Transcription-associated mutational asymmetry in mammalian evolution

Phil Green[1], Brent Ewing[1], Webb Miller[2], Pamela J. Thomas[3], NISC Comparative Sequencing Program[3,4] & Eric D. Green[3,4]

Published online 3 March 2003, doi:10.1038/ng1103

Although mutation is commonly thought of as a random process, evolutionary studies show that different types of nucleotide substitution occur with widely varying rates that presumably reflect biases intrinsic to mutation and repair mechanisms[1–4]. A strand asymmetry[5,6], the occurrence of particular substitution types at higher rates than their complementary types, that is associated with DNA replication has been found in bacteria[7] and mitochondria[8]. A strand asymmetry that is associated with transcription and attributable to higher rates of cytosine deamination on the coding strand has been observed in enterobacteria[9–11]. Here, we describe a qualitatively different transcription-associated strand asymmetry in mammals, which may be a byproduct of transcription-coupled repair[12] in germline cells. This mutational asymmetry has acted over long periods of time to produce a compositional asymmetry, an excess of G+T over A+C on the coding strand, in most genes. The mutational and compositional asymmetries can be used to detect the orientations and approximate extents of transcribed regions.

We obtained most of the genomic sequence orthologous to a locus of roughly 1.5 Mb on human chromosome 7 containing nine known genes (Fig. 1) from each of eight other mammals (chimpanzee, baboon, cow, pig, cat, dog, mouse and rat). In initial analyses (Fig. 2 and Supplementary Fig. 1 online), we tabulated substitutions that have occurred in this locus in the human and chimpanzee lineages since their last common ancestor. There was a significant strand asymmetry in substitution rates, with the transition A→G occurring at a 28% higher rate than the complementary transition T→C ($\chi^2_{1df} = 33.54$, $P < 0.00001$).

To examine a possible association with transcription, we tabulated separately the substitutions at transcribed and untranscribed positions, scoring the former with respect to the coding strand (that is, the strand complementary to the template strand for transcription). We saw pronounced asymmetries in the transcribed regions for transition substitutions (Fig. 2): A→G transitions were 58% more frequent than T→C ($\chi^2_{1df} = 72.4$, $P < 0.00001$), and G→A transitions were 18% more frequent than C→T ($\chi^2_{1df} = 10.01$, $P < 0.002$). These asymmetries were also seen when we considered only substitutions in interspersed repeats in the transcribed regions (Fig. 2). Because such sequences are thought to be non-functional, this indicates that the pattern reflects an asymmetry in neutral mutation, rather than selection. Purine transitions were more frequent and pyrimidine transitions less frequent in transcribed regions than in untranscribed regions (Fig. 2), such that the overall transition rate in interspersed repeats was essentially identical for the transcribed and untranscribed portions of the locus (0.00614 versus 0.00613).

To test whether this asymmetry was specific to transcribed regions, we did a 'maximal segment' analysis[13] to identify regions with a significant excess or deficit of purine transitions relative to pyrimidine transitions (Fig. 1). We used the baboon sequence rather than the chimpanzee sequence for this analysis, because its higher level of divergence from the human (about 6% versus about 1%) provides more statistical power and higher resolution in detecting asymmetries. Lacking a close outgroup for the human–baboon comparison, we could not reliably score transition directions, but we could classify transitions as purine (A↔G) or pyrimidine (C↔T) with respect to the top strand. We found six maximal segments of excess purine transitions (Fig. 1), and these correspond to the six known genes that are transcribed from left to right. Similarly, we found two segments showing the complementary pattern, an excess of pyrimidine transitions, and these correspond

