bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Structure of *D. melanogaster* ARC1 reveals a repurposed molecule with characteristics of retroviral Gag

Matthew A. Cottee[1], Suzanne C. Letham[1], George R. Young[2], Jonathan P. Stoye[2,3] and Ian A. Taylor[1*]

[1]Macromolecular Structure Laboratory, [2]Retrovirus-Host Interactions Laboratory, The Francis Crick Institute, Midland Road, London NW1 1AT, UK. [3]Department of Medicine, Imperial College London, London SW7 2AZ, UK

*To whom correspondence should be addressed

E-mail: ian.taylor@crick.ac.uk

Telephone: +44 (0) 20 3796 2288

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

16 **Running Title**: Structure of *D. melanogaster* ARC1

17

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ABSTRACT

The tetrapod neuronal protein ARC and its *D. melanogaster* homologue, dARC1, have important but differing roles in neuronal development. Both are thought to originate through exaptation of ancient Ty3/Gypsy retrotransposon Gag genes, with their novel function relying on an original capacity for self-assembly and encapsidation of nucleic acids. Here, we present the crystal structure of dARC1 CA and examine the relationship between dARC1, mammalian ARC and the CA protein of circulating retroviruses. We show that whilst the overall architecture is highly related to that of orthoretroviral and spumaretroviral CA, there are significant deviations in both N- and C-terminal domains, potentially affecting recruitment of partner proteins and particle assembly. The degree of sequence and structural divergence suggests that Ty3/Gypsy Gag has been exapted on two separate occasions and that, although mammalian ARC and dARC1 share functional similarity, the structures have undergone different adaptations after appropriation into the tetrapod and insect genomes.

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

33 **Keywords**: ARC; Gag; CA; Neuronal development; TY3/Gypsy.

34

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

INTRODUCTION

Activity-Regulated Cytoskeleton-Associated protein (ARC) is an immediate early gene product induced in response to high levels of synaptic activity and is directed to neuronal synapses through signalling sequences in its 3' UTR (*1*). Mammalian ARC (mam-ARC) is essential for neuronal plasticity and is involved in memory (*2, 3*) acting as a regulator of AMPA receptors (*4, 5*). ARC has also been implicated in neurological disorders including Alzheimer's disease (*6, 7*), Fragile-X syndrome (*8*) and schizophrenia (*9-11*). In *Drosophila melanogaster*, two homologues of mam-ARC are expressed, dARC1, and dARC2 (*12*). dARC1 is present at neuromuscular junctions and, along with its mRNA, has been implicated in regulating the behavioural starvation response but is not involved in synaptic plasticity (*13*). Therefore, comparing the structural and functional properties of mam-ARC and dARC1 might lead to a better understanding of cognition and memory consolidation.

The *ARC* gene is thought to be derived from the *gag* gene of a Ty3/Gypsy retrotransposon (*14*) that, subsequent to genomic insertion, has been repurposed to perform an advantageous function to the host (*15*). This connection between ARC and retrotransposons was made when sequence alignments revealed that the ARC proteins shared sequence similarity with the Gag protein of retroviruses or retrotransposons (*14*). These data also suggested that ARC is evolutionarily related to the Ty3/Gypsy family of retrotransposons. Further evidence came from crystal structures of two α-helical domains from *Rattus norvegicus* ARC (rARC) (*16*), which revealed that rARC N- and C-terminal capsid (CA) domains were structurally homologous to the N- and C-terminal CA domains of both *Orthoretrovirinae* (*16*) and also *Spumaretrovirinae* (*17*). Further phylogenetic analysis revealed that despite mam-ARC and dARC1 seemingly providing related functions in the host, dARC1 and the tetrapod ARCs most likely arose from separate

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

lineages of Ty3/Gypsy, since dARC1 clustered with insect Ty3/Gypsy retrotransposons and tetrapod ARCs clustered with fish Ty3/Gypsy retrotransposons (*15*).

The relevance of ARC's retrotransposon origin to its function in synaptic plasticity was not immediately obvious until the recent observation that mam-ARC and dARC1 can self-assemble into particles and package RNA for potential transfer between cells (*12, 15*), similarly to retrotransposons and retroviruses (*18, 19*). In *D. melanogaster*, it is proposed that dARC1 expressed at neuromuscular junction presynaptic boutons assembles into particles that encapsidate *dARC1* mRNA. Loaded particles might then be packaged and released as extracellular vesicles for intercellular transfer to the post synapse where mRNA release and translation can take place (*12, 15*), Similarly, mam-ARC can also encapsidate ARC mRNA into particles, allowing transfer from donor to recipient neurons, where ARC mRNA can be translated (*15*).

Since both dARC1 and mam-ARC are able to form capsid-like particles (*12, 15*), it seems likely that they share a degree of structurally similarity. To date, crystal structures of the individual domains from rARC have been determined (*16*) along with the solution NMR structure of the rARC CA (*20*). Here we report two crystal structures of the entire CA region of dARC1 at 1.7 Å and 2.3 Å and consider these structures in comparison to those of rARC and retroviral CA. dARC1 comprises two α-helical domains with a fold related to that observed in the CA-NtD and CA-CtD of orthoretroviral and spumaretroviral CA. However, we observe significant divergence in the NtD of dARC1 where an extended hydrophobic strand that packs against α1 and α3 of the core fold replaces the N-terminal β-hairpin and helix α1 found in orthoretroviral CAs. In the rARC structure, this hydrophobic strand is replaced by peptides from the binding partners Ca^{2+}/calmodulin-dependent protein kinase 2A (CamK2A) and transmembrane AMPAR regulatory protein γ2 (TARPγ2) and may represent a functional adaptation for the recruitment of partner proteins. We also show that dARC1 utilises the same CtD-CtD interface required for assembly of retroviral

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CA into mature particles and propose that this obligate dimer represents a building block for dARC1 particle assembly. Further examination of the relationship between dARC1, mam-ARC and Gag from Ty retrotransposon families, reveals that, although dARC1 and mam-ARC are functional orthologues, the structural divergence in dARC1 and mam-ARC CA domains is consistent with the notion of Ty3/Gypsy Gag exaptation on two separate occasions. We suggest they may have undergone different adaptations after appropriation into the tetrapod and insect genomes.

RESULTS

Structures of dARC1 CA

We determined the crystal structure of the capsid domain region of dARC1, residues S39-N205 (dARC1 CA) using single wavelength anomalous diffraction (SAD) and crystals of Se-Met substituted protein. The structure was determined in both an orthorhombic and a hexagonal crystal form. The orthorhombic crystals diffracted to higher resolution allowing the structure to be refined to a final resolution of 1.7 Å with R-factor and Free R-factor of 18.1% and 21.3% respectively. Details of data collection, phasing and refinement are presented in **Table S1**. The asymmetric unit (ASU) contains two chains, each containing an α-helical N- (CA-NtD) and C-terminal domain (CA-CtD) (**Fig. 1A**). The chains are arranged in a dimer with a distinct U-shape reminiscent of a glacial trough (**Fig. 1A, right**). The CA-CtDs form the base of the trough and pack together to form a homodimer interface and the CA-NtDs form the sides of the trough and are separated by ~45 Å. Inspection of each domain reveals the CA-NtD is made up from an extended N-terminal strand and a four-helix core (α1-α4) and the CA-CtD comprises a further five α-helix bundle (α5-α9) (**Fig. 1B, i, ii**). The tertiary folds of each domain are particularly similar and can be superimposed with an RMSD of 2.2 Å over 49 Cα atoms (**Fig. 1C**). Moreover, it can be seen that the dARC1 N-terminal β-strand is topologically equivalent to

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

112 α5 in the CtD, while NtD α1-4 are equivalent to CtD α6-9. This strong similarity of dARC1
113 CA domains provides further evidence for the notion that tandem domains of CA arose as
114 the result of a gene duplication event (*17*). The hexagonal crystal form was independently
115 solved and refined to a resolution of 2.3 Å and reveals an almost identical dimeric ASU
116 that aligns with a RMSD of only 0.247 Å over 133 Cα pairs (**Fig. S1A to C**). Both
117 structures appear especially stable around the CTD-mediated dimeric interface and when
118 aligned through their CTDs show only small differences in the positioning of NtDs with
119 respect to the CtDs (**Fig. S1D**).
120
121 *The dARC1 CA dimer interface*
122 The dARC1 CA-CtD monomer consists of a five-helix core comprising α5 (residues
123 A125-Q134), α6 (residues I143-Q156), α7 residues (E164-L171), α8 (residues I177-H182)
124 and α9 (residues F191-N204). The dimer interface is located between CA-CtDs where the
125 outer surfaces of α5 and α7 pack against α5' and α7' of the opposing monomer (**Fig. 2A)**.
126 The homodimer interface encompasses 1005 $Å^2$ of buried surface and is defined by
127 numerous intermolecular interactions. The interface is largely hydrophobic with
128 contributions from sidechain packing of the Y126, Y129, M130, F133, L170, F172 and
129 L174 hydrophobic and aromatic residues that are exposed on α5 and α7 and form a
130 continuous apolar network with a Y129 and F133 at its centre (**Fig. 2A**). This is apparent
131 in the analysis of the dARC1 CA surface hydrophobicity profile reveals a distinct apolar
132 patch that locates to the centre of the CA-CtD homodimer interface (**Fig. S2A**). In addition,
133 at the periphery of the interface, there is also a salt bridge between R161 on the α6-α7
134 connecting loop with D169 at the C-terminus of α7 providing further stabilisation (**Fig. 2A**).
135 The number and hydrophobic nature of interactions within the homodimer interface
136 suggest the dimer constitutes a relatively stable or obligate structure.
137

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Self-association of dARC1 CA

Given the unexpected nature of the dimer observed in the crystal structure, the solution molecular mass, conformation and self-association properties of dARC1 CA were examined using a variety of solution hydrodynamic methods. Initial assessment by Size Exclusion Chromatography coupled Multi-Angle Laser Light Scattering (SEC-MALLS) was performed with protein concentrations ranging from 25-400 μM that yielded an invariant solution molecular weight of 40.0 kDa for dARC1 CA (**Fig. 2B**). By comparison, the dARC1 CA sequence-derived molecular weight is 19.6 kDa. Given this value, together with the lack of a concentration dependency of the molecular weight, it is apparent that dARC1 CA also forms strong dimers in solution. To confirm and better analyse dARC1 CA oligomerisation, the hydrodynamic properties were analysed using sedimentation velocity (SV-AUC) and sedimentation equilibrium (SE-AUC) analytical ultracentrifugation. A summary of the experimental parameters, molecular weights derived from the data, and statistics relating to the quality of fits are shown in **Table S2**. Analysis of the sedimentation velocity data for dARC1 CA using both discrete component and the C(S) continuous sedimentation coefficient distribution function **Fig. 2C** revealed a predominant single species with $S_{20,w}$ of 2.92±0.03 S and no significant concentration dependency of the sedimentation coefficient over the range measured (25-90 μM). These data show that dARC1 CA comprises a single stable 2.92 S species with a molecular weight derived from either the C(S) function or discrete component analysis ($S_{20,w}/D_{20,w}$) of 38 kD, **Table S2**, consistent with a dARC1 CA dimer. The frictional ratio (f/f_o) obtained from the analysis of the sedimentation coefficients is 1.41 (**Table S2)**, suggesting the solution dimer has an elongated conformation and is consistent with the U-shaped conformation observed in the crystal structures. Moreover, analysis of the crystal structure using HYDROPRO (*21*) gives calculated $S_{20,w}$ and $D_{20,w}$ values, in close agreement with that observed in solution (**Table S2**), supporting the idea that the dimer observed in the crystal structures is wholly

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

representative of the solution conformation. To further ascertain the affinity of dARC1 CA self-association, multispeed SE-AUC studies at varying protein concentration were carried out and typical equilibrium distributions for dARC1 CA are presented in **Fig. 2D**. Analysis of individual gradient profiles showed no concentration dependency of the molecular weight and so all the data were fitted globally with a single ideal molecular species model, producing a weight averaged molecular weight of 38.9 kDa, **Table S2**. The lack of any concentration dependency precludes any analysis of homodimer affinity but confirms that dARC1 CA forms a stable dimeric structure that has the expected properties of the dimer we observe in the crystal structure.

Attempts to mildly disrupt the central apolar network by introduction of an F133A mutation had no effect on dimerization, when assessed by SEC-MALLS (**Fig. S2B**). More aggressive mutations F133A+Y129A and F133A+R161A resulted in complete loss of protein solubility and an inability to purify the constructs, further suggesting that in dARC1 CA homodimerization is requirement for protein folding/structural integrity and likely forms a key building block of dARC1 particle assembly. Analysis of the electrostatic surface potential of the dimeric structure reveals a differential distribution of charge where the surface of the glacial trough has a net negative charge that spreads across both domains of each dARC1 and the underside were the C-terminus projects has a more positively charged character (**Fig. S2C**), suggesting that upon assembly dARC1 particles would have a negatively charged exterior and more positively charged interior were nucleic acid is contained.

Comparison with mammalian ARC CA structure.

Given that mam-ARC and dARC1 share functional similarities, we assessed the relationship between rARC and dARC1 by comparing the dARC1 structure with the individual domains from rARC. Overall, the alignments are excellent, reflecting the

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

190 evolutionary relationship, but there are significant differences between dARC1 and rARC
191 in both their NTDs and CTDs.

192 There are two crystal structures of the rARC NTD in complex with peptide ligands
193 (PDB 4X3H and 4X3I) (Zhang et al., 2015), and a recent solution NMR structure (6GSE,
194 Nielsen et al., 2019) of the entire rARC CA domain that resolves the NTD in an apo form.
195 Superficially, the dARC1 CA-NTD aligns well with all available structures of the rARC CA-
196 NTD, with DALI Z-scores and RMSDs of 8-10 and 1.5-1.9 Å respectively (**Fig. 3A**)

197 Examination of the dARC1 CA-NtD reveals an N-terminal extended strand (NT-
198 strand), residues G43 to R56, with a short β-configuration that packs against the core of
199 the NtD. The NT-strand makes many interactions with the apolar and aromatic side chains
200 that extend from α1, α2 and α4, burying 803 $Å^2$ of surface in the interface (**Fig. 3A and Bi,**
201 **Fig. S3A**) and the same configuration is observed in all four instances of the NtDs that we
202 see in our two crystal structures (**Fig. S3B**). The NT-strand residues are highly conserved
203 in *dARC* genes across Drosophilidae but not with the Mam-ARCs (**Fig. S3C)**. In particular,
204 two highly conserved aromatic residues, F45 and F52, are entirely buried, surrounded by
205 the conserved side chains of F64, L89, I115, and F119 and act to anchor the NT-strand
206 into the hydrophobic α1-α4 cleft of the CA-NtD. In addition, there is a mainchain interaction
207 between the backbone amide and carbonyl of F52 with the carbonyl of L89 and the amide
208 of Y91 that further stabilises the conformation of the NT-strand (**Fig. 3Bi and Fig. S3A**).

