

Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences

Rita Gemayel,^{1,2,*} Marcelo D. Vences,^{1,2,*}
Matthieu Legendre,³ and Kevin J. Verstrepen^{1,2}

¹Laboratory for Systems Biology, VIB, B-3001 Heverlee, Belgium;
email: Kevin.Verstrepen@biw.vib-kuleuven.be

²Genetics and Genomics Group, Center of Microbial and Plant Genetics (CMPG),
K.U. Leuven, B-3001 Heverlee, Belgium

³Structural and Genomic Information Laboratory, CNRS, Université de la Méditerranée
Parc Scientifique de Luminy, FR-13288 Marseille, France

Annu. Rev. Genet. 2010. 44:445–77

First published online as a Review in Advance on
August 25, 2010

The *Annual Review of Genetics* is online at
genet.annualreviews.org

This article's doi:
10.1146/annurev-genet-072610-155046

Copyright © 2010 by Annual Reviews.
All rights reserved

0066-4197/10/1201-0445\$20.00

*These authors contributed equally to this work.

Key Words

evolvability, phenotype, satellite repeats, microsatellites, ataxia, SNP

Abstract

Genotype-to-phenotype mapping commonly focuses on two major classes of mutations: single nucleotide polymorphisms (SNPs) and copy number variation (CNV). Here, we discuss an underestimated third class of genotypic variation: changes in microsatellite and minisatellite repeats. Such tandem repeats (TRs) are ubiquitous, unstable genomic elements that have historically been designated as nonfunctional “junk DNA” and are therefore mostly ignored in comparative genomics. However, as many as 10% to 20% of eukaryotic genes and promoters contain an unstable repeat tract. Mutations in these repeats often have fascinating phenotypic consequences. For example, changes in unstable repeats located in or near human genes can lead to neurodegenerative diseases such as Huntington disease. Apart from their role in disease, variable repeats also confer useful phenotypic variability, including cell surface variability, plasticity in skeletal morphology, and tuning of the circadian rhythm. As such, TRs combine characteristics of genetic and epigenetic changes that may facilitate organismal evolvability.

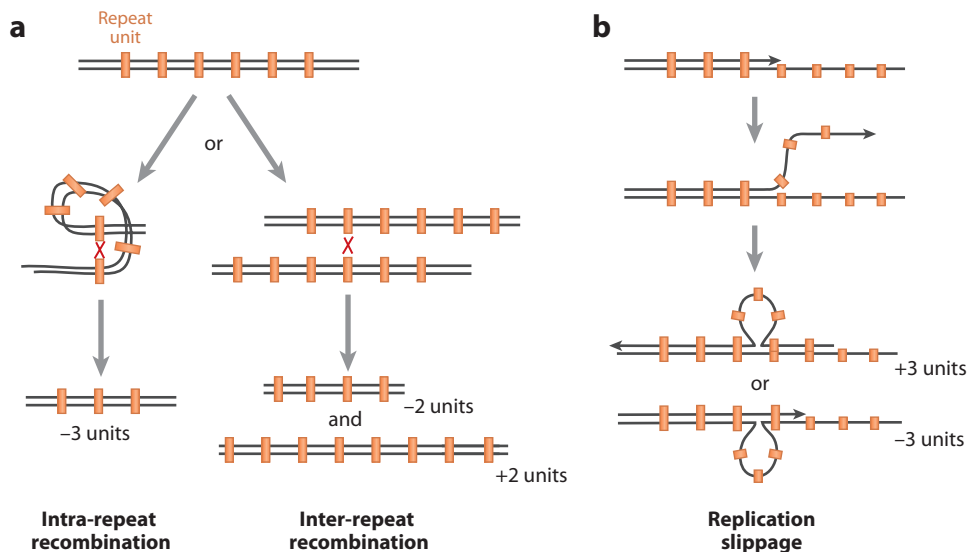


Figure 2

Mechanisms for tandem repeat (TR) expansions and contractions. Shown are extremely simplified illustrations of two major mechanisms of TR expansions and contractions. (a) Recombination. TR units can undergo intrachromosomal pairing and recombination that yields short alleles with a reduced number of TR units. Alternatively, TRs can be involved in unequal crossing over between chromosomes that can generate expanded or contracted alleles. (b) Strand slippage. Shown is a very simplified model of a replicative TR mutation mechanism. During DNA replication, single-stranded DNA containing repeat units can transiently dissociate and mispair, resulting in deletions or additions of repeat units. In reality, strand slippage mechanisms probably involve transient association of one strand with the other strand of the replication fork, and subsequent transient replication, as well as involvement of double-strand breaks and repair (for further details, see, for example, references 98, 135).

units changes. In other words, one or more repeat units are added or deleted. The vast majority of these changes consist of the addition or deletion of a number of full repeat units; additions or deletions of part of one unit are very rare.

There are currently two major models that describe the mechanisms by which TRs expand or contract: strand-slippage replication and recombination (Figure 2). Details of both mechanisms have been covered elsewhere (28, 98, 103, 118, 126). Briefly, strand-slippage replication, also known as slipped-strand mispairing, or DNA slippage, occurs during replication of the TR DNA when there is mispairing between the template and nascent DNA strands. When the newly synthesized strand denatures from the template strand during synthesis of the TR sequence, it will occasionally pair with

another part of the repeat sequence. If the template strand is looped out, then contraction of the TR occurs. If the nascent strand loops out, then an expansion will result. More elaborate models invoke DNA strand-breakage repair (98, 135). Double-strand breaks are generated during DNA replication, stalling at repeats. These breaks can be repaired by single-strand annealing to homologous TR sequences on the same branch of the replication fork. This would result in loss of repeat units. Single-strand annealing to homologous TR sequences on the other branch of the fork can also occur, and several TR units can be added to the strand before it returns to the broken replication fork. The importance of DNA strand breakage in the mutation of TR tracts is supported by studies demonstrating involvement of the double-strand

Repeat polymorphism: variability in repeat unit number

break repair pathway in TR expansion and contraction (98). Recombination events, including unequal crossing over and gene conversion, can also lead to contraction and expansions of TR sequences (**Figure 2**). It has been argued that recombination is more important in causing instability among minisatellites, and replicative mechanisms underlie microsatellite instability (103). Other studies (62, 118) suggest that replication-associated slippage mechanisms may underlie the majority of micro- and minisatellite instability. However, the precise molecular mechanism of slippage remains unclear. Several more detailed models have been proposed (for examples, see 28, 98, 103, 118, 126, 135), but it is uncertain which one or which ones are correct, what their relative contribution is, and how this depends on the repeat characteristics.

Although all TRs are inherently unstable, precise mutation rates vary greatly between TRs. Legendre and coworkers used machine learning techniques to develop a model that estimates the variability of a given repeat (68). Their analysis indicates that a repeat's stability depends mainly on its sequence. The most important factor defining a TR's variability is its number of repeat units: the more repeated units, the more unstable the TR. Other factors, such as repeat purity and the length of the repeat unit, also influence TR variability. In addition to a repeat's inherent instability, mutation rates of TR are also affected by external factors. For example, if a repeat tract is transcribed, higher transcription levels may lead to increased instability (143). Environmental factors may thus influence TR mutation rates. Another example is the contraction of a CT repeat in wheat that occurs after fungal infection (116) and short mononucleotide repeat tracts in bacteria, which are the frequent target of adaptive, environmentally-responsive mutations (108). Very recently it has been demonstrated that CAG repeat stability is modulated by the chaperone protein Hsp90 (84). The authors of Reference 84 show that a decrease in Hsp90 levels results in nearly tenfold increases in the rate of contraction of a

CAG repeat tract, while not affecting the rate of point mutations. As severe environmental stresses can overwhelm Hsp90 function, Hsp90 may be playing a role in mediating an influence by the environment on TR mutation rates.

Identification of Tandem Repeats

Whole genome sequencing now enables researchers to identify TRs quickly and at low cost. A comprehensive comparison and description of many of the existing TR detection algorithms and software have been previously published (80). One problem that arises in scanning a sequence for a TR is in formulating a definition for a true TR. Clearly, some small repeats with a small number of repeated units may not be biologically relevant. For example, GTACGTAC may occur many times in the genome just by chance. Additionally, TRs may be imperfect (interrupted by base substitutions, as in GTACGCAC), which also complicates search strategies. So how can we define a cutoff, a minimal requirement to call a sequence a TR? The numerous algorithms available to scan and identify DNA sequences for TRs (e.g., Tandem Repeat Finder, Sputnik, TROLL) often have specified thresholds that can be used to define what will be considered a TR. These thresholds are based on various repeat characteristics, including the number of repeated units, the length of a repeat unit, and the purity of the different repeated units. However, the exact scoring varies between algorithms, and thus different approaches may yield diverging TR frequencies as well as TRs of a different nature. The question now becomes how can we define a proper cutoff score?

One approach to defining what relevant TRs are in genomic DNA sequences is to use a TR-finding algorithm on several (e.g., 100 or 1,000) randomly shuffled sequences of the same size and composition as the sequence being analyzed. These genomes are searched for TRs using various different cutoff scores. The average number of TRs identified in the set of scrambled DNA sequences then serves as a value for background noise produced by

chance occurrence of TRs. Making the search criteria more stringent by increasing the cut-off score decreases the number of TRs found and increases the ratio of the number of repeats found in the real DNA sequence to the number of TRs found in the scrambled sequences. In many cases, a score at which 10 times more TRs are identified in the real DNA sequence than in the scrambled sequences can be considered a usable cutoff. The shuffling method is not only useful in determining whether, for example, a small repeat, such as ATAT, is meaningful, but also whether degenerate and impure repeats are significant.

In addition to sequence mining tools, a user-friendly and intuitive online applet (SERV) is available that allows users to identify TRs in DNA sequences and also predicts the variability of each repeat based on several parameters, such as repeat purity, unit size, and the number of repeat units (68). SERV can not only be used to find TRs in various sequences but also for selection of sufficiently variable TR markers for various applications, including forensics and population genetic analysis (see 68 for more details).

Tandem Repeats Are Often Located Within Genes and Regulatory Regions

While many repeats are located within gene deserts, the availability of complete genome sequences and increased knowledge about genome biology indicate that TRs also often occur within coding and regulatory regions (Figure 3) (Table 1) (68, 71). Approximately 17% of genes in the human genome contain repeats within their open reading frames (ORFs), and the percentage hovers around this figure in other organisms as well (Table 1). Among repeats found in coding sequences, repeats with units that contain a multiple of three nucleotides, such as tri- and hexanucleotide repeats, are by far the most common, presumably because of selection against frequent frameshift mutations that would occur when repeat units would not contain a multiple of three nucleotides (68, 81) (Figure 4). Interest-

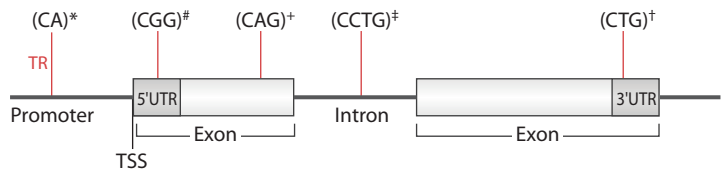


Figure 3

Location of tandem repeats (TRs) relative to a gene and examples. A hypothetical and simplified gene illustrates the locations where TRs are found in relation to genes in the genome. The different parts of a gene region are labeled, and red lines indicate examples of TRs found in each region, with one repeat unit shown between parentheses. Examples of each of these repeats are described below. *Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia (124). #Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in Fragile X syndrome (134). +A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington disease chromosomes (50). ‡Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of *ZNF9* (72). †Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member (9).

