16. National Toxicology Program *Review of Current DHHS, DOE, and EPA Research Related to Toxicology* (FY 1983) (US Department of HHS, Washington, 1983).
17. Shelby, M. D. & Stasiewicz, S., *Envir. Mutagen.* **6**, 871–876 (1984).
18. Waters, M. D., Garrett, N. E., Covone-de-Serres, C. M., Howard, B. E. & Stack, H. F. *Genetic Toxicology of Some Known or Suspected Human Carcinogens*, in *Chemical Mutagens: Principles and Methods for their Detection* Vol. 8 (ed. de Serres, F. J.) 261–341 (Plenum, New York, 1983).
19. Goldstein, M. & Dillon, W. R. *Discrete Discriminant Analysis* (Wiley, New York, 1978).
20. Lave, L. B., Omenn, G. S., Hefferman, K. D. & Dranoff, G. *J. Am. Col. Toxicol.* **2**, 125–130 (1983).
21. Chankong, V., Haimes, Y. Y., Rosenkranz, H. S. & Pet-Edwards, J. *Mutat. Res.* **153**, 135–166, 1985.
22. Pet-Edwards, J., Chankong, V., Rosenkranz, H. S. & Haimes, Y. Y. *Mutat. Res.* **153**, 187–200, 1985.
23. Williams, G. D., Laspia, M. F. & Dunkel, V. C. *Mutat. Res* **97**, 359, 1982.
24. *National Toxicology Program Tech. Bull.* **1**(3), 2 (1980).
25. Bridges, B. S. in *Screening Tests in Chemical Carcinogenesis* (eds Montesano, R., Bartsch, H., Tomatis, L.) 549–568 (International Agency for Research on Cancer, Lyon, 1976).
26. Marcus, W. L. in *Carcinogenesis, A Comprehensive Survey* Vol. 3: *Polynuclear Aromatic Hydrocarbons* (eds Jones, P. W. & Fruendenthal, R. I.) 469–472 (Raven, New York, 1978).
27. Office of Technology Assessment *Assessment of Technologies for Determining Cancer Risks from the Environment* 135–136 (US Government Printing Office, Washington, 1981).
28. Weisburger, J. H. & Williams, G. M. *Science* **214**, 401–407 (1981).
29. Hoel, D. G., Kaplan, N. L. & Anderson, M. W. *Science* **219**, 1032–1037 (1983).
30. Crump, K. S., Hoel, D. G., Langley, C. H. & Peto, R. *Cancer Res.* **36**, 2973–2979, 1976.
31. Omenn, G. S. in *Molecular and Cellular Approaches to Understanding Mechanisms of Toxicity* (ed Tashjian, A. H. Jr) 224–245 (Harvard School of Public Health, Boston, 1984).
32. National Research Council *Toxicity Testing* (National Academy Press, Washington, 1984).

ARTICLES

Introduction of homologous DNA sequences into mammalian cells induces mutations in the cognate gene

Kirk R. Thomas & Mario R. Capecchi

Department of Biology, University of Utah, Salt Lake City, Utah 84112, USA

Injection of homologous DNA sequences into nuclei of cultured mammalian cells induces mutations in the cognate chromosomal gene. It appears that these mutations result from incorrect repair of a heteroduplex formed between the introduced and the chromosomal sequence. This phenomenon is termed 'heteroduplex induced mutagenesis'. The high frequency of these events suggests that this method may prove useful for introducing mutations into specific mammalian genes.

GENE-TARGETING (homologous recombination between DNA sequences residing in the chromosome and newly introduced DNA sequences) allows the specific alteration of genes in the mammalian genome[1–4]. Recently, Smithies *et al.*[1] introduced a DNA sequence into the chromosomal β-globin locus by homologous recombination. We have corrected mutant neomycin resistance genes (*neo*) in the host genome via homologous recombination with an injected plasmid DNA carrying a different mutation in the *neo* gene[2]. In the process of analysing these gene-targeting events, we uncovered an unexpected class of 'corrected' chromosomal *neo*^{r} genes, which retained the original mutation but had acquired a second mutation that restored gene function. The acquisition of the compensating mutation apparently occurred as a result of incorrect repair of a heteroduplex formed between the *neo* gene in the chromosome and the newly introduced *neo* gene. These reactions occur at a surprisingly high frequency. In addition, we show that the *neo*^{r} genes containing the compensating mutations function in mammalian cells and in *Escherischia coli* as a result of a novel translation reinitiation mechanism.

Recombinant plasmids and recipient cell lines

We first established cell lines containing an amber mutant *neo* gene integrated into the genome of the mouse fibroblast line LMtk$^-$. We then corrected this mutation by injecting plasmid DNA carrying a non-overlapping deletion mutation in the *neo* gene. Cell lines containing a corrected *neo*^{r} gene were identified by selecting for cells resistant to the drug G418.

