Boeynaems et al. (2023): Aberrant phase separation is a common killing strategy of positively charged peptides.

SUPPLEMENTAL INFORMATION

Material & Methods

Supplemental text

Figs. S1 to S12

OTHER SUPPLEMENTAL MATERIAL

Table S1 (Excel): Mass spectrometry data.

Table S2 (Excel): Machine learning prediction scores.

Table S3 (Excel): Overview of abundant cationic proteins in model organisms.

MATERIAL & METHODS

EXPERIMENTAL MODELS

Human cell lines

U2OS (ATCC, HTB-96) cells were grown at 37°C in a humidified atmosphere with 5% CO_2 for 24 h in Dulbecco's Modified Eagle's Medium (DMEM), high glucose, GlutaMAX + 10 % Fetal Bovine Serum (FBS) and pen/strep (Thermo Fisher Scientific).

Bacteria

E. coli (Stbl3, Thermo Fisher) were grown at 37°C in a shaking incubator in LB medium and used for ex vivo and soft x-ray tomography assays. *E. coli* (NiCo21(DE3), New England Biolabs) transformed with plasmids for expression of GFP charge variants were grown at 37 °C in a shaking incubator in LB medium supplemented with 100 µg/mL ampicillin.

Mice and primary mouse cells

All mouse husbandry and procedures were performed in accordance with institutional guidelines and approved by the Stanford Administrative Panel on Animal Care (APLAC). C57BL/6 mice (Jackson Laboratory) were used. Mouse primary cortical neurons were grown in Neurobasal media (Gibco) supplemented with B-27 serum-free supplement (Gibco), GlutaMAX, and Penicillin-Streptomycin (Gibco) in a humidified incubator at 37°C, with 5% CO_2.

METHOD DETAILS

Plasmid construction

All constructs for human expression were generated through custom synthesis and subcloned into a pcDNA3.1 backbone by Genscript (Piscataway, USA). All constructs for GFP expression in *E. coli* were generated through custom synthesis and subcloned into an expression backbone (pETDuet-1) by Genscript (Piscataway, USA), as has been previously reported[110].

Peptides

All peptides were generated via chemical synthesis by Pepscan (Lelystad, the Netherlands), except BMAP-27 (AS-65598, Anaspec), LL-37 (AS-61302, Anaspec), 5-FAM-LC-LL-37 (AS-63694, Anaspec), Crotamine (CRO01-01000, Smartox-Biotech), and Cy3-Crotamine (CRO02-00500, Smartox-Biotech). Peptides were fluorescently labeled as described previously [111].

Recombinant proteins

Recombinant GFP variants

GFP variants were expressed in *E. coli* (NiCo21(DE3), NEB) in LB media supplemented with 100 µg/mL ampicillin. Overnight cultures were inoculated into 1 L of LB media supplemented with ampicillin and allowed to grow at 37 °C to an OD_{600} between 0.8-1.0. At this point, the cultures were induced with 1 mM IPTG. After induction, cultures were grown at 25 °C with shaking for an additional 16-18 h. At this point, cells were collected by centrifugation in a swinging bucket rotor at 4000 rpm for 15 min. The cell pellet was resuspended in a lysis buffer (50 mM NaH_2PO_4, pH 8.0 supplemented with 300 mM (GFP(-24)) or 1 M

NaCl (GFP(+24))). The cell pellet was then lysed by sonication (2s on, 4 s off) at 60% amplitude using a 0.5 in probe for 10 min. The insoluble fraction was separated via centrifugation in a fixed angle rotor at 10,000 rpm for 30 min at 4 °C. The GFP was purified from the soluble fraction using immobilized metal affinity chromatography (His-Pur Ni-NTA). The column was washed with lysis buffer containing 50 mM imidazole and GFP was eluted with lysis buffer supplemented with 250 mM imidazole. Fractions were collected and analyzed by SDS-PAGE for purity. Pure fractions were combined and dialyzed against 10 mM tris buffer, pH 7.4 before adjusting the concentration to 1 mg/mL GFP by ultrafiltration using a 10 kDa molecular weight cutoff spin filter (Amicon).

In vitro chromatin reconstitution

Chromatin was reconstituted by salt dialysis as described previously[112]. Histones for chromatin assembly were purified as described previously[112–114]. pUC19 plasmid DNA containing 19x "601" Widom positioning sequence[115] (ASP 696) was digested with EcoRI, XbaI, DraI and HaeII (New England Biolabs). 19x 601 array was then purified using polyethylene glycol (PEG) precipitation followed by dialysis into TE buffer (10 mM Tris pH 8.0, 0.5 mM EDTA). EcoRI and XbaI 5' overhangs were filled in with dGTP (New England Biolabs), dCTP (New England Biolabs), dTTP (New England Biolabs), and biotin-14-dATP (Thermo Fisher Scientific) using Large Klenow fragment 3'-5' exo- (New England Biolabs). Biotinylated 19x 601 array DNA, H2A/H2B histone dimer, and H3/H4 tetramer were added to high salt buffer (10 mM Tris-HCl, pH 7.5; 0.25 mM EDTA; 2 M NaCl) and gradually dialyzed over the course of ~67 hours at a rate of 0.5 mL/min from 500 mL high salt buffer into low salt buffer (10 mM Tris-HCl, pH 7.5; 0.25 mM EDTA; 2.5 mM NaCl). H3/H4 tetramer concentrations were titrated to vary chromatin saturation. Nucleosome assembly was verified using overnight digestion of chromatin at room temperature using AvaI which cuts between adjacent 601 positions in the array, followed by a native acrylamide gel shift analysis (gel shift of ~200 bp AvaI digestion product to ~600 bp suggests nucleosome positioned on 601 sequences as described previously[112]).

Recombinant NPM1

Recombinant NPM was a kind gift from Dr. Richard Kriwacki and Dr. Aaron Phillips (St. Jude Children's Research Hospital).

Human cell culture, treatments and microscopy

U2OS cells were grown at 37°C in a humidified atmosphere with 5 % CO2 for 24 h in Dulbecco's Modified Eagle's Medium (DMEM), high glucose, GlutaMAX + 10 % Fetal Bovine Serum (FBS) and pen/strep (Thermo Fisher Scientific). Cells were transiently transfected using Lipofectamine 3000 (Thermo Fisher Scientific) according to manufacturer's instructions. Cells grown on cover slips were fixed for 24 h after transfection in 4 % formaldehyde in PBS. For peptide uptake experiments, cells were treated with 1 µM of fluorescently labeled peptides and incubated for 45 min at 37°C, followed by a PBS wash and fixation with 4 % formaldehyde in PBS. Slides were mounted using ProLong Gold antifade reagent (Life Technologies). Confocal images were obtained using a Zeiss LSM 780 Meta NLO confocal microscope. Images were processed and analyzed using Fiji[116] and Adobe Photoshop.

Mouse primary neuron culture and microscopy

Primary mouse cortical neurons were dissociated into single cell suspensions from E16.5 C57BL/6 mice (Jackson Laboratory) cortices using a papain dissociation system (Worthington Biochemical Corporation). Neurons were seeded onto poly-L-lysine coated plates (0.1% w/v) and grown in Neurobasal media (Gibco)

supplemented with B-27 serum-free supplement (Gibco), GlutaMAX, and Penicillin-Streptomycin (Gibco) in a humidified incubator at 37°C, with 5% CO2. Cells were transiently transfected using Lipofectamine 3000 (Thermo Fisher Scientific) according to manufacturer's instructions. Cells grown on cover slips were fixed for 24 h after transfection in 4 % formaldehyde in PBS. For peptide uptake experiments, cells were treated with 1 µM of fluorescently labeled peptides and incubated for 1 h at 37°C, followed by a PBS wash and fixation with 4 % formaldehyde in PBS. Slides were mounted using ProLong Gold antifade reagent (Life Technologies). Confocal images were obtained using a Zeiss LSM 780 Meta NLO confocal microscope. Images were processed and analyzed using Fiji[116] and Adobe Photoshop.

Bacterial culture and microscopy

E. coli (NiCo21(DE3), New England Biolabs) transformed with plasmids for expression of GFP charge variants were grown at 37 °C in a shaking incubator in LB medium supplemented with 100 µg/mL ampicillin. For each variant, overnight cultures were prepared by inoculating a single colony from a plate into a snapcap tube containing 5 mL of LB media supplemented with ampicillin. After growth overnight, the cultures were then diluted to OD_{600} = 0.1 in LB media supplemented with ampicillin and were allowed to grow at 37 °C with shaking until they reached an OD_{600} of approximately 1.0. The OD_{600} of each culture was then normalized to 0.9 and 200 µL of the culture was then transferred to a 96-well plate for 10-fold serial dilutions with LB media supplemented with ampicillin. 5 µL of each culture and dilution (10^{-7}x diluted) were spotted onto LB plates supplemented with 100 µg/mL ampicillin with or without 1 mM isopropyl β-D-1-thiogalacopyranoside (IPTG, Gold Biotechnology). After drying, plates were incubated at 25 °C for 30 h before imaging on a scanner and transilluminator (Bio-Rad Gel Doc XR+).

