nature structural & molecular biology

Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome

Qin Yu[1], Renate König[1], Satish Pillai[2], Kristopher Chiles[1], Mary Kearney[3], Sarah Palmer[3], Douglas Richman[4,5], John M Coffin[3] & Nathaniel R Landau[1]

HIV-1 deleted for the *vif* accessory gene encapsidates the cellular cytidine deaminase APOBEC3G. Upon infection, the encapsidated APOBEC3G induces G→A mutations in the viral reverse transcripts. The G→A mutations result either from C→U deamination of the minus strand or deamination of both strands followed by repair of the plus strand. We report here that minus-strand deamination occurred over the length of the virus genome, preferentially at CCCA sequences, with a graded frequency in the 5′→3′ direction. APOBEC3G induced previously undetected C→T mutations in the 5′ U3 and the primer-binding site, both of which become transiently single-stranded during reverse transcription. *In vitro*, APOBEC3G bound and deaminated single-stranded DNA (ssDNA) but not double-stranded DNA (dsDNA) or DNA-RNA hybrids. We propose that the requirement for ssDNA accounts for the minus-strand mutations, the 5′→3′ graded frequency of deamination and the rare C→T mutations.

HIV-1 requires the accessory protein Vif to replicate in primary lymphocytes, monocytes and 'nonpermissive' transformed human CD4$^+$ cell lines[1–3]. In contrast, in transformed T-cell lines that are termed 'permissive,' Vif is not required for virus replication. The phenotype of nonpermissiveness to Δ*vif* HIV-1 is dominant in somatic cell fusion experiments, suggesting that nonpermissive cells express an inhibitor of Δ*vif* virus replication[4,5]. Sheehy *et al.*[6] have identified the cDNA CEM15, which is expressed only in nonpermissive cells and transfers the nonpermissive phenotype when transferred to a permissive cell. CEM15 is identical to the apolipoprotein B(Apo B) mRNA editing enzyme catalytic polypeptide-like 3G (APOBEC3G). These findings suggest that Vif neutralizes the antiviral activity of APOBEC3G.

Δ*vif* HIV-1 virions are assembled and released from nonpermissive cells, but their infectivity on a per-particle basis is reduced by as much as 1,000-fold (ref. 1). The defective virions can bind and fuse to new target cells. When reverse transcription is initiated to generate viral dsDNA, the infection aborts before integration[2,7,8]. Sequencing of the reverse transcripts has revealed many G→A mutations[9–13]. These mutations presumably result from APOBEC3G-catalyzed C→U deamination of the reverse transcript minus strand. Synthesis of the plus strand from a deaminated minus-strand template would cause plus-strand G→A mutations[9–13].

Δ*vif* virions contain readily detectable APOBEC3G[6,9]. Encapsidation of the enzyme would provide a means for it to be introduced into newly infected cells, where it could deaminate the viral reverse transcripts after their synthesis. In contrast, wild-type virions contain ~100-fold less APOBEC3G[9,14,15]. Thus, Vif seems to prevent the encapsidation of APOBEC3G, protecting the reverse transcripts from deamination. Vif coimmunoprecipitates with APOBEC3G in cotransfected cells, suggesting that the two proteins form a complex[9,14,15]. Vif binding to APOBEC3G induces its ubiquitination and degradation by a proteasomal pathway[14–17]. The degradation is mediated by the interaction of Vif with an Skp1-cullin-F-box (SCF)-like E3 ubiquitin ligase complex containing Cul-5, elongins B and C and Rbx-1 (ref. 16).

APOBEC3G is a member of the mammalian cytidine deaminase family, which in humans includes APOBEC1, APOBEC2 to APOBEC3A–G arrayed in tandem on chromosome 22 and the activation-induced deaminase (AID)[18,19]. APOBEC family members are expressed with a characteristic tissue specificity. APOBEC1 is an RNA-editing enzyme that deaminates C6666 to U in Apo B mRNA in intestinal tissues to generate a premature termination codon[20–22]. In contrast, AID is a B lymphoid protein that acts on DNA, catalyzing C→U deamination of immunoglobulin genes to induce somatic hypermutation and to stimulate class switch recombination[18,23–25]. APOBEC3G is expressed primarily in lymphoid and myeloid cell lineages[18]. In an *Escherichia coli* mutational assay system, APOBEC3G, as well as APOBEC1 and AID, catalyze C→U deamination of bacterial genes[19]. APOBEC3G is a homodimer in which each monomer consists of duplicated catalytic domains joined by a short stretch of linker amino acids[18]. The enzyme contains Zn^{2+} coordination motifs that are characteristic of cytidine deaminases, and recombinant APOBEC3G is active as a cytidine deaminase in an *in vitro* assay[13,19].

[1]Infectious Disease Laboratory, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. [2]Division of Biological Sciences, University of California, San Diego, La Jolla, California 92093-0679, USA. [3]HIV Drug Resistance Program, National Cancer Institute, National Institutes of Health, Frederick, Maryland 21702, USA. [4]Departments of Pathology and Medicine, University of California, San Diego, La Jolla, California 92093-0679, USA. [5]Veterans Administration, San Diego Healthcare System, San Diego, California 92161, USA. Correspondence should be addressed to N.R.L. (landau@salk.edu).

