The HML2 (HERV-K) group constitutes the most recently acquired family of human endogenous retroviruses, with many proviruses less than one million years old. Many maintain intact open reading frames and provirus expression together with HML2 particle formation are observed in early stage human embryo development and are associated with pluripotency as well as inflammatory disease, cancers and HIV-1 infection. Here, we reconstruct the core structural protein (CA) of an HML2 retrovirus, assemble particles in vitro and employ single particle cryogenic electron microscopy (cryo-EM) to determine structures of four classes of CA Fullerene shell assemblies. These icosahedral and capsular assemblies reveal at high-resolution the molecular interactions that allow CA to form both pentamers and hexamers and show how invariant pentamers and structurally plastic hexamers associate to form the unique polyhedral structures found in retroviral cores.
The human genome contains a large number of endogenous retroviruses (ERVs) that provide a fossil record of human–pathogen interactions over millions of years1–3. In most instances, ERV open reading frames (ORFs) are degraded, containing stop codons, deletions and rearrangements. However, several members of the HML2 group of ERVs have proviruses less than one million years old4–6 with intact ORFs5 and HML2 particle formation is observed in early-stage human embryo development8 and is associated with pluripotency9, as well as inflammatory disease10–12, cancers13–16 and HIV-1 infection17–19. Moreover, viruses created by reconstruction of HML2 consensus sequences have been shown to produce infectious particles20,21.

Among the circulating exogenous retroviruses, the Gag polyprotein is processed into the matrix (MA), capsid (CA) and nucleocapsid (NC) proteins that form the structural layers within a mature retroviral particle. CA forms the capsid shell surrounding the viral core that protects and transports the viral genome and also interacts with host cell factors that are necessary for trafficking, nuclear entry and proviral integration22–24. Given these essential functions of the capsid, several structural studies have been undertaken that have provided the molecular details of the capsid’s individual hexameric and pentameric CA building blocks25–28. In addition, electron cryomicroscopy studies of whole retroviral cores and in vitro assemblies have provided insight into capsid assembly at low and medium resolution29–34. However, despite these advances the pleomorphic nature of retroviral capsids has so far confounded attempts to determine high-resolution structures of entire closed capsid shells containing both hexameric and pentameric subunits.

Here we express and purify CA (HML2 CArec) from a consensus HML2 Gag ERV20,21 and determine crystal and solution nuclear magnetic resonance (NMR) structures of the N-terminal (CArec-NTD) and C-terminal (CArec-CTD) domains, respectively. Further, we assemble particles from HML2 CArec in vitro and use single-particle cryogenic electron microscopy (cryo-EM) to determine the high-resolution structures of four different types of Fullerene shell assemblies. Our data reveal that the structures of the (CArec-NTD) and C-terminal (CArec-NTD) domains are largely conserved with those of CA from exogenous retroviruses. Analysis of the shell structures reveals the intra- and inter-molecular interactions that drives CA assembly into pentamers and hexamers and their association into shells that encapsidate the retroviral genome.

Results

In vitro assembly and cryo-EM structures of HML2 CA shells. To assess if human ERVs have retained the capacity to assemble CA into shells, we synthesised a codon-optimised consensus HML2 CA coding sequence20,21 (Supplementary Fig. 1) and expressed the protein, HML2 CArec, in Escherichia coli. The addition of high salt (>1 M NaCl) to purified HML2 CA induced the formation of high-molecular mass particles (Fig. 1a) and a strong concentration dependency of the size distribution (1.5–2.5 MDa) measured by size-exclusion chromatography (SEC) coupled to multi-angle laser light scattering (SEC-MALLS) demonstrated the presence of a heterogeneous mixture (Fig. 1b). Further resolution of the assembly reaction using sedimentation velocity revealed major 32S and 42S species together with a minor 50S component (Fig. 1c). Analysis by cryo-EM revealed a distribution of regular particles, comprising a large proportion of small 160 Å diameter spherical particles and additional larger capsular particles of up to 300 Å in the largest dimension (Fig. 1d).

Cryo-EM combined with single-particle image reconstruction was employed to determine the structure of these HML2 CArec particles. After two- (2D) and three-dimensional (3D) classification of particle images, the structures of four types of particle were resolved to high-resolution (Fig. 2 and Supplementary Table 1). Details of classification, image processing and map resolution are presented in (Supplementary Figs. 2–5). The resolution of the maps varies from 2.7 to 4.3 Å, but show clear density for backbone and side chains and are of sufficient quality for de novo building of atomic models (Fig. 2a and Supplementary Fig. 6). The majority species comprises 12 HML2 CArec pentamers with all interfaces equivalent, obeying perfect T = 1 icosahedral symmetry (Fig. 2b). The structure is resolved at the highest resolution (2.7 Å), likely corresponding to the 32S species observed in sedimentation experiments. In addition, there are a smaller number of D5 and D6 symmetric capsular structures (Fig. 2c, d) and a T = 3 icosahedral particle (Fig. 2e) that together likely constitute the faster sedimenting 42–46S species. The D5 particle is an elongated or capsular expansion of the T = 1 particle with polar regions identical to T = 1, but containing five equatorially inserted distorted hexamers, consistent with local two-fold (dihedral) symmetry (Fig. 2c). The D6 particles are also capsular with fully six-fold symmetric hexamers at the poles each surrounded by a ring of six pentamers and with six equatorial hexamers with similar distortions to those observed in the D5 structure (Fig. 2d). The largest structure is a T = 3 particle (Fig. 2e), resolved to 4.3 Å and contains a T = 1 particle within the interior, although no fixed orientation between the shells is apparent (Supplementary Fig. 5a, b). The T = 3 shell is 300 Å in diameter and is comprised of 12 pentamers, each surrounded by five hexamers. Hexamers are located at icosahedral three-fold axes and interact across icosahedral two-fold axes. This combination of the four structures describe, at high resolution, the molecular interactions within CA pentamer and hexamer arrangements, as well as the inter-pentamer, pentamer–hexamer and inter-hexamer interactions that participate in the formation of closed Fullerene shell structures in vivo.

