DISEASE MECHANISMS

Tandem repeats mediating genetic plasticity in health and disease

Anthony J. Hannan[1,2]

Abstract | Accumulating evidence suggests that many classes of DNA repeats exhibit attributes that distinguish them from other genetic variants, including the fact that they are more liable to mutation; this enables them to mediate genetic plasticity. The expansion of tandem repeats, particularly of short tandem repeats, can cause a range of disorders (including Huntington disease, various ataxias, motor neuron disease, frontotemporal dementia, fragile X syndrome and other neurological disorders), and emerging data suggest that tandem repeat polymorphisms (TRPs) can also regulate gene expression in healthy individuals. TRPs in human genomes may also contribute to the missing heritability of polygenic disorders. A better understanding of tandem repeats and their associated repeatome, as well as their capacity for genetic plasticity via both germline and somatic mutations, is needed to transform our understanding of the role of TRPs in health and disease.

Genome-wide association studies
(GWAS). Studies that have been used extensively since the development of microchips that assay single nucleotide polymorphisms (SNPs) across the genome. These studies examine the association of particular polymorphisms and their linked genes with traits and disorders.

Repeatome
The entire collection of repetitive DNA sequences within a whole genome. Subsets of the repeatome can be transcribed and translated, producing equivalent RNA and protein repeatomes within the transcriptome and proteome.

[1]*Florey Institute of Neuroscience and Mental Health, University of Melbourne.*
[2]*Department of Anatomy and Neuroscience, University of Melbourne, Parkville, Victoria, Australia.*
e-mail: anthony.hannan@florey.edu.au

doi:10.1038/nrg.2017.115
Published online 5 Feb 2018

We are in the midst of a revolution in genomics as our understanding of the evolution and function of genomes from thousands of species and the genetics of human diseases rapidly expands. However, progress in our understanding of genetic polymorphisms and mutations has been limited by technologies that have been optimized for analysing single nucleotide polymorphisms (SNPs), particularly in genome-wide association studies (GWAS). Indeed, over half of the human genome is estimated to consist of repetitive elements that make up the repeatome [1], a large number of which lie within genes and their regulatory regions[2]. Furthermore, as a substantial proportion of repetitive sequences can be transcribed and translated, the repeatome also represents subsets of the transcriptome and proteome.

Repetitive DNA consists of tandem and interspersed repeats and copy number variants, which are major structural polymorphisms that can occur in either of these forms. Interspersed (also known as dispersed) repetitive DNA sequences include retrotransposons, as exemplified by short interspersed nuclear elements (SINEs) such as Alu repeats and long interspersed nuclear elements (LINEs) such as LINE-1 — an element that is still active in the human genome. Although much of the repeatome has been historically considered to be 'genomic junk', comparative genomics suggests that many classes of DNA repeats have evolved under tight selective constraints[2].

Tandem repeats have been referred to as microsatellite DNA and minisatellite DNA, variable number of tandem repeats (VNTRs) and simple sequence repeats[3,4]. The most commonly observed repeats in the genome are homonucleotide, dinucleotide and trinucleotide repeats, although tetranucleotides, pentanucleotides, hexanucleotides and longer motifs are also frequently seen[5]. Short tandem repeats (STRs) (also known as microsatellite DNA) consist of 1–6 bp motifs, and it has been estimated that there are over 1 million discrete STR loci in the human genome, constituting around 3% of genomic DNA[6]. Thousands of these loci have been implicated in the regulation of gene expression[7,8]. Minisatellite DNA consists of longer (>6 bp) motifs in tandem repetition.

Tandem repeats have also been implicated in molecular and cellular dysfunction associated with human diseases. For example, we have known for over 2 decades that specific tandem repeat expansions can cause Huntington disease, spinocerebellar ataxias, Friedreich ataxia (FRDA), fragile X syndrome (FXS), myotonic dystrophy and other diseases; the majority of these diseases are disorders of the nervous system[3]. Furthermore, tandem repeat expansion has recently been shown to be a major genetic contributor to frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS), the latter of which is the most common form of motor neuron disease[9,10]. Emerging genetic data also suggest roles for tandem repeat polymorphisms (TRPs) in the regulatory control of gene products associated with various polygenic disorders[7,8].

This Review discusses recurring themes in tandem repeat biology, with a focus on the repeat-associated genetics and pathogenesis of a range of human disorders. Furthermore, I discuss the urgent need to fully and accurately catalogue TRPs and disease-associated mutations involving repeat instability in the human

Short interspersed nuclear elements
(SINEs). A major class of interspersed repetitive DNA, with each element consisting of approximately 100–700 bp of DNA. SINEs are retrotransposons and are thus able to amplify themselves within genomes, usually via RNA intermediates and reverse transcription.

Alu repeats
Primate-specific SINEs that constitute the most abundant transposable elements in the human genome, which contains over 1 million Alu elements. Alu elements are retrotransposons consisting of interspersed repetitive DNA segments approximately 300 bp in length that constitute over 10% of the human genome.

genome and explain how doing this could enhance our understanding of the missing heritability in polygenic disorders. As STRs are the most intensively studied tandem repeats and have been most comprehensively linked to disease states, I focus on STRs in health and disease. I discuss the putative functions of STRs across diverse species and their causative roles in a range of human diseases, particularly those affecting the nervous system.

From tandem repeat mutations to disease

Many tandem repeats have historically been used as genetic markers owing to their unique polymorphic ranges. However, the discovery that specific tandem repeats were expanded via dynamic mutations in human disorders led researchers to consider their functions. Specific tandem repeats were first linked to FXS and spinobulbar muscular atrophy, followed by myotonic dystrophy, Huntington disease, dentatorubral-pallidoluysian atrophy and other neurological disorders involving ataxia[3,11,12] (TABLE 1). These tandem repeat disorders (TRDs) are caused by different tandem repeats that are located in various genic regions (FIG. 1), and they can be subdivided according to their genic location and pathogenic mechanisms (FIG. 2). A key aspect of TRDs is that they generally do not result simply in binary phenotypes (that is, individuals without the disease as compared to patients with the disease) but rather in phenotypes that are on a continuous quantitative scale (for example, the age of disease onset or the severity of disease) and that are modulated by the number of tandem repeats. I now briefly outline the different classes of TRDs and the insight they have provided into the biological functions and dysfunctions of tandem repeats.

The role of polyglutamine tracts in human disorders. A major class of TRDs is caused by trinucleotide CAG repeat expansions that encode polyglutamine tracts in various proteins[13]. The most common polyglutamine disease, and one of the first diseases found to be caused

Table 1 | **Tandem repeat disorders affecting the nervous system**

Human disorder	Gene	Tandem repeat motif (amino acid repeat)	Range of tandem repeat length Normal (expanded)
Polyglutamine diseases			
HD	*HTT*	CAG (Q) (RAN translation)	6–35 (36–250)
SCA1	*ATXN1*	CAG (Q)	6–38 (39–88)
SCA2	*ATXN2* and *ATXN2-AS*[a]	CAG•CTG (Q)	14–32 (33–200)
SCA3	*ATXN3*	CAG (Q)	12–44 (55–87)
SCA6	*CACNA1A*	CAG (Q)	4–18 (20–33)
SCA7	*ATXN7*	CAG (Q)	4–33 (37–460)
SCA17	*TBP*	CAG (Q)	25–40 (43–66)
DRPLA	*ATN1*	CAG (Q)	3–35 (48–93)
SBMA	*AR*	CAG (Q)	9–34 (38–68)
Neurodegenerative diseases other than polyglutamine diseases			
SCA8	*ATXN8OS* and *ATXN8*[a]	CTG•CAG (Q) (RAN translation)	15–50 (80–250)
SCA10	*ATXN10*	ATTGT (not translated)	10–32 (>280)
SCA12	*PPP2R2B*	CAG (RAN translation?)	4–32 (43–78)
Friedreich ataxia	*FXN*	GAA (not translated)	5–34 (66–1300)
HDL2	*JPH3*[b]	CTG•CAG (Q) (RAN translation?)	6–28 (41–58)
Fragile X disorders			
FXS	*FMR1*	CGG (not translated)	5–44 (>200)
FXTAS	*FMR1*	CGG (RAN translation)	5–44 (55–200)
C9ORF72 TRDs			
C9ORF72 ALS, C9ORF72 FTD and possibly other diseases	C9ORF72	GGGGCC•GGCCCC (RAN translation)	3–25 (>30)
Myotonic dystrophies			
DM1	*DMPK*	CTG•CAG (RAN translation)	5–34 (>50)
DM2	*CNBP*	CCTG•CAGG (RAN translation)	11–26 (>50)

This table is not an exhaustive summary of tandem repeat disorders affecting the nervous system but includes the major Mendelian disorders known to be caused by tandem repeat expansions. ALS, amyotrophic lateral sclerosis; DM, myotonic dystrophy; DRPLA, dentatorubral-pallidoluysian atrophy; FTD, frontotemporal dementia; FXS, fragile X syndrome; FXTAS, fragile X tremor-ataxia syndrome; HD, Huntington disease; HDL2, Huntington disease-like 2; RAN, repeat-associated non-ATG; SBMA, spinobulbar muscular atrophy; SCA, spinocerebellar ataxia; TRDs, tandem repeat disorders. [a]*ATXN2* and *ATXN2-AS* and *ATXN8OS* and *ATXN8* are the genes associated with SCA2 and SCA8, respectively, that are bidirectionally transcribed from opposite strands of DNA. [b]indicates translation off the antisense strand. ? indicates possible RAN translation.

