Title of Entry: Simple sequence repeats

Author Identification:
Calen P. Ryan
Northwestern University
Evanston, Illinois
CPR@u.nortwestern.edu

Synonyms:
Short tandem repeats, STRs, Microsatellites, Variable Number Tandem Repeats, VNTRs

Definition and Introduction:
Simple sequence repeats (SSRs), more often referred to as microsatellites or short tandem repeats, are short (1-9) tandemly repeated nucleotide sequences found ubiquitously across prokaryotic and eukaryotic genomes (Tóth et al. 2000). SSRs differ from minisatellite or satellite DNA, which consist of tandemly repeating sequences of 10-1000 and >1000 nucleotides, respectively. SSRs are highly-mutable, individually-variable sources of functionally-relevant genetic variation, with applications in medicine, forensics, population and behavioral genetics, and evolutionary biology.

SSR composition and genomic distribution:
SSRs are phylogenetically and genomically widespread, common to both prokaryotes and eukaryotes and occurring in subtelomeric, intragenic, intronic, and exonic regions (Tóth et al. 2000). SSRs often comprise large portions of host genomes, and may be associated with known genic regions. The most common SSR units are mono- and di-nucleotide repeats, but SSR unit length varies both by species and genomic location (Tóth et al. 2000). SSRs occur in both coding and non-coding DNA. Tri- and hexa-nucleotide repeats are more prevalent in exonic coding DNA, where they provide an important source of protein structural variability while maintaining reading frame. Poly (A/T) repeats tend to be more common than poly (C/G), but this also varies with genomic location.

While SSR variation in much of the genome is neutral, there is evidence for selection on repeat composition and length for many SSRs. In humans, trinucleotide CAGn motifs are overrepresented in exonic DNA, particularly in genes that code for proteins involved in neurodevelopment (Shimada et al. 2016). These CAG tracts, which code for the amino acid glutamine when expressed, may be evolutionarily favored at the level of the protein despite being under selection in DNA, and from their direct contribution to numerous human diseases (Shimada et al. 2016).

SSR 'birth' and evolution:
The 'birth' of SSRs involves de novo generation through conventional mutational processes (base-pair substitution and insertion-deletion events) or the modification of proto-microsatellites created during the dispersal of mobile transposable elements (Kelkar et al. 2008). Once formed, SSRs are hypothesized to change in length through replication slippage events, where disassociation of the replicating DNA strand is misaligned during reassociation (Viguera et al. 2001). While the genetic 'stutter' caused by replication slippage is usually corrected by the DNA mismatch repair system, failure to do so can result in rapid expansions or contractions in repeat number. The rapid expansion and contraction of SSRs leads to a great deal of mutability in repeat number in SSRs – up to 100,000x's higher than for point mutations within the a given species. High mutation rates result in hypervariability in repeat length for many SSRs, making them an important source of genetic diversity (Tóth et al. 2000). For this reason, SSRs have been used extensively as genetic markers for population genetics, paternity tests, and forensic applications.

At the level of the population, SSR repeat number often resembles a bell-shaped – or occasionally a multimodal – distribution (Shimada et al. 2016). Such distributions could arise from a combination of opposing molecular and/or selective processes. Longer repeats become increasingly unstable, and are more likely to accumulate additional repeats through replication slippage. Repeat expansion of SSRs also contributes to certain forms of protein toxicity and disease. However, slippage may result in the excision of numerous repeats in a single event, for example through the creation and exclusion of DNA hairpin loops. Such instability results in the rapid expansion or contraction of many SSRs through time.

Given the inherently unstable nature of most SSRs, especially under conditions of cellular stress (Chatterjee et al. 2015), their prevalence across host genomes is somewhat surprising. However, that same instability makes SSRs an important source of genetic diversity in challenging new environments, potentially facilitating adaptation, evolution, and speciation. These characteristics, as well as the frequently graded phenotypic effects of repeat number, have led to some researchers to refer to SSRs as 'evolutionary tuning knobs' (King et al. 1997).

Functional consequences of SSR variation:
Both in vitro and in vivo studies have shown that repeat number for some SSRs can affect the binding of transcription factors or non-coding regulatory RNA, and/or alter protein form and function. These regulatory or conformational changes can lead to a range of morphological, cognitive, and behavioral phenotypes. In *Drosophila melanogaster*, variation in a threnonine-glycine repeat in the *period* clock gene affects circadian response to changes in temperature, a finding supported by clinal geographic variation in repeat length for this SSR among natural populations (Sawyer et al. 1997). SSR repeat length variation in genes involved in head development affect head shape in Drosophila, as well as in the stalk-eyed fly (*Teleopsis dalmanni*) (Birge et al. 2010). Notably, a significant genotype-phenotype association exists between several repeat motifs and eye-stalk length in *T. dalmannni*, a rapidly evolving trait under strong sexual selection in this species (Birge et al. 2010). The effect of SSR repeat length on morphological characteristics has also been demonstrated in other taxa, including mammals. Variation in repeat length for an SSR in the *Alx-4* gene is linked to polydactyly in both mice and the domestic dog, while variation in *Runx-2* seems to be partly responsible for inter-breed variability in facial and cranial structure among domestic dogs (Fondon and Garner 2004).

