Published in final edited form as:

Nature. 2016 February 18; 530(7590): 358–361. doi:10.1038/nature16955.

Cryo-EM reveals a novel octameric integrase structure for β-retroviral intasome function

Allison Ballandras-Colas[1], **Monica Brown**[2], **Nicola J. Cook**[3], **Tamaria G. Dewdney**[1], **Borries Demeler**[4], **Peter Cherepanov**[3,5], **Dmitry Lyumkis**[2], and **Alan N. Engelman**[1]

[1]Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute and Department of Medicine, Harvard Medical School, 450 Brookline Avenue, Boston, Massachusetts 02215 USA

[2]Laboratory of Genetics and Helmsley Center for Genomic Medicine, The Salk Institute for Biological Studies, 10010 N. Torrey Pines Road, La Jolla, California 92037, USA

[3]Clare Hall Laboratories, The Francis Crick Institute, Blanche Lane, South Mimms EN6 3LD, UK

[4]Department of Biochemistry, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, Texas 78229, USA

[5]Division of Medicine, Imperial College London, St.-Mary's Campus, Norfolk Place, London W2 1PG, UK

Abstract

Retroviral integrase (IN) catalyzes the integration of viral DNA (vDNA) into host target (tDNA), which is an essential step in the lifecycle of all retroviruses[1]. Prior structural characterization of IN-vDNA complexes, or intasomes, from the spumavirus prototype foamy virus (PFV) revealed a functional IN tetramer[2–5], and it is generally believed that intasomes derived from other retroviral genera will employ tetrameric IN[6–9]. However, the intasomes of orthoretroviruses, which include all known pathogenic species, have not been characterized structurally. Using single-particle cryo-electron microscopy (cryo-EM) and X-ray crystallography, we determine here an unexpected octameric IN architecture for the β-retrovirus mouse mammary tumor virus (MMTV) intasome. The structure is composed of two core IN dimers, which interact with the vDNA ends and

Correspondence: (Alan N. Engelman) alan_engelman@dfci.harvard.edu, Tel: +1 617 632 4361, (Dmitry Lyumkis), dlyumkis@salk.edu, Tel: +1 858 453 4100.

Supplementary Information is available for this manuscript.

Authors Contributions A.B.-C. and A.N.E. discovered how to assemble MMTV intasomes; A.B.-C. and T.G.D. expressed and purified MMTV IN proteins for biochemical analysis; A.B.-C. assembled intasomes, characterized their biochemistry, supplied them for cryo-EM and centrifugation analyses and performed IN activity assays; M.B. and D.L. performed EM work, collected cryo-EM data and determined the structure; D.L. modeled the intasome structure; B.D. collected and analyzed the sedimentation velocity data; N.J.C. and P.C. expressed and purified IN_{CCD}, $IN_{NTD\text{-}CCD}$ and $IN_{CTD/212–266}$ constructs, established crystallization conditions and determined these structures.

Author Information Coordinates of cryo-EM density maps for the full and core intasome datasets were deposited with the EM databank under accession codes EMD-6440 and EMD-6441, respectively. X-ray diffraction data and the resulting IN_{CCD}, $IN_{NTD\text{-}CCD}$ and IN_{CTD} structures were deposited with the protein databank under accession codes 5CZ1, 5CZ2 and 5D7U, respectively. The core intasome structure was deposited with the protein database under accession code 3JCA.

The authors declare no competing financial interests.

structurally mimic the PFV IN tetramer, and two flanking IN dimers that engage the core structure via their IN C-terminal domains (CTDs). Contrary to the belief that tetrameric IN components are sufficient to catalyze integration, the flanking IN dimers were necessary for MMTV IN activity. The IN octamer solves a conundrum for the β- as well as α-retroviruses by providing critical CTDs to the intasome core that cannot be provided *in cis* due to evolutionarily restrictive catalytic core domain (CCD)-CTD linker regions. The octameric architecture of the MMTV intasome provides a new paradigm for the structural basis of retroviral DNA integration.

MMTV intasomes were assembled from IN and vDNA components by differential salt dialysis, akin to the strategy used for PFV intasomes[2]. Fractionation of assembly reactions by size exclusion chromatography (SEC) revealed a higher-order species with a distinct elution profile from those of IN and vDNA (Fig. 1a). To address biological relevance, reactions were conducted with supercoiled plasmid tDNA to monitor the concerted integration of two vDNA ends[10] (Fig. 1b). The SEC-purified complexes catalyzed efficient concerted integration activity, which was inhibited by the IN strand transfer inhibitor raltegravir (Fig. 1c). The sequencing of concerted integration products excised from agarose gels revealed that the majority contained 6 bp target site duplications flanking the integrated vDNA ends, which are known to occur during MMTV infection[11] (Fig. 1d). To address the specificity of complex formation, the invariant CA dinucleotide, which is essential for IN catalysis[12,13], was mutated to GT in the vDNA substrate. As the mutant vDNA failed to support complex formation (data not shown), we conclude that the higher-order species identified by SEC are *bona fide* MMTV intasomes. We note that divalent metal ion was a critical cofactor for MMTV intasome formation. Based on prior reports that Ca^{2+} promoted the assembly of active HIV-1 IN-vDNA complexes but was unable to support IN catalysis[14], it was utilized here for intasome assembly.

To determine the MMTV intasome structure, single-particle cryo-EM data was collected on a Titan Krios microscope equipped with a Gatan K2 direct detector. Computational processing of the data revealed the most stable structural conformation of the complex, which was refined to ~5–6 Å for different regions of the map (Fig. 2a and Extended Data Figs. 1 and 2). The MMTV intasome is composed of central core density, as well as flanking density regions that are conformationally mobile in comparison to the intasome core (Extended Data Fig. 3). Restricting data refinement to the core density region accordingly increased the resolution for the central portion of the structure to ~4 Å for the best-resolved regions (Extended Data Fig. 2d).

Each IN monomer is composed of an N-terminal domain (NTD), CCD, and CTD (Extended Data Fig. 4a), and the map was sufficiently detailed to readily assign these domains to their corresponding cryo-EM densities. Given a lack of MMTV IN structures, the different protein domains were crystallized as IN_{CCD}, IN_{CTD} and $IN_{NTD\text{-}CCD}$ fragments, and these structures were refined to 1.7 Å, 1.5 Å and 2.7 Å resolution, respectively (Extended Data Fig. 5 and Extended Data Table 1). MMTV DNA was modeled using PFV intasome DNA coordinates and by extending the modeled fragment by 3 bp in the region distal from the IN active sites to account for the different vDNA lengths. Using rigid body docking, the two vDNAs and eight NTDs, CCDs and CTDs were unambiguously positioned into the cryo-EM

map (Fig. 2b). Rosetta[15–17] was used to refine the X-ray structures and vDNA, and to build a subset of interdomain linker regions to best fit the density within the intasome core region (Extended Data Fig. 6 and Supplementary Videos 1–5). The resulting model revealed two molecules of vDNA and eight MMTV INs arranged as four IN dimers (Fig. 3a). Two catalytic IN dimers A and B are positioned in the core region in close contact to the vDNAs, whereas supportive IN dimers C and D locate to the flanking density regions, donating their CTDs to the core. Flexible linkers connect the IN domains, and the NTD-CCD linker, which is contracted in the majority of IN protomers, extends in IN_1 and IN_3 to donate these NTDs *in trans* to opposing CCDs (Fig. 3a and Extended Data Fig. 6e). Sedimentation velocity centrifugation indicated the molecular mass of active MMTV intasomes as 302.1 kDa, which is fully consistent with the octameric IN structure (calculated IN_8-$vDNA_2$ = 313.6 kDa; Extended Data Fig. 4b).

The structures of the MMTV and PFV intasomes were compared to ascertain aspects of the new structure important for DNA recombination (Fig. 3a). The PFV intasome is composed of two IN dimers A and B, with the inner protomers of each dimer (IN_1 and IN_3; red and green in Fig. 3a) adopting extended conformations[2]. The NTDs and CTDs of the outer IN protomers (chartreuse and orange in Fig. 3a) are unseen in PFV intasome density maps. The architecture in the core density region of the MMTV intasome is strikingly similar to the PFV structure. For example, the positions of the CCDs and NTDs that contact vDNA (red IN_1 and green IN_3 in Fig. 3a) are analogous. The two remaining NTDs in the core region stabilize the CCD dimer interface in an arrangement identical to that seen in the $IN_{NTD\text{-}CCD}$ crystal structure (Extended Data Figs. 5d and 6e). Both flanking density regions contain a CCD dimer that is also stabilized on each side by NTDs, mimicking the CCD dimer arrangements found in the core density region.

The arrangements of the CTDs differ dramatically between the MMTV and PFV structures, with MMTV IN residue Arg240 mediating several key contacts. For example, core protomer IN_1 Arg240 interacts with vDNA while IN_2 Arg240 interacts with IN_1 Asp233 (Fig. 3b). Flanking protomer IN_5 Arg240 engages its IN_6 neighbor whereas IN_6 Arg240 mediates an inter-dimeric interaction with core protomer IN_1 Asp223, docking the flanking IN dimer to the core structure (Fig. 3b). To test the functionality of the flanking IN dimers, complementation assays were performed by varying ratios of wild-type (IN_{WT}) and mutant IN proteins in strand transfer reactions. Similar strategies were used previously to dissect the division of labor within multimeric complexes of retroviral IN[18–21] as well as the related bacteriophage Mu transpososome[22].

IN_{R240E}, which like IN_{WT} purified as a dimer (Extended Data Fig. 7), was defective for strand transfer activity (Fig. 4a). To assess the functionality of Arg-240-mediated IN-IN interactions, we compared IN_{R240E} to IN_{K165E}, which carries a change that uniquely disrupts IN-vDNA binding[2,23]. In concordance with its inability to assume the roles of inner IN_1 and IN_3 subunits of the core tetramer, IN_{K165E} mildly stimulated the activity of limited IN_{WT} protein (Fig. 4b), presumably providing a source for other IN subunits within the functional complex. IN_{R240E} by contrast potently inhibited IN_{WT} function, confirming the importance of Arg240-mediated protein-protein interactions for MMTV IN activity. Two deletion mutant constructs, $IN_{CCD\text{-}CTD}$ and IN_{CTD}, which purified as dimers and monomers,

respectively (Extended Data Fig. 7), were additionally analyzed. The reaction composed of 75% $IN_{CCD-CTD}$ supported near IN_{WT} activity, indicating that this mutant could function when present in up to 6 of 8 octamer positions. This finding strongly supports flanking IN dimer functionality, as the absence of the NTD would likewise preclude $IN_{CCD-CTD}$ from assuming intasome core positions 1 and 3. As the IN_{CTD} response curve overlaid that predicted for non-functional protein, we moreover conclude that CCD-mediated dimerisation is critical for flanking IN CTD function (Fig. 4).

