5-2006

Simple Sequence Repeats as Advantageous Mutators in Evolution

Yechezkel Kashi
Israel Institute of Technology

David G. King
Southern Illinois University Carbondale

Published in *Trends in Genetics*, Vol. 22, No. 5 (May 2006) at 10.1016/j.tig.2006.03.005

Trends in Genetics, in press (anticipated publication, May 2006)

Simple Sequence Repeats as Advantageous Mutators in Evolution

Yechezkel Kashi
Department of Biotechnology and Food Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel

and

David G. King
Department of Anatomy
Southern Illinois University Carbondale
Carbondale, IL 62901, U.S.A.

Corresponding author: Kashi, Y. (kashi@techunix.technion.ac.il)

ABSTRACT Simple Sequence Repeats (SSRs) often serve to modify genes with which they are associated. The influence of SSRs on gene regulation, transcription, and protein function typically depends on the number of repeats, while mutations that add or subtract repeat units are both frequent and reversible. SSRs thus provide a prolific source of quantitative and qualitative variation. Over the past decade, a number of researchers have found that this spontaneous variation has been tapped by natural as well as artificial selection to adjust nearly every aspect of gene function. These studies support the hypothesis that SSRs, by virtue of their special mutational and functional qualities, play a major role in generating the genetic variation underlying adaptive evolution.

Introduction

Simple Sequence Repeats (SSRs, also called microsatellites and minisatellites) are mutation-prone DNA tracts composed of tandem repetitions of relatively short motifs. Although SSRs are commonly regarded as "junk" (i.e., with no significant role as genomic information), accumulating evidence for many molecular and phenotypic effects of SSR repeat-number variation has lent growing support to the hypothesis that SSRs could play a positive role in adaptive evolution [1-20]. Indeed, from an evolutionary perspective, the properties of these remarkable sequences [Box 1] confer virtually ideal "mutator" properties. SSR instability may be indirectly advantageous by supplying abundant quantitative genetic variation with minimal genetic load, while variation in repetition purity and motif length allow site-specific adjustment of both mutation rate and mutation effect.

Here we highlight positive evidence from a few recent reports that support an evolutionary role for SSRs as important sources of adaptive genetic variation, both within and between species. In contrast to many other studies that simply demonstrate effective functional differences between "normal" and "mutant" SSR alleles, these examples offer evidence that common SSRs alleles can offer potential selective advantages. This shall be followed by an overview of the molecular basis for the functional effects of SSRs in both coding and non-coding domains, and a brief consideration of the evolutionary benefit for SSR mutability.

Temperature compensation of circadian rhythm in *Drosophila*

The first thoroughly-documented eukaryotic case, with evidence not only for quantitative phenotypic effects of repeat-number alleles but also for natural selection acting upon those

alleles, came from investigation by Sawyer *et al.* of the clock gene *period* in the fruit fly *Drosophila melanogaster* [21]. This gene contains an SSR with a variable-length repeating hexanucleotide motif encoding threonine-glycine. Of the two most common alleles of this gene, at warm temperature the shorter $(Thr\text{-}Gly)_{17}$ allele yields a period closer to 24 hours, while the longer $(Thr\text{-}Gly)_{20}$ variant yields better temperature compensation so that temperature fluctuations have a lesser impact on circadian cycle. Across Europe and northern Africa, the frequencies of these two alleles display a significant latitudinal cline, with the longer allele predominating in colder regions. Such a pattern is to be expected if these alleles were selected by climate, based on the differential temperature sensitivity that they confer [21].

Additional evidence has recently come from the "Evolution Canyon" ecological study site at Mount Carmel, Israel. This canyon presents a dramatic microclimatic contrast, with the sunny, south-facing slope experiencing higher temperature and drought stress than the north-facing slope. Resulting biotic differences occur between ecological zones separated by only 100 m at the bottom and 400 m at the top. The longer, cold-climate allele of the *Drosophila per* gene was more than twice as abundant on the cooler, north-facing slope than on the warmer, sunny slope, supporting the conclusion that natural selection of these microsatellite alleles "fine-tunes" the *Drosophila* circadian clock to differing environmental conditions [22].

Adaptive divergence among barley and wheat populations

The "Evolution Canyon" site has also furnished much more extensive evidence that ecological (i.e., fitness-related) parameters affect SSR allele frequencies in a natural setting. Analysis of 19 nuclear and 4 chloroplast microsatellite loci in 7 populations of wild barley (*Hordeum spontaneum*) distributed across the canyon's north- and south-facing slopes has revealed significant interslope differentiation of SSR allele frequencies [23]. Similarly, analysis of 20 microsatellites in 15 emmer wheat populations (*Triticum dicoccoides*) at sites in Israel and Turkey also yielded SSR allele distribution patterns correlated with physical conditions [24]. These results indicate that frequencies of both coding and noncoding SSR alleles have been shaped by natural selection acting through microclimatic factors. Since the specific roles played by SSRs in these grasses remain unknown (like those for most SSRs), conclusive evidence that SSRs themselves are being selected will require further research.

