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

Allison Ballardas-Colas1, Monica Brown2, Nicola J. Cook3, Tamaria G. Dewdney1, Borries Demeler4, Peter Cherepanov1,5, Dmitry Lyumkis2 & Alan N. Engelman1

Retroviral integrase catalyses the integration of viral DNA into host target DNA, which is an essential step in the life cycle of all retroviruses1. Previous structural characterization of integrase–viral DNA complexes, or intasomes, from the spumavirus prototype foamy virus revealed a functional integrase tetramer2–5, and it is generally believed that intasomes derived from other retroviral genera use tetrameric integrase6–9. 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 asalpharetroviruses by providing critical carboxy-terminal domains to the intasome core that cannot be provided in cis because of evolutionarily restrictive catalytic core–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) intasomes2. 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 ends10 (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 infection11 (Fig. 1d). To address the specificity of complex formation, the invariant CA dinucleotide, which is essential for IN catalysis12,13, 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++ promoted the assembly of active HIV-1 IN–vDNA complexes but was unable to support IN catalysis14, 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 assembly 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 assembly density regions that are conformationally mobile compared with the intasome core (Extended Data Fig. 3).

**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. d, Sequenced intasome-mediated concerted integration products (n = 35).

---

1Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute and Department of Medicine, Harvard Medical School, 450 Brookline Avenue, Boston, Massachusetts 02215, USA. 2Laboratory of Genetics and Helmsley Center for Genomic Medicine, The Salk Institute for Biological Studies, 10010 N Torrey Pines Road, La Jolla, California 92037, USA. 3Clare Hall Laboratories, The Francis Crick Institute, Blanche Lane, South Mimms, Potters Bar, Hertfordshire EN6 3LD, UK. 4Department of Biochemistry, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, Texas 78229, USA. 5Division of Medicine, Imperial College London, St. Mary’s Campus, Norfolk Place, London W2 1PG, UK. © 2016 Macmillan Publishers Limited. All rights reserved
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 intasome structures, the different protein domains were crystallized as IN\(_{\text{CCD}}\)-IN\(_{\text{CTD}}\) and IN\(_{\text{NTD}}\)-CCD fragments, and these structures were refined to 1.7 Å, 1.5 Å and 2.7 Å resolution, respectively (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 \textit{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. 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\(_{\text{NTD}}\)-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\(_3\) Arg240 interacts with IN\(_1\) Asp233 (Fig. 3b). Flanking protomer IN\(_3\) 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\(_{\text{WT}}\)) 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\(_{\text{R240E}}\), which like IN\(_{\text{WT}}\) 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\(_{\text{R240E}}\) with IN\(_{\text{K165E}}\), which carries a change that uniquely disrupts IN–vDNA binding\(^{23}\). In concordance with its inability to assume the conformationality of Arg240-mediated IN–IN interactions, we compared IN\(_{\text{R240E}}\) with IN\(_{\text{K165E}}\) mildly stimulated the activity of limited IN\(_{\text{WT}}\) protein (Fig. 4b), presumably providing a source for other IN subunits within the functional complex. IN\(_{\text{R240E}}\) by contrast potently inhibited IN\(_{\text{WT}}\) function, confirming the importance of Arg240-mediated protein–protein interactions for MMTV IN activity. Two deletion mutant constructs, IN\(_{\text{CCD}}\)-CTD and IN\(_{\text{NTD}}\)-CTD, which purified as dimers and monomers, respectively (Extended Data Fig. 7), were additionally analysed. The reaction composed of 75% IN\(_{\text{CCD}}\)-CTD supported near IN\(_{\text{WT}}\) 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\(_{\text{CCD}}\)-CTD from assuming intasome core positions 1 and 3. As the IN\(_{\text{CTD}}\) response
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 IN tetramers for vDNA and tDNA engagement, 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 positions (green dashes), two positions (purple dashes) or is unable to complement INWT 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 INWT (grey line with red diamonds) probably reflects concentration-dependent cooperative multimerization of IN with vDNA.

The INWT alone and INWT + INCTD values were not significantly different (P > 0.1; two-tailed t-test).

