Extended data

The Mimivirus 1.2 Mb dsDNA genome is elegantly organized into a nuclear-like weapon

Alejandro Villalta[a#], Emmanuelle R. J. Quemin[b#], Alain Schmitt[a#], Jean-Marie Alempic[a], Audrey Lartigue[a], Vojtěch Pražák[c], Lucid Belmudes[d], Daven Vasishtan[c], Agathe M. G. Colmant[a], Flora A. Honoré[a], Yohann Couté[d], Kay Grünewald[b,c], Chantal Abergel[a*]

Extended data Methods

Extraction and purification of the Mimivirus genomic fibre

The genomic fibre was extracted from 12 mL of purified Mimivirus reunion virions at 1.5 x 10^{10} particles/mL, split into 12x1 mL samples and treated in parallel. Trypsin (Sigma T8003) in 40 mM Tris-HCl pH 7.5 buffer was added at a final concentration of 50 µg/mL and the virus enzyme mix was incubated for 2h at 30°C in heating dry block (Grant Bio PCH-1). DTT was then added at a final concentration of 10mM and incubated at 32°C for 16h. Finally, 0.001% Triton X-100 was added to the mix and incubated for 4h at 32°C. Each tube was vortexed for 20 s with 1.5 mm diameter stainless steel beads (CIMAP) to separate the fibres from the viral particles and centrifuged at 5,000 g for 15 min. to pellet the opened capsids. The supernatant was recovered, and the fibres were concentrated by centrifugation at 15,000 g, for 4h at 4°C. Most of the supernatant was discarded leaving 12x~200 µL of concentrated fibres that were pooled and layered on top of ultracentrifuge 4mL tubes (polypropylene centrifuge tubes beckman coulter) containing a discontinuous sucrose gradient (40%, 50%, 60%, 70% w/v in 40 mM Tris-HCl pH 7.5 buffer). The gradients were centrifuged at 200,000 g for 16h at 4 °C. Since no visible band was observed, successive 0.5 mL fractions were recovered from the bottom of the tube, the first one supposedly corresponding to 70% sucrose. Each fraction was dialyzed using 20 kDa Slide-A-Lyzers (ThermoFisher) against 40 mM Tris-HCl pH 7.5 to remove the sucrose. These were further concentrated by centrifugation at 15,000 g, 4°C for 4h and most of the supernatant was removed, leaving ~100 µL of sample at the bottom of each tube. Each fraction was imaged by negative staining transmission electron microscopy (TEM). For proteomic analysis, an additional step of concentration was performed by speedvac (Savant SPD131DDA, Thermo Scientific).

Negative staining

300 mesh ultra-thin carbon-coated copper grid (Electron Microscopy Sciences, EMS) were prepared for negative staining by adsorbing 4-7 µL of the sample for 3 min., followed by two washes with water before staining for 2 min. in 2% uranyl acetate. The grids were observed either on a FEI Tecnai G2 microscope operated at 200 keV and equipped with an Olympus Veleta 2k camera (IBDM microscopy platform, Marseille, France); a FEI Tecnai G2 microscope operated at 200 keV and equipped with a Gatan OneView camera (IMM, microscopy platform, France) or a FEI Talos L120c operated at 120 keV and equipped with a Ceta 16M camera (CSSB multi-user cryo-EM facility, Germany).

Agarose gel electrophoresis to assess the presence of DNA into the fibre

Purified Mimivirus reunion virions were opened following the method described earlier. Crude supernatant enriched in genomic fibres were digested with various enzyme combinations. Proteinase K was used at a final concentration of 1 mg/mL and incubated for 30 min at 55 °C. DNase was used at a final concentration of 200 µg/mL in combination with $MgCl_2$ at a concentration of 5 mM and incubated for 30 min at 37 °C. RNase H (NEB) was added to a final activity of 2.5 U per sample and incubated for 30 min at 37 °C. One sample was mock digested by heating 30 min at 37 °C then 30 min at 55 °C with no enzyme and one sample was kept on ice with no enzyme as negative controls. All samples were loaded on a 2% agarose gel and stained using ethidium bromide after migration. The same treatment was then applied to the purified fibre by adding 0.5 µl of PK (Takara ST 0341) to 10 µL of sample solution (1mg/mL final concentration) and incubating the reaction mix at 55°C for 30 min. DNase treatment was done by adding DNase (Sigma 10104159001) and $MgCl_2$ to a final concentration of 0.18 mg/mL and 2.3 mM, respectively in 10 µL of sample and incubated at 37°C for 30 min prior to PK treatment. For RNase treatment, RNase was added to 10 µL of sample solution to a final concentration of 0.95 mg/mL A (Sigma SLBW2866) and incubated at 37°C for 30 min prior to PK treatment. All the samples were then loaded on a 2% agarose gel and stained with ethidium bromide after migration.

Single-particle analysis by cryo-EM

Sample preparation

For single-particle analysis, 3 µL of the purified sample were applied to glow-discharged Quantifoil R 2/1 Cu G200F1 grids, blotted for 2 s. using a Vitrobot Mk IV (Thermo

Scientific) and the following parameters: 4°C, 100% humidity, blotting force 0, and plunge frozen in liquid ethane/propane cooled to liquid nitrogen temperature.

Data acquisition

The grids were loaded in a Titan Krios (Thermo Scientific) microscope operated at 300 keV and equipped with a K2 direct electron detector and a GIF BioQuantum energy filter (Gatan). 7,656 movie frames were collected using the EPU software (Thermo Scientific) at a nominal magnification of x130,000 with a pixel size of 1.09 Å and a defocus range of -1 to -3 µm. Each movie fractionated into 40 frames, was collected using EPU (Thermo Scientific) for a total exposure time of 8 s. with a dose of 7.5 electrons per physical pixel per second and a total dose of 50.6 e/Å² with a 20 eV slit for the GIF in zero-loss mode (Extended data Table 3).

Sorting and clustering of Relion 2D classification

All movie frames were aligned using MotionCor2[1] in Relion 3.0[2] and used for contrast transfer function (CTF) parameter calculation with CTFFIND-4.1[3]. Helical particles were manually picked with Relion 3.0[2,4], then extracted with different box sizes (450, 500, 700 pixels) to get a better estimate of the initial helical parameters. Particles were subjected to reference-free 2D classification in Relion 3.1.0[2,4], where multiple unwinding states of the fibre were identified (Extended data Fig. 3 & 4). Additional cluster analysis of the 200 initial 2D classes provided by Relion led to 3 homogeneous clusters corresponding to different unwinding states, as illustrated in Fig. 2 and Extended data Fig. 4. The clustering strategy, implemented in python language, using mainly numpy[5] and scikit-learn[6] libraries, was applied in 2 steps. First, a few main clusters were identified by applying a DBSCAN[7] clustering algorithm on the previously estimated fibre external width values (W1). The widths values, estimated by adjusting a parameterized cosine model on each 2D stack, range from roughly 290 Å to 330 Å. Then, each main cluster was subdivided into several sub-clusters by applying a KMEANS[8] clustering algorithm on a pairwise similarity matrix. This similarity metric was based on a Fast Fourier Transform-based implementation (FFT) of a 2D image correlation scheme, invariant to bounded image shifts, as well as left-right and up-down mirroring. The number of sub-clusters was manually chosen by visual inspection. For the most compacted main classes, the number of sub-clusters were small (2 sub-clusters, but only one homogeneous, Cl1, for the most compacted class and n=1 sub-clusters for the intermediate

class, Cl2), whereas the number of sub-clusters was higher (n=5) for the most unwound main class (from green to purple in Fig. 2), highlighting its overall heterogeneity going from unwound to unfolded ribbons.

Model-based cryo-EM data processing and 3D reconstruction

The different compaction states allowed us to identify a helical structure composed by a unique building block, directly from visual inspection of the most unwound 2D classes, of approximate dimensions of 90 Å x 45 Å (Extended data Fig. 3 to 5). Also, initial helical parameters could be estimated directly from measurements on the 2D classes for the compact state (rise 36.8 Å and twist 19.3°) and unwound state (rise 30.4 Å and twist 17°). These initial observations allowed us to build a theoretical 3D-model in Blender 2.81, using Python scripting, by assembling the composing element, so that their positions follow a helical geometry defined by a set of parameters {Radius, N-start, Rise, Twist, θ-start}, based on the initial helical parameters and building block deciphered from the 2D classes.

The first refined model parameters were used for 3D refinement in Relion 3.1.0[2,4] on separated datasets of 19,526 compact particles (box size 500 pixels) and 8,648 unwound particles (box size 500 pixels) with a featureless cylinder as reference (diameter of 300 Å for compacted classes and 330 for unwound classes). The helical parameters of the theoretical model in conjunction with the geometry of its constituting elements (Extended Fig. 5) were iteratively refined by adjusting the theoretical 3D model to Relion reconstructed map, or by comparing theoretical 2D classes to experimental ones, and applying in Relion refinements the new parameters suggested by the theoretical compaction model (Fig. 2, Extended data Fig. 4). The theoretical compaction model provided the position and the rotation of all composing building blocks for any compaction states (specifically for the three main compaction states identified by 2D classification: Cl1, Cl2 and Cl3) corresponding to a unique "width 1: W1" parameter value (fibre external width). The theoretical modelling not only provided a way to reject aberrant solutions or identifying candidates for various compaction states (to be confirmed or rejected by Relion 3D reconstruction), but it also gave us a mean to smoothly adjust the model with a flexible number of parameters, such as N-start (>1) and θ-start, which are not directly accessible as helical constraints in Relion.