To guide the structural interpretation of these particles at higher resolution, we determined the 1.8 Å crystal structure of the CArec N-terminal domain (CArec-NTD) (Supplementary Fig. 7a and Supplementary Table 2) and the solution NMR structure of CArec C-terminal domain (CArec-CTD) (Supplementary Fig. 7b, c and Supplementary Table 3). CArec-NTD has the same N-terminal β-hairpin and six-helix fold observed in exogenous orthoretroviral CA proteins. CArec-CTD comprises the same four-helix fold found in exogenous orthoretroviral CA-CTDs, but with an additional C-terminal distal helix (α11, residues A237–E248) that associates with the core of the CTD domain. Importantly, although α11 displays the sequential and medium-range inter-residue nuclear Overhauser effect (NOE) cross-peaks (Supplementary Fig. 7d) that define a helical conformation, the higher T2, lower T1 and lower heteronuclear 1H–15N NOE values observed in NMR relaxation experiments (Supplementary Fig. 8) together indicate it is more dynamic than the core domain (α7–α10). These high-resolution structures could be docked directly into the cryo-EM density and, with the additional building of the NTD–CTD linkers, the NTD–CTD interface and helix α8 at the CTD–CTD interface (Supplementary Fig. 6), refined well into each particle structure. Exceptions were the N-terminal β-hairpin that contains a large loop insertion and is more mobile than in the crystal structure, and the NTD–CTD linker region of D5 and D6 equatorial hexamers that displayed weaker density. In addition, the CTD α11 was not visible in these structures of mature particles, but given its proximity to the C terminus of CA, it may have a structural role in immature Gag assembly, similar to that observed for the SP1 spacer of HIV-135 and CAH helix of murine leukaemia virus (MLV) Gag2,36.

Structural superimposition with exogenous retroviral CA domains revealed the strongest similarity of CArec-NTD with the CA-NTD of the beta and alapharetroviruses Jaagsiekte sheep retrovirus (JSRV) and Rous sarcoma virus (RSV) (Z-scores 10.7
CA conformational switching in pentamers and hexamers. The T = 1 particle comprises 12 HML2 CA\textsuperscript{rec} pentamers. The NTDs of each pentamer are associated proximal to the five-fold axis and form a layer of structure at a radius of 86.3 Å (NTD centroid distance) from the particle centre. Beneath the NTDs, the CTDs form a layer of structure at a radius of 86.3 Å (CTD centroid distance), and more distal to the five-fold axis. The NTD layer does not contribute to inter-protomer interactions and so the entire inter-protomer interface is mediated through inner cage CTD–CTD interactions across icosahedral two-fold axes (Fig. 2b). Comparison of the HML2 CA\textsuperscript{rec} pentamer to CA pentamers studied at lower resolution for HIV\textsuperscript{29} and RSV\textsuperscript{33,37} shows a good alignment of CTDs, that they occupy similar radial positions and employ the same inner cage CTD–CTD interactions at icosahedral two-fold axes. However, the relationship between the NTD and CTD layers in the different retroviral genera varies substantially (Fig. 3). In HML2, the vertical displacement between the NTD and CTD, as judged by the centre of mass (CoM) of the two layers, is 21.6 Å. In the RSV pentamer, this increases to 26.6 Å, and in the HIV-1 pentamer, the layers are even further apart at 29.1 Å (Fig. 3a–d). Therefore, although the underlying CTD sub-lattice and individual domain structures (Supplementary Fig. 9) are conserved across retroviral genera, the differences in the NTD–CTD posture appears to be genera specific.

In the four shell types, regardless of the surrounding environment, the pentamers adopt an invariant configuration (root-mean-square deviation (RMSD) over 1082 Ca = 0.91 Å) and so a key question is how the CA\textsuperscript{rec} monomer is able to assemble into both pentamers and hexamers. In pentamers and hexamers, the packing of helices α1–α2–α3 around the centre of the ring mediates the inter-protomer NTD interface (Fig. 4a, b), and similar to HIV-1\textsuperscript{26} and MLV\textsuperscript{25}, the interface is largely polar, containing solvent facing accessible channels. However, a comparison of the pentamer (Fig. 4a) with the D6 polar hexamer (Fig. 4b), which obeys strict six-fold symmetry, shows that although helix α1–α2–α3 packing is maintained, in hexamers, the CA\textsuperscript{rec} monomers are positioned at a greater radius and are rotated away from the symmetry axis, relative...
to pentamers. As a result, the relative tilt angle between adjacent α1 helices is reduced and helix α2 comes closer to α3 of the neighbouring monomer allowing the additional monomer to be accommodated into the hexameric ring (Fig. 4c and Supplementary Movie 1).

The transformation from pentamer to hexamer also involves remodelling of both the intra-monomer NTD–CTD interface and the CTD interaction with the NTD of the neighbouring monomer in either the pentameric or hexameric ring. In the pentamer, residue R143 located at the C terminus of α6 in the NTD makes a hydrogen bond to the main-chain carbonyl of K182, located at the C terminus of α7 in the CTD. This configuration facilitates formation of a specific hydrogen bonding arrangement between CTD residues D171 on α7 and K229 on α10 with S79’ and D90’ on α4 of the neighbouring monomer NTD (‘ denote a neighbouring monomer) (Fig. 4d) and defines the pentamer conformation. In the D6 hexamer, because of the expanded radius, for the CTD to maintain the interaction with the neighbouring NTD it is rotated by a further 15° and there is an accompanying 2 Å shift and tilt between α6 and α7 (Fig. 4e). This results in an alternative hydrogen-bonding configuration that defines the hexamer conformation where R143 now makes an intra-monomer hydrogen bond with the side chain of Q181 on the penultimate turn of α7, D171 is hydrogen bonded with S79’ but with a shifted geometry and the K229-D90’ interaction is absent (Fig. 4f). Therefore, by selecting one of these hydrogen-bonding configurations, CArec monomers can adapt to either the pentamer or hexamer conformation. Examination of the sequence conservation of residues making intra-monomer and inter-monomer NTD–CTD interactions (Fig. 5) reveals substantial sequence conservation within the betaretroviruses, extending to a limited extent in the alpharetroviruses. However, conservation of K229-D90’ is betaretrovirus exclusive and appears related to the genera-specific NTD–CTD posture (Fig. 3) that allows the close approach of the CTD to the C-terminal of helix α4 in the neighbouring NTD.

In the T=3 particle, there are no inter-pentamer interactions, instead each pentamer is surrounded by five hexamers. The hexamers contact both pentamers and hexamers through the CTD–CTD interactions with three alternating CArec protomers linking to T=3 pentamers and the other three linking to T=3 hexamers (Fig. 6a). These non-equivalent contacts are accommodated by relative shifts of adjacent CA monomers reducing the hexamer to three-fold symmetry (Fig. 6b), consistent with its location at the icosahedral three-fold axis. Compared to the symmetrical D6 polar hexamer, this creates pairs of monomers with a repacking of their NTD–NTD interfaces and associated shifts in