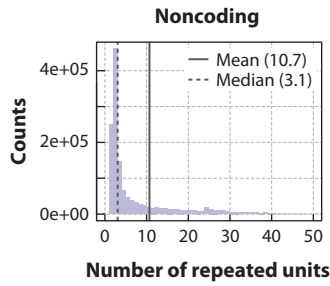
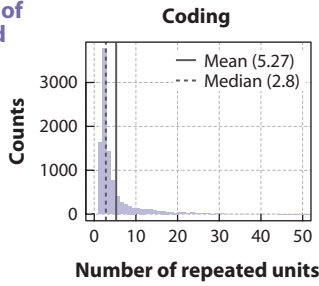
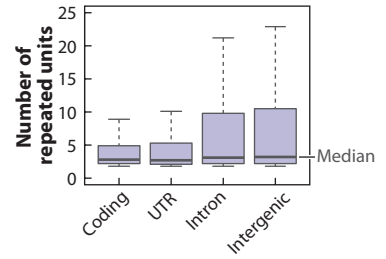
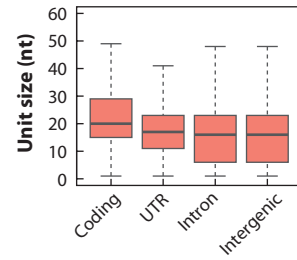
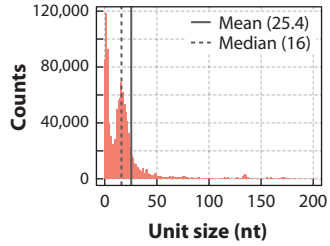
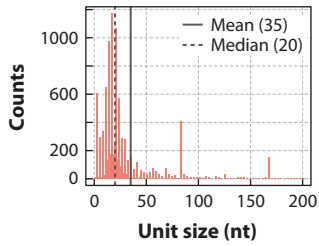
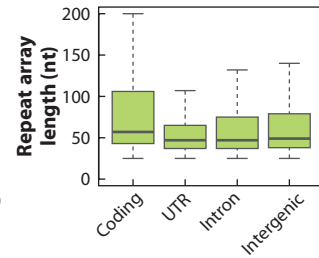
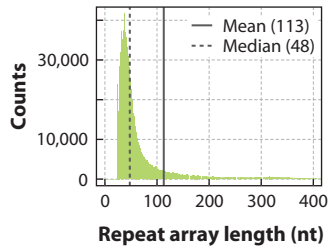
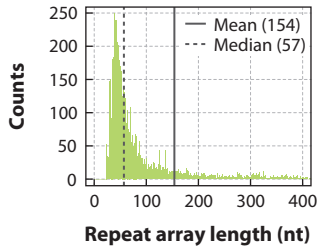
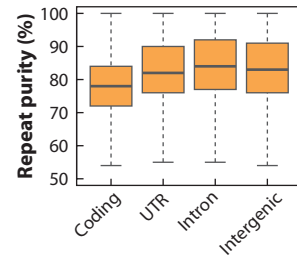
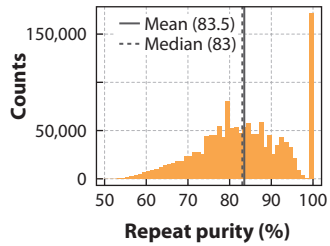
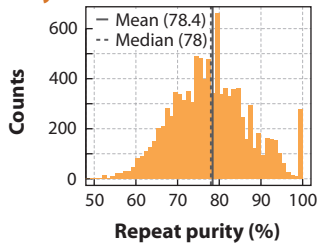
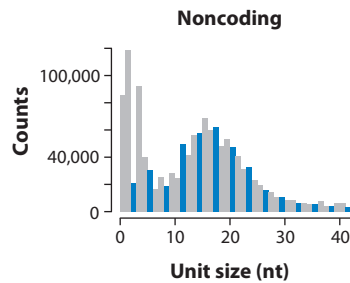
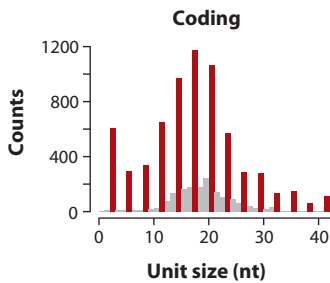
ingly, TRs are not uniformly represented in all categories of genes. For example, analysis of TRs in the human genome shows that genes in certain categories of biological function are enriched for variable TRs (Table 2) (68). Functional enrichment among genes with TRs has also been observed in yeast, where TRs tend to be found in genes involved in particular processes, with microsatellites being primarily found in regulatory genes such as those encoding transcription factors, and minisatellites in cell-wall genes (101, 135, 147).

Once believed to be neutral variations in junk DNA, TR polymorphisms are now known to be, in certain instances, linked to variation in function. An increasing body of evidence shows that the genomes of diverse organisms contain

Table 1 Genes with (coding) repeats in various species

Species	Genes with repeats within coding regions	
	(%)	
<i>Saccharomyces cerevisiae</i>	12.7	
<i>Arabidopsis thaliana</i>	13.6	
<i>Caenorhabditis elegans</i>	13.1	
<i>Drosophila melanogaster</i>	21	
<i>Homo sapiens</i>	17.1	

The percentage of genes with coding tandem repeats (TRs) compared to total number of genes is shown for yeast, plant, worm, fly, and human.

a**Number of repeated units****Distributions within noncoding****Unit size****Repeat array length****Repeat purity****b**

TR sequences that not only have specific biological functions, but by means of their intrinsic instabilities, may also confer faster rates of evolution of genes and their associated phenotypes. In this review, we compile and summarize examples in the literature that demonstrate evidence for these repetitive sequences having diverse functional roles among a large spectrum of living organisms, with special emphasis on the role of repeats in transcription regulation and protein function. We also make note of the few examples so far that demonstrate more rapid evolution of genes and biological functions, mediated by these highly variable DNA sequences.

First, we focus on the role of repeats in diseases. Second, we discuss the potentially beneficial role of variable TRs located within coding regions. Lastly, we summarize the effects of TRs located within regions regulating gene expression (promoters and other regulatory sequences).

TANDEM REPEATS AS HARMFUL ELEMENTS: NEURODEGENERATIVE DISEASES AND BEYOND

TRs in the human genome first became well known because they are the cause of several neurodegenerative diseases. The three most known examples of such diseases, undoubtedly because they were the first to be genetically characterized, are Fragile X Syndrome (FRAXA) (134), spinobulbar muscular atrophy (SBMA) (65), and Huntington disease (HD). The common genetic feature underlying these three pathologies is the expansion of the number of TRs located within a specific gene, with

Table 2 Enrichment of function among human genes containing tandem repeats (TRs)

Biological process	Adjusted <i>P</i> -value
Regulation of transcription from RNA polymerase II promoter	8.05×10^{-9}
Positive regulation of transcription; DNA dependent	6.09×10^{-4}
Forebrain development	4.15×10^{-3}
Negative regulation of metabolic processes	3.35×10^{-3}
Embryonic morphogenesis	7.90×10^{-3}
mRNA metabolic processes	9.28×10^{-3}
Sensory organ development	1.11×10^{-2}
Cell fate commitment	1.96×10^{-2}
Base-excision repair; DNA ligation	1.96×10^{-2}
Chromatin remodeling	2.26×10^{-2}
Organ morphogenesis	2.45×10^{-2}
Neurogenesis	2.55×10^{-2}
Anterior/posterior pattern formation	3.01×10^{-2}
Ribosome assembly	3.51×10^{-2}

Genes in the human genome containing TRs are enriched for particular functions and processes. Shown are the most enriched categories and the corresponding value of statistical significance. See Legendre et al. (68) for more details.

affected individuals carrying greater TR expansions having earlier onset and higher severity of the disease.

Since the identification of this specific mutation type in the early 1990s, more pathologies have been added to the category of diseases caused by expanded TRs and are now classified as disorders of unstable repeat expansion (reviewed in 39). These diseases present some other common features, such as an autosomal dominant type of inheritance [one exception being Friedreich ataxia (FRDA), see below] or an X-linked type of inheritance. The pathological state is manifested when the number of repeats exceeds a certain threshold. In the

Figure 4

Characteristics of tandem repeats (TRs) in the human genome. (a) Shown are the distributions of various TR characteristics within the genomes of *Homo sapiens*. Each row shows the distribution of different TR parameters (number of repeated units, size of repeat unit, length of the repeat array, and repeat purity). In the columns are the distributions between coding and noncoding regions, with the third column showing more detailed distributions within noncoding regions (UTRs, intronic and intergenic sequences). (b) A closeup of distributions of TR unit sizes in the human genome from part (a). Unit sizes in multiples of three have been colored red for coding sequences, blue for noncoding sequences. Note that the coding regions are depleted for TR units that are not in multiples of three, whereas the distribution in noncoding sequences does not show depletion for any repeat unit sizes.

unaffected human population, the repeats in question show extensive length variability, but their number never exceeds the critical pathological threshold. The number of repeats usually tends to increase, sometimes quite dramatically, in successive generations. These disorders have been extensively reviewed elsewhere (39, 96), but a brief description is given in this section with special emphasis being devoted to the role of the TR expansion in pathogenesis. We subdivide these disorders according to the location of the TR within the mutated gene (see **Figure 3** for illustration).

Human Diseases Caused by Repeat Expansions in Coding Regions

Some of the earliest studies showing that variation in TRs can result in phenotypic consequences comes from research into the origins of human genetic diseases. In particular, several neurodegenerative diseases have been linked to variation within specific repeat tracts located within coding regions.

Huntington disease. HD is probably the best-known example of a disorder caused by expansion of TRs in a gene's coding region. It is a devastating neurodegenerative disease with an autosomal dominant type of inheritance and a typical late onset manifestation. Patients with HD suffer from a number of symptoms including chorea, cognitive decline, and psychiatric disturbances (96). The joint efforts of a number of research groups led to the identification of the genetic cause of HD: expansion of a CAG repeat in exon 1 of the *IT15* gene (50). Repeat numbers ranging from 6 to 35 are found in healthy individuals, whereas alleles with 36 to 39 repeats are associated with an increased risk of developing HD, and 40 repeats or more cause HD (110). The number of CAG repeats in HD correlates with the age of onset and the severity of the disease. Long repeats cause early disease onset and more severe symptoms (26).

The *IT15* gene encodes a large protein named huntingtin (htt) with the CAG repeats

resulting in a stretch of glutamine residues at its amino-terminus. Huntingtin is a large multifunctional scaffold protein with many interacting partners, and it plays a role in numerous cellular functions, mainly transcription, transport, and signaling (39). The precise pathological mechanism of HD is still not entirely elucidated but strong evidence supports transcriptional dysregulation, transport defects, and mitochondrial dysfunction as the main causes of pathogenesis (96). The expanded polyglutamine tract interferes with the interaction of htt and its numerous partners in the nucleus, such as the global transcriptional activator CREB binding protein (CBP), resulting in aberrant expression of target genes (91). In addition, expanded (mutant) htt can no longer sequester the neuron-restrictive silencer element (NRSE) in the cytoplasm. As a consequence, NRSE accumulates in the nucleus and represses the expression of brain-derived neurotrophic factor (BDNF) (150). Decreased axonal transport of BDNF-containing vesicles and altered Ca^{+2} signaling at the mitochondrial membrane also contribute to pathogenesis (40, 97).

Spinobulbar muscular atrophy. Also known as Kennedy disease, SBMA is an X-linked recessive disorder, which therefore primarily affects males. It is caused by the expansion of a CAG repeat in the coding region of the androgen receptor gene (65). The normal allele length ranges between 9 and 36 repeats, whereas in affected patients it varies between 38 and 62 repeats (65). The polyglutamine residues, resulting from the CAG repeats, are located in the activation domain at the protein's amino-terminus. The androgen receptor activates transcription of androgen responsive genes when it translocates to the nucleus upon hormone binding. The expansion of the intragenic CAG repeat leads to a gain-of-function of the mutant protein (96). The pathogenesis of this disease is still unknown but recent studies point towards the inhibition of fast axonal transport in motor neurons by the mutant protein as a likely mechanism (96).

Spinocerebellar ataxias. Spinocerebellar ataxias (SCA) are a group of dominantly inherited, related disorders that can only be clearly distinguished at the genetic level because some of their symptoms overlap. They are all caused by the expansion of a CAG repeat in the coding region of specific genes. The resulting mutant proteins are ataxin-1, 2, 3, and 7 for SCA1, 2, 3, and 7, respectively, the α 1-A subunit of the voltage-dependent calcium channel (CACNA1A) for SCA6, and the TATA-binding protein (TBP) for SCA17. In all of these disorders, there is an inverse correlation between the number of repeats and age of onset of symptoms (96). The pathogenic mechanisms of these diseases are still not fully elucidated but aberrant interactions between the expanded proteins and their partners seem to be a common theme underlying some of these pathologies. In SCA1, for example, these aberrant interactions are believed to enhance transcriptional repression of specific neuronal genes and consequently lead to neurotoxicity (66). In SCA7, on the other hand, changes in chromatin structure at target promoters underlie disease pathogenesis. Mutated ataxin-7 enhances the recruitment of GCN5 histone acetyltransferase complexes to promoters of a subset of genes expressed in rod photoreceptors. This results in hyperacetylation of histone H3 and subsequent downregulation of gene expression (48).

Human Diseases Caused by Repeat Expansions in Introns and Untranslated Regions

Apart from repeats located within coding regions, variation of repeats located in introns and untranslated regions of genes can also lead to various diseases.

Fragile X syndrome. Also known as fragile site mental retardation 1 (FMR1), FRAXA is one of the most common forms of inherited mental retardation. This disease, with an X-linked type of inheritance, is caused by the expansion of a CGG repeat in the 5' untranslated

region (UTR) of the *FMR1* gene (134). In affected patients, the repeat number is larger than 200, whereas it varies between 6 and 60 in the unaffected population (96). This expansion results in transcriptional silencing of the *FMR1* gene as a result of increased methylation of the CpG island in the 5' UTR as well as decreased histone acetylation at the 5' end of *FMR1*. The result of these epigenetic changes is the inhibition of transcription factor binding (43). The FMRP protein plays a crucial role in controlling local protein synthesis in neurons by binding to specific mRNA targets and regulating their translation. The absence of this protein in affected individuals leads to an uncontrolled synthesis of proteins involved in cytoskeletal structure, synaptic transmission, and neuronal maturation (10).

Friedreich ataxia. The most common form of ataxia, FRDA is the only disorder of repeat expansion with an autosomal recessive type of inheritance. It is caused by the expansion of a GAA repeat in the first intron of the *FRDA* gene encoding a protein called frataxin (14). In affected patients, the number of TRs in the *FRDA* gene can exceed 1,000, whereas in the unaffected population it can be as low as 6 and rarely exceeds 32 repeats (31, 85). As for most other known repeat-associated diseases, the age of onset of this disease inversely correlates with the number of TRs. Symptom severity also correlates with repeat length, with the long repeats causing the most severe symptoms (31).