We performed further 3D classification and 3D refinement, imposing alternatively the refined parameters corresponding to a 6-start or 5-start in Relion 3.1.0[2,4] in order to obtain more

homogeneous subsets for each helical symmetry. This process did not perfectly separate 6-start and 5-start fibres particles as from 19,526 particles composing the most compacted fibre sub-cluster, 13,252 particles were selected from the 3D classification as the 5-start fibre subset and similarly 12,294 particles were selected as the 6-start fibre subset. Similarly, from the 8,648 particles in the unwound state, 5,036 particles were selected as the 5-start fibre subset and 6,320 particles as the 6-start fibre subset.

Protein modelling

Ultimately, the 3D classifications allowed us to precisely identify the parameters of an equivalent 1-start helix parameter, for both the compacted and unwound 5- and 6-start fibres (Extended data Table 1). This led to 3D reconstructions resolved enough to easily identify in the overall structure the building block of the fibre as the R135 dimeric structure[9] (PDB 4Z24) which was fitted into the maps using UCSF Chimera 1.13.1[10]. The qu_143 model corresponds to the R135 structure in which the 12 C-terminal residues were removed (99% identity between the two proteins over 690 amino-acids) (Extended Fig. 8). The model of qu_946 was obtained using SWISS-MODEL[11]. Further 3D refinement was performed with the 3D classification reconstruction lowpass filtered to 15 Å as reference and, the 5-start compacted state map was resolved enough to identify secondary structure elements and better fit the qu_946 model into the electron density. N-terminal residues were manually built using the extra density available in the cryo-EM map of the 5-start compact form and was further refined using the Real-space refinement program in PHENIX 1.18.2[12]. Validation was also performed into PHENIX 1.18.2[12] using the comprehensive validation program (Extended data Table 4). Due to the medium resolution of the structure, all measurements were performed by taking into account only the backbone atoms of the protein model.

DNA modelling

3D reconstructions corresponding to the most compacted states of the fibre suggested the presence of DNA strings early on, internally lining the proteinaceous shell. A theoretical model composed by 5 (or 6) helical dsDNA segments was built using rectilinear DNA in A-form (twist 32.7° and rise 2.548 Å) or B-form (twist 36.0° and rise 3.375 Å). These dsDNA segments were retrieved from the web3dna platform[13] and bent to follow the helical parameters of the protein shell. The diameter of the 5-start (or 6-start) DNA helix was first estimated from the 2D classes by fitting a sinusoidal model corresponding to the projection of

the 3D helix on the 2D tangential imaging plane and further refined by adjusting a parameterized 3D atomic model on the most refined volumetric maps obtained by 3D refinement in Relion as described in the previous paragraph. For the 5- and 6-start most compacted states (Cl1), the DNA helix outer diameter was estimated to be 132 Å and 144.5 Å, respectively. So far, the best fitted dsDNA model is based on a theoretical adjustment of the A-form DNA built on the hypothesis that the periodicity of the protein shell is constraining the periodicity of the dsDNA strings. For instance, since the 5-start proteinaceous shell is composed of repeating protein units every 49.5Å along its helical path at the DNA/protein interaction radius, we selected as the best candidate the A-form DNA compressed by a factor of 0.865 (leading to a final pitch of 24.75Å, showing 2 periods within each protein unit). Finally, each of the 5 (or 6) DNA strands was independently rotated and translated in the electronic map to optimise the phase and orientation according to the periodical contacts observed in the 3D maps between the DNA and the protein shell. Additional analysis would be needed to refine the DNA modelling parameters and validate its precise conformation inside the helical protein shell.

Bubblegram analysis

Samples were prepared as described for single-particle analysis and subjected to increasing dose of electrons in a Titan Krios (Thermo Scientific) microscope operated at 300 keV and equipped with a K3 direct electron detector and a GIF BioQuantum energy filter used in zero-loss mode with a 20 eV slit (Gatan). Data were recorded using SerialEM[14] at a nominal magnification of x81,000, a pixel size of 1.09 Å and a dose of 15 e-/pixel/s (Extended data Fig. 2 & Table 3). In a typical experiment, 12 to 15 exposures of 6 s applying a total dose of 75 e-/Å² per exposure were collected using frames of 0.1 s. and pre-aligned in SerialEM.

Cryo-electron tomography

Sample preparation

For cryo-ET of the genomic fibre, samples were prepared as described above for single-particle analysis except that 5 nm gold fiducials (UMC, Utrecht) were added to the sample right before plunge freezing at a ratio of 1:2 (sample : fiducials).

Data acquisition

Tilt-series were acquired using SerialEM[14] in a Titan Krios (Thermo Scientific) microscope operated at 300 keV and equipped with a K3 direct electron detector and a GIF BioQuantum energy filter used in zero-loss mode with a 20 eV slit (Gatan). We used the dose-symmetric tilt-scheme[15] starting at 0° with a 3° increment to +/-60° at a nominal magnification of x64,000, a pixel size of 1.8 Å and a total dose of 150 e-/Å² over the 41 tilts (*i.e.* ~3.7 e-/Å²/tilt for an exposure time of 0.8 s fractionated into 0.2 s per frame, so 4 frames that were pre-aligned using SerialEM, Extended data Table 3).

Data processing

Tilt series outputted from the microscope were aligned and reconstructed using the IMOD software package[16] with fiducial alignments, CTF correction and back-projection. For visualization purposes and help with initial sub-volume averaging, we applied a binning of 8 and SIRT-like filtering from IMOD[16] as well as a bandpass filter (200, 1000, 0.01) from bsoft[17]. The genomic fibres were picked manually from individual tomograms and pre-processed using an in-house developed PEX script [18] that evenly sampled at 0.5-pixel intervals through a straight line defined by the user. Each particle was oriented parallel to the line, prior to iterative refinement using PEET[19] and refined separately initially. Symmetry and approximate helical parameters of individual fibre were manually assessed in UCSF Chimera 1.13.1[10] in order to sort the different conformations present in the dataset.

Mass spectrometry-based proteomic analysis of Mimivirus reunion genomic fibre

8 µL of Laemmli 5X (125 mM Tris-HCl pH 6.8, 10% SDS, 20% glycerol, 25% β-mercaptoethanol and traces of bromophenol blue) were added to 32 µL of the purified fibre in water and heated for 10 min at 95 °C. Extracted proteins were stacked in the top of a SDS-PAGE gel (4-12% NuPAGE, Life Technologies), stained with Coomassie blue R-250 (Bio-Rad) before in-gel digestion using modified trypsin (Promega, sequencing grade) as previously described[20]. Resulting peptides were analyzed by online nanoliquid chromatography coupled to tandem MS (UltiMate 3000 RSLCnano and Q-Exactive Plus, Thermo Scientific). Peptides were sampled on a 300 µm x 5 mm PepMap C18 precolumn and separated on a 75 µm x 250 mm C18 column (Reprosil-Pur 120 C18-AQ, 1.9 µm, Dr. Maisch) using a 60-min gradient. MS and MS/MS data were acquired using Xcalibur (Thermo Scientific). Peptides and proteins were identified using Mascot (version 2.6.0) through concomitant searches against Mimivirus reunion and classical contaminant database

(homemade) and the corresponding reversed database. The Proline software[21] was used to filter the results: conservation of rank 1 peptides, peptide score ≥ 25, peptide length ≥ 6, peptide-spectrum-match identification false discovery rate < 1% as calculated on scores by employing the reverse database strategy, and minimum of 1 specific peptide per identified protein group. Proline was then used to perform a compilation and MS1-based quantification of the identified protein groups. Intensity-based absolute quantification (iBAQ)[22] values were calculated from MS intensities of identified peptides. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[23] partner repository with the dataset identifier PXD021585 and 10.6019/PXD021585".

RNA polymerase model building

The RNA polymerase model was created based on the vaccinia virus DNA-dependent RNA polymerase complex structure[24] (PDB: 6RIC), by selecting the subunits identified in the Mass spectrometry based proteomic analysis.

Model visualization

Molecular graphics and analyses were performed with UCSF Chimera 1.13.1[10], and UCSF ChimeraX 1.1[25], developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

References:

1. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
2. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
3. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).

4. He, S. & Scheres, S. H. W. Helical reconstruction in RELION. *J. Struct. Biol.* **198**, 163–176 (2017).
5. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
6. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
7. Hahsler, M., Piekenbrock, M. & Doran, D. **dbscan** : Fast Density-Based Clustering with *R*. *J. Stat. Softw.* **91**, (2019).
8. Mannor, S. *et al.* K-Means Clustering. in *Encyclopedia of Machine Learning* (eds. Sammut, C. & Webb, G. I.) 563–564 (Springer US, 2011). doi:10.1007/978-0-387-30164-8_425.
9. Klose, T. *et al.* A Mimivirus Enzyme that Participates in Viral Entry. *Struct. Lond. Engl. 1993* **23**, 1058–1065 (2015).
10. Pettersen, E. F. *et al.* UCSF Chimera-A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
11. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
12. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. *Acta Crystallogr. Sect. Struct. Biol.* **75**, 861–877 (2019).
13. Li, S., Olson, W. K. & Lu, X.-J. Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res.* **47**, W26–W34 (2019).
14. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).