---

**Fig. 2 Structure of HML2 CArec monomer and assembled particles.**

- **a** (Left) Cartoon representation of the HML2 CArec monomer from the T=1 particle. The CArec-NTD, residues P1 to S153, and CArec-CTD, residues A154 to Q246, are coloured in pink and cyan, respectively. (Right) Experimental electron potential map (grey mesh) and model for the region indicated between α6 and α6’. Residues G152 to P156 in the linker region were built de novo into the density. **b** (Left) Surface representation of the T=1 particle (60 monomers arranged as 12 pentamers) viewed along the five-fold symmetry axis with the NTDs and CTDs coloured as in a. (Right) The inner CTD “cage” viewed along the five-fold (upper) and the two-fold (lower). **c** (Left) D5 particle (90 monomers arranged as 12 pentamers and 5 hexamers) viewed along an equatorial pseudo-hexamer at the two-fold symmetry axis. Pentamer NTDs and CTDs are coloured as in a, and hexamer NTDs and CTDs are coloured in red and teal, respectively. (Right) Central slab of the D5 density map (grey surface) with the T=1 particle aligned on the D5 polar pentamer. **d** D6 particle (108 monomers arranged as 12 pentamers and 8 hexamers) viewed along an equatorial pseudo-hexamer at the two-fold symmetry axis. Pentamers and hexamers are coloured as in c. **e** T=3 particle (180 monomers arranged as 12 pentamers and 20 hexamers) viewed along an icosahedral two-fold symmetry axis. Pentamers and hexamers are coloured as in c.
the location of the CTD and its interaction with the NTD of the neighbouring monomer (Fig. 6c–e). When the CTD–CTD interaction is with a pentamer, the intra-hexamer CTD–NTD interaction has the same structural conformation and specific hydrogen bonding arrangement as the T = 1 pentamer conformation (Fig. 6c and Fig. 4d). Notably, all the pentamers in the T = 3 particle also maintain the same intra-pentamer hydrogen bonding arrangement as the T = 1 pentamer, even though each monomer now makes identical interactions with surrounding hexamers rather than pentamers. By contrast, when the CTD–CTD interaction is with a hexamer, the CTD shifts away from its NTD by 3.6 Å, the R143-mediated α6–α7 interaction is lost and the intra-hexamer CTD–NTD interaction has a structural conformation and hydrogen-bonding configuration that is comparable to that of the D6 hexamer (Fig. 6d, e). Thus, the same structural remodelling of inter-monomer interfaces that switches the pentamer to the hexamer conformation is exploited within the T = 3 hexamer to adapt to either the hexamer–pentamer or hexamer–hexamer environment.

This symmetry breaking of the T = 3 hexamer is also manifest at the level of whole particle geometry. Measurements of the radial distance of the centroid for T = 3 pentamers and hexamers show that pentamers are situated at a higher radius than hexamers, 143.9 and 136.5 Å, respectively. This difference in displacement is accommodated in T = 3 hexamers by the alternating shift-up of the CTDs (3.3 Å) to interact with adjacent pentamers and shift-down (3.3 Å) to interact with adjacent hexamers (Fig. 6f). Moreover, adaptation of hexamers to the local environment is also a feature of our other D5 and D6 capsular assemblies, where hexamers interact with each other in the equatorial belt. Again, the pentamer conformation remains invariant and the hexamer six-fold symmetry is broken by NTD domain movements, changes in the NTD–CTD linker conformation and in the intra-hexamer CTD–NTD hydrogen bonding configuration (Supplementary Fig. 10) and in the CTD radial position when interacting with either a neighbouring pentamer or hexamer. These observations regarding the plasticity of the hexamer structure are key when considering the pleotropic assemblies that constitute viral cores and

![Fig. 3 Comparison of HIV-1, RSV and HML2 CA pentamers.](image-url)

**Fig. 3 Comparison of HIV-1, RSV and HML2 CA pentamers.** a–c Top and side view cartoon representations of CA pentamers with α-helices shown as cylinders, from a HML2 CA<sup>α95</sup>, b RSV (built from PDB: 1EM9 and 3G1I docked into EMD 5772) and c HIV-1 (PDB: 5MCY). NTDs are coloured pink, red and purple and CTDs are coloured light cyan, light blue and cyan for HML2, RSV and HIV-1, respectively. Pentamers are aligned with respect to their CTDs. In the lower panels, the distance between the centre of mass (CoM) for each NTD and CTD ring is shown on the right. d Central section through CA pentamers aligned with respect to CTDs. The CoM for NTD and CTD rings is represented by the spheres, colour coded as in a–c and shows the coincident CoM for CTDs and the difference in CTD–NTD vertical displacement for each pentamer.
provide an explanation of how CA accommodates the extremes of curvature that are encountered in these structures.

**Building shells through CTD–CTD interactions.** Given the importance of the CTD inner cage, we also analysed the CTD–CTD interactions that connect pentamers to pentamers, hexamers to pentamers and hexamers to hexamers in all of the Fullerene structures (Fig. 7). To undertake this analysis and as a basis for comparison, we first determined the NMR ensemble structure of the solution dimer and measured HML2 CArec-CTD...
The solution structure CTD–CTD interface involves an outer network of polar interactions (residues S157-K214, N159-R162-E200′-N201′ and D186-K188) together with the packing of a core of hydrophobic side chains (residues: A189, V192, I193 and V194, L196, M197) that extend from opposing α8 helices (Fig. 7a).

Residues at the CTD interface are largely conserved amongst the betaretroviruses (Fig. 5) and introduction of a I193A/L196A mutant, analogous to the W184A/M185A that disrupts in HIV CTD–CTD interactions26, abolishes self-association of HML2 CArec-CTD in solution (Supplementary Fig. 11e).

In this configuration, the α8 helices are oriented with a shallow relative crossing angle of 20° and a helix centroid displacement of β1-β2-α1-α2-α3.

Fig. 5 Structure-based sequence alignment of HML2 CArec with alpha and betaretroviral CAs. The alignment was performed domain-wise using PROMALS3D, numbering refers to the HML2 CArec sequence and Uniprot accession codes for each aligned sequence are shown on the left. The five betaretroviral: HML2 CArec, JSRV, MPMV, mouse mammary tumour virus (MMTV), Simain type D retrovirus (SRV1) and squirrel monkey retrovirus (SMRV), and two alpharetroviral: RSV and avian leukosis virus (ALV) CAs are grouped separately (upper and lower). The regions corresponding to HML2 CArec secondary structures are shaded red for α-helices and green for β-strands and also displayed pictorially above the sequences. Residues with ≥50% identity are highlighted in black, and ≥50% similarity in grey. The major homology region (MHR) identified in the CA of all the orthoretroviruses is boxed. Residues in HML2 CArec that make inter-molecular CTD-CTD, NTD-CTD or intramolecular NTD-CTD contacts and those that form the conserved NTD β-hairpin hydrogen bond are indicated with the coloured asterisks (black, green, red and blue, respectively).
Fig. 6 Variation of hexamer–pentamer and hexamer–hexamer interfaces in the $T = 3$ particle. 

**a** A hexamer–pentamer–hexamer subassembly from the $T = 3$ particle shown in surface representation viewed from the exterior of the shell (upper) and from the side, highlighting the curvature (lower). NTDs and CTDs in the pentamer and those in the hexamer that interface with pentamers are coloured pink and cyan, respectively. NTDs and CTDs in the hexamer that interface with hexamers are coloured red and teal, respectively. 

**b** Cartoon representation of a hexamer from the $T = 3$ particle viewed down the icosahedral three-fold axis (colouring as in **a**). 