Analysis of IN primary sequences suggests an explanation for the octameric arrangement of IN within the MMTV intasome when an IN tetramer suffices for PFV integration. Whereas 50-residue CCD-CTD linkers afford the positioning of inner PFV IN CTDs for vDNA and tDNA engagement[2,3], the analogous 8-mer MMTV linker is simply too short to accomplish the task (Extended Data Fig. 8a). MMTV has accordingly evolved to employ flanking IN dimers to nestle CTDs into the core intasome structure to provide essential CTD function *in trans* for integration. As flanking IN dimer CTDs 6 and 8 structurally mimic the PFV domains (Fig. 3a and Extended Data Fig. 8a), we presume these CTDs will engage tDNA during MMTV integration. Extending our analysis to other retroviruses indicates that in addition to the spumaviruses, IN tetramers could suffice for γ- and ε-retroviral intasome activity, while an IN octamer will be required to catalyze α-retrovirus in addition to β-retrovirus integration (Extended Data Fig. 8b). We note that Aihara and colleagues have independently determined an octameric IN architecture for the α-retrovirus Rous sarcoma virus (RSV) intasome[24]. Whereas the majority of studies have highlighted a tetramer as the IN species that catalyzes concerted HIV-1 integration[9,25,26], others have implicated a role for octameric IN[27,28]. Given the intermediary length of lentiviral IN CCD-CTD linker regions (Extended Data Fig. 8b), the higher-order nature of IN in active HIV-1 intasomes may need to be reevaluated.

PFV IN, which cleaves tDNA phosphodiester bonds separated by 4 bp, preferentially integrates into flexible sequences, whereas MMTV and RSV IN, which cleave tDNA with 6 bp staggers, are relatively unconstrained by tDNA flexibility[3,29]. Superposition of the MMTV and PFV intasome structures revealed that the two sets of catalytic IN active sites almost perfectly aligned (Extended Data Fig. 8c). The same practical spacing of IN active sites therefore catalyzes PFV and MMTV integration into sharply bent versus relatively non-deformed tDNA, respectively (Extended Data Fig. 8d). Due to their positions in the structure, we note that the flanking IN dimers dramatically expand the potential contact area with tDNA, which is likely to have consequences for the docking of α- and β-retroviral intasomes to host chromatin.

METHODS

Statistical methods were not used to predetermine sample sizes. Experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

DNA constructs

Full-length (FL) MMTV IN[31] and IN_{CTD} ($IN_{212–266}$ and $IN_{212–319}$) expression constructs provided N-terminal His_6 tags followed by human rhinovirus (HRV) 3C protease cleavage sites. The $IN_{NTD-CCD}$ expression construct was made by introducing a stop codon after the TCA that encodes for IN residue Ser212. IN_{K165E} and IN_{R240E} expression constructs were made by PCR-directed mutagenesis. DNA fragments corresponding to $IN_{51–212}$ (IN_{CCD}) and $IN_{51–319}$ ($IN_{CCD-CTD}$) were amplified by PCR and subcloned into expression vector pET-20b (Novagen); these proteins harbored cleavable C-terminal His_6 tags. The sequences of all PCR amplified regions of plasmid DNAs were verified by sequencing.

Protein expression and purification for intasome and IN activity assays

FL INs, $IN_{CCD-CTD}$ and $IN_{CTD/212–319}$ were expressed in *Escherichia coli* strain PC2 (ref. 32) in LB broth (supplemented with 50 μM $ZnCl_2$ for FL INs) by induction with 0.4 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) (1 mM IPTG for $IN_{CCD-CTD}$) at 30 °C (37 °C for $IN_{CCD-CTD}$ and IN_{CTD}) for 4 h. Bacteria pellets were resuspended in 20 mM HEPES, pH 7.6, 1 M NaCl, 5 mM 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS), complete EDTA-free protease inhibitor (Roche). Following sonication for 5 min at 50 mA, cell lysates were centrifuged at 45,000 × *g* for 1 h. The supernatant, supplemented with 5 mM imidazole, was filtered through a 0.45 μm filter and purified using a Ni^{2+}-charged HisTrap 5 mL column (GE Healthcare) equilibrated with 20 mM HEPES, pH 7.6, 1 M NaCl, 5 mM CHAPS, 15 mM imidazole. Proteins were eluted by a linear gradient of imidazole (15–500 mM) containing a step wash at 65 mM imidazole using the ÄKTA purifier system (GE Healthcare; for $IN_{CCD-CTD}$, a second step wash was done at 115 mM imidazole). IN-containing fractions were diluted 1:5 with 20 mM HEPES, pH 7.6, 5 mM CHAPS, 2 mM dithiothreitol (DTT) and immediately loaded on a Heparin HiTrap 5 mL column equilibrated with 20 mM HEPES, pH 7.6, 200 mM NaCl, 5 mM CHAPS, 2 mM DTT. Proteins were eluted by a linear NaCl gradient from 200 mM to 2 M (IN_{CTD} was isolated in the column flow through). IN-containing fractions were pooled and cleaved with HRV 3C protease (GE Healthcare) overnight at 4 °C to remove the His_6 tag. In lieu of purification by Heparin HiTrap, $IN_{CCD-CTD}$ was dialysed against 20 mM HEPES, pH 7.6, 1 M NaCl, 5 mM CHAPS, 2 mM DTT, 2 mM EDTA at 4 °C for 2 h, cleaved with HRV 3C protease overnight at 4 °C, followed by dialysis against 20 mM HEPES, pH 7.6, 1 M NaCl, 5 mM CHAPS, 2 mM DTT, 0.5 mM EDTA (SEC1 buffer). Cleaved proteins were purified by SEC using a Superdex 200 10/300 column (GE Healthcare) equilibrated with SEC1 buffer. Purified INs were concentrated by ultracentrifugation using 10-kDa molecular weight cutoff Millipore concentrators and dialysed overnight against SEC1 buffer supplemented to contain 10% glycerol. Protein concentration was determined by spectrophotometry, and aliquots flash-frozen in liquid N_2 were stored at −80 °C. Purified INs were analyzed by SEC using a Superdex 3.2/300 column equilibrated with SEC1 buffer; protein standards were from Bio-Rad.

MMTV intasome assembly

Intasomes were assembled by mixing 128 μM MMTV IN with 38 μM 22 bp preprocessed vDNA (5′-CAGGTCGGCCGACTGCGGCA/5′-AATGCCGCAGTCGGCCGACCTG) in 20

mM HEPES, pH 7.6, 600 mM NaCl, 2 mM DTT, prior to dialysis for 16 h at 4 °C against 25 mM Tris-HCl, pH 7.4, 80 mM NaCl, 2 mM DTT, 25 μM $ZnCl_2$, 10 mM $CaCl_2$. The resulting milky white precipitate was dissolved by adding NaCl to the final concentration of 250 mM, followed by incubation on ice for 1 h. Following 10 min centrifugation at 20,000 g at 4 °C, soluble intasomes were purified by SEC using Superdex 200 10/300 equilibrated with 25 mM Tris-HCl, pH 7.4, 200 mM NaCl, 2 mM DTT, 25 μM $ZnCl_2$, 10 mM $CaCl_2$ (SEC2 buffer). Intasome-containing fractions, which eluted around 10.5 mL, were concentrated by ultracentrifugation using 10-kDa cut off concentrators.

In vitro integration assays

Strand transfer assays were performed as described previously[31]. Briefly, 1 μM intasome or 1 μM MMTV IN plus 0.5 μM vDNA were mixed with 300 ng pGEM-3 tDNA in 40 μL of 20 mM HEPES, pH 7.4, 60 mM NaCl, 5 mM $MgCl_2$, 4 μM $ZnSO_4$, 10 mM DTT. Reactions incubated for 1 h at 37 °C were terminated by adding 25 mM EDTA–0.5 % SDS. DNA products deproteinized by digestion with proteinase K and precipitated with ethanol were analyzed by electrophoresis through 1.5% agarose gels and visualized by staining with ethidium bromide. Raltegravir, which was used at the final concentration of 100 μM, was obtained from Selleck Chemicals. Proteins were premixed on ice prior to addition to reactions for biochemical complementation assays. Concerted integration products were measured by band intensity quantification relative to IN_{WT} product formation, which was set to 100% using Molecular Imager® Gel Doc TM XR+ System with Image Lab software (BioRad); the background across 8 gel images corresponded to 1.26% ± 0.47% of IN_{WT} function.

Concerted integration reaction products were cloned and sequenced essentially as previously described[32]. Briefly, DNA excised from agarose gels was repaired using Phi29 DNA polymerase (New England Biolabs) and ligated to a PCR-amplified kanamycin resistance cassette. Plasmids recovered following transformation of ligation mixtures into *E. coli* were sequenced using primers that annealed to the ends of the cassette DNA.

Analytical ultracentrifugation

Sedimentation velocity analysis was performed at 20 °C in a Beckman Optima XL-I analytical ultracentrifuge using an An60Ti rotor and standard 2-channel epon centerpieces (Beckman-Coulter). Samples were prepared in 20 mM phosphate buffer, pH 7.5, 150 mM NaCl at two loading concentrations, A_{280} values of 0.3 and 0.9 for MMTV IN and the intasome, and A_{280} values of 0.18 and 0.53 for vDNA, to exclude potential mass action oligomerization. IN and vDNA were spun simultaneously at 35,000 rpm for 22 h while the intasome was spun at 27,000 rpm for 12 h; the different rotor speeds were based on the predicted masses of the different macromolecules.

Data were analyzed using UltraScan-III version 2.2, release 2000 (ref. 33). Hydrodynamic corrections for buffer density and viscosity were estimated with UltraScan to be 1.041 g/mL and 1.101 cP, respectively. The partial specific volume of IN (0.728 mL/g) was estimated by UltraScan from its protein sequence using a method analogous to the methods outlined in Laue *et al.*[34]. Sedimentation velocity data were analyzed as described[35]. Optimization was

performed by two-dimensional spectral analysis (2DSA)[36] with simultaneous removal of time-invariant and radially-invariant noise contributions[37]. 2DSA solutions, which are subjected to parsimonious regularization by genetic algorithm analysis[38], were further refined using Monte Carlo analysis to determine confidence limits for the determined parameters[39]. Calculations were performed on the Lonestar cluster at the Texas Advanced Computing Center at the University of Texas at Austin.

Protein expression and X-ray crystallography

MMTV IN_{CCD}, $IN_{NTD\text{-}CCD}$ and IN_{CTD} fragments spanning MMTV IN residues 51–212, 1–212 and 212–266, respectively, were expressed in BL21(DE3)-CodonPlus cells (Stratagene) in LB medium (supplemented with 50 μM $ZnCl_2$ for $IN_{NTD\text{-}CCD}$) by induction with 0.01% (w/v) IPTG. Bacteria were lysed by sonication in 0.5 M NaCl, 50 mM Tris-HCl, pH 7.4 and the proteins were isolated by absorption to Ni-nitrilotriacetic acid agarose (Qiagen). Following digestion with HRV 3C protease to release His_6 tags, the proteins were further purified by ion exchange and size exclusion chromatography.

Crystals were grown by vapor diffusion in hanging drops by mixing 1 μL protein (6–10 mg/mL in 200 mM NaCl, 2 mM DTT, 25 mM Tris-HCl, pH 7.5) and 1 μL reservoir solution, which contained 12.5% PEG-3350, 0.15 M ammonium citrate, pH 6.5 (IN_{CCD}), 19% PEG-3350, 0.2 M $MgCl_2$, 5% (v/w) 1-butyl-3-methylimidazolium dicyanamide ($IN_{NTD\text{-}CCD}$), or 19% isopropanol, 50 mM ammonium acetate, 0.1 M HEPES-NaOH, pH 7.5 (IN_{CTD}). Crystals, cryoprotected with 25% glycerol (IN_{CCD}, $IN_{NTD\text{-}CCD}$) or 30% PEG-400 (IN_{CTD}), were frozen by immersion in liquid N_2. Diffraction data for the IN_{CCD} was collected using a charged coupled device detector at beamline BM14 (European Synchrotron Radiation Facility [ESRF]) whereas IN_{CTD} and $IN_{NTD\text{-}CCD}$ crystals were analyzed at beamline I03 (Diamond Light Source) equipped with a PILATUS direct detector. The data integrated with XDS[40] were scaled with Aimless[41]. The structures, which were each derived from a single crystal, were solved by molecular replacement in Phaser[42] using search models generated from PDB entries 1ASV (CCD)[43], 3F9K (NTD)[10] and 1EX4 (CTD)[44]. The models were rebuilt using ARP/wARP[45] and/or manually in Coot[46] and refined in Phenix[47] and/or Refmac[48]. Pseudo-merohedral twin law (-h,-k,l) was accounted for during refinement of the $IN_{NTD\text{-}CCD}$ structure. Final models, validated with MolProbity[49], had at least 96.9% of residues in the favored regions and none in the disallowed regions of the Ramachandran plot. Detailed X-ray data collection and refinement statistics are given in Extended Data Table 1.

