Figure S1

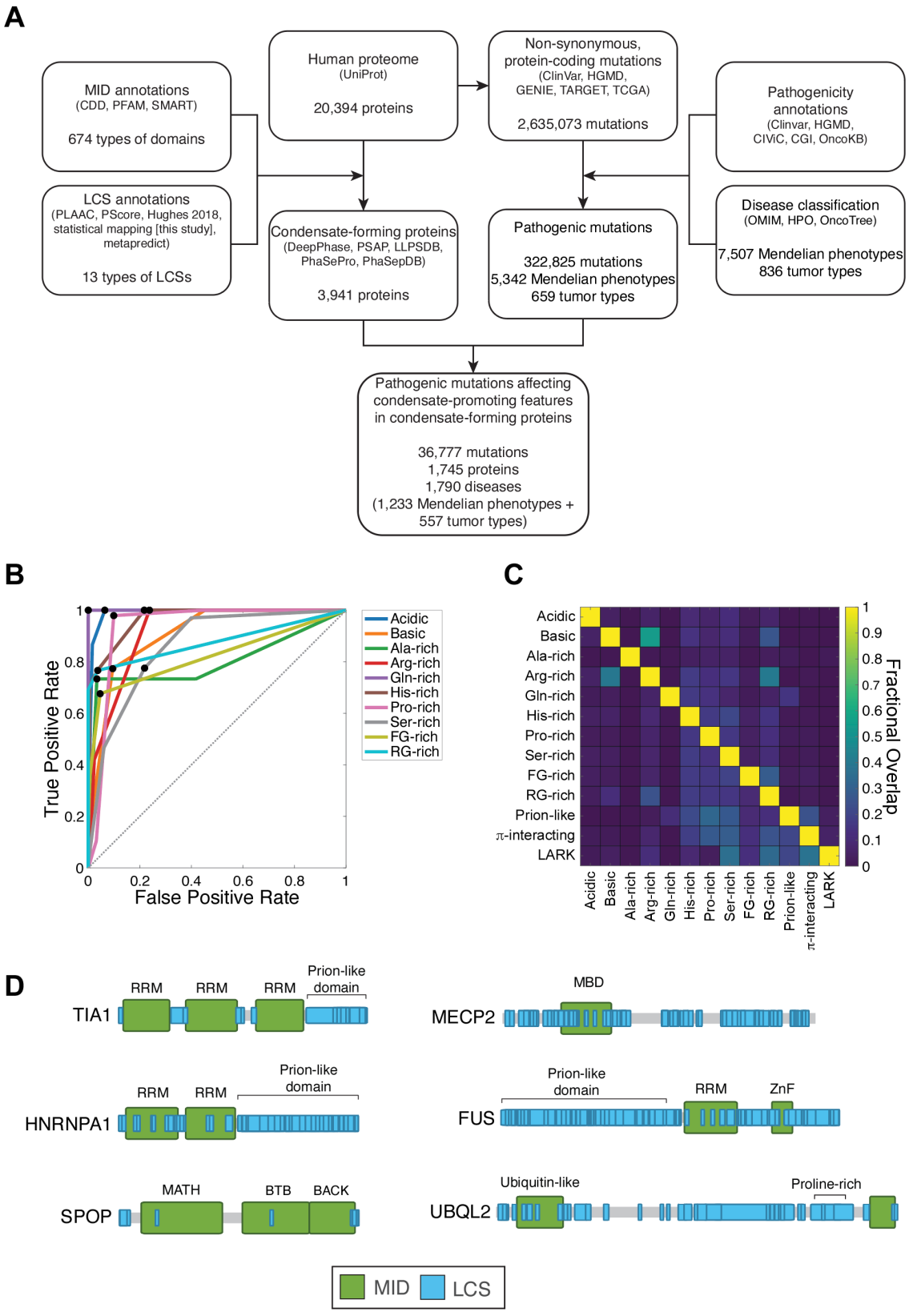
A
MID annotations
(CDD, PFAM, SMART)
674 types of domains
LCS annotations
(PLAAC, PScore, Hughes 2018, statistical mapping [this study], metapredict)
13 types of LCSs
Human proteome
(UniProt)
20,394 proteins
Condensate-forming proteins
(DeepPhase, PSAP, LLPSDB, PhaSePro, PhaSepDB)
3,941 proteins
Non-synonymous, protein-coding mutations
(ClinVar, HGMD, GENIE, TARGET, TCGA)
2,635,073 mutations
Pathogenic mutations
322,825 mutations
5,342 Mendelian phenotypes
659 tumor types
Pathogenicity annotations
(Clinvar, HGMD, CIViC, CGI, OncoKB)
Disease classification
(OMIM, HPO, OncoTree)
7,507 Mendelian phenotypes
836 tumor types
Pathogenic mutations affecting condensate-promoting features in condensate-forming proteins
36,777 mutations
1,745 proteins
1,790 diseases
(1,233 Mendelian phenotypes + 557 tumor types)
B
True Positive Rate
False Positive Rate
Acidic
Basic
Ala-rich
Arg-rich
Gln-rich
His-rich
Pro-rich
Ser-rich
FG-rich
RG-rich
C
Fractional Overlap
Acidic
Basic
Ala-rich
Arg-rich
Gln-rich
His-rich
Pro-rich
Ser-rich
FG-rich
RG-rich
Prion-like
π-interacting
LARK
D
TIA1
RRM
RRM
RRM
Prion-like domain
MECP2
MBD
HNRNPA1
RRM
RRM
Prion-like domain
FUS
Prion-like domain
RRM
ZnF
SPOP
MATH
BTB
BACK
UBQL2
Ubiquitin-like
Proline-rich
MID
LCS

