Get more information from challenging samples with next-generation sequencing of short tandem repeats

In this application note, we demonstrate:

- Usability of the Applied Biosystems™ Precision ID Next-Generation Sequencing (NGS) System and Converge™ Software v2.0.1 with NGS Data Analysis module v1.0 for short tandem repeat (STR) analysis in forensic DNA laboratories
- Performance of the Applied Biosystems™ Precision ID GlobalFiler™ NGS STR Panel v2 with respect to genotype concordance, sensitivity, coverage distribution, intralocus balance, and observed stutter for 320 reference population samples
- Increased discrimination and limit of detection (LOD) of NGS technology over traditional capillary electrophoresis (CE) methods with challenging mock casework samples of types commonly encountered in forensic laboratories [1-4].

Performance of Precision ID NGS System and Precision ID GlobalFiler NGS STR Panel v2

Fragment analysis of STRs with CE technology remains the gold standard for human identification testing in forensic laboratories. However, the ability of the CE STR technology to address difficult casework samples can be challenged by limited DNA quantity, degraded DNA, and sample mixtures. Recent advances in NGS technology demonstrate the potential to surpass the limits of CE-based fragment analysis and perform routine identification of missing persons and criminal casework when CE techniques fall short.

To evaluate the performance of the Precision ID NGS System, a set of 320 known reference samples from diverse populations was analyzed with the Precision ID GlobalFiler NGS STR Panel v2 and Converge Software with NGS Data Analysis module v1.0 to assess a number of key performance metrics (e.g., depth of coverage, intralocus balance, sensitivity, genotype concordance) and overall NGS system usability. Data were compared to the results obtained using the CE-based Applied Biosystems™ GlobalFiler™ PCR Amplification Kit where applicable, to measure concordance with the established methods for analyzing DNA fragment length. Additionally, five global test sites were selected to analyze a range of forensic casework specimens (N = 80) by both CE STR and NGS STR methods, to assess the utility of an early-access version of the Precision ID GlobalFiler NGS STR Panel v2 with challenging forensic samples. The data obtained in these studies demonstrated robust sensitivity (as low as 125 pg genomic DNA) and high marker concordance (98.96%) with the orthogonal CE data set. Results obtained from a selected group of challenging mock casework samples (e.g., vaginal swab at 37°C for 2 years, contact DNA swab from doorknob, and bone samples buried in soil for up to 18 months) highlight the ability of NGS STR analysis to generate more complete STR profiles and obtain additional genetic information with difficult samples when compared to fragment length analysis.

Materials and methods

DNA extraction and quantitation

Genomic DNA (gDNA) samples were extracted from 320 known reference samples using the Applied Biosystems™ PrepFiler™ *Express* Forensic DNA Extraction Kit. Quantification was performed with the Applied Biosystems™ Quantifiler™ Trio DNA Quantification Kit on the Applied Biosystems™ 7500 Real-Time PCR System. The methods were based on our commercially available recommendations.

NGS workflow: library preparation, template preparation, sequencing, and data analysis

The Precision ID GlobalFiler NGS STR Panel v2 includes 31 autosomal STRs, 1 Y-STR, amelogenin, Y-indel rs2032678, and a target in the *SRY* gene (Table 1).

The STRs were designed to contain small amplicons (~100–300 bp) in order to maximize allele recovery and depth of coverage. Reference samples were amplified using the Precision ID GlobalFiler NGS STR Panel v2 and Applied Biosystems™ Precision ID DL8 Kit on the Ion Chef™ System with 1 ng of DNA input (24 cycles with 4 min for annealing and extension). The resulting library pools were quantified using the Ion Library TaqMan® Quantitation Kit. Pooled libraries were templated with the Ion S5™ Precision ID Chef & Sequencing Kit, and sequence analysis was performed using either an Ion 520™ or Ion 530™ Chip on the Ion S5™ System. Primary analysis utilized Torrent Suite Software v5.6 and the Human Identification (HID) STR Genotyper Plugin, which contains optimized, targeted analysis algorithms for STRs on Converge Software v2.0.1 with NGS Data Analysis module v1.0.

Table 1. Markers in the Precision ID GlobalFiler NGS STR Panel v2 and allele number comparison with CE from analysis of 320 population samples. Note: An early-access version of the panel for evaluation at the test sites did not include the Penta D, Penta E, or SRY markers.

