Next Generation Sequencing of Microhaplotypes for Forensic DNA Typing

Kyoung-Jin Shin, Ph.D.

Dept. of Forensic Medicine

Yonsei University College of Medicine

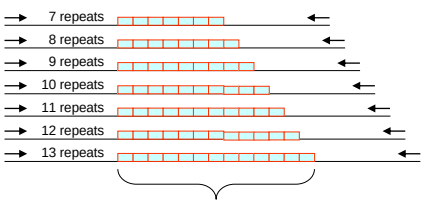
Seoul, Republic of Korea

Short Tandem Repeat (STR)

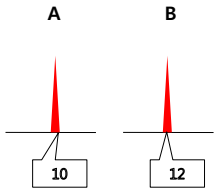
An accordion-like DNA sequence that occurs between genes

TCCCAAGCTCTTCCTCTTCCCTAGATCAATACAGACAGAAGACA
GGTGGATAGATAGATAGATAGATAGATAGATAGATAGATAGA
TAGATAGATATCATTGAAAGACAAAACAGAGATGGATGATAGAT
ACATGCTTACAGATGCACAC

= **12 GATA repeats** ("12" is all that is reported)



Usually 2-5bp per repeat unit



The number of consecutive repeat units can vary between people

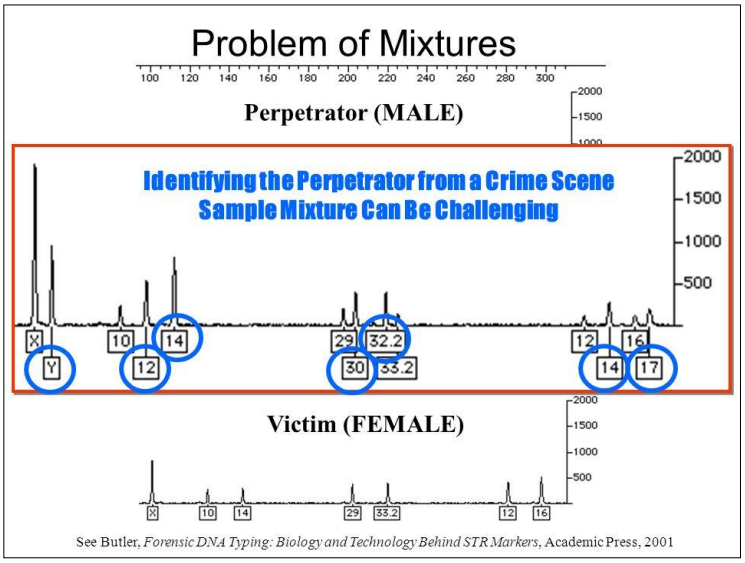
http://www.cstl.nist.gov/div831/strbase/training.htm

Microhaplotype

- New marker needed
  - To infer biogeographic ancestry where the STR data is limited
  - To aid in mixture deconvolution and avoid the issue of stutter
- Defined as at least three haplotypes (alleles) composed of multi-SNPs within a region smaller than 200 bp
- NGS enables high-level multiplexing and phase-known haplotyping of microhaplotypes

Kidd et al. FSIG, 2014 and 2017

Background

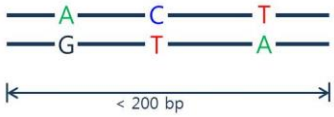


Background

➢ **Simple SNP vs. Microhaplotype**

Simple SNP

Microhaplotype



- Diallelic marker
- Low heterozygosity
- ID or ancestry

- Multi-allelic marker
- High heterozygosity
- ID and/or ancestry

Aims for this study

- Development of a NGS panel to analyze multiple microhaplotypes
- Reporting statistical value of micro-haplotypes in Koreans
- Evaluation of the potential value of microhaplotype in forensics

Methods

➢ Workflow of microhaplotype analysis

1. Target marker selection
2. Primer design to amplify target markers
3. Development of multiplex PCR
4. PCR-based library preparation
5. Library QC & NGS run
6. Data analysis

Microhaplotype

➢ Microhaplotype study

Nearest genes or other symbol	Number of hap alleles		rs-Number dbSNP	Chr	Build 37 nucleotide position	Extent of microhap in base pairs
	>10% freq avg 54 pops	Observed				
CEP104	4	4	rs6681344	1	3,743,132	188
			rs6603840	1	3,743,319	
USH2A ISR1G	4	4	rs4528199	1	214,634,428	18
			rs6904596	1	214,634,445	