Cryo-EM data acquisition

Sample containing MMTV intasomes in SEC2 buffer supplemented to contain 0.05% NP-40 was applied onto freshly plasma treated (6s, Gatan Solarus plasma cleaner) holey carbon C-flat grids (CF-1.2/1.3–4C, Protochips, Inc.), adsorbed for 30 sec and then plunged into liquid ethane using a manual cryo-plunger in an ambient environment of 4 °C.

Data were acquired over three separate sessions using Leginon software[50] installed on an FEI Titan Krios electron microscope operating at 300 kV, with a dose of 40 e-/$Å^2$ at the rate of ~6.9 e-/pix/sec and an estimated underfocus ranging from 1–4 μm (centered at 2.6 ± 0.6

μm). The dose was fractionated over 50 raw frames collected over 10-sec exposure time (200 ms per frame) on the Gatan K2 Summit direct detection device, with each frame receiving a dose of ~6.5 e^-/pixel/sec. 2714 movies were collected and recorded at a nominal magnification of 22,500, corresponding to a pixel size of 1.31 Å at the specimen level. The individual frames were gain corrected, aligned and summed using a GPU-enabled whole frame alignment program as previously described[51,52], and exposure filtered[53] according to the 6.9 e-/pix/sec dose rate. See Extended Data Table 2 for additional details on cryo-EM data collection.

Cryo-EM image analysis

Pre-processing operations prior to the refinement of the final models were performed using the Appion package[54] and are conceptually identical to those previously described[52]. Briefly, single intasome particles (244,315) were selected from the aligned and summed micrographs, from which 147,850 were used to create an initial raw particle stack after removing regions of the micrographs containing carbon and large areas of aggregation. 2D alignments and classifications were performed using the CL2D[55] and Relion[56] algorithms (Extended Data Fig. 1c), and an initial model was generated directly from the class averages using OptiMod[57] (Extended Data Fig. 1d). Following iterative rounds of 2D alignment and classification, 77,365 particles remained for 3D refinement and classification. 3D refinements and classifications were initially performed within Relion[56], after which the parameters were converted for use in Frealign[58]. The final map was refined in Frealign.

Several conformational states of the intasome were observed following 3D classification in both Relion and Frealign[59]. Whereas one of the resulting maps yielded the stable intasome structure from 41,475 particles (Fig. 2a, Extended Data Fig. 2c and Extended Data Table 2), all other maps (one of which is displayed in Extended Data Fig 3b) displayed mobility in the flanking regions, which did not resolve by further classifying the data. To improve the resolution of the core region, we ran Relion and recovered four models in the classification. For each of the resulting maps, the flanking regions were segmented and treated with a soft-edged mask that adopted the shape of the remaining density. Subsequently, for each raw particle, the flanking region from the respective conformational state to which that particle belonged was computationally subtracted from the raw particle image. The contrast transfer function was included in the computational subtraction process. In this manner, datasets lacking the majority of the flanking INs were created. Refinement of the core intasome dataset was then conducted using the likelihood-based approach in Frealign[59], effectively a focused classification of the core region. The best class was resolved to ~4 Å resolution in the most homogeneous regions using 30,307 particles (Extended Data Fig. 2d and Extended Data Table 2). Although slight ghost images remained for the flanking regions within certain particles, they did not dramatically affect the refinement; the use of a tighter mask facilitated the recovery of higher resolution information.

Assembly of the atomic model

Models of the core intasome and the full octamer structures were built and refined in a stepwise manner using Rosetta[15] starting with rigid-body fitted X-ray structures of individual domains as input. Rosetta protocols were used for all parts of the modeling[60]. To

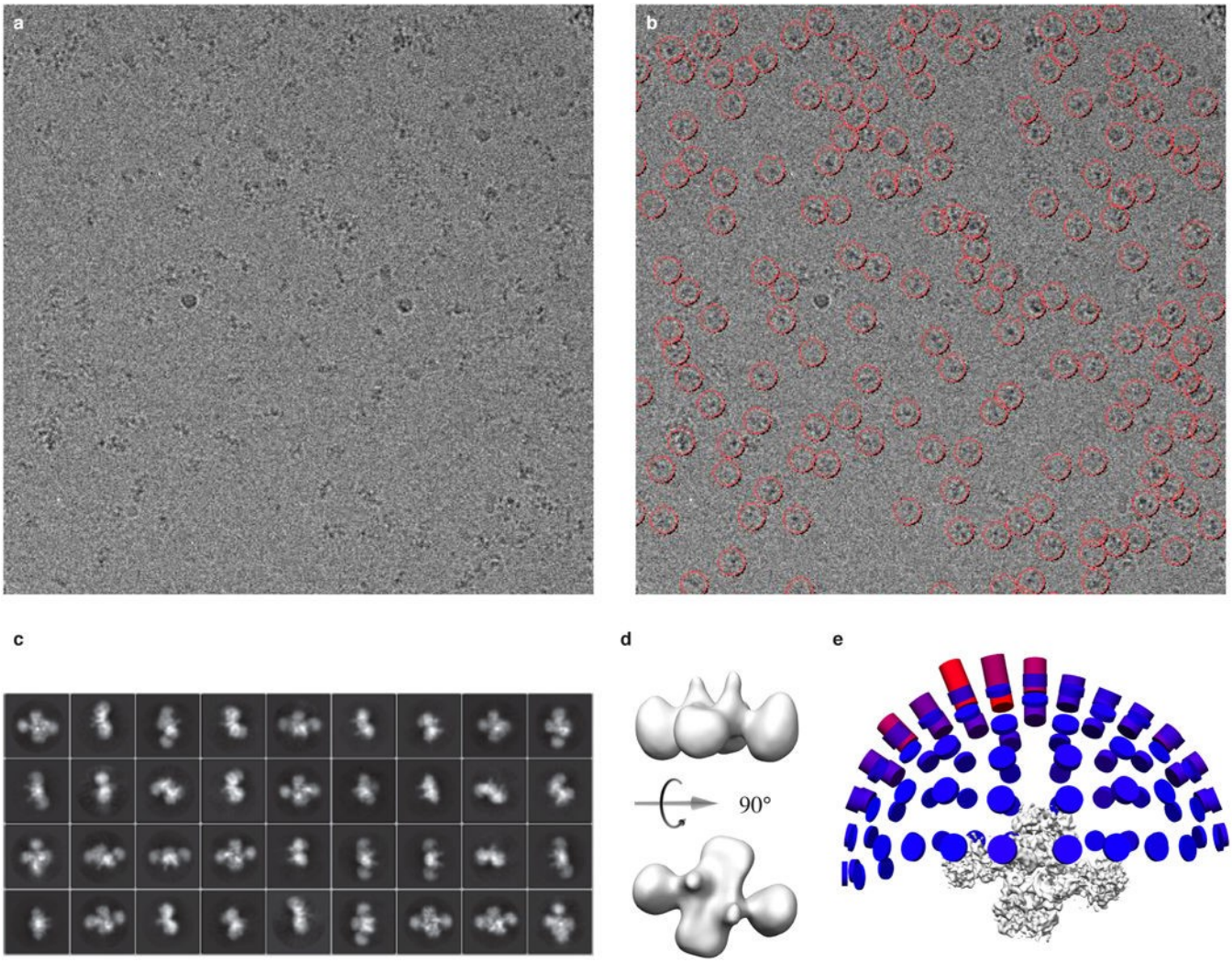
optimally fit X-ray models into the EM density, we first independently refined each individual domain (NTD, CCD and CTD) using multiple input starting seeds. CCD_1 and CCD_2 were each seeded with 6 starting X-ray models: independent CCD monomers from chains A-D of the IN_{CCD} structure and monomers A-B of the CCD portions of the $IN_{NTD\text{-}CCD}$ structures. CTDs 1, 2, 5 and 6 were seeded with subunits A and B of the IN_{CTD} X-ray model. Likewise, for NTD_1 and NTD_3, the two different NTDs of the $IN_{NTD\text{-}CCD}$ X-ray structure were used as input seeds. All models were refined against the core intasome structure resolved to ~4–5 Å resolution (Extended Data Fig. 2d). At least 2000 models were generated from each and the lowest energy model was selected for moving forward. Modeling quality was assessed by energy scores, structural similarity of the top scoring models, and visual inspection (Extended Data Fig. 6a). We next proceeded to independently model IN_1, IN_2, IN_5 and IN_6, thereby filling in the linker regions between individual domains using 7-mers from the PDB[15]. This enabled *de novo* modeling for linker residues 45–54 between NTD_1-CCD_1 and residues 211–213 between CCD_1-CTD_1 and CCD_2-CTD_2 (some residues, as well as outlier linker regions, were not modeled due to disorder; Extended Data Fig. 6b and 6c); modeling was facilitated by the presence of 'bumps' within the density that corresponded to bulky amino acid side chains, in particular within NTD_1-CCD_1, which is located in the best resolved region of the structure (Extended Data Fig. 2d). IN_1 and IN_2 were each seeded with combinations of the best models arising from refinement of individual domains and were subsequently refined against the core intasome density map. 2000 models were again generated for each, and the best were selected to move forward. This set of procedures produced FL models for IN_1 and IN_2 and models for CTD_5 and CTD_6 fitted to the EM protein density. MMTV DNA was modeled based on the X-ray structure of the PFV intasome (PDB 3L2Q). This model was rigid-body docked into the EM density and then relaxed with Rosetta. The complete intasome model was iteratively relaxed with Rosetta and then adjusted manually using Coot[46]. Several iterative rounds of refinement and inspection were performed using MolProbity[49] at the end of each round until a consensus model was obtained (Extended Data Fig. 6c,d and Extended Data Table 2).

