Effective Genotyping Strategy of Forensic Short Tandem Repeat Using Next Generation Sequencing

Eun Hye Kim[1] · Sang-Eun Jung[1] · Kyoung-Jin Shin[1] · Su Jeong Park[2] · Seung Hwan Lee[2] · In Seok Yang[1]

[1]Department of Forensic Medicine, Yonsei University College of Medicine, Seoul, South Korea
[2]DNA Analysis Laboratory, DNA Forensic Division, Supreme Prosecutors' Office, Seoul, South Korea

Introduction

STR markers have been typed by capillary electrophoresis (CE) assay based on their length variation among human individuals. STR amplicons are usually prepared by multiplex PCR with fluorescence dye labeled template specific primers. Despite rapid typing of STR markers by automation of the method, it has some limitations such as the number of STR loci to be measured simultaneously related to the number of fluorescence dyes and the maximum size of STR amplicons. Recently NGS has been on the spotlight as an ultimate genotyping tool to overcome the limitation of CE-based STR analysis in forensic field. STR profiling using NGS has become available along with advance of bioinformatics tools. NGS platforms produce shorter reads, but vastly greater numbers of reads than traditional Sanger sequencing. Hence, appropriate data analysis protocol may be required for STR profiling using NGS.

Materials and Methods

1. Preparation of STR amplicons and NGS libraries

1 ng of standard DNA–2800M, 9947A, and 1:1 mixture were used for PCR amplification. STR amplicons prepared by using primer set from PowerPlex 16 system without fluorescence dye. All PCR procedures were performed equally with 34 thermal cycles from each 1 ng of the sample. Then the amplicons were purified using column purification method. DNA libraries were prepared by ligation of adaptors with multiplex identifier (MID). Size selection using AMPure beads was subsequently performed to remove small fragments.

2. Generation and analysis of NGS data

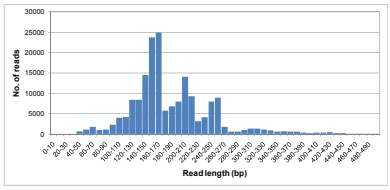
The NGS data was obtained by emulsion PCR of the NGS libraries and subsequent sequencing on PicoTiterPlate of Roche 454 GS Junior platform. All experimental procedures were followed by manufacturer's instruction. Obtained NGS data were sorted into three datasets according to the MID sequences. For building the reference sequences, known STR repeat structures were obtained from STRbase (http://www.cstl.nist.gov/strbase/) and flanking sequences were from human reference genome GRCh37/hg19.

NGS data analysis basically follows the protocol presented by Bornman et al. (Biotechniques, 2012) with modifications on coverage calculation and STR allele assignment. NGS reads were aligned onto the STR reference sequence using Bowtie 2 program on Linux operating system. Samtools and BEDTools were subsequently used for format conversion of alignment outputs. For coverage calculation, aligned sequence reads were filtered based on breadth and mapping quality of them.

Results

1. Size distribution of NGS reads

Totally 164,468 reads were generated from a sequencing run using Roche 454 GS Junior platform. Average read length was 183.64 bp.

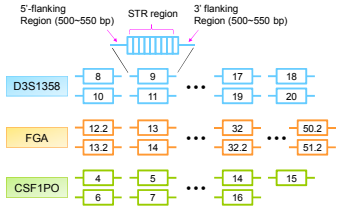


2. NGS datasets

Sample	MID sequences	No. of NGS reads
2800M	ACACGACGACT	51,475
9947A	ACACGTAGTAT	33,213
1:1 mixture	ACGACACGTAT	76,943
Unsorted		2,837

Obtained NGS data were sorted into three datasets according to the presence of MID sequences.

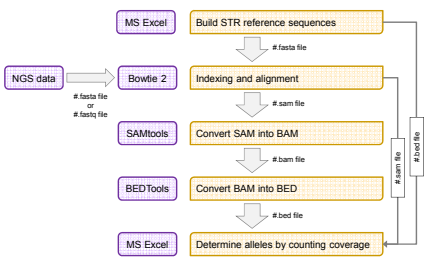
3. Schematic view of STR reference sequence



Long flanking sequences ranged between 500 bp and 550 bp in STR reference sequences were designed for complete alignment of sample sequences that generated with any primer combinations.