Figure S1. Mapping condensate-promoting features in condensate-forming proteins. Related to Figure 1.

A. Flow chart showing computational steps taken to define putative condensate-forming proteins, condensate-promoting features, pathogenic mutations, and disease annotations. Data sources are shown in parenthesis.
B. ROC curves for optimizing cutoffs for mapping indicated LCSs across the proteome, benchmarked against a set of validated LCSs (Supplementary Table 2).
C. Fractional overlap between two types of LCSs across the proteome. The fraction of residues mapped as a particular type of LCS (y-axis) that overlap with another type of LCS (x-axis) is indicated by the color scale.
D. Examples of known condensate-forming proteins and their condensate-promoting features, as identified by the analytical approach used in this study.

Figure S2

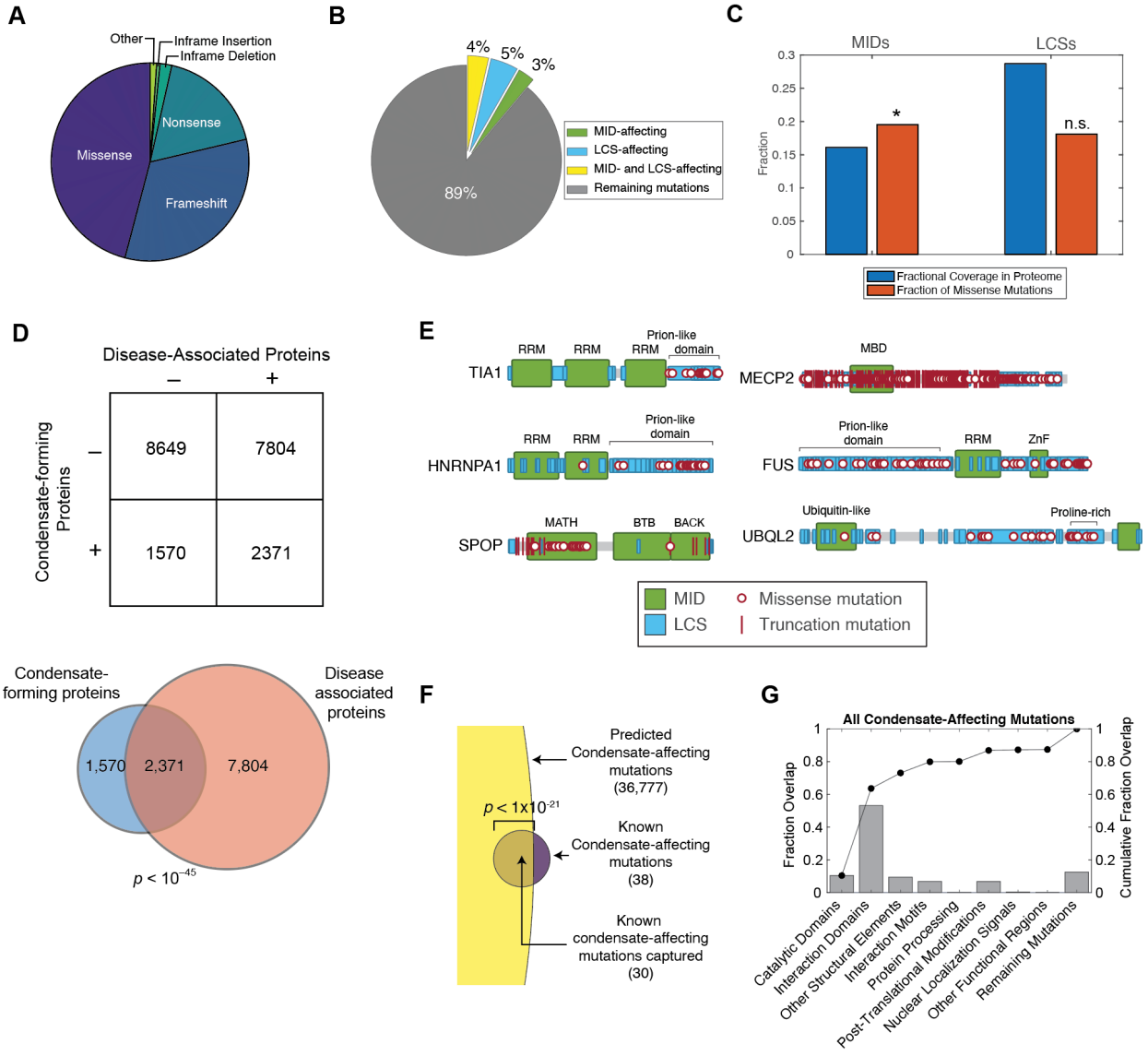
A
Other
Inframe Insertion
Inframe Deletion
Nonsense
Missense
Frameshift
B
4%
5%
3%
89%
MID-affecting
LCS-affecting
MID- and LCS-affecting
Remaining mutations
C
MIDs
LCSs
Fraction
*
n.s.
Fractional Coverage in Proteome
Fraction of Missense Mutations
D
Disease-Associated Proteins
–
+
Condensate-forming Proteins
8649
7804
1570
2371
Condensate-forming proteins
Disease associated proteins
1,570
2,371
7,804
p < 10-45
E
TIA1
RRM
RRM
RRM
Prion-like domain
MECP2
MBD
HNRNPA1
FUS
ZnF
SPOP
MATH
BTB
BACK
UBQL2
Ubiquitin-like
Proline-rich
MID
LCS
Missense mutation
Truncation mutation
F
Predicted Condensate-affecting mutations (36,777)
p < 1x10-21
Known Condensate-affecting mutations (38)
Known condensate-affecting mutations captured (30)
G
All Condensate-Affecting Mutations
Fraction Overlap
Cumulative Fraction Overlap
Catalytic Domains
Interaction Domains
Other Structural Elements
Interaction Motifs
Protein Processing
Post-Translational Modifications
Nuclear Localization Signals
Other Functional Regions
Remaining Mutations