Published online 18 April 2004; doi:10.1038/nsmb758

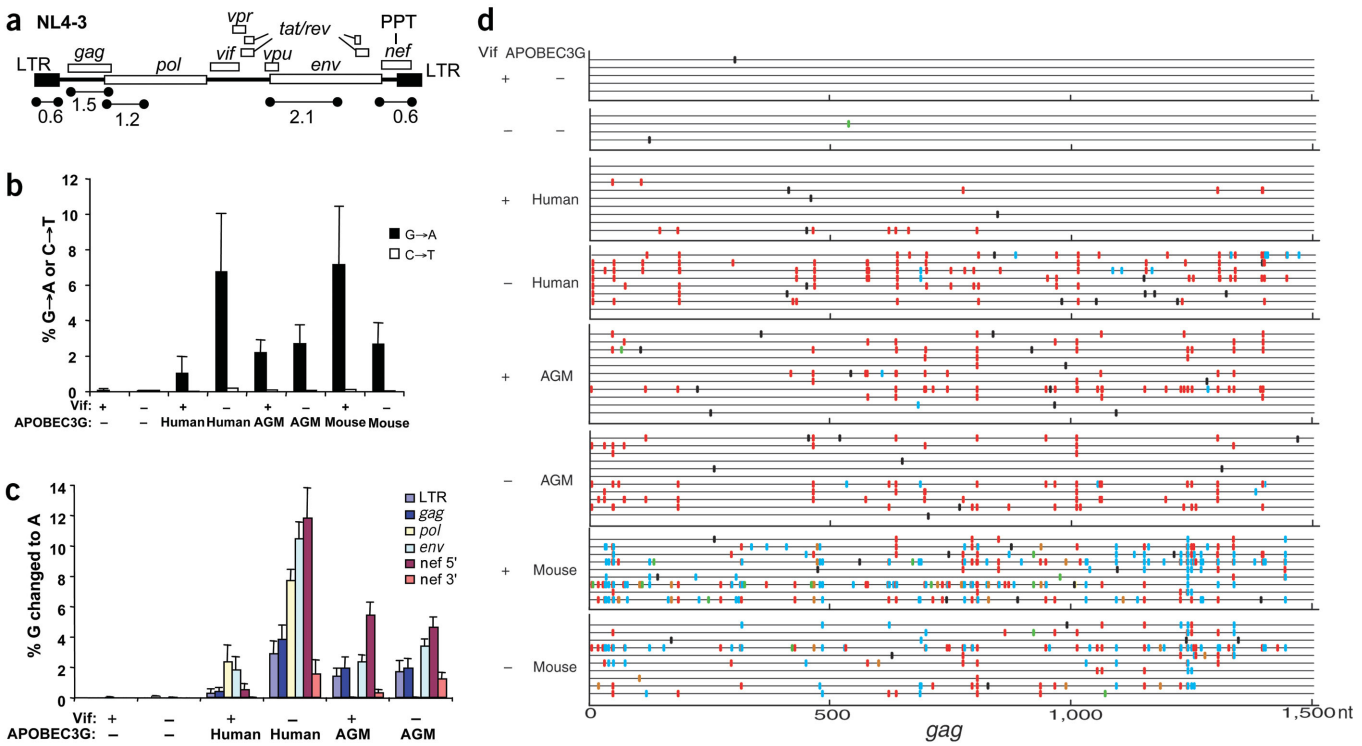

Figure 1 Graded frequency of G→A mutations induced by APOBEC3G. (**a**) Positions and sizes (in kb) of the sequenced viral DNA fragments, which contain the 5′ LTR, *gag*, the 5′ half of *pol*, most of *env* and *nef*. (**b**) The average frequency of G→A and C→T mutations in the five fragments sequenced. The *y*-axis indicates the percentage of the G nucleotides that were changed to A and the percentage of C nucleotides that were changed to T. Error bars represent s.d. across the sequenced fragments. (**c**) The frequency of plus-strand G→A mutation for each of the sequenced regions. The *nef* fragment is presented as two separate portions 5′ and 3′ of the PPT. The *pol* region was not sequenced for the viruses with AGM APOBEC3G. Error bars represent s.e.m. across clones within each treatment. (**d**) Sequences of the *gag* fragment from the different viruses. Mutations are color coded with respect to dinucleotide context: GG→AG, red; GA→AA, cyan; GC→AC, green; GT→AT, orange; non-G→A, black.

The mouse genome contains a single APOBEC3 gene that is ~30% identical to the human homolog. The mouse enzyme is highly active against HIV-1 produced in transfected cells, interfering with both wild-type and Δ*vif* virus replication[9]. APOBEC3G enzymes derived from African green monkey (AGM) and rhesus macaque were also active against HIV-1 (ref. 9). Neither the mouse nor the AGM APOBEC3G form a complex with HIV-1 Vif. The failure of HIV-1 Vif proteins to interact with primate APOBEC3G accounts for their ability to block the replication of wild-type HIV-1 (ref. 9). The species specificity of the interaction results from a single amino acid change between human and primate APOBEC3G: a K128D mutation[26].

To understand the mechanism by which G→A mutations are induced by APOBEC3G, we characterized the mutations that are generated by the enzyme *in vivo* on HIV-1 reverse transcripts and *in vitro* on model substrates. The exclusive induction of G→A mutations could conceivably be caused by specific deamination of the minus strand or by deamination of both strands followed by repair of the plus strand. We found that G→A mutations occur with a 5′→3′ graded frequency over the viral genome. Previously undetected C→T mutations were present in regions of the genome where the plus strand is thought to become transiently single-stranded during reverse transcription. On model substrates *in vitro*, APOBEC3G specifically deaminated ssDNA. We propose a model in which the length of time that each nucleotide remains single-stranded during reverse transcription determines the frequency with which it is mutated.

RESULTS

A 5′→3′ gradient of APOBEC3G deamination

Previous studies have reported sequences over limited portions of the HIV-1 genome[9–13]. To fully characterize APOBEC3G target sequences, portions of the reverse transcripts derived from the 5′ long terminal repeat (LTR), *gag*, *pol*, *env* and *nef* were cloned and sequenced (**Fig. 1a**). At least ten independent nucleotide sequences were determined for wild-type and Δ*vif* viruses that had been produced in the absence or presence of human, AGM or mouse APOBEC3G. Averaging across the genome, the reverse transcripts of Δ*vif* viruses produced in the presence of human, AGM or mouse APOBEC3G contained many G→A mutations. Human APOBEC3G activity was suppressed by HIV-1 Vif, but the AGM and mouse APOBEC3G were not (**Fig. 1b**), in accord with previous reports[9–13]. Vif did not fully suppress G→A mutations at the ratio of provirus to APOBEC3G plasmid used in the transfection. Mouse APOBEC3G induced more G→A mutations in wild-type than Δ*vif* virus. Wild-type virions also encapsidate more mouse APOBEC3G than Δ*vif* virions, a reproducible finding that remains unexplained[9].

Comparison of the G→A mutational frequency in the individual regions of the genome showed an increasing trend in the 5′→3′ direction (**Fig. 1c**). Mutations were least frequent in the 5′ LTR and increased toward *nef*. The percentage of G→A mutation is significantly higher in *pol* versus *gag* ($P < 0.01$) and in *env* versus *pol* ($P < 0.05$), as well as in the 5′ half versus the 3′ half of *pol* and the 5′ half versus the 3′ half of *env* ($P < 1 \times 10^{-7}$ and $P < 0.001$, respectively, using a paired *t*-test; data not shown). Visual inspection of the *nef* sequences

