Published online 30 July 2019

NAR Genomics and Bioinformatics, 2019, Vol. 1, No. 1 ***e3***
doi: 10.1093/nargab/lqz003

Open chromatin structure in PolyQ disease-related genes: a potential mechanism for CAG repeat expansion in the normal human population

Matan Sorek[1,2,†], Lea R. Z. Cohen[1,2,†] and Eran Meshorer[1,2,*]

[1]Edmond and Lily Safra Center for Brain Sciences, Edmond J. Safra Campus, Jerusalem, Hebrew University of Jerusalem, 9190401, Israel and [2]Department of Genetics, The Alexander Silberman Institute of Life Sciences, Edmond J. Safra Campus, Jerusalem, Hebrew University of Jerusalem, 9190401, Israel

Received May 17, 2019; Revised July 07, 2019; Editorial Decision July 12, 2019; Accepted July 16, 2019

ABSTRACT

The human genome contains dozens of genes that encode for proteins containing long poly-glutamine repeats (polyQ, usually encoded by CAG codons) of 10Qs or more. However, only nine of these genes have been reported to expand beyond the healthy variation and cause diseases. To address whether these nine disease-associated genes are unique in any way, we compared genetic and epigenetic features relative to other types of genes, especially repeat containing genes that do not cause diseases. Our analyses show that in pluripotent cells, the nine polyQ disease-related genes are characterized by an open chromatin profile, enriched for active chromatin marks and depleted for suppressive chromatin marks. By contrast, genes that encode for polyQ-containing proteins that are not associated with diseases, and other repeat containing genes, possess a suppressive chromatin environment. We propose that the active epigenetic landscape support decreased genomic stability and higher susceptibility for expansion mutations.

INTRODUCTION

A large fraction of the genome is composed of repetitions of short DNA sequences. These repetitive sequences challenge DNA replication as well as the transcription, recombination and DNA repair machineries. Repetitive sequences can form unusual DNA structural loops (e.g. hairpin structures), which can interfere with these processes, resulting in expanded repeat tracts inside or outside of genes (1–4). Repeat instability is therefore a relatively common form of genetic mutation. Interestingly, the longest TNR expansions are usually found in non-dividing cells, while the short TNR expansions are observed in both dividing and non-dividing cells, and the mechanism of expansion seems to be determined by the cell cycle stage (5,6). Similar to all other genetic mutations, repeat expansions are genetically transmitted; however, they are less stable throughout generations. Most notable in relation to human diseases is the family of trinucleotide repeat (TNR) expansions.

Several disorders are caused by TNR expansions inside genes, either in the reading frame, in the untranslated region (UTR), or in regulatory elements (2,4). Expansions in non-coding regions of the gene typically induce diseases when the expansion is extremely long (hundreds or even thousands of repeats), while shorter expansions have no effect. Examples for such disorders include Friedreich's ataxia (FRDA / FA), caused by ∼100–1000 GAA TNR expansions in the first intron of the FXN gene, and Fragile X syndrome (FXS), caused by >200 repeat expansions of the CGG TNR in the 5′UTR region of the FMR1 gene (7). When the repetitive tract resides inside the gene's coding region, the expansions are translated, and therefore alter the protein product, inducing an extended protein. The exact type of expansion within the protein is dependent on the type of the repeat inside the gene.

There are two known groups of disorders induced by TNR expansions in coding regions: poly-glutamine (polyQ) related disorders and poly-alanine related disorders. The repeat lengths associated with the disease in poly-alanine disorders are much shorter (∼12–33) compared with those of polyQ disorders (∼20–200). Poly-alanine diseases are congenital disorders caused by expansions of either GCA, GCG or GCC, resulting in expanded tracts of alanine (A) (7,8). By contrast, polyQ disorders are almost always late onset. They are caused by CAG (and more rarely CAA) expansions, resulting in proteins containing an expanded tract of glutamines (Q). In the normal human population, the length of the repeat is quite variable, but comparatively small. When the length of the repeat exceeds a certain threshold, the carrier will develop the disease, with full penetrance (6). Each different mutated polyQ protein causes

*To whom correspondence should be addressed. Tel: +972 2 6585161; Fax: +972 2 6586073; Email: eran.meshorer@mail.huji.ac.il
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

widespread neuronal degeneration, in addition to some specific brain areas that are affected the most. The diseases share several pathological characteristics, including the formation of mutant protein cellular aggregates, progressive degeneration that lasts around 15 years eventually leading to patients' death, and a dominant pattern of inheritance (9). In addition, the age of onset, usually around mid-life, is reversely correlated with the length of the repeat (10).

Several factors have been associated with repeat instability. First, long repeats are less stable than shorter repeats. Other intrinsic properties of the repeat sequence, such as the genomic orientation of the repeat and the TNR purity affect the level of repeat instability between generations (4,11). The level of instability of TNR length also depends on other features, for example the parental source of the gene affect repeat instability between generations (4). Finally, the genomic context of the genes, e.g. DNA methylation, can also affect repeat instability (6). In addition, in mice models, insertion of a human repeat tract that includes the human genomic sequence increases repeat instability compared to insertion of the repeat tract without the regulatory sequences (4,12–15). Also supporting the role of genomic context, several chromatin modulators were shown to be necessary for long CAG repeat expansions (16–18). Importantly, in many TNR diseases, the expanded version of the gene is associated with an increase in heterochromatin marks and a decrease in euchromatin marks, suggesting that chromatin structure contributes to repeat instability in the mutated version of the genes (19).

Most of the studies to date have been carried out on disease models that already contained an expanded sequence. However, less is known about the causes for repeat expansions that take place in genes that encode for proteins with repeat tracts within the normal range. In healthy cells, DNA methylation and active histone marks were suggested to be enriched around disease-associated TNRs (20). Interestingly, CAG repeat expansions in bacterial and yeast models, as well as *in vitro* models using human cell extracts, demonstrated that even short, normal range, repeat tracts are unstable, with a tendency to expand (21–23). Repeat lengths of the nine polyQ disease-related genes in the normal population are much more variable compared to the other CAG repeat containing genes (24), with relatively stable length of repeat. We do not know, however, which cellular mechanisms contribute to this variability and allow for the expansion of the normal allele. Because the stability of long repeat sequences is linked to chromatin structure, we hypothesized that the chromatin state can also influence the expansion dynamics of short, normal repeats. There are several relevant developmental stages for these expansions. Somatic TNR mutations occur (25) and can affect the severity of the disease pathology (26). However, recent studies revealed that mutations during early development in progenitor cells are far more frequent than was previously thought, and are more frequent than paternal gamete mutations (27), suggesting that the expansion mutations can occur very early in the embryonic stage of either the patient or his/her parents. Therefore, we set to investigate the genetic and epigenetic state of TNR-containing genes in human pluripotent stem cells, which represent an early developmental stage, and for which ample data are available (28).

