

# Microsatellites

Jörg T Epplen, *Human genetics, Ruhr-University, Bochum, Germany*

Amer D Akkad, *Human genetics, Ruhr-University, Bochum, Germany*

Based in part on the previous version of this *Encyclopedia of Life Sciences (ELS)* article, *Microsatellites* by Jörg T Epplen and Stefan Böhringer.

**Microsatellites are short, in-tandem arranged repetitive deoxyribonucleic acid (DNA) elements which are widespread in eukaryotic and prokaryotic genomes. The basic unit lengths of microsatellites (or simple DNA sequence repeats) comprise up to six (or eight) nucleotides, and these motifs are perfectly reiterated from 5 to more than 100 times. Longer and more imperfectly reiterated periodicities border on the definitions of the so-called minisatellites. Microsatellites have been utilised as markers for population genetic studies, forensic and relatedness testing, investigations on genetic diversity and the identification of genetic traits. Albeit constituting elusive model disorders, microsatellite expansion diseases are rather rare entities characterised by differential pathogenic pathways. Hence microsatellites exhibit a broad spectrum of biological relevance ranging from selfish or neutral to pathogenic elements.**

## Repetitive Sequences in DNA

Genomic redundancy is a ubiquitous phenomenon, in both eukaryotic and prokaryotic organisms as caused by duplication or multiplication of partial sequences or the entire genome. The full representation of all types of repetitive sequences could be evaluated exactly only after complete sequence analysis of a given genome. Yet the evaluation of repetitive deoxyribonucleic acids (DNAs) critically depends on the computational algorithm employed (Leclercq *et al.*, 2007).

In humans, for example, coding sequences comprise less than 5% of the DNA, whereas repeat sequences account for at least 50% (see e.g. <http://www.ensembl.org/>). The latter sequences may be classified in different categories:

- Segmental duplications of about 10 kilobases (kb) to several megabases, copied from one genomic region

ELS subject area: Molecular Biology

### How to cite:

Epplen, Jörg T; and Akkad, Amer D (April 2011) Microsatellites. In: *Encyclopedia of Life Sciences (ELS)*. John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a0005911.pub2

## Advanced article

### Article Contents

- Repetitive Sequences in DNA
- Definition of Microsatellites
- Occurrence and Properties of Microsatellites
- Markers or More?

Online posting date: 15<sup>th</sup> April 2011

- into another (Ji *et al.*, 2000), including paralogues and pseudogenes. **See also:** [Paralogous Genes and Chromosomal Regions](#); [Pseudogene Evolution in the Human Genome](#)
- Transposon-derived repeats interspersed ubiquitously including inactive (partially) retrotransposed copies of cellular genes (Smit, 1999). **See also:** [Transposons: Eukaryotic](#)
  - Classical satellites, as evidenced by buoyant density gradient centrifugation (for review see Skinner, 1977), located at centromeres, telomeres, short arms of acrocentric chromosomes and ribosomal gene clusters (underrepresented in whole-genome sequencing projects for methodological reasons). **See also:** [Evolution of Alpha Satellite DNA](#)
  - Minisatellites (Jeffreys *et al.*, 1985) interspersed throughout in many eukaryotic and bacterial genomes. **See also:** [Minisatellites](#)
  - Microsatellites or simple sequence repeats (Tautz, 1993; Weber and Wong, 1993) scattered genome wide in practically all genomes investigated. **See also:** [Simple Repeats](#)