Figure S2. Mapping mutations to condensate-promoting features. Related to Figure 1.

A. Pie chart showing the types of pathogenic mutations examined in the study.
B. Pie chart showing the proportion of pathogenic mutations that affect MIDs, LCSs, or both MIDs and LCSs within putative condensate-forming proteins.
C. (Blue) Bar graph of fraction of proteome comprised of MIDs (left) or LCSs (right) along with (red) fraction of all pathogenic missense mutations that occur within MIDs or LCSs. Missense mutations were significantly enriched among MIDs (*), but not among LCSs (n.s., not significant). p-value < 10^{-250} and p-value ~ 1, respectively, binomial test.
D. Venn diagram showing a significant overlap between condensate-forming and proteins with pathogenic mutations ($p < 10^{-45}$, Fisher exact test)
E. Examples of known condensate-forming proteins, their condensate-promoting features, and pathogenic mutations affecting these features as identified by the analytical approach used in this study.
F. Overlap of predicted condensate-affecting mutations with known condensate-affecting mutations curated from the literature (see also Table S4).
G. Bar graph of the cumulative proportion of pathogenic condensate-affecting mutations that overlap with canonical molecular-scale models. Moving from left to right, mutations in multiple categories were not double counted. Mutations with no overlap are denoted as “remaining mutations”.

