Is spatial occurrence of microsatellites in the genome a determinant of their function and dynamics contributing to genome evolution?

Atul Grover[1,2] and P. C. Sharma[1,*]

[1]University School of Biotechnology, Guru Gobind Singh Indraprastha University, Sector 16C, Dwarka, Delhi 110 075, India
[2]Present address: Molecular Biology and Genetic Engineering Laboratory, Defence Institute of Bio-Energy Research, Goraparao, Haldwani 263 139, India

The non-random distribution of microsatellites in the genome has been implicated in a number of cellular and evolutionary activities. Recently, microsatellites have gained much attention due to their suggested association with cancers, ageing and various other metabolic disorders. Microsatellites are thought to have evolved mainly through polymerase slippage with variable mutation rates. It is difficult to develop a typical molecular evolutionary model that may describe genomic dynamics of these sequence elements. Microsatellites may also be the accidental sites of action by selection forces in events like genome divergence and speciation. Substantial evidence is available to describe various life stages of microsatellite evolution. This review addresses the current state of knowledge suggesting interrelationships between genomic locations, functions and evolutionary dynamics of microsatellites, and their subsequent implications in genome evolution.

Keywords: Genome evolution, life-cycle concept, microsatellites, microsatellite instabilities, mutation models.

AVAILABILITY of whole genome sequences and a wealth of published literature reporting the analysis of genomic sequences facilitate studies that aim at understanding various aspects of genome organization and evolution in different forms of life. The present day eukaryotic genomes pack a bulk of non-coding DNA embedded with protein coding regions. A part of this non-coding DNA plays a regulatory role, whereas the other part simply provides structural stability to the chromosomes.

Repetitiveness of nucleotide sequences is an important feature of all genomes, however, the extent to which it occurs within genomes varies greatly. The only consensus reached so far is that the amount of repetitiveness exceeds the expected values of repetitiveness[1]. Repetitive sequences are now known to play important roles in a cell, define genome structure and drive the adaptive evolution of an organism[2–4]. These sequences are broadly classified into interspersed repeats and tandem repeats, and may constitute a significant proportion of some genomes[5–7]. Tandem repeats are broadly classified into satellites, minisatellites and microsatellites, and mainly distinguished on the basis of the length of the repeating unit. Here, we critically overview the genesis and propagation of microsatellites and how the evolutionary events involving microsatellites affect the genomic transitions leading to major changes including genetic drift and speciation over long periods.

Microsatellites are remarkably constituted of small repeating units, 1–6 bp in length. Such a unit formation is structurally simple and therefore these repeats are also called as simple sequence repeats. These sequences constitute hypervariable regions of the genome and undergo structural changes through addition or removal of repeat units or through point mutations therein[8,9]. The latter event can cause imperfections in these arrays, thus leading to the formation of perfect and imperfect microsatellites[10]. The idea whether microsatellites are evolutionary junks, or useful sequences that are repeated throughout the genome has been a topic of debate in the scientific community. Evidences are being gathered in favour of the hypothesis that the simplicity of these sequences in itself is a useful attribute of the genome[3] and also that they are strategically placed in the genomes. However, a recent study by Buschiazzo and Gemmell[11] indicates that for most of the microsatellites, survival in mammalian genomes is only by chance and there are no evolutionary designs behind their conservation over long evolutionary periods.

Genomic location, functions and hypervariability

Considering that not all of the microsatellite motifs have uniform distribution in the genomes[12–16], they are likely to be involved in different genomic activities with defined biological roles. Whether such roles are dictated by genomic location and motif characteristics of the microsatellites or by the specific genomic requirements[4] is not precisely clear. Based on their genomic locations, we have grouped microsatellites as gene-associated, mobile element-associated, telomere-associated, centromere-associated, and microsatellites present elsewhere in the

genome. Evidences suggesting a relationship between genomic position and functionality of the microsatellites have been presented here.

Gene-associated microsatellite sequences

A microsatellite may be present within the coding sequence of the gene, in the upstream region including 5′-UTR, in intron or in the 3′-UTR. It may be overlapping a gene, or present in close proximity. Thus, such a microsatellite may be directly affecting the corresponding protein structure or regulate the expression of the protein-coding gene. Their participation in the central dogma is at least known with two well-studied examples – contingency loci in prokaryotes[17] and disease-causing microsatellite mutations in human beings[18].

It is now well established that tandem repeats play an important role in generating structural variability in surface antigens in pathogenic bacteria and fungi[17–19], which help them to evade the host immune system, and thereby causing virulence. They are also involved in other functions in prokaryotes like participating in genetic recombination, protein–protein interactions, generating contingency loci, etc.[20]. Martin et al.[21] presented an experimental evidence of the participation of microsatellites in transcription modulation in the pathogen *Neisseria meningitides*, where upstream TAAA repeats to *nadA* interfered with the binding of transcriptional regulatory protein IHF.

A number of health disorders in humans have been related to microsatellite instability occurring anywhere in the genic regions[22]. Etiologically, around 35 human disorders are known when a simple sequence formation within a gene or in its vicinity becomes intolerable, if it roughly reaches 150 bp size[18]. Noticeably, this approximates the length of the DNA wrapped around a nucleosome core particle. Broadly, these diseases fall into two classes – class I characterized by the synthesis of toxic proteins upon disproportional expansion of microsatellite sequences; class II with cancers and other health disorders like diabetes and arthritis occurring due to general instability of the microsatellite sequence in the concerned gene sequences[22].

Among all the microsatellite repeats capable of causing human disorders on expansion within the human genes, repeats of type CCG, CTG/CAG and GAA are most common[18]. Further, their elongation is not tolerated anywhere along the gene sequence from 5′-UTR to 3′-UTR. Usdin[23] has extensively reviewed various disorders due to such expansions and the possible molecular mechanisms involved like replication, transcription, translation, protein traps, etc. Non-trinucleotide microsatellite repeats are also known to cause human diseases by the length variation mechanisms or simply due to microsatellite instabilities (Table 1).

Expansion events in 5′-regions of the genes lead to malfunctioning of the transcription machinery[4]. Microsatellites have been recognized as both, part of intrinsic promoters and transcription enhancers[4] and increase in the repeat number at upstream CCG repeat site is associated with enhanced *FMR1* transcription[24,25]. Interestingly, the same motif is also associated with upstream sequences of genes in rice displaying low levels of mutability[14]. Vinces et al.[4] reported that 25% of the yeast gene promoters harbour tandem repeats, which enabled the concerned genes to have a higher transcriptional divergence rate, simply by alterations in the repeat length. Understandably, a change in the length of the repeat can alter the organization of the genes relative to the nucleosomes. For several human metabolic disorders for which no point mutations are known screening for microsatellite expansions within the gene or in upstream regions has been suggested for molecular diagnosis[22].

