LETTER

doi:10.1038/nature16955

# Cryo–EM reveals a novel octameric integrase structure for betaretroviral 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] & Alan N. Engelman[1]

**Retroviral integrase catalyses the integration of viral DNA into host target DNA, which is an essential step in the life cycle of all retroviruses[1]. Previous structural characterization of integrase–viral DNA complexes, or intasomes, from the spumavirus prototype foamy virus revealed a functional integrase tetramer[2–5], and it is generally believed that intasomes derived from other retroviral genera use tetrameric integrase[6–9]. However, the intasomes of orthoretroviruses, which include all known pathogenic species, have not been characterized structurally. Here, using single-particle cryo-electron microscopy and X-ray crystallography, we determine an unexpected octameric integrase architecture for the intasome of the betaretrovirus mouse mammary tumour virus. The structure is composed of two core integrase dimers, which interact with the viral DNA ends and structurally mimic the integrase tetramer of prototype foamy virus, and two flanking integrase dimers that engage the core structure via their integrase carboxy-terminal domains. Contrary to the belief that tetrameric integrase components are sufficient to catalyse integration, the flanking integrase dimers were necessary for mouse mammary tumour virus integrase activity. The integrase octamer solves a conundrum for betaretroviruses as well as alpharetroviruses by providing critical carboxy-terminal domains to the intasome core that cannot be provided in *cis* because of evolutionarily restrictive catalytic core domain–carboxy-terminal domain linker regions. The octameric architecture of the intasome of mouse mammary tumour virus provides new insight into the structural basis of retroviral DNA integration.**

Mouse mammary tumour virus (MMTV) intasomes were assembled from integrase (IN) and viral DNA (vDNA) components by differential salt dialysis, akin to the strategy used for prototype foamy virus (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, strand transfer reactions were conducted with supercoiled plasmid target DNA (tDNA) to monitor the concerted integration of two vDNA ends[10] (Fig. 1b). The SEC-purified complexes catalysed 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 most contained 6 base pair (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. On the basis of 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 used here for intasome assembly.

To determine the MMTV intasome structure, single-particle cryo-electron microscopy (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 compared with the intasome core (Extended Data Fig. 3). Restricting data refinement to the core density region accordingly increased the resolution for the

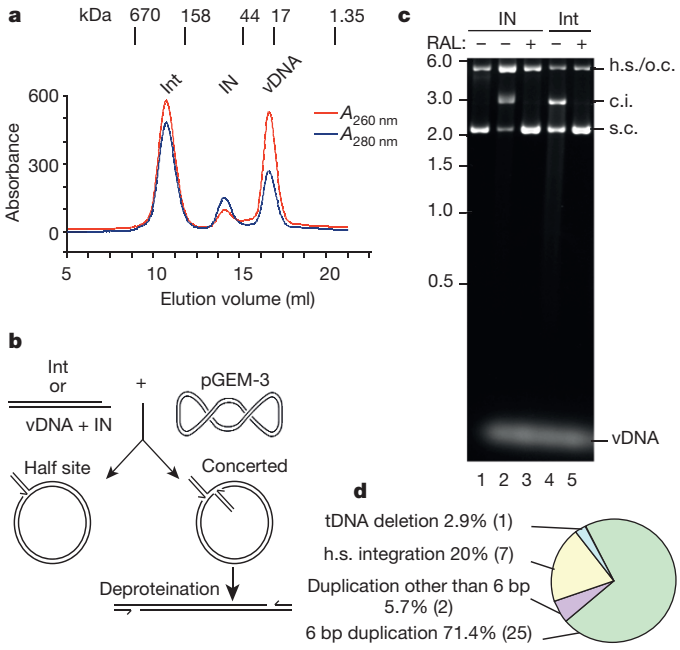



**Figure 1 | MMTV intasome (Int) characterization. a**, Purification by SEC. Elution positions of mass standards in kilodaltons are indicated. **b**, Integration assay schematic. Intasome or IN plus vDNA was reacted with supercoiled tDNA, which can yield half-site (h.s.) or concerted integration (c.i.)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

[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, Potters Bar, Hertfordshire 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.

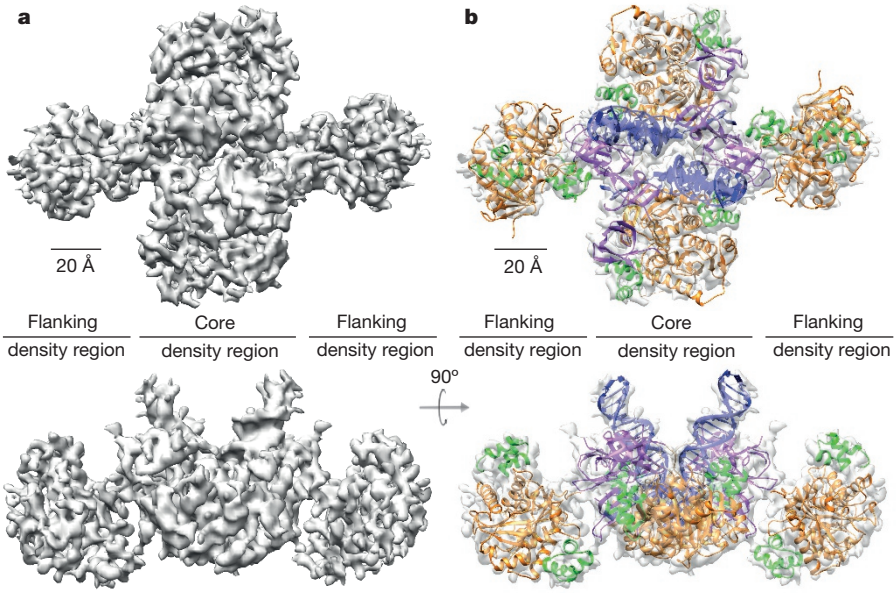



**Figure 2 | Cryo-EM structure of the MMTV intasome. a**, Top view (upper) of the cryo-EM map; the lower view is rotated by 90°. Core density and flanking density regions are indicated. **b**, Individual domain crystal structures (NTD, green; CCD, orange; CTD, purple) and vDNA (blue) model fitted by rigid body docking. Rulers demarcate 20 Å.

central portion of the structure to ~4 Å for the best-resolved regions (Extended Data Fig. 2d).

Each IN monomer is composed of an amino (N)-terminal domain (NTD), a catalytic core domain (CCD) and a carboxy (C)-terminal domain (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 $IN_{NTD–CCD}$ 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 modelled using PFV intasome DNA coordinates and by extending the modelled 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 most 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 (light green) 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 $IN_{NTD–CCD}$ 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 $IN_6$ neighbour 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 labour 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 Arg240-mediated IN–IN interactions, we compared $IN_{R240E}$ with $IN_{K165E}$, 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, $IN_{CCD–CTD}$ and $IN_{CTD}$, which purified as dimers and monomers, respectively (Extended Data Fig. 7), were additionally analysed. The reaction composed of 75% $IN_{CCD–CTD}$ supported near $IN_{WT}$ activity, indicating that this mutant could function when present in up to six of eight 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

