Published in final edited form as:
Nat Neurosci. 2020 February ; 23(2): 172–175. doi:10.1038/s41593-019-0569-y.

Structures of virus-like capsids formed by the *Drosophila* neuronal Arc proteins

Simon Erlendsson[1], **Dustin R. Morado**[1], **Harrison B. Cullen**[2], **Cedric Feschotte**[2], **Jason D. Shepherd**[3], **John A. G. Briggs**[1,*]

[1]Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge, UK

[2]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA

[3]Department of Neurobiology and Anatomy, The University of Utah School of Medicine, Salt Lake City, Utah, USA

Abstract

Arc, a neuronal gene critical for synaptic plasticity, originated through domestication of retrotransposon *Gag* genes and mediates intercellular mRNA transfer. We report high-resolution structures of retrovirus-like capsids formed by *Drosophila* dArc1 and dArc2 that have surface spikes and putative internal RNA-binding domains. These data demonstrate that virus-like capsid-forming properties of Arc are evolutionarily conserved and provide a structural basis for understanding their function in intercellular communication.

The immediate early gene *Arc* is a master regulator of synaptic plasticity[1] essential for consolidation of memory[2] and experience-dependent long-lasting changes in the mammalian brain[3]. The *Arc* gene is conserved throughout tetrapods, but absent in fish and basal chordates[4]. At least two Arc homologues (*dArc1* and *dArc2*) are found in the brachyceran fly lineage, which arose by genomic duplication of an ancestral *dArc* gene[4,5]. dArc1 is implicated in modulating synaptic plasticity at the neuro-muscular junction[5] and in controlling fat metabolism[6,7]. dArc2 has no assigned function. Tetrapod *Arc* and fly *dArc*

*Correspondence to: jbriggs@mrc-lmb.cam.ac.uk.

Data availability

Atomic coordinates have been deposited in PDB under accession numbers: 6TAP, 6TAQ, 6TAR, 6TAS, 6TAT, 6TAU. Cryo-EM density maps have been deposited in the Electron Microscopy Data Bank under accession numbers: EMD-10423, EMD-10424, EMD-10425, EMD-10426, EMD-10427, EMD-10428. Requests for raw data or materials should be addressed to jbriggs@mrc-lmb.cam.ac.uk.

Accession codes

Atomic coordinates have been deposited in PDB under accession numbers: 6TAP, 6TAQ, 6TAR, 6TAS, 6TAT, 6TAU. Cryo-EM density maps have been deposited in the Electron Microscopy Data Bank under accession numbers: EMD-10423, EMD-10424, EMD-10425, EMD-10426, EMD-10427, EMD-10428.

Author contributions:

S.E., J.D.S., and J.A.G.B. designed the project. S.E. prepared all samples and performed cryo-EM sample preparation/screening, reconstruction and model building with assistance from D.R.M. and J.A.G.B.. H.B.C. and C.F. designed and performed evolutionary sequence analyses. S.E., J.D.S., and J.A.G.B. prepared the manuscript. All authors commented on the final manuscript.

Competing interests:

J.D.S. is a paid consultant for a company exploring biotechnological applications of domesticated retrotransposons.

Europe PMC Funders Author Manuscripts

genes originated from independent domestication events from distinct members of the Ty3/*gypsy* superfamily of Long Terminal Repeat (LTR) retrotransposons[4,8,9]. Active and inactive retrotransposons make up large fractions of the genomes of multi-cellular organisms[10,11]. The Gag domain in Ty3/*gypsy* retrotransposons encodes the CA protein that forms a virus-like protein capsid with extensive structural homology to that of retroviruses such as Human Immunodeficiency Virus (HIV)[12–14]. Tetrapod Arc and dArc1 contain a CA fold[13,14] and form capsid-like structures, which are secreted in extracellular vesicles (EVs) that mediate transfer of Arc protein and mRNA between cells[4,5]. These observations suggest that Arc capsids support a new mechanism for intercellular communication.

We purified recombinant dArc1 and dArc2. Both proteins assembled homogenous, ~37 nm diameter spherical particles with a dense ~3 nm thick capsid layer (Fig. 1a,d). Using single particle cryo-EM we determined the structures of dArc1 and dArc2 capsids at 2.8 Å and 3.7 Å resolution, respectively (Fig. 1, Supplementary Fig. 1-3, Supplementary Table 1, Supplementary Movie 1 and 2). dArc1 and dArc2 assemble icosahedral capsids with triangulation number T=4, composed of 12 five-fold symmetric pentameric capsomeres and 30 two-fold symmetric hexameric capsomeres (Fig. 1). Short spikes (5-8 nm long) protrude from the centre of each capsomere.

We built atomic models for CA (dArc1 residues 42-205, dArc2 residues 29-192) into the structures (Fig. 2a-b, Extended Data Fig. 1a-b, Supplementary Table 1, Supplementary Movie 3). The dArc1 and dArc2 capsomeres are highly similar, with RMSDs <1 Å, reflecting similar sequences (~53% identity, ~73% similarity) (Extended Data Fig. 2-3). CA consists of two globular domains, CA_{NTD} and CA_{CTD}, connected by a flexible linker (Fig. 2b). The CA_{NTD} (dArc1 residues K42-S120) constitutes the central part of each capsomere (Fig. 2c-d) and consists of four α-helices and an extended chain. The CA_{CTD} (dArc1 residues A125-H203) forms the “valleys” between protruding capsomeres and consists of five α-helices (Fig. 2b-d).

The dArc1 CA_{NTD} and CA_{CTD} structures are similar to those of monomeric rat Arc (rArc) in solution[13] (RMSDs of 1.3 and 1.2 Å, respectively; Extended Data Fig. 4a-b). However, protein-protein interfaces in the dArc capsid differ from those predicted based on rArc crystal contacts[15]. The extended chain in dArc CA_{NTD} (dArc1 residues 42-55, dArc2 residues 29-32) binds and occludes a hydrophobic groove between α1 and the linker between α2 and α3 (Fig. 2c), which is the binding site for TARPγ2, CaMKII and NMDA receptor peptides in rArc[13,14] (Extended Data Fig. 4c). These observations are consistent with a requirement for domain reorganization or uncovering of interaction surfaces to initiate rArc CA oligomerization[13].

CA molecules are held together by two intracapsomer interfaces: (i) a CA_{NTD}:CA_{NTD} interface formed by acidic residues from α1 and basic residues from the neighboring α2 and α3 (Fig. 2d); (ii) a CA_{NTD}:CA_{CTD} interface involving I148 from CA_{CTD} α6 docking into a hydrophobic pocket between α3 and α4 from the neighboring CA_{NTD}, and a potential salt bridge between D144 in α6 and R56 in α1 (Fig. 2d). The dimeric CA_{CTD}:CA_{CTD} intercapsomere interface is dominated by hydrophobic and π-stacking interactions between adjacent α5 and α7 (Fig. 2e.i). The three-fold and pseudo-three-fold CA_{CTD} axes provide

the biggest accessible openings in the capsid surface. At these positions in dArc1, the sidechains of K181 and H182 extend towards the center of the opening and are positioned 6-10 Å apart (Fig. 2e.ii and Extended Data Fig. 5h). In dArc2, E168 and S169 extend towards the centre of the opening (Extended Data Fig. 1e.ii). We speculate that changes in protonation state of H182 in dArc1 could modulate capsid stability or transport into or out of the capsid.

The four independent copies of CA in the asymmetric unit differ only in the relative orientation of CA_{NTD} and CA_{CTD} (Extended Data Fig. 5d). Independent copies of the CA_{CTD} all form interfaces similar to those in the Ty3 and mature HIV-1 capsids (Extended Data Fig. 5g, 6-7). In contrast, the CA_{NTD}-CA_{NTD} interface differs among the independent copies of CA (Extended Data Fig. 5e), and between dArc and Ty3 (Extended Data Fig. 7f-g). This reflects divergence of CA_{NTD} structure and arrangement among retrotransposons and retroviruses, perhaps due to the need for CA_{NTD} to interact with divergent host factors[14].

dArc1 and dArc2 contain 41 and 28 amino acids upstream of CA_{NTD}, respectively. These sequences are similar except for a poly-alanine stretch present in dArc1 (Extended Data Fig. 2) and form the exposed spikes. The spikes are flexible and are not resolved (Fig. 1 and Supplementary Fig. 3). The N termini have predicted propensity to form amphipathic α-helices (Supplementary Fig. 4), with the hydrophobic face likely to mediate oligomerization, bind cellular proteins, or bind membranes. In dArc, the N termini occlude openings in the capsid at the five-fold and two-fold axes and could regulate access to the capsid interior.

We observe a disordered/diffuse density beneath the five-fold axis inside the dArc1 and dArc2 capsids (Fig. 1), although the dArc2 protein terminates at the end of α9 of CA_{CTD} (Extended Data Fig. 2). This suggests that some copies of the N-terminal region constitute the spike, while others protrude inwards. Tetrapod Arc has an extended ~200 amino acid N-terminal domain that may bind RNA and/or negatively charged membranes[16]. We speculate that this domain is similarly located both outside and inside the capsid, allowing membrane binding and RNA packaging to be facilitated by the same domain.

dArc1 has 48 residues downstream of CA_{CTD} (dArc1 NC) that extend into the capsid interior and are homologous to the nucleocapsid (NC) domains of retroviruses/retrotransposons (Fig. 3a), which mediate specific nucleic acid interactions. Below the two-fold axis, two copies of residues 224-252 in dArc1 NC are well resolved, forming anti-parallel single-knuckle zinc fingers (Fig. 3b-d). Residues 206-223, as well as the other copies of NC in the asymmetric unit, are not resolved and are presumably poorly ordered. The zinc fingers are connected to one another by salt bridges and the majority of the basic residues face electronegative patches in the interface between α1 and the adjacent α3 in CA_{NTD} (Supplementary Movie 4). The spatial separation of the two zinc fingers from adjacent copies of dArc1 NC is strikingly similar to the separation of the two zinc fingers within a single copy of HIV-1 NC when bound to SL2 or SL3 in the HIV-1 genomic packaging signal[17,18] (Fig. 3e). This arrangement facilitates binding of HIV-1 NC to exposed bases in the stem loop tips. The absence of zinc fingers in dArc2 suggests putative differences in RNA-binding specificity. We posit that dArc1 NC may facilitate specific

recognition of the 3' UTR of dArc1 mRNA, which is absent in dArc2. Consistent with this observation, dArc1 mRNA is enriched and more abundant in EVs than dArc2 mRNA, which lacks a long 3'UTR[5].

dArc1 and dArc2 CA have conserved sequence and structure (Extended Data Fig. 2-3). Codon selection analyses reveal that each gene has evolved under strong purifying selection since their emergence in a bracyceran fly ancestor ~100 million years ago (Extended Data Fig. 8 and Supplementary Data Files 1-4). Each individual structural domain (spike, CA_{NTD}, CA_{CTD}, NC) has experienced a comparable level of functional constraint during evolution (Extended Data Fig. 8). The hydrophobic cores and protein-protein interfaces of CA are particularly conserved at the amino acid level: of 36 more disruptive amino acid substitutions, 31 are exposed, and four are at CA_{NTD}:CA_{NTD} interfaces. Ten of the amino acid differences are additional basic residues in dArc2 dispersed in linear sequence but located on the surface of the capsid. Consequently, the interior surface of the dArc2 capsid is more electropositive than dArc1 (Fig. 3f,g, Extended Data Fig. 9), while dArc1 has additional basic residues in the disordered NC regions. The basic capsid interiors of both dArc1 and dArc2 may facilitate non-specific electrostatic RNA packaging.