Figure 1*a* shows the recombinant plasmids, pRH4-14/TK and pRH140ΔNae/TK, used for these experiments. These plasmids were derived from the parental plasmid pRH140[5], which contains sequences from the bacterial plasmid pBR322 and the *neo*^{r} gene coded by the bacterial Tn*5* transposon. The pBR322 sequences supply an ampicillin resistance gene (*amp*^{r}) and an origin of DNA replication which functions in bacteria. The *neo*^{r} gene was engineered to be functional both in bacteria and in mammalian cells[5]. In bacteria, the *neo*^{r} gene confers kanamycin resistance; in mammalian cells the *neo*^{r} gene confers resistance to the drug G418 (G418^r). The herpes simplex virus thymidine kinase gene, HSV-*tk*, was introduced into the above plasmid at the unique *Bam*HI site to generate the plasmid pRH140/TK.

pRH4-14/TK contains an amber codon near the 5′ end of the *neo* gene[5,6]. This premature polypeptide chain termination signal renders the gene product defective in both bacteria and mammalian cells; the mutation also creates a new *Dde*I site which is used as a test for the presence of the amber mutation. pRH140ΔNae/TK contains a deletion of 284 bp at the 3′ end of the *neo* gene which removes 52 amino acids from the carboxy-terminal end of the NEO protein, rendering it non-functional. The plasmid pRH4-14-ΔNae/TK contains both the 4-14 amber mutation and the ΔNae deletion.

We used two recipient mouse cell lines derived from LMtk$^-$ fibroblasts. Both cell lines habour the plasmid pRH4-14/TK, but differ in the copy number and chromosomal location of the integrated plasmid[2]. The cell line LM1 contains a single copy of the plasmid integrated into the host genome by its *Hin*dIII ends at a locus designated *J*-1. Cell line LM4 contains four copies of pRH-14/TK integrated at four independent chromosomal sites designated *J*-1–*J*-4.

Generation of G418^r cell lines

Plasmid DNA, (~5 molecules per cell) was injected into nuclei of LM1 and LM4 cells, which were then selected with G418. Injection of linear pRH140ΔNae/TK molecules into LM1 or LM4 resulted in G418^r cell lines at a frequency of ~1 per 1,000 cells injected (Table 1). This frequency is five orders of magnitude greater than the spontaneous reversion frequency[2] (to G418^r of either cell line). No G418^r cell lines were obtained following injection of LM1 or LM4 with pBR322/TK, pRH4-