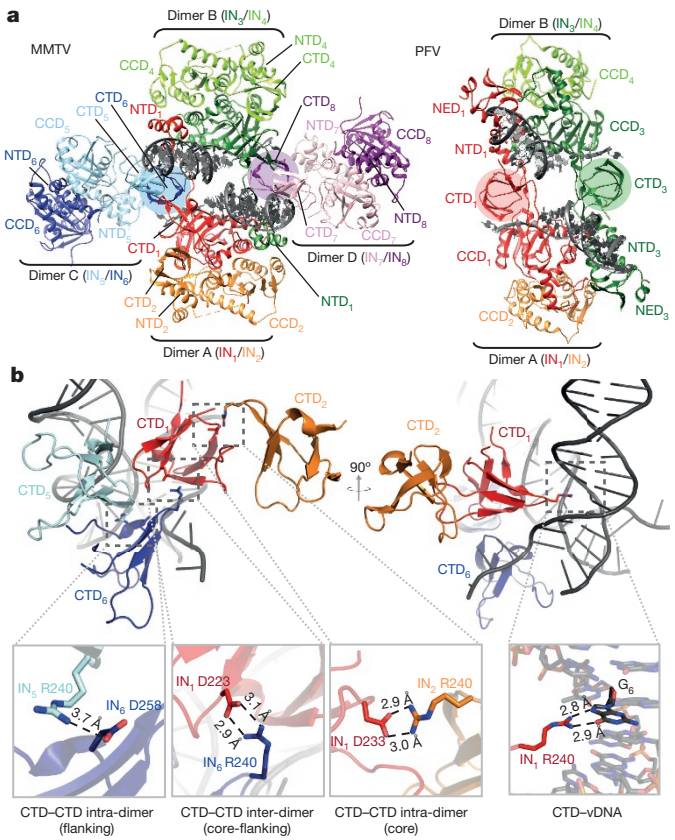



**Figure 3 | Comparison of MMTV and PFV intasome structures.** **a**, MMTV (left) and PFV (right) intasomes colour coded to highlight IN dimers and constituent protomers. Core dimers A and B are red–orange and green–chartreuse (light green), respectively, while MMTV flanking IN dimers C and D are blue–sky blue and purple–light pink, respectively. Coloured 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. Colours are conserved between **a** and **b**.

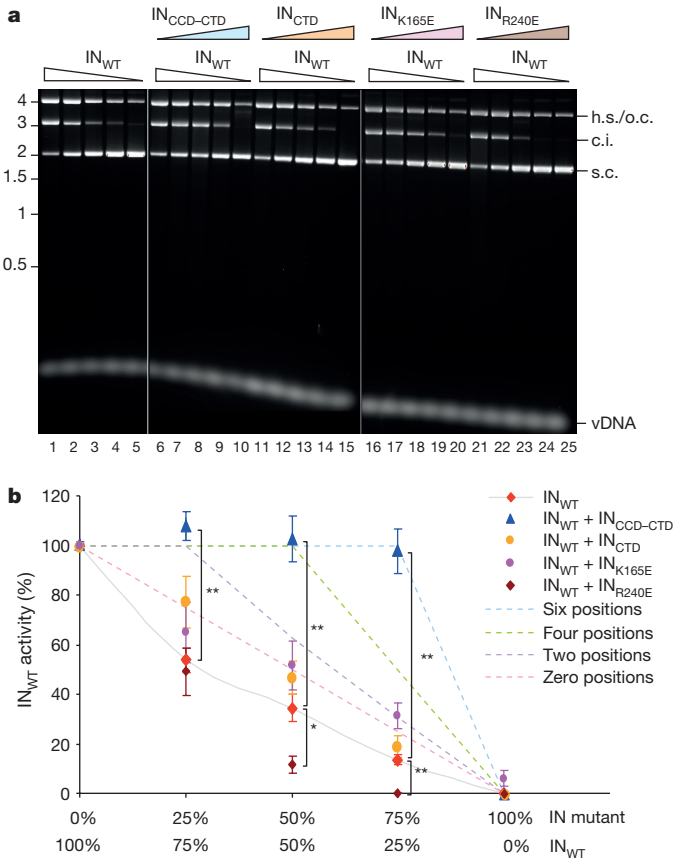



**Figure 4 | MMTV intasome functionality.** **a**, Representative agarose gels. The reactions in lanes 1–4 contained 1.0, 0.75, 0.5, 0.25 μM $IN_{WT}$, respectively; IN was omitted from the reaction in lane 5. Subsequent five-reaction sets contained the same $IN_{WT}$ concentrations with 0, 0.25, 0.5, 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 labelling as in Fig. 1. **b**, Dashed lines indicate theoretical activities (graphed as percentage $IN_{WT}$ activity) for mixtures that contain a mutant protein that supports full $IN_{WT}$ function when present in six of eight octamer positions (blue dashed line), four of eight positions (green dashes), two positions (purple dashes) or is unable to complement $IN_{WT}$ function (pink dashes). Actual activities are from four technical replicates (average ± s.e.m.; see Supplementary Table 1 for source data). The nonlinear response of $IN_{WT}$ (grey line with red diamonds) probably 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). $^*P < 0.05$; $^{**}P < 0.01$.

curve overlaid that predicted for non-functional protein, we moreover conclude that CCD-mediated dimerization 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 fifty-residue CCD–CTD linkers afford the positioning of inner PFV IN CTDs for vDNA and tDNA engagement[2,3], the analogous eight-amino-acid 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 gamma- and epsilonretroviral intasome activity, while an IN octamer will be required to catalyse alpharetrovirus in addition to betaretrovirus integration (Extended Data Fig. 8b). We note that an octameric IN architecture for the alpharetrovirus Rous sarcoma virus intasome has recently been independently determined[24]. Whereas most studies have highlighted a tetramer as the IN species that catalyses 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 re-evaluated.

PFV IN, which cleaves tDNA phosphodiester bonds separated by 4 bp, preferentially integrates into flexible sequences, whereas MMTV and Rous sarcoma virus 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 catalyses PFV and MMTV integration into sharply bent versus relatively non-deformed tDNA, respectively (Extended Data Fig. 8d). Owing 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 alpha- and betaretroviral intasomes to host chromatin.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 14 August; accepted 30 December 2015.**


