Identification of an Evolutionarily Conserved Domain in Human Lens Epithelium-derived Growth Factor/ Transcriptional Co-activator p75 (LEDGF/p75) That Binds HIV-1 Integrase

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 279, No. 47, Issue of November 19, pp. 48883–48892, 2004
Printed in U.S.A.

Identification of an Evolutionarily Conserved Domain in Human Lens Epithelium-derived Growth Factor/Transcriptional Co-activator p75 (LEDGF/p75) That Binds HIV-1 Integrase*[S]

Received for publication, June 7, 2004, and in revised form, August 24, 2004
Published, JBC Papers in Press, September 14, 2004, DOI 10.1074/jbc.M406307200

Peter Cherepanov‡§, Eric Devroe¶‖, Pamela A. Silver¶‖, and Alan Engelman‡§**

From the Departments of ‡Cancer Immunology and AIDS and ¶Cancer Biology, Dana-Farber Cancer Institute and the Departments of §Pathology and ‖Systems Biology, Harvard Medical School, Boston, Massachusetts 02115

Human lens epithelium-derived growth factor/transcriptional co-activator p75 (LEDGF/p75) protein was recently identified as a binding partner for HIV-1 integrase (IN) in human cells. In this work, we used biochemical and bioinformatic approaches to define the domain organization of LEDGF/p75. Using limited proteolysis and deletion mutagenesis we show that the protein contains a pair of evolutionarily conserved domains, assuming about 35% of its sequence. Whereas the N-terminal PWWP domain had been recognized previously, the second domain is novel. It is comprised of ~80 amino acid residues and is both necessary and sufficient for binding to HIV-1 IN. Strikingly, the integrase binding domain (IBD) is not unique to LEDGF/p75, as a second human protein, hepatoma-derived growth factor-related protein 2 (HRP2), contains a homologous sequence. LEDGF/p75 and HRP2 IBDs avidly bound HIV-1 IN in an *in vitro* GST pull-down assay and each full-length protein potently stimulated HIV-1 IN activity *in vitro*. LEDGF/p75 and HRP2 are predicted to share a similar domain organization and have an evident evolutionary and likely functional relationship.

Human immunodeficiency virus type 1 (HIV-1)[1] integrase (IN) accomplishes the joining of the reverse-transcribed viral genome to a host cell chromosome (for reviews see Refs 1–3). Its activity is essential for viral replication and spread in primary cells and contributes to the persistence of the viral infection *in vivo* (4, 5). Blocking HIV-1 IN activity by specific inhibitors was shown to arrest viral spread in cell culture (6). HIV-1 IN, born to the transposase family of DNA transferases, is structurally and mechanistically similar to Mu phage and Tn5 transposases. Its active site formed by the three acidic residues Asp^{64}, Asp^{116}, and Glu^{152} (known as the "DDE motif") is located within the structurally conserved catalytic core domain. The core domain is contained within residues 50–212 of the protein and flanked by the N-terminal HHCC-type zinc finger and the C-terminal DNA binding domains. Similar to the related bacterial transposases, retroviral INs form multimers, although the true stoichiometry of IN within the retroviral preintegration complex (PIC) is not known (7).

Mutations in HIV-1 IN display a wide range of phenotypes, affecting viral replication at the integration step (class I mutants), or causing various pleiotropic effects on virion morphogenesis and reverse transcription (class II) (8). Pleiotropic phenotypes of many IN mutants advocate that IN might have additional functions in viral replication. Thus, a role for IN in reverse transcription has been proposed (9). The complex phenotypes of class II mutants could potentially be explained by failure of the mutant INs to interact with viral reverse transcriptase and/or a host cell factor(s). A number of cellular and viral proteins were suggested to participate in retroviral integration (for a review see Ref. 10). Furthermore, several proteins were reported to directly interact with HIV-1 IN, including viral reverse transcriptase (9), a component of the SWI-SNF chromatin-remodeling complex INI1 (11), uracil DNA glycosylase UNG2 (12), heat shock protein HSP60 (13), a DNA repair protein Rad18 (14), a Polycomb group protein EED (15) and lens epithelium-derived growth factor/transcriptional co-activator p75 (LEDGF/p75) (Ref. 16, for a review see Ref. 17). The exact roles of these proteins and their importance to viral replication have yet to be determined. However, when HIV-1 or feline immunodeficiency virus (FIV) INs are expressed separately from other viral proteins, endogenous host-cell LEDGF/p75 appears to be the dominant interactor, accounting for their nuclear/chromosomal accumulation (16, 18, 19). LEDGF/p75 protein markedly stimulated HIV-1 IN activity *in vitro* and was recently reported to be associated with functional HIV-1 PICs (16, 19). These data cumulatively suggest that LEDGF/p75 and possibly its homologs pose as cellular host factors in retroviral replication likely acting at the levels of chromosomal targeting, and/or integration of viral cDNA (17).

LEDGF/p75 belongs to a family of hepatoma-derived growth factor (HDGF)-related proteins (HRPs). Five mammalian HRPs are known: HDGF, HRP1, HRP2, HRP3, HRP4, and

* This work was supported by an HHMI Predoctoral Fellowship (to E. D.), the Claudia Adams Barr Fund (to P. A. S.), and National Institutes of Health Grant AI39394 (to A. E.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains Supplementary Materials.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AY728140, AY728141, and AY728142.

** To whom correspondence should be addressed: Dept. of Cancer Immunology and AIDS, Dana-Farber Cancer Institute, 44 Binney St., Boston, MA 02115. Tel.: 617-632-4361; Fax: 617-632-3113; E-mail: alan_engelman@dfci.harvard.edu.

[1] The abbreviations used are: HIV-1, human immunodeficiency virus type 1; BLAST, basic local alignment search tool; BSA, bovine serum albumin; CypA, cyclophilin A; DTT, dithiothreitol; EST, expressed sequence tag; FIV, feline immunodeficiency virus; GST, glutathione *S*-transferase; HA, hemagglutinin; HDGF, hepatoma-derived growth factor; HMGA, high mobility group superfamily A; HRP, HDGF-related protein; IBD, integrase-binding domain; IN, integrase; LEDGF, lens epithelium-derived growth factor; MALDI MS, matrix-assisted laser desorption/ionization mass spectrometry; NLS, nuclear localization signal; NP40, Nonidet P40; ORF, open reading frame; PIC, preintegration complex; PMSF, phenylmethylsulfonyl fluoride; SMART, simple modular architecture research tool; TFIIS, transcription factor IIS; TR, trypsin-resistant; Tricine, *N*-[2-hydroxy-1,1-bis(hydroxymethyl)ethyl]glycine; HPLC, high performance liquid chromatography.

LEDGF/p75 (see Supplementary Table I) (20, 21). The characteristic feature of these proteins is a high degree of sequence homology within their N-terminal 90–95 residues, spanning the PWWP domain (InterPro accession number IPR000313) (20, 22). This domain, named for a conserved although not invariant Pro-Trp-Trp-Pro motif, extends for about 70 residues (22). Several dozen other PWWP domain-containing proteins have been described, including Wolf-Hirschhorn syndrome candidate 1 gene product WHSC1, mismatch repair protein MSH6, mammalian DNA methyltransferases Dnmt3a and Dnmt3b, and a plant homolog of ataxia telangiectasia-mutated protein kinase (22). PWWP domains seem to be distantly related to the Tudor and Chromo domains and are thought to mediate protein-protein interactions involved in regulation of chromatin structure (22, 23). Noteworthy, the PWWP domain of Dnmt3b methyltransferase was recently shown to be essential for chromatin association of the protein (24). Recognizable orthologs of mammalian HRPs seem to be present in all vertebrates, although proteins containing PWWP domains are more widespread and occur throughout *eukaryota*, including yeast (22). Apart from their homologous N-terminal PWWP domains, HRPs show little sequence similarity.

Cellular functions of LEDGF/p75 and other HRPs have not been studied in detail. Like other PWWP domain-containing proteins, HRPs are imported into the nucleus (21, 22, 25, 26). Chromatin association has so far been demonstrated only for LEDGF/p75 (16, 18, 27). LEDGF/p75 was implicated in the regulation of expression of stress response related genes, such as Hsp27, αB-crystallin, and antioxidant protein 2, presumably through binding to heat shock and stress-related regulatory elements in the promoters of the target genes (28, 29). Overexpression of LEDGF/p75 was reported to enhance cell viability under conditions of serum starvation, thermal, and oxidative stress (28, 30). During apoptosis, LEDGF/p75 is subject to cleavage by caspases that abolishes its activity as a cell-survival factor (30).

In this work, we studied the evolutionary conservation and domain organization of LEDGF/p75. We identified the HIV-1 IN binding domain (IBD) in this protein and found that another human protein, HRP2, can bind HIV-1 IN via a homologous domain and stimulate its activity *in vitro*.

EXPERIMENTAL PROCEDURES

Isolation and Sequence Analysis of LEDGF/p75 and HRP2 cDNAs—A wealth of expressed sequence tags (ESTs) representing fragments of cDNAs encoding homologs of human LEDGF/p75 from various vertebrate sources were readily identified by searching the NCBI sequence data base with translating basic local alignment search tool (BLAST) (www.ncbi.nlm.nih.gov/BLAST). ESTs with the following GenBank™ accession numbers gave sufficient sequence information to design primers for PCR amplification of the complete coding regions of *Gallus gallus* LEDGF/p75 cDNA: BU112216, AJ394255, BU332575, BU129859, and CN231179. ESTs for the *Xenopus laevis* ortholog were: BJ042207, BJ055881, BX852842, BU912001, and BQ729240. Total RNA isolated from *G. gallus* pro-B cell line DT40 or kidney tissue from an adult male *X. laevis* specimen was reverse-transcribed using random-primed SuperScript III reverse transcriptase (Invitrogen). The complete coding region of chicken LEDGF/p75 cDNA was amplified using Expand DNA polymerase (Roche Applied Science) and primers: 5′-CACGGCGGCGAGACAAC/5′-AGATTTCAAATGCAATCCTCTTC. The primers used to amplify the frog cDNA were: 5′-TGCCTGAATTTCGTCGAG and 5′-AATGACCACACGAGTGTGA. The resulting PCR fragments were subcloned into the pCR4-TOPO vector (Invitrogen), and three independent clones for each cDNA were sequenced.