Figure S3

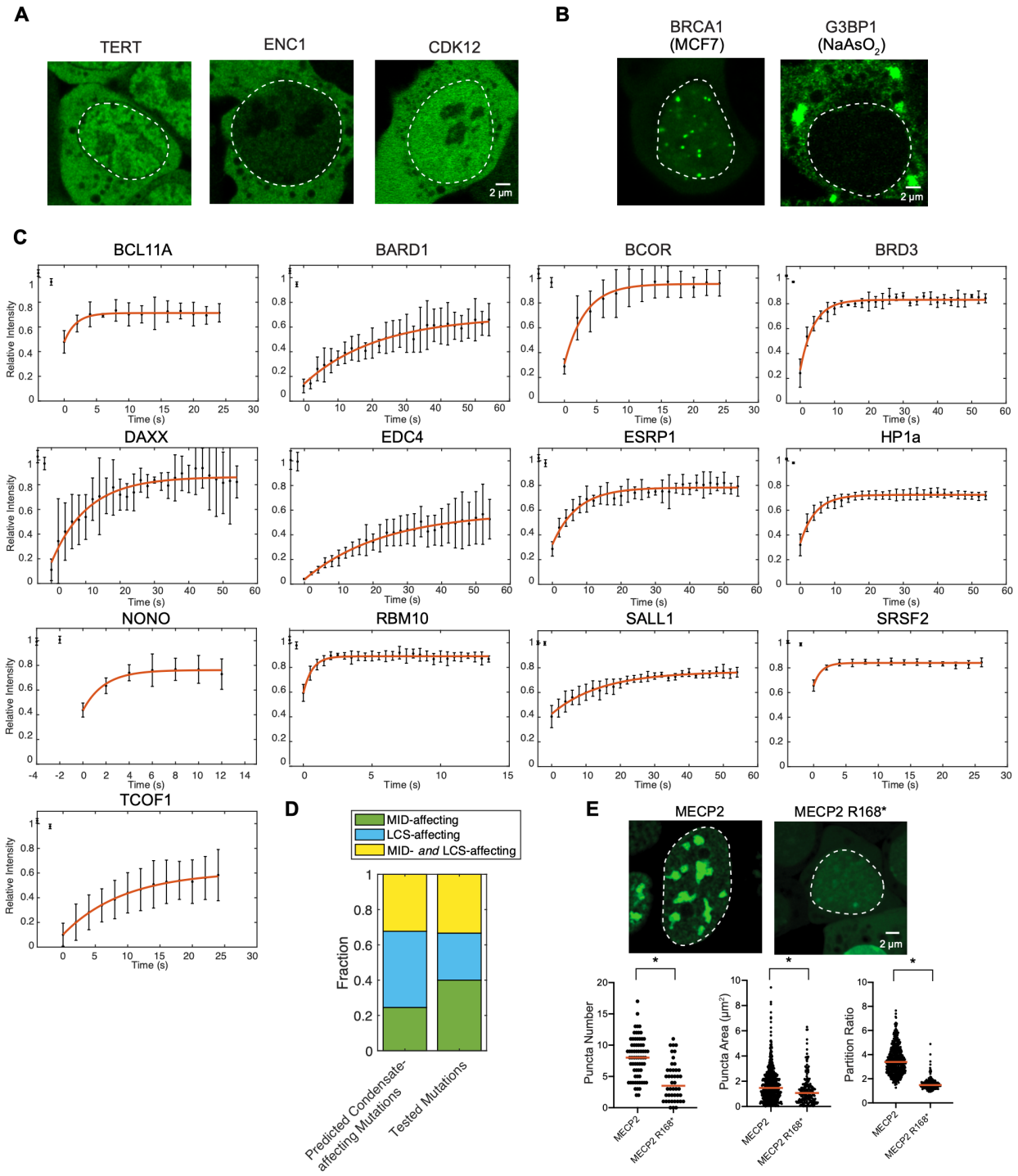
A
TERT
ENC1
CDK12
2 µm
B
BRCA1 (MCF7)
G3BP1 (NaAsO2)
2 µm
C
BCL11A
BARD1
BCOR
BRD3
DAXX
EDC4
ESRP1
HP1a
NONO
RBM10
SALL1
SRSF2
TCOF1
Relative Intensity
Time (s)
D
MID-affecting
LCS-affecting
MID- and LCS-affecting
Fraction
Predicted Condensate-affecting Mutations
Tested Mutations
E
MECP2
MECP2 R168*
2 µm
Puncta Number
Puncta Area (µm2)
Partition Ratio
MECP2
MECP2 R168*

Figure S3. Pathogenic mutations in condensate-promoting features alter condensate properties in live cells. Related to Figure 3.

A. Representative images of 3 of the 12 mEGFP-tagged candidate proteins that did not incorporate into punctate structures indicative of condensates in mESCs. Nuclei indicated with dotted white line. See also Table S5.

B. Representative images of mGFP-tagged BRCA1 in MCF7 cells (left) and mEGFP-tagged G3BP1 in NaAsO2 treated mESCs (right), showing incorporation into punctate structures. Both candidates were found to be not punctate in unstressed mESCs. Nuclei are shown with white dotted lines. See also Supplementary Table 5.

C. Fluorescence recovery after photobleaching (FRAP) of 13 candidate proteins that incorporated into punctate structures in mESCs. Fluorescence intensities normalized to pre-bleach values are plotted over time. Data points represent the mean and error bars represent standard deviation of normalized fluorescent values over n = 3-4 puncta in independent cells that were analyzed for each candidate.