Ty3 packages two copies of its 5.2 kb genome into a T=9 capsid with an internal volume of ~5 x 10^4 nm^3 [12]. The dArc1 capsid is smaller (volume ~1.7 x 10^4 nm^3), but would enclose RNA at a similar density to Ty3 if it packaged two copies of the ~2.3 kb dArc1 mRNA (including the 3' UTR). We suggest that shrinkage of the dArc capsid from T=9 to T=4 occurred post-domestication, reflecting shortening of the packaged RNA from a full-length LTR retrotransposon mRNA to the shorter dArc1 mRNA.

The structures of dArc1 and dArc2 presented here provide a foundation to interpret mutations, post-translational modifications or binding sites, and facilitate the design of experiments to specifically disrupt capsid assembly and differentiate cell-autonomous and non-autonomous functions. Do the N-terminal spikes mediate membrane and/or RNA binding? Do the Zn fingers and basic capsid interior determine specificity of RNA packaging, and would modulating RNA specificity alter dArc function? Our observations demonstrate that dArc1 and dArc2, despite their ancient origin and sequence divergence from each other and from tetrapod Arc, have preserved capsid-forming and RNA-packaging properties strikingly reminiscent of retrotransposon and retroviral Gag proteins. This supports the model that *Arc* genes have been repurposed from retrotransposons to package and transfer RNA between cells within virus-like capsids[4,5].

Methods

dArc1 and dArc2 expression and purification

DNA sequences corresponding to dArc1 residues 2-254 and dArc2 residues 2-193 were expressed as Glutathione-s-Transferase (GST) fusion constructs by subcloning into the pGEX 4T1 vector (GE28-9545-49). The plasmids were transform into E. coli BL21(DE3)/pLysS and protein was expressed in autoinduction medium (10 g tryptone, 5 g yeast extract, 5 g glycerol, 0.5 g glucose, 2 g α-lactose, 3.3 g $(NH_4)_2SO_4$, 6.8 g KH_2PO_4, 7.1 g Na_2HPO_4, 1 mM $MgSO_4$, 50 μM $ZnCl_2$ per liter - ampicillin and chloramphenicol for selection).

Cultures are grown to an OD_{600} of 0.6-0.8 at 37 °C and then shifted to 19 °C for autoinduction overnight (12-16 hrs.). After autoinduction cells were pelleted at 6000 rpm (Rotor JLA 8.100), for 15 min. at 4 °C and resuspended in Lysis buffer (50 mM Tris, 400 mM NaCl, 5 % Glycerol, 2 mM DTT, 50 μM $ZnCl_2$, 0.2 mM PMSF (Phenylmethanesulfonyl fluoride), pierce protease inhibitor cocktail (Thermo Scientific A32963), Nuclease Mix (GE80-6501-42) pH 8) and snap frozen in liquid nitrogen. The thawed lysates were sonicated and centrifuged at 16000 rpm (Rotor JA 25.50) for 30 min. The resulting supernatants were cleared by filtration (0.22 μM), and the conductivity of the solution was adjusted to 20 mS/cm before applying it to a Heparin Sepharose column (GE17-0407-01), and eluting using a NaCl gradient. The sample was immediately applied to a GSTrap column (GE17-5282-01) and eluted in TBS (50 mM Tris, 150 mM NaCl, 2 mM DTT, 50 μM $ZnCl_2$, pH 8) containing 10 mM Reduced L-Gluthatione (Sigma-Aldrich, G6529). The sample was dialysed into TBS overnight and the GST-fusion construct was cleaved using Thrombin (Novagen, Merch 69671). dArc capsids formed spontaneously, incorporating almost all cleaved Arc, and the cleaved GST tag was removed by washing the sample over a 100 kDa MWCO spin filtration column (GE vivaspin).

Cryo-EM

5 μl of dArc1 or dArc2 capsids at a concentration of ~1 mg/ml were applied to glow-discharged continuous carbon lacey grids (Lacey Carbon Films on 300 Mesh Copper Grids, Agar scientific), blotted and plunge-frozen in liquid ethane using a FEI Vitrobot Mark IV. Cryo-EM images were acquired on a 300 keV FEI Titan Krios microscope equipped with a Gatan K2-Summit 4Kx4K detector operated in counting mode. A GIF-quantum energy filter (Gatan) was used with a slit width of 20 eV to remove inelastically scattered electrons. Both the dArc1 and dArc2 datasets were collected using SerialEM[19] at a nominal magnification of 105,000 with calibrated pixel sizes of 1.211 Å for dArc1 and 1.388 Å for dArc2. Micrographs were recorded as movies divided into 75 frames. For dArc1 we used a total exposure time of 6.6 s. and an accumulated dose of 35.4 electrons per $Å^2$. For dArc2, the total exposure time used was 16.8 s and the accumulated dose 34.9 electrons per $Å^2$. Defocus values ranged from −1.0 to −4.0 μm. Data collection parameters are summarized in Supplementary Table 1.

Image processing

dArc1 and dArc2 image processing and reconstructions are summarized in Supplementary Fig. 1-2. Dataset parameters are presented in Supplementary Table 1. Acquired movies were aligned using MotionCor2[20] with 5 × 5 patches and applying dose-weighting to individual frames. Contrast transfer function (CTF) parameters were estimated using Ctffind4[21]. Particles were automatically picked using RELION 3[22,23] and extracted into 512×512 pixel boxes. Extracted particles were subjected to several rounds of 2D classification to remove false picks. We did not observe classes that might represent alternative immature and mature populations of capsids. 2D classes showed views characteristic for icosahedral symmetry, and icosahedral symmetry (I4) was applied for initial model generation and throughout subsequent 3D classification and refinement. A small number of further particles were discarded based on 3D classification. The 3D classification for dArc2 considered only at the CA capsid layer. The dArc2 capsids are less stable at high concentration, and this approach

retained more data and produced a higher resolution density after refinement. After refinement we performed per-particle CTF estimation, Bayesian polishing and Ewald sphere correction[24]. The effective resolutions of the cryo-EM density maps were estimated by Fourier shell correlation (FSC = 0.143) (Supplementary Fig. 1) according to the definition of Rosenthal and Henderson[25].

To further improve the resolution, we performed symmetry expansion as implemented in RELION, to calculate the positions and orientations for each of the 2,260,000 asymmetric units for dArc1 and 106,740 for dArc2, centered either at the five-fold or at the two-fold capsomeres. We extracted individual capsomeres using a box sizes of 148 pixels in all cases. For the five-fold capsomeres we removed the redundant five-fold symmetrized capsomeres leaving 453,276 particles for dArc1 and 21,348 particles for dArc2 which were further refined using C5 symmetry. For the dArc1 two-fold capsomeres we performed 3D classification without any applied symmetry and selected only classes displaying clear two-fold symmetric inner density for further refinements (1,685,349 particles). For dArc2 we did not observe any notable differences in classes from the 3D classification, and all particles were used for further refinement. This local refinement approach yielded structures with resolutions of 2.8 Å for the dArc1 capsomeres and 3.7 Å for the dArc2 capsomeres (Supplementary Fig. 2). For all maps, local resolutions were calculated using ResMap[26] (Supplementary Fig. 3) and the maps were locally sharpened using LocalDeblur[27].

Modelling

Model building was performed into the final locally-refined maps. The dArc CA domains were initially modelled based on the NMR structure of the rat Arc CA domain (PDB: 6GSE)[13] using MODELLER[28] and the flexible linker between the NTD and CTD as well as the N-terminal tails were *de novo* built using Coot[29]. The Coot models were then refined into the EM density using Phenix real-space refinement (program phenix.real_space_refine)[30,31]. The dArc1 zinc finger was initially modelled based on the HIV-1 zinc finger (PDB:1A1T)[17] using MODELLER, rigid-body fitted into the EM density in Chimera [32] and used as starting point for building additional amino acids in each direction using Coot. We refined the zinc fingers into the EM density using Zen implemented within PDB_REDO[33] for ideal tetrahedral coordination of the zinc molecules. Side-chain rotamers and potential clashes at points of contact between the zinc fingers and CA were manually fixed in Coot. Models built into the locally-refined maps were then fitted as rigid bodies into the lower-resolution full capsid maps. Residues making contacts between capsomeres were refined manually. The quality of all models were evaluated using MolProbity, Ramachandran statistics and Clashscore[34], both before and after fitting the into the lower resolution full capsid maps (Supplementary Table 1). See Supplementary Fig. 5 for per residue model–map correlation coefficients.

Alignments

All amino acid sequence alignments were performed using T-COFFEE[35] and figures prepared using ESPript 3.0[36]. Input sequences and PDB files are provided in Supplementary Table 2.

Codon Selection Analyses

dArc1 or dArc2 homologous sequences were mined with NCBI megaBLAST [https://blast.ncbi.nlm.nih.gov/Blast.cgi] using the *Drosophila melanogaster* dArc1 [NM_137111.3] and dArc2 [NM_137112.2] nucleotide sequence as queries. The 'refseq RNA' and 'WGS' (Whole Genome Shotgun) sequence databases, taxonomically limited to species belonging to order Diptera, were searched using default parameters. Each full mRNA sequence with an e-value less than 0.01 was extracted and their open reading frame (ORF) further extracted using the NCBI ORF Finder program [https://www.ncbi.nlm.nih.gov/orffinder/]. The nucleotide sequence for each species' ORF was aligned using the MUSCLE program[37]. A separate multiple sequence alignment was made for dArc1 and dArc2, and multiple identical or nearly identical sequences from a single species (likely database or biological duplicates) were filtered out of the alignment, so each species was represented only once for each gene alignment. The sequence with most similarity to the *Drosophila melanogaster* sequence was kept in the alignment. Sub-alignments corresponding to each of the domains of dArc1 and dArc2, as determined from the structures, were created to calculate domain-specific dN/dS values (see below).

