HHS Public Access

Author manuscript

Nat Struct Mol Biol. Author manuscript; available in PMC 2015 May 01.

Published in final edited form as:

Nat Struct Mol Biol. 2014 November ; 21(11): 969–975. doi:10.1038/nsmb.2895.

A Genome-Wide Map of AAV-Mediated Human Gene Targeting

David R. Deyle[1,7,8], **R. Scott Hansen**[1,8], **Anda M. Cornea**[2], **Li B. Li**[1], **Amber A. Burt**[1], **Ian E. Alexander**[3], **Richard S. Sandstrom**[4], **John A. Stamatoyannopoulos**[4], **Chia-Lin Wei**[5], and **David W. Russell**[1,6]

[1]Department of Medicine, University of Washington, Seattle, Washington, USA

[2]Department of Molecular and Cellular Biology, University of Washington, Seattle, Washington, USA

[3]Gene Therapy Research Unit, Children's Medical Research Institute, Westmead, New South Wales, Australia

[4]Department of Genome Sciences, University of Washington, Seattle, Washington, USA

[5]Genomic Technologies Department, Joint Genome Institute, Walnut Creek, California, USA

[6]Department of Biochemistry, University of Washington, Seattle, Washington, USA

Abstract

To determine which genomic features promote homologous recombination, we created a genome-wide map of gene targeting sites. An adeno-associated virus vector was used to target identical loci introduced as transcriptionally active retroviral vector proviruses. 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 found within transcription units, especially when the target loci were transcribed in the opposite orientation to their surrounding chromosomal genes. The impact of DNA replication was determined by mapping replication forks, which revealed a preference for recombination at target loci transcribed towards an incoming fork. Our results constitute the first genome-wide screen of gene targeting in mammalian cells, and they demonstrate a strong recombinogenic effect of colliding polymerases.

Correspondence should be addressed to D.W.R. (drussell@u.washington.edu).

[7]Present address: Department of Medical Genetics, Mayo Clinic, Rochester, Minnesota, USA.

[8]These authors contributed equally to this work.

Accession Codes

Replication timing and gene expression analyses of HT-1080 human fibrosarcoma cells were deposited in GEO Datasets under accession number GSE58907 (RepliSeq) and GSE58968 (microarray).

Author Contributions

D.R.D., R.S.H., A.M.C., C.L.W., I.E.A. and D.W.R designed experiments. A.M.C. performed gene targeting and plasmid rescue experiments. C.L.W. performed high-throughput sequencing. R.S.H and R.S.S. conducted the Repli-Seq experiments and processed this data. A.E.B provided bioinformatics support. D.R.D. performed target site mapping, bioinformatics processing, microarray analysis, and data collection. L.B.L. mapped the MVM 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.

INTRODUCTION

Homologous recombination is a fundamental biological process required for meiosis, DNA repair, and gene targeting. During meiosis, recombination rates can vary significantly at different chromosomal loci [1], 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 frequencies [2,3], while in mammalian cells some target loci alleles can be targeted at higher frequencies than others [4], and recombination hotspots have been identified such as that present in the murine IgH locus [5,6]. In addition, both transcription and DNA hypomethylation can increase targeting frequencies in mammalian cells [7,8], 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 our ability to manipulate mammalian genomes.

We previously used adeno-associated virus (AAV) vectors to study position effects on human gene targeting [9]. 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 conditions [10]. While these targeting frequencies can be orders of magnitude higher than those typically obtained in human cells by transfection or electroporation [11,12], both processes share common features, including stimulation by double strand breaks [13,14], involvement of the same homologous recombination proteins [15], and similar effects of mutation type on targeting frequencies [16]. When identical target sites were introduced at 16 different chromosomal positions in HT-1080 human fibrosarcoma cells, their AAV-mediated gene targeting frequencies varied as much as one log [9]. 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 AAV-mediated gene targeting map. Identical target sites were introduced at thousands of chromosomal positions with retroviral vectors, an AAV gene targeting vector was used to correct a neomycin resistance gene mutation in these targets, and the chromosomal locations of each target site were mapped by high-throughput DNA sequencing. A comparison of these targeted site positions 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 (Figure 1A). The murine leukemia virus (MLV) vector LHSN63Δ53O contains a nonfunctional neomycin phosphotransferase target gene (*neo*) with a 53 bp deletion at nucleotide 63 of its open reading 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-LHSN63Δ53O with a truncated *neo* gene that lacks the 53 bp deletion. Diploid HT-1080 human fibrosarcoma cells were transduced with MLV-LHSN63Δ53O, and the resulting hygromycin-resistant cells were selected as a polyclonal population. These cells were then transduced with AAV2-HSN5′ and cultured in G418 to select for targeting events. 7,950 G418-resistant, targeted clones were obtained and expanded as a polyclonal population for target site analysis. Assuming 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-LHSN63Δ53O except it contains a functional *neo* gene. HT-1080 cells transduced with MLV-LHSNO were cultured in G418 and this polyclonal population of approximately 75,000 independent clones was expanded for analysis.

Genomic DNA was isolated from both the targeted and untargeted populations of G418-resistant cells, digested with specific restriction enzymes to release the bacterial plasmid and flanking genomic sequences of each provirus, circularized, and rescued in *Escherichia coli* as kanamycin-resistant plasmids (Figure 1A). Individual, plasmid-containing bacterial colonies were amplified separately to prevent overgrowth of specific clones and then combined before plasmid purification and DNA sequencing. Table 1 summarizes these sequencing results. Nine thousand plasmids rescued from both gene-targeted cells and untargeted control cells were sequenced with an Illumina Genome Analyzer, and more than 12,000,000 reads of 75 nt were obtained in each case. After screening for reads containing a portion of the MLV long terminal repeat and eliminating duplicates, a total of 2,015 targeted and 1,928 control proviruses were uniquely localized to the Feb. 2009 assembly of the human genome from targeted and untargeted cells respectively (Supplementary Table 1). The rescued proviruses were present in all human chromosomes (Figure 1B), and the cytogenetic distributions of all mapped sites are shown in Figure 1C. There were no significant differences in the percentage of sites per chromosome, except 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). These “ambiguous” sites were characterized further based on the type of repetitive DNA they contained, which showed a similar distribution among targeted and untargeted sites (Supplementary Table 2).

Effects of flanking genomic elements on targeting

To determine if the chromosomal landscape can 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), Long Terminal Repeats (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 Figure 1). We focused specifically on the GT dinucleotide repeat subset of microsatellite repeats, since our

previous analysis of 16 target sites suggested that these elements might increase targeting frequencies [9], 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 found near these elements (Supplementary Figure 1), consistent with prior studies of gammaretrovirus integration profiles [17–19].

