中华人民共和国公安部物证鉴定中心
Institute of Forensic Science, Ministry of Public Security, P. R. China

Full resolution genotyping of forensic STRs using the Ion Chef™ and Ion S5™ XL systems

Le Wang

May 10th, 2016

PCR-CE vs NGS for STR Genotyping

PCR-CE (Capillary Electrophoresis):

- **Gold standard**
- **Simple, economic, fast……**
- **Well-accepted by court**
  Accreditated instruments, reagents, laboratories, documents and experimental procedures
- **DNA Databases all over the world**
  >40 million entries in China
- **Long-pending crime cases**

NGS (Next Generation Sequencing):

- **Full resolution of STR base composition**
- **Potential to be compatible with current DNA databases**
- **Potential to be advantageous for challenging (mixed, degraded, LCN) samples and complicated familial testing**

Reports

High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform

Sarah L. Fordyce[1,*], Maria C. Ávila-Arcos[1,*], Eszter Rockenbauer[2], Claus Børsting[2], Rune Frank-Hansen[2], Frederik Torp Petersen[2], Eske Willerslev[1], Anders J. Hansen[2], Niels Morling[2], and M. Thomas. P. Gilbert[1]
[1]*Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen, Denmark and* [2]*Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark*

BioTechniques 51:127-133 (August 2011) doi 10.2144/000113721
Keywords: Short tandem repeat (STR); high-throughput; sequencing; GS FLX

Supplementary material for this article is available at www.BioTechniques.com/article/113721.
*S.L.F. and M.C.A.-A. contributed to this work equally.

Full resolution of STR base composition

STR sequence polymorphism

D3S1358 allele frequency (length polymorphism)

Allele	Count	Frequency
11	1	0.0025
14	30	0.0761
15	111	0.2817
16	102	0.2588
17	83	0.2107
18	63	0.1599
19	3	0.0076
20	1	0.0025

LCN DNA:
Less STR loci required for individual identification

Frequency of allele 18 is 63 times higher than that of allele 18a

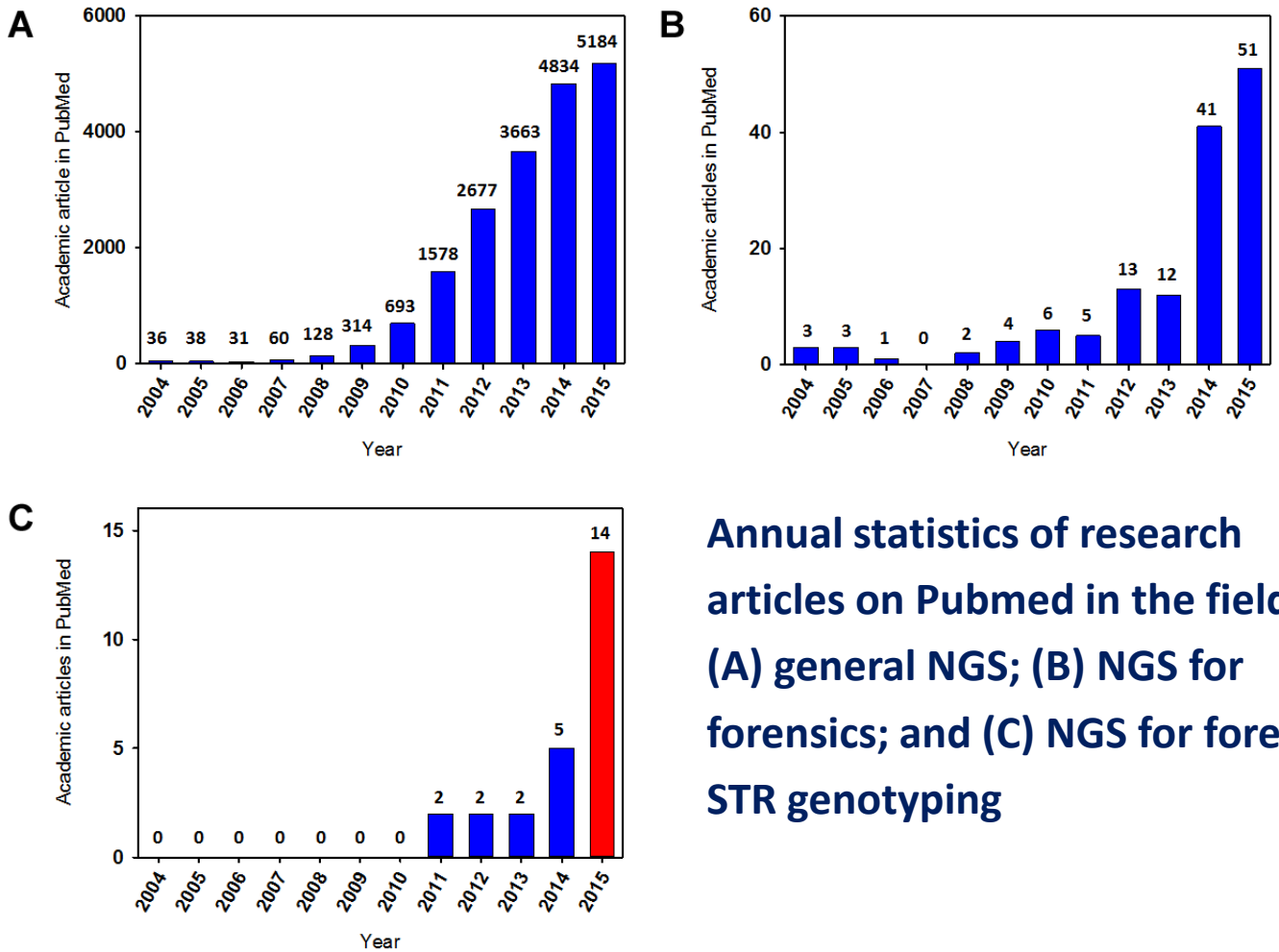
No. of alleles changed from 8 to 17

D3S1358 allele frequency (sequence polymorphism)

Allele	Repeat structure	Count	Frequency
11	$TCTA_{[1]}$ $TCTG_{[2]}$ $TCTA_{[8]}$	1	0.0025
14a	$TCTA_{[1]}$ $TCTG_{[3]}$ $TCTA_{[10]}$	1	0.0025
14b	$TCTA_{[1]}$ $TCTG_{[2]}$ $TCTA_{[11]}$	28	0.0711
14c	$TCTA_{[1]}$ $TCTG_{[1]}$ $TCTA_{[12]}$	1	0.0025
15a	$TCTA_{[1]}$ $TCTG_{[2]}$ $TCTA_{[12]}$	100	0.2538
15b	$TCTA_{[1]}$ $TCTG_{[1]}$ $TCTA_{[13]}$	11	0.0279
16a	$TCTA_{[1]}$ $TCTG_{[3]}$ $TCTA_{[12]}$	32	0.0812
16b	$TCTA_{[1]}$ $TCTG_{[2]}$ $TCTA_{[13]}$	64	0.1624
16c	$TCTA_{[1]}$ $TCTG_{[1]}$ $TCTA_{[14]}$	6	0.0152
17a	$TCTA_{[1]}$ $TCTG_{[3]}$ $TCTA_{[13]}$	29	0.0736
17b	$TCTA_{[1]}$ $TCTG_{[2]}$ $TCTA_{[14]}$	52	0.132
17c	$TCTA_{[1]}$ $TCTG_{[1]}$ $TCTA_{[15]}$	2	0.0051
18a	$TCTA_{[1]}$ $TCTG_{[4]}$ $TCTA_{[13]}$	1	0.0025
18b	$TCTA_{[1]}$ $TCTG_{[3]}$ $TCTA_{[14]}$	55	0.1396
18c	$TCTA_{[1]}$ $TCTG_{[2]}$ $TCTA_{[15]}$	7	0.0178
19	$TCTA_{[1]}$ $TCTG_{[3]}$ $TCTA_{[15]}$	3	0.0076
20	$TCTA_{[1]}$ $TCTG_{[4]}$ $TCTA_{[15]}$	1	0.0025

Data source: C. Gelardi et al. Forensic Sci. Int. Genet. 12 (2014): 38-41.

Trend analysis: small proportion but increasing



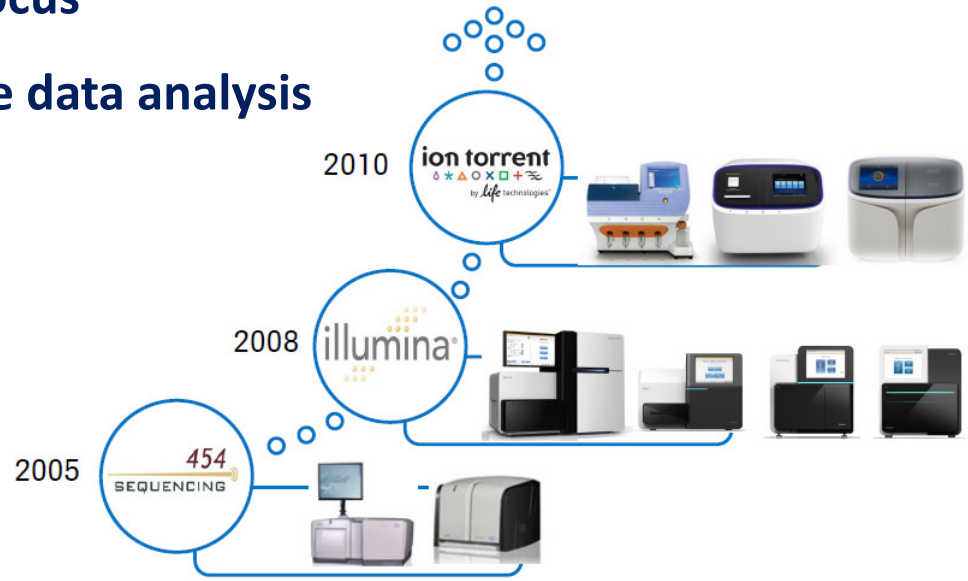
Annual statistics of research articles on Pubmed in the fields of (A) general NGS; (B) NGS for forensics; and (C) NGS for forensic STR genotyping