MATERIALS AND METHODS

Genomic data acquisition

DNA sequences for hg38 (human) and mm10 (mouse) were downloaded from ftp://ftp.ensembl.org/pub/release-90/fasta/homo_sapiens/dna/ and ftp://ftp.ensembl.org/pub/release-91/fasta/mus_musculus/dna/, respectively. Gene locations were downloaded from https://genome.ucsc.edu/cgi-bin/hgTables, and protein sequences were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/. Maximal numbers of different TNRs were calculated for whole genes and for coding sequence only for each gene variant and the maximal repeat length among all variants was used as the gene representative. Maximal repeat length for different amino-acids was calculated in a similar fashion. Genes and proteins were considered as containing long TNR or amino acids (AA) repeat if they had 10 repeats or more. In order to enable human–mouse comparison, only genes that exist in both datasets were taken into the analysis. For expression calculation, we downloaded expression data in human embryonic stem cells (hESCs) from the GREIN database (http://www.ilincs.org/apps/grein/). Overall, we have curated 35 samples from 12 independent experiments. The 12 datasets that were used were: GSE20301, GSE30567, GSE33480, GSE26880, GSE30995, GSE47117, GSE51607, GSE52133, GSE53094, GSE56152, GSE56796, and GSE61390. ChIP-Seq data were downloaded from BindDB (http://bind-db.huji.ac.il/, (28)). Binding score for each gene was permissively calculated based on all variants, where we considered a gene as bound by a factor if at least one of its variants was bound. Since the promoter region is the main regulatory site of the gene, and since longer genes have a larger probability to be bound by factors in general, we considered only the binding data in the promoter region, defined as −1000 to +1000 relative to the transcription start site (TSS).

Significance level calculations

To test significance for the larger repeat length of the nine polyQ disease-related proteins in human compared to mouse (Figure 2), we used a permutation test for the proteins with long (>10) polyQ tract and calculated human-to-mouse ratio of the median repeat length among 9 random genes from this group ($n = 67$, $P = 0.0013$). This was repeated 100,000 times. To test significance for the enriched binding of transcription factors (TFs) and histone modifications (HMs) on genes that encode for proteins with long AA tracts (Figure 3), we first calculated for each factor whether it binds a significantly ($P < 0.05$) large fraction of the genes with long AA tract, based on the one-sided proportion test null hypothesis defined by the rest of the genes. The expected number of such factors has a binomial distribution, which is used to deduce the significance levels of the actual number of these factors.

The number of factors is 222 in all cases. The number of genes with long repeats, the number of enriched factors (EF) and the corresponding one-sided P values for the 6 AA are: Q: 58, EF:67, 2.3×10^{-34}; S: 51, EF:42, 2.6×10^{-14}; E: 74,

EF:23, 0.00036; P: 55, EF:18, 0.016; G: 52, EF:90, 5.9×10^{-58}; A: 85, EF:90, 5.9×10^{-58}.

Heterochromatin and active signatures

Activators and repressors were defined based on the correlation-based distance from the repressive histone mark H3K27me3 (in Supplementary Figure S3 activators were defined based on the correlation-based distance from the active histone mark H3K4me3, and repressors were defined based on (29)). To calculate the Z-score of the active or suppressive signature of a given group, we first used for each factor the Z statistic of the proportion test, defined by

$$Z = \frac{p - p_0}{\sqrt{\frac{p_0 \cdot (1-p_0)}{n}}}$$

where p is the percentage of bound genes by the given factor among the group, p_0 is the percentage of bound genes among all genes and n is the size of the group. Next, we defined the combined Z-score as

$$Z_{\text{combined}} = \frac{\sum_{i=1}^{f} Z_i}{\sqrt{f}}$$

where f is the number of related-factors for the suppressive or active signature. The results of this test are translated to the significance of the enrichment/depletion observed by normal distribution, when absolute value of above 1.96 indicates $P < 0.05$, two-tailed t-test.

The nine longest polyQ encoding genes were defined as the nine genes that had the largest repeat lengths and are not polyQ disease-related. The nine similar expression genes (Figure 4F) were defined based on the median gene expression of the genes among all hESC datasets. For each of the polyQ disease-related gene, we used the long-polyQ encoding gene that had the smallest (in absolute value) $\log_2$ of the fold-change ratio.

To validate the difference between human and mouse in the Z-score, we ran a permutation test of 100 000 random permutations, where each time nine random genes were chosen and the Z-value difference between human and mouse was measured, for both active and repressive profiles. While the difference between human and mouse in the Z-score of the active signature was not significant, the difference in the repressive signature was significant ($P < 0.011$).

RESULTS

Number of amino acid repeat tract containing proteins is evolutionarily conserved

The basic feature of the polyQ disease-related genes is that they all contain TNRs encoding the amino acid (AA) glutamine (Q). Q is encoded by either CAG or CAA, and codon bias in humans show that 72% of Qs are encoded by CAGs (30). In polyQ-stretches, this bias is further exaggerated, with >80% of the Qs encoded by CAGs in stretches of 4Qs or longer (31). We first asked whether polyQ-containing proteins differ from other polyAA-containing proteins in their repeat length conservation. To examine the AA length conservation, we characterized AA repeat-containing proteins in the proteomes of different organisms. Comparing the proteomes of humans, mice and several primates, we observed that Q repeats are not the most common (Figure 1A) nor the longest (Figure 1B) tracts of AA repeats. Analysing the number of proteins containing long (>10) repeat tracts, we found, as expected, that not all of the 20 AAs form long repeat tracts, and in over 75% of the proteins that contain long repeat tracts, these repeat tracts are comprised of one of only 6 AAs (including glutamine) (Figure 1A), as was also previously shown (32). Among these AAs are Alanine and Serine, which are encoded, among others, by a shifted CAG codon (AGC encodes Serine and GCA encodes Alanine). In addition, the number of repeat-containing proteins for each AA is relatively conserved among primates, while mouse contains sometimes a larger number of AA repeat-containing proteins (Figure 1A). This was also true for coding regions at the nucleotide level (Supplementary Figure S1). To further evaluate the conservation of repeat tracts, we compared the length of repeat tract in human to primates and mouse (Figure 1C), and found that most of the repeat tract lengths are relatively similar between the species. Interestingly, most of the polyQ disease-related genes also contain a repeat tract of another AA (Supplementary Table S1), raising the intriguing idea that those genes might be more susceptible for accumulating repeats.

