October 16, 2019

TECHNICAL NOTE

Performance of the Precision ID GlobalFiler^{TM} NGS STR Panel v2: Artifacts, Thresholds and Chip Loading

Artifacts are a known phenomenon of DNA typing and may interfere with analysis and interpretation of capillary electrophoresis and next generation sequencing (NGS) data. Thresholds are applied to data to encourage generation of accurate and reproducible results, and should be determined through empirical testing. Thresholds may also be established differently for reference samples and casework samples. To assist with NGS STR data interpretation, this technical note provides characterization of known NGS artifacts, guidance on analytical threshold values, and reviews reported variability in chip loading efficiency following the commercial release of the Precision ID GlobalFiler^{TM} NGS STR Panel v2.

A. Observed Sequence Artifacts

Thermo Fisher Scientific conducted population studies with a set of 320 known reference population samples from diverse populations to assess key performance metrics, including depth of coverage, intralocus balance, sensitivity and genotyping concordance with CE STR data.[1] For the population samples, representative CE data were available for a total of 8,162 markers, of which 8,077 were concordant with NGS (98.96%) (Table 1 and Figure 1).

Table 1. Performance of the Precision ID GlobalFiler NGS STR Panel v2

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

A total of 45 artifacts and an additional 40 dropouts were identified for alleles using a 5% global cutoff for analytical and stochastic thresholds (AT and ST).

- Of the 45 false positives observed, the most affected STRs were: (1) D12S391 and D10S1248 (base insertions) and (2) Penta D, Penta E, and D18S51 (single-base indels and nonreproducible insertions). In addition, five of the observed discordances resulted from indels and/or single-nucleotide polymorphisms (SNPs) adjacent to the STR repeat in flanking-region sequences.
- All dropouts (false negatives) occurred at the Penta D locus due to a 13 bp deletion adjacent to the start of the STR repeat structure for alleles 2.2 or 3.2; this characterized deletion occurs at a frequency of 11% in the African American population.

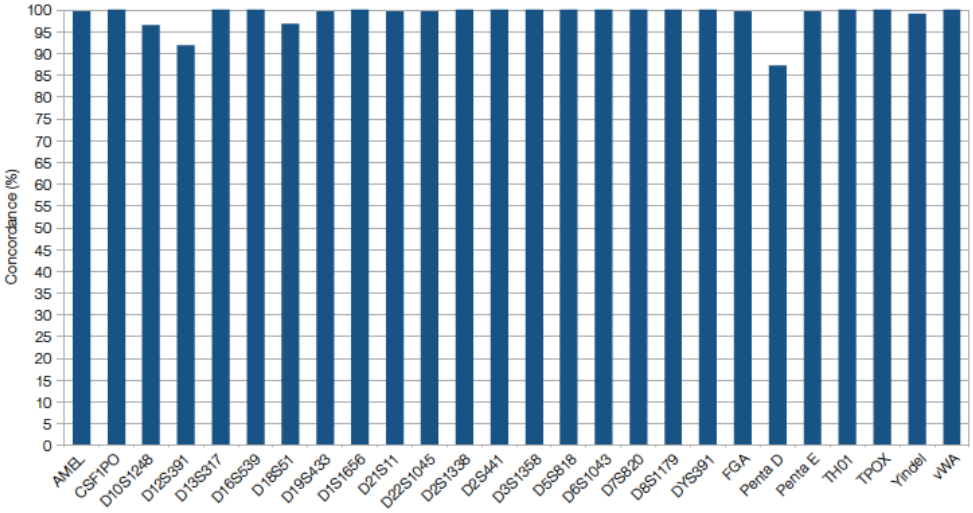


Figure 1. Percent genotyping concordance for NGS results of the 320 population samples (8,162 markers) as compared to traditional CE results

Closer inspection of the sequence data underlying the allele designations allowed for categorization of the artifacts in the data set. The majority of the observed artifacts were single base differences from the truth allele and resulted from either misincorporation of nucleotides in homopolymer stretches or stuttering effects that occur when templated DNA is adhered to Ion Sphere Particles (ISP's). Table 2 and Figures 2-6 provide more detailed information for the various sequence artifact types detected in this study and the STR markers impacted.