15. Hagen, W. J. H., Wan, W. & Briggs, J. A. G. Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. *J. Struct. Biol.* **197**, 191–198 (2017).

16. Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116**, 71–76 (1996).

17. Heymann, J. B. & Belnap, D. M. Bsoft: image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157**, 3–18 (2007).

18. Grange, M., Vasishtan, D. & Grünewald, K. Cellular electron cryo tomography and in situ sub-volume averaging reveal the context of microtubule-based processes. *J. Struct. Biol.* **197**, 181–190 (2017).

19. Nicastro, D. The Molecular Architecture of Axonemes Revealed by Cryoelectron Tomography. *Science* **313**, 944–948 (2006).

20. Casabona, M. G., Vandenbrouck, Y., Attree, I. & Couté, Y. Proteomic characterization of Pseudomonas aeruginosa PAO1 inner membrane. *Proteomics* **13**, 2419–2423 (2013).

21. Bouyssié, D. *et al.* Proline: an efficient and user-friendly software suite for large-scale proteomics. *Bioinforma. Oxf. Engl.* **36**, 3148–3155 (2020).

22. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

23. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

24. Hillen, H. S. *et al.* Structural Basis of Poxvirus Transcription: Transcribing and Capping Vaccinia Complexes. *Cell* **179**, 1525-1536.e12 (2019).

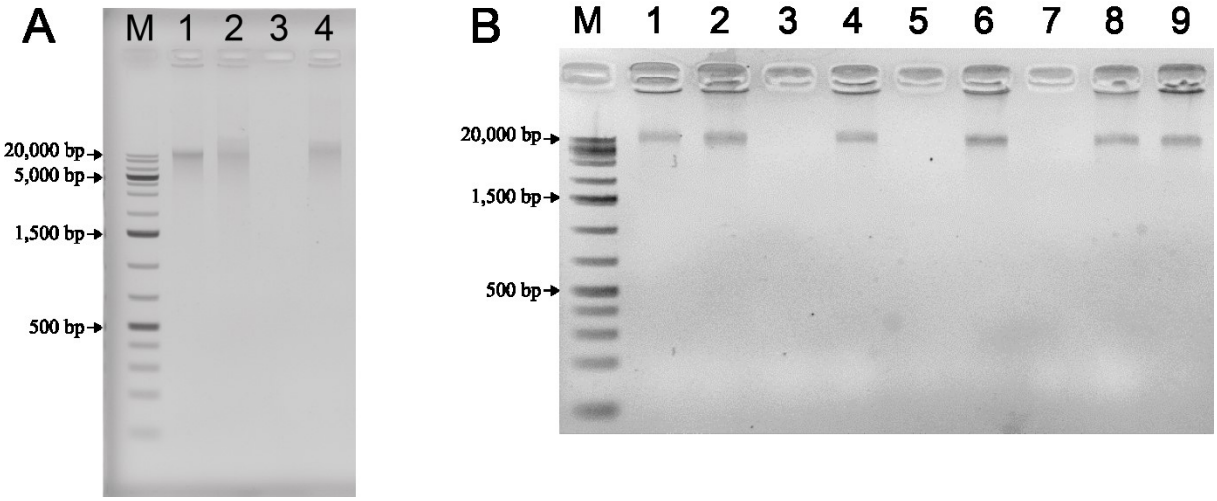
25. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis: UCSF ChimeraX Visualization System. *Protein Sci.* **27**, 14–25 (2018).

26. He, S. & Scheres, S. H. W. Helical reconstruction in RELION. *J. Struct. Biol.* **198**, 163–176 (2017).

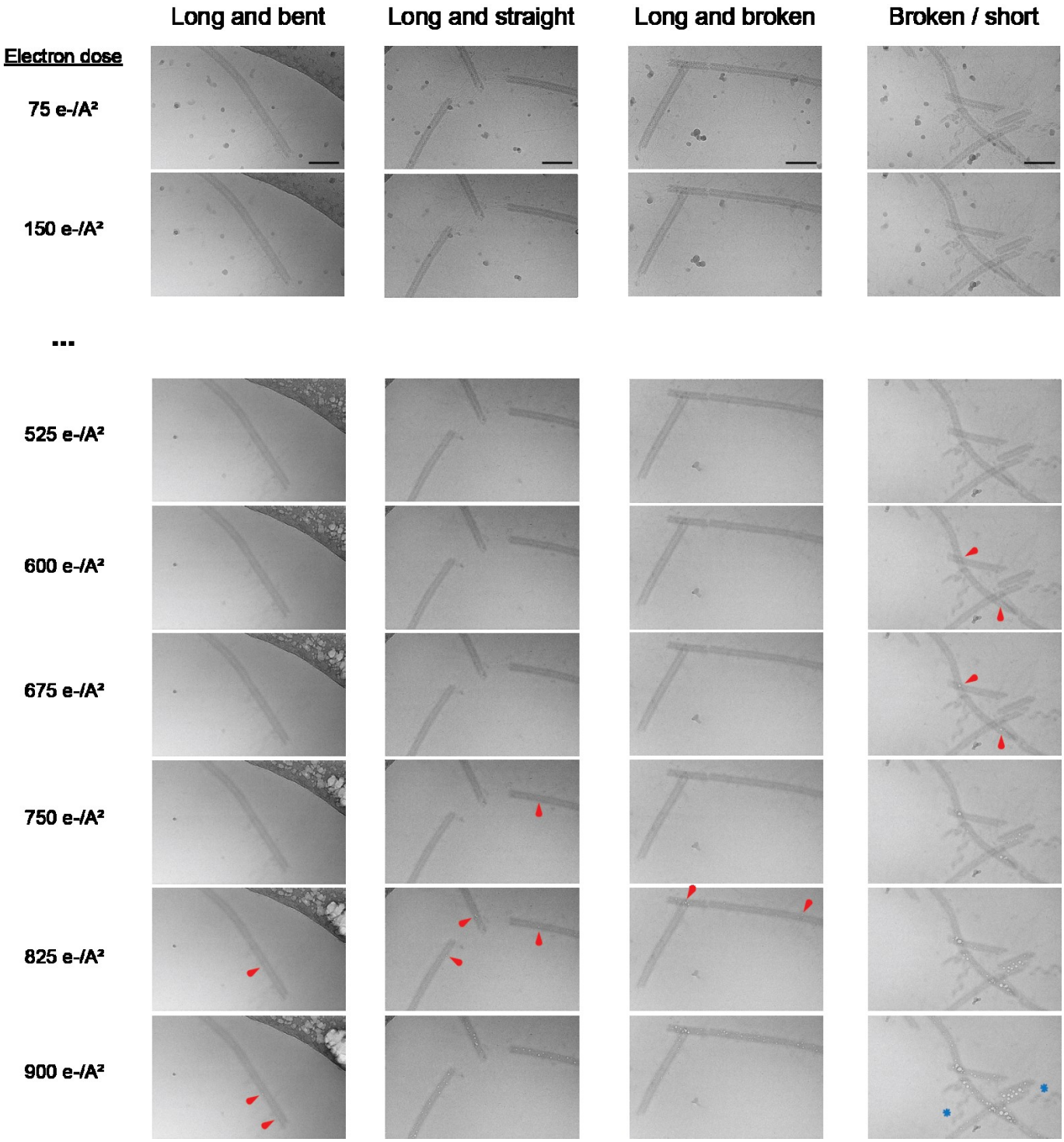
27. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10881–10890 (1988).

28. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).