For PFV IN, which cleaves tDNA phosphodiester bonds separated by 4bp, preferentially integrates into flexible sequences, whereas MMTV and Rous sarcoma virus IN, which cleave tDNA with 6bp staggerers, are relatively unconstrained by tDNA flexibility. Superposition of the MMTV and PFV intasome structures revealed that the two sets of catalytic IN active sites almost perfectly aligned (Extended Data Fig. 8b), the higher-order nature of IN in active HIV-1 intasomes may need to be re-evaluated.
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 intasome in crystalline. 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 intasome. J. Mol. Biol. 296, 535–548 (2000).
7. Wang, J. Y., Tong, H., Yang, W. & Craigie, R. Structure of a two-domain fragment of HIV-1 intasome: 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 intasome mutants in the presence of modified PSIP1/LEDGF/p75. PLoS Pathog. 5, e1000259 (2009).
11. Majors, J. E. & Varma, 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 ends mediates human immunodeficiency virus integration in vitro. Proc. Natl Acad. Sci. USA 91, 7316–7320 (1994).
13. 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).
14. Kudryashev, M. et al. Structure of the type VI secretion system contractile sheath. Cell 160, 952–962 (2015).
15. Wang, R. Y. et al. De novo protein structure determination from near-atomic-resolution cryo-EM maps. Nature Methods 12, 335–338 (2015).
16. van Gent, D. C., Vink, C., Groeneger, A. A. & Plasterk, R. H. Complementation between HIV intasome proteins mutated in different domains. EMBO J. 12, 3261–3267 (1993).
17. Engelman, A., Bushman, F. D. & Craigie, R. Identification of discrete functional domains of HIV-1 intasome and their organization within an active multimeric complex. EMBO J. 12, 3269–3275 (1993).
18. Yang, F. & Roth, M. J. Assembly and catalysis of concerted two-end integration events by Moloney murine leukemia virus intasomes. J. Virol. 75, 9561–9570 (2001).
19. Diamond, T. L. & Bushman, F. D. Division of labor within human immunodeficiency virus intasome complexes: determinants of catalysis and target DNA capture. J. Virol. 79, 15376–15387 (2005).
20. Faure, A. et al. HIV-1 intasome crosslinked oligomers are active in vitro. Nucleic Acids Res. 39 (2011).
21. Bera, S., Pandey, K. K., Vora, A. C. & Grandgenett, D. P. Molecular interactions between HIV-1 intasome and the two viral DNA ends within the synaptic complex that mediates concerted integration. J. Mol. Biol. 389, 183–196 (2009).
22. Heuer, T. S. & Brown, P. O. Photo-cross-linking studies suggest a model for the architecture of an active retroviral intasome. Nature 366, 667–668 (1990).
23. Serrao, E., Baillandras-Colas, A., Cherepanov, P., Maertens, G. N. & Engelman, A. N. Key determinants of target DNA recognition by retroviral intasomes. Retrovirology 12, 39 (2015).
24. Ying, 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 intasome crosslinked oligomers are active in vitro. Nucleic Acids Res. 39 (2011).
26. Yang, F. & Roth, M. J. Assembly and catalysis of concerted two-end integration events by Moloney murine leukemia virus intasomes. J. Virol. 75, 9561–9570 (2001).
27. Heuer, T. S. & Brown, P. O. Photo-cross-linking studies suggest a model for the architecture of an active retroviral intasome. Nature 366, 667–668 (1990).
28. Serrao, E., Baillandras-Colas, A., Cherepanov, P., Maertens, G. N. & Engelman, A. N. Key determinants of target DNA recognition by retroviral intasomes. Retrovirology 12, 39 (2015).

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 F05 GM103368 (to the Leona M. and Harry B. Helmsley Charitable Trust grant number 2012-PG-MED002 (to D.L., both funding sources provided equal support), NIH PG082251 (to P.C.), NIH P30 AI060354 (Harvard University Center for AIDS Research), and US National Science Foundation grants NSF-ACI-1252622 and TG-MCB0700539 (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 program (46) and visualized by pyMOL (56). N.J.C. and P.C. expressed and purified INCCD, INNTD-CCD and INCTD/212–266 intasomes; A.B.-C. assembled intasomes, characterized their biochemical analysis; A.B.-C. and T.G.D. expressed and purified MMTV IN proteins for 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; B.D. modelled the intasome structure; B.D. collected and analysed the sedimentation velocity data; N.J.C. and P.C. expressed and purified intasome constructs, established crystallization conditions and determined these structures.

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 intasome constructs, established crystallization conditions and determined these structures.

