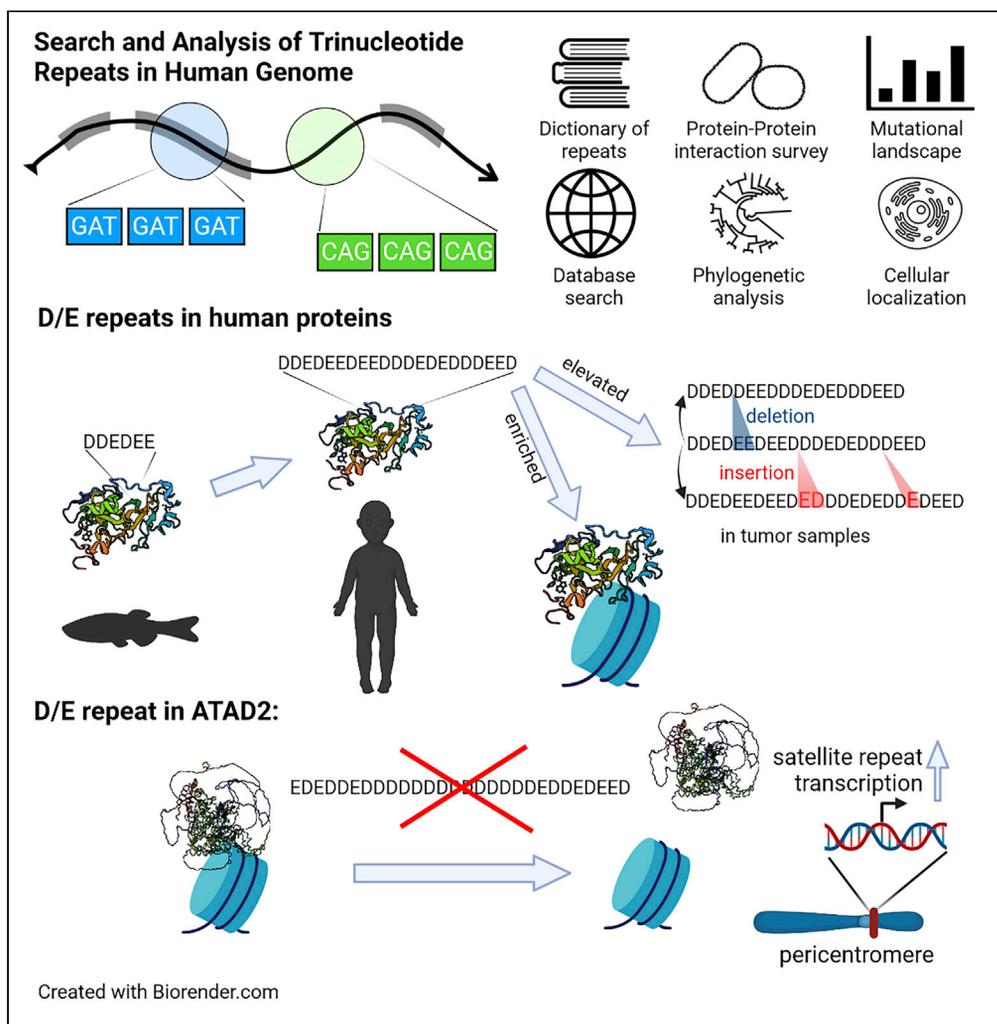


Article

Genome-wide survey of D/E repeats in human proteins uncovers their instability and aids in identifying their role in the chromatin regulator ATAD2



Shalabh Shukla,
Pavlo Lazarchuk,
Maria N. Pavlova,
Julia M. Sidorova

julias@uw.edu

Highlights

Many human proteins contain runs of aspartic/ glutamic acid residues (D/E repeats)

D/E repeats show increased incidence of in-frame insertions/deletions in tumors

Nuclear and histone-interacting proteins often have long D/E repeats

D/E repeat of the oncogene ATAD2 controls its binding to pericentric chromatin

Shukla et al., iScience 25, 105464
December 22, 2022 © 2022 The Author(s).
<https://doi.org/10.1016/j.isci.2022.105464>



Article

Genome-wide survey of D/E repeats in human proteins uncovers their instability and aids in identifying their role in the chromatin regulator ATAD2

Shalabh Shukla,¹ Pavlo Lazarchuk,¹ Maria N. Pavlova,¹ and Julia M. Sidorova^{1,2,*}

SUMMARY

D/E repeats are stretches of aspartic and/or glutamic acid residues found in over 150 human proteins. We examined genomic stability of D/E repeats and functional characteristics of D/E repeat-containing proteins vis-à-vis the proteins with poly-Q or poly-A repeats, which are known to undergo pathologic expansions. Mining of tumor sequencing data revealed that D/E repeat-coding regions are similar to those coding poly-Qs and poly-As in increased incidence of trinucleotide insertions/deletions but differ in types and incidence of substitutions. D/E repeat-containing proteins preferentially function in chromatin metabolism and are the more likely to be nuclear and interact with core histones, the longer their repeats are. One of the longest D/E repeats of unknown function is in ATAD2, a bromodomain family ATPase frequently overexpressed in tumors. We demonstrate that D/E repeat deletion in ATAD2 suppresses its binding to nascent and mature chromatin and to the constitutive pericentromeric heterochromatin, where ATAD2 represses satellite transcription.

INTRODUCTION

The ATPase family AAA Domain-containing protein 2 (ATAD2, also referred to as ANCCA, or AAA + Nuclear Coregulator Cancer-Associated), a member of a conserved AAA ATPase family,¹ has attracted major interest as a protein whose high expression correlates with poor prognosis in various cancers.^{2–8} Overexpression of ATAD2 is ubiquitous across many cancers,⁹ and its depletion inhibits growth of cancer cells, fueling an ongoing effort to develop clinically usable small molecule inhibitors of ATAD2.^{10,11}

Mammalian ATAD2 was first described as a transcriptional co-activator at genes regulated by MYC, AR, ER, and E2F,^{2,12–14} and later found to be associated with chromatin genome-wide as a global facilitator of transcription.^{3,15,16} We and others found human ATAD2 on the newly-replicated, immature chromatin.^{17,18} Yeast homologs of ATAD2 were implicated in maintaining heterochromatin (the *Schizosaccharomyces pombe* homolog of ATAD2, Abo1,^{19,20}), or limiting its spread (the *Saccharomyces cerevisiae* homolog Yta7,²¹ also reviewed in Cattaneo et al.¹). Heterochromatic localization of human ATAD2 has been inferred^{17,22} but not pursued in detail.

In addition to a two-partite AAA + ATPase domain and a bromodomain,^{3,14,23,24} ATAD2 contains a 28 amino acid-long uninterrupted run of aspartic and glutamic acid residues in its N-terminal half, hereafter referred to as a D/E repeat (in ATAD2 this region has also been called Acidic N terminal region, ANR,¹ however, this term does not distinguish it from the regions that are merely enriched for acidic residues). The function of the human ATAD2 D/E repeat is unknown, but indirect evidence from *S. cerevisiae* and mouse homologs suggests that it may participate in binding to chromatin.^{3,25} Repeats of Ds or Es, or any combination of both residues have been noted in many human proteins^{26,27}. Although in a handful of proteins D/E repeats have been implicated in binding to histones or other positively charged proteins, or mimicking a DNA substrate for DNA-binding proteins,^{27–29} overall these repeats have not received much attention as a group.

D/E repeats are encoded by runs of trinucleotides GAC, GAT, GAG, and GAA. Short tandem repeats in DNA are considered more unstable compared to non-repetitive DNA, and as a subcategory of short tandem repeats, trinucleotide repeats (TNRs) exhibit genetic instability manifesting as contractions and

¹Department of Laboratory Medicine and Pathology, University of Washington, 1959 NE Pacific St., Box 357705, Seattle, WA 98195, USA

²Lead contact

*Correspondence: julias@uw.edu

<https://doi.org/10.1016/j.isci.2022.105464>



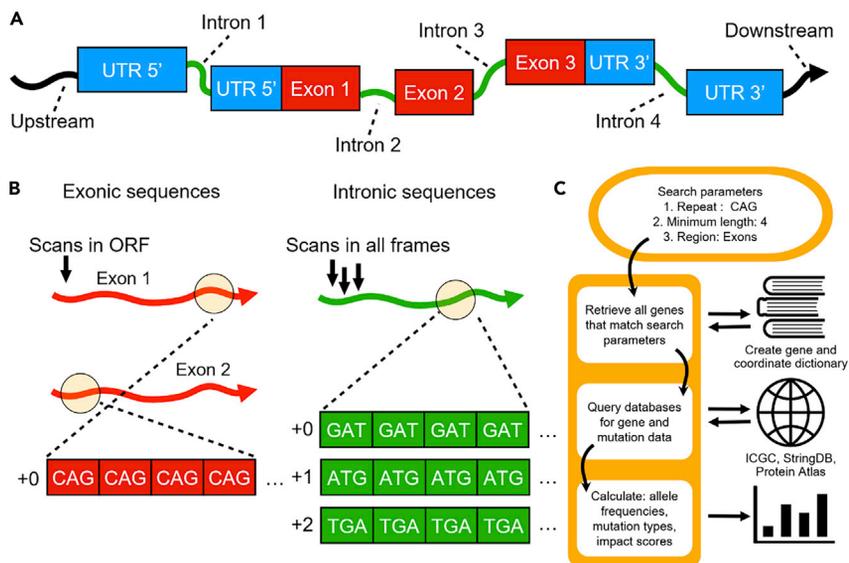


Figure 1. An overview of the TriRepSearch pipeline

(A) A schematic of the regions of a protein-coding gene that are included in the search. Note that only the sense (top) strand of a gene is searched.

(B) Region-specific sequences are extracted for analysis by TriRepSearch. Exonic sequences are scanned in the frame of the ORF only and continuously, accounting for the uninterrupted amino acid repeats that are encoded on more than one exon. Intronic regions are scanned in all three possible trinucleotide frames (+0, +1, and +2) and individually because an uninterrupted trinucleotide repeat can by definition span only one intron.

(C) The general workflow of the TriRepSearch package for search and analysis of triplet repeats in protein-coding genes. Search parameters specified by the user can allow for the search of non-homogeneous and discontinuous trinucleotide repeats.

expansions of the number of repeat units.^{30,31} Repeat variability can fuel the functional plasticity of repeat-containing proteins,^{32,33} which bears significance for cancers, where repeats may exhibit elevated instability.^{34,35} Extreme repeat expansions in non-coding and coding DNA can result in cellular dysfunction and disease.³⁶ However, the degree of instability can vary among different repeats, and it is unknown how D/E repeat-encoding sequences compare to others, for example, to the poly-Q repeats encoded by the CAG TNR and poly-A repeats encoded by the GCN TNR, both of which can undergo pathologic expansions in exons.³⁷ Although the TNR of GAA, a trinucleotide that encodes the E residue, is known to expand in introns of the genes *FXN* and *DMD* (reviewed in Khristich and Mirkin³⁸), exonic expansions of (GAA)*n* or any expansions of (GAG)*n*, (GAT)*n*, or (GAC)*n* are currently unknown. At the same time, insertions or deletions of 1–2 D/E residues have been noted, e.g., in the *COMP*³⁹ and *ASPN*^{40,41} genes, where they may be clinically significant.

Here we compare the distribution and instability of the D/E-coding TNRs in the human genome versus the well-known expandable TNRs, and derive an overview of D/E repeat function via proteome-wide analysis of subcellular localization, functional categorization, and protein-protein interactions using custom-designed computational tools. Based on the common themes emerging from these analyses, we address the function of the ATAD2 D/E repeat *in vivo* and demonstrate for the first time that it is critical for ATAD2 binding to chromatin, including the constitutive heterochromatin of the pericentromere, where it is important for transcriptional silencing.

RESULTS

An overview of D/E repeat-coding TNRs vis-a-vis other coding TNRs

Figure 1 outlines the implemented trinucleotide repeat search pipeline. To delineate the scope of our task, we began by identifying all the homopolymeric TNRs, i.e., uninterrupted repeats of the same trinucleotide, in the protein-coding fraction of the human genome (note that all trinucleotides are identified by their sequence in the sense DNA strand). The results demonstrate a wide range of the maximum lengths

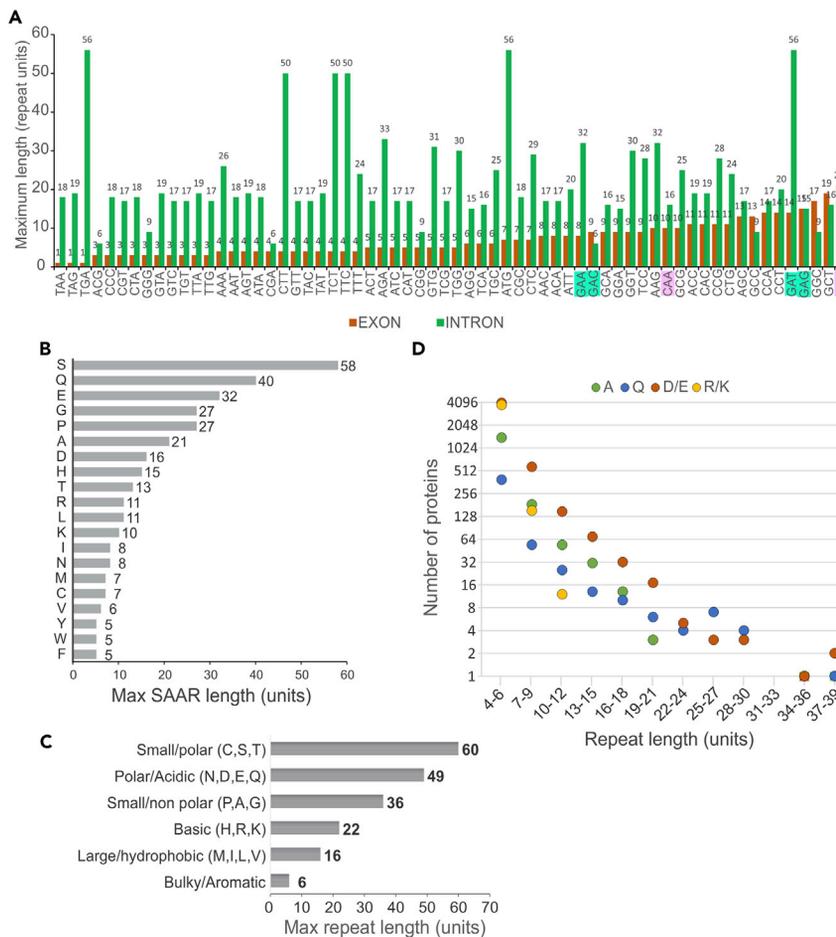


Figure 2. An overview of trinucleotide and amino acid repeats in the protein-coding subset of the human genome

(A) Maximum lengths of uninterrupted homopolymeric repeats of the indicated trinucleotides in exons and introns of human protein-coding genes. The trinucleotides coding for D or E are highlighted in green and those coding for Q – are in pink. Note that for the intronic trinucleotide repeats, each region with a repeat was effectively annotated three times, i.e., in each of the three reading frames (also see Figure 1).

(B) Maximum lengths of single amino acid repeats (SAARs).

(C) Maximum lengths of multiple amino acid repeats (MAARs) comprising amino acids of similar properties.

(D) Totals of human proteins containing A, Q, D/E, or R/K amino acids repeated the indicated number of times.