NGS-based STR:

- Longer sequencing read length
- Higher error rate for repetitive sequence
- Tens of full resolution alleles in each locus
- Complicated in data analysis, more parameters to be optimized

NGS-based SNP:

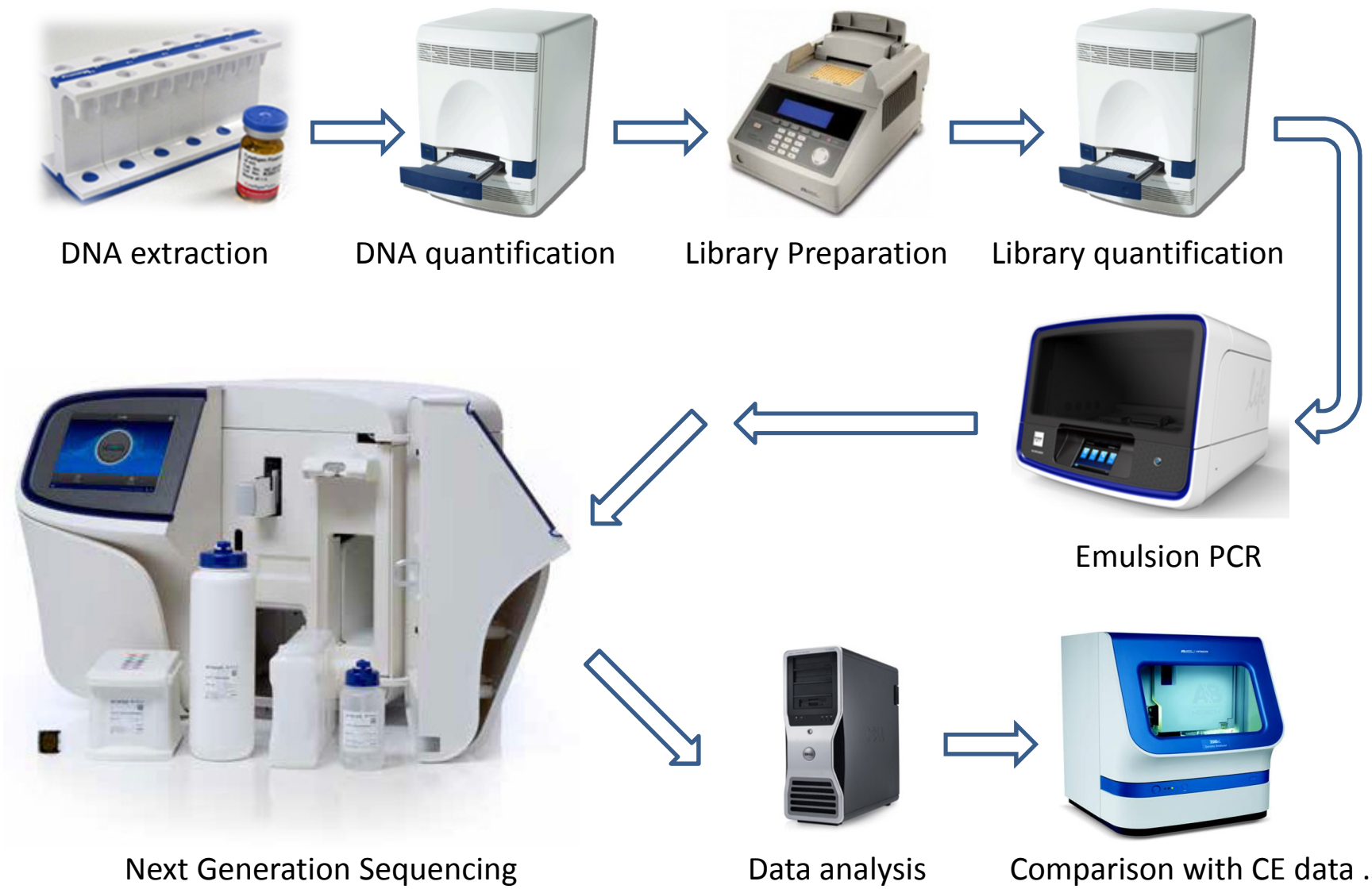
- Short-read
- Less repetitive sequence
- Two or three alleles in each locus
- Mature data analysis tools



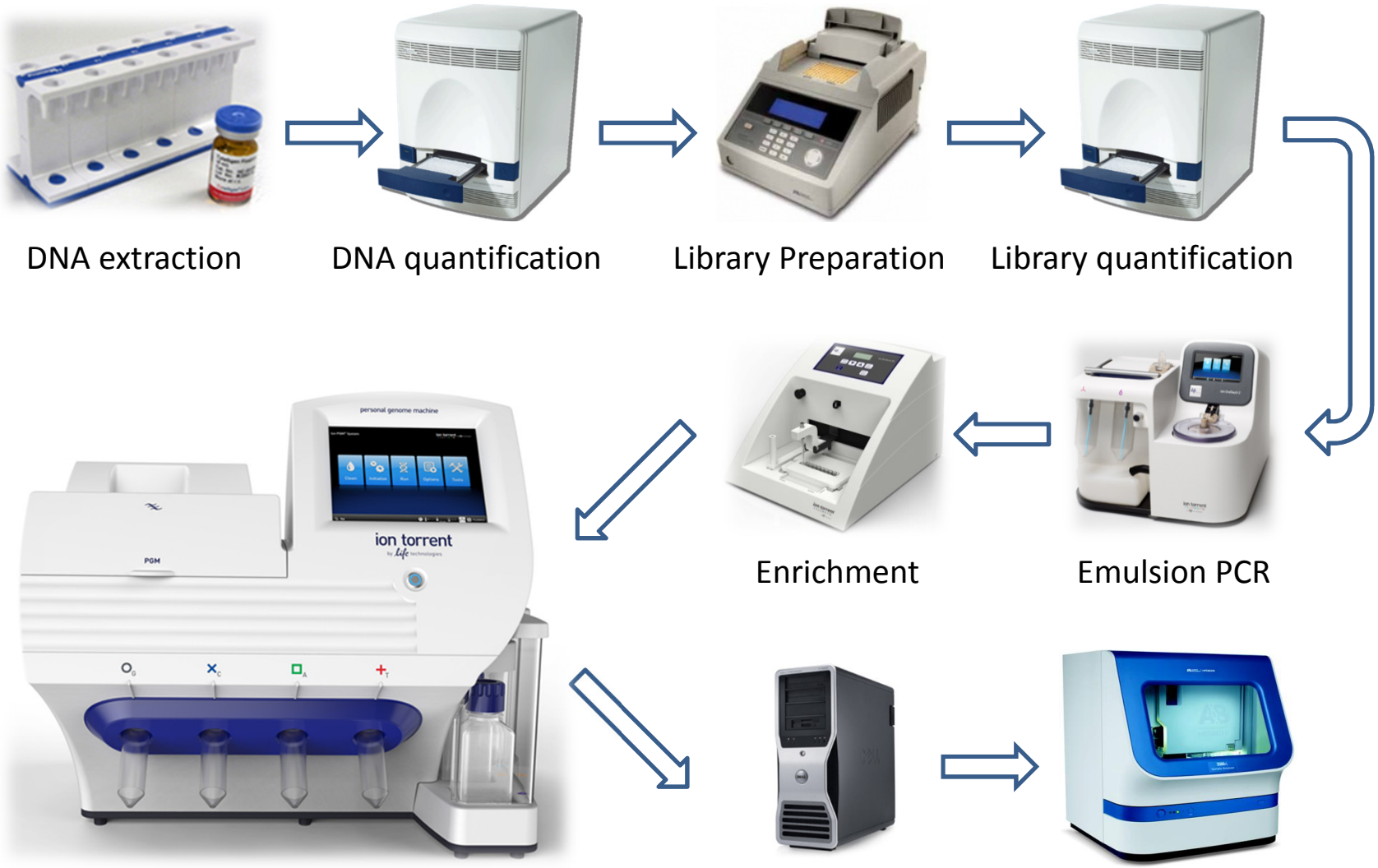
Literature summary for NGS-STR

- 20+ scientific publications from 10+ independent research groups
  - **Platforms:** 454 (2011-2014), Illumina Miseq™ and Ion Torrent™ (2015 till now)
  - **STR markers:** 48 autosomal STR, 29 Y-STR, 9 X-STR
  - **Data analysis software:** STRait Razor, MyFLq, STRinNGS
  - **Samples:** 9947A, 2800M, K562, blood, saliva, buccal swab, semen, vaginal secretion, skin, bone and muscle
  - **Population:** African American, Caucasian, Hispanic, Dane and Chinese Han
- Commercial solutions