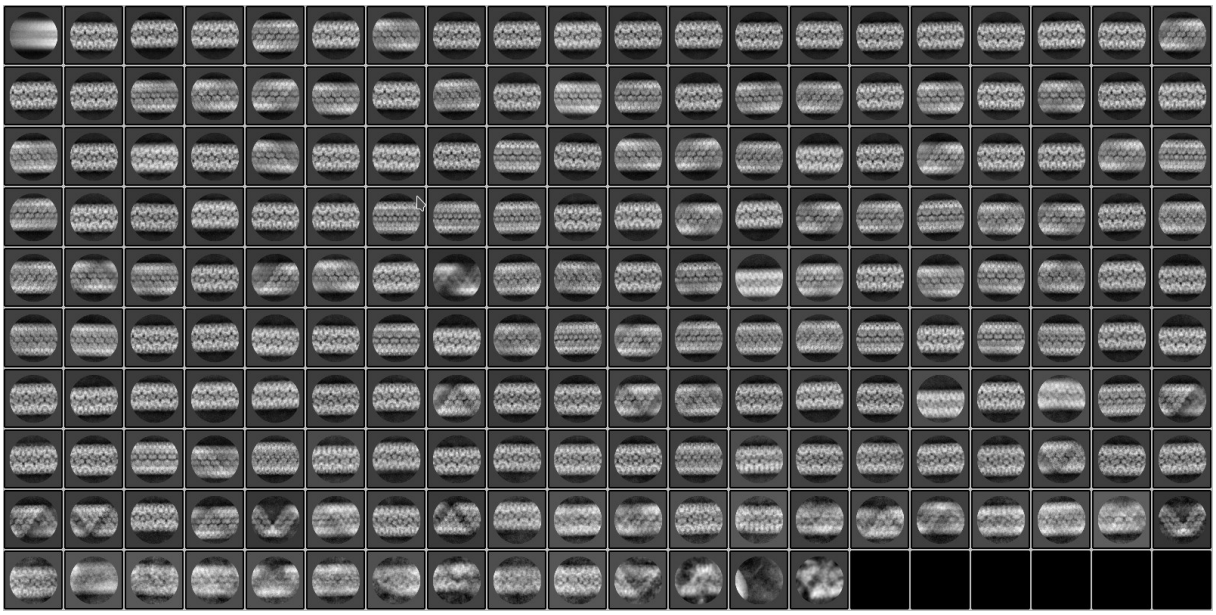
Extended data Figures



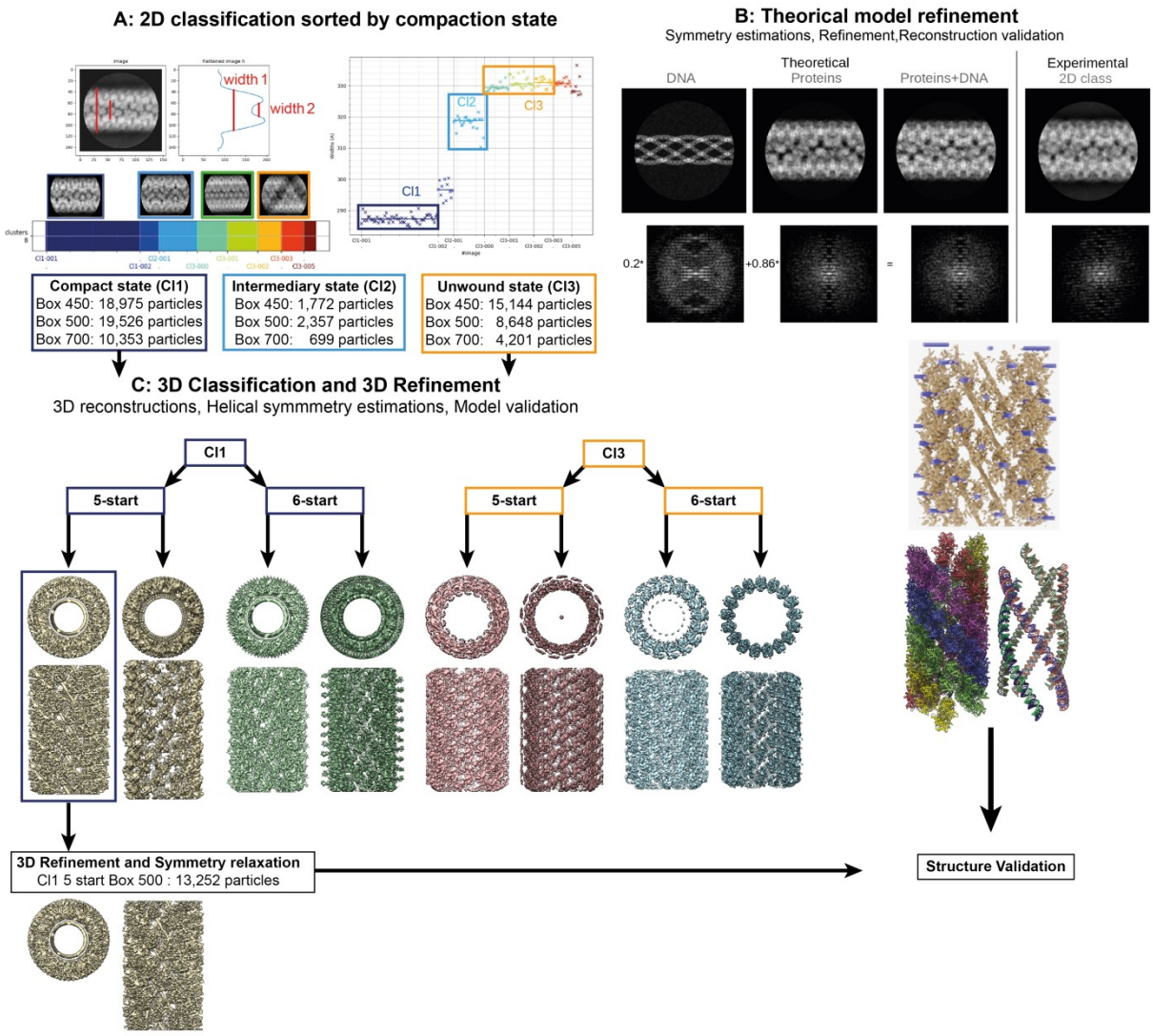
Extended data Figure 1: Agarose gel electrophoresis of the A] purified fibre: M: Molecular weight markers (1 kb DNA Ladder Plus, Euromedex). Lane 1, untreated. Lane 2: Proteinase K (PK) treated. Lane 3: DNase and PK treatment. Lane 4: RNase and PK treatment. B] Crude fibre extract: M: Molecular weight markers, Lane 1: heated with no enzyme, Lane 2: PK treated, Lane 3: DNase treated, Lane 4: RNase treated, Lane 5: PK then DNase treatment; Lane 6: PK then RNase treatment, Lane 7: DNase then PK treatment, Lane 8: RNase.then proteinase treatment, Lane 9: untreated



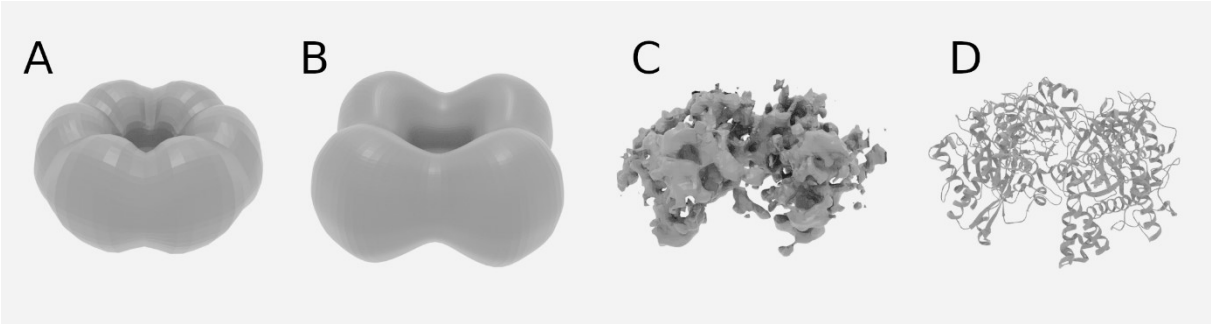
Extended data Figure 2: Bubblegrams on the Mimivirus genomic fibre. Field of view for genomic fibres either (from left to right) long and bent, long and straight, long and broken, or a mix of short and broken in low-dose after different exposures. The electron dose applied is given on the left in e-/Å². Red arrowheads indicate when and where the bubbles are first detected and further expand for some fibres as a sign of hydrogen gas trapped inside the DNA-protein complex upon protein damage. In unfolded ribbons highlighted by a blue asterisk, no bubbles are detected with a dose of up to 900 e-/Å². Scale bars, 100 nm.



Extended data Figure 3: 2D classification. 200 2D classes were obtained after reference-free 2D classification of fibres acquired for single-particle analysis and extracted with a box-size of 500 pixels in Relion 3.1.0 after motion correction, CTF estimation and manual picking (Extended data methods). The 2D classes are representative of the different compaction states of the genomic fibre observed in our highly heterogeneous dataset.