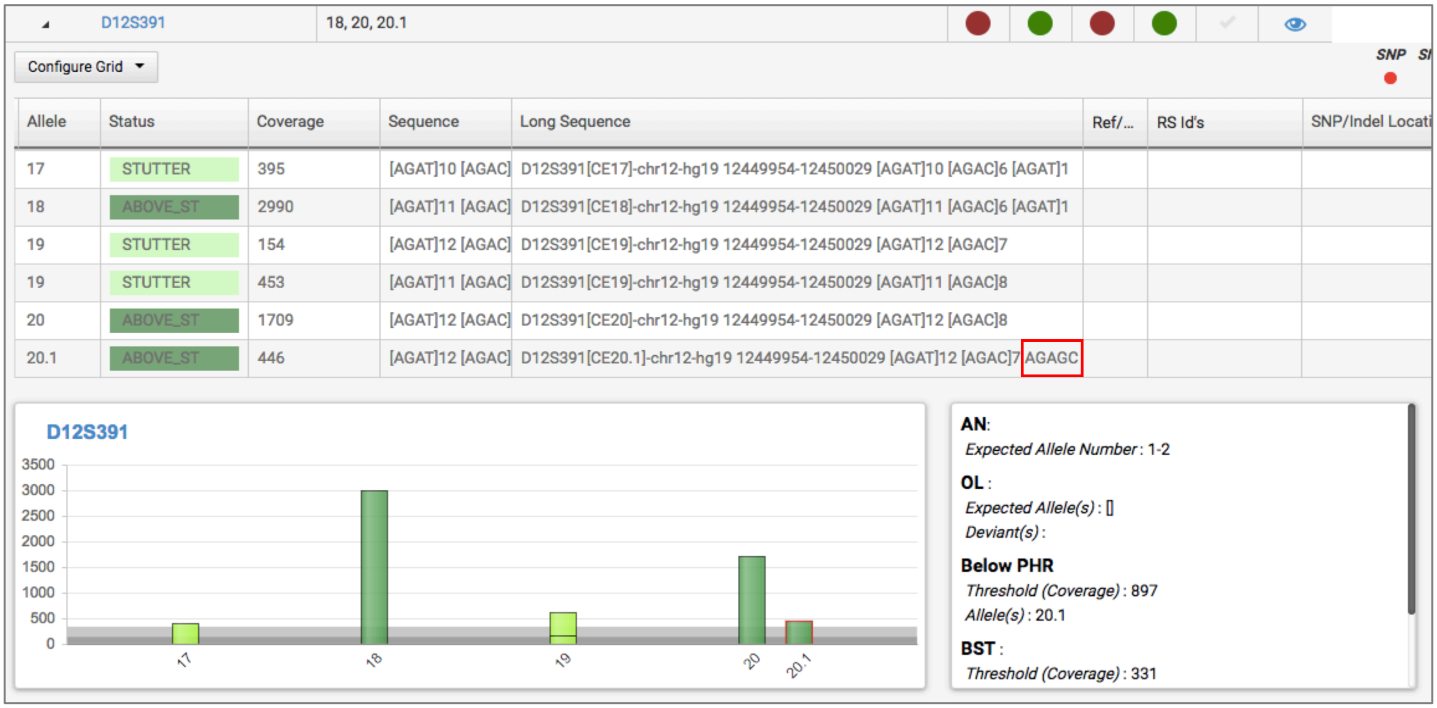
Table 2. Reported sequence artifacts above 5% AT

Artifact Type	Marker(s)	Allele(s) Observed with Artifact	Example Sequence	Example Artifact
Insertion G	D12S391	17.1, 20.1, 21.1, 25.1	[AGAT]11 [AGAC]8 AGA**G**C	20.1
Insertion GA	D10S1248	12.2	**GA** [GGAA]12	12.2
Overcall and Undercall A	D18S51	11.1, 12, 12.1, 13.1, 13.3, 14.1,15.2, 15.3, 19.3	[AGAA]1 **A**[AGAA]11 [AGAA]2 **A**GAA[AGAA]11	12.1 13.3
Overcall A	D12ATA63	18.1	[TAA]15 [CAA]2 **A**[CAA]1	18.1
	Penta E	16.1	[AAAGA]3 **A**[AAAGA]13	16.1
Overcall T	FGA	25.1	[TTTC]3 TTTT TTCT [CTTT]11 CCTT **T** [CTTT]5 CTCC [TTCC]2	25.1
Y allele in female sample	AMEL Y	Y	NA	Y
	Y indel	1,2		1,2

Reported sequence artifacts. The table shows artifact types observed with the corresponding markers affected by the sequence variants. The undercall (highlighted in **red**) at D18S51 shows the unincorporated base resulting in an artifact allele one base shorter than the truth allele. Whereas, the overcalls (highlighted in **blue**) show misincorporations resulting in artifact alleles one base longer than the truth allele (with exception of the two nucleotide base GA insertion at D10S1248).