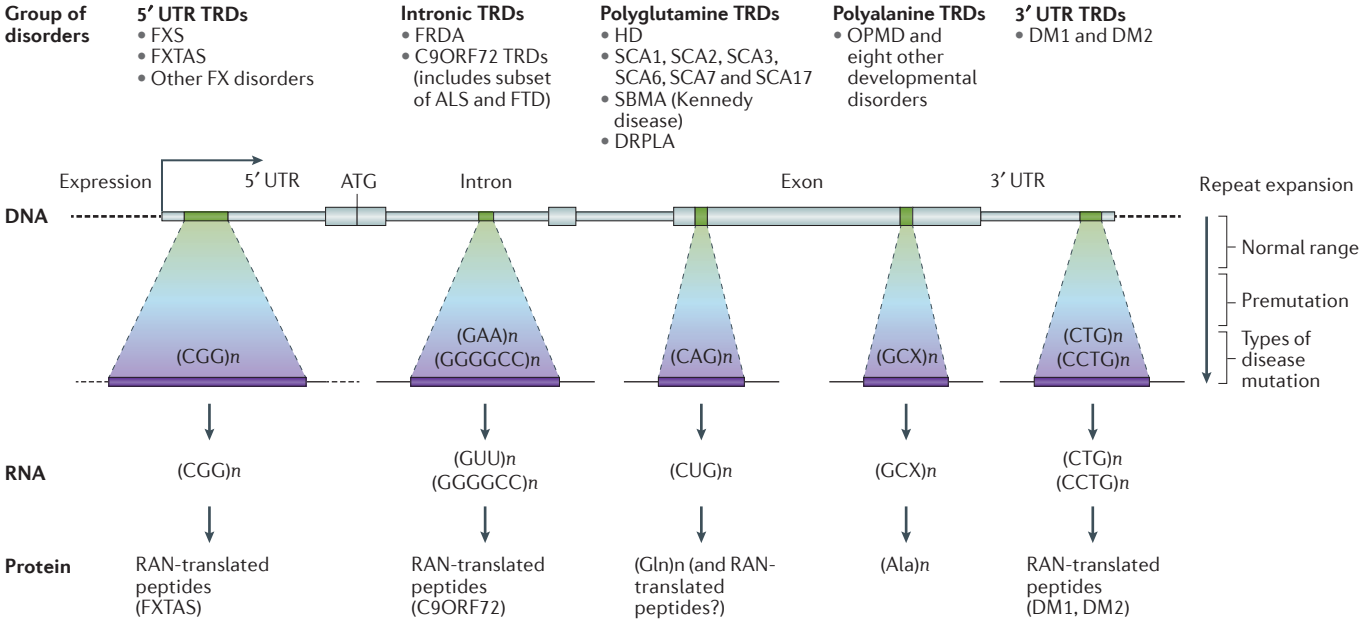


Figure 1 | **Genic locations and expression products of tandem repeats causing human disorders.** Tandem repeat disorders (TRDs) are a subset of human disorders that are caused by tandem repeat mutations, which are usually tandem repeat expansions. These TRDs can be classified according to whether the tandem repeat is located in the 5′ untranslated region (UTR), intron, exon or 3′ UTR. Exonic TRDs are usually polyglutamine TRDs (where the tandem repeat encodes polyglutamine tracts) or polyalanine TRDs (where the tandem repeat encodes polyalanine tracts), but they can encode other amino acid repeats when repeat-associated non-ATG (RAN) translation occurs. The genic location of the repeat, as well as the specific tandem repeat motifs, can have consequences at the DNA, RNA and protein levels that result in cellular and systemic pathogenic outcomes. Note that some of these TRDs are caused by bidirectional transcription from the disease gene; however, for simplicity, only the sense strand of DNA is shown in this diagram. For the 5′ UTR TRDs and intronic TRDs, fragile X tremor-ataxia syndrome (FXTAS) and C9ORF72 TRDs, respectively, are the only diseases thus far that have been shown to involve RAN-translated peptides. For the polyglutamine TRDs, Huntington disease (HD) has been shown to also involve RAN-translated peptides, and it is possible that this also occurs in the other polyglutamine TRDs. Similarly, for the 3′ UTR TRDs, myotonic dystrophy 1 (DM1) and DM2 have been shown to involve RAN-translated peptides. Ala, alanine; ALS, amyotrophic lateral sclerosis; DRPLA, dentatorubral-pallidoluysian atrophy; FRDA, Friedreich ataxia; FTD, frontotemporal dementia; FX, fragile X; FXS, fragile X syndrome; Gln, glutamine; OPMD, oculopharyngeal muscular atrophy; SBMA, spinobulbar muscular atrophy; SCA, spinocerebellar ataxia.

Long interspersed nuclear elements
(LINEs). Another major class of interspersed repetitive DNA. They consist of elements approximately 7,000 bp in length. LINEs constitute over 20% of the human genome and are transcriptionally and translationally active (encoding a reverse transcriptase), with recent evidence suggesting that they have evolved roles, including somatic mutation affecting brain development and function.

Tandem repeat disorders
(TRDs). Disorders caused by mutation of a tandem repeat sequence. These are usually Mendelian disorders with dominant or recessive inheritance patterns, although tandem repeats are increasingly being found to contribute to additional disorders with non-Mendelian inheritance patterns.

Homopeptide
A repeating sequence of amino acids encoded by a trinucleotide repeat. For example, a CAG repeat encodes a polyglutamine homopeptide and when the expansion of this homopeptide occurs in the huntingtin protein, it causes Huntington disease.

by a CAG repeat expansion, is Huntington disease. This disease is a fatal autosomal dominant neurodegenerative disease caused by a CAG repeat expansion encoding an extended polyglutamine tract in the huntingtin protein[14]. The length of the CAG repeat is inversely proportional to the age at disease onset, although genetic and environmental modifiers can also influence the age at disease onset[15]. Genetic modifiers, that is, other genes and their respective alleles that modify the onset of Huntington disease, have also been actively pursued. Intriguingly, one of the major genetic modifiers of Huntington disease regulates DNA repair, and thus it potentially regulates the somatic mutability of the CAG repeat[16]. It is not yet known whether specific polymorphisms in such modifier genes increase or decrease somatic mutability of the CAG repeat. The CAG expansion in the huntingtin gene, which results in a glutamine expansion in the protein, causes a complex cascade of molecular and cellular pathogenesis that leads to psychiatric, cognitive and motor symptoms[17].

In addition to Huntington disease, there are at least eight other polyglutamine diseases, including six types of spinocerebellar ataxias (SCA1, SCA2, SCA3, SCA6, SCA7 and SCA17; SCA3 is also known as Machado–Joseph disease), spinobulbar muscular atrophy (SBMA; also known as Kennedy disease) and dentatorubral-pallidoluysian atrophy (DRPLA)[12]. SBMA was the first of these diseases found to be caused by a CAG and glutamine repeat expansion[18]. The CAG repeat expansions in the different genes and the associated polyglutamine-expanded proteins that cause these other polyglutamine diseases have been reviewed in detail[19] and will not be discussed extensively here. Interestingly, some additional diseases found to involve CAG·CTG repeat expansions were only proposed to involve toxic homopeptide expansions when it was shown that the CAG repeat could be transcribed off the antisense strands of their respective genes, including those causing Huntington disease-like 2 (HDL2)[20] and SCA8 (REF. [21]). In the past few years, the discovery that repeat-associated non-ATG translation (RAN translation) of toxic peptides may contribute to the pathogenesis of SCA8, as well as Huntington disease, myotonic dystrophy 1 (DM1), FXS, C9ORF72-related ALS and C9ORF72-related FTD, has generated much excitement[22–24] (see below).

Although polyglutamine-expansion diseases were thought to involve proteins with toxic gain-of-function

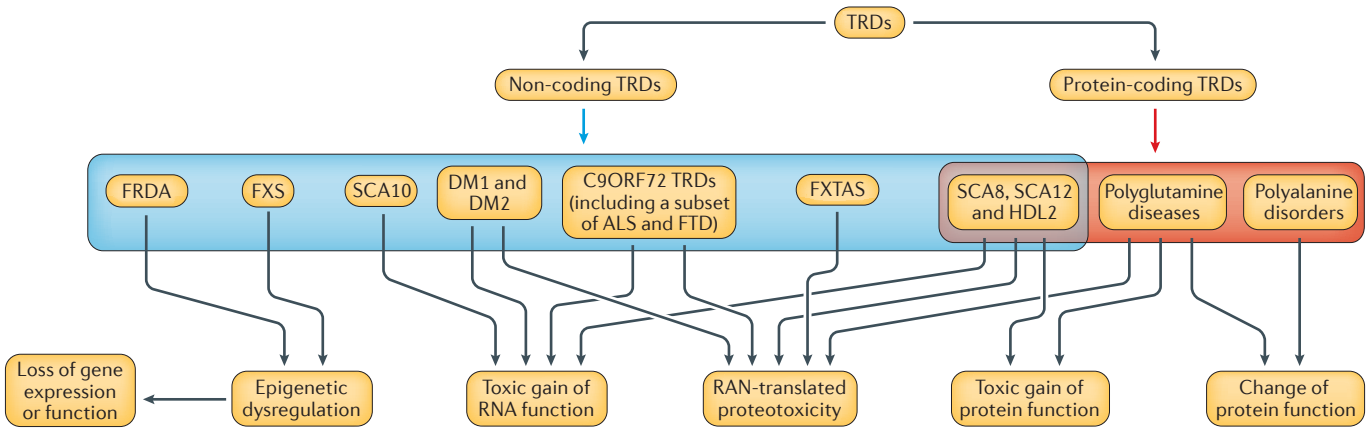


Figure 2 | **Pathways to pathogenesis in tandem repeat disorders.** The tandem repeat disorders (TRDs) can initially be classified according to whether they consist of protein-coding or non-coding tandem repeats and further classified according to various downstream molecular consequences and pathogenic mechanisms. There are aspects of pathogenesis for each disorder that remain to be elucidated. Therefore, these proposed pathways are only indicative and do not exhaustively encapsulate the expanding frontiers of research across the field of TRDs. ALS, amyotrophic lateral sclerosis; DM, myotonic dystrophy; FRDA, Friedreich ataxia; FTD, frontotemporal dementia; FXS, fragile X syndrome; FXTAS, fragile X tremor-ataxia syndrome; HDL2, Huntington disease-like 2; RAN, repeat-associated non-ATG; SCA, spinocerebellar ataxia.

mutations, evidence also suggests that native polyglutamines (that is, polyglutamines that have not been expanded) with improved function also contribute to pathogenesis[13]. Therefore, polyglutamine-expansion diseases could be considered to involve proteins with change-of-function mutations. Key studies demonstrating the importance of the polyglutamine tract in the native, wild-type huntingtin protein include evolutionary and comparative genomics[25,26], imaging genetics[27,28] and functional genomics, for example, the deletion of CAG repeats in a mutant mouse[29], although their precise functions within each of the respective proteins have not yet been elucidated. Interestingly, TRPs within *HTT* (the gene encoding huntingtin) that comprise CAG repeat lengths within the normal range have been associated with brain development and function, and incompletely penetrant CAG repeat lengths have been associated with depression[30]. A new study, involving two cohorts, found that both relatively short and relatively long CAG repeat lengths, within the normal range for the *HTT* gene, were associated with increased risk of depression[31]. This raises interesting questions regarding the genetic contributions of CAG repeats to depression in patients with Huntington disease and in healthy individuals[32]. Furthermore, CAG repeats with a length within the normal range have been associated with depression in two other genes associated with polyglutamine diseases, *TBP* (the gene encoding TATA-box binding protein) and *ATXN7* (the gene encoding ataxin 7)[33], adding to the evidence that tandem repeats in the normal range contribute to quantitative traits. Nevertheless, most evidence suggests that expanded polyglutamine tracts are neurotoxic in CAG repeat-expansion neurodegenerative diseases, and it appears that the spatiotemporal expression patterns of each disease gene, together with the intracellular localization of the disease protein and its ability to modify transcriptomes and proteomes within specific human cell populations, is likely to contribute to the unique and overlapping attributes of polyglutamine diseases. The disease-specific and shared mechanisms of pathogenesis in polyglutamine diseases have been extensively investigated and reviewed[3,5,12,17,19] and will not be the focus of this Review.