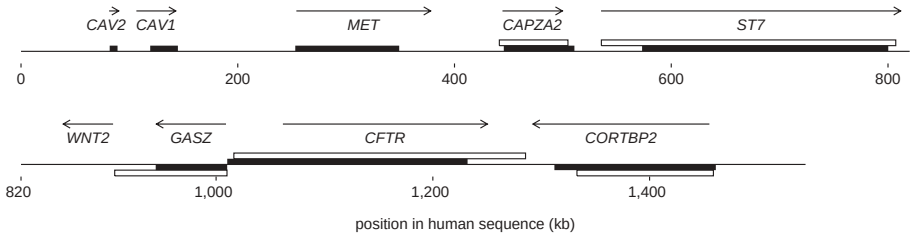

Fig. 1 Known transcripts, regions of excess purine or pyrimidine transition substitutions and regions of excess G+T or C+A composition in the sequenced locus. Arrows indicate transcript direction and extent. Solid bars indicate maximal segments in which approximately 55% of isolated transitions in the human–baboon alignment involve top-strand purines (bar shown above line) or top-strand pyrimidines (bar shown below line). Open bars indicate maximal segments in which approximately 52% of bases (ignoring repeats) in the human sequence are G or T (bar shown above line) or C or A (bar shown below line).

[1]*Howard Hughes Medical Institute and Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA.* [2]*Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA.* [3]*NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.* [4]*Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. Correspondence should be addressed to P.G. (e-mail: phg@u.washington.edu).*

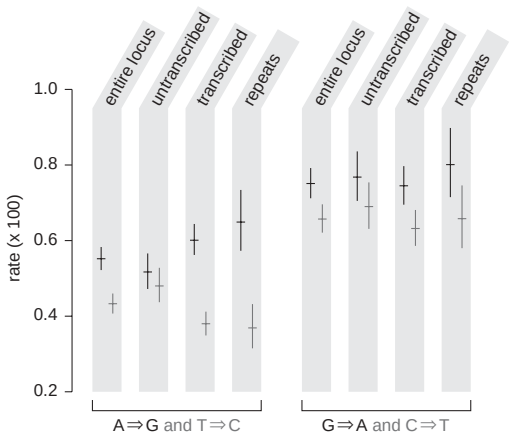

Fig. 2 Rates of isolated, non-CpG transition substitutions in the combined human and chimpanzee lineages since their last common ancestor. Substitutions are properly viewed as changes in base pairs rather than bases, as it is not possible with evolutionary data to infer on which DNA strand the mutation originally occurred. But for notational convenience, we arbitrarily chose a strand and scored the substitution by indicating the base change on that strand. Substitutions are scored on the top strand for the entire locus and untranscribed region estimates, and on the coding strand of each transcript for the transcribed region estimates. Rates for each substitution and its complement (for example, A→G and T→C) and their 95% confidence intervals are shown side by side, by region type. transcribed, the exons and introns of the nine known genes; untranscribed, all other positions in the locus; repeats, interspersed repeats in transcribed regions.

to two of the three known genes transcribed in the complementary direction (*WNT2* has no corresponding segment).

Although segment boundaries did not precisely align with the transcript boundaries (possibly owing to an insufficient density of informative sites), this analysis showed that the strand asymmetry was associated specifically with transcribed regions and extended throughout them. It did not show the type of pattern expected to result from mutational differences in leading- and lagging-strand synthesis starting from multiple replication origins, that is, complementary biases flanking particular sites[14]. (A substitution asymmetry associated with replication origins in the β-globin locus has been reported[15] but disputed[14,16].) The originally observed asymmetry in the locus as a whole may now be understood as arising from the fact that six of the eight genes showing the asymmetry are oriented in the same direction.

We then tested whether the asymmetry could be seen in non-primate mammalian lineages by looking at transition substitutions between cow and pig, cat and dog and mouse and rat for each of the nine genes (Fig. 3). Each species pair showed an excess of purine transitions on the coding strand for most of the genes, although there were lineage-specific differences as to which genes were affected. In particular, *CAV1* and *CAV2* showed the asymmetry only in the primates, and the rodents did not show the asymmetry for two genes (*ST7* and *CORTBP2*) that were affected in the other three species pairs.

