Published online 1 July 2022 *Nucleic Acids Research, 2022, Vol. 50, No. 13* ***7655–7668***
https://doi.org/10.1093/nar/gkac532

Molecular insights into the interaction of CAG trinucleotide RNA repeats with nucleolin and its implication in polyglutamine diseases

Ying An[1,3], Zhefan S. Chen[1], Ho Yin Edwin Chan[1,2,*] and Jacky Chi Ki Ngo [1,3,4,*]

[1]School of Life Sciences, The Chinese University of Hong Kong, Shatin N.T., Hong Kong SAR, China, [2]Gerald Choa Neuroscience Centre, The Chinese University of Hong Kong, Shatin N.T., Hong Kong SAR, China, [3]Center for Soybean Research of the State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin N.T., Hong Kong SAR, China and [4]Center for Novel Biomaterials, The Chinese University of Hong Kong, Shatin N.T., Hong Kong SAR, China

Received February 18, 2022; Revised May 08, 2022; Editorial Decision June 05, 2022; Accepted June 08, 2022

ABSTRACT

Polyglutamine (polyQ) diseases are a type of inherited neurodegenerative disorders caused by cytosine–adenine–guanine (CAG) trinucleotide expansion within the coding region of the disease-associated genes. We previously demonstrated that a pathogenic interaction between expanded CAG RNA and the nucleolin (NCL) protein triggers the nucleolar stress and neuronal cell death in polyQ diseases. However, mechanisms behind the molecular interaction remain unknown. Here, we report a 1.45 Å crystal structure of the r(CAG)$_5$ oligo that comprises a full A′-form helical turn with widened grooves. Based on this structure, we simulated a model of r(CAG)$_5$ RNA complexed with the RNA recognition motif 2 (RRM2) of NCL and identified NCL residues that are critical for its binding to CAG RNA. Combined with *in vitro* and *in vivo* site-directed mutagenesis studies, our model reveals that CAG RNA binds to NCL sites that are not important for other cellular functions like gene expression and rRNA synthesis regulation, indicating that toxic CAG RNA interferes with NCL functions by sequestering it. Accordingly, an NCL mutant that is aberrant in CAG RNA-binding could rescue RNA-induced cytotoxicity effectively. Taken together, our study provides new molecular insights into the pathogenic mechanism of polyQ diseases mediated by NCL–CAG RNA interaction.

INTRODUCTION

Polyglutamine (PolyQ) diseases are a type of heterogeneous hereditary neurodegenerative disorders (1,2). The causative mutation of polyQ diseases is the uninterrupted genomic expansion of trinucleotide cytosine–adenine–guanine (CAG) repeats in the coding region of the disease genes (3,4). To date, nine types of neurological diseases have been characterized and classified as polyQ diseases: spinal and bulbar muscular atrophy (SBMA) (5), Huntington's disease (HD) (6), dentatorubral-pallidoluysian atrophy (DRPLA) (7), and six types of spinocerebellar ataxias (SCA1, SCA2, SCA3, SCA6, SCA7 and SCA17, respectively) (8). The presence of CAG-repeated expansion in the mutant allele leads to the production of both elongated CAG RNAs and polyQ proteins in the patients. The misfolding and aggregation of polyQ proteins cause cellular toxicity in patients with polyQ diseases. However, growing evidence has demonstrated that the expanded CAG RNAs contribute to disease pathogenicity in a dominant manner by disrupting normal cellular processes (2,9).

Nucleolin (NCL) is an evolutionarily conserved protein that is predominantly localized in the cell nucleolus and regulates various cellular events, such as transcription, ribosomal biogenesis, mRNA stability and protein translation, by binding to different RNA targets (10–15). The dysregulation of NCL has been implicated in many pathological conditions, such as cancer (16–18) and neurodegenerative diseases (19–21). We and others have reported that the expanded CAG transcripts in polyQ diseases (22,23) and the GGGGCC RNA repeats in C9ALS/FTD (24–26), respectively, interact with RNA-binding proteins, including NCL, to trigger toxic RNA pathogenicity. In polyQ diseases, expanded CAG RNA sequesters NCL, thus inhibiting ribosomal RNA (rRNA) transcription. The suppression of rRNA synthesis impairs ribosomal biogenesis (27),

[*]To whom correspondence may be addressed. Tel: +852 3943 6346; Fax: +852 2603 7246; Email: jackyngo@cuhk.edu.hk
Correspondence may also be addressed to Ho Yin Edwin Chan. Tel: +852 3943 4021; Fax: +852 2603 7732; Email: hyechan@cuhk.edu.hk

eventually triggering nucleolar stress and p53-mediated cell apoptosis (22,27,28).

Given the essential role of the interaction between NCL and expanded CAG RNAs in the pathogenesis of polyQ diseases, elucidating structural mechanisms underlying the interaction between CAG RNA repeats and NCL can be beneficial for therapeutic development. The human NCL protein consists of 710 amino acids, and it can be divided into three functional domains: an N-terminal domain containing several stretches of acidic amino acids, a central domain carrying four RNA recognition motifs (RRMs), and an arginine/glycine-rich C-terminal domain (29–31). Our previous study reported that the RRM2 and RRM3, but not RRM1 and RRM4, of NCL are vital for the direct interaction with expanded CAG RNA (22). However, the detailed molecular mechanism through which mutant CAG RNA interacts with NCL remains largely elusive.

Studies have attempted to uncover the pathogenic roles of trinucleotide repeat RNAs by determining their three-dimensional structures. CAG repeat-containing transcripts would fold into stable hairpin structures that recruit RNA-binding proteins and form ribonucleoprotein complexes in cells (32,33). However, due to the high content of CAG repeats, the potential conformational heterogeneity of hairpin long stretches of CAG trinucleotides makes it difficult for X-ray crystallographic studies. To date, all the existing 3D structures of trinucleotide repeats, including (CUG)$_n$ (34–36) and (CAG)$_n$ (37–40), adopt double-stranded helical structures. In this study, we solved the structure of CAG RNA containing five repeats and investigated the molecular mechanisms underlying its interaction with NCL. Next, we built a molecular model of the complex of NCL–RRM2:CAG RNA and validated it both *in vitro* and *in vivo*. Furthermore, we confirmed that the NCL residues involved in the interaction between CAG RNA and NCL are dispensable for its normal functions in the regulation of gene expression and rRNA synthesis. This finding indicated that CAG RNA sequesters NCL, instead of competing with its cognate substrates, during the pathogenesis of polyQ diseases. Taken together, our work provides novel insights into how CAG RNA interacts with NCL at the molecular level and opens up a new opportunity for therapeutic development to combat the RNA toxicity of polyQ diseases.

MATERIALS AND METHODS

Crystallization, data collection and structure determination

The rGGGC(CAG)$_5$GUCC, referred to r(CAG)$_5$ oligomer, was purchased from Integrated DNA Technologies, Inc. The r(CAG)$_5$ oligomer was dissolved in diethylpyrocarbonate (DEPC)-treated water to the final concentration of 1.2 mM. Prior to crystallization, the oligomer was heated at 75°C for 10 min and cooled slowly to room temperature within 3 h. Crystals were obtained from the condition of 100 mM HEPES (pH 7.5), 1.6 M $(NH_4)_2SO_4$ and 10% PEG 400 at 16°C using vapor diffusion method. X-ray diffraction data were collected on the beamline TPS 05 at National Synchrotron Radiation Research Center (NSRRC) in Taiwan. The HKL2000 program (41) was applied for data processing.

Table 1. X-ray crystallographic data collection and refinement statistics

Data collection	r(CAG)$_5$ RNA
Data collection source	TPS 05
Wavelength (Å)	1.0000
Resolution (Å)	30.70–1.45
No. of measured reflections	223 072
No. of rejected reflections	1270
Space group	H 3 2
Cell dimensions	
a, *b*, *c* (Å)	47.21, 47.21, 185.98
α, β, γ (Å)	90, 90, 120
R_{sym}	0.04 (0.33)
I/σ_I	58.1 (6.6)
Completeness (%)	99.7(100)
Wilson *B*-factor (Å)	12.5
Refinement	
Resolution (Å)	22.28–1.45
No. of unique reflections	14676
R_{work}/R_{free}	0.195, 0.208
No. of atoms	601
Water	107
Average *B*-factor (Å)	19
RMS deviations	
Bond length (Å)	0.004
Bond angles	0.929

The RNA structure was solved by molecular replacement using the Phaser program embedded in the PHENIX suite (42), manually revised using Coot (43), and further refined using PHENIX (42). The truncated rUUGGGC(CAG)$_3$GUCC structure (PDB code: 4YN6) (40) was used as a search model. The 3DNA program (44,45) was used to calculate the helical parameters of the RNA structure. Data collection and refinement statistics are summarized in Table 1. All structural figures were drawn by the software PyMOL. The coordinate was deposited in the Protein Data Bank (PDB) with an accession code of 7VFT.

Construction of plasmids and RNA interference

The *pcDNA3.1(+)-S1-MJD*$_{CAG78}$ and *pcDNA3.1(+)-SCA2*$_{CAG42}$ constructs were previously described (22,46,47). To generate *pEGFP-NCL*^{WT}, the *NCL* DNA sequence was amplified from the HEK293 cDNA template. The DNA fragment was subsequently cloned into *pEGFP-C1* vector using KpnI and BamHI. To generate the construct of *pET15b-NCL–RRM2/3*, the *NCL–RRM2/3* fragment was amplified and subcloned into *pET15b* vector using NdeI and BamHI. The *NCL–RRM2/3* mutant constructs that include a double mutant (K427A/K429A), a single mutant (Y433A), and a triple mutant (K427A/K429A/K437A) were generated by site-directed mutagenesis and PCR cloning. The *NCL–RRM2/3* mutant constructs used for the interaction study were purchased from GenScript (Supplementary Table S1). The *pEGFP-NCL*^{MT} construct was generated via a two-step PCR cloning. In the first-round PCR, the *NCL (1–387)* and *NCL (560–710)* fragments were amplified from *pEGFP-NCL*^{WT}; the *NCL–RRM2/3*^{MT} fragment with multiple mutations (K398A, Y402A, K424A, K427A, K429A, and R457A) was amplified from *pET15b-NCL–RRM2/3*^{MT}. In the second-round PCR, the above fragments were mixed and used as templates for generating the full-length *NCL*^{MT}

fragment using primers NCL-1-F and NCL-710-R. The full-length *NCL^{MT}* insert was then subcloned into *pEGFP-C1* vector (Clontech) using KpnI and BamHI. The detailed sequences of the primers used in this study are listed in Supplementary Table S2.

Three NCL siRNAs were purchased from Sangon Biotech. Their sequences are 5′-GCAGCCUGUAUUC UGGAUATT-3′, 5′-GGACAUUCCAAGACAGUAU TT-3′ and 5′-GAGUUGAGUGAUAGAGCUATT-3′. Non-targeting siRNA (D-001210-01-50) was purchased from GE Dharmacon, which was used as a control.

