Supplementary Materials for

Cryo-EM reveals that a supramolecular assembly mediates lentiviral DNA integration

Allison Ballandras-Colas, Daniel P. Maskell, Erik Serrao, Julia Locke, Paolo Swuec, Stefán R. Jónsson, Abhay Kotecha, Nicola J. Cook, Valerie E. Pye, Ian A. Taylor, Valgerdur Andrésdóttir, Alan N. Engelman, Alessandro Costa, Peter Cherepanov

correspondence to: peter.cherepanov@crick.ac.uk; alessandro.costa@crick.ac.uk

This PDF file includes:

Other Supplementary Materials for this manuscript includes the following:

Materials and Methods

Protein expression and purification

To obtain pCPH6P-MVV-IN, used for bacterial expression of MVV IN with an N-terminal cleavable hexahistidine (His_6) tag, a DNA fragment encoding full-length IN from MVV clone KV1772 (*25*) was inserted between *Xma*I and *Xho*I sites of pCPH6P (*10*). The protein was produced in EndoI-null PC2 *Escherichia coli* cells (BL21(DE3), endA::Tet^R, $T1^R$, pLysS) (*10*) grown in LB medium in shaker flasks at 25 °C by induction with 0.01% w/v isopropyl β-D-1-thiogalactopyranoside. Cells were lysed by sonication in Buffer A containing 1 M NaCl, 0.5% 3-[(3-cholamidopropyl) dimethylammonio]-1-propane sulfonate (CHAPS) and 50 mM BisTris propane (BTP) - HCl pH 7.0, supplemented with 0.2 mM phenylmethylsulfonyl fluoride (PMSF). The supernatant, precleared by centrifugation, was incubated with NiNTA agarose (Qiagen) at 4 °C in the presence of 20 mM imidazole. The resin was washed using buffer A containing 20 mM imidazole and recombinant protein was eluted with 200 mM imidazole in buffer A. IN-containing fractions were supplemented with 5 mM dithiothreitol (DTT) and incubated with human rhinovirus 3C protease (1:50 w/w ratio) overnight at 7 °C to remove the His_6 tag. The protein, diluted with 8 volumes of 50 mM BTP-HCl pH 7.0, 0.5% CHAPS, was loaded onto a 5-ml HiTrap Heparin column (GE Healthcare) and eluted with a linear 0.1 – 1 M NaCl gradient over 7 column volumes. MVV IN was supplemented with 5 mM DTT and 370 mM NaCl (to a final NaCl concentration of 1M), concentrated to 20 mg/ml using a 10-kDa cutoff VivaSpin unit (Generon), and stored on ice. Human LEDGF was expressed in bacteria and purified as described (*10*).

Strand transfer assays

Activity assays and EM were done using MVV intasomes assembled with double-stranded oligonucleotide substrate mimicking the terminal 29 bp of the processed MVV U5 vDNA end prepared by annealing synthetic oligonucleotides EV272 (5'-CCGTGCA ACACCGGAGCGGATCTCGCA) and EV273 (5'- GCTGCGAGATCCGCTCCGGTGT TGCACGG). For integration site sequencing, intasomes were assembled with extended substrates mimicking 35 bp of the processed MVV U3 and/or U5 vDNA ends (EV276 5'-TTCCAGGGTAGGCATTTGTTCTCTGTCCTGACA with EV277 5'-ACTGTCAGGA CAGAGAACAAATGCCTACCCTGGAA, or EV245 5'-CTAATTCCGTGCAACACC GGAGCGGATCTCGCA with EV246 5'-GCTGCGAGATCCGCTCCGGTGTTGCACG GAATTAG, respectively). Supercoiled pGEM9 target DNA (300 ng) was incubated with 0 - 6 μM IN, 0 - 3 μM LEDGF, and 0.5 μM oligonucleotide vDNA substrate in 75 mM NaCl, 40 mM KCl, 5 mM $MgSO_4$, 2.5 mM DTT, 5 μM $ZnCl_2$, and 50 mM BisTris-HCl pH 6.5 in the final volume of 40 μl. The reactions were allowed to proceed for 1-2 h at 37 °C and stopped by addition of 0.25% SDS, 12.5 mM EDTA and 50 mM Tris-HCl pH 7.5. DNA products, deproteinized by digestion with proteinase K and deionized by ethanol precipitation, were separated in 1.5% (w/v) agarose gels and visualized by staining with ethidium bromide. Products of full-site integration into pGEM9 target were treated with phage Phi29 DNA polymerase, ligated to a kanamycin resistance cassette, and processed for sequencing as previously described (*26*).

Crystallization and structure determination of MVV IN CTD and CCD-CTD fragments

Constructs spanning residues 219-276 and 58-276 of MVV IN and fused to His_6-SUMO tag were expressed in *E. coli* BL21(DE3)-CodonPlus cells (New England BioLabs). His_6-tagged proteins were absorbed onto NiNTA agarose from precleared cell lysates in respective core buffers (0.15 M NaCl, 50 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid-NaOH pH 7.4 [CTD] or 500 mM NaCl, 20 mM Tris-HCl pH 7.5 [CCD-CTD]), supplemented with 20 mM imidazole and 0.2 mM PMSF. Following extensive washing in the presence of 20 mM imidazole, the proteins were eluted with 200 mM imidazole in respective core buffers and digested with SUMO protease. Tag-free MVV IN CTD was further purified by cation exchange chromatography. The proteins, polished by size exclusion chromatography in respective core buffers, were supplemented with 2 mM DTT and concentrated by ultrafiltration to 30 mg/ml (CTD) or 3 mg/ml (CCD-CTD). Crystals were grown by vapor diffusion at 18 °C in hanging drops by mixing 1 µl of protein with 1 µl reservoir solution containing 20% PEG-3350, 0.2 M potassium sodium tartrate, 0.1 M BTP, pH 7.5 (CTD) or 15-21% PEG-400, 15-100 mM calcium acetate, 0.1 M 2-(N-morpholino)ethanesulfonic acid-NaOH pH 6.0 (CCD-CTD). Prior to snap-freezing in liquid nitrogen, the CTD crystals were cryo-protected by supplementing mother liquor with 30% PEG-400 and CCD-CTD crystals were soaked in 30% PEG-400, 75 mM calcium acetate, 0.1 M 2-(N-morpholino)ethanesulfonic acid-NaOH pH 6.0.

X-ray diffraction data were collected on I04 and I04-1 beamlines of the Diamond Light Source (Oxford, UK) at 100 K and processed using XDS (*27*) and AIMLESS (*28*) as implemented in Xia2 (*29*). The structure of CTD was solved by molecular replacement in Phaser (*30*) using a search model derived from a crystal structure of HIV-1 IN CCD-CTD construct (*7*). Crystals of the CCD-CTD diffracted anisotropically, with resolution limits of 2.5, 2.5 and 3.5 Å along *a*, *b* and *c* axes, respectively. Therefore, the data were elliptically truncated and corrected using the STARANISO server (http://staraniso.globalphasing.org). The structure was solved by molecular replacement in Phaser using search models derived from the refined CTD structure and the CCD domain from a previously reported structure of the MVV IN NTD-CCD construct (*12*). The structures were built and refined using Coot (*31*) and Phenix (*32*) (table S1). The final models have good geometry, as assessed by MolProbity (*33*), and fit well to the electron density. The CTD structure has a dimer in the asymmetric unit, and CCD-CTD forms a dimer via crystallographic 2-fold symmetry. The coordinates and structure factors were deposited with the PDB under accession codes 5LLJ and 5T3A.

Nucleoprotein complex assembly and purification

For intasome assembly, a mixture containing 8 µM IN, 8 µM LEDGF, and 5 µM vDNA in 80 mM NaCl, 40 mM potassium acetate, 3 mM $CaCl_2$ 10 µM $ZnCl_2$, 1 mM DTT and 25 mM BisTris-HCl pH 6.0 was incubated at 37 °C for 10 min. Unlike Mg^{2+}, Ca^{2+} does not support IN enzymatic activities but is known to stabilize IN-DNA interactions (*5, 34*). The protein-DNA mixture was then supplemented with 50 mM BisTris-HCl pH 6.5 and 470 mM NaCl and placed on ice for 10 min. The presence of LEDGF and the initial incubation at 32-45 °C was essential for nucleoprotein complex formation (data not shown). The intasome was purified by two-step chromatography. First, the assembly mixtures were separated on a Superdex-200 16/60 GL column in 590 mM NaCl, 3 mM $CaCl_2$, and 25 mM

BisTris-HCl pH 6.5. The intasome eluted close to the void volume of Superdex-200, well-separated from free proteins and vDNA. The fractions were supplemented with NaCl to the final concentration of 1 M and concentrated using a 10-kDa cutoff VivaSpin device to the final volume of 200 µl and injected into a Superose-6 10/300 GL column, which was equilibrated with 1 M NaCl, 3 mM $CaCl_2$, and 25 mM BisTris-HCl pH 6.5. The intasome, eluting at ~11 ml (fig. S2A), was used immediately in activity assays and/or EM. The MVV strand transfer complex was assembled and purified in a similar way, using a DNA construct corresponding to the product of full-site integration of 23-bp MVV vDNA mimic into a palindromic 52-bp target DNA. The product DNA was made by annealing oligonucleotides EV226 (5'-GCTGCGAGATCCGCTCCGGTGTT), EV280 (5'- AACA CCGGAGCGGATCTCGCAGTCGACCACCCTAATCAAGTTTTTTGGGG; vDNA portions of the sequences are underlined) and EV281 (5'- CCCCAAAAAACTT GATTAGGGTG).

SEC-MALLS

Size exclusion chromatography coupled to multi-angle laser light scattering (SEC-MALLS) was carried out using a Jasco chromatography system equipped with a DAWN-HELEOS II laser photometer and an OPTILAB-TrEX differential refractometer. MVV IN (0.2 - 4 mg/mL in 1 M NaCl, 7.5 mM CHAPS, 2 mM DTT, 25 mM BTP-HCl pH 7.0) or intasome (~0.3 mg/mL in 1 M NaCl, 3 mM NaCl, 25 mM BisTris-HCl pH 6.5) were separated on a Superose 6 10/300 GL column (GE Healthcare) equilibrated in 1 M NaCl, 3 mM $CaCl_2$, 25 mM BisTris-HCl pH 6.5 at 0.3 ml/min, at 20 °C. The weight-averaged molecular masses of species contained in chromatographic peaks were calculated using the data from both detectors in ASTRA-6.1 software (Wyatt Technology).