1. Craigie, R. & Bushman, F. D. HIV DNA integration. *Cold Spring Harb. Perspect. Med.* **2,** a006890 (2012).
2. Hare, S., Gupta, S. S., Valkov, E., Engelman, A. & Cherepanov, P. Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature* **464,** 232–236 (2010).
3. Maertens, G. N., Hare, S. & Cherepanov, P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468,** 326–329 (2010).
4. Hare, S., Maertens, G. N. & Cherepanov, P. 3′-Processing and strand transfer catalysed by retroviral integrase *in crystallo*. *EMBO J.* **31,** 3020–3028 (2012).
5. Maskell, D. P. *et al.* Structural basis for retroviral integration into nucleosomes. *Nature* **523,** 366–369 (2015).
6. Yang, Z. N., Mueser, T. C., Bushman, F. D. & Hyde, C. C. Crystal structure of an active two-domain derivative of Rous sarcoma virus integrase. *J. Mol. Biol.* **296,** 535–548 (2000).
7. Wang, J. Y., 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.* **20,** 7333–7343 (2001).
8. Bao, K. K. *et al.* Functional oligomeric state of avian sarcoma virus integrase. *J. Biol. Chem.* **278,** 1323–1327 (2003).
9. Li, M., Mizuuchi, M., Burke, T. R., Jr & Craigie, R. Retroviral DNA integration: reaction pathway and critical intermediates. *EMBO J.* **25,** 1295–1304 (2006).
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.* **5,** e1000259 (2009).
11. Majors, J. E. & Varmus, H. E. Nucleotide sequences at host-proviral junctions for mouse mammary tumour virus. *Nature* **289,** 253–258 (1981).
12. Roth, M. J., Schwartzberg, P. L. & Goff, S. P. Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. *Cell* **58,** 47–54 (1989).
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* **62,** 829–837 (1990).
14. Ellison, V. & Brown, P. O. A stable complex between integrase and viral DNA ends mediates human immunodeficiency virus integration *in vitro*. *Proc. Natl Acad. Sci. USA* **91,** 7316–7320 (1994).
15. DiMaio, F. *et al.* Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nature Methods* **12,** 361–365 (2015).
16. Kudryashev, M. *et al.* Structure of the type VI secretion system contractile sheath. *Cell* **160,** 952–962 (2015).
17. Wang, R. Y. *et al. De novo* protein structure determination from near-atomic-resolution cryo-EM maps. *Nature Methods* **12,** 335–338 (2015).
18. van Gent, D. C., Vink, C., Groeneger, A. A. & Plasterk, R. H. Complementation between HIV integrase proteins mutated in different domains. *EMBO J.* **12,** 3261–3267 (1993).
19. Engelman, A., Bushman, F. D. & Craigie, R. Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. *EMBO J.* **12,** 3269–3275 (1993).
20. Yang, F. & Roth, M. J. Assembly and catalysis of concerted two-end integration events by Moloney murine leukemia virus integrase. *J. Virol.* **75,** 9561–9570 (2001).
21. Diamond, T. L. & Bushman, F. D. Division of labor within human immunodeficiency virus integrase complexes: determinants of catalysis and target DNA capture. *J. Virol.* **79,** 15376–15387 (2005).
22. Baker, T. A., Mizuuchi, M., Savilahti, H. & Mizuuchi, K. Division of labor among monomers within the Mu transposase tetramer. *Cell* **74,** 723–733 (1993).
23. Jenkins, T. M., 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.* **16,** 6849–6859 (1997).
24. Yin, Z., Shi, K., Banerjee, S., Grandgenett, D. P. & Aihara, H. Crystal structure of the Rous sarcoma virus intasome. *Nature* http://dx.doi.org/10.1038/nature16950 (this issue).
25. Faure, A. *et al.* HIV-1 integrase crosslinked oligomers are active *in vitro*. *Nucleic Acids Res.* **33,** 977–986 (2005).
26. Bera, S., Pandey, K. K., Vora, A. C. & Grandgenett, D. P. Molecular interactions between HIV-1 integrase and the two viral DNA ends within the synaptic complex that mediates concerted integration. *J. Mol. Biol.* **389,** 183–198 (2009).
27. Lee, S. P., Xiao, J., Knutson, J. R., Lewis, M. S. & Han, M. K. $Zn^{2+}$ promotes the self-association of human immunodeficiency virus type-1 integrase *in vitro*. *Biochemistry* **36,** 173–180 (1997).
28. Heuer, T. S. & Brown, P. O. Photo-cross-linking studies suggest a model for the architecture of an active human immunodeficiency virus type 1 integrase-DNA complex. *Biochemistry* **37,** 6667–6678 (1998).
29. Serrao, E., Ballandras-Colas, A., Cherepanov, P., Maertens, G. N. & Engelman, A. N. Key determinants of target DNA recognition by retroviral intasomes. *Retrovirology* **12,** 39 (2015).
30. Engelman, A. & Craigie, R. Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function *in vitro*. *J. Virol.* **66,** 6361–6369 (1992).


**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We acknowledge support from US National Institutes of Health (NIH) grants R01 AI070042 (to A.N.E.), NIH P50 GM103368 and the Leona M. and Harry B. Helmsley Charitable Trust grant number 2012-PG-MED002 (to D.L., both funding sources provided equal support), NIH P50 GM082251 (to P.C.), NIH P30 AI060354 (Harvard University Center for AIDS Research), 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 UCSF Chimera package (supported by NIH P41 GM103331). CryoEM data collection was in part facilitated by the National Resource for Automated Molecular Microscopy (9 P41 GM103310). We thank B. Anderson and J.-C. Ducom at The Scripps Research Institute for help with EM data collection and network infrastructure, J. Fitzpatrick and F. Dwyer for computational support at The Salk Institute, 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.

**Author 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. modelled the intasome structure; B.D. collected and analysed the sedimentation velocity data; N.J.C. and P.C. expressed and purified $IN_{CCD}$, $IN_{NTD–CCD}$ 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 have been deposited in the Electron Microscopy Data Bank under accession numbers EMD-6440 and EMD-6441, respectively. X-ray diffraction data and the resulting $IN_{CCD}$, $IN_{NTD–CCD}$ and $IN_{CTD}$ structures have been deposited in the Protein Data Bank (PDB) under accession numbers 5CZ1, 5CZ2 and 5D7U, respectively. The core intasome structure has been deposited in the Protein Data Bank under accession number 3JCA. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.L. (dlyumkis@salk.edu) or A.N.E. (alan_engelman@dfci.harvard.edu).

## 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 harboured 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). After 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 mass 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 analysed 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, before 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. After centrifugation for 10 min 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 analysed 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 before 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 eight 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 after transformation of ligation mixtures into *E. coli* were sequenced using primers that annealed to the ends of the cassette DNA.

**Analytical ultracentrifugation.** We analysed sedimentation velocity at 20 °C in a Beckman Optima XL-I analytical ultracentrifuge using an An60Ti rotor and standard two-channel Epon Centerpieces (Beckman-Coulter). Samples were prepared in 20 mM phosphate buffer, pH 7.5, 150 mM NaCl at two loading concentrations, absorbance ($A_{280\,nm}$) values of 0.3 and 0.9 for MMTV IN and the intasome, and $A_{280\,nm}$ values of 0.18 and 0.53 for vDNA, to exclude potential mass action oligomerization. IN and vDNA were spun simultaneously at 35,000 r.p.m. for 22 h while the intasome was spun at 27,000 r.p.m. for 12 h; the different rotor speeds were based on the predicted masses of the different macromolecules.

Data were analysed 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 $ml^{-1}$ and 1.101 centipoise, respectively. The partial specific volume of IN (0.728 ml $g^{-1}$) was estimated by UltraScan from its protein sequence using a method analogous to the methods outlined in ref. 34. Sedimentation velocity data were analysed as described[35]. Optimization was performed by two-dimensional spectrum analysis[36] with simultaneous removal of time-invariant and radially-invariant noise contributions[37]. Two-dimensional spectrum analysis 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 $IN_{CCD}$, $IN_{NTD–CCD}$ 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 $IN_{NTD–CCD}$) 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). After digestion with HRV 3C protease to release $His_6$ tags, the proteins were further purified by ion exchange and SEC.

Crystals were grown by vapour diffusion in hanging drops by mixing 1 μl protein (6–10 mg $ml^{-1}$ 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 ($IN_{NTD–CCD}$) or 19% isopropanol, 50 mM ammonium acetate, 0.1 M HEPES-NaOH, pH 7.5 ($IN_{CTD}$). Crystals, cryoprotected with 25% glycerol ($IN_{CCD}$, $IN_{NTD–CCD}$) or 30% PEG-400 ($IN_{CTD}$), were frozen by immersion in liquid nitrogen. Diffraction data for the $IN_{CCD}$ were collected using a charge-coupled device detector at beamline BM14 (European Synchrotron Radiation Facility) whereas $IN_{CTD}$ and $IN_{NTD–CCD}$ crystals were analysed 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 $IN_{NTD–CCD}$ structure. Final models, validated with MolProbity[49], had at least 96.9% of residues in the favoured 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 (6 s, Gatan Solarus plasma cleaner) holey carbon C-flat grids (CF-1.2/1.3-4C, Protochips), adsorbed for 30 s 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 electrons per pixel per square ångström at a rate of ~6.9 electrons per pixel per second and an estimated underfocus ranging from 1 to 4 μm (centred at 2.6 ± 0.6 μm). The dose was fractionated over 50 raw frames collected over a 10-s exposure time (200 ms per frame) on a Gatan K2 Summit direct detection device, with each frame receiving a dose of ~6.5 electrons per pixel per second. Two thousand seven hundred and fourteen 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 graphic processing unit-enabled whole-frame alignment program as previously described[51,52], and exposure filtered[53] according to a dose rate of 6.9 electrons per pixel per second. See Extended Data Table 2 for additional details on cryo-EM data collection.

