A genome-wide map of adeno-associated virus–mediated human gene targeting

David R Deyle1,7,8, R Scott Hansen1,8, Anda M Cornea2, Li B Li1, Amber A Burt1, Ian E Alexander3, Richard S Sandstrom4, John A Stamatoyannopoulos4, Chia-Lin Wei5 & David W Russell1,6

To determine which genomic features promote homologous recombination, we created a genome-wide map of gene targeting sites. We used an adeno-associated virus vector to target identical loci introduced as transcriptionally active retroviral vectors. A comparison of ~2,000 targeted and untargeted sites showed that targeting occurred throughout the human genome and was not influenced by the presence of nearby CpG islands, sequence repeats or DNase I–hypersensitive sites. Targeted sites were preferentially located within transcription units, especially when the target loci were transcribed in the opposite orientation to their surrounding chromosomal genes. We determined the impact of DNA replication by mapping replication forks, which revealed a preference for recombination at target loci transcribed toward an incoming fork. Our results constitute the first genome-wide screen of gene targeting in mammalian cells and demonstrate a strong recombinogenic effect of colliding polymerases.

Homologous recombination is a fundamental biological process required for meiosis, DNA repair and gene targeting. During meiosis, recombination rates can vary substantially at different chromosomal loci1; however, the effects of chromosomal position on gene-targeting frequencies are not as well characterized. In yeast, dispersed target loci present at different chromosomal sites are targeted at similar frequencies2,3, whereas in mammalian cells some alleles of target loci can be targeted at higher frequencies than others4, and recombination hotspots have been identified, such as the hotspot present in the mouse IgH locus5,6. In addition, both transcription and DNA hypomethylation can increase targeting frequencies in mammalian cells7,8, suggesting that localized variation in these processes could influence targeting at different loci. A better understanding of how targeting frequencies vary across the genome may lead to insights into DNA-recombination mechanisms as well as improvements in the ability to manipulate mammalian genomes.

We previously used adeno-associated virus (AAV) vectors to study position effects on human gene targeting9. AAV vectors have single-stranded, linear DNA genomes that efficiently recombine with homologous chromosomal sequences, with up to 1% of infected cells undergoing gene targeting under optimal conditions10. Although these targeting frequencies can be orders of magnitude higher than those typically obtained in human cells by transfection or electroporation11,12, both processes share common features, including stimulation by double-strand breaks13,14, involvement of the same homologous-recombination proteins15 and similar effects of mutation type on targeting frequencies16. In a previous study, when identical target sites were introduced at 16 different chromosomal positions in HT-1080 human fibrosarcoma cells, their AAV-mediated gene targeting frequencies varied by as much as one log2. This study demonstrated clear position effects on human gene targeting, but the number of targeted sites was too low to draw meaningful conclusions regarding the effects of surrounding sequences on gene targeting.

In the work described here, we set out to determine which genomic elements influence homologous recombination by creating a genome-wide map of AAV-mediated gene targeting. We introduced identical target sites at thousands of chromosomal positions with retroviral vectors, used an AAV gene targeting vector to correct a neomycin-resistance gene mutation in these targets and mapped the chromosomal locations of each target site by high-throughput DNA sequencing. A comparison of the positions of these targeted sites to a set of untargeted control sites allowed us to determine whether neighboring sequences can influence gene-targeting frequencies and how transcription and replication affect the process.

RESULTS

Genome-wide mapping of gene-targeting sites

We used a retroviral shuttle-vector system to introduce, rescue and map target loci that had undergone AAV-mediated gene targeting (Fig. 1a). The murine leukemia virus (MLV) vector LHSN63A53O contains a nonfunctional neomycin phosphotransferase target gene (neo) with a 53-bp deletion at nucleotide 63 of its open reading

1Department of Medicine, University of Washington, Seattle, Washington, USA. 2Department of Molecular and Cellular Biology, University of Washington, Seattle, Washington, USA. 3Gene Therapy Research Unit, Children’s Medical Research Institute, Westmead, New South Wales, Australia. 4Department of Genome Sciences, University of Washington, Seattle, Washington, USA. 5Genomic Technologies Department, Joint Genome Institute, Walnut Creek, California, USA. 6Department of Biochemistry, University of Washington, Seattle, Washington, USA. 7Present address: Department of Medical Genetics, Mayo Clinic, Rochester, Minnesota, USA. 8These authors contributed equally to this work. Correspondence should be addressed to D.W.R. (drussell@u.washington.edu).