209 In apo-rARC CA-NTD (6GSE) the helical core aligns very well with the
210 corresponding region of dARC1 (RMSD = 1.45 Å). However, here the rARC NT-strand,
211 residues D210-E216, has a disordered conformation (**Fig. 3, A and Bii**) and the α1-α4
212 hydrophobic cleft, which in dARC1 contains the native NT-strand, is unoccupied in rARC
213 suggesting there is a functional divergence for the NT-strand between the dARC1 and
214 mam-ARC families. This notion is supported by inspection of the rARC CA NtD-TARPγ2
215 and CA NtD-CaMK2B complexes (4X3H and 4X3I) where the α1-α4 cleft of rARC is now

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

occupied by the bound TARPγ2 or CaMK2B derived peptides (**Fig 3B iii and iv**) and the bound peptides adopt the same extended β-configuration as the native NT-strand in the dARC1 structure (**Fig. S3D)** and bury a comparable amount of surface, 772 Å and 641 Å, respectively. Moreover, both bound peptides contain an aromatic residue equivalent to dARC1 F52, Y229 in TARPγ2 and F313 in CaMK2B, that packs into the core of rARC CA-NtD and makes an identical main-chain interaction with the backbone carbonyl of H245 and the amide of N247 as that observed between the backbone amide and carbonyl of F52 with the carbonyl of L89 and the amide of Y91 in dARC1 (**Fig. S3D**). In these peptide-complex structures, the rARC NT-strand, D210-E216, that is disordered in the apo-structure now adopts a parallel β-configuration to pack against the bound peptides (**Fig. 3, Biii and Biv**) and it is possible that the propensity to form this stabilising beta configuration has been selected for. This notion is supported by inspection of the dARC and mam-Arc multiple sequence alignment (**Fig. S3C**) that reveals a conserved "TQIF" motif in Amniota that retains β-branched residues, favoured in beta structure, at the T and I position. This motif is not present in amphibians or in *L. ch* Gypsy2 the closest known relative to the transposon from which tetrapod ARC was exapted, suggesting that this feature, and possibly peptide binding ability, arose within Amniota.

The structures of dARC1 CA-CtD and rARC CA-CtD (PDB: 4X3X) also superimpose well (RMSD = 2.7 Å). However, the CtD of the apo-rARC CA NMR structure more closely matched the structure of dARC1 CA-CtD, (RMSD = 2.2 Å), with all 5 helices overlaying (**Fig. 4A**). However, in contrast to our solution studies of dARC1 (**Fig. 2, A to C and Fig S2A**) the rARC CA domain was monomeric in solution, even at the high concentrations under which NMR was performed (*20*).

In dARC1, a large proportion of the CTD dimer interface results from the packing of hydrophobic side chains projecting from helices 5 and 7 (**Fig. 2A**). However, upon comparison of external α5/α7 surfaces of dARC1 and rARC (**Fig. 4, B and C**) it is

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

apparent that the exposed Y126, Y129, M130, F133, L170, F172 and L174 side chains that are responsible for the hydrophobic character of the dARC1 dimer interface are not conserved in rARC and are replaced by E282, Q285, R286, D289, Y324, V326 and T328 in rARC. Indeed, the hydrophobic patch present on the surface of dARC1 is not evident in the same surface on rARC (**Fig. 4, D and E**). In addition, R161 and D169, which make a salt bridge interaction in the dARC1 interface, are also not conserved, being replaced by D315 and Q323 in rARC (**Fig. 4, B and C**). These sequence differences are also apparent throughout the entire dARC and mam-ARC families. Here, there is strong sequence conservation of residues that constitute the core fold of the CA-CtD across both dARC and mam-ARCs but the hydrophobic CA-CtD dimer interface residues are only present in the dARC lineage (**Fig. 5F**). Taken together, these data reveal that, while tertiary structure topology of dARC1 and rARC CA-CtDs is conserved, there are significant differences in the character of the surface that is presented around α5-α7; in dARC1, the hydrophobic nature of this surface drives the formation of a strong CTD dimer, whereas in rARC, the more polar nature of this surface may explain why the protein is monomeric in solution. Given these differences, although there is strong evidence for assembly of both dARC1 and mam-ARC into capsid-like particles (*12, 15*), it seems likely that if dARC1 and mam-ARC use the α5/α7 interface in a particle assembly pathway, the interface may be substantially weaker for mam-ARC.

Retroviral CA domain structures are related to both dARC1 CA domains

The topology of the alpha-helical two-domain fold of dARC1 is highly reminiscent of retroviral CA structures. Interrogation of the PDB database with dARC1 CA using the DALI alignment/search engine (*22*) produced an overwhelming number of matches to Gag proteins, (87%, Z-score ≥ 5.0) and identified rARC, together with many orthoretroviral and spumaretroviral CA-NtD and CA-CtD structures. Alignments with CA-NtDs and CA-CtDs from HIV CA, Rous Sarcoma virus (RSV) CA and prototypic foamy virus (PFV) CA are

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

presented in **Fig. 5.** The best structural alignments to dARC1-NtD were with retroviral Gag CA-CtD structures rather than Gag CA-NtD structures (**Fig. 5, A and D**) indicating that the dARC1 CA-NtD is more closely related to the orthoretroviral CA-CtD than it is to the orthoretroviral CA-NtD. Alignments with dARC1-CtD also had the best structural alignment with orthoretroviral Gag CA-CtD structures (**Fig. 5, B and D**), perhaps not surprising given the observation of close resemblance of the dARC1 CA-NtD to the dARC1 CA-CtD (**Fig. 1Biii**). Alignments with PFV CA-NtD and CA-CtD were also found (**Fig. 5C and D**) and although not as significant as with the orthoretroviral CA, these data support previous observations of a relationship of spumaretroviral Gag with mam-ARC (*17*).

These data provide evidence for a structural conservation between orthoretroviral CA and ARC proteins and the weaker alignments observed with orthoretroviral CA-NtDs suggest that orthoretroviral CA-NtDs have undergone much more structural divergence than has occurred in the Ty3 family or ARC proteins. Moreover, these data further support the previously proposed idea that a duplication of a CA-CtD progenitor first gave rise to double domain ancestors and that subsequent divergence of domains resulted in spumaretroviral, orthoretroviral, and Metaviridae-derived proteins such as ARC, that are found presently (*17, 23*).

The dARC1 CtD dimer is an ancient assembly interface conserved in orthoretroviridae

Given the existence of the dARC1 CA dimer and the distant relationship with orthoretroviral CA, we next looked to see if the dimer interface was conserved between dARC1 and the CtD dimers of HIV-1 CA and RSV CA that are known to be essential for capsid assembly in orthoretroviruses. Cartoon representations of the dARC1, HIV-1 and RSV CA-CtD dimers are shown in **Fig. 6, A to C**. In each, the domain arrangement that presents the dimer interface is the same and this also seen in the CA-CtD dimer of native Ty3 particles visualised by cryo-electron microscopy (*24*) but with some repositioning of

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the CA-NtDs (**Fig. S4**). The structures have been aligned to find the best Cα alignment over the entire dimer (HIV, RMSD = 2.8 Å over 117 Cα; RSV, RMSD = 3.1 Å over 101 Cα) (**Fig. 6, D and E**) and it is apparent that each interface is made up from interactions between residues on CtD helices α5 and α7 of dARC1, that correspond to α7' and α8 in the orthoretroviral CA-CtD structures. Notably in the orthoretroviruses, α7' is reduced to a single turn and the monomers are rotated with respect to each other. Therefore, in dARC1 residues on α5 and α7 contribute equally to the interface whilst in the orthoretroviruses α8 contributes more to the interface than α7'. This combination of the larger contribution of α5 in dARC1, together with the rotation and displacement of CA-CtDs seen in the orthoretroviruses, has the effect of reducing the surface area that is buried at the interface from 768 $Å^2$ in dARC1 to 452 $Å^2$ in HIV-1. Notably, the homodimer affinity for orthoretroviral CA-CtD dimers is much weaker than the dARC1 dimer. Equilibrium dissociation constants ranging between 10-20 μM have been reported for HIV-1 (*25, 26*) and CA-CTD dimerisation is undetectable for other genera (*27-29*). Nevertheless, given the domain organisation and the similarity in character of the orthoretroviral and dARC1 CA-CtD dimers, we suggest this interface is a key building block of capsid assembly, retained in dARC1 and conserved from Ty3/Gypsy transposable elements through to orthoretroviridae.

DISCUSSION

dARC1 Capsid structures

Our crystal structures demonstrate that the central region of dARC1 contains two largely α-helical domains that, despite the lack of sequence conservation, have the same predominantly α-helical folds observed in the structures of CA domains from the ortho- and spuma-retroviruses. A more detailed inspection of dARC1 CA-NtD and CA-CtD reveal they comprise 4- and 5-helix bundles, respectively, with a topology that aligns well with the

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

arrangement of secondary structure elements observed in orthoretroviral CA NtDs and CtDs (**Fig. 5**). However, it is apparent that both the ARC CA-NtD and CA-CtD are much more closely related to the orthoretroviral CA-CtDs than they are to orthoretroviral CA-NtDs (**Fig. 5**), consistent with our previous notion that an ancient domain duplication was a key event during retrotransposon evolution (*17*). Notably, orthoretroviral CA-NtDs contain an extra N-terminal β-hairpin and an additional two helices compared to the ARCs and the CA domains of Ty3/Gypsy transposons (*24*) (**Fig. S4**). This suggests that unique aspects of the retroviral life cycle might be driving specific changes in structure of the retroviral CA-NtD. One such pressure might be associated with the process of maturation that follows retrovirus budding from the cell. Maturation involves proteolytic cleavage of immature viral cores followed by CA reassembly to yield mature virions and although it is proposed that dARC1 and mam-ARC transport mRNA between cells it is thought likely that particles are packaged into extracellular vesicles for cell-to-cell transfer (*12, 15*). Similarly, maturation events do not occur in Ty3 elements, that also do not bud from the cell, and that have Gag which assembles directly into mature forms (*24*). Absence of maturation also characterizes spumaviruses and it was observed previously that the CA NtD-equivalent region of PFV Gag showed greater similarity to rARC than to orthoretroviral CA (*17*).

Structural differences between insect and mammalian ARCs

Our 3D superimpositions have demonstrated that there is a large degree of structural conservation between the dARC1 and mam-ARC CA structures. However, despite this strong similarity, two regions of distinct differences between the dARC1 and rARC structure are apparent. The first concerns the ARC CA-NtD and the interaction with potential binding partners, the second, the putative dimerization domain of the CtD.

Functionally important interactions between mam-ARC and a variety of neuronal proteins, including the TARPγ2 and CaMK2B proteins, as well as the NMDA receptor,

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

have been defined (*16, 20*). However, no such interactions have been reported for dARC1. In the rARC structures with bound TARPγ2 or CaMK2B peptides the disordered N-terminal region of rARC seen in the apo structure, now forms a short parallel β-sheet with the bound peptide stabilising the peptide binding within a hydrophobic cleft on rARC. It is apparent that the conformation of these rARC-bound peptides strongly resembles the conformation of the NT-strand of dARC1 NtD (**Fig. 3**). Therefore, given the sequence differences in the NT-strand region between the dARC and mam-ARCs (**Fig. S3C**) one notion is that mam-ARC has evolved an N-terminal strand that no longer binds into the CA-NtD hydrophobic cleft but has gained the ability to promote the binding of synaptic protein ligands, perhaps acting as a sensor of synaptic stimuli. This sensing property might then contribute control to a functional role for ARC based on assembly and mRNA trafficking.

There are also significant differences between the dARC1 and rARC CA-CtD, illustrated in **Fig. 4.** Overall, our crystal structure of dARC1 and the NMR structure of full length rARC (Nielsen et al., 2019) are very similar, with good overlay in all five helices. However, inspection of the dARC1 surface reveals a substantial hydrophobic patch absent in rARC (**Fig. 4D and E**). This hydrophobic patch is shared with the orthoretroviruses (*25, 30*) and seems to be associated with the formation of stable dARC1 dimers, whereas rARC is monomeric. Whether this translates to differences in the stability of assembled particles *in vivo* remains to be determined, however it is possible that differences in the physiological roles of dARC1 and mam-ARC may mean that mam-ARC has evolved to require a weaker interface that facilitates disassembly. Alternatively, it is possible that mam-ARC may require a conformational change to facilitate dimerization or uses a completely different assembly mechanism that employs other surfaces of the molecule.

ARC particle assembly

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The observation that residues at the dARC1 CA-CtD interface are not conserved between the insect and mammalian ARC lineages suggests the possibility that, although mam-ARC particles have been observed *in vitro* and in cells, their mode of assembly may not utilise an obligate CA-CtD dimer as a building block. This type of observation has been made with orthoretroviruses that assemble through a combination of NtD-NtD, NtD-CtD and CtD-CtD interactions to form the viral capsid, where the relative contribution that different types of CA interaction make to the overall formation of the viral core varies depending on the retroviral genera. For instance, in lentiviruses, it is apparent that capsid assembly requires a strong intrinsic CtD-CtD dimeric interaction (*25, 30*). However, more generally, capsid shell formation requires three types of interaction, intra-hexamer NtD-NtD self-association (*30-33*), intrahexamer NtD-CtD interactions between adjacent capsid monomers (*30, 34, 35*) and inter-hexamer CtD-CtD interactions (*25, 30*). Therefore, it is entirely possible that in dARC1 and mam-ARC particles the relative contribution of each type of interfaces may also differ.

