Investigation into the Y-STR Typing Using Next Generation Sequencing

So Yeun Kwon[1,2] · Eun Hye Kim[1,2] · Hwan Young Lee[1]· Kyoung-Jin Shin[1,2]

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

Introduction

Next generation sequencing (NGS) can produce massively parallel sequencing data for many targeted regions at high depths of coverage, which implies the possibility of successful application of NGS to forensic casework sample analysis. Until now, NGS studies were mainly progressed in autosomal STR than Y chromosomal STR (Y-STR) and it has been difficult to find studies of NGS on Y-STR previously. Therefore, in the present study, we constructed and evaluated NGS optimized Y-STR multiplex system including 24 Y chromosomal markers (DYS19, DYS385ab, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS481, DYS533, DYS549, DYS570, DYS576, DYS635, DYS643, GATA-H4 and M175). And we scrutinized the genotyping concordance between NGS and capillary electrophoresis method with 149 unrelated Korean males. Finally, we present the identified sequence variations and the results of statistical analysis of 23 Y-STRs in Korean males.

Materials and Methods

1. DNA samples

DNA was extracted from buccal swab samples of 149 unrelated Korean males using QIAamp DNA Mini Kit (Qiagen) and quantified using NanoDrop® ND-1000 Spectrophotometer (NanoDrop Technologies) according to the manufacturer's instructions. 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. Y-STR multiplex PCR system

A multiplex PCR system included the PowerPlex® Y23 (Promega) Y-STRs and M175 shown in Fig. 1. Primers were designed using the Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) program such that the amplicon sizes of 24 targeted markers were less than 253 bp, and the primers did not bind to the region with a greater than 1% mutation rated based on the NCBI SNP information (http://www.ncbi.nlm.nih.gov/SNP/). The multiplex system was designed for sufficient sensitivity and specificity from 100 pg of genomic DNA and a male-female mixture sample with a ratio of 1:1000.

3. Preparation of NGS libraries

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 Y chromosomal STR itself, and primer sequences included Y-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 30 thermal cycles from each 1 ng of the sample and appropriate concentration of primers.

2) The second PCR - indexing PCR was performed with 17 thermal cycles from each 1.0 μl of 100-fold diluted the first PCR products 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.

