ARTICLE

Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes

Haibao Tang,[1] Ewen F. Kirkness,[2] Christoph Lippert,[1] William H. Biggs,[2] Martin Fabani,[2] Ernesto Guzman,[2] Smriti Ramakrishnan,[1] Victor Lavrenko,[1] Boyko Kakaradov,[2] Claire Hou,[2] Barry Hicks,[1] David Heckerman,[1] Franz J. Och,[1] C. Thomas Caskey,[3] J. Craig Venter,[2,*] and Amalio Telenti[2,*]

Short tandem repeats (STRs) are hyper-mutable sequences in the human genome. They are often used in forensics and population genetics and are also the underlying cause of many genetic diseases. There are challenges associated with accurately determining the length polymorphism of STR loci in the genome by next-generation sequencing (NGS). In particular, accurate detection of pathological STR expansion is limited by the sequence read length during whole-genome analysis. We developed TREDPARSE, a software package that incorporates various cues from read alignment and paired-end distance distribution, as well as a sequence stutter model, in a probabilistic framework to infer repeat sizes for genetic loci, and we used this software to infer repeat sizes for 30 known disease loci. Using simulated data, we show that TREDPARSE outperforms other available software. We sampled the full genome sequences of 12,632 individuals to an average read depth of approximately 30× to 40× with Illumina HiSeq X. We identified 138 individuals with risk alleles at 15 STR disease loci. We validated a representative subset of the samples (n = 19) by Sanger and by Oxford Nanopore sequencing. Additionally, we validated the STR calls against known allele sizes in a set of GeT-RM reference cell-line materials (n = 6). Several STR loci that are entirely guanine or cytosines (G or C) have insufficient read evidence for inference and therefore could not be assayed precisely by TREDPARSE. TREDPARSE extends the limit of STR size detection beyond the physical sequence read length. This extension is critical because many of the disease risk cutoffs are close to or beyond the short sequence read length of 100 to 150 bases.

Introduction

Microsatellites, or short tandem repeats (STRs), are stretches of simple nucleotide repetitions in the genome; typical repeat units are 1–6 bp in length. Short tandem repeats are often polymorphic as a result of strand slippage during DNA replication and are a common source of rare genetic diseases.[1] The mutation rates of STRs are typically on the order of $\sim 10^{-4}$ mutations per generation per site,[2] as compared to point mutation rates, which are on the order of $\sim 10^{-8}$ mutations per generation per site for single-nucleotide variants (SNVs).[3] Because of the higher mutation rate, STRs offer a different level of resolution at which to study kinship and trait variations among individuals.

STRs are currently used in forensics to identify suspects from DNA traces left at a crime scene. The amplification targets the 13 CODIS (Combined DNA Index System)[4] STR loci, and the sizes of the amplicons are analyzed by electroporesis. The repeat number at each loci is inferred from the size of the amplicon, and a DNA profile is generated. STRs also have a role in revealing genealogy. For example, STR loci on the Y chromosomes (Y-STRs) are used for defining haplotypes that predated the use of Y-SNPs. The STR data, coupled with public genealogy databases such as Y-search, can be used for "surname inference."[5]

STRs have been shown to be involved in several human genetic diseases.[6] Several neural-degenerative disorders, known as the "polyglutamine" (PolyQ) diseases, are caused by variable stretches of the repeated trinucleotide CAG within protein-coding exons. Examples of PolyQ diseases are Huntington disease (HD [MIM: 143100]) and several forms of spinocerebellar ataxia (SCA). Huntington disease is caused by an expansion of the CAG repeats in the first exon of the Huntingtin gene (*HTT* [MIM: 613004]). Individuals carrying an expanded allele have motor, cognitive, and psychological symptoms that typically appear at the age of 40 years old or older, depending on the number of repeats.

STRs also occur in non-coding regions and can regulate gene expression and histone modifications, affecting the expression of nearby genes in *cis* to the STR sites.[7] Examples of these repeat disorders include Myotonic dystrophy (DM1 [MIM: 160900]), caused by CTG repeats; Friedreich Ataxia (FRDA [MIM: 229300]), caused by GAA repeats; and Fragile X syndrome (FXS [MIM: 300624]), caused by CGG repeats. STRs that regulate gene expression (e-STRs) are mostly enriched in genes responsible for cognitive functions and autoimmune responses.[8]

Whole-genome-scale analysis of human STR variation in the presumably healthy 1000 Genomes Project individuals suggests potential contributions of STRs to more complex traits.[8] Because most of these diseases are in the form of tri-nucleotide repeats, they are termed tri-nucleotide-repeat diseases (TREDs). Additionally, STR mutations are known to be associated with susceptibility to cancer.[9] Microsatellite instability is also a well-known hypermutability event that results from impaired DNA

[1]Human Longevity, Mountain View, CA 94041, USA; [2]Human Longevity, San Diego, CA 92121, USA; [3]Baylor College of Medicine, Houston, TX 77030, USA
*Correspondence: jcventer@jcvi.org (J.C.V.), atelenti@humanlongevity.com (A.T.)
https://doi.org/10.1016/j.ajhg.2017.09.013.

mismatch repair, as often occurs in colorectal and gastric cancer.[10]

Most STR loci are not usually included in routine analyses of the genetics of traits and diseases. Consequently, STRs might contribute to the "missing heritability" of complex diseases and traits.[11] One critical bottleneck in assaying STR loci by conventional experimental approaches is a dependency on targeting predefined sites for amplification and subsequently directly sequencing or measuring the size of the amplicons. Large-scale experimental assays are limited because it is difficult to measure multiplex loci within the same assay. Loci might also fail to amplify if they are highly expanded or if there are additional variants on the flanking regions. With whole-genome shotgun sequencing (WGS), it is now possible to type many STR loci on the basis of a single comprehensive sequencing run without the need to design separate genetic assays.

