# Combining *in Vitro* and *in Silico* Single-Molecule Force Spectroscopy to Characterize and Tune Cellulosomal Scaffoldin Mechanics

Tobias Verdorfer[1], Rafael C. Bernardi[2], Aylin Meinhold[1], Wolfgang Ott[1], Zaida Luthey-Schulten[2,3], Michael A. Nash*[4,5] and Hermann E. Gaub[1]

[1] Lehrstuhl für Angewandte Physik and Center for Nanoscience, Ludwig-Maximilians-Universität, 80799 Munich, Germany.
[2] Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States.
[3] Department of Chemistry, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States.
[4] Department of Chemistry, University of Basel, 4056 Basel, Switzerland.
[5] Department of Biosystems Science and Engineering, Swiss Federal Institute of Technology (ETH Zurich), 4058 Basel, Switzerland.

*Correspondence and requests for materials should be addressed to MAN (email: michael.nash@unibas.ch).

# Abstract

Cellulosomes are poly-protein machineries that efficiently degrade cellulosic material. Crucial to their function are scaffolds consisting of highly homologous cohesin domains, which serve a dual role by coordinating a multiplicity of enzymes as well as anchoring the microbe to its substrate. Here we combined two approaches to elucidate the mechanical properties of the main scaffold ScaA of *Acetivibrio cellulolyticus*. A newly developed parallelized one-pot *in vitro* transcription-translation and protein pulldown protocol enabled high-throughput atomic force microscopy (AFM)-based single-molecule force spectroscopy (SMFS) measurements of all cohesins from ScaA with a single cantilever, thus promising improved relative force comparability. Albeit very similar in sequence, the hanging cohesins showed considerably lower unfolding forces than the bridging cohesins, which are subjected to force when the microbe is anchored to its substrate. Additionally, all-atom steered molecular dynamics (SMD) simulations on homology models offered insight into the process of cohesin unfolding under force. Based on the differences among the individual force propagation pathways and their associated correlation communities, we designed mutants to tune the mechanical stability of the weakest hanging cohesin. The proposed mutants were tested in a second high-throughput AFM SMFS experiment revealing that in one case a single alanine to glycine point mutation suffices to more than double the mechanical stability. In summary, we have successfully characterized the force induced unfolding behaviour of all cohesins from the scaffoldin ScaA, as well as revealed how small changes in sequence can have large effects on force resilience in cohesin domains. Our strategy provides an efficient way to test and improve the mechanical integrity of protein domains in general.

# Introduction

Multidomain protein scaffolds organize cellulolytic enzymes and provide adhesion between the host cell and its substrate. In cellulosomes, these so-called scaffoldins utilize various orthogonal high-affinity receptor-ligand interactions between cohesins and dockerins to anchor themselves to the cell's exterior membrane and to coordinate a broad arsenal of cellulolytic enzymes.[1–3] Cellulosomes are extracellular polyprotein complexes produced by many microorganisms for the efficient degradation of cellulose and hemicellulose, two of nature's most abundant polymers. In addition to multiple repeats of cohesin domains, scaffoldins may also contain other ancillary domains, such as enzymatic subunits or carbohydrate-binding modules (CBMs).[4] Cellulosomes represent an advantageous strategy compared to the secretion of freely soluble enzymes for cellulose degradation, because they achieve both robust adhesion to the substrate as well as synergistic and cooperative interplay between the enzymes. This complex synergy is based on the spatial organization and enhanced local concentration. Due to their remarkable ability to achieve these complex tasks, cellulosomes have become a prime instructive example of molecular systems offering modularity, self-assembly, and highly efficient enzymatic catalysis.[5] Aside from the unique role they play in the degradation of cellulolytic material, cellulosomal scaffoldins are especially interesting as building blocks in a biotechnological toolbox. Cohesin-dockerin interactions with orthogonal specificities can be used to post-translationally incorporate selected enzymes or other auxiliary domains in specific locations by self-assembly within rationally designed multi-component complexes.[6–8]

In this study we focused on a scaffoldin of the cellulosome-producing organism *Acetivibrio cellulolyticus*, a gram-negative, anaerobic bacterium found in mechanically demanding environments, such as sewage sludge[9] or the bovine rumen.[10] Although *A. cellulolyticus* expresses an exceptionally branched and diverse cellulosome[11], the assembly of the majority of its cellulolytic enzymes relies mainly on a single scaffoldin (ScaA)[12] (Figure 1). Simultaneously, ScaA has the essential function of targeting specific substrates and anchoring the cell to the cellulose fibril via its single CBM. ScaA therefore serves a dual purpose for the organism by incorporating catalytic modules into the cellulosome complex and by adhering the bacterium to the cellulosic substrate, making it the key player in *A. cellulolyticus* cellulosome.
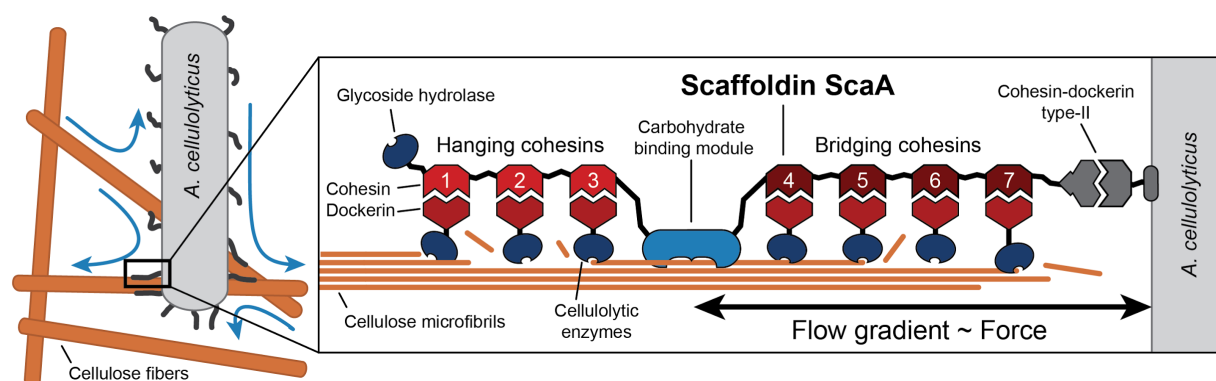


**Figure 1: Simplified schematics of the cellulosome of *A. cellulolyticus*.** The scaffoldin ScaA binds up to 7 cellulolytic enzymes via cohesin-dockerin interactions and attaches the multi-enzyme complex to the cellulose microfibril through a CBM module, ensuring close proximity of the enzymes and substrate. Simultaneously, ScaA anchors the assembled cellulosome to the cell wall of the microbe via a type-II cohesin-dockerin complex. The bridging cohesins 4-7 in particular have to withstand the forces between the cellulose fibril and microbe caused by flow gradients in the environment.

ScaA consists of an N-terminal glycoside hydrolase domain, seven type-I cohesins, a CBM located between cohesins 3 and 4, and a C-terminal type-II dockerin domain. Cohesins 4 through 7 are located between the CBM and the anchoring type-II Doc, and form the so-called 'bridging' region of the scaffoldin, while cohesins 1 through 3 are located exterior from the CBM and form the 'hanging' region of this scaffoldin. It had been hypothesized that the

cohesins of the bridging region will be subjected to higher mechanical stress compared to the hanging cohesins because the CBM anchors the cell to the cellulose substrate and the domains located within the hanging region are not expected to be mechanically stressed *in vivo.*[13] The organism may have adapted the bridging cohesins to these conditions and therefore, they should be able to withstand higher mechanical stress in order to maintain a folded structure.

This hypothesis of bridging vs. hanging cohesins within cellulosomal scaffoldins has previously been successfully tested in part for the CipA scaffoldin of *Clostridium thermocellum*. These prior results showed that cohesins from the bridging region withstood higher unfolding forces compared to those from the hanging region.[13] This behavior has been attributed to slight differences in the stability of the mechanical clamp motifs, which are structural elements formed by backbone hydrogen bonds between parallel β-strands of the N- and C-terminal ends of the cohesin protein domains.

Since the mechanobiology of cellulosomes is key to their function, we investigated the mechanostability of cohesins of the ScaA scaffoldin as well as several derived mutants at the single molecular level. The high sequence similarity between the selected cohesins raises the question on how differently these cohesins withstand mechanical stress. It is known that small variations in cohesin primary sequences are responsible for determining the specificity of interactions with their dockerin binding partners[14,15], but very little is known about how sequence variations affect mechanical stability. To address these points, we performed automated atomic force microscope (AFM)-based single molecule force spectroscopy (SMFS) experiments, a technique which has been established as a robust standard approach to investigate the mechanics of individual molecules.[16–22] The fast dynamics and particularly the large force range of AFM-SMFS made it a prime choice for our investigations. To ensure improved relative force comparability and high experimental throughput, we developed a parallelized AFM sample preparation method which utilizes a one-step protein expression and surface immobilization protocol which is a simplified and easy-to-use version of work previously introduced by our group[23]. At the same time, we carried out steered molecular dynamics (SMD) simulations on structural protein models derived from a homology modelling strategy, which improved our understanding of the cohesin mechanostability at the sub-molecular level. By employing cross-correlation based network analysis on simulation trajectories, we identified regions involved in structural stability outside the mechanical clamp motif and proposed mutations to pin down single crucial amino acids involved in fold strength. This so-called dynamical network analysis has been successfully employed before to investigate force propagation pathways in cohesin-dockerin complexes[24] and in filamins.[25] We tested the proposed mutants experimentally to verify the influence of the proposed alterations on mechanical stability. Employing this combined computational/experimental approach, we were able to predict and verify, among others, a single point mutation outside of the mechanical clamp motif of cohesin 1, which increased its mechanical stability 2.6-fold. The ability to predict such a remarkable difference in mechanostability reveals the potential of our combined approach to characterize and manipulate the mechanical properties of protein domains.

# Results

**AFM-based SMFS.** As specific pulling handles throughout this study we used a type-III cohesin-dockerin complex (Coh3:XDoc3) from *Ruminococcus flavefaciens*, which ensured a high yield of single-molecule interaction curves because of its high specificity and long-term stability. This pulling handle enabled the unfolding of mechanically rigid cohesins due to its high receptor-ligand rupture force of more than 600 pN.[26] It is important to note here that the Coh3:XDoc3 interaction, which we employed as a specific pulling handle, is orthogonal to the ScaA cohesin domains that were unfolded under force, and thus does not interact with them. ScaA cohesin domains from the bridging or hanging region were cloned in frame with XDoc3 and synthesized using cell-free expression (see below). The cohesins of interest therefore carried the XDoc3 domain at their C-termini which was able to bind to the

Coh3 on the cantilever. The cantilever carried covalently and site-specifically immobilized Coh3 domains which were recombinantly expressed as fusion proteins with the 4th immunoglobulin rod filamin domain from *Dictyostelium discoideum* (ddFLN4).[27,28] We used the ddFLN4 as fingerprint domain within the molecular chain to reliably rule out data traces with unspecific- or multiple interactions. The fingerprint domain unfolds at relatively low forces (~100 pN) and its contour length increment (~34 nm) differs significantly from the expected contour length increments of the cohesins (~45 nm). Furthermore, based on our experience we knew that ddFLN4 reliably refolds following mechanical unfolding when bound to the cantilever.

As calibration errors of up to 15% come with cantilever-based force measurements,[29] the precision and therefore the comparability of individual SMFS experiments is limited by the accuracy of the calibration of the cantilever spring constant. This limitation can be circumvented by presenting a set of molecules in a covalently linked microarray format on a single glass slide in predetermined positions.[23] Taking advantage of this setup, the proteins of interest can all be measured with a single cantilever in a high-throughput fashion, and large sets of molecular constructs can be compared in a single AFM measurement session. Although the systematic error of cantilever calibration will still be up to 15%, the relative stability of the proteins can be assessed with extremely high precision, allowing us to detect differences in stability well below 15%.

We have previously reported the use of microfluidics in combination with an intricate multi-step *in vitro* on-chip protein expression and covalent surface attachment protocol.[23] Here we were able to reduce the complexity of the setup significantly, gaining throughput as well as flexibility. We developed a new sample preparation protocol, which is schematized in Figure 2 (a). Briefly, a silicone mask was placed on a glass slide to form microwells, and the resulting spots were covalently functionalized with PEG-Coenzyme A. The individual wells were filled with an *in vitro* transcription/translation (IVTT) system, along with the plasmid DNA encoding for the fusion proteins, and phosphopantetheinyl transferase (Sfp).[30] In a single incubation step, this mixture resulted in cell-free protein synthesis, and simultaneous covalent ligation of the protein library onto the surface through *a* ybbR-tag[30] at the N-terminus. The mask was removed from the glass slide and the spatially separated protein spots were probed in series using a single functionalized cantilever. The combination of a site-specific N-terminal enzyme-mediated immobilization strategy, and a specific C-terminal pulling handle ensured that only fully expressed constructs were probed by the AFM. We recorded 2000 pulling cycles per protein spot while continuously cycling through the array in an automated fashion. No further user interaction was required after the start of the experiment, which allowed long term multi-day measurements to build up large statistics for each construct.
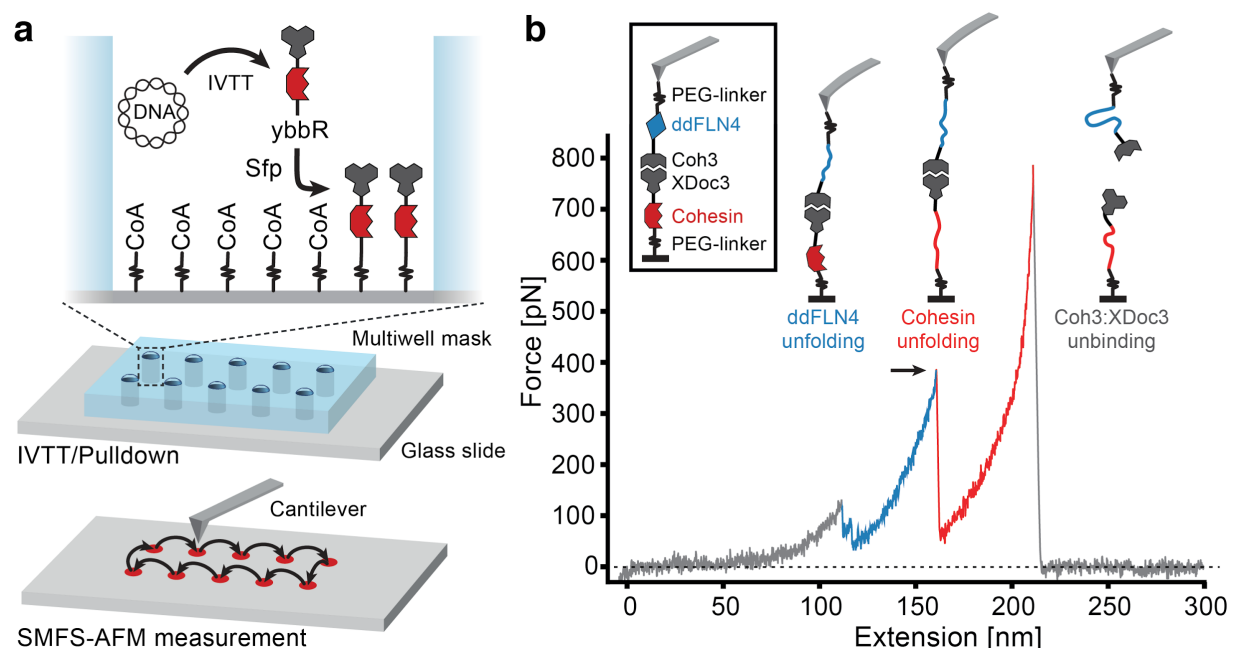
**Figure 2: Schematics of the experimental design and exemplary force curve.** (**a**) A multiwell mask is attached to a glass slide and the surface is functionalized with PEG-Coenzyme A. In a one-pot reaction, an IVTT-kit expresses the proteins of interest containing a ybbR-tag at their N-termini, and an XDoc3 domain at their C-termini. Sfp catalyzes a reaction to bind the constructs site-specifically and covalently to the PEG-CoA spacers. After removal of the multiwell mask, the different constructs are probed by the same AFM tip in an automated fashion. (**b**) Exemplary force-distance trace showing the unfolding of ybbR-Cohesin-XDoc3 immobilized on the surface and the Coh3-ddFLN4-ybbR bound to the cantilever, when the cantilever is retracted from the surface with constant speed. A typical curve shows a two-step unfolding and subsequent stretching of the ddFLN4 fingerprint domain (blue), followed by the unfolding (indicated by an arrow) and stretching of the cohesin domain under investigation (red) and a final rupture of the Coh3:XDoc3 complex.

A typical force vs. distance trace is shown in Figure 2 (b): The cantilever approached the surface and Coh3 bound to XDoc3. Upon retraction of the cantilever with constant speed, the polypeptide chain stretched until the ddFLN4 fingerprint unfolded in a distinct two-step pattern, followed by the unfolding of the respective cohesin under investigation. Finally the Coh3:XDoc3 bond ruptured typically at forces of ~780 pN at force loading rates of ~$10^5$-$10^6$ pN/s. The heights of the various peaks are plotted in the histograms of Figs. 3 and 5. Only force curves displaying a distinct combination of contour length increments (ddFLN4: ~34 nm[27,28], cohesin: ~45 nm) were included in our analysis. Alignments of all relative contour length increments from all pulling experiments can be found in the Supplemental Information (Supplemental Figure S1).

**AFM-based SMFS on ScaA's wild type cohesins.** To investigate the mechanical stability of *A. cellulolyticus*' scaffoldin ScaA, seven distinct surface spots (one for each cohesin) were prepared using the one-step expression/immobilization reaction described above. All data were collected using a single cantilever. Figure 3 shows the resulting unfolding- and rupture force distributions. The outer histograms show very similar force distributions of the fingerprint domain ddFLN4 and of the pulling handle Coh3:XDoc3 independent of the measured molecular construct, which agreed with previously reported literature values.[26,27] Coh3:XDoc3 showed a sharp peak at ~780 pN and a minor shoulder at ~600 pN, both of which are known features of this molecular complex.[26] The remarkable consistency in force among different molecular constructs indicated low force drift during the course of the experiment and confirmed our ability to achieve precise relative comparability of the unfolding forces of the cohesins. Following work done by the Schulten[31] group and Evans and Ritchie[32] in 1997, the cohesin unfolding force distributions can be fitted using a two-state model (from here on called the Bell-Evans model).

