Supporting Information

Ferris et al. 10.1073/pnas.0914142107

SI Materials and Methods

Bioinformatics Analysis of Integration Sites. Integration junction site sequences were first trimmed for LTR and linker sequences and then mapped to the mouse genome [mouse genome build mm9, July 2007, University of California Santa Cruz (UCSC) genome website] using BLAT and BLAST. Integration sites were considered to be authentic if the sequence (*i*) began within 3 bp of the end of the HIV-1 LTR, (*ii*) had a match to the mouse genome with at least 20 bp in length and 95% identity, and (*iii*) had a unique best hit to the mouse genome. Because PCR amplification can produce multiple copies of the same integration site, in any given experiment, each specific integration site was counted once, even if it appeared many times in the analysis. Genomic feature tables for mm9 were downloaded from the UCSC genome website and used for association analysis. Host sequences adjacent to integration sites were classified as repeats if there were more than five indistinguishable hits in the genome. These integrations could not be mapped to a unique location in the genome. Most of the integrations in repeats were in L1 or LTR/ERVK elements. To facilitate comparisons of datasets compiled for the different CBD–IBD fusions, the datasets were normalized to 10,000 integration sites. Initially, integration hotspots were calculated as three or more sites within a 10-kb window in the genome for each of the genomic features. Because most LEDGF and HP1α–IBD-directed integrations occurred inside genes, the data were recalculated to enumerate genes with three or more integration sites within a gene body (see *Results* in main text). Because most ING2–IBD-directed integrations were located near TSS, the data were recalculated according to the number of integration sites within ±2.5 kb of TSSs.

Gene Expression Profiling. Total RNA was isolated from MEF-KO cells using an RNAeasy mini kit (Qiagen), and the RNA quality was checked on an Agilent Bioanalyzer. All samples used for microarray analysis had a high quality score (RIN >9). Two hundred nanograms of total RNA was reverse transcribed with random hexamer primers, amplified, and terminally labeled with biotin using the Affymetrix Whole Transcript Sense Target Labeling Kit. Four replicates were prepared, labeled, and hybridized to Affymetrix Mouse Gene ST 1.0 GeneChips and scanned on the Affymetrix GeneChip Scanner 3000. Data were collected using Affymetrix GCOS software. Average signal intensity for each probe and gene was calculated with Partek Genomics Suite software using the RMA normalization algorithm. Gene expression levels were compiled for 17,306 RefSeq genes for which the genomic coordinates agree between the Affymetrix annotation file and the mm9 RefSeq database from the UCSC website. These data were used for the correlation analysis with the integration site data.