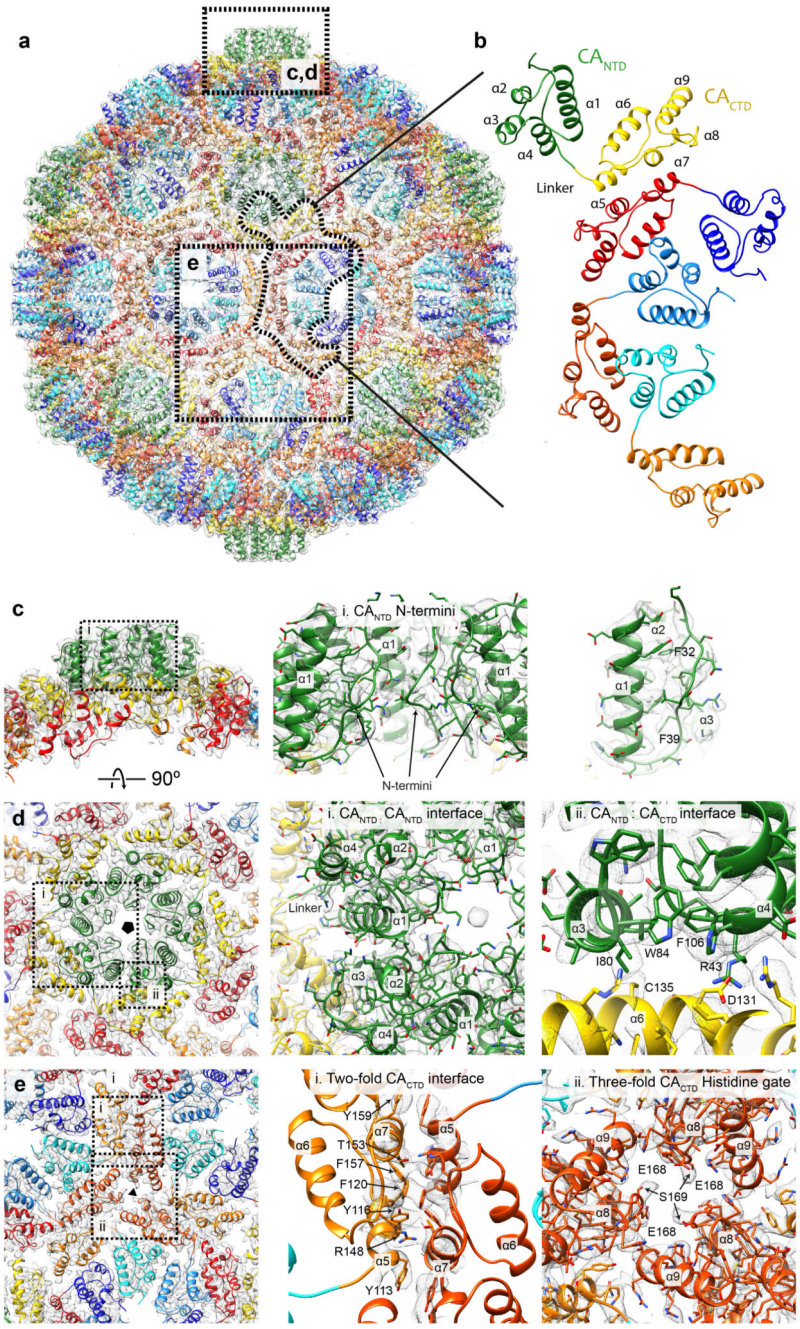
The ratio of the rate of nonsynonymous substitutions to the rate of synonymous substitutions (dN/dS) for each gene alignment was computed using the *codeml* function in the PAML program[38]. All partial codons, stop codons, or those with ambiguities were replaced with a gap in the alignment. The phylogenetic tree used by *codeml* to calculate dN/dS values along branches of the tree was calculated using MEGA7[39] with the maximum likelihood method based on the multiple nucleotide alignments described above.

codeml outputs the log-likelihood values of the different hypotheses tested by the analysis. First, the likelihood value is calculated under the under null model. Under the null hypothesis, it is assumed that the ORF is not under any functional constraint, i.e. the coding sequence is evolving neutrally. Then, the *codeml* process was repeated, under an alternate model with a variable dN/dS value calculated using Maximum Likelihood. The likelihood values of the dN/dS under each model are compared by performing a chi-square test to determine the significance of the dN/dS calculation under the alternate hypothesis. A p-value less than 0.05 was used to reject neutrality and support evidence of purifying selection having acted on the coding sequence. All P-values are < 0.0001.

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Extended Data

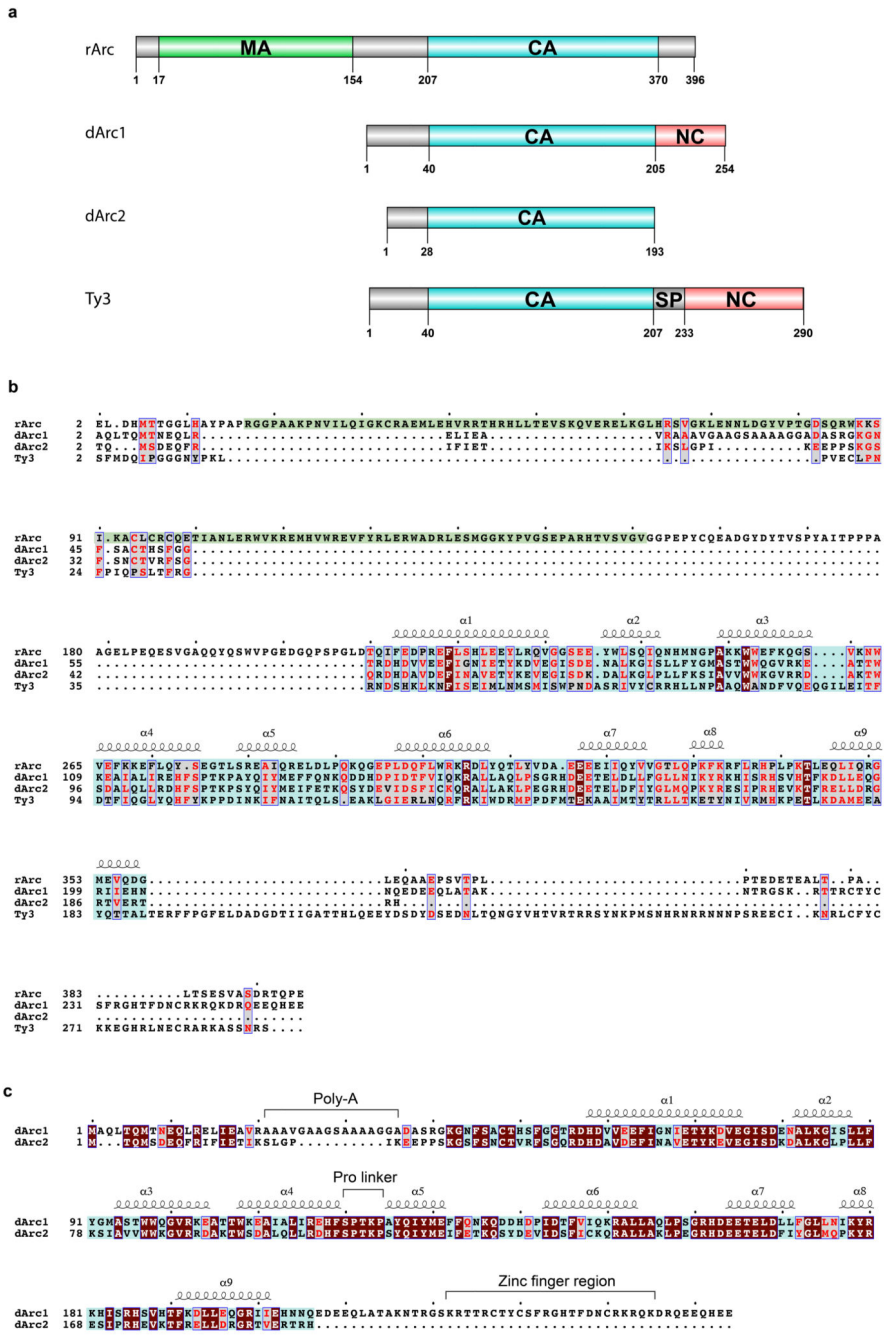


Extended Data Fig. 1. Full capsid atomic model of the dArc2.

The views and color scheme are similar to those shown in figure 2. **a)** The 12 five-fold capsomeres are coloured in green (CA_{NTD}), and yellow (CA_{CTD}). The 30 two-fold capsomeres are coloured in cyan to blue (CA_{NTD}) and orange to red (CA_{CTD}). **b)** The asymmetric unit containing four CA molecules. 60 asymmetric units including 240 individual CA molecules make up the T=4 capsid. **c)** Close-up of the five-fold capsomere (outlined in **a**) i) Cut-away showing three out of five N-termini in the centre of the capsomeres. The N-termini extend into and form the capsid spikes. The N-termini are

Europe PMC Funders Author Manuscripts

stabilized by docking into an extended hydrophobic groove adjacent to α1. **d)** External view of the five-fold capsomere i) The CA_{NTD}:CA_{NTD} interaction between α1, and α2 and α3 of the neighbouring CA molecule in the capsomere. Electronegative charged residues on the outside of α1 interact with electropositive charges in α2 and α3. ii) The CA_{NTD}:CA_{CTD} interface which involves α6 in the CA_{CTD} and α3 and α4 in the neighbouring CA_{NTD}. This interface relies on both hydrophobic and electrostatic interactions. Residues R43, I80, W84, F106, C135 and D131 are depicted in the figure. **e)** External view of the two and three-fold CA_{CTD} interfaces. i) The two-fold CA_{CTD} interface connects two adjacent capsomeres and is dominated by hydrophobic π-stacking interactions. The interface involves residues from α5 and α. Residues Y113, Y116, F120, R148, T153, F157 and Y159 are depicted in figure. EM density is shown only for the contact site. All CA_{CTD} interfaces are highly similar (Supplementary Fig. 7). ii) The three-fold CA_{CTD} axis. At this position, instead of the histidine present in dArc1, E168 and S169 from α8 surround the largest gap in the capsid. The corresponding views for dArc1 are shown in Fig. 2.