Using WGS, one can use the sequencing reads that map to STR loci to predict allele lengths. Variant calling software (for example,[12,13]) can identify some short indels in reads that span STRs. Other tools seek to identify STR variants by specifically examining the sequencing reads that are piled around a target STR region.[14,15] A popular caller, lobSTR, uses three separate steps: sensing, alignment, and allelotyping, which explicitly model two possible alleles (diploid) as well as sequencing errors typically associated with STRs (because of stutter noise).[15] However, lobSTR only considers reads that fully span an STR locus. Owing to the short length of Illumina reads (100–150 bases), this imposes a major limitation on the length of STR alleles that can be identified. A more recently developed STR caller, ExpansionHunter, incorporates additional evidence beyond spanning reads.[16]

It is also possible to estimate length variation at an STR by combining information from a prior estimate and the observed sizes of paired-end sequence fragments spanning the STR, as in STRViper.[14] However, STRViper assumes a single allele at each site, and this approach is inadequate for diploid human calls. Using long sequence reads such as PacBio[17] or Oxford Nanopore (ONP)[18,19] could potentially help to increase both the precision and the range of detectable variants.[20,21] However, the per-base cost of the long-read technologies is greater than for short-read technologies for whole-genome sequencing, limiting its widespread use for typing STRs. Indeed, few human genomes have been sequenced with PacBio or ONP because of the prohibitive costs associated with long-read sequencing platforms.

Despite recent progress,[14,15,20] high-throughput genotyping of STRs remains limited as a result of low effective coverage, sequencing stutters, and a lack of robust models with which to perform both haploid and diploid calls while distinguishing true variation from technical artifacts.[11] We built TREDPARSE to assess multiple sequence signatures suggested or implemented by previous methods.[14,15,20] Our testing on both simulated datasets, and more than 10,000 sequenced full human genomes,[22] demonstrates that TREDPARSE yields highly accurate typing of many disease-related STRs.

Material and Methods

TREDPARSE Overview

The goal of TREDPARSE is to identify each allele length at predefined STR loci by using Illumina WGS sequence data that are sampled at sufficient depth (discussion on the sequencing depth is provided in later sections). Given a set of observed reads that are mapped around a particular STR locus, our goal is to estimate up to two haplotypes h_1 and h_2, where $1 \leq h_1 \leq h_2 \leq h_{max}$, that represent the number of an individual's repeat units that maximize the likelihood in our model.

The TREDPARSE workflow involves a number of key steps—ploidy inference for a given locus, realignment of reads near the STR region, classification of the reads into four key types of evidence, and the deployment of a full probabilistic framework (Figure 1). The input for the workflow is typically a BAM file that contains mapped WGS reads, and the output is the maximum-likelihood size estimates, distributrions over the number of repeats, and the associated probability of having each of the 30 STR-related diseases. The full probabilistic model is partitioned

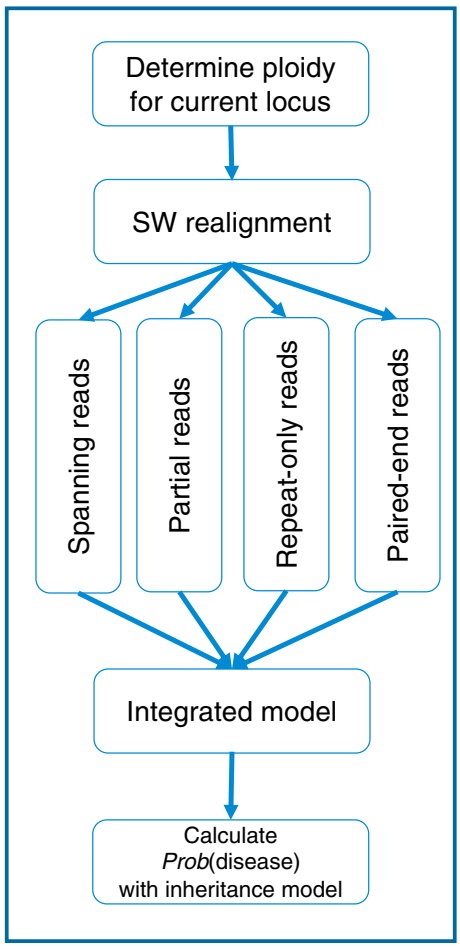

Figure 1. TREDPARSE Workflow for Calling STR-Related Genetic Disease
The workflow includes ploidy inference, read realignment, and integration of various types of evidence in a probabilistic model.