Amplicon preparation and data analysis for autosomal STR profiling using next generation sequencing

In Seok Yang · Eun Hye Kim · Kyoung-Jin Shin

Department of Forensic Medicine, Yonsei University College of Medicine, Seoul, Korea

Introduction

STR markers are highly variable among individuals, thereby making these STRs effective for human identification and kinship testing. Length-based genotyping using capillary electrophoresis (CE) has been the standard approach to forensic STR analysis. Recently emerged next generation sequencing (NGS) has the potential to be ultimate genotyping platform for forensic purpose. Still, amplicon sequencing method for STR profiling using NGS and its resultant data analysis protocol were not well established. Therefore, we tested 3 different amplicon preparation methods for NGS library and sought to present appropriate NGS data analysis protocol for PowerPlex 16 STRs profiling in this study.

Materials and Methods

1. STR loci and primer design

Dataset 1 : PCR product using PowerPlex 16 HS system

Dataset 2: PCR product using NGS adaptor fusion primers

Forward primer (Primer A-Key)
5'-CGTATCGCCTCCCTCGCGCCATCAG-*{template-specific-sequence}*-3'
Reverse primer (Primer B-Key)
5'-CTATCGCCTTGCCAGCCCGCTCAG-*{template-specific-sequence}*-3'

Dataset 3: PCR product using original primers of PowerPlex 16 system

	Type 1 primer set	Type 2 primer set	Type 3 primer set
	Dataset 1	Dataset 2	Dataset 3
Fluorescence dye attached	+	–	–
Adaptor sequence included	–	+	–

2. STR amplicon preparation

STR amplicons were prepared using PowerPlex 16 protocol with 34 thermal cycles from 1 ng of 2800M Control DNA sample. Then the amplicons were purified using Qiaquick PCR Purification Kit.

3. NGS library preparation

For PCR products of dataset 1 and 3, NGS libraries were prepared by ligation of NGS adaptors.

For PCR products of dataset 2, NGS library preparation was omitted, since PCR amplification was carried out with NGS adaptor fusion primer.

Size selection using AMPure beads was subsequently performed to remove small fragments.

4. NGS data generation

NGS datasets were obtained by emersion PCR of the NGS libraries and subsequent sequencing on 1/8 of a GS Pico TitierPlate using Roche (454) GS FLX Titanium Sequencer.

5. NGS data analysis

NGS data analysis in this study basically follows the protocol presented by Bornman et al. (Biotechniques, 2012) with modifications on coverage calculation and STR allele assignment.

STR reference sequences were built using known STR repeat structure in STRbase (http://www.cstl.nist.gov/strbase/). 5' and 3' flanking sequences were from human reference sequence GRCh37/hg19.

Sequence reads of 3 NGS datasets 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.

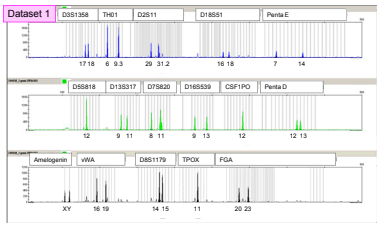
For coverage calculation, aligned sequence reads were filtered based on breadth and mapping quality of them.

6. Assignment of STR alleles

STR alleles were assigned by using coverage threshold of 20%, which were empirically determined for single-source samples.

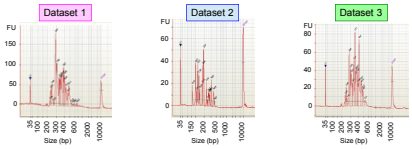
Results

1. CE profile of STR amplicon



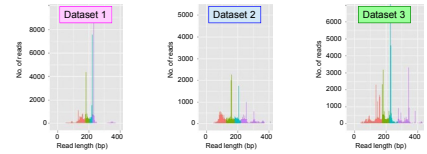
Full STR genotypes for 16 loci could be obtained from all STR amplicons that were generated with 3 different primer sets.

2. NGS library profiles



All NGS libraries were ranged between 150 and 600 bp. These library profiles were formed in relatively smaller size range than those of neubilized DNA fragments for sequencing of whole genome or mitochondrial DNA.

