Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

A genomic view of short tandem repeats

Melissa Gymrek[1,2]

Short tandem repeats (STRs) are some of the fastest mutating loci in the genome. Tools for accurately profiling STRs from high-throughput sequencing data have enabled genome-wide interrogation of more than a million STRs across hundreds of individuals. These catalogs have revealed that STRs are highly multiallelic and may contribute more *de novo* mutations than any other variant class. Recent studies have leveraged these catalogs to show that STRs play a widespread role in regulating gene expression and other molecular phenotypes. These analyses suggest that STRs are an underappreciated but rich reservoir of variation that likely make significant contributions to Mendelian diseases, complex traits, and cancer.

Addresses
[1] Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA
[2] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA

Corresponding author: Gymrek, Melissa (mgymrek@ucsd.edu)

Current Opinion in Genetics & Development 2017, **44**:9–16

This review comes from a themed issue on **Molecular and genetic bases of disease**

Edited by **Nancy Bonini**, **Edward Lee** and **Wilma Wasco**

http://dx.doi.org/10.1016/j.gde.2017.01.012

Introduction

Short tandem repeats (STRs), also known as microsatellites, consist of repeating motifs of 1–6 base pairs (bp) and comprise about 3% of the human genome [1]. Their repetitive nature induces slippage events during DNA replication that result in frequent mutations in the number of repeats. As a result, STRs exhibit mutation rates that are orders of magnitude higher than other types of variation [2], and thus contribute a large fraction of human genetic variation.

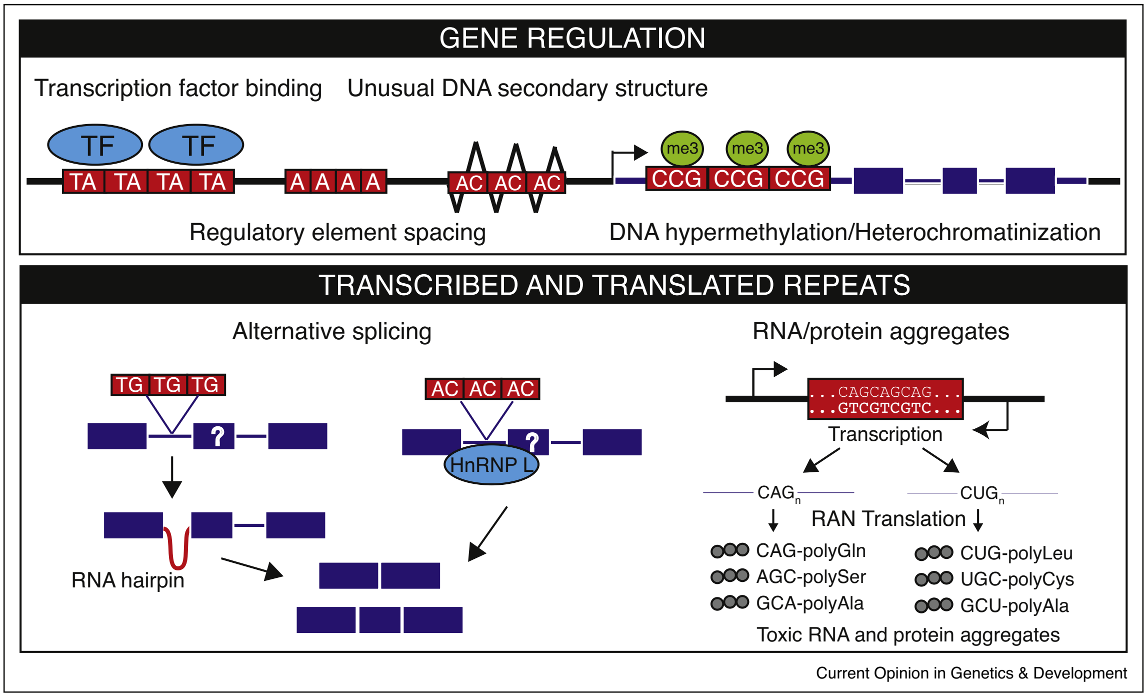
A role for STRs in human disease was established over two decades ago, with independent discoveries of trinucleotide expansions resulting in Fragile X Syndrome [3,4] and spinal and bulbar muscular atrophy [5]. Since then, STR expansions have been implicated in dozens of disorders [6]. Further work has shown that these expansions induce a variety of pathogenic effects (Figure 1), including polyglutamine aggregation [6], hypermethylation [7], RNA toxicity [8], and repeat associated non-ATG (RAN) translation [9]. Smaller pathogenic repeats have also been shown to affect RNA splicing (cystic fibrosis [10]) or regulate gene expression (progressive myoclonus epilepsy [11] and Gilbert syndrome [12]). Many of these mechanisms are present across multiple loci, indicating that they likely represent genome-wide phenomena.

