Massively Parallel Sequencing of Global STR Loci in Koreans

Eun Hye Kim[1,2] · Hwan Young Lee[1] · Woo Ick Yang[1] and 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

Because of its capability to produce massively parallel sequencing data and to provide sequence variations, the prospect of NGS in STR typing has been explored by many forensic laboratories using in-house multiplex PCR systems or commercial kits. Meanwhile, the expanded CODIS Core Loci were newly announced to increase the international data compatibility and to maximize the power of forensic DNA databases. Accordingly, we developed a multiplex PCR system for the NGS analysis of 25 forensic markers which include 20 expanded CODIS Core Loci (D1S1656, D2S441, D2S1338, D3S1358, D5S818, CSF1PO, D7S820, D8S1179, D10S1248, D12S391, D13S317, TPOX, D16S539, D18S51, D19S433, D21S11, D22S1045, FGA, TH01 and vWA) and additional 5 Loci (Penta D, Penta E, D6S1043, DYS391 and amelogenin). To investigate the sequence variations of 24 global STR loci, NGS data of 160 unrelated Koreans were analyzed. In addition, we present the observed sequence variations with allele frequencies and forensic parameters in Koreans.

Materials and Methods

1. DNA samples

DNA was extracted from buccal swab samples of 160 unrelated Koreans using a QIAamp DNA Mini Kit (Qiagen) and quantified using a NanoDrop® ND-1000 Spectrophotometer (NanoDrop Technologies). Finally diluted 1 ng/ul of DNA was prepared and used. The study was approved by the Institutional Review Board of Severance Hospital, Yonsei University in Seoul, Korea.

2. Construction of multiplex PCR system

The STR markers selected for a multiplex PCR were composed of 25 forensic markers that included 20 expanded CODIS Core Loci and additional 5 Loci (Penta D, Penta E, D6S1043, DYS391 and amelogenin) shown in Fig. 1. Primers for PCR amplification were designed using the Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) program such that the amplicon sizes of the 25 targeted markers were less than 250 bp, and the primers did not bind to the region with a greater than 1% mutation rate based on the NCBI SNP information (http://www.ncbi.nlm.nih.gov/SNP/).

3. Preparation of NGS library

We conducted two-step PCR amplifications to generate a library using primers with a modification referring to the sequence information of Nextera® system (Illumina). The first PCR targeted the autosomal STR itself, and primer sequences included STR-specific sequences and read sequences. A second PCR was performed to add indices and platform-specific sequences. See detailed information on poster No. 79.

1) The first PCR - multiplex PCR was performed with 29 thermal cycles from each 1 ng of the sample and appropriate concentration of primers.

2) The second PCR - indexing PCR was performed with 15 thermal cycles from each 1.0 µl of 100-fold diluted the first PCR amplicons and Nextera XT Index Kit (Illumina).

Following PCR cleanup with 1.2× Agencourt® AMPure® XP beads (Beckman Coulter), the libraries were quantified using KAPA library quantification kits (KAPA Biosystems) and Agilent 2100 Bioanalyzer.

4. NGS run and data analysis

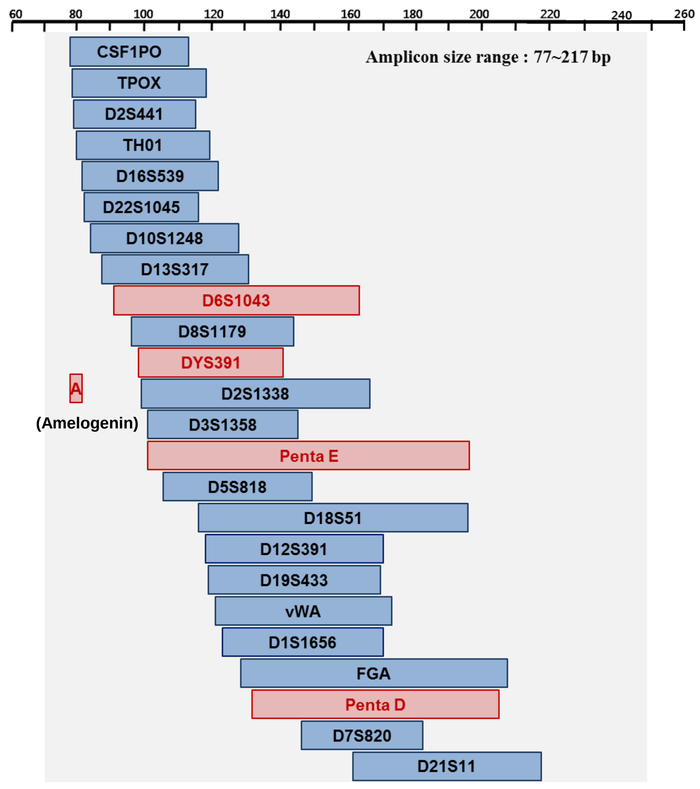
The barcoded libraries were normalized to 10nM and then pooled in equal volumes. Finally, the pooled library was sequenced on MiSeq™ (Illumina) using a MiSeq Reagent Kit v2, 2x250 bp (Illumina). NGS data analysis basically follows the protocol presented by Bornman et al (Biotechniques, 2012). The process of NGS data analysis used in this study was illustrated in Fig. 2. The STR profiles obtained by two CE methods – Kplex-23 and Euplex-13 system (BioQuest) – were used as reference data for comparing the STR typing results from NGS.