STR	Repeat structure	Source	Chromosome	CE alleles	NGS alleles	NGS alleles/CE alleles (%)
TPOX	AATG	CODIS	2	8	9	113
D3S1358	TCTA/TCTG	CODIS	3	9	21	233
FGA	CTTT/TTCC	CODIS	4	22	30	136
CSF1PO	AGAT	CODIS	5	8	9	113
D5S818	AGAT	CODIS	5	8	10	125
D7S820	GATA	CODIS	7	7	7	100
D8S1179	TCTA/TCTG	CODIS	8	10	26	260
TH01	TCAT	CODIS	11	7	7	100
vWA	TCTA/TCTG	CODIS	12	10	26	260
D13S317	TATC	CODIS	13	9	9	100
D16S539	GATA	CODIS	16	7	8	114
D18S51	AGAA	CODIS	18	15	19	127
D21S11	TCTA/TCTG	CODIS	21	17	55	324
AMEL-X	NA	Sex determination	X	NA	NA	NA
AMEL-Y	NA	Sex determination	Y	NA	NA	NA
rs2032678	NA	Sex determination	Y	NA	NA	NA
SRY	NA	Sex determination	Y	NA	NA	NA
D1S1656	TAGA	Expanded CODIS	1	15	27	180
D2S441	TCTA/TCAA	Expanded CODIS	2	8	14	175
D2S1338	TGCC/TTCC	Expanded CODIS	2	15	56	373
D10S1248	GGAA	Expanded CODIS	10	10	12	120
D12S391	AGAT/AGAC	Expanded CODIS	12	19	69	363
D19S433	AAGG/TAGG	Expanded CODIS	19	17	19	112
D22S1045	ATT	Expanded CODIS	22	10	10	100
DYS391	TCTA	Expanded CODIS	Y	NA	NA	NA
D1S1677	TTCC	Non-CODIS	1	9	11	122
D2S1776	AGAT	Non-CODIS	2	9	9	100
D3S4529	ATCT	Non-CODIS	3	8	12	150
D4S2408	TCT	Non-CODIS	4	7	8	114
D5S2800	GATA/GATT	Non-CODIS	5	9	15	167
D6S474	GATA/GACA	Non-CODIS	6	7	11	157
D6S1043	AGAT/AGAC	Non-CODIS	6	15	26	173
D12ATA63	TAA/CAA	Non-CODIS	12	10	15	150
D14S1434	CTGT/CTAT	Non-CODIS	14	8	11	138
Penta D	AAAGA	Other STR	21	13	14	108
Penta E	AAAGA	Other STR	15	17	22	129
		Total alleles		**343**	**597**	**174**

NA: not applicable.

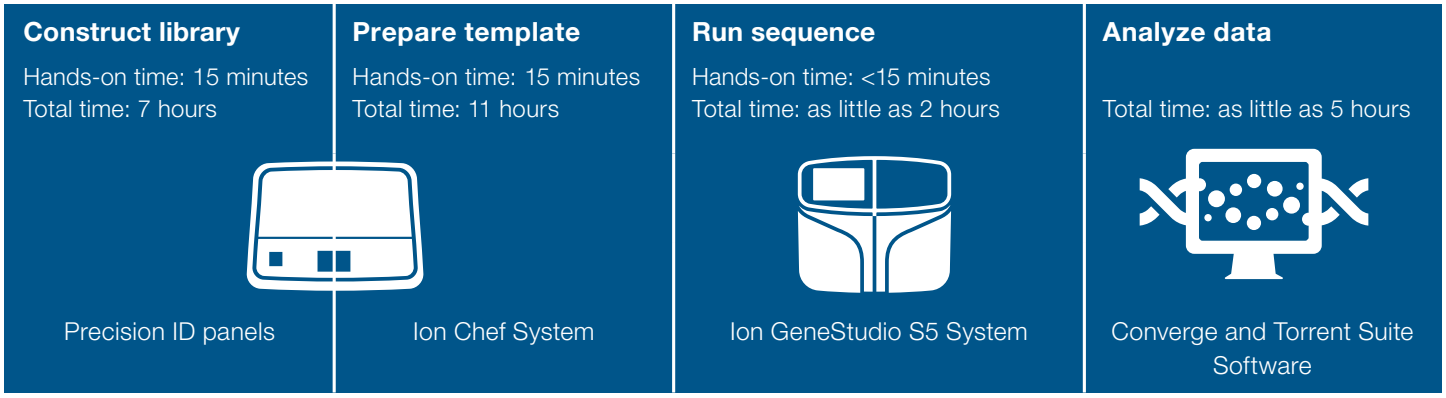

Figure 1. The Precision ID NGS System. The system comprises the Ion Chef System, Ion GeneStudio™ S5 series systems, Torrent Suite Software, and Converge Software with NGS Data Analysis module.

Population samples

A total of 1 ng of gDNA each was used to analyze 320 known African American and Caucasian reference samples using both the GlobalFiler PCR Amplification Kit and Precision ID GlobalFiler NGS STR Panel v2. Torrent Suite Software v5.6 was utilized for the NGS workflow in combination with the Converge Software v2.0.1 with NGS Data Analysis module v1.0.

