Structural studies of cilia and flagella associated protein 410 (CFAP410) reveal its bimodular organization with an N-terminal LRR motif and a C-terminal tetrameric helical bundle

Alexander Stadler[1], Heloisa B. Gabriel[2], Santiago Alonso-Gil[3], Xuan Deng[4], Robbie Crickley[2], Katharina Korbula[1], Barbora Mikolaskova[1], Kaiyao Huang[4], Bojan Zagrovic[3], Sue Vaughan[2], Jack D. Sunter[2], Gang Dong[1,*]

[1] Max Perutz Labs, Vienna Biocenter, Medical University of Vienna, 1030 Vienna, Austria

[2] Department of Biological and Medical Sciences, Oxford Brookes University, Oxford, OX3 0BP, UK

[3] Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna Biocenter 5, 1030, Vienna, Austria

[4] Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, Hubei, China.

[*]Correspondence: gang.dong@meduniwien.ac.at (G.D.)

Keywords: CFAP410, cilium, ciliopathies, flagellum, protein, structure

Abstract

Cilia and flagella are antenna-like organelles protruding from the surface of most cells in the human body. Defects in ciliary structure or function often lead to diseases that are collectively called ciliopathies. Cilia and flagella associated protein 410 (CFAP410) localizes at the basal body of cilia/flagella and plays essential roles in ciliogenesis, neuronal development and DNA damage repair. Multiple single amino acid mutations in CFAP410 have been identified in patients with two types of ciliopathies, spondylometaphyseal dysplasia, axial (SMDAX) and retinal dystrophy with or without macular staphyloma (RDMS). However, the molecular mechanism for how the mutations cause these disorders remains poorly understood due to a lack of high resolution structures of the protein. We have carried out extensive structural and biochemical investigations on *Trypanosoma brucei* CFAP410, as well as its homologs in *Homo sapiens* and *Chlamydomonas reinhardtii.* Multiple crystal structures solved by us reveal that CFAP410 is a bimodular protein comprising two distinct domains: a leucine rich repeat (LRR) motif at the N-terminus and a tetrameric helical bundle at the C-terminus. These two domains are interconnected via a long unstructured loop. We further examined how the disease-causing mutations may affect the folding and/or assembly of CFAP410 using biochemical methods and molecular dynamics simulation. Taken together with our *in vivo* localization data from *T. brucei*, we provide an explanation how the mutations in CFAP410 might cause retinal and skeletal ciliopathies.

Introduction

Cilia and flagella associated protein 410 (CFAP410) is coded by a gene that was mapped to chromosome 21 open reading frame 2 and thus named C21orf2 originally [1]. The gene was first thought as a candidate for a few genetic disorders based on its chromosomal location and the predicted mitochondrial localization of its coded protein [1]. Later mutational analysis excluded CFAP410 as the causative factor for those disorders and instead suggested it is a compelling candidate for retinal dystrophy [2, 3]. Immunofluorescence analysis showed that CFAP410 localizes to the basal body in mIMCD3 and hTERT-RPE1 cells as well as to the base of the connecting cilium in mouse photoreceptors, whereas genome-wide siRNA screen and biochemical studies suggest that it works together with NEK1 and SPATA7 as a functional module in ciliogenesis and DNA damage repair [4-6].

CFAP410 is also one of the genes implicated in amyotrophic lateral sclerosis (ALS), a fatal progressive neurodegenerative disease caused by the loss of motor neurons [7]. Moreover, expression level of CFAP410 is substantially reduced in the brain of Down syndrome patients, implicating its function in neuronal development [8]. The diversified functions of CFAP410 can be linked by its localization and function in cilia, as many neurological disorders are caused by defects in primary cilia [9].

CFAP410 exhibits significant sequence conservation across phyla and is present in all sequenced genomes of ciliates, including algae, protists, and animals. *Homo sapiens* CFAP410 (HsCFAP410) consists of 255 amino acids. A number of single residue mutations of HsCFAP410 have been identified in patients with skeletal and/or retinal disorders including spondylometaphyseal dysplasia, axial (SMDAX) and retinal dystrophy with or without macular staphyloma (RDMS). Both disorders are genetically autosomal recessive and manifested as dysplasia of the ribs, vertebral bodies, ilia, proximal femora in

skeletons, and retinitis pigmentosa in eyes [10, 11]. These clinically identified mutations include I35F, C61Y, R73P, Y107C, Y107H, V111M and L224P [4, 12-15]. However, it remains mysterious how these single-residue mutations of HsCFAP410 cause so deleterious defects due to a lack of high resolution structures of the protein.

Here we report our extensive structural and biochemical characterizations on three homologs of CFAP410 from *H. sapiens, Trypanosoma brucei* and *Chlamydomonas reinhardtii.* Our work revealed that CFAP410 is a bimodular protein comprising two highly conserved domains, including a leucine rich repeat (LRR) motif at its N-terminus and an oligomeric helical bundle at its C-terminus. We have determined crystal structures for the C-terminal domain (CTD) of all three homologs as well as the N-terminal domain (NTD) of *T. brucei* CFAP410 (TbCFAP410). These structures allow us to examine how the mutations might affect the folding and/or assembly of the protein. We further checked the localization of the wild-type and disease-causing mutations in *T. brucei*. Taken all these results, we provide an explanation how single-residue mutations in CFAP410 might cause ciliopathies.

Results

Structural studies of the C-terminal domain of CFAP410

To investigate the three-dimensional structure of CFAP10 proteins we cloned the coding genes from *H. sapiens*, *T. brucei* and *C. reinhardtii*. However, all full-length proteins were insoluble upon expression in *Escherichia coli*. Our folding analyses suggested that all these CFAP410 proteins contain two folded regions, including a large domain at the N-terminus and a small one at the C-terminus (**Fig. S1A-D**). The two domains are connected by a long disordered loop. Sequence alignments of homologs from various

Trypanosoma species showed that both domains are highly conserved, whereas the connecting loop has very low homology and is variable in length (**Fig. S1E**). Similar conservation patterns are also seen in homologs from mammals and algae (**Fig. S2A-C**), and such conservation pattern is also preserved across species (**Fig. S2D**). We therefore took a divide-and-conquer strategy to investigate the structure of either domain of CFAP410 separately.

We first tried to crystallize the short CTDs of *H. sapiens*, *T. brucei* and *C. reinhardtii* CFAP410 proteins. Multiple truncations were chosen based on sequence conservation and secondary structural predictions. Synthetic poplypeptides were purified and used for crystallization trials. The CTD truncations that finally generated diffracting crystals were fragments covering residues 256-295, 212-246 and 213-246 of *H. sapiens*, *T. brucei* and *C. reinhardtii* CFAP410 proteins, and their structures were determined at 1.50, 1.29 and 1.40 Å resolution, respectively (**Table 1**). $2F_o$-F_c electron density maps for all three structures show very high quality, with the side-chains of almost all residues in the three CTDs having been confidently built and refined (**Fig. S3**).

All three crystal structures show a similar homotetrameric arrangement, with each monomer consisting of two α helices that form an open nutcracker-like conformation (**Fig. 1A-C**). Despite small variations in the tetramers (**Fig. 1D**), all monomers in the three proteins are essentially the same, with root-mean-square deviation (RMSD) of ~0.5 Å for all aligned backbone atoms (**Fig. 1E**). Each of the two helices consists of 13-15 residues (**Fig. 1F**).

The helical bundle of CFAP410-CTD is a dimer of two individually folded dimers

Multiple single-residue mutations of HsCFAP410 have been identified in patients with SMDAX or RDMS, one of which is L224P located in its CTD [4]. The crystal structures

reported in the present study allowed us to examine how this mutation may affect the folding and/or assembly of the domain. We observed in our structures that two CTD monomers form a criss-cross dimer, which further packs into a tetramer in a head-to-head manner with another equivalent dimer. The centrally packed N-terminal helix (i.e. H1) consists of mostly hydrophobic residues that are involved in either intra- or inter-dimer interactions, whereas the peripheral C-terminal helix (i.e. H2) is less hydrophobic and contains multiple polar or charged residues (**Fig. 2A**). Besides the aforementioned mutation site L224 in H1, there is another highly conserved alanine (Ala) residue that is located in the center of the same helix (**Fig. 2B**). While the Leu residue is fully buried in the dimeric interface (**Fig. 2C**), the Ala residue is situated on the hydrophobic inter-dimer interface, with the four equivalent copies arranged in a diamond-like shape across both interfaces (**Fig. 2D**).

To find out how the conserved Leu and Ala residues contribute to the formation of the tetramer, we first examined the wild-type and mutants A219E and L224P of CrCFAP410-CTD using the static light scattering (SLS) method (**Fig. 2E**). Our data demonstrated that the wild-type protein formed a tetramer as revealed in the crystal structure, but mutant A219E became a dimer. Surprisingly, mutant L224P completely disrupted the oligomeric structure and formed only a monomer. The same results were also observed for the CTDs of *H. sapiens* and *T. brucei* CFAP410 (**Fig. 2F**).

We further checked whether either of these two mutants affects the folding of the proteins using the circular dichroism (CD) method. Our results showed that for HsCFAP410-CTD both the wild-type and mutant A219E adopted the same conformation (**Fig. 2G**). However, the disease-causing mutant L224P lost all characteristic helical features and became completely unfolded. Consistent results were also seen for both *T. brucei* and *C. reinhardtii* CFAP410 proteins (**Fig. 2H**).

We further carried out molecular dynamics simulation analyses on wild-type and mutant proteins. Our conformational analyses on HsCFAP410-CTD showed that while the simulated wild-type model agreed very well with the crystal structure (RMSD = 0.8 Å), simulated model of L224P was drastically different from the wild-type crystal structure, with the associated all-atom RMSD of 4.4 Å (**Fig. S4A**). Such structural changes were seen in all four subunits of the complex structure (**Fig. S4B**). The most dramatic difference can be attributated to the major disruption of intra-dimeric contacts (**Fig. S4C**). The simulation analysis also indicated significant changes in the hydrophilic and hydrophobic components in mutant L224P in comparison with the wild-type structure (**Fig. S4D, E**).

The N-terminal domain of TbCFAP410 contains a leucine rich repeat (LRR) with a long C-terminal hairpin-like loop essential for structural stability

To provide a full structural view of CFAP410, we next cloned the N-terminal region of CFAP410 from *H. sapiens*, *T. brucei* and *C. reinhardtii*. In each case we made multiple truncation constructs based on folding and secondary structure predictions and tried different fusion tags, including His_6, SUMO, and maltose-binding protein (MBP). No soluble proteins could be obtained for any constructs of HsCFAP410 and CrCFAP410. For TbCFAP410 we could produce soluble proteins for four truncations covering residues 1-177, 1-170, 1-164 and 1-160, whereas two other shorter constructs spanning residues 1-149 and 1-140 were insoluble upon expression.

Purified proteins of all four soluble TbCFAP410-NTD truncations were subjected to extensive crystallization screenings. In the end we obtained crystals only for the shortest construct spanning residues 1-160, which is what NTD refers to in all subsequent structural analyses of TbCFAP410. Single crystals of TbCFAP410-NTD diffracted X-rays to approximately 1.0 Å resolution at the ESRF synchrotron site (**Table 1**). Crystal structure

was determined by the molecular replacement method using a homology model for residues 16-138 of TbCFAP410 generated by SWISS-MODEL [16] based on the crystal structure of the *C. reinhardtii* axonemal dynein light chain-1 (PDB code: 6L4P) [17].