**c, d** NTD–CTD intra-hexamer interactions in the $T = 3$ hexamer. 

**c** A CA monomer, NTD (pink) and CTD (cyan) that interfaces with a neighbouring pentamer, and the adjacent NTD shown in red. 

**d** A CA monomer, NTD (red) and CTD (teal) that interfaces with a neighbouring hexamer, and the adjacent NTD is shown in pink. Residues making interactions are shown as sticks with hydrogen bonds shown as dashes, the prime (‘) notation indicates the adjacent NTD and “mc” indicates a main-chain interaction. 

**e** NTD (pink/red) alignment of monomers in $T = 3$ hexamers showing the relative displacement of the CTD when the protomer interacts with either a neighbouring pentamer (cyan) or a neighbouring hexamer (teal). 

**f** Hexamer surface representation viewed perpendicular to the symmetry axis showing alternate CTDs displaced upwards (cyan, pentamer facing) and downwards (teal, hexamer facing). NTDs are coloured pink or red as in **a**.
Although much of the polar network is lost, a new compensatory phobic interactions of A189, V192 and I193 are preserved, and icosahedral particle geometry, although remarkably, the hydro-

ergized to 65°. This has the effect of twisting the interface

Fig. 7 Analysis of HML2 CA\textsuperscript{occ} CTD–CTD interactions. a, b Structures of the CTD–CTD dimer interface in solution (lowest-energy NMR structure) (a) and in the \( T = 1 \) particle (b) are shown in cartoon representation. Residues involved in the interface are shown as sticks with hydrogen bonds shown as dashes. The boxed region is rotated 90° to show the hydrophobic interactions between helix \( \alpha 8 \) of each monomer, with the van der Waals radii shown as grey space fill. c Plot of the cross angles and displacements between \( \alpha 8 \) helices in all CTD–dimer pairs. The circles are colour-coded with respect to particle type, shown in the key. Groups 1, 2 and 3 are ringed and a schematic describing the nomenclature of POLAR, LONG and LAT is shown alongside. The class of interaction seen in each particle type and the group to which they belong is shown in the key below. d Panel of representative CTD dimers from the groups displayed in c. Top to bottom: NMR dimer (group 1); \( T = 1 \) pentamer (group 2); \( T = 3 \) pentamer on hexamer (group 2); \( T = 3 \) hexamer on hexamer (group 3). e Alignment of CTD monomers (grey) showing relative displacement of the interacting CTD monomer in \( T = 1 \) (red; group 2) and \( T = 3 \) hexamer on hexamer (orange; group 3).

11 Å, allowing extensive hydrophobic contacts along their entire length. By comparison, at the unique type of CTD–CTD interface in the \( T = 1 \) particle (Fig. 7b) the \( \alpha 8 \) centroid displacement is maintained at 12 Å, but the crossing angle is significantly enlarged to 65°. This has the effect of twisting the interface relative to the solution state to meet the constraint imposed by icosahedral particle geometry, although remarkably, the hydrophobic interactions of A189, V192 and I193 are preserved, and although much of the polar network is lost, a new compensatory S157–E200–N159 interaction contact is present.

Parameterisation of all the CTD–CTD interactions observed in \( T = 1, T = 3, D5 \) and D6 shells in terms of the crossing angle and \( \alpha 8 \) centroid displacements provides a way to understand the detailed conformations that CTD–CTD dimers adopt within the particles. These data (Fig. 7c, d) show that the NMR solution ensemble populates one region, group 1, and all family members have \( \alpha 8 \) crossing angles near to 20° and displacements close to 11 Å. The CTD–CTD dimers in the particles fall into two other classes: group 2 that contains all CTD–CTD interactions at pentamer–pentamer and pentamer–hexamer interfaces and have
crossing angles and displacements between 11 and 15 Å, and group 3 that contains all the CTD–CTD interactions at hexamer–hexamer interfaces and have crossing angles near to 95° and displacements between 13 and 16 Å. T = 3 particles have two CTD–CTD configurations, a group 2 set from hexamer–pentamer interfaces with a similar configuration to that observed at T = 1 pentamer interfaces and a group 3 set with the larger crossing angle found at hexamer–hexamer interfaces (Fig. 7d, e). In the D5 and D6 particles, further configurations are possible depending on whether pentamers and hexamers are located at poles or are equatorial and interact either longitudinally or laterally (Fig. 7c). Nevertheless, all hexamer–pentamer interfaces still fall into group 2, and all hexamer–hexamer interfaces still fall into group 3, albeit in D6, the α8 displacement for the equatorial hexamer–hexamer interactions are larger than in D5, due to the lower curvature of the equatorial region.

Previously, a coarse-grained theoretical model for a larger HIV-1 fulleren e one-shaped capsid was constructed with a continuum of α8 crossing angles covering the range of groups 2 and 3 observed in HML2 CArec. Our data now demonstrate how discrepancy of crossing angles between CTDs facilitates adaptation of CA to the different environments, particularly at the points of curvature that pentamers introduce into the structures.

Discussion

Our studies of HML2 CArec show that an ERV retains the ability to form capsid assemblies using the same architecture as exogenous viruses. The full diversity of interactions we observe in our structures reveals (1) plasticity of packing at the NTD–NTD interfaces, (2) alternative geometry and bonding configurations at the NTD–CTD interfaces and (3) the quantisation of the CTD–CTD association parameters. Plasticity at the NTD–NTD interface has been observed previously in crystal structures of bovine leukaemia virus CA, where hexamers are distorted from perfect 6-fold geometry to allow the CTD–CTD interfaces to maintain a hexagonal lattice that would otherwise be dislocated by crystal packing forces. Further, flexibility has been observed in intact HIV-1 virion cores where small rotations at the interdomain linker and CTD–CTD interface are distributed over the whole core to accommodate differences in curvature of the Fulleren e cone. Now our high-resolution structures enable us to classify and quantify these molecular interactions and provide a rule book for HML2 CArec pentamer and hexamer construction. Moreover, they define how switching of the intra-hexamer CTD–CTD configurations facilitates symmetry breaking of the hexamers to adapt to distinct local conditions by positioning the CTD for interaction with the CTD of the adjacent pentamer or hexamer. This provides the basis for building structures with a wide range of shapes and sizes consistent with Fulleren e geometry and more pleomorphic morphologies found in exogenous betaretroviruses, the highly similar alpharetroviruses and the more distant lentiviruses.