Author Information Coordinates of cryo-EM density maps for the full and core intasomes 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 INCCD, INNTD-CCD and INCTD structures have been deposited in the Protein Data Bank (PDB) under accession numbers 5CZ1, 5C2Z 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@dtci.harvard.edu).
PCRamplified 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-A 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 (A280nm) values of 0.3 and 0.9 for MMTV IN and the intasome, and 0.9 for INCCD–CTD and INNTD–CCD, 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 described35. Optimization was performed by two-dimensional spectrum analysis36 with simultaneous removal of time-invariant and radially-invariant noise contributions37. Two-dimensional spectrum analysis solutions, which are subjected to parsimonious regularization by genetic algorithm analysis38, were further refined using Monte Carlo analysis to determine confidence limits for the determined parameters39. Calculations were performed on the Texas Advanced Computing Center at the University of Texas at Austin.

**Protein expression and X-ray crystallography.** MMTV INCCD, INNTD–CCD and INCTD 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 μg/ml ZnCl2, for INCCD–CTD) by induction with 0.01% (v/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-nitrotriacetic acid agarose (Qiagen). After digestion with HRV 3C protease to release His 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 (INCCD), 19% PEG-3350, 0.2 M MgCl2, 5% (v/v) 1-butyl-3-methylimidazolium dicyanamide (INNTD–CCD) or 19% isopropanol, 50 mM ammonium acetate, 0.1 M HEPES-NaOH, pH 7.5 (INCCD–CTD crystals, cryo-protected with 25% glycerol (INCCD, INNTD–CCD) or 30% PEG-400 (INCTD), were frozen by immersion in liquid nitrogen. Diffraction data for the INCTD were collected using a charge-coupled device detector at beamline BM14 (European Synchrotron Radiation Facility) whereas INNTD–CCD and INCTD crystals were analysed at beamline I03 (Diamond Light Source) equipped with a PILATUS direct detector. The data, integrated with XDS40, were scaled with Aimless41. The structures, which were each derived from a single crystal, were solved by molecular replacement in Phaser42 using search models generated from PDB entries 1ASV (CCD)43, 3F9K (NTD)10 and 1EX4 (CTD)44. The models were rebuilt using ARP/wARP45 and/or manually in COOT46 and refined in CNS47. The protein crystals were cryo-protected using mother liquor with 25% glycerol. Data collection 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 g, 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 as separate sessions using Leginon software48 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 1.3 Å at the specimen level. The individual frames were gain corrected, aligned and summed using a geometric and refinement whole-frame alignment program as previously described49,50, and exposure filtered51 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 package48 and were conceptually...
identical to those previously described. 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<sup>59</sup> and Relion<sup>60</sup> algorithms (Extended Data Fig. 1c), and an initial model was generated directly from the class averages using OptiMod<sup>61</sup> (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<sup>60</sup>, after which the parameters were converted for use in Frealign<sup>62</sup>. The final map was refined in Frealign.

Several conformational states of the intasome were observed after three-dimensional classification in both Relion and Frealign<sup>60</sup>. 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<sup>60</sup>, effectively a focused classification of the core region. The best class was resolved to ~1 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<sup>63</sup> starting with rigid-body fitted X-ray structures of individual domains as input. Rosetta protocols were used for all parts of the modelling<sup>64</sup>. 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<sub>1</sub> and CCD<sub>2</sub>, each seeded with six starting X-ray models: independent CCD monomers from chains A–D of the IN<sub>NTD-CCD</sub> structure and monomers A–B of the CCD portions of the IN<sub>NTD-CCD</sub> structures. CTDs 1, 2, 5 and 6 were seeded with subunits A and B of the IN<sub>CTD</sub>-X-ray model. Likewise, for NTD<sub>1</sub> and NTD<sub>2</sub>, the two different NTDs of the IN<sub>NTD-CCD</sub>-X-ray structure were used as input seeds. All models were refined against the core intasome structure resolved to 4–5 Å resolution in the most homogeneous regions using 30,307 particles (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<sub>NTD</sub>, IN<sub>2</sub>, IN<sub>3</sub> and IN<sub>4</sub>, thereby filling in the linker regions between individual domains. Although certain INs corresponded to bulky amino-acid side chains, in particular within NTD<sub>1</sub>–CCD<sub>1</sub>, linker regions were converted for use in Frealign<sup>62</sup>. The final model was refined in Frealign.