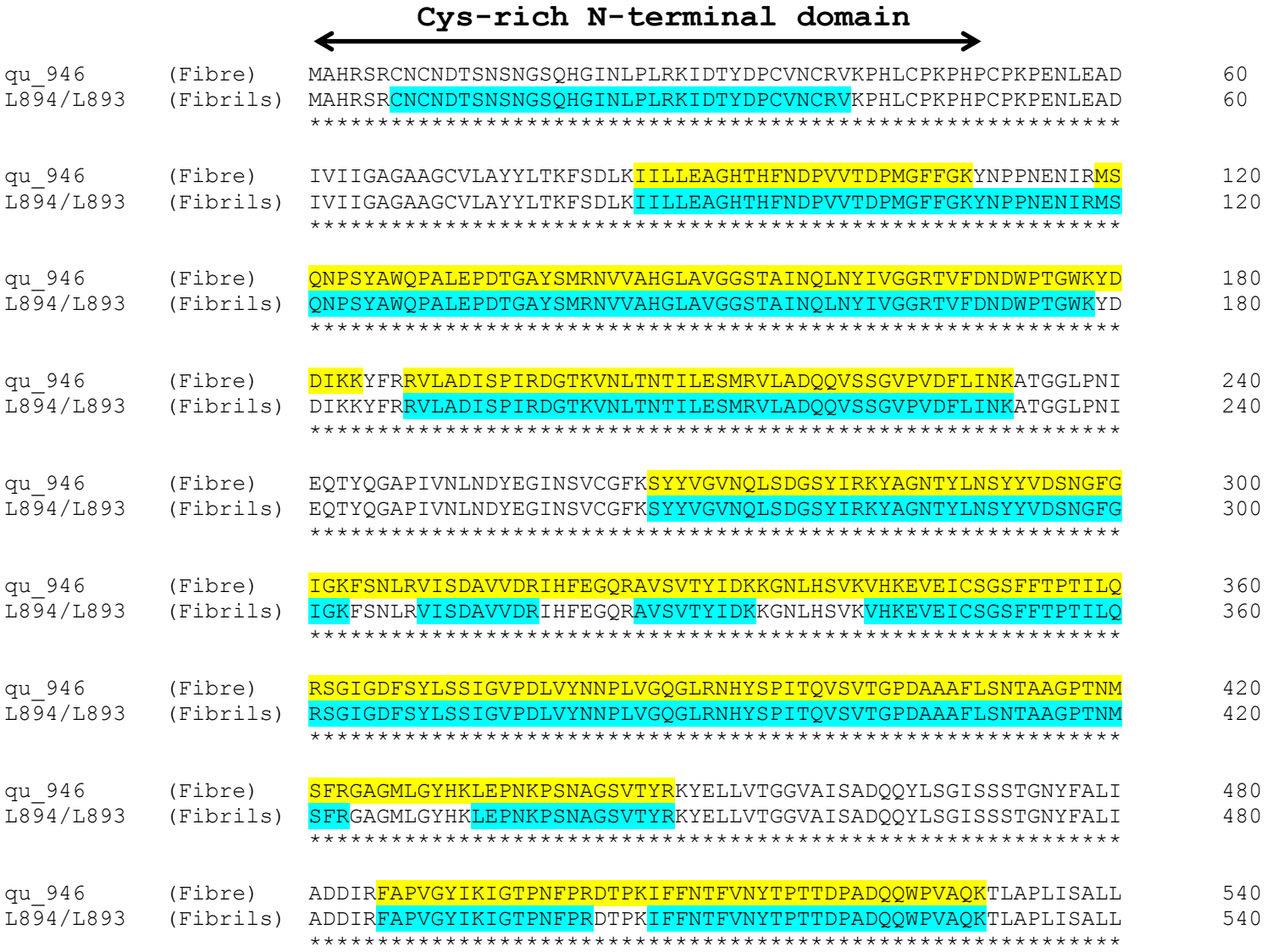


Extended data Fig. 4: Iterative helical 3D reconstruction based on optimized theoretical models for each compaction state of the genomic fibre. 1- A] Automatic sorting of the 2D classes using the fibre width W1 and pairwise correlations of re-oriented 2D classes resulting into to 3 main clusters (compact, Cl1 in blue; intermediate, Cl2 in cyan and unwound, Cl3 in orange). B] A theoretical model was then iteratively built (Extended data Fig. 5) to compute theoretical 2D classes further compared to the experimental ones. C] Validated model parameters are input into Relion[2,4] for 3D-refinement and classification. A final atomic model (Cl1 5 start: qu_946, dsDNA) was built into the best 3D-map prior symmetry relaxed 3D refinement after symmetry and structure validation.



Extended data Figure 5: Theoretical model improvement through iterative refinement. A) Initial "donut" shaped building block designed from the 2D classes images, B) optimized tetrameric building block, C] envelop of the building block which was manually extracted from the intermediate 3D reconstruction, D) model of the qu_946 dimer based on the 4z24 structure[9,11].

A


Cys-rich N-terminal domain
qu_946 (Fibre) MAHRSRCNCNDTSNSNGSQHGINLPLRKIDTYDPCVNCRVKPHLCPKPHPCPKPENLEAD 60
L894/L893 (Fibrils) MAHRSRCNCNDTSNSNGSQHGINLPLRKIDTYDPCVNCRVKPHLCPKPHPCPKPENLEAD 60
qu_946 (Fibre) IVIIGAGAAGCVLAYYLTKFSDLKIILLEAGHTHFNDPVVTDPMGFFGKYNPPNENIRMS 120
L894/L893 (Fibrils) IVIIGAGAAGCVLAYYLTKFSDLKIILLEAGHTHFNDPVVTDPMGFFGKYNPPNENIRMS 120
qu_946 (Fibre) QNPSYAWQPALEPDTGAYSMRNVVAHGLAVGGSTAINQLNYIVGGRTVFDNDWPTGWKYD 180
L894/L893 (Fibrils) QNPSYAWQPALEPDTGAYSMRNVVAHGLAVGGSTAINQLNYIVGGRTVFDNDWPTGWKYD 180
qu_946 (Fibre) DIKKYFRRVLADISPIRDGTKVNLTNTILESMRVLADQQVSSGVPVDFLINKATGGLPNI 240
L894/L893 (Fibrils) DIKKYFRRVLADISPIRDGTKVNLTNTILESMRVLADQQVSSGVPVDFLINKATGGLPNI 240
qu_946 (Fibre) EQTYQGAPIVNLNDYEGINSVCGFKSYYVGVNQLSDGSYIRKYAGNTYLNSYYVDSNGFG 300
L894/L893 (Fibrils) EQTYQGAPIVNLNDYEGINSVCGFKSYYVGVNQLSDGSYIRKYAGNTYLNSYYVDSNGFG 300
qu_946 (Fibre) IGKFSNLRVISDAVVDRIHFEGQRAVSVTYIDKKGNLHSVKVHKEVEICSGSFFTPTILQ 360
L894/L893 (Fibrils) IGKFSNLRVISDAVVDRIHFEGQRAVSVTYIDKKGNLHSVKVHKEVEICSGSFFTPTILQ 360
qu_946 (Fibre) RSGIGDFSYLSSIGVPDLVYNNPLVGQGLRNHYSPITQVSVTGPDAAAFLSNTAAGPTNM 420
L894/L893 (Fibrils) RSGIGDFSYLSSIGVPDLVYNNPLVGQGLRNHYSPITQVSVTGPDAAAFLSNTAAGPTNM 420
qu_946 (Fibre) SFRGAGMLGYHKLEPNKPSNAGSVTYRKYELLVTGGVAISADQQYLSGISSSTGNYFALI 480
L894/L893 (Fibrils) SFRGAGMLGYHKLEPNKPSNAGSVTYRKYELLVTGGVAISADQQYLSGISSSTGNYFALI 480
qu_946 (Fibre) ADDIRFAPVGYIKIGTPNFPRDTPKIFFNTFVNYTPTTDPADQQWPVAQKTLAPLISALL 540
L894/L893 (Fibrils) ADDIRFAPVGYIKIGTPNFPRDTPKIFFNTFVNYTPTTDPADQQWPVAQKTLAPLISALL 540

```
qu_946      (Fibre)    GYDAIYQIVQQMKVVAVNAGFNVTLQMAYPPNDLLVELHNGLNTYGINWWHYFVPSLVND   600
L894/L893   (Fibrils)  GYDAIYQIVQQMKVVAVNAGFNVTLQMAYPPNDLLVELHNGLNTYGINWWHYFVPSLVND   600
                       ************************************************************

qu_946      (Fibre)    DTPAGKLFASTLSKLSYYPRSGAHLDSHQSCSCSIGGTVDTELKVIGVENVRVTDLSAAP   660
L894/L893   (Fibrils)  DTPAGKLFASTLSKLSYYPRSGAHLDSHQSCSCSIGGTVDTELKVIGVENVRVTDLSAAP   660
                       ************************************************************

qu_946      (Fibre)    HPPGGNTWCTAAMIGARATDLILGKPLVANLPPEDVPVFTTS                     702
L894/L893   (Fibrils)  HPPGGNTWCTAAMIGARATDLILGKPLVANLPPEDVPVFTTS                     702
                       *****************************************
```