With the exception of cohesin 1, all unfolding force distributions could be fitted using the Bell-Evans model - indicating a stable fold and a single barrier unfolding pathway dominating the unfolding process - and most probable rupture forces were obtained. The measured unfolding forces of cohesin 1 resulted in a more complex force distribution. Individual AFM force-distance traces gave no hints of any peculiarities in comparison to the unfolding traces of the other cohesins. We hypothesize that cohesin 1 exhibits more than one distinct fold or has several multi-barrier unfolding pathways that precluded its unfolding force distribution to be fitted using a simple two-state model. In this case we used kernel density estimation (KDE) as a means to smooth the unfolding force histogram and obtain the most probable unfolding force. As originally hypothesized, the three hanging cohesins showed in fact a considerably lower most probable unfolding force (for FWHM errors see Figure 3) (cohesin 1: 139 pN, cohesin 2: 402 pN, cohesin 3: 346 pN) compared to the four bridging cohesins (cohesin 4: 578 pN, cohesin 5: 587 pN, cohesin 6: 461 pN, cohesin 7: 523 pN).
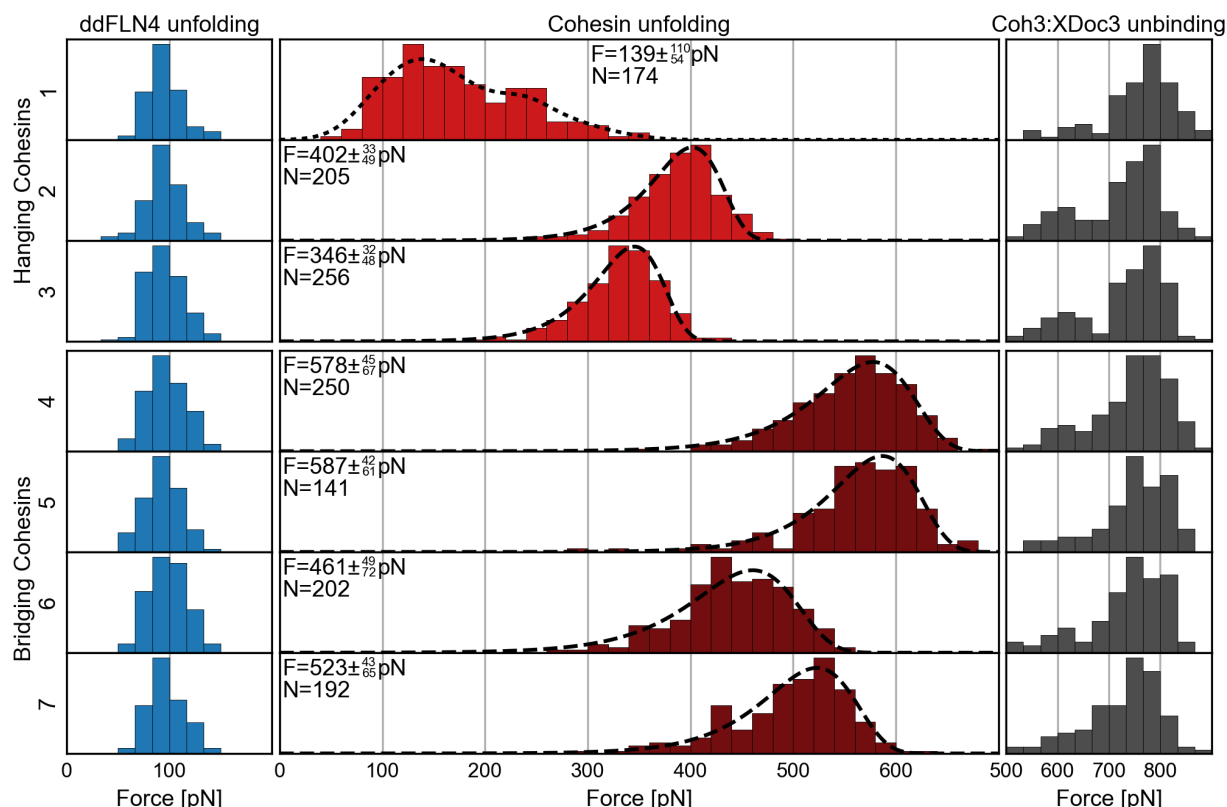


**Figure 3: Unfolding and rupture force histograms of wild type ScaA cohesins.** Histograms showing the unfolding and unbinding forces of the fingerprint domain ddFLN4 (blue), the hanging (light red) and bridging (dark red) cohesins of ScaA, and the Coh3:XDoc3 receptor ligand pulling handle (gray). The force distributions of the ddFLN4 fingerprint and the Coh3:XDoc3 handle are independent of the measured ScaA cohesin construct, which allows for improved relative force comparability of the ScaA cohesins. Unfolding force distributions of cohesins 2-7 were fitted following the Bell-Evans model (dashed lines). A kernel density estimation (KDE) was used to obtain the most probable unfolding force (±FWHM) of cohesin 1 (dotted line). All data were recorded using a single cantilever with a spring constant of 225 pN/nm at a retraction speed of 1600 nm/s during a 24 hour automated SMFS experiment.

**Cohesin homology models.** Since structural data were not available for any of the ScaA cohesins, a homology modeling strategy was adopted,[33] employing Modeller 9.17[34] to obtain structural models for all the cohesins investigated here (Figure 4 (b)). Using BLAST[35] we obtained homologous cohesin structures (PDB IDs: 1G1K, 4DH2, 2VN6 and 4UMS) within the Protein Data Bank[36] (PDB). These structures were then used as templates to derive the homology models that were further refined with molecular dynamics (MD) simulations. Equilibration for 100 ns was performed using NAMD[37] through its QwikMD interface[38] (see Supplemental Figure S2, S3 and S4).

Comparison of the aligned structures (Supplemental Figure S5 A) reveals that all seven cohesins of ScaA show a similar general fold.
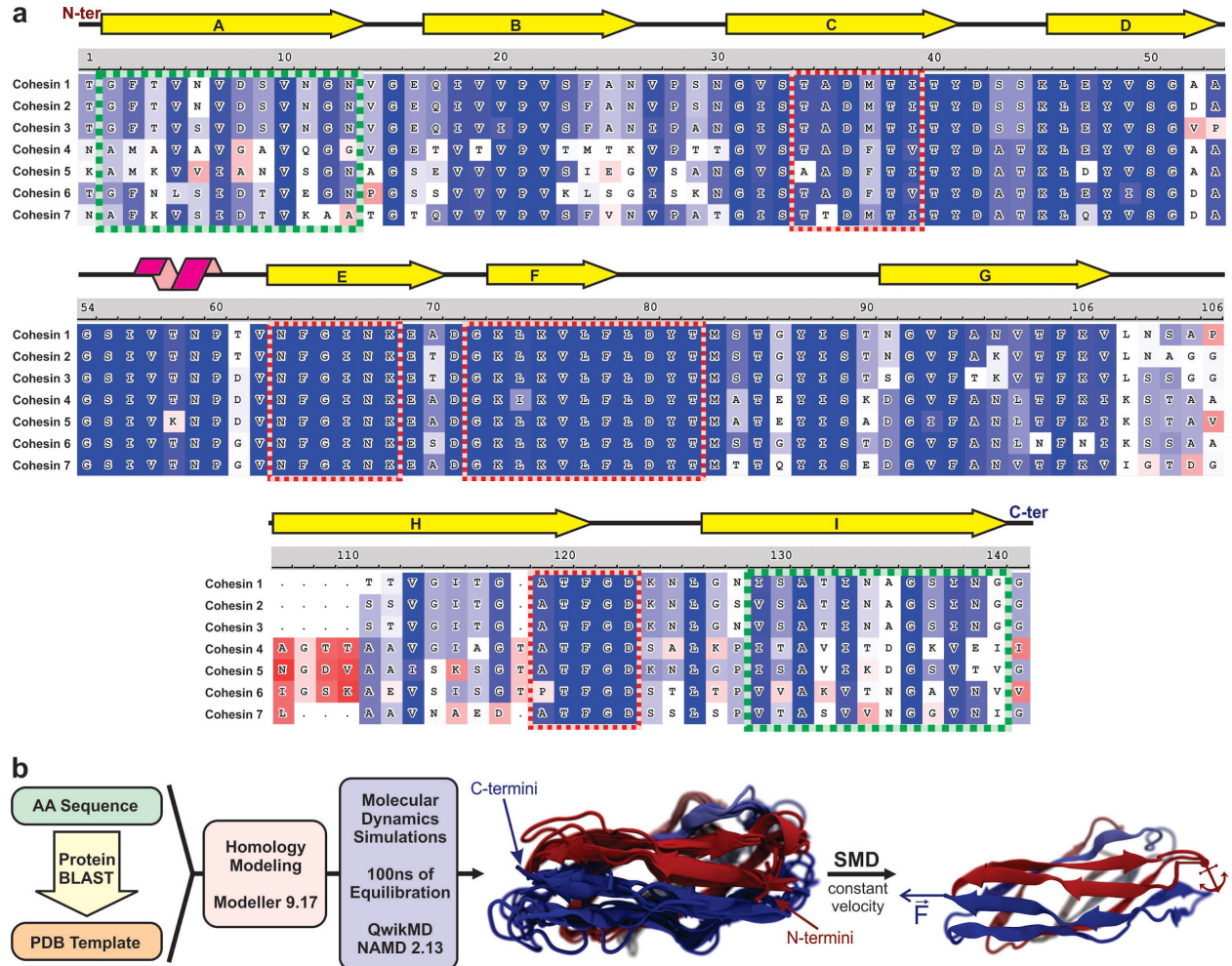


**Figure 4. Sequence alignment and modeling workflow.** (**a**) Sequence alignment of all cohesins of ScaA reveals high overall conservation. The amino acids thought to be primarily involved in mechanical stability[13] are represented by green boxes. Regions primarily involved in dockerin recognition and binding are represented in red boxes. The background colors of the letters represent BLOSUM 70 sequence alignment score: From high (dark blue) to low (red). A high resolution version can be found in Supplemental Figure S6. (**b**) Schematic representation of molecular modeling and dynamics protocol. Homologous protein structures were obtained by running the sequences of the cohesins against the Protein Data Bank using BLAST. Most highly identical structures with high sequence coverage were then used as templates to obtain structural models using Modeller 9.17. Models were properly solvated employing QwikMD and equilibrated for 100 ns of unbiased molecular dynamics simulations using NAMD. A very similar fold was observed for all seven cohesins, here colored ranging from red to blue based on its residue index number. SMD were performed by holding the N-termini and pulling the C-termini with constant velocity.

**SMD simulations and network analyses.** To evaluate the behavior of ScaA's cohesins under force we performed SMD simulations[31,39] using NAMD and QwikMD, where the N-termini of the constructs were fixed and the C-termini were pulled with constant velocity. Employing four different pulling speeds, we investigated first all the unfolding steps in long (on the order of microseconds) SMD simulations. For all cohesins, the first part to unfold with highest peak force is the C-terminal region with beta strand I, followed by beta strand H losing its structure (see Supplemental Figure S7). Next, in the N-terminal region, both beta strand A and B lose their structure almost at the same time in most simulations. Beta-strands C to G finally lose structure under relatively low force (see representative unfolding analysis in Supplemental Figure S8). The force necessary to break any of the cohesin folds was observed to be equal to the force required to unfold the C-terminal segment of these cohesins. Therefore, to

obtain statistically relevant mechanical stability behavior, we performed many 20 ns long simulation replicas (at least 25 replicas per construct). Figure 5 (a) shows the peak unfolding forces, as well as the sequence identity between the cohesins of ScaA and their respective best available structural template. With few exceptions, the simulation replicas reveale d qualitatively the same general trend of the unfolding force peak distributions as measured by SMFS (Figure 3). The absolute forces however are shifted in the simulations towards higher values (Supplemental Figure S9), which was to be expected due to much higher force loading rates of the simulations compared to the AFM experiments.
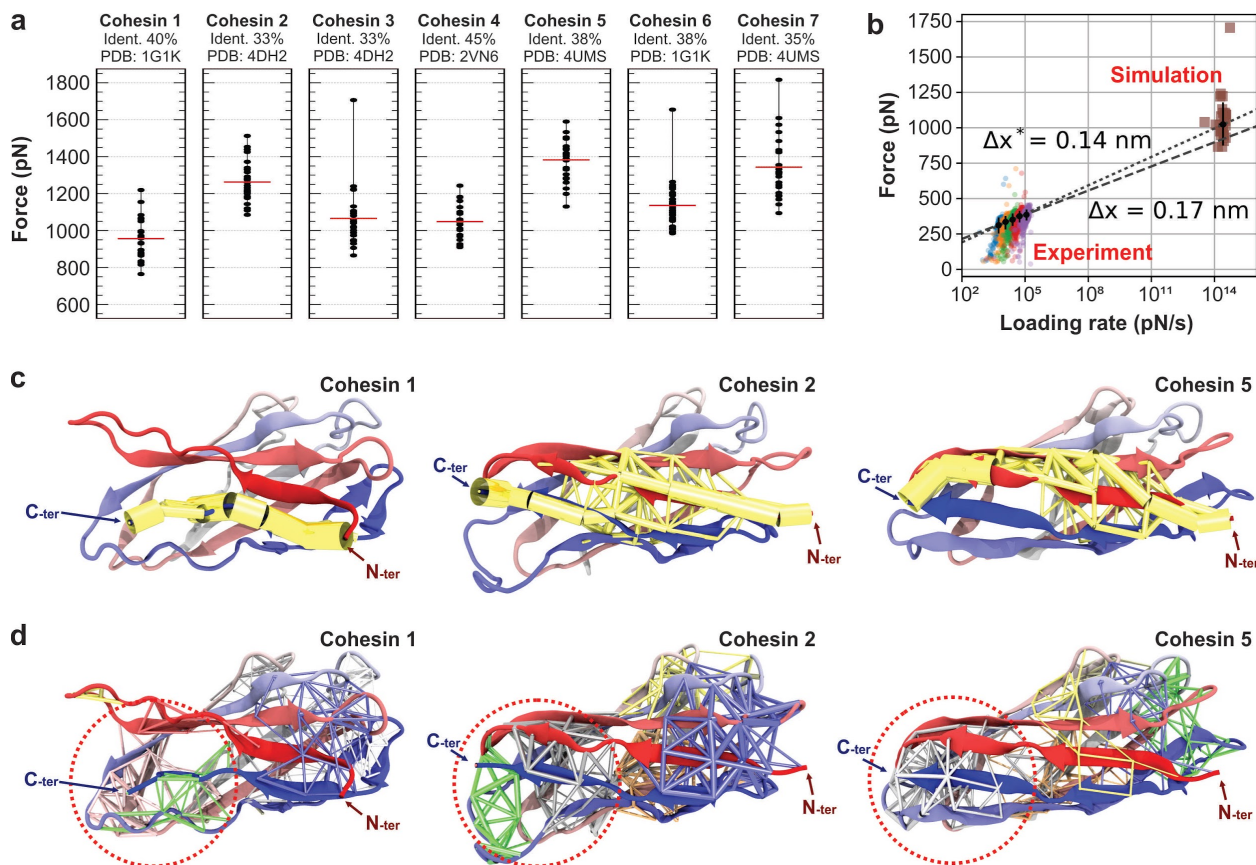


**Figure 5. SMD peak force results, dynamic force spectrum, force propagation pathways and community analysis** (**a**) SMD peak unfolding forces from each simulation replica (black dots) and average peak force per cohesin (red line ± SEM) for the different natural cohesins of ScaA. Sample force vs distance profiles are found in Supplemental Figure S10. Statistical significance between the unfolding forces of all cohesins can be found in Supplemental Figure S11. Also shown are the sequence identity to their respective PDB homology modeling template. (**b**) Dynamic force spectrum for unfolding events of cohesin 3. Varicolored points represent rupture force/loading rate data from an experiment with 5 different pulling speeds. Brown squares represent rupture force/loading rate data from SMD simulations. Black points represent the most probable rupture force/loading rate of each pulling speed obtained from kernel density estimates. Error bars represent the FWHM. Lines represent least-squares fits of the Bell-Evans model and resulting distances to the transition state (±SD) x=0.17 ± 0.012 nm and x*=0.14 ± 0.0015 nm (for detailed representation of experimental data see Supplemental Figure S12). Red line represents a least-squares fit of the DHS model. (**c**) Force propagation pathways through selected cohesins calculated using Pearson correlation (yellow tubes). The thickness of the pathway edges represents the normalized probability of force propagating through the particular edge. (**d**) Network-based community analysis in selected cohesins showing regions with high internal correlation during pulling simulations calculated using generalized correlation. Communities are colored individually and thick connections correspond to high correlation.

A force offset between MD-simulations and SMFS-experiments can easily be understood in view of the Bell-Evans model, which predicts a linear dependence of the unfolding force from the logarithm of the force loading rate. To corroborate this assumption, we varied the pulling rate in the SMFS experiments of cohesin 3 and plotted the

measured unfolding forces together with the corresponding MD simulations in Figure 5 (b). The dashed line represents the best fit to the experimental data extrapolated to the MD-time-domain; the dotted line fits both the experimental and the simulation data. The resulting distances to the transition state of 0.14 ± 0.012 nm and 0.17 ± 0.0015 nm agree very well with literature values of comparable cohesin protein structures.[13,15,26] It is important to note that the slope in the dynamic force spectrum changes with increasing pulling speeds, resulting in a nonlinear upturn at higher pulling velocities as shown by Rico *et al*.[18] This effect is caused by a change from the stochastic to the deterministic unfolding regime. In the former, the unfolding process is governed by spontaneous, thermal unfolding under a given force, while in the latter, the high pulling velocities leave the protein no time to explore its energy landscape. As shown by Dudko, Hummer and Szabo[40] (DHS model), the regime transition can happen at different loading rates and is strongly dependent on the free energy of activation ΔG and hence on the energy landscape of the protein (see Supplemental Figure S13). A fit of the DHS model to the data in Figure 5 (b) suggests that our SMD simulations were carried out at loading rates where unfolding is still dominated by stochastic fluctuations. This indicates that the SMD simulations provide an accurate description of the unfolding process of the system in this study.

The aforementioned results motivated a detailed analysis of the molecular structures and interactions, which could give rise to the particular properties of the different cohesins. Using Pearson cross-correlation-based force propagation analysis - a recently introduced protocol for the analysis of load distributions in molecular complexes[24] - we calculated the suboptimal force propagation pathways, revealing that force propagates mainly through beta-strands A, B and I (see Supplemental Figure S14). Figure 5 (c) shows the force propagation pathways through cohesin 1 and the two strongest cohesins within their respective group, cohesin 2 and 5. Cohesin 1 predominantly showed a single path between the N- and C-terminal beta-sheets that carried the entire mechanical load, suggesting a badly formed mechanical clamp motif. The other cohesins show multiple possible force pathways suggesting a better distribution of force propagating from the N- to the C-terminus through a multitude of backbone hydrogen bonds (see Supplemental Figure S14).