Social behavior in voles

Direct experimental evidence that allelic variation at an SSR locus is intimately involved in phenotypic variation at the interspecies as well as at the individual level has recently been provided by Hammock and Young's elegant study of social behavior in voles (*Microtus*) [18, 25]. Prairie and pine voles (*M. ochrogaster* and *M. pinetorum*) are highly social, monogamous rodent species, while montane and meadow voles (*M. montanus* and *M. pennsylvanicus*) are asocial and non-monogamous. These differing social behaviors depend on the pattern of expression for the vasopressin receptor *avpr1a* gene, with higher levels of expression in the ventral forebrain of the social voles. (Increasing expression of this gene, using viral vector transfer into the ventral pallidum, can increase partner preference behavior in a normally non-monogamous species [26].) Although the protein-coding region of the *avpr1a* gene is highly conserved among voles, the two social species have a long, compound SSR in the 5' regulatory region of this gene, much of which is absent in the two asocial species. (Interestingly, bonobos (*Pan paniscus*) and humans, two primate species characterized by high empathic and sexual bonding, also share a highly homologous SSR-rich tract upstream of the *avpr1a* gene, while the corresponding sequence of the less-empathic chimpanzee (*Pan troglodytes*) presents a substantial deletion of this region [25].)

Experiments transfecting two versions of the SSR locus from social and asocial species into cultured rat cells showed that the species divergence in SSR lengths at this locus is sufficient to alter expression of the *avpr1a* gene in a manner that is dependent on cell type. A transgenic mouse containing the social species'

version of the SSR locus displayed gene expression patterns, as well as behaviors in response to experimental vasopressin injection, that were more like those of the social species than those of the wild-type mouse [27]. Furthermore, the long, compound SSR locus of prairie voles also shows repeat number variation among individual animals. When two different alleles from this social species were transfected into rat A7r5 cells, while holding constant the rest of the regulatory region, the allele with an expanded GA repeat yielded higher levels of gene expression. And when individual prairie voles were selectively bred for longer and shorter alleles of this polymorphic SSR, the "fine-tuning" effect of this polymorphic SSR was demonstrated by correlation of repeat length with quantitative differences both in the distribution of the vasopressin receptor in individual brains and also in individual social behavior, with longer-allele males showing "greater probability of social engagement and bonding behavior" [25].

Such effects of SSR repeat number on cell-type-specific gene expression in culture together with correlation of SSR repeat length with social behavior and gene expression in intact animals support a strong presumption that SSR variation, mediated through expression of the vasopressin receptor gene, is at least partially responsible for both individual and interspecies variation in vole social behavior phenotypes.

Skeletal morphology in domestic dogs

A different line of evidence showing that variation generated by SSRs can supply raw material for evolutionary divergence in phenotype has recently been provided by Fondon and Garner's [17] analysis of 92 breeds of domestic dogs (*Canis lupus familiaris*).

When Fondon and Garner examined 17 genes known to influence morphological traits, they found "only a few silent SNPs". In contrast, the same genes showed "extraordinary levels of tandem repeat variation", with some polymorphism in nearly every gene examined. Furthermore, the exceptional purity of repetition in these morphogenetic genes, in contrast with less-perfect repeats at other sites, suggests that diversifying selection has followed purifying mutational slippage too recently to permit the accumulation of new point mutations.

Although the function of most of the observed SSR polymorphism remains unknown, Fondon and Garner [17] found that the length ratio of two adjacent SSRs in the runt-related transcription factor *Runx-2*, encoding 18-20 glutamines followed by 12-17 alanines, was correlated with measures of facial shape across breeds. In humans, the homologous *CBFA1* gene, which encodes osteoblast-specific transcription factor *OSF2*, is known to influence craniofacial structure, and an expansion of the alanine stretch from 17 to 27 has been found in members of one human family who are afflicted with cleidocranial dysplasia [28]. Fondon and Garner also found that in Great Pyrenees, a dog breed characterized by polydactyly, the presence of extra toes was consistently linked with a 51 bp contraction of a hexanucleotide repeat in *Alx-4*, a gene previously associated with polydactyly in mice.

This evidence strongly suggests that genetic variation supplied by SSRs is at least partially responsible for phenotypic differences among individual dogs and for morphological divergence among dog breeds. Although the traits that distinguish dog breeds have been shaped by human breeders, there is no reason to suppose that artificial selection draws on a source of variation any different from that which sustains natural selection.