Note that in all graphs repeat lengths are in numbers of repeated units, i.e., trinucleotides in (A) and amino acids in (B–D). See also Figure S1 and Tables S1 and S2.

attained by different homopolymeric TNRs in exons and introns (Figure 2A). At $n = 21$, (CAG) $_n$, which encodes poly-Q repeats, is the longest found in the human exome (of the GRCh37 release 105), and repeats of at least some of the D or E-encoding trinucleotides (GAG and GAT) are in the top five (Figure 2A). Intronic (GAT) $_n$ are found at a record-breaking $n = 56$. Notable examples of proteins with homopolymeric repeats in exons are listed in Table S1.

The majority of amino acid repeats are represented by heteropolymeric TNRs⁴² because the number of units in single amino acid repeats (SAARs) is typically larger than in the homopolymeric TNRs that encode them (Figures 2A and 2B). Among all the different SAARs, the poly-D and poly-E SAARs reach the maximal lengths that are comparable to the poly-A and poly-Q SAARs in the human proteome. Notably however, poly-Es can be twice as long as poly-Ds, which incidentally is similar to poly-Qs being much longer than poly-Ns (see discussion).

D/E repeats were also among the longest multiple amino acid repeats (MAARs), defined as uninterrupted runs of amino acids of one type (e.g., basic or acidic, Figure 2C and Table S2), where they were second only

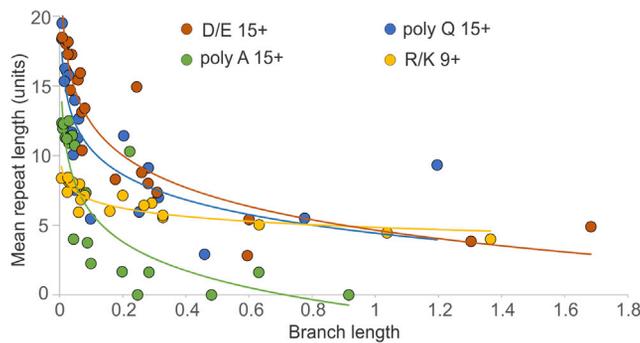


Figure 3. Expansion of repeat lengths in phylogeny

Shown are the homologs of the human proteins that carry repeats of poly-Q, poly-A, or D/E of 15 residues or longer, or R/K of 9 residues or longer. Mean repeat lengths of these homologs in each phylum are plotted as a function of evolutionary divergence (branch length) from *H. sapiens*. See also [Figure S2](#).

to the small/polar (C, S, T) MAARs and far longer than basic (H, R, K) MAARs, in agreement with.²⁶ Thus, at least in terms of their maximal uninterrupted length, D/E repeats are on par with the poly-Q repeats and exceed poly-A repeats, both of which are known to show instability and expandability.

D/E, poly-Q, and poly-A repeats display similar relationships between the length of the repeats and the number of proteins that contain a repeat of a given length ([Figure 2D](#)). The overwhelming majority of D/E and poly-Q proteins and the entirety of poly-A proteins contain uninterrupted repeats no longer than 33 aa. In all three sets, the number of proteins containing a repeat of a particular length drops sharply as the length of a repeat increases, and proteins with repeats less than 16 amino acids long comprise 99% or more of each set. Although there are very few poly-A proteins with repeats longer than 17 units, the subsets of poly-Q and D/E repeat proteins that contain repeats longer than 15 amino acids are of similar number. Thus, the overall similar values distribution and sample size make D/E, poly-Q, and to a degree, poly-A repeat sets suitable for comparative analyses. The R/K repeat set, on the other hand, represents proteins with even shorter repeat lengths than poly-A proteins, making them less comparable to D/E and poly-Q repeat sets. For comparison, the lengths distributions of the D/E- and Q-coding trinucleotides in introns are shown in [Figure S1](#).

Evolutionary expansion of D/E repeats in complex organisms

Phylogenetic comparisons of poly-Q SAAR lengths in protein homologs suggested a tendency for expansion in complex organisms.⁴² To extend this analysis to D/E and other repeats, we identified eukaryotic homologs for a subset of human proteins with poly-Q, poly-A, or D/E repeat lengths above 15 residues, or R/K repeat lengths above 9 residues. To validate our data retrieval and analysis tools, we reproduced the analysis conducted in Kumar et al., 2016⁴² for two poly-Q proteins, ATXN3 and HTT, and arrived at a very similar result ([Figure S2A](#)). Next, we plotted the mean repeat lengths as a function of evolutionary distance from *Homo sapiens* for non-human homologs of our selected human repeat proteins ([Figure 3](#)). As expected, average uninterrupted poly-Q repeat lengths increased dramatically in primates (divergences of up to 0.1 on the X-axis in [Figure 3](#)). Of interest, D/E repeats followed a similar tendency of recent expansion. Poly-A repeats showed lesser expansion, and R/K repeats exhibited no expansion trend. These results suggest a similarity of selection pressures acting on D/E and poly-Q repeats.

Assessing instability of TNRs encoding D/E, poly-Q, poly-A, and R/K repeats in human tumors

We reasoned that potential genetic instability of D/E repeats can be gleaned from a comparison of mutation frequencies and types in the TNRs encoding these repeats versus the TNRs encoding the poly-Q and poly-A repeats, the two disease-associated SAARs with a known propensity to expand, and the R/K repeats that have no evidence of expansion. We used the publicly available sequencing data for tumor samples, as repeat instability is documented in this material and is relatable to normal tissues.³⁵ Mutation calls were downloaded from the ICGC data portal.^{43,44} For each repeat-containing protein, we retrieved mutation data for the repeats, the corresponding repeat-containing exons, and all coding sequences combined. For each of these categories, we derived mean tumor allele frequencies (MTAF) and mutations per base

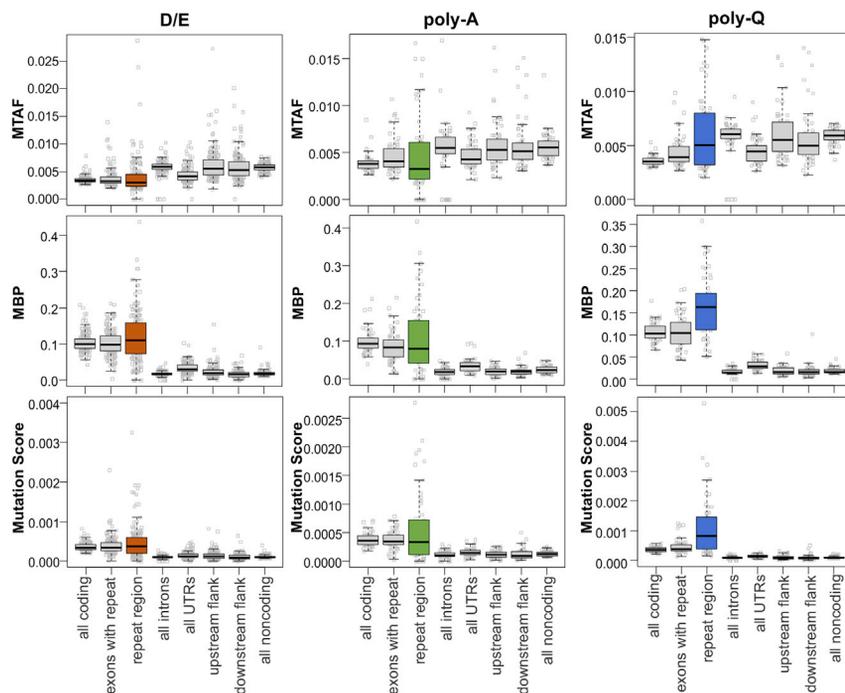


Figure 4. Mutation prevalence in exonic repeats in tumor samples

Mean tumor allele frequencies (MTAF), Mutations per base pair (MBP), and Mutation Scores (MTAF x MBP) were derived for the indicated coding and non-coding regions in the genes of proteins that contain poly-A, poly-Q, or D/E repeats of at least 12 residues, and plotted. All mutation types are included. Upstream and downstream flanks encompass 500 bp of DNA. See also [Figures S3](#) and [S4](#).

pair (MBP) metrics. For subsets of proteins with repeat lengths greater than 12 residues, we also derived MTAF and MBP for all non-coding sequences and 500 bp flanks upstream and downstream of the respective genes. Mutation scores, MS, for each region were derived as $MS = MTAF \times MBP$. These metrics were also compiled for intronic homopolymeric repeats of GAA, GAT, GAG (encoding D or E), and CAG (encoding Q). The GAC trinucleotide (encoding for E) was not included because it is found at lengths of only six units or shorter in intronic sequences ([Figure 2A](#)).

An overall comparison of the subset of genes for proteins with repeat lengths equal to or greater than 12 residues revealed a similar pattern across all repeats, in which a fraction of genes shows substantially elevated MTAF and/or MBP specifically within the repeat ([Figures S3](#), [S4](#), and [4](#)) compared to the repeat-encompassing exon or to all exons. Non-coding regions of these same genes scored overall higher MTAFs and lower MBPs than their coding regions, as expected, because non-coding sequences are less constrained and are of greater length overall than coding sequences. On average however, only poly-Q repeats showed as mutable enough to move the repeat-specific median MTAF, MBP, and MS values up compared to the all-exon baseline ([Figures 4](#) and [S3](#)).

Among the D/E proteins, Asporin (ASPN) showed a five times higher mutation score than the rest of the proteins, which overshadowed the rest of the variation. Excluding Asporin from the dataset allowed us to better see the elevated mutation scores for additional D/E proteins ([Figures 4](#) and [S3C](#)). Of note, ATAD2 scored an average MTAF but had the fifth highest MBP of the repeat versus the MBP of its entire coding sequence ([Figure S4A](#)). For all types of repeats, MTAF and MBP values did not correlate, and subsets of proteins had lower repeat MBPs than MBPs for the entire coding region ([Figures S4A–S4C](#)).

To reveal quantitative differences between the repeats at a finer level of detail, we asked how mutation scores (MSs) correlated with repeat length. For each repeat type, we plotted mean MSs for all proteins with repeat lengths no less than n residues over the range of n from 4 to 30 ([Figure 5A](#)). Given the shapes of the distributions in [Figure 2D](#), this allows the proteins with repeat lengths at the lower end of the range

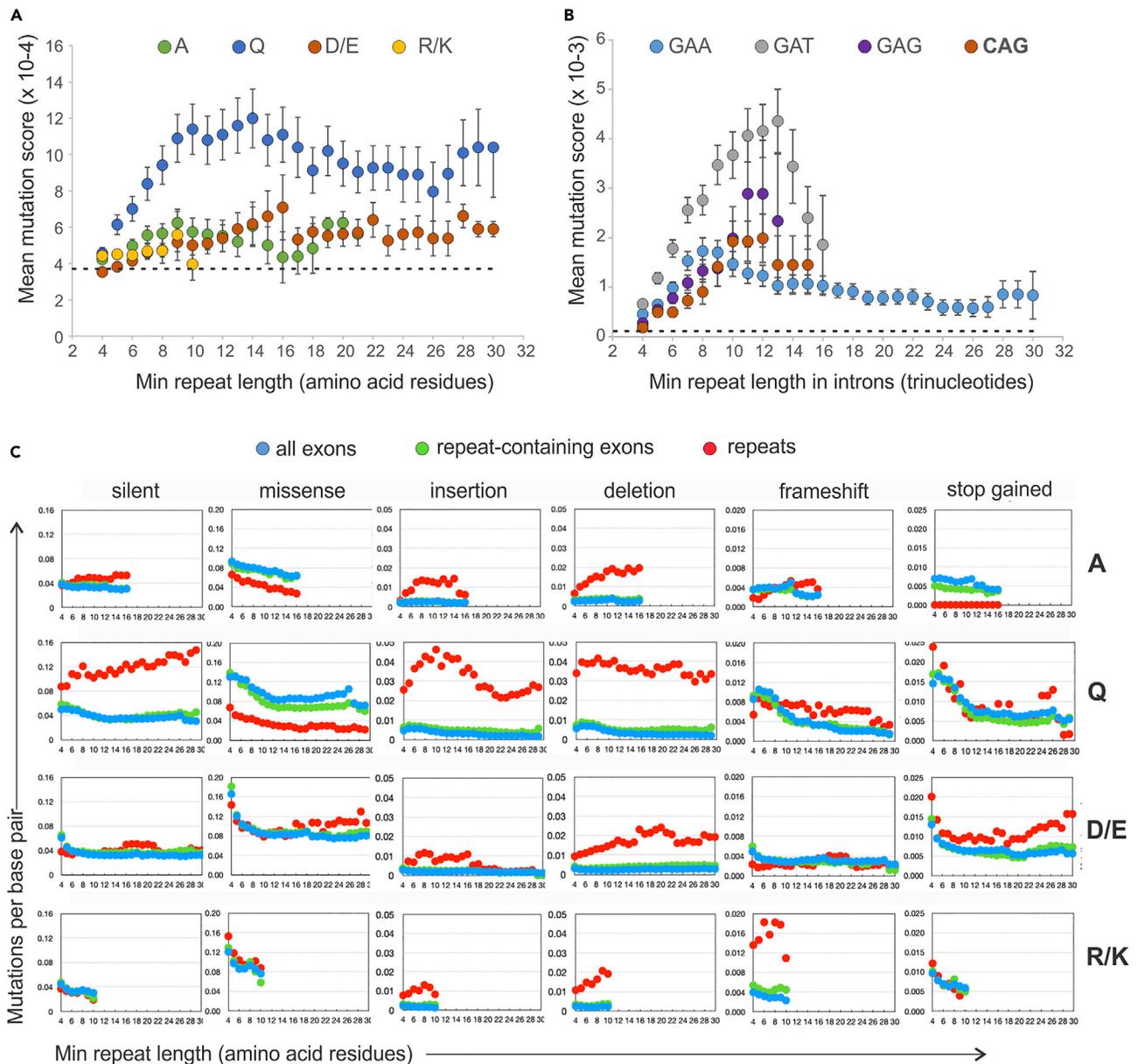


Figure 5. Mutation types and their dependence on repeat length

(A) Mutation scores of all within-repeat mutations were derived for the proteins with uninterrupted repeats no shorter than 4 and up to 30 residues long, and mean mutation scores were calculated for sets of proteins binned by repeat length, such that each bin contained the proteins with the repeat length at or longer than indicated by each X-axis value (expressed as number of residues). The minimal number of proteins in a bin was set at 3, and bins with one or two proteins were not included. Error bars are SEMs. A dotted line represents mean MS for all exons of all repeat-containing proteins in the sample, $3.6 \times 10^{-4} \pm 1.8 \times 10^{-6}$ SEM. p values are described in the main text.

(B) Mutation scores of all within-repeat mutations were derived for proteins with uninterrupted intronic repeats of (GAA)_n, (GAT)_n, (GAG)_n, or (CAG)_n, where n is between 4 and 30. Binning for the X-axis and minimal number of proteins in a bin are as in (A) except repeat lengths are expressed as the number of trinucleotides. Error bars are SEMs. A dotted line represents mean MS for all introns of all repeat-containing proteins in the sample, $1.3 \times 10^{-4} \pm 5.1 \times 10^{-5}$ SEM.

(C) For the same protein sets as in (A), means of mutations per base pair (MBPs) values for each mutation type were plotted for repeats (red), repeat-containing exons (green), and all exons (blue). X-axes represent sets of proteins binned by number of amino acid residues in the repeat, where each value is a minimal number of repeats allowed in the bin. Bins with fewer than 3 proteins were not included. p values adjusted for multiple comparisons (listed in Table 2) were derived in one-way ANCOVA tests with repeat length as a covariate.