A first report of repeat-expansion based disorder in plants, implicates GAA repeat in the intron of isopropyl malate isomerase large subunit 1 to be responsible for the stunted growth in *Arabidopsis thaliana*[26]. Microsatellite motifs like CCG, CTG and GAA have abilities to form a variety of stable secondary structures[27]. When present upstream of the transcription start site, these structures may have important regulatory effects. Alternatively, they may be acting as unique protein recognition motifs.

Phenotypic effects of mutations in microsatellites are not limited to metabolic disorders. Many of the altered phenotypic effects including those related to environmental adaptations can be attributed to mutations in microsatellite loci, sometimes even outside the genic regions[3]. An interesting example refers to the occurrence of two microsatellite containing duplicated blocks, 3.5 kb upstream of the arginine vasopressin V1a receptor gene *AVPR1A* in humans that modulates social awareness and behaviour. Each block is nearly 350 bp long. The first block contains a $(GT)_{25}$ repeat, and the second complex hosts $(CT)_4TT(CT)_8(GT)_{24}$. Polymorphisms in the latter loci are known to alter socio-behavioural traits in human beings, including autism spectrum disorders[28]. The upstream microsatellites to *V1aR* gene are known to occur in many animal species also. A hypermutable functional microsatellite may be responsible in creating allelic diversity of proteins rather producing a toxic protein.

Microsatellite expansions even when occurring in 3′-UTR can cause errors in biological processes. For example, they cause transcriptional slippage in *URA3* gene of yeast[29]. Myotonic dystrophy type 1 (DM1) is a human disorder caused due to transcriptional slippage resulting from an expansion of CTG repeats in 3′-UTR of DM protein kinase gene[30]. CTG expansion in mRNA synthesized due to transcriptional slippage causes disruption in the splicing process and creates misfolded proteins[30]. Li et al.[18] focused on detailed phenotypic effects of microsatellite expansion at each of the regions of a gene, and have discussed the underlying molecular mechanisms.

Microsatellites can also affect the enzymes controlling cell cycles, as evident by the fact that microsatellite

Table 1. Metabolic disorders in humans caused due to microsatellite instability of non-trinucleotide tandem repeats

Disorder	Wild type allele	Disease-causing allele	Genomic location
Diabetic retinopathy	$(AC)_{24}$ (Z allele)	Z-2	2.1 kb upstream of TSS of aldose reductase (*ALR2*) gene[89]
Diabetic retinopathy	$(CTTTT)_{24}$	$(CTTTT)_{8-18}$ except $(CTTTT)_{14}$	2.5 kb upstream of TSS of nitric oxide synthase (*NOS2A*) gene[90]
Lung allograft fibrosis	(CA) repeats	Allele 2	First intron of *IFN-γ*[91]
Aplastic anaemia	12 (CA) repeats	Homozygous state	First intron of *IFN-γ*[92]
Cardian artery disease	$(TAAAA)_{6-8}$	Longer repeats	Promoter of sex hormone binding globulin (*SHBG*) gene[93]
Polycistronic ovary syndrome	$(TAAAA)_{6-8}$	$(TAAAA)_{9-11}$	Promoter of sex hormone binding globulin (*SHBG*) gene[67]
Spinocerebellar ataxia type 10 (SCA10)	$(ATTCT)_{10-29}$	$(ATTCT)_n$, where $n \leq 4500$	Intron 9 of *ATXN10* gene[68]
Spinocerebellar ataxia type 31 (SCA31)	Null allele	$(TGGAA)_n$	Introns of thymidine kinase (*TK2*) gene and brain-expressed, associated with Nedd4 (*BEAN*) gene[94]
Myotonic dystrophy type 2	$(CCTG)_{36}$	$(CCTG)_{75-11000}$	*ZNF9*[95]
Myoclonus epilepsy 1 (EPM1)	$(CCCCGCCCCGCG)_{2,3}$	$(CCCCGCCCCGCG)_{12-17}$	Cystatin-B (*CSTB*) gene[96]
Diabetes type I	$(ATAGGGTGTGGGG)_{6-10}$	Null allele	Human insulin (*INS*) gene[97]

instabilities may cause cancers[31]. Occurrence of poly A repeats in the mis-match repair (MMR) genes is a common feature in eukaryotes[32]. The mononucleotide repeat present within the coding sequence of these genes is highly susceptible to single nucleotide mutations rendering these genes hypermutable and subsequently generating a modified protein[33]. In eukaryotes, genes with products participating in the central dogma host more number of microsatellite repeats than expected mathematically[33–35].

Microsatellites present at centromeres and telomeres

At the centromeres, microsatellite frequency takes a dip. However, human centromeres harbour a microsatellite repeat sequence $(AATGG)_n$ as in some other species. A general richness of repeats at these regions in different species is considered significant from an evolutionary point of view, in terms of their sequence–structure–function relationships[35]. In general, low complexity sequences at the centromeres facilitate chromatid cohesion and kinetochore formation[36]. The sequence of both the telomeric and centromeric microsatellite repeats may vary a great deal within a close taxonomic group, though in plants, the prevailing telomeric structure is determined by TTTAGGG[37]. Telomeric repeats provide a template for the formation of ribonucleoprotein complex, a chromosome capping structure and its length regulates the integrity of the chromosome. Mutation-driven allele lengths are not the only changes that occur at telomeres, but a sequence motif variation is also possible, as suggested by Sykorova *et al.*[37].

Microsatellites occurring elsewhere in the genome

Microsatellites may be present anywhere in the genome and can make secondary structures stable enough to cause a polymerase slippage at these points. If two microsatellites with same motif are present in the vicinity of each other but in the opposite orientation, so as to make inverted repeats, they form a structure that can stall the DNA replication process[38]. Some of the dinucleotide motif repeats can also form Z-DNA, that affects recombinational events during meiosis[39]. The participation of microsatellites in meiotic recombination has long been recognized and their frequency in recombination hotspots in the genome is over two-fold to that in other regions[40]. However, this is just another cause–effect paradox. Whether this frequency is an outcome of this mode of their genesis, or their presence at these regions facilitate recombination, needs to be studied. Nevertheless, the length of the microsatellite for such purposes plays a crucial role. Too small or too long, both kinds of microsatellites may be undesirable, and can cause several errors[39]. Thus, not only motif sequence influences the genomic activity, but microsatellite length can also influence the genomic activities including determining chromosome structure and participating in crossing over[40].

Effect of computational tool used in microsatellite mining

Insight into the genomic locations of microsatellites and associated features has been possible especially after the availability of a variety of computational tools facilitating scanning of the whole genome sequences[5]. Microsatellite

mining is a challenging field of computational biology research as the characteristics of the underlying algorithm can affect the efficiency of various models and hypothesis derived from the resulting datasets. There have been many attempts directed towards designing the microsatellite mining tools following different approaches. However, none of these approaches return a complete set of microsatellites present in a given genome[12]. Majority of the microsatellite scanning tools return a near complete dataset of microsatellites, based on a particular definition, used in the algorithm. As a result, the microsatellite dataset identified by a given tool is often different from that obtained by using the other tool. When the two datasets are compared, most of the microsatellites are generally found common to both the datasets, but only a few, especially the imperfect ones are restricted to any one of the given datasets (our unpublished results). Such numerical variations may slightly alter the overall microsatellite picture of a genome, however, studies designed and conclusions drawn with respect to genomic location of specific microsatellite loci remain unaffected by such inconsistencies with regard to the number of different microsatellites. We are of the opinion that any individual tool is capable of providing meaningful information and can be used for most of the genome-wide studies involving microsatellites.