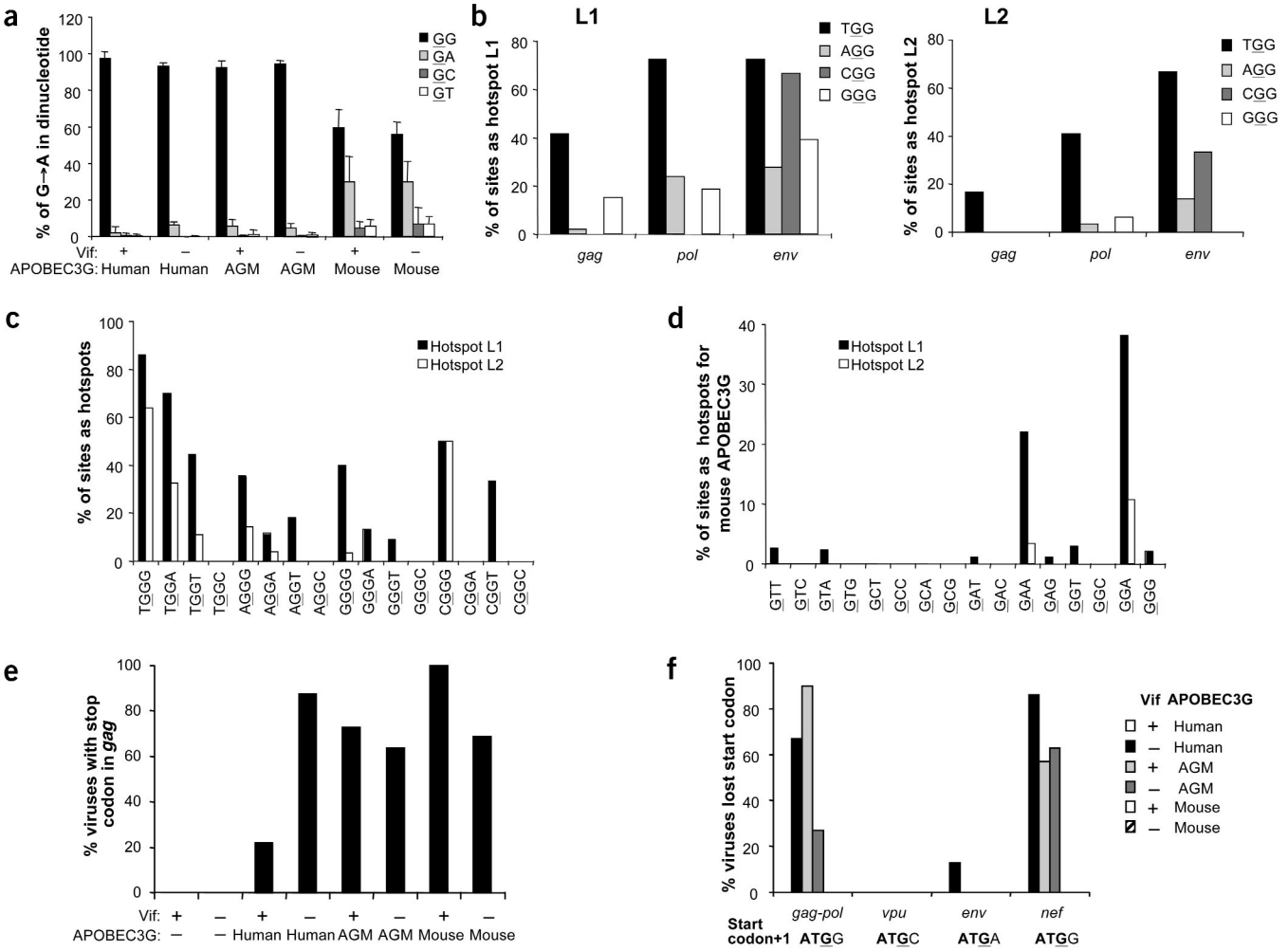

Figure 2 Target sequence preference of APOBEC3G. (**a**) The influence of a 3′ neighboring nucleotide on G→A mutation. Relative frequencies of GG→AG, GA→AA, GC→AC and GT→AT changes are displayed. Error bars represent s.d. across sequenced fragments (5′ LTR, *gag*, *pol*, *env* and *nef*). (**b**) Trinucleotide hotspots of human APOBEC3G deamination in *gag*, *pol* and *env*. Hotspots are defined as G nucleotides that are changed to A in ≥33% (L1) or ≥67% (L2) of viral sequences. All hotspots for human APOBEC3G contain a GG dinucleotide. (**c**) Tetranucleotide hotspots of human APOBEC3G deamination. The 16 tetranucleotides containing a central GG were classified as L1 or L2 hotspots based on mutational frequency. (**d**) Trinucleotide hotspots of mouse APOBEC3G deamination. The percentages of the 16 trinucleotides containing a 5′ G as L1 or L2 hotspots are shown. Nucleotide 5′ of the target G showed minimal influence on G→A mutation (data not shown). (**e**) The percentage of proviruses containing an in-frame stop codon in *gag* caused by G→A mutation. (**f**) The percentage of proviruses in which G→A mutation caused loss of a translational initiation codon. The nucleotide 3′ of the initiator ATG in *gag*, *vpu*, *env* and *nef* is indicated.

showed that the 5′ portion had abundant mutations. There were no mutations in the polypurine tract (PPT) and a low frequency of mutations in the portion 3′ of PPT, overlapping U3. To illustrate these differences, the *nef* sequence was plotted as separate portions 5′ and 3′ of the PPT (**Fig. 1c**). AGM APOBEC3G caused a similar 5′→3′ gradient of mutations but with a lower overall mutational frequency.

Whereas the primary target of APOBEC3G on HIV-1 is the newly synthesized reverse transcripts, virions also contain viral RNA and primer $tRNA^{Lys3}$ that are potential targets. To determine whether the genomic viral RNA or the tRNA primer were targeted, RNA was isolated from wild-type and Δ*vif* HIV-1 virions that had been produced in the presence of human APOBEC3G (termed NL4 hu-APOBEC3G$^+$ and Δ*vif* hu-APOBEC3G$^+$) or absence of APOBEC3G (termed NL4 APOBEC3G$^-$ and Δ*vif* APOBEC3G$^-$). Nucleotide sequence analysis of a 1.5-kb fragment of *pol* for >30 independent amplicons from the four classes of virus showed no significant differences in mutational frequency (NL4 APOBEC3G$^-$, 0.021%; Δ*vif* APOBEC3G$^-$, 0.055%; NL4 hu-APOBEC3G$^+$, 0.013%; Δ*vif* hu-APOBEC3G$^+$, 0.001%). The sequences of 20 independent amplicons of $tRNA^{Lys1,2,3}$ molecules derived from the four classes of virions were identical. We conclude that APOBEC3G is active only against the viral reverse transcripts.

Deamination hotspots

APOBEC3G deaminated nearly 4% of the minus-strand cytosines in the Δ*vif* viral genome but with very different frequencies. Some guanines were changed at frequencies of as high as 100%, whereas others were not changed in any of the clones (**Fig. 1d**). Occasional reverse transcripts of NL4 mouse APOBEC3G$^+$ virus were heavily mutated with a G→A frequency of as high as 24%. To determine the effect of sequence context on the frequency of deamination, we calculated the effect of neighboring nucleotide sequence on targeting frequency.

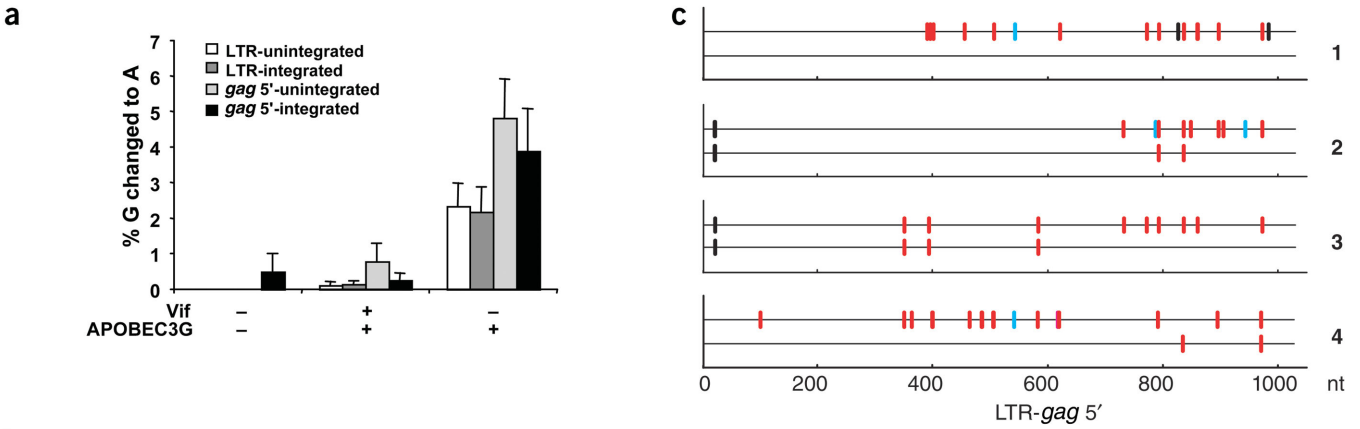

Figure 3 Analysis of integrated proviruses generated by Δ*vif* APOBEC3G+ viruses. Viral and cellular DNA junctions from Δ*vif* APOBEC3G−, NL4 APOBEC3G+ and Δ*vif* APOBEC3G+ were cloned and sequenced. (**a**) G→A mutational frequencies of the integrated and unintegrated viral DNA in the 5′ LTR and *gag* 5′ portion. Error bars represent s.e.m. across clones within each treatment. (**b**) The frequency of plus-strand C→T mutations in the integrated APOBEC3G+ proviruses at positions 20, 21 and 642. Sequence of U3 5′-end and primer-binding site (PBS) are shown with frequently mutated C nucleotides in bold. (**c**) Sequence of Δ*vif* APOBEC3G+ proviruses with noncomplementary DNA strands. Each of the four groups was an independently integrated provirus and yielded two nonidentical sequences. Shown is a region of the provirus (~1 kb) containing the 5′ LTR and 5′ portion of *gag*. GG→AG, red; GA→AA, cyan; non-G→A, black.