Protein expression and purification

Wild-type and mutant proteins of human NCL–RRM2/3 were purified by similar protocols. Optimal protein expression condition was obtained from *E. coli strain* BL21 (DE3). Cells were harvested and lysed by sonication in buffer containing 25 mM Tris–HCl (pH 8.0), 300 mM NaCl, 20 mM imidazole, 5% glycerol, 1 mM benzamidine and 1 mM phenylmethylsulfonyl fluoride (PMSF). The supernatant was subjected to Ni^{2+}-NTA column and subsequent eluate was further polished by gel filtration chromatography (Superdex-75 16/60 GL, GE Healthcare) and monoQ GL column (GE Healthcare). The final purified protein was kept in buffer containing 20 mM HEPES (pH 7.4) and 100 mM KCl. All the buffers used were prepared in DEPC-treated water.

Electrophoretic mobility shift assay

Electrophoretic mobility shift assays (EMSAs) were conducted as previously described (48,49). *$SCA2_{CAG42}$* RNA was synthesized by the MEGAscript® kit (Ambion) following the manufacturer's protocol (22). Prior to the EMSA, *$SCA2_{CAG42}$* RNA was first denatured at 90°C for 10 min and then refolded at 37°C for 50 min. Wild-type and mutant NCL–RRM2/3 proteins were prepared in buffer including 20 mM HEPES (pH7.4) and 100 mM KCl. Reactions were assembled in a final volume of 5 μl containing 166 nM *$SCA2_{CAG42}$* RNA, 20 mM HEPES (pH 7.4), 100 mM NaCl, 0.5 μl RNaseOUT Recombinant Ribonuclease Inhibitor (Thermo Fisher Scientific), and different concentrations of wild-type and mutants of NCL–RRM2/3. The reaction mixtures were then incubated at 25°C for 1 h. The samples were analyzed on 2.3% agarose gels and the results were visualized and recorded under Gel Doc XR + system.

For an apparent K_d determination of the protein–RNA complex, a set of equilibration reaction mixtures were assembled by adding various amount of NCL–RRM2/3 to a fixed concentration of *$SCA2_{CAG42}$* RNA. Different concentrations of NCL–RRM2/3 obtained by serial dilution were incubated with 92 nM of *$SCA2_{CAG42}$* RNA at 25°C for 1 h. The data were analyzed using GraphPad Prism 8.0 to perform non-linear regression.

Circular dichroism spectroscopy

All circular dichroism (CD) measurements were conducted by using a JASCO J-810 spectrometer with a peltier temperature controller. Wild-type and mutant NCL–RRM2/3 proteins were prepared in 10 mM sodium phosphate (pH 7.4). Each CD spectrum was recorded at 25°C in the 190–260 nm range with 1 s response time and 1 nm bandwidth, and a scanning speed of 50 nm/min. The final spectrum was the average of three accumulations.

Molecular docking

The docking of NCL–RRM2 to CAG RNA was conducted using the data-driven flexible docking server HADDOCK 2.2 (50) (https://alcazar.science.uu.nl/services/HADDOCK2.2/). The model of NCL–RRM2, prepared from the NCL–RRM1/2 structure (PDB code: 2KRR), was input as the protein template whereas our $r(CAG)_5$ RNA structure (PDB code: 7VFT) was served as the RNA template. To generate the ambiguous interaction restraints (AIRs) for driving the docking, residues K427 and K429 were defined as active residues of NCL–RRM2 based on our EMSA results, while A·A base pairs from the central 3-CAG-repeats of $r(CAG)_5$ RNA were input as active residues based on previous findings that NCL binds specifically to A·A base pairs (22). Passive residues which are solvent-accessible surface neighbors to the active ones were defined by the HADDOCK server automatically. All HADDOCK attempts were conducted with 1,000 rigid body solutions and with ten times of rigid body minimization. The best 200 structures were subjected to semi-flexible refinement followed by water refinement. The final structures were applied to cluster analysis using the Fraction of Common Contacts (FCC) with a cutoff value of 0.75 and a smallest cluster size of 4.

Cell culture, plasmid transfection and siRNA transfection

SK-N-MC cells (American Type Culture Collection) were cultured in the Dulbecco's Modified Eagle Medium containing 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin in a 37°C incubator with 5% CO_2. All plasmid transfections were carried out by using Lipofectamine 2000 (Thermo Fisher Scientific). All siRNAs were transfected with Lipofectamine® RNAIMAX (Thermo Fisher Scientific) transfection reagent.

RNA extraction, reverse transcription PCR and real time PCR

RNA was isolated from cells by using TRIzol reagent (Thermo Fisher Scientific) and treated with DNase I (Thermo Fisher Scientific) to remove contaminant genomic DNA before reverse transcription PCR. Purified RNA (0.5 μg) was used for reverse transcription (ImPromII Reverse Transcription System, Promega). Random hexamer (Thermo Fisher Scientific) or Oligo(dT) (Thermo Fisher Scientific) was used during the reverse transcription. Real time PCR was conducted on a Bio-Rad CFX96 Real Time PCR system by using TaqMan® probes to quantify gene expression as previously described (22). Probes used in the assays are listed: Human pre-45s rRNA (Assay ID: AILJIZM) and Human *actin* (Assay ID: Hs99999903_m1).

Protein sample preparation and western blotting

All protein samples were prepared from SK-N-MC cells by using SDS lysis buffer (100 mM Tris–HCl, pH 6.8, 2% SDS, 40% glycerol, 5% β-mercaptoethanol and 0.1% bromophenol blue). Samples were pre-heated at 99°C for 10 min before loaded onto a 12% SDS-PAGE gel. Primary antibodies used in this study were: anti-nucleolin 3G4B2 (Sigma-Aldrich; 1:1000) for endogenous NCL, anti-GFP (JL-8, Clontech; 1:2000) for EGFP-NCL, anti-Akt1 (B-1, Santa Cruz; 1:1000) for Akt1, anti-p53 (1C12, Cell Signaling Technology; 1:2000), anti-β-tubulin (E7, Developmental Studies Hybridoma Bank; 1:5000). HRP conjugated Goat anti-Rabbit IgG (H + L) (11-035-045) and HRP conjugated Goat anti-Mouse IgG (H + L) (115-035-062) secondary antibodies (Jackson ImmunoResearch) were used at 1:5000. The chemiluminescence signal was developed using Immobilon Forte Western HRP substrate (Merck Millipore), and the images were captured and processed using ChemiDoc™ Touch Imaging System (Bio-Rad).

Protein immunoprecipitation (IP) with subsequent RNA detection

SK-N-MC cells were seeded on 6 cm-dishes with a density of 1.6 × 10^6 cells/dish. When reaching 80–90% confluence, cells were transfected with 6 μg of *pcDNA3.1(+)-S1-MJD*$_{CAG78}$ and an equal amount of either *pEGFP-NCL* or *pEGFP* control plasmids simultaneously using Lipofectamine 2000 in accordance with the manufacturer's instructions. After 48 h post-transfection, cells were collected and lysed as previously described (51). Ten percents of total cell lysate was used as input. Binding buffer used in this assay contains 50 mM HEPES (pH 7.4), 150 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT (fresh), 0.5% Triton X-100, supplemented with 0.5% bovine serum albumin (BSA), 100 μg/ml yeast tRNA, 1 mM PMSF, protease inhibitor cocktail (Sigma-Aldrich) and RNAsin (Promega). Twenty microliters of Protein G dynabeads (Thermo Fisher Scientific) was prepared in 200 μl buffer containing 50 mM Tris-HCl (pH 8.0), 200 mM NaCl, 0.5 mM $Mg(OAc)_2$, 20 μg/ml heparin, 0.01% NP-40, 5% glycerol, 1 mM DTT (fresh), supplemented with 0.5% BSA, 100 μg/ml yeast tRNA, 1 mM PMSF, protease inhibitor cocktail (Sigma-Aldrich) and incubated with 3 μg anti-GFP antibody (A01388, GenScript) or normal rabbit IgG antibody (12-370, Millipore) for 2 h at room temperature with gentle rotation. Antibody bound beads were washed three times with buffer and then incubated with the remaining cell lysates at 4°C overnight. The beads with protein-RNA complex captured were washed with binding buffer twice followed by washing buffer (binding buffer containing 300 mM NaCl). For both the input and suspension, 20% of the samples was saved for protein analysis by western blot, and the rest was subjected to RNA extraction by using TRIzol reagent (Thermo Fisher Scientific) followed by reverse transcription PCR and conventional PCR.

Lactate dehydrogenase (LDH) cytotoxicity assay

SK-N-MC cells were seeded on a 24-well plate with a density of 1.2 × 10^5 cells/well. Upon reaching 80% confluency, cells were transfected with 1 μg of *pcDNA3.1(+)-MJD*$_{CAG78}$ and 0.4 μg of either *pEGFP-NCL* or *pEGFP* control constructs simultaneously. After 48 h, LDH enzyme activity in the cell culture medium was determined by using the Cytotox 96 non-radioactive cytotoxicity assay kit (Promega).

Statistical analyses

All data were subjected to one-way ANOVA followed *by post hoc* Tukey test. (∗), (∗∗) and (∗∗∗) represent $P < 0.05$, $P < 0.01$ and $P < 0.001$, respectively, which are considered statistically significant. NS indicates no significant difference was observed.

RESULTS

Overall structure of the double helical r(CAG)$_5$ RNA

The structures of CNG (N represents one of the four natural RNA nucleotides) repeats resolved through X-ray crystallography thus far have revealed only duplex conformation, despite their propensities to adopt multi-branched hairpin structures as well (34,36–40,52,53). In particular, the CAG RNA duplex containing two consecutive repeats of CAG adopts a regular A-form helical structure, whereas the duplex containing three repeats adopts an A′-form that comprises a wider major groove and a smaller inclination angle compared with the standard A-form RNA (37,40,54). To determine which helix form is adopted by longer CAG repeats, we solved a 1.45 Å resolution crystal structure of a 23-nt-long RNA oligo containing five repeats of CAG, which is the longest CAG RNA structure solved to date (Figure 1A). Our RNA model adopted an anti-parallel double-stranded helical structure with the two strands related by a 2-fold crystallographic axis. The surface electrostatic potential of our CAG RNA structure calculated using the Adaptive Poisson–Boltzmann Solver (APBS) (55) indicated that the major groove is predominantly electronegative. The minor groove contains alternating positive and negative patches along the helix axis; this finding is similar to that reported by Kiliszek *et al.* (37) (Figure 1B). In the crystal lattice, each double-stranded helix was packed with another two helices in an end-to-end manner to form pseudo-infinite CAG helices (Figure 1C).