IN linker regions

Linker lengths for Extended Data Figure 8b were assessed by aligning published[30] or in-house generated IN sequence alignments against alignments based on known domain structures[2] (Extended Data Fig. 4a). The following sequences were included: γ-retroviruses, Moloney murine leukemia virus (Genbank accession number J02255.1), reticuloendotheliosis virus strain A (DQ237900.1), feline leukemia virus (NC_001940.1); ε-retroviruses, walleye dermal sarcoma virus (NC_001867.1), walleye epidermal hyperplasia virus types 1 and 2 (AF133051.1 and AF133051.2, respectively); spumaviruses, PFV (U21247.1), macaque simian foamy virus (NC_010819.1), spider monkey foamy virus (EU010385.1); lentiviruses, HIV-1 strain NL4–3 (U26942.1), HIV-2 strain ROD (X05291.1), simian immunodeficiency virus strain agm.tan-1 (U58991.1), equine infectious anemia virus (M16575.1), feline immunodeficiency virus (M25381.1), caprine arthritis encephalitis virus (M33677.1), bovine immunodeficiency virus (NC_001413.1); δ-retroviruses, bovine leukemia virus (K02120.1), human T-cell lymphotropic virus types 1 and 2 (NC_001436.1 and NC_001488.1, respectively); β-retroviruses, MMTV (NC_001503.1), Mason Pfizer monkey virus (NC_001550.1), Jaagsiekte sheep retrovirus

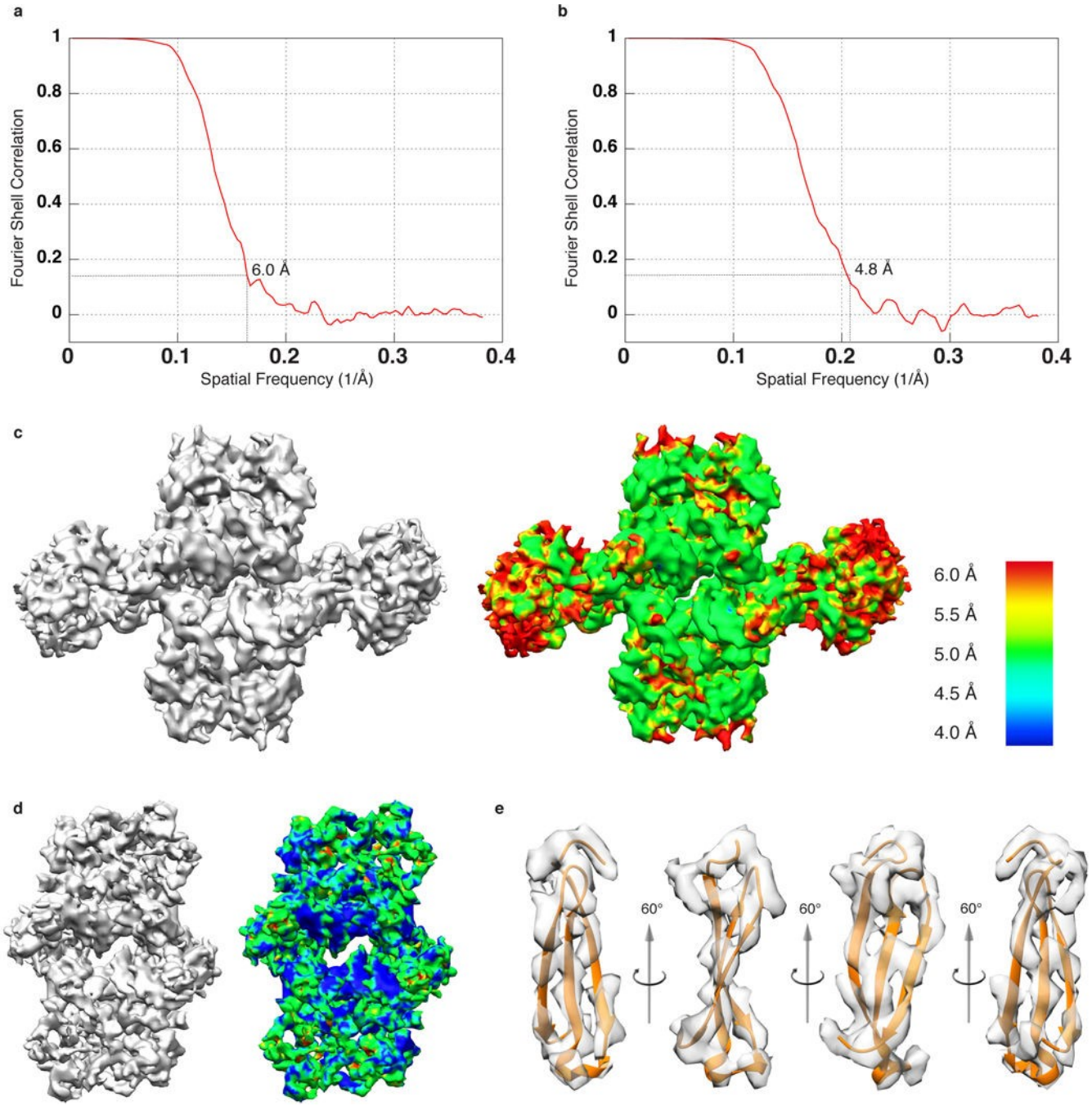
(NC_001494.1); α-retroviruses, RSV (J02342.1), lymphoproliferative disease virus (KC802224.1).

Extended Data



Extended Data Figure 1. Cryo-EM data and refinement

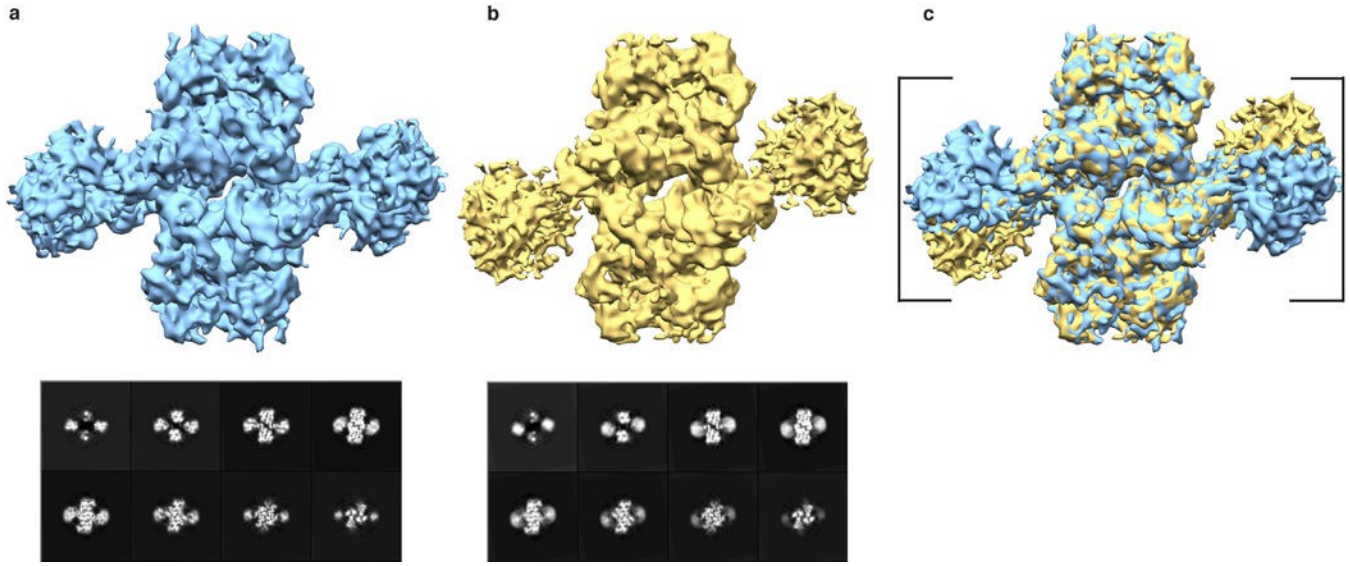
a, Representative cryo-electron micrograph of MMTV intasomes, taken at 2.7 μm underfocus. **b**, Same as in panel a, marked to show selected particles. **c**, 2D class averages calculated using Relion[56]. **d**, Initial model from the class averages calculated using OptiMod[57]. **e**, Refined reconstruction from the full dataset, with an Euler angle distribution plot showing the relative orientations of the particles.



Extended Data Figure 2. Cryo-EM resolution analysis of reconstructed intasome maps

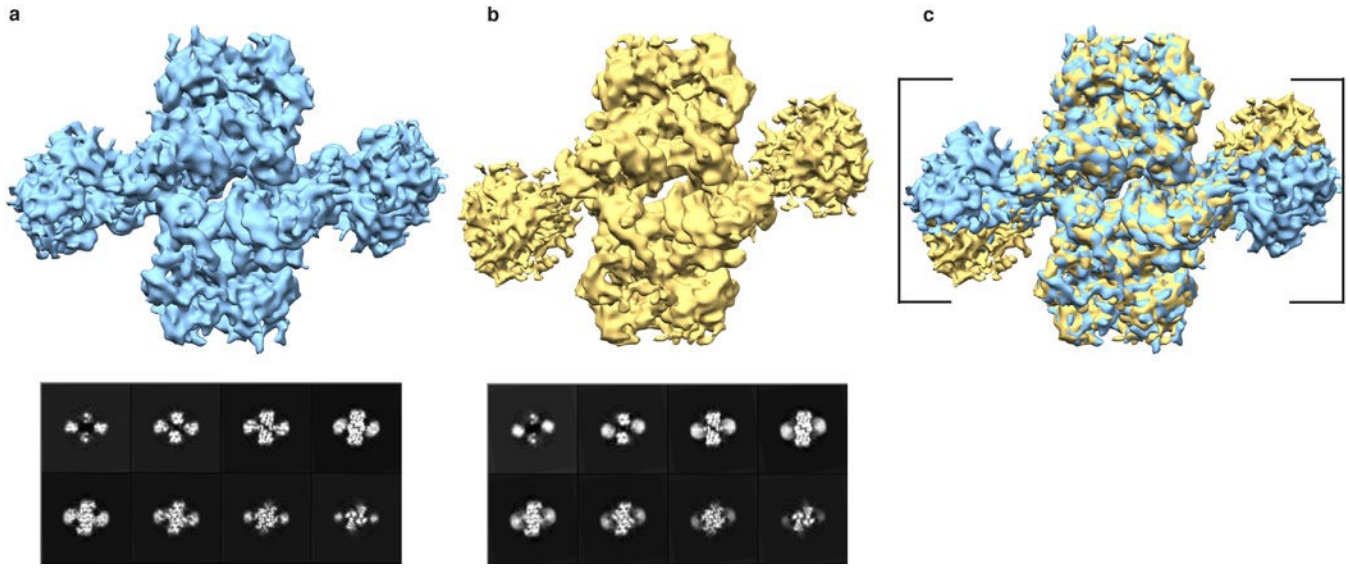
a, Fourier shell correlation (FSC) curve corresponding to the refined map generated from the full intasome dataset. **b**, FSC curve corresponding to the refined map generated from the core intasome dataset with the NTDs, CCDs and interdomain linker regions of the flanking IN dimers computationally subtracted. Average global resolutions in panels a and b are indicated. **c**, Refined map generated from the full dataset (left) displayed side-by-side with the same map colored for local resolution (right). **d**, Refined map generated from the core intasome dataset (left) displayed side-by-side with the same map colored for local resolution

(right) using the coloring scheme in panel c. **e**, Rotational snapshots of segmented density of CCD_1 with the fit of the refined model (see Extended Data Fig. 6) highlighting structural features evident at ~4–5 Å resolution. Partial separation of β-strands, which is typically evident at or beyond 4.5 Å resolution, is apparent.



