# Supplementary Protocols

## Protocol 1: Detailed image processing:

An initial 3-D model of the integrase/LEDGF complex was constructed from a negatively stained data set of 3732 images. Two lines of evidence showed that the data set contains a two-fold symmetry; firstly the eigenvectors obtained upon multi statistical analysis of the starting, unaligned, image dataset indicated that a two-fold symmetry operator was present in the particles. The two first eigenvectors are identical, show two fold symmetry and are rotated one with respect to the other by 45° and secondly, after alignment, several class averages clearly showed a two-fold symmetric projection.

These two observations were instrumental for 3-D reconstruction of the first model since they revealed the existence and the position of the two-fold symmetry axis. Using this information a starting model was constructed using the common lines method and the sinogram correlation function to assign the projection directions and the standard weighted back projection algorithm for 3-D reconstruction.

In order to improve the resolution, a dataset of 11441 negatively stained frozen hydrated particles was recorded and analysed independently. The normalized and floated images were clustered into 1000 class prior alignment after multivariate statistical analysis and hierarchical ascendant classification in order to avoid any bias introduced by the choice of the alignment references. The 1000 class averages were aligned against each other and used as alignment references for the original dataset. This refinement procedure was iterated 4 times in order to obtain a stable clustering. The angular assignment of the class averages was performed by using reprojections of the negatively stained model as anchor sets. Angular assignment was done using the common line methods. After several rounds of refinement, the 3-D model was stable and a good correspondence was found between the input class averages and the reprojections of the 3-D model.

For the structure in presence of DNA, we proceeded the same way up to the point where the 3-D model had to be reconstructed. An independent reconstruction was performed using the two-fold symmetry in order to avoid any bias introduced by the DNA-free model. We further

hypothized that the data set contained not only integrase/LEDGF complexes that had bound DNA but also DNA-free complexes. We therefore used both the DNA-free and the DNA-containing models to sort the original images into two categories: those who are likely to have bound DNA and those who have not. This sorting protocol reduced the dataset by about 25% and yielded the final 3-D model.

**Protocol 2: IN-LEDGF-DNA complex preparation:**

The HPLC-purified synthetic U5A and U5B were purchased from Eurogentec S.A. Double-stranded U5-substrate was obtained by mixing equimolar amounts of the 21-mer U5A

(5'-ACTGCTAGAGATTTTCCACAC)

and the complementary strand U5B

(5'-GTGTGGAAAATCTCTAGCAGT)

in a 10mM Cacodylate-Na buffer (pH6.0) containing 50mM NaCl. The mixture was heated to 90°C for 3 min and annealing was allowed by slow cooling overnight at 4°C. The DNA duplex and the $IN_4$-$LEDGF_2$ complex where mixed in a 10:1 ratio and left 10 mn at room temperature. The preparation was then processed for cryo-EM.

**Protocol 3: Cross-linking experiments**

U5 substrate analogs composed of a 40-mer and two complementary 19-mer and 21-mer (U5B or U5A) oligonucleotides, form a nick upon annealing. An uridine analog containing a 2,3-dihydroxypropyl group in the 2'-position of the sugar ring (**U***), was placed at different positions of both U5B and U5A and a set of dsDNA molecules containing this modification either in the processed or the non-processed strands was prepared and used for cross-linking. Oxidation of the 2,3-dihydroxypropyl group by $NaIO_4$ leads to an aldehyde able to form

Schiff bases with proximal lysine residues (Lys). The cross-linking efficiency was found to be highest when the 6$^{th}$ nucleotide from the U5A end (T6) was modified and therefore all further experiments were performed using this modified substrate. After trypsin digestion and purification, the cross-linked products were separated by gel electrophoresis and the main product was analyzed by mass-spectrometry, revealing one major peak (**Fig S9A)**. The mass of a cross-linked peptide was calculated as 1350 g/mole that is close to two products of integrase trypsin digestion, M(RKAKIIRDYGK) = 1348 and M(DSRDPVWKGPAK) = 1356. Therefore, the conjugate was treated with hydrofluoric acid (HF) to degrade the DNA part. The mass-spectrum after the treatment showed two groups of peaks corresponding to DNA fragments and an intensive peak with the mass 1618 (**Fig S9B)**. Its fragmentation spectrum analyzed in tandem MS/MS mode showed that the cross-linked peptide had the RKAKIIRDYGK sequence corresponding to amino acids 263-273 in the C-terminal domain. This result indicates that the 6$^{th}$ nucleotide from the non-processed strand end is located near Lys264 or Lys266.

*Oligonucleotides preparation:* The 40-mer integrase substrate for 3'-processing (U5-DNA) was composed of oligonucleotides 40-U5B,

5'-GACTACGGTTCAAGTCAGCGTGTGGAAAATCTCTAGCAGT-3'

and 40-U5A,

5'-ACTGCTAGAGATTTTCCACACGCTGACTTGAACCGTAGTC-3',

which were synthesized using a 380B Applied Biosystems synthetizer by the standard cyanoethyl phosphoramidite procedure.