Microsatellite mutation models

Polymerase slippage

Many opinions and models have been put forward to explain the microsatellite hypermutability and instability in the genomes. Debate continues on how microsatellites come into existence; why these loci remain hypermutable; and what eventually happens after a few generations. Nevertheless, it is generally agreed that microsatellites are the outcome of replicative errors which failed to get fixed by the mismatch repair system of the cell. Evidently, the position of the microsatellites relative to the origin of replication has a direct impact on microsatellite mutability[41]. Mutations in mismatch repair system genes also tend to increase the microsatellite instability[9,31].

DNA strand bias exists for polymerase slippage within complementary microsatellite DNA sequences with error frequencies higher at AG and AAGG alleles than at CT and CCTT alleles[42]. This is possible if the microsatellite on the single-stranded DNA during DNA replication forms a secondary structure thereby causing an error during replication[9,42]. Another probable explanation offered for polymerase slippage is the misalignment of permutational intermediates during DNA replication[42]. This phenomenon also explains the fact that polymerase slippage occurs only at microsatellite loci, and not at purine/pyrimidine-rich stretches elsewhere[42]. Such an observation probably makes a strong point to explain the occurrence of CT and CTT repeats near 5′-end of the genes in plant genomes in the direction of transcription, while AG and AGG fall on the non-transcribed strand of the genes[43]. *In vivo* experiments involving *MutS* mutants of *Escherichia coli* displayed increased instability of AC and TG repeats, with instability of TG repeats 1.6 times to that of AC repeats[44]. Orientation dependence was also observed in constructs introduced in yeast and *E. coli* for CTG repeats, but not in the case of GT repeats[41]. Thus, polymerase slippage is dependent both on motif length and sequence making some of the motifs more conserved than others[15]. Moreover, DNA polymerases can also introduce single nucleotide errors in specific microsatellite motifs like GT, CT, etc.[8]. The types of interruptions introduced within a microsatellite sequence include single-base deletions and substitutions as well as complex deletions and substitutions. Morel *et al.*[44] on the basis of their *in vivo* studies involving $(AC)_{51}$ reported 70% and 30% deletions and insertions, respectively. However, the length of microsatellites is also supposed to be often regulated by unequal crossing over and other recombinational events[45]. A recent review on mismatch repair assays by Spampinato *et al.*[46] accumulates a number of examples of enhanced mutation rates in different organisms.

Mutation rates

Due to the combined effect of polymerase slippage and unequal crossing over, the microsatellite mutation rates may be up to 10^6 times higher than those estimated for other regions elsewhere in the genome[8]. The mutation rates of microsatellites may vary at intra- and inter-specific levels depending upon the age, sex and even genomic location and structure of the microsatellite repeat[9,45,47–49], as depicted in Figure 1. As an example, loss of a single C from an iteration, the mononucleotide was reported thrice more frequently than loss of A in a similar length of the A-repeat, and also faster than rate of loss of CA from a CA-repeat in *E. coli*[50]. Further, placing a $(GT)_{16}G$ repeat at different locations in the genome affected the mutation rates in *MSH2* deficient and proficient strains of yeast[51].

A relevant question probes whether any relationship exists between microsatellite abundance in the genome and mutation rate at which they are evolving. A tangible hypothesis suggests that high mutation rates should correspond to microsatellite richness in a genome. However, even microsatellite rich genomes do not always show high mutation rates[13,14]. Therefore, the microsatellite richness in a genome should be viewed as the absence of purifying selection pressures to get rid of microsatellites, and not as the abundance of mutational events in the genome. Whether the richness of microsatellites in and

around housekeeping genes[32] is due to the absence of mutational events at these sites owing to vitality of the function of these genes, or because of functional vitality of microsatellites themselves is not clear at present. The fact that there are certain regions in the genomes which are less susceptible to mutations[52] does not overrule the possibility that microsatellites accumulate in these regions due to absence of mutational events. In such a case, hypermutability should reflect a characteristic of the genomic location and not that of the microsatellite alone.

Evolutionary models

Majority of the mutations operating at microsatellite loci are simply minor expansions and contractions per generation, probably caused by repeated replication slippage at these sites[45]. The stepwise mutation model[53] proposed for microsatellite variation explains this change by a single repeat number per slippage event. However, this model is too simple to explain all kinds of changes occurring at microsatellite sites and does not alone explain all of the mutations at a microsatellite site. Therefore, other models such as infinite allele model (IAM)[54] that assume that each mutation creates a new allele in the population and two-phase models (TPM)[55] have been invoked to explain various mutations occurring at microsatellite loci. According to IAM, the frequent forward and backward mutations create identical alleles[54]. The benefit of adopting this model is that it describes the mutational dynamics at complex microsatellites better than other models. TPM assumes that single nucleotide mutations are more likely to occur and multi-step mutations follow a truncated geometric distribution[55].
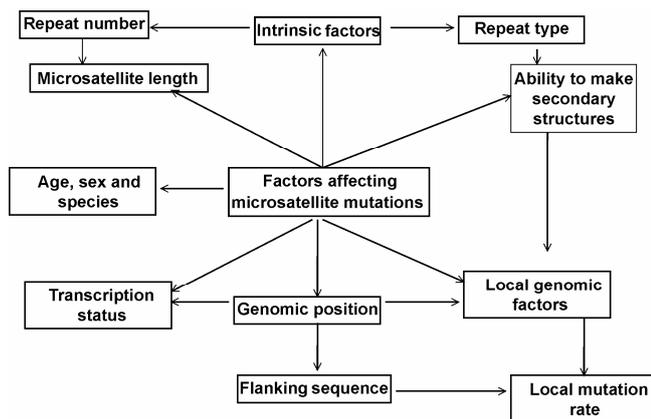

Figure 1. Various factors influencing the mutations and mutation rates of microsatellites. Intrinsic factors include repeat type and repeat number. Some motifs are more mutable because of the inherent self-complementarity that causes DNA polymerase to slip more often at these sites. Similarly, a longer repeat is likely to provide more space to DNA polymerase to undergo slippage. However, microsatellites falling in certain genomic regions like in a gene or on a hitchhiked region are likely to show lower mutation rates compared to microsatellites falling distant to these regions.