Sporulation efficiency and cell adhesion in yeast

A recent study of quantitative trait loci controlling sporulation efficiency in a cross of two differing strains of budding yeast (*Saccharomyces cerevisiae*) identified *RAS2* (a homologue of the RAS proto-oncogenes) as one of the genes affecting this trait (G. Ben-Ari, PhD Thesis, the Hebrew University of Jerusalem, 2005). The promoter regions of the high- and low-efficiency alleles were distinguished by the presence of A_9 and A_{10} poly-A tracts, respectively. Replacement of the original *RAS2* allele in a laboratory strain (S288c) by the corresponding longer allele, using "knock-in" technology, reduced sporulation efficiency from 17.1% to 0.7%. In a parallel study of ten wine-

yeast strains, found to be almost identical genetically and characterized for sporulation efficiency, the A_9 tract was found in six strains with sporulation efficiencies of 15-55% while the A_{10} tract was found in four strains that did not sporulate at all. These findings strongly implicate this mononucleotide-repeat polymorphism as a causal basis for differentiation in sporulation efficiency, a significant life-history trait for yeast. More generally, a regulatory role for mononucleotide SSRs could be extremely important, since mononucleotide repeats comprise the most numerous class of SSRs in most genomes [29, 30, 31].

Much longer repeats (minisatellites) have also been investigated in *S. cerevisiae*, where they seem to occur predominantly in genes encoding cell-surface proteins involved in cell adhesion and flocculation [32]. These genes display substantial in-frame repeat-number variation among yeast strains, with the frequency of repeat-number mutations being dependent on several *RAD* genes. Experimental manipulation of repeat length has demonstrated a linear correlation between repeat number and the extent of cell adhesion. Variation in repeat length thus appears capable of permitting gradual and fully reversible functional changes, in turn allowing rapid evolutionary adaptation to particular environments [32].

Repeat-related diseases in man

Allelic differences in SSR repeat number are known to cause a wide range of hereditary disorders and disease susceptibilities in humans, most notoriously the "triplet repeat diseases" [e.g., 6, 9, 15, 16, 33]. Although such cases effectively illustrate many of the ways in which repeat number can affect genetic function, they can also convey a misleading impression that any non-neutral effects of repeat-number mutation are predominantly deleterious. One might expect that such deleterious effects would lead to evolutionary elimination, or at least to selection for reduced mutability of such sites. However, the widespread occurrence of unstable SSRs in many functional sites argues against such an impression. Some evidence hints that even apparently deleterious SSR alleles might convey some unexpected advantage and be preserved by evolutionary selection. For example, the long "premutation" allele of a CAG repeat in the human spinocerebellar ataxia gene *SCA2* occurs at unusually high frequency, given its propensity for pathological expansion. Preliminary evidence from extended haplotype analysis suggests recent positive selection in a human population with northern European ancestry [34]. Similarly, haplotype data suggest that positive selection in northern Europe may have increased the frequency of the shorter of two alleles of a thymidine repeat at a transcription factor binding site in a human matrix metalloproteinase gene (*MMP3*), in spite of this allele's association with heart disease risk [35]. Although such evidence remains weak, it does suggest the possibility that even disease-related SSR alleles may contribute evolutionarily advantageous effects.

Molecular basis for adaptive effects of SSRs

The studies described above highlight the potential adaptive significance of variation generated by SSRs. Documenting the functional effects of SSR alleles remains challenging, however, even when they appear within genes whose functions have been established, such as fruit fly *period*, vole *avpr1a*, dog *Alx-4* and *Runx-2*, and yeast *RAS2*. Ideally, an incremental effect of repeat number should be demonstrated over a range of quantitative phenotypic differences. Although a few studies have provided data from multiple alleles [e.g., 4, 17, 32], and the triplet repeat diseases also show dependence on repeat number, many more examples report effect differences between two alleles only. Nevertheless, current evidence indicates that the number of repeats in many different SSRs can affect gene function in any of several different ways.

Triplets (i.e., individual codons) comprise by far the most common motif length for SSRs located within protein-coding domains [29, 30, 36, 37]. Triplet repeats are especially common in genes encoding transcription factors [4, 6, 13, 15, 33, 38, 39]. Variation in the number of repeated codons yields variation in the length of homopolymeric amino acid stretches that in turn can affect such properties as protein flexibility

and binding affinity. Examples associated with human triplet repeat diseases are the most thoroughly studied, with literature too extensive to review here [e.g., 6, 15, 33]. Motif lengths that are multiples of three are also common. For example, many eukaryotic structural and cell surface proteins appear to have evolved by repeat expansion of minisatellites, with each motif encoding an oligopeptide [32, 40, 41].

SSRs with motif lengths that are not multiples of three bp can also encode protein segments. Although such SSRs have not received nearly as much attention as triplet repeats, they are nevertheless found in many genes. Repeat number mutations in coding non-triplet SSRs cause frameshifts that can effectively inactivate gene expression or code for different or shorter protein sequences in the alternative form. Because frameshifting based on SSR mutation is readily reversible by subsequent mutation, such SSRs can function like on/off switches for their genes. Although this SSR effect can cause cancer [42], some bacteria apply it in "contingency genes" to control variable expression of surface antigens [14, 43]. Nontriplet (mononucleotide) repeats are also exceptionally prevalent in coding regions of minor mismatch repair system genes of many eukaryotes [44], where repeat number variation would permit mutation rates to be modulated over evolutionary time.