The final structure shows that all residues in the construct could be traced and confidently built (**Fig. 3A**). Residues 1-138 fold into a canonical leucine-rich repeat (LRR) motif with five tandem repeats forming a calzone-like structure. The whole domain contains five parallelly arranged β strands forming a flat sheet on one side and an array of four short α helices on the other. The region between the α helices and the β strands constitutes the hydrophobic core of the LRR motif and is tightly packed with highly conserved leucine residues. Additionally, there is a short helix on either end of the LRR motif that acts as a cap and is oriented roughly perpentidular to the β sheet. The overall structure is slightly bent toward the β sheet side as seen in other LRR structures. Additionally, we found that residues 139-160 of TbCFAP410 form a long hairpin-like loop that is tightly packed on top of the helical surface of the LRR domain (**Fig. 3B**). The loop makes extensive contacts with the LRR motif via multiple hydrogen bonds and many hydrophobic interactions (**Fig. S5A**). Those residues in the loop involved in its interaction with the LRR motif are highly conserved in all homologs from various kinetoplastids, as are the interacting residues on the LRR side (**Fig. S5B-D**).

The N-terminal LRR domain of CFAP410 contains a highly conserved surface patch

Sequence alignments of CFAP410 homologs from mammals, trypanosomes and algae revealed an extensive surface patch on the flat β sheet of the LRR motif, which includes eighteen hydrophobic or charged residues that are highly conserved in all homologs across phyla (**Fig. 3C**; **Fig. S6A, B**). As shown before, the opposite side of the LRR motif is covered by a long hairpin-like loop in TbCFAP410. Although such loop is

conserved in kinetoplastids, it is essentially devoid of any conserved residues in all other homologs from animals and algae (**Fig. 3D**; **Fig. S6C**).

Structural basis for the disease-causing mutations on CFAP410

Multiple single-residue mutations in HsCFAP410 have been identified in patients. Except for L224P discussed above, all other mutations are located within the N-terminal LRR domain of HsCFAP410, including V37F, C63Y, R75P, Y107C, Y107H and V111M. Sequence alignment shows that the LRR motif is highly conserved in mammals, trypanosomes and algae (**Fig. S6A**). Except for R75, all sites struck by disease-causing mutations are highly conserved hydrophobic residues. The LRR motifs of *H. sapiens* and *T. brucei* CFAP410 share 41% identity and 60% overall similarity (**Fig. S7A**). Comparison of the crystal structure of TbCFAP410-NTD with homologous models of HsCFAP410 (aa1-130) generated by SWISS-MODEL (template-based) [16] and AlphaFold2 (*ab initio*) [18] shows that they all share a very similar three-dimensional structure (**Fig. S7B**). Particularly, all the five disease-causing mutation sites mentioned above adopt essentially the same conformation in these structures. Statistics reports showed that both predicted models of HsCFAP410 were built with high confidence (**Fig. S7C-E**). These allowed us to examine how the mutations of single residues in HsCFAP410 might affect its structure and function based on the crystal structure we have determined for TbCFAP410-NTD.

Location of individual residues corresponding to the five disease-causing mutations in HsCFAP410 can be clearly visualized in TbCFAP410 (**Fig. 4A, B**). These residues can be grouped into three categories, including one surface residue that is fully exposed to solvent (R75), two semi-exposed residues (V37 and Y107), and two hidden residues that are completely buried in the hydrophobic core of the structure (C63 and V111) (**Fig. 4C-E**). Given that most of these residues are hydrophobic and partially or fully buried in the

structure, we therefore examined whether these disease-causing mutations affect the folding or structural stability of the protein. Wild-type and mutant TbCFAP410-NTD proteins with an N-terminal His_6 tag were expressed in *E. coli*. After breaking the cells and spinning down the cell lysate, we checked the soluble fractions in the supernatants and the insoluble parts in the pellets. Our data show that wild-type TbCFAP410-NTD were present in both fractions, suggesting that it is at least partially soluble (**Fig. 4F**, lanes 2-4). In contrast, however, all the six mutants were present only in the pellets but not in the supernants (**Fig. 4F**, lanes 5-22), demonstrating that they are insoluble and thus likely mis-folded.

In vivo **assays revealed abnormal localizations of the disease-causing mutants of TbCFAP410**

To determine the impact of these disease-causing mutations on the function of TbCFAP410 *in vivo,* we generated individual mutant constructs with an N-terminal mNeonGreen fluorescent protein tag and expressed them in *T. brucei* using a tetracycline inducible expression system. All expressions were induced for 24 h. The cells, which also expressed SAS6 endogenously tagged with mScarlet as a marker of the basal body and pro-basal body, were then examined by wide-field fluorescence microscopy (PMID: 33165561). We observed mNeonGreen fluorescence signals in approximately 80% of the cells examined. The wild-type protein localized to the basal and pro-basal bodies as well as the posterior cell tip, together with a weak cytoplasmic signal (**Fig. 5A**, left), matching the localization observed in the TrypTag project [19]. The localization of the majority of the mutants was similar, with a clear posterior cell tip signal and a basal body signal but a variable signal from the cytoplasm.

For the mutants C63Y and L272P the pattern was different with no clear basal body signal. However, the cytoplasm signal was strong in these mutants, which could have

masked the basal body signal (**Fig. 5A**, left). Additionally, we examined two clustered mutations within the surface patch that is highly conserved across all phyla, either mutating only the central ridge of the patch (4 mut: E68K/Y70A/R72E/K73E) or a more extensive one with two extra flanking residues all mutated (6 mut: E68K/Y70A/R72E/K73E/N26A/R120E) (**Fig. 5B**). The localization patterns of these two mutants were similar to that of the wild-type protein (**Fig. 5A**, left).

To investigate whether these mutants were able to localize to the basal body and how stably they were associated with the cytoskeleton, we next generated detergent extracted cytoskeletons of the induced cells and quantified the localization of these mutants. Similar to what was seen in live cells, the wild-type protein remained stably associated with both the basal body and the posterior cell tip, with a small number of cytoskeletons only having a basal body signal (**Fig. 5A**, right; **Fig. 5C**). The same pattern was also observed for Y70A, R75P, 4 mut and 6 mut. The Y107C mutant in approximately half the cytoskeletons had only a weak signal at the posterior cell tip and this weak signal was associated with a basal body localization, with the mutant localizing clearly to the posterior cell tip and basal body in an additional ~25% of cytoskeletons (**Fig. 5C**). For mutants C63Y, L272P and Y107H there was a significant reduction in the cytoskeletons with a basal body and posterior cell tip signal, and this was most extreme in the L272P mutant in which no signal observed at either the basal body or the posterior cell tip. This demonstrates that these mutants (L272P, C63Y, Y107H) were unable to stably associate with the basal body and posterior cell tip, which suggests that in human cells these equivalent mutations could consequently cause a defect in cilia function.

Discussion

CFAP410 is a protein present in nearly all sequenced ciliates and plays an essential

role in ciliogenesis [4]. Located mainly at the basal body, the protein is highly conserved across phyla including algae, protista and animals. Here we report our structural studies on three homologs of CFAP410 from *T. brucei*, *C. reinhardtii*, and *H. sapiens*. Our results show that CFAP410 is a bimodular protein comprising two folded domains at both ends, i.e. NTD and CTD, which are connected via a variable unstructured loop (**Fig. S1**). All clinically identified single-residue mutations causing ciliopathies are located in the folded NTD or CTD, but not in the variable linker. We have determined crystal structures of the CTD of these three CFAP410 homologs, which all adopt a similar nutcracker-like conformation composed of two helices (**Fig. S3**). The CTDs from two molecules first form an intercalated homodimer; two such dimers further assemble into a tetramer by packing head-to-head via their first helices (**Fig. 1**).

Multiple single-residue mutations in HsCFAP410 have been identified in patients with SMS or RDMS diseases [13-15]. The only mutation within the CTD of HsCFAP410 is L224P. Our *in vitro* biochemical studies showed that such mutation disrupts its helical property and causes complete disassembly of the tetramer (**Fig. 2**). The same observation was also seen for the corresponding mutation in the CTD of both *T. brucei* and *C. reinhardtii* CFAP410 proteins, which further confirmed the highly conserved structure of CFAP410 in all distantly related ciliates.

Except for L224P, all other six ciliopathy mutations locate in the NTD of HsCFAP410. Although we did not obtain soluble HsCFAP410-NTD, we have managed to solve a high-resolution crystal structure of TbCFAP410-NTD (**Fig. 3**). The two proteins share 41% identities and adopt essentially the same conformation (**Fig. S7**), which allowed us to extrapolate from the structure of TbCFAP410-NTD how the disease-causing mutations might affect the structure and function of HsCFAP410. We found that all the corresponding disease-causing mutations in TbCFAP410-NTD cause structural instability as overexpressed proteins in *E. coli* were completely insoluble (**Fig. 4**). We further examined

the structure trying to understand why these mutations cause such dramatic change in the folding of the protein. Firstly, for mutation R73P of HsCFAP410, the equivalent residue is R75 of TbCFAP410, which is located in a short linker tightly packed between an α helix and a β strand in the crystal structure (**Fig. 6A**). We found that its fully solvent exposed side-chain forms a stacking interaction with that of residue E53 via two hydrogen bonds. Additionally, the amide proton of residue R75 forms an additional hydrogen bond with the carbonyl group of residue L54. Mutating R75 to Pro would disrupt all these interactions. Furthermore, the circular side-chain of Pro would additionally cause clashes with the neighboring residues in the tightly packed region of the structure, and thus destabilizes the structure of the protein.

All other five ciliopathy mutations within HsCFAP410-NTD occur at highly conserved hydrophobic residues. Our structural analyses on the equivalent residues in TbCFAP410-NTD show that residue V37 (corresponding to V35 in HsCFAP410) is packed tightly against a hydrophobic cleft; its replacement by Phe (i.e. V35F in HsCFAP410) would cause severe clashes between the bulky aromatic side-chain and the neighboring residues around the cleft (**Fig. 6B**). Regarding the two mutations Y107C and Y107H, residue Y107 is located in a semi-open eggshell-like pocket. The hydroxyl group of Y107 forms a hydrogen bond with the side-chain of a neighboring residue. Mutation Y107C would apparently disrupt this hydrogen bond as well as generate a void cavity (**Fig. 6C**). It is similar for mutation Y107H, but this mutation would additionally cause clashes between the side-chain of His and other hydrophobic residues around the pocket. For the other two mutants C63Y and V111M, both residues are fully buried inside the hydrophobic core of the structure (**Fig. 6D, E**). Their replacements by residues with a bulkier side-chain would inevitably cause severe clashes and thus disrupt the folding of CFAP410-NTD. Our molecular dynamics simulation analyses indeed demonstrated significant changes in all these disease-causing mutants, particularly C63Y, in comparison with the wild-type crystal

structure (**Fig. 7A**). Similar simulation results were also observed for all other disease-causing mutants (**Fig. 7B**).

