Sequence analysis of 25 autosomal STRs including SE33 using in-house MPS panel in four populations

Ye-Lim Kwon[1,2] · Bo Min Kim[1,2] · Eun Young Lee[1] · Kyoung-Jin Shin[1,2]

[1]Department of Forensic Medicine, Yonsei University College of Medicine, Seoul, Korea
[2]Brain Korea 21 PLUS Project for Medical Science, Yonsei University, Seoul, Korea

Introduction

The diversity of autosomal STRs can be increased by identifying sequence variation using massively parallel sequencing (MPS), which is applicable to the analysis of degraded and mixture DNA. In order to apply MPS analysis to forensic caseworks, it is important to compile sequence data to calculate statistics. Particularly, SE33 is known as highly polymorphic STR locus, but a few sequence-based data have been reported. Therefore, we developed MPS panel simultaneously amplified the 28 markers, consisted of 20 expanded CODIS core loci, five additional autosomal STR loci (D4S2408, D6S1043, Penta E, Penta D and SE33) and three sex typing markers (Amelogenin, DYS391 and Y-M175). Using this MPS panel and conventional capillary electrophoresis (CE) kits, we analyzed sequence variation of the 25 autosomal STRs for 350 samples across four populations (African Americans, Caucasians, Hispanics and Koreans). Here, we show genotype concordance between two methods and the gains of MPS method by comparing the number of alleles obtained by CE (length-based) and MPS (sequence-based) analysis. Especially, we present the sequence analysis results on SE33 and their usefulness in forensic practice.

Materials and Methods

Samples

- A total of 350 unrelated samples used in this study were consisted of four populations: African American (AfAm, N=83), Caucasian (Cauc, N=82), Hispanic (Hisp, N=82), Korean (Kor, N=103).

MPS Library Preparation and Run

- The MPS library was generated using the developed in-house MPS panel which simultaneously amplified the 25 autosomal STRs including SE33 by two-step PCR with 1 ng of genomic DNA as template.
- The MPS libraries were purified with 1.2 x Agencourt® AMPure® XP beads (Beckman Coulter) and quantified using KAPA library quantification kits (KAPA Biosystems) and 2100 Bioanalyzer (Agilent).
- The barcoded libraries were normalized to 10nM and then pooled in equal volumes. The pooled library was sequenced on the MiSeq System (Illumina).

MPS Data Analysis

- FASTQ files were analyzed using the bioinformatic pipeline STRait Razor v3.0. We confirmed the genotype concordance between CE and MPS method and investigated sequence variation.