Received 31 March; accepted 27 August; published online 5 October 2014; doi:10.1038/nsmb.2895
Figure 1 Genome-wide gene targeting. (a) Experimental design, with top right schematic showing the structures of the AAV2-HSN5′ targeting vector with a neo gene truncated at bp 629, MLV-LHSN63A53O target-site provirus containing a 53-bp deletion at bp 63 of neo and control vector MLV-LHSNO with a wild-type neo gene. Locations of the AAV inverted terminal repeats (ITR), retrovirus long terminal repeats (LTR), simian virus 40 (SV40) and Tn5 promoters, transcriptional start sites (arrows), hph and neo genes and p15A replication origin are indicated. (b) Localized targeted (n = 2,015) and untargeted (n = 1,928) provirus sites graphed per chromosome as a percentage of all mapped sites. There were no significant differences (P > 0.05, chi-square test). (c) Locations of mapped sites, shown as red (targeted) or blue (untargeted) circles adjacent to each human chromosome ideogram.

frame, a hygromycin phosphotransferase gene (hph) for selection and a plasmid replication origin for recovering the target sites. The AAV2-HSN5′ gene targeting vector contains sequences homologous to MLV-LHSN63A53O with a truncated neo gene that lacks the 53-bp deletion. We transduced diploid HT-1080 human fibrosarcoma cells with MLV-LHSN63A53O and selected the resulting hygromycin-resistant cells as a polyclonal population. We then transduced these cells with AAV2-HSN5′ and cultured them in G418 to select for targeting events. We obtained 7,950 G418-resistant, targeted clones and expanded them as a polyclonal population for target-site analysis. Assuming that all G418-resistant clones were targeted, the targeting frequency was 0.054% of infected cells. To generate an appropriate untargeted control data set, we transduced HT-1080 cells with the MLV vector LHSNO, which is identical to MLV-LHSN63A53O except that it contains a functional neo gene. We cultured HT-1080 cells transduced with MLV-LHSNO in G418 and expanded this polyclonal population of approximately 75,000 independent clones for analysis.

We isolated genomic DNA from both the targeted and untargeted populations of G418-resistant cells, digested it with specific restriction enzymes to release the bacterial plasmid and flanking genomic sequences of each provirus, circularized these fragments and transformed Escherichia coli to obtain kanamycin-resistance plasmids (Fig. 1a). We amplified individual plasmid-containing bacterial colonies separately to prevent overgrowth of specific clones and then combined them before plasmid purification and DNA sequencing (summary of sequencing results in Table 1). We sequenced 9,000 plasmids rescued from both gene-targeted cells and untargeted control cells with an Illumina Genome Analyzer and obtained more than 12,000,000 reads of 75 nt in each case. After screening for reads containing a portion of the MLV long terminal repeat (LTR) and eliminating duplicates, we found a total of 2,015 targeted and 1,928 control proviruses that uniquely localized to the February 2009 assembly of the human genome from targeted and untargeted cells, respectively (Supplementary Table 1). The rescued proviruses were present in all human chromosomes (Fig. 1b; cytogenetic distributions of all mapped sites in Fig. 1c). There were no significant differences in the percentage of sites per chromosome, although too few sites were present on the Y chromosome for a meaningful analysis. Sequencing also identified targeted and untargeted sites that mapped to repetitive regions of the genome and could not be uniquely localized (Table 1).

We further characterized these ‘ambiguous’ sites on the basis of the type of repetitive DNA they contained and found a similar distribution among targeted and untargeted sites (Supplementary Table 2).

Effects of flanking genomic elements on targeting

To determine whether the chromosomal landscape could influence targeting frequencies, we compared the proportion of targeted and untargeted sites found within and near specific types of genetic elements.
There were no significant differences in the percentages of sites found within repetitive sequences, including short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), LTRs, DNA repeats, simple repeats and microsatellite repeats (Table 2). Similarly, there were no significant differences in the percentages of sites located within 10 kb of these repeat elements, even when binned by increasing distance (Supplementary Fig. 1). We focused specifically on the GT dinucleotide–repeat subset of microsatellite repeats because our previous analysis of 16 target sites suggested that these elements might increase targeting frequencies, but in this more comprehensive study they had no significant effect on targeting (Table 2). CpG islands and DNase I–hypersensitive sites also had no influence on targeting frequencies (Table 2), although both targeted and untargeted MLV proviruses were preferentially located near these elements (Supplementary Fig. 1), in agreement with prior studies of gammaretrovirus integration profiles.

Convergent transcription increases targeting

Gammaretroviral vectors preferentially integrate near the transcription start sites of active genes, and the majority of provirus sites in our experiment were located within RefSeq transcription units (Table 2). There was a slight but statistically significant difference between targeted and untargeted sites (58.6% versus 54.1%, respectively, found in genes), showing that proviruses embedded within chromosomal transcription units were targeted at higher frequencies. Both targeted and untargeted sites were preferentially located near transcription start sites, with no significant differences (Fig. 2a). We explored the role of transcription further and determined the baseline expression levels of provirus-containing genes by global transcription analysis of infected HT-1080 cells, using Illumina HumanHT-12 v3 microarrays. 810 of the 1,180 intragenic targeted sites (69%) and 724 of the 1,042 intragenic untargeted sites (70%) were represented on the array. The median expression level of genes containing targeted sites was significantly higher than that of genes containing untargeted sites (8.39 versus 7.98 arbitrary units, respectively; \(P < 7.98 \times 10^{-5}\)), both of which were higher than the median expression level of the full transcript set represented on the array (6.69 arbitrary units). We binned these provirus-containing genes into subsets according to their expression level, and we detected a statistically significant increase in the percentage of targeted sites found in the most highly expressed genes (Fig. 2b).