In addition, we investigated the communities formed in the systems by employing dynamic network analysis[41] and generalized correlation[42] (see Supplemental Figure S15). These communities correspond to sets of residues that move in concert with each other, and can be used to investigate regions that are generally more strongly connected during pulling simulations. Figure 5 (d) shows the individual communities of cohesins 1, 2 and 5 in different colors, where thicker connections between the amino acids correspond to higher correlation between them. While cohesins 2 and 5 showed pronounced communities connecting the C-terminus to its surroundings (Figure 5 (d), red dashed circle region), cohesin 1 showed weak communities in this area of the protein, suggesting high and uncorrelated flexibility and therefore loose intra-protein contacts between the N- and C-terminal beta-sheet. Particularly, the area between beta sheets G and H turned out to be most flexible in the case of cohesin 1. The same region of cohesin 1 was observed to be highly flexible also during the 100 ns MD equilibration, and was not as flexible in any other cohesin investigated here. Taken together our results suggested that this is a critical region responsible for cohesin 1 lower stability under mechanical force.

**Mutant design and SMD simulations.** Since the region between beta sheets G and H in cohesin 1 was found to be the most flexible during the equilibrium MD and the analysis of cross-correlation communities suggested that weak communities in this same area could be responsible for the badly formed mechanical clamp motif of cohesin 1, we aligned the sequences of all hanging cohesins in this region (Figure 6 (a)). Despite the high overall sequence similarity of ~85% among *A. cellulolyticus'* hanging cohesins, major differences exist between the weaker cohesin 1 and the stronger cohesins 2 and 3 in the region between beta-strands G and H, more precisely from amino acid 103 to 108. Considering the results of the SMD simulations, force propagation, community analysis and the sequence comparison with cohesin 2 and 3, we proposed four mutants of cohesin 1, namely: A105G, P106G, T107S, and a triple mutant that we refer here as GGS (A105G, P106G, and T107S). To investigate if these mutants would show a

higher force resilience compared to their wild type counterparts, we followed the same modeling and simulation steps as previously discussed for the wild type cohesins.
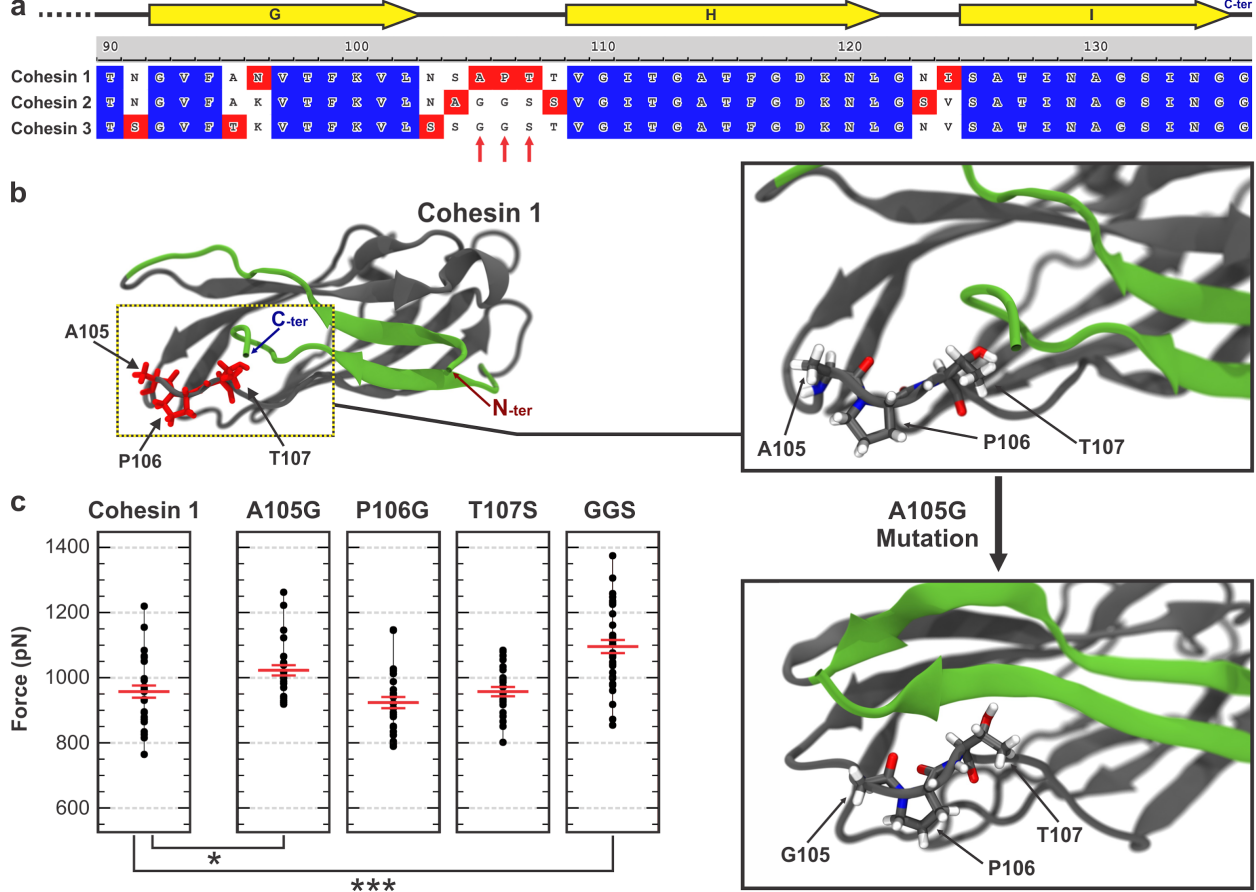


**Figure 6. Mutant design and SMD peak force results.** (**a**) Sequence alignment of the C-terminal end of the hanging cohesins 1-3. From the community analysis we deduced that the low force resilience of cohesin 1 originated from the loss of sequence identity in the flexible area around amino acids 103 - 108. Based on comparison with cohesins 2 and 3, we designed three point mutations (A105G, P106G, T107S) and a triple mutant (A105G P106G T107S (referred to as mutant 'GGS')). A high resolution version can be found in Supplemental Figure S16. (**b**) Left: Homology model of cohesin 1 showing the beta sheets involved in the mechanical clamp motif in green and unique amino acids within the group of hanging cohesins in red. The linker region around amino acids A105, P106 and T107 was observed to be most flexible in the SMD simulations and was therefore suggested for mutation studies. Right: Comparison of the homology models of cohesin 1 and its mutant A105G with molecular representation of the aforementioned region of interest. A seemingly small point mutation from an alanine to a glycine (from a methyl to a proton side chain) at position 105 changed the fold of the protein in this region significantly, resulting in much closer beta strands in the mechanical clamp motif. (**c**) SMD peak force from each simulation replica (black dot) and average peak force per cohesin (red line ± SEM). The significantly increased peak forces of the mutants A105G and GGS suggest an increased stability compared to wild type cohesin 1, while mutants P106G and T107S showed no significant change in average peak unfolding force. The single A105G mutation was able to recover forces in the same range of cohesin 3 (see Figure 5 (a)). Statistical significance between the unfolding forces of all cohesins can be found in Supplemental Figure S17.

After the 100ns MD equilibration, mutated cohesins showed significant structural differences compared to the wild type cohesin 1. A single alanine to glycine mutation (A105G), for example, already stabilized the nearby regions of the protein, resulting in a flawlessly folded beta-stranded C-terminus, as shown in Figure 6 (b). SMD simulations, as shown in Figure 6 (c), revealed that A105G and GGS displayed a significantly higher unfolding force than cohesin 1, with A105G showing a mean unfolding force similar to that of cohesin 3. Analysis of the force propagation profiles and communities of the A105G mutant revealed a behavior that resembles one of the stronger cohesins, as shown in Supplemental Figures S14, S15.

10

**AFM-based SMFS on hanging cohesins and mutants of Cohesin 1.** In order to test the predictions from the SMD simulations and to identify the amino acids responsible for the low force resilience of cohesin 1, we prepared a second set of AFM-based SMFS experiments similar to the ones described in Figure 3. We compared the mechanical stability of the proposed mutants A105G, P106G, T107S and the triple mutant GGS with the three hanging cohesins, again using a single cantilever to ensure improved relative force comparability (Figure 7). Similar to Figure 3, force distributions not following the Bell-Evans model (cohesin 1, P106G and T107S) were smoothed using a KDE to obtain meaningful most probable rupture forces. Most remarkably, mutant A105G showed a dramatic increase in most probable rupture force to 370 pN, making it around 2.6 times stronger than its parent structure cohesin 1 and therefore about as strong as cohesin 3. While the mutant unfolding simulations predicted a slight decrease in mechanical stability for the P106G mutant, the experimentally obtained unfolding force histogram shows no considerable change compared to wild type. Mutant T107S exhibited a bimodal unfolding force distribution with most probable rupture forces of 138 pN and 339 pN roughly similar to the most probable unfolding forces of wild type cohesins 1 and 3. A detailed examination of individual unfolding traces from different force regimes showed no distinctive features which could explain its bimodal unfolding force distribution. We can only theorize that this construct might exhibit a combination of strongly differing folded conformations or unfolding pathways. Such behavior was not observed in the simulations. The triple mutant GGS showed a most probable rupture force of 440 pN, making it as strong as cohesin 2.
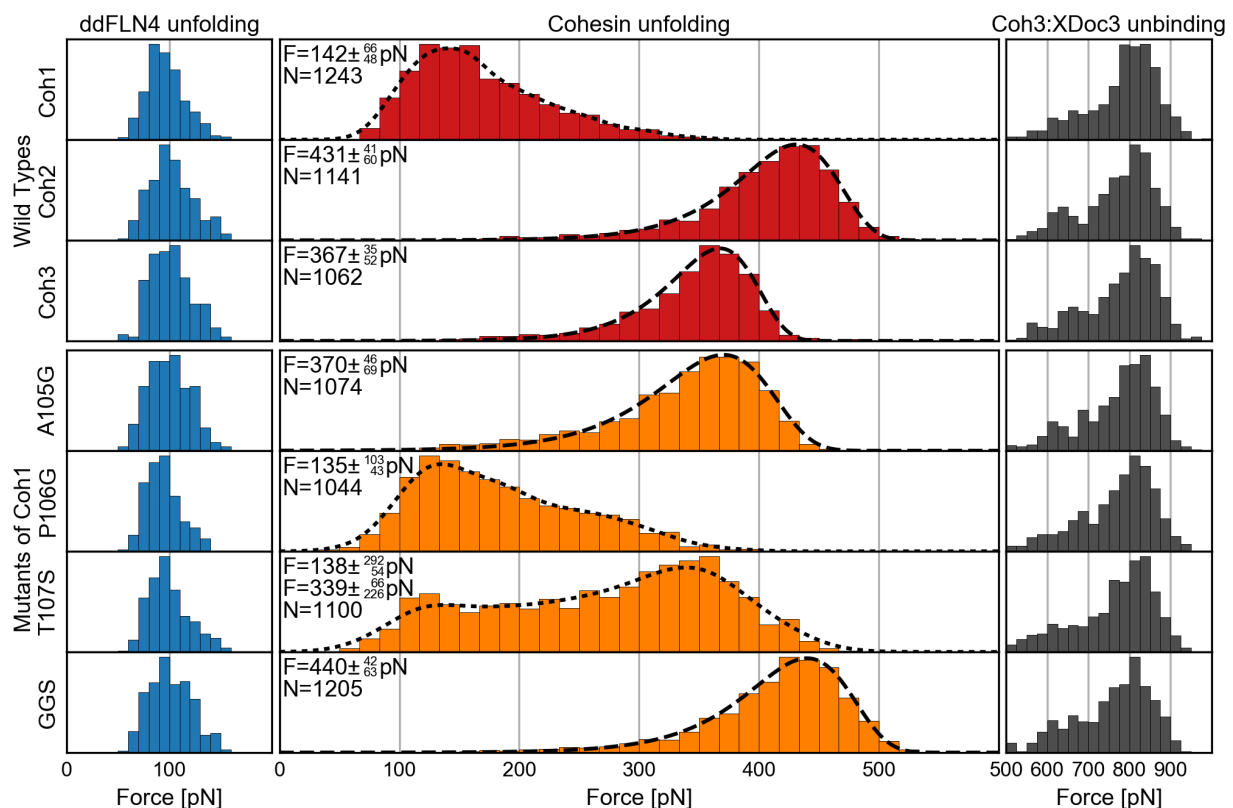


**Figure 7: Unfolding and rupture force histogram of the hanging cohesins and mutants of cohesin 1.** Histograms showing the unfolding and unbinding forces of the fingerprint domain ddFLN4, the wild type hanging cohesins 1-3 (red), mutants of cohesin 1 (orange) and the Coh3:XDoc3 receptor ligand binding handle (gray). The force distributions of the ddFLN4 fingerprint and the Coh3:XDoc3 handle are independent of the measured ScaA cohesin construct. Similar to Figure 3, force histograms were fitted following a Bell-Evans model where possible (cohesin 2, cohesin 3, A105G and GGS; dashed line). A KDE was used to find the most probable rupture forces (±FWHM) in all other cases (cohesin 1, P106G and T107S; dotted line). All data were recorded using a single cantilever with a spring constant of 163 pN/nm at a retraction speed of 1600 nm/s during a 72 hour automated SMFS experiment.

# Discussion

**Mechanical stability of highly homologous ScaA cohesins.** The high precision comparison of the mechanical stability of seven homologous cohesin domains from *A. cellulolyticus'* scaffoldin ScaA was enabled by the development of a novel SMFS sample preparation method, where several constructs were produced by cell-free *in vitro* expression and covalent linkage to the surface in parallel in individual microwells. The proteins were probed sequentially with a single cantilever, enabling precise comparison of unfolding force distributions between multiple domains by eliminating relative error in calibration of cantilever spring constant values. All constructs contained a ddFLN4 domain which served as a molecular fingerprint, and allowed clear identification of single-molecule force-distance traces showing specific interactions through their unique contour length increments. This overall approach facilitated high-throughput SMFS of multiple proteins, and allowed for fast and automated data analysis.

A clear trend can be seen in the most probable unfolding force of all cohesin domains from ScaA (Figure 3). The three cohesins on the far side of the substrate-anchoring CBM (i.e., hanging cohesins) showed in fact a considerably lower most probable unfolding force compared to the four bridging cohesins, which are exposed to force *in vivo*. This result strongly supports the hypothesis that higher mechanical stability is a repeatable feature of cohesins in the bridging region of cellulosomal scaffoldins, despite the fact that all domains tested show high levels of sequence homology (see Figure 4 (a)).

To elucidate the origins of the large differences in mechanostability of the cohesin domains, we generated structural homology models for all seven cohesins under investigation. The combination of knowledge-based information from structural templates together with modern force fields and molecular dynamics allowed us to employ a real-space structural prediction and refinement strategy to obtain all ScaA cohesin structures.[33] However, the computational prediction of three-dimensional protein structures has its limitations, and the accuracy of the predicted models is strongly dictated by the availability of close structural templates.[43] The range of identity between our model cohesins and the best available structural templates was between 33% and 45%, implying rather poor homologues. To check the fold stability, all structural models were subjected to 100ns of equilibrium MD and the final structures were superimposed. The results showed that, even though four different templates were employed, all cohesins generated highly similar structural models (see Figure 4 (b) and Supplemental Figure S5).

The seven structural models were then stretched *in silico* using a constant velocity SMD protocol. It is noteworthy that, with the exception of cohesin 4, the forces of all distributions were shifted by a constant value (±SD) of 782 $\pm$ 29 pN (see Supplemental Figure S9). This finding is remarkable if one takes into account the relatively low identity between the modeled systems and their templates (33%-45%). Simulations showed that although cohesin 4 has the highest identity to its template (45%, PDB: 2VN6), it might have been a suboptimal choice resulting in a non-ideal folding state, as it shows an N-terminal region with wobbly beta-strand formation (see Supplemental Figures S4, S5).

Comparing the force peaks between the simulations and experiments served as a validation for the homology structures. Figure 5 (b) shows that a direct comparison of AFM SMFS, where we varied the force loading rate in a range between ~$10^3$ pN/s and ~$10^5$ pN/s, and *in silico* SMFS at ~$10^{14}$ pN/s. A fit of the DHS model suggests that our SMD simulations unfolded the proteins at loading rates still dominated by the stochastic regime. This finding, which bridges 11 orders of magnitude in force loading rate, suggests that the homology models can provide an accurate description of the unfolding process, validating the predictive power of both comparative modeling and *in silico* SMFS. We want to emphasize, that this only holds true for remarkably strong proteins like the cohesins investigated

here. At similar loading rates of ~$10^{14}$ pN/s, weaker systems would be unfolded in the deterministic regime. In this case, a slower pulling velocity would have to be chosen, requiring considerably more computational time.

**Investigation of the low force resilience of cohesin 1.** Our simulation results in combination with calculated force propagation pathways and correlation communities suggested that the high flexibility in the region around amino acids 100-110 could be responsible for a badly formed mechanical clamp between the N- and C-terminal beta sheets of cohesin 1. We proposed the aforementioned mutants A105G, P106G, T107S and the triple mutant GGS carrying all three mutations, in an attempt to affect the folding and the formation of the mechanical clamp motif, and, ideally, improve mechanical stability.

Following the same modeling and equilibration protocol followed previously, we obtained structural models for the mutants. An inspection of these structures, after 100 ns of MD, revealed how the fold can be affected by a single A105G mutation, as shown in Figure 6 (b). The A105G point mutation resulted in a longer mechanical clamp between the N- and C-terminal beta sheets. SMD simulations predicted an increase in unfolding forces both for mutant A105G and GGS, the latter being the most promising as shown in Figure 6 (c). It is noteworthy that the simulations revealed that the single A105G mutation already increases the force necessary to unfold cohesin 1 up to the same levels of the cohesin 3.

We want to emphasize, that a strategy of carrying out only a couple of SMD simulations with low pulling velocity might give an incomplete picture of a biomolecular system under shear force. The approach adopted here, of simulating many fast pulling simulations (totaling over 350 independent SMD runs), showed that the force distribution in simulations is as widely distributed as in experiments, and therefore a small *in silico* sampling might reveal differing trends to those observed experimentally. A possibility to sample both with slow pulling and many replicas would be to employ coarse-grained methods, which are less computationally demanding. However, our simulations revealed that seemingly small mutations, like changing a methyl group to a hydrogen, can cause enormous differences in folding and therefore force resilience. With such minor changes in the biomolecule, exploratory studies to design new mutants using coarse-grained molecular dynamics simulations would be hardly reliable, as they would lack of atomic detail.