5. NGS run and data analysis

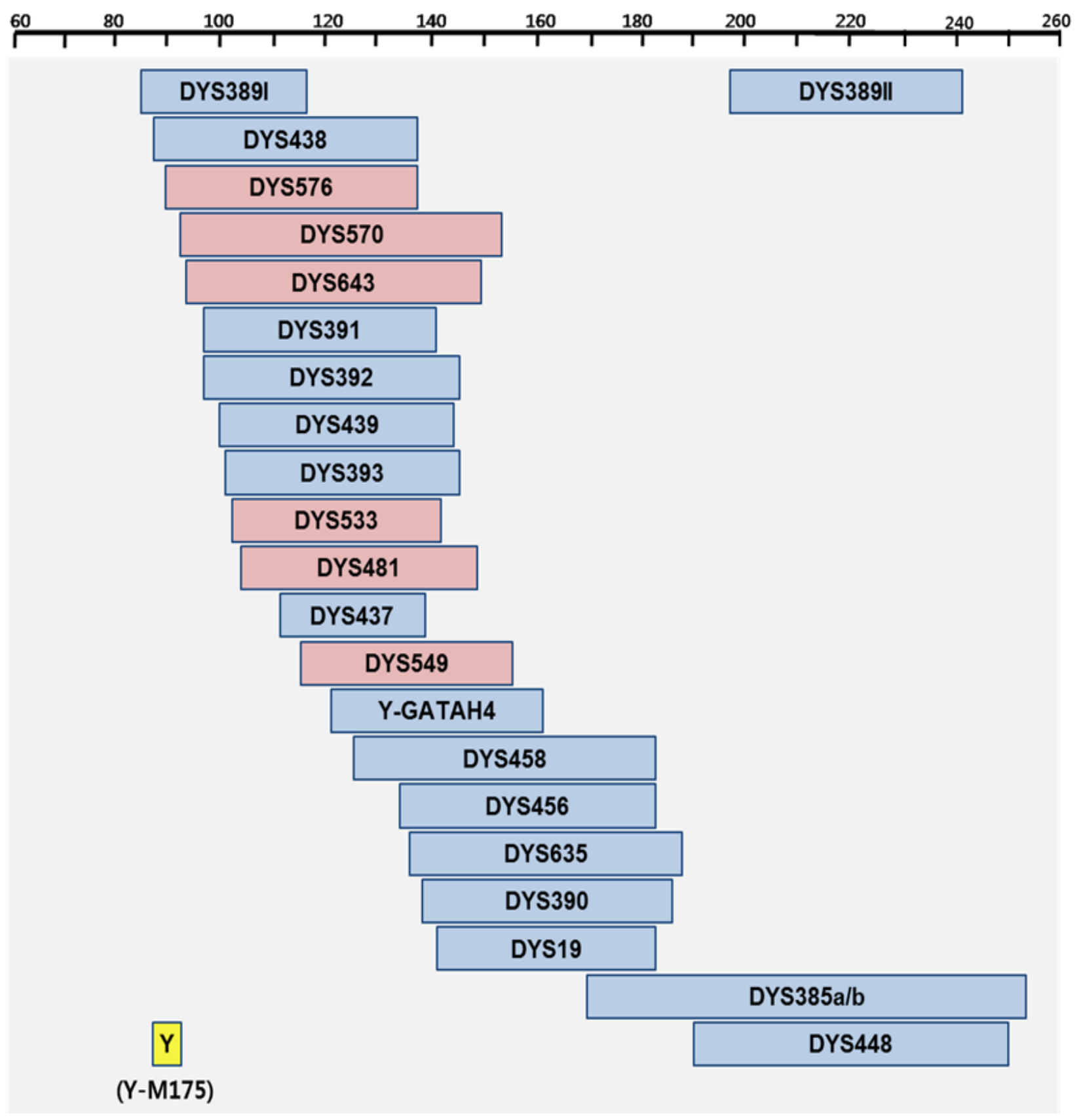
The barcoded libraries were normalized to 10nM and then pooled in equal volumes. Finally, the pooled library was sequenced on MiSeq™ (Illumina) using a MiSeq Reagent Kit v2, 2x250bp (Illumina). NGS data analysis basically follows the protocol presented by Bornman et al (Biotechniques, 2012). The process of NGS data analysis used in this study were illustrated in Fig. 2. The STR profiles obtained by two CE methods – the AmpFℓSTR® Yfiler™ Kit (Applied Biosystems) and in-house Euplex Y15 system were used as reference data for comparing the STR typing results from NGS.

6. Statistical analysis

Different haplotypes and unique haplotypes were calculated using counting method. Haplotype diversities and discrimination capacities were estimated using Nei's formulas. The statistical parameters were compared between 17 Yfielr™ and 23 Powerplex® Y23 loci. Moreover, forensic efficiency information was assessed by gene diversity and haplotype diversity between CE and NGS.