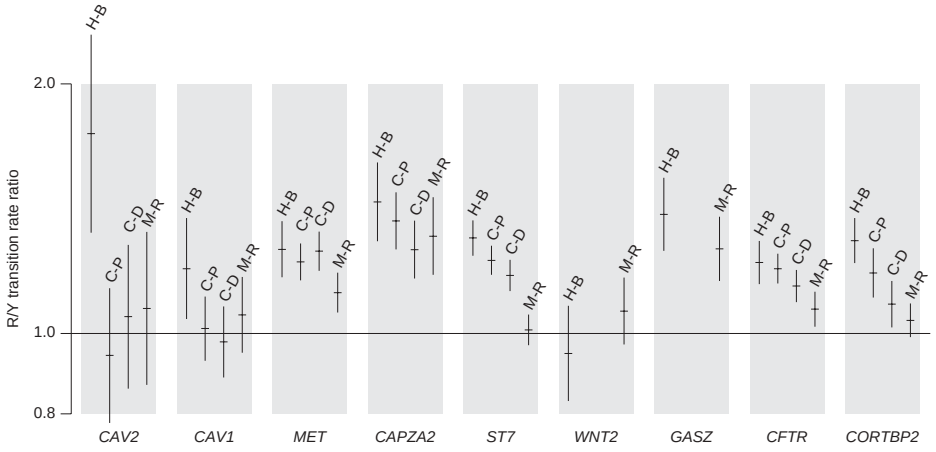

Fig. 3 Transcript-specific ratios of A↔G to C↔T transition rates for different pairs of mammals. For each gene, the rate ratio and 95% confidence interval are shown for the human–baboon (H-B), cow–pig (C-P), cat–dog (C-D) and mouse–rat (M-R) sequence alignments, in that order. All substitutions are scored on the coding strand. *WNT2* and *GASZ* sequences are not yet available for pig and dog.

Because the A→G versus T→C asymmetry is stronger than the G→A versus C→T asymmetry (Fig. 2), over long periods of time it should produce an excess of G nucleotides relative to C and of T relative to A; a quantitative analysis[17] taking into account the observed rates predicted the equilibrium G+T frequency to be 52.7%. We examined this prediction in the human sequence, excluding interspersed repeats (which may have inserted too recently to have reached compositional equilibrium) and protein-coding exons (which are under selection) from the analysis. Six of the nine genes showed a significant ($P < 0.05$) G+T excess (Fig. 4), the three exceptions being *WNT2*, which did not show the mutational asymmetry, and *CAV1* and *CAV2*, for which the mutational asymmetry seems to be specific to the primate lineage (Fig. 3) and so may not have been acting long enough to produce a compositional bias. For the remaining genes, the G+T excess was generally of roughly the predicted magnitude, although it was weak for *MET*. Maximal segment analysis (Fig. 1) indicated that, like the mutational asymmetry, the regions of G+T excess were specific to, and spanned, the transcribed regions.

To explore the extent of this compositional asymmetry in human genes, we analyzed the sequence of human chromosome 22 (ref. 18). Of 275 annotated transcripts for which mRNA data was available, 187 (68%) showed a significant ($P < 0.01$) G+T excess. For transcripts longer than 10 kb, this proportion rose to 82% (159 of 195), and for transcripts longer than 20 kb, it was 91% (127 of 139). The average G+T content over all transcripts was 52.6%, close to the predicted value. We note that a G+T compositional excess has also been observed in several bacterial genomes[7], but it seems to arise from replication rather than transcription[7,19].

Our analyses identified a strand asymmetry in neutral substitution patterns in most mammalian genes. This probably explains a previously observed strand asymmetry in disease-causing mutations in human genes[20] in which the pattern was less clear owing to the effects of selection. In contrast with the known transcription-associated substitution asymmetry in enterobacteria, which is characterized by an excess of C→T coding-strand transitions attributable to cytosine deamination[9–11], the one we found is characterized by an excess of purine transitions and a deficit of pyrimidine transitions relative to untranscribed DNA. As there is no difference in overall substitution rate, the asymmetry is probably not due to differences in mutation rates or repair efficiency in transcribed regions. Rather, we

believe that it is a byproduct of transcription-coupled repair (TCR; refs. 12,21) acting on the mismatched base pairs that result from uncorrected DNA polymerase substitution errors during DNA replication. The usual fate of such mismatches in untranscribed regions is presumably to persist until the next replication round, at which time the two mispaired bases segregate into the daughter DNA duplexes, implying a 50% chance that a given daughter cell will inherit the mutation. Mispairs in transcribed regions, however, may be resolved by TCR before the next replication round.