**Cryo-EM image analysis.** Pre-processing operations before the refinement of the final models were performed using the Appion package[54] and were conceptually

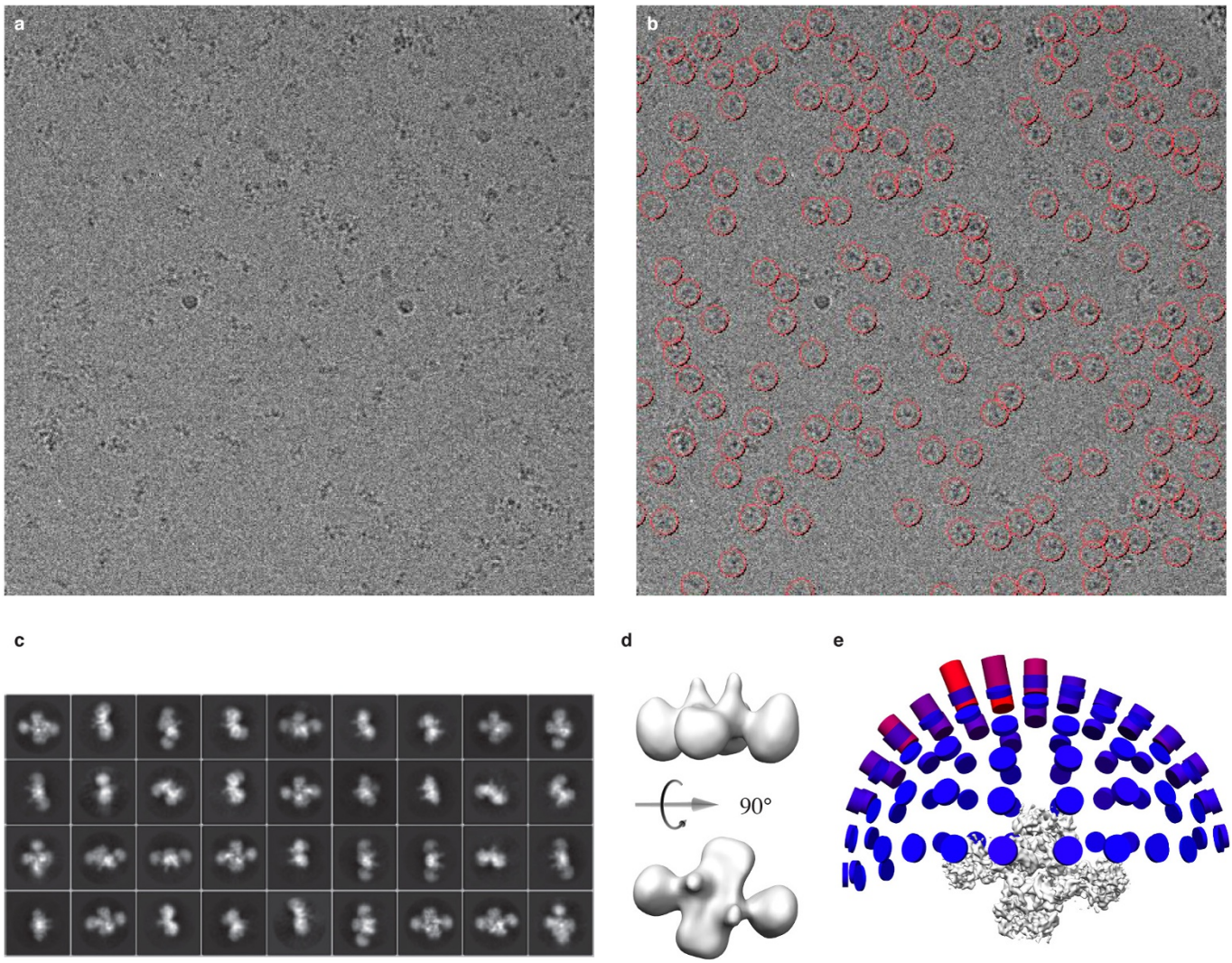
identical to those previously described[52]. Briefly, single intasome particles (244,315) were selected from the aligned and summed micrographs, from which 147,850 were used to create an initial raw particle stack after removing regions of the micrographs containing carbon and large areas of aggregation. Two-dimensional 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). After iterative rounds of two-dimensional alignment and classification, 77,365 particles remained for three-dimensional refinement and classification. Three-dimensional 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 after three-dimensional 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 shown 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, data sets lacking most of the flanking INs were created. Refinement of the core intasome data set 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.

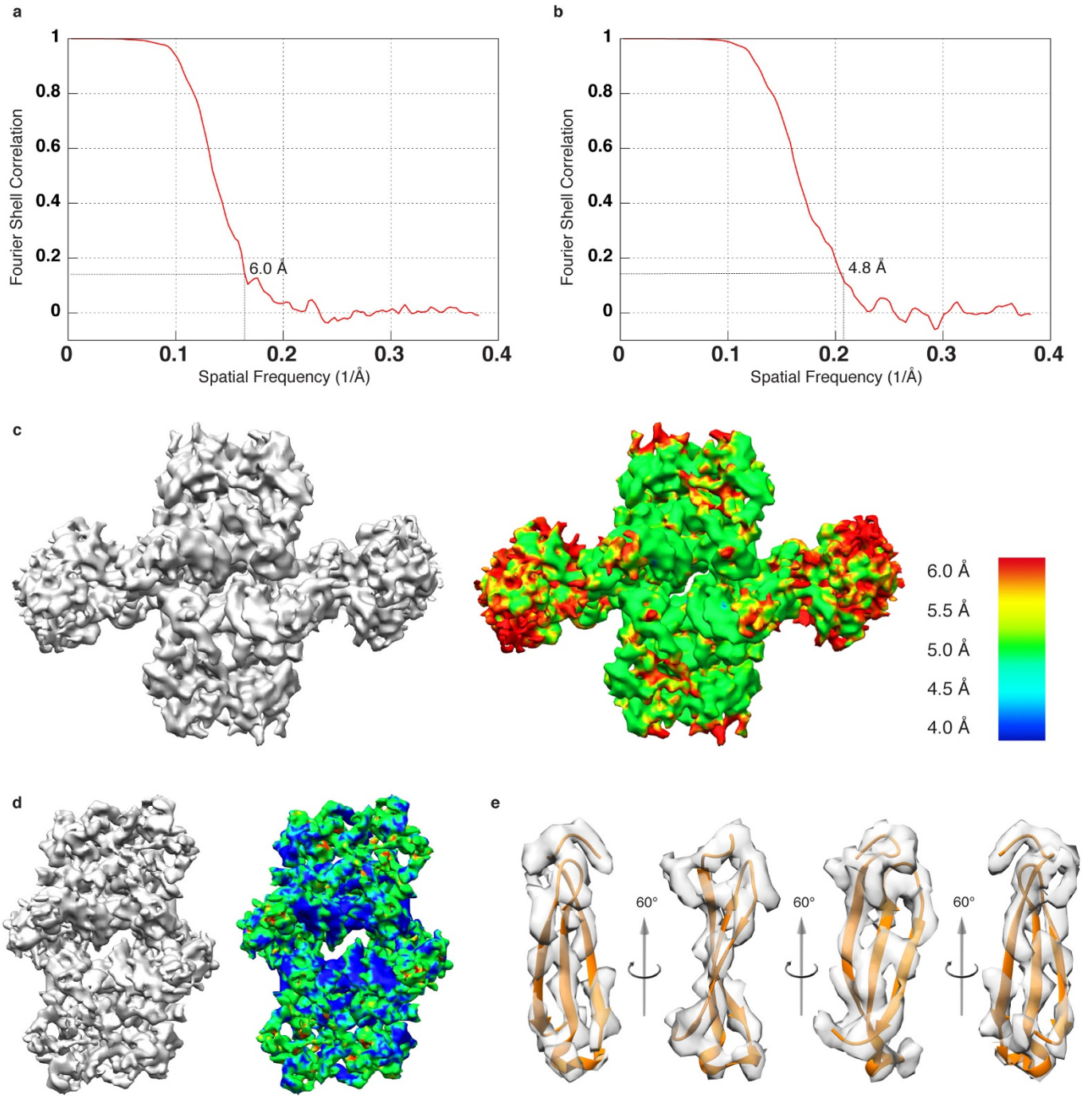
**Assembly of the atomic model.** Models of the core intasome and the full octamer structures were built and refined in a stepwise manner using Rosetta[15] starting with rigid-body fitted X-ray structures of individual domains as input. Rosetta protocols were used for all parts of the modelling[60]. 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 six 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 $IN_{NTD–CCD}$ 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 $IN_{NTD–CCD}$ 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 2,000 models were generated from each and the lowest-energy model was selected for moving forward. Modelling 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 seven-amino-acid oligopeptides from the PDB[15]. This enabled *de novo* modelling 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 modelled owing to disorder; Extended Data Fig. 6b, c); modelling 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. Two thousand 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 modelled on the basis of the X-ray structure of the PFV intasome (PDB accession number 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 Fig. 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: gammaretroviruses: Moloney murine leukaemia virus (GenBank accession number J02255.1), reticuloendotheliosis virus strain A (DQ237900.1), feline leukaemia virus (NC_001940.1); epsilonretroviruses: 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 anaemia virus (M16575.1), feline immunodeficiency virus (M25381.1), caprine arthritis encephalitis virus (M33677.1), bovine immunodeficiency virus (NC_001413.1); deltaretroviruses: bovine leukaemia virus (K02120.1), human T-cell lymphotropic virus types 1 and 2 (NC_001436.1 and NC_001488.1, respectively); betaretroviruses: MMTV (NC_001503.1), Mason Pfizer monkey virus (NC_001550.1), Jaagsiekte sheep retrovirus (NC_001494.1); alpharetroviruses: Rous sarcoma virus (J02342.1), lymphoproliferative disease virus (KC802224.1).