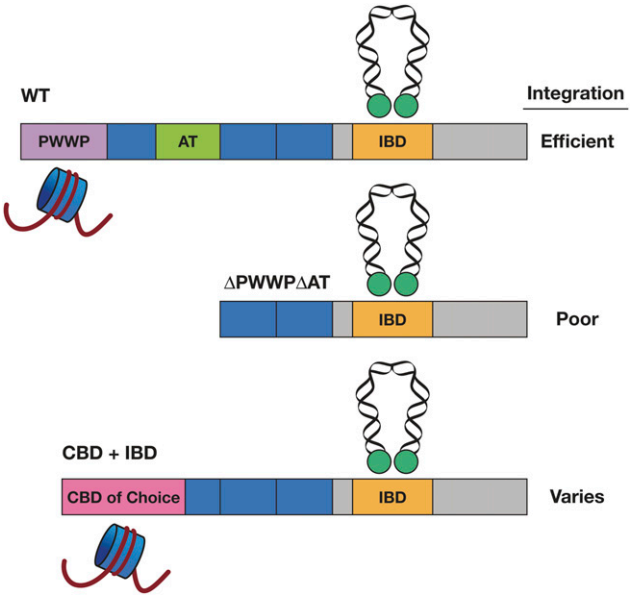

Fig. S1. Design of the CBD–IBD fusions. The figure shows a schematic diagram (not to scale) of the structure of LEDGF and the design of the CBD–IDB fusions. The PWWP domain is in purple, the AT hooks in green, and the IBD near the C terminus is in orange. The PWWP domain is shown interacting with a nucleosome (the nucleosome core is in blue, DNA in red); the IBD is shown interacting with an HIV-1 PIC. The linear viral DNA, which is ≈10 kb, is shown in black, IN is green. The ends of the linear viral DNA are held together by protein; IN is bound to both ends of the viral DNA. Other viral proteins may also be present in the PIC. In the center of the panel, an amino-terminally truncated LEDGF is shown. Because the PWWP domain and the AT hooks have been removed, the truncated LEDGF does not bind chromatin/DNA, and HIV-1 integration efficiency is poor (1–3). However, as shown at the bottom, if the missing PWWP domain and AT hooks are replaced with a CBD from another protein, the resulting CBD–IBD fusion protein binds chromatin, restoring the ability of HIV-1 to infect cells efficiently (4).

1. Shun MC, et al. (2008) Identification and characterization of PWWP domain residues critical for LEDGF/p75 chromatin binding and human immunodeficiency virus type 1 infectivity. *J Virol* 82:11555–11567.
2. Llano M, et al. (2006) An essential role for LEDGF/p75 in HIV integration. *Science* 314:461–464.
3. Shun MC, et al. (2007) LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev* 21:1767–1778.
4. Meehan AM, et al. (2009) LEDGF/p75 proteins with alternative chromatin tethers are functional HIV-1 cofactors. *PLoS Pathog* 5:e1000522.

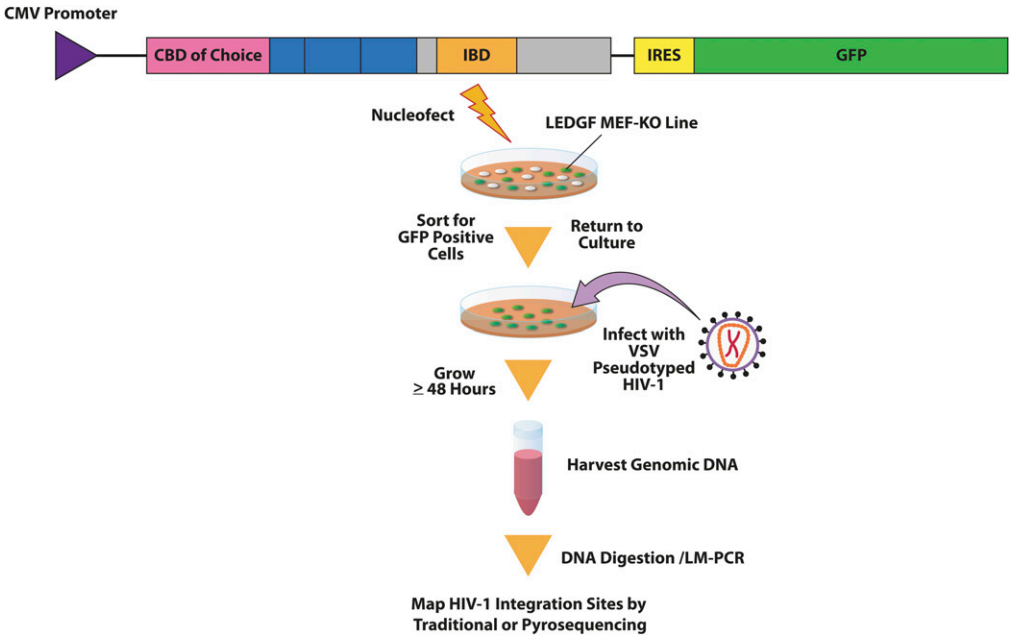

Fig. S2. Schematic diagram of the techniques used to isolate the integration sites. *Top:* Diagram of the CBD–IBD expression construct. Expression is driven by a CMV promoter (*Top Left*). Expression of the CBD–IBD fusion is linked, using an IRES, to the expression of GFP. The construct that expresses both the CBD–IBD fusion and GFP is introduced into MEF-KO cells by nucleofection. The cells are allowed to recover, and GFP-expressing cells are isolated by cell sorting. These cells are infected with a VSV-G pseudotyped HIV-1 vector and then cultured for 48 h. The cells are harvested and genomic DNA is prepared. The genomic DNA is digested, linkers are added, and the integration sites are selectively amplified and sequenced.