Methods

Details of constructs. A DNA sequence coding for the consensus HML2 CArec protein was synthesised codon optimised for E. coli expression (Genentech) (Supplementary Fig. 1). The sequence coding for the full-length (HML2 CArec, P1-Q246), amino terminal (P1-P151; CArec-NTD) and C-terminal domain (P156-Q246; CArec-CTD) were amplified by PCR and the products inserted into a pET22b expression vector (Novagen) between the NdeI and XhoI restriction sites in order to produce C-terminus hexa-histidine fusions. The 193Aa/L196A double mutation was introduced into CArec-CTD using the Quikchange II Site-directed Mutagenesis Kit (Agilent) according to the manufacturer’s instructions. The correct sequence of expression constructs was verified by automated DNA sequencing (Beckman Coulter Genomics), primer sequences used for PCR cloning and mutation are presented in Supplementary Table 5. Proteins were expressed in the E. coli strain BL21 (DE3) by the addition of 0.1 mM isopropyl β-D-thiogalactopyranoside (IPTG) to log-phase cultures followed by continued growth at 20 °C overnight. The bacteria were harvested and resuspended in 10 mL of lysis buffer per gram of cells (50 mM Tris, pH 8, 250 mM NaCl, 10 mM imidazole, 5 mM MgCl2), 0.5 mM TCEP (Tris(2-carboxyethyl)phosphine) and 0.2% v/v Triton X-100. The cells were lysed by sonication and the clarified supernatant was injected onto a 5 mL HisTrap Column (GE Healthcare). Bound sample was washed with wash buffer (50 mM Tris, pH 8, 750 mM NaCl, 20 mM imidazole, 5 mM MgCl2, 0.5 mM TCEP, 0.2% v/v Triton X-100, 4 mM ATP) and eluted with His B buffer (20 mM Tris, pH 8, 250 mM NaCl, 10 mM imidazole, 0.5 mM TCEP). For HML2 CArec-CTD, the eluent was concentrated to ~3 mL and the protein further purified by SEC on a Superdex 75 (26/60) column equilibrated in SEC buffer (50 mM Tris, pH 8, 100 mM NaCl, 0.5 mM TCEP). For HML2 CArec-NTD the eluents from HiTrap column were diluted 2.5-fold in IEX A buffer (20 mM Tris pH 8, 0.5 mM TCEP) and applied a 6 mL Resource Q ion exchange column. Proteins were eluted using a 40 column-volume gradient into IEX B buffer (20 mM Tris pH 8, 1 M NaCl, 0.5 mM TCEP). Fractions containing HML2 CArec or HML2 CArec-NTD were concentrated to ~3 mL and further purified by SEC on a Superdex 200/26/60 column equilibrated in SEC buffer. All purified proteins were concentrated to 15–30 mg mL−1, flash frozen in liquid nitrogen and stored at ~80 °C. Selenium was incorporated into the N-terminal domain CArec-NTD construct by replacement of methionine with selenomethionine in defined culture medium and by inhibition of methionine biosynthesis just prior to IPTG induction38. For CArec-CTD NMR experiments, 15N and 13C-2H uniformly labelled protein was expressed in M9 minimal media with 13NH4Cl or 15NH4Cl and 13C-glucose, as required, as sole nitrogen or nitrogen and carbon sources. To obtain triple-labelled, 1H-13C-15N samples, the M9 media containing 13NH4Cl and 13C-glucose was prepared in 2H2O instead of H2O. Isotope-labelled and selenium incorporated samples were purified in the same way as unlabelled protein. Verification of N-terminal methionine processing, correct molecular mass, degree of selenium and isotopic label incorporation was obtained by electrospray ionisation mass spectrometry.

Cryo-EM sample preparation and data collection. HML2 CArec (16 mg mL−1) was adjusted with high salt buffer (20 mM Tris-HCl, 5 M NaCl, 0.5 mM TCEP, pH 8.0) to final salt and protein concentrations of 1.4 M and 10 mg mL−1, respectively. Samples were incubated at 4 °C for 1 h prior to plunge freezing. Quantifoil R2/2 200 mesh copper grids were prepared by glow discharge at 25 mA for 1 min in air and washed with ethanol (96%). All grids were filmed on a FEI Tecnai G2 F20 (300 kV) at 0.75% relative humidity. Two-microtitre sample was added to carbon side of grid and incubated for 30 s in the Vitrobot chamber before blitting for 5 s. Immediately after, 2.2 µL of low salt buffer (20 mM Tris, pH 8, 100 mM NaCl) was applied, followed by a 5 s blot and plunge freezing into liquid ethane. Data from frozen-hydrated samples were collected on a Titan Krios (Thermo Fisher) in nanoprobe mode on a Falcon III detector operating in counting mode. Movies were collected at a magnification corresponding to a calibrated pixel size of 1.09 Å. Movies were recorded with an exposure time of 59.4 s, corresponding to 30 frames and a total dose of 30 e− Å−2.

Image processing. All movies were motion corrected on the fly using Scipion v1.139 with exposure weighting in MotionCor240. Initial parameters for the two-dimensional electron density maps were generated from a subset of particles (20,000) using cryoSPARC. From the best class were then subject to manual refinement in RELION 2.1 imposing I2 symmetry yielding 1,218,119 negatively stained and 1,218,119 negatively stained 2D particles for further analysis. The particle classes were further subdivided based on particle diameter guided by known sizes for spherical capsular particles from prior negative stain studies: 715,082 (T = 3). A total of 1,498,437 particles from 10,935 dose-weighted micrographs were identified by template-based particle picking using RELION 2.1. Two cycles of reference-free 2D classification yielded 1,218,119 particles for further analysis, and all particles were further subdivided based on particle diameter guided by known sizes for spherical capsular particles from prior negative stain studies: 715,082 (T = 1), 211,787 (D5 and D6e) and 398 (T = 3).

For the reconstruction of the T = 1 particle, an ab initio model without symmetry was generated from a subset of particles (20,000) using cryoSPARC241. Two cycles of 3D classification were carried out to identify best particles for subsequent refinement. 3D auto-refinement in RELION 2.1 imposing I2 symmetry was performed on 121,357 particles belonging to the most populated, highest-resolution 3D classes, followed by 3D classification without alignment and auto-refinement of the highest resolution class. Particles and the map corresponding to the highest resolution final refinement were further docked using cistEM44 with an initial high-resolution limit for refinement set to 10 Å and final limit of 5 Å. The final reconstruction, performed on 64,731 particles extracted into
a 576 × 576 pixel box, has a resolution of 2.75 Å (0.143 Fourier shell correlation (FSC) threshold criterion). For the analysis of the 2D maps of the D5 and D6 capsular particles, an ab initio model was generated without symmetry in cryoSPARC using a subset of particles. One cycle of 3D classification yielded two distinct species of capsular structures. Particles belonging to each species were separated into two stacks of 153,442 (D5) and 88,345 (D6) particles, respectively. Ab initio models were generated using cryoSPARC for each stack. The final models for 3D classification in RELION, were performed both without symmetry as well as imposing D5 symmetry for the 153,442-particle stack and D6 symmetry for the 88,345-particle stack. For the D5 particle, a final stack of 93,221 particles was first refined in cryoSPARC using a homogeneous refinement procedure and then manually refined in cisTEM to produce a final map at 3.7 Å resolution (0.143 FSC threshold criterion).