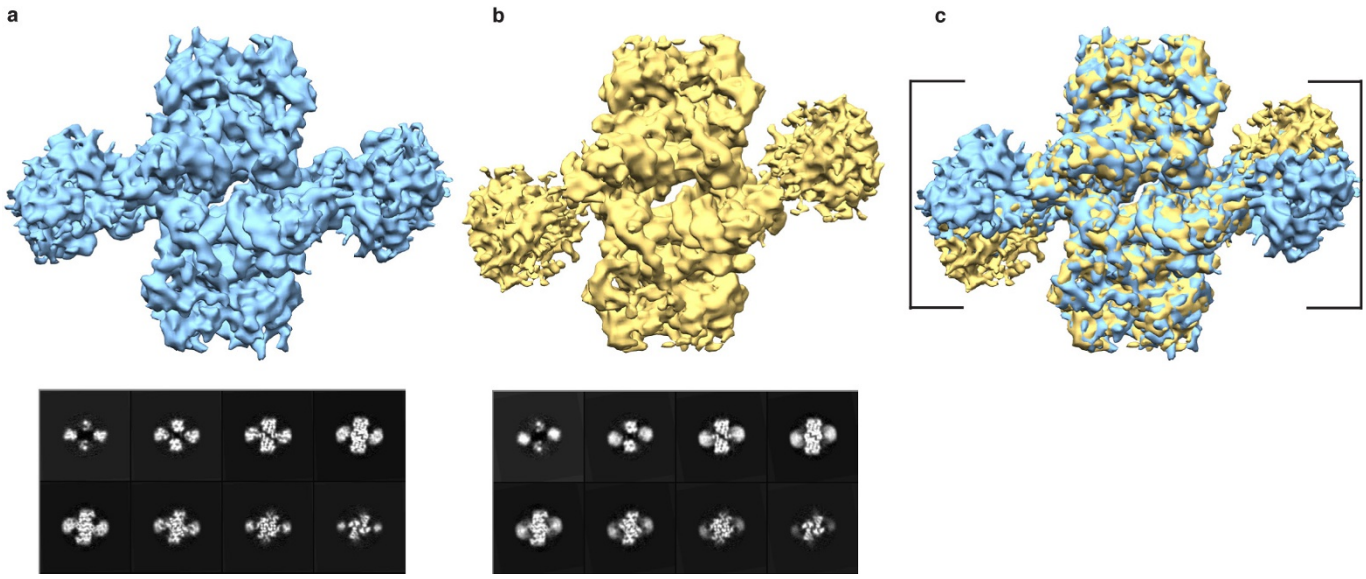

31. Ballandras-Colas, A., Naraharisetty, H., Li, X., Serrao, E. & Engelman, A. Biochemical characterization of novel retroviral integrase proteins. *PLoS ONE* **8,** e76638 (2013).
32. Cherepanov, P. LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity *in vitro*. *Nucleic Acids Res.* **35,** 113–124 (2007).
33. Demeler, B. *et al.* UltraScan-III version 2.2: a comprehensive data analysis software package for analytical ultracentrifugation experiments <http://www.ultrascan3.uthscsa.edu/> (2014).
34. Laue, T. M., Shah, B. D., Ridgeway, T. M. & Pelletier, S. L. in *Analytical Ultracentrifugation in Biochemistry and Polymer Science* (eds Harding, S. E., Rowe, A. J. & Horton, J. C.) 90–125 (Royal Society of Chemistry, 1992).
35. Demeler, B. in *Current Protocols in Protein Science* (eds Coligan, J. E. *et al.*) Ch. 7, Unit 7.13, 7.13.1–7.13.24 (Wiley, 2010).
36. Brookes, E., Cao, W. & Demeler, B. A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. *Eur. Biophys. J.* **39,** 405–414 (2010).
37. Schuck, P. & Demeler, B. Direct sedimentation analysis of interference optical data in analytical ultracentrifugation. *Biophys. J.* **76,** 2288–2296 (1999).
38. Brookes, E. H. & Demeler, B. in *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation* 361–368 (Association for Computing Machinery, 2007).
39. Demeler, B. & Brookes, E. Monte Carlo analysis of sedimentation experiments. *Colloid Polym. Sci.* **286,** 129–137 (2008).
40. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
41. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D* **69,** 1204–1214 (2013).
42. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
43. Bujacz, G. *et al.* High-resolution structure of the catalytic domain of avian sarcoma virus integrase. *J. Mol. Biol.* **253,** 333–346 (1995).
44. Chen, J. C.-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* **97,** 8233–8238 (2000).
45. Morris, R. J., Perrakis, A. & Lamzin, V. S. ARP/wARP and automatic interpretation of protein electron density maps. *Methods Enzymol.* **374,** 229–244 (2003).
46. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
47. Adams, P. D. *et al.* The Phenix software for automated determination of macromolecular structures. *Methods* **55,** 94–106 (2011).
48. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
49. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
50. Suloway, C. *et al.* Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151,** 41–60 (2005).
51. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10,** 584–590 (2013).
52. Lyumkis, D. *et al.* Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. *Science* **342,** 1484–1490 (2013).
53. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife* **4,** e06980 (2015).
54. Lander, G. C. *et al.* Appion: an integrated, database-driven pipeline to facilitate EM image processing. *J. Struct. Biol.* **166,** 95–102 (2009).
55. Sorzano, C. O. *et al.* A clustering approach to multireference alignment of single-particle projections in electron microscopy. *J. Struct. Biol.* **171,** 197–206 (2010).
56. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
57. Lyumkis, D., Vinterbo, S., Potter, C. S. & Carragher, B. Optimod – an automated approach for constructing and optimizing initial models for single-particle electron microscopy. *J. Struct. Biol.* **184,** 417–426 (2013).
58. Grigorieff, N. FREALIGN: high-resolution refinement of single particle structures. *J. Struct. Biol.* **157,** 117–125 (2007).
59. Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. Likelihood-based classification of cryo-EM images using FREALIGN. *J. Struct. Biol.* **183,** 377–388 (2013).
60. DiMaio, F., Zhang, J., Chiu, W. & Baker, D. Cryo-EM model validation using independent map reconstructions. *Protein Sci.* **22,** 865–868 (2013).
61. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42,** W320–W324 (2014).