Convergent transcription increases targeting

Gammaretroviral vectors preferentially integrate near the transcription start sites of active genes [17], and the majority of provirus sites in our experiment were found within RefSeq transcription units (Table 2). There was a slight but statistically significant difference between targeted and untargeted sites (58.6% vs 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 (Figure 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 1042 intragenic untargeted sites (70%) were represented on the array. The median expression level of genes containing targeted sites was significantly higher than that of untargeted sites (8.39 vs 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). These provirus-containing genes were binned into subsets based on their expression level, and we detected a statistically significant increase in the percent of targeted sites found in the most highly expressed genes (Figure 2B).

Since the provirus target sites 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 (Figure 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, since a non-specific 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 [20,21]. To assess the role of replication in gene targeting, we determined the locations of 4173 replication initiation zones and the directions of replication forks

throughout the genome of HT-1080 cells with the Repli-Seq method [22]. Newly replicated DNAs in exponentially growing cells were pulse-labeled with 5-bromo-2-deoxyuridine (BrdU), and the cells were then sorted into six cell cycle subsets by flow cytometry (late G1, four subsets of S phase, and early G2). The BrdU-labeled, newly replicated DNA strands were then isolated from each cell cycle population, sequenced by Illumina-based methods for massively parallel sequencing, and mapped to the human genome. Uniquely mapped sequences were used to calculate local signal densities for each of the cell cycle fractions from which average replication time values at each genomic coordinate were determined (see Methods). Figure 3A shows the 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. 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. A portion of chromosome 1 is expanded in Figure 3B, with fork direction indicated as well the locations and transcriptional orientation of target sites.

In comparison to random chromosomal positions, both targeted and untargeted sites were more likely to be present in early replicating regions (Figure 4A) and closer to initiation zones (Figure 4B). This can be explained by the preferential integration of gammaretroviral vectors in expressed genes [17], which are known to replicate earlier in S phase [23]. A similar correlation could also account for the slightly earlier replication timing of targeted sites in comparison to untargeted sites, since high chromosomal gene expression favors targeting (Figure 2B). Targeted sites were more likely to be transcribed in the opposite direction of replication fork movement, while this had no effect on untargeted sites (Figure 4C). When this analysis was performed on the subset of forks defined by peaks and valleys with replication time differences of at least 10%, 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 (Figure 4D). As noted above for overlapping transcription units (Figure 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 Figure 2), which require more time to complete transcription and therefore must be transcribed during replication [24]. In addition, this analysis showed that long chromosomal genes are preferentially transcribed in the same direction as replication (Supplementary Figure 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. Figure 5 shows examples of these target sites with their corresponding Repli-Seq patterns. Fourteen sites were located in regions where replication fork direction could be unambiguously established (Supplementary Table 3), allowing us to directly determine its impact on targeting frequencies. On average, sites transcribed in the opposite direction of fork movement had ~two times higher targeting frequencies.

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 reported stimulatory effects of neighboring repeat sequences on homologous recombination in mammalian cells [25–28], we did not observe a significant effect of sequence elements on targeting 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 the dynamic activity of the genome that were associated with the directional effects of transcription and replication.

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 of its surrounding chromosomal gene. While this type of “convergent transcription” has not been directly linked to homologous recombination, the topological consequences are similar to those of RNA and DNA polymerase collisions discussed below. In addition, convergent transcription through repetitive sequences may induce genotoxicity via the ATR (ataxia-telangiectasia mutated [ATM] and Rad3-related) signaling pathway [29–31], which could promote recombination during the DNA repair process, and in yeast, cohesin accumulates at sites of convergent transcription [32], which 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 to 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 (Figure 5C), which was consistent with the genome-wide data showing that ~70% of targeted sites were transcribed in the opposite direction of fork movement when fork direction was limited to high confidence calls (Figure 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].

Figure 6A models how convergent replication and transcription might stimulate AAV-mediated gene targeting, 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, chickenfoot structures formed by fork collapse can be processed by an exonuclease such as EXO1 [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 though 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 anti-sense vector strand, while 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 a ~one log difference in targeting frequencies depending on which strand was packaged [50]. In this system, a mutant Human Placental Alkaline Phosphatase (*ALPP*) reporter gene was introduced into HT-1080 cells with a gammaretroviral vector, and then corrected with MVM vectors containing a truncated *ALPP* gene. By inverting the vector termini, we generated vector stocks containing either the sense or anti-sense strands of the target site (MVM-s and MVM-as), as well as a flanking GFP cassette used to control for transduction frequencies. We have now mapped the location of this target site and confirmed that it is transcribed in the opposite orientation to the replication fork (Figure 6B). Since the sense strand vector targeted at higher frequencies (Figure 6C), these data support a model in which the incoming AAV (or MVM) vector genome pairs with the exonuclease-processed chicken foot or the lagging strand of the fork, and not the R-loop.

In this study, the high targeting frequencies of AAV vectors allowed us to generate thousands of targeted clones without introducing recombinogenic double stand breaks. The targeting frequencies of transfected plasmid-based constructs would typically be 2–4 orders of magnitude lower under these conditions [51], raising the possibility that unique features of AAV vector biology influenced our results. One distinction is the single-stranded AAV vector genome, which appears to be the recombination substrate based on the strand preferences observed with other parvoviral targeting vectors [50] and the lack of targeting by double-stranded, encapsidated AAV vector genomes [52]. Double-stranded plasmid molecules may be unable to pair with exposed single-stranded chromosomal regions in the same way. Alternatively, the AAV capsid could promote targeting through specific aspects of its processing, as suggested by microinjection experiments showing that purified AAV vector genomes did not target efficiently even when delivered directly to the cytoplasm or nucleus [53]. Despite these differences, AAV vectors and transfected plasmid constructs also share mechanistic similarities, including higher targeting frequencies when introducing

insertions [16], stimulation by double-strand breaks [13,14], and the participation of the same recombination proteins [15].

In summary, our genome-wide screen of AAV-mediated human gene targeting demonstrated a stimulatory effect of colliding polymerases. Our findings support a model in which the single-stranded vector genome pairs with exposed single stranded chromosomal regions found at stalled replication forks, and when the target site is transcribed in the opposite direction to the incoming replication fork, the sense strand vector is more efficient at targeting. This increases homologous recombination above the baseline levels observed in the absence of transcription, which must also occur since AAV vectors target silent genes at reduced frequencies [54–56]. Replication forks could also play a key role in targeting these silent loci, because targeting requires S phase [53,57], and the process exposes single-stranded chromosomal regions that may pair with vector DNA.