ARC exaptation

Mam-ARC and dARC1 appear to have different biological properties. However, it remains to be determined whether these differences result from the capture of two different Ty3/Gypsy elements or whether they reflect evolutionary adaptations. Perhaps the best studied example of the appropriation of retroelement encoded genes by mammalian hosts is the case of syncytin, a fusagenic protein essential for proper placenta formation (*36*). It is evident that syncytin capture appears to have occurred on multiple independent occasions, involving envelope proteins from clearly different retroviruses (*37, 38*), resulting in placentae with subtly different morphologies (*39*). Determining whether this is also the case with the ARC genes, as well as their close relatives in the mammalian genome (*14*) will require further characterization of existing retrotransposon elements

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

399 utilizing structural methods not reliant on the comparative similarities in related nucleic

400 acid sequences that have disappeared with the passage of time.

401

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

MATERIALS AND METHODS

Protein expression and purification

dARC1 residues S39-N205 were determined to represent the CA domain according to MSA and secondary structural analysis performed in ClustalX (*40*) and Pspired (*41*). An *E. coli* codon-optimised cDNA for *Drosophila melanogaster* dARC1 (Uniprot Q7K1U0) was synthesised (Geneart), the relevant sequence PCR-amplified, and subcloned into a pET22b plasmid (Novagen). The resulting construct comprises residues 39-205 of dARC1, with an N-terminal Met, and a C-terminal PLEHHHHHH motif. Proteins were expressed in *E. coli* strain BL21 (DE3) grown in LB broth, by induction of log-phase cultures with 1 mM IPTG, incubated overnight at 20°C. Cells were pelleted and resuspended in 50 mM Tris-HCl, 150 mM NaCl, 10 mM Imidazole, 5mM $MgCl_2$, 1mM DTT, pH 8.0, supplemented with 1 mg/mL lysozyme (Sigma-Aldrich), 10 µg/mL DNase I (Sigma-Aldrich) and 1 Protease Inhibitor cocktail tablet (EDTA free, Pierce) per 40 mL of buffer. Cells were lysed using an EmulsiFlex-C5 homogeniser (Avestin) and dARC1 CA captured from clarified lysate using immobilised metal ion affinity on a 5 mL Ni^{2+}-NTA superflow column (Qiagen). Bound dARC1 CA was eluted in non-reducing buffer (50 mM Tris-HCl, 150 mM NaCl, 300 mM Imidazole) and Carboxypeptidase A (CPA, Sigma C9268) was added at a ratio of ~ 100 mg dARC1 per mg CPA. The resulting mixture was incubated overnight at 4 ˚C to allow digestion of the C-terminal His-tag. The CPA was inactivated via the addition of TCEP to 2mM. dARC1 CA was further purified by size exclusion chromatography using a Superdex 75 (26/60) (GE healthcare) column, equilibrated in 20 mM Tris-HCl, 150 mM NaCl, 1mM TCEP pH 8.0. Purified protein eluted in a single peak. Seleno-methionine derivative protein was produced using an identical procedure, although *E. coli* B834 (DE3) cells, grown in seleno-methionine Medium (Molecular Dimensions, Newmarket, UK), were used to express the protein. Electrospray-ionisation mass spectrometry was used to confirm the identity of dARC1, and where applicable, seleno-methionine incorporation. It also

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

confirmed that the N-terminal Met had been processed, and that the His-tag had been completely digested, leaving the motif “PLE” at the C-terminus. Protein was concentrated by centrifugal ultrafiltration (Vivaspin (MWCO 10 kDa), then snap frozen and stored at -80°C. Concentrations were determined by UV absorbance spectroscopy using an extinction coefficient at 280 nm.

Protein crystallisation and structure determination

dARC1 CA was crystallised using sitting drop vapour diffusion at 18 ˚C, using Swissci MRC 2-drop trays (Molecular Dimensions) with drops set using a Dragonfly robot with humidity chamber (TTP Labtech). Native protein was initially concentrated to 20 mg/mL. Typically drops were 200-300 nL, made by mixing protein:mother liquor in a 3:1 or 1:1 ratio, with a 75 µL reservoir. Initial crystal hits were obtained using the Structure Screen 1&2 (Molecular Dimensions) in a condition containing 4.3 M NaCl, 0.1 M HEPES pH 7.5. Two crystal forms could be observed in such conditions. Thin rods which turned out to be oP or as hexagonal disks or trapezoidal prisms which were hP. Datasets were collected for these native crystals, but they could not be solved by molecular replacement methods. SeMet dARC1 CA was crystallised in conditions that optimised protein concentration, NaCl concentration, and pH. The best crystals grew in 300-400 nL drops set with protein at 12.5-16 mg/mL, with mother liquor NaCl ranging between 2.8-3.3 M. Rods were ~400x30x30 µm and hexagons/trapezoids were ~130 µm across and up to 30 µm thick. Crystals were harvested using MiteGen lithographic loops. The best cryoprotection was achieved using sodium malonate, mixed into mother liquor to a concentration of 1.6 M. This was added directly to the drop, or crystals were bathed in this solution before flash freezing in liquid nitrogen.

Data collection and structure determination

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data were collected at the tuneable SLS beamline, PXIII. For the orthorhombic crystal form, a peak dataset was collected to 2.06 Å (see Table S1). The data were processed by the SLS GoPy pipeline in $P2_12_12_1$, using XDS (*42*) and showed significant anomalous signal to 2.82 Å. The resultant dataset was solved using SAD methods with Phenix (*43*) and despite a relatively low FOM the experimental map was readily interpretable and it was possible to almost completely autobuild an initial structure with BUCCANEER (*44*). A higher resolution (1.55 Å) dataset was collected at a non-anomalous, low energy remote wavelength (Table S1). This dataset was processed using the Xia2 (*45*) pipeline, using DIALS (*46*) for indexing and integration, and AIMLESS (*47*) or scaling and merging. This dataset was initially used for refinement to 1.7 Å and manual model building in COOT (*48*). It was evident that the data were anisotropic, and that they might benefit from anisotropic correction. Diffraction images were reprocessed using the autoPROC pipeline (*49*), using XDS, POINTLESS (*50*), AIMLESS and STARANISO (http://staraniso.globalphasing.org/cgi-bin/staraniso.cgi). This dataset was used for further refinement of the model and indeed there was an improvement in map quality, and in agreement between model and data. For the hexagonal crystal form, a highly redundant peak dataset was collected to 2.14 Å. This was processed using the Xia2 pipeline, using DIALS for indexing and integration, and AIMLESS for scaling and merging, showing significant anomalous signal to 2.59 Å, in $P6_122$. This dataset was solved using SAD methods in Phenix. Again, the experimental map was readily interpretable and it was possible to almost completely autobuild an initial structure with BUCCANEER. Refinement and model building were carried out in Phenix and COOT respectively. Anomalous signal was very strong in this dataset and so Freidel pairs were treated separately during refinement. Molprobity (*51*) and PDB_REDO (*52*) were used to monitor and assess model geometry. Details of data collection, phasing and structure refinement statistics are presented in **Table S1**.

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

480

481 *SEC-MALLS*

482 Size exclusion chromatography coupled multi-angle laser light scattering (SEC-
483 MALLS) was used to determine the molar mass of dARC CA. Samples ranging from 25
484 µM to 400 µM were applied in a volume of 100 µL to a Superdex INCREASE 200 10/300
485 GL column equilibrated in 20 mM Tris-HCl, 150 mM NaCl and 0.5 mM TCEP, 3 mM NaN_3,
486 pH 8.0, at a flow rate of 1.0 mL/min. The scattered light intensity and the protein
487 concentration of the column eluate were recorded using a DAWN-HELEOS laser
488 photometer and OPTILAB-rEX differential refractometer respectively. The weight-
489 averaged molecular mass of material contained in chromatographic peaks was determined
490 from the combined data from both detectors using the ASTRA software version 6.0.3
491 (Wyatt Technology Corp., Santa Barbara, CA, USA).

492

493 *Analytical Ultracentrifugation*

494 Sedimentation velocity experiments were performed in a Beckman Optima Xl-I
495 analytical ultracentrifuge using conventional aluminium double sector centrepieces and
496 sapphire windows. Solvent density and the protein partial specific volumes were
497 determined as described (*53*). Prior to centrifugation, dARC1 CA samples were prepared
498 by exhaustive dialysis against the buffer blank solution, 20 mM Tris-HCl pH 8, 150 mM
499 NaCl and 0.5 mM TCEP (Tris Buffer). Samples (420 µL) and buffer blanks (426 µL) were
500 loaded into the cells and centrifugation was performed at 50,000 rpm and 293 K in an
501 An50-Ti rotor. Interference data were acquired at time intervals of 180 s at varying sample
502 concentration (25, 50 and 100 µM). Data recorded from moving boundaries was analysed
503 in terms of the size distribution functions C(S) using the program SEDFIT (*54-56*).

504 Sedimentation equilibrium experiments were performed in a Beckman Optima XL-I
505 analytical ultracentrifuge using aluminium double sector centrepieces in an An-50 Ti rotor.

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Prior to centrifugation, samples were dialyzed exhaustively against the buffer blank (Tris Buffer). Samples (150 µL) and buffer blanks (160 µL) were loaded into the cells an after centrifugation for 30 hours, interference data was collected at 2 hourly intervals until no further change in the profiles was observed. The rotor speed was then increased and the procedure repeated. Data were collected on samples of different concentrations of dARC1 CA (25, 50 and 70 µM) at three speeds and the program SEDPHAT (*57, 58*) was used to determine weight-averaged molecular masses by nonlinear fitting of individual multi-speed equilibrium profiles to a single-species ideal solution model. Inspection of these data revealed that the molecular mass of dARC1 CA showed no significant concentration dependency and so global fitting incorporating the data from multiple speeds and multiple sample concentrations was applied to extract a final weight-averaged molecular mass.

Structure analysis and alignments

Molecular interfaces were analysed using the EBI protein structure interface analysis service PDBePISA (https://www.ebi.ac.uk/pdbe/pisa). Electrostatic surface potential and the surface hydrophobicity/hydrophilicity distribution of the dARC1 CA dimer were calculated with APBS (*59*) and using the pymol script (https://pymolwiki.org/index.php/Color_h) respectively. The DALI (*60*) comparison server (http://ekhidna2.biocenter.helsinki.fi/dali) was used to search for and align structural homologues from the PDB.

Sequence alignment

Amino acid alignments were produced with MAFFT v7.271 (*61*), within tcoffee v11.00.8cbe486 (*62*), weighting alignments using 3-state secondary-structure predictions produced with RaptorX Property v1.02 (*63*). Alignment images were produced with ESPript (*64*).

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

532

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. J. L. Dynes, O. Steward, Dynamics of bidirectional transport of Arc mRNA in neuronal dendrites. *J Comp Neurol* **500**, 433-447 (2007); published online EpubJan 20 (10.1002/cne.21189).
2. N. Plath, O. Ohana, B. Dammermann, M. L. Errington, D. Schmitz, C. Gross, X. Mao, A. Engelsberg, C. Mahlke, H. Welzl, U. Kobalz, A. Stawrakakis, E. Fernandez, R. Waltereit, A. Bick-Sander, E. Therstappen, S. F. Cooke, V. Blanquet, W. Wurst, B. Salmen, M. R. Bosl, H. P. Lipp, S. G. Grant, T. V. Bliss, D. P. Wolfer, D. Kuhl, Arc/Arg3.1 is essential for the consolidation of synaptic plasticity and memories. *Neuron* **52**, 437-444 (2006); published online EpubNov 9 (10.1016/j.neuron.2006.08.024).
3. J. D. Shepherd, M. F. Bear, New views of Arc, a master regulator of synaptic plasticity. *Nat Neurosci* **14**, 279-284 (2011); published online EpubMar (10.1038/nn.2708).
4. S. Chowdhury, J. D. Shepherd, H. Okuno, G. Lyford, R. S. Petralia, N. Plath, D. Kuhl, R. L. Huganir, P. F. Worley, Arc/Arg3.1 interacts with the endocytic machinery to regulate AMPA receptor trafficking. *Neuron* **52**, 445-459 (2006); published online EpubNov 9 (10.1016/j.neuron.2006.08.033).
5. J. D. Shepherd, G. Rumbaugh, J. Wu, S. Chowdhury, N. Plath, D. Kuhl, R. L. Huganir, P. F. Worley, Arc/Arg3.1 mediates homeostatic synaptic scaling of AMPA receptors. *Neuron* **52**, 475-484 (2006); published online EpubNov 9 (10.1016/j.neuron.2006.08.034).
6. J. Wu, R. S. Petralia, H. Kurushima, H. Patel, M. Y. Jung, L. Volk, S. Chowdhury, J. D. Shepherd, M. Dehoff, Y. Li, D. Kuhl, R. L. Huganir, D. L. Price, R. Scannevin, J. C. Troncoso, P. C. Wong, P. F. Worley, Arc/Arg3.1 regulates an endosomal pathway essential for activity-dependent beta-amyloid generation. *Cell* **147**, 615-628 (2011); published online EpubOct 28 (10.1016/j.cell.2011.09.036).
7. R. Bi, L. L. Kong, M. Xu, G. D. Li, D. F. Zhang, T. Li, Y. Fang, C. Zhang, B. Zhang, Y. G. Yao, The Arc Gene Confers Genetic Susceptibility to Alzheimer's Disease in Han Chinese. *Mol Neurobiol* **55**, 1217-1226 (2018); published online EpubFeb (10.1007/s12035-017-0397-6).
8. S. Park, J. M. Park, S. Kim, J. A. Kim, J. D. Shepherd, C. L. Smith-Hicks, S. Chowdhury, W. Kaufmann, D. Kuhl, A. G. Ryazanov, R. L. Huganir, D. J. Linden, P. F. Worley, Elongation factor 2 and fragile X mental retardation protein control the dynamic translation of Arc/Arg3.1 essential for mGluR-LTD. *Neuron* **59**, 70-83 (2008); published online EpubJul 10 (10.1016/j.neuron.2008.05.023).
9. M. Fromer, A. J. Pocklington, D. H. Kavanagh, H. J. Williams, S. Dwyer, P. Gormley, L. Georgieva, E. Rees, P. Palta, D. M. Ruderfer, N. Carrera, I. Humphreys, J. S. Johnson, P. Roussos, D. D. Barker, E. Banks, V. Milanova, S. G. Grant, E. Hannon, S. A. Rose, K. Chambert, M. Mahajan, E. M. Scolnick, J. L. Moran, G. Kirov, A. Palotie, S. A. McCarroll, P. Holmans, P. Sklar, M. J. Owen, S. M. Purcell, M. C. O'Donovan, De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014); published online EpubFeb 13 (10.1038/nature12929).
10. S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'Dushlaine, K. Chambert, S. E. Bergen, A. Kahler, L. Duncan, E. Stahl, G. Genovese, E. Fernandez, M. O. Collins, N. H. Komiyama, J. S. Choudhary, P. K. Magnusson, E. Banks, K. Shakir, K. Garimella, T. Fennell, M. DePristo, S. G. Grant, S. J. Haggarty, S. Gabriel, E. M. Scolnick, E. S. Lander, C. M. Hultman, P. F. Sullivan, S. A. McCarroll, P. Sklar, A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-190 (2014); published online EpubFeb 13 (10.1038/nature12975).
11. F. Manago, M. Mereu, S. Mastwal, R. Mastrogiacomo, D. Scheggia, M. Emanuele, M. A. De Luca, D. R. Weinberger, K. H. Wang, F. Papaleo, Genetic Disruption of Arc/Arg3.1 in Mice Causes Alterations in Dopamine and Neurobehavioral Phenotypes Related to Schizophrenia. *Cell Rep* **16**, 2116-2128 (2016); published online EpubAug 23 (10.1016/j.celrep.2016.07.044).