PolyQ disease-related proteins are evolutionarily unstable

Analysis of the human genome revealed about 50 human genes that contain polyQ-encoding repeat tracts (of 10 or more Qs), but which are not associated with any known polyQ-related disease. Although mice have an overall larger number of polyQ containing proteins (Figure 1A), the nine polyQ disease-related proteins have significantly longer repeat tracts in humans compared with mice ($P < 0.002$, see 'Materials and Methods' section), already in the healthy variants of the proteins (Figure 2A). The length of the polyQ tract of most of the polyQ-containing proteins that are non-disease-related fall on the diagonal, suggesting an overall high conservation between human and mouse for polyQ proteins ($R^2 = 0.77$), further highlighting the nine polyQ disease-related proteins as evolutionarily unique.

As in the case of polyQ, there are also nine genes that encode for proteins with large poly-alanine repeat tracts, which are known to cause diseases when the poly-alanine tract is abnormally expanded (8). But interestingly, by contrast to the polyQ-related proteins, the lengths of all poly-alanine repeat tracts are conserved between mice and human (Figure 2B). Together, these data highlight the nine polyQ disease-related genes as evolutionarily distinct from other polyQ genes, as well as from poly-alanine related genes.

Genes that encode for proteins with amino-acid repeat tracts are highly bound

As noted above, over 50 proteins in humans contain polyQ tracts of 10Qs or more, but these repeats are likely not expanded beyond the normal range. The selective expansion in the nine polyQ disease-related genes can be a result of a non-genetic characteristic of the genes, such as regulatory

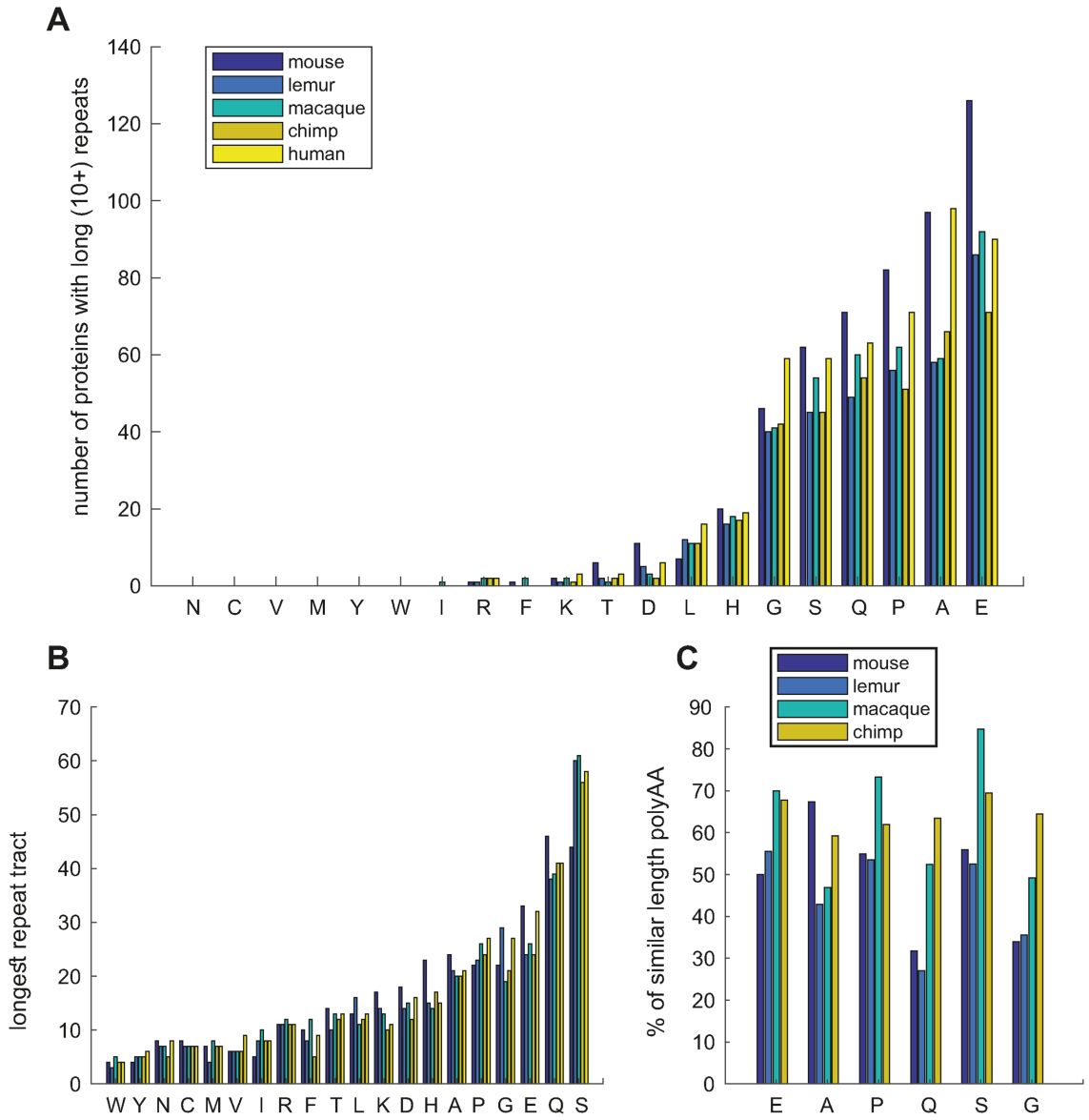

Figure 1. Amino-acid (AA) repeats are conserved in evolution. (**A**) The number of proteins containing 10 or more repeats for each of the AAs in the five organisms listed. (**B**) The length of the longest repeat tracts of each AA in the five organisms listed. (**C**) The fraction of genes with repeat length difference of <25% compared to human for every organism and for every one of the 6 most frequent AA repeats. This is likely an underestimate due to the sequencing quality of the non-human non-mouse genomes.

elements that can influence the stability of repeat tracts (4), or the epigenetic environment in general, which determines accessibility of transcription factors and other chromatin binding proteins. Therefore, we turned to chromatin structure to seek for a potential explanation for the selective instability of the nine polyQ disease-related genes in comparison to the other polyQ encoding genes.