MVV culture and integration site sequencing

Sheep choroid plexus cells were established from healthy animals by explants of choroid plexus as described previously (*35*). The cells were grown at 37 °C in Dulbecco's modified Eagles medium supplemented with 200 U/ml penicillin, 100 U/ml streptomycin, 2 mM glutamine and either 10% lamb serum (growth medium) or 1% lamb serum (maintenance medium). Cells were infected with pathogenic MVV molecular clone KV1772 (*25*) at a multiplicity of infection of 5, and DNA was extracted from infected cells 40 h post-infection prior to the onset of the cytopathic effect using a Gentra Puregene kit (Qiagen). MVV U5-chromsomal DNA junctions were recovered using linker-mediated PCR and sequenced on an Illumina platform as described (*36*). The first round and the nested MVV U5 primers were 5'-CTAATTCCGTGCAACACCG and 5'-AATGATACGGCGAC CACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNCAA CACCGGAGCGGATC, respectively (the underlined stretch of nucleotides represents the Illumina index). Viral long terminal repeat and linker reads were cropped to isolate chromosomal sequences and aligned to the oviAri3 version of sheep genome with BWA-MEM (*37*). The resulting alignments, processed with Samtools (*38*), were parsed for conversion to BED format using custom python scripts. Target genomic sequences were extracted with BEDTools (*39*), and sequence logos were generated using WebLogo (*40*). To obtain *in vitro* integration sites, genomic DNA isolated from non-infected sheep cells was incubated with purified recombinant MVV intasomes. Reaction products were repaired with Phi29 DNA polymerase and processed for linker-mediated PCR as for the

infection sample. Integration site sequencing data were deposited with the NCBI GEO repository (accession GSE87786).

Negative stain grid preparation, data collection, and processing

A 4-6 nm thick carbon layer was evaporated onto freshly cleaved mica sheets (EMS) using a Q150TE carbon coater (Quorum Technologies). Carbon was incubated overnight at 50 °C before floating on 400-mesh copper grids (Agar Scientific). Prior to application of the sample, the carbon-coated grids were glow-discharged for 30 s at 45 mA using a K100X glow discharger (EMS). A 4-µl droplet of the freshly purified sample solution at 50 ng/µl (concentration of the intasome was estimated based on A_{260}) was applied onto the glow-discharged grid and incubated for 1 min. Approximately 2 µl of sample solution was absorbed by gentle side blotting and the grid was immediately stained with four subsequent applications onto 4x 50 µl drops of a 2% (w/v) uranyl formate solution with gentle stirring for 1 min in total. After staining, the grids were blotted dry and stored at room temperature for imaging on an Tecnai G2 Spirit LaB6 transmission electron microscope (FEI) operating at 120 keV. Images were collected with a 2k×2k Ultrascan-1000 camera (Gatan) at a nominal magnification of 30,000× (corresponding to a pixel size of 3.45 Å at the specimen level), with an average electron dose of 35 e/Å^2 and defocus values of 0.5-2.0 µm. A representative negatively stained micrograph is shown in fig. S4A.

Particles, picked semi-automatically in EMAN2 (*41*), were subjected to reference-free 2D classification in Relion-1.4 (*42*) (fig. S4B). The classification highlighted subtle heterogeneity in the MVV intasome samples, with about 10% of the particles belonging to the elongated classes (fig. S4B, right). The effect was subsequently explained by the internal symmetry of the intasome, which allows a flanking tetramer to nucleate formation of another core intasome in the presence of excess vDNA ends. The elongated particles contained more than one pair of vDNA ends and, therefore, were not biologically relevant. These classes were not detected in the samples of the strand transfer complex. The 2D classes corresponding to the major type of the intasome complex were used to generate an initial 3D model using the *e2inimodel.py* program in the EMAN2 package. The initial volume was used as reference model for 3D refinements in Relion-1.4 and refined to 20 Å resolution using a dataset of ~25,000 particles (fig. S4C).

Cryo-grid preparation and data collection

A 4-µl drop of freshly prepared intasome or the strand transfer complex at a concentration of 0.3 mg/ml (estimated based on A_{260}) in 1 M NaCl, 3 mM $CaCl_2$ and 25 mM BisTris-HCl pH 6.5 was applied onto glow-discharged lacey carbon grids coated with ultrathin carbon (product 01824, Ted Pella). The grids were incubated for 30 s under 100% humidity in a Vitrobot Mark IV (FEI) at 20 °C. To lower salt concentration before plunge-freezing, the grids were blotted for 0.5 s, immediately hydrated with a 4-µl drop of 200 mM NaCl, 3 mM $CaCl_2$, and 25 mM BisTris-HCl pH 6.5 and blotted again for 2.5 s, followed by plunging into liquid ethane. The grids were screened for sample integrity and ice quality on a Tecnai T12 microscope.

An initial preliminary intasome dataset was acquired at the Institut de Biologie Structural (Grenoble, France) on a Tecnai G2 Polara (FEI) microscope operating at 300 kV

and equipped with a K2 Summit direct electron detector (Gatan). Movies containing 50 frames were recorded using the Digital Micrograph image acquisition software (Gatan) at a magnified physical pixel size of 1.21 Å over 50 frames. A total dose of 50 e/Å^2 distributed over 50 frames was used, and frames were recorded in single-electron-counting mode, with a nominal defocus range of ~3.5-5.0 μm. The initial dataset yielded 36,000 particles that allowed refinement of the structure to ~14 Å resolution revealing the density for vDNA (data not shown). The datasets used to refine the intasome and strand transfer complex structures were acquired at the Division of Structural Biology (STRUBI) in Oxford, using a Tecnai G2 Polara (FEI) microscope operating at 300 kV with a K2 Summit direct electron detector in single-electron-counting mode (Gatan) and in-built Quantum energy filter in zero-loss mode (Gatan). The movies were acquired manually using SerialEM (*43*) at a magnification of 37,037x, resulting in a pixel size of 1.35 Å at the specimen level. The intasome and the strand transfer complex datasets contained a total of 1,651 movies (with a total dose of 50 e/Å^2 distributed over 40 frames) and 722 movies (40 e/Å^2 over 30 frames), respectively.

Image processing and structure refinement

The movies were corrected for drift and beam-induced motion using MotionCorr-2.0 with implemented dose weighting (*44*), skipping the first two frames. Contrast transfer function was estimated using CTFFIND4 (*45*) from non-weighted aligned frame sums. A small subset of particles using semi-automatic picking with e2boxer.py in EMAN2 (*41*) was used to generate an initial subset of six reference-free 2D classes, which, after low-pass filtering to 20 Å, were used for automated particle picking in Relion-1.4 (*42*). All particles were subjected to two rounds of reference-free classification in Relion-1.4 resulting in two hundred 2D classes. Particles from classes containing poor or no structural features and aggregated complexes were eliminated after each round to obtain 253,785- and 37,021-particle datasets for the intasome and strand transfer complex, respectively.

Refinement of the intasome dataset in Relion-1.4 without imposing symmetry, followed by B-factor sharpening, resulted in a 5.8-Å resolution map with a clear 2-fold symmetry (fig. S6, top). The particles were subjected to classification into three 3D classes in 50 iterations with the regularization parameter T set to 4, without imposing symmetry. Refinement of the best resultant subset (class 1, which contained 37.1% of the particles) yielded reconstructions to 6.20 and 5.19 Å resolution, with and without two-fold symmetry, respectively. The final reconstruction was obtained using a solvent mask with a 10-pixel soft edge covering the whole complex during refinement. The map was sharpened by applying a B factor of -330 to obtain the final resolution of 4.94 Å (fig. S6, bottom), based on the gold-standard Fourier shell correlation of 0.143 criterion (*46*) (fig. S7A). The local resolution (fig. S7C) was estimated with ResMap (*47*). Euler angle distribution of the particles used in the final refinement and re-projections of the refined cryo-EM map and matched 2D classes are shown in figs S7B and S7D, respectively. Refinement of the remaining two 3D intasome classes produced considerably less detailed and symmetric reconstructions, indicating partial disorder and/or dissociation of the particles (fig. S6, middle panel).

37,021 strand transfer complex particles selected upon reference-free 2D classification were used in the 3D refinement, using the intasome structure, low-pass

filtered to 40 Å, as an initial model. 3D classification/refinement approaches did not improve resolution, likely due to the modest size of the strand transfer complex dataset. The final strand transfer complex reconstruction at 8.6 Å resolution was obtained by imposing two-fold symmetry during refinement followed by postprocessing using a B factor of -1,100, as implemented in Relion-1.4 (Fig. 3A, fig. S13). UCSF Chimera (*48*) and PyMol (https://www.pymol.org/) were used to prepare the figures.

Pseudo-atomic models of the MVV intasome and the strand transfer complex

Crystal structures of the individual NTD, CCD, and CTD domains of MVV IN were docked into the map as rigid bodies in UCSF Chimera (*48*), revealing 16 IN subunits; 12 NTDs, 16 CCDs, and 14 CTDs could be unambiguously positioned in the map. The CCD-CTD linkers predicted to form an α helix (fig. S9C) were modeled *de novo* in Coot for all IN subunits in agreement with the electron density map. NTD-CCD linkers from chains A and I, located in the best-resolved region of the map, could be manually built and refined in Coot (fig. S10B). Residues 228-236 of the loop in CTD A, D, I, and L were manually adjusted in Coot and residues 140-159 of CCD A and I were modeled on the basis of the PFV intasome crystal structure (PDB code 3OY9) to fit into the density and to avoid clashes with vDNA. The initial intasome model was refined for 5 cycles in Phenix (using phenix.real_space_refine). The final adjustments were made in Coot (table S2). Movie S1 demonstrates the quality of the map and details of the model fitting. The strand transfer complex model was constructed starting with the intasome model; the target DNA portion, initially modeled based on the structure of the RSV strand transfer complex (PDB code 5EJK), was fitted into the density and extended in Coot (table S2). The final cryo-EM reconstructions and fitted coordinates were deposited with the EM data bank under accession codes EMD-4138 and EMD-4139, with the associated PDB codes 5M0Q and 5M0R, respectively.