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

12. J. Ashley, B. Cordy, D. Lucia, L. G. Fradkin, V. Budnik, T. Thomson, Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell* **172**, 262-274 e211 (2018); published online EpubJan 11 (10.1016/j.cell.2017.12.022).
13. M. D. Mattaliano, E. S. Montana, K. M. Parisky, J. T. Littleton, L. C. Griffith, The Drosophila ARC homolog regulates behavioral responses to starvation. *Molecular and cellular neurosciences* **36**, 211-221 (2007); published online EpubOct (10.1016/j.mcn.2007.06.008).
14. M. Campillos, T. Doerks, P. K. Shah, P. Bork, Computational characterization of multiple Gag-like human proteins. *Trends in genetics : TIG* **22**, 585-589 (2006); published online EpubNov (10.1016/j.tig.2006.09.006).
15. E. D. Pastuzyn, C. E. Day, R. B. Kearns, M. Kyrke-Smith, A. V. Taibi, J. McCormick, N. Yoder, D. M. Belnap, S. Erlendsson, D. R. Morado, J. A. G. Briggs, C. Feschotte, J. D. Shepherd, The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell* **172**, 275-288 e218 (2018); published online EpubJan 11 (10.1016/j.cell.2017.12.024).
16. W. Zhang, J. Wu, M. D. Ward, S. Yang, Y. A. Chuang, M. Xiao, R. Li, D. J. Leahy, P. F. Worley, Structural basis of arc binding to synaptic proteins: implications for cognitive disease. *Neuron* **86**, 490-500 (2015); published online EpubApr 22 (10.1016/j.neuron.2015.03.030).
17. W. R. Taylor, J. P. Stoye, I. A. Taylor, A comparative analysis of the foamy and ortho virus capsid structures reveals an ancient domain duplication. *BMC structural biology* **17**, 3 (2017); published online EpubApr 4 (10.1186/s12900-017-0073-0).
18. M. J. Curcio, S. Lutz, P. Lesage, The Ty1 LTR-retrotransposon of budding yeast, Saccharomyces cerevisiae. *Microbiol Spectr* **3**, 1-35 (2015); published online EpubApr 1 (10.1128/microbiolspec.MDNA3-0053-2014).
19. E. O. Freed, HIV-1 assembly, release and maturation. *Nature reviews. Microbiology* **13**, 484-496 (2015); published online EpubAug (10.1038/nrmicro3490).
20. L. D. Nielsen, C. P. Pedersen, S. Erlendsson, K. Teilum, The Capsid Domain of Arc Changes Its Oligomerization Propensity through Direct Interaction with the NMDA Receptor. *Structure*, (2019); published online EpubApr 16 (10.1016/j.str.2019.04.001).
21. A. Ortega, D. Amoros, J. Garcia de la Torre, Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. *Biophysical journal* **101**, 892-898 (2011); published online EpubAug 17 (10.1016/j.bpj.2011.06.046).
22. L. Holm, C. Sander, Dali/FSSP classification of three-dimensional protein folds. *Nucleic acids research* **25**, 231-234 (1997); published online EpubJan 1 (10.1093/nar/25.1.231).
23. N. J. Ball, G. Nicastro, M. Dutta, D. J. Pollard, D. C. Goldstone, M. Sanz-Ramos, A. Ramos, E. Mullers, K. Stirnnagel, N. Stanke, D. Lindemann, J. P. Stoye, W. R. Taylor, P. B. Rosenthal, I. A. Taylor, Structure of a Spumaretrovirus Gag Central Domain Reveals an Ancient Retroviral Capsid. *PLoS pathogens* **12**, e1005981 (2016); published online EpubNov (10.1371/journal.ppat.1005981).
24. S. O. Dodonova, S. Prinz, V. Bilanchone, S. Sandmeyer, J. A. G. Briggs, Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 10048-10057 (2019); published online EpubMay 14 (10.1073/pnas.1900931116).
25. T. R. Gamble, S. Yoo, F. F. Vajdos, U. K. von Schwedler, D. K. Worthylake, H. Wang, J. P. McCutcheon, W. I. Sundquist, C. P. Hill, Structure of the carboxyl-terminal dimerization domain of the HIV-1 capsid protein. *Science* **278**, 849-853 (1997); published online EpubOct 31 (10.1126/science.278.5339.849).
26. S. Yoo, D. G. Myszka, C. Yeh, M. McMurray, C. P. Hill, W. I. Sundquist, Molecular recognition in the HIV-1 capsid/cyclophilin A complex. *Journal of molecular biology* **269**, 780-795 (1997); published online EpubJun 27 (

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

27. S. Khorasanizadeh, R. Campos-Olivas, M. F. Summers, Solution structure of the capsid protein from the human T-cell leukemia virus type-I. *Journal of molecular biology* **291**, 491-505 (1999); published online EpubAug 13 (10.1006/jmbi.1999.2986).
28. G. B. Mortuza, D. C. Goldstone, C. Pashley, L. F. Haire, M. Palmarini, W. R. Taylor, J. P. Stoye, I. A. Taylor, Structure of the capsid amino-terminal domain from the betaretrovirus, Jaagsiekte sheep retrovirus. *Journal of molecular biology* **386**, 1179-1192 (2009); published online EpubMar 6 (10.1016/j.jmb.2008.10.066).
29. R. L. Kingston, T. Fitzon-Ostendorp, E. Z. Eisenmesser, G. W. Schatz, V. M. Vogt, C. B. Post, M. G. Rossmann, Structure and self-association of the Rous sarcoma virus capsid protein. *Structure* **8**, 617-628 (2000); published online EpubJun 15 (
30. B. K. Ganser-Pornillos, A. Cheng, M. Yeager, Structure of full-length HIV-1 CA: a model for the mature capsid lattice. *Cell* **131**, 70-79 (2007); published online EpubOct 5 (10.1016/j.cell.2007.08.018).
31. S. Li, C. P. Hill, W. I. Sundquist, J. T. Finch, Image reconstructions of helical assemblies of the HIV-1 CA protein. *Nature* **407**, 409-413 (2000); published online EpubSep 21 (10.1038/35030177).
32. J. A. Briggs, T. Wilk, R. Welker, H. G. Krausslich, S. D. Fuller, Structural organization of authentic, mature HIV-1 virions and cores. *The EMBO journal* **22**, 1707-1715 (2003); published online EpubApr 1 (10.1093/emboj/cdg143).
33. G. B. Mortuza, L. F. Haire, A. Stevens, S. J. Smerdon, J. P. Stoye, I. A. Taylor, High-resolution structure of a retroviral capsid hexameric amino-terminal domain. *Nature* **431**, 481-485 (2004); published online EpubSep 23 (10.1038/nature02915).
34. J. Lanman, T. T. Lam, M. R. Emmett, A. G. Marshall, M. Sakalian, P. E. Prevelige, Jr., Key interactions in HIV-1 maturation identified by hydrogen-deuterium exchange. *Nature structural & molecular biology* **11**, 676-677 (2004); published online EpubJul (10.1038/nsmb790).
35. J. Lanman, T. T. Lam, S. Barnes, M. Sakalian, M. R. Emmett, A. G. Marshall, P. E. Prevelige, Jr., Identification of novel interactions in HIV-1 capsid protein assembly by high-resolution mass spectrometry. *Journal of molecular biology* **325**, 759-772 (2003); published online EpubJan 24 (
36. S. Mi, X. Lee, X. Li, G. M. Veldman, H. Finnerty, L. Racie, E. LaVallie, X. Y. Tang, P. Edouard, S. Howes, J. C. Keith, Jr., J. M. McCoy, Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785-789 (2000); published online EpubFeb 17 (10.1038/35001608).
37. M. Funk, G. Cornelis, C. Vernochet, O. Heidmann, A. Dupressoir, A. Conley, S. Glickman, T. Heidmann, Capture of a Hyena-Specific Retroviral Envelope Gene with Placental Expression Associated in Evolution with the Unique Emergence among Carnivorans of Hemochorial Placentation in Hyaenidae. *Journal of virology* **93**, (2019); published online EpubFeb 15 (10.1128/JVI.01811-18).
38. G. Cornelis, O. Heidmann, S. A. Degrelle, C. Vernochet, C. Lavialle, C. Letzelter, S. Bernard-Stoecklin, A. Hassanin, B. Mulot, M. Guillomot, I. Hue, T. Heidmann, A. Dupressoir, Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E828-837 (2013); published online EpubFeb 26 (10.1073/pnas.1215787110).
39. P. Wooding, G. Burton, *Comparative Placentation: Structures, Functions and Evolution*. (Springer Berlin Heidelberg, 2008).
40. M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, D. G. Higgins, Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007); published online EpubNov 1 (10.1093/bioinformatics/btm404).

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

41. D. T. Jones, Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* **292**, 195-202 (1999); published online EpubSep 17 (10.1006/jmbi.1999.3091).
42. W. Kabsch, Xds. *Acta crystallographica. Section D, Biological crystallography* **66**, 125-132 (2010); published online EpubFeb (10.1107/S0907444909047337).
43. P. H. Zwart, P. V. Afonine, R. W. Grosse-Kunstleve, L. W. Hung, T. R. Ioerger, A. J. McCoy, E. McKee, N. W. Moriarty, R. J. Read, J. C. Sacchettini, N. K. Sauter, L. C. Storoni, T. C. Terwilliger, P. D. Adams, Automated structure solution with the PHENIX suite. *Methods in molecular biology* **426**, 419-435 (2008)10.1007/978-1-60327-058-8_28).
44. K. Cowtan, The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta crystallographica. Section D, Biological crystallography* **62**, 1002-1011 (2006); published online EpubSep (10.1107/S0907444906022116).
45. G. Winter, xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography* **43**, 186-190 (2009)10.1107/s0021889809045701).
46. G. Winter, D. G. Waterman, J. M. Parkhurst, A. S. Brewster, R. J. Gildea, M. Gerstel, L. Fuentes-Montero, M. Vollmar, T. Michels-Clark, I. D. Young, N. K. Sauter, G. Evans, DIALS: implementation and evaluation of a new integration package. *Acta Crystallogr D Struct Biol* **74**, 85-97 (2018); published online EpubFeb 1 (10.1107/S2059798317017235).
47. P. R. Evans, G. N. Murshudov, How good are my data and what is the resolution? *Acta crystallographica. Section D, Biological crystallography* **69**, 1204-1214 (2013); published online EpubJul (10.1107/S0907444913000061).
48. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta crystallographica. Section D, Biological crystallography* **60**, 2126-2132 (2004); published online EpubDec (10.1107/S0907444904019158).
49. C. Vonrhein, C. Flensburg, P. Keller, A. Sharff, O. Smart, W. Paciorek, T. Womack, G. Bricogne, Data processing and analysis with the autoPROC toolbox. *Acta crystallographica. Section D, Biological crystallography* **67**, 293-302 (2011); published online EpubApr (10.1107/S0907444911007773).
50. P. Evans, Scaling and assessment of data quality. *Acta crystallographica. Section D, Biological crystallography* **62**, 72-82 (2006); published online EpubJan (10.1107/S0907444905036693).
51. V. B. Chen, J. R. Wedell, R. K. Wenger, E. L. Ulrich, J. L. Markley, MolProbity for the masses-of data. *Journal of biomolecular NMR* **63**, 77-83 (2015); published online EpubSep (10.1007/s10858-015-9969-9).
52. R. P. Joosten, J. Salzemann, V. Bloch, H. Stockinger, A. C. Berglund, C. Blanchet, E. Bongcam-Rudloff, C. Combet, A. L. Da Costa, G. Deleage, M. Diarena, R. Fabbretti, G. Fettahi, V. Flegel, A. Gisel, V. Kasam, T. Kervinen, E. Korpelainen, K. Mattila, M. Pagni, M. Reichstadt, V. Breton, I. J. Tickle, G. Vriend, PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *Journal of applied crystallography* **42**, 376-384 (2009); published online EpubJun 1 (10.1107/S0021889809008784).
53. T. M. Laue, B. D. Shah, T. M. Ridgeway, S. L. Pelletier, in *Analytical Ultracentrifugation in Biochemistry and Polymer Science,* S. E. Harding, A. J. Rowe, J. C. Horton, Eds. (The Royal Society of Chemistry, Cambridge United Kingdom, 1992), pp. 90-125.
54. P. Schuck, M. A. Perugini, N. R. Gonzales, G. J. Howlett, D. Schubert, Size-distribution analysis of proteins by analytical ultracentrifugation: strategies and application to model systems. *Biophysical journal* **82**, 1096-1111 (2002); published online EpubFeb (10.1016/S0006-3495(02)75469-6).
55. P. Schuck, Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophysical journal* **78**, 1606-1619 (2000); published online EpubMar (10.1016/S0006-3495(00)76713-0).