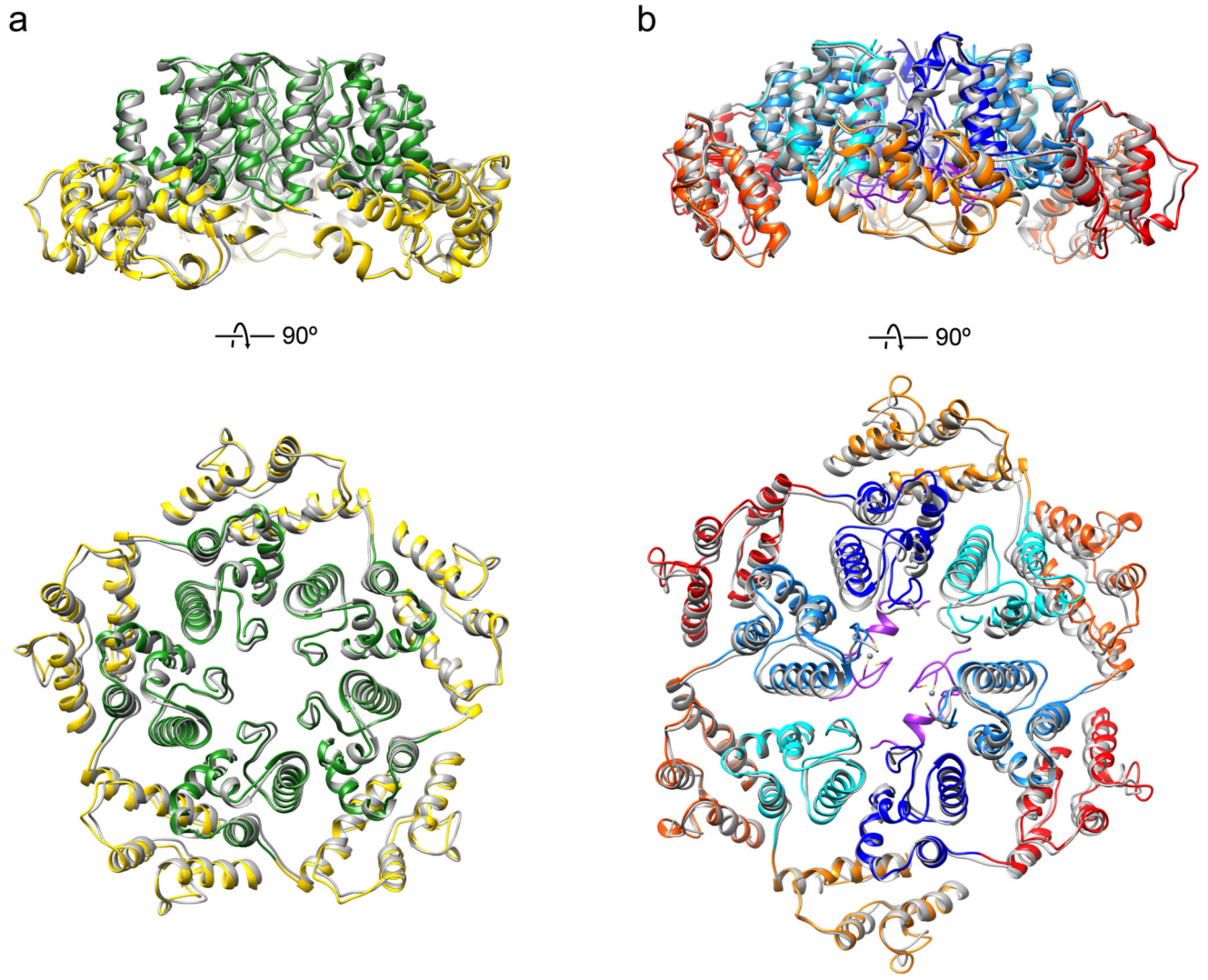
Extended Data Fig. 2. Sequence alignment of dArc1 and dArc2, rArc and Ty3.

a) Domain overview. Matrix-like domain (MA; Green), capsid domain (CA; blue), nucleocapsid domain (NC; red). Only rArc contains a putative MA domain but lacks the NC domain. dArc2 is shortest of the four and only codes for the CA domain. **b)** The amino acid sequence alignment shows good overall alignment of the CA coding region for all proteins. Conserved residues are Brown, Equivalent residues (T-Coffee equivalence score >0.7) are grey. **c)** The amino acid sequence alignment between dArc1 and dArc2. Except for the Poly-

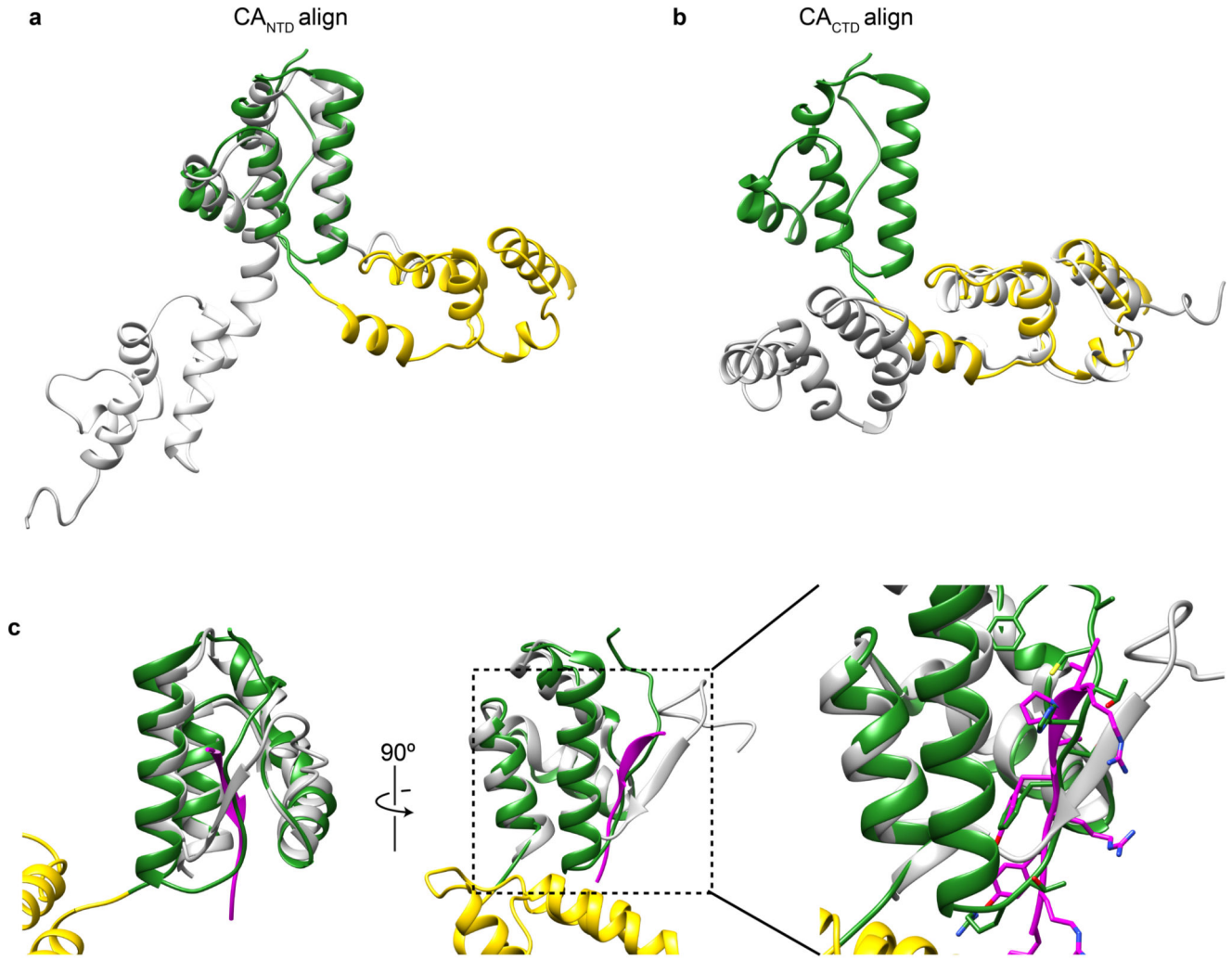
Europe PMC Funders Author Manuscripts

alanine stretch and the NC domain, the sequences are highly conserved. The zinc finger region is marked.

Europe PMC Funders Author Manuscripts



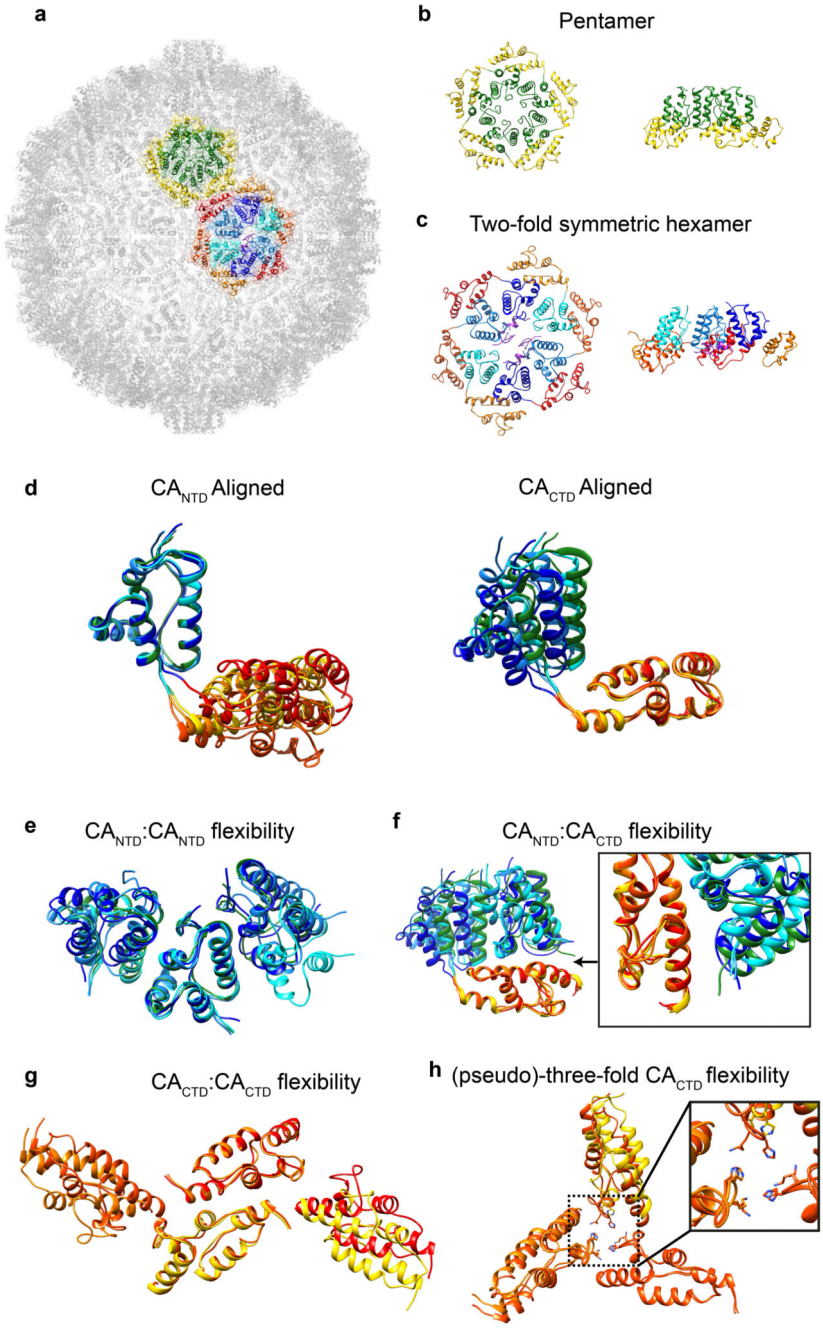
Extended Data Fig. 3. Comparison of dArc1 and dArc2 capsomer structures.
a) Comparison of five-fold symmetric capsomeres of dArc1 (colored) and dArc2 (grey). Full model C^{α}-RMSD: 0.7 Å. **b)** Comparison of two-fold symmetric capsomeres of dArc1 (colored) and dArc2 (grey). Full model C^{α}-RMSD: 0.9 Å.