To check how the disease-causing mutations affect the function of CFAP410 in the native environment, we carried out *in vivo* assays on wild-type and mutant TbCFAP410 in the parasite. Our results showed that wild-type TbCFAP410 localized to both the basal body and the posterior tip of the cell. However, mutation L272P completely lost its localization at the basal body and the posterior cell tip, suggesting that the CTD, possibly its oligomerization, plays an essential role in directing the specific localization of TbCFAP410 in the cell (**Fig. 5**). In contrast, mutations in TbCFAP410-NTD only partially disrupted their *in vivo* localization, with mutations C63Y and Y107H showing the most severe changes. The inability of these mutants to stably associate with the basal body and posterior cell tip in *T. brucei* suggests that in human cells these equivalent mutations are likely also affect the localization of HsCFAP410 and hence cause a defect in cilia function.

In the crystal structure of TbCFAP410-NTD there is also an extensive surface patch across all five strands of the β sheet. This patch consists of eighteen highly conserved residues that are arranged in an “X” shape, with an invariable aromatic residue (Y70 in TbCAFP410) in the central crossing point of the patch (**Fig. 3C**; **Fig. S6**). We speculate that this conserved patch might play an important role for the function of CFAP410 in the cell. Therefore, we made three mutations in this patch, including a single-residue mutant Y70A and two other more extensive mutants, “4 mut” (E68K/Y70A/R72E/K73E) and “6 mut” (N26A/E68K/Y70A/R72E/K73E/R120E). However, none of these mutations showed significant defects in cellular localization of the protein in *T. brucei* (**Fig. 5**). One of the reasons might be that the function of the patch is not for the localization of CFAP410 but instead for its other functions that could not be detected in this localization assay we performed. Another reason why we did not see a severe defect in flagellar structure or cell growth could also be explained by the presence of a close paralog of CFAP410 in *T. brucei*

(Tb927.10.5130), which also localizes to the basal body, pro-basal body and posterior cell tip [19]. Sequence alignment shows that the paralog has a similar size to TbCFAP410 and their N-terminal regions adopt a similar conformation (**Fig. S8A, B**). We also found that most residues in the aforementioned surface patch on TbCFAP410 are also conserved in the paralog (**Fig. S8C, D**). It is possible that the two proteins might function redundantly *in vivo* and the paralog could have compensated the loss of the function of TbCFAP410 in all the three patch mutants we tested.

We also carried out depletion tests in *C. reinhardtii* to check if CFAP410 plays an essential function in ciliary assembly in algae. We have successfully generated two targeted knockout mutants of CrCFAP410 using CRISPR/Cas9 (**Fig. S9A-C**). However, although there was notable reduction in both flagellar length and percentage of flagellated cells in the two mutants in comparison to the wild-type cells, the differences were not as dramatic as what would be expected for an essential protein (**Fig. S9D, E**). Interestingly, we found there is a truncated version of CrCFAP410, which we named “CrCFAP410-short”. CrCFAP410-short contains exactly the same N-terminal region (aa1-132) as CrCFAP410, but it lacks the whole C-terminal oligomeric domain present in CrCFAP410 (**Fig. S9F**). We hypothesize that CrCFAP410-short may work redundantly to CrCFAP410 and could thus compensate the depleted CrCFAP410 to maintain normal flagellar structure and function.

Although only a single paralog or truncated form of CFAP410 is present in *T. brucei* and *C. reinhardtii*, there are totally eleven isoforms of CFAP410 in human cells. In contrast to the paralogs in *T. brucei* and *C. reinhardtii* that either contains a different CTD or completely lack the CTD present in CFAP410, all the eleven HsCFAP410 isoforms contain an identical CTD (**Fig. S10A**). But these isoforms are diversified in their N-terminal part comprising the LRR motif or the linker region that connects the two folded domains. A tetrameric model of HsCFAP410 generated by AlphaFold showed that the CTD forms a tightly packed helical bundle like those seen in our crystal structures (**Fig. 1**), while each

N-terminal LRR motif of the four molecules is loosely connected to the helical bundle via a long unstructured loop (**Fig. S10B**). This large number of CFAP410 isoforms in humans might be formed from alternative splicings, variable promoter usage, or other post-transcriptional modifications of the same gene, with each unique mRNA sequence subsequently produces a specific form of the protein [20-22]. Whether these isoforms are expressed in a tissue-specific manner like many other human proteins would need to be further investigated in the future. The fact that they all contain the same CTD that forms an oligomeric structure suggests that this domain likely plays an important role for the function of CFAP410 in humans. Despite variability in the the connecting loops, most of these isoforms also contain an intact NTD with nearly identical sequences in their structural cores. It suggests that the NTD likely also plays a critical role in their function.

As mentioned above, the helical side of the LRR motif lacks conservation across phyla, but a cross-species conservation pattern does exist in all kinetoplastids (**Fig. S5**). The hairpin-like loop in the crystal structure of TbCFAP410-NTD spans over 20 residues (aa139-160) and makes extensive contacts with the LRR motif via multiple hydrogen bonds as well as numerous hydrophobic interactions (**Fig. S5**). Those residues in the loop involved in its interaction with the LRR motif are highly conserved in all kinetoplastid homologs, so as the corresponding interaction residues on the LRR motif. All these suggest that this extensive loop is a unique feature present only in kinetoplastid species. As mentioned above, all TbCFAP410 truncations with the intact loop were soluble, whereas those lacking parts or the whole loop, i.e. constructs covering residues 1-149 and 1-140, were all insoluble upon expression in *E. coli*. This further suggests that the loop is essential to facilitate the folding and/or to stablize the structure of the LRR motif in TbCFAP410. The stabilizing loop might be an ancient element retained during evolution in the early branched flagellated Kinetoplastida.

In summary, taken together our extensive structural, biochemical and *in vivo* studies,

we conclude that CFAP410 forms a tetramer via its CTD that forms a tightly packed eight-helix bundle, which fusions the four globular NTDs together to form a clustered protein complex (**Fig. 8**). The NTD of CFAP410 adopts the conformation of a canonica LRR motif and contains an extended surface patch; the patch is highly conserved across distant phyla including algae, protist parasites and mammals, and thus likely plays an essential role in the *in vivo* function of CFAP410. The molecular mechnism for how the signle-residue mutations in HsCFAP410 cause human diseases seems to be either the disruption of the oligomeric assembly of the CTD (L224P) or via destroying the three-dimensional conformation of the NTD (I35F, C61Y, R73P, Y107C, Y107H and V111M).

Materials and Methods

Cloning and site-directed mutagenesis

Sequence encoding full-length TbCFAP410 was amplified from *T. brucei* genomic DNA using PCR and ligated into the pET15b vector (Novagen), which provides an N-terminal His_6 tag followed by a thrombin cleavage site prior to the target protein. Gene fragments of all truncated constructs were subsequently generated in a similar manner using the full-length construct as the PCR template and ligated into either pET15b or a custom vector that provides a maltose-binding protein (MBP) and His_{10} tag to the N-terminus of the target proteins. The MBP-His_{10} tag can be cleaved off by the tobacco etch virus (TEV) protease. Expression constructs for HsCFAP410 and CrCFAP410 were made from synthetic DNA fragments ordered from Twist Bioscience. All single- and multi-residue mutations of TbCFAP410-NTD were generated using the megaprimer-based mutagenesis strategy as reported [23]. Incorporation of mutations was confirmed by DNA sequencing.

Protein expression and purification

All recombinant CFAP410 proteins were expressed in *E. coli* (strain BL21-DE3). The cloned constructs for protein expression were used to transform competent bacterial cells. The cells were grown in Luria-Bertani (LB) medium at 37°C to an OD_{600} of 0.6-0.8, and then subjected to cold shock on ice for 30 min. Protein expression was induced by addition of 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), and cell cultures were further incubated at 18°C overnight. Cells were harvested by centrifugation in a Sorvall GS3 rotor (6,000×g, 12 min, 4°C) and then resuspended in 10 ml of lysis buffer [20 mM HEPES pH 7.5, 100 mM NaCl, 20 mM imidazole, 1mM β-Mercaptoethanol, 1 mM phenylmethylsulfonyl fluoride (PMSF)] per L of cell culture.

Resuspended cells were lysed in an EmulsiFlex-C3 homogenizer (Avestin) at 40 bars and cell debris was pelleted by centrifugation (40,000×g, 40 min, 4°C). The supernatant was filtered (0.45-µm pore size, Amicon) and loaded onto a Ni-HiTra® HP column pre-equilibrated with the same lysis buffer in an ÄKTA™ Pure chromatography system (Cytiva). The column with bound proteins was washed with 2× column volume (cv) of lysis buffer, and bound protein was subsequently eluted using a linear gradient concentration of imidazole (20 - 500 mM, 20×cv) in the same lysis buffer.

The N-terminal MBP-His_{10} tag of all TbCFAP410 truncated proteins were cleaved off by incubating the pooled fractions of interest (20-40 ml) with ~1% (w/w) of the TEV protease in a dialysis bag against 2-4 L of dialysis buffer containing 20 mM HEPES (pH 7.5) and 100 mM NaCl (4°C, overnight). The dialyzed samples were loaded on a Ni-HiTra® HP column to get rid of the cleaved tag and any uncut proteins. Target proteins in the flow-through were further purified on a Superdex-200 16/60 column (Cytiva) with a running

buffer containing 20 mM HEPES (pH 7.5), 100 mM NaCl, and 1 mM dithiothreitol (DTT). The eluted proteins were pooled and concentrated, and final concentration was determined using Ultraviolet (UV) absorbance at 280 nm. The samples were either used freshly or stored at -80°C for subsequent experiments.

Preparation of synthetic polypeptides

Wild-type and mutant polypetides of the C-terminal regions of *T. brucei, H. sapiens* and *C. reinhardtii* CFAP410 were chemically synthesized and purified by an in-house facility. All HsCFAP410-CTD fragments were soluble in the sample buffer (20 mM HEPES pH 7.5, 100 mM NaCl, 1 mM DTT), but those of TbCFAP410 and CrCFAP410 only partially dissolved in the buffer. We therefore carried out refolding for the insoluble polypeptides using the following protocol. Dry peptide powder was first dissolved in 8 M urea at a concentration of 10 mg/ml, the solution was then quickly diluted by 10× fold using the sample buffer. After dialysis against the same buffer overnight at 4°C, the samples were loaded onto a Superdex 75 Increase 10/300 GL column (Cytiva). Fractions containing the target polypetides were pooled and concentrated to 10-15 mg/ml. Sample homogeneity was confirmed on both SDS PAGE and native polyacrylamide gels.

Crystallization and structural determination

Purified TbCFAP410-NTD and synthetic CTDs of all CFAP410 proteins were subjected to extensive crystallization trials using several commercial crystallization kits (Hampton Research). Conditions of the initial hits were further optimized to obtain single crystals. The final condition for crystallizing TbCFAP410-NTD contained 0.1 M sodium

citrate (pH 5.5) and 1 M $NH_4H_2PO_4$ at room temperature (RT). The conditions for crystallizing the CTDs of *T. brucei, H. sapiens* and *C. reinhardtii* CFAP410 were 2.0 M $(NH_4)_2SO_4$ and 5% (v/v) iso-propanol (4°C), 1.5 M NaCl and 10% (v/v) ethanol (RT), and 0.02 M $CaCl_2$, 0.1 M NaOAc (pH 5.0) and 30% (v/v) 2-Methyl-2,4-pentanediol (4°C), respectively.