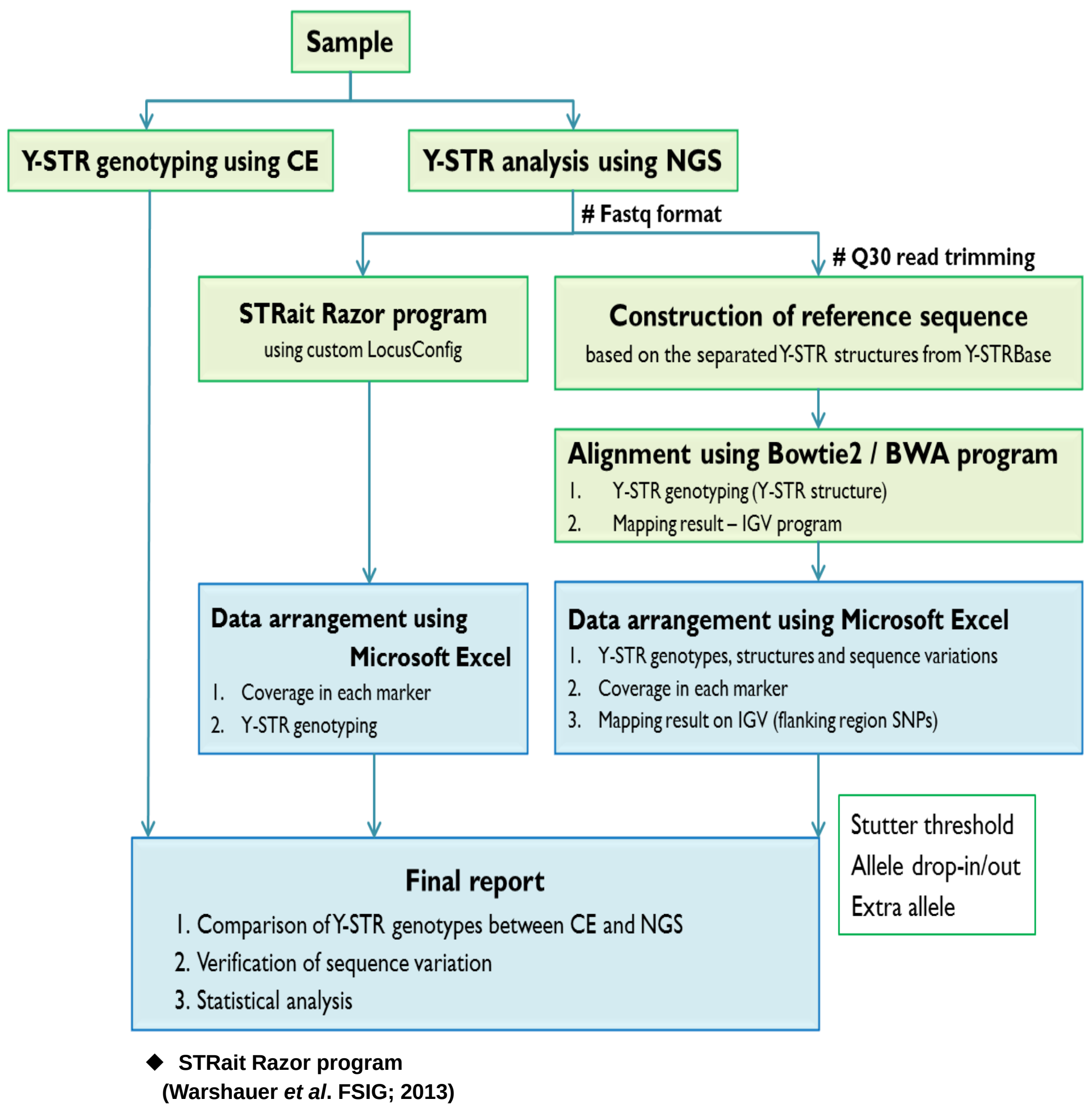
Results

Fig. 1. Allelic size range of 24 Y chromosomal markers



The 23 Y-STR and Y-M175 markers were amplified simultaneously in a size range of 85-253bp. AmpFℓSTR® Yfiler™ loci are marked in blue boxes and additional 6 loci from PowerPlex® Y23 are marked in red boxes. M175 marker is indicated within yellow box, which is a representative Y-SNP marker of Y-haplogroup O. Most of amplicons were less than 190bp and those of DYS389II, DYS385ab and DYS448 were less than 253bp.