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

56. P. H. Brown, P. Schuck, Macromolecular size-and-shape distributions by sedimentation velocity analytical ultracentrifugation. *Biophysical journal* **90**, 4651-4661 (2006); published online EpubJun 15 (10.1529/biophysj.106.081372).
57. J. Vistica, J. Dam, A. Balbo, E. Yikilmaz, R. A. Mariuzza, T. A. Rouault, P. Schuck, Sedimentation equilibrium analysis of protein interactions with global implicit mass conservation constraints and systematic noise decomposition. *Analytical biochemistry* **326**, 234-256 (2004); published online EpubMar 15 (10.1016/j.ab.2003.12.014).
58. P. Schuck, On the analysis of protein self-association by sedimentation velocity analytical ultracentrifugation. *Analytical biochemistry* **320**, 104-124 (2003); published online EpubSep 1 (10.1016/s0003-2697(03)00289-6).
59. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10037-10041 (2001); published online EpubAug 28 (10.1073/pnas.181342398).
60. L. Holm, C. Sander, Dali: a network tool for protein structure comparison. *Trends in biochemical sciences* **20**, 478-480 (1995); published online EpubNov (
61. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013); published online EpubApr (10.1093/molbev/mst010).
62. C. Notredame, D. G. Higgins, J. Heringa, T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology* **302**, 205-217 (2000); published online EpubSep 8 (10.1006/jmbi.2000.4042).
63. S. Wang, J. Peng, J. Ma, J. Xu, Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep* **6**, 18962 (2016); published online EpubJan 11 (10.1038/srep18962).
64. X. Robert, P. Gouet, Deciphering key features in protein structures with the new ENDscript server. *Nucleic acids research* **42**, W320-324 (2014); published online EpubJul (10.1093/nar/gku316).

ACKNOWLEDGEMENTS

General: We thank the Swiss Light Source for beamtime, and the staff of beamline PXIII.

Funding: This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001162, FC001178), the UK Medical Research Council (FC001162, FC001178), and the Wellcome Trust (FC001162, FC001178); and by the Wellcome Trust (108014/Z/15/Z and 108012/Z/15/Z).

Author contributions: M.A.C., S.C.L and I. A. T. performed experiments. M.A.C., S.C.L, G.R.Y., J.P.S. and I.A.T. contributed to experimental design, data analysis and manuscript writing.

Competing Interests: The authors declare they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

776 **Data and materials availability**: The coordinates and structure factors for dARC1 CA
777 (S39-N205) have been deposited in the Protein Data Bank under accession numbers
778 6S7X and 6S7Y.
779

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

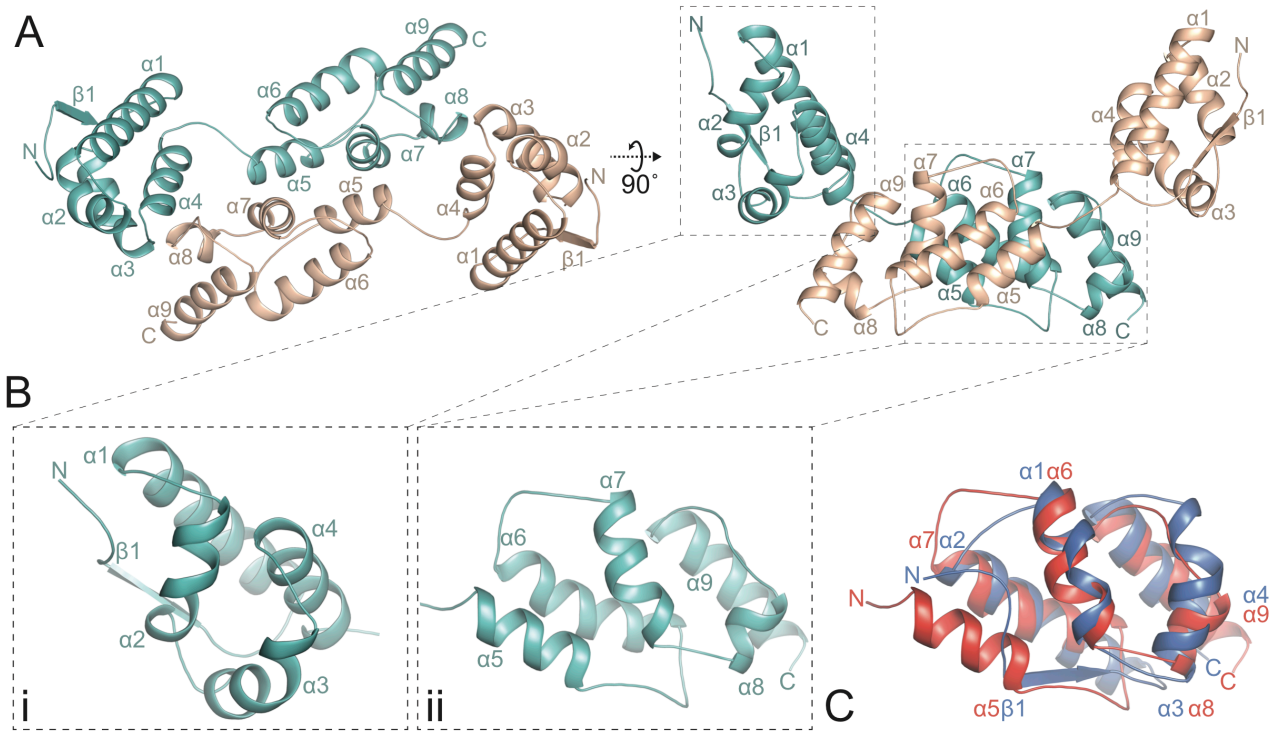


Fig. 1. Crystal structure of the dARC1 CA domain. (**A**) Cartoon representation of the dARC1 CA dimer, the N-terminal extended β-strand and α-helices are numbered sequentially from the N- to the C-terminus. Monomer-A is coloured cyan and Monomer-B in wheat. The right-hand panel is a view at 90° relative to the left-hand panel. (**B**) Close-up cartoon representations of dARC1 CA-NtD (left) and dARC CA-CtD (right) showing the helical topology of each domain. (**C**) 3D Cα structural alignment of dARC1 CA-NtD (blue cartoon) with dARC1 CA-CtD (red cartoon) with secondary structure elements labelled.

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

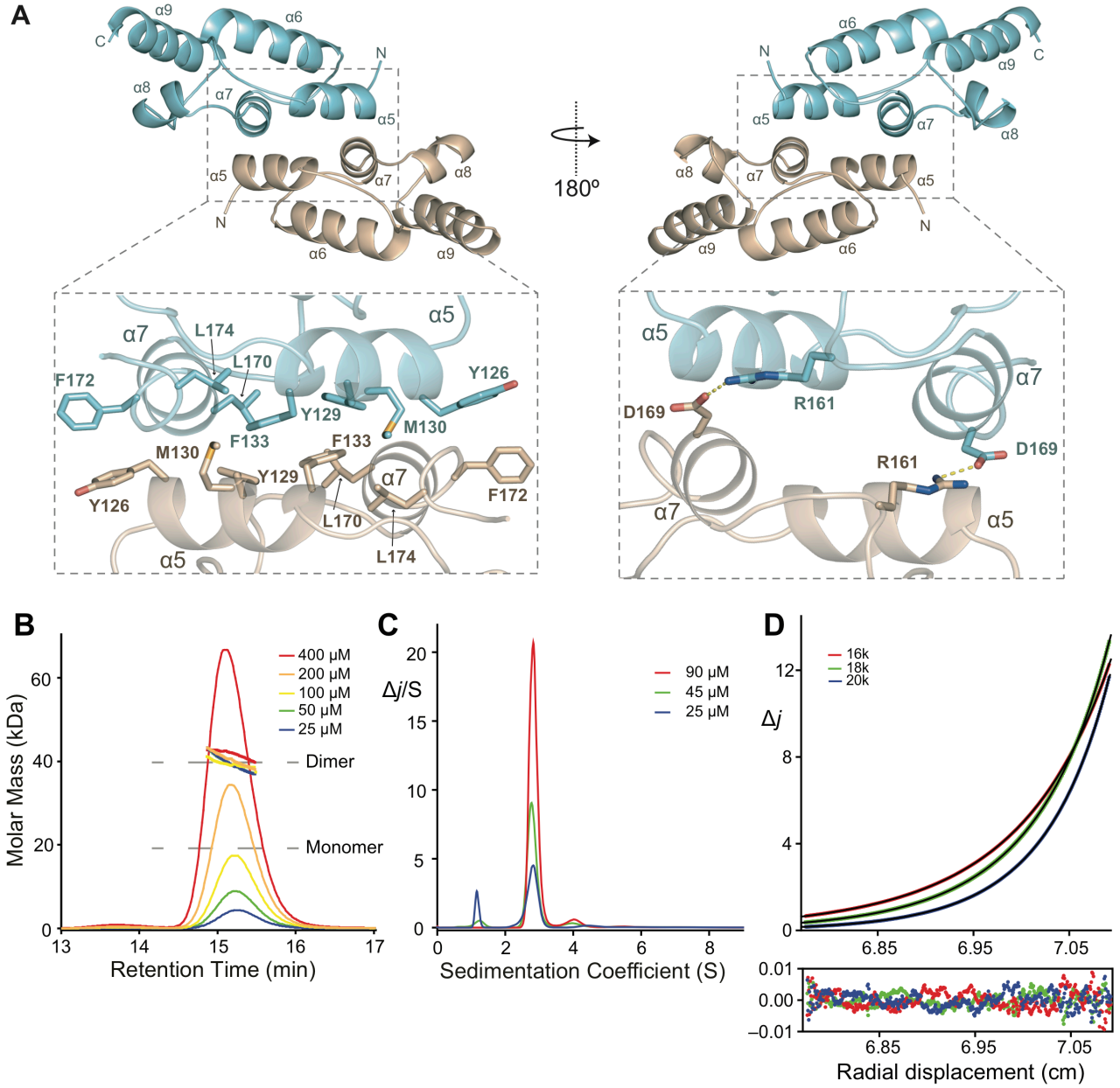


Fig. 2. The dARC1 CA dimer interface and solution conformation. (**A**) A cartoon representation of dARC1 CA-CtD dimer, α-helices are numbered sequentially from the N- to the C-terminus. Monomer-A is coloured cyan and Monomer-B in wheat. The right-hand panel is a view at 180° relative to the left-hand panel. **Inset,** close up of molecular details of interactions at the dARC1 dimer interface. Residues that make interactions are shown in stick representation coloured by atom type. Salt-bridge interactions between R161 and D169 are shown as dashed lines. (**B**) SEC-MALLS analysis of dARC1 CA. The sample loading concentrations were 400 μM (8 mg/mL) (red), 200 μM (4 mg/mL) (orange), 100 μM

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

798 (2 mg/mL) (yellow), 50 μM (1 mg/mL) (green) and 25 μM (0.5 mg/mL) (blue). The
799 differential refractive index (dRI) is plotted against column retention time and the molar
800 mass, determined at 1-second intervals throughout the elution of each peak, is plotted as
801 points. The dARC1 CA monomer and dimer molecular mass is indicated with the grey
802 dashed lines. (**C**) C(S) distributions derived from sedimentation velocity data recorded
803 from dARC1 CA at 25 μM (blue), 50 μM (green) and 100 μM (red). The curves represent
804 the distribution of the sedimentation coefficients that best fit the sedimentation data (f/f_0 =
805 1.41). (**D**) Multi-speed sedimentation equilibrium profile determined from interference data
806 collected on dARC1 CA at 70 μM. Data was recorded at the three speeds indicated. The
807 solid lines represent the global best fit to the data using a single species model (Mw =
808 38.9±1 kD). The lower panel shows the residuals to the fit.
809