The 3′-terminal part of the *Danio rerio* (zebrafish) HRP2 cDNA was obtained from the following ESTs: BI886683, AW154327, CD586524, BQ480774, and AL916221. The resulting contig was used to search through the zebrafish genome assembly (www.ensembl.org/Danio_rerio/) using nucleotide BLAST. The gene was identified on chromosome 22 and four exons containing the available portion of HRP2 cDNA sequence could be readily matched to the genomic sequence. Only one fragment predicted to encode a PWWP domain could be identified within the upstream 30 kb genomic sequence using PROSITE (us.expasy.org/prosite/). Assuming this sequence to represent the beginning of the HRP2 open reading frame (ORF), two sets of PCR primers were designed to amplify the cDNA: 5′-GTGGACGGATAGAAACG/5′-GAAGGAAGCCAAGGTGTG and 5′-AACGAGCAGAACGAGGAG/5′-GTTTGTGAGCATAAAAGGAG. Random-primed cDNA prepared from a sample of *D. rerio* kidney total RNA was used as template. PCRs with both primer pairs readily amplified fragments with the expected size of about 2.1 kb. Sequence analysis agreed with the chromosomal sequence and confirmed homology to human and mouse HRP2. The *D. rerio* HRP2 gene spans about 24.6 kb on the chromosome 22; the coding region of the cDNA is derived from 18 exons. The complete ORF from the *X. laevis* HRP2 cDNA was reconstructed from the following ESTs: CA789647, AW643477, BJ039312, BJ626283, BJ619541, BF612425, BU916767, CA981423, BJ054046, BJ642669, BE678817, BE678996, BX853753, BF426654, BJ622360, CD363094, BJ639036, BG234506, BJ086552, BJ050372, and BG812065. Partial *G. gallus* HRP2 cDNA sequence was obtained from a contig of the following ESTs: BU261466, BU324804, BU392278, CD727797, BU351707, BU347241, AI981158, BU141299, BU428133, BU236024.

Protein Secondary Structure Prediction and Sequence Analysis—Secondary structure prediction was done using the PROFsec and NORSp programs accessed through the PredictProtein server (cubic.bioc.columbia.edu) (31, 32). Hydrophobicity profiles (33) were analyzed using BioAnnotator software (InforMax Inc.). Multiple sequence alignments were done with AlignX (InforMax Inc.) using BLOSUM62 or GONNET matrices (34, 35). Homology between IBDs and the N-terminal domain of transcription factor IIS (TFIIS) was found using InterProScan release 7.2 (www.ebi.ac.uk/InterProScan/) (36) and SMART version 4.0 (smart.embl-heidelberg.de/) (37).

DNA Constructs for Bacterial Protein Expression—All glutathione *S*-transferase (GST)-LEDGF/p75 fusion constructs used for protein expression in this work were based on the pGEX-4T1 vector (Amersham Biosciences). The full-length LEDGF/p75 ORF and its fragments were PCR-amplified using *Pfu*-Ultra DNA polymerase (Stratagene). Sense primers were designed to incorporate a BamHI restriction site followed directly by the first codon of the relevant LEDGF/p75 fragment; antisense primers contained a stop codon (TGA) directly following the last codon. PCR fragments were digested with BamHI and subcloned between BamHI and SmaI sites of pGEX-4T1.

To clone the putative IBD of HRP2, a fragment coding for residues 470–593 of the human protein was PCR-amplified from random-primed HeLa cDNA using Expand DNA polymerase and the following primers: 5′-GCGTGGATCCTCCGTGGAGGAGAAGCTGCAG/5′-CCCTCACTTGTCCTCCGCCTTCTCC. The resulting PCR fragment was digested with BamHI and ligated into BamHI/SmaI-digested pGEX-4T1. The full-length HRP2 ORF was PCR-amplified from cDNA clone MGC2641 (American Type Culture Collection) using primers 5′-GCGTGGATCCATGCCACACGCCTTCAAGCC and 5′-GCTCAGCTCTCCTCGTCCAGGGCCTC; the PCR fragment digested with BamHI was subcloned between BamHI and SmaI sites of pGEX-6P3, resulting in pCP-GSTHRP2. For expression of non-tagged full-length HRP2, the entire HRP2 ORF was amplified using primers 5′-TGCCACACGCCTTCAAGCC and 5′-GTTTTCACCGTCATCAC, the resulting PCR fragment was digested with XhoI and cloned between NdeI and XhoI sites of pRSETB (Invitrogen) (the vector NdeI terminus was filled-in using T4 DNA polymerase) giving pCP-NatHRP2. Non-modified pGEX-4T1 was used to produce GST as a control for pull-down experiments. Plasmids pCPNat75 and pKB-IN6H were described previously (18).

DNA Constructs for Expression in Human Cells—Plasmids pBHA-P75 and pCPHA-HRP2 expressed human LEDGF/p75 and HRP2 with N-terminal influenza hemagglutinin (HA) tags, respectively, under the control of the human cytomegalovirus immediate-early promoter. To make pBHA-P75, the LEDGF/p75 ORF was PCR-amplified using 5′-CCGCGGATCCGACACCATGGCATACCCATACGACGTCCCAGACTACGCTACTCGCGATTTCAAACCTGGAGACC/5′-ATAAGAATGCGGCCGCCTAGTTATCTAGTGTAGAATCC and *Pfu*-Ultra DNA polymerase. The resulting amplicon was digested with BamHI and NotI and ligated into BamHI/NotI-digested pcDNA6-V5-HisB (Invitrogen). The BamHI/XhoI fragment of pCP-GSTHRP2 carrying the entire HRP2 ORF was re-cloned between BglII/XhoI sites of the pCPHA-NLS vector, fusing the 5′-end of the HRP2 ORF directly to the HA tag coding sequence, resulting in pCPHA-HRP2. The HA tag fusion vector pCPHA-NLS was made by first disrupting the BglII site in pcDNA6-V5-HisB by digesting it with BglII, filling-in using *Pfu* polymerase, and religation resulting in pcDNA6ΔBgl. A DNA fragment obtained by annealing synthetic oligonucleotides 5′-CGGGAAGCTTAGACACCATGGCCTAC-

CCTTACGACGTGCCCGACTACGCCAGATCTG and 5′-GGTGGGATCCCTCCACCTTCCGCTTCTTCTTGGGAGGGCCAGATCTGGCGTAGTCG followed by extension using Sequenase Version 2.0 T7 DNA polymerase (Amersham Biosciences) was restricted with HindIII and BamHI and then ligated with HindIII/BamHI-digested pcDNA6ΔBgl. The resulting pCPHA-NLS vector encodes for the HA tag fused to the simian virus 40 large T antigen nuclear localization signal (NLS), with an intervening BglII restriction site. The construct from Bram *et al.* (38) was used to express human cyclophilin A (CypA) with a C-terminal HA tag. This plasmid will be referred to here as pCypA-HA. The construct pED-FLAG-IN was used to express FLAG-tagged HIV-1 IN (39). All expression constructs were verified to be free of inadvertent mutations by sequencing.

Protein Expression and Purification—GST fusion proteins were produced in *Escherichia coli B* strain BL21. Shake-flask cultures grown to an optical density of 0.9–1.0 at 600 nm were induced for 3 h by addition of 0.5 mM isopropyl-thio-β-D-galactopyranoside at 37 °C. The temperature of induction was reduced to 28 °C to increase the stability of full-length GST-LEDGF/p75, GST-LEDGF-(1–325), and GST-LEDGF-(1–471). Bacteria were disrupted by sonication in 500 mM NaCl, 5 mM dithiothreitol (DTT), 1 mM EDTA, 0.2 mM phenylmethylsulfonyl fluoride (PMSF), 50 mM Tris-HCl, pH 7.2. Because the fusion proteins differed in terms of stability, expression levels, and solubility, purification procedures had to be adjusted accordingly. Briefly, GST fusions containing full-length LEDGF/p75 and fragments LEDGF-(1–471) and LEDGF-(1–325) were isolated from soluble fractions of bacterial lysates by adsorption onto glutathione-Sepharose (Amersham Biosciences). Proteins eluted with 25 mM reduced glutathione (Sigma-Aldrich) in 500 mM NaCl/50 mM Tris-HCl, pH 7.2 were further purified by chromatography on 5-ml HiTrap heparin and SP-Sepharose columns (Amersham Biosciences) to partially remove proteolytic fragments. Both columns were operated in 50 mM NaH_2PO_4, pH 7.2, and proteins were eluted with a linear gradient of 0.2 M to 0.8 M NaCl. Peak fractions were pooled and diluted 1:3 with 50 mM NaH_2PO_4, pH 7.2 before injection onto SP-Sepharose. GST fusions of LEDGF/p75 fragments 326–530, 326–471, 347–471, 366–471, the HRP2 fragment 326–530, as well as free GST protein were purified from soluble fractions in one step on glutathione-Sepharose. GST fused to LEDGF-(347–429) was expressed in an insoluble form. To purify this protein, inclusion bodies were dissolved in 8 M urea, 100 mM NaCl, 1 mM DTT, 0.5 mM EDTA, 25 mM Tris-HCl, pH 7.2 and refolded by dilution into 10-fold excess of cold 100 mM Tris-HCl, pH 8.5. The protein was purified by chromatography on glutathione-Sepharose. GST fusion proteins dialyzed against excess 200 mM NaCl/25 mM Tris-HCl, pH 7.2 were concentrated using Centricon-30 (Millipore) when necessary, supplemented with 10% glycerol and stored at −70 °C after flash-freezing in liquid nitrogen.

Non-tagged HRP2 was induced in Rosetta2 (DE3) cells (Novagen) for 3 h at 28 °C by addition of 0.25 mM isopropyl-thio-β-D-galactopyranoside. The bacteria were disrupted by sonication in 1 M NaCl, 50 mM NaH_2PO_4, 5 mM DTT, 0.3 mM PMSF, pH 7.7. The lysate precleared by centrifugation at 15,000 rpm for 30 min was diluted with 50 mM NaH_2PO_4, pH 7.2 to reduce conductivity to 24 mS/cm and injected into a 5-ml HiTrap heparin column. Bound proteins were eluted with a linear salt gradient in 50 mM NaH_2PO_4, pH 7.2. HRP2 protein eluting at ~500 mM NaCl was collected, the peak fractions were pooled, diluted 1:3 in 50 mM NaH_2PO_4, pH 7.2 and injected into a 5-ml HiTrap SP-Sepharose column. The protein was eluted with a linear gradient of NaCl from 0.15 to 1.0 M in 50 mM NaH_2PO_4, pH 7.2. Fractions containing HRP2 were pooled, concentrated using a Centricon device, and further separated on a Superdex 200HR column (Amersham Biosciences) at 0.25 ml/min in 250 mM NaCl, 50 mM NaH_2PO_4, pH 7.2. The purified protein was concentrated to 3.5 mg/ml, supplemented with 10% glycerol and stored at −70 °C after flash-freezing in liquid nitrogen. Non-tagged LEDGF/p75 and His_6-tagged HIV-1 IN were produced in *E. coli* strain BL21(DE3), pLysS using pCPNat75 and pKB-IN6H, respectively, and purified according to published procedures (18, 40). LEDGF-(326–530), LEDGF-(347–471), and HRP2-(470–593) fragments released from GST by digestion with thrombin (Sigma-Aldrich) were further purified by cation exchange chromatography on SP-Sepharose using a linear 0.1–0.5 M NaCl gradient in 50 mM sodium phosphate buffer, pH 7.2. Protein concentrations were determined using the Bradford colorimetric assay (Bio-Rad) employing bovine serum albumin (BSA) as a standard.