For imaging GFP co-localization with ribosomes in *E. coli*, an rplI-mScarlet fusion was integrated into the native genome locus of rplI in the NiCo21(DE3) *E. coli* strain. This process began with QC101 (MG1655 *rplI-mCherryKan*R) cells, which contain a genomically integrated ribosome-mCherry fusion protein and were previously used to investigate the localization of ribosomes in *E. coli* during various phases of cell growth[117]. We received these cells as a gift from the Sanyal lab and amplified the rplI-mCherry_KanR cassette via colony PCR using primers, CGAACACGAAGTGAGCTTCC and AAGCAAAACGCCGACCAA, and Phusion polymerase as recommended by New England Biolabs (NEB). mCherry was replaced with mScarletI for better compatibility with our imaging system. HiFi assembly was used to assemble a vector with a rplI-mScarletI_KanR cassette.

Construction of p2-39 rplI-mScarletI_KanR plasmid

Fragment	Forward Primer	Reverse Primer	DNA template
p2-39 backbone	CATTGGTCGGCGTTTTGCTTACTTAATTAACGGCACTCC	GGAAGCTCACTTCGTGTTCGAGAACCCCGCATATGTATATC	p2-39 mScarletI [82]
rplI-upstream homology	CGAACACGAAGTGAGCTTCC	ATCACAGCTTCTCCTTTACTTTCAGCTACTACGTTTACG	rplI-mCherry_Kan^R cassette
mScarletI	AGTAAAGGAGAAGCTGTGATTAAAGAGTTC	TTTGTATAGTTCATCCATGCCACCG	p2-39 mScarletI [82]

Kan-downstream homology	GCATGGATGAACTATAC AAATAAGAATTCAAACA GTAATACAAG	AAGCAAAACGCCGACCAA	rplI-mCherry_Kan^R cassette

All fragments were PCR purified (NEB), ligated using NEB HiFi assembly kit, transformed into NEB5α cells according to the manufacturer's instructions, and plated onto LB agar plates supplemented with 25 µg/mL chloramphenicol and 50 µg/mL kanamycin. Single colonies were inoculated into LB media supplemented with chloramphenicol and kanamycin and grown at 37 °C overnight with shaking. Plasmid DNA was isolated from overnight cultures using the Qiagen Spin Miniprep Kit and plasmid sequences were verified by Genewiz.

The dsDNA cassette for genomic integration was amplified from the p2-39 rplI-mScarletI_KanR plasmid using primers, CGAACACGAAGTGAGCTTCC and AAGCAAAACGCCGACCAA. The PCR-amplified product was treated with DpnI, PCR purified, and verified by gel electrophoresis.

To prepare cells for genomic integration, electrocompetent NiCo21(DE3) cells (NEB) were prepared and transformed with 50 ng pTKRED. pTKRED was a gift from Edward Cox and Thomas Kuhlman (Addgene plasmid # 41062)[118]. Briefly, overnight cultures of NiCo21(DE3) cells were diluted 100-fold in LB media and grown at 37 °C, 225 rpm until OD_{600} = 0.35-0.4. Cells were then incubated on ice for 10-20 min and centrifuged at 400 rcf for 10 min at 4 °C. Cells were washed twice in ice cold sterile milliQ water and then washed once in ice cold 10% glycerol before resuspending in ice cold 10% glycerol, aliquoting, and snap freezing with liquid nitrogen.

Electrocompetent NiCo21(DE3) cells were transformed with pTKRED using a MicroPulser Electroporator (Bio-Rad). 1 µL plasmid was added to 100 µL electrocompetent cells in a chilled Gene Pulser/MicroPulser electroporation cuvette with 0.1 cm gap (Bio-Rad) and incubated on ice for 1 min. Cells were then subjected to a single pulse and were then recovered in SOC media at 30 °C with shaking for 3 h. A portion of the recovered cells was then plated on LB plates supplemented with 50 µg/mL spectinomycin and incubated at 30 °C overnight. Single colonies were inoculated into LB media supplemented with 100 µg/mL spectinomycin and the presence of the plasmid was confirmed by sequencing (Genewiz).

Electrocompetent pTKRED NiCo21(DE3) cells were prepared. Cells were grown in SOB media supplemented with 0.5% glucose, 100 µg/mL spectinomycin, and 2 mM IPTG at 30 °C. Cells were then harvested as described above when OD_{600} = 0.5 – 0.6. 1 µg of the dsDNA rplI-mScarletI-KanR cassette was added to 100 µL electrocompetent cells, and electroporation was performed as described above. Cells were recovered in 1 mL SOC media for 3 h at 30 °C with shaking and 250 µL of electroporated cells were plated onto LB agar plates supplemented with 100 µg/mL spectinomycin and 25 µg/mL kanamycin[118]. Plates were incubated at 30 °C for 1-2 days. Single colonies were selected and grown in LB media supplemented with 25 µg/mL kanamycin and incubated at 42 °C with shaking for approximately 20 h before they were diluted and spread onto LB plates supplemented with 25 µg/mL kanamycin. Plates were incubated overnight at 37 °C. We verified that these colonies did not grow on LB plates with spectinomycin at 30 °C, indicating that the pTKRED plasmid had been cured. Additionally, we performed colony PCR with primers upstream (CGCCATCAGTAATCGGTCA) and downstream (CGCGAAGTTCTTCCACGAT) of the cassette, confirming successfully integration into the *E. coli* genome.

Chemically competent rplI-mScarletI_KanR NiCo21(DE3) cells were prepared and transformed with a plasmid encoding each GFP variant. Briefly, overnight cultures were diluted 100-fold and grown to OD_{600} = 0.3 -0.4 at 37 °C. At this point, cells were harvested by centrifugation at 2500 rcf for 10 min at 4 °C in a cold fixed angle rotor. The pellet was washed once in ice-cold transformation buffer (10 mM PIPES, 15 mM $CaCl_2$, 250 mM KCl, 55 mM $MnCl_2$, pH 6.7) and DMSO was added dropwise to a final concentration of 7%. Competent cells were aliquoted and snap frozen in liquid nitrogen. 2 µL of GFP plasmid (GFP(-18) or GFP(+18))[82] was added to 100 µL chemically competent cells. Cells were incubated on ice for 30 min, heat shocked at 42 °C for 30 s, and incubated on ice for 5 min. Cells were recovered in 200 µL SOC for 1 h at 37 °C and 100 µL of this mixture was plated onto LB plates supplemented with 50 µg/mL kanamycin and 100 µg/mL ampicillin. Cells were grown overnight at 37 °C.

For each variant, cultures were grown as above until they reached an OD_{600} between 0.8-1.0. At this point GFP expression was induced by the addition of 1 mM IPTG and the cultures were grown at 25 °C with shaking. Protein expression was induced for 24 h and then cells were imaged on agarose pads using a 100X oil 1.40 NA UPlanSApo objective (Olympus) on GFP (λ_{ex} = 470-522 nm; λ_{em} = 525-550 nm; EVOS GFP light cube) and Texas Red (λ_{ex} = 585-629 nm; λ_{em} = 628-632 nm; EVOS Texas Red light cube) channels using an EVOS FL Auto 2 inverted fluorescence microscope. Raw microscopy images were background subtracted (rolling ball algorithm with 100 pixel radius) in ImageJ.

Immunofluorescence

U2OS cells and primary neurons were fixed 24h after transfection (or after 45 min – 1 h of peptide treatment) in 4% formaldehyde in PBS and stained according standard protocols. Following primary antibodies were used: G3BP1 (ab56574, Abcam), H3K9me3 (ab8898, Abcam), SON (NBP1-88706, Novus Biologicals), FIBL (ab4566, Abcam), NPM1 (ab10530, Abcam), PABPC (ab21060, Abcam), FLAG (F7425 and F1804, Sigma-Aldrich). Alexa-labeled secondary antibodies were used (Thermo Fisher Scientific).

Neurotoxicity assays

Five days after seeding the neurons, recombinant GFP mutants were added at a concentration of 3 µM for 24h. Cytotoxicity in primary neuron cultures was measured with an alamarBlue cell viability (Thermo Fisher Scientific), according to manufacturer's instructions. Readout was measured using a SPARK Multimode microplate reader (Tecan Life Sciences). Data was analyzed using Microsoft Excel and GraphPad Prism.

Peptide coacervation experiments

Peptide experiments were performed as described previously with slight modifications [111]. In brief, peptides were generated via chemical synthesis by Pepscan (Lelystad, the Netherlands). Peptides were dissolved in milli-Q water and stored at -20°C. To test for coacervation, peptides were diluted in human or *E. coli* cell lysate at the indicated concentrations in PBS at pH 7.4. NPM1 phase separation assays were done as described in [56]. Nucleosome arrays, plasmid DNA or E. coli ribosomes (New England Biolabs) were added at a concentration of 200 ng/µl and 1 µg/µl, respectively. Samples were transferred to an imaging chamber (Grace Bio-Labs, Bend, USA) and imaged on a Zeiss LSM 710 confocal microscope. Pictures of turbid solutions in PCR strips were taken using a Google Pixel 3a.