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

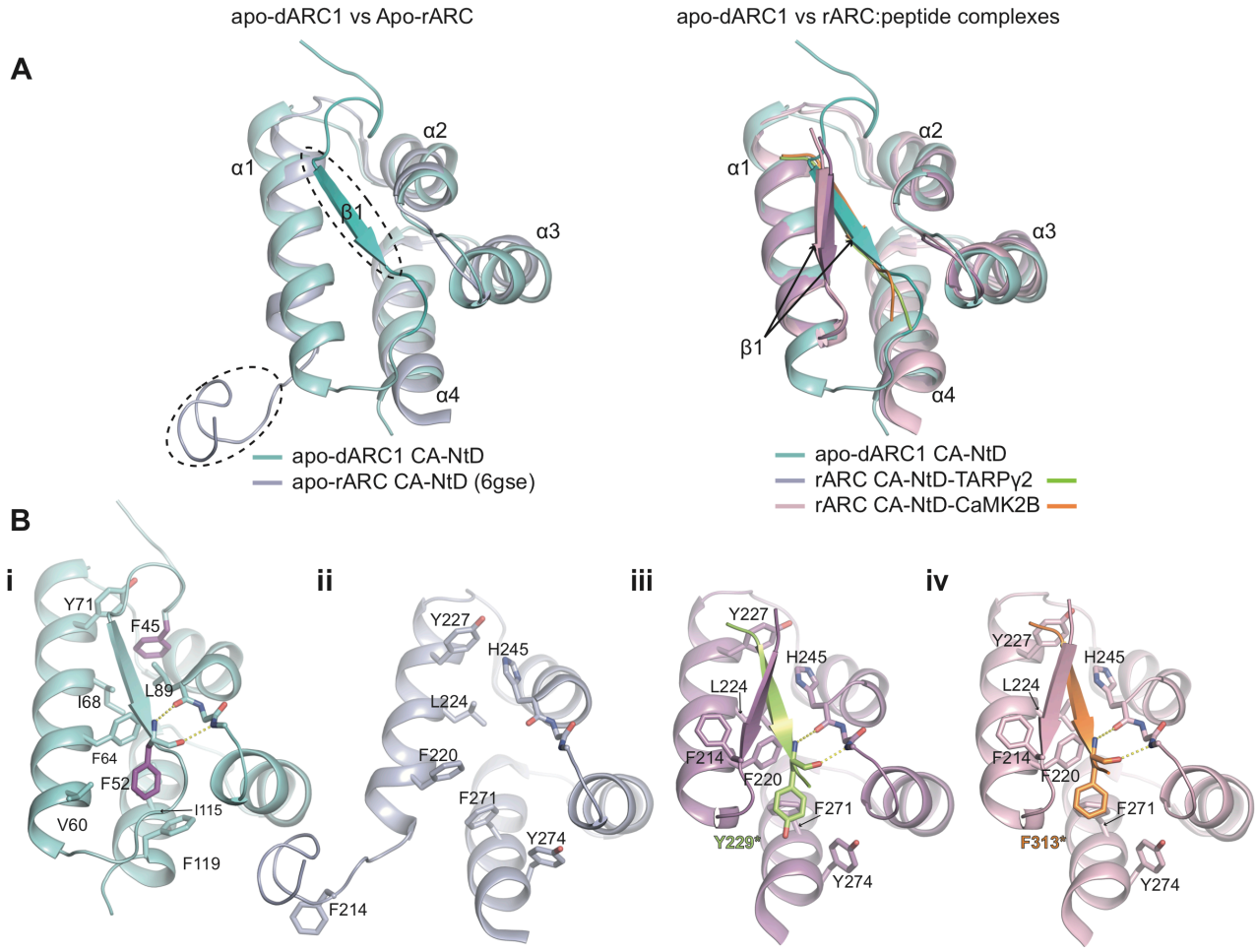
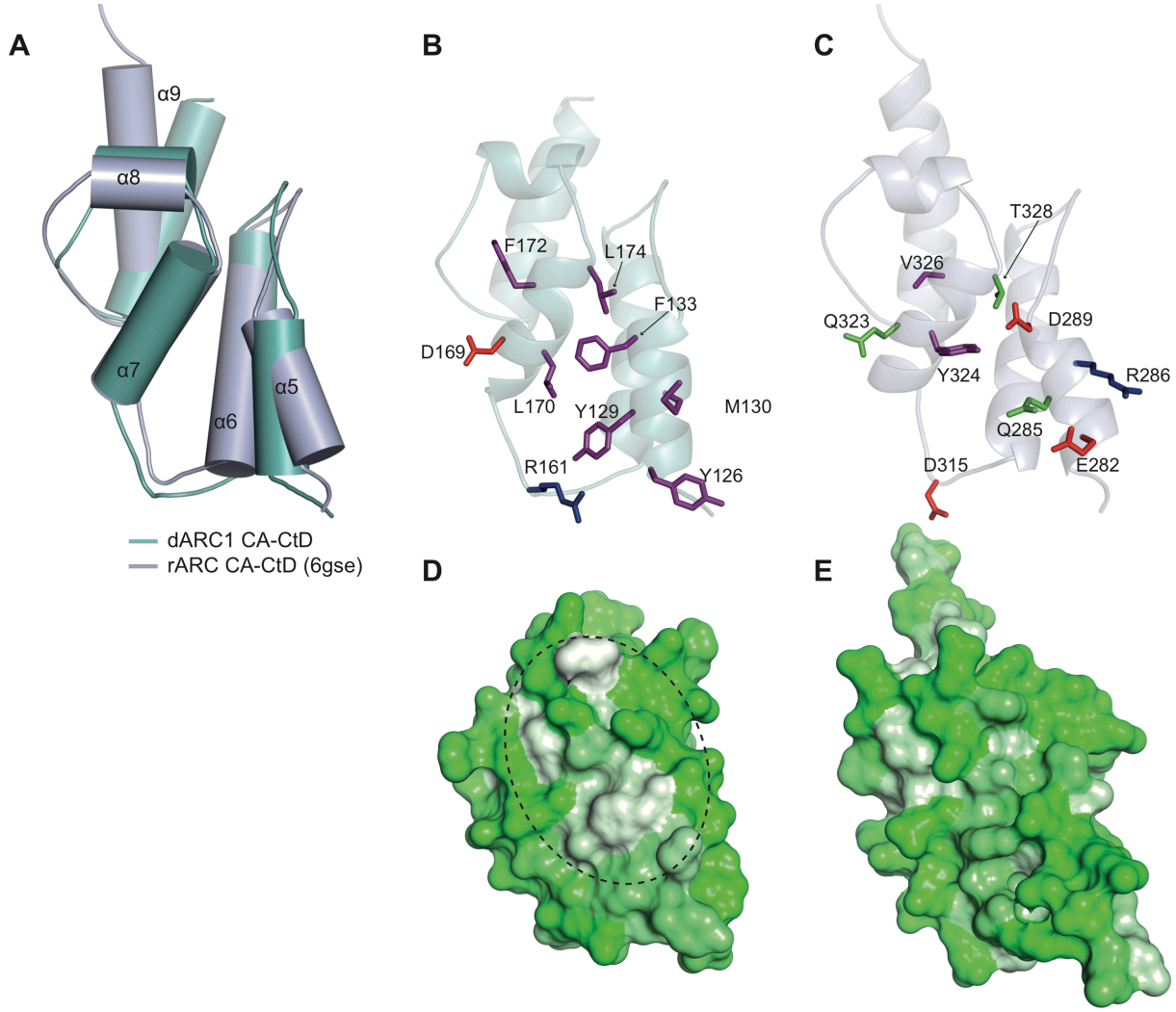


Fig. 3. Comparison of dARC1 and rARC CA-NTD structures. (**A**) Left, 3D structural alignment of dARC1 CA-NtD (teal cartoon) and the apo-rARC CA-NtD (PDB: 6GSE, lilac cartoon). Secondary structure elements are labelled. Circled are the ordered, N-terminal β-strand of dARC1, and the disordered N-terminal strand of apo-rARC. Right, 3D structural alignment of dARC1 CA-NtD and the peptide-complex structures of rARC CA-NtDs (PDB: 4X3H and 4X3I). The protein backbones are shown in cartoon representation coloured according to the legend. Secondary structure elements are labelled. The arrow indicates the different positioning of the extended N-terminal β-strand between the dARC1 and rARC structures. (**B, i to iv**) Individual views of the structures presented in **A**. **i**) apo-dARC1, **ii**) apo-rARc, **iii**) rARC-TARPγ2, **iv**) rARC-CaMK2B. Residues that constitute the hydrophobic NtD cleft are shown in stick format, coloured by atom type. In each structure, the side chain of the aromatic residues buried in the interface (F45 and

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

823 F52, dARC1 CA-NtD; Y229*, rARC CA-NtD-TARPγ2; F313*, rARC CA-NtD-CaMK2B) are
824 coloured purple, yellow and orange respectively. The conserved mainchain hydrogen
825 bonding interactions between the backbone amide and carbonyl of F52 with the carbonyl
826 of L89 and the amide of Y91 (dARC1), of Y229 with the carbonyl of H245 and the amide
827 of N247 (rARC CA-NtD-TARPγ2) and of F313 (rARC CA NtD-CaMK2B) with the carbonyl
828 of H245 and the amide of N247 are shown as dashed lines.
829

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



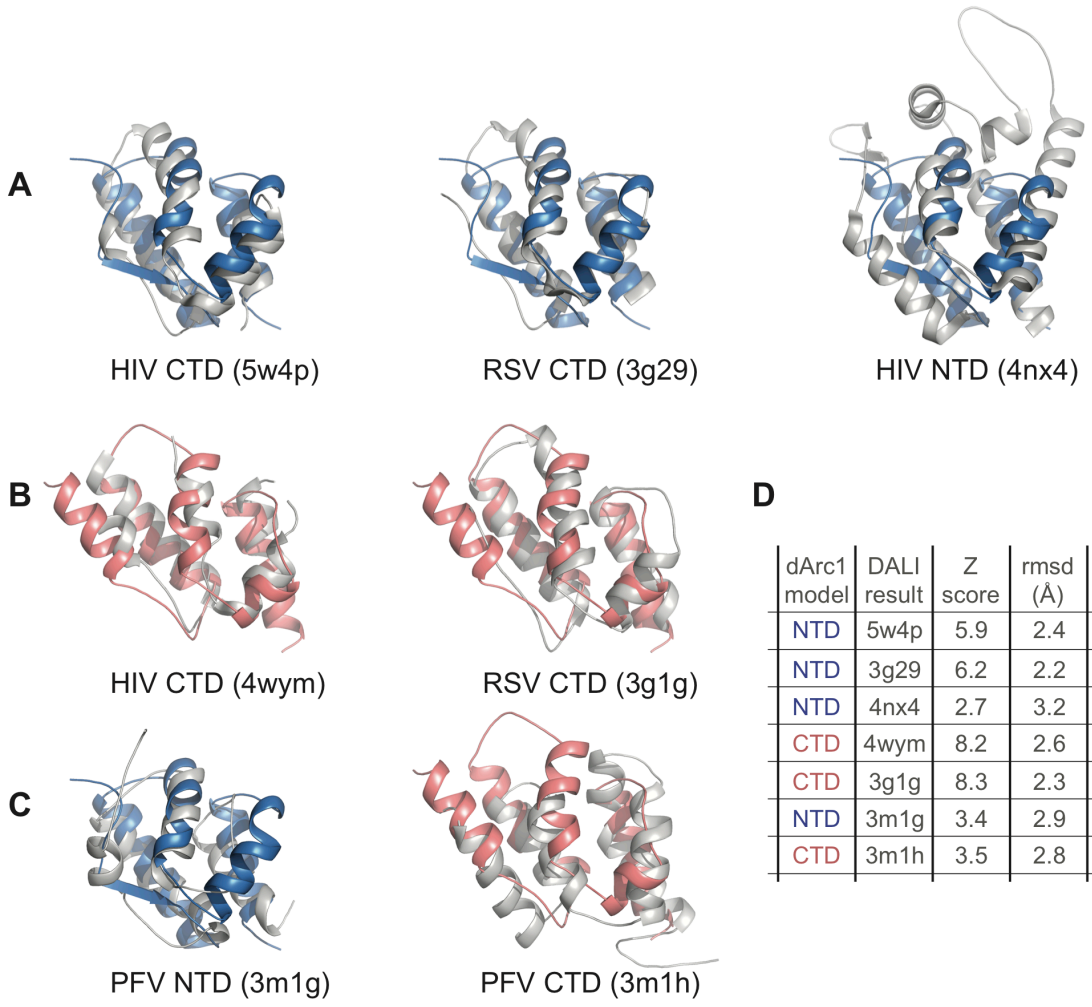
831 **Fig. 4. Comparison of the dARC1 and rARC CA-CTD structures.** (**A**) 3D structural
832 alignment of dARC1 CA-CtD and the rARC CA-CtD from apo rARC (PDB: 6GSE). The
833 structures are shown in cartoon with equivalent helices labelled and shown as cylinders.
834 dARC1 is coloured cyan and rARC is coloured light blue. (**B and C**) Details of the CA-CtD
835 homodimer interfaces. Cartoon representations of the protein backbone of dARC1 CA-CtD
836 (**B**) and rARC CA-CtD (**C**) are shown, coloured as in **A**. The view is of one monomer

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

837 looking into the dimer interface, Residues that make interactions in dARC1 CA and their
838 equivalents in rARC are shown in stick representation colour-coded by residue type
839 (purple, hydrophobic/aromatic; green, polar; red, acidic; blue, basic). (**D and E**)
840 Hydrophobic surface representations of **B** and **C** respectively. Circled in **D** is a distinct
841 hydrophobic patch on the surface of the dARC1 CA-CtD which is absent in rARC. (**F**)
842 Multiple sequence alignment of ARC, dARC1 and dARC2 CA-CtDs, and parent
843 retrotransposon sequences. Group 1 contains tetrapod ARC (tARC) sequences and the
844 closely related *Latimeria chalumnae* (*L. ch*). Above, secondary structure of rARC, numbers
845 according to the rARC (*R. norvegicus*) sequence. Group 2 contains dARC1, dARC2 and
846 closely related *Linepithema humile* (*L. h*) Gypsy11 retrotransposon. Below, secondary
847 structure of dARC1, numbers according to the dARC1 (*D. melanogaster*) sequence). Red
848 box, white text, invariant residues shared between groups. Red text, residues conserved
849 within a group. Asterisks mark the residues at the dARC1 CtD dimer interface and their
850 equivalents in tARCs, as shown in **B** and **C**.
851

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

852



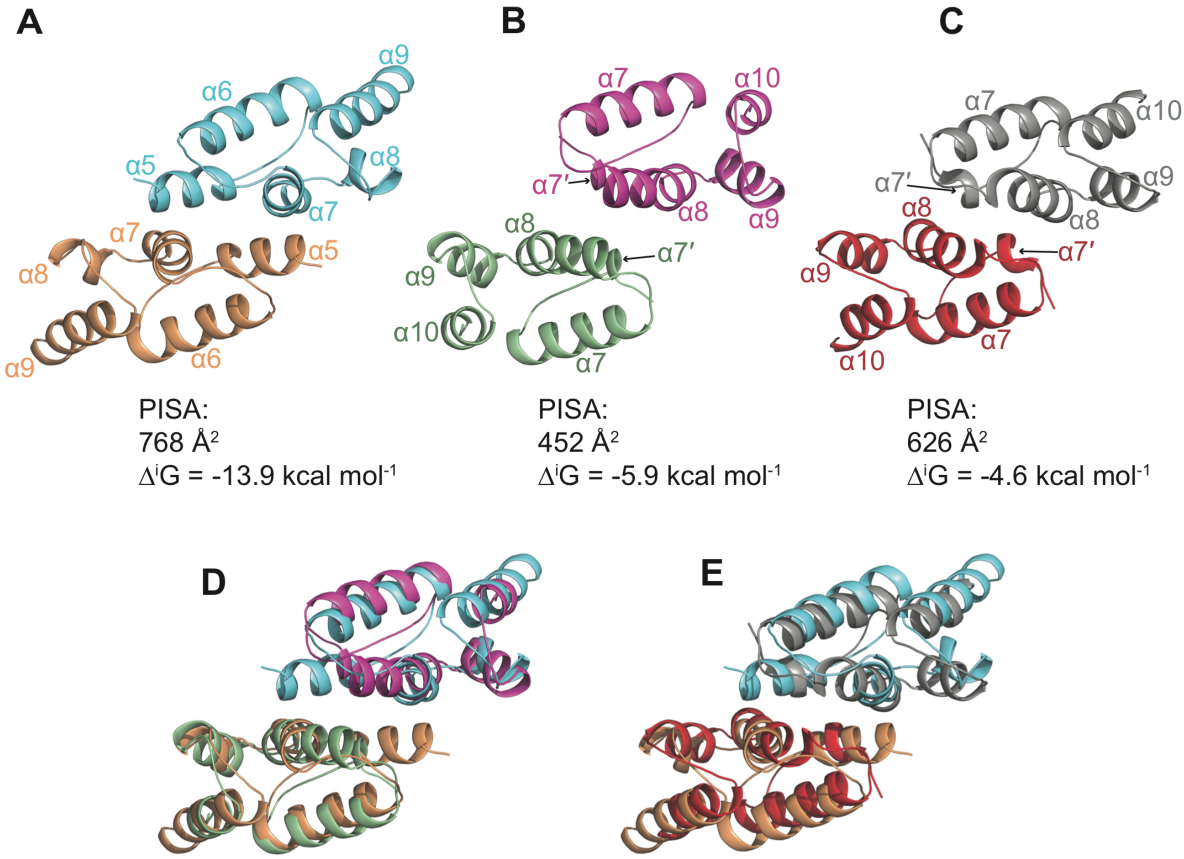
D

dArc1 model	DALI result	Z score	rmsd (Å)	#Cα pairs	seqID (%)
NTD	5w4p	5.9	2.4	64	6
NTD	3g29	6.2	2.2	64	8
NTD	4nx4	2.7	3.2	68	9
CTD	4wym	8.2	2.6	68	12
CTD	3g1g	8.3	2.3	73	7
NTD	3m1g	3.4	2.9	65	12
CTD	3m1h	3.5	2.8	66	10

853 **Fig. 5. Structural similarity with ortho and spumaretroviral CA**. (**A**) Pairwise DALI 3D
854 Cα structural alignment of dARC1 CA-NtD with HIV CA-CtD (Left), RSV CA-CtD (middle)
855 and HIV- NtD (right). In each panel, the cartoon of the dARC1 CA-NtD backbone is shown
856 in blue, the backbone of the aligned structures is shown in grey. (**B**) Pairwise 3D Cα
857 structural alignment of dARC1 CA-CtD with HIV CA-CtD (Left) and RSV CA-CTD (right). In
858 each panel the cartoon of the dARC1 CA-CtD backbone is shown in red, the backbone of
859 the aligned structures is shown in grey. (**C**) Pairwise 3D Cα structural alignment of dARC1
860 CA-NtD with Prototypic Foamy virus (PFV) CA-NtD (Left) and dARC1 CA-CtD with PFV
861 CA-CTD (right). (**D**) DALI Z scores, RMSD, number of aligned residues and sequence
862 identities for 3D Cα alignments.