For the remaining 398 particles from the original 2D classification showing T = 3 icosahedral symmetry, ab initio model generation in cryoSPARC and homogeneous refinement yielded a map of 4.66 Å resolution (0.143 FSC threshold criterion) from 395 particles. Manual refinement of this structure and corresponding particles in cisTEM extracted into a 648 × 648 pixel box size. This produced a final map with resolution of 3.18 Å (0.143 FSC threshold criterion). For the D6 particle, a final stack of 16,723 particles was refined in an identical manner using cisTEM to produce a final map at 3.7 Å resolution (0.143 FSC threshold criterion).

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-019-13786-y | www.nature.com/naturecommunications

Model building. Maps from the T = 1 and D5 particles were segmented using the Seggar tool, implemented in UCSF Chimera46 to extract density for a pentamer (T = 1) and hexamer (D5), respectively. The extracted map segments were then used for model building. Initially, the CArec–CTD structures were docked into density using rigid body refinement in chimera. The fits were then further optimised using the ligandfit tools and missing sections of the model built into density in COOT47 with iterative rounds of real-space refinement. The monomer structure, extracted from an in-house X-ray source (Rigaku Micromax-007HF with a Raxis-IV detector), was then employed as a molecular replacement model in Phaser54 and located the two copies of HML2 CArec-NTD in the asymmetric unit (AU). The model was validated with the Protein Structure Validation Suite-PSVS62 and Procheck-NMR63. The Ramachandran plot for the family of structures showed 80% of residues are in the most favoured region, 11% are in the additionally allowed region and only 1% are outliers. Details of the structure determination are summarised in Supplementary Table 1.

Protein crystallisation and structure determination. HML2 CArec-NTD was crystallised using hanging drop vapour diffusion. Typically, a 40 mg mL−1 solution of HML2 CArec-NTD in 150 mM NaCl, 20 mM HEPES and 0.5 mM TCEP, pH 8.0, was mixed with an equal volume of crystallisation solution containing 0.2 M sodium acetate trihydrate, 0.1 M Tris hydrochloride and 15% (v/v) PEG 4000 and suspended over a reservoir of the crystallisation solution. Drops were incubated at 18°C. Crystals appeared within 2–3 days and were transferred into fresh crystallisation solution containing 20% (v/v) glycerol and flash frozen in liquid nitrogen. The crystals belonged to either the space group C2221 with three copies of HML2 CArec-NTD in the asymmetric unit (AU) or space group C2 with two copies of HML2 CArec-NTD in the AU. Seleno-methionine-derived protein was crystallised under the same conditions.

The structure of HML2 CArec–CTD was solved by a combination of single wavelength anomalous diffraction (SAD) and molecular replacement. Initially a 3.2 Å dataset was recorded on a C2 monoclinic crystal at 100 K using Cu Kα X-rays. The monomer from the C2221 crystal was then employed as a molecular replacement model in Phaser54 and located the two copies of HML2 CArec-NTD in the AU. The model was complete to 2.3 Å resolution, and has good geometric quality with a figure of merit of 0.67 and a map of sufficient quality to unambiguously build a near complete model of a single protomer using Arp/Warp53. A further high-resolution 1.8 Å dataset was collected on a C2221 crystal at 100 K using Cu Kα X-rays. The monomer from the C2221 crystal was then employed as a molecular replacement model in Phaser54 and located the two copies of HML2 CArec-CTD in the AU. The model was complete to 1.8 Å resolution, and has good geometric quality with a figure of merit of 0.67 and a map of sufficient quality to unambiguously build a near complete model of a single protomer using Arp/Warp53. This model was submitted to intra- and inter-molecular distance restraints defining the HML2 CArec–CTD homodimeric interface were derived from inter-proton NOE correlations observed in the filtered NOE experiments, while the corresponding NOEs were removed from the 3D 13C-NOSY-HSQC constraint list to avoid duplications. A set of 100 structures were calculated and the 20 lowest-energy structures of the set were taken to represent the converged ensemble and refined in an explicit water shell. The superposition of the 20 water-refined structures is shown in Supplementary Fig. 11b. The quality of the calculated structure ensemble was assessed and validated with the Protein Structure Validation Suite-PSVS52 and Procheck-NMR53. For the final 20 lowest-energy NMR structures, no distance or torsional restraints were imposed. The Ramachandran plot for the family of structures showed 88% of residues are in the most favoured region, 11% are in the additionally allowed region and only 1% are outliers. Details of the structure determination are summarised in Supplementary Table 3.

3D structural alignments and interface analysis. The European Bioinformatics Institute (EBI) protein structure comparison service (PDBeFold) (http://www.ebi.ac.uk/msd-srv/ssm/) was used to perform searches with HML2 CArec–NTD and HML2 CArec–CTD structures against the non-redundant protein data bank (PDB) (https://www.rcsb.org) and Centroid Chimera (https://www.c SHIELD, 2019) were calculated and the 20 lowest-energy structures of the set were taken to represent the converged ensemble and refined in an explicit water shell. The superposition of the 20 water-refined structures is shown in Supplementary Fig. 11b. The quality of the calculated structure ensemble was assessed and validated with the Protein Structure Validation Suite-PSVS52 and Procheck-NMR53. For the final 20 lowest-energy NMR structures, no distance or torsional restraints were imposed. The Ramachandran plot for the family of structures showed 88% of residues are in the most favoured region, 11% are in the additionally allowed region and only 1% are outliers. Details of the structure determination are summarised in Supplementary Table 3.

The final model comprises residues 1–6 and 23–151 (chain A) and residues 1–5 and 21–149 (chain B) and was refined to a Ramachandram plot of 15.7195%. The model quality was assessed using the program MolProbity50 and has good geometric quality with a figure of merit of 0.67 and a map of sufficient quality to unambiguously build a near complete model of a single protomer using Arp/Warp53. A further high-resolution 1.8 Å dataset was collected on a C2 monoclinic crystal at 100 K using Cu Kα X-rays. The monomer from the C2221 crystal was then employed as a molecular replacement model in Phaser54 and located the two copies of HML2 CArec-NTD in the AU. The model was complete to 2.3 Å resolution, and has good geometric quality with a figure of merit of 0.67 and a map of sufficient quality to unambiguously build a near complete model of a single protomer using Arp/Warp53. This model was submitted to intra- and inter-molecular distance restraints defining the HML2 CArec–CTD homodimeric interface were derived from inter-proton NOE correlations observed in the filtered NOE experiments, while the corresponding NOEs were removed from the 3D 13C-NOSY-HSQC constraint list to avoid duplications. A set of 100 structures were calculated and the 20 lowest-energy structures of the set were taken to represent the converged ensemble and refined in an explicit water shell. The superposition of the 20 water-refined structures is shown in Supplementary Fig. 11b. The quality of the calculated structure ensemble was assessed and validated with the Protein Structure Validation Suite-PSVS52 and Procheck-NMR53. For the final 20 lowest-energy NMR structures, no distance or torsional restraints were imposed. The Ramachandran plot for the family of structures showed 88% of residues are in the most favoured region, 11% are in the additionally allowed region and only 1% are outliers. Details of the structure determination are summarised in Supplementary Table 3.