Workflow Chart

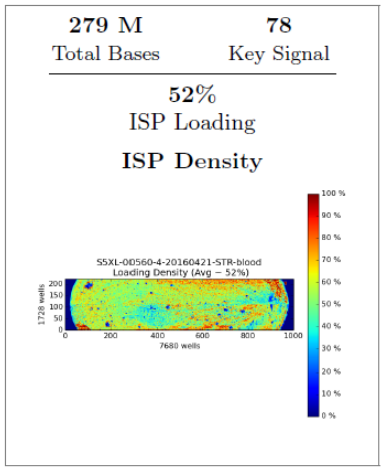


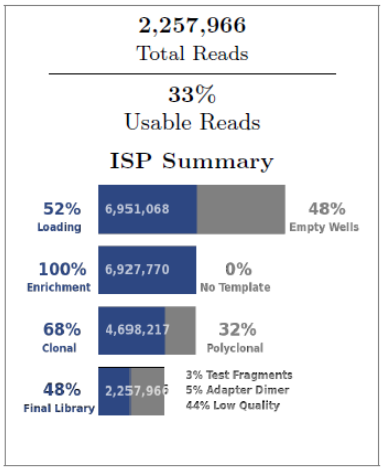
Workflow Chart



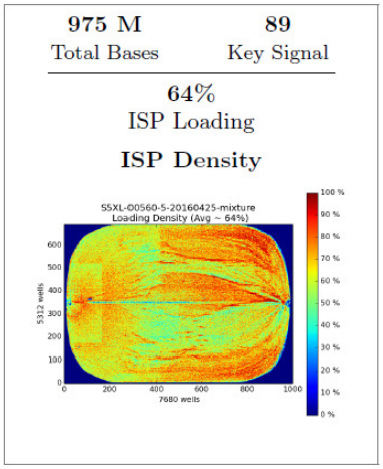
Data Quality

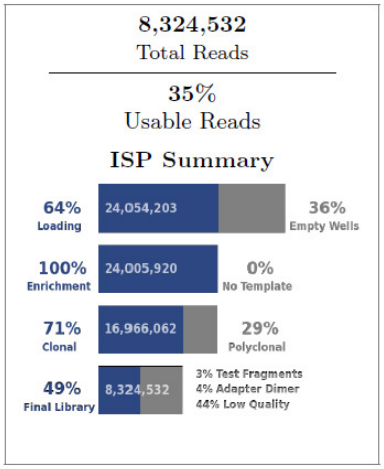
Run Summary for Ion 520™ CHIP





Run Summary for Ion 530™ CHIP





Data Quality

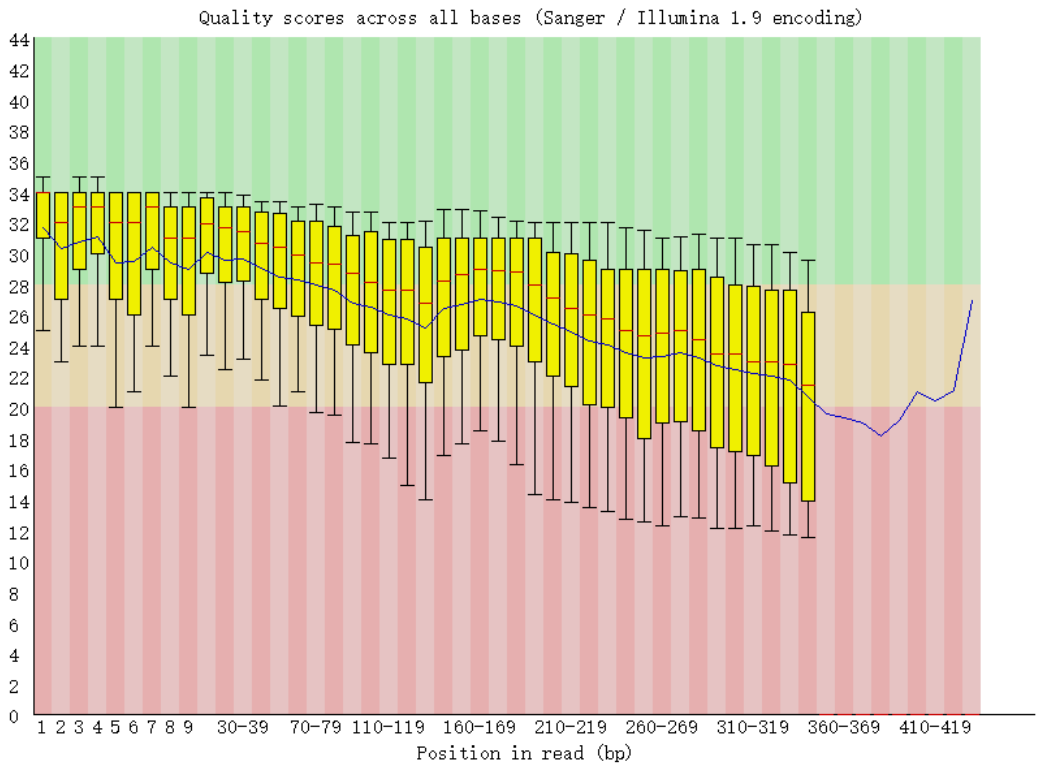


Fig. 1 Overview of the range of quality values across all bases at each position in the FastQ file.

The central red line is the median value.

The yellow box represents the inter-quartile range (25-75%).

The upper and lower whiskers represent the 10% and 90% points.

The blue line represents the mean quality.

Reference Samples: All yielded concordant genotypes with the PCR-CE results

Sample	Alleles (%)	Stutters (%)	Noise (%)
9947A	78.83	4.05	17.12
1	75.24	4.34	20.42
2	76.34	4.94	18.72
3	75.41	5.75	18.84
4	73.68	6.00	20.32
5	74.91	5.82	19.27
6	77.66	4.45	17.89
7	74.85	5.65	19.50
8	73.78	5.96	20.26
9	76.37	4.29	19.34
10	77.19	3.98	18.83
11	74.46	5.85	19.69
Average	75.73	5.09	19.18

Table 1 Percentage of reads representing alleles, stutters (including backstutters) and noise for 12 reference samples

Reference Samples

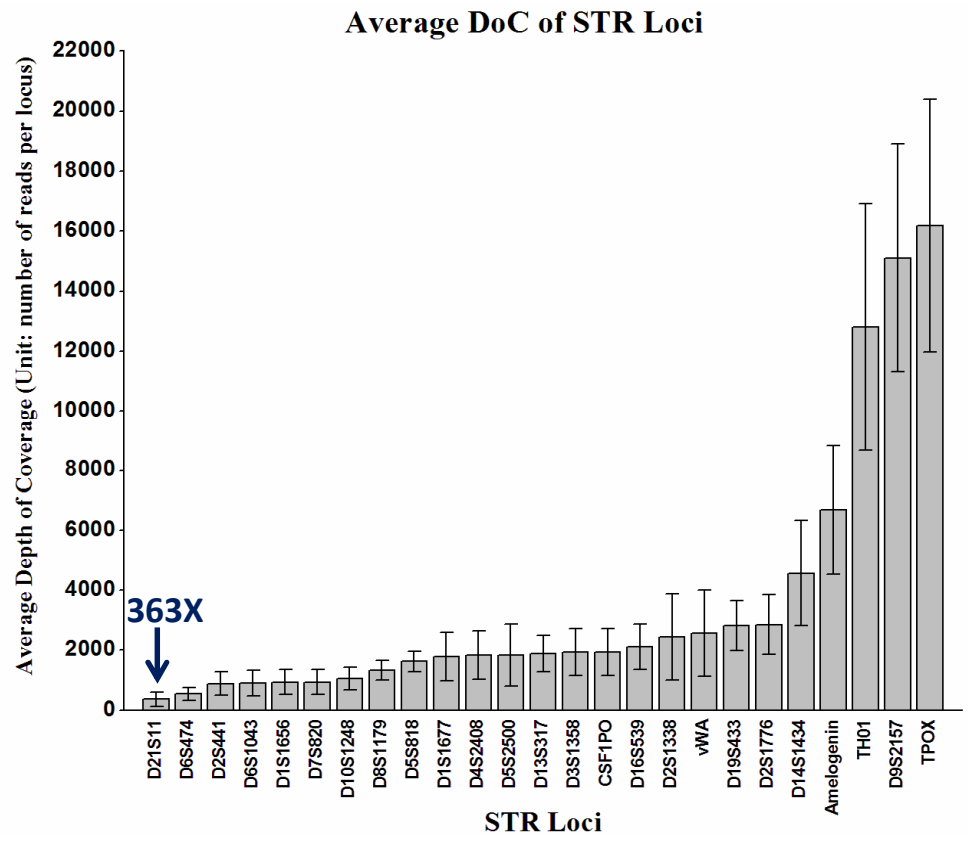


Fig. 2 Average DoC (Depth of Coverage) for the 12 reference samples amplifiedwith one ng of DNA. Error bars represent standard deviation.