Fig. 2. Workflow of NGS data processing



Sample

Y-STR genotyping using CE

Y-STR analysis using NGS

Fastq format

Q30 read trimming

STRait Razor program
using custom LocusConfig

Construction of reference sequence
based on the separated Y-STR structures from Y-STRBase

Alignment using Bowtie2 / BWA program
1. Y-STR genotyping (Y-STR structure)
2. Mapping result – IGV program

Data arrangement using Microsoft Excel
1. Coverage in each marker
2. Y-STR genotyping

Data arrangement using Microsoft Excel
1. Y-STR genotypes, structures and sequence variations
2. Coverage in each marker
3. Mapping result on IGV (flanking region SNPs)

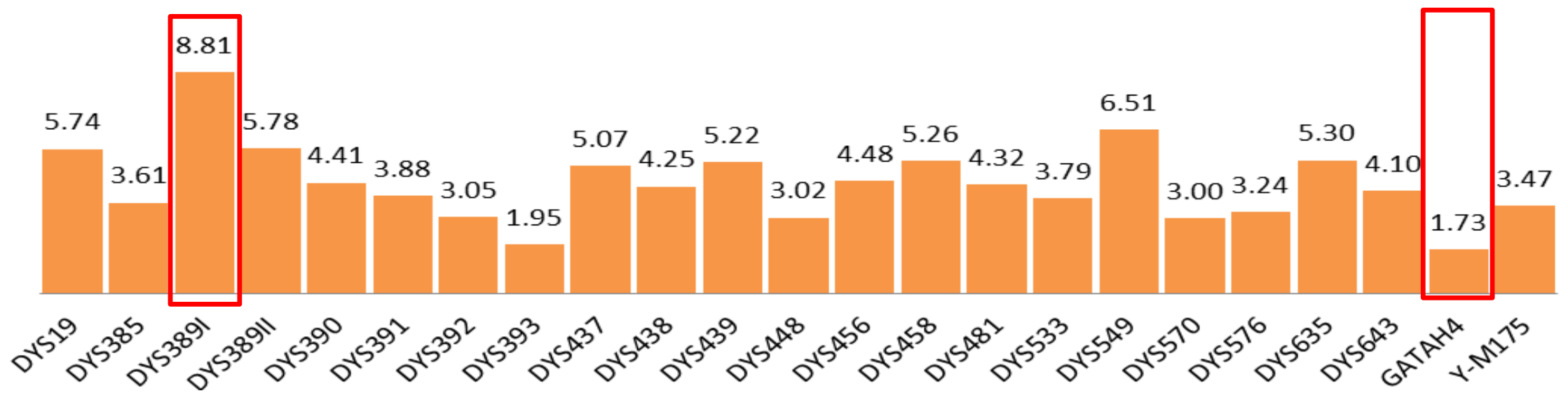
Stutter threshold
Allele drop-in/out
Extra allele

Final report
1. Comparison of Y-STR genotypes between CE and NGS
2. Verification of sequence variation
3. Statistical 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. Y-STR alleles could be determined when 20% of total coverage was used as a threshold