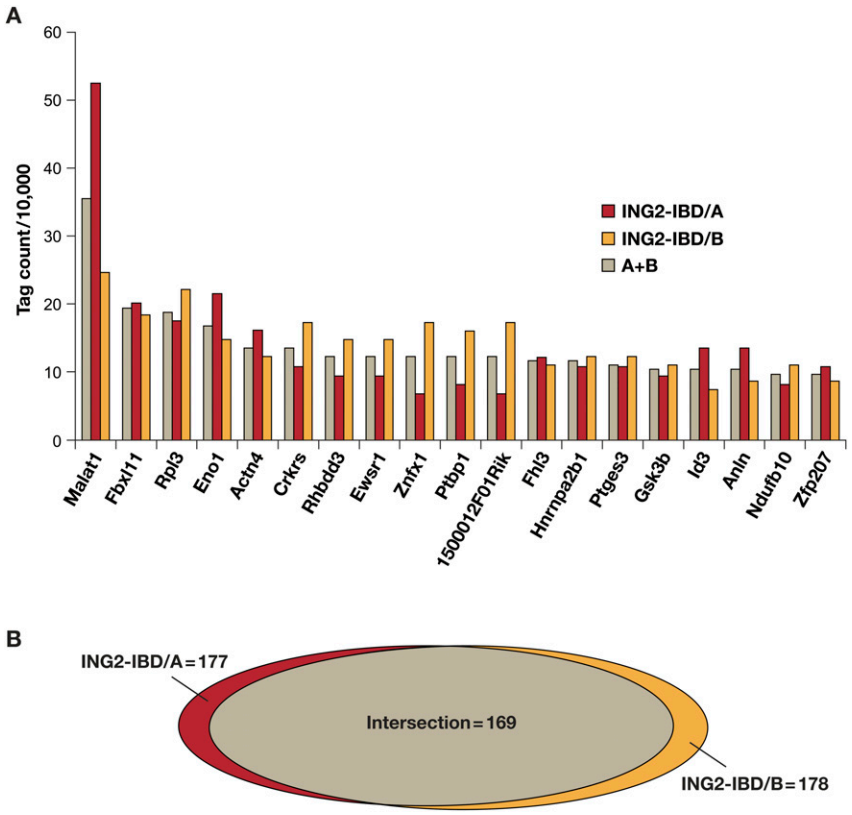

Fig. S3. The hotspots in the two ING2–IBD 454 experiments are similar. (*A*) The 19 most favored hotspots for HIV-1 DNA integration directed by the ING2–IBD fusion (the genes are identified on the *x* axis). The sites were ranked based on the number of integrations in the combined ING2–IBD dataset from the two independent 454 experiments (A+B). To simplify the comparison, all three datasets (run A, run B, and A+B) were normalized to 10,000 integrations. Although there is some variation in the rank order of the hotspots, there is good agreement between the data in the two separate experiments (run A and run B). (*B*) Venn diagram of the highly favored hotspots (five or more integrations from the two separate 454 experiments (run A and run B) after normalization to 10,000 total integrations). Of the 177 and 178 highly favored hotspots in the two experiments, 169 are in common.