In order to test the predictions from the SMD simulations, we compared the proposed mutants to the wild type hanging cohesins experimentally (Figure 7). We found that the two promising mutants - A105G and GGS - showed a considerable increase in mechanical stability. Mutant A105G showed an increase of most probable rupture force by nearly 2.6-fold to 370 pN, relative to its wild type cohesin 1, which unfolded at 142 pN. As predicted by the SMD simulations, this seemingly small change from an alanine to a glycine outside of the mechanical clamp motif influenced the fold of the protein enough to make it as strong as cohesin 3. The triple mutant GGS showed, again as predicted by the SMD simulations, the largest increase in unfolding force to 440 pN, making as strong as cohesin 2, the strongest cohesin within the group of hanging cohesins, which unfolded at 431 pN. The experimental results confirm the amino acids responsible for the low force resilience of cohesin 1 and the predicted increase in mechanostability of the proposed mutants, thus corroborating the *in silico* approach.

# Conclusions

Multienzyme molecular devices like cellulosomes rely on scaffoldins for the organization of their active constituents. Since these large protein structures can sometimes be subjected to sizable forces, their mechanical stability is a prerequisite for proper and sustained function. This holds particularly true for the scaffoldin ScaA of *A. cellulolyticus*, which anchors the microbe to a cellulose fibril through a CBM in addition to spatially organizing an ensemble of cellulolytic enzymes. Another interesting aspect is that cellulosome scaffoldins are mainly composed of

highly similar cohesin domains with very different mechanical properties. Our *in vitro* and *in silico* SMFS studies not only show that the mechanical stability of all cohesins from the scaffoldin ScaA is consistent with the hypothesis proposed by Valbuena *et al.*[13] that bridging cohesins are mechanically stronger than hanging cohesins, but, moreover, how minimal differences in protein sequence can lead to very different behavior under shear force.

We elucidated the surprisingly low unfolding force of cohesin 1, when compared to the other ScaA cohesins. We found that the point mutation A105G increased the mechanical stability of cohesin 1 more than two-fold when compared to wild type. The remarkably strong influence on the mechanical stability of cohesin 1 of a single alanine to glycine mutation, which effectively only substitutes a methyl group by a hydrogen atom, raises the question why evolutionary pressure has not favored this mutant, as it comes at virtually no additional cost for the organism. Possibly not all cohesins are supposed to display high mechanical stability, since cellulosomal organisms have already been shown to be able to regulate their gene expression patterns depending on potentially varying substrates[44,45]. Thus, occasionally un- and re-folding cohesins would ensure that cellulosomal components can be exchanged in case of changing environmental conditions.

Both approaches, *in silico* and *in vitro*, of our combined approach started from the genetic information coding for the protein, from which the homology models for the former were derived and the samples for the latter were expressed. Given the large number of cellulosome producing microorganisms with sequenced genomes, a wide spectrum of novel combinations e.g. cohesin-dockerin pairs with similar or orthogonal affinities and tunable strengths may be analyzed, modified and combined. The fact that cellulosomes are extracellular organelles of microbes that live in largely diverse ambient environments, including the human gut,[46,47] guarantees robustness of its molecular building blocks and their interactions. This is reflected in the extremely high unfolding barriers and rupture forces of its molecular constituents and qualifies them for a large range of potential applications.

As viable candidates for source materials in a rationally designed artificial protein nanomachine, cellulosomes have demonstrated large potential in molecular engineering applications.[1,6,8] The development of recombinant designer cellulosomes using so-called chimeric scaffoldins allowed control over the position of each enzyme in the cellulosomal complex.[7] Synthetic scaffolds containing orthogonal cohesin domains have furthermore been successfully displayed on the surface of yeast cells, allowing dockerin-tagged cellulases to bind and improve ethanol production almost by a factor of three compared to free enzymes.[48] In terms of industrial cellulose degradation, the incorporation of mechanically stronger cohesins domains and cohesin-dockerin interactions with higher affinities will make designer cellulosomes more durable and efficient. A better understanding of individual cellulosomal components can improve upcoming designs and lead to more efficient and reliable multi-enzyme molecular devices. For example, the new-found properties of ScaA recommend this scaffoldin and its cohesin domains to be part of a potential versatile molecular breadboard for the programmed self-assembly of molecular devices with designed properties.

From a technical point of view, we were able to measure seven constructs using a single cantilever in two separate experiments by utilizing a fast and parallelized sample preparation method, while still achieving sufficient statistics (N=1420 in 24 hrs and N=7869 in 72 hrs). Furthermore, we have shown that even in the absence of crystallized protein structures, SMD simulations, when combined with protein homology modelling, are a powerful tool to investigate the intricate mechanisms governing protein mechanics. Particularly force propagation and community analyses have proven instrumental, allowing us to not only analyze the origins of a particular molecular property, such as the low mechanical stability of cohesin 1, but also opening new means to identify crucial regions for point mutations aiming at locally altering the mechanics of the protein of choice. In summary, our newly developed methods are enabling novel investigations of protein unfolding and rational modification of structural aspects of proteins based on common design principles across different families of proteins well beyond the cellulosome community. Our results demonstrate a strategy that can be applied in fine-tuning mutations that can change the mechanostability of protein domains, and also raise further questions about the evolutionary pressures that can result

in mechanically stronger or weaker proteins. Considering the vast number of cellulosomal constituents yet to be explored, the combination of techniques presented here can potentially accelerate the probing and design of scaffolding domains, starting from nothing more than their genetic code, presenting new opportunities in molecular engineering and biotechnology.

# Associated Content

# Acknowledgments

# References

(1) Bayer, E. A.; Morag, E.; Lamed, R. *Trends Biotechnol.* **1994**, *12* (9), 379.

(2) Bayer, E. a.; Lamed, R.; White, B. a.; Flint, H. J. *Chem. Rec.* **2008**, *8* (6), 364.

(3) Carvalho, A. L.; Dias, F. M. V.; Prates, J. A. M.; Nagy, T.; Gilbert, H. J.; Davies, G. J.; Ferreira, L. M. A.; Romão, M. J.; Fontes, C. M. G. A. *Proceedings of the National Academy of Sciences* **2003**, *100* (24), 13809.

(4) Ding, S.-Y.; Bayer, E. A.; Steiner, D.; Shoham, Y.; Lamed, R. *J. Bacteriol.* **1999**, *181* (21), 6720.

(5) Ding, S.-Y.; Xu, Q.; Crowley, M.; Zeng, Y.; Nimlos, M.; Lamed, R.; Bayer, E. A.; Himmel, M. E. *Curr. Opin. Biotechnol.* **2008**, *19* (3), 218.

(6) Xu, Q.; Ding, S.-Y.; Brunecky, R.; Bomble, Y. J.; Himmel, M. E.; Baker, J. O. *Biotechnol. Biofuels* **2013**, *6* (1), 126.

(7) Fierobe, H.-P.; Mingardon, F.; Mechaly, A.; Bélaïch, A.; Rincon, M. T.; Pagès, S.; Lamed, R.; Tardif, C.; Bélaïch, J.-P.; Bayer, E. a. *J. Biol. Chem.* **2005**, *280* (16), 16325.

(8) Gefen, G.; Anbar, M.; Morag, E.; Lamed, R.; Bayer, E. a. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (26), 10298.

(9) Khan, A. W. *Microbiology* **1980**, *121* (2), 499.

(10) Pandya, P. R.; Singh, K. M.; Parnerkar, S.; Tripathi, A. K.; Mehta, H. H.; Rank, D. N.; Kothari, R. K.; Joshi, C. G. *J. Appl. Genet.* **2010**, *51* (3), 395.

(11) Dassa, B.; Borovok, I.; Lamed, R.; Henrissat, B.; Coutinho, P.; Hemme, C. L.; Huang, Y.; Zhou, J.; Bayer, E. a. *BMC Genomics* **2012**, *13* (1), 210.

(12) Hamberg, Y.; Ruimy-Israeli, V.; Dassa, B.; Barak, Y.; Lamed, R.; Cameron, K.; Fontes, C. M. G. a.; Bayer, E. a.; Fried, D. B. *PeerJ* **2014**, *2*, e636.

(13) Valbuena, A.; Oroz, J.; Hervás, R.; Vera, A. M.; Rodríguez, D.; Menéndez, M.; Sulkowska, J. I.; Cieplak, M.; Carrión-Vázquez, M. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (33), 13791.

(14) Mechaly, A.; Fierobe, H.-P.; Belaich, A.; Belaich, J.-P.; Lamed, R.; Shoham, Y.; Bayer, E. A. *J. Biol. Chem.* **2001**, *276* (13), 9883.

(15) Stahl, S. W.; Nash, M. a.; Fried, D. B.; Slutzki, M.; Barak, Y.; Bayer, E. a.; Gaub, H. E. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (50), 20431.

(16) Florin, E. L.; Moy, V. T.; Gaub, H. E. *Science* **1994**, *264* (5157), 415.

(17) Rief, M.; Gautel, M.; Oesterhelt, F.; Fernandez, J. M.; Gaub, H. E. *Science* **1997**, *276* (5315), 1109.

(18) Rico, F.; Gonzalez, L.; Casuso, I.; Puig-Vidal, M.; Scheuring, S. *Science* **2013**, *342* (6159), 741.

(19) King, G. M.; Carter, A. R.; Churnside, A. B.; Eberle, L. S.; Perkins, T. T. *Nano Lett.* **2009**, *9* (4), 1451.

(20) Müller, D. J.; Engel, A. *Nat. Protoc.* **2007**, *2* (9), 2191.

(21) Fernandez, J. M.; Li, H. *Science* **2004**, *303* (5664), 1674.

(22) Walder, R.; LeBlanc, M.-A.; Van Patten, W. J.; Edwards, D. T.; Greenberg, J. A.; Adhikari, A.; Okoniewski, S. R.; Sullan, R. M. A.; Rabuka, D.; Sousa, M. C.; Perkins, T. T. *J. Am. Chem. Soc.* **2017**, *139* (29), 9867.

(23) Otten, M.; Ott, W.; Jobst, M. a.; Milles, L. F.; Verdorfer, T.; Pippig, D. a.; Nash, M. a.; Gaub, H. E. *Nat. Methods* **2014**, No. september, 1.

(24) Schoeler, C.; Bernardi, R. C.; Malinowska, K. H.; Durner, E.; Ott, W.; Bayer, E. A.; Schulten, K.; Nash, M. A.; Gaub, H. E. *Nano Lett.* **2015**, *15* (11), 7370.

(25) Seppälä, J.; Bernardi, R. C.; Haataja, T. J. K.; Hellman, M.; Pentikäinen, O. T.; Schulten, K.; Permi, P.; Ylänne, J.; Pentikäinen, U. *Sci. Rep.* **2017**, *7* (1), 4218.

(26) Schoeler, C.; Malinowska, K. H.; Bernardi, R. C.; Milles, L. F.; Jobst, M. a.; Durner, E.; Ott, W.; Fried, D. B.; Bayer, E. a.; Schulten, K.; Gaub, H. E.; Nash, M. a. *Nat. Commun.* **2014**, *5*, 5635.

(27) Schwaiger, I.; Kardinal, A.; Schleicher, M.; Noegel, A. A.; Rief, M. *Nat. Struct. Mol. Biol.* **2004**, *11* (1), 81.

(28) Milles, L. F.; Bayer, E. A.; Nash, M. A.; Gaub, H. E. *J. Phys. Chem. B* **2016**, acs.jpcb.6b09593.

(29) Brand, U.; Gao, S.; Engl, W.; Sulzbach, T.; Stahl, S. W.; Milles, L. F.; Nesterov, V.; Li, Z. *Measurement Science and Technology* **2017**, *28* (3), 034010.

(30) Yin, J.; Lin, A. J.; Golan, D. E.; Walsh, C. T. *Nat. Protoc.* **2006**, *1* (1), 280.

(31) Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. *Biophys. J.* **1997**, *72* (4), 1568.

(32) Evans, E.; Ritchie, K. *Biophys. J.* **1997**, *72* (4), 1541.

(33) Goh, B. C.; Hadden, J. A.; Bernardi, R. C.; Singharoy, A.; McGreevy, R.; Rudack, T.; Cassidy, C. K.; Schulten, K. *Annu. Rev. Biophys.* **2016**, *45*, 253.

(34) Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M.-Y.; Pieper, U.; Sali, A. *Curr. Protoc. Protein Sci.* **2007**, *Chapter 2*, Unit 2.9.

(35) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215* (3), 403.

(36) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. In *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*; Rossmann, M. G., Arnold, E., Eds.; International Tables for Crystallography; Springer Netherlands, 2006; pp 675–684.

(37) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26* (16), 1781.

(38) Ribeiro, J. V.; Bernardi, R. C.; Rudack, T.; Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Schulten, K. *Sci. Rep.* **2016**, *6*, 26536.

(39) Izrailev, S.; Stepaniants, S.; Isralewitz, B.; Kosztin, D.; Lu, H.; Molnar, F.; Wriggers, W.; Schulten, K. In *Computational Molecular Dynamics: Challenges, Methods, Ideas*; Springer, Berlin, Heidelberg, 1999; pp 39–65.

(40) Dudko, O. K.; Hummer, G.; Szabo, A. *Phys. Rev. Lett.* **2006**, *96* (10), 108101.

(41) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (16), 6620.

(42) Lange, O. F.; Grubmüller, H. *Proteins* **2006**, *62* (4), 1053.

(43) Zhang, Y. *Curr. Opin. Struct. Biol.* **2008**, *18* (3), 342.

(44) Riederer, A.; Takasuka, T. E.; Makino, S.-I.; Stevenson, D. M.; Bukhman, Y. V.; Elsen, N. L.; Fox, B. G. *Appl. Environ. Microbiol.* **2011**, *77* (4), 1243.

(45) Han, S. O.; Yukawa, H.; Inui, M.; Doi, R. H. *J. Bacteriol.* **2003**, *185* (20), 6067.

(46) Chassard, C.; Delmas, E.; Robert, C.; Lawson, P. A.; Bernalier-Donadille, A. *Int. J. Syst. Evol. Microbiol.* **2012**, *62* (Pt 1), 138.

(47) Moraïs, S.; Ben David, Y.; Bensoussan, L.; Duncan, S. H.; Koropatkin, N. M.; Martens, E. C.; Flint, H. J.; Bayer, E. A. *Environ. Microbiol.* **2016**, *18* (2), 542.

(48) Tsai, S.-L.; Goyal, G.; Chen, W. *Appl. Environ. Microbiol.* **2010**, *76* (22), 7514.

# Supporting Information for

# Combining *in Vitro* and *in Silico* Single Molecule Force Spectroscopy to Characterize and Tune Cellulosomal Scaffoldin Mechanics

Tobias Verdorfer[1], Rafael C. Bernardi[2], Aylin Meinhold[1], Wolfgang Ott[1], Zaida Luthey-Schulten[2,3], Michael A. Nash*[4,5] and Hermann E. Gaub[1]

[1] Lehrstuhl für Angewandte Physik and Center for Nanoscience, Ludwig-Maximilians-Universität, 80799 Munich, Germany.
[2] Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States.
[3] Department of Chemistry, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States.
[4] Department of Chemistry, University of Basel, 4056 Basel, Switzerland.
[5] Department of Biosystems Science and Engineering, Swiss Federal Institute of Technology (ETH Zurich), 4058 Basel, Switzerland.

# Materials and Methods

All reagents were at least of analytical purity grade and all buffers were filtered using a 0.2 µm polyethersulfone membrane filter (Nalgene, Rochester, NY, USA) prior to use. All incubation steps were done at room temperature, if not otherwise stated.

## Gene construction, protein expression and purification

All genes were codon optimized for *E. coli* and synthesized (Invitrogen GeneArt Gene Synthesis - Thermo Fisher Scientific Messtechnik GmbH, Regensburg, Germany). All constructs were cloned into pET28a vectors using the Gibson assembly strategy[1] (New England Biolabs, MA, USA). All protein sequences can be found in the Supplementary Information.

ScaA cohesin mutant plasmid DNA was constructed using individually designed primers (Eurofins Genomics GmbH, Ebersberg, Germany) and the Phusion High-Fidelity PCR Kit (Thermo Fisher Scientific Messtechnik GmbH, Regensburg, Germany). The resulting double stranded linear DNA was ligated, phosphorylated and the template DNA was digested, in parallel, using a homemade reaction mix (1µl CutSMART buffer, 1µl ATP, 1µl T4 Polynucleotide Kinase, New England Biolabs, MA, USA, 1µl *Dpn*I, 1µl T4 DNA ligase and 0.5µl PEG-6k, Thermo Fisher Scientific GmbH, Regensburg, Germany, combined with 4.5µl unpurified PCR product) incubated at 37°C for 15min, 22°C for 45min and finally at 80°C for 5min. All plasmids used in *in vitro* protein expression were amplified in DH5-alpha cells, purified using the QIAprep Spin Miniprep Kit (Qiagen, Hilden, Germany), eluted with ultrapure water and stored at -20° C. All sequences were finally checked by DNA sequencing (Eurofins Genomics GmbH, Ebersberg, Germany).

Coh3-ddFLN$_4$-HIS-ybbR protein was expressed in *E. coli* NiCo21(DE3) cells (New England Biolabs, MA, USA). Precultures of 5 mL in LB medium, grown overnight at 37°C, were inoculated in ZYM-5052 auto-induction media containing kanamycin and grown for 6 h at 37°C followed by 24 h at 25°C[2]. Bacteria were spun down, and stored at -80°C. The pellet was resuspended and cells were lysed through sonication followed by centrifugation at 18000 g for 1 h at 4°C. The supernatant was applied to a Ni-NTA column (GE Healthcare, MA, USA) for HIS-Tag purification and washed extensively using HIS wash buffer (25mM TRIS, 500mM NaCl, 0.25% Tween-20, 10 % (v/v) Glycerol, 20mM imidazole, pH 8.5 @ 4°C), followed by a elution using HIS elution buffer (HIS wash buffer with 200mM imidazole instead of 20mM). Fractions containing protein were concentrated over regenerated cellulose filters (Amicon, Merck KGaA, Darmstadt, Germany), exchanged into measurement buffer (TBS- Ca: 25 mM Tris, 72 mM NaCl, 1mM CaCl2, pH 7.2) using desalting columns (Zeba, Thermo Scientific, MA, USA), and frozen with 25 % (v/v) glycerol in liquid nitrogen to be stored at -80°C until used in experiments.