One of the reasons why any of these evolutionary models does not completely explain microsatellite dynamics lies in the fact that at micro-level both mutational patterns and mutation rates are biased at microsatellite sites, whereas the evolutionary models are designed assuming normal distribution of the alleles. Microsatellites have a tendency to expand[56] and in the absence of any selection pressure, microsatellites can attain infinite growth, and achieve significant lengths. At least in exonic regions, a strong purifying selection force is consistently under operation to keep a check on the array length of microsatellites[33]. In non-coding regions of the genome, no such selection pressure might be operating and yet infinite lengths of microsatellites are not observed[5]. Evidences indicate that mutational bias towards the longer repeats is responsible for restricting the growth of the microsatellites[57]. Whether any 'equilibrium length' of microsatellite alleles exists wherein repeat elongation and contraction are equally participating is an important question to be resolved yet. Increasing evidences indicate that microsatellite lengths are regulated by an equilibrium between polymerase slippage and point mutations (PS/PM model)[58,59]. A single nucleotide mutation within a pure formation of microsatellite renders it imperfect (microsatellite splitting). The observation that an imperfect repeat displays lower mutation rates compared to a long stretch of a perfect repeat contributed in building another model for mutational dynamics in microsatellites called 'proportional-rate linear-biased one-phase model'[60]. The model correlates mutation rates with the length of the repeat and imperfections harboured[60]. There are certain suggestions that point mutations systematically reduce the length of the microsatellite repeat by preferably occurring at the extremes of an array (microsatellite trimming)[60]. Thus, such models based on biased mutational processes are more realistic but less popular due to complex mathematical procedures involved.

Life cycle model

It is clear from the above discussion that some selection forces are in regular operation to keep a vigil on microsatellite dynamics in the genome. To explain such an evolutionary dynamics, birth–death models have been described in which selection maintains the repeats within a genome, with occasional creation of new repeat and elimination of some existing repeats. Unlike interspersed repeats, microsatellites do not confine to a birth–maturity–death type life cycle. The evidence in favour of 'proto-microsatellites' indicates that they do have 'embryonic life', and at times they do have 're-incarnation' as well (Figure 2). Rarely in the middle of their lives, one kind of microsatellites may also undergo transition into another kind and therefore they do have 'metamorphosis' as well during their lifetime. Nevertheless, the life cycle concept outlines the

mutational biases at microsatellite loci very well and mirrors microsatellite evolution in a realistic way.

The birth of a microsatellite is indicative of dominance of forces that promote the microsatellites over those which suppress their growth. The process is more or less non-random as has been discussed above indicating that polymerase slippage is a biased process, and for *de novo* synthesis of a microsatellites, existence of proto-microsatellites is necessary (Figure 2). A proto-microsatellite is a very short sequence of tandem repetition, thought to arise frequently by random base substitutions and indels in the genome[61]. Proto-microsatellites might also be developed from low complexity sequences present at the ends of transposable elements[62,63]. A proto-microsatellite is born as polymerase slippage occurs in the microsatelite at this site and its mismatch repair system either fails or sets undone by various forces operating at these sites. Natural selection may prevail within the same or subsequent generations to keep a particular microsatellite or to eliminate it. This is evident by the fact that non-trinucleotide microsatellites are hardly tolerated in the exons (discussed here). The threshold length for slippage to occur at proto-microsatellite is considered as four repeat units[64].

The growth of the microsatellite after its birth, is thought to grow following a stepwise mutation model mainly through polymerase slippage. Some loci may be more prone to multi-step mutations. Unequal crossing over can also contribute to increase or decrease in the lengths of the microsatellites, sometimes even leading to the elimination of the repeat from a specific site. Ellegren[62] suggested that the contribution of unequal crossing over towards microsatellite propagation is significantly lesser than the polymerase slippage.

Once the microsatellite has attained a certain length, selective forces start working against it, thereby creating several interruptions in the sequence of a long microsatellite. Ellegren[45] recognized such events to be the initial

steps contributing towards the death of the microsatellites, whereas we describe it as 'ageing'. Accumulation of interruptions breaks the repeat array. Large deletions occur at the sites of interruptions causing acceleration in the process of elimination of a microsatellite[65]. Despite the two-fold action of interruptions and deletions, the death rate of microsatellites is presumably lower than the birth rate considering enormity of microsatellite sequences in eukaryotes and occurrence of a majority of them as interrupted or imperfect microsatellites[16].

It is worthwhile to consider what we should call as 'death' of a microsatellite. In principle, even if small formations of a previously occurring microsatellite motif are occurring at a given genomic location, but the repeat number is lesser than the standard size described under definitions of microsatellites, it must be considered as a dead microsatellite. In such an eventuality, there can be two possibilities – either the 'new size' has become lesser than the initial threshold of a proto-microsatellite (4-repeat unit), or greater than that (Figure 2). In the latter case, it is again available as a potential site for polymerase slippage, and hence for 're-incarnation' of the microsatellite. In the earlier case as well, proto-microsatellite can be recreated through indels and substitutions[61] leading to the formation of a microsatellite once again.

It must be noted that all the stages in the 'life cycle' of a microsatellite are transitional. While a microsatellite can meet its death prematurely as an 'accident', it can also elongate its life term by eliminating interruptions. Further, immortality of some of the microsatellites too is not overruled, as evidenced by conservation of microsatellites over periods spanning as long as 450 million years[11]. There are suggestions that an imperfect repeat can revert to a perfect repeat through single nucleotide deletions in a single or more generations[65]. The fact that DNA polymerases are capable of creating a microsatellite, as well as can start the events[56] leading to eventual death of a microsatellite, strongly favours the concept of a defined life cycle of the microsatellites on an evolutionary time scale. Barrier *et al.*[66] also reported 'hotspots' for microsatellite formation in the intron of the genes *ASAP1* and *ASAP3* in Madiinae (Asteraceae), as 39 microsatellites could be located within these introns. Single nucleotide mutations (substitutions, deletions, insertions) at a single site (or periodically multiple sites) within a microsatellite sequence coupled with polymerase slippage can create a complex repeat from a simple microsatellite repeat. At times, this may lead to replacement of one motif by the other, as has been observed in case of rice[14]. Complex and compound microsatellite repeats are interesting from the point of view of evolution, as they represent areas of high microsatellite frequency. There is a possibility that each individual participating microsatellite in a compound repeat influences the mutability of the other partner(s) also[8].
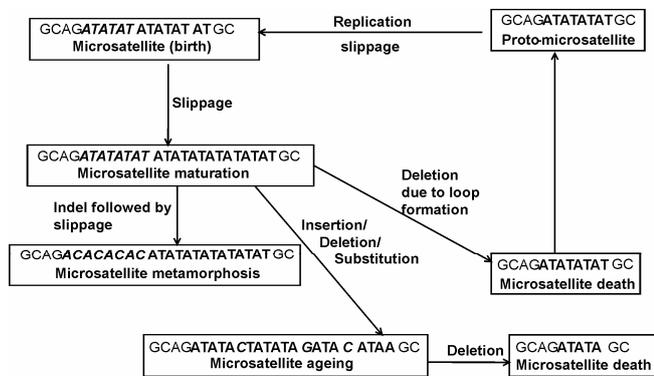

Figure 2. Life cycle of a microsatellite. A metamorphosed microsatellite follows its own independent course. Microsatellite repeat is shown in bold.