This spreadsheet contains all 129 microhaps that have been typed on 55 populations and haplotyped statistically. We have used Informativeness of Rosenberg as a measure of ancestry (I_n) and Global Average Effective Number of Alleles (Ae) (Invest Gen 2015) as a standardized measure of variation. It is obvious that some microhaplotypes rank very low by both existing statistics. However, we note them in case other populations may change the rankings and/or full sequencing may reveal additional variation making them useful in some situations. Also, some of the microhaplotypes are very close molecularly and have been kept in this listing pending future data than may indicate one of them is significantly better than the other(s). A couple of SNPs occur in overlapping microhaplotypes because we have arbitrarily limited a microhaplotype to an extent of 200 basepairs.

standardized name using POSIX characters	Global I_n	Glob Avg Ae	I_n rank	Ae rank	chr	p-most chr_position Build GRCh37/hg19	SNPs_Included	KiddGene use_at own_peril	# SNPs
mh13KK-218	0.37857	5.54820	6	**1**	13	54060827	rs1927847/rs9536429/rs7492234/rs9536430	**D13S169**	4
mh05KK-170	0.41105	4.89784	4	**2**	5	2448024	rs74865590/rs438055/rs370672/rs6555108	**D5S1970**	4
mh21KK-315	0.30144	4.76321	14	**3**	21	21880086	rs8126597/rs6517970/rs8131148/rs6517971	**D21S1263**	4
mh21KK-320	0.26139	4.39290	27	**4**	21	43062859	rs2838081/rs2838082/rs78902658/rs2838083	**LINC00111**	4
mh11KK-180	0.26272	4.00795	25	**5**	11	1690791	rs12802112/rs28631755/rs7112918/rs4752777	**FAM99A**	4
mh13KK-217	0.27577	3.78958	20	**6**	13	46865930	rs7320507/rs9562648/rs9562649/rs2765614	**LRRC63**	4
mh01KK-117	0.29949	3.70283	15	**7**	1	204633340	rs17413714/rs2772234/rs1610401/rs1610400	**LRRN2**	4
mh13KK-223	0.24080	3.66781	34	**8**	13	110806699	rs1192204/rs1192205/rs3825483/rs3825481	**COL4A1**	4
mh01KK-205	0.15142	3.65696	85	**9**	1	18722692	rs11810587/rs1336130/rs1533623/rs1533622	**IGSF21**	4
mh19KK-299	0.25782	3.65057	28	**10**	19	22729551	rs4932999/rs4932769/rs2361019/rs2860462	**LINC01233**	4
mh21KK-324	0.41930	3.63751	2	**11**	21	46714549	rs6518223/rs2838868/rs7279250/rs8133697	**LOC642852**	4
mh02KK-136	0.21666	3.57664	49	**12**	2	228092389	rs6714835/rs6756898/rs12617010	**COL4A3**	3
mh13KK-213	0.22618	3.56103	45	**13**	13	23765541	rs8181845/rs679482/rs9510616	**SGCG**	3
mh16KK-255	0.22448	3.45043	46	**14**	16	81970353	rs16956011/rs3934955/rs3934956/rs4073828	**PLCG2**	4
mh02KK-134	0.29594	3.35710	17	**15**	2	161079435	rs3101043/rs3111398/rs72623112	**ITGB6**	3
mh20KK-307	0.24022	3.28681	35	**16**	20	16513260	rs6044080/rs17674942/rs6044081/rs16997830	**KIF16B**	4
mh18KK-293	0.33334	3.24533	8	**17**	18	76089886	rs621320/rs621340/rs678179/rs621766	**D18S1122**	4
mh22KK-061	0.20551	3.19248	59	**18**	22	44763606	rs763040/rs5764924/rs763041	**D22S1159**	3
mh10KK-169	0.29695	3.18111	16	**19**	10	14208510	rs10796164/rs10796165/rs17154765/rs10796166	**FRMD4A**	4
mh09KK-157	0.18911	3.16004	65	**20**	9	135862479	rs606141/rs8193001/rs56256724/rs633153	**GFI1B**	4
mh03KK-150	0.20618	3.08902	58	**21**	3	131645972	rs1225051/rs1225050/rs1225049/rs1225048	**CPNE4**	4
mh11KK-191	0.24528	2.97384	32	**22**	11	99880163	rs12421109/rs12289401/rs12420819/rs770566	CNTN5	4