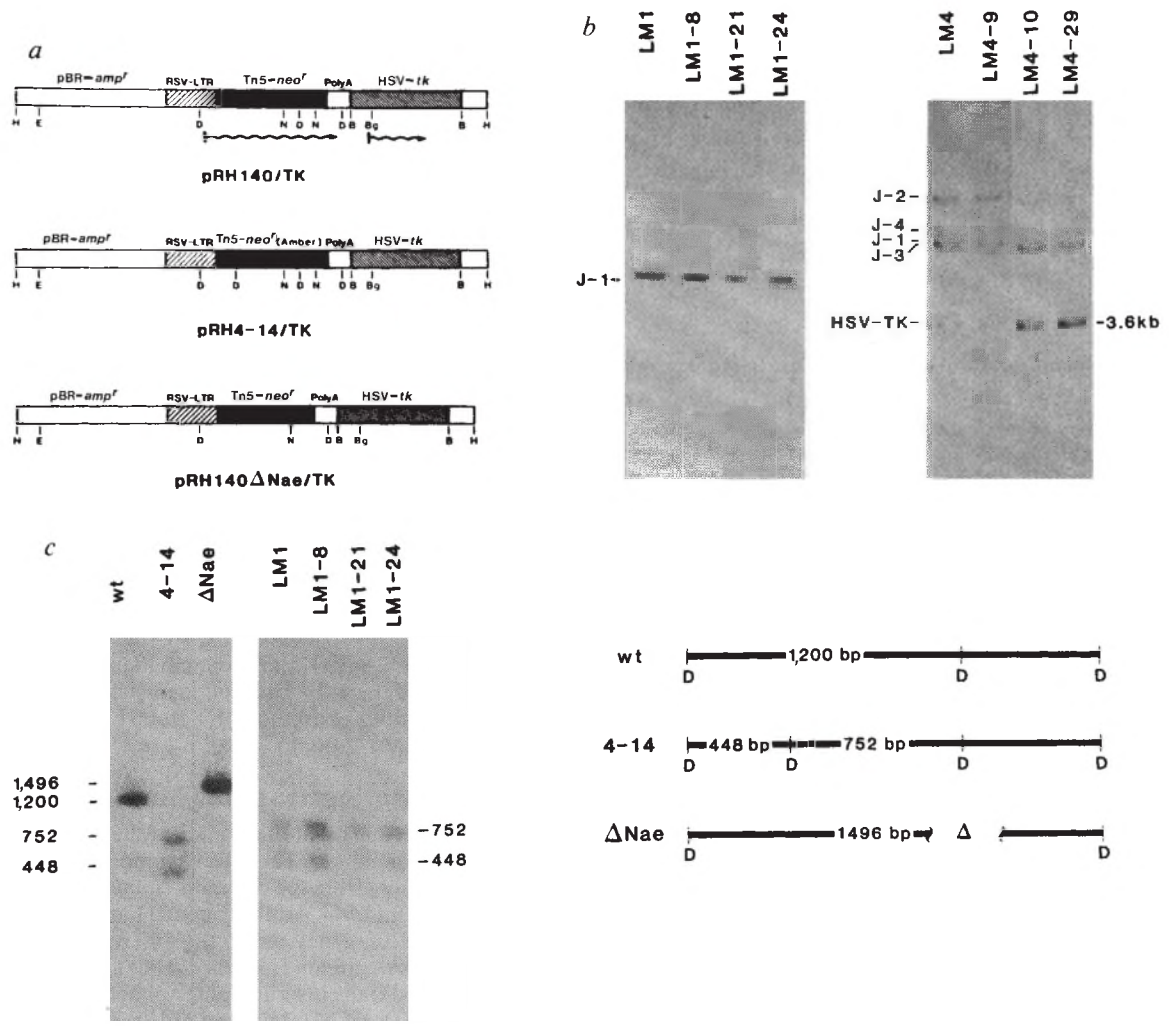

Fig. 1 *a*, Maps of the plasmids pRH140/TK, pRH4-14/TK and pRH140ΔNae/TK. *b*, *Bgl*II and *Bam*HI Southern transfer patterns of LM1, LM4 and their G418^r derivatives. *c*, *Dde*I Southern transfer pattern of LM1, LM1-8, LM1-21 and LM1-24. *a*, The plasmids contain sequences derived from pBR322, the *neo^r* gene coded for by the bacterial Tn*5* transposon and the herpes simplex virus thymidine kinase gene (HSV-*tk*). The *neo^r* gene is expressed from a bifunctional promoter, RSV-LTR, which allows expression of the *neo^r* gene in *E. coli* and in mammalian cells[5]. pRH4-14/TK contains an amber mutation that also creates a new *Dde*I site. pRH140ΔNae/TK contains a 284 base pair deletion at the 3′ end of the *neo* gene. Each vector is represented in linear form from the unique *Hin*dIII site. The restriction sites are designated: H, *Hin*dIII; E, *Eco*RI; Bg, *Bgl*II; N, *Nae*I; B, *Bam*HI and D, *Dde*I. Poly A, polyadenylation sequence from HSV-*tk*. For a more extensive restriction map of these plasmids see refs 2, 6. *b*, DNA was purified from each cell line and digested with either *Bgl*II (LM1, LM1-8, LM1-21, LM1-24) or *Bam*HI (LM4, LM4-9, LM4-10, LM4-29). Aliquots of DNA (5 μg) were electrophoresed through 0.75% agarose, transferred to nitrocellulose paper and probed with either ^{32}P labelled, nick-translated pRH140 (LM1) or pRH140/TK (LM4). kb, kilobases. *c*, DNA (10 μg) from each cell line was digested with *Dde*I, electrophoresed through 0.75% agarose, transferred to nitrocellulose and probed with a ^{32}P-labelled, nick-translated, 1200-bp *Dde*I fragment from the *neo^r* gene of pRH140. The sizes (in base pairs) of restriction fragments generated by *Dde*I digestion of the three *neo^r* alleles (wild type, the 4–14 amber mutation and the ΔNae deletion mutation) are given.

14/TK or pRH-14-ΔNae/TK (see Table 1).

Two classes of G418^r cell lines were obtained from LM1 and LM4. In the first class a wild type *neo^r* gene was generated by homologous recombination between the newly introduced pRH140ΔNae/TK plasmid sequence and the pRH4-14/TK chromosomal sequence. This class has been described previously[2]. The second class became resistant to G418^r by virtue of 'heteroduplex induced (het-induced) mutagenesis'.

Fig. 1*b* shows the Southern transfer patterns of LM1 and LM4 and several of their G418 derivatives. Digestion of genomic DNA with either *Bam*HI or *Bgl*II will isolate the integrated *neo* genes onto single fragments linked to chromosomal sequences. In the case of LM1, *Bgl*II digestion reveals a single hybridizing band representing the single *neo* locus, *J*-1. The *Bgl*II Southern transfer patterns of the G418^r derivative cell lines LM1-8, LM1-21 and LM1-24, are indistinguishable from the parental pattern. Digestion of LM4 DNA with *Bam*HI reveals 4 bands which hybridize to pRH140 sequences. *Bam*HI digestion of DNA from the G418^r derivatives LM4-9, LM4-10 and LM4-29 show identical patterns. Thus conversion to G418^r was not the result of acquisition of new *neo* sequences or rearrangements of old ones.