D. Fraction of predicted condensate-affecting mutations (left) or those selected for experimental evaluation (right) that were MID-affecting, LCS-affecting, or MID- and LCS-affecting.

E. Representative images (top) and quantification of puncta number (bottom, left), puncta area (bottom, middle), and partition ratio (bottom, right) for mEGFP-tagged MECP2 and mEGFP-tagged MECP2 R168* mutant in mESCs.

Figure S4

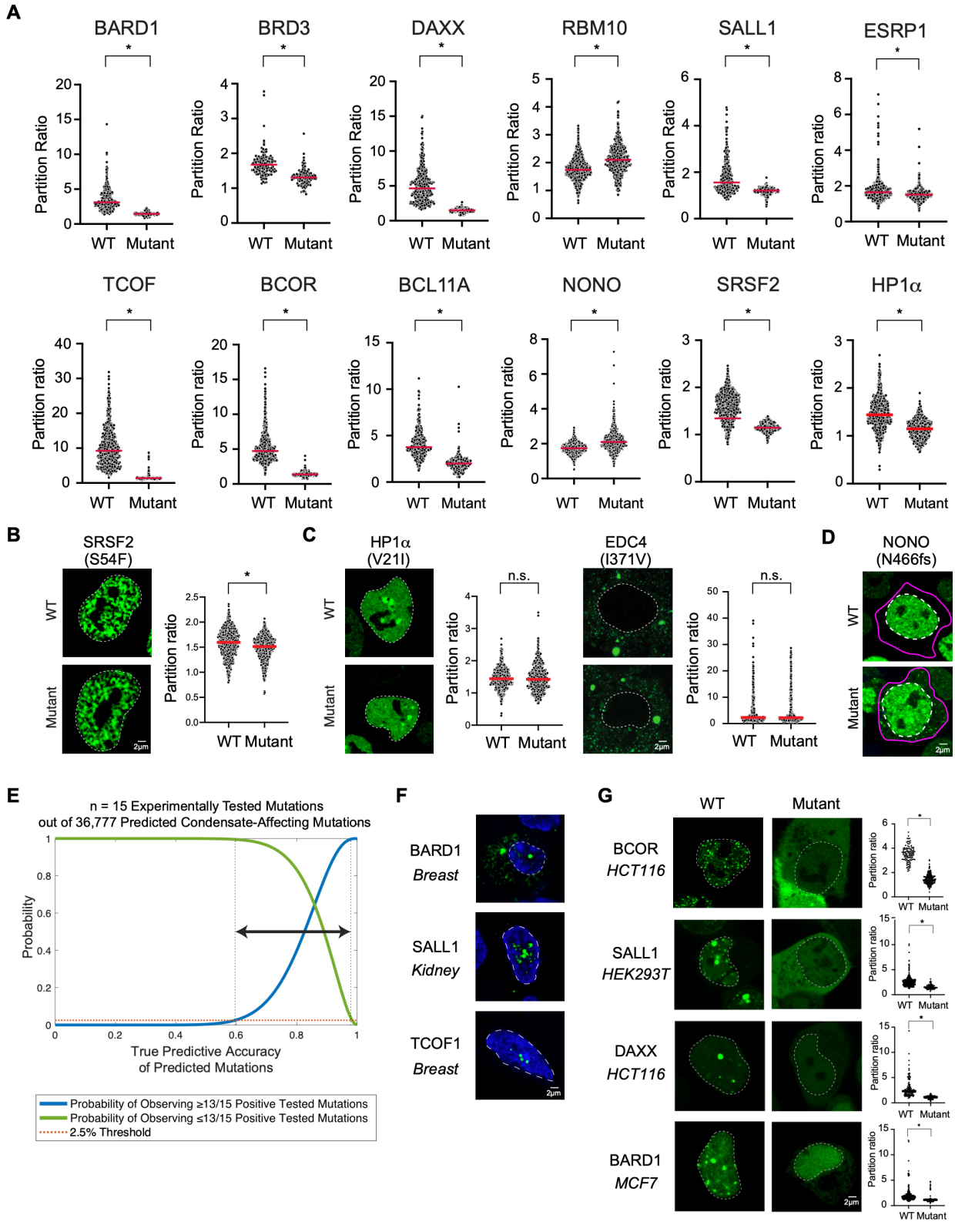
A
BARD1
BRD3
DAXX
RBM10
SALL1
ESRP1
TCOF
BCOR
BCL11A
NONO
SRSF2
HP1α
Partition Ratio
Partition ratio
WT
Mutant
B
SRSF2 (S54F)
C
HP1α (V21I)
EDC4 (I371V)
n.s.
D
NONO (N466fs)
E
n = 15 Experimentally Tested Mutations
out of 36,777 Predicted Condensate-Affecting Mutations
Probability
True Predictive Accuracy of Predicted Mutations
Probability of Observing ≥13/15 Positive Tested Mutations
Probability of Observing ≤13/15 Positive Tested Mutations
2.5% Threshold
F
BARD1 Breast
SALL1 Kidney
TCOF1 Breast
G
BCOR HCT116
SALL1 HEK293T
DAXX HCT116
BARD1 MCF7