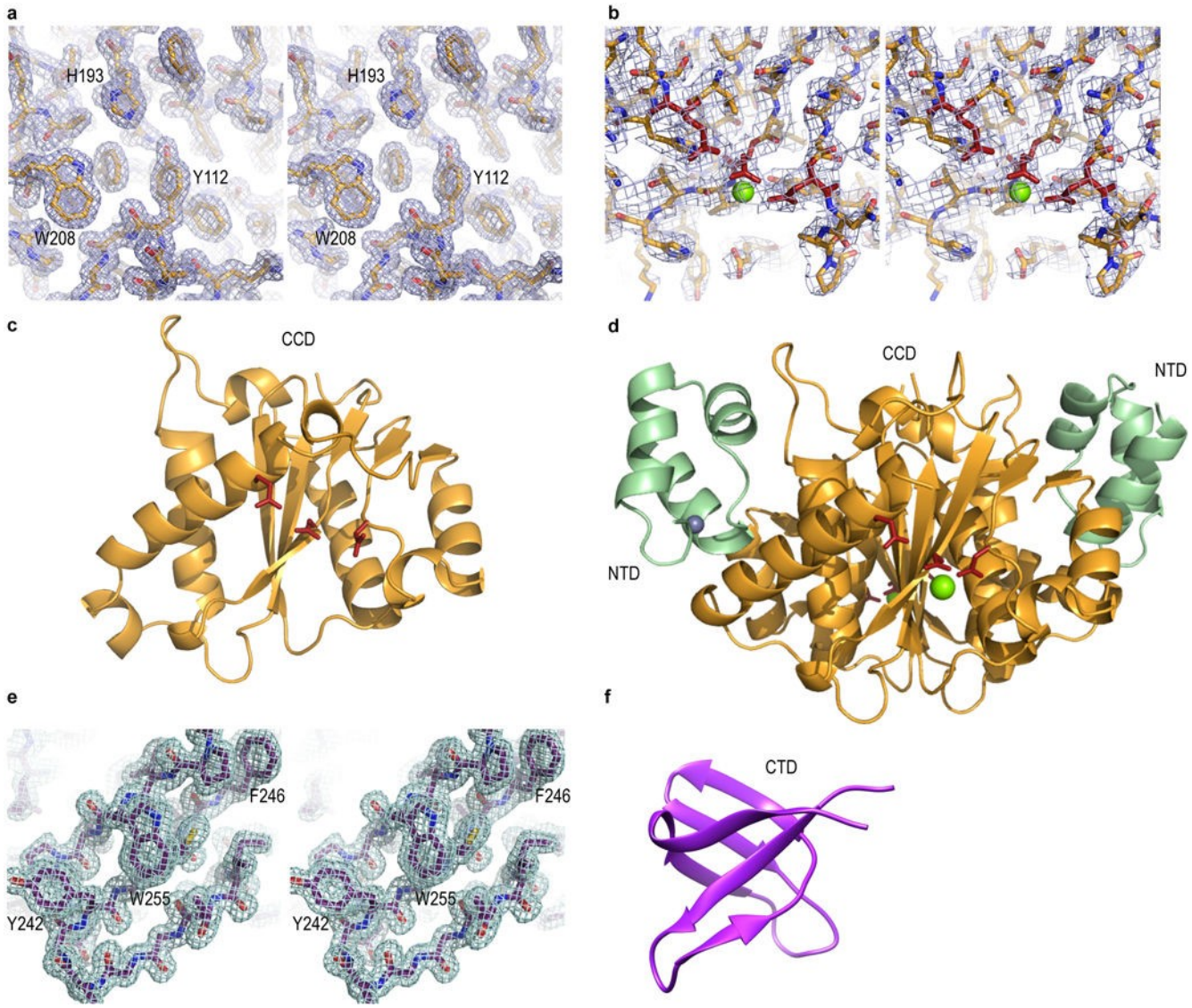
Extended Data Figure 3. Structural heterogeneity of the MMTV intasome

a, Stable structural conformation of the MMTV intasome after 3D classification of the data. Slices from the density map are displayed below. **b**, One of several conformations of MMTV intasome refinement after 3D classification of the data. Slices from the density map are displayed below. Multiple fuzzy regions in the flanking INs are apparent in b, which are indicative of remaining heterogeneity within the data and/or continuous structural mobility of the region. **c**, Overlay of the two reconstructed maps, highlighting the extent of mobility within the flanking regions (brackets).



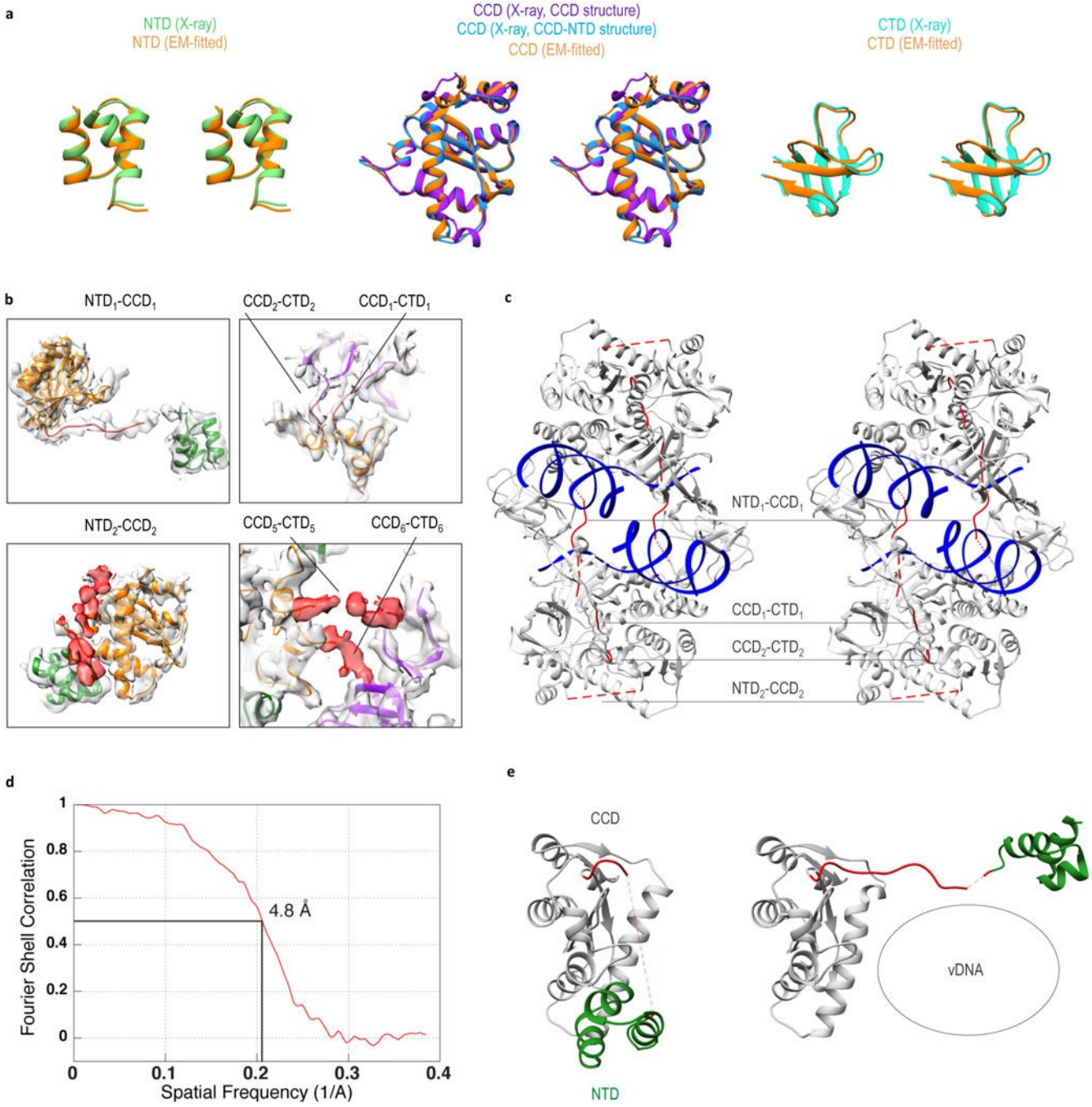
Extended Data Figure 4. MMTV IN domains and intasome sedimentation coefficient distribution

a, Primary IN sequence alignment with boxes denoting canonical IN structural domains. The N-terminal extension domain (NED) occurs in spuma-, γ- and ε-retroviral IN proteins. Identical residues between MMTV, RSV, HIV-1 and PFV INs are highlighted by red background; residues that are minimally conserved in three of the sequences are in red. PFV IN secondary structure elements are from PDB code 3L2Q; MMTV elements are from the $IN_{NTD\text{-}CCD}$ and IN_{CTD} crystal structures described here (PDB codes 5CZ2 and 5D7U, respectively). α, β, η, TT and TTT represent α helix, β strand, 3_{10} helix, α-turn and β-turn, respectively. Figure generated with ESPript 3.0 (ref. 61). **b**, Monte Carlo analysis of sedimentation velocity data for the higher loading concentrations of vDNA (green), MMTV IN (blue) and intasome (red). A clear shift to a discrete species at 10.5 s is observed for the intasome, with minor IN and vDNA populations evident. Different centrifugation parameters for IN and vDNA versus intasomes (see Methods) likely attributed to the minor variations in sedimentation coefficient between major and minor IN and vDNA species. Measured sedimentation coefficients and calculated molar masses compared to theoretical molar masses are shown beneath the graph.



Extended Data Figure 5. MMTV IN domain crystal structures

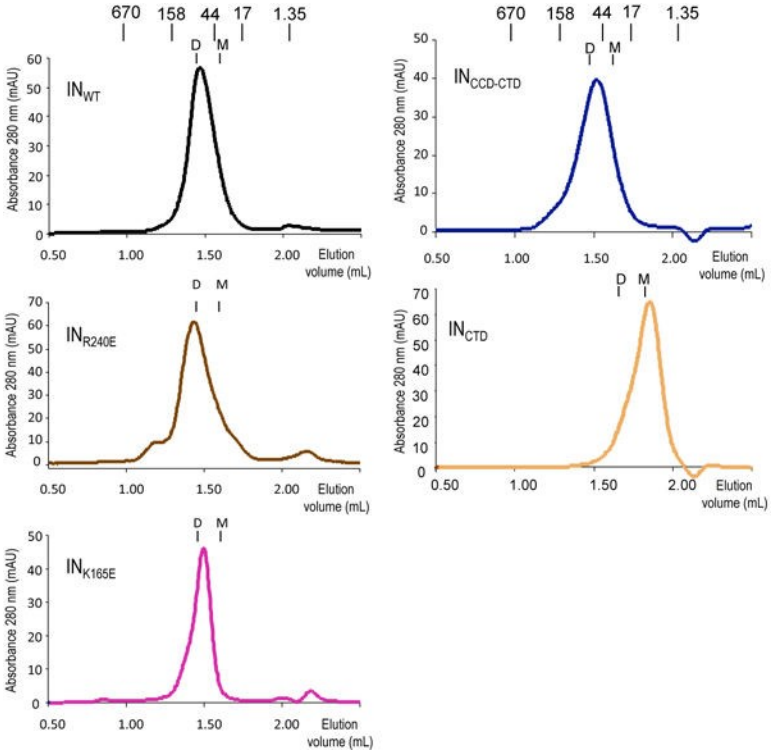
a, Stereo view of the final *2Fo-Fc* density map of the IN_{CCD} crystal structure with blue mesh contoured at 1σ. Amino acid side chains are readily evident at the 1.7 Å resolution. **b**, Stereo view of the final *2Fo-Fc* density map of the 2.7 Å resolution $IN_{NTD-CCD}$ crystal structure with blue mesh contoured at 1σ. The map is centered on the DDE catalytic triad (red sticks); green spheres, Mg^{2+} ions. **c**, Cartoon representation of the IN_{CCD} monomer (one of 4 in the crystallographic asymmetric unit) colored in gold. Active site residues are shown as red sticks. **d**, Cartoon representation of the $IN_{NTD-CCD}$ dimer structure (one of 3 in the asymmetric unit). The NTD and CCD are colored green and gold, respectively. Red sticks, active site residues; grey and green spheres, Zn^{2+} and Mg^{2+} ions, respectively. **e**, Stereo view of the final *2Fo-Fc* density map of the 1.5 Å resolution IN_{CTD} crystal structure, shown as a green mesh contoured at 1σ. **f**, Cartoon representation of one of the two CTD monomers present in the asymmetric unit.