Although polyglutamine diseases constitute the largest subgroup of TRDs, various other diseases are caused by pathological repeat expansions. Polyalanine disorders, which predominantly affect embryogenesis and other aspects of development, are biochemically most similar to polyglutamine diseases[34,35]. In general, polyalanine disorders do not appear to be progressive or neurodegenerative, and they seem to disrupt the function of normal polyalanine tracts in the respective proteins; normal polyalanine tracts most often regulate transcription[34,36–38].

Fragile X-associated and other neurological tandem repeat disorders. Some non-coding tandem repeat expansions have been associated with other neurodegenerative diseases, including FRDA[12]. FRDA is an autosomal recessive ataxia caused by a GAA trinucleotide repeat expansion in an intronic region of *FXN*, the gene encoding the mitochondrial protein frataxin[39]. In patients with FRDA, *FXN* is repressed by the presence of GAA tracts, resulting in insufficient levels of frataxin[40]. The pathogenesis of FRDA and other neurodegenerative diseases that involve non-coding tandem repeat expansions, such as SCA10 (which is caused by an intronic ATTCT pentanucleotide expansion in *ATXN10*, the gene encoding ataxin 10)[41], thus appears to be distinct from the pathogenesis of polyglutamine diseases[42].

FXS was one of the first disorders shown to be caused by a tandem repeat expansion mutation. An expansion of the CGG trinucleotide in *FMR1*, the gene encoding fragile X mental retardation 1 protein (FMRP), beyond approximately 200 repeats was found to cause FXS[43,44]. FXS is caused by silencing of the *FMR1* gene due to

Repeat-associated non-ATG translation
(RAN translation). Type of translation of a peptide from a tandem repeat that occurs in the absence of an ATG start codon. The resultant peptides, consisting of repeating amino acid sequences, have been shown to exert toxic effects in specific human diseases.

hypermethylation of the CGG repeat expansion[45], and it appears to be primarily a neuronal and synaptic disorder characterized by mental retardation, autistic-like features and other developmental abnormalities. Following the discovery of the cause of FXS, premutation CGG repeat polymorphisms of ~55–200 repeats in *FMR1* were found to be associated with fragile X tremor-ataxia syndrome[46] (FXTAS), ovarian failure and other disorders[47]. RAN translation can occur from the trinucleotide repeat in *FMR1* in the absence of a canonical ATG start codon[22], a discovery that has relevance for TRDs in addition to fragile X-associated disease[23,24]. Finally, a rarer neurodevelopmental disorder, fragile XE syndrome, is caused by a CCG repeat expansion in *FMR2* (also known as *AFF2*)[48,49].

DM1 is also caused by a tandem repeat expansion. Specifically, DM1 is caused by a CTG trinucleotide repeat expansion in *DMPK*, the gene encoding DM1-protein kinase, and this expansion appears to produce a toxic RNA species that disrupts RNA splicing[50,51]; however, other potential molecular mechanisms for how this expansion disrupts *DMPK*, including RAN translation, have been implicated[22]. Interestingly, DM2 is caused by a CCTG tetranucleotide repeat expansion in *CNBP* (also known as *ZNF9*), which encodes CCHC-type zinc-finger nucleic-acid binding protein (CNBP)[52]; RAN translation that produces both LPAC and QAGR tetrapeptide proteins from the CCTG·CAGG repeats has also been implicated in the pathogenesis of DM2 (REF. [53]).

Insights from amyotrophic lateral sclerosis and frontotemporal dementia. The various types of TRDs and the mechanisms of pathogenesis that are associated with them have been the subject of intense investigation in recent years[54–56]. Most recently, a large subset of cases in two major neurodegenerative diseases, ALS and FTD, was found to be caused by tandem repeat expansions[9,10,57] (see below).

ALS is the most common form of motor neuron disease, and it progresses rapidly in patients, leading to early death. Specific polymorphisms in a CAG repeat tract in *ATXN2*, which encodes polyglutamine tracts of intermediate size in ataxin 2, have been associated with an increased risk of ALS[58]. These CAG repeat tracts are usually shorter (~29–32 repeats) than those that cause SCA2 (>32 repeats)[58]; however, a small number of patients with ALS have an *ATXN2* expansion in the size range of those present in SCA2 (REF. [59]), supporting a genetic link between ALS and SCA2. Furthermore, intermediate size CAG repeats in *ATXN2* may act as a disease modifier in combination with a hexanucleotide repeat expansion in another gene, *C9ORF72*, which causes *C9ORF72*-related ALS[60].

Hexanucleotide GGGGCC repeat expansions in *C9ORF72* are major genetic contributors to ALS and another neurodegenerative disease, FTD[9,10]. Although GGGGCC repeat expansions in *C9ORF72* are associated with approximately 40% of familial ALS cases, they have also been linked to sporadic ALS cases[9,10,57]. The role of the *C9ORF72* hexanucleotide expansion in ALS and FTD is under intense investigation, and there is evidence that this expansion is associated with other neurodegenerative disorders, including Alzheimer disease[57,61]. However, as the association of hexanucleotide expansion with other neurodegenerative disorders may reflect the clinical heterogeneity and diagnostic ambiguity of some neurodegenerative diseases, systematic large-scale investigations are required to address the full pathogenic impact of *C9ORF72* hexanucleotide expansions. This tandem repeat expansion in *C9ORF72* may contribute to the pathogenesis of neurodegenerative diseases because it is transcribed as part of a disease-associated C9ORF72 RNA that may be involved in multiple RNA-mediated pathogenic pathways[57] and/or as a result of RAN translation of dipeptide repeats, which causes proteotoxicity[62]. The details of these mechanisms have been recently discussed[24,57], although there are many key aspects of pathogenesis to be elucidated for hexanucleotide repeat expansions and other TRDs.

The overview of TRDs in this section is not exhaustive, and tandem repeat instability has been implicated in other disorders, such as SCA36 (REF. [63]), myoclonic epilepsy[64,65] (also known as EPM1 or Unverricht–Lundborg disease) and in many types of cancer[66].

RAN-translated peptides as major contributors to human disease. The capacity for RAN translation to occur in multiple reading frames of tandem repeats, particularly in trinucleotide and hexanucleotide repeats, suggests that this mechanism has roles in TRDs other than those discussed above[22–24]. For example, as SCA12, which is thought to be caused by a non-coding CAG repeat expansion, has parallels with neurodegenerative disorders (for example, with SCA8) associated with RAN translation[22], it is worth investigating whether RAN translation is a pathogenic mediator of SCA12. Furthermore, the similarity of the TRD FRDA to TRDs involving peptide-mediated neurotoxicity (for example, various spinocerebellar ataxias)[12] could potentially be explained by RAN translation as a common mechanism. Therefore, the known Mendelian TRDs need to be systematically assessed for the possible involvement of RAN translation. Furthermore, considering the large number of tandem repeats in the human genome and the current uncertainty regarding the number and type of repeats needed to trigger RAN translation, the capacity of RAN translation to cause or modify human diseases is substantial. Increasing our understanding of the molecular and cellular regulators of RAN translation, the specificity of RAN translation for tandem repeat motifs and repeat lengths, and the length-dependent effects of each species of RAN-translated peptide are a high priority for future research in the field of tandem repeats.

Tandem repeats in healthy genomes

Tandem repeats fine-tune gene expression. Many non-coding tandem repeats are located in gene regulatory regions and are thought to fine-tune gene transcription through several mechanisms[7,8,67–72]. For example, non-coding tandem repeats can modulate transcription[73,74], and one mechanism involves the formation of binding sites for transcription factors[69]. Furthermore,

tandem repeats can cause the separation of transcription regulatory elements to indirectly affect gene expression[75]. Tandem repeats might also drive the formation of the Z-DNA secondary structure[76] and induce heterochromatin formation and epigenetic modifications, including DNA methylation[42,77].

A striking demonstration that STRs modulate gene expression genome-wide was provided via analyses of human genomic and transcriptomic data[7]. Using an STR profiler called lobSTR[78], a genome-wide survey of STRs was conducted on data from the 1000 Genomes Project, and from a subset of these individuals, the gene expression data had been profiled in lymphoblastoid cell lines (LCLs). Over 2,000 human expression STRs (eSTRs) were identified from approximately 190,000 analysed STRs, the tandem repeat lengths of which correlated to the expression level of genes located near them in LCLs. The authors concluded that eSTRs account for 10–15% of the *cis* heritability contributed by all common variants[7]. Furthermore, these investigators found an enrichment of eSTRs associated with a number of complex polygenic disorders[7]. This landmark study, together with other findings, including a related study that associates STRs with gene expression and DNA methylation[8], provides indirect evidence for the previous proposal[79] that tandem repeats could be a major contributor to missing heritability of human traits and disorders.

Tandem repeats modulate RNA structure and function. A substantial proportion of tandem repeats are transcribed, which might enable them to modify the structure and function of RNAs. There is evidence that tandem repeats, particularly CA and TG dinucleotides, in introns can modify alternative splicing[80,81]. Furthermore, tandem repeats, particularly expanded trinucleotide and hexanucleotide repeats, have been associated with the formation of R loops, which consist of a DNA–RNA hybrid and a displaced single-stranded DNA[82,83]. Additionally, the recent discovery of tandem repeat-induced RNA phase transitions[84] may have implications for the understanding of tandem repeat disorders. Tandem repeats can also alter the structure of mRNA and non-coding RNA transcripts and thus potentially alter their stability, degradation, transport and translation, as well as their ability to bind to DNA, RNA, proteins and other molecules[50].

Homopeptides modulate protein structure and function. A subset of STRs — predominantly mononucleotide, trinucleotide and hexanucleotide repeats — resides in open reading frames and is translated into homopeptide sequences, such as polyglutamine or polyalanine tracts or dipeptide repeats. Tandem repeats with motifs >6 bp are translated less often. It has been proposed that single amino acid repeats (that is, homopeptides) have evolved specific roles that are associated with protein structure and function, including protein–DNA, protein–RNA and protein–protein interactions. Homopeptide sequences are thought to confer unique structural properties to the protein domains in which they reside, including flexibility[5], which have functional consequences. These potential roles of tandem repeats in proteins have been investigated[5,85–87] but will not be discussed in detail in this Review.