All diffraction data were collected at the ESRF synchrotron site and processed using XDS [24]. Crystal structure of TbCFAP410-NTD was determined by the molecular replacement method using a homology model of TbCFAP410 (aa16-138) generated by SWISS-MODEL [16] based on the crystal structure of the *C. reinhardtii* axonemal dynein light chain-1 (PDB code: 6L4P) [17]. Crystal structure of TbCFAP410-CTD was determined *de novo* using ARCIMBOLDO [25]. Those of HsCFAP410-CTD and CrCFAP410-CTD were determined subsequently by molecular replacement using the crystal structure of TbCFAP410-CTD as search template. All resulting models were optimized by multiple rounds of manual rebuilding in COOT [26] and refinement in Phenix [27], and finally validated in MolProbity [28]. Details of data collection and refinement are summarized in **Table 1**.

Static light scattering (SLS)

SLS measurements were carried out by coupling SEC with mass determination. 50 µl of purified polypeptides at 2 mg/ml was analyzed on a Superdex S-200 10/300 GL column (Cytiva) pre-equilibrated with a buffer containing 20 mM HEPES (pH 7.5), 100 mM NaCl, 1 mM DTT, and 1% (v/v) glycerol. The polypeptides were run at a flow rate of 0.5 ml/min on a high performance liquid chromatography (HPLC) system (Agilent Technologies 1260 infinity) which was directly connected to a Mini-DAWN Treos light-scattering instrument (Wyatt Technology Corp., Santa Barbara, CA). Data analyses

were carried out using ASTRA software provided by the manufacturer. Final curves were built in SigmaPlot [29].

Circular dichroism (CD)

Far-UV CD spectra of all polypeptides between 195 and 260 nm were measured on a Chirascan plus CD spectrometer (Applied Photophysics) in a cuvette with a 0.5-mm path length. Samples were diluted to approximately 0.2 mg/ml using a buffer containing 10 mM HEPES (pH 7.5) and 100 mM NaF. Data points were corrected for buffer signal and drifts. CD curves were generated using SigmaPlot by averaging data collected from two scans for each protein sample.

Generation of constructs for *in vivo* assays in *T. brucei*

DNA sequence of TbCFAP410 was recoded in wild-type and all mutations for *in vivo* assays. The recoded DNA sequences were chemically synthesized (Twist Bioscience) and cloned into the plasmid pDEX577-mNG. All constructs were subsequently linearized with Not I and then electroporated using a BTX Gemini Twin Wave with 3 pulses of 1.7 kV for 100 µs as previously described [30]. Cells were recovered in SDM-79 for 8 hours before selection with phleomycin (5 µg/ml).

Cell culture of *T. brucei*

SmOxP9 procyclic *T. brucei* expressing TbSAS6 endogenously tagged with mScarlet were used for all experiments [31]. These were derived from TREU 927 strain, expressing T7 RNA polymerase and tetracycline repressor [32] and were grown at 28°C in

SDM-79 medium supplemented with 10% FCS [33]. Cell concentration was determined in a Z2 Coulter Counter particle counter.

Fluorescence microscopy

Cells were induced overnight with doxycycline (0.5 µg/ml) harvested by centrifugation at 800×g for 5 min. For live-cell microscopy, cells were washed two times in PBS supplemented with 10 mM glucose and 46 mM sucrose (vPBS). In the last wash, the DNA was stained using 10 µg/ml Hoechst 33342 (Sigma-Aldrich), re-suspended in vPBS and then mounted with a coverslip and immediately imaged. For cytoskeletons, cells were washed in PBS, settled onto glass slides for 5 min and treated with 1% (v/v) NP-40 in PEME for 1 min. Cytoskeletons were then fixed in −20°C methanol for 20 min and rehydrated in PBS for 10 min. After fixation, the DNA was stained with 20 µg/ml Hoechst 33342 in PBS, washed in PBS for 10 min or overnight and mounted before imaging. Images were taken using a Zeiss Axio Imager.Z1 microscope equipped with an ORCA-Flash 4.0 CMOS camera using a Plan-Apochromat 63×/1.4 NA oil objective. Images were acquired and analysed with ZEN 2 PRO software and assembled for publication in Adobe illustration CS6. For each induction 3 sets of >90 cytoskeletons were counted for each induced cell line using ImageJ (NIH, MD) and the mean and standard deviation calculated.

Molecular dynamics (MD) simulations

DSF measurements were performed on a CFX Connect Real-Time PCR machine (Bio-rad). Proteins were diluted to 0.4 mg/ml using a buffer containing 10 mM Tris-HCl pH 8.0 and 100 mM NaCl. To prepare for the measurements, 24 µl of the protein sample was

mixed with 1 µl of 20× SYBR orange dye (ThermoFisher Scientific). Each sample was prepared and measured in duplicates, and the two measurements were averaged to produce the final curve. Melting temperature (T_m) values were obtained using the software provided by the manufacturer.

Atomistic models for wild-type TbCFAP410-NTD and all mutants (V37F, C63Y, R75P, Y107C, Y107H and V111M) were constructed starting from the crystal structure by *in silico* mutagenesis using Pymol (http://www.pymol.org). The same strategy was followed with the wild-type and single-residue mutants of the tetrameric CTD structures. All further calculations were performed using GROMACS 2019 package [34] and Amber99SB-ILDN force-field parameters [35]. In all cases, the simulated system was placed into a 7×7×7-nm^3 octahedral box, energy minimized and solvated with TIP3P water molecules. Additionally, Na^+ and Cl^- ions were added to achieve electroneutrality at the final salt concentration of 150 mM. The complete system was then energy-minimized using position restraints placed on protein Cα-atoms, and subjected to 5,000 steps of NVT molecular dynamics (MD) with a 0.5-fs time step and subsequent 250,000 steps of NPT MD with a 1-fs time step. After this initial equilibration, a 1 µs NPT production MD run was simulated (5×10^8 steps with a 2-fs time step) for each system. All analysis was performed on the final 750 ns of each simulated trajectory. A twin-range (10 Å, 10.5 Å) spherical cut-off function was used to truncate the van der Waals interactions. Electrostatic interactions were treated using the particle-mesh Ewald summation (real space cut-off of 10-Å and 1.2-Å grid with fourth-order spline interpolation). MD simulations were carried out using 3D periodic boundary conditions, in the isothermal−isobaric (NPT) ensemble with an isotropic pressure of 1.013 bar and a constant temperature of 295 K. The pressure and temperature were controlled using a Nose-Hoover thermostat [36, 37] and a Parrinello–Rahman barostat with 0.5-ps and 10-ps relaxation parameters, respectively, and a compressibility of 4.5×10^{-5} bar^{-1} for the barostat.

Accession code

Coordinates and structure factors have been deposited in the Protein Data Bank (PDB) under accession codes 8AXJ, 8AXO, 8AXQ, and 8AXR.

Acknowledgements

The authors are grateful to the staff at the beamlines ID23-1 and ID23-2 at the European Synchrotron Radiation Facility (ESRF) for their help with X-ray diffraction data collection. We also thank Dr. A. Burga for helping us order all synthetic polypeptides. This work was supported by the grant I5960-B2 from the Austrian Science Fund (FWF) to GD and a Newton International Fellowship (NIF\R1\191618) to HG. SA acknowledges the VIP2 fellowship 2020 funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 847548.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Krohn, K., et al., *Immunochemical characterization of a novel mitochondrially located protein encoded by a nuclear gene within the DFNB8/10 critical region on 21q22.3.* Biochem Biophys Res Commun, 1997. **238**(3): p. 806-10.
2. Abu-Safieh, L., et al., *Autozygome-guided exome sequencing in retinal dystrophy patients reveals pathogenetic mutations and novel candidate disease genes.* Genome Res, 2013. **23**(2): p. 236-47.
3. Scott, H.S., et al., *Characterization of a novel gene, C21orf2, on human chromosome 21q22.3 and its exclusion as the APECED gene by mutation analysis.* Genomics, 1998. **47**(1): p. 64-70.
4. Wheway, G., et al., *An siRNA-based functional genomics screen for the identification of regulators of ciliogenesis and ciliopathy genes.* Nat Cell Biol, 2015. **17**(8): p. 1074-1087.
5. Wang, Z., et al., *Axial spondylometaphyseal dysplasia is also caused by NEK1 mutations.* J Hum Genet, 2017. **62**(4): p. 503-506.
6. Fang, X., et al., *The NEK1 interactor, C21ORF2, is required for efficient DNA damage repair.* Acta Biochim Biophys Sin (Shanghai), 2015. **47**(10): p. 834-41.
7. Chia, R., A. Chio, and B.J. Traynor, *Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications.* Lancet Neurol, 2018. **17**(1): p. 94-102.
8. Shim, K.S., et al., *Reduction of chromatin assembly factor 1 p60 and C21orf2 protein, encoded on chromosome 21, in Down syndrome brain.* J Neural Transm Suppl, 2003(67): p. 117-28.
9. Ma, X., R. Peterson, and J. Turnbull, *Adenylyl cyclase type 3, a marker of primary cilia, is reduced in primary cell culture and in lumbar spinal cord in situ in G93A SOD1 mice.* BMC Neurosci, 2011. **12**: p. 71.
10. Ehara, S., et al., *Axial spondylometaphyseal dysplasia.* Eur J Pediatr, 1997. **156**(8): p. 627-30.
11. Warman, M.L., et al., *Nosology and classification of genetic skeletal disorders: 2010 revision.* Am J Med Genet A, 2011. **155A**(5): p. 943-68.
12. Wang, Z., et al., *Axial Spondylometaphyseal Dysplasia Is Caused by C21orf2 Mutations.* PLoS One, 2016. **11**(3): p. e0150555.
13. McInerney-Leo, A.M., et al., *Homozygous variant in C21orf2 in a case of Jeune syndrome with severe thoracic involvement: Extending the phenotypic spectrum.* Am J Med Genet A, 2017. **173**(6): p. 1698-1704.
14. Khan, A.O., et al., *C21orf2 is mutated in recessive early-onset retinal dystrophy with macular staphyloma and encodes a protein that localises to the photoreceptor primary cilium.* Br J Ophthalmol, 2015. **99**(12): p. 1725-31.
15. Suga, A., et al., *Identification of Novel Mutations in the LRR-Cap Domain of C21orf2 in Japanese Patients With Retinitis Pigmentosa and Cone-Rod Dystrophy.* Invest Ophthalmol Vis Sci, 2016. **57**(10): p. 4255-63.
16. Waterhouse, A., et al., *SWISS-MODEL: homology modelling of protein structures and complexes.* Nucleic Acids Res, 2018. **46**(W1): p. W296-W303.
17. Toda, A., et al., *The complex of outer-arm dynein light chain-1 and the microtubule-binding domain of the gamma heavy chain shows how axonemal dynein tunes ciliary beating.* J Biol Chem, 2020. **295**(12): p. 3982-3989.
18. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
19. Dean, S., J.D. Sunter, and R.J. Wheeler, *TrypTag.org: A Trypanosome Genome-wide Protein Localisation Resource.* Trends Parasitol, 2017. **33**(2): p. 80-82.
20. Brett, D., et al., *Alternative splicing and genome complexity.* Nat Genet, 2002. **30**(1): p. 29-30.
21. Kpebe, A. and L. Rabinow, *Alternative promoter usage generates multiple evolutionarily conserved isoforms of Drosophila DOA kinase.* Genesis, 2008. **46**(3): p. 132-43.
22. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
23. Tyagi, R., R. Lai, and R.G. Duggleby, *A new approach to 'megaprimer' polymerase chain reaction mutagenesis without an intermediate gel purification step.* BMC Biotechnol, 2004. **4**: p. 2.
24. Kabsch, W., *Xds.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 125-32.