Reference Samples

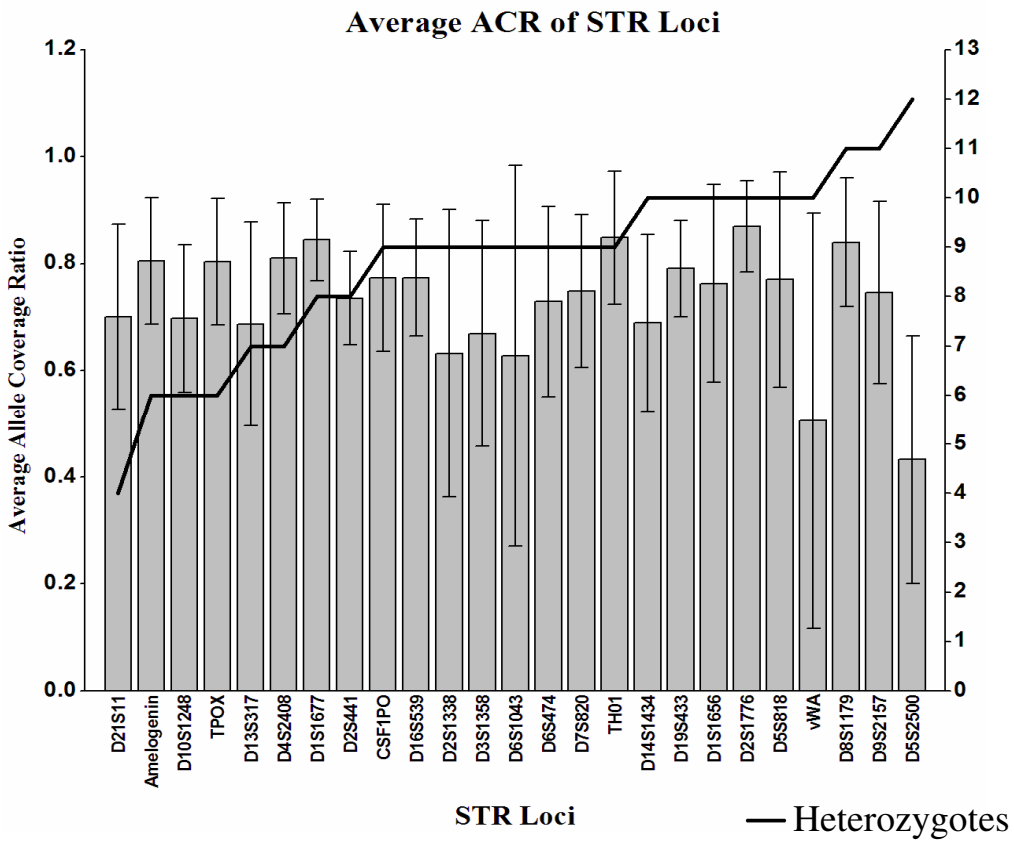


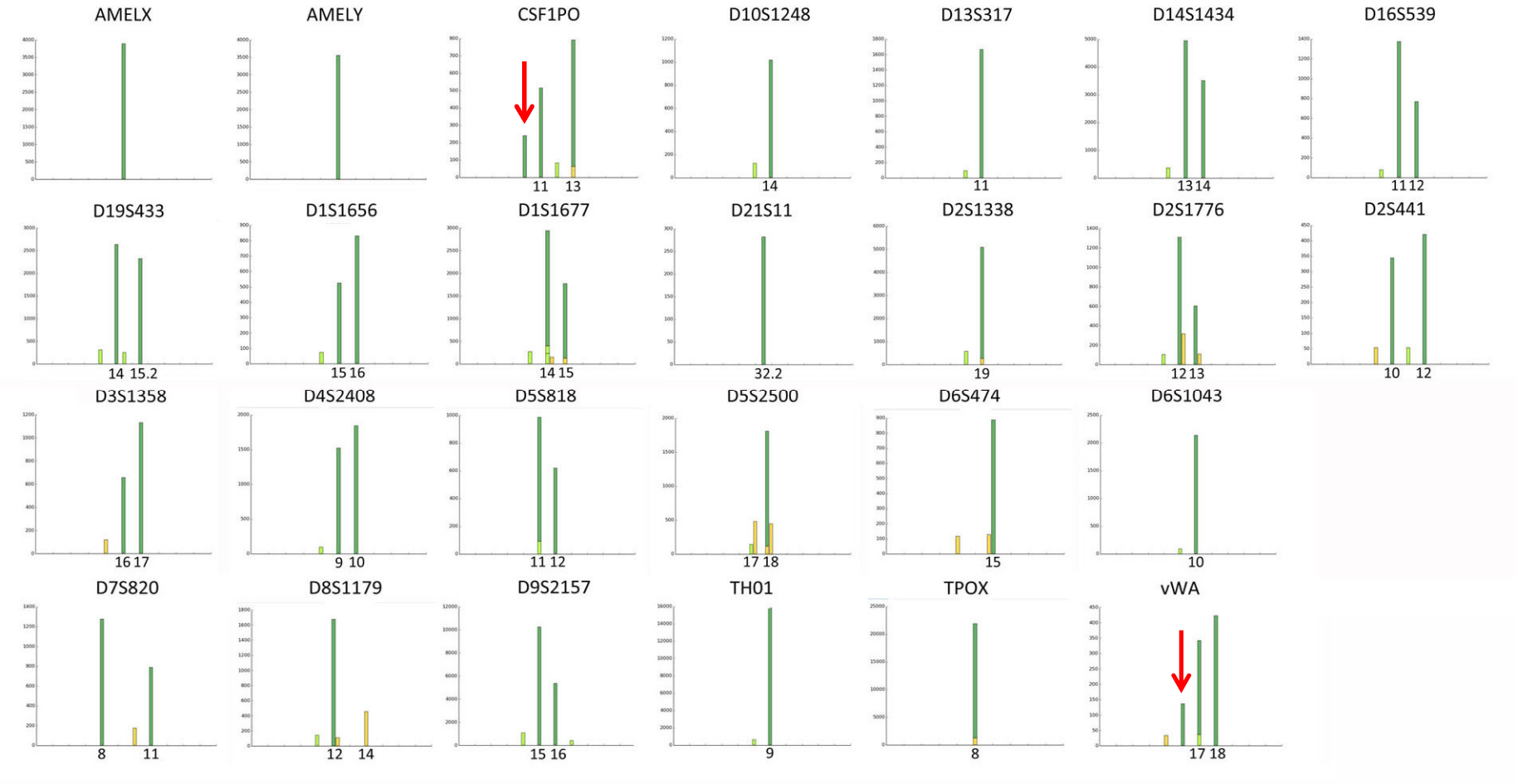
Fig. 3 Average ACRs (Allele Coverage Ratio) for the 12 reference samples amplified with one ng of DNA. ACRs were calculated by dividing the lower coverage allele by the higher coverage allele at that locus, with 1.0 indicating equal coverage. Horizontal black line indicates the number of heterozygotes for each calculated ACR. Error bars represent standard deviation.

Sensitivity

Table2 Number of alleles observed for dilution series of two control samples

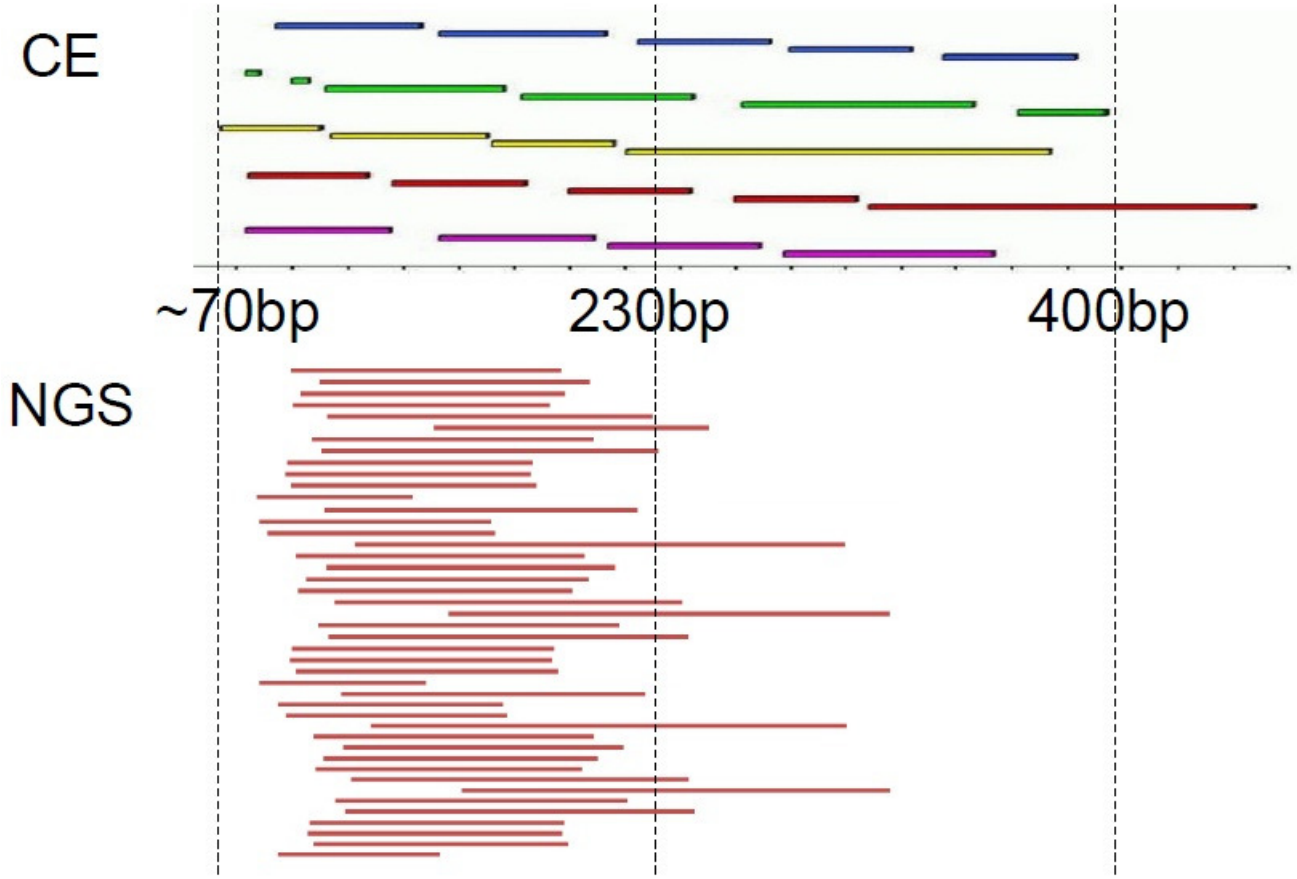
	Alleles Observed		Alleles Observed
9947A-500 pg	43 (100%)	R1-500 pg	42 (100%)
9947A-250 pg	43 (100%)	R1-250 pg	42 (100%)
9947A-125 pg	43 (100%)	R1-125 pg	42 (100%)
9947A-62.5 pg	43 (100%)	R1-62.5 pg	42 (100%)
9947A-31 pg	39 (90.70%)	R1-31 pg	36 (85.71%)
9947A-16 pg	29 (67.44%)	R1-16 pg	35 (83.33%)