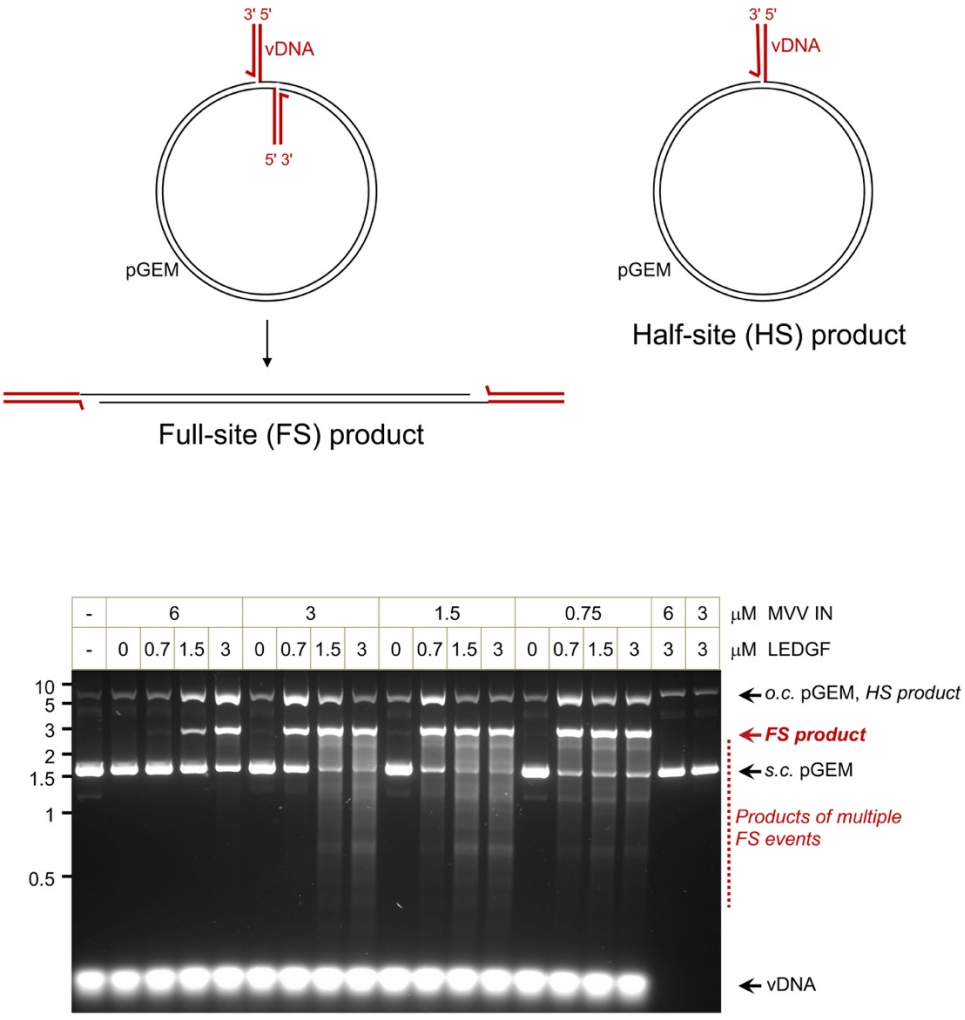
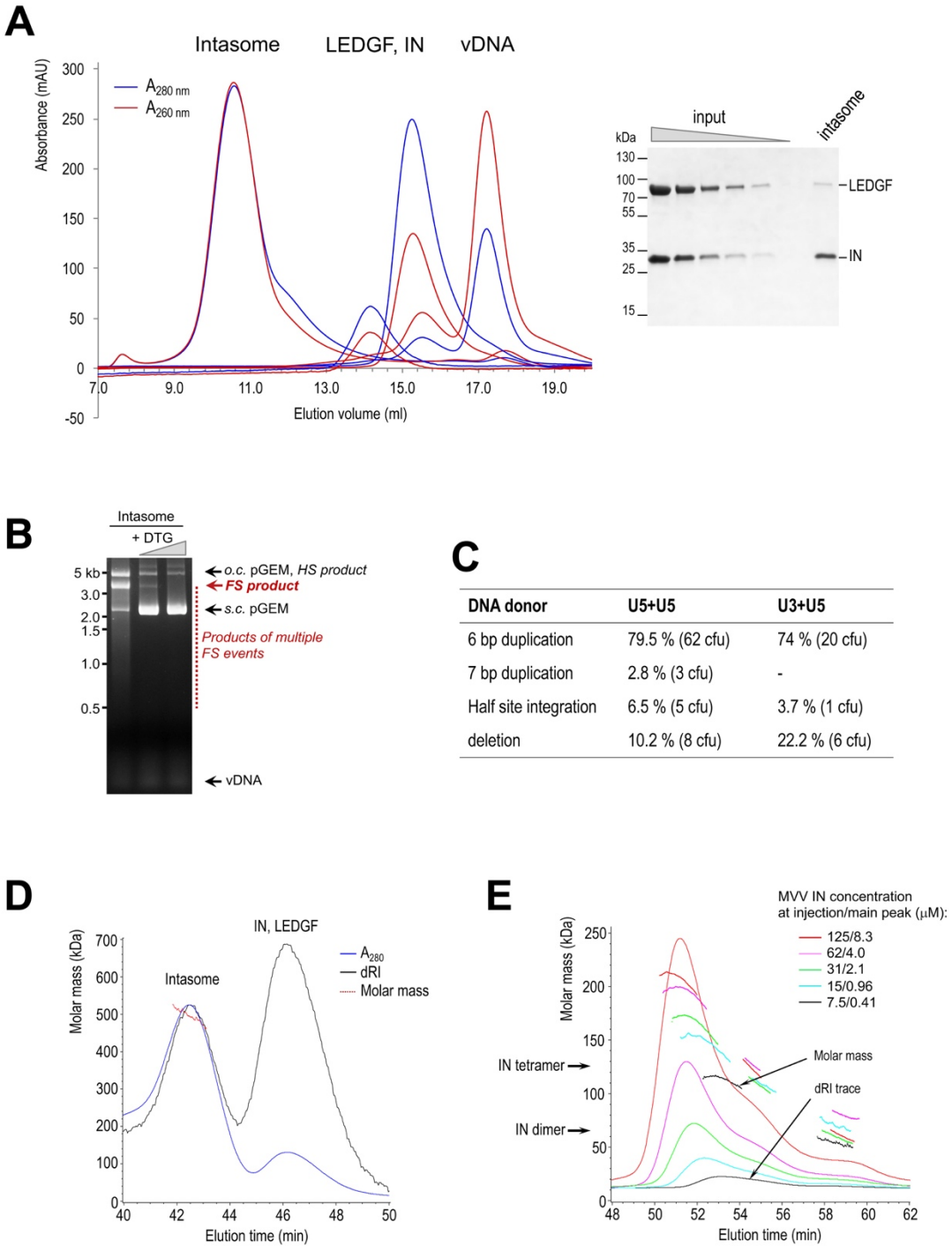


Fig. S1. LEDGF-dependent strand transfer activity of MVV IN.

Top, the major types of IN-mediated strand transfer products with an oligonucleotide mimic of the pre-processed MVV U5 vDNA end and a plasmid as target DNA. Concerted insertion of a pair of oligonucleotide ends leads to the full-site (FS) product, which after deproteinization migrates in agarose gels as a linearized plasmid. Integration of a single vDNA end leads to formation of branched half-site (HS) products, which co-migrate with the open circular form of the target plasmid. Bottom, strand transfer assays using indicated MVV IN and LEDGF protein concentrations. The reaction products were separated in an agarose gel and detected by staining with ethidium bromide. The vDNA mimic was omitted in the two right-most lanes. Migration positions of the reaction products, vDNA and open circular (o.c.) and supercoiled (s.c.) pGEM are indicated on the right of the gel; positions of DNA molecular size markers (kb) are shown to the left.



DNA donor	U5+U5	U3+U5
6 bp duplication	79.5 % (62 cfu)	74 % (20 cfu)
7 bp duplication	2.8 % (3 cfu)	-
Half site integration	6.5 % (5 cfu)	3.7 % (1 cfu)
deletion	10.2 % (8 cfu)	22.2 % (6 cfu)

Fig. S2. Purification, activity and sizing the MVV intasome.

A. Left, Chromatogram of the MVV intasome, LEDGF, IN, and vDNA through a Superose-6 10/300 column in the presence of 1 M NaCl. The complex was assembled and purified in the absence of Mg^{2+}, which is essential for IN catalytic activities. A_{280} and A_{260} readings are shown as blue and red lines, respectively. Elution positions of free LEDGF,

MVV IN, and vDNA are indicated above their corresponding peaks. The highly proteinaceous nature of the MVV intasome explains its observed A_{260}/A_{280} ratio of ~1. Right, serial 2-fold dilutions of an equimolar IN-LEDGF mixture used as input in intasome assembly in comparison to the purified intasome on a Coomassie-stained SDS-PAGE gel. Based on densitometry of the protein bands, the ratio of LEDGF to IN in the intasomal peak did not exceed 1:10. Migration positions of LEDGF and IN are labelled on the right of the gel with molecular mass standards (kDa) to the left. **B.** Strand transfer assays using purified MVV intasomes. Supercoiled pGEM was incubated with the intasome in the absence (left-most lane) or presence of 0.5 μM and 5 μM dolutegravir (DTG, middle and right-most lanes, respectively). **C.** Sequence analysis of strand transfer products obtained with purified intasomes and pGEM target DNA. Intasomes were assembled either as symmetrized complexes with MVV vDNA U5 ends (U5+U5) or with an equimolar mixture of U5 and U3 ends (U5+U3). DNA migrating in agarose gels as FS products was cloned and sequenced to establish precise insertion positions and identities (U3 or U5) of the vDNA end(s) in each product. The table shows fractions and totals (colony forming units, cfu) of duplications or deletions in target DNA, as well as half-site integration events. **D.** SEC-MALLS analysis of the MVV intasome. The chromatograms represent A_{280} (blue line) and differential refractive index (dRI, black line). Red dots indicate averaged molecular mass (kDa, vertical axis) of molecules present in the peak, determined at 1-sec intervals. **E.** SEC-MALLS of free MVV IN, injected at concentrations of 7.5-125 μM, as indicated. Color-coded chromatogram traces and dots represent dRI and averaged molecular mass measurements, respectively. Main peak protein concentrations, determined from the dRI measurements, are indicated in the legend. At the lowest MVV IN input (7.5 μM), the mean molecular mass of molecules eluting in the main peak (retention time of 53-54 min) approached that of tetrameric MVV IN species (4 x 32 kDa) with a measured average concentration of ~0.41 μM. Higher MVV IN inputs resulted in earlier elution positions and higher aggregation states, probably corresponding to mixtures of tetrameric and octameric species.