Here we globally

COMT

http://medicine.yale.edu/lab/kidd/publications/publications.aspx

Design of multiplex PCR systems

◆ **Criteria**

- Target marker
  - Microhaplptypes suggested by Kidd et al., 2014 and 2015
- Small sized amplicons is adapted as possible
  - while primer is not overlapping with core SNP region
  - finally to be ranged in 90bp ~ 270bp
- Avoid SNP with ≥ 1% variation reported in primer binding area

◆ **Resource**

- UCSC genome browser (http://genome.ucsc.edu/)
- Primer 3 v.0.4.0 (http://frodo.wi.mit.edu/primer3/)

Microhaplotype NGS panel

➢ **Single tube multiplex PCR system for 51 microhaplotypes**

➢ **Amplicon sizes of 51 microhaplotypes (KplexSeq-MH51)**

Group	Marker	Target size (bp)	Group	Marker	Target size (bp)	Group	Marker	Target size (bp)	Group	Marker	Target size (bp)
1	TAS2R1	96	2	CCR2	98	3	TTC12	99	4	DRD3	115
1	USH2A	124	2	PAPD7	130	3	COMT	137	4	ACN9	146
1	COG2	151	2	LINC01233	156	3	GNGT2	157	4	ITGB6	158
1	EDAR	166	2	COL4A3	167	3	CDH4	166	4	D18S1122	169
1	NPEPPS	169	2	SUDS3	180	3	TYRP1	184	4	GFI1B	195
1	D13S169	193	2	D21S1263	198	3	GATA4	200	4	D5S1970	210
1	PLCG2	205	2	D22S1159	216	3	ATXN	220	4	LOC642852	220
1	IGSF21	227	2	KIF16B	221	3	COL4A1	226	4	*ADH7*	225
1	LRRC	221	2	C14ORF43	230	3	RXRA	233	4	CEP104	232
1	NCAM1	231	2	FAM99A	236	3	SGCG	236	4	FRMD4A	239
1	LINC0111	240	2	OR52S1P	240	3	DLEU2	244	4	LRRN2	246
1	ARHGAP27	249	2	CPNE4	247/252	3	LRRC63	254	4	LYPD6B	250
1	PLIN3	269				3	PAH	270	4	FAT1	268

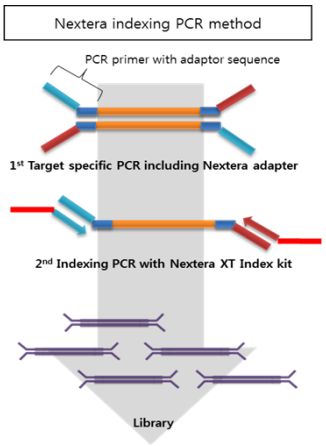
➢ **DNA samples**

- 1ng of 2800M Standard DNA and 122 Korean samples

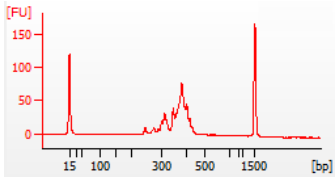
Methods

➢ Library preparation

- PCR-based library preparation



- Assess the barcoded library



- Quantify the barcoded library using KAPA library quantification kit

Two-step PCR for MPS library preparation

◆ **The first PCR using Target specific primers**

PCR component	Volume
dH_20	2.18
Gold ST*R 10× Buffer	2.0
3× Primer Mix-I &, II	6.66
3× Primer Mix-III & IV	6.66
AmpliTaq Gold (5U/uL)	1.5
Template DNA (1ng/uL)	1.0
Total	20.0

◆ **Thermal Cycle**

95℃	11 min	
94℃	20 sec	× 26 cycles
59℃	60 sec	
72℃	45 sec	
72℃	5 min	
4℃	forever	

◆ **The second PCR using Nextera XT Index primers**

PCR component	Volume
dH_20	12.3
Gold ST*R 10× Buffer	2.0
Index 1 (i5)	2.0
Index 2 (i7)	2.0
AmpliTaq Gold (5U/uL)	0.7
1/10 diluted PCR product	1.0
Total	20.0