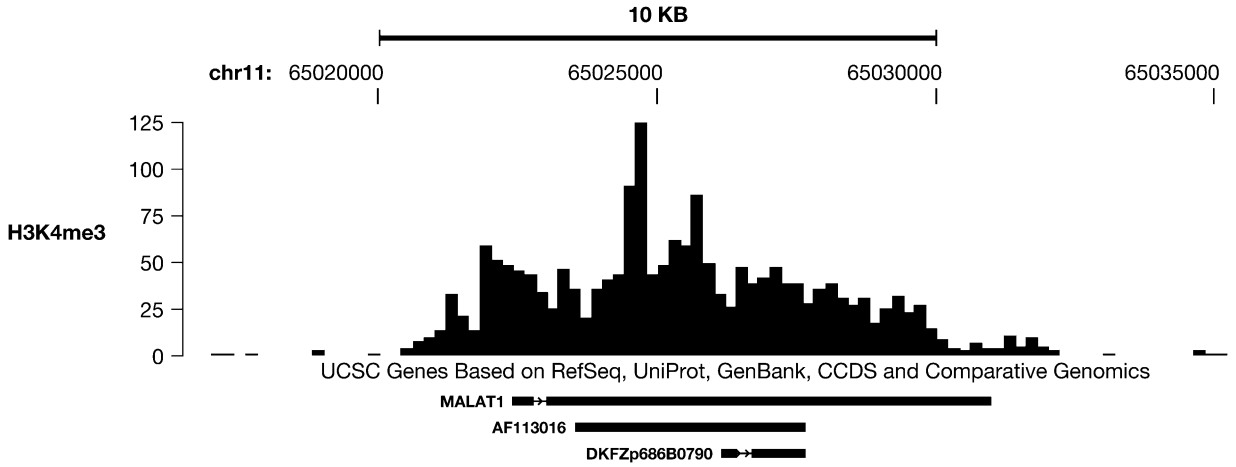

Fig. S4. H3K4me3 distribution in the Malat1 gene in human CD4+ T cells. The figure was generated from the data from Barski et al. [Barski A, et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129:823–837] and shows that the H3K4me3 marks are distributed throughout *Malat1*.

Table S1. Titer of an HIV-1 vector in MEF-KO cells expressing the CBD–IBD fusions

CBD in the fusion protein	Titer	Chromatin mark
Mock transfected MEF-KOs	0.01	None
Human LEDGF	1.0	??
ING2 PHD finger	0.6	H3K4me3
HP1α chromodomain	0.8	H3K9me2,3

Table S2. Sequence of linkers and primers

Linker upper strand	5′ GTAATACGACTCACTATAGGGCTCCGCTTAAGGGAC 3′
Linker lower strand MseI	5′ –PO4-TAGTCCCTTAAGCGGAG-NH2-C3 3′
Linker lower strand ASN (AvrII, SpeI, NheI)	5′ –PO4-CTAGGTCCCTTAAGCGGAG-NH2-C3 3′
Linker lower strand Tsp509I	5′ –PO4-AATTGTCCCTTAAGCGGAG-NH2-C3 3′
Linker primer P1	5′ GTAATACGACTCACTATAGGGC 3′
HIV primer P1	5′ GCCTCAATAAAGCTTGCCTTG 3′
Linker primer P2 w/ 454 sequencing primerB + key + MID	5′ GCCTTGCCAGCCCGCTCAGACGAGTGCGTAGGGCTCCGCTTAAGGGAC 3′
HIVp2.A w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGACGAGTGCGTTGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.B w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGACGCTCGACATGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.C w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGAGACGCACTCTGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.D w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGAGCACTGTAGTGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.E w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGATCAGACACGTGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.F w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGCGTGTCTCTATGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.G w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGCTCGCGTGTCTGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.H w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGTAGTATCAGCTGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.I w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGTCTCTATGCGTGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.J w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGTGATACGTCTTGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.K w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGTACTGAGCTATGTGACTCTGGTAACTAGAGATCCCTC 3′
HIVp2.L w/ 454 sequencing primerA + key + MID	5′ GCCTCCCTCGCGCCATCAGATATCGCGAGTGTGACTCTGGTAACTAGAGATCCCTC 3′