Sensitivity



Reference sample No.1 _62.5 pg genomic DNA

Almost all STRs are mini-STRs with NGS



Degraded Samples

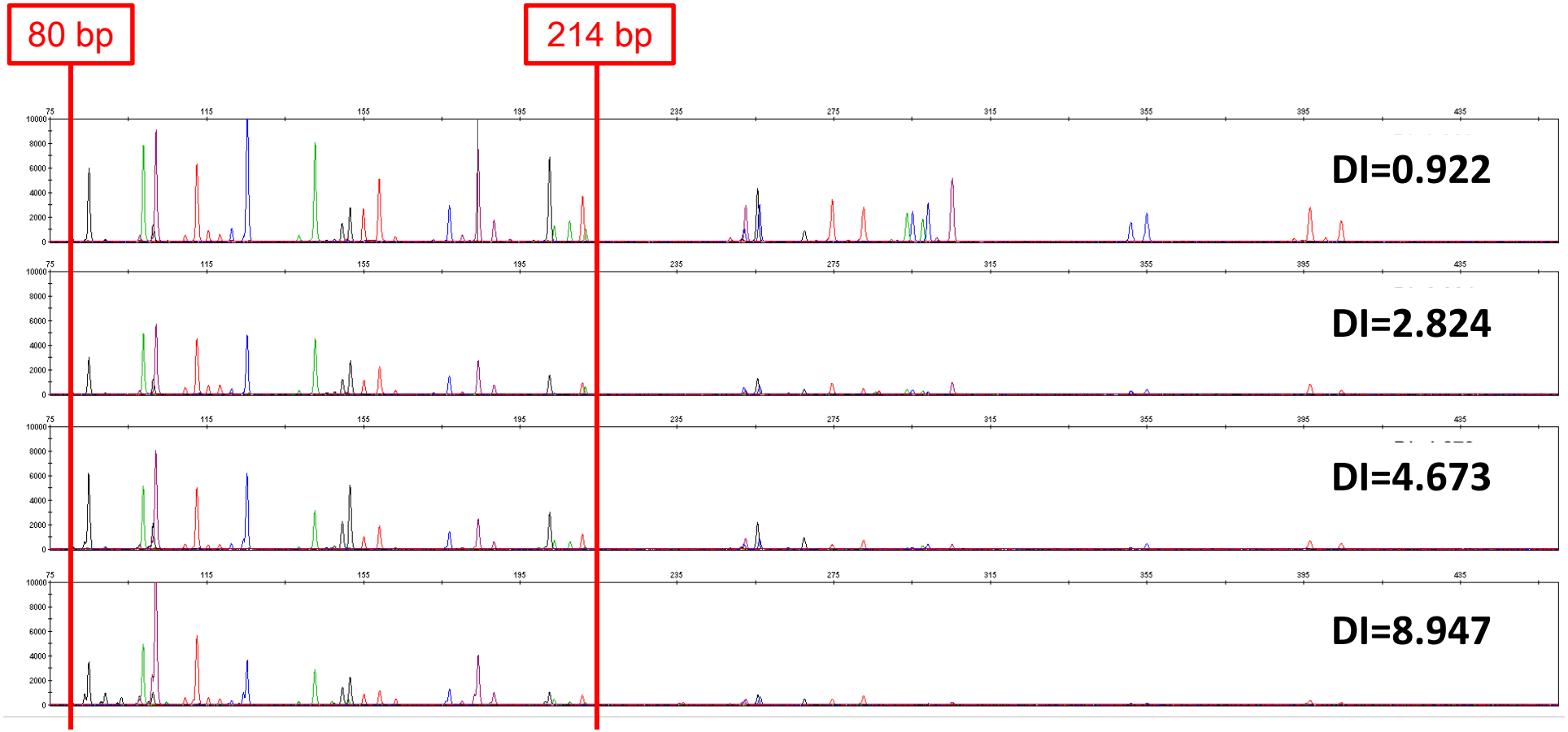
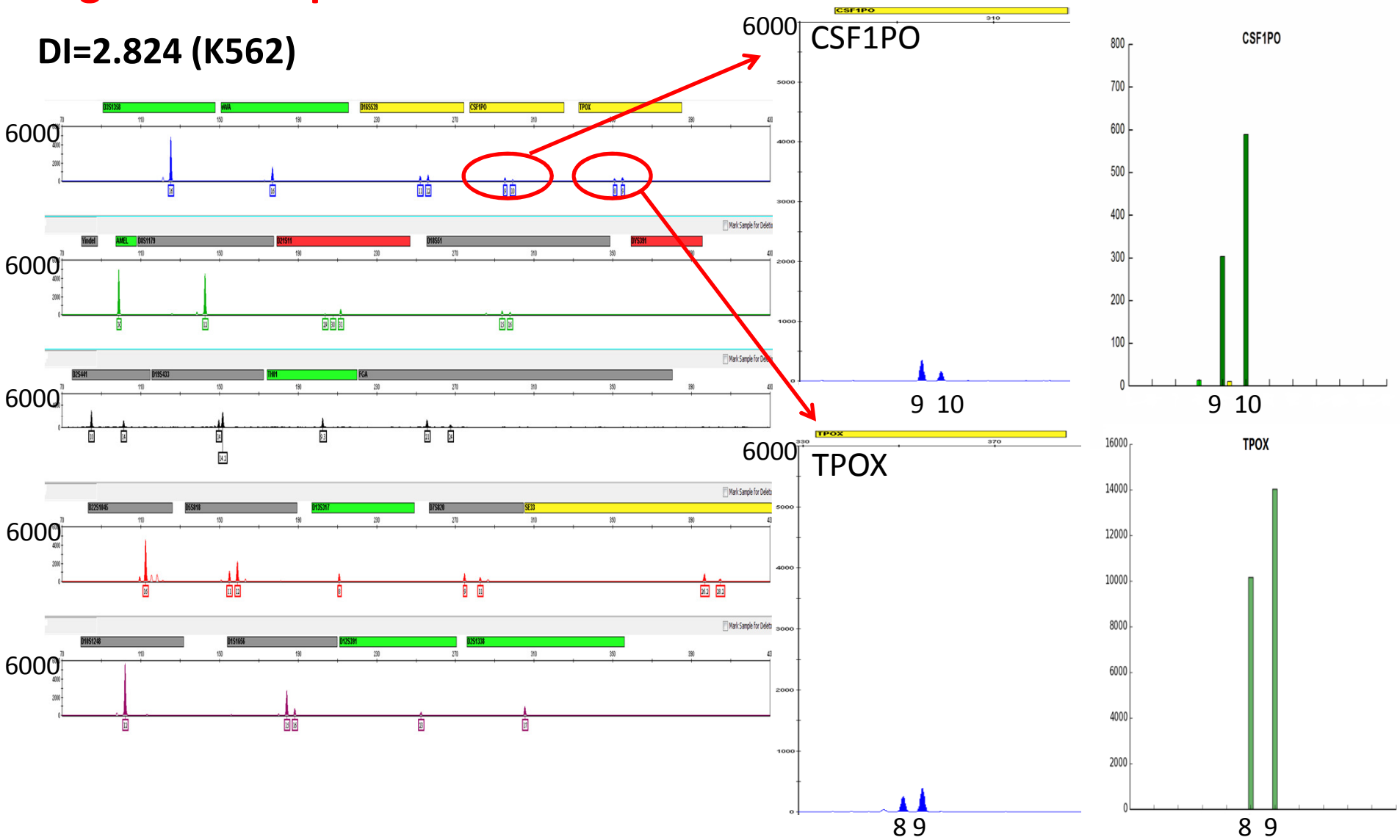


Fig. 5 K562 genomic DNA digested with DNase I for serial time durations to mimic different degradation levels of forensic biological samples. DI (Degradation Index) was calculated by dividing the small probe (80 bp amplicon) concentration by the larger one (214 bp amplicon).

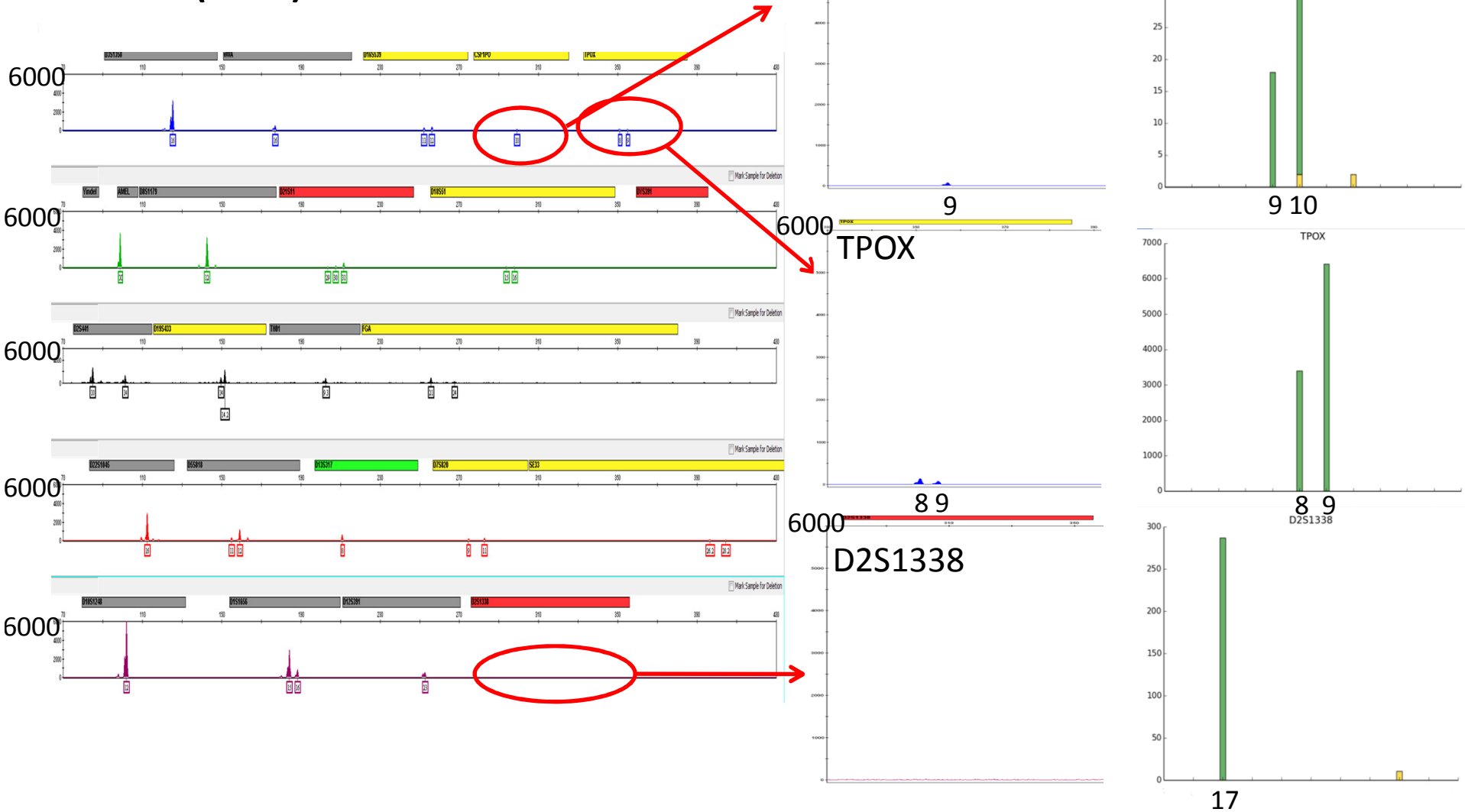
Degraded Samples

DI=2.824 (K562)