In affected patients, the transcription of *FRDA* appears to be impaired by the presence of the long stretches of intronic GAA repeats and as a result levels of frataxin are extremely reduced (13). Several studies have implicated changes in chromatin methylation and acetylation profiles in the expanded *FRDA* gene to explain the low transcription in affected individuals (1), whereas other studies point to stalled transcriptional elongation (100). The exact physiological role of frataxin, a highly conserved mitochondrial protein, is not fully elucidated, but several studies point to a role in heme biosynthesis, the regulation of iron

UTR: untranslated region

levels, and the formation of iron-sulfur clusters on mitochondrial proteins (12, 70). Hence, the pathological mechanism of this severe disease most likely involves a decrease in cellular energy levels and an increase in free radical production resulting from impaired synthesis of iron-sulfur-containing proteins in the mitochondria.

Myotonic dystrophy type 1 and 2. Two TR expansion diseases with overlapping phenotypes, yet distinct genetic bases are myotonic dystrophy type 1 (DM1) and type 2 (DM2). DM1 is caused by the expansion of a CTG repeat in the 3' UTR of the protein serine-threonine kinase DMPK (9). Repeat number in affected patients varies between 50 and 1,000, whereas in unaffected individuals this number lies between 5 and 37. DM2 is caused by the expansion of a CCTG repeat in intron 1 of the *ZNF9* gene encoding zinc finger protein 9 (ZNF9) (72). In this case too, repeat expansion can be dramatic, reaching 11,000 repeats in some patients, whereas 26 repeats is the highest number detected in the unaffected population (72).

A common pathogenic mechanism, mediated by gain-of-function at the RNA level, has been proposed to explain the overlapping clinical features of DM1 and DM2. In fact, the expanded repeats in *DMPK* and *ZNF9* genes do not affect the processing of the resulting RNAs (as seen in other disorders caused by untranslated repeat expansion) but rather alter the activity of two important RNA processing proteins, CUG-BP1 and MNBL. This *trans* effect of the expanded TRs is achieved via the adoption of stable hairpin RNA structures, which sequester MNBL on the one hand (83) and trigger the hyperphosphorylation and overexpression of CUG-BP1 on the other (64). This results in the missplicing of a certain number of RNAs that, as in the case of the muscle chloride channel *ClC-1* gene, can lead to a premature stop codon and loss of the protein product (141).

Unstable Tandem Repeats Cause Diseases in Other Organisms

The pathogenic phenotypes caused by unstable TRs have been, understandably, almost exclusively described in humans. However, a few notable examples of diseases caused by variable TRs have been described in other species. Canine epilepsy can be caused by the expansion of a 12-nt TR in the single exon of the *Epm2b* gene. The GC-pure nature of the repeat itself is believed to induce formation of secondary hairpin structures obstructing transcription and resulting in significantly reduced levels of the *Epm2b* mRNA in affected dogs (74).

Canine ectodermal dysplasia, a disease characterized by the absence of hair and abnormal teeth formation, is another example of a disease involving TRs in dogs. The hairless phenotype is caused by a 7-nt duplication within the GCC repeat tract of the *FOXI3* gene (25). This duplication produces a frameshift and a premature stop codon resulting in a nonfunctional *FOXI3* protein. The most likely explanation of the disease phenotype is the haploinsufficiency of *FOXI3* because homozygous *FOXI3* mutations are embryonically lethal (25).

A third example has been described in a wild strain of *Arabidopsis thaliana*. The expansion of a GAA repeat in the intron of the *ILL1* gene causes severe growth impairment at a temperature higher than the one normally encountered by this species in its natural environment. The repeat expansion causes a significant reduction in *ILL1* mRNA levels, whereas the spontaneous contraction of the expanded GAA tract reverses the mutant phenotype (125).

TANDEM REPEATS IN CODING REGIONS CONFER FUNCTIONAL VARIABILITY

Tandem Repeats as Useful Genomic Elements

The examples of TR-associated diseases and syndromes described in the previous section

seem to indicate that variable repeats located within coding or regulatory loci are dangerous and potentially harmful. Why then are TRs so abundant in functional regions of the genome? One would expect that if repeats are indeed detrimental, they would be removed in the course of evolution. One might argue that some proteins need the repetitive amino acid patterns encoded by TRs. However, in most cases, it is possible to encode amino acid repeats without having unstable TRs on the DNA level because most amino acids can be encoded by various codons; and even a few mutations (impurities) in a TR can dramatically reduce its propensity to mutate (68). Still, many repeats located in coding and regulatory regions are very pure, without any evidence for the third-position codon variation that would be indicative of selective pressure against unstable TRs (135). Moreover, some repeats have been conserved over considerable evolutionary distances (M. Legendre and K.J. Verstrepen, unpublished results). In addition, some repeats seem to have evolved independently in homologous genes. For example, some homologs, such as the *Saccharomyces cerevisiae* *SIZ1* and *SIZ2* genes, show the same repeat sequence that is shifted a few positions in the gene; whereas in other cases, such as the *WSC2* and *WSC3* genes, a similar (but not identical) repeat occurs at the same position in a homologous gene pair. Although standard methods to investigate positive selection, such as Ka/Ks measurement, are difficult for TRs, these facts seem to argue that, instead of being selected against, some repeats may in fact be selected for. Taken together, circumstantial evidence points towards a beneficial role of (certain) unstable TRs. We systematically discuss the growing evidence supporting this hypothesis, starting with phase variation in microbes and ending with potentially beneficial repeats in the human genome. We review the studies supported by solid experimental evidence and mention those that report correlation data and those for which experimental evidence is still lacking.

Phase Variations in Host-Adapted Prokaryotes

The first hard proof of a beneficial role of variable TRs was found in bacteria and particularly in pathogenic species. One of the multiple strategies that these bacteria employ to evade the host defense system is called phase variation (the reversible, random, high frequency gain or loss of a phenotype). This phenotypic switching is mediated by changes in expression of one or multiple genes. Some pathogenic bacteria have made valuable use of variable TRs as mediators for this rapid switching between phenotypes. The mechanism through which this is achieved relies on the high mutability of TRs (Figure 5). These can be located in promoters or ORFs of genes encoding the variant proteins. In this section, we focus on a number of reports documenting phase variation mediated by variable TRs in ORFs. In the subsequent section of the review, we discuss a few examples of phase

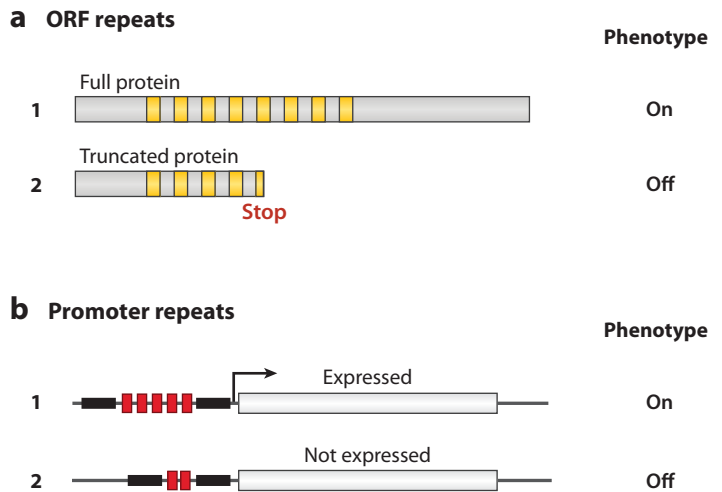


Figure 5

Mechanism of phase variation driven by tandem repeats (TRs). The high mutability of TRs is exploited by pathogenic bacteria to rapidly switch between phenotypes. (a) Variation in TRs located within the gene coding region can either yield a fully functional protein (phenotype ON) or cause frameshifts resulting in a mistranslated, nonfunctional protein (phenotype OFF). In the example shown here, there is a stop codon generated by the change in TR unit. (b) Variation in TRs located within promoters can free or block RNA polymerase binding sites resulting in efficient expression (phenotype ON) or blocked expression (phenotype OFF).

variation mediated by changes in TRs located in promoters.

One of the earliest reports on TR-mediated phase variation came from a study on surface genes in *Neisseria gonorrhoeae*. In this bacterium, members of the P.II gene family contain a variable CTCTT repeat in the region encoding the membrane signal peptide (123). Changes in the number of units of this pentameric repeat can cause frameshifts, and consequently the proteins are either correctly translated (phenotype ON) or not correctly translated (phenotype OFF) in different cells of the population (123) (**Figure 5**). This ON/OFF switching of phenotype occurs during infection and is believed to create variants in the bacterial population capable of surviving in the host (61). Other compelling examples come from *Haemophilus influenzae*, in which the outer membrane structure is also phase variable. The *lic1* gene, responsible for the addition of phosphorylcholine moieties to the membrane lipopolysaccharide (LPS), contains an intragenic CAAT repeat. Variation in repeat number results in different LPS structures (i.e., switching of phenotypes) depending on whether the protein is correctly translated or not (140). In this case too, the switching is essential for the survival of the bacterium in the human host because it leads to variants that are resistant to complement-mediated lysis (140). Recently, the *loxA* gene, also involved in LPS biosynthesis in *H. influenzae*, was shown to undergo phase variation driven by an octameric CGAGCATA repeat. According to the number of repeated units, the LosA protein is predicted either to be fully translated as a functional protein or as a truncated, nonfunctional version. The consequence of this switching is the creation of serum-resistant bacterial clones when the full-length, functional protein is present (27).

Early studies on phase variation in pathogenic bacteria have focused on structural cell-surface genes, but recent genome-wide sequencing has uncovered that these genes represent only a portion of the total phase variable genes in bacteria. DNA-methyltransferases have now emerged as yet another class of

genes capable of rapid phenotypic switching. In *Neisseria* and *Haemophilus* species, the *modA* gene is an example of a phase-variable DNA-methyltransferase. The signature TR in this gene is an AGTC or AGCC repeat whose number of units directly influences the rate of phase variation (22). Depending on the number of repeats, the *modA* gene has two potential start codons, but only the one giving the longest peptide yields a fully active protein. The consequence of *modA* phase variation (i.e., functional versus nonfunctional protein) is a change in the expression pattern of a subset of genes, such as those encoding outer membrane proteins, iron transporters, and heat-shock proteins. Interestingly, some of the *modA* targets are also predicted to be phase variable themselves. On the population level, the phenotypic consequence of this multilevel phase variation is a significant increase in the species' fitness under different stress conditions (e.g., heat shock, antimicrobial agent) when *modA* is switched OFF (121). In *N. gonorrhoeae* particularly, increased biofilm thickness and density as well as better survival in human cervical epithelial cells are a direct consequence of a *modA* in an OFF state (121). As this example shows, TR-mediated switching of a global regulator can have pleiotropic effects on a number of genes. This mechanism can generate a concerted change in gene expression patterns of different coregulated genes, which might increase the fitness of the population or facilitate adaptation.

More recent genome-wide studies suggest that the few well-studied variable repeats in prokaryotes may only represent the proverbial tip of the iceberg. Particularly in host-adapted (virulent) prokaryotes, the number of TRs is high, suggesting that in these genomes, TRs are selected for rather than against (86). All these reports argue for the utility of TRs for prokaryotes, particularly through the immense adaptive capacity that they confer to these organisms.

These fascinating examples of phase variation were the first to illustrate the role of TRs as mediators for phenotypic change, which allows for rapid adaptation of the organism

(e.g., bacterium) to a hostile environment (e.g., immune system). In the next paragraphs, we describe TRs as mediators of rapid evolution in eukaryotes, starting with the simplest species, the budding yeast (*S. cerevisiae*), up to the more complex multicellular eukaryotes.

Variable Tandem Repeats Generate Functional Variability in Cell Surface and Extracellular Proteins

Like the prokaryotes described above, eukaryotic microbes also contain a large number of TRs located within coding regions. The best-studied case is the benign brewer's yeast *S. cerevisiae*. Depending on the threshold values used to determine if a sequence is a TR, variable TRs are found in as many as 12.7%–22% of all genes (coding regions) (Table 1) (68, 135). Most coding TRs are located within genes encoding cell-surface proteins and genes encoding regulatory proteins such as transcription factors and chromatin remodelers. Interestingly, the former gene class contains mostly long minisatellite repeats (i.e., unit size ≥ 10 nt), whereas regulatory genes contain mostly microsatellites (i.e., unit size ≤ 9 nt) (101, 135, 147). Even in higher eukaryotes, including humans, minisatellites are often found in genes encoding large, extracellular proteins such as mucins, whereas microsatellites are mostly present in regulatory genes (68; M. Legendre and K.J. Verstrepen, unpublished results).

Experimental studies in yeast have described several unstable phenotypes linked to genes containing variable TRs in their coding region. The report on the *FLO1* gene in *S. cerevisiae* is arguably one of the most compelling and experimentally supported studies. The *FLO1* gene encodes a cell-surface adhesin, a mannoprotein mediating the adherence of yeast cells to each other (i.e., flocculation) as well as to plastic and other surfaces. This gene contains a variable number of repeats of approximately 100 nt, and the repeat region shows particular length variability in different *S. cerevisiae* strains (135). Experimentally altering the number of repeats in a lab strain leads to a variable and quantifiable

phenotype: As the number of repeats in *FLO1* increases, cell adherence to plastic becomes stronger and the fraction of flocculating cells becomes larger. Variable TRs, in this case, lead to gradual, quantitative functional changes that would allow for a rapid adaptation of the organism to changes in the environment (135).