◆ **Thermal Cycle**

95℃	15 min	
94℃	20 sec	× 15 cycles
61℃	30 sec	
72℃	45 sec	
72℃	5 min	
4℃	forever	

Methods

➢ Library validation

Methods

➢ NGS run

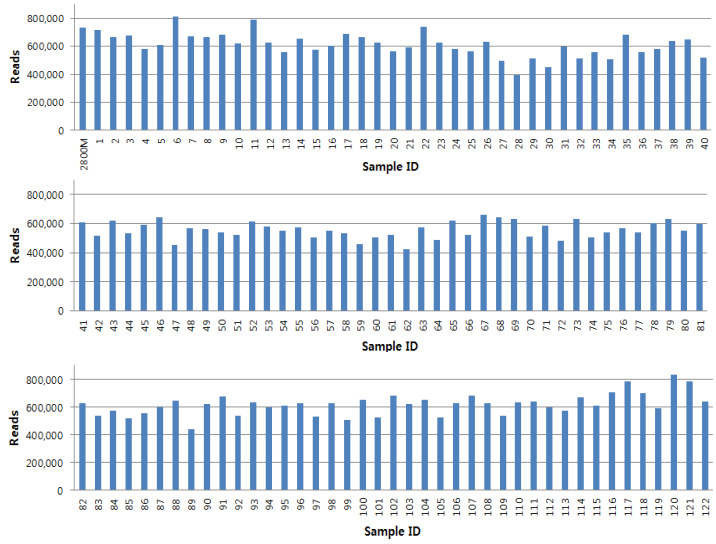
- Library pooling; final conc. 10 nM
- NGS run on an MiSeq system (Illumina)
- MiSeq Reagent Kit v3, 600 Cycles (2x300 bp)

➢ Data analysis

- STRait Razor v2.0 and in-house sequence parsing program
- Data analysis supported by Yonsei Genome Center

Results

➢ Observed reads per sample



Results

➢ Average depth per marker

Marker	ACN9	ADH7	ARHGAP27	ATXN1	C14ORF43	CCR2	CDH4	CEP104	COG2	COL4A1	COL4A3	COMT	CPNE4
Read depth	13753	10502	16976	11374	15277	9113	11331	9065	9547	13051	10447	13875	5318
상대적 비율(%)	2.41	1.84	2.97	1.99	2.67	1.59	1.98	1.59	1.67	2.28	1.83	2.43	0.93

Marker	D13S169	D18S1122	D21S1263	D22S1159	D5S1970	DLEU1	DRD3	EDAR	FAM99A	FAT1	FRMD4A	GATA4	GFI1B
Read depth	13642	12341	15300	12394	10665	5702	9363	10958	10063	5305	12382	10136	13795
상대적 비율(%)	2.39	2.16	2.68	2.17	1.87	1.00	1.64	1.92	1.76	0.93	2.17	1.77	2.41

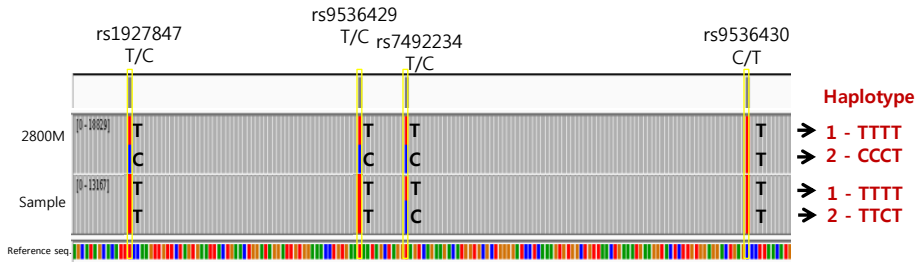
Marker	GNGT2	IGSF21	ITGB6	KIF16B	LINC00111	LINC01233	LOC642852	LRRC2	LRRC63	LRRN2	LYPD6B	NCAM1	NPEPPS
Read depth	11435	11422	14464	10251	10809	10627	13461	14725	10450	12491	15778	11932	11391
상대적 비율(%)	2.00	2.00	2.53	1.79	1.89	1.86	2.36	2.58	1.83	2.19	2.76	2.09	1.99