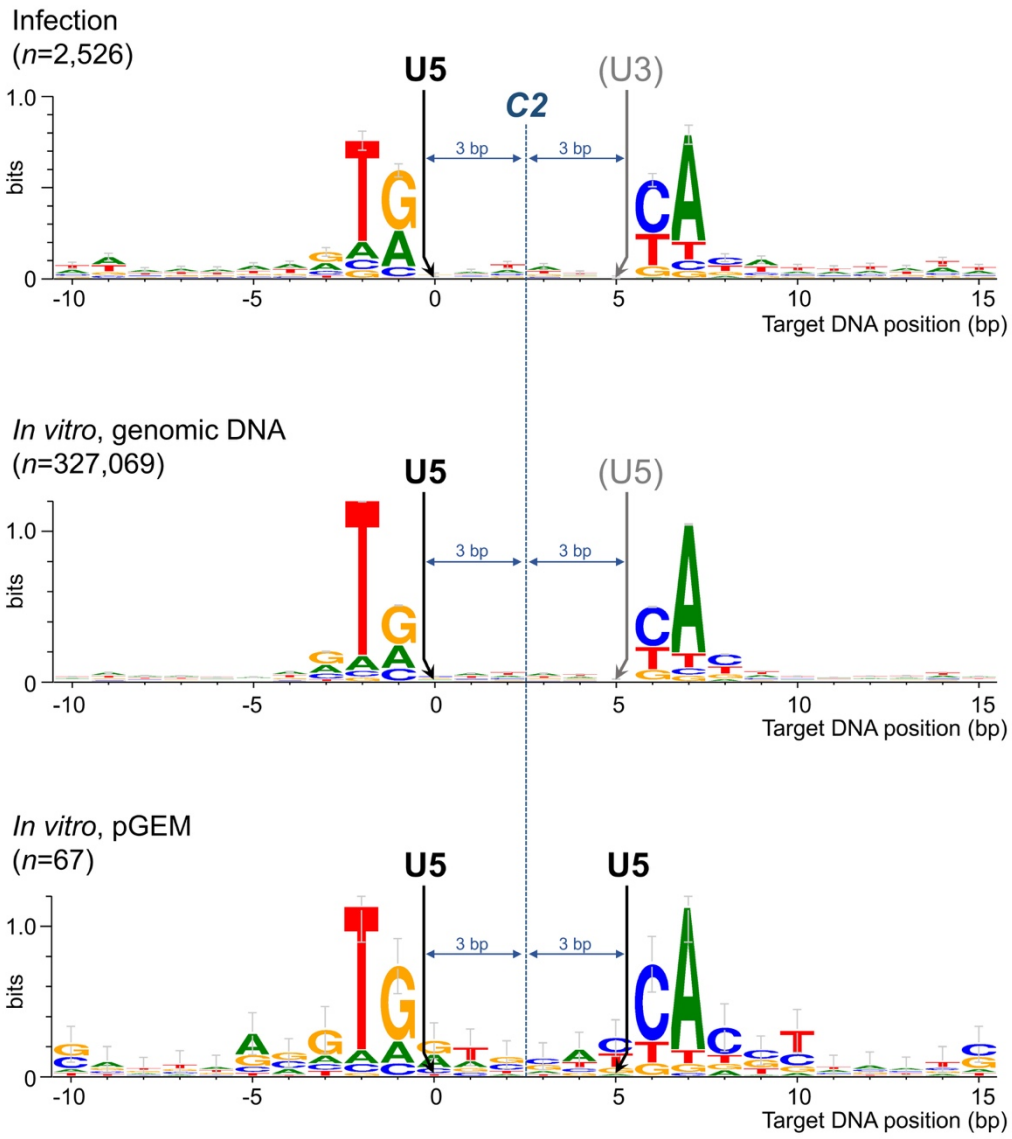


Fig. S3. Nucleotide sequence preferences at MVV integration sites.

Sequence logos generated from alignments of integration target sites in virus-infected sheep cells (top) or by incubating recombinant MVV intasomes with deproteinized sheep DNA (middle) or pGEM plasmid (bottom; only integration sites with 6-bp duplications were used in this case). Position 0 corresponds to the nucleotide of the target sequence joined to the U5 end (indicated). Because both vDNA ends were recovered in the experiment with pGEM target, the position of the second vDNA end insertion in the complementary strand is also known precisely and is separated by 6-bp. Although the position of the second vDNA end (U3 in the case of the infection experiment, or U5 in the case of *in vitro* integration) are not directly known, the two-fold symmetry (*C2*, indicated as a dashed line) of the logos in each case can only arise if the second vDNA end integrated 3 bp from the symmetry axis, which is 6 bp from the mapped insertion.

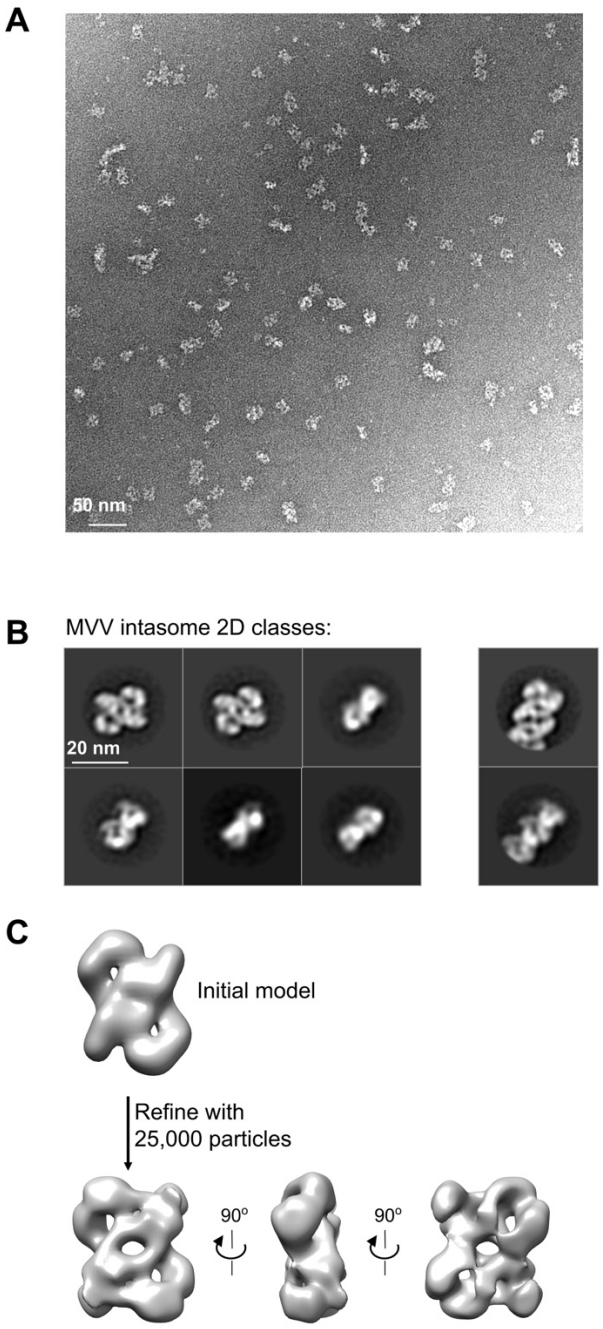


Fig. S4. Negative stain EM of MVV intasomes.
A. A representative electron micrograph of negatively stained MVV intasome particles. **B.** Left, reference-free 2D class averages representing the major species in the sample. Right, two representative 2D classes of expanded complexes due to aggregation of the core intasome. **C.** The initial 3D model (top) and the refined map from negatively-stained particles (bottom).

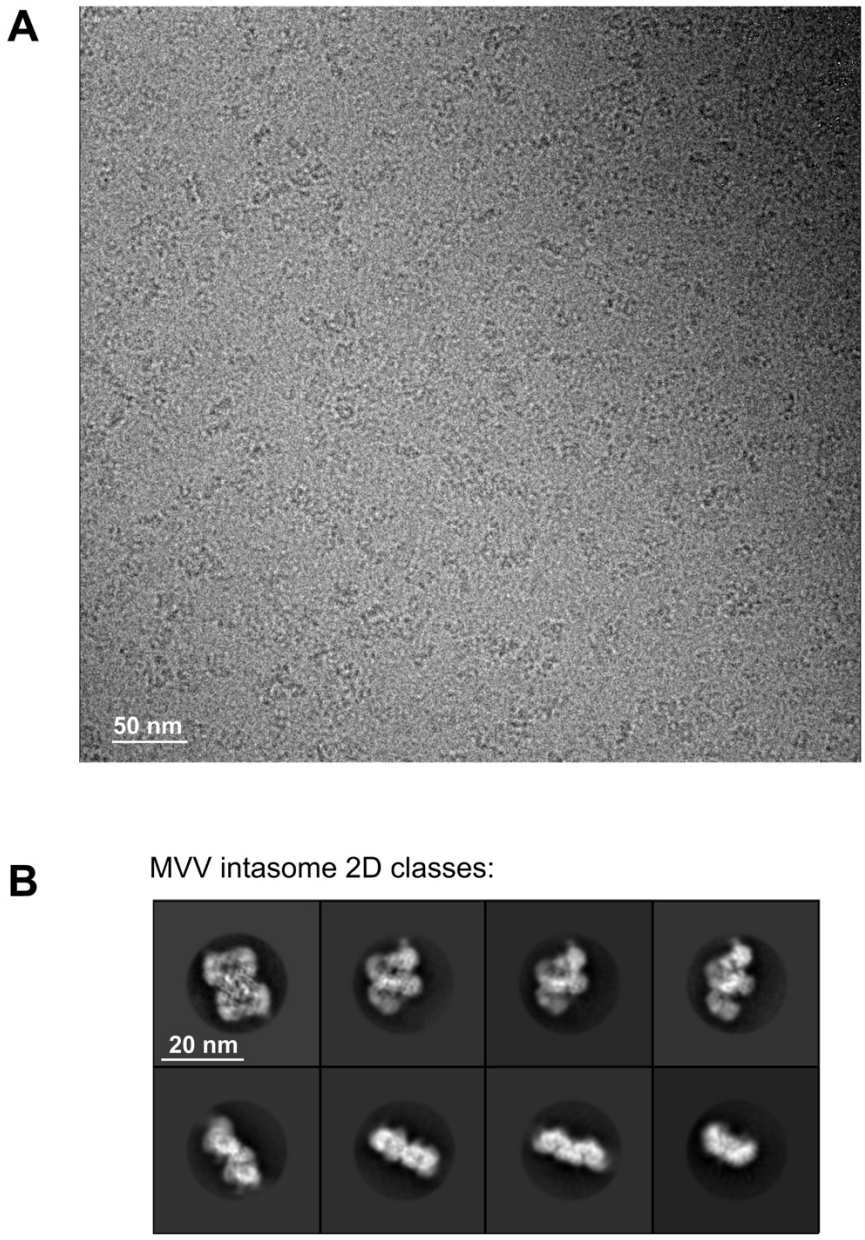


Fig. S5. Overview of the MVV intasome cryo-EM data.
A. A cryo-electron micrograph of MVV intasomes on ultrathin carbon support.
B. Representative reference-free 2D class averages.