Degraded Samples

DI=8.947(K562)

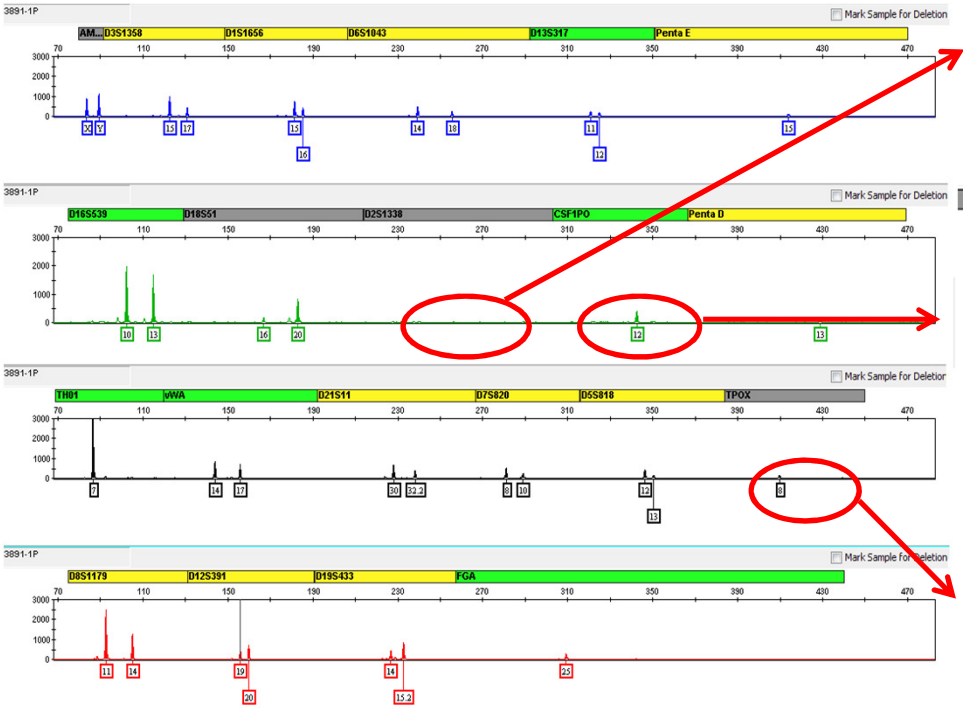


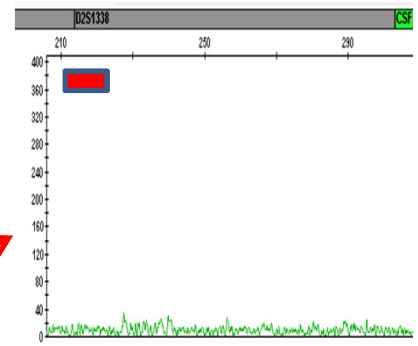
Casework Samples

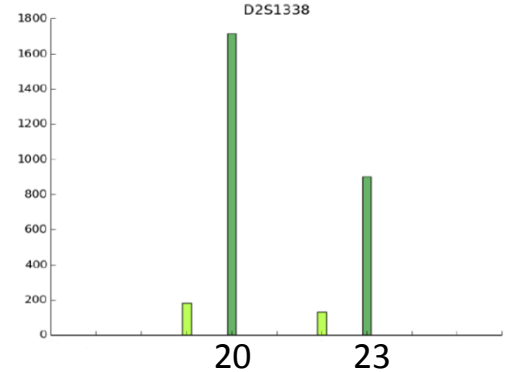
Sample	Description	Con. (ng/μL)	DI	Globalfiler	PP21	NGS
Case 1	Bone from anonymous body	0.014	1.507	N/A	Partial	Full
Case 2	Bone from anonymous body	0.03	2.139	N/A	Partial	Full
Case 3	Bone from anonymous body	0.045	1.799	N/A	Partial	Full
Case 4	Bone from stolen body	0.097	3.324	N/A	Partial	Full
Case 5	Cigarette butt taken from homicide crime scene	0.1174	4.916	Partial	N/A	Full
Case 6	Touch DNA taken from clothes	1.345	1.588	Full	N/A	Full
Case 7	Touch DNA taken from clothes	1.81	1.884	Full	N/A	Full
Case 8	Buccal swab	3.15	0.761	Full	N/A	Full
Case 9	Nail from anonymous body	5.21	0.837	Full	N/A	Full
Case 10	Nail from anonymous body	10.295	0.721	Full	N/A	Full

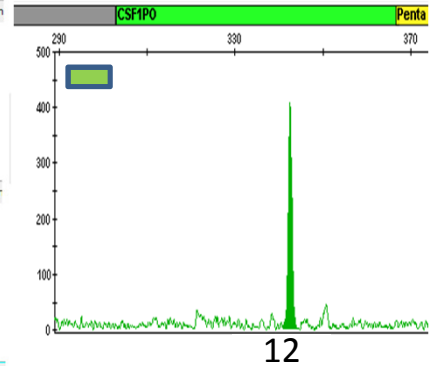
Casework Samples

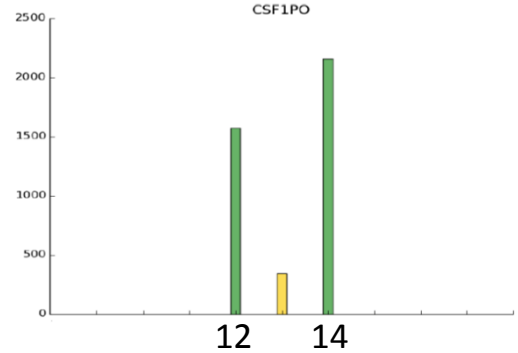
Case 1: 6 μL of template DNA in 10 μL PCR reaction mixture, 30 cycles