Another intriguing possibility for SSR-based gene switching is suggested by a short poly-C tract in the *MC1R* gene for a melanocortin receptor expressed in pig melanocytes. Frameshifting caused by germ-line addition of an extra C in this SSR leads to loss of pigmentation, while somatic cell reversions to the original tract length occur at relatively high frequency during skin development, creating a pattern of black spots [45]. A similar mechanism could usefully generate somatic cell variety during embryogenesis of other tissues.

Effects of coding SSRs may be surprisingly sophisticated. As noted above, the *Runx-2* gene analyzed by Fondon and Garner contains a compound repeat in which the length ratio of two adjacent SSRs correlates with facial shape much more strongly than does the length of either repeat alone. This suggests that precise modulation of transcription by the *Runx-2* protein could be facilitated by the pairing of repeats with opposing activities [17]. In effect, a compound SSR appears to represent the functional equivalent of a micrometer in which two relatively coarse screws of slightly different pitch work in opposite directions to allow finer adjustment than could be attained with either screw by itself.

SSRs effects are not limited to coding sequences. Repeat variation commonly exerts a functional influence on DNA structure and transcription activity even when the SSRs are located in introns or other noncoding sites where they do not affect protein structure directly. Examples of several such SSR effects are presented in Box 2. Additional examples are reviewed elsewhere [e.g., 16, 19, 20]. Three basic principles extend through all this diversity. (1) First, whatever role an SSR plays within genes, whether coding or noncoding, whether within transcripts or regulatory sequences, changing the number of repeats can modulate its genetic function. (2) Second, mutations which alter repeat number typically occur at rates orders of magnitude higher than single-nucleotide point mutations. (3) Third, the mutation rates associated with SSR sites are influenced, among other factors, by site-specific features including motif length, number of repeats, and purity of repetition [33, 46-49].

Evolutionary utility of SSRs

Any genomic variable that routinely affects genetic function must surely play an evolutionary role as well. It is time to abandon the presumption that SSRs are "junk DNA" [Box 3]. Our 1997 proposal, that SSRs "provide a ready and virtually inexhaustible supply of new quantitative variation for rapid evolutionary adaptation" [7] has been echoed by Fondon and Garner's recent hypothesis that "gene-associated tandem repeats function as facilitators of evolution, providing abundant, robust variation and thus enabling extremely rapid evolution of new forms" [17].

A metaphorical characterization of SSRs as "evolutionary tuning knobs" [8] expresses each SSRs' potential to facilitate the efficient adaptive adjustment of a quantitative trait. Yet the sheer

number of SSRs is staggering. The human genome contains close to a million mononucleotide repeats longer than 9 bp, while longer motifs account for many more SSR sites [31]. If even a small fraction of these many, diverse SSRs are functionally active, their high mutability implies that the quantitative genome is in a constant state of mutational ferment. Indeed, we believe not only that SSRs contribute adaptively significant variation, but that provision of such variation may be SSRs' evolved "function". That is, indirect selection (see Glossary) may encourage the presence of large numbers of SSR tracts in the genome and endow these tracts with their special mutator properties [8, 12, 20, 50; also see Box 1].

In a changeable world, long-term stability of fitness is found in the adaptive variation that mutability provides. Implicit in the genome are many "ingenious and unexpected mechanisms", or "protocols" [51, 52], for regulating, modifying, and restructuring genetic information with minimal risk to ongoing adaptation. The quantitative adjustment and on/off switching provided by site specific mutation of SSRs may be one of the simplest of these protocols, but it may also be one of the most widespread and powerful means of providing genetic variation for evolution. This hypothesis raises several questions (see Questions Box) which should be addressed by direct experiment as well as by comparative analysis of genome sequence data.

ACKNOWLEDGEMENT

We thank *TiG*'s editor and referees for pertinent and inciteful advice, especially John Fondon III for calling attention to the importance of data from multiple repeat-number alleles at any given "tuning knob" locus.

REFERENCES

1. Hamada, H. *et al*. (1984) Enhanced gene expression by the poly(dT-dG) · poly(dC-dA) sequence. *Mol. Cell Biol.* **4**, 2622-2630

2. Trifonov, E. N. (1989) The multiple codes of nucleotide sequences. *Bull. Math. Biol.* **51**, 417-432

3. Kashi, Y. *et al*. (1990) Large restriction fragments containing poly-TG are highly polymorphic in a variety of vertebrates. *Nucleic Acids Res*. **18**, 1129-1132.