Because the provirus–target site contained both LTR and SV40 promoters that were presumably expressed at the time of targeting, the preference for targeting sites in active chromosomal genes could not simply be explained by transcription at the target site. Instead, we hypothesized that opposing transcription units might stimulate targeting if colliding RNA polymerases somehow exposed single-stranded regions of chromosomal DNA. When we determined the orientation of each RefSeq gene in relation to that of its embedded provirus, we found that 60.1% of targeted sites were in opposite orientation, as compared to 48.8% of untargeted sites (\(P = 1.4 \times 10^{-7}\)). Binning these genes by expression level showed that this effect was mainly due to preferential targeting of sites embedded in highly expressed chromosomal genes that were transcribed in the opposite direction (Fig. 2c). These data demonstrate that overlapping, convergent transcription at target sites stimulates homologous recombination, and they suggest that in some cases both genes were transcribed at the same time, because a nonspecific targeting enhancement of transcription-related chromatin opening should not depend on transcription direction.

Replication-fork direction influences gene targeting

Homologous recombination occurs in S and G2 phases and is coordinated with DNA replication. To assess the role of replication in gene targeting, we determined the locations of 4,173 replication initiation zones and the directions of replication forks throughout the genome of HT-1080 cells with the Repli-Seq method. We pulse-labeled newly replicated DNAs in exponentially growing cells with 5-bromo-2-deoxyuridine (BrdU) and then sorted the cells into six cell-cycle subsets (late G1, four subsets of S phase, and early G2) by flow cytometry. We then isolated the BrdU-labeled, newly replicated DNA strands from each cell-cycle population, sequenced them by Illumina-based methods for massively parallel sequencing and mapped them to the human genome. We used uniquely mapped sequences to calculate local signal densities for each of the cell-cycle fractions from which we determined average replication-time values at each genomic coordinate (Online Methods). As an example, we show results of Repli-Seq analysis for the entire chromosome 1, aligned with the locations of targeted and untargeted provirus sites, and a set of computer-generated random chromosomal positions (Fig. 3a). The patterns of newly
replicated DNAs found in each cell-cycle subset reflect replication timing, allowing initiation zones and fork movement to be established across the genome. As an example, we show an expanded portion of chromosome 1 with fork direction indicated as well the locations and transcriptional orientation of target sites (Fig. 3b).

In comparison to random chromosomal positions, both targeted and untargeted sites were more likely to be present in early-replicating regions (Fig. 4a) and closer to initiation zones (Fig. 4b). This can be explained by the preferential integration of gammaretroviral vectors in expressed genes17, which are known to replicate earlier in S phase23. A similar correlation could also account for the slightly earlier replication timing of targeted sites in comparison to untargeted sites, because high chromosomal gene expression favors targeting (Fig. 2b). Targeted sites were more likely to be transcribed in the opposite direction to replication fork movement, although this had no effect on untargeted sites (Fig. 4c). When we performed this analysis on the subset of forks defined by peaks and valleys with replication-time differences of at least 10%, we found that 63.3% of targeted and 50.3% of untargeted sites were transcribed in the opposite orientation. Limiting the analysis to forks with >20% or >30% peak-valley timing differences improved the accuracy of fork-direction calls and increased the proportion of sites transcribed in the opposite orientation to over 70% for targeted sites (Fig. 4d). As noted above for overlapping transcription units (Fig. 2c), these results suggest that colliding polymerases stimulate gene targeting. This was also true when we limited our analysis to the subset of sites found in chromosomal genes ≥500 kb in length (Supplementary Fig. 2), which require more time to complete transcription and therefore must be transcribed during replication25. In addition, this analysis showed that long chromosomal genes are preferentially transcribed in the same direction as replication (Supplementary Fig. 2a), suggesting that the genome may have evolved to avoid these obligate head-on polymerase collisions.

We previously isolated HT-1080 subclones containing distinct, mapped provirus-target sites and determined the targeting frequency at each site. (Examples of these target sites with their corresponding Repli-Seq patterns are shown in Fig. 5.) 14 sites were located in regions where replication fork direction could be unambiguously established (Supplementary Table 3), thus allowing us to determine the impact of this directionality on targeting frequencies at specific sites. On average, sites transcribed in the opposite direction of fork movement had targeting frequencies approximately two times higher.