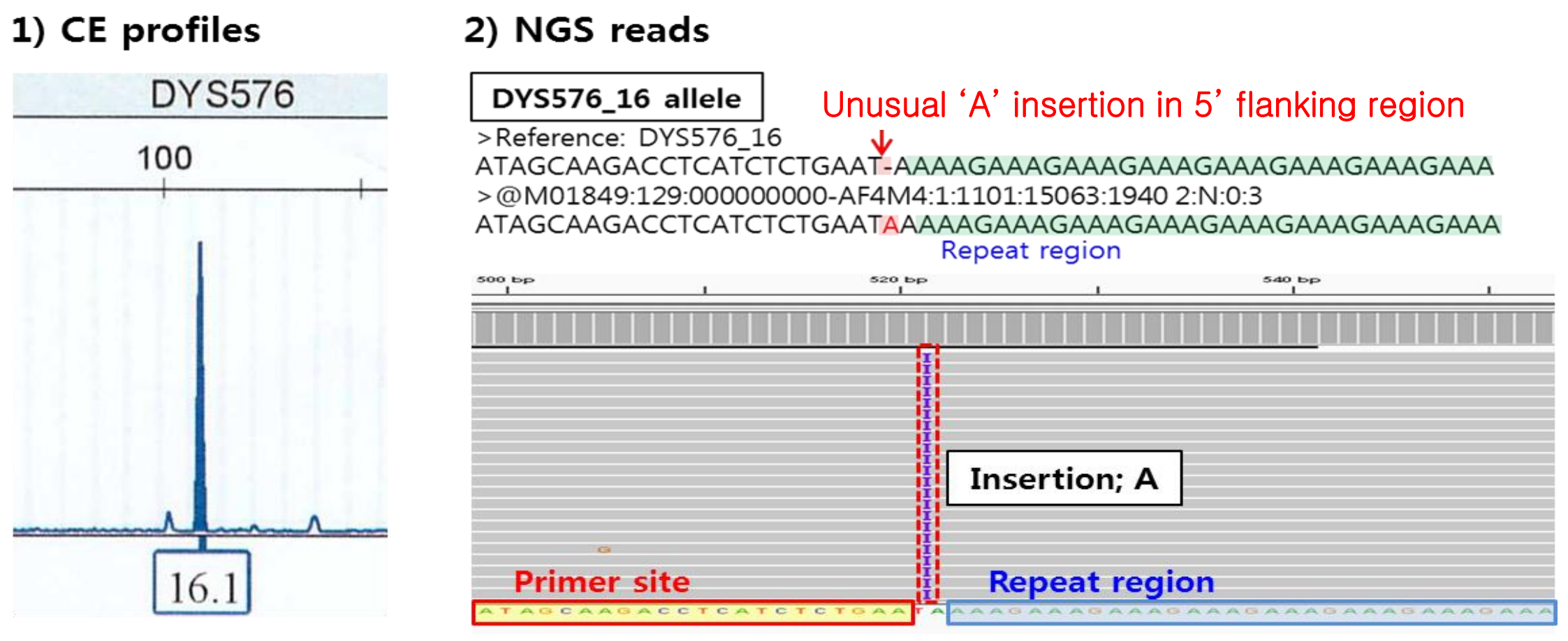
Fig. 3. Relevant coverage between 24 Y chromosomal marker



The relative reads counts in average coverage were calculated by STRait Razor program. The minimum and maximum coverage were observed on GATA-H4 and DYS389I, respectively. The difference was in less than five times.

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

Y-STR genotypes of 149 unrelated Korean males obtained from NGS were concordant with CE profiles except for a sample. One sample was identified that an adenine was inserted at 5' flanking region in 16 allele of DYS576 but shown as16.1 in CE profile.


1) CE profiles

2) NGS reads

DYS576
100
16.1
DYS576_16 allele
Unusual 'A' insertion in 5' flanking region
>Reference: DYS576_16
ATAGCAAGACCTCATCTCTGAAT-AAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA
>@M01849:129:000000000-AF4M4:1:1101:15063:1940 2:N:0:3
ATAGCAAGACCTCATCTCTGAATAAAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA
Repeat region
500 bp
520 bp
540 bp
Insertion; A
Primer site
Repeat region