Figures 2-6. Representative examples of observed STR artifacts

Insertion G (D12S391)



D12S391 | 18, 20, 20.1

Configure Grid ▾

Allele	Status	Coverage	Sequence	Long Sequence	Ref/...	RS Id's	SNP/Indel Locati
17	STUTTER	395	[AGAT]10 [AGAC]	D12S391[CE17]-chr12-hg19 12449954-12450029 [AGAT]10 [AGAC]6 [AGAT]1			
18	ABOVE_ST	2990	[AGAT]11 [AGAC]	D12S391[CE18]-chr12-hg19 12449954-12450029 [AGAT]11 [AGAC]6 [AGAT]1			
19	STUTTER	154	[AGAT]12 [AGAC]	D12S391[CE19]-chr12-hg19 12449954-12450029 [AGAT]12 [AGAC]7			
19	STUTTER	453	[AGAT]11 [AGAC]	D12S391[CE19]-chr12-hg19 12449954-12450029 [AGAT]11 [AGAC]8			
20	ABOVE_ST	1709	[AGAT]12 [AGAC]	D12S391[CE20]-chr12-hg19 12449954-12450029 [AGAT]12 [AGAC]8			
20.1	ABOVE_ST	446	[AGAT]12 [AGAC]	D12S391[CE20.1]-chr12-hg19 12449954-12450029 [AGAT]12 [AGAC]7 AGAGC			

AN:
Expected Allele Number : 1-2

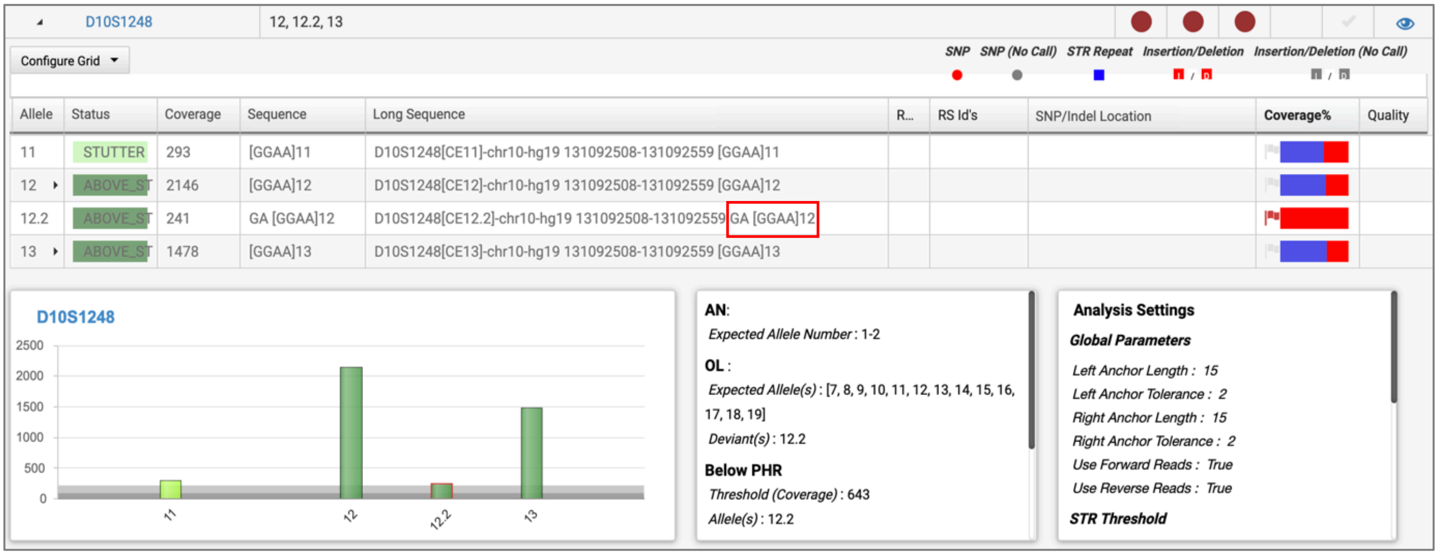
OL :
Expected Allele(s) : []
Deviant(s) :

Below PHR
Threshold (Coverage) : 897
Allele(s) : 20.1

BST :
Threshold (Coverage) : 331

Figure 2. An insertion of nucleotide base G at D12S391 resulting in the 20.1 artifact allele.

Insertion GA (D10S1248)



D10S1248 | 12, 12.2, 13

Configure Grid ▾

SNP ● SNP (No Call) ● STR Repeat ■ Insertion/Deletion I / D Insertion/Deletion (No Call) I / D

Allele	Status	Coverage	Sequence	Long Sequence	R...	RS Id's	SNP/Indel Location	Coverage%	Quality
11	STUTTER	293	[GGAA]11	D10S1248[CE11]-chr10-hg19 131092508-131092559 [GGAA]11					
12 ▸	ABOVE_ST	2146	[GGAA]12	D10S1248[CE12]-chr10-hg19 131092508-131092559 [GGAA]12					
12.2	ABOVE_ST	241	GA [GGAA]12	D10S1248[CE12.2]-chr10-hg19 131092508-131092559 GA [GGAA]12					
13 ▸	ABOVE_ST	1478	[GGAA]13	D10S1248[CE13]-chr10-hg19 131092508-131092559 [GGAA]13					