Microsatellite evolution in the light of natural selection

The above discussion focuses on the selection forces governing the length of the microsatellite array, but we still need to discuss what is 'selectively neutral' and what role is played by natural selection at hypermutable sites like microsatellites. Further, the role of genetic drift needs to be discussed. Like elsewhere in the genome, most changes at microsatellite sites too are driven by natural selection. Since, most of the microsatellites lie at non-genic sites, most nucleotide substitutions within microsatellites are selectively neutral. But, those occurring within a gene or at its regulatory regions may not be neutral. Such changes either produce an abnormal phenotype as in the case of metabolic errors[67,68], or provide an inherent adaptability to the organism against environmental stress[3]. According to the nearly neutral theory of molecular evolution, genetic drift leads to fixation of deleterious mutations.

Microsatellite variability and genome divergence

Do microsatellites actually represent neutral loci? The present belief is that some of the microsatellites are not neutral[3,4,8]. In such a case, does their variability derive the genome divergence or the evolutionary processes generate their variability? The clues to answer these questions may be derived from the understanding of mechanisms of evolution and inferring various leads gained through 'genome scans' either bioinformatically or using microsatellites as molecular markers. In this section, we review the recent studies on genome-scans, and analyse the results keeping microsatellites in the centrestage considering them to be other than 'merely molecular markers'. Whereas, the intra-specific genome scans provide vital information of adaptive alleles segregating among populations, interspecific studies provide more authentic and easy to interpret picture of long-term effects of microsatellite variability on major evolutionary events like speciation, adaptation, etc.

Copy number variations of microsatellites

A common way to create genetic variation within and among populations is through copy number variation. Many of the polygenic traits and certain metabolic disorders arise through such variations in human beings[69]. Microsatellites also differ in natural populations by copy numbers[70]. Largely, microsatellite evolution either in terms of expansion or contraction is considered neutral. However, considering millions of microsatellite loci in the eukaryotic genomes, such changes should generate random variation in genome size and a normal distribution for microsatellite number should exist in agreement

with the principles of population dynamics[70]. Schlotterer[71] suggested that owing to random sampling in population studies especially between two generations, the most recent common ancestor for alleles of some of the loci is traced just a few generations back, while for others it is traceable several generations back. Similar comparisons at intra-species level in the rice genome[14,16] hinted that the sub-species genome with poor representation of microsatellites on an average hosts longer repeats than the sub-species genome with higher number of repeats, thereby making the overall genome under microsatellite cover almost constant. Further, it was observed that orthologous microsatellites maintaining the same length between two subspecies genomes occurred at a higher frequency comparative to those displaying length variations[16]. There could be two reasons for such an observation – either a selection pressure returned these microsatellites to the same length or these microsatellites themselves maintained length sanctity.

One of the sources for microsatellite copy number variation in eukaryotes might be through the action of the transposable genetic elements, capable of mediating the dispersal of microsatellites in the genomes[72]. Elements like SINEs, LINEs, retro-pseudogenes and others are potential sources for proto-microsatellites due to the occurrence of poly-(A) tract at the 3′-ends[72]. The poly-(A) site is amenable to reverse transcription errors explaining the genesis of proto-microsatellites, which may further expand as replication error. Interestingly, more than 50% of the human Y-chromosome microsatellites are thought to have evolved through retrotransposition[73].

Despite the various evidences and appealing explanations about copy number variation among different genomes, an important unanswered question remains whether microsatellites are the means of creating genetic divergence or they are simply the by-products of ongoing evolutionary events in genomes? Genome evolution or divergence is frequently an outcome of the response to environmental factors. Loci participating in this genomic activity are often identified by changes in the phenotype[74], and thus the impact of loci-like microsatellites often go unnoticed and they are largely considered as neutral. Undoubtedly, microsatellites are evolutionarily important elements and participate in defining the structure of a chromosome[39]. In terms of evolutionary systematics, they play a role probably at the lower levels, and patterns of their distribution are conserved only at the level of a closed taxonomic group[5]. The fact that microsatellites could be used for cross-species genomic comparisons[11,61] and that longer microsatellites are overabundant[61] indicates that they are important for providing stability to the chromatin structure. The non-uniform distribution of different microsatellite types is also an indication of their involvement in various genomic activities.

Evolutionary significance of low microsatellite variability

Microsatellites are likely to complete their life cycles in long evolutionary times, however, as we learnt from our experience of comparing microsatellites in two subspecies of rice[14,16], 0.44 Mya (time of divergence of two subspecies of rice) is not long enough for all the microsatellite loci to complete their life cycle. Microsatellites showing lesser or no polymorphisms are considered to be of recent origin, compared to microsatellites showing higher degree of polymorphism[71]. In rice, most of the microsatellites seem to have conserved their lengths across the subspecies genomes[16]. It can be argued that several bottlenecks existed in rice much earlier to its divergence into *indica* and *japonica* subspecies, and majority of microsatellites in rice have thus played no role in its evolution as a highly successful agricultural crop species. However, one has to be conscious in drawing such conclusions, as a microsatellite associated with important region of the genome (from evolutionary point of view) can significantly display lower variability during genetic drift and selective sweeps[71], leading to allele excess. The prevailing notion suggests that the fixation of a beneficial allele to a given environment will alter patterns of polymorphism at a nearby microsatellite also[75]. The low mutation rates of microsatellites thus may not only be associated with its proximity to genic sequences, but also if the microsatellite has been hitchhiked and alleles have been fixed[71,76]. Possibly, false positives may appear (under stepwise mutation model) if the common ancestor of the sampled individual is traced to only a few generations back[77]. Therefore, occurrence of a particular allele in a population must be dealt with caution. There have also been evidences of accumulation of rare alleles at these sites[71,75]. Hitchhiking generates a genomic region of high polymorphism between species (or populations during early stages of divergence), and the polymorphism level is comparable to total differentiation between the two groups under study and characterized by alternative alleles[78]. In case of microsatellites, this would mean that each of the two groups under study would show different size alleles, and this difference can easily be typed by picking any random individuals from each of the populations. Gradually, hitchhiking leads to (ecological) speciation[79], creating intra-genomic heterogeneity in terms of differential selection of some of the genomic regions. Only when a (sub-) population is at such an early stage of speciation, microsatellites associated with genomic regions involved in reproductive isolation can be studied. Owing to their mutabilities, microsatellites mirror recent selective sweeps, while other nucleotide sequences detect ancient selective sweeps[78].

Islands of divergence model

Recently, Nosil *et al.*[79] classified genomic regions into three categories on the basis of signatures of divergence.