In order to test whether the nine polyQ disease-related genes are characterized by a unique epigenetic landscape separating them from the other polyQ genes, we analysed their chromatin state in embryonic stem cells (ESCs), which reflect early embryonic stages. We used the BindDB database (28), which contains ChIP-Seq datasets from more than 200 experiments for human and mouse ESCs. We defined groups of genes by the length and type of repeats in the genes/protein products, or according to expansion diseases. Using BindDB, we analysed the binding of transcription factors (TFs) and histone modifications (HMs) to promoter regions on several groups of genes in human ESCs. We first found that genes that encode for proteins with large poly-alanine or polyQ repeat tracts are bound by more TFs and enriched for a larger number of HMs compared to the rest of the genes (Figure 3A and F). This also characterized genes that encode for proteins with other AA repeat tracts (Figure 3). These results demonstrate that genes that encode AA repeat-containing proteins, (especially alanine, glycine, glutamine and serine) are highly bound, i.e. bound by a larger number of factors.

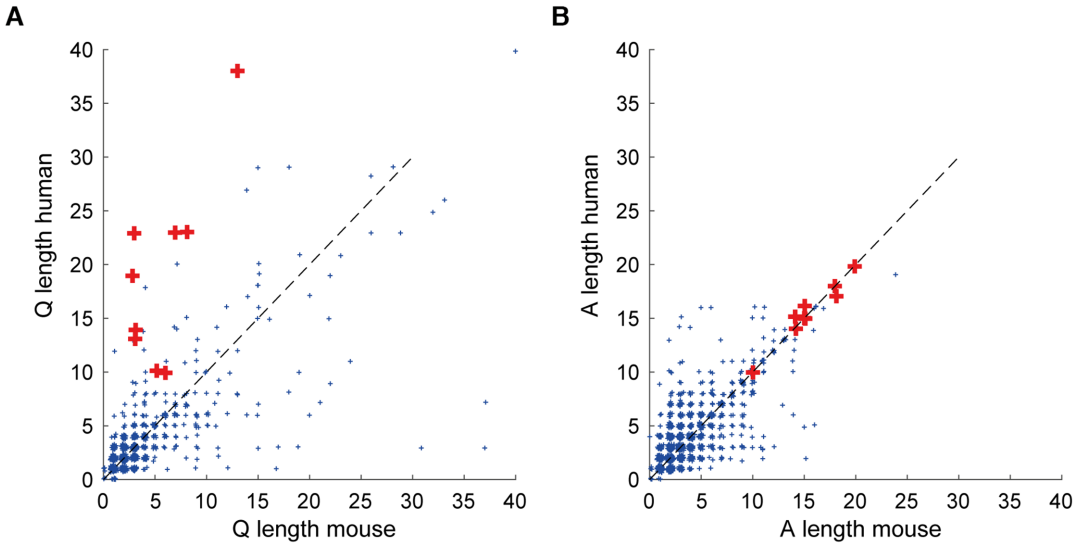

Figure 2. Conservation of alanine and glutamine repeats in human and mouse. (**A**) The length of the maximal glutamine repeat tract of proteins in healthy humans versus mice. Red crosses denote the nine polyQ disease-related genes. (**B**) Same as (A) for Alanine. Red crosses denote the nine poly-alanine disease-related genes.

Human polyQ disease-related genes have an active epigenetic profile

Next, we analysed the differences in binding to promoters of activators and repressors for each of the groups of genes we had defined. Examination of the TFs and HMs revealed that many of them generally correlate with each other. We therefore divided the HM and TFs into groups based on their similarity to H3K27me3, a well-known histone mark associated with repression and the Polycomb Repressor Complex II (PRC2). The first group consists of TFs and HMs most similar to H3K27me3 (Figure 4A, black box, top left). This group contains mostly TFs and HMs associated with heterochromatin and repression (Supplementary Table S2), hereafter referred to as 'repressors'. The second group consists of TFs and HMs most distinct from H3K27me3 (Figure 4A, blue box, bottom right), and, as expected, mostly contains TFs and HMs associated with transcription and activation (Supplementary Table S2), hereafter referred to as 'activators'. We then defined the epigenetic profile based on these two groups. For each set of genes, we calculated an epigenetic profile score based on the combined enrichment/depletion of these two groups (see 'Materials and Methods' section).

Given a group of genes, there are four types of possible profiles: 1. 'Bivalent': enrichment of both active and repressive marks; 2. 'Depleted': depletion of both active and repressive marks; 3. 'Super-Repressed': enrichment of repressive marks and depletion of active marks; and 4. 'Super-Active': enrichment of active marks and depletion of repressive marks. Accordingly, we examined these profiles for different gene sets in mouse and human ESCs.

In line with our previous findings (Figure 3), we found that human genes that encode for proteins with a long polyQ tract (Figure 4B, middle), as well as genes that encode for proteins with a long poly-alanine repeat tract (Figure 4C), are characterized by a 'Bivalent' profile, i.e. highly enriched for both repressive (red bars) and active (blue bars) marks. This profile also characterized genes that encode for proteins with a long AA repeat tract of any type, as well as genes that contain a long TNR in their DNA sequence (Figure 4C). For mouse genes, the corresponding AA repeat related gene groups, as well as genes that contain a long TNR in their DNA sequence, all showed a 'Bivalent' or a 'Super-Repressed' profile (Figure 4D and E).

Strikingly, the nine polyQ disease-related genes in human behave differently than all other sets of genes; these genes have a 'Super-Active' profile, enriched for active marks and depleted for repressive marks (Figure 4B, left). This is in contrast with the profile of these genes in mouse, where they are both active and repressive (Figure 4D). This was also in sharp contrast compared with the poly-alanine disease-related genes, which are characterized by a predominantly repressive profile, showing a 'Bivalent' profile for the human genes (Figure 4B, right) and a 'Super-Repressed' profile for the mouse genes (Figure 4D, right). Permutation tests for significance revealed that the polyQ disease-related genes were significantly more depleted for repressors when compared to the genes that encode for proteins with a long polyQ tract not related to disease ($P = 0.0016$), as well as to genes that encode for long AA repeat ($P = 0.0074$), and to genes containing long TNR ($P = 0.0046$) (Supplementary Figure S2). These results suggest that polyQ disease-related genes are unique, among all related groups of genes, in their 'Super-Active' profile. The results were robust to the groups' definition of repressors and activators (Supplementary Figure S3).