A number of behavioral phenotypes have also been shown to vary with variation in SSR repeat number. A polymorphic SSR in the *arginine vasopression 1a receptor* (*avpr1a*) gene is associated with differences in this receptor's binding affinity for vasopression, and is linked to both inter- and intra-specific differences in social behavior in among voles (Hammock and Young 2005). Similar variation in the *avpr1a* gene may correspond to socio-sexual differences between Chimpanzees (*Pan troglodytes*) and Bonobos (*Pan paniscus*), and may contribute to autism-associated socio-behavioral deficits in humans. Other genes with nearby or interveneing SSRs that have been associated with behavioral and cognitive phenotypes include genes coding for the serotonin transporter (SLC6A4), the dopamine transporter (SLC6A3), the dopamine receptor (DRD4), and the androgen receptor (AR). SSR variation in a polyglutamine tract in the human AR has also been associated with a range of morphological and behavioral phenotypes, and is a risk factor for several diseases.

SSRs and human disease:
Over 40 human diseases, referred to as 'repeat expansion disorders', have been linked to pathological SSR expansion. These disorders include Huntington's disease, myotonic

dystrophy, spinal and bulbar muscular atrophy, fragile X, and a number of spinocerebellar ataxias. The link between repeat number and disease include polyglutamine toxicity, altered RNA/protein levels, and protein gain/loss of function. Typically, symptom severity increases and age at onset decreases as repeat number expands with each generation. SSRs with longer repeat lengths are less stable and accumulate additional repeats more rapidly than shorter repeats. As a result, disease symptoms may arise at an earlier age with each generation, a process referred to as 'anticipation'. Notably, a disproportionate number of repeat expansion disorders are neurological in nature. Some researchers have argued that highly mutable SSRs may have played a fundamental role in the rapid evolution of brain and cognitive development in the human lineage, while also bearing a burden in the form of a susceptibility to a number of neuropathological diseases (Shimada et al. 2016).

Conclusion:
The biological significance of SSRs arises from their ubiquity, mutability, and molecular and phenotypic effects. While the pathways through which SSRs can affect health, behavior, and evolution are becoming clearer, the causes and consequences of repeat variation for the majority SSRs are still unknown. Rapid technological and analytical developments will no doubt improve the scale and resolution of data necessary to answer these questions, and aid in elucidating the role of SSRs in gene regulation, phenotypic variability, and evolution.

Cross-References:
Minisatellites, Transposable Elements, Repeat Sequences, Drosophila, Canine Morphology, Gene Expression, Genetic Variation, Microsatellite Marker, Mutations, Phenotype, Internal Clock, Biological Clocks, Insect Morphology, Dopamine Receptor, Serotonin, Androgens

References

Birge, L., M. Pitts, R. Baker, and G. Wilkinson. 2010. Length polymorphism and head shape association among genes with polyglutamine repeats in the stalk-eyed fly, Teleopsis dalmanni. *BMC Evolutionary Biology* 10: 227.

Chatterjee, N., Y. Lin, B. A. Santillan, P. Yotnda, and J. H. Wilson. 2015. Environmental stress induces trinucleotide repeat mutagenesis in human cells. *Proceedings of the National Academy of Sciences*: 201421917.

Fondon, J. W., and H. R. Garner. 2004. Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences* 101: 18058–18063.

Hammock, E. A. D., and L. J. Young. 2005. Microsatellite Instability Generates Diversity in Brain and Sociobehavioral Traits. *Science* 308: 1630–1634.

Kelkar, Y. D., S. Tyekucheva, F. Chiaromonte, and K. D. Makova. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research* 18: 30–38.

King, D. G., M. Soller, and Y. Kashi. 1997. Evolutionary tuning knobs. *Endeavour* 21: 36–40.

Sawyer, L. A., J. M. Hennessy, A. A. Peixoto, E. Rosato, H. Parkinson, R. Costa, and C. P. Kyriacou. 1997. Natural Variation in a Drosophila Clock Gene and Temperature Compensation. *Science* 278: 2117–2120.

Shimada, M. K., R. Sanbonmatsu, Y. Yamaguchi-Kabata, C. Yamasaki, Y. Suzuki, R. Chakraborty, T. Gojobori, and T. Imanishi. 2016. Selection pressure on human STR loci and its relevance in repeat expansion disease. *Molecular Genetics and Genomics* 291: 1–19.

Tóth, G., Z. Gáspári, and J. Jurka. 2000. Microsatellites in Different Eukaryotic Genomes: Survey and Analysis. *Genome Research* 10: 967–981.

Viguera, E., D. Canceill, and S. D. Ehrlich. 2001. Replication slippage involves DNA polymerase pausing and dissociation. *The EMBO Journal* 20: 2587–2595.