3. Read length distributions



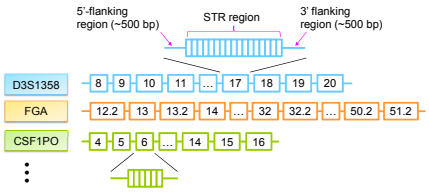
Through duplicate experiment, three different NGS datasets could be obtained consistently. Different patterns on read length distribution were observed among 3 NGS datasets.

4. Total number of sequence reads

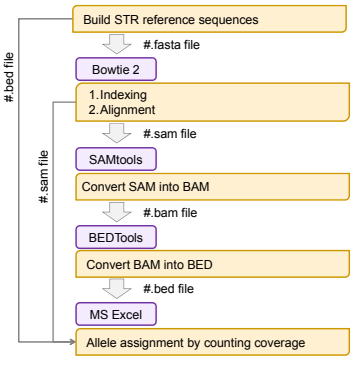
	Dataset 1	Dataset 2	Dataset 3
Data quantity (total number of sequence reads)	143,404 ★★★	82,371 ★	155,781 ★★★

NGS datasets obtained by PCR amplification f target STR loci followed by adaptor ligation (datasets 1 and 3) showed higher data quantity than those from PCR products generated with NGS adaptor fusion primers (dataset 2).

5. Schematic view of STR reference sequence



6. STR profiling protocol from NGS data



7. Alignment results of NGS datasets

STRs	Dataset 1 Number of aligned reads Total	Forward	Reverse	Dataset 2 Number of aligned reads Total	Forward	Reverse	Dataset 3 Number of aligned reads Total	Forward	Reverse
D3S1358	3135	1431	1704	11453	5680	5773	9087	1552	1535
D5S818	1548	675	873	2182	1274	908	4446	1893	2553
D7S820	60540	28989	31551	3682	3205	2477	25520	12393	13127
D8S1179	60	36	24	4844	2853	1991	23146	11274	11872
D13S317	16730	8221	10509	6337	3732	2605	19020	7588	11432
D16S539	87	48	39	2775	1557	1218	13392	8034	5358
D18S51	259	113	146	2738	1661	1077	7452	3406	4046
D21S11	39999	19093	20906	12134	6245	5889	17141	7772	9369
CSF1PO	113	61	52	3569	1822	1747	10105	5368	4737
FGA	205	83	122	2605	1278	1327	8474	6133	3341
Penta_D	227	91	136	907	425	482	9299	1895	1404
Penta_E	87	26	61	968	540	428	9547	1509	2038
TH01	17260	7127	10133	13031	5141	7890	7504	3318	4186
TPOX	59	24	35	2143	1266	877	4286	2078	2208
VWA	85	59	26	5300	2463	2837	9574	2128	1446
Total	140394	64077	76317	76668	39142	37526	154993	76341	78652
Unaligned	1010			5703			988		

The number of sequence reads mapping to each STR locus appeared differently among 3 NGS datasets.

8. Coverage of STR alleles

	STRs	Start	End	Alleles	Dataset 1	Dataset 2	Dataset 3
CSF1PO	CSF1PO	508	537	CSF1PO_5	0	2	0
	CSF1PO	1576	1609	CSF1PO_6	0	0	1
	CSF1PO	2648	2685	CSF1PO_7	0	0	0
	CSF1PO	3724	3765	CSF1PO_8	0	5	0
	CSF1PO	4804	4849	CSF1PO_9	0	17	4
	CSF1PO	5888	5937	CSF1PO_10	2	44	40
	CSF1PO	6976	7029	CSF1PO_11	12	200	676
	CSF1PO	8068	8125	CSF1PO_12	78	2180	8834
	CSF1PO	9164	9225	CSF1PO_13	2	35	101
	CSF1PO	10264	10329	CSF1PO_14	0	0	1
	CSF1PO	11368	11437	CSF1PO_15	0	0	0
	CSF1PO	12476	12549	CSF1PO_16	0	1	0
D13S317	D13S317	505	534	D13S317_5	2	0	1
	D13S317	1581	1614	D13S317_6	5	2	0
	D13S317	2661	2698	D13S317_7	25	7	6
	D13S317	3745	3786	D13S317_8	343	208	272
	D13S317	4833	4878	D13S317_9	8000	2982	8956
	D13S317	5925	5974	D13S317_10	711	268	586
	D13S317	7021	7074	D13S317_11	6880	1439	8373
	D13S317	8121	8178	D13S317_12	120	48	139
	D13S317	9225	9286	D13S317_13	4	0	1
	D13S317	10333	10398	D13S317_14	0	0	4
	D13S317	11445	11514	D13S317_15	1	0	0
	D13S317	12561	12634	D13S317_16	0	0	0
	D13S317	13681	13758	D13S317_17	1	0	0
					Dataset 1	Dataset 2	Dataset 3
				Data quality (coverage)	★	★★	★★★