To test whether long repeat sequences are generally associated with a unique chromatin state, we examined the epigenetic profile of the nine genes with the longest glutamine repeat tract, which are not related to polyQ disease (Supplementary Table S3). The polyQ tracts encoded by these genes are longer than the repeat tracts of the nine polyQ disease-related genes in WT as a group (mean and median

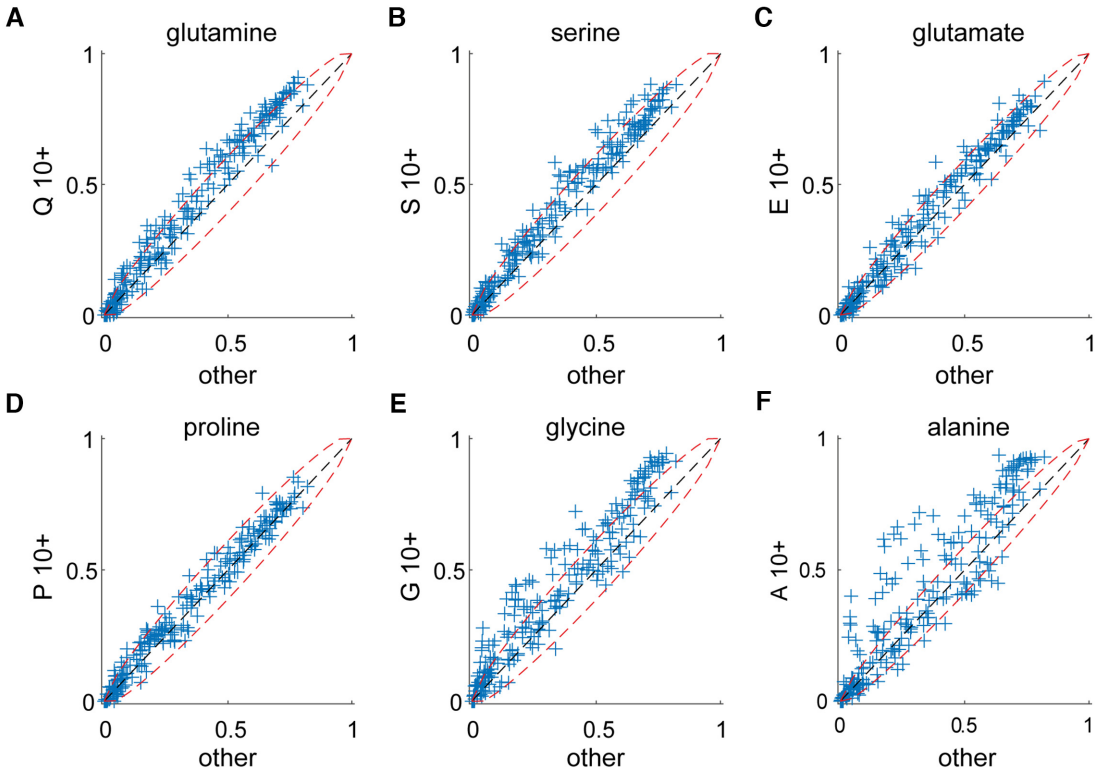

Figure 3. Human genes that encode for proteins with polyAA repeat tracts are highly bound in human ESCs. Shown is the fraction of genes from a defined group bound by a factor (*y*-axes), compared with the fraction of the rest of the genes, which are bound by that factor (*x*-axes). Each marker represents one factor. Black dashed line represents the 'expected by chance' $y = x$ line. Red dashed lines represent 1.64 standard deviations (equivalent to $P = 0.05$) from the null hypothesis defined by the proportion test. The defined groups are genes that encode for proteins with repeat tract whose length is 10 or more of either glutamine ($P = 2.3 \times 10^{-34}$), serine ($P = 2.6 \times 10^{-14}$), glutamate ($P = 0.00036$), proline ($P = 0.016$), glycine ($P = 5.9 \times 10^{-58}$) and alanine ($P = 5.9 \times 10^{-58}$).

length of repeat around 29Qs compared to 19Qs, respectively). We found that the epigenetic profile of these genes was slightly (non-significantly) depleted for activators and indistinguishable from the rest of the genome for repressors (Figure 4F, middle).

To further exclude the possibility that the Super-Active profile of the nine polyQ disease genes is merely an artifact of these genes being functionally important in hESCs and therefore expressed, we investigated the relative expression levels of these genes. While all nine polyQ disease genes are expressed above average in hESCs (Figure 5A), the expression level of the nine longest polyQ encoding genes, which do not have a 'Super-Active' profile (Figure 5B, orange crosses) was similar to the polyQ disease-related genes. In addition, when we examined the nine polyQ encoding genes with the most similar expression level compared to the nine polyQ disease-related genes, we found a 'Bivalent', but not a 'Super-Active', open chromatin structure (Figure 4F, right). Finally, the number of bound repressors was reversely correlated with repeat tract length in disease, a proxy to the tendency of repeat expansion (Supplementary Figure S4), pointing at a quantitative relation between the chromatin state and the expansion mutation rate of the gene.

These data suggest that unlike all other repeat-containing genes in mouse and human, and in contrast with other polyQ encoding highly expressed genes, the polyQ disease-related genes specifically possess an active epigenetic profile in ESCs. Such an active conformation may render them more prone for repeat expansions in early embryonic development.

DISCUSSION

We observed that the nine polyQ disease-related genes, in their WT form, are characterized by a unique epigenetic profile in human ESCs, characterized by both an open (depleted for heterochromatic marks) and an active (enriched for activation marks) state. This observation raises the hypothesis that the mechanism of repeat expansion is enhanced by this open, super-active state (Figure 6). One potential direct consequence of the open chromatin structure of the polyQ disease-related genes is reduced protection from expansion type mutations. We hypothesize that suppressive chromatin marks may better protect against such mutations than active chromatin in the early embryo, rendering the nine polyQ disease-related genes prone for instability.

Supporting this idea, a recent study showed that the DNA:RNA hybrids, also known as R-loops, exist *in vivo* in WT hESC in the repeat-containing genes FMR1 and C9orf72 (33). These R-loops are composed of DNA and a nascent RNA, and are thought to induce genome instability in G/C rich sequences (34,35). The highly active and open chromatin structure we observed in the nine polyQ disease