Extended Data Fig. 4. Alignment of the dArc1 and rat Arc structures.

a-b) The dArc1 CA structure from the five-fold capsomere, compared to the full-length rat Arc (rArc) CA structure (obtained by Nuclear Magnetic Resonance, PDB ID: 6GSE)[13]. The rArc CA_{NTD} and CA_{CTD} are depicted in grey and light grey, respectively. The individual CA_{NTD} and CA_{CTD} folds are completely conserved. The flexible linker connecting CA_{NTD} and CA_{CTD} in the dArc CA capsid structure is more rigid in the monomeric rArc and the interdomain orientation is different. **c)** Alignment between dArc1 CA_{NTD} (green) and the rArc CA_{NTD} (grey) crystal structure (PDB ID: 4X3H)[14] bound to the transmembrane AMPAR regulatory protein γ2 (TARPγ2) (pink). The binding site for TARPγ2, CaMKII and NMDA peptides in rArc, is occupied by the N-terminus in dArc1 and dArc2.

Europe PMC Funders Author Manuscripts

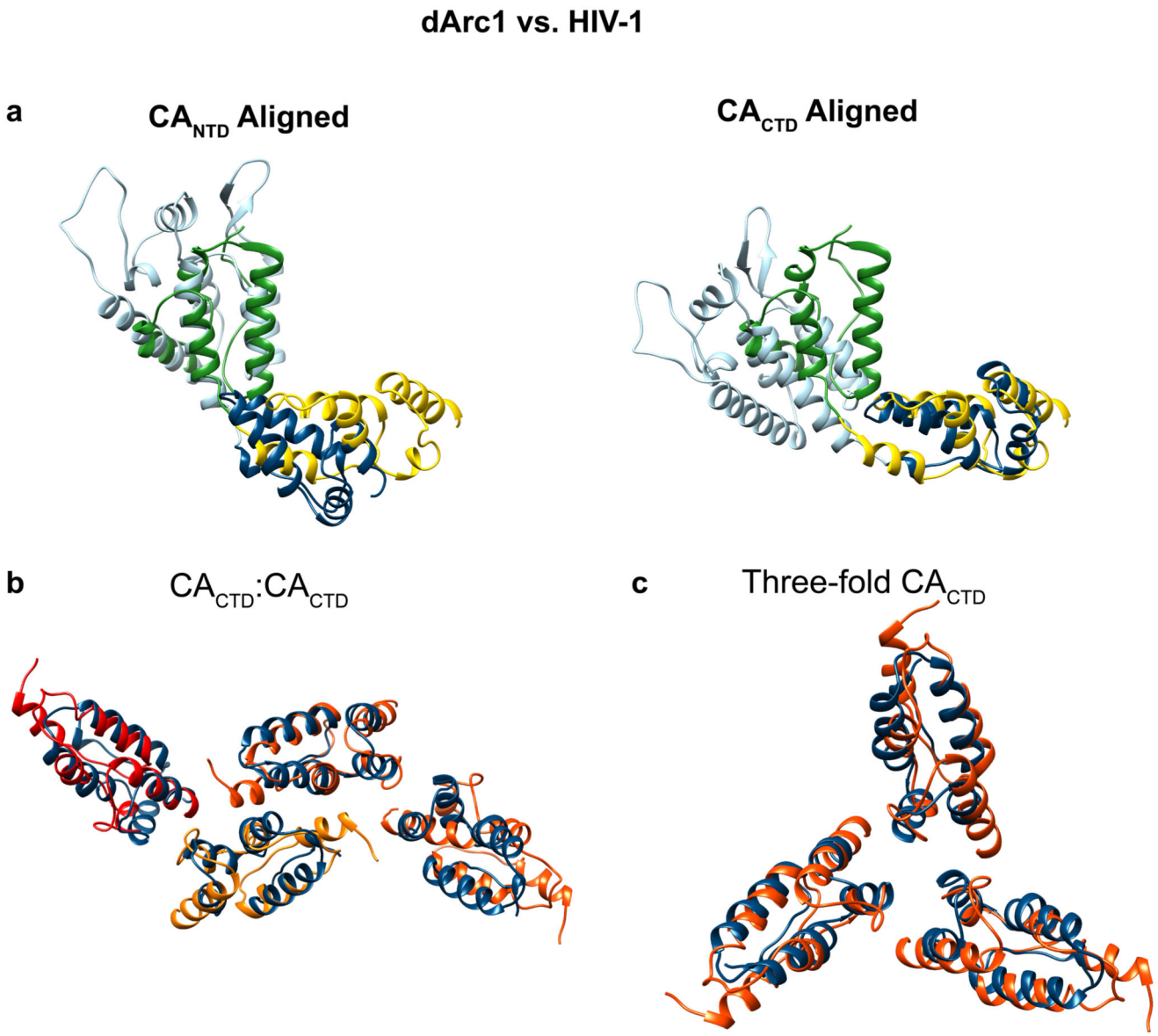


Extended Data Fig. 5. Flexibility of the CA domain within the dArc1 capsid.

a) Overview of the atomic structure of the dArc1 capsid with one five-fold and one two-fold capsomere highlighted. **b-c)** Top and side views of the five-fold and two-fold capsomeres to indicate the color coding in the following panels. **d)** The four different conformations of CA in the asymmetric unit aligned by either the CA_{NTD} or the CA_{CTD} domain. **e)** Flexibility of the four different CA_{NTD}:CA_{NTD} interfaces in the capsomeres, with the central CA_{NTD} aligned. For the adjacent CA_{NTD} C^{α}-RMSD: 8.6 Å. The interface accommodates relative displacements up to 55° between α1 and α2-3 in the adjacent CA_{NTD}. **f)** Flexibility of the

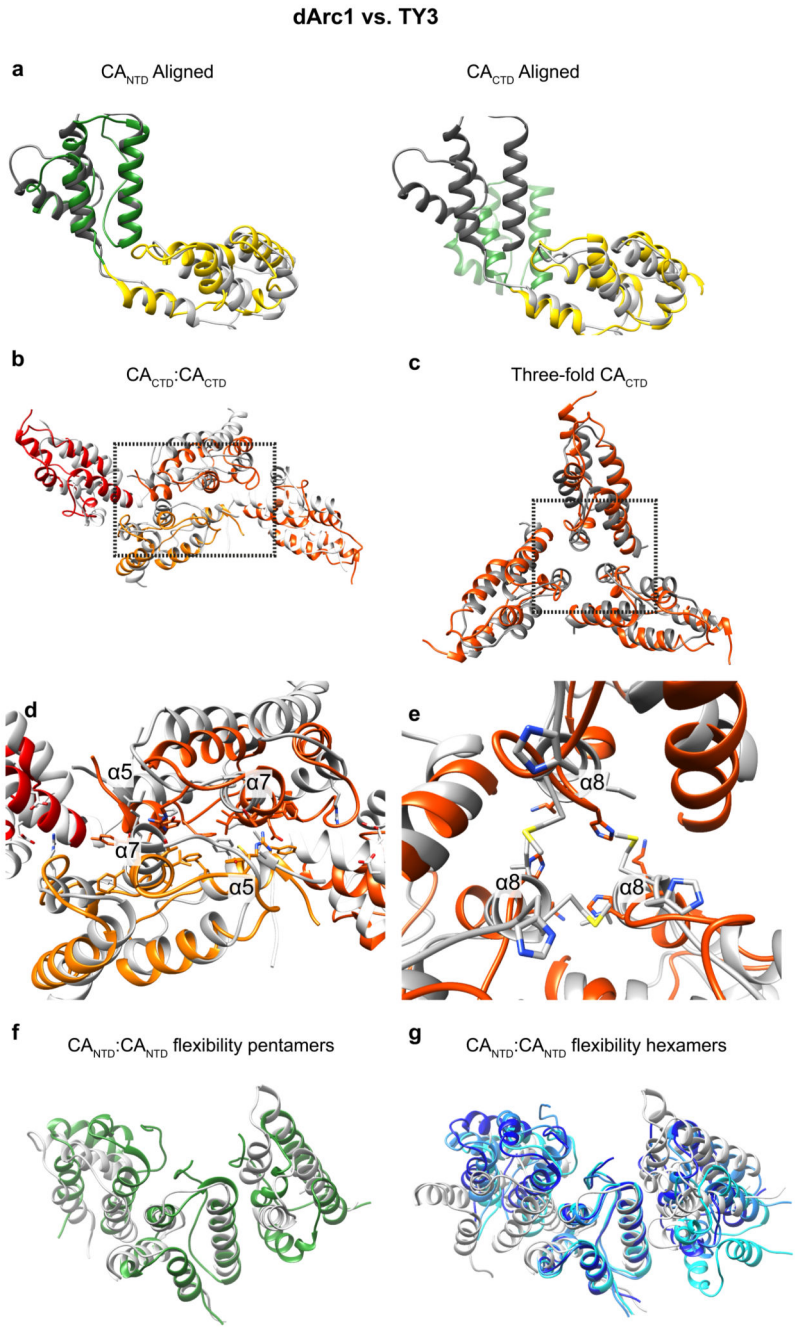
four different CA_{NTD}:CA_{CTD} interfaces, with the CA_{CTD} aligned. There are only subtle movements of the neighboring residues in the CA_{NTD} relative to the CA_{CTD}. For CA_{NTD} residues involved in the interface (97-103, 117-119, 53-58), C^{α}-RMSD: 4.8 Å. For the full adjacent CA_{NTD}, C^{α}-RMSD: 6.9 Å. **g)** The two different conformations of the CA_{CTD}:CA_{CTD} interface. The CA_{CTD} domains forming the interfaces between adjacent capsomeres are less variable than the interfaces created between CA_{NTD}:CA_{CTD} and CA_{NTD}:CA_{NTD} shown above. C^{α}-RMSD for the two CTD domains forming the dimeric interface: 0.4 Å. **h)** Alignment of the three-fold CA_{CTD} and pseudo-three-fold CA_{CTD} axes. The histidines are positioned 6 Å apart in the true three-fold and 6 or 10 Å apart in the pseudo three-fold axes.