As shown in Fig. 1*c*, digestion of genomic DNA with the restriction enzyme *Dde*I generates a series of fragments that allow the detection of the wild-type *neo^r* gene and each of the *neo* mutant alleles, the 4–14 amber mutation and the ΔNae deletion mutation. The Southern transfer patterns of three G418^r cell lines derived from LM1 following digestion of genomic DNA with *Dde*I are also shown. Only the *Dde*I fragments characteristic of the 4–14 amber mutation are seen. Similar results were obtained when DNA from LM4 and its derivative G418^r cell lines, LM4-9, LM4-10 and LM4-29 were digested with *Dde*I. These results were quite unexpected because all the derivative cell lines were G418^r.

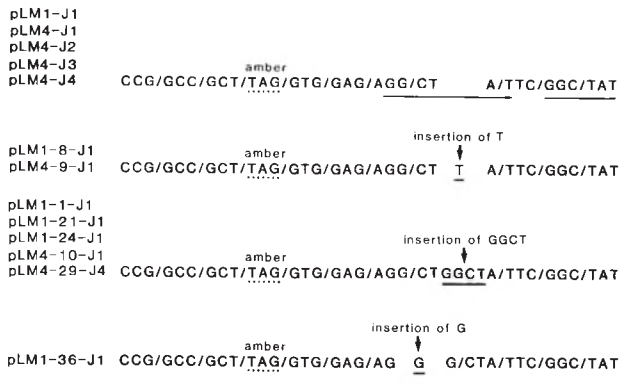

Fig. 2 A summary of pertinent DNA sequences from plasmids rescued from LM1 and LM4 and their G418^r derivatives. Plasmids were rescued from each cell line as described[2]. 5′ fragments of the *neo^r* genes from these plasmids were subcloned into the M13 vector, mp8, and sequenced by the chain termination method[8]. The sequences of the protein coding strands representing codons 12–22 are shown. The amber mutations as well as the 6 base pair repeats are underlined.

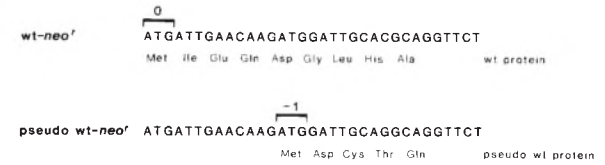

Fig. 3 The amino-terminal sequences of wild-type and pseudo-wild-type *neo^r* proteins are shown with the DNA sequence of the 5′ end of the *neo^r* gene. To facilitate protein purification, hybrid proteins were made by in-frame fusions of the 5′ ends of *neo^r* genes with the *lacZ* gene. Wild-type fusions were made with the *neo^r* gene from the plasmid *p*RH140 (ref. 5); pseudo-wild-type fusions were made with the *neo^r* gene from pLM1-1-J1. Both fusions were constructed by inserting a blunt-ended *Bcl*I–*Pvu*II restriction fragment, containing nucleotides −31 to +233 of *neo^r*, into a blunt-ended *Apa*I site in the 8th codon of the *lacZ* gene in p93.94 (R. Weiss, unpublished). *E. coli* cells containing the hybrid genes were grown to stationary phase, lysed in a French Press, and the extract clarified by centrifugation at 280,000*g* for one h. The β-galactosidase activity was purified either by adhesion to antibodies (mouse anti- β-galactosidase, Promega Biotech), followed by acrylamide gel electrophoresis, or by affinity chromatography[9]. The sequence of the first seven amino acids of each protein was determined on an ABI microprotein sequenator.

Plasmids were rescued from the genome of LM1, LM1-8, LM1-21 and LM1-24. For these experiments genomic DNA was digested with *Bam*HI or *Bgl*II and ligated under conditions that favoured intramolecular ligation. This DNA was used to obtain ampicillin resistant bacteria by transfection[6]. From each cell line we rescued a class of plasmids containing the same 5′ junction with chromosomal DNA. Plasmids derived from LM1, as expected, do not provide resistance to kanamycin, but the plasmids rescued from the G418^r cell lines do. When we examined the *Dde*I polymorphism present in these rescued plasmids, all had the 4–14 *Dde*I polymorphism including the plasmids rescued from LM1-8, LM1-21 and LM1-24. Thus all the plasmids retained the amber mutation, yet those rescued from the G418^r cell lines conferred kanamycin resistance on bacteria.

Plasmids were also rescued from LM4, LM4-9, LM4-10 and LM4-29 after digestion of genomic DNA with either *Bam*HI or *Bgl*II. One plasmid from each G418r line was found to confer kanamycin resistance on *E. coli*. When these plasmids were reintroduced into LMtk$^-$ cells G418^r colonies were obtained at a frequency similar to that resulting from injection of the wild-type *neo^r* gene. Thus these plasmids contain a *neo^r* gene that functions in both bacteria and mammalian cells. The rescued plasmids all had the 4–14 *Dde*I polymorphism, including those that conferred drug resistance (*J*-1 from LM4-9 and LM4-10 and *J*-4 from LM4-29).