PR protein precipitation

U2OS cells were trypsinized (Thermo Fisher Scientific) and harvested by centrifugation. Cell pellets were washed three times with ice-cold PBS. Pellets were redissolved in PBS buffer with Halt Protease inhibitor (Thermo Fisher Scientific) and lysed using a probe sonicator (Branson Sonifier 250, VWR Scientific) on ice. The lysate was cleared from the insoluble fraction by centrifugation for 15 min at 10,000 rpm at 4°C. The supernatant was retrieved and the procedure was repeated until no pellet was visible after centrifugation. The protein concentration was measured using Micro BCA assay (Thermo Fisher Scientific). PR30 (Pepscan, The Netherlands) was added to a final concentration of 50 mM to 0.5 mg of soluble lysate in a total volume of 400 µl and incubated at room temperature for 15 min. The volume of the samples was increased to 1 ml with PBS, before gently spinning down the PR droplets at 4,000 rpm for 5 min. The supernatans was transferred to a new tube. Pellets were subsequently washed with 1 ml PBS and vortexed before spinning down again. Washing steps were repeated three times. To each pellet a volume of buffer corresponding to the supernatans was added. The resulting samples were processed for LC-MS/MS. Samples for analysis by silver staining were generated identically. Silver staining (Thermo Fisher Scientific) was performed according to the manufacturer's instructions.

Proteomics sample preparation and LC-MS/MS analysis

Proteins in the supernatant and pellet samples were first reduced by addition of DTT to a concentration of 5 mM and incubation for 30 minutes at 55˚C and then alkylated by addition of iodoacetamide to a concentration of 10 mM for 15 minutes at room temperature in the dark. Samples were diluted with 50 mM Tris pH 7.9 to a urea concentration of 4 M and proteins were digested with 5 µg lysyl endopeptidase (Wako) (1/100, w/w) for 4 hours at 37°C. Samples were further diluted with 50 mM Tris pH 7.9 pH 8.0 to a final urea concentration of 2 M and proteins were digested with 5 µg trypsin (Promega) (1/100, w/w) overnight at 37˚C. Next, peptides were purified on SampliQ SPE C18 cartridges (Agilent). Columns were first washed with 1 ml 100% acetonitrile (ACN) and equilibrated with 3 ml of solvent A (0.1% TFA in water/ACN (98:2, v/v)) before samples were loaded on the column. After peptide binding, the column was washed again with 2 ml of solvent A and peptides were eluted twice with 750 µl elution buffer (0.1% TFA in water/ACN (40:60, v/v)). Purified peptides were dried under vacuum in HPLC inserts and stored at −20 °C until LC-MS/MS analysis.

Peptides were re-dissolved in 20 µl loading solvent A (0.1% TFA in water/acetonitrile (ACN) (98:2, v/v)) of which 2.5 µl (supernatant) or 5 µl (pellet) was injected for LC-MS/MS analysis on an Ultimate 3000 RSLC nanoLC in-line connected to a Q Exactive HF Biopharma mass spectrometer (Thermo Fisher Scientific) equipped with a pneu-Nimbus dual ion source (Phoenix S&T). Trapping was performed at 10 µl/min for 4 min in loading solvent A on a 20 mm trapping column (made in-house, 100 µm internal diameter (I.D.), 5 µm beads, C18 Reprosil-HD, Dr. Maisch) and the sample was loaded on a reverse-phase column (made in-house, 75 µm I.D. x 400 mm length, 3 µm beads C18 Reprosil-HD, Dr. Maisch). Peptides were loaded with loading solvent A and were separated with a non-linear 145 min gradient from 2% to 56% MS solvent B (0.1% FA in water/ACN 20:80 (v/v)) at a flow rate of 250 nl/min followed by a 15 min wash reaching 99% MS solvent B and re-equilibration with 98% MS solvent A (0.1% FA in water).

The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS acquisition for the 16 most abundant ion peaks per MS spectrum. Full-scan MS spectra (375-1,500 m/z) were acquired at a resolution of 60,000 in the Orbitrap analyzer after accumulation to a target value of 3,000,000. The 16 most intense ions above a threshold value of 13,000 (minimum AGC of 1,000) were isolated for fragmentation at a normalized collision energy of 28%. The C-trap was filled at a target value of 100,000 for maximum 80 ms and the MS/MS spectra (200-2,000 m/z) were acquired at a

resolution of 15,000 in the Orbitrap analyzer with a fixed first mass of 145 m/z. Only peptides with charge states ranging from +2 to +6 were included for fragmentation and the dynamic exclusion was set to 12 s. QCloud was used to control instrument longitudinal performance during the project[119].

Protein identification and quantification

Data analysis was performed with MaxQuant (version 1.5.8.3) using the Andromeda search engine with default search settings including a false discovery rate set at 1% on PSM and protein level. The spectra of all LC-MS/MS runs were interrogated against the human proteins in the Swiss-Prot Proteome database (database release version of January 2018 containing 20,243 human protein sequences, (http://www.uniprot.org)). The mass tolerance for precursor and fragment ions was set to 4.5 and 20 ppm, respectively, during the main search. Enzyme specificity was set as C-terminal to arginine and lysine, also allowing cleavage at proline bonds with a maximum of two missed cleavages. Variable modifications were set to oxidation of methionine residues and acetylation of protein N-termini. Matching between runs was enabled with a matching time window of 0.7 minutes and an alignment time window of 20 minutes. Only proteins with at least one unique or razor peptide were retained leading to the identification of 1599 proteins. Proteins were quantified by the MaxLFQ algorithm integrated in the MaxQuant software. A minimum ratio count of two unique or razor peptides was required for quantification. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD040619.

Soft X-ray tomography

PR30 (Pepscan, The Netherlands) was added to a final concentration of 50 mM to 0.5 mg of soluble lysate in a total volume of 400 µl and incubated at room temperature for 15 min. Samples were subsequently loaded in thin-wall glass capillaries and rapidly frozen into liquid nitrogen cooled liquid propane. The specimens were imaged by XM-2, a soft x-ray microscope in the National Center for X-ray tomography (http://ncxt.org) located at the Advanced Light Source of Lawrence Berkeley National Laboratory (ALS, LBNL). The x-ray microscope operates at the “water-window” region of soft x-rays, providing natural contrast of biomolecules with respect to water. The condenser and objective lenses, chosen for this experiment, guaranteed isotropic 32nm voxel size[120,121]. For 3D reconstructions, 92 projection images, with 200-300 ms exposure time each, were acquired sequentially around a rotation axis with 2° increment angles. After normalization and alignment, tomographic reconstructions were calculated using iterative reconstruction methods[122]. The segmentation and visualization of multi-droplets was done based on linear attenuation coefficient of x-rays in Amira 6.3.0.

Clustering and machine learning

To group proteins according to their relative solubility profiles, we used both the three experimental solubility values and calculated the change in solubility between 0 and 0.4, and 0.4 and 2. These delta values were then clustered at several different k values using the kmeans algorithm. Aggregate solubilities were plotted and visualized, and k=9 was selected based on a high level of intra-cluster concordance and specificity of GO enrichment.

UniRep had been demonstrated to be an effective model to extract fundamental features of amino-acid sequences that are structurally, evolutionarily and biophysically grounded[64]. We obtained the UniRep average hidden state representations of the entire human proteome (UP000005640_9606) using the 256-unit model. We then set aside the proteins in the mass spec experiment to train a support vector

machine model with a linear kernel (sklean.svn.SVC) that was able to distinguish blue from grey+red proteins. The mass spec proteins were randomly split into 80-20 train-validation sets. We then classified the rest of the proteome using the trained model, and the confidence scores were used to rank the proteins in each category.

Since we performed the mass spectrometry experiments on soluble cell lysate without the addition of detergents—to avoid any effects of the detergent on condensation—our input sample was strongly depleted from membrane proteins. Nonetheless, as can be seen in the GO analysis of cluster 9, we got a small membrane contaminant fraction specifically in the blue class of proteins. This makes sense as these membrane proteins were insoluble in all conditions of the experiment due to the lack of detergent, leading the algorithm to rank membrane proteins incorrectly with high blue scores. Therefore, we excluded all membrane bound proteins from our prediction analyses as our experimental setup was not designed to study the phase behavior of this class of proteins.