N-terminal Microsequencing and Mass Spectrometry—To determine the N-terminal residues of trypsin-resistant (TR) 1 and TR2 peptides, tryptic fragments resulting from digestion of 10 μg of LEDGF/p75 were separated by 10–20% Tricine-SDS-PAGE (Invitrogen) and electroblotted onto Sequi-Blot polyvinylidene difluoride membrane (Bio-Rad). Bands excised from Coomassie Blue R250-stained membranes were subjected to Edman degradation in a Procise protein sequencer (Applied Biosystems). In-gel trypsin digestion and peptide extraction were done as described (41). To determine molecular masses of the intact TR1 and TR2 peptides, a mixture of digestion products was separated on a 2.1 × 250 mm Vidyac C8 column. Fractions containing the TR1 and TR2 fragments were analyzed by matrix-assisted laser desorption/ionization mass spectrometry (MALDI MS).

GST Pull-down Assay—Purified GST fusion proteins were adsorbed onto glutathione-Sepharose beads (Amersham Biosciences) in 200 mM NaCl, 5 mM DTT, 25 mM Tris-HCl, pH 7.3, using 125 μl (settled volume) beads per 40 μg of protein. After 4 h at 4 °C, the beads were washed in excess buffer and stored on ice. To test for IN binding, 10 μl of glutathione-Sepharose beads carrying GST fusion proteins were resuspended in 200 μl of cold PD buffer (150 mM NaCl, 5 mM $MgCl_2$, 5 mM DTT, 0.1% Nonidet P40, 25 mM Tris-HCl, pH 7.4) containing 10 μg of BSA. After addition of 3.8 μg of His_6-tagged HIV-1 IN the samples were gently rocked for 1.5–2 h at 4 °C and left for an additional 15–30 min without mixing. After careful aspiration of the supernatant, the settled beads were resuspended in 700 μl of fresh PD buffer, and allowed to sediment without centrifugation. The wash was repeated twice and bound proteins were eluted in SDS-containing sample buffer and analyzed by SDS-PAGE. In certain cases IN pull-down was confirmed by Western blotting using polyclonal anti-IN serum (42).

Cell Transfection and Immunoprecipitation—293T cells were maintained in Dulbecco's modified Eagle's medium containing 10% fetal calf serum (Invitrogen), 5 units/ml penicillin and 5 μg/ml streptomycin. 293T cells grown in 6-well dishes to 30–50% confluency were transfected with 0.5 μg of pCypA-HA, pBHA-P75, or pCPHA-HRP2 along with 0.5 μg of pED-FLAG-IN per well using FuGENE 6 transfection reagent (Roche Applied Science). Twenty-four hours post-transfection, cells were washed in cold phosphate-buffered saline, and lysed in 400 μl of cell lysis buffer (500 mM NaCl, 0.5% Triton X-100, 50 mM HEPES pH 7.9, 5% glycerol, 2 mM $MgCl_2$, 25 mM β-glycerophosphate, 1 mM sodium orthovanadate, supplemented with complete protease inhibitor mixture (Roche Applied Science)). The extracts were centrifuged at 19,000 × *g* to remove cell debris and precleared by incubation with 4 μl (settled volume) of protein G-Sepharose beads (Amersham Biosciences). Precleared supernatants were incubated with 4 μg of mouse anti-HA 12CA5 antibody (Roche Applied Science) at 4 °C, 4 μl of protein G-Sepharose beads were added, and the samples were left rocking for an additional hour. The beads were washed three times in cell lysis buffer, four times in reduced salt buffer (cell lysis buffer modified to contain 150 mM NaCl, 0.1% Triton X-100, and 0.1% Nonidet P-40). Whole cell extracts and immunoprecipitated proteins were resolved in 4–20% SDS-polyacrylamide gels. Following semi-dry transfer to polyvinylidene difluoride membranes, HA-tagged CypA, LEDGF/p75, and HRP2 proteins were detected by Western blotting using anti-HA 3F10 antibody conjugated to horseradish peroxidase (Roche Applied Science) and Western Lightning chemiluminescent reagent plus (PerkinElmer Life Sciences). FLAG-tagged IN was detected with anti-FLAG M2 antibody (Sigma-Aldrich) and goat anti-mouse IgG horseradish peroxidase conjugate (Jackson ImmunoResearch Laboratories).

RESULTS

Conservation of LEDGF/p75 Protein—Sequences of several mammalian LEDGF/p75 orthologs were available in public sequence databases. We identified ESTs representing partial cDNA sequences of *G. gallus* and *X. laevis* LEDGF/p75 cDNAs, which allowed us to clone and sequence complete LEDGF/p75 cDNAs from these species. On the basis of the obtained cDNA sequences, chicken and frog LEDGF/p75 were predicted to be composed of 579 and 564 amino acids, respectively, both somewhat larger than the 530-residue human ortholog. Alignment of the predicted amino acid sequences revealed ~48% identity between mammalian, avian, and amphibian LEDGF/p75 proteins (Supplementary Fig. S1). The plot in Fig. 1*A* summarizes this alignment by showing the degree of conservation along the protein sequence. Three regions of homology were evident (highlighted as *shaded boxes* in Fig. 1*A*). The most conserved fragment spanning residues 1–94 (conserved region I), which showed about 89% identity between human, chicken, and frog, corresponded to the PWWP domain (22). A 105-residue region spanning residues 351–455 displayed about 87% identity (re-

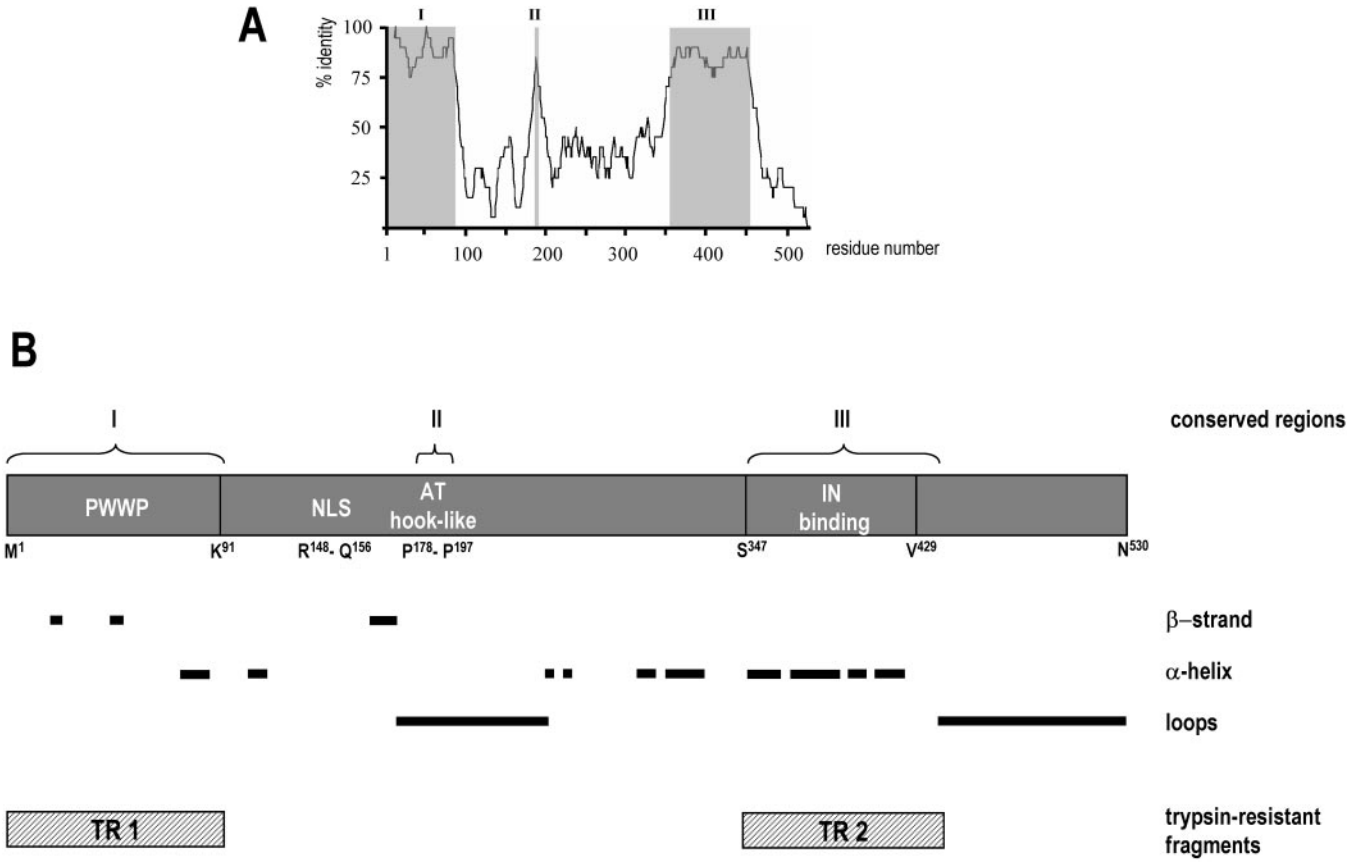

FIG. 1. **Sequence conservation and structural and functional elements within LEDGF/p75.** *A*, plot represents percentages of identical residues, calculated in windows of 20 residues from the alignment of predicted amino acid sequences of human, chicken, and frog orthologs. The original alignment was produced using BLOSUM matrix and can be viewed in Supplementary Fig. S1. Residue numbers correspond to the human sequence. Three regions of homology (I, II, and III) are highlighted as *shaded boxes*. The *G. gallus* and *X. laevis* sequences for this analysis were obtained in this study; the human LEDGF/p75 sequence was GenBank™ entry AAC25167. *B*, locations of the conserved PWWP domain (residues Met^{1}–Lys^{91}) (22), NLS (Arg^{148}–Gln^{156}) (43), high mobility group superfamily A (HMGA) AT hook-like sequence (Pro^{178}–Pro^{197}), and IBD (Ser^{347}–Val^{429}) are indicated. Positions of homology regions I, II, and III, as defined in Fig. 1*A*, are shown above the protein diagram. The β-strand and α-helical regions predicted by PROFsec with a confidence value higher than 6, and loop regions predicted by NORSp, are indicated as *thick lines* below the diagram. The *hatched boxes* indicate the positions of TR1 and TR2.

gion III). In addition, a short fragment involving residues 178–197 (region II) showed significant homology. Intuitively, these most conserved regions likely represent functional and/or structural determinants within the protein. The most variable regions encompassed an internal fragment flanking the PWWP domain (residues 94–177 in human LEDGF/p75, showing only about 13% identity) and the 60 C-terminal residues of the protein (20% identity). The single conserved feature of the first hypervariable region was a 7-residue sequence, ^{146}RRGRKRK^{152}, which partially overlaps the NLS in human LEDGF/p75 (residues 148–156) (43). Both chicken and frog LEDGF/p75 contain an insertion of 39 amino acids within the first hypervariable region (Supplementary Fig. S1).