Table 1 Transformation frequencies to G418^r by injecting LM1 and LM4 cells with plasmid DNAs

DNA injected	Frequency of G418^r cell lines generated by hetero-duplex induced mutagenesis	Number of cells injected
pBR/TK	0	2×10^4
pRH140ΔNae/TK	1.2×10^{-3}	10^4
pRH4-14/TK	0	3×10^4
pRH4-14-ΔNae/TK	0	2×10^4

LM1 or LM4 cells were grown on glass cover slips (10 mm × 10 mm) in 35 mm Petri dishes. Twenty-five cells per dish received nuclear injections of ~5 plasmid molecules per cell[7]. Plasmid molecules were linearized with *Hin*dIII (pBR/TK) or *Bcl*I (pRH140ΔNae/TK; pRH-14/TK; pRH4-14-ΔNae/TK). After the injection, the cells were incubated for 24 h in nonselective medium at 37° in a 5% CO_2 incubator and then switched to minimum essential medium supplemented with 400 μg ml^{-1} G418. The dishes were scored for colonies after 3 weeks. pBR/TK contains the 3.6 kb *Bam*HI fragment of the herpes *tk* gene inserted at the *Bam*HI site of pBR322. pRH4-14/TK and pRH140ΔNae/TK are described in Fig. 1 legend. pRH4-14-ΔNae/TK was created by the deletion of the 284 bp *Nae* fragment from the *neo^r* gene in pRH4-14/TK.

The compensating mutations are insertions

We sequenced ~300 bp at the 5′ ends of the *neo^r* genes isolated from the G418^r cell lines. These were compared with the corresponding sequences from the plasmids isolated from the parental cell lines LM1 and LM4. As illustrated in Fig. 2, the only changes found in the *neo^r* genes were: the insertion of a thymidine residue 11 bp downstream from the amber mutation in pLM1-8-J1 and pLM4-9-J1; the insertion of four bases, GGCT, in pLM1-1-J1, pLM1-21-J1, pLM1-24-J1, pLM4-10-J1 and pLM4-29-J4; and the insertion of a guanosine residue 8 or 9 bp downstream from the amber mutation in pLM1-36-J1.

Reinitiation of translation

To determine how *neo^r* genes containing both an amber mutation and a frameshift mutation confer *neo^r* activity in both bacteria and mammalian cells, we analysed the amino-terminal sequence of the *neo^r* gene products. We will refer to the gene containing both the amber mutation and an insertion mutation as the pseudo-wild-type gene. To facilitate analysis, the 5′ 200 nucleotides of the *neo* genes were fused to the *lacZ* gene. Hybrid β-galactosidase containing the amino termini from the wild-type and pseudo-wild-type NEO protein containing the amber mutation and the GGCT insertion were purified from transformed *E. coli* and subjected to amino-terminal sequencing. The fusion protein from wild-type *neo^r-lacZ* gene has the sequence Met-Ile-Glu-Gln-Asp-Gly . . . as predicted (Fig. 3). However, the protein from the pseudo-wild-type *lacZ* gene fusion began with the sequence Met-Asp-Cys-Thr-Gln. This protein was initiated from an AUG codon 14 bp downstream from the normal AUG codon in the −1 translation reading frame (see Fig. 3). It now becomes clear how the insertion mutations reverse the amber mutation. After initiation of protein synthesis in the −1 frame, the ribosome passes through the amber mutation, which is in the 0 reading frame and therefore not read, and regains proper phase when it reaches the +1 frameshift mutation (insertion of a T, G or GGCT). Although this alters the amino-terminal amino acid sequence of the NEO protein, this portion of the protein has been shown to be dispensable[10].

Figure 4*a* shows two models that could explain how translation of the pseudo-wild-type gene begins at the −1 AUG codon.

In model I we assume that the ribosome can enter at two sites, AUG codons in the 0 and −1 reading frame. Model II uses a single entry point followed by translation reinitiation. In this model the ribosome enters at the AUG codon in the 0 reading frame and translates the messenger RNA until it reaches the amber codon. There it terminates and releases the polypeptide fragment. The ribosome then scans back until it reaches the −1 AUG codon, where it reinitiates protein synthesis.

We performed two sets of experiments to distinguish between these models. First, the amber mutation was removed, by *in vitro* site-directed mutagenesis, from the pseudo-wild-type gene. Second, a four bp insertion mutation was created *in vitro* in the wild-type *neo*^{r} gene just downstream from the −1 AUG codon. Plasmids containing these altered pseudo-wild-type and wild-type genes were introduced into *E. coli* and cultured mammalian cells to evaluate their ability to confer kanamycin and G418 resistance (see Fig. 4*b*). The two models make opposite predictions of the resultant phenotypes. If the ribosome can enter at either AUG then both the above plasmids should be functional; the ribosomes entering at the −1 AUG codon would regain proper phase on reaching either of the +1 frameshift mutations. However, neither set of plasmids would be functional if only the initial entry site is used because: (1) after entering the pseudo-wild-type mRNA lacking the amber mutation, the ribosome would be unable to terminate and therefore unable to reinitiate protein synthesis; (2) after entering a wild type mRNA containing a four base-pair insertion, the ribosome would encounter a +1 frameshift and generate a mutant gene product. As shown in Fig. 4*b*, neither of these *neo* genes is functional in either bacteria or mammalian cells. These results strongly support the single entry-termination-reinitiation model.