Charge clustering metric

The charge clustering metric implemented here was developed and adapted from an inverse weighted distance metric, which was originally developed in the context of clustering of aromatic residues[63]. This metric quantifies the clustering of target amino acids relative to other target residues, and is complimentary, yet distinct, from other current charge distribution metrics[59]. While other metrics quantify the relative patterning of charged residues, our focus here is not on the patterning of residues relative to charged residues being evenly distributed, but instead on identifying folded domains or disordered regions where positively or negatively charged residues are clustered together, regardless of the presence of well-mixed regions elsewhere in the same protein. The IWD metric also enables the same functional form to be used for folded domains and disordered regions. To adapt the clustering in the context of charge residues, a net-charge per residue (NCPR) is calculated for all charged residues using a sliding window of five residues. Next, positively charged or negatively charged residues are selected as so-called *target residues*. Then the NCPR of each target residue is weighted by the sum of the inverse distances between all pairs of other target residues for each residue (Fig. S3). To calculate a single value for the sequence an average across the target residues is computed. In the positive charge clustering, the target residues are the “true” positively charged (arginine and lysine) residues, where “true positive” here means the net charge of the residue is positive, as determined by the sliding-window NCPR. Conversely, in the negative charge clustering, the target residues are the “true” negatively charged residues (aspartic acid and glutamic acid) residues, where “true negative” again means the net charge of the residue is negative as determined by the sliding window NCPR.

To calculate the bivariate-charge clustering, a metric which reports on how well interspersed the positive and negative residues are relative to each other, the target residues are both true positive and true negative residues (again where ‘true’ here reflects the use of the sliding window NCPR). Clustering is then calculated only between pairs of opposite charge residues. The sum of the inverse distance between all pairs of oppositely charged target residues is then weighted by the absolute value of the difference in NCPR between target residues in the pair. In the IDRs, inverse distance is calculated in linear sequence space, while for the surface of folded domains inverse distance is calculated in cartesian space[123].

Enrichment of protein sequence feature calculation.

The protein sequence analysis was conducted using SHEPHARD, a Python-based framework designed for integrating and analyzing large-scale amino acid sequence properties[124]. IDRs were predicted and annotated using metapredict (V2)[125]. Protein structures for folded domains were taken from AlphaFold2 structure predictions[126]. The folded domain structures were analyzed to calculate the per-residue solvent-accessible surface area using SOURSOP and the 3D position of solvent-accessible residues were used to calculate 3D charge clustering (see above)[127]. Sparrow (https://github.com/idptools/sparrow) was used to calculate the sequence features for both IDRs and folded domains. While our analysis focused on charged residue clustering, we performed a full analysis of different sequence features to perform an unbiased screen to identify features that delineated grey, red, and blue class proteins. To determine which sequence features were most discriminatory in separating the three groups, an enrichment test was conducted comparing a given feature's values to the whole dataset. To evaluate how well correlated a specific sequence feature is with a given class, a linear-discriminant analysis (LDA) was conducted. The weight coefficients from the LDA were then used to inform how well correlated a specific sequence feature was with a given group. This was conducted for both sequence features on the surface of the folded domains and in the IDRs.

Bioinformatics analysis of protein abundance

Protein abundance analyses (Fig. S7–11) were performed using SHEPHARD[124] and sparrow (https://github.com/idptools/sparrow). Mass spectrometry data were obtained for humans[128], *X. laevis*[129], *A. thaliana*[130], *E. coli*[131], *S. pombe*[132], and *S. cerevisiae*[133]. Phosphorylation data for the human proteome was obtained from ProteomeScout[134]. All data and code for this analysis are provided at https://github.com/holehouse-lab/supportingdata/tree/master/2023/boeynaems_2023.

Peptide charge calculation

Peptide charge as a function of pH was calculated using Protein Calculator v3.4 (http://protcalc.sourceforge.net/).

GO enrichment analysis

Go enrichment analysis was performed using DAVID[135]. Enriched terms are depicted in word clouds with font size correlated to -log(p-value). P-values were Bonferroni corrected.

QUANTIFICATION AND STATISTICAL ANALYSIS

All data was analyzed using Graphpad Prism 8.4.1 and Excel. Statistical tests, p values, number of samples, replicates, and experiments are indicated in the figure legends.

SUPPLEMENTAL TEXT

Proteome-wide net charge analysis

We and others have previously indicated that proteomes tend to be—on average—more negatively charged[89,90,136]. We confirmed this conclusion in six different model organisms (**Fig. S7**). Since our work thus far suggests that polycationic folded and disordered proteins drive cellular toxicity across the kingdoms of life, one might expect a strong selection against positively charged proteins at the proteome level. To our surprise, bioinformatic analysis of these six different model organisms found little obvious evidence of this signature, with approximately equal numbers of positively and negatively charged proteins, despite the modest trend towards negatively charged proteins (**Fig. S7**). Yet, an implicit assumption in these analyses is that all proteins are equally abundant, which is obviously false. Using previously published quantitative mass-spectrometry data, we used the per-protein copy numbers to re-assess our analysis. Focusing on the set of proteins that comprise the top 85% of all proteins in a cell by abundance, we re-analyzed our proteomes using the overall biomolecular mass (i.e., copy number times protein molecular weight) (**Fig. S8**). We took this mass-weighted approach because we reasoned that poisons are intrinsically dose-dependent, such that we should expect an effect to depend on the overall mass of polycations, not simply the number of molecules. Our analysis identified 1063 highly abundant proteins, of which 154 (~15%) had a net positive charge above +0.05. Of these polycationic proteins, 75% were histones or ribosomal proteins, 8% were membrane proteins, 6% were constitutive components of ribonucleoprotein bodies (e.g., the nucleolus, the spliceosome, or the signal recognition particle), and the remaining 11% were nuclear RNA- or zinc-binding proteins (**Fig. S9**, **Table S3**). A key feature of these proteins is that none exist as soluble cytosolic proteins; all but 24 are constitutively sequestered in assemblies with either nucleic acids (ribosomal proteins or histones) or phospholipids (membrane proteins). The remaining 24 (zinc-binding proteins, RNA-binding proteins, and nuclear body proteins) also undergo extensive phosphorylation. On average, these 24 proteins have one phosphosite for every ten residues, five times more than expected by random chance (**Fig. S10**). In short, we find for essentially every single example of an abundant, positively charged protein, that its charge is neutralized by a phosphate moiety—either in *trans* (nucleic acids or phospholipid headgroups) or in *cis* (through phosphorylation).

To examine the generality of this finding, we extended our analysis across five other model organisms (*A. thaliana*, *X. laevis*, *S. pombe*, *S. cerevisiae*, and *E. coli*). In all five cases, the same trends held true, with almost no examples of abundant free cytosolic positively charged proteins identified (**Fig. S11**, **Table S3**). Taken together, our integrative bioinformatic analysis suggests that while proteomes do possess many examples of proteins with a net positive charge, these are either expressed at low copy numbers, sequestered from the cellular environment via constitutive binding to anionic biomolecules, or neutralized through phosphorylation.