MATERIALS AND METHODS

Cell culture

HT-1080 cells [58], 293 cells [59], and 293T cells [60] were grown at 37°C in 5% CO_2 in Dulbecco's modified Eagle's medium containing 4 g glucose/liter (Invitrogen), 10% heat-inactivated fetal bovine serum, 100 U/ml penicillin, 100 μg/ml streptomycin, and 1.25 μg/ml amphotericin. To generate cells containing proviral target sites, HT-1080 cells were seeded on day 1 at 3×10^5 cells/dish in two 6-cm-diameter dishes, and on day 2 they were infected with MLV vector LHSN63Δ53O at a multiplicity of infection of 10 transducing units/cell in the presence of 4 μg/ml Polybrene (Sigma). On day 3, the infected cells were treated with trypsin, pooled, and expanded into twenty 10-cm-diameter dishes. Selection with 0.2 mg/ml hygromycin B (Calbiochem) was begun on day 4, and on day 6, all infected cells were pooled and frozen down 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 7.9×10^5 cells/dish in one 10-cm-diameter dish, and on day 2 infected with MLV vector LHSNO at an MOI of 0.1 transducing units/cell in the presence of 4 μg/ml Polybrene. On day 3, the infected cells were expanded into ten 10-cm-diameter dishes. Selection with 0.7 mg active compound/ml G418 (Invitrogen) was begun on day 4, and cells were cultured with media changes every 3–4 days 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 pLHSN63Δ53O is based on plasmid pLHSNO [61] and contains the following: an MLV retroviral vector backbone, *hph* gene, SV40/Tn5 hybrid promoter, *neo* gene with a 53 bp deletion at bp 63 of the open reading frame, and p15A plasmid replication origin [62]. VSV-G-pseudotyped MLV vector stocks were made by cotransfection of 293 cells

with each MLV vector plasmid and helper plasmids [63], and titered on HT-1080 cells as described [9]. The AAV vector plasmid pA2HSN5′ contains pLHSNO sequences including a 309 bp fragment 5′ to the *hph* 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 293T cells and density gradient purification as described [64]. The AAV vector titer was based on the amount of full-length single-stranded vector genomes detected on Southern blots. The MVM vector system and proviral target vector were described previously [50].

Gene targeting

On day 1, polyclonal HT-1080/LHSN63Δ53O cells were thawed in two 10-cm-diameter dishes, in medium containing 0.2 mg/ml hygromycin B. On day 3, the cells were trypsinized, pooled, and seeded at 1.4×10^6 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/ml active compound) was added to the medium of all dishes. Cells were cultured with media changes every 3–4 days 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/total number of colonies obtained. The spontaneous *neo* reversion frequency for the polyclonal HT-1080/LHSN63Δ53O 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/LHSN63Δ53O cells present in twenty 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 50 μg kanamycin/ml. Nine thousand 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 50 μg kanamycin/ml, to avoid overgrowth of individual clones. Nine thousand colonies from transformations of AAV-targeted HT-1080/LHSN63Δ53O 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-LAP375Δ4SP as previously described [66] 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 by GoTaq Flexi DNA Polymerase (Promega). The first round of PCR was performed with the primers 5′-CCTGAAATGACCCTGTGCCTTA and 5′-GGGCAGGAACTGCTTACCAC. The PCR products were further amplified in the second round PCR by the primers 5′-AGTTCGCTTCTCGCTTCTGTTC and 5′-TGGCCCATATTCAGCTGTTCCA. The second round PCR products were sequenced and the integration sites were mapped with BLAT on the UCSC genome browser.

High-throughput plasmid sequencing

Ten μg of plasmid DNA from each of the two pooled samples (HT-1080/LHSNO or AAV-targeted HT-1080/LHSN63Δ53O) were 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*G1b)+(0.750*S1)+(0.583*S2)+(0.417*S3)+(0.250*S4)+(0*G2). These weighted average data were smoothed by wavelet transformation [J7 level, corresponding to a scale of 128 kb for the Repli-Seq 1 kb genomic intervals [67]. 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 using the BEDOPS suite of programs [68]. The data are submitted as GEO series GSE58907 with accession numbers GSM1422157-GSM1422162.

Mapping of integration sites and comparisons with genomic features

Illumina sequence reads were processed with Biopython [69] and aligned with Bowtie [70]. Reads containing the final ten base pairs 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 Feb. 2009 assembly of the human genome (GRCh37.78). The locations of

genomic features were determined by using tables available from the University of California Santa Cruz (UCSC) database [71]. The analysis was performed with the tools available on the Galaxy website (http://main.g2.bx.psu.edu/) [72] and processed using 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 base pair 1 of the p arm. The positions were used to extract 50 bp of sequence from the Human Feb. 2009 assembly (GRCh37.78), and the resulting sequences were aligned to the human genome using 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 non-overlapping upstream or downstream element was identified using 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 4×10^6 parental HT-1080 cells using the RNeasy Mini kit (Qiagen). The RNA was processed using 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 by using the lumi 1.14.0 software package [73], and ranked based on the average expression level from three biological replicates. The data are submitted as GEO accession number GSE58968.

Statistical analysis

Statistics were performed with the R statistical analysis software version 2.15.2012-03-29 [74]. Mean gene expression data were analyzed using two-tailed Student's t-test and comparisons between targeted sites, untargeted sites, and randomly generated sites were analyzed by using the Chi-square test. *P*-values <0.05 were considered statistically significant.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank J. Delrow, A. Dawson, and R. Basom for microarray analysis, P. Hendrie for MVM data, T. Canfield for Repli-Seq processing and R. Hirata and R. Stolitenko for technical assistance. This work was supported by grants from the U.S. National Institutes of Health (R01DK55759, P01HL53750, and R01AR48328) to D.W.R, (K08AR053917) to D.R.D. (U54HG007010 and P01HL53750) to RSH and JAS. This work was also supported by grants from the Australian Department of Innovation, Industry, Science and Research (CG130052) to D.W.R, I.E.A. and C.L.W., and the Genome Institute of Singapore (GIS) funded by the Agency for Science, Technology and Research (A*STAR) Singapore to C.L.W.