Figure S4. Mutations in condensate-promoting features cause diverse condensate dysregulation phenotypes. Related to Figure 3.

A. Quantification of partition ratios of wild-type and mutant mEGFP-tagged cell lines for indicated candidates. *, *p*-values < 0.0001. See also Table S6.

B. Representative images (left) and quantification (right) for a second SRSF2 mutant tested in mEGFP-tagged mESC lines, showing a mild but significant effect on partitioning. *, *p*-value < 0.0001.

C. Representative images and quantification for a second HP1a mutant (left) and a WT/mutant pair for EDC4 (right) showing no significant difference in partitioning between the WT versus the mutant protein.

D. Representative images of mESC lines expressing WT or mutant NONO, reproduced from Fig. 3c, showing the cytoplasmic boundary of the cell of interest (magenta). Nuclear outline is shown as a white dashed line.

E. Estimation of the 95% confidence interval of the accuracy of the predicted catalog of condensate-affecting mutations. Under the assumption of randomly selected mutations, the probability of observing ≥ 13/15 (blue) or ≤ 13/15 (green) positive experimental outcomes from a catalog of n = 36,777 mutations as a function of all possible true catalog accuracies. Probabilities are computed using a hypergeometric distribution with the assumption of random sampling. Red dashed line indicates probability of 2.5% and its intersections with the red and blue curves denote an estimate for the 95% confidence interval for the true catalog accuracy to be between 60-98%.

F. Representative images of BARD1, SALL1, and TCOF1 immunofluorescence in human breast, kidney, and breast tissue, respectively. Nuclear outline is shown as a white dashed line.

G. Representative images of disease-relevant human cell lines expressing WT or mutant BCOR, SALL1, DAXX, and BARD1 (left) with quantification of partition ratios of wild-type and mutant cell lines for indicated candidates (right). *, *p*-values < 0.0001. Specific cell lines are shown adjacent to the image. Nuclear outline is shown as a white dashed line.

Table S1. LCSs used in this study and optimization of LCS mapping against a set of benchmark proteins. Related to Figure 1.

LCS	Amino acids	Test Protein	Curated coordinates	Refs.	Mapped coordinates	AUC	Optimal Frequency in 5-amino acid Window	TPR	FPR
Acidic patch	Asp, Glu	NPM1	34-39, 120-130, 161-188	(Mitrea et al., 2016)	118-134,159-189	0.99	0.5	1.0	0.06
Basic patch	Arg, Lys	MECP2	170-181, 184-194, 246-258, 263-274, 282-289, 301-310, 340-348	(Li et al., 2020)	26-39,105-115, 167-178, 249-258, 264-272,302-310	0.91	0.5	0.77	0.09
Alanine-rich region	Ala	HXD13	57-71	(Basu et al., 2020)	18-39,46-75,92-126	0.80	1.0	0.73	0.03
Arginine-rich region	Arg	SURF6	29-31, 56-58, 81-82, 118-120, 145-148, 152-159, 216-217, 221-223, 234-237, 246-249, 299-306, 321-326, 330-331, 336-345	(Mitrea et al., 2016, 2018)	142-163, 213-225, 233-250, 295-310, 320-349	0.92	0.3	1.0	0.24
Glutamine-rich region	Gln	HTT	17-40	(Peskett et al., 2018)	15-41, 49-68, 498-506, 593-601	1.00	0.5	1.0	$3x10^{-4}$
Histidine-rich region	His	DYR1A	590-616	(Lu et al., 2018)	531-541, 596-622, 648-656	0.97	0.1	1.0	0.22
Proline-rich region	Pro	UBQL2	491-538	(Dao et al., 2018)	7-17, 313-327, 470-478, 490-537, 575-580	0.94	0.3	0.98	0.10
Serine-rich region	Ser	MED1	1078-1482	(Sabari et al., 2018)	808-816, 1021-1029, 1077-1150, 1163-1171, 1223-1285, 1366-1375, 1463-1471, 1532-1543	0.86	0.3	0.78	0.22
FG-rich region	Gly, Phe	NUP98	1-469	(Schmidt and Görlich, 2015)	3-68, 75-102, 113-153, 224-278, 286-306, 315-395, 405-424, 433-453, 461-483, 871-884, 1052-1060	0.83	0.3	0.69	0.05
RG-rich region	Arg, Gly	FUS	211-285, 371-422, 453-526	(Wang et al., 2018)	211-222, 230-270, 375-411, 469-507	0.88	0.3	0.77	0.04
Prion-like domain	-	-	-	(Lancaster et al., 2014; Martin et al., 2020; Wang et al., 2018)	-	-	-	-	-
pi-interacting residues	-	-	-	(Vernon et al., 2018)	-	-	-	-	-
LARKS	-	-	-	(Hughes et al., 2018)	-	-	-	-	-