Sequences are shown in the plus-strand polarity with the targeted G underlined. For human and AGM APOBEC3G, the targeted G was nearly always flanked by a second G at the 3′ terminus (**Figs. 1d** and **2a**). Mouse APOBEC3G was less stringent, targeting both GG and GA (**Figs. 1d** and **2a**). To further characterize APOBEC3G sequence preference, we defined two categories of mutational hotspots, level 1 (L1) and level 2 (L2), in which a given base was changed in ≥33% and ≥67% of the sequenced clones, respectively. In all of the Δ*vif* hu-APOBEC3G+ sequences, L1 and L2 hotspots always contained GG. The most frequent nucleotide 5′ of the GG in L1 and L2 hotspots was T, indicating that TGG was the preferred trinucleotide target of human APOBEC3G (**Fig. 2b** and **Supplementary Table 1** online). The number of L1 and L2 hotspots increased in the 5′→3′ direction, consistent with the gradient of G→A mutations. Of the 16 possible tetrameric sequences that contain a central GG dinucleotide, TGGG was preferentially targeted by human APOBEC3G (**Fig. 2c** and **Supplementary Table 2** online). A C at the +2 position (NGGC) was strongly disfavored. The mouse enzyme preferentially targeted GGA and GAA without regard for the nucleotide 5′ of the targeted G (**Fig. 2d**). These findings were not the result of overabundance or uneven distribution of target sequences in the HIV-1 genome (**Supplementary Tables 1** and **2** online).

A G→A mutation at TGG generates TAG, TAA or TGA, the three translational termination codons. As a result, nearly all of the viral open reading frames from Δ*vif* hu-APOBEC3G+ virions were prematurely terminated (**Fig. 2e**). In addition, *gag* and *nef* initiation codons were mutated in 60% and 90% of Δ*vif* hu-APOBEC3G+ reverse transcripts, respectively (**Fig. 2f**). The ATG initiation codons of *gag* and *nef* are followed by a G, forming a target TGG. *env* and *vpu* initiation codons are not followed by G and were not targeted. AGM but not mouse APOBEC3G also attacked these initiation codons.

Partial repair of deaminated proviruses

A few completed double-stranded reverse transcripts generated by Δ*vif* APOBEC3G+ virus escape degradation[9]. These may have been protected from deamination or may have escaped by chance. To distinguish these possibilities, a 1-kb fragment of the integrated Δ*vif* hu-APOBEC3G+ virus containing the 5′ LTR and a portion of *gag* along with the flanking genomic sequence was amplified and sequenced. As controls, 23 NL4 hu-APOBEC3G+ and 8 Δ*vif* APOBEC3G− integrated provirus clones were also sequenced. Of these, all except one contained unique flanking human genomic sequence indicative of independent integration events. In contrast, for Δ*vif* hu-APOBEC3G+ virus, some individual clones contained identical integration sites. Of 28 proviruses sequenced, 14 had flanking sequences identical to one of the other clones. These were presumably derived from the same integration event, because independent integrations into exactly the same site of the genome have not been found in large-scale analyses[27]. The mean frequencies of G→A mutations in the proviral sequences of Δ*vif* hu-APOBEC3G+ viruses were similar to those of the unintegrated sequences ($P = 0.44$ for 5′ LTR and $P = 0.30$ for *gag*) (**Fig. 3a**), resulting in frequent loss of the *gag* initiation codon and the accumulation of termination codons (data not shown).

Notably, three previously undetected APOBEC3G-catalyzed C→T mutations were identified in the Δ*vif* hu-APOBEC3G+ proviral sequences. Two of these, C20T and C21T in the 5′ LTR, had been missed in the unintegrated viral DNA because of their proximity to the 5′ end of the viral genome. C21T was found in 43% of Δ*vif* hu-APOBEC3G+ integrated proviruses but was not found in sequences of the integrated APOBEC3G− virus (**Fig. 3b**). The third C→T mutation, 642C→T, was in the primer-binding site. This mutation was also detected in the unintegrated Δ*vif* APOBEC3G+ virus. These were the only three C→T mutations that appeared more than

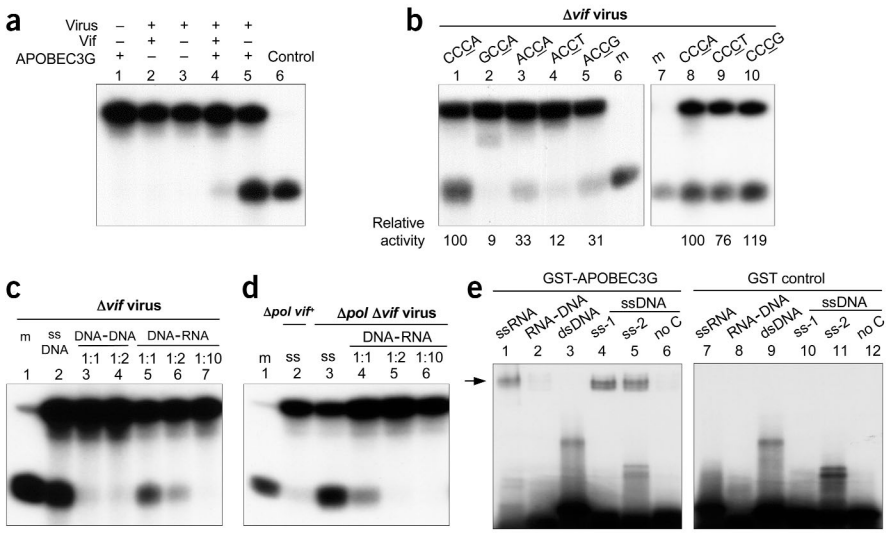

Figure 4 Human APOBEC3G deaminates and binds ssDNA. (**a**) The deaminase activity of APOBEC3G released from pelleted virions was measured on ^{32}P-labeled oligonucleotide model substrates. Mock virus (lane 1) was material pelleted from the supernatant of cells transfected with pcAPOBEC3G but without viral DNA. An oligonucleotide containing a dU in place of the target dC was used as an assay control (lane 6, control). (**b**) APOBEC3G from Δ*vif* hu-APOBE3G$^+$ virus lysate was incubated with deoxyoligonucleotide containing target sequence (marked above lanes). Oligonucleotide containing a CCUA served as a marker and positive control (lane 6). (**c**) Deaminase activity of APOBEC3G on ssDNA (ss), dsDNA (ds) or RNA-DNA hybrids. Labeled deoxyoligonucleotide CCCA annealed to unlabeled complementary DNA or complementary RNA at the indicated molar ratio was incubated with Δ*vif* hu-APOBE3G$^+$ virus lysate. (**d**) The deaminase activity of APOBEC3G released from Δ*pol* viruses. Δ*polvif*$^+$ hu-APOBEC3G and Δ*pol* Δ*vif* hu-APOBEC3G$^+$ and Δ*pol*Δ*vif* hu-APOBEC3G$^+$ virus lysates were tested on ssDNA or RNA-DNA hybrids. (**e**) Binding of APOBEC3G to RNA and DNA. Recombinant GST-APOBEC3G (left) or GST control (right) was incubated with C-rich ssRNA, RNA-DNA, dsDNA, ssDNA (ss-1 and ss-2) and ssDNA lacking C (no C) in standard EMSA. ss-1 DNA contained CCCA and CCA target sequences. ss-2 had the same base composition as ss-1 but was scrambled to avoid APOBEC3G target sequences. Binding was resistant to a 100-fold excess of double-stranded competitor DNA.

once in all of the sequences analyzed. The implication of these unusual plus-strand mutations is considered in Discussion.