See also Figures S5 and S6, and Table S3.

drive the sample mean for each bin. Repeat lengths with sample sizes of fewer than 3 proteins were not included. As [Figure 5A](#) shows, poly-Q repeat proteins displayed the highest mean MSs and the most dramatic MS increase dependent on length. D/E repeat proteins were in the second place, and poly-A and R/K proteins exhibited no significant correlation. For both poly-Q and D/E proteins, mean MSs strongly positively correlated with repeat lengths for n between 4 and 16 residues (Pearson coefficients 0.804 and 0.895, and p values 0.0005 and 0.00001, respectively). Of interest, for repeats longer than 16 residues, MS values peaked and leveled off for poly-Q proteins but not nearly as much for D/E proteins, in which the positive correlation between n and MS remained significant over the full range of repeat lengths (Pearson coefficient 0.623, p value 0.0005). Thus, the data suggest that although poly-Q repeats appear to be the most mutation-prone, D/E repeats definitely show an elevated level of mutations that correlates with repeat length.

A parallel analysis performed on intronic runs of GAA, GAT, GAG, and CAG TNRs revealed markedly elevated mutation scores compared to bulk introns, and these mutation scores rose and fell for all of the examined TNRs ([Figure 5B](#)). Note that these TNRs were identified by their sequence in the sense (top) DNA strand for the purpose of direct comparison with the exonic TNRs. Intriguingly, the expansion-associated (CAG) n and (GAA) n had comparable though not the highest MS, whereas (GAT) n showed the highest MS.

Repeat-preserving and repeat-disrupting mutation tendencies distinguish D/Es in exons and the cognate TNRs in introns from the pathologic expansion-associated TNRs

To address which types of mutations were elevated in the repeats and whether this varied with repeat length, we plotted mean MBPs for each mutation type within the repeats, repeat-containing exons, and whole coding regions of poly-A, poly-Q, R/K, and D/E proteins over a range of repeat lengths ([Figure 5C](#) and see [Table 1](#) for p values; only the cases in which repeats were significantly different both from all exons and from the repeat-containing exon are detailed below). First, we found that in-frame insertions and deletions within repeats were significantly elevated for all four types of repeats, albeit poly-Q repeats showed the highest MBPs for these mutation types. Second, silent mutations were significantly elevated, and missense mutations were reduced in poly-Qs and poly-As, though MBPs values were lower overall in poly-As compared to poly-Qs. Third, D/E and R/K repeats were similar to poly-As and lower than poly-Qs for within-repeat insertion and deletion MBP values, however, they did not show repeat-associated alteration in silent or missense mutations. Fourth, uniquely, R/Ks had the highest frameshift MBPs of all the examined repeats, and D/Es, also uniquely, showed elevated stop codon gains.

Changes in repeat length markedly contributed to the repeat-specific MBP variation in several cases. First, insertions increased in poly-Q and D/E repeats and then fell after a certain repeat length. In contrast, deletions increased with repeat length in poly-A and D/E repeats while starting out high and declining in poly-Q repeats. Second, missense mutations strongly declined with repeat length in poly-Q and poly-A repeats.

The overall trend of mutations in poly-Qs to be the least disruptive to the repeat, gleaned from the above results, could also be traced by comparing BLOSUM62 scores of single base substitutions for different repeats ([Figure S5](#)). Although all repeats except R/Ks displayed higher BLOSUM62 score averages compared to the adjacent gene regions, poly-Q substitutions tended to be more conservative compared to other repeats. Longer repeat lengths were associated with more conservation in poly-Qs and declined in D/Es.

Prevalence of insertions and deletions was also significantly higher than in the adjacent non-repeat sequences for all intronic TNRs examined ([Figure S6](#), note that in this case, both the whole and partial repeat unit insertions/deletions are included). Surprisingly, the GAA TNR showed the lowest prevalence of insertions among the four TNRs. Repeat length strongly contributed to both insertions and deletions in (GAA) n and (CAG) n , to insertions only in (GAG) n and deletions only in (GAT) n . On the other hand, the behavior of single base substitutions distinguished the expansion-associated TNRs (GAA, CAG) from those not known to expand (GAT, GAG). The CAG TNR was the only TNR with a suppressed prevalence of single base substitutions. The GAA TNR had an elevated prevalence of substitutions.

Collectively, the results demonstrate that insertion/deletion instability of repeat sequences is elevated over the cognate non-repeat sequences whether they are known to undergo pathologic expansions or not.

Table 1. Lengths of Poly-A, poly-Q, D/E, and R/K repeats contribute to their mutability

Ancova (1-way), $p < 0.05$

Repeat type	Variables	Mutation type					
		silent	missense	insertion	deletion	frameshift	stop gained
poly-A	Regions	1.54E-06	1.03E-09	6.93E-11	1.37E-15	0.487	1.1E-21
poly-A	Min repeat length	0.106	4.47E-11	0.194	0.000638	0.809	1.65E-06
poly-Q	Regions	4.82E-39	9.89E-29	1.42E-37	1.06E-57	9.65E-07	0.0643
poly-Q	Min repeat length	0.0264	3.44E-15	0.000416	0.00522	2.44E-19	3.74E-14
D/E	Regions	0.00165	0.015	7.09E-12	3.66E-41	0.00416	1.49E-13
D/E	Min repeat length	0.0397	0.00189	3.71E-08	1.07E-05	0.000737	0.119
R/K	Regions	0.0605	0.0291	3.78E-10	1.53E-10	1.21E-10	0.997
R/K	Min repeat length	0.000009	0.000014	0.589	0.0125	0.522	0.000016

Multiple Comparison of Means, Tukey HSD, FWER = 0.01

Repeat type	Region1	Region2	Mutation type					
			silent	missense	insertion	deletion	frameshift	stop gained
poly-A	Coding	Exons	0.538	0.828	0.975	0.782	0.759	0.0003
poly-A	Coding	Repeat	3.57E-06	0.0002	9.96E-10	1E-12	0.867	6.44E-15
poly-A	Exons	Repeat	0.0001	0.0009	1.92E-09	6.44E-12	0.448	5.48E-13
poly-Q	Coding	Exons	0.59	0.0259	0.403	0.0395	0.996	0.734
poly-Q	Coding	Repeat	<4.44E-16	<4.44E-16	<4.44E-16	<4.44E-16	0.0123	0.643
poly-Q	Exons	Repeat	<4.44E-16	<4.44E-16	<4.44E-16	<4.44E-16	0.0098	0.232
D/E	Coding	Exons	0.0399	0.661	0.956	0.0623	0.539	0.991
D/E	Coding	Repeat	0.0018	0.0199	6.1E-08	<4.44E-16	0.0064	1.45E-11
D/E	Exons	Repeat	0.528	0.15	2.05E-07	<4.44E-16	0.0989	2.58E-11
R/K	Coding	Exons	0.93	1	0.297	0.668	0.239	1
R/K	Coding	Repeat	0.358	0.311	<4.44E-16	<4.44E-16	<4.44E-16	1
R/K	Exons	Repeat	0.56	0.322	<4.44E-16	<4.44E-16	<4.44E-16	0.999

Listed are the p values for the differences between the mutations per base pair (MBP) values of repeats, repeat-containing exons, and all exons (collectively, "Regions"), with minimum repeat length as a covariate.

Furthermore, despite their lengths and recent evolutionary expansion, both of which are on par with poly-Q repeats, the instability of D/E repeat-coding TNRs in exons and the same TNRs in introns is quantitatively and qualitatively different from poly-Qs and their TNR. And lastly, patterns of repeat-supporting versus repeat-disrupting mutations, e. g. silent substitutions versus frameshifts, were very different among the examined TNRs (see [Table S3](#) and [discussion](#)).

Long D/E repeats are prevalent in proteins with nuclear functions

The list of D/E repeat proteins includes many chromatin regulators and transcription factors, of which ATAD2 is a perfect example. To quantify the prevalence of nuclear function proteins among D/E repeat proteins, we determined which GO IDs (retrieved from STRING DB) were enriched two-fold or more among these proteins ([Figure 6](#), these enrichments were highly significant with Fisher exact p values <0.001, unless marked otherwise). For D/E repeats, the enriched GO IDs corresponded to nuclear functions in DNA and RNA metabolism. Poly-Q and poly-A proteins followed the same overall tendency though some distinctions were notable. For example, D/E proteins were absent from categories centered on DNA-binding and were the only group enriched in nucleolar and chromatin categories ([Figure 6B](#)).

Querying the Protein Atlas database of IF *in situ* and IHC-based protein localization data also revealed that proteins with D/E repeats at or above 12 residues were clearly enriched in the nucleus vs. the cytoplasm ([Figure 7A](#)). Here, categorical values of low, mid, and high confidence of localization to a particular cellular compartment obtained from the Protein Atlas were converted to numerical values of 0, 0.5, and 1 to derive

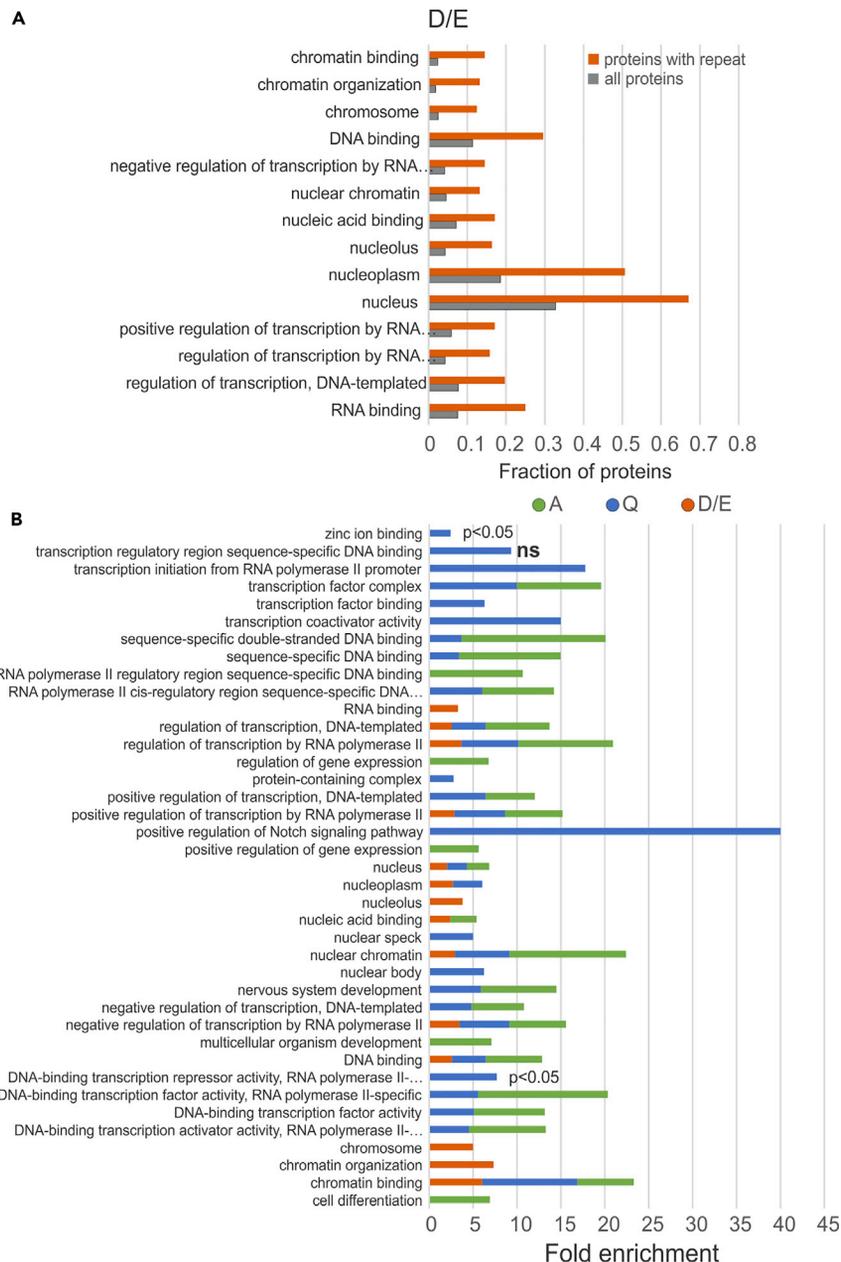


Figure 6. The GO categories enriched for D/E, poly-Q, or poly-A repeat-containing proteins

(A) The GO categories that were enriched more than 2-fold for D/E repeat-containing proteins were retrieved from STRING DB and plotted.

(B) Fold enrichment values for the GO categories in which proteins with the indicated repeats were enriched more than 2-fold. All enrichment values are statistically significant at $p \leq 0.01$ or less in Fisher exact tests, except where indicated otherwise in the graph; ns, not significant at $p \leq 0.05$.

protein group averages. Using these values, we then asked how subcellular localization of repeat-containing proteins changed with their repeat length, by plotting localization confidence averages for D/E, poly-Q, poly-A, and R/K proteins as a function of minimal repeat length. Figure 7B shows that for D/E, poly-Q, and poly-A proteins, repeat length significantly positively correlated with nuclear localization, with Pearson coefficients for all three groups measuring above 0.9 and p values at or below 1.5E-05. In contrast, repeat length did not correlate with the likelihood of non-nuclear localization of these groups of proteins or in some cases showed a negative correlation. R/K proteins were more likely to be nuclear than cytoplasmic,

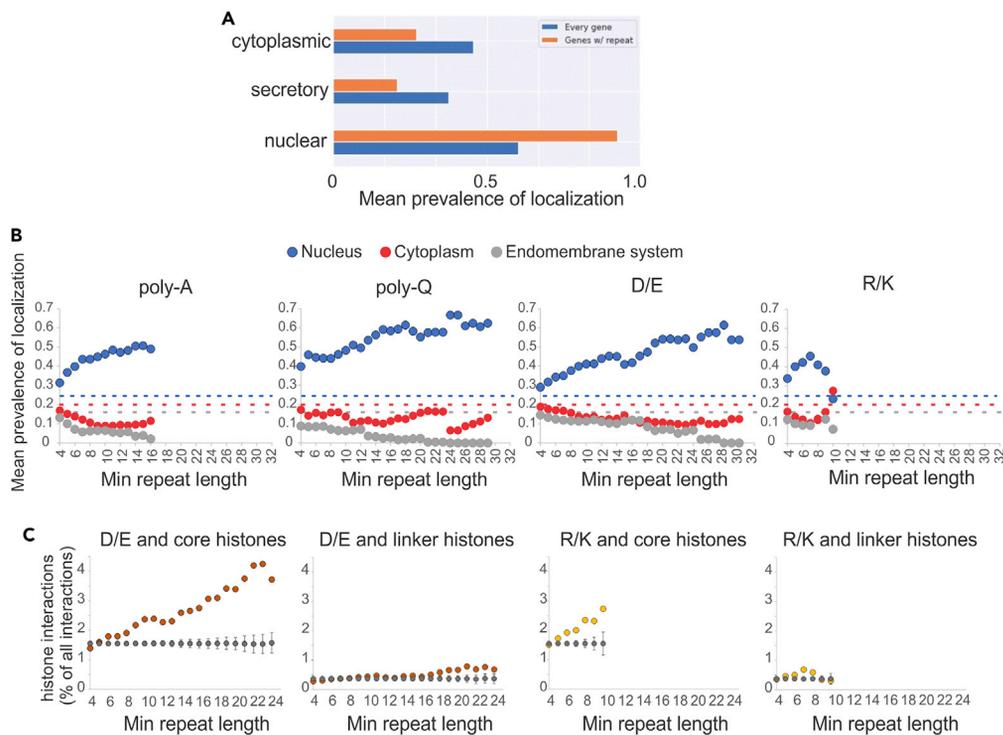


Figure 7. Cellular localization and interactions of D/E and other repeat proteins

(A) Predominant localization of the proteins containing D/E repeats of 12 or more residues compared to all proteins. Localization calls, expressed as low, medium, and high confidence of localization to a given subcellular compartment, were converted to numerical values of 0, 0.5, and 1 for the purpose of averaging across multiple proteins. The averaged values are plotted as Mean localization prevalence on X-axis.