Genomic regions that show high mutation rates are placed at a different stratum separate from the ones that show low mutation rates under this model. The same classification can be extended to microsatellites also by identifying them as falling in the regions of genomic divergence, nearby to these regions and the ones away in the highly conserved regions. Accordingly, a genome becomes a mosaic of microsatellites with heterogeneous rates of mutations. It must be understood that natural populations living in isolation of each other diverge genetically as the combined outcome of selection forces acting independently on two different populations and due to random genetic drift[80]. The latter will affect all microsatellite loci across the genome. On the other hand, natural selection will act only at specific loci. The patterns created by the two forces are different and detectable as directional selection leads to higher differentiation of allele frequencies at the loci under natural selection, exceeding the expected values or the ones caused by random drift[81], thus making them outlier loci, i.e. more active in divergence. This eventually creates linkage disequilibrium between the microsatellites present at random sites and the targets of selection.

The 'genomic islands of divergence' described by Nosil *et al.*[79] for different genomic loci, in general, can be well adapted to explain the participation of microsatellites in adaptive evolution as well. According to this model, if all the microsatellite loci are sequentially plotted according to the heterogeneity occurring at these loci in such a way that heterogeneity is plotted on the Y-axis, then the part of the curve closer to the X-axis is called 'sea floor', while the 'sea level (or surface)' is determined as the threshold above which the microsatellite heterogeneity is greater than expected by neutral evolution alone[79]. Natural selection determines both the island elevation and the island size. Tightly-linked loci fall into these regions, and the island size is an indication of the contiguous loci naturally selected for. In this model, loosely-linked loci remain closer to the sea surface and are differentiated by selection forces[79]. The sea floor is constituted of microsatellite loci unlinked to the genomic regions involved in adaptive divergence. However, these loci too may show variable mutation rates, thereby creating peaks and valleys in the sea floor as well. These are determined by genetic drift alone[79], and adaptive selection forces have little to do with these peaks and valleys. Analogous to geographical islands, genomic islands of divergence keep evolving with time due to recurrent mutations, and life cycle patterns of microsatellites themselves. Islands may also grow if the genomic sequences adjacent to the regions already under divergent selection show further adaptive mutations[82].

Proving evolutionary and ecological significance of microsatellites remains a difficult task, especially in the absence of empirically detectable phenotype in majority of the cases. If a microsatellite is occurring within a gene

sequence or is conserved in its regulatory region, its homology and functionality to model organism can be studied[83–86]. Nevertheless, non-coding regions are also equally affected by selection forces as coding regions are affected[87]. There are evidences of frequent positive selections in the genomes of several species including humans[88]. Edelist *et al.*[85] similarly demonstrated skewness in allele distribution of microsatellites in the genomic regions selected for salt tolerance in *Helianthus paradoxus*. However, such genome scan methods can only provide candidate loci participating in adaptations and genome divergence, subject to the verification by selection experiments.

Concluding remarks

Modern genomes particularly eukaryotic genomes are smarter structures than used to be thought earlier. A large chunk of non-coding DNA and repetitive DNA which was earlier thought of as junk is now being recognized as sequences with specific roles, which are strategically located and carefully selected during evolution and divergence. Microsatellites represent such structures, which may even constitute parts of the genes and from therein regulate a number of biological processes of the cell. Substantial efforts have been made to understand their genomic locations, associated functions and the mode of their evolution. Genome comparisons are now possible for such studies especially after the availability of large amount of genome sequence data in the public domain. Such comparisons facilitate understanding of the conservation and divergence patterns of microsatellites over long evolutionary scales. This would help drawing consensus on modes of evolution and evolutionary biases that determine microsatellite variability and functionality in the genomes.

1. Haubold, B. and Wiehe, T., How repetitive are genomes? *BMC Bioinformatics*, 2006, **7**, 541.
2. Shapiro, J. A. and von Sternberg, R., Why repetitive DNA is essential to the genome function? *Biol. Rev.*, 2005, **80**, 227–250.
3. Kashi, Y. and King, D. J., Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.*, 2006, **22**, 253–259.
4. Vinces, M. D. *et al.*, Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*, 2009, **324**, 1213–1216.
5. Sharma, P. C., Grover, A. and Kahl, G., Mining microsatellites in eukaryotic genomes. *Trends Biotechnol.*, 2007, **25**, 490–498.
6. Piegu, B. *et al.*, Doubling genome size without polyploidization: dynamics of retrotransposon-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.*, 2006, **16**, 1262–1269.
7. Warren, W. C. *et al.*, Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 2008, **453**, 175–183.
8. Eckert, K. A. and Hile, S. E., Every microsatellite is different: intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol. Carcinog.*, 2009, **48**, 379–388.
9. Shah, S. N., Hile, S. E. and Eckert, K. A., Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. *Cancer Res.*, 2010, **70**, 431–435.
10. Kofler, R., Schlotterer, C., Luschutzky, E. and Lelley, T., Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. *BMC Genomics*, 2008, **9**, 612.
11. Buschiazzo, E. and Gemmell, N. E., Conservation of human microsatellites across 450 million years of evolution. *Genome Biol. Evol.*, 2010, **2**, 153–165.
12. Merkel, A. and Gemmell, N., Detecting short tandem repeats from genome data: opening the software black box. *Brief. Bioinform.*, 2008, **9**, 355–366.
13. McConnell, R. *et al.*, An unusually low microsatellite mutation rate in *Dictyostelium discoideum*, an organism with unusually abundant microsatellites. *Genetics*, 2007, **177**, 1499–1507.
14. Grover, A., Aishwarya, V. and Sharma, P. C., Biased distribution of microsatellite motifs in the rice genome. *Mol. Genet. Genomics*, 2007, **277**, 469–480.
15. Grover, A. and Sharma, P. C., Microsatellite motifs with moderate GC content are clustered around genes on *Arabidopsis thaliana* chromosome 2. *In Silico Biol.*, 2007, **7**, 201–213.
16. Roorkiwal, M., Grover, A. and Sharma, P. C., Genome-wide analysis of conservation and divergence of microsatellites in rice. *Mol. Genet. Genomics*, 2009, **282**, 205–215.
17. Bayliss, C. D., Field, D. and Moxon, E. R., The simple sequence contingency loci of *Haemophilus influenza* and *Neisseria meningitides. J. Clin. Invest.* 2001, **107**, 657–666.
18. Li, Y.-C. *et al.*, Microsatellites within genes: structure, function and evolution. *Mol. Biol. Evol.*, 2004, **21**, 991–1007.
19. Levdansky, E., Romano, J. and Shadkchan, Y., Coding tandem repeats generate diversity in *Aspergillus fumigatus* genes. *Eukaryot. Cell*, 2007, **6**, 1380–1391.
20. Mrazek, J., Analysis of distribution indicates diverse functions of simple sequence repeats in *Mycoplasma* genomes. *Mol. Biol. Evol.*, 2006, **23**, 1370–1385.
21. Martin, P. *et al.*, Microsatellite instability regulates transcription factor binding and gene expression. *Proc. Natl. Acad. Sci. USA*, 2005, **102**, 3800–3804.
22. Debrauwere, H. *et al.*, Differences and similarities between various tandem repeat sequences: minisatellites and microsatellites. *Biochimie*, 1997, **79**, 577–586.
23. Usdin, K., The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res.*, 2008, **18**, 1011–1019.
24. Ladd, P. D., Smith, L. E. and Rabia, N. A., An antisense transcript spanning the CGG repeat region of FMR1 is upregulated in permutation carriers but silenced in full mutation individuals. *Hum. Mol. Genet.*, 2007, **16**, 3174–3187.
25. Tassone, F. *et al.*, Elevated FMR1 mRNA in permutation carriers is due to increased transcription. *RNA*, 2007, **13**, 555–562.
26. Sureshkumar, S. *et al.*, A genetic defect caused by a triplet repeat expansion in *Arabidopsis thaliana. Science*, 2009, **323**, 1060–1063.
27. Kovtun, I. V. and McMurray, C. T., Features of trinucleotide instability *in vivo. Cell Res.*, 2008, **18**, 383–395.
28. Donaldson, Z. R. *et al.*, Evolution of a behaviour-linked microsatellite containing element in the 5′-flanking region of the primate *AVPR1A* gene. *BMC Evol. Biol.*, 2008, **8**, 180.
29. Fabre, E., Dujon, B. and Richard, G. F., Transcription and nuclear transport of CAG/CTG trinucleotide repeats in yeast. *Nucl. Acid Res.*, 2002, **30**, 3540–3547.
30. Mankodi, A. *et al.*, Expanded CUG repeats trigger aberrant splicing of CLC-1 chloride channel pre-mRNA and hyperexcitability of skeletal muscle in mytonic dystrophy. *Mol. Cell*, 2002, **10**, 35–44.
31. Hsieh, P. and Yamane, K., DNA Mismatch repair: molecular mechanism, cancer and ageing. *Mech. Ageing Dev.*, 2009, **129**, 391–407.