25. Rodriguez, D.D., et al., *Crystallographic ab initio protein structure solution below atomic resolution.* Nat Methods, 2009. **6**(9): p. 651-3.
26. Emsley, P. and K. Cowtan, *Coot: model-building tools for molecular graphics.* Acta Crystallogr D Biol Crystallogr, 2004. **60**(Pt 12 Pt 1): p. 2126-32.
27. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 213-21.
28. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 1): p. 12-21.
29. Kornbrot, D., *Statistical software for microcomputers: SigmaPlot 2000 and SigmaStat2.* Br J Math Stat Psychol, 2000. **53 (Pt 2)**: p. 335-7.
30. Dean, S., et al., *A toolkit enabling efficient, scalable and reproducible gene tagging in trypanosomatids.* Open Biol, 2015. **5**(1): p. 140197.
31. Atkins, M., et al., *CEP164C regulates flagellum length in stable flagella.* J Cell Biol, 2021. **220**(1).
32. Poon, S.K., et al., *A modular and optimized single marker system for generating Trypanosoma brucei cell lines expressing T7 RNA polymerase and the tetracycline repressor.* Open Biol, 2012. **2**(2): p. 110037.
33. Brun, R. and Schonenberger, *Cultivation and in vitro cloning or procyclic culture forms of Trypanosoma brucei in a semi-defined medium. Short communication.* Acta Trop, 1979. **36**(3): p. 289-92.
34. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. **26**(16): p. 1701-18.
35. Lindorff-Larsen, K., et al., *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins, 2010. **78**(8): p. 1950-8.
36. Nose, S., *A Unified Formulation of the Constant Temperature Molecular-Dynamics Methods.* Journal of Chemical Physics, 1984. **81**(1): p. 511-519.
37. Hoover, W.G., *Canonical Dynamics - Equilibrium Phase-Space Distributions.* Physical Review A, 1985. **31**(3): p. 1695-1697.

Figure legends

Figure 1. CFAP410-CTD forms a highly conserved tetrameric helical bundle. (**A-C**) Crystal structures of the CTDs of *H. sapiens*, *C. reinhardtii* and *T. brucei* CFAP410. Two orthogonal views are shown in each case. The four copies in each structure are colored differently with their N- and C-terminal ends labeled. (**D**) Superposition of the three tetrameric structures shown in A-C. (**E**) Superposition of all unique monomeric structures from above three tetramers. (**F**) Sequence alignment of the CTDs of *H. sapiens*, *C. reinhardtii* and *T. brucei* CFAP410 with residue ranges shown at both ends. The two α-helices of the crystal structures are shown above the alignments. The alignments were carried out using the option of "Tcoffee with defaults" in Jalview, with residues highlighted using the Clustal color scheme.

Figure 2. Mutational studies of CFAP410-CTDs. (**A**) Sequence alignment of the CTDs of three CFAP410 proteins. Boxes under the alignments represent the two α helices in the crystal structure. Residues involved in intra- and inter-dimer interactions are denoted by asterisks and dots, respectively. Green arrow indicates the conserved alanine located at the center of the dimer-dimer interface; red arrow indicates the conserved disease-causing Leu residue. (**B**) Crystal structures of the three CFAP410-CTDs. All helices are shown as cylinders with the conserved Ala and Leu residues shown as sticks. (**C**) Two orthogonal views of the TbCFAP410-CTD dimer with one chain shown as an electrostatic surface plot and the other as ribbons. The conserved Leu residue (L272) is deeply buried between the two chains. (**D**) An open-book view of the TbCFAP410-CTD tetramer shown in (B), with the dimer on the left shown as an electrostatic surface plot and the right one as ribbons. Marked by green circles (left) and shown as sticks (right) are the four conserved Ala residues (A267) situated at the center of the inter-dimer interface. (**E**) SLS results of wild-type (WT), A219E and L224P of CrCFAP410-CTD. (**F**) SEC profiles of WT and mutants of the CTDs of all three CFAP410 proteins. (**G**) CD spectra of WT, A219E and L224P of HsCFAP410-CTD. (**H**) CD spectra of WT and Leu mutants of the CTDs of all three CFAP410 proteins.

Figure 3. Crystal structure of the N-terminal LRR motif of TbCFAP410. (**A**) Ribbon diagram of TbCFAP410-NTD (aa1-160) color-ramped from blue (N-terminus) to red (C-terminus). (**B**) Structure of TbCFAP410-NTD with its C-terminal loop (aa141-160)

shown as sticks together with $2F_o$-F_c electron density map (green mesh) contoured at 1.5 σ level. (**C**) A conserved surface patch on the β sheet of the LRR motif. Solvent-exposed surface residues are colored in blue, marine and light blue based on their sequence identities. (**D**) The opposite side of the LRR motif covered by the loop is devoid of conserved residues. An electrostatic surface plot in the same orientation is shown at the right side in both (C) and (D).

Figure 4. Structural environment and expression test of the ciliopathy mutations in TbCFAP410-NTD. (**A**) Crystal structure of TbCFAP410-NTD with residues corresponding to all known ciliopathy mutations in HsCFAP410 shown as sticks and labeled. All other side-chains are shown as lines. (**B**) Front and back views of TbCFAP410-NTD shown as semi-transparent surface and ribbons. Only three of the five residues in (A) that are fully or partially exposed to solvent are labeled. (**C**-**E**) The five ciliopathy mutation sites are grouped into three categories: fully exposed (C), semi-buried (D), and fully buried (E). (**F**) Solubility test of wild-type and mutants of TbCFAP410-NTD upon over-expression in *E. coli*. Expression with only the vector is included as a negative control.

Figure 5. Fluorescence microscopy images from wild-type and mutants of TbCFAP410. (**A**) Images acquired from live cells or methanol fixed cytoskeletons from doxycycline induced cells expressing TbCFAP410 wild-type or TbCFAP410 mutants (green) with mSc::TbSAS6 - basal body marker (magenta). Scale bar: 5 µm. (**B**) Conservation of the conserved patch on the surface of TbCFAP410-NTD in the same orientation as in Fig. 3C. The six residues in the core of the patch that were selected for mutagenesis analyses are labeled. (**C**) Histogram of TbCFAP410 signal counts. The percentage of cytoskeletons with a basal body and posterior signal, a basal body only signal and a basal body with a very weak posterior signal was calculated for each cell line and the mean ± s.d. was plotted. Induction was repeated twice, with a similar result both times and data from one is shown. *BB: basal body.*

Figure 6. Structural examination of how the disease-causing mutants on TbCFAP410-NTD may affect protein folding or structural stability. (**A**) Residue R75 is in a rigid linker and forms hydrogen bonds with residues from the neighboring linker. Its mutation to Pro would disrupt the all hydrogen bonds and cause severe clashes.

(**B**) Residue V37 is located in a shallow cleft. Its replacement by Phe would clash with the surrounding residues. (**C**) Residue Y107 is in a semi-open eggshell-like pocket and the hydroxyl group on its side-chain forms a hydrogen bond with a neighboring residue. Both Y107C and Y107H mutants would disrupt this hydrogen bond and generate a cavity or clash with other residues surrounding the pocket. (**D**, **E**) Residues C63 and V111 are both buried deeply in the hydrophobic core of the structure. Their replacements by residues with bulkier side-chains (i.e. Y and M) would cause sever clashes with the surrounding residues. All images were generated using Pymol (http://www.pymol.org).

Figure 7. Simulation analyses of TbCFAP410-NTD mutants. (**A**) Conformational analysis of the TbCFP410 C63Y mutant simulations. *Left:* central configuration of the most populated cluster in the TbCFP410-NTD C63Y simulations (cluster occupancy 21%); *Right:* central configuration of the second most populated cluster in the C63Y simulations showing the large opening of the C-terminal cleft, which coincides with the formation of the E61-K156 salt bridge. (**B**) Simulation analyses of wild type and mutant TbCFP410-NTD with the number of conformational clusters (with cluster occupancy P > 1%). The table also shows cluster occupancy of the most populated cluster (P_{MPC}), all-atom RMSD between the central structures of the most populated clusters in mutant and wild type simulations, change in conformational entropy upon mutation ($\Delta S_{conf} = S_{mutant} - S_{wt}$) whereby S_{mutant} was rescaled by the relative number of degrees of freedom, and change in the total solvent-accessible surface area ($\Delta SASA_T$) and its hydrophilic ($\Delta SASA_{Hphi}$) and hydrophobic components ($\Delta SASA_{Hpho}$) upon mutation ($\Delta SASA = SASA_{mutant} - SASA_{wt}$).

Figure 8. Bimodular organization of CFAP410. CFAP410-CTD forms tetrameric helical bundle, whereas CFAP410-NTD folds into a monomeric LRR motif. The two domains are loosely interconnected via a long disordered loop. There is a conserved surface patch (colored in blue) on CFAP410-NTD that likely plays an important role in the *in vivo* function of the protein. The insert is a tetrameric structural model generated by Alphafold2, with each molecule color ramped from blue at the N-terminus to red at the C-terminus.

Table 1. Data collection and refinement statistics

	TbCFAP410-NTD	TbCFAP410-CTD	HsCFAP410-CTD	CrCFAP410-CTD
Wavelength (Å)	0.972	0.972	0.873	0.972
Resolution range (Å)	19.90 - 1.0 (1.04 - 1.0)	19.10 - 1.29 (1.34 - 1.29)	19.94 - 1.50 (1.55 - 1.50)	19.56 - 1.40 (1.45 - 1.40)
Space group	P 2_1 2_1 2_1	P 3_2 2 1	C 1 2 1	P 2_1 2_1 2_1
Unit cell *a, b, c* (Å) α, β, γ (°)	39.80, 52.96, 61.43 90, 90, 90	40.11, 40.11, 62.68 90, 90, 120	56.12, 50.63, 22.15 90, 101.08, 90	43.44, 48.10, 53.16 90, 90, 90
Total reflections	789,670 (30,477)	141,666 (12,689)	60,771 (5,760)	275,741 (19,011)
Unique reflections	69,533 (5,835)	15,025 (1,430)	9,761 (948)	22,140 (1,924)
Multiplicity	11.4 (5.2)	9.4 (8.9)	6.2 (6.1)	12.5 (9.9)
Completeness (%)	98.17 (83.75)	99.65 (97.15)	99.60 (98.13)	98.19 (87.57)
Mean I/sigma(I)	10.84 (0.61)	8.16 (0.76)	11.05 (0.99)	9.40 (1.11)
R-merge	0.099 (1.695)	0.132 (2.047)	0.077 (1.477)	0.136 (1.019)
R-meas	0.104 (1.882)	0.140 (2.173)	0.084 (1.616)	0.142 (1.074)
R-pim	0.030 (0.792)	0.046 (0.717)	0.0335 (0.646)	0.040 (0.331)
$CC_{1/2}$	0.998 (0.386)	0.991 (0.43)	0.999 (0.519)	0.995 (0.747)
CC*	1.00 (0.746)	0.998 (0.775)	1.00 (0.827)	0.999 (0.925)
Reflections used in refinement	69,519 (5,834)	15,021 (1,430)	9,750 (946)	22,131 (1,924)
Reflections used for R-free	1,989 (168)	1,516 (141)	977 (95)	1,988 (173)
R-work	0.152	0.187	0.213	0.165
R-free	0.162	0.204	0.249	0.190
Number of heavy atoms in proteins	1,302	505	452	1,084
Number of solvent	157	33	28	91
RMS (bonds)	0.010	0.005	0.006	0.009
RMS (angles)	1.22	0.62	0.93	0.93
Ramachandran favored (%) allowed (%) outliers (%)	96.23 3.77 0.00	100.00 0.00 0.00	100.00 0.00 0.00	98.40 1.60 0.00

Statistics for the highest-resolution shell are shown in parentheses.