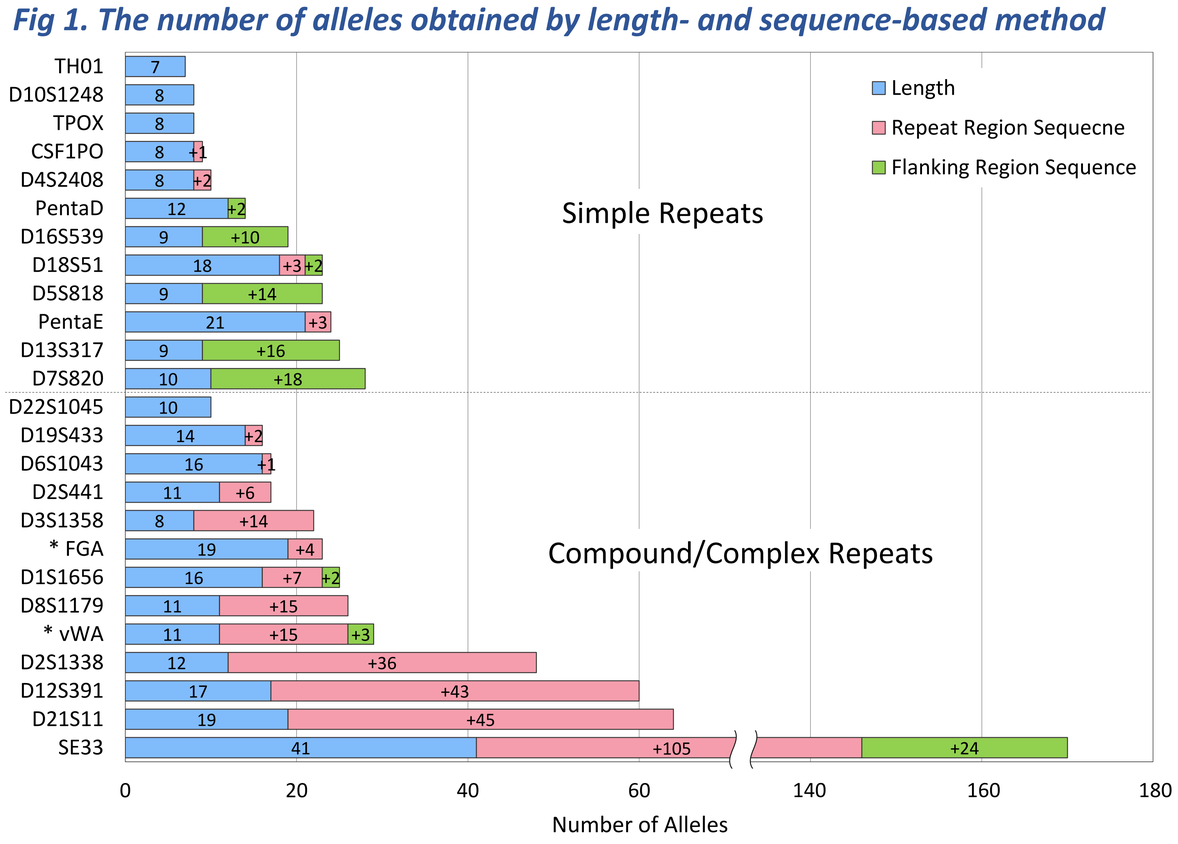
Results

Amplicon size and MPS coverage

- The amplicon size of SE33 was smaller than 258bp, and the others were less than 220bp.
- The depth of coverage (DoC) for each marker were calculated by read count obtained by STRait Razor. The minimum and maximum coverage were observed in Penta D and D1S1656, respectively. The difference of DoC was in less than twice.

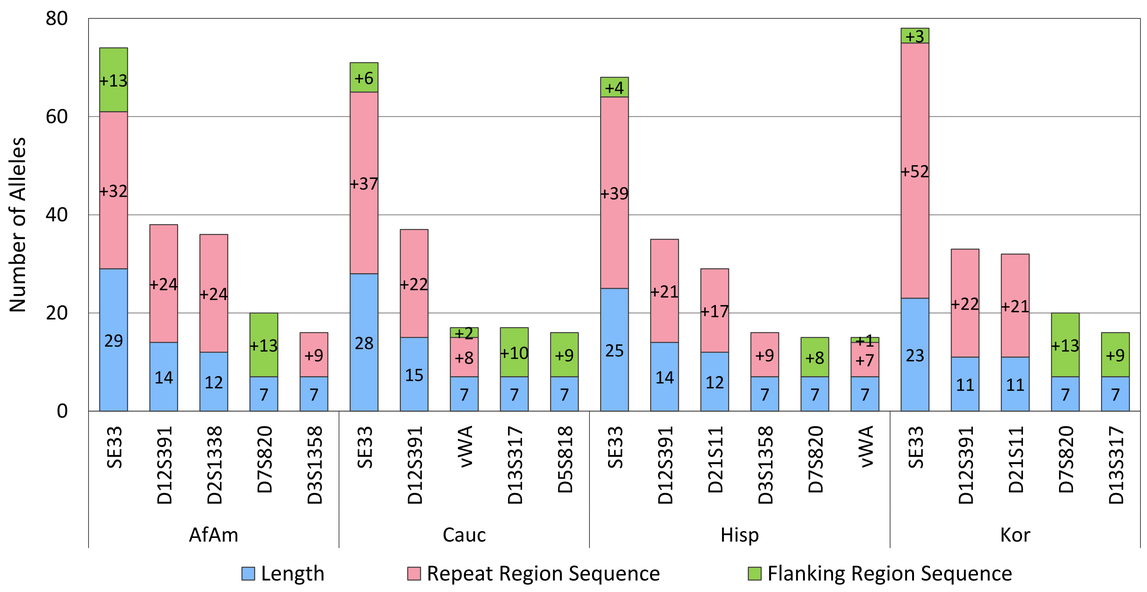
The gains of MPS method

Fig 1. The number of alleles obtained by length- and sequence-based method



Repeat region variations are mainly observed in D21S11, D12S391, and D2S1338. Most of flanking region variations are observed at D7S820, D13S317, D5S818 and D16S539. In particular, SE33 has a large number of repeat and flanking region variation. (The markers for which dropped out allele was observed are FGA and vWA and marked with *.)

Fig 2. The number of alleles obtained by length- and sequence-based method for each population

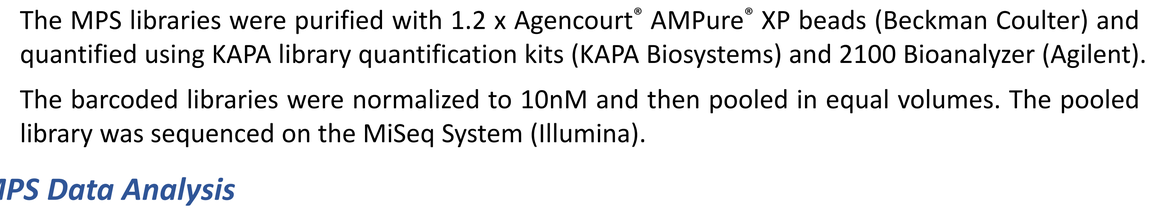


Of 25 autosomal STRs, only the top five markers with the highest rate of increase are selected in this figure. The SE33 markers across all populations showed the greatest increase in observed allele numbers with 2.5-fold or more.