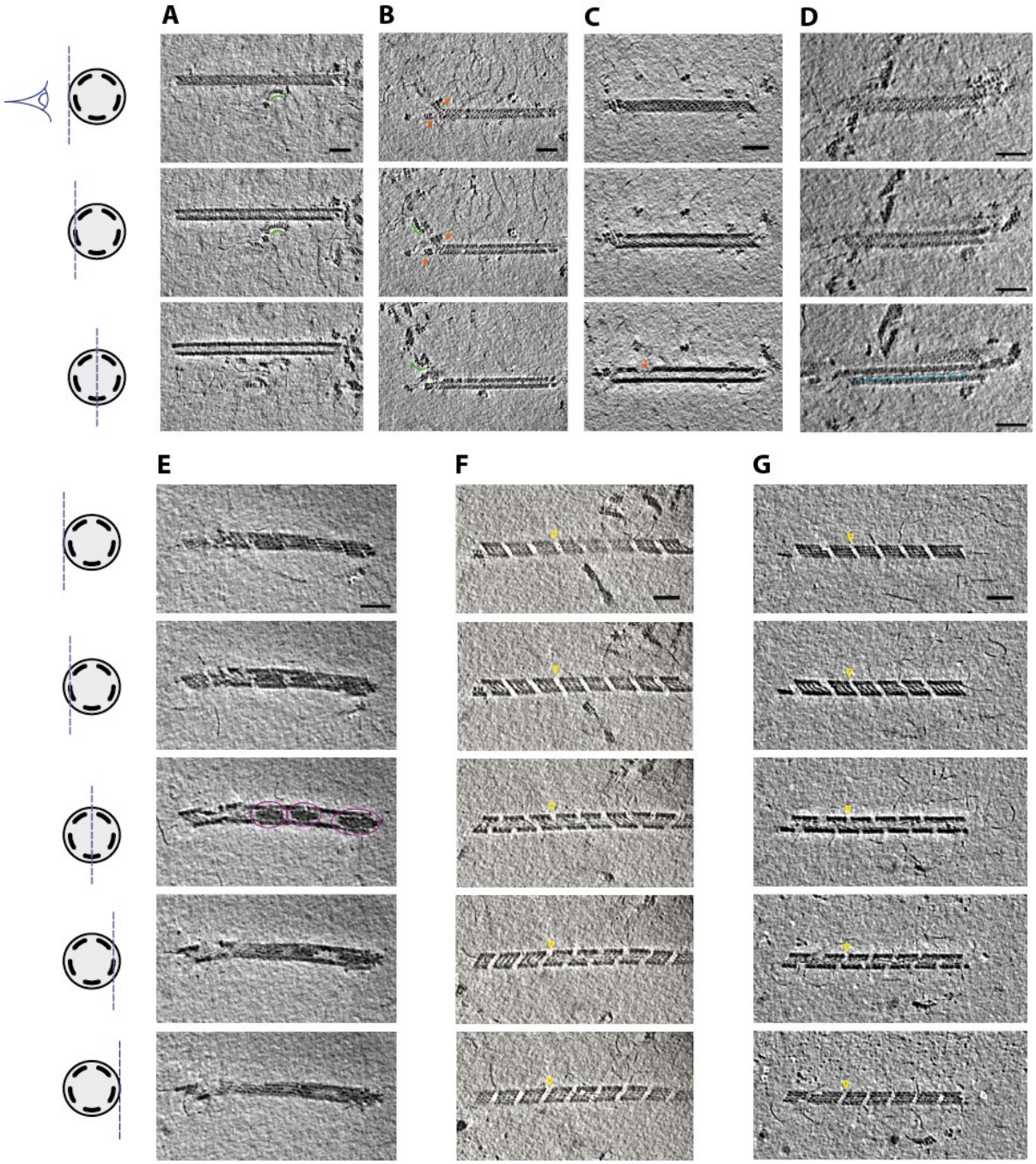
B

Cys-rich N-terminal domain

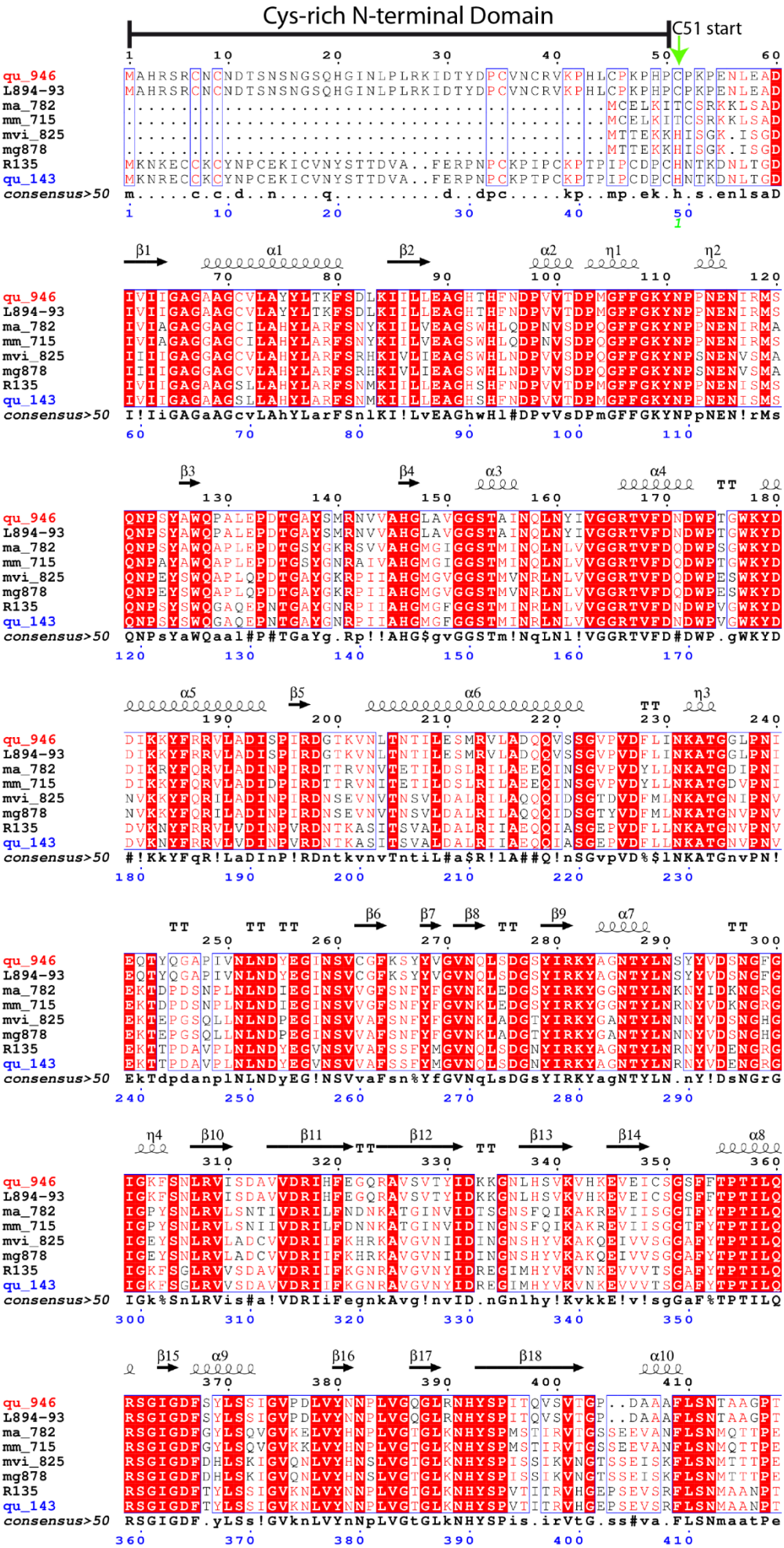
```
qu_143  (Fibre)          MKNRECCKCYNPCEKICVNYSTTDVAFERPNPCKPTPCKPTPIPCDPCHNTKDNLTGDIV    60
R135    (Fibrils)        MKNKECCKCYNPCEKICVNYSTTDVAFERPNPCKPIPCKPTPIPCDPCHNTKDNLTGDIV    60
                         ***:******************************* ************************

qu_143  (Fibre)          IIGAGAAGSLLAHYLARFSNMKIILLEAGHSHFNDPVVTDPMGFFGKYNPPNENISMSQN    120
R135    (Fibrils)        IIGAGAAGSLLAHYLARFSNMKIILLEAGHSHFNDPVVTDPMGFFGKYNPPNENISMSQN    120
                         ************************************************************

qu_143  (Fibre)          PSYSWQGAQEPNTGAYGNRPIIAHGMGFGGSTMINRLNLVVGGRTVFDNDWPVGWKYDDV    180
R135    (Fibrils)        PSYSWQGAQEPNTGAYGNRPIIAHGMGFGGSTMINRLNLVVGGRTVFDNDWPVGWKYDDV    180
                         ************************************************************

qu_143  (Fibre)          KNYFRRVLVDINPVRDNTKASITSVALDALRIIAEQQIASGEPVDFLLNKATGNVPNVEK    240
R135    (Fibrils)        KNYFRRVLVDINPVRDNTKASITSVALDALRIIAEQQIASGEPVDFLLNKATGNVPNVEK    240
                         ************************************************************

qu_143  (Fibre)          TTPDAVPLNLNDYEGVNSVVAFSSFYMGVNQLSDGNYIRKYAGNTYLNRNYVDENGRGIG    300
R135    (Fibrils)        TTPDAVPLNLNDYEGVNSVVAFSSFYMGVNQLSDGNYIRKYAGNTYLNRNYVDENGRGIG    300
                         ************************************************************

qu_143  (Fibre)          KFSGLRVVSDAVVDRIIFKGNRAVGVNYIDREGIMHYVKVNKEVVVTSGAFYTPTILQRS    360
R135    (Fibrils)        KFSGLRVVSDAVVDRIIFKGNRAVGVNYIDREGIMHYVKVNKEVVVTSGAFYTPTILQRS    360
```