Sensitivity and casework samples

For the sensitivity study, a range of gDNA inputs (1 ng, 250 pg, and 125 pg) were processed for known reference samples (N = 40) with both CE and NGS workflows to measure the percentage of truth alleles detected. Alleles were compared to either CE STR data, if available, or NGS-curated truth alleles.

For mock casework sample testing, each of the five laboratories in this study selected 16 samples to represent a range of challenging, mixed, and degraded specimens. All samples were quantified using the Quantifiler Trio kit, and 1 ng of total gDNA was targeted for CE and NGS workflows when available. Where applicable, samples genotyped with CE-based methods were amplified with the GlobalFiler PCR Amplification Kit for analysis on the Applied Biosystems™ 3500 Genetic Analyzer and GeneMapper™ *ID-X* Software v1.2 or higher.

Results

Usability and system performance

The Precision ID NGS System provides a robust, automated workflow for library preparation, template preparation, and massively parallel sequencing. DNA sequencing results can be obtained within a 2–3 day period with minimal hands-on time (Figure 1). The Ion Chef System also provides greater consistency with chip loading relative to the manual processing methods, and improved reproducibility of sequence data output. From a representative sample set of Ion S5 sequencing runs (N = 10), on average 7.2 million reads were obtained for 32 sample barcodes sequenced on the Ion 530 Chip (Figure 2). With this system, usable sequencing reads are maximized due to the overall consistency provided by the highly automated Ion Chef workflow.

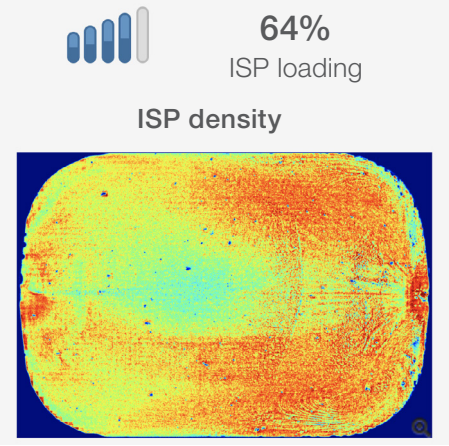

Figure 2. Typical chip-loading density for the Precision ID GlobalFiler NGS STR Panel v2 (~50–70%).

Relative read depth of coverage per marker (expressed as a percentage of total reads obtained) is shown in Figure 3. For the marker set, average read depth was 2,730x, with the greatest depth-of-coverage difference of 9.2-fold between the lowest-performing markers (Penta D, Penta E, and FGA) and the highest-performing markers (D16S539, D8S1179, TH01, and TPOX). The remaining markers in the panel exhibited similar coverage distribution.

Intralocus balance

Intralocus balance (ILB) was measured for heterozygote pairs using the 1 ng DNA samples (N = 40) in this study. With the exception of D12ATA63, the markers in the panel exhibited ILB values above 60% (Figure 4). Compared to CE-based STRs, NGS data were expected to show a larger range for the ILB across the marker set with an increase in the number of outlying data points. This result is due to the decreased success in the number of "end-to-end" sequence reads for larger alleles in a given heterozygote pair.

Stutter

For the 320 population samples analyzed, stutter percentages with the Precision ID GlobalFiler NGS STR Panel v2 averaged 13% of the truth allele peak. Data varied by marker from a low stutter value of 6.2% (TH01) to a high stutter percentage of 17.2% (D1S1656). In general, while the level of stutter is ~2-fold greater in NGS over CE, the pattern of stutter correlated to allele size was conserved (data not shown). Increased NGS stutter compared to CE methods results from the additional rounds of PCR cycling with the NGS assay.