5. Statistical analysis

Statistical parameters, including the value of the allele frequencies and observed heterozygosity (H_{obs}), were calculated using counting methods. Forensic efficiency information was assessed by calculating expected heterozygosity (H_{exp}), match probability (MP), power of discrimination (PD), polymorphic information content (PIC) and power of exclusion (PE).

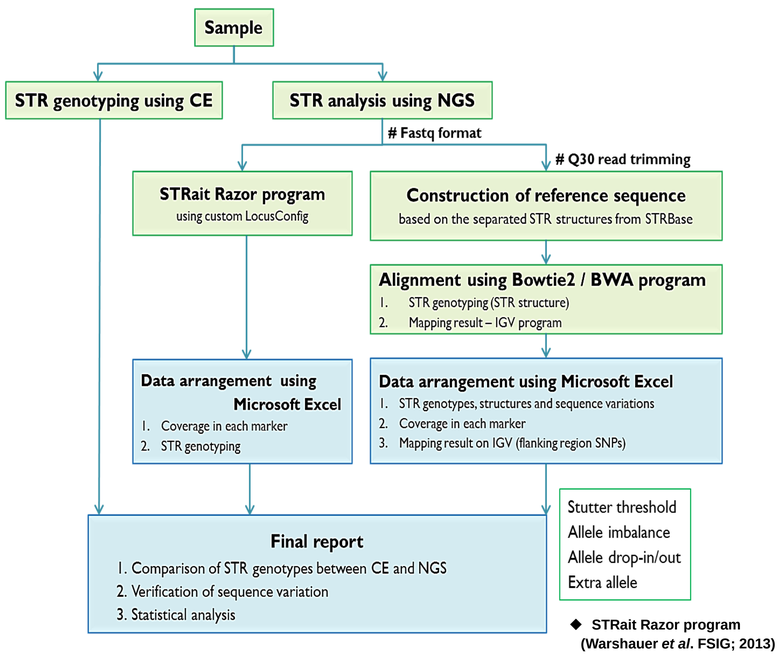
Results

Fig. 1. Allelic size range for 25 forensic markers



Autosomal 20 expanded CODIS Core Loci were marked in blue boxes and additional 5 loci were marked in red boxes. Most markers were less than 200 bp. Those of Penta D and FGA were less than 210 bp. D21S11 showed a maximum size range less than 217 bp.

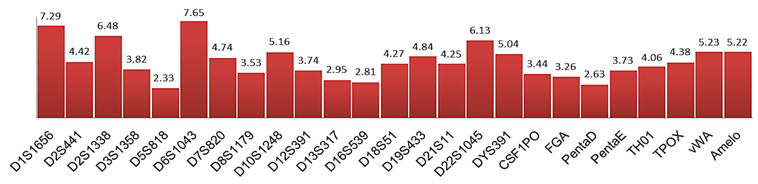
Fig. 2. Pipeline of NGS data analysis



◆ STRait Razor program (Warshauer *et al*. FSIG; 2013)

Basically, the genotypes were compared with CE method and NGS data. The percentage coverage values were determined by dividing an assigned coverage for each allele by the total coverage of the locus. STR alleles could be determined when 20% of total coverage was used as a threshold.

Fig . 3. Read count from NGS data on each marker



The relative reads counts in average coverage by STRait Razor. The minimum and maximum coverage were observed in D5S818 and D6S1043, respectively. The difference was in less than 3 times.

Fig. 4. Comparison of genotype between CE and NGS analysis

STR genotypes of 160 Koreans obtained from NGS were concordant with CE profiles except 14 samples. Fourteen samples were identified that an adenine was inserted at 5' flanking region in 9 allele of D2S441, but shown as 9.1 allele in CE profile.