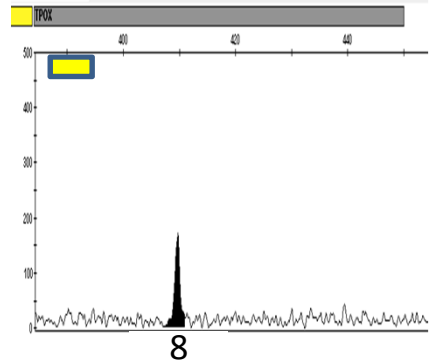


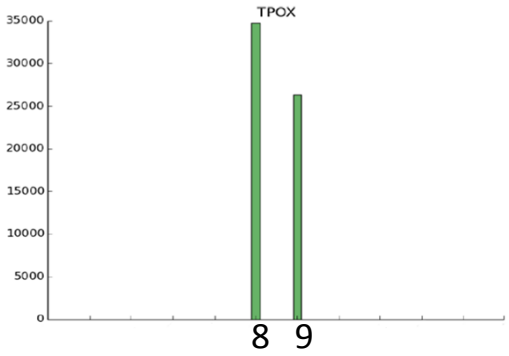








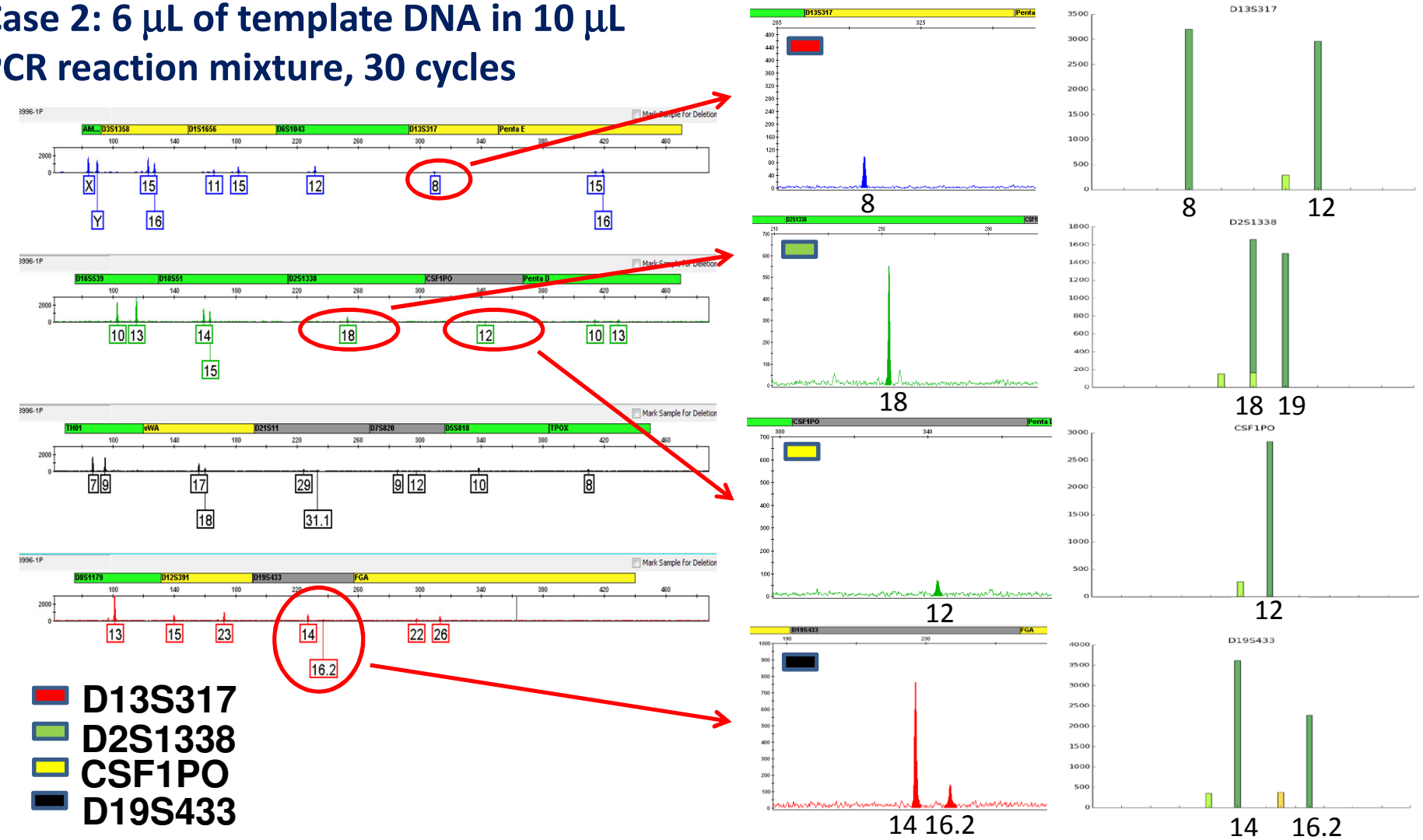




D2S1338
CSF1PO
TPOX

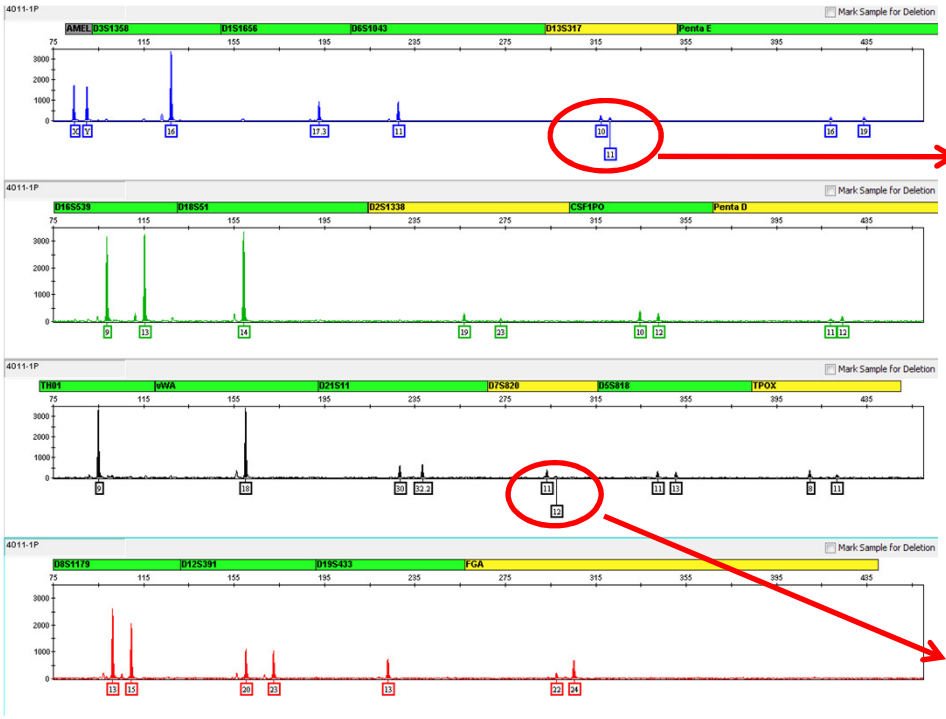
Casework Samples

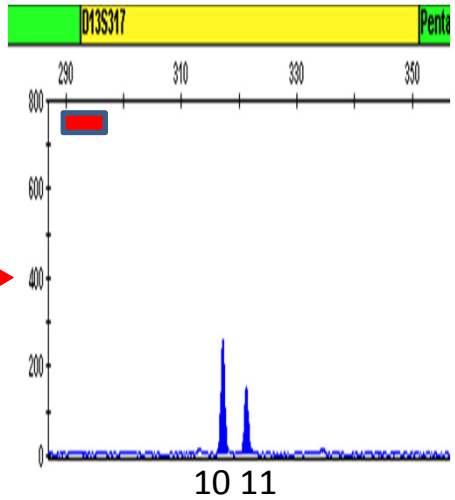
Case 2: 6 µL of template DNA in 10 µL PCR reaction mixture, 30 cycles

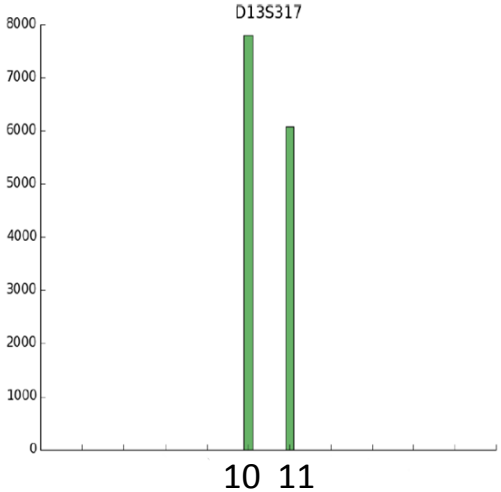


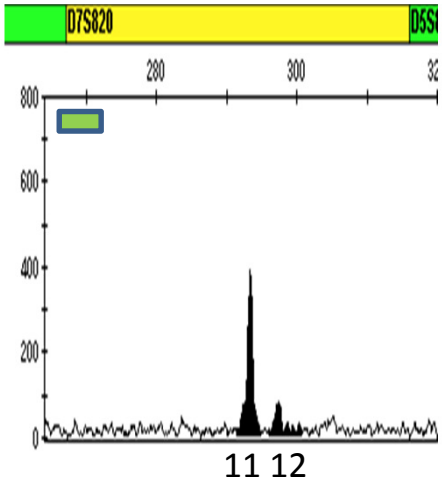
Casework Samples

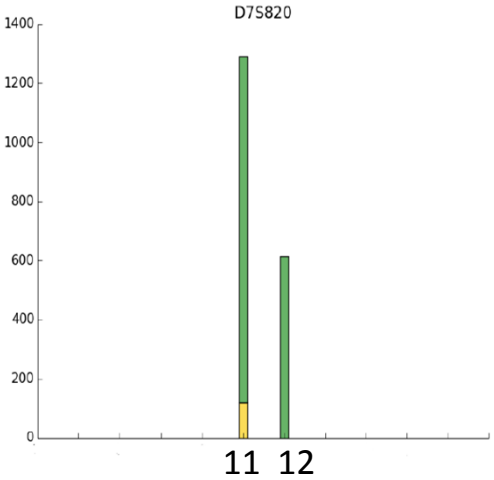
Case 3: 6 μL of template DNA in 10 μL PCR reaction mixture, 30 cycles