**DISCUSSION**

In this report, we describe a genome-wide analysis of gene targeting in mammalian cells and expand the region of the human genome that has been targeted. We introduced thousands of identical target sites at different chromosomal locations, mapped the locations of over 2,000 targeted loci and compared these genomic positions to a control set of untargeted sites. Although several prior studies have reported stimulatory effects of neighboring repeat sequences to homologous recombination in mammalian cells25–28, we did not observe a significant effect of sequence elements on target frequencies, including several of the same types of repeats. This may reflect differences in experimental design, the fact that all the chromosomal sequence elements flanked an identical target site or the behavior of specific target loci that could not be reproduced on a genome-wide scale. Instead, we detected consistent differences in targeting, due to
Transcription increases targeting frequencies in mammalian cells\(^7\), and we observed preferential targeting at sites present within transcriptionally active chromosomal genes. However, in our system there were no completely silent sites because every target locus also contained an actively transcribed neo gene to facilitate its recovery as a shuttle vector. Instead, we observed increased targeting when the target neo gene was transcribed in the opposite direction to that of its surrounding chromosomal gene. Although this type of ‘convergent transcription’ has not been directly linked to homologous recombination, the topological consequences are similar to those of collisions of RNA and DNA polymerases discussed below. In addition, convergent transcription through repetitive sequences may induce genotoxicity via the ataxia telangiectasia–mutated and Rad3-related (ATR) signaling pathway\(^29\–31\), which could promote recombination during the DNA repair process. Also, in yeast, cohesin accumulates at sites of convergent transcription\(^32\), and this could promote sister-chromatid pairing and recombination\(^33\). The same phenomenon may also induce chromosomal recombination events, given that overlapping transcription units are common in the human genome\(^34\).

DNA replication had the greatest impact on gene targeting in our experiments, as indicated by correlating target sites with a genome-wide map of replication fork movements. Proximity to replication initiation zones had no significant effect, but target sites transcribed in the opposite direction from the incoming replication fork were preferentially targeted. The magnitude of this effect may have been underestimated because it is difficult to reliably assess replication fork direction at some positions, and there could be subsets of cells with variations in fork movements\(^35\). However, the targeting frequencies measured at 14 mapped sites showed that opposing fork direction doubled targeting on average (Fig. 5c); this was consistent with the genome-wide data showing that ~70% of targeted sites were transcribed in the opposite direction from fork movement when fork direction was limited to high-confidence calls (Fig. 4d). Our results suggest that human gene targeting is mechanistically related to transcription-associated recombination (TAR), which in the case of yeast has been attributed to head-on collisions between replication forks and all three types of RNA polymerases\(^36\–38\).

We present a model for how convergent replication and transcription might stimulate AAV-mediated gene targeting (Fig. 6a), based on current knowledge of the chromosomal structures produced by head-on collisions\(^39\–41\). The advancing polymerases increase superhelical strain, stall replication forks and expose regions of single-stranded DNA\(^42\,43\) that could pair with AAV vector genomes in three ways. First, R loops containing RNA–DNA heteroduplexes leave one strand unpaired and are known to be recombinogenic\(^44\,45\). Second, ‘chicken-foot’ structures formed by fork collapse can be processed by an exonuclease such as EXO1 (ref. \(46\)) to expose single-stranded regions. And third, the chicken-foot structure could reopen, leaving a single-stranded gap on the lagging strand of the fork\(^46\). Once paired, the AAV vector genome could introduce site-specific sequence changes into the chromosomal template through further recombination and/or repair processes. These alternative possibilities are not mutually exclusive, but they can be distinguished in part by which vector strand pairs with the genome: R-loop pairing occurs with the antisense vector strand, whereas chicken-foot and lagging-strand pairing occurs with the sense vector strand.

AAV virions contain single-stranded DNA genomes of either orientation\(^47\), so we could not determine which strand participated in the recombination reaction. However, autonomous parvoviruses such as minute virus of mice (MVM) contain distinct left and right termini, and they package only one strand orientation into virions\(^48\,49\). We previously used MVM vectors to demonstrate an approximately one-log difference in targeting frequencies depending on which strand was packaged\(^50\). In this system, we introduced a mutant human placental alkaline phosphatase (ALPP) reporter gene into HT-1080 cells with a reporter gene into HT-1080 cells with a


Figure 6 | Stalled replication forks may promote vector pairing at target loci. (a) Model of a target locus transcribed in the opposite direction from an incoming replication fork, which stalls the incoming fork and produces a chicken-foot structure. This exposes single-stranded regions in the target locus in three possible ways, with distinct consequences for vector pairing. (b) An MVM targeting system, shown with vectors containing sense or antisense targeting strands (MVM-s and MVM-as) that can correct an ALPP (Alk Phos) reporter gene with a 4-bp deletion at bp 375 of its reading frame that was introduced by gamma retroviral vector MLV-LAP375Δ4SP with puromycin selection50. Arrows indicate transcription start sites. The HT-1080 Repi-Seq data for this portion of human chromosome 9 is shown below the vector maps, with the target site located at bp 114505235. MSCV, murine stem-cell virus promoter; SV40, SV40 viral promoter; GFP, green fluorescent protein gene; puro, puromycin-resistance gene. (c) Average targeting frequencies (with s.d.; n = 3 independent experiments) of the sense and antisense MVM vectors when infections were done at the indicated multiplicities of infection50. The mixture contained 2.5 × 10^5 vector genomes per cell of both MVM-s and MVM-as.
microarray analysis and data collection. L.B.L. mapped the JVM vector target site. J.A.S. and I.E.A. provided support for the project. D.R.D., R.S.H., A.M.C. and D.W.R. analyzed the data and wrote the manuscript. All authors commented on the manuscript. D.W.R. supervised the project.