On the basis of their integration site, the Δ*vif* hu-APOBEC3G$^+$ clones formed eight groups that each contain two to four members. For four of the groups, all of the clones within each group were identical. For the other four groups, the clones of each group shared an identical integration site but contained two different yet related viral sequences (**Fig. 3c**). These were attributed to sequencing of proviruses with mismatched strands. In group 1, one strand contained G→A mutations, whereas the other was wild type. In groups 2 and 4, one strand contained a set of mutations, whereas the other strand had a subset of these. These findings suggest that the uracil-containing minus strand of the proviruses may become partially repaired either before or after integration.

APOBEC3G is specific for ssDNA

The predominance of G→A mutations could have been caused either by repair of plus-strand C→U deamination or by a requirement of APOBEC3G for ssDNA. In addition, the targeting of hotspots could have been caused by an intrinsic preference of the enzyme for particular target sequences or by extrinsic factors such as pausing during reverse transcription or nonrandom degradation of the plus-strand template by RNase H. To distinguish between these possibilities, we measured APOBEC3G deaminase activity and binding on model substrate oligonucleotides. As a source of native APOBEC3G, we characterized the properties of APOBEC3G released by the lysis of virions. The enzyme was then incubated with 5′-^{32}P-labeled oligonucleotide containing a target C. The oligonucleotide was cleaved at the site of deamination by treatment with uracil DNA glycosylase (UDG) followed by high pH, and the amount of cleaved product was quantified by autoradiography. Control virus lysates were measured to confirm that the assay specifically detected encapsidated APOBEC3G (**Fig. 4a**). Cytidine deaminase activity was not detected in material pelleted from the culture supernatant of cells transfected with APOBEC3G expression plasmid but no viral DNA (lane 1). NL4 APOBEC3G$^-$ and Δ*vif* APOBEC3G$^-$ viruses also lacked cytidine deaminase activity (lanes 2 and 3), whereas NL4 APOBEC3G$^+$ virions contained low but detectable cytidine deaminase activity (lane 4). This finding was consistent with the low frequency of G→A mutations detected in the reverse transcripts of NL4 APOBEC3G$^+$ and the small amount of APOBEC3G encapsidated in such virions. In contrast, Δ*vif* APOBEC3G$^+$ virions generated an intense band (lane 5) of the correct size (lane 6). These data demonstrate that the assay detected encapsidated APOBEC3G.

This *in vitro* assay was used to determine the APOBEC3G target sequence by altering the target CCCA of the oligonucleotide. The results showed that CCCA and CCCG were preferred, GCCA was not targeted, and ACCA, ACCT and ACCG were inefficiently targeted (**Fig. 4b**). This order mirrored the frequency with which the tetranucleotide sequences were changed on the reverse transcripts in infected cells, suggesting that hotspots for APOBEC3G *in vivo* result from the intrinsic specificity of the enzyme for DNA target sequences.

The strand preference of APOBEC3G was determined by comparing the deamination of double- and single-stranded oligonucleotides containing the favored CCCA target sequence. Double-stranded oligonucleotides were formed by annealing unlabeled complementary oligonucleotide to the labeled probe at specific ratios. DNA-DNA hybrids were almost completely resistant to APOBEC3G at a 1:1 ratio of the two strands (**Fig. 4c**). DNA-RNA hybrids were relatively resistant, although deamination was detectable at 1:1 and 1:2 ratios. Virions contain reverse transcriptase RNase H activity that could have partially degraded the RNA in the DNA-RNA hybrids to expose single-stranded oligonucleotide. To remove this potential complication, APOBEC3G$^+$ Δ*vif*Δ*pol* virions that lack reverse transcriptase were tested. Deamination of DNA-RNA hybrids by APOBEC3G derived from these virions was substantially reduced (**Fig. 4d**, lanes 4–6). Notably, the absence of *pol* did not affect the APOBEC3G activity in the virions, indicating that *pol* gene products were not required for APOBEC3G encapsidation.

Binding of APOBEC3G to nucleic acid was measured by electrophoretic mobility shift assay (EMSA), in which recombinant glutathione-*S*-transferase–APOBEC3G (GST-APOBEC3G) or GST

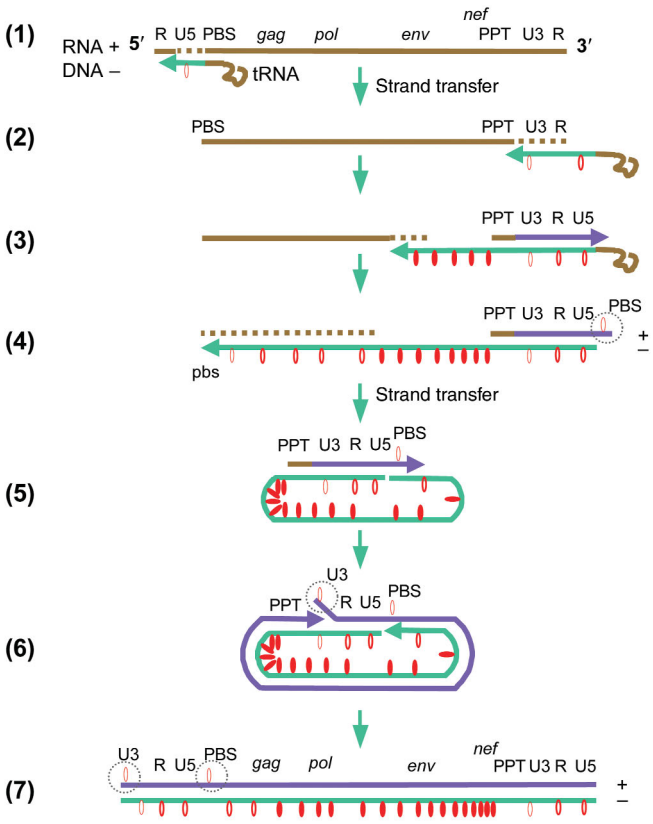

Figure 5 Proposed mechanism of APOBEC3G-mediated deamination of HIV reverse transcripts. Reverse transcriptase extends the annealed $tRNA^{Lys3}$ primer from the primer-binding site (step 1). RNase H degrades the genomic RNA template as it is reverse transcribed. The minus-strand strong-stop DNA is transferred to the 3′ end of the viral RNA and continues to elongate (step 2). APOBEC3G deaminates C→U of the single-stranded minus-strand DNA (red oval). Plus-strand DNA synthesis initiates from the PPT, an RNase H–resistant fragment of the viral genome, and is extended to copy the 3′ 18 nucleotides of $tRNA^{Lys3}$ to regenerate the primer-binding site (step 3). The tRNA is removed by RNase H, causing the plus-strand primer-binding site to become transiently single-stranded and then deaminated (step 4). Transfer of the plus-strand strong-stop DNA to the 3′ end of minus-strand DNA leads to circularization of the DNA strands (step 5) and displacement synthesis (step 6). The 5′ end of the plus-strand DNA becomes transiently single-stranded during the displacement synthesis and is then deaminated (step 6). The linear dsDNA with LTRs is completed (step 7). The RNA is brown; minus- and plus-strand DNA are cyan and green, respectively. The dashed lines indicate RNase H cleavage of the viral RNA. Open and closed red ovals, dC→dU deamination at low and high frequency, respectively. Plus-strand dC→dU events, dotted circles.

was incubated with 5′-^{33}P-labeled single- or double-stranded oligonucleotides and bound complexes were detected by their mobility on nondenaturing PAGE. GST-APOBEC3G bound ssRNA and ssDNA but not dsDNA, and bound poorly to the DNA-RNA hybrid (**Fig. 4e**, left panel). GST control was not shifted, demonstrating that the binding was caused by the APOBEC3G portion of the protein (**Fig. 4e**, right panel). Binding required that the DNA contains C but did not depend upon sequence context.