Marker	OR52S1P1	PAH	PAPD7	PLCG2	PLIN3	RXRA	SGCG	SUDS3	TAS2R1	TTC12	TYRP1	USH2A
Read depth	8203	7326	10403	7009	11826	9723	11190	10681	11968	10763	12556	8851
상대적 비율(%)	1.44	1.28	1.82	1.23	2.07	1.70	1.96	1.87	2.09	1.88	2.20	1.55

Results

- **Phase-known haplotyping of microhaplotype marker**

Example) D13S169



Results

- **Genotypes**

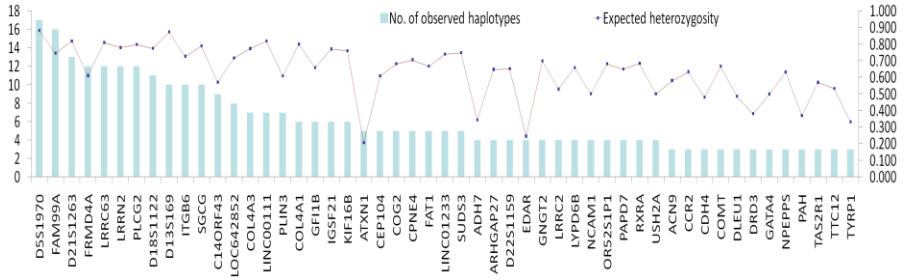
Marker	SNPs_Included
D13S169	rs1927847/rs9536429/rs7492234/rs95
D5S1970	rs74865590/rs438055/rs370672/rs65
D21S1263	rs8126597/rs6517970/rs8131148/rs63
LINC00111	rs2838081/rs2838082/rs78902658/rs2
FAM99A	rs12802112/rs28631755/rs7112918/rs4
LRRC63	rs9562648/rs9562649/rs2765614
LRRN2	rs17413714/rs2772234/rs1610401/rs1
COL4A1	rs1192204/rs1192205/rs3825483/rs38
IGSF21	rs11810587/rs1336130/rs1533623/rs1
LINC01233	rs4932999/rs4932769/rs2361019/rs28
LOC642852	rs6518223/rs2838868/rs7279250/rs81
COL4A3	rs6714835/rs6756898/rs1261701
SGCG	rs8181845/rs679482/rs9510616
PLCG2	rs16956011/rs3934955/rs3934956/rs4
ITGB6	rs3101043/rs3111398/rs7262311
KIF16B	rs6044080/rs17674942/rs6044081/rs16
D18S1122	rs621320/rs621340/rs678179/rs621
D22S1159	rs763040/rs5764924/rs763041

Marker	SNP_1	Frequency	SNP_2	Frequency	SNP_3	Frequency	SNP_4	Frequency
D13S169	rs1927847		rs9536429		rs7492234		rs9536430	
	Allele	C 0.5902	T	0.9344	T	0.4836	C	0.3975
		T 0.4098	C	0.0656	C	0.5164	T	0.6025
D5S1970	rs748							
	Allele	C						
		T						
D21S1263	rs812							
	Allele	G						
		A						
LINC00111	rs283							
	Allele	G						
		A						
FAM99A	rs128							
	Allele	A						
		G						
LRRC63	rs956							
	Allele	G						
		A						
LRRN2	rs174							
	Allele	C						
		A						
COL4A1	rs119							
	Allele	C						
		T						
IGSF21	rs118							
	Allele	T						
		C						
LINC01233								
	Allele	C						
		T						
LOC642852	rs651							
	Allele	C						
		T						
COL4A3	rs671							
	Allele	G						
		T						
SGCG	rs818							
	Allele	T						
		C	0.4221	C	0.8156	A	0.5820	

Marker	SNPs_Included	No. of observed haplotypes	Observed haplotypes	Haplotype frequency
D13S169	rs1927847/rs9536429/rs7492234/rs9536430	10	CTTC	0.1393
			TTCT	0.1680
			CTCT	0.1680
			TTTC	0.0943
			CTCC	0.0984
			TTTT	0.1311
			CCCT	0.0164
			CTTT	0.1189
			CCCC	0.0492
			TTCC	0.0164
D5S1970	rs74865590/rs438055/rs370672/rs6555108	14	CGAA	0.0861
			CGAG	0.1148
			TAAG	0.2090
			CGGA	0.0492
			TAAA	0.1557
			CAAG	0.1189
			CAGG	0.0328
			CAAA	0.1557
			CGGG	0.0574
			TGAG	0.0041
			TGGG	0.0041
			TAGA	0.0041
			TGAA	0.0041
			CAGA	0.0041
D21S1263	rs8126597/rs6517970/rs8131148/rs6517971	12	GATT	0.3279
			GCCC	0.0656
			GACC	0.0328
			ACCC	0.1270
			GCTC	0.1311
			ACTC	0.1844
			GACT	0.0656
			GCCT	0.0205
			ACTT	0.0205
			GATC	0.0164
			AATC	0.0041
			ACCT	0.0041

Results

➢ **Observed haplotypes & frequencies of 51 MHs in Koreans**



Results

Preparing manuscript...