Secondary Structure Prediction—We used the protein structure prediction program package available through the PredictProtein server (31) to analyze possible structural elements in LEDGF/p75. The protein is uncommonly rich in charged amino acids, accounting for about 42% of its sequence. Thus, it was not surprising that two extensive loop regions were predicted by the NORSp program (residues 177–250 and 440–530, Fig. 1*B*). Furthermore, analysis by PROFsec identified only relatively short regions, which are likely to be involved in stable secondary structure (Fig. 1*B*). By homology, the N-terminal 90 residues of the protein are known to constitute a PWWP domain (22). Prediction of β-strand elements followed by α-helices in that region is accurate, since a five-stranded β-barrel core and a C-terminal bundle of α-helices are conserved structural features of PWWP domains (44). The region encompassing residues 347–423 and matching homology region III (Fig. 1*A*) was predicted with high confidence to pack into four or five α-helices. Of note, this fragment, along with the N-terminal PWWP domain, span two mostly hydrophobic regions of LEDGF/p75 with average hydrophobicity indices above zero (data not shown).

Limited Proteolysis of LEDGF/p75—We used limited proteolysis (45) to probe the domain organization of LEDGF/p75. As the protein is rich in charged amino acids, a cleavage site for trypsin is predicted on average every 4–5 residues. Considering all Lys and Arg residues, the largest hypothetical LEDGF/p75 tryptic peptide was just 25 residues (Thr^{477}–Lys^{501}) with a molecular mass of about 2.6 kDa. We found that recombinant human LEDGF/p75 was indeed very sensitive to trypsin. A mass ratio of 250:1 of LEDGF/p75:protease yielded final proteolyzed products as well as semi-stable intermediates (Fig. 2*A*). The protease was quenched at different time points by addition of PMSF and reaction products were analyzed using Tris-glycine or Tricine SDS-PAGE. As quantified by densitometry of Coomassie-stained gels, ~60–70% of the protein became extinct after a relatively short exposure to trypsin (compare *lanes 1* and *6* in Fig. 2*A*). As proteolysis proceeded, two distinct polypeptides TR1 and TR2 with apparent molecular masses close to 10 kDa gradually accumulated at the expense of the intermediate cleavage products (Fig. 2*A*). Both TR1 and TR2 fragments persisted even after overnight digestion under these conditions (data not shown).

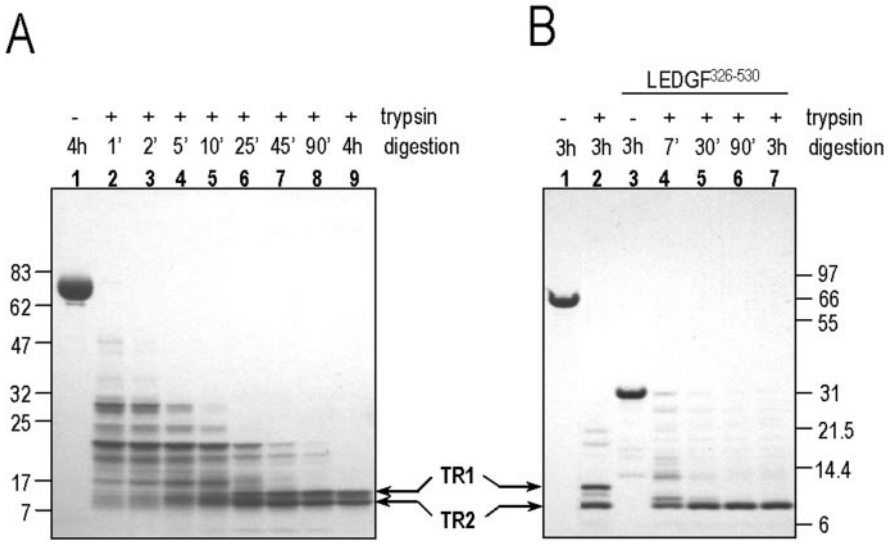

FIG. 2. **Proteolysis of recombinant LEDGF/p75 with trypsin.** *A*, LEDGF/p75 (25 μg) was incubated with 0.1 μg of trypsin (Sigma-Aldrich) in 60 μl of 100 mM NaCl, 2 mM $MgCl_2$, 4 mM DTT, 25 mM Tris-HCl, pH 7.4 at 22 °C. Aliquots taken at the indicated time points were quenched with PMSF. The reaction products were analyzed by SDS-PAGE on a 4–20% gel and detected by staining with Coomassie R250. *Lane 1*, mock digest with protease omitted. Positions of molecular mass markers (83, 62, 47, 32, 25, 17, and 7 kDa) are indicated. *B*, full-length LEDGF/p75 (*lanes 1* and *2*) or its fragment containing residues 326–530 (*lanes 3–7*) were incubated with (*lanes 2* and *4–7*) or without (*lanes 1* and *3*) trypsin, and the reactions were quenched at the indicated time points. Reactions were at 22 °C in the same buffer as in *A*. LEDGF/p75:trypsin ratios were 250:1 and 750:1 for full-length and the deletion mutant, respectively. The 10–20% Tricine gel was stained with Coomassie-R250. Positions of molecular mass markers (97, 66, 55, 31, 21.5, 14.4, and 6 kDa) are indicated. The migration positions of TR1 and TR2 are shown.

N-terminal sequencing of TR1 and TR2 fragments revealed that TR1 was derived from the N terminus part of LEDGF/p75, having the same N-terminal sequence as the full protein, *i.e.* NH_2-Met-Thr-Arg-Asp-Phe. TR2 originated from the C-terminal portion of the protein and contained two overlapping N termini: NH_2-Lys-Arg-Glu-Thr-Ser-Met- and NH_2-Glu-Thr-Ser-Met-Asp-Ser- corresponding to trypsin cleavage at peptide bonds Lys^{342}–Lys^{343} and Arg^{344}–Glu^{345}, respectively. To identify the C termini of the fragments, TR1 and TR2 were purified by reverse phase high performance liquid chromatography and their masses were determined by MALDI-MS. The molecular mass of the TR1 fragment was 11,424 ± 10 Da. The TR2 product represented a mixture of fragments of 11,348 ± 10 and 11,632 ± 10 Da. These data allowed us to unambiguously map the C termini of the TR fragments to LEDGF/p75 residues Lys^{100} for TR1 and Lys^{442} for TR2. Indeed, the calculated molecular mass of Met^1–Lys^{100} was 11,429.3 Da, whereas the masses of Lys^{343}–Lys^{442} and Glu^{345}–Lys^{442} fragments were 11,641.2 and 11,356.8 Da, respectively, which matched the experimentally determined masses well within confidence intervals. When a deletion mutant retaining the 206 C-terminal residues of LEDGF/p75 (residues 326–530) was exposed to trypsin, only the TR2 fragment was obtained (Fig. 2*B*, *lanes 3–7*). This result confirmed that TR1 and TR2 resided in the N- and C-terminal regions of LEDGF/p75, respectively. Although more than a dozen potential trypsin cleavage sites exist within fragments 1–100 and 345–442 of LEDGF/p75, both appeared to resist proteolysis, indicating that both are involved in stable structures.

In addition to trypsin, we tested proteinase K, thrombin, chymotrypsin, and Arg C proteases (data not shown). Unlike trypsin, digestion with proteinase K did not result in stable proteolytic products, however transient fragments of about 10 kDa in size were observed. Incubation of GST-LEDGF/p75 with thrombin resulted in multiple cuts within the putative loop region adjoining the PWWP domain. Chymotrypsin and Arg C proteases appeared less active than trypsin and although the fragments obtained confirmed the tryptic map, the cleavage patterns were more complex and longer incubation times were necessary to allow for accumulation of final products.

TR2 Is the Functional LEDGF/p75 IBD—To identify region(s) of LEDGF/p75 involved in the interaction with HIV-1 IN, we prepared a series of LEDGF/p75 deletion mutants. Mutants were expressed and purified as GST fusions, preadsorbed onto glutathione-Sepharose beads, and tested for their ability to pull-down recombinant HIV-1 IN. As can be seen from Fig. 3*A*, both the full-length protein (residues 1–530) and the mutant lacking the variable 59 C-terminal residues (1–471) readily bound HIV-1 IN (Fig. 3*A*, *lanes 9* and *12*). However, a more extended deletion from the C terminus disrupted interaction with IN, as LEDGF-(1–325) lacking 205 residues failed to pull down IN (*lane 10*). This result corroborates the previous finding that LEDGF/p52, an alternative splice form containing a unique 8-residue tail in place of LEDGF/p75 residues 326–530, did not bind HIV-1 IN (18). Furthermore, the C-terminal fragment of LEDGF/p75 containing residues 326–530 was sufficient to pull down HIV-1 IN (*lane 11*). By making another set of deletions, the IN binding function of LEDGF/p75 was mapped to just 83 amino acids, spanning residues 347–429 (Fig. 3*B*, *lane 15*; see also Supplementary Fig. S1). Importantly, this fragment lies within conserved region III of LEDGF/p75 (Fig. 1*A*) and the TR2 fragment (Fig. 2; see also Fig. 1*B* for summary). We found that further truncations from the N terminus of 347–429 abolished the interaction with IN (*lane 16*) and reduced the solubility of the recombinant protein (data not shown). Deletions from the C terminus of this fragment, on the other hand, profoundly affected stability of GST fusion proteins in *E. coli* (data not shown). These observations indicated that residues 347–429 of LEDGF/p75 span the IBD and comprise the minimal sequence required for its proper folding.