DISCUSSION

Analysis of the mutations generated by APOBEC3G in HIV-1 DNA in this study has provided insight into the mechanism by which the enzyme deaminates cytosines during reverse transcription. HIV-1 reverse transcripts were deaminated over the length of the viral genome. The frequency of mutations increased over the genome in the 5′→3′ direction, with the lowest frequency in the 5′ LTR and the highest in the 5′ portion of *nef*. The consensus target sequence for human APOBEC3G on model substrates was CCCA/G, and this sequence was preferentially targeted in the viral reverse transcripts. APOBEC3G was active on ssDNA but did not attack DNA-DNA or DNA-RNA hybrids, a property that is shared by the related cytidine deaminase AID[28]. Our findings strongly argue against a model in which both DNA strands are deaminated and then the plus strand is subsequently repaired. Instead, the plus strand is protected because it does not become single-stranded during reverse transcription. The genomic viral RNA and the $tRNA^{Lys3}$ primer were not affected by APOBEC3G.

In light of these findings, we propose a model for APOBEC3G-catalyzed deamination of Δ*vif* HIV-1 reverse transcripts based on the accepted model of retroviral reverse transcription (**Fig. 5**). Reverse transcription is initiated by the extension of the $tRNA^{Lys3}$ primer annealed to the plus-stranded viral genomic RNA to generate minus-strand strong-stop DNA. The minus-strand strong-stop DNA is then translocated to the 3′ end of the genome and further extended to complete the minus strand. As the minus strand is synthesized, the plus-strand RNA template is degraded by the RNase H of reverse transcriptase, exposing single-stranded minus-strand DNA, which is then deaminated by APOBEC3G, preferentially at consensus sequence cytosines. Plus-strand synthesis is primed by RNase H–resistant PPTs. The plus-strand strong-stop DNA is then translocated to the 5′ end of the genome and extended to completion. Because plus-strand synthesis is not accompanied by template degradation, it does not become single-stranded and is therefore protected from APOBEC3G. Thus, the single-strand specificity of APOBEC3G accounts for (i) the preponderance of minus-strand mutations (ii) the 5′→3′ gradient of mutations and (iii) the location of the rare plus-strand mutations.

The proposed model predicts that the probability that a given minus-strand cytosine is deaminated depends upon the length of time that it remains single-stranded. The amount of time that a region of the genome remains single-stranded is determined by the time span between degradation of its plus-strand RNA template and the synthesis of its plus-strand DNA complement. Over most of the genome, this time span is determined by the distance (in number of nucleotides) of the region from the primer. For the region 3′ of the PPT, the situation is more complex. Plus-strand synthesis is thought to initiate from the PPT as soon as its minus-strand copy has been generated, causing the minus-strand sequence 3′ of the PPT to rapidly become double-stranded. The timing of these events is imprinted in the observed mutational frequency. The frequency of G→A mutations increased in the 5′→3′ direction until the PPT was reached, after which it lessened considerably. The PPT itself contained no mutations despite the presence of a string of G nucleotides. The absence of mutations in the PPT suggests that this region is only briefly single-stranded. This suggests that the PPT RNA is not removed until it is displaced by plus-strand DNA synthesis. For HIV, plus-strand synthesis is thought also to initiate from a central PPT, called the cPPT, located near the 3′ end of *pol*[29]. The proposed model predicts that the mutational frequency should also drop 3′ of the cPPT. This region was not sequenced in our study. However, two groups have reported on a clade O HIV-1 isolate that exhibits an extremely high frequency of G→A mutations[30]. Notably, those sequences show a complete protection of the PPT and the cPPT from G→A hypermutation[31] and a pronounced decrease in the frequency of G→A mutations just 3′ of cPPT[30]. It is highly probable that

these mutations were caused by APOBEC3G, because they tend to occur at sequences similar to the APOBEC3G consensus target sequences that we have defined here.

Unexpectedly, two sites of APOBEC3G-induced C→T mutation were found, one in the primer-binding site and another in U3 of the 5′ LTR. Careful consideration of the mechanism of reverse transcription shows that both are in regions in which the plus strand is predicted to become transiently single-stranded during reverse transcription. The plus strand of the primer-binding site becomes single-stranded as it is synthesized and its template RNA, the tRNA primer, is removed by RNase H. The 5′ U3 plus strand would become single-stranded during displacement synthesis (**Fig. 5**, step 6). The 5′ U3 would remain single-stranded for the longest time, because it is displaced first and its complementary strand is synthesized last. The C→T mutation was not in the 3′ U3, suggesting that the deamination occurred on the plus-strand DNA that forms the 5′ LTR.

The mutational frequency of G nucleotides was also affected by sequence context. Categorizing target sequences into L1 and L2 hotspots allowed us to identify the APOBEC3G consensus target sequence as T/CGGG (plus-sense). Deamination hotspots could conceivably be caused by several mechanisms, including (i) an intrinsic preference of the enzyme for a consensus nucleotide target sequence (ii) secondary structure of the viral RNA or DNA (iii) pause sites during reverse transcription that expose single-stranded target sites or (iv) nonrandom RNase H degradation of the plus-strand template to expose ssDNA. Our findings suggest that the primary cause of the differential targeting is the intrinsic sequence specificity of the enzyme. This conclusion is based on the similarity of the sequence preference of the enzyme for model oligonucleotide substrates and HIV-1 reverse transcripts. Oligonucleotides containing CCCG and CCCA were preferred *in vitro*, and the sequence T/CGGG was preferentially changed on the plus-strand of the reverse transcripts.

Evolutionary pressure from APOBEC3G may provide an explanation for earlier findings that noted the skewing of nucleotides and codons in the HIV-1 genome[32,33]. The genome of HIV-1 is A-rich, unlike those of other retroviruses, such as human T-cell leukemia virus type I (HTLV-I). HIV-1 coding sequences are made up of 36% A, and the third base positions of its codons contain 60% A[32,33]. Analysis of the number of APOBEC3G consensus target sequences in the HIV genome shows that the CGGG tetranucleotide but not TGGG tetranucleotide consensus target sequence is underrepresented, resulting in a skewing of codon usage. CGG encodes arginine, which can be encoded by six codons, CGN and AGG/A. The four CGN codons together are used in only 9% of arginine codons. AGA, which is not an APOBEC3G target, is used in 65% of arginine codons, consistent with selection by APOBEC3G. In contrast, HTLV-1, which does not have Vif, is not skewed[32,33] (**Supplementary Fig. 1** online). TGGG is a favored target of APOBEC3G but is relatively abundant in HIV-1. This may be because TGG is the single codon for tryptophan and therefore cannot be selected against without losing protein coding capacity.