➢ **Forensic statistical parameters for 51 MHs in Koreans**

	H_{obs}	H_{exp}	Forensic Statistics			Paternity Statistics	
			MP	PD	PIC	PE	TPI
D13S169	0.869	0.875	0.035	0.965	0.857	0.732	3.81
D5S1970	0.885	0.870	0.037	0.963	0.852	0.765	4.36
D21S1263	0.811	0.818	0.060	0.940	0.793	0.621	2.65
LINC00111	0.811	0.819	0.067	0.933	0.790	0.621	2.65
COL4A1	0.779	0.801	0.074	0.926	0.767	0.560	2.26
PLCG2	0.779	0.783	0.080	0.920	0.747	0.560	2.26
LRRN2	0.828	0.780	0.083	0.917	0.748	0.652	2.90
LRRC63	0.787	0.769	0.093	0.907	0.730	0.575	2.35
COL4A3	0.811					0.621	2.65
KIF16B	0.697					0.423	1.65
IGSF21	0.664					0.375	1.49
⋮							
CDH4	0.525					0.210	1.05
CEP104	0.549	0.525	0.345	0.655	0.416	0.234	1.11
LRRC2	0.426	0.475	0.370	0.630	0.365	0.131	0.87
PAH	0.402	0.369	0.427	0.573	0.334	0.115	0.84
DRD3	0.385	0.380	0.440	0.560	0.326	0.105	0.81
ADH7	0.344	0.343	0.456	0.544	0.316	0.083	0.76
TYRP1	0.344	0.332	0.492	0.508	0.283	0.083	0.76
EDAR	0.246	0.245	0.592	0.408	0.233	0.044	0.66
ATXN1	0.074	0.103	0.832	0.168	0.100	0.005	0.54

Match Probability

$= 3.82 \times 10^{37}$

Summary

- By investigation of 51 microhaplotypes, a total of 196 SNPs and 316 haplotype alleles were observed from 122 Koreans.
- Each microhaplotype consisted of 2 to 7 SNPs, and exhibited 3 to 17 haplotype alleles.
- Overall match probability (3.98×10^{39}) of the 51 microhaplotypes was much lower than that of the expanded autosomal STRs.

Ongoing work

Forensic Science International: Genetics 29 (2017) 29–37

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

Evaluating 130 microhaplotypes across a global set of 83 populations

Kenneth K. Kidd[a,*], William C. Speed[a], Andrew J. Pakstis[a], Daniele S. Podini[b], Robert Lagacé[c], Joseph Chang[c], Sharon Wootton[c], Eva Haigh[a], Usha Soundararajan[a]

[a] Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT, 06520-8005, USA
[b] Department of Forensic Sciences, The George Washington University, 2100 Foxhall Road, NW, Washington, D.C., 20007, USA
[c] Human Identification Group, ThermoFisher Scientific, 180 Oyster Point Blvd, South San Francisco, CA, 94080, USA

➢ **Developing modified NGS panel to analyze microhalptypes …**

Future work

- Resolution in mixture deconvolution
  - Mixed DNA with ratio of 1:1, 1:3, 1:6, 1:9, 1:19, 1:29, 1:50 and 1:100
- Evaluation of biogeographic inference
  - DNA from 4 population including Koreans
  - Adapted new informative microhaplotypes
- Application to complex kinship case

Acknowledgements

❖ **Yonsei DNA profiling group**

Eun Young Lee, MS
So Yeun Kwon, MS
Sang Eun Jung, MS
Hwan Young Lee, Ph.D.

❖ **Yonsei Genome Center**

Sora Kim, Ph.D. candidate
Sang Woo Kim, Ph.D.

➢ This work was supported by a fund (No. NFS2016DNA02) from the Forensic Research Program of the National Forensic Service, Korea.