Furthermore, this approach allows NGS reads containing STR region with simple and even complex or compound repeat structure to be genotyped exactly.

4. STR genotyping protocol from NGS data



A short read aligners, Bowtie 2, was used to align NGS reads with the reference sequences. Its output (SAM file) was used to determine STR alleles.

5. Alignment results of NGS datasets

STR locus	Amplicon size range (bp)	2800M		9947A		1:1 mixture	
		All	Entire STR	All	Entire STR	All	Entire STR
D3S1358	115-147	9470	8743	6341	6012	14261	13306
D5S818	119-155	9485	8705	5523	5011	9347	8531
D7S820	215-247	3676	3476	1868	1780	4815	4603
D8S1179	203-247	4458	4017	1967	1805	3368	3054
D13S317	169-201	4897	4631	4060	3868	12839	12140
D16S439	264-304	967	877	708	655	2497	2361
D18S51	209-366	739	332	1284	546	1117	481
D21S11	203-259	3045	2313	2996	2525	4873	3871
CSF1PO	321-357	291	244	596	522	862	742
FGA	322-444	956	460	666	255	3137	1440
Penta D	376-441	142	31	267	56	403	75
Penta E	379-474	193	84	356	116	563	309
TH01	156-195	5503	4620	3324	2811	6712	5518
TPOX	262-290	269	230	215	183	679	576
vWA	123-171	3153	2782	1014	919	8565	7649
AMEL	106, 112	3416	3247	1773	1741	2334	2247
Total		50550	44792	32958	28805	76372	66903
Unaligned		1105		410		834	

This table shows the number of aligned reads at each locus. It is observed that the numbers are inversely related to the amplicon size ranges of STR loci.

"All" indicates all aligned reads regardless of the presence of entire STR region. "Entire STR" represents aligned reads containing entire STR region.

6. Determination of STR alleles

(1) 2800M (D3S1358)

Alleles	Allele coverage	Locus coverage	Percentage of allele coverage	Assigned allele
11	0	8743		
12	0	8743		
13	2	8743	0.02%	
14	13	8743	0.15%	
15	71	8743	0.81%	
16	495	8743	5.66%	
17	4355	8743	49.81%	O
18	3757	8743	42.97%	O
19	24	8743	0.27%	
20	26	8743	0.30%	

(2) 9947A (D3S1358)

Alleles	Allele coverage	Locus coverage	Percentage of allele coverage	Assigned allele
11	2	6012	0.03%	
12	12	6012	0.20%	
13	217	6012	3.61%	
14	2868	6012	47.70%	O
15	2879	6012	47.89%	O
16	34	6012	0.57%	
17	0	6012		
18	0	6012		
19	0	6012		
20	0	6012		

(3) 1:1 mixture (D3S1358)

Alleles	Allele coverage	Locus coverage	Percentage of allele coverage	Assigned allele
11	0	13309		
12	5	13309	0.04%	
13	103	13309	0.77%	
14	1519	13309	11.41%	O
15	2245	13309	16.87%	O
16	541	13309	4.06%	
17	4936	13309	37.09%	O
18	3906	13309	29.35%	O
19	33	13309	0.25%	
20	21	13309	0.16%	

STR alleles for single source sample could be determined when 20% of total coverage was used as a threshold.

But, the threshold had to be lowered to 10% to assign alleles for 1:1 mixture sample.

These threshold values allowed STR allele to be determined consistently in most loci.