NGS datasets in high quality order were dataset 3, dataset 2, and dataset 1. Thus, it revealed that NGS dataset with the highest quality was obtained by PCR amplification using original primers of PowerPlex 16 system (no fluorescence dye and NGS adaptor).

9. An example of STR allele assignment

STRs	Start	End	Alleles	Coverage	Total coverage	Proportion	Alleles	
D13S317	2661	2698	D13S317_7	6	18241	0.00033		
D13S317	3745	3786	D13S317_8	272	18241	0.01491		
D13S317	4833	4878	D13S317_9	8856	18241	0.48550	O	P ≥ 20%
D13S317	5925	5974	D13S317_10	586	18241	0.03213	O	
D13S317	7021	7074	D13S317_11	8273	18241	0.45902	O	
D13S317	8121	8178	D13S317_12	139	18241	0.00762		
D13S317	9225	9286	D13S317_13	1	18241	0.00005		

Most STR alleles could be assigned exactly with coverage threshold of 20%.

10. Accuracy of STR allele assignment

STRs	CE results	Dataset 1 Alleles P ≥ 20%	Dataset 1 20% > P ≥ 10%	Dataset 2 Alleles P ≥ 20%	Dataset 2 20% > P ≥ 10%	Dataset 3 Alleles P ≥ 20%	Dataset 3 20% > P ≥ 10%
D3S1358	17,18	17, 18		17, 18		17, 18	
D5S818	12	12		12		12	
D7S820	8, 11	8, 11		8, 11		8, 11	
D8S1179	14, 15	14, 15		14, 15	16	14, 15	
D13S317	9, 11	9, 11		9, 11		9, 11	
D16S539	9, 13	9, 13		9	13	9, 13	
D18S51	16, 18	16, 18		16	16.2, 18	16, 18	
D21S11	29, 31.2	29, 31.2		29, 31.2		29, 31.2	
CSF1PO	12	12		12		12	
FGA	20, 23	20, 23		20, 23		20, 23	
Penta_D	12, 13	12, 13		12, 13		12, 13	
Penta_E	7, 14	7	14	7	14	7, 14	
TH01	6, 9.3	6, 9.3		6, 9.3		6, 9.3	
TPOX	11	11		11		11	
VWA	16, 19	16(18'), 19(21)	18(20)	16(18'), 19(21)		16(18'), 19(21)	
				Dataset 1		Dataset 2	Dataset 3
				Accuracy of STR allele assignment		★★	★	★★★

NGS datasets in high accuracy order on STR allele assignment were dataset 3, dataset 1, and dataset 2. Thus, it revealed that NGS dataset with accurate STR allele call was obtained by PCR amplification using original primers of PowerPlex 16 system (no fluorescence dye and NGS adaptor).

Conclusion

Through comparison among three NGS datasets, it revealed that NGS dataset with the highest quality for STR profiling was obtained by amplicon preparation with template specific primer followed by adaptor ligation. Therefore, the amplicon preparation method and NGS data analysis protocol presented from this study will be helpful to interpret and analyze STR profile for forensic purpose using a NGS platform.

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, 2012).