Whether APOBEC3G evolved to serve as an antiviral defense mechanism or whether it fortuitously interferes with HIV-1 reverse transcription is not clear. Nevertheless, evolutionary pressure from APOBEC3G seems to have molded the HIV-1 genome. In addition to its role as a cellular antiviral protein that is targeted by Vif, APOBEC3G has served as an unexpected tool to probe the dynamics of the molecular events in retrovirus replication.

METHODS

Plasmids. Δ*vif* HIV-1 (pNL4-3-Δ*vif*) contained two in-frame stop codons near the 5′ end of NL4-3 *vif*[9]. Expression vectors for human, AGM and mouse APOBEC3G were constructed in pcDNA-III (Invitrogen)[9]. To construct Δ*vif*Δ*pol* pNL4-3, the *Age*I-*Eco*RI fragment (nucleotides (nt) 3485–5743) of Δ*vif*-NL4-3 was cloned into the corresponding region pNL4.3-R$^-$E$^-$pol$^-$ (ref. 34).

Viruses and infections. Virus stocks were generated by cotransfection of 293T cells in 10-cm dishes with 8.0 μg wild type or pNL4-3-Δ*vif* and 4.0 μg of APOBEC3G expression vector using Lipofectamine 2000 reagent (Invitrogen). Virus-containing supernatant was harvested 48 h later, filtered through a 0.45-μm filter, treated with DNase to remove contaminating input plasmid and quantified by P24 ELISA.

Sequencing of viral reverse transcripts. HOS.CD4.X4 cells (1×10^6) were infected with viruses containing 10 ng P24. After 4 h, the cells were washed with medium, and at 12 h after infection, DNA was isolated using DNeasy DNA isolation kit (Qiagen). The 5′ LTR (nt 1–654), *gag* (nt 793–2295), *pol* (nt 2088–3355), *vpu* (nt 6070–6302), *env* (nt 6230–8129) and *nef* (nt 8807–9431) fragments were amplified with Expand High Fidelity DNA polymerase (Roche), cloned into the TOPO TA-cloning vector pCR4 (Invitrogen), and the nucleotide sequence was determined (**Fig. 1a**) (see **Supplementary Table 3** online for primer sequences). Sequence data were analyzed using Hypermut[35]. More than 95% of the clones from the Δ*vif* viruses with APOBEC3G and wild-type virus with AGM or mouse APOBEC3G gave unique G→A mutation patterns. The identical sequences from different clones of these samples were judged as amplification of a single viral DNA.

Integration site analysis. HIV-1 integration sites were cloned as described[27]. Briefly, HOS.CD4.X4 cells (1×10^6) were infected with DNase-treated virus (10 ng P24), and free virus was removed by changing the culture medium after 4 h. After 48 h, DNA was purified using DNeasy kit (Qiagen) and cleaved with a mixture of *Xba*I, *Spe*I and *Nhe*1, which produce compatible cohesive ends but do not cleave the NL4-3 LTR or *gag*. A double-stranded oligonucleotide linker with a compatible end was ligated to the cleaved DNA, and the fragments were amplified with Expand High Fidelity polymerase (Roche) using one primer that was complementary to the linker and a second one complementary to the *gag* plus strand (see **Supplementary Table 4** online for oligonucleotide sequences). The amplicons were purified via agarose gel and cloned into a TOPO-TA cloning vector (Invitrogen). The nucleotide sequence was determined using primers in *gag* and LTR.

Cellular sequences at the integration sites were identified using the BLAT search engine of the University of California Santa Cruz human genome browser (http://genome.ucsc.edu/cgi-bin/hgBlat, assembly date July 2003). Integration sites were considered authentic if the sequence joined to the 5′ LTR had a unique best hit in the BLAT search with a >98% match in the database. Independent clones with identical integration sites were considered to be several isolates of a single integration event.

Production of recombinant APOBEC3G. Human APOBEC3G cDNA was amplified from pcAPOBEC3G with primers containing *Eco*RI and *Xho*I sites. The amplicon was cloned into the *Eco*RI and *Xho*I sites of pGEX-6P3 (Amersham Pharmacia) to generate an in-frame fusion to the C terminus of GST. BL21(DE3) bacteria were transformed with the plasmid, and the bacteria were grown in 500 ml of rich medium to an A_{595} of 0.4. The culture was induced with 0.5 mM IPTG and incubated 16 h at 25 °C. The bacteria were pelleted by centrifugation, resuspended in the lysis buffer (10 mM EDTA, 5 mM DTT, 1% (v/v) Triton X-100 and 1 mM PMSF–protease inhibitor cocktail (Sigma) in PBS), and lysed by sonication. The lysate was cleared by centrifugation at 18,000*g* for 30 min and then incubated with glutathione-Sepharose beads (Amersham) for 1 h at 4 °C. The beads were washed three times with the lysis buffer, and bound protein was eluted by rotating for 30 min in 2 ml of elution buffer (20 mM reduced glutathione and 50 mM Tris, pH 8.0) at 4 °C. The protein was >95% pure as judged on Coomassie blue–stained PAGE analysis.

APOBEC3G binding assay. Standard EMSAs were carried out to examine APOBEC3G and RNA-DNA binding. Single-stranded oligonucleotides were 5′-end-labeled with [^{33}P]ATP using T4 polynucleotide kinase (Promega) (see **Supplementary Table 5** online for oligonucleotide sequences). Unincorporated label was removed on a G25 Quick Spin column (Roche). To generate double-

stranded probes, two complementary oligonucleotides were annealed in buffer containing 100 mM NaCl by heating to 90 °C and then gradually cooling to room temperature over 30 min. The probe (0.2 pmol) was mixed with 0.5 μg GST-APOBEC3G in binding buffer containing 10 mM HEPES, pH 7.6, 100 mM KCl, 10 mM $MgCl_2$, 0.1 mM EDTA, 0.5 mM DTT, 10% (v/v) glycerol, with or without 1 μg poly(dI-dC), and incubated at 37 °C for 30 min. The reactions were separated on 4–20% Tris-glycine gradient acrylamide gels (Invitrogen) in Tris-glycine native running buffer at 100 V. The gels were dried and exposed to X-ray film with a phosphor screen. Band intensities were quantified by densitometry.

Cytidine deaminase assay. Virions were generated by transfection of 293T cells, filtered and then pelleted through 2.0 ml of 20% (w/v) sucrose-PBS by ultracentrifugation for 1 h at 150,000*g*. The pelleted virions were lysed in 100 μl buffer containing 50 mM Tris, pH 8.0, 40 mM KCl, 50 mM NaCl, 5 mM EDTA, 10 mM DTT and 0.1% (v/v) Triton X-100. The amount of p24 in the lysates was measured by ELISA. Oligonucleotides containing APOBEC3G target sites were 5′-end-labeled with [^{32}P]ATP and annealed with complementary DNA or RNA oligonucleotides (see **Supplementary Table 5** online for oligonucleotide sequences). The virus lysate (1 μg p24) was mixed with labeled oligonucleotide (1×10^5 c.p.m.) in deaminase buffer (40 mM Tris, pH 8.0, 40 mM KCl, 50 mM NaCl, 5 M EDTA, 10% (v/v) glycerol, 1 mM DTT)[36]. After 4 h at 37 °C, the reactions were terminated by heating to 90 °C for 5 min. The oligonucleotide was purified on a G25 Quick Spin column (Roche) and then treated with UDG for 30 min at 37 °C in UDG buffer (20 mM Tris, pH 8.0, 1 mM DTT) to remove the uracil bases generated by deamination and cleaved at the abasic site by treatment with 0.15 M NaOH at 37 °C for 30 min. The cleaved product was separated on 15% TBE-urea PAGE and detected by autoradiography. A labeled oligonucleotide containing a dU instead of the dC was treated with UDG and NaOH in parallel to serve as a size marker for the deaminated cleavage product.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