The majority of repeat disorders identified so far follow autosomal dominant inheritance patterns that were readily identified using linkage analysis in pedigrees. However, STRs may contribute to a variety of inheritance modes not amenable to traditional linkage techniques. For instance, STRs are predicted to contribute a higher number of *de novo* mutations per generation than any other type of variation [13], but the role of *de novo* STRs in spontaneous conditions such as autism and neurodevelopmental disorders has so far not been interrogated. Furthermore, STRs are often highly multi-allelic, and thus may generate complex inheritance patterns not well captured by linkage or analysis of bi-allelic single nucleotide polymorphisms (SNPs).

Despite the clear implication of STRs in disease, they have been notably missing from medical sequencing studies. Next-generation sequencing (NGS) has the potential to profile more than a million STRs, but genotyping STRs from NGS has proven challenging. Thus, STRs are often filtered from sequencing pipelines due to their low quality calls [14,15], and even known pathogenic STR mutations cannot be detected in most cases [16]. The intronic GGGGCC repeat implicated in FTD and ALS [8,17] identified via a combination of linkage and NGS is the only exception known to the author. Notably, this repeat was not identified through repeat-aware genotyping methods, but rather through anomalies in coverage and a cluster of erroneous SNPs resulting from poor sequence alignment at the expanded repeat.

New bioinformatics tools and advances in sequencing technologies are beginning to overcome these challenges and are providing the first genome-wide portrait of STR variation at a population scale. Here, I review advances over the last several years in STR profiling and how these are leading to an improved understanding of the role of STRs in human traits. Finally, I comment on remaining challenges in analyzing low complexity regions of the genome and prospects of emerging long read technologies to help overcome these hurdles.

Figure 1



Mechanisms by which STRs affect phenotypes.

A schematic view of known and proposed mechanisms by which STRs might regulate gene expression and function. Top: from left to right, STRs may form transcription factor binding sites [12,36], affect spacing between regulatory elements [38], induce unusual DNA secondary structures such as Z-DNA [61], or modulate epigenetic properties such as DNA methylation [62] and heterochromatinzation [63]. Bottom: from left to right, STRs may mediate effects at the RNA and protein level by modulating alternative splicing through RNA secondary structure [10], affecting RNA protein binding sites [64], or forming toxic RNA and protein aggregates [9]. (Purple boxes = genes; black lines = DNA; red boxes = STRs; blue circles = DNA/RNA binding proteins; gray circles = amino acids; green circles = DNA modifications).

Genotyping STRs from high-throughput sequencing data

STRs are challenging to genotype from NGS (Figure 2). First, short reads often do not span entire repeats, effectively reducing the number of informative reads. Second, STR variations present as large insertions or deletions that may be difficult to align to a reference genome, and thus introduce significant mapping bias toward shorter alleles. Finally, PCR amplification during library preparation often introduces “stutter” noise in the number of repeats at STRs.

A variety of bioinformatic methods have been developed to overcome these challenges, many of which are summarized in Table 1. Some use custom alignment techniques to avoid mapping biases imposed by standard short read aligners. One example, lobSTR, [18] rapidly detects reads with a repetitive signature using an entropy metric. It then aligns only non-repetitive flanking regions of the read to the reference genome and employs a model of STR stutter errors to determine the maximum likelihood genotype at each locus. STR-FM [19] uses a similar technique, with an improved detection method based on string matching that shows higher sensitivity to pick up shorter repeats and homopolymer runs.

Other tools, such as Repeatseq [20], save substantial run time by using pre-existing alignments from indel-tolerant aligners such as BWA [21]. This approach is often more sensitive, but is highly affected by the quality of the upstream alignments and may be strongly biased toward shorter alleles if the aligner cannot identify large insertions or deletions. The updated BWA-MEM [22] algorithm exhibits higher sensitivity to larger indels, eliminating much of this bias. A new generation of STR genotyping tools uses BWA-MEM alignments as input combined with improved error models to obtain greater genotyping accuracy. popSTR [23•] uses population information to train locus and individual-specific error profiles. HipSTR [24•] uses a repeat-aware Hidden Markov Model to realign reads, trains locus-specific stutter models, and uses flanking SNPs to physically phase STRs. These methods are more computationally expensive, but show more than 97% accuracy on high coverage data against gold standard techniques.