Adapting to difficult environmental conditions is a property that, at times, is more clearly deciphered in wild microorganisms than in their lab-domesticated counterparts. An example of such a case, in which intragenic TRs are involved, comes from the study of the wild flor *S. cerevisiae* strains. These flor strains have the unique capacity to form buoyant biofilms at the liquid surface, an adaptive mechanism to oxygen-limiting conditions. This unique feature separating flor strains from most lab strains, which tend to sediment rather than float, is due to variable TRs in the *FLO11* gene (30). *FLO11* is another member of the *FLO* gene family of cell-surface adhesins. In its coding region, an approximately 100-nt unit is repeated several times resulting in a serine- and threonine-rich protein. In the flor strains, the number of repeats is much higher than in the lab strains and, as reported by Fidalgo et al. (30), the flor *FLO11* TR region is sufficient to confer the floating biofilm phenotype to a lab *S. cerevisiae* strain. In this particular case, increasing the number of serine and threonine residues, which are highly glycosylated, results in a more hydrophobic Flo11 protein and subsequently a more hydrophobic cell surface capable of maintaining the cells in a biofilm at the air-liquid interface (30).

Adhesins in pathogenic fungi, such as *Candida albicans*, also contain variable repeats. Unlike those in *S. cerevisiae*, adhesins in pathogenic fungi are important virulence factors that confer adhesion to host tissues. As with the repeats in the *S. cerevisiae* *FLO* genes, variable TRs in pathogenic adhesin genes seem to modulate adhesion strength (49). These reports highlight the ability of TRs to generate highly dynamic protein interaction surfaces that can be adjusted, by means of their variable nature, to

create more robust and adaptable variants within a species.

Repeats in fungi do not just occur in cell-surface genes but also in genes encoding proteins that regulate cell-surface properties. In *S. cerevisiae*, for example, the *MSS11* and *MSB2* genes, which encode regulators of cell surface genes, contain variable repeats (135). There is, no doubt, more to be discovered on TRs in these genes and others with similar functions.

Repeats as Tuning Knobs for Circadian Clocks

Circadian clocks are defined as the cellular timing mechanisms that synchronize different

biological processes with the external environment. The duration of a circadian cycle, also known as the period, is usually approximately 24 h. This period is highly affected by external conditions (light/dark cycles, temperature), and different molecular mechanisms responsible for maintaining the period length constant have been identified (148). Still, the key question is how a biochemical system can be tuned to robustly keep in phase with the environment. Interestingly, several studies have elegantly illustrated a role for variable TRs in genes involved in the control and tuning of circadian rhythm in a number of organisms. We first describe two studies, one in the fungus *Neurospora crassa* and the other in the fly *Drosophila melanogaster*, in which initial correlation data are supported by direct experimental results. We also describe a few correlations-only studies in birds.

White collar-1 (WC-1) is a transcription factor inducing the expression of a key component of the *N. crassa* circadian clock (35). The amino-terminus of the WC-1 protein contains a polyglutamine TR that is required for a functional circadian clock in continuous dark conditions (67). In natural isolates of *N. crassa*, circadian period length and temperature compensation vary between strains and this variability is associated with the length of the polyglutamine repeat in WC-1 (Figure 6). Longer polyglutamine stretches correlate with shorter circadian periods in strains collected from low latitudes (i.e., close to the equator) (82). This association is validated by experimental data from a cross between two strains with different polyglutamine repeat lengths. Progeny resulting from this cross have period lengths that cosegregate with the lengths of the polyglutamine stretches (82). Although this report went further than just correlating genotype and phenotype data, a more solid experimental approach would be to create transgenic strains with variable repeat number in the *wc-1* gene and measure the different circadian clock parameters in these strains.

In *D. melanogaster*, the *period* (*per*) gene is an essential regulatory component of the circadian

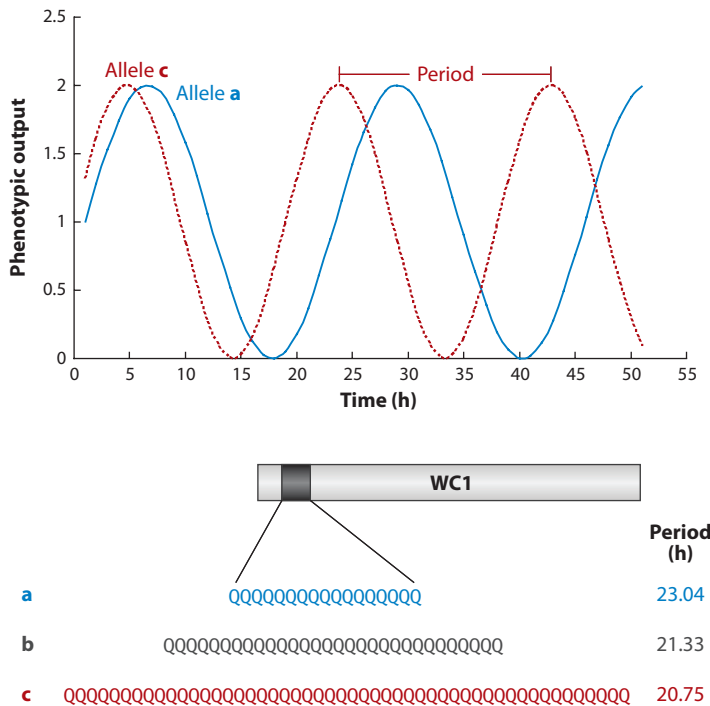


Figure 6

Tandem repeats (TRs) as a tuning mechanism for circadian clocks. TRs are often found in genes controlling circadian rhythms. Variation in the number of these repeats is frequently found across populations, and different sized alleles of the repeat tracts are associated with different characteristics of the circadian rhythm, such as phase and period. An example of a gene with repeat lengths that associate with phenotypic variation in the circadian clock is the *wc-1* gene of filamentous fungus *Neurospora crassa*. Natural variation in an N-terminal poly-Q repeat results in variation in the circadian clock controlling asexual reproduction (conidiation) in this fungus (82).

clock. In natural populations, the *per* gene has a variable threonine-glycine TR, the two most common alleles having 17 and 20 repeats. At warm temperatures, natural isolates and transgenic flies with 17 repeats have a 24 h circadian clock that gets shorter at cooler temperatures. The flies with 20 repeats, however, show better temperature compensation at low temperatures and are thus more favored in colder environments. Such subtle differences allow a better adaptation of the flies with 17 repeats to warmer environments and the flies with 20 repeats to the colder environments. This is supported by the findings that in the warmer parts of Europe and Australia the 17-repeat allele is predominant over the 20-repeat allele in natural populations, whereas the reverse is observed in populations from the colder regions of these continents (114), although it is sometimes difficult to distinguish between neutral variation that segregates with geographic isolation and true adaptation. Similar allele/environment correlations have been reported in another study involving the distribution of *Drosophila* populations (115). Interestingly, a variable threonine-glycine repeat in the *per* homolog of the rat mole (*Spalax ehrenbergi*) has also been implicated in differences in circadian behavior in these blind animals (7).

The avian *CLOCK* gene is another example of a TR-containing gene involved in circadian rhythm. Studies on this gene's TRs are more recent, and only statistical correlations have been reported (55). There is, however, a striking similarity with the results reported on the *Drosophila per* gene. In the avian genome, the Clock protein contains a variable polyglutamine TR at its carboxy-terminus. Sequence analysis of the *CLOCK* gene in two species of birds, the migratory bluethroat and the non-migratory blue tit, revealed some surprising differences in the variability patterns of the TR between the two species. TR copy number variation, for instance, is more present in the blue tit population than in the bluethroat population sampled across the European continent. In addition, TR allele frequencies in blue tits have a latitudinal cline: North European

populations have alleles with higher mean repeat number than their southern counterparts (55). Although this study only reports correlation data, it is of interest to highlight the differences in behavior between the blue tit and the bluethroat species. The nonmigratory blue tits need to adapt to seasonal changes in light intensity and other climatic factors, whereas for the migratory bluethroats the need for adaptation is probably not as strong. Variable TR in the *CLOCK* gene in the blue tits population could provide the means to tune the internal clock in response to changes in the environmental conditions.

The examples reported here point, once again, to the rich opportunities nature has with the use of variable TRs. More particularly for the circadian clock, TRs may provide the mechanism through which a synchronized internal clock is tuned to variable environmental conditions.

Tandem Repeats Mediate Evolvability in Organismal Morphology

In higher vertebrates, TRs are often enriched in genes controlling body morphology, suggesting that these unstable repeats may confer evolutionary flexibility to organ and/or body shape (68). A compelling example comes from a study by Fondon & Garner (32). Comparing genomic and morphological data from different dog breeds, the authors provide strong evidence that implicates variable TRs behind the rapid evolution of canine skeletal morphology.

Two regulatory genes, *Alx-4* and *Runx-2*, which also have homologs in the human genome, were singled out in this study. Dogs with a 51-nt deletion in the TR (coding for a proline-glutamine tract) of *Alx-4* have an additional rear claw (i.e., bilateral rear polydactyly, a signature feature of the Great Pyrenees breed), whereas dogs with the full-length repeat do not present polydactyly. Deletion of 17 amino acids in this repeat is believed to cause a reduction in *Alx-4*-dependent gene expression in limb buds during development (32).

Similarly, the *Runx-2* gene in dogs contains two variable TRs, one coding for polyglutamine and the other for polyalanine. The ratio of glutamine repeats over alanine repeats strongly correlates with the degree of dorsoventral nose bend and midface length in different dog breeds (**Figure 7a,b**). These correlation data are consolidated by the phenotypic characteristics resulting from mutations in the human *Runx-2* homolog (including a polyalanine tract expansion). Clinical features observed in these patients include craniofacial malformations, supporting an evolutionary conserved, functional role for *Runx-2* TRs (32).

This study provided a striking observation on the rapid, continuous evolution of skull morphology in dogs. Despite a strong selection against genetic diversity as a result of domestic breeding, gross morphological evolution is evident in as little as 50 years. This rate of evolution can probably not be sustained by the rate of single point mutations. This fact, in addition to the presence of variable (and evolutionarily conserved) TRs in a high number of genes controlling body morphology, as well as the experimental data on the *Alx-4* and *Runx-2* genes, provide a compelling argument for the role of TRs as mediators of rapid morphological evolution in mammals (32). However, the ultimate evidence for the role of repeats, a series of transgenic animals that differ only in TR length, is still lacking.

A role for TRs in the molecular evolution of vertebrate skeleton has been recently investigated using, as a model, the molecular structure of the tooth in different vertebrates. Teeth consist of a protein matrix upon which the calcium phosphate minerals are assembled. In vertebrate tooth enamel, the major protein forming this matrix is amelogenin. This protein contains a central domain composed of proline-rich tripeptide TRs (PXX). The proline-rich domain adopts a flexible but well-defined helical conformation, which is central to its role as a structural matrix (54). Analysis of amelogenin orthologs shows that the number of repeats increases from amphibians to mammals. The authors argue that as a

result, the mineral binding capacity and, subsequently, bone thickness and strength increase. They provide *in vivo* evidence by showing that transgenic mice expressing frog amelogenin have a significantly less compact matrix and thinner enamel than wild-type mice (54).

This study shows how the variable nature of TRs may have been exploited for the emergence of the solid vertebrate skeleton and teeth, which were no doubt decisive in giving the vertebrate species a survival edge. It is important to note that the repeats in the amelogenin gene are not very pure, i.e., the repeat unit itself is not highly conserved on the DNA level, which implies that the TR may not be hypervariable. However, the TR sequence most likely was pure and highly variable when it first developed, and then it acquired point mutations that reduced its variability.

How Does Coding Repeat Variation Lead to Functional Changes?

As mentioned in the introduction, most TRs in coding regions have a unit size that is a multiple of three. Identical triplet repeats encoding a stretch of a single amino acid are relatively abundant. In eukaryotes, there is an overrepresentation of five amino acids in these stretches: glutamine, arginine, glutamate, alanine, and serine (87; M. Legendre and K.J. Verstrepen, unpublished results). In prokaryotes, however, the picture is slightly different with the most represented amino acids in single stretches being serine, glycine, alanine, and proline (29). Interestingly, in both prokaryotes and eukaryotes, repeats coding for hydrophobic amino acids, such as isoleucine, methionine, and tryptophan, are largely absent. These observations might indicate that selection could be playing a role in shaping the landscape of coding repeats. This active role has, indeed, been suggested by a recent comparative study on intragenic TRs in 12 vertebrate genomes. In particular, a number of trinucleotide repeats within ORFs, but not their equivalent sequences found in non-coding regions, have been retained over considerable evolutionary distances, arguing for a

positive selection being exerted on these sequences (87).

How does repeat variation lead to functional changes? Although there are a number of examples where variable repeats located within coding regions affect certain phenotypes, little is known about the underlying molecular mechanism. TR domains in proteins are predicted to be mostly flexible and unstructured (119); this is supported by the absence of any crystal structure of a repeat-containing domain in the Protein Data Bank, which contains the majority of known protein crystal structures (M. Legendre and K.J. Verstrepen, unpublished results). Thus, the roles of a large number of TR-containing domains in proteins are still not clearly understood, but the general consensus is that these domains mediate protein-protein interactions (58). In support of this, TR-containing proteins are usually part of multiprotein complexes (29) and have also been found in highly connected proteins (i.e., hub proteins) when protein interaction networks from a number of organisms were analyzed (24). Together, this indicates that intragenic microsatellite repeats might encode unstructured, hydrophilic loops that stick out of the protein surface and serve as interaction domains. By contrast, longer, minisatellite repeats could play a more structural role as already indicated by several studies (30, 135).