AN:
Expected Allele Number : 1-2

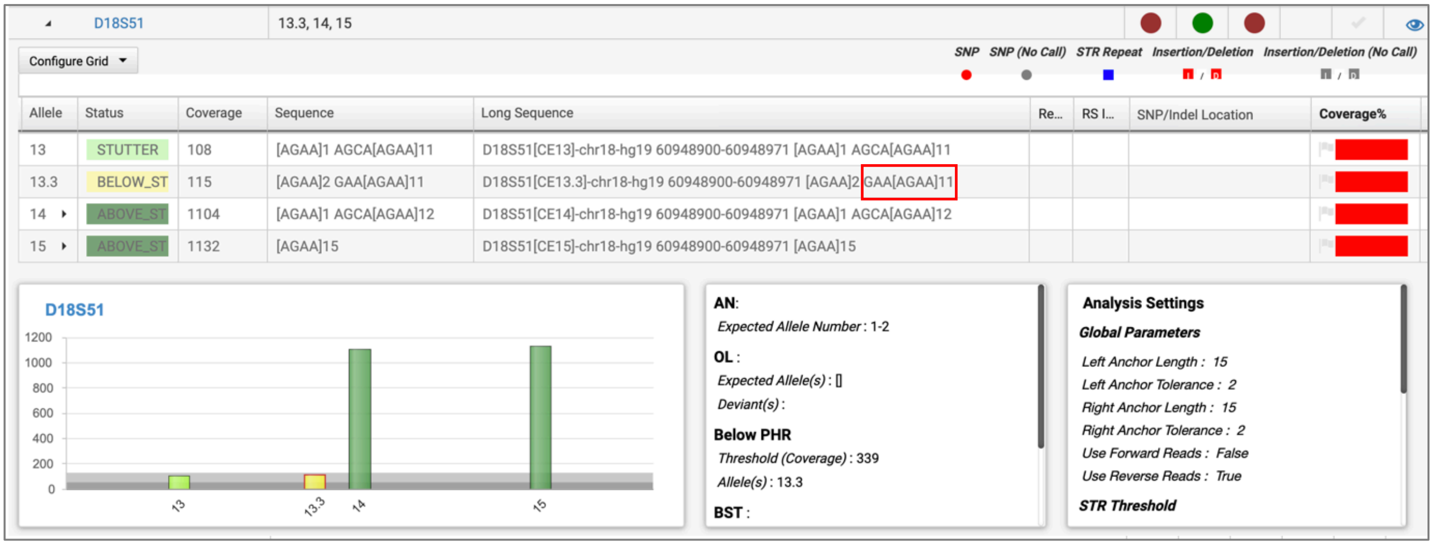
OL :
Expected Allele(s) : [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]
Deviant(s) : 12.2

Below PHR
Threshold (Coverage) : 643
Allele(s) : 12.2

Analysis Settings
Global Parameters
Left Anchor Length : 15
Left Anchor Tolerance : 2
Right Anchor Length : 15
Right Anchor Tolerance : 2
Use Forward Reads : True
Use Reverse Reads : True
STR Threshold

Figure 3. An insertion of nucleotide bases GA at D10S1248 resulting in the 12.2 allele artifact.

Undercall A (D18S51)



D18S51
13.3, 14, 15
Configure Grid
SNP ● SNP (No Call) ● STR Repeat ■ Insertion/Deletion Insertion/Deletion (No Call)
D18S51
1200
1000
800
600
400
200
0
13
13.3
14
15