(B) Mean localization prevalence values for the nuclear, cytoplasmic, or membrane system compartments were derived for poly-A, poly-Q, D/E, and R/K proteins and plotted as a function of repeat length. X-axes represent sets of proteins binned by minimal length, i.e., a minimal number of amino acid residues in the repeat. Dotted lines represent mean prevalences of localization to the respective compartments derived for the whole proteome.

(C) Average percentage of interactions with histones among all protein-protein interactions of D/E and R/K proteins was plotted as a function of minimal repeat length. Note that the minimal repeat length values that yielded fewer than 10 proteins per sample were not included because of the increasing variance in such samples. For the random sample control (grey circles), a sample of interactions equal in number to that of D/E proteins' interactions for each minimal length bin was randomly drawn 1,000 times out of all protein/protein interactions. Percentages of interactions with histones found in each of these draws were fitted to a normal distribution, and μ and σ values of these distributions were plotted as gray circles (μ) with error bars (σ). p values for the differences between random samples and D/E samples derived from this analysis are listed in Table 2.

See also Figures S7 and S8.

however, length of R/K repeats did not correlate with nuclear localization (Pearson $r = -0.39$, $p = 0.38$). The finding that proteins with longer D/E repeats are more likely to be nuclear supports the assumption that these repeats, as poly-Qs and poly-As, are important for the nuclear function of the proteins that carry them.

In some cases, D/E repeats were shown to be involved in association with histones. Examples include D/E regions of the proteins PELP1,⁴⁵ SET,⁴⁶ and DAXX.⁴⁷ To systematically address chromatin function of D/E repeats, we retrieved from BioGRID the physical interactomes of human proteins that contain these repeats. Most up-to-date lists of all histone variants were obtained from.⁴⁸ We then plotted the percentages of interactions of D/E-carrying proteins with histones among all their interactions as a function of D/E repeat length (Figure 7C). Interactions with core histones comprised the majority of histone interactions of D/E proteins. Proteins with longer D/E repeats associated with core histones up to three times more frequently compared to proteins with short repeats (Figure 7C, left panel) and compared to the proteome-wide average, where only 1.53% of interactions were with core histones. Also, for all repeat lengths

Table 2. Human proteins with longer D/E repeats are more likely to interact with core histones

Min repeat length	Core histones with		Linker histones with	
	D/E	R/K	D/E	R/K
4	4.57E-21	ns	2.55E-23	ns
5	ns	4.52E-05	6.58E-05	4.64E-05
6	7.91E-10	7.44E-08	ns	5.19E-04
7	6.97E-07	1.54E-04	ns	5.29E-09
8	3.89E-11	4.47E-07	ns	ns
9	4.86E-21	3.29E-03	ns	ns
10	1.10E-26	1.01E-02	ns	ns
11	5.59E-21		ns	
12	7.64E-14		ns	
13	2.06E-11		ns	
14	2.92E-15		ns	
15	1.43E-13		ns	
16	2.16E-14		ns	
17	1.39E-17		ns	
18	1.92E-15		ns	
19	2.16E-17		ns	
20	1.60E-17		ns	
21	1.01E-15		ns	
22	1.41E-17		ns	
23	9.81E-16		ns	
24	5.06E-09		ns	

Listed are the p values for the prevalence of interactions with histones among D/E or R/K repeat-containing proteins. Ns, not significant, denotes p values > 0.01.

above 6 residues the enrichment for interactions with core histones was highly significant compared to samples of random interactions of the same sample size (Figure 7C, left panel, and Table 2 for p values). In contrast, interactions with linker histones were overall not enriched over random samples at the upper cutoff of 0.01 for p values, though they became somewhat more frequent for proteins with the longest repeats (Figure 7C, center-left panel, Table 2). As a control, histone associations of R/K repeat proteins showed a more modest enrichment (Figure 7C, center-right and right panels). Statistically significant enrichment over random interactions samples was seen only for the R/K repeat length range of 5 to 8 residues with p values < 0.01, and for both core and linker histone interactions. This likely reflects interactions between histones because histone proteins themselves carry R/K repeats.

Of interest, proteins with poly-Es of a specified length exhibited a more consistent enrichment for interactions with core histones than proteins with poly-Ds (Figure S7). Despite its limitation (poly-Ds only reach half the length of poly-Es), this result may suggest that D/E repeats' negative charge is not a sole determinant of interaction with histones. We thus probed the secondary structure properties of poly-D and poly-E peptides using PEP-FOLD3 *de novo* structure prediction application (<https://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD>). Poly-Qs and poly-Ns were included for comparison. The results showed that poly-Es (and poly-Qs) adopted an α -helical conformation whereas poly-Ds (and poly-Ns) were disordered (Figure S8). This highlights the molecular differences between these residues that may affect their conformation within proteins, and suggests that D/E repeats of variable length and D to E ratio and distribution may in fact have different propensity for structuring.

D/E repeat of ATAD2 is important for its association with chromatin and maintains silencing of pericentromeric constitutive heterochromatin

ATAD2 has a 28 aa-long D/E repeat that consists of a stretch of Ds flanked by mixed Ds and Es, and the histones H2, H3 and H1 are listed among its interactors in BioGRID. To test if the D/E repeat of ATAD2

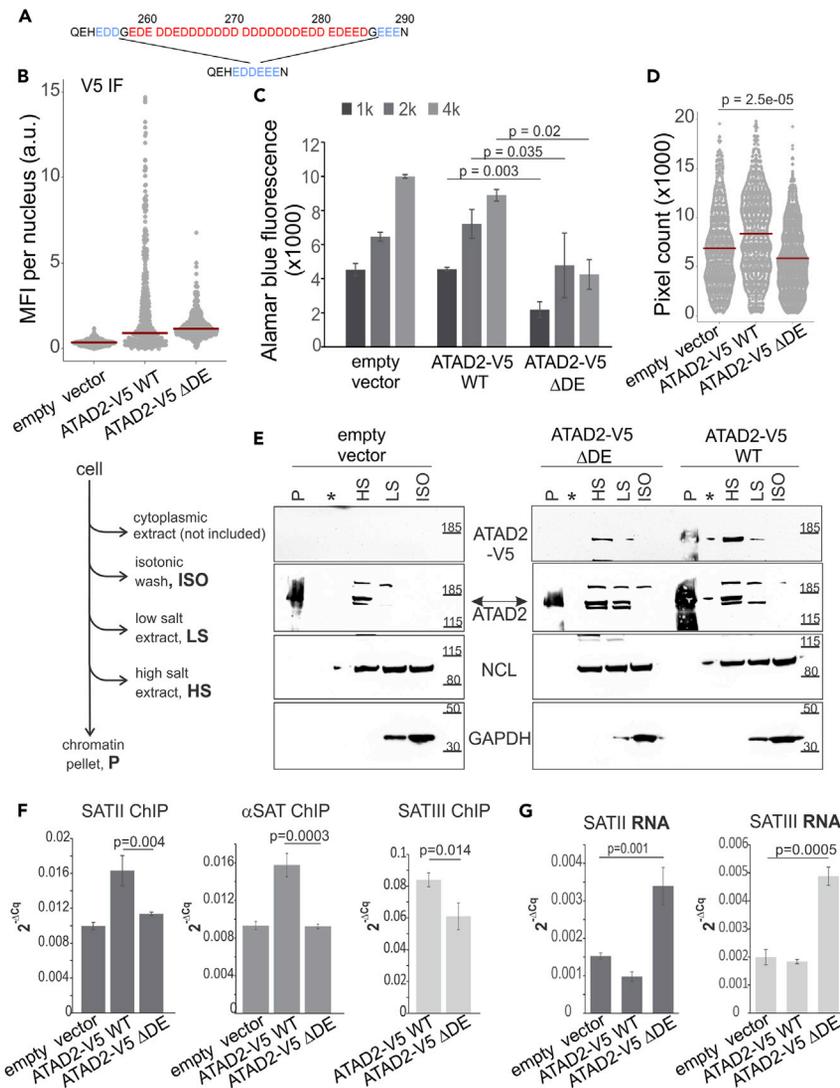


Figure 8. Chromatin-binding phenotypes of the D/E repeat deletion in the ATAD2 protein

All experiments were performed in GM639 SV40-transformed fibroblasts. An empty vector, or the vectors expressing the wild type or D/E deletion mutant ATAD2, C-terminally tagged with a V5 epitope, were stably integrated into these cells via lentiviral transduction.

(A) A schematic of the D/E repeat deletion generated in ATAD2-V5. The 28 residues (red) comprising the uninterrupted run of Ds and Es, and the two flanking G residues were removed, creating a 6 amino acid D/E repeat (blue).

(B) Immunofluorescence *in situ* staining of cells using the V5 antibody and a DAPI counterstain. Nuclear V5 signal was quantified as Mean Fluorescence Intensity (MFI) per nucleus in digital microscopy images. Red bars are distribution medians.

(C) Quantitation of Alamar Blue fluorescence. Cells were seeded into 96-well plates at 1000, 2000, and 4000 cells/well in triplicates, grown for 3 days, and subjected to the Alamar Blue assay for growth/viability. Error bars are standard deviations. p values are calculated in paired one-tailed t-tests.

(D) A representative result from two independent experiments performed on GM639 with the indicated transgenes. Nuclear area measurements were done on digital images of cells stained with DAPI *in situ*. Red bars are medians and the p value was determined in a KS test.

(E) A schematic of fractionation and Western blots of cells expressing the empty vector or the indicated transgenes. Asterisks denote blank lanes. Samples were resolved on 4–12% Bis-Tris gels, and approximate positions and molecular weights (kDa) of the protein ladder are shown on the right of the gels.

(F) qPCR measurements on ChIPs carried out with the V5 tag antibody and chromatin extracts from the indicated cells. Each pulldown and input was analyzed in triplicates. Pulldowns were normalized to inputs using the ΔCq method. Error

Figure 8. Continued

bars are standard deviations. p values were calculated in paired t-tests performed on Δ Cq values. The results represent three (left and center panels) and two (right panel) independent experiments. (G) qPCR measurements on reverse-transcribed RNA preps isolated from the indicated cells. SATII (left panel) or SATIII (right panel) Cq values were normalized to GAPDH Cq values using the Δ Cq method. p values were calculated in paired t-tests performed on Δ Cq values. Error bars are standard deviations. The results represent three (left panel) and two (right panel) independent experiments. See also [Figure S9](#).

is involved in association with chromatin, we deleted the entire uninterrupted repeat from D256 to E287 ([Figure 8A](#)). The flanking G residues were also removed, generating a 6 aa-long D/E stretch. The mutant and wild type ATAD2 with C-terminal V5 tags were stably expressed in SV40-transformed human fibroblasts that we previously used to study ATAD2¹⁸ ([Figure 8](#)). These cells express a high level of endogenous ATAD2.¹⁸ By western blotting, ATAD2-V5 Δ DE was expressed at a modestly higher level than the wild type ATAD2-V5 ([Figure S9A](#)). Immunofluorescence *in situ* (IF, [Figure 8B](#)) showed that Δ DE and wild type proteins were nuclear-localized and confirmed their overall expression levels, albeit on a single-cell basis the wild type ATAD2 expression was more heterogeneous. Remarkably, despite the presence of the endogenous ATAD2, ectopic expression of ATAD2-V5 Δ DE markedly slowed cell population growth when compared to empty vector control or to the wild type ATAD2-V5 ([Figure 8C](#)). ATAD2-V5 Δ DE cells also had decreased nuclear areas compared to the other two cell lines ([Figure 8D](#)). These phenotypes suggest that the mutant exerts a dominant negative effect on the cells.

To address if association of ATAD2 with chromatin was altered by the D/E repeat deletion, we followed the multistep nuclear extraction procedure described in Federation et al.²² ([Figure 8E](#)). Probing with the V5 antibody revealed that ATAD2-V5 was predominantly found in the high salt extract and pellet fractions of cells, whereas ATAD2-V5 Δ DE was mostly in the high salt extract ([Figure 8E](#)). We next used the ATAD2 antibody raised against the N-terminal fragment of ATAD2 (amino acids 1 to 150), which recognizes the endogenous ATAD2,¹⁸ ([Figures S9B](#) and [S9C](#)) and should also recognize ATAD2-V5 and its D/E deletion mutant. Probing with the ATAD2 antibody produced an overall higher signal and revealed the presence of ATAD2 in the low salt extract of cells carrying ATAD2-V5 Δ DE (see [Figures S5B](#) and [S5C](#) for more detail on the antibody specificity). Together, these results are consistent with a reduced affinity of ATAD2-V5 Δ DE for chromatin. To note, the pellet and high salt fractions, to which ATAD2 fractionated above, were shown to be enriched for heterochromatic and structural chromatin proteins in the original study.²²

We next used ChIP with the V5 antibody to examine ATAD2 association with chromatin more quantitatively. We probed for association of ATAD2 with heterochromatic repetitive DNA based on the ATAD2 fractionation pattern (see above) and because the high copy number of repetitive DNA may facilitate detection of ATAD2. The wild type and DE mutant ATAD2-V5 occupancy on SATII and α SAT repeats that comprise pericentromere (SATII and α SAT) and centromere (α SAT), and account for the majority of constitutive heterochromatin, was measured by qPCR in V5 pull-downs against the background of empty vector cells that express no V5 tag ([Figure 8F](#)). Above-background levels of SATII and α SAT sequences were clearly associated with the wild type ATAD2-V5. In contrast, ATAD2-V5 Δ DE pull-downs yielded virtually no SATII and α SAT DNA above the background, suggesting a markedly reduced association of the mutant ATAD2 with α SAT and SATII sequences. Association of ATAD2-V5 Δ DE with another pericentromeric repeat, SATIII, was also reduced compared to the wild type ATAD2-V5 ([Figure 8F](#), right panel). To note, chromatin association defect of ATAD2-V5 Δ DE may not necessarily be limited to pericentromeric heterochromatin. Indeed, ATAD2-V5 Δ DE occupancy was also reduced, albeit not eliminated, on Alu repeats, which are not heterochromatic⁴⁹ ([Figure S5D](#)).