DNA duplex 40-U5ald used for the cross-linking experiment was composed of a 40-mer oligonucleotide 40-U5B, 19-mer oligonucleotide 19-U5A, 5'-GTCGACTTGAACCGTAGTC-3' and 21-mer oligonucleotide U5A6-ald, 5'-ACTGC(**U*)**AGAGATTTTCCACAC-3' containing 2,3-dihydroxypropyl group in the 2'-

position of the uridine residue U* which was synthesized as in (Zatsepin *et al.*, 2002). All oligonucleotides were further purified on 20% denaturing acrylamide/urea gel. The U5 substrate analogs were composed of one 40-mer and two complementary oligonucleotides, 19-mer and 21-mer, U5B or U5A. We verified that the nick presence had no effect on the 3'-processing efficiency (data not shown).

*MALDI-TOF mass-spectrometry analysis:* Mass spectra were recorded in positive ion mode on a Bruker Ultraflex TOF/TOF instrument (Bremen, Germany). The peptide samples were analyzed in reflectron mode using 2, 5-dihydroxybenzoic acid (DHB) as a matrix; modified oligonucleotides and their derivatives with cross-linked peptides were analyzed in linear mode using 2,4,6-trihydroxyacetophenone containing di-ammonium citrate (THAP/DAC) as a matrix.

**Protocol 4: Detailed fitting procedure:**

To fit atomic structures in the low resolution cryoEM map we used normal mode flexible fitting (NMFF) [64,65].

The procedure we used has been validated by [64] in a similar context by fitting the Xray structure of GroEL in its open conformation into the 15 Å resolution cryo negatively stained EM envelope of its closed structure and by comparing these fitting results to the known atomic structure of the closed form. In our case, this procedure mainly results in a curvature of the α-helix linking the C-terminus to the catalytic core, the structures of the domains remaining unchanged. The fitting of the N-terminus was done by rigid body.

The fitting was performed in a multi step approach to avoid trapping in local minima. We therefore computed in the first step the optimal fit using the lowest frequency mode with NORMA (mode 7), but we applied only 30% of the corresponding amplitude to generate a first intermediate model. This step was completed by a model regularization using REFMAC. The resulting model was then used to initiate the second step, where we now applied 50% of the computed optimal perturbation followed by model regularization. A third step followed

where 100% of the perturbation was applied. At this stage the N-terminus part was introduced in the model and manually placed in the density left empty after the previous fit and its position was then refined by rigid body fitting. A last step of fitting was done using 12 normal modes (mode 7 to 18) using a map masked around the molecule positions. The correlation coefficient between the model and the structure increases from 0.531 (model 0) to 0.83 (model 5) (**Supplementary table 1**).

For the IN/LEDGF/DNA structure, model 4 was used as a starting model. It was first fitted in density using the O program. A map masked around this molecule position was used to exclude DNA density for fitting. A first round of one-mode fitting was done and 30 % of the amplitudes were applied and the resulting structure was regularized with REFMAC and used for a new one-mode fitting. 50 % of the amplitudes computed from this last step were used for a 12 mode fitting and the resulting model was calculated with 100% of the amplitudes. A last step of rigid body fitting was used to refine the N-terminal positions. The correlation coefficient between the model and the structure increases from 0.471 (model 4) to 0.772 (model 9) (**Supplementary table 2**). DNA molecules were placed in the empty density remaining in the top region with the O program. The models at each step of the fitting procedure are shown in **Supplementary movies 1, 2 and 3**. The procedure is summarized in **Supplementary figure 6 and upplementary tables 1, 2**.

The fitting of the composite model in IN/LEDGF map resulted in a small displacement of one C-terminus of the IN dimer (**Supplementary figure 7 and movie 1**). The N-terminus where fitted by rigid body in the remaining empty density. The positions of the Cter parts of the two N-terminuses are at a distance of the Nter of the catalytic core in agreement with the 10 residues missing in the structure. Fitting of the model in the IN/LEDGF/DNA map resulted in a displacement of the C and N-terminus (**Supplementary figure 8 and movie 2**). DNA fitting was done manually by rigid body translation using a difference map between the envelope and the protein model (**Supplementary movie 3**).

The NMFF procedure resulted in small displacement of the C-terminus part of the IN structure, mainly bending the linking helices between the catalytic core and the C terminal domain. The positions of the N-terminus domains where refined by rigid body and are at a distance of the Nter of the catalytic core in agreement with the 10 residues missing in the structure. The final positions of the domains and DNA are validated by the cross linking and mutational experiments. The movements of the C-terminus and N-terminus between the initial and final models are summarized in **Supplementary table 2**.