We thank D. Chen, R. Mariani and B. Schröfelbauer for laboratory assistance. This work was funded by the US National Institutes of Health (AI058864, DA014494, AI27670, AI38858, AI43638, AI36214 and AI29164), the Universitywide AIDS Research Program of California (IS02-SI-704 and F03-SIBS-215), the Elizabeth Glaser Pediatric AIDS Foundation (EGPAF 28-PF-77491) and the Research Center for AIDS and HIV Infection of the San Diego Veterans Affairs Healthcare System. N.R.L. is an Elizabeth Glaser Scientist of the Pediatric AIDS Foundation.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 2 February; accepted 15 March 2004
Published online at http://www.nature.com/natstructmolbiol/

1. Strebel, K. *et al.* The HIV 'A' (sor) gene product is essential for virus infectivity. *Nature* **328**, 728–730 (1987).
2. von Schwedler, U., Song, J., Aiken, C. & Trono, D. Vif is crucial for human immunodeficiency virus type 1 proviral DNA synthesis in infected cells. *J. Virol.* **67**, 4945–4955 (1993).
3. Gabuzda, D.H. *et al.* Role of *vif* in replication of human immunodeficiency virus type 1 in CD4+ T lymphocytes. *J. Virol.* **66**, 6489–6495 (1992).
4. Madani, N. & Kabat, D. An endogenous Inhibitor of human immunodeficiency virus in human lymphocytes is overcome by the viral Vif protein. *J. Virol.* **72**, 10251–10255 (1998).
5. Simon, J.H. *et al.* The regulation of primate immunodeficiency virus infectivity by Vif is cell species restricted: a role for Vif in determining virus host range and cross-species transmission. *EMBO J.* **17**, 1259–1267 (1998).
6. Sheehy, A.M., Gaddis, N.C., Choi, J.D. & Malim, M.H. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418**, 646–650 (2002).
7. Goncalves, J., Korin, Y., Zack, J. & Gabuzda, D. Role of Vif in human immunodeficiency virus type 1 reverse transcription. *J. Virol.* **70**, 8701–8709 (1996).
8. Simon, J.H. & Malim, M.H. The human immunodeficiency virus type 1 Vif protein modulates the postpenetration stability of viral nucleoprotein complexes. *J. Virol.* **70**, 5297–5305 (1996).
9. Mariani, R. *et al.* Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. *Cell* **114**, 21–31 (2003).
10. Lecossier, D., Bouchonnet, F., Clavel, F. & Hance, A.J. Hypermutation of HIV-1 DNA in the absence of the Vif protein. *Science* **300**, 1112 (2003).
11. Zhang, H. *et al.* The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. *Nature* **424**, 94–98 (2003).
12. Mangeat, B. *et al.* Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* **424**, 99–103 (2003).
13. Harris, R.S. *et al.* DNA deamination mediates innate immunity to retroviral infection. *Cell* **113**, 803–809 (2003).
14. Stopak, K., de Noronha, C., Yonemoto, W. & Greene, W.C. HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. *Mol. Cell* **12**, 591–601 (2003).
15. Marin, M., Rose, K.M., Kozak, S.L. & Kabat, D. HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. *Nat. Med.* **9**, 1398–1403 (2003).
16. Yu, X. *et al.* Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. *Science* **302**, 1056–1060 (2003).
17. Sheehy, A.M., Gaddis, N.C. & Malim, M.H. The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. *Nat. Med.* **9**, 1404–1407 (2003).
18. Jarmuz, A. *et al.* An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics* **79**, 285–296 (2002).
19. Harris, R.S., Petersen-Mahrt, S.K. & Neuberger, M.S. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol. Cell* **10**, 1247–1253 (2002).
20. Yang, Y., Sowden, M.P. & Smith, H.C. Induction of cytidine to uridine editing on cytoplasmic apolipoprotein B mRNA by overexpressing APOBEC-1. *J. Biol. Chem.* **275**, 22663–22669 (2000).
21. Lau, P.P. *et al.* A DnaJ protein, apobec-1-binding protein-2, modulates apolipoprotein B mRNA editing. *J. Biol. Chem.* **276**, 46445–46452 (2001).
22. Anant, S., MacGinnitie, A.J. & Davidson, N.O. apobec-1, the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme, is a novel RNA-binding protein. *J. Biol. Chem.* **270**, 14762–14767 (1995).
23. Muto, T., Muramatsu, M., Taniwaki, M., Kinoshita, K. & Honjo, T. Isolation, tissue distribution, and chromosomal localization of the human activation-induced cytidine deaminase (AID) gene. *Genomics* **68**, 85–88 (2000).
24. Wedekind, J.E., Dance, G.S., Sowden, M.P. & Smith, H.C. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends Genet.* **19**, 207–216 (2003).
25. Petersen-Mahrt, S.K., Harris, R.S. & Neuberger, M.S. AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature* **418**, 99–103 (2002).
26. Schrofelbauer, B., Chen, D. & Landau, N.R. A single amino acid of APOBEC3G controls its species-specific interaction with virion infectivity factor (Vif). *Proc. Natl. Acad. Sci. USA* **101**, 3927–3932 (2004).
27. Schroder, A.R. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
28. Dickerson, S.K., Market, E., Besmer, E. & Papavasiliou, F.N. AID mediates hypermutation by deaminating single-stranded DNA. *J. Exp. Med.* **197**, 1291–1296 (2003).
29. Charneau, P. & Clavel, F. A single-stranded gap in human immunodeficiency virus unintegrated linear DNA defined by a central copy of the polypurine tract. *J. Virol.* **65**, 2415–2421 (1991).
30. Vartanian, J.P., Henry, M. & Wain-Hobson, S. Sustained G→A hypermutation during reverse transcription of an entire human immunodeficiency virus type 1 strain Vau group O genome. *J. Gen. Virol.* **83**, 801–805 (2002).
31. Borman, A.M., Quillent, C., Charneau, P., Kean, K.M. & Clavel, F. A highly defective HIV-1 group O provirus: evidence for the role of local sequence determinants in G→A hypermutation during negative-strand viral DNA synthesis. *Virology* **208**, 601–609 (1995).
32. Berkhout, B. & van Hemert, F.J. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. *Nucleic Acids Res.* **22**, 1705–1711 (1994).
33. Berkhout, B., Grigoriev, A., Bakker, M. & Lukashov, V.V. Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. *AIDS Res. Hum. Retroviruses* **18**, 133–141 (2002).
34. Paxton, W., Connor, R.I. & Landau, N.R. Incorporation of Vpr into human immunodeficiency virus type-1 virions: requirement for the p6 region of gag and mutational analysis. *J. Virol.* **67**, 7229–7237 (1993).
35. Rose, P.P. & Korber, B.T. Detecting hypermutations in viral sequences with an emphasis on G→A hypermutation. *Bioinformatics* **16**, 400–401 (2000).
36. Petersen-Mahrt, S.K. & Neuberger, M.S. *In vitro* deamination of cytosine to uracil in single-stranded DNA by apolipoprotein B editing complex catalytic subunit 1 (APOBEC1). *J. Biol. Chem.* **278**, 19583–19586 (2003).