According to current models[21], TCR is triggered by DNA damage–induced stalling of RNA polymerase II, but its targeting to specific sites is thought to require the mismatch repair proteins MSH2 and MSH6 (refs. 22,23). As the MSH2–MSH6 heterodimer MutSα recognizes mispairs as well as damaged bases, the TCR repair machinery may be directed to any mispair in the vicinity of the stalled polymerase. It is, moreover, plausible that MutSα bound to the mispair could itself trigger TCR by obstructing the RNA polymerase II complex, even in the absence of DNA damage. In any case, repair then proceeds by excision of an oligonucleotide patch on the transcribed DNA strand, followed by resynthesis using the coding strand as a template. This will resolve the mismatched base pair into a proper base pair, which will be mutant if, and only if, the originally misinserted base is on the coding strand.

Assuming that misinsertions occur with equal frequency on the two strands, the above mechanism again implies a 50% chance that a daughter cell inherits the mutation, so it does not change the overall mutation rate. But the spectrum of resulting mutational events, viewed as base changes on the coding strand, will correspond to the spectrum of DNA polymerase base misinsertion errors, which may be strand-asymmetric. Our observation of an excess of purine transitions is consistent with data from both prokaryotic and eukaryotic systems[24,25] that the misinserted base in a purine–pyrimidine mispair is more likely to be the purine than the pyrimidine. Moreover, the fact that the strongest asymmetry occurs for A→G transitions, which in this model would result from the resolution of G–T mispairs arising from misinserted G, is consistent with the observation that MutSα is particularly efficient at recognizing G–T mispairs[26].

The above explanation can thus account for each of the key observations regarding the mutational asymmetry. If it is correct, then TCR must be fairly active in the mammalian germ line, affecting most (but not all) genes. The observation that most genes show the G+T compositional bias suggests that the mutational bias has been acting for much of mammalian evolution, although differences among lineages (Fig. 3) indicate some changes in germline gene expression may have occurred.

The maximal segment analyses illustrated in Figure 1 offer two new methods to detect genes in mammalian genomes. Traditional comparative genomic approaches depend on sequence conservation reflecting purifying selection to identify biological features, and require relatively diverged sequences (>30%). In contrast, our approach uses neutral mutation patterns to detect the extent and orientation of transcribed regions, and may be used either with single sequences or with sequence pairs too closely related for the traditional approach to analyze.

Methods

Sequencing and annotation. We isolated BAC clones for each species as described[27] and sequenced them at the National Institutes of Health Intramural Sequencing Center as part of a larger comparative sequencing program (E.D.G. *et al.* & NISC Comparative Sequencing Program, unpublished data). We identified genes in the human and mouse sequences using the National Center for Biotechnology Information tool spidey to align reference cDNAs from RefSeq to the genomic sequence. We similarly generated gene locations for the other mammalian genomic sequences when cDNAs were available, but otherwise we inferred them by transfer from the human or mouse sequence by pairwise genomic sequence alignment using the program transform-pos. We used sequin to validate and check the inferred exon structure. We identified interspersed repeats in the human sequence using Repeatmasker version 07072001, run in sensitive mode, using RepBase Update version 6.10. GenBank accession numbers for the clone sequences from each organism are listed in Supplementary Table 1 online.

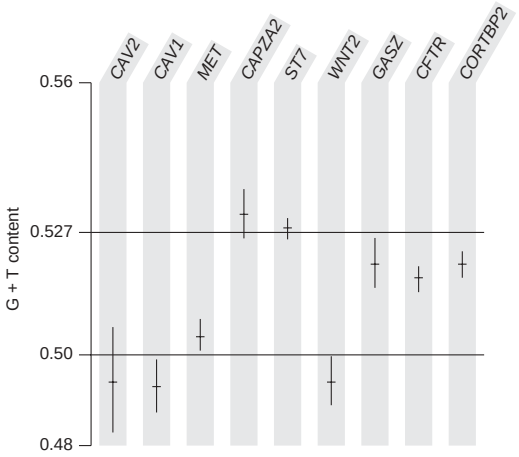

Fig. 4 G+T composition in transcribed regions. For each gene, the G+T composition as determined from non-coding, non-repetitive sequence in the coding strand of the transcript and the 95% confidence interval are shown. Horizontal lines indicate unbiased (0.5) and predicted (0.527) compositions.

Sequence alignments. We computed pairwise alignments using BLASTZ[28] and the multiple alignment using MultiPipMaker.