## AFM Sample preparation

The sample preparation in these experiments follows in principle previously published protocols.[3–5] In brief, both the AFM cantilevers (Biolever Mini, Olympus, Tokyo, Japan) and the microscope slides (76mmx26mm, Carl Roth GmbH, Karlsruhe, Germany) were cleaned and silanized using (3-aminopropyl)-dimethyl-ethoxysilane (APDMES, abcr GmbH, Karlsruhe, Germany) by baking at 80°C for 1h. A multiwell mask (CultureWell Gasket, Grace Bio-Labs, Bend, USA) was cleaned by sonication in a 1:1 mixture of isopropyl alcohol (IPA) and ultrapure water and then dried in a stream of nitrogen. The mask was attached to the glass slide to allow compartmentalization of the surface. The cantilevers were incubated with heterobifunctional NHS-PEG-Maleimide (5 kDa; Rapp Polymere, Tübingen, Germany) in 100mM HEPES buffer pH 7.4 for 45 min. The surfaces in the wells however were incubated with a 1:100

mixture of NHS-PEG-Maleimide and NHS-PEG-CH$_3$ (both 5 kDa; Rapp Polymere, Tübingen, Germany) in 100mM HEPES buffer pH 7.4, which, as experience has shown, will later result in the right surface density of immobilized protein for SMFS measurements in these experiments. After rinsing with ultrapure water, both the cantilevers and the surfaces were incubated with 1 mM Coenzyme A (CoA) in a 1 mM sodium phosphate pH 7.2, 50 mM NaCl, 10 mM EDTA buffer for at least 1 h. After a final ultrapure water rinse the cantilevers were incubated with 40µM Coh3-ddFLN$_4$-HIS-ybbR and 5 µM phosphopantetheinyl transferase (Sfp) for 2 h with magnesium chloride supplemented measurement buffer (TBS- Ca: 25 mM Tris, 72 mM NaCl, 1mM CaCl$_2$, 20mM MgCl$_2$ pH 7.2). The glass slide with the multiwell mask still attached was stored under Argon for later use. The cantilevers were rinsed extensively with measurement buffer (TBS- Ca: 25 mM Tris, 72 mM NaCl, 1mM CaCl2, pH 7.2) and finally stored in it until use in measurement.

## One-step *in vitro* expression and protein pulldown

PURExpress® IVTT-kit was thawed on ice and supplemented with 5 µM Sfp, 0.8 U/µl RNase inhibitor (NEB #M0314), 10 ng/µl Plasmid-DNA, 0.05% v/v Triton X-100 (Sigma-Aldrich Chemie GmbH, Taufkirchen, Germany) and ultrapure water resulting in total volumes of 10µl for each reaction mix. There was no need to supply this reaction mix with additional MgCl$_2$ for the Sfp coupling reaction, since the PURExpress® IVTT-kit already contains 13mM MgCl$_2$[6]. The reaction mixes were transferred to the wells onto the CoA functionalized glass slide and incubated at 37°C for 3h. During this time the cell free expression kit is constantly producing proteins, while at the same time the Sfp couples expressed protein to the surface via the ybbR tags. Both cantilevers and surfaces were rinsed extensively with measurement buffer (TBS- Ca: 25 mM Tris, 72 mM NaCl, 1mM CaCl2, pH 7.2) before measurement and finally the multiwell mask was removed from the surface and stored in a 1:1 mixture of IPA and ultrapure water for further use.

## AFM SMFS measurements

A custom build AFM connected to a MFP3D controller (Asylum Research, Santa Barbara, CA, USA) was used for all measurements. Acquisition- and instrument control software was written in Igor Pro 6 (Wavemetrics, OR, USA). The cantilever was aligned to each measurement spot by moving the sample using a 25mmx25mm piezomotor stage (PI, Karlsruhe, Germany) using a camera mounted below the sample. These positions were saved in the software for later use. The cantilever was brought in close proximity to the surface and constant speed measurements with retraction speeds of 1600 nm/s were started. The glass surface was moved horizontally by 100 nm in a snail-like-pattern within each protein spot. After 2000 approach- and retract-cycles the AFM-head was automatically lifted by a linear piezo actuator (Newport, CA, USA) and the surface was moved horizontally by typically ~300 µm to expose the cantilever to the next protein spot. Cantilevers were calibrated using the equipartition theorem method[7].

## AFM SMFS Data Analysis

Data analysis was carried out following previous work[8]. In short, data were transformed into physical units and corrected for cantilever bending, laser spot- and baseline-drift. Force peaks and rupture events were detected and transformed to contour length space. The Worm Like Chain model (WLC)[9] was used to fit relevant peaks. All curves showing a ddFLN4 and cohesin contour length increment (ddFLN4: 34nm[8,10], cohesin: 45nm) were used to assemble unfolding force histograms, which were then fitted following the Bell-Evans model[11,12], which is commonly used to estimate the distance to the transition state $\Delta x$ and the natural off-rate $k_0$ of mechanically induced receptor ligand dissociation from single-molecule force spectroscopy experiments.

Bell-Evans probability density function at given loading rate r:

$$p(F) = \frac{k_0}{r} \, exp[\frac{\Delta x}{k_B T}F - \frac{k_0 \, k_B T}{r \, \Delta x}(e^{\frac{\Delta x}{k_B T}F} - 1)]$$

The Bell-Evans model predicts a linear dependence between the most probable rupture force <F> and the logarithm of the force loading rate r:

$$< F(r) >= (\frac{k_B T}{\Delta x}) \, ln(\frac{r \, \Delta x}{k_0 \, k_B T})$$

The Dudko-Hummer-Szabo model[13] describes a non-linear dependence for the most probable rupture force on loading rate:

$$< F(r) >= \frac{\Delta G}{v \Delta x} \left\{ 1 - \left[ \frac{k_B T}{\Delta G} ln \left( \frac{k_0 \, k_B T}{r \, \Delta x} exp \left( \frac{\Delta G}{k_B T} + \gamma \right) \right) \right]^{v} \right\}$$

where $\Delta G$ is the free energy of activation and $\gamma \approx 0.577...$ is the Euler-Mascheroni constant. The model parameter $v$ defines the single-well free-energy surface model used ($v$ = 2/3 for linear-cubic and $v$ = 1/2 for cusp free-energy).

## Structural Model Determination

The amino acid sequence of all seven cohesins under investigation were obtained from the GenBank (GenBank: AAF06064.1) proteomic server[14,15]. The template search was performed employing the similarity search algorithm in the protein Blast server (http://blast.ncbi.nlm.nih.gov/Blast.cgi)[16] using the Protein Data Bank[17] (http://www.pdb.org) as database and the default options. Using VMD's[18] multiseq[19] analysis tool, sequences were aligned to templates employing ClustalW algorithim[20]. The construction of cohesin models were performed using MODELLER 9.17 software[21] that employs spatial restriction techniques based on the 3D-template structure. The best model was selected by analyzing the stereochemical quality check using PROCHECK[22] and overall quality by ERRAT server.[23] All structures were subjected to 100 ns of equilibrium MD, as described below, to ensure conformational stability. All structures shown in this manuscript are from post-equilibration simulations.

## Molecular dynamics simulations

Employing advanced run options of QwikMD,[24] structural models were solvated and the net charge of the proteins were neutralized using a 75 mM salt concentration of sodium chloride, which were randomly arranged in the solvent. The overall number of atoms included in MD simulations varied from 50,000 in the equilibrium simulations to near 300,000 in the pulling simulations. The MD simulations in the present study were performed employing the NAMD molecular dynamics package.[25] The CHARMM36 force field[26,27] along with the TIP3 water model[28] was used to describe all systems. The simulations were performed assuming periodic boundary conditions in the NpT ensemble with temperature maintained at 300 K using Langevin dynamics for pressure, kept at 1 bar, and temperature coupling. A distance cut-off of 11.0 Å was applied to short-range, non-bonded interactions, whereas long-range electrostatic interactions were treated using the particle-mesh Ewald (PME)[29] method. The equations of motion were integrated using the r-RESPA multiple time step scheme[25] to update the van der Waals interactions every two steps and electrostatic interactions every four steps. The time step of integration was chosen to be 2 fs for all simulations performed. Before the MD simulations all the systems were submitted to an energy minimization protocol for 1,000 steps. MD simulations with position restraints in the protein backbone atoms were performed for 10 ns and served to pre-equilibrate systems before the 100 ns equilibrium MD runs, which served to evaluate structural model stability. During the 10 ns pre-equilibration the initial temperature was set to zero and was constantly increased by 1 K every 1,000 MD steps until the desired temperature (300 K) was reached.

With structures properly equilibrated and checked, solvent boxes were enlarged in the Z coordinate to allow space for protein unfolding during SMD simulations. The new solvent boxes were equilibrated for 10 ns keeping the protein atoms restrained in space. SMD simulations[11] were performed using a constant velocity stretching (SMD-CV protocol), employing four different pulling speeds: 250, 25, 2.5 and 0.5 Å/ns.

Simulation replicas (at least 25 per system), used in all the plots in this manuscript, were performed with constant pulling speed of 2.5 Å/ns. Values for force over the pulling spring were saved every 50 steps. The spring constant of the pulling spring was set to 5.0 kcal/mol/Å$^2$, while the holding spring had a constant of 10 kcal/mol/Å$^2$. In all simulations, totaling over 350 SMD simulations, SMD was employed by harmonically restraining the position of N-terminal amino acid residue of the cohesin domain, and moving a second restraint point, at the C-terminal of the cohesin domain, with constant velocity in the +z direction. The procedure is equivalent to attaching one end of a harmonic spring to the end of a domain and pulling on the other end of the spring. The force applied to the harmonic spring is then monitored during the time of the molecular dynamics simulation. The pulling point was moved with constant velocity along the z-axis and due to the single anchoring point and the single pulling point the system is quickly aligned along the z-axis. Owing to the flexibility of the linkers between the cohesins and fingerprint domains, this approach reproduces the experimental set-up. All analyses of MD trajectories were carried out employing VMD[18] and its plug-ins. Secondary structures were assigned using the Timeline plug-in, which employs STRIDE criteria.[30]

The Network View plugin[31] on VMD was employed to perform dynamical network analysis. A network was defined as a set of nodes, all α-carbons, with connecting edges. Edges connect pairs of nodes if corresponding monomers are in contact, and 2 non-consecutive monomers are said to be in contact if they fulfill a proximity criterion, namely any heavy atoms (non-hydrogen) from the 2 monomers are within 4.5Å of each other for at least 75% of the frames analyzed. As suggested by Sethi et al.[31] nearest neighbors in sequence are not considered to be in contact as they lead to a number of trivial suboptimal paths, which can be understood as allosteric signaling pathways or force propagation pathways [32]. Suboptimal paths are defined as paths that are slightly longer than the optimal path, with a given suboptimal path visiting a node not more than once. These multiple communication paths are nearly equal in length, and not all residues along these paths need be considered important for allostery. Instead, only residues or interactions that occur in the highest number of suboptimal pathways need to be conserved to guarantee an effective pathway for allosteric communication. The thickness of the edges connecting the nodes reveals the least and most used paths. Allostery can be understood in terms of pathways of residues that efficiently transmit energy, here in the form of mechanical stress, between different binding sites[33]. The dynamical networks were constructed from 20 ns windows of the total trajectories sampled every 400 ps. The probability of information transfer across an edge is set as wij = −log (| Cij |), where Cij is the correlation matrix calculated with Carma[34]. Using the Floyd-Warshall algorithm, the suboptimal paths were then calculated. The tolerance value used for any path to be included in the suboptimal path was −log (0.5) = 0.69. To calculate the relevance of off-diagonal terms in the correlation matrix we employed Carma to calculate a correlation matrix where x, y, z components of each atom were considered independently. As previously investigated by our group [32], Pearson correlation is ideal for force propagation calculation. However, due to its nature, communities analysis would benefit from an information-theory-based method, so here we employed generalized correlation[35] to the community analysis. Tightly correlated groups of atoms are clustered into communities, indicating functional domains of biomolecules and important interfaces between multi-molecule complexes.
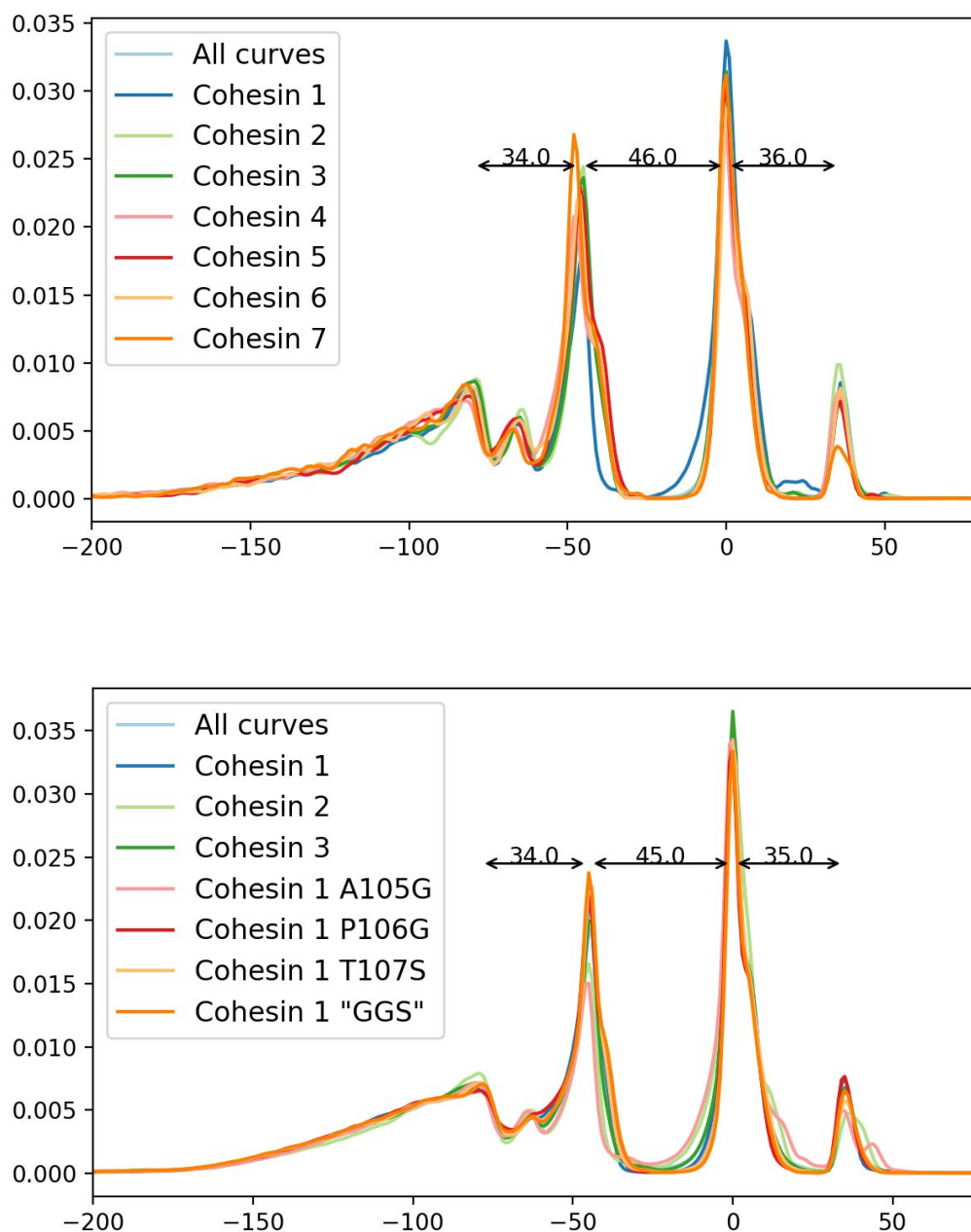
**Figure S1:** Frequency of observed relative contour lengths increments determined by transforming multiple force traces into contour length space via the worm-like chain model and aligning them. The individual increments (f.l.t.r.) correspond to the unfolding of the ddFLN4 fingerprint domain, the ScaA cohesins and the occasional unfolding of the X-module of the Coh3.XDoc3 complex[36].
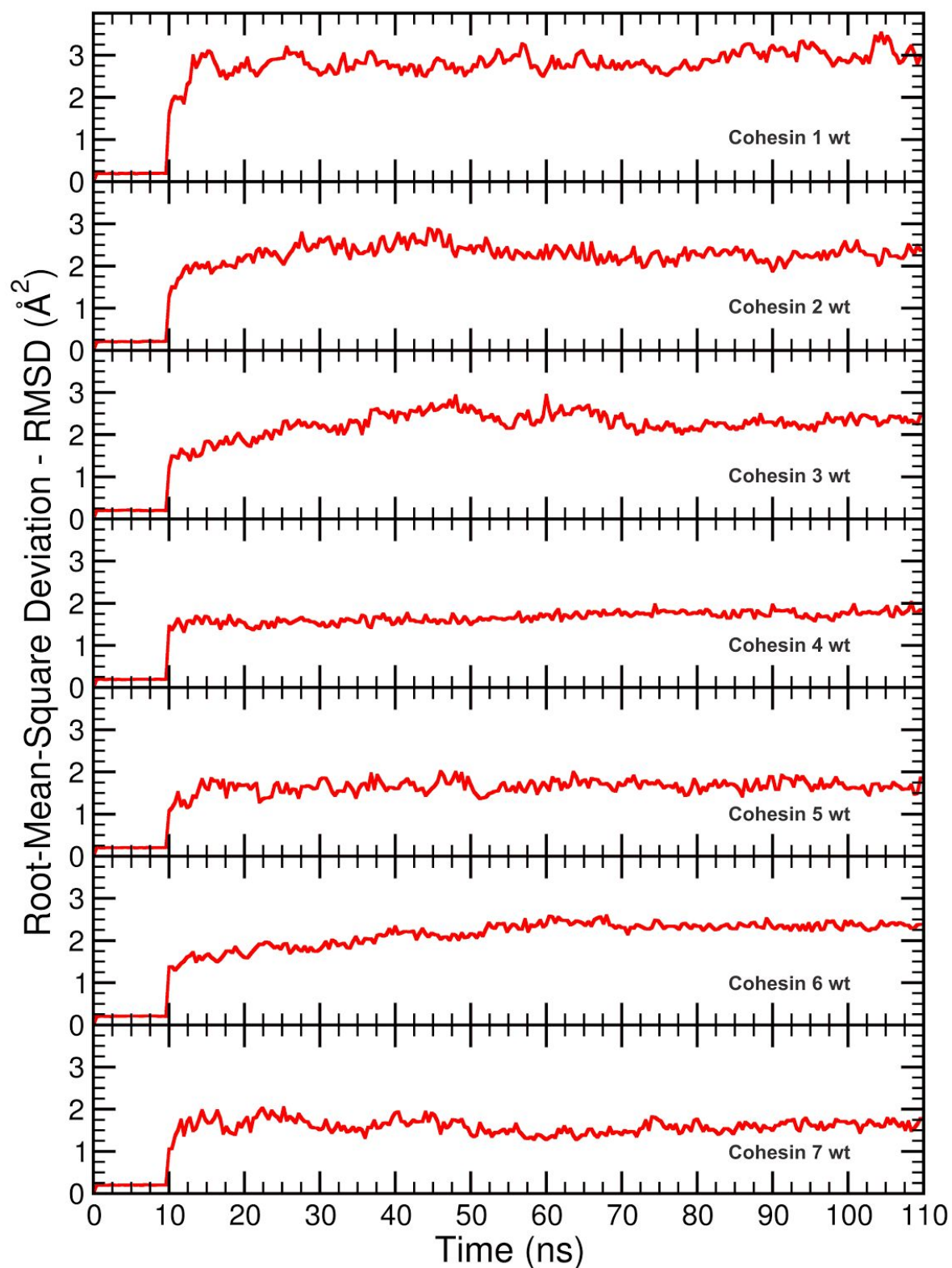
**Figure S2: Root-Mean-Square Deviation (RMSD) for equilibrium simulations.** All constructs were simulated with position restraints of the backbone atoms during 10 ns and free of restraints during 100 ns. All plots show stable structures after approximately 30ns. It is noteworthy that hanging cohesins have a higher RSMD value, particularly cohesin 1.
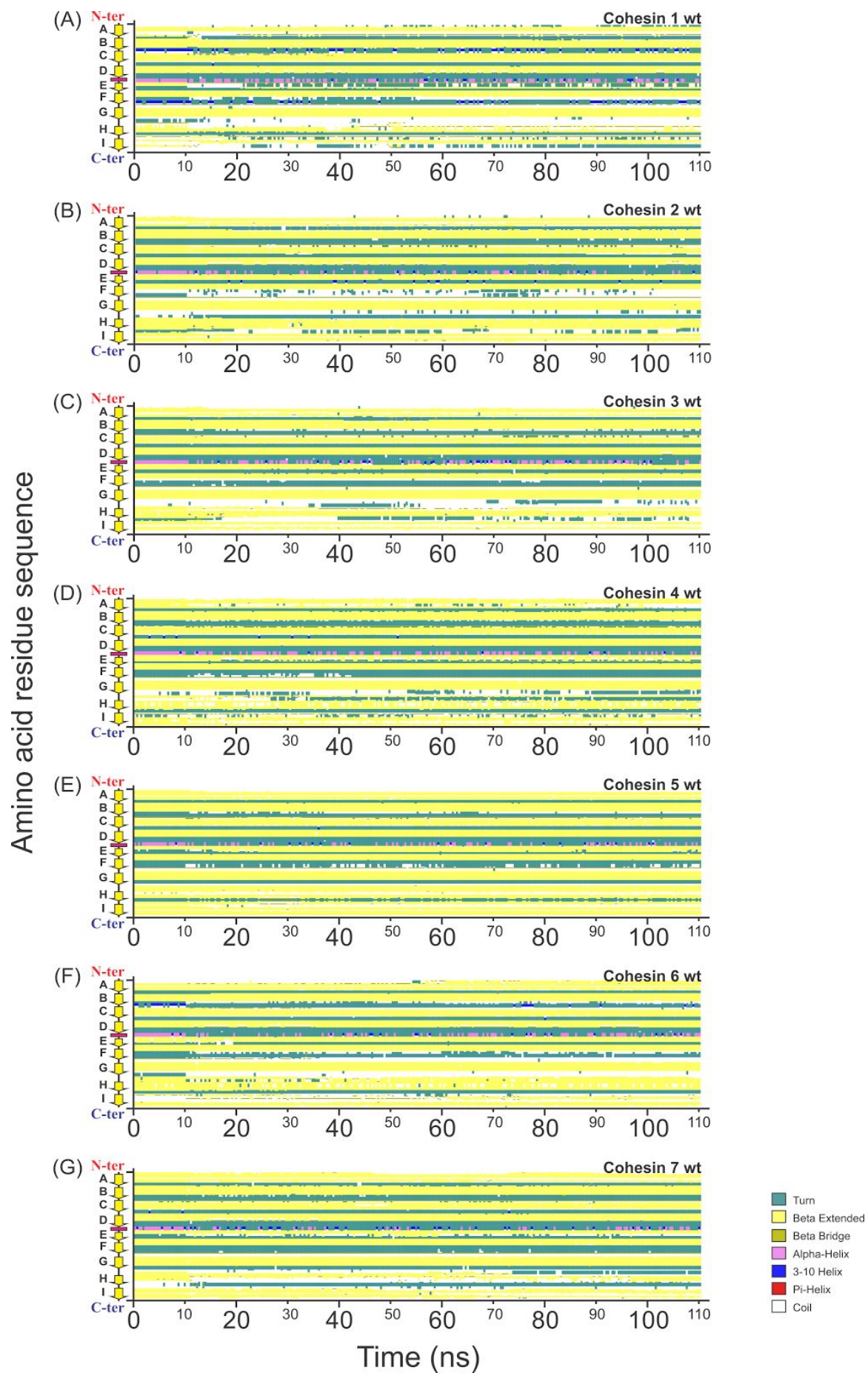
**Figure S3: Secondary Structure evolution during equilibration simulations.** Secondary structure content was evaluated using VMD's Timeline during the equilibration simulations. All constructs were simulated with position restraints of the backbone atoms during 10 ns and free of restraints during 100 ns. All conformations show stable structures.
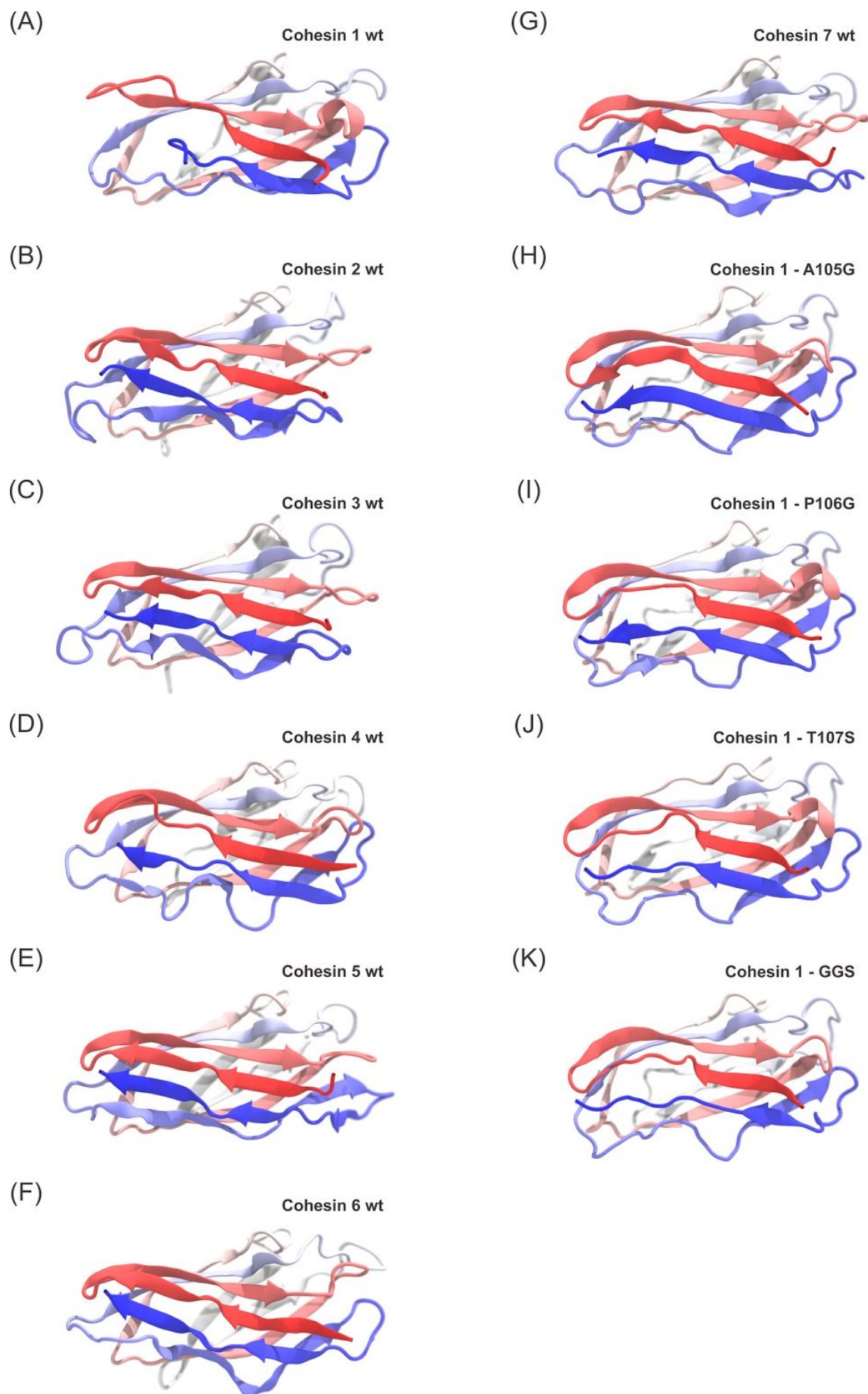
**(A)** Cohesin 1 wt