## Definition of Microsatellites

Among the five categories of repeated DNA elements aforementioned, microsatellites are most striking genomic hallmarks. In the otherwise amorphous sequence string of A's, C's, G's and T's, perfect blocks such as ...GTGTGTGTGTGTGTGTGT... or ...GATAGATA-GATAGATAGATA... are immediately obvious to the naked eye. Extensive use of the polymerase chain reaction had sparked interest in these microsatellites as readily detectable and informative locus markers. Investigations using DNA sequencing as well as *in vivo* models (see below) have furthered the understanding of the occurrence and in part also of the functional relevance of the corresponding mechanisms of microsatellites. By definition microsatellites consist of tandem repeat unit lengths of one to six (or even eight) nucleotides/base pairs (bp) (Tautz, 1993). At each microsatellite locus 5–100 or more repeat units can be found. This definition of microsatellites or simple sequence repeats is not used consistently. Different designations

and abbreviations are found in the literature describing microsatellites such as STR (for short tandem repeats) or SSR (for simple sequence repeats). Exceptionally, repeat unit lengths of up to 13 bases have been included under the microsatellite designation (see International Genome Sequencing Consortium, 2001). Such broader categorisations infringe the demarcation from minisatellites (Jeffreys *et al.*, 1985), an entity defined earlier than microsatellites. Another rarely employed designation for such repetitive DNA sequences is the abbreviation VNTR (for variable number of tandem repeats). Since any locus showing variable tandem repetition could be interpreted as falling within this unspecific definition, it is not a useful designation for a certain class of sequences. Other abbreviations to be discerned from microsatellites are LINEs (for long interspersed nuclear elements in the range of 6–8 kbp origin) or SINEs (for short interspersed repetitive elements in the range of 260–300 bp as composed of two 130 bp monomers separated by a short A-rich linker). Those genomic elements can be distinguished from microsatellites by size and sequence content. **See also:** [Long Interspersed Nuclear Elements \(LINEs\): Evolution](#); [Short Interspersed Elements \(SINEs\)](#)

## Occurrence and Properties of Microsatellites

Many microsatellites are more or less randomly scattered throughout the genomes, both in protein-coding and noncoding regions (Tóth *et al.*, 2000). In exons, such elements are less abundant than in noncoding regions, but they may occur even in the smallest bacterial genomes. In man microsatellites are present on average every 2000 bp (International Genome Sequencing Consortium, 2001). Despite of potential representation bias, a few rules for the overall evolutionary success of simple repeats can be formulated on the basis of their existence in present-day genomes:

- Microsatellites are observed as simple or perfect repetitions  $(X)_n$  as well as in complex or imperfect  $(X)_n(Y)_m(X)_o\dots$  patterns.
- The nomenclature of microsatellites is derived from number of bases in the basic repeated nucleotide unit: di-, tri-, tetra-, ...
- Longer CpG-containing simple repeat motifs  $((GC)_n, (GGC)_n, (GAC)_n, (GGGC)_n, (GAAC)_n)$  are exceedingly rare, probably because of CpG methylation and increased mutability of these dinucleotides.
- Long AT-rich simple repeats  $((GT)_n, (GA)_n, (AT)_n, (GAA)_n, (GATA)_n, (GAAA)_n, (AAAT)_n)$  outnumber GC-rich elements  $((GGT)_n, (GGGT)_n)$  by far. Microsatellites with selected polypurine compositions on one strand  $((GAA)_n, (GGAA)_n, (GAAA)_n)$  are regularly contained in particularly long blocks.
- Perfect trinucleotide periodicity is observed in coding as well as in non coding regions. Exonic trinucleotide

repeats may lead to the expansion of defined amino acid blocks. So far in coding regions mostly poly-glutamine and poly-alanine stretches have been identified.

- Perfect di- or tetranucleotide periodicity warrants evolutionary success for simple repeats  $((GT)_n, (GA)_n, (AT)_n, (GATA)_n, (GGAA)_n, (GAAA)_n, (AAAT)_n)$  (Epplen *et al.*, 1998).

The nucleotide composition of the most abundant simple repeat motifs is highly reminiscent of the AT/GC content of spacer and intron DNA in many eukaryotes. To date, no satisfactory explanation exists for the abundance of these repeated genomic elements, and they remain puzzling objects for future investigation.