Substitution rates. To minimize the effects of alignment artifacts and eliminate double substitution events, we tabulated only isolated substitutions, defined as those for which the 5′ and 3′ neighboring sites were identical in the species being compared. We also ignored possible CpG mutations, identified as transitions at CR (representing CA or CG) and YG (representing CG or TG) sites. We calculated rates by dividing the number of substitution events of the appropriate type by the number of potentially mutable sites that meet the same criteria (that is, that are flanked by sites that are identical in the species being compared and are not of the form CR or YG). We estimated confidence intervals for rates and rate ratios using standard procedures based on approximate normality of the log-transformed values. We tabulated substitutions in the human and chimpanzee lineages (Fig. 2) using baboon as an outgroup to infer the ancestral nucleotide using parsimony; positions where baboon differed from both human and chimpanzee sequences were ignored. We did not attempt to correct for multiple substitutions[4]; the effect of these is trivial for the human–chimpanzee comparison (roughly 1% diverged), and, though somewhat larger for the more diverged species pairs (Fig. 3), the effects on purine and pyrimidine transition rates are approximately proportionate and therefore cancel when the ratio is taken.

Maximal segment analysis. To detect regions with a relative excess of purine transitions in the human–baboon genomic sequence alignment, we first assigned scores to each isolated transition using a scoring system based on log-likelihood ratios that is theoretically optimal for discriminating regions in which at least 55% of transitions involve purines. Each purine transition was scored as $\log_2 (0.55/0.5) = 0.138$ and each pyrimidine transition as $\log_2 (0.45/0.5) = -0.152$. All other alignment positions were scored as 0. To detect regions of pyrimidine transition excess, we reversed these scores. We then identified maximal scoring segments of the alignment (that is, segments whose scores could not be increased by extending in either direction) whose score exceeded a given score threshold, *S*, using a

simple dynamic programming algorithm analogous to that used in BLAST[29]. The algorithm avoids merging distinct high-scoring segments separated by a region of negative score by imposing a 'dropoff' threshold, *D*. Appropriate thresholds were determined by simulating 1,000 copies of the 1.5-Mb alignment, each having an identical number of transitions to the original but randomly assigned as purine or pyrimidine, and identifying maximal scoring segments in each. We used $S = 12$ and $D = 7$, such that only 5% of the simulated alignments have a segment exceeding that score. Note that accuracy in predicting transcript boundaries is limited by the density of informative sites, as well as by the effects of selection at unknown features in the sequence.

Nucleotide compositional analysis. The K/M composition ratio (where K denotes G or T and M denotes A or C) at equilibrium should equal the substitution rate ratio M→K/K→M, which we estimated to be 0.00382/0.00343 = 1.11 from the rates for interspersed repeats in transcribed regions (Fig. 2 and Supplementary Fig. 1 online). This yields a predicted equilibrium frequency for K of 0.527. For the compositional maximal segment analysis shown in Figure 1, we used a slightly lower target value of 0.52, resulting in scores of $\log_2 (0.52/0.5) = 0.057$ for K and $\log_2 (0.48/0.5) = -0.059$ for M. Positions in interspersed repeats were scored as 0. We used conservative threshold and dropoff values of $S = 45$ and $D = 30$.

Chromosome 22 analysis. We used the repeat-masked sequence file Chr_22_19-05-2000.masked.fa, along with the associated annotation release 2.3 from the Chromosome 22 Gene Annotation Group[30]. We considered only those genes with supporting mRNA data (annotated as 'GD_mRNA'). For determining composition, we ignored all positions in interspersed repeats, in coding exons or in an overlapping transcript on the opposite strand.

URLs. National Institutes of Health Intramural Sequencing Center, http://www.nisc.nih.gov/; spidey, http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/; Repeatmasker, http://repeatmasker.genome.washington.edu/; RepBase, http://www.girinst.org/; blastz and transform-pos source code and MultiPipMaker, http://bio.cse.psu.edu/; chromosome 22 data and annotations, http://www.sanger.ac.uk/HGP/Chr22/. The mammalian sequences, annotations, and alignments used in this paper are available at http://www.nisc.nih.gov/data/.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

The following individuals were key contributors in the NIH Intramural Sequencing Center Comparative Sequencing Program: J. Thomas (BAC isolation and mapping); J. Touchman & R. Blakesley (BAC sequencing); G. Bouffard, S. Beckstrom-Sternberg, J. McDowell & B. Maskeri (computational analyses). We thank A. Smit and D. Haussler for helpful suggestions. This work was supported by the Howard Hughes Medical Institute and the National Human Genome Research Institute.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 30 December 2002; accepted 19 January 2003.