Of note, full-length LEDGF/p75, as well as deletion mutants containing the interdomain region (residues 150–325) were only marginally stable when expressed in bacteria, even when the temperature of induction was reduced. The bulk of the GST fusions recovered by adsorption to glutathione-Sepharose represented various proteolytic fragments. Due to dimerization of GST, it was not feasible to completely remove proteolytic fragments from preparations of GST-LEDGF/p75

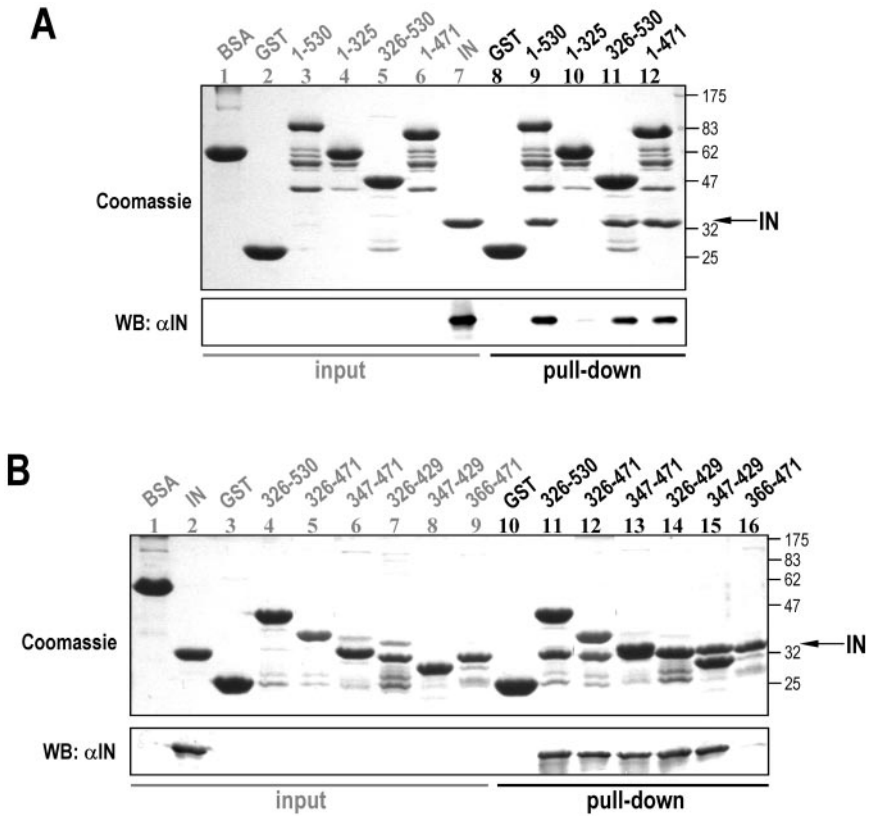

FIG. 3. **Mapping the IN binding determinant of LEDGF/p75.** *A*, eluted proteins following GST pull-down were separated through a 12% SDS-PAGE gel and stained with Coomassie R250 or analyzed by Western blot with polyclonal anti-IN serum. *Lanes 1–7*, input quantities of proteins used in the binding assays. BSA (*lane 1*), used as a nonspecific control, was included in each reaction. *Lanes 8–12* show results of pull-downs with GST (*lane 8*), GST fused to full-length LEDGF-(1–530) (*lane 9*), LEDGF-(1–325) (*lane* 10), LEDGF-(326–530) (*lane 11*), and LEDGF-(1–471) (*lane 12*). *B*, pull-down with GST (*lane 10*), or GST fused to LEDGF-(326–530) (*lane 11*), LEDGF-(326–471) (*lane* 12), LEDGF-(347–471) (*lane 13*), LEDGF-(326–429) (*lane 14*), LEDGF-(347–429) (*lane 15*), and LEDGF-(366–471) (*lane 16*). *Lanes 1–9* contained input quantities of protein. Samples were analyzed as i*n panel A*. The migration position of IN is indicated.

or fusions with LEDGF-(1–325) or LEDGF-(1–471) even after additional heparin affinity and cation exchange chromatography (Fig. 3*A*).

Identification of HRP2 as a Second IBD-containing Protein—Using translated BLAST to search for human cDNAs encoding polypeptides with homology to the LEDGF/p75 IBD we found that a second HDGF-related protein, HRP2, contains a very similar sequence within its C-terminal region. Because this region of homology is relatively short and occurs within largely divergent sequences, the similarity within C-terminal regions of LEDGF/p75 and HRP2 remained unnoticed until now. Fig. 4*A* presents an alignment of the human LEDGF/p75 IBD with the related sequence from HRP2 and includes their respective orthologs from different species. Human LEDGF/p75 and HRP2 proteins are about 48% identical within this region, and, considering conservative amino acid substitutions, the similarity exceeds 70%. Furthermore, predicted secondary structural elements within the two putative IBDs matched very well, with both domains demonstrating high α-helical content (Fig. 4*A*). We identified several ESTs encoding an HRP2 ortholog from *D. rerio*, which allowed us to clone and sequence its complete coding region. In addition, HRP2 cDNA from *X. laevis* could be completely reconstructed from available ESTs (see "Experimental Procedures"). Sequence alignment of human, frog, and fish HRP2 revealed high degrees of sequence conservation within the PWWP and IBD-like regions (regions I and III, Fig. 4*B*) (for a complete alignment see supplementary Fig. S2). An approximate 20-amino acid region of homology (region II) was similar to homology region II in LEDGF/p75, with each region containing several conserved Pro, Arg, and Lys residues. HRP2 region IV, however, appears unique to this protein. In addition, we also identified a hypothetical 475-residue protein CG7946 from *Drosophila melanogaster* (GenBank™ accession NP_651768, UniGene cluster Dm.4512) that contains an IBD-related sequence. This fragment, spanning CG7946 residues 318–400, shared about 21% identical and 46% similar residues with the HRP2 IBD (not shown). Intriguingly, since this protein is also predicted to possess an N-terminal PWWP domain, it likely represents an insect ortholog of HRP2. Additional searches using InterProScan and SMART (Simple Modular Architecture Research Tool) revealed homology between the IBDs and the N-terminal domain of TFIIS (SMART accession SM00509). Although the E-values reported by SMART for these hits were relatively high, equating to 3.1 and 1.2 for human LEDGF/p75 and HRP2 IBDs, respectively, the N-terminal domain of TFIIS seems to represent their closest relative among known protein domains. The TFIIS domain family includes four-helix bundle domains of TFIIS, elongin A, and CRSP70 (46).

To find out whether the putative IBD of HRP2 has affinity for HIV-1 IN, we fused a fragment spanning HRP2 residues 470–593 to GST and tested it in our pull-down assay. As seen in Fig. 5*A*, the HRP2 fragment readily interacted with recombinant HIV-1 IN, suggesting that HRP2 contains a functional IBD. To determine if HRP2 protein interacted with HIV-1 IN in human cells, 293T cells were transiently transfected with a plasmid encoding FLAG-tagged HIV-1 IN and a second plasmid encoding either HA-tagged LEDGF/p75, HRP2, or CypA. Cell extracts prepared 24 h post-transfection were immunoprecipitated with anti-HA 12CA5 antibody, and recovery of FLAG-IN was monitored by Western blotting. As shown in Fig. 5*B*, FLAG-IN was readily co-immunoprecipitated with HA-tagged LEDGF/p75 (*lane* 5). Significantly lower, but detectable amounts of FLAG-IN were recovered with HA-HRP2 (*lane 6*). In contrast, only negligible binding of FLAG-IN was detected with HA-CypA, which served as negative control (*lane 4*). We estimated that recovery of FLAG-IN upon co-immunoprecipitation with HA-HRP2 was 5–10-fold lower than with HA-LEDGF/p75, suggesting that at least under the conditions of this experiment, LEDGF/p75 had a greater affinity for HIV-1 IN than did HRP2.

Stimulation of HIV-1 IN Activity by LEDGF/p75 and HRP2—In accordance with previously reported results (16), recombinant LEDGF/p75 potently stimulated HIV-1 IN strand transfer

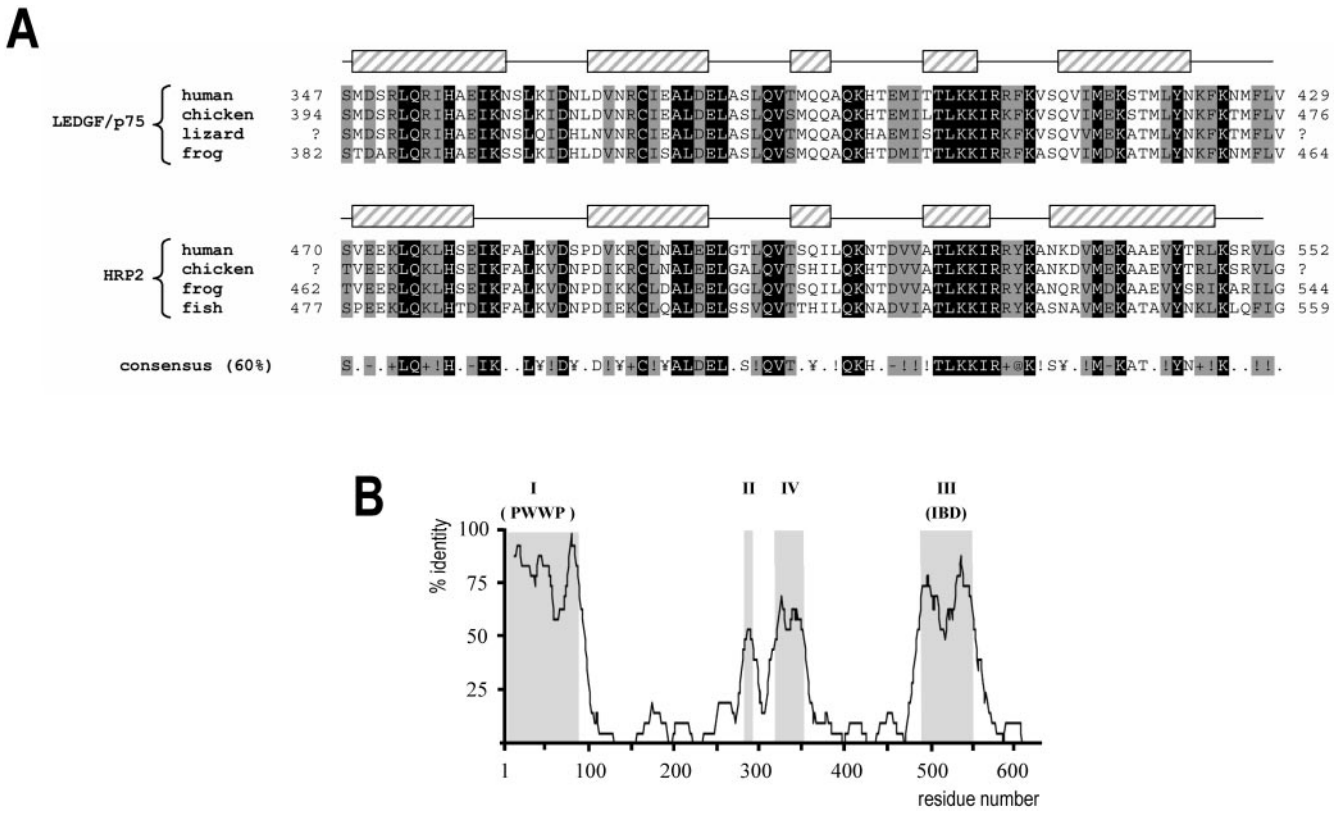

FIG. 4. **HRP2 contains a conserved IBD.** *A*, multiple sequence alignment of the conserved IBD regions of LEDGF/p75 and HRP2 orthologs. Amino acid coordinates were given when complete sequence information was available. Amino acid residues identical between all aligned sequences are highlighted in *black*. Positions conserved through substitution are shown in *gray*. The code for the consensus is: −, negatively charged; +, positively charged; ±, charged; !, hydrophobic; ¥, polar; @, aromatic. The taxonomic alignment is: chicken, *G. gallus*; lizard, *Anolis sagrei*; frog, *X. laevis*; fish, *D. rerio*. Sequences with the following GenBank™ accession numbers were used: AAC25167 (human LEDGF/p75), NP_116020 (human HRP2), and CF775831 (EST from *A. sagrei*). The coding regions of *G. gallus* and *X. laevis* LEDGF/p75 and *D. rerio* HRP2 were determined in this work. The sequence of the complete *X. laevis* HRP2 cDNA and a 3′-portion of the *G. gallus* HRP2 cDNA coding region were reconstructed from available ESTs. α-Helical regions for human LEDGF/p75 and HRP2 as predicted by PROFsec are shown as *boxes* above the alignments. *B*, sequence conservation between mammalian, amphibian, and fish orthologs of HRP2. Percentages of identical residues along the alignment of human, *X. laevis*, and *D. rerio* HRP2 proteins, calculated in windows of 20 residues. The complete alignment is shown in Supplementary Fig. S2.