Discussion

The unexpected finding that correction of the pRH4-14/TK sequences in the host genome of LM1 and LM4 often results from insertion of a few base pairs downstream from the amber mutation raises a number of questions. How did the insertions of these base pairs occur? How did injection of pRH140ΔNae/TK into either LM1 or LM4 mediate these events? The DNA sequence surrounding the position of the base-pair insertions contains the six bp direct repeat GGCTAT (see Fig. 2). The fact that each of the insertions, T, G or GGCT is a tandem repeat of part of this sequence suggests a duplication-generating mechanism.

A number of observations argue that the insertion or duplication events at this site were induced by the initiation of recombination between the pRH4-14/TK sequence residing in the chromosome and the incoming pRH140ΔNae/TK sequence. First, the frequency of generating this class of G418^r cell lines was comparable to the frequency of generating cell lines by legitimate gene replacement or gene conversion[2]. Second, this frequency is five orders of magnitude greater than the spontaneous reversion frequency of LM1 or LM4 to G418^r. Third, we previously isolated a cell line in which both a homologous recombination event and a het-induced mutagenic event occurred[2]. Because the homologous recombination events and the mutagenic events each occur at a frequency of ~1 per 1000 cells injected, the predicted frequency of both events occurring independently in the same cell line is 1 per 10^6 injected cells. As we obtained such a cell line after a few thousand injections, it is tempting to postulate that the two events occurred as a result of a concerted reaction. Fourth, we did not obtain G418^r cells from LM1 or LM4 following injection of a plasmid DNA, pBR322/TK, that does not contain sequences similar to the *neo* gene, so injection of DNA does not *per se* induce rampant mutagenesis. Similarly, we did not obtain G418^r cells from LM1 or LM4 following injection of pRH4-14/TK or pRH4-14ΔNae/TK. These experiments demonstrate the specificity of the reaction. Because injection of the pRH4-14/TK vector does not induce the mutations, an important factor for triggering the mutagenic response may be the single base-pair mismatches at the amber mutation and/or the large mismatch at the ΔNae deletion. The observation that injection of the double mutant, pRH4-14-ΔNae/TK, also does not induce mutations argues that at least the single base-pair mismatch is required. The only mechanism that we can envision by which single base-pair mismatches between a chromosomal sequence and an