Our crystal structure closely resembles an A′-form double helix with wider major groove (Supplementary Table S3). The sugar puckers in the duplex are biased toward a C3′-endo conformation, as observed in other RNA structures (37) (Supplementary Table S4). The average helical rise for the CAG duplex is 2.98 Å, which is larger than the helical rise of A-form RNA (2.81 Å); the helical rise of 2.98 Å is more favorably matched with the helical rise of A′-form RNA (3.00 Å) (Supplementary Table S3). These observations indicate that long CAG repeat transcripts preferentially adopt the A′-form over the regular A-form.

The C1′–C1′ distances between the non-canonical paired residues ranged from 10.8 to 11.6 Å, with the center A12:A12 base pair being the widest (Figure 1D). These widths are larger than those in the standard Watson–Crick base pair (10.5 Å on average) (Figure 1D). Such displacement of the glycosidic bond from their Watson–Crick po-

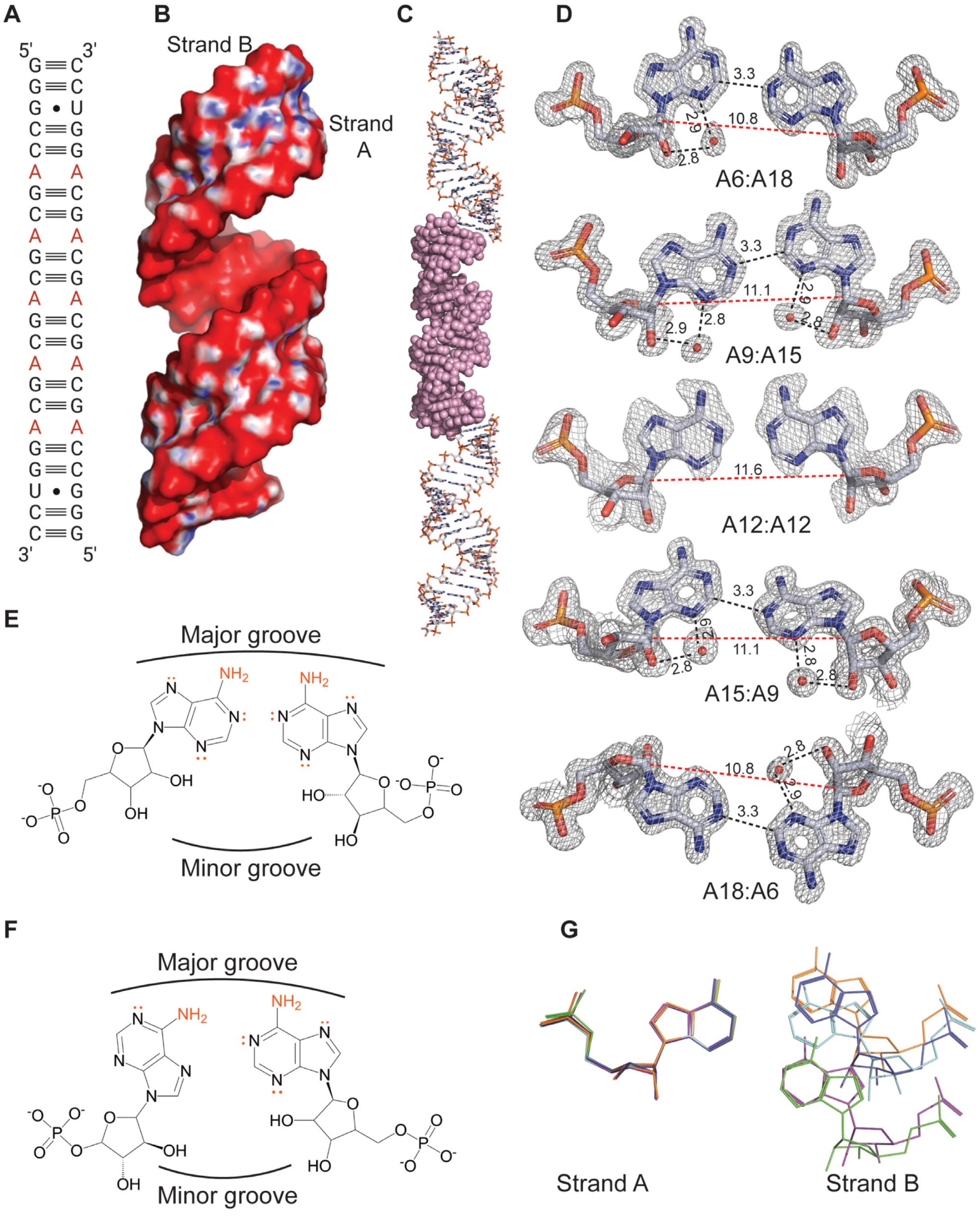

Figure 1. r(CAG)$_5$ RNA adopts A′-form double-stranded helix. (**A**) Secondary structure of CAG RNA duplex containing five constitutive CAG trinucleotide repeats. (**B**) Solvent-accessible surfaces of the r(CAG)$_5$ RNA duplex are colored by electrostatic potential. (–5 to +5 kT/e in *blue* to *red*) computed by the APBS program (55). (**C**) r(CAG)$_5$ RNA duplexes pack head-to-tail in a pseudo-continuous helix. Three duplexes are shown and the middle duplex is colored in pink. (**D**) Electron density maps (grey mesh) of five pairs of A·A base pair contoured at 1.0 σ level. The base pairs are shown in the stick mode colored according to atom type (N, blue; O, red; and P, orange), red spheres represent water molecules. Hydrogen bonds are denoted with black dashed lines. Red dashed lines denoted C1′ atoms. The distances of the interactions are given in Angstroms. (E, F) Arrangement of H-bond donors and acceptors at the major and minor grooves of mismatched A·A base pairs adopting anti-anti conformation (**E**) and syn-anti conformation (**F**), respectively. (**G**) Comparison of the orientations of adenosines in the mismatched A·A base pairs. A6:A18 pair is colored in green. A9:A15 pair is in orange. A12:A12 pair is in cyan. A15:A9 pair is in purple. A18:A6 pair is in blue.

sitions prevents a steric clash between the two adenosines and might alter the accessible surface on the minor groove side, particularly in the case of the A12:A12 base pair, which represents a transition state between two anti–anti conformations (39). The C–G and G–C base pairs in our structure formed standard Watson–Crick interactions. By contrast, the non-canonical A·A base pairs, except for the A12:A12 base pair, are stabilized by weak C2H2···N1 hydrogen bonds as reported in previous studies (Figure 1D) (37,39,40). In addition, water molecules also appear to play some roles in stabilizing the A·A base pair. Again, with the exception of A12:A12, at least one water molecule helps to maintain the orientation of adenine ring(s) in the A·A base pair by mediating hydrogen bonds between the N3 of the adenine and the O2′ of ribose (Figure 1D).

Similar to the CAG RNA structures reported by Kiliszek *et al.*, all non-canonical A·A base pairs in our r(CAG)$_5$ RNA structure uniformly adopt the anti–anti conformation, which has been found to be a favorable arrangement in double-stranded helical structure *in silico* studies (39). Compared with the syn–anti conformation observed in other CAG RNA structures, the anti-anti conformation appears to position an additional hydrogen bond acceptor on the minor groove side for the binding of proteins or ligands (37,39,40) (Figure 1E and F; Supplementary Figure S1). The A·A base pairs appear to adopt diverse conformations in our structure. We, therefore, superimposed the different adenosines in strand A to compare the conformations of the base pairs (Figure 1G). The A12 of strand B, which is considered to adopt a transitional conformation, protrudes into neither the major nor minor groove side when compared with the other four A bases. Such conformation increases the accessibility of the bases on the minor groove side and may provide a new binding site for ligands or proteins. Taken together, no dominant mode of the A·A base pair was observed (Figure 1G).

Model of NCL–RRM2 complexed to r(CAG)$_5$ RNA

We previously demonstrated that the RRM2 and RRM3 of NCL, rather than RRM1 and RRM4, are responsible for the interaction with expanded CAG RNA (22). We also identified a peptide inhibitor P3, derived from the first RNP region of NCL–RRM2, that inhibited the interaction between NCL and expanded CAG RNA and reduced RNA-induced toxicity both *in vitro* and *in vivo* (46). A structure–activity relationship (SAR) study of the inhibitor revealed the residues K3, K5, Y9 and K13, which correspond to the K427, K429, Y433 and K437 of NCL–RRM2, respectively, are key residues responsible for CAG RNA binding. To examine whether these residues are crucial for RNA binding in the context of NCL, we mutated K427, K429, Y433 and K437 in NCL–RRM2/3 to alanines (Figure 2A). Ataxin-2 CAG transcripts, which have been shown to interact with an NCL-derived peptide in a length-dependent manner (47), were applied in the electrophoretic mobility shift assay (EMSA) to study the effects of the mutations. The EMSA results revealed that K427 and K429 were crucial for the interaction between NCL–RRM2/3 and the *SCA2*$_{CAG42}$ transcripts (Figure 2A). By contrast, neither the alanine substitution of Y433 nor the additional mutation of K437 to alanine in the context of the K427/K429A mutant significantly reduced RNA-binding activity. On the basis of these results, we performed data-driven molecular docking by using the HADDOCK 2.2 server (50) to predict the possible interface between NCL–RRM2 and expanded CAG RNA. The three-dimensional structures of the r(CAG)$_5$ RNA we solved (Figure 1) and the NCL–RRM2 derived from PDB code 2KRR were used for docking. Since NCL specifically binds CAG repeats (22), all three A·A base pairs within the middle section of our CAG RNA structure were specified for the docking. K427 and K429 were selected as the active NCL residues to perform HADDOCK docking runs based on our EMSA results.

On the basis of the findings of cluster analysis performed using the highest HADDOCK scores, two potential models that bind to the minor and major grooves, respectively, were obtained (Figure 2B and C). Among the two models, the one with NCL–RRM2 that bound to the minor groove of the double-helical CAG RNA appeared to be more favorable based on the larger interaction surface area and more direct interactions between NCL residues and the CAG RNA (Figure 2B, Supplementary Table S5). To determine which model represents the true binding mode, we identified the RNA-binding NCL residues in both models and performed alanine substitution mutagenesis of each residue independently (Figure 2B and C). EMSA was performed to analyze the effects of these alanine substitutions on expanded CAG RNA binding (Figure 2D, Table 2). Our results revealed that the alanine substitution of D425, which was predicted to be a key residue interacting with the A·A base pair directly in the major groove-binding model, failed to affect CAG RNA binding (Figure 2C and D). Q406, which appears to interact with the backbone of the RNA in the major groove-binding model only, also did not affect the interaction with CAG RNA when mutated to alanine. However, the mutation of R420, another residue that interacts with the RNA in the major groove-binding model, inhibited RNA binding (Figure 2D). The effect of R420 on RNA binding is likely due to the alteration of the secondary structure of RRM, as revealed through circular dichroism (CD) spectroscopy (Supplementary Figure S2).