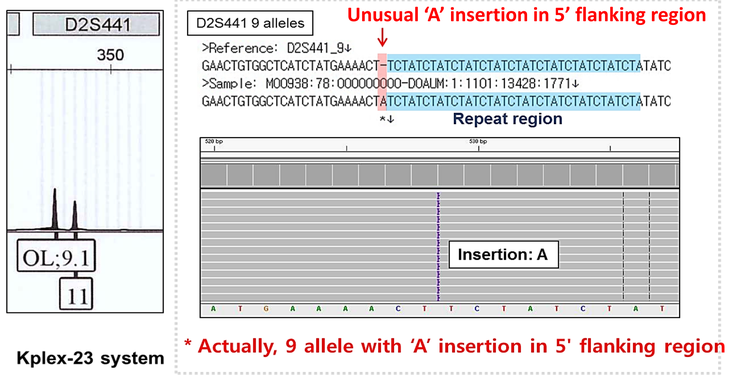


Table 1. Examples of observed sequence variations

1) D12S391

allele	Sub-allele	Structure	Frequency
15	15	$[AGAT]_8$ $[AGAC]_6$ AGAT	0.025
16	16a	$[AGAT]_8$ $[AGAC]_7$ AGAT	0.006
	16b	$[AGAT]_9$ $[AGAC]_6$ AGAT	0.003
17	17	$[AGAT]_{10}$ $[AGAC]_6$ AGAT	0.066
18	18a	$[AGAT]_9$ $[AGAC]_8$ AGAT	0.003
	18b	$[AGAT]_{10}$ $[AGAC]_7$ AGAT	0.025
	18c	$[AGAT]_{11}$ $[AGAC]_6$ AGAT	0.266
	18d	$[AGAT]_{12}$ $[AGAC]_5$ AGAT	0.006
19	19a	$[AGAT]_{11}$ $[AGAC]_7$ AGAT	0.031
	19b	$[AGAT]_{11}$ $[AGAC]_8$	0.003
	19c	$[AGAT]_{12}$ $[AGAC]_6$ AGAT	0.228
	19d	$[AGAT]_{13}$ $[AGAC]_5$ AGAT	0.003
19.2	19.2	$[AGAT]_4$ AT $[AGAT]_7$ $[AGAC]_7$ AGAT	0.003
20	20a	$[AGAT]_{11}$ $[AGAC]_8$ AGAT	0.009
	20b	$[AGAT]_{11}$ $[AGAC]_9$	0.006
	20c	$[AGAT]_{12}$ $[AGAC]_7$ AGAT	0.025
	20d	$[AGAT]_{12}$ $[AGAC]_8$	0.006
	20e	$[AGAT]_{13}$ $[AGAC]_6$ AGAT	0.091
	20f	$[AGAT]_{13}$ $[AGAC]_7$	0.003
21	21a	$[AGAT]_{12}$ $[AGAC]_8$ AGAT	0.047
	21b	$[AGAT]_{12}$ $[AGAC]_9$	0.019
	21c	$[AGAT]_{13}$ $[AGAC]_7$ AGAT	0.016
	21d	$[AGAT]_{13}$ $[AGAC]_8$	0.003
	21e	$[AGAT]_{14}$ $[AGAC]_6$ AGAT	0.031
22	22a	$[AGAT]_{12}$ $[AGAC]_{10}$	0.003
	22b	$[AGAT]_{13}$ $[AGAC]_8$ AGAT	0.016
	22c	$[AGAT]_{13}$ $[AGAC]_9$	0.028
	22d	$[AGAT]_{14}$ $[AGAC]_7$ AGAT	0.006
	22e	$[AGAT]_{15}$ $[AGAC]_6$ AGAT	0.006
23	23a	$[AGAT]_{12}$ $[AGAC]_{10}$ AGAT	0.003
	23b	$[AGAT]_{12}$ $[AGAC]_{11}$	0.003
24	24a	$[AGAT]_{14}$ $[AGAC]_9$ AGAT	0.003
	24b	$[AGAT]_{15}$ $[AGAC]_8$ AGAT	0.003
25	25	$[AGAT]_{15}$ $[AGAC]_9$ AGAT	0.003

2) D2S441

allele	Sub-allele	Structure	Frequency
9	9	$[TCTA]_9$	0.041
10	10a	$[TCTA]_{10}$	0.069
	10b	$[TCTA]_8$ TCTG TCTA	0.147
11	11a	$[TCTA]_{11}$	0.391
	11b	$[TCTA]_9$ TCTG TCTA	0.009
11.3	11.3	$[TCTA]_4$ TCA $[TCTA]_7$	0.016
12	12	$[TCTA]_{12}$	0.169
13	13a	$[TCTA]_{13}$	0.009
	13b	$[TCTA]_{10}$ TTTA $[TCTA]_2$	0.013
14	14a	$[TCTA]_{14}$	0.003
	14b	$[TCTA]_{11}$ TCTG $[TCTA]_2$	0.003
	14c	$[TCTA]_{11}$ TTTA $[TCTA]_2$	0.122
15	15	$[TCTA]_{12}$ TTTA $[TCTA]_2$	0.009

The sequence variations were observed in 12 STRs (D1S1656, D2S441, D2S1338, D3S1358, D6S1043, D7S820, D8S1179, D12S391, D13S317, D21S11, CSF1PO and vWA) of 160 Koreans. D12S391 and D2S441 were presented as examples. And the remaining of sequence variations in 10 STRs will be published in a paper.