COMPE T I N G F I N A N C I A L I N E R E S T S
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Kong, A. et al. A high-resolution recombination map of the human genome. Nat. Genet. 37, 241–247 (2005).
2. Wilson, J.H., Leung, W.Y., Bosco, G., Dieu, D. & Haber, J.E. The frequency of gene targeting in yeast depends on the number of target copies. Proc. Natl. Acad. Sci. USA 91, 177–181 (1994).
3. Gray, M. & Honigberg, S.M. Effect of chromosomal locus, GC content and length of homology on PCR-mediated targeted gene replacement in Saccharomyces. Nucleic Acids Res. 29, 5156–5162 (2001).
4. Yáñez, R.J. & Porter, A.C. A chromosomal position effect on gene targeting in human cells. Nucleic Acids Res. 30, 4892–4901 (2002).
5. Raynard, S.J., Read, L.R. & Baker, M.D. Evidence for the murine IgH mu locus acting as a hot spot for intrachromosomal homologous recombination. J. Immunol. 168, 2332–2339 (2002).
6. Buzina, A. & Shuman, M.J. An element in the endogenous IgH locus stimulates gene targeting in hybridoma cells. Nucleic Acids Res. 24, 1525–1530 (1996).
7. Thayagarajan, B., Johnson, B.L. & Campbell, C. The effect of target site transcription on gene targeting in human cells in vivo. Nucleic Acids Res. 23, 2784–2790 (1995).
8. Dominguez-Bendala, J.C. & McWhir, J. Enhanced gene targeting frequency in ES cells with low genomic methylation levels. Transgenic Res. 13, 69–74 (2004).
9. Cornea, A.M. & Russell, D.W. Chromosomal position effects on AAV-mediated gene targeting. Nucleic Acids Res. 35, 3852–3859 (2010).
10. Hirata, R., Chamberlain, J., Dong, R. & Russell, D.W. Targeted transgene insertion into human chromosomes by adeno-associated virus vectors. Nat. Biotechnol. 20, 735–738 (2002).
11. Zwaika, T.P. & Thomson, J.A. Homologous recombination in human embryonic stem cells. Nat. Biotechnol. 21, 319–321 (2003).
12. Brook, A. & Leung, W.Y. Direct recruitment of JmJ3 to replication fork regions of mouse genome. EMBO J. 24, 2301–2309 (2005).
13. Miller, D.G., Petek, L.M. & Russell, D.W. Human gene targeting by adeno-associated virus vectors is enhanced by DNA double-strand breaks. Mol. Cell. Biol. 23, 3550–3557 (2003).
14. Porteus, M.H., Cathomen, T., Weitzman, M.D. & Baltimore, D. Efficient targeting vectors. Nat. Genet. 29, 4612–4620 (2000).
15. Vasileva, A., Linden, R.M. & Jessberger, R. Homologous recombination is required for AAV-mediated gene targeting. Nucleic Acids Res. 34, 3345–3360 (2006).
16. Russell, D.W. & Hirata, R.K. Human gene targeting favors insertions over deletions. Hum. Gene Ther. 19, 907–914 (2008).
17. Wu, X., Li, Y., Crise, B. & Burgess, S.M. Transcription start site regions in the human genome are favored targets for MLV integration. Science 300, 1749–1751 (2003).
18. Mitchell, R.S. et al. Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS Pathog. 2, e60 (2006).
19. Costes, A. & Lambert, A.C. Homologous recombination as a replication fork escort: acting together to maintain genome stability. Chromosoma 122, 401–413 (2013).
20. Hansen, R.S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc. Natl. Acad. Sci. USA 107, 139–144 (2010).
21. Goldman, M.A., Holmquist, G.P., Gray, M.C., Caston, L.A. & Nag, A. Replication timing of genes and multiple repetitive elements. Science 224, 686–692 (1984).
22. Heimrich, A., Ballarino, M. & Yora, L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. Mol. Cell 44, 966–977 (2011).
23. Bullock, P., Miller, J. & Botchan, M. Effects of poly(d(GpT)·d(GpC)) and poly(d(CpG)·d(CpG)) repeats on homologous recombination in somatic cells. Nucleic Acids Res. 36, 3948–3955 (2008).
24. Benet, A., Molla, G. & Azorin, F. dIGA x TC), microsatellite DNA sequences enhance homologous DNA recombination in SV40 minichromosomes. Nucleic Acids Res. 28, 4617–4622 (2000).
25. Wu, X., Li, Y., Crise, B. & Burgess, S.M. Transcription start site regions in the human genome are favored targets for MLV integration. Science 300, 1749–1751 (2003).
26. Mitchell, R.S. et al. Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS Pathog. 2, e60 (2006).
27. Wahl, W.P., Wallace, L.J. & Moore, P.D. The Z-DNA motif d(1g30) promotes replication initiation at gene conversion events while stimulating homologous recombination in human cells in culture. Mol. Cell. Biol. 10, 785–793 (1990).
28. Wahl, W.P., Wallace, L.J. & Moore, P.D. Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells. Cell 60, 95–103 (1990).
29. Lin, H., Gundersen, J.H. Transcription-induced DNA toxicity at trinucleotide repeats: double bubble is trouble. Cell Cycle 10, 611–618 (2011).
30. Lin, Y., Leng, M., Wan, M. & Wilson, J.H. Convergent transcription through a long CAG tract destabilizes repeats and induces apoptosis. Mol. Cell. Biol. 30, 8435–8441 (2010).
ONLINE METHODS