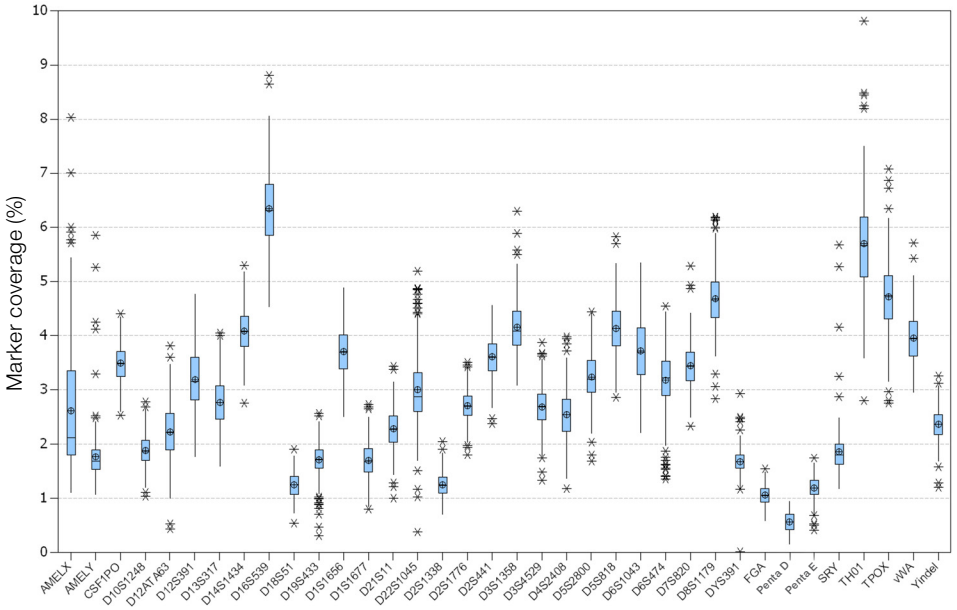

Figure 3. Coverage per marker with 1 ng gDNA input. Data are plotted by reads per marker divided by total reads for eight DNA samples over four Precision ID DL8 library preparations. The libraries were pooled prior to template preparation.

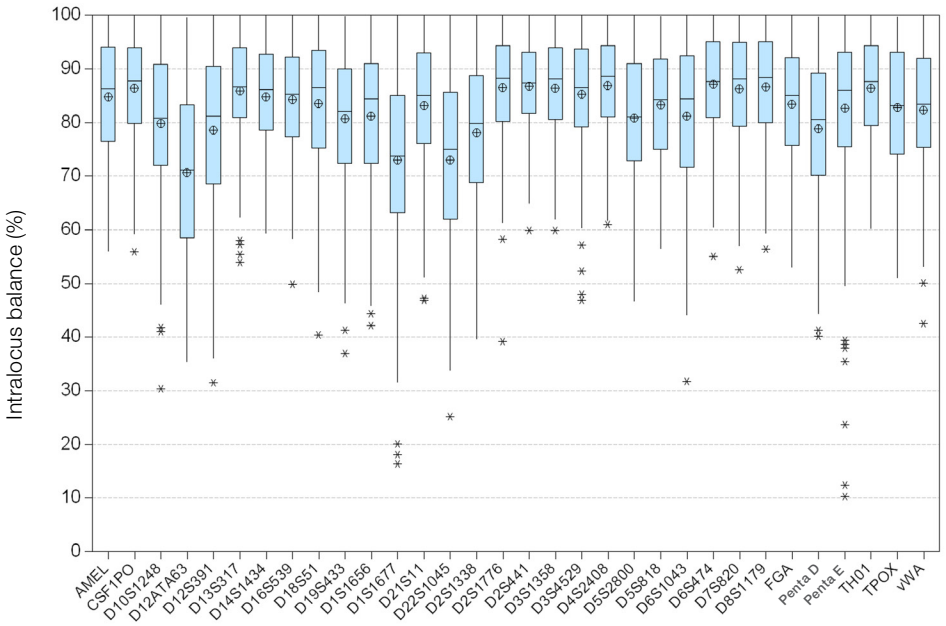

Figure 4. Intralocus balance (ILB) with 1 ng DNA input. ILB was calculated for heterozygote pairs using the 1 ng DNA input test samples in the study (ratio of minimum reads over maximum reads obtained in a heterozygote pair) for a total of 320 samples. Blue boxes show the middle 50% or interquartile range (IQR). "Whiskers" indicate 1.5 IQR from the upper and lower margins of the IQR. Asterisks are outlier data that are more than 1.5 IQR from the median.

Marker concordance with orthogonal CE data

Genotypes from 320 reference population samples from individuals of African American and Caucasian descent were compared with traditional CE analysis results. Orthogonal CE data were available for a total of 8,162 markers, of which 8,077 were accurately identified with NGS (98.96%) (Figure 5 and Table 2). A total of 45 artifacts were detected and 40 dropouts were identified with a stochastic threshold set at 5%. All dropouts (false negatives) occurred within the Penta D locus due to a 13 bp deletion adjacent to the start of the Penta D repeat structure for alleles 2.2 or 3.2, which occurs at a frequency of 11% in the African American population.

Of the 45 artifacts observed above the stochastic threshold, the most frequently affected STRs were as follows: 1) D12S391 and D10S1248 (base insertions at these markers caused discordance due to sequence complexity), and 2) nonreproducible artifacts at Penta D, Penta E, and D18S51 (due to single-base insertions and deletions). In addition, 5 of the observed discordances arose from the indels and/or single-nucleotide polymorphisms (SNPs) adjacent to the STR repeat in flanking-region sequences. Laboratories are advised to evaluate thresholds with verification data obtained from in-house studies, to set analytical threshold (AT) and stochastic threshold (ST) values appropriate for their respective interpretation criteria.