4. Gerber, H.-P. *et al*. (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263**, 808-811.

5. King, D.G. (1994) Triplet repeat DNA as a highly mutable regulatory mechanism. *Science* 263:595-596.

6. Künzler, P. *et al.* (1995) Pathological, physiological, and evolutionary aspects of short unstable DNA repeats in the human genome. *Biol. Chem. Hoppe Seyler* **376**, 201-211

7. Kashi, Y. *et al*. (1997) Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* **13**, 74-78

8. King, D. G. *et al.* (1997) Evolutionary Tuning Knobs. *Endeavour* **21**, 36-40

9. Comings, D. E. (1998) Polygenic inheritance and micro/minisatellites. *Mol. Psychiatry* **3**, 21-31

10. Nakamura, Y. *et al.* (1998) VNTR (variable number of tandem repeat) sequences as transcriptional, translational, or functional regulators. *J. Hum. Genet.* **43**, 149-152

11. Kashi, Y. and Soller, M. (1999) Functional roles of microsatellites and minisatellites. In *Microsatellites Evolution and Applications* (Goldstein, D. B., & Schlötterer, C., eds), pp. 10-23, Oxford University Press

12. King, D. G., and Soller, M. (1999) Variation and fidelity: The evolution of simple sequence repeats as functional elements in adjustable genes. In *Evolutionary Theory and Processes: Modern Perspectives* (Wasser, S. P., ed), pp. 65-82, Kluwer Academic Publishers

13. Young, E. T. *et al*. (2000) Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. *Genetics* **154**, 1053-1068

14. Bayliss, C. D. *et al*. (2001) The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. *J. Clin. Invest.* **107**, 657–662

15. Karlin, S. (2002) Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl. Acad. Sci. USA* **99**, 333-338

16. Rockman, M. V., and Wray, G. A. (2002) Abundant raw material for *cis*-regulatory evolution in humans. *Mol. Biol. Evol.* **19**, 1991-2004

17. Fondon III, J. W., and Garner, H. R. (2004) Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. USA* **101,** 18058-18063

18. Hammock, E. A. D., & Young, L. J. (2004) Functional microsatellite polymorphism associated with divergent social structure in vole species. *Mol. Biol. Evol.* **21**, 1057-1063

19. Li, Y.-C. *et al*. (2004) Microsatellites within genes: Structure, function, and evolution. *Mol. Biol. Evol.* **21**, 991-1007

20. King, D. G. *et al*. (2006) Tuning knobs in the genome: Evolution of simple sequence repeats by indirect selection. In *The Implicit Genome* (Caporale, L.H., ed), pp. 77-90, Oxford University Press

21. Sawyer, L. A. *et al.* (1997) Natural variation in a *Drosophila* clock gene and temperature compensation. *Science* **278**, 2117-2120

22. Zamorzaeva, I., *et al*. (2005) Sequence polymorphism of candidate behavioural genes in *Drosophila melanogaster* flies from 'Evolution Canyon'. *Mol. Ecol.* **14**, 3235–3245

23. Nevo, E., *et al*. (2005) Genomic microsatellite adaptive divergence of wild barley by microclimatic stress in 'Evolution Canyon', Israel. *Biol. J. Linn. Soc*. **84**, 205-224

24. Fahima, T. *et al.* (2002) Microsatellite polymorphism in natural populations of wild emmer wheat, *Triticum dicoccoides*, in Israel. *Theor. Appl. Genet.* **104**, 17-29

25. Hammock, E. A. D., and Young, L. J. (2005) Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* **308**, 1630-1634

26. Lim, M. M. *et al*. (2004) Enhanced partner preference in a promiscuous species by manipulating the expression of a single gene. *Nature* **429**, 754-757

27. Young, L. J.. *et al*. (1999) Increased affiliative response to vasopressin in mice expressing the V_{1a} receptor from a monogamous vole. *Nature* **400**, 766-768

28. Mundlos, S. *et al*. (1997) Mutations involving the transcription factor *CBFA1* cause cleidocranial dysplasia. *Cell* **89**, 773-779

29. Tóth, G. *et al*. (2000) Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* **10**, 967-981

30. Katti, M. V. *et al*. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* **18**, 1161-1167

31. Cohen, H. *et al*. (2004) Mono-nucleotide repeats (MNRs): A neglected polymorphism for generating high density genetic maps *in silico*. *Hum. Genet.* **115**, 213-220

32. Verstrepen, K. J. *et al*. (2005) Intragenic tandem repeats generate functional variability. *Nature Genet.* **37**, 986-990

33. Brown, L. Y., and Brown, S. A. (2004) Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends Genet.* **20**, 51-58

34. Yu, F., *et al*. (2005) Positive selection of a pre-expansion CAG repeat of the human *SCA2* gene. *PLoS Genetics* 1(3):e41

35. Rockman, M. V. *et al*. (2004) Positive selection on *MMP3* regulation has shaped heart disease risk. *Curr. Biol.* **14**, 1531-1539

36. Gentles, A. J., and Karlin, S. (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res.* **11**, 540-546

37. Morgante, M. *et al*. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **30**, 194-200

38. Eichinger, L. *et al*. (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **435**, 43-57

39. Lavoie, H. *et al*. (2003) Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains. *Hum. Mol. Genet.* **12**, 2967-2979

40. Katti, M. V. *et al*. (2000) Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications. *Protein Sci.* **9**, 1203-1209

41. Tompa, P. (2003) Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* **25**, 847-855

42. Laken, S. J. *et al*. (1997) Familial colorectal cancer in Ashkenazim due to a hypermutable tract in *APC*. *Nat. Genet.* **17**, 79-83