Cell culture. HT-1080 cells, HEK293 cells, and HEK293T cells were grown at 37 °C in 5% CO2 in Dulbecco’s modified Eagle’s medium containing glucose (4 g/l; Invitrogen), 10% heat-inactivated FBS, penicillin (100 U/ml), streptomycin (100 µg/ml), and amphotericin (1.25 µg/ml). To generate cells containing proviral target sites, HT-1080 cells were seeded on day 1 at 3 × 10³ cells/dish in two 6-cm-diameter dishes, and on day 2 they were infected with MLV vector LHSN63∆5O at a multiplicity of infection of 10 transducing units/cell in the presence of polybrene (4 µg/ml; Sigma). On day 3, the infected cells were treated with trypsin, pooled, and expanded into 20 10-cm-diameter dishes. Selection with hygromycin B (0.2 mg/ml; Calbiochem) was begun on day 4, and on day 6, all infected cells were pooled and frozen for gene-targeting experiments. Southern blots showed that each cell contained an average of 6.8 provirus copies (not shown). To generate control cells containing untargeted retroviral proviruses, HT-1080 cells were seeded on day 1 at 7.9 × 10⁵ cells/dish in one 10-cm-diameter dish and on day 2 were infected with MLV vector LHSNO at an MOI of 0.1 transducing units/cell per dish in the presence of polybrene (4 µg/ml). On day 3, the infected cells were expanded into ten 10-cm-diameter dishes. Selection with 0.7 mg G418 active compound (Invitrogen) per ml was begun on day 4, and cells were cultured with medium changes every 3–4 d until all cells in control dishes had detached. On day 11, DNA was isolated. Hygromycin selection was omitted from the untargeted control population to reflect the fact that the targeted population contained multiple proviruses per cell, so any single targeted provirus would not have had to confer hygromycin resistance. Antibiotic-selection experiments confirmed that G418-selected cells transduced with LHSNO were also hygromycin resistant.

Vector stocks. The MLV vector plasmid pLHSN63∆5O is based on plasmid pLHSNO65 and contains the following: an MLV retroviral vector backbone, pLHSNO65. Each contains pLHSNO sequences (5′-GCCAGAAGCTGCTTACAC-3′). The AAV vector plasmid pA2HSN5 contains pLHSNO sequences including a 309-bp fragment 5′ to the lph gene, the hph gene, the SV40/Tn5 hybrid promoter, and the 5′ portion of the neo gene (truncated at bp 629). AAV vector AAV2-HSN5 (serotype 2) was made by calcium phosphate transfection of HEK293 cells with each MLV vector plasmid and helper plasmids63 and titered on HT-1080 cells as described previously.9 The AAV vector plasmid pA2HSN5 contains plHSNO sequences including a 309-bp fragment 5′ to the lph gene, the hph gene, the SV40/Tn5 hybrid promoter, and the 5′ portion of the neo gene (truncated at bp 629). AAV vector AAV2-HSN5 (serotype 2) was made by calcium phosphate transfection of HEK293 cells and density-gradient purification as previously described.9,64 The AAV vector titer was based on the amount of full-length single-stranded vector genomes detected on Southern blots. The MVV vector system and proviral target vector were described previously.96

Gene targeting. On day 1, polyclonal HT-1080/LHSNO63∆5O cells were thawed in two 10-cm-diameter dishes, in medium containing 0.2 mg hygromycin B per ml. On day 3, the cells were triedpsinized, pooled, and seeded at 1.4 × 10⁶ cells/dish in six 10-cm dishes. On day 4, the cells in five dishes were infected with AAV2-HSN5 at an MOI of 10,000 genome-containing particles/cell. On day 5, the cells in all dishes were expanded to four 10-cm dishes. Each dilutions were plated for each original dish for plating-efficiency and targeting-frequency calculations. On day 6, G418 (0.7 mg active compound per ml) was added to the medium of all dishes. Cells were cultured with medium changes every 3–4 d until all cells in control dishes had detached, and on day 15, the dilution dishes were stained with Coomassie brilliant blue G, and DNA was isolated from the AAV-targeted cells. The targeting frequency was expressed as the number of G418-resistant colonies per total number of colonies obtained. The spontaneous neo reversion frequency for the polyclonal HT-1080/LHSN63∆5O population was <10⁻⁷.