References

1. Kong A, et al. A high-resolution recombination map of the human genome. Nat Genet. 2002; 31:241–7. [PubMed: 12053178]
2. Wilson JH, Leung WY, Bosco G, Dieu D, Haber JE. The frequency of gene targeting in yeast depends on the number of target copies. Proc Natl Acad Sci U S A. 1994; 91:177–81. [PubMed: 8278360]
3. Gray M, Honigberg SM. Effect of chromosomal locus, GC content and length of homology on PCR-mediated targeted gene replacement in Saccharomyces. Nucleic Acids Res. 2001; 29:5156–62. [PubMed: 11812849]
4. Yanez RJ, Porter AC. A chromosomal position effect on gene targeting in human cells. Nucleic Acids Res. 2002; 30:4892–901. [PubMed: 12433992]
5. Raynard SJ, Read LR, Baker MD. Evidence for the murine IgH mu locus acting as a hot spot for intrachromosomal homologous recombination. J Immunol. 2002; 168:2332–9. [PubMed: 11859123]
6. Buzina A, Shulman MJ. An element in the endogenous IgH locus stimulates gene targeting in hybridoma cells. Nucleic Acids Res. 1996; 24:1525–30. [PubMed: 8628687]
7. Thyagarajan B, Johnson BL, Campbell C. The effect of target site transcription on gene targeting in human cells in vitro. Nucleic Acids Res. 1995; 23:2784–90. [PubMed: 7651841]
8. Dominguez-Bendala J, McWhir J. Enhanced gene targeting frequency in ES cells with low genomic methylation levels. Transgenic Res. 2004; 13:69–74. [PubMed: 15070077]
9. Cornea AM, Russell DW. Chromosomal position effects on AAV-mediated gene targeting. Nucleic Acids Res. 2010; 38:3582–94. [PubMed: 20185563]
10. Hirata R, Chamberlain J, Dong R, Russell DW. Targeted transgene insertion into human chromosomes by adeno-associated virus vectors. Nat Biotechnol. 2002; 20:735–8. [PubMed: 12089561]
11. Zwaka TP, Thomson JA. Homologous recombination in human embryonic stem cells. Nat Biotechnol. 2003; 21:319–21. [PubMed: 12577066]
12. Brown JP, Wei W, Sedivy JM. Bypass of senescence after disruption of p21CIP1/WAF1 gene in normal diploid human fibroblasts. Science. 1997; 277:831–4. [PubMed: 9242615]
13. Miller DG, Petek LM, Russell DW. Human gene targeting by adeno-associated virus vectors is enhanced by DNA double-strand breaks. Mol Cell Biol. 2003; 23:3550–7. [PubMed: 12724413]
14. Porteus MH, Cathomen T, Weitzman MD, Baltimore D. Efficient gene targeting mediated by adeno-associated virus and DNA double-strand breaks. Mol Cell Biol. 2003; 23:3558–65. [PubMed: 12724414]
15. Vasileva A, Linden RM, Jessberger R. Homologous recombination is required for AAV-mediated gene targeting. Nucleic Acids Res. 2006; 34:3345–60. [PubMed: 16822856]
16. Russell D, Hirata R. Human Gene Targeting Favors Insertions Over Deletions. Hum Gene Ther. 2008; 19:907–914. [PubMed: 18680404]
17. Wu X, Li Y, Crise B, Burgess SM. Transcription start regions in the human genome are favored targets for MLV integration. Science. 2003; 300:1749–51. [PubMed: 12805549]
18. Mitchell RS, et al. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. 2004; 2:E234. [PubMed: 15314653]
19. Lewinski MK, et al. Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS Pathog. 2006; 2:e60. [PubMed: 16789841]
20. Costes A, Lambert AE. Homologous recombination as a replication fork escort: fork protection and recovery. Biomolecules. 2013; 3:39–71. [PubMed: 24970156]
21. Aze A, Zhou JC, Costa A, Costanzo V. DNA replication and homologous recombination factors: acting together to maintain genome stability. Chromosoma. 2013; 122:401–13. [PubMed: 23584157]
22. Hansen RS, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A. 2010; 107:139–44. [PubMed: 19966280]
23. Goldman MA, Holmquist GP, Gray MC, Caston LA, Nag A. Replication timing of genes and middle repetitive sequences. Science. 1984; 224:686–92. [PubMed: 6719109]

24. Helmrich A, Ballarino M, Tora L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. Mol Cell. 2011; 44:966–77. [PubMed: 22195969]
25. Bullock P, Miller J, Botchan M. Effects of poly[d(pGpT).d(pApC)] and poly[d(pCpG).d(pCpG)] repeats on homologous recombination in somatic cells. Mol Cell Biol. 1986; 6:3948–53. [PubMed: 3025620]
26. Benet A, Molla G, Azorin F. d(GA x TC)(n) microsatellite DNA sequences enhance homologous DNA recombination in SV40 minichromosomes. Nucleic Acids Res. 2000; 28:4617–22. [PubMed: 11095670]
27. Wahls WP, Wallace LJ, Moore PD. The Z-DNA motif d(TG)30 promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture. Mol Cell Biol. 1990; 10:785–93. [PubMed: 2405255]
28. Wahls WP, Wallace LJ, Moore PD. Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells. Cell. 1990; 60:95–103. [PubMed: 2295091]
29. Lin Y, Wilson JH. Transcription-induced DNA toxicity at trinucleotide repeats: double bubble is trouble. Cell Cycle. 2011; 10:611–8. [PubMed: 21293182]
30. Lin Y, Leng M, Wan M, Wilson JH. Convergent transcription through a long CAG tract destabilizes repeats and induces apoptosis. Mol Cell Biol. 2010; 30:4435–51. [PubMed: 20647539]
31. Nakamori M, Pearson CE, Thornton CA. Bidirectional transcription stimulates expansion and contraction of expanded (CTG)*(CAG) repeats. Hum Mol Genet. 2011; 20:580–8. [PubMed: 21088112]
32. Lengronne A, et al. Cohesin relocation from sites of chromosomal loading to places of convergent transcription. Nature. 2004; 430:573–8. [PubMed: 15229615]
33. Sjogren C, Nasmyth K. Sister chromatid cohesion is required for postreplicative double-strand break repair in Saccharomyces cerevisiae. Curr Biol. 2001; 11:991–5. [PubMed: 11448778]
34. Yelin R, et al. Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol. 2003; 21:379–86. [PubMed: 12640466]
35. Tuduri S, Tourriere H, Pasero P. Defining replication origin efficiency using DNA fiber assays. Chromosome Res. 2010; 18:91–102. [PubMed: 20039120]
36. Prado F, Aguilera A. Impairment of replication fork progression mediates RNA polII transcription-associated recombination. Embo J. 2005; 24:1267–76. [PubMed: 15775982]
37. Takeuchi Y, Horiuchi T, Kobayashi T. Transcription-dependent recombination and the role of fork collision in yeast rDNA. Genes Dev. 2003; 17:1497–506. [PubMed: 12783853]
38. de la Loza MC, Wellinger RE, Aguilera A. Stimulation of direct-repeat recombination by RNA polymerase III transcription. DNA Repair (Amst). 2009; 8:620–6. [PubMed: 19168400]
39. Helmrich A, Ballarino M, Nudler E, Tora L. Transcription-replication encounters, consequences and genomic instability. Nat Struct Mol Biol. 2013; 20:412–8. [PubMed: 23552296]
40. Kim N, Jinks-Robertson S. Transcription as a source of genome instability. Nat Rev Genet. 2012; 13:204–14. [PubMed: 22330764]
41. Branzei D, Foiani M. Maintaining genome stability at the replication fork. Nat Rev Mol Cell Biol. 2010; 11:208–19. [PubMed: 20177396]
42. Postow L, et al. Positive torsional strain causes the formation of a four-way junction at replication forks. J Biol Chem. 2001; 276:2790–6. [PubMed: 11056156]
43. Sogo JM, Lopes M, Foiani M. Fork reversal and ssDNA accumulation at stalled replication forks owing to checkpoint defects. Science. 2002; 297:599–602. [PubMed: 12142537]
44. Huertas P, Aguilera A. Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. Mol Cell. 2003; 12:711–21. [PubMed: 14527416]
45. Yu K, Chedin F, Hsieh CL, Wilson TE, Lieber MR. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. Nat Immunol. 2003; 4:442–51. [PubMed: 12679812]

