism, recent mutation, and contamination of genetic material could all cause differences between Genbank accessions and the reference sequence. C57BL/6J was created by 125 generations of brother–sister matings (Foster *et al.*, 1981), after which the expected proportion of

TABLE I Numbers of coding and intronic nucleotides sampled, along with numbers and frequencies of nucleotide differences and gaps between Genbank accessions and the mouse genome

	Coding sequences	Introns	5′ UTR	3′ UTR
No. of accessions	340	46	149	285
Nucleotides	408, 419	31, 256	42, 933	212, 407
Mismatches	413	68	61	332
Gaps	22	67	43	185
$k_n \times 100$ (SE)	0.101 (0.012)	0.217 (0.051)	0.143 (0.039)	0.157 (0.019)
$k_g \times 100$ (SE)	0.00540 (0.0013)	0.215 (0.049)	0.100 (0.026)	0.0872 (0.012)
θ (SE)	0.0541 (0.014)	0.985 (0.236)	0.739 (0.237)	0.557 (0.071)

Standard errors were estimated from 1,000 bootstrap samples in which sequence data for individual loci were resampled with replacement.

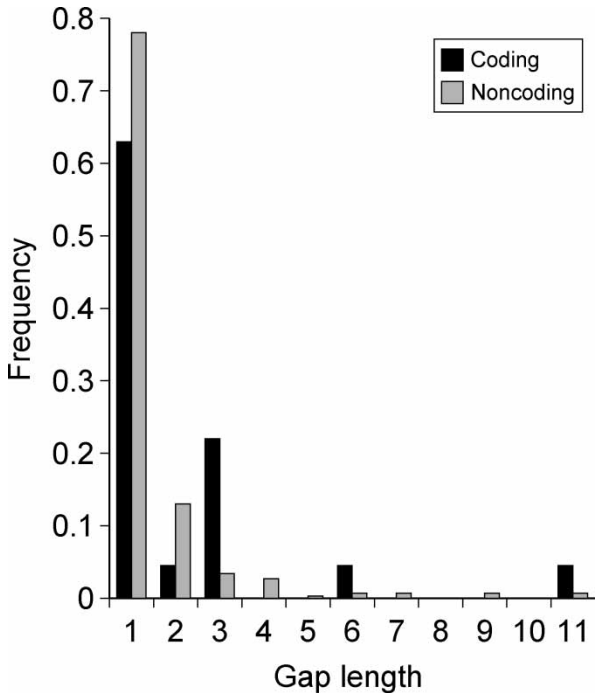

FIGURE 1 Frequency distribution of gap length in coding and non-coding sequences.

the original genetic variation is $(3/4)^{125} = 2.4 \times 10^{-16}$. If the original genetic variation was at the upper end of the range of polymorphism rates between extant mouse inbred strains (about 0.4%; Wade *et al.*, 2002), the predicted residual polymorphism after the inbreeding process would therefore be several orders of magnitude below the levels of differences we observe. The Jackson Laboratory started distributing C57BL/6J mice in the late 1940s, and in 1981 the C57BL/6J strain was known to be maintained in 57 separate locations worldwide (Foster *et al.*, 1981). As a result, some new variation will have arisen from new mutations in the locally maintained sublines. However, assuming a maximum cumulative subline divergence of 100 years, and a point mutation rate equal to the estimated synonymous substitution rate of 7×10^{-9} per year (Keightley and Eyre-Walker, 2000), the predicted polymorphism rate between sublines is less than 10^{-6}, a figure three orders of magnitude below our observed rate (Table I). Finally, contamination by non-C57BL/6J inbreds could seriously inflate our estimates of the sequencing error rate. However, the C57BL/6J strain carries the recessive non-agouti coat colour gene, which makes contamination from most common inbreds (with the notable exception of C57BL/10) relatively easy to detect. Furthermore, genetic monitoring by suppliers of inbred mouse strains, initially with electrophoretic markers, and now with DNA markers, has been routine since the 1980s.

Whether our estimates of error rates in murine sequences are likely to be representative of gene sequence accessions deposited in Genbank depends on whether investigators sequencing murine DNA are more or less accurate than the average. Although our estimate of k_n is at the lower end of the range of previous estimates of the nucleotide error rate for coding sequences, it is sufficiently high as to imply that polymorphism studies require sequencing accuracy substantially above what is routinely achieved in single sequencing runs. The problem of sequence error is about twice as acute for non-coding DNA, and is exacerbated by the relatively high frequency of indels.

Acknowledgements

We thank Julian Christians for helpful comments on the manuscript.

References

Adams, M.D., Celniker, S.E., Holt, R.A., *et al.* (2000) "The genome sequence of *Drosophila melanogaster*", *Science* **287**, 2185–2195.
Beck, S. (1993) "Accuracy of DNA sequencing: should the sequence quality be monitored?", *DNA Sequence* **4**, 215–217.
Clark, A.G. and Whittam, T.S. (1992) "Sequencing errors and molecular evolutionary analysis", *Molecular Biology and Evolution* **9**, 744–752.
Foster, H.L., Small, J.D. and Fox, J.G. (1981) *The Mouse in Biomedical Research* (Academic Press, New York), Vol. 1.
Hill, F., Gemünd, C., Benes, V., Ansorge, W. and Gibson, J. (2000) "An estimate of large-scale sequencing accuracy", *EMBO Reports* **1**, 29–31.
Karlin, S., Bergman, A. and Gentles, A.J. (2001) "Annotation of the *Drosophila* genome", *Nature* **411**, 259–260.
Keightley, P.D. and Eyre-Walker, A. (2000) "Deleterious mutations and the evolution of sex", *Science* **290**, 331–333.
Krawetz, S.A. (1989) "Sequence errors described in Genbank—a means to determine the accuracy of DNA sequence interpretation", *Nucleic Acids Research* **17**, 3951–3957.
Kristensen, T., Lopez, R. and Prydz, H. (1992) "An estimate of the sequencing error frequency in the DNA sequence databases", *DNA Sequence* **2**, 343–346.
Lamperti, E.D., Kittelberger, J.M., Smith, T.F. and Villakomaroff, L. (1992) "Corruption of genomic databases with anomalous sequence", *Nucleic Acids Research* **20**, 2741–2747.
Li, W.H. and Sadler, L.A. (1991) "Low nucleotide diversity in man", *Genetics* **129**, 513–523.
Morse, H.C. (1978) *Origins of Inbred Mice* (Academic Press, New York), pp 1–21.
Mouse Genome Sequencing Consortium (2002) "Initial sequencing and comparative analysis of the mouse genome", *Nature* **420**, 520–562.
Rat Genome Sequencing Project Consortium (2004) "Genome sequence of the Brown Norway rat yields insights into mammalian evolution", *Nature* **428**, 493.
Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) "CLUSTAL W—Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", *Nucleic Acids Research* **22**, 4673–4680.
Wade, C.M., Kulbokas, E.J., Kirby, A.W., Zody, M.C., Mullikin, J.C., Lander, E.S., Lindblad-Toh, K. and Daly, M.J. (2002) "The mosaic structure of variation in the laboratory mouse genome", *Nature* **420**, 574–578.