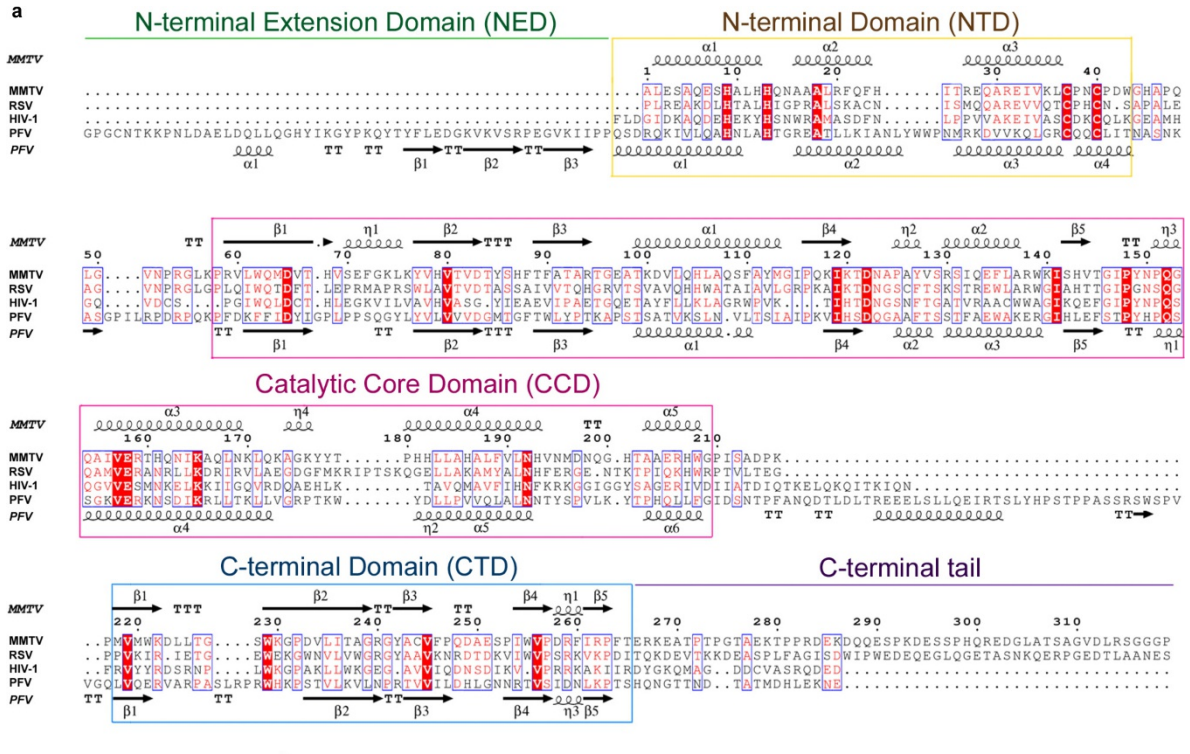

**Extended Data Figure 1 | Cryo-EM data and refinement.** **a**, Representative cryo-electron micrograph of MMTV intasomes, taken at 2.7 μm underfocus. **b**, Same as in **a**, marked to show selected particles. **c**, Two-dimensional class averages calculated using Relion[56]. **d**, Initial model from the class averages calculated using OptiMod[57]. **e**, Refined reconstruction from the full data set, with an Euler angle distribution plot showing the relative orientations of the particles.

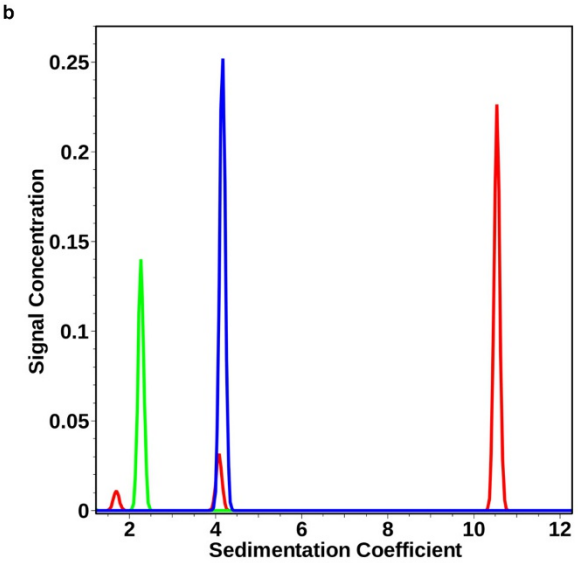



**Extended Data Figure 2 | Cryo-EM resolution analysis of reconstructed intasome maps. a**, Fourier shell correlation curve corresponding to the refined map generated from the full intasome data set. **b**, Fourier shell correlation curve corresponding to the refined map generated from the core intasome data set with the NTDs, CCDs and interdomain linker regions of the flanking IN dimers computationally subtracted. Average global resolutions in **a** and **b** are indicated. **c**, Refined map generated from the full data set (left) displayed side-by-side with the same map coloured for local resolution (right). **d**, Refined map generated from the core intasome data set (left) displayed side-by-side with the same map coloured for local resolution (right) using the colouring scheme in **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 three-dimensional classification of the data. Slices from the density map are displayed below. **b**, One of several conformations of MMTV intasome refinement after three-dimensional 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).