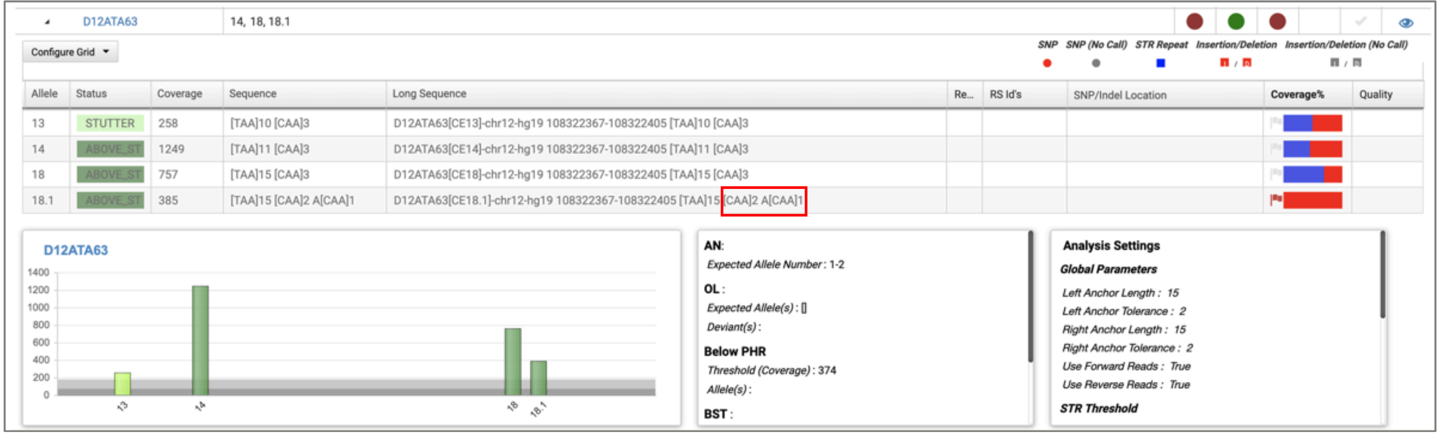
Allele	Status	Coverage	Sequence	Long Sequence	Re...	RS I...	SNP/Indel Location	Coverage%
13	STUTTER	108	[AGAA]1 AGCA[AGAA]11	D18S51[CE13]-chr18-hg19 60948900-60948971 [AGAA]1 AGCA[AGAA]11				
13.3	BELOW_ST	115	[AGAA]2 GAA[AGAA]11	D18S51[CE13.3]-chr18-hg19 60948900-60948971 [AGAA]2 GAA[AGAA]11				
14	ABOVE_ST	1104	[AGAA]1 AGCA[AGAA]12	D18S51[CE14]-chr18-hg19 60948900-60948971 [AGAA]1 AGCA[AGAA]12				
15	ABOVE_ST	1132	[AGAA]15	D18S51[CE15]-chr18-hg19 60948900-60948971 [AGAA]15				

AN:
Expected Allele Number : 1-2
OL :
Expected Allele(s) : []
Deviant(s) :
Below PHR
Threshold (Coverage) : 339
Allele(s) : 13.3
BST :

Analysis Settings
Global Parameters
Left Anchor Length : 15
Left Anchor Tolerance : 2
Right Anchor Length : 15
Right Anchor Tolerance : 2
Use Forward Reads : False
Use Reverse Reads : True
STR Threshold

Figure 4. An undercall of nucleotide base A at D18S51 resulting in the 13.3 artifact allele.

Overcall A (D12ATA63)



D12ATA63
14, 18, 18.1
Configure Grid
SNP ● SNP (No Call) ● STR Repeat ■ Insertion/Deletion Insertion/Deletion (No Call)
D12ATA63
1400
1200
1000
800
600
400
200
0
13
14
18
18.1

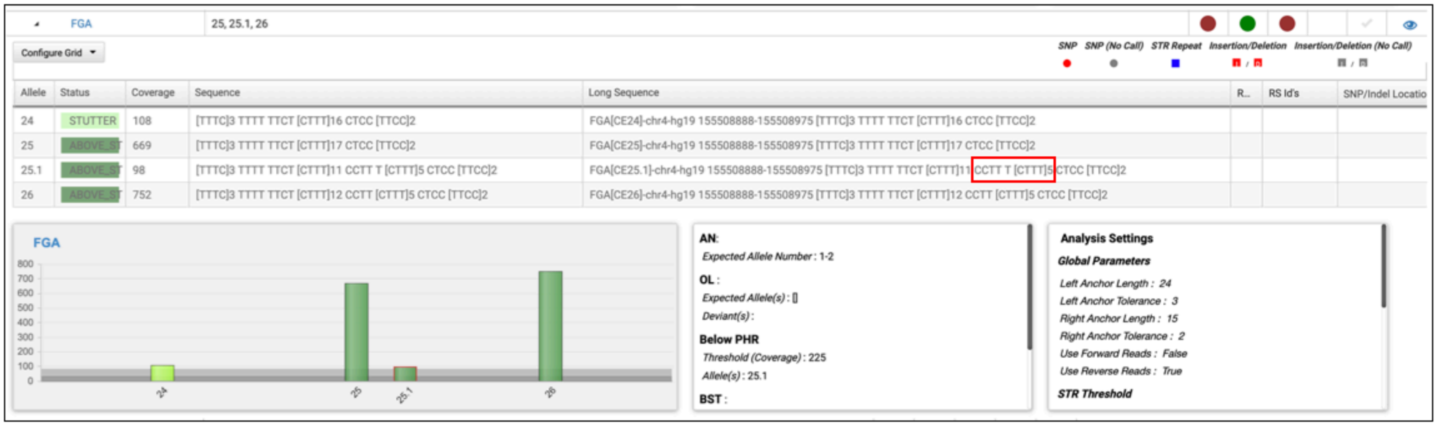
Allele	Status	Coverage	Sequence	Long Sequence	Re...	RS Id's	SNP/Indel Location	Coverage%	Quality
13	STUTTER	258	[TAA]10 [CAA]3	D12ATA63[CE13]-chr12-hg19 108322367-108322405 [TAA]10 [CAA]3					
14	ABOVE_ST	1249	[TAA]11 [CAA]3	D12ATA63[CE14]-chr12-hg19 108322367-108322405 [TAA]11 [CAA]3					
18	ABOVE_ST	757	[TAA]15 [CAA]3	D12ATA63[CE18]-chr12-hg19 108322367-108322405 [TAA]15 [CAA]3					
18.1	ABOVE_ST	385	[TAA]15 [CAA]2 A[CAA]1	D12ATA63[CE18.1]-chr12-hg19 108322367-108322405 [TAA]15 [CAA]2 A[CAA]1					