Extended Data Figure 6. Molecular modeling of cryo-EM density

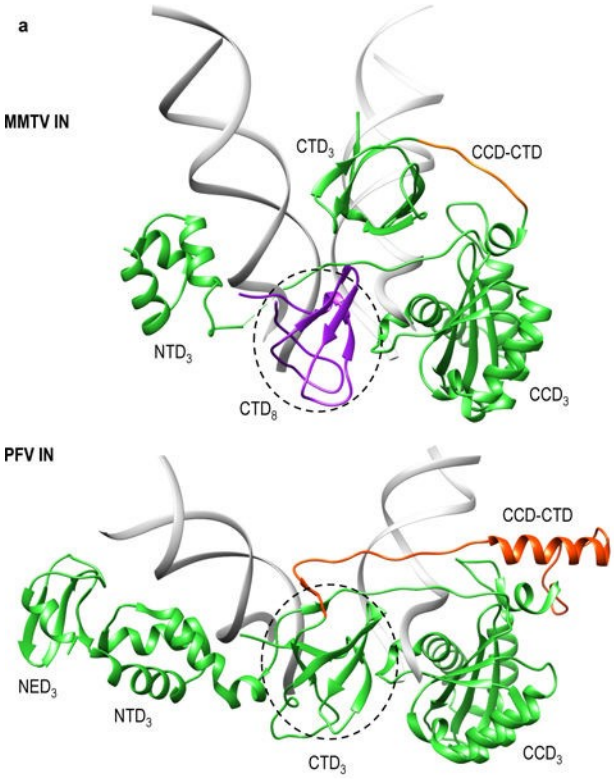
a, Stereo views showing comparisons between the starting X-ray domain models and refined cryo-EM domain models for IN_1 highlight relatively minor structural perturbations that are evident only in the most flexible regions of the intasome. **b**, Linker region snapshots. Atomic models were built *de novo* from the cryo-EM density for the indicated linkers in the top two panels (residues 45–54 connecting NTD_1 and CCD_1 and CCD-CTD residues 211–213). Linkers NTD_2-CCD_2, CCD_5-CTD_5 and CCD_6-CTD_6 were not modeled, but are shown as cryo-EM density (red) in the lower panels. **c**, Stereo view of the cryo-EM model for the

MMTV intasome core region (Extended Data Fig. 2d), generated using Rosetta[15–17]. All domains were refined starting with the X-ray crystal structures (Extended Data Fig. 5). Specific linker regions were built *de novo* (continuous red lines) from the cryo-EM density, whereas lower resolution linker regions (red dotted lines) were omitted from the model. **d**, FSC curve between the refined cryo-EM core intasome model and map, showing an average resolution of 4.8 Å. **e**, Comparison of two NTD-CCD conformations in the intasome highlights the NTD-CCD linker, which assumes a retracted state in the outer IN_2 and IN_4 monomers of core intasome dimers A and B, respectively, as well as in flanking IN dimers C and D (left). The linker extends in core IN molecules IN_1 and IN_3, which interact with the vDNA (right).



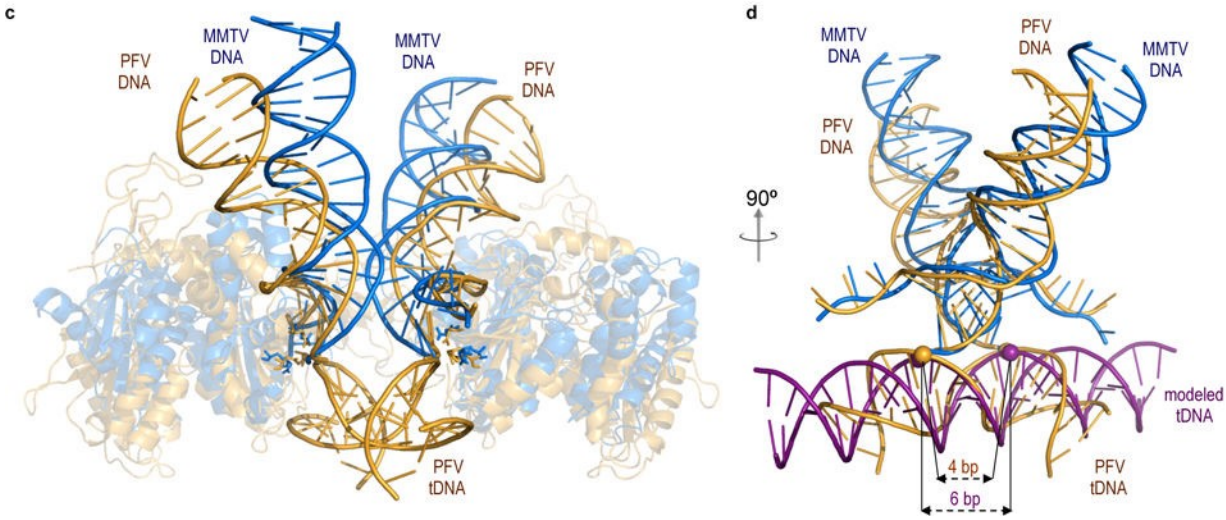
Extended Data Figure 7. Gel filtration profiles of IN_{WT} and IN mutant proteins

Migration positions of mass standards in kDa as well as theoretical protein monomer (M) and dimer (D) positions are indicated.



b

Genus	NED	NTD-CCD	CCD-CTD	IN multimer
Gamma	+	19-21	61	Tetramer
Epsilon	+	19-21	55	Tetramer
Spuma	+	18	50	**Tetramer**
Lenti	-	11-12	20-22	?
Delta	-	18	19	?
Beta	-	14	8	**Octamer**
Alpha	-	14	8	**Octamer**



Extended Data Figure 8. Comparisons of PFV and MMTV intasome structures

a, Cartoon representations of the inner IN_3 green subunits of the MMTV and PFV intasomes (Fig. 3a; vDNA strands are in grey). CCD-CTD linker regions are highlighted in orange, and dashed lines circle analogously positioned CTDs. Of note, this CTD in the MMTV structure is colored differently because it originates from a separate IN molecule (IN_8 from flanking dimer D). **b**, Lengths of NTD-CCD and CCD-CTD interdomain linker regions across retroviral IN proteins; '+' indicates the presence of an NED. The multimeric state of IN in known intasome structures is indicated by bold type. **c**, The PFV intasome with bound tDNA

(PDB code 3OS2; orange) was superimposed with the MMTV intasome (blue). The distance between overlaid active sites is in each case ~26 Å. **d**, 90° rotation of superimposed structures, with proteins omitted for clarity. Canonical B-form tDNA (magenta) was superimposed with PFV intasome tDNA. The positions of phosphodiester bonds staggered by 4 bp in the PFV crystal structure or by 6 bp in the modeled tDNA are indicated by spheres.

Extended Data Table 1

X-ray crystallography data collection and refinement statistics.

Construct	CCD	NTD-CCD	CTD
Data collection			
Space group	P1	$P12_11$	$C222_1$
Cell dimensions			
a, *b*, *c* (Å)	51.89, 53.71,69.65	54.37, 83.15, 141.14	35.99,42.28, 139.09
a, *b*, *g* (°)	69.69, 82.08, 63.97	90,90.19,90	90, 90, 90
Resolution (Å) *	46.6 – 1.70(1.73–1.70)	70.6 – 2.72 (2.79 – 2.72)	40.4 – 1.50(1.53–1.50)
R_{merqe}	0.060 (0.57)	0.08 (0.534)	0.043 (0.585)
Ilsl	21.0(2.0)	9.5(2.0)	29.2 (3.8)
Completeness (%)	99.1 (95.6)	99.3 (99.0)	99.8 (99.9)
Redundancy	5.2 (2.8)	3.2(3.1)	12.2 (8.9)
Refinement			
Resolution (Å)	32.8 – 1.70	70.6 – 2.72	40.4 – 1.50
No. reflections used	69,075	32,115	17,448
R_{work}/R_{free}	0.189/0.222	0.245/0.266	0.165/0.202
No. atoms			
Protein	4,983	9,110	890
Ligand/ion	0	12	8
Water	437	0	69
B-factors			
Protein	26.0	70.9	28.5
Ligand/ion	–	45.6	46.4
Water	33.5	–	46.9
R.m.s deviations			
Bond lengths (Å)	0.007	0.010	0.005
Bond angles (°)	0.954	1.281	0.911

* Data for the highest resolution shells are given in parenthesis.

Extended Data Table 2

Cryo-EM data statistics.

Construct	core MMTV intasome	full MMTV intasome
EM data collection/processing		
Microscope	Titan Krios	Titan Krios
Voltage	300	300

Construct	core MMTV intasome	full MMTV intasome
Camera	Gatan K2 Summit	Gatan K2 Summit
Defocus range (μm)	1.0–4.0	1.0–4.0
Defocus mean ± std (μm)	2.6 ± 0.6	2.6 ± 0.6
Exposure time (s)	10	10
Dose rate (e-/pixel/s)	6.9	6.9
Total dose (e-/Å^2)	40	40
Pixel size (Å)	1.31	1.31
Number of micrographs	2,714	2,714
Number of particles (processed)	147,850	147,850
Number of particles (refined)	77,365	77,365
Number of particles (in final map)	30,307	41,475
Symmetry	C2	C2
Resolution (global) (Å)*	4.8	6.0
Resolution range (local) (Å)	4–5	5–6
Map sharpening B-factor (Å^2)	−300	−460
Model refinement		
Space group	P1	–
Cell dimensions		
a = *b* = *c* (Å)	151.2	–
a=*b* = *g* (°)	90	–
Number of atoms (modeled)	11,462	–
Validation		
MolProbity score	1.46 (96th percentile)	–
Clashscore, all atoms	2.27 (99th percentile)	–
Protein		
Ramachandran favored (%)	1,115(92.76)	–
allowed (%)	87 (7.24)	–
Disallowed (%)	0(0)	–
Good rotamers (%)	1,035(99.71)	–
Cβ deviations >0.25Å (%)	0(0)	–
Cis Prolines (%)	8/88(9.09)	–
Bad bonds (%)	2/10,140(0.02)	–
Bad angles (%)	3/13,810(0.02)	–
DNA		
Bad bonds (%)	0/1,834(0)	–
Bad angles (%)	1/2,822 (0.04)	–
r.m.s. deviations		
Bond lengths (Å)	0.012	–
Bond angles (°)	1.334	–

* Resolution assessment based on frequency-limited refinement using the 0.143-threshold for resolution analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge support from US National Institutes of Health (NIH) grants R01 AI070042 (to A.N.E.), P50 GM103368 (to D.L.), P50 GM082251 (to P.C.) and P30 AI060354 (Harvard University Center for AIDS Research), Leona M. and Harry B. Helmsley Charitable Trust grant #2012-PG-MED002 (to D.L.) and US National Science Foundation grants NSF-ACI-1339649 and TG-MCB070039 (to B.D.). B.D. acknowledges support from San Antonio Cancer Institute grant CA054174 for the Center for Analytical Ultracentrifugation of Macromolecular Assemblies at the University of Texas Health Science Center at San Antonio. Molecular graphics and analyses were performed with the USCF Chimera package (supported by NIH P41 GM103331). We thank Bill Anderson and Jean-Christophe Ducom at The Scripps Research Institute for help with EM data collection and network infrastructure, James Fitzpatrick and Frank Dwyer for computational support at The Salk Institute, Dr. V. Pye for help with X-ray structure refinement and the staff of BM14 (European Synchrotron Radiation Facility, Grenoble, France) and I03 (Diamond Light Source, Oxfordshire, UK) beamlines for assistance with data collection.