Tandem repeats and genetic plasticity

TRPs exhibit a range of attributes that distinguish them from SNPs and other genetic variants[79], including the fact that tandem repeats are more liable to mutation than single nucleotides[88]. In recent years, the variability of tandem repeats has been catalogued on a genome-wide scale[89–94]. The mutation rate at tandem repeat loci can be orders of magnitude higher than the mutation rate at single nucleotide loci, thus tandem repeat instability and associated disorders are often referred to as dynamic mutations[54]. The motif composition, motif length, tandem repeat length and epigenetic modifications of tandem repeats all influence their mutation rates[11]. Furthermore, TRPs exhibit extended digital (that is, multiallelic) polymorphic ranges, whereas SNPs generally involve only binary (that is, biallelic) variants (FIG. 3), which has various evolutionary, developmental and functional implications[95]. This array of possible variants at tandem repeat loci may result in a range of biological consequences[79]. Variability of specific genic and intergenic tandem repeats has functional consequences, for example, in cell biology[96] and animal behaviour[97]. Non-coding tandem repeats that are transcribed but not translated have the potential to modulate RNA structure and function. Furthermore, there is abundant evidence that tandem repeats encoding amino acid runs (for example, polyglutamine, polyalanine or polyproline tracts) have an impact on protein structure and function[5,13,25,29,34,36,37,98,99] (FIG. 4).

Somatic mutation of tandem repeats. The extent of somatic mutations in tandem repeats and their possible functions are unclear. However, many tandem repeats are highly liable to mutation during both meiosis and mitosis, which can lead to a high rate of somatic mutations in tandem repeats in healthy and diseased tissues[79,100–103]; this high rate of mutation has implications for genetic plasticity in development, biological functions and human disease[95]. However, traditional approaches to DNA sequencing, which involve sampling specific cell populations that are usually derived from blood, saliva or buccal cells, may miss somatic mutations and the genomic plasticity and cellular heterogeneity that they are associated with. Somatic variability of tandem repeats may have evolved various functions in developmental and adult biological processes. The effect of somatic mutation on another component of the repeatome, retrotransposons[104], is also being pursued, although the extent and heterogeneity of genome-wide somatic tandem repeat mutation remains largely unexplored. Although somatic tandem repeat mutations have been heavily implicated in cancer[66], they may also have implications for a range of disorders, particularly those affecting the

Somatic mutations
Changes in the DNA sequence that occur in somatic (non-germline) cells after conception, either during development or in adulthood. The gene mutation can occur either during mitosis (somatic cell division) or in non-dividing cells.

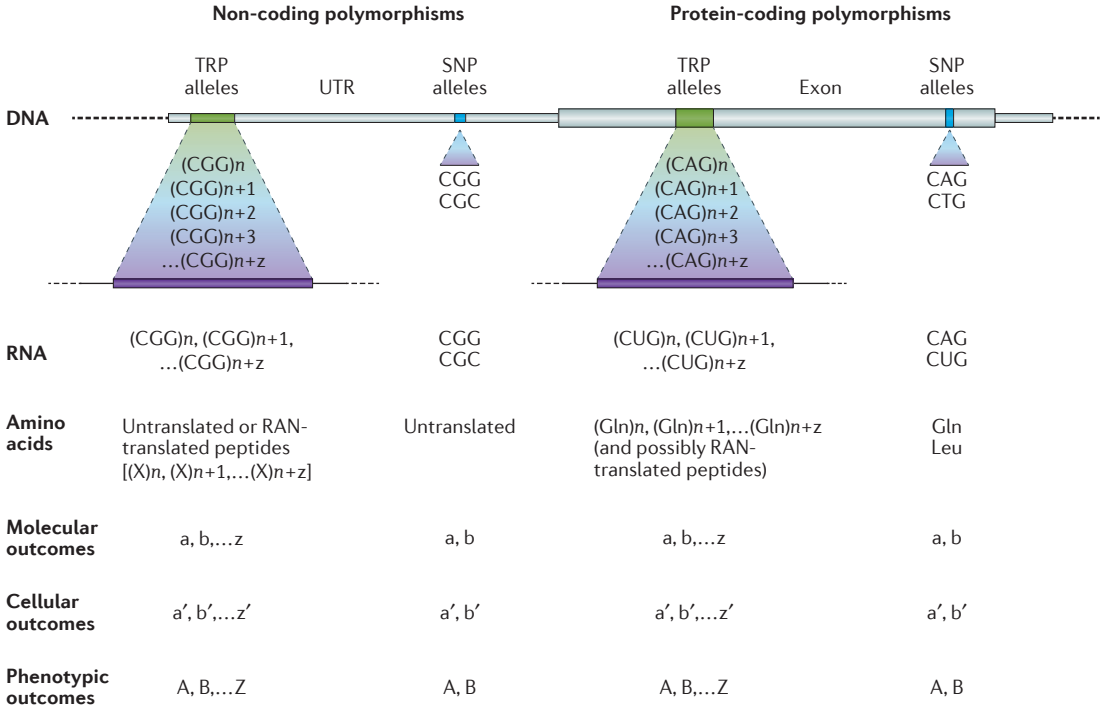


Figure 3 | **Tandem repeat polymorphisms as dynamic sources of phenotypic plasticity.** The comparative effects of tandem repeat polymorphisms (TRPs) and single nucleotide polymorphisms (SNPs) are illustrated. Examples of TRPs and SNPs that are located in either non-coding or coding regions are shown, demonstrating the potential impact of the extended digital (that is, multiallelic) distribution of TRPs versus the potential impact of binary (that is, biallelic) SNPs on molecular, cellular and phenotypic outcomes. In the examples of TRPs provided, if the smallest allele has *n* tandem repeats, then other alleles can have a wide (that is, extended digital) range of repeat lengths (up to *n* + z) and outcomes (A…Z). By comparison, a single nucleotide change to an equivalent trinucleotide leads to binary (that is, A or B) outcomes. Gln, glutamine; Leu, leucine; RAN, repeat-associated non-ATG; UTR, untranslated region.

nervous system in which somatic variability of tandem repeats could contribute to cellular selection and other developmental mechanisms[95].

Epigenetic changes may modulate the stability of tandem repeats. There is evidence that epigenetic factors can regulate the stability and mutational repair rates of tandem repeats[105–107], and this has been most clearly demonstrated for genes associated with TRDs[42]. The epigenetic regulation of tandem repeats may control their somatic mutability, which is a form of genetic plasticity, in response to environmental stimuli[106–108]. Indeed, environmental stimuli transduced as epigenetic signals have been shown to regulate the mutagenesis of trinucleotide repeats[108,109]; this type of mechanism might lead to long-term changes to cellular function in response to the environmental experience. The hypothesized role of the epigenetic modulation of tandem repeats would be most relevant to the brain, in which cognitive processes, such as learning and memory, require long-term changes to cell function in response to experience and the neural activity that is associated with it[95]. It is possible that the somatic mutation of tandem repeats in the developing and mature brain could be regulated epigenetically in response to experience-dependent modulation of neural activity. Such evolved neural functions of tandem repeats and their somatic mutability might also help to explain the preponderance of central nervous system diseases that dominate the known tandem repeat disorders. Furthermore, there is evidence that specific tandem repeat lengths can affect various epigenetic states and associated gene expression[110–114].

Tandem repeats and missing heritability

Following many disappointing association studies for complex polygenic disorders in which only a fraction of the expected disease alleles has been identified, discussion has focused on possible sources of missing heritability[115]. It has been proposed that repetitive DNA and tandem repeats in particular comprise a major component of the missing heritability associated with common polygenic disorders[79]. Testing this hypothesis will require improved sequencing and bioinformatic methodologies that are specifically designed to identify tandem repeats in exomes and genomes, which thus permit a systematic genome-wide approach to the cataloguing of tandem repeats in human genomes[4,79,116] (BOX 1).

Tandem repeats as major contributors to common polygenic disorders. There is much to be done in this relatively new field of TRD pathogenesis. The discovery

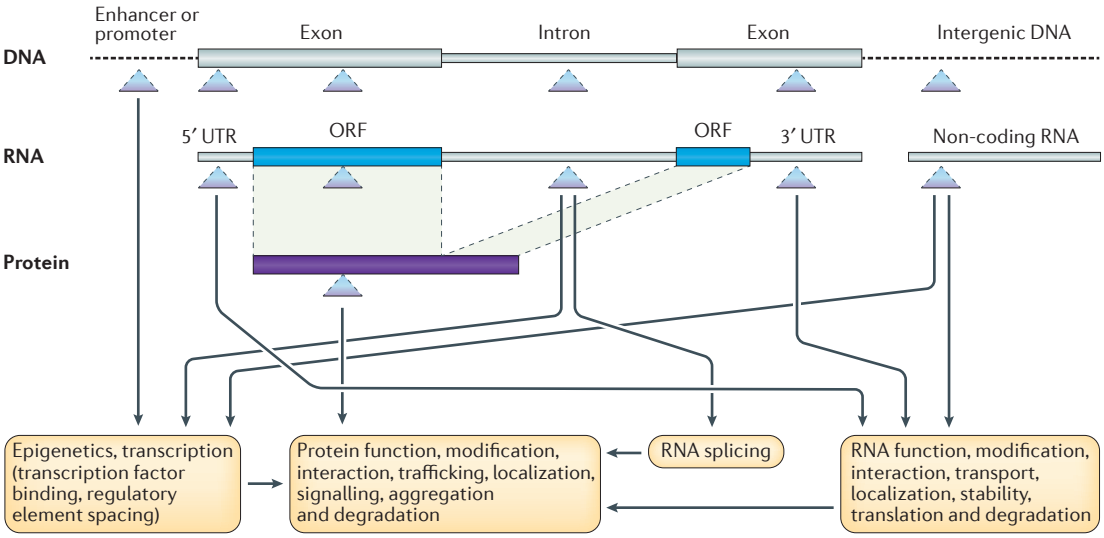


Figure 4 | **Impacts of tandem repeats on the structure and function of DNA, RNA and proteins.** Depending on their location within or between genes, tandem repeats, which are represented here by triangles, can have a variety of effects in healthy organisms at the DNA, RNA and protein levels. Tandem repeats in enhancers or promoters, as well as non-coding RNAs generated from tandem repeats, can affect epigenetic modifications and gene transcription, including transcriptional modulation via altered transcription factor binding. Tandem repeats may also be affected via the modulation of DNA methylation, chromatin states and non-coding RNA structure and function. Tandem repeats in 5′ or 3′ untranslated regions (UTRs) can affect RNA metabolism, including the transport, stability and translation of RNA. The subset of tandem repeats located in open reading frames (ORFs) can be translated into amino acid repeats, including homopeptides and longer motifs, within proteins, thus potentially affecting various aspects of protein structure and/or function. Finally, tandem repeats in introns can modify RNA splicing, which leads to changes at the protein level.