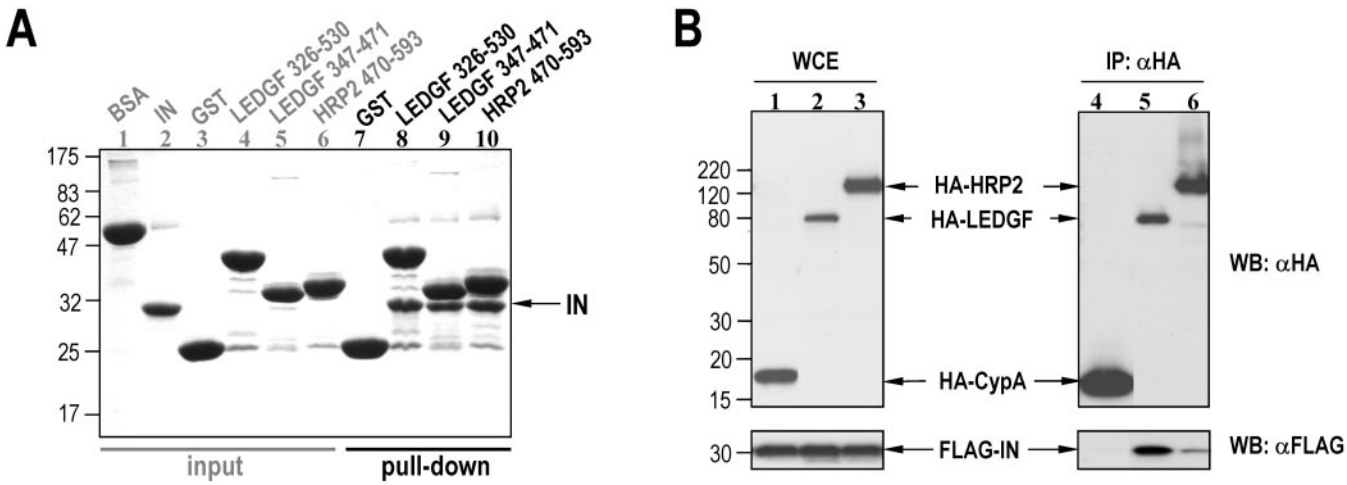

FIG. 5. **The HRP2 IBD can bind HIV-1 IN**. *A*, GST pull-down assay. *Lanes 1–6* contained input proteins. *Lanes* 7–10, pull-down with GST (*lane* 7), GST fused to LEDGF-(326–530) (*lane 8*) or LEDGF-(347–471) (*lane 9*) and GST fused to a fragment spanning residues 470–593 of human HRP2. The gel was stained with Coomassie R250. *B*, co-immunoprecipitation of FLAG-tagged HIV-1 IN with HA-tagged LEDGF/p75 and HRP2. *Lanes 1–3* contained whole cell extracts (WCE); *lanes 4–6*, proteins precipitated with anti-HA antibody (IP). Cells were transfected with FLAG-IN along with pCypA-HA (*lanes 1* and *4*), pBHA-P75 (*lanes* 2 and *5*), or pCPHA-HRP2 (*lanes 3* and *6*). HA-tagged proteins and FLAG-IN (*bottom panels*) were detected by Western blotting. Positions corresponding to migration of HA-HRP2, HA-LEDGF/p75, HA-CypA, and FLAG-IN are indicated.

activity in the absence of organic solvents and polyethylene glycol (*lanes 1–7*, Fig. 6*A*). The DNA substrate was a linearized plasmid containing HIV-1 U3 and U5 sequences at its termini (mini-HIV, see Ref. 47). Strand-transfer products represented a range of branched molecules resulting from inter- and intramolecular integration of the substrate DNA and were readily

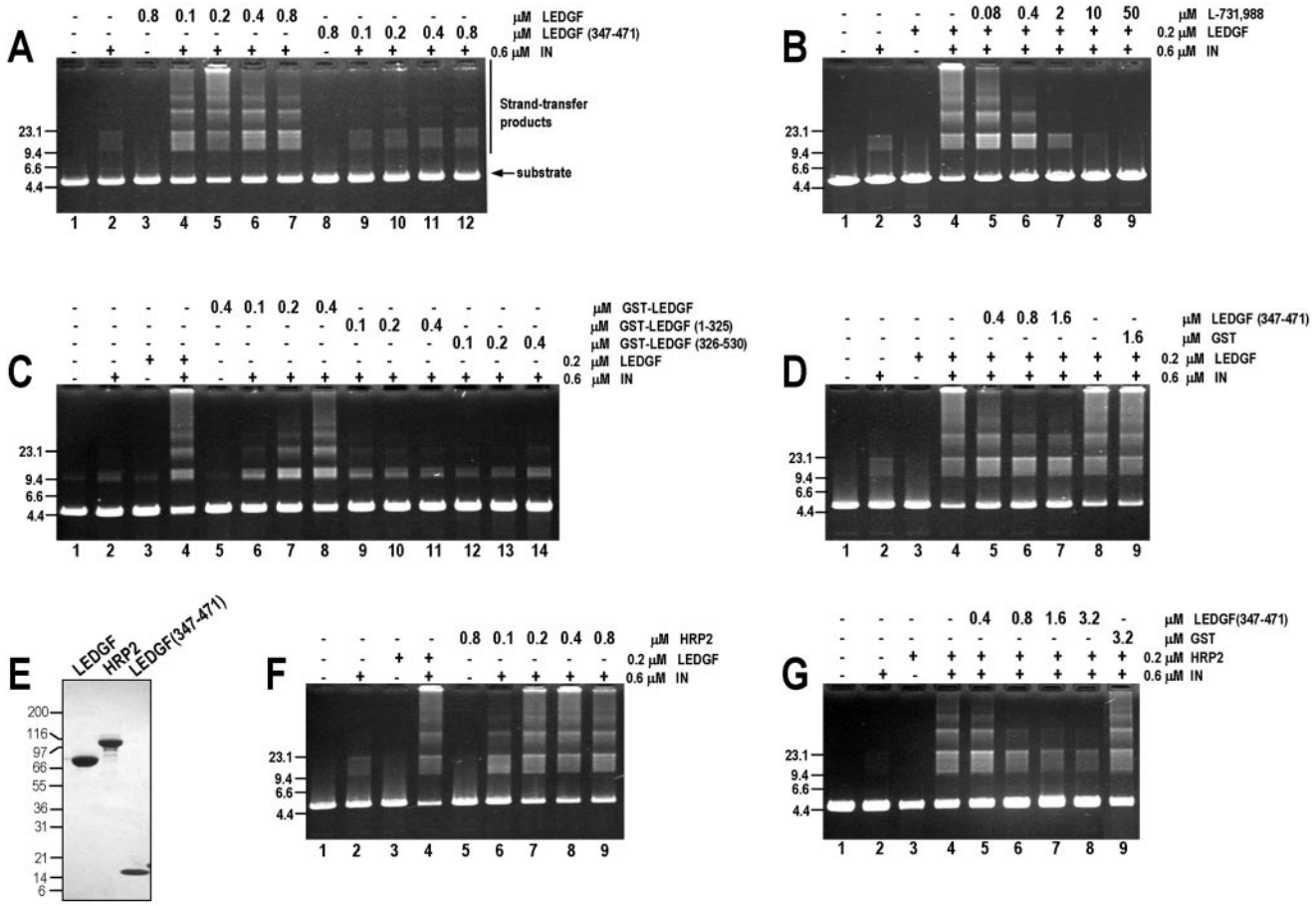

FIG. 6. **Stimulation of HIV-1 IN strand transfer activity by LEDGF/p75 and HRP2.** *A*, substrate DNA was preincubated with (+) 0.6 μM or without (−) IN for 7 min at room temperature. LEDGF/p75 (*lanes 3–7*) or LEDGF-(347–471) (*lanes 8–12*) was added at the indicated final concentrations. The reaction mixtures (20 μl) contained 150 ng mini-HIV DNA, 110 mM NaCl, 5 mM $MgCl_2$, 10 mM DTT, 2 μM $ZnCl_2$, 10 mM Hepes, pH 7.45. The DNA substrate was prepared as in Ref. 47. Following incubation for 90 min at 37 °C the reactions were stopped by addition of 25 mM EDTA and 0.5% SDS. Products were treated with 0.5 μg/μl proteinase K (Roche Applied Sciences) for 45 min at 50 °C, ethanol-precipitated, re-dissolved in Tris-EDTA and separated in 0.8% agarose. The gel was stained with ethidium bromide. Migration of the substrate DNA, strand transfer products, and molecular mass markers (23.1, 9.4, 6.6, and 4.4 kb) are indicated. *B*, inhibition of LEDGF/p75-dependent integration by diketo acid L-731,988. L-731,988, diluted in 50% dimethyl sulfoxide, was added to samples 5–9 at the indicated final concentrations. IN (*lanes* 2 and *4–9*) was preincubated with substrate DNA and drug for 7 min at room temperature. Next, LEDGF/p75 was added to samples in *lanes 3–9*, and the reactions proceeded for 90 min at 37 °C. The composition of the final reaction mixtures was as explained in Fig. 6*A*, however all reactions were adjusted to contain 1% dimethyl sulfoxide. Reaction products were treated and visualized as in Fig. 6*A*. *C*, IN (0.6 μM) was added to *lanes 2*, *4*, and *6–14*. LEDGF/p75 (0.2 μM) was present in *lane 4*. Full-length GST-LEDGF/p75 (*lanes 5–8*), GST-LEDGF-(1–325) (*lanes* 9-*11*) or GST-LEDGF-(326–530) (*lanes 12–14*) were added at the indicated concentrations. *D*, LEDGF-(347–471) (*lanes 5–7*) or GST (*lane 9*) were added together with full-length protein as competitors at the indicated concentrations. *E*, purified LEDGF/p75, HRP2, and LEDGF-(347–471) (~7.5 μg each) were separated by 4–20% SDS-PAGE. The gel was stained with Coomassie R250. Positions of molecular mass markers are indicated in kDa. *F*, reactions were set up as in Fig. 6*A*, and supplemented with HRP2 or LEDGF/p75 at the indicated concentrations. Following 90 min at 37 °C the reactions were stopped by addition of 700 mM NaCl, 20 mM EDTA and extracted twice with phenol/chloroform (1:1). The products precipitated with ethanol were dissolved in Tris-EDTA and separated in 0.8% agarose. *G*, LEDGF-(347–471) (*lanes 5–8*) or GST (*lane* 9) were added as competitors. Reaction products were treated and visualized as in Fig. 6*F*.

visualized in agarose gels by staining with ethidium bromide (Fig. 6*A*). In the absence of LEDGF/p75, only a marginal level of product formation was detected (compare *lanes 2* and *4–7* in Fig. 6*A*). Maximum stimulation was seen at 0.2 μM LEDGF/p75 (*lane* 5, Fig. 6*A*). LEDGF/p75-dependent stimulation of IN strand transfer activity was efficiently blocked by a specific IN inhibitor, diketo acid L-731,988 (6), at submicromolar concentrations (Fig. 6*B*).