46. Cotta-Ramusino C, et al. Exo1 processes stalled replication forks and counteracts fork reversal in checkpoint-defective cells. Mol Cell. 2005; 17:153–9. [PubMed: 15629726]
47. Berns KI, Adler S. Separation of two types of adeno-associated virus particles containing complementary polynucleotide chains. J Virol. 1972; 9:394–6. [PubMed: 5014934]
48. Bourguignon GJ, Tattersall PJ, Ward DC. DNA of minute virus of mice: self-priming, nonpermuted, single-stranded genome with a 5′-terminal hairpin duplex. J Virol. 1976; 20:290–306. [PubMed: 789912]
49. Crawford LV, Follett EA, Burdon MG, McGeoch DJ. The DNA of a minute virus of mice. J Gen Virol. 1969; 4:37–46. [PubMed: 4975639]
50. Hendrie PC, Hirata RK, Russell DW. Chromosomal integration and homologous gene targeting by replication-incompetent vectors based on the autonomous parvovirus minute virus of mice. J Virol. 2003; 77:13136–45. [PubMed: 14645570]
51. Russell DW, Hirata RK. Human gene targeting by viral vectors. Nat Genet. 1998; 18:325–30. [PubMed: 9537413]
52. Hirata RK, Russell DW. Design and packaging of adeno-associated virus gene targeting vectors. J Virol. 2000; 74:4612–20. [PubMed: 10775597]
53. Liu X, et al. Targeted correction of single-base-pair mutations with adeno-associated virus vectors under nonselective conditions. J Virol. 2004; 78:4165–75. [PubMed: 15047832]
54. Wang PR, et al. Induction of hepatocellular carcinoma by in vivo gene targeting. Proc Natl Acad Sci U S A. 2012; 109:11264–11269. [PubMed: 22733778]
55. Rogers CS, et al. Production of CFTR-null and CFTR-DeltaF508 heterozygous pigs by adeno-associated virus-mediated gene targeting and somatic cell nuclear transfer. J Clin Invest. 2008; 118:1571–7. [PubMed: 18324337]
56. Sun X, et al. Adeno-associated virus-targeted disruption of the CFTR gene in cloned ferrets. J Clin Invest. 2008; 118:1578–83. [PubMed: 18324338]
57. Trobridge G, Hirata RK, Russell DW. Gene targeting by adeno-associated virus vectors is cell-cycle dependent. Hum Gene Ther. 2005; (16):522–6. [PubMed: 15871683]
58. Rasheed S, Nelson Rees WA, Toth EM, Arnstein P, Gardner MB. Characterization of a newly derived human sarcoma cell line (HT-1080). Cancer. 1974; 33:1027–33. [PubMed: 4132053]
59. Graham FL, Smiley J, Russell WC, Nairn R. Characteristics of a human cell line transformed by DNA from human adenovirus type 5. J Gen Virol. 1977; 36:59–74. [PubMed: 886304]
60. DuBridge RB, et al. Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. Mol Cell Biol. 1987; 7:379–87. [PubMed: 3031469]
61. Inoue N, Hirata RK, Russell DW. High-fidelity correction of mutations at multiple chromosomal positions by adeno-associated virus vectors. J Virol. 1999; 73:7376–7380. [PubMed: 10438827]
62. Chang AC, Cohen SN. Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. J Bacteriol. 1978; 134:1141–56. [PubMed: 149110]
63. Burns JC, Friedmann T, Driever W, Burrascano M, Yee JK. 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 U S A. 1993; 90:8033–7. [PubMed: 8396259]
64. Khan IF, Hirata RK, Russell DW. AAV-mediated gene targeting methods for human cells. Nat Protoc. 2011; 6:482–501. [PubMed: 21455185]
65. Rutledge EA, Russell DW. Adeno-associated virus vector integration junctions. J Virol. 1997; 71:8429–36. [PubMed: 9343199]
66. Josephson NC, et al. Transduction of human NOD/SCID-repopulating cells with both lymphoid and myeloid potential by foamy virus vectors. Proc Natl Acad Sci U S A. 2002; 99:8295–300. [PubMed: 12060773]
67. Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA. Identification of higher-order functional domains in the human ENCODE regions. Genome Res. 2007; 17:917–27. [PubMed: 17568007]
68. Neph S, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012; 28:1919–20. [PubMed: 22576172]

69. Cock PJA, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25:1422–1423. [PubMed: 19304878]

70. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009; 10:R25. [PubMed: 19261174]

71. Karolchik D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004; 32:D493–6. [PubMed: 14681465]

72. Taylor, J.; Schenck, I.; Blankenberg, D.; Nekrutenko, A. Current Protocols in Bioinformatics. John Wiley & Sons, Inc; 2002. Using Galaxy to Perform Large-Scale Interactive Data Analyses.

73. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008; 24:1547–8. [PubMed: 18467348]