AN:
Expected Allele Number : 1-2
OL :
Expected Allele(s) : []
Deviant(s) :
Below PHR
Threshold (Coverage) : 374
Allele(s) :
BST :

Analysis Settings
Global Parameters
Left Anchor Length : 15
Left Anchor Tolerance : 2
Right Anchor Length : 15
Right Anchor Tolerance : 2
Use Forward Reads : True
Use Reverse Reads : True
STR Threshold

Figure 5. An overcall of nucleotide base A at D12ATA63 resulting in the 18.1 artifact allele.

Overcall T (FGA)



FGA
25, 25.1, 26
Configure Grid
SNP ● SNP (No Call) ● STR Repeat ■ Insertion/Deletion Insertion/Deletion (No Call)
FGA
800
700
600
500
400
300
200
100
0
24
25
25.1
26

Allele	Status	Coverage	Sequence	Long Sequence	R...	RS Id's	SNP/Indel Locatio
24	STUTTER	108	[TTTC]3 TTTT TTCT [CTTT]16 CTCC [TTCC]2	FGA[CE24]-chr4-hg19 155508888-155508975 [TTTC]3 TTTT TTCT [CTTT]16 CTCC [TTCC]2			
25	ABOVE_ST	669	[TTTC]3 TTTT TTCT [CTTT]17 CTCC [TTCC]2	FGA[CE25]-chr4-hg19 155508888-155508975 [TTTC]3 TTTT TTCT [CTTT]17 CTCC [TTCC]2			
25.1	ABOVE_ST	98	[TTTC]3 TTTT TTCT [CTTT]11 CCTT T [CTTT]5 CTCC [TTCC]2	FGA[CE25.1]-chr4-hg19 155508888-155508975 [TTTC]3 TTTT TTCT [CTTT]11 CCTT T [CTTT]5 CTCC [TTCC]2			
26	ABOVE_ST	752	[TTTC]3 TTTT TTCT [CTTT]12 CCTT [CTTT]5 CTCC [TTCC]2	FGA[CE26]-chr4-hg19 155508888-155508975 [TTTC]3 TTTT TTCT [CTTT]12 CCTT [CTTT]5 CTCC [TTCC]2			

AN:
Expected Allele Number : 1-2
OL :
Expected Allele(s) : []
Deviant(s) :
Below PHR
Threshold (Coverage) : 225
Allele(s) : 25.1
BST :

Analysis Settings
Global Parameters
Left Anchor Length : 24
Left Anchor Tolerance : 3
Right Anchor Length : 15
Right Anchor Tolerance : 2
Use Forward Reads : False
Use Reverse Reads : True
STR Threshold

Figure 6. An overcall of nucleotide base T at FGA resulting in the 25.1 artifact allele.

As shown in the examples above, the truth allele is the predominant call and the corresponding sequence artifacts fall within one base from the parent allele (except for the GA insertion at D10S1248, which is a two nucleotide base difference). For several of the examples highlighted, the artifacts fall near the AT or ST, with exception of D12S391 and D12ATA63 where the sequence artifacts are approximately 26% and 50% of the parent allele, respectively. When analyzing sequence data output, laboratories are advised to pay particular attention to the markers listed above as they may be prone to these sequence artifacts with some regularity.

B. Analytical Threshold

A thorough analysis of system noise and sequence artifacts was performed in order to refine analytical thresholds for the system. In this study, a set of population samples (N = 568) from Thermo Fisher Scientific (including the 320 population samples described above) and a collaborator site were processed at 1ng total DNA input and analyzed with the commercial panel using recommended analysis settings. For the study, the AT was set to remove 99% of the artifacts for each marker. With this setting, ≤1% of the alleles observed above the AT would be expected to result from non-truth alleles (e.g., noise or artifacts) assuming inter-run and population specific variation were captured in the data set. Figure 7 shows the artifacts by marker expressed as a percent of the total marker coverage.