SUPPLEMENTAL FIGURES

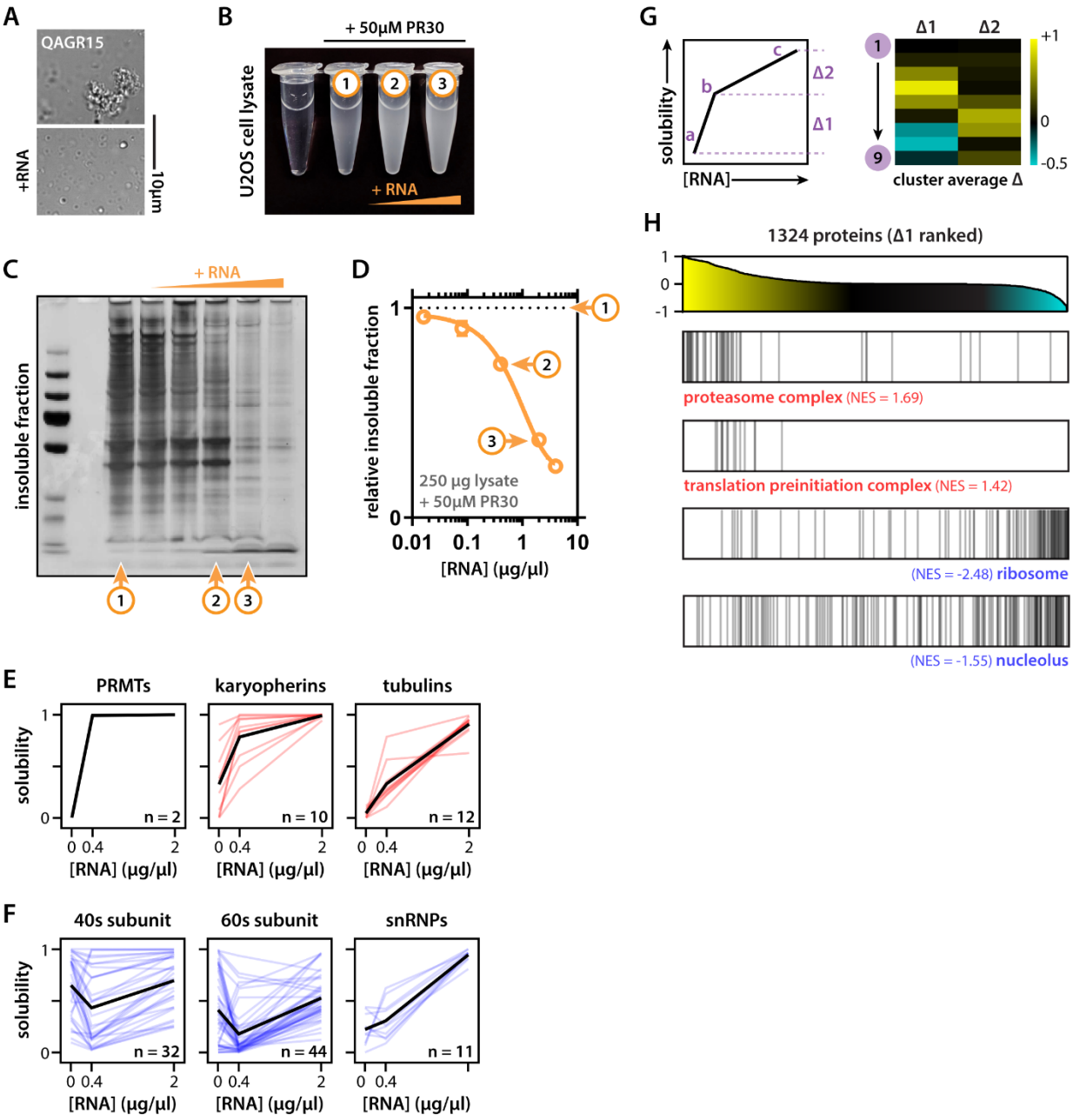

Figure S1: Disease-related basic repeat peptides target biomolecular condensates. (A) QAGR + lysate condensates are modulated by RNA addition (2 µg/µl). (B) PR + lysate mixtures with (1) no, (2) 0.4 and (3) 2µg/µl of RNA added. (C) Silver stain of condensate fraction highlighting position of the three mixtures. Same picture as in **Fig. 1**. (D) Quantification of total condensate fraction. (E-F) Example solubility profiles of red and blue protein families. Colored lines indicate individual proteins, black lines indicate family average. (G-H) Solubility profiles can be broken down in solubility steps. Δ1 indicates the first solubility step (i.e., response to the addition of 0.4 µg/µl RNA) and has opposite values for blue versus red clusters

(convex vs concave profiles). Different protein classes enrich for high or low Δ1 values in our data set. GSEA analysis[137], NES = normalized enrichment score.

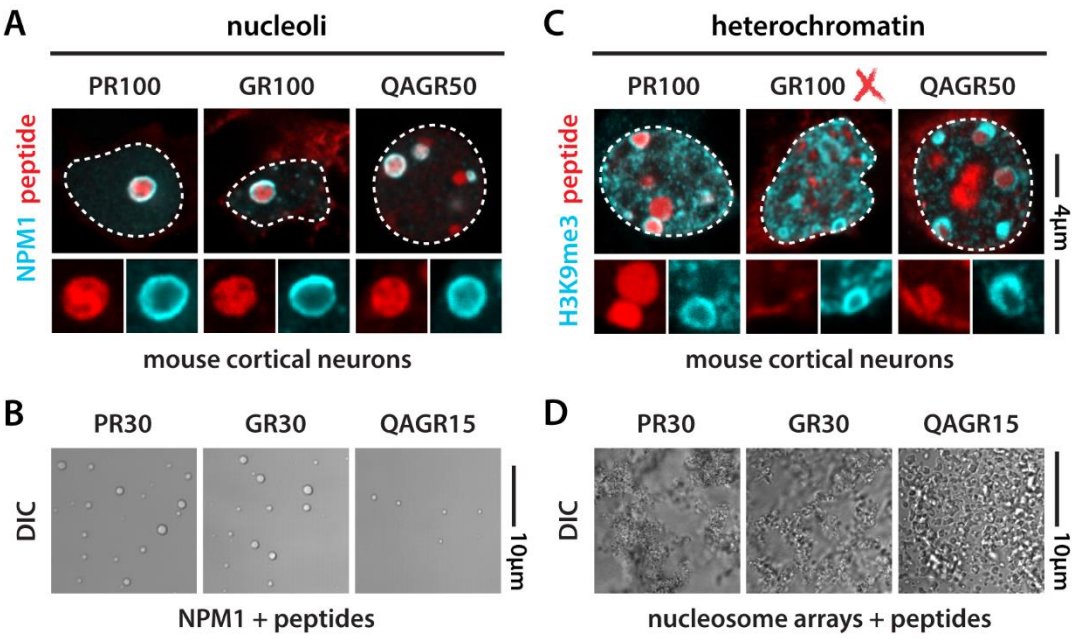

Figure S2: Disease-related basic repeat peptides target biomolecular condensates. (A) mCherry-tagged PR100, GR100 and QAGR50 target nucleoli in mouse primary cortical neurons. (B) PR30, GR30 and QAGR15 peptides phase separate with NPM1 *in vitro*. (C) mCherry-tagged PR100 and QAGR50, but not GR100, target heterochromatin in primary cortical neurons. (D) PR30, GR30 and QAGR15 peptides condense nucleosome arrays *in vitro*. This shows that despite GR can engage with chromatin in the test tube, in cells it seems to prefer other compartments.

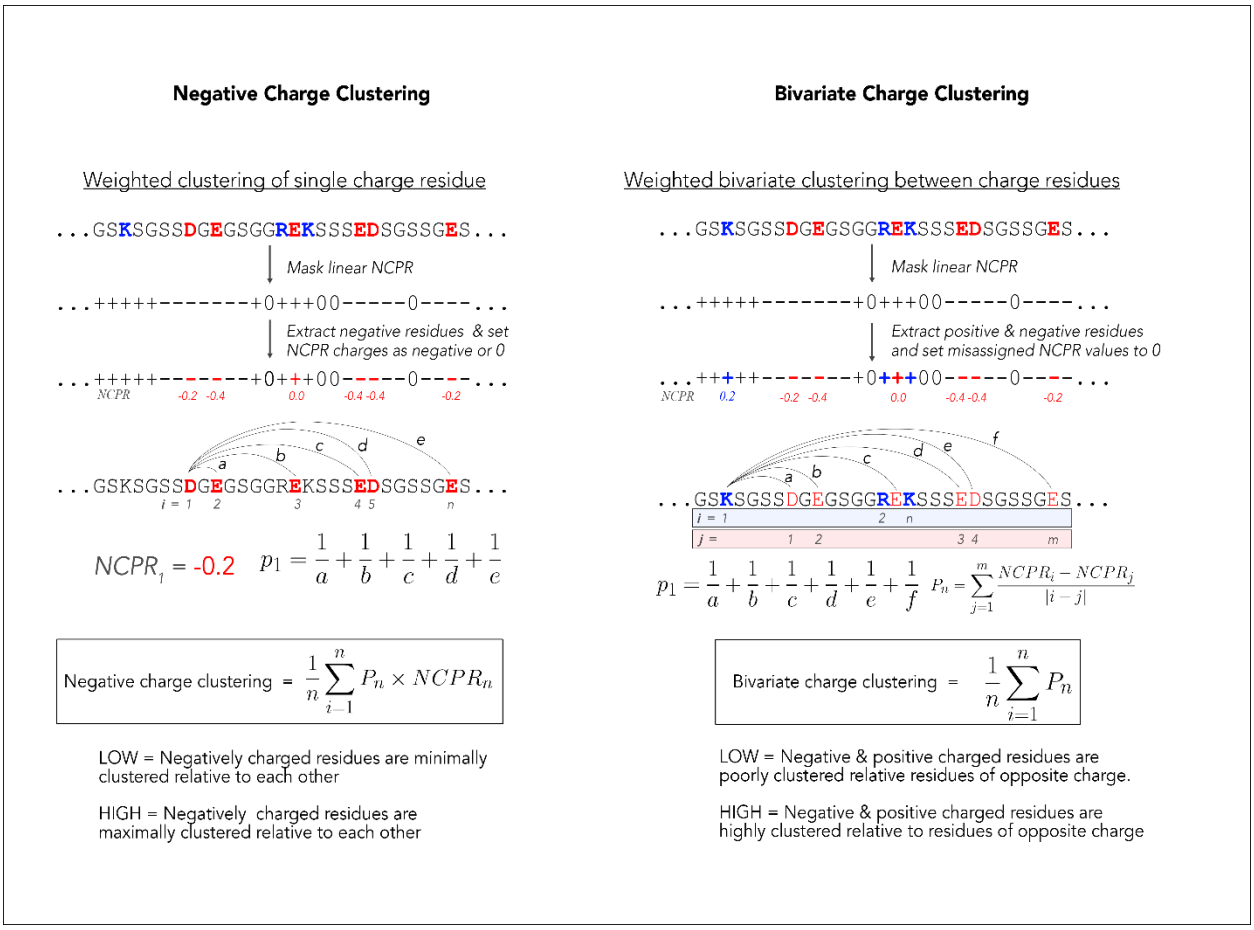

Figure S3: Overview of our computational approach for calculating charge clustering.