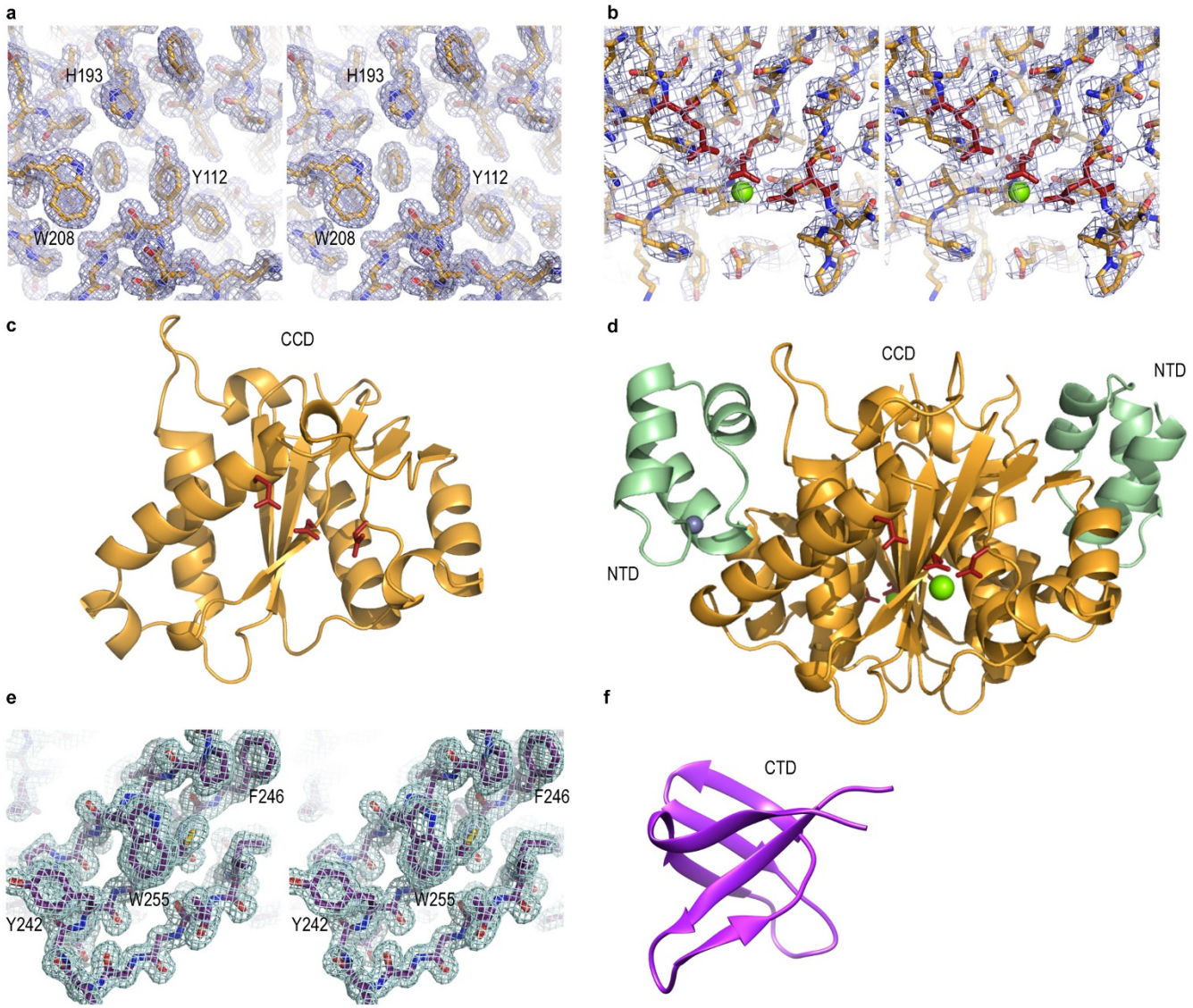


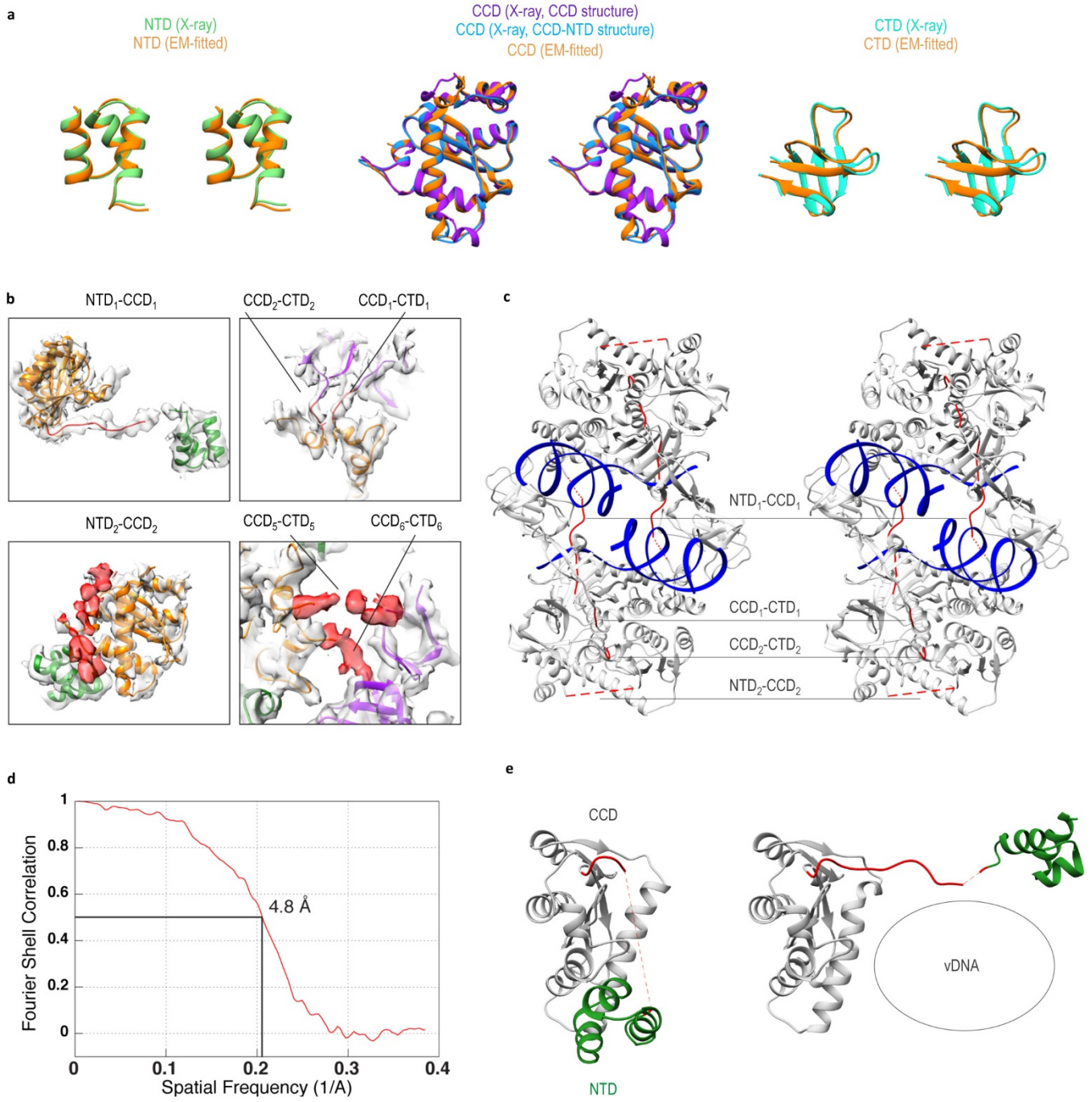

| Species | Molar mass (kDa) (measured) | Molar mass (kDa) (theorical) | s ($x10^{-13}$ sec) |
|---|---|---|---|
| vDNA | 11.9 | 13.2 | 2.27 |
| IN | 40.9 | 35.9 | 3.89 |
| intasome | 302.1* | 313.6 | 10.53 |

** assuming a partial specific volume weight-average of 0.713 ml/g for an 8:2 protein:DNA complex.*

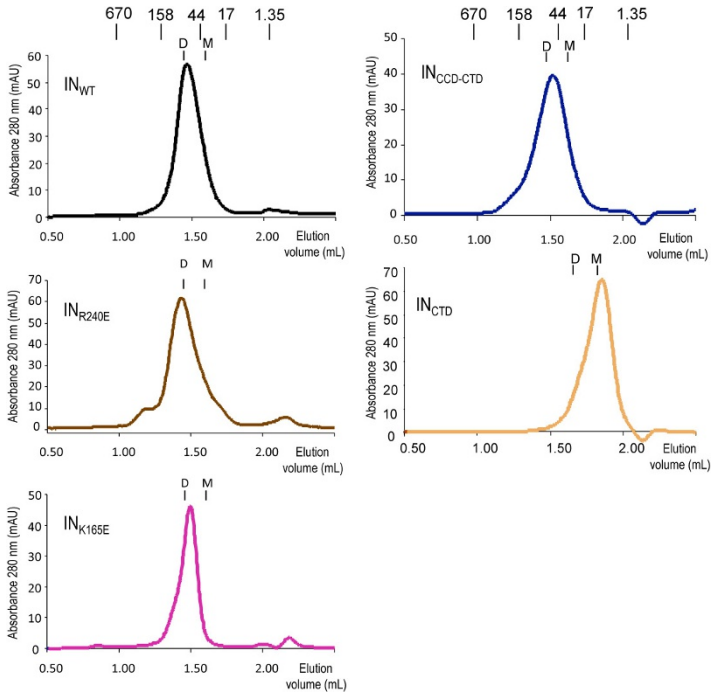
**Extended Data Figure 4 | MMTV IN domains and intasome sedimentation coefficient distribution. a**, Primary IN sequence alignment with boxes denoting canonical IN structural domains. The N-terminal extension domain occurs in spuma-, gamma- and epsilonretroviral IN proteins. Identical residues between MMTV, Rous sarcoma virus, 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 accession number 3L2Q; MMTV elements are from the $IN_{NTD-CCD}$ and $IN_{CTD}$ crystal structures described here (PDB accession numbers 5CZ2 and 5D7U, respectively). Symbols 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) probably attributed to the minor variations in sedimentation coefficient between major and minor IN and vDNA species. Measured sedimentation coefficients and calculated molar masses compared with theoretical molar masses are shown beneath the graph.