43. Meyer, T.F. (1989) Molecular basis of surface antigen variation in *Neisseria*. *Trends in Genetics* **3**, 319-324

44. Chang, D. K. *et al*. (2001) Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. *Genome Res.* **11**, 1145-1146

45. Kijas, J. M. H. *et al*. (2001) A frameshift mutation in *MC1R* and a high frequency of somatic reversions cause black spotting in pigs. *Genetics* **158**, 779-785

46. Armour, J. A. L. *et al*. (1999) Minisatellites and mutation processes in tandemly repetitive DNA. In *Microsatellites Evolution and Applications* (Goldstein, D. B., and Schlötterer, C., eds.), pp. 24-33, Oxford University Press

47. Chambers, G. K. and MacAvoy, E. S. (2000) Microsatellites: consensus and controversy. *Comp. Biochem. Physiol. B* **126**, 455-476

48. Vergnaud, G., and Denoeud, F. (2000) Minisatellites: Mutability and genome architecture. *Genome Res.* **10**, 899-907

49. Ellegren, H. (2004) Microsatellites: Simple sequences with complex evolution. *Nature Rev. Genet*. **5**, 435-445

50. Caporale, L. H. (2003) Natural selection and the emergence of a mutation phenotype: An update of the evolutionary synthesis considering mechanisms that affect genomic variation. *Ann. Rev. Microbiol*. **57**:465-485

51. Caporale, L. H. (2000) Mutation is modulated: implications for evolution. *Bioessays* **22**, 388-395

52. Doyle, J. *et al*. (2006) An engineering perspective: The implicit protocols. In *The Implicit Genome* (Caporale, L.H., ed), pp. 294-298, Oxford University Press

53. Levins, R. (1968) *Evolution in Changing Environments*, Princeton University Press

54. Gebhardt, F. *et al*. (1999) Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J. Biol. Chem.* **274**, 13176-13180

55. Albanèse, V. *et al*. (2001) Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite. *Hum. Mol. Genet.* **10**, 1785-1792

56. Kühn, C., *et al*. (2004) Evidence for multiple alleles at the *DGAT1* locus better explains a quantitative trait locus with major effect on milk fat content in cattle. *Genetics* **167**, 1873-1881

57. Tian, B. *et al*. (2000) Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA* **6**, 79-87

58. Rothenburg, S. *et al*. (2001) DNA methylation and Z-DNA formation as mediators of quantitative differences in the expression of alleles. *Immunol. Rev.* **184**, 286-298

59. Suter, B. *et al*. (2000) Poly(dA·dT) sequences exist as rigid DNA structures in nucleosome-free yeast promotors *in vivo*. *Nucleic Acids Res*. 28, 4083-4089.

60. Caserta, M. *et al*. (2002) Aspects of nucleosomal positional flexibility and fluidity. *ChemBioChem* **3**, 1172-1182.

61. Sniegowski, P. D. *et al*. (2000) The evolution of mutation rates: separating causes from consequences. *Bioessays* **22**, 1057-1066

62. Nevo, E. *et al*. (2001) Evolution of genome–phenome diversity under environmental stress. *Proc. Natl. Acad. Sci. USA* **98**, 6233-6240

63. Jackson, A. L. (1998) Induction of microsatellite instability by oxidative DNA damage. *Proc. Natl. Acad. Sci. USA* **95,** 12468-12473

64. Schmidt, A. L., and Mitter, V. (2004) Microsatellite mutation directed by an external stimulus. *Mutat. Res.* **568**, 233-243

GLOSSARY BOX

Coding sequences -- DNA sequences which are translated into proteins. In conventional usage, all other sequences are "non-coding".

Gene -- A tract of DNA consisting of coding sequences (exons) and associated non-coding introns and upstream and downstream regulatory regions, all concerned with biosynthesis of a specific protein (or a family of related proteins generated by alternative splicing).

Imperfect repeats -- see "purity of repetition".

Indirect selection -- the effective preservation or elimination of genomic features that do not directly affect phenotype, through causal linkage to associated phenotypic traits; also called "second order selection" [53]. The mutability of an SSR locus is not "visible" to direct selection acting on phenotype, but mutability is nevertheless a characteristic property of the locus. So direct selection acting upon a particular SSR allele, on the basis of its associated phenotype, necessarily but indirectly acts likewise upon the mutability of that allele [12, 20, 50]. If a population contains alleles that differ in mutability, then selection will favor those alleles, whether more or less mutable, that are most consistently associated with the more fit phenotypes. Whenever alleles conferring a favorable phenotype arise as a result of those alleles' high mutability, then that high mutability will itself be selected indirectly.

Microsatellite -- an SSR with a very short motif, generally from one to six bp. Definitions vary; some exclude mononucleotides and/or put the upper limit as low as five bp or as high as ten [47, 49].