To determine if the reduced occupancy of ATAD2-V5 Δ DE on constitutive heterochromatin is functionally significant, we asked if transcription from these sites was altered in cells expressing ATAD2-V5 Δ DE. Transcripts from pericentromeric satellites are detectable in human cells and their levels can be modulated by heterochromatin defects or by certain biological perturbations.^{50,51} We were able to detect SATII and SATIII transcripts by RT-qPCR in RNA preparations subjected to two rounds of DNA elimination ([Figure S9E](#)). Remarkably, cells expressing ATAD2-V5 Δ DE had markedly elevated levels of SATII and SATIII transcripts compared to cells with the wild type ATAD2 or the empty vector control ([Figure 8G](#)). This suggests that reduction of binding of ATAD2-V5 Δ DE to the satellite repeat heterochromatin is associated with a disruption of transcriptional silencing despite the presence of the endogenous ATAD2. Lastly, depletion

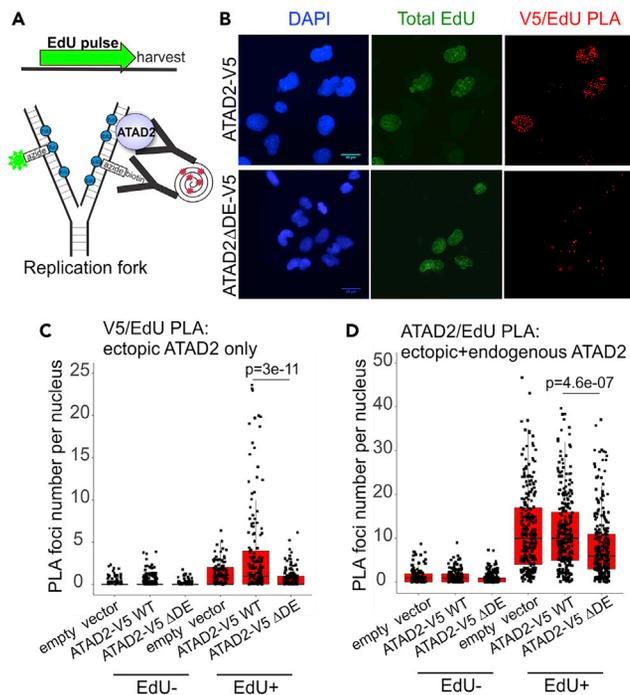


Figure 9. D/E repeat deletion reduces binding of ATAD2 to nascent chromatin

(A) A schematic of the Proximity Ligation Assay (PLA) for the association of ATAD2 with EdU-labeled, newly-replicated chromatin.
 (B) Microscopy images of DAPI (blue), EdU-Alexa488 (green), and V5/EdU PLA (red) channel fluorescence in the indicated cells. Scale (left panels) is 20 μ m.
 (C) Quantitation of V5/EdU PLA foci numbers per nucleus in the indicated cells.
 (D) Quantitation of ATAD2/EdU PLA foci numbers per nucleus in the indicated cells.
 In (C) and (D) foci numbers in EdU-negative and EdU-positive cells are shown separately. p values were calculated in Wilcoxon tests. (C) and (D) each represent three independent experiments.

of 90% of ATAD2 by siRNA in the same SV40 fibroblast cell line also showed elevated SATII expression compared to cells with control siRNA (Figure S9F). Collectively, the results implicate ATAD2 and its D/E repeat in maintaining proper heterochromatin function and, furthermore, support the idea that ATAD2-V5 Δ DE has a dominant-negative effect on the endogenous ATAD2.

Another way to explore chromatin binding activity of ATAD2 is to measure ATAD2 association with newly-replicated chromatin, as previously documented by us¹⁸ and others.¹⁷ We used Proximity Ligation Assay (PLA) to detect ATAD2 on nascent DNA. In this assay, a fluorescent signal develops only if the assayed counterparts, in this case ATAD2 and EdU-labeled DNA, are in close proximity to each other (approx. 40 nm, Figure 9A). By “clicking” EdU to a biotin-azide spiked with fluorescent azide, we are able to identify EdU-positive cells and verify that the PLA signal is seen in these cells only, thus confirming its specificity. We performed V5/EdU PLA to detect exclusively the ectopic wild type and D/E mutant ATAD2 (Figure 9B). ATAD2-V5 Δ DE displayed markedly reduced association with nascent, EdU-labeled DNA compared to ATAD2-V5 (Figure 9C). This low PLA signal could conceivably be explained by the fact that expression level of ATAD2-V5 Δ DE is never as high as that seen in the top quartile of the wild type ATAD2-V5 expressing cells (Figure 8B). However, PLA with the ATAD2 antibody showed that ATAD2-V5 Δ DE expression also reduced the ATAD2/EdU signal. This suggests that the endogenous wild type ATAD2 was also less able to associate with EdU-labeled DNA in the presence of the D/E mutant (Figure 9D), and therefore that the expression level of the D/E mutant was sufficient to interfere with the endogenous ATAD2. Thus, the data are consistent with the notion that D/E repeat deletion compromises the ability of ATAD2 to associate with nascent chromatin/DNA in a dominant-negative manner. Taken together, the data of this section argue that the D/E repeat of ATAD2 is required for its association with nascent and mature chromatin as well as with at least a subset of constitutive heterochromatin, where ATAD2 contributes to transcriptional silencing.

DISCUSSION

Our analysis provides an overview and an information resource on the human proteins that contain D/E repeats. The computational tool we have generated can be applied to other user-specific protein libraries and also to the databases other than used in this study (provided, a database-specific link is first established for the latter). Further, we highlighted several general features of D/E repeats as a group and established their function in the protein ATAD2, which carries one of the longest uninterrupted runs of D/E residues in the human proteome.

D/E repeats are as widespread as poly-Q repeats, which are well-known for their disease-associated expansions,³⁸ achieve comparable lengths to poly-Qs, and exceed poly-A and R/K repeats. D/E repeats' lengths have undergone a recent evolutionary expansion on par with poly-Q repeats. These and other observations raised a question of whether there should exist as yet undiscovered D/E expansion diseases or whether D/E repeat expansions do not occur or are selected against.

One approach to this question is to frame it in a deliberately agnostic fashion with regard to the potential mutation mechanism(s) involved, and ask how D/E repeats compare to poly-Qs and other repeats in terms of their genetic instability in tumors. Analysis of the PCAWG tumor mutation data revealed significantly elevated levels of in-frame insertions and deletions in all four examined repeats (poly-Q, poly-A, D/E, and R/K), though the poly-Qs had the highest values among all. Nevertheless, the levels of in-frame insertions and deletions in D/E repeats were more similar to poly-A repeats, for which the instances of expansion are known but are less common, and even to R/K repeats, for which expansions are not known.

One curious feature observed for poly-Q and D/E repeats was that the occurrence of in-frame insertions increased with repeat length up to a point and then declined. It is uncertain however, if this reflects repeat biology. Insertions into the longest of the repeats may be underrepresented in the data because the corresponding reads were not mappable or did not span the expanded region. Detection of large repeat expansions by next generation sequencing is a recognized challenge.⁵² Also, according to the findings of inverse correlation between cancer susceptibility and (CAG)_n length,⁵³ expanded (CAG)_n repeats may be specifically underrepresented in cancers. On the other hand, the level of deletions remained high and even somewhat declined with length in poly-Qs, whereas in all other repeats it increased with length. This may suggest an expansion-limiting pressure at work on all the examined repeats except poly-Qs, which is in alignment with the fact that these repeats hold a record of pathologic expansions in multiple different proteins.

D/Es differed from both of the expandable repeats, poly-Qs and poly-As, in two notable ways. First, unlike poly-Qs and poly-As, they exhibited no suppression of missense mutations. Second, only D/E repeats showed elevated levels of premature terminations. In summary, if mutations can be broadly categorized into repeat-maintaining and repeat-disrupting by whether or not they support or disrupt the purity and length of the repeat, the two factors that are thought to be crucial for a repeat's potential to expand,^{38,54} then it follows that D/Es and R/Ks, unlike poly-Qs and poly-As, lacked the repeat-preserving mutation trends and showed repeat-disrupting trends (Table S3). These mutation trends could be observed both in exonic and intronic instances of D/E-coding TNRs, which likely are subjected to very different selective pressures. To conclude, despite the similarities between poly-Qs and D/Es in terms of their length and recent evolutionary expansion, the results of the mutation survey may suggest that D/E repeat expansions could in fact be non-existent or extremely rare, whereas small insertion/deletion mutagenesis is likely common and may have physiologic significance.

Although structural properties of D/E repeats were beyond the scope of this work, a couple of observations are worth discussing in the context of the repeats' prevalence and function. When compared as separate subsets, poly-E and poly-D SAARs in the human proteome appeared to follow the tendency of poly-Q and poly-N SAARs in that the maximum attainable lengths of both poly-Es and poly-Qs exceeded poly-D and poly-Ns (by two-fold for poly-Es and five-fold for poly-Qs). Poly-Ds and poly-Ns of any length are also found in far fewer proteins than, respectively, poly-Es and poly-Qs.^{55,56} Poly-Ns have been highlighted before as a unique example of the SAAR that is longer and more prevalent in invertebrates than vertebrates,^{55,57,58} and to our knowledge, this underrepresentation of poly-Ns in vertebrates remains unexplained. Nevertheless, modeling of all four of these SAARs as isolated peptides highlights the fact that poly-Qs and poly-Es are different from, respectively, poly-Ns and poly-Ds: they are far likelier to adopt alpha-helical structure.

Poly-Ns and poly-Ds, on the other hand, model as disordered, with poly-Ns adopting a more compact shape than poly-Ds. Indeed, the alpha-helical propensity of poly-E⁵⁹ and its increase over poly-D⁶⁰ have been noted. Poly-Qs can be alpha-helical,⁶⁰ though their structures are influenced by the flanking sequences.^{61,62} Although more research is clearly needed, it is possible to suggest from the above considerations that, rather than being ubiquitously disordered, D/E repeats likely have a structuring potential that informs their function. By the same token, pure D/E repeats may be structurally different from D/E-rich domains, and not all D/E repeats may be DNA mimics, as the latter category has been defined in terms of a regular structure with the periodicity of a DNA double helix⁶³ and this is likely not the case for all D/E repeats.

To reveal common themes in the cellular functioning of D/E repeats, we analyzed publicly available protein function, localization and interaction data and examined if the latter two features varied with repeat length. D/E repeat proteins were strongly enriched for nuclear localization and the GO categories associated with nuclear functions, prominently including chromatin regulation. Importantly, the likelihood of D/E repeat proteins localizing to the nucleus and interacting with core histones increased with their repeat length. These data point to the direct involvement of D/E repeats in nuclear functions of the repeat-bearing proteins.

Consistent with the above trends, we found that the 28 amino acid-long D/E stretch in the human chromatin regulator ATAD2 critically participates in the association of ATAD2 with chromatin, including the pericentromeric constitutive heterochromatin (PCH). We propose that association with PCH, which is hypoacetylated, may be more strongly dependent on D/E repeat because here ATAD2 cannot rely on its bromodomain, which binds to acetylated histones.^{14,23,24} To note, the D/E deletion that we generated in ATAD2 essentially reduces its D/E stretch from the exceptional length of 28 residues to merely 6 residues, a length that in other D/E proteins corresponds to the lowest likelihood of nuclear localization and histone interaction. It is noteworthy that this mutation had a dramatic, dominant-negative phenotype. When expressed in wild type cells that have endogenous ATAD2, our D/E mutant inhibited population growth, reduced binding of the wild type ATAD2 to newly-replicated chromatin, and increased PCH transcript level while also being defective in binding to nascent chromatin and PCH. The *S. pombe* homolog of ATAD2, Abo1, was shown to form multimers,⁶⁴ and, if this is the case for human ATAD2, it is possible that ATAD2-V5ΔDE may interfere with the endogenous wild type ATAD2 by sequestering it into inactive complexes. Alternatively, the D/E mutant can titrate the proteins that cooperate or interact with ATAD2, for example, the FACT complex.¹⁶

Our finding that disruption of ATAD2 association with PCH by D/E repeat deletion correlates with an increase in PCH transcription, while is agreement with the heterochromatic role of the *S. pombe* ATAD2 homolog Abo1,^{19,20} is surprising since mammalian ATAD2 has been described as a transcriptional activator and (in mouse ES cells) a global transcriptional facilitator that binds acetylated histones, supports open chromatin, and enables histone turnover at active genes by countering histone loading by HIRA and FACT.^{1,15,16} Potential differences between mouse and human ATAD2 and the cell types used can be responsible for the above paradox. It is also possible to reconcile all reported activities of ATAD2 by considering its role in histone turnover. Histone turnover may contribute to maintaining silenced heterochromatin, and, generally, a lower rate of histone turnover is thought to support heterochromatic silencing.⁴⁷ Incidentally, both facilitation of the turnover, by mouse ATAD2 in ES cells,¹⁵ and its suppression, by human ATAD2 in cancer cells,³ have been reported. We propose that ATAD2 may suppress histone turnover in PCH and facilitate it in euchromatin in a manner dependent on local regulatory cues. One precedent to this dual role is the *S. pombe* FACT.^{65,66} Further, participation in the maintenance of PCH is not mutually exclusive with the presence of ATAD2 on nascent chromatin throughout the genome, where it can contribute to the reestablishment of local chromatin environment, whether at PCH or euchromatic loci.

The proposed PCH maintenance activity of ATAD2 may also be advantageous to the cancer cells in which ATAD2 is overexpressed. As PCH needs re-establishment after each round of replication, rapidly dividing cancer cells may require extra means to maintain it. In fact, elevated satellite repeat expression was noted for epithelial cancers.^{67–69} Excessive satellite repeat expression can interfere with pericentromere function and cause chromosomal and genomic instability,^{70–73} and overexpression of ATAD2 may mitigate it. Overall, our focus on D/E repeats, the extent to which their length is a correlate of their function across the proteome, and their role in ATAD2 in particular, has opened routes for further mechanistic dissection of ATAD2

roles in promoting cell proliferation. Understanding of the mechanism of ATAD2- Δ DE interference with the wild type ATAD2 and a further analysis of how the D/E repeat length affects ATAD2 affinity for chromatin, may hold future significance for cancer therapy.

Limitations of the study

Large expansions of TNRs may not be captured in the PCAWG cancer mutations database because of methodological and biological constraints, as discussed. Our focus on uninterrupted repeats and the coding fraction of the genome allowed us to generate meaningful data with limited computing resources, however, a more extensive analysis may be recommended. Proteome-wide averages, e.g., nuclear localization or function, while overall informative, may not reveal more subtle, subpopulation-specific tendencies which will require more granular re-inspection of the data.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
 - Lead contact
 - Materials availability
 - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
 - Bioinformatics
 - Alamar Blue cell viability/growth assay
 - Constructs
 - Western blotting
 - Cellular fractionation
 - Chromatin immunoprecipitation (ChIP)
 - RNA isolation and qPCR
 - Proximity Ligation Assay (PLA) and immunofluorescence *in situ* (IF)
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2022.105464>.

ACKNOWLEDGMENTS

We are grateful to Dr. Monnat for encouragement and discussions. This work was supported by the NIH grant R01 GM115482 to J.S.

AUTHOR CONTRIBUTIONS

S.S.: Conceptualization, methodology, software, investigation, formal analysis, visualization, and writing – original draft. P.L.: Investigation. M.P.: Investigation. J.S.: Conceptualization, investigation, formal analysis, visualization, writing – original draft, writing – review and editing, supervision, and funding acquisition.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse and equitable conduct of research.