863

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



865 **Fig. 6. Comparison with retroviral CA-CtD dimers. (A to C)** Cartoon representations of
866 CA-CtD dimers. (**A**) dARC1, cyan and wheat, (**B**) HIV-1, magenta and pale green, (**C**)
867 RSV, grey and red. The orthoretroviral structures are aligned with respect to the dARC1
868 dimer. CtD helices α5 to α9 are labelled in the dARC1 structure and the equivalent α7' to
869 α10 in the orthoretroviral structures. The buried surface area (Å^2) and free energy of
870 interaction (Δ^iG) of each interface, calculated in PDBePISA is displayed below each
871 structure. (**D and E**) Structural alignment of dARC CA with HIV-1 CA and RSV CA dimers
872 respectively. Protein backbones are coloured as in **A, to C**.
873

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Materials

Structure of *D. melanogaster* ARC1 reveals a repurposed molecule with characteristics of retroviral Gag

Matthew A. Cottee[1], Suzanne C. Letham[1], George R. Young[2], Jonathan P. Stoye[2,3] and Ian A. Taylor[1*]

[1]Macromolecular Structure Laboratory, [2]Retrovirus-Host Interactions Laboratory, The Francis Crick Institute, Midland Road, London NW1 1AT, UK. [3]Department of Medicine, Imperial College London, London SW7 2AZ, UK

*To whom correspondence should be addressed

E-mail: ian.taylor@crick.ac.uk

Telephone: +44 (0) 20 3796 2288

File includes

Table S1. dARC1 CA Statistics of data collection, phasing and refinement.

Table S2. Hydrodynamic parameters of dARC1 CA

Fig. S1. Crystal structures of the dARC1 CA.

Fig. S2. dARC1 CA dimer

Fig. S3. Comparison of dARC1and mam-ARC CA-NtDs

Fig. S4. Comparison of dARC1 and Ty3 CA

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

898 **Table S1. dARC1 CA Statistics of data collection, phasing and refinement.**

	dARC1 CA oP (Se-Met, peak)	dARC1 CA oP (Se-Met, lrem)	dARC1 CA oP (Se-Met, lrem STARANISO)	dARC1 CA hP (Se-Met, anom)
Data collection				
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P6_122$
Cell dimensions				
a, b, c (Å)	62.43, 70.715, 142.07	62.35, 70.39, 142.84	62.33, 70.39, 142.82	62.18, 62.18, 401.02
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 120
Wavelength (Å)	0.9794	0.9840	0.9840	0.9794
Resolution (Å)	47.36-2.06 (2.12-2.06)	142.84-1.70 (1.74-1.70)	71.41-1.55 (1.70-1.55)	57.29-2.14 (2.20-2.14)
Anisotropic diff. limits (a*, b*, c*)	-	-	1.55, 2.40, 1.62	
Unique reflections	39409 (2872)†	70026 (5125)	48232 (2412)	26972 (1886)
R_{meas} (%)	21.5 (254.6)‡	13.5 (105.8)	9.3 (150.1)	12.2 (111.0)‡
R_{pim} (%)	8.2 (106.1)‡	4.3 (33.9)	2.9 (47.8)	2.5 (31.6)‡
$CC_{1/2}$	0.998 (0.595)	0.998 (0.568)	0.999 (0.557)	1.000 (0.456)
I/σ(I)	10.3 (1.1)	9.1 (1.5)	16.4 (1.5)	22.0 (2.7)
Completeness (spherical, %)	99.5 (95.4)	100.0 (100.0)	52.9 (11.5)	99.9 (99.3)
Completeness (ellipsoidal, %)	-	-	84.1 (66.0)	-
Multiplicity	12.9 (10.5)	9.6 (9.5)	10.0 (9.7)	40.8 (22.4)
Anom Multiplicity	6.8 (5.5)	-	-	22.4 (11.8)
Phasing				
Significant anomalous signal to: (Aimless)	2.82 Å	-		2.59 Å
No. sites (found/expected)	7/4	-		10/4
Mean FOM (Phenix)	0.28	-		0.324
Local rms density corr (Phenix)	0.89			0.83
Refinement				
Resolution (Å)			71.41-1.70 (1.74-1.70)	53.86-2.30 (2.36-2.30)
refl working/free			41243/2241 (1680/69)	36685/1844§ (2798/121)
R_{work}/R_{free}/Test set size (%)			18.1/21.3/4.9 (25.2/28.2/3.9)	24.2/27.8/4.8 (31.6/31.8/4.1)
No residues/atoms				
Protein			328/2730	321/2658
Ligands			5	
Water			391	20
B-factors (Å²)				
Wilson			19.74	44.15
Protein			23.40	66.04
Ligands			33.17	
Water			31.52	46.10
Overall			24.43	65.89
Geometry				
Bond lengths RMSD (Å)			0.010	0.004
Bond angles RMSD (deg)			1.71	0.697
Ramachandran Outliers			0 %	0 %
Ramachandran Favoured			99.38 %	97.79 %
Molprobity score (N number/percentile)			0.92 (9248/100th)	1.44 (8909/99th)

899 †Values in parenthesis refer to the highest resolution shell, ‡Values quoted for within I^+/I^-, §Friedel pairs separated.

900

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table S2. Hydrodynamic parameters of dARC1 CA

Protein	dARC1 CA
Protein Parameter	
v ($mL.g^{-1}$)	0.731
ρ ($g.mL^{-1}$)	1.005
[a]M_r	19,633
ε_{280} (M^{-1} cm^{-1})	23,600
Sed velocity	
C_{range} (μM)	25-100
[b]$S_{20,w}$ ($x10^{13}$) sec	2.92±0.03
[c]$S_{20,w}$ (calc) ($x10^{13}$) sec	2.93
[d]$D_{20,w}$ ($x10^{7}$) cm^2 sec^{-1}	6.95±0.05
[e]$D_{20,w}$ (calc) ($x10^{7}$) cm^2 sec^{-1}	6.77
[f]M_w (S/D) kD	38.2±1
[g]f/f_0 ($S_{20,w}$)	1.41
[h]f/f_0 ($D_{20,w}$)	1.37
[i]M_w C(S) kD	38.7±3
[j]f/f_0 C(S)	1.41
[k]RMSD C(S)	0.005 – 0.008
Sed eqm	
C_{range} (μM)	25-70
[l]M_w kD	38.9±1
[m]RMSD	0.002 – 0.004
[n]χ^2	0.31

[a]Molar mass calculated from the protein sequence after removal of the C-terminal His-tag
[b]Sedimentation coefficient at standard conditions (water @ 20° C) determined by combining all data from discrete component and C(S) analysis.
[c]Sedimentation coefficient calculated using HYDROPRO-10 from the co-ordinates of the dARC1 CA crystal structure using an atomic element radius (AER) of 2.84 Å.
[d]Translational diffusion coefficient for dARC1 CA determined from combined data from discrete component analysis.
[e]Translational diffusion coefficient calculated from the crystal structure using HYDROPRO-10, AER = 2.84 Å.
[f]The weight averaged molecular weight from discrete component analysis. Error is the range of three measurements
[g]The frictional ratio calculated from $S_{20,w}$ from discrete component analysis.
[h]The frictional ratio calculated from $D_{20,w}$ from discrete component analysis.
[i]The weight averaged molecular weight from C(S) analysis. Error is the range of three measurements
[j]The weight-averaged frictional ratio from the best fit C(S) distribution function.
[k]The range of the rms deviations observed when data were fitted using a continuous sedimentation coefficient distribution model.
[l]The weight averaged molecular weight from Global SE analysis. The error is the range of molecular weights observed for each multi-speed sample when fitted individually.
[m]The range of the rms deviations observed for each multi-speed sample when fitted individually.
[n]The reduced chi-squared for the global fitting of all multispeed data.

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

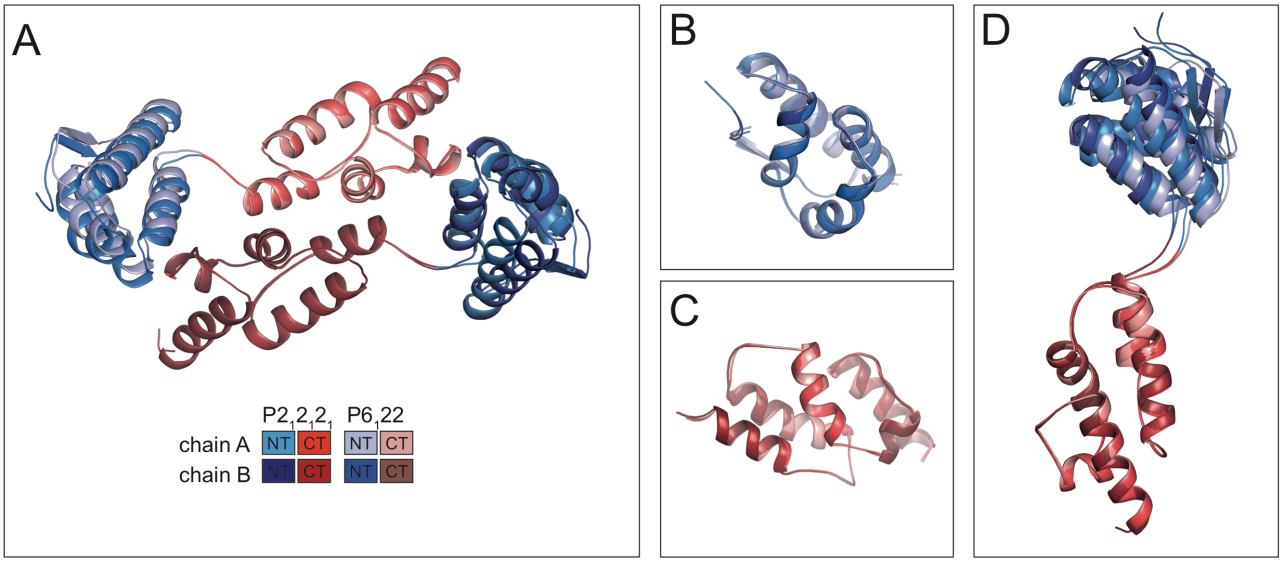


Fig. S1. Crystal structures of the dARC1 CA. (**A**) Structural alignment of orthorhombic ($P2_12_12_1$) and hexagonal ($P6_122$) crystal forms of dARC1 CA that contain an almost identical dimer. The ASUs are shown in cartoon representation, aligned using the dimeric CtDs (RMSD of 0.247 Å across 133 Cα pairs). NtDs are coloured in blue shades, CtDs in red shades, according to the legend. (**B**) Alignment of all four NtDs only, average RMSD 0.34 ± 0.09Å over 66±2 Cαs. (**C**) Alignment of all four CtDs only, average rmsd 0.32 ± 0.12 Å over 74±4 Cαs. (**D**) Alignment of all four chains through the CTD, highlighting that structural differences between chains are due to slight flexibility of the NTD-CTD linker region.