Minisatellite -- an SSR with a longer motif, up to several dozen bp in length. The lower limit has been defined at various values from six to ten bp [47, 49]. For most examples in the literature minisatellite motif-length is twelve or more. The upper limit for minisatellite motif-length is not precisely defined. Functional effects of minisatellite SSRs have been investigated much less extensively than have those of microsatellites. Although less abundant, minisatellites share the same fundamental characteristics of frequent repeat number mutations and of repeat number influencing gene function [46, 48].

Motif -- a particular sequence of DNA basepairs. The number of possible motif sequences increases with motif length. Thus there are two distinct SSR mononucleotide motifs (A/T and C/G), six distinct dinucleotide motifs (AA/TT, AC/TG, AG/TC, AT/TA, CC/GG, CG/GC), ten distinct trinucleotide motifs, etc. (Note that SSR motifs are treated as equivalent if they can be matched by choosing either strand or by starting with any basepair in the sequence.)

Noncoding sequences -- see "coding sequences".

Perfect repeats -- see "purity of repetition".

Polymorphism -- two or more alleles at a locus, each occurring at appreciable frequencies within a population.

Premutation -- a lengthy repeat allele that is prone to extreme expansion, leading to pathological mutation as seen in the "triplet repeat diseases".

Purity of repetition -- the degree to which all motifs within an SSR are identical. In a "pure" or "perfect" repeat, none of the motif copies displays any variation. In contrast, an "imperfect" repeat has some substitutions in the sequence of one or more of the repeating motifs. Imperfect repeats are more stable (less prone to slippage mutations) than pure repeats.

Simple Sequence Repeat (SSR) -- a DNA tract consisting of a relatively short base-pair motif that is repeated over and over in tandem.

Triplet repeat diseases -- A class of hereditary disorders (including Fragile-X, Huntington's disease, spinocerebellar ataxia, and cleidocranial dysplasia) originally characterized by "genetic anticipation", a peculiar pattern of inheritance in which symptoms become more severe and appear at an earlier age as the disease is passed from one generation to the next. The cause is now understood to be extreme pathological expansion of DNA triplets that encode homopolymeric amino acid stretches, commonly glutamine or alanine.

Box 1 Characteristic properties of simple sequence repeats (SSRs)

- **SSRs experience an extremely high rate of reversible, length-altering mutations.** Motif repetition makes SSRs prone to mutation by replication slippage, unequal crossing over, or related processes [46-49]. The resulting mutations, which typically add or subtract one or a few copies of the repeating motif, can be readily reversed by a subsequent mutation at the same or any other point in the repetitive sequence.
- **The mutability of SSRs is a site-specific, adjustable characteristic**. Mutation size can vary from single base-pairs (sometimes inappropriately listed as indels) at mononucleotide repeats up to multiples of much longer motifs in minisatellite repeats. SSR mutation rate is affected by motif length, motif sequence, number of repeats, and purity of repetition [46-49]. Point mutations can degrade repeat purity and stabilize an SSR; whereas active mutational slippage tends to eliminate imperfect repeats. Therefore, SSRs represent sites where selection can indirectly shape the site-specific mutation rates at which new alleles arise.
- **Most SSRs are polymorphic, with extensive allelic variation in repeat number.** In the human genome for example, the proportion of AC repeats that are polymorphic is estimated to exceed 90 percent [16]. SSR polymorphism is familiar as the basis for DNA fingerprinting, lineage analysis, and gene mapping.
- **Normal variation in repeat number can be functionally significant.** The number of repeats at SSR loci can influence on several aspects of genetic function (see main text), although small allelic differences in repeat number commonly exert small quantitative phenotypic effects (many alleles may indeed be effectively neutral).
- **SSRs are ubiquitous.** SSRs are found in genomes of all species examined. They are abundant in both coding and noncoding domains. They occur within many open reading frames, but they are even more frequent in non-coding regulatory regions [16]. Many genes are associated with more than one SSR; those containing at least one coding SSR often contain two or more [15].
- **SSRs are diverse.** SSRs are based on many different motifs and occur in various functional domains.
- **SSR distribution is non-random.** The frequency distribution of SSRs with different motifs varies by functional domain, with triplet motifs much more common within coding regions [29, 30, 37, 49]. Different species have different motif frequency distributions; for example the most common dinucleotide repeats in human, *Caenorhabditis elegans*, and *Arabidopsis thaliana* genomes are, respectively, AC_n, AG_n, and AT_n [29, 30].

Box 2 Some examples of non-coding effects of SSRs.

- **Transcription factor binding.** The first intron of the gene for human epidermal growth factor includes an AC repeat that influences transcription activity both *in vivo* and *in vitro* [54], while a polymorphic TCAT repeat in the first intron of the human tyrosine hydroxylase gene binds a zinc finger transcription factor (*ZNF191*) [55]. In both cases, effects are quantitatively correlated with the number of repeats. Milk fat production in Holstein dairy cattle (*Bos taurus*) correlates with the number of 18 bp repeats, each containing a potential transcription factor binding site, in the promoter for an enzyme regulating triglyceride synthesis [56].
- **RNA shape.** Hairpin folds of RNA transcribed from trinucleotide CTG repeats in the 3′ UTR of the myotonic dystrophy protein kinase gene bind to and activate the dsRNA-activated protein kinase [57].
- **DNA structure and packaging.** AC_n or AT_n repeats can form Z-DNA [1, 58], while repeats of several types can influence nucleosome formation [59, 60].
- **DNA length and orientation.** In any regulatory region, SSR mutations that change repeat number will necessarily change the length of the DNA in that region, thereby rotating the flanking sequences and altering the local spatial relationships of transcription factor interactions.