Received: July 4, 2022

Revised: August 3, 2022

Accepted: October 26, 2022

Published: December 22, 2022

REFERENCES

- Cattaneo, M., Morozumi, Y., Perazza, D., Boussouar, F., Jamshidikia, M., Rousseaux, S., Verdel, A., and Khochbin, S. (2014). Lessons from yeast on emerging roles of the ATAD2 protein family in gene regulation and genome organization. *Mol. Cells* 37, 851–856. <https://doi.org/10.14348/molcells.2014.0258>.
- Ciró, M., Prosperini, E., Quarto, M., Grazini, U., Walfridsson, J., McBlane, F., Nucifero, P., Pacchiana, G., Capra, M., Christensen, J., and Helin, K. (2009). ATAD2 is a novel cofactor for MYC, overexpressed and amplified in aggressive tumors. *Cancer Res.* 69, 8491–8498. <https://doi.org/10.1158/0008-5472.can-09-2131>.
- Caron, C., Lestrat, C., Marsal, S., Escoffier, E., Curtet, S., Virolle, V., Barbry, P., Debernardi, A., Brambilla, C., Capra, M., et al. (2010). Functional characterization of ATAD2 as a new cancer/testis factor and a predictor of poor prognosis in breast and lung cancers. *Oncogene* 29, 5171–5181. <https://doi.org/10.1038/ncr.2010.259>. <https://www.nature.com/articles/ncr2010259#supplementary-information>.
- Kalashnikova, E.V., Revenko, A.S., Gemo, A.T., Andrews, N.P., Tepper, C.G., Zou, J.X., Cardiff, R.D., Borowsky, A.D., and Chen, H.W. (2010). ANCCA/ATAD2 overexpression identifies breast cancer patients with poor prognosis, acting to drive proliferation and survival of triple-negative cells through control of B-Myb and EZH2. *Cancer Res.* 70, 9402–9412. <https://doi.org/10.1158/0008-5472.can-10-1199>.
- Boussouar, F., Jamshidikia, M., Morozumi, Y., Rousseaux, S., and Khochbin, S. (2013). Malignant genome reprogramming by ATAD2. *Biochim. Biophys. Acta* 1829, 1010–1014. <https://doi.org/10.1016/j.bbtagrm.2013.06.003>.
- Raeder, M.B., Birkeland, E., Trovik, J., Krakstad, C., Shehata, S., Schumacher, S., Zack, T.I., Krohn, A., Werner, H.M., Moody, S.E., et al. (2013). Integrated genomic analysis of the 8q24 amplification in endometrial cancers identifies ATAD2 as essential to MYC-dependent cancers. *PLoS One* 8, e54873. <https://doi.org/10.1371/journal.pone.0054873>.
- Zheng, L., Li, T., Zhang, Y., Guo, Y., Yao, J., Dou, L., and Guo, K. (2015). Oncogene ATAD2 promotes cell proliferation, invasion and migration in cervical cancer. *Oncol. Rep.* 33, 2337–2344. <https://doi.org/10.3892/or.2015.3867>.
- Krakstad, C., Tangen, I.L., Hoivik, E.A., Halle, M.K., Berg, A., Werner, H.M., Raeder, M.B., Kusunmano, K., Zou, J.X., Øyan, A.M., et al. (2015). ATAD2 overexpression links to enrichment of B-MYB-translational signatures and development of aggressive endometrial carcinoma. *Oncotarget* 6, 28440–28452. <https://doi.org/10.18632/oncotarget.4955>.
- Han, H.J., Huang, Q.Y., Huang, L.J., Chang, F., and Diao, Q.Z. (2019). Prognostic value of ATPase family, AAA+ domain containing 2 expression in human cancers: a systematic review and meta-analysis. *Medicine* 98, e17180. <https://doi.org/10.1097/md.00000000000017180>.
- Hussain, M., Zhou, Y., Song, Y., Hameed, H.M.A., Jiang, H., Tu, Y., and Zhang, J. (2018). ATAD2 in cancer: a pharmacologically challenging but tractable target. *Expert Opin. Ther. Targets* 22, 85–96. <https://doi.org/10.1080/14728222.2018.1406921>.
- Nayak, A., Dutta, M., and Roychowdhury, A. (2021). Emerging oncogene ATAD2: signaling cascades and therapeutic initiatives. *Life Sci.* 276, 119322. <https://doi.org/10.1016/j.lfs.2021.119322>.
- Zou, J.X., Guo, L., Revenko, A.S., Tepper, C.G., Gemo, A.T., Kung, H.J., and Chen, H.W. (2009). Androgen-induced coactivator ANCCA mediates specific androgen receptor signaling in prostate cancer. *Cancer Res.* 69, 3339–3346. <https://doi.org/10.1158/0008-5472.can-08-3440>.
- Zou, J.X., Revenko, A.S., Li, L.B., Gemo, A.T., and Chen, H.W. (2007). ANCCA, an estrogen-regulated AAA+ ATPase coactivator for ERalpha, is required for coregulator occupancy and chromatin modification. *Proc. Natl. Acad. Sci. USA* 104, 18067–18072. <https://doi.org/10.1073/pnas.0705814104>.
- Revenko, A.S., Kalashnikova, E.V., Gemo, A.T., Zou, J.X., and Chen, H.W. (2010). Chromatin loading of E2F-MLL complex by cancer-associated coregulator ANCCA via reading a specific histone mark. *Mol. Cell Biol.* 30, 5260–5272. <https://doi.org/10.1128/mcb.00484-10>.
- Morozumi, Y., Boussouar, F., Tan, M., Chaikuad, A., Jamshidikia, M., Colak, G., He, H., Nie, L., Petosa, C., de Dieuleveult, M., et al. (2016). Atad2 is a generalist facilitator of chromatin dynamics in embryonic stem cells. *J. Mol. Cell Biol.* 8, 349–362. <https://doi.org/10.1093/jmcb/mjv060>.
- Wang, T., Perazza, D., Boussouar, F., Cattaneo, M., Boudgour, A., Chuffart, F., Barral, S., Vargas, A., Liakopoulou, A., Puthier, D., et al. (2021). ATAD2 controls chromatin-bound HIRA turnover. *Life Sci. Alliance* 4, e202101151. <https://doi.org/10.26508/lsa.202101151>.
- Koo, S.J., Fernández-Montalván, A.E., Badock, V., Ott, C.J., Holton, S.J., von Ahlsen, O., Toedling, J., Vittori, S., Bradner, J.E., and Gorjánác, M. (2016). ATAD2 is an epigenetic reader of newly synthesized histone marks during DNA replication. *Oncotarget* 7, 70323–70335. <https://doi.org/10.18632/oncotarget.11855>.
- Lazarchuk, P., Hernandez-Villanueva, J., Pavlova, M.N., Federation, A., MacCoss, M., and Sidorova, J.M. (2020). Mutual balance of histone deacetylases 1 and 2 and the acetyl reader ATAD2 regulates the level of acetylation of histone H4 on nascent chromatin of human cells. *Mol. Cell Biol.* 40, 004211–e519. <https://doi.org/10.1128/mcb.004211-19>.
- Gal, C., Murton, H.E., Subramanian, L., Whale, A.J., Moore, K.M., Paszkiewicz, K., Codlin, S., Bähler, J., Creamer, K.M., Partridge, J.F., et al. (2016). Abo1, a conserved bromodomain AAA-ATPase, maintains global nucleosome occupancy and organisation. *EMBO Rep.* 17, 79–93. <https://doi.org/10.15252/embr.201540476>.
- Dong, W., Oya, E., Zahedi, Y., Prasad, P., Svensson, J.P., Lennartsson, A., Ekwall, K., and Durand-Dubief, M. (2020). Abo1 is required for the H3K9me2 to H3K9me3 transition in heterochromatin. *Sci. Rep.* 10, 6055. <https://doi.org/10.1038/s41598-020-63209-y>.
- Tackett, A.J., Dilworth, D.J., Davey, M.J., O'Donnell, M., Aitchison, J.D., Rout, M.P., and Chait, B.T. (2005). Proteomic and genomic characterization of chromatin complexes at a boundary. *J. Cell Biol.* 169, 35–47. <https://doi.org/10.1083/jcb.200502104>.
- Federation, A.J., Nandakumar, V., Searle, B.C., Stergachis, A., Wang, H., Pino, L.K., Merrihew, G., Ting, Y.S., Howard, N., Kutayin, T., et al. (2020). Highly parallel quantification and compartment localization of transcription factors and nuclear proteins. *Cell Rep.* 30, 2463–2471.e5. <https://doi.org/10.1016/j.celrep.2020.01.096>.
- Lloyd, J.T., McLaughlin, K., Lubula, M.Y., Gay, J.C., Dest, A., Gao, C., Phillips, M., Tonelli, M., Cornilescu, G., Marunde, M.R., et al. (2020). Structural insights into the recognition of mono- and diacetylated histones by the ATAD2B bromodomain. *J. Med. Chem.* 63, 12799–12813. <https://doi.org/10.1021/acs.jmedchem.0c01178>.
- Evans, C.M., Phillips, M., Malone, K.L., Tonelli, M., Cornilescu, G., Cornilescu, C., Holton, S.J., Gorjánác, M., Wang, L., Carlson, S., et al. (2021). Coordination of di-acetylated histone ligands by the ATAD2 bromodomain. *Int. J. Mol. Sci.* 22, 9128. <https://doi.org/10.3390/ijms22179128>.
- Gradolatto, A., Smart, S.K., Byrum, S., Blair, L.P., Rogers, R.S., Kolar, E.A., Lavender, H., Larson, S.K., Aitchison, J.D., Taverna, S.D., and Tackett, A.J. (2009). A noncanonical bromodomain in the AAA ATPase protein Yta7 directs chromosomal positioning and barrier chromatin activity. *Mol. Cell Biol.* 29, 4604–4611. <https://doi.org/10.1128/mcb.00160-09>.
- Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J., and Gentles, A.J. (2002). Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl. Acad. Sci. USA* 99, 333–338. <https://doi.org/10.1073/pnas.012608599>.
- Chou, C.C., and Wang, A.H.J. (2015). Structural D/E-rich repeats play multiple roles especially in gene regulation through DNA/RNA mimicry. *Mol. Biosyst.* 11, 2144–2151. <https://doi.org/10.1039/c5mb00206k>.
- Lee, C.-H., Shih, Y.-P., Ho, M.-R., and Wang, A.H.-J. (2018). The C-terminal D/E-rich domain of MBD3 is a putative Z-DNA mimic that competes for Zα-DNA-binding activity.

- Nucleic Acids Res. 46, 11806–11821. <https://doi.org/10.1093/nar/gky933>.
29. Wang, H.C., Chou, C.C., Hsu, K.C., Lee, C.H., and Wang, A.H.J. (2019). New paradigm of functional regulation by DNA mimic proteins: recent updates. *IUBMB Life* 71, 539–548. <https://doi.org/10.1002/iub.1992>.
30. Kovtun, I.V., and McMurray, C.T. (2008). Features of trinucleotide repeat instability in vivo. *Cell Res.* 18, 198–213. <https://doi.org/10.1038/cr.2008.5>.
31. Depienne, C., and Mandel, J.-L. (2021). 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* 108, 764–785. <https://doi.org/10.1016/j.ajhg.2021.03.011>.
32. Verstreppe, K.J., Jansen, A., Lewitter, F., and Fink, G.R. (2005). Intragenic tandem repeats generate functional variability. *Nat. Genet.* 37, 986–990. <https://doi.org/10.1038/ng1618>.
33. Gemayel, R., Chavali, S., Pougach, K., Legendre, M., Zhu, B., Boeynaems, S., van der Zande, E., Gevaert, K., Rousseau, F., Schymkowitz, J., et al. (2015). Variable glutamine-rich repeats modulate transcription factor Activity. *Mol. Cell* 59, 615–627. <https://doi.org/10.1016/j.molcel.2015.07.003>.
34. Sonay, T.B., Koletou, M., and Wagner, A. (2015). A survey of tandem repeat instabilities and associated gene expression changes in 35 colorectal cancers. *BMC Genom.* 16, 702. <https://doi.org/10.1186/s12864-015-1902-9>.
35. Persi, E., Prandi, D., Wolf, Y.I., Pozniak, Y., Barnabas, G.D., Levanon, K., Barshack, I., Barbieri, C., Gasperini, P., Beltran, H., et al. (2019). Proteomic and genomic signatures of repeat instability in cancer and adjacent normal tissues. *Proc. Natl. Acad. Sci. USA* 116, 16987–16996. <https://doi.org/10.1073/pnas.1908790116>.
36. Fujino, Y., and Nagai, Y. (2022). The molecular pathogenesis of repeat expansion diseases. *Biochem. Soc. Trans.* 50, 119–134. <https://doi.org/10.1042/bst20200143>.
37. Bunting, E.L., Hamilton, J., and Tabrizi, S.J. (2022). Polyglutamine diseases. *Curr. Opin. Neurobiol.* 72, 39–47. <https://doi.org/10.1016/j.conb.2021.07.001>.
38. Khristich, A.N., and Mirkin, S.M. (2020). On the wrong DNA track: molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.* 295, 4134–4170. <https://doi.org/10.1074/jbc.REV119.007678>.
39. Goldberg, P. (1999). The ups and downs of parting COMPany. Poly-aspartic acid tract instability: a novel expansion-contraction mutation. *Clin. Genet.* 56, 259–261. <https://doi.org/10.1034/j.1399-0004.1999.560402.x>.
40. Torres, B., Orozco, G., García-Lozano, J.R., Oliver, J., Fernández, O., González-Gay, M.A., Balsa, A., García, A., Pascual-Salcedo, D., López-Nevot, M.A., et al. (2007). Asporin repeat polymorphism in rheumatoid arthritis. *Ann. Rheum. Dis.* 66, 118–120. <https://doi.org/10.1136/ard.2006.055426>.
41. Sobhan, M.R., Mehdinejad, M., Jamaladini, M.H., Mazaheri, M., Zare-Shehneh, M., and Neamatzadeh, H. (2017). Association between aspartic acid repeat polymorphism of the asporin gene and risk of knee osteoarthritis: a systematic review and meta-analysis. *Acta Orthop. Traumatol. Turc.* 51, 409–415. <https://doi.org/10.1016/j.aott.2017.08.001>.
42. Kumar, A.S., Sowpati, D.T., and Mishra, R.K. (2016). Single amino acid repeats in the proteome world: structural, functional, and evolutionary insights. *PLoS One* 11, e0166854. <https://doi.org/10.1371/journal.pone.0166854>.
43. International Cancer Genome Consortium, Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabé, R.R., Bhan, M.K., Calvo, F., Eerola, I., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. <https://doi.org/10.1038/nature08987>.
44. Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L.D., and Ferretti, V. (2019). The international cancer genome Consortium data portal. *Nat. Biotechnol.* 37, 367–369. <https://doi.org/10.1038/s41587-019-0055-9>.
45. Choi, Y.B., Ko, J.K., and Shin, J. (2004). The transcriptional corepressor, PELP1, recruits HDAC2 and masks histones using two separate domains. *J. Biol. Chem.* 279, 50930–50941. <https://doi.org/10.1074/jbc.M406831200>.
46. Seo, S.-b., McNamara, P., Heo, S., Turner, A., Lane, W.S., and Chakravarti, D. (2001). Regulation of histone acetylation and transcription by INHAT, a human cellular complex containing the set oncoprotein. *Cell* 104, 119–130. [https://doi.org/10.1016/S0092-8674\(01\)00196-9](https://doi.org/10.1016/S0092-8674(01)00196-9).
47. Drané, P., Ouarrhni, K., Depaux, A., Shuaib, M., and Hamiche, A. (2010). The death-associated protein DAXX is a novel histone chaperone involved in the replication-independent deposition of H3.3. *Genes Dev.* 24, 1253–1265. <https://doi.org/10.1101/gad.566910>.
48. Khare, S.P., Habib, F., Sharma, R., Gadwal, N., Gupta, S., and Galande, S. (2012). H1stome—a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res.* 40, D337–D342. <https://doi.org/10.1093/nar/gkr1125>.
49. Zhang, J., Bellani, M.A., James, R.C., Pokharel, D., Zhang, Y., Reynolds, J.J., McNeer, G.S., Jackson, A.P., Stewart, G.S., and Seidman, M.M. (2020). DONSON and FANCM associate with different replisomes distinguished by replication timing and chromatin domain. *Nat. Commun.* 11, 3951. <https://doi.org/10.1038/s41467-020-17449-1>.
50. Saksouk, N., Simboeck, E., and Déjardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. *Epigenet. Chromatin* 8, 3. <https://doi.org/10.1186/1756-8935-8-3>.
51. Smurova, K., and De Wulf, P. (2018). Centromere and pericentromere transcription: roles and regulation, in sickness and in health. *Front. Genet.* 9, 674. <https://doi.org/10.3389/fgene.2018.00674>.
52. Bahlo, M., Bennett, M., Degorski, P., Tankard, R., Delatycki, M., and Lockhart, P. (2018). Recent advances in the detection of repeat expansions with short-read next-generation sequencing [version 1; peer review: 3 approved]. *F1000Res.* 7. <https://doi.org/10.12688/f1000research.13980.1>.
53. Murmann, A.E., Yu, J., Opal, P., and Peter, M.E. (2018). Trinucleotide repeat expansion diseases, RNAi, and cancer. *Trends Cancer* 4, 684–700. <https://doi.org/10.1016/j.trecan.2018.08.004>.
54. Legendre, M., Pochet, N., Pak, T., and Verstreppe, K.J. (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.* 17, 1787–1796. <https://doi.org/10.1101/gr.6554007>.
55. Kreil, D.P., and Kreil, G. (2000). Asparagine repeats are rare in mammalian proteins. *Trends Biochem. Sci.* 25, 270–271. [https://doi.org/10.1016/S0968-0004\(00\)01594-2](https://doi.org/10.1016/S0968-0004(00)01594-2).
56. Mier, P., Alanis-Lobato, G., and Andrade-Navarro, M.A. (2017). Context characterization of amino acid homorepeats using evolution, position, and order. *Proteins* 85, 709–719. <https://doi.org/10.1002/prot.25250>.
57. Faux, N.G., Bottomley, S.P., Lesk, A.M., Irving, J.A., Morrison, J.R., de la Banda, M.G., and Whistock, J.C. (2005). Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res.* 15, 537–551. <https://doi.org/10.1101/gr.3096505>.
58. Kamel, M., Mier, P., Tari, A., and Andrade-Navarro, M.A. (2019). Repeatability in protein sequences. *J. Struct. Biol.* 208, 86–91. <https://doi.org/10.1016/j.jsb.2019.08.003>.
59. Ogasawara, N., Kasahara, K., Iwai, R., and Takahashi, T. (2018). Unfolding of α -helical 20-residue poly-glutamic acid analyzed by multiple runs of canonical molecular dynamics simulations. *PeerJ* 6, e4769. <https://doi.org/10.7717/peerj.4769>.
60. Gonçalves-Kulik, M., Mier, P., Kastano, K., Cortés, J., Bernadó, P., Schmid, F., and Andrade-Navarro, M.A. (2022). Low complexity induces structure in protein regions predicted as intrinsically disordered. *Biomolecules* 12, 1098. <https://doi.org/10.3390/biom12081098>.
61. Urbanek, A., Popovic, M., Morató, A., Estaña, A., Elena-Real, C.A., Mier, P., Fournet, A., Allemand, F., Delbecq, S., Andrade-Navarro, M.A., et al. (2020). Flanking regions determine the structure of the poly-glutamine in huntingtin through mechanisms common among glutamine-rich human proteins. *Structure* 28, 733–746. <https://doi.org/10.1016/j.str.2020.04.008>.
62. Mier, P., and Andrade-Navarro, M.A. (2021). Between interactions and aggregates: the PolyQ balance. *Genome Biol. Evol.* 13,