Contrarily, alanine mutations at N399 and K424, which were identified from the minor groove–binding model, significantly inhibited the NCL:CAG RNA interaction. Mutation of Y402, which binds RNA in both models, also weakened RNA binding (Figure 2D). To determine whether Y402 truly binds to the minor groove of CAG RNA, we generated a double alanine substitution mutant of Y402 and K403, whose side-chain interacts with the RNA in the minor groove–binding model but not the major groove–binding model. The results of EMSA indicated that the interaction between CAG RNA and the double mutant Y402A/K403A was further decreased, suggesting that both Y402 and K403 participate in minor groove binding (Figure 2D). Moreover, mutations of K398, despite to a lower extent, and R457 to alanine also inhibited RNA binding. These results indicated that NCL binds to the minor groove of CAG RNA.

To ensure that changes in the RNA-binding property are not caused by alterations in the overall protein struc-

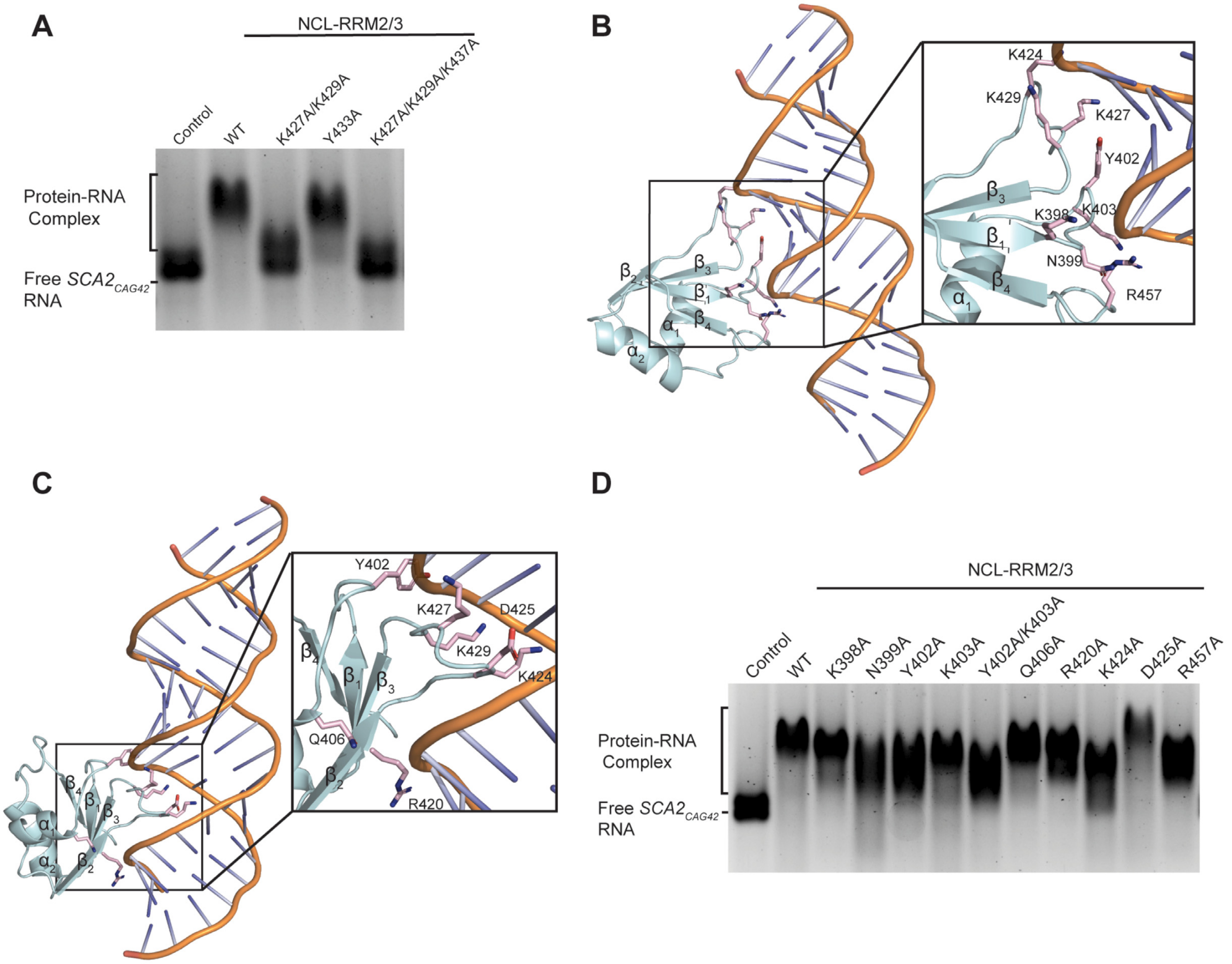

Figure 2. Identification of NCL residues crucial for CAG RNA binding. (**A**) Representative image of EMSA analysis on the interaction of *SCA2*$_{CAG42}$ RNA with NCL–RRM2/3 or different NCL–RRM2/3 mutants that were generated based on the previously identified P3 inhibitor. (B, C) Two docking models of NCL–RRM2 with r(CAG)$_5$ RNA generated by HADDOCK server. Models of NCL–RRM2 binding to the minor groove (**B**) and major groove (**C**) are shown respectively. Residues involved in the protein–RNA interaction are depicted in sticks. (**D**) Representative image of EMSA analysis on the interaction of *SCA2*$_{CAG42}$ RNA with NCL–RRM2/3 or different NCL–RRM2/3 mutants generated based on the interacting residues identified from the two docking models. All EMSAs were repeated at least three times with similar results.

Table 2. Summary of the effects of different NCL–RRM2/3 mutants on *SCA2*$_{CAG42}$ RNA binding

NCL–RRM2/3 mutants	Predicted binding site	Disruption of protein–RNA interaction
K398A	Minor groove	Yes
N399A	Minor groove	Yes
Y402A	Both	Yes
K403A	Minor groove	Yes
Y402A/K403A	Minor groove	Yes
Q406A	Major groove	No
R420A	Major groove	Yes
K424A	Both	Yes
D425A	Major groove	No
R457A	Minor groove	Yes

ture after mutagenesis, CD spectroscopy was performed to compare the secondary structural compositions of wild-type and mutant NCL–RRM2/3 (Supplementary Figure S2). Our CD spectroscopy results revealed that the overall protein structures of most of the NCL–RRM2/3 mutants remained intact. One of the exceptions is the aforementioned residue R420, in which the alanine substitution resulted in a change in its secondary structure content. This finding explains the observed inhibitory effect of R420 on RNA binding despite it being identified from the major groove–binding model. Another exception is Y433A, which did not affect the interaction between the RRMs and CAG RNA and thus was not included in subsequent experiments. Finally, the mutation of N399, which indirectly mediates the RRM–RNA interaction, altered the secondary structure of NCL–RRM2/3. Hence, this mutation was excluded from our final mutant construct. Overall, our results indicated that the residues we identified are key determinants for CAG RNA binding.

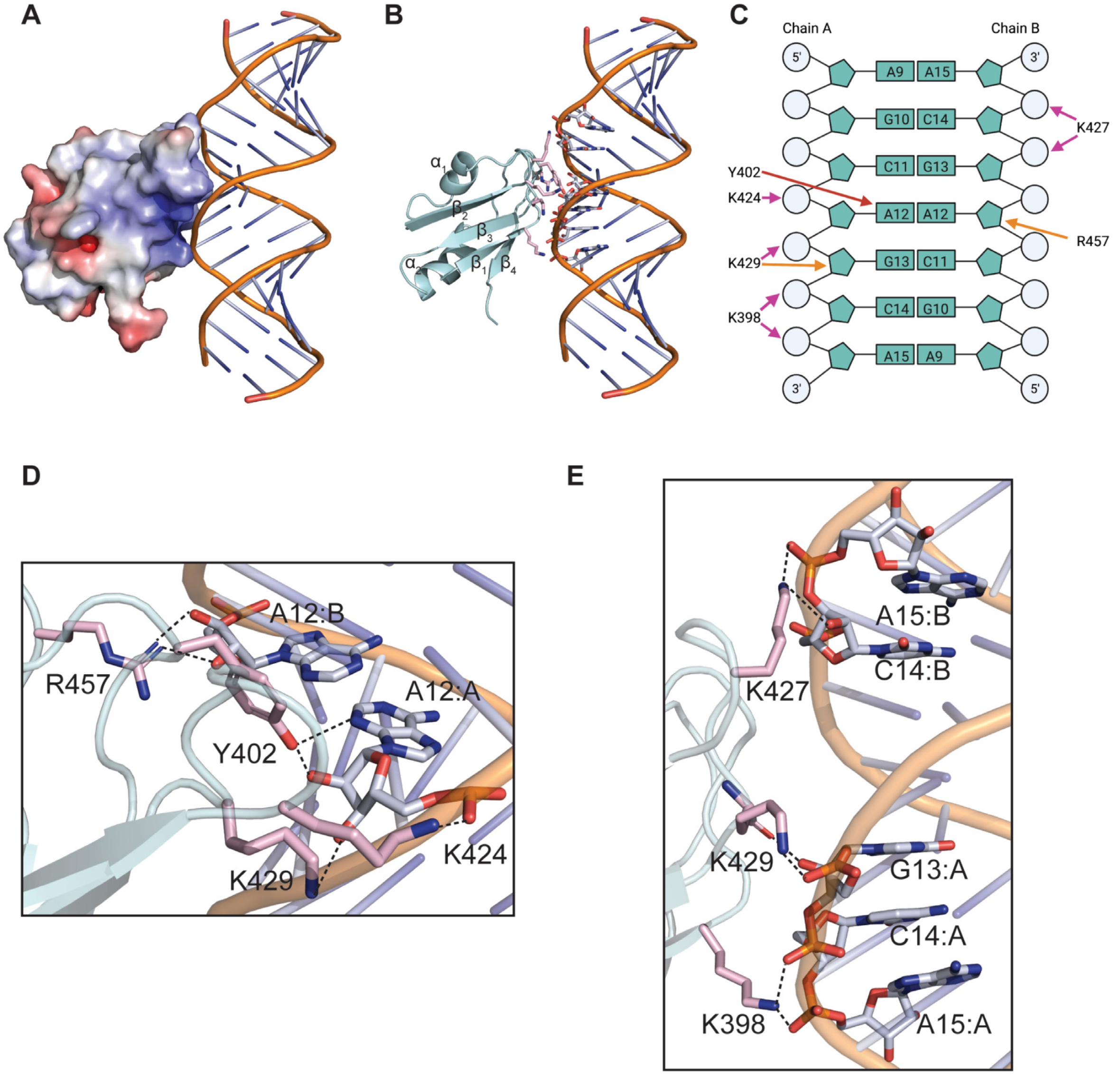

Figure 3. Potential interaction between NCL–RRM2:r(CAG)$_5$ RNA docking model. (**A**) Overview of the interaction model of NCL–RRM2 with r(CAG)$_5$ RNA obtained after performing HADDOCK docking simulations based on the EMSA analysis. NCL–RRM2 solvent-accessible surfaces are colored by the electrostatic surface potential computed by APBS (55). (**B**) Three finger-like projections from NCL–RRM2 interact with the minor groove of the RNA duplex. Residues involved in the protein-RNA interaction are depicted in sticks. (**C**) Base-pairing diagram of the RNA region that interacts with NCL–RRM2. Red arrow denotes the interaction between the NCL–RRM2 residues and the nitrogenous base of the CAG RNA duplex, pink arrows denote the interactions between NCL–RRM2 and the phosphate backbone; and orange arrows denote the interactions between NCL–RRM2 residues and the ribose sugars. (**D, E**) Detailed view on the interaction between amino acids on NCL–RRM2 and the non-canonical A·A pair at the minor groove of r(CAG)$_5$ RNA.