Several conformational states of the intasome were observed after three-dimensional classification in both Relion and Frealign<sup>60</sup>. 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<sup>60</sup>, effectively a focused classification of the core region. The best class was resolved to ~1 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<sup>63</sup> starting with rigid-body fitted X-ray structures of individual domains as input. Rosetta protocols were used for all parts of the modelling<sup>64</sup>. 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<sub>1</sub> and CCD<sub>2</sub>, each seeded with six starting X-ray models: independent CCD monomers from chains A–D of the IN<sub>NTD-CCD</sub> structure and monomers A–B of the CCD portions of the IN<sub>NTD-CCD</sub> structures. CTDs 1, 2, 5 and 6 were seeded with subunits A and B of the IN<sub>CTD</sub>-X-ray model. Likewise, for NTD<sub>1</sub> and NTD<sub>2</sub>, the two different NTDs of the IN<sub>NTD-CCD</sub>-X-ray structure were used as input seeds. All models were refined against the core intasome structure resolved to ~1–5 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<sub>NTD</sub>, IN<sub>2</sub>, IN<sub>3</sub> and IN<sub>4</sub>, thereby filling in the linker regions between individual domains. Although certain INs corresponded to bulky amino-acid side chains, in particular within NTD<sub>1</sub>–CCD<sub>1</sub>, linker regions were converted for use in Frealign<sup>62</sup>. The final model was refined in Frealign.
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. d, Initial model from the class averages calculated using OptiMod. 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 $\sim$4–5 Å resolution. Partial separation of β-strands, which is typically evident at or beyond 4.5 Å resolution, is apparent.

© 2016 Macmillan Publishers Limited. All rights reserved
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).
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 INCTD crystal structures described here (PDB accession numbers 5CZ2 and 5D7U, respectively). Symbols $\alpha$, $\beta$, $\eta$, TT and TTT represent $\alpha$-helix, $\beta$-strand, $3_{10}$-helix, $\alpha$-turn and $\beta$-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$_{NDT-CCD}$ 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. d, Cartoon representation of the IN$_{NDT-CCD}$ 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. Active site residues are shown as red sticks.
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 IN1 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 NTD1 and CCD1 and CCD–CTD residues 211–213). Linkers NTD2–CCD2, CCD3–CTD3 and CCD6–CTD6 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 Rosetta15–17. 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 IN2 and IN4 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 IN1 and IN3, which interact with the vDNA (right).
Extended Data Figure 7 | Gel filtration profiles of IN<sub>WT</sub> and IN mutant proteins. Elution profiles of mass standards in kilodaltons as well as theoretical protein monomer (M) and dimer (D) positions are indicated.
Extended Data Figure 8 | Comparisons of PFV and MMTV intasome structures. a, Cartoon representations of the inner IN, 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 (IN₈ 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 | P12₁,1 | C222₁ |
| Cell dimensions |     |         |     |
| a, b, c (Å) | 51.89, 53.71, 69.65 | 54.37, 83.16, 141.14 | 35.89, 42.28, 139.09 |
| a, b, c (Å) | 69.69, 82.08, 63.97 | 90, 90, 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) |
| Rmerge | 0.080 (0.57) | 0.08 (0.534) | 0.043 (0.585) |
| R/σf | 21.0 (2.0) | 9.5 (2.0) | 22.2 (3.8) |
| Completeness (%) | 99.1 (95.6) | 99.3 (98.0) | 98.8 (98.9) |
| Redundancy | 5.2 (2.8) | 3.2 (3.1) | 12.8 (2.8) |
| **Refinement** |     |         |     |
| Resolution (Å) | 32.8 - 1.70 | 70.6 - 2.72 | 40.4 - 1.50 |
| No. reflections used | 69,075 | 32,115 | 17,448 |
| Rmerge/Rfree | 0.189/0.222 | 0.245/0.266 | 0.165/0.202 |
| No. atoms |     |         |     |
| Protein | 4,983 | 9,110 | 890 |
| Ligand | 0 | 12 | 8 |
| Water | 437 | 0 | 69 |
| B-factors |     |         |     |
| Protein | 26.0 | 70.9 | 28.5 |
| Ligand | - | 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±0.0 | 1.0±0.0 |
| 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-/Å²) | 46 | 46 |
| 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 (Å²) | -300 | -460 |
| **Model refinement** | | |
| Space group | P1 | - |
| Cell dimensions | | |
| a ≠ b ≠ c (Å) | 151.2 | - |
| a = b = c (°) | 90 | - |
| Number of atoms (modeled) | 11,462 | - |
| **Validation** | | |
| MolProbity score | 1.46 (95th percentile) | - |
| Clashscore, all atoms | 2.27 (99th percentile) | - |
| Protein | | |
| Ramachandran favored (%) | 1,115 (92.76) | - |
| allowed (%) | 87 (7.24) | - |
| Disallowed (%) | 0 (0) | - |
| Good rotamers (%) | 1,039 (99.71) | - |
| CB 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.