Title: Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results

Authors: Rebecca S. Just, Jodi A. Irwin

PII: S1872-4973(18)30110-8
DOI: https://doi.org/10.1016/j.fsigen.2018.02.016
Reference: FSIGEN 1857

To appear in: *Forensic Science International: Genetics*

Received date: 9-11-2017
Revised date: 2-2-2018
Accepted date: 18-2-2018

Please cite this article as: Rebecca S.Just, Jodi A.Irwin, Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results, Forensic Science International: Genetics https://doi.org/10.1016/j.fsigen.2018.02.016

Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results

Rebecca S. Just[a]*, Jodi A. Irwin[a]

[a]*DNA Support Unit, Federal Bureau of Investigation Laboratory, 2501 Investigation Parkway, Quantico, VA 22135, USA*

*Corresponding author. Tel: +1 (703) 632-8230; Email: rsjust@fbi.gov

rsjust@fbi.gov
jairwin@fbi.gov

Highlights

- Use of LUS length in allele designations captures >80% of aSTR sequence variation
- LUS length reference regions are straightforward to define
- LUS-based alleles can be interpreted in some existing probabilistic programs
- Use of sequence information increases LRs for DNA donors as compared to repeat unit
- LUS-based designations can accommodate multiple stutter products by sequence

Abstract

Some of the expected advantages of next generation sequencing (NGS) for short tandem repeat (STR) typing include enhanced mixture detection and genotype resolution via sequence variation among non-homologous alleles of the same length. However, at the same time that NGS methods for forensic DNA typing have advanced in recent years, many caseworking laboratories have implemented or are transitioning to probabilistic genotyping to assist the interpretation of complex autosomal STR typing results. Current probabilistic software programs are designed for length-based data, and were not intended to accommodate sequence strings as the product input. Yet to leverage the benefits of NGS for enhanced genotyping and mixture deconvolution, the sequence variation among same-length products must be utilized in some form.

Here, we propose use of the longest uninterrupted stretch (LUS) in allele designations as a simple method to represent sequence variation within the STR repeat regions and facilitate - in the near-term - probabilistic interpretation of NGS-based typing results. An examination of published population data indicated that a reference LUS region is straightforward to define for most autosomal STR loci, and that using repeat unit plus LUS length as the allele designator can represent greater than 80% of the alleles detected by sequencing. A proof of concept study performed using a freely available probabilistic software demonstrated that the LUS length can be used in allele designations when a program does not require alleles to be integers, and that utilizing sequence information improves interpretation of both single-source and mixed contributor STR typing results as compared to using repeat unit information alone. The LUS concept for allele designation maintains the repeat-based allele nomenclature that will permit backward compatibility to extant STR databases, and the LUS lengths themselves will be concordant regardless of the NGS assay or analysis tools employed. Further, these biologically-based,