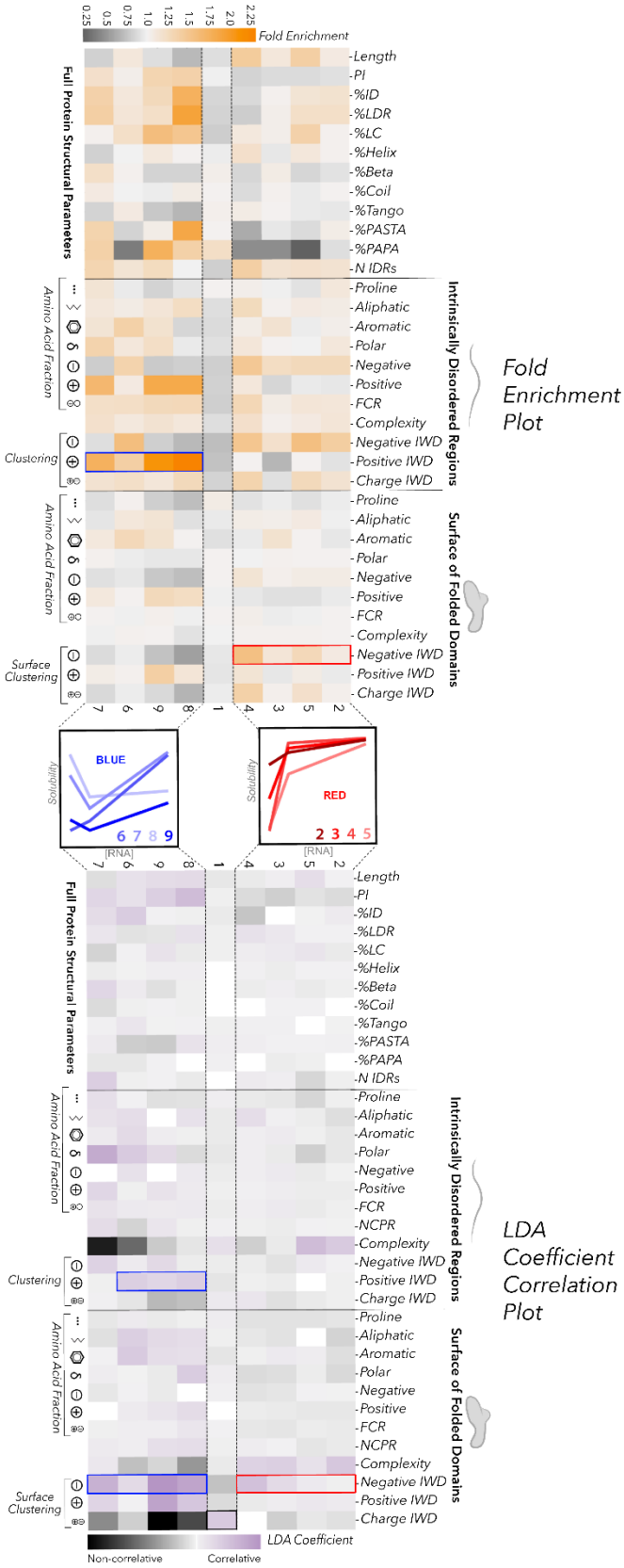

Figure S4: Enrichment and correlation of sequence features in different clusters. The upper heat map shows the fold enrichment of sequence features, while the bottom heatmap reports on the correlation of the sequence feature with the cluster (coefficients of linear discriminants as calculated from an LDA). In both heatmaps, each column denotes a cluster with blue clusters on the left and red clusters on the right, and each row denotes a different sequence feature. Sequence features are grouped by features calculated using residues in just the IDRs, features calculated based on residues on the surface of the folded domains, and features calculated on the entire protein sequence. The magnitude of the LDA coefficients in the bottom heatmap describe how well correlative a feature is with each cluster, while the sign of the coefficient denotes correlation (negative coefficients describe non-correlative features and do not inform on anti-correlation).

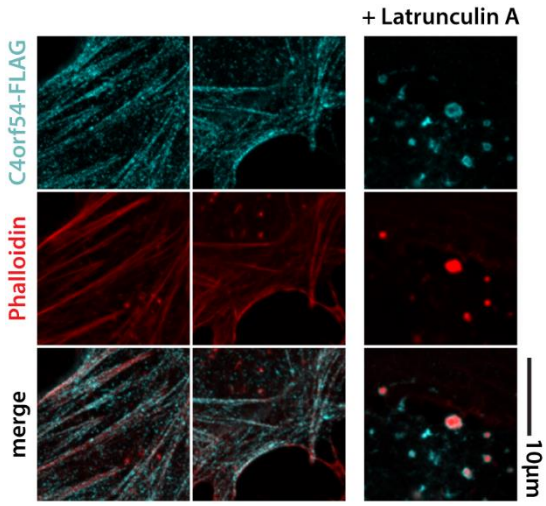

Figure S5: C4orf54 is an actin-binding protein. Expression of C4orf54-FLAG in U2OS cells shows colocalization with the actin cytoskeleton (phalloidin staining). C4orf54 forms shells around condensed actin after treatment with latrunculin A.

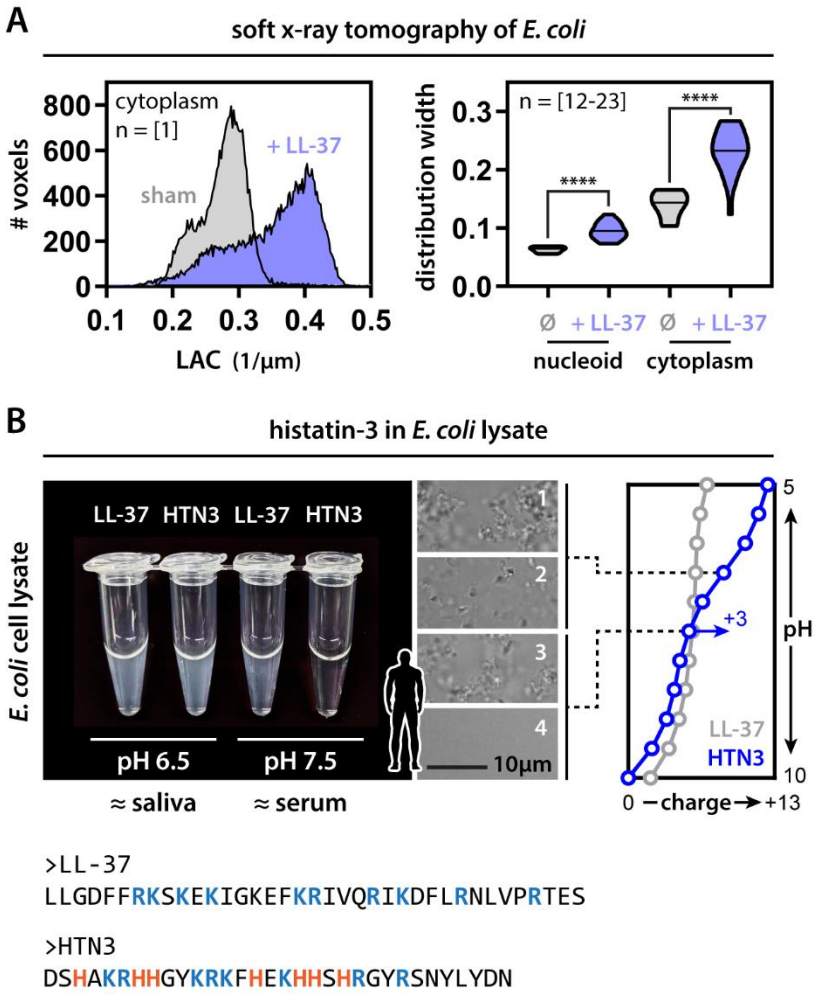

Figure S6: LL-37 and HTN3. (A) Molecular density (reported as the X-ray linear absorption coefficient or LAC) distribution of individual *E. coli* cells imaged with soft x-ray tomography shows increase in heterogeneity upon LL-37 treatment (i.e., wider distribution). This change in heterogeneity is observed for both nucleus and cytoplasm. Mann-Whitney. **** p-value < 0.0001. (B) HTN3 only drives *ex vivo* condensate formation under conditions similar to those of saliva (i.e., mildly acidic). Under these conditions HTN3 gains +3 net charge, while LL-37 retains the same charge.