Although LEDGF-(347–471) efficiently bound HIV-1 IN (*lane 13* in Fig. 3*B*), it failed to significantly stimulate strand transfer activity (compare *lanes 2*, *5*, and *10* in Fig. 6*A*). While full-length GST-LEDGF/p75 predictably stimulated IN (*lanes 6–8*, Fig. 6*C*), neither GST-LEDGF-(1–325) nor GST-LEDGF-(326–530) significantly effected the reaction (compare *lanes 2*, *8*, *11*, and *14* in Fig. 6*C*). Moreover, the isolated IBD (LEDGF-(347–471)) competitively counteracted the stimulatory effect of full-length LEDGF/p75 (*lanes 4–7* in Fig. 6*D*), confirming that LEDGF/p75 requires binding to IN to stimulate its activity. GST, used as a negative control, did not affect the level of LEDGF/p75-dependent strand transfer (*lanes* 8 and *9*, Fig. 6*D*). These data demonstrate that the IBD is essential but not sufficient to activate HIV-1 IN.

To test whether HRP2 can stimulate HIV-1 IN *in vitro*, we purified full-length HRP2 protein (Fig. 6*E*). As can be seen from Fig. 6*F*, HRP2 was proficient in activating HIV-1 IN-mediated strand transfer at a concentration range similar to LEDGF/p75. Akin to the LEDGF/p75 IBD fragment (LEDGF-(347–471)), IBD-containing HRP2-(470–593) (*lane 6*, Fig. 5*A*) did not increase HIV-1 IN strand transfer (data not shown). In addition, LEDGF-(347–471) competitively inhibited HRP2-dependent stimulation (*lanes 4–8* in Fig. 6*G*), arguing that LEDGF/p75 and HRP2 share a common binding site on HIV-1 IN.

DISCUSSION

Sequence alignments, *in silico* secondary structure prediction, and limited proteolysis collectively suggest that LEDGF/p75 contains a pair of small structural domains: an N-terminal PWWP domain (residues 1–90), the existence of which had been recognized on the basis of sequence homology, and a novel domain that mediates interaction with HIV-1 IN. Remarkably, these two domains encompass only about 35% of the protein sequence. Recombinant LEDGF/p75 displays high sensitivity to proteolysis suggesting that a large portion of the protein exists as flexible regions or loops. Of note, we did not detect a stable interaction between the PWWP and IBD domains in a GST pull-down assay (data not shown), suggesting that the domains are relatively independent in the full-length protein. We think that such flexibility might be related to the function of the protein *in vivo*, allowing the domains to associate with and link together components of various complexes. Two putative loop regions, with no regular secondary structure were suggested by *in silico* analysis of LEDGF/p75 (Fig. 1*B*). Interestingly, proteins containing extended loops are statistically associated with transcription regulatory functions (48). In addition to the PWWP and the IBD domains, an internal 20-residue fragment of LEDGF/p75 (residues 178–197) displayed significant sequence conservation (*region II*, Fig. 1*A*). This 20-amino acid fragment contains five Pro residues and is thus unlikely to adopt an independent secondary structure. High Pro and Arg content makes it similar to the AT hook motif of the HMGA proteins. Due to the recognized sequence conservation of region II, we speculate that it is important for LEDGF/p75 function. One likely possibility is that it represents a part of the DNA binding determinant of LEDGF/p75.

The LEDGF/p75 IBD is comprised of about 80 residues and is predicted to fold into four or five α-helices (Figs. 1*B* and 4). The minimal fragment that bound HIV-1 IN via GST pull-down spanned residues Ser^{347}–Val^{429}. This is in agreement with a previous report that LEDGF/p52 protein lacking residues 326–530 neither bound HIV-1 IN *in vitro* nor co-localized with it in live cells (18). Intriguingly, we identified a homologous sequence within another HDGF-related protein, HRP2, which likewise displayed affinity for HIV-1 IN. Thus, in addition to the N-terminal PWWP domains, LEDGF/p75 and HRP2 share conserved C-terminal domains, suggesting a close evolutionary and probable functional relationship between these proteins. Although we did not analyze susceptibility of HRP2 to proteases, analysis of its predicted amino acid sequence suggests that domain organization is similar to that of LEDGF/p75. Alignment of HRP2 orthologs from mammalian, amphibian, and fish sources showed a high degree of sequence conservation within the PWWP and IBD regions (Fig. 4*B*, see also Supplementary Fig. S2). Two additional fragments with significant interspecies homology (*regions II* and *IV*, Fig. 4*B*) were present in this protein. While HRP2 homology region II was clearly related to LEDGF/p75 region II, containing similarly spaced Pro and charged residues (Supplementary Figs. S1 and S2), region IV appears unique to HRP2. An extended α-helix involving residues Glu^{321}–Arg^{356} is predicted in this fragment. Thus, it is likely that HRP2 possesses an additional small structural domain. The sequences connecting the conserved regions in HRP2 contain multiple low complexity elements comprised of Pro, Ser, or Ser-Asp repeats, suggesting high flexibility (Supplementary Fig. S2). Low complexity sequences are common to eukaryotic proteins and are thought to be natively disordered (49). Such sequences are usually not conserved, in accordance with their putative roles as flexible hinges. A high prevalence of simple sequences in HRP2 explains the overall low degree of sequence conservation between orthologs compared with that of LEDGF/p75 (see Supplementary Table I and Figs. S1 and S2). *In silico* analysis of amino acid sequences of other HRPs suggest that although they do not possess IBD-like domains, α-helical elements are located within C-terminal regions of HDGF and HRP1 (data not shown), suggesting the presence of a second functional domain within these proteins as well.

Like HDGF, all HRPs seem to have mitogenic activity in cell culture (21, 25, 30). It is presently unclear whether the growth factor activity of such proteins that lack classical secretory signals is related to their functions *in vivo* (20). The original observation that LEDGF/p75 co-purified from HeLa nuclear extracts together with the transcription co-activator PC4 provided a clue that the protein might be involved in transcriptional regulation (50). More recently, LEDGF/p75 was reported to bind to heat shock and stress-related elements within promoters regions of the AOP2, Hsp27, and αB-crystallin genes and trans-activate their expression (28, 29). Although an earlier study isolated LEDGF/p75 from a lens epithelial cDNA library, expression of the protein is clearly not limited to lens. In contrast to the protein's name, cDNA clones encoding LEDGF/p75 have been isolated from a wide range of primary and transformed mouse and human tissues at all stages of development (refer to EST collections associated with the UniGene entries from Supplementary Table I). Sequences derived from 215 cDNA clones suggesting several alternative LEDGF splice variants exist in the AceView data base (for up to date information consult www.ncbi.nlm.nih.gov/IEB/Research/Acembly/). While the most abundant splice form, supported by 170 cDNA clones, encodes for LEDGF/p75, only 12 cDNAs are derived from p52 mRNA. Although a detailed expression analysis of individual splice forms will require a specialized study, it would appear that LEDGF/p75 is the dominant protein product of the *PSIP1* gene in most tissues.

According to the large numbers of human and mouse ESTs corresponding to LEDGF/p75 and HRP2, these proteins are ubiquitously expressed at relatively high levels (see Supplementary Table I). Although the HRP2 IBD displayed an apparent high affinity for HIV-1 IN by GST pull-down (Fig. 5*A*), results of co-immunoprecipitation experiments suggested that LEDGF/p75 was a more potent IN interactor than was full-length HRP2 in human cells (Fig. 5*B*). This was not entirely unexpected, as depletion of endogenous LEDGF/p75 alone by siRNA efficiently disrupted the nuclear and chromosomal accumulation of HIV-1 and FIV IN in cells (18, 19). However, LEDGF/p75 and HRP2 proteins stimulated HIV-1 IN to a comparable degree *in vitro* (Fig. 6*F*). Based on this result we speculate that binding of IN to HIV-1 cDNA termini might stabilize the HRP2-IN interaction. HRP2 could potentially explain the failure of persistent siRNA-mediated knockdowns of LEDGF/p75 to reduce viral replication (19). It would also be interesting to determine if LEDGF/p75 and/or HRP2 modulate the enzymatic activity of FIV and other retro/lenti-viral INs (19).

It was demonstrated that HIV-1 displays a significant bias toward integration into active genes (51, 52). Somewhat similar, but not identical integration specificity was observed for murine leukemia virus, which prefers to integrate within transcription start regions in the human genome (52). On a practical level, specificity for integration within or near active genes poses a problem in developing retroviral vector-based gene therapies (53). Distant relatives of retroviruses, yeast retrotransposons present the best studied paradigm of targeted integration in eukaryotes (reviewed in Ref. 54). At least in the case of the Ty5 retrotransposon, a specific interaction between Ty5 IN and the chromosomal protein Sir4p determines the specificity of retrotransposition into silent chromatin (55, 56).

Integration of another yeast retrotransposon, Ty3, which has a preference for RNA polymerase III transcription start sites, is controlled by a TFIIIB transcription factor complex, although the interacting determinant on the retrotransposon side is not known (57). Putative chromodomains were identified in the C-terminal regions of INs from many LTR retrotransposons, such as fungal Cft1 and Skippy, and were hypothesized to mediate the targeting of their integration (58). In this context, a model involving a chromatin binding protein as a targeting factor for retroviral integration seems quite plausible. LEDGF/p75, a chromosomal protein and a putative regulator of transcription that binds lentiviral INs in live cells, represents such a candidate factor (16–19). Identification of LEDGF/p75 as a component of HIV-1 PICs encourages further research, as it remains to be seen whether LEDGF/p75 and/or its close relative HRP2 play role(s) in PIC formation or targeting during retroviral infection (19).