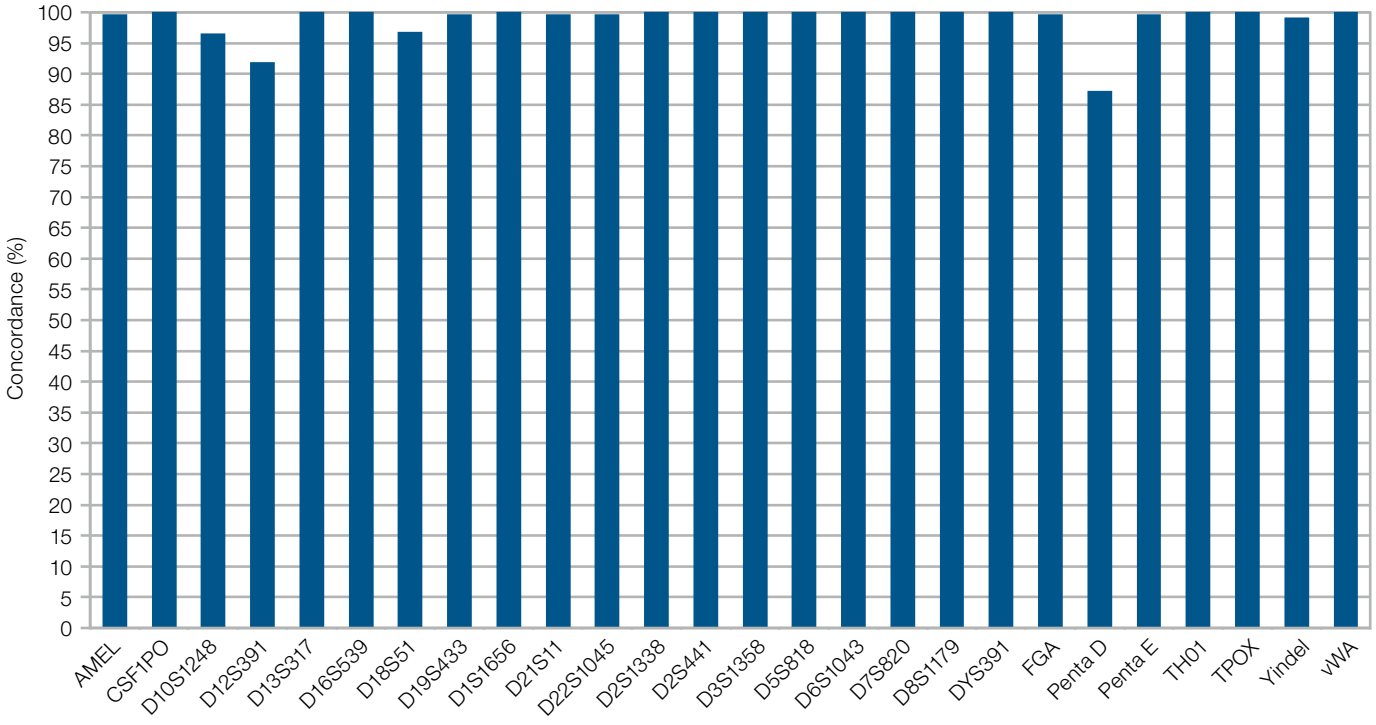

Figure 5. Concordance of the Precision ID GlobalFiler NGS STR Panel v2 data with traditional CE results. Marker concordance was compared to the orthogonal data based on traditional CE analysis using the GlobalFiler PCR Amplification Kit for 320 population samples.

Table 2. Performance of the Precision ID GlobalFiler NGS STR Panel v2. Comparison of NGS genotyping accuracy with orthogonal CE data was obtained for a total of 8,162 markers; a 5% threshold was used to determine false positives, false negatives, and allele concordance.

Data type	N	%
Total CE truth markers	8,162	100.00
NGS false negatives	40	0.49
NGS false positives	45	0.55
Overall NGS concordance	8,077	98.96

Sensitivity and challenging mock casework samples

As described, an early-access version of the Precision ID GlobalFiler NGS STR Panel v2 was provided to five global test sites to run a series of tests with the Precision ID NGS System. In the initial study, the test sites performed a DNA template titration to assess sensitivity in their respective laboratories. For these data, allele detection provided >99.9% accuracy (truth alleles detected) for 1 ng, 250 pg, and 125 pg DNA template inputs (Figure 6).

For each of the five global test sites, a group of 16 challenging nonprobative casework samples was analyzed to measure performance of the panel. Several samples from this study demonstrate the benefits of NGS STR analysis with challenging forensic specimens.

