B46 lusSTR®: An Open-Source Python® Package for Translating Forensic Short Tandem Repeat (STR) Sequences Into Interpretable Formats

Rebecca Mitchell, PhD, NBACC, Fort Detrick, MD 21702; Daniel S. Standage, PhD, NBACC, Fort Detrick, MD 21702; Rebecca Just, PhD, NBACC, Fort Detrick, MD 21702*

Learning Overview: After attending this presentation, attendees will understand the capabilities of lusSTR®, a new software program for translating human DNA sequences of STR loci into more user-friendly representations.

Impact on the Forensic Science Community: lusSTR® implements an automated method for converting human DNA sequences of STR loci derived from Next Generation Sequencing (NGS) into more intelligible annotation. Laboratories can utilize the lusSTR® program to assist validation of STR sequencing protocols, to facilitate interpretation of STR sequence data, and to more easily compare STR sequences produced by different NGS assays and analysis packages.

While recent advances in human forensic DNA technology have provided significant improvement in identification capabilities, they have also created additional challenges in analyzing and interpreting the resulting data. The ability of DNA sequencing to detect nucleotide-level variation at conventional forensic STR loci has revealed further individualization capabilities beyond the traditional length-based allele designation. However, the long sequence strings that result from NGS typing of STR loci are difficult for analysts to interpret in a consistent manner.

Currently, there is no standard for translating STR sequences into more human-readable forms, nor is there community consensus regarding what those forms should be or the portions of the amplified STR region they should include (e.g., repeat region only or including flanking sequence). In addition, depending on the NGS assay and analysis program employed, sequences for some loci may be reported according to previously acceptable reverse strand (3' to 5') alignment rather than the GRCh38-consistent forward strand reporting recommended by an International Society for Forensic Genetics (ISFG) DNA commission.[1] As a result, forensic laboratories investigating and validating NGS technologies for sequencing the STR loci face challenges in analyzing the sequence data produced in their internal studies, as well as difficulties comparing results across NGS assays, analysis programs, and STR datasets. The use of reference tables for sequence translation, while straightforward, requires that an exact match to the sequence be present in the reference set. Sequences not present in the table may require manual handling, which can produce inconsistent and error-prone annotation of these complex strings. An automated and flexible method for sequence string translation ensures consistent annotation and facilitates analysis of STR sequences both internal to a lab and across laboratories and datasets.

The open-source software program lusSTR® was developed to convert STR sequence strings to several different representations to allow for easier and more reliable comparisons between sequences. lusSTR® accommodates the autosomal, Y- and X-chromosome STR sequences produced using two commercial NGS assays: the ForenSeq™ Signature Prep kit and PowerSeq™ 46GY and uses as input either: a standard report produced from one of two STR alignment packages (ForenSeq™ UAS and STRait Razor v3) or a simple .csv file of sequences.[2] lusSTR® partitions sequences into core and flanking regions, reverse complementing as necessary, and produces an output file containing multiple locus-specific annotations per sequence.

This presentation will describe the lusSTR® program, detail the development and testing of lusSTR®, and depict how lusSTR® functions and may be used in the forensic setting. The multiple annotation types produced by lusSTR®, derived by an automated method, will provide analysts a more straightforward and reliable way to evaluate and compare forensic STR sequences.

This work was funded under Contract No. HSHQDC-15-C-00064, awarded by the Department of Homeland Security Science and Technology (DHS S&T) to the National Biodefense Analysis and Countermeasures Center (NBACC), a DHS federal laboratory operated by the Battelle National Biodefense Institute (BNBI). Views and conclusions contained herein are those of the authors and should not be interpreted to represent policies, expressed or implied, of the DHS or S&T.

Reference(s):

1. Parson, Walther, David Ballard, Bruce Budowle, John M. Butler, Katherine B. Gettings, Peter Gill, Leonor Gusmao, et al. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Science International: Genetics* 22 (2016). 54-63.
2. Woerner, August E., Jonathan L. King, and Bruce Budowle. Fast STR allele identification with STRait Razor 3.0. *Forensic Science International: Genetics* 30 (2017). 18-23.

Next Generation Sequencing, Short Tandem Repeats, Bioinformatics

*Presenting Author