Shuttle-vector rescue in bacteria. DNA was isolated from the G418-resistant, polyclonal HT-1080/LHSNO cells present in ten 10-cm dishes and the G418-resistant, polyclonal AAV-targeted HT-1080/LHSNO63∆5O cells present in 20 10-cm dishes. The shuttle-vector target sites, along with flanking chromosomal DNA, were rescued as bacterial plasmids as described previously,65 except that three pairs of compatible, cohesive restriction enzymes were used: EcoRI/MfeI, BsrGI/BsiWI, and PciI/BspHI. Transformed bacteria were selected on agar containing kanamycin (50 µg/ml). 9,000 colonies from bacterial transformations of HT-1080/LHSNO DNA were inoculated into individual wells of a 48-well plate (BD Biosciences), each containing 500 µl LB medium with kanamycin (50 µg/ml), to avoid overgrowth of individual clones. 9,000 colonies from transformations of AAV-targeted HT-1080/LHSN63∆5O DNA were inoculated in an identical way. Cultures were grown overnight at 37 °C, at which point the colonies were pooled, and plasmid DNA was purified for sequencing.

MLV provirus integration-site mapping. Inverse PCR was performed to map the integration site of MLV-LAP375∆45P as previously described except for the choice of restriction enzymes and PCR primers. The genomic DNA was digested with Sau3A1 (New England BioLabs) and circularized with T4 DNA Ligase (New England BioLabs). Ligation products were further digested with SpeI (New England BioLabs). Nested PCRs were performed with GoTag Flexi DNA Polymerase (Promega). The first round of PCR was performed with the primers 5′-CCTGAAATGACCCCTGGCCCTTA-3′ and 5′-GGCGAGGAAAGCTGCTTACAC-3′. The PCR products were further amplified in the second-round PCRs by the primers 5′-ATGGCGCTCTGCTTCTGGTTC-3′ and 5′-TGGCCCATATTCAGCTGTGTTCCA-3′. The second-round PCR products were sequenced, and the integration sites were mapped with BLAT on the UCSC genome browser.

High-throughput plasmid sequencing. 10 µg of plasmid DNA from each of the two pooled samples (HT-1080/LHSNO or AAV-targeted HT-1080/LHSN63∆5O) was used for sequencing and prepared according to the manufacturer’s instructions (Illumina). Standard Illumina genomic DNA was constructed from the plasmid DNA, and sequencing was performed on the Genome Analyzer GAIIx with SBS chemistry (Illumina) at the Genome Institute of Singapore to generate 76-bp single-read sequences.

Replication time and replication fork mapping. Replication fork direction was defined from replication-time patterns such that ‘left’ forks were those traveling from a valley of late replication to an adjacent peak of early replication, and ‘right’ forks were those traveling from a peak of early replication to a valley of late replication (as oriented by the plus-strand reference genome). HT-1080 replication time was determined by Repli-Seq as previously described.22 To facilitate replication fork and bioinformatic analysis, we used the normalized 50-kb Repli-Seq densities for each cell-cycle fraction to determine a single scalar replication value at each genomic coordinate according to the following formula that is based on weighting the cell-cycle signals by the average progress of the cell cycle as determined by DNA content (DAPI cytometry fluorescence): $(0.917 \times G1b) + (0.750 \times S1) + (0.583 \times S2) + (0.417 \times S3) + (0.250 \times S4) + (0.000 \times G2)$. These weighted-average data were smoothed by wavelet transformation (7 level, corresponding to a scale of 128 kb for the Repli-Seq 1-kb genomic intervals)77. Wavelet-smoothed replication-time profiles were globally normalized by percentile for further analysis. Replication peaks were defined as local maxima and valleys as local minima in the wavelet-smoothed profiles. Replication forks were defined by a difference in replication time between adjacent peak and valley of >10 (potential range is 0–100; left forks have negative values, and right forks have positive values). Increasing the threshold to >20 or higher allowed for more-confident fork-direction calls. A filter for gaps and other low-signal regions was applied to avoid inclusion of false-replication-fork regions. Overlaps and proximities between replication features and integration sites were determined with the BEDOPS suite.68 The data have been deposited as GEO series GSE58907 with accession numbers GSM1422157, GSM1422158, GSM1422159, GSM1422160, GSM1422161 and GSM1422162.