CE and MPS analysis of SE33

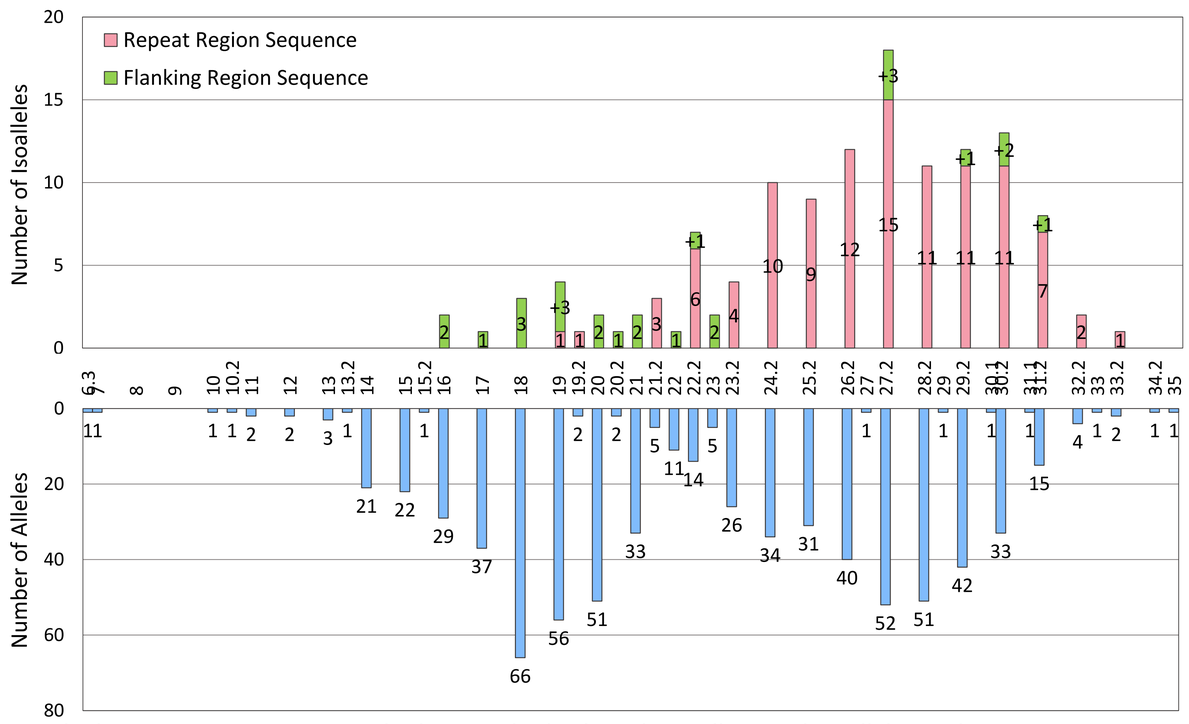
- By identifying the sequence variation, additional 129 unique alleles were detected, which shows 4.15-fold increase in alleles. Also, SE33 showed 97.3% heterozygosity by sequence-based analysis.
- In African American, eight samples (nine alleles , 0.08%) showed genotype discrepancy between length- and sequence-based method. Five alleles showed [TCTT/-] deletion (rs1371483225) and four alleles showed [TTTT/-] deletion (rs369314007). (See *Fig 3.* detailed)

Fig 3. Illustration of deletion position that causes genotype discordance between length- and sequence-based method in SE33



The forward primer of SE33 shares with each other on CE and MPS panels. The MPS reverse primer was positioned in upper stream to rs1371483225 and rs369314007 deletions, and the CE reverse primer covered both deletions.

Fig 4. Allele distribution pattern of SE33 alleles



Flanking region variation mainly distinguished relatively smaller sized isoalleles without microvariants, while repeat region variation was predominantly observed in larger sized isoalleles with microvariants.

Conclusion

- We successfully analyzed 25 autosomal STRs for 350 samples across four representative populations using the developed in-house MPS panel.
- We identified sequence variation which located in repeat and flanking region. SE33 is the most polymorphic locus and showed the largest increase of observed allele by sequence variations.
- Analyzing STRs using the MPS method provides additional sequence information and is useful to increase the power of discrimination in challenging caseworks than CE method.

Acknowledgement

This study was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (NRF-2014M3A9E1069989)