## Markers or More?

Microsatellites represent most polymorphic genetic markers, and they are also highly abundant, only outnumbered by single nucleotide polymorphisms. This high degree of informativeness is due to inherent length instability (Weber and Wong, 1993). Different mechanisms, actually all mechanisms involved in DNA synthesis, have been postulated and partially shown to be involved in microsatellite instability using *Escherichia coli*, yeast and rodents as model systems. Changes in the secondary DNA structure are caused by the repetitive DNA elements and microsatellites may lead to sequence instabilities during replication (slippage), repair (hairpin structures cleaved by DNA nucleases) and uneven recombination events of homologous regions (Richard *et al.*, 2008). Yet our understanding of how the different classes of tandem repetitive DNA elements arise and how they produce polymorphisms is still incomplete (Richard *et al.*, 2008). Mutation rates are estimated to range from  $10^{-6}$  to  $10^{-2}$  per locus and generation, depending on species, locus and repeat number (Schlötterer, 2000; Eckert and Hile, 2009). Mutations at microsatellite loci therefore outnumber base substitutions by far. Microsatellites have initially been valuable markers for population genetics, especially for large-scale studies and are still used in some diagnostic areas, for example paternity, forensic and genome instability testing. As markers microsatellites have disadvantages based on their highly polymorphic character and their limited utility for high throughput. Therefore, microsatellites are replaced as markers for genetic association studies by SNPs for technical, abundance as well as efficiency reasons. **See also:** [Single Nucleotide Polymorphism \(SNP\)](#)

What is the biological meaning of microsatellites in genomic DNA? Traditional molecular genetic and biological reasoning has not yielded a satisfactory explanation for the existence and preponderance of simple repetitive DNA elements. Yet, the intrinsic polymorphisms of simple repeats have been exploited in many aspects of research. It has been believed that genomic redundancy may lead to

increased fitness in the case of beneficial or obstructive point mutations (Kafri *et al.*, 2006). There is also evidence that different microsatellites have specialised properties, i.e. different secondary structures, different protein coding or binding properties with impact on protein function as, for example discussed for dogs: Fondon and Garner (2004) presented evidence for correlations of microsatellite lengths in developmentally relevant as well as brain genes with morphological skull changes and influence on brain functioning (Fondon *et al.*, 2008). Specific DNA structures may represent structural hallmarks in the nucleus, since different tandem repeats exhibit various secondary structures (Dion and Wilson, 2009). These structural properties may relate to the differential evolutionary potentials of microsatellites.

Some of the simple trinucleotide repeats in microsatellites and in certain minisatellites are not functionally neutral DNA elements. Depending on the exact location with respect to a gene, microsatellites can be detrimental to gene function, as evidenced by the involvement of microsatellites in repeat expansion diseases in humans (Richard *et al.*, 2008). Noncoding trinucleotide microsatellites may modulate methylation, transcription or influence splicing processes. Myotonic dystrophy type 2 (DM2) for example is caused by the expansion of a complex repeat motif,  $(TG)_n(TCTG)_m(CCTG)_o$  in intron 1 of the *CNBP* gene. Crucial is the expansion of the CCTG repeat which causes DM2, whereby expansions exceeding 75 units are pathogenic. Interestingly in affected individuals repeat numbers can vary between 75 to more than 11 000, with a mean of approximately 5000 repeats (Dalton *et al.*, 2007). In contrast, expansion of coding trinucleotide microsatellites leads to poly amino acid tracts. In coding regions exclusively poly alanine and poly-glutamine have been identified leading among gene-specific effects to fatal protein aggregations. Huntington disease is one of the best characterised dominantly inherited trinucleotide expansion disorders. Individuals with expanded alleles (>26 CAG repeats) are classified in three subcategories, so-called carriers of intermediate alleles ( $\geq 27$ –35 CAG repeats), carriers of HD-causing alleles with reduced penetrance (36–39 CAG repeats) and carriers of HD-causing alleles with full penetrance ( $\geq 40$  CAG repeats). Individuals with intermediate alleles are usually not at risk of developing symptoms of HD. Yet, due to instability of the CAG tract during meiosis, their offspring is prone to develop HD in case of further CAG expansion. Carriers of reduced-penetrance alleles develop in rare cases symptoms of HD, and individuals with full-penetrance alleles will develop symptoms of HD with high probability (Warby *et al.*, 2010). Despite the abundance of microsatellites and their polymorphic properties up to now only three dozens of diseases are associated with simple repeat expansions. **See also: Trinucleotide Repeat Expansions: Disorders**