TANDEM REPEATS AND THE EVOLUTION OF PROMOTERS

An increasing number of studies describe polymorphic TR sequences situated within the promoters of genes. As with many TR polymorphisms, these were once regarded as neutral mutations, but greater interest in studying the genetic underpinnings of variation in gene expression has led to a reexamination of these hypermutable sequences. In the following sections, we review studies that have uncovered a link between TR polymorphisms in promoters and variation in gene expression, gene function, or both. In some notable examples, the studies have gone beyond correlating particular TR

alleles with phenotypes and have demonstrated that specific changes in the TRs of promoters lead to quantifiable changes in gene expression. Here, we begin with a sampling of correlative studies, followed by sections summarizing reports where mechanisms for TR-mediated gene expression regulation are revealed.

Tandem Repeats and Gene Expression

A plethora of studies have revealed correlations between particular TR alleles and gene expression levels and phenotypes. For many of these examples, it remains to be confirmed whether variation in the repeats themselves are the cause of gene expression variation rather than being merely linked to expression-determining loci. It is nonetheless worth mentioning them to highlight the prevalence of association between variable TRs and gene expression variation. The common occurrence of such repeats in promoters and their association with expression variability suggest future directions in the study of the roles of these unstable sequences in the genomes of diverse organisms.

A survey of the literature suggests that TRs may be involved in introducing variation in the expression of a huge variety of genes and phenotypes. For example, in tilapia, an important aquacultural fish, variable CA repeats in the promoter of *prl1*, a gene encoding a hormone involved in osmoregulation, show an association with both *prl1* expression as well as salt response (124). The authors of this study crossed individuals with the two differing naturally occurring allele lengths, CA₃₁ and CA₁₄, and found that individuals in the F₂ generation, homozygous for the long allele, had high expression of *prl1* and low fish mass under high-salt conditions.

Correlative observations between TR alleles and gene expression have been further substantiated in instances where investigators have manipulated the TRs in vivo or in vitro. One of the earliest reports of TRs modulating gene expression (44) showed that repeated tracts of TG or CA are capable of enhancing expression of a reporter gene in vivo in cultured simian and

human cells in a repeat-length dependent manner. Another study found similar results with the promoter of the human serotonin transporter gene, *5HTT*, which contains a polymorphic TR of 20–23-nt units. Shorter alleles of this TR have reduced transcriptional efficiency in both a reporter gene assay as well as in a native context, and are associated with anxiety-related personality traits (69).

Similarly, shorter alleles of TRs upstream of the human insulin gene (*INS*) are linked to altered expression of *INS* both in vitro and in vivo, and to predisposition to insulin-dependent diabetes mellitus (6). This same TR was found in another study to affect expression of the nearby *IGF2*, a gene encoding insulin-like growth factor II (99). The exact mechanism by which this enhancement was produced was not determined in the study, but it was suggested that the TG elements may potentiate Z-DNA formation, which is known to alter gene expression (see below).

The CCTTT repeats in the promoter of primate inducible nitric oxide gene (*iNOS* or *NOS2A*) are variable across many primate species and multiallelic in humans (145). Transcription of *iNOS* is increased with increasing number of CCTTT repeats in the promoter as shown by a luciferase reporter gene assay (138). Interestingly, the various *iNOS* promoter TR alleles have been linked to disease: the 14-repeat unit alleles are associated with the absence of diabetic retinopathy among diabetic patients (138), and longer alleles (≥ 15 -repeat units) have been associated with severe malaria (93).

Promoter TRs might contribute to variation in complex mammalian behaviors. A study in voles demonstrated that variable TRs in the promoter of the *vasopressin 1a receptor* (*avpr1a*) gene contribute to variation in gene expression and social-behavioral traits such as pair bonding and parental care (45). Larger alleles of the TR are found in more social species of voles and promote higher expression of a luciferase reporter gene than the shorter alleles, which are found in asocial vole species. The association between TR size and sociability is observed

even within a population of a single vole species, with individual male voles that carry longer alleles showing increased social behaviors such as pup grooming and partner preference. The allele lengths are also able to predict differences in the vasopressin receptor distribution patterns in vole brains. The existence of polymorphic alleles in primate *avpr1a* loci suggests a role for TRs may also exist in human behavioral trait variation.

The relationship between TR size and gene expression or phenotypes is not always straightforward. An example of how the relationship can be complex is found in the pathogenic fungus, *C. albicans*. This fungus expresses, among other virulence factors, secreted proteases that help it attain nutrients and invade host tissue. One such protease is encoded by *SAP2*, which contains in its promoter two variable pentameric TRs (GCTTT and TTGAT/A). *C. albicans* is a diploid organism, and allelic differences in the repeat number exist between the two copies of the *SAP2* promoter. Alteration of the number of repeat units of these promoter TRs affected the timing of expression of each allele during infection, as well as the level of induction in laboratory conditions (122). What is interesting is that in this promoter, it is not the specific number of units in each individual TR that determines expression but rather the total combined unit copy number of the two TRs. This is reminiscent of the finding in the *Runx-2* gene in dogs, where the ratio of two TR lengths is what correlates with a phenotype (Figure 7a,b) (32). It is also often the case in many of the studies mentioned here that a certain size of repeat is needed for optimal gene expression, above or below which expression levels decrease (for example, see Figure 7d). Such patterns are more readily discerned when several alleles of the TR containing a wide range of unit numbers are available for analysis whether they are naturally occurring alleles or engineered ones. The bases for such size-dependent expression patterns may be explained by the various mechanisms at work in promoter TR-mediated gene expression modulation (see further sections below; summarized in Table 3).

Table 3 Mechanisms for gene expression regulation by tandem repeats (TRs)

Properties or functions influenced by tandem repeats	References
Overlap with regulatory protein binding sites	19, 56, 78, 142
Chromatin structure	42, 112, 132, 136
Z-DNA formation	89, 109
Spacing of promoter elements	133, 144, 146
RNA structure	37, 129

A summary of known and proposed mechanisms for TR function in gene expression regulation.

changes of the gene encoding the NadA adhesion protein. The stochastic loss or gain of the repeats results in variable binding of the IHF transcription factor (78). Similarly, among eukaryotes, concomitant variation of repeats and transcription factor binding sites has been shown to influence the level of expression of at least one gene in *S. cerevisiae* (136). Variation in TCC repeats in the promoter of the human epidermal growth factor (EGF) gene alters binding of the transcriptional regulator Sp1 to the promoter and EGF expression (56). The promoter of the human reduced folate carrier gene (hRFC) contains a polymorphic 61mer TR. An increase of just one unit in this repeat tract is correlated with higher expression levels of the gene (142). The TRs were shown to bind in vitro two transcription factors, Sp1 and AP2. Similarly, variation in the number of copies of a pentanucleotide TR (TG[T/C]CC) in the promoter of the p53-induced quinone oxidoreductase homolog *PIG3* results in a variable number of binding sites for p53, which in turn determines the amount of *PIG3* expression (19). A screen of polymorphic TRs in the human genome reveals there are many more that are able to bind regulatory proteins (51). It is important to note that apart from TRs in promoter regions, TR polymorphisms affecting transcription factor binding may also be at work in intronic regions (see below).

Tandem Repeats Can Change Spacing Between Functional Elements in Promoters

In some instances, changes in repeat unit numbers do not alter the number of binding sites for

a regulatory protein but rather change important spacing between different regulatory sites. Variable TRs modulating regulatory site spacing often underlie phase variation in bacteria (see above and **Figure 5**). In the pathogenic bacterium *H. influenzae*, loss or gain of TA repeats in the divergent promoters of *bifA* and *bifB* (genes required for the formation of adhesive appendages called fimbriae) alters the spacing between the -10 and -35 consensus sequences, which in turn alters the efficiency of RNA polymerase binding (133). Frequent changes in the repeat unit number thus provide a rapid on-off switching mechanism resulting in a mixed population of bacteria in which only some are fimbriated. Such repeat-mediated phase variation, in which unstable repeats vary the distances between two regulatory sites that affect expression of adhesive and immunogenic surface proteins, appears to be common among pathogenic bacteria. This form of regulation is also present in the regulation of *Bordetella pertussis* fimbriae (144), an adhesive outer membrane protein of *N. meningitidis* (113), and surface lipoproteins in *Mycoplasma hyorhinis* (146). Just as is the case with TRs in coding regions, the high frequency variability in TR size in promoters allows for the rapid production of variants in a pathogenic bacterial population, which increases the chances of evading the host immune system.

Tandem Repeats Influence Chromatin Structure

Whereas some variable TRs can affect gene expression by altering the number of transcription factor binding sites or spacing between promoter elements, a much larger fraction of TRs may influence promoter activity by altering the structure of chromatin, the complex of DNA and protein that makes up the genomes of eukaryotes and archaeans. The basic unit of chromatin is the nucleosome, which consists of chromosomal DNA wrapped around a complex of histone proteins. Binding of histone complexes to DNA greatly influences gene expression by controlling accessibility of transcription and regulatory proteins to the

Cis-regulatory site/element: a local DNA sequence which directly influences gene expression/function (e.g., transcription factor binding site)

DNA (reviewed in 53). The mononucleotide TRs, poly(A) and poly(T), inhibit nucleosome formation and are important promoter regulatory elements (52, 57, 117, 149). A study by Vines et al. showed that TR sequences are enriched within the nucleosome-free sequences of yeast and human promoters (136), hinting at the common role of TRs as nucleosome positioning elements. To test this possibility, the repeats of several of the TR-containing promoters in yeast were mutated, and both gene expression and nucleosome positioning were affected, demonstrating that in yeast promoters, TRs affect gene expression by acting as nucleosome inhibitory sequences that maintain an open chromatin structure in specific regions of the promoter. The study's finding that nearly 25% of all yeast promoters contain TRs suggests these sequences may be a common nucleosome positioning element in eukaryotes. As pointed out in the study, apart from influencing nucleosome position, these TRs may also be influencing transcription through additional mechanisms, such as by influencing DNA structure and melting.

Findings by others studying TRs in specific loci support the role of TRs as nucleosome determining sequences. As previously mentioned in this review, polymorphisms in the CTG repeats downstream of the myotonic dystrophy protein kinase gene (*DMPK*) affect gene expression by influencing chromatin structure, which in turn determines subsequent mRNA splicing patterns (34). Such CTG repeats have been documented to position nucleosomes in eukaryotes (42) and archaeans (112). The exact nature of the repeating units is important in determining the effects TRs have on nucleosomes. For example, (CTG)₁₂ promotes nucleosome formation, (CGG)₁₂ excludes nucleosomes, while (GAA)₁₂ has little effect on nucleosome formation. Of these, only the first repeat, (CTG)₁₂, was reported to affect gene expression in a reporter system (132). TRs of the pentanucleotide CCGNN have been found to inhibit nucleosome formation in vitro (137) and activate expression of a reporter gene in yeast (60).

Tandem Repeats Affect DNA Structure

Z-DNA, an unusual left-handed form of the normally right-handed (B-DNA) helical structure of DNA, is able to act as a *cis*-regulatory element in gene regulation (92). Because Z-DNA formation is favored in sequences with alternating pyrimidines and purines, several classes of TR sequences are able to generate Z-DNA structures. Naylor & Clark (89) describe a CA repeat upstream of the rat prolactin gene that forms Z-DNA and inhibits gene transcription. Interestingly, the ortholog of this gene, *prll*, in tilapia also contains promoter TR polymorphisms associated with variable gene expression (described above) (124). TR sequences that form Z-DNA have been proposed to be an abundant source of quantitative gene expression variation among mammals (109).

The mechanism by which Z-DNA formation affects gene expression is not entirely clear. The structure of Z-DNA itself is so drastically different from the normal B-DNA that its formation may block binding of regulatory proteins to the promoter. Z-DNA is generated as a result of negative supercoiling of DNA, and it may be this torsional strain that inhibits proper transcription factor binding to the DNA. There are also known Z-DNA-binding proteins which may act on gene expression. For example, polymorphisms in the Z-DNA-forming GT tracts of the promoter of the human macrophage immune response gene *SLC11A1* influence gene expression. These same sequences are bound by the transcription factor HIF-1alpha, both in vivo and in vitro (4). The ZalphaADAR domain of the double-stranded RNA adenosine deaminase (ADAR1) protein has also been shown to both bind promoter Z-DNA and activate gene expression (92). The same study also found that the Z-DNA-forming sequences by themselves weakly enhance gene expression, suggesting that these sequences may both recruit expression regulators as well as promote transcription through intrinsic properties. As Z-DNA has been demonstrated to inhibit

nucleosome placement (38), it is difficult to distinguish between the direct effect of Z-DNA formation and its effect on chromatin structure.

Tandem Repeats and the Evolution of Other Regulatory Sequence Elements (Introns, UTRs)

The location of TR sequences involved in the regulation of gene expression and function need not be limited to promoter sequences. TRs have been found in other expression regulatory sequences, including the 5' and 3' UTRs of transcripts, and introns. Many of these TRs have been shown to have a function in regulating gene expression, and variation in repeat units often affects activity. Most described so far have been associated with neurodegenerative diseases (see above), although it is likely they also occur in genes that are not linked to diseases. These TRs may have been retained for an unknown, possibly regulatory, function. As with promoters, TRs in these sequences may be facilitating rapid evolution of gene expression levels.