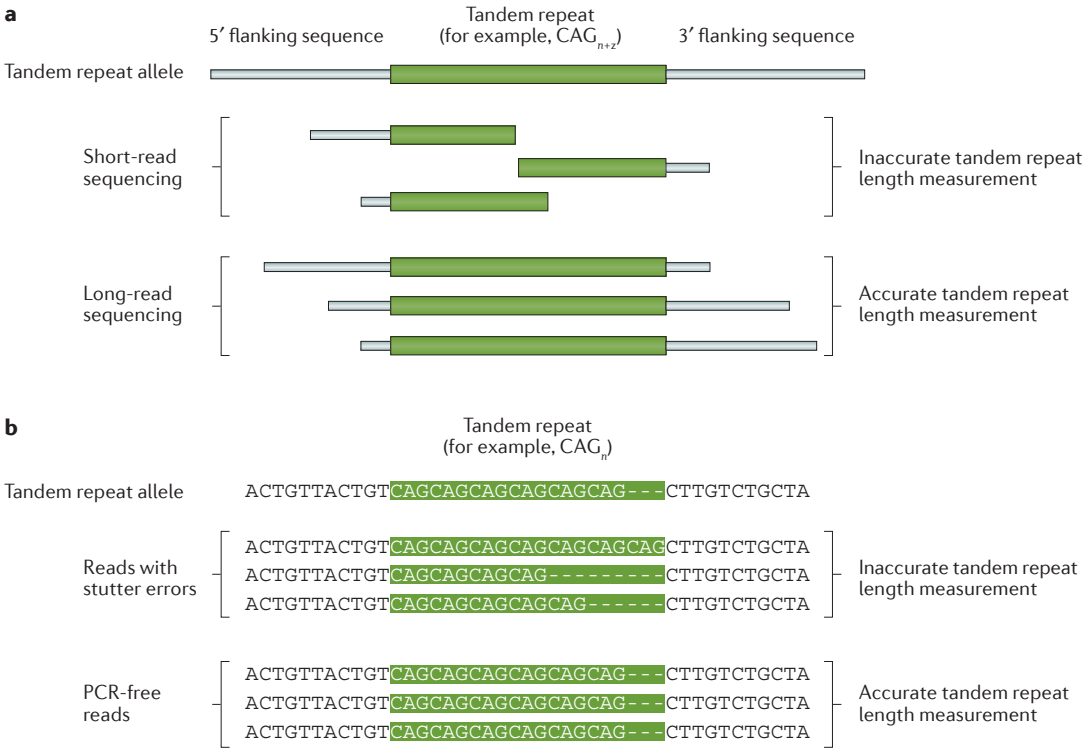
that C9ORF72 tandem repeat expansion has a major role in neurodegenerative diseases, along with other recent discoveries, suggests that ALS, FTD and the other known TRDs are the tip of the iceberg with respect to the role of tandem repeats in disease. The vast number of GWAS investigating the genetic contributions to traits and disorders have utilized microchips that detect SNPs or single nucleotide variants, which have no doubt led to important gene discoveries despite the remaining missing heritability[115]. However, these studies do not capture polymorphisms of the repeatome, particularly TRPs or tandem repeat variants[79]. Furthermore, as tandem repeats generally have higher spontaneous mutation rates than single nucleotides, SNPs do not reliably identify the majority of TRPs. Therefore, in order to discover TRPs that are associated with common polygenic diseases and non-disease traits in the healthy population, an approach that views tandem repeats as potentially functional polymorphisms, rather than genetic stutters or microsatellite markers, is required.

Tandem repeat polymorphisms associated with traits and endophenotypes. The field of tandem repeat genetics has focused largely on the association between tandem repeats and various human disorders (see above). However, well over a million discrete tandem repeats reside within our genomes, with a largely unexplored degree of polymorphic variability and heterogeneity[116]. Spatiotemporal expression patterns, together with structural and functional data, suggest that polymorphic tandem repeats also contribute to healthy brain development and function and associated complex traits[3,25,27,29,37,95,96]. In fact, in studying the genetics of neurological and psychiatric disorders, we appear to be looking at the pathological ends of phenotypic bell curves in which healthy individuals occupy the middle territory. TRPs have been associated with a wide variety of traits and endophenotypes underpinning human cognition, affective states and behaviours[3,25,27,33,37,79,95,117–128]. However, one major caveat of many of these previous studies is that they were underpowered, and therefore well-powered replication and follow-up studies, as well as systematic GWAS and meta-analyses, are needed. Despite this caveat, it is clear that aspects of individual differences, whether they relate to cognition, behaviour or other phenotypic features, can be associated with polymorphic variability in various tandem repeats. The extent and robustness of these genetic associations, as well as the mechanisms underlying how they modulate brain development and function, remain to be elucidated. Nevertheless, this field of research promises to deliver major insights into the genetic underpinnings of many neurological diseases, psychiatric disorders and other complex diseases. Furthermore, there are examples of TRPs modulating traits and endophenotypes in other species, such as the modulation of social and reproductive behaviour in voles[96] and the modulation of morphology between dog breeds[36]. A comparative genomics approach to tandem repeat evolution and function will be invaluable in discerning signal from

Box 1 | **Technical and bioinformatic challenges in genome-wide repeat-length sequencing**

The determination of tandem repeat length from whole exome and whole genome sequences requires solutions to various technical and bioinformatic challenges. First, the majority of current sequencing approaches (for example, Illumina technology) involve short reads, which often do not completely span tandem repeats (see the figure, part **a**)[130]. For example, a trinucleotide repeat with a repeat length of 120 (that is, 360 bp long) is not fully captured by a short read, as these capture only up to ~200 nucleotides, confounding accurate tandem repeat sequencing. Recently developed sequencing approaches, such as PacBio and Oxford Nanopore technologies, which can read thousands of nucleotides, could be used to capture tandem repeats and their correct alignment and possibly to provide information beyond sequence and repeat lengths, such as secondary structure[131]. Second, the use of PCR to amplify DNA before sequencing can introduce stutter errors, leading to inaccurate reporting of repeat lengths (see the figure, part **b**). More recent advances facilitate PCR-free sequencing, which can minimize this issue when trying to sequence tandem repeats genome-wide. Third, standard bioinformatic approaches, which have been routinely applied to whole exome and whole genome data sets, are inadequate in tandem repeat genomics because tandem repeats often lead to increases in local sequence misalignment. A solution to this problem is the use of repeat-aware bioinformatic software, such as LobSTR[78], MIPSTER[132], popSTR[133], HipSTR[134], RepeatSeq[135], STRviper[136], STR-FM[137] and other approaches[138,139], which have been recently reviewed[116]. Systematic comparison of these evolving software tools will facilitate their optimization for ongoing studies of tandem repeats in health and disease. Part of the bioinformatic challenge is to be able to capture both pure and interrupted tandem repeats and to be able to infer mutation rates. Solving such problems could inform the trajectories of tandem repeat evolution and functional constraints[140–145], as well as polygenic and epistatic interactions[146,147].

As the importance of tandem repeats in health and disease is more widely recognized, there will be increased impetus to capture all tandem repeats within genomes. One way to achieve this without resequencing whole genomes is to utilize capture techniques that specifically capture and sequence all tandem repeats[148,149]. Another key challenge in bioinformatics is to be able to detect and characterize mutations in somatic tandem repeats, which can be difficult to analyse in tissue samples and other large populations of somatic cells, especially for low-frequency somatic tandem repeat mutations. Finally, characterizing the tandem repeat transcriptome and tandem repeat proteome presents its own technical challenges, and a range of different bioinformatic, biophysical and biochemical approaches are required to advance the field of tandem repeat functional genomics[150].



noise and in extending the accumulated evidence that many tandem repeats have evolved as digital genetic modulators[95] rather than simply as genetic stutters.

Conclusions and future directions

Tandem repeats first drew the attention of the human genetics community in the early 1990s with the realization that diseases such as FXS, SBMA and Huntington disease were caused by repeat expansions. However, the discovery that dynamic mutations caused these human disorders did little to counteract the commonly held view that most tandem repeats were simply microsatellite markers that reflect the fallibility of DNA replication. The revolutions in genome sequencing and

functional genomics have provided evidence that many tandem repeats are functional sequences with complex roles at the DNA, RNA and protein levels, many of which we have yet to fully comprehend.

Despite the plethora of SNP-based GWAS investigating the genetics of traits and diseases, the potential functional roles of the repeatome (which constitutes approximately half of the genome), and tandem repeats in particular, have been largely overlooked. Comparative genomics, both between individuals and between species, indicates that many tandem repeats might have evolved under tight selective restraints, suggesting that they have highly conserved structures and functions. In particular, tandem repeats can fine-tune gene expression as well as the functions of RNA and proteins. Roles for tandem repeats in modulating diverse biological functions in healthy organisms are being discovered. Furthermore, it is possible that genome-wide investigations of complex traits and disorders will reveal that tandem repeats help to explain missing heritability[79].

One future direction will be to fully catalogue tandem repeats across all extant species to gain insights into their evolved biological functions. It is possible that key evolutionary milestones, such as the emergence of nervous systems, vertebrates and mammals, might be reflected in shifts in the structures, sizes and functions of tandem repeats across the genome and their variability, for example, within human populations and between wild and domesticated populations of animal and plant species. This comparative repeatomics could lead to a paradigm shift in the perception of tandem repeats as functional and multiallelic modulators. An extension of this work, which would focus on interspecies and germline variability, is to define the extent of tandem repeat mutation in somatic cells of developing and adult tissues. Somatic mutation of tandem repeats could mediate intercellular selection during development, particularly during brain development[95,101,103].

It will also be important to explore how the somatic mutation of tandem repeats influences downstream epigenetics and, conversely, how epigenetic modifications that mediate environmental signals modulate somatic tandem repeats. This will affect our understanding of gene–environment interactions and the role of both somatic epigenetic and genetic variability in developing and adult organisms. This in turn could lead to another paradigm shift, whereby tandem repeat modulation is seen not only as a genetic source of disease risk but also as a highly evolved and healthy form of adaptive mutation.