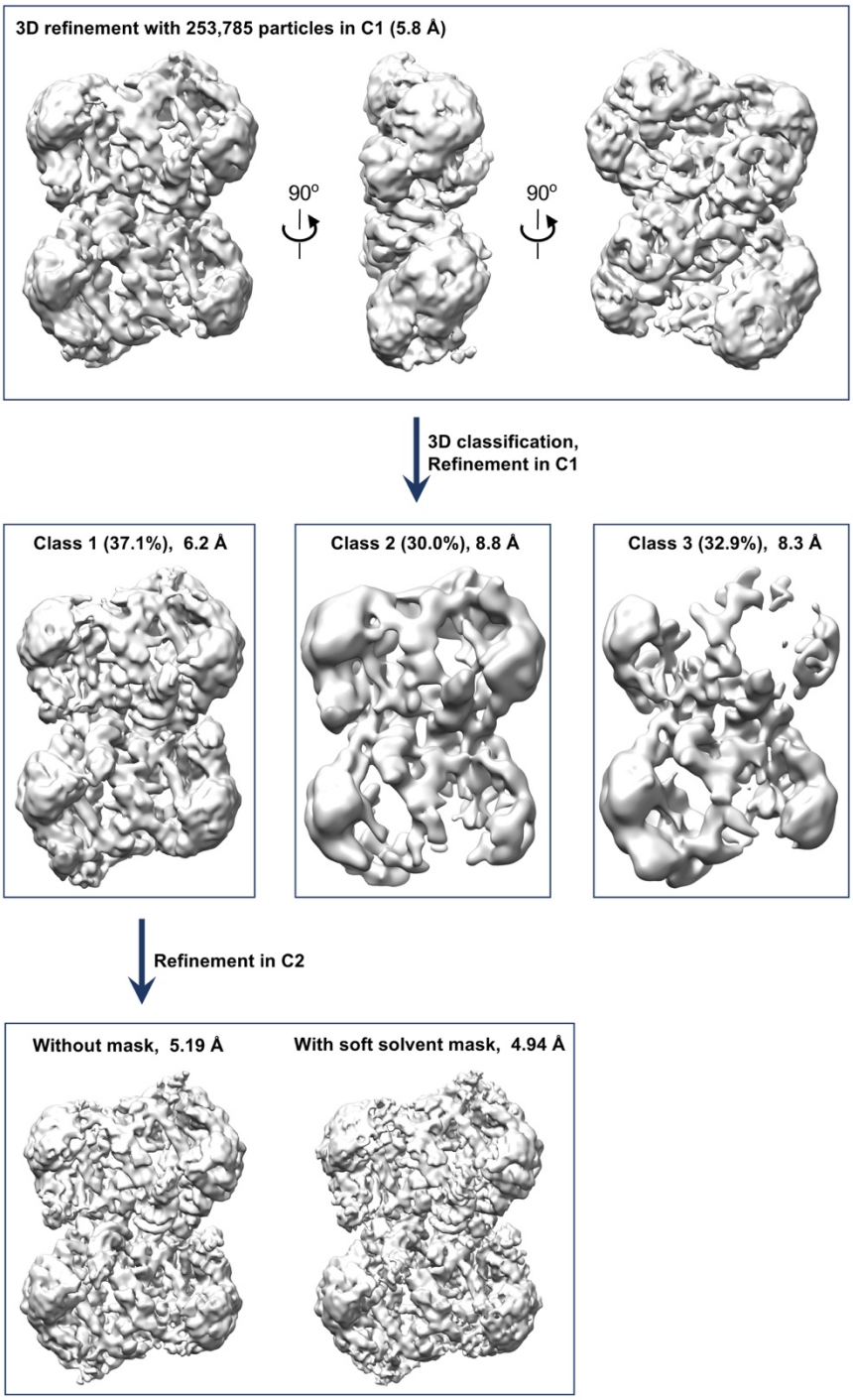


Fig. S6. MVV intasome cryo-EM classification and refinement strategy.
Refinement using all particles selected by reference-free 2D classification without imposing symmetry resulted in reconstruction at 5.8 Å. Classification of the initial particle set into three 3D classes isolated a subset of particles that were used to refine a 5.19 Å resolution structure with 2-fold symmetry (lower panel, left). The final reconstruction at 4.94 Å resolution was obtained by imposing a soft solvent mask during refinement.

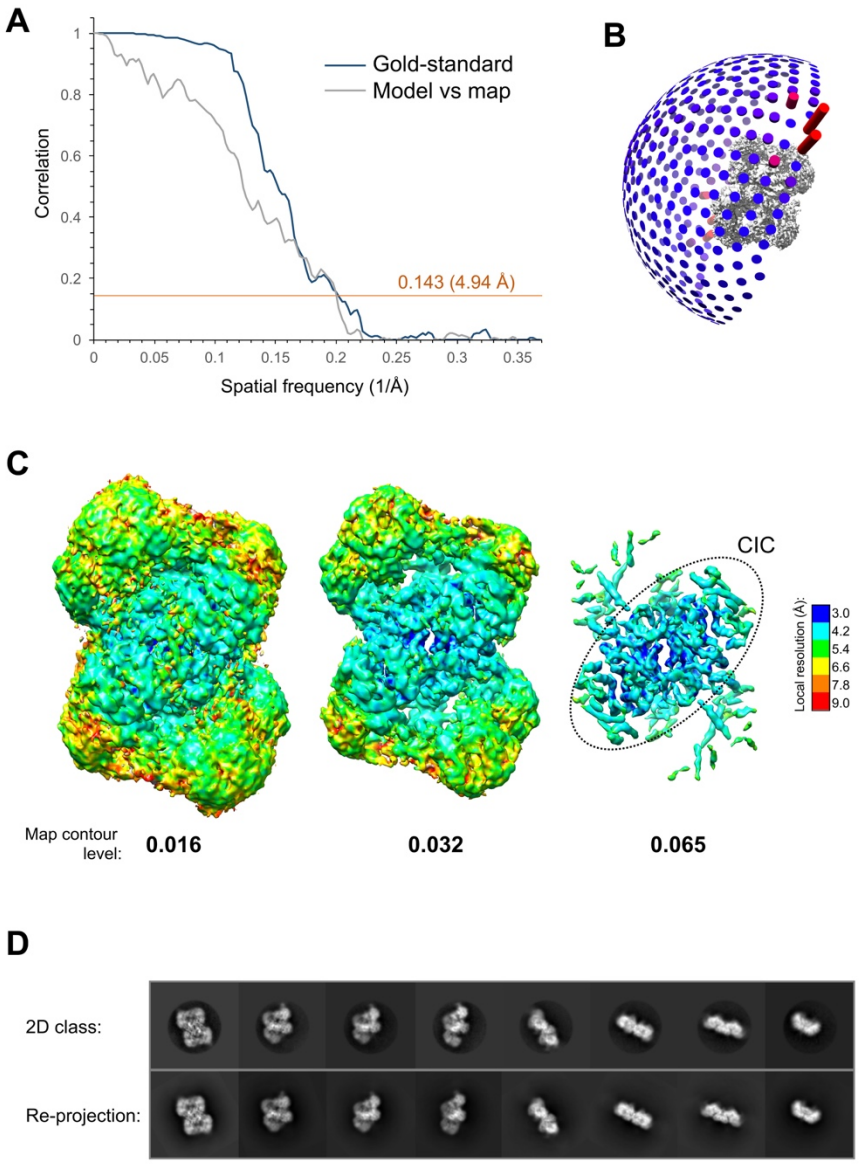


Fig. S7. Overview of cryo-EM data refinement for the intasome complex.
A. Gold standard Fourier-shell correlation (FSC) between two independently refined models (one for each half of the data, blue line) and model versus final map (grey line). The orange line demarcates the 0.143 correlation cutoff (corresponding to 4.94 Å resolution). **B.** Euler angle distribution plot showing the relative orientations of all the particles included in the final three-dimensional reconstruction. **C.** The final map contoured at level 0.016 (left), 0.032 (middle), or 0.065 (right), colored by local resolution estimated with ResMap. The conserved intasome core (CIC) structure is indicated with an oval. Resolution color scale is shown to the right. **D.** Re-projections of the refined electron density maps and matched reference-free 2D class averages.

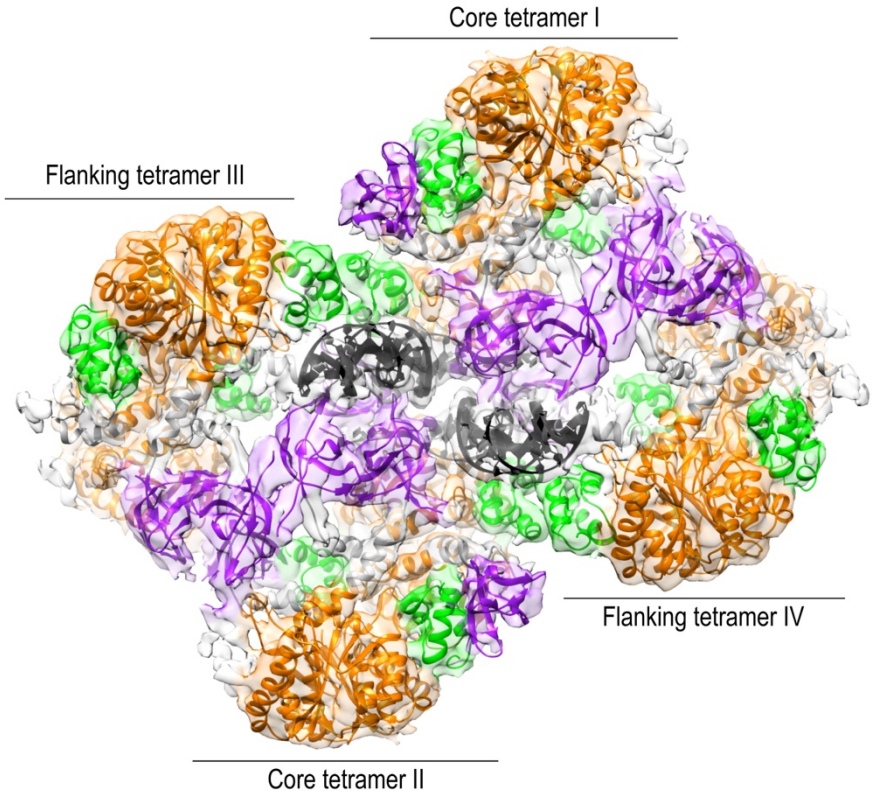


Fig. S8. Cryo-EM density and pseudo-atomic model colored by domain.
NTDs are shown in green, CCDs in orange, CTDs in purple, α-helical CCD-CTD linkers in grey, and vDNA in black.