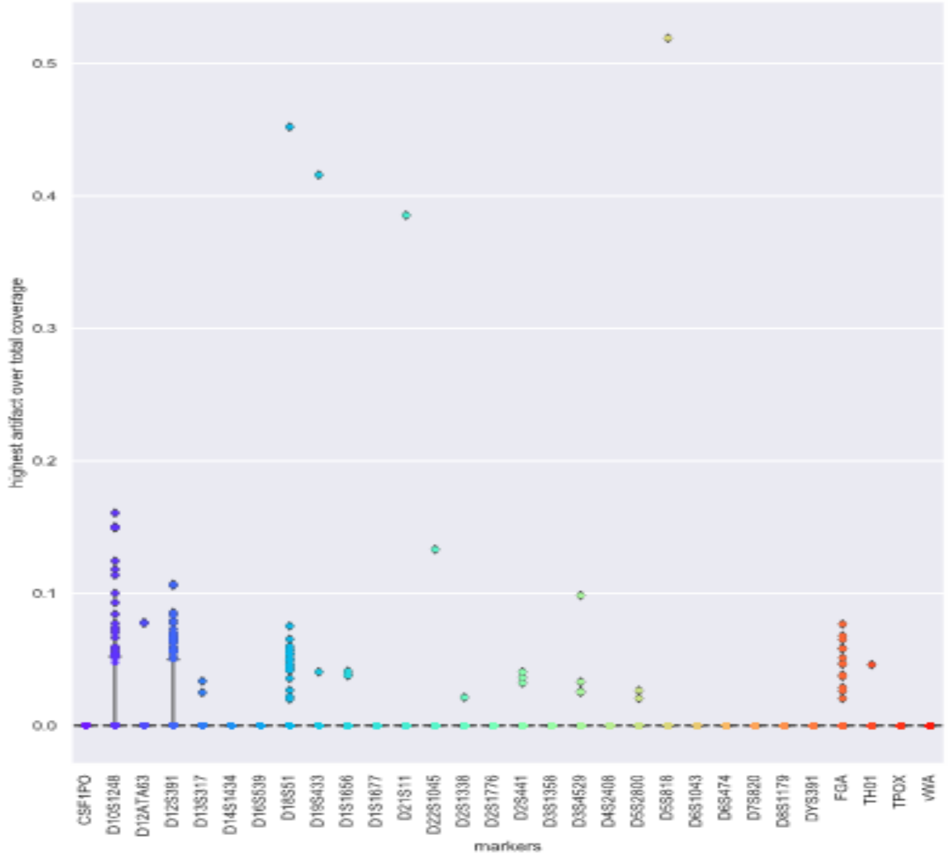


Figure 7. Analysis of Artifacts on a Per Marker Basis
(Analytical Threshold)

Based on this analysis, AT values were adjusted for the markers listed as shown in Table 3. These values are also reflected in an updated .json analysis parameters file[2] available from your Field Application Scientist upon request. (Note: ST values were also adjusted to match the AT's in Table 3).

Table 3. Adjusted AT values

Marker	AT
D10S1248	0.12
D12S391	0.12
D18S51	0.06
D1S1677	0.06
D22S1045	0.06
D5S2800	0.06
FGA	0.06

For the markers listed in Table 3, higher AT values provide greater specificity for the detection of truth alleles and the removal of observed sequence artifacts. Figure 8 demonstrates the use of the elevated AT value at D10S1248.