Mechanism of RNA binding by NCL–RRM2

To refine our NCL–RRM2:CAG RNA complex model, the experimentally determined key amino acids, namely K398, Y402, K424, K427, K429 and R457, were selected as the active residues to repeat the molecular docking. The comparison of HADDOCK scores and the analysis of both NCL–RRM2 complex models again suggested that the minor groove binding model is more favorable and the resulting model likely represents the true complex (Supplementary Table S5).

Examination of the surface electrostatic potential of the minor groove–bound complex indicated that the selected active residues clustered together to form a positively charged surface on NCL–RRM2 that would be favorable for nucleic acid binding (Figure 3A). By contrast, Y433 and K437 are positioned on strand β_3, which is not part of the interaction interface of the RRM2 domain (Supplementary Figure S3); this explains why both point mutations did not affect CAG binding. R420, which was found to be crucial in our EMSA, remains solvent-exposed on strand β_2 and does not participate in RNA binding (Supplementary Figure S3). Since its mutation to alanine altered the secondary structure of RRM2/3 in our CD studies (Supplementary Figure S2), we speculate that R420 might interact

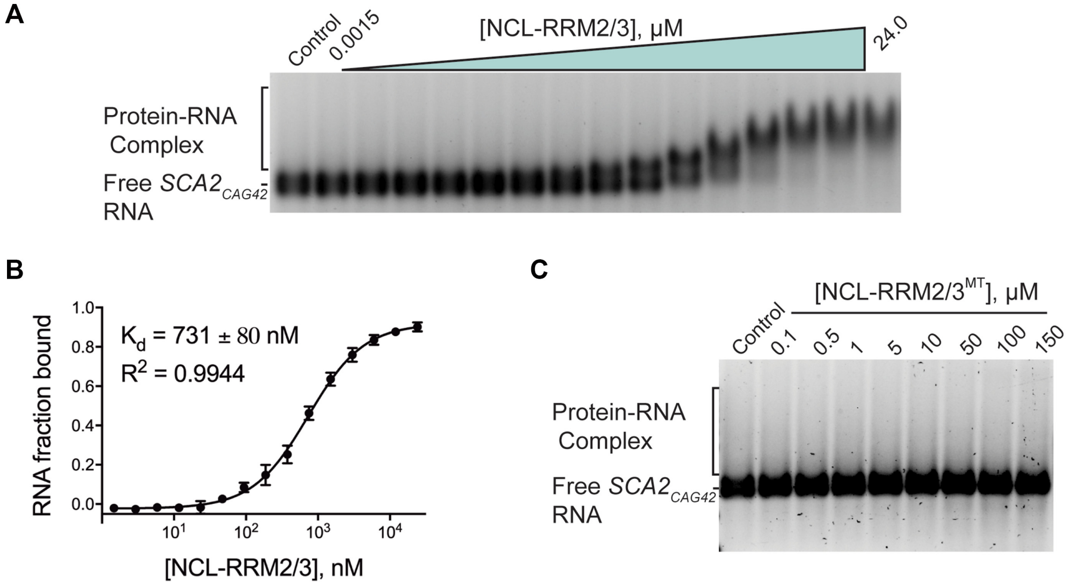

Figure 4. Effect of the mutations of six residues identified from the docking model. (A, B) EMSA was performed to measure the binding affinity between NCL–RRM2/3 and *SCA2*$_{CAG42}$ RNA. (**A**) Representative gel shift image. (**B**) NCL–RRM2/3 binds to *SCA2*$_{CAG42}$ RNA with an apparent K_d value is 731 ± 80 nM. The binding curve represents the overall fitting against three independent experiments. (**C**) NCL–RRM2/3^{MT}, a mutant containing six residues identified from the docking model failed to interact with *SCA2*$_{CAG42}$ RNA. All EMSAs were repeated at least three times with similar results.

with RRM3 to facilitate the formation of NCL:CAG RNA complex.

The refined model (Figure 3B) revealed that all selected active residues are distributed on the β_1–α_1 loop, β_2–β_3 loop and α_2–β_4 loop of NCL–RRM2. These loops form three finger-like protrusions that bind to the RNA and engage the sugar–phosphate backbone and base pairs within the minor groove (Figure 3B). We observed that K403 within the β_1–α_1 loop interacts with the phosphate backbone of C14 in the chain B RNA and forms an ionic pair with the side-chain of D455 located on the α_2–β_4 loop (Supplementary Figure S4A). Because the mutation of K403 to alanine enhanced the inhibition of the interaction between NCL–RRM2/3 and RNA in the context of Y402, we speculated that in addition to interacting with the RNA, the observed ionic pair of K403–D455 in our model is crucial for maintaining and stabilizing the three-dimensional structure or the RNA-binding conformation of NCL–RRM2. To test this, we mutated D455 to alanine to disrupt its ionic interaction with K403, or both K403 and D455 to oppositely charged side-chains (K403D and D455K) to maintain the ionic pair yet disrupt the interaction with the RNA backbone. The EMSA results demonstrated that D455A interfered the interaction between NCL–RRM2/3 and RNA, but the K403D/D455K double mutant partially diminished the effect (Supplementary Figure S4B), indicating that the ionic interaction between K403 and D455 plays a role in RNA binding.

The selected active residues in NCL–RRM2 form direct interactions with seven nucleotides located at the mid-region of the r(CAG)$_5$ RNA (Figure 3C). Intriguingly, Y402, hydrogen bonds to one of the accessible adenine N7 atoms on the minor groove side, is the only residue that appears to directly interact with the adenine within the A·A base pair, suggesting it may play a vital role in the specificity of NCL towards CAG RNA (Figure 3D, Supplementary Figure S5A). On the other hand, the other active residues mostly mediate direct interactions with the backbone phosphates and riboses of the RNA helix (Figure 3E, Supplementary Figure S5B–D). Taken together, the active residues help to orient NCL–RRM2 in a favorable position for binding to CAG RNA through an extensive network of interactions.

Mutation of six key residues of NCL–RRM2/3 abolishes CAG RNA binding

To further validate our NCL–RRM2:CAG RNA model, we selected and mutated six aforementioned NCL–RRM2 residues that directly interact with the CAG RNA, namely K398, Y402, K424, K427, K429 and R457, in the context of NCL–RRM2/3, to alanines, hereafter referred to as NCL–RRM2/3^{MT}. K403 was excluded due to its potential role in stabilizing the RNA-binding conformation of RRM2.

The ability of NCL–RRM2/3^{MT} to bind to expanded CAG RNA was examined. The binding affinity of wild-type NCL–RRM2/3 (NCL–RRM2/3^{WT}) for *SCA2*$_{CAG42}$ RNA was first determined with an apparent K_d value of 731 ± 80 nM by performing the EMSA (Figure 4A and B). By contrast, under the same experimental condition, no shift was observed even when a 6-fold higher amount of NCL–RRM2/3^{MT} was titrated to *SCA2*$_{CAG42}$ RNA (Figure 4C). This result demonstrated that the mutated residues are the key determinants for the NCL–RRM2/3:CAG RNA interaction during the pathogenesis of polyQ diseases.

Binding interface for CAG RNA is not important for the normal gene expression regulation function of NCL

To confirm the functional importance of the six identified residues, we mutated them to alanines in the context of

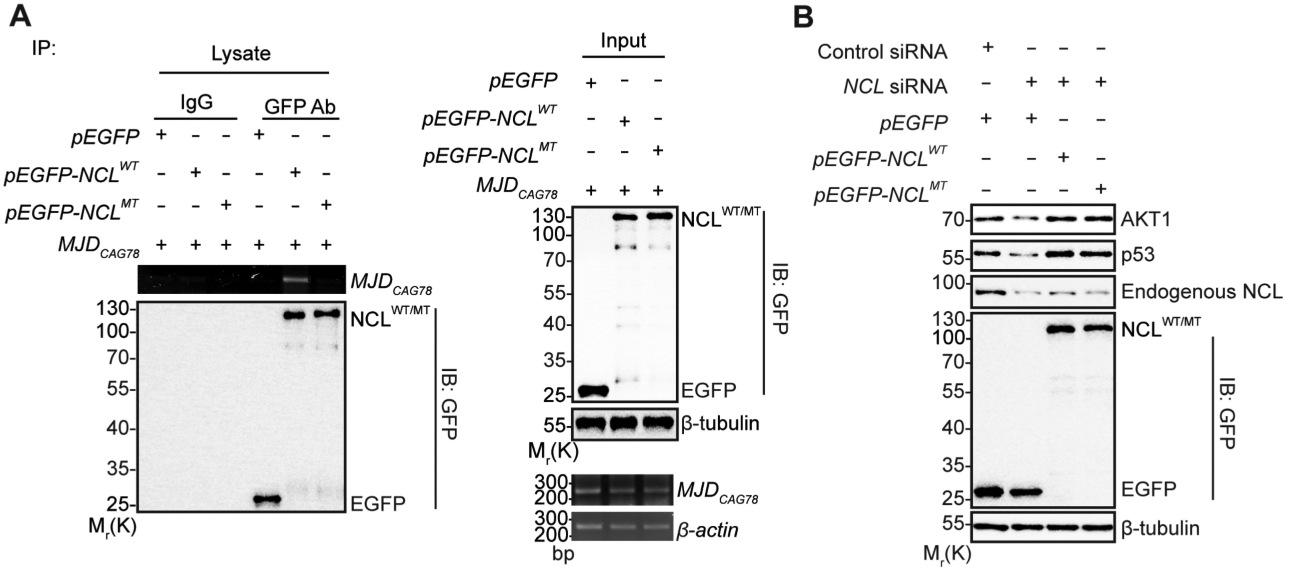

Figure 5. NCL^{MT} disrupts the binding with expanded CAG RNA but retains the normal gene expression regulation function of NCL. (**A**) NCL^{WT} protein interacted with the expanded *MJD*$_{CAG78}$ RNA, while NCL^{MT} protein failed to interact with the RNA. immunoprecipitation (IP) assay was performed using total cell lysates from SK-N-MC cells co-transfected with *pcDNA3.1(+)-S1-MJD*$_{CAG78}$ plasmid and *pEGFP* vector or *pEGFP-NCL* constructs to investigate the binding of NCL^{WT} or NCL^{MT} to expanded CAG RNA. EGFP or EGFP-NCL proteins were immunoprecipitated using anti-GFP antibody and 20% of the samples were subjected to immunoblotting. RNA was extracted from the rest of the samples and analyzed by RT-PCR. 10% total cell lysate was used as input. The expression of *MJD*$_{CAG78}$ RNA, NCL^{WT}, and NCL^{MT} were detected in the input samples. (**B**) The protein levels of AKT1 and p53 were diminished upon knockdown of endogenous *NCL*. Both NCL^{WT} and NCL^{MT}, but not the EGFP protein, restored the expression of AKT1 and p53. The EGFP, NCL^{WT} and NCL^{MT} proteins were expressed at comparable levels. All experiments were repeated at least three times with similar results.