74. Team, R.D.C. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2012.

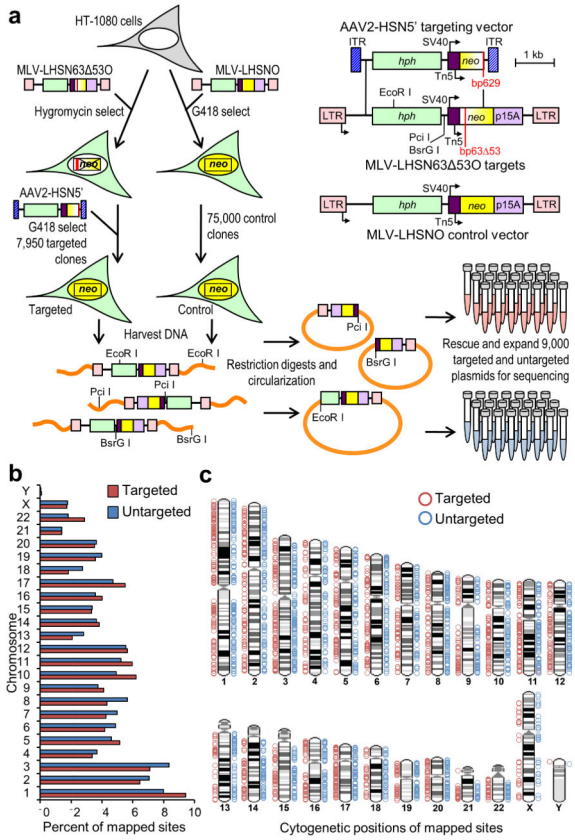

Figure 1. Genome-wide gene targeting

(a) Experimental design with the inset showing the structures of the AAV2-HSN5′ targeting vector with a *neo* gene truncated at bp 629, MLV-LHSN63Δ53O target site provirus containing a 53 bp deletion at bp 63 of *neo*, and control vector MLV-LHSNO with a wild-type *neo* gene. The 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 are graphed per chromosome as a percentage of all mapped sites. There were no significant differences ($P > 0.05$, Chi-square test). **(c)** The locations of mapped sites are shown as red (targeted) or blue (untargeted) circles adjacent to each human chromosome ideogram.

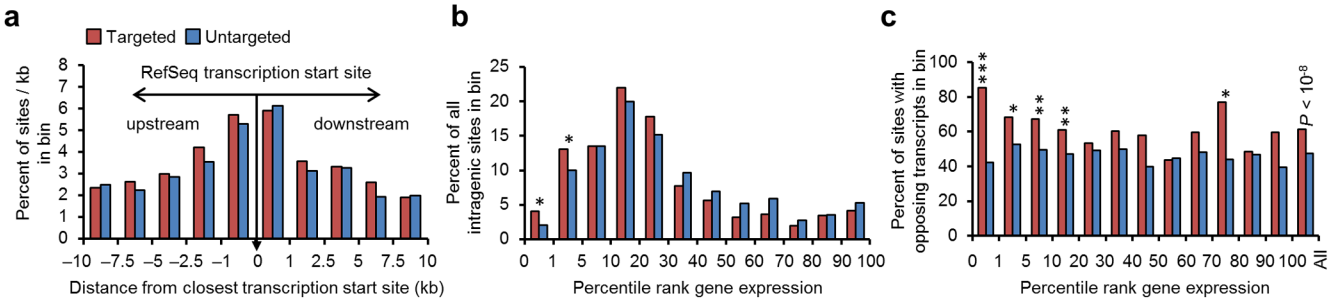

Figure 2. Transcriptional effects on targeting

(a) The percent of targeted and untargeted sites (per kb) found within each interval relative to RefSeq gene transcription start sites. 1180 targeted sites were found in 895 genes, and 1042 untargeted sites were found in 889 genes. **(b)** The percent of intragenic sites found in genes binned into different expression levels by global gene expression ranking of HT-1080 cells, with low percentile rank indicating a higher expression level. **(c)** The percent of intragenic targeted and control provirus sites found in opposite transcriptional orientation to the chromosomal gene they are embedded in is shown after ranking and binning genes by expression level. *P* values were determined by Chi-square test and significant values (**P* < 0.05, **P* < 0.01,**P* < 0.002) are shown by asterisks.

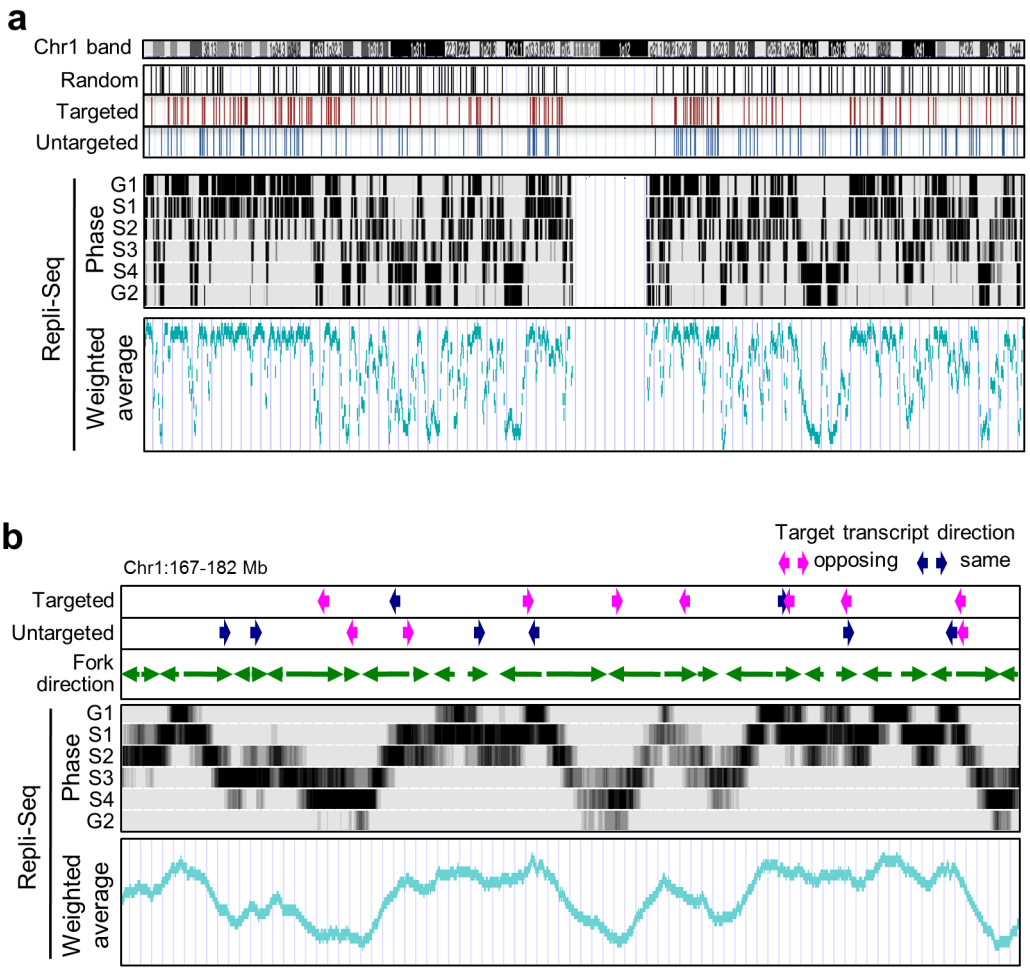

Figure 3. Genome-wide replication fork mapping

(a) Repli-Seq replication timing results are shown for chromosome 1 and aligned to the positions of random, targeted and untargeted sites used in our study. Sequence read tracings are shown for each cell cycle phase (late G1, four subset of S phase, and early G2), as are the weighted averages of these reads. **(b)** A 15 Mb close-up of these results is shown in the same format, except the target site transcription directions of targeted and of untargeted proviruses are shown with replication fork directions underneath, and transcript orientation relative to fork movement indicated (opposing in red, and same in blue).