**(B)** Cohesin 2 wt

**(C)** Cohesin 3 wt

**(D)** Cohesin 4 wt

**(E)** Cohesin 5 wt

**(F)** Cohesin 6 wt

**(G)** Cohesin 7 wt

**(H)** Cohesin 1 - A105G

**(I)** Cohesin 1 - P106G

**(J)** Cohesin 1 - T107S

**(K)** Cohesin 1 - GGS

**Figure S4. Structural model of studied cohesins after 100ns of MD simulation.** All structures were obtained using Modeller 9.17 and subjected to 100 ns of molecular dynamics equilibration using QwikMD and NAMD. All images were prepared using VMD.
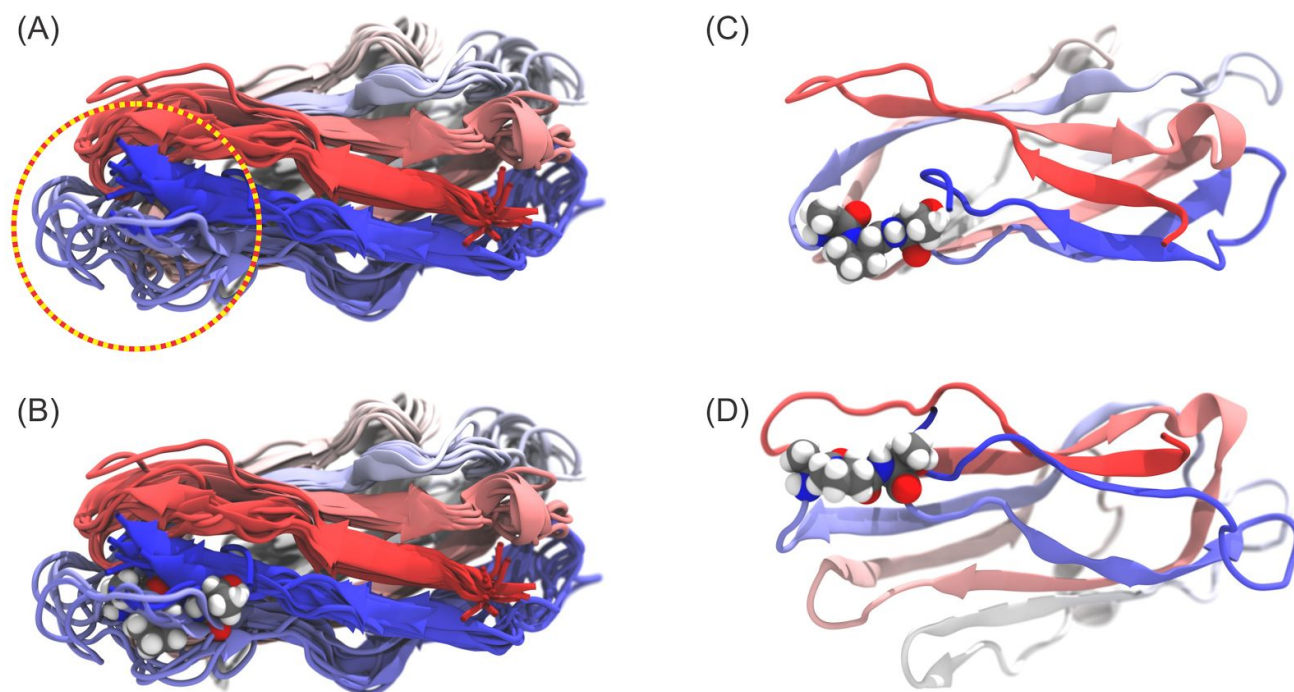
**Figure S5. Structural model for ScaA cohesins after 100ns of MD simulation.** (A) Using Modeller, very similar model structures were obtained for ScaA cohesins. The region highlighted in the circle was observed to be the most flexible one in cohesin 1, presenting a different behavior than the other cohesins. (B) Sequence alignment shows that, in the highlighted region of (A), 3 amino acid residues of cohesin 1 were different, compared to cohesin 2 and 3, namely ALA105, PRO106 and THR107. (C) (D) Two different viewpoints of cohesin 1 with highlighted ALA105, PRO106 and THR107.
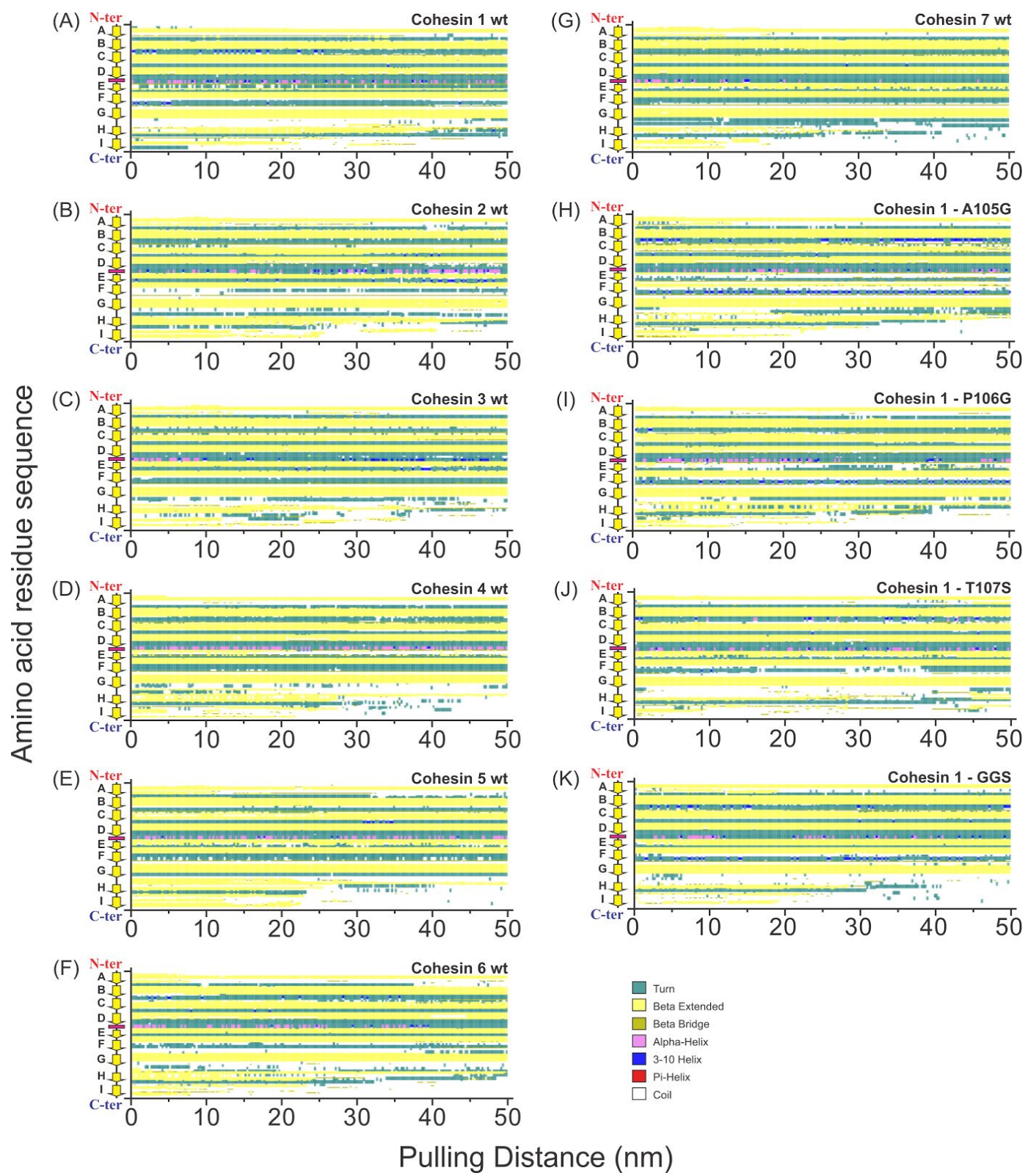
**Figure S6. Sequence alignment of all cohesins of ScaA.** The amino acids thought to be primarily involved in mechanical stability are represented by green boxes. Regions primarily involved in dockerin recognition and binding are represented in red boxes. The background colors of the letters represent BLOSUM 70 sequence alignment score: From high (dark blue) to low (red).