The UTRs of transcripts are sites of unstable TRs in a number of organisms, including humans. Examples of regulatory TRs in 5' UTRs include a variable 28-nt TR in the 5' UTR of the human *TS* gene (59). In this particular example, it is thought that the number of TRs in the UTR affects translation efficiency rather than transcription or transcript stability. As mentioned earlier in this review, expansion of CG repeats in the 5' UTR of the *FMRI* gene in humans causes the neurodegenerative Fragile X syndrome (134).

Expansion of a CTG repeat in the 3' UTR of *DMPK*, discussed above in the sections on human disease and on chromatin structure, results in the inherited neuromuscular disorder myotonic dystrophy (DM1) (9, 77). The repeats may be causing disease through various mechanisms, including recruitment of RNA-binding proteins, influencing splicing patterns (131) and influencing gene expression of neighboring genes (76). Type 2 myotonic dystrophy

(DM2) has also been reported to be caused by expansion of a TR in the intron of another gene, the zinc finger protein 9 (*ZNF9*) gene. In this case, the expansion is of the tetranucleotide CCTG repeat (72). Despite the different genes and repeat units in both these cases, a similar disease results, suggesting that both repeat expansions trigger similar physiological changes. We have discussed some details of the mechanisms that may underlie these similarities in the section above on human diseases.

Expanded CTG repeats in a 3' UTR have also been strongly associated with type 8 spinocerebellar ataxia (SCA8) (63). Interestingly, expansion of a very different repeat, ATTCT, within an intron has been associated with another form of spinocerebellar ataxia (SCA10) (more on TRs in introns below) (79). The human dopamine transporter (*DATI*) gene, associated with several neuropsychiatric disorders, including attention-deficit hyperactivity disorder, contains in its 3' UTR a repeat of 40-nt units. Expression of a luciferase reporter gene with a promoter containing these repeats increases with the number of repeats (36).

Regulatory TRs can also be found in introns, including the SCA10 repeat mentioned in the previous paragraph. A polymorphic dinucleotide repeat in intron 1 of *EFGR* (a proto-oncogene) modulates transcription of the gene. The number of CA repeat units is inversely proportional to the level of transcript, as observed by both in vitro and in vivo assays (41). Increasing the copy number of a 72-nt repeat in intron 5 of *SIRT3* (the mammalian ortholog of yeast silencing and longevity gene *SIR2*) increases gene expression, and particular alleles of the TR have an association with longevity in human males (5). Changes in copy numbers of a TCAT repeat in the first intron of the tyrosine hydroxylase (*TH*) gene is correlated in a quantitative fashion to both silencing of the *TH* gene and to binding of the transcription factor ZNF191 (2). The optimal number of TR units for silencing was eight, with greater or fewer units giving progressively less silencing activity. This pattern of activity, where there is an

optimal TR unit number for maximal activity, has been observed for several regulatory TRs (Figure 7) and highlights the importance of studying a range of TR allele sizes when ascertaining the relationship between TR length and gene expression.

TRs in UTRs and introns are often retained within highly conserved sequences in mammalian genomes, even though the repeats themselves are variable (i.e., showing variation in the number of repeat units), and sometimes even replaced by different repeats (i.e., repeats with different unit sequences) (104). Although such TR replacements have been observed in disease genes, they may have a conserved role among a variety of genes, particularly those involved in embryonic and nervous system development (105).

Tandem Repeats and RNA Structure

The expansion of several triplet repeats associated with human neurological diseases (CAG, CCG, CGG, CUG, and AUUCU) has been shown to result in hairpin RNA structures (120). It may be the case that both the RNA structures and proteins recruited to such structures play a role in disease pathogenesis (37), and this possibility is subject to further examination.

The effects of TRs found in transcribed regions on RNA structure may in turn affect mRNA processing, stability, and translation. Expanded CUG repeats in the *DMPK* gene transcript, for example, can form very stable hairpin structures, and these structures may be able to recruit and activate PKR, a double-stranded RNA-binding and proapoptotic kinase (129). A TG repeat in the cystic fibrosis transmembrane conductance (*CFTR*) gene forms stable secondary structures. Expanded alleles of this TR can result in decreased RNA splicing efficiency and cause nonclassic cystic fibrosis (47).

There are also examples of repeats in mRNA that affect translation efficiency rather than transcription levels. An example is the TR in

the 5' UTR of the human thymidylate synthase (*TS*) gene, in which longer TR alleles are associated with higher protein levels but not with different levels of gene expression (59). It may be that alterations in RNA structure also play a role in this form of regulation. Certain TR DNAs are also known to mediate gene silencing through a two-step process involving siRNA generated from the TR sequences and subsequent DNA methylation (16). Some of these RNA mechanisms may be at work in the TR-mediated regulation discussed in the following paragraph.

Regulatory Tandem Repeats in Promoters Are Associated With Rapid Evolution of Gene Expression

The findings presented here paint a story of abundant DNA elements in diverse genomes that by means of their intrinsic instabilities act as mutational hot spots. Because of the presence of these highly variable sequences within promoters and other regulatory sequences that control levels of gene expression, it is possible that unstable TRs accelerate evolution of gene expression, generating variability in populations that allows quick Darwinian evolution and adaptation. This possibility merits further detailed investigation. Several reports have presented data suggesting this is the case. Computational analysis of gene expression among yeast strains and among species of yeast indicates that promoters that contain repeats have evolved faster so that their gene expression patterns are more divergent across the strains and species studied (136). It will be of interest to see if this is the case on a genome-wide level in other organisms. So far, only anecdotal reports indicate that this is the case for particular genes. For example, the promoter of *MMP3*, a gene associated with heart disease, contains mononucleotide (polyT) stretches, one of which is polymorphic among humans. A one-nt decrease in the tract size results in higher levels of *MMP3* expression and a significant association with myocardial infarction and aneurysm, whereas an allele just one

nt larger drives lower expression of the gene and is associated with coronary artery disease (106). In much of the world, the distribution of the *MMP3* allele appears to be determined by genetic drift, but in Europe the evidence suggests there has been positive selection for the smaller (5T) allele. The polymorphic tracts of *MMP3* were also shown to be rapidly evolving among several primate species, as the *MMP3* TR stretch is polymorphic across all the primate species observed in the study. That these sequences are rapidly changing among primates suggests that this mutational hot spot may also be driving rapid evolution of *MMP3* gene expression and its associated phenotypes. A similar situation may also occur with the CCTTT repeats of the *iNOS* promoter, which are polymorphic within and between several primate species (145). As already mentioned, the abundance of polymorphic TR sequences in the promoters of the human genome, and their contribution to expression variation, has led to the speculation that such sequences serve as an important source of genetic variation and can accelerate gene expression evolution (107, 130).

At least one study has experimentally confirmed that TRs in promoters can promote rapid evolution of gene expression (136). In this study, promoters containing repeats were cloned into reporter genes that facilitated selection for different levels of gene expression. After just a few rounds of selection, promoter constructs containing TR sequences yielded many variants with higher expression levels linked to specific changes in the number of promoter TR units. Promoters with nonrepetitive control sequences showed no such gene expression evolution in the course of the experiment. More studies of these kinds, using genomic sequence and gene expression data from numerous strains and species, as well as experimental evolution studies, are necessary to illuminate other instances in which TRs are generating variation at high rates that can be acted upon by selection to accelerate gene expression evolution.

CONCLUSION: VARIABLE TANDEM REPEATS MAY BE ACCELERATING THE EVOLUTION OF CODING AND REGULATORY SEQUENCES

Considering all lines of evidence, variation in TR tracts is an underestimated and understudied source of phenotypic variation. They represent, in addition to single nucleotide polymorphisms (SNPs) and copy number variations (CNVs), a third, mostly ignored, category of genetic variation. Many of the examples demonstrating or suggesting that variable TRs influence phenotypes have been discovered by research teams focusing on a specific phenotype or gene rather than large-scale studies specifically aimed at characterizing the global role of repeats in genome evolution. Hence, it is likely that the few instances where unstable repeat tracts have been reported to play a functional role only represent the proverbial tip of the iceberg.

A central question that keeps returning is: Are TRs selected for? Do genomes need repeats, or do they just occur by chance, a nuisance that is difficult to get rid of? In her thought-provoking essay *Chance Favors the Prepared Genome*, Lynn Helena Caporale outlines the many ways that genomes maintain a balance between fidelity of replication and the generation of variability (15). One strategy is for localized gene amplification—duplications that facilitate subsequent genetic changes and phenotypic evolution—to occur (3). Another strategy is to maintain elements in the genome that mutate at rates higher than the rest, while keeping other regions of the genome stable and robust (60a). It is tempting to speculate that TRs represent such hypervariable DNA sequences that confer variability, plasticity and evolvability. Repeats may form completely by chance, for example, as a result of replication slippage. However, it is possible that some of these random unstable repeat tracts may confer useful variability. Such repeats may in fact be selected for, allowing them to spread through the population, reach fixation, and even be conserved

among various species. The exact number of repeat units in such adaptive TRs may differ between individuals, populations, and species, depending on the acting selective pressure. Repeats are also likely to form in regions of the genome where instability is negative, but in such instances, the tracts will be selected against and quickly disappear. Last but not least, some repeat tracts may be virtually neutral, representing true junk DNA.

Apart from their high variability and instability, TRs are also unique in the type of phenotypic variation they generate. In contrast to SNPs and other conventional DNA mutations, which can generate seemingly endless phenotypic variation, TRs seem to be a bit more limited. TR variation does not lead to gross changes in function or completely novel characteristics. As any given TR locus has a wide range of allelic variations, they allow digital rather than binary fine-tuning of specific phenotypes such as binding activities or transcription levels (90). Although some TRs do mediate a binary switch (e.g., ON-OFF switching of cell-surface proteins in certain microbes), in most cases gradual changes in repeat numbers result in gradual changes in the corresponding gene function and the resulting phenotype (Figure 7). Such gradual, quantitative changes are often believed to depend on multiple genes, or quantitative trait loci (QTLs), and complex genetic interactions. However, the examples summarized in this review show that TRs may in fact provide a simple, monogenic mechanism underlying quantitative changes.

It is interesting to note that variable repeats share both genetic and epigenetic characteristics. As they are DNA sequences, TRs are by definition genetic elements. However, with mutation frequencies around 10^{-2} to 10^{-5} per generation, the instability of repeat tracts lies between that of genetic point mutations (typically 10^{-8} to 10^{-9} per generation) and epigenetic switches (typically 10^{-1} to 10^{-2}). Furthermore, like epigenetic mechanisms, variation in TRs is readily reversible; repeat tracts can expand and shrink without loss of information.

Finally, like epigenetic changes, the phenotypic variation conferred by repeat variation is typically limited to a switch (ON or OFF) or to fine-tuning of a given phenotype. Taken together, it seems that variable repeat tracts may help to bridge the gap between highly variable but ultimately functionally limited epigenetic changes, and the typically less variable but virtually unlimited genetic variation conferred by conventional DNA mutations.

Here, we have focused on the role of TRs in the evolution of gene function and expression, but TRs also influence other cellular processes and functions. For example, TRs affect the sites and rates of recombination (60, 88). They may also play an important role in chromosome structure, especially of telomeric and centromeric regions (20). It has been reported that expanded alleles of certain disease-associated TR sequences can act as origins of replication (73), whereas others are known to promote chromosome breakage (33). Chromosomal fragile sites, often characterized by sequences rich in TRs, may underlie specific cases of rapid phenotypic evolution, such as the frequent loss of pelvic spines in stickleback fish populations (17), but more generally will impact chromosomal rearrangements and evolution. Indeed, it has been suggested that the evolution of mammalian genomes, including our own, has been driven by TR-rich fragile sites prone to breakage and chromosomal rearrangements (111). Changes in certain TRs found in DNA mismatch repair genes may even lead to variation in genomic mutation rates (18).

It is most fitting that recent evidence in favor of a role for TRs mediating rapid evolution has been found among dogs (32), as it was more than 125 years ago that Darwin first pondered the underpinnings of variation among domestic canines (21). One hundred years after Darwin's speculation, TRs were first identified as a major source of genetic variation in genomes (127). It is now a wide-open field for further study into the phenotypic consequences of so much variation and their influence on evolutionary trajectories.

SUMMARY POINTS

1. Tandem repeats are found in virtually every genome. They often occur within coding and regulatory regions.
2. Tandem repeats are highly unstable. More specifically, the number of repeated units expands or contracts at high frequencies.
3. Variable tandem repeats located in promoters or coding sequences can act as mediators for rapid phenotypic changes because the frequent contraction or expansion of the repeat tract generates quantitative, gradual changes in expression or function of the respective genes. As such, tandem repeats provide a simple, monogenic mechanism that allows tuning of gene expression or function.
4. Despite their prevalence and their functional role, tandem repeats have been largely ignored in comparative genomics. However, tandem repeat polymorphisms are now emerging as a third major class of genetic mutation, alongside single nucleotide polymorphisms (SNPs) and copy number variations (CNVs).
5. Whereas tandem repeat variation is by definition a genetic change, it also shares certain characteristics with epigenetic changes (e.g., the high instability and complete reversibility).

FUTURE ISSUES

1. Which molecular functions do the peptides encoded by tandem repeats have? How do changes in these coding tandem repeats translate into changes in the structures and activities in the proteins they encode, and how does this yield changes in the resulting phenotype?
2. How can unstable tandem repeats accelerate evolution of gene expression? Do the repeats affect other features apart from DNA and chromatin structure and transcription factor binding sites?
3. How common are tandem repeat mutations in somatic and germline cells of multicellular organisms? And are they adaptive?
4. Do tandem repeats underlie or affect more diseases than the handful of neurodegenerative diseases to which they have already been linked?
5. Do environmental conditions alter mutation rates of tandem repeats? If so, are there mechanisms in place that change the rates specifically for tandem repeats?