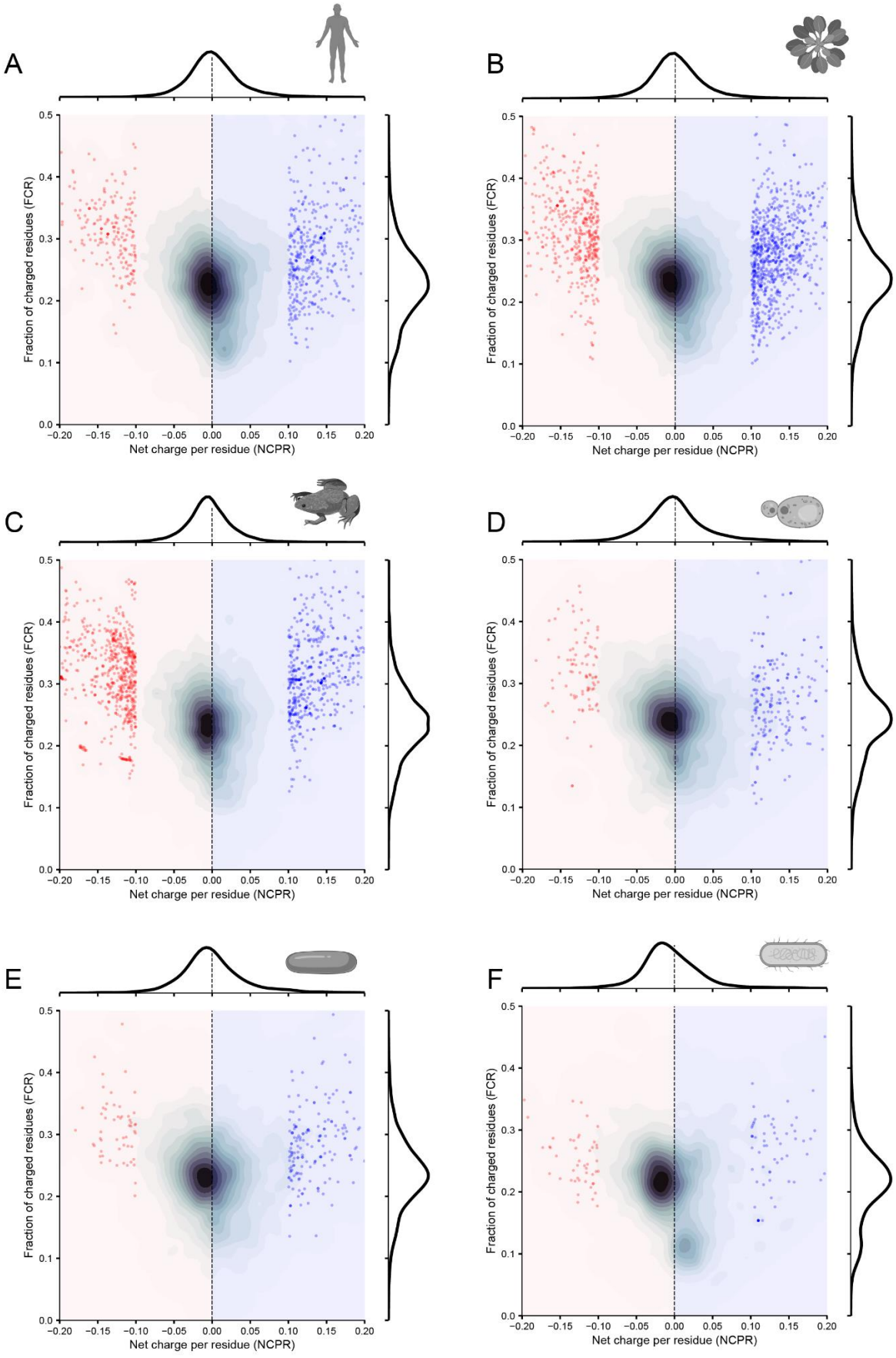

Figure S7: Proteome-wide analysis for six different organisms reveals a small bias towards negatively charged proteins, yet many examples of positively charged proteins exist. This analysis examined (A) *Homo sapiens* (n=20,393), (B) *Arabidopsis thaliana* (n=39,319), (C) *Xenopus laevis* (n=49,880), (D) *Saccharomyces cerevisiae* (n=6,060), (E) *Schizosccharomyces pombe* (n=5,122), and (F) *Escherichia coli* (n=4,438). In all cases, there are approximately an equal number of proteins with a net positive charge as a net negative charge. Proteins with a net charge per residue above 0.10 or below –0.10 are identified as individual markers, while all other proteins are reported as contour density data. Marginal distributions for the net charge per residue (NCPR) and fraction of charged residues (FCR) are shown alongside the two-dimensional distributions.

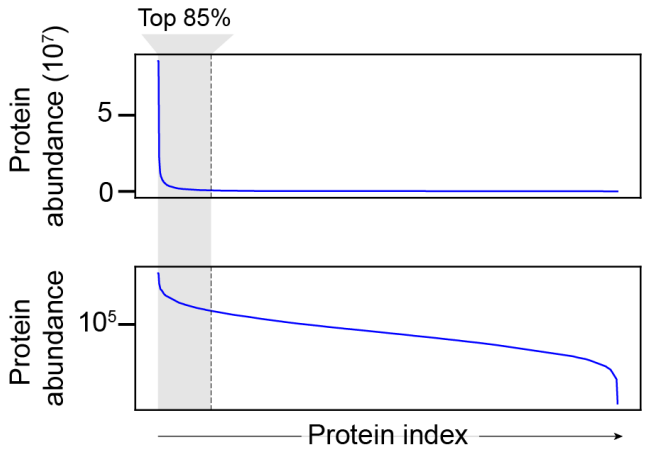

Figure S8: Quantitative mass-spectrometry data from Hein *et al.*[128] provides copy number information for 9,209 proteins. We rank-ordered those proteins by most to least abundant and took the set of proteins that comprise the top 85% of proteins in a human cell by copy number. This ensures our analysis only focuses on a subset of proteins that are found in high abundance (i.e., in human cells, ~10,000 copies or more). This same approach was used to select high-abundance proteins across five other organisms.

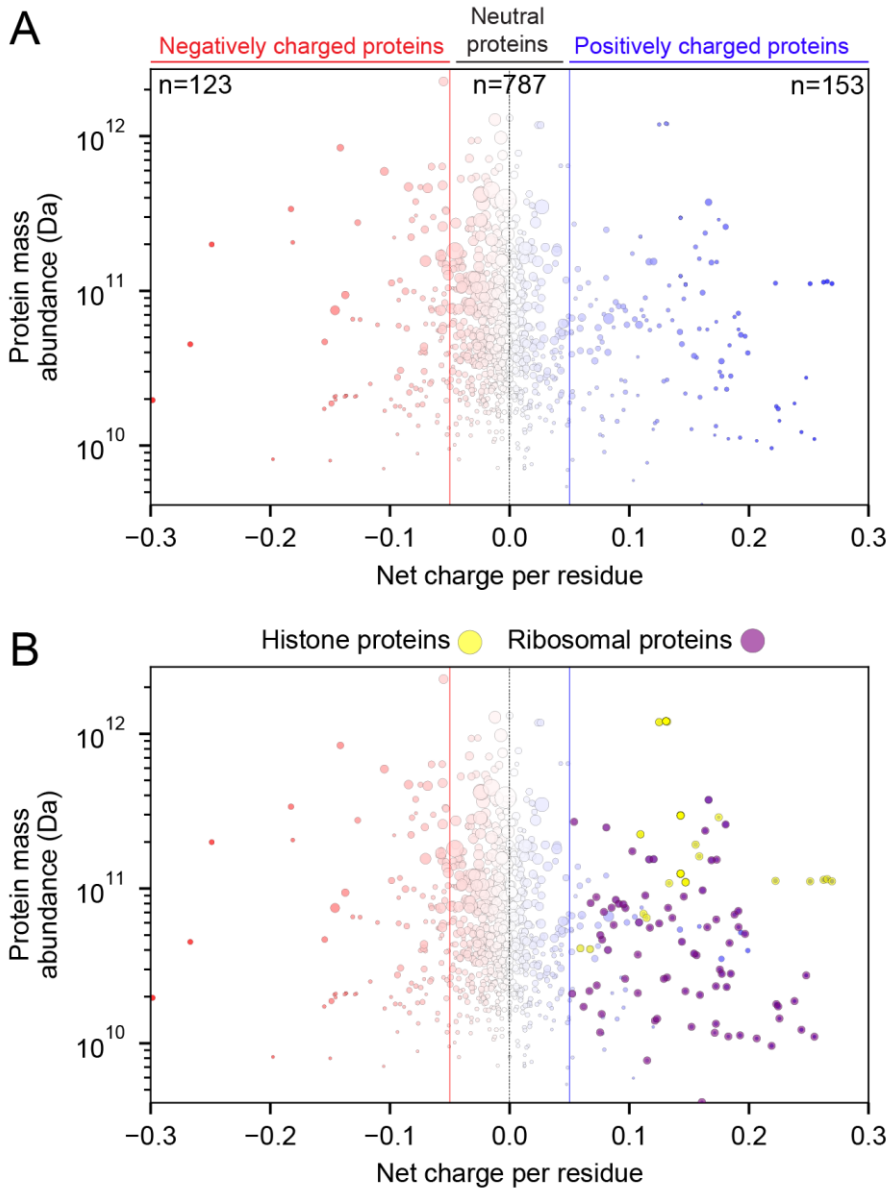

Figure S9: Assessment of highly abundant proteins across the human. (A) Of the 1063 highly abundant proteins identified, just 153 have a net positive charge, and of those, the majority (~90%) are sequestered in constitutive molecular complexes with RNA (ribosomal proteins, RNP proteins), DNA (histones) or phospholipids (membrane proteins). The remaining proteins are nuclear RNA-binding proteins. (B) Same data as shown in panel A with histones and ribosomal proteins explicitly labelled for convenience.

