Forensic Science International: Genetics 45 (2020) 102225

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Characterizing stutter variants in forensic STRs with massively parallel sequencing

Ran Li[a,b,1], Riga Wu[a,b,1], Haixia Li[a,b], Yinming Zhang[a,b], Dan Peng[a,b], Nana Wang[a,b], Xuefeng Shen[a,b], Zhiyuan Wang[c], Hongyu Sun[a,b,*]

[a] Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, Guangdong, People's Republic of China
[b] Guangdong Province Translational Forensic Medicine Engineering Technology Research Center, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510089, Guangdong, People's Republic of China
[c] Forensic Science Center of Foshan Municipal Public Security Department, Foshan 528000, Guangdong, People's Republic of China

ARTICLE INFO

Keywords:
Massively parallel sequencing (MPS)
Short tandem repeat (STR)
Stutter

ABSTRACT

Despite improvements in characterizing stutters of short tandem repeats (STRs), the relationships among the amounts of stutter variants and the relationships among motifs are not well understood yet. In the present study, 750 peripheral blood samples from human subjects were included to characterize the stutters of 58 STRs via the ForenSeq DNA Signature Prep Kit on a MiSeq FGx instrument. Alleles and corresponding stutter products were identified with a sequence simplification procedure. After screening, 26,921 alleles were included, that resulted in over 50 million reads, among which 8.69% were stutter products. Among these stutter products, 83.44% were N-1 stutters. Additionally, N-4, N-3, N-2, N0, N + 1, and N + 2 variants accounted for 0.11%, 0.77%, 6.45%, 3.01%, 5.95%, and 0.25% of the stutter products, respectively. For backward stutters, stutter products correlated best with the corresponding one-unit-longer stutter (or parental allele), which may represent a good predictor for backward stutters. For forward stutters, the N + 2 stutter correlated best with the N + 1 stutter, whereas the N + 1 stutter correlated best with the N-1 stutter rather than the expected parental allele, which indicated that the patterns were more complex for forward stutters. Additionally, some interesting findings were obtained for D21S11. For two adjacent contiguous motifs, co-stuttering patterns were observed where one motif tended to increase one repeat unit while the other motif decreased one repeat unit, whereas the inter-motif dependency was not significant for interrupted motifs. In conclusion, with massively parallel sequencing technology and our sequence simplification strategy, sequence variations within alleles and stutter products were identified, which was useful to determine the origin of stutters, identify more stutter variants, and explore the relationships among motifs. These findings may be helpful for allele designation, a deeper understanding of the mechanism of stutter, and improving resolution in forensic mixture analyses.

1. Introduction

Short tandem repeats (STRs) are prevalent genetic markers in forensic genetics due to their high degree of repeat-number polymorphisms in the human population [1,2]. High levels of discriminations can be achieved for forensic purposes, such as individual identification and paternity testing, by genotyping several to dozens of STR markers via capillary electrophoresis (CE). However, unexpected signals and/or artifacts frequently occur. A common and well-known artifact is stutter, which is presumed to be the result of slipped strand mispairing (SSM) during the polymerase chain reaction (PCR) [1,3,4]. Generally, the template strand "loops out" and results in a new strand one repeat unit shorter (N-1) than the parental alleles (PAs). Stutter variants of two or more repeat units smaller or one unit larger (N + 1) have also been reported [5–7]. Sometimes, stutter products have the same length as the actual alleles, which further complicates the interpretation of mixture profiles7–10]. Therefore, it is important to accurately characterize, predict, and filter this kind of artifact.

Usually, a threshold is set for the differentiation between stutter products and real alleles. This threshold is critically important; if it is too low, stutter products can be falsely identified as alleles, whereas if it is too high, a true allele of minor contributors can be lost [3]. Additionally, since different sequences stutter differently, locus-specific thresholds are preferable. Kalafut et al. introduced an allele-specific

* Corresponding author at: No. 74 Zhongshan 2nd Road, 510089, Guangzhou, People's Republic of China.
E-mail addresses: sunhy@mail.sysu.edu.cn, sunhongyu2002@163.com (H. Sun).
[1] These authors contributed equally to the article.

https://doi.org/10.1016/j.fsigen.2019.102225
Received 8 April 2019; Received in revised form 19 November 2019; Accepted 8 December 2019
Available online 09 December 2019
1872-4973/ © 2019 Elsevier B.V. All rights reserved.