32. Lawson, M. J. and Zhang, L., Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5′-UTR region. *Gene*, 2008, **407**, 54–63.

33. Loire, E. *et al.*, Hypermutability of genes in *Homo sapiens* due to the hosting of long mono-SSR. *Mol. Biol. Evol.*, 2009, **26**, 111–121.

34. van Passel, M. W. J. and de Graaf, L. H., Mononucleotide repeats are asymmetrically distributed in fungal genes. *BMC Genomics*, 2008, **9**, 596.

35. Eichler, E. E., Repetitive conundrums of centromere structure and function. *Hum. Mol. Genet.*, 1999, **8**, 151–155.

36. Frydrychova, R. *et al.*, Phylogenetic distribution of TTAGG telomeric repeats in insects. *Genome*, 2004, **47**, 163–178.

37. Sykorova, E., Lim, K. Y. and Kunicka, Z., Telomere variability in the monocotyledonous plant order Asparagales. *Proc. R. Soc. Lond. B Biol. Sci.*, 2003, **270**, 1893–1904.

38. Field, D. and Wills, C., Long, polymorphic microsatellites in simple organisms. *Proc. R. Soc. Lond. B Biol. Sci.*, 1996, **263**, 209–215.

39. Li, Y.–C. *et al.*, Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.*, 2002, **11**, 2453–2465.

40. Bagshaw, T. M., Pitt, J. P. W. and Gemmell, N. J., High frequency of microsatellites in *S. cerevisiae* meiotic recombination hotspots. *BMC Genomics*, 2008, **9**, 49.

41. Freudenreich, C. H., Stavenhagen, J. B. and Zakian, V. A., Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. *Mol. Cell. Biol.*, 1997, **17**, 2090–2098.

42. Hile, S. E. and Eckert, K. A., Positive correlation between DNA polymerase alpha-primase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences. *J. Mol. Biol.*, 2004, **335**, 745–759.

43. Fujimori, S. *et al.*, A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. *FEBS Lett.*, 2003, **554**, 17–22.

44. Morel, P. *et al.*, The role of SOS and flap processing of microsatellite instability in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 10003–10008.

45. Ellegren, H., Microsatellite mutations in the germline: implications for the evolutionary inference. *Trends Genet.*, 2000, **16**, 551–558.

46. Spampinato, C. P. *et al.*, From bacteria to plants: A compendium of mismatch repair assays. *Mutat. Res.*, 2009, **682**, 110–128.

47. Kelkar, Y. D., The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.*, 2008, **18**, 30–38.

48. Marriage, T. N. *et al.*, Direct estimation of the mutation rate at dinucleotide microsatellite loci in *Arabidopsis thaliana* (Brassicaceae). *Heredity*, 2009, **107**, 310–317.

49. Phillips, N. *et al.*, Spontaneous mutation and standing genetic (co) variation at dinucleotide microsatellites in *Caenorhabditis briggsae* and *Caenorhabditis elegans*. *Mol. Biol. Evol.*, 2009, **26**, 659–669.

50. Sagher, D., Hsu, A. and Strauss, B., Stabilization of the intermediate in frameshift mutation. *Mutat. Res.*, 1999, **423**, 73–77.

51. Hawk, J. D. *et al.*, Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. *Proc. Natl. Acad. Sci. USA*, 2005, **102**, 8639–8643.

52. Prendergast, J. G. D. *et al.*, Chromatin structure and evolution in the human genome. *BMC Evol. Biol.*, 2007, **7**, 72.

53. Ohta, T. and Kimura, M., A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.*, 1973, **22**, 201–204.

54. Kimura, M. and Crow, J. F., The number of alleles that can be maintained in a finite population. *Genetics*, 1964, **49**, 725–738.

55. Di Rienzo, A. *et al.*, Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 3166–3170.

56. Weetman, D., Hauser, L. and Carvalho, G. R., Reconstruction of microsatellite mutation history reveals a strong and consistent deletion bias in invasive clonal snails, *Potamopyrgus antipodarum*. *Genetics*, 2002, **162**, 813–822.

57. Lai, Y. and Sun, F., The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol. Biol. Evol.*, 2003, **20**, 2123–2131.

58. Kruglyak, S. *et al.*, Equillibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 10774–10778.

59. Calabrese, P. P., Durrett, R. T. and Aquadro, C. A., Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation model. *Genetics*, 2001, **159**, 839–852.

60. Sainudiin, R. *et al.*, Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics*, 2004, **168**, 383–395.

61. Dieringer, D. and Schlotterer, C., Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res.*, 2003, **13**, 2242–2251.

62. Ellegren, H., Microsatellites: simple sequences with complex evolution. *Nature Rev. Genet.*, 2004, **5**, 435–445.

63. Tay, W. T. *et al.*, Generation of microsatellite repeat families by RTE retrotransposons in lepidopteran genomes. *BMC Evol. Biol.*, 2010, **10**, 144.