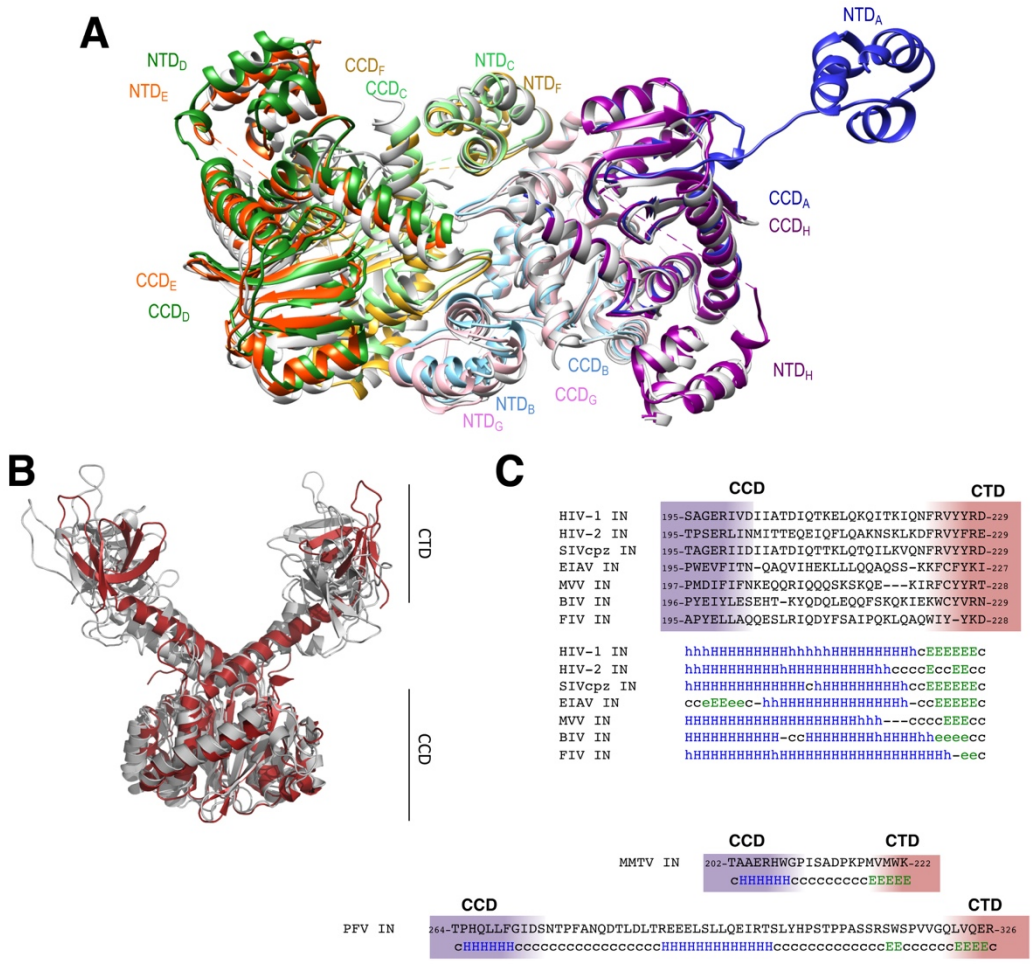


Fig. S9. Composition of MVV IN tetramers within the intasome.
A. Superposition of the core tetramer I (dark green, light green, sky blue and blue) and flanking tetramer III (red, yellow, pink and purple) with the crystal structure of the isolated MVV IN NTD-CCD construct (PDB code 3HPH; grey). NTD_A of the core tetramer reaches across the synaptic interface to interact with the opposing vDNA end. CTDs are omitted for clarity. **B.** Superposition of the crystal structure of HIV-1 IN CCD-CTD construct (PDB code 1EX4, dark red) with MVV IN dimers A-B, C-D, E-F, and G-H from the intasome (grey). **C.** Top, sequence alignment of lentiviral IN CCD-CTD linkers and their secondary structure from crystal structures, when available, or predicted by the PHD algorithm (*49*) when full or partial structure information is missing (PDB accession codes 1EX4, 3F9K, 1C6V, 3HPH, 3KKR, and 4PA1 for HIV-1, HIV-2, SIV, MVV, BIV, and FIV INs, respectively). Bottom, sequences and secondary structures of MMTV and PFV IN CCD-CTD linkers (PDB accession codes 3JCA and 3OA9, respectively).

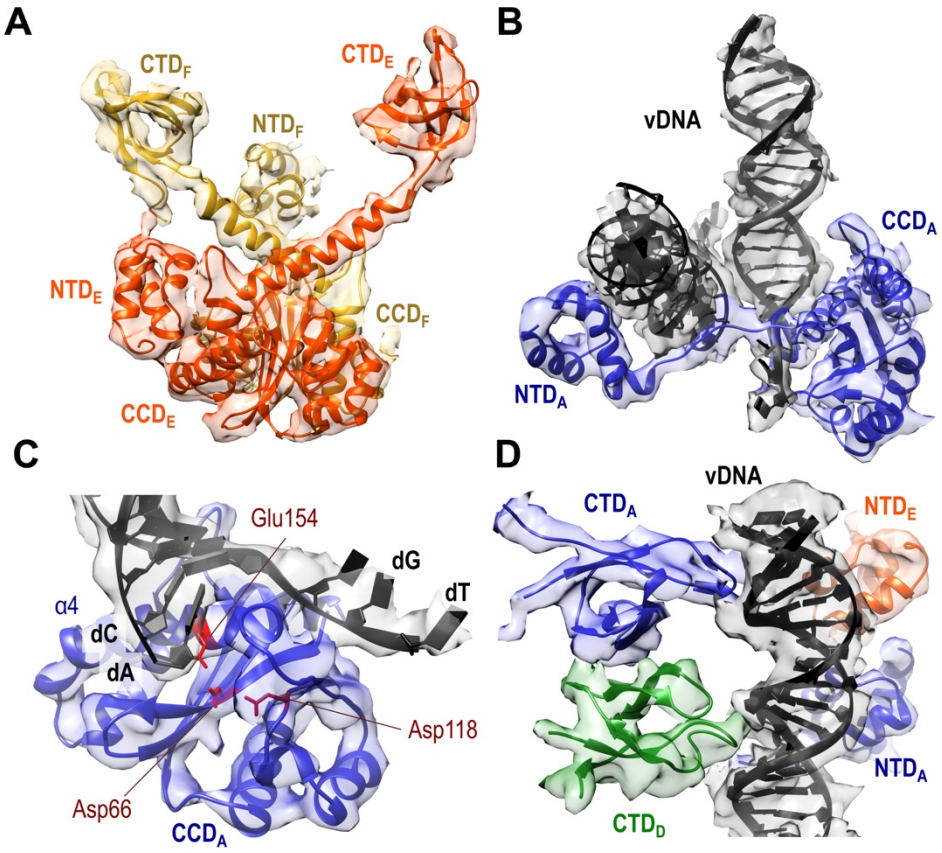


Fig. S10. Detailed views of the cryo-EM map and fitted structures.
A. Representative IN dimer (chains E and F). **B.** Viral DNA and IN_A structures shown in dark grey and blue, respectively, highlighting the NTD_A-CCD_A linker defined in the electron density. **C.** View of the active site of CCD_A engaged with vDNA. The electron density map and pseudo-atomic model are shown as semitransparent surface and cartoons, respectively. Residues comprising the D,D-35-E active site motif (Asp66, Asp118, and Glu154) of MVV IN are shown as sticks. The α4 helix of the MVV IN CCD and the invariant vDNA CA dinucleotide and its complement are indicated. **D.** Viral DNA anchored with two CTDs of the core tetramer (blue and green) and two NTDs, orange and blue, from a flanking and core tetramer, respectively.

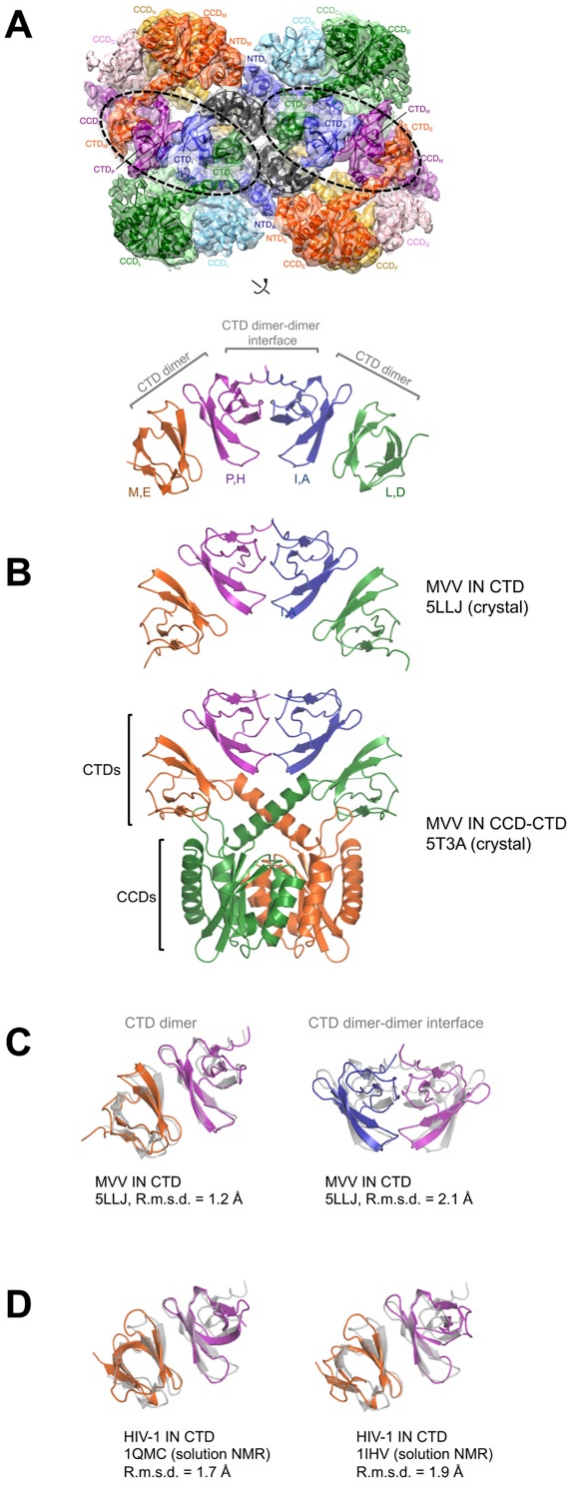


Fig. S11. CTDs bridge core and flanking IN tetramers of the intasome.
A. Top: Locations of the CTD quartets bridging IN tetramers I and IV (and the symmetry-related bridge between tetramers II and III) are indicated with ovals. Bottom: the CTD quartet containing a pair of interacting dimers. Each CTD dimer buries ~1,000 Å^2 of

surface area, and an additional ~1,250 Å^2 is buried in the interface between the dimers. **B.** Analogous quartet CTD arrangements observed in crystals of MVV CTD (top, contents of two asymmetric units are shown) and CCD-CTD (bottom, the asymmetric unit contains a single protein chain) constructs. **C.** Superposition of the CTD dimer and CTD dimer-dimer interfaces observed in the intasome (grey) and in the MVV IN CTD crystal structure. **D.** Superposition of the MVV intasome CTD dimer (grey) and the dimeric HIV-1 IN CTD structures observed in solution (PDB IDs 1QMC and 1IHV).