An additional priority for future repeatome research is to fully catalogue the polymorphic extent of tandem repeats in health and disease. An overarching hypothesis is that TRPs will be associated with, and causatively involved in, a large number and wide range of human traits and diseases. Tandem repeats have been discovered to be both mediators and modulators of disease states. In order to systematically test this hypothesis and potentially discover substantial missing heritability, we require a number of technical advances, some of which have been discussed elsewhere[4,79]. First, new sequencing technologies that ensure that tandem repeats are accurately sequenced at a genome-wide scale are needed. Second, bioinformatic approaches are required that incorporate the extended digital (that is, multiallelic) nature of TRPs[95], as opposed to the majority of binary (that is, biallelic) SNPs (FIG. 4), to draw valid statistical conclusions from sequence-based GWAS (BOX 1). Third, the incorporation of all aspects of genomic variability, whether SNPs, TRPs or other variants, into sophisticated and dynamic databases will be required in order to allow the international community to seamlessly exchange and interpret information. A comprehensive understanding of the roles of tandem repeats in disease will then facilitate their selective targeting with novel therapeutic approaches (BOX 2).

Twenty-first-century genomics promises to revolutionize human health and medicine. The utilization of personalized genomics and precision medicine, supported by an understanding of gene–environment interactions and 'enviromics'[129], has enormous potential in facilitating novel approaches to preventive strategies, treatments and cures. Historically, we have been limited in our capacity to accurately investigate all forms of repetitive DNA and its transcriptomic and proteomic counterparts as well as its functional relevance. It is now time to shine a light on this area to illuminate the many mysteries of tandem repeat biology in health and disease.

Box 2 | **Tandem repeats as therapeutic targets**

As it becomes clearer that tandem repeats contribute to a range of disorders, we must question how tandem repeats can be targeted to prevent and treat diseases. The most obvious approach for disorders driven by autosomal dominant and toxic gain-of-function mechanisms, such as Huntington disease, is to silence the mutant allele. Silencing could be attempted using antisense oligonucleotides, RNA interference or gene editing using CRISPR. CRISPR technology could also be used as a form of gene therapy to edit the mutant tandem repeat to a healthy length.

One challenge in attempting to silence or edit a tandem repeat-expanded allele is distinguishing between the mutant and healthy alleles, as the majority of patients will be heterozygous at the target locus and silencing or editing both alleles may have negative side effects. A second challenge is distinguishing the target tandem repeat from similar or identical tandem repeats located in other genes. Using patient-unique sequences that flank the tandem repeat being targeted could be one solution, although such personalized genomic approaches may present their own challenges given that traditional therapeutic development strategies have largely taken a 'one pill suits all' approach to a given disease.

As individual tandem repeat disorders involve specific post-transcriptional and post-translational mechanisms of pathogenesis, the mutant RNA and protein products could be therapeutically targeted. This may have advantages at the pharmaceutical level as RNA transcripts and proteins derived from expanded tandem repeats may have unique secondary or tertiary structures that can be targeted using traditional structural biology and medicinal chemistry approaches.

An additional challenge where the tissues of interest are not readily accessible for biopsies is targeting somatic mutations in tandem repeats. Therapeutic approaches based on tandem repeat lengths identified within readily accessible biosamples (such as blood and buccal cells) may not effectively target diseased tissues (such as the brain) in the same patient if this tissue harbours divergent tandem repeat lengths. Tandem repeat disorders that primarily involve somatic mutation, such as some cancers[66], could be treated by targeting the key molecules that control the accurate replication and repair of tandem repeats, such as mismatch repair proteins.

1. Hannan, A. J. Tandem repeat polymorphisms: mediators of genetic plasticity, modulators of biological diversity and dynamic sources of disease susceptibility. *Adv. Exp. Med. Biol.* **769**, 1–9 (2012).
2. Liang, K. C., Tseng, J. T., Tsai, S. J. & Sun, H. S. Characterization and distribution of repetitive elements in association with genes in the human genome. *Comput. Biol. Chem.* **57**, 29–38 (2015).
3. Fondon, J. W., Hammock, E. A., Hannan, A. J. & King, D. G. Simple sequence repeats: genetic modulators of brain function and behavior. *Trends Neurosci.* **31**, 328–334 (2008).
4. Press, M. O., Carlson, K. D. & Queitsch, C. The overdue promise of short tandem repeat variation for heritability. *Trends Genet.* **30**, 504–512 (2014).
5. Faux, N. Single amino acid and trinucleotide repeats: function and evolution. *Adv. Exp. Med. Biol.* **769**, 26–40 (2012).
6. Subramanian, S., Mishra, R. K. & Singh, L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol.* **4**, R13 (2003).
7. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
This study provides striking evidence that short tandem repeats are key regulators of gene expression in humans. Although the data are from lymphocytes, the implication is that this is likely to occur in other cell types as well.
8. Quilez, J. *et al.* Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* **44**, 3750–3762 (2016).
This study is a key demonstration that tandem repeats are important regulators of human gene expression and associated DNA methylation. Together with reference 7, it provides genome-wide evidence for a crucial role of tandem repeats as 'tuning knobs' that regulate gene expression.
9. Dejesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-Linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
10. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
This study and that described in reference 9 are the first demonstration that this GGGGCC hexanucleotide repeat expansion is a major genetic contributor to both ALS and FTD.
11. López Castel, A., Cleary, J. D. & Pearson, C. E. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat. Rev. Mol. Cell. Biol.* **11**, 165–170 (2010).
12. La Spada, A. R. & Taylor, J. P. Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.* **11**, 247–258 (2010).
13. Orr, H. T. Polyglutamine neurodegeneration: expanded glutamines enhance native functions. *Curr. Opin. Genet. Dev.* **22**, 251–255 (2012).
14. Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
15. Mo, C., Hannan, A. J. & Renoir, T. Environmental factors as modulators of neurodegeneration: insights from gene-environment interactions in Huntington's disease. *Neurosci. Biobehav. Rev.* **52**, 178–192 (2015).
16. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. Identification of genetic factors that modify clinical onset of Huntington's disease. *Cell* **162**, 516–526 (2014).
17. Bates, G. P. *et al.* Huntington disease. *Nat. Rev. Dis. Primers* **1**, 15005 (2015).
18. La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E. & Fischbeck, K. H. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79 (1991).
19. Rüb, U. *et al.* Clinical features, neurogenetics and neuropathology of the polyglutamine spinocerebellar ataxias type 1, 2, 3, 6 and 7. *Prog. Neurobiol.* **104**, 38–66 (2013).
20. Wilburn, B. *et al.* An antisense CAG repeat transcript at JPH3 locus mediates expanded polyglutamine protein toxicity in Huntington's disease-like 2 mice. *Neuron* **70**, 427–440 (2011).
21. Moseley, M. L. *et al.* Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8. *Nat. Genet.* **38**, 758–769 (2006).
22. Zu, T. *et al.* Non-ATG-initiated translation directed by microsatellite expansions. *Proc. Natl Acad. Sci. USA* **108**, 260–265 (2011).
This study describes the first description of RAN translation that does not require ATG initiation and that can produce amino acid repeat peptides that may contribute to dysfunction in various tandem repeat disorders.
23. Pearson, C. E. Repeat associated non-ATG translation initiation: one DNA, two transcripts, seven reading frames, potentially nine toxic entities! *PLOS Genet.* **7**, e1002018 (2011).
24. Cleary, J. D. & Ranum, L. P. New developments in RAN translation: insights from multiple diseases. *Curr. Opin. Genet. Dev.* **44**, 125–134 (2017).
25. Nopoulos, P., Epping, E. A., Wassink, T., Schlaggar, B. L. & Perlmutter, J. Correlation of CAG repeat length between the maternal and paternal allele of the Huntingtin gene: evidence for assortative mating. *Behav. Brain Funct.* **7**, 45 (2011).
26. Rubinsztein, D. C., Amos, B. & Cooper, G. Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. *Phil. Trans. R. Soc.* **354**, 1095–1099 (1999).
27. Mühlau, M. *et al.* Variation within the Huntington's disease gene influences normal brain structure. *PLOS ONE* **7**, e29809 (2012).
28. Lee, J. K. *et al.* Sex-specific effects of the Huntington gene on normal neurodevelopment. *J. Neurosci. Res.* **95**, 398–408 (2017).
29. Zheng, S. *et al.* Deletion of the huntingtin polyglutamine stretch enhances neuronal autophagy and longevity in mice. *PLOS Genet.* **6**, e1000838 (2010).
30. Perlis, R. H. *et al.* Prevalence of incompletely penetrant Huntington's disease alleles among individuals with major depressive disorder. *Am. J. Psychiatry* **167**, 574–579 (2010).
31. Gardiner, S. L. et. al. Huntingtin gene repeat size variations affect risk of lifetime depression. *Transl Psychiatry* **7**, 1277 (2017).
32. Du, X., Pang, T. Y. & Hannan, A. J. A tale of two maladies? Pathogenesis of depression with and without the Huntington's disease gene mutation. *Front. Neurol.* **4**, 81 (2013).
33. Gardiner, S. L. *et al.* Large normal-range TBP and ATXN7 CAG repeat lengths are associated with increased lifetime risk of depression. *Transl Psychiatry* **7**, e1143 (2017).
34. Shoubridge, C. & Gecz, J. Polyalanine tract disorders and neurocognitive phenotypes. *Adv. Exp. Med. Biol.* **769**, 185–203 (2012).
35. Hughes, J. N. & Thomas, P. Q. Molecular pathology of polyalanine expansion disorders: new perspectives from mouse models. *Methods Mol. Biol.* **1017**, 135–151 (2013).
36. Fondon, J. W. & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl Acad. Sci. USA* **101**, 18058–18063 (2004).
This study provides evidence that tandem repeats and their encoded polyalanine tracts in proteins are implicated in the development, evolution and morphology of dogs.
37. Nasrallah, M. P. *et al.* Differential effects of a polyalanine tract expansion in Arx on neural development and gene expression. *Hum. Mol. Genet.* **21**, 1090–1098 (2012).
38. Polling, S. *et al.* Polyalanine expansions drive a shift into α-helical clusters without amyloid-fibril formation. *Nat. Struct. Mol. Biol.* **22**, 1008–1015 (2015).
39. Campuzano, V. *et al.* Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271**, 1423–1427 (1996).
40. Evans-Galea, M. V., Lockhart, P. J., Galea, C. A., Hannan, A. J. & Delatycki, M. B. Beyond loss of frataxin: the complex molecular pathology of Friedreich ataxia. *Discov. Med.* **17**, 25–35 (2014).
41. Matsuura, T. *et al.* Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat. Genet.* **26**, 191–194 (2000).
42. Evans-Galea, M. V., Hannan, A. J., Carrodus, N., Delatycki, M. B. & Saffery, R. Epigenetic modifications in trinucleotide repeat diseases. *Trends Mol. Med.* **19**, 655–663 (2013).
43. Kremer, E. J. *et al.* Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science* **252**, 1711–1714 (1991).
44. Verkerk, A. J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
45. Sutcliffe, J. S. *et al.* DNA methylation represses FMR-1 transcription in fragile X syndrome. *Hum. Mol. Genet.* **1**, 397–400 (1992).
46. Hagerman, R. J. *et al.* Intention tremor, parkinsonism, and generalized brain atrophy in male carriers of fragile X. *Neurology* **57**, 127–130 (2001).
47. Loesch, D. & Hagerman, R. Unstable mutations in the FMR1 gene and the phenotypes. *Adv. Exp. Med. Biol.* **769**, 78–114 (2012).
48. Gecz, J., Gedeon, A. K., Sutherland, G. R. & Mulley, J. C. Identification of the gene FMR2, associated with FRAXE mental retardation. *Nat. Genet.* **13**, 105–108 (1996).
49. Gu, Y., Shen, Y., Gibbs, R. A. & Nelson, D. L. Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. *Nat. Genet.* **13**, 109–113 (1996).
50. Ciesiolka, A., Jazurek, M., Drazkowska, K. & Krzyzosiak, W. J. Structural characteristics of simple RNA repeats associated with disease and their deleterious protein interactions. *Front. Cell. Neurosci.* **11**, 97 (2017).
51. Fu, Y. H. *et al.* An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* **255**, 1256–1258 (1992).
52. Liquori, C. L. *et al.* Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science* **293**, 864–867 (2001).
53. Zu, T. *et al.* RAN translation regulated by muscleblind proteins in myotonic dystrophy type 2. *Neuron* **95**, 1292–1305.e5 (2017).
54. Hannan, A. J. (ed.) *Tandem Repeat Polymorphisms: Genetic Plasticity, Neural Diversity and Disease* Vol. 769, 208 (Springer-Verlag, 2012).
55. Van Eyk, C. L. & Richards, R. I. Dynamic mutations: where are we now? *Adv. Exp. Med. Biol.* **769**, 55–77 (2012).
56. Batra, R., Charizanis, K. & Swanson, M. S. Partners in crime: bidirectional transcription in unstable microsatellite disease. *Hum. Mol. Genet.* **19**, R77–R82 (2010).
57. Haeusler, A. R., Donnelly, C. J. & Rothstein, J. D. The expanding biology of the C9orf72 nucleotide repeat expansion in neurodegenerative disease. *Nat. Rev. Neurosci.* **17**, 383–395 (2016).
58. Elden, A. C. *et al.* Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature* **466**, 1069–1075 (2010).
59. Daoud, H. *et al.* Association of long ATXN2 CAG repeat sizes with increased risk of amyotrophic lateral sclerosis. *Arch. Neurol.* **68**, 739–742 (2011).
60. van Blitterswijk, M. *et al.* Ataxin-2 as potential disease modifier in C9ORF72 expansion carriers. *Neurobiol. Aging* **35**, 2421.e13–2421.e17 (2014).
61. Majounie, E. *et al.* Repeat expansion in C9ORF72 in Alzheimer's disease. *N. Engl. J. Med.* **366**, 283–284 (2012).
62. Zu, T. *et al.* RAN proteins and RNA foci from antisense transcripts in C9ORF72 ALS and frontotemporal dementia. *Proc. Natl Acad. Sci. USA* **110**, E4968–E4977 (2013).
63. Kobayashi, H. *et al.* Expansion of intronic GGCCTG hexanucleotide repeat in *NOP56* causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. *Am. J. Hum. Genet.* **89**, 121–130 (2011).
64. Lafrenière, R. G. *et al.* Unstable insertion in the 5′ flanking region of the cystatin B gene is the most common mutation in progressive myoclonus epilepsy type 1, EPM1. *Nat. Genet.* **15**, 298–302 (1997).
65. Virtaneva, K. *et al.* Unstable minisatellite expansion causing recessively inherited myoclonus epilepsy, EPM1. *Nat. Genet.* **15**, 393–396 (1997).
66. Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat. Med.* **22**, 1342–1350 (2016).
This study demonstrates that tandem repeat instability, particularly that of STRs, is a major contributor to a variety of different cancers. This finding provides insights into the extent of somatic instability of tandem repeats and is a major incentive to improve sequencing and bioinformatics approaches to capture all tandem repeat mutations in oncological disorders, which may provide novel therapeutic targets.
67. Gebhardt, F., Zänker, K. S. & Brandt, B. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J. Biol. Chem.* **274**, 13176–13180 (1999).