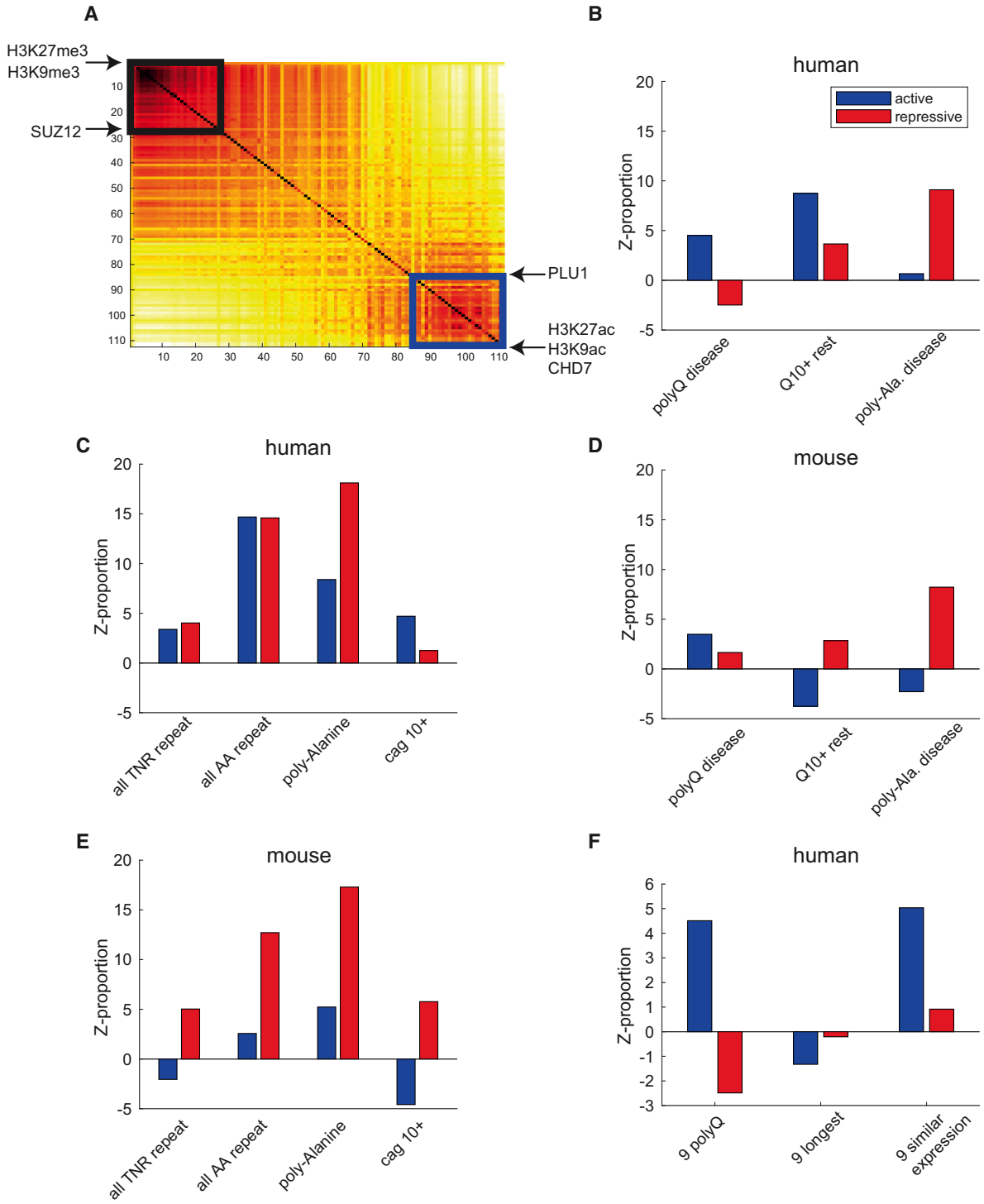

Figure 4. PolyQ disease-related genes have a unique epigenetic profile. (**A**) Clustering of transcription factors (TFs) and histone modifications (HMs) based on similarity to H3K27me3. Shown is a similarity matrix sorted by similarity to H3K27me3. Each row/column relates to a different TF/HM. The upper row and the left column represent H3K27me3. The top-left box contains repressors, and the lower-right box contains activators. Highest similarity in darker red, lowest in brighter yellow. (**B–F**) Combined *Z*-score for active (blue) and repressive (red) marks (see 'Materials and Methods' section for details) for different gene groups. Positive combined *Z*-score represents enrichment and negative combined *Z*-score represents depletion. The height of the bars represent significance, when absolute value of above 1.96 indicates $P < 0.05$, two-tailed *t*-test. The significance of the difference between human and mouse in the *Z*-score of the repressive signature was validated ($P < 0.011$, see 'Materials and Methods' section). (**B**) Epigenetic profile of polyQ disease genes is 'Super-Active' in human ESCs (left), in contrast to human genes encoding for proteins with long polyQ tracts, which do not cause diseases (middle), and to poly-alanine disease-related genes (right). (**C**) Epigenetic profile of human genes that contain long TNRs of any type, genes that encode for proteins with long AA repeat tract of any type, genes that encode for proteins with long poly-alanine repeat tract and genes that contain long CAG repeats. (**D**) Same as (B) for mouse epigenetic profile of mouse genes. (**E**) Same as (C) for mouse epigenetic profile and mouse genes. (**F**) Epigenetic profile of the 9 polyQ disease-related genes, the longest Q tract containing genes, and the highly expressed polyQ genes that are not associated with disease.

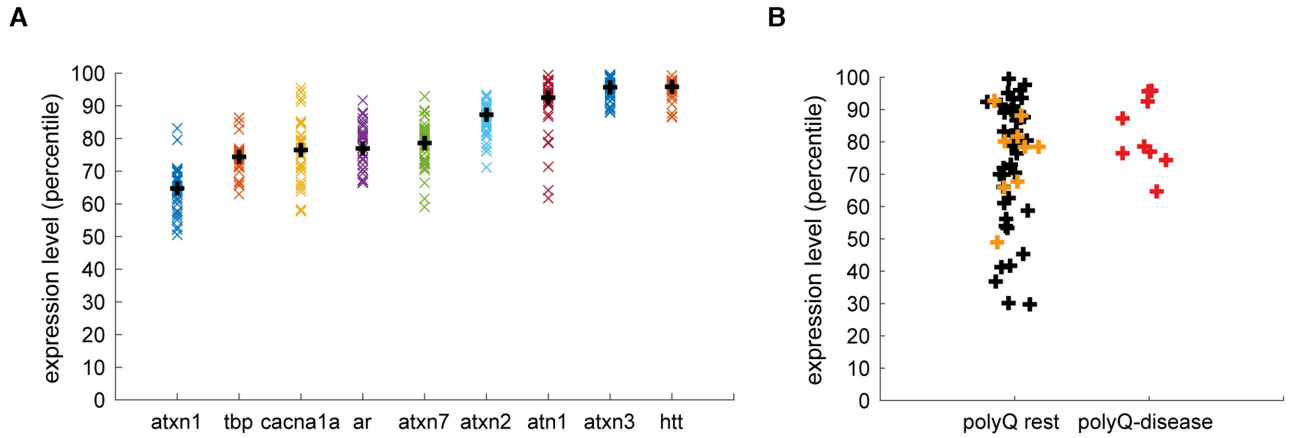

Figure 5. PolyQ disease-related genes are expressed in hESCs. (**A**) Expression level percentile from multiple published sources (see 'Materials and Methods' section) of the nine polyQ disease-related genes. All nine genes are expressed above the median expression level. (**B**) Average expression level percentile from (A) of the nine polyQ disease-related genes (red, right) compared with the rest of the polyQ containing genes (black, left). Orange marks represent the nine genes encoding for the longest polyQ tracts that are not disease-related. They show similar expression levels to the nine polyQ disease-related genes.