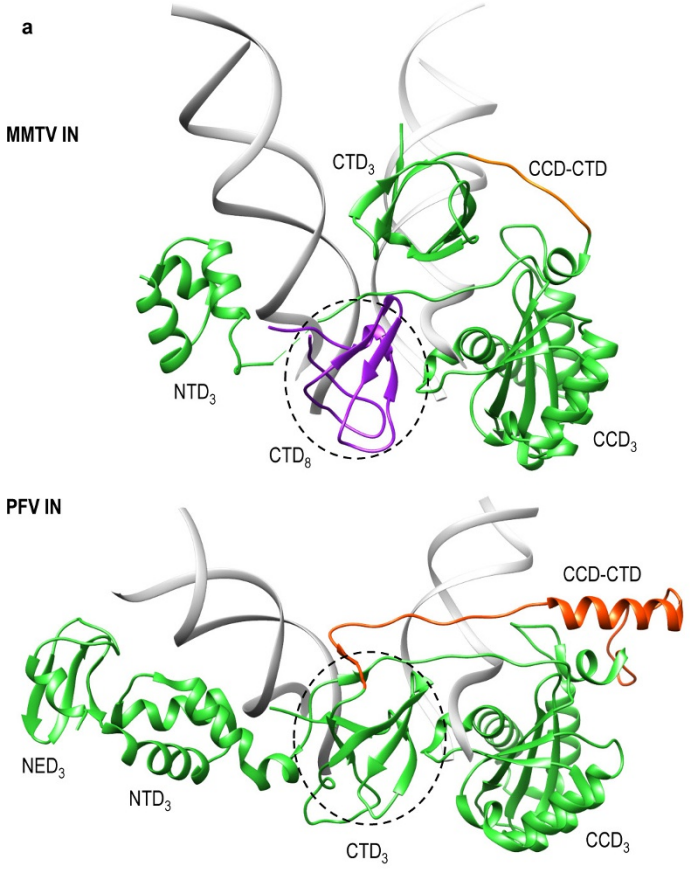

**Extended Data Figure 5 | MMTV IN domain crystal structures. a**, Stereo view of the final $2F_o - F_c$ density map of the $IN_{CCD}$ crystal structure with blue mesh contoured at $1\sigma$. Amino-acid side chains are readily evident at the 1.7 Å resolution. **b**, Stereo view of the final $2F_o - F_c$ density map of the 2.7 Å resolution $IN_{NTD-CCD}$ crystal structure with blue mesh contoured at $1\sigma$. The map is centred on the DDE catalytic triad (red sticks); green spheres, $Mg^{2+}$ ions. **c**, Cartoon representation of the $IN_{CCD}$ monomer (one of four in the crystallographic asymmetric unit) coloured in gold. Active site residues are shown as red sticks. **d**, Cartoon representation of the $IN_{NTD-CCD}$ dimer structure (one of three in the asymmetric unit). The NTD and CCD are coloured 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 $2F_o - F_c$ density map of the 1.5 Å resolution $IN_{CTD}$ crystal structure, shown as a green mesh contoured at $1\sigma$. **f**, Cartoon representation of one of the two CTD monomers present in the asymmetric unit.




**Extended Data Figure 6 | Molecular modelling 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 modelled, 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**, Fourier shell correlation 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).

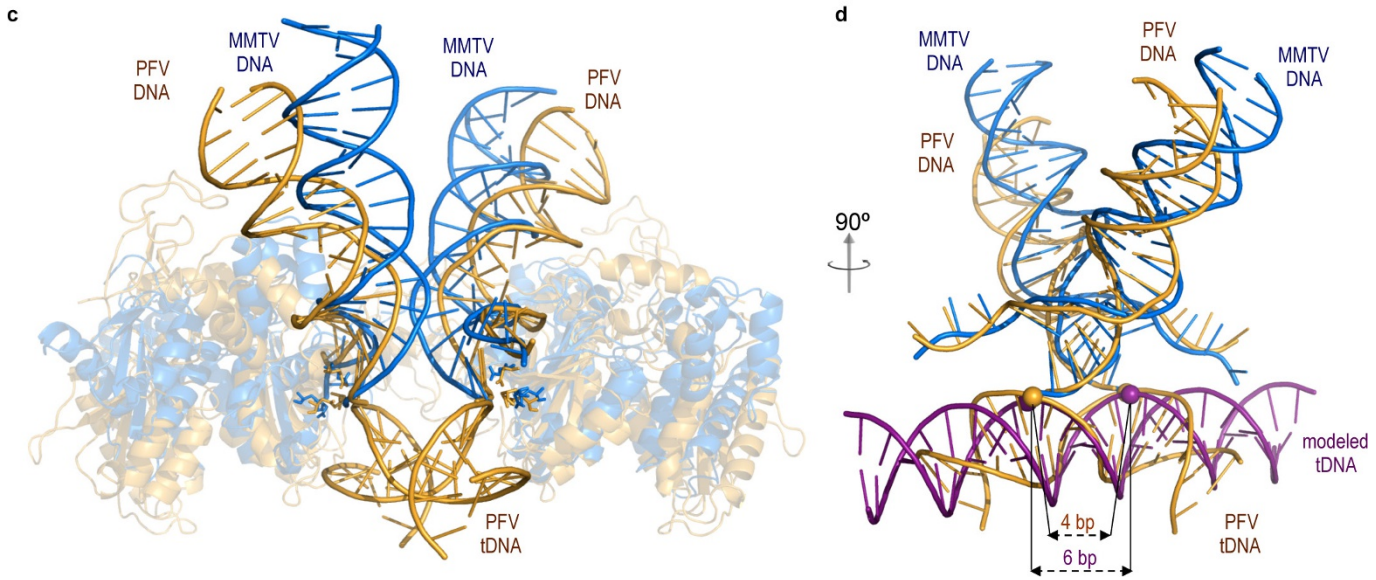



**Extended Data Figure 7 | Gel filtration profiles of $IN_{WT}$ and IN mutant proteins.** Elution profiles of mass standards in kilodaltons as well as theoretical protein monomer (M) and dimer (D) positions are indicated.




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 coloured 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 N-terminal extension domain (NED). The multimeric state of IN in known intasome structures is indicated by bold type. **c**, The PFV intasome with bound tDNA (PDB accession number 3OS2; orange) was superimposed with the MMTV intasome (blue). The distance between overlaid active sites is in each case ~26 Å. **d**, Ninety-degree 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 modelled 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) |
| $R_{merge}$ | 0.060 (0.57) | 0.08 (0.534) | 0.043 (0.585) |
| *I*/s*I* | 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 |
| 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 (96$^{th}$ percentile) | - |
| Clashscore, all atoms | 2.27 (99$^{th}$ 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.