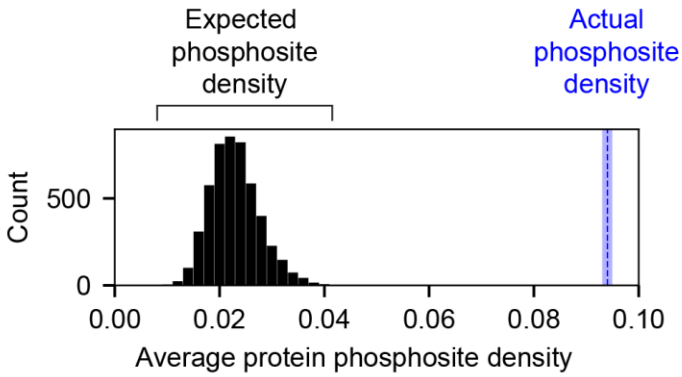

Figure S10: The expected phosphosite density was generated by randomly selecting 24 proteins and calculating the average phosphosite density (number of residues that possess phosphosites / total number of residues) 5000 times with replacement to construct a null distribution. This null distribution gives an expected average phosphosite density of 0.02 (i.e., one in every fifty residues is phosphorylated). The value and the associated distribution can be compared with the actual phosphosite density for the 24 non-RNP/membrane-bound proteins identified from our 153 highly abundant positively charged human proteins. The phosphosite density for these 24 proteins is 0.94, or one in every ten residues.

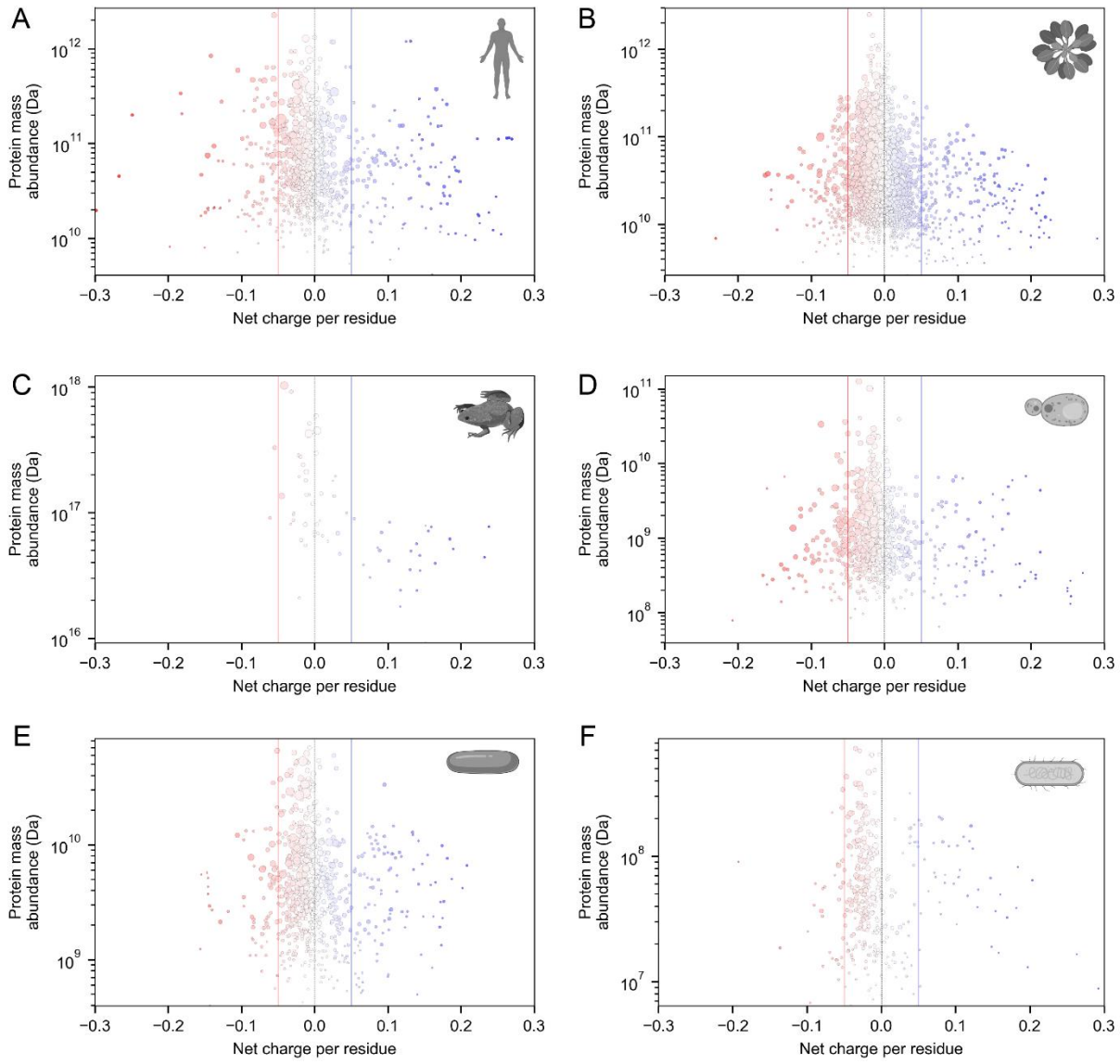

Figure S11: Analysis of abundant, positively charged proteins across the tree of life consistently identifies positively charged proteins as histones, ribosomal proteins, membrane proteins, or proteins engaged as parts of ribonuclear protein (RNP) complexes (see **Table S3**).

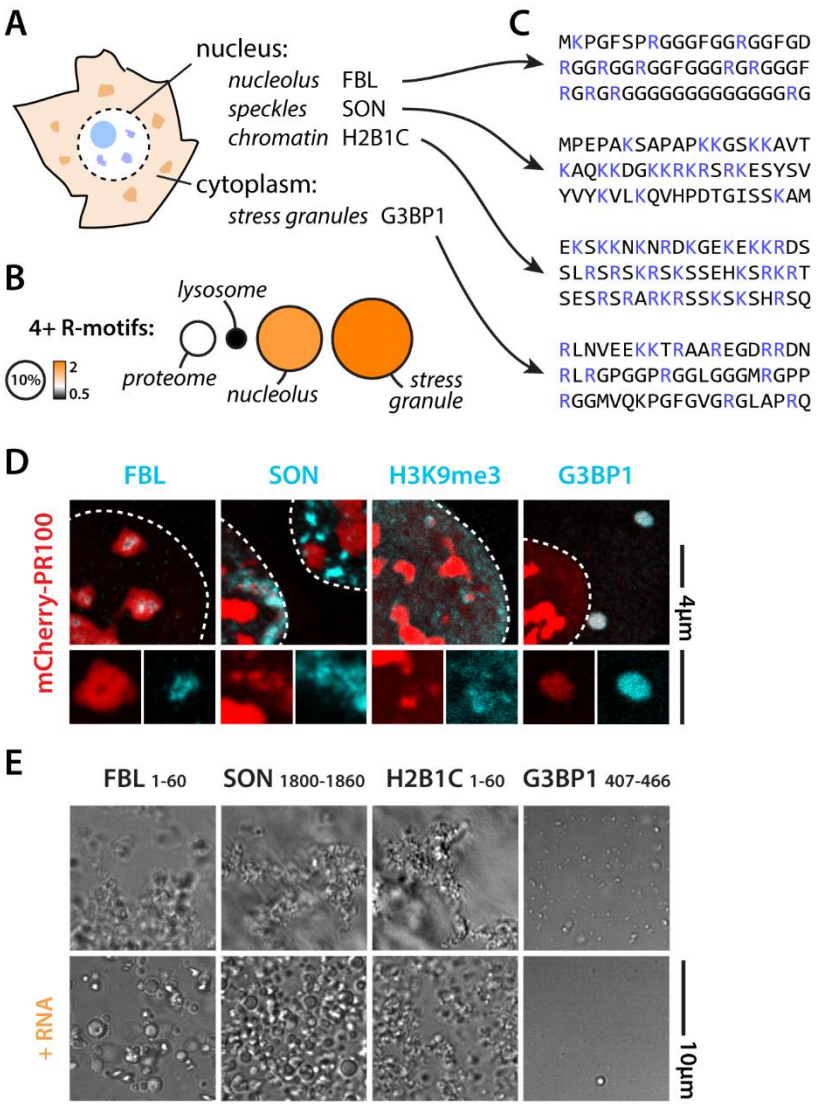

Figure S12: Endogenous cationic IDRs show similar behavior to disease ones when in isolation. (A) Key proteins from several RNA-centric condensates (B) RNA-centric condensates are enriched in arginine-rich motifs (fraction of proteins with at least four motifs[138]). Circle size indicated percentage of proteins with arginine-rich motif. Color indicated fold enrichment. (C) Example cationic IDRs from key condensate proteins. (D) These condensate proteins all colocalize with PR when expressed in cells. (E) Isolated cationic IDR peptides form irregular gel-like condensates when added to cell lysate, but are modulated into more spherical assemblies upon addition of RNA (2 μg/μl).