Acknowledgments—We thank G. Maertens for constructing the pBHA-P75 plasmid and for critical reading of the manuscript, T. Ellenberger and O. Tsodikov (Harvard Medical School) for advice on LEDGF/p75 domain mapping and use of their fermentation facility, and J. Obenchain (Merck Research Laboratories) for L-731,988. We are grateful to J. Walter and C. Cvetic (Harvard Medical School) for providing us with *X. laevis* tissues, T. Look and T. Liu (Dana-Farber Cancer Institute) for their generous gift of *D. rerio* RNA samples, J.-M. Buerstedde (Heinrich-Pette Institute for Experimental Virology and Immunology) for the *G. gallus* DT40 cell line and H. Göttlinger (Dana-Farber Cancer Institute) for the HA-CypA expression construct. We thank J. Lee and E. Gillespie from the Molecular Biology Core facility at the Dana-Farber Cancer Institute for HPLC, MALDI MS and Edman analyses of our samples and J. Gibson-Brown (Washington University) for sharing the supporting raw data file for the lizard CF775831 EST.

REFERENCES

1. Craigie, R. (2001) *J. Biol. Chem.* **276,** 23213–23216
2. Asante-Appiah, E., and Skalka, A. M. (1997) *Antiviral Res.* **36,** 139–156
3. Brown, P. O. (1997) in *Retroviruses* (Coffin, J. M., Huges, S. H., and Varmus, H. E., eds) pp. 161–203, Cold Spring Harbor Laboratory, Cold Spring Harbor
4. Nakajima, N., Lu, R., and Engelman, A. (2001) *J. Virol.* **75,** 7944–7955
5. LaFemina, R. L., Schneider, C. L., Robbins, H. L., Callahan, P. L., LeGrow, K., Roth, E., Schleif, W. A., and Emini, E. A. (1992) *J. Virol.* **66,** 7414–7419
6. Hazuda, D. J., Felock, P., Witmer, M., Wolfe, A., Stillmock, K., Grobler, J. A., Espeseth, A., Gabryelski, L., Schleif, W., Blau, C., and Miller, M. D. (2000) *Science* **287,** 646–650
7. Rice, P. A., and Baker, T. A. (2001) *Nat. Struct. Biol.* **8,** 302–307
8. Engelman, A. (1999) *Adv. Virus Res.* **52,** 411–426
9. Wu, X., Liu, H., Xiao, H., Conway, J. A., Hehl, E., Kalpana, G. V., Prasad, V., and Kappes, J. C. (1999) *J. Virol.* **73,** 2126–2135
10. Engelman, A. (2003) *Curr. Top. Microbiol. Immunol.* **281,** 209–238
11. Kalpana, G. V., Marmon, S., Wang, W., Crabtree, G. R., and Goff, S. P. (1994) *Science* **266,** 2002–2006
12. Willetts, K. E., Rey, F., Agostini, I., Navarro, J. M., Baudat, Y., Vigne, R., and Sire, J. (1999) *J. Virol.* **73,** 1682–1688
13. Parissi, V., Calmels, C., De Soultrait, V. R., Caumont, A., Fournier, M., Chaignepain, S., and Litvak, S. (2001) *J. Virol.* **75,** 11344–11353
14. Mulder, L. C., Chakrabarti, L. A., and Muesing, M. A. (2002) *J. Biol. Chem.* **277,** 27489–27493
15. Violot, S., Hong, S. S., Rakotobe, D., Petit, C., Gay, B., Moreau, K., Billaud, G., Priet, S., Sire, J., Schwartz, O., Mouscadet, J. F., and Boulanger, P. (2003) *J. Virol.* **77,** 12507–12522
16. Cherepanov, P., Maertens, G., Proost, P., Devreese, B., Van Beeumen, J., Engelborghs, Y., De Clercq, E., and Debyser, Z. (2003) *J. Biol. Chem.* **278,** 372–381
17. Turlure, F., Devroe, E., Silver, P. A., and Engelman, A. (2004) *Front. Biosci.* **9,** 3187–3208
18. Maertens, G., Cherepanov, P., Pluymers, W., Busschots, K., De Clercq, E., Debyser, Z., and Engelborghs, Y. (2003) *J. Biol. Chem.* **278,** 33528–33539
19. Llano, M., Vanegas, M., Fregoso, O., Saenz, D., Chung, S., Peretz, M., and Poeschla, E. M. (2004) *J. Virol.* **78,** 9524–9537
20. Izumoto, Y., Kuroda, T., Harada, H., Kishimoto, T., and Nakamura, H. (1997) *Biochem. Biophys. Res. Commun.* **238,** 26–32
21. Dietz, F., Franken, S., Yoshida, K., Nakamura, H., Kappler, J., and Gieselmann, V. (2002) *Biochem. J.* **366,** 491–500
22. Stec, I., Nagl, S. B., van Ommen, G. J., and den Dunnen, J. T. (2000) *FEBS Lett.* **473,** 1–5
23. Maurer-Stroh, S., Dickens, N. J., Hughes-Davies, L., Kouzarides, T., Eisenhaber, F., and Ponting, C. P. (2003) *Trends Biochem. Sci.* **28,** 69–74
24. Ge, Y. Z., Pu, M. T., Gowher, H., Wu, H. P., Ding, J. P., Jeltsch, A., and Xu, G. L. (2004) *J. Biol. Chem.* **279,** 25447–25454
25. Kishima, Y., Yamamoto, H., Izumoto, Y., Yoshida, K., Enomoto, H., Yamamoto, M., Kuroda, T., Ito, H., Yoshizaki, K., and Nakamura, H. (2002) *J. Biol. Chem.* **277,** 10315–10322
26. Ikegame, K., Yamamoto, M., Kishima, Y., Enomoto, H., Yoshida, K., Suemura, M., Kishimoto, T., and Nakamura, H. (1999) *Biochem. Biophys. Res. Commun.* **266,** 81–87
27. Nishizawa, Y., Usukura, J., Singh, D. P., Chylack, L. T., Jr., Shinohara, T., and Kimura, A. (2001) *Cell Tissue Res.* **305,** 107–114
28. Fatma, N., Singh, D. P., Shinohara, T., and Chylack, L. T., Jr. (2001) *J. Biol. Chem.* **276,** 48899–48907
29. Singh, D. P., Fatma, N., Kimura, A., Chylack, L. T., Jr., and Shinohara, T. (2001) *Biochem. Biophys. Res. Commun.* **283,** 943–955
30. Wu, X., Daniels, T., Molinaro, C., Lilly, M. B., and Casiano, C. A. (2002) *Cell Death Differ.* **9,** 915–925
31. Rost, B., and Liu, J. (2003) *Nucleic Acids Res.* **31,** 3300–3304
32. Liu, J., and Rost, B. (2003) *Nucleic Acids Res.* **31,** 3833–3835
33. Cowan, R., and Whittaker, R. G. (1990) *Pept. Res.* **3,** 75–80
34. Henikoff, S., and Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. U. S. A.* **89,** 10915–10919
35. Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992) *Science* **256,** 1443–1445
36. Zdobnov, E. M., and Apweiler, R. (2001) *Bioinformatics* **17,** 847–848
37. Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. (2004) *Nucleic Acids Res.* **32,** D142–144
38. Bram, R. J., Hung, D. T., Martin, P. K., Schreiber, S. L., and Crabtree, G. R. (1993) *Mol. Cell. Biol.* **13,** 4760–4769
39. Limon, A., Devroe, E., Lu, R., Ghory, H. Z., Silver, P. A., and Engelman, A. (2002) *J. Virol.* **76,** 10598–10607
40. Cherepanov, P., Este, J. A., Rando, R. F., Ojwang, J. O., Reekmans, G., Steinfeld, R., David, G., De Clercq, E., and Debyser, Z. (1997) *Mol. Pharmacol.* **52,** 771–780
41. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) *Anal. Chem.* **68,** 850–858
42. Cherepanov, P., Pluymers, W., Claeys, A., Proost, P., De Clercq, E., and Debyser, Z. (2000) *FASEB J.* **14,** 1389–1399
43. Maertens, G., Cherepanov, P., Debyser, Z., Engelborghs, Y., and Engelman, A. (2004) *J. Biol. Chem.*, **279,** 33421–33429
44. Slater, L. M., Allen, M. D., and Bycroft, M. (2003) *J. Mol. Biol.* **330,** 571–576
45. Hubbard, S. J. (1998) *Biochim. Biophys. Acta* **1382,** 191–206
46. Booth, V., Koth, C. M., Edwards, A. M., and Arrowsmith, C. H. (2000) *J. Biol. Chem.* **275,** 31266–31268
47. Cherepanov, P., Surratt, D., Toelen, J., Pluymers, W., Griffith, J., De Clercq, E., and Debyser, Z. (1999) *Nucleic Acids Res.* **27,** 2202–2210
48. Liu, J., Tan, H., and Rost, B. (2002) *J. Mol. Biol.* **322,** 53–64
49. Huntley, M. A., and Golding, G. B. (2002) *Proteins* **48,** 134–140
50. Ge, H., Si, Y., and Roeder, R. G. (1998) *EMBO J.* **17,** 6723–6729
51. Schroder, A. R., Shinn, P., Chen, H., Berry, C., Ecker, J. R., and Bushman, F. (2002) *Cell* **110,** 521–529
52. Wu, X., Li, Y., Crise, B., and Burgess, S. M. (2003) *Science* **300,** 1749–1751
53. Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M. P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C. S., Pawliuk, R., Morillon, E., Sorensen, R., Forster, A., Fraser, P., Cohen, J. I., de Saint Basile, G., Alexander, I., Wintergerst, U., Frebourg, T., Aurias, A., Stoppa-Lyonnet, D., Romana, S., Radford-Weiss, I., Gross, F., Valensi, F., Delabesse, E., Macintyre, E., Sigaux, F., Soulier, J., Leiva, L. E., Wissler, M., Prinz, C., Rabbitts, T. H., Le Deist, F., Fischer, A., and Cavazzana-Calvo, M. (2003) *Science* **302,** 415–419
54. Bushman, F. D. (2003) *Cell* **115,** 135–138
55. Xie, W., Gai, X., Zhu, Y., Zappulla, D. C., Sternglanz, R., and Voytas, D. F. (2001) *Mol. Cell. Biol.* **21,** 6606–6614
56. Zhu, Y., Dai, J., Fuerst, P. G., Voytas, D. F., Xie, W., Gai, X., Zappulla, D. C., and Sternglanz, R. (2003) *Proc. Natl. Acad. Sci. U. S. A.* **100,** 5891–5895
57. Yieh, L., Hatzis, H., Kassavetis, G., and Sandmeyer, S. B. (2002) *J. Biol. Chem.* **277,** 25920–25928
58. Malik, H. S., and Eickbush, T. H. (1999) *J. Virol.* **73,** 5186–5190
59. Gouet, P., Courcelle, E., Stuart, D. I., and Metoz, F. (1999) *Bioinformatics* **15,** 305–308

Identification of an Evolutionarily Conserved Domain in Human Lens Epithelium-derived Growth Factor/Transcriptional Co-activator p75 (LEDGF/p75) That Binds HIV-1 Integrase

Peter Cherepanov, Eric Devroe, Pamela A. Silver and Alan Engelman

J. Biol. Chem. 2004, 279:48883-48892.
doi: 10.1074/jbc.M406307200 originally published online September 14, 2004