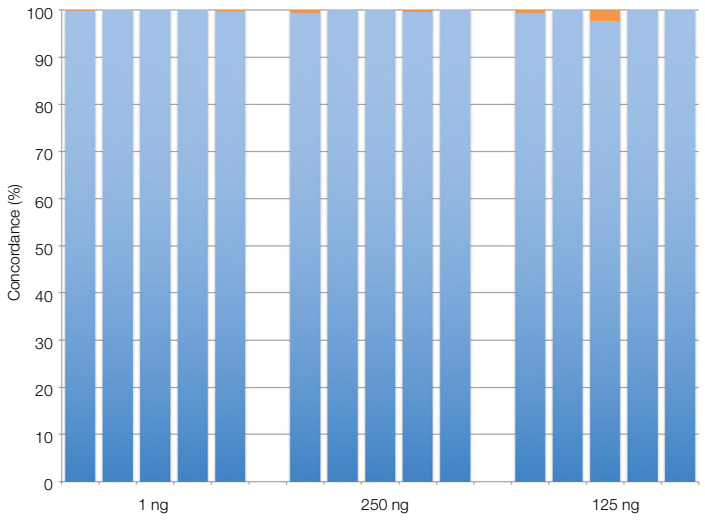

Figure 6. Sensitivity and concordance of the Precision ID GlobalFiler NGS STR Panel v2. gDNA samples of 1 ng, 250 pg, and 125 pg were analyzed with the Precision ID GlobalFiler NGS STR Panel v2 by five global test sites. Data were compared with alleles generated on the 3500 or 3500xL Genetic Analyzer using the GlobalFiler PCR Amplification Kit.

Casework example 1

A vaginal swab stored at 37°C for 2 years provided a full NGS STR profile (Figure 7) concordant with CE STR data and an average allele depth of coverage of 2,725x with corresponding intralocus balance greater than 49% for all markers except D5S2800 (ILB = 36%).

Figure 7. Genotyping results from vaginal swab. A full NGS STR profile concordant with CE STR data was obtained from a vaginal swab stored at 37°C for 2 years.

Casework example 2

This example shows results for contact DNA from a doorknob at a suspected burglary scene (Table 3). The corresponding electropherogram exhibits a partial DNA profile with peak heights at or below stochastic thresholds in most cases and overall low data quality (Figure 8).

The corresponding NGS STR data (Figure 9) provided interpretable DNA genotypes at an additional 12 loci. In this example, NGS analysis could render this particular profile suitable for comparison to CODIS or a known individual reference sample.

Table 3. Comparison of NGS and CE profiles from a doorknob swab. Additional alleles can be recovered by using a hybrid approach for highly degraded samples.

Marker	NGS profile	CE profile
D2S441	10,10	10
AMEL	X,Y	Y
D5S818	12	12
D2S1338	25	25
D3S4529	15	–
D12S391	17,18	–
D14S1434	13	–
D16S539	12	–
D18S51	–	14
D19S433	–	–
D1S1656	17.3	–
D1S1677	14,15	–
D21S11	31.2	–
D22S1045	–	11
D12ATA63	12	–
D2S1776	8,9	–
CSF1PO	10,13	–
D3S1358	16	–
D13S317	11	–
D4S2408	10	–
D5S2800	17	–
D10S1248	14,14	–
D6S1043	11	–
D6S474	–	–
D7S820	11	–
D8S1179	12	11
DYS391	–	–
FGA	22,24	–
TH01	6,9.3	–
TPOX	9	–
vWA	16	16,17

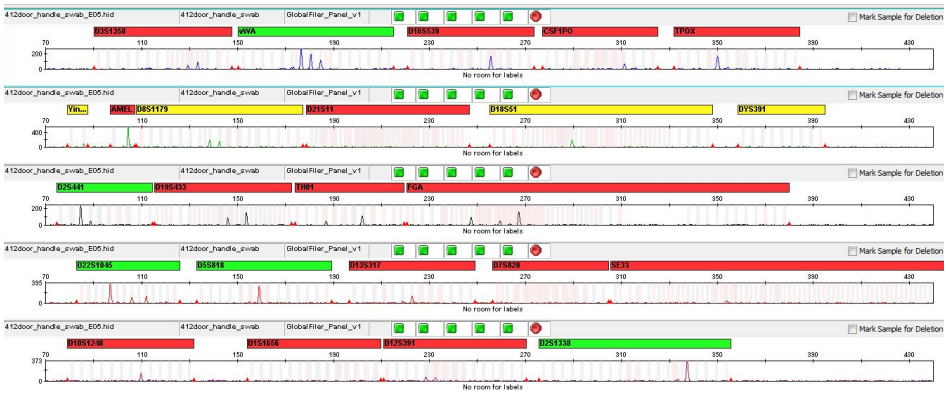

Figure 8. CE genotyping results from a doorknob swab. CE STR analysis recovered a partial DNA profile with peak heights at or below the stochastic thresholds and overall low data quality.

Figure 9. NGS genotyping results from a doorknob swab. For this swab, 0.114 ng of gDNA was analyzed with both systems. The data show partial DNA profiles and increased stochastic sampling effects due to low starting template. Alleles for the NGS profile in Table 3 were hand-curated for final interpretation; a known 0.1 artifact at D12S391 and elevated stutter at CSF1PO were removed from the table.