Box 3 Correcting some Misconceptions about SSRs.

- **SSRs are not just genetic "junk".** The repetitiveness and mutability which once suggested that SSRs could not be serving any critical function are the very features that make SSRs useful. The genetic "meaning" of a specific SSR allele, whether as a coding sequence or in cis relation to a coding sequence, resides not only in its motif sequence and repeat number, which together represent a particular quantitative effect, but also in repetitiveness itself [2]. Repetition, by conferring mutability, represents an SSR's ability to vary reversibly in subsequent generations.
- **SSR alleles are not always adaptively neutral.** SSR alleles are commonly analyzed under the presumption that allele frequencies are determined solely by mutational processes and genetic drift. Although this may often be an appropriate null hypothesis, the possibility of adaptively relevant function should be explicitly recognized and tested. In natural populations, the most frequent SSR alleles have already been winnowed by selection and are thus expected to fall within a range where fitness differences may not be noticeable. Nevertheless, adaptively significant effects may readily emerge as ongoing mutation yields variants whose length falls outside this currently-favored range.
- **SSR sites with functional effects are not just rare exceptions.** The relevant literature is dispersed across many disciplines, with many studies focussed not on SSRs *per se* but on the functions of particular genes or the genomic bases for particular phenotypes. Repeat number variants of mononucleotide repeats are often reported as SNPs (i.e., single bp indels) rather than SSR alleles.
- **Functional effects of SSR mutability are not always harmful.** A commonplace prejudice that mutation must, on average, be predominantly deleterious appears to be reinforced by the association of certain SSRs with human disease. But these are exceptions. Disease associations receive disproportionate attention but they clearly represent pathological aberrations of normal SSR function. SSRs variation within a normal (i.e., non-pathological) range of repeat number commonly yields small, quantitative functional effects.
- **Evolutionary theory does not prohibit selection favoring mutability.** The classic argument that natural selection necessarily minimizes mutation rates is based on assumptions that do not apply to SSRs [12, 20, 50]. Indirect selection for mutability is unlikely to occur unless special circumstances obtain [61], but appropriate special circumstances are exactly what SSRs provide. Widespread prevalence and evolutionary conservation of mutable SSR sites imply that at least some SSRs have been retained because their mutability yields advantageous variation [12, 20, 50].

Questions Outstanding

- **In association tests of candidate genes, when specific SSR alleles consistently correspond with particular trait values, could the trait differences be caused by the SSRs themselves?** Positive evidence that SSR alleles are responsible should include experimental testing of alternative SSR alleles, preferably more than two, against a controlled genetic background (e.g., by genetic knock-in). Alternatively, extensive sequencing is needed to demonstrate the absence of any other associated polymorphism.
- **What is the quantitative relationship between phenotypic variation and the number of repeats in a corresponding SSR?** This question can only be addressed by measuring the incremental effects of repeat-number alleles representing three or more different lengths.
- **To what extent do SSRs contribute to adaptive divergence among populations?** Innumerable studies, not reviewed here, have reported differentiation of SSR allele frequencies among natural populations and species. Although such alleles are usually presumed to be neutral, the possibility of small but appreciable fitness differences needs to be explicitly tested [62].
- **To what extent is SSR function regulated by other aspects of the genome?** Evidence that other genetic elements have adapted to accommodate and regulate the mutability of SSRs would strongly support a positive evolutionary role for SSRs themselves. Such evidence is already available for bacteria; moreover, the regulating mismatch repair elements themselves contain SSRs that allow their own adjustment [44, 50].
- **Is the mutability of particular SSRs adjusted by indirect selection?** Selective retention of a favorable SSR allele necessarily preserves the repeat-based mutability by which it arose. But when allele stability is beneficial, single base pair substitutions can stabilize the SSR by reducing the purity of repetition. For example, the repeat sequence in the longer and more frequent allele of a human tyrosine hydroxylase gene is interrupted by single nucleotide deletion, which presumably discourages further expansion [55].
- **Can mutability of SSRs be induced by stress conditions?** A stress-inducible increase in mutation rate, specifically directed to SSR loci, could "adjust" the fitness of individual cells. Oxidative stress can destabilize microsatellites in prokaryotes [63]. One preliminary report suggests that targetted SSR mutations may be elicited by fungal infection in plants [64].
- **Does SSR mutation play a role during the life span of individual organisms?** The intriguing example of somatic SSR mutation causing pigs' pattern of black spots [see main text] suggests that the mutability of SSRs may play a role generating cellular diversity during normal development.