In summary, the biological relevance, if any, of most simple tandem repetitive DNA sequences has still to be characterised. Indeed, conventional biological categories may not be appropriate for a full understanding of this

conspicuous form of genetic redundancy. These abundant elements may also be involved in complex multifactorial disorders and in the characters of quantitative trait loci since incremental length increases may result in gradual accentuation of symptoms of disease and changes in quantitative phenotype traits. **See also: Megasatellite DNA; Minisatellites; Simple Repeats**

## References

- Dalton JC, Ranum LPW and Day JW (2007) Myotonic dystrophy type 2. *Gene Reviews* PMID: 20301639 [Internet].
- Dion V and Wilson JH (2009) Instability and chromatin structure of expanded trinucleotide repeats. *Trends in Genetics* **25**: 288–297.
- Eckert KA and Hile SE (2009) Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Molecular Carcinogenesis* **48**: 379–388.
- Epplen JT, Mäueler W and Santos EJM (1998) On GATAGATA and other ‘junk’ in the barren stretch of genomic desert. *Cytogenetics and Cell Genetics* **80**: 75–82.
- Fondon JW 3rd, Hammock EA, Hannan AJ and King DG (2008) Simple sequence repeats: genetic modulators of brain function and behavior. *Trends in Neuroscience* **31**: 328–334.
- Fondon JW and Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the USA* **101**: 18058–18063.
- International Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- Jeffreys AJ, Wilson V and Thein SL (1985) Hypervariable minisatellite regions in human DNA. *Nature* **314**: 67–73.
- Ji Y, Eichler EE, Schwartz S and Nicholls RD (2000) Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Research* **10**: 597–610.
- Kafri R, Levy M and Pilpel Y (2006) The regulatory utilization of genetic redundancy through responsive backup circuits. *Proceedings of the National Academy of Sciences of the USA* **103**(31): 11653–11658.
- Leclercq S, Rivals E and Jarne P (2007) Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics* **8**: 125.
- Richard GF, Kerrest A and Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews* **72**: 686–727.
- Schlötterer C (2000) Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**: 365–371.
- Skinner D (1977) Satellite DNAs. *Bioscience* **27**: 790–796.
- Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics and Development* **9**: 657–663.
- Tautz D (1993) Notes on the definition and nomenclature of tandemly repetitive DNA sequences. In: Pena SDJ, Chakraborty R, Epplen JT and Jeffreys AJ (eds) *DNA Fingerprinting: State of the Science*, pp. 21–28. Basel: Birkhäuser.

- Tóth G, Gaspari Z and Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research* **10**: 967–981.
- Warby SC, Graham RK and Hayden MR (2010) Huntington disease. *Gene Reviews* PMID: 20301482 [Internet].
- Weber J and Wong C (1993) Mutation of human short tandem repeats. *Human Molecular Genetics* **2**: 1123–1128.

## Further Reading

- Hancock JM (1996) Simple sequences in a 'minimal' genome. *Nature Genetics* **14**: 14–15.
- Ohno S (1972) Evolutional reason for having so much junk DNA. In: Pfeiffer RA (ed.) *Modern Aspects of Cytogenetics: Constitutive Heterochromatin in Man*, pp. 169–180. Stuttgart: FK Schattauer.