Table 1. Examples of alleles with observed sequence variation

1) DYS389II

Allele	Sub-allele	Structure	Frequency
26	26	$[TCTG]_4$ $[TCTA]_{10}$ N_{48} $[TCTG]_3$ $[TCTA]_9$	0.007
27	27a	$[TCTG]_4$ $[TCTA]_{10}$ N_{48} $[TCTG]_3$ $[TCTA]_{10}$	0.013
	27b	$[TCTG]_4$ $[TCTA]_{11}$ N_{48} $[TCTG]_3$ $[TCTA]_9$	0.074
	27c	$[TCTG]_4$ $[TCTA]_{12}$ N_{48} $[TCTG]_3$ $[TCTA]_8$	0.007
	27d	$[TCTG]_5$ $[TCTA]_{10}$ N_{48} $[TCTG]_3$ $[TCTA]_9$	0.007
28	28a	$[TCTG]_4$ $[TCTA]_{10}$ N_{48} $[TCTG]_3$ $[TCTA]_{11}$	0.007
	28b	$[TCTG]_4$ $[TCTA]_{11}$ N_{48} $[TCTG]_3$ $[TCTA]_{10}$	0.067
	28c	$[TCTG]_4$ $[TCTA]_{12}$ N_{48} $[TCTG]_3$ $[TCTA]_9$	0.148
	28d	$[TCTG]_4$ $[TCTA]_{13}$ N_{48} $[TCTG]_3$ $[TCTA]_8$	0.007
29	29a	$[TCTG]_4$ $[TCTA]_{11}$ N_{48} $[TCTG]_3$ $[TCTA]_{11}$	0.154
	29b	$[TCTG]_4$ $[TCTA]_{12}$ N_{48} $[TCTG]_3$ $[TCTA]_{10}$	0.060
	29c	$[TCTG]_4$ $[TCTA]_{13}$ N_{48} $[TCTG]_3$ $[TCTA]_9$	0.128
	29d	$[TCTG]_5$ $[TCTA]_{10}$ N_{48} $[TCTG]_3$ $[TCTA]_{11}$	0.027
	29e	$[TCTG]_5$ $[TCTA]_{11}$ N_{48} $[TCTG]_3$ $[TCTA]_{10}$	0.034
30	30a	$[TCTG]_4$ $[TCTA]_{11}$ N_{48} $[TCTG]_3$ $[TCTA]_{12}$	0.007
	30b	$[TCTG]_4$ $[TCTA]_{12}$ N_{48} $[TCTG]_3$ $[TCTA]_{11}$	0.087
	30c	$[TCTG]_4$ $[TCTA]_{13}$ N_{48} $[TCTG]_3$ $[TCTA]_{10}$	0.013
	30d	$[TCTG]_4$ $[TCTA]_{14}$ N_{48} $[TCTG]_3$ $[TCTA]_9$	0.020
	30e	$[TCTG]_5$ $[TCTA]_{11}$ N_{48} $[TCTG]_3$ $[TCTA]_{11}$	0.054
	30f	$[TCTG]_5$ $[TCTA]_{12}$ N_{48} $[TCTG]_2$ $[TCTA]_{11}$	0.007
	30g	$[TCTG]_5$ $[TCTA]_{12}$ N_{48} $[TCTG]_3$ $[TCTA]_{10}$	0.007
31	31a	$[TCTG]_4$ $[TCTA]_{12}$ N_{48} $[TCTG]_3$ $[TCTA]_{12}$	0.007
	31b	$[TCTG]_4$ $[TCTA]_{13}$ N_{48} $[TCTG]_3$ $[TCTA]_{11}$	0.027
	31c	$[TCTG]_5$ $[TCTA]_{12}$ N_{48} $[TCTG]_3$ $[TCTA]_{11}$	0.034

N_{48} : CATTATACCTACTTCTGTATCCAACTCTCATCTGTATTATCTATGTA

Sequence variations were observed in 8 Y-STRs (DYS389II, DYS448, DYS635, DYS390, DYS437, Y-GATA-H4, DYS389I and DYS438) in 149 Korean males. The largest number of sequence variation was observed on DYS389II. DYS448 and DYS635 also had sequence variations in the following order.