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

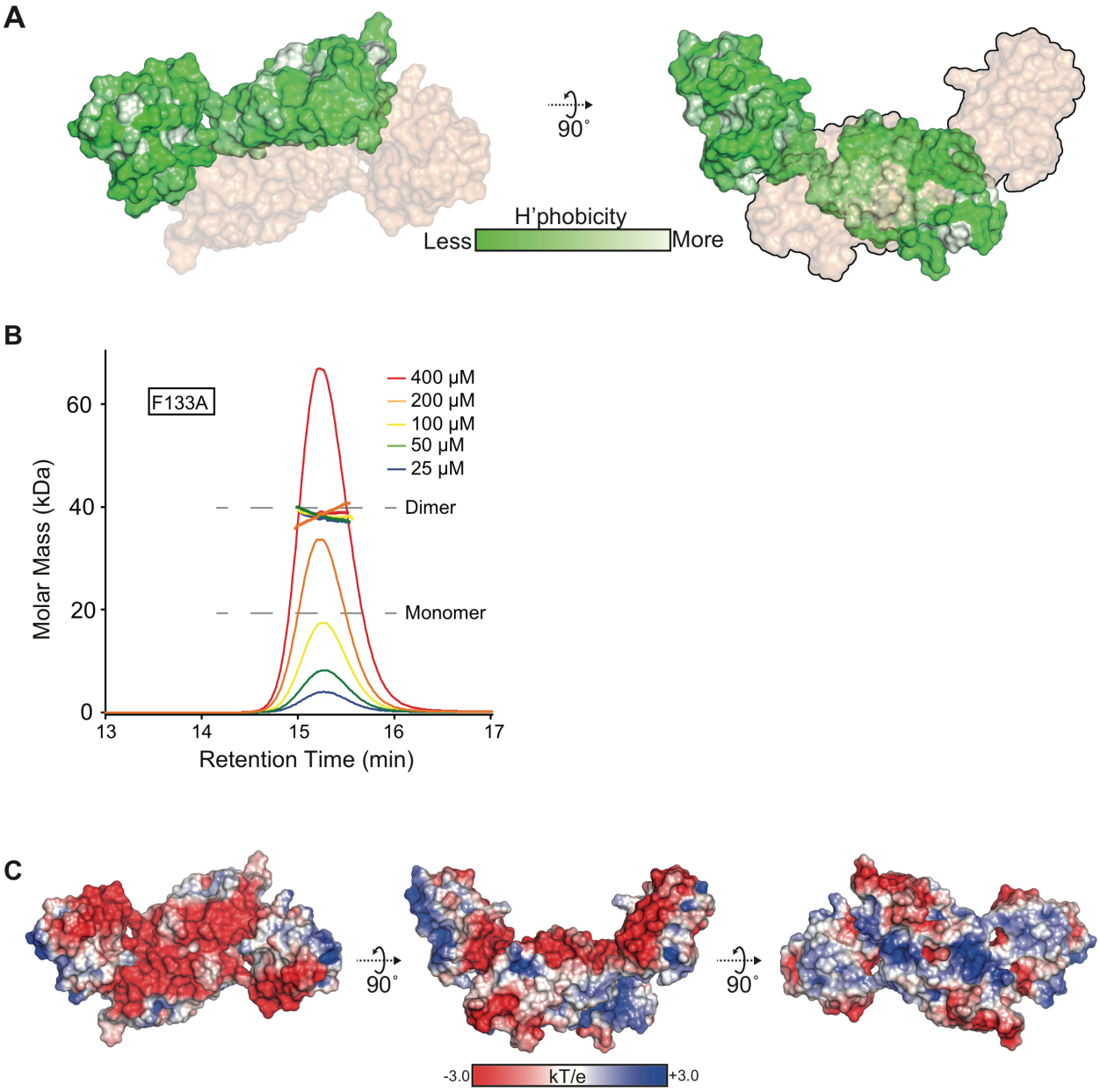


Fig. S2. dARC1 CA dimer (**A**) Surface representation of dARC1 CA-CtD dimer displaying the distribution of surface hydrophobicity/hydrophilicity calculated using the pymol script (https://pymolwiki.org/index.php/Color_h). Greater Hydrophilicity is represented by darker green shading, the lighter and non-coloured regions represent the most hydrophobic areas of the molecule. The orientation in the left- and right-hand panels is the same as in **Fig. 1A** with monomer A displaying surface hydrophobicity, and monomer B shown as a wheat surface. **(B)** SEC-MALLS analysis of dARC1 CA (F133A). The sample loading concentrations were 400 μM (8 mg/mL) (red), 200 μM (4 mg/mL) (orange), 100 μM (2

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

946 mg/mL) (yellow), 50 μM (1 mg/mL) (green) and 25 μM (0.5 mg/mL) (blue). The differential
947 refractive index (dRI) is plotted against column retention time and the molar mass,
948 determined at 1-second intervals throughout the elution of each peak, is plotted as points.
949 The dARC1 CA monomer and dimer molecular mass is indicated with the grey dashed
950 lines. (**C**) Electrostatic surface potential of the dARC1 CA dimer, calculated with APBS.
951 The model was modified to replace SeMet with native Met residues. The view in the left-
952 hand and centre panels is the same as in **A**. The right-hand panel is a view from the C-
953 terminal “underside” of the dimer.
954

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

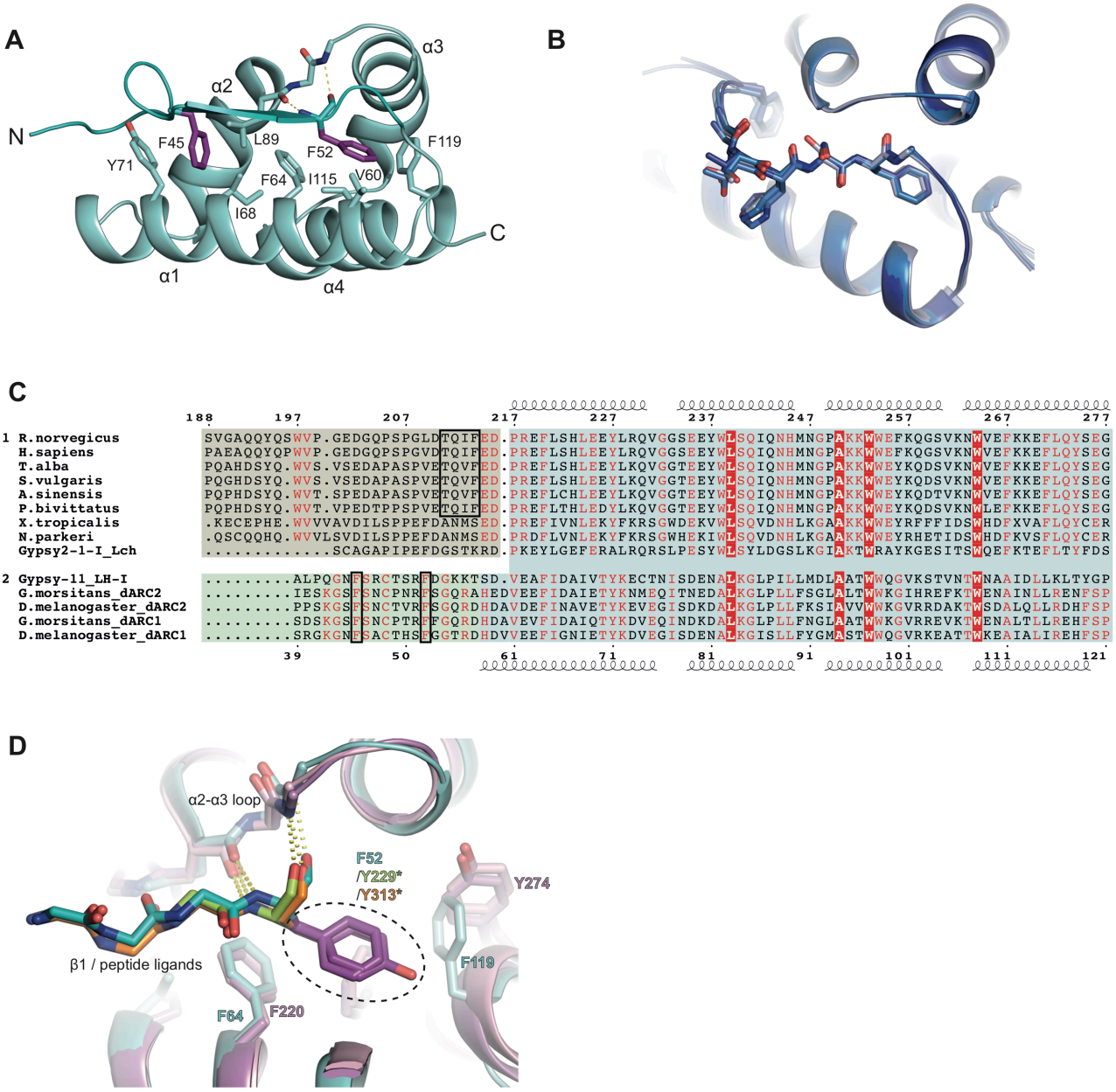


Fig. S3. Comparison of dARC1and mam-ARC CA-NtDs

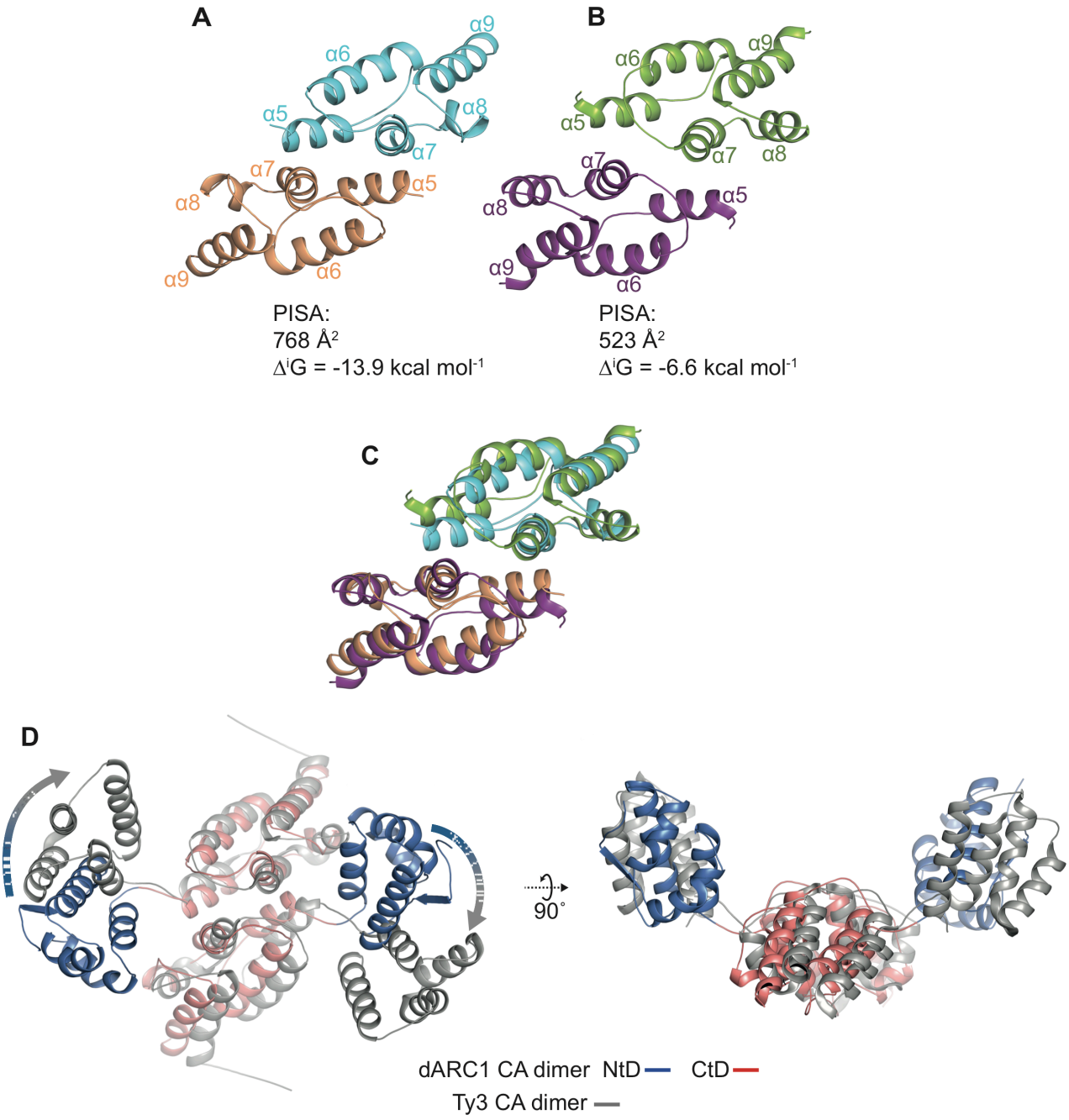
(**A**) The backbone of the dARC1 CA-NtD is shown in cartoon representation in cyan. Residues that contribute to the hydrophobic interface between the core domain, and the native N-terminal strand are shown in stick representation, coloured by atom type, with the conserved aromatic residues F45 and F52 that are buried in the interface coloured purple. The mainchain hydrogen bonding interactions between the F52 backbone amide and carbonyl with the carbonyl of L89 and the amide of Y91 are shown as dashed lines. (**B**) Overlay of the four dARC1 CA NtDs from the two crystal structures, shown in different shades of blue as in **Fig S1B**, highlighting the identical conformation of the native N-

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

965 terminal strand (sticks). (**C**) Multiple sequence alignment of NtDs of ARC and dARC CA
966 sequences. Group 1 contains tetrapod ARC (tARC) sequences and the closely related *L.*
967 *ch* Gypsy2 retrotransposon. Above, rARC secondary structure, numbers according to the
968 rARC (*R. norvegicus*) sequence). Group 2 contains dARC CA sequences and the closely
969 related *L. h* Gypsy11 retrotransposon. Below, dARC1 secondary structure, numbers
970 according to the dARC1 (*D. melanogaster*) sequence). Red box, white text, invariant
971 residues shared between groups. Red text, residues conserved within a group. Blue
972 highlight, alpha-helical core which is conserved between the two groups. Yellow highlight,
973 native N-terminal strand which is conserved within group 1. TQIF motif is boxed in black.
974 Green highlight, native N-terminal strand which is conserved within group 2. Conserved,
975 strand-burying aromatic residues are boxed in black. (**D**) Close-up view of the dARC1 CA-
976 NtD (cyan), aligned with the rARC CA NtD-CaMK2B (light pink-orange) and rARC CA NtD-
977 TARPγ2 (dark pink-lime) structures, showing identical mechanisms of strand burial in the
978 core domain. The native strand (dARC1) or rARC-bound peptides (TARPγ2 and CaMK2B)
979 as well as the NtD α2-α3 loop are shown in stick representation, the equivalent strand-
980 anchoring aromatic residue (circled) in each case is coloured purple.
981

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

982



983 **Fig. S4. Comparison of dARC1 and Ty3 CA. (A and B)** Cartoon representations of CA-
984 CtD dimers. (**A**) dARC1, cyan and wheat, (**B**) Ty3 (PDB: 6r23), magenta and pale green.
985 The structures are aligned with respect to the dARC1 dimer (RMSD = is 4.0 Å, 149 Cα).
986 CtD helices are labelled α5 to α9 according to the dARC1 crystal structure. The buried
987 surface area (Å^2) and free energy of interaction (Δ^{i}G) of each interface, calculated in
988 PDBePISA is displayed below each structure. (**C**) Structural alignment of dARC1 CA-CtD
989 and Ty3 CACtD, protein backbones are coloured as in **A and B**. (**D**) View of the dARC1

bioRxiv preprint doi: https://doi.org/10.1101/691972; this version posted July 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and Ty3 dimers showing the positioning of NtDs with respect to the CtD dimer. The structures are aligned on the CtD dimer as in **A**. Ty3 is coloured in grey, dARC1 CA-NtD and CA-CtD are coloured blue and red respectively. The dARC1 CA-NtDs have to undergo a slight rotational operation to match the conformation seen in the Ty3 dimer (represented by the shaded arrows). Despite this difference, both dimers have the same "glacial trough" arrangement and the dARC1 crystallographic dimer has an NtD-CtD orientation very close to that observed Ty3 icosahedral particles.