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Christopher Brown, David King, David Mittelman, Aaron New and Karin Voordeckers for their help, comments, and suggestions. We also thank John W. Fondon for providing the pictures in **Figure 7**. We apologize for the omission of several relevant studies we could not cite due to space limitations. Research in the lab of KJV is supported by NIH grant P50GM068763, Human Frontier Science Program HFSP RGY79/2007, ERC Young Investigator Grant 241426, VIB, K.U.Leuven, the FWO-Odysseus program and the AB InBev Baillet-Latour foundation.

LITERATURE CITED

1. Al-Mahdawi S, Pinto RM, Ismail O, Varshney D, Lymperti S, et al. 2008. The Friedreich ataxia GAA repeat expansion mutation induces comparable epigenetic changes in human and transgenic mouse brain and heart tissues. *Hum. Mol. Genet.* 17:735–46
2. Albanese V, Biguet NF, Kiefer H, Bayard E, Mallet J, Meloni R. 2001. Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite. *Hum. Mol. Genet.* 10:1785–92
3. Andersson DI, Hughes D. 2009. Gene amplification and adaptive evolution in bacteria. *Annu. Rev. Genet.* 43:167–95
4. Bayele HK, Peyssonnaud C, Giatromanolaki A, Arrais-Silva WW, Mohamed HS, et al. 2007. HIF-1 regulates heritable variation and allele expression phenotypes of the macrophage immune response gene SLC11A1 from a Z-DNA forming microsatellite. *Blood* 110:3039–48
5. Bellizzi D, Rose G, Cavalcante P, Covello G, Dato S, et al. 2005. A novel VNTR enhancer within the SIRT3 gene, a human homologue of SIR2, is associated with survival at oldest ages. *Genomics* 85:258–63
6. Bennett ST, Lucassen AM, Gough SC, Powell EE, Undlien DE, et al. 1995. Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nat. Genet.* 9:284–92
7. BenShlomo R, Ritte U, Nevo E. 1996. Circadian rhythm and the per ACNGGN repeat in the mole rat, *Spalax ehrenbergi*. *Behav. Genet.* 26:177–84
8. Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.* 62:1408–15
9. Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, et al. 1992. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell* 69:385
10. Brown V, Jin P, Ceman S, Darnell JC, O'Donnell WT, et al. 2001. Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. *Cell* 107:477–87
11. Buard J, Bourdet A, Yardley J, Dubrova Y, Jeffreys AJ. 1998. Influences of array size and homogeneity on minisatellite mutation. *EMBO J.* 17:3495–502
12. Bulteau AL, O'Neill HA, Kennedy MC, Ikeda-Saito M, Isaya G, Szwed LI. 2004. Frataxin acts as an iron chaperone protein to modulate mitochondrial aconitase activity. *Science* 305:242–45
13. Campuzano V, Montermini L, Lutz Y, Cova L, Hindelang C, et al. 1997. Frataxin is reduced in Friedreich ataxia patients and is associated with mitochondrial membranes. *Hum. Mol. Genet.* 6:1771–80
14. Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, et al. 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 271:1423–27
15. Caporale LH. 1999. Chance favors the prepared genome. *Ann. N. Y. Acad. Sci.* 870:1–21
16. Chan SW, Zhang X, Bernatavichute YV, Jacobsen SE. 2006. Two-step recruitment of RNA-directed DNA methylation to tandem repeats. *PLoS Biol.* 4:e363
17. Chan YF, Marks ME, Jones FC, Villarreal G Jr, Shapiro MD, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* 327:302–5
18. Chang DK, Metzgar D, Wills C, Boland CR. 2001. Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. *Genome Res.* 11:1145–46
19. Contente A, Dittmer A, Koch MC, Roth J, Dobbstein M. 2002. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nat. Genet.* 30:315–20

20. Csink AK, Henikoff S. 1998. Something from nothing: the evolution and utility of satellite repeats. *Trends Genet.* 14:200–4
21. Darwin C. 1883. *The Variation of Animals and Plants Under Domestication*. New York: D. Appleton Co.
22. De Bolle X, Bayliss CD, Field D, van de Ven T, Saunders NJ, et al. 2000. The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. *Mol. Microbiol.* 35:211–22
23. Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284:601–3
24. Dosztanyi Z, Chen J, Dunker AK, Simon I, Tompa P. 2006. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* 5:2985–95
25. Drogemuller C, Karlsson EK, Hytonen MK, Perloski M, Dolf G, et al. 2008. A mutation in hairless dogs implicates FOXP3 in ectodermal development. *Science* 321:1462
26. Duyao M, Ambrose C, Myers R, Novelletto A, Persichetti F, et al. 1993. Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat. Genet.* 4:387–92
27. Erwin AL, Bonthuis PJ, Geelhood JL, Nelson KL, McCrea KW, et al. 2006. Heterogeneity in tandem octanucleotides within *Haemophilus influenzae* lipopolysaccharide biosynthetic gene *losA* affects serum resistance. *Infect. Immun.* 74:3408–14
28. Fan H, Chu JY. 2007. A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics* 5:7–14
29. Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, et al. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res.* 15:537–51
30. Fidalgo M, Barrales RR, Ibeas JI, Jimenez J. 2006. Adaptive evolution by mutations in the FLO11 gene. *Proc. Natl Acad. Sci. USA* 103:11228–33
31. Filla A, DeMichele G, Cavalcanti F, Pianese L, Monticelli A, et al. 1996. The relationship between trinucleotide (GAA) repeat length and clinical features in Friedreich ataxia. *Am. J. Hum. Genet.* 59:554–60
- 32. Fondon JW 3rd, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. USA* 101:18058–63**
33. Freudenreich CH, Kantrow SM, Zakian VA. 1998. Expansion and length-dependent fragility of CTG repeats in yeast. *Science* 279:853–56
34. Frisch R, Singleton KR, Moses PA, Gonzalez IL, Carango P, et al. 2001. Effect of triplet repeat expansion on chromatin structure and expression of DMPK and neighboring genes, SIX5 and DMWD, in myotonic dystrophy. *Mol. Genet. Metab.* 74:281–91
35. Froehlich AC, Liu Y, Loros JJ, Dunlap JC. 2002. White collar-1, a circadian blue light photoreceptor, binding to the frequency promoter. *Science* 297:815–19
36. Fuke S, Suo S, Takahashi N, Koike H, Sasagawa N, Ishiura S. 2001. The VNTR polymorphism of the human dopamine transporter (DAT1) gene affects gene expression. *Pharmacogenomics J.* 1:152–56
37. Galvao R, Mendes-Soares L, Camara J, Jaco I, Carmo-Fonseca M. 2001. Triplet repeats, RNA secondary structure and toxic gain-of-function models for pathogenesis. *Brain Res. Bull.* 56:191–201
38. Garner MM, Felsenfeld G. 1987. Effect of Z-DNA on nucleosome placement. *J. Mol. Biol.* 196:581–90
39. Gatchel JR, Zoghbi HY. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.* 6:743–55
40. Gauthier LR, Charrin BC, Borrell-Pages M, Dompierre JP, Rangone H, et al. 2004. Huntingtin controls neurotrophic support and survival of neurons by enhancing BDNF vesicular transport along microtubules. *Cell* 118:127–38
41. Gebhardt F, Zanker KS, Brandt B. 1999. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J. Biol. Chem.* 274:13176–80
42. Godde JS, Wolffe AP. 1996. Nucleosome assembly on CTG triplet repeats. *J. Biol. Chem.* 271:15222–29
43. Grabczyk E, Kumari D, Usdin K. 2001. Fragile X syndrome and Friedreich's ataxia: Two different paradigms for repeat induced transcript insufficiency. *Brain Res. Bull.* 56:367–73
- 44. Hamada H, Seidman M, Howard BH, Gorman CM. 1984. Enhanced gene expression by the poly(dT-dG)-poly(dC-dA) sequence. *Mol. Cell. Biol.* 4:2622–30**

32. Provides strong evidence implicating variable tandem repeats as mediators of rapid evolution of skull morphology in domestic dogs.

44. One of the earliest demonstrations that tandem repeats can act as promoter elements that influence gene expression level.

45. Hammock EA, Young LJ. 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* 308:1630–34
46. Hartl DL. 2000. Molecular melodies in high and low C. *Nat. Rev. Genet.* 1:145–49
47. Hefferon TW, Groman JD, Yurk CE, Cutting GR. 2004. 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* 101:3504–9
48. Helmlinger D, Hardy S, Abou-Sleymane G, Eberlin A, Bowman AB, et al. 2006. Glutamine-expanded ataxin-7 alters TFTC/STAGA recruitment and chromatin structure leading to photoreceptor dysfunction. *PLoS Biol.* 4:432–45
49. Hoyer LL. 2001. The ALS gene family of *Candida albicans*. *Trends Microbiol.* 9:176–80
50. The Huntington's Disease Collaborative Research Group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72:971–83
51. Iglesias AR, Kindlund E, Tammi M, Wadelius C. 2004. Some microsatellites may act as novel polymorphic cis-regulatory elements through transcription factor binding. *Gene* 341:149–65
52. Iyer V, Struhl K. 1995. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* 14:2570–79
53. Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.* 10:161–72
54. Jin TQ, Ito Y, Luan XH, Dangaria S, Walker C, et al. 2009. Elongated polyproline motifs facilitate enamel evolution through matrix subunit compaction. *PLoS Biol.* 7:e1000262
55. Johnsen A, Fidler AE, Kuhn S, Carter KL, Hoffmann A, et al. 2007. Avian clock gene polymorphism: evidence for a latitudinal cline in allele frequencies. *Mol. Ecol.* 16:4867–80
56. Johnson AC, Jinno Y, Merlino GT. 1988. Modulation of epidermal growth factor receptor proto-oncogene transcription by a promoter site sensitive to S1 nuclease. *Mol. Cell. Biol.* 8:4174–84
57. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458:362–66
58. Karlin S, Burge C. 1996. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl. Acad. Sci. USA* 93:1560–65
59. Kawakami K, Salonga D, Park JM, Danenberg KD, Uetake H, et al. 2001. Different lengths of a polymorphic repeat sequence in the thymidylate synthase gene affect translational efficiency but not its gene expression. *Clin. Cancer Res.* 7:4096–101
60. Kirkpatrick DT, Wang YH, Dominska M, Griffith JD, Petes TD. 1999. Control of meiotic recombination and gene expression in yeast by a simple repetitive DNA sequence that excludes nucleosomes. *Mol. Cell. Biol.* 19:7661–71
- 60a. King DG, Soller M, Kashi Y. 1997. Evolutionary tuning knobs. *Endeavour* 21:36–40**
61. Kita E, Katsui N, Emoto M, Sawaki M, Oku D, et al. 1991. Virulence of transparent and opaque colony types of *Neisseria gonorrhoeae* for the genital tract of mice. *J. Med. Microbiol.* 34:355–62
62. Kokoska RJ, Stefanovic L, Tran HT, Resnick MA, Gordenin DA, Petes TD. 1998. Destabilization of yeast micro- and minisatellite DNA sequences by mutations affecting a nuclease involved in Okazaki fragment processing (rad27) and DNA polymerase delta (pol3-t). *Mol. Cell. Biol.* 18:2779–88
63. Koob MD, Moseley ML, Schut LJ, Benzow KA, Bird TD, et al. 1999. An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nat. Genet.* 21:379–84
64. Kuyumcu-Martinez NM, Wang GS, Cooper TA. 2007. Increased steady-state in levels of CUGBP1 in myotonic dystrophy 1 are due to PKC-mediated hyperphosphorylation. *Mol. Cell* 28:68–78
65. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. 1991. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* 352:77–79
66. Lam YC, Bowman AB, Jafar-Neiad P, Lim J, Richman R, et al. 2006. ATAXIN-1 interacts with the repressor capicua in its native complex to cause SCA1 neuropathology. *Cell* 127:1335–47
67. Lee K, Dunlap JC, Loros JJ. 2003. Roles for WHITE COLLAR-1 in circadian and general photoperception in *Neurospora crassa*. *Genetics* 163:103–14
- 68. Legendre M, Pochet N, Pak T, Verstrepen KJ. 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.* 17:1787–96**

60a. Proposes for tandem repeats the role of genetic tuning knobs that evolution can act upon.