full-length NCL and investigated whether the interaction with expanded CAG RNA would be affected in SK-N-MC cells. An *MJD*$_{CAG78}$-transfected cell model was used because neurotoxicity in polyQ diseases occurs in a length-dependent manner (47) and our previous studies have reported that the *MJD*$_{CAG78}$ construct induces more severe cell death (22,56). In brief, *EGFP-NCL* wild-type (*NCL*^{WT}) or mutant (*NCL*^{MT}) was co-expressed with *MJD*$_{CAG78}$ construct in SK-N-MC cells, respectively. Immunoprecipitation of EGFP or EGFP-NCL proteins followed by RT-PCR analysis revealed that NCL^{WT}, but not NCL^{MT}, interacted with *MJD*$_{CAG78}$ RNA (Figure 5A). This result demonstrated that the mutated residues are the key determinants for the NCL:expanded CAG RNA interaction.

Although it is well-established that the interaction between NCL and expanded CAG RNA sequesters it from its cognate substrates during the pathogenesis of polyQ diseases, whether expanded CAG RNA competes with other substrates for the same binding site(s) on NCL remains unclear. Hence, we examined whether the mutations of the CAG RNA-interacting residues affect NCL's normal cellular functions in the regulation of the expression of *AKT1* (17) and *TP53* (57) genes. We first performed the gene knockdown of *NCL* in SK-N-MC cells by RNAi using siRNA that would not interfere with the expression of exogenous *NCL*^{WT}or *NCL*^{MT}. The knockdown efficiency was then determined through Western blot (Figure 5B). Consistent with previous findings (17,57), the expression of both AKT1 and p53 were significantly decreased upon NCL silencing. Next, we reconstituted the *NCL*-knockdown cells with EGFP-NCL^{WT} and EGFP-NCL^{MT}, respectively. Our results revealed that the expression of EGFP-NCL^{MT} restored the levels of AKT1 and p53 in the same manner as NCL^{WT} did in SK-N-MC cells (Figure 5B), indicating that the mutations of the CAG RNA–interacting residues did not affect the normal gene expression regulation function of NCL.

NCL^{MT} is more effective in rescuing expanded CAG RNA-induced cell death

Our previous study demonstrated that expanded CAG RNA sequesters NCL from binding with the upstream control element (*UCE*) of the rRNA promoter, thereby inhibiting rRNA transcription and subsequently leading to nucleolar stress–dependent cell apoptosis (22). Inhibition of the direct interaction between NCL and expanded CAG RNA by using either our competitive inhibitor P3 (46) or BIND (47) alleviated nucleolar stress in expanded CAG RNA-expressing cells (22). Because the mutations of six key residues identified in this study could disrupt the binding of NCL to expanded CAG RNA, we speculated that the expression of NCL^{MT}, which still possesses normal cellular functions, could alleviate nucleolar stress in expanded CAG RNA-expressing cells by binding to the *UCE* of the rRNA promoter to facilitate and restore rRNA synthesis. To verify this speculation, we expressed *NCL*^{WT} or *NCL*^{MT} in cells transfected with the *MJD*$_{CAG78}$ plasmid and performed RT-PCR to compare the level of pre-45s rRNA (Figure 6A). Our results indicated that the expression of expanded CAG RNA downregulated rRNA transcription, as expected. When *NCL*^{WT} was co-expressed in the *MJD*$_{CAG78}$-expressing cells, its pre-rRNA level was partially recovered. More importantly, the pre-rRNA level was

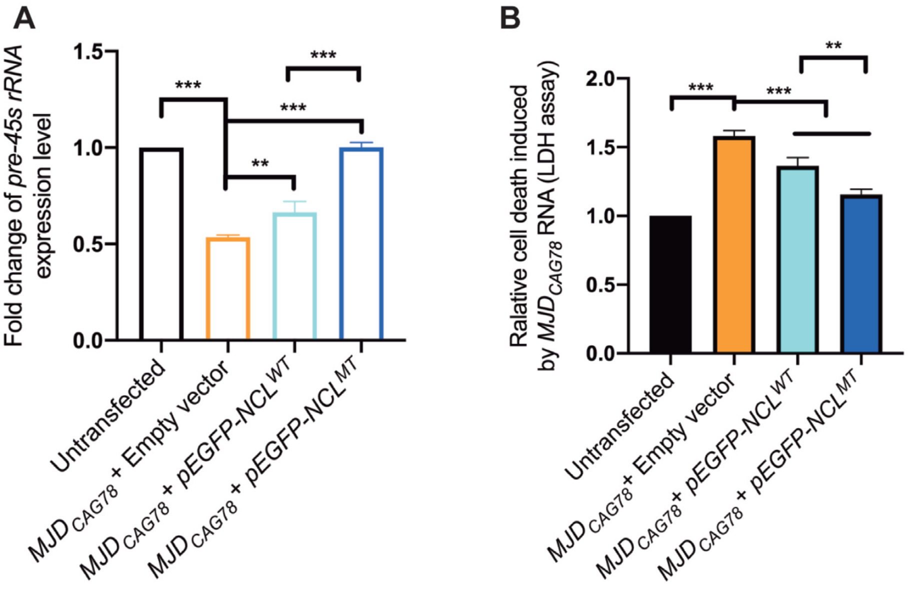

Figure 6. Mutant NCL^{MT} effectively restores the expression level of pre-45s rRNA and alleviates cytotoxicity in polyQ disease model. (**A**) The pre-45s rRNA expression level was reduced in MJD_{CAG78}-expressing SK-N-MC cells. When compared to NCL^{WT}, NCL^{MT} showed enhanced capability in recovering the pre-45s rRNA expression level. The qRT-PCR data from the transfected groups were normalized to that from the untransfected control group. (**B**) The MJD_{CAG78}-induced cell death was suppressed upon co-overexpression of NCL^{WT} or NCL^{MT}, while NCL^{MT} was more effective in inhibiting the cell death than NCL^{WT}. The LDH assay was carried out following the manufacturer's instructions 48 h post-transfection. The results from the transfected groups were normalized to that from the untransfected control group. n=3; error bars represent standard deviation.

fully restored when NCL^{MT} was co-expressed, suggesting that the NCL^{MT} not only restored NCL's cellular function during rRNA transcription but also functioned more favorably than the wild-type protein.

We next investigated whether the expression of NCL^{WT} can inhibit CAG RNA-induced cell death. The expression of MJD_{CAG78} alone led to severe cell death when compared with the untransfected control (Figure 6B). By contrast, the co-expression of either NCL^{WT} or NCL^{MT} significantly suppressed CAG RNA-induced cell death. With effects similar to those observed in the restoration of rRNA synthesis, NCL^{MT} performed better than its wild-type counterpart in rescuing cell death.

DISCUSSION

Expansion of CNG repeats in the human genome leads to the development of various hereditary trinucleotide repeat expansion diseases (58,59). Studies on the secondary structure of expanded CNG repeats–containing RNA that have used various biochemical and biophysical models have suggested that all CNG repeats adopt moderately stable hairpin structures (33,60,61). On the other hand, all the existing crystal structures of the short repeats of different CNG indicate that the RNA forms double-helical duplexes (34,36–40,52,53). Thermodynamic studies of RNA oligomers containing different numbers of trinucleotide repeats provide insights into such differences. Broda *et al.* reported that duplexes were formed when RNA oligomers contain only 2–3 CNG repeats (62). When trinucleotide repeats increased to 4–5 units, the oligomers formed both duplexes and hairpins. However, only hairpins were observed when the length of oligomers further increased. To solve the crystal structure of CAG repeats in its hairpin form, we performed crystallographic studies by using RNA oligomers that contain five or six repeats of CAG. Despite extensive trials, only oligomers containing five CAG repeats but not six CAG repeats were crystallized. Our crystal structure revealed that the five CAG repeat-containing RNA forms duplexes and adopts an A′-form RNA double helix (Figure 1). Although the oligomer does not adopt a hairpin structure, it exhibits obvious differences when compared with other CNG RNA structures that contain at least five repeats. For instance, the non-canonical G·G base pairs preferentially adopt the syn–anti conformation in CGG transcript (PDB code: 4KQ0) (Supplementary Figure S6A and C, and Table S3) to fit two guanines in the helix more snuggly (53,63), while the equally bulky A·A base pairs in our structure adopt the anti–anti 1 × 1 nucleotide internal loop as the major conformation. On the other hand, when considering the CUG RNA structure with six repeats, all uridines adopt an anti conformation (34) (PDB code: 3GM7, Supplementary Figures S6B and D, and Table S3). The differences in the identity and conformation of N alter the H-bond acceptor/donor positioning and cause significant changes in h-rise and inclination, resulting in distinct groove widths that may confer unique protein/ligand-binding properties to the different CNG repeats RNA.