1. Gojobori, T., Li, W.H. & Graur, D. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**, 360–369 (1982).
2. Li, W.H., Wu, C.I. & Luo, C.C. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* **21**, 58–71 (1984).
3. Hess, S.T., Blake, J.D. & Blake, R.D. Wide variations in neighbor-dependent substitution rates. *J. Mol. Biol.* **236**, 1022–1033 (1994).
4. Li, W.H. *Estimating the number of nucleotide substitutions between sequences.* in *Molecular Evolution* (Sinauer Associates, Sunderland, Massachusetts, 1997).
5. Francino, M.P. & Ochman, H. Strand asymmetries in DNA evolution. *Trends Genet.* **13**, 240–245 (1997).
6. Frank, A.C. & Lobry, J.R. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**, 65–77 (1999).
7. Lobry, J.R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660–665 (1996).
8. Tanaka, M. & Ozawa, T. Strand asymmetry in human mitochondrial DNA mutations. *Genomics* **22**, 327–335 (1994).
9. Francino, M.P., Chao, L., Riley, M.A. & Ochman, H. Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science* **272**, 107–109 (1996).
10. Beletskii, A. & Bhagwat, A.S. Correlation between transcription and C to T mutations in the non-transcribed DNA strand. *Biol. Chem.* **379**, 549–551 (1998).
11. Francino, M.P. & Ochman, H. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. *Mol. Biol. Evol.* **18**, 1147–1150 (2001).
12. Hanawalt, P.C. Transcription-coupled repair and human disease. *Science* **266**, 1957–1958 (1994).
13. Karlin, S. & Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268 (1990).
14. Francino, M.P. & Ochman, H. Strand symmetry around the β-globin origin of replication in primates. *Mol. Biol. Evol.* **17**, 416–422 (2000).
15. Wu, C.I. & Maeda, N. Inequality in mutation rates of the two strands of DNA. *Nature* **327**, 169–170 (1987).
16. Bulmer, M. Strand symmetry of mutation rates in the β-globin region. *J. Mol. Evol.* **33**, 305–310 (1991).
17. Sueoka, N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.* **40**, 318–325 (1995).
18. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
19. Tillier, E.R. & Collins, R.A. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* **50**, 249–257 (2000).
20. Krawczak, M., Ball, E.V. & Cooper, D.N. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.* **63**, 474–488 (1998).
21. Svejstrup, J.Q. Mechanisms of transcription-coupled DNA repair. *Nat. Rev. Mol. Cell Biol.* **3**, 21–29 (2002).
22. Leadon, S.A. & Avrutskaya, A.V. Differential involvement of the human mismatch repair proteins, hMLH1 and hMSH2, in transcription-coupled repair. *Cancer Res.* **57**, 3784-3791 (1997).
23. Mellon, I., Rajpal, D.K., Koi, M., Boland, C.R. & Champe, G.N. Transcription-coupled repair deficiency and mutations in human mismatch repair genes. *Science* **272**, 557–560 (1996).
24. Bebenek, K., Joyce, C.M., Fitzgerald, M.P. & Kunkel, T.A. The fidelity of DNA synthesis catalyzed by derivatives of *Escherichia coli* DNA polymerase I. *J. Biol. Chem.* **265**, 13878–13887 (1990).
25. Mendelman, L.V., Boosalis, M.S., Petruska, J. & Goodman, M.F. Nearest neighbor influences on DNA polymerase insertion fidelity. *J. Biol. Chem.* **264**, 14415–11423 (1989).
26. Jiricny, J. Replication errors: cha(lle)nging the genome. *EMBO J.* **17**, 6427–6436 (1998).
27. Thomas, J.W. *et al.* Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res.* **12**, 1277–1285 (2002).
28. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
29. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
30. Collins, J.E. *et al.* Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13**, 27–36 (2003).