68. Presents an experimentally validated model for predicting tandem repeat variability in a wide range of organisms.

69. Lesch KP, Bengel D, Heils A, Sabol SZ, Greenberg BD, et al. 1996. Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science* 274:1527–31
70. Lesuisse E, Santos R, Matzanke BF, Knight SAB, Camadro JM, Dancis A. 2003. Iron use for haeme synthesis is under control of the yeast frataxin homologue (Yfh1). *Hum. Mol. Genet.* 12:879–89
71. Li YC, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.* 11:2453–65
72. Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, et al. 2001. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science* 293:864–67
73. Liu G, Bissler JJ, Sinden RR, Leffak M. 2007. Unstable spinocerebellar ataxia type 10 (ATTCT)*(AGAAT) repeats are associated with aberrant replication at the ATX10 locus and replication origin-dependent expansion at an ectopic site in human cells. *Mol. Cell. Biol.* 27:7828–38
74. Lohi H, Young EJ, Fitzmaurice SN, Rusbridge C, Chan EM, et al. 2005. Expanded repeat in canine epilepsy. *Science* 307:81
75. Lopes J, Debrauwere H, Buard J, Nicolas A. 2002. Instability of the human minisatellite CEB1 in rad27Delta and dna2–1 replication-deficient yeast cells. *EMBO J.* 21:3201–11
76. Lu X, Timchenko NA, Timchenko LT. 1999. Cardiac clav-type RNA-binding protein (ETR-3) binds to RNA CUG repeats expanded in myotonic dystrophy. *Hum. Mol. Genet.* 8:53–60
77. Mahadevan M, Tsilfidis C, Sabourin L, Shutler G, Amemiya C, et al. 1992. Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* 255:1253–55
78. **Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER. 2005. Microsatellite instability regulates transcription factor binding and gene expression. *Proc. Natl. Acad. Sci. USA* 102:3800–4**
79. Matsuura T, Yamagata T, Burgess DL, Rasmussen A, Grewal RP, et al. 2000. Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat. Genet.* 26:191–94
80. Merkel A, Gemmell N. 2008. Detecting short tandem repeats from genome data: opening the software black box. *Brief. Bioinform.* 9:355–66
81. Metzgar D, Liu L, Hansen C, Dybvig K, Wills C. 2002. Domain-level differences in microsatellite distribution and content result from different relative rates of insertion and deletion mutations. *Genome Res.* 12:408–13
82. Michael TP, Park S, Kim TS, Booth J, Byer A, et al. 2007. Simple sequence repeats provide a substrate for phenotypic variation in the *Neurospora crassa* circadian clock. *PLoS ONE* 2:10
83. Miller JW, Urbinati CR, Teng-umnuay P, Stenberg MG, Byrne BJ, et al. 2000. Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J.* 19:4439–48
84. Mittelman D, Sykoudis K, Hersh M, Lin Y, Wilson JH. 2010. Hsp90 modulates CAG repeat instability in human cells. *Cell Stress Chaperones* 15:753–59
85. Montermini L, Andermann E, Labuda M, Richter A, Pandolfo M, et al. 1997. The Friedreich ataxia GAA triplet repeat: premutation and normal alleles. *Hum. Mol. Genet.* 6:1261–66
86. Mrazek J, Guo XX, Shah A. 2007. Simple sequence repeats in prokaryotic genomes. *Proc. Natl. Acad. Sci. USA* 104:8472–77
87. Mularoni L, Ledda A, Toll-Riera M, Alba MM. 2010. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res.* 20:745–54
88. Myers S, Freeman C, Auton A, Donnelly P, McVean G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* 40:1124–29
89. Naylor LH, Clark EM. 1990. d(TG)n-d(CA)n sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. *Nucleic Acids Res.* 18:1595–601
90. Nithianantharajah J, Hannan AJ. 2007. Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. *BioEssays* 29:525–35
91. Nucifora FC Jr, Sasaki M, Peters MF, Huang H, Cooper JK, et al. 2001. Interference by huntingtin and atrophin-1 with cbp-mediated transcription leading to cellular toxicity. *Science* 291:2423–28
92. Oh DB, Kim YG, Rich A. 2002. Z-DNA-binding proteins can act as potent effectors of gene expression in vivo. *Proc. Natl. Acad. Sci. USA* 99:16666–71
93. Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Looreesuwan S, Tokunaga K. 2002. Significant association of longer forms of CCTTT microsatellite repeat in the inducible nitric oxide synthase promoter with severe malaria in Thailand. *J. Infect. Dis.* 186:578–81

78. Illustrates how tandem repeats can alter regulatory binding sites and mediate phase variation.

94. Ohno S. 1972. So much “junk” DNA in our genome. *Brookhaven Symp. Biol.* 23:366–70
95. Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* 284:604–7
- 96. Orr H, Zoghbi H. 2007. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* 30:575–621**
97. Panov AV, Gutekunst CA, Leavitt BR, Hayden MR, Burke JR, et al. 2002. Early mitochondrial calcium defects in Huntington’s disease are a direct effect of polyglutamines. *Nat. Neurosci.* 5:731–36
98. Paques F, Leung WY, Haber JE. 1998. Expansions and contractions in a tandem repeat induced by double-strand break repair. *Mol. Cell. Biol.* 18:2045–54
99. Paquette J, Giannoukakis N, Polychronakos C, Vafiadis P, Deal C. 1998. The INS 5’ variable number of tandem repeats is associated with IGF2 expression in humans. *J. Biol. Chem.* 273:14158–64
100. Punga T, Buhler M. 2010. Long intronic GAA repeats causing Friedreich ataxia impede transcription elongation. *EMBO Mol. Med.* 2:120–29
101. Richard GF, Dujon B. 2006. Molecular evolution of minisatellites in hemiascomycetous yeasts. *Mol. Biol. Evol.* 23:189–202
- 102. Richard GF, Kerrest A, Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.* 72:686–727**
103. Richard GF, Paques F. 2000. Mini- and microsatellite expansions: the recombination connection. *EMBO Rep.* 1:122–26
104. Riley DE, Krieger JN. 2005. Short tandem repeat (STR) replacements in UTRs and introns suggest an important role for certain STRs in gene expression and disease. *Gene* 344:203–11
105. Riley DE, Krieger JN. 2009. Embryonic nervous system genes predominate in searches for dinucleotide simple sequence repeats flanked by conserved sequences. *Gene* 429:74–79
106. Rockman MV, Hahn MW, Soranzo N, Loisel DA, Goldstein DB, Wray GA. 2004. Positive selection on MMP3 regulation has shaped heart disease risk. *Curr. Biol.* 14:1531–39
107. Rockman MV, Wray GA. 2002. Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* 19:1991–2004
108. Rosenberg SM. 2001. Evolving responsively: adaptive mutation. *Nat. Rev. Genet.* 2:504–15
109. Rothenburg S, Koch-Nolte F, Rich A, Haag F. 2001. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc. Natl. Acad. Sci. USA* 98:8985–90
110. Rubinsztein DC, Leggo J, Coles R, Almqvist E, Biancalana V, et al. 1996. Phenotypic characterization of individuals with 30–40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36–39 repeats. *Am. J. Hum. Genet.* 59:16–22
111. Ruiz-Herrera A, Castresana J, Robinson TJ. 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biol.* 7:R115
112. Sandman K, Reeve JN. 1999. Archaeal nucleosome positioning by CTG repeats. *J. Bacteriol.* 181:1035–38
113. Sarkari J, Pandit N, Moxon ER, Achtman M. 1994. Variable expression of the Opc outer membrane protein in *Neisseria meningitidis* is caused by size variation of a promoter containing poly-cytidine. *Mol. Microbiol.* 13:207–17
114. Sawyer LA, Hennessy JM, Peixoto AA, Rosato E, Parkinson H, et al. 1997. Natural variation in a *Drosophila* clock gene and temperature compensation. *Science* 278:2117–20
115. Sawyer LA, Sandrelli F, Pasetto C, Peixoto AA, Rosato E, et al. 2006. The period gene Thr-Gly polymorphism in Australian and African *Drosophila melanogaster* populations: implications for selection. *Genetics* 174:465–80
116. Schmidt AL, Mitter V. 2004. Microsatellite mutation directed by an external stimulus. *Mutat. Res.* 568:233–43
117. Sekinger EA, Moqtaderi Z, Struhl K. 2005. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell* 18:735–48
118. Sia EA, Jinks-Robertson S, Petes TD. 1997. Genetic control of microsatellite stability. *Mutat. Res.* 383:61–70
119. Simon M, Hancock JM. 2009. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol.* 10(6):R59
120. Sobczak K, de Mezer M, Michlewski G, Krol J, Krzyzosiak WJ. 2003. RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Res.* 31:5469–82

96. Covers the genetics, pathologies and molecular mechanisms of disorders of trinucleotide repeat expansions.

102. A comprehensive review of the various types of repeats in eukaryotic genomes.

121. Srikhanta YN, Dowideit SJ, Edwards JL, Falsetta ML, Wu HJ, et al. 2009. Phasevarions mediate random switching of gene expression in pathogenic *Neisseria*. *PLoS Pathogens* 5(4):e1000400
122. Staib P, Kretschmar M, Nichterlein T, Hof H, Morschhauser J. 2002. Host versus in vitro signals and intrastrain allelic differences in the expression of a *Candida albicans* virulence gene. *Mol. Microbiol.* 44:1351–66
123. Stern A, Brown M, Nickel P, Meyer TF. 1986. Opacity genes in *Neisseria gonorrhoeae*: control of phase and antigenic variation. *Cell* 47:61–71
124. Streelman JT, Kocher TD. 2002. Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. *Physiol. Genomics* 9:1–4
125. Sureshkumar S, Todesco M, Schneeberger K, Harilal R, Balasubramanian S, Weigel D. 2009. A genetic defect caused by a triplet repeat expansion in *Arabidopsis thaliana*. *Science* 323:1060–63
126. Tachida H, Iizuka M. 1992. Persistence of repeated sequences that evolve by replication slippage. *Genetics* 131:471–78
- 127. Tautz D, Trick M, Dover GA. 1986. Cryptic simplicity in DNA is a major source of genetic variation. *Nature* 322:652–56**
128. Thierry A, Bouchier C, Dujon B, Richard GF. 2008. Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in *Candida glabrata*. *Nucleic Acids Res.* 36:5970–82
129. Tian B, White RJ, Xia T, Welle S, Turner DH, et al. 2000. Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA* 6:79–87
130. Tirosch I, Barkai N, Verstrepen KJ. 2009. Promoter architecture and the evolvability of gene expression. *J. Biol.* 8:95
131. Tiscornia G, Mahadevan MS. 2000. Myotonic dystrophy: the role of the CUG triplet repeats in splicing of a novel DMPK exon and altered cytoplasmic DMPK mRNA isoform ratios. *Mol. Cell* 5:959–67
132. Tomita N, Fujita R, Kurihara D, Shindo H, Wells RD, Shimizu M. 2002. Effects of triplet repeat sequences on nucleosome positioning and gene expression in yeast minichromosomes. *Nucleic Acids Res.* (Suppl. 2):231–32
133. van Ham SM, van Alphen L, Mooi FR, van Putten JP. 1993. Phase variation of *H. influenzae* fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. *Cell* 73:1187–96
134. Verkerk A, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DPA, et al. 1991. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile-X syndrome. *Cell* 65:905–14
- 135. Verstrepen KJ, Jansen A, Lewitter F, Fink GR. 2005. Intragenic tandem repeats generate functional variability. *Nat. Genet.* 37:986–90**
- 136. Vincs MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* 324:1213–16**
137. Wang YH, Griffith JD. 1996. The [(G/C)3NN]n motif: a common DNA repeat that excludes nucleosomes. *Proc. Natl. Acad. Sci. USA* 93:8863–67
138. Warpeha KM, Xu W, Liu L, Charles IG, Patterson CC, et al. 1999. Genotyping and functional analysis of a polymorphic (CCTTT)(n) repeat of NOS2A in diabetic retinopathy. *FASEB J.* 13:1825–32
139. Weber JL, Wong C. 1993. Mutation of human short tandem repeats. *Hum. Mol. Genet.* 2:1123–28
140. Weiser JN, Love JM, Moxon ER. 1989. The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. *Cell* 59:657–65
141. Wheeler TM, Lueck JD, Swanson MS, Dirksen RT, Thornton CA. 2007. Correction of CIC-1 splicing eliminates chloride channelopathy and myotonia in mouse models of myotonic dystrophy. *J. Clin. Investig.* 117:3952–57
142. Whetstine JR, Witt TL, Matherly LH. 2002. The human reduced folate carrier gene is regulated by the AP2 and sp1 transcription factor families and a functional 61-base pair polymorphism. *J. Biol. Chem.* 277:43873–80
143. Wierdl M, Greene CN, Datta A, Jinks-Robertson S, Petes TD. 1996. Destabilization of simple repetitive DNA sequences by transcription in yeast. *Genetics* 143:713–21
144. Willems R, Paul A, van der Heide HG, ter Avest AR, Mooi FR. 1990. Fimbrial phase variation in *Bordetella pertussis*: a novel mechanism for transcriptional regulation. *EMBO J.* 9:2803–9

127. First to suggest the importance of tandem repeats as a significant source of genetic diversity.

135. Provides experimental evidence for the role of variable tandem repeats in mediating quantitative phenotypic changes.

136. Experimentally shows that variable tandem repeats in promoters can mediate the evolution of gene expression.

145. Xu W, Liu L, Emson PC, Harrington CR, Charles IG. 1997. Evolution of a homopurine-homopyrimidine pentanucleotide repeat sequence upstream of the human inducible nitric oxide synthase gene. *Gene* 204:165–70
146. Yogev D, Rosengarten R, Watson-McKown R, Wise KS. 1991. Molecular basis of *Mycoplasma* surface antigenic variation: a novel set of divergent genes undergo spontaneous mutation of periodic coding regions and 5' regulatory sequences. *EMBO J.* 10:4069–79
147. Young ET, Sloan JS, Van Riper K. 2000. Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. *Genetics* 154:1053–68
148. Young MW, Kay SA. 2001. Time zones: a comparative genetics of circadian clocks. *Nat. Rev. Genet.* 2:702–15
149. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, et al. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 309:626–30
150. Zuccato C, Tartari M, Crotti A, Goffredo D, Valenza M, et al. 2003. Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nat. Genet.* 35:76–83