68. Shimajiri, S. *et al.* Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett.* **455**, 70–74 (1999).
69. Contente, A., Dittmer, A., Koch, M. C., Roth, J. & Dobbelstein, M. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nat. Genet.* **30**, 315–320 (2002).
70. King, D. G. Evolution of simple sequence repeats as mutable sites. *Adv. Exp. Med. Biol.* **769**, 10–25 (2012).
71. Sawaya, S. *et al.* Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLOS ONE* **8**, e54710 (2013).
72. Sawaya, S. M., Bagshaw, A. T., Buschiazzo, E. & Gemmel, N. J. Promoter microsatellites as modulators of human gene expression. *Adv. Exp. Med. Biol.* **769**, 41–44 (2012).
73. Hsieh, T. Y. *et al.* Molecular pathogenesis of Gilbert's syndrome: decreased TATA-binding protein binding affinity of *UGT1A1* gene promoter. *Pharmacogenet. Genomics* **17**, 229–236 (2007).
74. Borel, C. *et al.* Tandem repeat sequence variation as causative *cis*-eQTLs for protein-coding gene expression variation: the case of CSTB. *Hum. Mutat.* **33**, 1302–1309 (2012).
75. Rockman, M. V. & Wray, G. A. Abundant raw material for *cis*-regulatory evolution in humans. *Mol. Biol. Evol.* **19**, 1991–2004 (2002).
76. Rothenburg, S., Koch-Nolte, F., Rich, A. & Haag, F. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc. Natl Acad. Sci. USA* **98**, 8985–8990 (2001).
77. Stöger, R., Kajimura, T. M., Brown, W. T. & Laird, C. D. Epigenetic variation illustrated by DNA methylation patterns of the fragile-X gene FMR1. *Hum. Mol. Genet.* **6**, 1791–1801 (1997).
78. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 1154–1162 (2012).
79. Hannan, A. J. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet.* **26**, 59–65 (2010).
This article includes the first proposal that tandem repeats, and their polymorphic variants, can help explain the missing heritability associated with complex polygenic disorders. It also proposes key roles of somatic tandem repeat variability and predicts a new era of tandem repeat associations in human genetics.
80. Hefferon, T. W., Groman, J. D., Yurk, C. E. & Cutting, G. R. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proc. Natl Acad. Sci. USA* **101**, 3504–3509 (2004).
81. Hui, J. *et al.* Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J.* **24**, 1988–1998 (2005).
82. Santos-Pereira, J. M. & Aguilera, A. R loops: new modulators of genome dynamics and function. *Nat. Rev. Genet.* **16**, 583–597 (2015).
83. Schmidt, M. H. & Pearson, C. E. Disease-associated repeat instability and mismatch repair. *DNA Repair* **38**, 117–126 (2016).
84. Jain, A. & Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243–247 (2017).
85. Schaefer, M. H., Wanker, E. E. & Andrade-Navarro, M. A. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. *Nucleic Acids Res.* **40**, 4273–4287 (2012).
86. Pellegrini, M. Tandem repeats in proteins: prediction algorithms and biological role. *Front. Bioeng. Biotechnol.* **3**, 143 (2015).
87. Kumar, A. S., Sowpati, D. T. & Mishra, R. K. Single amino acid repeats in the proteome world: structural, functional, and evolutionary insights. *PLOS ONE* **11**, e0166854 (2016).
88. Willems, T. *et al.* Population-scale sequencing data enable precise estimates of Y-STR mutation rates. *Am. J. Hum. Genet.* **98**, 919–933 (2016).
89. O'Dushlaine, C. T. & Shields, D. C. Marked variation in predicted and observed variability of tandem repeat loci across the human genome. *BMC Genomics* **9**, 175 (2008).
90. Payseur, B. A., Jing, P. & Haasl, R. J. A genomic portrait of human microsatellite variation. *Mol. Biol. Evol.* **28**, 303–312 (2011).
91. McIver, L. J., Fondon, J. W., Skinner, M. A. & Garner, H. R. Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics* **97**, 193–199 (2011).
92. McIver, L. J., McCormick, J. F., Martin, A., Fondon, J. W. & Garner, H. R. Population-scale analysis of human microsatellites reveals novel sources of exonic variation. *Gene* **516**, 328–334 (2013).
93. Willems, T. *et al.* The landscape of human STR variation. *Genome Res.* **24**, 1894–1904 (2014).
94. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
95. Nithianantharajah, J. & Hannan, A. J. Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. *Bioessays* **29**, 525–535 (2007).
This paper proposes that tandem repeats serve novel functions in evolution, ontology, neural development and disease. It is the first proposal of a key role of somatic tandem repeat variability in development, particularly that of the brain and associated aspects of behaviour and cognition.
96. Verstrepen, K. J., Jansen, A., Lewitter, F. & Fink, G. R. Intragenic tandem repeats generate functional variability. *Nat. Genet.* **37**, 986–990 (2005).
97. Hammock, E. A. & Young, L. J. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* **308**, 1630–1634 (2005).
The authors of this study demonstrate that a specific short tandem repeat can regulate neural systems underlying social behaviour using monogamous and polygamous voles as model species.
98. Matsushima, N. *et al.* Flexible structures and ligand interactions of tandem repeats consisting of proline, glycine, asparagine, serine, and/or threonine rich oligopeptides in proteins. *Curr. Protein Pept. Sci.* **9**, 591–610 (2008).
99. Van Eyk, C. L., McLeod, C. J., O'Keefe, L. V. & Richards, R. I. Comparative toxicity of polyglutamine, polyalanine and polyleucine tracts in *Drosophila* models of expanded repeat disease. *Hum. Mol. Genet.* **21**, 536–547 (2012).
100. Gonitel, R. *et al.* DNA instability in postmitotic neurons. *Proc. Natl Acad. Sci. USA* **105**, 3467–3472 (2008).
101. McMurray, C. T. Mechanisms of trinucleotide repeat instability during human development. *Nat. Rev. Genet.* **11**, 786–799 (2010).
102. Lee, J. M., Pinto, R. M., Gillis, T., St Claire, J. C. & Wheeler, V. C. Quantification of age-dependent somatic CAG repeat instability in Hdh CAG knock-in mice reveals different expansion dynamics in striatum and liver. *PLOS ONE* **6**, e23647 (2011).
103. Lokanga, R. A. *et al.* Somatic expansion in mouse and human carriers of fragile X premutation alleles. *Hum. Mutat.* **34**, 157–166 (2013).
104. Richardson, S. R., Morell, S. & Faulkner, G. J. L1 retrotransposons and somatic mosaicism in the brain. *Annu. Rev. Genet.* **48**, 1–27 (2014).
105. Dion, V., Lin, Y., Hubert, L., Waterland, R. A. & Wilson, J. H. Dnmt1 deficiency promotes CAG repeat expansion in the mouse germline. *Hum. Mol. Genet.* **17**, 1306–1317 (2008).
106. Libby, R. T. *et al.* CTCF *cis*-regulates trinucleotide repeat instability in an epigenetic manner: a novel basis for mutational hot spot determination. *PLOS Genet.* **4**, e1000257 (2008).
107. Slean, M. M., Panigrahi, G. B., Ranum, L. P. & Pearson, C. E. Mutagenic roles of DNA "repair" proteins in antibody diversity and disease-associated trinucleotide repeat instability. *DNA Repair* **7**, 1135–1154 (2008).
108. Fonville, N. C., Ward, R. M. & Mittelman, D. Stress-induced modulators of repeat instability and genome evolution. *J. Mol. Microbiol. Biotechnol.* **21**, 36–44 (2011).
109. Chatterjee, N., Lin, Y., Santillan, B. A., Yotnda, P. & Wilson, J. H. Environmental stress induces trinucleotide repeat mutagenesis in human cells. *Proc. Natl Acad. Sci. USA* **112**, 3764–3769 (2015).
110. Greene, E., Mahishi, L., Entezam, A., Kumari, D. & Usdin, K. Repeat-induced epigenetic changes in intron 1 of the frataxin gene and its consequences in Friedreich ataxia. *Nucleic Acids Res.* **35**, 3383–3390 (2007).
111. Ruan, H., Wang, Y. H. Friedrich's ataxia GAA.TTC duplex and GAA.GAA.TTC triplex structures exclude nucleosome assembly. *J. Mol. Biol.* **383**, 292–300 (2008).
112. Evans-Galea, M. V. *et al.* FXN methylation predicts expression and clinical outcome in Friedreich ataxia. *Ann. Neurol.* **71**, 487–497 (2012).
113. Colak, D. *et al.* Promoter-bound trinucleotide repeat mRNA drives epigenetic silencing in fragile X syndrome. *Science* **343**, 1002–1005 (2014).
114. Pretto, D. I. *et al.* CGG allele size somatic mosaicism and methylation in FMR1 premutation alleles. *J. Med. Genet.* **51**, 309–318 (2014).
115. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
116. Gymrek, M. A genomic view of short tandem repeats. *Curr. Opin. Genet. Dev.* **44**, 9–16 (2017).
117. Cerasa, A. *et al.* MAO A VNTR polymorphism and amygdala volume in healthy subjects. *Psychiatry Res.* **191**, 87–91 (2011).
118. Grube, S. *et al.* A CAG repeat polymorphism of KCNN3 predicts SK3 channel function and cognitive performance in schizophrenia. *EMBO Mol. Med.* **3**, 309–319 (2011).
119. Kopf, J. *et al.* NOS1 ex1f-VNTR polymorphism influences prefrontal brain oxygenation during a working memory task. *Neuroimage* **57**, 1617–1623 (2011).
120. Squitieri, F., Esmaeilzadeh, M., Ciarmiello, A. & Jankovic, J. Caudate glucose hypometabolism in a subject carrying an unstable allele of intermediate CAG(33) repeat length in the Huntington's disease gene. *Mov. Disord.* **26**, 925–927 (2011).
121. Sonuga-Barke, E. J. *et al.* A functional variant of the serotonin transporter gene (SLC6A4) moderates impulsive choice in attention-deficit/hyperactivity disorder boys and siblings. *Biol. Psychiatry* **70**, 230–236 (2011).
122. Van Holstein, M. *et al.* Human cognitive flexibility depends on dopamine D2 receptor signaling. *Psychopharmacology* **218**, 567–578 (2011).
123. Simmons, Z. L. & Roney, J. R. Variation in CAG repeat length of the androgen receptor gene predicts variables associated with intrasexual competitiveness in human males. *Horm. Behav.* **60**, 306–312 (2011).
124. Abshire, M. Y. *et al.* Role of androgen receptor CAG repeat polymorphism length in hypothalamic progesterone sensitivity in hyperandrogenic adolescent girls. *Endocrine* **41**, 156–158 (2012).
125. Shumay, E. *et al.* Repeat variation in the human PER2 gene as a new genetic marker associated with cocaine addiction and brain dopamine D2 receptor availability. *Transl Psychiatry* **2**, e86 (2012).
126. Zilles, D. *et al.* Genetic polymorphisms of 5-HTT and DAT but not COMT differentially affect verbal and visuospatial working memory functioning. *Eur. Arch. Psychiatry Clin. Neurosci.* **262**, 667–676 (2012).
127. Eisenegger, C. *et al.* DAT1 polymorphism determines L-DOPA effects on learning about others' prosociality. *PLOS ONE* **8**, e67820 (2013).
128. Bagshaw, A. T., Horwood, L. J., Fergusson, D. M., Gemmell, N. J. & Kennedy, M. A. Microsatellite polymorphisms associated with human behavioural and psychological phenotypes including a gene-environment interaction. *BMC Med. Genet.* **18**, 12 (2017).
129. McOmish, C. E., Burrows, E. L. & Hannan, A. J. Identifying novel interventional strategies for psychiatric disorders: integrating genomics, 'enviromics' and gene-environment interactions in valid preclinical models. *Br. J. Pharmacol.* **171**, 4719–4728 (2014).
130. Zavodna, M., Bagshaw, A., Brauning, R. & Gemmell, N. J. The accuracy, feasibility and challenges of sequencing short tandem repeats using next-generation sequencing platforms. *PLOS ONE* **9**, e113862 (2014).
131. Sawaya, S., Boocock, J., Black, M. A. & Gemmell, N. J. Exploring possible DNA structures in real-time polymerase kinetics using Pacific Biosciences sequencer data. *BMC Bioinformatics* **16**, 21 (2015).
132. Carlson, K. D. *et al.* MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Res.* **25**, 750–761 (2015).
133. Kristmundsdóttir, S., Sigurpálsdóttir, B. D., Kehr, B. & Halldórsson, B. V. popSTR: population-scale detection of STR variants. *Bioinformatics* **33**, 4041–4048 (2016).
134. Willems, T. *et al.* Genome-wide profiling of heritable and *de novo* STR variations. *Nat. Methods* **14**, 590–592 (2017).
135. Highnam, G. *et al.* Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res.* **41**, e32 (2013).