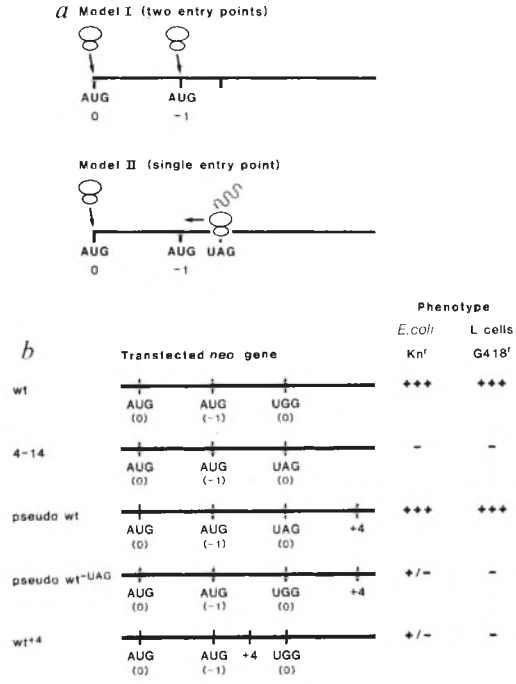

Fig. 4 *a* Two models explaining the synthesis of pseudo-wild-type protein. *b*, Phenotypic analysis of transfected *neo*^{r} genes. *a*, In model I, ribosomes entering *neo*^{r} messenger RNA bind at two binding sites, one on the 0 reading frame at nucleotide 1, the other on the −1 reading frame at nucleotide 14. Proteins made from the latter site would be shifted onto the appropriate reading frame by downstream +1 frameshifts. Model II proposes a single ribosome entry site in the 0 reading frame at nucleotide 1. To use the −1 AUG codon at position 14, the ribosome must first translate in the 0 reading frame up to the UAG (amber) codon. At this point, the ribosome will release the newly synthesized peptide and then scan the message until it locates the alternative AUG codon in the −1 reading frame. *b*, Plasmids containing various *neo*^{r} genes were assayed for the ability to confer drug resistance to *E. coli* or to mouse L cells. The wild type *neo*^{r} gene (wt) from pRH140 contains a tryptophan codon, UGG, at codon 15. 4-14 is the amber mutation from the plasmid pRH4-14 and contains UAG at codon 15. Pseudo-wt is a sequence containing both the 4-14 amber mutation and a downstream, four base-pair insertion of GGCT, following nucleotide 56. Pseudo-wt$^{(-UAG)}$ was derived from the pseudo-wt gene, by removal of the UAG codon by site-directed mutagenesis[11]. The mutant wt^{+4} was derived from the wild-type gene following the insertion of four base pairs at position 35. GGCC was inserted at this position by filling in the ends of an *Xma*III restriction site. The figure depicts the 5′ end of the *neo*^{r} mRNA from each plasmid. AUG (0), the translational initiation codon in the 0 reading frame; AUG (−1), the proposed initiation codon in the −1 reading frame at nucleotide 14; the sites of insertional mutagenesis of GGCC following base 35 or GGCT following base 56 are indicated (+4); UGG and UAG are the alternative states of the fifteenth codon. Plasmids containing these mutations were introduced into *E. coli* strain MH1 by $CaCl_2$-mediated transformation or into mouse L cells by nuclear microinjection. Sensitivity of *E. coli* to kanamycin or of L cells to G418, was determined as previously described[6]. +/−, Resistance to low levels of kanamycin. We believe that this low level of resistance was again due to reinitiation of the −1 AUG as these +1 frameshift mutations generate a new nonsense mutation in the +1 reading frame much further downstream from the − AUG (a UGA codon 53 bp downstream from the −1 AUG codon).

exogeneous sequence can influence the mutation process requires formation of a heteroduplex.

A number of models can be drawn in which the strands in the heteroduplex transiently misalign at the direct repeats leading to partial duplication of this sequence as a consequence of DNA repair. A model mechanism for the insertions found in the LM1 and LM4 G418^r cell lines is given in Fig. 5. A misaligned heteroduplex between the *neo* gene in the chromosome and the injected *neo* gene is formed; a nick is then introduced near one of the loops, generating a primer for strand extension. The condition that the insertion mutation produces a functional gene restricts both the position of the nicks and the number of nucleotides that can be incorporated at the nick. We have shown that insertions that generate a frameshift mutation in the +1 translation reading frame will compensate for the upstream amber mutation. This condition limits the number of bases incorporated at the nick to one base, 4 bases, 7 bases etc. The position of the nicks is also restricted. Nicks at some positions followed by insertion of one or four bp introduce new nonsense mutations. Thus, only a limited set of possible mutations exist which generate a functional *neo^r* gene. After primer extension, the DNA strands are ligated and the insertion fixed into the genome by a round of DNA replication. In prokaryotic recombination and/or DNA repair, reactions homologous to the ligation depicted in the model may not occur; however, such reactions do occur in mammalian cells[12].

We stress that all the insertion mutations that we analysed resulted from independent events. They involve insertions of different bases, in different cell lines, at different chromosomal loci and were generated at different times. It is interesting that all eight insertions occurred in the first repeat. The proximity and/or the nature of the mismatch at the amber mutation may account for this polarity.

The translation reinitiation mechanism described in this paper could be a general mechanism for translating polycistronic messenger RNAs in eukaryotic cells. After the termination of the first protein, the ribosome could scan, forward or backwards, for the AUG codon used to initiate translation of the second protein. Recent reports[13,14] support such a model for translation of a polycistronic messenger. Further, the *src* mRNA of the Rous sarcoma virus and some genes of the cauliflower mosaic virus may be translated by such a mechanism[15,16]. A similar mechanism can also explain the observation[17,18] made with *in vitro* engineered templates, that translation-initiation at an internal AUG codon is stimulated by the presence of a termination codon in frame with the upstream AUG codon.

Conclusion

As a consequence of correcting a gene residing in the host genome by 'gene-targeting', we uncovered an unexpected class of corrected genes. These genes still retained the original mutation but acquired a compensating mutation. The frequency of this event was surprising and was shown to depend not only on homology but also on mismatched base pairs between the newly introduced DNA sequence and the corresponding sequence residing in the genome. All the insertion mutations occurred in a small direct repeat proximal to the base-pair mismatch, suggesting that the direct repeat is also an important component for generating high frequency 'het-induced mutagenesis'. The actual frequency of mutagenesis may be much higher than the observed frequency of 1 per 1000 cells receiving an injection because we imposed the condition that only mutations that converted the amber mutant into a functional gene will be identified. Most such mutations would not be expected to restore gene activity.