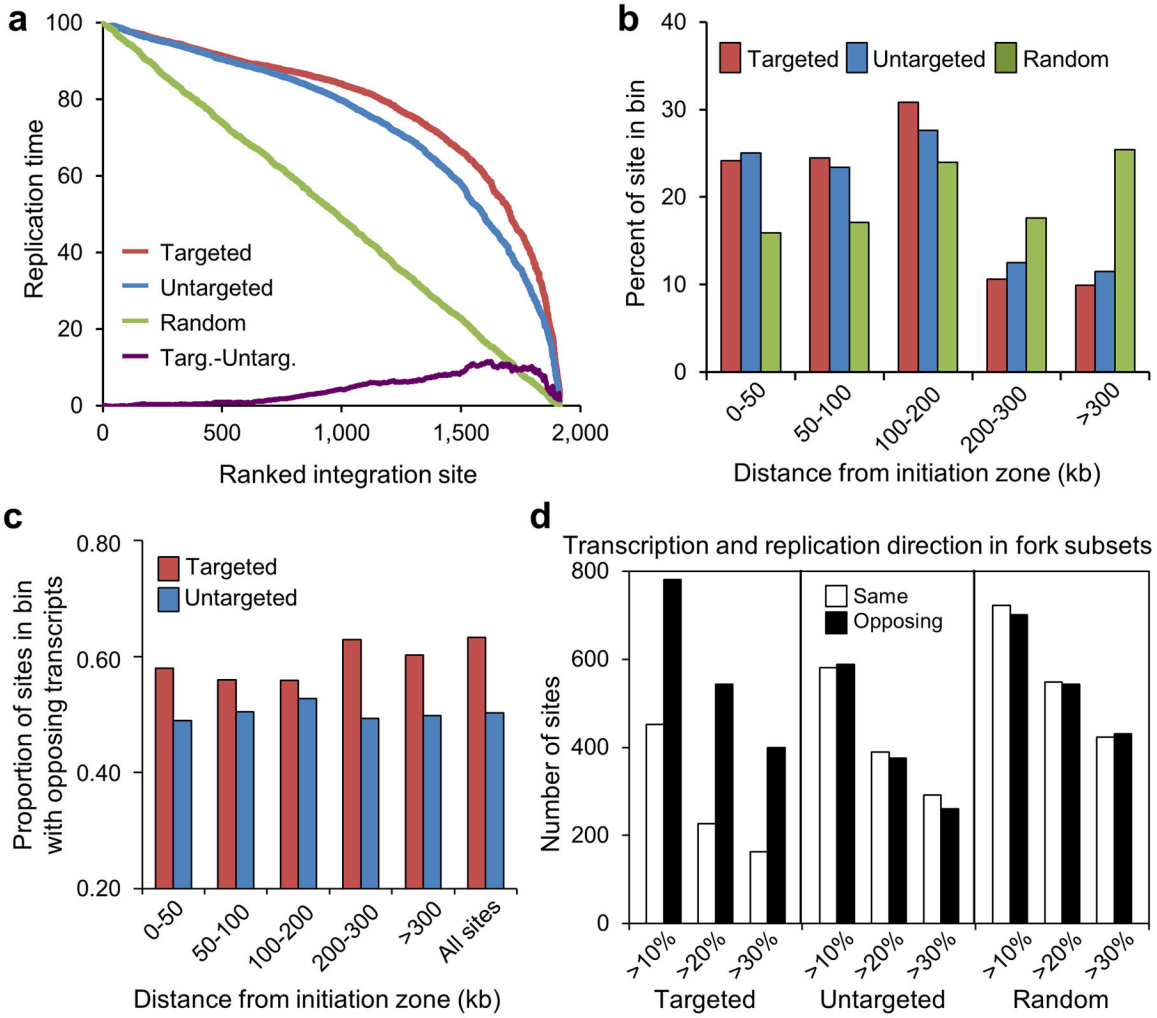

Figure 4. DNA replication effects on targeting

(a) The ranked replication time distribution is shown for targeted, untargeted, and random sites, along with the difference between targeted and untargeted sites showing slightly earlier replication of targeted sites. **(b)** The percent of sites found at different distances from replication initiation zones. There were no statistical differences between targeted and untargeted sites, except for the 100–200 kb window ($P = 0.03$, Chi-square test). **(c)** The proportion of sites transcribed in the opposite direction of fork movement is shown at different distances from initiation zones. *P <0.05, **P <10^{-5}, Chi-square test. **(d)** The number of sites transcribed in the opposite or same direction as fork movement are shown when replication timing differed by at least 10%, 20% or 30%, to increase confidence in fork direction calls. The total number of targeted, untargeted and random sites analyzed in A–C was 2007, 1909 and 2001 respectively.

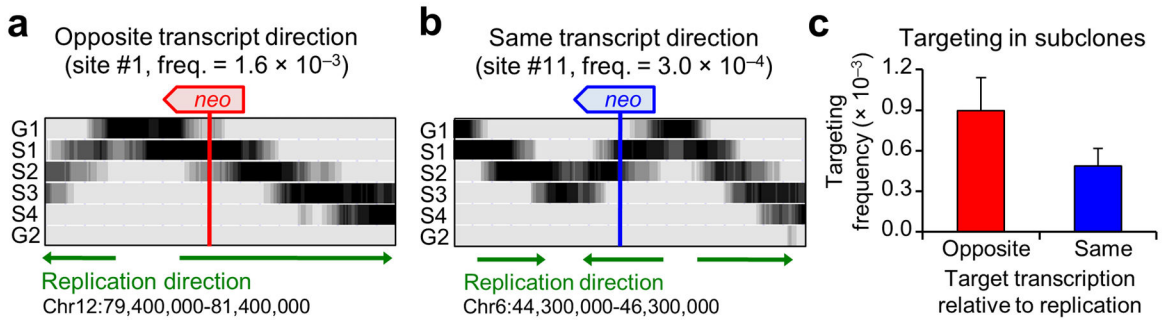

Figure 5. Targeting frequencies in subclones with specific, mapped integration sites

Examples of Repli-Seq data for target sites transcribed in the opposite **(a)** or same **(b)** direction as replication fork movement. Site numbers refer to Supplementary Table 3. **(c)** Average targeting frequencies are shown for all target sites with discernable replication fork directions (Error bars, s.e.m. for n = 7 for each group). The two groups were significantly different ($p<0.05$ by Student's test).