NMR spectroscopy and structure determination. NMR experiments were recorded on 298 K on Bruker Avance 600-, 700-, 800- and 950-MHz spectrometers. 1H-15N, 13C-15N and 1H-13C-15N-labelled CArec–CTD samples were prepared in buffer containing 20 mM Tris-HCl, 50 mM NaCl and 0.5 mM TCEP, pH 7.0. For the D5 particle, a final stack of 93,221 particles was refined in an identical manner using cisTEM to produce a final map at 3.7 Å resolution (0.143 FSC threshold criterion).

For the remaining 398 particles from the original 2D classification showing T = 3 icosahedral symmetry, ab initio model generation in cryoSPARC and homogeneous refinement yielded a map of 4.66 Å resolution (0.143 FSC threshold criterion). For the D6 particle, a final stack of 16,723 particles was refined in an identical manner using cisTEM to produce a final map at 3.7 Å resolution (0.143 FSC threshold criterion).

The remaining 398 particles from the original 2D classification showing T = 3 icosahedral symmetry, ab initio model generation in cryoSPARC and homogeneous refinement yielded a map of 4.66 Å resolution (0.143 FSC threshold criterion). For the D6 particle, a final stack of 16,723 particles was refined in an identical manner using cisTEM to produce a final map at 3.7 Å resolution (0.143 FSC threshold criterion).

For the remaining 398 particles from the original 2D classification showing T = 3 icosahedral symmetry, ab initio model generation in cryoSPARC and homogeneous refinement yielded a map of 4.66 Å resolution (0.143 FSC threshold criterion). Re
the helix were calculated and used as the start and end points of the vector. Angles between the vectors were obtained by calculating the dot product of the two vectors and dividing by the product of the vector magnitudes.

**Structure-based sequence alignment.** Structure-based sequence alignment of alpha and betaretroviral CA-NTDs and CA-CTDs was performed using the PROMALS3D4 server. Where necessary, alignments were adjusted manually based on the position of secondary structures observed in seed structures used for the individual CA-NTD and CA-CTD alignments.

**Analytical ultracentrifugation.** Sedimentation velocity experiments were performed in a Beckman Optima XL-A analytical ultracentrifuge using conventional aluminium double sector centrepieces and sapphire windows. Solvent density and the protein partial specific volumes were determined from tabulated values65. Prior to centrifugation, samples were prepared by exhaustive dialysis against the buffer blank 20 mM Tris-HCl, pH 8, 150 mM NaCl and 0.5 mM TCEP. After centrifugation for 30 h at 22,000 r.p.m. (35,235 × g) and 293 K in an An-50 Ti rotor. Interference data were accessed at intervals of 180 s at a sample concentration of 1.5 mg mL$^{-1}$. Data recorded from moving boundaries were analysed using the size distribution function CS(λ) in the program SEDFIT64. Sedimentation equilibrium experiments were performed in a Beckman Optima XL-A analytical ultracentrifuge using aluminium double sector centrepieces in an An-50 Ti rotor. Prior to centrifugation, samples were dialysed exhaustively against the buffer blank 20 mM Tris-HCl, pH 8, 150 mM NaCl and 0.5 mM TCEP. Centrifugation for 30 h at 22,000 r.p.m. (35,235 × g) and the procedure repeated each time. Data were collected on samples of HML2 CArec$^{6}$-CTD and CArec$^{6}$-CTD (193A/1L96A) at varying protein concentration (50–200 µM) and at three speeds. The program SEDPHAT68,70 was used to determine weight-averaged molecular masses by nonlinear fitting to individual multi-speed equilibrium profiles to a single-species ideal solution model. Inspection of these data revealed that the HML2 CArec$^{6}$-CTD molecular mass showed significant concentration dependency and so global fitting to a monomer–dimer equilibrium model incorporating the data from the three speeds and three sample concentrations was applied to extract the dimerisation association constant ($K_d$). Details of protein hydrodynamic parameters and sedimentation equilibrium data are presented in Supplementary Table 4.

To compare the self-association properties of HML2 CArec$^{6}$-CTD with the dimer interface mutant HML2 CArec$^{6}$-CTD (193A/L196A), MSTAR analysis was performed on the equilibrium distributions recorded at 30,000 r.p.m. (65,520 × g) on 200 µM samples using the program SEDFIT-MSTAR71. The point average molecular weight ($M_w$) showed the expected increase with increasing radius for CArec$^{6}$-CTD and yielded a weight-averaged molecular mass at the cell bottom ($M_w$) intermediate between monomer and dimer. For CArec$^{6}$-CTD (193A/L196A), no radial dependency of $M_w$ was observed and the $M_w$ obtained equated to monomer molecular mass.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

The structural data that support the findings of this study have been deposited in the Protein Data Bank, BioMagResBank and EM Data Bank. The coordinates for HML2 CArec$^{6}$-NTD and HML2 CArec$^{6}$-CTD have the PDB accession numbers 6SA9 and 6SAI, respectively. Chemical shift assignments for HML2 CArec-CTD in the BioMagResBank have the accession number 34419. The coordinates for HML2 CArec and 12