Casework example 3

In this example, NGS STR results obtained from a bone sample buried in soil for 18 months demonstrate the utility of the small-amplicon panel design. Both CE STR and NGS STR analyses provided concordant, full DNA profiles. However, multiple CE STR loci exhibited off-scale fluorescence data and a significant ski-slope effect, which indicates DNA degradation (Figure 10). The representative marker set from the DNA profile generated with the Precision ID GlobalFiler NGS STR Panel v2 (Figure 11) provided greater intra- and interlocus balance overall. This result demonstrates that the smaller amplicons of the Precision ID GlobalFiler NGS STR Panel v2 are less adversely affected by DNA degradation than the range of larger amplicons in the CE-based GlobalFiler PCR Amplification Kit.

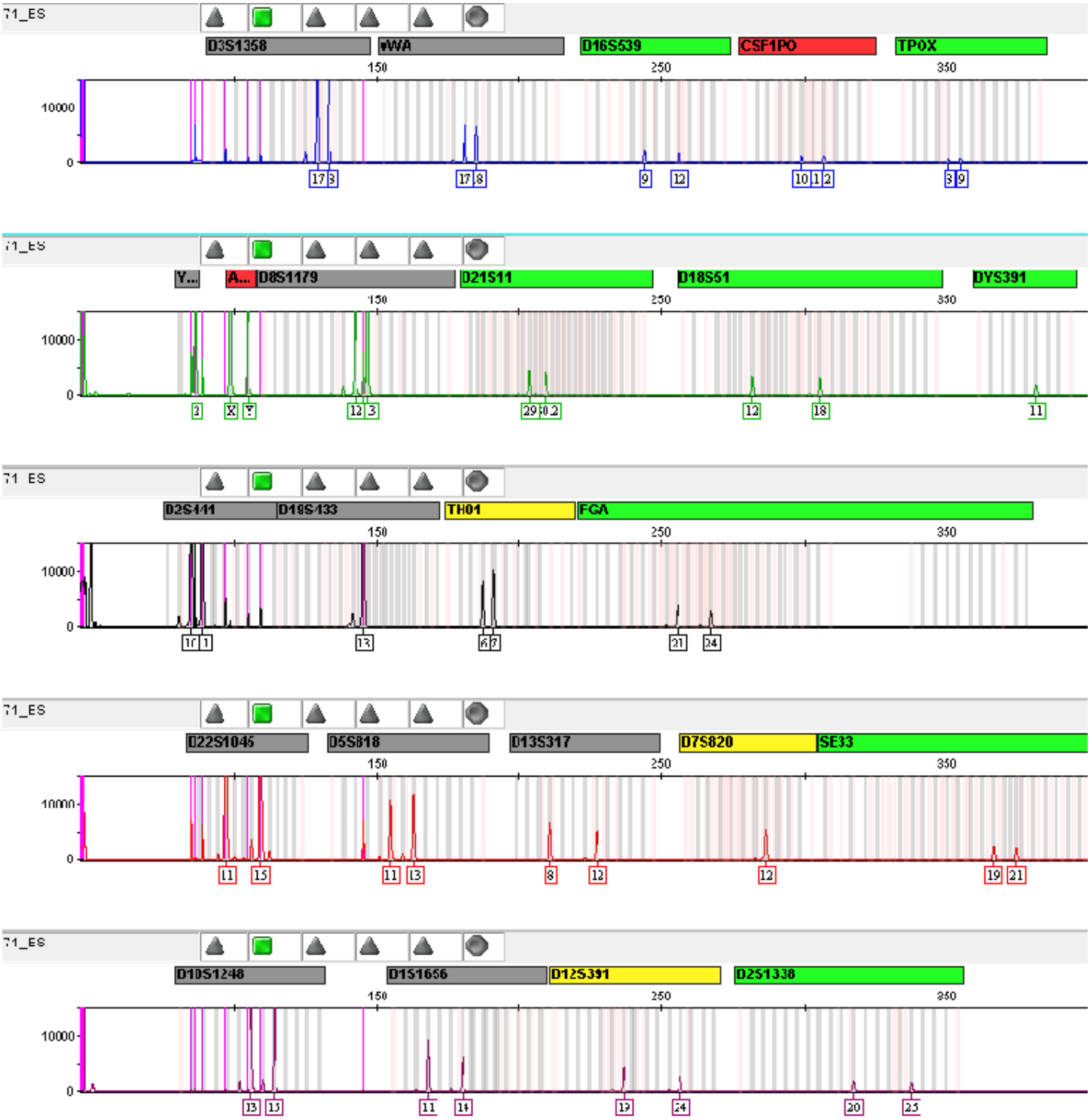

Figure 10. Results obtained from a bone sample buried in soil for 18 months, using GeneMapper *ID-X* Software. The results show off-scale data and the characteristic ski-slope effect observed with degraded samples.