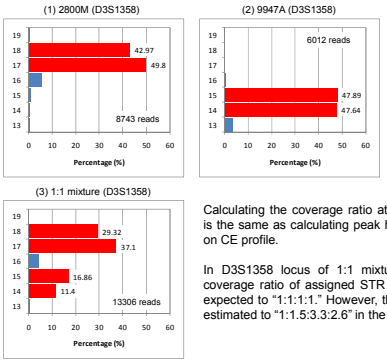
7. STR genotyping results from NGS data

STR locus	2800M		9947A		1:1 mixture	
	CE	NGS	CE	NGS	CE	NGS
D3S1358	17, 18	17, 18	14, 15	14, 15	14, 15, 17, 18	14, 15, 17, 18
D5S818	12	12	11	11	11, 12	11, 12
D7S820	8, 11	8, 11	10, 11	10, 11	8, 10, 11	8, 10, 11
D8S1179	14, 15	14, 15	13	13	13, 14, 15	13, 14, 15
D13S317	9, 11	9, 11	11	11	9, 11	9, 11
D16S539	9, 13	9, 13	11, 12	11, 12	9, 11, 12, 13	9, 11, 12, 13
D18S51	16, 18	16, 18	15, 19	15, 19	15, 16, 18, 19	15, 16, 18, 19
D21S11	29, 31.2	29, 31.2	30	30	29, 30, 31.2	29, 30, 31.2
CSF1PO	12	12	10, 12	10, 12	10, 12	10, 12
FGA	20, 23	20, 23	23, 24	23, 24	20, 23, 24	20, 23, 24
Penta D	12, 13	12, 13	12	12	12, 13	12, 13
Penta E	7, 14	7, 14	12, 13	12, 13	7, 12, 13, 14	7, 13, 14
TH01	6, 9.3	6, 9.3	8, 9.3	8, 9.3	6, 8, 9.3	6, 8, 9.3
TPOX	11	11	8	8	8, 11	8, 11
vWA	16, 19	16, 19	17, 18	17, 18	16, 17, 18, 19	16, 17, 19

STR alleles for 2800M and 9947A samples were exactly determined by NGS. However, allele drop-out (orange background) was observed in Penta E and vWA loci of 1:1 mixture sample.

8. Estimation of actual mixture ratio

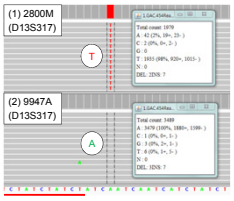
8a. Using coverage ratios of assigned STR alleles



Calculating the coverage ratio at each locus is the same as calculating peak height ratios on CE profile.

In D3S1358 locus of 1:1 mixture sample, coverage ratio of assigned STR alleles was expected to "1:1:1:1." However, the ratio was estimated to "1:1.5:3.3:2.6" in the locus.

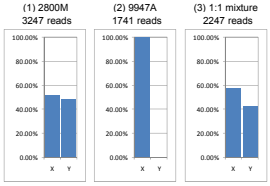
8b. Using reference/variant ratios from observed sequence variation



Sample (coverage)	A	T	G	C
1:1 mixture (5683)	2642 (46%)	3037 (53%)	2	2

A sequence variation of A to T was detected in flanking region of D13S317 locus. When mixture ratio was calculated using the information, it was exactly estimated to "1:1" as expected.

9. Estimation of male/female ratio



Male/female ratio was estimated using coverage ratios of amelogenine locus.

The ratio between X and Y alleles in 1:1 mixture sample was obtained to "1.36:1" unlike theoretically expected one.

10. Determination of STR repeat structures

(1) Two STR alleles with the same length, but with different sequence

Sample	STR locus	Allele	Repeat structure
9947A	D8S1179	13	13a: TCTA TCTG [TCTA]11 13b: [TCTA]13

(2) Different repeat structure between samples

Sample	STR locus	Allele	Repeat structure
2800M	D3S1358	17, 18	17: TCTA [TCTG]3 [TCTA]13 18: TCTA [TCTG]3 [TCTA]14
9947A	D3S1358	14, 15	14: TCTA [TCTG]2 [TCTA]11 15: TCTA [TCTG]2 [TCTA]12

(3) Sequence variation in flanking region

Sample	STR locus	Allele	Repeat structure
9947A	D8S1179	9, 11	9: [TATC]9 AATC AATC 11: [TATC]11 TATC AATC

As shown in left tables, three different types of sequence variations were observed in target STR or near flanking regions.

These results reveal that STR alleles with the same length, but with different repeat structure can be distinguished using NGS.

Discussion

- **Successful STR allele call from single source and 1:1 mixture sample could be achieved based on 20% and 10% of total coverage, respectively, using reference alignment method.**
- **Actual mixture ratio of a mixed sample was estimated by not only analyzing coverage ratios of the assigned alleles but also examining reference/variant ratios from observed sequence variations.**
- **Repeat structures of the assigned STR alleles were successfully determined by simultaneously examining repeat length and sequence variation in target STR regions.**
- **Therefore, NGS data analysis method presented from this study will be helpful to interpret and analyze STR profile from single source and even mixed samples for forensic investigation using NGS.**

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 (1333-304-260, 2012 and 2013).