**Figure S7. Secondary Structure evolution during pulling simulations.** Secondary structure content was evaluated using VMD's Timeline during the SMD simulations. Here we show the evolution of the secondary structure during the first 50nm of pulling, which corresponds to the region where the peak force is observed. The plots present a representative simulation (one of the replicas) for each system studied. In all simulations presented the C-terminal region is the first to unfold, showing that the highest peak corresponds to losing the last beta-strand structure.
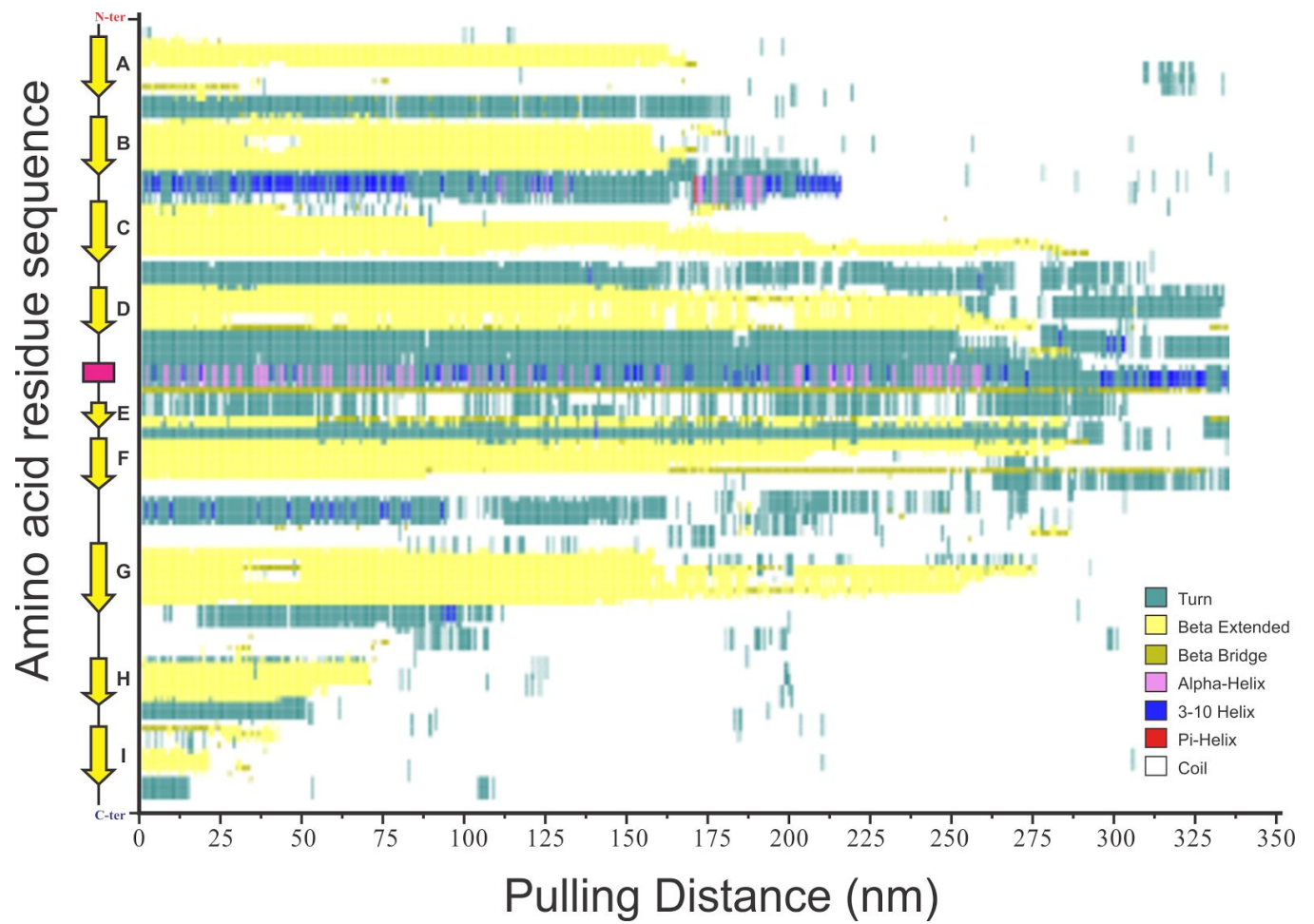
**Figure S8. Cohesin 1 secondary structure evolution during pulling simulations.** Secondary structure content was evaluated using VMD's Timeline during the SMD simulations. Here we show the evolution of the secondary structure during the whole unfolding process.
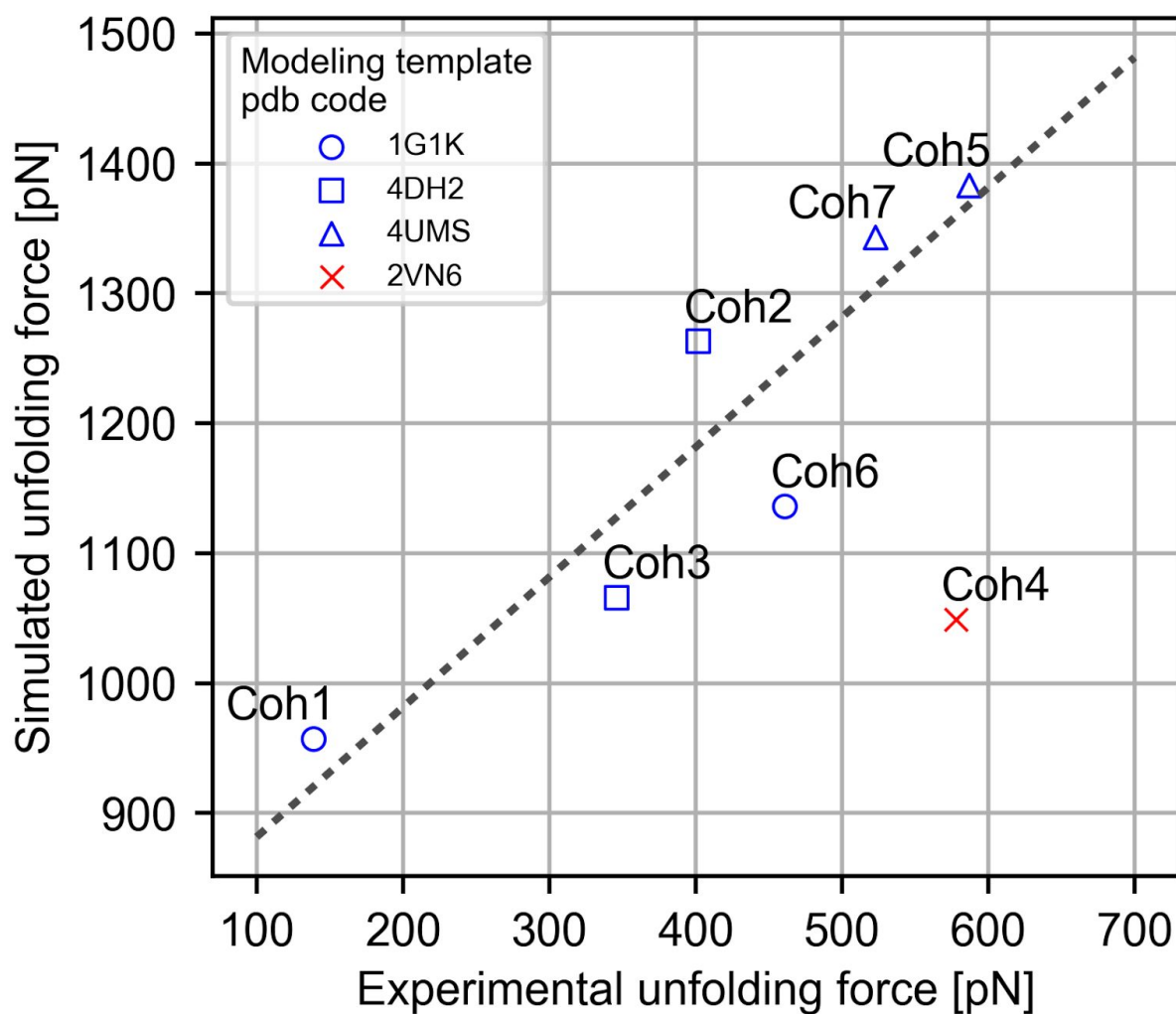
**Figure S9. Experimental unfolding force vs. simulated peak unfolding force of all wild type cohesins.** A linear fit shows a clear trend between measured and simulated unfolding forces with offset fitting parameter (±SD) $F_0$ = 782 ± 29 pN. Cohesin 4 (red) was excluded from the fit since, as explained in the main text, its homology modeling template was a suboptimal and likely resulted in a non-ideal initial folded state.
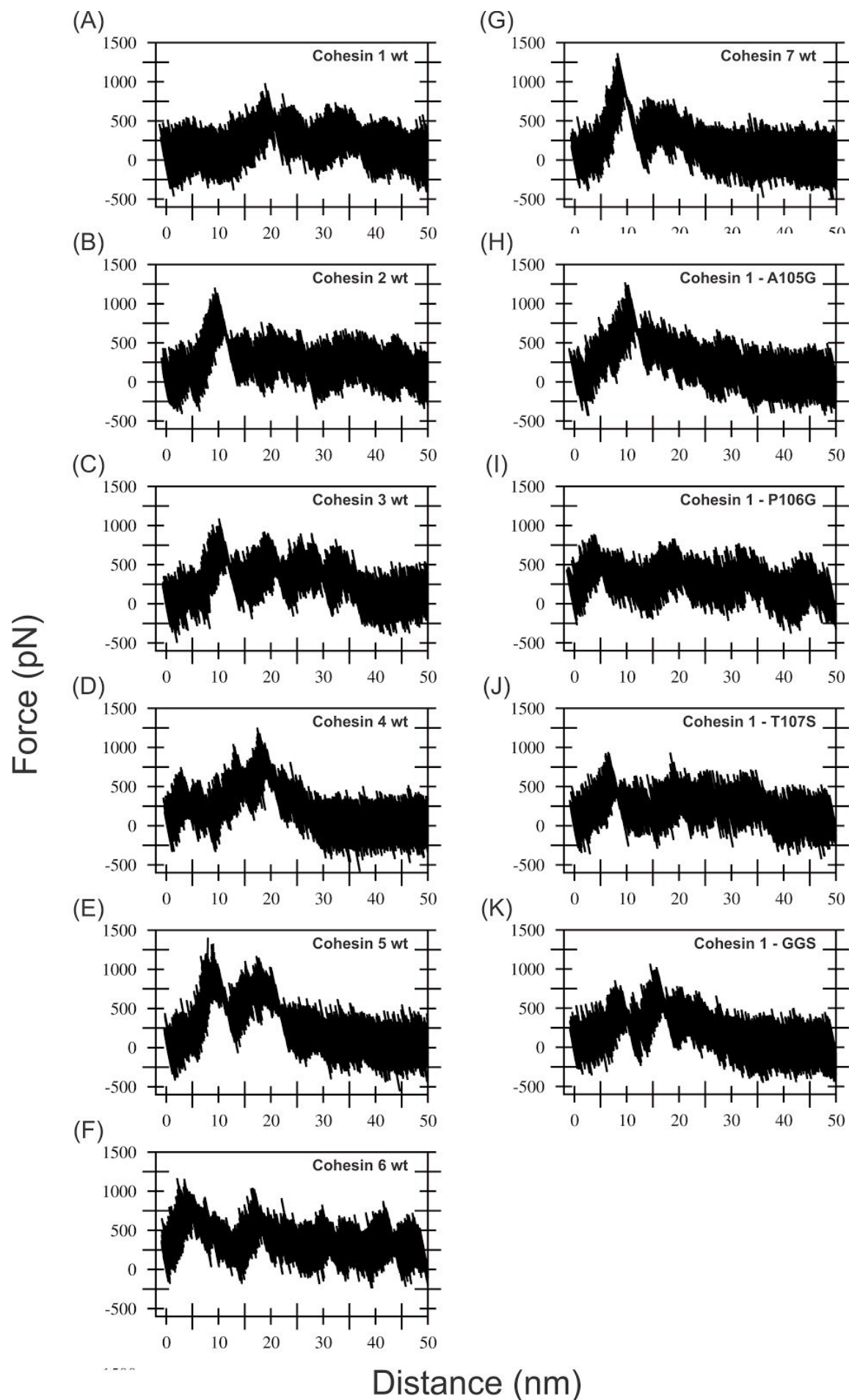
**Figure S10. Force profile during pulling simulations.** For the first 50 nm of pulling, the plots present a representative simulation (one of the replicas) for each system studied.

| | Coh1 | Coh2 | Coh3 | Coh4 | Coh5 | Coh6 | Coh7 |
|---|---|---|---|---|---|---|---|
| Coh1 | 1 | **** 4.7E-11 | ** 0.00949 | ** 0.0072 | **** 1.13E-10 | **** 1.61E-07 | **** 1.13E-10 |
| Coh2 | **** 4.7E-11 | 1 | **** 4.14E-08 | **** 1.51E-07 | ** 0.004 | **** 1.28E-05 | 0.0717 |
| Coh3 | ** 0.00949 | **** 4.14E-08 | 1 | 0.985 | **** 3.5E-09 | ** 0.00222 | **** 1.72E-08 |
| Coh4 | ** 0.0072 | **** 1.51E-07 | 0.985 | 1 | **** 2.04E-10 | ** 0.00233 | **** 9.36E-09 |
| Coh5 | **** 1.13E-10 | ** 0.004 | **** 3.5E-09 | **** 2.04E-10 | 1 | **** 8.55E-11 | 0.124 |
| Coh6 | **** 1.61E-07 | **** 1.28E-05 | ** 0.00222 | ** 0.00233 | **** 8.55E-11 | 1 | **** 4.15E-07 |
| Coh7 | **** 1.13E-10 | 0.0717 | **** 1.72E-08 | **** 9.36E-09 | 0.124 | **** 4.15E-07 | 1 |

*: $p < 0.05$     **: $p < 0.01$     ***: $p < 0.001$     ****: $p < 0.0001$

**Figure S11. Statistical significance between the simulated unfolding forces of all wild type cohesins.** P-values were calculated using the Kolmogorov-Smirnov test.
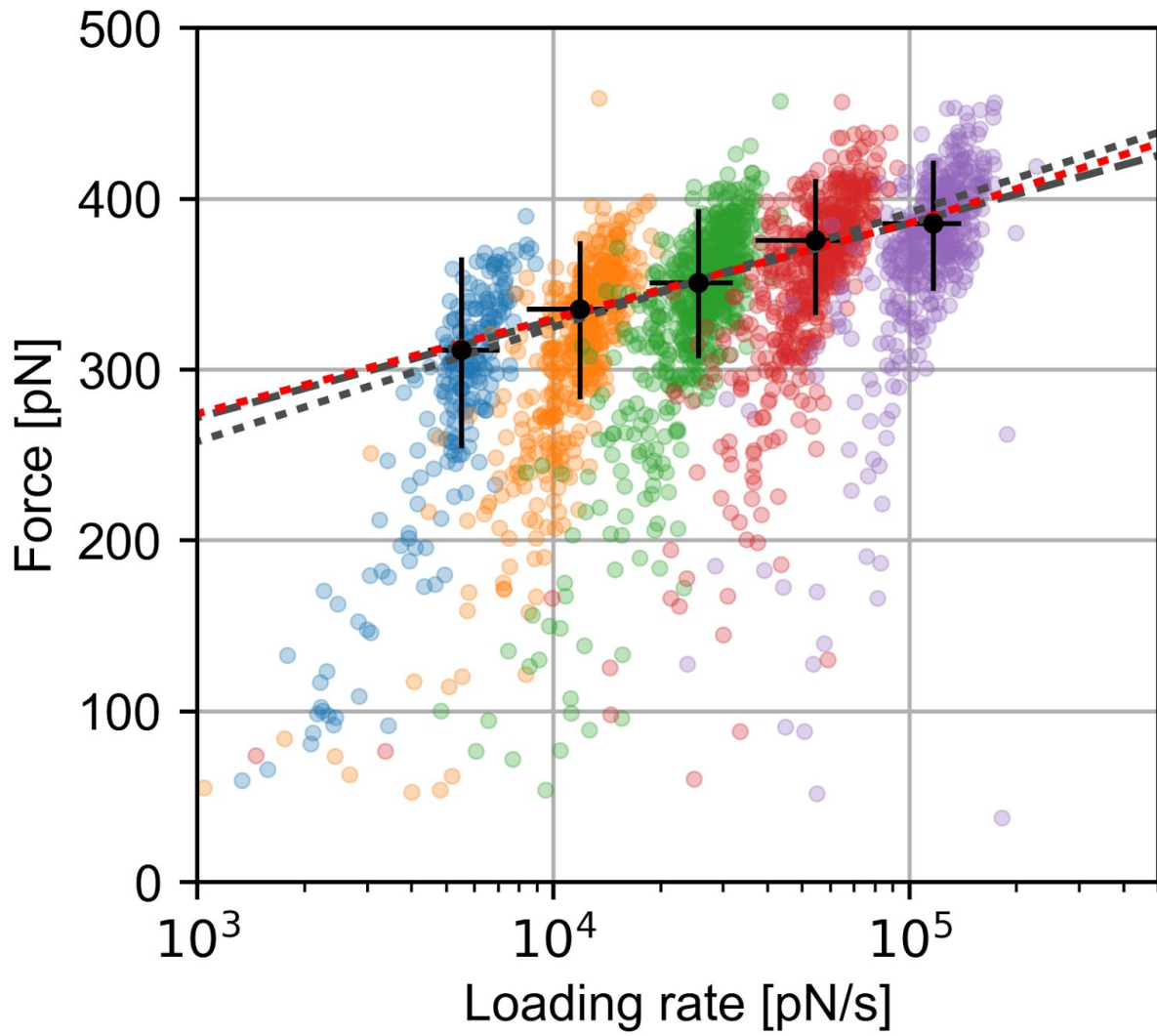
**Figure S12. Experimental dynamic force spectrum for unfolding events of cohesin 3.** Varicolored points represent rupture force/loading rate data from an experiment with 5 different pulling speeds. Black points represent the most probable rupture force/loading rate of each pulling speed obtained from kernel density estimates. Error bars represent the full width at half maximum. Gray lines represent least-squares fits of the Bell-Evans model to the experimental, and to both the experimental and the simulation data with fitting parameters (±SD) $\Delta x=0.17 \pm 0.012$ nm, $k_0=6.7 \cdot 10^{-4} \pm 6.3 \cdot 10^{-4}$ s$^{-1}$ and $\Delta x=0.14 \pm 0.0015$ nm, $k_0=4.9 \cdot 10^{-3} \pm 8.9 \cdot 10^{-4}$ s$^{-1}$, respectively. The red dotted line represents a least-squares fit of the DHS model to both the combined experimental and the simulation data with fitting parameters (±SD) of $\Delta x=0.19 \pm 0.024$ nm, $k_0=1.4 \cdot 10^{-4} \pm 2 \cdot 10^{-4}$ s$^{-1}$ and $\Delta G=60 \pm 13$ k$_B$T. In this range of loading rates the Bell-Evans fit through the experimental data falls along the DHS fit.
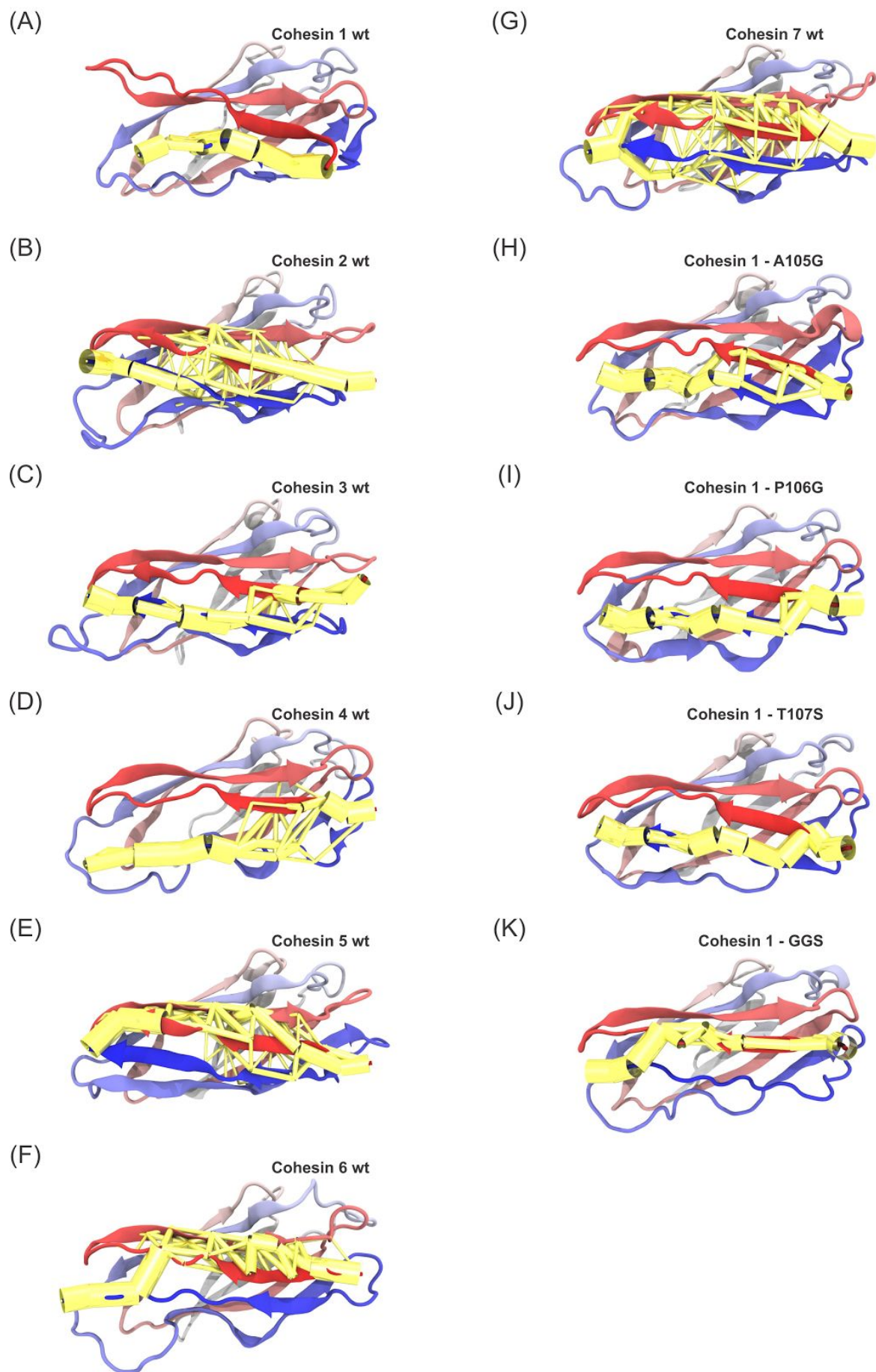
**Figure S13. Network-based force propagation analysis calculated using Pearson correlation**. Suboptimal force paths were calculated using VMD during the first 5nm of pulling simulation. Note that for all the systems beta-strands A,B and I are the main regions involved in the force propagation.
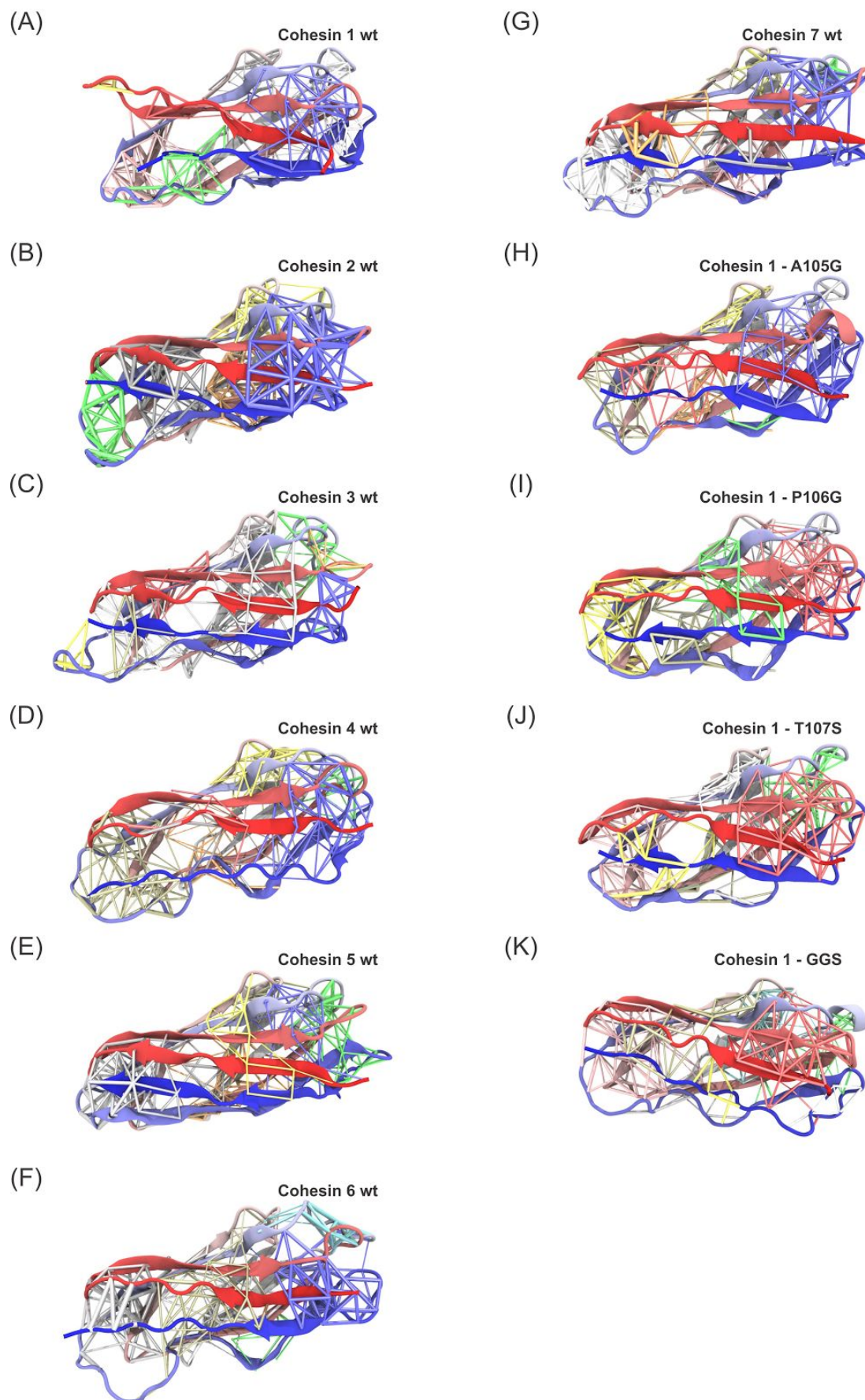
**Figure S14. Network-based community analysis calculated using generalized correlation**. Communities were calculated using VMD during the first 5nm of pulling simulation. Each color represents a different community. Colors of the communities in different systems are not related, and should not be compared as being the same community in different systems. Thickness of the network represents the *log* of the normalized correlation value. Thick connections represent highly correlated regions.