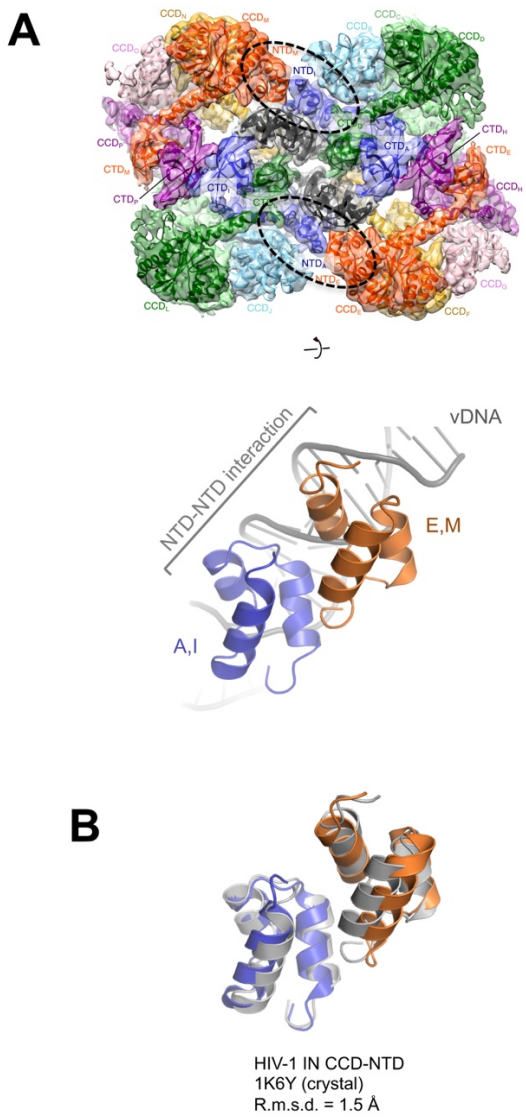


Fig. S12. NTD_A-NTD_E and NTD_I-NTD_M interactions within the intasome.
A. Top: The locations of the NTD dimers (indicated with ovals). Bottom: the NTD dimer (chain A in blue and chain E in orange) interacting with vDNA (grey). **B.** Superposition of the MVV intasome NTD dimer (grey) and the analogous dimer within the asymmetric unit in the HIV-1 IN NTD-CCD crystal structure (PDB ID 1K6Y, chains C and D)(*14*). The NTD-NTD contact buries ~315 $Å^2$ of surface area.

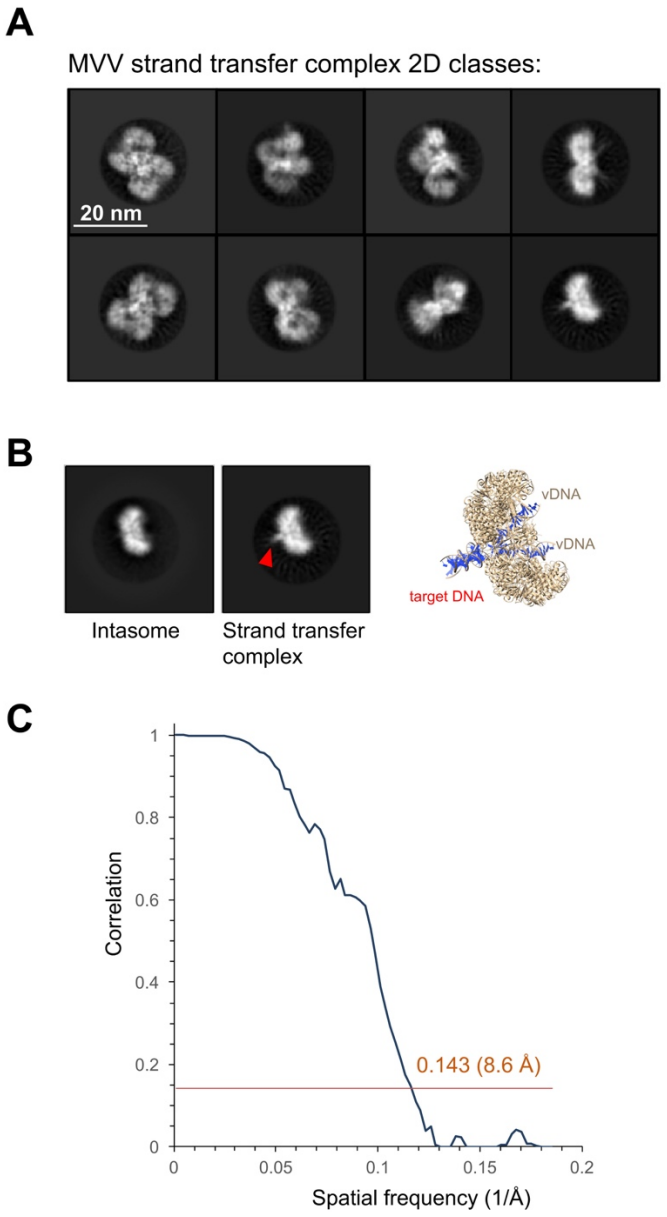


Fig. S13. Cryo-EM analysis of the MVV strand transfer complex.
A. Representative reference-free 2D class averages of the strand transfer complex. **B.** Comparison of 2D class averages of the MVV intasome and the strand transfer complex. Density corresponding to the target DNA that is clearly visible in the strand transfer complex is indicated with a red arrowhead. An image of the strand transfer complex model in the same orientation as the class average is shown to the right. **C.** Gold-standard FSC between refined strand transfer complex reconstructions from half sets; the orange line demarcates the 0.143 correlation cutoff, corresponding to a resolution of 8.6 Å.

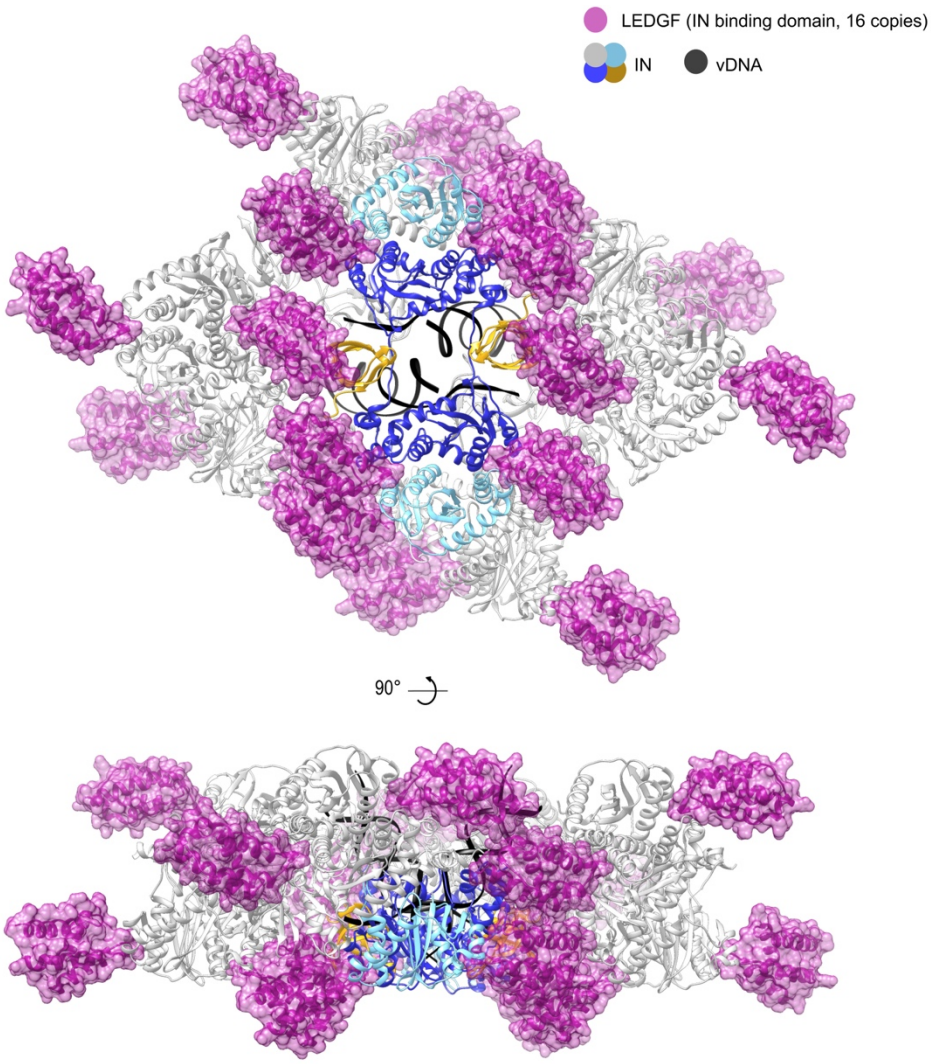


Fig. S14. Model of the MVV intasome-LEDGF complex.
The intasome is shown as cartoons (with CIC in color and the remainder of the intasome in grey, as in Fig. 2) and 16 copies of the LEDGF IN binding domain in space-fill mode (magenta). The model was constructed by iterative superposition of the co-crystal structure of the MVV IN NTD-CCD construct with the IN binding domain of LEDGF (PDB accession code 3HPH) (*12*) onto the pseudo-atomic model of the MVV intasome.

Table S1. X-ray data collection and refinement statistics.

	CTD	CCD-CTD [a]
Data collection		
Wavelength (Å)	0.97795	0.9282
Space group	$P4_32_12$	$C222_1$
Unit cell parameters		
a, b, c (Å)	64.25, 64.25, 72.33	78.05, 124.83, 62.38
α, β, γ (°)	90, 90, 90	90, 90, 90
Number of crystals used	1	1
Resolution (Å)	48.04-1.78 (1.83-1.78) [b]	62.42-2.50 (2.57-2.50) [b]
Number of reflections		
measured	184,115 (10,797)	102,586 (7,386)
unique	15,111 (1,088)	10,859 (773)
Completeness (%)	99.9 (99.4)	99.7 (98.4)
Multiplicity	12.2 (9.9)	9.4 (9.6)
$\langle I/\sigma(I)\rangle$	23.7 (1.5)	13.0 (1.7)
$CC_{1/2}$	1.0 (0.53)	0.99 (0.81)
R_{merge}	0.057 (1.46)	0.130 (1.54)
R_{pim} (all I^+ and I^-)	0.024 (0.70)	0.064 (0.55)
Refinement statistics		
Resolution range (Å)	38.47-1.78 (1.84-1.78)	45.39-2.50 (2.59-2.50)
Number of reflections		
work set	15,075 (1,468)	8,771 (555)
free	778 (82)	521 (48)
R_{work}/R_{free}	0.186/0.234	0.179/0.240
Number of atoms		
total	1,083	1,796
protein	940	1,726
ligands	1	24
solvent	142	46
R.m.s. deviations from ideal		
bond lengths (Å)	0.004	0.002
bond angles (°)	0.67	0.52
Average B-factor (Å^2)	42.2	62.4
Ramachandran plot quality (%) [c]		
favored	100	99
disallowed	0	0

[a] Data were corrected for anisotropy using STARANISO server (http://staraniso.globalphasing.org).
[b] Values in parentheses correspond to the highest resolution bin.
[c] Analyzed using MolProbity (http://molprobity.biochem.duke.edu/).