List of LCSs within IDRs associated with condensate formation used in this study. Preexisting approaches were used when available to map certain LCSs across the proteome (prion-like domains, pi-interacting residues, and LARKS). The remainder of LCSs were mapped using a statistical approach (Methods) benchmarked against protein regions corresponding to previously characterized LCSs in test proteins identified in prior studies. These curated LCSs were used as 'gold standards' for benchmarking our LCS mapping parameter of frequency of corresponding amino acid types within 5-amino acid windows (see Methods). The optimal cutoff and its performance are indicated. AUC, area under the curve from ROC curves (Figure S1B). FG, phenylalanine-glycine; RG, arginine-glycine; LARKS, low-complexity, aromatic-rich kinked segments; AUC, area under the curve; TPR, true positive rate; FPR, false positive rate; Refs., references.

Table S2. Pathological mutations known to affect condensates. Related to Figure 3.

Protein	Diseases	Mutations	References	Notes for uncaptured mutations
ANXA11	Amyotrophic lateral sclerosis	**Asp40Gly**, **Gly38Arg**, **His390Pro**, **Arg456His**	(Nahm et al., 2020)	
αSYNUCLEIN	Parkinson's disease	Ser129Glu, Ala53Thr, Glu46Lys	(Ray et al., 2020)	Mutation not in variant databases sourced in this study; Mutation does not affect defined MIDs or LCSs
DDX3X	Medulloblastoma, Intellectual disability	Ala222Pro, **Thr275Met**, Gly302Val, Leu353Phe, Met370Arg, Leu351Trp, Leu556Ser	(Fonseca et al., 2021; Valentin-Vega et al., 2016)	Mutation not in variant databases sourced in this study; Mutation does not affect defined MIDs or LCSs
FUS	Amyotrophic lateral sclerosis and frontotemporal lobar dementia	**Gly399Val**, **Gly187Ser**, **Gly156Glu**	(Burke et al., 2015; Patel et al., 2015)	
HNRNPA1	Amyotrophic lateral sclerosis	***Asp314Val**	(Molliex et al., 2015)	
HNRNPDL	Limb-girdle muscular dystrophy 1G	**Asp378His**, **Asp378Asn**	(Batlle et al., 2020)	
KEAP1	Lung squamous cell carcinoma	Arg320Glu, **Arg470Cys**	(Cloer et al., 2018)	Mutation not in variant databases sourced in this study
MECP2	Rett syndrome	**Arg168Ter**, **Arg255Ter**, **Arg270Ter**, **Arg294Ter**, Pro389Ter, **Arg306Cys**, Pro322Leu, Pro225Arg	(Li et al., 2020)	Mutation does not affect defined MIDs or LCSs
MLL4	Kabuki Syndrome	Gln4092Ter	(Fasciani et al., 2020)	Mutation not in variant databases sourced in this study
NF2	Cancers	Leu46Arg, Leu64Pro, Leu141Pro	(Meng et al., 2021)	Protein not captured in set of defined condensate-forming proteins
RBM20	Congenital dilated cardiomyopathy	**Arg636Ser**	(Schneider et al., 2020)	
SHP2	Noonan syndrome	**Glu76Lys**, **Arg498Leu**, Gln506Pro, Gly464Ala, **Tyr279Cys**, Tyr468Met	(Zhu et al., 2020)	Mutation not in variant databases sourced in this study; Mutation does not affect defined MIDs or LCSs
SPOP	Prostate cancer	**Phe133Val**, **Trp131Gly**	(Bouchard et al., 2018)	
TDP43	Amyotrophic lateral sclerosis	**Ala321Gly**, **Ala321Val**, Glu331Lys, **Met337Val**, Ala326Pro, Met337Pro	(Conicella et al., 2016)	Mutation not in variant databases sourced in this study;
TIA1	Amyotrophic lateral sclerosis and frontotemporal dementia	**Pro362Leu**, **Ala381Thr**, **Glu384Lys**	(Mackenzie et al., 2017)	
TAU	Alzheimer's disease	Pro301Leu	(Kanaan et al., 2020)	Mutations not in variant databases sourced in this study
UBQL2	Amyotrophic lateral sclerosis	**Pro506Ser**, **Pro506Thr**, **Pro506Ala**, Thr487Ile, **Pro497Leu**, **Pro497His**, **Pro497Ser**	(Dao et al., 2019; Sharkey et al., 2018)	Mutation does not affect defined MIDs or LCSs