Europe PMC Funders Author Manuscripts



Extended Data Fig. 6. Alignment of the dArc1 and mature HIV-1 structures.

a) CA from the five-fold capsomere of the dArc1 capsid aligned with CA from the HIV-1 CA pentamer from mature virions (PDB ID: 5MCY) [5]by either CA_{NTD} (HIV-1; light blue) or CA_{CTD} (HIV-1; dark blue). The HIV-1 CA_{NTD} is composed by 7 helical segments and an N-terminal β-Hairpin. α1-α4 in dArc CA_{NTD} correspond to α2-α4 and α7 in the HIV-1 CA_{NTD}. The CA_{CTD} fold is well conserved. **b-c)** The arrangement of CA_{CTD} at the two and three-fold interfaces in the dArc capsid is similar to the arrangement at the corresponding positions in mature HIV-1 (PDB ID: 5MCX)[40]. The HIV-1 two-fold CA_{CTD} interface is constituted by a short 3_{10} helical segment and α9 which align with dArc α5 and α7, respectively. Mutations in HIV-1 CA α9 are known to disrupt virus assembly and maturation[41]



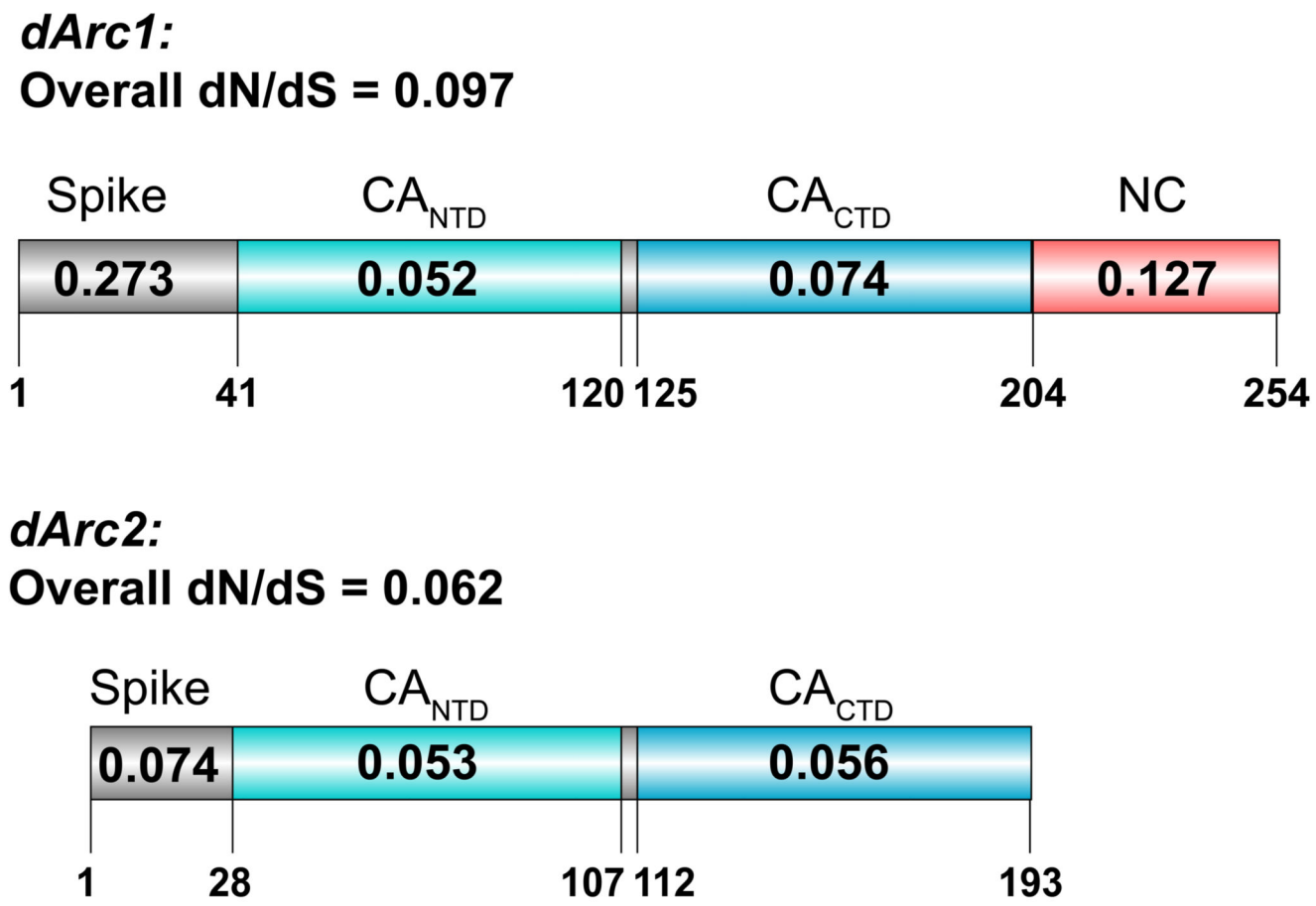
Extended Data Fig. 7. Alignment of the dArc1 and Ty3 structures.

a) CA from the five-fold capsomere of the dArc1 capsid aligned with CA from the five-fold capsomere of Ty3, either aligned on CA_{NTD} (Ty3; dark grey) or CA_{CTD} (Ty3; light grey). The structures of CA_{NTD} and the CA_{CTD} domains are very similar but differ in the inter-domain orientation. **b-c)** The arrangement of CA_{CTD} at the two and three-fold dArc interfaces is very similar to the arrangement at the corresponding positions in Ty3. **d)** The two-fold CA_{CTD}:CA_{CTD} contact surfaces for dArc1 and dArc2 are facilitated by hydrophobic stacking interactions, primarily between α5 and α7. The Ty3 interface also

involves hydrophobic residues in α5 and α7, but lacks direct α5 - α5 contacts. **e)** The Ty3 threefold CA_{CTD} interface also contains a three-fold symmetric histidine cluster (H170) positioned at the outer edge of the pore. **f)** Comparison of dArc1 (green) and the Ty3 (grey) CA_{NTD}:CA_{NTD} interfaces from pentameric capsomeres aligned by the central CA_{NTD}. **g)** Comparison of dArc1 (shades of blue) and the Ty3 (shades of grey) CA_{NTD}:CA_{NTD} interfaces from hexameric capsomeres aligned by the central CA_{NTD}.

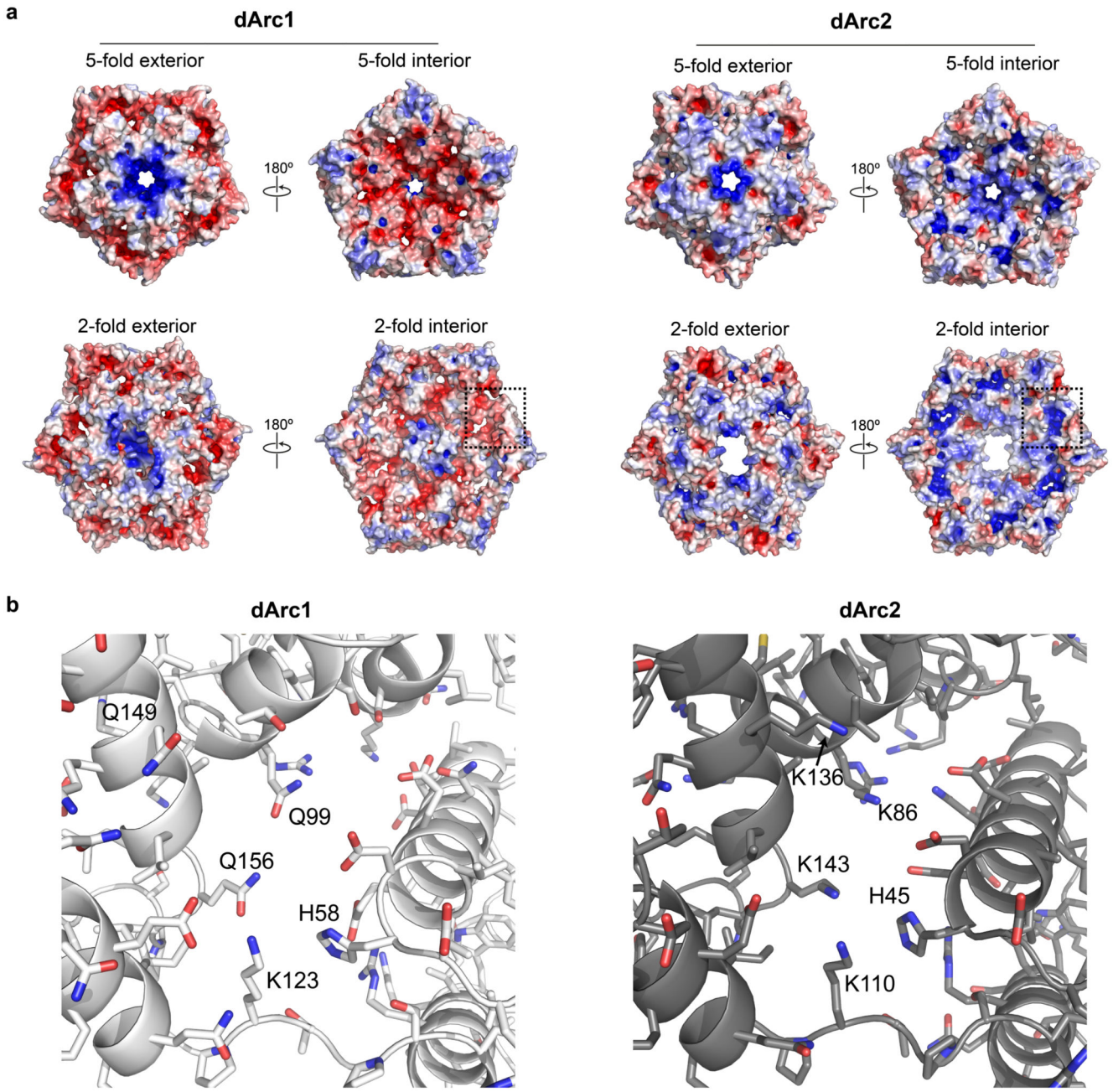
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts



Extended Data Fig. 8. Purifying selection is acting on all dArc1 and dArc2 domains.

The diagram shows dN/dS values computed using PAML (see Methods) for each of the structural domains of dArc1 and dArc2. A dN/dS ratio less than 1 is indicative of purifying selection. A Chi-Square Likelihood Ratio Test was used to compare the likelihood value of the calculated dN/dS to the likelihood that the sequence is under neutral evolution. All dN/dS values on dArc1 (n = 32 species) and dArc2 (n = 33 species) show statistically significant ($P<0.0001$) purifying selection compared with a neutral evolution model.