Figure 1

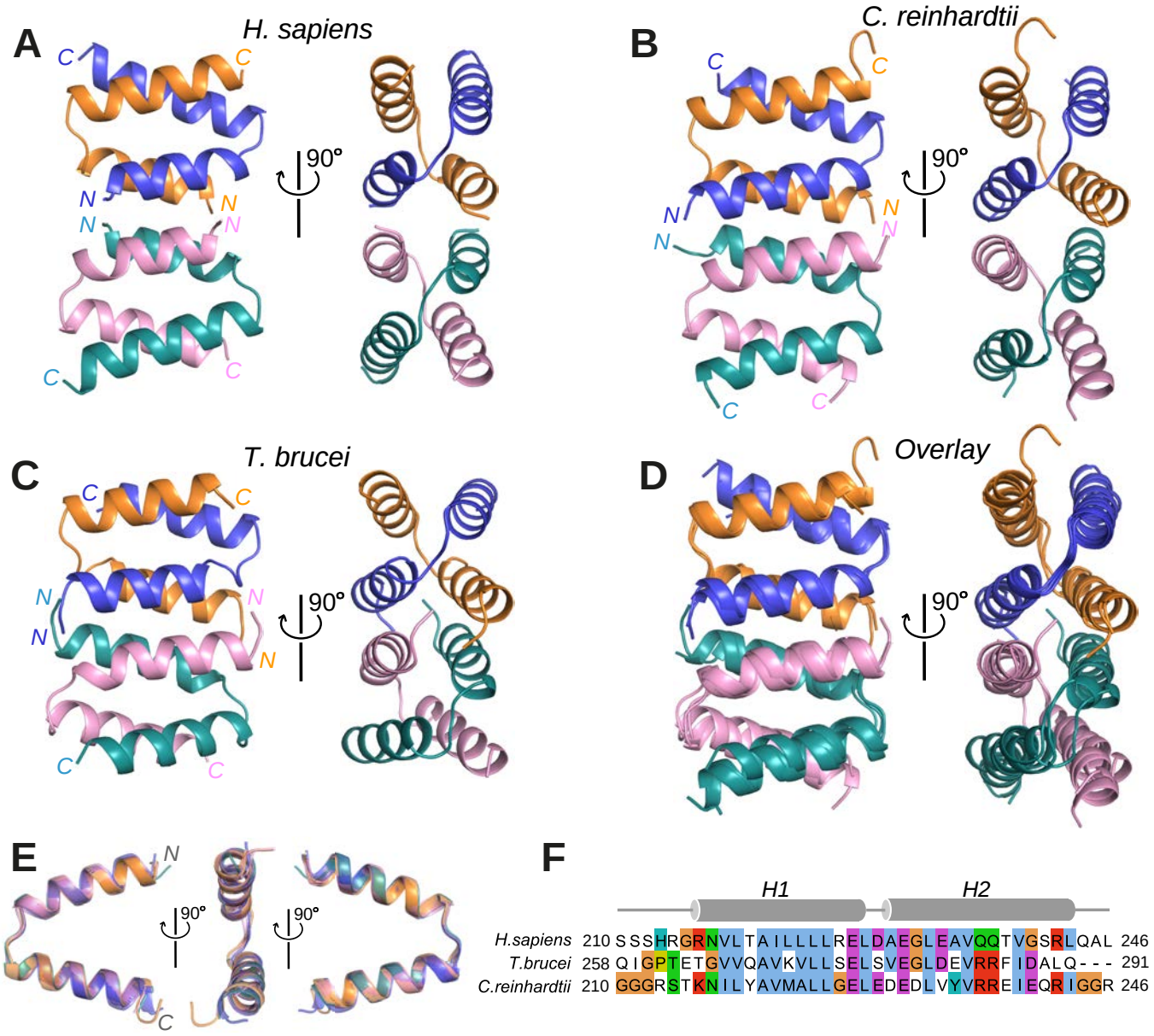

A
H. sapiens
B
C. reinhardtii
C
T. brucei
D
Overlay
E
F
H1
H2
90°
N
C
H.sapiens 210 SSSHRGRNVLTAILLLLRELDAEGLEAVQQTVGSRLQAL 246
T.brucei 258 QIGPTETGVVQAVKVLLSELSVEGLDEVRRFIDALQ--- 291
C.reinhardtii 210 GGGRSTKNILYAVMALLGELEDEDLVYVRREIEQRIGGR 246

Figure 2

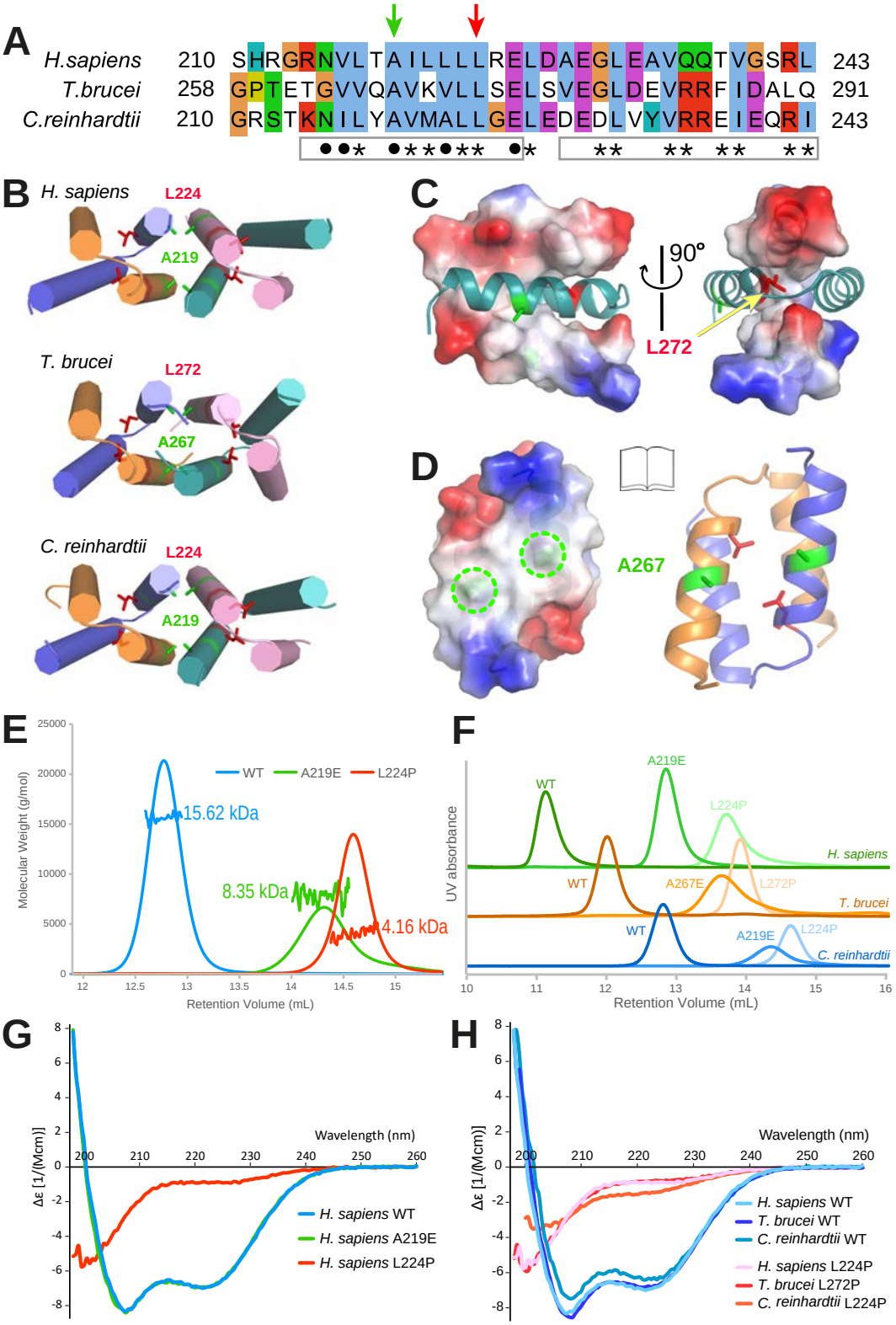

A
H.sapiens 210 SHRGRNVLTAILLLLRELDAEGLEAVQQTVGSRL 243
T.brucei 258 GPTETGVVQAVKVLLSELSVEGLDEVRRFIDALQ 291
C.reinhardtii 210 GRSTKNILYAVMALLGELEDEDLVYVRREIEQRI 243
B
H. sapiens
L224
A219
T. brucei
L272
A267
C. reinhardtii
L224
A219
C
90°
L272
D
A267
E
Molecular Weight (g/mol)
WT
A219E
L224P
15.62 kDa
8.35 kDa
4.16 kDa
Retention Volume (mL)
F
UV absorbance
WT
A219E
L224P
H. sapiens
WT
A267E
L272P
T. brucei
WT
A219E
L224P
C. reinhardtii
Retention Volume (mL)
G
Δε [1/(Mcm)]
Wavelength (nm)
H. sapiens WT
H. sapiens A219E
H. sapiens L224P
H
Δε [1/(Mcm)]
Wavelength (nm)
H. sapiens WT
T. brucei WT
C. reinhardtii WT
H. sapiens L224P
T. brucei L272P
C. reinhardtii L224P