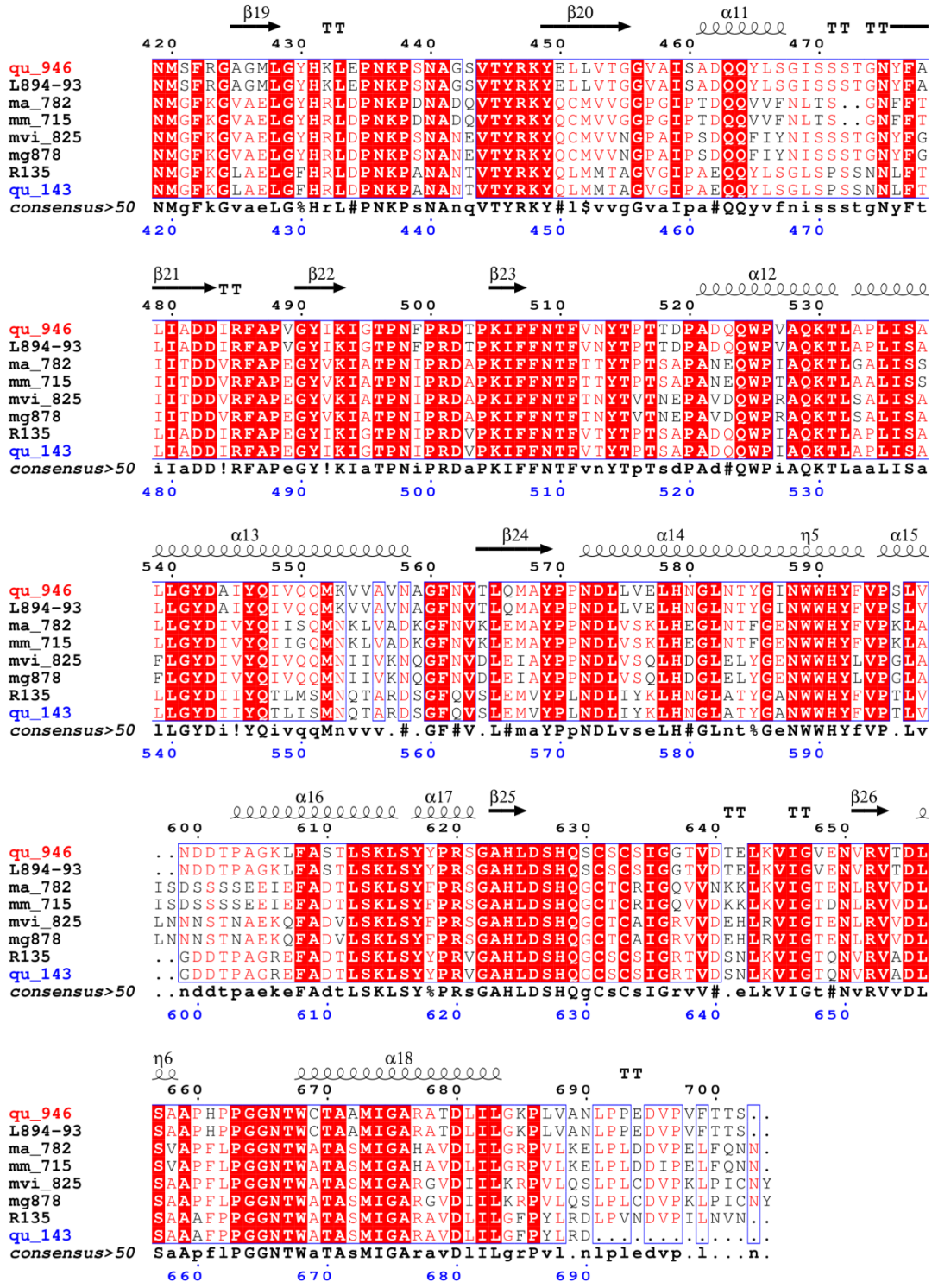
```
                    ************************************************************

qu_143 (Fibre)      GIGDFTYLSSIGVKNLVYNNPLVGTGLKNHYSPVTITRVHGEPSEVSRFLSNMAANPTNM  420
R135   (Fibrils)    GIGDFTYLSSIGVKNLVYNNPLVGTGLKNHYSPVTITRVHGEPSEVSRFLSNMAANPTNM  420
                    ************************************************************

qu_143 (Fibre)      GFKGLAELGFHRLDPNKPANANTVTYRKYQLMMTAGVGIPAEQQYLSGLSPSSNNLFTLI  480
R135   (Fibrils)    GFKGLAELGFHRLDPNKPANANTVTYRKYQLMMTAGVGIPAEQQYLSGLSPSSNNLFTLI  480
                    ************************************************************

qu_143 (Fibre)      ADDIRFAPEGYIKIGTPNIPRDVPKIFFNTFVTYTPTSAPADQQWPIAQKTLAPLISALL  540
R135   (Fibrils)    ADDIRFAPEGYIKIGTPNIPRDVPKIFFNTFVTYTPTSAPADQQWPIAQKTLAPLISALL  540
                    ************************************************************

qu_143 (Fibre)      GYDIIYQTLISMNQTARDSGFQVSLEMVYPLNDLIYKLHNGLATYGANWWHYFVPTLVGD  600
R135   (Fibrils)    GYDIIYQTLMSMNQTARDSGFQVSLEMVYPLNDLIYKLHNGLATYGANWWHYFVPTLVGD  600
                    *********:**************************************************

qu_143 (Fibre)      DTPAGREFADTLSKLSYYPRVGAHLDSHQGCSCSIGRTVDSNLKVIGTQNVRVADLSAAA  660
R135   (Fibrils)    DTPAGREFADTLSKLSYYPRVGAHLDSHQGCSCSIGRTVDSNLKVIGTQNVRVADLSAAA  660
                    ************************************************************

qu_143 (Fibre)      FPPGGNTWATASMIGARAVDLILGFPYLRDLPVNDVPILNVN                    702
R135   (Fibrils)    FPPGGNTWATASMIGARAVDLILGFPYLRDLPVNDVPILNVN                    702
                    *****************************************
```

Extended data Figure 6: Comparative proteomic analysis of the purified fibrils and genomic fibres. Peptide coverage of A] qu_946 B] qu_143 identified in the purified fibrils (cyan) and genomic fibres (yellow).



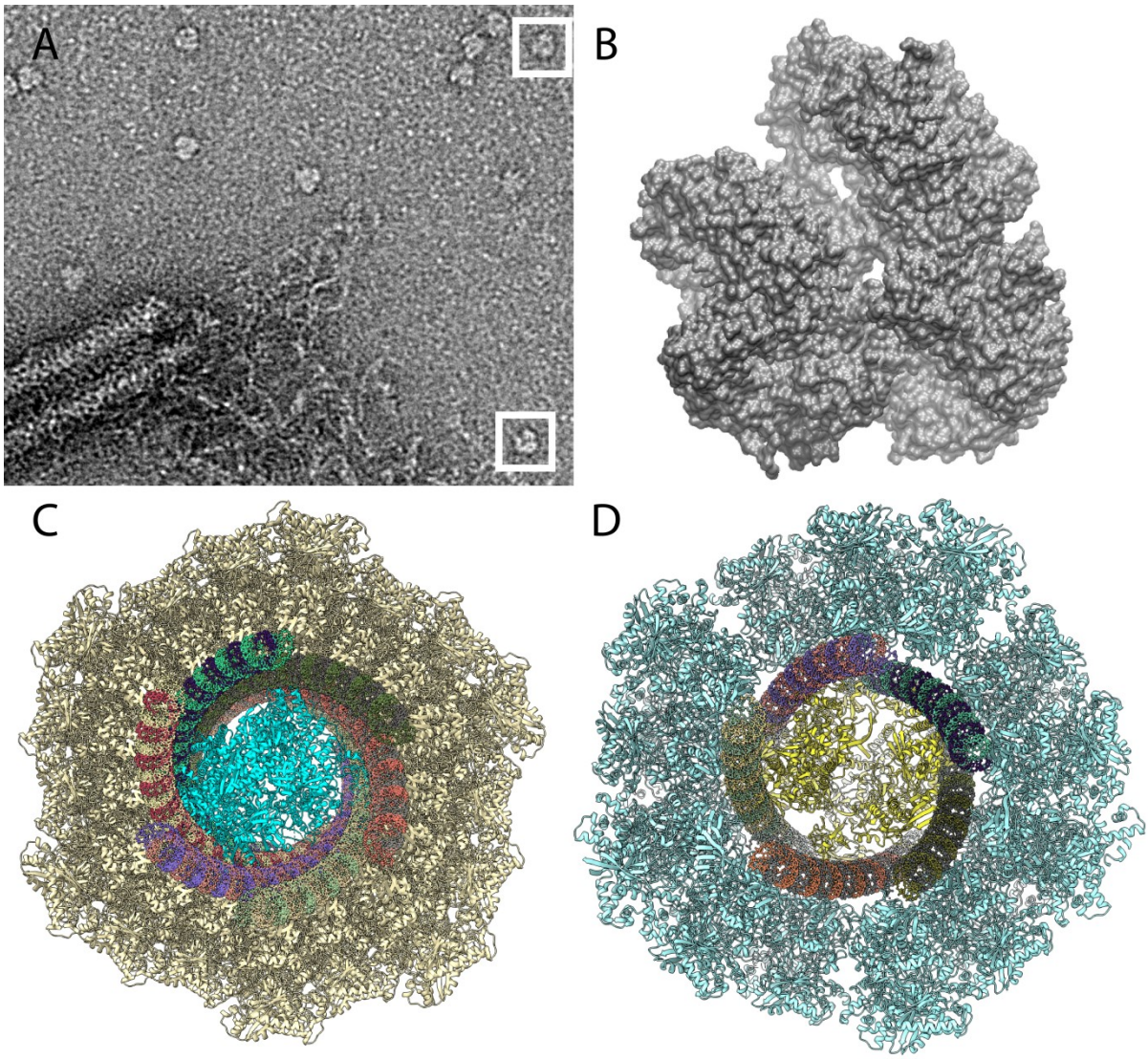
Extended data Figure 7: Cryo-ET of Mimivirus genomic fibre. Slices through tomograms exhibiting different features of compact (upper panel) and broken (lower panel) genomic fibres. (Upper panel) Free DNA strands are bent when they remain covered by the shell protein (green overlays in A and B); punctual breaks in the fibre (orange arrows in B and C); sometimes DNA strand is visible in the centre of the fibre (blue dashed line in D). (Lower panel) electron dense masses inside the lumen (highlighted with purple circles in E); breaks along the genomic fibre are equally interspaced and lead to rearrangement of the DNA strands (yellow arrowheads in F and G). For display purposes, the tomograms were binned 8 times and band-pass filtered using bsoft[17] (see extended data methods), images of the slices were prepared using the slicer function in 3dmod from the IMOD package[16]. Thickness of the slices is 1.1 nm for all. Distance between the tomographic slices (from top to bottom) is 4.4 nm between first and second (all), and fourth and fifth (E-G); and 6.6 nm between the third and second (all) or fourth (E-G). The scheme on the left indicates the plane through which the genomic fibre is viewed. Scale bars, 50 nm.