- evab246. <https://doi.org/10.1093/gbe/evab246>.
63. Wang, H.C., Ho, C.H., Hsu, K.C., Yang, J.M., and Wang, A.H.J. (2014). DNA mimic proteins: functions, structures, and bioinformatic analysis. *Biochemistry* 53, 2865–2874. <https://doi.org/10.1021/bi5002689>.
64. Cho, C., Jang, J., Kang, Y., Watanabe, H., Uchihashi, T., Kim, S.J., Kato, K., Lee, J.Y., and Song, J.J. (2019). Structural basis of nucleosome assembly by the Abo1 AAA+ ATPase histone chaperone. *Nat. Commun.* 10, 5764. <https://doi.org/10.1038/s41467-019-13743-9>.
65. Murawska, M., and Braun, S. (2022). Chaperoning heterochromatin: new roles of FACT in chromatin silencing. *Trends Genet.* 38, 646–649. <https://doi.org/10.1016/j.tig.2022.02.011>.
66. Hirai, H., Takemata, N., Tamura, M., and Ohta, K. (2022). Facultative heterochromatin formation in rDNA is essential for cell survival during nutritional starvation. *Nucleic Acids Res.* 50, 3727–3744. <https://doi.org/10.1093/nar/gkac175>.
67. Zhu, Q., Pao, G.M., Huynh, A.M., Suh, H., Tonnu, N., Nederlof, P.M., Gage, F.H., and Verma, I.M. (2011). BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature* 477, 179–184. <https://doi.org/10.1038/nature10371>.
68. Ting, D.T., Lipson, D., Paul, S., Brannigan, B.W., Akhavanfard, S., Coffman, E.J., Contino, G., Deshpande, V., Iafrate, A.J., Letovsky, S., et al. (2011). Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* 331, 593–596. <https://doi.org/10.1126/science.1200801>.
69. Bersani, F., Lee, E., Kharchenko, P.V., Xu, A.W., Liu, M., Xega, K., MacKenzie, O.C., Brannigan, B.W., Wittner, B.S., Jung, H., et al. (2015). Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proc. Natl. Acad. Sci. USA* 112, 15148–15153. <https://doi.org/10.1073/pnas.1518008112>.
70. Zeller, P., and Gasser, S.M. (2017). The importance of satellite sequence repression for genome stability. *Cold Spring Harb. Symp. Quant. Biol.* 82, 15–24. <https://doi.org/10.1101/sqb.2017.82.033662>.
71. Black, E.M., and Giunta, S. (2018). Repetitive fragile sites: centromere satellite DNA as a source of genome instability in human diseases. *Genes* 9, E615. <https://doi.org/10.3390/genes9120615>.
72. Zhu, Q., Hoong, N., Aslanian, A., Hara, T., Benner, C., Heinz, S., Miga, K.H., Ke, E., Verma, S., Soroczynski, J., et al. (2018). Heterochromatin-encoded satellite RNAs induce breast cancer. *Mol. Cell* 70, 842–853.e7. <https://doi.org/10.1016/j.molcel.2018.04.023>.
73. Francastel, C., and Magdinier, F. (2019). DNA methylation in satellite repeats disorders. *Essays Biochem.* 63, 757–771. <https://doi.org/10.1042/ebc20190028>.
74. Gish, W.R., and Botchan, M.R. (1987). Simian virus 40-transformed human cells that express large T antigens defective for viral DNA replication. *J. Virol.* 61, 2864–2876.
75. Shen, J.Z., Qiu, Z., Wu, Q., Finlay, D., Garcia, G., Sun, D., Rantala, J., Barshop, W., Hope, J.L., Gimple, R.C., et al. (2021). FBXO44 promotes DNA replication-coupled repetitive element silencing in cancer cells. *Cell* 184, 352–369.e23. <https://doi.org/10.1016/j.cell.2020.11.042>.
76. Ding, L., Kim, H.J., Wang, Q., Kearns, M., Jiang, T., Ohlson, C.E., Li, B.B., Xie, S., Liu, J.F., Stover, E.H., et al. (2018). PARP inhibition elicits STING-dependent antitumor immunity in brca1-deficient ovarian cancer. *Cell Rep.* 25, 2972–2980.e5. <https://doi.org/10.1016/j.celrep.2018.11.054>.
77. Edelstein, A., Amodaj, N., Hoover, K., Vale, R., and Stuurman, N. (2010). Computer control of microscopes using µManager. *Curr. Protoc. Mol. Biol.* 92, 14–20. <https://doi.org/10.1002/0471142727.mb1420s92>.
78. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods* 9, 676–682. <https://doi.org/10.1038/nmeth.2019>.
79. Carpenter, A., Jones, T., Lamprecht, M., Clarke, C., Kang, I., Friman, O., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology* 7 (R100). <https://doi.org/10.1186/gb-2006-7-10-r100>.
80. Swanson, C., Saintigny, Y., Emond, M.J., Monnat, R.J., Jr., and Raymond, J. (2004). The Werner syndrome protein has separable recombination and survival functions. *DNA Repair* 3, 475–482.
81. Sidorova, J.M., Li, N., Folch, A., and Monnat Jr, R.J. (2008). The RecQ helicase WRN is required for normal replication fork progression after DNA damage or replication fork arrest. *Cell Cycle* 7, 796–807.
82. Juarez, E., Chambwe, N., Tang, W., Mitchell, A.D., Owen, N., Kumari, A., Monnat, R.J., Jr., and McCullough, A.K. (2018). An RNAi screen in human cell lines reveals conserved DNA damage repair pathways that mitigate formaldehyde sensitivity. *DNA Repair* 72, 1–9. <https://doi.org/10.1016/j.dnarep.2018.10.002>.
83. Lazarchuk, P., Roy, S., Schlacher, K., and Sidorova, J. (2019). Detection and quantitation of acetylated histones on replicating DNA using in situ proximity ligation assay and click-it Chemistry. *Methods Mol. Biol.* 1983, 29–45. https://doi.org/10.1007/978-1-4939-9434-2_3.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
mouse α -biotin	Vector Laboratories	RRID:AB_2336110
rabbit α -ATAD2	Proteintech	RRID:AB_2879352
mouse α -NCL Cat. No. 396400	Life Technologies	RRID:AB_1501816
mouse α -NCL Cat. No. 87792	CST	RRID:AB_2800106
mouse α -V5 tag Cat. No. 80076	CST	RRID:AB_2920661
rabbit α -V5 tag Cat. No. 13202	CST	RRID:AB_2687461
mouse α -GAPDH Cat. No. 97166	CST	RRID:AB_2756824
Alexa Fluor 594 goat α -rabbit antibody	Invitrogen	RRID:AB_2534079
HRP-linked α -mouse antibody	CST	RRID:AB_330924
HRP-linked α -rabbit antibody	CST	RRID:AB_2099233
Chemicals, peptides, and recombinant proteins		
resazurin sodium salt	Acros Organics	CAS 62758-13-8
EdU	Click Chemistry Tools	CAS 61135-33-9
Blasticidin HCl	Thermo Scientific	CAS 3513-03-9
Pierce Protease inhibitor tablets	Thermo Scientific	Cat. No. A32965
Proteinase K recombinant	Invitrogen	Cat. No. 25530-015
RNAse A	Qiagen	Cat. No. 19101
mouse a-V5 tag antibody-conjugated magnetic beads	CST	Cat. No. 53475
Biotin azide	Click Chemistry Tools	Cat. No. 1265
AZDye 488 azide	Click Chemistry Tools	Cat. No. 1275
Bovine serum albumin (BSA)	Millipore-Sigma	Cat. No A9647
Lipofectamine RNAiMAX Invitrogen	Thermo Scientific	Cat. No. 13778150
Critical commercial assays		
DuoLink red detection kit	Millipore-Sigma	Cat. No. DUO92008
DuoLink anti-mouse antibody		DUO92002
Duo-Link anti-rabbit antibody		
DUOLINK DAPI mounting medium	Millipore-Sigma	Cat. No. DUO82040
High Capacity cDNA Reverse Transcription kit	Applied Biosystems	Cat.No.43-688-14
RNeasy Plus RNA isolation kit	Qiagen	Cat. No.74134
DNA-free kit	Thermo Fisher/Invitrogen	Cat. No. AM1906
iTaq Universal SYBR Green supermix	Bio-Rad	Cat.No.1725121
Pierce™ ECL Western Blotting Substrate	Thermo Fisher	Cat. No. 32209
Q5 site directed mutagenesis kit	NEB	Cat. No. E05545
Gibson assembly cloning kit	NEB	Cat. No. E5510S
mouse a-V5 tag antibody-conjugated magnetic beads	CST	Cat. No. 53475
NucleoSpin Gel and PCR clean up	Macherey-Nagel	Cat.No.740609.50
Buffer NTB	Macherey-Nagel	Cat.No.740595.150

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
<i>Experimental models: Cell lines</i>		
GM000639	(Gish and Botchan ⁷⁴ , 1987)	Cellosaurus ID CVCL_7299
<i>Oligonucleotides</i>		
5'GAAGAAGAGAATCAGAAGC3' to introduce D/E domain deletion	This work	N/A
5'ATCATCTTCATGTTCTTGG3' to introduce D/E domain deletion	This work	N/A
Alu ChIP 1: 5'AATGGTACGATCTCGGCTCA3'	(Shen et al., ⁷⁵ 2021)	N/A
Alu ChIP 2: 5'TAGCCAGGTGTGGTGAATG3'	(Shen et al., ⁷⁵ 2021)	N/A
SATIII ChIP 1: 5'AATCAACCCGAGTGCAATC NGAATGGAATCG3'	(Shen et al., ⁷⁵ 2021)	N/A
SATIII ChIP 2: 5'TCCATTCCATTCTGTAC TCGG3'	(Shen et al., ⁷⁵ 2021)	N/A
SATII-1: 5'TCATCGAATGGAAATGAAAG GAGTCATCATCT3'	This work	GenBank: X72623.1
SATII-2: 5'CGACCATTGGATGATTGCA GTCAA3'	This work	GenBank: X72623.1
α SAT primer pair	CST	Cat. No. 4486
GAPDH RT qPCR-1: 5'ACAACCTTGGCATT GAA3'	(Ding et al., ⁷⁶ 2018)	N/A
GAPDH RT qPCR-2: 5'GATGCAGGGATGA TGTCTG3'	(Ding et al., ⁷⁶ 2018)	N/A
SATIII RT qPCR: 5'AATCAACCCGAGTG CAATCGAATGGAATCG3'	(Shen et al., ⁷⁵ 2021)	N/A
SATIII RT qPCR-2: 5'TCCATTCCATTCT GTACTCGG3'	(Shen et al., ⁷⁵ 2021)	N/A
Hs_ATAD2_3 siRNA, target sequence 5'AAGGCATTATAAAGATCGAA3'	Qiagen	Cat. No. SI00103138
Hs_ATAD2_5 siRNA, target sequence 5'CCGGATAAAGAAGACGGACA-3'	Qiagen	Cat. No. SI02640232
<i>Recombinant DNA</i>		
GFP-ATAD2-V5 plasmid	Addgene	RRID:Addgene_65370
pLenti EFS T2A BSD ATAD2-V5	This work	N/A
pLenti EFS T2A BSD ATAD2 Δ DE-V5	This work	N/A
<i>Software and algorithms</i>		
RStudio	RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA	https://www.rstudio.com/
MicroManager	(Edelstein et al., ⁷⁷ 2010)	https://micro-manager.org/
FIJI ImageJ	(Schindelin et al., ⁷⁸ 2012)	https://fiji.sc/
CellProfiler	(Carpenter et al., ⁷⁹ 2006)	https://cellprofiler.org/
TriRepSearch	This work	https://doi.org/10.5281/zenodo.7140652
PhyloRepeatSearch	This work	https://doi.org/10.5281/zenodo.7242893

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Julia M. Sidorova (julias@uw.edu).