136. Cao, M. D. *et al.* Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Res.* **42**, e16 (2014).
137. Fungtammasan, A. *et al.* Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res.* **25**, 736–749 (2015).
138. Anisimova, M., Pecerska, J. & Schaper, E. Statistical approaches to detecting and analyzing tandem repeats in genomic sequences. *Front. Bioeng. Biotechnol.* **3**, 31 (2015).
139. Gelfand, Y., Hernandez, Y., Loving, J. & Benson, G. VNTRseek-a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic Acids Res.* **42**, 8884–8894 (2014).
140. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44**, 1161–1165 (2012).
141. Fondon, J. W., Martin, A., Richards, S., Gibbs, R. A. & Mittelman, D. Analysis of microsatellite variation in *Drosophila melanogaster* with population-scale genome sequencing. *PLOS ONE* **7**, e33036 (2012).
142. Ananda, G. *et al.* Microsatellite interruptions stabilize primate genomes and exist as population-specific single nucleotide polymorphisms within individual human genomes. *PLOS Genet.* **10**, e1004498 (2014).
143. Bilgin Sonay, T. *et al.* Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Res.* **25**, 1591–1599 (2015).
144. Abe, H. & Gemmell, N. J. Evolutionary footprints of short tandem repeats in avian promoters. *Sci. Rep.* **6**, 19421 (2016).
145. Shimada, M. K. *et al.* Selection pressure on human STR loci and its relevance in repeat expansion disease. *Mol. Genet. Genomics* **291**, 1851–1869 (2016).
146. Wray, N. R. *et al.* Research review: polygenic methods and their application to psychiatric traits. *J. Child. Psychol. Psychiatry* **55**, 1068–1087 (2014).
147. Sackton, T. B. & Hartl, D. L. Genotypic context and epistasis in individuals and populations. *Cell* **166**, 279–287 (2016).
148. Guilmatre, A., Highnam, G., Borel, C., Mittelman, D. & Sharp, A. J. Rapid multiplexed genotyping of simple tandem repeats using capture and high-throughput sequencing. *Hum. Mutat.* **34**, 1304–1311 (2013).
149. Duitama, J. *et al.* Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Res.* **42**, 5728–5741 (2014).
150. Hatters, D. M. & Hannan, A. J. (eds) *Tandem Repeats in Genes, Proteins and Disease: Methods and Protocols* Vol. 1017, 258 (Humana Press, 2013).

Acknowledgements
The author thanks C. Pearson for comments on an early outline of the manuscript and past and present members of the Hannan laboratory for useful discussions. Apologies to the authors of the many excellent relevant articles that could not be cited and discussed due to space constraints. The author is supported by a Principal Research Fellowship and Project Grants from the National Health and Medical Research Council (NHMRC), as well as the Australian Research Council (ARC) and DHB Foundation, Equity Trustees.

Competing interests statement
The author declares no competing interests.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

FURTHER INFORMATION
GenBank: http://www.ncbi.nlm.nih.gov/genbank
US National Center for Biotechnology Information (NCBI) genome resource: http://www.ncbi.nlm.nih.gov/genome
Online Mendelian Inheritance in Man (OMIM) database: http://omim.org/
HipSTR: https://hipstr-tool.github.io/HipSTR/
LobSTR: http://lobstr.teamerlich.org/
MIPSTR: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4417122/
PopSTR: https://github.com/DecodeGenetics/popSTR
RepeatSeq: http://github.com/adaptivegenome/repeatseq
STRviper: http://bioinf.scmb.uq.edu.au:8080/STRViper/
STR-FM: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4417121/
1000 Genomes Project: http://www.internationalgenome.org/home

ALL LINKS ARE ACTIVE IN THE ONLINE PDF