1. Stoye, J. P. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat. Rev. Microbiol. 10, 395–406 (2012).
2. Johnson, W. E. Origins and evolutionary consequences of ancient endogenous retroviruses. Nat. Rev. Microbiol. 17, 355–370 (2019).
3. Beldshaw, R. et al. Long-term reinfection of the human genome by endogenous retroviruses. Proc. Natl Acad. Sci. USA 101, 4894–4899 (2004).
4. Barbulescu, M. et al. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr. Biol. 9, 861–868 (1999).
5. Turner, G. et al. Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr. Biol. 11, 1531–1535 (2001).
6. Hughes, J. F. & Coffin, J. M. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. Proc. Natl Acad. Sci. USA 101, 1688–1682 (2004).
7. Plater-Zyberk, E., Ruscetti, F. W. & Coffin, J. M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of endogenous retroviruses. Retrovirology 8, 9 (2011).
8. Grow, E. J. et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature 522, 221–225 (2015).
9. Fuchs, N. V. et al. Human endogenous retrovirus K (HERV-K) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. Retrovirology 10, 115 (2013).
10. Reynier, F. et al. Increase in human endogenous retrovirus HERV-K (HML-2) viral load in active rheumatoid arthritis. Scand. J. Immunol. 70, 295–299 (2009).
11. Stoye, J. P. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Proc. Natl Acad. Sci. USA 101, 4894–4899 (2004).
12. Barbulescu, M. et al. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr. Biol. 9, 861–868 (1999).
13. Turner, G. et al. Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr. Biol. 11, 1531–1535 (2001).
14. Hughes, J. F. & Coffin, J. M. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. Proc. Natl Acad. Sci. USA 101, 1688–1682 (2004).
15. Plater-Zyberk, E., Ruscetti, F. W. & Coffin, J. M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of endogenous retroviruses. Retrovirology 8, 9 (2011).
16. Grow, E. J. et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature 522, 221–225 (2015).
17. Fuchs, N. V. et al. Human endogenous retrovirus K (HERV-K) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. Retrovirology 10, 115 (2013).
18. Reynier, F. et al. Increase in human endogenous retrovirus HERV-K (HML-2) viral load in active rheumatoid arthritis. Scand. J. Immunol. 70, 295–299 (2009).
19. Plater-Zyberk, E., Ruscetti, F. W. & Coffin, J. M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of endogenous retroviruses. Retrovirology 8, 9 (2011).
20. Grow, E. J. et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature 522, 221–225 (2015).
21. Fuchs, N. V. et al. Human endogenous retrovirus K (HERV-K) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. Retrovirology 10, 115 (2013).
22. Reynier, F. et al. Increase in human endogenous retrovirus HERV-K (HML-2) viral load in active rheumatoid arthritis. Scand. J. Immunol. 70, 295–299 (2009).
33. Cardone, G., Purdy, J. G., Cheng, N., Craven, R. C. & Steven, A. C. Visualization of a missing link in retrovirus capsid assembly. *Nature* **457**, 6998–6999 (2009).

34. Zhao, G. et al. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **497**, 643–646 (2013).

35. Bharat, T. A. et al. Cryo-electron microscopy of tubular arrays of HIV-1 Gag resolves structures essential for immature virus assembly. *Proc. Natl Acad. Sci. USA* **111**, 8233–8238 (2014).

36. Gifford, R. J. et al. Nomenclature for endogenous retrovirus (ERV) loci. *Virology* **59**, 15 (2018).

37. Keller, P. W. et al. A two-pronged structural analysis of retroviral maturation indicates that core formation proceeds by a disassembly–reassemble pathway rather than a disparate transition. *J. Virol.* **87**, 13655–13664 (2013).

38. Doublie, S. Preparation of selenomethionyl proteins for phase determination. *Methods Enzymol.* **276**, 523–530 (1997).

39. de la Rosa-Trevín, J. M. et al. Scipion: a software framework toward integration, reproducibility and validation in 3D electron microscopy. *J. Struct. Biol.* **195**, 93–99 (2016).

40. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).

41. Rohou, A. & Grigorieff, N. CTFFind4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).

42. Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle cryomicroscopy. *Methods Enzymol.* **533**, 244 (2018).

43. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. CryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

44. Grant, T., Rohou, A., Grigorieff, N. cis-TEM, user-friendly software for single-particle image processing. *Elife* **7**, https://doi.org/10.7554/eLife.35383 (2018).

45. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).

46. Zweiter, N. et al. Structure solution and refinement of albumin by molecular replacement. *Acta Crystallogr. D* **66**, 1605–1612 (2010).

47. Emsley, P. & Lunt, A. Recent improvements in Coot. *Acta Crystallogr. D* **69**, 1441–1452 (2013).

48. Zwart, P. H. et al. Automated structure solution with the PHENIX suite. *Methods Enzymol.* **533**, 226 (2018).

49. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).

50. Lowe, C. S. et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* **50**, 437–450 (2003).

51. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).

52. Sheldrick, G. M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr. D* **66**, 479–483 (2010).

53. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

54. Pettersen, E. F. et al. UCSF chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **26**, 1605–1612 (2005).

55. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).

56. Hwang, T. L., van Zijl, P. C. & Mori, S. Accurate quantitation of water-amide exchange by T2 relaxation dispersion analysis. *J. Magn. Reson.* **160**, 160–167 (2003).

57. Painter, J. & Merritt, E. A. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr. D* **66**, 439–450 (2000).

58. Hwang, T. L., van Zijl, P. C. & Mori, S. Accurate quantitation of water-amide proton exchange rates using the phase-modulated CLEAN beam exchange (CLEANEX-PM) approach with a fast-HQHC (FHSQC) detection scheme. *J. Magn. Reson.* **111**, 221–226 (1998).

59. Angst, J. Protein structure prediction and analysis. *Methods Biochem. Anal.* **225**, 103–132 (1998).

60. Linge, J. P., Habeck, M., Rieping, W. & Wuthrich, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **6**, 1–10 (1995).

61. Linge, J. P., Habeck, M., Rieping, W. & Nilges, M. ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **19**, 315–316 (2003).

62. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **66**, 778–795 (2007).

63. Laskowski, R. A., Bullmann, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. QUANTA and PROCHECK-NMR: programs for checking the structure of macromolecular crystals. *Acta Crystallogr. D* **60**, 213–223 (2004).

64. Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).

65. Parse, B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

66. Schuck, P. On the analysis of protein self-association by sedimentation velocity analytical ultracentrifugation. *Anal. Biochem.* **320**, 104–124 (2003).

67. Vistica, J. et al. Sedimentation equilibrium analysis of protein interactions with global implicit mass conservation constraints and systematic noise decomposition. *Anal. Chem.* **86**, 234–256 (2004).

68. Schuck, P. On the analysis of protein self-association by sedimentation velocity analytical ultracentrifugation. *Anal. Biochem.* **320**, 104–124 (2003).

69. Vistica, J. et al. Sedimentation equilibrium analysis of protein interactions with global implicit mass conservation constraints and systematic noise decomposition. *Anal. Chem.* **86**, 234–256 (2004).

70. Schuck, P. et al. SEDFIT-MSTAR: molecular weight and molecular weight distribution analysis of polymers by sedimentation equilibrium in the ultracentrifuge. *Analyst* **139**, 79–92 (2014)

Acknowledgements

We gratefully acknowledge the Francis Crick Structural Biology Science Technology platform, Diamond Light Source (Grant No. MX13775) and the MRC UK biomedical NMR facility for access. We thank members of the Rosenthal and Taylor Laboratory for advice and support. This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001143, FC001178), the UK Medical Research Council (FC001143, FC001178) and the Wellcome Trust (FC001143, FC001178), and by grants from the Wellcome Trust (108014/Z/15/Z) (IAT) and the UK Medical Research Council (U117574558, MC_PC_13051) (A.R.).

Author contributions

O.A., T.G., N.J.B., L.R., D.C.G., L.R., A.N. and I.A.T. performed experiments. O.A., T.G., N.J.B., L.R., D.C.G., K.S., A.N. and I.A.T. contributed to data analysis and manuscript writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-13786-y.

Correspondence and requests for materials should be addressed to I.A.T. or P.B.R.

Peer review information Nature Communications thanks Sébastien Alphose and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2019