64. Vowles, E. J. and Amos, W., Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. *Mol. Biol. Evol.*, 2006, **23**, 598–607.

65. Harr, B., Zangerl, B. and Schlotterer, C., Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from *Drosophila*. *Mol. Biol. Evol.*, 2000, **17**, 1001–1009.

66. Barrier, M. *et al.*, Interspecific evolution in plant microsatellite structure. *Gene*, 2000, **241**, 101–105.

67. Xita, N. *et al.*, Association of the $(TAAAA)_n$ repeat polymorphism in the sex hormone-binding globulin (SHBG) gene with polycystic ovary syndrome and relation to SHBG serum levels. *J. Clin. Endocrinol. Metab.*, 2003, **88**, 5976–5980.

68. White, M. C. *et al.*, Inactivation of hnRNP by expanded intronic AUUCU repeat induces apoptosis via translocation of PKCdelta to mitochondria in spinocerebellar ataxia 10. *PLoS Genet.*, 2010, **6**, e1000984.

69. Hollox, E. *et al.*, Psoriasis is associated with increased beta-defensin genomic copy number. *Nature Genet.*, 2008, **40**, 23–25.

70. Takezaki, N. and Nei, M., Genomic drift and evolution of microsatellite DNAs in human populations. *Mol. Biol. Evol.*, 2009, **26**, 1835–1849.

71. Sclotterer, C., Hitchhiking mapping-functional genomics from the population genetic perspective. *Trends Genet.*, 2003, **19**, 32–38.

72. Smykal, P. *et al.*, Evolutionary conserved lineage of Angela-family retrotransposon as a genome-wide microsatellite repeat dispersal agent. *Heredity*, 2009, **103**, 157–167.

73. Kayser, M. *et al.*, Microsatellite length differences between humans and chimpanzees at autosomal loci are not found at equivalent haploid Y chromosomal loci. *Genetics*, 2006, **173**, 2179–2186.

74. Jensen, J. D., Wong, A. and Aquadro, C. F., Approaches for identifying targets of positive selection. *Trends Genet.*, 2007, **23**, 568–577.

75. Schlotterer, C., Kauer, M. and Dieringer, D., Allele excess at neutrally evolving microsatellites and the implications for tests of neutrality. *Proc. R. Soc. Lond. B Biol. Sci.*, 2004, **271**, 869–874.

76. Samdja, C., Galindo, J. and Butlin, R., Hitching a lift on the road to speciation. *Mol. Ecol.*, 2008, **17**, 4177–4180.

77. Via, S. and West, J., The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Mol. Ecol.*, 2008, **17**, 4334–4345.

78. Storz, J. F., Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol. Ecol.*, 2005, **14**, 671–688.

79. Nosil, P., Funk, D. J. and Orbitz-Barrientos, D., Divergent selection and heterogenous genomic divergence. *Mol. Ecol.*, 2009, **18**, 375–402.

80. Nielsen, R., Molecular signatures of natural selection. *Annu. Rev. Genet.*, 2005, **39**, 197–218.

81. Egan, S. P., Nosil, P. and Funk, D. J., Selection and genomic differentiation during ecological speciation: isolating the contributions of host association via a comparative genome scan of *Neochlamisus bebbianae* leaf beetles. *Evolution,* 2008, **62**, 1162–1181.

82. Kirkpatrick, M. and Barton, N. H., Chromosome inversions, local adaptation and speciation. *Genetics*, 2006, **173**, 419–434.

83. Olafsdottir, G. A., Snorrason, S. S. and Ritchie, M. G., Parallel evolution? Microsatellite variation of recently isolated marine and freshwater three-spined stickleback. *J. Fish Biol.*, 2007, **70**, 125–131.

84. Makinen, H. S. *et al.*, Hitchhiking mapping reveals a candidate genomic region for natural selection in three-spined stickleback chromosome VIII. *Genetics*, 2008, **178**, 435–465.

85. Edelist, C. *et al.*, Microsatellite signature of ecological selection for salt tolerance in a wild sunflower hybrid species, *Helianthus paradoxus*. *Mol. Ecol.*, 2006, **15**, 4623–4634.

86. Vasemagi, A., Nilsson, J. A. and Primmer, C. R., Expressed sequence tagged-linked microsatellites as a source of gene associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). *Mol. Biol. Evol.*, 2005, **22**, 1067–1076.

87. Wright, S. I. And Andolfatto, P., The impact of natural selection on the genome: emerging patterns of *Drosophila* and *Arabidopsis*. *Annu. Rev. Ecol. Evol. Syst.*, 2008, **39**, 193–213.

88. Hermission, J., Who believes in whole-genome scans for selection? *Heredity,* 2009, **103**, 283–284.

89. Richeti, F. *et al.*, Evaluation of $(AC)_n$ and C(-106)T polymorphisms of the aldose reductase gene in Brazilian patients with DM1 and susceptibility to diabetic retinopathy. *Mol. Vis.*, 2007, **13**, 740–745.

90. Warpeha, K. M. *et al.*, Genotyping and functional analysis of a polymorphic $(CCTTT)_n$ repeat of *NOS2A* in diabetic retinopathy. *FASEB J.*, 1999, **13**, 1825–1832.

91. Awad, M. *et al.*, CA repeat allele polymorphism in the first intron of the human interferon-gamma gene is associated with lung allograft fibrosis. *Hum. Immunol.*, 1999, **60**, 343–346.

92. Dufor, C., Homozygosis for 12 (CA) repeats in the first intron of the human IFN-gamma gene is significantly associated with the risk of aplastic anaemia in Caucasian population. *J. Haematol.*, 2004, **126**, 682–685.

93. Mantzou, E., Doukas, C. and Georgiou, I., The importance of the $(TAAAA)_n$ alleles at the SHBG gene promoter for the severity of coronary artery disease in postmenopausal women. *Menopause*, 2008, **15**, 461–468.

94. Sato, N. *et al.*, Spinocerebellar ataxia type 31 is associated with 'inserted' penta-nucleotide repeats containing $(TGGAA)_n$. *Am. J. Hum. Genet.,* 2009, **85**, 544–557.

95. Edwards, S. F. *et al.*, A Z-DNA sequennce reduces slipped-strand structure formation in myotonic dystrophy type 2 $(CCTG) \times (CAGG)$ repeat. *Proc. Natl. Acad. Sci. USA*, 2009, **106**, 3270–3275.

96. Lalioti, M. D., Antonarakis, S. E. and Scott, H. S., The epilepsy, the protease inhibitor and the dodecamer: progressive myoclonus epilepsy, cystatin b and a 12-mer repeat expansion. *Cytogenet. Genome Res.*, 2003, **100**, 213–223.

97. Overbach, D. and Gabbay, K. H., Localization of a type I diabetes susceptibility locus to the variable tandem repeat region flanking the insulin gene. *Diabetes*, 1993, **42**, 1708–1714.