Mapping of integration sites and comparisons with genomic features. Illumina sequence reads were processed with Biopython99 and aligned with Bowtie20. Reads containing the final 10 bp of the MLV LTR were parsed and aligned to the MLV genome. Reads with 100% identity to the MLV genome were removed from the data set, and the remaining reads were trimmed to remove MLV sequences. Trimmed reads were mapped to the February 2009 assembly of the human genome (GRCh37.78). The locations of genomic features were determined by tables available from the University of California Santa Cruz (UCSC) database. The analysis was performed with the tools available on the Galaxy website (https://usegalaxy.org)172 and processed with Microsoft Excel. To produce a randomly localized set of genomic positions, we generated random numbers between 1 and 5,976,710,698 (the size of the build 37 diploid male genome) with the Excel ‘RANDBETWEEN’ function. We converted the random numbers to chromosomal positions by dividing the numeric range into separate chromosomes, with
each starting at bp 1 of the P arm. The positions were used to extract 50 bp of sequence from the Human February 2009 assembly (GRCh37.78), and the resulting sequences were aligned to the human genome with Bowtie. We extracted 2,217 random sequences, 19% of which corresponded to gapped or repetitive sequences that could not be uniquely mapped. We used the remaining set of 1,798 uniquely localized positions for comparison to the targeted and control data sets. The closest nonoverlapping upstream or downstream element was identified with the 'Fetch closest non-overlapping feature' tool on the Galaxy website.

**Microarray expression analysis.** Global gene expression analysis was performed on Illumina HumanHT-12 v3 microarrays. Total RNA was isolated in triplicate from $4 \times 10^6$ parental HT-1080 cells with the RNeasy Mini kit (Qiagen). The RNA was processed with the Expression BeadChip kit and run per the manufacturer's instructions on the HumanHT-12v3 chip (Illumina). The expression levels of target genes were determined with lumi 1.14.0 (ref. 73) and ranked on the basis of the average expression level from three biological replicates of three independent cell cultures. The data have been submitted as GEO accession number GSE58968.

**Statistical analysis.** Statistics were performed with the R statistical analysis software version 2.15.2012-03-29 (http://www.r-project.org/). Mean gene expression data were analyzed by two-tailed Student's t-test, and comparisons between targeted sites, untargeted sites, and randomly generated sites were analyzed by chi-square test. P values <0.05 were considered statistically significant.

58. Rasheed, S., Nelson Rees, W.A., Toth, E.M., Arnstein, P. & Gardner, M.B. Characterization of a newly derived human sarcoma cell line (HT-1080). Cancer 33, 1027–1033 (1974).

59. Graham, F.L., Smiley, J., Russell, W.C. & Nairn, R. Characteristics of a human cell line transformed by DNA from human adenovirus type 5. J. Gen. Virol. 36, 59–74 (1977).

60. DuBridge, R.B. et al. Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. Mol. Cell. Biol. 7, 379–387 (1987).

61. Inoue, N., Hirata, R.K. & Russell, D.W. High-fidelity correction of mutations at multiple chromosomal positions by adeno-associated virus vectors. J. Virol. 73, 7376–7380 (1999).

62. Chang, A.C. & Cohen, S.N. Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. J. Bacteriol. 134, 1141–1156 (1978).

63. Burns, J.C., Friedmann, T., Driever, W., Burrascano, M. & Yee, J.K. Vesicular stomatitis virus G glycoprotein pseudotyped retroviral vectors: concentration to very high titer and efficient gene transfer into mammalian and nonmammalian cells. Proc. Natl. Acad. Sci. USA 90, 8033–8037 (1993).

64. Khan, I.F., Hirata, R.K. & Russell, D.W. AAV-mediated gene targeting methods for human cells. Nat. Protoc. 6, 482–501 (2011).

65. Rutledge, E.A. & Russell, D.W. Adeno-associated virus vector integration junctions. J. Virol. 71, 8429–8436 (1997).

66. Josephson, N.C. et al. Transduction of human NOD/SCID-repopulating cells with both lymphoid and myeloid potential by foamy virus vectors. Proc. Natl. Acad. Sci. USA 99, 8295–8300 (2002).

67. Thurman, R.E., Day, N., Noble, W.S. & Stamatoyannopoulos, J.A. Identification of higher-order functional domains in the human ENCODE regions. Genome Res. 17, 917–927 (2007).

68. Neph, S. et al. BEDOPS: high-performance genomic feature operations. Bioinformatics 28, 1919–1920 (2012).

69. Cock, P.J.A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423 (2009).

70. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).

71. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, D493–D496 (2004).

72. Taylor, J., Schenck, I., Blankenberg, D. & Nekrutenko, A. Using Galaxy to perform large-scale interactive data analyses. Curr. Protoc. Bioinformatics 19, 10.5 (2002).

73. Du, P., Kibbe, W.A. & Lin, S.M. lumi: a pipeline for processing Illumina microarray. Bioinformatics 24, 1547–1548 (2008).