applied
biosystems

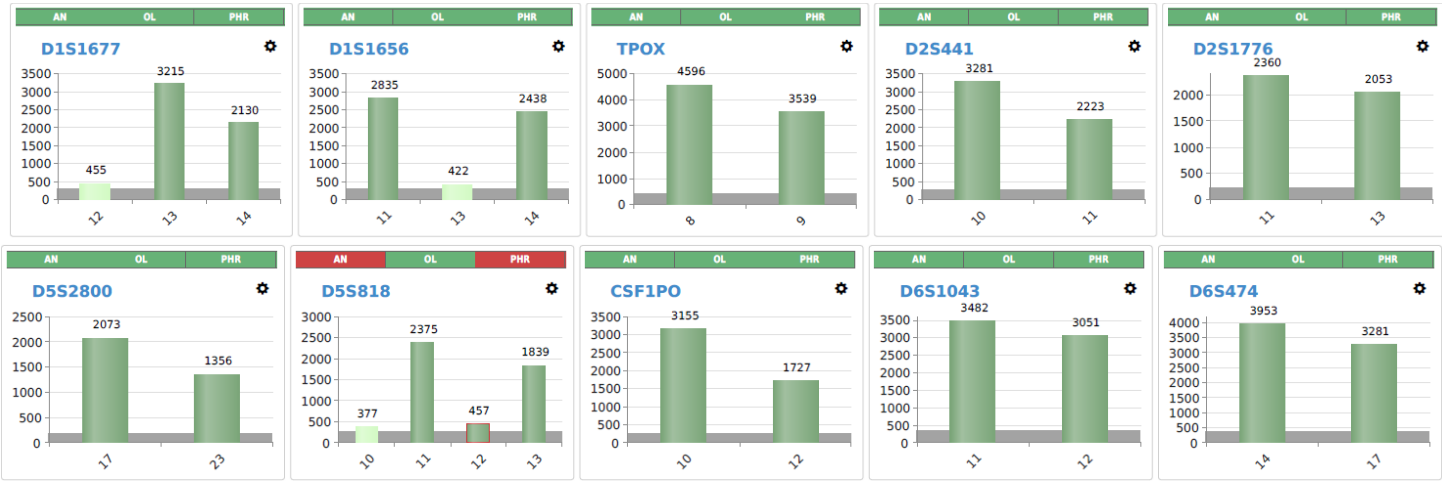

Figure 11. NGS STR results from a bone sample buried in soil for 18 months. A total of 10 NGS STR markers from the full DNA profile were selected to demonstrate increased intralocus balance relative to CE STR data, and substantial read depth of coverage.

Conclusion

The data presented in this study demonstrate the usability of the Precision ID NGS System and Precision ID GlobalFiler NGS STR Panel v2 for routine human identification of forensic casework samples. The Precision ID NGS System reliably generated correct genotypes for the marker set in the panel over a range of starting DNA inputs (1 ng–125 pg) of known references and commonly encountered mock casework sample types.

Mixed, degraded, inhibited, and low-input forensic DNA samples present many challenges for human identification laboratories using the current CE methods. In these instances, NGS may allow for greater allelic recovery and a deeper interrogation of STR amplicons. Full nucleotide sequencing of STR motifs may provide increased discrimination in routine casework analysis. Distinguishing between alleles with the same fragment length but different sequences—commonly referred to as isometric heterozygotes—also enables NGS methods to enhance interpretation approaches for complex mixtures.

The Converge Software with NGS Data Analysis module can detect SNPs in the STR-flanking sequences or within the STR itself to increase allelic resolution even further. Forensic researchers analyzing large population data sets can provide population frequencies for the expanded NGS STRs and the newly discovered sequence-based alleles to better understand the increased discrimination of NGS STR markers. These data can lay the foundation for relevant match statistics that could potentially be used in forensic laboratories and could further define the utility of sequence-based STR analysis in routine interpretation of casework samples [5].

References

1. HIDS 2016 presentations and posters.
2. HIDS 2017 presentations and posters.
3. HIDS 2018 presentations and posters.
4. Müller P, Alonso A, Barrio PA et al. (2018) Systematic evaluation of the early access Applied Biosystems Precision ID GlobalFiler Mixture ID and GlobalFiler NGS STR Panels for the Ion S5 System. *Forensic Sci Int Genet* 36:95–103.
5. Gettings KB, Borsuk LA, Steffen CR et al. (2018) Sequence-based U.S. population data for 27 autosomal STR loci. *Forensic Sci Int Genet* 37:106–115.

Find out more at **thermofisher.com/hid-ngs**

ThermoFisher SCIENTIFIC