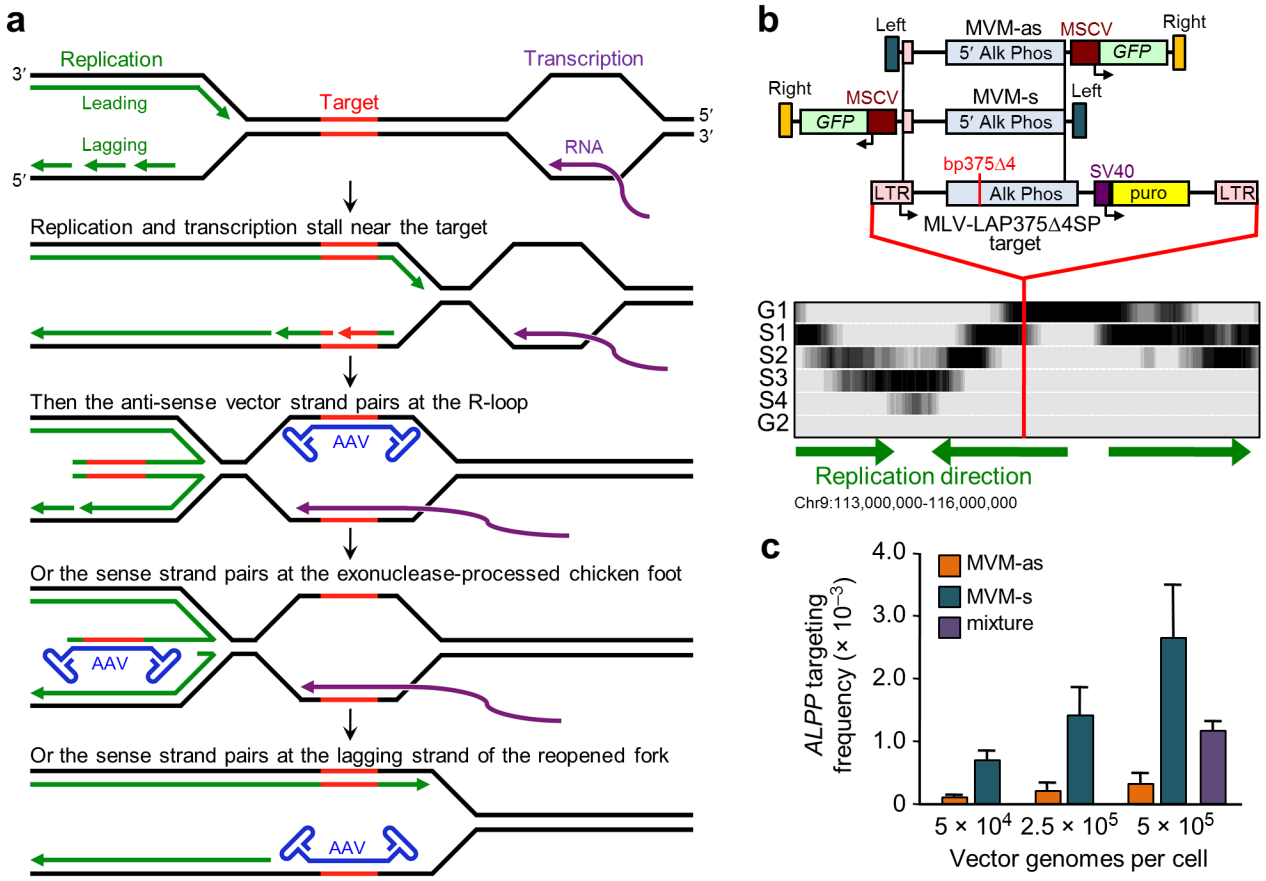

Figure 6. Stalled replication forks may promote vector pairing at target loci

(a) Model of a target locus transcribed in the opposite direction to 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 is shown with vectors containing sense or anti-sense 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 gammaretroviral vector MLV-LAP375Δ4SP with puromycin selection [50]. Arrows indicate transcription start sites. The HT-1080 Repli-Seq data for this portion of human chromosome 9 is shown below the vector maps, with the target site located at bp 114,505,235. MSCV, murine stem cell virus promoter; SV40, SV40 viral promoter; *GFP*, green fluorescent protein gene; *puro*, puromycin resistance gene. **(c)** Average targeting frequencies (with standard deviations) of the sense and antisense MVM vectors when infections were done at the indicated multiplicities of infection [50]. The “mixture” contained 2.5×10^5 vector genomes per cell of both MVM-s and MVM-as.

Table 1

Summary of targeting, plasmid rescue, and sequencing results.

Experimental step	Targeted	Untargeted
Transduced cells	HT-1080/LHSN63Δ53O*	HT-1080/LHSNO
Targeting vector	AAV2-HSN5′	none
Total G418-resistant colonies	7,950	75,000
Bacterial colonies expanded	9,000	9,000
Total sequence reads	12,181,202	16,985,709
Reads containing terminal 10 bp of provirus LTR	41,498	29,914
Reads with perfect alignment of junction sequence to human genome	16,462	4,393
Total distinct sites identified†	3,853	2,665
Uniquely mappable sites‡	2,015	1,928
Ambiguous sites§	1,838	737

*1.4×10^7 HT-1080/LHSN63Δ53O cells were transduced with AAV2-HSN5′ at a multiplicity of infection of 10,000. The average targeting frequency, expressed as the number of G418-resistant colony-forming units divided by the total number of colony-forming units, was 5.37×10^{-4}.

†Number obtained by eliminating duplicates, reads without adjacent vector sequence, and reads for which the human genome contained the terminal 10 bp of the proviral LTR.

‡These loci were used to compare targeted and control sites.

§Ambiguous sites are characterized in Supplementary Table 2.

Table 2

Chromosomal features associated with targeted and untargeted sites.

Chromosomal feature*	Percent of targeted sites	Percent of untargeted sites	*P* value†
SINEs	10.8	9.4	
LINEs	10.0	8.8	
LTRs	4.8	5.8	
DNA repeats	3.9	3.2	
Simple repeats‡	2.3	1.9	
Microsatellite repeats§	0.05	0.10	
GT repeats	<0.05	<0.05	
CpG islands	5.7	5.9	
DNase I hypersensitive sites	16.7	17.0	
RefSeq transcription units	58.6	54.1	<0.005
Introns	53.6	49.1	<0.006
Exons	5.8	6.0	

*Repeat elements defined as in the UCSC Genome Table Browser. SINE, short interspersed nuclear element; LINE, long interspersed nuclear element.

†*P* values of ≥ 0.05 are not shown and were not considered statistically significant.

‡Simple repeats are defined as simple tandem repeats of any period ≥ 1 nucleotide.

§Microsatellites are simple di- and trinucleotide repeats with at least 15 copies of the repeat.