**Figure S15.** High resolution version of the sequence alignment of the C-terminal end of the hanging cohesins 1-3.
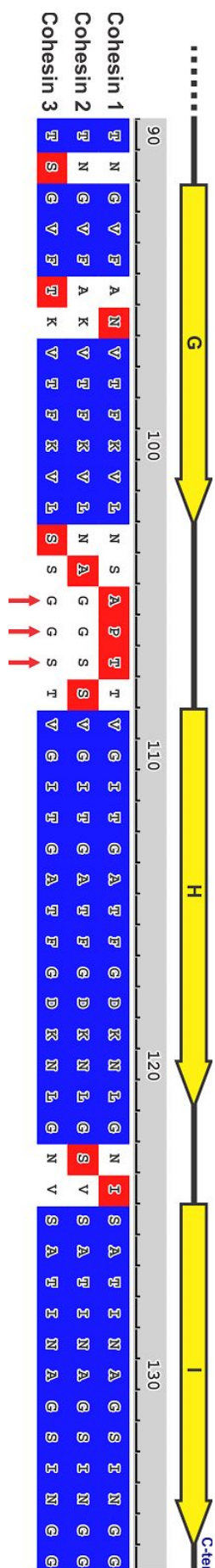
|       | Coh1 | A105G | P106G | T107S | GGS |
|-------|------|-------|-------|-------|-----|
| Coh1  | 1 | *<br>0.0333 | 0.523 | 0.714 | ***<br>0.000295 |
| A105G | *<br>0.0333 | 1 | **<br>0.00128 | 0.124 | **<br>0.00999 |
| P106G | 0.523 | **<br>0.00128 | 1 | 0.414 | ****<br>1.33E-05 |
| T107S | 0.714 | 0.124 | 0.414 | 1 | ***<br>0.000235 |
| GGS   | ***<br>0.000295 | **<br>0.00999 | ****<br>1.33E-05 | ***<br>0.000235 | 1 |

*: p < 0.05    **: p < 0.01    ***: p < 0.001    ****: p < 0.0001

**Figure S16. Statistical significance between the simulated unfolding forces Cohesin 1 and its four mutants.** P-values were calculated using the Kolmogorov-Smirnov test.

# Protein Sequences

ybbR-tag - linker and additional residues - Cohesin - linker - XDoc3

MGTDSLEFIASKLALEVLFQGPLQHHHHHHPWTSAS

Cohesin 1
TGFTVNVDSVNGNVGEQIVVPVSFANVPSNGVSTADMTITYDSSKLEYVSGAAGSIVTNPTVNFGINKEA
DGKLKVLFLDYTMSTGYISTNGVFANVTFKVLNSAPTTVGITGATFGDKNLGNISATINAGSINGG

Cohesin 2
TGFTVNVDSVNGNVGEQIVVPVSFANVPSNGISTADMTITYDSSKLEYVSGDAGSIVTNPTVNFGINKETD
GKLKVLFLDYTMSTGYISTNGVFAKVTFKVLNAGGSSVGITGATFGDKNLGSVSATINAGSINGG

Cohesin 3
TGFTVSVDSVNGNVGEQIVIPVSFANIPANGISTADMTITYDSSKLEYVSGVPGSIVTNPDVNFGINKETDG
KLKVLFLDYTMSTGYISTSGVFTKVTFKVLSSGGSTVGITGATFGDKNLGNVSATINAGSINGG

Cohesin 4
NAMAVAVGAVQGGVGETVTVPVTMTKVPTTGVSTADFTVTYDATKLEYVSGAAGSIVTNPDVNFGINKEA
DGKIKVLFLDYTMATEYISKDGVFANLTFKIKSTAAAGTTAAVGIAGTATFGDSALKPITAVITDGKVEII

Cohesin 5
KAMKVVIANVSGNAGSEVVVPVSIEGVSANGVSAADFTITYDATKLDYVSGAAGSIVKNPDVNFGINKEAD
GKLKVLFLDYTMATEYISADGIFANLTFKIKSTAVNGDVAAISKSGTATFGDKNLGPISAVIKDGSVTVG

Cohesin 6
TGFNLSIDTVEGNPGSSVVVPVKLSGISKNGISTADFTVTYDATKLEYISGDAGSIVTNPGVNFGINKESDG
KLKVLFLDYTMSTGYISTDGVFANLNFNIKSSAAIGSKAEVSISGTPTFGDSTLTPVVAKVTNGAVNVV

Cohesin 7
NAFKVSIDTVKAATGTQVVVPVSFVNVPATGISTTDMTITYDATKLQYVSGDAGSIVTNPGVNFGINKEAD
GKLKVLFLDYTMTTQYISEDGVFANVTFKVIGTDGLAAVNAEDATFGDSSLSPVTASVVNGGVNIG

Cohesin 1 A105G
TGFTVNVDSVNGNVGEQIVVPVSFANVPSNGVSTADMTITYDSSKLEYVSGAAGSIVTNPTVNFGINKEA
DGKLKVLFLDYTMSTGYISTNGVFANVTFKVLNS**G**PTTVGITGATFGDKNLGNISATINAGSINGG

Cohesin 1 P106G
TGFTVNVDSVNGNVGEQIVVPVSFANVPSNGVSTADMTITYDSSKLEYVSGAAGSIVTNPTVNFGINKEA
DGKLKVLFLDYTMSTGYISTNGVFANVTFKVLNSA**G**TTVGITGATFGDKNLGNISATINAGSINGG

Cohesin 1 T107S
TGFTVNVDSVNGNVGEQIVVPVSFANVPSNGVSTADMTITYDSSKLEYVSGAAGSIVTNPTVNFGINKEA
DGKLKVLFLDYTMSTGYISTNGVFANVTFKVLNSAP**S**TVGITGATFGDKNLGNISATINAGSINGG

Cohesin 1 A105G P106G T107S
TGFTVNVDSVNGNVGEQIVVPVSFANVPSNGVSTADMTITYDSSKLEYVSGAAGSIVTNPTVNFGINKEA
DGKLKVLFLDYTMSTGYISTNGVFANVTFKVLNS**GGS**TVGITGATFGDKNLGNISATINAGSINGG

VVPNTVTSAVKTQYVEIESVDGFYFNTEDKFDTAQIKKAVLHTVYNEGYTGDDGVAVVLREYESEPVDITA
ELTFGDATPANTYKAVENKFDYEIPVYYNNATLKDAEGNDATVTVYIGLKGDTDLNNIVDGRDATATLTYY
AATSTDGKDATTVALSPSTLVGGNPESVYDDFSAFLSDVKVDAGKELTRFAKKAERLIDGRDASSILTFYT
KSSVDQYKDMAANEPNKLWDIVTGDAEEE

Coh3 - linker - ddFLN4 - linker and additional residues - ybbR

MGTALTDRGMTYDLDPKDGSSAATKPVLEVTKKVFDTAADAAGQTVTVEFKVSGAEGKYATTGYHIYWD
ERLEVVATKTGAYAKKGAALEDSSLAKAENNGNGVFVASGADDDFGADGVMWTVELKVPADAKAGDVY
PIDVAYQWDPSKGDLFTDNKDSAQGKLMQAYFFTQGIKSSSNPSTDEYLVKANATYADGYIAIKAGEPGS
VVPSTGSADPEKSYAEGPGLDGGESFQPSKFKIHAVDPDGVHRTDGGDGFVVTIEGPAPVDPVMVDNG
DGTYDVEFEPKEAGDYVINLTLDGDNVNGFPKTVTVKPAPGSELKLPRSRHHHHHHGSLEVLFQGPDSL
EFIASKLA

# References

(1)  Gibson, D. G.; Young, L.; Chuang, R.-Y.; Venter, J. C.; Hutchison, C. a.; Smith, H. O.; Iii, C. A. H.; America, N. *Nat. Methods* **2009**, *6* (5), 343.
(2)  Studier, F. W. *Protein Expr. Purif.* **2005**, *41* (1), 207.
(3)  Jobst, M. A.; Schoeler, C.; Malinowska, K.; Nash, M. A. *J. Vis. Exp.* **2013**, No. 82, 1.
(4)  Otten, M.; Ott, W.; Jobst, M. a.; Milles, L. F.; Verdorfer, T.; Pippig, D. a.; Nash, M. a.; Gaub, H. E. *Nat. Methods* **2014**, No. september, 1.
(5)  Zimmermann, J. L.; Nicolaus, T.; Neuert, G.; Blank, K. *Nat. Protoc.* **2010**, *5* (6), 975.
(6)  Shimizu, Y.; Kanamori, T.; Ueda, T. *Methods* **2005**, *36* (3), 299.
(7)  Hutter, J. L.; Bechhoefer, J. *Rev. Sci. Instrum.* **1993**, *64* (7), 1868.
(8)  Milles, L. F.; Bayer, E. A.; Nash, M. A.; Gaub, H. E. *J. Phys. Chem. B* **2016**, acs.jpcb.6b09593.
(9)  Siggia, E. D.; Bustamante, C.; Marko, J. F.; Smith, S. B. *Science* **1994**, 5.
(10)  Schwaiger, I.; Kardinal, A.; Schleicher, M.; Noegel, A. A.; Rief, M. *Nat. Struct. Mol. Biol.* **2004**, *11* (1), 81.
(11)  Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. *Biophys. J.* **1997**, *72* (4), 1568.
(12)  Evans, E.; Ritchie, K. *Biophys. J.* **1997**, *72* (4), 1541.
(13)  Dudko, O. K.; Hummer, G.; Szabo, A. *Phys. Rev. Lett.* **2006**, *96* (10), 108101.
(14)  Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. *Nucleic Acids Res.* **2017**, *45* (D1), D37.
(15)  Ding, S.-Y.; Bayer, E. A.; Steiner, D.; Shoham, Y.; Lamed, R. *J. Bacteriol.* **1999**, *181* (21), 6720.
(16)  Gish, W.; States, D. J. *Nat. Genet.* **1993**, *3* (3), 266.
(17)  Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. In *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*; Rossmann, M. G., Arnold, E., Eds.; International Tables for Crystallography; Springer Netherlands, 2006; pp 675–684.
(18)  Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14* (1), 33.
(19)  Roberts, E.; Eargle, J.; Wright, D.; Luthey-Schulten, Z. *BMC Bioinformatics* **2006**, *7*, 382.
(20)  Thompson, J. D.; Gibson, T. J.; Higgins, D. G. *Curr. Protoc. Bioinformatics* **2002**, *Chapter 2*, Unit 2.3.
(21)  Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M.-Y.; Pieper, U.; Sali, A. *Curr. Protoc. Protein Sci.* **2007**, *Chapter 2*, Unit 2.9.
(22)  Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26* (2), 283.
(23)  MacArthur, M. W.; Laskowski, R. A.; Thornton, J. M. *Curr. Opin. Struct. Biol.* **1994**, *4* (5), 731.
(24)  Ribeiro, J. V.; Bernardi, R. C.; Rudack, T.; Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Schulten, K. *Sci. Rep.* **2016**, *6*, 26536.

(25)    Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26* (16), 1781.

(26)    Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257.

(27)    MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102* (18), 3586.

(28)    Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79* (2), 926.

(29)    Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98* (12), 10089.

(30)    Frishman, D.; Argos, P. *Proteins* **1995**, *23* (4), 566.

(31)    Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (16), 6620.

(32)    Schoeler, C.; Bernardi, R. C.; Malinowska, K. H.; Durner, E.; Ott, W.; Bayer, E. A.; Schulten, K.; Nash, M. A.; Gaub, H. E. *Nano Lett.* **2015**, *15* (11), 7370.

(33)    Ribeiro, A. A. S. T.; Ortiz, V. *Chem. Rev.* **2016**, *116* (11), 6488.

(34)    Glykos, N. M. *J. Comput. Chem.* **2006**, *27* (14), 1765.

(35)    Lange, O. F.; Grubmüller, H. *Proteins* **2006**, *62* (4), 1053.

(36)    Schoeler, C.; Malinowska, K. H.; Bernardi, R. C.; Milles, L. F.; Jobst, M. a.; Durner, E.; Ott, W.; Fried, D. B.; Bayer, E. a.; Schulten, K.; Gaub, H. E.; Nash, M. a. *Nat. Commun.* **2014**, *5*, 5635.

**Table of contents graphic:**