Table 2. Comparison between CE and NGS analysis

◆ Comparison of the number of alleles (n=160)

Loci	CE method	NGS method	Fold change
D12S391	12	34	+2.83x
D21S11	13	29	+2.23x
D2S1338	13	29	+2.23x
D8S1179	9	17	+1.89x
D3S1358	8	15	+1.88x
D2S441	8	13	+1.63x
vWA	7	11	+1.57x
D1S1656	13	17	+1.31x
D13S317	8	10	+1.25x
D7S820	7	8	+1.14x
CSF1PO	8	9	+1.13x
D6S1043	12	13	+1.08x

The most variable sequences were observed in D12S391, following by D21S11 and D2S1338.

◆ Comparison of the number of genotypes (n=160)

Loci	CE method	NGS method	Fold change
D21S11	31	82	+2.65x
D3S1358	14	31	+2.21x
D12S391	31	68	+2.19x
D8S1179	31	65	+2.10x
D2S1338	43	80	+1.86x
D2S441	22	32	+1.45x
vWA	22	29	+1.32x
D1S1656	40	50	+1.25x
D13S317	23	25	+1.09x
CSF1PO	17	18	+1.06x
D7S820	21	22	+1.05x
D6S1043	50	51	+1.02x

The most variable genotypes were observed in D21S11, and following by D3S1358.

Table 3. Statistical analysis of forensic parameters in Koreans

Loci	Method	H_{obs}	H_{exp}	MP	PD	PIC	PE
D1S1656	CE	0.831	0.834	0.056	0.944	0.812	0.658
	NGS	0.856	0.846	0.048	0.952	0.827	0.707
D2S441	CE	0.813	0.749	0.108	0.893	0.712	0.622
	NGS	0.825	0.794	0.079	0.921	0.751	0.646
D2S1338	CE	0.844	0.865	0.039	0.961	0.847	0.683
	NGS	0.869	0.907	0.020	0.980	0.896	0.732
D3S1358	CE	0.738	0.712	0.136	0.864	0.658	0.489
	NGS	0.794	0.775	0.080	0.920	0.743	0.587
D6S1043	CE	0.869	0.877	0.033	0.967	0.861	0.732
	NGS	0.869	0.877	0.033	0.967	0.861	0.732
D7S820	CE	0.813	0.769	0.097	0.903	0.730	0.622
	NGS	0.813	0.771	0.095	0.905	0.733	0.622
D8S1179	CE	0.806	0.841	0.047	0.953	0.818	0.611
	NGS	0.881	0.906	0.022	0.978	0.895	0.757
D12S391	CE	0.744	0.801	0.072	0.928	0.772	0.499
	NGS	0.825	0.861	0.037	0.963	0.846	0.646
D13S317	CE	0.781	0.799	0.075	0.925	0.766	0.565
	NGS	0.788	0.801	0.073	0.927	0.769	0.576
D21S11	CE	0.781	0.780	0.080	0.920	0.750	0.565
	NGS	0.919	0.922	0.021	0.979	0.893	0.834
CSF1PO	CE	0.794	0.721	0.146	0.854	0.672	0.587
	NGS	0.794	0.723	0.143	0.857	0.675	0.587
vWA	CE	0.806	0.791	0.080	0.920	0.758	0.611
	NGS	0.806	0.799	0.076	0.924	0.769	0.611

The most significant change of PD between CE and NGS was D21S11 with +0.059. Although the sequence variation was observed in D6S1043, the PD value was not changed.

Conclusion

- We constructed a single tube new multiplex PCR system that is optimized for NGS analysis of 25 forensic markers with small-sized amplicon.
- STR genotyping results obtained from NGS analysis were mostly consistent with those from CE-based analyses for 160 Koreans except for 14 samples with 9.1 allele of D2S441.
- Sequence variations which differentiate alleles with the same length were observed in 12 STR loci, and the most variable sequences were observed in D12S391.
- Therefore, NGS analysis of global STRs using newly developed multiplex PCR system could provide additional genetic information for forensic investigation.

Acknowledgments

This work was supported by the research project for practical use and advancement of forensic DNA analysis of Supreme Prosecutors' Office, Republic of Korea (No. 1333-304-260, 2014).