Figure 3

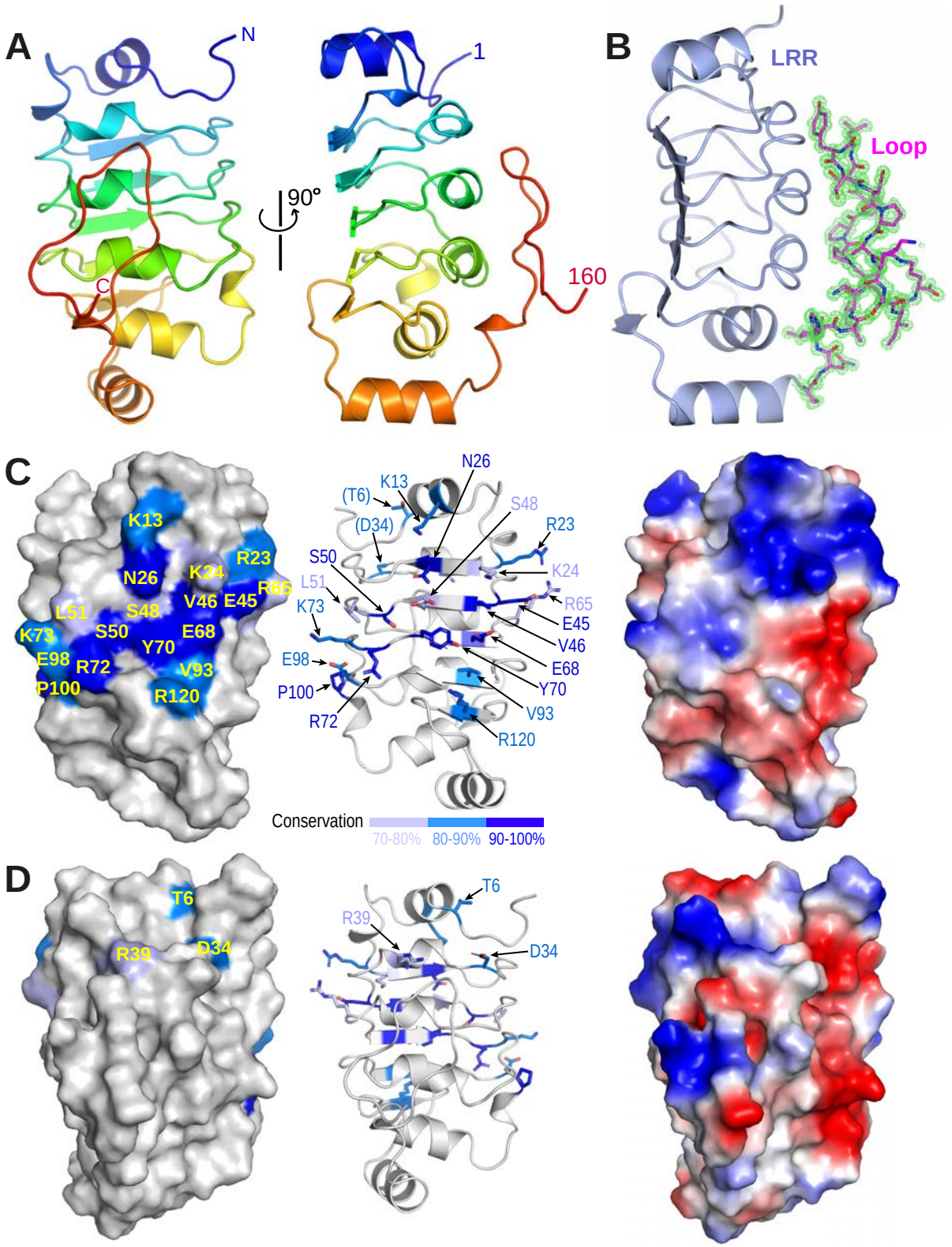

A
N
C
90°
1
160
B
LRR
Loop
C
K13
N26
K24
R23
R65
V46
E45
L51
S48
K73
S50
E68
Y70
E98
R72
V93
P100
R120
(T6)
K13
N26
(D34)
S48
R23
S50
K24
L51
R65
K73
E45
V46
E68
E98
Y70
P100
V93
R72
R120
Conservation
70-80%
80-90%
90-100%
D
T6
R39
D34
T6
R39
D34

Figure 4

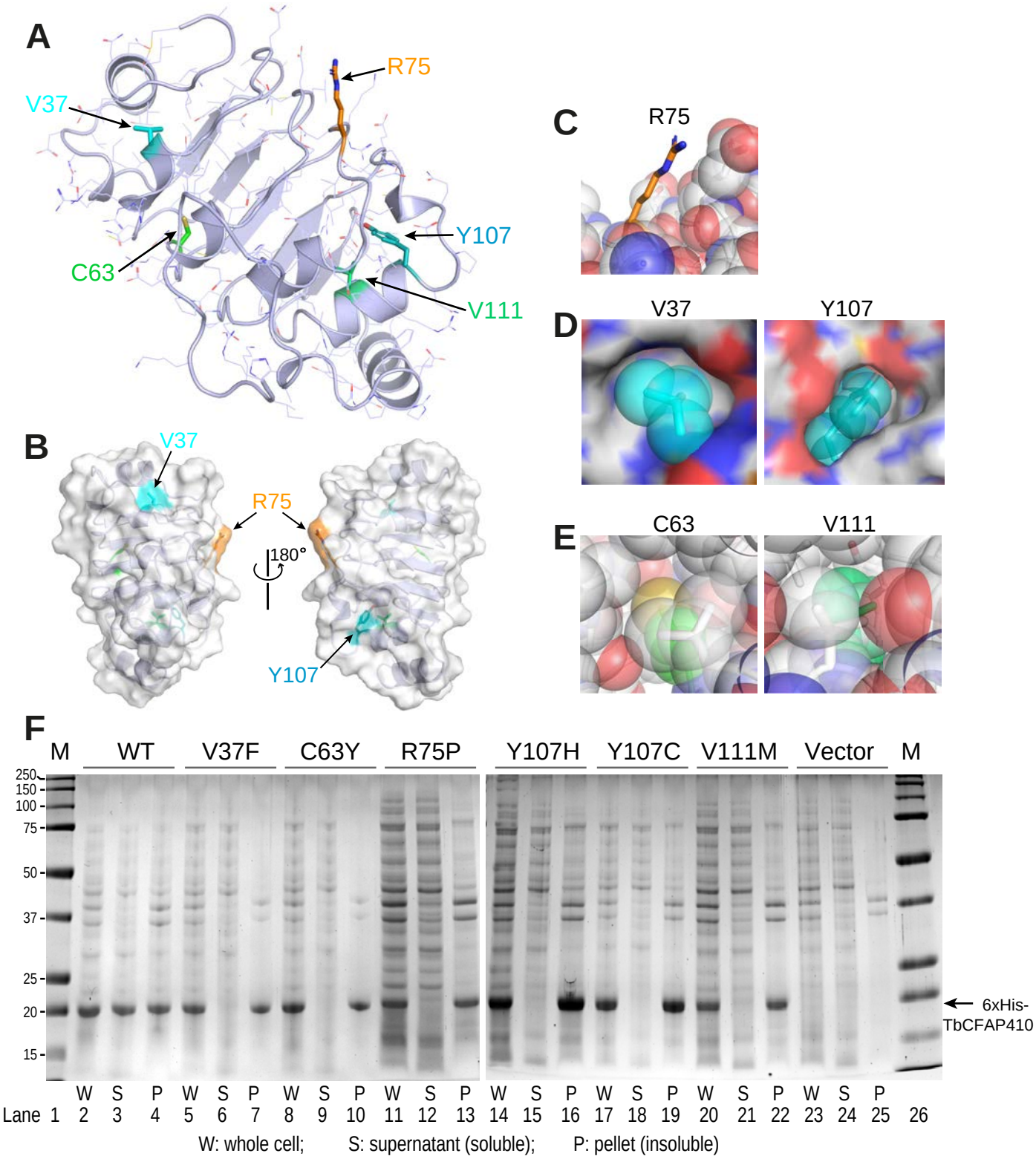

A
R75
V37
Y107
C63
V111
B
V37
R75
180°
Y107
C
R75
D
V37
Y107
E
C63
V111
F
M
WT
V37F
C63Y
R75P
Y107H
Y107C
V111M
Vector
M
250
150
100
75
50
37
25
20
15
6xHis-
TbCFAP410
Lane 1 W 2 S 3 P 4 W 5 S 6 P 7 W 8 S 9 P 10 W 11 S 12 P 13 W 14 S 15 P 16 W 17 S 18 P 19 W 20 S 21 P 22 W 23 S 24 P 25 26
W: whole cell; S: supernatant (soluble); P: pellet (insoluble)

Figure 5

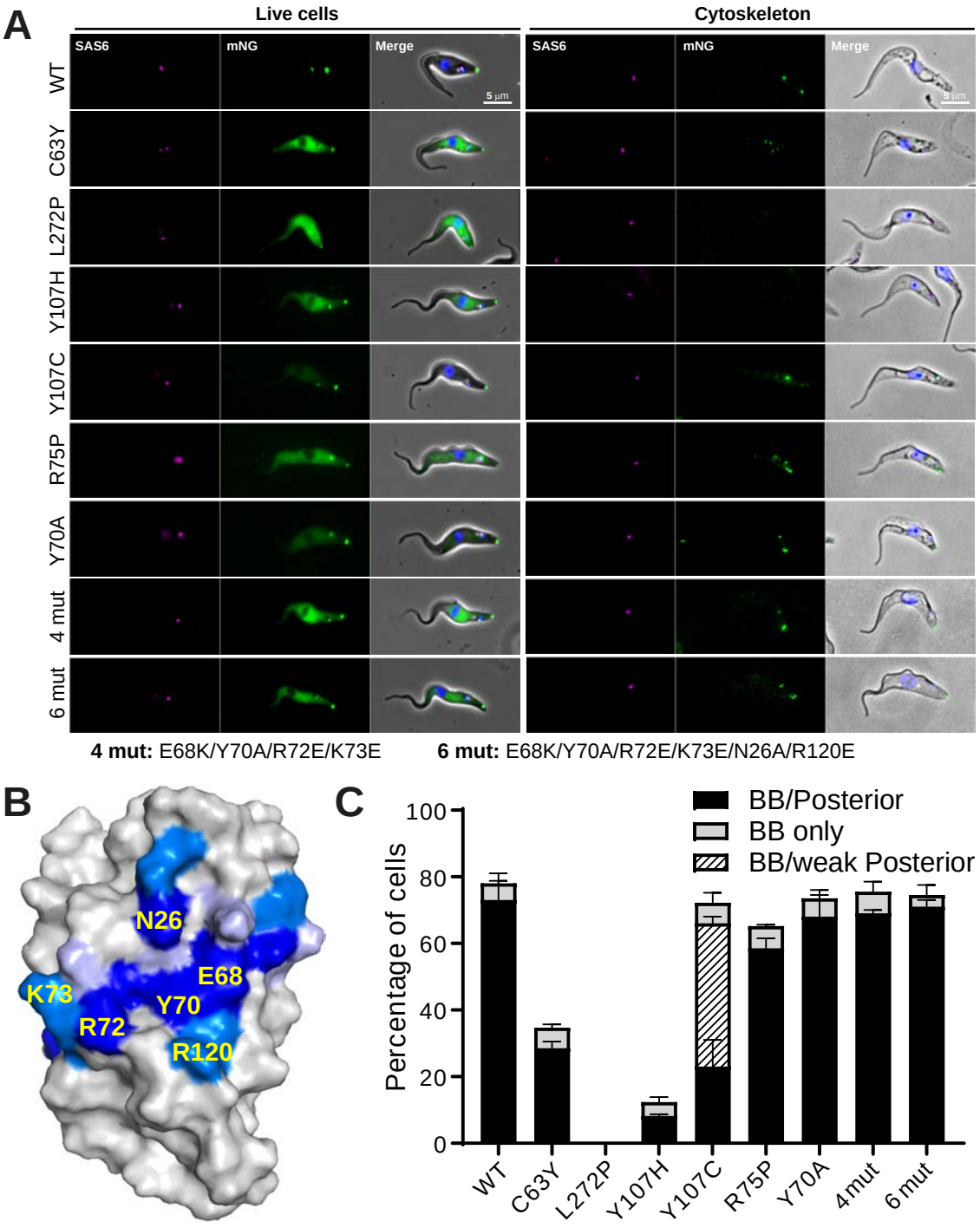

A
Live cells
Cytoskeleton
SAS6
mNG
Merge
5 µm
WT
C63Y
L272P
Y107H
Y107C
R75P
Y70A
4 mut
6 mut
4 mut: E68K/Y70A/R72E/K73E
6 mut: E68K/Y70A/R72E/K73E/N26A/R120E
B
N26
E68
K73
Y70
R72
R120
C
BB/Posterior
BB only
BB/weak Posterior
Percentage of cells
100
80
60
40
20
0
WT
C63Y
L272P
Y107H
Y107C
R75P
Y70A
4mut
6mut