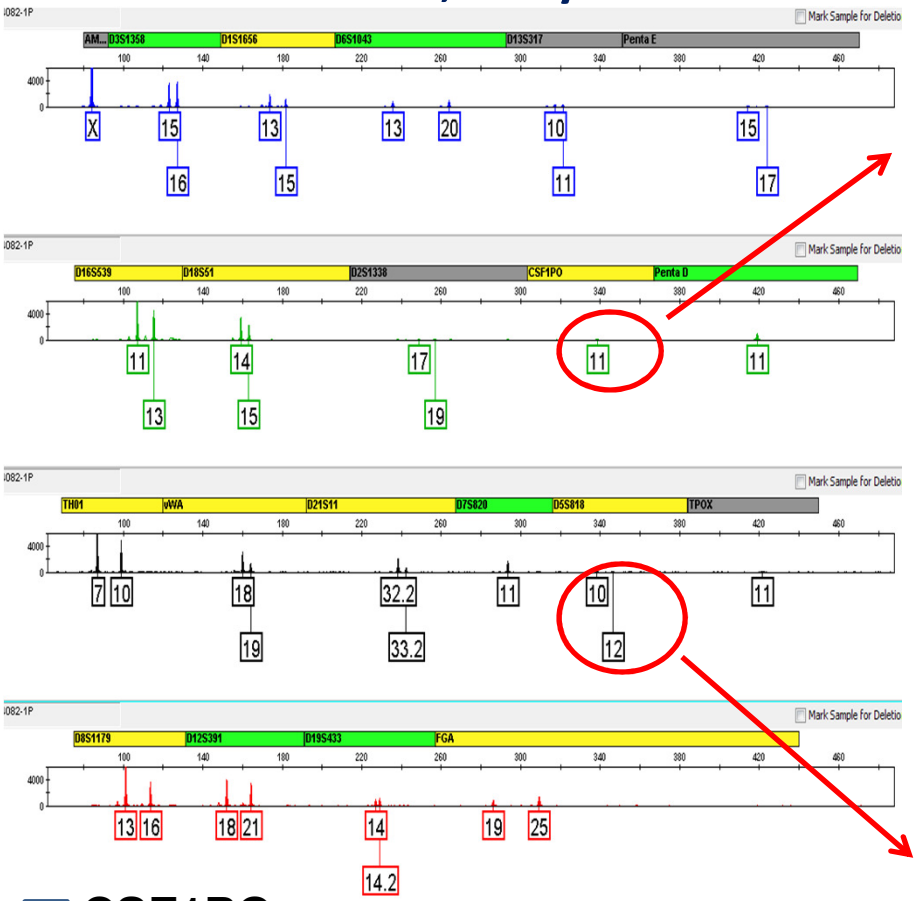


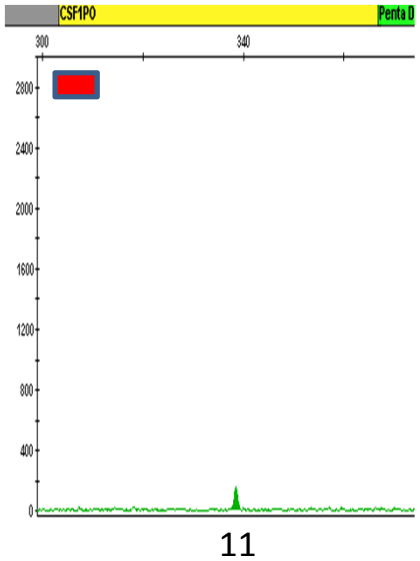
D13S317

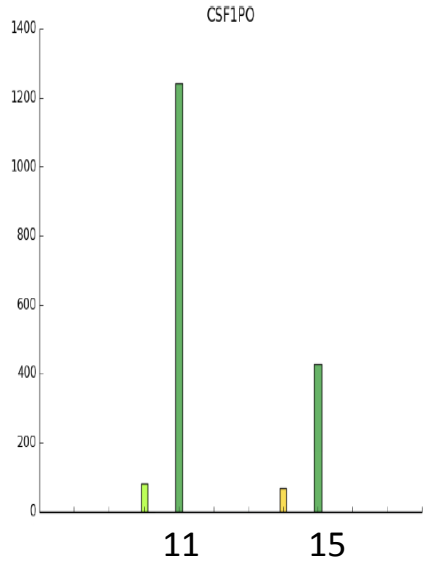
D7S820

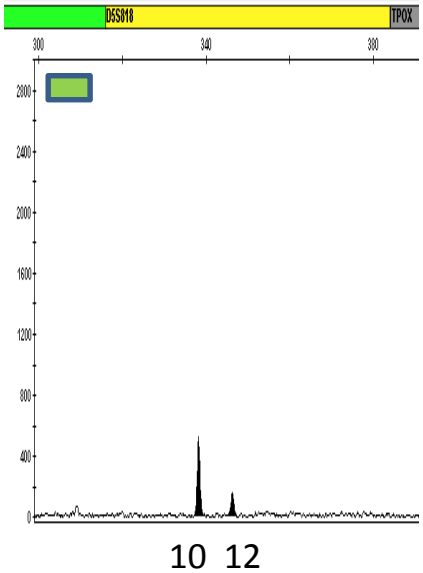
Casework Samples

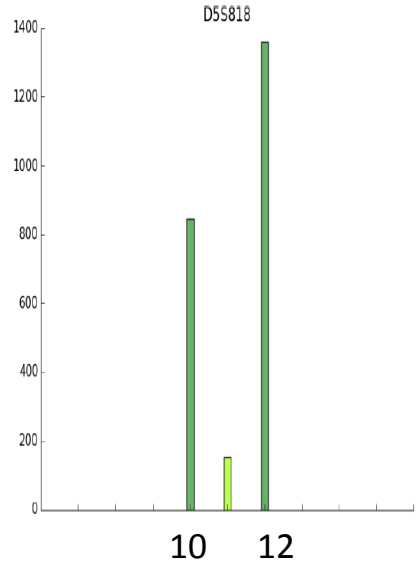
Case 4: 6 µL of template DNA in 10 µL PCR reaction mixture, 30 cycles









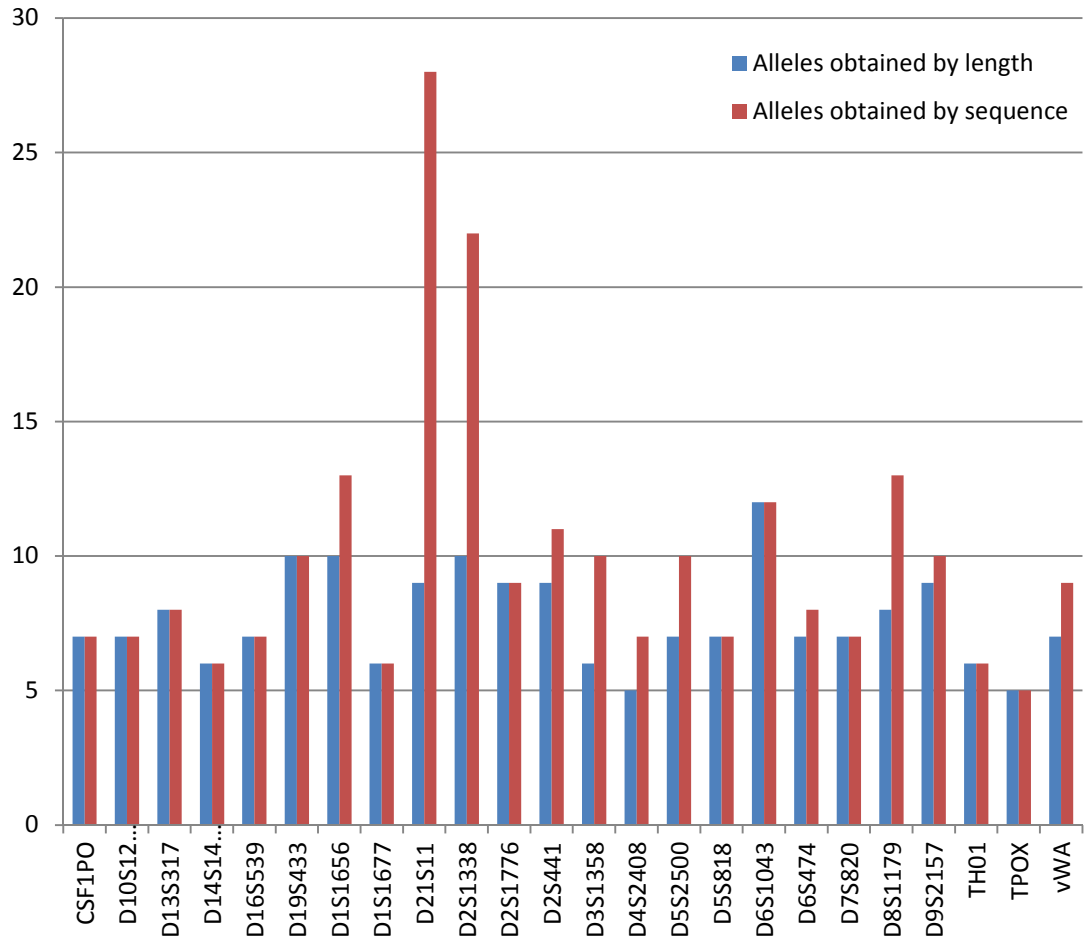


CSF1PO

D5S818

Population Statistics (N=73)

Locus	Alleles obtained by length	Alleles obtained by sequence
CSF1PO	7	7
D10S1248	7	7
D13S317	8	8
D14S1434	6	6
D16S539	7	7
D19S433	10	10
D1S1656	10	13
D1S1677	6	6
D21S11	9	28
D2S1338	10	22
D2S1776	9	9
D2S441	9	11
D3S1358	6	10
D4S2408	5	7
D5S2500	7	10
D5S818	7	7
D6S1043	12	12
D6S474	7	8
D7S820	7	7
D8S1179	8	13
D9S2157	9	10
TH01	6	6
TPOX	5	5
vWA	7	9



Population Statistics

Sample Size Requirements for Addressing the Population Genetic Issues of Forensic Use of DNA Typing

RANAJIT CHAKRABORTY[1]

Abstract DNA typing offers a unique opportunity to identify individuals for medical and forensic purposes. Probabilistic inference regarding the chance occurrence of a match between the DNA type of an evidentiary sample and that of an accused suspect, however, requires reliable estimation of genotype and allele frequencies in the population. Although population-based data on DNA typing at several hypervariable loci are being accumulated at various laboratories, a rigorous treatment of the sample size needed for such purposes has not been made from population genetic considerations. It is shown here that the loci that are potentially most useful for forensic identification of individuals have the intrinsic property that they involve a large number of segregating alleles, and a great majority of these alleles are rare. As a consequence, because of the large number of possible genotypes at the hypervariable loci that offer the maximum potential for individualization, the sample size needed to observe all possible genotypes in a sample is large. In fact, the size is so large that even if such a huge number of individuals could be sampled, it could not be guaranteed that such a sample was drawn from a single homogeneous population. Therefore adequate estimation of genotypic probabilities must be based on allele frequencies, and the sample size needed to represent all possible alleles is far more reasonable. Further economization of sample size is possible if one wants to have representation of only the frequent alleles in the sample, so that the rare allele frequencies can be approximated by an upper bound for forensic applications.

Table 4. Conservative Estimates of the Number of Individuals Needed to Represent All Possible Genotypes in a Sample for a *k*-Allelic Codominant Locus

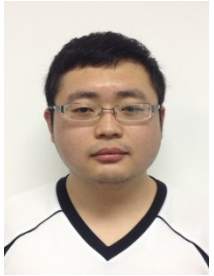
	Minimum Sample Size[a] *Needed for:*				
k	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.25$
5	213	156	116	99	77
10	921	691	532	465	379
15	2,164	1,647	1,288	1,137	944
20	3,962	3,043	2,406	2,137	1,794
25	6,330	4,893	3,899	3,479	2,943
30	9,279	7,210	5,778	5,174	4,402
40	16,955	13,278	10,733	9,659	8,287
50	27,051	21,305	17,329	15,651	13,507
100	115,134	92,153	76,248	69,539	60,970

a. Sample size in these computations refers to the number of individuals to be typed for each *k*-allelic codominant locus. The values of $\alpha = 0.001$, 0.01, 0.05, 0.10, and 0.25 represent 99.9%, 99%, 95%, 90%, and 75% confidence, respectively, of being assured that all possible genotypes are represented in the sample.

Source: Human Biology, Vol. 64. No. 2 (April 1992). pp. 141-159

Acknowledgements

- **Group members**



Man Chen **Bo Wu** **Yi-Cheng Liu**

- **Guang-Feng Zhang and Xiu-Lan Xu for providing the casework samples**

- **Generous and reliable support from ThermoFisher Scientific, special thanks to WenWen Tang, Qingqing Zhang and Mao-Fen Yan for technical assistance**

Thank you for your attention!

Le Wang

wangle_02@163.com

Institute of Forensic Science

Ministry of Public Security

P. R. China

The Ion S5™ and Ion Chef™, Ion 520™ Chip & Ion 530™ are for Research, Forensic or Paternity Use Only.
Miseq is a trademarke of Illumina Inc.

Speaker was provided travel and hotel support by Thermo Fisher Scientific for this presentation, but no remuneration