Materials availability

- Plasmids generated in this study are available on request from the [lead contact](#).

Data and code availability

- Original western blot images have been deposited at Zenodo and are publicly available as of the date of publication. The DOI is <https://doi.org/10.5281/zenodo.7136678>. Microscopy data reported in this article will be shared by the [lead contact](#) on request.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the [key resources table](#).
- This article analyzes the existing, publicly available data. The datasets used for analyses are freely available from ICGC (<https://dcc.icgc.org/>), StringDB (<https://string-db.org/>), Protein Atlas (<https://www.proteinatlas.org/>), BioGRID (<https://thebiogrid.org>), and Ensembl (<https://www.ensembl.org>).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The SV40-transformed human fibroblast GM639⁷⁴ (GM00639, Cellosaurus ID CVCL_7299) has been used by us and others previously^{80,81} and was obtained from the NIGMS Human Genetic Mutant Cell Repository. Exome sequencing of this cell line was reported in.⁸² GM639 was grown in high glucose Dulbecco Modified Minimal Essential Medium (DMEM) with L-glutamine, 10% fetal bovine serum, FBS, (Hyclone) and antibiotics in a humidified 5% CO₂, 37°C incubator. Mycoplasma testing was performed regularly using the UW/FHCRC Cancer Consortium Shared Resource Specimen processing service <https://sharedresources.fredhutch.org/services/mycoplasma-testing>.

METHOD DETAILS

Bioinformatics

TriRepSearch (Figure 1) is a python package for identifying and analyzing repetitive stretches of trinucleotides in protein-coding genes. The search parameters are flexible and robust allowing for the search of non-homogeneous and discontinuous repeats. The user can specify the minimum and maximum length of the repeat, the trinucleotides that are allowed in the repeat (i.e., can enter multiple trinucleotides to search for repeats that have mixed trinucleotide content), the region where the repeat is found (either exonic or intronic), and the number of gaps allowed in the repeated sequence (by default the gap allowance is set to 0).

TriRepSearch searches against all protein-coding genes found in the Ensembl GRCh37 release 105 (<https://grch37.ensembl.org/index.html>). After the search is complete, a dictionary containing all the genes found as well as the genomic coordinates of the individual repeats is generated. Using this dictionary, TriRepSearch queries various online databases such as ICGC (<https://dcc.icgc.org/>), StringDB (<https://string-db.org/>), and Protein Atlas (<https://www.proteinatlas.org/>) through their dedicated inbuilt APIs. Alternatively, TriRepSearch can use local copies of the databases which are hosted on the GitHub or provided by the user as CSVs for offline search or to maintain version cohesion in case of dataset depreciation. Biological data retrieved from these databases are stored as CSVs for the user. In addition, multiple statistical analyses are performed on the retrieved mutational and expression data to determine the tumor allele frequencies in repeat regions, the various types of mutational events that occur in repeat regions, and the overall score of mutability of the repeat regions. Severity of single base substitutions and multi region comparisons of repeats are also calculated and output as CSVs for the user.

TriRepSearch was developed in Python 3.8 (<https://www.python.org/downloads/release/python-380/>) and is deployable as a dedicated python package for all users. The TriRepSearch package is open source and distributed under the MIT license via GitHub (<https://github.com/SidorovaLab/TriRepSearch>). For easy installation TriRepSearch can be installed using the 'PIP' command (instructions on github).

Phylogenetic tree construction

Additional code is required to reconstruct the relation trees from Ensembl phylogenetic JSON data. This was done in python using the in-built urllib module. We then employed a recursive tree traversal algorithm

to extract the relevant relationship data from the given dataset. The code is available on GitHub (<https://github.com/SidorovaLab/PhyloRepeatSearch>).

Protein-protein interactions

Protein-protein interaction data was collected from the most recent release of the full online BioGRID biomedical interaction repository dataset, obtained using their restful API. We then filtered for all physical human protein-protein interactions, and tallied the interactions for and between D/E containing proteins, R/K containing proteins and histone proteins.

Alamar Blue cell viability/growth assay

10 mg/mL stock solution of Resazurin sodium salt (Acros Organics) in PBS was stored at 4°C. For the assays, the stock was diluted to 36 µg/mL in tissue culture media, and added to cells. After incubation for 1.5–2 h at 37°C, fluorescence was measured in a microplate reader using a 540/35 excitation filter and a 590/20 emission filter.

Constructs

GFP-ATAD2-V5 plasmid was a gift from Kyle Miller (Addgene plasmid # 65370; <http://n2t.net/addgene:65370>; RRID:Addgene_65370). A complete D/E domain deletion was introduced into this plasmid using Q5 site directed mutagenesis kit (NEB Cat. No. E0554S) with the primers 5'GAAGAAGAGAATCAG AAGC-3' and 5'ATCATCTTCATGTTCTTGG-3'. Wildtype and mutant ORFs (including the V5 tag but excluding GFP) were then transferred into pLenti EFS T2A BSD backbone using Gibson assembly cloning kit (NEB Cat. No. E5510S). Virus generation from these constructs was done as described⁸¹ by calcium phosphate transfection into human HEK293T with helper plasmids pCMV-dR8.2 dvpr and pCMV-VSVG. GM639 transduced with empty vector or ATAD2-expressing lentiviral constructs were selected with blasticidin for 6 days before analyses. Expression of transgenes was verified in Western blots with mouse α-V5 antibody.

Western blotting

Proteins were visualized on Western blots by ECL (Thermo Scientific) and quantified using FluorChem Imager (Alpha Inotech). For presentation, images were saved in TIFF format, adjusted for brightness/contrast and cropped in Adobe Photoshop, then assembled into figures in CorelDraw. Image brightness/contrast adjustments were made across all lanes of each protein measured.

Cellular fractionation

Fractionation was performed according to²² with minor modifications. Nuclei from 1–5 × 10⁶ cells were isolated by incubation with buffer A (15 mM Tris-Cl, pH 8.0, 15 mM NaCl, 60 mM KCl, 1 mM EDTA, pH 8.0) supplemented with 0.25% NP40 and protein inhibitors, for 15 min on ice followed by pelleting at 400 × g for 7 min at 4°C. Nuclear pellets were extracted with 100–200 µL of buffer A at 4°C with gentle agitation for 5 min and pelleted as above. Supernatants were saved as the isotonic wash fraction. Nuclei were further extracted with 100–200 µL of low salt extraction buffer (10 mM Tris-Cl, pH 8.0, 250 mM NaCl, 0.5 mM EDTA, protease inhibitors) at 4°C with gentle agitation for 30 min and pelleted at 1000 × g at 4°C for 5 min. Supernatants were saved as the low salt fraction. Nuclei were extracted with 100–200 µL of high salt extraction buffer (10 mM Tris-Cl, pH 8.0, 600 mM NaCl, 0.5 mM EDTA, protease inhibitors) at 4°C with gentle agitation for 30 min and pelleted at max speed at 4°C for 10 min. Supernatants were saved as the high salt fraction, and chromatin pellets were resuspended in LDS sample buffer (Invitrogen), boiled for 10 min, and sonicated in Bioruptor Pico (Diagenode).

Chromatin immunoprecipitation (ChIP)

ChIP was performed according to the Abcam protocol from 1–2 × 10⁷ cells crosslinked in 1% formaldehyde in PBS for 10 min. Chromatin was isolated in ChIP lysis buffer (50 mM HEPES-KOH pH7.5, 140 mM NaCl, 1 mM EDTA pH8.0, 1% Triton X-100, 0.1% Sodium Deoxycholate, 0.1% SDS, protease inhibitors) added at 750 µL per 1 × 10⁷ cells, incubated on ice for 10 min and sonicated in 100 µL aliquots in Bioruptor Pico for 10 cycles (30 s ON, 30 s OFF). Lysates were cleared by centrifugation at 16,000 × g for 10 min at 4°C and supplemented with 10% glycerol for storage at –80°C whereas 50 µL of lysates were processed for input DNA isolation as follows. Lysate aliquots were supplemented with 70 µL of elution buffer (1% SDS, 100 mM NaHCO₃), 4.8 µL of 5 M NaCl and 2 µL RNase A (10 mg/mL), and incubated while shaking at

65°C overnight. Then 2 μ L of proteinase K (20 mg/mL) was added to samples and incubated for 1 h while shaking at 60°C. Sample DNA was then purified using the NucleoSpin Gel and PCR clean up kit (Macherey-Nagel) with the SDS-compatible NTB buffer (Macherey-Nagel) and quantified to calculate the concentration of chromatin preps.

5 μ g of chromatin was diluted 1:10 in ChIP dilution buffer (50 mM Tris-HCl pH8.0, 150 mM NaCl, 2 mM EDTA pH8.0, 1% NP-40, 0.5% Sodium Deoxycholate, 0.1% SDS, Protease Inhibitors) and immunoprecipitated with 15 μ L of mouse α -V5 tag antibody-conjugated magnetic beads (CST Cat. No. 53475) overnight at 4°C. Beads were washed 2 \times in Low Salt Wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 150 mM NaCl), 2 \times in High Salt Wash Buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 500 mM NaCl), 1 \times in LiCl Wash Buffer (0.25 M LiCl, 1% NP-40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8.0), and 1 \times in PBS. Beads were eluted with 120 μ L of elution buffer by shaking for 30 min at 30°C. Eluates were then processed the same way as inputs to isolate DNA. Input and pulldown DNA samples were analyzed in triplicates by qPCR with iTaq Universal SYBR Green supermix (Bio-Rad) and the following pairs of primers:

Alu ChIP: 5'AATGGTACGATCTCGGCTCA-3' and 5'TAGCCAGGTGTGGTACTTG-3'

SATIII ChIP: 5'AATCAACCCGAGTGCAATCNGAATGGAATCG-3' and 5'TCCATTCCATTCTGTACTCGG-3'. Alu and SATIII primers were described in Shen et al.⁷⁵

SATII: 5'TCATCGAATGGAAATGAAAGGAGTCATCATCT-3' and 5'CGACCATTGGATGATTGCAGTCAA-3' were derived from the chr.1 SATII sequence, GenBank: X72623.1.

α SAT: (CST Cat. No. 4486).

RNA isolation and qPCR

RNAs were isolated using RNeasy Plus RNA isolation kit (Qiagen) and additionally purified of DNA contamination using DNA-free kit (Thermo Fisher). 2 μ g of RNA was reverse-transcribed using High Capacity cDNA Reverse Transcription kit (Applied Biosystems) per manufacturer's protocol, either with random primers, or with a GAPDH-reverse primer 5'GATGCAGGGATGATGTTCTG-3' and one of the following: an HSATII primer 5'ATCGAATGGAAATGAAAGGAGTCA-3' or an HSATIII primer 5'TCCATTCCATTCTGTACTCGG-3'. cDNAs were diluted 1:2 and 1 μ L of diluted cDNA was used per qPCR reaction. Reactions were run in triplicates with iTaq Universal SYBR Green supermix (Bio-Rad) and the following pairs of primers: 5'ACAACCTTGGCATTGAA-3' and 5'GATGCAGGGATGATGTTCTG-3' for GAPDH, described in⁷⁶; 5'TCATCGAATGGAAATGAAAGGAGTCA-3' and 5'CGACCATTGGATGATTGCAGTCAA-3' for HSATII; 5'AATCAACCCGAGTGCAATCGAATGGAATCG-3' and 5'TCCATTCCATTCTGTACTCGG-3' for HSATIII, described in Shen et al.⁷⁵ Triplicate Ct values for GAPDH were averaged and subtracted from HSAT values to derive Δ Ct values. The results were plotted as mean $2^{-\Delta$ Ct} values. Significance was determined in paired two-tailed t-tests.

Proximity Ligation Assay (PLA) and immunofluorescence *in situ* (IF)

EdU (Click Chemistry Tools) labeling of cells grown on coverslips was for 20 min at a final concentration of 20 μ M from a 10 mM stock in DMSO stored at -20° C. PLA was performed using Duolink red detection kit and Duolink anti-mouse and anti-rabbit antibodies (Millipore-Sigma) in accordance with our detailed methods paper,⁸³ except that after 2% formaldehyde fixation for 20 min cells were washed twice in PBS, permeabilized by addition of 4°C 90% methanol in PBS and used immediately or stored at -20° C before staining. For PLA, cells were washed twice for 5 min in PBS and rinsed once in 3% BSA in PBS. The following mix was prepared before use (in the order of addition to PBS): 20 μ M biotin azide and 0.4 μ M AZDye 488 azide (Click Chemistry Tools); 11 mM Na ascorbate; 2.2 mM CuSO₄; and incubated on cells for 30 min at room temperature (RT) in the dark. Cells were washed twice in 3% BSA in PBS and blocked with PBS/3% BSA/3% Normal goat serum for 20 min at room temperature then stained with mouse α -biotin antibody at 1:200 dilution and rabbit α -ATAD2 antibody at 1:200 dilution, or mouse α -biotin antibody at 1:200 dilution and rabbit α -V5 antibody at 1:1000 dilution for 1 h at RT. After washing in three changes of 0.3% BSA/0.1% Tween 20 in PBS, coverslips were incubated with PLA secondary antibodies and Duolink red detection reagents per manufacturer's manual and mounted in DUOLINK DAPI mounting medium (Millipore-Sigma).

Immunofluorescence *in situ* (IF) was performed as in¹⁸ and similar to the procedure above, except cells on coverslips were not EdU-labeled or clicked. After fixation, permeabilization and blocking as above, cells were incubated with rabbit α -V5 antibody at 1:1000 dilution for 1 h at RT, washed as above and incubated with Alexa 594-conjugated goat α -rabbit secondary antibody (Invitrogen) at 1:500 dilution for 1 h at RT. PLA and IF images were collected under Zeiss Axiovert 200 M microscope with 40 \times magnification objective using MicroManager software.⁷⁷ Digital images were analyzed with the Fiji ImageJ software package⁷⁸ with custom macros as described in⁸³ or with the CellProfiler software package⁷⁹.

RNAi-mediated depletion. siRNAs against ATAD2 (Hs_ATAD2_3, Cat. No. SI00103138, and Hs_ATAD2_5, Cat. No. SI02640232) and AllStars negative control SI03650318 siRNAs were from Qiagen. siRNA was transfected with lipofectamine RNAiMAX (Invitrogen) according to the manufacturer's protocol. Experiments were performed 36 to 48 h after transfection with individual siRNAs, which produced similar results. Depletion was verified in each transfection by western blotting.

QUANTIFICATION AND STATISTICAL ANALYSIS

Details of quantification and statistical tests are included in Figure Legends. Wilcoxon and Kolmogorov-Smirnov (KS) tests were done in RStudio and t-tests in Excel or RStudio. ANCOVA analyses for Figure 5 were done in python using the pingouin package's ancova module, and pairwise Tukey's HSD analysis was done in python using the statsmodels package's pairwise_tukeyhsd module. Fisher exact tests for Figure 6 were done in python using the SciPy package's fisher_exact module. For enrichment tests for Figure 7, the dataset obtained via random draws from the human interactome was fitted against probability density function for normal distribution in python using SciPy package's norm module (<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html> [docs.scipy.org]) to derive the mean and variance.