References

1. Craigie R, Bushman FD. HIV DNA integration. Cold Spring Harb Perspect Med. 2012; 2:a006890. [PubMed: 22762018]
2. Hare S, Gupta SS, Valkov E, Engelman A, Cherepanov P. Retroviral intasome assembly and inhibition of DNA strand transfer. Nature. 2010; 464:232–236. [PubMed: 20118915]
3. Maertens GN, Hare S, Cherepanov P. The mechanism of retroviral integration from X-ray structures of its key intermediates. Nature. 2010; 468:326–329. [PubMed: 21068843]
4. Hare S, Maertens GN, Cherepanov P. 3′-processing and strand transfer catalysed by retroviral integrase in crystallo. EMBO J. 2012; 31:3020–3028. [PubMed: 22580823]
5. Maskell DP, et al. Structural basis for retroviral integration into nucleosomes. Nature. 2015; 523:366–369. [PubMed: 26061770]
6. Yang ZN, Mueser TC, Bushman FD, Hyde CC. Crystal structure of an active two-domain derivative of Rous sarcoma virus integrase. J Mol Biol. 2000; 296:535–548. [PubMed: 10669607]
7. Wang JY, Ling H, Yang W, Craigie R. Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. EMBO J. 2001; 20:7333–7343. [PubMed: 11743009]
8. Bao KK, et al. Functional oligomeric state of avian sarcoma virus integrase. J Biol Chem. 2003; 278:1323–1327. [PubMed: 12446721]
9. Li M, Mizuuchi M, Burke TR Jr, Craigie R. Retroviral DNA integration: reaction pathway and critical intermediates. EMBO J. 2006; 25:1295–1304. [PubMed: 16482214]
10. Hare S, et al. A novel co-crystal structure affords the design of gain-of-function lentiviral integrase mutants in the presence of modified PSIP1/LEDGF/p75. PLoS Pathog. 2009; 5:e1000259. [PubMed: 19132083]
11. Majors JE, Varmus HE. Nucleotide sequences at host-proviral junctions for mouse mammary tumour virus. Nature. 1981; 289:253–258. [PubMed: 6256658]
12. Roth MJ, Schwartzberg PL, Goff SP. Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. Cell. 1989; 58:47–54. [PubMed: 2546673]
13. Craigie R, Fujiwara T, Bushman F. The IN protein of Moloney murine leukemia virus processes the viral DNA ends and accomplishes their integration in vitro. Cell. 1990; 62:829–837. [PubMed: 2167180]
14. Ellison V, Brown PO. A stable complex between integrase and viral DNA ends mediates human immunodeficiency virus integration in vitro. Proc Natl Acad Sci USA. 1994; 91:7316–7320. [PubMed: 8041787]
15. DiMaio F, et al. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. Nat Methods. 2015; 12:361–365. [PubMed: 25707030]

16. Kudryashev M, et al. Structure of the type VI secretion system contractile sheath. Cell. 2015; 160:952–962. [PubMed: 25723169]
17. Wang RY, et al. De novo protein structure determination from near-atomic-resolution cryo-EM maps. Nat Methods. 2015; 12:335–338. [PubMed: 25707029]
18. van Gent DC, Vink C, Groeneger AA, Plasterk RH. Complementation between HIV integrase proteins mutated in different domains. EMBO J. 1993; 12:3261–3267. [PubMed: 8344263]
19. Engelman A, Bushman FD, Craigie R. Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. EMBO J. 1993; 12:3269–3275. [PubMed: 8344264]
20. Yang F, Roth MJ. Assembly and catalysis of concerted two-end integration events by Moloney murine leukemia virus integrase. J Virol. 2001; 75:9561–9570. [PubMed: 11559787]
21. Diamond TL, Bushman FD. Division of labor within human immunodeficiency virus integrase complexes: Determinants of catalysis and target DNA capture. J Virol. 2005; 79:15376–15387. [PubMed: 16306609]
22. Baker TA, Mizuuchi M, Savilahti H, Mizuuchi K. Division of labor among monomers within the Mu transposase tetramer. Cell. 1993; 74:723–733. [PubMed: 8395353]
23. Jenkins TM, Esposito D, Engelman A, Craigie R. Critical contacts between HIV-1 integrase and viral DNA identified by structure-based analysis and photo-crosslinking. EMBO J. 1997; 16:6849–6859. [PubMed: 9362498]
24. Yin Z, Shi K, Banerjee S, Grandgenett DP, Aihara H. Crystal structure of the Rous sarcoma virus intasome. Nature. 2015
25. Faure A, et al. HIV-1 integrase crosslinked oligomers are active in vitro. Nucleic Acids Res. 2005; 33:977–986. [PubMed: 15718297]
26. Bera S, Pandey KK, Vora AC, Grandgenett DP. Molecular interactions between HIV-1 integrase and the two viral DNA ends within the synaptic complex that mediates concerted integration. J Mol Biol. 2009; 389:183–198. [PubMed: 19362096]
27. Lee SP, Xiao J, Knutson JR, Lewis MS, Han MK. Zn2+ promotes the self-association of human immunodeficiency virus type-1 integrase in vitro. Biochemistry. 1997; 36:173–180. [PubMed: 8993331]
28. Heuer TS, Brown PO. Photo-cross-linking studies suggest a model for the architecture of an active human immunodeficiency virus type 1 integrase-DNA complex. Biochemistry. 1998; 37:6667–6678. [PubMed: 9578550]
29. Serrao E, Ballandras-Colas A, Cherepanov P, Maertens GN, Engelman AN. Key determinants of target DNA recognition by retroviral intasomes. Retrovirology. 2015; 12:39. [PubMed: 25924943]
30. Engelman A, Craigie R. Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function in vitro. J Virol. 1992; 66:6361–6369. [PubMed: 1404595]
31. Ballandras-Colas A, Naraharisetty H, Li X, Serrao E, Engelman A. Biochemical characterization of novel retroviral integrase proteins. PloS one. 2013; 8:e76638. [PubMed: 24124581]
32. Cherepanov P. LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity in vitro. Nucleic Acids Res. 2007; 35:113–124. [PubMed: 17158150]
33. Demeler, B., et al. UltraScan-III version 2.2: A comprehensive data analysis software package for analytical ultracentrifugation experiments. 2014. <http://www.ultrascan3.uthscsa.edu/>
34. Laue, TM.; Shah, BD.; Ridgeway, TM.; Pelletier, SL. Analytical ultracentrifugation in biochemistry and polymer science. Harding, SE.; Rowe, AJ.; Horton, JC., editors. Royal Society of Chemistry; 1992. p. 90-125.
35. Demeler B. Methods for the design and analysis of sedimentation velocity and sedimentation equilibrium experiments with proteins. Curr Protoc Prot Sci. 2010:13. **Chapter 7** Unit 7.13.
36. Brookes E, Cao W, Demeler BA. two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. Eur Biophys J. 2010; 39:405–414. [PubMed: 19247646]
37. Schuck P, Demeler B. Direct sedimentation analysis of interference optical data in analytical ultracentrifugation. Biophys J. 1999; 76:2288–2296. [PubMed: 10096923]

38. Brookes, EH.; Demeler, B. Proceedings of the 9th annual conference on Genetic and evolutionary computation. ACM; London, England: 2007. p. 361-368.
39. Demeler B, Brookes E. Monte Carlo analysis of sedimentation experiments. Colloid Polym Sci. 2008; 286:129–137.
40. Kabsch W, XDS. Acta Crystallogr D Biol Crystallogr. 2010; 66:125–132. [PubMed: 20124692]
41. Evans PR, Murshudov GN. How good are my data and what is the resolution? Acta Crystallogr D Biol Crystallogr. 2013; 69:1204–1214. [PubMed: 23793146]
42. McCoy AJ, et al. Phaser crystallographic software. J App Crystallogr. 2007; 40:658–674.
43. Bujacz G. High-resolution structure of the catalytic domain of avian sarcoma virus integrase. J Mol Biol. 1995; 253:333–346. [PubMed: 7563093]
44. Chen JC-H, et al. Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: A model for viral DNA binding. Proc Natl Acad Sci USA. 2000; 97:8233–8238. [PubMed: 10890912]
45. Morris RJ, Perrakis A, Lamzin VS. ARP/wARP and automatic interpretation of protein electron density maps. Methods Enzymol. 2003; 374:229–244. [PubMed: 14696376]
46. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr. 2004; 60:2126–2132. [PubMed: 15572765]
47. Adams PD, et al. The Phenix software for automated determination of macromolecular structures. Methods. 2011; 55:94–106. [PubMed: 21821126]
48. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr. 1997; 53:240–255. [PubMed: 15299926]
49. Chen VB, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010; 66:12–21. [PubMed: 20057044]
50. Suloway C, et al. Automated molecular microscopy: the new Leginon system. J Struct Biol. 2005; 151:41–60. [PubMed: 15890530]
51. Li X, et al. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat Methods. 2013; 10:584–590. [PubMed: 23644547]
52. Lyumkis D, et al. Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. Science. 2013; 342:1484–1490. [PubMed: 24179160]
53. Grant T, Grigorieff N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. eLife. 2015; 4:e06980. [PubMed: 26023829]
54. Lander GC, et al. Appion: an integrated, database-driven pipeline to facilitate EM image processing. J Struct Biol. 2009; 166:95–102. [PubMed: 19263523]
55. Sorzano CO, et al. A clustering approach to multireference alignment of single-particle projections in electron microscopy. J Struct Biol. 2010; 171:197–206. [PubMed: 20362059]
56. Scheres SH. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol. 2012; 180:519–530. [PubMed: 23000701]
57. Lyumkis D, Vinterbo S, Potter CS, Carragher B. Optimod–an automated approach for constructing and optimizing initial models for single-particle electron microscopy. J Struct Biol. 2013; 184:417–426. [PubMed: 24161732]
58. Grigorieff N. FREALIGN: high-resolution refinement of single particle structures. J Struct Biol. 2007; 157:117–125. [PubMed: 16828314]
59. Lyumkis D, Brilot AF, Theobald DL, Grigorieff N. Likelihood-based classification of cryo-EM images using FREALIGN. J Struct Biol. 2013; 183:377–388. [PubMed: 23872434]
60. DiMaio F, Zhang J, Chiu W, Baker D. Cryo-EM model validation using independent map reconstructions. Protein Sci. 2013; 22:865–868. [PubMed: 23592445]
61. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res. 2014; 42:W320–W324. [PubMed: 24753421]