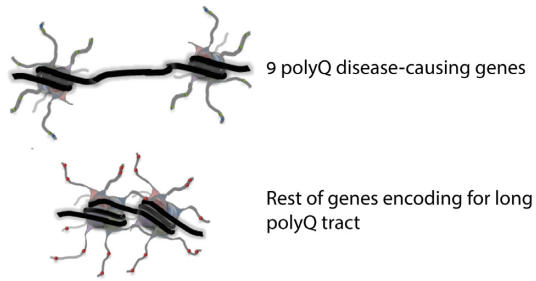

Figure 6. A model for CAG repeat expansions. The nine polyQ disease genes, at their normal form, have an open chromatin structure that contributes to the expansion of CAG repeats. This is in contrast to the rest of the genes that encode for proteins with long polyQ repeat tracts that have a more closed chromatin structure, which is protected from these processes, therefore inhibiting these genes from repeat instability. Blue dots represent active histone marks and red dots represent repressive histone marks.

genes in healthy hESCs may further facilitate R-loops formation, which can further lead to repeat expansion.

Interestingly, another recent work showed that most of the disease-associated short tandem repeats in the human genome are located at the boundaries between topologically associated domains (TADs) (36). This characteristic, which is stable across different cell types during early human development, was shown to be disrupted in the FMR1 gene in Fragile X syndrome patients, which correlated with the silencing of that gene. Another study had revealed the absence of repressive marks in CAG expansions related genes (20). We also provide exemplar density plots for the HM in their genomic coordinates for the polyQ disease-related gene ATXN7 and for another long polyQ-containing gene, which is not disease-associated, FOXP2 (Supplementary Figure S5). This supports the notion that polyQ disease genes acquire the observed repressed heterochromatic structure only upon expansion to disease cases, while the normal form is much more open and active, as we hypothesize.

Expansion from WT to disease length is often attributed to slippage mutations. Slippage mutations occur in repetitive regions during replication due to the formation of secondary structures and the capacity of the two strands to anneal to one another in different places (37). It is possible that the open chromatin allows the formation of those secondary structures more often than the repressed, closed heterochromatin. If this is indeed the case, the heterochromatin signature reported in TNR expansion disorders might serve as a protection mechanism from acquiring even more mutations (19).

Our model provides explanations for several known unexplained observations. First, it explains why human genes that encode for proteins with poly-alanine repeat tracts normally have a similar number of repeats in human and mouse, as these genes have a repressive profile in both organisms, suggested to protect them from acquiring expansions during early embryonic stages. This contrasts with polyQ disease-related genes in humans, where the 'Super Active' profile provides less protection and allows the polyQ disease-related genes to acquire expansions over time in human but not in mouse. This also explains why other genes that encode for proteins with polyQ tracts in human do not undergo expansions, as their epigenetic profile is protective, preventing them from acquiring expansions in these early developmental stages. Germ line mutation and instability can be explained by our model as well, but the data to test open chromatin in those stages is not yet available.

In addition to explaining some of the more established observations, our model also provides a few testable predictions. First, as our model assumes that Super-Active state increases the probability for an expansion, we expect that genes, which have a Super-Active profile and contain CAG repeats, will have the potential to expand and eventually even cause a polyQ disease. Searching for such potential proteins, with a repeat tract size close to the threshold of the polyQ disease-related genes, yielded a few potential hits: PRDM10, BMP2k and NCOA3 with polyQ repeats, and SRP14 with poly-alanine repeats. Our model also predicts that other polyQ containing proteins may cause a late onset disease too should their repeat tract expand further. Taken together, we uncover a human-specific, uniquely 'open', epigenetic signature for the nine polyQ disease-related genes

in ESCs, which we suggest may support repeat expansions during evolution.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Shahar Simon and Ayelet Hashahar-Cohen for help with data acquisition. E.M. is the Arthur Gutterman Chair for Stem Cell Research.

FUNDING

Azrieli PhD Fellowship, Azrieli Foundation (to M.S.); The Israel Science Foundation [1140/17 to E.M]. Funding for open access charge: Israel Science Foundation
Conflict of interest statement. None declared.

