Short Tandem Repeats:

Variation, Evolution and Cancer

# What are short tandem repeats?

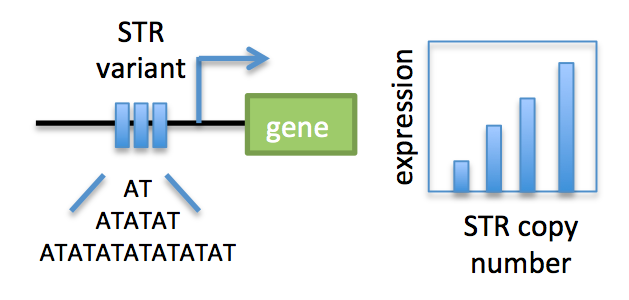
Short tandem repeats (STRs), stretches of repetitive DNA elements emerge as one of the major missing links in heritability and disease. STRs are short (1-6 base pair long) DNA tracks repeating 5 to 50 times in a tandem fashion. They are common in the human genome, contributing about 3%, a larger proportion than the entire sum of the protein coding sequences. STR change copy number 10−3 to 10−5 per cell division. Most such mutations result from replication slippage that escaped the proofreading of mismatch repair systems, providing a dynamic source of genetic and phenotypic variation. They have mutation rates at least 1000 times more often than that of SNPs. This extreme polymorphism posed significant obstacles in sequencing of STRs, which are overcome only recently by STR-compatible genotyping algorithms and PCR-free sequencing technologies.

# Why do they matter?

Due to their high polymorphism between individuals, STRs have been long used as markers in conservation studies, paternity tests, forensic and disease prognosis.

Also, studies show that STRs are enriched in gene regulatory regions. Copy number changes of those STRs can have an impact on the phenotype through various mechanisms including altered protein sequence, and gene regulation by various mechanisms, either genetic (such as changing the number of transcription factor binding sites) or epigenetic (i.e. by marking sites by methylation) (Figure 1). Recent genome-wide studies found association between STR length alterations and greater expression divergence in humans. Considering STR’s higher capacity of mutating, they are well able to challenge SNPs, when it comes to recent evolution. Indeed, STRs in gene promoters can enhance gene expression divergence between closely-related species such as humans and reflect local adaptations.

.



**Figure 1. Schematic representation of the impact of STR variants on gene expression**

# Approaches to annotate them

The most widely used tools to annotate STRs de novo include Tandem Repeats Finder, Repeat Masker. TRAL developed by Schaper et al in Anisimova group integrates information on the phylogeny through using known domains of proteins and advanced statistical methods.

# Approaches to genotype them

Lobstr is currently the most widely used tool to genotype and allelotype STRs in a population. It allows the user to apply filters on the number of reads, quality, position of the STR allele.

# Aim of the practice

You are the bioinformatician in a cancer research group studying STRs surrounding cancer genes. Your job is to identify how much polymorphism STRs around cancer genes show.

# Roadmap

1. Explore the cancer gene files and STR files.
2. Identify STRs surrounding cancer genes.
3. How many alleles do cancer genes show on average?