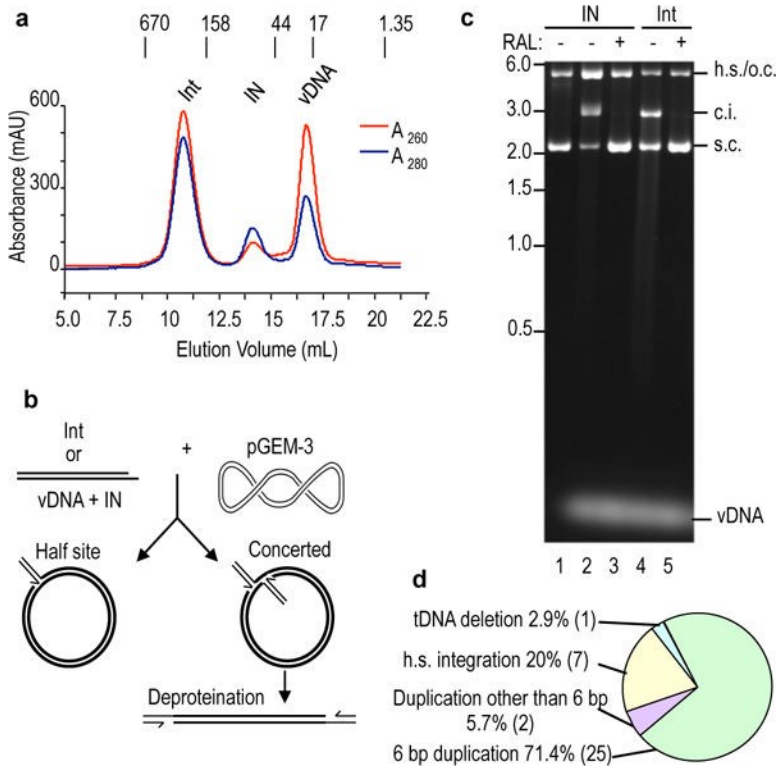


Figure 1. MMTV intasome (Int) characterization

a, Purification by SEC. Elution positions of mass standards in kDa are indicated. **b**, Integration assay schematic. Int or IN plus vDNA was reacted with supercoiled tDNA, which can yield half site (h.s.) or concerted integration (c.i.) products. **c**, Ethidium bromide-stained agarose gel. Lane 1–3 reactions were initiated with IN; vDNA was omitted from lane 1. Raltegravir (RAL) was included as indicated. Lanes 4 and 5, Int reactions. Migrations positions of h.s. products that co-migrate with open circular (o.c.) tDNA, c.i. products, supercoiled (s.c.) tDNA and mass standards in kb are indicated. For gel source data, see Supplementary Figure 1. **d**, Sequenced Int-mediated concerted integration products (n=35).

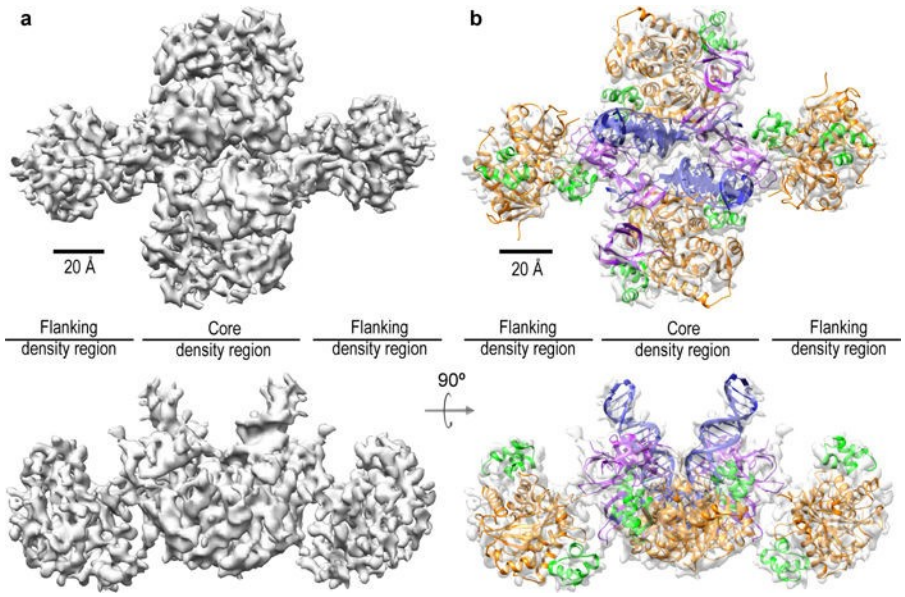


Figure 2. Cryo-EM structure of the MMTV intasome

a, top view (upper) of the cryo-EM map; the lower view is rotated 90°. Core density and flanking density regions are indicated. **b**, Individual domain crystal structures (NTD, green; CCD, orange; CTD, purple) and vDNA (blue) model fitted by rigid body docking. Rulers demarcate 20 Å.

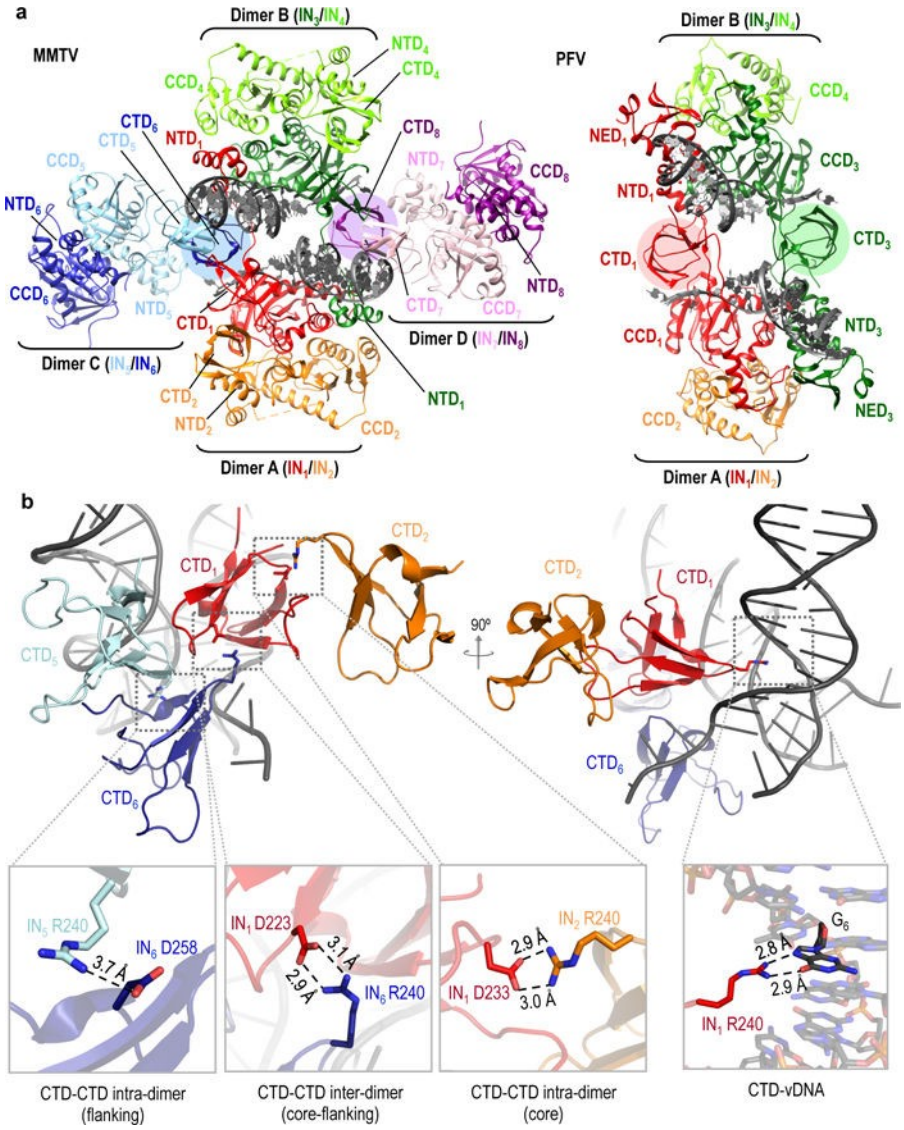


Figure 3. Comparison of MMTV and PFV intasome structures

a, MMTV (left) and PFV (right) intasomes color coded to highlight IN dimers and constituent protomers. Core dimers A and B are red-orange and green-chartreuse, respectively, while MMTV flanking IN dimers C and D are blue-sky blue and purple-light pink, respectively. Colored circles highlight similarly positioned CTDs between structures. **b**, Close-up views of Arg240-mediated protein (left) and vDNA (right; G6 of plus-strand) interactions. For simplicity, only one set of asymmetric interactions is shown. The interaction of IN_5 with residues 258–261 of IN_6 varied during model refinement, with the indicated interaction (as well as other atomic distances) observed in the final refined model. Colors are conserved between panels **a** and **b**.

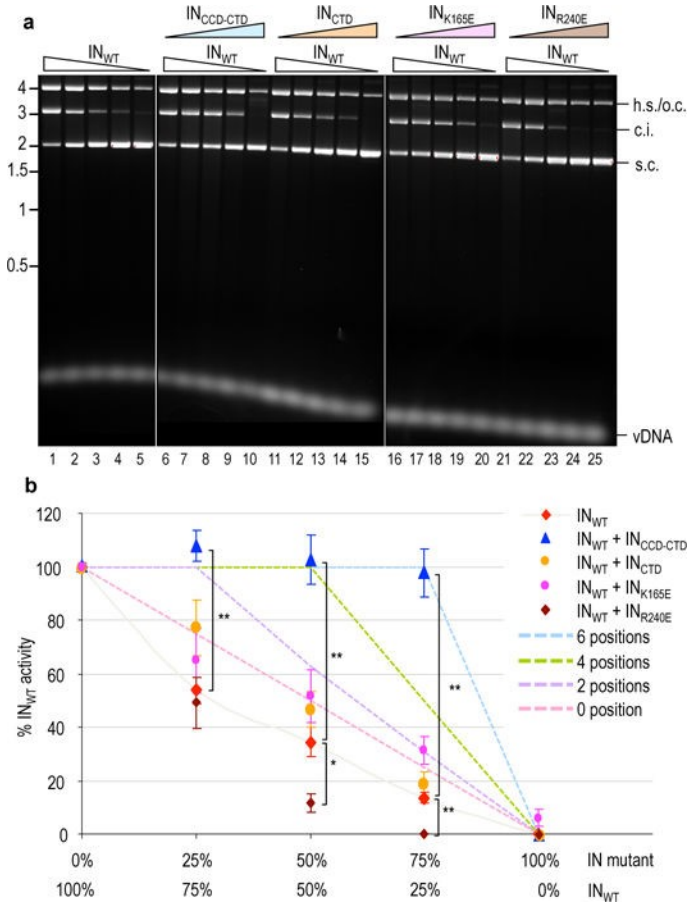


Figure 4. MMTV intasome functionality

a, Representative agarose gels. The reactions in lanes 1–4 contained 1,0, 0.75, 0.5, 0.25 μM IN_{WT}, respectively; IN was omitted from the reaction in lane 5. Subsequent 5-reaction sets contained the same IN_{WT} concentrations with 0, 0.25, 0.5, 0.75, 1.0 μM of the indicated mutant protein, for a total concentration of 1 μM IN in lanes 6–25. Lanes 1–5 versus lanes 6–15 and 16–25 were from separate agarose gels (see Supplementary Fig. 1 for gel source data); other labeling as in Fig. 1. **b**, Dashed lines indicate theoretical activities (graphed as percent IN_{WT} activity) for mixtures that contain a mutant protein that supports full IN_{WT} function when present in 6 of 8 octamer positions (blue dashed line), 4 of 8 positions (green dashes), 2 positions (purple dashes) or is unable to complement IN_{WT} function (pink dashes). Actual activities are from 4 technical replicates (average ± s.e.m.; see Supplementary Table 1 for source data). The non-linear response of IN_{WT} (grey line with red diamonds) likely reflects concentration-dependent cooperative multimerization of IN with vDNA[30]. The IN_{WT} alone and IN_{WT} + IN_{CTD} values were not significantly different (P >0.1; two-tailed t-test). *, P<0.05; **, P<0.01.