REFERENCES

1. Mirkin,S.M. (2006) DNA structures, repeat expansions and human hereditary disorders. *Curr. Opin. Struct. Biol.*, **16**, 351–358.
2. Mirkin,S.M. (2007) Expandable DNA repeats and human disease. *Nature*, **447**, 932–940.
3. Cohen-Carmon,D. and Meshorer,E. (2012) Polyglutamine (polyQ) disorders: the chromatin connection. *Nucleus*, **3**, 433–441.
4. Pearson,C.E., Edamura,K.N. and Cleary,J.D. (2005) Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.*, **6**, 729–742.
5. Gomes-Pereira,M., Hilley,J.D., Morales,F., Adam,B., James,H.E. and Monckton,D.G. (2014) Disease-associated CAG·CTG triplet repeats expand rapidly in non-dividing mouse cells, but cell cycle arrest is insufficient to drive expansion. *Nucleic Acids Res.*, **42**, 7047–7056.
6. McMurray,C.T. (2010) Mechanisms of trinucleotide repeat instability during human development. *Nat. Rev. Genet.*, **11**, 786–799.
7. La Spada,A.R. and Taylor,J.P. (2010) Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.*, **11**, 247–258.
8. Shoubridge,C. and Gecz,J. (2012) Polyalanine tract disorders and neurocognitive phenotypes. In: Hannan,AJ (ed). *Tandem Repeat Polymorphisms*. Springer, NY, pp. 185–203.
9. Fan,H.C., Ho,L.I., Chi,C.S., Chen,S.J., Peng,G.S., Chan,T.M., Lin,S.Z. and Harn,H.J. (2014) Polyglutamine (PolyQ) Diseases: Genetics to treatments. *Cell Transplant*, **23**, 441–458.
10. Kuiper,E.F.E., de Mattos,E.P., Jardim,L.B., Kampinga,H.H. and Bergink,S. (2017) Chaperones in polyglutamine aggregation: Beyond the Q-Stretch. *Front. Neurosci.*, **11**, 145.
11. Usdin,K., House,N.C.M. and Freudenreich,C.H. (2015) Repeat instability during DNA repair: Insights from model systems. *Crit. Rev. Biochem. Mol. Biol.*, **50**, 142–167.
12. Cleary,J.D. and Pearson,C.E. (2003) The contribution of *cis*-elements to disease-associated repeat instability: clinical and experimental evidence. *Cytogenet. Genome Res.*, **100**, 25–55.
13. Brock,G. (1999) Cis-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands. *Hum. Mol. Genet.*, **8**, 1061–1067.
14. Gourdon,G., Dessen,P., Lia,A.S., Junien,C. and Hofmann-Radvanyi,H. (1997) Intriguing association between disease associated unstable trinucleotide repeat and CpG island. *Ann. Genet.*, **40**, 73–77.
15. Libby,R.T. (2003) Genomic context drives SCA7 CAG repeat instability, while expressed SCA7 cDNAs are intergenerationally and somatically stable in transgenic mice. *Hum. Mol. Genet.*, **12**, 41–50.
16. Jung,J. and Bonini,N. (2007) CREB-binding protein modulates repeat instability in a drosophila model for PolyQ disease. *Science*, **315**, 1857–1859.
17. Dion,V., Lin,Y., Hubert,L., Waterland,R.A. and Wilson,J.H. (2008) Dnmt1 deficiency promotes CAG repeat expansion in the mouse germline. *Hum. Mol. Genet.*, **17**, 1306–1317.
18. Koch,M.R., House,N.C.M., Cosetta,C.M., Jong,R.M., Salomon,C.G., Joyce,C.E., Philips,E.A., Su,X.A. and Freudenreich,C.H. (2018) The chromatin remodeler Isw1 prevents CAG repeat expansions during transcription in *Saccharomyces cerevisiae*. *Genetics*, **208**, 963–976.
19. Dion,V. and Wilson,J.H. (2009) Instability and chromatin structure of expanded trinucleotide repeats. *Trends Genet. TIG*, **25**, 288–297.
20. Essebier,A., Vera Wolf,P., Cao,M.D., Carroll,B.J., Balasubramanian,S. and Bodén,M. (2016) Statistical enrichment of epigenetic states around triplet repeats that can undergo expansions. *Front. Neurosci.*, **10**, 92.
21. Henricksen,L.A., Tom,S., Liu,Y. and Bambara,R.A. (2000) Inhibition of flap endonuclease 1 by flap secondary structure and relevance to repeat sequence expansion. *J. Biol. Chem.*, **275**, 16420–16427.
22. Liu,Y., Prasad,R., Beard,W.A., Hou,E.W., Horton,J.K., McMurray,C.T. and Wilson,S.H. (2009) Coordination between polymerase β and FEN1 can modulate CAG repeat expansion. *J. Biol. Chem.*, **284**, 28352–28366.
23. Pluciennik,A., Burdett,V., Baitinger,C., Iyer,R.R., Shi,K. and Modrich,P. (2013) Extrahelical (CAG)/(CTG) triplet repeat elements support proliferating cell nuclear antigen loading and MutL endonuclease activation. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 12277–12282.
24. Butland,S.L., Devon,R.S., Huang,Y., Mead,C.-L., Meynert,A.M., Neal,S.J., Lee,S.S., Wilkinson,A., Yang,G.S., Yuen,M.M. *et al.* (2007) CAG-encoded polyglutamine length polymorphism in the human genome. *BMC Genomics*, **8**, 126.
25. Shelbourne,P.F., Keller-McGandy,C., Bi,W.L., Yoon,S.-R., Dubeau,L., Veitch,N.J., Vonsattel,J.P., Wexler,N.S., Arnheim,N. and Augood,S.J. (2007) Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. *Hum. Mol. Genet.*, **16**, 1133–1142.
26. Swami,M., Hendricks,A.E., Gillis,T., Massood,T., Mysore,J., Myers,R.H. and Wheeler,V.C. (2009) Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum. Mol. Genet.*, **18**, 3039–3047.
27. D'Gama,A.M. and Walsh,C.A. (2018) Somatic mosaicism and neurodevelopmental disease. *Nat. Neurosci.*, **21**, 1504–1514.
28. Livyatan,I., Aaronson,Y., Gokhman,D., Ashkenazi,R. and Meshorer,E. (2015) BindDB: An integrated database and webtool platform for 'Reverse-ChIP' epigenomic analysis. *Cell Stem Cell*, **17**, 647–648.
29. Saksouk,N., Simboeck,E. and Déjardin,J. (2015) Constitutive heterochromatin formation and transcription in mammals. *Epigenetics Chromatin*, **8**, 3.
30. Athey,J., Alexaki,A., Osipova,E., Rostovtsev,A., Santana-Quintero,L.V., Katneni,U., Simonyan,V. and Kimchi-Sarfaty,C. (2017) A new and updated resource for codon usage tables. *BMC Bioinformatics*, **18**, 391.
31. Mier,P. and Andrade-Navarro,M.A. (2018) Glutamine codon usage and polyQ evolution in primates depend on the Q stretch length. *Genome Biol. Evol.*, **10**, 816–825.
32. Alba,M.M. (2004) Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.*, **14**, 549–554.
33. Abu Diab,M., Mor-Shaked,H., Cohen,E., Cohen-Hadad,Y., Ram,O., Epsztejn-Litman,S. and Eiges,R. (2018) The G-rich repeats in *FMR1* and *C9orf72* loci are hotspots for local unpairing of DNA. *Genetics*, **210**, 1239–1252.
34. Jonkers,I. and Lis,J.T. (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.*, **16**, 167–177.
35. Lin,Y., Dent,S.Y.R., Wilson,J.H., Wells,R.D. and Napierala,M. (2010) R loops stimulate genetic instability of CTG·CAG repeats. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 692–697.
36. Sun,J.H., Zhou,L., Emerson,D.J., Phyo,S.A., Titus,K.R., Gong,W., Gilgenast,T.G., Beagan,J.A., Davidson,B.L., Tassone,F. *et al.* (2018) Disease-Associated short tandem repeats Co-localize with chromatin domain boundaries. *Cell*, **175**, 224–238.
37. Bzymek,M. and Lovett,S.T. (2001) Instability of repetitive DNA sequences: The role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 8319–8325.