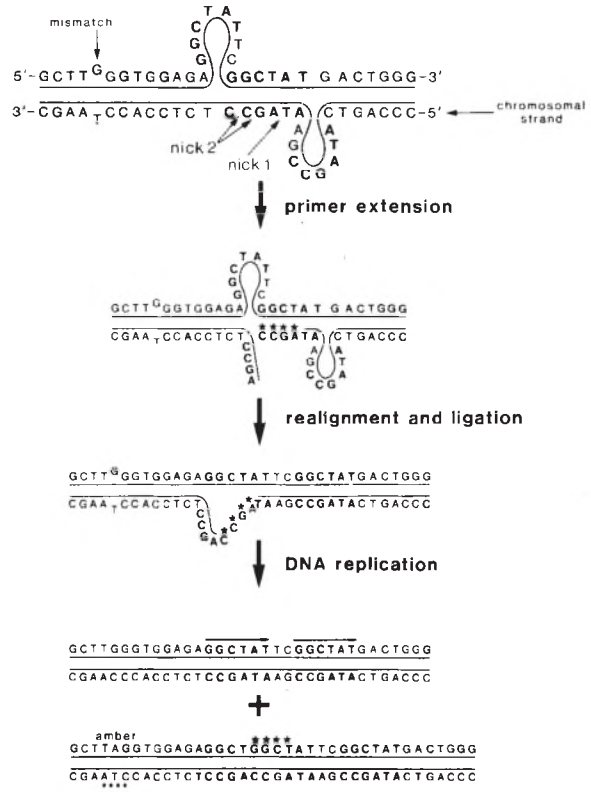

Fig. 5 A model for the mechanism of inserting GGCT into the chromosomal pRH-14/TK sequence. The first step features the formation of a misaligned heteroduplex at the six base-pair direct repeat, between the *neo^r* gene in the chromosome and the injected *neo^r* gene. In the second step a nick is introduced near one of the displaced loops to generate a primer for strand extension. Following extension, the strands are ligated and the insertion fixed into the genome by a round of DNA replication. The positions of the nicks were chosen to account for the insertions isolated: nick 1 would generate either the +T or the +GGCT insertions; nick 2 would generate the +G insertion. The site of the nick and the length of strand extension is limited by the requirement to produce a functional gene. Bold letters, the bases involved in the six base-pair repeats; *, the bases added during strand extension.

In future, we will examine 'het-induced mutagenesis' that directs the loss of gene function rather than the correction of gene function. These events should occur at a higher frequency and reveal a wider spectrum of changes at the DNA sequence level. The only apparent requirements for a 'hot spot' for het-induced mutagenesis are small direct repeats proximal to base-pair mismatches between the newly introduced plasmid sequence and the homologous sequence in the genome. Comparable small direct repeats are encountered in the DNA coding sequence of most genes, making them susceptible to this type of mutagenesis. Permutations of this methodology should provide the means for efficiently introducing mutations into specific mammalian cellular genes.

We thank Laurie Fraser for assistance with tissue culture and Regina Zeikus, Robert Weiss and Diane Dunn for assistance with protein sequencing.

Received 18 July; accepted 2 October 1986.

1. Smithies, O., Gregg, R. G., Boggs, S. S., Koralewski, M. A. & Kucherlapati, R. S. *Nature* **317**, 230-234 (1985).
2. Thomas, K. R., Folger, K. R. & Capecchi, M. R. *Cell* **44**, 419-428 (1986).
3. Lin, F. L., Sperle, K. and Sternberg, N. *Proc. natn. Acad. Sci. U.S.A.* **82**, 1391-1395 (1985).
4. Smith, A. J. H. & Berg, P. *Cold Spring Harb. Symp. Quant. Biol.* **49**, 171-181 (1984).
5. Hudziak, R. *et al. Cell* **31**, 137-146 (1982).
6. Folger, K. R., Thomas, K. R. & Capecchi, M. R. *Molec. cell. Biol.* **5**, 59-69 (1985).
7. Capecchi, M. R. *Cell* **22**, 479-488 (1980).
8. Sanger, F., Nicklen, S. & Coulson, R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463-5467 (1977).
9. Ullman, A. *Gene* **29**, 27-31 (1984).
10. Beck, E., Ludwig, G., Auerswald, E. A., Riess, B. & Schaller, H. *Gene* **19**, 327-336 (1982).
11. Hutchinson, C. A. *et al. J. biol. Chem.* **253**, 6551-6560 (1978).
12. Roth, D. B., Porter, T. N. & Wilson, J. H. *Molec. cell. Biol.* **5**, 2599-2607 (1985).
13. Peabody, D. S. & Berg, P. *Molec. cell. Biol.* **6**, 2695-2703 (1986).
14. Peabody, D. S., Subramani, S. & Berg, P. *Molec. cell. Biol.* **6**, 2704-2711 (1986).
15. Hughes, S. *et al. Molec. cell. Biol.* **4**, 1738-1746 (1984).
16. Dixon, L. K. & Hohn, T. *EMBO J.* **3**, 2731-2736 (1984).
17. Johansen, H., Schumperli, D. & Rosenberg, M. *Proc. natn. Acad. Sci. U.S.A.* **81**, 7698 (1984).
18. Liu, C., Simonsen, C. C. & Levinson, A. D. *Nature* **309**, 82-85 (1984).