Table S2. Cryo-EM data collection and refinement.

	Intasome	**Strand transfer complex**
Data collection		
Microscope	Polara G2	Polara G2
Operating voltage (kV)	300	300
Detector	Gatan K2	Gatan K2
Pixel size (Å)	1.35	1.35
Defocus range (μm)	2.0-5.5	2.0-5.9
Number of frames per movie	40	30
Total electron dose (e/Å^2)	50	40
Reconstruction		
Software	Relion-1.4	Relion-1.4
Number of particles used	94,283	37,021
Accuracy of rotation (°)	4.28	7.62
Accuracy of translation (pixels)	1.76	1.78
Map sharpening B-factor (Å^2)	-330	-1,100
Final resolution (Å) [a]	4.94	8.6
Modeling and real-space refinement		
Protein residues	3,903	3,908
DNA nucleotides	80	156
Software for real-space refinement [b]	phenix.real_space_refine	-
Real-space correlation coefficient		
prior to refinement	0.55	0.83
post refinement	0.80	-
MolProbity clash score [c]		
prior to refinement	59.87	28.01
post refinement	25.79	-
Ramachandran plot quality (%) [c]		
favored	96.91	95.85
disallowed	0.21	0.41

[a] Based on the Fourier shell correlation of 0.143 between half-sets.
[b] Only the intasome model was subjected to real-space refinement.
[c] Analyzed with MolProbity server (http://molprobity.biochem.duke.edu/).

Movie S1. Overview of the segmented electron density map and model fitting.

References

1. P. Lesbats, A. N. Engelman, P. Cherepanov, Retroviral DNA Integration. *Chem. Rev.*, **116**, 12730-12757 (2016).
2. S. Hare, S. S. Gupta, E. Valkov, A. Engelman, P. Cherepanov, Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature* **464**, 232-236 (2010).
3. G. N. Maertens, S. Hare, P. Cherepanov, The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468**, 326-329 (2010).
4. Z. Yin, K. Shi, S. Banerjee, K. K. Pandey, S. Bera, D. P. Grandgenett, H. Aihara, Crystal structure of the Rous sarcoma virus intasome. *Nature* **530**, 362-366 (2016).
5. A. Ballandras-Colas, M. Brown, N. J. Cook, T. G. Dewdney, B. Demeler, P. Cherepanov, D. Lyumkis, A. N. Engelman, Cryo-EM reveals a novel octameric integrase structure for betaretroviral intasome function. *Nature* **530**, 358-361 (2016).
6. F. Dyda, A. B. Hickman, T. M. Jenkins, A. Engelman, R. Craigie, D. R. Davies, Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science* **266**, 1981-1986 (1994).
7. J. C. Chen, J. Krucinski, L. J. Miercke, J. S. Finer-Moore, A. H. Tang, A. D. Leavitt, R. M. Stroud, Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 8233-8238 (2000).
8. M. Li, K. A. Jurado, S. Lin, A. Engelman, R. Craigie, Engineered hyperactive integrase for concerted HIV-1 DNA integration. *PLoS One* **9**, e105078 (2014).
9. R. Craigie, F. D. Bushman, Host Factors in Retroviral Integration and the Selection of Integration Target Sites. *Microbiol. Spectr.* **2**, DOI: 10.1128/microbiolspec.MDNA1123-0026-2014 (2014).
10. P. Cherepanov, LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity in vitro. *Nucleic Acids Res.* **35**, 113-124 (2007).
11. B. A. Johns, T. Kawasuji, J. G. Weatherhead, T. Taishi, D. P. Temelkoff, H. Yoshida, T. Akiyama, Y. Taoda, H. Murai, R. Kiyama, M. Fuji, N. Tanimoto, J. Jeffrey, S. A. Foster, T. Yoshinaga, T. Seki, M. Kobayashi, A. Sato, M. N. Johnson, E. P. Garvey, T. Fujiwara, Carbamoyl pyridone HIV-1 integrase inhibitors 3. A diastereomeric approach to chiral nonracemic tricyclic ring systems and the discovery of dolutegravir (S/GSK1349572) and (S/GSK1265744). *J. Med. Chem.* **56**, 5901-5916 (2013).
12. S. Hare, F. Di Nunzio, A. Labeja, J. Wang, A. Engelman, P. Cherepanov, Structural basis for functional tetramerization of lentiviral integrase. *PLoS Pathog.* **5**, e1000515 (2009).
13. P. J. Lodi, J. A. Ernst, J. Kuszewski, A. B. Hickman, A. Engelman, R. Craigie, G. M. Clore, A. M. Gronenborn, Solution structure of the DNA binding domain of HIV-1 integrase. *Biochemistry* **34**, 9826-9833 (1995).

14. J. Y. Wang, H. Ling, W. Yang, R. Craigie, Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. *EMBO J.* **20**, 7333-7343 (2001).
15. D. R. Davies, I. Y. Goryshin, W. S. Reznikoff, I. Rayment, Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. *Science* **289**, 77-85 (2000).
16. S. P. Montano, Y. Z. Pigli, P. A. Rice, The mu transpososome structure sheds light on DDE recombinase evolution. *Nature* **491**, 413-417 (2012).
17. P. Cherepanov, G. Maertens, P. Proost, B. Devreese, J. Van Beeumen, Y. Engelborghs, E. De Clercq, Z. Debyser, HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J. Biol. Chem.* **278**, 372-381 (2003).
18. C. J. McKee, J. J. Kessl, N. Shkriabai, M. J. Dar, A. Engelman, M. Kvaratskhelia, Dynamic modulation of HIV-1 integrase structure and function by cellular lens epithelium-derived growth factor (LEDGF) protein. *J. Biol. Chem.* **283**, 31802-31812 (2008).
19. S. P. Lee, J. Xiao, J. R. Knutson, M. S. Lewis, M. K. Han, Zn2+ promotes the self-association of human immunodeficiency virus type-1 integrase in vitro. *Biochemistry* **36**, 173-180 (1997).
20. A. Faure, C. Calmels, C. Desjobert, M. Castroviejo, A. Caumont-Sarcos, L. Tarrago-Litvak, S. Litvak, V. Parissi, HIV-1 integrase crosslinked oligomers are active in vitro. *Nucleic Acids Res.* **33**, 977-986 (2005).
21. T. Jacks, M. D. Power, F. R. Masiarz, P. A. Luciw, P. J. Barr, H. E. Varmus, Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* **331**, 280-283 (1988).
22. J. A. Briggs, M. N. Simon, I. Gross, H. G. Krausslich, S. D. Fuller, V. M. Vogt, M. C. Johnson, The stoichiometry of Gag protein in HIV-1. *Nat. Struct. Mol. Biol.* **11**, 672-675 (2004).
23. A. Engelman, In vivo analysis of retroviral integrase structure and function. *Adv. Virus Res.* **52**, 411-426 (1999).
24. D. P. Maskell, L. Renault, E. Serrao, P. Lesbats, R. Matadeen, S. Hare, D. Lindemann, A. N. Engelman, A. Costa, P. Cherepanov, Structural basis for retroviral integration into nucleosomes. *Nature* **523**, 366-369 (2015).
25. O. S. Andresson, J. E. Elser, G. J. Tobin, J. D. Greenwood, M. A. Gonda, G. Georgsson, V. Andresdottir, E. Benediktsdottir, H. M. Carlsdottir, E. O. Mantyla, Nucleotide sequence and biological properties of a pathogenic proviral molecular clone of neurovirulent visna virus. *Virology* **193**, 89-105 (1993).
26. E. Valkov, S. S. Gupta, S. Hare, A. Helander, P. Roversi, M. McClure, P. Cherepanov, Functional and structural characterization of the integrase from the prototype foamy virus. *Nucleic Acids Res.* **37**, 243-255 (2009).
27. W. Kabsch, Xds. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125-132 (2010).
28. P. R. Evans, G. N. Murshudov, How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204-1214 (2013).
29. G. Winter, C. M. Lobley, S. M. Prince, Decision making in xia2. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1260-1273 (2013).

30. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).
31. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486-501 (2010).
32. P. D. Adams, P. V. Afonine, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010).
33. V. B. Chen, W. B. Arendall, 3rd, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, D. C. Richardson, MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21 (2010).
34. V. Ellison, P. O. Brown, A stable complex between integrase and viral DNA ends mediates human immunodeficiency virus integration in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 7316-7320 (1994).
35. G. Petursson, N. Nathanson, G. Georgsson, H. Panitch, P. A. Palsson, Pathogenesis of visna. I. Sequential virologic, serologic, and pathologic studies. *Lab. Invest.* **35**, 402-412 (1976).
36. E. Serrao, P. Cherepanov, A. N. Engelman, Amplification, Next-generation Sequencing, and Genomic DNA Mapping of Retroviral Integration Sites. *J. Vis. Exp.*, (2016).
37. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
38. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Genome Project Data Processing, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
39. A. R. Quinlan, BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 11-34 (2014).
40. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).
41. J. M. Bell, M. Chen, P. R. Baldwin, S. J. Ludtke, High resolution single particle refinement in EMAN2.1. *Methods* **100**, 25-34 (2016).
42. S. H. Scheres, Semi-automated selection of cryo-EM particles in RELION-1.3. *J Struct Biol* **189**, 114-122 (2015).
43. D. N. Mastronarde, Automated electron microscope tomography using robust prediction of specimen movements. *J Struct Biol* **152**, 36-51 (2005).
44. S. Q. Zheng, E. Palovcak, J.-P. Armache, Y. Cheng, D. A. Agard, Anisotropic Correction of Beam-induced Motion for Improved Single-particle Electron Cryo-microscopy. *BioArxiv*, DOI: http://dx.doi.org/10.1101/061960 (2016).
45. A. Rohou, N. Grigorieff, CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216-221 (2015).
46. S. H. Scheres, S. Chen, Prevention of overfitting in cryo-EM structure determination. *Nat Methods* **9**, 853-854 (2012).

47. A. Kucukelbir, F. J. Sigworth, H. D. Tagare, Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63-65 (2014).
48. Z. Yang, K. Lasker, D. Schneidman-Duhovny, B. Webb, C. C. Huang, E. F. Pettersen, T. D. Goddard, E. C. Meng, A. Sali, T. E. Ferrin, UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J. Struct. Biol.* **179**, 269-278 (2012).
49. B. Rost, C. Sander, Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584-599 (1993).