Extended Data Fig. 9. Electrostatic potential of individual capsomeres of dArc1 and dArc2.
a) As in Fig 3, the surface is coloured with the electrostatic potential from -5 (red) to +5 k_bT/e (blue). The dArc2 capsomeres are more electropositive than the dArc1 capsomeres. **(b)** Close-ups showing that additional neutral-to-basic residue substitutions facilitate electrostatic charge differences in the capsid interior.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Sjors Scheres, Takanori Nakane and Garib Murshudov (MRC-LMB) for advice on data processing and analysis, and Jake Grimmett and Toby Darling (MRC-LMB) for support with computing infrastructure (MRC-LMB). This study was supported by the MRC-LMB EM Facility. This work was funded by the Novo Nordisk Foundation (NNF17OC0030788: S.E.), the National Institute of General Medical Sciences (R35-GM122550: C.F.), the National Institute of Mental Health (R01-MH112766: J.D.S.), the Chan Zuckerberg Initiative (J.D.S), and the Medical Research Council (MC_UP_1201/16: J.A.G.B.).

References

1. Shepherd JD, Bear MF. New views of Arc, a master regulator of synaptic plasticity. Nat Neurosci. 2011; 14:279–284. [PubMed: 21278731]
2. Plath N, et al. Arc/Arg3.1 is essential for the consolidation of synaptic plasticity and memories. Neuron. 2006; 52:437–444. [PubMed: 17088210]
3. McCurry CL, et al. Loss of Arc renders the visual cortex impervious to the effects of sensory experience or deprivation. Nat Neurosci. 2010; 13:450–457. [PubMed: 20228806]
4. Pastuzyn ED, et al. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. Cell. 2018; 172:275–288.e18. [PubMed: 29328916]
5. Ashley J, et al. Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. Cell. 2018; 172:262–274.e11. [PubMed: 29328915]
6. Mosher J, et al. Coordination between Drosophila Arc1 and a specific population of brain neurons regulates organismal fat. Dev Biol. 2015; 405:280–290. [PubMed: 26209258]
7. Mattaliano MD, Montana ES, Parisky KM, Littleton JT, Griffith LC. The Drosophila ARC homolog regulates behavioral responses to starvation. Mol Cell Neurosci. 2007; 36:211–221. [PubMed: 17707655]
8. Abrusán G, Szilágyi A, Zhang Y, Papp B. Turning gold into 'junk': transposable elements utilize central proteins of cellular networks. Nucleic Acids Research. 2013; 41:3190–3200. [PubMed: 23341038]
9. Campillos M, Doerks T, Shah PK, Bork P. Computational characterization of multiple Gag-like human proteins. Trends Genet. 2006; 22:585–589. [PubMed: 16979784]
10. Huang CRL, Burns KH, Boeke JD. Active Transposition in Genomes. Annu Rev genet. 2012; 46:651–675. [PubMed: 23145912]
11. Bourque G, et al. Ten things you should know about transposable elements. Genome Biol. 2018; 19:199–12. [PubMed: 30454069]
12. Dodonova SO, Prinz S, Bilanchone V, Sandmeyer S, Briggs JAG. Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. Proc Natl Acad Sci USA. 2019; 116:10048–10057. [PubMed: 31036670]
13. Nielsen LD, Pedersen CP, Erlendsson S, Teilum K. The Capsid Domain of Arc Changes Its Oligomerization Propensity through Direct Interaction with the NMDA Receptor. Structure. 2019; 27:1071–1081.e5. [PubMed: 31080121]
14. Zhang W, et al. Structural Basis of Arc Binding to Synaptic Proteins: Implications for Cognitive Disease. Neuron. 2015; 86:490–500. [PubMed: 25864631]
15. Zhang W, et al. Arc Oligomerization Is Regulated by CaMKII Phosphorylation of the GAG Domain: An Essential Mechanism for Plasticity and Memory Formation. Molecular Cell. 2019; 75:13–25.e5. [PubMed: 31151856]
16. Eriksen MS, et al. Molecular determinants of Arc oligomerization and formation of virus-like capsids. bioRxiv. 2019; 73
17. De Guzman RN, et al. Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element. Science. 1998; 279:384–388. [PubMed: 9430589]
18. Amarasinghe GK, et al. NMR structure of the HIV-1 nucleocapsid protein bound to stem-loop SL2 of the psi-RNA packaging signal. Implications for genome recognition. Journal of Molecular Biology. 2000; 301:491–511. [PubMed: 10926523]

19. Mastronarde DN. SerialEM: A Program for Automated Tilt Series Acquisition on Tecnai Microscopes Using Prediction of Specimen Position. Microscopy and Microanalysis. 2003; 9:1182–1183.
20. Zheng SQ, et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat Meth. 2017; 14:331–332.
21. Rohou A, Grigorieff N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. J Struct Biol. 2015; 192:216–221. [PubMed: 26278980]
22. Zivanov J, et al. RELION-3: new tools for automated high-resolution cryo-EM structure determination. bioRxiv. 2018
23. Scheres SHW. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol. 2012; 180:519–530. [PubMed: 23000701]
24. Zivanov J, Nakane T, Scheres SHW. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. IUCrJ. 2019; 6:5–17.
25. Rosenthal PB, Henderson R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. Journal of Molecular Biology. 2003; 333:721–745. [PubMed: 14568533]
26. Kucukelbir A, Sigworth FJ, Tagare HD. Quantifying the local resolution of cryo-EM density maps. Nat Meth. 2014; 11:63–65.
27. Ramírez-Aportela E, et al. Automatic local resolution-based sharpening of cryo-EM maps. bioRxiv. 2018
28. Webb B, Sali A. Protein Structure Modeling with MODELLER. Methods Mol Biol. 2017; 1654:39–54. [PubMed: 28986782]
29. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta crystallographica Section D, Biological crystallography. 2004; 60:2126–2132. [PubMed: 15572765]
30. Adams PD, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta crystallographica Section D, Biological crystallography. 2010; 66:213–221. [PubMed: 20124702]
31. Afonine PV, et al. Real-space refinement in PHENIX for cryo-EM and crystallography. Acta Crystallogr D Struct Biol. 2018; 74:531–544. [PubMed: 29872004]
32. Pettersen EF, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004; 25:1605–1612. [PubMed: 15264254]
33. Touw WG, van Beusekom B, Evers JMG, Vriend G, Joosten RP. Validation and correction of Zn-CysxHisy complexes. Acta Crystallogr D Struct Biol. 2016; 72:1110–1118. [PubMed: 27710932]
34. Williams CJ, et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. 2018; 27:293–315. [PubMed: 29067766]
35. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology. 2000; 302:205–217. [PubMed: 10964570]
36. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Research. 2014; 42:W320–4. [PubMed: 24753421]
37. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004; 5:113–19. [PubMed: 15318951]
38. Li T, et al. The Changes of Positive Selection Within env Gene of HIV-1 B', CRF07_BC and CRF08_BC from China Over Time. Curr HIV Res. 2017; 15:31–37. [PubMed: 27917706]
39. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016; 33:1870–1874. [PubMed: 27004904]
40. Mattei S, Glass B, Hagen WJH, Kräusslich H-G, Briggs JAG. The structure and flexibility of conical HIV-1 capsids determined within intact virions. Science. 2016; 354:1434–1437. [PubMed: 27980210]
41. Joshi A, Nagashima K, Freed EO. Mutation of dileucine-like motifs in the human immunodeficiency virus type 1 capsid disrupts virus assembly, gag-gag interactions, gag-membrane binding, and virion maturation. J Virol. 2006; 80:7939–7951. [PubMed: 16873251]

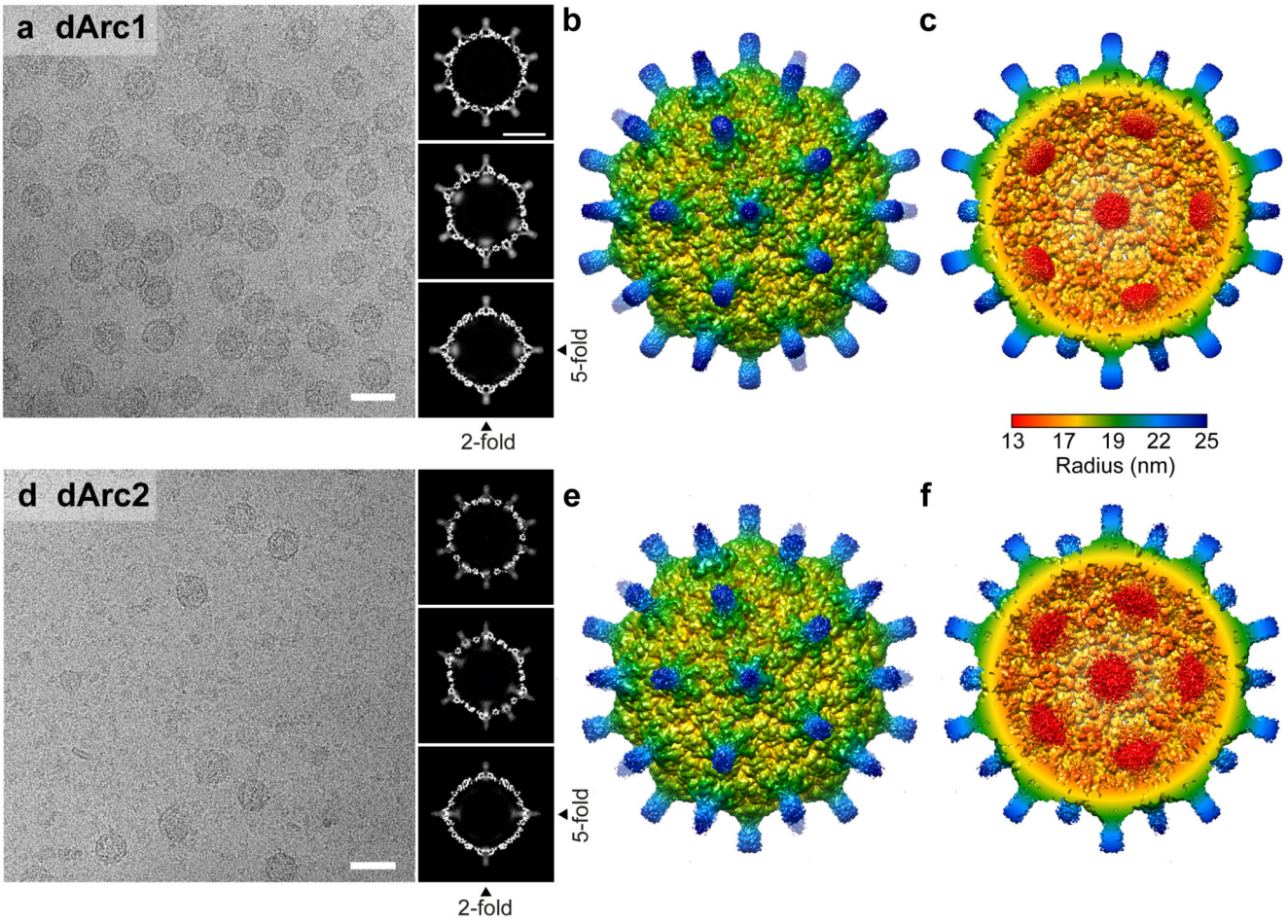


Fig. 1. The cryo-EM structures of dArc1 and dArc2.

a) Representative dArc1 micrograph and central sections through three-dimensional reconstructions perpendicular to the five-, three-, and two-fold axes. Scale bars are 50 nm for micrographs and 20 nm for sections. **b**) Surface representation of dArc1, viewed down the five-fold axis and colored by radius. **c**) As in **b**, with the front half of the capsid removed to reveal internal features. **d-f)** As in **a-c**, but for dArc2.