Cys-rich N-terminal Domain
C51 start
qu_946
L894-93
ma_782
mm_715
mvi_825
mg878
R135
qu_143
consensus>50



Extended data Figure 8: Multiple alignment of selected Mimiviruses GMC-oxydoreductases using Multalin[27]. The 4Z24 structure[9] was used as reference for secondary structure elements, qu_946 and qu_143 numbering are indicated at the top and bottom of the alignment, respectively. The figure was prepared using ESPript 3.0[28] (http://espript.ibcp.fr).

.



Extended data Figure 9: A] Negative staining micrograph of unprotected DNA still connected to an unwinding fibre with scattered macromolecular complexes. B] Surface representation of the model of the Mimivirus RNA polymerase built with the subunits identified by MS-proteomic of the purified fibres based on the vaccinia virus structure (6RIC Chains ABCDEF[24]). C] Cartoon representation of the RNA polymerase subunits (in cyan) docked into the theoretical model of the 6-start structure. The protein shell is cartooned in gold and the dsDNA strands are represented as ball & sticks and individually coloured. D] Cartoon representation of the RNA polymerase subunits (in yellow) manually docked into the theoretical model of the 5-start structure. The protein shell is cartooned in cyan and the dsDNA strands are represented as ball & sticks and individually coloured. The figure was produced using ChimeraX[25].

Extended data Table 1: Parameters of the different genomic fibre structures			
Compacted 5-start			
Shell external diameter	283.4 Å ± 4e-4	DNA external diameter	132 Å ± 8e-2
Shell internal diameter	115.21 Å ± 5e-4	DNA-internal diameter	93.06 Å ± 8.1e-2
Thickness	84.09 Å	DNA interspacing	39.53 Å ± 3.6e-1
Spacing Shell-DNA	6.6 Å ± 5.3e-1	Rise, Twist	7,93 Å, -221,07°
Unwound 5-start			
Shell external diameter	331.92 Å ± 3e-4	Rise, Twist	5,49 Å, -148,26°
Shell internal diameter	165.38 Å ± 4e-4		
Thickness	83.97 Å		
Compacted 6-start			
Shell external diameter	290 Å ± 3e-4	DNA external diameter	144.48 Å ± 9.6e-2
Shell internal diameter	133.61 Å ± 4e-4	DNA-internal diameter	105.52 Å ± 9.6e-2
Thickness	78 Å	DNA interspacing	36.4 Å ± 4.6e-1
Spacing Shell-DNA	5.8 Å ± 1.8	Rise, Twist	6,22 Å, -63,26°,
Unwound 6-start			
Shell external diameter	331.90 Å ± 4e-4	Rise, Twist	5,56 Å, -63,52°
Shell internal diameter	173.90 Å ± 4e-4		
Thickness	79 Å		

Extended table 2: Mass spectrometry-based proteomic analysis of the Mimivirus reunion.

gene name	protein name	molecular weight (Da)	identified peptides	spectral counts	coverage	iBAQ
qu_946	GMC oxidoreductase	78892	52	211	63.66	209431097
qu_143	GMC-type oxidoreductase	76360	47	119	57.55	64288450
qu_734	hypothetical protein	25119	2	4	15.25	2238815
qu_772	hypothetical protein	24047	6	8	31.65	1111931
qu_446	Major capsid protein	67238	8	9	17.73	500153
qu_623	hypothetical protein	29447	3	3	13.9	325738
qu_384	Thioredoxin	39459	5	6	18.5	297567
qu_629	Thiol oxidoreductase E10R	34053	1	1	3.08	280834
qu_741	hypothetical protein	16247	1	1	5.44	236574
qu_409	hypothetical protein	30040	3	3	13.21	199984
qu_880	hypothetical protein	49215	3	3	9.47	166941
qu_431	Predicted Major core protein	75338	2	2	2.87	137622
qu_205	Glutaredoxin	11539	1	1	9.8	88231
qu_736	hypothetical protein	40921	2	2	4.28	79433
qu_738	hypothetical protein	39143	2	2	8.62	52064
qu_368	hypothetical protein	56316	1	1	2.06	44971
qu_76	collagen-like protein 1	88293	2	2	2.58	35680
qu_418	hypothetical protein	64964	1	1	1.41	32102
qu_824	putative PAN domain-containing protein	22435	1	1	6	31963
qu_748	hypothetical protein	22233	1	1	4.43	30058
qu_245	RNA polymerase subunit 5	23541	1	1	7.32	29204
qu_366	putative regulator of chromosome condensation	106242	2	2	2.06	28934
qu_544	hypothetical protein	10721	1	1	11.83	27501
qu_220	DNA-directed RNA polymerase subunit 6	46141	1	1	2.18	25534
qu_511	putative PAN domain-containing protein	24425	1	1	5.38	21622
qu_600	hypothetical protein	37211	2	2	7.19	20856
qu_464	hypothetical protein	139105	1	2	0.95	20731
qu_495	hypothetical protein	199755	5	5	2.82	19382
qu_261-259-257-255	DNA directed RNA polymerase subunit 2	135691	3	3	2.68	18709
qu_543	hypothetical protein	18141	1	1	7.14	18247
qu_617	hypothetical protein	11815	1	1	10.28	12489
qu_351	hypothetical protein	30607	1	1	3.4	12330
qu_68	hypothetical protein	67161	2	3	3.54	11619
qu_616	uncharacterized N-acetyltransferase	46994	1	1	2.7	8000
qu_530-532	DNA directed RNA polymerase subunit 1	119306	2	2	2.01	7691
qu_493	DNA directed RNA polymerase subunit 3/11	41630	1	1	3.92	6470
qu_404	probable mRNA-capping enzyme	136590	1	1	0.6	2560

Extended Table 3: Cryo-EM data acquisition parameters

	Single particle analysis	**Tomograms**	**Bubblegrams**
Hardware			
Microscope	Krios1	Krios1	Krios1
Detector	K2	K3	K3
Accelerating voltage (keV)	300	300	300
Pixel size (Å)	1.09 Å	1.80 Å	1.09 Å
Data acquisition parameters			
Nominal magnification	130,000	64,000	81,000
Square pixel (Å²)	1.1881	3.24	1.1881
Dose per physical pixel per second	7.5	15	15.017
Dose (e-/Å²/sec)	6.3	4.6	12.64
Exposure time (sec)	8	0.8	2-6
Total Dose (e-/Å²)	50.5	3.68	25-75
Number of frames	40	4	0.1
Dose per fraction (e-/Å²)	1.25	0.92	1.25
Defocus range (μm)	1 to 3	2 to 4	3 to 5
Apertures (Size in microns)			
C1	2,000	2,000	2,000
C2	70	50	50
Microprobe/Nanoprobe	Np	Np	Np
Objective	100	100	100
Energy filter			
Slit (eV)	20	20	20

Extended Table 4: Mimivirus Reunion 5-start fibre compact state data statistics[12]

Mimivirus Reunion 5-start fibre compacted state	
Data collection and processing	
Magnification	130,000
Voltage (kV)	300
Electron exposure (e^{-} Å^{-2})	50,1
Defocus range (μm)	−1 to −3.0
Pixel size (Å)	1.09
Symmetry imposed	Helical
Initial particle images (no.)	49,384
Final particle images (no.)	13,252
Map resolution (Å)	6.3
FSC threshold	0.143
Map resolution range (Å)	4.3-12
3D Refinement	
Initial model used	3D class model low-pass filtered to 15 Å
Model resolution (Å)	5.68 Å
FSC threshold	0.143
Model composition	
Non-hydrogen atoms	447,930
Protein residues	58,410
R.m.s. deviations	
Bond lengths (Å)	0,003
Bond angles (°)	0,625
Validation	
MolProbity score	2.21
Clashscore	21.77
Rotamers outliers (%)	0.18
Ramachandran plot	
Favoured (%)	94.40
Allowed (%)	5.60
Disallowed (%)	0.0