Figure 6

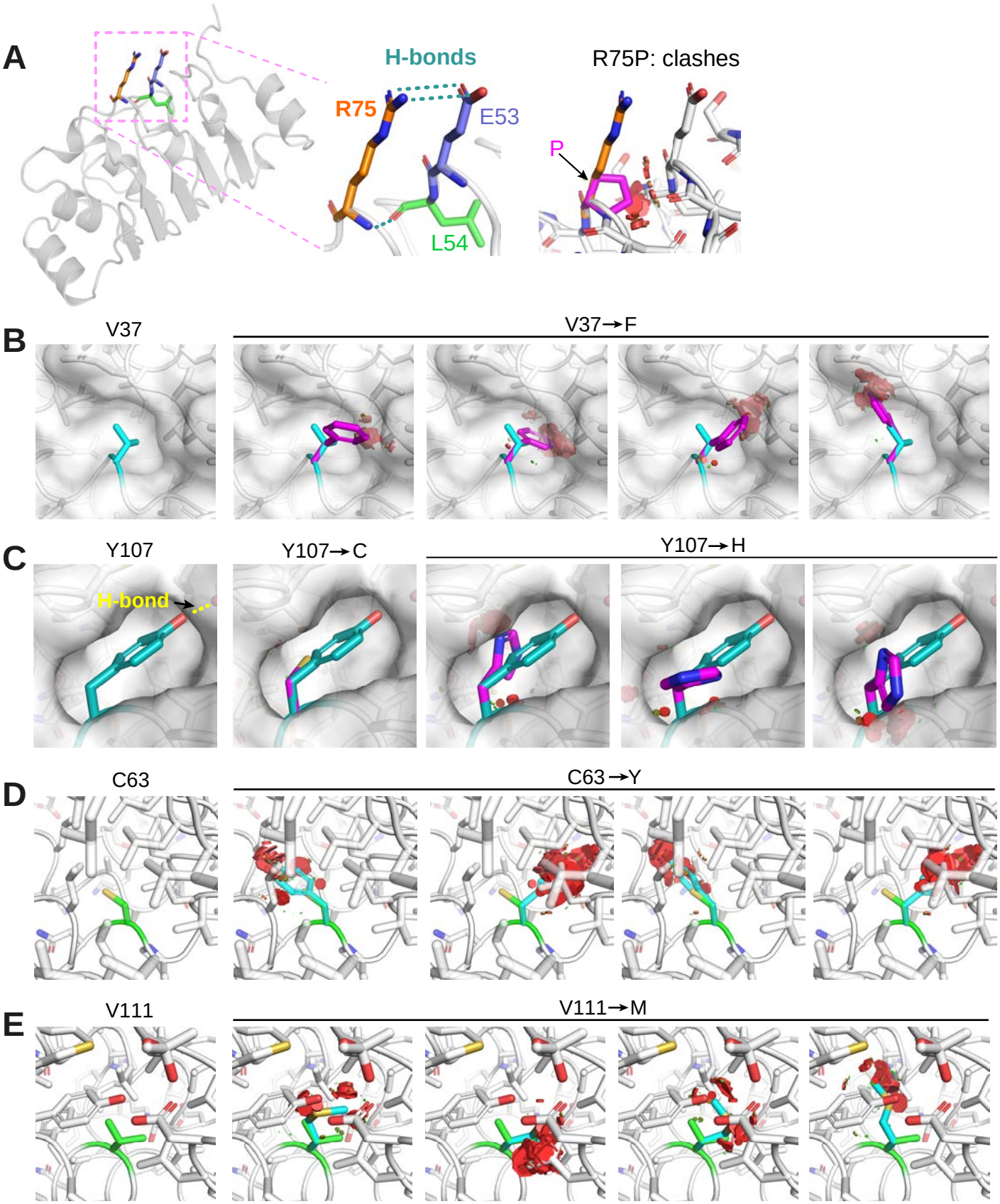
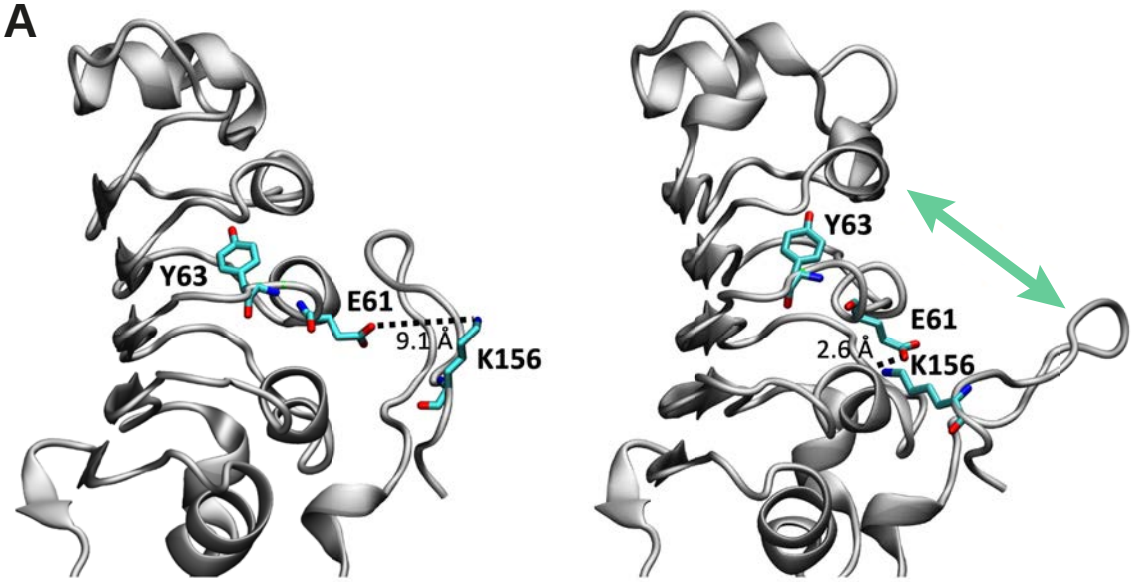

A
H-bonds
R75
E53
L54
R75P: clashes
P
B
V37
V37→F
C
Y107
H-bond
Y107→C
Y107→H
D
C63
C63→Y
E
V111
V111→M

Figure 7

A



C63Y closed (21%)
RMSD vs WT: **2.0 Å**

C63Y open (17%)
RMSD vs WT: **3.1 Å**

B

system	# clusters (P > 1%)	P_{MPC}	RMSD	ΔS_{conf}	$\Delta SASA_T$	$\Delta SASA_{Hphi}$	$\Delta SASA_{Hpho}$
WT	6	59.1	-	-	-	-	-
V37F	1	99.9	0.9	-34.1	-1.0	0.0	-1.0
C63Y	10	20.5	2.0	15.4	2.0	0.5	1.5
R75P	5	59.2	1.1	8.1	-1.0	-1.5	0.5
Y107C	8	31.3	1.1	25.7	0.5	0.3	0.2
Y107H	7	76.0	1.1	3.2	0.5	0.5	0.0
V111M	2	94.3	0.9	2.9	0.0	0.0	0.0
		%	Å	J/mol K	nm^2		

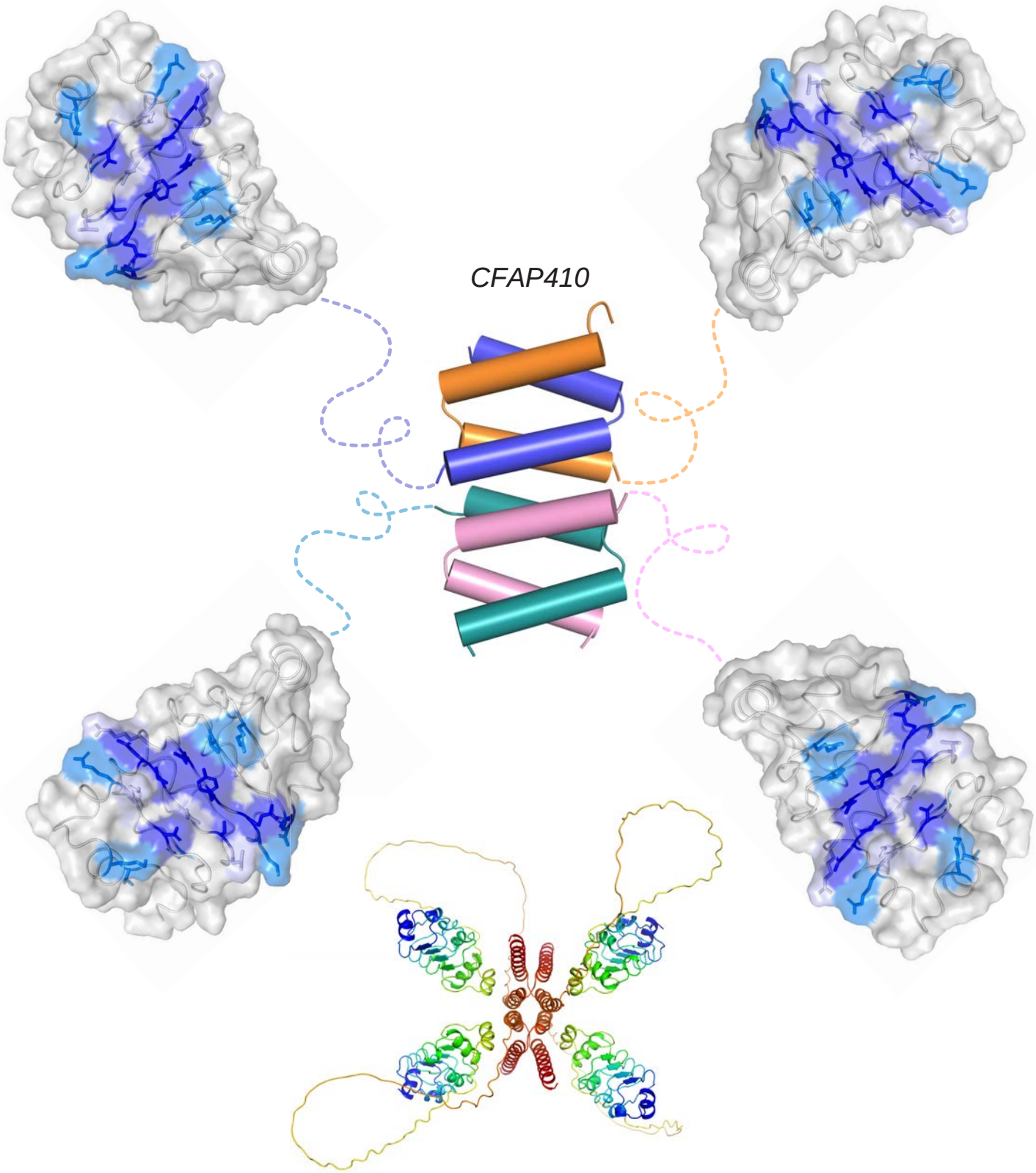



Figure 8