Molecular docking techniques are widely used to study the structural basis of the interaction between biological molecules (64–66). The data-driven HADDOCK applied in this study is one of the best performing docking methods given its ability to integrate experimental data during the process. On the basis of the results of our interaction studies here and those of previous studies, a new model of NCL–RRM2 in complex with the r(CAG)$_5$ RNA duplex was successfully generated to examine the molecular basis of their interaction. Our docking model showed that the CAG repeats RNA utilizes its minor groove to bind to the NCL–RRM2, providing new structural insights in how the NCL–RRM2, despite adopting the canonical RRM fold, functions as a double-stranded RNA binding domain (dsRBD) under the pathological conditions of CAG repeat expansion diseases. This result is consistent with the binding modes of other dsRBD, where they mainly interact with the minor groove side of their double-stranded RNA targets via direct readout of the minor groove sequence (67). In our model, NCL–RRM2 mostly interacts with the sugar–phosphate backbone of the r(CAG)$_5$ RNA, and only one residue, Y402 of NCL, directly contacts an adenine within the A·A base pair. We hypothesize that the A·A base pair might function by altering the widths of both the major and minor grooves of CAG repeats to create new sites for the binding and sequestration of cellular proteins instead of mediating an extensive network of unique interactions with its binding partners. However, because our r(CAG)$_5$ RNA crystal structure shows a duplex instead of a hairpin structure, we could not rule out that the A·A base pairs within expanded CAG repeats might adopt very different conformations to provide new protein binding sites that have yet to be uncovered. Although no protein:CAG RNA complex structure is available for comparison, a search for protein:CNG RNA complex structures on the PDB database revealed that the crystal structures of p19, a viral RNA silencing suppressor, in complex with double-helical UUG(CUG)$_5$CU 20mer or C(CUG)$_6$C 20mer RNAs have previously been solved (PDB codes: 4JGN and 4JNX). Similar to our model, the protein binds to the RNAs mainly via electrostatic interactions and hydrogen bond formation with the negatively charged phosphate–sugar backbone but not the non-canonical base pairs (Supplementary Figure S7). No direct interaction was observed between the side-chains of p19 and the nitrogenous bases of the RNA duplex. This finding concurs with our hypothesis that the non-canonical base pairs might alter the overall RNA structure for protein binding instead of providing unique site(s) for extensive interactions. Additional structural studies on protein:CNG RNA complexes at high resolution are required to validate our hypothesis.

Our previous study reported that NCL RRM2 and RRM3 are crucial for the interaction with expanded CAG RNA (22). Although this study did not investigate the molecular mechanism underlying the RRM3–CAG RNA interaction, the sequence alignment of all four NCL RRMs indicated that four out of the six critical residues we identified in RRM2 are conserved in RRM3. Moreover, the tyrosine residue that interacts with the A·A base pair is conserved in both RRM2 and RRM3, suggesting that RRM3 might bind to CAG RNA in a similar manner (Supplementary Figure S8). Two critical lysine residues, K398 and K424, in NCL–RRM2 are replaced by serine and glutamine residues, respectively, in NCL–RRM3. The presence of two extra positively charged residues might confer RRM2 a more crucial role in CAG RNA binding and explain why mutations in RRM2 alone are adequate to abolish the interaction with CAG RNA.

The results of our docking model and interaction studies revealed the molecular basis of how RNA-binding protein containing canonical RRM interacts with trinucleotide RNA repeats or RNA duplexes in general. In addition, we identified an NCL mutant that was not sequestered by expanded CAG RNA both *in vitro* and *in vivo* and could perform normal cellular functions of NCL like gene expression and rRNA synthesis regulation. More importantly, the mutant could alleviate nucleolar stress by restoring pre-45s rRNA level and rescue expanded CAG RNA-induced cell death more effectively than wild-type NCL. NCL binds to different cellular substrates including pre-rRNA (11,13); *c-MYC* promoter (68); pre-mRNA (69); *GADD45A* mRNA (RNA encoding for the growth arrest- and DNA damage inducible- 45α protein [Gadd45α]) (70); *AKT1* mRNA (17); and Hdm2 protein that regulates p53 protein expression level (57). These interactions are vital for the multifunctional roles of NCL in the regulation of rRNA transcription and ribosome assembly, mRNA stability, and translation. Our works provide evidence that NCL likely binds these cognate substrates through mechanism(s) different from that of CAG RNA and mutations at the CAG RNA-binding interface do not affect its essential cellular functions.

NCL has been shown to be sequestered by hexanucleotide RNA repeats during the pathogenesis of C9ORF72-mediated amyotrophic lateral sclerosis and frontotemporal dementia (ALS/FTD) (24–26). It will be of high interest to investigate whether the RNA-binding mechanism of NCL is conserved and whether our mutant can rescue the RNA-toxicity in ALS/FTD. We believe our NCL mutant has provided a new basis and opportunity for the development of novel therapy to treat polyQ diseases. Therapeutic strategies that facilitate the expression of the NCL mutant or its direct cellular delivery might alleviate the nucleolar stress induced by the expanded CAG RNA. Such strategies will have the advantage of escaping sequestration by the disease-causing RNA.

DATA AVAILABILITY

All relevant data are available from the Lead Contact upon request. The coordinates of r(CAG)$_5$ RNA have been deposited in the Protein Data Bank under the accession number 7VFT.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the technical services provided by the ‘Synchrotron Radiation Protein Crystallography Facility of the National Core Facility Program for Biotechnology, Ministry of Science and Technology’ and the ‘National Synchrotron Radiation Research Center’, a national user facility supported by the Ministry of Science and Technology of Taiwan, ROC.

FUNDING

General Research Fund from the Research Grants Council of Hong Kong SAR [14122319]; State Key Laboratory of Agrobiotechnology ITC Fund (to J.C.K.N.); CUHK Lui Che Woo Institute of Innovative Medicine BRAIN Initiative [8303404]; CUHK Group Research Scheme [3110102]; CUHK Vice-Chancellor's One-Off Discretionary Fund [4930713, VCF2014011]; CUHK One-off Funding for Joint Lab/Research Collaboration [3132980]; CUHK Gerald Choa Neuroscience Centre [7105306]; CUHK Faculty of Science Strategic Development Fund, and donations from the Hong Kong Spinocerebellar Ataxia Association [6903291 to H.Y.E.C.]. Funding for open access charge: General Research Fund from the Research Grants Council of Hong Kong SAR [14122319].

Conflict of interest statement. None declared.

REFERENCES

1. Bauer,P.O. and Nukina,N. (2009) The pathogenic mechanisms of polyglutamine diseases and current therapeutic strategies. *J. Neurochem.*, **110**, 1737–1765.
2. Lieberman,AndrewP., Shakkottai,VikramG. and Albin,R.L. (2019) Polyglutamine repeats in neurodegenerative diseases. *Annu. Rev. Pathol. Mech. Dis.*, **14**, 1–27.
3. Fiszer,A. and Krzyzosiak,W.J. (2013) RNA toxicity in polyglutamine disorders: concepts, models, and progress of research. *J. Mol. Med. (Berl)*, **91**, 683–691.
4. Nalavade,R., Griesche,N., Ryan,D.P., Hildebrand,S. and Krauss,S. (2013) Mechanisms of RNA-induced toxicity in CAG repeat disorders. *Cell Death. Dis.*, **4**, e752.
5. Spada,A.R.L., Wilson,E.M., Lubahn,D.B., Harding,A.E. and Fischbeck,K.H. (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, **352**, 77–79.
6. Finkbeiner,S. (2011) Huntington's disease. *Cold Spring Harb. Perspect. Biol.*, **3**, a007476.
7. Koide,R., Ikeuchi,T., Onodera,O., Tanaka,H., Igarashi,S., Endo,K., Takahashi,H., Kondo,R., Ishikawa,A., Hayashi,T. *et al.* (1994) Unstable expansion of CAG repeat in hereditary dentatorubral–pallidoluysian atrophy (DRPLA). *Nat. Genet.*, **6**, 9–13.
8. Di Prospero,N.A. and Fischbeck,K.H. (2005) Therapeutics development for triplet repeat expansion diseases. *Nat. Rev. Genet.*, **6**, 756–765.
9. Lee,D., Lee,Y.I., Lee,Y.S. and Lee,S.B. (2020) The mechanisms of nuclear proteotoxicity in polyglutamine spinocerebellar ataxias. *Front. Neurosci.*, **14**, 489.
10. Abdelmohsen,K. and Gorospe,M. (2012) RNA-binding protein nucleolin in disease. *RNA Biol.*, **9**, 799–808.
11. Allain,F.H., Bouvet,P., Dieckmann,T. and Feigon,J. (2000) Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin. *EMBO J.*, **19**, 6870–6881.
12. Ginisty;,H., Amalric;,F. and Bouvet,P. (1998) Nucleolin functions in the first step of ribosomal RNA processing. *EMBO J.*, **17**, 1476–1486.
13. Ginisty,H., Serin,G., Ghisolfi-Nieto,L., Roger,B., Libante,V., Amalric,F. and Bouvet,P. (2000) Interaction of nucleolin with an evolutionarily conserved pre-ribosomal RNA sequence is required for the assembly of the primary processing complex. *J. Biol. Chem.*, **275**, 18845–18850.
14. Bouvet,P., Diaz,J.J., Kindbeiter,K., Madjar,J.J. and Amalric,F. (1998) Nucleolin interacts with several ribosomal proteins through its RGG domain. *J. Biol. Chem.*, **273**, 19025–19029.
15. Jia,W., Yao,Z., Zhao,J., Guan,Q. and Gao,L. (2017) New perspectives of physiological and pathological functions of nucleolin (NCL). *Life Sci.*, **186**, 1–10.
16. Qiu,W., Zhou,F., Zhang,Q., Sun,X., Shi,X., Liang,Y., Wang,X. and Yue,L. (2013) Overexpression of nucleolin and different expression sites both related to the prognosis of gastric cancer. *APMIS*, **121**, 919–925.
17. Abdelmohsen,K., Tominaga,K., Lee,E.K., Srikantan,S., Kang,M.J., Kim,M.M., Selimyan,R., Martindale,J.L., Yang,X., Carrier,F. *et al.* (2011) Enhanced translation by nucleolin via G-rich elements in coding and non-coding regions of target mRNAs. *Nucleic Acids Res.*, **39**, 8513–8530.
18. Pichiorri,F., Palmieri,D., De Luca,L., Consiglio,J., You,J., Rocci,A., Talabere,T., Piovan,C., Lagana,A., Cascione,L. *et al.* (2013) In vivo NCL targeting affects breast cancer aggressiveness through miRNA regulation. *J. Exp. Med.*, **210**, 951–968.
19. Becherel,O.J., Gueven,N., Birrell,G.W., Schreiber,V., Suraweera,A., Jakob,B., Taucher-Scholz,G. and Lavin,M.F. (2006) Nucleolar localization of aprataxin is dependent on interaction with nucleolin and on active ribosomal DNA transcription. *Hum. Mol. Genet.*, **15**, 2239–2249.
20. Parlato,R. and Kreiner,G. (2013) Nucleolar activity in neurodegenerative diseases: a missing piece of the puzzle?*J. Mol. Med. (Berl)*, **91**, 541–547.
21. Caudle,W.M., Kitsou,E., Li,J., Bradner,J. and Zhang,J. (2009) A role for a novel protein, nucleolin, in parkinson's disease. *Neurosci. Lett.*, **459**, 11–15.
22. Tsoi,H., Lau,T.C., Tsang,S.Y., Lau,K.F. and Chan,H.Y. (2012) CAG expansion induces nucleolar stress in polyglutamine diseases. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 13428–13433.
23. Tsoi,H. and Chan,H.Y. (2013) Expression of expanded CAG transcripts triggers nucleolar stress in huntington's disease. *Cerebellum*, **12**, 310–312.
24. Haeusler,A.R., Donnelly,C.J., Periz,G., Simko,E.A., Shaw,P.G., Kim,M.S., Maragakis,N.J., Troncoso,J.C., Pandey,A., Sattler,R. *et al.* (2014) C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature*, **507**, 195–200.
25. Babic Leko,M., Zupunski,V., Kirincich,J., Smilovic,D., Hortobagyi,T., Hof,P.R. and Simic,G. (2019) Molecular mechanisms of neurodegeneration related to C9orf72 hexanucleotide repeat expansion. *Behav. Neurol.*, **2019**, 2909168.
26. Tao,Z., Wang,H., Xia,Q., Li,K., Li,K., Jiang,X., Xu,G., Wang,G. and Ying,Z. (2015) Nucleolar stress and impaired stress granule formation contribute to C9orf72 RAN translation-induced cytotoxicity. *Hum. Mol. Genet.*, **24**, 2426–2441.
27. Turi,Z., Lacey,M., Mistrik,M. and Moudry,P. (2019) Impaired ribosome biogenesis: mechanisms and relevance to cancer and aging. *Aging (Albany NY)*, **11**, 2512–2540.
28. Deisenroth,C. and Zhang,Y. (2010) Ribosome biogenesis surveillance: probing the ribosomal protein-Mdm2-p53 pathway. *Oncogene*, **29**, 4253–4260.
29. Srivastava,M., Fleming,P.J., Pollard,H.B. and Burns,A.L. (1989) Cloning and sequencing of the human nucleolin cDNA. *FEBS Lett.*, **250**, 99–105.
30. Ginisty,H., Sicard,H., Roger,B. and Bouvet,P. (1999) Structure and functions of nucleolin. *J. Cell Sci.*, **112**, 761–772.
31. Tajrishi,M.M., Tuteja,R. and Tuteja,N. (2011) Nucleolin: the most abundant multifunctional phosphoprotein of nucleolus. *Commun. Integr. Biol.*, **4**, 267–275.
32. Krzyzosiak,W.J., Sobczak,K., Wojciechowska,M., Fiszer,A., Mykowska,A. and Kozlowski,P. (2012) Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nucleic Acids Res.*, **40**, 11–26.
33. Sobczak,K., de Mezer,M., Michlewski,G., Krol,J. and Krzyzosiak,W.J. (2003) RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Res.*, **31**, 5469–5482.
34. Kiliszek,A., Kierzek,R., Krzyzosiak,W.J. and Rypniewski,W. (2009) Structural insights into CUG repeats containing the 'stretched U-U wobble': implications for myotonic dystrophy. *Nucleic Acids Res.*, **37**, 4149–4156.
35. Kiliszek,A., Blaszczyk,L., Kierzek,R. and Rypniewski,W. (2017) Stabilization of RNA hairpins using non-nucleotide linkers and circularization. *Nucleic Acids Res.*, **45**, e92.
36. Mooers,B.H., Logue,J.S. and Berglund,J.A. (2005) The structural basis of myotonic dystrophy from the crystal structure of CUG repeats. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 16626–16631.
37. Kiliszek,A., Kierzek,R., Krzyzosiak,W.J. and Rypniewski,W. (2010) Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. *Nucleic Acids Res.*, **38**, 8370–8376.
38. Mukherjee,S., Blaszczyk,L., Rypniewski,W., Falschlunger,C., Micura,R., Murata,A., Dohno,C., Nakatani,K. and Kiliszek,A. (2019) Structural insights into synthetic ligands targeting A-A pairs in disease-related CAG RNA repeats. *Nucleic Acids Res.*, **47**, 10906–10913.