Europe PMC Funders Author Manuscripts

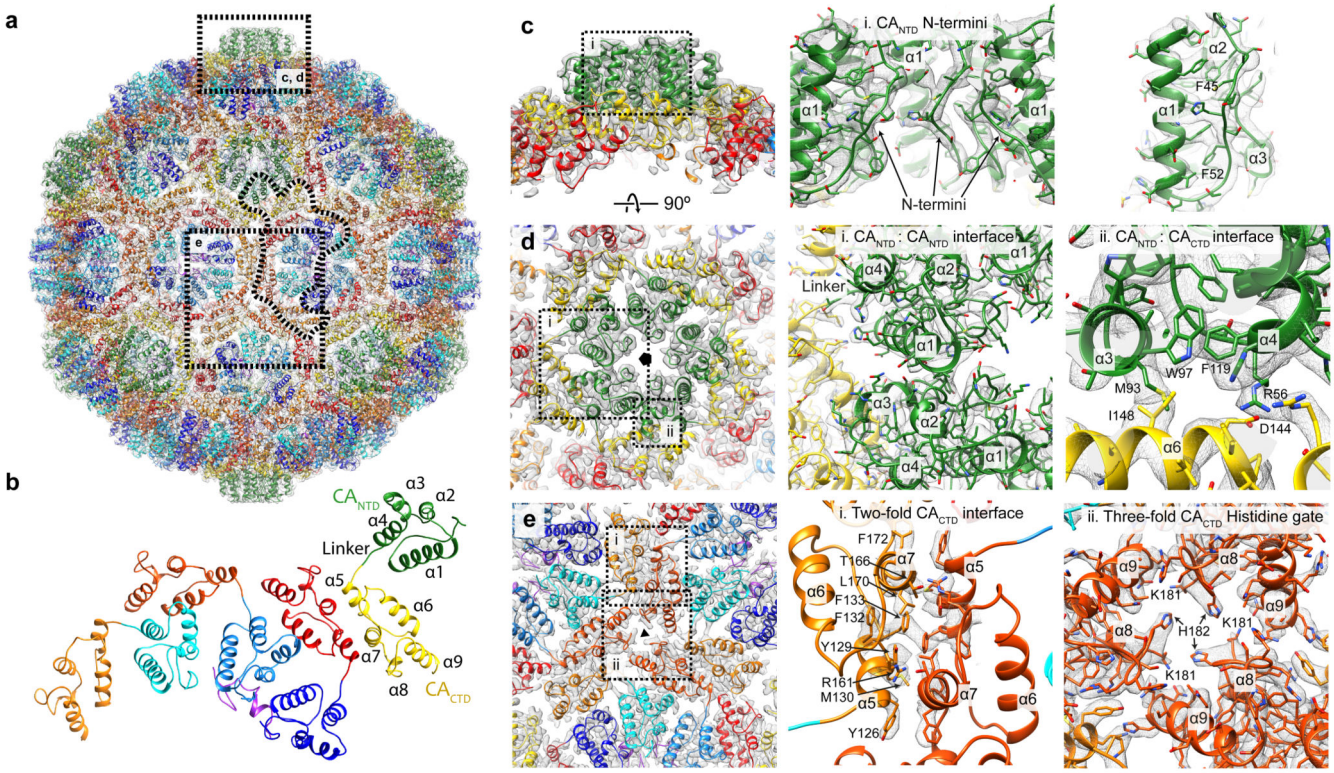


Fig. 2. Atomic model of the dArc1 capsid.

a) The 12 five-fold capsomeres are colored green (CA_{NTD}) and yellow (CA_{CTD}). The 30 two-fold capsomeres are colored cyan to blue (CA_{NTD}) and orange to red (CA_{CTD}). The zinc fingers are colored purple. EM density is transparent grey. **b)** The asymmetric unit containing four CA molecules and one zinc finger. 60 asymmetric units including a total of 240 CA molecules and 60 zinc fingers make up the T=4 capsid. **c)** Close-up of the five-fold capsomere (outlined in **a**) i) Cut-away showing three of the five N termini in the centre of the capsomeres. The N termini extend into and form the capsid spikes. The N termini are stabilized by docking into an extended hydrophobic groove adjacent to α1. **d)** External view of the five-fold capsomere i) The CA_{NTD}:CA_{NTD} interaction between α1, and α2 and α3 of the neighbouring CA molecule in the capsomere. ii) The CA_{NTD}:CA_{CTD} interface involves α6 in the CA_{CTD} and α3 and α4 in the neighbouring CA_{NTD}. This interface involves hydrophobic and electrostatic interactions. **e)** External view of the two and three-fold CA_{CTD} interfaces. i) The two-fold CA_{CTD} interface connects two adjacent capsomeres and is dominated by hydrophobic π-π stacking interactions. The interface involves residues from α5 and α7. EM density is only shown for the interface contact area. All CA_{CTD} interfaces are highly similar (Extended Data Fig. 5g). ii) The three-fold CA_{CTD} axis. At this position, basic residues from α8 surround the largest gap in the capsid. Corresponding views for dArc2 are in Extended Data Fig. 1.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

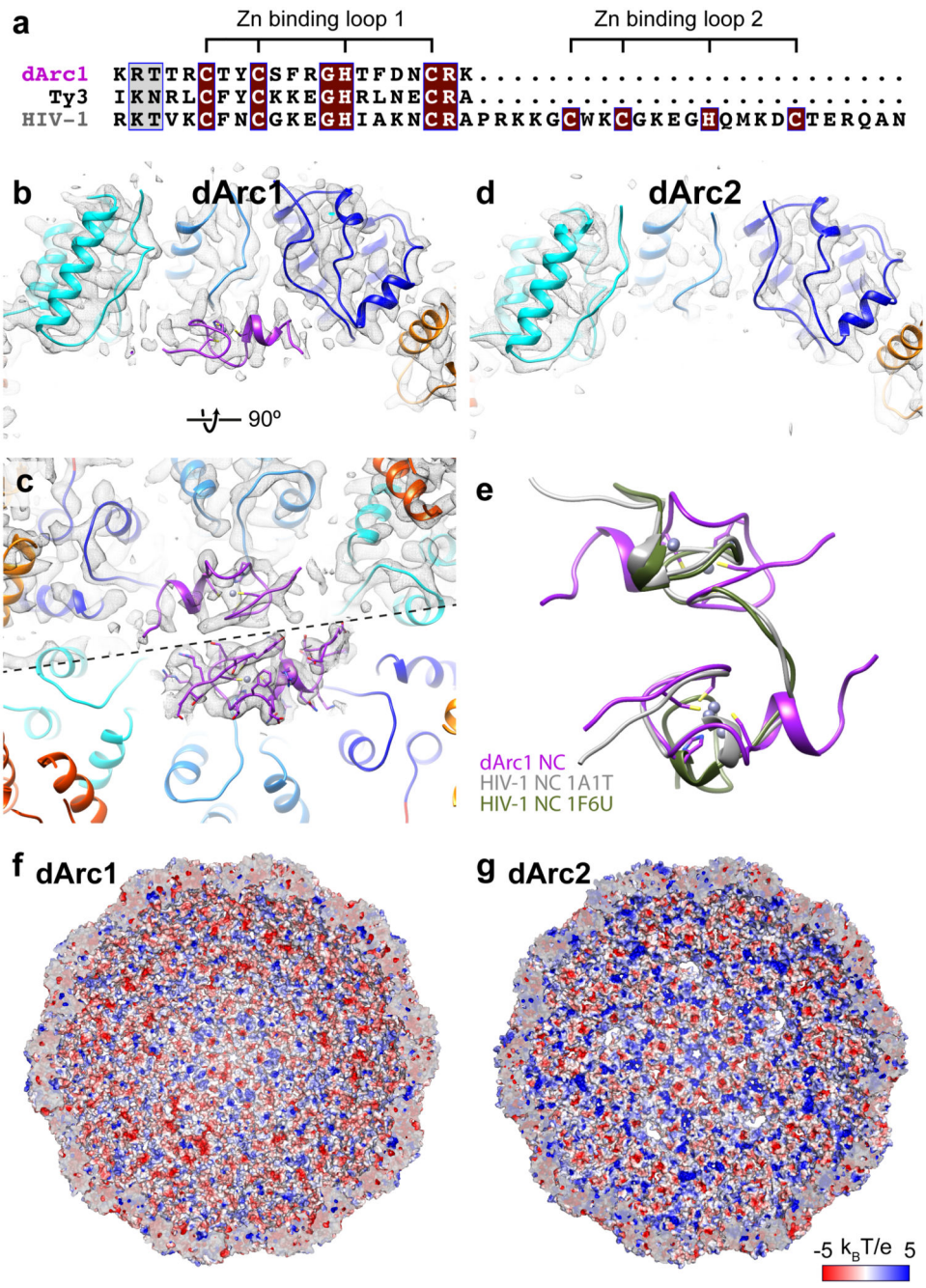


Fig. 3. The dArc1 NC domain forms a zinc finger bound to CA.

a) sequence alignment between the zinc finger regions of dArc1, Ty3 and HIV-1. dArc1 coordinates zinc via C227, C230, H235 and C240. **b)** Cross-section of the two-fold capsomere showing the fit and position of the zinc finger (purple). **c)** Viewed from the inside of the capsid, two zinc-fingers bind beneath the two-fold capsomeres. The lower half of the panel shows EM density for the zinc finger at a lower isosurface threshold. **d**) dArc2 lacks the NC domain, and EM density for the zinc finger is absent. **e)** Overlay of the two copies of the dArc1 zinc fingers at the two-fold capsomere (purple), and the HIV-1 double zinc finger

motif in the SL2-bound (green) and SL3-bound (grey) configurations. **f-g**) Inner surfaces of the dArc1 and dArc2 capsids colored by electrostatic potential from -5 (red) to +5 k_bT/e (blue).

Europe PMC Funders Author Manuscripts