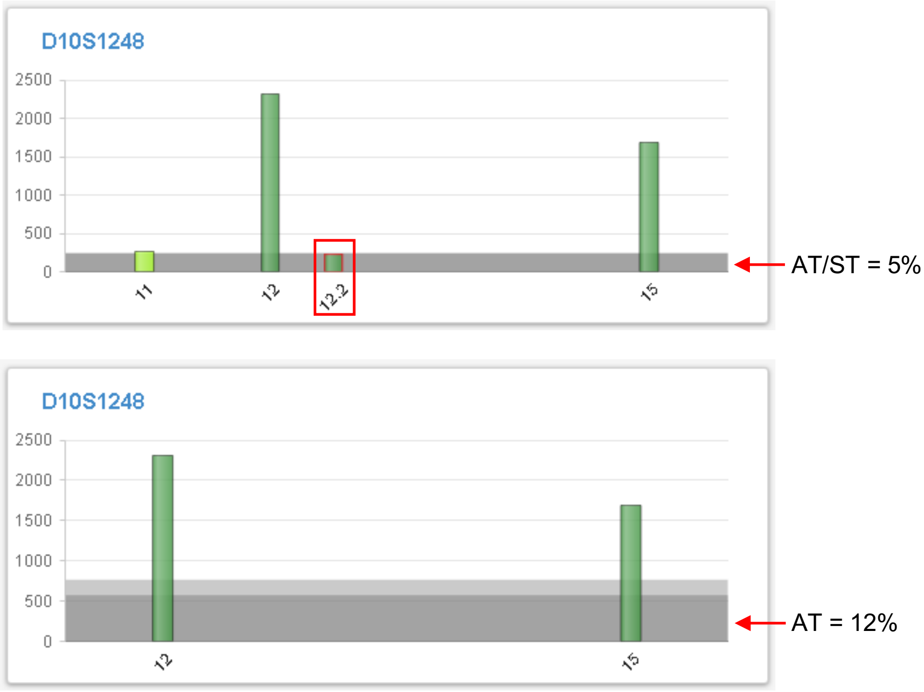


Figure 8. Increased AT to remove sequence artifacts

The increased marker-specific AT values are designed to improve accuracy for single source samples. However, there is an inherent trade-off in sensitivity when using this approach. The ability to detect low level contributors in sensitivity, mixture, and mock casework studies will be diminished with these thresholds. Laboratories should carefully consider the use of the elevated thresholds for studies other than population sample testing and single source evaluations. Individual results for AT and ST may vary; thresholds should be derived empirically to match system performance based on previously characterized samples and the laboratory's validation guidelines and interpretation criteria[3].

C. Chip loading variability

With the current Ion Chef loading script in Torrent Suite Software v5.10, optimal chip loading density with GlobalFiler NGS STR Panel v2 falls within a ~50-70% range and exhibits the 'normal' chip image (Figure 9, left panel). Some Ion Chef instruments have been reported to exhibit decreased loading densities and a visible distortion of the normal loading pattern from slight to severe 'tadpole' pattern on the chip image with the STR panel. Presence of this tadpole phenotype indicates that the affected area on the chip does not contain templated Ion Sphere Particles (ISPs). Depending on the severity of the tadpole, there may be an impact on total usable reads and genotype data quality. Customers are advised to review chip loading densities, loading patterns and control genotype results to monitor Chef loading performance. If this issue is seen with regularity or there appears to be an impact in genotype data quality, please contact your Field Application Specialist (FAS) or HID Technical Support for more information. To date, there have been no reported tadpole observations with the Precision ID Mitochondrial DNA panels (Catalog Number A30938 and A31443) or Ancestry and Identity SNP panels (Catalog Number A25642 and A25643).

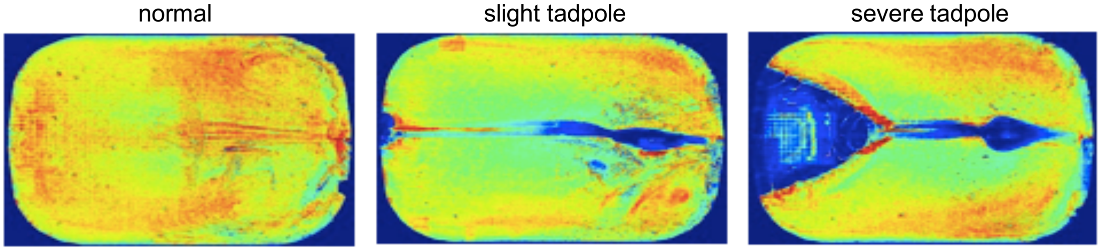


Figure 9. Chip images

Conclusions

Thermo Fisher Scientific's Precision ID NGS STR v2 and Converge^{TM} analysis solution is subjected to rigorous development specifications and criteria to balance the overall performance of the panel (e.g., coverage, locus balance) while minimizing artifacts across a variety of informative STR markers. However, the complexity of sequence analysis of STR's and population sample diversity has revealed a small percentage of artifacts not previously encountered during initial kit development. Understanding the presence of such artifacts can help labs to develop interpretation criteria for future implementation of this panel.

This document will be periodically updated with additional reports and characterized artifacts to assist in the analysis and interpretation of GlobalFiler NGS STR results.

References

1. Thermo Fisher Scientific Application Note "Get more information from challenging samples with next-generation sequencing of short tandem repeats". COL32762 1118.
2. Precision_ID_GlobalFiler_NGS_STR_Panel_AnalysisParams_v2.1.1. *Available upon request from local FAS.*
3. Massively parallel sequence data of 31 autosomal STR loci from 496 Spanish individuals revealed concordance with CE-STR technology and enhanced discrimination power. Barrio, Pedro A. et al. Forensic Science International: Genetics, Volume 42, 49 – 55.

Revision History

Revision	Date	Description
A	10.16.19	Initial publication.