39. Yildirim,I., Park,H., Disney,M.D. and Schatz,G.C. (2013) A dynamic structural model of expanded RNA CAG repeats: a refined X-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations. *J .Am. Chem. Soc.*, **135**, 3528–3538.
40. Tawani,A. and Kumar,A. (2015) Structural insights reveal the dynamics of the repeating r(CAG) transcript found in huntington's disease (HD) and spinocerebellar ataxias (SCAs). *PLoS One*, **10**, e0131788.
41. Otwinowski,Z. and Minor,W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.*, **276**, 307–326.
42. Adams,P.D., Grosse-Kunstleve,R.W., Hung,L.-W., Ioerger,T.R., McCoy,A.J., Moriarty,N.W., Read,R.J., Sacchettini,J.C., Sauter,N.K. and Terwilliger,T.C. (2002) PHENIX- building new software for automated crystallographic structure determination. *Acta Crystallogr. D. Biol. Crystallogr.*, **58**, 1948–1954.
43. Emsley,P. and Cowtan,K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D. Biol. Crystallogr.*, **60**, 2126–2132.
44. Lu,X.J. and Olson,W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, **3**, 1213–1227.
45. Li,S., Olson,W.K. and Lu,X.J. (2019) Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res.*, **47**, W26–W34.
46. Zhang,Q., Tsoi,H., Peng,S., Li,P.P., Lau,K.F., Rudnicki,D.D., Ngo,J.C. and Chan,H.Y. (2016) Assessing a peptidylic inhibitor-based therapeutic approach that simultaneously suppresses polyglutamine RNA- and protein-mediated toxicities in patient cells and drosophila. *Dis. Model Mech.*, **9**, 321–334.
47. Zhang,Q., Chan,Z.S., An,Y., Liu,H., Hou,Y., Li,W., Lau,K.F., Koon,A.C., Ngo,J.C.K. and Chan,H.Y.E. (2018) A peptidylic inhibitor for neutralizing expanded CAG RNA-induced nucleolar stress in polyglutamine diseases. *RNA*, **24**, 486–498.
48. Ream,J.A., Lewis,L.K. and Lewis,K.A. (2016) Rapid agarose gel electrophoretic mobility shift assay for quantitating protein: RNA interactions. *Anal. Biochem.*, **511**, 36–41.
49. Ryder,S.P., Recht,M.I. and Williamson,J.R. (2008) Quantitative analysis of protein-RNA interactions by gel mobility shift. *Methods Mol. Biol.*, **488**, 99–115.
50. van Zundert,G.C.P., Rodrigues,J., Trellet,M., Schmitz,C., Kastritis,P.L., Karaca,E., Melquiond,A.S.J., van Dijk,M., de Vries,S.J. and Bonvin,A. (2016) The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.*, **428**, 720–725.
51. Zhang,Q., An,Y., Chen,Z.S., Koon,A.C., Lau,K.F., Ngo,J.C.K. and Chan,H.Y.E. (2019) A peptidylic inhibitor for neutralizing r(GGGGCC)exp-Associated neurodegeneration in C9ALS-FTD. *Mol. Ther. Nucleic Acids*, **16**, 172–185.
52. Tamjar,J., Katorcha,E., Popov,A. and Malinina,L. (2012) Structural dynamics of double-helical RNAs composed of CUG/CUG- and CUG/CGG-repeats. *J. Biomol. Struct. Dyn.*, **30**, 505–523.
53. Kiliszek,A., Kierzek,R., Krzyzosiak,W.J. and Rypniewski,W. (2011) Crystal structures of CGG RNA repeats with implications for fragile X-associated tremor ataxia syndrome. *Nucleic. Acids. Res.*, **39**, 7308–7315.
54. Arnott,S., Hukins,D.W. and Dover,S.D. (1972) Optimised parameters for RNA double-helices. *Biochem. Biophys. Res. Commun.*, **48**, 1392–1399.
55. Baker,N.A., Sept,D., Joseph,S., Holst,M.J. and McCammon,J.A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 10037–10041.
56. Tsoi,H., Lau,C.K., Lau,K.F. and Chan,H.Y. (2011) Perturbation of U2AF65/NXF1-mediated RNA nuclear export enhances RNA toxicity in polyQ diseases. *Hum. Mol. Genet.*, **20**, 3787–3797.
57. Saxena,A., Rorie,C.J., Dimitrova,D., Daniely,Y. and Borowiec,J.A. (2006) Nucleolin inhibits hdm2 by multiple pathways leading to p53 stabilization. *Oncogene*, **25**, 7274–7288.
58. Moore,H., Greenwell,P.W., Liu,C.P., Arnheim,N. and Petes,T.D. (1999) Triplet repeats form secondary structures that escape DNA repair in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 1504–1509.
59. Li,L.B. and Bonini,N.M. (2010) Roles of trinucleotide-repeat RNA in neurological disease and degeneration. *Trends Neurosci.*, **33**, 292–298.
60. Sobczak,K., Michlewski,G., De Mezer,M., Kierzek,E., Krol,J., Olejniczak,M., Kierzek,R. and Krzyzosiak,W.J. (2010) Structural diversity of triplet repeat RNAs. *J. Biol. Chem.*, **285**, 12755–12764.
61. Tian,B., White,R.J., Xia,T., Welle,S., Turner,D.H., Mathews,M.B. and Thornton,C.A. (2000) Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA*, **6**, 79–87.
62. Broda,M., Kierzek,E., Gdaniec,Z., Kulinski,T. and Kierzek,R. (2005) Thermodynamic stability of RNA structures formed by CNG trinucleotide repeats. Implication for prediction of RNA structure. *Biochemistry*, **44**, 10873–10882.
63. Ciesiolka,A., Jazurek,M., Drazkowska,K. and Krzyzosiak,W.J. (2017) Structural characteristics of simple RNA repeats associated with disease and their deleterious protein interactions. *Front Cell Neurosci.*, **11**, 97.
64. De Vries,S.J., Van Dijk,M. and Bonvin,A.M. (2010) The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.*, **5**, 883–897.
65. Meng,X.Y., Zhang,H.X., Mezei,M. and Cui,M. (2011) Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided Drug Des.*, **7**, 146–157.
66. Pagadala,N.S., Syed,K. and Tuszynski,J. (2017) Software for molecular docking: a review. *Biophys. Rev.*, **9**, 91–102.
67. Masliah,G., Barraud,P. and Allain,F.H. (2013) RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cell. Mol. Life Sci.*, **70**, 1875–1895.
68. Gonzalez,V. and Hurley,L.H. (2010) The C-terminus of nucleolin promotes the formation of the c-MYC G-quadruplex and inhibits c-MYC promoter activity. *Biochemistry*, **49**, 9706–9714.
69. Ishikawa,F., Matunis,M.J., Dreyfuss,G. and Cech,T.R. (1993) Nuclear proteins that bind the pre-mRNA 3′ splice site sequence r(UUAG:G) and the human telomeric DNA sequence d(TTAGGG)n. *Mol. Cell Biol.*, **13**, 4301–4310.
70. Zhang,Y., Bhatia,D., Xia,H., Castranova,V., Shi,X. and Chen,F. (2006) Nucleolin links to arsenic-induced stabilization of GADD45alpha mRNA. *Nucleic Acids Res.*, **34**, 485–495.