2) DYS448

Allele	Sub-allele	Structure	Frequency
Null	Null		0.007
17	17a	$[AGAGAT]_{10}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_7$	0.020
	17b	$[AGAGAT]_9$ AGAGAG ATAGAG $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_8$	0.007
18	18a	$[AGAGAT]_{10}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_8$	0.369
	18b	$[AGAGAT]_{11}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_7$	0.013
19	19a	$[AGAGAT]_{10}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_9$	0.020
	19b	$[AGAGAT]_{11}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_8$	0.221
20	20a	$[AGAGAT]_{11}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_9$	0.141
	20b	$[AGAGAT]_{12}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_8$	0.087
21	21a	$[AGAGAT]_{11}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_{10}$	0.020
	21b	$[AGAGAT]_{12}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_9$	0.074
22	22a	$[AGAGAT]_{12}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_{10}$	0.013
	22b	$[AGAGAT]_{13}$ $[ATAGAG]_2$ $[AGATAG]_3$ ATAGAT AGAGAA $[AGAGAT]_9$	0.007

3) DYS635

Allele	Sub-allele	Structure	Frequency
19	19	$[TAGA]_{10}$ TACA $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.034
20	20a	$[TAGA]_{10}$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.255
	20b	$[TAGA]_{11}$ TACA $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.013
21	21a	$[TAGA]_{11}$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.450
	21b	$[TAGA]_7$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.007
	21c	$[TAGA]_{12}$ TACA $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.007
22	22a	$[TAGA]_8$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.020
	22b	$[TAGA]_{12}$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.121
	22c	$[TAGA]_{11}$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_5$	0.007
23	23a	$[TAGA]_{13}$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.034
	23b	$[TAGA]_9$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.007
24	24	$[TAGA]_{14}$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.040
25	25	$[TAGA]_{15}$ $[TACA]_2$ $[TAGA]_2$ $[TACA]_2$ $[TAGA]_4$	0.007

Table 2. Comparison between CE and NGS analysis

◆ Comparison of the number of alleles (n=149)

Loci	CE	NGS	Fold change
DYS389II	6	24	+4.00x
DYS448	7	13	+1.86x
DYS635	7	13	+1.86x
DYS390	6	10	+1.67x
DYS437	3	5	+1.67x
GATA-H4	4	5	+1.25x
DYS389I	5	6	+1.20x
DYS438	6	7	+1.17x

The most variable sequences were observed in DYS389II, and following by DYS448 and DYS635.

◆ Comparison of gene diviersities (n=149)

Loci	CE	NGS	Changes
DYS389II	0.738	0.915	+0.177
DYS437	0.423	0.599	+0.176
DYS438	0.652	0.726	+0.074
DYS448	0.738	0.785	+0.047
DYS635	0.692	0.718	+0.026
DYS390	0.660	0.674	+0.014
GATA-H4	0.620	0.625	+0.006
DYS389I	0.666	0.668	+0.003

The most significant change of gene diversity was observed in DYS389II with +0.177, and following by DYS437 and DYS438.

Table 3. Haplotype analysis of Yfiler and PowerPlex Y23

	AmpFℓSTR® Yfiler™ (17)[a]	PowerPlex® Y23 (23)
No. of samples	149	149
No. haplotypes	145	149
No. unique haplotypes	142	149
Discrimination Capacity (%)	97.32	100.00
Haplotype diversity	0.99955	1.00000

[a] Number of parenthesis is number of Y-STRs

Haplotype analysis for Yfiler™ and PowerPlex® Y23 loci produced the same statistical values in CE and NGS methods.

Conclusion

- We constructed a new multiplex PCR system optimized for NGS analysis including 24 Y chromosomal markers with small-sized amplicon.
- Y-STR genotypes from NGS analysis were consistent with CE profiles in a total of 149 unrelated Korean males except for a sample with 16.1 allele of DYS576.
- Sequence variations which differentiate alleles with the same length were observed in 8 Y-STR. The largest number of sequence variation was observed on DYS389II in Korean males.
- Therefore, NGS analysis of Y-STR using newly developed multiplex PCR system could provide additional genetic information such as discriminative allele information for forensic investigation.

Acknowledgments

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation funded (NRF) by the Ministry of Science, ICT & Future Planning (No. 2012-0009833)