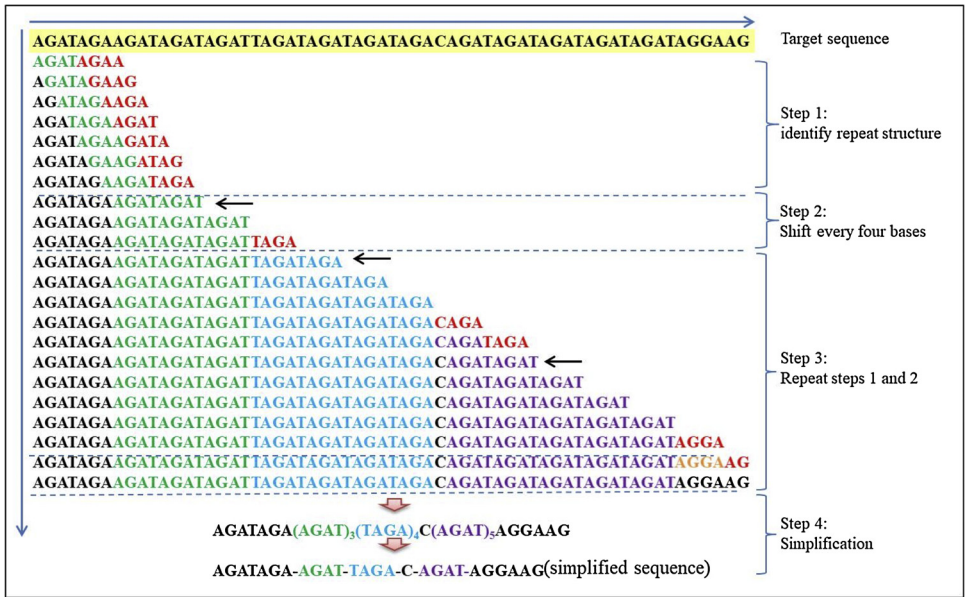

Fig. 1. Illustration of sequence simplification. In step 1, the first eight bases were split (by two-fold of the assumed period) and were shifted base by base until a repeat structure was detected (black arrow). In step 2, we shifted every four bases and judged whether these shifts were the same as the present motif. This step was completed once different sequences were identified (red sequences). In step 3, the procedures in step 1 and step 2 were repeated until the end of the sequence. Finally, in step 4, we retained non-repeat sequences (black sequences) and motifs (color), obtaining the simplified sequence.

stutter model that reduced both over- and under-filtering rates compared with those yielded from traditional locus-specific models [3]. However, these studies were based on a CE method and, thus, had some limitations. First, as only information regarding allele length is obtained via a CE method, it is often difficult to determine the origin and proportion of some stutter products. Additionally, small signals derived from less frequently occurring stutters are often masked by background noise. To identify such small stutter signals, a lower analytic threshold of 25 RFU or 35 RFU has been applied [5,7], which concomitantly increases the risk of including spurious noise.

Massively parallel sequencing (MPS) technology provides new possibilities to resolve the inherent limitations of traditional CE methodology. Since more detailed sequence information of alleles and stutter products can be obtained, MPS may help to better differentiate stutter variants, especially for variants with identical lengths but different sequences. Therefore, MPS is expected to yield a higher resolution for stutter analysis. For example, it has been reported that stutters could be identified via MPS with coverage as low as only one read [11]. Recently, a new predictor of stuttering—the block length of the missing motif (BLMM)—has been introduced, with which the mean square error decreased by a factor of up to 17.5 for compound and complex autosomal STR markers [11]. Woerner et al. has shown that flanking variations also influence the rates of stutter products in simple repeats [8] and they found that there was a lack of independence between stutter products in compound STRs [9]. These studies indicate that—in addition to the longest uninterrupted sequence (LUS), A-T content, PCR cycle number, the temperature of annealing/extension, and DNA polymerase processivity [1,10,12,13]—many more factors are associated with stutter products.

Despite these improvements, the relationships among the amounts of stutter variants and the relationships among motifs both remain poorly understood. For sequences with multi-motifs (compound or complex sequences), alleles may generate different stutter products corresponding to different parts (motifs) of the parental alleles. Does each motif stutter in the same way? How do the motifs of compound and complex STRs influence each other? Is there any relationship between stutter variants? To answer these questions, further analyses were conducted in the present study based on a large sample size using MPS technology.

2. Materials and methods

2.1. Sample collection

With informed consent, peripheral blood samples from 750 individuals were collected and dried on an FTA card at room temperature. Additionally, 2800 M control DNA (Promega Corporation, Madison WI) was used as a positive control. This study was approved by the Human Subjects Committee of Sun Yat-sen University.

2.2. Library preparation and sequencing

DNA libraries were constructed using the ForenSeq DNA Signature Prep Kit (Illumina Inc., San Diego CA) following the manufacturer's instructions, with which 27 autosomal STRs, 24 Y-STRs, and seven X-STRs were co-amplified. Sequencing was performed on a MiSeq FGx™ instrument (Illumina Inc., San Diego CA) via the MiSeq FGx Reagent Kit (Illumina Inc., San Diego CA), according to the manufacturer's instructions.

2.3. Data analysis

2.3.1. Genotype calling and allele screening

STRait Razor 3.0 [14] was employed for locus-specific sequence extraction and a text file was obtained for each sample that contained the locus name, length-based allele, sequences and the corresponding reads. A minimum depth of 50× and a heterozygote ratio of 0.4 were applied for genotype calling. To assign stutter products to the alleles that generated them for heterozygotes, the following filtration strategy was used: (1) The sequences of both alleles were abbreviated according to the repeat structure; (2) Repeat number was discarded, resulting in strings consisting of the flanking sequence and repeat unit only (i.e. the simplified sequence); (3) Heterozygotes with the same simplified sequences were excluded for further analysis (both alleles). This procedure is illustrated in Fig. 1 and was carried out using in-house scripts written via Visual-Basic-based software. In our present study, the allele structure was defined differently from that recommended by Parson et al. [15], such that all motifs were recognized within the flanking and routine core repeat region if they had ≥ 2 repeats. Each motif was determined according to the repeat structure from the beginning. Hence, different motif components could be obtained for one locus. For example, three alleles at CSF1PO—namely Allele 1 (CE 12) as

Download English Version:

https://daneshyari.com/en/article/13408623

Download Persian Version:

https://daneshyari.com/article/13408623

Daneshyari.com