Known pathological mutations that affect condensates curated from the literature. **Bolded** mutations, were captured among the set of mutations predicted to affect condensates in the catalog. Reasons for why certain mutations were not captured in our catalog are

mentioned in the right-most column. *, this HNRNPA1 mutation is described in the corresponding study as Asp262Val, affecting a non-canonical isoform and maps to position 314 in the canonical isoform used in our analyses. See also Figure S2D.

Table S3. Selected protein candidates and mutations used in experimental tests. Related to Figure 3.

Protein	cDNA source	Disease(s)	Distribution in mESCs	Selected Mutation(s)	Mutation effect
MECP2	(Li et al., 2020)	Rett syndrome	Punctate	Arg186Ter	Reduced condensate incorporation
BARD1	MHS6278-211689242	Breast, ovarian, prostate, Pancreatic cancers	Punctate	Arg406Ter	Reduced condensate incorporation
BCL11A	Addgene 139809	Intellectual development disorders	Punctate	Gln177Ter	Reduced condensate incorporation
BCOR	MHS6278-202757783	Various cancers	Punctate	Tyr657Ter	Reduced condensate incorporation
BRD3	Unpublished	Intellectual disability	Punctate	Phe334Ser	Reduced condensate incorporation
HP1α	(Li et al., 2020)	Developmental disorder, Autism	Punctate	Val21Ile, Trp142Cys	Reduced condensate incorporation
DAXX	Addgene 52023	Various cancers	Punctate	Arg318Ter	Reduced condensate incorporation
EDC4	Addgene 66597	Congenital heart disease	Punctate	Ile371Val	Reduced condensate incorporation
ESRP1	MHS6278-202833454	Deafness, Ear malformation	Punctate	Leu259Val	Reduced condensate incorporation
NONO	Addgene 127655	Developmental delay	Punctate	Asn466fs	Enhanced condensate incorporation; Altered condensate localization
RBM10	Addgene 81958	Lung, bladder, colon, pancreatic cancers	Punctate	Val354Met	Enhanced condensate incorporation
SALL1	MMM1013-202859719	Townes-Brocks syndrome	Punctate	Ser372Ter	Reduced condensate incorporation
SRSF2	(Guo et al., 2019)	Acute myeloid leukemia, Myelodysplastic syndrome	Punctate	Ser54Phe, Pro95His	Reduced condensate incorporation
TCOF	MHS1010-202695722	Treacher-Collins syndrome	Punctate	Gln55Ter	Reduced condensate incorporation
ASXL1	MHS6278-213245938	Acute myeloid leukemia, Myelodysplastic syndrome	Not punctate	-	-
BCL6	Addgene 81869	Various cancers	Not punctate	-	-
BRCA1	Addgene 14999	Breast, ovarian cancers	Not punctate	-	-
DVL2	Addgene 24802	Neural tube defects	Not punctate	-	-
DYR1A	Addgene 101770	Autism, Intellectual disability	Not punctate	-	-
ENC1	MHS6278-202826591	Autism	Not punctate	-	-
G3BP1	Addgene 127104	Autism	Not punctate	-	-
HMGA2	Addgene 52727	Silver-Russel syndrome	Not punctate	-	-
NIPBL	Addgene 107716	Cornelia-deLange syndrome	Not punctate	-	-
NKX21	Addgene 119173	Choreoathetosis	Not punctate	-	-
SNCAP	MHS6278-202809062	Parkinson disease	Not punctate	-	-
TERT	Addgene 114315	Dyskeratosis congenita	Not punctate	-	-

25 protein candidates from catalog selected for experimental study, not including MECP2, which was used as a positive control (Li et al., 2020). cDNA source indicates Addgene catalog number, cDNA clone ID (Team et al., 2009), or a prior study. mESCs, mouse

embryonic stem cells. See also Supplemental Discussion, *Selection of Candidates and Mutations for Experimental Validation*.